WeData Data Development Platform Operation Guide



©2013–2025 Tencent Cloud. All rights reserved.

Copyright Notice

©2013-2025 Tencent Cloud. All rights reserved.

The complete copyright of this document, including all text, data, images, and other content, is solely and exclusively owned by Tencent Cloud Computing (Beijing) Co., Ltd. ("Tencent Cloud"); Without prior explicit written permission from Tencent Cloud, no entity shall reproduce, modify, use, plagiarize, or disseminate the entire or partial content of this document in any form. Such actions constitute an infringement of Tencent Cloud's copyright, and Tencent Cloud will take legal measures to pursue liability under the applicable laws.

Trademark Notice

Tencent Cloud

This trademark and its related service trademarks are owned by Tencent Cloud Computing (Beijing) Co., Ltd. and its affiliated companies("Tencent Cloud"). The trademarks of third parties mentioned in this document are the property of their respective owners under the applicable laws. Without the written permission of Tencent Cloud and the relevant trademark rights owners, no entity shall use, reproduce, modify, disseminate, or copy the trademarks as mentioned above in any way. Any such actions will constitute an infringement of Tencent Cloud's and the relevant owners' trademark rights, and Tencent Cloud will take legal measures to pursue liability under the applicable laws.

Service Notice

This document provides an overview of the as-is details of Tencent Cloud's products and services in their entirety or part. The descriptions of certain products and services may be subject to adjustments from time to time.

The commercial contract concluded by you and Tencent Cloud will provide the specific types of Tencent Cloud products and services you purchase and the service standards. Unless otherwise agreed upon by both parties, Tencent Cloud does not make any explicit or implied commitments or warranties regarding the content of this document.

Contact Us

We are committed to providing personalized pre-sales consultation and technical after-sale support. Don't hesitate to contact us at 4009100100 or 95716 for any inquiries or concerns.

Contents

Operation Guide
- Console operation
  - Console Overview
  - Overview Page
  - Project List
  - Execute Resource Group configuration
    - Resource Group List
    - Scheduling Resource Group Configuration
    - Integrated Resource Group Configuration
    - Data Service Resource Group
    - Stream Compute Resource Group
    - Data Security Resource Group Configuration
    - Resource Group Configuring Public Network
  - Approval management
    - My Approvals
    - My Applications
- Project Management
  - Simple mode and standard mode
  - Basic Information Configuration
  - Member and Role Management
  - Data Source Management
  - Project Execution Resource Group
    - Scheduling resource
    - Integration resource
    - Data Service Resources
    - Stream Compute Resource Group
  - Computational Storage Engine Configuration
  - Variable Settings
- Data Integration
  - Overview of DataInLong
  - DataInLong Quick Start
    - MySQL Real-time Sync to DLC Guide
    - Guide for Real-time Synchronization from MySQL to TChouse-D
    - MySQL Offline Synchronization to Hive Guide
    - MySQL Offline Synchronization to Hive Guidance Documentation
  - Supported data sources and read-write capability
  - Integrated resource configuration and management
    - Integrated Resource Billing Instructions
    - Integration of resource purchase and configuration
    - Integration resource connectivity and usage planning
    - Integrate Resource Group to configure the public network
  - Real-time synchronization task configuration and operation and maintenance
    - Data sources supported by real-time synchronization
      - Data source overview
      - Data Source List
        - MySQL Data Source
        - TDSQL-C for MySQL Data Source
        - Oracle Data Source
        - PostgreSQL Data Source
        - SQL Server Data Source
        - Kafka Data Source

©2013-2025 Tencent Cloud. All rights reserved.

DTS-Kafka Data Source
Mongodb Data Source
TDSQL for MySQL Data Source
OceanBase Data Source
Dameng DM Data Source
TiDB Data Source
Doris Data Source
StarRocks Data Source
DLC Data Source
Iceberg Data Source
TCHouse-P Data Source
ElasticSearch Data Source
ClickHouse Data Source
Greenplum Data Source
Hive Data Source
HBase Data Source
TBase Data Source
HDFS Data Source
Whole Database Synchronization Task Configuration
Data sources supported by whole database synchronization
Overview of Whole Database Task Configuration
Supported DML and DDL Operations
Single table sync task configuration
Data sources supported by Single Table Synchronization
Overview of Single Table Task Configuration
Conversion Node Configuration
Configuration of log collection task
Log Collection Configuration Overview
TKE Source Collection Task Configuration
SDK Source Collection Task Configuration
CVM Source Collection Task Configuration
Collector Management
Real-time synchronization and operation and maintenance
Real-Time Operations and Maintenance
OPS Operations
Metrics Statistics
Key Events
Viewing running logs
Alarm Subscription
Operation History
Monitoring alarm
Rule Management
Alarm record
Real-time Node Advanced Parameters
Offline synchronization task configure and operation and maintenance
Supported Data Sources for Offline Synchronization
Data Source Overview
Supported Data Sources for Offline Synchronization
MySQL Data Source
TDSQL-C for MySQL Data Source
Oracle Data Source
SQL Server Data Source
PostgreSQL Data Source

©2013-2025 Tencent Cloud. All rights reserved.

TDSQL for PostgreSQL Data Source
TDSQL for MySQL Data Source
DB2 Data Source
Dameng DM Data Source
OceanBase Data Source
SAP HANA Data Source
SAP IQ (Sybase) Data Source
Doris/TCHouse-D Data Source
StarRocks Data Source
DLC Data Source
Iceberg Data Source
TCHouse-P Data Source
ClickHouse Data Source
Greenplum Data Source
Hive Data Source
HDFS Data Source
HBase Data Source
TBase Data Source
GBase Data Source
GaussDB Data Source
Impala Data Source
Kudu Data Source
BigQuery Data Source
FTP Data Source
SFTP Data Source
COS Data Source
Rest API Data Source
Elasticsearch/Tencent Cloud Elasticsearch Data Source
Mongo/Tencent Cloud Mongo Data Source
Redis Data Source
Kafka/CKafka data source
Conversion Node Configuration
Time Parameter Description
Offline node advanced parameters
Automatic Table Creation Capability in Synchronization Tasks
Practice Tutorial
Sync MySQL Sharding to Hive
SQL Server Offline Data Synchronization to Hive Partition
MySQL + gh-ost Real-time Synchronization to Kafka
FAQs
Common Issues with Real-Time Sync
Offline Synchronization Frequently Asked Questions
Others
Data Development
Data Development Overview
Data Development feature guidance
Feature Index
Overall Data Development Feature Index
Data Development Workflow Feature Index
Quick Start on Data Development
Engine Usage Instructions
Elastic MapReduce (EMR)
Using Data Lake Compute (DLC)

©2013-2025 Tencent Cloud. All rights reserved.

Tencent Cloud Data Warehouse TCHouse-P
Data Management
Task Development
Task Development Introduction
Cycle Workflow Universal Development Process
Task Node Type
Data Integration
EMR
Hive SQL
Spark SQL
PySpark
Spark
Impala
Trino
DLC
DLC SQL
DLC PySpark
TCHouse-P
TDSQL for PostgreSQL
General Node
Branch Node
Merge Node
JDBC SQL
Shell
Other nodes
Cross-workflow node
Kettle
Task Scheduling Configuration
Basic Configuration
Event Scheduling
Parameter Transfer
Advanced Settings
Scheduling Parameters
Function Development
Resource Management
Event Management
Development Space
Overview
Development Space feature
Script Development
Collaboration between Development Space and Orchestration Space
Recycle Bin
Task Publishing
Cross-project Cloning
Release Center
Task OPS
Operation Dashboard
Task OPS
OPS of Instances
Data Supplement
Task Alert
SQL Tag Statement
Variable Usage

©2013-2025 Tencent Cloud. All rights reserved.

Batch operation
Data Analysis
SQL Explore
Notebook Exploration
Data Assets
Data Assets Overview
Data map
Data Overview
Data Extraction
Search
Table Details
Database Details
Data Source Details
Collection Task Details
Asset Catalog
Asset Inventory
Asset Catalog Self-Definition
Batch Inventory Assets
Modify Business Information
Data Discovery
Metadata collection
OPS of Instances
Ownership Management
Database
Data Table
Data Quality
Data Quality Overview
Getting Started with Data Quality
Viewing Quality Overview
Manage Rule Templates
Feature Overview
View system built-in templates
Add and manage custom Definition templates
Configure Quality Task
Features Overview
Add Rule under Monitoring Task
Add Rules in Batches
Investigate quality issues
Feature Overview
View Monitoring Results
Configure Task Alerts
Viewing Alarms
View Quality Report
Data security
Introduction to COS Data Security Solution
Safe Mode
Asset Inventory Check
Classification and Grading Configuration
Sensitive data identification
Data protection
Dynamic Masking
Static Masking
Tables I Manage

©2013-2025 Tencent Cloud. All rights reserved.

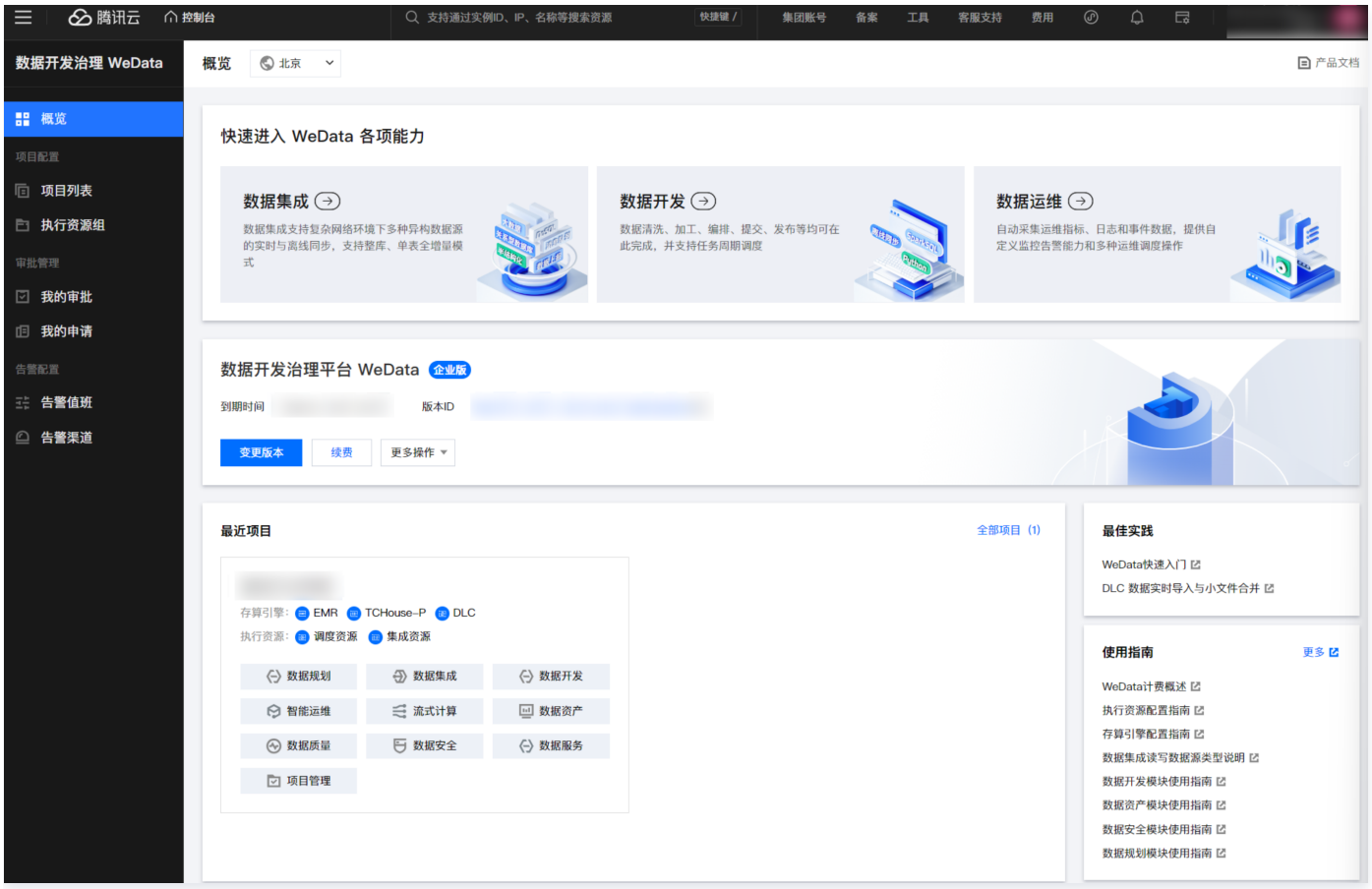
My Applied Tables
Data services
Data Service Overview
Preparations
Getting Started
Test API
Release API
API authorization
Calling APIs
Other Features
Data Service Alarm Feature Description
Data Service Monitoring feature Description
Data Service Authentication Feature Description
Data Service Log Analysis Feature Description
Practice Tutorial
Script Mode Advanced SQL Usage Guide
Appendix
API Error Code Description

©2013–2025 Tencent Cloud. All rights reserved.

Operation Guide
Console operation
Console Overview

Last updated: 2024-11-15 20:19:25

The management console is the first page users access on the Tencent Cloud website to enter WeData. It provides an overview and entry point for all WeData-related asset information under the tenant (main account), helping tenants to manage and view asset information, quickly understand product features, and enter workflows. This module includes: Overview, Project List, Execution Resource Group, Approval Management, Alarm Configuration, and User Management.

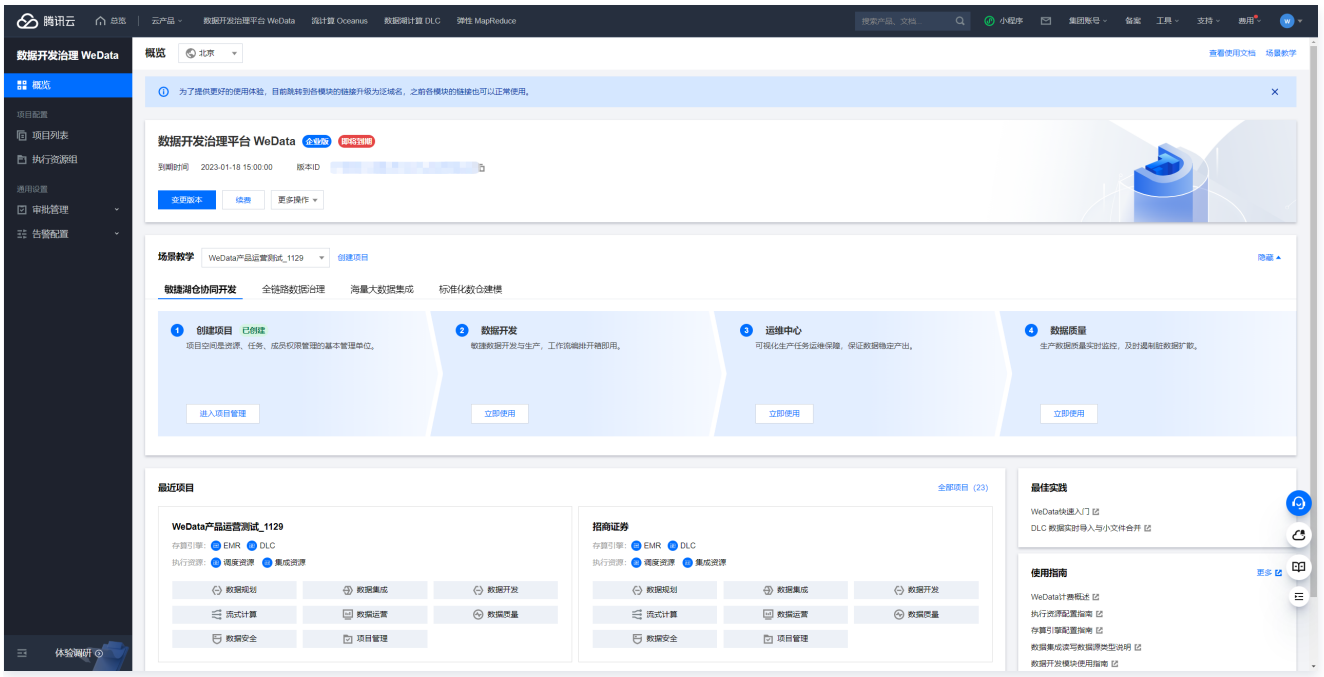


©2013-2025 Tencent Cloud. All rights reserved.

Overview Page

Last updated：2024-11-15 20:19:52

The Overview Page is the default opening page of the console. This page provides product usage tutorials, recent project access, best practices for products, and a product usage guide.



Platform Operations

Provides an option to change the current product version's features and also offers a portal for renewal and refund.

Use Cases

Offers tutorials for Agile Lakehouse Collaborative Development, Full-link Data Governance, Massive Data Integration, and Standardized Data Warehouse Modeling product use cases.

Recent Projects

Displays the projects recently accessed under the current account. It also provides access to data development, operation and maintenance, data quality, and project management modules for the project. You can click to enter specific feature modules for detailed operations. Additionally, you can click **View All Projects** to go to the Project List.

Best Practice

Provides an entry for best practices guidance documents for product usage, which can be used as a reference for product use.

User Guide

Provides detailed configuration documents and usage guides, enabling a comprehensive understanding of product features and settings.

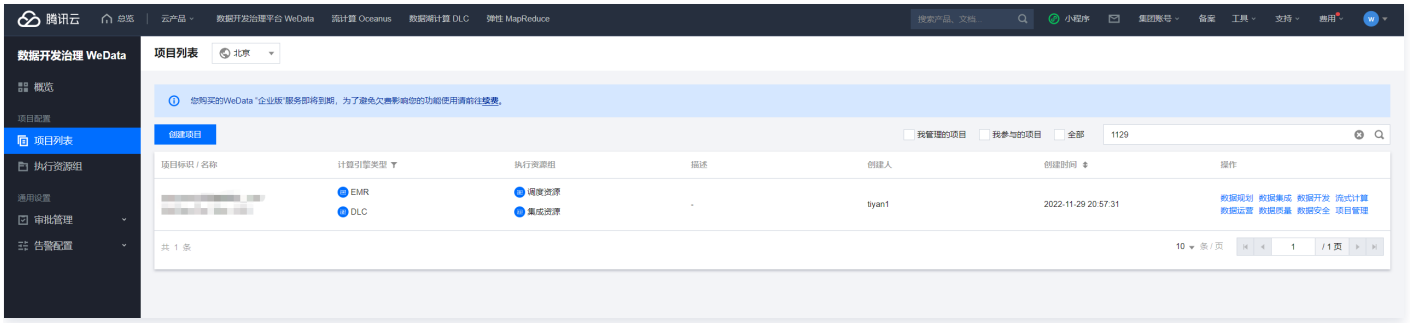
©2013-2025 Tencent Cloud. All rights reserved.

Project List

Last updated：2025-04-09 14:50:04

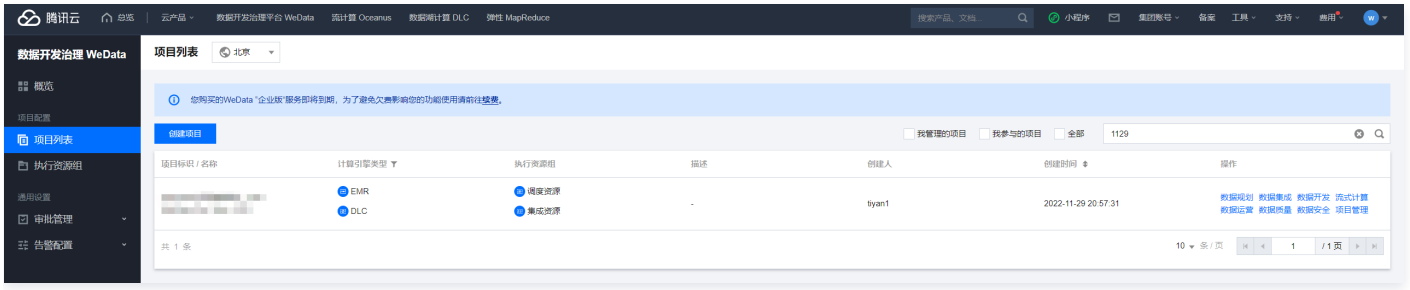
Project List

Click on **Project List** in the left directory tree of the console to view the project list and perform operations such as project creation. The project list provides basic information of all projects under the account, including region, project identifier/name, creator, creation time, bound compute engine type, execution resource group, and module operation entry under the project.



Enter each feature module

In the project list, select a project, and click Data Development, Operations Center, Data Operations, Data Quality, Project Management, Data Security, DataInLong, and Data Planning after the corresponding project to enter the respective page for operation.



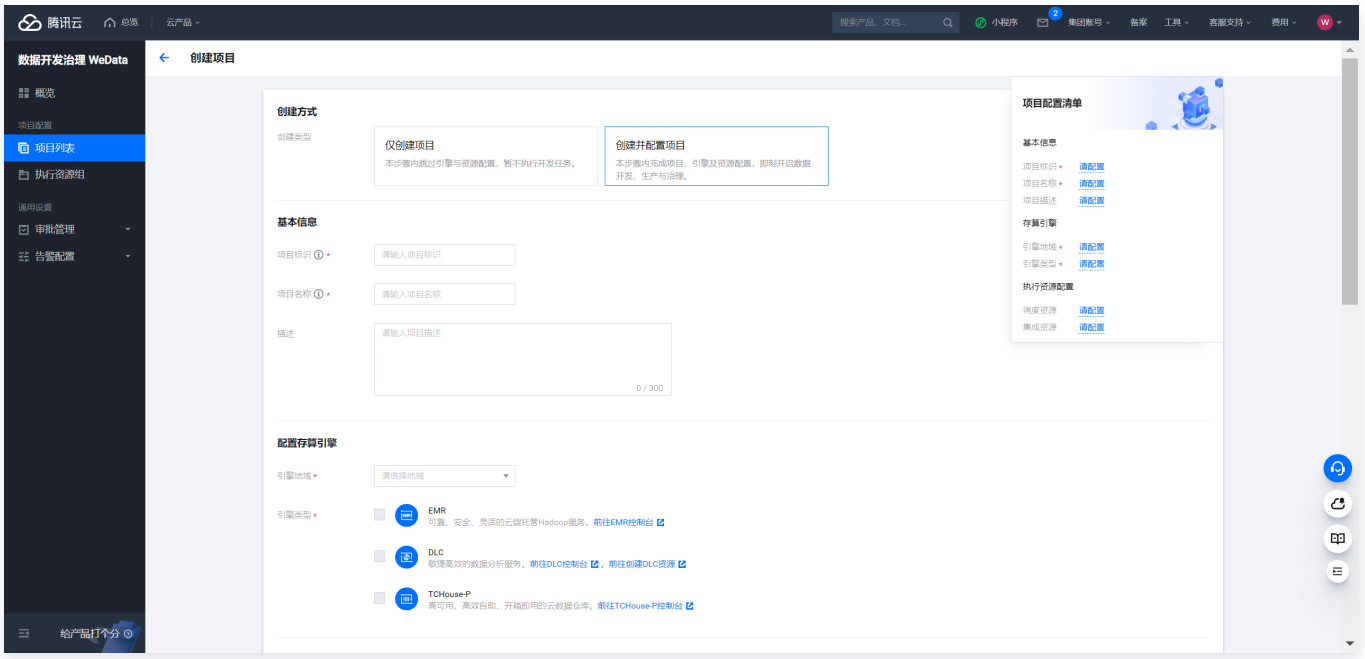
> **Note**
> Only members added to the project can enter the feature module. Non-project members do not have permission to enter the corresponding project's feature module, which will be grayed out.

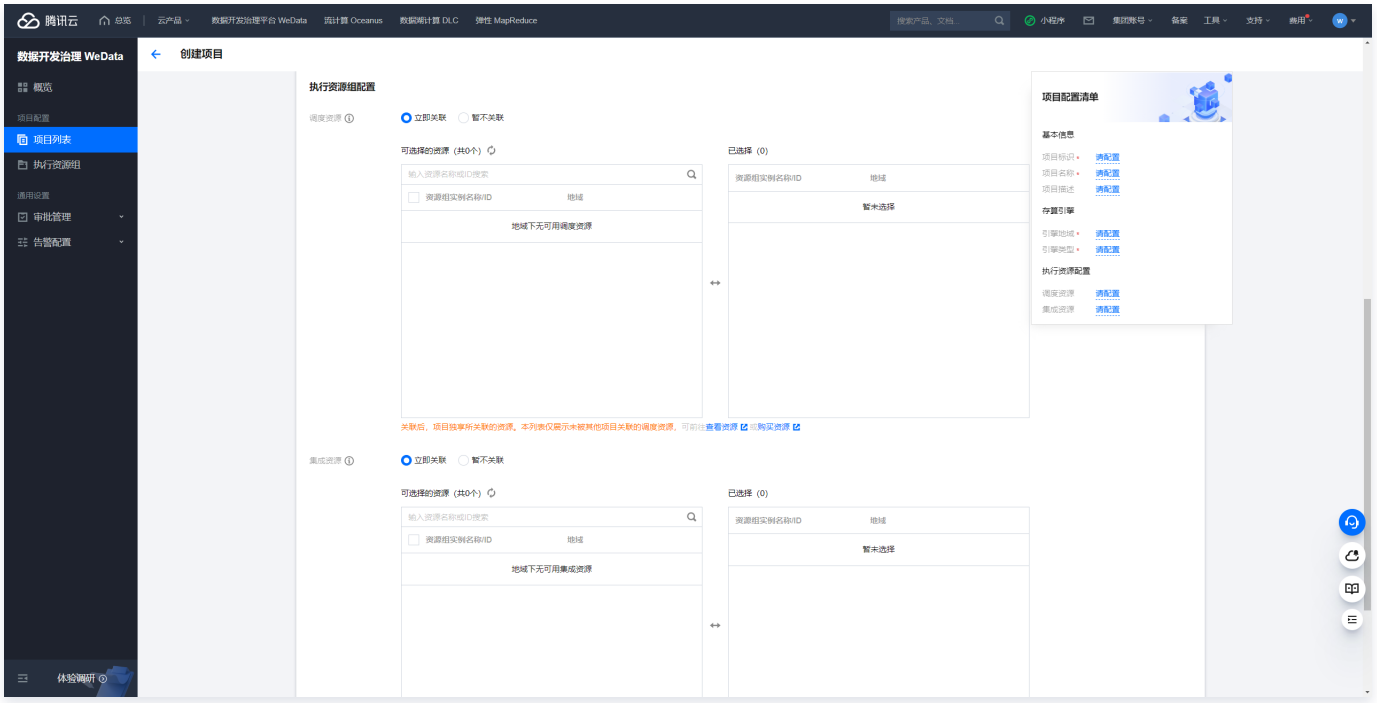
Create a project

WeData supports users to create projects in the console. The steps are as follows:

1. Select the region of your work. WeData currently supports creating projects under the Beijing region only.
2. Click **Create Project**, and configure the project in the pop-up window on the right according to your needs.

©2013-2025 Tencent Cloud. All rights reserved.





Classification	Parameter	Description
Creation Method	Creation Type	You can choose between "Create and Configure Project" and "Create Project Only" as the two project creation methods.
Basic information	Project Identifier	Project English ID, unique within the region. Starts with a letter and can contain letters, digits, and underscores, not exceeding 20 characters.
	Project name	Project Chinese Display Name, unique within the region. Starts with a letter or Chinese character and can contain letters, Chinese characters, digits, and underscores.
	Description	Provide a brief description of the created workspace.
Select Engine Type	Elastic MapReduce	After activation, you can use EMR in WeData to develop big data processing tasks. For details, refer to Go to EMR Console.

©2013–2025 Tencent Cloud. All rights reserved.

	DLC		After activation, you can use Tencent Cloud DLC in WeData. For details, see Go to DLC Console.
	Cloud Data Warehouse TCHouse-P		After activation, you can use Cloud Data Warehouse TCHouse-P in WeData. For details, see Go to TCHouse-P Console.
Configure Computational Storage Engine	Region		Select the region where the compute engine instance is located. Different types of compute engine instances in WeData must be in the same region.
	EMR Cluster		Select an available EMR cluster in the selected region for the current tenant. WeData only supports EMR with the Hive component. If no clusters are available, you can proceed to purchase an instance.
	DLC	Computing Resources	Select one from the available TCHouse-P computing resources of the current tenant in the selected region.
		secretId	Please go to the Tencent Cloud Certificate Authority M CAM console API Key Management Interface to copy the SecretId and SecretKey into the current input box. If there are restrictions on the permission control range, you can apply for a sub-account and fill in the corresponding SecretId and SecretKey.
		secretKey	
		Testing connectivity	Log in to Tencent Cloud test password to see if resources can be connected.
		Engine Metadata Collection	Once enabled, it will automatically create a collection task. After configuring the execution resource group, the collection task will gather and analyze metadata from the engine. You can go to the Data Operation Module to view and modify the task.
	TCHouse-P Cluster	Version	You can choose either the TCHouse-P1.0 or TCHouse-P2.0 version.
		Cluster Name	The name of the purchased TCHouse-P cluster in the selected region under this account.
		username & password	Account password used to connect to TCHouse-P cluster.
		Testing connectivity	Test the password to see if it can connect to the cluster. If successful, you can create a project (if the connectivity test fails, it may be because WeData is blocked by the network firewall where the cluster is located. Please refer to add TCHouse-P cluster allowlist).
Execute resource configuration	Scheduling resource	Scheduling resource	Scheduling resources are primarily used for scheduling data development tasks (including SQL tasks, shell tasks, etc.) at regular intervals.
		Associated resource	Scheduling resources must be located in the same region as EMR. Once associated, the project exclusively uses the associated resources. This list only shows scheduling resources not associated with other projects. You can go to view resources or purchase resources.
	Integration resource	Integration resource	The integration resource group mainly runs DataInLong tasks.
		Associated resource	After association, the project exclusively owns the associated resources. This list only shows integration resources that have not been associated with other projects. You can go to view resources.

3. After configuration, click **OK** to generate the project.

> **Note**
> - The different types of engines bound under the project will affect the corresponding available task types in the Data Development module.

©2013-2025 Tencent Cloud. All rights reserved.

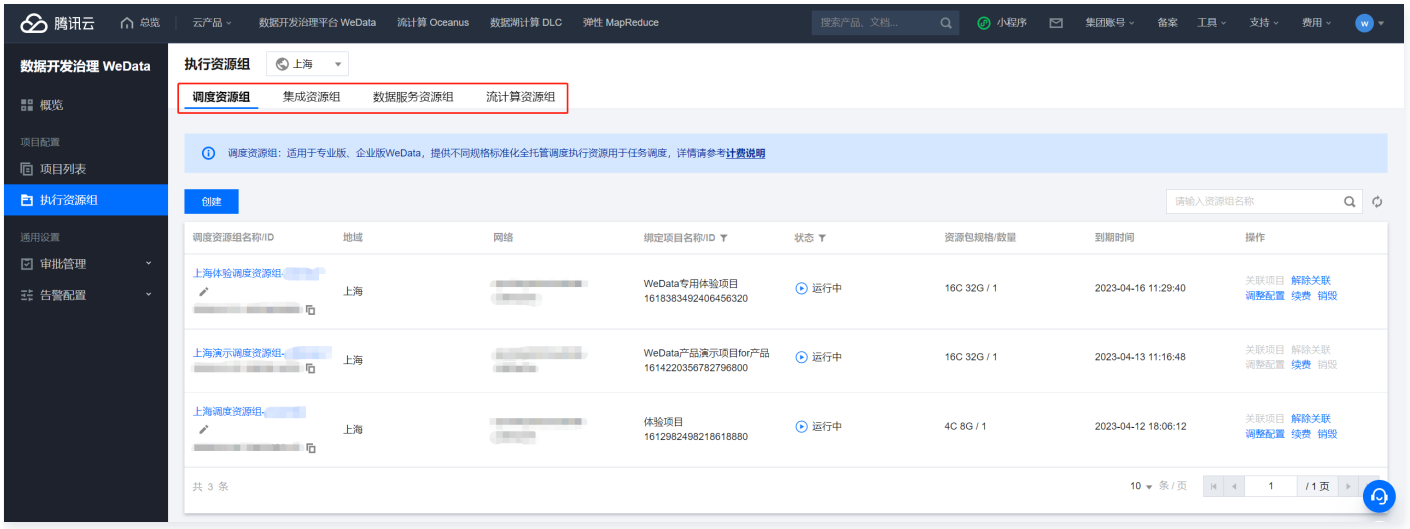
- WeData currently supports three types of computing engines: EMR (EMR), Tencent Cloud Data Warehouse TCHouse-P (TCHouse-P), and DLC (Data Lake Compute, DLC).

©2013-2025 Tencent Cloud. All rights reserved.

Execute Resource Group configuration Resource Group List

Last updated: 2024-09-09 12:04:09

Go to WeData Console, click **Execution Resource Group** to enter the resource group list page. The page includes management features for schedule resource group, integrated resource group, data service resource group, and stream computing resource group. The list displays all purchased resource groups and their status information under the current root account. Detailed parameters and descriptions in the list are as follows:



Parameter	Description
Resource Group Name/ID	Displayed name and ID of the resource group, the name supports modification
Region	Physical region to which the resource group belongs
Network	VPC information bound to the resource group
Bound Project Name/ID	Name and ID of the project space to which the resource group is bound
Status	Current running status of the resource group: • Running: Resources are operating normally • Creating: Order placed, the resource package within the current resource is under creation and configuration process, which may take 1-2 minutes • Creation failed: Resource allocation or formation failed. In this state, you can click to retry • Initialization in progress: Resource configuration is being initialized • Runtime Exception: An error occurred during the operation of the resource group • Updating: After changes or other operations, the current resource group's configuration is updating • Update Failed: Resource group configuration change failed. In this state, you can click Retry • Expired: Resource validity time has been reached. Freezing • Releasing: Resource validity time has exceeded 7 days or the user has triggered resource termination
Resource Pack Specifications/Quantity	Specifications of individual resource packs within the resource group and the number of resource packs
Expiration Time	Resource group expiration time. After expiration, the resource group will be frozen, affecting task execution
Operation	Perform operations on the resource group such as project association, disassociation, configuration

©2013-2025 Tencent Cloud. All rights reserved.

	adjustment, renewal, termination, and monitoring

Status and Operation Status Association Status Set Description:

Resource Group Status | Operations	Bind Project	Unbinding	Termination/Deletion	Renew	Adjusting Configurations
Running	✓	✓	✓	✓	✓
Creating	–	–	–	–	–
Creation Failed	✓	✓	In this state, the operation column displays the delete operation. Click Delete to default delete the unsuccessfully created resource and refund	–	–
Initializing	–	–	–	–	–
Running Abnormally	✓	✓	✓	✓	✓
Updating	–	–	–	–	–
Update Failed	✓	✓	✓	✓	✓
Expired	–	–	–	✓	–
Releasing	–	–	–	–	–
Released	–	–	–	–	–

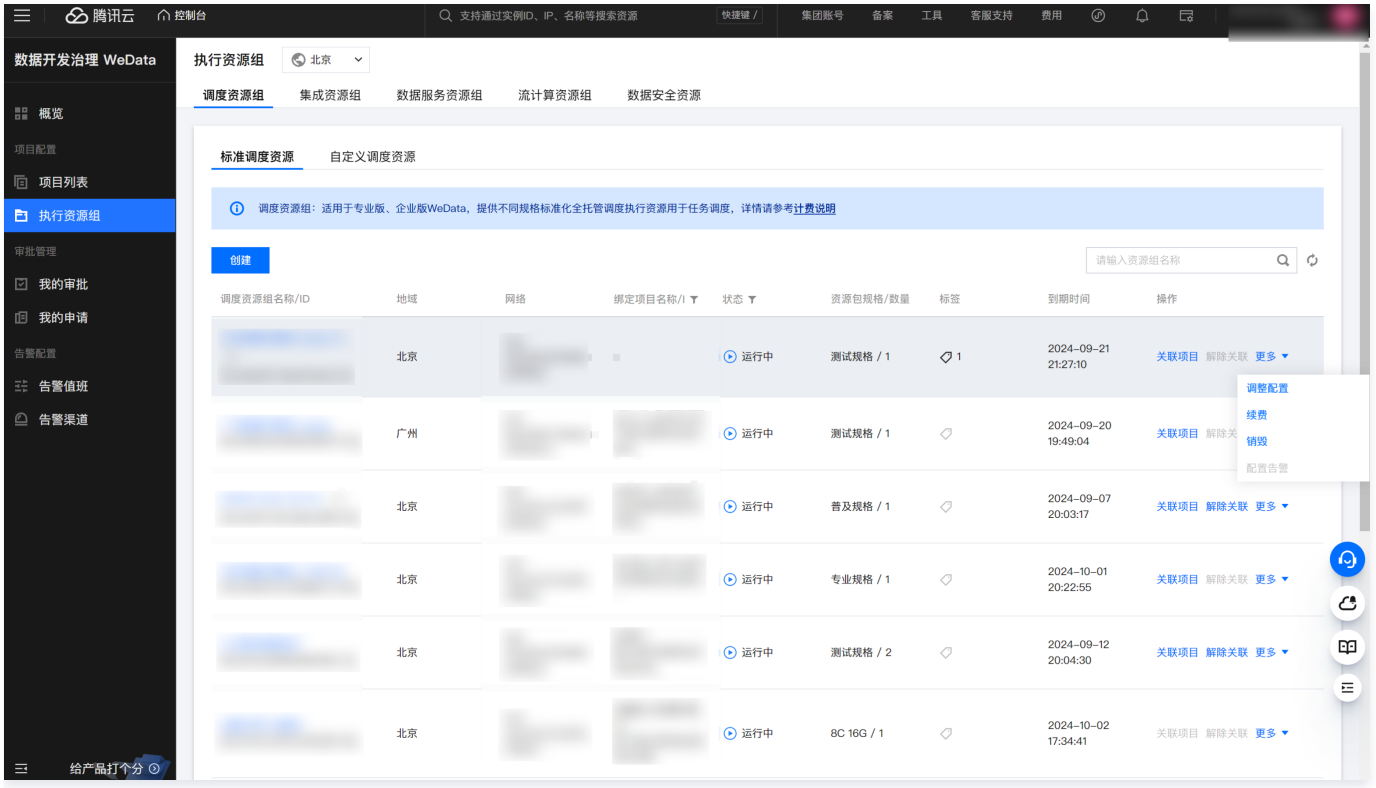
©2013–2025 Tencent Cloud. All rights reserved.

Scheduling Resource Group Configuration

Last updated：2024-12-12 20:31:42

Scheduling Resource Group List

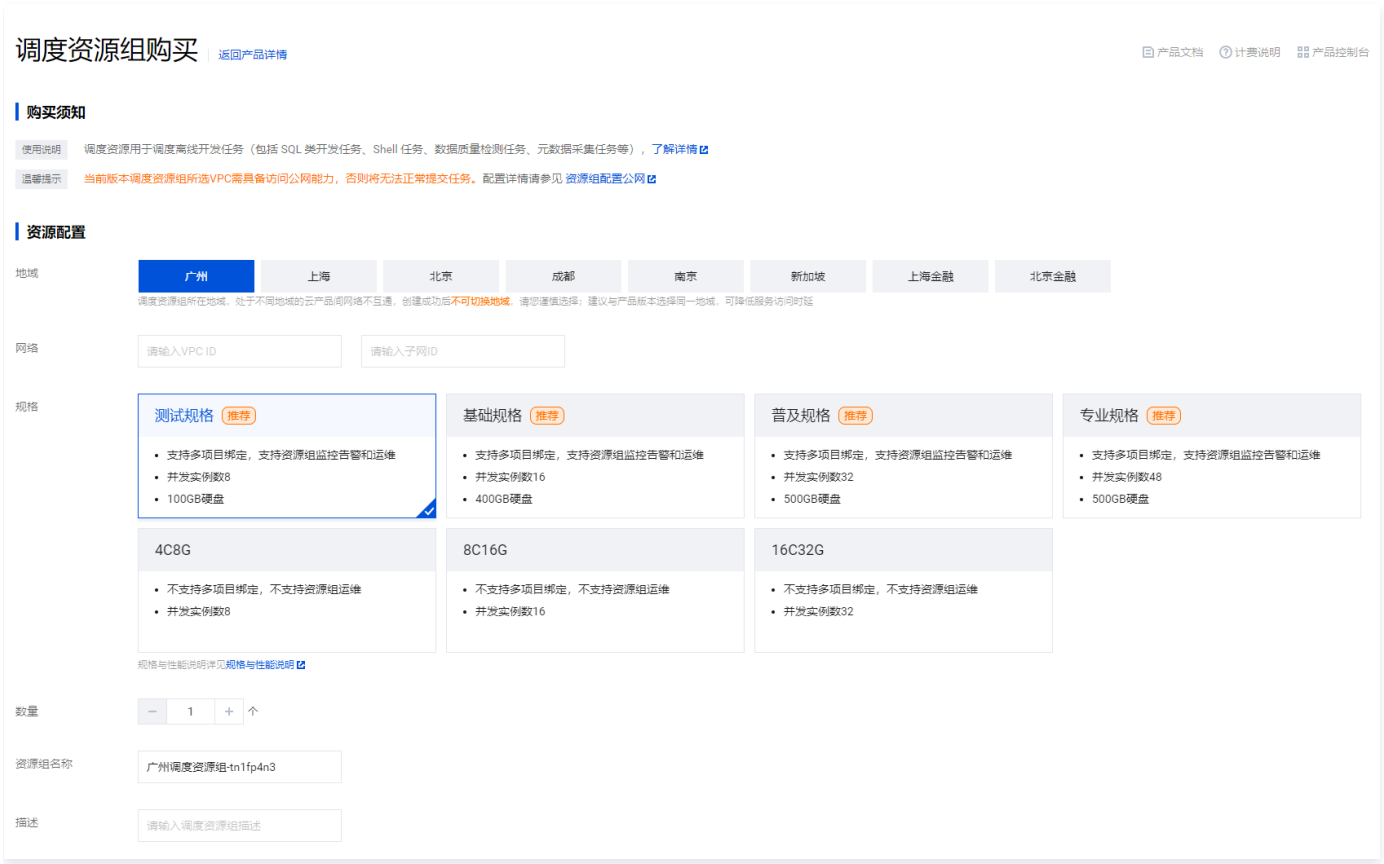
Execution Resource Group – Scheduling Resource Group, supporting creation, editing, and maintenance of scheduling resource groups. Scheduling resource groups are divided into Standard Scheduling Resource Groups and Custom Scheduling Resource Groups.



Creating Scheduling Resource Group

On the Resource Group Purchase page, after configuring the Resource Group, billing settings, and project association, the resource group creation is completed.

©2013–2025 Tencent Cloud. All rights reserved.



Parameter	Description
Region	The physical region of the resource group. The region of the scheduling resource group is in a different region from the cloud products, and the network between different regions is not interconnected. The region cannot be switched after successful creation, so please choose carefully; it is recommended to select the same region as the product version to reduce service access latency
Network	Specify the VPC and subnet information required for the current resource group configuration. The VPC selected by the scheduling resource needs to have public network access capability. If the existing network is not suitable, you can go to the console to create a new VPC or to create a new subnet
Specification	The specifications of the cores and memory of a single resource package. For details on specifications and performance, please refer to Specifications and Performance Description
Quantity	Total number of resource packages contained in the resource group
Resource Group Name	Set the display name of the resource group, the name cannot be empty and supports Chinese, English, numbers, -, _
Description	Set remarks for the current resource group
Billing mode	Monthly subscription
Purchase period	Support for purchase periods of 1 month, 2 months, 3 months, 4 months, 5 months, 6 months, 7 months, 8 months, 9 months, 1 year, 2 years, and 3 years, 4 years
Renewal	After selecting auto-renewal, if the account balance is sufficient, the device will be automatically renewed monthly upon expiration
Associated Project	After associating the project space, only tasks within this project can use this resource group for execution. Subsequent modifications can be made in the product console.

After configuration, click **Buy now** and complete the payment to view the scheduling resource group information on the console page.

©2013-2025 Tencent Cloud. All rights reserved.

Scheduling Resource Group Details

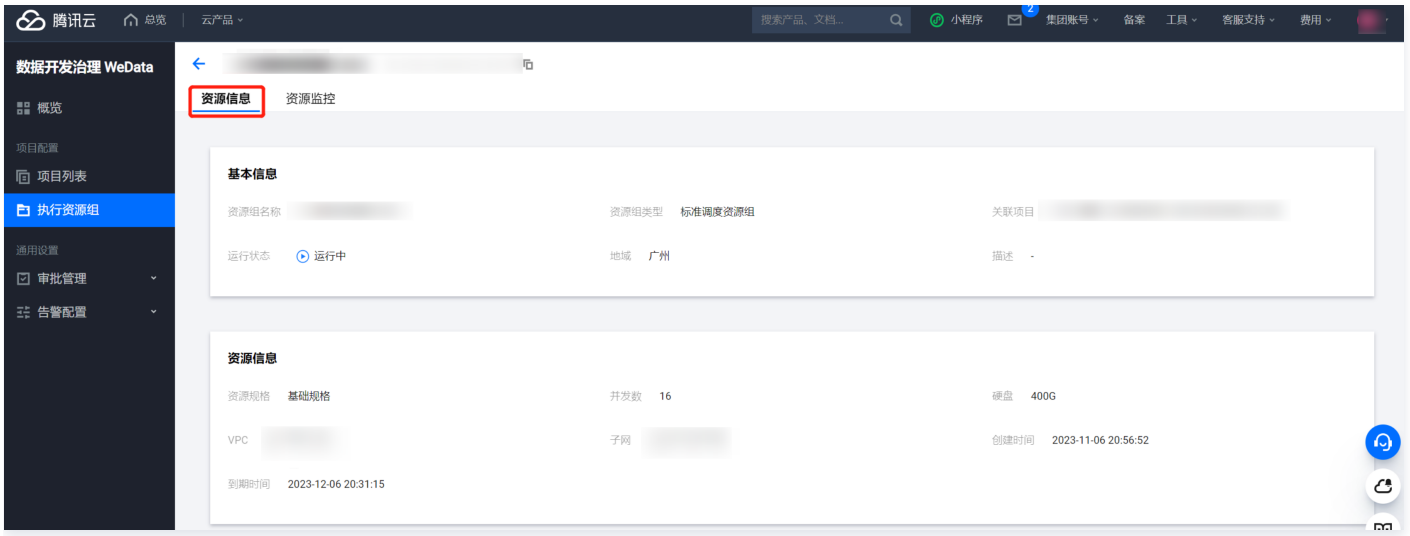
Click **Resource Group Name** to access the resource group information page, where you can view resource information, resource monitoring, and perform resource group operations and maintenance.



Resource Group Information

The resource information tab provides basic and detailed information about the resource.

- **Basic Information:** Resource Group Name, Resource Group Type, Associated Project, Running Status, Region, Description.
- **Resource Information:** Resource Specifications, Concurrency, Hard Disk, VPC, Subnet, Creation Time, Expiration Time.



Parameter Description:

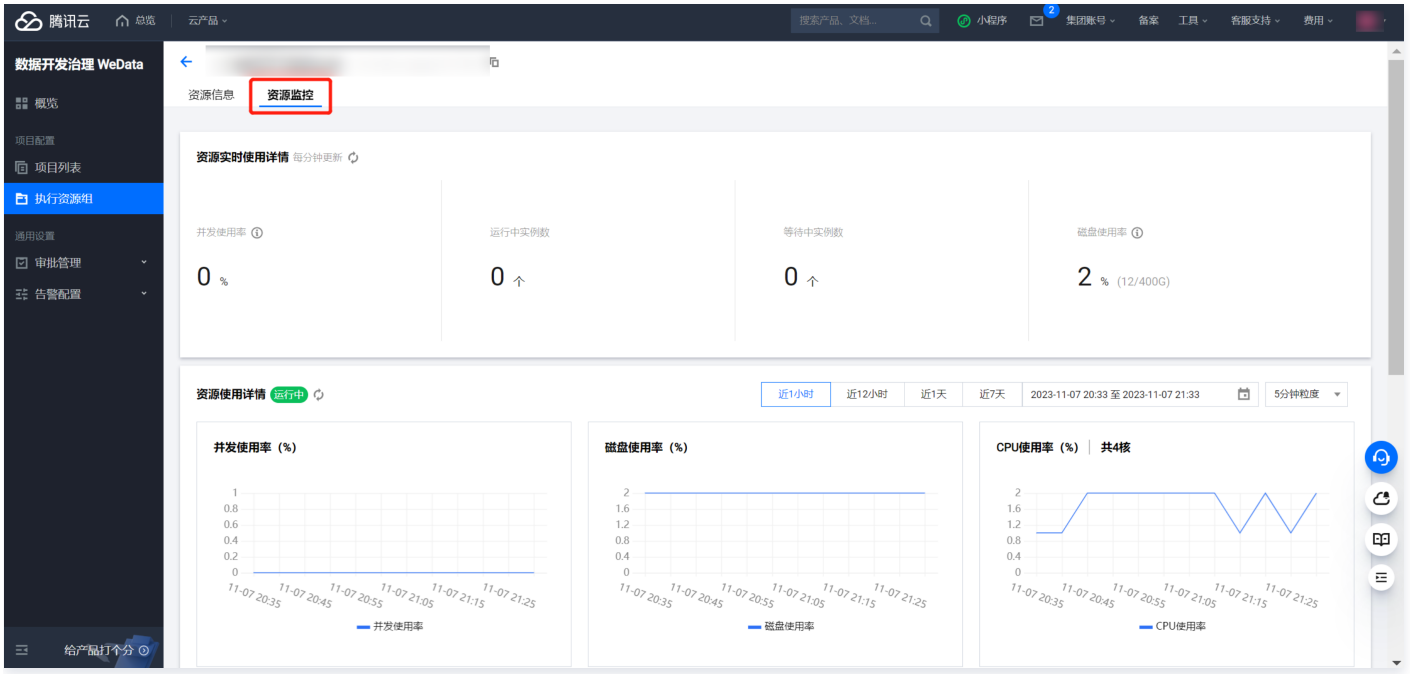
Parameter		Description
Basic information	Resource Group Name	User-defined scheduling resource group name.
	Resource Group Type	The type of scheduling resource group.
	Associated Project	Projects associated with the current scheduling resource group.
	Running Status	The current running status of the scheduling resource group includes types such as running, creating, creation failed, initializing, running exception, updating, update failed, expired, releasing, and released.
	Region	Region of ownership for the current scheduling resource group.
	Description	User-defined description information of the scheduling resource group.

©2013–2025 Tencent Cloud. All rights reserved.

Resource Information	Resource specification	Specification type of the scheduling resource group.
	Concurrent number	The maximum concurrency supported by the current scheduling resource group.
	Disk	Hard disk space size of the current resource group.
	VPC	VPC network information of the current scheduling resource group.
	Subnet	Subnet information of the current scheduling resource group.
	Creation Time	Creation time of the current scheduling resource group.
	Expiration Time	Expiration time of the current scheduling resource group.

Resource Group Monitoring

The resource monitoring tab can monitor the usage of the scheduling resource, providing real-time monitoring of key indicators, including concurrent usage rate, number of instances running, disk utilization, etc. It displays resource usage trends over user-defined time intervals through charts.



Parameter description:

Parameter		Description
Real-time Resource Usage Details	Concurrent Usage Rate	Shows the percentage of concurrent usage of the scheduling resource group, number of running instances/concurrent number. (This is an estimated usage rate, influenced by the size of resource usage by tasks, for reference only)
	Number of Running Instances	Shows the number of running instances of the scheduling resource.
	Number of Waiting Instances	Shows the number of waiting instances of the scheduling resource.
	Disk Utilization	Shows the current disk utilization percentage, along with usage and total volume.
Resource Usage Details	Concurrent Usage Rate (%)	Shows the fluctuation of concurrent usage rate within the last 1-hour statistical interval.

©2013-2025 Tencent Cloud. All rights reserved.

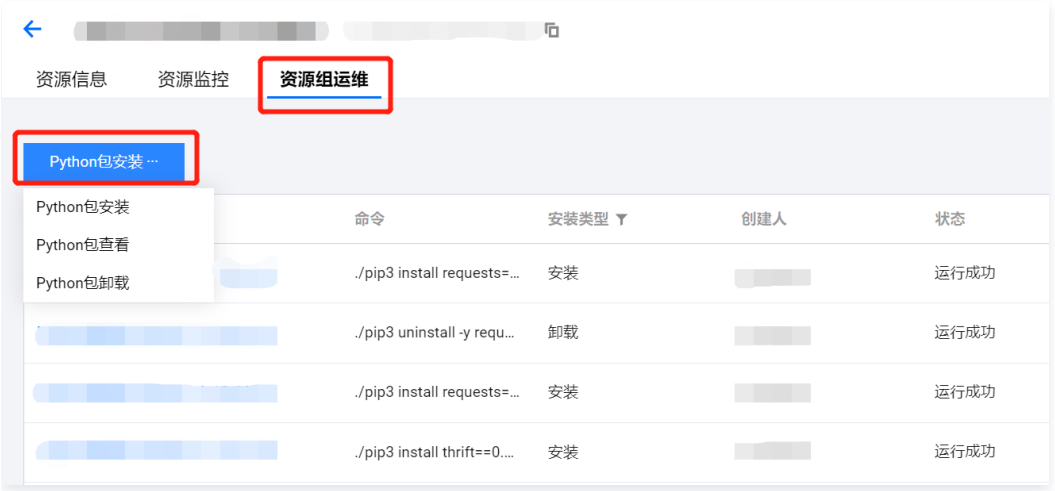
	Disk utilization (%)	Shows the fluctuation of disk utilization within the last 1-hour statistical interval.
	CPU utilization (%)	Shows the fluctuation of CPU utilization within the last 1-hour statistical interval.
	System Average Load	Shows the fluctuation of CPU system average load within the last 1-hour statistical interval.
	Memory utilization (%)	Shows the fluctuation of memory utilization within the last 1-hour statistical interval.
	Memory usage (MB)	Shows the fluctuation of memory usage within the last 1-hour statistical interval.
	Number of running instances	Shows the fluctuation of running instance count within the last 1-hour statistical interval.
	Number of tasks in scheduling	Shows the fluctuation of tasks in scheduling within the last 1-hour statistical interval.

Resource Group OPS

Resource Group OPS currently provides a management feature for installing Python packages, offering support for Python tasks in data development.



The Resource Group OPS interface provides entries for installing, uninstalling, and viewing Python packages, as well as a list of command execution records. Users can view relevant information about Python installation and uninstallation here, including name, installation type, creator, status, description, and time.



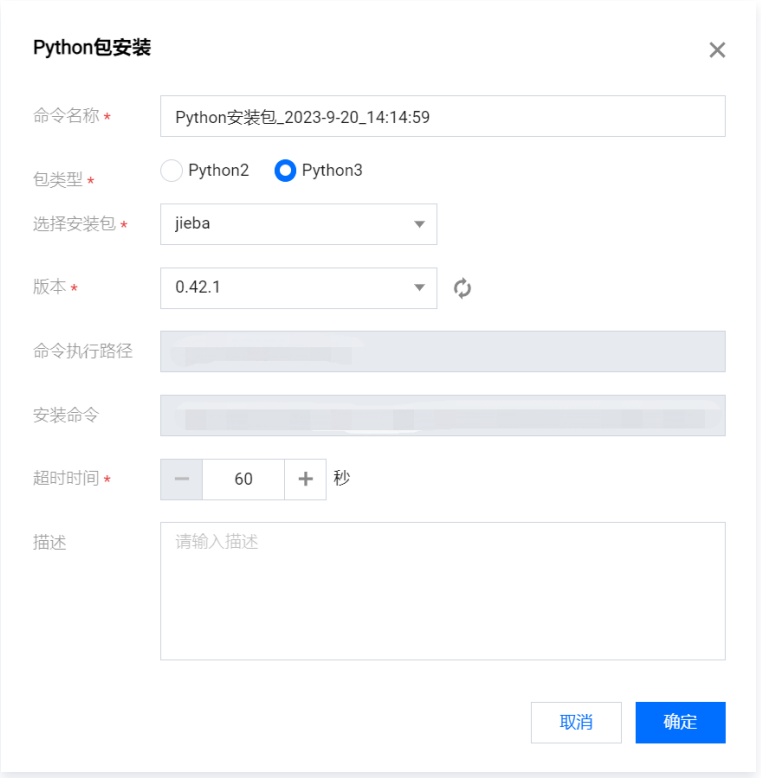
Install Python package

1. Click 'Install Python package' and configure Python package information in the popup.

> ⚠ **Note:**

©2013-2025 Tencent Cloud. All rights reserved.

> This feature is only effective in the New version specifications of scheduling resources. Old specifications do not support installation.



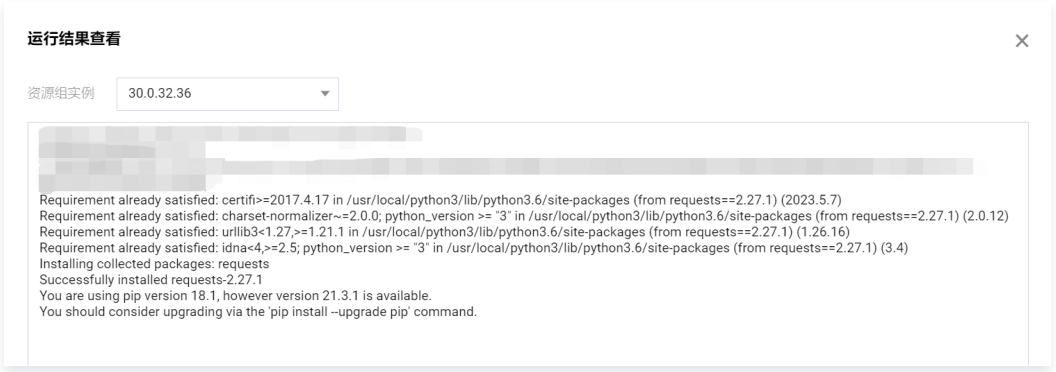
Information	Description
Command Name	Automatically generated, default is "Python Installation Package_Current Time", users can modify it.
Package Type	Python Version, including Python2 and Python3.
Select Installation Package	Supports keyword search. Currently supported installation packages are as follows: Python2:jieba,numpy,openpyxl,pandas,pyhive,requests,scikit-learn,thrift,thrift_sasl,tencentcloud-sdk-python. Python3:jieba,numpy,openpyxl,pandas,pyhive,requests,scikit-learn,thrift,thrift_sasl,tencentcloud-sdk-python,pymysql,cos-python-sdk-v5,threadpool,tencentcloud-dlc-connector,matplotlib.
Version	Different versions of the corresponding Python package can be selected.
Command Execution Path	Default path provided by WeData, Python package installation path.
Installation Command	Automatically generated.
Timeout Period	Maximum execution time of the installation command. The execution will be terminated if it exceeds the time period.
Description	Self-Definition Description Information.

2. After the configuration of Python package installation steps is completed, the installation command can be generated. On the resource group operation and maintenance page, click Run under the corresponding installation command operation column to execute the Python package installation command. After a successful installation, the status column will indicate that the command ran successfully.

©2013-2025 Tencent Cloud. All rights reserved.



3. Click **View Results** under the operation column to see the detailed execution of the installation command.



4. Click **Command Name** in the list to view command details.



View Python Package

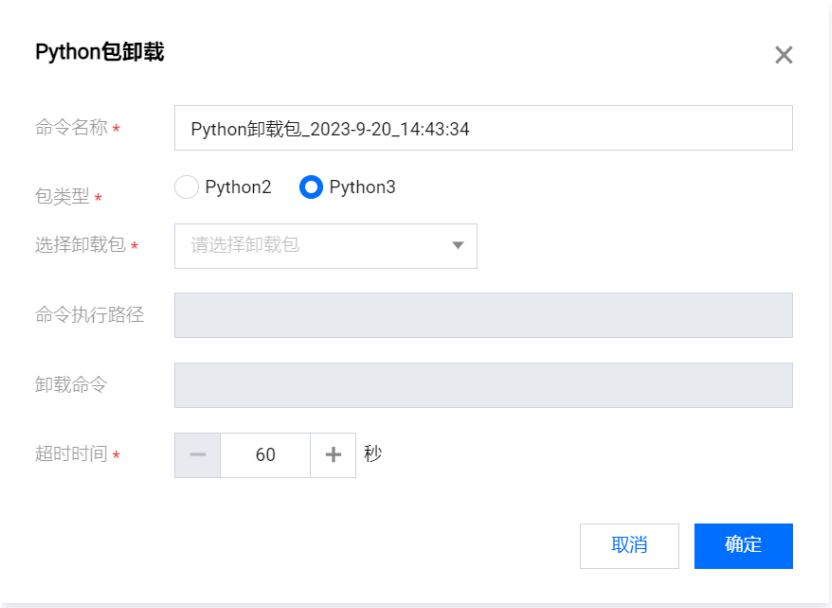
Click 'View Python Package' to see the installed Python2 and Python3 packages in the popup.



Uninstall Python Package

Click 'Uninstall Python Package' to choose the installed Python packages for uninstallation in the popup.

©2013–2025 Tencent Cloud. All rights reserved.



Information	Description
Command Name	Automatically generated, default is "Python Uninstall Package_Current Time", users can modify it.
Package Type	Python Version, including Python2 and Python3.
Select Uninstall Package	Choose from installed Python packages.
Command Execution Path	Default path provided by WeData, uninstallation path for Python packages.
Uninstall Command	Automatically generated.
Timeout Period	Maximum execution time of the installation command. The execution will be terminated if it exceeds the time period.

Scheduling Resource Group Operations

Associate/Disassociate Projects

Associate the current scheduling resource group with project space information. After assignment, tasks in the project space can use this resource group.

> **Note:**
> Old version resource groups can only associate with one project space. The new version resource group can associate with multiple project spaces.

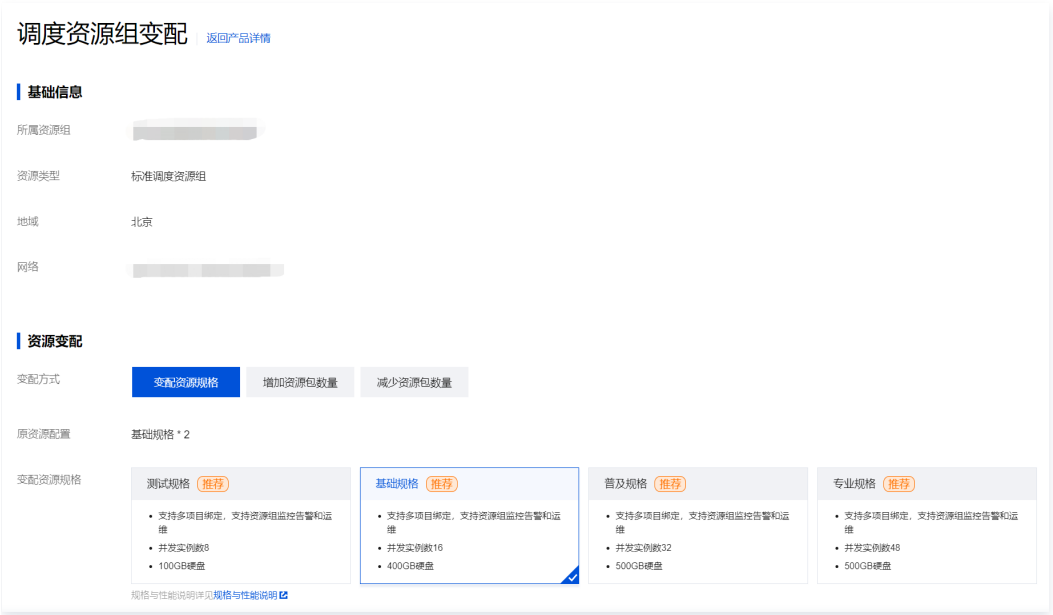
©2013–2025 Tencent Cloud. All rights reserved.



Disassociating removes the scheduling resource group from the available resources of the current space. Ensure no tasks are running in the current project space, otherwise disassociation will fail.



Adjusting Configurations



Change the basic specifications or the quantity of resource packages in the current resource group. Operations include changing resource specifications, increasing, or decreasing the number of resource packages.

- Change Resource Specifications: Expand or reduce the specifications of the basic resource packages within the resource group, while keeping the quantity unchanged.
- Increase Resource Package Quantity: Once selected, only the total number of resource packages under the current resource group can be increased.
- Decrease Resource Package Quantity: Once selected, only the total number of resource packages under the current resource group can be adjusted.

> **Note**

©2013–2025 Tencent Cloud. All rights reserved.

Decreasing the number of resource packages may affect the tasks associated with the resource group. Proceed with caution.

Renew

Renew Operation: When performing renew operations on the current resources, if multiple resource packages are included in the resource group, the operation will affect all resource packages.

Terminate

Termination refers to forcibly destroying the current resource group, which cannot be recovered afterward.

Note
If there are ongoing tasks associated with the resource group, this operation will directly impact task execution and data output. Please proceed with caution.

Configure alarms

Perform alarm monitoring on the concurrent usage rate and disk utilization of the resource group. When the usage rate exceeds a certain threshold, send alarm notifications to help identify insufficient resource inventory in time.



Parameter		Description
Alarm Condition	Concurrent Usage Rate	The resource occupancy rate during concurrent execution of tasks within the resource. Under high concurrency, tasks need to be queued for execution.
	Disk Utilization	Disk utilization of the executing resources.
Alarm notification	Recipient	You can specify account users or users scheduled according to the duty roster as recipients when alarm information is triggered.
	Alarm Method	Currently supported alarm information sending methods include Email, SMS, WeChat, phone, Corporate WeChat, HTTP, Corporate WeChat Group, and Feishu Group. **Note:** • Mobile, WeChat, and Email accounts can be configured in Tencent Cloud User Management .

©2013–2025 Tencent Cloud. All rights reserved.

		• Corporate WeChat accounts can be configured in Tencent Cloud Federated Accounts .
	Alarm Status	Alarm trigger switch: When the switch is turned on, the alarm conditions take effect.

©2013–2025 Tencent Cloud. All rights reserved.

Integrated Resource Group Configuration

Last updated：2024-09-04 22:16:10

Integrated Resource Group Management

Enter WeData Console, click **Execution Resource Group**> **Integrated Resource Group**, and go to the resource group list page.



The resource group list displays all purchased resource groups and their status information under the current main account. Detailed parameters and descriptions in the list are as follows:

Parameter	Description
Resource Group Name	Resource Group Display Name
Region	Physical region to which the purchased resource group belongs
Network	Network information configured and bound within the current resource group
Bind Project	Name and ID of the project space to which the current resource group is bound
Status	Current running status of the resource group: • Running: Resources are operating normally • Creating: Order placed, the resource package within the current resource is under creation and configuration process, which may take 1-2 minutes • Creation failed: Resource allocation or formation failed. In this state, you can click to retry • Initialization in progress: Resource configuration is being initialized • Runtime Exception: An error occurred during the operation of the resource group • Updating: After changes or other operations, the current resource group's configuration is updating • Update Failed: Resource group configuration change failed. In this state, you can click Retry • Expired: Resource validity time has been reached. Freezing • Releasing: Resource validity time has exceeded 7 days or the user has triggered resource termination
Resource Pack Specifications/Quantity	Specification of a single resource package included in the current resource group and the total number of resource packages
Tag	Resource values support Profit Sharing by Tag
Expiration Time	Resource group expiration time. After expiration, the resource group will be frozen, affecting task execution
Operation	Perform operations such as project association or resource reallocation for the current resource group. Permitted operations are associated with the resource status

©2013-2025 Tencent Cloud. All rights reserved.

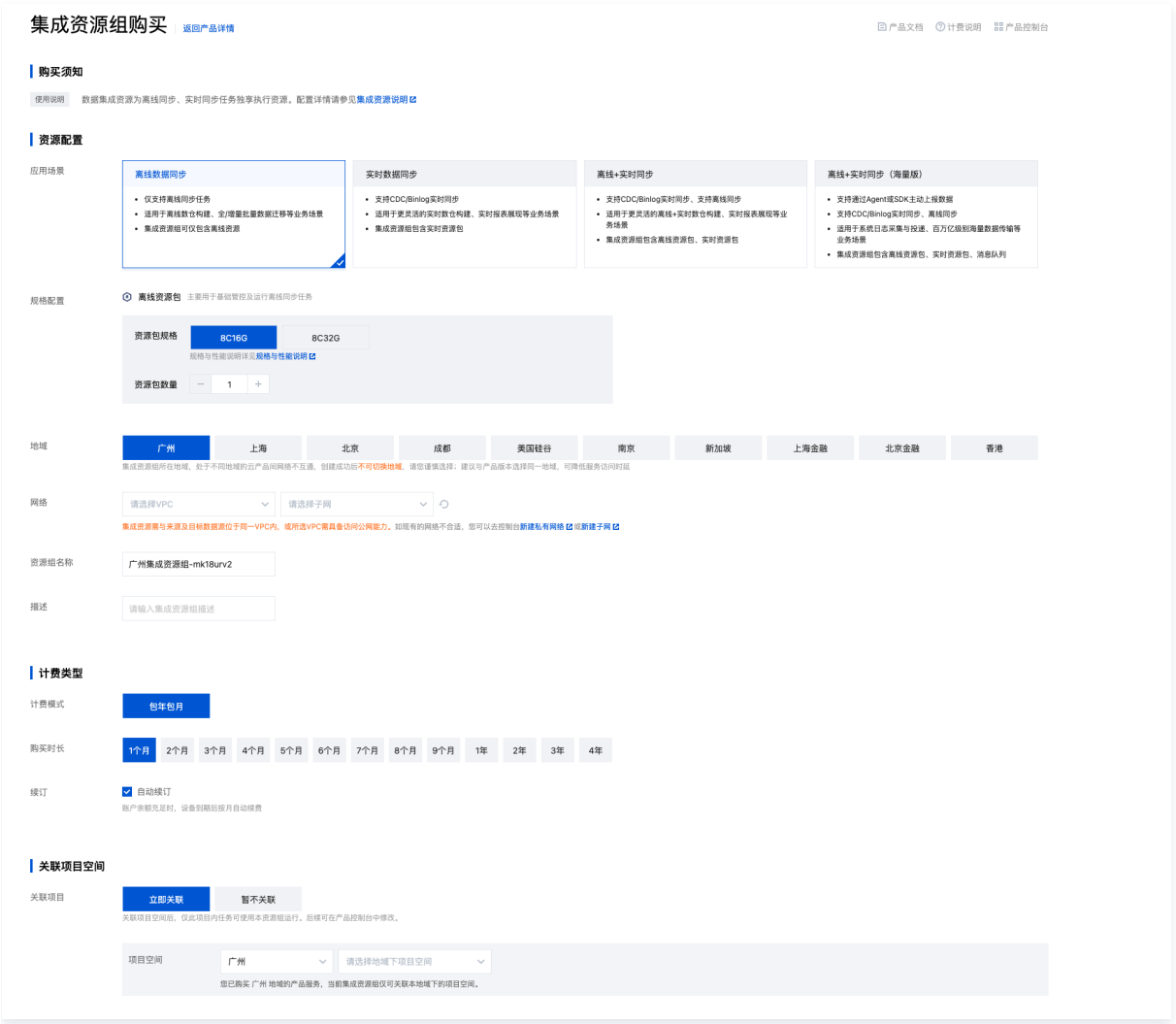
Status and operational status association and description:

Resource Group Status | Operations	Bind Project	Unbinding	Termination/Deletion	Renew	Adjusting Configurations
Running	✓	✓	✓	✓	✓
Creating	–	–	–	–	–
Creation Failed	✓	✓	In this state, the operation column displays the delete operation. Click Delete to default delete the unsuccessfully created resource and refund	–	–
Initializing	–	–	–	–	–
Running Abnormally	✓	✓	✓	✓	✓
Updating	–	–	–	–	–
Update Failed	✓	✓	✓	✓	✓
Expired	–	–	–	✓	–
Releasing	–	–	–	–	–
Released	–	–	–	–	–

Integrated Resource Group Creation/Purchase

On the selected Integrated Resource Group page, click **Create** to enter the resource configuration page.

©2013–2025 Tencent Cloud. All rights reserved.



Parameter	Description
Application Scenario	Supports four application scenarios: Offline Data Synchronization, Real-time Data Synchronization, Offline + Real-time Synchronization, and Offline + Real-time Synchronization (Massive Edition)
Specification Configuration	For detailed specifications and performance, see Specifications and Performance
Region	Physical region of the resource group. Integrated Resource Groups located in different regions do not have network connectivity between cloud products in different regions. The region cannot be switched after creation, so please choose carefully; it is recommended to select the same region as the product version to reduce service access latency
Network	Integrated resources must be in the same VPC as the source and destination data sources, or the selected VPC must have the ability to access the public network. If the current network is unsuitable, you can go to the console to create a new VPC or create a new subnet
Resource Group Name	Set the display name of the resource group, the name cannot be empty and supports Chinese, English, numbers, -, _
Description	Set remarks for the current resource group
Billing mode	Monthly subscription
Purchase period	Support for purchase periods of 1 month, 2 months, 3 months, 4 months, 5 months, 6 months, 7 months, 8 months, 9 months, 1 year, 2 years, and 3 years, 4 years

©2013-2025 Tencent Cloud. All rights reserved.

Renewal	After selecting auto-renewal, if the account balance is sufficient, the device will be automatically renewed monthly upon expiration
Associated Project	After associating the project space, only tasks within this project can use this resource group for execution. Subsequent modifications can be made in the product console.

After completing the configuration, click **Buy Now** and complete the payment. You can then go to the console page to view the information of the integrated resource group.

Associate/Disassociate Projects

The associated project specifies the project space information of the current resource group. After specifying, tasks under this project space can use this resource group.

- A resource group can be associated with multiple project spaces.
- Disassociation removes the resource group from the resources available to the current space.
- Disassociation requires that no tasks are running under the current project space; otherwise, the disassociation will fail.

The offline and real-time resource packages included in the DataInLong resource group must be associated with the same project space. The association or disassociation operation will simultaneously affect the resource group.



Integrated Resource Group Configuration Adjustment



Change the basic specifications or the quantity of resource packages in the current resource group. Operations include changing

©2013-2025 Tencent Cloud. All rights reserved.

resource specifications, increasing, or decreasing the number of resource packages.

- Resource Specification Adjustment: Expand or reduce the specifications of the offline resource package within the resource group, keeping the number of resource packages unchanged.
- Increase Resource Package Quantity: Once selected, only the total number of resource packages under the current resource group can be increased.
- Decrease Resource Package Quantity: Once selected, only the total number of resource packages under the current resource group can be adjusted.

> **Note**
> Decreasing the number of resource packages may affect the tasks associated with the resource group. Proceed with caution.

Integrated Resource Group Renewal

Renew Operation: When performing renew operations on the current resources, if multiple resource packages are included in the resource group, the operation will affect all resource packages.

Integrated Resource Group Termination



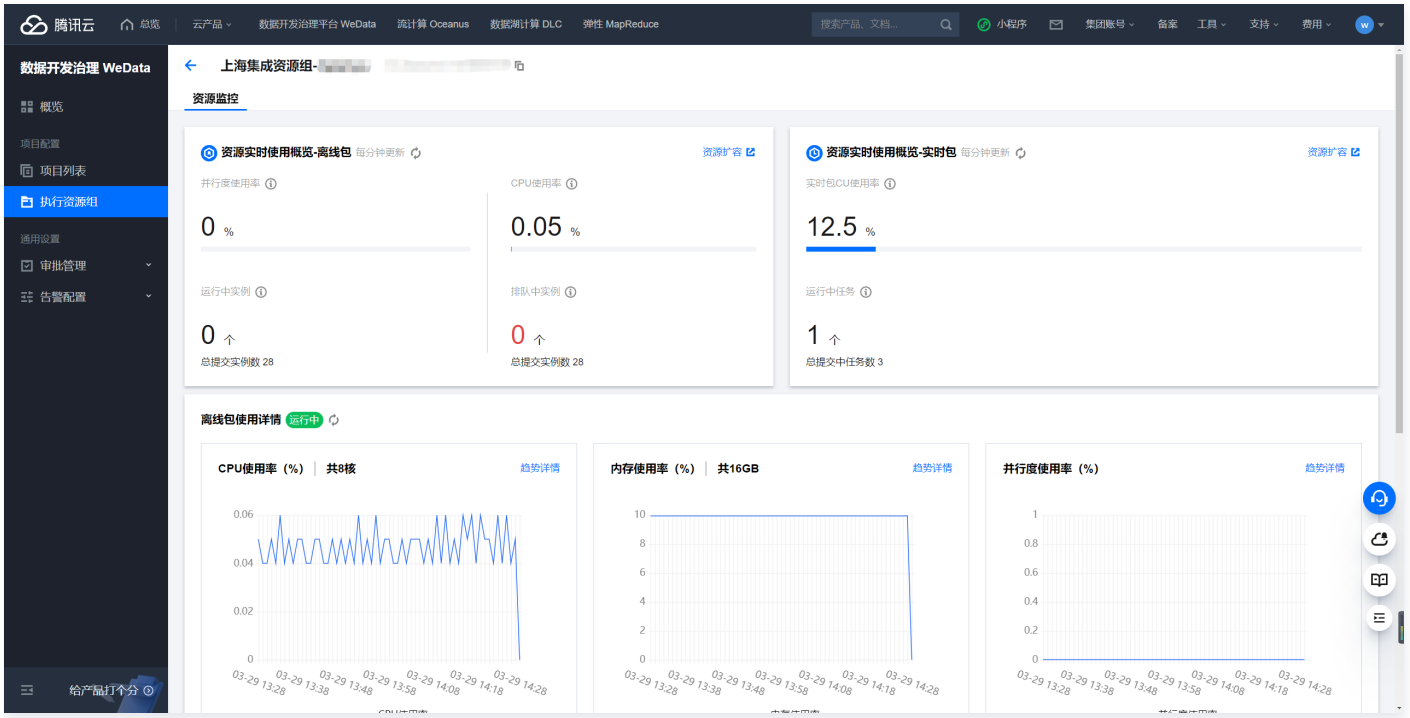
Destruction refers to forcibly terminating the current resource group. After destruction, the resource group cannot be recovered. If the current resource group is associated with any running tasks, this operation will directly affect task execution and data output. Please proceed with caution!

> **Note**
> If the resource integration group contains both Offline Packages and Real-time Packages, destroying the Offline Package will also destroy the Real-time Package.

Integrated Resource Group Monitoring

The Monitoring Page provides an overview of the current resource group's usage, including resource package utilization, running instances, and task execution statuses.

©2013–2025 Tencent Cloud. All rights reserved.

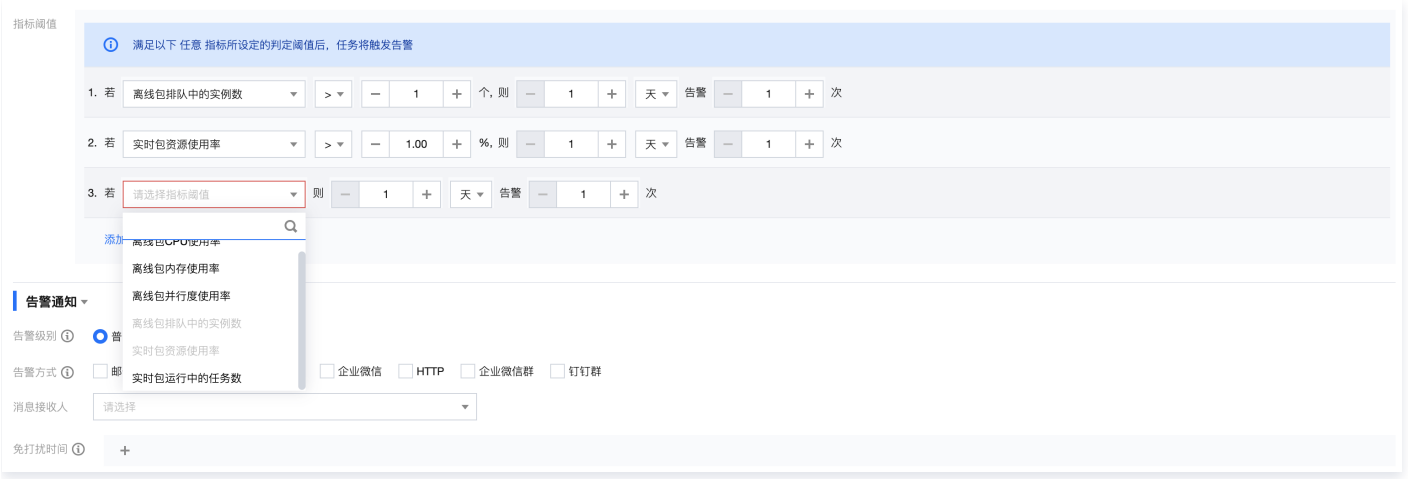


Integrated Resource Group Configuring Alarm

> **Note**
> Only Integrated Resource Groups bound to a project support configuring alarms.

Click **Configure Alerts** to enter the first project bound to the resource group to create alert rules.

- Offline Packages and Real-time Packages support different alert metrics. Offline Packages support alerts based on CPU Utilization, Memory Usage Rate, Parallelism Utilization Rate, and Number of Instances in Queue.
- Real-time Packages support alerts based on Resource Utilization and Number of Tasks Running.



©2013-2025 Tencent Cloud. All rights reserved.

Data Service Resource Group

Last updated：2024-08-30 17:36:48

Creating Data Service Resource Group

1. In the **executing resource group** > **Data Service Resource Group** tab page, click **Resource Initialization**, and in the pop-up window, configure the region, network, subnet, and API Gateway instance to add the data service resources to the list.
   If no API Gateway is created, you can jump to Cloud Native Gateway for creation. For detailed usage of the data service feature, refer to Creating API.



2. Enter the resource initialization interface and fill in the parameter information.



Parameter	Description
Specifications and Performance	For details on specifications and performance, see Service Resource Billing Explanation.
VPC	Data service resources must be located in the same VPC as the data source, or the selected VPC must have the ability to access the public network. If the existing network is not suitable, you can go to the console to Create a VPC or Create a Subnet.
Region	The physical region of the resource group's location. Network connectivity between cloud products in different regions is not available. It is recommended to choose the same region as the product version to reduce service access latency.
API Gateway instance	Select the API Gateway instance for API service calls. If it is empty, please jump to Cloud Native API Gateway to confirm and create.

3. After configuration, click **OK**, to view the resource group information on the Data Service Resource Group page.

©2013-2025 Tencent Cloud. All rights reserved.

Managing Data Service Resource Group

In the **executing resource group** > **Data Service Resource Group** tab page, click **Service Resource Group name,** then in the right frame, click **API Gateway Service,** to jump to the API Gateway Service page for gateway service management. It provides complete lifecycle management for APIs, including creating, maintaining, publishing, running, and decommissioning features.

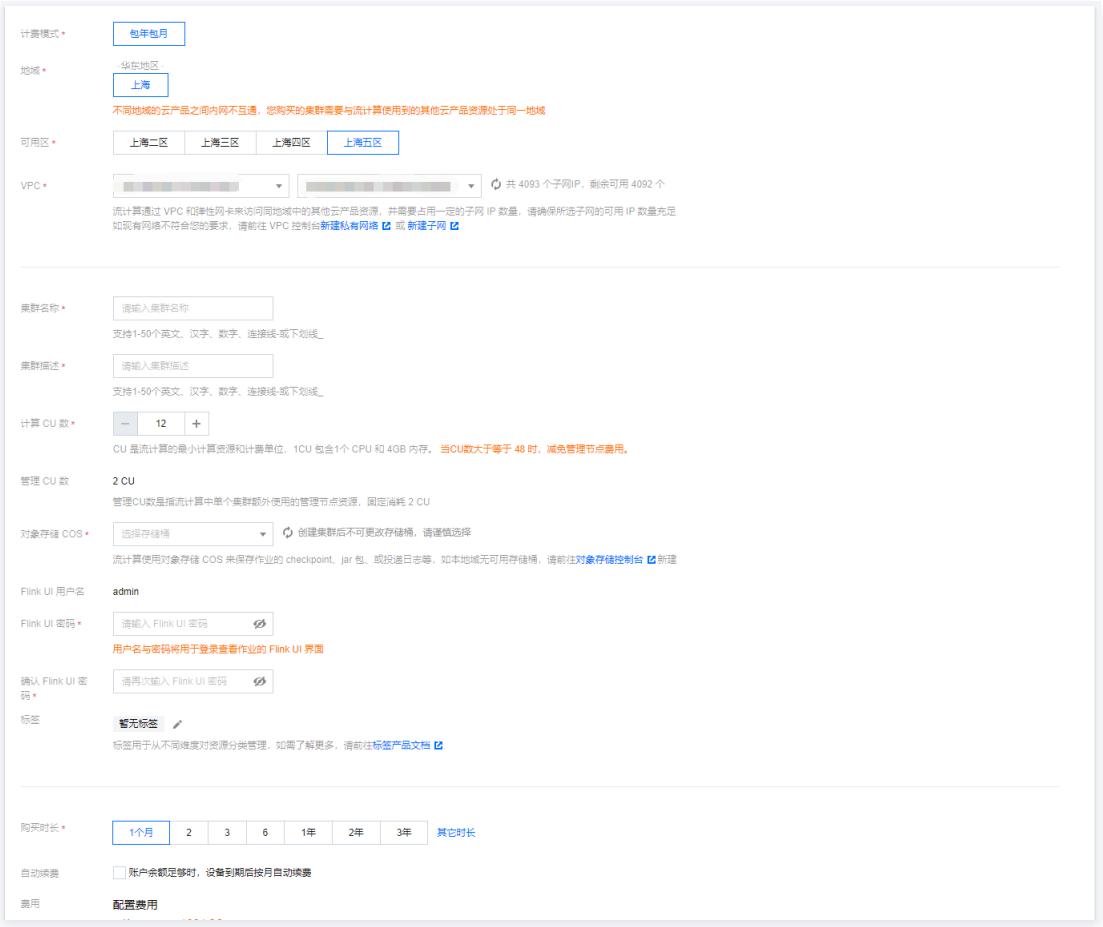


Stream Compute Resource Group

Last updated：2024-11-22 16:48:34

1. In the Stream Compute Resource Group tab page, click **Creating Cluster** to access the resource configuration page.





Parameter	Description
Billing mode	Monthly subscription
Region	Cloud products in different regions cannot communicate over private network. The cluster you purchase needs to be in the same region as other cloud product resources used for stream computing
Availability Zone	Different partitions in the region
VPC	Stream computing accesses other cloud product resources in the same region through VPC and ENI, and it requires some subnet IP addresses. Please ensure that the selected subnet has sufficient available IP addresses. If the existing network does not meet your requirements, please go to the VPC Console to Create a VPC or Create a Subnet

©2013-2025 Tencent Cloud. All rights reserved.

Cluster Name	Supports 1–50 characters, including English, Chinese, digits, hyphens (-), or underscores (_)
Cluster Description	Supports 1–50 characters, including English, Chinese, digits, hyphens (-), or underscores (_)
Computing CU Count	CU is the smallest computing resource and billing unit in stream computing. 1 CU includes 1 CPU and 4GB of memory. When the number of CUs is 48 or more, management node fees are waived.
Management CU Count	Management CU Count refers to additional management node resources used by a single cluster in stream computing, which is a fixed consumption of 2 CUs
Cloud Object Storage (COS)	Stream computing uses Cloud Object Storage (COS) to save job checkpoints, jar packages, or delivery logs. If there is no available bucket in this region, please go to the COS Console to create one
Flink UI username and password	Username and password will be used to log in to view the Flink UI interface
Tag	Tags are used to categorize resources from different dimensions. For more information, please go to Tag Product Documentation
Purchase period	1 month, 2 months, 3 months, 6 months, 1 year, 2 years, 3 years, and custom Definition for duration
Auto-renewal	After selecting auto-renewal, if the account balance is sufficient, the device will be automatically renewed monthly upon expiration
Fees	Based on the selected specifications, duration costs are calculated, including compute resource costs and manage resource costs. Total cost = compute resource costs + manage resource costs

2. After configuration is complete, click **Buy Now** and complete the payment to proceed to the console page to view the Stream Computing Resource Group information.

©2013–2025 Tencent Cloud. All rights reserved.

Data Security Resource Group Configuration

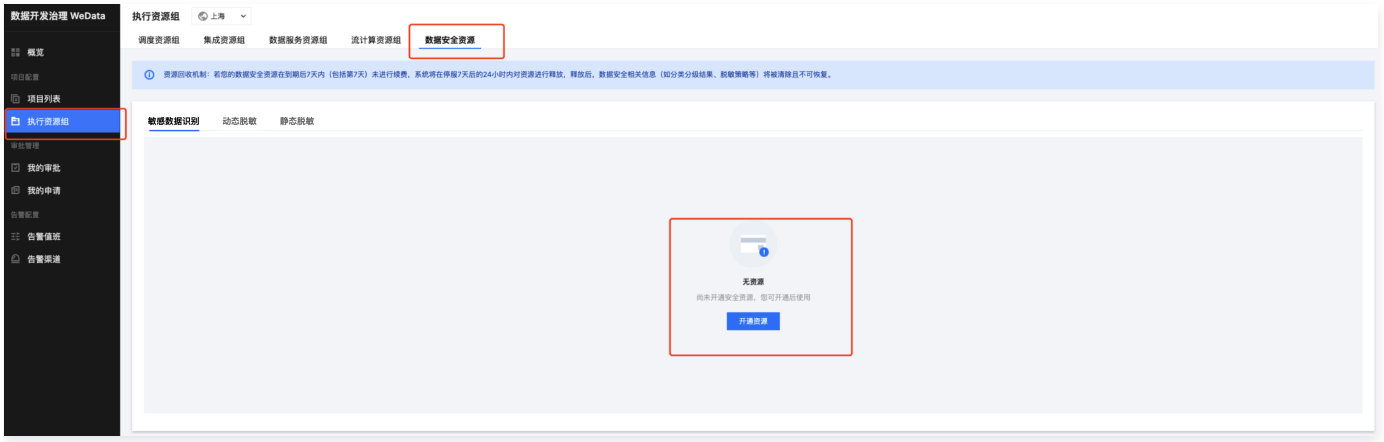
Last updated：2024-08-22 16:30:59

The Data Security Resource Group consists of three different types: Sensitive Data Identification, Dynamic Desensitization, and Static Data Masking. You can choose to purchase the features you need according to your requirements.

Create Data Security Resource Group

Sensitive data identification

1. Go to **Execution Resource Group** > **Data Security Resources** page, click on **Sensitive Data Identification**, click **Provision Resources**. You can choose different data security resources according to your needs:



2. Go to the **Sensitive Data Identification Resources Purchase** page and choose configurations according to your actual needs.



Parameter	Description
Package	Currently only the "Basic" Edition is supported, which includes the features: Classification and Grading Configuration, Classification and Grading Task, Sensitive Data Distribution.
Applicable Regions	Currently supported regions: Beijing, Shanghai, Chengdu, Guangzhou. (Please refer to the actual product purchase page for details) **Description:**

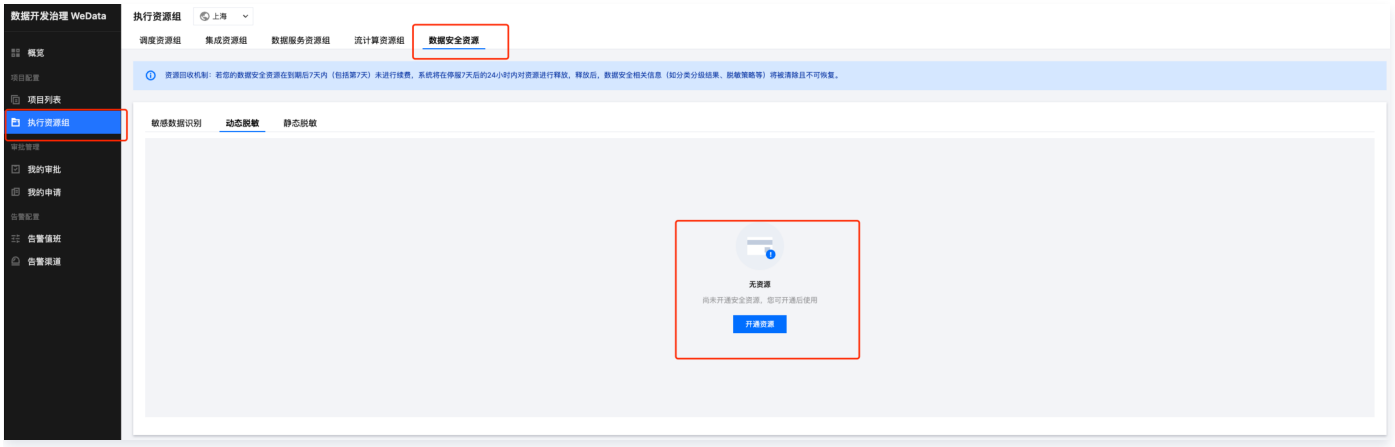
©2013-2025 Tencent Cloud. All rights reserved.

	Sensitive Data Identification is regionally shared, except for the international site, which means that after purchasing the package in the supported regions, it can be used in all regions.
Number of Database DBs	A minimum of 3 DBs are required by default, which can be increased as needed, sharing the DBs across regions.
duration	Supports purchasing duration of 3 months, 4 months, 5 months, 6 months, 1 year, 2 years, 3 years
Auto-renewal	After selecting auto-renewal, if the account balance is sufficient, the device will be automatically renewed monthly upon expiration

3. Once the configuration is complete, click **Confirm** to view the resource group information on the Data Security Resource Group page.

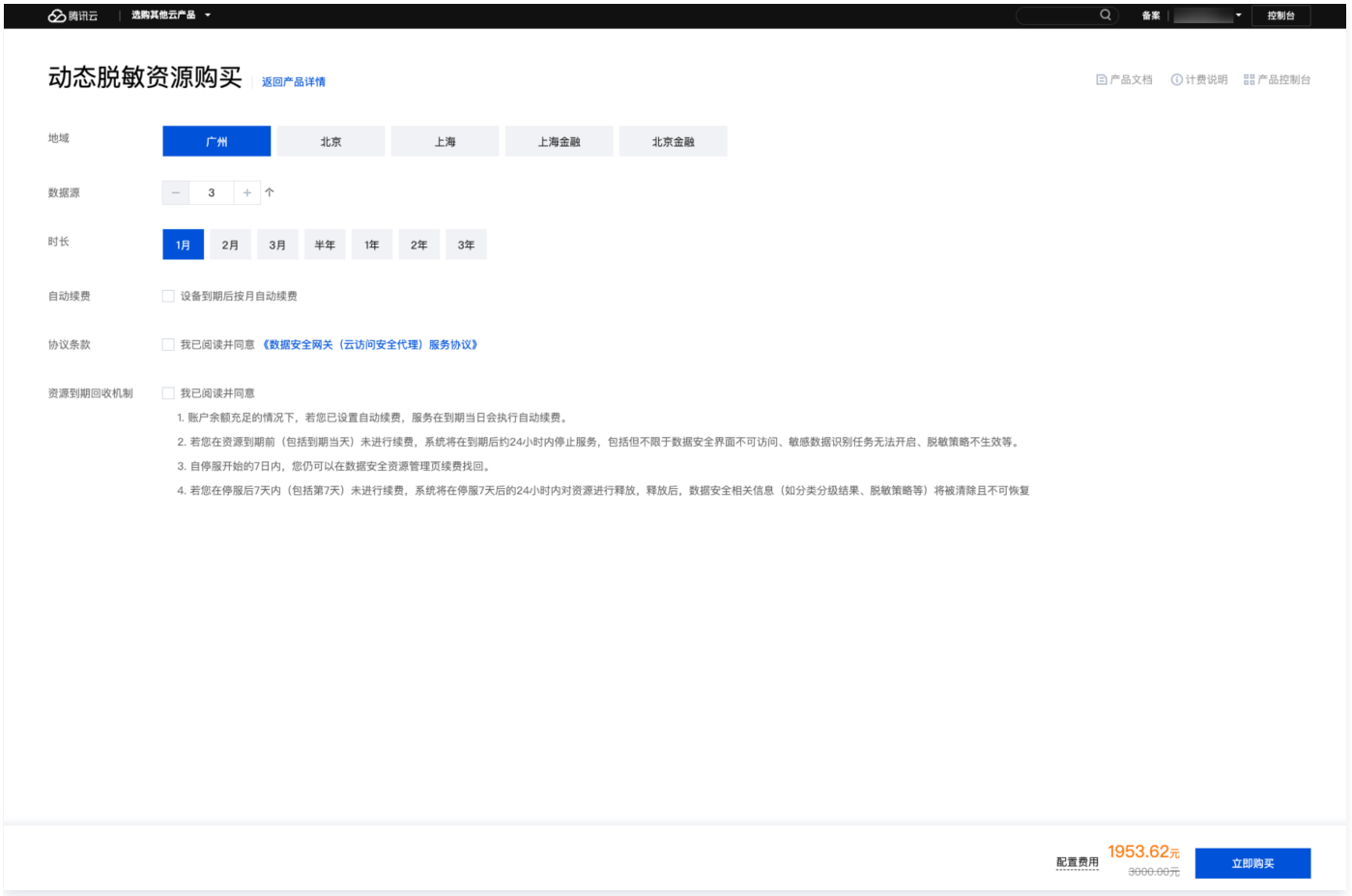
Dynamic Desensitization

1. Go to **Execution Resource Group** > **Data Security Resources** page, click on **Dynamic Desensitization**, click **Provision Resources**. You can choose different data security resources according to your needs:



2. Go to the **Dynamic Desensitization Resources Purchase** page and choose configurations according to your actual needs.

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Region	Currently supported regions: Beijing, Shanghai, Guangzhou, Shanghai Finance, Beijing Finance. (Please refer to the actual product purchase page for details)
Data Source	The default minimum is 3, which can be increased as needed. The data source refers to the data source information in WeData. For details, please refer to Data Source Management.
duration	Supports purchasing duration of 1 month, 2 months, 3 months, 6 months, 1 year, 2 years, 3 years
Auto-renewal	After selecting auto-renewal, if the account balance is sufficient, the device will be automatically renewed monthly upon expiration

3. Once the configuration is complete, click **Confirm** to view the resource group information on the Data Security Resource Group page.

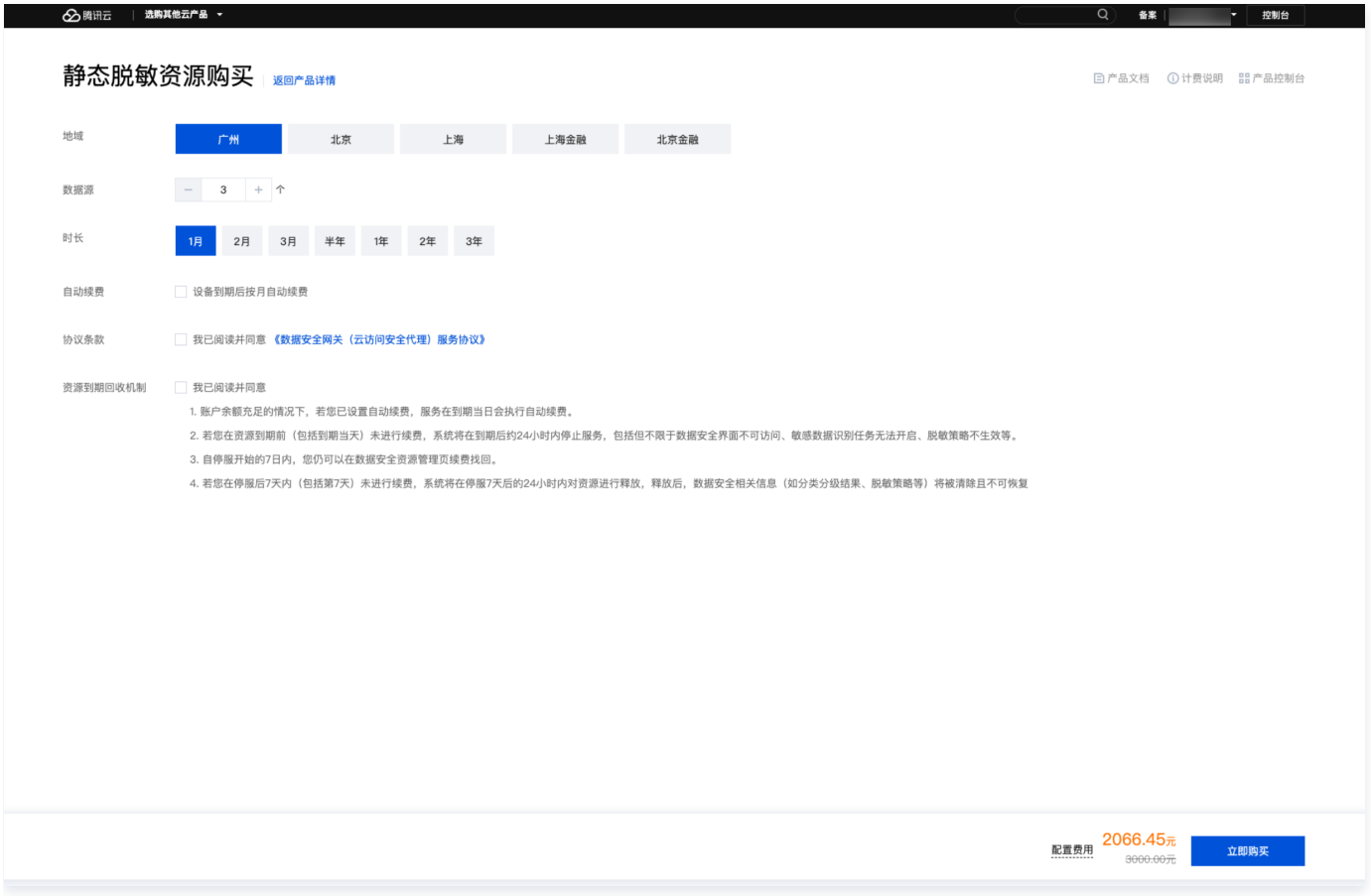
Static Data Masking

1. Go to **Execution Resource Group** > **Data Security Resources** page, click on **Static Data Masking**, click **Provision Resources**. You can choose different data security resources according to your needs:

©2013-2025 Tencent Cloud. All rights reserved.



2. Go to **Static Data Masking Resources Purchase** page and choose configurations according to your actual needs.



Parameter	Description
Region	Currently supported regions: Beijing, Shanghai, Guangzhou, Shanghai Finance, Beijing Finance. (Please refer to the actual product purchase page for details)
Data Source	The default minimum is 3, which can be increased as needed. The data source refers to the data source information in WeData. For details, please refer to Data Source Management.
duration	Supports purchasing duration of 1 month, 2 months, 3 months, 6 months, 1 year, 2 years, 3 years

©2013–2025 Tencent Cloud. All rights reserved.

Auto-renewal	After selecting auto-renewal, if the account balance is sufficient, the device will be automatically renewed monthly upon expiration

3. Once the configuration is complete, click **Confirm** to view the resource group information on the Data Security Resource Group page.

Expand Data Security Resource Group

When resources are insufficient, you can expand resources without shutting down the service.



You can expand the Data Security Resource Group as needed, click the + button to add an expansion pack.
Sensitive Data Identification:



Dynamic Desensitization:



Static Data Masking:

©2013-2025 Tencent Cloud. All rights reserved.



Renew Data Security Resource Group



If auto-renewal is not needed, you can turn off the auto-renewal switch:



Associate Data Source/Database

1. After purchasing resources, they need to be associated with the corresponding data sources/databases for **sensitive data identification** or **data desensitization operations**.
2. Go to the Associate Data Source page, click the **Increase Binding** button to bind the data source/database.



3. Go to the Increase Binding Data Source page, select the data source and database to associate, then click **OK** to complete the binding.

©2013-2025 Tencent Cloud. All rights reserved.



Unbind Data Source/Database

If the data source/database no longer needs security-related operations, you can unbind it to free up space.

1. Go to the Associate Data Source page, select the data source you need to unbind, click the **Unbind** button in the action bar to complete the unbinding.
2. You can also unbind in bulk, select multiple data sources to unbind, and click **Bulk Unbind** to unbind all selected data sources.



©2013-2025 Tencent Cloud. All rights reserved.

Resource Group Configuring Public Network

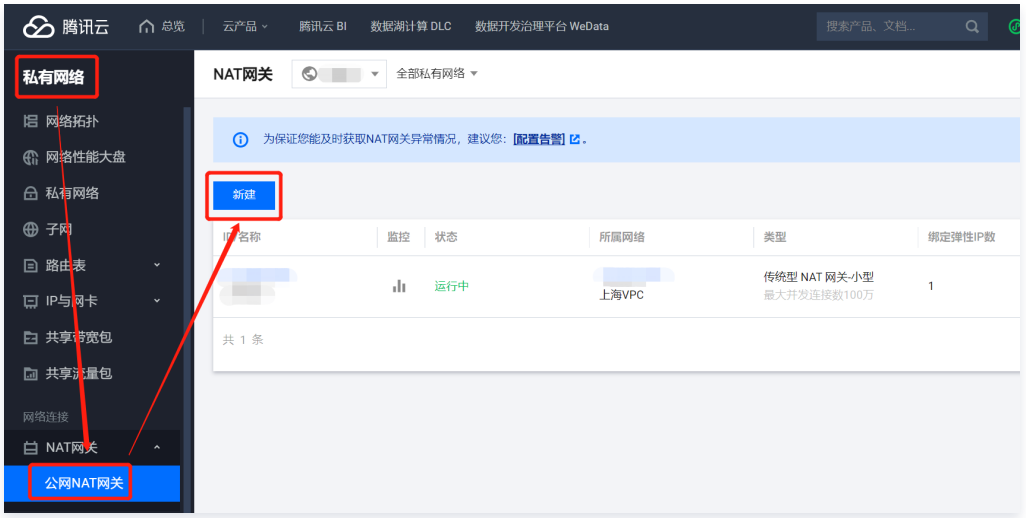
Last updated：2024-11-22 16:48:58

Operation scenarios

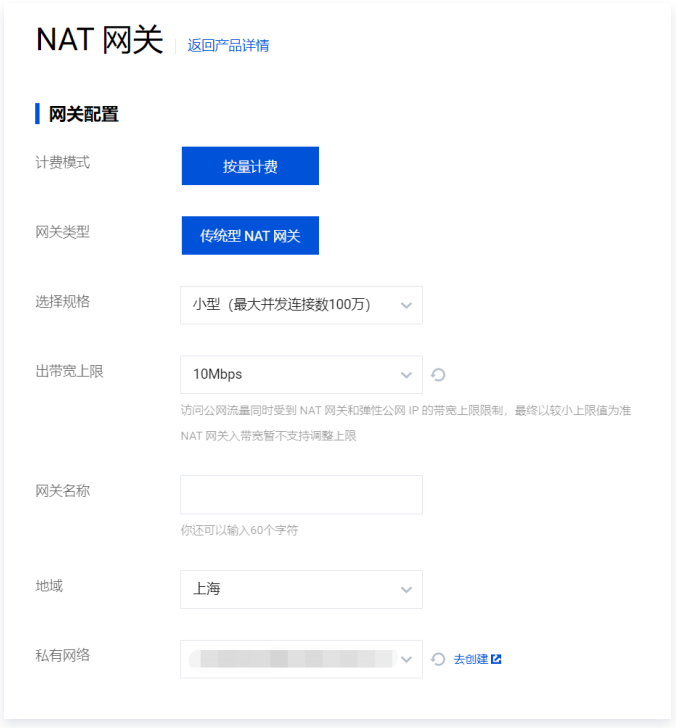
Applicable for configuring public network access for a Resource Group's VPC.

Operation step

1. Enter Tencent Cloud **VPC** > **NAT Gateway** . In the corresponding region, click **New** NAT Gateway.

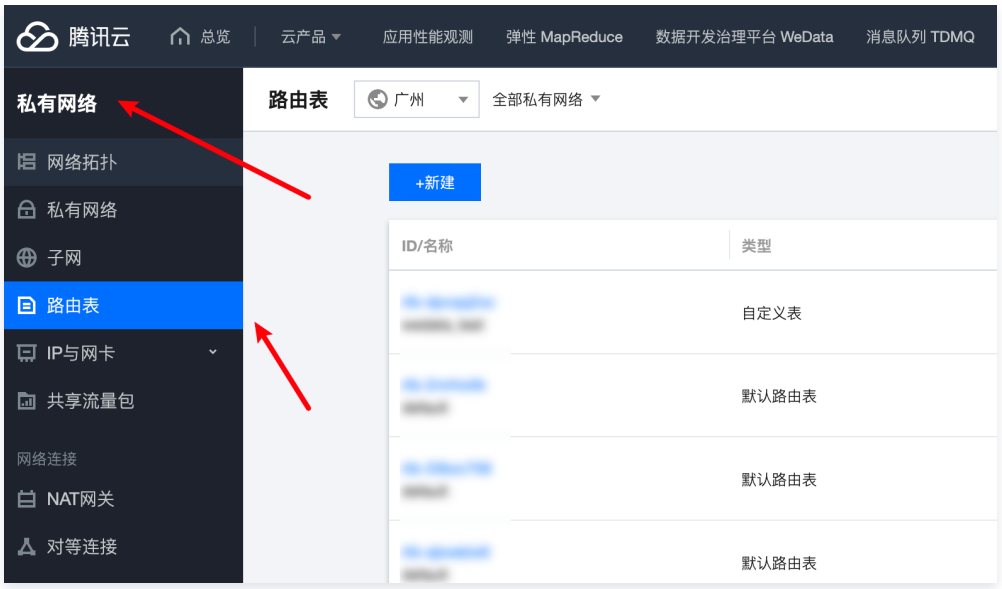


2. Purchase NAT Gateway:

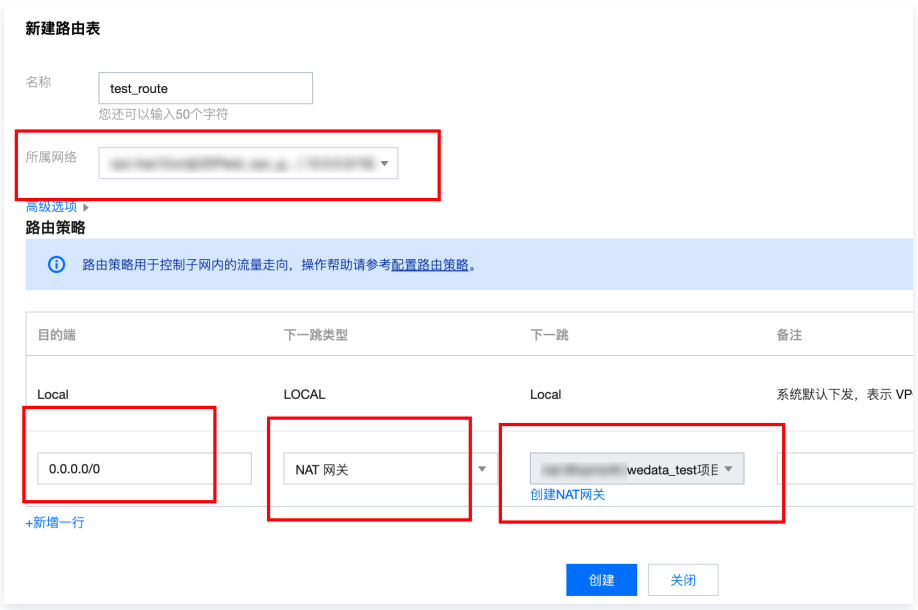


3. Enter Tencent Cloud **VPC** > **Route Table**, create and bind a route table pointing to the NAT Gateway.

©2013-2025 Tencent Cloud. All rights reserved.



4. Create a custom route table, select the VPC network, set the destination to 0.0.0.0/0 or a fixed route, select **NAT Gateway** as the Next Hop Type, and choose the newly created NAT Gateway as the next hop:



5. After the route table is created, enter the route table details and associate the subnet:



6. Select the subnet to be associated and bind it:

©2013–2025 Tencent Cloud. All rights reserved.



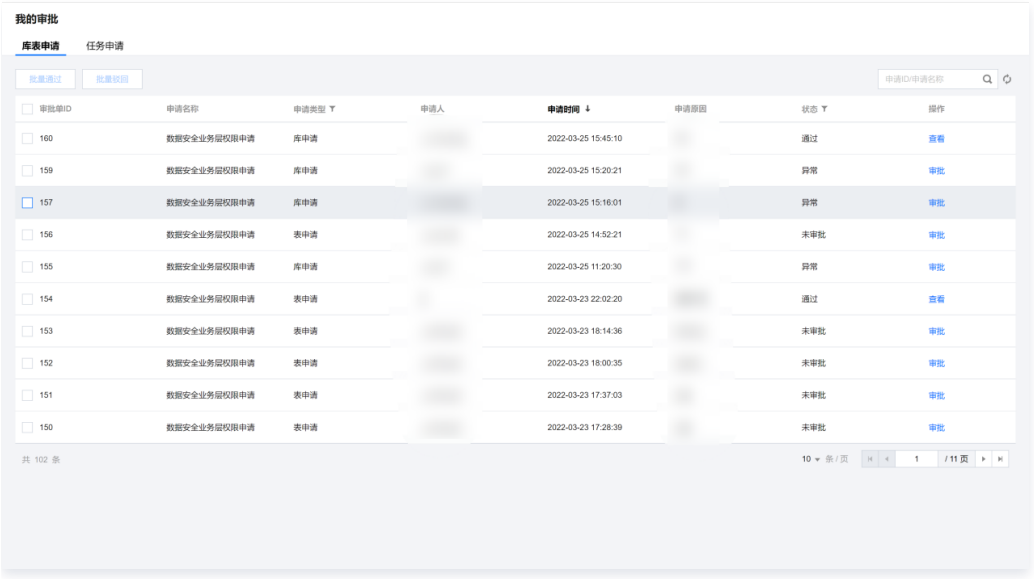
©2013–2025 Tencent Cloud. All rights reserved.

Approval management
My Approvals

Last updated: 2024-11-22 16:49:22

Database Table Applications

Enter the WeData Console, click **Approval Management** to enter the **My Approvals** page. Users can centrally manage database and table permission application approval forms that require this account's approval through the library table application page.



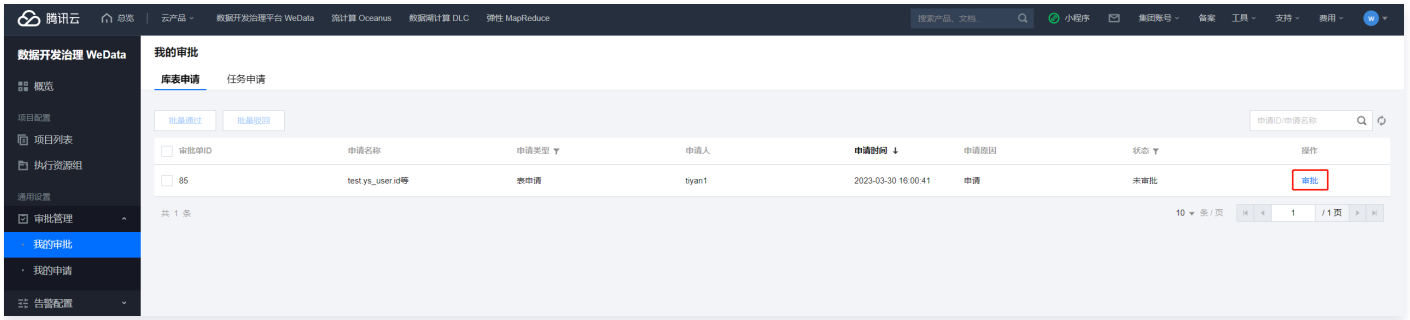
Information	Details
Approval Form ID	Approval form ID
Application Name	Application name
Application Type	Application type: Database application, Table application
Applicant	Username of the account creating the application
Application Time	Time of application and approval form creation
Application Reason	Applicant's Application Reason
Status	6 Approval Statuses: Not Approved, In Approval, Approved, Rejected, Canceled, Exception. Supports filtering approval forms based on approval status
Operation	You can perform viewing and approval actions on approval forms

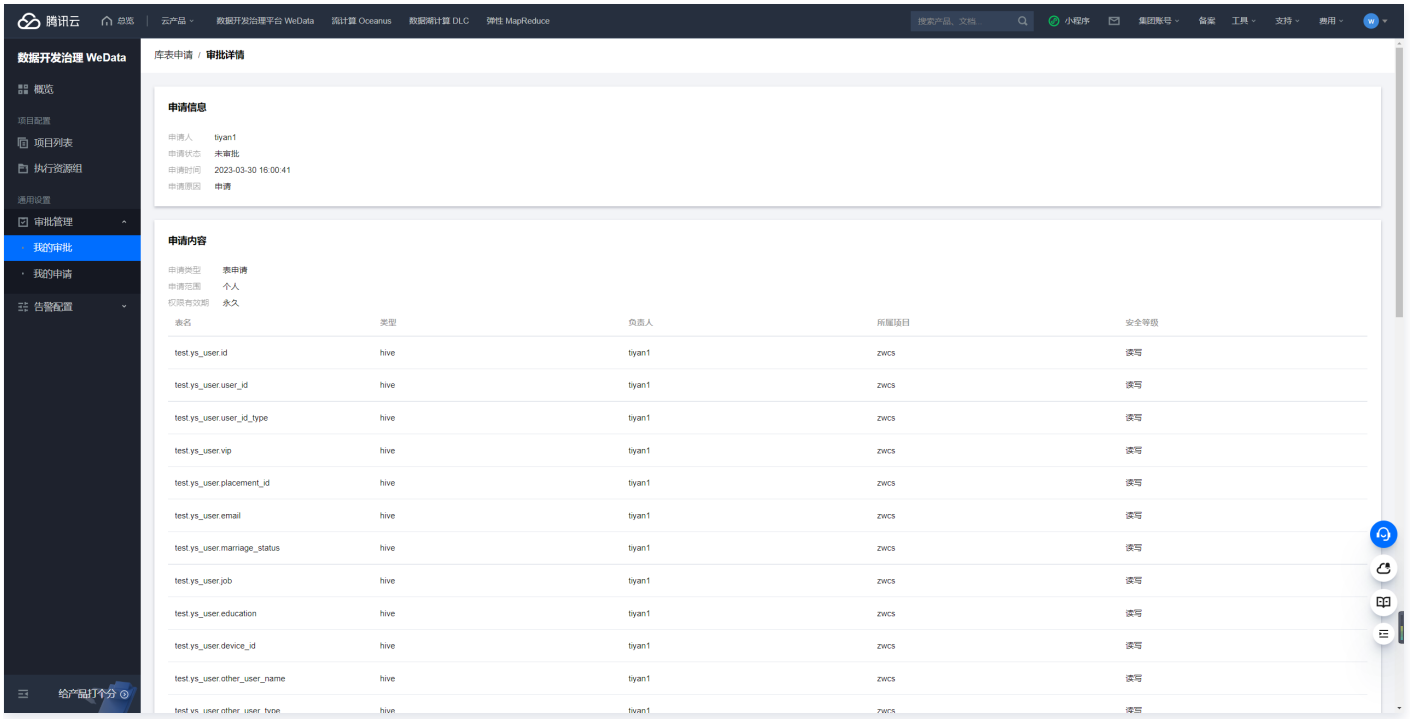
Approval

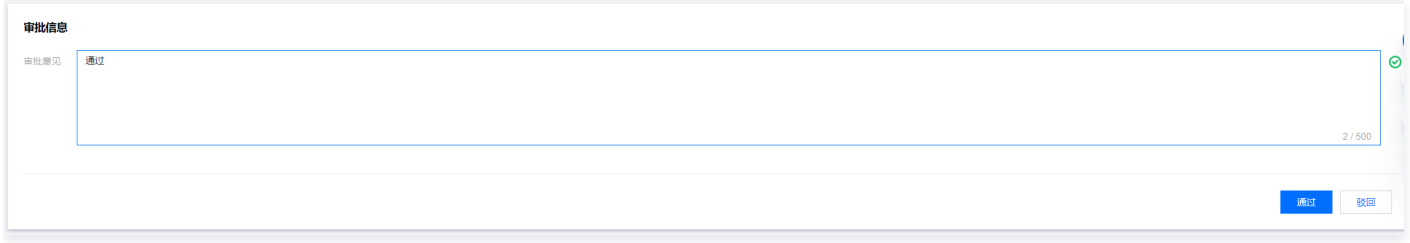
Table Application Approval

On the page, select the table application approval form that needs to be approved. Click the **Approve** button in the Operation column to approve the form. The information provided in the approval form includes applicant-related information and data table information. Users can choose to **Approve** or **Reject** the approval form.

©2013-2025 Tencent Cloud. All rights reserved.

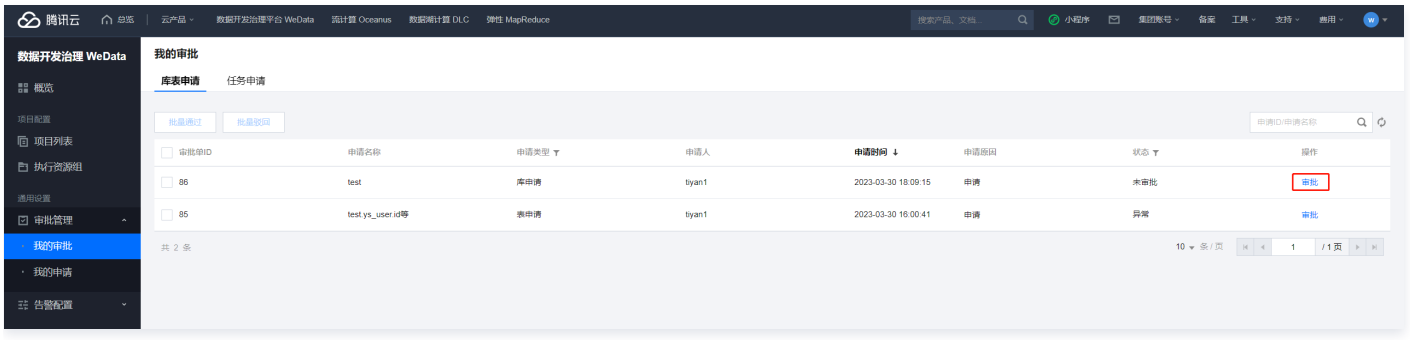




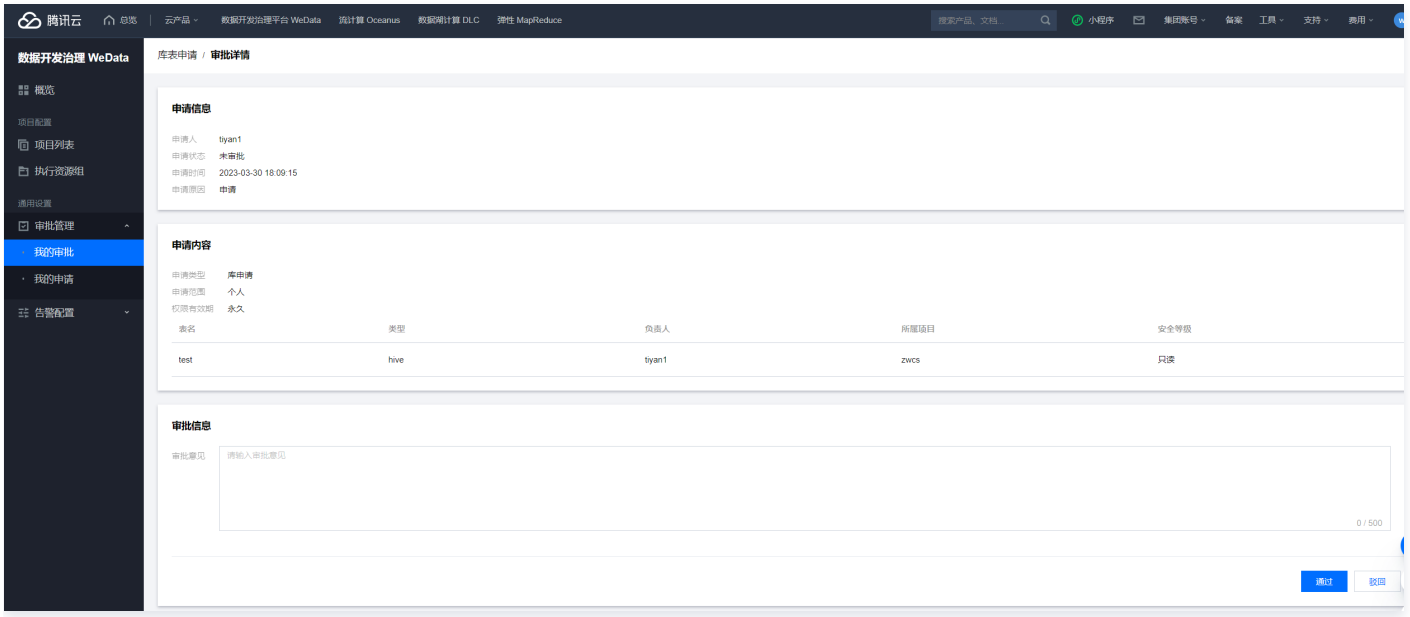


Database Application Approval

On the page, select the database application approval form that needs to be approved. Click the **Approve** button in the Operation column to approve the form. The information provided in the approval form includes applicant-related information and database information. Users can choose to **Approve** or **Reject** the approval form.

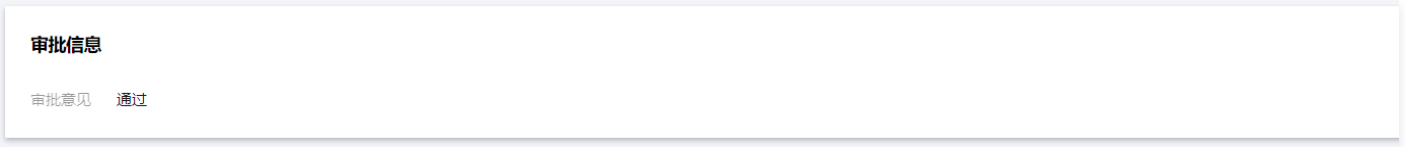


©2013–2025 Tencent Cloud. All rights reserved.



Viewing

On the page, click **View** to view the information of the approved approval forms, including the applicant-related information, approved database and table information, and approval comments.



Task Application

Users can centrally manage task run approval forms that require this account's approval through the Task Application page. For task application process settings, please go to **Wedata Project Management > Basic Information Configuration > Approval Configuration**.



Information	Details
Approval Form ID	Approval form ID
Application Name	Application name

©2013-2025 Tencent Cloud. All rights reserved.

Application Type	Application Types: Cross-Workflow Task Dependency Application, Task Submission Application
Applicant	Username of the account creating the application
Application Time	Time of application and approval form creation
Application Reason	Applicant's Application Reason
Status	6 Approval Statuses: Not Approved, In Approval, Approved, Rejected, Canceled, Exception. Supports filtering approval forms based on approval status
Operation	You can perform viewing and approval actions on approval forms

Viewing

Click **View** to enter the details page of the current approval form.

Approval

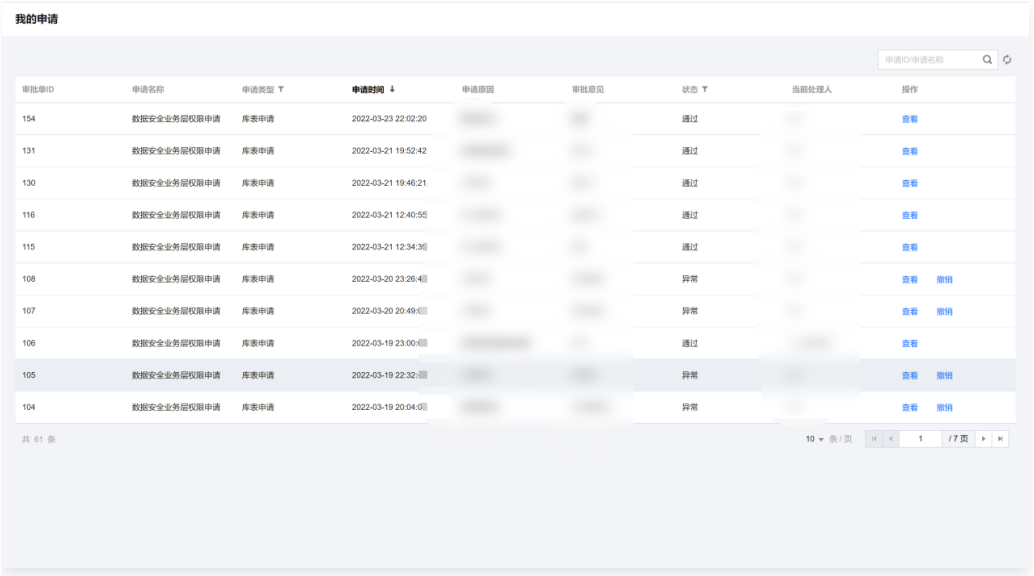
After filling in approval comments, you can approve or reject the current approval form.

©2013-2025 Tencent Cloud. All rights reserved.

My Applications

Last updated：2024-11-22 16:49:42

Enter WeData Console, click **Approval Management** to go to the **My Applications** page. This page includes database table application permissions, database application permissions, API application permissions, task application permissions submitted by the current user. Application forms can be managed through the application form list.

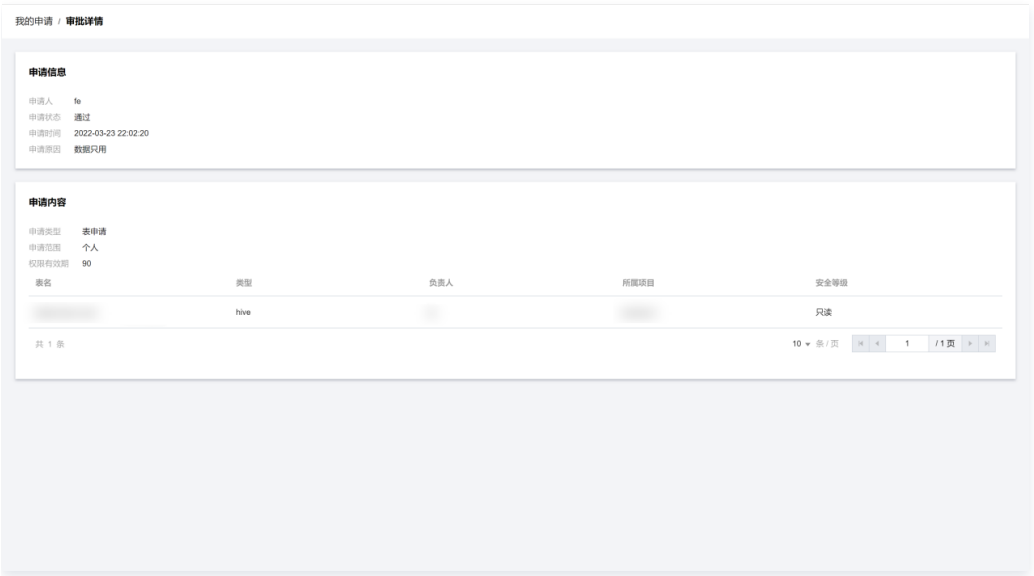


Information	Details
Approval Form ID	Approval form ID number
Application Name	Application name
Application Type	Application Type: Database Table Application, Task Application
Application Time	Time of application and approval form creation
Application Reason	Request reason provided by the applicant
Approval Comments	Approval comments provided by the approver
Status	6 Approval Statuses: Not Approved, In Approval, Approved, Rejected, Canceled, Exception. Supports filtering approval forms based on approval status
Current Handler	The username handling this approval form
Operation	You can view or revoke the approval form

Viewing

Select the corresponding permission application form, click **View** to enter the approval details page. This page displays the **Application Information** and **Application Content** of the application form.

©2013-2025 Tencent Cloud. All rights reserved.



> **Note**
> Only permission application forms in the Completed (Approved or Rejected) status can view approval details.

Revoke

Select the corresponding permission application form, click **Revoke**. In the pop-up confirmation box, click **Confirm** to complete the revocation operation.



> **Note**
> Only permission application forms that are not approved can be revoked.

©2013–2025 Tencent Cloud. All rights reserved.

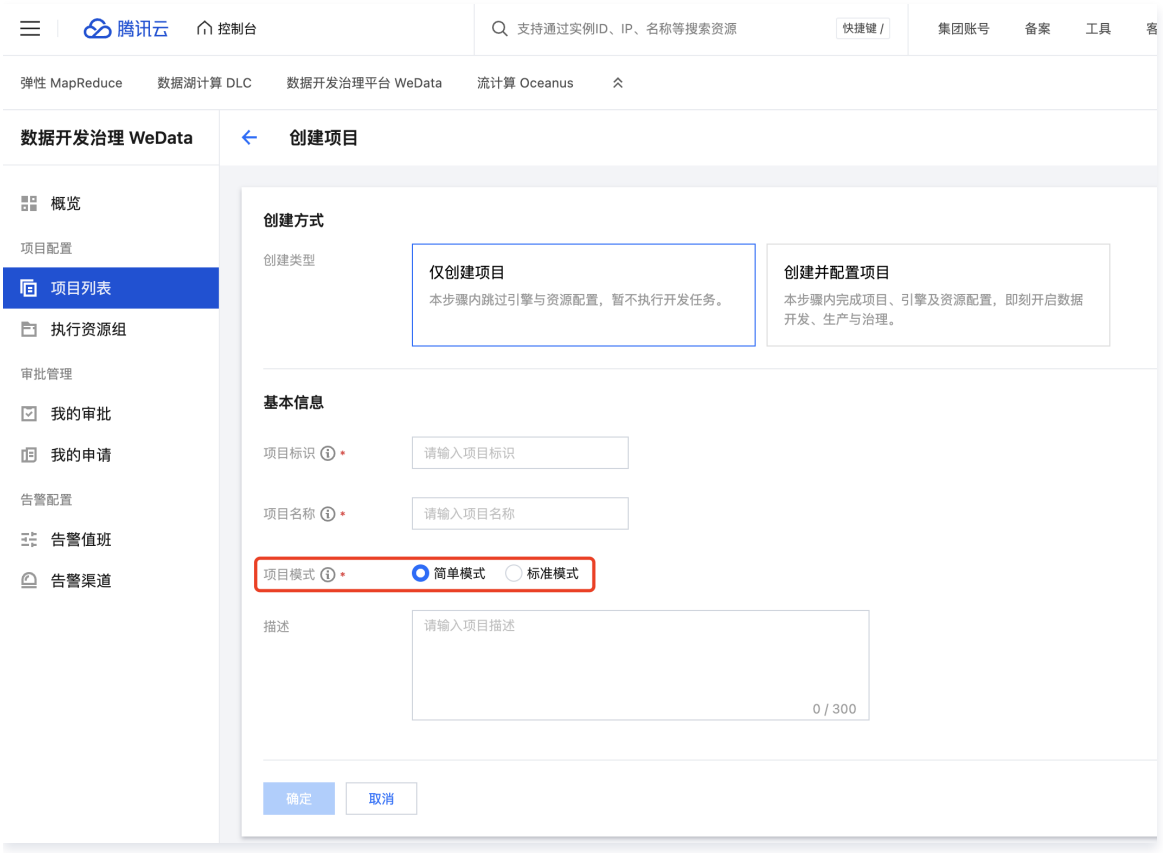
Project Management
Simple mode and standard mode

Last updated: 2024-08-24 18:16:16

To meet customer needs in various data development and data management scenarios, WeData Project Space Management offers Simple Mode and Standard Mode. Different project modes correspond to different engine configurations, data management, and development processes. This article will provide a detailed introduction to the two modes of the WeData project.

> **Note**
> Currently, the standard project mode is in the invitation testing phase. You can click Trial Application to apply.

After the trial is activated, you can see the project mode selection when creating a project. Additionally, the project mode can also be seen in the basic project information, supporting upgrades from Simple Mode to Standard Mode.



Standard Project Mode Usage Restrictions

> **Note:**
> 1. In the WeData standard project mode, it currently does not support binding different EMR instances for the development environment and production environment, different TCHouse-P clusters, or DLC engines in different regions.
> 2. In the WeData standard project mode, operations like data table version management and rollback are not supported in Data Management. Therefore, tables deleted in the development environment cannot be recovered via the Data Management visualization interface.
> 3. After enabling the WeData standard project mode, the cross-project cloning feature is no longer provided. Tasks and other objects in a standard mode project cannot be published to other projects through cross-project cloning, nor can tasks and other objects from other projects be published to a project with standard mode enabled through cross-project cloning.
> 4. For SQL tasks, SparkSQL and Trino are currently not supported for development and production isolation.

©2013-2025 Tencent Cloud. All rights reserved.

Differences between the two modes

Below, we will compare the differences between the two modes in terms of project management, task development processes, other objects involved in data development, DataInLong, etc.

Affected Modules	Simple Mode	Standard Mode
Project Management – Storage and Computing Engine Configuration	A project-bound engine has only one default environment, with no distinction between development and production environments.	An engine bound to a project has two environment configurations: Development Environment and Production Environment. Currently, an engine instance supports two databases, e.g., the development environment database is set to db_dev and the production environment database is set to db_prod.
Project Management – System Source	A system source automatically generates a data source environment configuration with no distinction between development and production environments.	A system source automatically generates a data source. Corresponding to two environment configurations: Development Environment Configuration and Production Environment Configuration.
Project Management – Custom Data Source	A custom data source corresponds to only one environment configuration with no distinction between development and production environments.	A custom data source corresponds to two environment configurations: Development Environment Configuration and Production Environment Configuration.
Project Management – Approval Configuration	Task submission approval is enabled by default. It can be disabled.	Task submission approval is disabled by default. Release approval configuration is enabled by default and cannot be disabled.
Task Development Process	Task development process includes: **Create** > **Develop** > **Debug\Save** > **Submit** > **Schedule OPS** 	Task development process includes: **Create** > **Develop** > **Debug\Save** > **Submit** > **Publish**>**Schedule OPS**. 
Task Data Access	Access production data using identities configured in project management.	Development and production environments can use different access accounts. Task debugging uses personal accounts, while scheduling runs can use a specified unified account.
Task development involves other objects	Function: Takes effect immediately after submission. Workflow, project parameters, resources, data table take effect immediately after saving.	Function and data table take effect in development environment instance\database after submission, and in production environment instance\database after publishing. Workflow, project parameters, resources take effect in development environment after upload and saving and submission, and in production environment after publishing.
DataInLong – Offline Synchronization	Task process: **Create** > **Develop** > **Debug\Save** > **Submit** > **Production Environment** Development debugging uses datasource's environment configuration. After submission to scheduling, it still uses the	Task process: **Create** > **Develop** > **Debug\Save** > **Submit** > **Production Environment** Development and debugging use the data source's development environment. After submission to the production environment, it is replaced with the data source's production environment. If the database

©2013–2025 Tencent Cloud. All rights reserved.

	datasource's environment configuration.	configured during the development and debugging stage is not the data source's development and production environment database, no replacement occurs. For example, the development environment database is set to db_dev, the production environment database is set to db_prod. • Offline Synchronization Task configuration uses db_dev, Cron Scheduling uses db_prod. • Offline Synchronization Task configuration uses db_prod, Cron Scheduling uses db_prod. • Offline Synchronization Task configuration uses db_other, Cron Scheduling still uses db_other. **Note:** The current Project Standard Mode does not affect the DataInLong > Offline Synchronization Task development flow. It affects the Orchestration Space > Offline Synchronization Task development flow.
DataInLong – Real-time Synchronization	Task: **Create** > **Develop** > **Debug\Save** > **Submit** > **Production Environment** Development debugging uses datasource's environment configuration. After submission to scheduling, it still uses the datasource's environment configuration.	Task: **Create** > **Develop** > **Debug\Save** > **Submit** > **Production Environment** Development and debugging use the data source's production environment configuration. After submission to the production environment, it still uses the data source's production environment configuration. For example, the development environment database is set to db_dev, the production environment database is set to db_prod. • Real-time Synchronization Task configuration can only be selected to db_prod. After submission of the real-time task, it uses db_prod. • Real-time Synchronization Task configuration uses db_other. After submission of the real-time task, it uses db_other. **Note:** The current Project Standard Mode does not affect the DataInLong > Real-time Synchronization Task development flow.

Use cases for two modes

- **Simple Mode**: Suitable for small data development teams that do not require a very strict data development process. It is simple and fast to use, and the roles of data development and data operation and maintenance can complete all the data development and maintenance work.
- **Standard Mode**: Suitable for medium to large data development teams with more standardized data development process requirements and stricter production environment data permissions. The development process is more standardized and secure, requiring collaboration between data development, data operation and maintenance, and publisher approver.

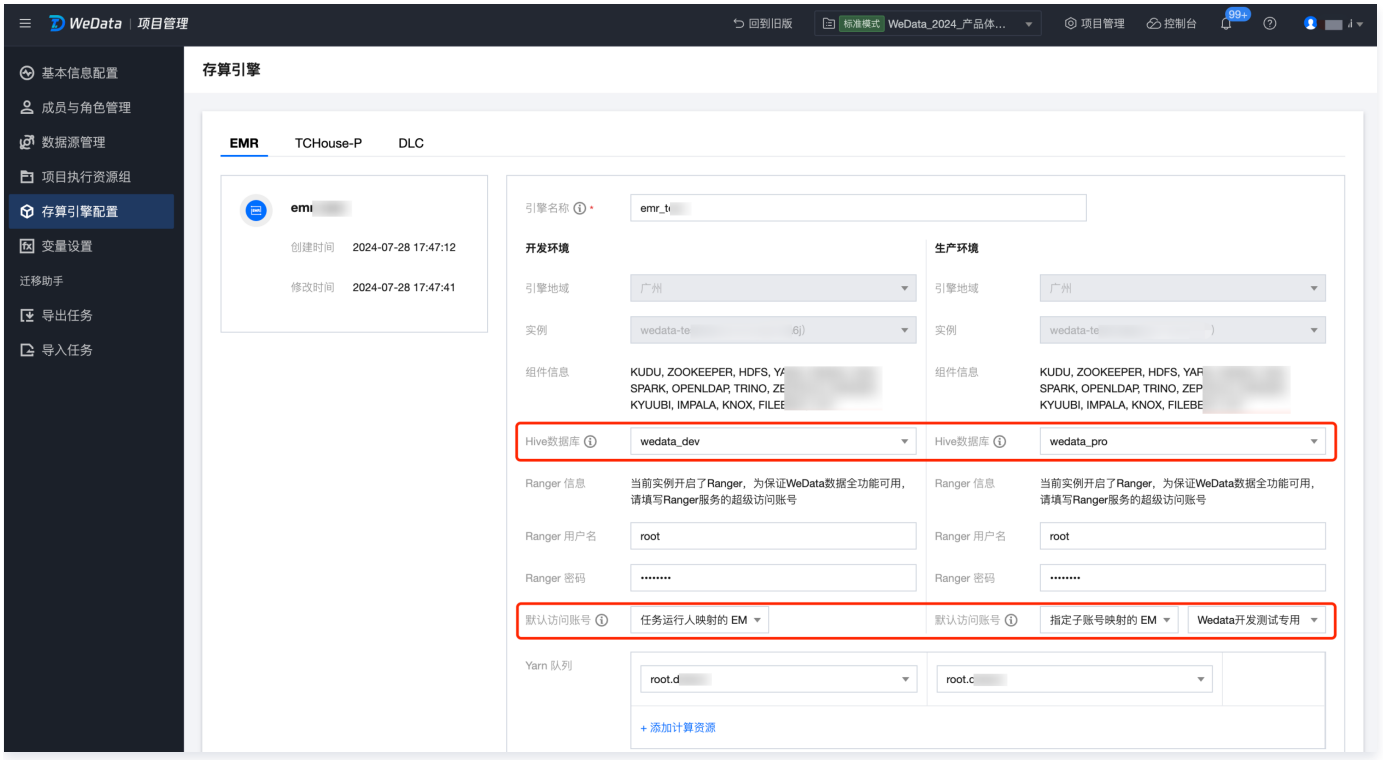
Configuration methods and working principle of Project Standard Mode

When binding the engine and creating the data source in Standard Mode, it provides configuration for both the development environment and the production environment, along with account configuration for accessing these environments. For the computational storage engine bound to the project, the system automatically generates data sources based on the engine configuration information. Data sources have their own development and production environment attributes, typically corresponding to two different JDBC connection strings.

©2013–2025 Tencent Cloud. All rights reserved.

The following is a simple example using the configuration of EMR.

1. In the computational storage engine configuration, set the development environment and production environment to use the same EMR Cluster Instance but with different databases. For example, the development environment database is wedata_dev, and the production environment database is wedata_pro. The default account for the development environment is the EMR account mapped by the task operation, while the account for accessing the production environment is the EMR account mapped by the specified sub-account.



2. According to the above engine configuration, the system generates system data sources by default. Since the cluster for the development and production environments is the same but the databases are different, the generated Hive data sources are as follows:

- Development Environment: `jdbc:hive2://ip:port/wedata_dev`
- Production Environment: `jdbc:hive2://ip:port/wedata_pro`

3. When developing and testing running HiveSQL tasks in the orchestration space, the development environment configuration of the data source (`jdbc:hive2://ip:port/wedata_dev`) will be used, and the account used is the corresponding EMR account of the person running it. After the task is submitted and published, during the cron scheduling of task operation and maintenance, the production environment configuration of the data source (`jdbc:hive2://ip:port/wedata_pro`) will be used, and the account used is the specified sub-account configured in the production environment.
4. Therefore, in HiveSQL's SQL statements, if the corresponding database name is not written, it will automatically read/write the wedata_dev database during development and testing, and read/write the wedata_pro database during cron scheduling. This achieves the goal of using the development database during development and the production database during production.

```
insert into user_info
select * from  table_1
-- Debug run reads data from wedata_dev.table_1 and writes to wedata_dev.user_info
-- Cron run reads data from wedata_pro.table_1 and writes to wedata_pro.user_info
```

5. If the corresponding database is written in the SQL statements of HiveSQL, the database in the SQL statement will still be used, and the effect of isolating the development environment from the production environment cannot be achieved.

```
insert into other_db.user_info
select * from  other_db.table_1
-- Debugging run reads data from other_db.table_1 and writes it into other_db.user_info
```

©2013-2025 Tencent Cloud. All rights reserved.

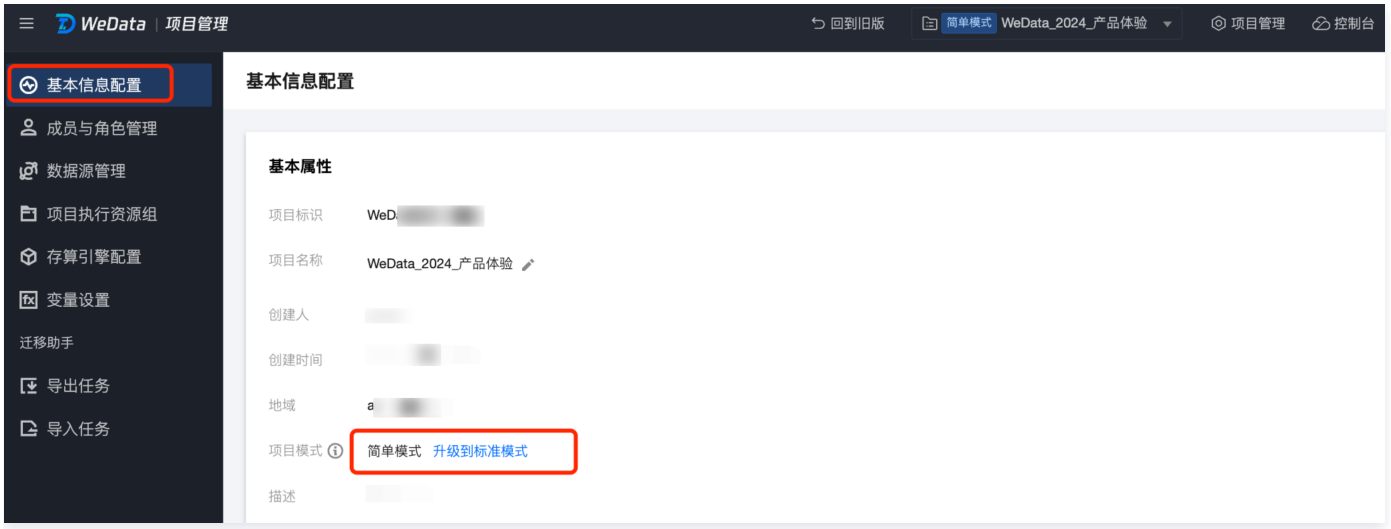
```
-- Periodic run reads data from other_db.table_1 and writes it into other_db.user_info
```

> **Note:**
> Currently, in the SQL task types in WeData orchestration space, the Hive data source for SparkSQL tasks and Trino tasks cannot achieve isolation between development and production environments through the above method. It is recommended to use project parameters to achieve this isolation.

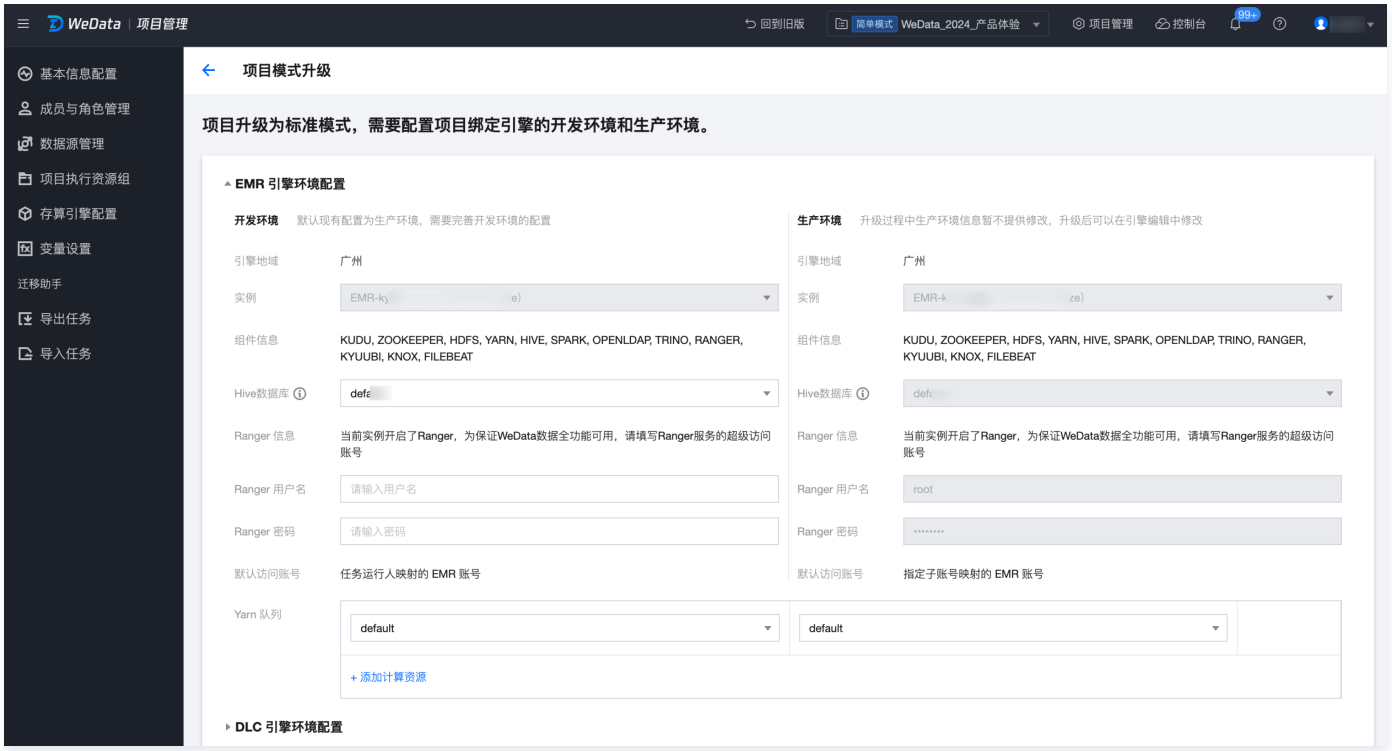
Simple mode upgrade to standard mode

WeData provides a feature to upgrade project simple mode to standard mode. Users can upgrade existing simple mode projects to standard mode. The upgrade process is irreversible, so please evaluate and proceed with caution.

1. In WeData **Project Management** > **Basic Information Configuration** > **Project Mode** attributes, you can check the current project mode. If it is simple mode, click the **Upgrade to Standard Mode** button next to proceed with the upgrade.

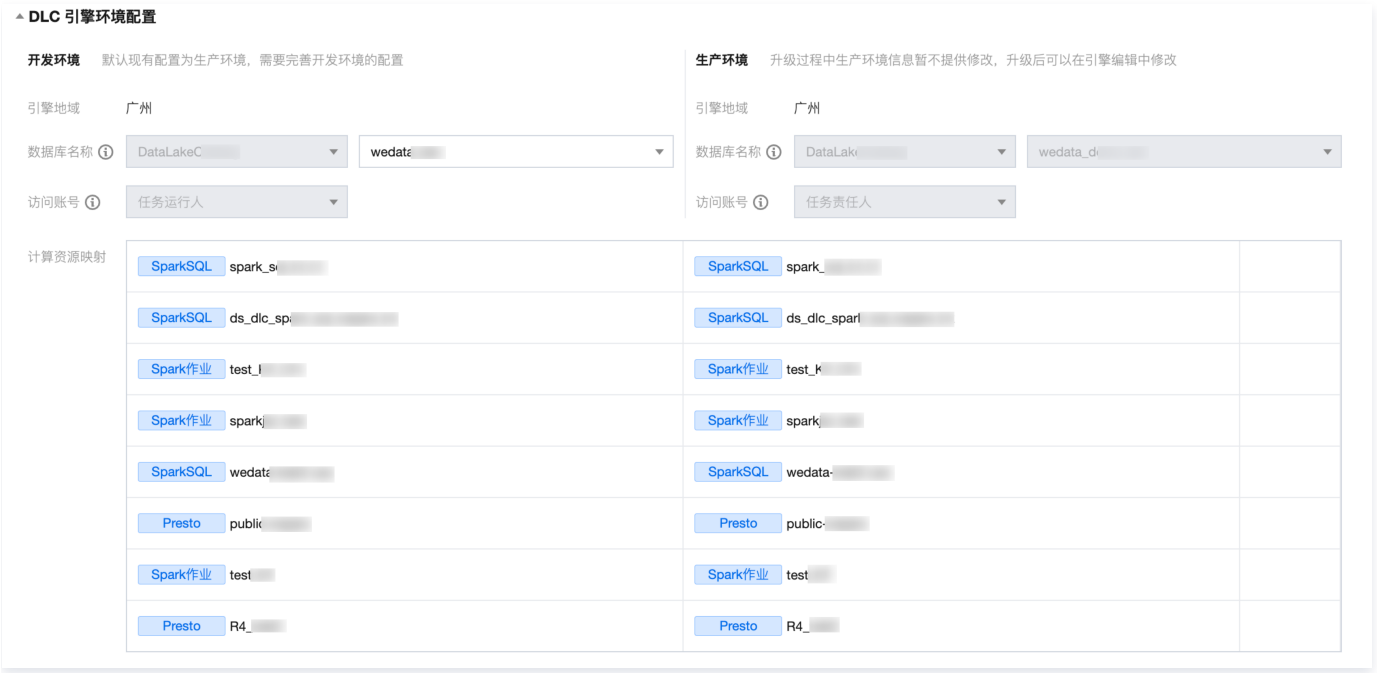


2. Upgrade conditions will be checked before the upgrade process. After passing the check, you can enter the upgrade page. The following images show the upgrade interface for EMR, DLC, and self-defined data sources.

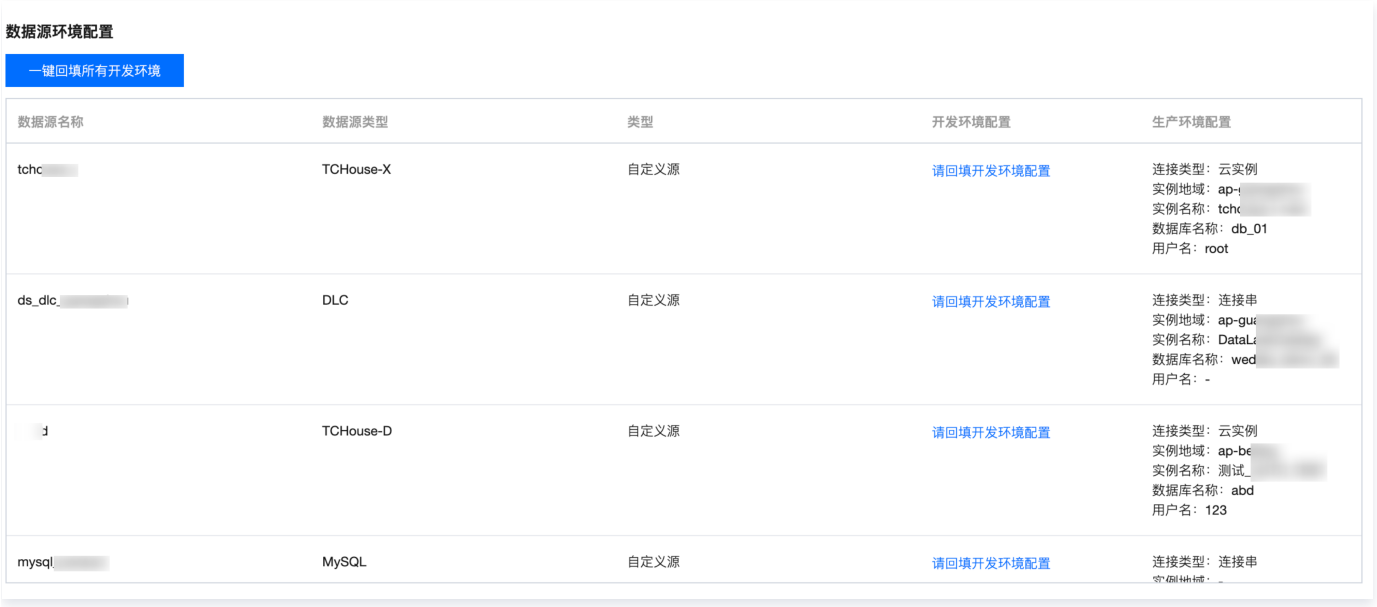


- DLC Engine Environment Configuration:

©2013–2025 Tencent Cloud. All rights reserved.



- EMR Data Source Environment Configuration:



3. Complete and confirm the compulsory parameters during the upgrade process, which include configuring the computational storage engine development environment. Confirm the backfill data source development environment configuration, check the **Agree to upgrade** checkbox, and click the **Upgrade** button to complete the upgrade. The upgrade duration depends on the number of computational storage engines and data sources in the project.

Upgrade Notice

Note

1. Before upgrading, the system will automatically check whether the current project meets the upgrade conditions. Only if the upgrade conditions are met can the upgrade proceed.
2. During the upgrade process, the current computational storage engine configuration information of the project will be set to the production environment configuration by default, and users need to fill in the development environment configuration. At the same time, the current data source configuration information of the project will be set to the production environment configuration by default. Users need to confirm the one-click backfill development environment configuration to proceed with the upgrade normally.

©2013–2025 Tencent Cloud. All rights reserved.

3. After upgrading, the cross-project cloning feature will no longer provide the new feature. The historical cloning information will be retained. The upgraded project can no longer be used as a target project for cross-project cloning.
4. After upgrading, task submission approval will be turned off by default, but approval for publishing tasks and other objects will be enabled and cannot be turned off.
5. The upgrade process is irreversible. After upgrading, you cannot return to Simple mode.

©2013-2025 Tencent Cloud. All rights reserved.

Basic Information Configuration

Last updated：2024-12-12 20:17:55

After clicking **Basic Information Configuration** in the left directory tree, you can view the basic attributes of the current project, as well as the data development scheduling and approval-related permission management switches.

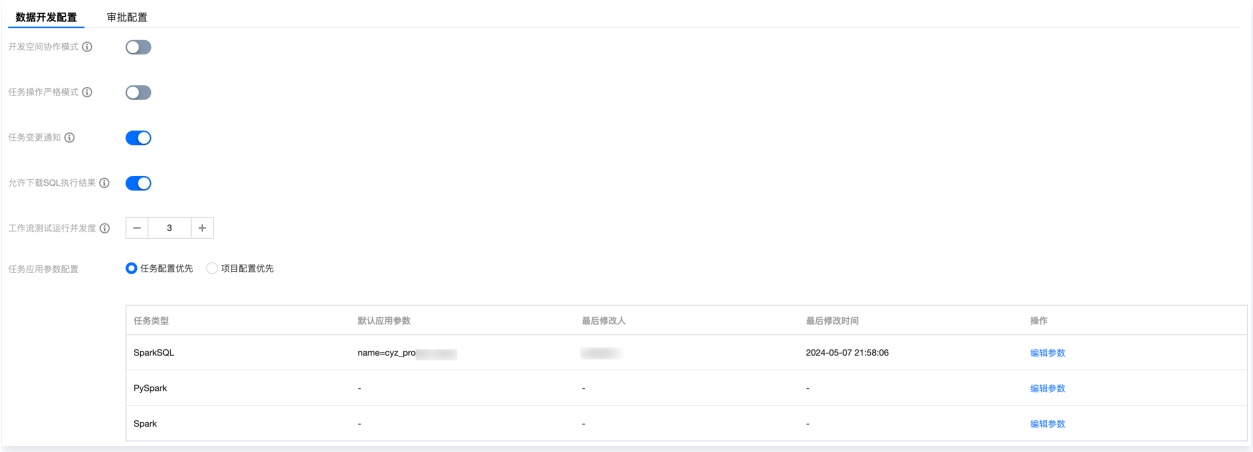
Basic Attributes

Basic attributes allow you to view the project's project identification, project name, creator, creation time, region, and description information. Among them, the project name and project description support editing and modification.



Data Development Configuration

Basic Configuration



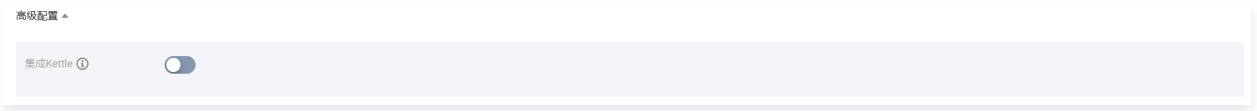
1. **Development Space Collaboration Mode**: Once enabled, the development space can edit and run scripts referenced by tasks created in the orchestration space.
2. **Strict Task Operation Mode**: Once enabled, tasks that are dependencies for downstream tasks cannot be paused, stopped, or deleted.
3. **Task Change Notification**: Once enabled, when tasks are changed and submitted by someone other than the person responsible, notifications can be sent to the task owner and the project admin.
4. **Allow SQL Execution Result Download**: Once enabled, data exploration and SQL execution results within the development space can be downloaded.
5. **Workflow Test Run Concurrency**: The number of tasks running concurrently during workflow test runs. (Note: workflow test runs also consume resources from the task scheduling resource group)
6. **Task Application Parameter Configuration:**

©2013-2025 Tencent Cloud. All rights reserved.

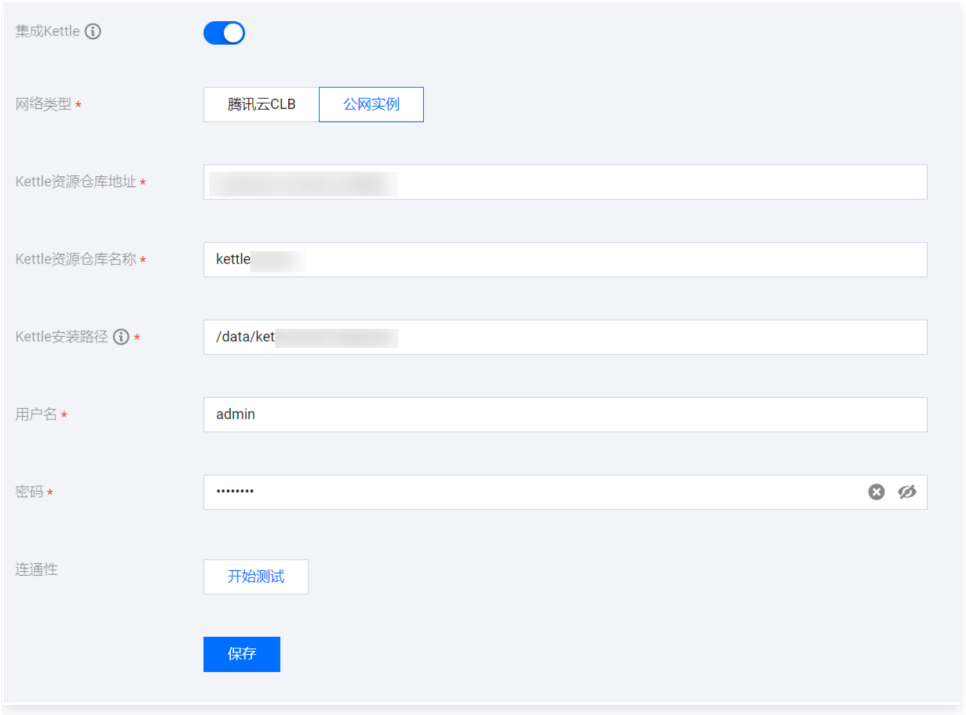
- Task Configuration Priority: Once selected, during the next instance run, the project's configured application parameters will automatically be added in front of the task application parameters.
- Project Configuration Priority: Once selected, during the next instance run, the project's configured application parameters will automatically be appended to the task application parameters.

Advanced Configuration

Integrate Kettle: Once enabled, a **Kettle** task node will be added in the orchestration space, allowing Kettle resource files to be selected and Kettle tasks to be executed within WeData's data development.



When enabled, you need to configure the Kettle resource repository address. You can use Tencent Cloud CLB or public network instances to ensure that WeData can access the Kettle resource repository through the configured VPC.



Approval Configuration

Supports configuring approval processes for data development tasks, database tables, and Data Service API within projects. The approval scenarios are task submission scheduling, database table permissions application, and Data Service API permissions application.

©2013–2025 Tencent Cloud. All rights reserved.



Property Name	Attribute Item Description
Approval Object	Supported approval objects include tasks, database tables, and APIs
Approval Scenes	When the approval object is a task, the approval scenario is data development task submission scheduling/task submission backfill plan, requiring approval from the approver When the approval object is a database table, the approval scene is applying for related permissions of the database and data table When the approval object is an API, the approval scene is applying for related permissions of the Data Service API
Enable Approval	Once enabled, the approval process will follow the approval conditions, approvers, and approval notifications set below
Approval Conditions	When the approval object is a task, approval conditions can be configured. • Task Priority: Supports specifying a task priority (high, medium, low) for approval • Workflow: Supports specifying one or more workflows for approval • Task Directory: Supports specifying one or more task directories for approval
Approver	Supports selecting the task owner, project admin, or other project members as approvers
Approval Notifications	Once enabled, approval notifications will be sent according to the notification methods below
Notification Methods	Message Notification is selected by default. Supports choosing Email, SMS, Telephone, WeCom, Enterprise WeChat group, FeiShu group for notifications

©2013–2025 Tencent Cloud. All rights reserved.

Member and Role Management

Last updated：2024-08-24 18:13:37

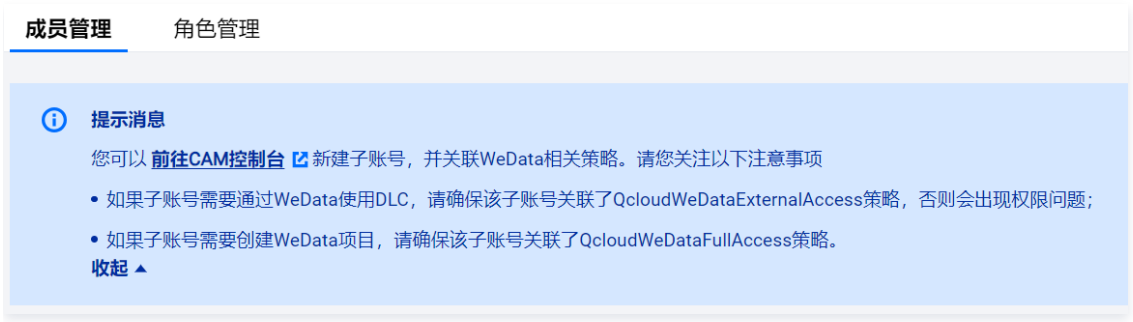
Member management

To add members to the project, click **Member and Role Management** in the left directory tree. Under the **Member Management** tab, you can view and manage member information for this project.



Add User

If there are no available members to add or the target member cannot be found, you can click the prompt link to go to the CAM console to create a new member. For detailed steps, refer to CAM How to Create Sub-user.



Project Member List

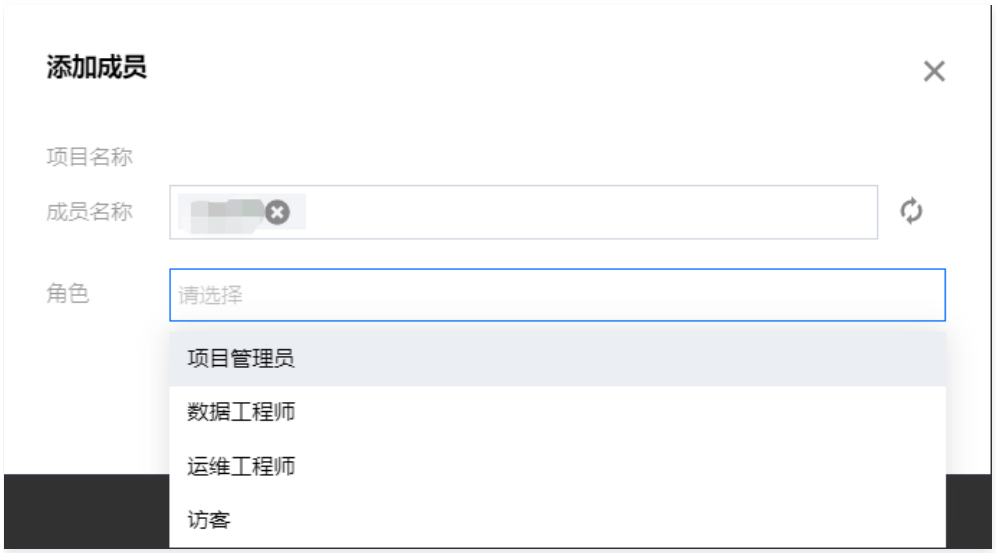
The project member list contains the names of all added project members, cloud account IDs, member roles, addition times, and operation information. For project members who are neither the owner nor the system administrator, you can edit or delete them in the operation column.



Add Member

©2013-2025 Tencent Cloud. All rights reserved.

On the page, click **Add** to add a user member to the current project. You need to select a user and configure role permissions for the user.



By default, the table owner and project administrator have data query and data change permissions for tables under this project. Other project members (such as data engineers, operation and maintenance engineers, etc.) need to be granted permissions in this module. Detailed role permission information can be viewed through Role Management.

Edit

1. Select the member and click **Edit/Batch Modify** in the operation column.
2. In the pop-up window, you can re-edit the roles corresponding to the selected members and click **OK**.



Delete

Select the member and click **Delete/Batch Delete** in the operation column.



Confirm member information and click **OK**. After deleting the member, the member will be removed from the member list.

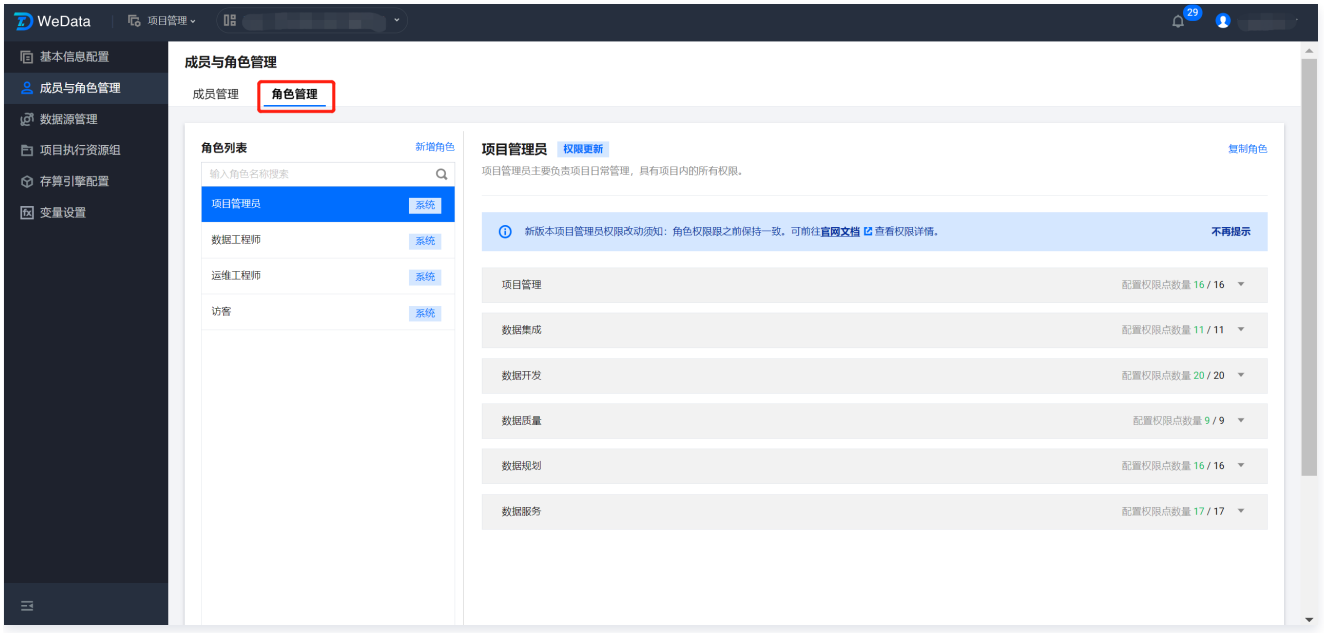
©2013-2025 Tencent Cloud. All rights reserved.

Note:
Deleting members will cause task failures, and tasks must be handed over.
1. **Change the person in charge in batch operations within the orchestration space.**
2. **Use advanced filtering to change the person in charge in task operations and maintenance.**
3. **If a member is in the personal area of the development space, export the tasks.**

Role management

Note
Different roles have different project feature operation permissions. WeData presets role permissions for project administrators, data engineers, operations engineers, and visitors. It also supports custom roles and custom role permissions.

Click **Member and Role Management** in the left directory tree. In the **Role Management** tab, you can view and manage roles and role permissions for this project.



Role List

The role list displays the existing roles in the current project, including preset roles and custom roles.
- Preset roles and their corresponding permissions cannot be edited or deleted.
- Custom roles and their corresponding permissions can be edited and deleted.

Selecting different roles allows you to view the specific permission information for that role under various feature modules, including read, write, and delete.

©2013–2025 Tencent Cloud. All rights reserved.

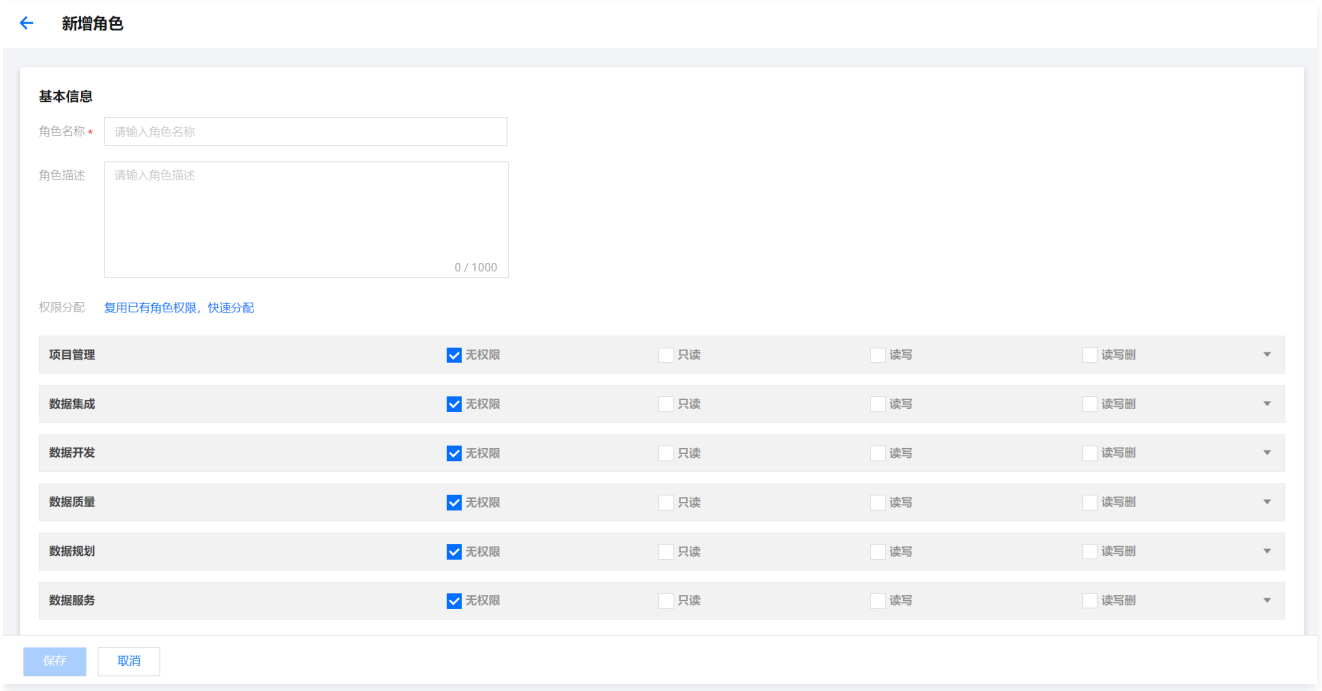


Note

" √ " indicates the default permissions for that type of user.

Add Role

On the page, click **Add Role**. Enter the basic information of the role and allocate permissions on the configuration page. Permissions should include project management, DataInLong, data development, data quality, data planning, and data service.



During the permission configuration process for the new role, you can quickly allocate permissions by reusing the permission configurations of existing roles. After reusing, only partial changes are needed to flexibly design role permissions that meet business needs.

©2013–2025 Tencent Cloud. All rights reserved.



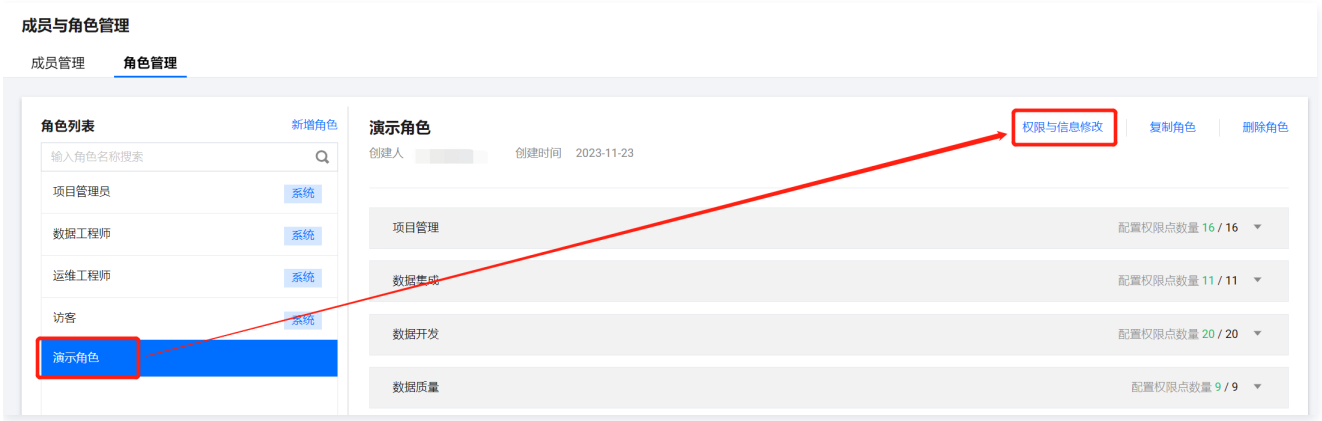
Copy Role

In the role list, select any role, click the upper right corner **Copy Role**, enter the name and description of the new copied role in the pop-up box to quickly create a new custom-defined role. This new role will have the same permissions configuration as the original role. This feature is generally used for quickly generating custom-defined roles to facilitate permission configuration.



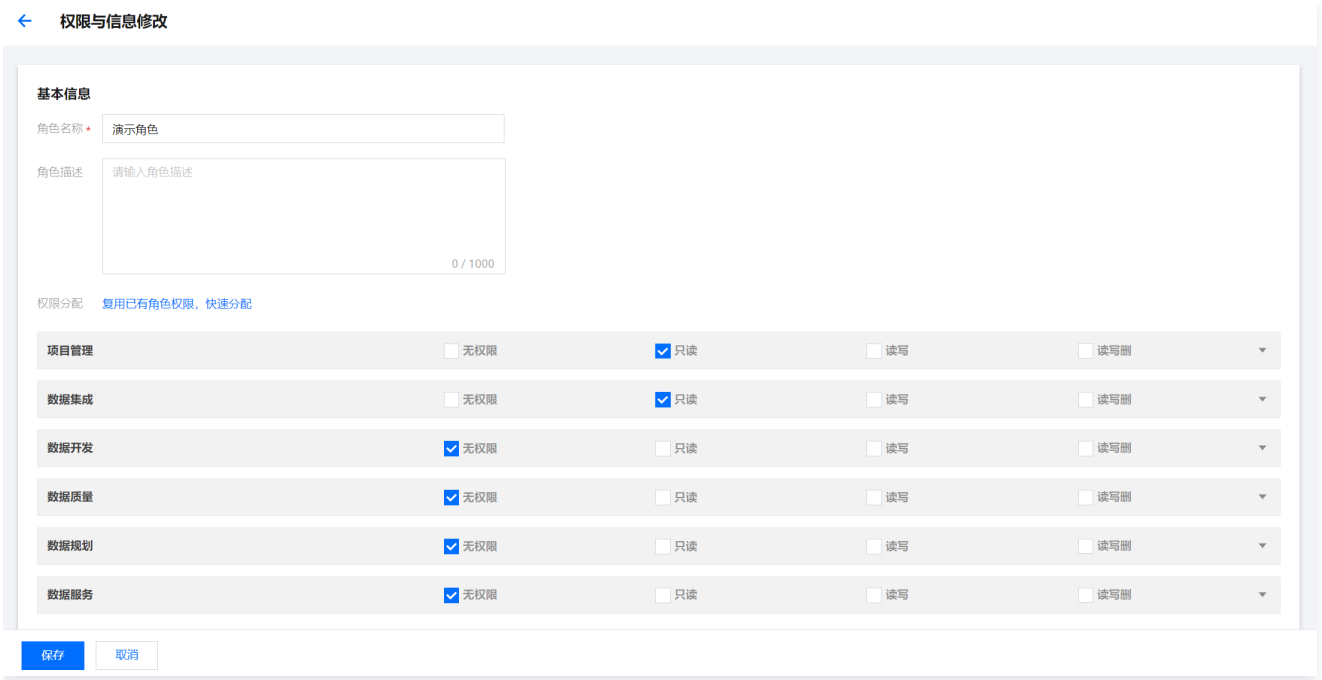
Edit

1. Select the custom-defined role member, click the upper right corner **Modify Permissions and Information**.



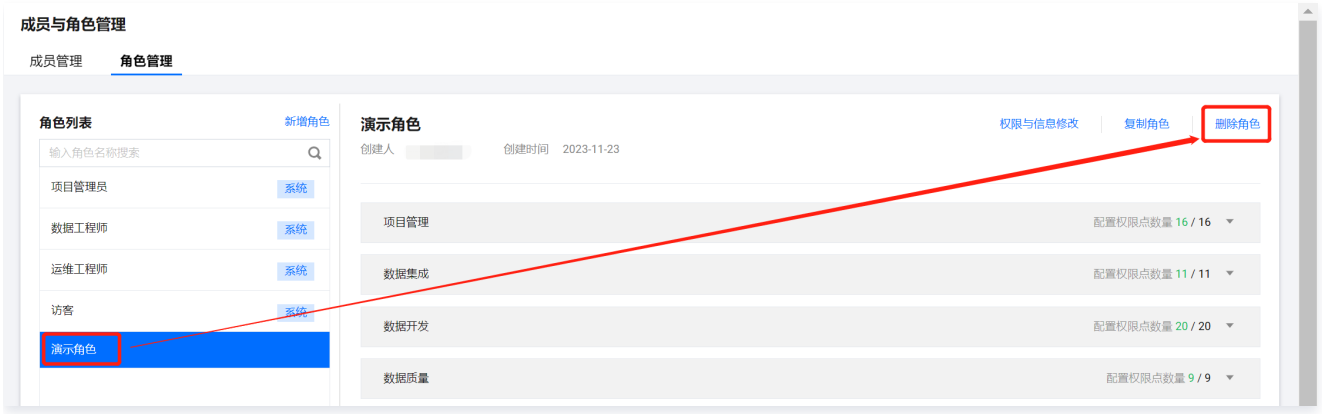
2. In the modification page, you can change the role name, description, and feature permissions. After making changes, click **Save** to complete the role editing.

©2013-2025 Tencent Cloud. All rights reserved.



Delete

1. Select the custom-defined role member, click the upper right corner **Delete Role**.



2. In the pop-up box, click **OK** to delete the corresponding custom-defined role.



©2013-2025 Tencent Cloud. All rights reserved.

Data Source Management

Last updated：2024-11-21 21:28:49

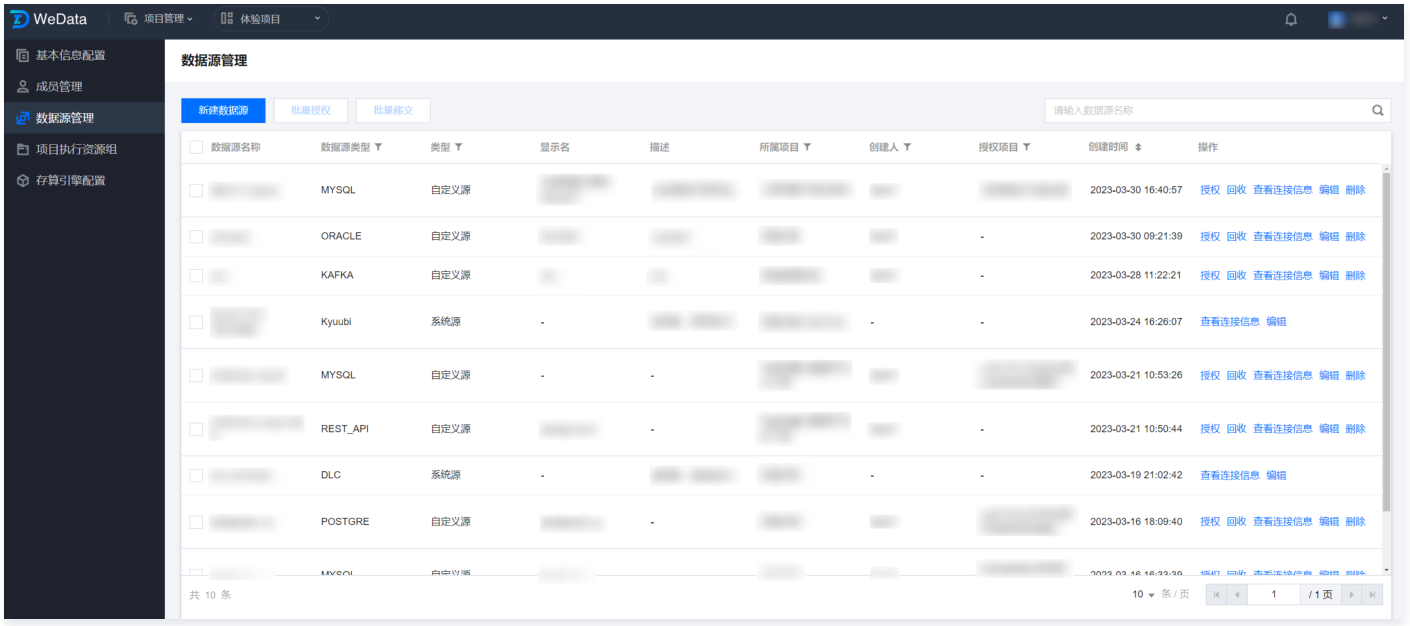
Supported Data Source Types

In the current version, Tencent Cloud WeData supports more than 40 types of data sources, including:

- **Relational Database**: TDSQL-C MySQL, TDSQL MySQL, TDSQL PostgreSQL, Tencent Cloud MySQL, MySQL, TCHouse-P, PostgreSQL, Oracle, SQL Server, IBM DB2, Dameng DM, SAP HANA, Sybase, TiDB;
- **Big Data**: HDFS, HBase, Hive, Kudu, Greenplum, TCHouse-C, Clickhouse, GaussDB, TCHouse-D, Doris, DLC, Impala, Gbase, Tbase, Iceberg, StarRocks, Trino, Kyuubi, TCHouse-X;
- **Semi-structured Data**: FTP, SFTP, COS, REST API;
- **NoSQL Data**: Redis, Tencent Cloud ElasticSearch, ElasticSearch, Tencent Cloud MongoDB, MongoDB, Graph Database;
- **Message Queue**: CKafka, DTS-Kafka, Kafka.

Data Source List

Log in to the WeData console, click the left menu **Data Source Management**, on the data source management page, you can view and manage the data source information under this project.



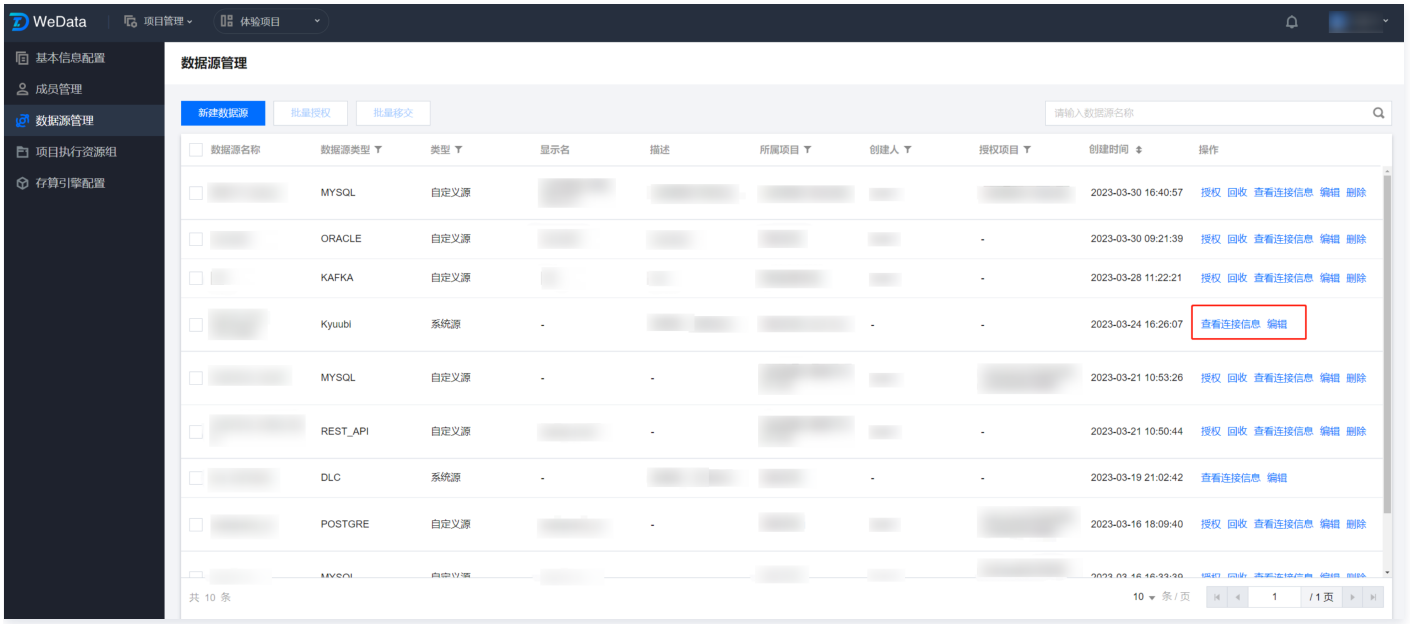
The project includes two types of data sources: system sources and custom sources. System sources are automatically generated after binding a computational storage engine; custom sources are created by the user.

System Source

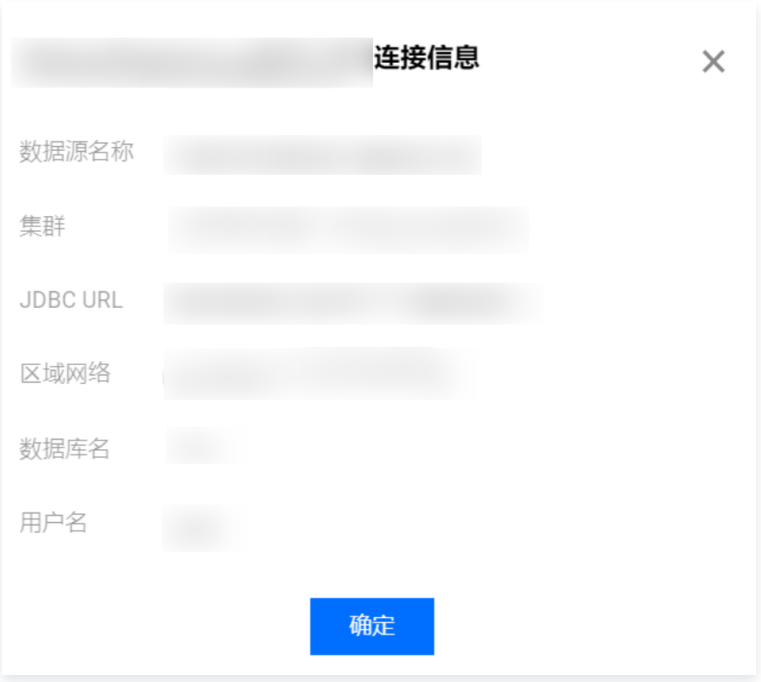
System sources are data sources automatically created when the user binds a computational engine in the project. Currently, system sources only support viewing and editing data source connection information, and do not support deletion, authorization, and recovery operations.

- Binding the EMR engine will automatically create EMR Hive data sources, EMR HBase data sources, EMR Kyuubi data sources, EMR Impala data sources, and EMR Trino data sources;
- Binding the DLC engine will automatically create DLC data sources;
- Binding the TCHouse-P engine will automatically create TCHouse-P data sources.

©2013-2025 Tencent Cloud. All rights reserved.



- View Connection Information: Select the row where the data source is located and click **View Connection Information.**



- Editing: Select the row of the data source and click **Editing.** Only the display name can be edited by the root account.

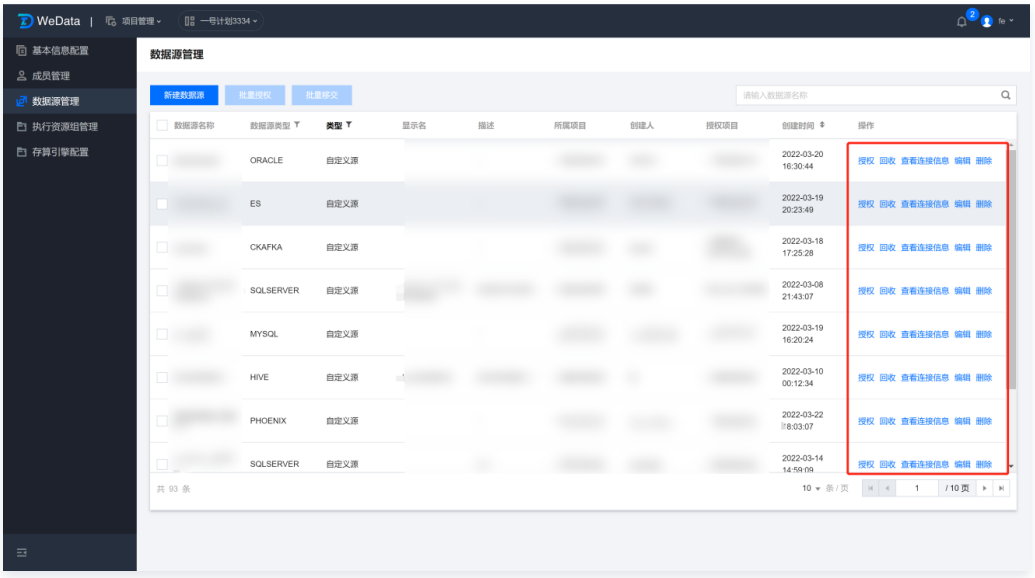
©2013–2025 Tencent Cloud. All rights reserved.



Custom Definition Source

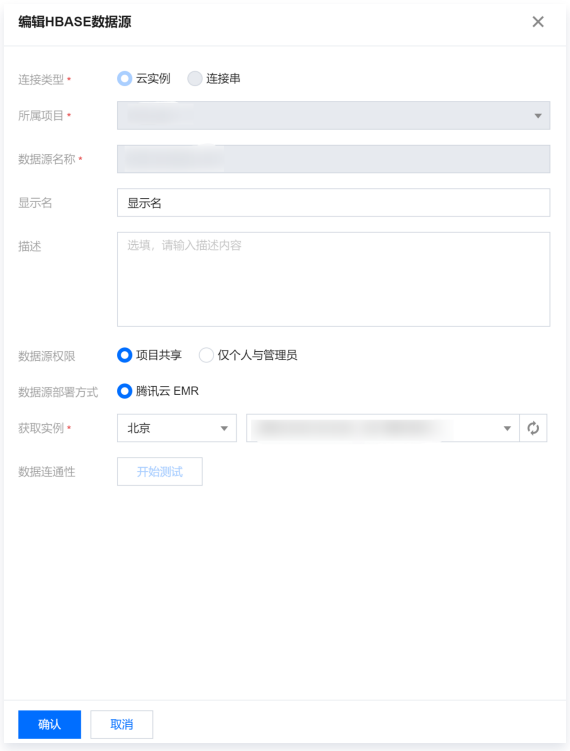
The Custom Definition Source is the data source introduced by the user in WeData. Currently, more than 20 types of data sources such as MySQL, Kafka, and HDFS are supported. The Custom Definition Source supports viewing connection information, editing, deleting, authorization, and recycling operations.

- View Connection Information: Select the row where the data source is located and click **View Connection Information**.



©2013–2025 Tencent Cloud. All rights reserved.

- Edit: Select the row where the data source is located and click **Edit**.
  Except for connection type, data source name, and belonging project, other configuration items are editable.
  - Cloud Instance: For data sources introduced via cloud instance, the data source description, username, and password can be modified.



  - Connection String: For data sources introduced via connection string, the description, JDBC URL parameters, username, and password can be modified.
  - After editing the data source, click **Start Test** to test the connectivity of the edited data source, and click **OK** to save the edited content.
- Delete
  - Select the data source and click **Delete** in the operation column.
  - Confirm the data source information and click **Delete**.
- Authorization/Batch Authorization
  - Select the data source and click **Authorize** in the operation column.
  - Select **Authorization Object**, either user or project.
  - Select the target user or project, click **Confirm**.

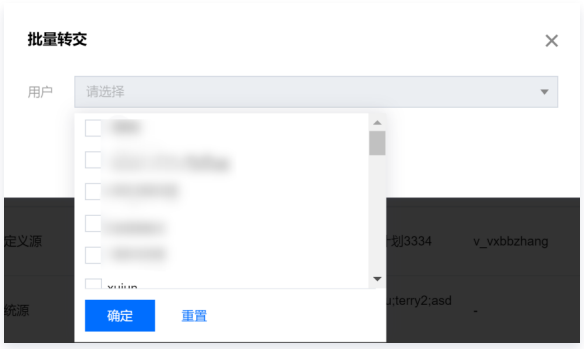


- Recycle
  - Select the data source, and click **Recycle** in the operation column.
  - Select the target user or project for which you need to revoke permissions, click **Recycle** and then **Confirm** in the pop-up box.

©2013–2025 Tencent Cloud. All rights reserved.



- Batch transfer
  - Select the data source, and click **Batch Transfer** in the top-left corner.
  - Select the target user to whom you need to transfer permissions, click **Confirm**. After the transfer, the owner of the data source will be the transferee.

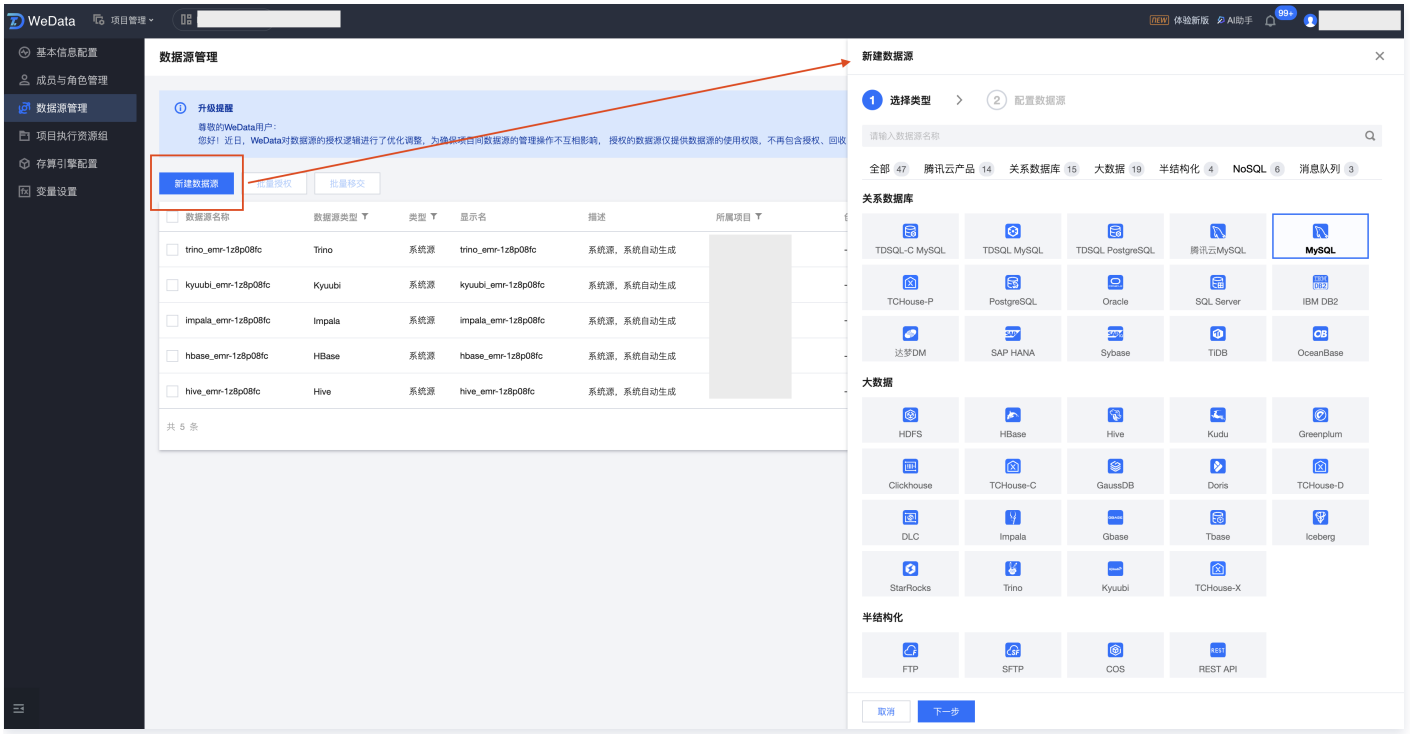


Create Relational Database Data Source

Tencent Cloud WeData currently supports the creation of over 10 types of relational databases. Tencent Cloud relational database products (such as TDSQL-C MySQL, TDSQL MySQL, TDSQL PostgreSQL, Tencent Cloud MySQL, etc.) support connection through cloud instances, while general relational database data sources (such as MySQL, PostgreSQL, Oracle, SQL Server, etc.) support connection through connection strings.

Select data source type

Click **New Data Source**, select the data source type, and then click **Next**.



©2013-2025 Tencent Cloud. All rights reserved.

Configure Data Source: TDSQL-C MySQL

Taking Tencent Cloud relational database product TDSQL-C MySQL as an example, the same applies to other data sources. TDSQL-C MySQL data source supports both cloud instance and connection string connection methods.

- Create a data source using the cloud instance method.



Parameter	Description
Connection Type	Select the data source connection method using cloud instance or connection string
Associated item	The item to which the data source belongs when created
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Display name	The name displayed when the data source is used in the product. If not filled, the default is the data source name
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
Get instance	Select the region, instance name, and ID information of the TencentDB instance under the account
Database name	Database name to connect to
Username	Username to connect to the database

©2013-2025 Tencent Cloud. All rights reserved.

Password	Password to connect to the database
Data connectivity	Test whether the configured database is connected **Note** • If the connectivity test fails, the data source can still be saved. However, it cannot be used if the connectivity test fails and the data source is saved. • If the connectivity test fails, it might be because WeData is blocked by the firewall of the network where the database is located. Please refer to Add Tencent Cloud MySQL Database Security Group.

- Create a data source via connection string.



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Display name	The name displayed when the data source is used in the product. If not filled, the default is the data source name
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
Deployment mode	Supports two deployment modes: self-built instances and public network instances. Self-built instances are data source instances deployed on Tencent CVM, while public network instances are those in the

©2013-2025 Tencent Cloud. All rights reserved.

	customer's local IDC or other cloud resources, accessible via the public network.
Region and Network	When choosing self-built instances, you need to select the region and vpcID where the data source instance is located
JDBC URL	Connection string information for connecting to TDSQL-C MySQL data source instances, including host IP, port, database name, etc
Database name	Database name to connect to
Username	Username to connect to the database
Password	Password to connect to the database
Data connectivity	Test whether the configured database is connected. **Note** • If the connectivity test fails, the data source can still be saved. However, it cannot be used if the connectivity test fails and the data source is saved. • If the connectivity test fails, it might be because WeData is blocked by the firewall of the network where the database is located. Please refer to Add Tencent Cloud MySQL Database Security Group.

Configure Data Source: PostgreSQL

Generic relational database data source, explained with PostgreSQL as an example. The same applies to other data sources. PostgreSQL data source only supports connection string as the connection method.

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Display name	Optional display name. If not filled, the default is the data source name
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
Deployment mode	Supports two deployment methods: self-built instances and public network instances. Self-built instances are data source instances deployed on Tencent CVM, while public network instances are those in a customer's local IDC or other cloud resources, accessible via the public network.
Region and Network	When selecting self-built instances, the region and vpcID of the data source must be selected
JDBC URL	Connection string information for connecting to PostgreSQL data source instances, including host IP, port, database name, etc
Username	Username to connect to the database
Password	Password to connect to the database

©2013-2025 Tencent Cloud. All rights reserved.

Data connectivity	Test whether the configured database is connected

Configure data source: IBM DB2

DB2 data sources only support connection string as the connection method.



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Display name	Optional display name. If not filled, the default is the data source name
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
Region and Network	Select the region, instance name, and ID information of the TencentDB instance under the account.
DB2 model	Select the corresponding DB2 model. The current version supports DB2 iSeries/AS 400, DB2 for LUW, and DB2 z/OS models
IP and port number	Enter the IP and port number information of the DB2 data source respectively
Username	Username to connect to the database
Password	Password to connect to the database
Data connectivity	Test whether the configured database is connected

Configure data source: Sybase

Sybase data sources only support connection string as the connection method.

©2013–2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Display name	Optional display name. If not filled, the default is the data source name
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
Deployment mode	Select from self-built instances or public network instances
Region and Network	Select the region, instance name, and ID information of the TencentDB instance under the account.
IP and port number	Enter the IP and port number information of the Sybase data source separately
Database name	Database name to connect to
Username	Username to connect to the database
Password	Password to connect to the database
Data connectivity	Test whether the configured database is connected

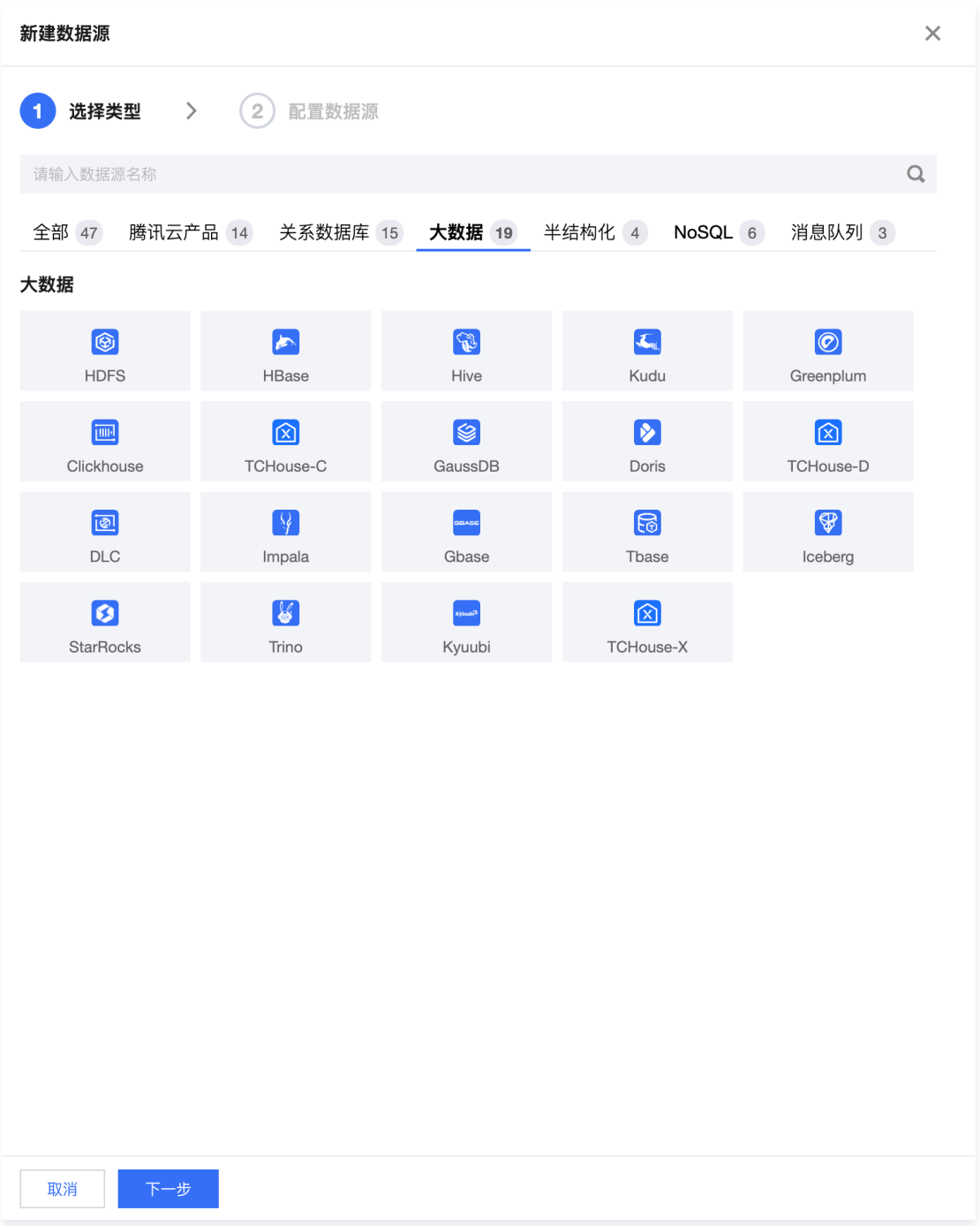
Create Big Data Data Source

The current version of Tencent Cloud WeData supports creating more than 10 types of Big Data sources, among which TCHouse-C and TCHouse-D are Tencent Cloud Big Data products.

Select data source type

©2013-2025 Tencent Cloud. All rights reserved.

Click **New Data Source**, select the data source type, and then click **Next**.



Configuring a Data Source.

Configuring a Data Source: HDFS

WeData currently supports importing HDFS data sources through both Tencent Cloud EMR cloud instances and connection strings.

©2013–2025 Tencent Cloud. All rights reserved.

- Importing Data Source via Cloud Instance.



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Display name	Optional display name. If not filled, the default is the data source name
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
Deployment mode	Only supports Tencent Cloud EMR
Get instance	Select the region, instance name, and ID information of the EMR instance under the account
Data connectivity	Test whether the configured database can be connected (if the connectivity test fails, an error message will be provided for troubleshooting. The data source can still be saved, but an exception will occur when using this data source)

- Importing Data Source via Connection String

©2013–2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Description	Optional, description of this data source
Deployment mode	Currently, it only supports adding HDFS from Tencent Cloud EMR
Region and Network	First select the region, then choose the EMR instance under that region. Click the refresh button on the right to retrieve instances in that region again
defaultFS	Used to specify the default file system and defaults to the local file system
Extended configuration	Optional, optional extensions related to the HDFS connection string

©2013–2025 Tencent Cloud. All rights reserved.

core-site.xml file	HDFS core-site.xml file
hdfs-site.xml file	HDFS hdfs-site.xml file
Authentication Method	Whether HDFS connection has Kerberos authentication
keytab file	Keytab file required for Kerberos authentication
conf file	Conf file required for Kerberos authentication
principal	Principal required for Kerberos authentication
Data connectivity	Test whether the configured database can be connected (if the connectivity test fails, an error message will be provided for troubleshooting. The data source can still be saved, but an exception will occur when using this data source)

Configuring a Data Source: HBase

WeData currently supports importing HBase data sources through both Tencent Cloud EMR cloud instances and connection strings.

- Importing Data Source via Cloud Instance



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Display name	Optional display name. If not filled, the default is the data source name
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it

©2013-2025 Tencent Cloud. All rights reserved.

Deployment mode	The current version only supports Tencent Cloud EMR
Get instance	Select the region, instance name, and ID information of the EMR instance under the account
Data connectivity	Test whether the configured database can be connected (if the connectivity test fails, an error message will be provided for troubleshooting. The data source can still be saved, but an exception will occur when using this data source)

- Connecting string method to import data source



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Description	Optional, description of this data source
Data source deployment mode	Currently, it only supports adding HBase from Tencent Cloud EMR
Get instance	First select the region, then choose the EMR instance under that region. Click the refresh button on the right to retrieve instances in that region again
Data connectivity	Test whether the configured database can be connected (if the connectivity test fails, an error message will be provided for troubleshooting. The data source can still be saved, but an exception will occur when using this data source)

©2013–2025 Tencent Cloud. All rights reserved.

Configuring a Data Source: Hive

WeData currently supports importing Hive data sources through connection strings.







Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Description	Optional, description of this data source

©2013–2025 Tencent Cloud. All rights reserved.

Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
JDBC URL	Connection string information for connecting to the Hive data source
Username	Username for connecting to the data source
Password	Password for connecting to the data source
Metadata collection	After selection, the current data source will automatically create a metadata collection task in the data discovery module
Data connectivity	Test whether the configured database is connected

Configuring a Data Source: Kudu

WeData currently supports importing Kudu data sources through both Tencent Cloud EMR and connection strings.

- Importing Data Source via Cloud Instance.



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Display name	Optional display name. If not filled, the default is the data source name
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
Deployment mode	The current version only supports Tencent Cloud EMR

©2013–2025 Tencent Cloud. All rights reserved.

Get instance	Select the region, instance name, and ID information of the EMR instance under the account.
Data connectivity	Test whether the configured database can be connected (if the connectivity test fails, an error message will be provided for troubleshooting. The data source can still be saved, but an exception will occur when using this data source)

- Importing Data Source via Connection String.



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
Kudu Server List	Address information for connecting to Kudu servers
Username	Username for connecting to the data source
Password	Password for connecting to the data source
Data connectivity	Test whether the configured database is connected

Configuring Data Source: Doris

WeData currently supports importing Doris DB and StarRocks data sources via connection strings.

©2013–2025 Tencent Cloud. All rights reserved.



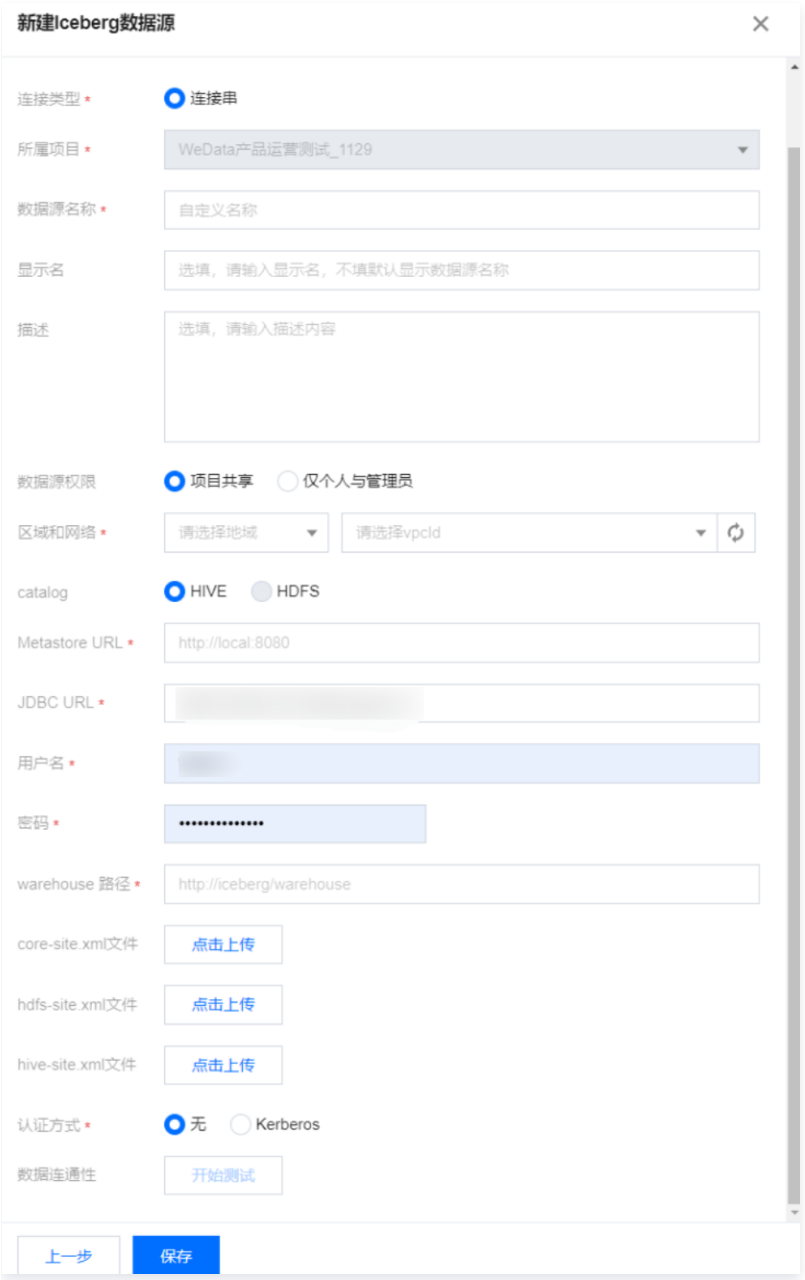
Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
Deployment mode	You can choose self-built instances or public instances. For self-built instances, you need to enter the region and network. Public instances do not require region and network information
Region and Network	Select the region, instance name, and ID information of the TencentDB instance under the account.
JDBC URL	Connection string information for connecting to the Hive data source
FE URL	Enter the fe HTTP address in the format: IP address:http port. Separate multiple addresses with commas (,), e.g., 172.17.16.3:8030,172.17.16.4:8030
Username	Username for connecting to the data source
Password	Password for connecting to the data source

©2013-2025 Tencent Cloud. All rights reserved.

Data connectivity	Test whether the configured database is connected

Configuring Data Source: Iceberg

WeData currently supports importing Iceberg data sources via connection strings.



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
Region and Network	Select the region, instance name, and ID information of the TencentDB instance under the account.

©2013–2025 Tencent Cloud. All rights reserved.

catalog	You can choose HIVE or HDFS
Metastore URL	Please enter the Metastore URL
JDBC URL	Connection string information for connecting to the Hive data source
Username	Username for connecting to the data source
Password	Password for connecting to the data source
Warehouse path	Please enter the warehouse path
core-site.xml	Supports uploading core-site.xml
hdfs-site.xml	Supports uploading hdfs-site.xml
hive-site.xml	Supports uploading hive-site.xml
Authentication Method	Kerberos authentication is optional
Data connectivity	Test whether the configured database is connected

Configuring a Data Source: DLC

WeData currently supports connecting to DLC data sources via connection strings.



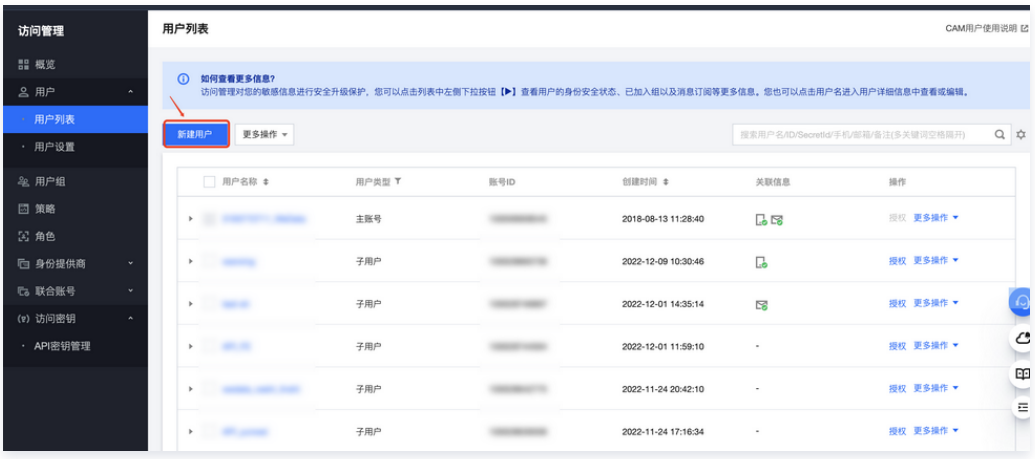
Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters

©2013-2025 Tencent Cloud. All rights reserved.

Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
JDBC URL	Connection string information for connecting to the DLC data source
Username	Username for connecting to the data source
SecretId	Username for connecting to the data source
SecretKey	Password for connecting to the data source
Data connectivity	Test whether the configured database is connected

DLC Permission Scope Operation Guide

1. In the **CAM** > **Users** > **User List** page, click **Create User** to start creating a new user (i.e., sub-account).



2. Create a user by following the steps on the Create User page: fill in the username, select the Tag key and Tag value, and complete the creation of the new user.

©2013-2025 Tencent Cloud. All rights reserved.



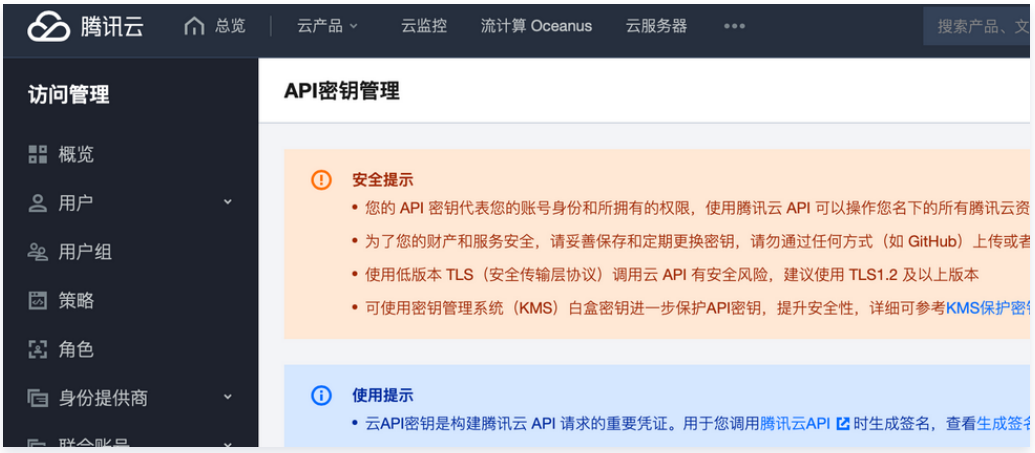
3. In the **CAM** > **Policies** page, set the corresponding DLC data policies for the specified role or sub-account to grant permissions.



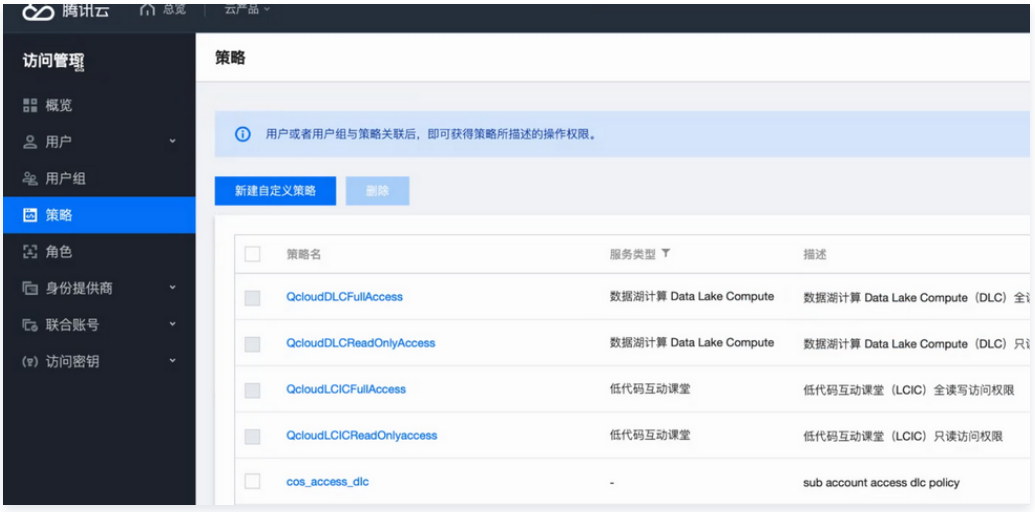


4. Apply for a key: If there are no sub-accounts, guide and redirect to the corresponding API key page for creating keys for the sub-account as shown below:

©2013-2025 Tencent Cloud. All rights reserved.



5. DLC platform CAM permissions add DLC policy permissions. For details, see CAM CAM Service.



Configuring a Data Source: Clickhouse, Kyuubi, TChouse-X

WeData currently supports importing Clickhouse, Kyuubi, and TChouse-X data sources via cloud instance and connection strings.

©2013-2025 Tencent Cloud. All rights reserved.

- Importing Data Source via Cloud Instance.



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Display name	Optional display name. If not filled, the default is the data source name
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
Deployment mode	The current version only supports Tencent Cloud EMR
Get instance	Select the region, instance name, and ID information of the TencentDB instance under the account
Database name	Target database name
Username	Username for connecting to the data source
Password	Password for connecting to the data source
Data connectivity	Test whether the configured database can be connected (if the connectivity test fails, an error message will be provided for troubleshooting. The data source can still be saved, but an exception will occur when using this data source)

- Importing data sources via connection string.

©2013–2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
JDBC URL	Connection string information for connecting to the Hive data source
Username	Username for connecting to the data source
Password	Password for connecting to the data source
Data connectivity	Test whether the configured database is connected

Configuring a Data Source: Trino

Tencent Cloud WeData currently supports importing Trino data sources via connection strings.

©2013–2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data source name	The name of the new data source, defined by the user, cannot be empty. Naming must start with a letter, can contain letters, numbers, and underscores. Length within 20 characters.
Display name	Optional display name, if not filled, the data source name will be displayed.
Description	Optional, description of this data source.
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it.
Deployment mode	Supports both self-built instance and public network instance deployment modes
Region and Network	Select the region, instance name, and ID information of the TencentDB instance under the account.
IP and port number	Optional, enter the IP address and port number of the current data source
catalog	Selectable catalog
Schema	Please enter sybase
Username	Username for connecting to the data source
Password	Password for connecting to the data source

©2013-2025 Tencent Cloud. All rights reserved.

Data connectivity	Test whether the configured database is connected

Configuring a Data Source: Greenplum, GauseDB, Impala, Tbase, Gbase

For data source types: Greenplum, GauseDB, Impala, Tbase, Gbase, Tencent Cloud WeData currently supports access via connection strings.



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
JDBC URL	Connection string information for connecting to the Hive data source
Username	Username for connecting to the data source
Password	Password for connecting to the data source
Data connectivity	Test whether the configured database is connected

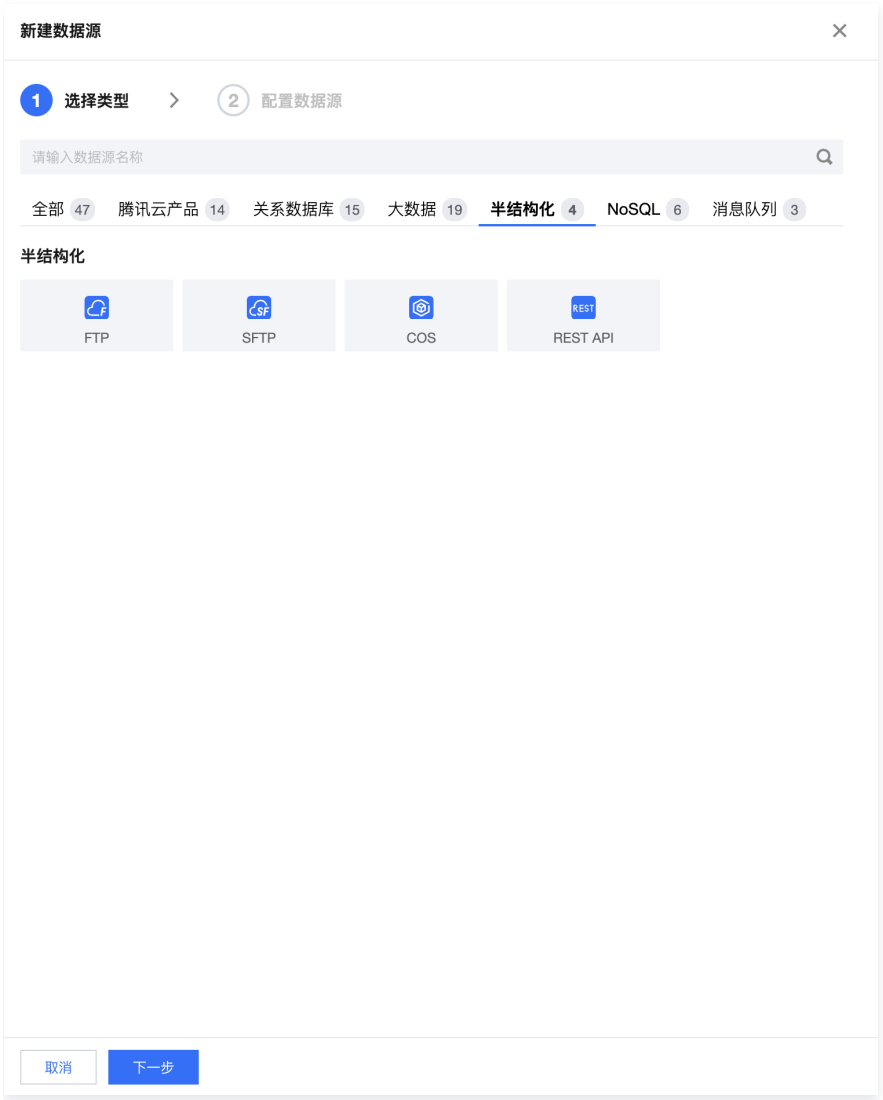
Create a semi-structured data source

In the current version, Tencent Cloud WeData supports four types of semi-structured data sources: FTP, SFTP, COS, and REST API, with COS being a Tencent Cloud product.

Select data source type

Click **New Data Source**, select the data source type, and then click **Next**.

©2013-2025 Tencent Cloud. All rights reserved.



Configuring a Data Source: FTP

©2013–2025 Tencent Cloud. All rights reserved.

Tencent Cloud WeData currently supports importing FTP data sources via connection strings.



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
Deployment mode	You can choose self-built instances or public instances. For self-built instances, you need to enter the region and network. Public instances do not require region and network information
Region and Network	Choose the region and the corresponding vpcid in the current region
IP and port number	Enter the IP and port number of the FTP data source
Username	Username to connect to the data source
Password	Password for connecting to the data source
Data connectivity	Test whether the configured database is connected

Configuring a Data Source: SFTP

©2013-2025 Tencent Cloud. All rights reserved.

Tencent Cloud WeData currently supports importing SFTP data sources via connection strings.



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
Deployment mode	You can choose self-built instances or public instances. For self-built instances, you need to enter the region and network. Public instances do not require region and network information
Region and Network	Choose the region and the corresponding vpcid in the current region
IP and port number	Enter the IP and port number of the SFTP data source
Username	Username to connect to the data source
Password	Password for connecting to the data source
Data connectivity	Test whether the configured database is connected

Configuring a Data Source: COS

©2013-2025 Tencent Cloud. All rights reserved.

Tencent Cloud WeData currently supports importing COS data sources via cloud instances.



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
Deployment mode	You can choose self-built instances or public instances. For self-built instances, you need to enter the region and network. Public instances do not require region and network information
Region and Network	Choose the region and the corresponding vpcid in the current region
secertId	Username for connecting to the data source
secertKey	Password for connecting to the data source
Bucket	Optional, add the access path of the current data source
Data connectivity	Test whether the configured database can be connected (if the connectivity test fails, an error message will be provided for troubleshooting. The data source can still be saved, but an exception will occur when using this data source)

Configuring a Data Source: REST API

Tencent Cloud WeData currently supports importing REST API data sources via connection strings.

©2013–2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
Deployment mode	You can choose self-built instances or public instances. For self-built instances, you need to enter the region and network. Public instances do not require region and network information
Region and Network	Choose the region and the corresponding vpcid in the current region
URL	Connection string information for connecting to the data source
Default Request Headers	Optional Request Headers

©2013-2025 Tencent Cloud. All rights reserved.

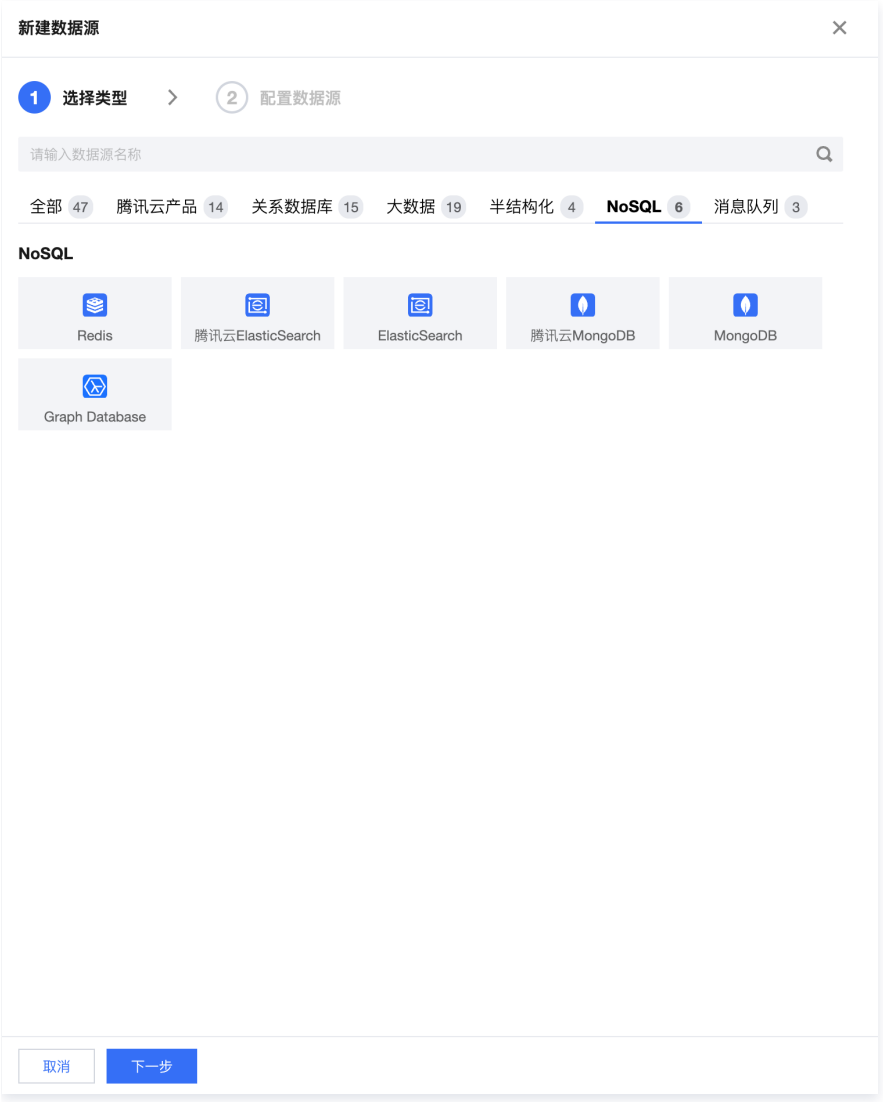
Authentication Method	Authentication methods can be selected: Basic, Token
Data connectivity	Test whether the configured database can be connected (if the connectivity test fails, an error message will be provided for troubleshooting. The data source can still be saved, but an exception will occur when using this data source)

Create NoSQL Data Source

In the current version, Tencent Cloud WeData supports NoSQL data sources such as Redis, Tencent Cloud ElasticSearch, ElasticSearch, Tencent Cloud MongoDB, MongoDB, and Graph Database. Among them, Tencent Cloud ElasticSearch and Tencent Cloud MongoDB are Tencent Cloud NoSQL database products.

Select data source type

Click **New Data Source**, select the data source type, and then click **Next**.



Configuring a Data Source: Tencent Cloud ElasticSearch

Tencent Cloud ElasticSearch and Tencent Cloud MongoDB data sources can be connected through cloud instances. Taking Tencent Cloud ElasticSearch as an example, the same applies to Tencent Cloud MongoDB data sources.

©2013–2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Display name	Optional display name. If not filled, the default is the data source name
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
Get instance	Select the region, instance name, and ID information of the Tencent Cloud ElasticSearch instance under the account
Security Authentication	If you choose Anonymous Login, there's no need to provide username and password. If you choose Username and Password Login, you must provide the username and password for security authentication
Username	Username for connecting to the data source
Password	Password for connecting to the data source
Data connectivity	Test whether the configured database can be connected (if the connectivity test fails, an error message will be provided for troubleshooting. The data source can still be saved, but an exception will occur when using this data source)

Configuring a Data Source: Redis

Redis, ElasticSearch, MongoDB, and Graph Database data sources are currently supported only via connection strings. Taking Redis as an example, the same applies to other data sources.

©2013–2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
Deployment mode	Supports two deployment methods: self-built instances and public network instances. Self-built instances are data source instances deployed on Tencent CVM, while public network instances are those in a customer's local IDC or other cloud resources, accessible via the public network.
Region and Network	When selecting self-built instances, the region and vpcID of the data source must be selected
IP and port number	Optional, enter the IP address and port number of the current data source
Account Types	• If there is no username, it can be left blank

©2013-2025 Tencent Cloud. All rights reserved.

	• Tencent Cloud Redis User-Defined Account, please fill in the account • For Tencent Cloud Redis instances purchased before January 2018, please fill in the instance ID
Password	Password for connecting to the data source
Data connectivity	Test whether the configured database is connected

Create message queue data source

The current version of Tencent Cloud WeData supports the following message queue type data sources: Tencent Cloud CKafka, DTS-Kafka, and Kafka. Among them, CKafka and DTS-Kafka are Tencent Cloud message queue products.

Select data source type

Click **New Data Source**, select the data source type, and then click **Next**.



Configuring the Data Source: Tencent Cloud CKafka

WeData supports connecting CKafka type data sources via cloud instances.

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Display name	Optional display name. If not filled, the default is the data source name
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
Get instance	Select the region, instance name, and ID information of the CKafka instance under the account
Connection Method	Tencent Cloud CKafka supports two access methods: PLAINTEXT and SASL_PLAINTEXT. For details, please refer to CKafka User Access Control
Data connectivity	Test whether the configured database can be connected (if the connectivity test fails, an error message will be provided for troubleshooting. The data source can still be saved, but an exception will occur when using this data source)

©2013–2025 Tencent Cloud. All rights reserved.

Configuring the Data Source: DTS-Kafka

WeData supports connecting DTS-Kafka type data sources via cloud instances.



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Display name	The name displayed when the data source is used in the product. If not filled, the default is the data source name
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
Cluster address	Cluster address, please go to **DTS Console** > **Subscription Information** > **Private Network Address Copy**
Topic	Name of the Topic to connect to
Consumer group name	Consumer group name
Account/Password	Enter Account/Password
Data connectivity	Test whether the configured database can be connected (if the connectivity test fails, an error message will be provided for troubleshooting. The data source can still be saved, but an exception will occur when using this data source)

Configure Data Source: Kafka

WeData supports connecting Kafka type data sources via connection string.

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Description	Optional, description of this data source
Data source permissions	Select project sharing or use only by individuals and administrators
Deployment mode	Supports two deployment methods: self-built instances and public network instances. Self-built instances are data source instances deployed on Tencent CVM, while public network instances are those in a customer's local IDC or other cloud resources, accessible via the public network.
Region and Network	When selecting self-built instances, the region and vpcID of the data source must be selected
Kafka service list	Please enter the service list, e.g., ip1:9092,ip2:9092
Kafka Authentication Method	Four authentication methods are supported: PLAINTEXT, SASL_PLAINTEXT, SASL_SSL, JAAS • If you choose the SASL_PLAINTEXT authentication method, username and password need to be filled in; • If you choose the SASL_SSL authentication method, TrustStore authentication file, TrustStore password, username, and password need to be filled in; • If you choose the JAAS authentication method, Kafka sasl jaas configuration needs to be filled in.

©2013-2025 Tencent Cloud. All rights reserved.

Data connectivity	Test whether the configured database can be connected (if the connectivity test fails, an error message will be provided for troubleshooting. The data source can still be saved, but an exception will occur when using this data source)

©2013–2025 Tencent Cloud. All rights reserved.

Project Execution Resource Group Scheduling resource

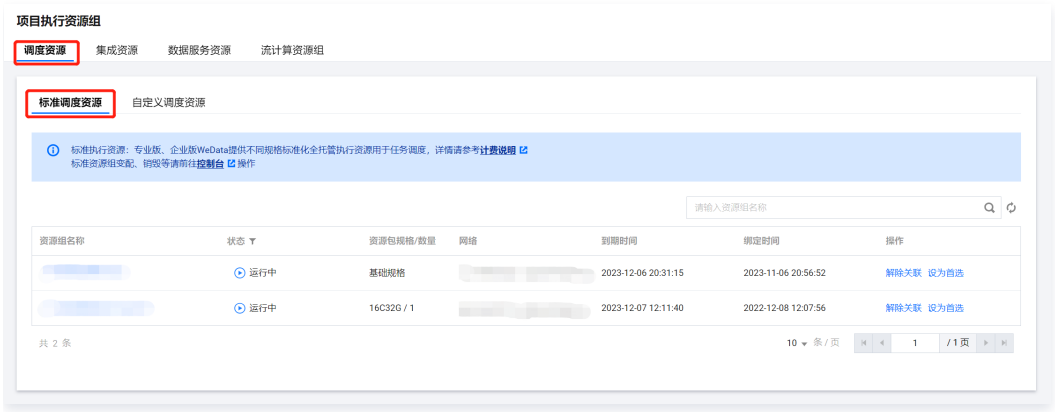
Last updated: 2024-12-12 20:17:04

This page displays information about various resource groups bound to the current project space, including Schedule Resource Group, Integrated Resource Group, Data Service Resource Group, and Stream Computing Resource Group. For detailed information, please refer to Execute Resources.



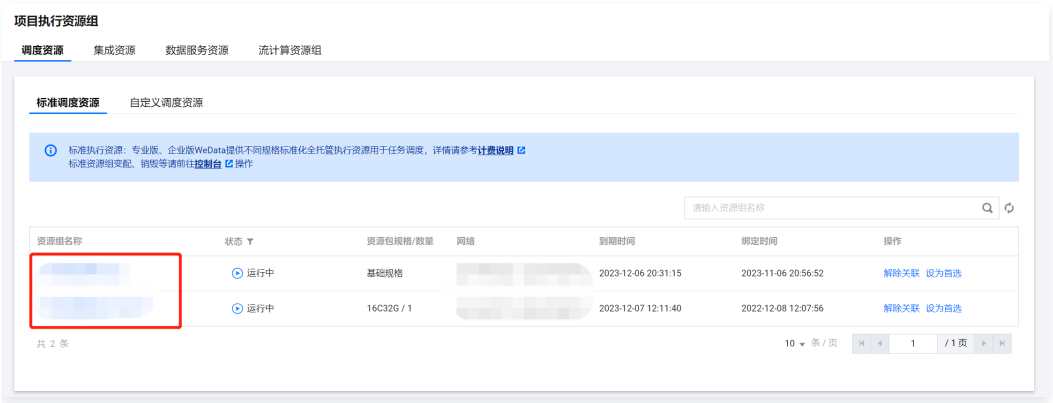
Standard Scheduling Resources

Professional Version and Enterprise Edition WeData provide different specifications of standardized, fully managed execution resources for task scheduling. For details, please refer to Billing Instructions. For changes, decommissioning, and other operations on standard resource groups, please refer to Schedule Resource Group Configuration.



After the project binds the Standard Scheduling Resource Group, click the Resource Group Name to enter the Resource Information page to view resource information, Resource Monitoring, and manage the Resource Group.

©2013-2025 Tencent Cloud. All rights reserved.



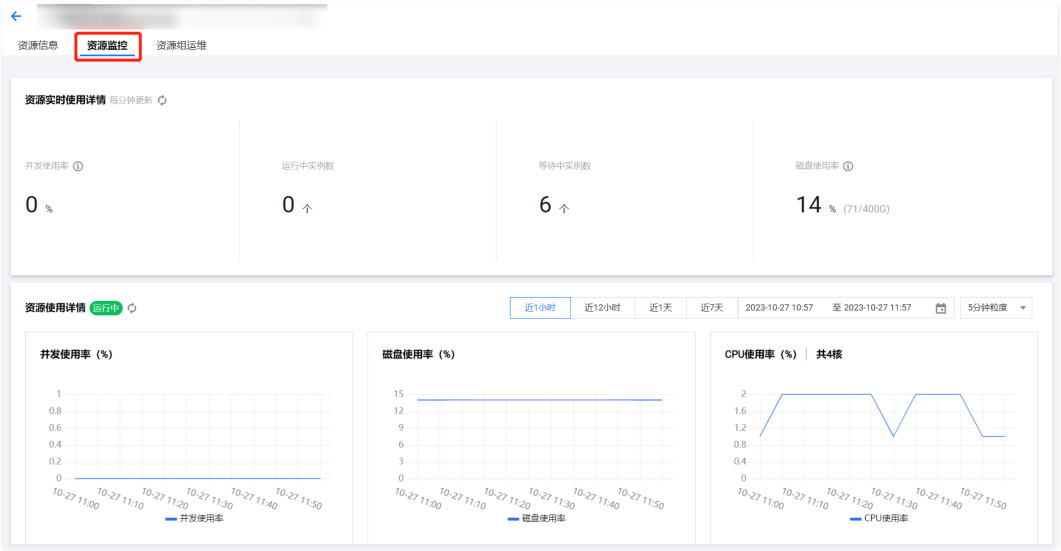
Resource Information

On the Resource Information page, you can view the basic and detailed information of the resources. For details, please refer to Resource Information.



Resource Monitoring

On the Resource Monitoring page, you can monitor the usage of the scheduling resources. For details, please refer to Resource Monitoring.



Resource Group OPS

Resource Group OPS currently provides a management feature for installing Python packages, offering support for Python tasks in data development.

©2013–2025 Tencent Cloud. All rights reserved.



The Resource Group OPS interface provides entries for installing, uninstalling, and viewing Python packages, as well as a list of command execution records. Users can view relevant information about Python installation and uninstallation here, including name, installation type, creator, status, description, and time.

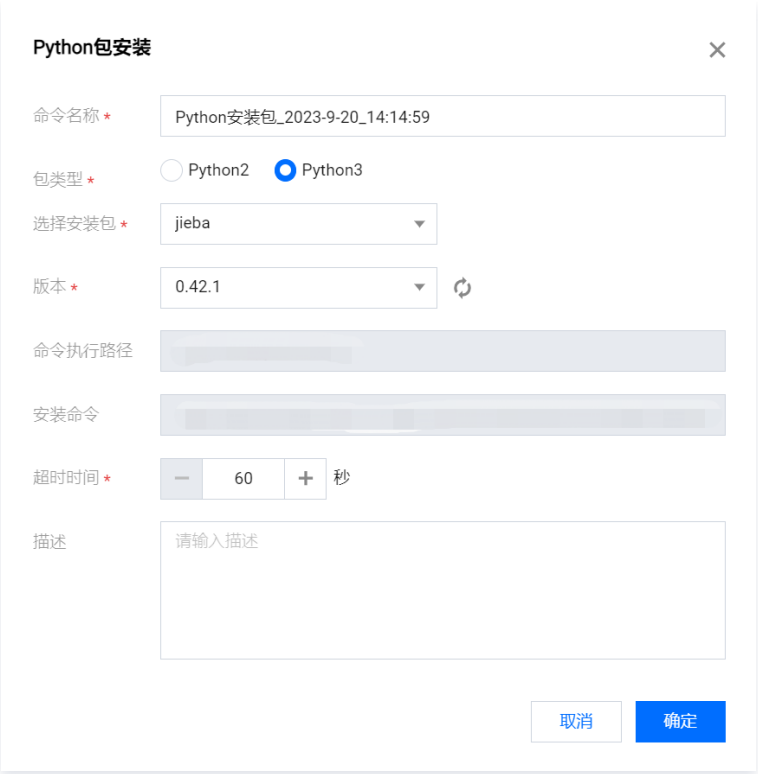


Install Python package

1. Click 'Install Python package' and configure Python package information in the popup.

> **Note:**
> This feature is only effective in the New version specifications of scheduling resources. Old specifications do not support installation.

©2013–2025 Tencent Cloud. All rights reserved.



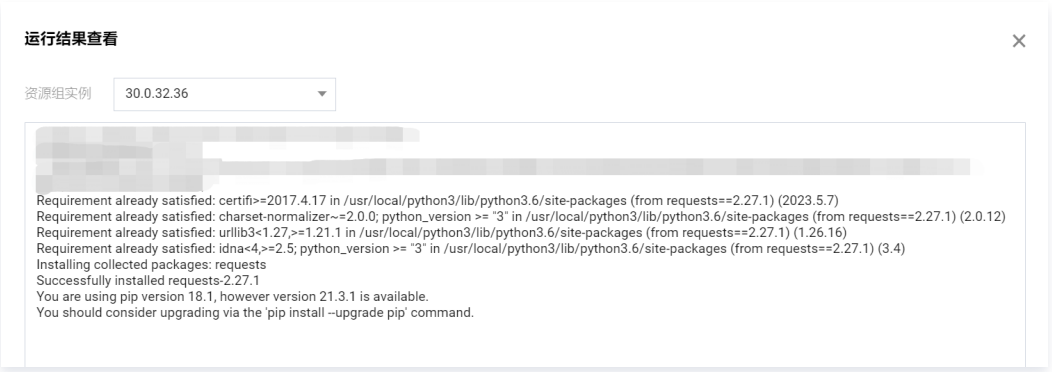
Information	Description
Command Name	Automatically generated, default is "Python Installation Package_Current Time", users can modify it.
Package Type	Python Version, including Python2 and Python3.
Select Installation Package	Supports keyword search. Currently supported installation packages are as follows: Python2:jieba,numpy,openpyxl,pandas,pyhive,requests,scikit-learn,thrift,thrift_sasl,tencentcloud-sdk-python. Python3:jieba,numpy,openpyxl,pandas,pyhive,requests,scikit-learn,thrift,thrift_sasl,tencentcloud-sdk-python,pymysql,cos-python-sdk-v5,threadpool,tencentcloud-dlc-connector,matplotlib.
Version	Different versions of the corresponding Python package can be selected.
Command Execution Path	Default path provided by WeData, Python package installation path.
Installation Command	Automatically generated.
Timeout Period	Maximum execution time of the installation command. The execution will be terminated if it exceeds the time period.
Description	Self-Definition Description Information.

2. After the configuration of Python package installation steps is completed, the installation command can be generated. On the resource group operation and maintenance page, click Run under the corresponding installation command operation column to execute the Python package installation command. After a successful installation, the status column will indicate that the command ran successfully.



©2013-2025 Tencent Cloud. All rights reserved.

3. Click **View Results** under the operation column to see the detailed execution of the installation command.

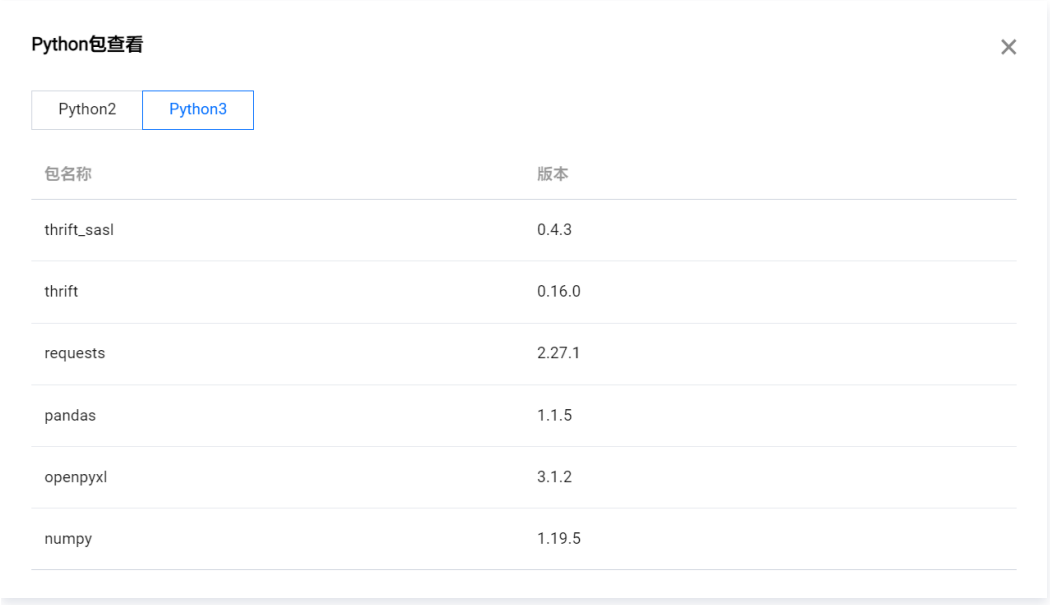


4. Click **Command Name** in the list to view command details.



View Python Package

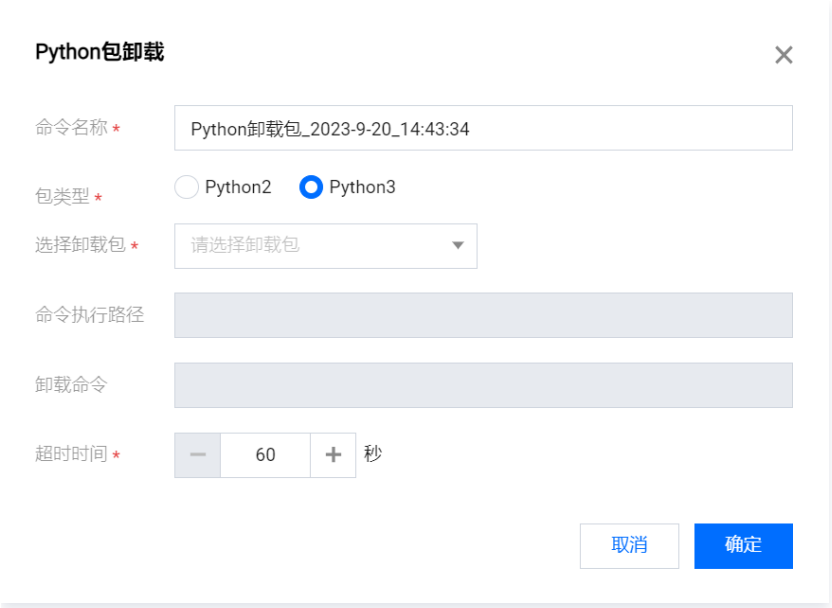
Click 'View Python Package' to see the installed Python2 and Python3 packages in the popup.



Uninstall Python Package

Click 'Uninstall Python Package' to choose the installed Python packages for uninstallation in the popup.

©2013–2025 Tencent Cloud. All rights reserved.



Information	Description
Command Name	Automatically generated, default is "Python Uninstall Package_Current Time", users can modify it.
Package Type	Python Version, including Python2 and Python3.
Select Uninstall Package	Choose from installed Python packages.
Command Execution Path	Default path provided by WeData, uninstallation path for Python packages.
Uninstall Command	Automatically generated.
Timeout Period	Maximum execution time of the installation command. The execution will be terminated if it exceeds the time period.

Custom Definition Scheduling Resources

The Enterprise edition of WeData allows associating the current user account's CVM with the Schedule Resource Group. Custom Definition Scheduling Resource Groups will be self-maintained by the user.



Creating Custom Definition Scheduling Resource Group

©2013-2025 Tencent Cloud. All rights reserved.

1. Click **Create Scheduling Resource Group** and fill in the basic information.



> **Note:**
> The Resource Group and the configured Compute Engine in this space must be located in the same region.

2. Click **Create Execution Resource Group** and fill in the basic information of the Resource Group. After completion, click **Next**.

Parameter	Description
Region	This parameter defaults to displaying the region where the project's Compute Engine is located. The Execution Resource Group must be established in the same region as the Compute Engine.
Resource Group Name	Resource Group English Name.
Resource Group Remarks	Resource Group Remarks display name.

©2013–2025 Tencent Cloud. All rights reserved.

3. Configure the server and select the associated network and server.



> **Note**
> The server and the configured Compute Engine in this space must be located in the same network.

4. Click **Save** and follow the instructions to install the Agent.

Manage from Definition Scheduling Resource Group

Add server

1. Click **Resource Group Management > Add Server** to add servers available for task execution to the current execution resource group.



©2013–2025 Tencent Cloud. All rights reserved.

2. You can choose existing CVMs (CVMs) or purchase new CVMs. After filling in the basic information of the server to be added, click **Next**. The recommended server configuration is as follows:
   - Network: In the same VPC as the project's Compute Engine (such as EMR).
   - Configuration: 32 cores and 64 GB memory (at least 2 units are recommended for production, 1 unit for testing).
   - Public bandwidth: 2 Mbps.
   - System Image: CentOS 64-bit CentOS 7.9 64-bit.
   - Hard Disk: 500 GB.

> **Note**
> Do not add EMR nodes to the execution resource group to avoid environmental conflicts.

Parameter description

Associated Network
Select the network to which the CVM server belongs. The CVM must be in the same VPC environment as the compute engine bound to the project.
Private IP
Enter the primary IPv4 private IP of the CVM server.
1. Getting method description:
   1.1 Go to the CVM Console, click **Instances,** to enter the instance list.



   1.2 In the instance list, select the server you added and click to enter its CVM details page.



   1.3 Confirm the associated network information and copy the primary IPv4 private IP.



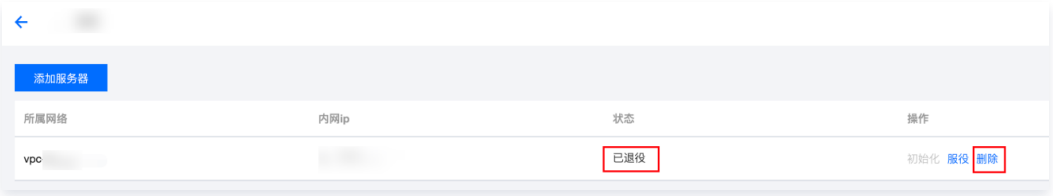
   1.4 Go to the Agent installation page and follow the prompts to install the Agent on your server.
2. Retiring/Serving Servers
   For servers in normal use, if you want to perform maintenance operations, you can choose to retire them. After retiring, the

©2013-2025 Tencent Cloud. All rights reserved.

server will enter a retiring state. Running tasks will continue to operate until completion, but no new tasks will be accepted. Once all running tasks are finished, the server will change to retired status, and you can start maintenance. After maintenance, click **Serve** to restore the server to normal use and start accepting new tasks again.



3. Deleting Servers
   If you want to remove a server from the execution resource group, you can first click the **Retire** button. Once the server status changes to retired (time taken depends on the current task duration), click the **Delete** button to delete the server for other uses.

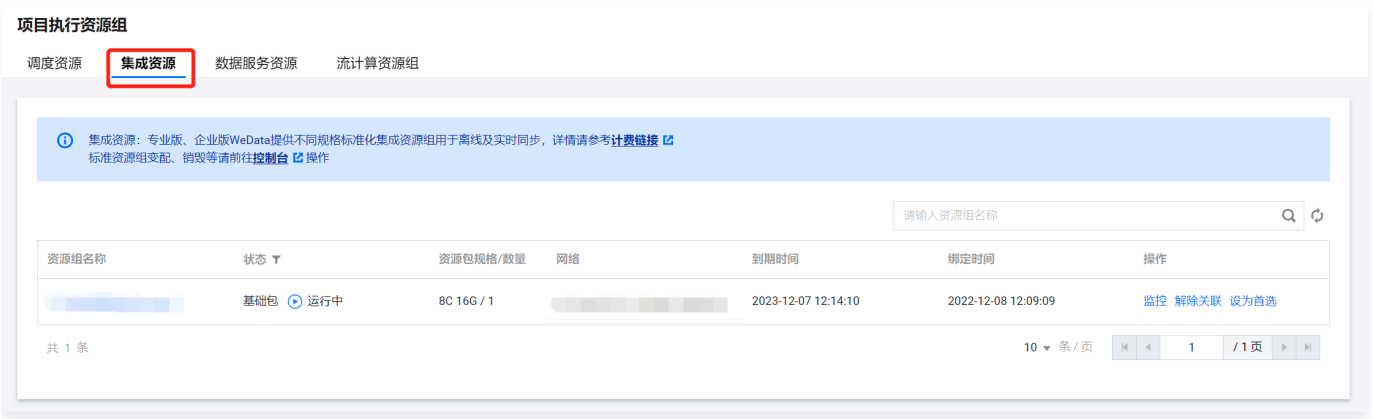


©2013–2025 Tencent Cloud. All rights reserved.

Integration resource

Last updated：2024-11-22 16:50:43

Professional Version and Enterprise Edition WeData provide standardized integration resource groups for offline and real-time synchronization in different specifications. For details, please refer to Billing Instructions for Integrated Resources. For resource group changes, destruction, etc., please go to the Console to operate.



Monitoring

Under the operation column of the integrated resource group, click **Monitoring** to dynamically monitor the usage of the corresponding resource group. When resource usage approaches saturation, you can allocate resources through Resource Scale-Out.

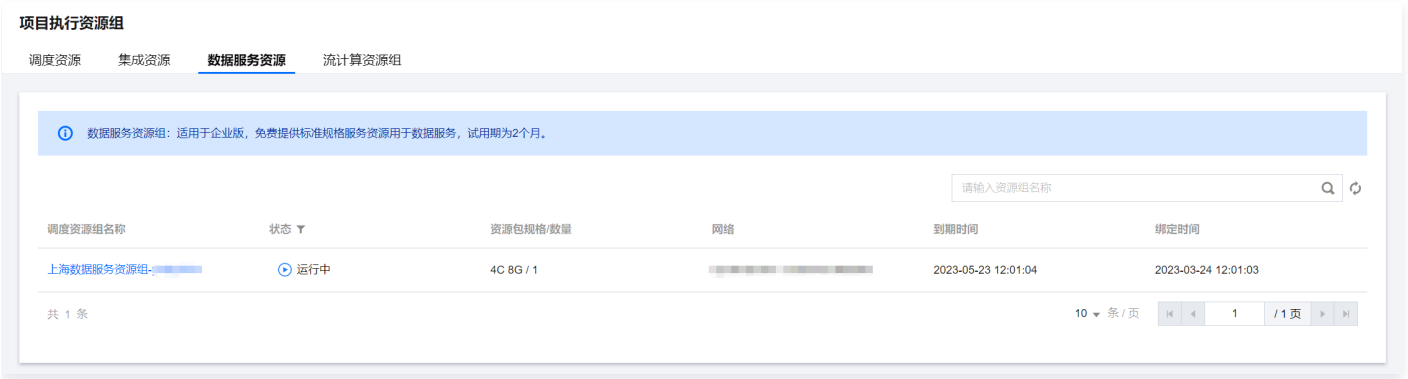


©2013-2025 Tencent Cloud. All rights reserved.

Data Service Resources

Last updated：2024-08-22 17:26:04

Applicable to the Enterprise Edition of WeData, providing standard specification service resources for data service API support. For details on data service resource billing and specifications, please refer to Service Resource Billing Explanation. For using the data service API feature, please refer to Create API.



©2013-2025 Tencent Cloud. All rights reserved.

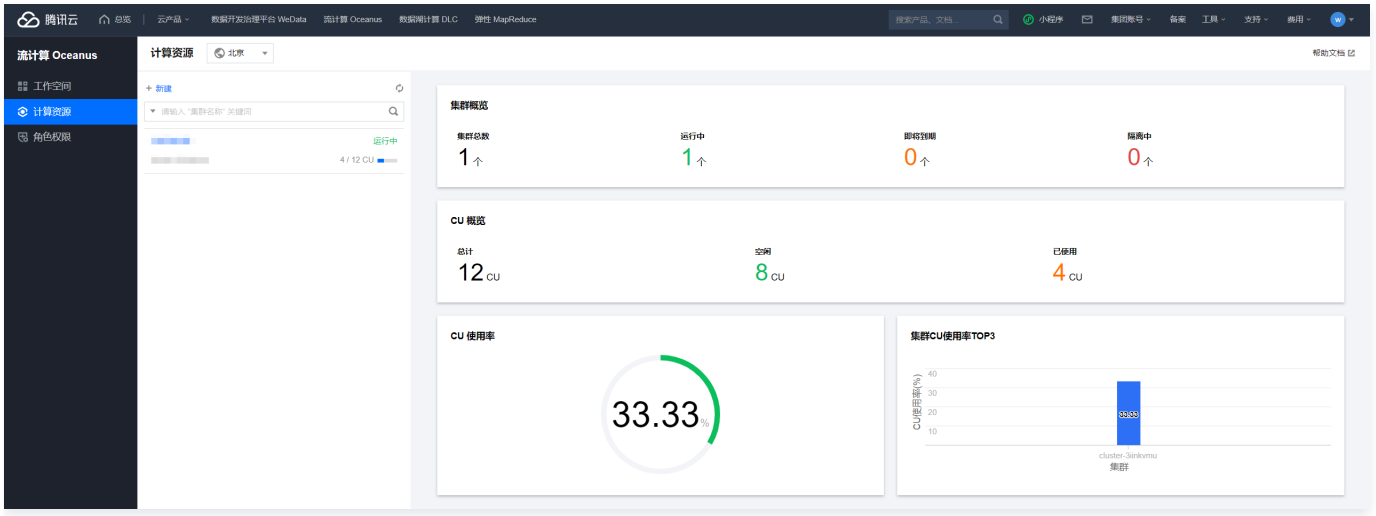
Stream Compute Resource Group

Last updated: 2024-08-22 17:26:46

Applicable to the enterprise version of WeData, SCS is a cloud-based stream data aggregation and computing service that provides fully managed cloud services. You do not need to worry about the operation and maintenance of the infrastructure, and you can easily connect to cloud data sources and get comprehensive supporting facilities.



1. Log in to SCS Console. Select Computing Resources from the left menu bar of the console.
2. Select the Region at the top left of the page, then click **Create Cluster** to create a Dedicated Cluster. For details, see Creating Dedicated Cluster.



©2013-2025 Tencent Cloud. All rights reserved.

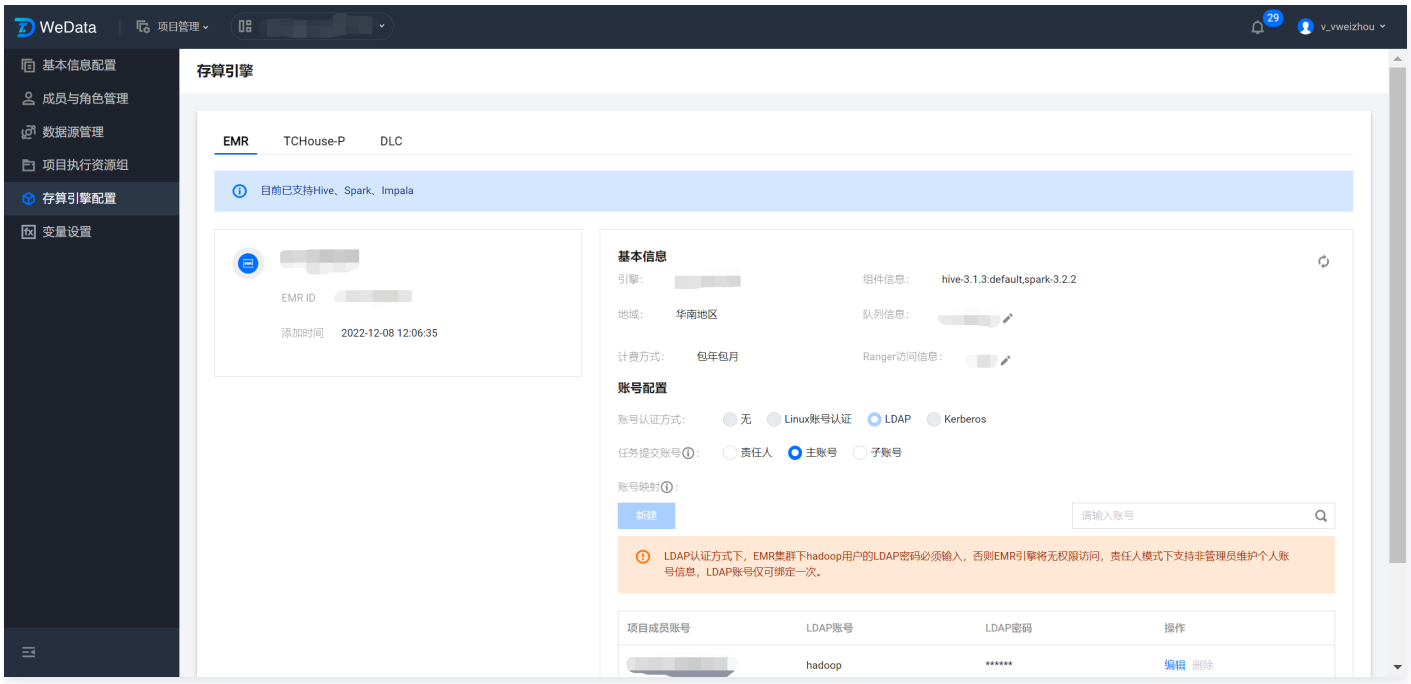
Computational Storage Engine Configuration

Last updated: 2024-09-10 16:30:03

Configure the computational storage engine, including EMR engine configuration, TCHouse-P engine configuration, and DLC engine configuration.

EMR Engine Configuration

- **Basic Information Management:** Supports configuring the basic information of the EMR (EMR) engine, including YARN queue settings and refresh.
- **Account Configuration:** Supports configuring task submission accounts and account mapping relationships for the EMR engine, adaptable to different authentication methods of EMR clusters.



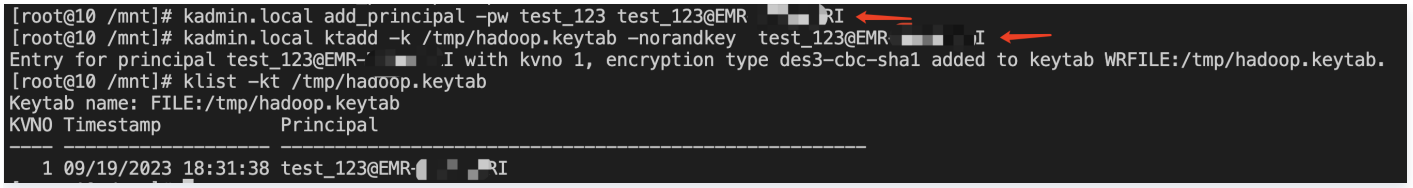
Account Authentication Methods

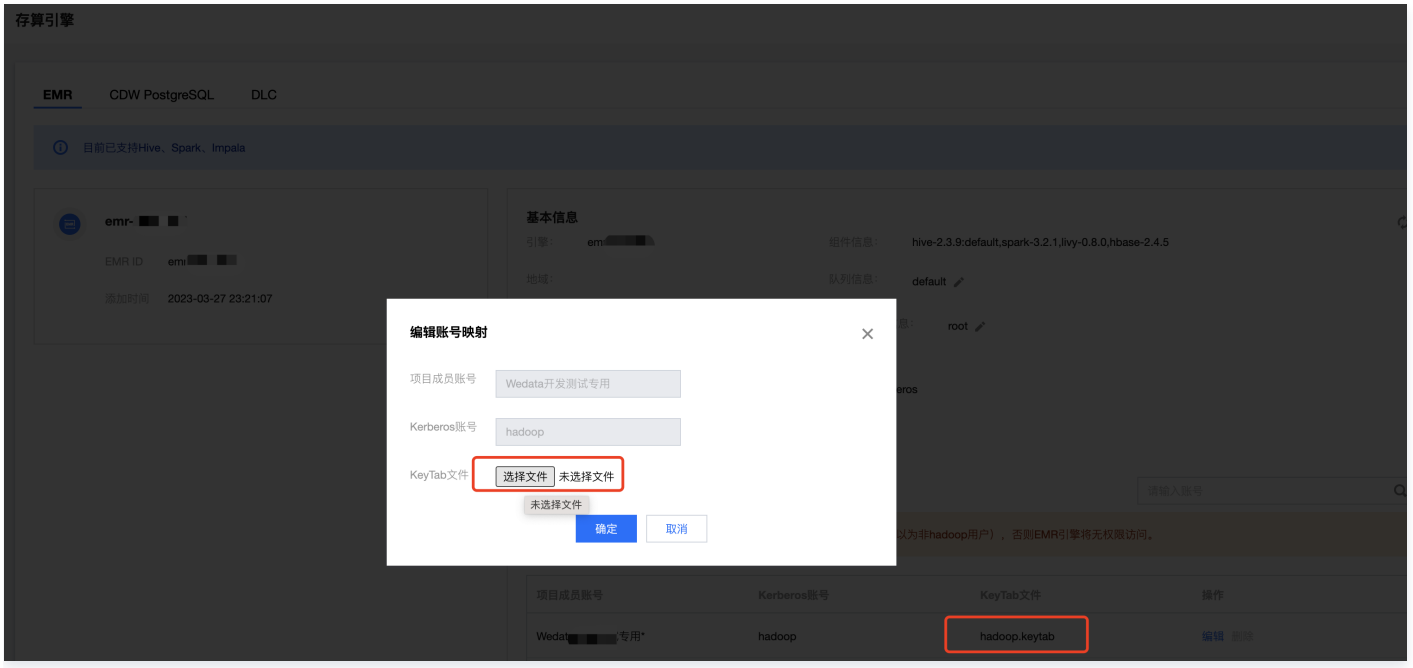
- **No Authentication**
  Indicates that the current EMR cluster has no authentication enabled. WeData uniformly submits tasks as the hadoop user.
- **Linux Account Authentication**
  Indicates that the current EMR cluster has Simple authentication enabled, submitting tasks as a Linux user. This authentication method can be one of two possibilities:
  1.1 WeData cloud account and Linux account are identical: When selecting the task submitter account as "responsible person", there is no need to configure mapping separately.
  1.2 WeData cloud account and Linux account are not identical: When selecting the task submitter account as "responsible person", it is necessary to configure mapping separately, otherwise the submission will fail on Linux due to the absence of a cloud account user.
- **Account Authentication**
  Indicates that the current EMR cluster has LDAP authentication enabled, submitting tasks as an LDAP user. Unlike Linux account authentication, LDAP user password configuration is required under this authentication.
- **Kerberos Account Authentication**
  Indicates that the current EMR cluster has Kerberos authentication enabled, submitting tasks as a Kerberos account. This authentication requires users to download the keytab file from EMR and configure the mapping relationship to allow the scheduling system to submit tasks with a legitimate identity.

> **Note**

©2013-2025 Tencent Cloud. All rights reserved.

- When the EMR cluster has Kerberos authentication enabled, the keytab file for the hadoop user is not supported for direct download from EMR and must be manually created on the EMR client. The command for creation is as follows:
  - kadmin.local add_principal -pw xxx hadoop@EMR-XXXXXXXX
  - kadmin.local ktadd -k /tmp/hadoop.keytab -norandkey hadoop@EMR-XXXXXXXX
- xxx is the keytab's password
- Download the generated hadoop.keytab configuration file, upload it to WeData, refer to the figure below.
- ⚠ The created principal is in a two-part format. The three-part hadoop principal (hadoop/_HOST@EMR-XXXXXXXX) in the emr service node cannot be directly downloaded and used.

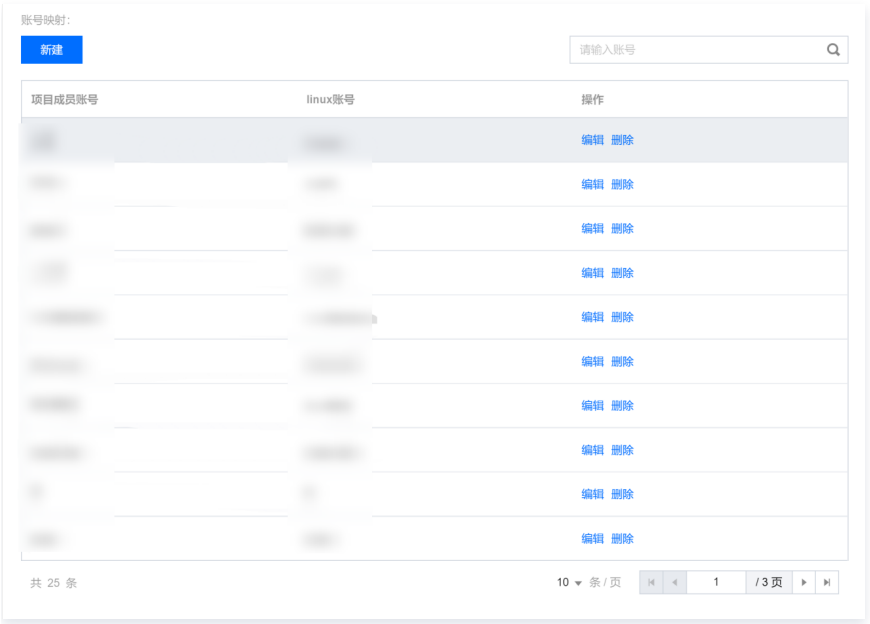




Account Mapping

When the task submission account is a Sub-account, you need to select a unified Sub-account for submission and configure the corresponding account mapping relationship. Users can create, edit, or delete account mappings here.

©2013-2025 Tencent Cloud. All rights reserved.



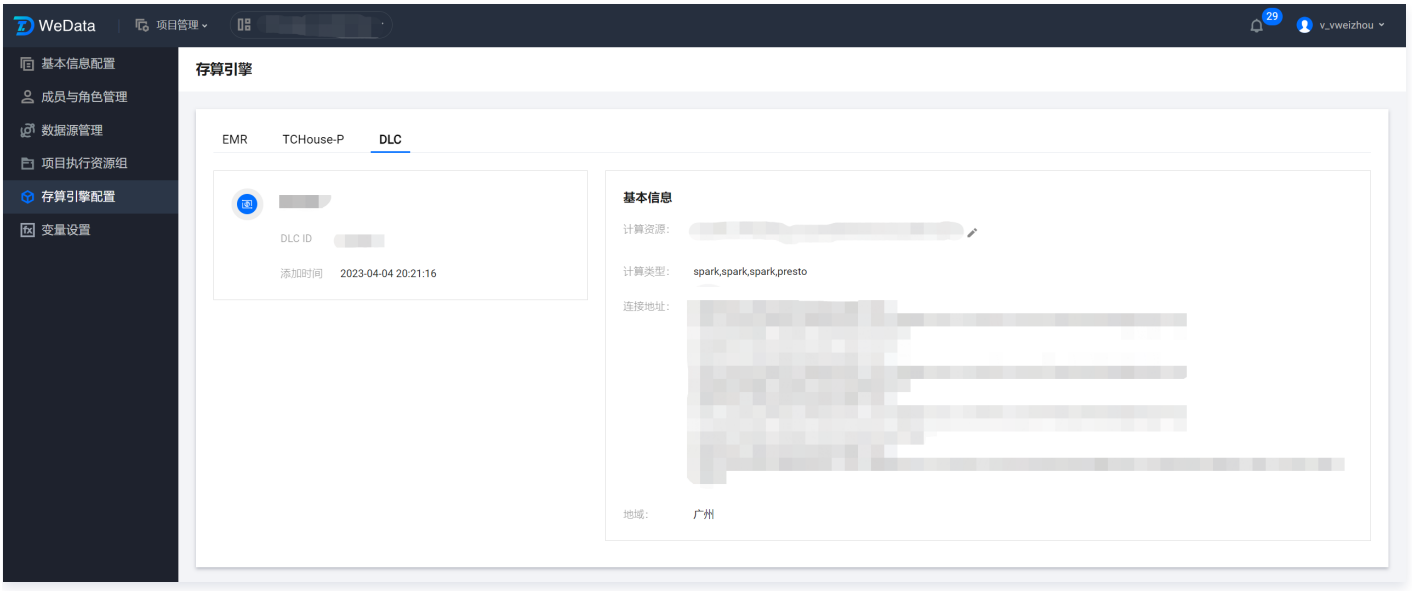
TCHouse-P Engine Configuration

Supports Cloud Data Warehouse TCHouse-P access configuration.



DLC Engine Configuration

Supports DLC access configuration.



©2013-2025 Tencent Cloud. All rights reserved.

Variable Settings

Last updated: 2025-04-18 21:59:03

Provides the feature for creating and managing project-level global variables. Users can pre-create global variables using this feature and invoke them later during Script Development and Workflow Development. It is essential to avoid variable naming conflicts and accidental modifications to prevent adverse effects on the data processing workflow.
The main purposes of variable settings are:

- Data Transmission: Facilitates data sharing between different nodes in the data processing workflow, making the actions between workflow nodes more flexible and capable of performing more complex operations.
- Improved Readability: Defining commonly used variables as global variables can make the code more readable and allow these variables to be used across multiple development nodes. This also makes the code more concise, maintainable, and easier to debug.
- Reduced Code Complexity: In some cases, several related variables with a certain logical relationship may need to be used. Defining these variables as global variables can reduce code complexity and simplify the program's logical structure.

Access the Variable Settings Page

1. Log in to Tencent Cloud WeData Data Development Platform.
2. Click **Project List** in the left-side menu.
3. Select the corresponding project and click to enter the **Project Management** module.
4. Click **Variable Settings** in the left-side menu.

Variable Operations

Add New Variables

Provides the feature to add new project-level variables. Click the **Add** button in the upper-left corner of the page, enter the parameter name and value in the list below, and then click the **Save** button to create the project variable.
For example: Add a parameter variable to get the previous day's date.
Parameter name: T-1
Parameter value: ${yyyy-MM-dd -1d}



Editing Variable

Provides the editing feature to modify variable parameters in the list. Select the variable you want to modify, click the **Edit** button in its action column to change the parameter value, and then click the **Save** button to complete the variable editing.

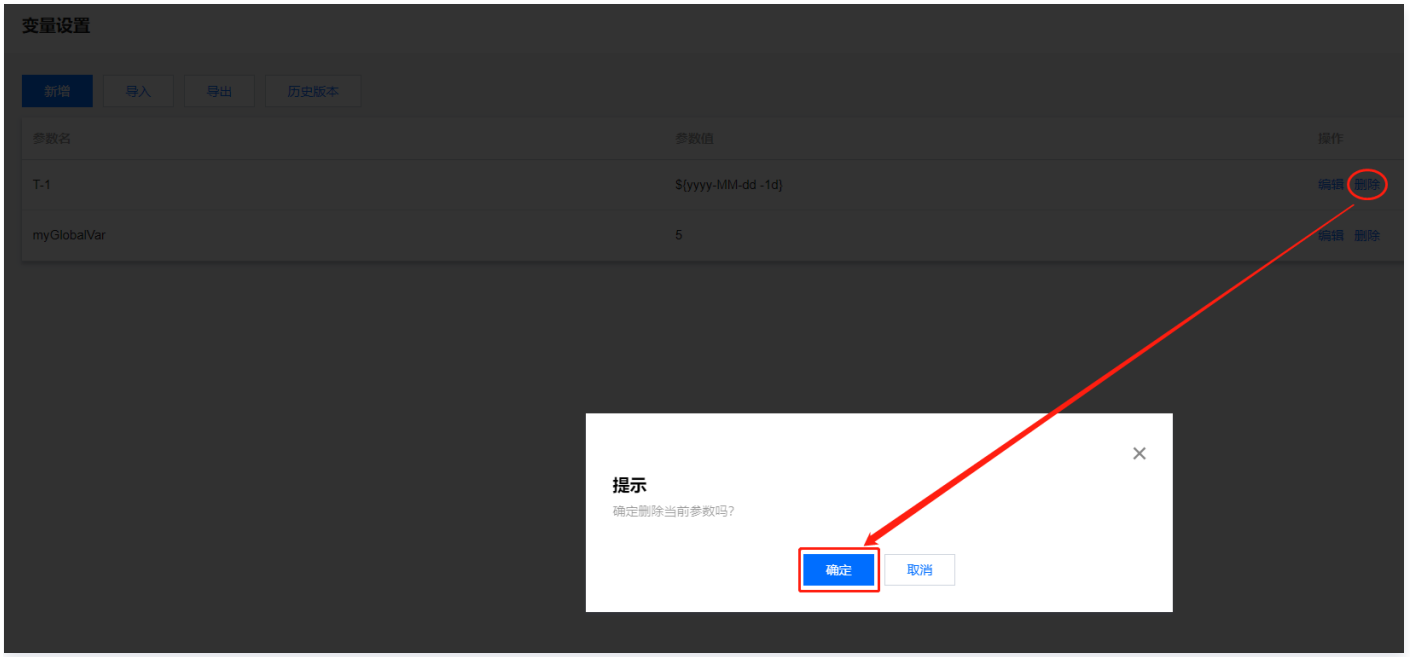




©2013-2025 Tencent Cloud. All rights reserved.

Delete Variable

Provides the delete feature to remove the created variable parameters from the list. Select the variable you want to delete, click the **Delete** button in its action column, and confirm in the pop-up box to delete the parameter.



©2013-2025 Tencent Cloud. All rights reserved.

Data Integration
Overview of DataInLong

Last updated: 2024-08-23 12:39:34

DataInLong connects and integrates various self-built data on and off the cloud quickly, addressing data platform construction, database migration backup, business upgrade, integration, data access acceleration, full-text search, and other scenarios of data integration and synchronization problems. The Tencent Cloud product DataInLong (DataInLong) provides DataInLong capabilities for WeData, supporting offline synchronization, real-time database monitoring, data reporting, and other synchronization features.

Use Limits

1. Data Synchronization: DataInLong only supports the transfer of data objects that can be abstracted as logical two-dimensional tables, such as structured, semi-structured, and unstructured (COS, etc., requiring that specific synchronized data must be abstracted as structured data) data content synchronization. The FTP method supports synchronizing completely unstructured files (such as meteorological files) to HDFS, but this transfer method does not support data content extraction.
2. Network Connectivity: Supports data storage synchronization within a single region and some cross-region data synchronization needs. Some regions can use the classic network for transmission, but connectivity is not guaranteed. If the classic network test fails, it is recommended to use the public network for connection.
3. Task Operation: Running DataInLong tasks requires using DataInLong resource groups. Complete the creation of integration resource groups before using DataInLong features. Integration resource groups include offline packages, real-time packages, etc., and can be purchased as needed based on the task type.
4. Data Consistency: DataInLong synchronization supports at least once delivery, but does not completely guarantee exactly once delivery (i.e., complete non-redundancy of data). Ensuring non-redundancy of data relies on primary keys and target-end capabilities.
5. Data Types and Precision: In offline or real-time synchronization, pay attention to type matching and precision conversion for source and target fields. If the source and target types are incompatible, or if the target field type's maximum value is smaller than the source field's maximum value (or the minimum value is larger than the source end's minimum value, or the precision is lower than the source end's precision), there is a risk of write failure or precision truncation.

Offline Synchronization

DataInLong provides offline data synchronization capabilities, which bulk read data from source tables and synchronize them to the target end periodically. For details, refer to Offline Synchronization.

Real-Time Synchronization

DataInLong offers real-time data synchronization capabilities, supporting streaming data transmission. Real-time synchronization supports real-time data consumption at single-table, sharded table, and multi-database multi-table granularity. Task types include single table synchronization, whole database synchronization, and log collection.

- Single Table Synchronization: The source end is a single table or sharded table, and the target end supports only one table. Single table synchronization uses a fixed schema pairing method, requiring you to specify the field mapping relationship between the source and target tables in the task. During task operation, only the specified source field content is written to the target field. For details, refer to Single Table Synchronization Task Configuration.
- Whole Database Synchronization: Whole database synchronization supports synchronizing all data from the entire instance or specified multiple database and table objects of the source end to multiple tables of the target end. This task does not need to specify the field mapping relationship between the source and target ends; by default, all source table fields are read, and fields are matched by name by default. For details, refer to Whole Database Synchronization Task Configuration.
- Log Collection: Log collection actively reports log file data from CVM cloud instances, self-built servers, or TKE to external target ends using Agent and SDK methods. For details, refer to Log Collection Task Configuration.

Concepts

- Data Source

  DataInLong uses data sources as target objects for reading/writing during the process. A data source can be a database or a data warehouse (e.g., EMR engine instance). Before configuring a DataInLong synchronization task, configure the relevant

©2013-2025 Tencent Cloud. All rights reserved.

information of the source and target databases or data warehouses on the data source management page. After configuration, you can control the synchronized read and write operations of databases or data warehouses by selecting the data source name in the synchronization task.

- Network Connectivity

  Before using DataInLong synchronization tasks, ensure network connectivity between the data source (including read and write ends) and the DataInLong resource group. Resources should not be blocked due to allowlist restrictions; otherwise, data transmission synchronization cannot be completed. For details, refer to Integration Connectivity and Usage Planning.
  - If the data source has public internet access: Purchase and create a NAT Gateway to allow integration resources to connect to the data source's VPC through the gateway. For detailed instructions, see the related NAT Gateway documentation.
  - If the data source is within a VPC:
    - If the data source is in the same VPC as the integration resource: You can use it directly.
    - If the data source is in a different VPC from the integration resource: Purchase a Peering Connection to interconnect the VPCs of the integration resource and the data source.
  - If the data source is in an IDC or another classic network environment: Purchase a VPN or Direct Connect Gateway to interconnect the VPCs of the integration resource and the data source.
- Rate Limit

  Rate limit is the maximum transmission speed allowed for DataInLong synchronization tasks.
- Concurrent number

  Concurrent number refers to the maximum number of parallel read or write operations allowed in a data synchronization task. The concurrency setting impacts the efficiency of data synchronization. Higher concurrency settings result in higher resource usage. Due to resource limitations or the characteristics of the task itself, the actual concurrency during execution may be less than or equal to this value.
- Dirty Data

  Dirty data refers to data that failed to be written during synchronization due to field type mismatches or errors in writing to the target data source. All failed write operations are classified as dirty data. For example, if a string data type from the source is written to an INT type target field and fails due to unreasonable type conversion, it is considered dirty data.
  In offline synchronization, you can set a dirty data threshold in the task to control the maximum number of dirty data entries during synchronization. If this threshold is exceeded, the task will be interrupted.
  In real-time synchronization, you can configure how dirty data is archived, ensuring that failed write operations are archived into storage to maintain an uninterrupted real-time data flow.

©2013-2025 Tencent Cloud. All rights reserved.

DataInLong Quick Start
MySQL Real-time Sync to DLC Guide

Last updated: 2024-09-05 17:43:54

This document provides a step-by-step guide to help you complete the process of real-time MySQL sync to DLC, including whole database migration and single table synchronization. The general process is as follows:
1. Data Source Compatibility.
2. MySQL Environment Preparation.
3. DLC Environment Preparation.
4. Project Configuration and Resource Group Preparation.
5. Real-time whole database: MySQL to DLC configuration steps.
6. Real-time single table: MySQL to DLC configuration steps.
7. Real-time Node Advanced Parameters.
8. FAQs.

1. Data Source Compatibility

MySQL Data Source

If you want to use MySQL for real-time data synchronization, you need to first confirm the MySQL data source version support, usage limitations, and supported data type conversion.

Supported Versions

Currently, DataInLong supports real-time reading at both single table and database levels for MySQL. To use real-time reading capabilities, you must follow these version restrictions:

Type	Version	Driver
MySQL	5.6,5.7,8.0.x	JDBC Driver:8.0.21
RDS MySQL	5.6,5.7, 8.0.x	
PolarDB MySQL	5.6,5.7,8.0.x	
Aurora MySQL	5.6,5.7,8.0.x	
MariaDB	10.x	
PolarDB X	2.0.1	

Use Limits

- Binlog logging needs to be enabled, and only MySQL server Binlog configuration format ROW is supported for synchronization.
- Tables without a primary key may have duplicate data because exact once cannot be guaranteed, so real-time synchronization tasks must ensure a primary key exists.
- XA ROLLBACK is not supported. Real-time synchronization tasks will not rollback data for XA PREPARE. To handle XA ROLLBACK scenarios, you need to manually remove the XA ROLLBACK tables from real-time synchronization tasks and then re-sync the tables after adding them back.
- Set MySQL Session Timeout:

  When creating an initial consistent snapshot for a large database, the connection you establish may time out while reading tables. You can prevent this by configuring interactive_timeout and wait_timeout in the MySQL configuration file.
  - interactive_timeout: The number of seconds the server waits for activity on an interactive connection before closing it. Please refer to MySQL :: MySQL 8.0 Reference Manual :: 7.1.8 Server System Variables.
  - wait_timeout: The number of seconds the server waits for activity on a non-interactive connection before closing it. Please refer to MySQL :: MySQL 8.0 Reference Manual :: 7.1.8 Server System Variables.

MySQL Data Type Conversion

©2013-2025 Tencent Cloud. All rights reserved.

The supported data types and internal mapping fields for reading MySQL data are shown below. (When processing MySQL, the data types from the MySQL data source will be mapped to the data processing engine data types first):

Field Type	Supported	Internal mapping fields	Remarks
TINYINT	Yes	TINYINT	TINYINT(1) maps to BOOLEAN Options need to be added to support TINYINT(1) mapping to bool or tinyint
SMALLINT	Yes	SMALLINT	–
TINYINT_UNSIGNED	Yes	SMALLINT	–
TINYINT_UNSIGNED_ZEROFILL	Yes	SMALLINT	–
INT	Yes	INT	–
INTEGER	Yes	INT	–
YEAR	Yes	INT	–
MEDIUMINT	Yes	INT	–
SMALLINT_UNSIGNED	Yes	INT	–
SMALLINT_UNSIGNED_ZEROFILL	Yes	INT	–
BIGINT	Yes	LONG	–
INT_UNSIGNED	Yes	LONG	–
MEDIUMINT_UNSIGNED	Yes	LONG	–
MEDIUMINT_UNSIGNED_ZEROFILL	Yes	LONG	–
INT_UNSIGNED_ZEROFILL	Yes	LONG	–
BIGINT_UNSIGNED	Yes	DECIMAL	DECIMAL(20,0)
BIGINT_UNSIGNED_ZEROFILL	Yes	DECIMAL	DECIMAL(20,0)
SERIAL	Yes	DECIMAL	DECIMAL(20,0)
FLOAT	Yes	FLOAT	–
FLOAT_UNSIGNED	Yes	FLOAT	–
FLOAT_UNSIGNED_ZEROFILL	Yes	FLOAT	–
DOUBLE	Yes	DOUBLE	–
DOUBLE_UNSIGNED	Yes	DOUBLE	–
DOUBLE_UNSIGNED_ZEROFILL	Yes	DOUBLE	–
DOUBLE_PRECISION	Yes	DOUBLE	–
DOUBLE_PRECISION_UNSIGNED	Yes	DOUBLE	–
ZEROFILL	Yes	DOUBLE	–
REAL	Yes	DOUBLE	–
REAL_UNSIGNED	Yes	DOUBLE	–

©2013–2025 Tencent Cloud. All rights reserved.

REAL_UNSIGNED_ZEROFILL	Yes	DOUBLE	–
NUMERIC	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
NUMERIC_UNSIGNED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
NUMERIC_UNSIGNED_ZEROFILL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
DECIMAL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
DECIMAL_UNSIGNED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
DECIMAL_UNSIGNED_ZEROFILL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
FIXED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
FIXED_UNSIGNED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
FIXED_UNSIGNED_ZEROFILL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
BOOLEAN	Yes	BOOLEAN	–
DATE	Yes	DATE	–
TIME	Yes	TIME	–
DATETIME	Yes	TIMESTAMP	–
TIMESTAMP	Yes	TIMESTAMP	–
CHAR	Yes	STRING	–
JSON	Yes	STRING	–
BIT	Yes	STRING	BIT(1) maps to BOOLEAN
VARCHAR	Yes	STRING	–
TEXT	Yes	STRING	–
BLOB	Yes	STRING	–
TINYBLOB	Yes	STRING	–
TINYTEXT	Yes	STRING	–
MEDIUMBLOB	Yes	STRING	–
MEDIUMTEXT	Yes	STRING	–
LONGBLOB	Yes	STRING	–
LONGTEXT	Yes	STRING	–
VARBINARY	Yes	STRING	–
GEOMETRY	Yes	STRING	–
POINT	Yes	STRING	–

©2013–2025 Tencent Cloud. All rights reserved.

LINESTRING	Yes	STRING	-
POLYGON	Yes	STRING	-
MULTIPOINT	Yes	STRING	-
MULTILINESTRING	Yes	STRING	-
MULTIPOLYGON	Yes	STRING	-
GEOMETRYCOLLECTION	Yes	STRING	-
ENUM	Yes	STRING	-
BINARY	Yes	BINARY	BINARY(1)
SET	No	-	-

DLC Data Source

If you want to perform real-time data sync to DLC, you need to confirm the supported data type conversion limits of the DLC data source first.

Use Limits

- Currently, only DLC Iceberg format tables are supported.
- When choosing Upsert mode synchronization, the DLC table must be a V2 table, with the setting write.upsert.enabled=true enabled.
- When choosing Upsert mode synchronization, you must specify the primary key. If it is a partition table, the partition field must also be added to the primary key.

Data type conversion is supported

- **Read**

DLC Data Type	Internal Type
INTEGER	Long
FLOAT,DOUBULE,DECIMAL	Double
STRING,VARCHAR,CHAR,ARRAY	String
TIMESTAMP,DATE,TIMESTAMP WITH TIME ZONE	Date
BOOLEAN	Boolean

- **Write**

Internal Type	DLC Data Type
Long	INTEGER
Double	FLOAT,DOUBULE,DECIMAL
String	STRING,MAP,STRUCT
Date	TIMESTAMP,DATE,TIMESTAMP WITH TIME ZONE
Boolean	BOOLEAN
Bytes	BINARY

2. MySQL Environment Preparation

Confirm MySQL Version

©2013-2025 Tencent Cloud. All rights reserved.

DataInLong has requirements for MySQL versions. Check whether the MySQL version to be synced meets these requirements. You can check the current MySQL database version with the following statement.

```
select version();
```

Setting MySQL Server Permissions

You must define a MySQL user with appropriate permissions for all databases monitored by the Debezium MySQL connector.

1. **Create a MySQL User (Optional)**

```
mysql> CREATE USER 'user'@'localhost' IDENTIFIED BY 'password';
```

2. **Grant Necessary Permissions to the User:**

For real-time data synchronization, the account must have SELECT, REPLICATION SLAVE, and REPLICATION CLIENT permissions on the database. Refer to the following command for execution:

```
mysql> GRANT SELECT, SHOW DATABASES, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO 'user' IDENTIFIED
BY 'password';
```

> **Note**
> Enabling scan.incremental.snapshot.enabled no longer requires the RELOAD permission (enabled by default).

3. **Refresh the User's Permissions:**

```
mysql>FLUSH PRIVILEGES;
```

See more about Permission Explanation.

Enable MySQL Binlog

1. **Check if Binlog is Enabled**

```
show variables like "log_bin"
```

- If the result is ON, it means Binlog is enabled. If it's a slave, use the following statement:

```
show variables like "log_slave_updates";
```

- If the return value is ON, Binlog is enabled. If Binlog is already enabled, you can skip the following steps.

2. **Enable Binlog**

If Binlog is confirmed not to be enabled, the following operations are required.

- For Tencent Cloud MySQL / TDSQL-C MySQL, binlog is enabled by default.
- For open-source MySQL, refer to the official documentation to enable binlog.

3. **Change Binlog format to Row**

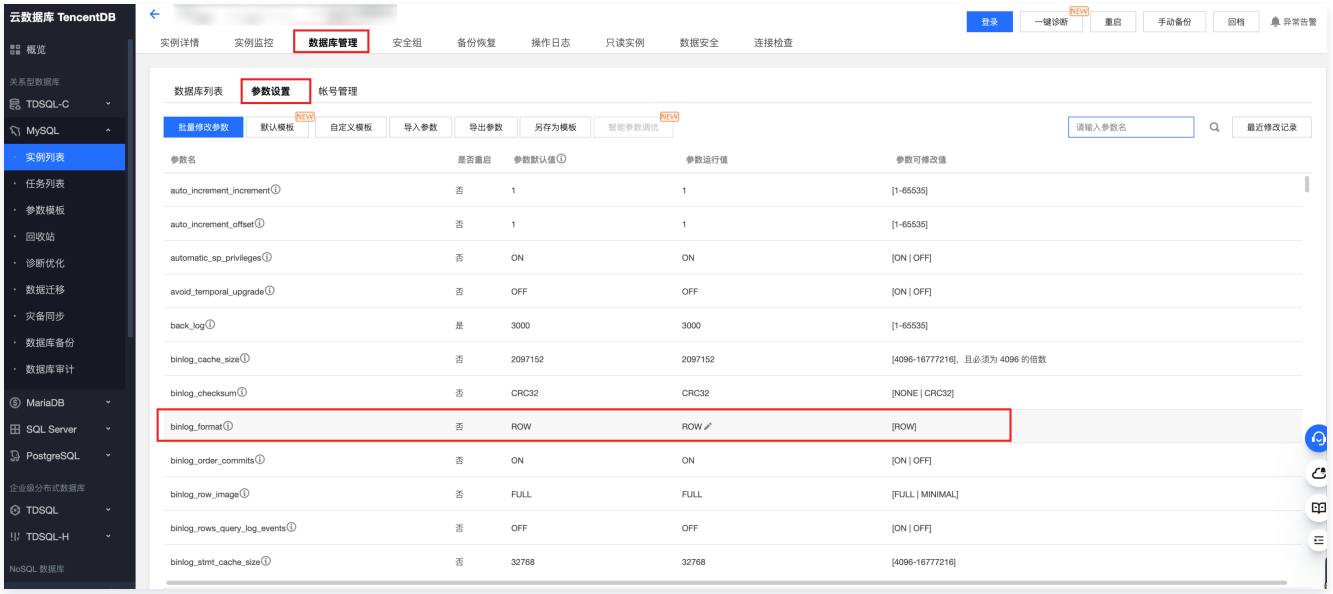
3.1 Real-time synchronization only supports MySQL server Binlog configuration format as ROW. Use the following statement to check the Binlog format.

```
show variables like "binlog_format";
```

3.2 If the return is non-ROW, please change the Binlog Format.

©2013-2025 Tencent Cloud. All rights reserved.

- For open-source MySQL, refer to the official documentation: MySQL :: MySQL 8.0 Reference Manual :: 19.1.6.4 Binary Logging Options and Variables.
- For Tencent Cloud MySQL / TDSQL-C MySQL:
  - Log in to the Tencent Cloud MySQL / TDSQL-C MySQL console, find the instance to enable Binlog, and click to enter the detailed information page of the instance.
  - In the tabs above, select DMC and find the Parameter Settings tab.
  - In the Parameter Settings tab, find the binlog_format parameter and set it to "ROW".



4. **binlog_row_image**

Real-time synchronization only supports MySQL server binlog_row_image configuration format as FULL or full.

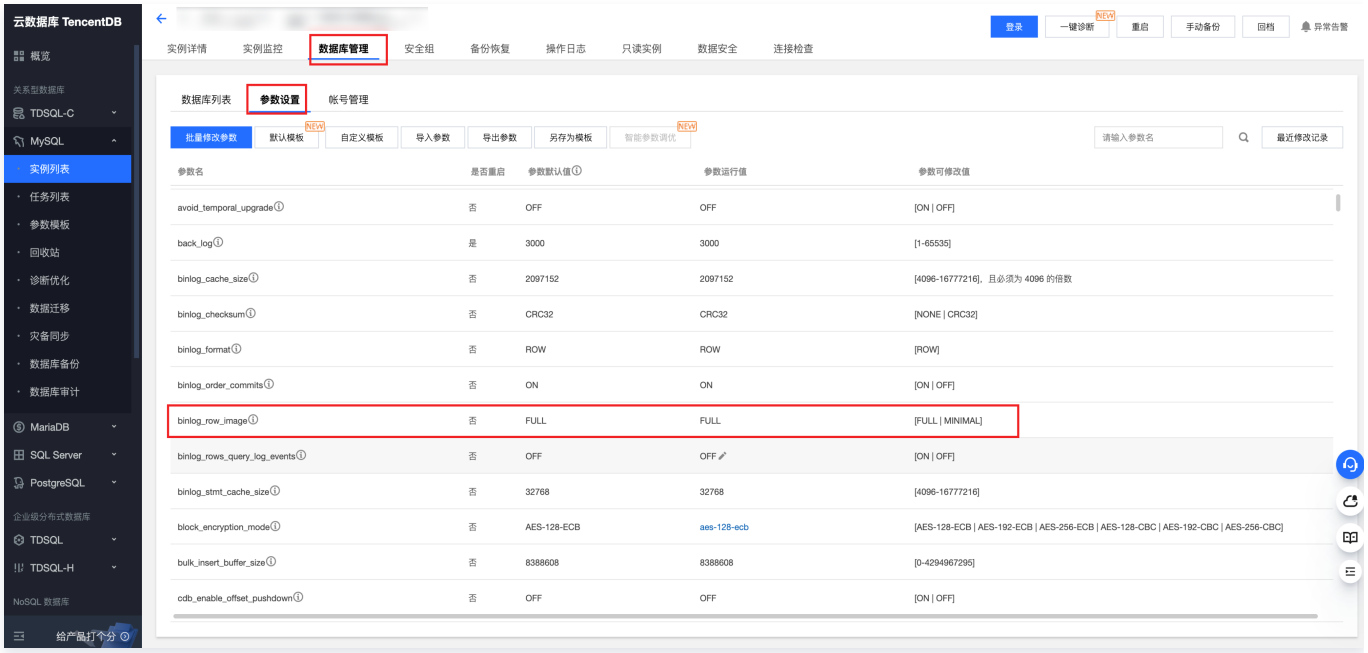
4.1 Use the following statement to query the usage format of binlog_row_image.

```
show variables like "binlog_row_image";
```

4.2 If the return is not FULL/full, please modify binlog_row_image

- For open-source MySQL, refer to the official documentation: MySQL :: MySQL 8.0 Reference Manual :: 19.1.6.4 Binary Logging Options and Variables.
- For Tencent Cloud MySQL / TDSQL-C MySQL:
  - Log in to the Tencent Cloud MySQL / TDSQL-C MySQL console, find the instance to enable Binlog, and click to enter the detailed information page of the instance.
  - In the tabs above, select DMC and find the Parameter Settings tab.
  - In the parameter settings tab, find the binlog_row_image parameter and set it to "FULL".

©2013-2025 Tencent Cloud. All rights reserved.



Enable GTIDs (optional)

GTID (Global Transaction Identifier), used to uniquely identify a transaction in binlog, can avoid data confusion or master-slave inconsistency caused by duplicate transaction execution.

1. **Check if GTID is enabled**

```
show global variables like '%GTID%';
```

If the return result is similar to the following, it proves that GTID is enabled.

```
+--------------------------+-------+
| Variable_name                      | Value |
+--------------------------+-------+
| enforce_gtid_consistency   | ON     |
| gtid_mode                                | ON     |
+--------------------------+-------+
```

2. **Enable GTID**
   - For open-source MySQL, refer to the official documentation MySQL :: MySQL 8.0 Reference Manual :: 19.1.4.2 Enabling GTID Transactions Online.
   - For Tencent Cloud MySQL / TDSQL-C MySQL instances, GTID is enabled by default and not supported to be disabled.

Add Tencent Cloud MySQL Database Security Group

A security group is a stateful virtual firewall with filtering features used to set network access control for one or more cloud databases. It is an important means of network security isolation provided by Tencent Cloud. If you use a Tencent Cloud MySQL database above the basic edition, you need to add the access IPs below to the target database's security group. If a security group is not configured, the connectivity test may fail when configuring the MySQL data source in WeData. For detailed steps, refer to Managing Cloud Database Security Groups.

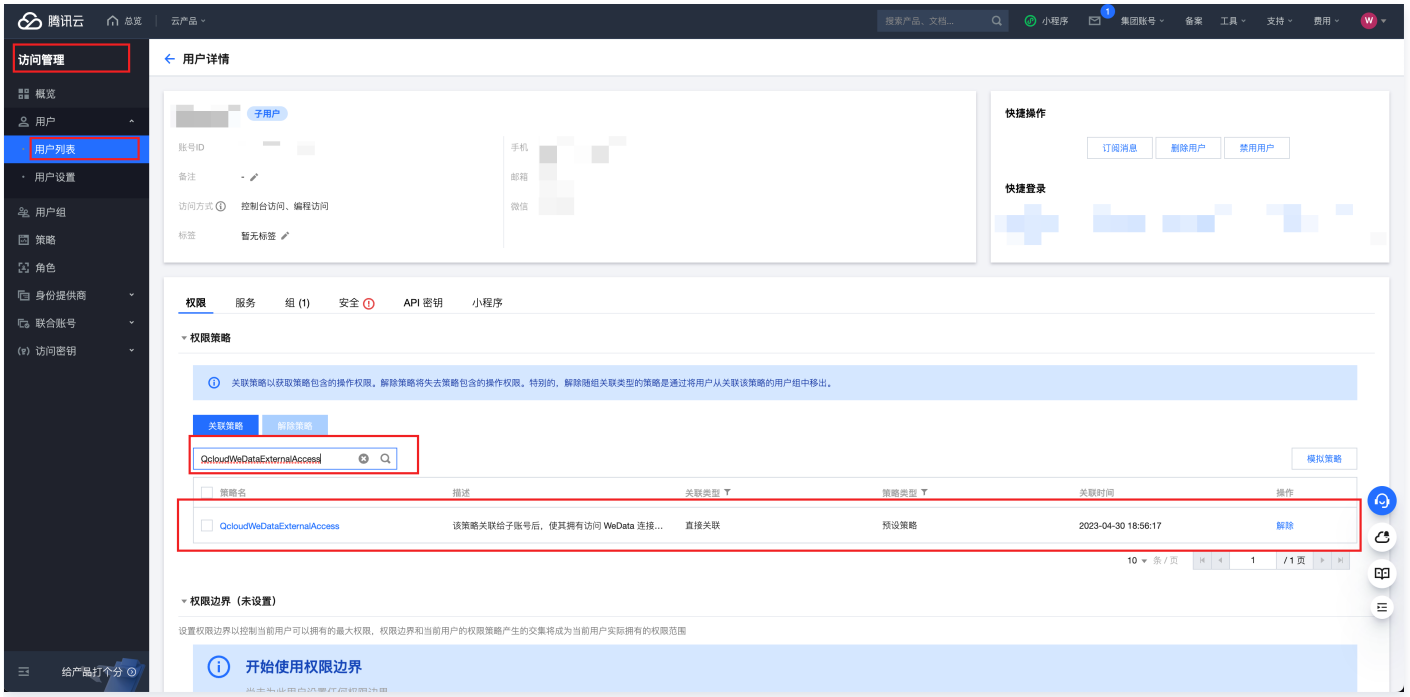
```
118.89.220.0/24, 139.199.116.0/24, 140.143.68.0/24, 152.136.131.0/24, 81.70.150.0/24, 81.70.161.0/24,
81.70.195.0/24, 81.70.198.0/24, 82.156.22.0/24, 82.156.221.0/24, 82.156.23.0/24, 82.156.24.0/24,
82.156.27.0/24, 82.156.82.0/24, 82.156.84.0/24, 82.157.119.0/24
```

3. DLC Environment Preparation

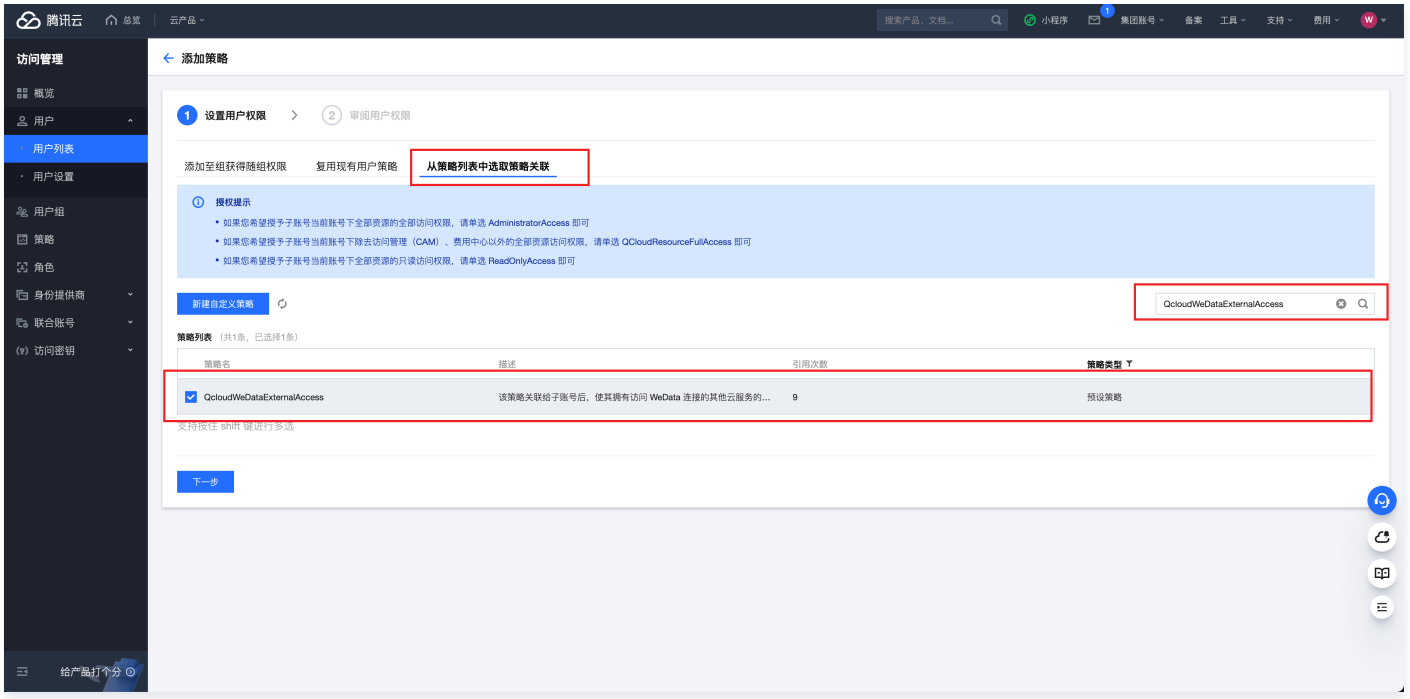
©2013-2025 Tencent Cloud. All rights reserved.

Check if WeData has access to DLC

1. Log in to Tencent Cloud Console, go to the CAM (Cloud Access Management) page.
2. In the left menu, select **Users** and find the user you need to bind the policy to.
3. Click the user's name to go to the user details page. In the permissions section, search for QcloudWeDataExternalAccess. If this policy is already associated, you can skip the following steps.



4. In the user details page, find the **Permission Management** module under **Associated Policy**.
5. Click the **Associate Policy** button under **Create Association**. Select **Select Policy Association from Policy List** and enter **QcloudWeDataExternalAccess**.
6. Select **QcloudWeDataExternalAccess**, and click the **Next** button to complete the policy binding.



After a successful binding, the user can utilize the permissions defined in the **QcloudWeDataExternalAccess** policy.

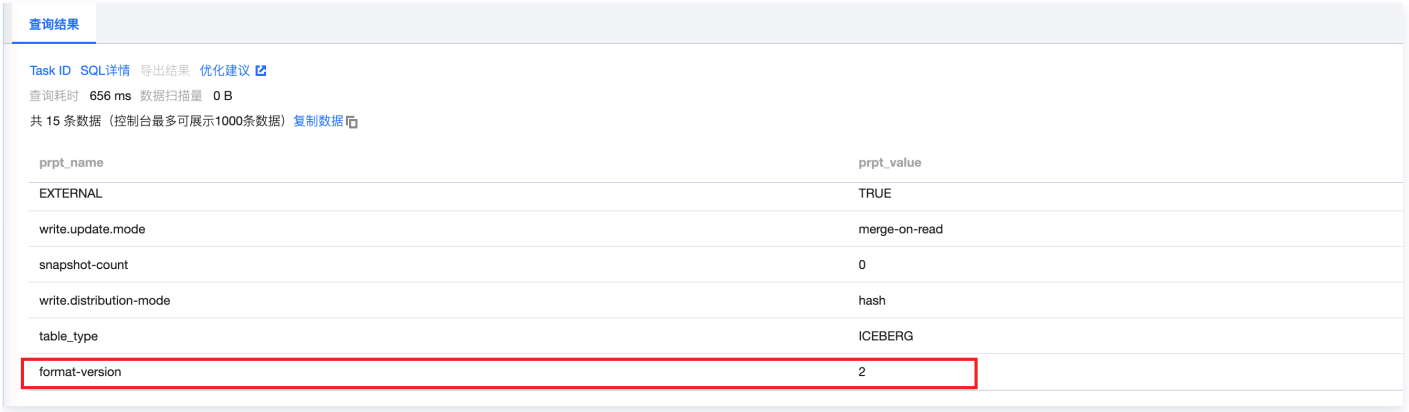
Check whether the table is a V1 Table or V2 Table (optional)

©2013–2025 Tencent Cloud. All rights reserved.

Real-time synchronization to DLC supports two modes: **Append Mode** and **Upsert Mode**. Upsert Mode must be a V2 Table. Use the command below to check the properties of the created table.

```
SHOW TBLPROPERTIES DataLakeCatalog.databases_name.table_name;
```

The returned result is as follows: format-version = 2 means the created table is a V2 table. If you created a V1 table and want to change it to a V2 table, refer to Modify Table Attributes.


查询结果

Task ID SQL详情 导出结果 优化建议

查询耗时 656 ms 数据扫描量 0 B

共 15 条数据（控制台最多可展示1000条数据）复制数据

prpt_name	prpt_value
EXTERNAL	TRUE
write.update.mode	merge-on-read
snapshot-count	0
write.distribution-mode	hash
table_type	ICEBERG
format-version	2

Create DLC Table (optional)

Real-time whole database tasks support automatically creating target tables that do not exist on the DLC during task execution, as well as using already existing DLC target tables. Users can choose to pre-create the target table in DLC, or allow the real-time whole database migration task to create the table automatically.

1. **Create V1 Table**

By default, DLC creates V1 tables. V1 tables do not support upsert mode. Refer to the following table creation statement:

```
CREATE TABLE IF NOT EXISTS DataLakeCatalog.dbname.test_v1 (id int, name string, ts date) PARTITIONED BY
(ts);
```

2. **Create V2 Table**

Using Upsert Mode to write to DLC requires creating a V2 Table, which must be specified during table creation. Refer to the table creation statement below:
When creating a V2 table, you must set 'write.upsert.enabled' = 'true', otherwise it will still be in Append mode.

```
CREATE TABLE IF NOT EXISTS DataLakeCatalog.dbname.test_v2 (id int, name string, ts date) PARTITIONED BY
(ts) TBLPROPERTIES (
  'format-version' = '2', -- Create V2 Table
  'write.upsert.enabled' = 'true', -- Perform upsert operations when writing, supported only for V2 tables
  'write.distribution-mode' = 'hash', -- Determines the distribution of written data. Set to hash for
supporting multi-threaded writing
  'write.update.mode' = 'merge-on-read' -- Update mode for writing, performing merge operations during
write, supported only for V2 tables
)
```

3. **Modify Table Attributes (optional)**

If you need to modify the attributes of an already created table, refer to the following syntax:

```
SHOW TBLPROPERTIES table_name [('property_name')]
```

For more DLC SQL syntax, refer to SQL Syntax Overview.
Here is an example of converting an existing V1 Table to a V2 Table:

©2013-2025 Tencent Cloud. All rights reserved.

```
ALTER TABLE
  DataLakeCatalog.database_name.table_name
SET
  TBLPROPERTIES (
    'format-version' = '2',
    'write.upsert.enabled' = 'true'
  )
```

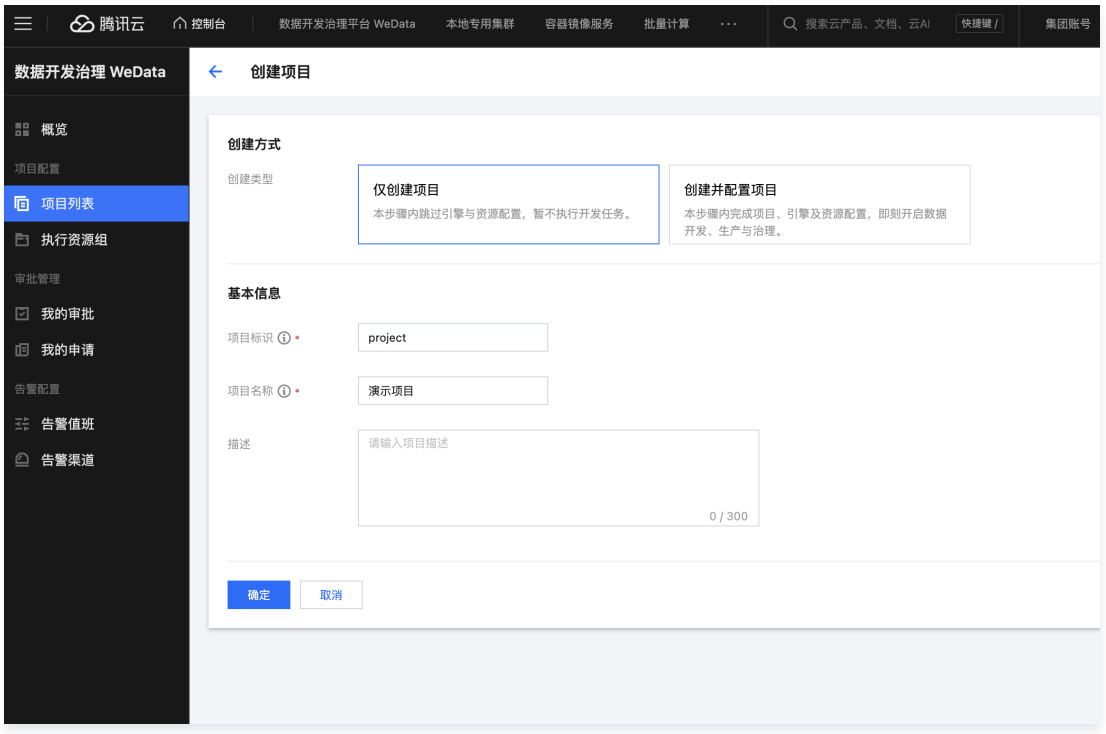
4. Project Configuration and Resource Group Preparation

Before configuring the synchronization task, a project needs to be created. The following steps are based on a new environment. If you have already created project members or completed project creation, you can skip this step.

Project Creation

If you haven't created a project space in WeData, you need to complete the project creation before performing data synchronization tasks.

1. The main account holder logs into the CAM console to add a sub-account. Refer to Preparing a CAM Sub-account. Once created, the member can use the set sub-account to log in to the WeData big data platform.
2. To create a project space, go to the WeData homepage, click **Project List** > **Create Project**. If you are only doing DataInLong, you can choose **Create Project Only** and enter the project identifier and name.

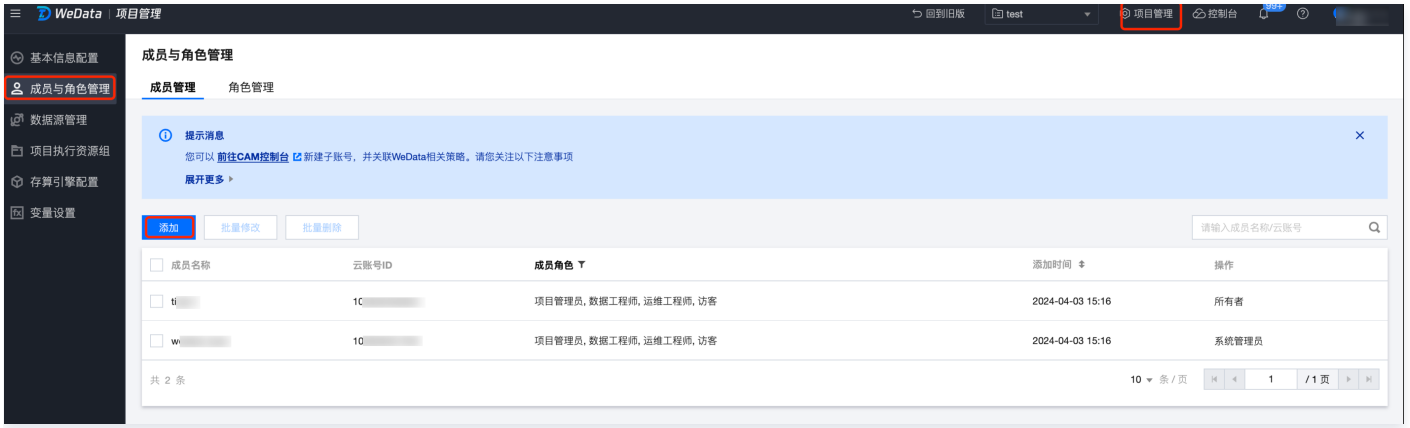


Add Member

Only the project owner and main account have permission to enter the newly created project. Other sub-accounts need to be added by the project owner or the main account in the project management module. For details, please see Project Management.

1. Click **Project Management** > **Member and Role Management** to enter the Member and Role Management interface. Click the **Add** button.

©2013-2025 Tencent Cloud. All rights reserved.



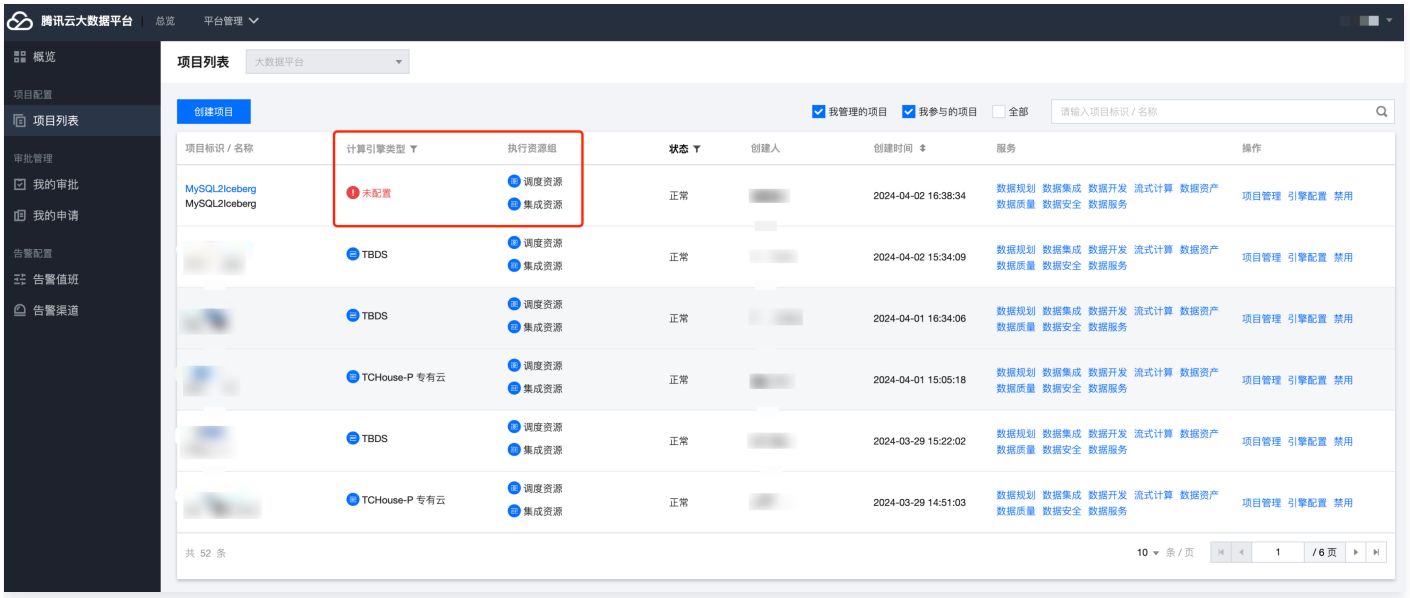
2. In the Add Member interface, set the required roles for the newly created member. Once set, the member can enter the project.



Resource Group Preparation

Confirm whether the Resource Group is configured

1. Before officially starting the DataInLong task, please confirm whether your project is equipped with **Integration Resources**.
   - If it prompts that the integration resource is not configured, the DataInLong task cannot run normally.
   - When you finish creating a new project, the column for **Execution Resource Group** will show that the integration resource is not configured. You just need to go to the WeData homepage > **Execution Resource Group** > **Integration Resource Group** and associate resources for the new project.



2. Enter the Integrated Resource Group interface, click on **Associated Project** after the corresponding resource group.

©2013–2025 Tencent Cloud. All rights reserved.

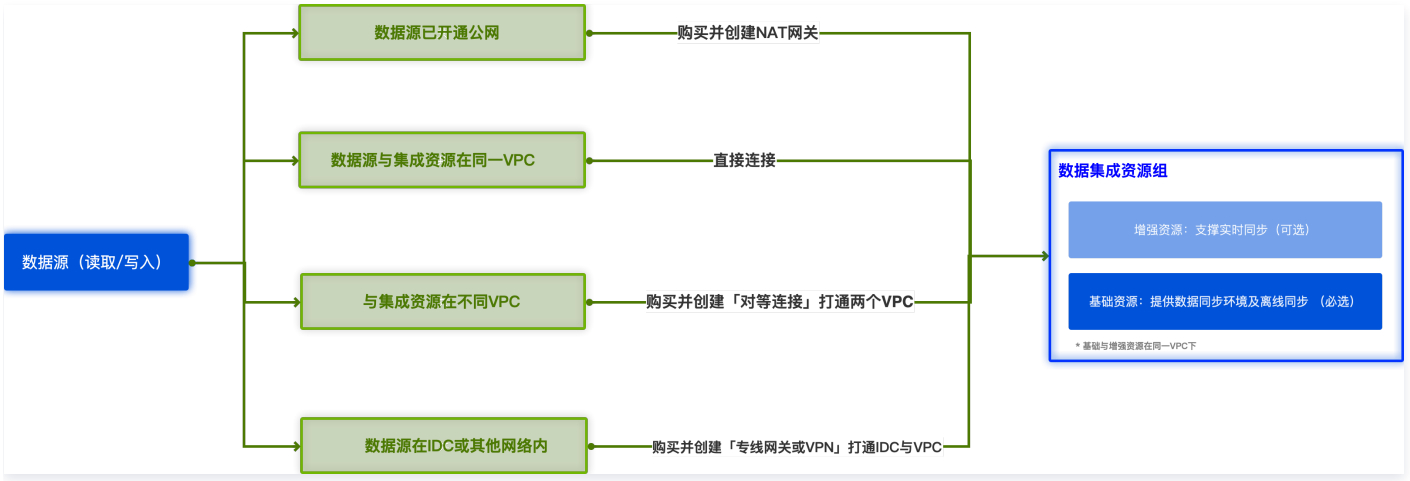


Purchase Integration Resource Group

If you find that there is currently no integration resource group, you need to go for purchase. The Integration Resource Group is dedicated to using computing resources when running DataInLong tasks. These resources are mainly presented in the form of resource groups. Before configuring the synchronization task, you need to confirm whether the DataInLong Resource Group is purchased. For details, please refer to Configure the Integration Resource Group.

Confirm Network Connectivity

Ensure network connectivity between the data source network (including the read endpoint and write endpoint) and the DataInLong resource group. Resources should not be denied access due to allowlist restrictions, as this will prevent data transmission and synchronization.
The machine resources contained in the DataInLong Resource Group must be in the same VPC network environment by default:



- **If the data source has public network access**: You need to purchase and create a NAT Gateway to allow the integration resources to connect to the data source VPC through the gateway. For detailed operations, please refer to Resource Group Configuration for Public Network.
- **If the data source is within a VPC:**
  - If the data source is in the same VPC as the integration resource: You can use it directly.
  - If the data source is in a different VPC from the integration resource: Purchase a Peering Connection to interconnect the VPCs of the integration resource and the data source.
- **If the data source is located in an IDC or other traditional network environment**: You need to purchase a VPN or Direct Connect Gateway to interconnect the VPCs of the integration resource and the data source.

©2013–2025 Tencent Cloud. All rights reserved.

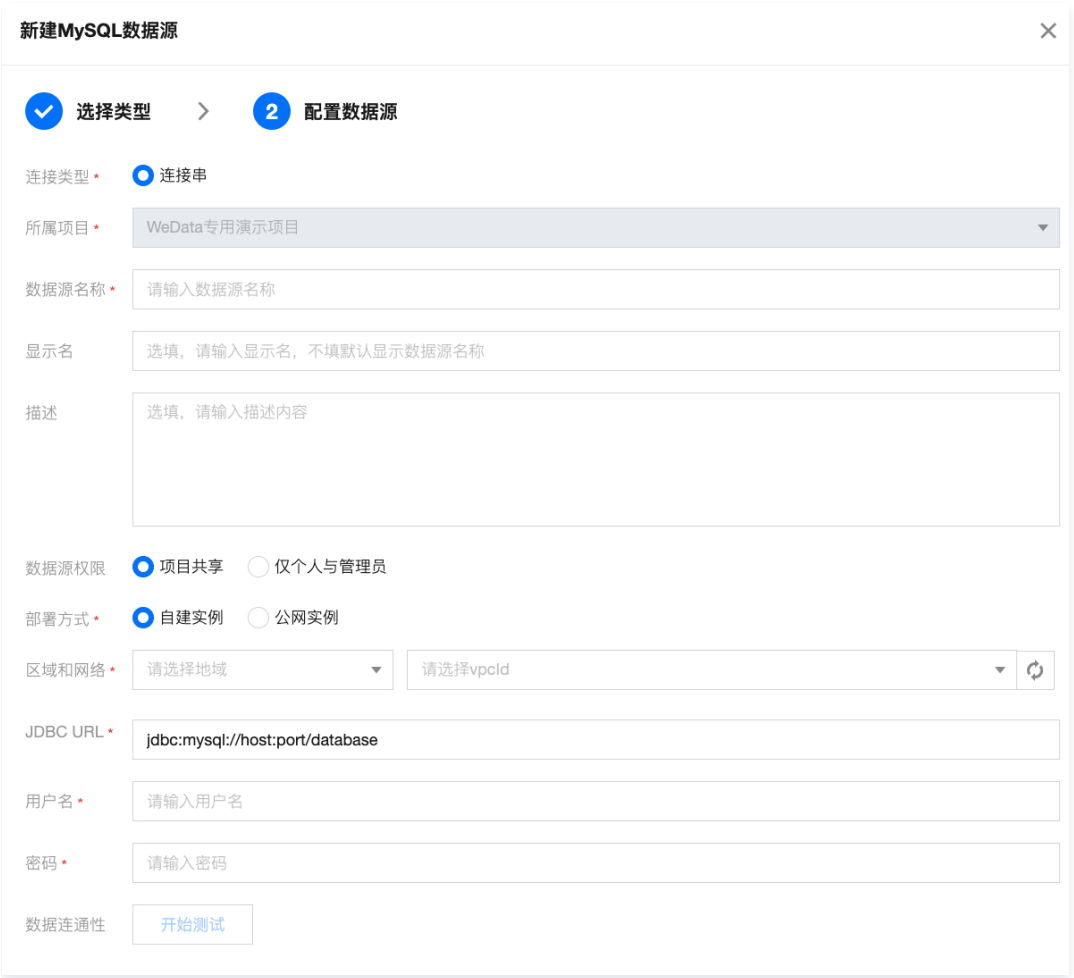
> **Note**
> For more information on the connectivity and usage planning of the Integrated Resource Group, refer to Integrated Resource Configuration and Management.

Data source configuration

MySQL Data Source Configuration

Enter the data source configuration interface, MySQL data sources support two connection methods: cloud instance and connection string.
click **Project Management** > **Data Source Management** > **New Data Source** > **Choose MySQL Data Source**.

- Create a data source via connection string.



The parameters are described as follows:

Parameter	Description
Data source name	The name of the new data source, defined by the user, cannot be empty. Naming must start with a letter, can contain letters, numbers, and underscores. Length within 20 characters.
Description	Optional, description of this data source.
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it.
Deployment mode	Supports two deployment modes: self-built instances and public network instances. Self-built instances are data source instances deployed on Tencent CVM, while public network instances are those in the customer's local IDC or other cloud resources, accessible via the public network.

©2013–2025 Tencent Cloud. All rights reserved.

Region and Network	When choosing a self-built instance, you need to select the region and vpcID where the data source instance is located.
JDBC URL	The connection string information used to connect to MySQL data source instances, including host IP, port, database name, etc.
Database name	Name of the database to connect to.
Username	Username for connecting to the database.
Password	Password for connecting to the database.
Data Connectivity (Legacy)	Test whether the configured database is connected. **Note** • If the connectivity test fails, the data source can still be saved. However, it cannot be used if the connectivity test fails and the data source is saved. • If the connectivity test fails, it may be because WeData is blocked by the database network firewall. You need to add Tencent Cloud MySQL Database Security Group. Refer to Chapter 2 Add Tencent Cloud MySQL Database Security Group. • After the Chinese New Year in 2024, newly purchased WeData does not include this feature by default. This feature has been upgraded to Resource Group Connectivity.
Resource Group Connectivity (New)	Support the connectivity test between specific resource groups and data sources. Users only need to ensure that their data source and the VPC network where the Integrated Resource Group is located are connected without additional network interconnection.  **Note** Users who purchased WeData before the Chinese New Year in 2024 do not have this feature. If you need to use it, please contact Tencent's operation and maintenance personnel.

DLC Data Source Configuration

WeData currently supports introducing DLC data sources via connection strings.
click **Project Management > Data Source Management > Create Data Source > Select DLC Data Source.**

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data source name	The identifier name for the new data source is defined by the user and cannot be empty. The name should start with a letter, can include letters, numbers, and underscores, and must be within 20 characters. It is similar to a unique identifier and cannot be modified
Display name	The display name for the new data source is defined by the user and can be modified. The display name is optional. If not provided, the data source name is displayed by default.
Description	Optional, description of this data source
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it
JDBC URL	Connection string information for connecting to the DLC data source
Data Connectivity (Legacy)	Test whether the configured database is connected. **Note** • If the connectivity test fails, the data source can still be saved. However, it cannot be used if the connectivity test fails and the data source is saved. • After the Chinese New Year in 2024, newly purchased WeData does not include this feature by default. This feature has been upgraded to Resource Group Connectivity.
Resource Group Connectivity (New)	Support the connectivity test between specific resource groups and data sources. Users only need to ensure that their data source and the VPC network where the Integrated Resource Group is located are connected without additional network interconnection.

©2013–2025 Tencent Cloud. All rights reserved.



Note

Users who purchased WeData before the Chinese New Year in 2024 do not have this feature. If you need to use it, please contact Tencent's operation and maintenance personnel.

5. Real-time whole database: MySQL to DLC

When an enterprise needs to migrate an entire database to a new platform or environment, whole database synchronization can synchronize all tables and data at once, ensuring data integrity and consistency.

Step 1: Create a whole database migration task

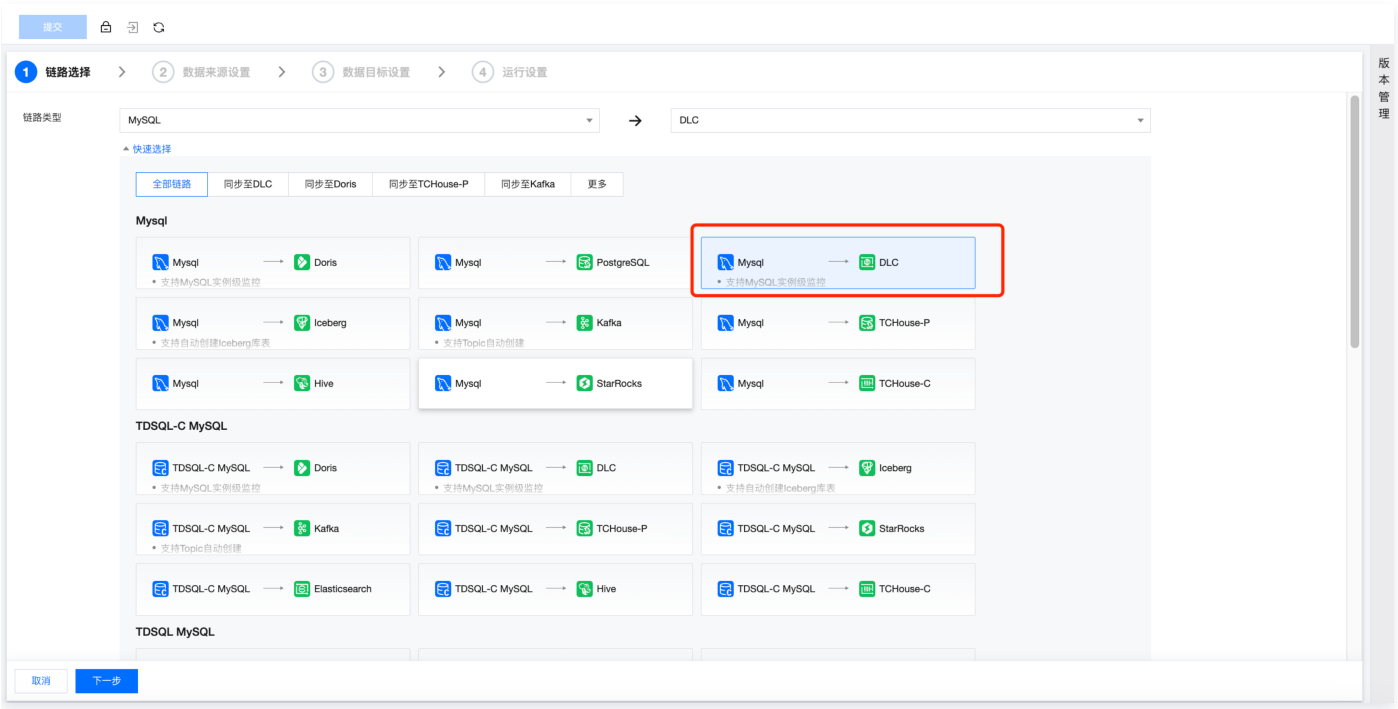
Log in to WeData, open the DataInLong module, then click **Configuration Center** > **Real-time Sync Task** page, and click **New** whole database migration task.



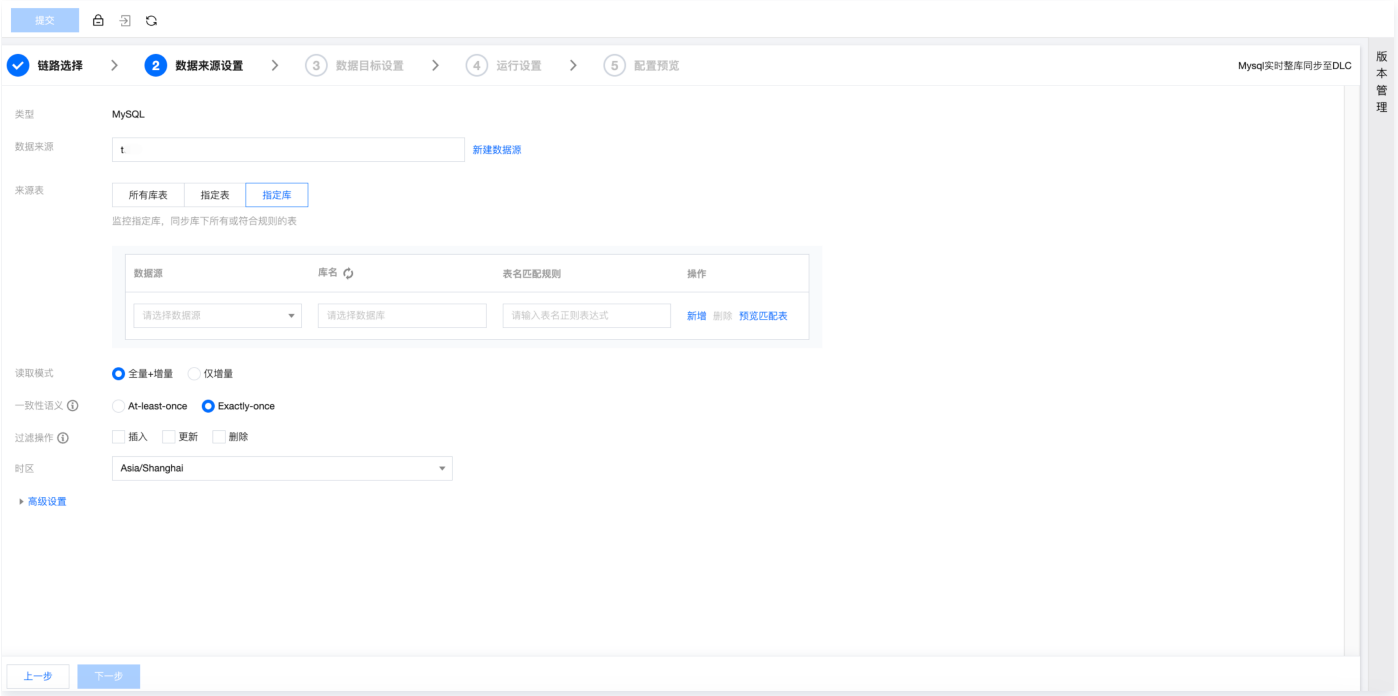
Step 2: Link Selection

In the link selection interface, choose **MySQL** > **DLC Link**. After selecting, click **Next** .

©2013-2025 Tencent Cloud. All rights reserved.

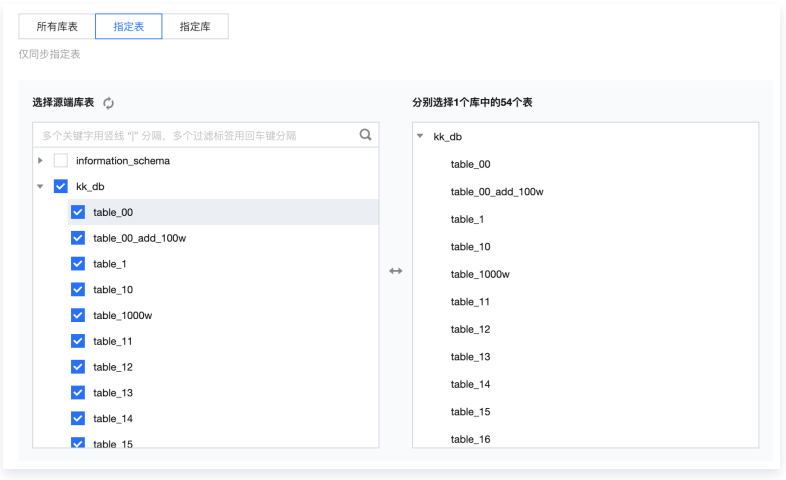


Step 3: Data Source Settings



Parameter	Description
Data Source	Select the MySQL data source for synchronization.
Source Table	All databases and tables: Monitor all databases under the data source. During the task execution, new databases and tables will be synchronized to the target end by default. Specified Table: Under this option, you need to specify the specific table name. After setting, the task will only synchronize the specified table; if a new table needs to be synchronized, the task must be stopped and restarted.

©2013–2025 Tencent Cloud. All rights reserved.



Specify Database: Under this option, you need to specify the specific database name with a table name regular expression. After setting, new tables that match the table name expression will be synchronized to the target end by default during task execution.



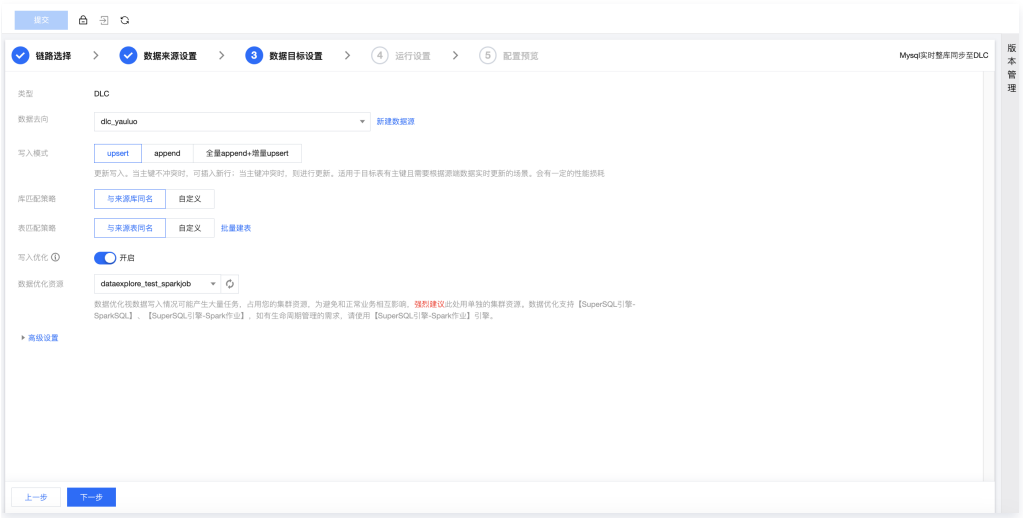
Common table name regular expression:

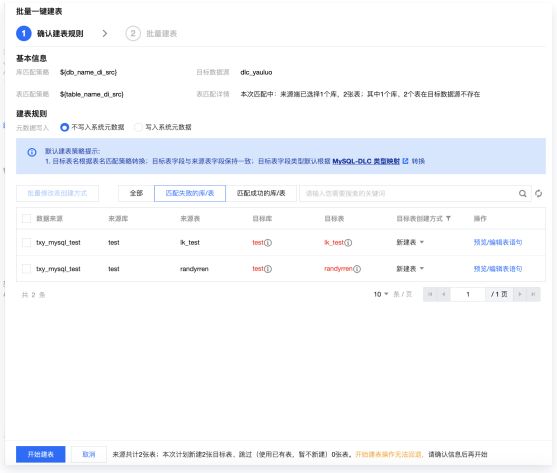
- All tables in a single database: .*
- Specific tables in a single database: table1|table2|table3
- Exclude specific tables: ^(?!(exclude1|exclude2|exclude3)$).*$
- Exclude specific tables but include all tables starting with 'a': ^(?!(exclude1|exclude2|exclude3)$)(^a.*)$

Read mode	• Full + Incremental: Data synchronization is divided into full and incremental synchronization phases. After the full phase is completed, the task enters the incremental phase. The full phase will synchronize historical data in the database, and the incremental phase will start synchronizing from the binlog CDC location after the task starts. • Increment only: Only sync data starting from the binlog cdc point after the task starts.
Filter operations	Support three operations: Insert, Update, and Delete. After setting, data of the specified operation types will not be synchronized.
Time zone	Set the log time timezone, default is Shanghai.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Step 4: Data Target Settings

©2013–2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Flow Direction	Select the target data source for synchronization.
Write mode	• upsert: Update write. If there is no primary key conflict, a new row is inserted; if there is a primary key conflict, the data is updated. Suitable for scenarios where the target table has a primary key and needs to be updated in real-time based on source data. There will be some performance overhead. • append: Append write. Regardless of whether there is a primary key, data is appended by inserting new rows. Primary key conflicts depend on the target. Suitable for scenarios without a primary key and allowing data duplication. No performance overhead. • Full append + Incremental upsert: Automatically switches data writing mode based on the synchronization phase of the source data. Uses append for the full phase to improve performance and upsert for the incremental phase for real-time updates. This mode currently supports data sources from MySQL, Oracle, and TDSQL MySQL.
Database/Table Matching Strategy	Name Matching Rules for database and data table objects in DLC.
Batch Table Creation	The DLC target end supports one-click batch table creation based on the database/table matching strategy: • Confirm Key Table Rules: Confirm key table information like data source and source database, supporting preview/edit table statements.  ○ Basic Information: Displays the selected or filled-in database matching strategy, table matching strategy, target data source, and table matching details. ○ Table Creation Rules:

©2013–2025 Tencent Cloud. All rights reserved.

- Metadata Write: You can choose whether to write metadata. Metadata supported for MySQL to DLC is as follows:
  - database_wedata_di (parameter: database, indicating the database name)
  - table_wedata_di (parameter: table, indicating the table name)
  - type_wedata_di (parameter: type, indicating the operation type)
  - ts_wedata_di (parameter: ts, indicating the log time)
  - processing_time_wedata_di (processing time, the local time of the machine processing this record)
- Unmatched Databases/Tables: Displays the unmatched databases/tables in the target data source based on the database/table matching rules
- Matched Databases/Tables: Displays the matched databases/tables in the target data source based on the database/table matching rules
- Preview/Edit Table Statements: You can preview automatically generated table creation statements and edit them directly

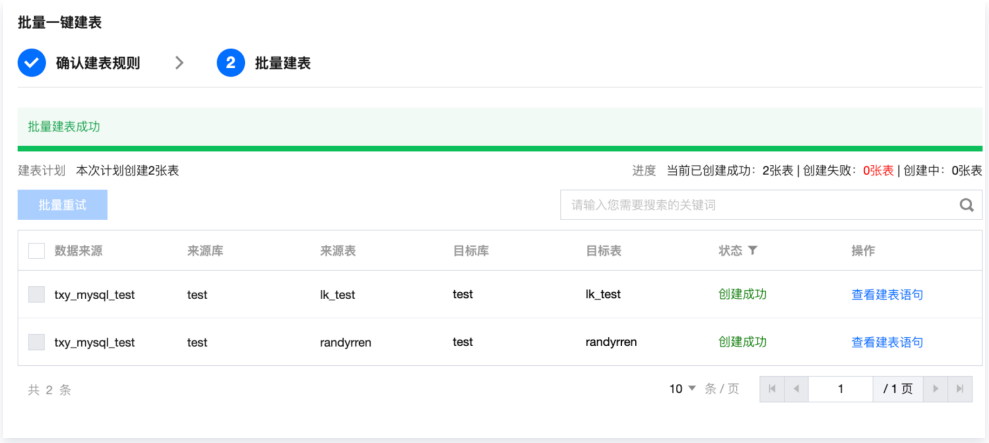


- Batch Modification of Table Creation Method: Supports batch modification of table creation methods
  - Failed Match Table Creation Method:
    - Create New Table: Support switching to Do Not Create New
    - Do Not Create New: Support switching to Create New Table
  - Successful Match Table Creation Method
    - Use Existing Table: Support switching to Delete Existing Table and Create New
    - Delete Existing Table and Create New: Support switching to Use Existing Table



- Batch Key Table: Display table creation results, support viewing table creation statements, retry.

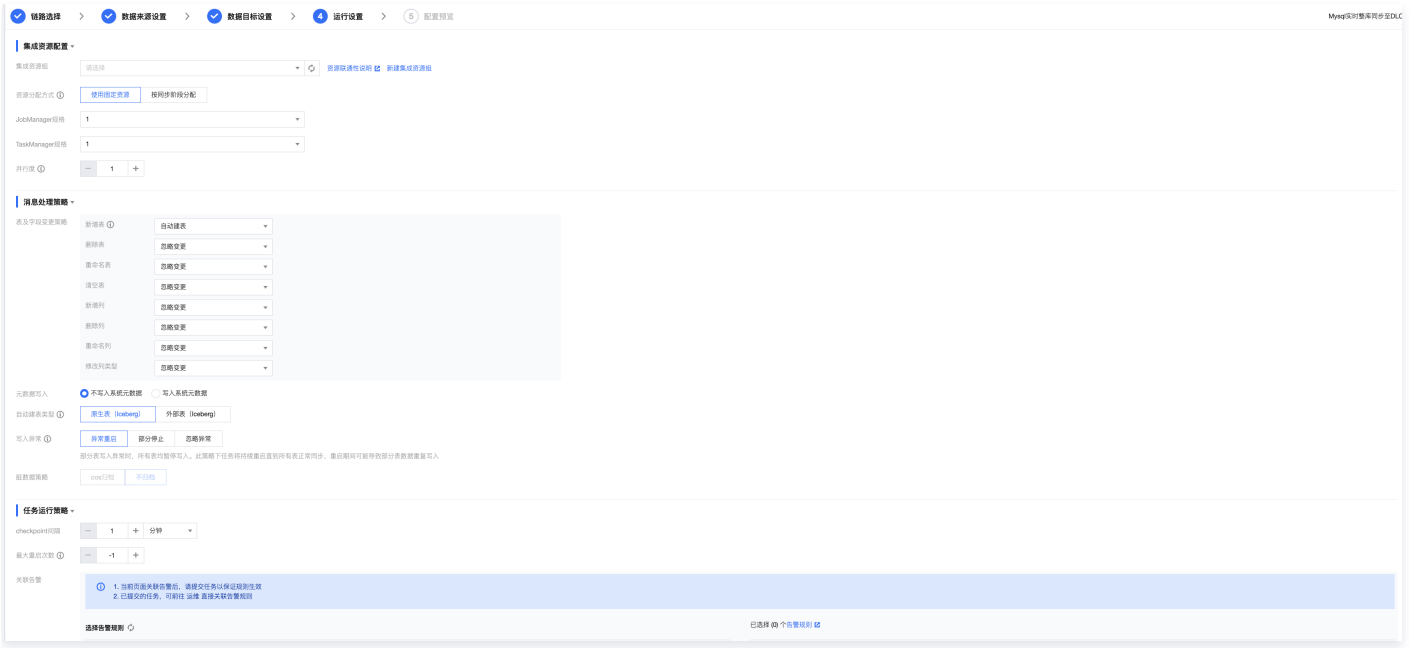
©2013–2025 Tencent Cloud. All rights reserved.



Write Optimization	Write optimization is suitable for scenarios of frequent writing/real-time writing. Enabling write optimization will automatically merge files and delete expired snapshots. It is highly recommended to enable write optimization in real-time Upsert writing scenarios. For details, please refer to Data Optimization.
Data Optimization Resources	Data optimization may generate a large number of tasks based on data write conditions, occupying your cluster resources. To avoid interference with normal business, it is strongly recommended to use separate cluster resources here. Data optimization supports **SuperSQL Engine** > **SparkSQL**, **SuperSQL Engine** > **Spark Job**. If you have lifecycle management needs, please use **SuperSQL Engine** > **Spark Job** engine.
Advanced Settings	You can configure parameters based on business requirements.

Step 5: Run Settings

This step mainly configures resources for the task, DDL and exception data processing strategies, and advanced runtime parameters.



1. Integration Resource Configuration:
   - Associate the current task with the corresponding integration resource group and set the runtime JM, TM specifications, and task parallelism. The actual CU count used during task execution = JobManager specification + TaskManager specification × parallelism.
   - Resource Allocation Method: Integration resources support multiple allocation methods
     - Fixed Allocation: In this method, resource allocation remains constant throughout the full and incremental synchronization phases, preventing resource preemption among tasks. It is suitable for scenarios where data may significantly change during task execution.

©2013–2025 Tencent Cloud. All rights reserved.

- Allocation by synchronization phase: Allocates planned resource usage based on different synchronization phases (full and incremental) to save overall resource usage.

2. Table and Field Change Strategy:

MySQL > DLC Whole Database Link supports the following DDL changes:

Table and Field Change Type	Support Policies
Add Table	Automatically Create Table, Ignore Change, Log Alerts, Task Error
Deleting a table	Ignore Change, Log Alerts, Task Error
Rename Table	Ignore Change, Log Alerts, Task Error
Clearing table data	Ignore Change, Log Alerts, Task Error
Add Column	Automatically Add Columns, Ignore Change, Log Alerts, Task Error
Delete Column	Automatic Change, Ignore Change, Log Alerts, Task Error
Rename Column	Automatic Change, Ignore Change, Log Alerts, Task Error
Modify Column Type	Automatic Change, Ignore Change, Log Alerts, Task Error

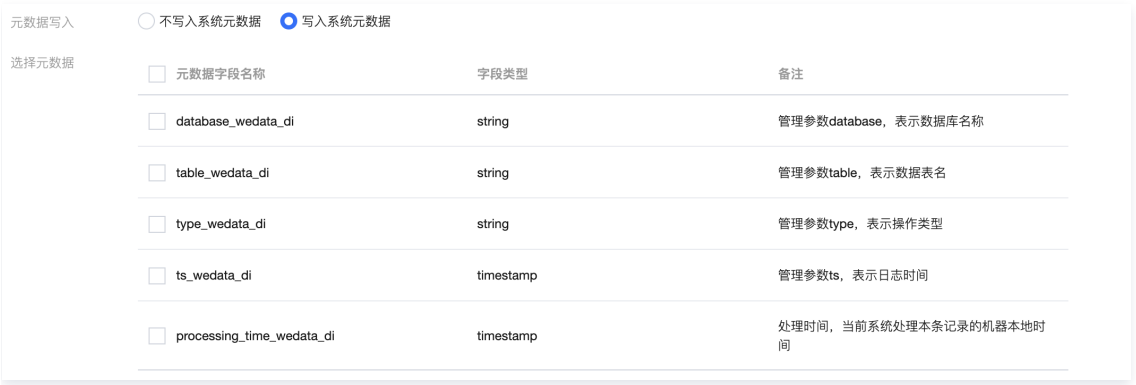
The following are the policies and display effects supported in the table above:

Policy name	Display Effect
Ignore Change	1. No change to target table/field 2. The task runs normally, changes are ignored by default
Logging and Alarming	1. No change to target table/field 2. The task runs normally, but the log prompts DDL change message details
Stop the task	1. No change to target table/field 2. Abnormal Task Restart
Auto change	The target end automatically changes according to the source end and automatically creates target objects. For example, adding tables, adding columns, etc. **Note** To avoid structural changes that might affect user business due to accidental operations, only additions, such as new tables/columns, support automated processing. Other high-risk operations, such as clearing tables, dropping tables, and renaming tables, only support the "reminder-style" response policies (e.g., alarm, stop)

3. Metadata Writing

When the new table option in the DDL strategy is set to **Automatically Create Table**, you can choose whether to write metadata. Selecting the corresponding metadata field will create the respective metadata field in the target table, and the corresponding system metadata will be written during synchronization. Currently, the MySQL > DLC whole database link supports the following metadata fields:

©2013-2025 Tencent Cloud. All rights reserved.



4. Write Exception Policy:

The DLC whole database synchronization task provides task-level runtime resources and data failure write processing strategies. Four data write failure handling strategies are supported

Policy name	Policy Description
Partial Stop	When a table write exception occurs, only the data writing to that table will be stopped; other tables will continue to synchronize normally. The stopped table cannot resume writing during this task's operation.
Abnormal Restart	When a table write exception occurs, all tables will pause writing. The task will continuously restart under this policy until all tables synchronize normally. During the restart period, some table data may be redundantly written.
Ignore Exception	Ignore the data that cannot be written within the table and mark it as dirty data. Other data in the table, as well as other tables in the task, will synchronize normally. Dirty data can be archived to COS or not archived. • COS Archiving: Archives dirty data that cannot be written. Requires configuration of COS data source, bucket, storage directory, content separator, and line break. • Non-Archiving: No additional operations required.

Scenario Example:

In Task Task1, 50 tables are planned to be synchronized. During the task operation, Table A has new fields added or field type changes:

- Partial Stop: After running the task for Table A, the field "DEMO" underwent a type change, and the new field type cannot match the target end type for writing. Under this policy, the task will stop Data Reading from the source table A, and subsequent tasks will only synchronize the remaining 49 tables to the target end.
- Abnormal Restart: After running the task for Table A, the field "DEMO" underwent a type change, and the new field type cannot match the target end type for writing. Under this policy, the task will continuously restart, and all 50 tables configured in the task will pause data write until the field in Table A is corrected.
- Ignore Abnormal: After running the task for Table A, the field "DEMO" underwent a type change, and the new field type cannot match the target end type for writing. Under this policy, the task will ignore the abnormal data that cannot be written, mark it as Dirty Data, and synchronize the other data in the table normally.

5. **Task Operation Strategy:**

Set the commit interval, maximum restart times, and task-level runtime parameters for the current task.

©2013-2025 Tencent Cloud. All rights reserved.



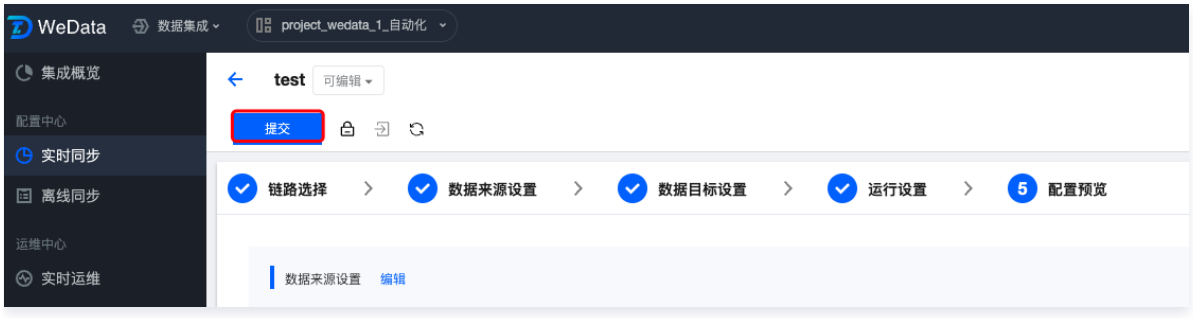
Parameter	Description
Checkpoint Interval	The maximum checkpoint interval for the current task submission.
Maximum Restart Times	Set the maximum restart threshold for failures during execution. If the restart count exceeds this threshold, the task status will be set to **Failed**. The setting range is [−1, 100], where • a value of 0 indicates no restart, • and −1 indicates no limit on maximum restart times.
Parameter	Set task-level runtime parameters. Different source and target ends support different task-level parameters. For details, please refer to Advanced Real-time Node Parameters.

Step 6. Configure Preview

Through configuration preview, you can view the entire results of link selection, data source settings, data target settings, and run settings at once to ensure the accuracy and reasonableness of the task configuration information.

Step 7: Task Configuration Detection and Submission

After completing the above six steps, you can perform task configuration detection and officially submit the task. At this point, you can click the **Submit** button in the upper left corner to proceed to the task configuration check and submission stage.



Steps	Step Description	
Task Configuration Detection	This step involves detection of the read end, write end, and resources in the task: • Detection passed: No configuration errors. • Detection failed: Configuration has issues that need to be fixed for further configuration. • Detection Alert: This is a system-recommended modification item. After modification, click **Retry** to recheck, or you can click **Ignore Exceptions** to proceed without blocking further configuration. The currently supported check items are listed in the table below.	

©2013–2025 Tencent Cloud. All rights reserved.

Submission Policy Selection	In this step, you can choose the submission policy for this task:
	• **First Submission: The first submission supports synchronizing data from the default or specified point.**
	  ○ **Start Now and synchronize from the default point position**: If the source-end configuration is set to "Full + Incremental" mode, it will first synchronize existing data (full phase) by default. After completion, binlog will be consumed to obtain changed data (incremental phase). If the source-end configuration is set to "Incremental Only", it will start reading from the latest binlog point position by default.
	  ○ **Start Now and synchronize from a designated time point**: The task will synchronize data according to the configured time and timezone. If the designated time point cannot be found, the task will start synchronizing from the earliest binlog point position by default. If the source-end configuration is "Full + Incremental", the task will skip the full phase and start synchronizing from the designated time point in the incremental phase by default.
	  ○ **Do not start now:** After submission, the task will not start running for the time being, and can be manually started later in the operation list.
	• **Non-First Submission: Supports starting or continuing tasks with running status.**
	  ○ **Continue running**: Under this strategy, after the new version of the task is submitted, it will continue running from the last sync point.
	  ○ **Restart and synchronize from the specified position:** Under this strategy, you can specify the position to start reading. The task will ignore the old version and restart reading from the specified position. If the specified time position is not found, the task will start synchronizing from the earliest binlog position by default.
	  ○ **Restart and run from the default point position**: Under this strategy, reading will start from the default point position as per the source-end configuration. If the source-end configuration is set to "Full + Incremental", it will first synchronize existing data (full phase) by default. After completion, binlog will be consumed to obtain changed data (incremental phase). If the source-end configuration is set to "Incremental Only", it will start reading from the latest binlog point position by default.
	The supported submit run strategies vary with different task statuses. Details can be found in the subsequent tables.
	
	Each submission will generate a new version of the real-time task. You can configure the version description in the dialog box.
		1. First Submission
		 
		2. Non-First Submission
		 

©2013-2025 Tencent Cloud. All rights reserved.

Submitting the job	After successful submission, you can click **Go to Operations** to view the task's status.	

Task Configuration Detection supports the following check items:

Check Classification	Check Items	Description
Task Configuration Detection	Source Configuration	Check if any required fields are missing on the source side
	Target Configuration	Check if any required fields are missing on the target side
	Mapping Configuration	Check if field mappings are configured
	Resource Group Configuration	Check if the resource group is configured
Data Source Detection	Source end connectivity detection	Check if the data source at the source end is networked with the resource group in the task configuration. If the check fails, view diagnostic information. After resolving network issues, recheck; otherwise, the task is likely to fail.
	Target end connectivity detection	Check if the data source at the target end is networked with the resource group in the task configuration. If the check fails, view diagnostic information. After resolving network issues, recheck; otherwise, the task is likely to fail.
Resource Detection	Resource status detection	Check if the resource group is in an available status. If the resource status is unavailable, please switch the resource group in the task configuration; otherwise, the task is likely to fail.
	Resource Capacity Detection	Check if the remaining resources in the resource group meet the resource requirements of the task configuration. If the check fails, please reduce the task resource configuration or expand the resource group.

Submission and running strategies supported by different task running statuses:

Task Status	Submission and Running Strategies	Description
1. First Submission 2. Stopped/Anomaly Detected/Initialization (Non-first submission)	Start immediately, synchronize from the default locus	Under this policy, reading will commence from the default position according to the configuration of the source. If the source is configured to the "full + incremental" reading mode, it will by default first synchronize the existing data (full phase), and then it can consume binlog to get the changed data (incremental phase); if the source is configured to "incremental only" reading, it will by default start reading from the latest binlog position.
	Start immediately, synchronize from the designated time point	With this strategy, you need to choose a specific start time and match the locus by time. 1. Start reading data from the designated time point. If the designated locus is not matched, the task will start synchronizing from the earliest binlog locus by default

©2013–2025 Tencent Cloud. All rights reserved.

		2. If your source reading mode is 'Full + Incremental', choosing this strategy will skip the full phase and start synchronizing from the designated incremental locus
	Do not start now, manually start the task later in real-time task operation and maintenance	Under this strategy, the task will be submitted to real-time operation and maintenance without starting. The task can be started in bulk from the real-time operation and maintenance page later.
Running (Non-First Submission)	Continue running, retain job status data, continue running from the last synchronized position	Under this strategy, after the new version of the task is submitted, it will continue running from the last synchronized position.
	Restart, continue running from the specified time point	Under this strategy, you can specify the position to start reading. The task will ignore the old version and restart reading from the specified position. If the specified time position is not found, the task will start synchronizing from the earliest binlog position by default.
	Restart, stop the running task and discard the task status, start running from the default position	Under this strategy, the running task will be stopped and the task status will be discarded. The task will then start reading from the default position based on the source-end configuration. If the source-end configuration is "Full + Increment" reading mode, it will synchronize the stock data (Full phase) first, and after completion, it will consume the binlog to get the change data (Increment phase); if the source-end configuration is "Increment Only" reading mode, it will start reading from the latest binlog position by default.
Suspended (Non-First Submission)	Continue running, retain job status data, continue running from the last synchronized position	Under this strategy, after the new version of the task is submitted, it will continue running from the last synchronized position. ⚠ **Note:** • **Pausing** will create a snapshot, and upon resubmission, the task supports resuming from the last checkpoint. • When **forcibly pausing**, no snapshot is created, but upon resubmission, the task can run from the most recent snapshot created during task execution. This type of pause may result in partial data replay. If the target write operation is Append, duplicate data might occur; if the target write operation is Upsert, there won't be a duplication issue.
	Restart, continue running from the specified time point	Under this strategy, you can specify the position to start reading. The task will ignore the old version and restart reading from the specified position. If the specified time position is not found, the task will start synchronizing from the earliest binlog position by default.
	Restart, stop the running task and discard the task status, start running from the default position	Under this strategy, the running task will be stopped and the task status will be discarded. The task will then start reading from the default position based on the source-end configuration. If the source-end configuration is "Full + Increment" reading mode, it will synchronize the stock data (Full phase) first, and after completion, it will consume the binlog to get the change data (Increment phase); if the source-end configuration is "Increment Only" reading mode, it will start reading from the latest binlog position by default.

©2013-2025 Tencent Cloud. All rights reserved.

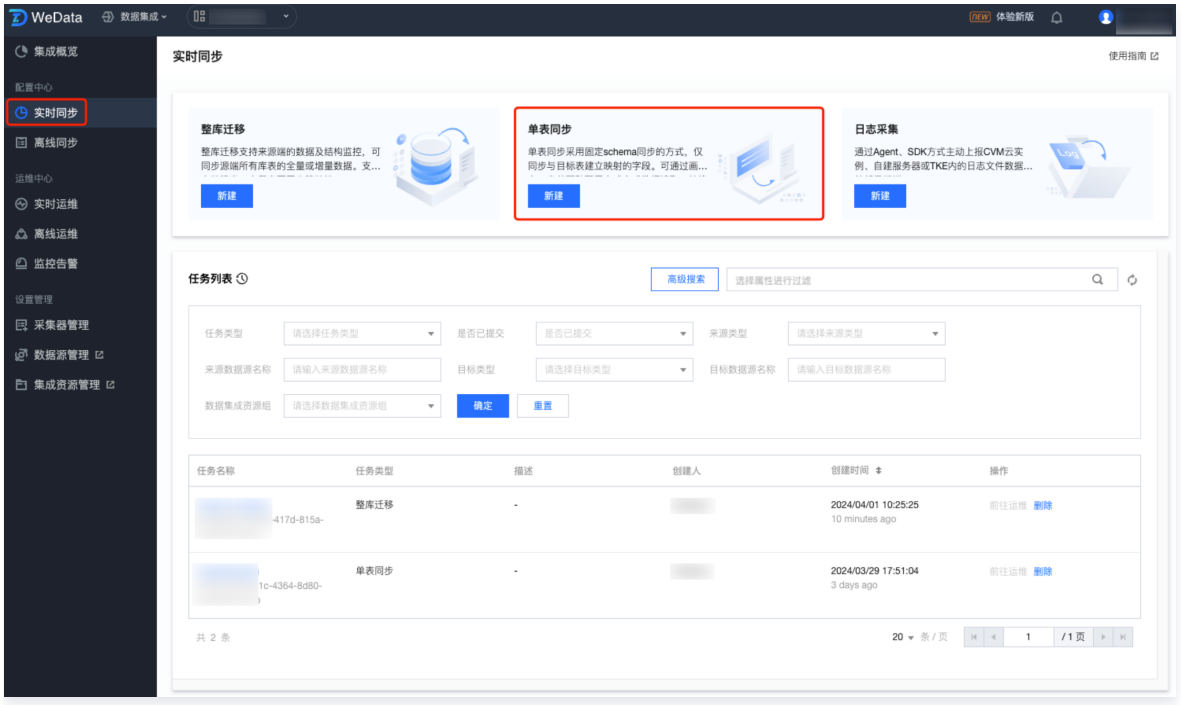
Failed (Non-First Submission)	Resume operation from the last execution failure (checkpoint) location	Under this policy, the task will continue running from the location of the last execution failure
	Restart, and start running from the default location based on the task reading configuration	Under this policy, it will start reading from the default location based on the source configuration. If the source configuration is set to "Full + Incremental" reading mode, it will first synchronize the stored data (Full phase), and upon completion, consume the binlog to acquire change data (Incremental phase); if the source configuration is set to "Incremental only," it will start reading using the latest binlog location by default.
Operating (Non-First Submission)	Not supported	When there is an existing task with the same name and its status is **Operating**, re-submitting the task is not supported

Step 8: Real-time Task Operations and Maintenance

For details, please refer to Real-time operation and maintenance.

6. Real-time single table: MySQL to DLC

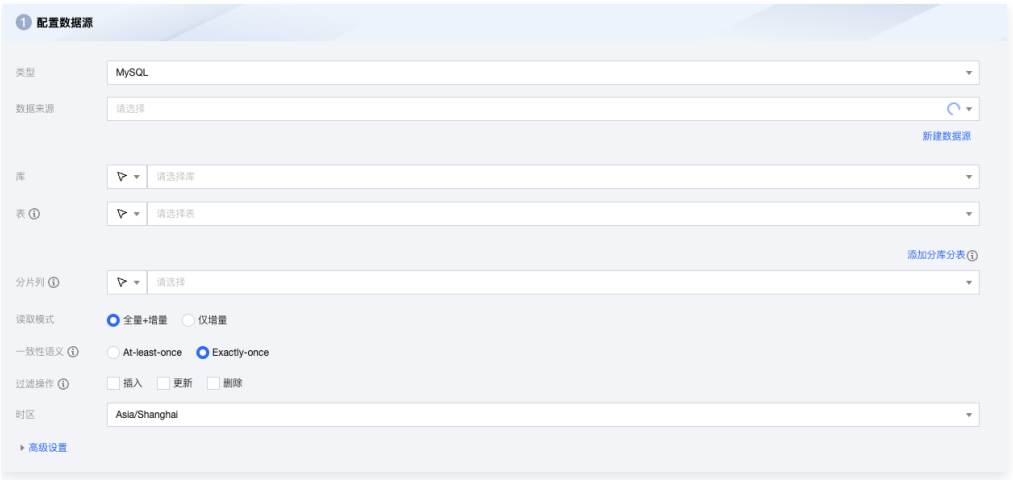
When the business needs to focus only on a specific table, single table synchronization can target the data of that table. This reduces unnecessary data processing and transmission, improving sync efficiency.
On the left sidebar of the DataInLong page, click **Real-time Synchronization**. At the top of the Real-time Synchronization page, select Single Table Synchronization to create a new one and enter the configuration page. When configuring single table synchronization, two modes are supported: **Single Table Mode** and **Canvas Mode**. The configuration process will be explained using the form mode as an example.



Step 1: Configure MySQL Read Node

1. In the DataInLong page, click **Real-time Synchronization** in the left sidebar.
2. On the Real-time Synchronization page, select **Single Table Synchronization** at the top, create a new one (choose Form or Canvas mode), and enter the configuration page.
3. Select the MySQL node and configure the node information.

©2013-2025 Tencent Cloud. All rights reserved.



4. Parameter information:

Data Source	Select the available MySQL data source for the project.
Database	Support selection or manual input of the library name to be read. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Support selecting or manually entering the Table Name to be read. • In the case of table partitioning, you can select or enter multiple table names on the MySQL source end. Multiple tables must have consistent structures. • In the case of table partitioning, supports configuring the table serial number range. For example, 'table_[0-99]' indicates reading 'table_0', 'table_1', 'table_2' up to 'table_99'; if the length of your table's numeric suffix is consistent, such as 'table_000', 'table_001', 'table_002' up to 'table_999', you can configure it as '"table": ["table_00[0-9]", "table_0[10-99]", "table_[100-999]"]'. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Adding Sharding	Suitable for sharding scenarios, click to configure multiple data sources, databases, and table information. In sharding and partitioning scenarios, ensure all table structures are consistent. The task configuration will default to display and use the first table structure for data acquisition.
Shard Column	The shard column is used to divide the table into multiple shards for synchronization. For tables with primary keys, it is recommended to select the table's primary key as the shard column first; For tables without primary keys, it is recommended to choose an indexed column as the shard column and ensure that no data update operations are performed on the shard column; otherwise, only At-Least-Once semantics can be guaranteed.
Read mode	Supports both Full + Incremental and Incremental only modes.
Consistency Semantics	• Exactly-once • At-least-once **Note:** Represents the consistency semantics of the reading end only. In the current version, the statuses of the two modes are not compatible. After a task is submitted, if the mode is changed, restarting with the state is not supported.
Filter operations	After setting, data of the specified operation types will not be synchronized. Supports Insert, Update, and Delete.

©2013-2025 Tencent Cloud. All rights reserved.

Time zone	Set the log time timezone, default is Shanghai.
Advanced Settings (optional)	You can configure parameters based on business requirements.

5. Preview data fields, click **Save**.

Step 2: Configure the DLC Write Node

1. In the DataInLong page, click **Real-time Synchronization** in the left sidebar.
2. On the Real-time Synchronization page, select **Single Table Synchronization** at the top, create a new one (choose Form or Canvas mode), and enter the configuration page.
3. Select the DLC node and configure the node information.



4. Parameter information:

Parameter	Description
Data Source	DLC data source to be written to.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Supports selecting or manually entering the name of the table to be written to. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.

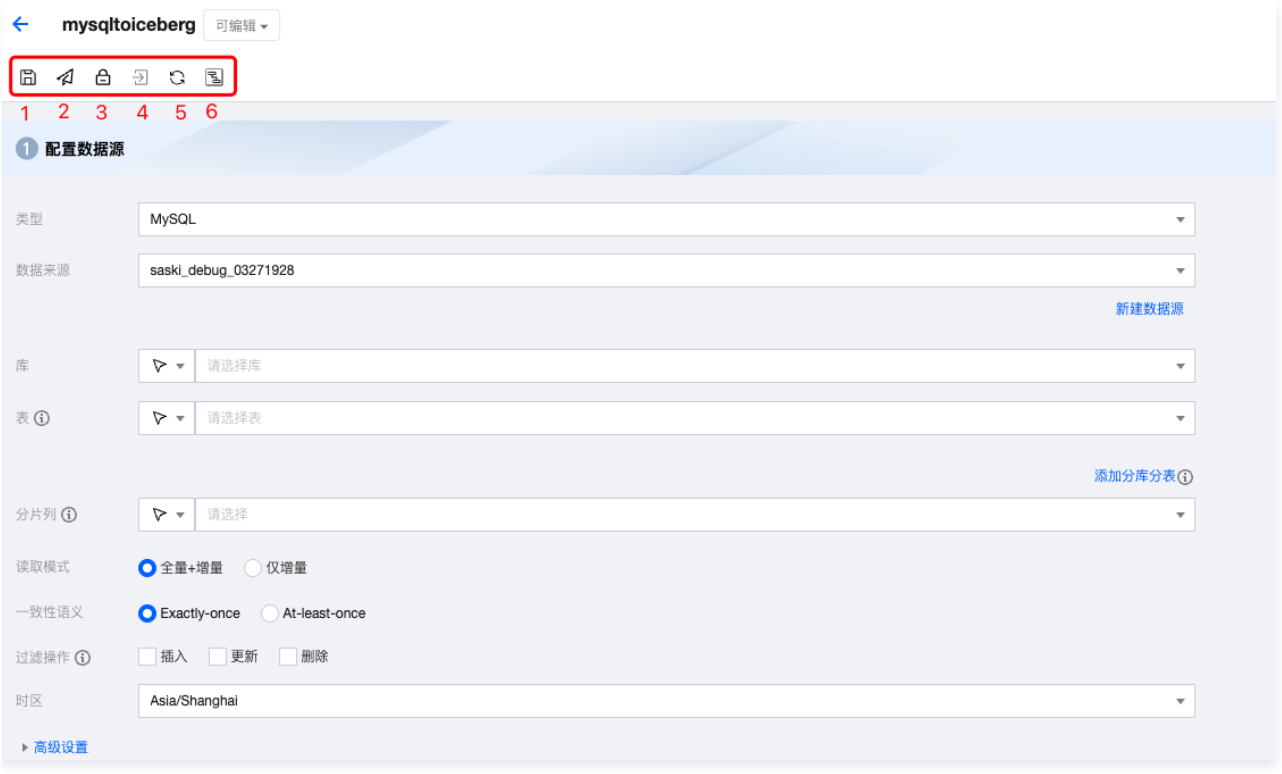
©2013-2025 Tencent Cloud. All rights reserved.

Write mode	DLC real-time synchronization supports three modes: • Upsert: Update write. If there is no primary key conflict, a new row is inserted; if there is a primary key conflict, the data is updated. Suitable for scenarios where the target table has a primary key and needs to be updated in real-time based on source data. There will be some performance overhead. • Append: Append write. Regardless of whether there is a primary key, data is appended by inserting new rows. Primary key conflicts depend on the target. Suitable for scenarios without a primary key and allowing data duplication. No performance overhead. • Full Append + Incremental Upsert: Automatically switches the data writing mode based on the synchronization phase of the source data. The full phase uses Append write to improve performance, and the incremental phase uses Upsert write for real-time data updates. This mode currently supports data sources from MySQL, Oracle, TDSQL MySQL
Unique key	In upsert write mode, you need to set a unique key to ensure data ordering. Multiple selections are supported.
Write Optimization	Write optimization is suitable for scenarios with frequent/real-time writes. Enabling write optimization will automatically merge files and delete expired snapshots. It is highly recommended to enable write optimization in real-time Upsert write scenarios. For more details, please refer to the documentation Data Optimization. **Note** Data Optimization Resources: Data optimization may generate many tasks based on data writing conditions, occupying your cluster resources. To avoid mutual impact with normal business, it is strongly recommended to use separate cluster resources here. Data optimization supports **SuperSQL Engine > SparkSQL**, **SuperSQL Engine > Spark Job**. If lifecycle management is needed, please use **SuperSQL Engine > Spark Job** engine.
Advanced Settings	You can configure parameters based on business requirements.

5. Preview data fields and map them with the fields configured in the read node, click **save**.

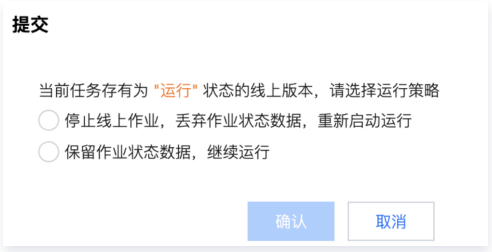
Step 3: Task Configuration Check and Submission

After completing the above steps, you can proceed with a task configuration check and formally submit the task. At this point, you can refer to the related buttons in the upper left corner of the task configuration interface for subsequent submission operations. The position and features of each button are as follows:



©2013–2025 Tencent Cloud. All rights reserved.

Real-time synchronization tasks can configure operation strategies and be submitted to the production environment after configuration is completed. Currently, the task configuration page supports save, submit, lock/unlock, go to operations, and form/canvas conversion operations.

Serial number	Parameter	Description
1	Save	Save current task configuration information, including data node configuration, node connection, and task attribute configuration.
2	Submit	Submit the current task to the production environment. At the time of submission, different operation strategies can be selected based on whether there is an active production task associated with the current task. • If there are no effective online tasks for the current task, i.e., it's the first submission or the online task is in a "failed" status, you can submit it directly. • If there are "Running" or "Paused" online tasks, you need to choose different strategies. Stopping the online job will abandon the previous task's running location, consuming data from the beginning. Retaining the job status will resume from the last consumed location after restarting.  **Note** Clicking to start the task immediately will make it run immediately after submission. Otherwise, it will require manual trigger to start officially.
3	Lock/Unlock	The default creator is the first lock holder, and only the lock holder can edit task configurations and run tasks. If the lock holder does not perform any editing operations within 5 minutes, others can click the icon to acquire the lock. If successful, they can perform editing operations.
4	Go to Operations	Quickly jump to the real-time operations page based on the current task name.
5	Refresh	Refresh task configuration
6	Canvas conversion/Form conversion	Form Mode and Canvas Mode can be converted to each other. The form mode provides a canvas conversion feature, and the canvas mode provides a form conversion feature. **Note** • Currently, conversion is only allowed if the task is in a successfully saved state. • If the canvas mode contains transition nodes, conversion to form mode is not supported.

Task Configuration Detection

When submitting a task, basic environment and configuration checks will be performed. Pay close attention to the failed items and warning items. Failed items are likely to cause task execution failure, while warning items should be judged based on the business situation. The current check results will not block task submission, and you can choose to recheck or ignore the exceptions and continue submitting.

©2013-2025 Tencent Cloud. All rights reserved.



Currently supported inspection items:

Check Classification	Check Items	Description
Task Configuration Detection	Source Configuration	Check if any required fields are missing on the source side
	Target Configuration	Check if any required fields are missing on the target side
	Mapping Configuration	Check if field mappings are configured
	Resource Group Configuration	Check if the resource group is configured
	Same data table detection	Detect whether all real-time single-table tasks under the current project (including submitted and unsubmitted) have the same source table or target table or both the source and target tables are the same. The basis for judging the same table is whether it is a table with the same name under the same data source and the same DB. This inspection item is mainly used for scenarios where multiple tasks are not expected to repeatedly read the same table.
Data Source Detection	Source end connectivity detection	Check if the data source at the source end is networked with the resource group in the task configuration. If the check fails, view diagnostic information. After resolving network issues, recheck; otherwise, the task is likely to fail.
	Target end connectivity detection	Check if the data source at the target end is networked with the resource group in the task configuration. If the check fails, view diagnostic information. After resolving network issues, recheck; otherwise, the task is likely to fail.
Resource Detection	Resource status detection	Check if the resource group is in an available status. If the resource status is unavailable, please switch the resource group in the task configuration; otherwise, the task is likely to fail.
	Resource Capacity Detection	Check if the remaining resources in the resource group meet the resource requirements of the task configuration. If the check fails, please reduce the task resource configuration or expand the resource group.

Submission Policy Selection

©2013–2025 Tencent Cloud. All rights reserved.

After normal submission or ignoring exception submission, you will enter the submission policy selection page, where you can choose different operation strategies based on business needs:



The submission operation strategy supported by different task operation statuses varies:

Task Status	Submission and Running Strategies	Description
1. First Submission 2. Stopped/Anomaly Detected/Initialization (Non-first submission)	Start immediately, synchronize from the default locus	Under this policy, reading will commence from the default position according to the configuration of the source. If the source is configured to the "full + incremental" reading mode, it will by default first synchronize the existing data (full phase), and then it can consume binlog to get the changed data (incremental phase); if the source is configured to "incremental only" reading, it will by default start reading from the latest binlog position.
	Start immediately, synchronize from the designated time point	With this strategy, you need to choose a specific start time and match the locus by time. 1. Start reading data from the specified time point. If the specified position is not matched, the task will default to synchronizing from the earliest binlog position 2. If your source read mode is **Full + Increment**, selecting this strategy will default to skipping the full stage and starting synchronization from the specified increment time position
	Do not start now, manually start the task later in real-time task operation and maintenance	Under this strategy, the task will be submitted to real-time operation and maintenance without starting. The task can be started in bulk from the real-time operation and maintenance page later.
Running (Non-First Submission)	Continue running, retain job status data, continue running from the last synchronized position	Under this strategy, after the new version of the task is submitted, it will continue running from the last synchronized position.
	Restart, continue running from the specified time point	Under this strategy, you can specify the position to start reading. The task will ignore the old version and restart reading from the specified position. If the specified time position is not found, the task will start synchronizing from the earliest binlog position by default.
	Restart, stop the running task and discard the task status, start running from the	Under this strategy, the running task will be stopped and the task status will be discarded, then according to the source

©2013-2025 Tencent Cloud. All rights reserved.

	default position	configuration, it will start reading from the default position. If the source configuration is "**Full + Increment**" read mode, it will default to synchronize existing data first (full stage), and then start consuming binlog to get change data (increment stage). If the source configuration is "Increment only" read mode, it will default to start reading from the latest binlog position.
Suspended (Non-First Submission)	Continue running, retain job status data, continue running from the last synchronized position	Under this strategy, after the new version of the task is submitted, it will continue running from the last synchronized position. **Note:** • **Pausing** will create a snapshot, and upon resubmission, the task supports resuming from the last checkpoint. • When **forcibly pausing**, no snapshot is created, but upon resubmission, the task can run from the most recent snapshot created during task execution. This type of pause may result in partial data replay. If the target write operation is Append, duplicate data might occur; if the target write operation is Upsert, there won't be a duplication issue.
	Restart, continue running from the specified time point	Under this strategy, you can specify the position to start reading. The task will ignore the old version and restart reading from the specified position. If the specified time position is not found, the task will start synchronizing from the earliest binlog position by default.
	Restart, stop the running task and discard the task status, start running from the default position	Under this strategy, the running task will be stopped and the task status will be discarded, then according to the source configuration, it will start reading from the default position. If the source configuration is "**Full + Increment**" read mode, it will default to synchronize existing data first (full stage), and then start consuming binlog to get change data (increment stage). If the source configuration is "Increment only" read mode, it will default to start reading from the latest binlog position.
Failed (Non-First Submission)	Resume operation from the last execution failure (checkpoint) location	Under this policy, the task will continue running from the location of the last execution failure
	Restart, and start running from the default location based on the task reading configuration	Under this policy, reading will commence from the default position according to the configuration of the source. If the source is configured to the "full + incremental" reading mode, it will by default first synchronize the existing data (full phase), and then it can consume binlog to get the changed data (incremental phase); if the source is configured to "incremental only" reading, it will by default start reading from the latest binlog position.
Operating (Non-First Submission)	Not supported	When there is an existing task with the same name and its status is **Operating**, re-submitting the task is not supported

Each submission will generate a new version of the real-time task. You can configure the version description in the dialog box.

Submission Result

After clicking the **Submit** button above, the submission progress and results will appear:

©2013-2025 Tencent Cloud. All rights reserved.



- Task Submitting:
  - Shows the submission progress percentage.
- Task Submission Result – Successful:
  - Show task submission successful result.
  - You can go to Operations for task operation and maintenance management or return to the Task List
- Task submission result – Failed:
  - Show reasons for task submission failure

Step 4: Real-time Task Operation and Maintenance

For details, please refer to Real-time operation and maintenance.

7. Real-time Node Advanced Parameters

When configuring read-write nodes, you can configure advanced settings. The configuration and usage scenarios are as follows:

MySQL Node Level

Below are the advanced settings and their descriptions supported by the MySQL source in usage scenarios:

Read/Write	Applicable Scenario	Configuration:	Description
Read	Single Table + Whole Database	scan.newly-added-table.enabled=true	**Parameter Description:** Set this parameter to sense new tables after **Suspension** > **Continue**. The default is false 1. Use this parameter during full incremental synchronization. New tables will read the stock data first, then read the incremental data 2. Use this parameter during incremental synchronization. New tables will only read incremental data
Read	Single Table + Whole Database	scan.incremental.snapshot.chunk.size=20000	**Parameter Description:** For tasks with uniform data distribution, this parameter represents the approximate number of records in one chunk. You can estimate the number of chunks by dividing the total data volume by the chunk size. The number of chunks affects whether jobmanager experiences an out-of-memory error. Currently, with 2CU, it can support more than 100,000 chunks. If the data volume is too large, you can increase the chunk size to reduce the number of chunks

©2013-2025 Tencent Cloud. All rights reserved.

			Precautions: For large data volume tasks (e.g., total data volume over 100 million, single record larger than 0.1M), it is generally recommended to set 20,000
Read	Single Table + Whole Database	split-key.even-distribution.factor.upper-bound=10.0d	**Parameter Description:** During the MySQL existing data reading phase, if the data is quite scattered and the maximum value of the primary key field is extremely large, you can modify this parameter to use non-uniform splitting. This helps reduce the problem of having too many chunks due to the excessively large primary key value, which could lead to JM OOM issues **Precautions:** Default value is 10.0d, generally no need to modify
Read	Single Table + Whole Database	debezium.query.fetch.size=0	**Parameter Description:** Represents the number of records fetched from the database each time. The default value is 0, which indicates the JDBC default fetch size **Precautions:** 1. For large tasks (e.g., total data volume over 100 million, individual record size over 0.1M) with only one reading instance, it is recommended to set this to 1024 records 2. For tasks with multiple reading instances, it is recommended to lower this value to reduce memory consumption. Setting it to 512 records is advised
Read	Single Table + Whole Database	debezium.max.queue.size=8192	**Parameter Description:** This attribute defines the maximum number of events stored in the internal queue. If this limit is reached, Debezium will pause reading new events until the unprocessed events are handled and committed. This helps prevent an excessive backlog of events in the queue, thus avoiding memory exhaustion and performance degradation. The default is 8192 **Precautions:** 1. For large tasks (e.g., total data volume over 100 million, individual record size over 0.1M) with only one reading instance, it is recommended to set this to 4096 2. For tasks with multiple reading instances, it is recommended to lower this value to reduce memory consumption. Setting it to 1024 is advised

DLC Type Node Level

The following are the advanced settings supported in the DLC target end under usage scenarios and their descriptions:

Read/Write	Applicable Scenario	Configuration:	Description
Write	Single table	write.distribution-mode=none	DLC concurrent writing supports the parameters none|hash|range. Parameter descriptions: 1. none: If a primary key exists, concurrent writing is based on the primary key; otherwise, single concurrent writing. 2. hash: If a partition field exists, concurrent writing is based on the partition field; otherwise, writing follows the strategy of the none parameter. 3. range: Not supported yet; the strategy is consistent with none.

8. FAQs

1. MySQL server ID conflict

Error message:

©2013-2025 Tencent Cloud. All rights reserved.

```
com.github.shyiko.mysql.binlog.network.ServerException: A slave with the same server_uuid/server_id as
this slave has connected to the master.
```

Solution: We have now optimized to randomly generate serverid. If in previous tasks the server-id was explicitly specified in MySQL advanced parameters, it is recommended to delete it, as multiple tasks may use the same data source with the same server-id, causing a conflict.

2. Binlog File Not Found Error:

Error message:

```
Caused by: org.apache.kafka.connect.errors.ConnectException: The connector is trying to read binlog
starting at GTIDs xxx and binlog file 'binlog.xxx', pos=xxx, skipping 4 events plus 1 rows, but this is no
longer available on the server. Reconfigure the connector to use a snapshot when needed.
```

Error Causes:
An error will occur if the binlog file being read by the job has already been cleaned up on the MySQL server. There are several reasons for Binlog cleanup, such as a too-short Binlog retention period, or the job processing speed not keeping up with the Binlog generation speed, exceeding the maximum retention period of MySQL Binlog files. This results in the Binlog files being cleaned up on the MySQL server, causing the Binlog positions being read to become invalid.
Solution: If the job processing speed cannot keep up with the Binlog generation speed, you can consider increasing the Binlog retention period or optimizing the job to reduce backpressure and accelerate source consumption. If there are no anomalies in the job status, it might be that other operations on the database caused the Binlog to be cleaned up, leading to inaccessibility. You need to refer to the MySQL database-side information to determine the reason for Binlog cleanup.

3. MySQL Connection Reset

Error message:

```
EventDataDeserializationException: Failed to deserialize data of EventHeaderV4 .... Caused by:
java.net.SocketException: Connection reset.
```

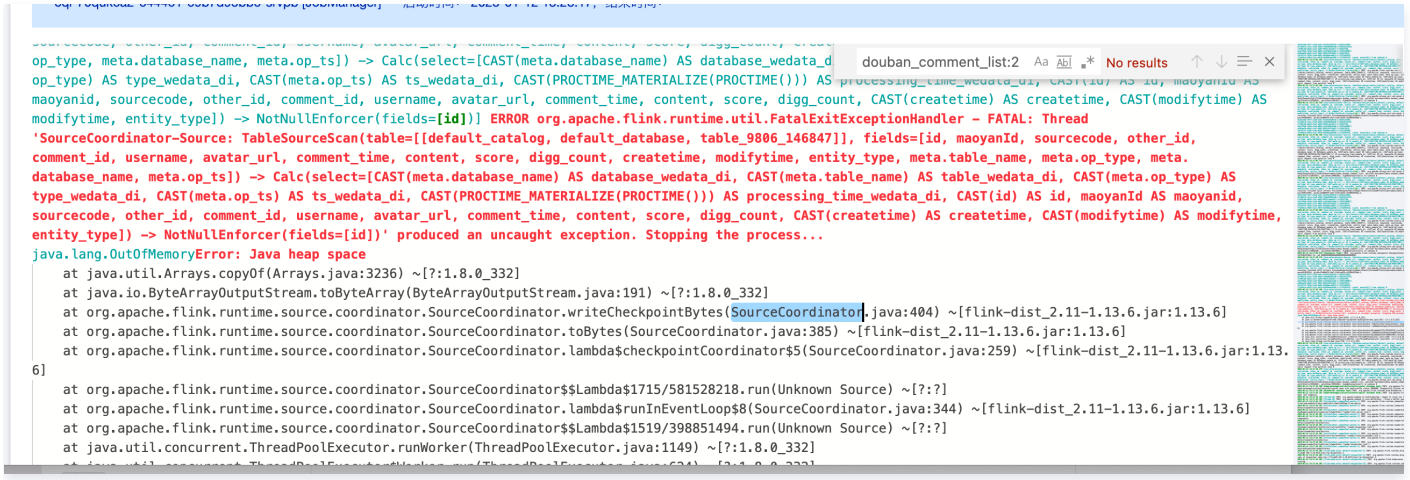
Error Causes:
1. Network Issues.
2. The job has backpressure, causing the source to be unable to read data, and the binlog client to be idle. If the binlog connection remains idle after the timeout, the MySQL server will disconnect the idle connection.

Solution:
1. If it is a network issue, you can increase the MySQL network parameters set global slave_net_timeout = 120; (default 30s) set global thread_pool_idle_timeout = 120.
2. If the backpressure is caused by the job, you can alleviate it by adjusting the job, such as increasing parallelism, improving write speed, and increasing TaskManager memory to reduce GC.

4. Mysql2dlc task JobManager OOM

Error message:

©2013-2025 Tencent Cloud. All rights reserved.



1. If the user data volume is large, you can increase the JobManager CU number and use the MySQL advanced parameter scan.incremental.snapshot.chunk.size to increase the chunk size, which defaults to 8096.
2. The user's data volume is not large, but there is a large difference between the maximum and minimum values of the primary key, resulting in many chunks being divided when using the equal chunk strategy. Modify the distribution factor so that user data follows an uneven data splitting logic, split-key.even-distribution.factor.upper-bound=5.0d. The default distribution factor has been modified to 10.0d.

5. The format of the user's binlog data is incorrect, causing a parsing exception in Debezium

Error message:

```
ERROR io.debezium.connector.mysql.MySqlStreamingChangeEventSource [] - Error during binlog processing.
Last offset stored = null, binlog reader near position = mysql-bin.000044/211839464.
2023-02-20 21:37:28.480 [blc-172.17.48.3:3306] ERROR io.debezium.pipeline.ErrorHandler [] - Producer
failure
io.debezium.DebeziumException: Error processing binlog event.
```

Solution:
After modifying binlog_row_image=full, you need to restart the database.

6. Is gh-ost supported?

Yes, it does not migrate the temporary table data generated by Online DDL changes. It only migrates the original DDL data executed using gh-ost in the source database. You can use the default or configure your own regular expressions for gh-ost shadow tables and unused tables.

7. Error when synchronizing incremental data (changelog data) to DLC

Error message:

©2013-2025 Tencent Cloud. All rights reserved.

```
2022-07-12 21:03:57.161 [IcebergFilesCommitter -> Sink: IcebergSink HYBRIS.xny_sl03_pro.subs_crm_clue (1/1)#34] WARN  org.apache.flink.runtime.taskmanager.Task [] -
IcebergFilesCommitter -> Sink: IcebergSink HYBRIS.xny_sl03_pro.subs_crm_clue (1/1)#34 (ced                              b5) switched from INITIALIZING to FAILED with failure cause:
java.lang.IllegalArgumentException: Cannot write delete files in a v1 table
    at org.apache.iceberg.ManifestFiles.writeDeleteManifest(ManifestFiles.java:154)
    at org.apache.iceberg.SnapshotProducer.newDeleteManifestWriter(SnapshotProducer.java:374)
    at org.apache.iceberg.MergingSnapshotProducer.lambda$newDeleteFilesAsManifests$10(MergingSnapshotProducer.java:681)
    at java.util.HashMap.forEach(HashMap.java:1289)
    at org.apache.iceberg.MergingSnapshotProducer.newDeleteFilesAsManifests(MergingSnapshotProducer.java:678)
    at org.apache.iceberg.MergingSnapshotProducer.prepareDeleteManifests(MergingSnapshotProducer.java:664)
    at org.apache.iceberg.MergingSnapshotProducer.apply(MergingSnapshotProducer.java:533)
    at org.apache.iceberg.SnapshotProducer.apply(SnapshotProducer.java:164)
    at org.apache.iceberg.SnapshotProducer.lambda$commit$2(SnapshotProducer.java:283)
    at org.apache.iceberg.util.Tasks$Builder.runTaskWithRetry(Tasks.java:404)
    at org.apache.iceberg.util.Tasks$Builder.runSingleThreaded(Tasks.java:214)
    at org.apache.iceberg.util.Tasks$Builder.run(Tasks.java:198)
    at org.apache.iceberg.util.Tasks$Builder.run(Tasks.java:190)
    at org.apache.iceberg.SnapshotProducer.commit(SnapshotProducer.java:282)
    at org.apache.iceberg.flink.sink.IcebergFilesCommitter.commitOperation(IcebergFilesCommitter.java:312)
    at org.apache.iceberg.flink.sink.IcebergFilesCommitter.commitDeltaTxn(IcebergFilesCommitter.java:299)
    at org.apache.iceberg.flink.sink.IcebergFilesCommitter.commitUpToCheckpoint(IcebergFilesCommitter.java:218)
    at org.apache.iceberg.flink.sink.IcebergFilesCommitter.initializeState(IcebergFilesCommitter.java:153)
    at org.apache.flink.streaming.api.operators.StreamOperatorStateHandler.initializeOperatorState(StreamOperatorStateHandler.java:118)
    at org.apache.flink.streaming.api.operators.AbstractStreamOperator.initializeState(AbstractStreamOperator.java:290)
    at org.apache.flink.streaming.runtime.tasks.OperatorChain.initializeStateAndOpenOperators(OperatorChain.java:441)
    at org.apache.flink.streaming.runtime.tasks.StreamTask.restoreGates(StreamTask.java:585)
    at org.apache.flink.streaming.runtime.tasks.StreamTaskActionExecutor$1.call(StreamTaskActionExecutor.java:55)
```

Cause:
The DLC table is a V1 table, which does not support changelog data.
Solution:
1. Modify the DLC table to V2 and enable upsert support.
2. To enable upsert support: ALTER TABLE tblname SET TBLPROPERTIES ('write.upsert.enabled'='true').
3. To modify the table to V2: ALTER TABLE tblname SET TBLPROPERTIES ('format-version'='2').

Check if the table properties are set successfully: SHOW TBLPROPERTIES tblname.

8. Cannot write incompatible dataset to table with schema

Error Details:

```
Caused by: java.lang.IllegalArgumentException: Cannot write incompatible dataset to table with schema:
* mobile should be required, but is optional
    at org.apache.iceberg.types.TypeUtil.checkSchemaCompatibility(TypeUtil.java:364)
    at org.apache.iceberg.types.TypeUtil.validateWriteSchema(TypeUtil.java:323)
```

Error Causes:
When creating a dlc table, the user set a NOT NULL constraint on the field.
Resolution:
Do not set a NOT NULL constraint when creating the table.

9. Synchronizing mysql2dlc reports an array out of bounds

Problem Details:

```
java.lang.ArrayIndexOutOfBoundsException: 1
    at
org.apache.flink.table.data.binary.BinarySegmentUtils.getLongMultiSegments(BinarySegmentUtils.java:736) ~
[flink-table-blink_2.11-1.13.6.jar:1.13.6]
    at org.apache.flink.table.data.binary.BinarySegmentUtils.getLong(BinarySegmentUtils.java:726) ~[flink-
table-blink_2.11-1.13.6.jar:1.13.6]
    at
org.apache.flink.table.data.binary.BinarySegmentUtils.readTimestampData(BinarySegmentUtils.java:1022) ~
[flink-table-blink_2.11-1.13.6.jar:1.13.6]
    at org.apache.flink.table.data.binary.BinaryRowData.getTimestamp(BinaryRowData.java:356) ~[flink-
table-blink_2.11-1.13.6.jar:1.13.6]
    at org.apache.flink.table.data.RowData.lambda$createFieldGetter$39385f9c$1(RowData.java:260) ~[flink-
table-blink_2.11-1.13.6.jar:1.13.6]
```

Error Causes:
The issue caused by using the time field as the primary key.

©2013-2025 Tencent Cloud. All rights reserved.

Solution:

1. Do not use the time field as the primary key.
2. The user still needs the Time Field to ensure Uniqueness. It is recommended that the user add a Redundant String Field in the DLC, use Function Transformation on the upstream Time Field, and map it to the redundant field.

10. DLC task reports it is not an Iceberg Table

Error message:

```
[flink-dist_2.11-1.13.6.jar:1.13.6]
    ... 10 more
Caused by: org.apache.iceberg.exceptions.NoSuchIcebergTableException: Not an iceberg table: HYBRIS.origindb.
movieunifiedorder01_unified_order_sharding__maoyan_order_v2 (type=null)
    at org.apache.iceberg.exceptions.NoSuchIcebergTableException.check(NoSuchIcebergTableException.java:33) ~
[pool-1-thread-1-1678868648050-prod-sort-connector-iceberg-dlc-2.6.29-1-1.4.0.jar:1.4.0]
```

Error Causes:

1. The user can run statements in dlc to check the type of the table.
2. Statement: desc formatted table name.
3. The table type can be viewed.

Resolution:

Choose the correct engine to create an Iceberg type table.

11. flink sql field order does not match the field order of the DLC target table, causing an error

Solution:

Do not modify the field order of the task table.

©2013–2025 Tencent Cloud. All rights reserved.

Guide for Real-time Synchronization from MySQL to TChouse-D

Last updated: 2024-09-05 17:45:27

This document will guide you step-by-step through the process of real-time synchronization from MySQL to TChouse-D, including whole database migration and single table synchronization. The general process is as follows:
1. Data Source Compatibility.
2. MySQL Environment Preparation.
3. TChouse-D Environment Preparation.
4. Project Configuration and Resource Group Preparation.
5. Real-time Whole Database: Steps for Configuring MySQL Synchronization to TChouse-D.
6. Real-time Single Table: Steps for Configuring MySQL Synchronization to TChouse-D.
7. Real-time Node Advanced Parameters.
8. FAQs.

1. Data Source Compatibility

MySQL Data Source

If you want to use MySQL for real-time data synchronization, you need to first confirm the MySQL data source version support, usage limitations, and supported data type conversion.

Supported Versions

Currently, DataInLong supports real-time reading at both single table and database levels for MySQL. To use real-time reading capabilities, you must follow these version restrictions:

Type	Version	Driver
MySQL	5.6,5.7,8.0.x	JDBC Driver:8.0.21
RDS MySQL	5.6,5.7, 8.0.x	
PolarDB MySQL	5.6,5.7,8.0.x	
Aurora MySQL	5.6,5.7,8.0.x	
MariaDB	10.x	
PolarDB X	2.0.1	

Use Limits

- Binlog logging needs to be enabled, and only MySQL server Binlog configuration format ROW is supported for synchronization.
- Tables without a primary key may have duplicate data because exact once cannot be guaranteed, so real-time synchronization tasks must ensure a primary key exists.
- XA ROLLBACK is not supported. Real-time synchronization tasks will not rollback data for XA PREPARE. To handle XA ROLLBACK scenarios, you need to manually remove the XA ROLLBACK tables from real-time synchronization tasks and then re-sync the tables after adding them back.
- Set MySQL Session Timeout:

When creating an initial consistent snapshot for a large database, the connection you establish may time out while reading tables. You can prevent this by configuring interactive_timeout and wait_timeout in the MySQL configuration file.

  - interactive_timeout: The number of seconds the server waits for activity on an interactive connection before closing it. Please refer to MySQL :: MySQL 8.0 Reference Manual :: 7.1.8 Server System Variables.
  - wait_timeout: The number of seconds the server waits for activity on a non-interactive connection before closing it. Please refer to MySQL :: MySQL 8.0 Reference Manual :: 7.1.8 Server System Variables.

MySQL Data Type Conversion

©2013-2025 Tencent Cloud. All rights reserved.

The supported data types and internal mapping fields for reading MySQL data are shown below. (When processing MySQL, the data types from the MySQL data source will be mapped to the data processing engine data types first):

Field Type	Supported	Internal mapping fields	Remarks
TINYINT	Yes	TINYINT	TINYINT(1) maps to BOOLEAN Options need to be added to support TINYINT(1) mapping to bool or tinyint
SMALLINT	Yes	SMALLINT	–
TINYINT_UNSIGNED	Yes	SMALLINT	–
TINYINT_UNSIGNED_ZEROFILL	Yes	SMALLINT	–
INT	Yes	INT	–
INTEGER	Yes	INT	–
YEAR	Yes	INT	–
MEDIUMINT	Yes	INT	–
SMALLINT_UNSIGNED	Yes	INT	–
SMALLINT_UNSIGNED_ZEROFILL	Yes	INT	–
BIGINT	Yes	LONG	–
INT_UNSIGNED	Yes	LONG	–
MEDIUMINT_UNSIGNED	Yes	LONG	–
MEDIUMINT_UNSIGNED_ZEROFILL	Yes	LONG	–
INT_UNSIGNED_ZEROFILL	Yes	LONG	–
BIGINT_UNSIGNED	Yes	DECIMAL	DECIMAL(20,0)
BIGINT_UNSIGNED_ZEROFILL	Yes	DECIMAL	DECIMAL(20,0)
SERIAL	Yes	DECIMAL	DECIMAL(20,0)
FLOAT	Yes	FLOAT	–
FLOAT_UNSIGNED	Yes	FLOAT	–
FLOAT_UNSIGNED_ZEROFILL	Yes	FLOAT	–
DOUBLE	Yes	DOUBLE	–
DOUBLE_UNSIGNED	Yes	DOUBLE	–
DOUBLE_UNSIGNED_ZEROFILL	Yes	DOUBLE	–
DOUBLE_PRECISION	Yes	DOUBLE	–
DOUBLE_PRECISION_UNSIGNED	Yes	DOUBLE	–
ZEROFILL	Yes	DOUBLE	–
REAL	Yes	DOUBLE	–
REAL_UNSIGNED	Yes	DOUBLE	–

©2013–2025 Tencent Cloud. All rights reserved.

REAL_UNSIGNED_ZEROFILL	Yes	DOUBLE	–
NUMERIC	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
NUMERIC_UNSIGNED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
NUMERIC_UNSIGNED_ZEROFILL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
DECIMAL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
DECIMAL_UNSIGNED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
DECIMAL_UNSIGNED_ZEROFILL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
FIXED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
FIXED_UNSIGNED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
FIXED_UNSIGNED_ZEROFILL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
BOOLEAN	Yes	BOOLEAN	–
DATE	Yes	DATE	–
TIME	Yes	TIME	–
DATETIME	Yes	TIMESTAMP	–
TIMESTAMP	Yes	TIMESTAMP	–
CHAR	Yes	STRING	–
JSON	Yes	STRING	–
BIT	Yes	STRING	BIT(1) maps to BOOLEAN
VARCHAR	Yes	STRING	–
TEXT	Yes	STRING	–
BLOB	Yes	STRING	–
TINYBLOB	Yes	STRING	–
TINYTEXT	Yes	STRING	–
MEDIUMBLOB	Yes	STRING	–
MEDIUMTEXT	Yes	STRING	–
LONGBLOB	Yes	STRING	–
LONGTEXT	Yes	STRING	–
VARBINARY	Yes	STRING	–
GEOMETRY	Yes	STRING	–
POINT	Yes	STRING	–

©2013–2025 Tencent Cloud. All rights reserved.

LINESTRING	Yes	STRING	-
POLYGON	Yes	STRING	-
MULTIPOINT	Yes	STRING	-
MULTILINESTRING	Yes	STRING	-
MULTIPOLYGON	Yes	STRING	-
GEOMETRYCOLLECTION	Yes	STRING	-
ENUM	Yes	STRING	-
BINARY	Yes	BINARY	BINARY(1)
SET	No		-

Doris/TChouse-D Environment Preparation and Database Configuration

Supported Versions

Currently, DataInLong supports real-time writing for both single tables and whole databases in Doris. To use real-time synchronization capabilities, you must follow these version restrictions:

Type	Version
Doris	0.15,1.x

Use Limits

- Doris supports three data models: [DUPLICATE KEY|UNIQUE KEY|AGGREGATE KEY]. To write in Upsert mode in Doris, ensure the data model is UNIQUE KEY. For details, see Data Model - Apache Doris.
- If the source table has a primary key, the automatically created Doris target table will be UNIQUE KEY; if the source table has no primary key, the automatically created Doris target table will be DUPLICATE KEY.

2. MySQL Environment Preparation

Confirm MySQL Version

DataInLong has requirements for MySQL versions. Check whether the MySQL version to be synced meets these requirements. You can check the current MySQL database version with the following statement.

```
select version();
```

Setting MySQL Server Permissions

You must define a MySQL user with appropriate permissions for all databases monitored by the Debezium MySQL connector.

1. **Create a MySQL User (Optional)**

```
mysql> CREATE USER 'user'@'localhost' IDENTIFIED BY 'password';
```

2. **Grant Necessary Permissions to the User:**

For real-time data synchronization, the account must have SELECT, REPLICATION SLAVE, and REPLICATION CLIENT permissions on the database. Refer to the following command for execution:

```
mysql> GRANT SELECT, SHOW DATABASES, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO 'user' IDENTIFIED BY
'password';
```

©2013-2025 Tencent Cloud. All rights reserved.

> **Note**
> Enabling scan.incremental.snapshot.enabled no longer requires the RELOAD permission (enabled by default).

3. **Refresh the User's Permissions:**

```
mysql>FLUSH PRIVILEGES;
```

See more about permission descriptions.

Enable MySQL Binlog

1. **Check if Binlog is Enabled**

```
show variables like "log_bin"
```

   - If the result is ON, it means Binlog is enabled. If it's a slave, use the following statement:

```
show variables like "log_slave_updates";
```

   - If the return value is ON, Binlog is enabled. If Binlog is already enabled, you can skip the following steps.

2. **Enable Binlog**

If Binlog is confirmed not to be enabled, the following operations are required.

- For Tencent Cloud MySQL / TDSQL-C MySQL, binlog is enabled by default.
- For open-source MySQL, refer to the official documentation to enable binlog.

3. **Change Binlog Format to Row**

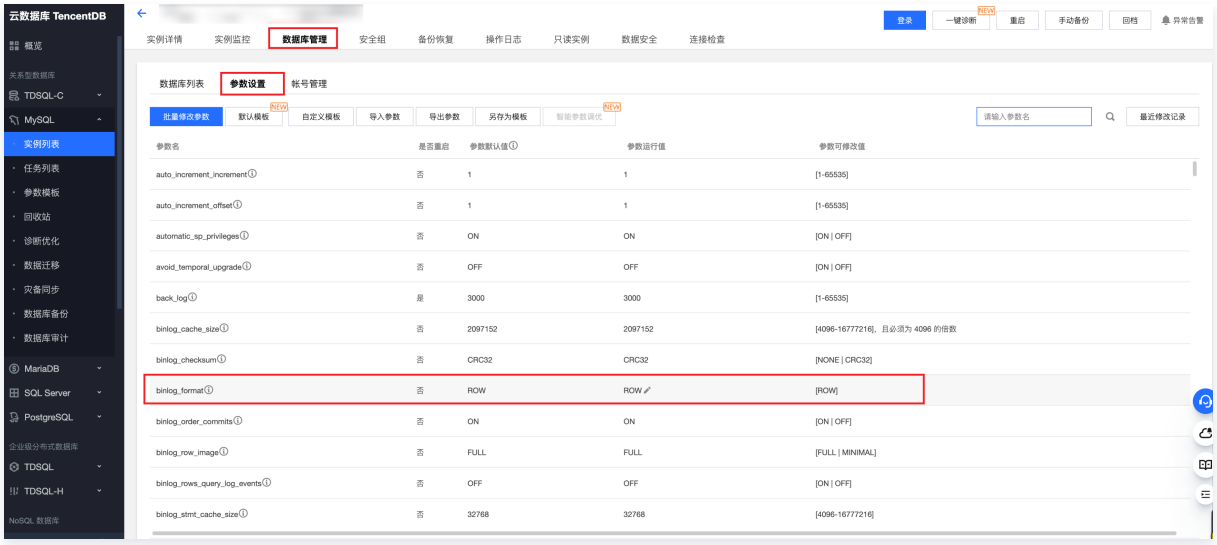
3.1 Real-time synchronization only supports MySQL server Binlog configuration format as ROW. Use the following statement to check the Binlog format.

```
show variables like "binlog_format";
```

3.2 If the return is non-ROW, please change the Binlog Format.

   - For open-source MySQL, refer to the official documentation: MySQL :: MySQL 8.0 Reference Manual :: 19.1.6.4 Binary Logging Options and Variables.
   - For Tencent Cloud MySQL / TDSQL-C MySQL:
     - Log in to the Tencent Cloud MySQL / TDSQL-C MySQL console, find the instance to enable Binlog, and click to enter the detailed information page of the instance.
     - In the tabs above, select DMC and find the Parameter Settings tab.
     - In the Parameter Settings tab, find the binlog_format parameter and set it to "ROW".

©2013-2025 Tencent Cloud. All rights reserved.



4. **binlog_row_image**

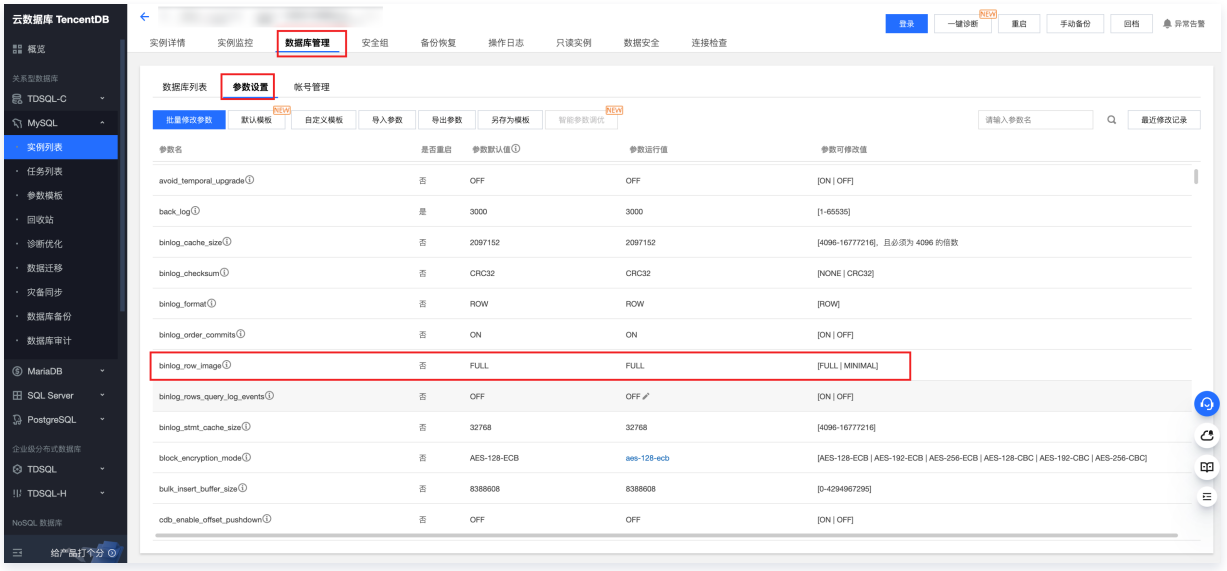
Real-time synchronization only supports MySQL server binlog_row_image configuration format as FULL or full.

4.1 Use the following statement to query the usage format of binlog_row_image.

```
show variables like "binlog_row_image";
```

4.2 If the return is not FULL/full, please modify binlog_row_image

- For open-source MySQL, refer to the official documentation: MySQL :: MySQL 8.0 Reference Manual :: 19.1.6.4 Binary Logging Options and Variables .
- For Tencent Cloud MySQL / TDSQL-C MySQL:
  - Log in to the Tencent Cloud MySQL / TDSQL-C MySQL console, find the instance to enable Binlog, and click to enter the detailed information page of the instance.
  - In the tabs above, select DMC and find the Parameter Settings tab.
  - In the parameter settings tab, find the binlog_row_image parameter and set it to "FULL".



Enable GTIDs (optional)

GTID (Global Transaction Identifier), used to uniquely identify a transaction in binlog, can avoid data confusion or master-slave inconsistency caused by duplicate transaction execution.

1. **Check if GTID is enabled**

©2013-2025 Tencent Cloud. All rights reserved.

```
show global variables like '%GTID%';
```

If the return result is similar to the following, it proves that GTID is enabled.

```
+--------------------------+-------+
| Variable_name                      | Value |
+--------------------------+-------+
| enforce_gtid_consistency   | ON     |
| gtid_mode                                | ON     |
+--------------------------+-------+
```

2. **Enable GTID**
   - For open-source MySQL, refer to the official documentation MySQL :: MySQL 8.0 Reference Manual :: 19.1.4.2 Enabling GTID Transactions Online.
   - For Tencent Cloud MySQL / TDSQL-C MySQL instances, GTID is enabled by default and not supported to be disabled.

Add Tencent Cloud MySQL Database Security Group

A security group is a stateful virtual firewall with filtering features used to set network access control for one or more cloud databases. It is an important means of network security isolation provided by Tencent Cloud. If you use a Tencent Cloud MySQL database above the basic edition, you need to add the access IPs below to the target database's security group. If a security group is not configured, the connectivity test may fail when configuring the MySQL data source in WeData. For detailed steps, refer to Managing Cloud Database Security Groups.

```
118.89.220.0/24, 139.199.116.0/24, 140.143.68.0/24, 152.136.131.0/24, 81.70.150.0/24, 81.70.161.0/24,
81.70.195.0/24, 81.70.198.0/24, 82.156.22.0/24, 82.156.221.0/24, 82.156.23.0/24, 82.156.24.0/24,
82.156.27.0/24, 82.156.82.0/24, 82.156.84.0/24, 82.157.119.0/24
```

3. TChouse-D Environment Preparation

Database Environment Configuration

1. Check the network connectivity between Doris/TChouse-D Database and DataInLong Resource Group.

Refer to the connectivity test in **Data Source Configuration** to ensure the DataInLong resource group has normal access to the Doris/TChouse-D database.

2. Check the data model of the table to be written into Doris/TChouse-D.

Doris/TChouse-D writing supports two modes: Append/Upsert. To write in Upsert mode, the Doris/TChouse-D data model must be Unique key. You can execute the following SQL in the SQL client:

```
show create table example_db.table_hash;
// The result after executing the above SQL
CREATE TABLE example_db.table_hash
(
    k1 BIGINT,
    k2 LARGEINT,
    v1 VARCHAR(2048),
    v2 SMALLINT DEFAULT "10"
)
UNIQUE KEY(k1, k2)
DISTRIBUTED BY HASH (k1, k2)
// If it's a Unique key model, the keyword: UNIQUE KEY will appear in the results
```

3. Create Doris/TChouse-D Table.
   - Create a detailed model table.

©2013-2025 Tencent Cloud. All rights reserved.

```
CREATE TABLE example_db.table_hash
(
    k1 TINYINT,
    k2 DECIMAL(10, 2) DEFAULT "10.5",
    k3 CHAR(10) COMMENT "string column",
    k4 INT NOT NULL DEFAULT "1" COMMENT "int column"
)
COMMENT "my first table"
DISTRIBUTED BY HASH(k1)
```

- Create a primary key unique model table.

```
CREATE TABLE example_db.table_hash
(
    k1 BIGINT,
    k2 LARGEINT,
    v1 VARCHAR(2048),
    v2 SMALLINT DEFAULT "10"
)
UNIQUE KEY(k1, k2)
DISTRIBUTED BY HASH (k1, k2)
```

For more details, see: CREATE-TABLE - Apache Doris

4. Project Configuration and Resource Group Preparation

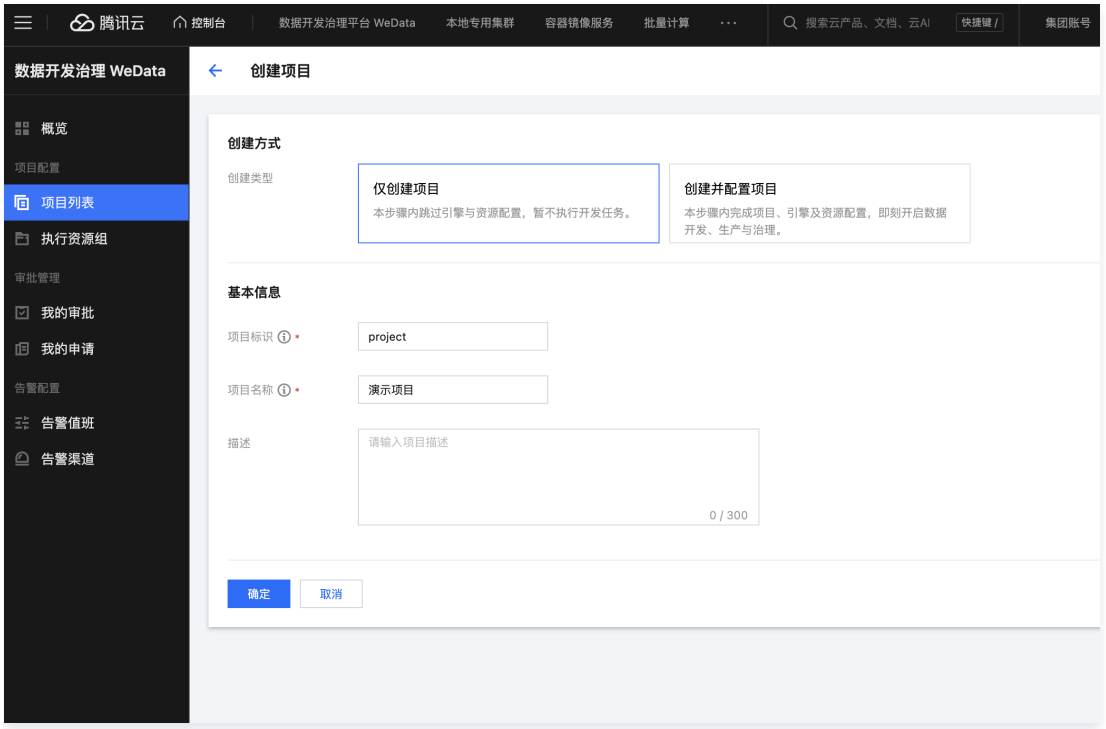
Before configuring the synchronization task, a project needs to be created. The following steps are based on a new environment. If you have already created project members or completed project creation, you can skip this step.

Project Creation

If you haven't created a project space in WeData, you need to complete the project creation before performing data synchronization tasks.

1. The main account holder logs into the CAM console to add a sub-account. Refer to Preparing a CAM Sub-account. Once created, the member can use the set sub-account to log in to the WeData big data platform.
2. To create a project space, go to the WeData homepage, click **Project List** > **Create Project**. If you are only doing DataInLong, you can choose **Create Project Only** and enter the project identifier and name.

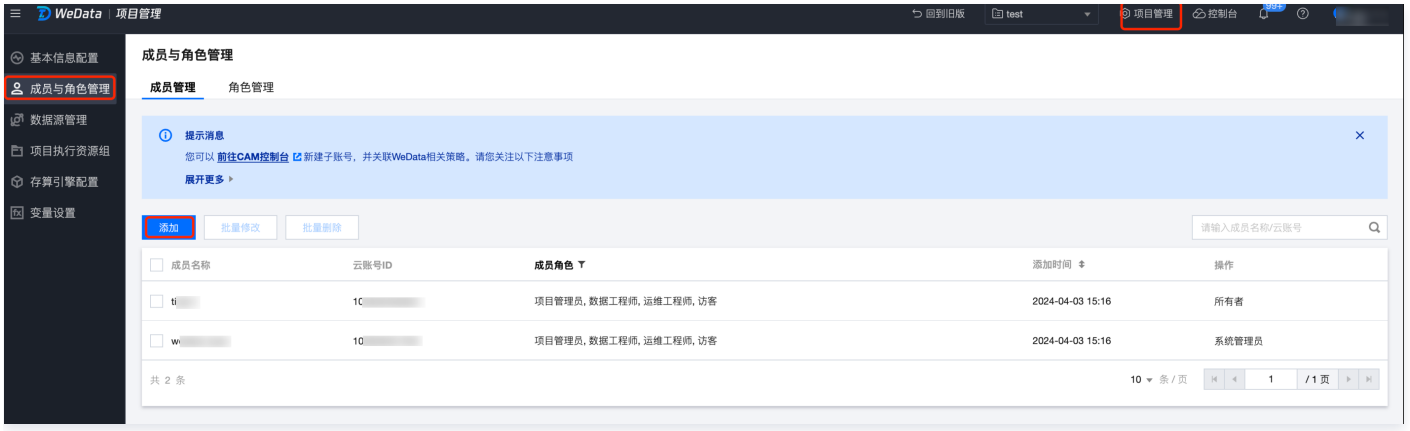
©2013-2025 Tencent Cloud. All rights reserved.



Add Member

Only the project owner and main account have permission to enter the newly created project. Other sub-accounts need to be added by the project owner or the main account in the project management module. For details, please see Project Management.

1. Click **Project Management** > **Member and Role Management** to enter the Member and Role Management interface. Click the **Add** button.



2. In the Add Member interface, set the required roles for the newly created member. Once set, the member can enter the project.



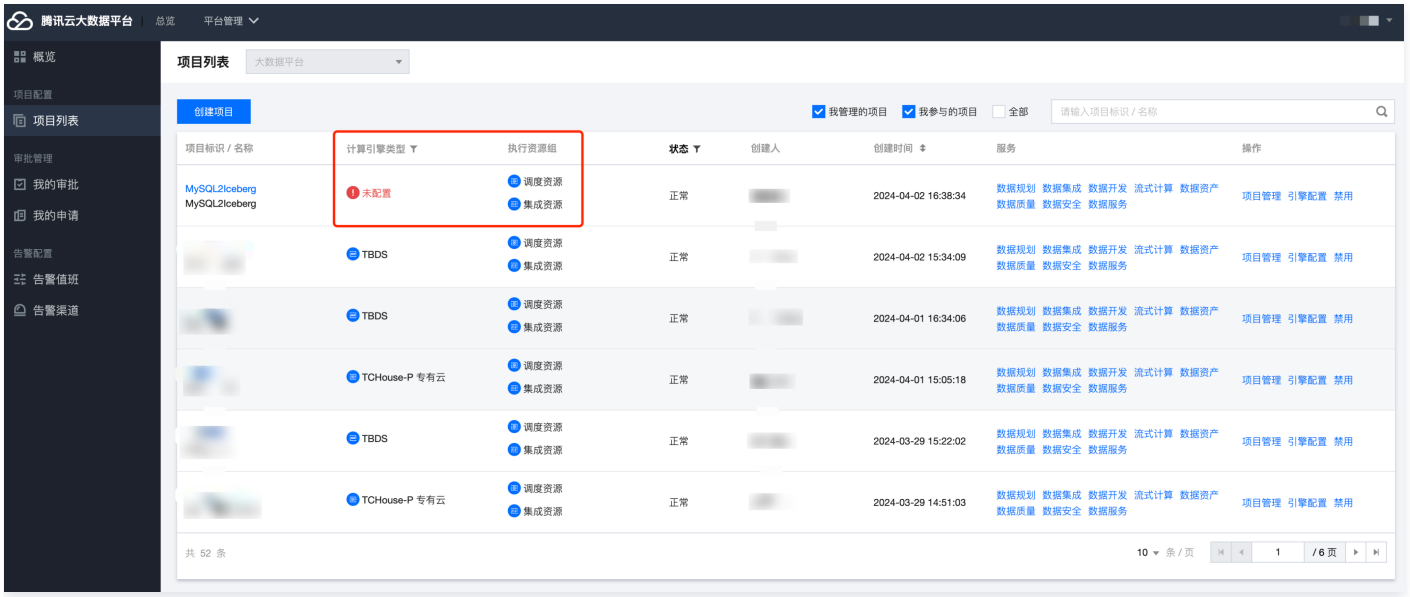
©2013–2025 Tencent Cloud. All rights reserved.

Resource Group Preparation

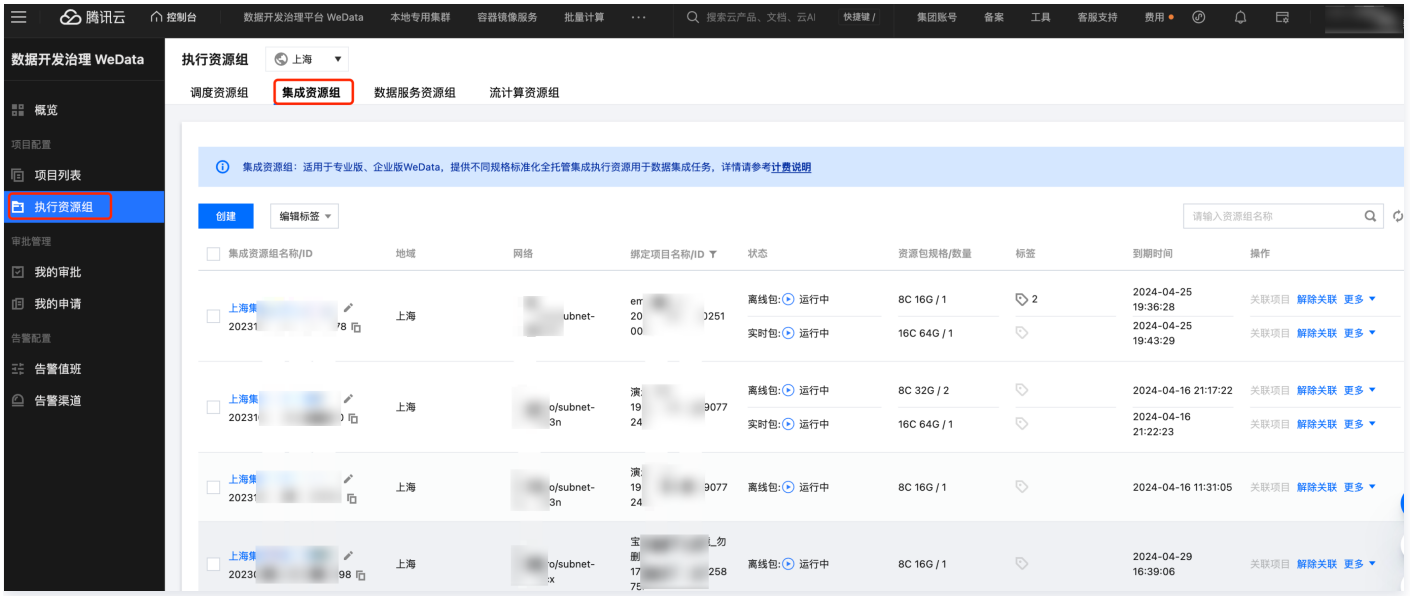
Confirm whether the Resource Group is configured

Before officially starting the DataInLong task, please confirm whether your project is equipped with **Integration Resources**.

- If it prompts that the integration resource is not configured, the DataInLong task cannot run normally.
- When you finish creating a new project, the column for **Execution Resource Group** will show that the integration resource is not configured. You just need to go to the WeData homepage > **Execution Resource Group** > **Integration Resource Group** and associate resources for the new project.



Enter the Integration Resource Group interface and click **Associate Project** after the corresponding Resource Group.



Purchase Integration Resource Group

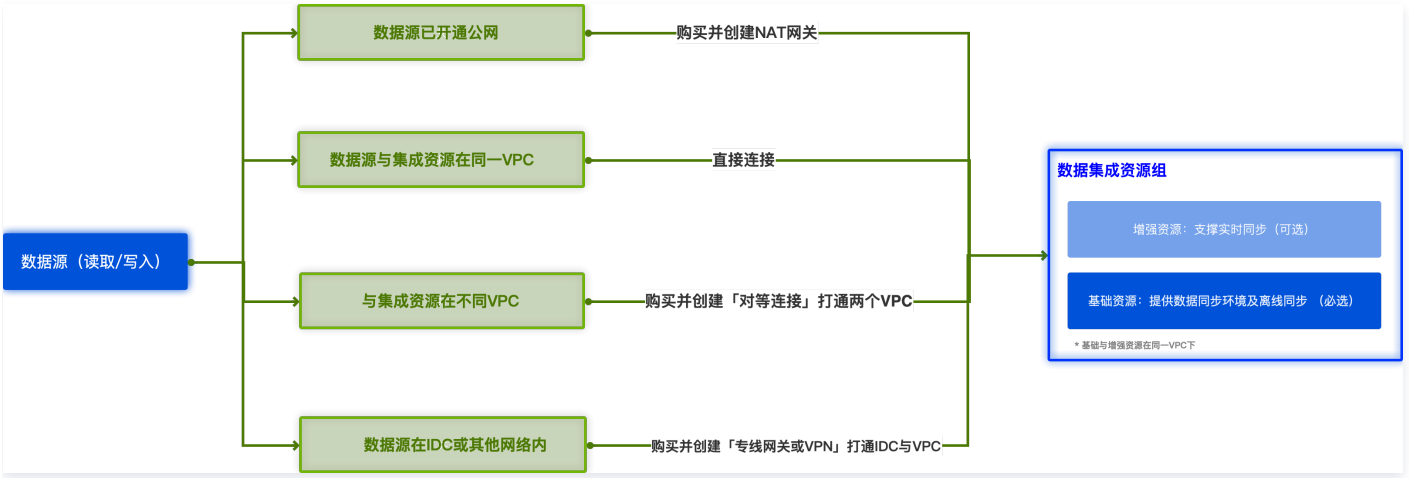
If you find that there is currently no integration resource group, you need to go for purchase. The Integration Resource Group is dedicated to using computing resources when running DataInLong tasks. These resources are mainly presented in the form of resource groups. Before configuring the synchronization task, you need to confirm whether the DataInLong Resource Group is purchased. For details, please refer to Configure the Integration Resource Group.

Confirm Network Connectivity

Ensure network connectivity between the data source network (including the read endpoint and write endpoint) and the DataInLong resource group. Resources should not be denied access due to allowlist restrictions, as this will prevent data

©2013–2025 Tencent Cloud. All rights reserved.

transmission and synchronization.
The machine resources contained in the DataInLong Resource Group must be in the same VPC network environment by default:



- **If the data source has public network access**: You need to purchase and create a NAT Gateway to allow the integration resources to connect to the data source VPC through the gateway. For detailed operations, please refer to Resource Group Configuration for Public Network.
- **If the data source is within a VPC:**
  - If the data source is in the same VPC as the integration resource: You can use it directly.
  - If the data source is in a different VPC from the integration resource: Purchase a Peering Connection to interconnect the VPCs of the integration resource and the data source.
- **If the data source is located in an IDC or other traditional network environment**: You need to purchase a VPN or Direct Connect Gateway to interconnect the VPCs of the integration resource and the data source.

> **Note**
> For documentation on the configuration and management of the Integrated Resource Group, refer to Integrated Resource Configuration and Management.

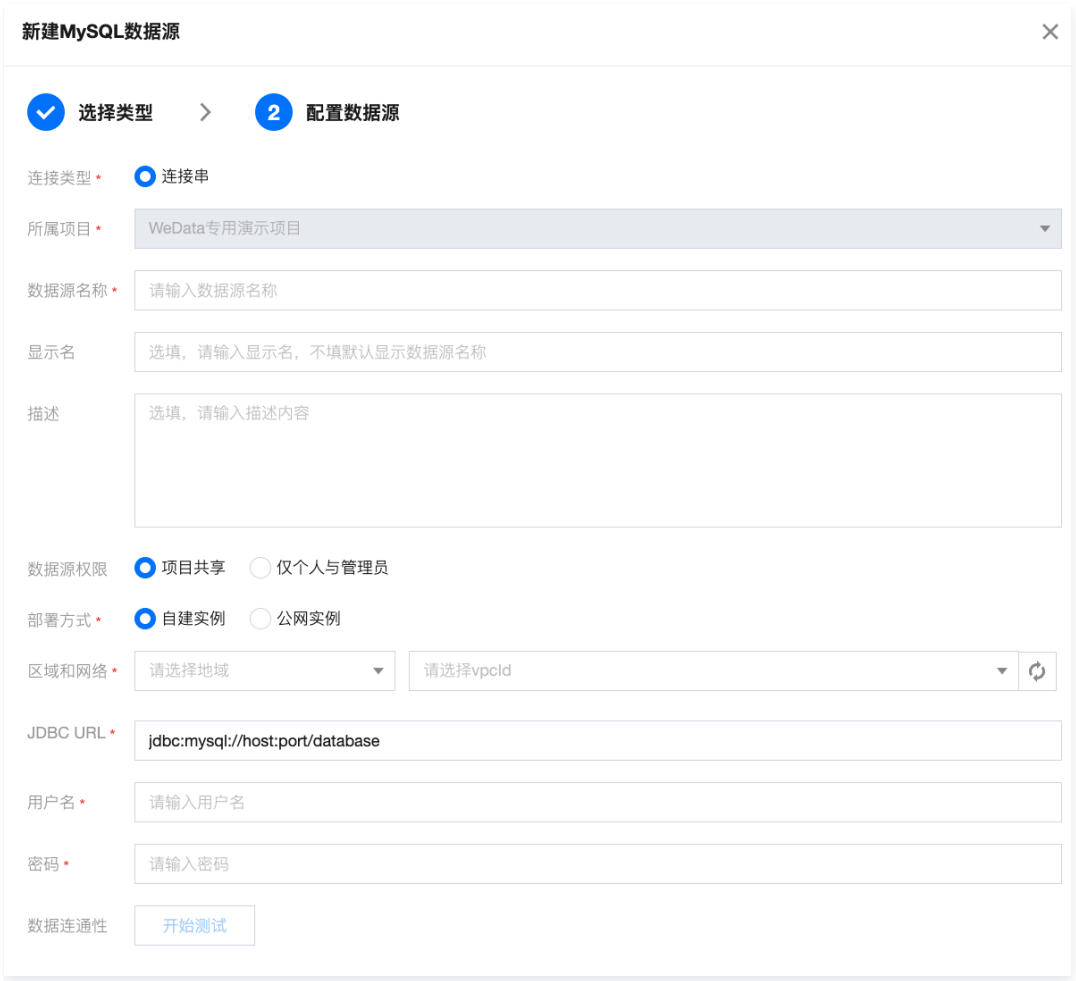
Data source configuration

MySQL Data Source Configuration

Enter the data source configuration interface, MySQL data sources support two connection methods: cloud instance and connection string.
click **Project Management** > **Data Source Management** > **New Data Source** > **Choose MySQL Data Source**.

- Create a data source via connection string.

©2013–2025 Tencent Cloud. All rights reserved.



The parameters are described as follows:

Parameter	Description
Data source name	The name of the new data source, defined by the user, cannot be empty. Naming must start with a letter, can contain letters, numbers, and underscores. Length within 20 characters.
Description	Optional, description of this data source.
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it.
Deployment mode	Supports two deployment modes: self-built instances and public network instances. Self-built instances are data source instances deployed on Tencent CVM, while public network instances are those in the customer's local IDC or other cloud resources, accessible via the public network.
Region and Network	When choosing a self-built instance, you need to select the region and vpcID where the data source instance is located.
JDBC URL	The connection string information used to connect to MySQL data source instances, including host IP, port, database name, etc.
Database name	Name of the database to connect to.
Username	Username for connecting to the database.
Password	Password for connecting to the database.
Data Connectivity (Legacy)	Test whether the configured database is connected. **Note**

©2013-2025 Tencent Cloud. All rights reserved.

- If the connectivity test fails, the data source can still be saved. However, it cannot be used if the connectivity test fails and the data source is saved.
- If the connectivity test fails, it may be because WeData is blocked by the database network firewall. You need to add Tencent Cloud MySQL Database Security Group. Refer to Chapter 2 Add Tencent Cloud MySQL Database Security Group.
- After the Chinese New Year in 2024, newly purchased WeData does not include this feature by default. This feature has been upgraded to Resource Group Connectivity.

Resource Group Connectivity (New)	Support the connectivity test between specific resource groups and data sources. Users only need to ensure that their data source and the VPC network where the Integrated Resource Group is located are connected without additional network interconnection.



Note
Users who purchased WeData before the Chinese New Year in 2024 do not have this feature. If you need to use it, please contact Tencent's operation and maintenance personnel.

TChouse-D Data Source Configuration

WeData currently supports introducing TChouse-D data sources through cloud instances.
click **Project Management > Data Source Management > New Data Source > Select TChouse-D Data Source.**

©2013-2025 Tencent Cloud. All rights reserved.



The parameters are described as follows:

Parameter	Description
Associated item	Project name to which the data source belongs.
Data source name	The name of the new data source, defined by the user, cannot be empty. Naming must start with a letter, can contain letters, numbers, and underscores. Length within 20 characters.
Display name	Optional: Display name of the data source. If not filled, the data source name will be used by default.
Description	Optional, description of this data source.
Data source permissions	Project sharing means that all members of the project can use the data source. Only personal and administrators mean that only the creator and project administrators can use it.
Get instance	First select the region, then choose the EMR instance under that region. Click the refresh button on the right to retrieve instances in that region again
Database name	Name of the database to connect to.
Username	Username for connecting to the database.
Password	Password for connecting to the database.
Data Connectivity (Legacy)	Test whether the configured database is connected. **Note** • If the connectivity test fails, the data source can still be saved. However, it cannot be used if the connectivity test fails and the data source is saved. • If the connectivity test fails, it may be because WeData is blocked by the database network firewall. You need to add Tencent Cloud MySQL Database Security Group. Refer to Chapter 2 Add Tencent Cloud MySQL Database Security Group.

©2013–2025 Tencent Cloud. All rights reserved.

	• After the Chinese New Year in 2024, newly purchased WeData does not include this feature by default. This feature has been upgraded to Resource Group Connectivity.
Resource Group Connectivity (New)	Support the connectivity test between specific resource groups and data sources. Users only need to ensure that their data source and the VPC network where the Integrated Resource Group is located are connected without additional network interconnection.  **Note** Users who purchased WeData before the Chinese New Year in 2024 do not have this feature. If you need to use it, please contact Tencent's operation and maintenance personnel.

5. Real-time Whole Database: MySQL Synchronization to TChouse-D

Step 1: Create a whole database migration task

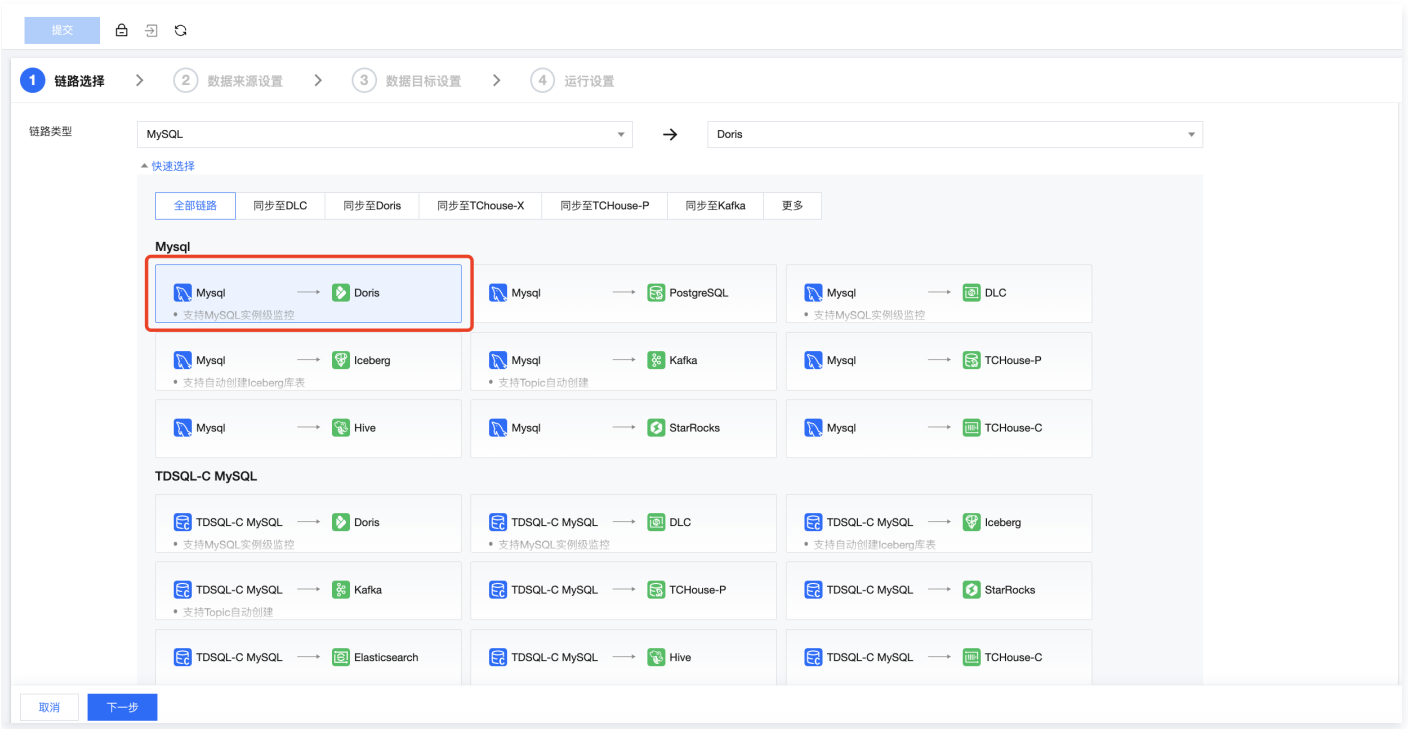
Log in to WeData, open the DataInLong module, then click **Configuration Center** > **Real-time Sync Task** page, and click **New** whole database migration task.



Step 2: Link Selection

In the link selection interface, select **MySQL** > **Doris Link**, after selection, click **Next**.

©2013-2025 Tencent Cloud. All rights reserved.

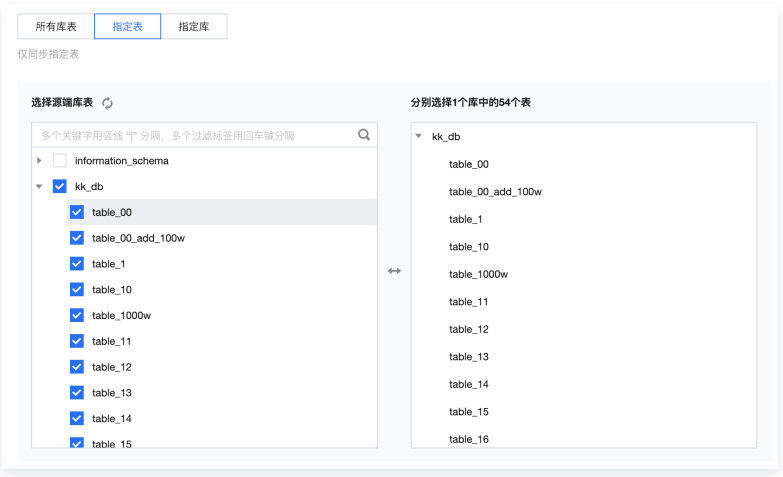


Step 3: Data Source Settings



Parameter	Description
Data Source	Select the MySQL data source for synchronization.
Source Table	• All databases and tables: Monitor all databases under the data source. During the task execution, new databases and tables will be synchronized to the target end by default. • Specified Table: Under this option, you need to specify the specific table name. After setting, the task will only synchronize the specified table; if a new table needs to be synchronized, the task must be stopped and restarted.

©2013–2025 Tencent Cloud. All rights reserved.



- Specify Database: Under this option, you need to specify the specific database name with a table name regular expression. After setting, new tables that match the table name expression will be synchronized to the target end by default during task execution.



Common table name regular expression:

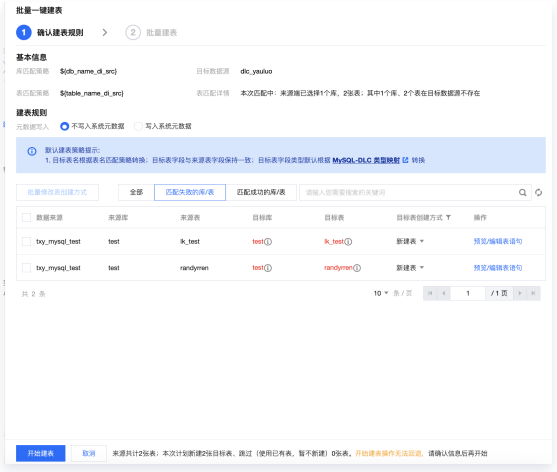
- All tables in a single database: .*
- Specific tables in a single database: table1|table2|table3
- Exclude specific tables: ^(?!(exclude1|exclude2|exclude3)$).*$
- Exclude specific tables but include all tables starting with 'a': ^(?!(exclude1|exclude2|exclude3)$)(^a.*)$

Read mode	• Full + Incremental: Data synchronization is divided into full and incremental synchronization phases. After the full phase is completed, the task enters the incremental phase. The full phase will synchronize historical data in the database, and the incremental phase will start synchronizing from the binlog CDC location after the task starts. • Increment only: Only sync data starting from the binlog cdc point after the task starts.
Filter operations	Support three operations: Insert, Update, and Delete. After setting, data of the specified operation types will not be synchronized.
Time zone	Set the log time timezone, default is Shanghai.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Step 4: Data Target Settings

©2013–2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Flow Direction	Select the target data source for synchronization.
Database/Table Matching Strategy	Name Matching Rules for databases and data table objects in Doris.
Batch Table Creation	Doris target end supports one-click batch table creation based on database and table matching strategies: • Confirm Key Table Rules: Confirm key table information like data source and source database, supporting preview/edit table statements.  ○ Basic Information: Displays selected or filled matching strategies for database, table, target data source, and table matching details. ○ Table Creation Rules: ○ Metadata Writing: You can choose whether to write metadata. The supported metadata from MySQL to Doris includes: ○ database_wedata_di (manages parameter database, representing the database name) ○ table_wedata_di (manages parameter table, representing the data table name) ○ type_wedata_di (manages parameter type, representing the operation type) ○ ts_wedata_di (manages parameter ts, representing the log time) ○ processing_time_wedata_di (processing time, local time of the machine processing the current record) ○ Unmatched Database/Table: Displays databases and tables in the target data source that could not be matched according to the database table matching rules

©2013–2025 Tencent Cloud. All rights reserved.

- Matched Database/Table: Displays databases and tables in the target data source that have been matched according to the database table matching rules
- Preview/Edit Table Statement: You can preview automatically generated table statements and directly edit them



- Batch Modify Table Creation Method: Supports batch modification of table creation methods
  - Table Creation Method for Unmatched Tables:
    - Create New Table: Supports changing to Do Not Create New
    - Do Not Create New: Supports changing to Create New Table
  - Table Creation Method for Matched Tables
    - Use Existing Table: Supports changing to Delete Existing Table and Create New
    - Delete Existing Table and Create New: Supports changing to Use Existing Table



- Batch Key Table: Display key table results, supporting viewing key table statements and retrying.

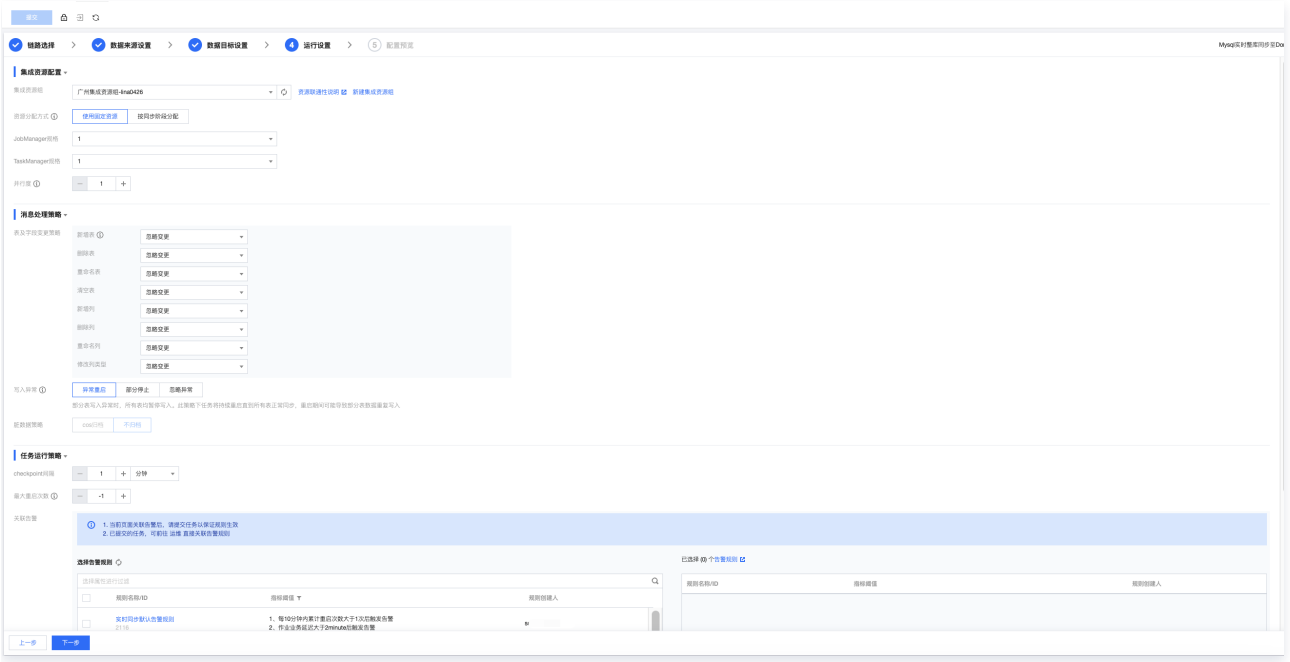


Batch Create Table Field Mapping: See MySQL to Doris Table Field Type Conversion

©2013–2025 Tencent Cloud. All rights reserved.

Advanced Settings	You can configure parameters based on business requirements.

Step 5: Run Settings



Integrated Resource Configuration:

- Associate the current task with the corresponding integrated resource group, and set the runtime JM, TM specifications and task parallelism. The actual CU number occupied by the current task during runtime is calculated as = JobManager specification + TaskManager specification × parallelism.
- Resource Allocation Method: Integration resources support multiple allocation methods
  - Fixed Allocation: In this method, resource allocation remains constant throughout the full and incremental synchronization phases, preventing resource preemption among tasks. It is suitable for scenarios where data may significantly change during task execution.
  - Allocation by synchronization phase: Allocates planned resource usage based on different synchronization phases (full and incremental) to save overall resource usage.
- Table and Field Change Strategy:

Table and Field Change Type	Support Policies
Add Table	Ignore Change, Automatic Table Creation, Log Alerts, Task Error
Deleting a table	Ignore Change
Rename Table	Ignore Change, Log Alerts
Clearing table data	Ignore Change, Log Alerts
Add Column	Automatically Add Columns, Ignore Change, Log Alerts, Task Error
Delete Column	Automatic Change, Ignore Change, Log Alerts, Task Error
Rename Column	Automatic Change, Ignore Change, Log Alerts, Task Error
Modify Column Type	Automatic Change, Ignore Change, Log Alerts, Task Error

The following are the policies and display effects supported in the table above:

Policy name	Display Effect

©2013–2025 Tencent Cloud. All rights reserved.

Ignore Change	1. No change to target table/field 2. Task is running normally
Logging and Alarming	1. No change to target table/field 2. The task runs normally, changes are ignored by default
Stop the task	1. No change to target table/field 2. Abnormal Task Restart
Automatic Processing	The target end automatically changes according to the source end and automatically creates target objects. For example, adding tables, adding columns, etc. **Note** To avoid structural changes due to misoperation that could impact user business, only related changes such as adding tables/columns are currently supported for automatic processing; other high-risk operations like deleting columns, clearing tables, and dropping tables are only supported by "reminder-style" response policies (e.g., alarms, stop)

Write Failure Handling Strategy:
DLC Whole Database Synchronization Task provides task-level runtime resources and data failure handling strategies. The data write failure handling strategy supports four types:

Policy name	Policy Description
Partial Stop	When a table write exception occurs, only the data writing to that table will be stopped; other tables will continue to synchronize normally. The stopped table cannot resume writing during this task's operation.
Abnormal Restart	When a table write exception occurs, all tables will pause writing. The task will continuously restart under this policy until all tables synchronize normally. During the restart period, some table data may be redundantly written.
Ignore Exception	Ignore the data that cannot be written within the table and mark it as dirty data. Other data in the table, as well as other tables in the task, will synchronize normally. Dirty data can be archived to COS or not archived. • COS Archiving: Archives dirty data that cannot be written. Requires configuration of COS data source, bucket, storage directory, content separator, and line break. • Non-Archiving: No additional operations required.

Scenario Example:
In Task Task1, 50 tables are planned to be synchronized. During the task operation, Table A has new fields added or field type changes:
- Partial Stop: After running the task for Table A, the field "DEMO" underwent a type change, and the new field type cannot match the target end type for writing. Under this policy, the task will stop Data Reading from the source table A, and subsequent tasks will only synchronize the remaining 49 tables to the target end.
- Abnormal Restart: After running the task for Table A, the field "DEMO" underwent a type change, and the new field type cannot match the target end type for writing. Under this policy, the task will continuously restart, and all 50 tables configured in the task will pause data write until the field in Table A is corrected.
- Ignore Abnormal: After running the task for Table A, the field "DEMO" underwent a type change, and the new field type cannot match the target end type for writing. Under this policy, the task will ignore the abnormal data that cannot be written, mark it as Dirty Data, and synchronize the other data in the table normally.

Step 6. Configure Preview

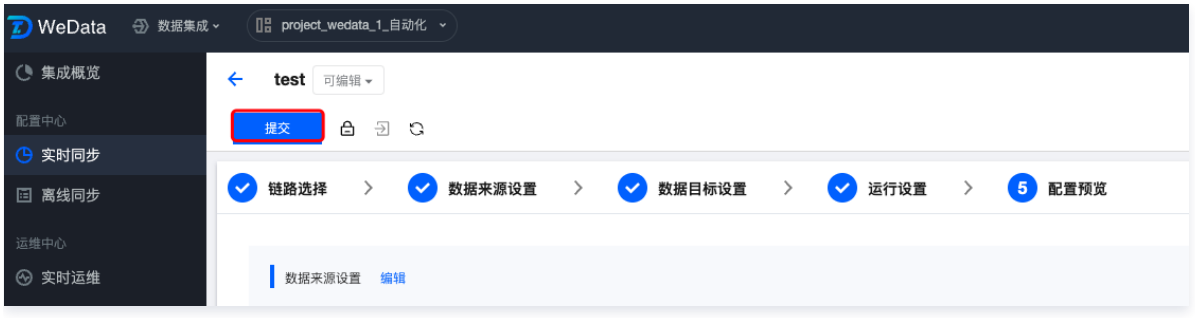
Through configuration preview, you can view all results of link selection, data source settings, data target settings, and run settings at once to ensure the accuracy and rationality of task configuration information.

©2013-2025 Tencent Cloud. All rights reserved.



Step 7: Task Configuration Detection and Submission

After completing the above six steps, you can perform task configuration detection and officially submit the task. At this point, you can click the **Submit** button in the upper left corner to proceed to the task configuration check and submission stage.



Steps	Step Description	

©2013–2025 Tencent Cloud. All rights reserved.

Task Configuration Detection	This step involves detection of the read end, write end, and resources in the task: • Detection passed: No configuration errors. • Detection failed: Configuration has issues that need to be fixed for further configuration. • Detection Alert: This is a system-recommended modification item. After modification, click **Retry** to recheck, or you can click **Ignore Exceptions** to proceed without blocking further configuration. The supported inspection items are shown in the subsequent table.	

Submission Policy Selection	In this step, you can choose the submission policy for this task:
	• **First Submission: The first submission supports synchronizing data from the default or specified point.**
	  ○ **Start Now, synchronize from the default point:** If the source configuration is "Full + Incremental" reading mode, it will first synchronize the existing data (full phase), and after completion, consume the binlog to obtain change data (incremental phase). If the source configuration is "Incremental Only" reading mode, it will read from the latest binlog point by default.
	  ○ **Start Now, synchronize from the designated time point:** The task will synchronize data according to the configured time and time zone. If the designated time point is not found, the task will start from the earliest binlog point by default. If the source reading mode is "Full + Incremental", the task will skip the full phase and start synchronizing from the designated incremental time point by default.
	  ○ **Do not start now:** After submission, the task will not start running for the time being, and can be manually started later in the operation list.
	• **Non-First Submission: Supports starting or continuing tasks with running status.**
	  ○ **Continue running**: Under this strategy, after the new version of the task is submitted, it will continue running from the last sync point.
	  ○ **Restart and synchronize from the specified position:** Under this strategy, you can specify the position to start reading. The task will ignore the old version and restart reading from the specified position. If the specified time position is not found, the task will start synchronizing from the earliest binlog position by default.
	  ○ **Restart, run from the default point:** Under this strategy, reading will start from the default point based on the source configuration. If the source configuration is "Full + Incremental" reading mode, it will first synchronize the existing data (full phase) and then consume the binlog to obtain change data (incremental phase). If the source configuration is "Incremental Only" reading mode, it will read from the latest binlog point by default.
	Different task statuses support different submit run strategies, as detailed in the subsequent table.
		1. First Submission
		 
		2. Non-First Submission
		 

©2013-2025 Tencent Cloud. All rights reserved.

	Each submission will generate a new version of the real-time task. You can configure the version description in the dialog box.	
Submitting the job	After successful submission, you can click **Go to Operations** to view the task's status.	

Supported inspection items for task configuration detection:

Detection Category	Inspection Item	Description
Task Configuration Detection	Source Configuration	Check if any required fields are missing on the source end
	Target Configuration	Check if any required fields are missing on the target end
	Mapping Configuration	Check if field mapping has been configured
	Resource Group Configuration	Check if the resource group is configured
Data Source Detection	Source end connectivity detection	Check if the source data source is network-connected with the task configuration's resource group. If detection fails, check the diagnostic information. Recheck after network connectivity is established, otherwise the task is likely to fail.
	Target end connectivity detection	Check if the target data source is network-connected with the task configuration's resource group. If detection fails, check the diagnostic information. Recheck after network connectivity is established, otherwise the task is likely to fail.
Resource Detection	Resource Status Detection	Check if the resource group is in an available status. If the resource status is unavailable, please change the resource group in the task configuration, otherwise the task is likely to fail.
	Resource Availability Detection	Check if the current remaining resources of the resource group meet the resource requirements of the task configuration. If the detection fails, please reduce the task resource configuration or expand the resource group.

Different submission and operation strategies supported by various task running statuses:

Task Status	Submission and Operation Strategies	Description
1. First Submission 2. Stopped/Detect Anomalies/Initialization (Non-First Submission)	Start immediately, synchronize from the default locus	Under this strategy, data will be read from the default locus based on the source configuration. If the source configuration is set to "Full + Increment" mode, the default is to first synchronize the stock data (full phase), and then consume binlog to get the changed data (incremental phase) after completion. If the source configuration is set to "Increment Only" mode, the default is to start reading from the latest binlog position.
	Start immediately, synchronize from a designated time point	Under this strategy, a specific start time needs to be selected, and synchronization will be based on time matching points.

©2013-2025 Tencent Cloud. All rights reserved.

		1. Read data from the designated time point. If the specified point is not matched, the task will default to synchronize from the earliest binlog point 2. If your source reading method is Full + Increment, selecting this strategy will default to skip the full phase and start synchronization from the designated incremental time point
	Do not start now, go to Real-time Task Operation and Maintenance later to manually start the task	Under this strategy, the task is only submitted to Real-time Operation and Maintenance without starting, and tasks can be started in batches from the Real-time Operation and Maintenance page later.
Running (Non-First Submission)	Continue running, retain job status data, and continue running from the last sync point	Under this strategy, after the new version of the task is submitted, it will continue running from the last sync point.
	Restart, continue running from the designated time point	Under this strategy, you can specify the position to start reading. The task will ignore the old version and restart reading from the specified point. If the specified time position is not found, the task will start synchronizing from the earliest binlog position by default.
	Restart, stop the running task and discard its status, then start running from the default position	Under this policy, running tasks will be stopped, and their status discarded. Then, according to the source configuration, reading will commence from the default checkpoint. If the source configuration is set to "full + incremental" reading mode, it will by default first synchronise the stock data (full phase), and upon completion, consume binlog to obtain changed data (incremental phase); if the source configuration is set to "incremental only", it will by default start reading from the latest binlog checkpoint.
Suspended (Non-First Submission)	Continue running, retain job status data, and continue running from the last sync point	Under this strategy, after the new version of the task is submitted, it will continue running from the last sync point. ⚠ **Note:** • **When pausing**, a snapshot will be generated. Task resubmission supports continuing from the last sync point. • **When force pausing**, no snapshot is generated. Task resubmission supports running from the most recently generated snapshot. This type of pause may result in partial data replay of the task. If the target write is Append, there may be duplicate data. If the target write is Upsert, there will be no duplication issue.
	Restart, continue running from the designated time point	Under this strategy, you can specify the position to restart reading. The task will ignore the old version and restart reading from the specified position. If the specified time position is not found, the task will default to synchronize from the earliest binlog position.
	Restart, stop the running task and discard its status, then start running from the default position	Under this policy, running tasks will be stopped, and their status discarded. Then, according to the source configuration, reading will commence from the default checkpoint. If the source configuration is set to "full + incremental" reading mode, it will by default first synchronise the stock data (full phase), and upon completion, consume binlog to obtain changed data (incremental phase); if the source configuration is set to

©2013-2025 Tencent Cloud. All rights reserved.

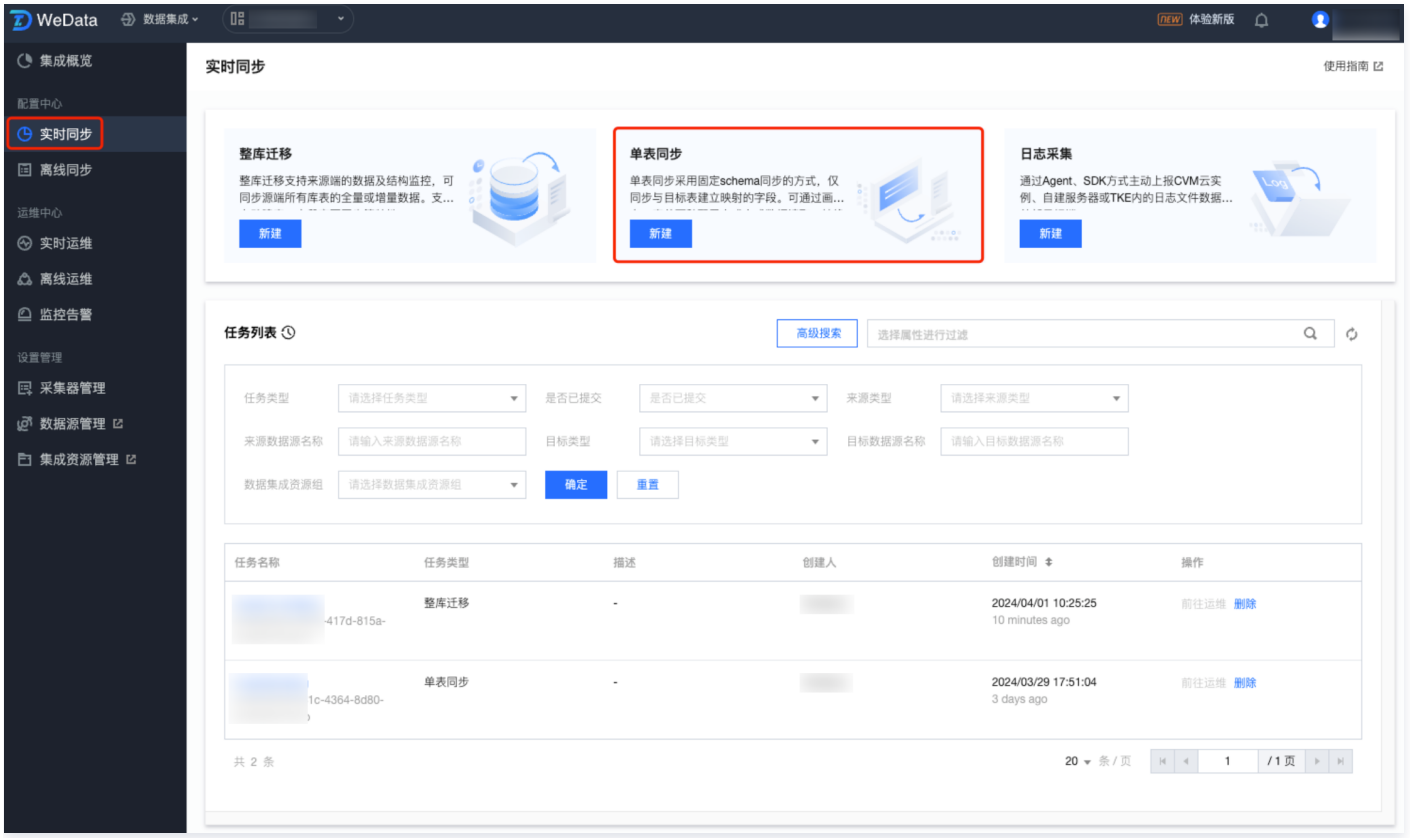
		"incremental only", it will by default start reading from the latest binlog checkpoint.
Failed (Non-First Submission)	Resume operation from the previous failed (checkpoint) position	Under this strategy, it will continue running from the last failed position of the task
	Restart and run based on the default position according to the task reading configuration	Under this policy, it will start reading from the default position according to the configuration of the source. If the configuration of the source is set to "full + incremental" mode, it will by default first sync the existing data (full phase) and, upon completion, consume binlog to obtain the changed data (incremental phase); if the source is configured as "incremental only", it will by default start reading from the latest position of binlog.
Operating (Non-First Submission)	Not supported	When there is an existing task with the same name and its status is **In operation**, re-submitting the task is not supported

Step 8: Real-time Task Operations and Maintenance

For details, please refer to Real-time Operation and Maintenance Operations.

6. Real-time Single Table: MySQL Synchronization to TChouse-D

When the business needs to focus only on a specific table, single table synchronization can target the data of that table. This reduces unnecessary data processing and transmission, improving sync efficiency.
On the left sidebar of the DataInLong page, click **Real-time Synchronization**. At the top of the Real-time Synchronization page, select Single Table Synchronization to create a new one and enter the configuration page. When configuring single table synchronization, two modes are supported: **Single Table Mode** and **Canvas Mode**. The configuration process will be explained using Form mode as an example.

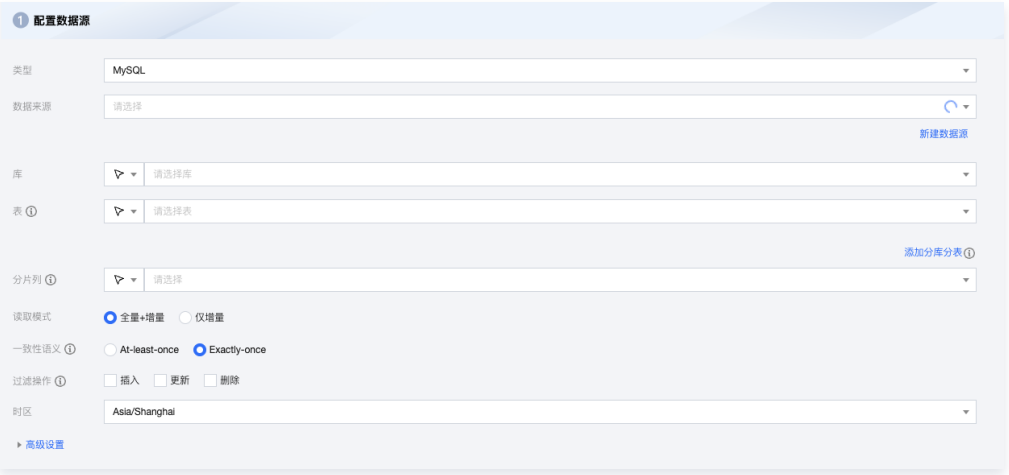


Step 1: Configure MySQL Read Node

1. In the DataInLong page, click **Real-time Synchronization** in the left sidebar.

©2013-2025 Tencent Cloud. All rights reserved.

2. On the Real-time Synchronization page, select **Single Table Synchronization** at the top, create a new one (choose Form or Canvas mode), and enter the configuration page.
3. Select the MySQL node and configure the node information.



4. Parameter information is listed in the table below:

Data Source	Select the available MySQL data source for the project.
Database	Support selection or manual input of the library name to be read. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Support selecting or manually entering the Table Name to be read. • In the case of table partitioning, you can select or enter multiple table names on the MySQL source end. Multiple tables must have consistent structures. • In the case of table partitioning, you can configure the table serial number range. For example, 'table_[0-99]' means reading 'table_0', 'table_1', 'table_2' up to 'table_99'; if your table numeric suffixes are consistent in length, such as 'table_000', 'table_001', 'table_002' up to 'table_999', you can configure it as '"table": ["table_00[0-9]", "table_0[10-99]", "table_[100-999]"]'. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Adding Sharding	Applicable to sharding scenarios, you can configure multiple data sources, databases, and table information by clicking. In sharding scenarios, the structure of all tables must be consistent. The task configuration will default to displaying and using the first table structure for data retrieval.
Shard Column	The shard column is used to divide the table into multiple shards for synchronization. For tables with primary keys, it is recommended to select the table's primary key as the shard column first; For tables without primary keys, it is recommended to choose an indexed column as the shard column and ensure that no data update operations are performed on the shard column; otherwise, only At-Least-Once semantics can be guaranteed.
Read mode	Supports both Full + Incremental and Incremental only modes.
Consistency Semantics	• Exactly-once • At-least-once **Note:** Represents the consistency semantics of the reading end only. The current version's two modes have incompatible states. If the mode is changed after task submission, stateful restart is not supported.

©2013-2025 Tencent Cloud. All rights reserved.

Filter operations	After setting, data of the specified operation types will not be synchronized. Supports Insert, Update, and Delete.
Time zone	Set the log time timezone, default is Shanghai.
Advanced Settings (optional)	You can configure parameters based on business requirements.

5. Preview data fields, click **Save**.

Step 2: Configure TChouse-D Write Node

1. In the DataInLong page, click **Real-time Synchronization** in the left sidebar.
2. On the Real-time Synchronization page, select **Single Table Synchronization** at the top, create a new one (choose Form or Canvas mode), and enter the configuration page.
3. Select the TChouse-D node and configure the node information.



4. Parameter information is listed in the table below:

Parameter	Description
Data Flow Direction	Required TChouse-D Data Source.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Supports selecting or manually entering the name of the table to be written to. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.

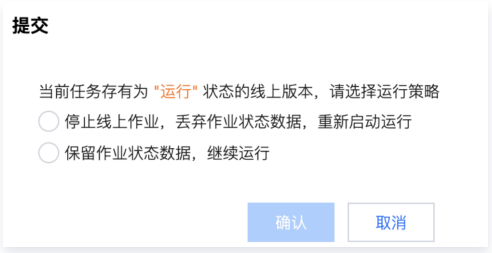
Step 3: Task Configuration Check and Submission

After completing the above steps, you can proceed with a task configuration check and formally submit the task. At this point, you can refer to the related buttons in the upper left corner of the task configuration interface for subsequent submission operations. The position and features of each button are as follows:

©2013-2025 Tencent Cloud. All rights reserved.



After the real-time synchronization task is configured, you can configure the operation strategy and submit it to the production environment. Currently, the task configuration page supports save, submit, lock/unlock, go to O&M, and form/canvas conversion operations.

Serial number	Parameter	Description
1	Save	Save current task configuration information, including data node configuration, node connection, and task attribute configuration.
2	Submit	Submit the current task to the production environment. At the time of submission, different operation strategies can be selected based on whether there is an active production task associated with the current task. • If there are no effective online tasks for the current task, i.e., it's the first submission or the online task is in a "failed" status, you can submit it directly. • If there are "Running" or "Paused" online tasks, you need to choose different strategies. Stopping the online job will abandon the previous task's running location, consuming data from the beginning. Retaining the job status will resume from the last consumed location after restarting.  Note

		Clicking to start the task immediately will make it run immediately after submission. Otherwise, it will require manual trigger to start officially.
3	Lock/Unlock	The default creator is the first lock holder, and only the lock holder can edit task configurations and run tasks. If the lock holder does not perform any editing operations within 5 minutes, others can click the icon to acquire the lock. If successful, they can perform editing operations.
4	Go to Operations	Quickly navigate to the real-time O&M page based on the current task name.
5	Refresh	Refresh Task Configuration
6	Canvas Conversion/Form Conversion	Form Mode and Canvas Mode can be converted to each other. Form Mode provides the canvas conversion feature, and Canvas Mode provides the form conversion feature. **Note** • Currently, only tasks in the saved state are allowed to be converted. • If the canvas mode contains conversion nodes, it is not supported to switch to the form mode.

Task Configuration Check

When you submit a task, a basic environment and configuration check will be performed. Pay attention to the failed items and warning items. Failed items are likely to cause task execution failure, while warning items may affect the task based on the business scenario. The current check results will not block task submission; you can choose to recheck or ignore the exceptions and continue submitting.



Currently supported inspection items:

Detection Category	Inspection Item	Description
Task Configuration Detection	Source Configuration	Check if any required fields are missing on the source end
	Target Configuration	Check if any required fields are missing on the target end

©2013–2025 Tencent Cloud. All rights reserved.

	Mapping Configuration	Check if field mapping has been configured
	Resource Group Configuration	Check if the resource group is configured
	Same Data Table Detection	This check ensures that all real-time single-table tasks (including submitted and unsubmitted) under the current project do not have the same source table or target table, or both source and target tables are the same. The judgment basis for the same table is whether it is the same-named table under the same data source and the same DB. This inspection item mainly prevents multiple tasks from reading the same table repeatedly.
Data Source Detection	Source end connectivity detection	Check if the source data source is network-connected with the task configuration's resource group. If detection fails, check the diagnostic information. Recheck after network connectivity is established, otherwise the task is likely to fail.
	Target end connectivity detection	Check if the target data source is network-connected with the task configuration's resource group. If detection fails, check the diagnostic information. Recheck after network connectivity is established, otherwise the task is likely to fail.
Resource Detection	Resource Status Detection	Check if the resource group is in an available status. If the resource status is unavailable, please change the resource group in the task configuration, otherwise the task is likely to fail.
	Resource Availability Detection	Check if the current remaining resources of the resource group meet the resource requirements of the task configuration. If the detection fails, please reduce the task resource configuration or expand the resource group.

Submission Policy Selection

After normal submission or ignoring exceptions, you will enter the submission policy selection page, where you can choose different operation strategies according to your business needs:



The submission run strategy supported by different task running statuses varies:

Task Status	Submission and Operation Strategies	Description
1. First Submission 2. Stopped/Detect Anomalies/I	Start immediately, synchronize from the default locus	Under this strategy, data will be read from the default locus based on the source configuration. If the source configuration is set to "Full + Increment" mode, the default is to first synchronize the stock data (full phase), and then consume binlog to get the changed data (incremental phase) after

©2013-2025 Tencent Cloud. All rights reserved.

nitialization (not the first submission)		completion. If the source configuration is set to "Increment Only" mode, the default is to start reading from the latest binlog position.
	Start immediately, synchronize from a designated time point	Under this strategy, a specific start time needs to be selected, and synchronization will be based on time matching points. 1. Start reading data from the designated time point. If the specified position is not matched, the task will start synchronizing from the earliest binlog position by default 2. If your source end reading method is **Full + Incremental**, selecting this strategy will skip the full phase by default and start synchronizing from the designated incremental time position
	Do not start now, go to Real-time Task Operation and Maintenance later to manually start the task	Under this strategy, the task is only submitted to Real-time Operation and Maintenance without starting, and tasks can be started in batches from the Real-time Operation and Maintenance page later.
Running (Non-First Submission)	Continue running, retain job status data, and continue running from the last sync point	Under this strategy, after the new version of the task is submitted, it will continue running from the last sync point.
	Restart, continue running from the designated time point	Under this strategy, you can specify the position to restart reading. The task will ignore the old version and restart reading from the specified position. If the specified time position is not found, the task will default to synchronize from the earliest binlog position.
	Restart, stop the running task and discard its status, then start running from the default position	Under this policy, running tasks will be stopped, and their status discarded. Then, according to the source configuration, reading will commence from the default checkpoint. If the source configuration is set to "full + incremental" reading mode, it will by default first synchronise the stock data (full phase), and upon completion, consume binlog to obtain changed data (incremental phase); if the source configuration is set to "incremental only", it will by default start reading from the latest binlog checkpoint.
Suspended (Non-First Submission)	Continue running, retain job status data, and continue running from the last sync point	Under this strategy, after the new version of the task is submitted, it will continue running from the last sync point. ⚠ **Note:** • **When pausing**, a snapshot will be generated. Task resubmission supports continuing from the last sync point. • **When force pausing**, no snapshot is generated. Task resubmission supports running from the most recently generated snapshot. This type of pause may result in partial data replay of the task. If the target write is Append, there may be duplicate data. If the target write is Upsert, there will be no duplication issue.
	Restart, continue running from the designated time point	Under this strategy, you can specify the position to restart reading. The task will ignore the old version and restart reading from the specified position. If the specified time position is not found, the task will default to synchronize from the earliest binlog position.

©2013-2025 Tencent Cloud. All rights reserved.

	Restart, stop the running task and discard its status, then start running from the default position	Under this policy, running tasks will be stopped, and their status discarded. Then, according to the source configuration, reading will commence from the default checkpoint. If the source configuration is set to "full + incremental" reading mode, it will by default first synchronise the stock data (full phase), and upon completion, consume binlog to obtain changed data (incremental phase); if the source configuration is set to "incremental only", it will by default start reading from the latest binlog checkpoint.
Failed (Non-First Submission)	Resume operation from the previous failed (checkpoint) position	Under this strategy, it will continue running from the last failed position of the task
	Restart and run based on the default position according to the task reading configuration	Under this strategy, data will be read from the default locus based on the source configuration. If the source configuration is set to "Full + Increment" mode, the default is to first synchronize the stock data (full phase), and then consume binlog to get the changed data (incremental phase) after completion. If the source configuration is set to "Increment Only" mode, the default is to start reading from the latest binlog position.
Operating (Non-First Submission)	Not supported	When there is an existing task with the same name and its status is **In operation**, re-submitting the task is not supported

Each submission will generate a new version of the real-time task. You can configure the version description in the dialog box.

Submission Result

After clicking the **Submit** button in the above image, the submission progress and submission results will appear:



- Task Submitting:
  - Shows the submission progress percentage.
- Task Submission Result – Successful:
  - Show task submission successful result.
  - You can go to Operations for task O&M management or return to the task list
- Task Submission Result – Failed:
  - Display the reason for task submission failure

Step 4: Real-time Task Operation and Maintenance

©2013-2025 Tencent Cloud. All rights reserved.

For more information, please refer to Real-time operation and maintenance.

7. Real-time Node Advanced Parameters

When configuring read-write nodes, you can configure advanced settings. The configuration and usage scenarios are as follows:

MySQL Node Level

The following are the advanced settings and their descriptions supported by the MySQL source end in the usage scenario:

Read/Write	Applicable Scenario	Configuration:	Description
Read	Single Table + Whole Database	scan.newly-added-table.enabled=true	**Parameter Description:** Set this parameter to sense new tables after **Suspension** > **Continue**. The default is false 1. Use this parameter during full incremental synchronization. New tables will read the stock data first, then read the incremental data 2. Use this parameter during incremental synchronization. New tables will only read incremental data
Read	Single Table + Whole Database	scan.incremental.snapshot.chunk.size=20000	**Parameter Description:** For tasks with uniform data distribution, this parameter represents the approximate number of records in one chunk. You can estimate the number of chunks by dividing the total data volume by the chunk size. The number of chunks affects whether jobmanager experiences an out-of-memory error. Currently, with 2CU, it can support more than 100,000 chunks. If the data volume is too large, you can increase the chunk size to reduce the number of chunks **Precautions:** For large data volume tasks (e.g., total data volume over 100 million, single record larger than 0.1M), it is generally recommended to set 20,000
Read	Single Table + Whole Database	split-key.even-distribution.factor.upper-bound=10.0d	**Parameter Description:** During the MySQL existing data reading phase, if the data is quite scattered and the maximum value of the primary key field is extremely large, you can modify this parameter to use non-uniform splitting. This helps reduce the problem of having too many chunks due to the excessively large primary key value, which could lead to JM OOM issues **Precautions:** Default value is 10.0d, generally no need to modify
Read	Single Table + Whole Database	debezium.query.fetch.size=0	**Parameter Description:** Represents the number of records fetched from the database each time. The default value is 0, which indicates the JDBC default fetch size **Precautions:** 1. For large tasks (e.g., total data volume over 100 million, individual record size over 0.1M) with only one reading instance, it is recommended to set this to 1024 records 2. For tasks with multiple reading instances, it is recommended to lower this value to reduce memory consumption. Setting it to 512 records is advised
Read	Single Table + Whole Database	debezium.max.queue.size=8192	**Parameter Description:** This attribute defines the maximum number of events stored in the internal queue. If this limit is reached, Debezium will pause reading new events until the unprocessed events are handled and committed. This helps prevent an excessive backlog of events in the queue, thus avoiding memory exhaustion and performance degradation. The default is 8192 **Precautions:**

©2013-2025 Tencent Cloud. All rights reserved.

			1. For large tasks (e.g., total data volume over 100 million, individual record size over 0.1M) with only one reading instance, it is recommended to set this to 4096 2. For tasks with multiple reading instances, it is recommended to lower this value to reduce memory consumption. Setting it to 1024 is advised

Doris/TCHouse-D type node level

The following are the advanced settings and their descriptions supported by the Doris/TCHouse-D target end in the usage scenario:

Read/Write	Applicable Scenario	Configuration:	Description
Write	Single Table Only	sink.properties.*=xxx	**Parameter Description:** Import Parameters for Stream Load. For example, 'sink.properties.column_separator' = ', ' **For detailed configuration reference:** Flink Doris Connector
Write	Single Table Only	sink.properties.columns=xxx	**Parameter Description:** Configure the function mapping relationship of columns. For example, 'sink.properties.columns' = 'dt,page,user_id,user_id=to_bitmap(user_id)' **Refer to:** Doris Stream Load
Write	Single Table + Whole Database	• sink.batch.size = 100000 • sink.batch.bytes= 83886080 • sink.batch.interval= 10s	**Parameter Description:** Improve the efficiency of writing to Doris **Note:** tm cu is recommended to be set to 2CU to avoid tm oom

8. FAQs

1. MySQL server ID conflict

Error message:

```
com.github.shyiko.mysql.binlog.network.ServerException: A slave with the same server_uuid/server_id as
this slave has connected to the master.
```

Solution: We have now optimized to randomly generate serverid. If in previous tasks the server-id was explicitly specified in MySQL advanced parameters, it is recommended to delete it, as multiple tasks may use the same data source with the same server-id, causing a conflict.

2. Binlog File Not Found Error:

Error message:

```
Caused by: org.apache.kafka.connect.errors.ConnectException: The connector is trying to read binlog
starting at GTIDs xxx and binlog file 'binlog.xxx', pos=xxx, skipping 4 events plus 1 rows, but this is no
longer available on the server. Reconfigure the connector to use a snapshot when needed.
```

Error Causes:
An error will occur if the binlog file being read by the job has already been cleaned up on the MySQL server. There are several reasons for Binlog cleanup, such as a too-short Binlog retention period, or the job processing speed not keeping up with the Binlog generation speed, exceeding the maximum retention period of MySQL Binlog files. This results in the Binlog files being cleaned up on the MySQL server, causing the Binlog positions being read to become invalid.

©2013-2025 Tencent Cloud. All rights reserved.

Solution: If the job processing speed cannot keep up with the Binlog generation speed, you can consider increasing the Binlog retention period or optimizing the job to reduce backpressure and accelerate source consumption. If there are no anomalies in the job status, it might be that other operations on the database caused the Binlog to be cleaned up, leading to inaccessibility. You need to refer to the MySQL database-side information to determine the reason for Binlog cleanup.

3. MySQL Connection Reset

Error message:

```
EventDataDeserializationException: Failed to deserialize data of EventHeaderV4 .... Caused by:
java.net.SocketException: Connection reset.
```

Error Causes:

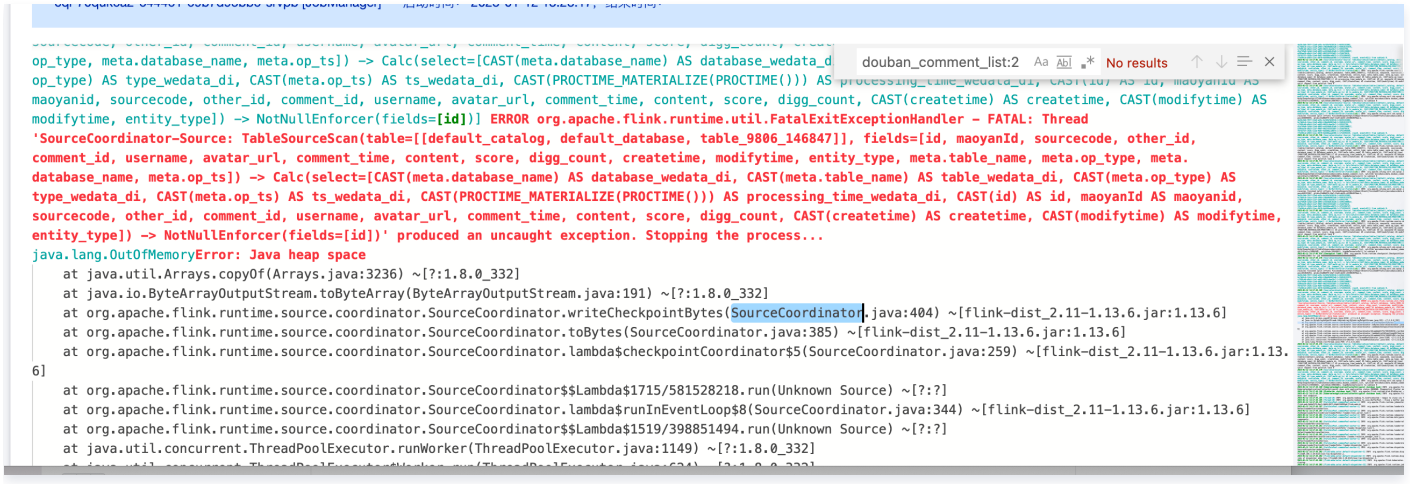
1. Network Issues.
2. The job has backpressure, causing the source to be unable to read data, and the binlog client to be idle. If the binlog connection remains idle after the timeout, the MySQL server will disconnect the idle connection.

Solution:

1. If it is a network issue, you can increase the MySQL network parameters set global slave_net_timeout = 120; (default 30s) set global thread_pool_idle_timeout = 120.
2. If the backpressure is caused by the job, you can alleviate it by adjusting the job, such as increasing parallelism, improving write speed, and increasing TaskManager memory to reduce GC.

4. Mysql2dlc task JobManager OOM

Error message:



1. If the user data volume is large, you can increase the JobManager CU number and use the MySQL advanced parameter scan.incremental.snapshot.chunk.size to increase the chunk size, which defaults to 8096.
2. The user's data volume is not large, but there is a large difference between the maximum and minimum values of the primary key, resulting in many chunks being divided when using the equal chunk strategy. Modify the distribution factor so that user data follows an uneven data splitting logic, split-key.even-distribution.factor.upper-bound=5.0d. The default distribution factor has been modified to 10.0d.

5. The format of the user's binlog data is incorrect, causing a parsing exception in Debezium

Error message:

```
ERROR io.debezium.connector.mysql.MySqlStreamingChangeEventSource [] - Error during binlog processing.
Last offset stored = null, binlog reader near position = mysql-bin.000044/211839464.
2023-02-20 21:37:28.480 [blc-172.17.48.3:3306] ERROR io.debezium.pipeline.ErrorHandler [] - Producer
failure
io.debezium.DebeziumException: Error processing binlog event.
```

©2013-2025 Tencent Cloud. All rights reserved.

Solution:
After modifying binlog_row_image=full, you need to restart the database.

6. Is gh-ost supported?

Yes, it does not migrate the temporary table data generated by Online DDL changes. It only migrates the original DDL data executed using gh-ost in the source database. You can use the default or configure your own regular expressions for gh-ost shadow tables and unused tables.

7. How to select and optimize the Doris specifications?

Please refer to Doris Resource Specifications Model Selection and Optimization Suggestions .

9. Too many import tasks. A new import task submission error "current running txns on db xxx is xx, larger than limit xx "?

Adjust the FE parameter: max_running_txn_num_per_db, default is 100, can be appropriately increased, recommended to control within 500.

10. Import frequency too fast causing error err=[E-235]?

- Parameter optimization suggestion: You can temporarily resolve it by appropriately increasing the max_tablet_version_num parameter. The default is 200, recommended to control within 2000.
- Business optimization suggestion: Reduce the import frequency to fundamentally solve this problem.

11. The import file is too large, restricted by parameters. Error "The size of this batch exceeds the max size"?

Adjust the BE parameter: streaming_load_max_mb, recommended to exceed the size of the file to be imported.

12. Import data error: "[-238]"?

- Reason: The -238 error usually occurs when the volume of data imported in the same batch is too large, resulting in too many Segment files in a single tablet.
- Parameter optimization suggestion: You can appropriately increase the BE parameter max_segment_num_per_rowset. The default value is 200, can be increased in multiples (such as 400, 800), recommended to control within 2000
- Business optimization suggestion: It is recommended to reduce the data volume of a single batch import.

13. Import failure, error: "too many filtered rows xxx," "ErrorURL":", or Insert has filtered data in strict mode, tracking url=xxxx."?

Reason: Schema, partition, etc., of the table do not match the imported data. You can execute the doris command in TCHouse-D Studio or client to check the specific reason: show load warnings on `<tracking url>`, `<tracking url>` is the error URL returned in the error message.

©2013-2025 Tencent Cloud. All rights reserved.

MySQL Offline Synchronization to Hive Guide

Last updated: 2024-09-05 17:46:38

This document will guide you step-by-step through the process of real-time synchronization from MySQL to Iceberg, including whole database migration and single table synchronization. The general process is as follows:
1. Data Source Compatibility.
2. MySQL Environment Preparation.
3. Iceberg Environment Preparation.
4. Project Configuration and Resource Group Preparation.
5. Real-time Whole Database: MySQL to Iceberg Configuration Steps.
6. Real-time Single Table: MySQL to Iceberg Configuration Steps.
7. Real-time Node Advanced Parameters.
8. FAQs.

1. Data Source Compatibility

MySQL Data Source

If you want to use MySQL for real-time data synchronization, you need to first confirm the MySQL data source version support, usage limitations, and supported data type conversion.

Supported Versions

Currently, DataInLong supports real-time reading at both single table and database levels for MySQL. To use real-time reading capabilities, you must follow these version restrictions:

Type	Version	Driver
MySQL	5.6,5.7,8.0.x	JDBC Driver:8.0.21
RDS MySQL	5.6,5.7, 8.0.x	
PolarDB MySQL	5.6,5.7,8.0.x	
Aurora MySQL	5.6,5.7,8.0.x	
MariaDB	10.x	
PolarDB X	2.0.1	

Use Limits

- Binlog logging needs to be enabled, and only MySQL server Binlog configuration format ROW is supported for synchronization.
- Tables without a primary key may have duplicate data because exact once cannot be guaranteed, so real-time synchronization tasks must ensure a primary key exists.
- XA ROLLBACK is not supported. Real-time synchronization tasks will not rollback data for XA PREPARE. To handle XA ROLLBACK scenarios, you need to manually remove the XA ROLLBACK tables from real-time synchronization tasks and then re-sync the tables after adding them back.
- Set MySQL Session Timeout:

When creating an initial consistent snapshot for a large database, the connection you establish may time out while reading tables. You can prevent this by configuring interactive_timeout and wait_timeout in the MySQL configuration file.
  - interactive_timeout: The number of seconds the server waits for activity on an interactive connection before closing it. Please refer to MySQL :: MySQL 8.0 Reference Manual :: 7.1.8 Server System Variables.
  - wait_timeout: The number of seconds the server waits for activity on a non-interactive connection before closing it. Please refer to MySQL :: MySQL 8.0 Reference Manual :: 7.1.8 Server System Variables.

MySQL Data Type Conversion

The supported data types and internal mapping fields for reading MySQL data are shown below. (When processing MySQL, the data types from the MySQL data source will be mapped to the data processing engine data types first):

©2013-2025 Tencent Cloud. All rights reserved.

Field Type	Supported	Internal mapping fields	Remarks
TINYINT	Yes	TINYINT	TINYINT(1) maps to BOOLEAN Options need to be added to support TINYINT(1) mapping to bool or tinyint
SMALLINT	Yes	SMALLINT	–
TINYINT_UNSIGNED	Yes	SMALLINT	–
TINYINT_UNSIGNED_ZEROFILL	Yes	SMALLINT	–
INT	Yes	INT	–
INTEGER	Yes	INT	–
YEAR	Yes	INT	–
MEDIUMINT	Yes	INT	–
SMALLINT_UNSIGNED	Yes	INT	–
SMALLINT_UNSIGNED_ZEROFILL	Yes	INT	–
BIGINT	Yes	LONG	–
INT_UNSIGNED	Yes	LONG	–
MEDIUMINT_UNSIGNED	Yes	LONG	–
MEDIUMINT_UNSIGNED_ZEROFILL	Yes	LONG	–
INT_UNSIGNED_ZEROFILL	Yes	LONG	–
BIGINT_UNSIGNED	Yes	DECIMAL	DECIMAL(20,0)
BIGINT_UNSIGNED_ZEROFILL	Yes	DECIMAL	DECIMAL(20,0)
SERIAL	Yes	DECIMAL	DECIMAL(20,0)
FLOAT	Yes	FLOAT	–
FLOAT_UNSIGNED	Yes	FLOAT	–
FLOAT_UNSIGNED_ZEROFILL	Yes	FLOAT	–
DOUBLE	Yes	DOUBLE	–
DOUBLE_UNSIGNED	Yes	DOUBLE	–
DOUBLE_UNSIGNED_ZEROFILL	Yes	DOUBLE	–
DOUBLE_PRECISION	Yes	DOUBLE	–
DOUBLE_PRECISION_UNSIGNED	Yes	DOUBLE	–
ZEROFILL	Yes	DOUBLE	–
REAL	Yes	DOUBLE	–
REAL_UNSIGNED	Yes	DOUBLE	–
REAL_UNSIGNED_ZEROFILL	Yes	DOUBLE	–

©2013–2025 Tencent Cloud. All rights reserved.

NUMERIC	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
NUMERIC_UNSIGNED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
NUMERIC_UNSIGNED_ZEROFILL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
DECIMAL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
DECIMAL_UNSIGNED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
DECIMAL_UNSIGNED_ZEROFILL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
FIXED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
FIXED_UNSIGNED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
FIXED_UNSIGNED_ZEROFILL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
BOOLEAN	Yes	BOOLEAN	–
DATE	Yes	DATE	–
TIME	Yes	TIME	–
DATETIME	Yes	TIMESTAMP	–
TIMESTAMP	Yes	TIMESTAMP	–
CHAR	Yes	STRING	–
JSON	Yes	STRING	–
BIT	Yes	STRING	BIT(1) maps to BOOLEAN
VARCHAR	Yes	STRING	–
TEXT	Yes	STRING	–
BLOB	Yes	STRING	–
TINYBLOB	Yes	STRING	–
TINYTEXT	Yes	STRING	–
MEDIUMBLOB	Yes	STRING	–
MEDIUMTEXT	Yes	STRING	–
LONGBLOB	Yes	STRING	–
LONGTEXT	Yes	STRING	–
VARBINARY	Yes	STRING	–
GEOMETRY	Yes	STRING	–
POINT	Yes	STRING	–
LINESTRING	Yes	STRING	–

©2013–2025 Tencent Cloud. All rights reserved.

POLYGON	Yes	STRING	-
MULTIPOINT	Yes	STRING	-
MULTILINESTRING	Yes	STRING	-
MULTIPOLYGON	Yes	STRING	-
GEOMETRYCOLLECTION	Yes	STRING	-
ENUM	Yes	STRING	-
BINARY	Yes	BINARY	BINARY(1)
SET	No		-

Iceberg Data Source

If you want to perform real-time data synchronization to Iceberg, you need to first confirm the Iceberg data source version support, usage limitations, and supported data type conversion.

Supported Versions

Iceberg supports version 0.11 and above.

Use Limits

After writing to Iceberg, to improve query performance, it is generally required that the downstream performs small file merging through timed Spark Action and reasonably configures the Checkpoint interval.

Iceberg Data Type Conversion for Writing

The supported data types and internal mapping field types for Iceberg writing are as follows (when processing Iceberg, Iceberg data source data types will be mapped with those of the data processing engine):

Internal Type	Iceberg Type
CHAR	STRING
VARCHAR	STRING
STRING	STRING
BOOLEAN	BOOLEAN
BINARY	FIXED(L)
VARBINARY	BINARY
DECIMAL	DECIMAL(P,S)
TINYINT	INT
SMALLINT	INT
INTEGER	INT
BIGINT	LONG
FLOAT	FLOAT
DOUBLE	DOUBLE
DATE	DATE
TIME	TIME
TIMESTAMP	TIMESTAMP

©2013-2025 Tencent Cloud. All rights reserved.

TIMESTAMP_LTZ	TIMESTAMPTZ
INTERVAL	–
ARRAY	LIST
MULTISET	MAP
MAP	MAP
ROW	STRUCT
RAW	–

2. MySQL Environment Preparation

Confirm MySQL Version

DataInLong has requirements for MySQL versions. Check whether the MySQL version to be synced meets these requirements. You can check the current MySQL database version with the following statement.

```
select version();
```

Setting MySQL Server Permissions

You must define a MySQL user with appropriate permissions for all databases monitored by the Debezium MySQL connector.

1. **Create a MySQL User (Optional)**

```
mysql> CREATE USER 'user'@'localhost' IDENTIFIED BY 'password';
```

2. **Grant Necessary Permissions to the User:**

For real-time data synchronization, the account must have SELECT, REPLICATION SLAVE, and REPLICATION CLIENT permissions on the database. Refer to the following command for execution:

```
mysql> GRANT SELECT, SHOW DATABASES, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO 'user' IDENTIFIED BY
'password';
```

> **Note**
> Enabling scan.incremental.snapshot.enabled no longer requires the RELOAD permission (enabled by default).

3. **Refresh the User's Permissions:**

```
mysql>FLUSH PRIVILEGES;
```

See more about Permission Explanation.

Enable MySQL Binlog

1. **Check if Binlog is Enabled**

```
show variables like "log_bin"
```

If the result is ON, it means Binlog is enabled. If it's a slave, use the following statement:

©2013–2025 Tencent Cloud. All rights reserved.

```
show variables like "log_slave_updates";
```

If the return value is ON, Binlog is enabled. If Binlog is already enabled, you can skip the following steps.

2. **Enable Binlog**

If Binlog is confirmed not to be enabled, the following operations are required.

- For Tencent Cloud MySQL / TDSQL-C MySQL, binlog is enabled by default.
- For open-source MySQL, refer to the official documentation to enable binlog.

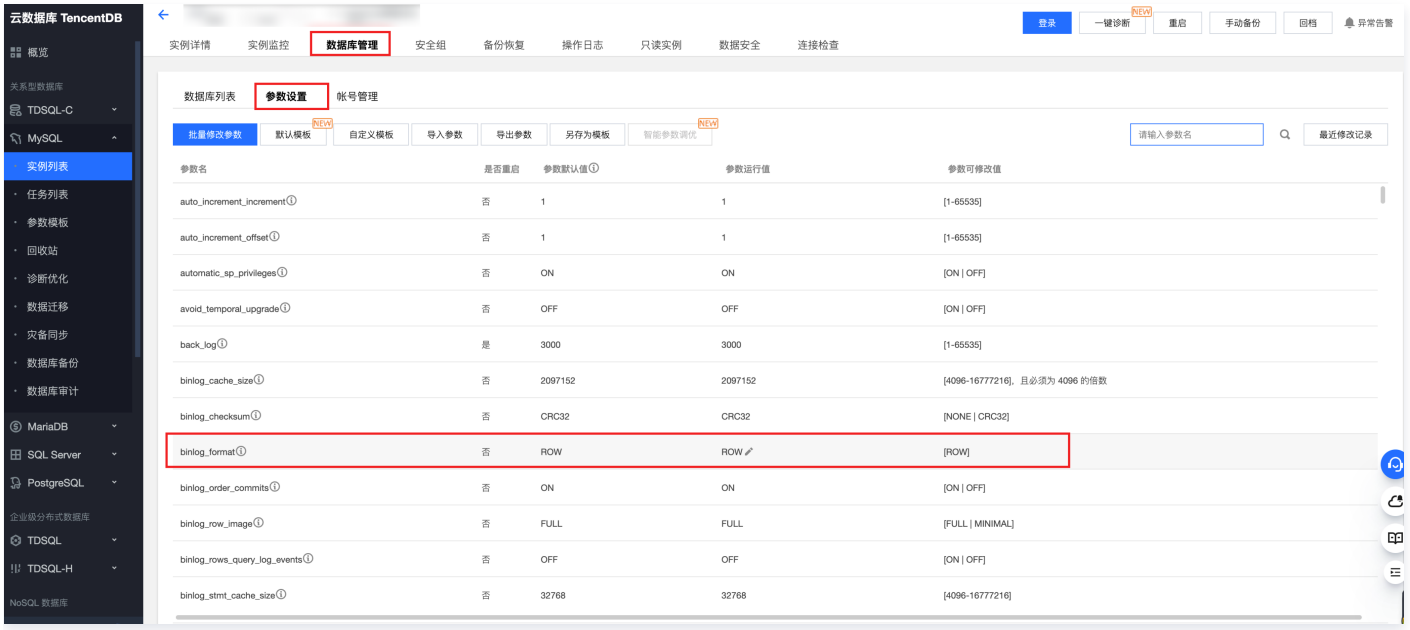
3. **Change Binlog Format to Row**

Real-time synchronization only supports MySQL server Binlog configuration format as ROW. Use the following statement to check the Binlog format.

```
show variables like "binlog_format";
```

If the return is non-ROW, please change the Binlog Format.

- For open-source MySQL, refer to the official documentation: MySQL :: MySQL 8.0 Reference Manual :: 19.1.6.4 Binary Logging Options and Variables.
- For Tencent Cloud MySQL / TDSQL-C MySQL:
  - Log in to the Tencent Cloud MySQL / TDSQL-C MySQL console, find the instance to enable Binlog, and click to enter the detailed information page of the instance.
  - In the tabs above, select DMC and find the Parameter Settings tab.
  - In the Parameter Settings tab, find the binlog_format parameter and set it to "ROW".



4. **binlog_row_image**

Real-time synchronization only supports MySQL server binlog_row_image configuration format as FULL or full.
Use the following statement to query the usage format of binlog_row_image.

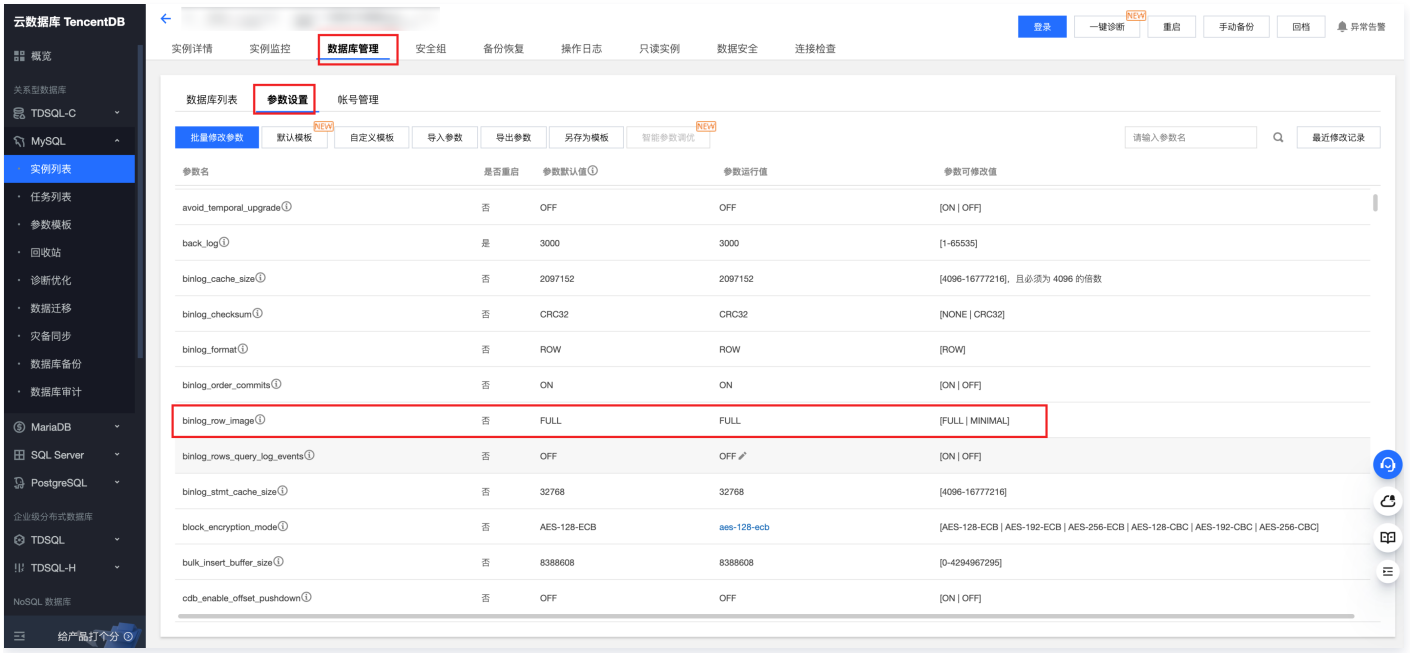
```
show variables like "binlog_row_image";
```

If the return is not FULL/full, please modify binlog_row_image

- For open-source MySQL, refer to the official documentation: MySQL :: MySQL 8.0 Reference Manual :: 19.1.6.4 Binary Logging Options and Variables.
- For Tencent Cloud MySQL / TDSQL-C MySQL:

©2013-2025 Tencent Cloud. All rights reserved.

- Log in to the Tencent Cloud MySQL / TDSQL-C MySQL console, find the instance to enable Binlog, and click to enter the detailed information page of the instance.
- In the tabs above, select DMC and find the Parameter Settings tab.
- In the parameter settings tab, find the binlog_row_image parameter and set it to "FULL".



Enable GTIDs (optional)

GTID (Global Transaction Identifier), used to uniquely identify a transaction in binlog, can avoid data confusion or master-slave inconsistency caused by duplicate transaction execution.

1. **Check if GTID is enabled**

```
show global variables like '%GTID%';
```

If the return result is similar to the following, it proves that GTID is enabled.

```
+--------------------------+-------+
| Variable_name            | Value |
+--------------------------+-------+
| enforce_gtid_consistency | ON    |
| gtid_mode                | ON    |
+--------------------------+-------+
```

2. **Enable GTID**

- For open-source MySQL, refer to the official documentation MySQL :: MySQL 8.0 Reference Manual :: 19.1.4.2 Enabling GTID Transactions Online.
- For Tencent Cloud MySQL / TDSQL-C MySQL instances, GTID is enabled by default and not supported to be disabled.

Add Tencent Cloud MySQL Database Security Group

A security group is a stateful virtual firewall with filtering features used to set network access control for one or more cloud databases. It is an important means of network security isolation provided by Tencent Cloud. If you use a Tencent Cloud MySQL database above the basic edition, you need to add the access IPs below to the target database's security group. If a security group is not configured, the connectivity test may fail when configuring the MySQL data source in WeData. For detailed steps, refer to Managing Cloud Database Security Groups.

©2013-2025 Tencent Cloud. All rights reserved.

```
118.89.220.0/24, 139.199.116.0/24, 140.143.68.0/24, 152.136.131.0/24, 81.70.150.0/24, 81.70.161.0/24,
81.70.195.0/24, 81.70.198.0/24, 82.156.22.0/24, 82.156.221.0/24, 82.156.23.0/24, 82.156.24.0/24,
82.156.27.0/24, 82.156.82.0/24, 82.156.84.0/24, 82.157.119.0/24
```

3. Iceberg Environment Preparation

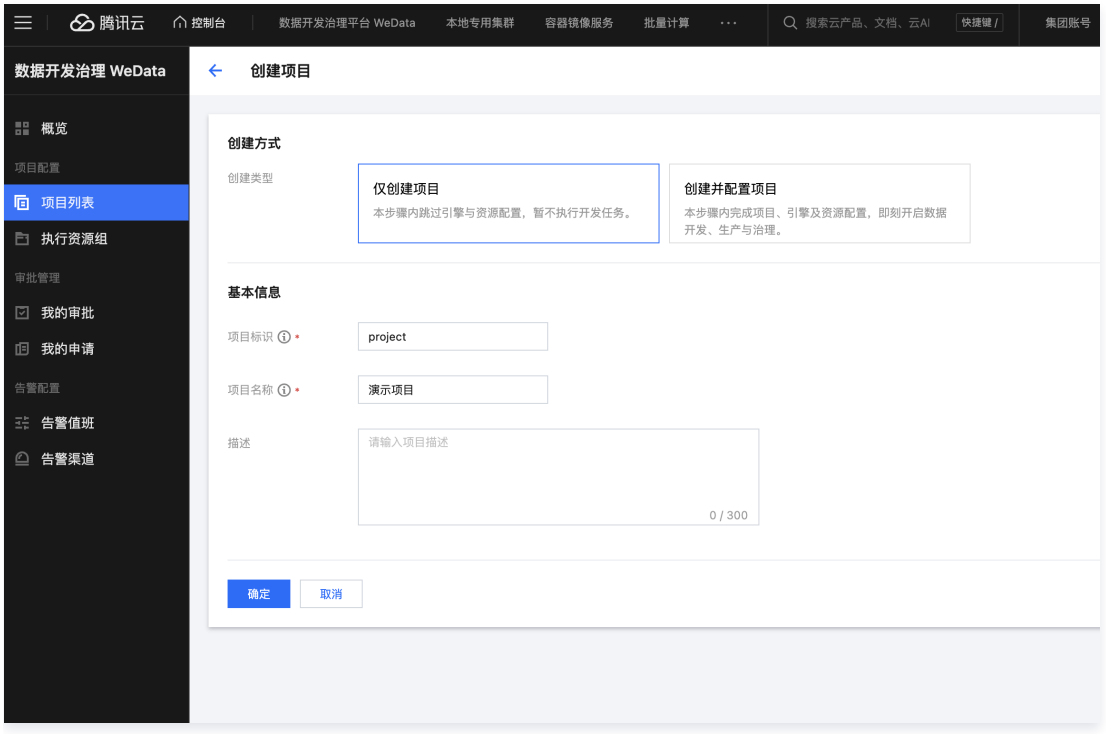
Real-time integration can automatically create target tables based on the upstream table structure. If you do not want to use this feature, you can also create the target table with the corresponding structure in Iceberg before running the job.

4. WeData Prerequisites

Before configuring the synchronization task, a project needs to be created. The following steps are based on a new environment. If you have already created project members or completed project creation, you can skip this step.

Project Creation

If you haven't created a project space in WeData, you need to complete the project creation before performing data synchronization tasks.
The main account holder logs into the CAM console to add a sub-account. Refer to Preparing a CAM Sub-account. Once created, the member can use the set sub-account to log in to the WeData big data platform.
To create a project space, go to the WeData homepage, click **Project List** > **Create Project**. If you are only doing DataInLong, you can choose **Create Project Only** and enter the project identifier and name.

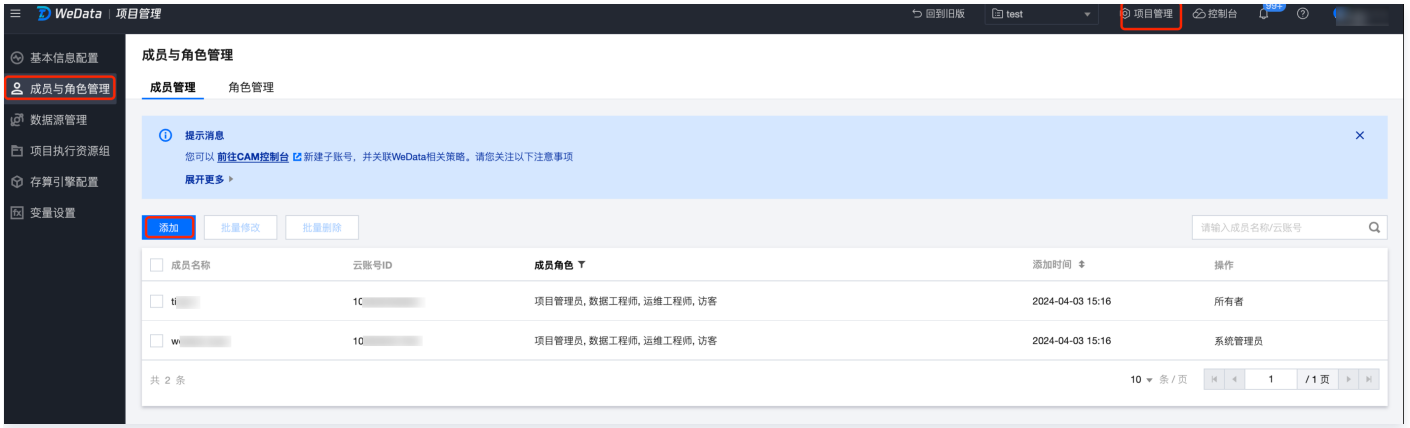


Add Member

Only the project owner and main account have permission to enter the newly created project. Other sub-accounts need to be added by the project owner or the main account in the project management module. For details, please see Project Management.

1. Click **Project Management** > **Member and Role Management** to enter the Member and Role Management interface. Click the **Add** button.

©2013-2025 Tencent Cloud. All rights reserved.



2. In the Add Member interface, set the required roles for the newly created member. Once set, the member can enter the project.

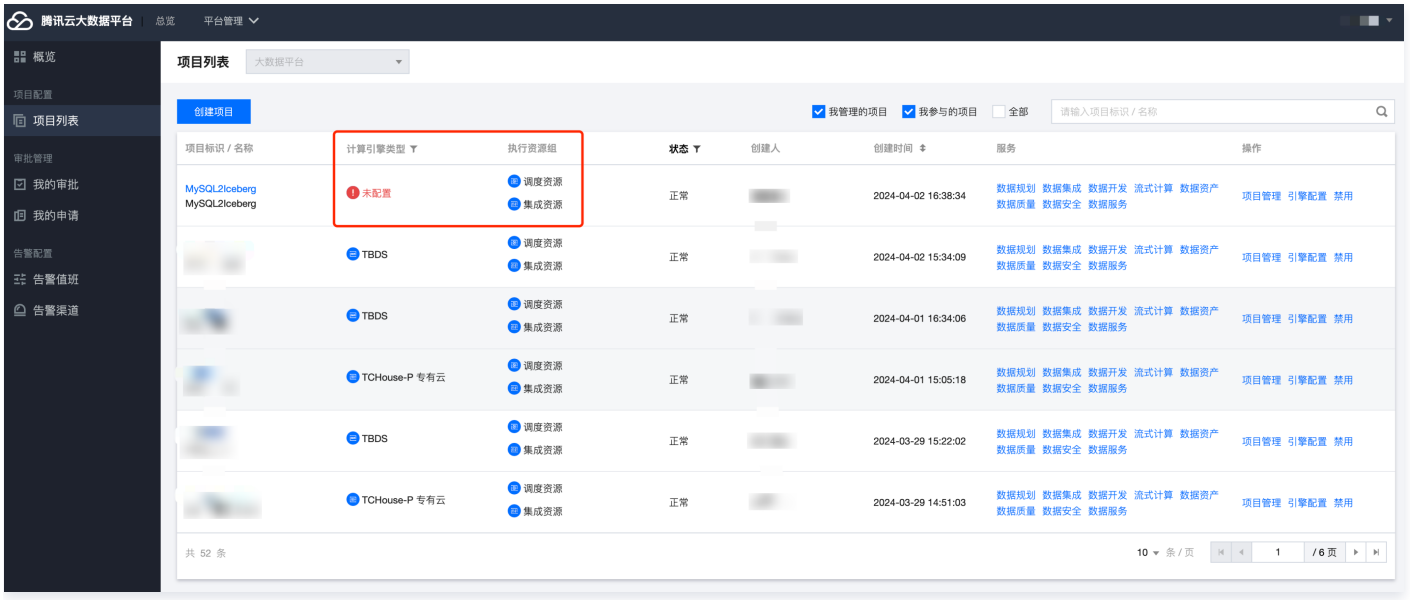


Resource Group Preparation

Confirm whether the Resource Group is configured

Before officially starting the DataInLong task, please confirm whether your project is equipped with **Integration Resources**.

- If it prompts that the integration resource is not configured, the DataInLong task cannot run normally.
- When you finish creating a new project, the column for **Execution Resource Group** will show that the integration resource is not configured. You just need to go to the WeData homepage > **Execution Resource Group** > **Integration Resource Group** and associate resources for the new project.



Enter the Integration Resource Group interface and click **Associate Project** after the corresponding Resource Group.

©2013–2025 Tencent Cloud. All rights reserved.

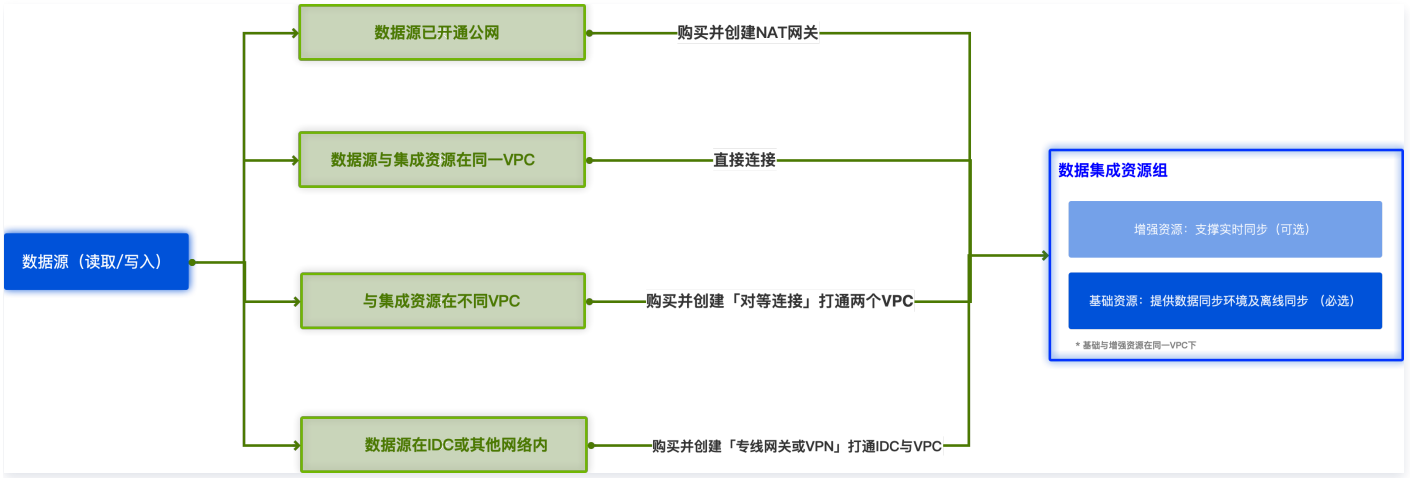


Purchase Integration Resource Group

If you find that there is currently no integration resource group, you need to go for purchase. The Integration Resource Group is dedicated to using computing resources when running DataInLong tasks. These resources are mainly presented in the form of resource groups. Before configuring the synchronization task, you need to confirm whether the DataInLong Resource Group is purchased. For details, please refer to Configure the Integration Resource Group.

Confirm Network Connectivity

Ensure network connectivity between the data source network (including the read endpoint and write endpoint) and the DataInLong resource group. Resources should not be denied access due to allowlist restrictions, as this will prevent data transmission and synchronization.
The machine resources contained in the DataInLong Resource Group must be in the same VPC network environment by default:



- **If the data source has public network access**: You need to purchase and create a NAT Gateway to allow the integration resources to connect to the data source VPC through the gateway. For detailed operations, please refer to Resource Group Configuration for Public Network.
- **If the data source is within a VPC:**
  - If the data source is in the same VPC as the integration resource: You can use it directly.
  - If the data source is in a different VPC from the integration resource: Purchase a Peering Connection to interconnect the VPCs of the integration resource and the data source.
- **If the data source is located in an IDC or other traditional network environment**: You need to purchase a VPN or Direct Connect Gateway to interconnect the VPCs of the integration resource and the data source.

©2013–2025 Tencent Cloud. All rights reserved.

> **Note**
> For documentation on the configuration and management of the Integrated Resource Group, refer to Integrated Resource Configuration and Management.

Data source configuration

MySQL Data Source Configuration

Enter the data source configuration interface, MySQL data sources support two connection methods: cloud instance and connection string.
click **Project Management** > **Data Source Management** > **New Data Source** > **Choose MySQL Data Source.**

- Create a data source through a cloud instance.



Parameter	Description
Connection Type	Select the data source connection method using cloud instance or connection string
Associated item	The item to which the data source belongs when created
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Display name	The name displayed when the data source is used in the product. If not filled, the default is the data source name
Description	Optional, description of this data source

©2013–2025 Tencent Cloud. All rights reserved.

Data source permissions	Project Sharing means that all members of the project can use the current data source. 'Personal and Administrator' means that only the creator and the project administrator can use the data source
Get instance	Select the region, instance name, and ID information of the TencentDB instance under the account
Database name	Database name to connect to
Username	Username to connect to the database
Password	Password to connect to the database
Data Connectivity (Legacy)	Test whether the configured database is connected **Note** • If the connectivity test fails, the data source can still be saved. However, it cannot be used if the connectivity test fails and the data source is saved. • If the connectivity test fails, it may be because WeData is blocked by the database network firewall. You need to add Tencent Cloud MySQL Database Security Group. Refer to Chapter 2 Add Tencent Cloud MySQL Database Security Group. • After the Chinese New Year in 2024, newly purchased WeData does not include this feature by default. This feature has been upgraded to Resource Group Connectivity.
Resource Group Connectivity (New)	Support the connectivity test between specific resource groups and data sources. Users only need to ensure that their data source and the VPC network where the Integrated Resource Group is located are connected without additional network interconnection.  **Note** Users who purchased WeData before the Chinese New Year in 2024 do not have this feature. If you need to use it, please contact Tencent's operation and maintenance personnel.

- Create a data source via connection string.

©2013–2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data source name	The name of the new data source, defined by the user, cannot be empty. Naming must start with a letter, can contain letters, numbers, and underscores. Length within 20 characters.
Description	Optional, description of this data source.
Data source permissions	Project Sharing means that all members of the project can use the current data source. 'Personal and Administrator' means that only the creator and the project administrator can use the data source.
Deployment mode	Support three deployment methods: CDB, self-built instances, and public network instances.
Region and Network	Region and VPCid of the data source.
JDBC URL	Connection string information for connecting to MySQL database. URL supports IP and domain name formats.
Database name	Name of the database to connect to.
Username	Username for connecting to the database.
Password	Password for connecting to the database.
Data Connectivity (Legacy)	Test whether the configured database is connected. **Note** • If the connectivity test fails, the data source can still be saved. However, it cannot be used if the connectivity test fails and the data source is saved.

©2013–2025 Tencent Cloud. All rights reserved.

- If the connectivity test fails, it may be because WeData is blocked by the database network firewall. You need to add Tencent Cloud MySQL Database Security Group. Refer to Chapter 2 Add Tencent Cloud MySQL Database Security Group.
- After the Chinese New Year in 2024, newly purchased WeData does not include this feature by default. This feature has been upgraded to Resource Group Connectivity.

Resource Group Connectivity (New)	Support the connectivity test between specific resource groups and data sources. Users only need to ensure that their data source and the VPC network where the Integrated Resource Group is located are connected without additional network interconnection.



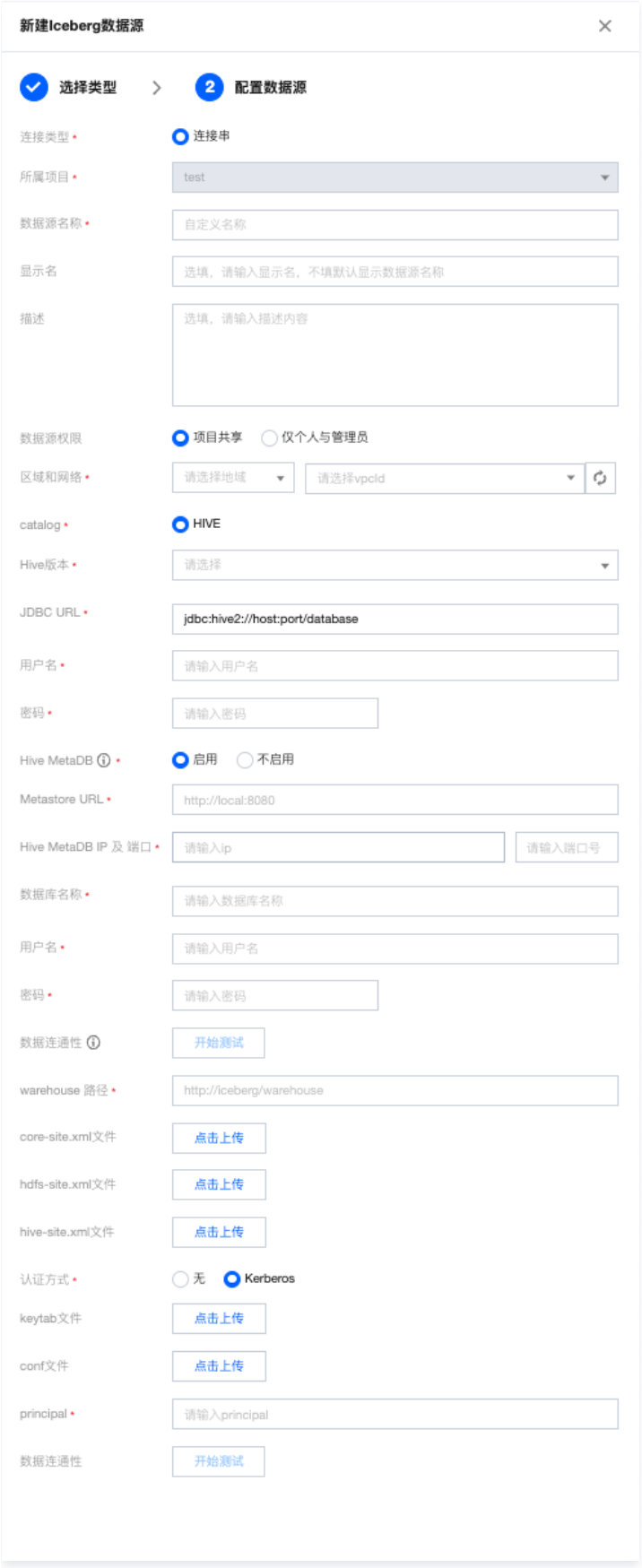
Note
Users who purchased WeData before the Chinese New Year in 2024 do not have this feature. If you need to use it, please contact Tencent's operation and maintenance personnel.

Iceberg Data Source Configuration

Enter the data source configuration interface, Iceberg data source only supports connection string method.
Click **Project Management** > **Data Source Management** > **New Data Source** > choose **Iceberg Data Source**.

- Create a data source via connection string.

©2013–2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data source name	The name of the new data source is defined by the user and cannot be empty. It must start with a letter and can contain letters, numbers, and underscores. Length within 20 characters. Cannot be

©2013–2025 Tencent Cloud. All rights reserved.

	modified!
Description	Optional, description of this data source.
Data source permissions	Project Sharing means that all members of the project can use the current data source. 'Personal and Administrator' means that only the creator and the project administrator can use the data source.
Region and Network	Region and VPCid of the data source.
catalog	Specify the directory service type used by Iceberg
Hive Version	Specified Hive version
JDBC URL	Connection string information for connecting to the Iceberg database.
Username	Username for connecting to the database.
Password	Password for connecting to the database.
Hive MetaDB	Select whether to enable the Hive metadata database. To get Iceberg databases, tables, and other metadata, Hive MetaDB must be enabled.
Metastore URL	Service address where Hive metadata is stored, used to access the metadata storage
Hive MetaDB IP and port	After enabling Hive MetaDB, you need to configure the IP and port. The IP and port are used to specify the location and access port of the Hive MetaDB server
Database name	After enabling Hive MetaDB, you need to configure the database name created in Hive MetaDB for storing Hive metadata
Username	Account name for the database connected to Hive MetaDB
Password	The password is the account password associated with the aforementioned username. When configuring Hive, you need to provide the correct password to verify identity and authorize access to MetaDB.
Warehouse path	The location of Iceberg table data storage, usually a directory path on HDFS
core-site.xml file	The core configuration file of Hadoop, containing runtime settings of the Hadoop system, such as HDFS access control. **This parameter is required!**
hdfs-site.xml file	The HDFS configuration file of Hadoop, used to specify related configurations of HDFS, such as replication strategy, permission settings, etc. **This parameter is required!**
hive-site.xml file	The configuration file of Hive, containing runtime configuration information of Hive, such as connection parameters, storage paths, etc. **This parameter is required!**
Authentication Method	Specify the authentication method when connecting to Hive, which can be no authentication or using Kerberos authentication
keytab file	When Kerberos authentication is enabled, the keytab file needs to be configured. The keytab file contains the key table for Kerberos authentication.
conf file	When Kerberos authentication is enabled, the conf file needs to be configured. The conf file is the configuration file for Hive, containing various configuration parameters required for Hive operation, including Kerberos-related configuration information.
principal	When Kerberos authentication is enabled, the principal of the Hive service also needs to be specified so that Kerberos can verify the identity of the Hive service.
Data Connectivity (Legacy)	Test whether the configured database is connected. **Note** • If the connectivity test fails, the data source can still be saved. However, it cannot be used if the connectivity test fails and the data source is saved.

©2013-2025 Tencent Cloud. All rights reserved.

- After the Chinese New Year in 2024, newly purchased WeData does not include this feature by default. This feature has been upgraded to Resource Group Connectivity.

Resource Group Connectivity (New)	Support the connectivity test between specific resource groups and data sources. Users only need to ensure that their data source and the VPC network where the Integrated Resource Group is located are connected without additional network interconnection.  **Note** Users who purchased WeData before the Chinese New Year in 2024 do not have this feature. If you need to use it, please contact Tencent's operation and maintenance personnel.

5. Real-time Whole Database: MySQL to Iceberg

When an enterprise needs to migrate an entire database to a new platform or environment, whole database synchronization can synchronize all tables and data at once, ensuring data integrity and consistency.

Step 1: Create a whole database migration task

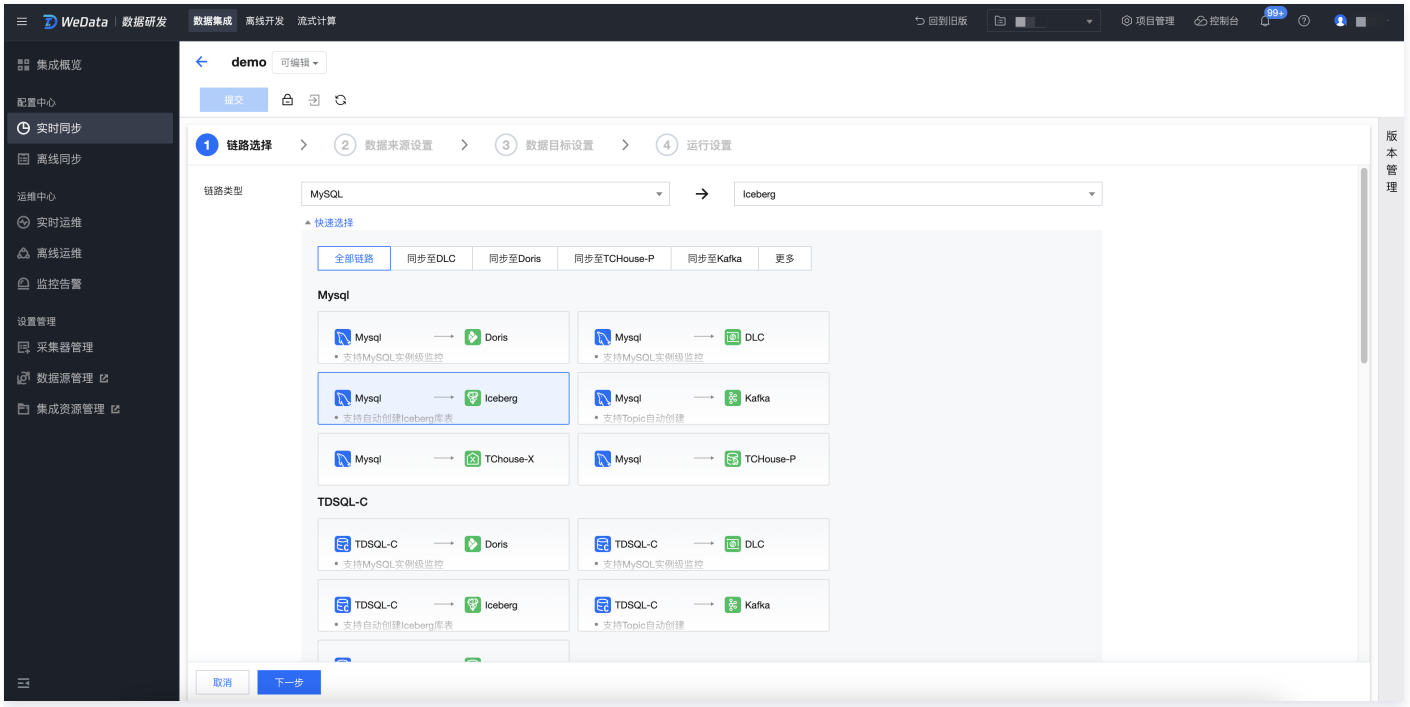
Log in to WeData, open the DataInLong module, then click **Configuration Center** > **Real-time Sync Task** page, and click **New** whole database migration task.



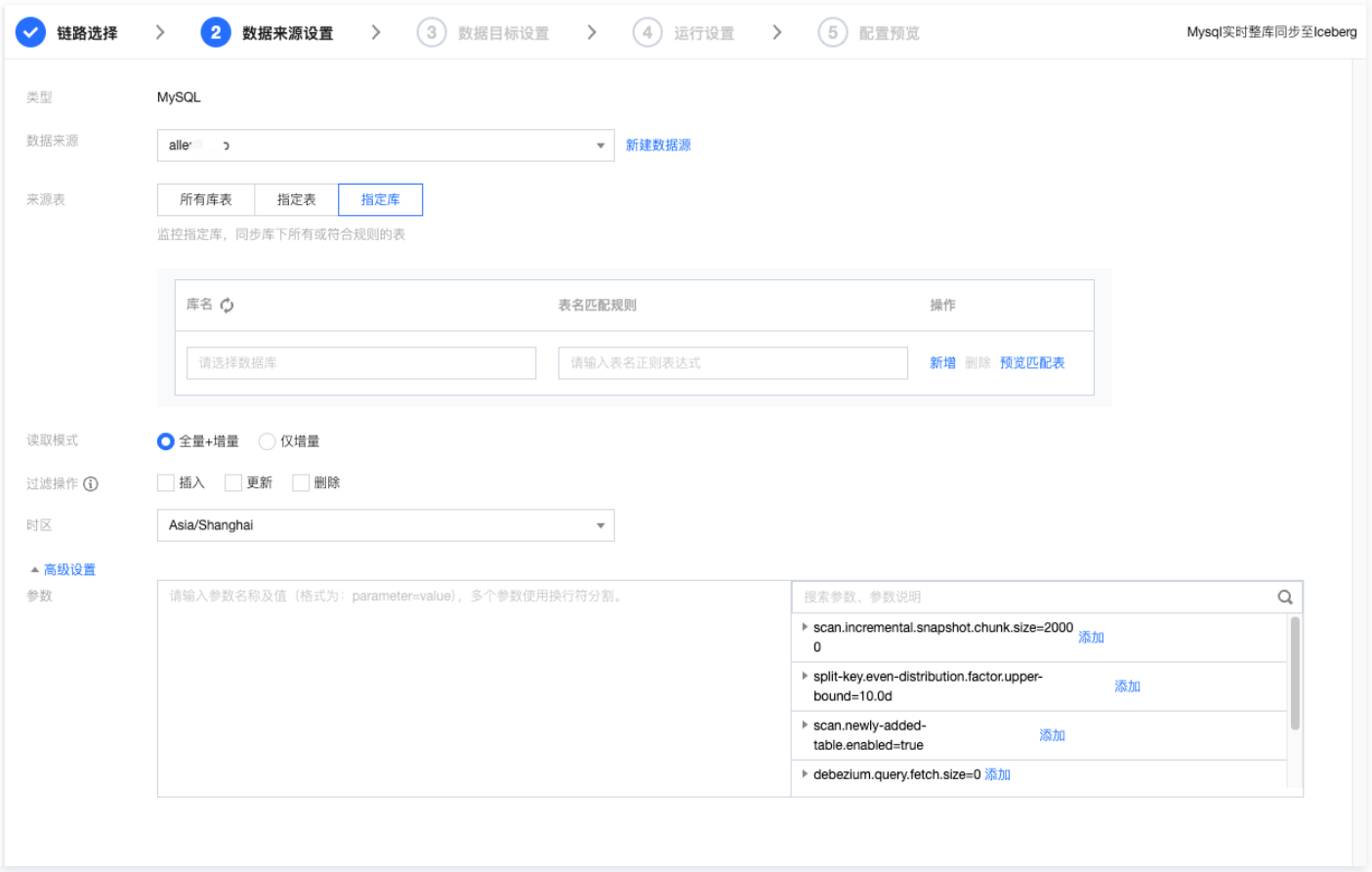
Step 2: Link Selection

On the link selection interface, select **MySQL** > **Iceberg link**, after selection, click **Next**.

©2013-2025 Tencent Cloud. All rights reserved.

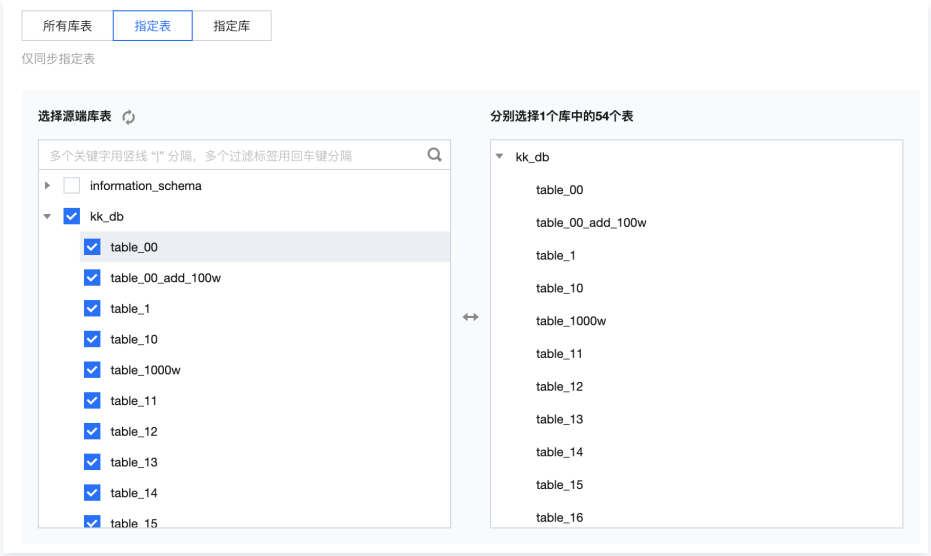


Step 3: Data Source Settings



Parameter	Description
Data Source	Select the MySQL data source for synchronization.
Source Table	All databases and tables: Monitor all databases under the data source. During the task execution, new databases and tables will be synchronized to the target end by default.

©2013–2025 Tencent Cloud. All rights reserved.

	Specified Table: Under this option, you need to specify the specific table name. After setting, the task will only synchronize the specified table; if a new table needs to be synchronized, the task must be stopped and restarted.  Specify Database: Under this option, you need to specify the specific database name with a table name regular expression. After setting, new tables that match the table name expression will be synchronized to the target end by default during task execution.  Common table name regular expression: • All tables in a single database: .* • Specific tables in a single database: table1|table2|table3 • Exclude specific tables: ^(?!(exclude1|exclude2|exclude3)$).*$ • Exclude specific tables but include all tables starting with 'a': ^(?!(exclude1|exclude2|exclude3)$)(^a.*)$
Read mode	• Full + Incremental: Data synchronization is divided into full and incremental synchronization phases. After the full phase is completed, the task enters the incremental phase. The full phase will synchronize historical data in the database, and the incremental phase will start synchronizing from the binlog CDC location after the task starts. • Incremental: Only synchronize data from the binlog CDC location after the task starts.
Filter operations	Support three operations: Insert, Update, and Delete. After setting, data of the specified operation types will not be synchronized.
Time zone	Set the log time timezone, default is Shanghai.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

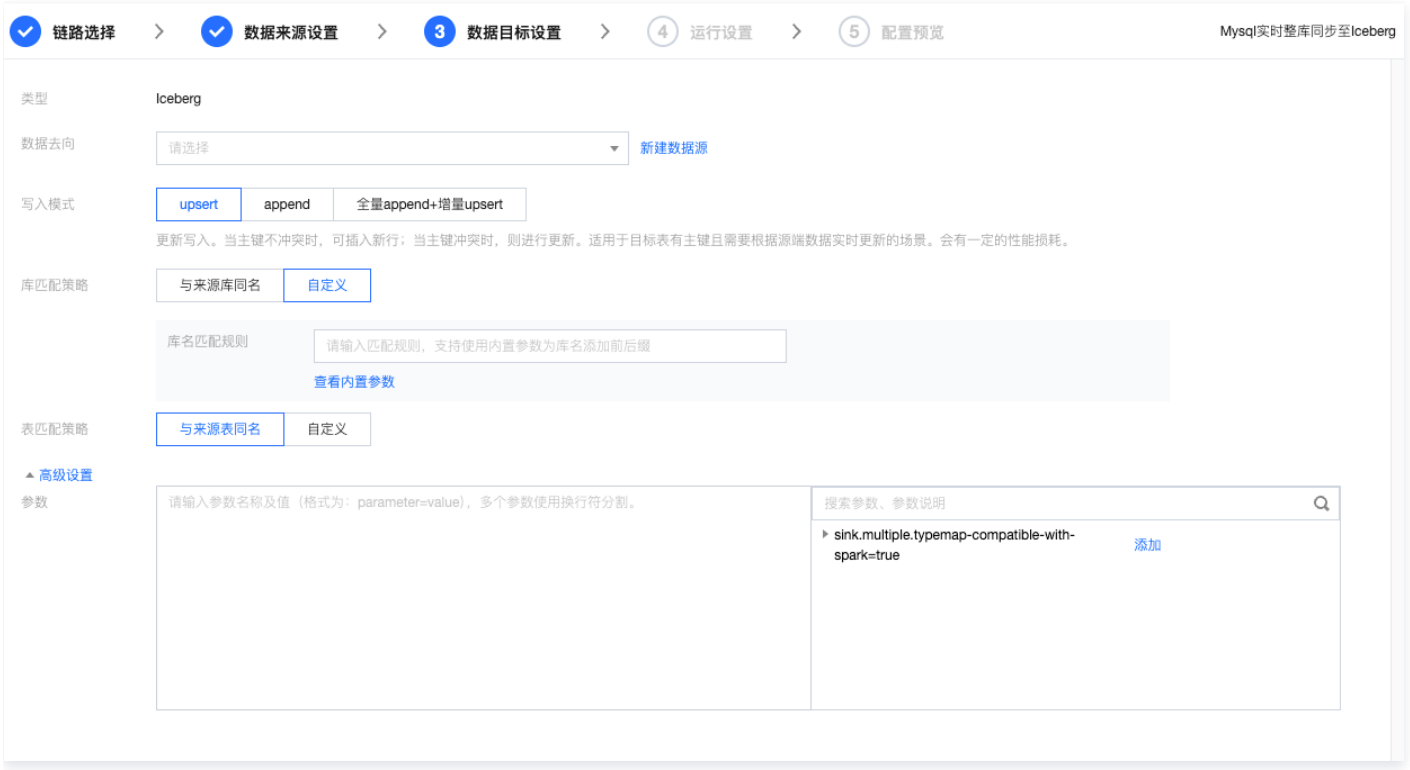
During MySQL full database synchronization, tasks use primary keys or specified primary keys (if no primary key, specify in advanced settings; currently, only index columns are supported) for splitting. Currently, the following types are supported:

> **Primary key support range type:**

©2013–2025 Tencent Cloud. All rights reserved.

TINYINT,TINYINT_UNSIGNED,SMALLINT,SMALLINT_UNSIGNED,INT,MEDIUMINT,INT_UNSIGNED,MEDIUMINT_UNSIGNED,BIGINT,BIGINT_UNSIGNED,FLOAT,DOUBLE,DECIMAL,TIME,DATE,DATETIME,TIMESTAMP,CHAR,VARCHAR,TEXT,BINARY,VARBINARY,BLOB.

Step 4: Data Target Settings



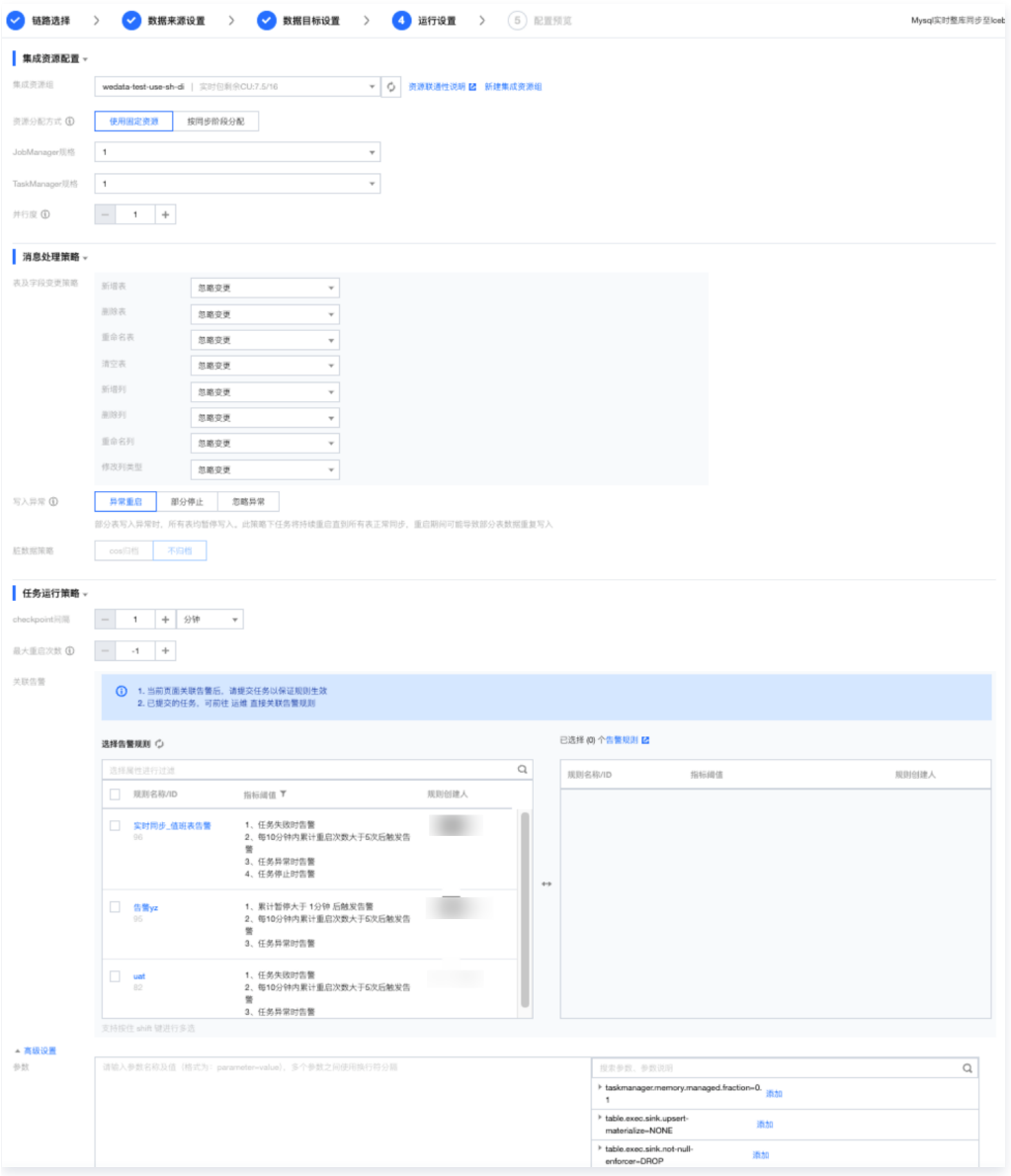
Parameter	Description
Data Source	Select the target data source for synchronization.

Write mode	• upsert: Update mode writes to the target table. This requires a primary key to be set in the target table:
	  ○ The task will use the primary key as the unique key to update records by default.
	  ○ If the table has no primary key, data will be written in append mode.
	• append: Append mode writes to the data table.
	• Full append + incremental upsert: Automatically switches data write methods based on the source data synchronization phase. The full phase uses append for better performance, and the incremental phase uses upsert for real-time data updates. This mode currently supports MySQL, Oracle, and TDSQL MySQL data sources.

Library and Table Matching Strategy	Name matching rules for database and data table objects in Iceberg.
Advanced Settings	Configuration parameters can be set according to business needs. Refer to Chapter 7 Real-time Node Advanced Parameters.

Step 5: Run Settings

©2013-2025 Tencent Cloud. All rights reserved.



Integration Resource Configuration:

- Associate the current task with the corresponding integration resource group and set the runtime JM, TM specifications, and task parallelism. The actual CU count used during task execution = JobManager specification + TaskManager specification × parallelism.
- Resource Allocation Method: Integration resources support multiple allocation methods
  - Fixed Allocation: In this method, resource allocation remains constant throughout the full and incremental synchronization phases, preventing resource preemption among tasks. It is suitable for scenarios where data may significantly change during task execution.
  - Allocation by synchronization phase: Allocates planned resource usage based on different synchronization phases (full and incremental) to save overall resource usage.
- Table and Field Change Strategy:

Table and Field Change Type	Support Policies
Add Table	Ignore Change, Automatic Table Creation, Log Alerts, Task Error
Deleting a table	Ignore Change, Log Alerts, Task Error
Rename Table	Ignore Change, Log Alerts, Task Error
Clearing table data	Ignore Change, Log Alerts, Task Error

©2013–2025 Tencent Cloud. All rights reserved.

Add Column	Automatically Add Columns, Ignore Change, Log Alerts, Task Error
Delete Column	Automatic Column Deletion, Ignore Change, Log Alerts, Task Error
Rename Column	Automatic Column Renaming, Ignore Change, Log Alerts, Task Error
Modify Column Type	Automatic Column Type Modification, Ignore Change, Log Alerts, Task Error

The following are the policies and display effects supported in the table above:

Policy name	Display Effect
Ignore Change	1. No change to target table/field 2. Task is running normally
Logging and Alarming	1. No change to target table/field 2. The task runs normally, changes are ignored by default
Stop the task	1. No change to target table/field 2. Abnormal Task Restart
Automatic Processing	The target end automatically changes according to the source end and automatically creates target objects. For example, adding tables, adding columns, etc. **Note** To avoid structural changes due to misoperation that could impact user business, only related changes such as adding tables/columns are currently supported for automatic processing; other high-risk operations like deleting columns, clearing tables, and dropping tables are only supported by "reminder-style" response policies (e.g., alarms, stop)

Write Exception Policy:
Iceberg whole database synchronization tasks provide task-level runtime resources and data write failure handling strategies. Four types of data write failure handling strategies are supported:

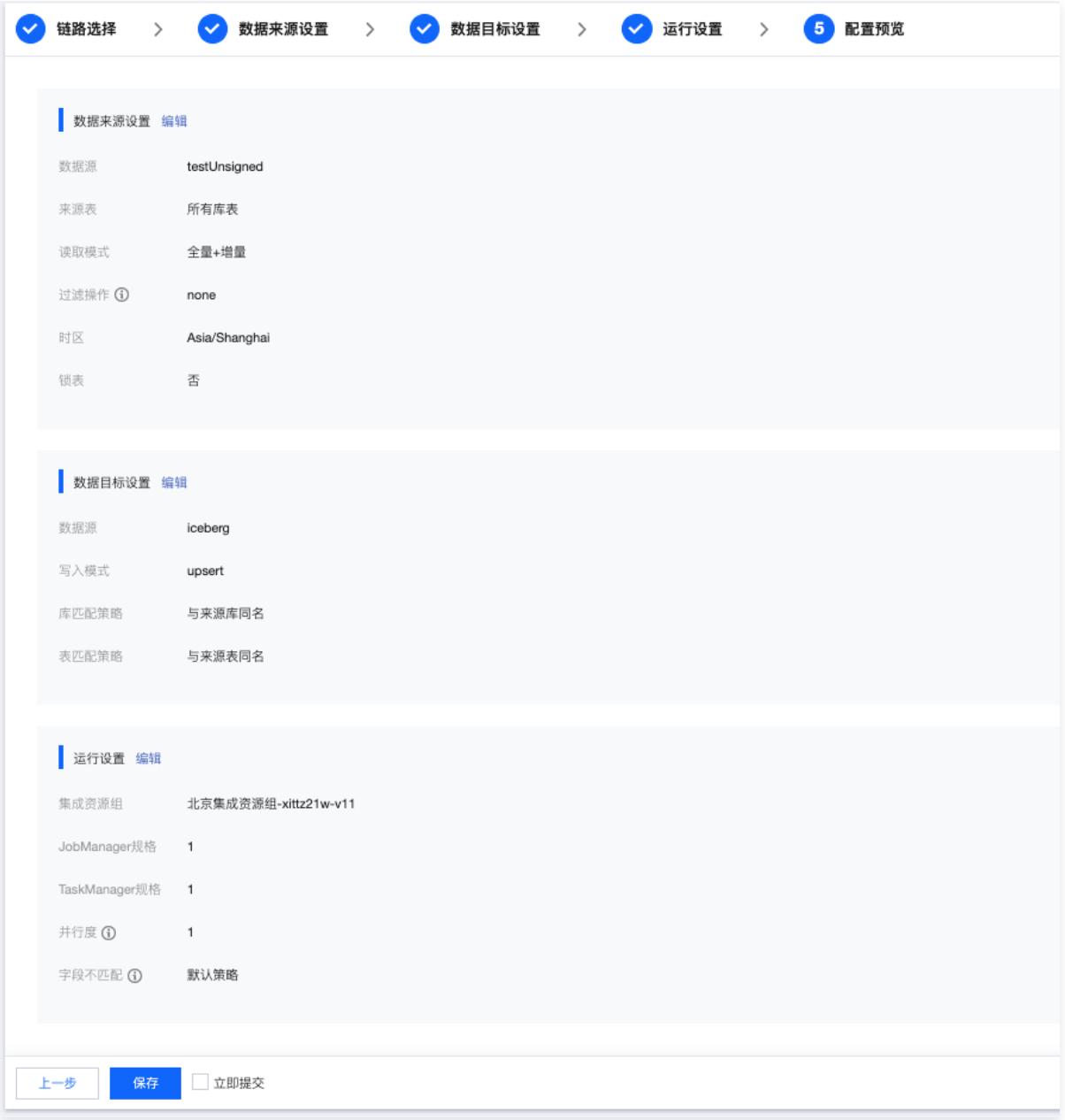
Policy name	Policy Description
Partial Stop	When a table write exception occurs, only the data writing to that table will be stopped; other tables will continue to synchronize normally. The stopped table cannot resume writing during this task's operation.
Abnormal Restart	When a table write exception occurs, all tables will pause writing. The task will continuously restart under this policy until all tables synchronize normally. During the restart period, some table data may be redundantly written.
Ignore Exception	Ignore the data that cannot be written within the table and mark it as dirty data. Other data in the table, as well as other tables in the task, will synchronize normally. Dirty data can be archived to COS or not archived. • COS Archiving: Archives dirty data that cannot be written. Requires configuration of COS data source, bucket, storage directory, content separator, and line break. • Non-Archiving: No additional operations required.

Scenario Example:
In Task Task1, 50 tables are planned to be synchronized. During the task operation, Table A has new fields added or field type changes:
- Partial Stop: After running the task for Table A, the field "DEMO" underwent a type change, and the new field type cannot match the target end type for writing. Under this policy, the task will stop Data Reading from the source table A, and subsequent tasks will only synchronize the remaining 49 tables to the target end.
- Abnormal Restart: After running the task for Table A, the field "DEMO" underwent a type change, and the new field type cannot match the target end type for writing. Under this policy, the task will continuously restart, and all 50 tables configured in the task will pause data write until the field in Table A is corrected.

©2013-2025 Tencent Cloud. All rights reserved.

- Ignore Abnormal: After running the task for Table A, the field "DEMO" underwent a type change, and the new field type cannot match the target end type for writing. Under this policy, the task will ignore the abnormal data that cannot be written, mark it as Dirty Data, and synchronize the other data in the table normally.

Step 6. Configure Preview



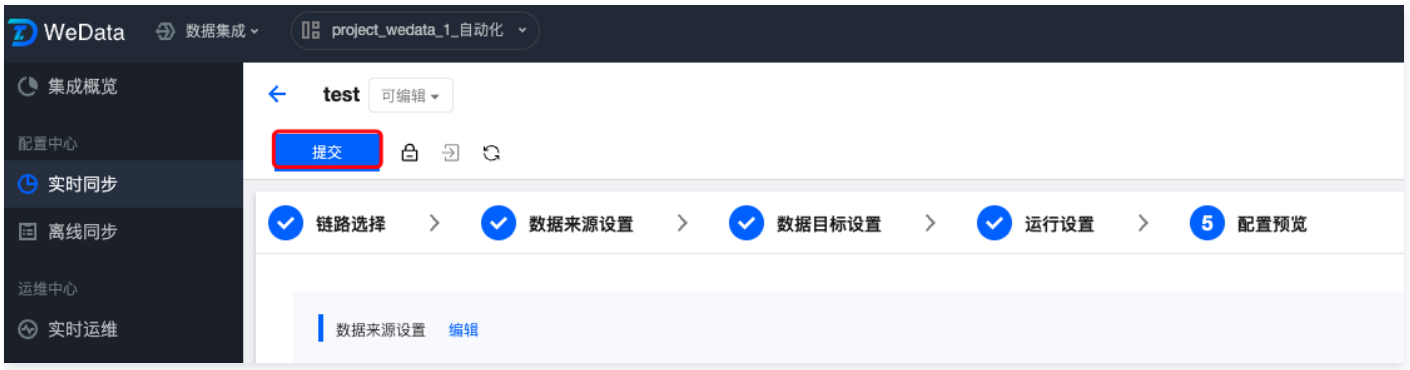
Serial number	Parameter	Description
1	Submit	Submit the current task to the production environment. On submission, different operation strategies can be selected based on the presence of production tasks • If there are no effective online tasks for the current task, i.e., it's the first submission or the online task is in a "failed" status, you can submit it directly. • If there are "Running" or "Paused" online tasks, you need to choose different strategies. Stopping the online job will abandon the previous task's running location, consuming data from the beginning. Retaining the job status will resume from the last consumed location after restarting.

©2013–2025 Tencent Cloud. All rights reserved.

		 **Note** Click **Start Now** to immediately run the task after submission, otherwise, it needs to be manually triggered to start running.
2	Lock/Unlock	The default creator is the first lock holder, and only the lock holder is allowed to edit task settings and run tasks. If the lock holder does not make any editing operations within 5 minutes, others can independently click the icon to acquire the lock. Upon successfully acquiring the lock, editing operations can be performed.
3	Go to Operations	Quickly jump to the task operations page based on the current task name.
4	Save	After the preview is complete, you can click the **Save** button to save the whole database task configuration. If only saved, the task will not be submitted to the operations center.

Step 7: Task Configuration Detection and Submission

After completing the above six steps, you can perform task configuration detection and officially submit the task. At this point, you can click the **Submit** button in the upper left corner to proceed to the task configuration check and submission stage.



Steps	Step Description	
Task Configuration Detection	This step involves detection of the read end, write end, and resources in the task: • Detection passed: No configuration errors. • Detection failed: Configuration has issues that need to be fixed for further configuration. • Detection Alert: This is a system-recommended modification item. After modification, click **Retry** to recheck, or you can click **Ignore Exceptions** to proceed without blocking further configuration.	
Submission Policy Selection	In this step, you can choose the submission policy for this task: • **First Submission: The first submission supports synchronizing data from the default or specified point.**	1. First Submission

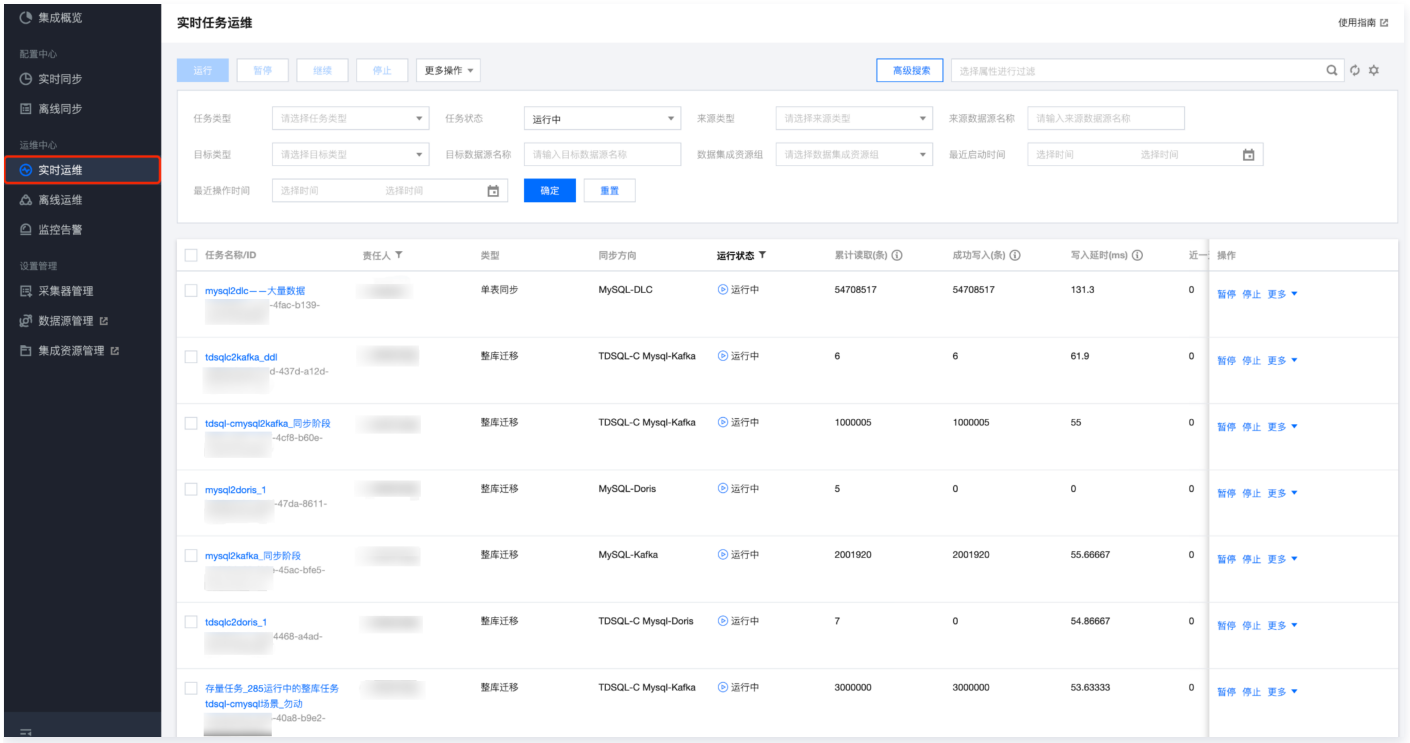
©2013–2025 Tencent Cloud. All rights reserved.

	• **Start immediately and synchronize from the default position:** If the source end is configured to read as "Full + Increment", it will first synchronize the existing data (full stage) and then consume binlog to get the changed data (incremental stage). If the source end is set to read as "Increment Only", it will use the latest binlog position by default. • **Start immediately and synchronize from the specified time point:** The task will synchronize data according to the configured time and timezone. If the specified time location is not found, the task will start synchronizing from the earliest binlog position by default. If the source end is configured to read as "Full + Increment", the task will skip the full stage and start synchronizing from the specified time position in the incremental stage. • **Do not start now:** After submission, the task will not start running for the time being, and can be manually started later in the operation list. • **Non-first submission:** Supports starting or continuing the task with running status • **Continue running:** Under this strategy, the new version of the task will continue running from the last synchronized position after submission. • **Restart and synchronize from the specified position:** Under this strategy, you can specify the position to start reading. The task will ignore the old version and restart reading from the specified position. If the specified time position is not found, the task will start synchronizing from the earliest binlog position by default. • **Restart and run from the default position:** Under this strategy, it will read from the default position based on the source end configuration. If the source end is configured to read as "Full + Increment", it will first synchronize the existing data (full stage) and then consume binlog to get the changed data (incremental stage). If the source end is set to read as "Increment Only", it will use the latest binlog position by default. Each submission will generate a new version of the real-time task. You can configure the version description in the dialog box.	 2.Non-First Submission 
Submitting the job	After submission, you can click to go to operations to view the task running status.	

Step 8: Real-time Task Operations and Maintenance

For real-time tasks that have been configured and submitted, you can enter real-time operations to view and operate the corresponding tasks. You can check the basic running status, task running metrics, running logs, and configure abnormal alarms in real-time operations. This document enumerates common operations and maintenance for real-time synchronization tasks.
The operations list supports managing and controlling all submitted tasks, including start and stop operations.

©2013-2025 Tencent Cloud. All rights reserved.



Real-time task status and meaning:

Status Types	Description
Initialization	Task has been submitted but has not yet started running.
Operating	Status is being changed. This status often appears shortly after completing task operations and the task is in the process of status changing.
Running	Task is running.
Suspended	The current task has been suspended, and both the task status and read-end checkpoint are preserved.
Stopped	Task has been manually stopped. **Note** If a user manually stops a task, it will switch to this status. Other non-exception cases of stopping.

Real-time task supported operations and meanings:

Operation Type	Description
Running	Run the current task to start reading and writing
Stop	Manually stop the current task. Once stopped, the current task status and location will not be preserved
Pause	Pause the current task. After pausing, the task status and location are preserved. Subsequent operations can resume from the completed location.
Resume	Resume from the last paused location. **Note** It is recommended not to pause tasks for too long, otherwise restarting may fail due to expired source end logs or lost locations.

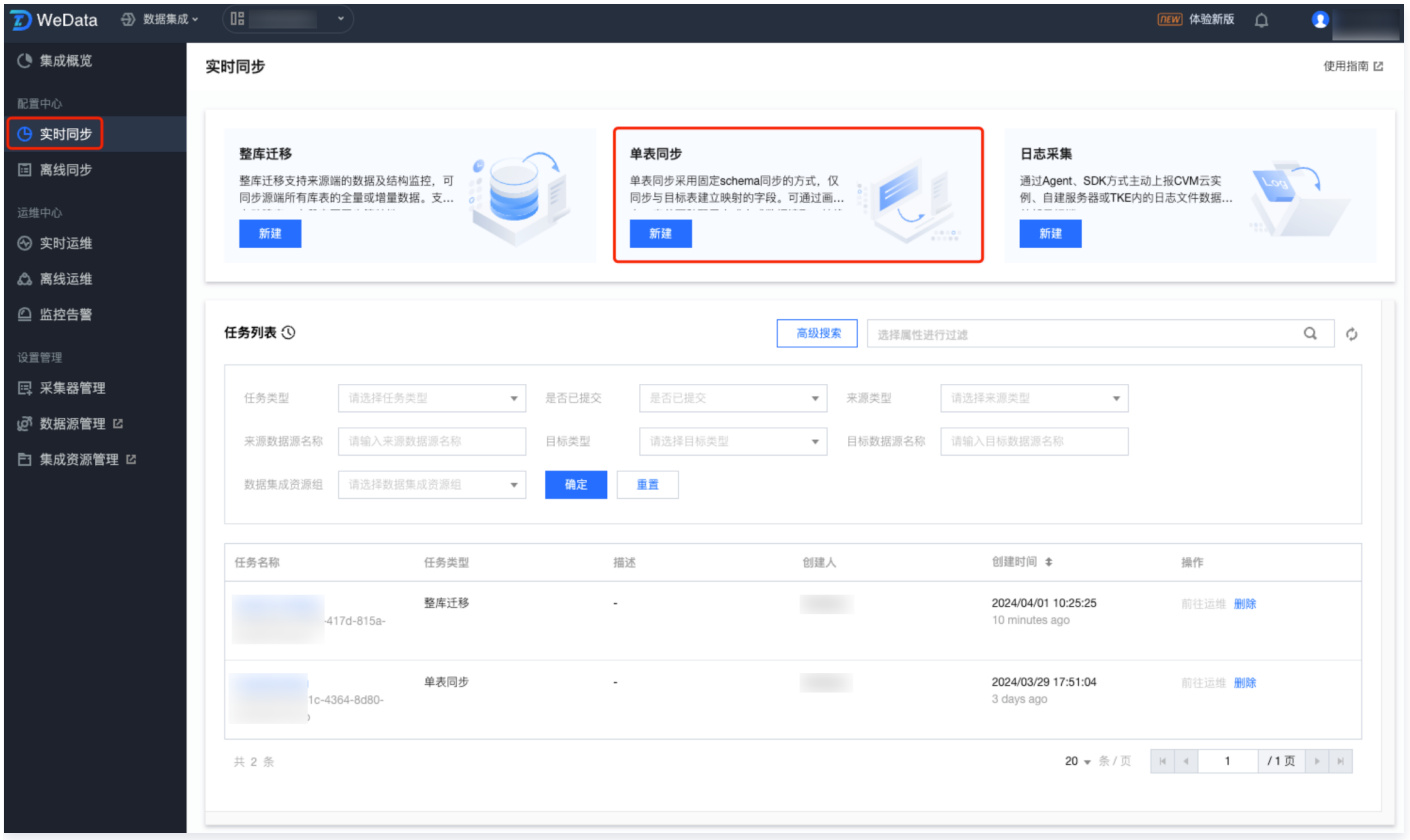
©2013-2025 Tencent Cloud. All rights reserved.

Real-time task status and corresponding allowed operations:

Status Types	Operation			
	Running	Pause	Resume	Stop
Initialization	✓	–	–	–
Operating	–	–	–	–
Running	–	✓	–	✓
Suspended	–	–	✓	✓
Stopped	✓	–	–	–

VI. Real-time single table: MySQL to Iceberg sync

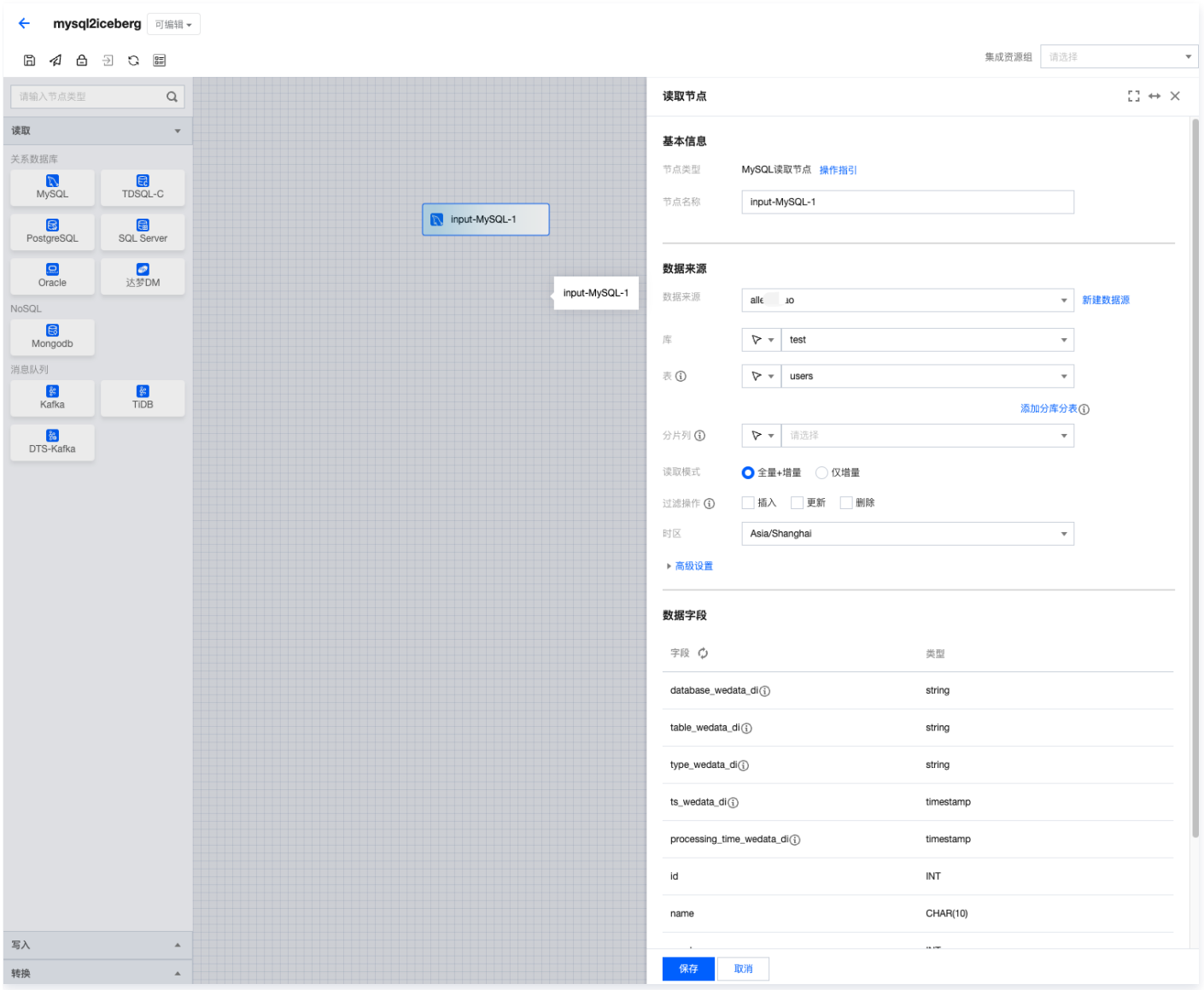
When the business needs to focus only on a specific table, single table synchronization can target the data of that table. This reduces unnecessary data processing and transmission, improving sync efficiency.
On the DataInLong page, click **Real-time Synchronization** in the left navigation bar. On the Real-time Synchronization page, select Single Table Sync at the top to create a new sync and enter the configuration page. For single table sync, two modes are supported:**Single Table Mode** and **Canvas Mode**. The subsequent guide will use Canvas Mode to explain the configuration process.



Step 1: Configure MySQL Read Node

1. Click **Read on the left side**, click to select the MySQL node, and configure the node information.

©2013–2025 Tencent Cloud. All rights reserved.



2. Fill in the parameter information. You can refer to the table below for parameter configuration.

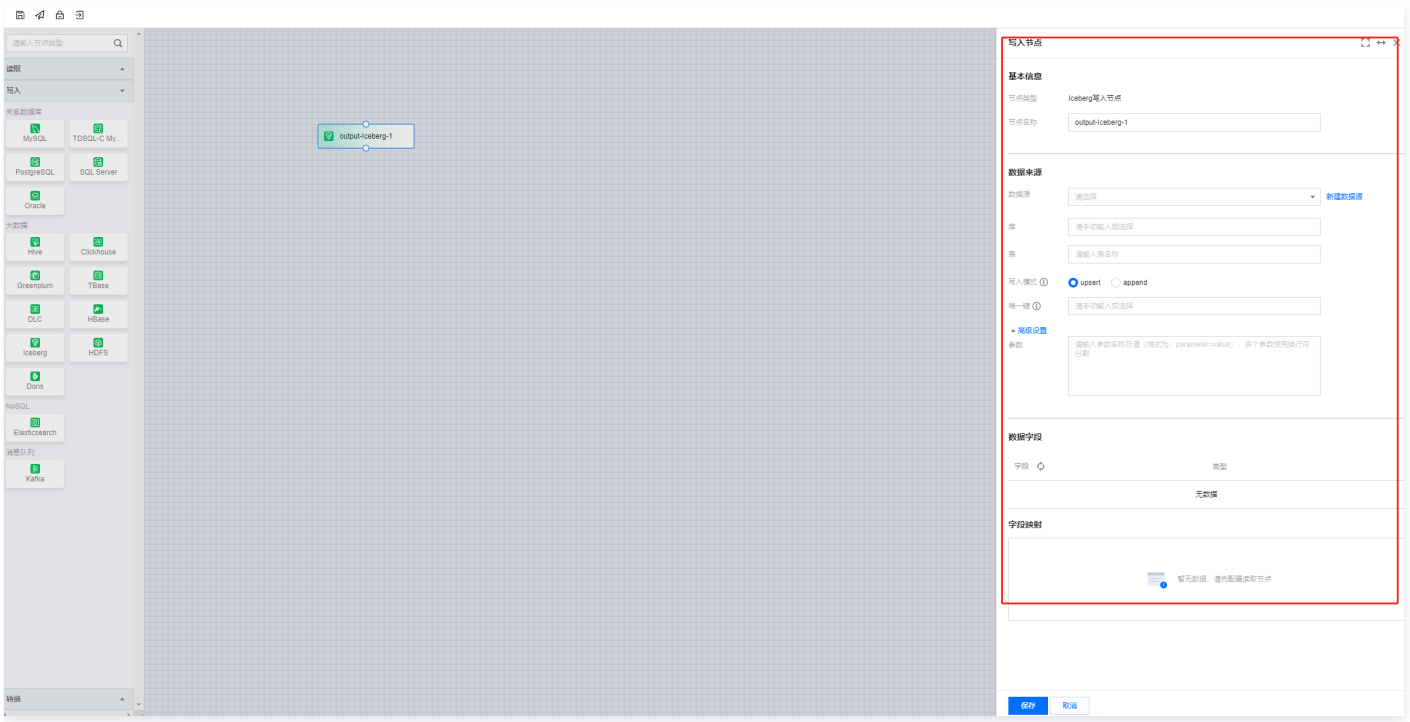
Parameter	Description
Node name	Enter MySQL node name
Data Source	Select the data source where the table to be synced is located
Database	Select the database where the table to be synced is located
Table	Multiple tables can be selected; ensure schema consistency across tables
Table primary key	In sharding mode, default table schema is consistent. The system will use the primary key of the first table; please select or enter the table primary key field name
Format	Specify MySQL log encoding format (utf-8, gbk, Latin1, utf8mb4)
Read mode	Both Full and Incremental modes are supported
Filter operations	After setting, it will not sync data of specified operation types; supports Insert, Update, and Delete

3. Preview data fields, click **Save**.

Step 2: Configure Iceberg Write Node

©2013–2025 Tencent Cloud. All rights reserved.

1. Click **Write on the left side**, click to select the Iceberg node, and configure the node information.



2. Fill in the parameter information. You can refer to the table below for parameter configuration.

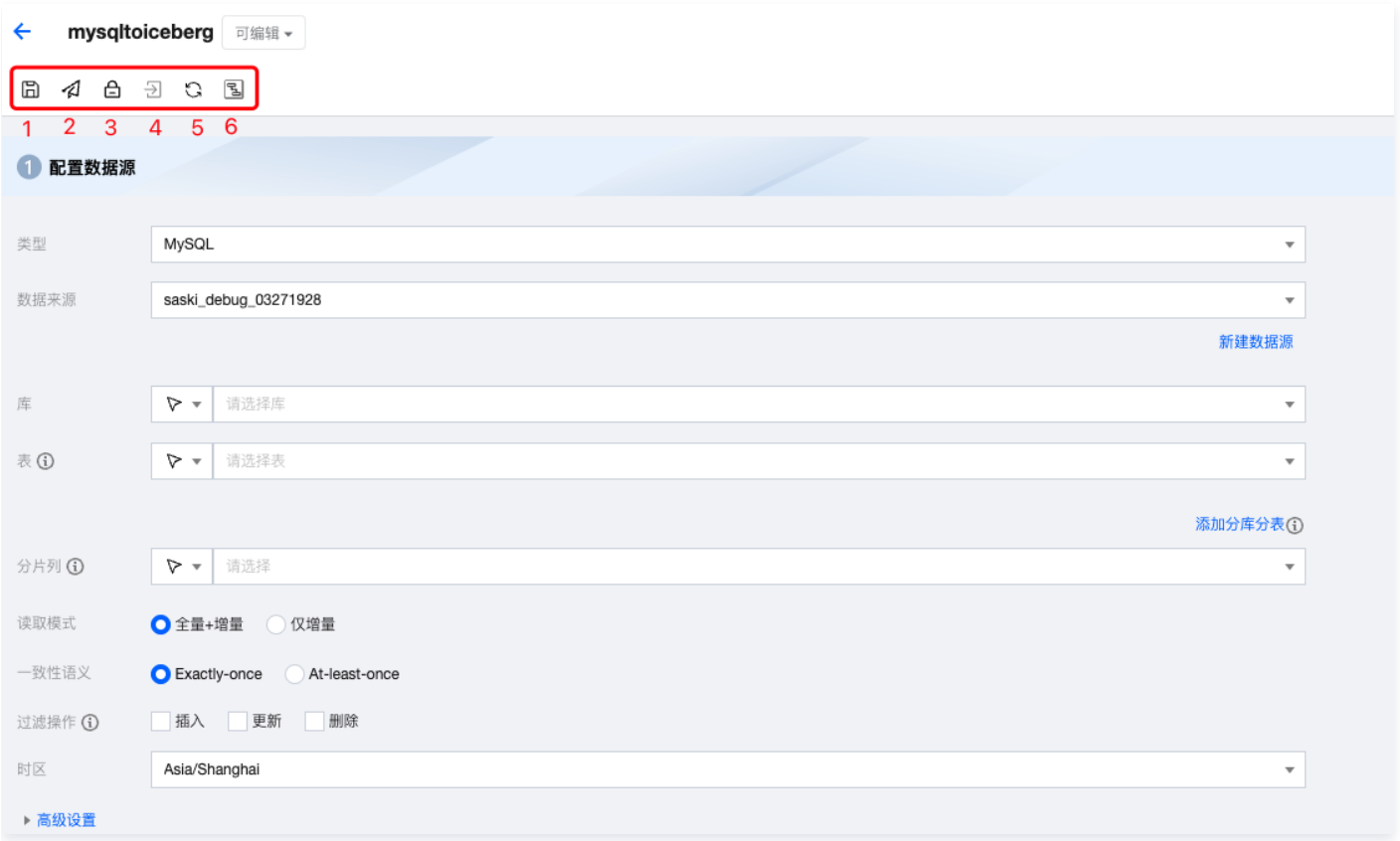
Parameter	Description
Data Source	Iceberg data source to be written.
Database	Supports selecting or manually entering the name of the database to be written to. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Supports selecting or manually entering the name of the table to be written to. When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Write mode	Iceberg real-time sync write supports two modes: append: appends data, throws an error on primary key conflict. upsert: updates data, updates on primary key conflict.
Unique key	Select a field as the primary key for the write table.
Advanced Settings	Configure parameters based on your business needs. You can refer to section VII Advanced Parameters for Real-time Nodes in this document.

3. Preview data fields and map them with the fields configured in the read node, click **save**.

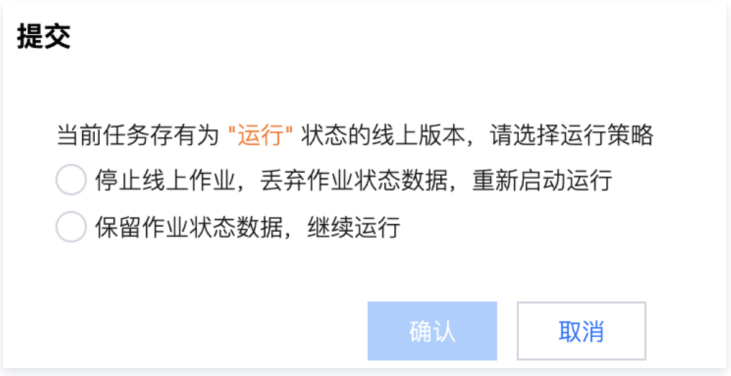
Step 3: Task Configuration Check and Submission

After completing the above steps, you can proceed with a task configuration check and formally submit the task. At this point, you can refer to the related buttons in the upper left corner of the task configuration interface for subsequent submission operations. The position and features of each button are as follows:

©2013-2025 Tencent Cloud. All rights reserved.



After completing the configuration of a real-time synchronization task, you can configure the operation strategy and submit it to the production environment for running. Currently, you can save, submit, lock/unlock, and proceed to operations on the task configuration page.

Serial number	Parameter	Description
1	Save	Save the current task configuration information, including data node configuration, node connection, task attributes, and task scheduling configuration.
2	Submit	Submit the current task to the production environment. At the time of submission, different operation strategies can be selected based on whether there is an active production task associated with the current task. • If there are no effective online tasks for the current task, i.e., it's the first submission or the online task is in a "failed" status, you can submit it directly. • If there are "Running" or "Paused" online tasks, you need to choose different strategies. Stopping the online job will abandon the previous task's running location, consuming data from the beginning. Retaining the job status will resume from the last consumed location after restarting.  Note

©2013-2025 Tencent Cloud. All rights reserved.

		Click **Start Now** to immediately run the task after submission, otherwise, it needs to be manually triggered to start running.
3	Lock/Unlock	The default creator is the first lock holder and only the lock holder is allowed to edit the task configuration and run the task. If the lock holder does not perform any edit operations within 5 minutes, others can click the icon to seize the lock. Once successful, they can proceed with editing.
4	Go to Operations	Quickly jump to the task operations page based on the current task name.
6	Canvas Conversion	Switch the current edit mode to Canvas Mode

Task Submission Detection



Parameter	Description
Detection Abnormal	Supports skipping anomalies for direct submission, or halting the submission.
Detection contains only warnings and below	Can be submitted directly.

Submit Results



- Task Submitting:
  - Shows the submission progress percentage.
  - Prompt the user not to refresh/close the page. Message: The current task has been successfully submitted, you can go to OPS for task status and data management.
- Task Submission Result – Success:

©2013–2025 Tencent Cloud. All rights reserved.

- Show task submission successful result.
- Prompt for successful submission and subsequent redirection: "Submission successful, you will be redirected to the current task's operation and maintenance detail page in 10 seconds" "The current task has been successfully submitted, you can go to operations and maintenance for task status and data management".

- Show the reason for task submission failure:
  - Return the failure reason.

Step 4: Real-time Task Operation and Maintenance

For real-time tasks that have been configured and submitted, you can enter real-time operations to view and operate the corresponding tasks. You can check the basic running status, task running metrics, running logs, and configure abnormal alarms in real-time operations. This document enumerates common operations and maintenance for real-time synchronization tasks.
The operations list supports managing and controlling all submitted tasks, including start and stop operations.
Refer to Section 5 in this article Real-time Whole Database: MySQL to Iceberg and Step 8 Real-time Task Operation and Maintenance.

7. Real-time Node Advanced Parameters

When configuring read-write nodes, you can configure advanced settings. The configuration and usage scenarios are as follows:

MySQL Node Level

The following are the advanced settings supported by the MySQL read nodes under usage scenarios and their descriptions:

Read/Write	Applicable Scenario	Configuration:	Description
Read	Single Table + Whole Database	scan.newly-added-table.enabled=true	**Parameter Description:** Set this parameter to sense new tables after **Suspension** > **Continue**. The default is false 1. Use this parameter during full incremental synchronization. New tables will read the stock data first, then read the incremental data 2. Use this parameter during incremental synchronization. New tables will only read incremental data
Read	Single Table + Whole Database	scan.incremental.snapshot.chunk.size=20000	**Parameter Description:** For tasks with uniform data distribution, this parameter represents the approximate number of records in one chunk. You can estimate the number of chunks by dividing the total data volume by the chunk size. The number of chunks affects whether jobmanager experiences an out-of-memory error. Currently, with 2CU, it can support more than 100,000 chunks. If the data volume is too large, you can increase the chunk size to reduce the number of chunks **Precautions:** For large data volume tasks (e.g., total data volume over 100 million, single record larger than 0.1M), it is generally recommended to set 20,000
Read	Single Table + Whole Database	split-key.even-distribution.factor.upper-bound=10.0d	**Parameter Description:** During the MySQL existing data reading phase, if the data is quite scattered and the maximum value of the primary key field is extremely large, you can modify this parameter to use non-uniform splitting. This helps reduce the problem of having too many chunks due to the excessively large primary key value, which could lead to JM OOM issues **Precautions:** Default value is 10.0d, generally no need to modify

©2013-2025 Tencent Cloud. All rights reserved.

Read	Single Table + Whole Database	debezium.query.fetch.size=0	**Parameter Description:** Represents the number of records fetched from the database each time. The default value is 0, which indicates the JDBC default fetch size **Precautions:** 1. For large tasks (e.g., total data volume over 100 million, individual record size over 0.1M) with only one reading instance, it is recommended to set this to 1024 records 2. For tasks with multiple reading instances, it is recommended to lower this value to reduce memory consumption. Setting it to 512 records is advised
Read	Single Table + Whole Database	debezium.max.queue.size=8192	**Parameter Description:** This attribute defines the maximum number of events stored in the internal queue. If this limit is reached, Debezium will pause reading new events until the unprocessed events are handled and committed. This helps prevent an excessive backlog of events in the queue, thus avoiding memory exhaustion and performance degradation. The default is 8192 **Precautions:** 1. For large tasks (e.g., total data volume over 100 million, individual record size over 0.1M) with only one reading instance, it is recommended to set this to 4096 2. For tasks with multiple reading instances, it is recommended to lower this value to reduce memory consumption. Setting it to 1024 is advised

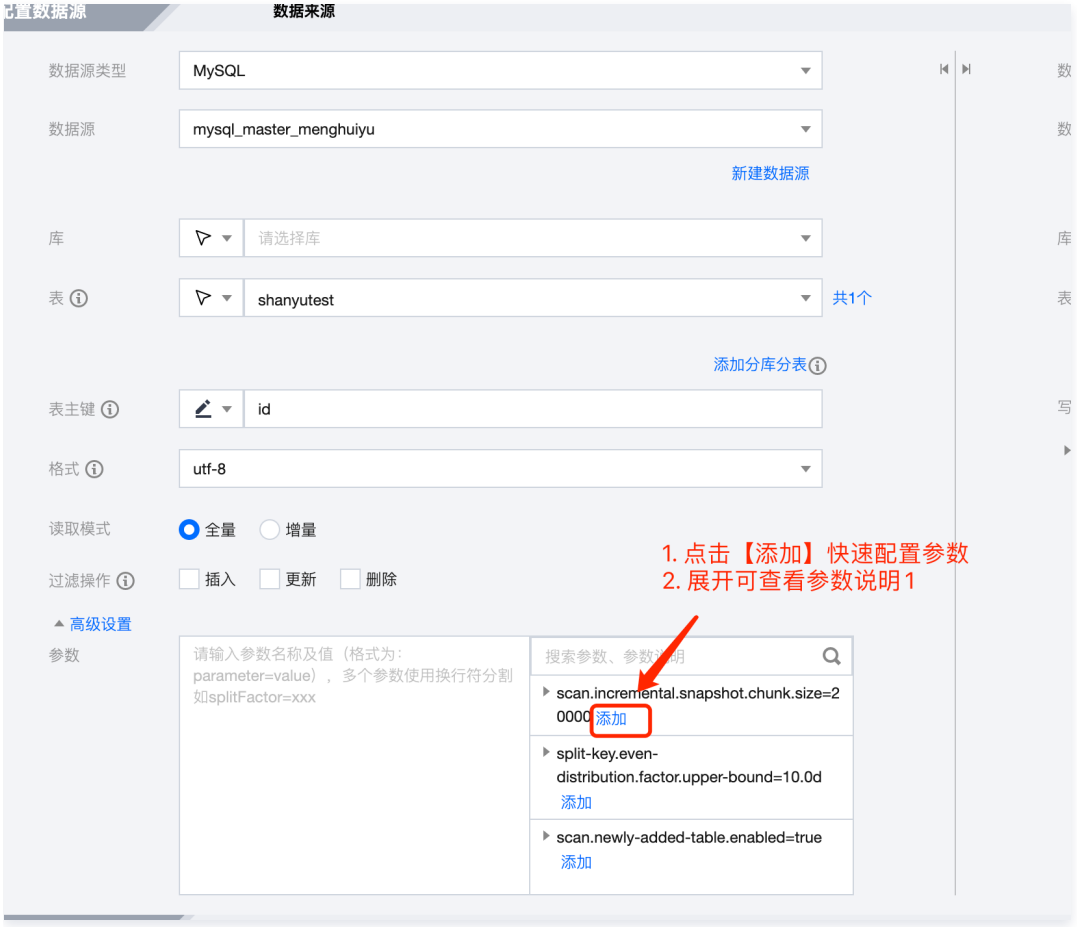
Iceberg Type Node Level

The following are advanced settings and their descriptions supported by Iceberg Write Nodes in usage scenarios:

Read/Write	Applicable Scenario	Configuration:	Description
Write	Entire Database	switch.append.upsert.enable	**Parameter Description:** Set this parameter, default 1. Use Append Mode for full-phase writes to reduce merge overhead 2. Switch to Upsert Mode writing in the incremental phase. Default is true.
Write	Entire Database	sink.partitioner	**Parameter Description:** Specify the data exchange method between Source and Sink sides. When the number of tables is large, use forward to improve performance. When the number of tables is less than the job concurrency number, set to rebalance to fully utilize multiple concurrent resources. Default is forward

When configuring advanced read-write node parameters, refer to the following diagram for guidance:

©2013-2025 Tencent Cloud. All rights reserved.



Note

1. One parameter per line; parameters that need to be used together should be written on the same line.
2. Each parameter comes with a default value.

8. FAQs

1. MySQL server ID conflict

Error message:

```
com.github.shyiko.mysql.binlog.network.ServerException: A slave with the same server_uuid/server_id as
this slave has connected to the master.
```

Solution: We have now optimized to randomly generate serverid. If in previous tasks the server-id was explicitly specified in MySQL advanced parameters, it is recommended to delete it, as multiple tasks may use the same data source with the same server-id, causing a conflict.

2. Binlog File Not Found Error:

Error message:

```
Caused by: org.apache.kafka.connect.errors.ConnectException: The connector is trying to read binlog
starting at GTIDs xxx and binlog file 'binlog.xxx', pos=xxx, skipping 4 events plus 1 rows, but this is no
longer available on the server. Reconfigure the connector to use a snapshot when needed.
```

Error Causes:
An error will occur if the binlog file being read by the job has already been cleaned up on the MySQL server. There are several reasons for Binlog cleanup, such as a too-short Binlog retention period, or the job processing speed not keeping up with the Binlog

©2013-2025 Tencent Cloud. All rights reserved.

generation speed, exceeding the maximum retention period of MySQL Binlog files. This results in the Binlog files being cleaned up on the MySQL server, causing the Binlog positions being read to become invalid.
Solution: If the job processing speed cannot keep up with the Binlog generation speed, you can consider increasing the Binlog retention period or optimizing the job to reduce backpressure and accelerate source consumption. If there are no anomalies in the job status, it might be that other operations on the database caused the Binlog to be cleaned up, leading to inaccessibility. You need to refer to the MySQL database-side information to determine the reason for Binlog cleanup.

3. MySQL Connection Reset

Error message:

```
EventDataDeserializationException: Failed to deserialize data of EventHeaderV4 .... Caused by:
java.net.SocketException: Connection reset.
```

Error Causes:

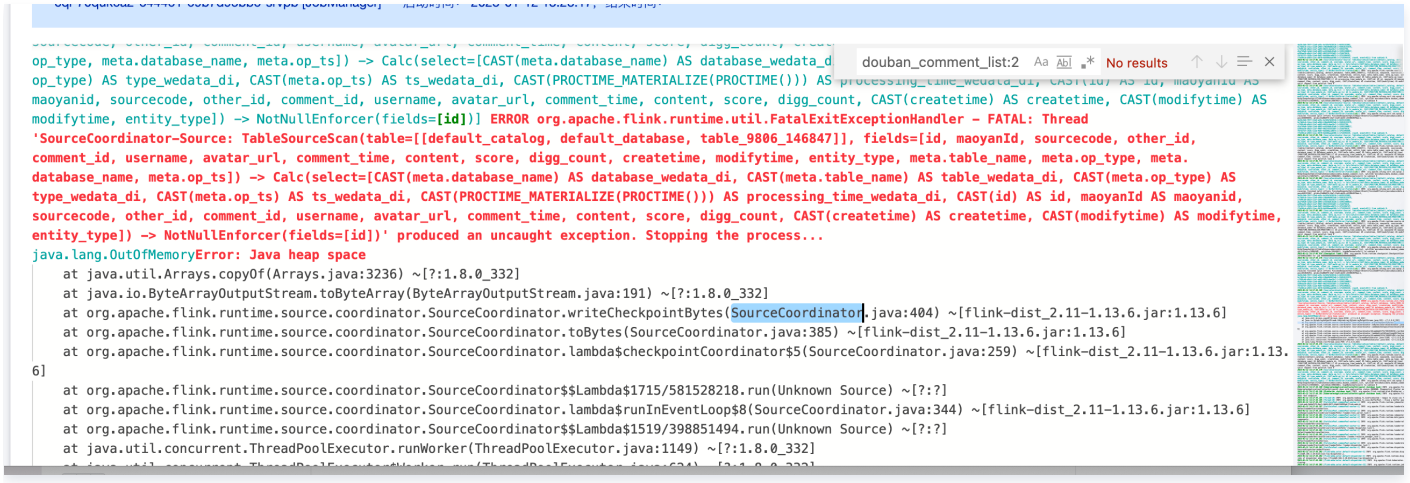
1. Network Issues.
2. The job has backpressure, causing the source to be unable to read data, and the binlog client to be idle. If the binlog connection remains idle after the timeout, the MySQL server will disconnect the idle connection.

Solution:

1. If it is a network issue, you can increase the MySQL network parameters set global slave_net_timeout = 120; (default 30s) set global thread_pool_idle_timeout = 120.
2. If the backpressure is caused by the job, you can alleviate it by adjusting the job, such as increasing parallelism, improving write speed, and increasing TaskManager memory to reduce GC.

4. Mysql2dlc task JobManager OOM

Error message:



1. If the user data volume is large, you can increase the JobManager CU number and use the MySQL advanced parameter scan.incremental.snapshot.chunk.size to increase the chunk size, which defaults to 8096.
2. The user's data volume is not large, but there is a large difference between the maximum and minimum values of the primary key, resulting in many chunks being divided when using the equal chunk strategy. Modify the distribution factor so that user data follows an uneven data splitting logic, split-key.even-distribution.factor.upper-bound=5.0d. The default distribution factor has been modified to 10.0d.

5. The format of the user's binlog data is incorrect, causing a parsing exception in Debezium

Error message:

```
ERROR io.debezium.connector.mysql.MySqlStreamingChangeEventSource [] - Error during binlog processing.
Last offset stored = null, binlog reader near position = mysql-bin.000044/211839464.
2023-02-20 21:37:28.480 [blc-172.17.48.3:3306] ERROR io.debezium.pipeline.ErrorHandler [] - Producer
```

©2013-2025 Tencent Cloud. All rights reserved.

```
failure
io.debezium.DebeziumException: Error processing binlog event.
```

Solution:
After modifying binlog_row_image=full, you need to restart the database.

6. Is gh-ost supported?

Yes, it does not migrate the temporary table data generated by Online DDL changes. It only migrates the original DDL data executed using gh-ost in the source database. You can use the default or configure your own regular expressions for gh-ost shadow tables and unused tables.

©2013-2025 Tencent Cloud. All rights reserved.

MySQL Offline Synchronization to Hive Guidance Documentation

Last updated: 2024-09-05 17:47:21

This document will guide you step-by-step through the process of offline synchronization from MySQL to Hive. The general process is as follows:
1. Data Source Compatibility.
2. MySQL Environment Preparation.
3. Hive environment preparation.
4. WeData preliminary preparation.
5. Offline single table: MySQL to Hive synchronization configuration steps.
6. Offline node advanced parameters.
7. Frequently Asked Questions.

1. Data Source Compatibility

MySQL Data Source

If you want to perform offline data synchronization using MySQL, you need to first confirm the supported MySQL data source versions and the supported data type conversions for reading.

Supported Versions

You need to follow these version limitations:

Type	Version	Driver
MySQL	5.6,5.7,8.0.x	JDBC Driver:8.0.21
RDS MySQL	5.6,5.7, 8.0.x	
PolarDB MySQL	5.6,5.7,8.0.x	
Aurora MySQL	5.6,5.7,8.0.x	
MariaDB	10.x	
PolarDB X	2.0.1	

MySQL Data Type Conversion

The supported MySQL data types and their corresponding type conversions for reading are as follows (when processing MySQL, the MySQL data source types are mapped to data processing engine types):

MySQL Data Type	Internal Type
int, tinyint, smallint, mediumint, int, bigint	Long
float, double, decimal	Double
varchar, char, tinytext, text, mediumtext, longtext, year	String
date, datetime, timestamp, time	Date
bit, bool	Boolean
tinyblob, mediumblob, blob, longblob, varbinary	Bytes

Hive Data Source

If you want to perform offline data synchronization using Hive, you need to first confirm the supported Hive data source versions and the supported data type conversions for writing.

©2013-2025 Tencent Cloud. All rights reserved.

Supported Versions

You need to follow these version limitations:

Data Source	Supported Versions
Hive	2.1.1,2.3.2,2.3.5,2.3.8,2.3.9,3.1.1,3.1.2,3.1.3

Hive data type conversion for writing

The supported Hive data types and their corresponding type conversions for writing are as follows (when processing Hive, the Hive data source types are mapped to data processing engine types):

Internal Type	Hive Data Type
Long	TINYINT,SMALLINT,INT,BIGINT
Double	FLOAT,DOUBLE
String	String,CHAR,VARCHAR,STRUCT,MAP,ARRAY,UNION,BINARY
Boolean	BOOLEAN
Date	Date,TIMESTAMP

2. MySQL Environment Preparation

Confirm MySQL Version

DataInLong has requirements for MySQL versions. Check whether the MySQL version to be synced meets these requirements. You can check the current MySQL database version with the following statement.

```
select version();
```

Setting MySQL Server Permissions

You can define a MySQL user with appropriate permissions.

1. **Create a MySQL User (Optional):**

```
mysql> CREATE USER 'user'@'localhost' IDENTIFIED BY 'password';
```

2. **Grant Necessary Permissions to the User:**

In the case of offline synchronization, the account must have SELECT privileges on the database. You can refer to the following command:

```
mysql> GRANT SELECT ON *.* TO 'user' IDENTIFIED BY 'password';
```

3. **Refresh the User's Permissions:**

```
mysql>FLUSH PRIVILEGES;
```

Add Tencent Cloud MySQL Database Security Group

A security group is a stateful virtual firewall with filtering features. It is used to set network access control for one or more cloud databases and is an important network security isolation method provided by Tencent Cloud. If you are using a Tencent Cloud MySQL database above the basic version, you need to add the following access IPs to the security group of the target database. If you do not configure the security group, the connectivity test during MySQL data source configuration may fail. For specific operations, see Managing Cloud Database Security Groups.

©2013–2025 Tencent Cloud. All rights reserved.

```
118.89.220.0/24, 139.199.116.0/24, 140.143.68.0/24, 152.136.131.0/24, 81.70.150.0/24, 81.70.161.0/24,
81.70.195.0/24, 81.70.198.0/24, 82.156.22.0/24, 82.156.221.0/24, 82.156.23.0/24, 82.156.24.0/24,
82.156.27.0/24, 82.156.82.0/24, 82.156.84.0/24, 82.157.119.0/24
```

3. Hive Environment Preparation

If you configured EMR as the computational engine when configuring the storage engine, it will allocate a system Hive data source for you. If you want to self-configure the Hive data source, you need to prepare the following related files in advance for data source configuration on WeData:

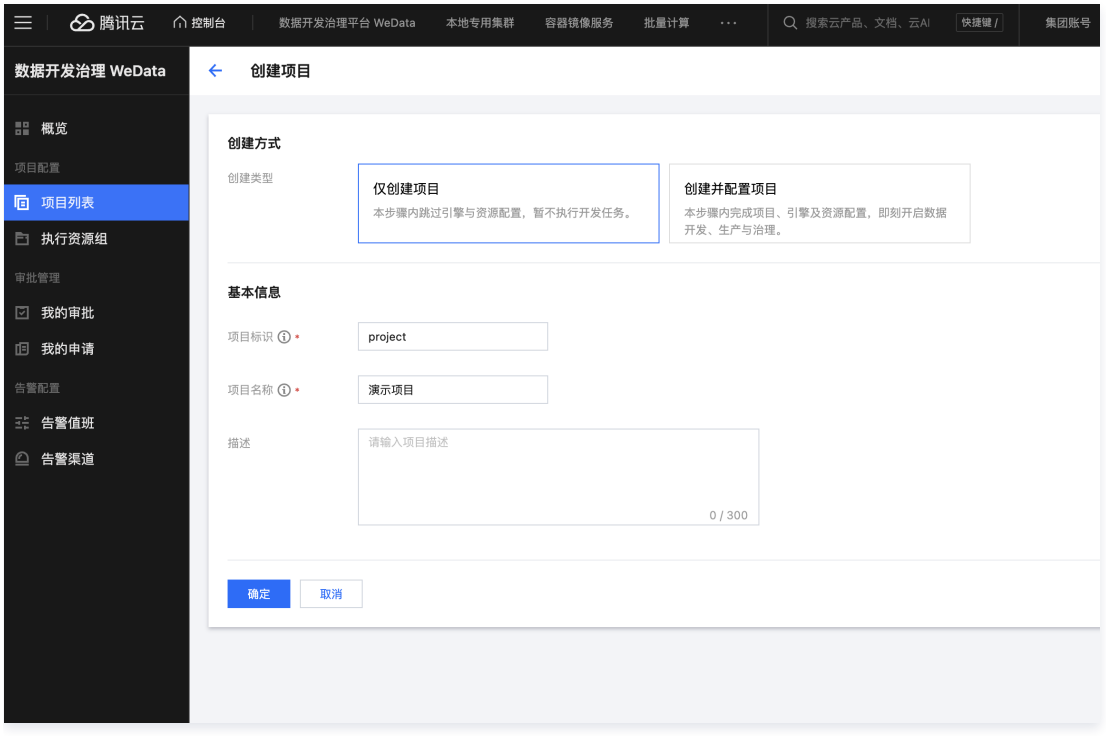
- core-site.xml, hdfs-site.xml: in the /usr/local/service/hadoop/etc/hadoop directory of the EMR machine.
- hive-site.xml: in the /usr/local/service/hive/conf/ directory.

4. WeData Prerequisites

Before configuring the synchronization task, a project needs to be created. The following steps are based on a new environment. If you have already created project members or completed project creation, you can skip this step.

Project Creation

If you haven't created a project space in WeData, you need to complete the project creation before performing data synchronization tasks.
The main account holder logs into the CAM console to add a sub-account. Refer to Preparing a CAM Sub-account. Once created, the member can use the set sub-account to log in to the WeData big data platform.
To create a project space, go to the WeData homepage, click **Project List** > **Create Project**. If you are only doing DataInLong, you can choose **Create Project Only** and enter the project identifier and name.

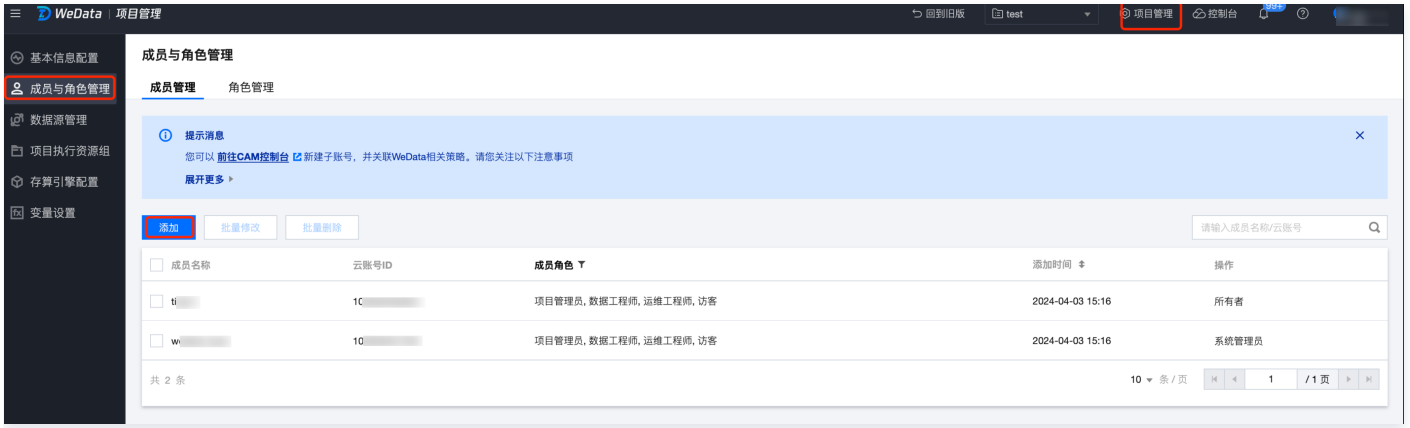


Add Member

Only the project owner and main account have permission to enter the newly created project. Other sub-accounts need to be added by the project owner or the main account in the project management module. For details, please see Project Management.

1. Click **Project Management** > **Member and Role Management** to enter the Member and Role Management interface. Click the **Add** button.

©2013-2025 Tencent Cloud. All rights reserved.



2. In the Add Member interface, set the required roles for the newly created member. Once set, the member can enter the project.

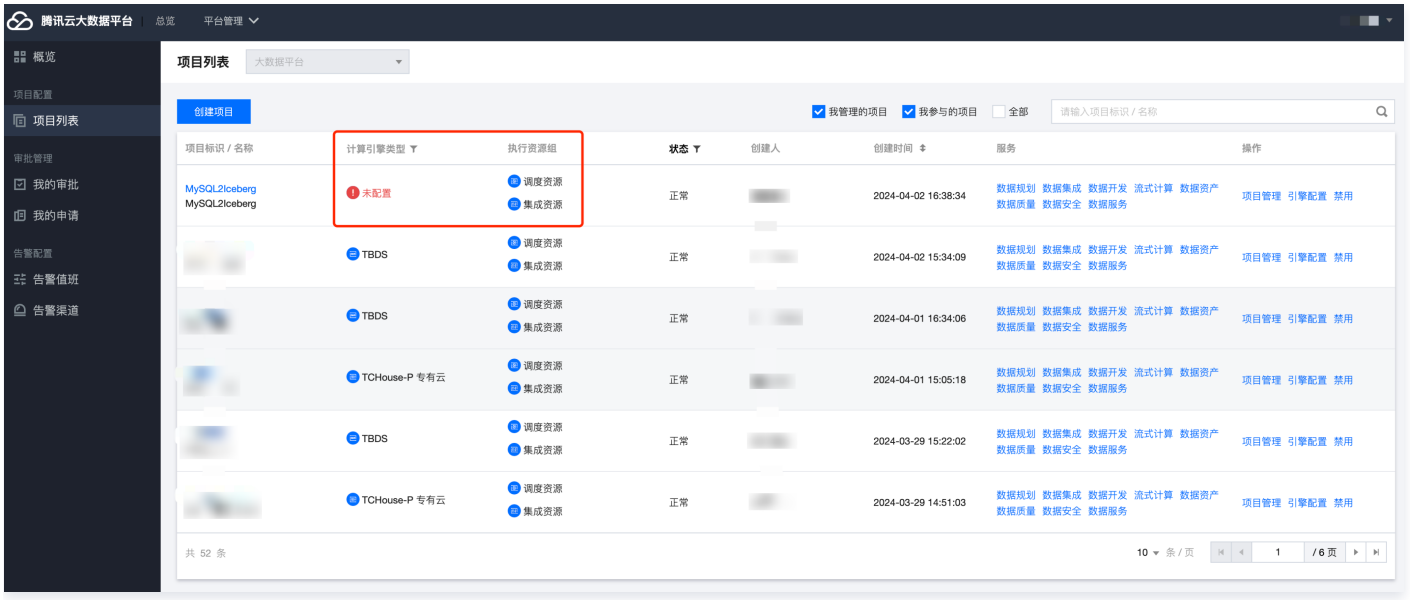


Resource Group Preparation

Confirm whether the Resource Group is configured

Before officially starting the DataInLong task, please confirm whether your project is equipped with **Integration Resources**.

- If it prompts that the integration resource is not configured, the DataInLong task cannot run normally.
- When you finish creating a new project, the column for **Execution Resource Group** will show that the integration resource is not configured. You just need to go to the WeData homepage > **Execution Resource Group** > **Integration Resource Group** and associate resources for the new project.



Enter the Integration Resource Group interface and click **Associate Project** after the corresponding Resource Group.

©2013–2025 Tencent Cloud. All rights reserved.

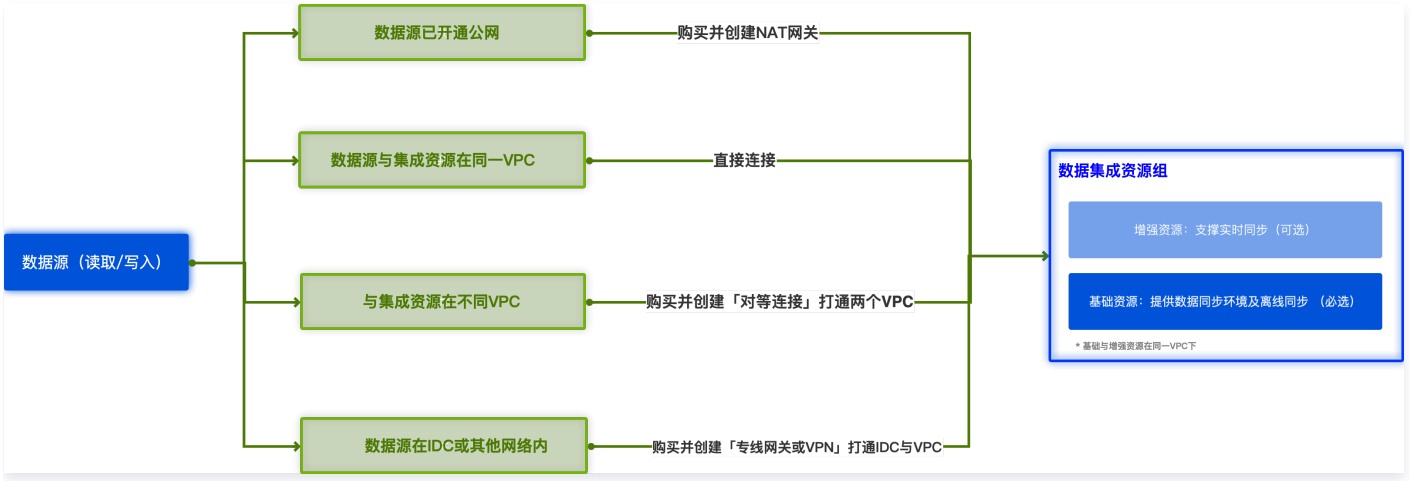


Purchase Integrated Resource Group

If you find that there is currently no integration resource group, you need to go for purchase. The Integration Resource Group is dedicated to using computing resources when running DataInLong tasks. These resources are mainly presented in the form of resource groups. Before configuring the synchronization task, you need to confirm whether the DataInLong Resource Group is purchased. For details, please refer to Configure the Integration Resource Group.

Confirm Network Connectivity

Once the Integrated Resource Group is configured, ensure the data source network (including reading and writing) is intercommunicable with the DataInLong Resource Group, and that resources are not inaccessible due to allowlist restrictions or similar reasons, or else data transfer synchronization cannot be completed. The machine resources included in the DataInLong Resource Group should by default be within the same VPC network environment:



- **If the data source has public network access**: You need to purchase and create a NAT Gateway to allow the integration resources to connect to the data source VPC through the gateway. For detailed operations, please refer to Resource Group Configuration for Public Network.
- **If the data source is within a VPC**:
  - If the data source is in the same VPC as the integration resource: You can use it directly.
  - If the data source is in a different VPC from the integration resource: Purchase a Peering Connection to interconnect the VPCs of the integration resource and the data source.
- **If the data source is located in an IDC or other traditional network environment**: You need to purchase a VPN or Direct Connect Gateway to interconnect the VPCs of the integration resource and the data source.

©2013–2025 Tencent Cloud. All rights reserved.

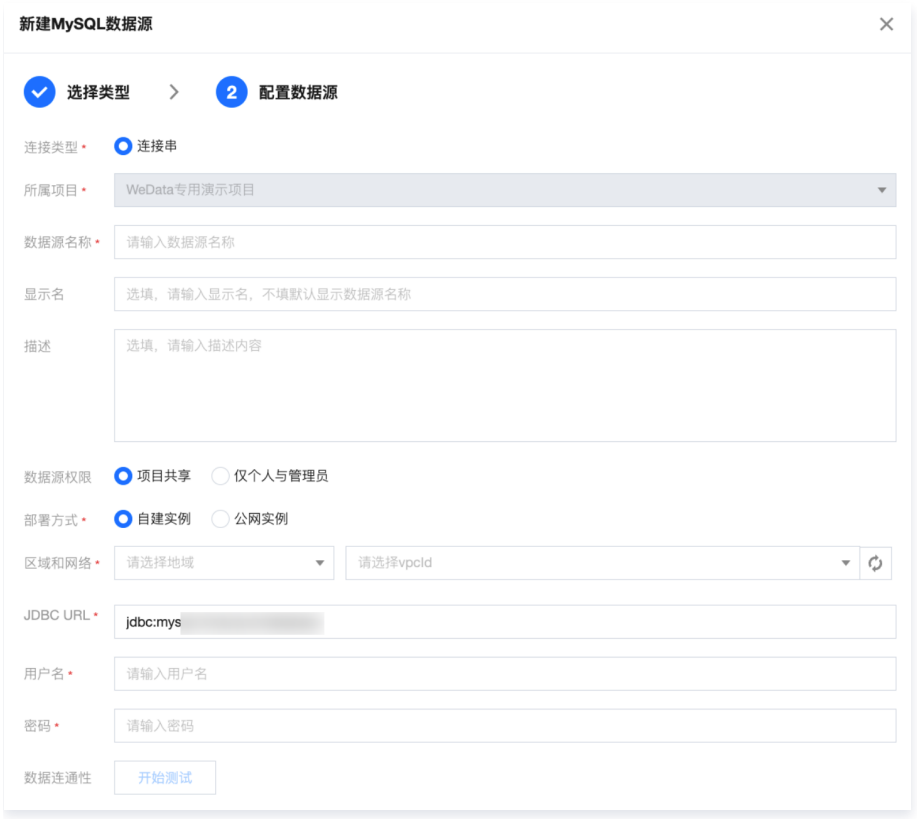
> **Note**
> For documentation on the configuration and management of the Integrated Resource Group, refer to Integrated Resource Configuration and Management.

Data source configuration

MySQL Data Source Configuration

Enter the data source configuration interface, MySQL data sources support two connection methods: cloud instance and connection string.
click **Project Management** > **Data Source Management** > **New Data Source** > **Choose MySQL Data Source**.

- Create a data source via connection string.



Parameter	Description
Connection Type	Selecting the connection mode for the Data Source connection string
Associated item	The item to which the data source belongs when created
Data source name	The name of the new data source defined by the user, cannot be empty, starts with a letter, can contain letters, numbers, and underscores. Length within 20 characters
Display name	The name displayed when the data source is used in the product. If not filled, the default is the data source name
Description	Optional, description of this data source
Deployment mode	Supports two deployment modes: self-built instances and public network instances. Self-built instances are data source instances deployed on Tencent CVM, while public network instances are those in the customer's local IDC or other cloud resources, accessible via the public network.
Region and Network	When choosing a self-built instance, you need to select the region and vpcID where the data source instance is located.

©2013-2025 Tencent Cloud. All rights reserved.

JDBC URL	The connection string information used to connect to MySQL data source instances, including host IP, port, database name, etc.
Username	Username to connect to the database
Password	Password to connect to the database
Data Connectivity (Legacy)	Test whether the configured database is connected **Note** • If the connectivity test fails, the data source can still be saved. However, it cannot be used if the connectivity test fails and the data source is saved. • If the connectivity test fails, it may be because WeData is blocked by the database network firewall. You need to add Tencent Cloud MySQL Database Security Group. Refer to Chapter 2 Add Tencent Cloud MySQL Database Security Group. • After the 2024 Spring Festival, newly purchased WeData will not have this feature by default. This feature has been upgraded to Resource Group Connectivity.
Resource Group Connectivity (New)	Support the connectivity test between specific resource groups and data sources. Users only need to ensure that their data source and the VPC network where the Integrated Resource Group is located are connected without additional network interconnection.  **Note** Users who purchased WeData before the Chinese New Year in 2024 do not have this feature. If you need to use it, please contact Tencent's operation and maintenance personnel.

Hive Data Source Configuration

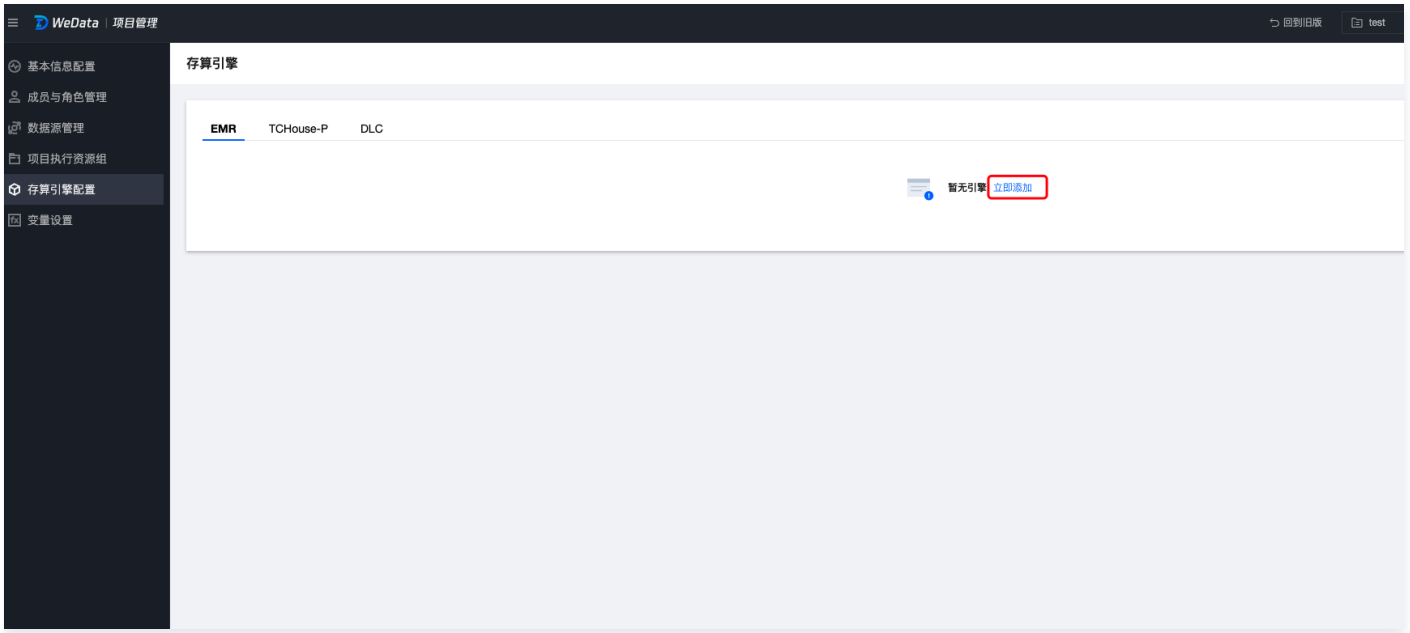
Hive data sources are divided into system sources and custom sources. System sources are automatically generated after binding with the EMR engine; users can use them directly but cannot delete or edit connection information. The same EMR cluster will generate only one Hive system source. Custom sources are similar to the aforementioned MySQL and are registered by users themselves, allowing the registration of multiple Hive custom sources as needed.

Hive System Source

After creating a new project, if the project's computing engine type is not associated, you can go to the **Storage and Computing Engine Configuration** page to configure the storage and computing engine for the project. Follow the instructions below to configure the storage and computing engine. For detailed configuration information, refer to Storage and Computing Engine Configuration.

©2013–2025 Tencent Cloud. All rights reserved.





Hive Custom Source

Click **Project Management** > **Data Source Management** > **New Data Source** > **Select Hive Data Source.**

- Create a data source via connection string.

©2013–2025 Tencent Cloud. All rights reserved.



Parameter	Description

©2013–2025 Tencent Cloud. All rights reserved.

Data source name	The name of the newly created data source should be defined by the user and cannot be empty. The name should start with a letter and can include letters, numbers, and underscores. It should be within 20 characters in length and cannot be modified
Description	Optional, description of this data source.
Data source permissions	Project Sharing means that all members of the project can use the current data source. 'Personal and Administrator' means that only the creator and the project administrator can use the data source.
Deployment mode	You can choose either a self-built instance or a public network instance as the deployment mode
Hive Version	The specific version number of the Hive software being used
Region and Network	Choose Region and Network
Hive Server 2 Connection Address	This is the JDBC URL for the Hive Server 2 database connection, used to specify how clients connect to the Hive service. The format typically includes the protocol jdbc:hive2://, followed by the host address, port number, and the default database name to connect to.
Username	Username for connecting to the database.
Password	Password for connecting to the database.
ZooKeeper Connection	Indicates whether to connect via ZooKeeper. When enabled, ZooKeeper can be used to manage Hive Server 2 instances in a distributed environment
ZooKeeper Connection Address	Example:jdbc:hive2://zk01:2181,zk02:2181,zk03:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2
Hive Metadata Address	The network address where Hive metadata is stored, usually pointing to the Hive Metastore Service.
Hive MetaDB	Indicates whether to enable connection to the Hive Metadata Database. The metadata database stores information about all tables and columns in Hive.
Hive MetaDB IP and port	After enabling Hive MetaDB, you need to configure the IP and port. The IP and port specify the location and access port of the Hive MetaDB server
Database name	After enabling Hive MetaDB, you need to configure the name of the database you created in Hive MetaDB for storing Hive metadata
Username	Account name for the database connected to Hive MetaDB
Password	The password is the account password associated with the aforementioned username. When configuring Hive, you need to provide the correct password to verify identity and authorize access to MetaDB.
core-site.xml file	The core configuration file of Hadoop includes the basic system settings such as security settings and Hadoop runtime I/O settings.**This parameter is mandatory if the data source is used as a write node!**
hdfs-site.xml file	The configuration file of Hadoop Distributed File System (HDFS), includes specific HDFS configuration details such as replication policy and permission settings.**This parameter is mandatory if the data source is used as a write node!**
hive-site.xml file	The configuration file for Hive includes runtime configuration details such as Metastore connection information and storage path.**This parameter is mandatory if the data source is used as a write node!**
Authentication Method	Specify the authentication method used for connecting to Hive. You can choose "None" for no authentication or "Kerberos" for using the Kerberos authentication protocol
keytab file	When Kerberos authentication is enabled, the keytab file needs to be configured. The keytab file contains the key table for Kerberos authentication.
conf file	When Kerberos authentication is enabled, the conf file must be configured. The conf file is Hive's configuration file, which includes various configuration parameters needed for Hive operations,

©2013-2025 Tencent Cloud. All rights reserved.

	including Kerberos-related configuration information
principal	When Kerberos authentication is enabled, the principal of the Hive service also needs to be specified so that Kerberos can verify the identity of the Hive service.
Data Connectivity (Legacy)	Test whether the configured database is connected. **Note** • If the connectivity test fails, the data source can still be saved. However, it cannot be used if the connectivity test fails and the data source is saved. • After the Chinese New Year in 2024, newly purchased WeData does not include this feature by default. This feature has been upgraded to Resource Group Connectivity.
Resource Group Connectivity (New)	Support the connectivity test between specific resource groups and data sources. Users only need to ensure that their data source and the VPC network where the Integrated Resource Group is located are connected without additional network interconnection.  **Note** Users who purchased WeData before the Chinese New Year in 2024 do not have this feature. If you need to use it, please contact Tencent's operation and maintenance personnel.

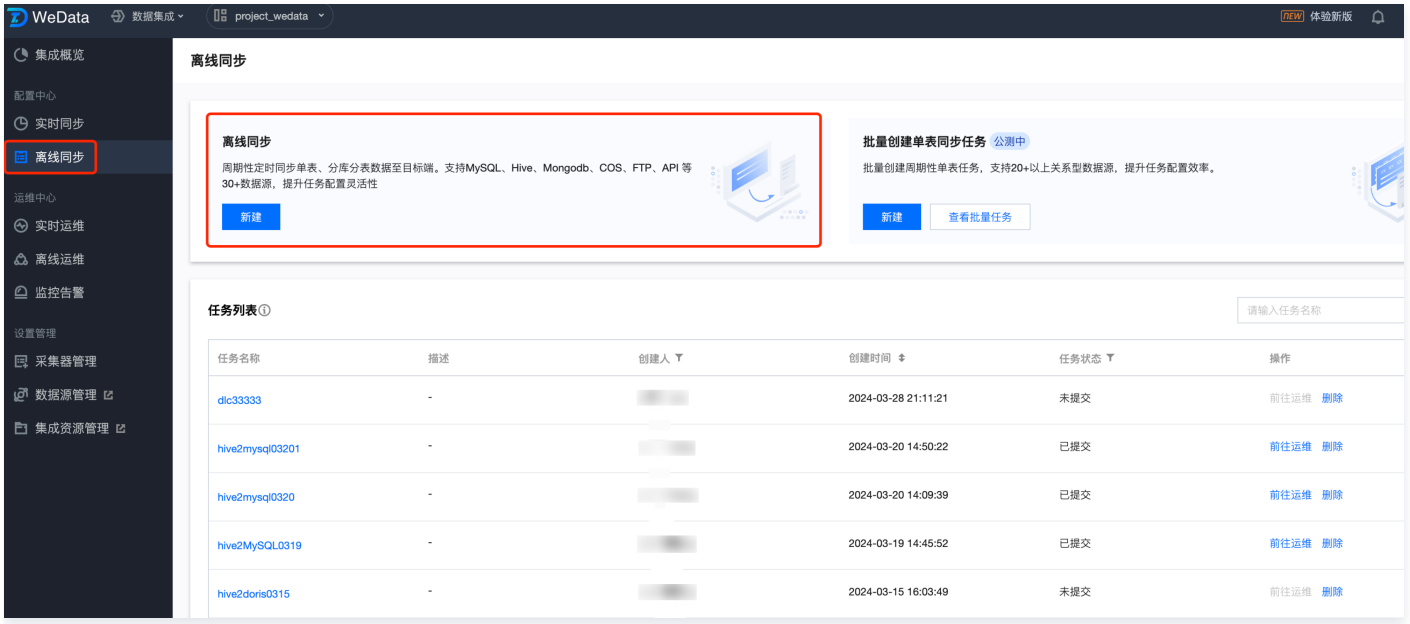
5. Offline Single Table: MySQL to Hive synchronization configuration steps

Log in to WeData Console, then click **DataInLong** > **Offline Synchronization**. Then create a new offline synchronization task, you can choose Form Mode, Canvas Mode, or Script Mode.

- Form Mode: Provides only read and write nodes, suitable for fixed field synchronization from single table to single table. Applicable to ODS layer data synchronization without the need for data cleaning.
- Canvas Mode: Provides three types of nodes for reading, writing, and transformation. Suitable for workflows that include cleaning processes.
- Script Mode: Supports an initialized script mode configuration page. Users can choose different data sources and data targets, displaying the corresponding script templates
  - Users need to select the data source and data target first. Editing is not allowed until a selection is made.
  - Once selected, the corresponding script module is displayed.
  - In the script, users can manually write parameters such as data sources and connection information.
  - Supports writing SQL statements in the script, inserting querysql into the connection.

This example explains the configuration steps using Form Mode.

©2013-2025 Tencent Cloud. All rights reserved.



Step 1: MySQL Read Node Configuration

Using Form Mode as an example, the details of MySQL read node configuration are explained.

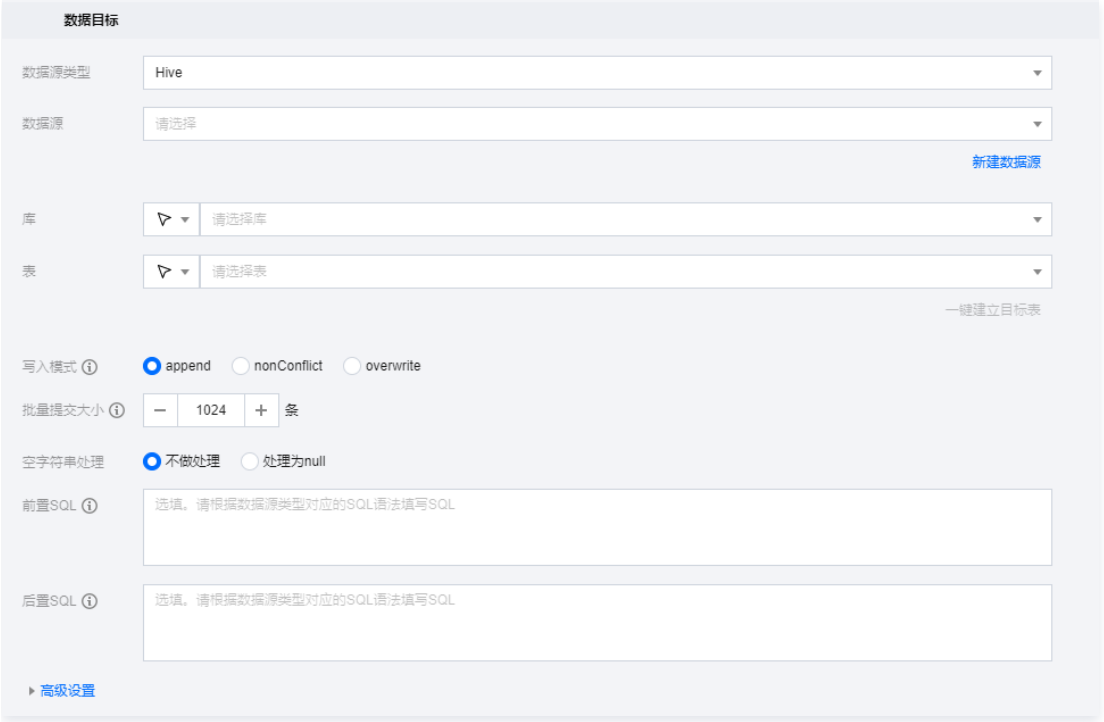


Parameter	Description
Data Source	Available MySQL data sources.
Database	Support selecting or manually entering the name of the database to be read • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Support selecting or manually entering the name of the table to be read • In the case of table partitioning, multiple table names can be selected or entered at the MySQL source end. The structure of multiple tables must be consistent. • In the case of table partitioning, you can configure the table serial number range. For example, 'table_[0-99]' means reading 'table_0', 'table_1', 'table_2' up to 'table_99'; if your table numeric suffixes

©2013–2025 Tencent Cloud. All rights reserved.

	are consistent in length, such as 'table_000', 'table_001', 'table_002' up to 'table_999', you can configure it as '"table": ["table_00[0-9]", "table_0[10-99]", "table_[100-999]"]'. When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Adding Sharding	Suitable for sharding scenarios, click to configure multiple data sources, databases, and table information. In sharding and partitioning scenarios, ensure all table structures are consistent. The task configuration will default to display and use the first table structure for data acquisition.
Split Key	Specify the field used for data sharding. After specifying, concurrent tasks will be started for data synchronization to improve efficiency. You may use a column in the source data table as the split key; it is recommended to use a primary key or an indexed column as the split key. Only integer-type fields are supported.
Filter Condition (Optional)	In actual business scenarios, data from the current day is often chosen for synchronization, with the where condition specified as gmt_create > $bizdate. The where condition can effectively increase business incremental synchronization. • If the where clause is not specified, including not providing the key or value for where, data synchronization will be considered as full data synchronization. • The where condition cannot be specified as limit 10, as this does not conform to MySQL WHERE clause constraints.
Advanced Settings (optional)	Parameters can be configured according to business needs. You may refer to Section 6 Offline Node Advanced Parameters in this article

Step 2: Hive Write Node Configuration



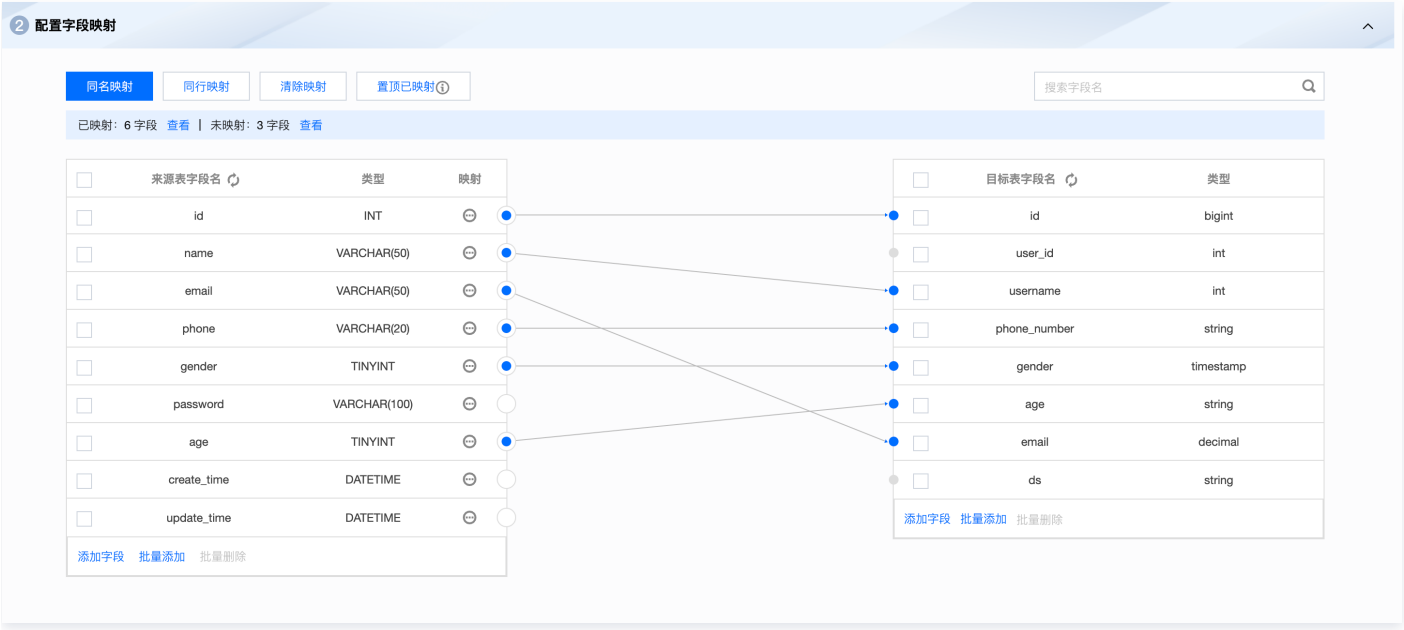
Parameter	Description
Data Source	The Hive data source to write into.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be

©2013-2025 Tencent Cloud. All rights reserved.

	performed given network connection in DataInLong.
Table	Supports selection, or manual input of the table name to be written to When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Clear Table	You can manually choose whether to clear the Hive data table before writing.
Write mode	Hive writing supports three modes: • Append: Retain original data, new rows are appended • nonConflict: Reports an error on data conflict • Overwrite: Deletes original data and writes new data writeMode is a high-risk parameter. Please pay attention to the data write-out directory and write mode to avoid accidental data deletion. Data loading behavior needs to be used in conjunction with hiveConfig. Please pay attention to your configuration.
Batch Submission Size	The batch size of the records to be submitted at one time. This value can greatly reduce the number of network interactions between the data synchronization system and Hive, and improve the overall throughput. If the value is set too large, it may cause the data synchronization process to run an OOM exception.
Pre-SQL (optional)	SQL command executed before the synchronization task. Fill in the SQL according to the correct SQL syntax corresponding to the data source type. For example, clear old data in the table before execution (truncate table tablename).
Post-SQL (optional)	SQL statements executed after the synchronization task are written according to the correct SQL syntax corresponding to the data source type, for example, adding a timestamp with `ALTER TABLE tablename ADD colname TIMESTAMP DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP`.
Advanced Settings	Advanced settings can be referenced in Section 6 Offline Node Advanced Parameters of this article.

Step 3: Configure Field Mapping

Since the table structures of the source and target databases may not be exactly the same, such as differences in field names, data types, or field order, field mapping can convert the source database fields to the corresponding fields in the target database to ensure correct data synchronization.
Field mapping aims to specify the source of target field content by connecting lines. It supports three methods to configure the relationships between source and target nodes: same name mapping, peer mapping, and manual connections.



Step 4: Configure Task Attributes

©2013-2025 Tencent Cloud. All rights reserved.

After completing the synchronization information configuration, you need to set the basic attributes of the task, utilize resources, and data link channel information.



Category	Parameter	Description
Basic Attributes	Task Name/Type	Displays basic information of the current task name and type.
	Responsible Person	The name of one or more space members responsible for this task, defaulting to the task creator.
	Description	Displays current task remark information.
	Scheduling Parameters	Scheduling parameters are used during task scheduling. They are automatically replaced based on the task scheduling business time and the format of the scheduling parameters, achieving dynamic parameter values during the task scheduling period
Resource configuration	Integrated Resource Group	Specifies the name of the integrated resource group used by the current task; only one resource group can be bound to a task.

©2013–2025 Tencent Cloud. All rights reserved.

Task Running Policy	Associate alarm	Supports associating alarm rules with the current task
Channel Settings	Dirty Data Threshold	Dirty data refers to data that fails to be written during synchronization. The dirty data threshold is the maximum number of dirty data records or bytes that can be tolerated during synchronization. Once this threshold is exceeded, the task will automatically end. The default threshold is 0, meaning no dirty data is tolerated.
	Concurrent number	Maximum concurrency for the expected task during actual execution. Due to reasons such as resources, data source types, and task optimization results, the actual concurrency may be less than or equal to this value. The larger this value, the more pre-allocated execution machine resources. **Note:** When the concurrency > 1, if the source data source supports setting a split key, then the split key is a mandatory field; otherwise, the set concurrency will not take effect.
	Synchronization Rate Limiting	Limit the synchronization rate based on traffic or the number of records to protect the read and write pressure on the data source end or data destination end. This value is the maximum running rate, with a default of -1 indicating no rate limit.

Step 5: Configure Task Scheduling

To ensure that the task execution time meets expectations, you need to set the current task's periodic operation plan, including scheduling time and dependency attributes.



Category	Parameter	Description
Scheduling Time	Scheduling Method	Periodic Scheduling: The task runs periodically according to the configured scheduling plan. One-time Execution: The task runs only once at the specified time.
	Effective Date	The effective time period for scheduling time configuration. The system will automatically schedule according to the time configuration within this period. It will not be automatically

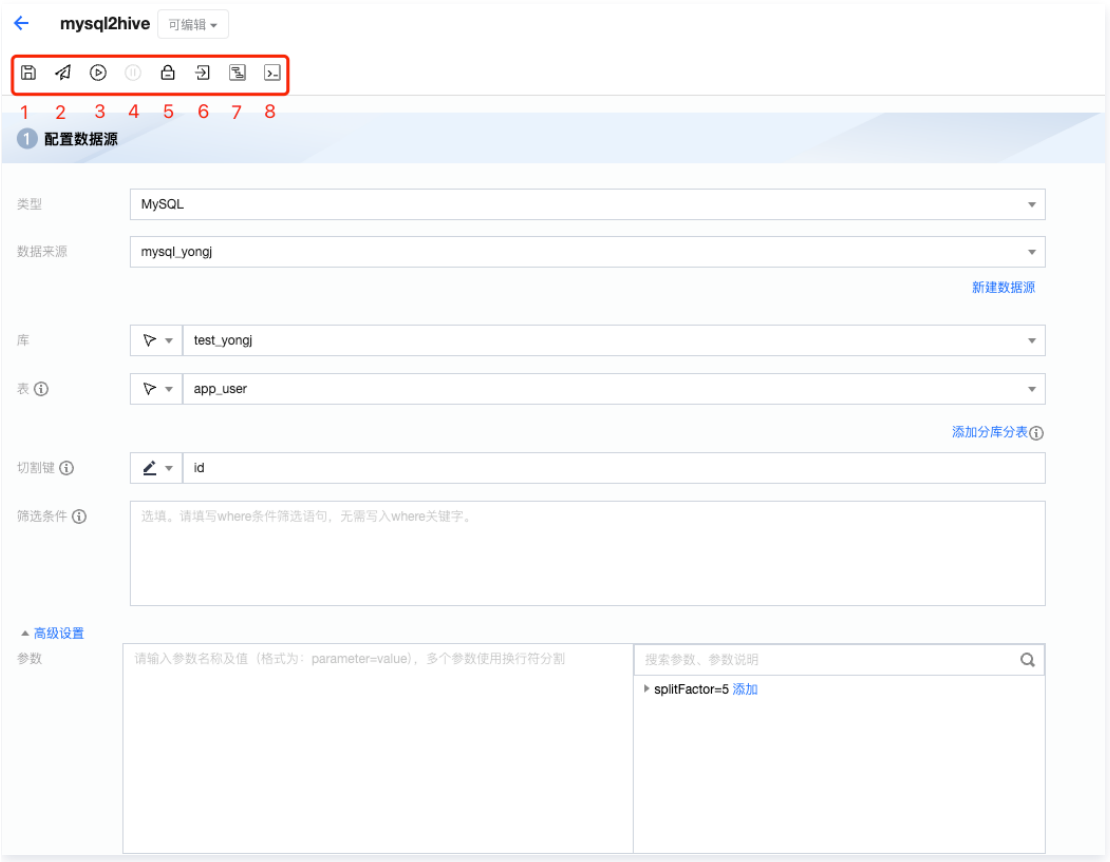
©2013-2025 Tencent Cloud. All rights reserved.

		scheduled beyond the validity period.
	Scheduling Period	Scheduling plan interval step unit, supporting Year, Month, Week, Day, Hour, Minute: • Minute: You need to specify the specific execution start time and interval. The task will run periodically from the start minute of each hour according to the interval. For example, if the execution time is 02:00 to 23:59 and the interval is 5 minutes, the task will run an instance every 5 minutes starting from 02:00. • Hour: You need to specify the specific execution start and end time and the interval. For example, if the execution time is 02:20 to 05:00 and the interval is 1 hour, the task will run instances respectively at 02:20, 03:20, and 04:20. • Day: You need to specify the specific execution time each day. The task runs only once at that time each day. • Week: You need to specify the fixed days for weekly execution (supporting multiple selections) and the time. The task runs only once at the specified time on those days. • Month: Specify the fixed run date and time for each month. If 'Month End' is selected, it will run on the last day of the month. • Year: Specify the fixed run date and time for each year.
Dependency Attributes	Self-dependency	Self-dependency refers to the dependency relationship between different instances of the same task: • Ordered Serialization: The current instance depends on the state of the previous cycle instance. • Unordered Serial: The current instance has no dependency relationship with the previous cycle instance. If a task has multiple instances simultaneously, the system randomly selects one instance to run. Only one instance will be in the running state at any time. • Parallel: There is no dependency relationship between the previous cycle instance and the subsequent cycle instance. If a task has multiple instances simultaneously, multiple instances will run concurrently.
	Retry Wait Time	The maximum wait time interval for each retry run after an instance fails. If the instance is not retried within this interval, it will be marked as failed.
	Retry Attempts	The maximum number of retries after an instance fails. If this limit is exceeded, the task will be marked as failed.

Step Six: Task Test Run and Submission

After configuring the Offline Synchronization Task, you can proceed with an online test run or submit it to the production scheduling environment. Currently, the task configuration page supports Save, Submit, Test Run, Debug Stop, Lock/Unlock, and Go to Operations and Maintenance.

©2013-2025 Tencent Cloud. All rights reserved.



Serial number	Parameter	Description
1	Save	Save the current task configuration information, including data node configuration, node connection, task attributes, and task scheduling configuration.
2	Submit	Submit the current task to the production environment. After submission, the task will run periodically according to the scheduling attributes. The submitted task will also generate task and instance records in **Task Operation and Maintenance** > **Offline Maintenance** . **Note** • Before submission, the task will save the latest configuration by default. • Before submission, the task will undergo necessary checks, including node configuration, task connections, and resource groups. If the checks do not pass, the task submission will fail and a prompt will be shown.
3	Test Run	Debug and run the current task.
4	Debug Stop	Terminate the task currently in test run.
5	Lock/Unlock	The default creator is the first lock holder and only the lock holder is allowed to edit the task configuration and run the task. If the lock holder does not perform any edit operations within 5 minutes, others can click the icon to seize the lock. Once successful, they can proceed with editing.
6	Go to Operations	Quickly jump to the task operations page based on the current task name.
7	Canvas Conversion	Switch the current edit mode to Canvas Mode

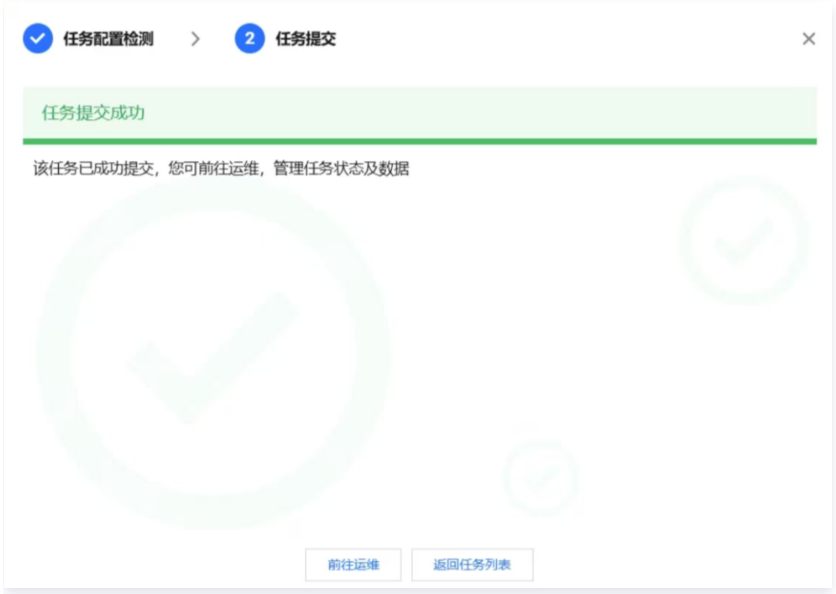
©2013–2025 Tencent Cloud. All rights reserved.

8	Script Conversion	Switch the current edit mode to script mode



Parameter	Description
Detection Abnormal	Supports skipping anomalies for direct submission, or halting the submission.
Detection contains only warnings and below	Can be submitted directly.

Submit Results

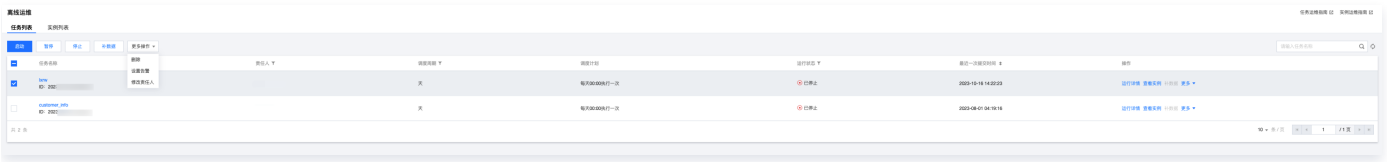


- Task Submitting:
  - Show submission progress.
- Task Submission Result – Success:
  - Show task submission successful result.
- Show the reason for task submission failure:
  - Return the failure reason.

Step Seven: Offline Task Operation and Maintenance

©2013–2025 Tencent Cloud. All rights reserved.

The task list page displays all periodic running tasks submitted to the scheduling system under the current account by default in a list format.



Parameter		Description
Task Name		Current Record Ownership for Task Name
Responsible Person		Responsible Person's Name configured when the current task was created
Scheduling Period		Cycle Scheduling Frequency configured for the current task
Scheduling Plan		Detailed Scheduling Plan for the current task
Running Status		Scheduling Running Status of the current task Scheduling: The task has been submitted for scheduling and is under normal scheduling Paused: The current task schedule is temporarily interrupted, and can be restarted later Stopped: The current task schedule has been terminated Stopping: The stop operation is being performed on the current task, and the state is in transition
Most Recent Submission Time		The last time the task was submitted to the scheduling system
Operation	Monitoring Management	Including task metric statistics, monitoring rule configuration, etc.: For details on metric statistics, see Offline Synchronization Metric Statistics. For monitoring rule configuration, see Alarm Subscription
	Viewing Instance	click to jump to the task instance information list
	Backfill Data	Perform batch backfill data for the task, only effective for tasks "Scheduling In Progress"
	Startup	Start scheduling tasks of the node, effective only for tasks "Paused" and "Stopped"
	Pause	Pause the scheduling tasks of the node, only effective for tasks "Scheduling In Progress". No new instances will be generated after being paused, but existing instances will continue to run
	Delete	Delete the task and all instances under it, only effective for tasks "Stopped"
	Stop	Terminate all "Pending Execution" and "Running" instances of this node's tasks, and no new instances will be generated
	Alarm settings	Set the task's alert information, supporting batch operations
	Change Responsible Person	Change the task's Responsible Person, supporting batch operations

6. Offline node advanced parameters

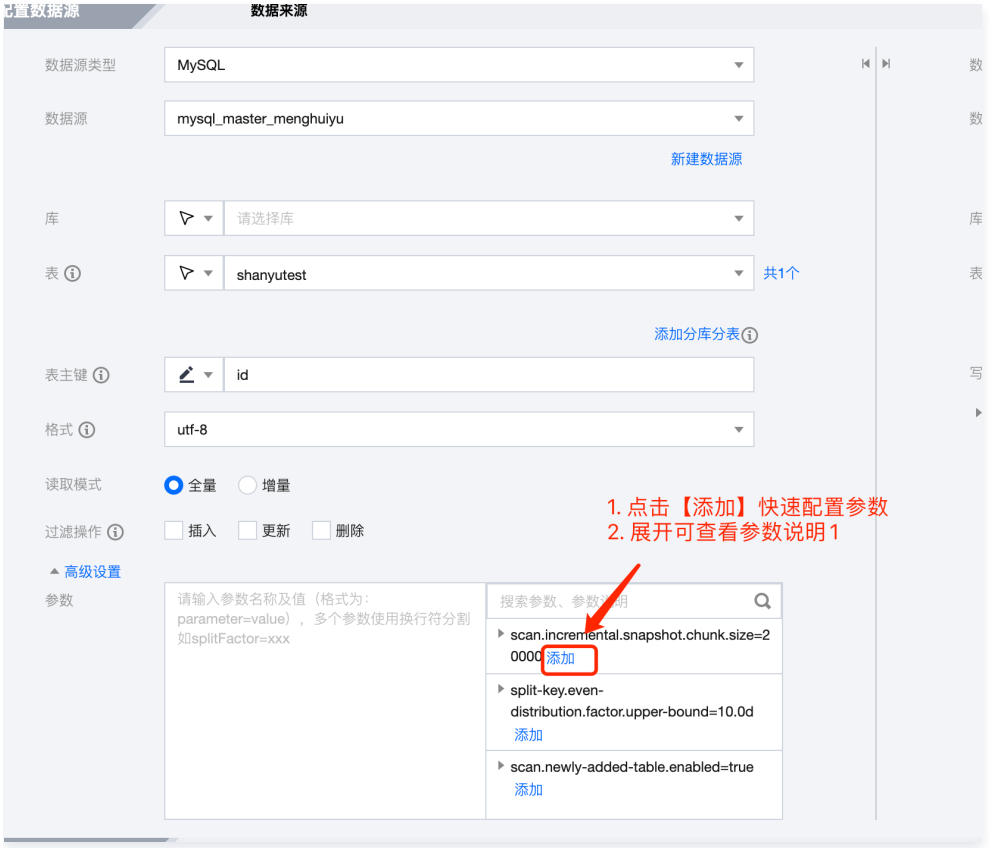
When configuring read-write nodes, you can configure advanced settings. The configuration and usage scenarios are as follows:

Offline type	Read/ Write	Configuration	Applicable Scenario	Description
mysql	Read	splitFactor=5	Single table	Control the granularity of data synchronization tasks being split. During the data synchronization process, if the amount of data to be synchronized is large, it may lead to inefficiency or excessive

©2013-2025 Tencent Cloud. All rights reserved.

				resource consumption for a single task. By setting the splitFactor, a large data synchronization task can be divided into multiple smaller tasks to be processed in parallel, thereby improving synchronization efficiency and performance
hive	Write	compress=none/snappy/lz4/bzip2/gzip/deflate	Single table	Default is none. Effective only for textfile format, not for ORC/Parquet (ORC/Parquet requires specifying compression in the create table statement)
	Write	format=orc/parquet	Single table	The format of HDFS temporary files defaults to ORC and is unrelated to the final Hive table format

When configuring advanced read-write node parameters, refer to the following diagram for guidance:



7. Frequently Asked Questions

1. Error when writing to Hive On CHDFS table: Permission denied: No access rules matched

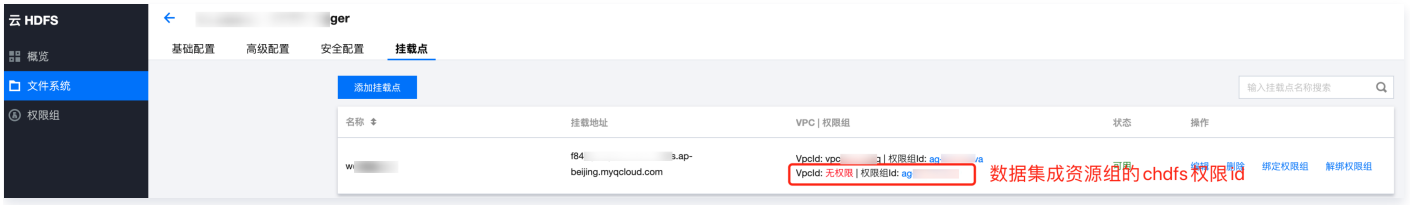
©2013-2025 Tencent Cloud. All rights reserved.

```
  at com.tencent.cloud.wedata.runner.common.api.AbstractTaskRunner.executeWithCommonLog(AbstractTaskRunner.java:508)
  at com.tencent.cloud.wedata.runner.common.api.AbstractTaskRunner.startWork(AbstractTaskRunner.java:546)
  at com.tencent.teg.scheduler.runner.runners.WeDataDiRunner.main(WeDataDiRunner.java:75)
[2022-12-29 01:00:52]-[ERROR] 获取FileSystem时发生网络IO异常,请检查您的网络是否正常!HDFS地址：[message:defaultFS =ofs://f4n .chdfs.ap-beijing.myqcloud.com]
[2022-12-29 01:00:52]-[ERROR] Exception when job run: com.tencent.wedata.common.exception.wedataException: Code:[HiveWriter-06], Description:[与HDFS建立连接时出现IO异常.]. - IOException: P
  at com.tencent.wedata.common.exception.wedataException.aswedataException(wedataException.java:26)
  at com.tencent.wedata.plugin.writer.hivewriter.HdfsHelper.getFileSystem(HdfsHelper.java:368)
  at com.tencent.wedata.plugin.writer.hivewriter.HiveWriter$Job.init(HiveWriter.java:194)
  at com.tencent.wedata.core.job.JobContainer.initJobWriter(JobContainer.java:717)
  at com.tencent.wedata.core.job.JobContainer.init(JobContainer.java:307)
  at com.tencent.wedata.core.job.JobContainer.start(JobContainer.java:114)
  at com.tencent.wedata.core.Engine.start(Engine.java:232)
  at com.tencent.wedata.core.Engine.entry(Engine.java:122)
  at com.tencent.teg.scheduler.runner.runners.wedata.wedataHelper.startwedata(wedataHelper.java:126)
  at com.tencent.teg.scheduler.runner.runners.WeDataDiRunner.execute(WeDataDiRunner.java:161)
  at com.tencent.cloud.wedata.runner.common.api.AbstractTaskRunner.executeWithCommonLog(AbstractTaskRunner.java:508)
  at com.tencent.cloud.wedata.runner.common.api.AbstractTaskRunner.startWork(AbstractTaskRunner.java:546)
  at com.tencent.teg.scheduler.runner.runners.WeDataDiRunner.main(WeDataDiRunner.java:75)

[2022-12-29 01:00:52]-[INFO] Total 0 records, 0 bytes | Speed 0B/s, 0 records/s | Error 0 records, 0 bytes |  All Task WaitWriterTime 0.000s |  All Task WaitReaderTime 0.000s | Percenta
[2022-12-29 01:00:52]-[INFO]

该任务最可能的错误原因是:
wedataException: Code:[HiveWriter-06], Description:[与HDFS建立连接时出现IO异常.]. - IOException: Permission denied: No access rules matched
[2022-12-29 01:00:52]-[ERROR] Task execution failed
[2022-12-29 01:00:52]-[INFO] commitTask success.
[2022-12-29 01:00:52]-[INFO] taskId: <2 37>
[2022-12-29 01:00:52]-[INFO] taskState: <FAILED>
```

Error reason:

DataInLong resource group is a fully managed resource group. Its outbound IP range is not within the customer's VPC internal network segment, so you need to open security groups for the DataInLong resource group.

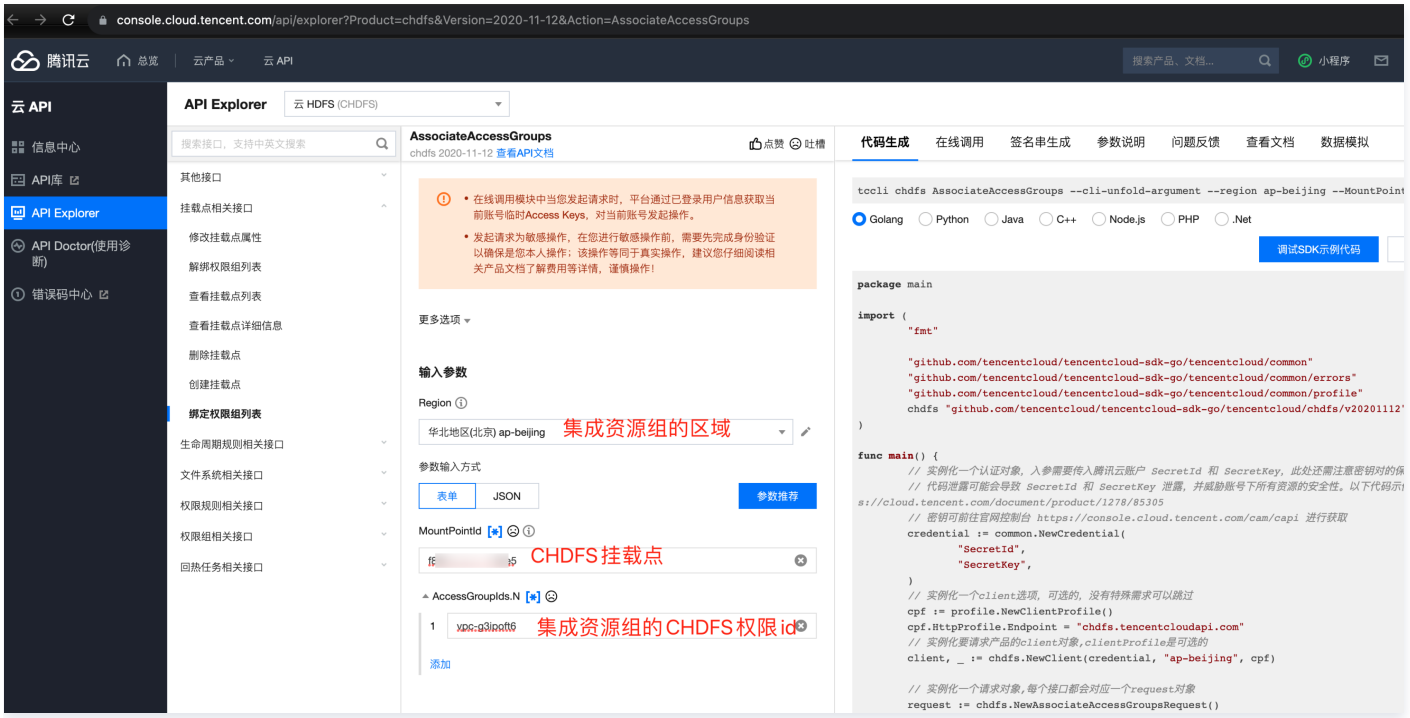
Please authorize the CHDFS permission ID for the integration resource group at the CHDFS mount point.



Solution:

Go to the API Explorer interface to fill in the parameters, where MountPointId is the customer’s CHDFS mount point, and AccessGroupIds.N is the CHDFS permission ID for the DataInLong resource group.

> **Note:**
> Different regions have different permission IDs.



Supported data sources and read-write capability

Last updated: 2024-12-19 10:14:39

©2013-2025 Tencent Cloud. All rights reserved.

DataInLong includes both offline and real-time data links, currently supporting relational databases, big data storage, semi-structured data, NoSQL, message queues, and various other data sources. You can select the corresponding feature modules based on the support of each module for the data sources to configure synchronization tasks.

Data Source		Offline Synchronization		Real-Time Synchronization					Version
		Offline Read	Offline Write	Single Table Read	Single Table Write	Database-wide Read	Database-wide Write	Log Collection	
Relational Database	MySQL	✓	✓	✓	✓	✓	-	✓	5.6，5.7，8.0.x
	TDSQL-C MySQL	✓	✓	✓	✓	✓	-	✓	-
	TDSQL MySQL	✓	✓	✓	✓	✓	-	✓	InnoDB Type
	PostgreSQL	✓	✓	✓	✓	✓	✓	✓	9.6，10，11，12
	TDSQL PostgreSQL	✓	✓	-	-	-	-	-	5.21.6 Row and Column Storage Fusion Edition
	SQL Server	✓	✓	✓	✓	✓	-	✓	2008，2008R2，2012，2014，2016，2017，2019
	Oracle	✓	✓	✓	✓	✓	-	✓	11，12，19
	DB2	✓	✓	-	-	-	-	-	-
	SAP HANA	✓	✓	-	-	-	-	-	-
	SAP IQ(Sybase)	✓	-	-	-	-	-	-	-
	DM	✓	✓	✓	✓	-	-	-	7,8
	OceanBase (beta usage)	✓	✓	✓	-	✓	-	-	3.1.4，4.2.1.3
Big data	TCHouse-P	✓	✓	-	✓	-	✓	✓	-
	Hive	✓	✓	-	✓	-	✓	✓	1.x，2.x，3.x
	DLC	✓	✓	-	✓	-	✓	✓	-
	Doris	✓	✓	-	✓	-	✓	✓	0.15、1.x
	StarRocks	✓	✓	-	-	-	✓	-	-
	ClickHouse	✓	✓	-	✓	-	-	✓	20.7+
	Iceberg	✓	✓	-	✓	-	✓	✓	0.13.1+
	HBase	✓	✓	-	✓	-	-	✓	2.2.x
	HDFS	✓	✓	-	✓	-	-	✓	2.x，3.x
	Kudu	✓	✓	-	-	-	-	-	-

©2013-2025 Tencent Cloud. All rights reserved.

	TBase	✓	✓	–	✓	–	–	✓	–
	GaussDB	✓	✓	–	–	–	–	–	–
	GBase	✓	✓	–	–	–	–	–	–
	Impala	✓	✓	–	–	–	–	–	–
	Greenplum	✓	✓	–	✓	–	–	✓	4.x，5.x，6.x
Semi-structured	COS	✓	✓	–	–	–	–	–	–
	Rest API	✓	–	–	–	–	–	–	–
	FTP	✓	✓	–	–	–	–	–	–
	SFTP	✓	✓	–	–	–	–	–	–
NoSQL	Redis	–	✓	–	–	–	–	–	–
	Elasticsearch	✓	✓	–	✓	–	–	✓	6.x，7.x
	CTSDB influxDB	✓	✓	–	–	–	–	–	–
	Mongo	✓	✓	✓	–	✓	–	–	4.x
	Alibaba Graph Database	✓	✓	–	–	–	–	–	Only Script Mode supported
Cloud Message Queue	Kafka	✓	✓	✓	✓	✓	✓	✓	2.4.1,2.7.1,2.8.1,2.8.2 (including Kafka, Ckafka data sources)
	TDMQ Pulsar	–	✓	–	–	–	–	–	Virtual Cluster Edition
	TiDB-Kafka	–	–	✓	–	–	–	–	–
	DTS-kafka	–	–	✓	–	–	–	–	–
Data Reporting	SDK	–	–	–	–	–	–	✓	–
	TKE	–	–	–	–	–	–	✓	–
	CVM	–	–	–	–	–	–	✓	–

Note:
Data link read/write ends must meet the corresponding data source versions.

©2013–2025 Tencent Cloud. All rights reserved.

Integrated resource configuration and management Integrated Resource Billing Instructions

Last updated: 2024-11-25 15:06:41

For billing-related information on integrated resources, please see Integrated Resource Billing Instructions.

©2013-2025 Tencent Cloud. All rights reserved.

Integration of resource purchase and configuration

Last updated: 2024-11-25 15:07:05

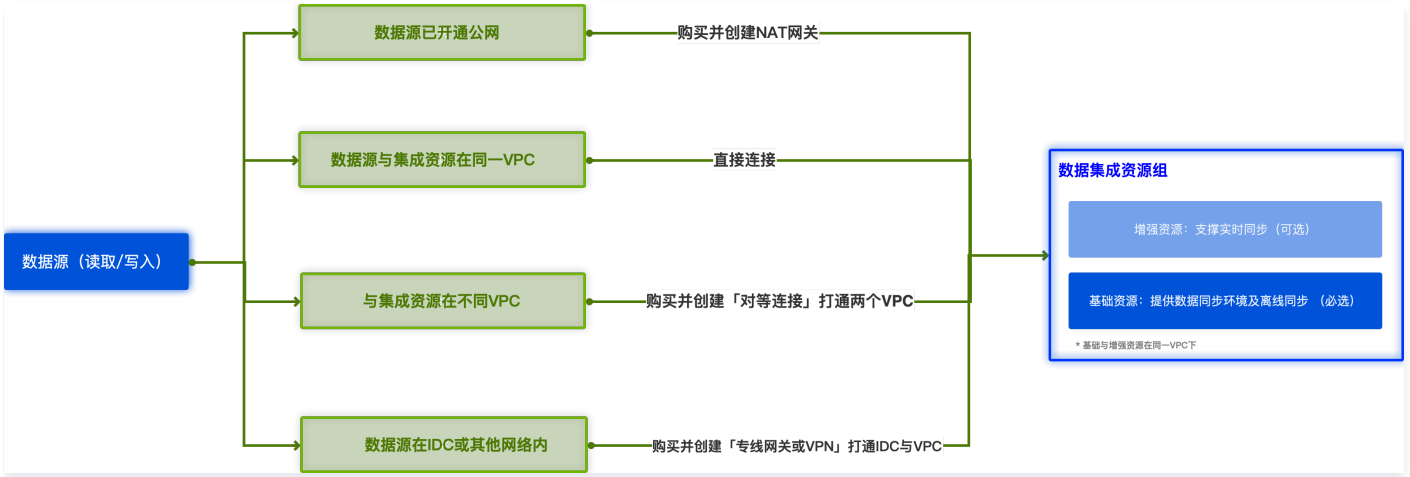
For details on integration of resource purchase and configuration, please refer to Integration of resource purchase and configuration.

©2013-2025 Tencent Cloud. All rights reserved.

Integration resource connectivity and usage planning

Last updated：2024-11-25 15:07:59

Before using DataInLong synchronization tasks, ensure that the network of the data source (including the read and write ends) and the DataInLong resource group can communicate with each other, and that resources are not denied access due to allowlist restrictions or other reasons; otherwise, data transmission synchronization cannot be completed.
Machines within the DataInLong resource group must be in the same VPC network environment by default:



- **If the data source is accessible via the public network**: You need to purchase and create a NAT Gateway to allow integration resources to connect to the VPC where the data source is located. For detailed operations, see Configuring Public Network for Resource Group.
- **If the data source is within a VPC**:
  - If the data source is in the same VPC as the integration resource: You can use it directly.
  - If the data source is in a different VPC from the integration resource: Purchase a Peering Connection to interconnect the VPCs of the integration resource and the data source.
- **If the data source is located in an IDC or other traditional network environment**: You need to purchase a VPN or Direct Connect Gateway to interconnect the VPCs of the integration resource and the data source.

©2013-2025 Tencent Cloud. All rights reserved.

Integrate Resource Group to configure the public network

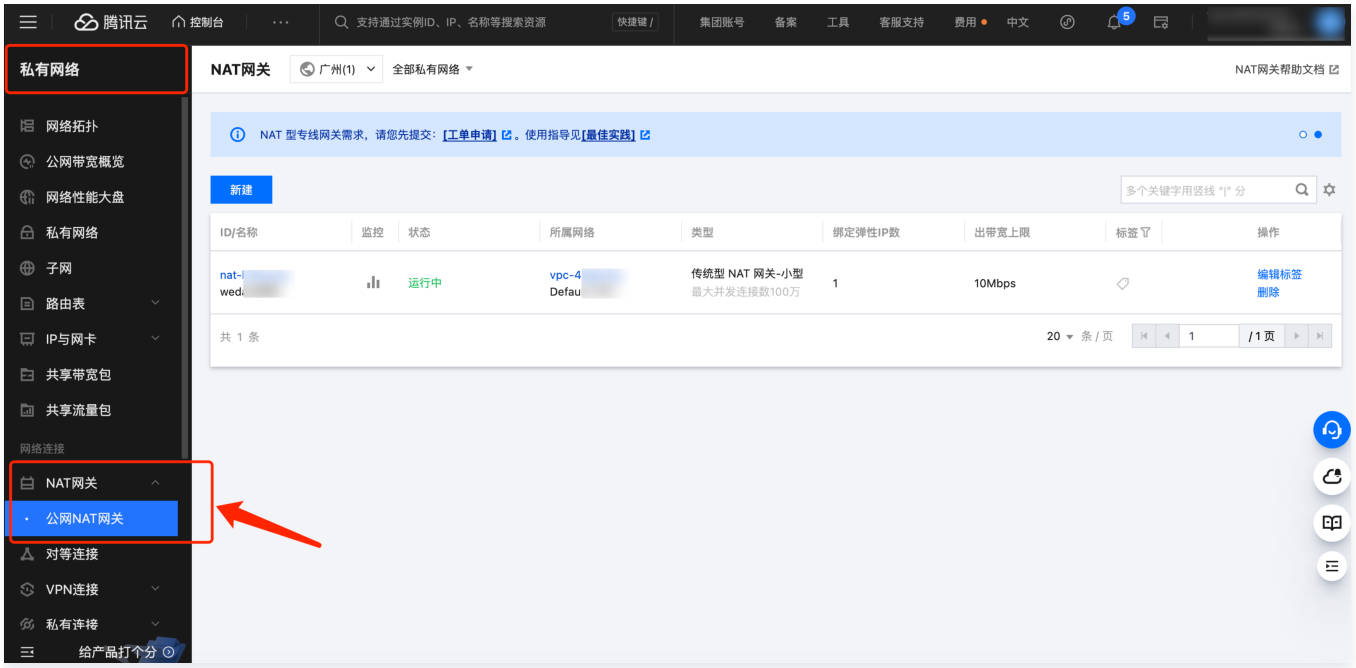
Last updated：2024-11-25 15:09:07

Operation scenarios

Applicable for configuring public network access for a Resource Group's VPC.

Operation step

1. Enter Tencent Cloud **VPC** > **NAT Gateway** . In the corresponding region, click **New** NAT Gateway.

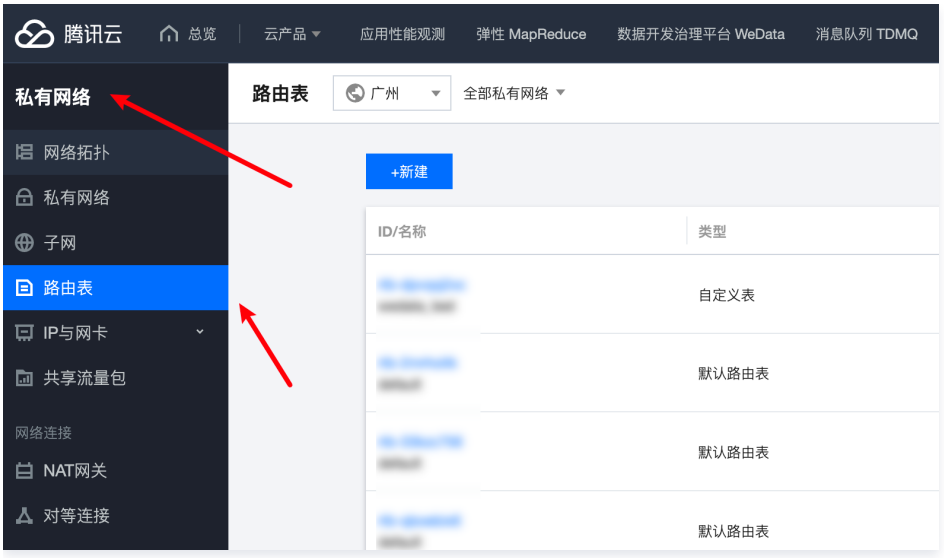


2. Purchase NAT Gateway:

©2013-2025 Tencent Cloud. All rights reserved.

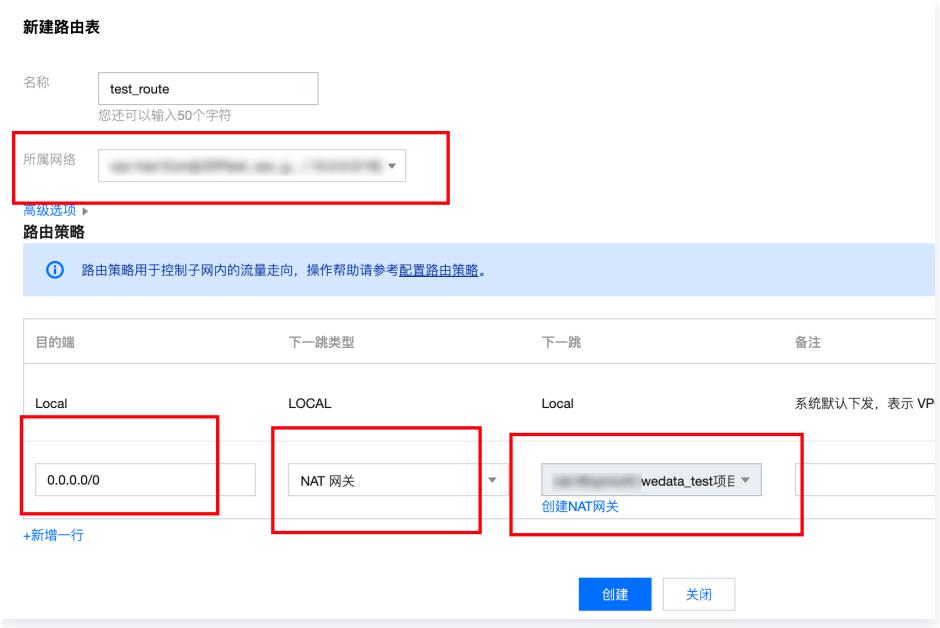


3. Enter Tencent Cloud **VPC** > Route Table, create and bind the route table to point to the NAT Gateway.



4. Create a custom route table, select the corresponding VPC network, and set the destination to 0.0.0.0/0 or a fixed route. For Next Hop Type, select **NAT Gateway**, and for Next Hop, choose the newly created NAT Gateway:

©2013–2025 Tencent Cloud. All rights reserved.



5. After the route table is created, enter the route table details and associate the subnet:



6. Select the subnet to be associated and bind it:



©2013–2025 Tencent Cloud. All rights reserved.

Real-time synchronization task configuration and operation and maintenance
Data sources supported by real-time synchronization
Data source overview

Last updated: 2024-11-25 15:25:08

Background

Real-time synchronization supports various input and output data sources to form a synchronization link. You can configure real-time synchronization tasks based on the data sources supported by DataInLong.

Supported Data Sources

1. Supported Data Sources for Real-Time Whole Database Sync

Data Source		Real-Time Synchronization	
		Database-wide Read	Database-wide Write
Relational database	MySQL/TencentDB for MySQL	✓	-
	TDSQL-C MySQL	✓	-
	TDSQL MySQL	✓	-
	SQL Server	✓	-
	PostgreSQL	✓	✓
	Oracle	✓	-
	OceanBase	✓	-
Big data	DLC	-	✓
	Doris/TCHouse-D	-	✓
	Iceberg	-	✓
	StarRocks	-	✓
	TCHouse-P	-	✓
	Hive	-	✓
NoSQL	Elasticsearch/Tencent Cloud Elasticsearch	-	✓
	MongoDB/Tencent Cloud MongoDB	✓	-
Message Queue	Kafka	✓	✓

2. Supported Data Sources for Real-time Single Table Sync

Data Source		Real-Time Synchronization	
		Single Table Read	Single Table Write
Relational database	MySQL/TencentDB for MySQL	✓	✓
	TDSQL-C MySQL	✓	✓

©2013-2025 Tencent Cloud. All rights reserved.

	TDSQL MySQL	✓	✓
	PostgreSQL	✓	✓
	SQL Server	✓	✓
	Oracle	✓	✓
	OceanBase	✓	–
	Dameng DM	✓	✓
Big data	Hive	–	✓
	DLC	–	✓
	Doris/TCHouse-D	–	✓
	ClickHouse/TCHouse-C	–	✓
	Iceberg	–	✓
	HBase	–	✓
	HDFS	–	✓
	TBase	–	✓
	Greenplum	–	✓
	TCHouse-P	–	✓
NoSQL	Elasticsearch	–	✓
	MongoDB/Tencent Cloud MongoDB	✓	–
Message Queue	Kafka/CKafka	✓	✓
	TiDB-Kafka	✓	–
	DTS-kafka	✓	–

3. Data sources supported by log collection

Data Source		Log Collection	
		Log Reading	Log Writing
Relational database	MySQL	–	✓
	TDSQL-C MySQL	–	✓
	PostgreSQL	–	✓
	SQL Server	–	✓
	Oracle	–	✓
Big data	Hive	–	✓
	DLC	–	✓
	Doris	–	✓
	ClickHouse	–	✓
	Iceberg	–	✓

©2013-2025 Tencent Cloud. All rights reserved.

	HBase	–	✓
	HDFS	–	✓
	TBase	–	✓
	Greenplum	–	✓
	TCHouse-P	–	✓
NoSQL	Elasticsearch	–	✓
Message Queue	Kafka	–	✓
Data Reporting	SDK	✓	–
	TKE	✓	–
	CVM	✓	–

©2013–2025 Tencent Cloud. All rights reserved.

Data Source List
MySQL Data Source

Last updated: 2024-11-25 15:25:54

DataInLong provides MySQL read and write capabilities. This article introduces the pre-environment configuration and current capability support for real-time data synchronization using MySQL.

Supported Versions

Currently, DataInLong supports real-time reading at both single table and database levels for MySQL. To use real-time reading capabilities, you must follow these version restrictions:

Type	Version	Driver
MySQL	5.6,5.7,8.0.x	JDBC Driver:8.0.21
RDS MySQL	5.6,5.7, 8.0.x	
PolarDB MySQL	5.6,5.7,8.0.x	
Aurora MySQL	5.6,5.7,8.0.x	
MariaDB	10.x	
PolarDB X	2.0.1	

Use Limits

- Binlog logging needs to be enabled, and only MySQL server Binlog configuration format ROW is supported for synchronization.
- Tables without primary keys might experience data duplication due to the inability to ensure exactly once. Therefore, it is best to have a primary key for real-time synchronization tasks.
- XA ROLLBACK is not supported. Real-time synchronization tasks will not rollback data for XA PREPARE. To handle XA ROLLBACK scenarios, you need to manually remove the XA ROLLBACK tables from real-time synchronization tasks and then re-sync the tables after adding them back.
- Set MySQL Session Timeout:

When creating an initial consistent snapshot for a large database, the connection you establish may time out while reading tables. You can prevent this by configuring interactive_timeout and wait_timeout in the MySQL configuration file.

  - interactive_timeout: The number of seconds the server waits for activity on an interactive connection before closing it. Please refer to MySQL :: MySQL 8.0 Reference Manual :: 7.1.8 Server System Variables.
  - wait_timeout: The number of seconds the server waits for activity on a non-interactive connection before closing it. Please refer to MySQL :: MySQL 8.0 Reference Manual :: 7.1.8 Server System Variables.

Database Environment Preparation

Confirm MySQL Version

DataInLong has requirements for MySQL versions. Check whether the MySQL version to be synced meets these requirements. You can check the current MySQL database version with the following statement.

```
select version();
```

Setting MySQL Server Permissions

You must define a MySQL user with appropriate permissions for all databases monitored by the Debezium MySQL connector.

1. **Create a MySQL User (Optional):**

```
mysql> CREATE USER 'user'@'localhost' IDENTIFIED BY 'password';
```

©2013-2025 Tencent Cloud. All rights reserved.

2. **Grant Necessary Permissions to the User:**

For real-time data synchronization, the account must have SELECT, REPLICATION SLAVE, and REPLICATION CLIENT permissions on the database. Refer to the following command for execution:

```
mysql> GRANT SELECT, SHOW DATABASES, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO 'user' IDENTIFIED BY 'password';
```

> ⚠ **Note:**
> Enabling scan.incremental.snapshot.enabled no longer requires the RELOAD permission (enabled by default).

3. **Refresh the User's Permissions:**

```
mysql>FLUSH PRIVILEGES;
```

See more about permission descriptions.

Enable MySQL Binlog

1. **Check if Binlog is Enabled**

```
show variables like "log_bin"
```

- If the result is ON, it means Binlog is enabled. If it's a slave, use the following statement:

```
show variables like "log_slave_updates";
```

- If the return value is ON, Binlog is enabled. If Binlog is already enabled, you can skip the following steps.

2. **Enable Binlog**

If you confirm that Binlog is not enabled, you need to perform the following steps:

- For TencentDB for MySQL instances, binlog is enabled by default.
- For open-source MySQL, refer to the official documentation to enable binlog.

3. **Change Binlog Format to Row**

Real-time synchronization only supports MySQL server Binlog configuration format as ROW. Use the following statement to check the Binlog format.

```
show variables like "binlog_format";
```

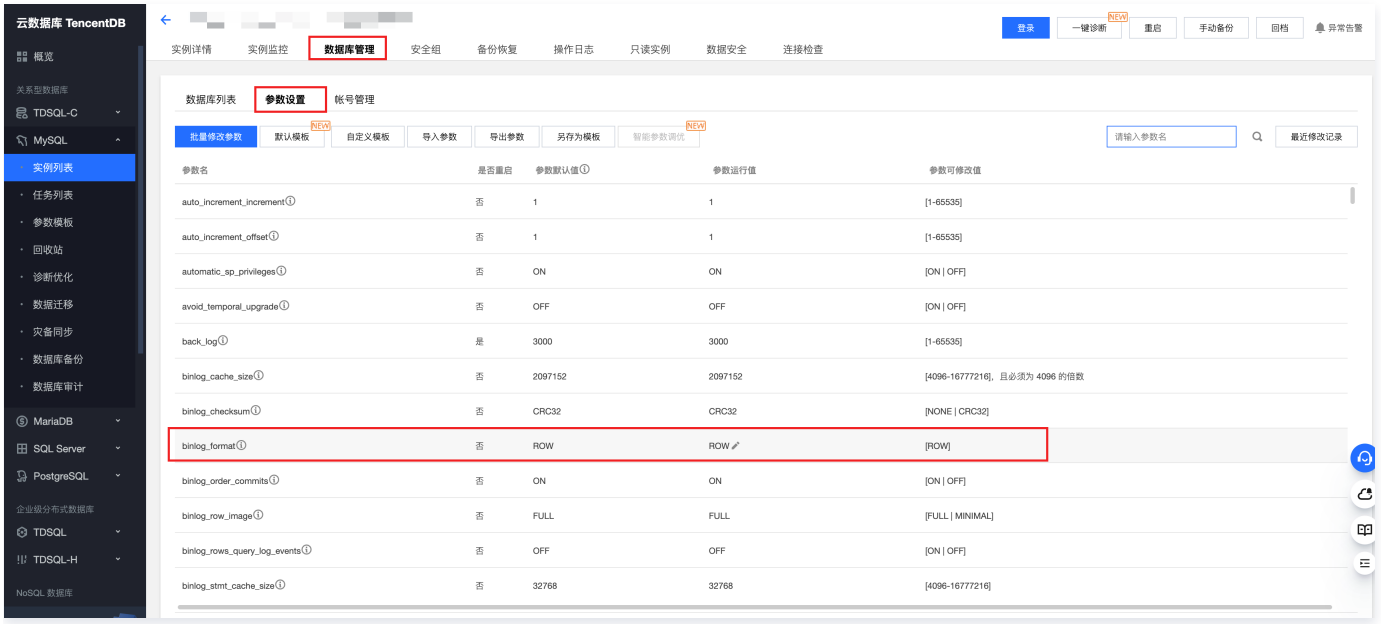
If the return is non-ROW, please change the Binlog Format.

- For open-source MySQL, refer to the official documentation:

MySQL :: MySQL 8.0 Reference Manual :: 17.1.6.4 Binary Logging Options and Variables

- For TencentDB for MySQL instances:
  - Log in to the TencentDB for MySQL console, find the instance for which you want to enable binlog, and click to enter the detailed information page of the instance.
  - In the tabs above, select **DMC** and find the**Parameter Settings** tab.
  - In the**Parameter Settings** tab, find the **binlog_format** parameter and set it to "ROW".

©2013-2025 Tencent Cloud. All rights reserved.



4. **binlog_row_image**

Real-time synchronization only supports MySQL server binlog_row_image configuration format as FULL or full.
Use the following statement to query the usage format of binlog_row_image.

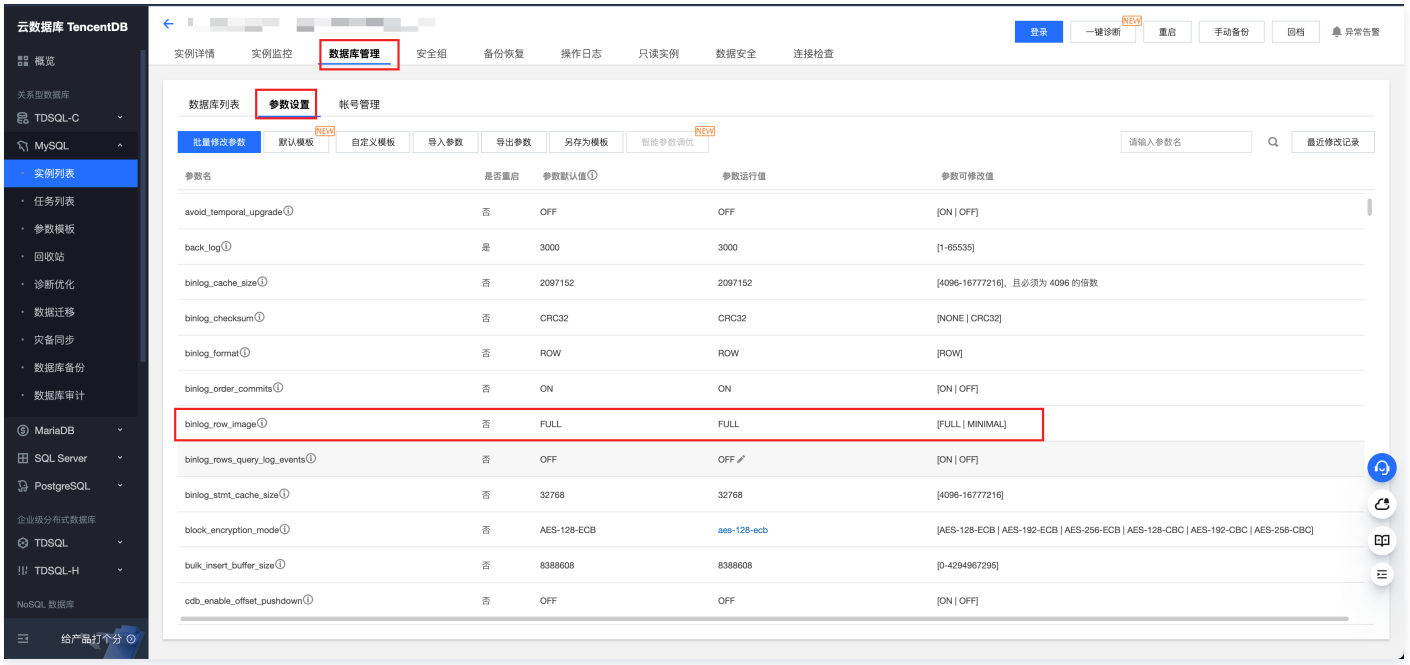
```
show variables like "binlog_row_image";
```

If the return is not FULL/full, please modify binlog_row_image

- For open-source MySQL, refer to the official documentation:

MySQL :: MySQL 8.0 Reference Manual :: 17.1.6.4 Binary Logging Options and Variables

- For TencentDB for MySQL instances:
  - Log in to the TencentDB for MySQL console, find the instance for which you want to enable binlog, and click to enter the detailed information page of the instance.
  - In the tabs above, select **DMC** and find the**Parameter Settings** tab.
  - In the **Parameter Settings** tab, find the **binlog_row_image** parameter and set it to "FULL".



Enable GTIDs (optional)

©2013-2025 Tencent Cloud. All rights reserved.

GTID (Global Transaction Identifier), used to uniquely identify a transaction in binlog, can avoid data confusion or master-slave inconsistency caused by duplicate transaction execution.

Enable Process

1. Check if GTID is enabled

```
show global variables like '%GTID%';
```

If the return result is similar to the following, it proves that GTID is enabled.

```
+--------------------------+-------+
| Variable_name                      | Value |
+--------------------------+-------+
| enforce_gtid_consistency   | ON     |
| gtid_mode                                | ON     |
+--------------------------+-------+
```

2. Enable GTID

- For open-source MySQL, refer to the official documentation MySQL :: MySQL 8.0 Reference Manual :: 17.1.4.2 Enabling GTID Transactions Online.
- For TencentDB for MySQL instances, binlog is enabled by default and cannot be disabled.

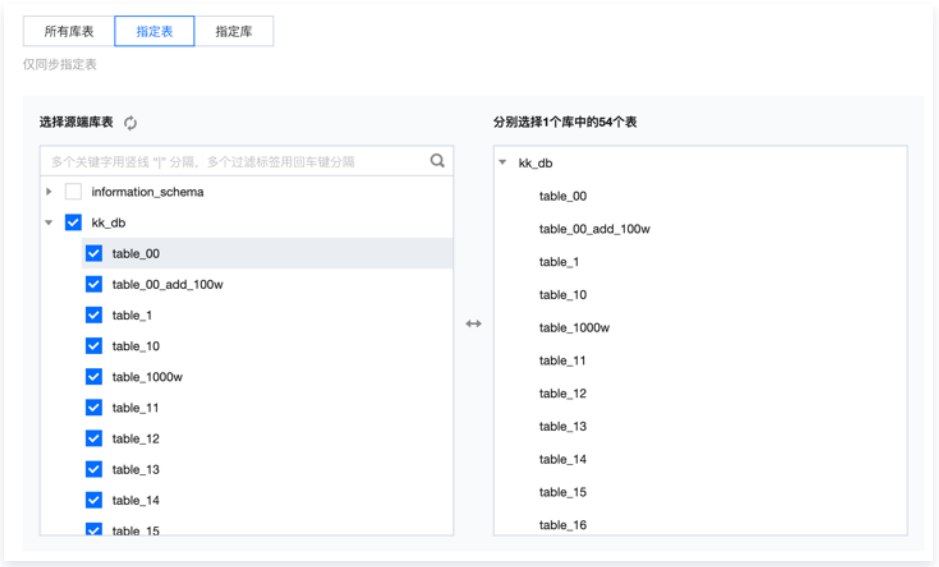
Real-time whole database source configuration

Data Source Settings



Parameter	Description
Data Source	Select the MySQL data source for synchronization.
Source Table	• All databases and tables: Monitor all databases under the data source. During the task execution, new databases and tables will be synchronized to the target end by default. • Specified Table: Under this option, you need to specify the specific table name. After setting, the task will only synchronize the specified table; if a new table needs to be synchronized, the task must be stopped and restarted.

©2013-2025 Tencent Cloud. All rights reserved.



- Specify Database: Under this option, you need to specify the specific database name with a table name regular expression. After setting, new tables that match the table name expression will be synchronized to the target end by default during task execution.



Read mode	• Full + Incremental: Data synchronization is divided into full and incremental synchronization phases. After the full phase is completed, the task enters the incremental phase. The full phase will synchronize historical data in the database, and the incremental phase will start synchronizing from the binlog CDC location after the task starts. • Increment only: Only sync data starting from the binlog cdc point after the task starts.
Consistency Semantics	Represents the consistency semantics of the reader. Supports At-least-once and Exactly-once. • At-least-once: Data may be read multiple times, relying on the destination to support deduplication to ensure data consistency. Suitable for scenarios where the full data volume is large and non-numeric primary keys are used, with high performance requirements for synchronization. • Exactly-once: Data is strictly read only once, with some performance loss, and does not support tables without primary keys or unique index columns. Suitable for general scenarios where the source table has numeric primary keys or unique index columns. The current version does not support mode switching with state retention. Mode change after task submission requires a state reset.
Filter operations	Support three operations: Insert, Update, and Delete. After setting, data of the specified operation types will not be synchronized.
Time zone	Set the log time timezone, default is Shanghai.
Syncing gh-ost temporary tables	The business scenario for gh-ost is to perform online table structure changes in MySQL, specifically Online DDL, without affecting the normal operation of the business. It can resolve issues caused by traditional commands like alter table or create index, such as table locks, performance degradation, and synchronization delays. It is suitable for scenarios where table modifications are necessary, such as adding new columns, adding indexes, modifying column types, etc. gh-ost prerequisites: • gh-ost must be able to access MySQL.

©2013-2025 Tencent Cloud. All rights reserved.

	• If MySQL is TencentDB's CDB, you need to add --aliyun-rds to the parameters of the gh-ost execution command. • During the execution of the gh-ost tool, the naming convention for temporary tables is ^_(.*)_(gho|ghc|del)$, where (.*) is the name of the change table. Self-definition of temporary table names is not supported. • For other gh-ost limitations, refer to gh-ost/requirements-and-limitations.md at master · github/gh-ost · GitHub. To synchronize gh-ost temporary tables, you need to enable this switch. When enabled, it will monitor '_tableName_gho', '_tableName_ghc', '_tableName_del', and synchronize changes to '_tableName_gho' table.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

During MySQL whole database synchronization, the task uses the primary key or unique index (if no primary key) for splitting. The currently supported primary key types include:

Primary key supported range types	TINYINT,TINYINT_UNSIGNED,SMALLINT,SMALLINT_UNSIGNED,INT,MEDIUMINT,INT_UNSIGNED,MEDIUMINT_UNSIGNED,BIGINT,BIGINT_UNSIGNED,FLOAT,DOUBLE,DECIMAL,TIME,DATE,DATETIME,TIMESTAMP,CHAR,VARCHAR,TEXT,BINARY,VARBINARY,BLOB.

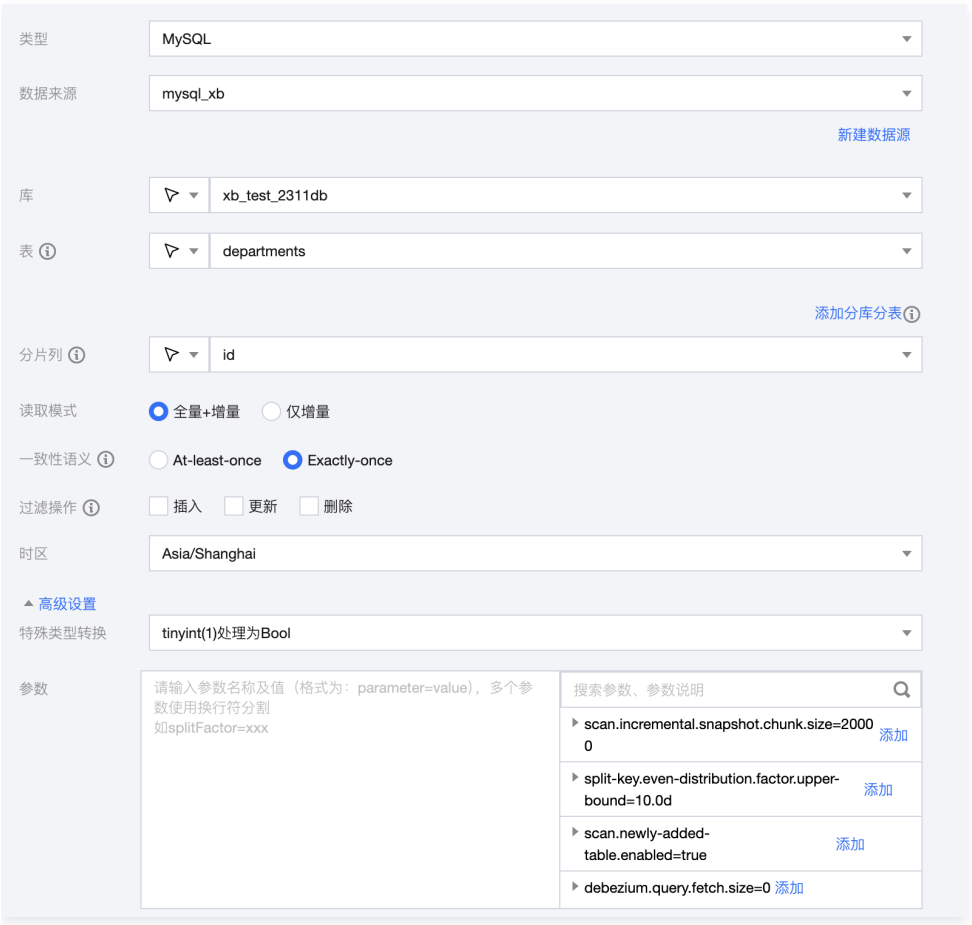
Supported Target Data Sources

Currently supports real-time whole database synchronization of MySQL data to the following target types:

- Kafka Whole Database Target Configuration Details
- Doris Whole Database Target Configuration Details
- DLC Whole Database Target Configuration Details
- Iceberg Whole Database Target Configuration Details
- TCHouse-P Whole Database Target Configuration Details
- PostgreSQL Whole Database Target Configuration Details
- Hive Whole Database Target Configuration Details
- StarRocks Whole Database Target Configuration Details

Real-time Single Table Read Node Configuration

©2013-2025 Tencent Cloud. All rights reserved.

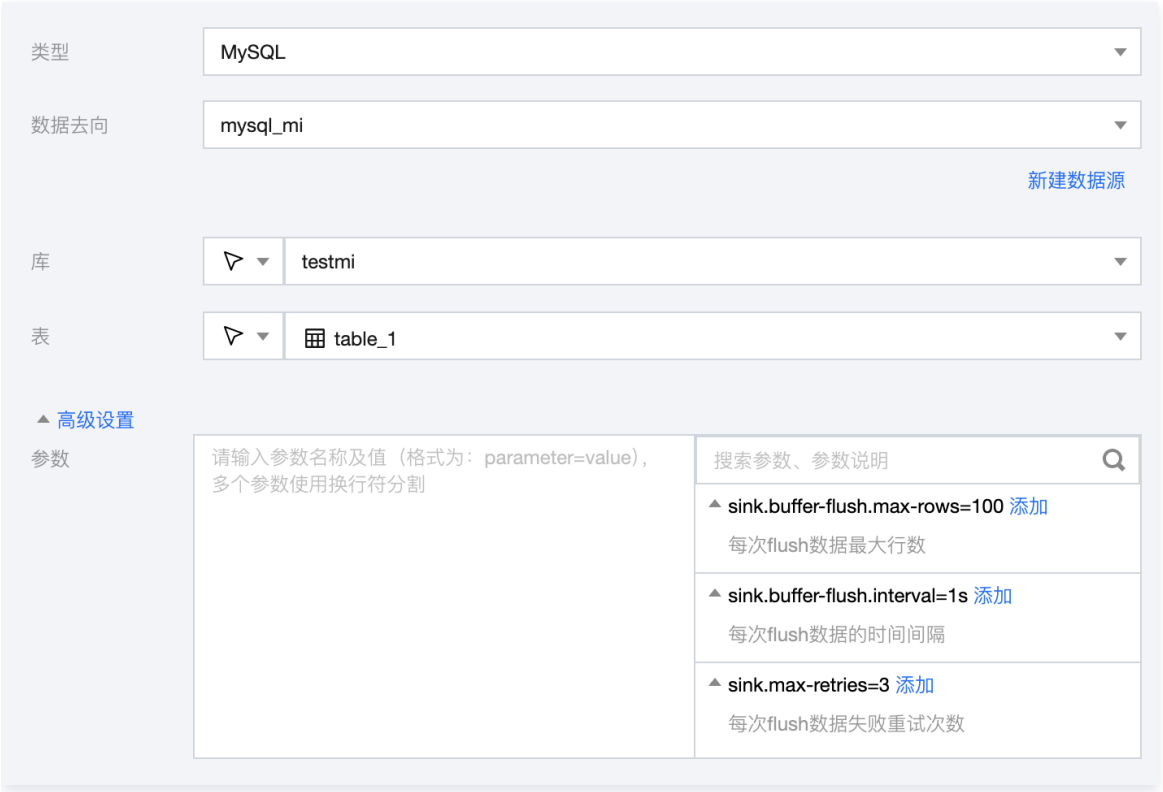


Parameter	Description
Data Source	Select MySQL Data Source.
Database	Support selecting or manually entering the library name to be read. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Support selecting or manually entering the table name to be read. • In the case of table partitioning, you can select or enter multiple table names on the MySQL source end. Multiple tables must have consistent structures. • In the case of table partitioning, you can configure the table serial number range. For example, 'table_[0-99]' means reading 'table_0', 'table_1', 'table_2' up to 'table_99'; if your table numeric suffixes are consistent in length, such as 'table_000', 'table_001', 'table_002' up to 'table_999', you can configure it as '"table": ["table_00[0-9]", "table_0[10-99]", "table_[100-999]"]'. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Adding Sharding	Suitable for sharding scenarios, click to configure multiple data sources, databases, and table information. In sharding and partitioning scenarios, ensure all table structures are consistent. The task configuration will default to display and use the first table structure for data acquisition.
Shard Column	Shard columns are used to divide the table into multiple shards for synchronization. For tables with primary keys, it is recommended to prioritize the primary key as the shard column; for tables without primary keys, it is recommended to select an indexed column as the shard column, and ensure that no updates are performed on the shard column, otherwise only At-Least-Once semantics can be guaranteed.

©2013-2025 Tencent Cloud. All rights reserved.

Read mode	• Full + Incremental: Data synchronization is divided into full and incremental synchronization phases. After the full phase is completed, the task enters the incremental phase. The full phase will synchronize historical data in the database, and the incremental phase will start synchronizing from the binlog CDC location after the task starts. • Increment only: Only sync data starting from the binlog cdc point after the task starts.
Consistency Semantics	Represents the consistency semantics of the reader. Supports At-least-once and Exactly-once. • At-least-once: Data may be read multiple times, relying on the destination to support deduplication to ensure data consistency. Suitable for scenarios where the full data volume is large and non-numeric primary keys are used, with high performance requirements for synchronization. • Exactly-once: Data is strictly read only once, with some performance loss, and does not support tables without primary keys or unique index columns. Suitable for general scenarios where the source table has numeric primary keys or unique index columns. The current version does not support mode switching with state retention. Mode change after task submission requires a state reset.
Filter operations	After setting, data of the specified operation types will not be synchronized. Support filtering Insert, Update, and Delete operations.
Time zone	Set the log time timezone, default is Shanghai.
Special Type Conversion	Map tinyint(1) to Bool or tinyint. Default is to map to Bool.
Parameters (Optional)	You can configure parameters based on business requirements.

Real-time Single Table Write Node Configuration



Parameter	Description
Data Flow Direction	Select MySQL Data Source.
Database	Support selecting or manually entering the library name to be read.

©2013-2025 Tencent Cloud. All rights reserved.

	• By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Support selecting or manually entering the table name to be read. When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Parameters (Optional)	You can configure parameters based on business requirements.

Log Collection Write Node

1. Enter the **DataInLong Real-time Synchronization > Log Collection** page, select MySQL type for the target.



Parameter	Description
Data Flow Direction	Select an available MySQL data source from the current project.
Database	Select the corresponding database in the data source.
Table	Select the corresponding table in the data source.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Read/Write Data Type Conversion Support

Read

The supported data types and conversion relationships for MySQL read are as follows (when processing MySQL, the data type of MySQL data source will first be mapped to the data processing engine's data type)

Field Type	Supported	Internal mapping fields	Remarks
TINYINT	Yes	TINYINT	TINYINT(1) maps to BOOLEAN Options need to be added to support TINYINT(1) mapping to bool or tinyint
SMALLINT	Yes	SMALLINT	-

©2013-2025 Tencent Cloud. All rights reserved.

TINYINT_UNSIGNED	Yes	SMALLINT	-
TINYINT_UNSIGNED_ZEROFILL	Yes	SMALLINT	-
INT	Yes	INT	-
INTEGER	Yes	INT	-
YEAR	Yes	INT	-
MEDIUMINT	Yes	INT	-
SMALLINT_UNSIGNED	Yes	INT	-
SMALLINT_UNSIGNED_ZEROFILL	Yes	INT	-
BIGINT	Yes	LONG	-
INT_UNSIGNED	Yes	LONG	-
MEDIUMINT_UNSIGNED	Yes	LONG	-
MEDIUMINT_UNSIGNED_ZEROFILL	Yes	LONG	-
INT_UNSIGNED_ZEROFILL	Yes	LONG	-
BIGINT_UNSIGNED	Yes	DECIMAL	DECIMAL(20,0)
BIGINT_UNSIGNED_ZEROFILL	Yes	DECIMAL	DECIMAL(20,0)
SERIAL	Yes	DECIMAL	DECIMAL(20,0)
FLOAT	Yes	FLOAT	-
FLOAT_UNSIGNED	Yes	FLOAT	-
FLOAT_UNSIGNED_ZEROFILL	Yes	FLOAT	-
DOUBLE	Yes	DOUBLE	-
DOUBLE_UNSIGNED	Yes	DOUBLE	-
DOUBLE_UNSIGNED_ZEROFILL	Yes	DOUBLE	-
DOUBLE_PRECISION	Yes	DOUBLE	-
DOUBLE_PRECISION_UNSIGNED	Yes	DOUBLE	-
ZEROFILL	Yes	DOUBLE	-
REAL	Yes	DOUBLE	-
REAL_UNSIGNED	Yes	DOUBLE	-
REAL_UNSIGNED_ZEROFILL	Yes	DOUBLE	-
NUMERIC	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
NUMERIC_UNSIGNED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
NUMERIC_UNSIGNED_ZEROFI	Yes	DECIMAL	Use the actual precision of the user database p<=38

©2013–2025 Tencent Cloud. All rights reserved.

LL			maps to DECIMAL 38 < p <= 65 maps to String
DECIMAL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
DECIMAL_UNSIGNED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
DECIMAL_UNSIGNED_ZEROFILL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
FIXED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
FIXED_UNSIGNED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
FIXED_UNSIGNED_ZEROFILL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
BOOLEAN	Yes	BOOLEAN	–
DATE	Yes	DATE	–
TIME	Yes	TIME	–
DATETIME	Yes	TIMESTAMP	–
TIMESTAMP	Yes	TIMESTAMP	–
CHAR	Yes	STRING	–
JSON	Yes	STRING	–
BIT	Yes	STRING	BIT(1) maps to BOOLEAN
VARCHAR	Yes	STRING	–
TEXT	Yes	STRING	–
BLOB	Yes	STRING	–
TINYBLOB	Yes	STRING	–
TINYTEXT	Yes	STRING	–
MEDIUMBLOB	Yes	STRING	–
MEDIUMTEXT	Yes	STRING	–
LONGBLOB	Yes	STRING	–
LONGTEXT	Yes	STRING	–
VARBINARY	Yes	STRING	–
GEOMETRY	Yes	STRING	–
POINT	Yes	STRING	–
LINESTRING	Yes	STRING	–
POLYGON	Yes	STRING	–
MULTIPOINT	Yes	STRING	–
MULTILINESTRING	Yes	STRING	–
MULTIPOLYGON	Yes	STRING	–

©2013–2025 Tencent Cloud. All rights reserved.

GEOMETRYCOLLECTION	Yes	STRING	-
ENUM	Yes	STRING	-
BINARY	Yes	BINARY	BINARY(1)
SET	No		-

Write

The supported data types and conversion relationships for MySQL write are as follows:

Internal Type	MySQL Type
TINYINT	TINYINT
SMALLINT	SMALLINT,TINYINT UNSIGNED
INT	INT,MEDIUMINT,SMALLINT UNSIGNED
BIGINT	BIGINT,INT UNSIGNED
DECIMAL(20, 0)	BIGINT UNSIGNED
FLOAT	FLOAT
DOUBLE	DOUBLE,DOUBLE PRECISION
DECIMAL(p, s)	NUMERIC(p, s),DECIMAL(p, s)
BOOLEAN	BOOLEAN,TINYINT(1)
DATE	DATE
TIME [(p)][WITHOUT TIMEZONE]	TIME [(p)]
TIMESTAMP [(p)][WITHOUT TIMEZONE]	DATETIME [(p)]
STRING	CHAR(n),VARCHAR(n),TEXT
BYTES	BINARY,VARBINARY,BLOB
ARRAY	-

FAQs

MySql server ID conflict

Error message:

```
com.github.shyiko.mysql.binlog.network.ServerException: A slave with the same server_uuid/server_id as
this slave has connected to the master.
```

Solution: We have now optimized to randomly generate serverid. If in previous tasks the server-id was explicitly specified in MySQL advanced parameters, it is recommended to delete it, as multiple tasks may use the same data source with the same server-id, causing a conflict.

Error message reporting binlog file not found:

Error message:

```
Caused by: org.apache.kafka.connect.errors.ConnectException: The connector is trying to read binlog
starting at GTIDs xxx and binlog file 'binlog.xxx', pos=xxx, skipping 4 events plus 1 rows, but this is no
longer available on the server. Reconfigure the connector to use a snapshot when needed.
```

©2013-2025 Tencent Cloud. All rights reserved.

Error Causes:
An error will occur if the binlog file being read by the job has already been cleaned up on the MySQL server. There are several reasons for Binlog cleanup, such as a too-short Binlog retention period, or the job processing speed not keeping up with the Binlog generation speed, exceeding the maximum retention period of MySQL Binlog files. This results in the Binlog files being cleaned up on the MySQL server, causing the Binlog positions being read to become invalid.
Solution: If the job processing speed cannot keep up with the Binlog generation speed, you can consider increasing the Binlog retention period or optimizing the job to reduce backpressure and accelerate source consumption. If there are no anomalies in the job status, it might be that other operations on the database caused the Binlog to be cleaned up, leading to inaccessibility. You need to refer to the MySQL database-side information to determine the reason for Binlog cleanup.

MySQL reports connection reset

Error message:

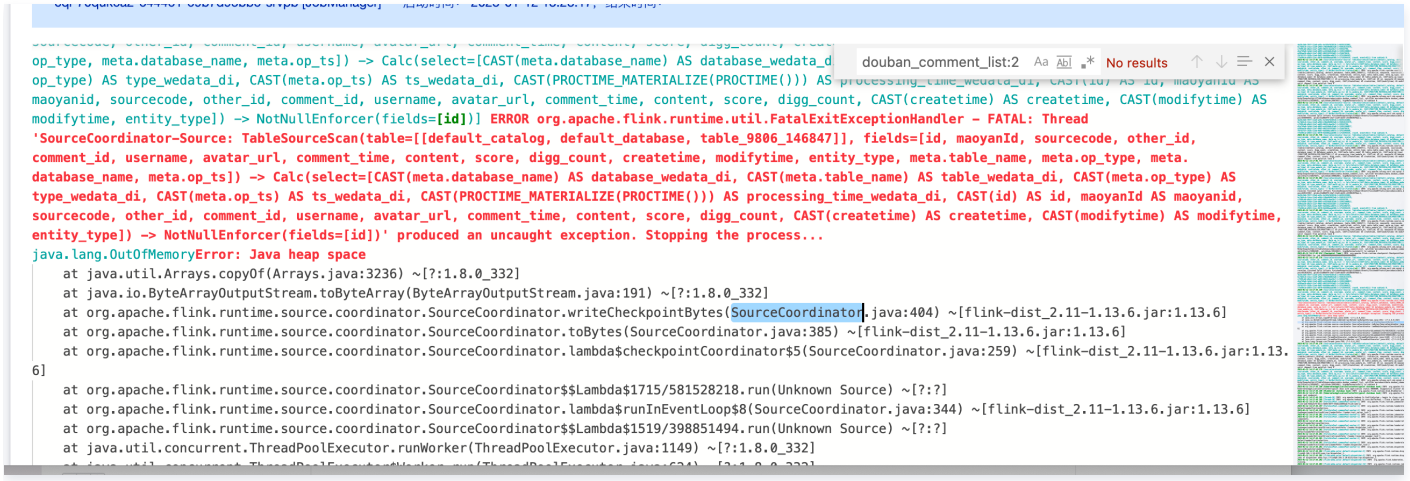
```
EventDataDeserializationException: Failed to deserialize data of EventHeaderV4 .... Caused by:
java.net.SocketException: Connection reset.
```

Error Causes:
1. Network Issues.
2. The job has backpressure, causing the source to be unable to read data, and the binlog client to be idle. If the binlog connection remains idle after the timeout, the MySQL server will disconnect the idle connection.

Solution:
1. If it is a network issue, you can increase the MySQL network parameters set global slave_net_timeout = 120; (default 30s) set global thread_pool_idle_timeout = 120.
2. If the backpressure is caused by the job, you can alleviate it by adjusting the job, such as increasing parallelism, improving write speed, and increasing TaskManager memory to reduce GC.

Mysql2dlc task JobManager OOM

Error message:



Causes and Solutions:
1. If the user data volume is large, you can increase the JobManager CU number and use the MySQL advanced parameter scan.incremental.snapshot.chunk.size to increase the chunk size, which defaults to 8096.
2. The user's data volume is not large, but there is a large difference between the maximum and minimum values of the primary key, resulting in many chunks being divided when using the equal chunk strategy. Modify the distribution factor so that user data follows an uneven data splitting logic, split-key.even-distribution.factor.upper-bound=5.0d. The default distribution factor has been modified to 10.0d.

User's binlog data format is incorrect, causing Debezium parse exception

Error message:

©2013-2025 Tencent Cloud. All rights reserved.

```
ERROR io.debezium.connector.mysql.MySqlStreamingChangeEventSource [] - Error during binlog processing.
Last offset stored = null, binlog reader near position = mysql-bin.000044/211839464.
2023-02-20 21:37:28.480 [blc-172.17.48.3:3306] ERROR io.debezium.pipeline.ErrorHandler [] - Producer
failure
io.debezium.DebeziumException: Error processing binlog event.
```

Solution:
After modifying binlog_row_image=full, you need to restart the database.

Is gh-ost supported?

Yes, it does not migrate the temporary table data generated by Online DDL changes. It only migrates the original DDL data executed using gh-ost in the source database. You can use the default or configure your own regular expressions for gh-ost shadow tables and unused tables.

©2013-2025 Tencent Cloud. All rights reserved.

TDSQL-C for MySQL Data Source

Last updated: 2024-09-06 14:45:44

DataInLong provides reading and writing capabilities for TDSQL-C for MySQL. This article introduces the pre-environment configuration for real-time data synchronization using TDSQL-C for MySQL and the current capability support.

Supported Versions

Currently, DataInLong supports real-time reading of single tables and database-level real-time reading and single table writing in TDSQL-C for MySQL. To use real-time synchronization capability, follow these version restrictions:

Engine types	Version	Driver
MySQL	5.7,8.0	JDBC Driver:8.0.21

Use Limits

- Binlog logging needs to be enabled, and only MySQL server Binlog configuration format ROW is supported for synchronization.
- **It is not recommended to read from TDSQL-C for MySQL read-only replicas due to potential stability issues.**
- Tables without primary keys might experience data duplication due to the inability to ensure exactly once. Therefore, it is best to have a primary key for real-time synchronization tasks.
- XA ROLLBACK is not supported. Real-time synchronization tasks will not rollback data for XA PREPARE. To handle XA ROLLBACK scenarios, you need to manually remove the XA ROLLBACK tables from real-time synchronization tasks and then re-sync the tables after adding them back.
- Set MySQL Session Timeout:

When creating an initial consistent snapshot for a large database, the connection you establish may time out while reading tables. You can prevent this by configuring interactive_timeout and wait_timeout in the MySQL configuration file.

  - interactive_timeout: The number of seconds the server waits for activity on an interactive connection before closing it. Please refer to MySQL :: MySQL 8.0 Reference Manual :: 7.1.8 Server System Variables.
  - wait_timeout: The number of seconds the server waits for activity on a non-interactive connection before closing it. Please refer to MySQL :: MySQL 8.0 Reference Manual :: 7.1.8 Server System Variables.

Database Environment Preparation

Confirm MySQL Version

DataInLong has requirements for MySQL versions. Check whether the MySQL version to be synced meets these requirements. You can check the current MySQL database version with the following statement.

```
select version();
```

Setting MySQL Server Permissions

You must define a MySQL user with appropriate permissions for all databases monitored by the Debezium MySQL connector.

1. **Create a MySQL User (Optional)**

```
mysql> CREATE USER 'user'@'localhost' IDENTIFIED BY 'password';
```

2. **Grant the required permissions to the user.**

For real-time data synchronization, the account must have SELECT, REPLICATION SLAVE, and REPLICATION CLIENT permissions on the database. Refer to the following command for execution:

```
mysql> GRANT SELECT, SHOW DATABASES, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO 'user' IDENTIFIED BY
'password';
```

©2013-2025 Tencent Cloud. All rights reserved.

> **Note:**
> Enabling scan.incremental.snapshot.enabled no longer requires the RELOAD permission (enabled by default).

3. **Refresh the user's permissions.**

```
mysql>FLUSH PRIVILEGES;
```

For more information, please see Debezium connector for MySQL.

Enable MySQL Binlog

1. **Check if Binlog is Enabled**

```
show variables like "log_bin"
```

- If the result is ON, it means Binlog is enabled. If it's a slave, use the following statement:

```
show variables like "log_slave_updates";
```

- If the return value is ON, Binlog is enabled. If Binlog is already enabled, you can skip the following steps.

2. **Enable Binlog**

If you confirm that Binlog is not enabled, you need to perform the following steps:

- For TDSQL-C for MySQL, binlog is enabled by default.

3. **Change Binlog Format to Row**

Real-time synchronization only supports MySQL server Binlog configuration format as ROW. Use the following statement to check the Binlog format.

```
show variables like "binlog_format";
```

If the return is non-ROW, please change the Binlog Format.

- For TDSQL-C for MySQL:
  - Log in to the TDSQL-C for MySQL console, find the instance for which you want to enable Binlog, and click to enter the detailed information page of the instance.
  - In the tabs above, select **DMC** and find the**Parameter Settings** tab.
  - In the**Parameter Settings** tab, find the **binlog_format** parameter and set it to "ROW".

4. **binlog_row_image**

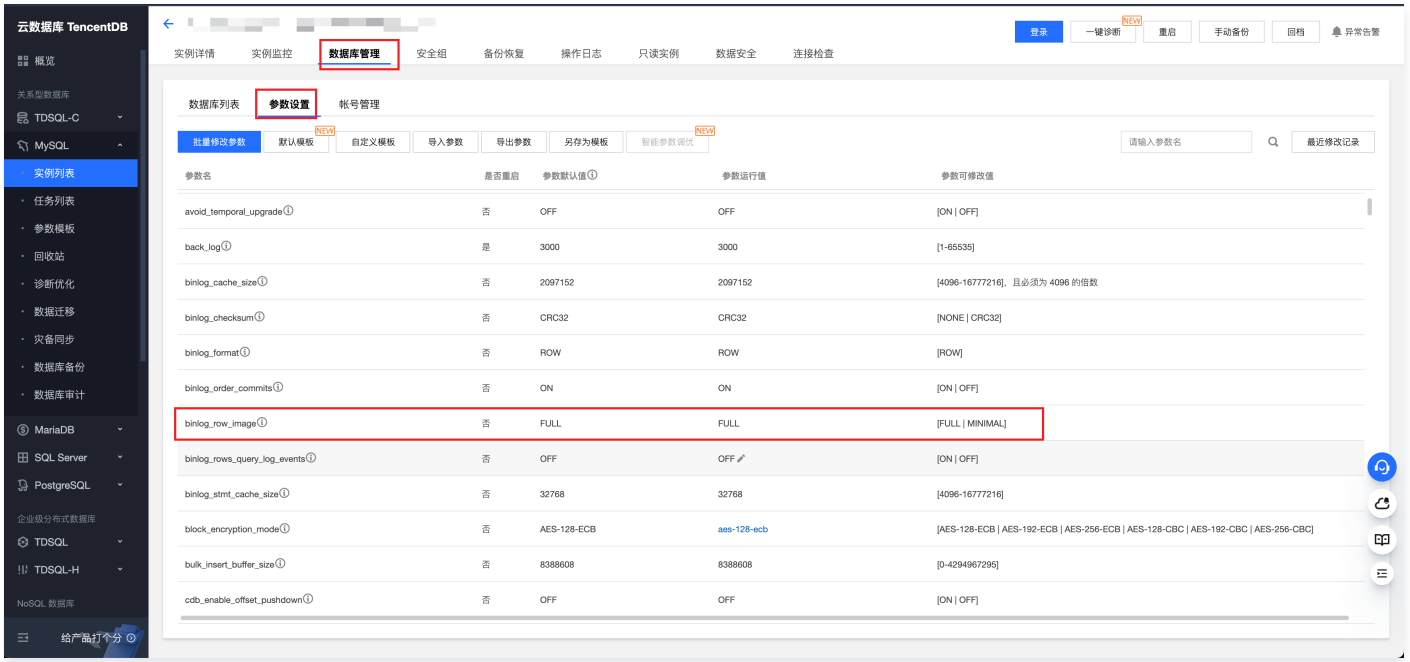
Real-time synchronization only supports MySQL server binlog_row_image configuration format as FULL or full.
Use the following statement to query the usage format of binlog_row_image.

```
show variables like "binlog_row_image";
```

If the return is not FULL/full, please modify binlog_row_image

- For TDSQL-C for MySQL:
  - Log in to the TDSQL-C for MySQL console, find the instance for which you want to enable Binlog, and click to enter the detailed information page of the instance.
  - In the tabs above, select **DMC** and find the**Parameter Settings** tab.
  - In the **parameter settings** tab, find the **binlog_row_image** parameter and set it to "FULL".

©2013-2025 Tencent Cloud. All rights reserved.



Enable GTIDs (optional)

GTID (Global Transaction Identifier), used to uniquely identify a transaction in binlog, can avoid data confusion or master-slave inconsistency caused by duplicate transaction execution.

Enable Process

1. Check if GTID is enabled.

```
show global variables like '%GTID%';
```

If the return result is similar to the following, it proves that GTID is enabled.

```
+--------------------------+-------+
| Variable_name                | Value |
+--------------------------+-------+
| enforce_gtid_consistency   | ON    |
| gtid_mode                      | ON    |
+--------------------------+-------+
```

2. Enable GTID

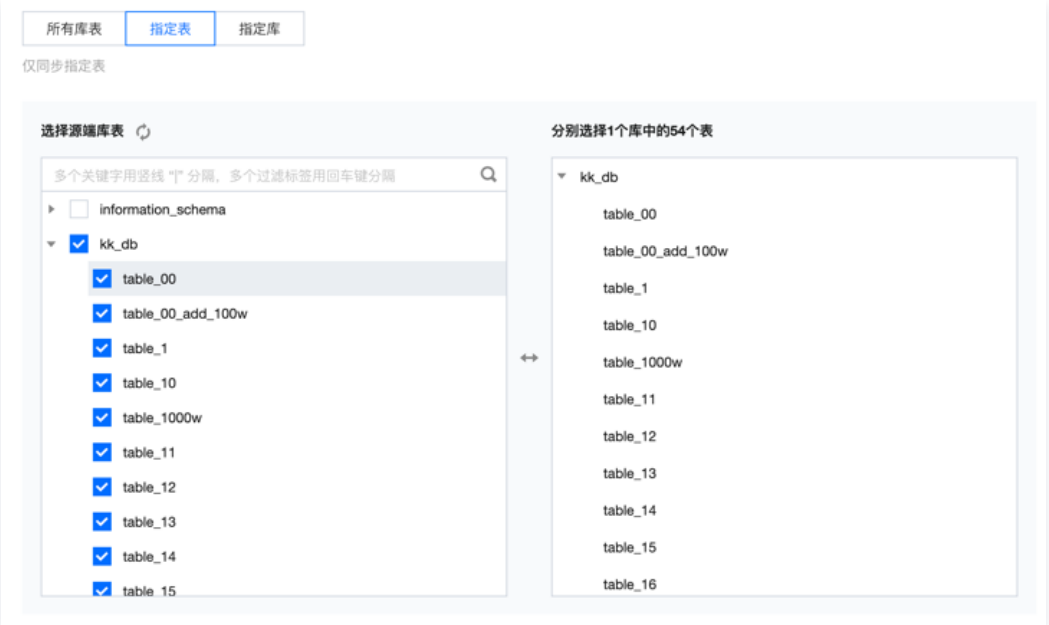
- For TDSQL-C for MySQL, it is enabled by default and not supported to disable.

Database-level source configuration

Data Source Settings

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Source	Select the TDSQL-C MySQL data source to synchronize.
Source Table	• All databases and tables: Monitor all databases under the data source. During the task execution, new databases and tables will be synchronized to the target end by default. • Specified Table: Under this option, you need to specify the specific table name. After setting, the task will only synchronize the specified table; if a new table needs to be synchronized, the task must be stopped and restarted.  • Specify Database: Under this option, you need to specify the specific database name with a table name regular expression. After setting, new tables that match the table name expression will be synchronized to the target end by default during task execution.

©2013–2025 Tencent Cloud. All rights reserved.

	
Read mode	• Full + Incremental: Data synchronization is divided into full and incremental synchronization phases. After the full phase is completed, the task enters the incremental phase. The full phase will synchronize historical data in the database, and the incremental phase will start synchronizing from the binlog CDC location after the task starts. • Increment only: Only sync data starting from the binlog cdc point after the task starts.
Consistency Semantics	It only represents the consistency semantics of the reading end. Supports At-least-once and Exactly-once. • At-least-once: Data may be read repeatedly, and relies on the target end to support deduplication to ensure data consistency. Suitable for scenarios with a large volume of data in the full stage and non-numeric primary keys, requiring high synchronization performance. • Exactly-once: Data is read strictly once, with some performance loss, and does not support tables without primary keys and unique index columns. Suitable for general scenarios where the source table has numeric primary keys or unique index columns. The two modes are incompatible in the current version. If the mode is changed after task submission, stateful restart is not supported.
Filter operations	Support three operations: Insert, Update, and Delete. After setting, data of the specified operation types will not be synchronized.
Time zone	Set the log time timezone, default is Shanghai.
Synchronize gh-ost Temporary Table	The business scenario for gh-ost is to perform online table structure changes in MySQL, specifically Online DDL, without affecting the normal operation of the business. It can resolve issues caused by traditional commands like alter table or create index, such as table locks, performance degradation, and synchronization delays. It is suitable for scenarios where table modifications are necessary, such as adding new columns, adding indexes, modifying column types, etc. gh-ost Prerequisites: • gh-ost must be able to access MySQL. • If MySQL is Tencent Cloud's CDB, the parameter --aliyun-rds needs to be added to the gh-ost command. • During the execution of the gh-ost tool, the generated temporary table rules are ^_(.*)_(gho|ghc|del)$, where (.*) is the name of the changing table. It does not support user-defined temporary table names. • For other gh-ost limitations, please refer to `gh-ost/requirements-and-limitations.md at master · github/gh-ost · GitHub.` To synchronize gh-ost temporary tables, this switch needs to be turned on. After enabling, it will monitor ‘_tablename_gho’, ‘_tablename_ghc’, ‘_tablename_del’, and synchronize changes in ‘_tablename_gho’.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Supported Target Data Sources

Currently supports real-time synchronization of TDSQL-C MySQL databases to the following types of target ends:

- Kafka Database Synchronization Configuration Details
- Doris Whole Database Target Configuration Details
- DLC Whole Database Target Configuration Details

©2013-2025 Tencent Cloud. All rights reserved.

- Iceberg Whole Database Target Configuration Details
- TCHouse-P Whole Database Target Configuration Details
- Hive Whole Database Target Configuration Details
- StarRocks Whole Database Target Configuration Details
- Elasticsearch Whole Database Target Configuration Details

Single Table Read Node Configuration

1. On the DataInLong page, click Real-time synchronization in the left-hand navigation bar.
2. On the Real-time synchronization page, select Single Table Synchronization to create a new configuration (you can choose between Form mode and Canvas mode) and enter the configuration page.



Parameter	Description
Data Source	Select the TDSQL-C MySQL data source.
Database	You can select or manually enter the database name to be read. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	You can select or manually enter the table name to be read. • In the case of table partitioning, you can select or enter multiple table names on the MySQL source end. Multiple tables must have consistent structures. • In the case of table partitioning, you can configure the table serial number range. For example, 'table_[0-99]' means reading 'table_0', 'table_1', 'table_2' up to 'table_99'; if your table numeric suffixes

©2013-2025 Tencent Cloud. All rights reserved.

	are consistent in length, such as 'table_000', 'table_001', 'table_002' up to 'table_999', you can configure it as '"table": ["table_00[0-9]", "table_0[10-99]", "table_[100-999]"]'. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Adding Sharding	Applicable to sharding scenarios, you can configure multiple data sources, databases, and table information by clicking. In sharding scenarios, the structure of all tables must be consistent. The task configuration will default to displaying and using the first table structure for data retrieval.
Shard Column	The shard column is used to divide the table into multiple shards for synchronization. For tables with primary keys, it is recommended to choose the primary key as the shard column; for tables without primary keys, it is recommended to choose a column with an index as the shard column and ensure no data update operations occur on the shard column, otherwise only At-Least-Once Semantics can be guaranteed.
Read mode	• Full + Incremental: Data synchronization is divided into full and incremental synchronization phases. After the full phase is completed, the task enters the incremental phase. The full phase will synchronize historical data in the database, and the incremental phase will start synchronizing from the binlog CDC location after the task starts. • Increment only: Only sync data starting from the binlog cdc point after the task starts.
Consistency Semantics	It only represents the consistency semantics of the reading end. Supports At-least-once and Exactly-once. • At-least-once: Data may be read repeatedly, and relies on the target end to support deduplication to ensure data consistency. Suitable for scenarios with a large volume of data in the full stage and non-numeric primary keys, requiring high synchronization performance. • Exactly-once: Data is read strictly once, with some performance loss, and does not support tables without primary keys and unique index columns. Suitable for general scenarios where the source table has numeric primary keys or unique index columns. The two modes are incompatible in the current version. If the mode is changed after task submission, stateful restart is not supported.
Filter operations	After setting, the specified operation types will not be synchronized; support filtering for insert, update, and delete operations.
Time zone	Set the log time timezone, default is Shanghai.
Special Type Conversion	Map tinyint(1) to Bool or tinyint, with default mapping to Bool.
Parameter (optional)	You can configure parameters based on business requirements.

Single Table Write Node Configuration

1. On the DataInLong page, click Real-time synchronization in the left-hand navigation bar.
2. On the Real-time synchronization page, select Single Table Synchronization to create a new configuration (you can choose between Form mode and Canvas mode) and enter the configuration page.

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Source	TDSQL-C for MySQL data source to be written.
Database	Supports selecting or manually entering the name of the database to be written to. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Supports selecting or manually entering the name of the table to be written to. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Primary key	Select a field as the primary key for the write table.
Advanced Settings (optional)	You can configure parameters based on business requirements.

Log Collection Write Node Configuration

©2013–2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Source	Select an available TDSQL-C for MySQL data source in the current project.
Database/Table	Select the corresponding database table in this data source.
Primary key	Select a field as the primary key for the data table.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Read/Write Data Type Conversion Support

Read

The supported data types and conversion relationships for TDSQL-C MySQL read are as follows (when processing TDSQL-C MySQL, the data types of the TDSQL-C MySQL data source are mapped to the data types of the data processing engine first)

Field Type	Supported	Internal mapping fields	Remarks
TINYINT	Yes	TINYINT	TINYINT(1) is mapped to BOOLEAN by default Single table tasks can map TINYINT(1) to tinyint through advanced settings
SMALLINT	Yes	SMALLINT	-
TINYINT_UNSIGNED	Yes	SMALLINT	-
TINYINT_UNSIGNED_ZEROFILL	Yes	SMALLINT	-
INT	Yes	INT	-
INTEGER	Yes	INT	-
YEAR	Yes	INT	-
MEDIUMINT	Yes	INT	-
SMALLINT_UNSIGNED	Yes	INT	-

©2013-2025 Tencent Cloud. All rights reserved.

SMALLINT_UNSIGNED_ZEROFILL	Yes	INT	–
BIGINT	Yes	LONG	–
INT_UNSIGNED	Yes	LONG	–
MEDIUMINT_UNSIGNED	Yes	LONG	–
MEDIUMINT_UNSIGNED_ZEROFILL	Yes	LONG	–
INT_UNSIGNED_ZEROFILL	Yes	LONG	–
BIGINT_UNSIGNED	Yes	DECIMAL	DECIMAL(20,0)
BIGINT_UNSIGNED_ZEROFILL	Yes	DECIMAL	DECIMAL(20,0)
SERIAL	Yes	DECIMAL	DECIMAL(20,0)
FLOAT	Yes	FLOAT	–
FLOAT_UNSIGNED	Yes	FLOAT	–
FLOAT_UNSIGNED_ZEROFILL	Yes	FLOAT	–
DOUBLE	Yes	DOUBLE	–
DOUBLE_UNSIGNED	Yes	DOUBLE	–
DOUBLE_UNSIGNED_ZEROFILL	Yes	DOUBLE	–
DOUBLE_PRECISION	Yes	DOUBLE	–
DOUBLE_PRECISION_UNSIGNED	Yes	DOUBLE	–
ZEROFILL	Yes	DOUBLE	–
REAL	Yes	DOUBLE	–
REAL_UNSIGNED	Yes	DOUBLE	–
REAL_UNSIGNED_ZEROFILL	Yes	DOUBLE	–
NUMERIC	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
NUMERIC_UNSIGNED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
NUMERIC_UNSIGNED_ZEROFILL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
DECIMAL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
DECIMAL_UNSIGNED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
DECIMAL_UNSIGNED_ZEROFILL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
FIXED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String

©2013–2025 Tencent Cloud. All rights reserved.

FIXED_UNSIGNED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
FIXED_UNSIGNED_ZEROFILL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
BOOLEAN	Yes	BOOLEAN	–
DATE	Yes	DATE	–
TIME	Yes	TIME	–
DATETIME	Yes	TIMESTAMP	–
TIMESTAMP	Yes	TIMESTAMP	–
CHAR	Yes	STRING	–
JSON	Yes	STRING	–
BIT	Yes	STRING	BIT(1) maps to BOOLEAN
VARCHAR	Yes	STRING	–
TEXT	Yes	STRING	–
BLOB	Yes	STRING	–
TINYBLOB	Yes	STRING	–
TINYTEXT	Yes	STRING	–
MEDIUMBLOB	Yes	STRING	–
MEDIUMTEXT	Yes	STRING	–
LONGBLOB	Yes	STRING	–
LONGTEXT	Yes	STRING	–
VARBINARY	Yes	STRING	–
GEOMETRY	Yes	STRING	–
POINT	Yes	STRING	–
LINESTRING	Yes	STRING	–
POLYGON	Yes	STRING	–
MULTIPOINT	Yes	STRING	–
MULTILINESTRING	Yes	STRING	–
MULTIPOLYGON	Yes	STRING	–
GEOMETRYCOLLECTION	Yes	STRING	–
ENUM	Yes	STRING	–
BINARY	Yes	BINARY	BINARY(1)
SET	No	–	–

Write

The supported data types and conversion relationships for MySQL write are as follows:

©2013–2025 Tencent Cloud. All rights reserved.

Internal Type	MySQL Type
TINYINT	TINYINT
SMALLINT	SMALLINT,TINYINT UNSIGNED
INT	INT,MEDIUMINT,SMALLINT UNSIGNED
BIGINT	BIGINT,INT UNSIGNED
DECIMAL(20, 0)	BIGINT UNSIGNED
FLOAT	FLOAT
DOUBLE	DOUBLE,DOUBLE PRECISION
DECIMAL(p, s)	NUMERIC(p, s),DECIMAL(p, s)
BOOLEAN	BOOLEAN,TINYINT(1)
DATE	DATE
TIME [(p)][WITHOUT TIMEZONE]	TIME [(p)]
TIMESTAMP [(p)][WITHOUT TIMEZONE]	DATETIME [(p)]
STRING	CHAR(n),VARCHAR(n),TEXT
BYTES	BINARY,VARBINARY,BLOB
ARRAY	–

FAQs

1. MySQL server ID conflict

Error message:

```
com.github.shyiko.mysql.binlog.network.ServerException: A slave with the same server_uuid/server_id as
this slave has connected to the master.
```

Solution: We have now optimized to randomly generate serverid. If in previous tasks the server-id was explicitly specified in MySQL advanced parameters, it is recommended to delete it, as multiple tasks may use the same data source with the same server-id, causing a conflict.

2. Binlog File Not Found Error:

Error message:

```
Caused by: org.apache.kafka.connect.errors.ConnectException: The connector is trying to read binlog
starting at GTIDs xxx and binlog file 'binlog.xxx', pos=xxx, skipping 4 events plus 1 rows, but this is no
longer available on the server. Reconfigure the connector to use a snapshot when needed.
```

Error Causes:
An error will occur if the binlog file being read by the job has already been cleaned up on the MySQL server. There are several reasons for Binlog cleanup, such as a too-short Binlog retention period, or the job processing speed not keeping up with the Binlog generation speed, exceeding the maximum retention period of MySQL Binlog files. This results in the Binlog files being cleaned up on the MySQL server, causing the Binlog positions being read to become invalid.

Solution: If the job processing speed cannot keep up with the Binlog generation speed, you can consider increasing the Binlog retention period or optimizing the job to reduce backpressure and accelerate source consumption. If there are no anomalies in the

©2013-2025 Tencent Cloud. All rights reserved.

job status, it might be that other operations on the database caused the Binlog to be cleaned up, leading to inaccessibility. You need to refer to the MySQL database-side information to determine the reason for Binlog cleanup.

3. MySQL Connection Reset

Error message:

```
EventDataDeserializationException: Failed to deserialize data of EventHeaderV4 .... Caused by:
java.net.SocketException: Connection reset.
```

Error Causes:

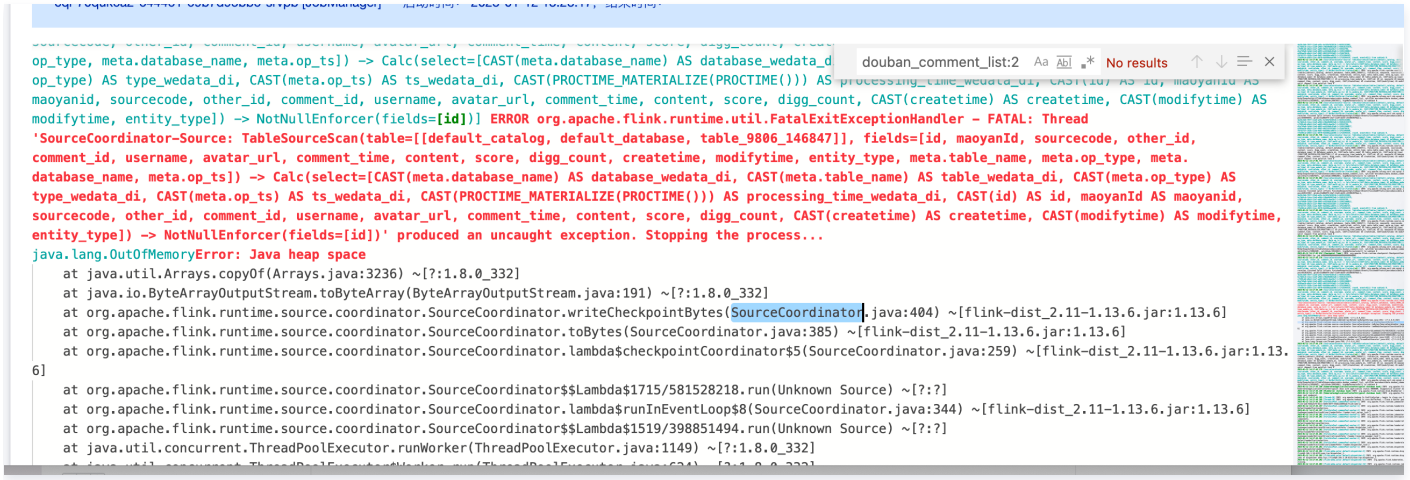
1. Network Issues.
2. The job has backpressure, causing the source to be unable to read data, and the binlog client to be idle. If the binlog connection remains idle after the timeout, the MySQL server will disconnect the idle connection.

Solution:

1. If it is a network issue, you can increase the MySQL network parameters set global slave_net_timeout = 120; (default 30s) set global thread_pool_idle_timeout = 120.
2. If the backpressure is caused by the job, you can alleviate it by adjusting the job, such as increasing parallelism, improving write speed, and increasing TaskManager memory to reduce GC.

4. Mysql2dlc task JobManager OOM

Error message:



Causes and Solutions:

1. If the user data volume is large, you can increase the JobManager CU number and use the MySQL advanced parameter scan.incremental.snapshot.chunk.size to increase the chunk size, which defaults to 8096.
2. The user's data volume is not large, but there is a large difference between the maximum and minimum values of the primary key, resulting in many chunks being divided when using the equal chunk strategy. Modify the distribution factor so that user data follows an uneven data splitting logic, split-key.even-distribution.factor.upper-bound=5.0d. The default distribution factor has been modified to 10.0d.

5. The format of the user's binlog data is incorrect, causing a parsing exception in Debezium

Error message:

```
ERROR io.debezium.connector.mysql.MySqlStreamingChangeEventSource [] - Error during binlog processing.
Last offset stored = null, binlog reader near position = mysql-bin.000044/211839464.
2023-02-20 21:37:28.480 [blc-172.17.48.3:3306] ERROR io.debezium.pipeline.ErrorHandler [] - Producer
failure
io.debezium.DebeziumException: Error processing binlog event.
```

©2013-2025 Tencent Cloud. All rights reserved.

Solution:
After modifying binlog_row_image=full, you need to restart the database.

6. Is gh-ost supported?

Yes, it does not migrate the temporary table data generated by Online DDL changes. It only migrates the original DDL data executed using gh-ost in the source database. You can use the default or configure your own regular expressions for gh-ost shadow tables and unused tables.

©2013-2025 Tencent Cloud. All rights reserved.

Oracle Data Source

Last updated: 2024-11-06 20:03:59

DataInLong provides Oracle read and write capabilities. This article introduces the pre-environment configuration for real-time data synchronization using Oracle and the current capability support status.

Supported Versions

Currently, DataInLong supports real-time reading and single table writing for both individual Oracle tables and entire databases. The following version limitations must be observed when using real-time synchronization capabilities:

Type	Version
Oracle	11, 12, 19

Use Limits

- Oracle read-only standby databases only support logical standby, not physical standby.
- Tables without primary keys might experience data duplication due to the inability to ensure exactly once. Therefore, it is best to have a primary key for real-time synchronization tasks.
- To monitor table field changes on the Oracle side, do not configure the system/sys accounts for the data source. Otherwise, all tables (including newly added tables) will require logging to be enabled for synchronization.

Database Environment Preparation

Granting database permissions

Primarily, this involves creating basic synchronization users and granting permissions to pull existing data. Users must be granted read permissions for required views and tables.
The primary function of the ANALYZE ANY permission is to read table structures and unique indexes. When a table lacks a primary key, a unique index can be used as a substitute.

```
sqlplus sys/password@host:port/SID AS SYSDBA;
CREATE USER flinkuser IDENTIFIED BY flinkpw DEFAULT TABLESPACE LOGMINER_TBS QUOTA UNLIMITED ON
LOGMINER_TBS;
-- Allows users to connect to the database
  GRANT CREATE SESSION TO flinkuser;
-- Grants users SELECT privileges on all tables in the database to read data
  GRANT SELECT ANY TABLE TO flinkuser;
--SELECT_CATALOG_ROLE is a pre-defined database role that provides permissions to view database system
tables. Additionally, during the incremental phase, the logminer session relies on this permission to
query table structures.
  GRANT SELECT_CATALOG_ROLE TO flinkuser;

--GRANT ANALYZE ANY is an authorization statement granting users the ability to analyze all tables,
indexes, partitions, and table functions owned by any user. When a table lacks a primary key, this
permission is relied upon to obtain unique indexes and use them as primary key substitutes.
  GRANT ANALYZE ANY TO flinkuser;
```

Enabling Log Archiving

This enables the generation of logical logs. Without this, Oracle cannot correctly generate logs, thereby preventing incremental data synchronization.

1. Use a DBA account to connect.

```
sqlplus sys/password@host:port/SID AS SYSDBA
```

2. Enable log archiving**Without enabling, incremental data cannot be read.**

©2013-2025 Tencent Cloud. All rights reserved.

```
alter system set db_recovery_file_dest_size = 10G;
alter system set db_recovery_file_dest = '/opt/oracle/oradata/recovery_area' scope=spfile;
shutdown immediate;
startup mount;
alter database archivelog;
alter database open;

Notes: Enabling log archiving requires a database restart and will use disk space
```

3. Check if it is enabled.

```
-- Should now "Database log mode: Archive Mode"
archive log list;
```

4. Enable supplemental log
   - If disabled, the fields not updated in the Update will all be empty.

```
-- Enable supplemental logging for a specific table:
ALTER TABLE inventory.customers ADD SUPPLEMENTAL LOG DATA (ALL) COLUMNS;
```

   - Alternatively, you can enable it globally:

```
-- Enable supplemental logging for database
ALTER DATABASE ADD SUPPLEMENTAL LOG DATA;
```

5. Create Tablespace

It is recommended to create a separate tablespace for real-time synchronization users, but existing tablespaces can also be reused.

```
 CREATE TABLESPACE logminer_tbs DATAFILE '/opt/oracle/oradata/SID/logminer_tbs.dbf' SIZE 25M REUSE
AUTOEXTEND ON MAXSIZE 2048M;
```

Grant permission to read logs through logminer

```
 --A special grant statement that gives the user SELECT permission on the V$DATABASE view, primarily used
to check whether the database's archived log is enabled (LOG_MODE) and to obtain the current database's
SCN locus.
  GRANT SELECT ON V_$DATABASE to flinkuser;

--EXECUTE_CATALOG_ROLE is a predefined role that primarily provides the ability to allow logminer to write
the dictionary into the binlog, so changes to the table structure are also recorded in the logical log.
  GRANT EXECUTE_CATALOG_ROLE TO flinkuser;

--Allows the user to read data through the LOGMINING component
  GRANT LOGMINING TO flinkuser;

--EXECUTE ON DBMS_LOGMNR permission is used to allow the user to execute programs within the DBMS_LOGMNR
package. The DBMS_LOGMNR package is used to analyze and extract changes in the database log files.
  GRANT EXECUTE ON DBMS_LOGMNR TO flinkuser;

--GRANT EXECUTE ON DBMS_LOGMNR_D is a grant statement that gives the user execute permissions on the
DBMS_LOGMNR_D package. DBMS_LOGMNR_D is an Oracle-provided package used to analyze online and archived
redo log files at the database level
  GRANT EXECUTE ON DBMS_LOGMNR_D TO flinkuser;
```

©2013-2025 Tencent Cloud. All rights reserved.

```
--The data query interface of logminer consists of a set of system-provided views. To read data from
logminer, you need to grant access permissions to the following views.
  GRANT SELECT ON V_$LOG TO flinkuser;
  GRANT SELECT ON V_$LOG_HISTORY TO flinkuser;
  GRANT SELECT ON V_$LOGMNR_LOGS TO flinkuser;
  GRANT SELECT ON V_$LOGMNR_CONTENTS TO flinkuser;
  GRANT SELECT ON V_$LOGMNR_PARAMETERS TO flinkuser;
  GRANT SELECT ON V_$LOGFILE TO flinkuser;
  GRANT SELECT ON V_$ARCHIVED_LOG TO flinkuser;
  GRANT SELECT ON V_$ARCHIVE_DEST_STATUS TO flinkuser;

  exit;
```

Database-level source configuration

Data Source Settings



Parameter	Description
Data Source	Select the Oracle data source to synchronize
Source Table	According to business requirements, choose "All Database Tables", "Specified Table", or "Specified Database" • All Database Tables: Monitor all databases under the data source. Newly added databases and tables during task execution will be synchronized to the target by default • Specified Table: Synchronize only the specified table • Designated library: Monitor specific libraries and schemas, and synchronize all or rule-compliant tables under the schema
Read mode	• Full + Incremental: Data synchronization is divided into full and incremental synchronization phases. After the full phase is completed, the task enters the incremental phase. The full phase will synchronize historical data in the database, and the incremental phase will start synchronizing from the binlog CDC location after the task starts. • Incremental: Sync data only from the binlog CDC locus after the task starts.
Consistency Semantics	It only represents the consistency semantics of the reading end. Supports At-least-once and Exactly-once. • At-least-once: Data may be read repeatedly, and relies on the target end to support deduplication to ensure data consistency. Suitable for scenarios with a large volume of data in the full stage and non-numeric primary keys, requiring high synchronization performance.

©2013-2025 Tencent Cloud. All rights reserved.

	• Exactly-once: Data is read strictly once, with some performance loss, and does not support tables without primary keys and unique index columns. Suitable for general scenarios where the source table has numeric primary keys or unique index columns. The two modes are incompatible in the current version. If the mode is changed after task submission, stateful restart is not supported.
Advanced Settings (Optional)	You can configure parameters based on business requirements

Note
To monitor table field changes on the Oracle side, do not configure the system/sys accounts for the data source. Otherwise, all tables (including newly added tables) will require logging to be enabled for synchronization.

Start command:

```
ALTER TABLE SCHEMA_NAME.TBALE_NAME ADD SUPPLEMENTAL LOG DATA (ALL) COLUMNS
```

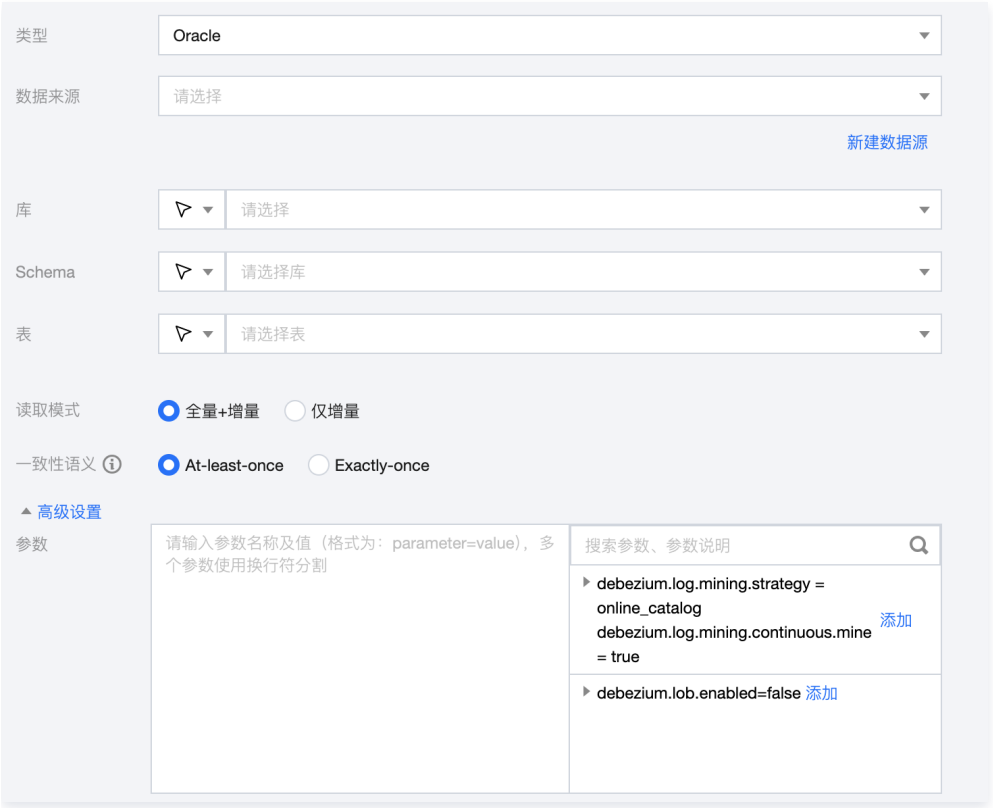
Supported Target Data Sources

Currently supports real-time whole database synchronization of Oracle data to the following types of target ends:

- DLC Whole Database Target Configuration Details
- Iceberg Whole Database Target Configuration Details
- TCHouse-P Whole Database Target Configuration Details
- Hive Whole Database Target Configuration Details
- StarRocks Whole Database Target Configuration Details
- Kafka Database Synchronization Configuration Details

Single Table Read Node Configuration

1. On the DataInLong page, click Real-time synchronization in the left-hand navigation bar.
2. On the Real-time synchronization page, select Single Table Synchronization to create a new configuration (you can choose between Form mode and Canvas mode) and enter the configuration page.



©2013-2025 Tencent Cloud. All rights reserved.

Parameter	Description
Data Source	Select the data source where the table to be synced is located.
Database	Support selection or manual input of the library name to be read. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Support selection or manual input of the available Schemas under this data source.
Table	Support selecting or manually entering the Table Name to be read.
Read mode	• Full + Incremental: Data synchronization is divided into full and incremental synchronization phases. After the full phase is completed, the task enters the incremental phase. The full phase will synchronize historical data in the database, and the incremental phase will start synchronizing from the binlog CDC location after the task starts. • Increment only: Only sync data starting from the binlog cdc point after the task starts.
Consistency Semantics	It only represents the consistency semantics of the reading end. Supports At-least-once and Exactly-once. • At-least-once: Data may be read repeatedly, and relies on the target end to support deduplication to ensure data consistency. Suitable for scenarios with a large volume of data in the full stage and non-numeric primary keys, requiring high synchronization performance. • Exactly-once: Data is read strictly once, with some performance loss, and does not support tables without primary keys and unique index columns. Suitable for general scenarios where the source table has numeric primary keys or unique index columns. The two modes are incompatible in the current version. If the mode is changed after task submission, stateful restart is not supported.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Single Table Write Node Configuration

1. On the DataInLong page, click Real-time synchronization in the left-hand navigation bar.
2. On the Real-time synchronization page, select Single Table Synchronization to create a new configuration (you can choose between Form mode and Canvas mode) and enter the configuration page.

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Source	The Oracle data source to be written to is required.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Supports selection or manual input of the Oracle data pattern to be written.
Table	Supports selecting or manually entering the name of the table to be written to. When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Primary key	Select a field as the primary key for the write table.
Advanced Settings (optional)	You can configure parameters based on business requirements.

Log Collection Write Node

©2013–2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Source	Select an available Oracle data source in the current project.
Database/Table	Select the corresponding database table in this data source.
Primary key	Select a field as the primary key for the data table.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Data type conversion is supported

Read

The supported data types for reading in Oracle and their corresponding conversion relationships are as follows (When processing Oracle, it first maps the data types of the Oracle data source to the data types of the data processing engine):

Oracle Type	Internal Type
NUMBER(p, s <= 0), p − s < 3	TINYINT
NUMBER(p, s <= 0), p − s < 5	SMALLINT
NUMBER(p, s <= 0), p − s < 10	INT
NUMBER(p, s <= 0), p − s < 19	BIGINT
NUMBER(p, s <= 0), 19 <= p − s <= 38	DECIMAL(p − s, 0)
NUMBER(p, s > 0)	DECIMAL(p, s)
NUMBER(p, s <= 0), p − s > 38	STRING
FLOAT,BINARY_FLOAT	FLOAT
DOUBLE PRECISION,BINARY_DOUBLE	DOUBLE
NUMBER(1)	BOOLEAN
DATE,TIMESTAMP [(p)]	TIMESTAMP [(p)] [WITHOUT TIMEZONE]

©2013–2025 Tencent Cloud. All rights reserved.

TIMESTAMP [(p)] WITH TIME ZONE	TIMESTAMP [(p)] WITH TIME ZONE
TIMESTAMP [(p)] WITH LOCAL TIME ZONE	TIMESTAMP_LTZ [(p)]
CHAR(n),NCHAR(n),NVARCHAR2(n),VARCHAR(n),VARCHAR2(n),CLOB,NCLOB,XMLType	STRING
BLOB,ROWID	BYTES
INTERVAL DAY TO SECOND,INTERVAL YEAR TO MONTH	BIGINT

Write

The supported data types for writing in Oracle and their corresponding conversion relationships are as follows:

Internal Type	Oracle Type
FLOAT	BINARY_FLOAT
DOUBLE	BINARY_DOUBLE
DECIMAL(p, s)	SMALLINT,FLOAT(s),DOUBLE PRECISION,REAL,NUMBER(p, s)
DATE	DATE
DECIMAL(20, 0)	–
FLOAT	REAL,FLOAT4
DOUBLE	FLOAT8,DOUBLE PRECISION
DECIMAL(p, s)	NUMERIC(p, s),DECIMAL(p, s)
BOOLEAN	BOOLEAN
DATE	DATE
TIMESTAMP [(p)][WITHOUT TIMEZONE]	TIMESTAMP [(p)]WITHOUT TIMEZONE
STRING	CHAR(n),VARCHAR(n),CLOB(n)
BYTES	RAW(s),BLOB
ARRAY	–

©2013–2025 Tencent Cloud. All rights reserved.

PostgreSQL Data Source

Last updated: 2024-09-06 14:48:06

DataInLong provides reading and writing capabilities for PostgreSQL. This document introduces the pre-environment configuration and the current capability support for using PostgreSQL for real-time data synchronization.

Supported Versions

Currently, DataInLong supports real-time reading and writing for PostgreSQL single tables and database-level. To use real-time synchronization capabilities, the following version limitations must be observed:

Type	Version
PostgreSQL	Version 9.6 or above

Use Limits

- In PostgreSQL, one data source can only correspond to one database. Therefore, a real-time database-level task can only synchronize tables under one database and does not support selecting tables across databases.
- Tables without primary keys might experience data duplication due to the inability to ensure exactly once. Therefore, it is best to have a primary key for real-time synchronization tasks.
- Schema changes on the PostgreSQL source rely on data transmission. Schema changes can only be detected when there is data sending after the changes.
- Column renaming on the PostgreSQL source will be recognized as column deletion and new column creation
- Deleting tables and clearing tables on the PostgreSQL source cannot be synchronized to the target end; such changes are ignored by default.

Database Environment Preparation

1. Modify postgres.conf to enable WAL:

```
wal_level = logical
```

2. For tables that need to be synchronized, enable full replication:

```
ALTER TABLE <table> REPLICA IDENTITY FULL
```

3. Required permissions: (replace items starting with my_ with the actual database, schema, table, username)

```
GRANT CONNECT ON DATABASE my_db TO my_user;
GRANT USAGE ON SCHEMA my_schema TO my_user;
GRANT SELECT ON ALL TABLES IN SCHEMA my_schema TO my_user;
ALTER USER my_user REPLICATION;
GRANT CREATE ON DATABASE my_db TO my_user;
```

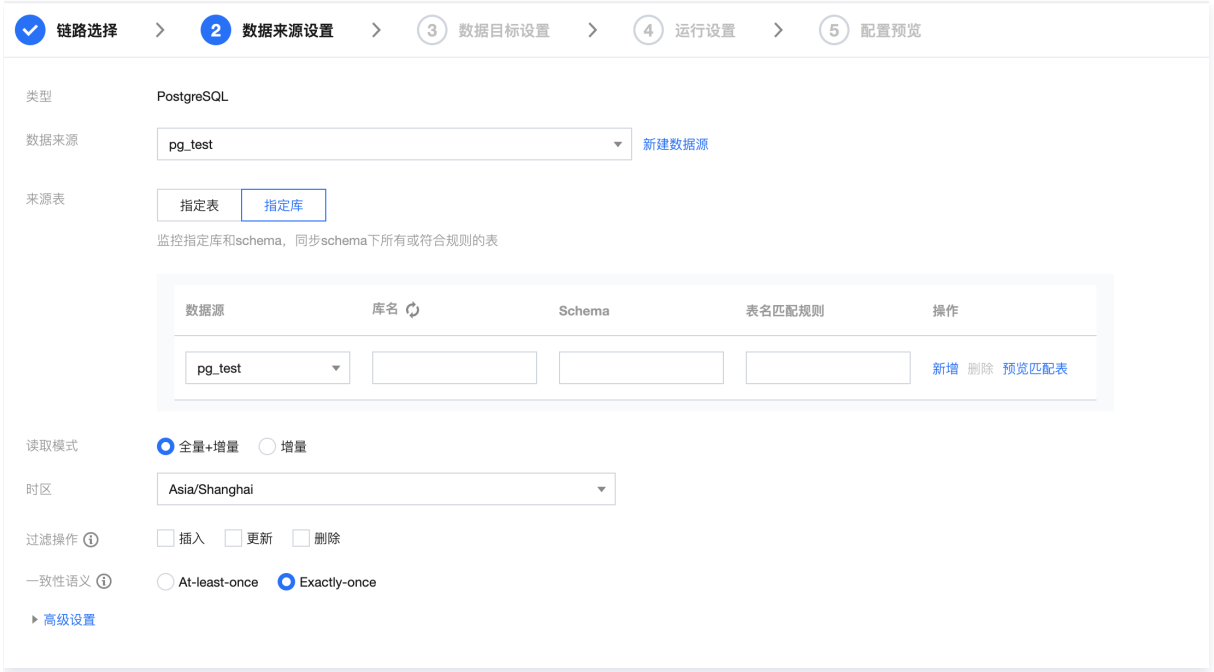
4. Pre-create a pg_publication named dbz_publication to avoid using the following permissions:

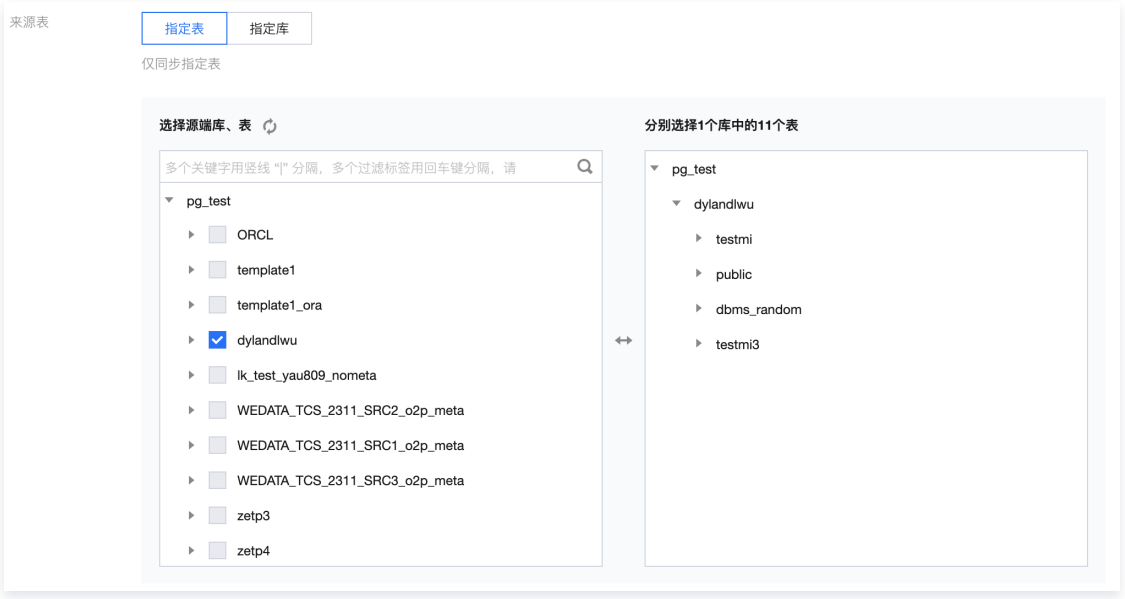
```
GRANT CREATE ON DATABASE my_db TO my_user;
```

Database-level source configuration

Data Source Settings

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Source	Select the PostgreSQL data source to be synchronized.
Source Table	PostgreSQL supports only specified tables and specified databases, not all databases and tables: • Specified Table: Under this option, you need to specify the exact table name. Once set, the task only synchronizes the specified table. Only tables under one database can be selected, and cross-database selection is not supported. To add new tables for synchronization, you need to pause and restart the task.  • Specified Database: Under this option, you need to specify the exact database name and table name regex expression. The database name must be the same; multiple databases are not supported. Once set, new tables that fit the table name expression will be synchronized to the target end by default during task runtime.

©2013–2025 Tencent Cloud. All rights reserved.

	
Read mode	• Full + Incremental: Data synchronization is divided into full and incremental synchronization phases. After the full phase is completed, the task enters the incremental phase. The full phase will synchronize historical data in the database, and the incremental phase will start synchronizing from the binlog CDC location after the task starts. • Incremental: Only synchronize data from the binlog CDC location after the task starts.
Time zone	Set the log time timezone, default is Shanghai.
Filter operations	Support three operations: Insert, Update, and Delete. After setting, data of the specified operation types will not be synchronized.
Consistency Semantics	It only represents the consistency semantics of the reading end. Supports At-least-once and Exactly-once. • At-least-once: Data may be read repeatedly, and relies on the target end to support deduplication to ensure data consistency. Suitable for scenarios with a large volume of data in the full stage and non-numeric primary keys, requiring high synchronization performance. • Exactly-once: Data is read strictly once, with some performance loss, and does not support tables without primary keys and unique index columns. Suitable for general scenarios where the source table has numeric primary keys or unique index columns. The two modes are incompatible in the current version. If the mode is changed after task submission, stateful restart is not supported.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Supported Target Data Sources

Currently, it supports real-time database-level synchronization of PostgreSQL data to the following types of target ends:

- Doris Whole Database Target Configuration Details
- DLC Whole Database Target Configuration Details
- Iceberg Whole Database Target Configuration Details
- TCHouse-P Whole Database Target Configuration Details
- Hive Whole Database Target Configuration Details
- StarRocks Whole Database Target Configuration Details
- Kafka Database Synchronization Configuration Details

Whole Database Write Configuration

©2013-2025 Tencent Cloud. All rights reserved.



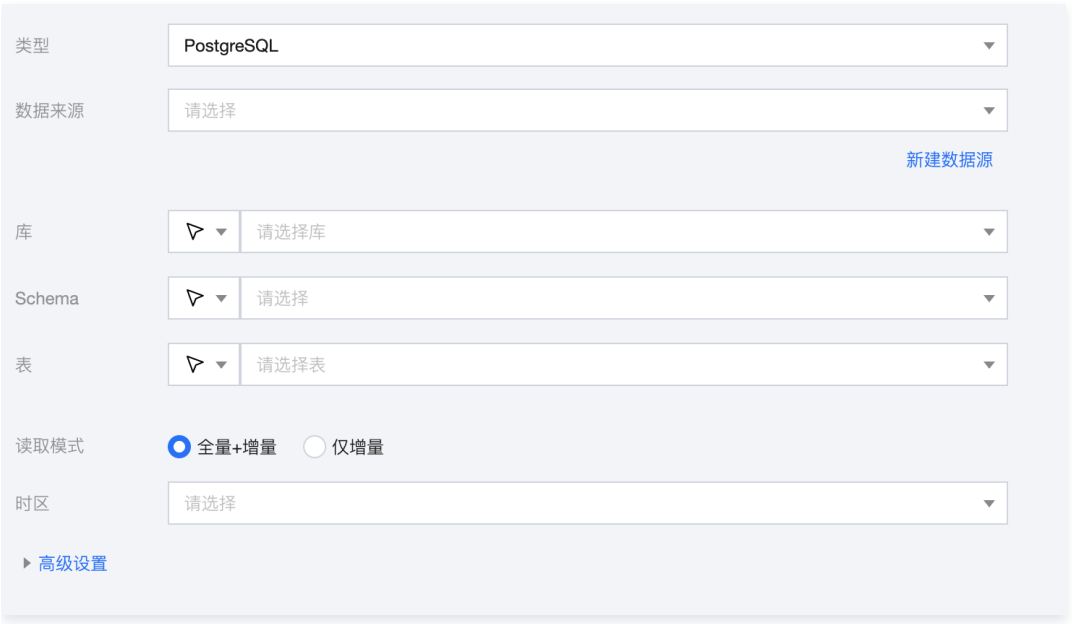
Parameter	Description
Data Flow Direction	Choose the PostgreSQL target data source to synchronize.
Write mode	• Upsert: Update mode writes to the target table. In this mode, the target table must have a primary key set. The task will use the primary key as the unique key for record updates by default. If the table does not have a primary key, data will be appended. • Append: Append mode writes to the data table.
Database/Schema/Table Matching Strategy	The default is to match the same name as the source database/table, or you can configure the rules manually. Hover over "View Built-in Parameters" to see the matching rules. The PostgreSQL target end currently does not support automatic database/schema/table creation. It only supports synchronization to existing databases, schemas, and tables. Make sure the target table exists before the task runs.

Note
Currently, only database-level synchronization of MySQL data to PostgreSQL is supported.

Single Table Read Node Configuration

1. On the DataInLong page, click Real-time synchronization in the left-hand navigation bar.
2. On the Real-time synchronization page, select Single Table Synchronization to create a new configuration (you can choose between Form mode and Canvas mode) and enter the configuration page.

©2013-2025 Tencent Cloud. All rights reserved.

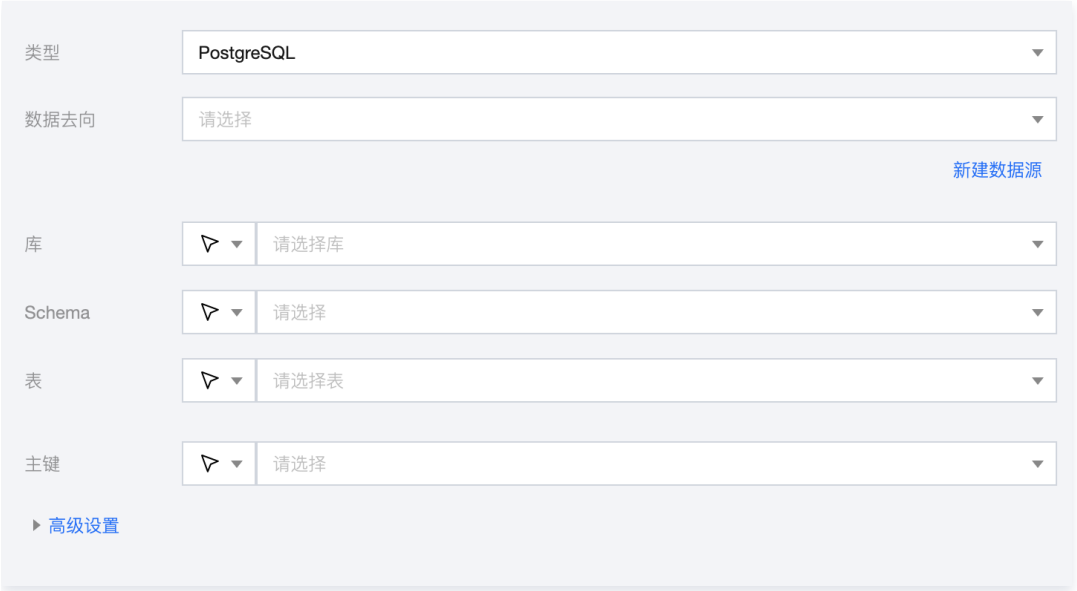


Parameter	Description
Data Source	Select the data source where the table to be synced is located.
Database	Support selection or manual input of the library name to be read. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Support selecting or manually entering the available schema under the data source.
Table	Support selecting or manually entering the Table Name to be read.
Read mode	• Full + Incremental: Data synchronization is divided into full and incremental synchronization phases. After the full phase is completed, the task enters the incremental phase. The full phase will synchronize historical data in the database, and the incremental phase will start synchronizing from the binlog CDC location after the task starts. • Increment only: Only sync data starting from the binlog cdc point after the task starts.
Advanced Settings (optional)	You can configure parameters based on business requirements.

Single Table Write Node Configuration

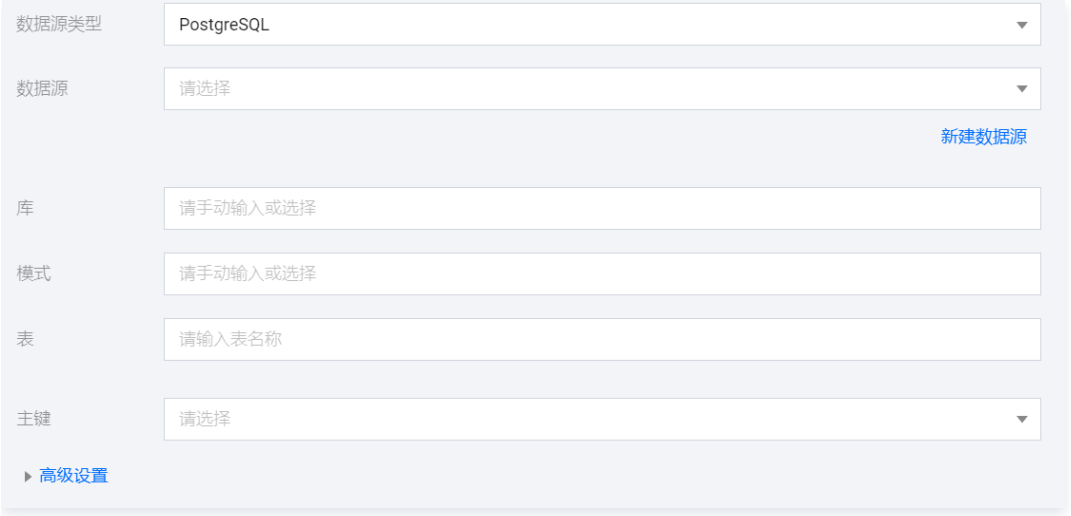
1. On the DataInLong page, click Real-time synchronization in the left-hand navigation bar.
2. On the Real-time synchronization page, select Single Table Synchronization to create a new configuration (you can choose between Form mode and Canvas mode) and enter the configuration page.

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Flow Direction	Select the data source containing the table to be written.
Database	Supports selecting or manually entering the name of the database to be written to. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Support selecting or manually entering the available schema under the data source to be written.
Table	Support selecting or manually entering the table name to be written.
Primary key	Select a field as the primary key for the write table.
Advanced Settings (optional)	You can configure parameters based on business requirements.

Log Collection Write Node Configuration



©2013–2025 Tencent Cloud. All rights reserved.

Parameter	Description
Data Source	Select the available PostgreSQL data source in the current project.
Database/Table	Choose the corresponding database table from the PostgreSQL data source.
Mode	Select the schema in the PostgreSQL data source.
Primary key	Select a field as the primary key of the data table
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Data type conversion is supported

Read

The supported data types and their conversion mappings for PostgreSQL read are as follows:

PostgreSQL Type	Internal Type
–	TINYINT
SMALLINT,INT2,SMALLSERIAL,SERIAL2	SMALLINT
INTEGER,SERIAL	INT
BIGINT,BGSERIAL	BIGINT
–	DECIMAL(20, 0)
REAL,FLOAT4	FLOAT
FLOAT8,DOUBLE PRECISION	DOUBLE
NUMERIC(p, s),DECIMAL(p, s)	DECIMAL(p, s)
BOOLEAN	BOOLEAN
DATE	DATE
TIME [(p)][WITHOUT TIMEZONE]	TIME [(p)][WITHOUT TIMEZONE]
TIMESTAMP [(p)]WITHOUT TIMEZONE	TIMESTAMP [(p)][WITHOUT TIMEZONE]
CHAR(n),CHARACTER(n),VARCHAR(n),CHARACTER ,VARYING(n),TEXT	STRING
BYTEA	BYTES

Write

The supported data types and their conversion mappings for PostgreSQL write are as follows:

Internal Type	PostgreSQL Type
TINYINT	–
SMALLINT	SMALLINT,INT2,SMALLSERIAL,SERIAL2
INT	INTEGER,SERIAL
BIGINT	BIGINT,BIGSERIAL
DECIMAL(20, 0)	–
FLOAT	REAL,FLOAT4

©2013–2025 Tencent Cloud. All rights reserved.

DOUBLE	FLOAT8,DOUBLE PRECISION
DECIMAL(p, s)	NUMERIC(p, s),DECIMAL(p, s)
BOOLEAN	BOOLEAN
DATE	DATE
TIME [(p)][WITHOUT TIMEZONE]	TIME [(p)][WITHOUT TIMEZONE]
TIMESTAMP [(p)][WITHOUT TIMEZONE]	TIMESTAMP [(p)]WITHOUT TIMEZONE
STRING	CHAR(n),CHARACTER(n),VARCHAR(n),CHARACTER VARYING(n),TEXT
BYTES	BYTEA

©2013–2025 Tencent Cloud. All rights reserved.

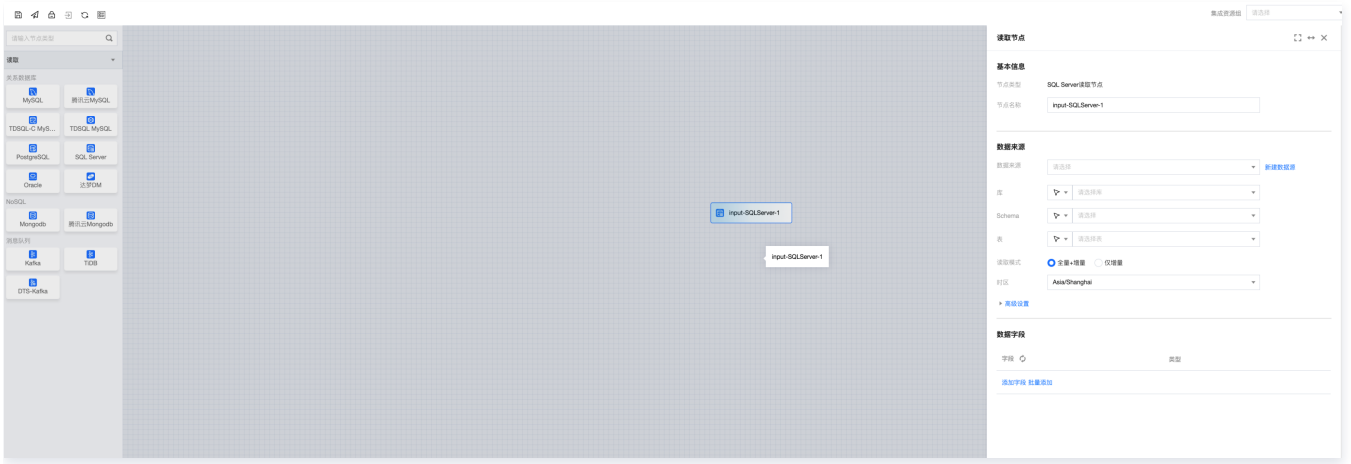
SQL Server Data Source

Last updated：2024-08-24 13:16:04

SQL Server Single Table Read Node Configuration

SQL Server Read Configuration Parameter Description

1. In the DataInLong page, click **Real-time Synchronization** in the left sidebar.
2. On the Real-time Synchronization page, select **Single Table Synchronization** at the top, create a new one (choose Form or Canvas mode), and enter the configuration page.
3. Click on the left **Read**, then click **SQL Server Node** and configure the node information.



4. You can refer to the table below for parameter configuration.

Parameter	Description
Node name	Enter the SQL Server node name.
Data Source	Select the available SQL Server data source in the project
Database	Support selection or manual input of the library name to be read. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
schema	Support selecting or manually entering the schema name to be read.
Table	Support selecting or manually entering the Table Name to be read.
Read mode	• Full + Incremental: Data synchronization is divided into full and incremental synchronization phases. After the full phase is completed, the task enters the incremental phase. The full phase will synchronize historical data in the database, and the incremental phase will start synchronizing from the binlog CDC location after the task starts. • Increment only: Only sync data starting from the binlog cdc point after the task starts.
Time zone	Set the log time timezone, default is Shanghai.
Advanced Settings	You can configure parameters based on business requirements.

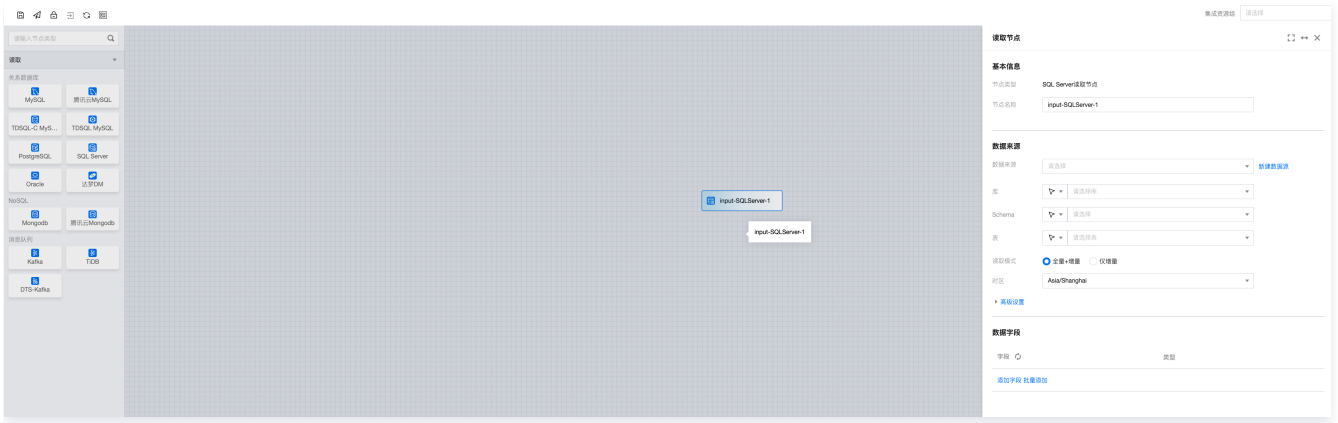
5. Preview data fields, click **Save**.

©2013-2025 Tencent Cloud. All rights reserved.

SQL Server Single Table Write Node Configuration

Configure the SQL Server node

1. In the DataInLong page, click **Real-time Synchronization** in the left sidebar.
2. On the Real-time Synchronization page, select **Single Table Synchronization** at the top, create a new one (choose Form or Canvas mode), and enter the configuration page.
3. Click on the left **Write**, then click **SQL Server Node** and configure the node information.

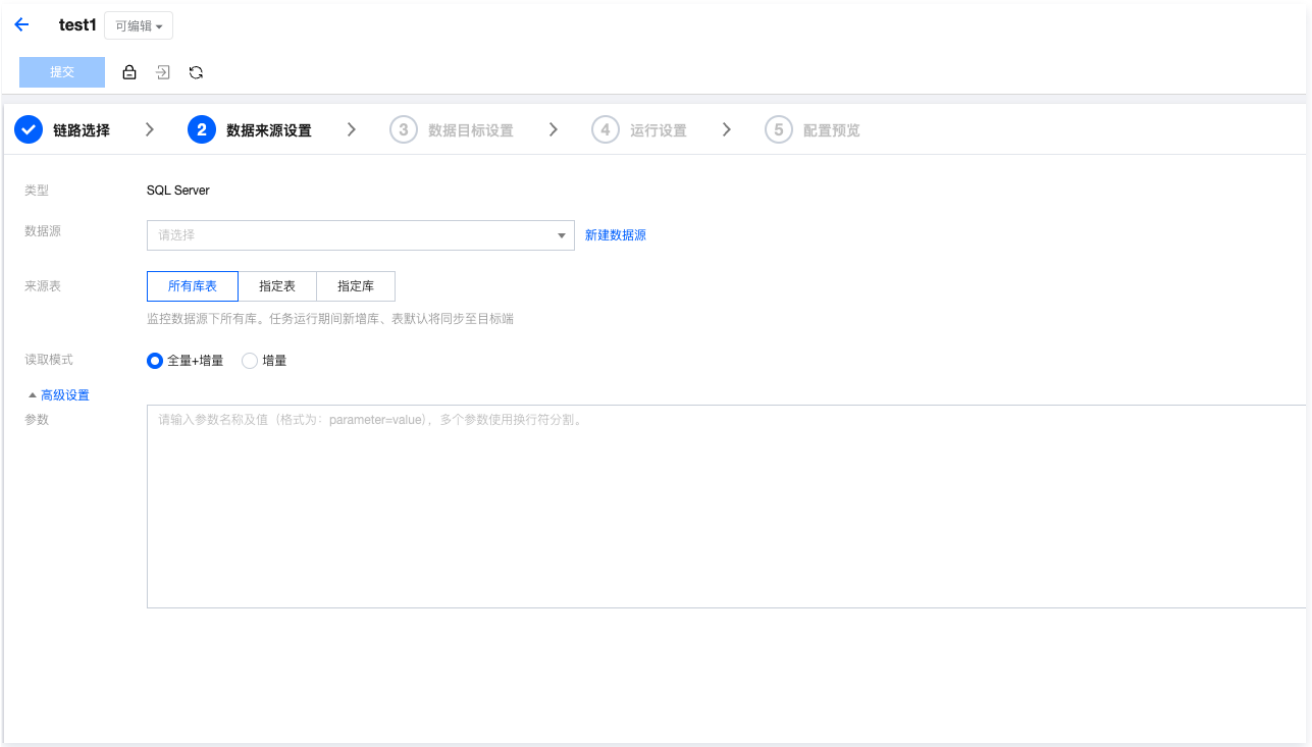


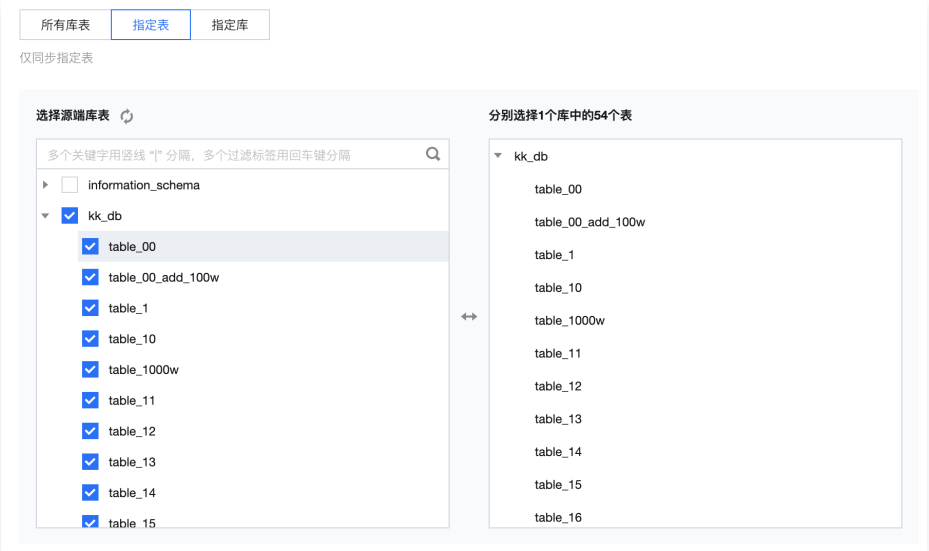
4. You can refer to the table below for parameter configuration.

Parameter	Description
Data Flow Direction	Select the available SQL Server data source in the project
Database	Supports selection, or manual input of the library name to be written to • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
schema	Support selecting or manually entering the schema name to be written.
Table	Supports selecting or manually entering the name of the table to be written to. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Primary key	Select a field as the primary key for the write table.
Advanced Settings (optional)	You can configure parameters based on business requirements.

SQL Server Real-time Whole Database Source Configuration

©2013-2025 Tencent Cloud. All rights reserved.

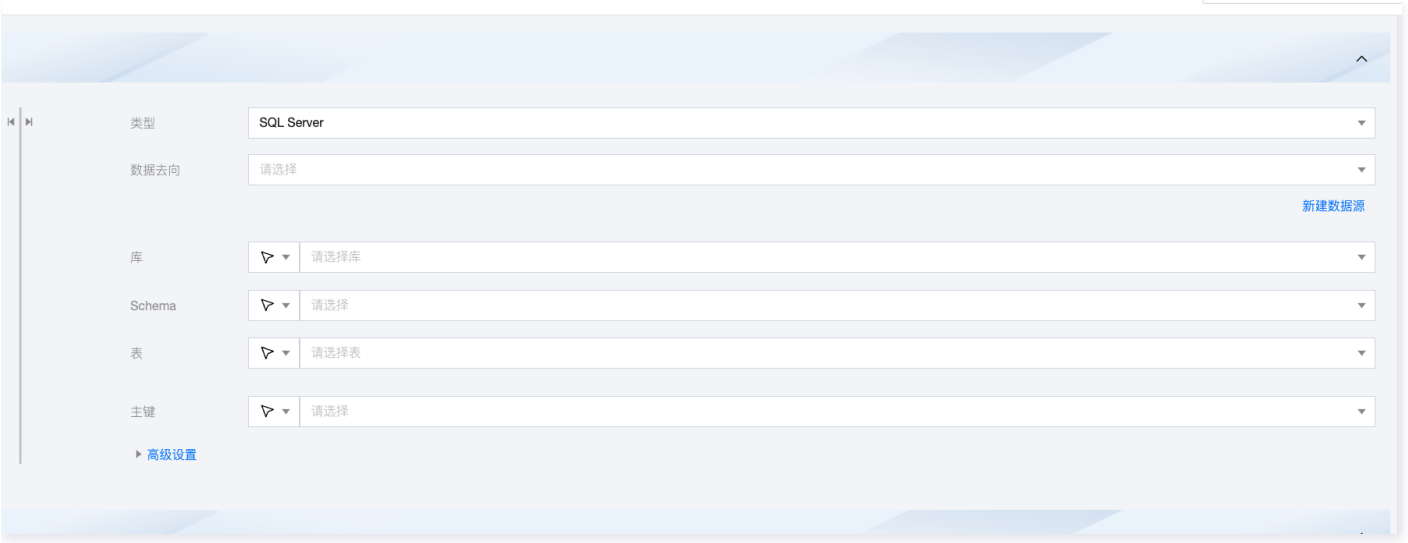


Parameters	Description
Data Source	Select the SQL Server data source for whole database synchronization.
Source Table	• All databases and tables: Monitor all databases under the data source. During the task execution, new databases and tables will be synchronized to the target end by default. • Specified Table: Under this option, you need to specify the specific table name. After setting, the task will only synchronize the specified table; if a new table needs to be synchronized, the task must be stopped and restarted.  • Specify Database: Under this option, you need to specify the specific database name with a table name regular expression. After setting, new tables that match the table name expression will be synchronized to the target end by default during task execution.

©2013–2025 Tencent Cloud. All rights reserved.

	
Read mode	• Full + Incremental: Data synchronization is divided into full and incremental synchronization phases. After the full phase is completed, the task enters the incremental phase. The full phase will synchronize historical data in the database, and the incremental phase will start synchronizing from the binlog CDC location after the task starts. • Incremental: Only synchronize data from the binlog CDC location after the task starts.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

SQL Server Log Collection Write Node Configuration



Parameter	Description
Data Flow Direction	Select the available SQL Server data source in the current project.
Database	Supports selection, or manual input of the library name to be written to • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
schema	Support selecting or manually entering the schema name to be read.
Table	Supports selecting or manually entering the name of the table to be written to. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Primary key	Select a field as the primary key of the data table

©2013–2025 Tencent Cloud. All rights reserved.

Advanced Settings (Optional)	You can configure parameters based on business requirements.

Data type conversion is supported

Read

The supported data types for SQL Server reading and their corresponding conversion relationships are as follows (when processing SQL Server, the data types of the SQL Server data source will be mapped to the data processing engine's data types first):

SQLServer Type	Internal Type
char(n)	CHAR(n)
varchar(n),nvarchar(n),nchar(n)	VARCHAR(n)
text,ntext,xml	STRING
decimal(p, s),money,smallmoney	DECIMAL(p, s)
numeric	NUMERIC
REAL,FLOAT	FLOAT
bit	BOOLEAN
int	INT
tinyint	TINYINT
smallint	SMALLINT
time (n)	TIME (n)
bigint	BIGINT
date	DATE
datetime2,datetime,smalldatetime	TIMESTAMP(n)
datetimeoffset	TIMESTAMP_LTZ(3)

Write

The supported data types for SQL Server writing and their corresponding conversion relationships are as follows:

Internal Type	SQLServer Type
CHAR(n)	char(n)
VARCHAR(n)	varchar(n),nvarchar(n),nchar(n)
STRING	text,ntext,xml
BIGINT	BIGINT,BIGSERIAL
DECIMAL(p, s)	decimal(p, s),money,smallmoney
NUMERIC	numeric
FLOAT	float,real
BOOLEAN	bit
INT	int

©2013–2025 Tencent Cloud. All rights reserved.

TINYINT	tinyint
SMALLINT	smallint
BIGINT	bigint
TIME(n)	time(n)
TIMESTAMP(n)	datetime2,datetime,smalldatetime
TIMESTAMP_LTZ(3)	datetimeoffset

©2013–2025 Tencent Cloud. All rights reserved.

Kafka Data Source

Last updated: 2024-11-25 15:57:04

DataInLong provides Kafka's read and write capabilities. This article introduces the pre-environment configuration for real-time data synchronization using Kafka and the current capability support.

Supported Versions

Currently, DataInLong supports single-table and whole-database real-time reads and writes with Kafka. To use the real-time synchronization capability, the following version limitations must be adhered to:

Node	Version
Kafka	2.4.1,2.7.1,2.8.1,2.8.2

Use Limits

- The Kafka whole-database source end must support canal-json, debezium, ogg-json serialization formats.
- One whole-database task supports multiple Topics. You need to switch to manual input mode and separate by commas.
- The canal-json/debezium format needs to include the mysqltype, sqltype, and primarykey fields.
- When Kafka is the source end, it only supports synchronizing add column, delete column, and new table events to the target end.

Kafka Environment Preparation

The process of establishing a connection between Kafka client and server-side is as follows:

1. The client uses the bootstrap.servers address you specified to connect to the Kafka server. The Kafka server returns the metadata of the brokers in the cluster to the client based on the configuration, including the connection addresses of the brokers.
2. The client uses the connection addresses returned by the brokers in the first step to connect to each broker for reading or writing:

We need to be cautious when the bootstrap.servers address is reachable, but network connectivity issues still occur. You can refer to the following methods for troubleshooting.

- Check if the broker connection address returned by the Kafka server has connectivity issues.
- Check if the addresses of listeners and advertised.listeners in the Kafka broker configuration file server.properties can connect to the integrated resource group network.

3. The Kafka target end supports automatic Topic creation and specifying Partitioning Strategy. If you need to use the auto-create Topic capability, please set it up in advance on the Kafka server side:

```
auto.create.topics.enable=true
```

- After enabling the automatic topic creation feature, the target topic must comply with CKafka/Kafka Topic Naming Rules to prevent creation failure during task execution.
- When Kafka enables automatic topic creation, please configure the number of partitions appropriately to avoid performance issues.

Whole Database Read Configuration

Data Source Settings

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Source	Select the Kafka data source to be synchronized.
Source Topic	Select or enter the Topic name for task scheduling consumption.
Serialization Format	Set the original message format in Kafka. Currently, canal-json, debezium, and ogg-json parsing are supported. **Note** The set format must be consistent with the actual message format.
Read Position	Set Kafka data read position: • From the earliest: earlist. • From the latest: latest. • From a specified time: Set the specific task start time point.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Supported Target Data Sources

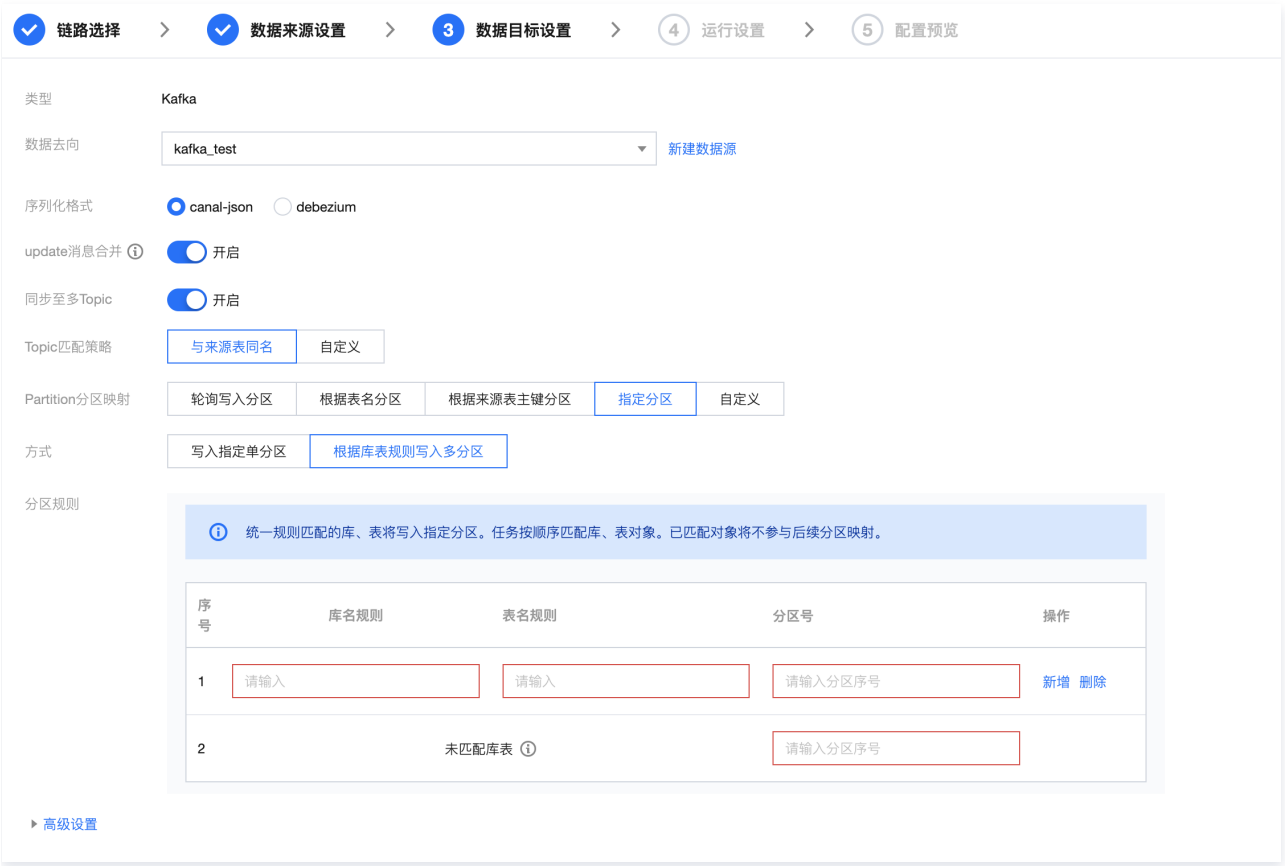
Currently, Kafka data can be synchronized in real-time to the following types of target ends:

- DLC Whole Database Target Configuration Details
- Iceberg Whole Database Target Configuration Details
- Doris Whole Database Target Configuration Details

Whole Database Write Configuration

Data Target Settings

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Source	Select the target data source for synchronization
Serialization Format	Supports canal-json and debezium formats
Update Message Merge	• Switch Off: A single Update change of a record on the source end corresponds to two Kafka records, representing the data before and after the change. • Switch On: A single Update change of a record on the source end corresponds to one Kafka record, containing both the data before and after the change.
Synchronization to Multiple Topics	• Default On: Under this option, many-to-many mapping between source data and target topics can be achieved. During task execution, the corresponding topic names are matched according to the strategy. If the topic does not exist, the system will automatically create the topic based on topic name matching rules. • Disabled: Manually input or select the target topic name. All subsequent data will be written to the specified topic.
Topic Matching Strategy	• Same Name as Source Table: By default, a topic with the same name as the source table is used. • Custom Definition: Match topics according to custom definition strategy rules.
Sharding rules	Configure topic partition mapping (polling partition writing, partition by table name, partition by source table primary key, designated partition, self Definition): • Polling Write to Partition: Polling (Round Robin) upstream data writes to each partition. • Table Name Partitioning: Hash mapping based on table name in upstream data writes to each partition. • Primary Key Partition: Hash mapping based on primary key data content in upstream data writes to each partition. • Designated Partition: ○ Write to specified single partition: Enter the partition number; all messages are written only to a fixed partition.

©2013-2025 Tencent Cloud. All rights reserved.

	○ Write to multiple partitions based on table rules: Input database and table regular expressions for object matching. Objects that match the rules are written to specified partitions. Rules are executed sequentially, and matched tables are not involved in subsequent rule matching. • Self Definition: Supports using "built-in parameters" to concatenate partitioning rules. Once set, messages will be hash partitioned based on the values corresponding to the partition rules.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Supported Data Sources

Currently supports real-time whole database synchronization of data from the following source types to Kafka targets:

- MySQL Whole Database Source Configuration Details
- TDSQL-C MySQL Whole Database Source Configuration Details
- Oracle Whole Database Source Configuration Details
- PostgreSQL Whole Database Source Configuration Details
- MongoDB Whole Database Source Configuration Details

Appendix: Canal-json/Debezium data format samples

- **Canal-json**


```
{
  "data": [
    {
      "id": "2",
      "name": "scooter33",
      "description": "Big 2-wheel scooter233",
      "weight": "5.11"
    }
  ],
  "database": "pacino99",
  "es": 1589373560000,
  "id": 9,
  "isDdl": false,
  "mysqlType": {
    "id": "INTEGER",
    "name": "VARCHAR(255)",
    "description": "VARCHAR(512)",
    "weight": "FLOAT"
  },
  "old": [
    {
      "weight": "5.12"
    }
  ],
  "pkNames": [
    "id"
  ],
  "sql": "",
  "sqlType": {
    "id": 4,
    "name": 12,
    "description": 12,
    "weight": 7
  },
  "table": "products999",
  "ts": 1589373560798,
  "type": "UPDATE"
```


©2013-2025 Tencent Cloud. All rights reserved.

```
}
```

- **Debezium**

```
{
  "schema": {
    "type": "struct",
    "fields": [
      {
        "type": "struct",
        "fields": [
          {
            "type": "int32",
            "optional": false,
            "field": "id"
          },
          {
            "type": "string",
            "optional": false,
            "field": "first_name"
          },
          {
            "type": "string",
            "optional": false,
            "field": "last_name"
          },
          {
            "type": "string",
            "optional": false,
            "field": "email"
          }
        ],
        "optional": true,
        "name": "mysql-server-1.inventory2.customers2.Value",
        "field": "before"
      },
      {
        "type": "struct",
        "fields": [
          {
            "type": "int32",
            "optional": false,
            "field": "id"
          },
          {
            "type": "string",
            "optional": false,
            "field": "first_name"
          },
          {
            "type": "string",
            "optional": false,
            "field": "last_name"
          },
          {
            "type": "string",
            "optional": false,
            "field": "email"
          }
        ],
```

©2013–2025 Tencent Cloud. All rights reserved.

```
      "optional": true,
      "name": "mysql-server-1.inventory2.customers2.Value",
      "field": "after"
    },
    {
      "type": "struct",
      "fields": [
        {
          "type": "string",
          "optional": false,
          "field": "version"
        },
        {
          "type": "string",
          "optional": false,
          "field": "connector"
        },
        {
          "type": "string",
          "optional": false,
          "field": "name"
        },
        {
          "type": "int64",
          "optional": false,
          "field": "ts_ms"
        },
        {
          "type": "boolean",
          "optional": true,
          "default": false,
          "field": "snapshot"
        },
        {
          "type": "string",
          "optional": false,
          "field": "db"
        },
        {
          "type": "string",
          "optional": true,
          "field": "table"
        },
        {
          "type": "int64",
          "optional": false,
          "field": "server_id"
        },
        {
          "type": "string",
          "optional": true,
          "field": "gtid"
        },
        {
          "type": "string",
          "optional": false,
          "field": "file"
        },
        {
          "type": "int64",
          "optional": false,
```


©2013–2025 Tencent Cloud. All rights reserved.

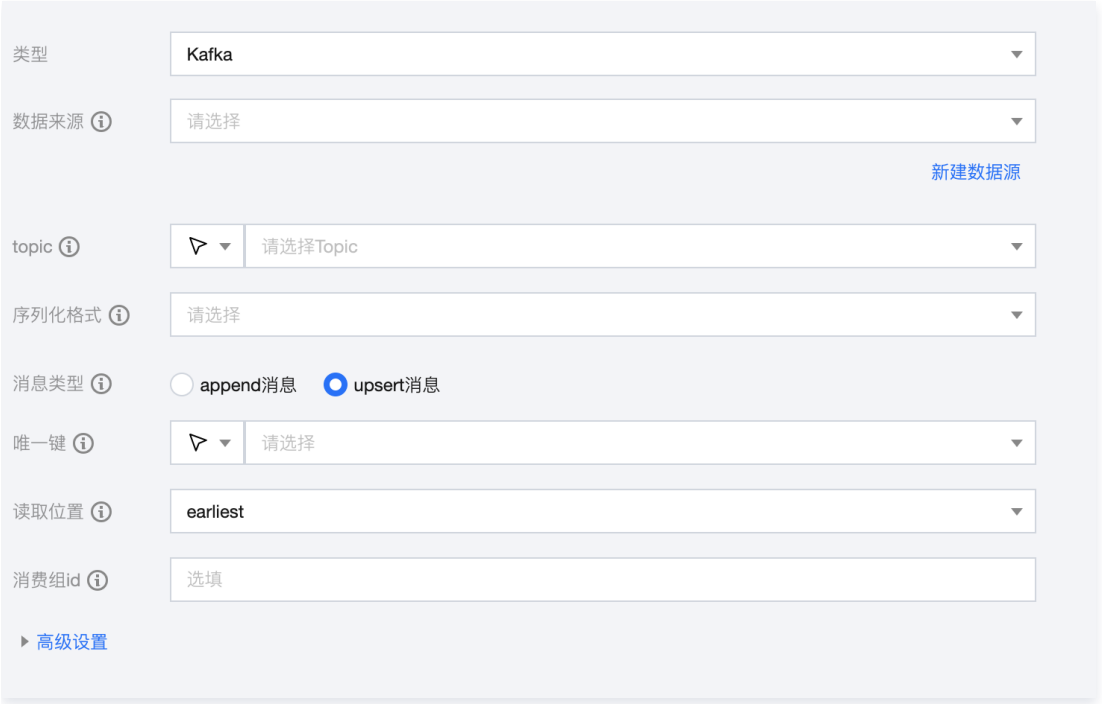
```
            "field": "pos"
          },
          {
            "type": "int32",
            "optional": false,
            "field": "row"
          },
          {
            "type": "int64",
            "optional": true,
            "field": "thread"
          },
          {
            "type": "string",
            "optional": true,
            "field": "query"
          }
        ],
        "optional": false,
        "name": "io.debezium.connector.mysql.Source",
        "field": "source"
      },
      {
        "type": "string",
        "optional": false,
        "field": "op"
      },
      {
        "type": "int64",
        "optional": true,
        "field": "ts_ms"
      }
    ],
    "optional": false,
    "name": "mysql-server-1.inventory.customers.Envelope"
  },
  "payload": {
    "op": "c",
    "ts_ms": 1465491411815,
    "before": null,
    "after": {
      "id": 12003,
      "first_name": "Anne322",
      "last_name": "Kretchmar3222",
      "email": "annek@noanswer.org3222"
    },
    "source": {
      "version": "1.9.6.Final",
      "connector": "mysql",
      "name": "mysql-server-1",
      "ts_ms": 0,
      "snapshot": false,
      "db": "inventory333",
      "table": "customers433",
      "server_id": 0,
      "gtid": null,
      "file": "mysql-bin.000003",
      "pos": 154,
      "row": 0,
      "thread": 7,
      "query": ""
```


©2013–2025 Tencent Cloud. All rights reserved.

```
        }
    }
}
```

Single Table Read Node Configuration

1. On the DataInLong page, click Real-time synchronization in the left-hand navigation bar.
2. On the Real-time synchronization page, select Single Table Synchronization to create a new configuration (you can choose between Form mode and Canvas mode) and enter the configuration page.



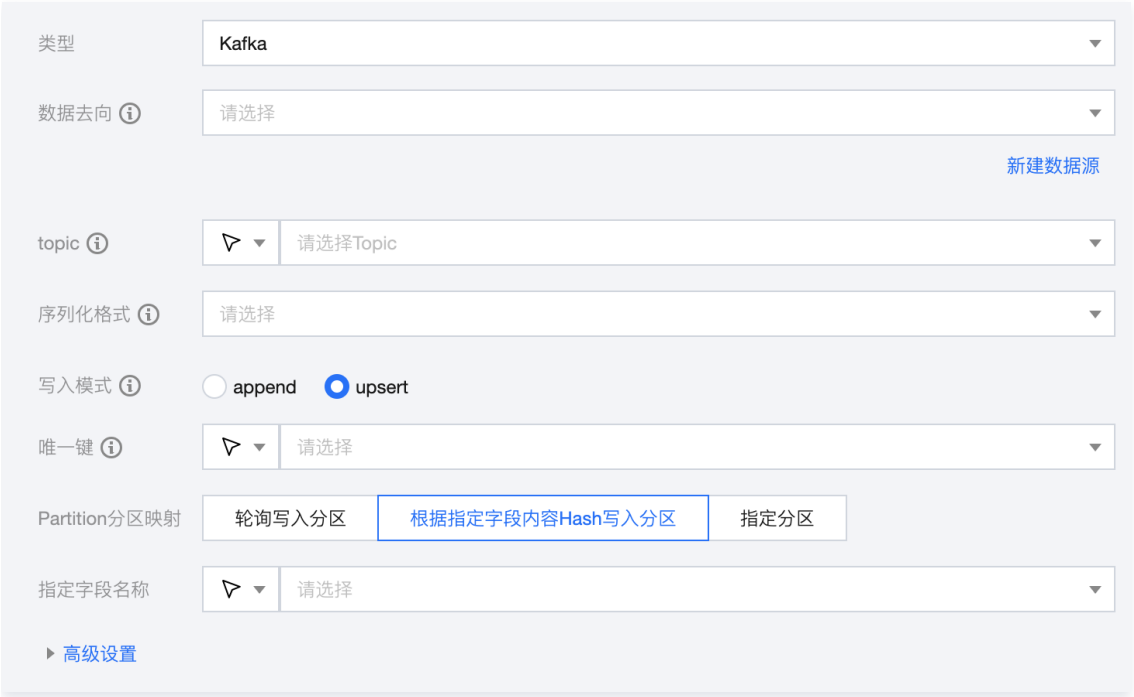
Parameter	Description
Data Source	The Kafka read-side data source type supports Kafka.
Topic	Kafka topic in the data source.
Serialization Format	Kafka message serialization format type supports: canal-json, ogg-json, json, avro, csv, raw.
Message Type	• Append Message: Messages in Kafka originate from the Append message stream, typically without carrying a unique key. It is recommended to use the Append write mode for write nodes. • Upsert Message: Messages in Kafka originate from the Upsert message stream, typically carrying a unique key. Once set, messages can ensure Exactly-Once delivery. It is recommended to use the Upsert write mode for write nodes.
Unique key	In Upsert write mode, you need to set a unique key to ensure data orderliness
Read Position	The starting position for synchronizing data when the sync task starts. Supports earliest and latest
Consumer Group ID	Please avoid this parameter overlapping with other consumption processes to ensure the correctness of the consumption point. If not specified, the default setting is group.id=WeData_ group_${task_id}.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Single Table Write Node Configuration

1. On the DataInLong page, click Real-time synchronization in the left-hand navigation bar.

©2013-2025 Tencent Cloud. All rights reserved.

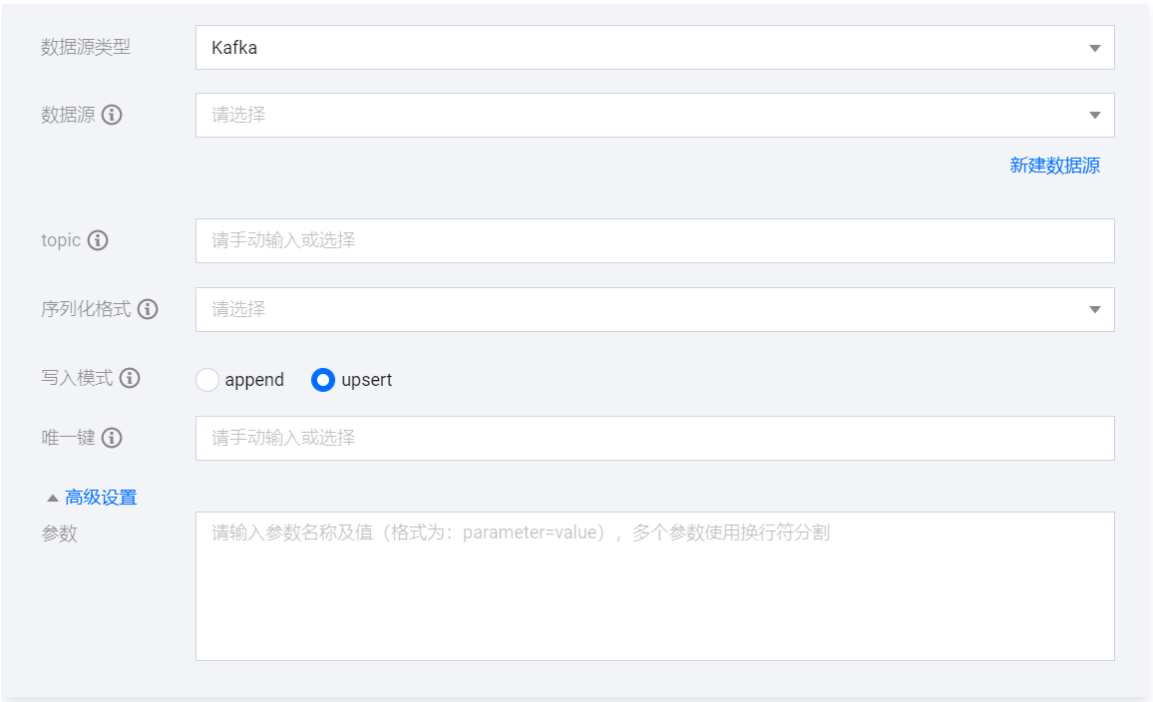
2. On the Real-time synchronization page, select Single Table Synchronization to create a new configuration (you can choose between Form mode and Canvas mode) and enter the configuration page.



Parameter	Description
Data Source	The Kafka write-side data source type supports Kafka.
Topic	Kafka topic in the data source.
Serialization Format	Kafka message serialization format type supports: canal-json, json, avro.
Write mode	• Upsert: Update write. If there is no primary key conflict, a new row is inserted; if there is a primary key conflict, the data is updated. Suitable for scenarios where the target table has a primary key and needs to be updated in real-time based on source data. There will be some performance overhead. • Append: Append write. Regardless of the presence of a primary key, data is appended by inserting new rows. Whether there is a primary key conflict depends on the target end. Suitable for scenarios without primary keys and allowing duplicate data. No performance loss.
Unique key	In Upsert write mode, you need to set a unique key to ensure data orderliness
Partition Partition Mapping	Configure topic partition mapping (polling partition writing, hash partitioning based on specified field, designated partition): • Polling Write to Partition: Polling (Round Robin) upstream data writes to each partition. • Hash based on specified field content: ○ Write partition: Hash mapping based on specified field content writes to each partition. ○ Designated partition: Enter the partition number; all messages are written only to a fixed partition.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Log Collection Write Node Configuration

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Source	Select the available Kafka data source in the current project. The Kafka write-end data source type supports Kafka and CKafka.
topic	Kafka topic in the data source.
Serialization Format	Kafka message serialization format type supports three types: • canal-json • json • avro
Write mode	Kafka supports two write modes: • Append: Append data. • Upsert: Insert messages in upsert mode; once set, messages can only be processed by the message endpoint once to ensure Exactly-Once delivery.
Unique key	In Upsert mode, a unique key must be set to ensure data order, and multiple selections are supported. The Append mode does not require setting up a unique key.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Data type conversion for reading and writing supported

Internal Type	Kafka Type
SMALLINT	SMALLINT, TINYINT UNSIGNED, TINYINT UNSIGNED ZEROFILL
INTEGER	INT, INTEGER, YEAR, SHORT, MEDIUMINT, SMALLINT UNSIGNED, SMALLINT UNSIGNED ZEROFILL
BIGINT	BIGINT, INT UNSIGNED, MEDIUMINT UNSIGNED, MEDIUMINT UNSIGNED ZEROFILL, INT UNSIGNED ZEROFILL

©2013-2025 Tencent Cloud. All rights reserved.

DECIMAL	BIGINT UNSIGNED, BIGINT UNSIGNED ZEROFILL, SERIAL, NUMERIC, NUMERIC UNSIGNED, NUMERIC UNSIGNED ZEROFILL, DECIMAL, DYNAMIC DECIMAL, DECIMAL UNSIGNED, DECIMAL UNSIGNED ZEROFILL, FIXED, FIXED UNSIGNED, FIXED UNSIGNED ZEROFILL
FLOAT	FLOAT, FLOAT UNSIGNED, FLOAT UNSIGNED ZEROFILL
DOUBLE	DOUBLE, DOUBLE UNSIGNED, DOUBLE UNSIGNED ZEROFILL, DOUBLE PRECISION, DOUBLE PRECISION UNSIGNED, ZEROFILL, REAL, REAL UNSIGNED, REAL UNSIGNED ZEROFILL
TIMESTAMP	ATETIME, TIMESTAMP WITH LOCAL TIME ZONE, TIMESTAMP WITH TIME ZONE
TIMESTAMP_WITH_TIMEZONE	TIMESTAMP, TIMESTAMP WITH LOCAL TIME ZONE, TIMESTAMP WITH TIME ZONE
BLOB	BLOB, TINYBLOB, MEDIUMBLOB, LONGBLOB
VARCHAR	JSON, VARCHAR, TEXT, TINYTEXT, MEDIUMTEXT, LONGTEXT

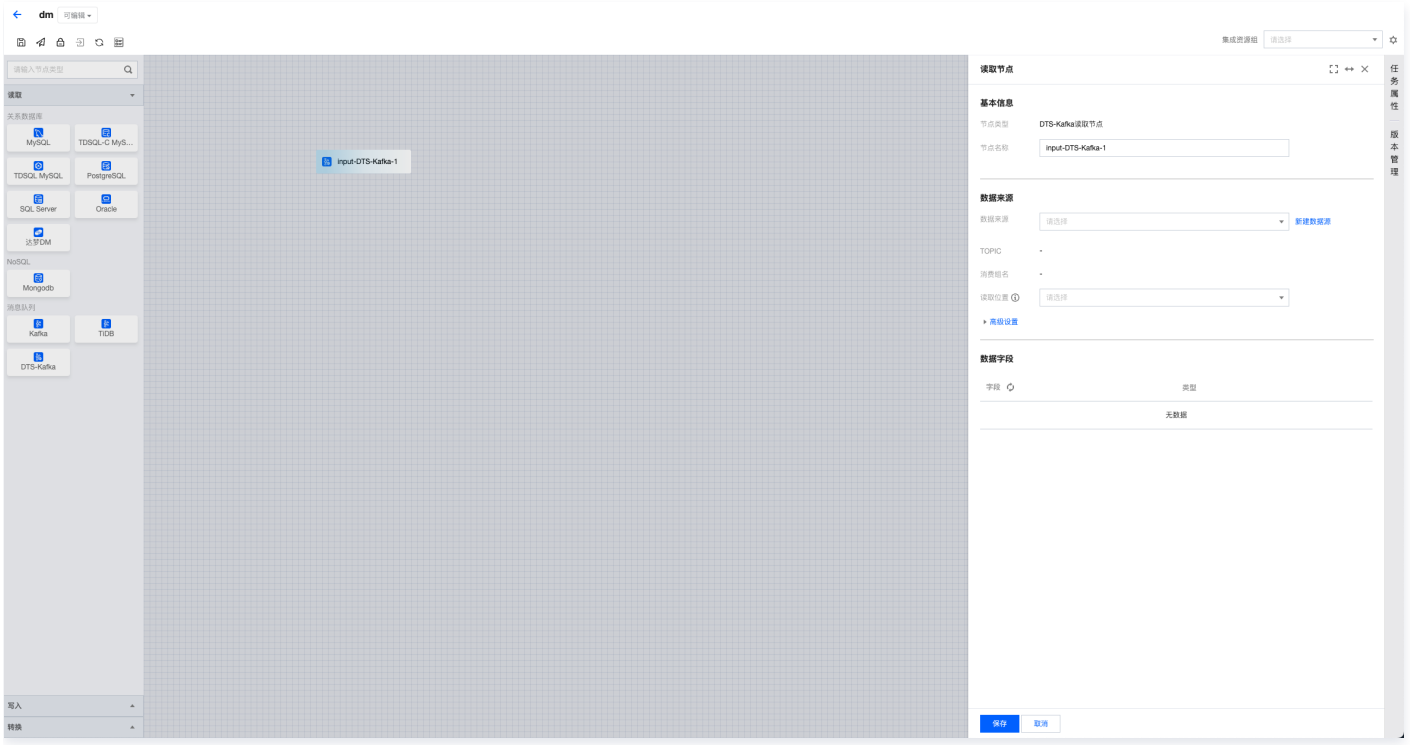
©2013–2025 Tencent Cloud. All rights reserved.

DTS-Kafka Data Source

Last updated: 2024-09-06 14:55:32

DTS-Kafka real-time single table read cloud data disks configuration

1. click the left **Read**, click **DTS-Kafka cloud data disks** and configure the node information.



2. You can refer to the table below for parameter configuration:

Parameter	Description
Node name	DTS-Kafka read cloud data disks
Data Source	Select the data source where the table to be synced is located
TOPIC	-
Consumer group name	-
Read Position	Starting point for data synchronization when the synchronization task is initiated
Advanced Settings	You can configure parameters based on business requirements.

3. Preview data fields, click **Save**.

©2013-2025 Tencent Cloud. All rights reserved.

Mongodb Data Source

Last updated: 2024-09-06 14:55:54

Mongodb Data Source Version Limitation

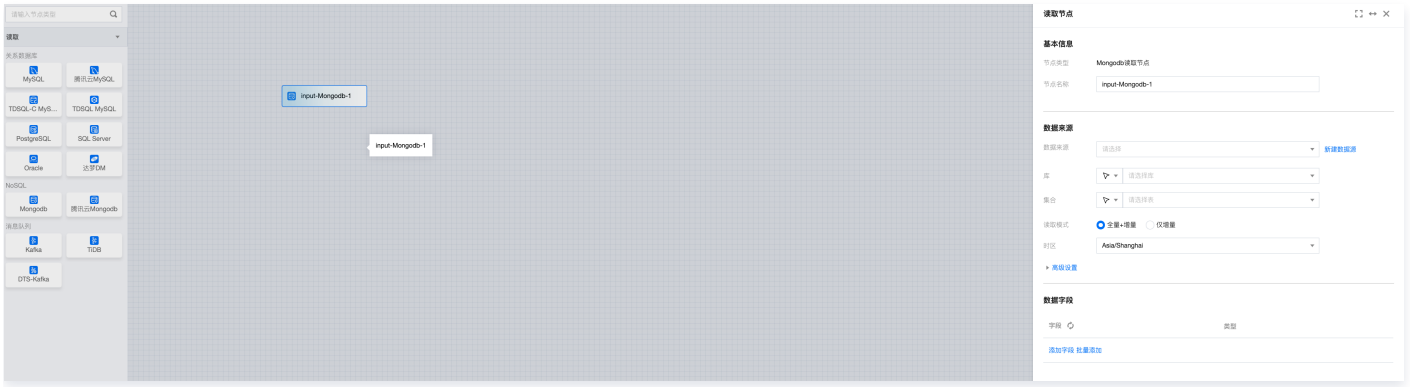
Supported MongoDB Versions Details:

Node	Version
MongoDB-CDC	MongoDB>=3.6

Mongodb Single Table Read Node Configuration

Mongodb Read Configuration Parameter Instructions

1. In the DataInLong page, click **Real-time Synchronization** in the left sidebar.
2. On the Real-time Synchronization page, select **Single Table Synchronization** at the top, create a new one (choose Form or Canvas mode), and enter the configuration page.
3. Click the left side **Read**, then click **Select Mongo Node** and configure node information.



4. Parameter description:

Parameter	Description
Node name	Enter the Mongo node name.
Data Source	Select the available Mongo data source for the project
Database	Support selecting or manually entering the name of the database to be read • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Collection	Support selecting or manually entering the name of the collection to be read
Read mode	• Full + Incremental: Data synchronization is divided into full and incremental synchronization phases. After the full phase is completed, the task enters the incremental phase. The full phase will synchronize historical data in the database, and the incremental phase will start synchronizing from the binlog CDC location after the task starts. • Incremental: Only synchronize data from the binlog CDC location after the task starts.
Time zone	Set the log time timezone, default is Shanghai.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

©2013-2025 Tencent Cloud. All rights reserved.

5. Preview data fields, click **Save**.

Mongodb Whole Database Source Configuration

Mongodb Read Configuration Parameter Instructions



Parameter	Description
Data Source	Select the Mongo data source to synchronize.
Source Table	According to business needs, choose "All Databases and Collections," "Specific Table," or "Specific Database": • All Databases and Collections: Monitor all databases under the data source. New databases and tables added during the task run will be synchronized to the target by default. • Specific Collection: Only synchronize the specified collection. • Specify Database: Monitor the specified database and synchronize all or matching tables under it.
Read mode	Full + Increment, Increment.
Filter operations	Provide multiple filter operations, including Insert, Update, Delete, Delete Collection, Delete Database, Rename Collection. After setting, data of the specified operation type will not be synchronized.

Data type conversion is supported

Read

The supported data types and corresponding transformations for Mongodb reading are as follows:

BSON Type	Internal Type
–	TINYINT
–	SMALLINT
Int	INT
Long	BIGINT
–	FLOAT
Double	DOUBLE
Decimal128	DECIMAL(p, s)
Boolean	BOOLEAN
Date Timestamp	DATE
Date Timestamp	TIME

©2013–2025 Tencent Cloud. All rights reserved.

Date	TIMESTAMP(3) TIMESTAMP_LTZ(3)
Timestamp	TIMESTAMP(0) TIMESTAMP_LTZ(0)
String,ObjectId,UUID ,Symbol ,MD5 ,JavaScript,Regex	STRING
BinData	BYTES
Object	ROW
Array	ARRAY
DBPointer	ROW&ltSTRING, STRING>

©2013–2025 Tencent Cloud. All rights reserved.

TDSQL for MySQL Data Source

Last updated: 2024-09-06 14:56:26

DataInLong provides read and write capabilities for TDSQL for MySQL. This article introduces the pre-environment configuration for real-time data synchronization using TDSQL-C for MySQL and the current capability support.

Supported Versions

Currently, DataInLong supports real-time reading of single tables and database-level real-time reading and single table writing in TDSQL-C for MySQL. To use real-time synchronization capability, follow these version restrictions:

Engine types	Version
MySQL	Only the original InnoDB type is supported; TDStore type is not supported

Use Limits

- Dynamic shard addition is not supported. You need to pause the task first, then add the shard.
- You need to enable Binlog logs. Only TDSQL for MySQL server Binlog configuration format set to ROW is supported for synchronization.
- Tables without primary keys might experience data duplication due to the inability to ensure exactly once. Therefore, it is best to have a primary key for real-time synchronization tasks.
- XA ROLLBACK is not supported. Real-time synchronization tasks will not rollback data for XA PREPARE. To handle XA ROLLBACK scenarios, you need to manually remove the XA ROLLBACK tables from real-time synchronization tasks and then re-sync the tables after adding them back.
- Set the TDSQL for MySQL session timeout:

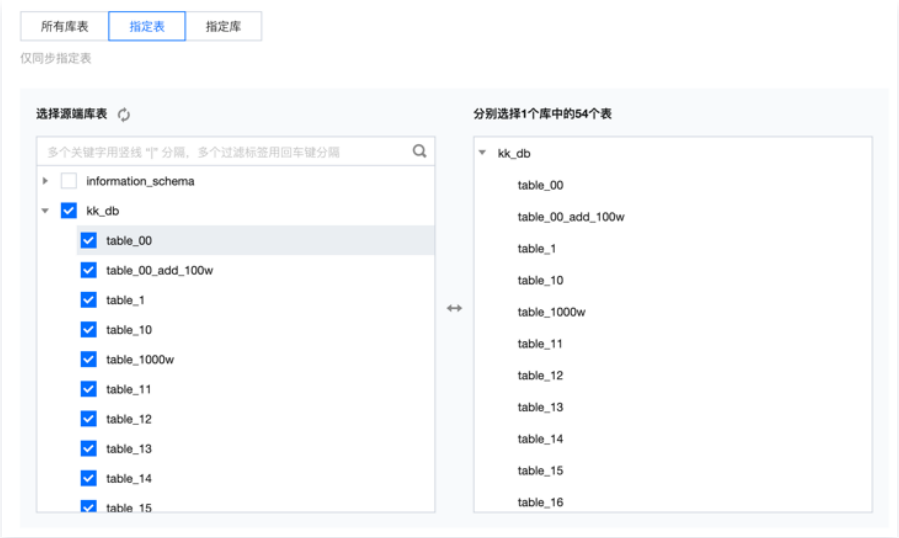
  When creating an initial consistent snapshot for large databases, the connections you establish might time out while reading the tables. You can prevent this behavior by configuring interactive_timeout and wait_timeout in the TDSQL for MySQL configuration file.
  - interactive_timeout: The number of seconds the server waits for activity on an interactive connection before closing it. Refer to Documentation.
  - wait_timeout: The number of seconds the server waits for activity on a non-interactive connection before closing it. Refer to Documentation.

Database-level source configuration

Data Source Settings



©2013-2025 Tencent Cloud. All rights reserved.

Parameter	Description
Data Source	Select the TDSQL for MySQL data source to be synchronized.
Source Table	• All databases and tables: Monitor all databases under the data source. During the task execution, new databases and tables will be synchronized to the target end by default. • Specified Table: Under this option, you need to specify the specific table name. After setting, the task will only synchronize the specified table; if a new table needs to be synchronized, the task must be stopped and restarted.  • Specify Database: Under this option, you need to specify the specific database name with a table name regular expression. After setting, new tables that match the table name expression will be synchronized to the target end by default during task execution. 
Read mode	• Full + Incremental: Data synchronization is divided into full and incremental synchronization phases. After the full phase is completed, the task enters the incremental phase. The full phase will synchronize historical data in the database, and the incremental phase will start synchronizing from the binlog CDC location after the task starts. • Increment only: Only sync data starting from the binlog cdc point after the task starts.
Consistency Semantics	It only represents the consistency semantics of the reading end. Supports At-least-once and Exactly-once. • At-least-once: Data may be read repeatedly, and relies on the target end to support deduplication to ensure data consistency. Suitable for scenarios with a large volume of data in the full stage and non-numeric primary keys, requiring high synchronization performance. • Exactly-once: Data is read strictly once, with some performance loss, and does not support tables without primary keys and unique index columns. Suitable for general scenarios where the source table has numeric primary keys or unique index columns. The two modes are incompatible in the current version. If the mode is changed after task submission, stateful restart is not supported.
Filter operations	Support three operations: Insert, Update, and Delete. After setting, data of the specified operation types will not be synchronized.

©2013-2025 Tencent Cloud. All rights reserved.

Time zone	Set the log time timezone, default is Shanghai.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Supported Target Data Sources

Currently, TDSQL for MySQL supports real-time whole database synchronization to the following types of target ends:

- DLC Whole Database Target Configuration Details
- Iceberg Whole Database Target Configuration Details
- Doris Whole Database Target Configuration Details

Real-time Single Table Read Node Configuration



Parameter	Description
Data Source	Select the data source where the table to be synced is located
Database	Select the database where the table to be synced is located
Table	Supports selecting or manually entering the name of the table to be written to.
Shard Column	Shard columns are used to divide the table into multiple shards for synchronization. • For tables with primary keys, it is recommended to choose the primary key as the shard column. • For tables without primary keys, it is recommended to choose an indexed column as the shard column and ensure that no data update operations are performed on the shard column; otherwise, only At-Least-Once semantics can be guaranteed.
Read mode	• Full + Incremental: Data synchronization is divided into full and incremental synchronization phases. After the full phase is completed, the task enters the incremental phase. The full phase will synchronize historical data in the database, and the incremental phase will start synchronizing from the binlog CDC location after the task starts. • Increment only: Only sync data starting from the binlog cdc point after the task starts.

©2013-2025 Tencent Cloud. All rights reserved.

Consistency Semantics	It only represents the consistency semantics of the reading end. Supports At-least-once and Exactly-once. • At-least-once: Data may be read repeatedly, and relies on the target end to support deduplication to ensure data consistency. Suitable for scenarios with a large volume of data in the full stage and non-numeric primary keys, requiring high synchronization performance. • Exactly-once: Data is read strictly once, with some performance loss, and does not support tables without primary keys and unique index columns. Suitable for general scenarios where the source table has numeric primary keys or unique index columns. The two modes are incompatible in the current version. If the mode is changed after task submission, stateful restart is not supported.
Filter operations	After setting, data of the specified operation types will not be synchronized. Supports Insert, Update, and Delete.
Time zone	Set the log time timezone, default is Shanghai.
Advanced Settings (optional)	You can configure parameters based on business requirements.

Real-time Single Table Write Node Configuration



Parameter	Description
Data Flow Direction	The TDSQL for MySQL data source that needs to be written.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Supports selecting or manually entering the name of the table to be written to. When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Advanced Settings (optional)	You can configure parameters based on business requirements.

Data type conversion is supported

©2013-2025 Tencent Cloud. All rights reserved.

Read

Field Type	Supported	Internal mapping fields	Remarks
TINYINT	Yes	TINYINT	TINYINT(1) is mapped to BOOLEAN by default Single table tasks can map TINYINT(1) to tinyint through advanced settings
SMALLINT	Yes	SMALLINT	–
TINYINT_UNSIGNED	Yes	SMALLINT	–
TINYINT_UNSIGNED_ZEROFILL	Yes	SMALLINT	–
INT	Yes	INT	–
INTEGER	Yes	INT	–
YEAR	Yes	INT	–
MEDIUMINT	Yes	INT	–
SMALLINT_UNSIGNED	Yes	INT	–
SMALLINT_UNSIGNED_ZEROFILL	Yes	INT	–
BIGINT	Yes	LONG	–
INT_UNSIGNED	Yes	LONG	–
MEDIUMINT_UNSIGNED	Yes	LONG	–
MEDIUMINT_UNSIGNED_ZEROFILL	Yes	LONG	–
INT_UNSIGNED_ZEROFILL	Yes	LONG	–
BIGINT_UNSIGNED	Yes	DECIMAL	DECIMAL(20,0)
BIGINT_UNSIGNED_ZEROFILL	Yes	DECIMAL	DECIMAL(20,0)
SERIAL	Yes	DECIMAL	DECIMAL(20,0)
FLOAT	Yes	FLOAT	–
FLOAT_UNSIGNED	Yes	FLOAT	–
FLOAT_UNSIGNED_ZEROFILL	Yes	FLOAT	–
DOUBLE	Yes	DOUBLE	–
DOUBLE_UNSIGNED	Yes	DOUBLE	–
DOUBLE_UNSIGNED_ZEROFILL	Yes	DOUBLE	–
DOUBLE_PRECISION	Yes	DOUBLE	–
DOUBLE_PRECISION_UNSIGNED	Yes	DOUBLE	–
ZEROFILL	Yes	DOUBLE	–
REAL	Yes	DOUBLE	–
REAL_UNSIGNED	Yes	DOUBLE	–
REAL_UNSIGNED_ZEROFILL	Yes	DOUBLE	–

©2013–2025 Tencent Cloud. All rights reserved.

NUMERIC	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
NUMERIC_UNSIGNED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
NUMERIC_UNSIGNED_ZEROFILL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
DECIMAL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
DECIMAL_UNSIGNED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
DECIMAL_UNSIGNED_ZEROFILL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
FIXED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
FIXED_UNSIGNED	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
FIXED_UNSIGNED_ZEROFILL	Yes	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
BOOLEAN	Yes	BOOLEAN	–
DATE	Yes	DATE	–
TIME	Yes	TIME	–
DATETIME	Yes	TIMESTAMP	–
TIMESTAMP	Yes	TIMESTAMP	–
CHAR	Yes	STRING	–
JSON	Yes	STRING	–
BIT	Yes	STRING	BIT(1) maps to BOOLEAN
VARCHAR	Yes	STRING	–
TEXT	Yes	STRING	–
BLOB	Yes	STRING	–
TINYBLOB	Yes	STRING	–
TINYTEXT	Yes	STRING	–
MEDIUMBLOB	Yes	STRING	–
MEDIUMTEXT	Yes	STRING	–
LONGBLOB	Yes	STRING	–

©2013–2025 Tencent Cloud. All rights reserved.

LONGTEXT	Yes	STRING	-
VARBINARY	Yes	STRING	-
GEOMETRY	Yes	STRING	-
POINT	Yes	STRING	-
LINESTRING	Yes	STRING	-
POLYGON	Yes	STRING	-
MULTIPOINT	Yes	STRING	-
MULTILINESTRING	Yes	STRING	-
MULTIPOLYGON	Yes	STRING	-
GEOMETRYCOLLECTION	No	-	-
ENUM	Yes	STRING	-
BINARY	Yes	BINARY	BINARY(1)
SET	No	-	-

Write

Internal Type	MySQL Type
TINYINT	TINYINT
SMALLINT	SMALLINT,TINYINT UNSIGNED
INT	INT,MEDIUMINT,SMALLINT UNSIGNED
BIGINT	BIGINT,INT UNSIGNED
DECIMAL(20, 0)	BIGINT UNSIGNED
FLOAT	FLOAT
DOUBLE	DOUBLE,DOUBLE PRECISION
DECIMAL(p, s)	NUMERIC(p, s),DECIMAL(p, s)
BOOLEAN	BOOLEAN,TINYINT(1)
DATE	DATE
TIME [(p)][WITHOUT TIMEZONE]	TIME [(p)]
TIMESTAMP [(p)][WITHOUT TIMEZONE]	DATETIME [(p)]
STRING	CHAR(n),VARCHAR(n),TEXT
BYTES	BINARY,VARBINARY,BLOB
ARRAY	-

©2013-2025 Tencent Cloud. All rights reserved.

OceanBase Data Source

Last updated：2024-09-06 14:56:58

DataInLong provides the capability to read from OceanBase. This document introduces the pre-environment configuration for real-time data synchronization using OceanBase and the current support status.

Supported Versions

Currently, DataInLong supports real-time reading at the single table and full database level for OceanBase. To use the real-time reading capability, the following version limitations must be adhered to:

Type	Version
OceanBase	Community Edition 3.1.4; Enterprise Edition 4.2.1.3

Use Limits

- Go live with whitelisting. To use it, please submit a ticket for application.
- Incremental collection requires the deployment of logproxy; without it, the task cannot enter the incremental phase.
- logproxy must be registered with the primary account, but the primary account cannot be used when configuring data sources.
- When the source is OceanBase for a full database task, the target end only supports using existing databases and tables; automatic database and table creation are not supported.
- DDL changes are not supported; schema changes at the source will not be automatically synchronized to the target.
- Only At-least-once semantics are supported. Target-side deduplication capability is needed to ensure data uniqueness.

Real-time full database reading node configuration

Data Source Settings

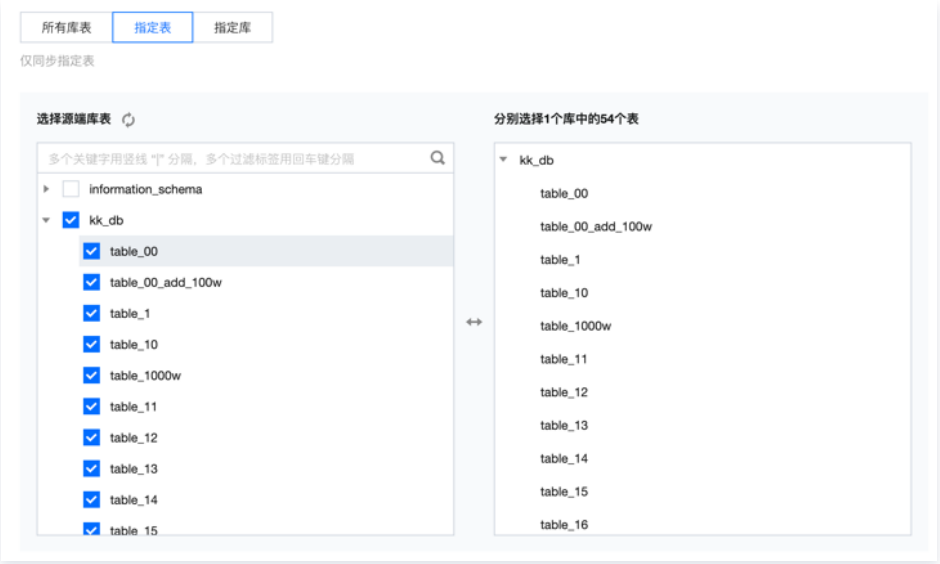
- There are differences in configuration between OceanBase MySQL Mode and Oracle Mode. MySQL Mode configuration is as follows:



- Oracle Mode configuration is as follows:

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Source Mode	Supports MySQL Compatibility Mode and Oracle Compatibility Mode
Data Source	Select the OceanBase data source that needs synchronization. This will be filtered based on the selected data source mode. For example, if MySQL Mode is selected, only OceanBase data sources in MySQL Mode will be displayed, not Oracle Mode data sources.
Source Table	• All databases and tables: Monitor all databases under the data source. During the task execution, new databases and tables will be synchronized to the target end by default. • Specified Table: Under this option, you need to specify the specific table name. After setting, the task will only synchronize the specified table; if a new table needs to be synchronized, the task must be stopped and restarted.  • Specified Database: Under this option, you need to specify the exact database name and table name regular expression. After setting, any new tables conforming to the table name expression will be synchronized to the target end during task execution.

©2013–2025 Tencent Cloud. All rights reserved.

	
Read mode	• Full + Incremental: Data synchronization is divided into full and incremental synchronization phases. After the full phase is completed, the task enters the incremental phase. The full phase will synchronize historical data in the database, and the incremental phase will start synchronizing from the binlog CDC location after the task starts. • Incremental: Only synchronize data from the binlog CDC location after the task starts.
Filter operations	Supports three operations: Insert, Update, and Delete. After setting, specific operation types of data won't be synchronized. Only MySQL Mode supports filtering operations.
Time zone	Set the timezone for log timestamps, default is Shanghai. Only MySQL Mode supports timezones.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Supported Target Data Sources

Currently supports real-time full database synchronization of OceanBase data to the following target ends:

- Doris Whole Database Target Configuration Details
- DLC Whole Database Target Configuration Details
- Iceberg Whole Database Target Configuration Details

Real-time Single Table Read Node Configuration



Parameter	Description
Data Source Mode	Supports MySQL Compatibility Mode and Oracle Compatibility Mode.

©2013-2025 Tencent Cloud. All rights reserved.

Data Source	Select the data source where the table to be synced is located.
Database	Select the database where the table to be synchronized is located.
Table	Select the table to be synced.
Table primary key	Used to split tables into multiple shards for synchronization, only applicable for MySQL Mode.
Read mode	• Full + Incremental: Data synchronization is divided into full and incremental synchronization phases. After the full phase is completed, the task enters the incremental phase. The full phase will synchronize historical data in the database, and the incremental phase will start synchronizing from the binlog CDC location after the task starts. • Incremental: Only synchronize data from the binlog CDC location after the task starts.
Advanced Settings (optional)	You can configure parameters based on business requirements.

Supports Data Type Conversion

The supported data types by OceanBase for reading and their corresponding conversion relationships are as follows (when processing OceanBase, it will first map the data types of OceanBase data sources and the data processing engine):

OceanBase MySQL Mode Type	Internal Type
TINYINT/INT/MEDIUMINT/INTEGER	INTEGER
FLOAT,BINARY_FLOAT	FLOAT
DECIMAL/NUMETRIC	DECIMAL
DOUBLE PRECISION,BINARY_DOUBLE	DOUBLE
BOOL/BOOLEAN	BOOLEAN
DATE,TIMESTAMP [(p)]	TIMESTAMP [(p)] [WITHOUT TIMEZONE]
TIMESTAMP [(p)] WITH TIME ZONE	TIMESTAMP [(p)] WITH TIME ZONE
TIMESTAMP [(p)] WITH LOCAL TIME ZONE	TIMESTAMP_LTZ [(p)]
CHAR(n),NCHAR(n),NVARCHAR2(n),VARCHAR(n),VARCHAR2(n),CLOB,NCLOB,XML, BIT Type	STRING
BLOB,ROWID	BYTES
INTERVAL DAY TO SECOND,INTERVAL YEAR TO MONTH	BIGINT

OceanBase Oracle Mode Type	Internal Type
NUMBER	DOUBLE
FLOAT,BINARY_FLOAT	FLOAT
DOUBLE PRECISION,BINARY_DOUBLE	DOUBLE
NUMBER(1)	BOOLEAN
DATE,TIMESTAMP [(p)]	TIMESTAMP [(p)] [WITHOUT TIMEZONE]
TIMESTAMP [(p)] WITH TIME ZONE	TIMESTAMP [(p)] WITH TIME ZONE
TIMESTAMP [(p)] WITH LOCAL TIME ZONE	TIMESTAMP_LTZ [(p)]
CHAR(n),NCHAR(n),NVARCHAR2(n),VARCHAR(n),VARCHAR2(n),CL	STRING

©2013–2025 Tencent Cloud. All rights reserved.

OB,NCLOB,XML Type	
BLOB,ROWID	BYTES
INTERVAL DAY TO SECOND,INTERVAL YEAR TO MONTH	BIGINT

©2013–2025 Tencent Cloud. All rights reserved.

Dameng DM Data Source

Last updated: 2024-09-06 14:57:30

DataInLong provides single table read and write capabilities for Dameng DM. This article introduces the pre-environment configuration for real-time data synchronization using Dameng DM and the current capability support situation.

Supported Versions

Currently, DataInLong supports real-time single table read and write for Dameng DM. The following version limitations must be adhered to when using real-time synchronization capabilities:

Type	Version
Dameng DM	7,8

Use Limits

- Currently, only single table real-time read and write are supported. Whole database tasks are not supported.
- Only At-least-once semantics are supported. Target-side deduplication capability is needed to ensure data uniqueness.
- When DM is used as the target, it does not support one-click creation of DM target-side table structures based on source table structures. Target tables need to be created in the DM database in advance.

Database Environment Preparation

Enable archive mode

1. Use the following statements to check whether ARCH_INI and RLOG_APPEND_LOGIC parameters are correctly set to ensure archive mode is enabled:

```
select para_name, para_value
from v$dm_ini
where para_name in ('ARCH_INI','RLOG_APPEND_LOGIC');
```

2. If the parameters show 0, follow the steps below to set the archive mode and modify the parameters.

   2.1 Modify the parameters in dm.ini as follows:

```
ARCH_INI = 1
RLOG_APPEND_LOGIC = 1
```

   2.2 dmarch.ini needs to configure local archiving. For example:

```
[ARCHIVE_LOCAL1]
ARCH_TYPE = LOCAL
ARCH_DEST = d:\dmdbms\arch
ARCH_FILE_SIZE = 128 #unit: Mb
ARCH_SPACE_LIMIT = 0 #unit: Mb, 0 means unlimited, range: 1024~4294967294M
ARCH_INI and RLOG_APPEND_LOGIC
```

   2.3 The parameters are described as follows:

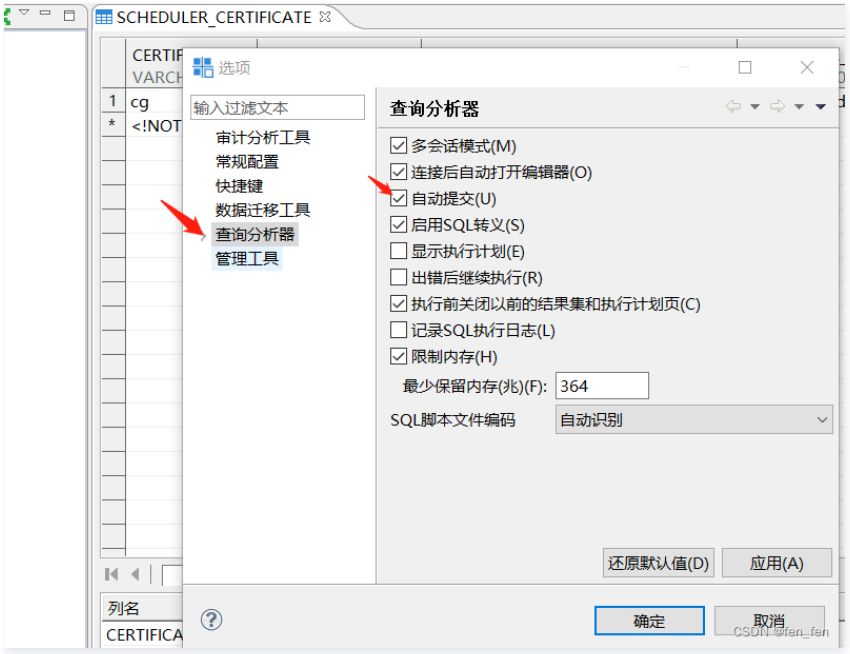
©2013-2025 Tencent Cloud. All rights reserved.

参数名	缺省值	属性	说明
RLOG_APPEND_LOGIC	0	动态，系统级	是否启用在日志中记录逻辑操作的功能，取值范围 0、1、2、3 0：不启用；1、2、3 启用。1：如果有主键列，记录 UPDATE 和 DELETE 操作时只包含主键列信息，若没有主键列则包含所有列信息；2：不论是否有主键列，记录 UPDATE 和 DELETE 操作时都包含所有列的信息；3：记录 UPDATE 时包含更新列的信息以及 ROWID，记录 DELETE 时只有 ROWID
ARCH_INI	0	动态，系统级	是否启用归档，0：不启用；1：启用

Enable automatic transaction commitment

Most DMC tools enable automatic transaction commitment by default. After executing a Select, Insert, Delete, or Update statement, the commit operation is automatically performed. However, Dameng DMC tools do not work this way. Whether you use TCCLI disql or visual management tools, automatic commitment is not performed by default. You must manually commit the operation.

Method 1: In the management tool menu, select **Window** > **Options**, click **Query Analyzer**, and check **Autocommit**.



Method 2: Use the SQL command method.

The temporary modification can be configured as follows, from the official documentation:

1. Set whether the current session automatically commits modified data.
   - ON: Enables autocommit. The transactions of all executed SQL statements will be automatically committed.
   - OFF: Disables autocommit. The transactions of all executed SQL statements will require explicit user commitment. This is the default setting.

```
SET AUTO[COMMIT] <ON|OFF>
```

Permission granting

```
grant create session to my_user;
grant select on all_users to  my_user ;
grant select on all_objects to  my_user ;
grant select on all_tab_cols to  my_user ;
```

©2013–2025 Tencent Cloud. All rights reserved.

```
grant select on all_indexes to  my_user ;
grant select on dba_objects to  my_user ;
grant select on V$LOGMNR_CONTENTS to  my_user ;
grant select on V$ARCHIVED_LOGS to  my_user ;
```

Single Table Read Node Configuration

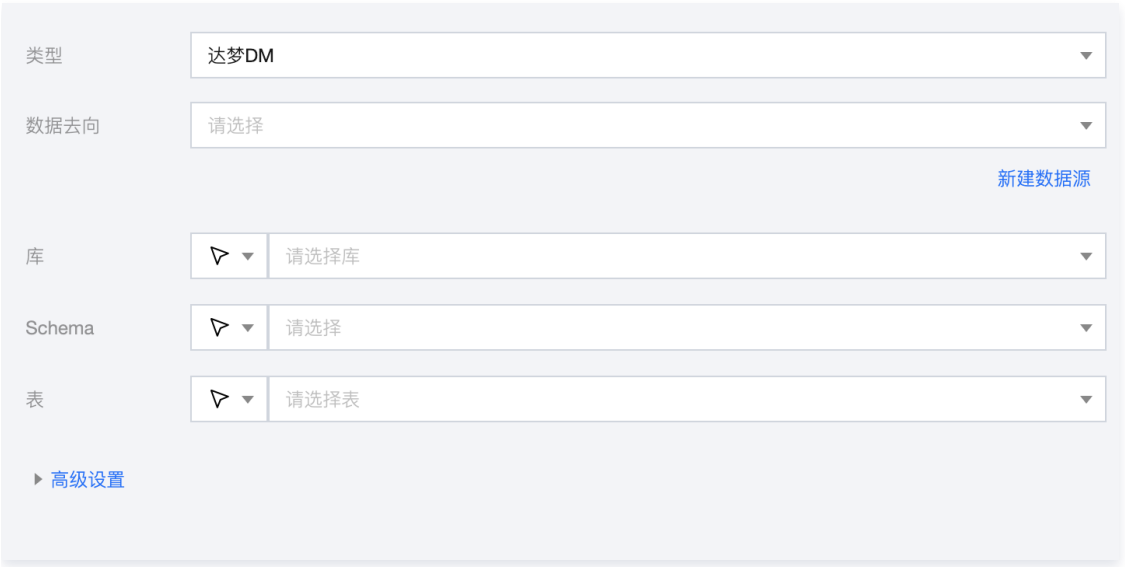


Parameter	Description
Data Source	Select Dameng DM Data Source.
Database	You can select or manually enter the database name to be read. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	You can select or manually enter the available schema under the data source to be read.
Table	You can select or manually enter the table name to be read. • In the case of table partitioning, you can select or enter multiple table names in the Dameng DM source. Multiple tables must have consistent structures. • In the case of table partitioning, you can configure the table serial number range. For example, 'table_[0-99]' means reading 'table_0', 'table_1', 'table_2' up to 'table_99'; if your table numeric suffixes are consistent in length, such as 'table_000', 'table_001', 'table_002' up to 'table_999', you can configure it as '"table": ["table_00[0-9]", "table_0[10-99]", "table_[100-999]"]'. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Adding Sharding	Suitable for sharding scenarios, click to configure multiple data sources, databases, and table information. In sharding and partitioning scenarios, ensure all table structures are consistent. The task configuration will default to display and use the first table structure for data acquisition.
Table primary key	Select a field as the primary key for the write table.
Read mode	• Full + Incremental: Data synchronization is divided into full and incremental synchronization phases. After the full phase is completed, the task enters the incremental phase. The full phase will synchronize

©2013-2025 Tencent Cloud. All rights reserved.

	historical data in the database, and the incremental phase will start synchronizing from the binlog CDC location after the task starts. • Increment only: Only sync data starting from the binlog cdc point after the task starts.
Parameter (optional)	You can configure parameters based on business requirements.

Real-time Single Table Write Node Configuration



Parameter	Description
Data Flow Direction	The Dameng DM Data Source to be written.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	You can select or manually enter the data schema for Dameng DM that needs to be written. • When the data source network is not connected, causing schema information to be unavailable for direct fetching, you can manually enter the schema name. Given network connection in DataInLong, data synchronization can still be performed.
Table	Supports selecting or manually entering the name of the table to be written to. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Advanced Settings (optional)	You can configure parameters based on business requirements.

Supported Data Types

Read

Dameng DM supports the following data types and conversion mappings (when processing Dameng DM, it first maps the data types from the Dameng DM data source to the data types of the data processing engine):

©2013–2025 Tencent Cloud. All rights reserved.

DM Types	Internal Type	Remarks
INT	INT	
TINYINT	TINYINT	TINYINT(1) maps to BOOLEAN Options need to be added to support TINYINT(1) mapping to bool or tinyint
SMALLINT	SMALLINT	
BIGINT	Long	
REAL	DOUBLE	
FLOAT	FLOAT	
DOUBLE	DOUBLE	
DECIMAL	DECIMAL	Use the actual precision of the user database p<=38 maps to DECIMAL 38 < p <= 65 maps to String
CHAR	STRING	
VARCHAR	STRING	
TEXT	STRING	
DATE	DATE	
TIMESTAMP	TIMESTAMP	

Write

The data types supported by DM Write and their corresponding conversion relationships are as follows:

Internal Type	DM Types
INT	INT,NUMBER,INTEGER
TINYINT	TINYINT(3)
SMALLINT	SMALLINT
BIGINT	BIGINT
FLOAT	FLOAT
DOUBLE	REAL,DOUBLE,DECIMAL,DOUBLE PRECISION
BOOLEAN	BIT(1)
CHAR	CHARACTER(1)
VARCHAR	VARBINARY,BYTE,VARCHAR(8188),VARCHAR2(8188),LONGVARCHAR(2147483647)
TEXT	CLOB,BINARY,TIME WITH TIME ZONE
TINYTEXT	TIME
DATE	DATE,DATETIME,DATETIME WITH TIME ZONE
TIMESTAMP	TIMESTAMP,TIMESTAMP WITH LOCAL TIME ZONE
BLOB	BLOB,IMAGE,LONGVARBINARY

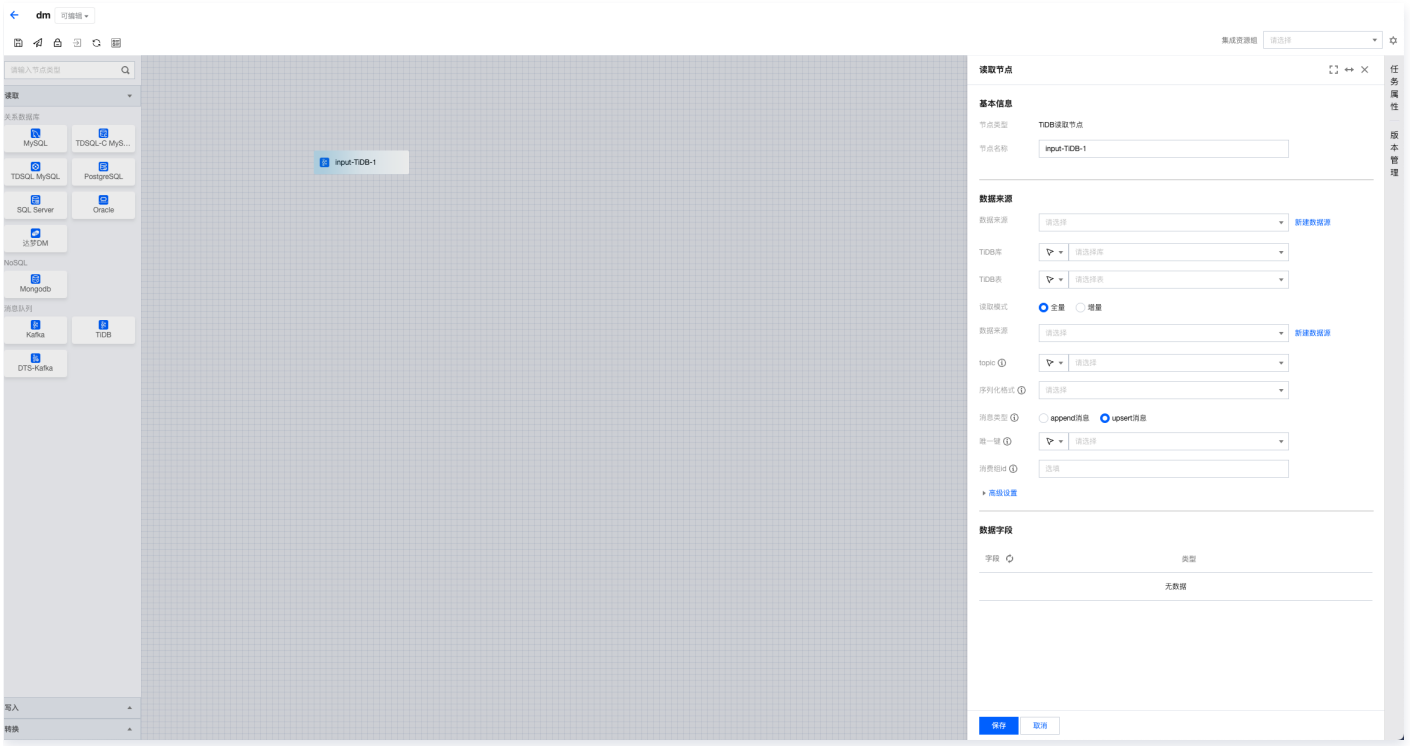
©2013–2025 Tencent Cloud. All rights reserved.

TiDB Data Source

Last updated：2024-09-06 14:57:53

TiDB Real-time Single Table Read Node Configuration

1. In the DataInLong page, click **Real-time Synchronization** in the left sidebar.
2. On the Real-time Synchronization page, select **Single Table Synchronization** at the top, create a new one (choose Form or Canvas mode), and enter the configuration page.
3. click Left Read, click Select **TiDB** Node and configure the node information.



4. You can refer to the table below for parameter configuration:

Parameter	Description
Node name	TiDB Read Node Name
Data Source	Select the TiDB data source where the table to be synchronized is located
TiDB Library	Select TiDB Library
TiDB Table	Select TiDB Table
Read mode	Supports Full and Incremental Read Modes
Data Source	Select the Kafka data source where the table to be synchronized is located
Topic	Topic in Kafka Data Source
Serialization Format	Kafka Message Serialization Format Type
Message Type	• Append Message: The message in Kafka comes from the append message stream, usually without a unique key. It is recommended to use the append write mode for write nodes • Upsert Message: The message in Kafka comes from the upsert message stream, usually with a unique key. Setting this ensures Exactly-Once. It is recommended to use the upsert write mode for write nodes
Unique key	In Upsert write mode, you need to set a unique key to ensure data orderliness

©2013-2025 Tencent Cloud. All rights reserved.

Consumer Group ID	Please avoid duplicate parameters with other consumption processes to ensure the correctness of the consumption location. If this parameter is not specified, the default setting is group.id=WeData_group_${Task ID}
Advanced Settings	You can configure parameters based on business requirements.

5. Preview data fields, click **Save**.

©2013–2025 Tencent Cloud. All rights reserved.

Doris Data Source

Last updated: 2024-09-06 14:58:21

DataInLong provides real-time writing capabilities for Doris. This article introduces the current support for real-time data synchronization using Doris.

Supported Versions

Currently, DataInLong supports real-time writing for both single tables and whole databases in Doris. To use real-time synchronization capabilities, you must follow these version restrictions:

Type	Version
Doris	0.15,1.x

Use Limits

- Doris supports three data models: [DUPLICATE KEY|UNIQUE KEY|AGGREGATE KEY]. To write in Upsert mode in Doris, ensure the data model is UNIQUE KEY. For details, see Data Model - Apache Doris.
- If the source table has a primary key, the automatically created Doris target table will be UNIQUE KEY; if the source table has no primary key, the automatically created Doris target table will be DUPLICATE KEY.
- When the source end is Kafka, Doris only supports synchronization to existing databases and tables. Ensure the target databases and tables exist before the task runs.

Whole Database Write Node Configuration

Supported Data Sources

Currently supported source types for real-time whole database synchronization to the Doris target end:

- MySQL Whole Database Source Configuration Details
- TDSQL-C MySQL Whole Database Source Configuration Details
- TDSQL MySQL Whole Database Source Configuration Details
- Oracle Whole Database Source Configuration Details
- PostgreSQL Whole Database Source Configuration Details
- Kafka Whole Database Source Configuration Details

Data Target Settings



Parameter	Description
Data Source	Select the target data source for synchronization.

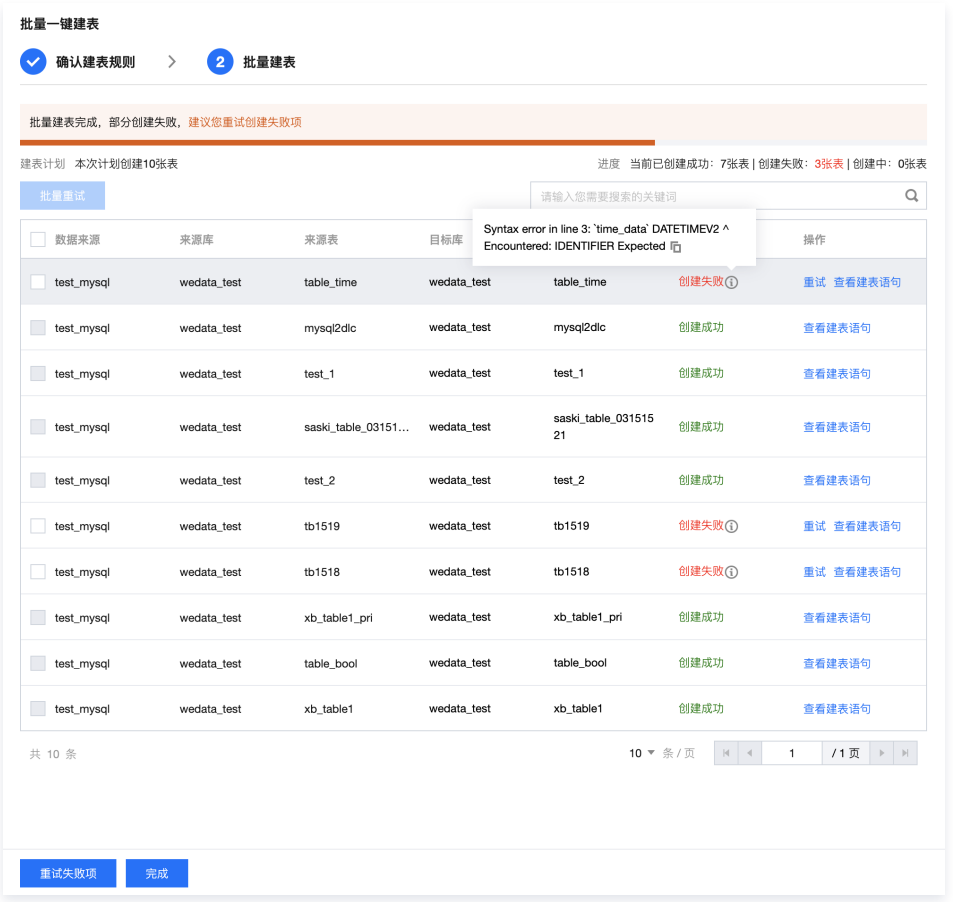
©2013-2025 Tencent Cloud. All rights reserved.

Database/Table Matching Strategy	Set the naming matching rules for database and table objects in Doris during task execution: • Same name as source database/table: The system will match objects with the same name as the source database/table in the target data source by default during task execution. • Custom Definition: The custom definition rule supports setting specific relationships between the source and the target, such as adding a uniform prefix or suffix to the source database or table name when writing to the target database or table during task execution. Under this strategy, the system will match the target objects according to the naming rules by default during task execution. **Note** When the source end is Kafka, the full synchronization phase of the whole database sync task to Doris does not support automatic database/table creation. Please pre-build the stock database and table objects on the Doris end to ensure the task runs properly.
Batch Table Creation	When the source end is MySQL, TDSQL-C MySQL, TDSQL MySQL, batch creation of Doris target tables is supported during the task configuration stage. Click **Batch Table Creation** to display the following pop-up. The system will match target databases and tables based on the user's set matching rules. Tables that do not exist on the target end will appear in **Unmatched Databases/Tables**, while existing tables will appear in **Matched Databases/Tables**. 

©2013-2025 Tencent Cloud. All rights reserved.



For unmatched tables or tables matched but need to be recreated (change the target table creation method to delete existing tables and create new ones), you can proceed with table creation. Click **Start Creating Tables** to create target tables based on the source end table structure. The creation results are displayed below. You can view the failed tables and reasons, modify the table creation statements, and retry the failed items.



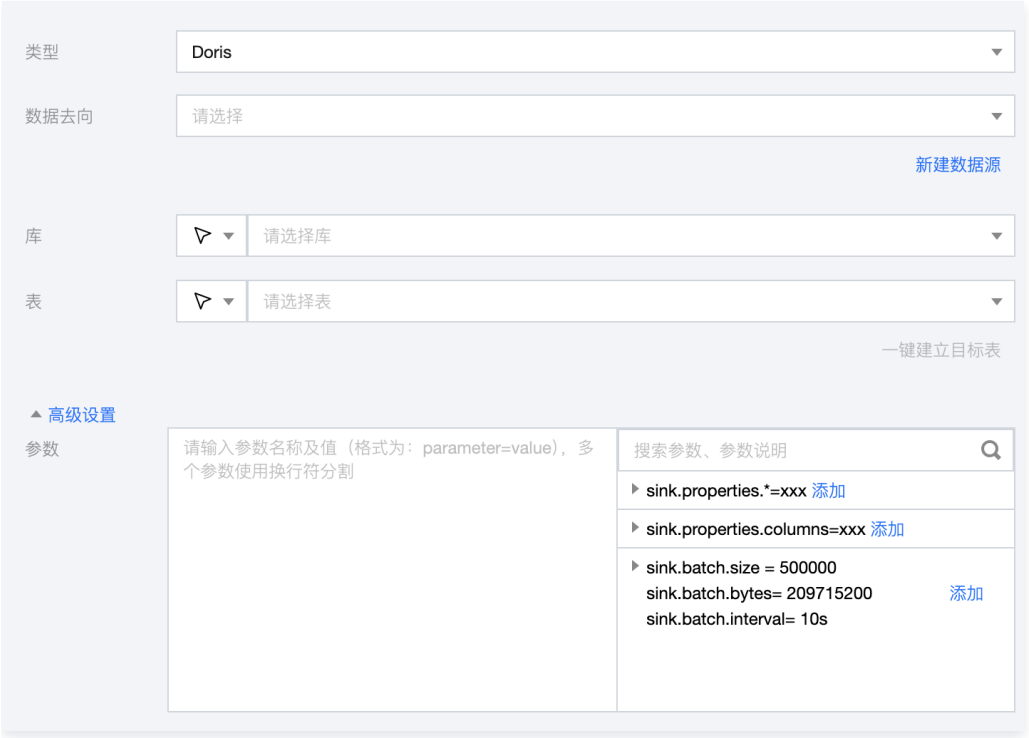
Advanced Settings	Set the runtime parameters for the Doris writer end. These parameters can be configured based on business requirements

Single Table Write Node Configuration

1. On the DataInLong page, click Real-time synchronization in the left-hand navigation bar.

©2013-2025 Tencent Cloud. All rights reserved.

2. On the Real-time synchronization page, select Single Table Synchronization to create a new configuration (you can choose between Form mode and Canvas mode) and enter the configuration page.



Parameter	Description
Data Flow Direction	The Doris data source to be written.
Database	Supports selecting or manually entering the name of the database to be written to. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Supports selecting or manually entering the name of the table to be written to. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Advanced Settings (Optional)	Parameters can be configured according to business needs. Requirements are as follows: 1. One parameter per line; parameters that need to be used together should be written on the same line. 2. Each parameter comes with a default value.

Log Collection Write Node Configuration

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Source	Select an available Doris data source in the current project.
Database/Table	Select the corresponding database table in this data source.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Data type conversion support for writing

The supported data types and conversion relationships for Doris writing are as follows:

Internal Type	Doris Type
NULL	NULL_TYPE
BOOLEAN	BOOLEAN
TINYINT	TINYINT
SMALLINT	SMALLINT
INT	INT
BIGINT	BIGINT
FLOAT	FLOAT
DOUBLE	DOUBLE
DATE	DATE
TIMESTAMP	DATETIME
DECIMAL	DECIMAL
STRING	CHAR
STRING	LARGEINT
STRING	VARCHAR
STRING	STRING

©2013–2025 Tencent Cloud. All rights reserved.

DECIMAL	DECIMALV2
ARRAY	ARRAY
MAP	MAP
STRING	JSON
STRING	VARIANT
STRING	IPV4
STRING	IPV6

FAQs

How to select and optimize Doris specifications?

Please refer to Doris Resource Specifications Model Selection and Optimization Suggestions.

Too many import tasks, new import task submission error "current running txns on db xxx is xx, larger than limit xx"?

Adjust the FE parameter: max_running_txn_num_per_db, default is 100, can be appropriately increased, recommended to control within 500.

Import frequency too fast resulting in err=[E-235] error?

- Parameter optimization suggestion: You can temporarily resolve it by appropriately increasing the max_tablet_version_num parameter. The default is 200, recommended to control within 2000.
- Business optimization suggestion: Reduce the import frequency to fundamentally solve this problem.

Import file too large and limited by parameters. Error "The size of this batch exceeds the max size"?

Adjust the BE parameter: streaming_load_max_mb, recommended to exceed the size of the file to be imported.

Import data error:"[-238]"?

- Reason: The -238 error usually occurs when the volume of data imported in the same batch is too large, resulting in too many Segment files in a single tablet.
- Parameter optimization suggestion: You can appropriately increase the BE parameter max_segment_num_per_rowset. The default value is 200, can be increased in multiples (such as 400, 800), recommended to control within 2000
- Business optimization suggestion: It is recommended to reduce the data volume of a single batch import.

Import failed, error:"too many filtered rows xxx," "ErrorURL":" or Insert has filtered data in strict mode, tracking url=xxxx."?

Reason: The schema, partition, etc. of the table does not match the imported data. You can execute doris commands in TCHouse-P Studio or the client to view the specific reason: show load warnings on `<tracking url>`, `<tracking url>` is the error URL returned in the error message.

©2013-2025 Tencent Cloud. All rights reserved.

StarRocks Data Source

Last updated：2024-09-06 14:58:46

DataInLong provides real-time writing capabilities for StarRocks. This article introduces the current support status for StarRocks real-time data synchronization.

Supported Versions

Currently, DataInLong supports database-level real-time writing to StarRocks. Using the real-time writing capability requires compliance with the following version limitations:

Type	Version
StarRocks	2.5.3,3.1.2,3.1.12

Use Limits

- DDL changes do not support modifying field names. If the source end modifies a field name, automatic modification of the field name on the StarRocks target end is not supported.
- Importing using CSV format is not supported.
- Advanced parameters are not supported.

Real-time whole database writing configuration

Supported Data Sources

Currently, the following source types of data are supported for real-time, database-level synchronization to the StarRocks target end:

- MySQL Whole Database Source Configuration Details
- TDSQL-C MySQL Whole Database Source Configuration Details
- Oracle Whole Database Source Configuration Details
- PostgreSQL Whole Database Source Configuration Details

Data Target Configuration



Parameter	Description
Data Flow Direction	Select the target data source for synchronization.

©2013-2025 Tencent Cloud. All rights reserved.

Database/Table Matching Strategy	Name matching rules for databases and data table objects in StarRocks: • Default to the same name as the source database/source table. • Self Definition: Supports generating target database and table names using built-in parameters and string combinations. **Note** Example: If the source table name is table1, and the mapping rule is ${table_name_di_src}_inlong, then the data from table1 will eventually be mapped and written into table1_inlong.

Data Type Conversion Supported

Flink data type	StarRocks data type
BOOLEAN	BOOLEAN
TINYINT	TINYINT
SMALLINT	SMALLINT
INTEGER	INTEGER
BIGINT	BIGINT
FLOAT	FLOAT
DOUBLE	DOUBLE
DECIMAL	DECIMAL
BINARY	INT
CHAR	STRING
VARCHAR	STRING
STRING	STRING
DATE	DATE
TIMESTAMP_WITHOUT_TIME_ZONE(N)	DATETIME
TIMESTAMP_WITH_LOCAL_TIME_ZONE(N)	DATETIME
ARRAY<T>	ARRAY<T>
MAP<KT,VT>	JSON STRING
ROW<arg T...>	JSON STRING

©2013–2025 Tencent Cloud. All rights reserved.

DLC Data Source

Last updated：2024-09-06 14:59:17

DataInLong provides real-time whole database and single table writing capabilities to DLC. This document introduces the current support status for real-time data synchronization using DLC.

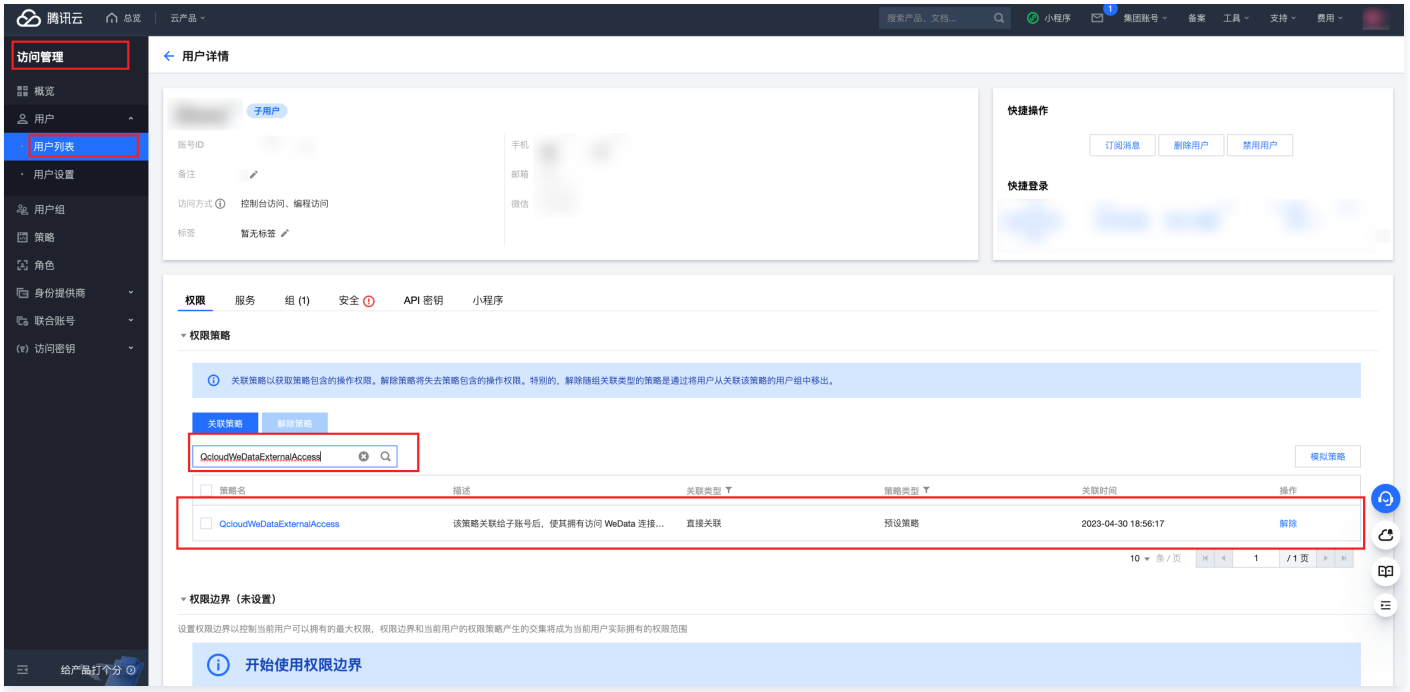
Use Limits

- Currently, only DLC Iceberg format tables are supported.
- When choosing Upsert mode synchronization, the DLC table must be a V2 table, with the setting write.upsert.enabled=true enabled.
- When selecting the Upsert mode for synchronization, a primary key must be specified. If it is a partitioned table, the partition field must also be added to the primary key.

Database Environment Preparation

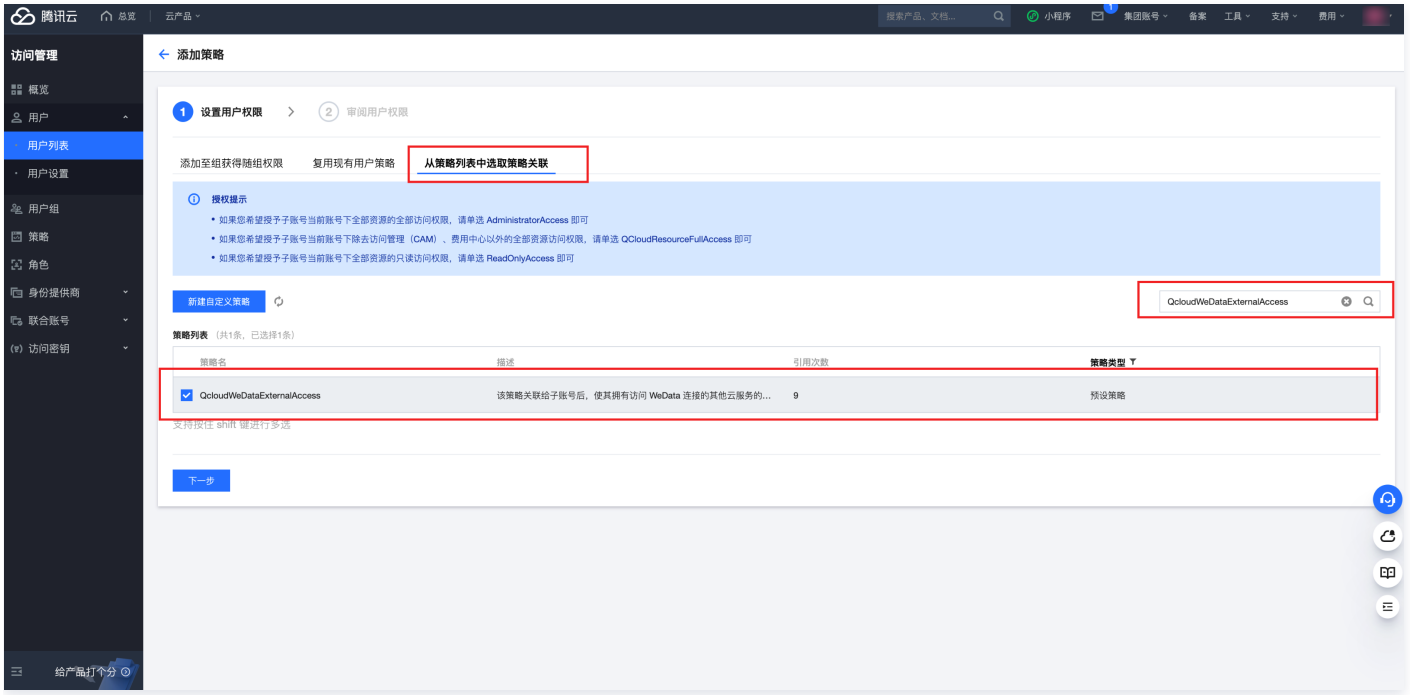
Check if WeData has access to DLC

1. Log in to the Tencent Cloud Console and go to the CAM (CAM) page.
2. In the left menu, select **Users** and find the user you need to bind the policy to.
3. Click the name of the user to enter the user's details page. In the Permissions section, search for **QcloudWeDataExternalAccess**. If the policy is already associated, you can skip the following steps.



4. In the user details page, find the **Permission Management** module under **Associate Policy**.
5. Click the **Associate Policy** button under **Create Association**. Select **Select policy from the policy list to associate** and enter **QcloudWeDataExternalAccess**.
6. Select **QcloudWeDataExternalAccess**, and click the **Next** button to complete the policy binding.

©2013-2025 Tencent Cloud. All rights reserved.



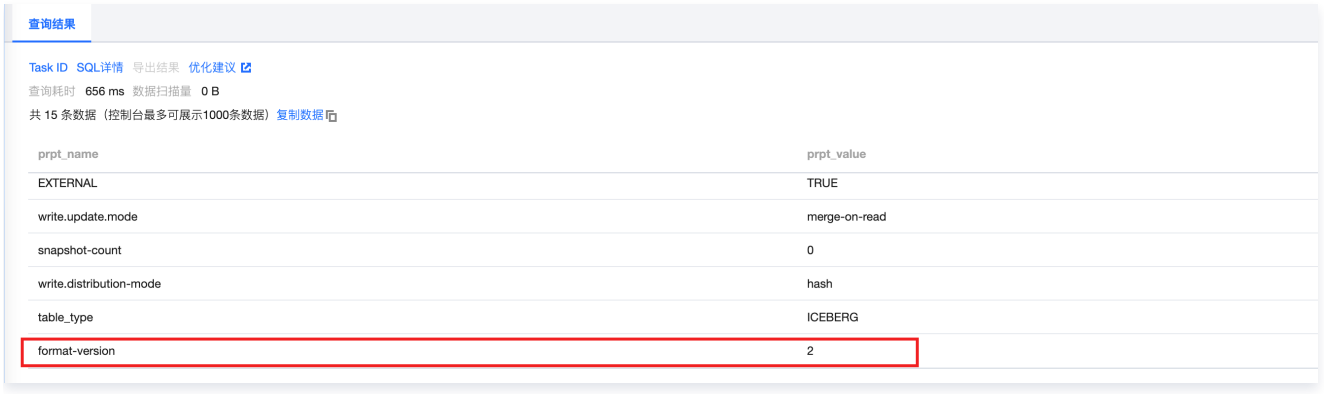
7. After a successful binding, the user can use the permissions defined in the **QcloudWeDataExternalAccess** policy.

Check whether the table is a V1 or V2 table (optional)

Real-time synchronization to DLC supports two modes: **Append Mode** and **Upsert Mode**. The Upsert mode must be for Iceberg V2 tables and requires the setting write.upsert.enabled=true to be enabled. Append mode supports both Iceberg V1 and V2 tables with the setting write.upsert.enabled=false.
You can use the following command to view the created table properties:

```
SHOW TBLPROPERTIES DataLakeCatalog.databases_name.table_name;
```

The returned result 'format-version = 2' indicates that a V2 table was created. If you created a V1 table and want to change it to a V2 table, refer to Modify Table Attributes.



Create a DLC Table (optional)

Real-time whole database tasks support automatically creating target tables that do not exist on the DLC during task execution, as well as using already existing DLC target tables. Users can choose to pre-create the target table in DLC, or allow the real-time whole database migration task to create the table automatically.

1. **Create V1 Table**

By default, DLC creates V1 tables, which do not support Upsert mode. Refer to the following for the create table statement:

```
CREATE TABLE IF NOT EXISTS DataLakeCatalog.dbname.test_v1 (id int, name string, ts date) PARTITIONED BY
```

©2013-2025 Tencent Cloud. All rights reserved.

```
(ts);
```

2. **Create V2 Table**

To use the Upsert mode to write to DLC, you need to create a V2 table. It must be specified when creating the table. Refer to the following for the create table statement:
When creating a V2 table, you must set 'write.upsert.enabled' = 'true' to ensure it is not in Append mode.

```
CREATE TABLE IF NOT EXISTS DataLakeCatalog.dbname.test_v2 (id int, name string, ts date) PARTITIONED BY
(ts) TBLPROPERTIES (
  'format-version' = '2', -- Create V2 Table
  'write.upsert.enabled' = 'true', -- Perform upsert operations when writing, supported only for V2 tables
  'write.distribution-mode' = 'hash', -- Determines the distribution of written data. Set to hash for
supporting multi-threaded writing
  'write.update.mode' = 'merge-on-read' -- Update mode for writing, performing merge operations during
write, supported only for V2 tables
)
```

3. **Modify Table Attributes (Optional)**

For already created tables, you need to modify their attributes. Refer to the following syntax:

```
SHOW TBLPROPERTIES table_name [('property_name')]
```

For more DLC SQL syntax, refer to SQL Syntax Overview.
Here is an example of converting an existing V1 Table to a V2 Table:

```
ALTER TABLE
  DataLakeCatalog.database_name.table_name
SET
  TBLPROPERTIES (
    'format-version' = '2',
    'write.upsert.enabled' = 'true'
  )
```

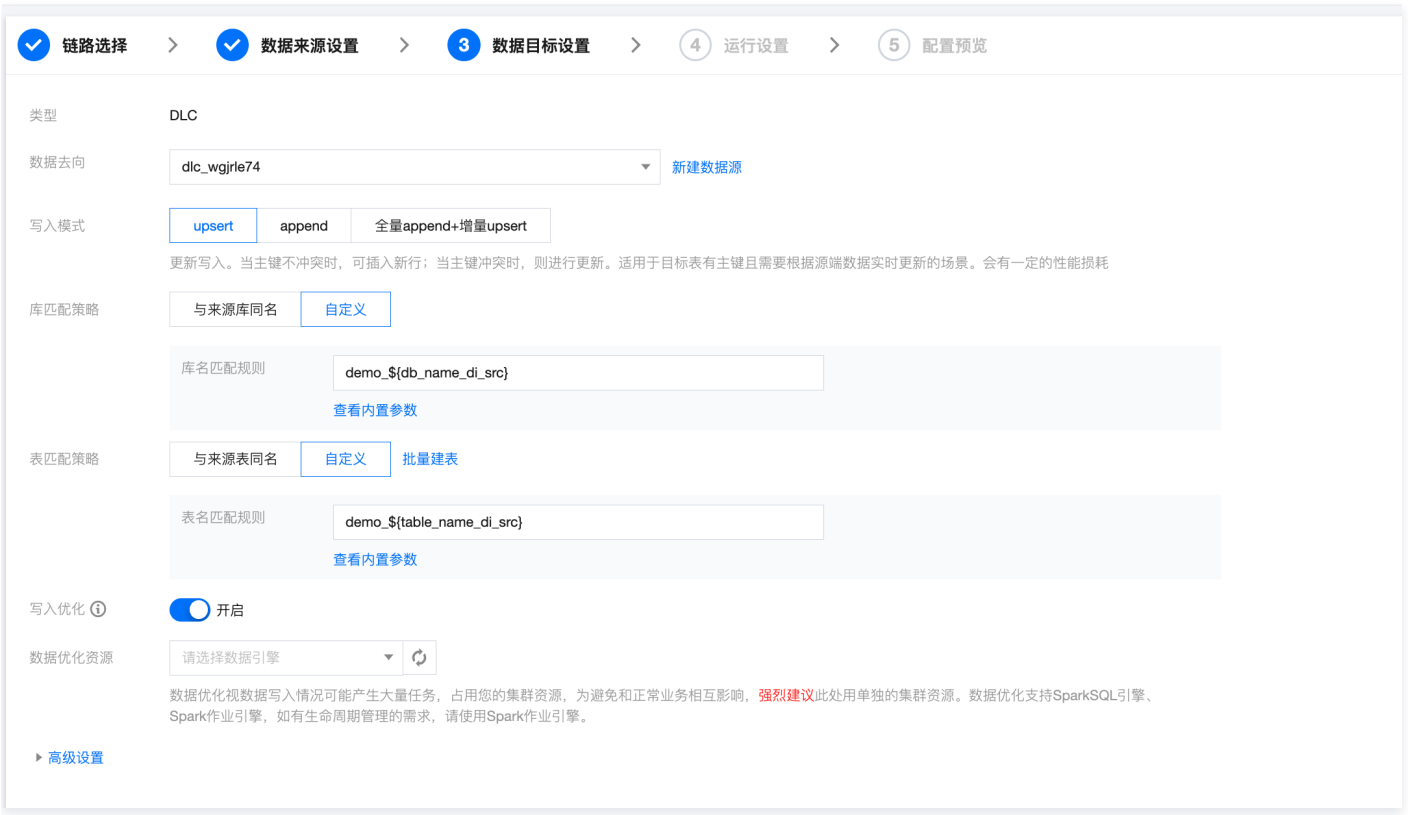
Whole Database Synchronization to DLC Configuration Details

Supported Data Sources

- MySQL Whole Database Source Configuration Details
- TDSQL-C MySQL Whole Database Source Configuration Details
- TDSQL MySQL Whole Database Source Configuration Details
- Oracle Whole Database Source Configuration Details
- PostgreSQL Full Database Source Configuration Details
- Kafka Whole Database Source Configuration Details
- Mongo Full Database Source Configuration Details
- OceanBase Full Database Source Configuration Details

Data Target Settings

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Flow Direction	Select the target data source for synchronization.
Write mode	• Upsert: Update write. If there is no primary key conflict, a new row is inserted; if there is a primary key conflict, the data is updated. Suitable for scenarios where the target table has a primary key and needs to be updated in real-time based on source data. There will be some performance overhead. • Append: Append write. Regardless of the presence of a primary key, data is appended by inserting new rows. Whether there is a primary key conflict depends on the target end. Suitable for scenarios without primary keys and allowing duplicate data. No performance loss. • Full Append + Incremental Upsert: Automatically switch data writing methods based on the synchronization phase of the source data. Append method is used during the full phase to improve performance and Upsert method is used during the incremental phase for real-time data updates. This mode currently only supports source data from MySQL, TDSQL-C MySQL, TDSQL MySQL, Oracle, and PostgreSQL.
Database/Table Matching Strategy	Name Matching Rules for database and data table objects in DLC.
Write Optimization	Write optimization is suitable for frequent or real-time write scenarios. Enabling write optimization will automatically merge files and delete expired snapshots. It is strongly recommended to enable write optimization in real-time Upsert write scenarios. For more details, refer to Data Optimization. **Note** Currently, integration tasks only support the default configuration using DLC write optimization. If users need to modify the corresponding configuration, please go to the DLC page to make changes.
Data Optimization Resources	Data optimization may generate a large number of tasks based on data write conditions, occupying your cluster resources. To avoid interference with normal business, it is strongly recommended to use separate cluster resources here. Data optimization supports **SuperSQL Engine** > **SparkSQL**, **SuperSQL Engine** > **Spark Job**. If you have lifecycle management needs, please use **SuperSQL Engine** > **Spark Job** engine.

©2013-2025 Tencent Cloud. All rights reserved.

Advanced Settings	You can configure parameters based on business requirements.

Single Table Write Node Configuration

1. In the DataInLong page, click **Real-time Synchronization** in the left sidebar.
2. On the Real-time Synchronization page, select **Single Table Synchronization** at the top, create a new one (choose Form or Canvas mode), and enter the configuration page.



3. You can refer to the table below for parameter configuration.

Parameter	Description
Data Flow Direction	DLC data source to be written to.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Supports selecting or manually entering the name of the table to be written to. When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Write mode	DLC real-time synchronization write supports two or three modes: • Upsert: Update write. If there is no primary key conflict, a new row is inserted; if there is a primary key conflict, the data is updated. Suitable for scenarios where the target table has a primary key and needs to be updated in real-time based on source data. There will be some performance overhead. • Append: Append write. Regardless of the presence of a primary key, data is appended by inserting new rows. Whether there is a primary key conflict depends on the target end. Suitable for scenarios without primary keys and allowing duplicate data. No performance loss. • Full Append + Incremental Upsert: Automatically switch data writing methods based on the synchronization phase of the source data. Append method is used during the full phase to improve

©2013-2025 Tencent Cloud. All rights reserved.

	performance and Upsert method is used during the incremental phase for real-time data updates. This mode currently only supports source data from MySQL, TDSQL-C MySQL, TDSQL MySQL, Oracle, and PostgreSQL.
Unique key	In Upsert write mode, a unique key must be set to ensure data order. Multiple selections are supported.
Write Optimization	Write optimization is suitable for frequent or real-time write scenarios. Enabling write optimization will automatically merge files and delete expired snapshots. It is strongly recommended to enable write optimization in real-time Upsert write scenarios. For more details, refer to Data Optimization. **Note** Currently, integration tasks only support the default configuration using DLC write optimization. If users need to modify the corresponding configuration, please go to the DLC page to make changes.
Data Optimization Resources	Data optimization may generate a large number of tasks based on data writing conditions, occupying your cluster resources. To avoid mutual interference with normal business operations, it is strongly recommended to use separate cluster resources here. Data optimization supports **SuperSQL Engine** > **SparkSQL**, **SuperSQL Engine** > **Spark Job**. If you need lifecycle management, please use the **SuperSQL Engine** > **Spark Job** engine.
Advanced Settings	You can configure parameters based on business requirements.

Data Type Conversion Supported

Internal Type	Iceberg Type
CHAR	STRING
VARCHAR	STRING
STRING	STRING
BOOLEAN	BOOLEAN
BINARY	FIXED(L)
VARBINARY	BINARY
DECIMAL	DECIMAL(P,S)
TINYINT	INT
SMALLINT	INT
INTEGER	INT
BIGINT	LONG
FLOAT	FLOAT
DOUBLE	DOUBLE
DATE	DATE
TIME	TIME
TIMESTAMP	TIMESTAMP
TIMESTAMP_LTZ	TIMESTAMPTZ
INTERVAL	-
ARRAY	LIST

©2013-2025 Tencent Cloud. All rights reserved.

MULTISET	MAP
MAP	MAP
ROW	STRUCT
RAW	–

FAQs

Error occurred when synchronizing incremental data (changelog data) to DLC

Error message:

```
2022-07-12 21:03:57.161 [IcebergFilesCommitter -> Sink: IcebergSink HYBRIS.xny_sl03_pro.subs_crm_clue (1/1)#34] WARN  org.apache.flink.runtime.taskmanager.Task [] -
IcebergFilesCommitter -> Sink: IcebergSink HYBRIS.xny_sl03_pro.subs_crm_clue (1/1)#34 (ced[redacted]b5) switched from INITIALIZING to FAILED with failure cause:
java.lang.IllegalArgumentException: Cannot write delete files in a v1 table
    at org.apache.iceberg.ManifestFiles.writeDeleteManifest(ManifestFiles.java:154)
    at org.apache.iceberg.SnapshotProducer.newDeleteManifestWriter(SnapshotProducer.java:374)
    at org.apache.iceberg.MergingSnapshotProducer.lambda$newDeleteFilesAsManifests$10(MergingSnapshotProducer.java:681)
    at java.util.HashMap.forEach(HashMap.java:1289)
    at org.apache.iceberg.MergingSnapshotProducer.newDeleteFilesAsManifests(MergingSnapshotProducer.java:678)
    at org.apache.iceberg.MergingSnapshotProducer.prepareDeleteManifests(MergingSnapshotProducer.java:664)
    at org.apache.iceberg.MergingSnapshotProducer.apply(MergingSnapshotProducer.java:533)
    at org.apache.iceberg.SnapshotProducer.apply(SnapshotProducer.java:164)
    at org.apache.iceberg.SnapshotProducer.lambda$commit$2(SnapshotProducer.java:283)
    at org.apache.iceberg.util.Tasks$Builder.runTaskWithRetry(Tasks.java:404)
    at org.apache.iceberg.util.Tasks$Builder.runSingleThreaded(Tasks.java:214)
    at org.apache.iceberg.util.Tasks$Builder.run(Tasks.java:198)
    at org.apache.iceberg.util.Tasks$Builder.run(Tasks.java:190)
    at org.apache.iceberg.SnapshotProducer.commit(SnapshotProducer.java:282)
    at org.apache.iceberg.flink.sink.IcebergFilesCommitter.commitOperation(IcebergFilesCommitter.java:312)
    at org.apache.iceberg.flink.sink.IcebergFilesCommitter.commitDeltaTxn(IcebergFilesCommitter.java:299)
    at org.apache.iceberg.flink.sink.IcebergFilesCommitter.commitUpToCheckpoint(IcebergFilesCommitter.java:218)
    at org.apache.iceberg.flink.sink.IcebergFilesCommitter.initializeState(IcebergFilesCommitter.java:153)
    at org.apache.flink.streaming.api.operators.StreamOperatorStateHandler.initializeOperatorState(StreamOperatorStateHandler.java:118)
    at org.apache.flink.streaming.api.operators.AbstractStreamOperator.initializeState(AbstractStreamOperator.java:290)
    at org.apache.flink.streaming.runtime.tasks.OperatorChain.initializeStateAndOpenOperators(OperatorChain.java:441)
    at org.apache.flink.streaming.runtime.tasks.StreamTask.restoreGates(StreamTask.java:585)
    at org.apache.flink.streaming.runtime.tasks.StreamTaskActionExecutor$1.call(StreamTaskActionExecutor.java:55)
```

Cause:
DLC tables that are V1 tables do not support changelog data.
Solution:
1. Modify DLC Table to V2 Table and enable Upsert support.
2. Enable Upsert for the table: ALTER TABLE tblname SET TBLPROPERTIES ('write.upsert.enabled'='true').
3. Alter table to V2 Table: ALTER TABLE tblname SET TBLPROPERTIES ('format-version'='2').

Check if the table properties are set successfully: SHOW TBLPROPERTIES tblname.

Cannot write incompatible dataset to table with schema

Error Details:

```
Caused by: java.lang.IllegalArgumentException: Cannot write incompatible dataset to table with schema:
* mobile should be required, but is optional
    at org.apache.iceberg.types.TypeUtil.checkSchemaCompatibility(TypeUtil.java:364)
    at org.apache.iceberg.types.TypeUtil.validateWriteSchema(TypeUtil.java:323)
```

Error Causes:
When users create a DLC Table, the fields are set with NOT NULL Constraint.
Resolution:
Do not set NOT NULL Constraint when creating the table.

Sync mysql2dlc, reports array index out of bounds

Problem Details:

```
java.lang.ArrayIndexOutOfBoundsException: 1
    at
org.apache.flink.table.data.binary.BinarySegmentUtils.getLongMultiSegments(BinarySegmentUtils.java:736) ~
[flink-table-blink_2.11-1.13.6.jar:1.13.6]
```

©2013–2025 Tencent Cloud. All rights reserved.

```
    at org.apache.flink.table.data.binary.BinarySegmentUtils.getLong(BinarySegmentUtils.java:726) ~[flink-
table-blink_2.11-1.13.6.jar:1.13.6]
    at
org.apache.flink.table.data.binary.BinarySegmentUtils.readTimestampData(BinarySegmentUtils.java:1022) ~
[flink-table-blink_2.11-1.13.6.jar:1.13.6]
    at org.apache.flink.table.data.binary.BinaryRowData.getTimestamp(BinaryRowData.java:356) ~[flink-
table-blink_2.11-1.13.6.jar:1.13.6]
    at org.apache.flink.table.data.RowData.lambda$createFieldGetter$39385f9c$1(RowData.java:260) ~[flink-
table-blink_2.11-1.13.6.jar:1.13.6]
```

Error Causes:

The issue caused by using the time field as the primary key.

Solution:

1. Do not use the time field as the primary key.
2. Users still need a timestamp to ensure uniqueness. It is recommended that users add a redundant string field in DLC and map the upstream timestamp to it using a function.

DLC Task reports that it is not an Iceberg Table

Error message:

```
[flink-dist_2.11-1.13.6.jar:1.13.6]
    ... 10 more
Caused by: org.apache.iceberg.exceptions.NoSuchIcebergTableException: Not an iceberg table: HYBRIS.origindb.
movieunifiedorder01_unified_order_sharding__maoyan_order_v2 (type=null)
    at org.apache.iceberg.exceptions.NoSuchIcebergTableException.check(NoSuchIcebergTableException.java:33) ~
[pool-1-thread-1-1678868648050-prod-sort-connector-iceberg-dlc-2.6.29-1-1.4.0.jar:1.4.0]
```

Error Causes:

1. Users can execute a statement in DLC to check the exact table type.
2. Statement: desc formatted table name.
3. You can check the table type.

Resolution:

Select the correct engine to create an Iceberg Table.

Flink SQL field order does not match the DLC target table fields, causing an error

Solution:

Do not modify the field order of the task table.

©2013–2025 Tencent Cloud. All rights reserved.

Iceberg Data Source

Last updated: 2024-09-06 14:59:48

DataInLong provides real-time writing capabilities for Iceberg. This document introduces the current capability support for using Iceberg for real-time data synchronization.

Supported Versions

Currently, DataInLong supports real-time writing for single tables and whole databases with Iceberg. The following version limitations must be adhered to when using real-time synchronization capabilities:

Type	Version
Iceberg	0.13.1+

Use Limits

- Upsert writing only supports Iceberg V2 tables.
- After writing to Iceberg, to improve query performance, it is generally required that the downstream performs small file merging through timed Spark Action and reasonably configures the Checkpoint interval.
- Iceberg column type changes only support changing int to long, float to double, and increasing the precision of the Decimal type when the scale is the same.

Whole Database Write Configuration

Supported Data Sources

Currently, the following source types support real-time whole database synchronization to the Iceberg target end:

- MySQL Whole Database Source Configuration Details
- TDSQL-C MySQL Whole Database Source Configuration Details
- TDSQL MySQL Whole Database Source Configuration Details
- Oracle Whole Database Source Configuration Details
- PostgreSQL Whole Database Source Configuration Details
- SQL Server Whole Database Source Configuration Details
- Kafka Whole Database Source Configuration Details
- OceanBase Whole Database Source Configuration Details

Data Target Settings

©2013-2025 Tencent Cloud. All rights reserved.

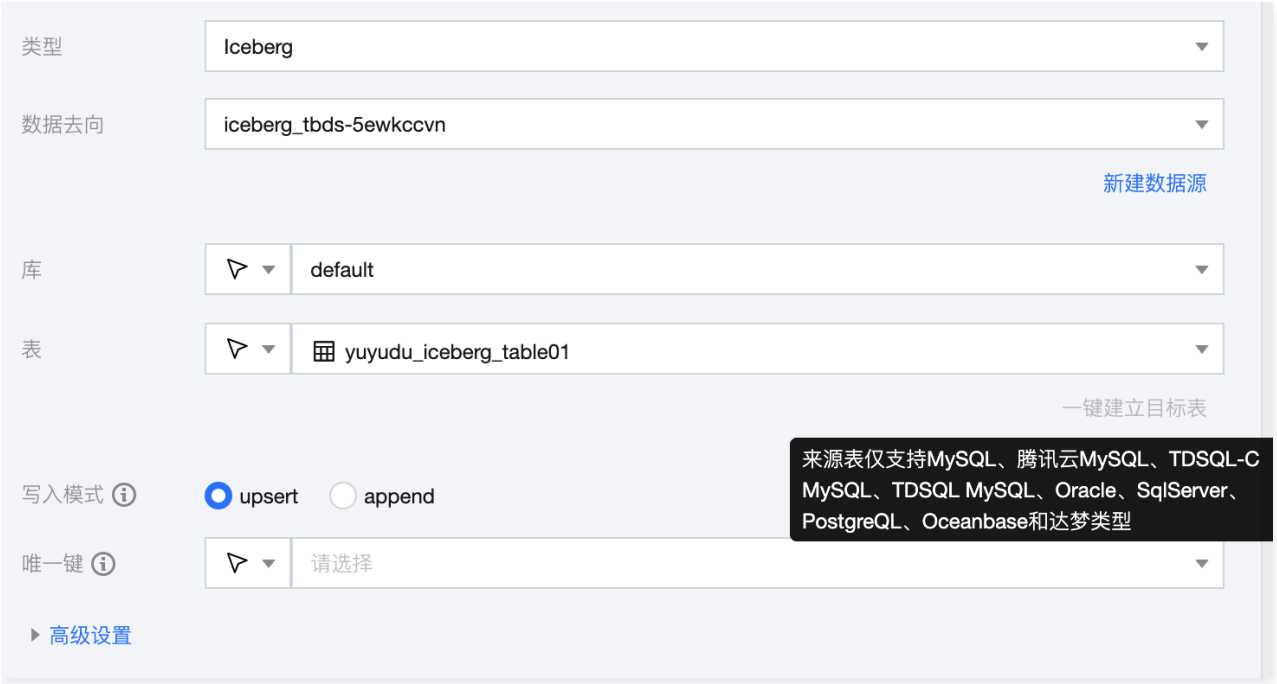


Parameter	Description
Data Flow Direction	Select the target data source for synchronization.
Write mode	• Upsert: Update writing. When there is no primary key conflict, a new row can be inserted; when there is a primary key conflict, the row is updated. This is suitable for scenarios where the target table has a primary key and needs to be updated in real-time based on source data. There will be some performance overhead. Upsert writing mode only supports Iceberg V2 tables and must have a unique key. • Append: Append write. Regardless of the presence of a primary key, data is appended by inserting new rows. Whether there is a primary key conflict depends on the target end. Suitable for scenarios without primary keys and allowing duplicate data. No performance loss. • Full Append + Incremental Upsert: Automatically switches the data writing method based on the data synchronization phase of the source. The full phase uses Append writing to enhance performance, while the incremental phase uses Upsert writing for real-time data updates
Library and Table Matching Strategy	Name matching rules for databases and data table objects in Iceberg: • Default to the same name as the source database/source table. • Self Definition: Supports generating target database and table names using built-in parameters and string combinations. **Note** Example: If the source table name is table1, and the mapping rule is ${table_name_di_src}_inlong, then the data from table1 will eventually be mapped and written into table1_inlong. The system will match the target database/table according to the matching rules: • If the matching database/table does not exist at the Iceberg target end, it will automatically create the database/table. • If the matching database/table already exists at the Iceberg target end, it will not automatically create the database/table, and will use the existing database/table by default.
Advanced Settings	You can configure parameters based on business requirements.

Single Table Write Node Configuration

©2013-2025 Tencent Cloud. All rights reserved.

1. On the DataInLong page, click Real-time synchronization in the left-hand navigation bar.
2. On the Real-time synchronization page, select Single Table Synchronization to create a new configuration (you can choose between Form mode and Canvas mode) and enter the configuration page.



Parameter	Description
Data Flow Direction	Iceberg data source to be written.
Database	Supports selecting or manually entering the name of the database to be written to. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Supports selecting or manually entering the name of the table to be written to. When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
One-click create target table	When the source is MySQL, TDSQL-C MySQL, TDSQL MySQL, Oracle, PostgreSQL, Oceanbase, or Dameng, it supports one-click creation of Iceberg target tables based on the source table structure.
Write mode	• Upsert: Insert or update. When there is no primary key conflict, a new row can be inserted; when there is a primary key conflict, it will be updated. Suitable for scenarios where the target table has a primary key and needs to be updated in real time based on source data. There will be some performance overhead. Only supports Iceberg V2 tables and must have a unique key. • Append: Append write. Regardless of the presence of a primary key, data is appended by inserting new rows. Whether there is a primary key conflict depends on the target end. Suitable for scenarios without primary keys and allowing duplicate data. No performance loss.
Unique key	In Upsert write mode, a unique key must be set to ensure data order. Multiple selections are supported.
Advanced Settings	You can configure parameters based on business requirements.

Log Collection Write Node Configuration

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Source	Select the available Iceberg data source in the current project.
Database/Table	Select the corresponding database table in this data source.
Write mode	Iceberg supports two write modes: • Append: Append data. • Upsert: Insert messages in Upsert mode. After setting, messages can only be processed once by the consumer to ensure Exactly-Once.
Unique key	In Upsert mode, a unique key must be set to ensure data order, and multiple selections are supported. The Append mode does not require setting up a unique key.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Data Type Conversion Supported

Internal Type	Iceberg Type
CHAR	STRING
VARCHAR	STRING
STRING	STRING
BOOLEAN	BOOLEAN
BINARY	FIXED(L)
VARBINARY	BINARY
DECIMAL	DECIMAL(P,S)
TINYINT	INT
SMALLINT	INT
INTEGER	INT
BIGINT	LONG

©2013-2025 Tencent Cloud. All rights reserved.

FLOAT	FLOAT
DOUBLE	DOUBLE
DATE	DATE
TIME	TIME
TIMESTAMP	TIMESTAMP
TIMESTAMP_LTZ	TIMESTAMPTZ
INTERVAL	-
ARRAY	LIST
MULTISET	MAP
MAP	MAP
ROW	STRUCT
RAW	-

FAQs

Submission failed due to excessive table field length

```
org.apache.iceberg.exceptions.CommitStateUnknownException:
MetaException(message:Put request failed : UPDATE `TABLE_PARAMS`
SET `PARAM_VALUE` = ? WHERE `TBL_ID`=? AND `PARAM_KEY`=?)
Cannot determine whether the commit was successful or not, the
underlying data files may or may not be needed. Manual
intervention via the Remove Orphan Files Action can remove these
files when a connection to the Catalog can be re-established if
the commit was actually unsuccessful.
Please check to see whether or not your commit was successful
before retrying this commit. Retrying an already successful
operation will result in duplicate records or unintentional
modifications.
At this time no files will be deleted including possibly unused
manifest lists.
    at
org.apache.iceberg.hive.HiveTableOperations.doCommit(HiveTableOper
ations.java:329)
    at
org.apache.iceberg.BaseMetastoreTableOperations.commit(BaseMetasto
reTableOperations.java:133)
```

Solution:

1. First, backup the TABLE_PARAMS table:
   mysqldump -hxxx -uroot -pxxx hivemetastore TABLE_PARAMS > table_params.sql
2. Change the length to 40000:
   alter table TABLE_PARAMS MODIFY PARAM_VALUE VARCHAR(40000);

> **Note**
> UTF-8 format does not support a length of 40000, you can change it to text type or reduce it to 20000.

©2013-2025 Tencent Cloud. All rights reserved.

TCHouse-P Data Source

Last updated: 2024-09-06 15:00:15

©2013-2025 Tencent Cloud. All rights reserved.

DataInLong provides the writing capability for TCHouse-P. This document introduces the current capabilities supported for real-time data synchronization using TCHouse-P.

Supported Versions

Currently, DataInLong supports real-time writing of single tables and entire databases to TCHouse-P. To use real-time synchronization capabilities, the following version limitations must be observed:

Type	Version
TCHouse-P	3.16.X

Use Limits

It is recommended to write to row-store tables instead of columnar storage tables (which may cause data inflation).

Whole Database Write Node Configuration

Supported Data Sources

Currently, the following source types are supported for real-time whole-database synchronization to the TCHouse-P target end:

- MySQL Whole Database Source Configuration Details
- TDSQL-C MySQL Whole Database Source Configuration Details
- Oracle Whole Database Source Configuration Details
- PostgreSQL Whole Database Source Configuration Details
- SQL Server Whole Database Source Configuration Details

Data Target Configuration



Parameter	Description
Data Flow Direction	Select the target data source for synchronization.
Write mode	• Append: Append write. Regardless of the presence of a primary key, data is appended by inserting new rows. Whether there is a primary key conflict depends on the target end. Suitable for scenarios without primary keys and allowing duplicate data. No performance loss.

©2013-2025 Tencent Cloud. All rights reserved.

	• Upsert: Update write. If there is no primary key conflict, a new row is inserted; if there is a primary key conflict, the data is updated. Suitable for scenarios where the target table has a primary key and needs to be updated in real-time based on source data. There will be some performance overhead.
Database/Schema/Table Matching Strategy	• With the same name as the source database/table: By default, in synchronization tasks, the source database and data table will be written into the destination database/table with the same name. Under this strategy, when the task runs, the system will default to match objects with the same name in the target data source. • Custom Definition: The Custom Definition rule supports setting special relationships between the source and target, such as uniformly adding a fixed prefix or suffix to the source database or table name when writing to the target database or table. Under this strategy, when the task runs, the system will default to match the target object according to the naming rule.
Advanced Settings	You can configure parameters based on business requirements.

Single Table Write Node Configuration



Parameter	Description
Data Flow Direction	The TCHouse-P data source to be written to.
Database	Supports selecting or manually entering the name of the database to be written to. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Supports selecting or manually entering the name of the Schema to be written to.
Table	Supports selecting or manually entering the name of the table to be written to. When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
One-click create target table	When the source end is MySQL, TDSQL-C MySQL, TDSQL MySQL, Oracle, PostgreSQL, Oceanbase, or Dameng, it supports quickly creating TCHouse-P target tables through the source table structure
Primary key	Supports selecting or manually entering the primary key name. Upsert writes based on the primary key.

©2013-2025 Tencent Cloud. All rights reserved.

Advanced Settings (optional)	You can configure parameters based on business requirements.

Data Type Conversion Supported

The following data types and conversion relationships are supported for TCHouse-P writing:

Internal Type	TDSQL for PostgreSQL type
TINYINT	-
SMALLINT	SMALLINT,INT2,SMALLSERIAL,SERIAL2
INT	INTEGER,SERIAL
BIGINT	BIGINT,BGSERIAL
FLOAT	REAL,FLOAT4
DOUBLE	FLOAT8,DOUBLE PRECISION
DECIMAL(p, s)	NUMERIC(p, s),DECIMAL(p, s)
BOOLEAN	BOOLEAN
DATE	DATE
TIME [(p)][WITHOUT TIMEZONE]	TIME [(p)][WITHOUT TIMEZONE]
TIMESTAMP [(p)][WITHOUT TIMEZONE]	TIMESTAMP [(p)]WITHOUT TIMEZONE
BYTES,VARYING(n)	BYTEA
CHAR(n)	CHAR(n),CHARACTER(n)
VARCHAR(n)	VARCHAR(n)
STRING	TEXT,POINT,MONEY,JSON,UUID,XML

©2013-2025 Tencent Cloud. All rights reserved.

ElasticSearch Data Source

Last updated：2024-09-06 15:05:49

DataInLong provides real-time writing capability for Elasticsearch. This article introduces the current support status of Elasticsearch real-time data synchronization.

Supported Versions

Currently, DataInLong supports real-time whole database writing to Elasticsearch at the database level. The following version restrictions apply to real-time writing capability:

Type	Version
Elasticsearch	5.x, 6.x, 7.x

Use Limits

- Currently, only a real-time whole database link from **TDSQL-C MySQL** to **Elasticsearch** is supported.
- The primary key values in Elasticsearch are generated as _id by default and cannot be modified.
- Currently, only the document's _id is supported as the partition column.
- Elasticsearch currently uses dynamic field mapping to create tables. The DDL change strategy does not support automatic modification of column types, deletion of columns, or renaming columns (which will result in the addition of new columns).
- Elasticsearch with Kerberos Authentication enabled is not supported.

Real-time whole database writing configuration

Supported Data Sources

Currently, the following source types of data are supported for real-time whole database synchronization to the Elasticsearch target:
TDSQL-C MySQL Whole Database Source Configuration Details

Data Target Configuration

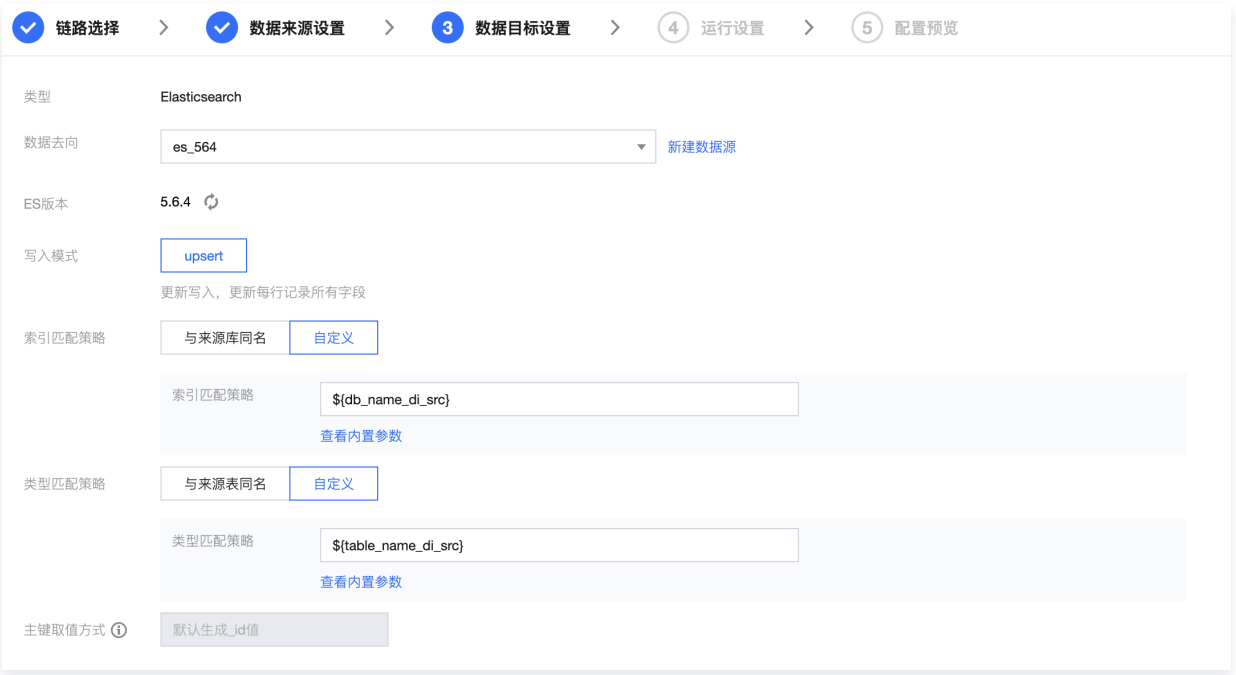
There are some configuration differences between Elasticsearch version 6.x and above (inclusive) and the following versions:

- Configuration details for versions 6.x and above (inclusive):



- Configuration details for versions below 6.x:

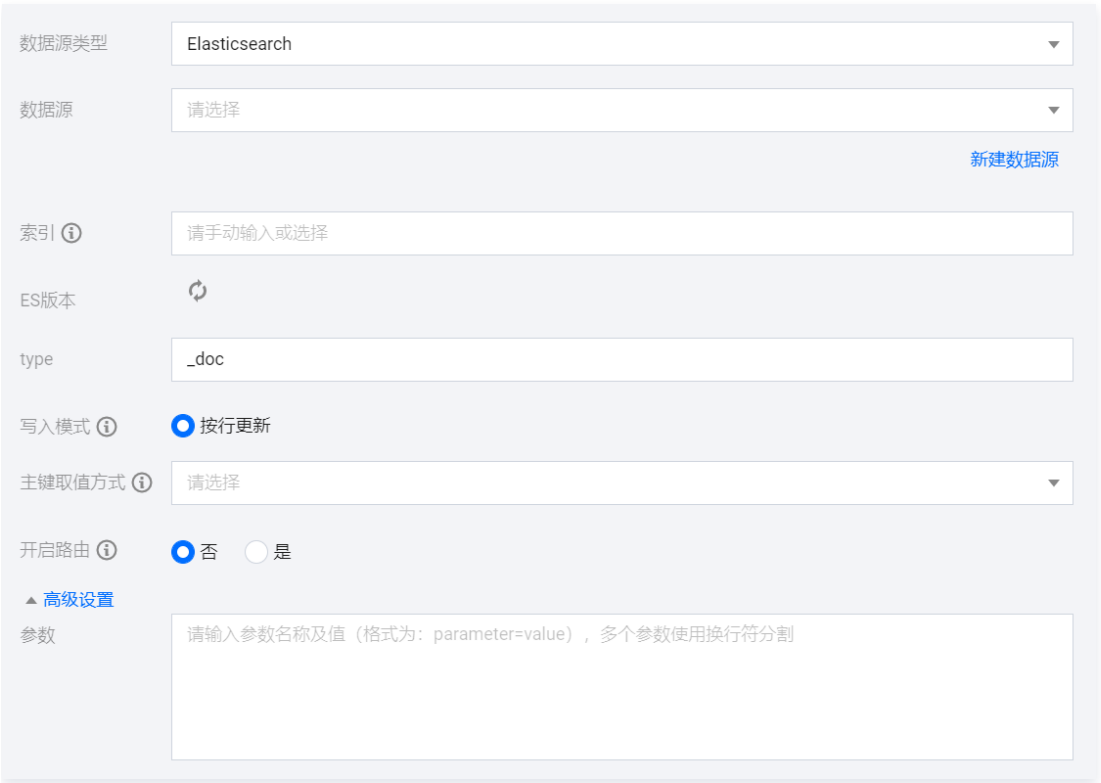
©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Flow Direction	Select the target data source for synchronization.
ES Version	The corresponding version is automatically displayed based on the selected data source
Write mode	Upsert: Update and write, updating all fields of each record (this is the only mode currently supported).
Index/Type matching strategy	Index and Type Naming Rules in ES: • For versions 6.x and above (inclusive), the index defaults to the same name as the source table; the type defaults to _doc and cannot be modified • For versions below 6.x, the index defaults to the same name as the source database; the type defaults to the same name as the source table • Custom Definition: Supports using a combination of built-in parameters and strings to generate target index/type names. **Note** Example: If the source table name is table1, and the mapping rule is ${table_name_di_src}_inlong, then the data from table1 will eventually be mapped and written into table1_inlong.
Gradual Value Retrieval Method	Currently, only default generation of _id values is supported

Log Collection Write Node Configuration

©2013–2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Source	Select an available Elasticsearch data source in the current project.
Index	Index Name in the Elasticsearch data source.
type	Automatically identify indices. For Elasticsearch 7.X, the default type is _doc.
Write mode	Elasticsearch only supports row-wise updates, updating all fields of each record.
Primary key value method	Three value methods are supported: • Source Table Primary Key: The document's id uses the primary key of the source table. • Composite Primary Key: The document's id is determined by multiple columns of the source table. • No Primary Key: The _id value is generated by default.
Enable routing	Enable routing partition index data in Elasticsearch. After enabling the routing feature, you can control which partition to use for storing documents in Elasticsearch.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Data Type Conversion Supported

Internal Type	JSON Types
CHAR / VARCHAR / STRING	string
BOOLEAN	boolean
BINARY / VARBINARY	string with encoding: base64
DECIMAL	number
TINYINT	number
SMALLINT	number

©2013–2025 Tencent Cloud. All rights reserved.

INT	number
BIGINT	number
FLOAT	number
DOUBLE	number
DATE	string with format: date
TIME	string with format: time
TIMESTAMP	string with format: date-time
TIMESTAMP_WITH_LOCAL_TIME_ZONE	string with format: date-time (with UTC time zone)
INTERVAL	number
ARRAY	array
MAP / MULTISET	object
ROW	object

©2013-2025 Tencent Cloud. All rights reserved.

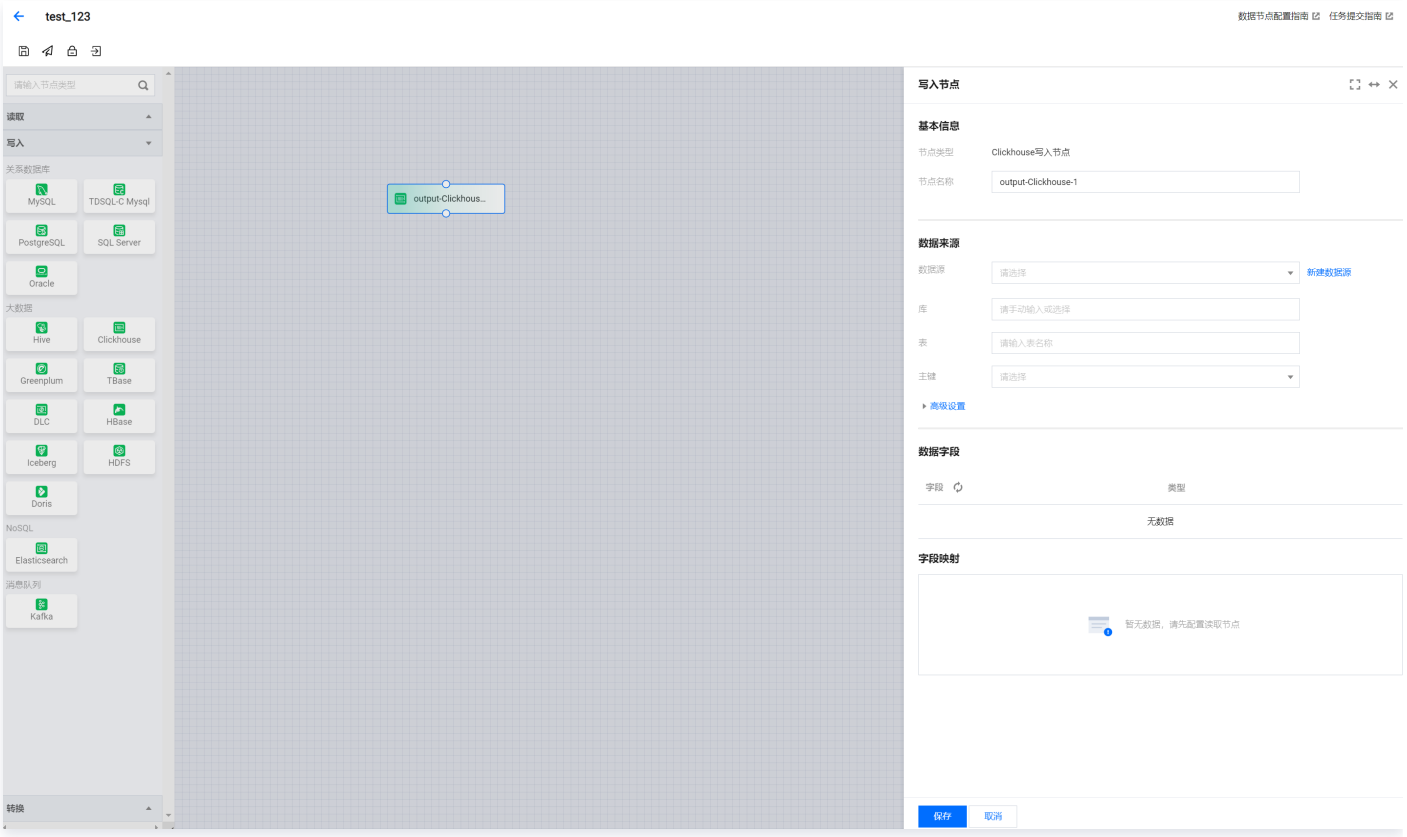
ClickHouse Data Source

Last updated：2024-09-06 15:06:16

ClickHouse Single Table Write Node Configuration

Configure ClickHouse Node

1. In the DataInLong page, click **Real-time Synchronization** in the left sidebar.
2. At the top of the Real-time Sync page, select **Single Table Sync**. Create (choose between Form and Canvas Mode) and enter the configuration page.
3. Click the left-side**write**, click to select the ClickHouse node and configure the node information.



4. You can refer to the table below for parameter configuration.

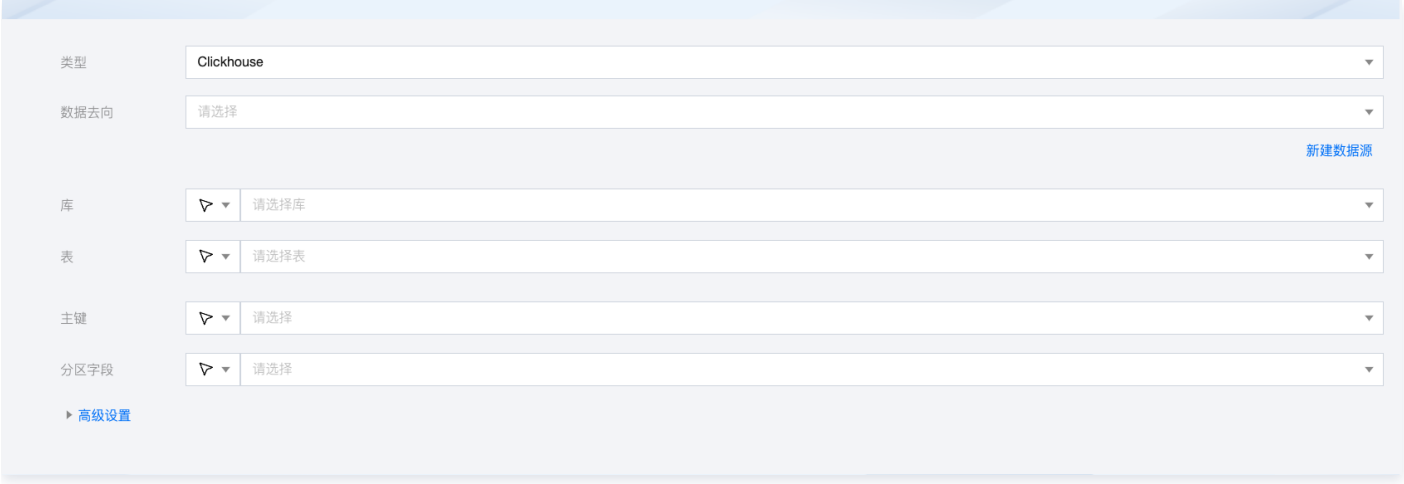
Parameter	Description
Node name	Enter the ClickHouse node name.
Data Source	Select the ClickHouse data source to write to.
Database	Supports selecting or manually entering the name of the database to be written to. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • If the data source network is not connected, making it impossible to directly fetch the database information, you can manually enter the database name. • Data synchronization can still be performed given network connection in DataInLong.
Table	Supports selecting or manually entering the name of the table to be written to. • If the data source network is not connected, making it impossible to directly fetch the table information, you can manually enter the table name. • Data synchronization can still be performed given network connection in DataInLong.
Primary key	Select a field as the primary key for the write table.

©2013-2025 Tencent Cloud. All rights reserved.

Partition Field	Select one or more fields as partition fields.
Advanced Settings	You can configure parameters based on business requirements.

5. Preview fields and map them to the write node. Click **Save**.

ClickHouse Log Collection and Writing Node Configuration



Parameter	Description
Data Flow Direction	Select an available ClickHouse data source in the current project.
Database/Table	Select the corresponding database table in this data source.
Primary key	Select a field as the primary key for the data table.
Partition Field	Select one or more fields as partition fields.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Data type conversion is supported

Write

Internal Type	ClickHouse Type
CHAR	String
VARCHAR	String,IP,UUID
STRING	String,EnumL
BOOLEAN	UInt8
BYTES	FixedString
DECIMAL	Decimal,Int128,Int256,UInt64,UInt128,UInt256
TINYINT	Int8
SMALLINT	Int16,UInt8
INTEGER	Int32,UInt16,Interval

©2013–2025 Tencent Cloud. All rights reserved.

BIGINT	Int64,UInt32
FLOAT	Float32
DATE	Date
TIME	DateTime
TIMESTAMP	DateTime
TIMESTAMP_LTZ	DateTime
INTERVAL_YEAR_MONTH	Int32
INTERVAL_DAY_TIME	Int64
ARRAY	Array
MAP	Map

©2013–2025 Tencent Cloud. All rights reserved.

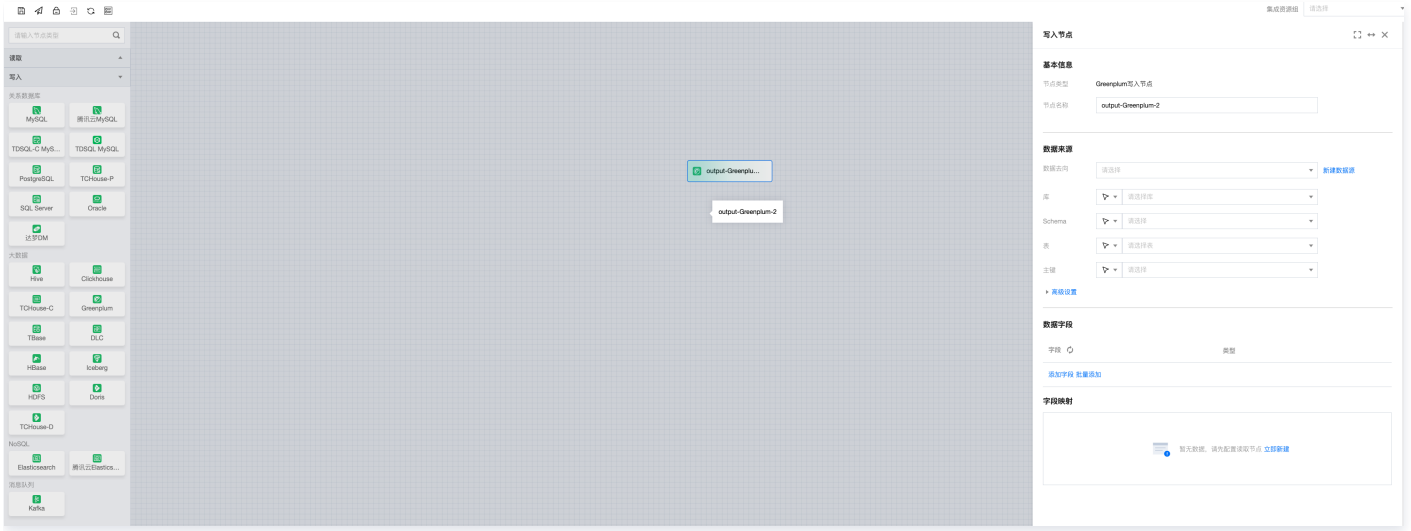
Greenplum Data Source

Last updated：2024-09-06 15:06:42

Greenplum single table write node configuration

Configure Greenplum node

1. In the DataInLong page, click **Real-time Synchronization** in the left sidebar.
2. On the Real-time Synchronization page, select **Single Table Synchronization** at the top, create a new one (choose Form or Canvas mode), and enter the configuration page.
3. Click the left side **Write**, click to select **Greenplum node** and configure node information.



4. You can refer to the table below for parameter configuration.

Parameter	Description
Data Flow Direction	Greenplum data source to be written to.
Database	Supports selecting or manually entering the name of the database to be written to. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
schema	Support selecting or manually entering the Schema Name to be read.
Table	Supports selecting or manually entering the name of the table to be written to. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Primary key	Select a field as the primary key for the write table.
Advanced Settings	You can configure parameters based on business requirements.

5. Preview data fields and map them with the fields configured in the read node, click **save**.

Greenplum log collection write node configuration

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Flow Direction	Select an available Greenplum data source in the current project.
Database	Supports selecting or manually entering the name of the database to be written to. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
schema	Support selecting or manually entering the Schema Name to be read.
Table	Supports selecting or manually entering the name of the table to be written to. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Primary key	Select a field as the primary key for the data table.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Greenplum write data type conversion support

Write

Internal Type	Greenplum Type
TINYINT	–
SMALLINT	SMALLINT,INT2,SMALLSERIAL,SERIAL2
INT	INTEGER,SERIAL
BIGINT	BIGINT,BIGSERIAL
DECIMAL(20, 0)	–
FLOAT	REAL,FLOAT4
DOUBLE	FLOAT8,DOUBLE PRECISION
DECIMAL(p, s)	NUMERIC(p, s),DECIMAL(p, s)

©2013–2025 Tencent Cloud. All rights reserved.

BOOLEAN	BOOLEAN
DATE	DATE
TIME [(p)][WITHOUT TIMEZONE]	TIME [(p)][WITHOUT TIMEZONE]
TIMESTAMP [(p)][WITHOUT TIMEZONE]	TIMESTAMP [(p)]WITHOUT TIMEZONE
STRING	CHAR(n),CHARACTER(n),VARCHAR(n),CHARACTER VARYING(n),TEXT
BYTES	BYTEA

©2013–2025 Tencent Cloud. All rights reserved.

Hive Data Source

Last updated：2024-09-06 15:07:09

DataInLong provides real-time writing capabilities to Hive. This document introduces the current capabilities of Hive real-time data synchronization.

Supported Versions

Currently, DataInLong supports real-time writing to the entire Hive database and single tables. The following version restrictions apply for using real-time writing capabilities:

Type	Version
Hive	1.x, 2.x, 3.x

Use Limits

- Only Append writing is supported, and target data duplicates cannot be avoided.
- DDL changes are not supported; schema changes at the source will not be automatically synchronized to the target.
- Automatically created tables during the full phase do not support syncing comment information.

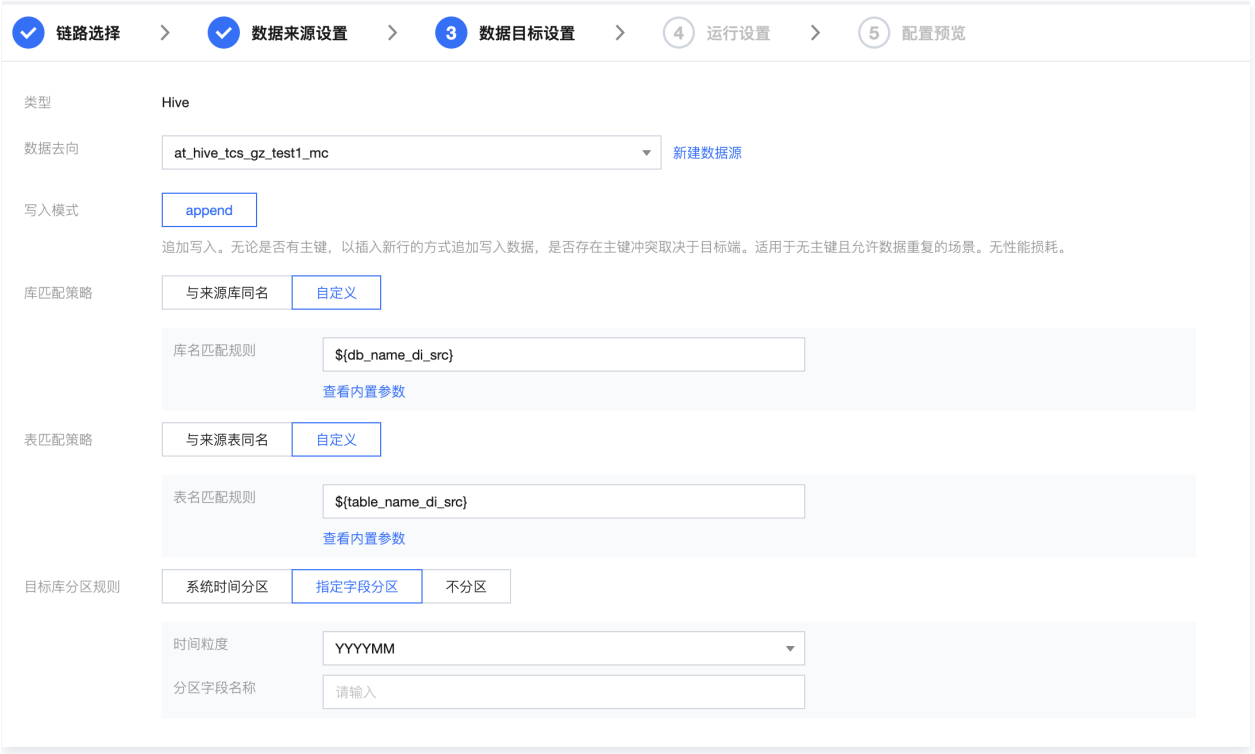
Whole Database Write Configuration

Supported Data Sources

Currently, real-time synchronization of entire databases from the following source types to Hive target databases is supported:

- MySQL Whole Database Source Configuration Details
- TDSQL-C MySQL Whole Database Source Configuration Details
- Oracle Whole Database Source Configuration Details
- PostgreSQL Whole Database Source Configuration Details

Data Target Configuration

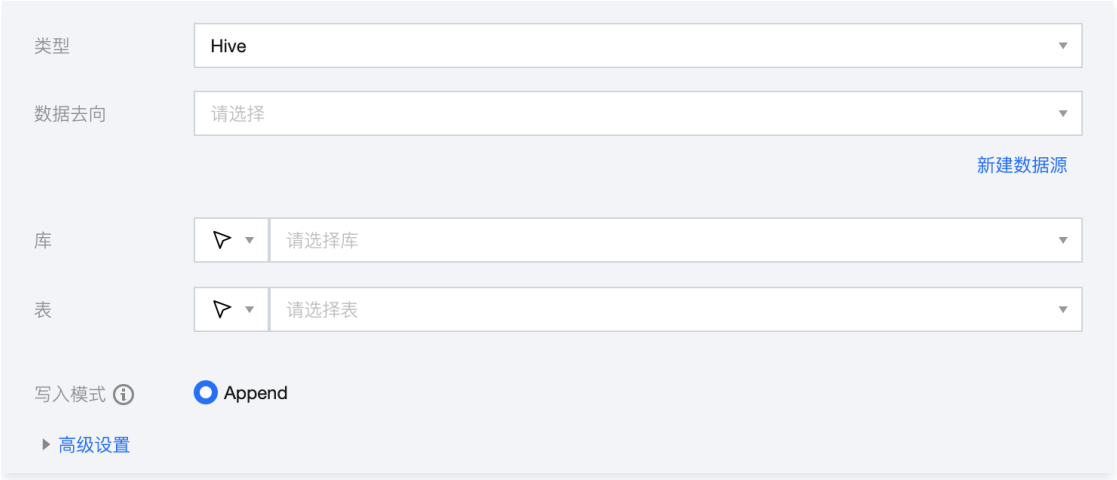


Parameter	Description

©2013-2025 Tencent Cloud. All rights reserved.

Data Flow Direction	Select the target data source for synchronization.
Write mode	Append: Append mode to write to data tables (currently, only this mode is supported).
Database/Table Matching Strategy	Naming rules for databases and data table objects in Hive: • Default to the same names as the source database/source table. • Self Definition: Supports generating target database and table names using built-in parameters and string combinations. **Note** Example: If the source table name is table1, and the mapping rule is ${table_name_di_src}_inlong, then the data from table1 will eventually be mapped and written into table1_inlong.
Target database partition rules	• System Time Partitioning: ○ Time Granularity: Single-choice, users can select from four time formats: YYYYMMDD, YYYYMM, YYYY, YYYY-MM-DD HH. Specified Field Partitioning: ○ Time Granularity: Single-choice, users can select from four time formats: YYYYMMDD, YYYYMM, YYYY, YYYY-MM-DD HH. ○ Partition Field Name: Input box, where users can enter the name of the field to be used as the partition field. This must be a time-type field. • No Partition: No partition processing.

Single Table Write Configuration

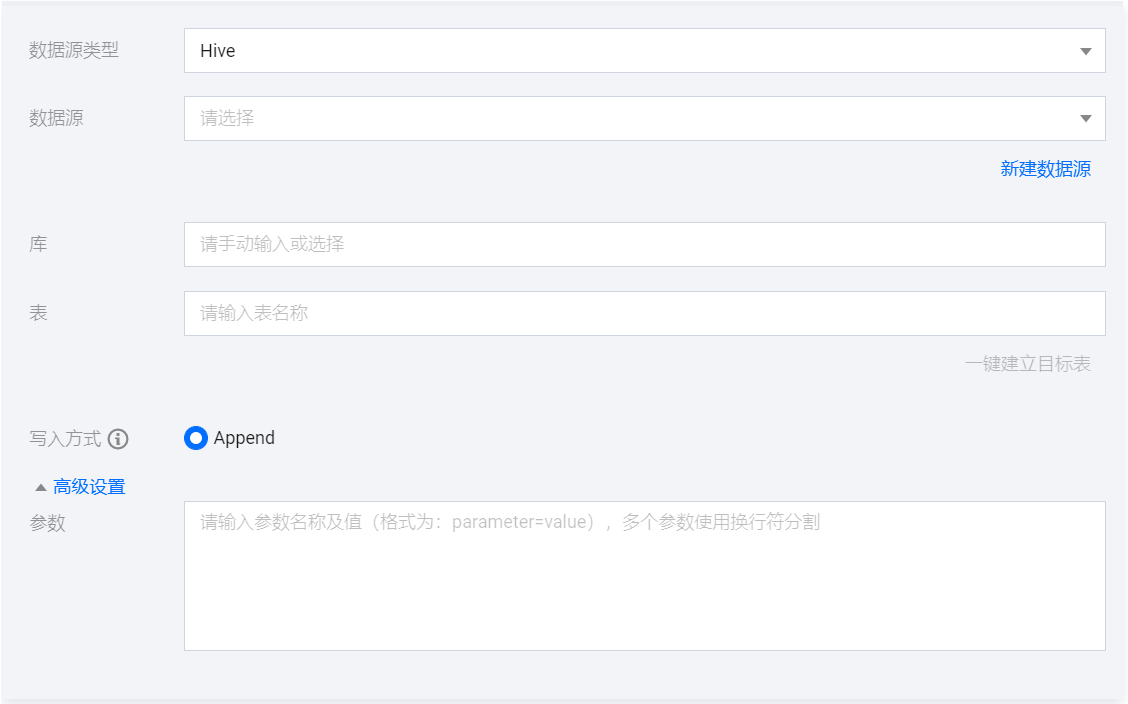


Parameter	Description
Data Flow Direction	The Hive data source to write into.
Database	Supports selecting or manually entering the name of the database to be written to. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Supports selecting or manually entering the name of the table to be written to. When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.

©2013-2025 Tencent Cloud. All rights reserved.

Write mode	Hive only supports Append write
Advanced Settings	You can configure parameters based on business requirements.

Log Collection Write Configuration



Parameter	Description
Data Source	Select an available HIVE data source in the current project.
Database/Table	Select the corresponding database table in this data source.
Write mode	Hive only supports Append write.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Data Type Conversion Supported

Internal Type	Hive Type
CHAR(p)	char(p)
VARCHAR(p)	varchar(p)
STRING	string
BOOLEAN	boolean
TINYINT	tinyint
SMALLINT	smallint
INT	int
BIGINT	bigint
FLOAT	float

©2013–2025 Tencent Cloud. All rights reserved.

DOUBLE	double
DECIMAL(p, s)	decimal(p, s)
DATE	date
TIMESTAMP	timestamp(9)
BINARY	bytes
LIST	array
MAP	map
STRUCT	row

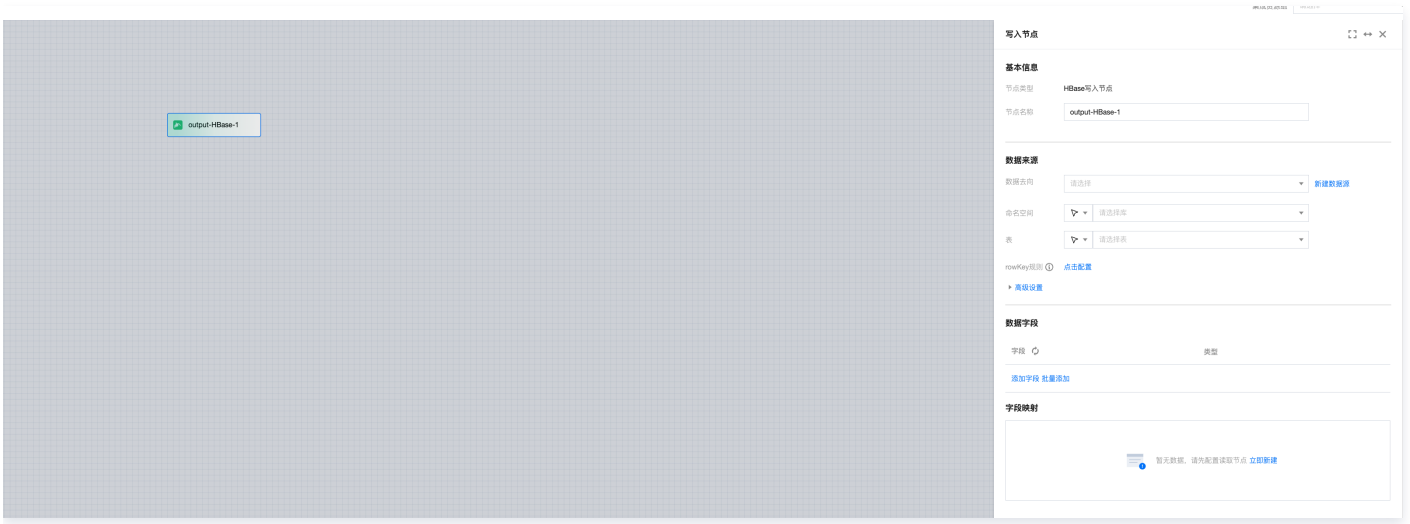
©2013–2025 Tencent Cloud. All rights reserved.

HBase Data Source

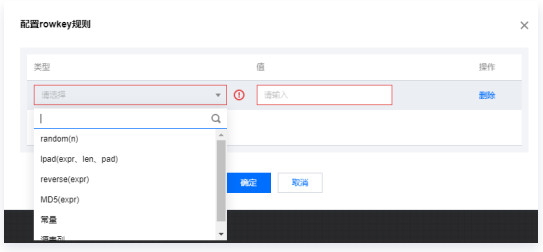
Last updated：2024-09-06 15:07:43

HBase single table write node configuration

1. In the DataInLong page, click **Real-time Synchronization** in the left sidebar.
2. On the Real-time Synchronization page, select **Single Table Synchronization** at the top, create a new one (choose Form or Canvas mode), and enter the configuration page.
3. Click the left side**Write**, click to select **HBase node** and configure the node information.



4. You can refer to the table below for parameter configuration.

Parameter	Description
Data Source	HBase data source to be written.
Namespace	Support selection or manual input of the space to be written.
Table	Supports selecting or manually entering the name of the table to be written to. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
rowkey Regulations	Click Configure to enter the rowkey configuration page, configure the type and corresponding value. 
Advanced Settings	You can configure parameters based on business requirements.

5. Preview data fields and map them with the fields configured in the read node, click **save**.

HBase log collection write node configuration

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Source	Select an available HBase data source in the current project.
Namespace	Manually input or select a namespace.
Table	Select the corresponding table in the data source.
rowkey Regulations	HBase data source needs to configure rowkey regulations.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

HBase write data type conversion support

Write

The supported data types and conversion relationships for HBase writes are as follows (when processing HBase, the data types of the HBase data source will be mapped to the data types of the data processing engine):

Internal Type	HBase Type
CHAR,VARCHAR,STRING	byte[] toBytes(String s),String toString(byte[] b)
BOOLEAN	byte[] toBytes(boolean b),boolean toBoolean(byte[] b)
BINARY VARBINARY	Returns byte[] as is.
DECIMAL	byte[] toBytes(BigDecimal v),BigDecimal toBigDecimal(byte[] b)
TINYINT	new byte[] { val },bytes[0] // returns first and only byte from bytes
SMALLINT	byte[] toBytes(short val),short toShort(byte[] bytes)
INT	byte[] toBytes(int val),int toInt(byte[] bytes)
BIGINT	byte[] toBytes(long val),long toLong(byte[] bytes)
FLOAT	byte[] toBytes(float val),float toFloat(byte[] bytes)
DOUBLE	byte[] toBytes(double val),double toDouble(byte[] bytes)

©2013–2025 Tencent Cloud. All rights reserved.

DATE	Stores the number of days since epoch as int value.
TIME	Stores the number of milliseconds of the day as int value.
TIMESTAMP	Stores the milliseconds since epoch as long value.

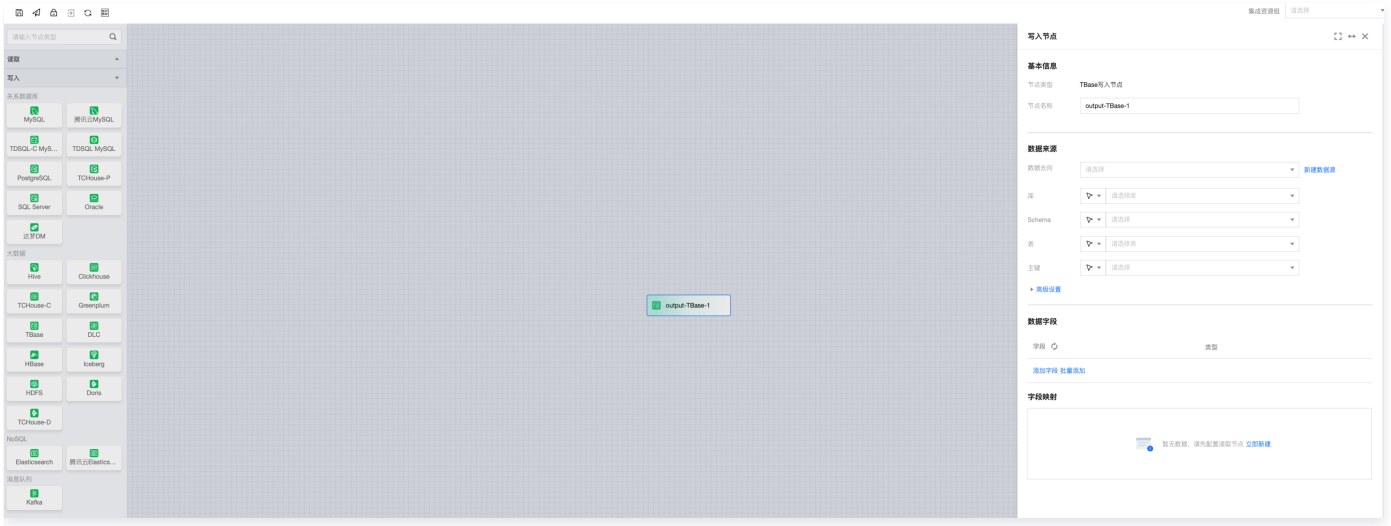
©2013–2025 Tencent Cloud. All rights reserved.

TBase Data Source

Last updated：2024-09-06 15:08:30

TBase Single Table Write Node Configuration

1. In the DataInLong page, click **Real-time Synchronization** in the left sidebar.
2. On the Real-time Synchronization page, select **Single Table Synchronization** at the top, create a new one (choose Form or Canvas mode), and enter the configuration page.
3. click on the left **write**, click select **TBase Node** and configure node information.



4. You can refer to the table below for parameter configuration.

Parameter	Description
Data Flow Direction	Select the available TBase data source for the project
Database	Support selecting or manually entering the name of the library to be written to. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
schema	Support selecting or manually entering the schema name to be written.
Table	Supports selecting or manually entering the name of the table to be written to.
Primary key	Select a field as the primary key for the write table.
Advanced Settings	You can configure parameters based on business requirements.

5. Preview data fields and map them with the fields configured in the read node, click **save**.

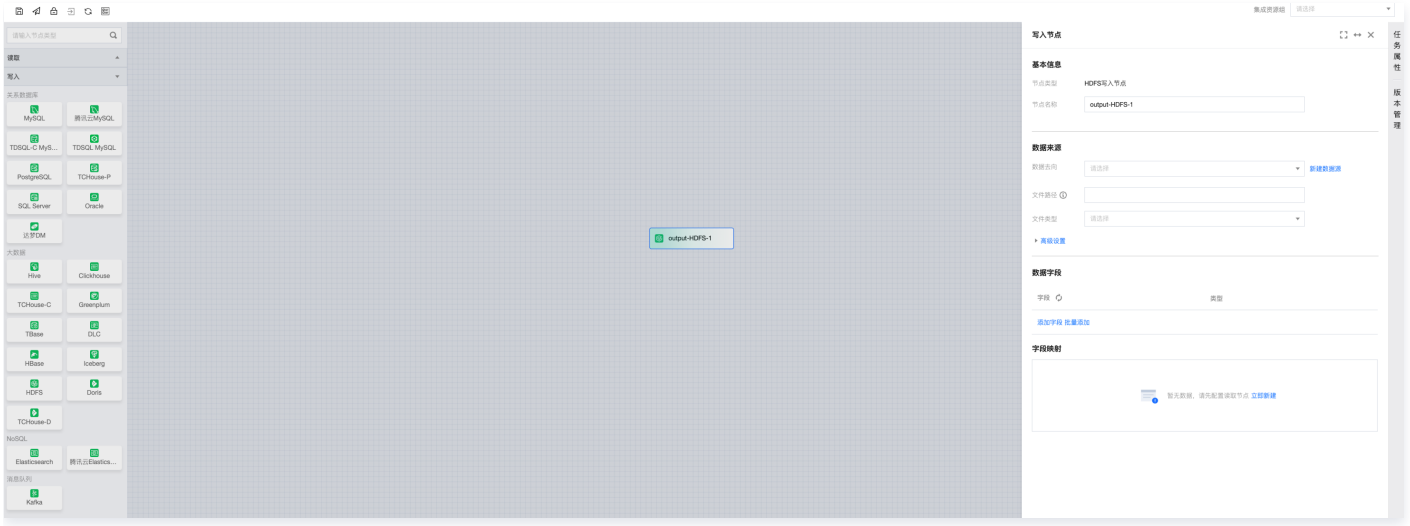
©2013-2025 Tencent Cloud. All rights reserved.

HDFS Data Source

Last updated：2024-09-06 15:09:16

HDFS Single Table Write Node Configuration

1. In the DataInLong page, click **Real-time Synchronization** in the left sidebar.
2. On the Real-time Synchronization page, select **Single Table Synchronization** at the top, create a new one (choose Form or Canvas mode), and enter the configuration page.
3. click **Write** on the left, click **Select HDFS Node** and configure the node information.



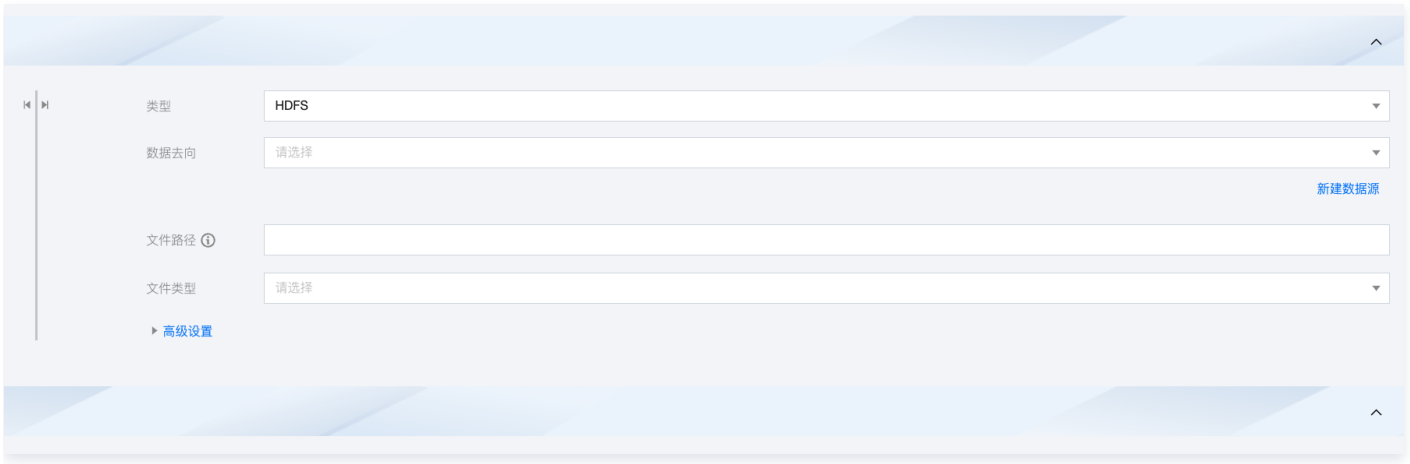
4. You can refer to the table below for parameter configuration.

Parameter	Description
Data Flow Direction	Select the available HDFS data source in the current project.
File path	Path information of the file system. The path supports using ‘*’ as a wildcard, specifying the wildcard will traverse multiple file information.
File Type	HDFS supports four file types: txt, orc, parquet, csv. • txt: Represents TextFile file format. • orc: Represents ORCFile file format. • parquet: Represents standard Parquet file format. • csv: Represents standard HDFS file format (logical two-dimensional table).
Advanced Settings	You can configure parameters based on business requirements.

5. Preview data fields and map them with the fields configured in the read node, click **save**.

HDFS Log Collection Write Node Configuration

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Flow Direction	Select the available HDFS data source in the current project.
File path	Path information of the file system. The path supports using ‘*’ as a wildcard, specifying the wildcard will traverse multiple file information.
File Type	HDFS supports four file types: txt, orc, parquet, csv.
Advanced Settings (Optional)	You can configure parameters based on business requirements.

©2013–2025 Tencent Cloud. All rights reserved.

Whole Database Synchronization Task Configuration
Data sources supported by whole database synchronization

Last updated: 2024-11-25 15:58:38

This product supports instance-level synchronization migration for data sources such as MySQL, TDSQL-C MySQL, and Kafka. Currently, it supports Doris, DLC, Kafka, and other target databases. Details of supported whole database links are as follows:

Data Source	Data Target
MySQL	Whole database sync to DLC configuration details
TDSQL-C MySQL	
TDSQL MySQL	
Kafka	
Mongo	
Oracle	
PostgreSQL	
OceanBase	
MySQL	Whole library sync to Doris configuration details
TDSQL-C MySQL	
TDSQL MySQL	
PostgreSQL	
Kafka	
OceanBase	
MySQL	Whole Database Sync to Kafka Configuration Details
TDSQL-C MySQL	
Mongo	
PostgreSQL	
Oracle	
MySQL	Whole Database Synchronization to TCHouse-P Configuration Details
TDSQL-C MySQL	
PostgreSQL	
SQLServer	
Oracle	
MySQL	Whole database sync to Iceberg configuration details
TDSQL-C MySQL	
TDSQL MySQL	

©2013-2025 Tencent Cloud. All rights reserved.

Kafka	
Oracle	
OceanBase	
PostgreSQL	
SQLServer	
MySQL	Whole Database Synchronization to StarRocks Configuration Details
TDSQL-C MySQL	
PostgreSQL	
Oracle	
MySQL	Whole database sync to Hive configuration details
TDSQL-C MySQL	
Oracle	
PostgreSQL	
TDSQL-C MySQL	Whole Database Synchronization to Elasticsearch Configuration Details
Mongodb	Whole Database Synchronization to PostgreSQL Configuration Details
MySQL	

©2013-2025 Tencent Cloud. All rights reserved.

Overview of Whole Database Task Configuration

Last updated: 2024-09-06 15:10:29

Background

Whole database migration supports the monitoring of source-end data and structure, allowing the real-time synchronization of full or incremental data from all databases and tables of the source to the target end. It also supports automatic table creation at the target end, field change synchronization, and other features. Supported data sources include MySQL, Doris, DLC, and Kafka.

Prerequisites

1. The data sources for both the source and target ends have been configured for subsequent tasks. For details, see Data Source Management and Configuration Method.
2. DataInLong resource group has been purchased. For details, see Configure Integrated Resource Group.
3. Network connectivity between the DataInLong resource group and data sources has been completed. For details, see Integrated Connectivity and Usage Plan.
4. Data source environment preparation has been completed. You can grant the database account configured for data sources the necessary permissions for corresponding operations before executing the sync task, based on your sync configuration requirements.

Operation step

Step 1: Create Whole Database Synchronization Task

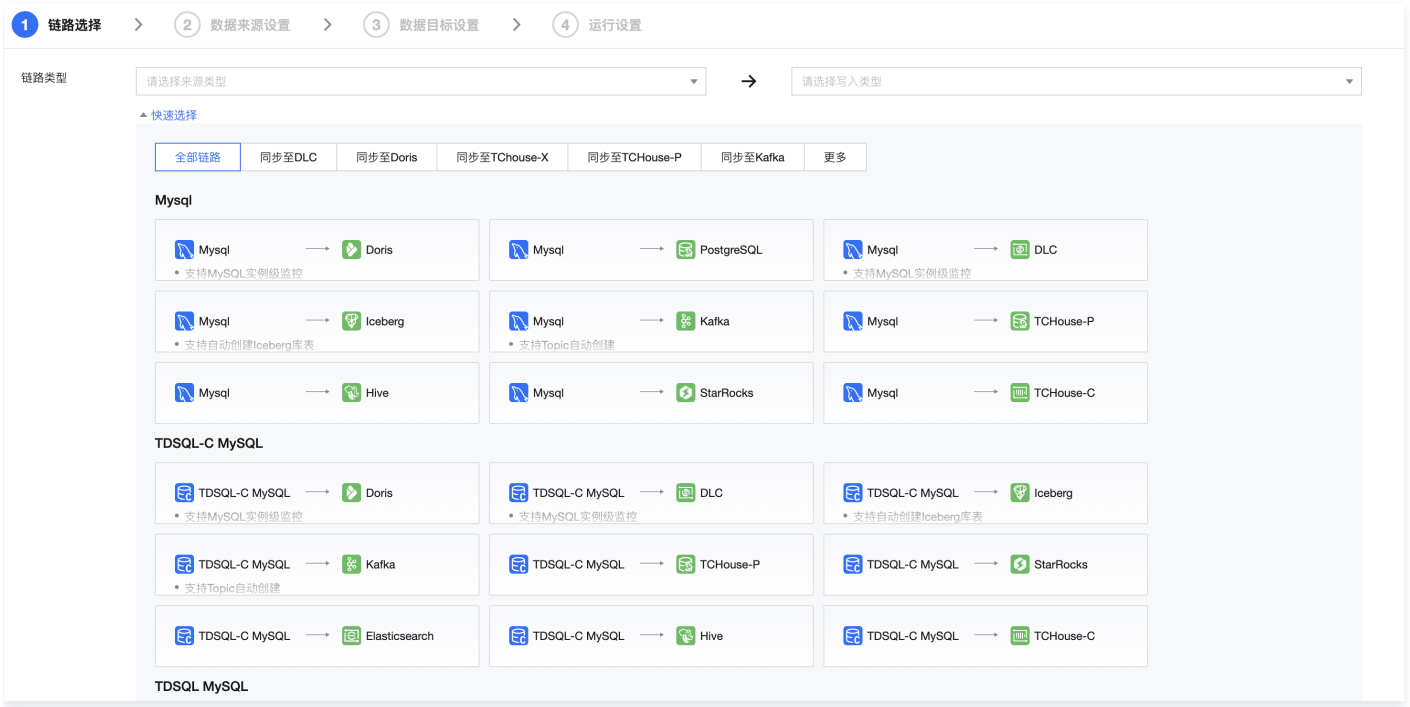
Enter the **Configuration Center > Real-time Synchronization** task page, then click **New** Whole Database Migration Task.



Step 2: Link Selection

You can select the data types for the source and destination according to actual business needs. After selection, the subsequent steps will display the corresponding source and target configuration parameters based on the selected types. Ensure that the selected data source type is consistent with the actual configured data source type.

©2013-2025 Tencent Cloud. All rights reserved.



Step 3: Data Source Settings

In this step, you can select the databases and tables from the source data source that you want to sync, and configure the reading method from the source, consistency semantics, filter operation, timezone, and more.

> **Note**
> The configuration item here varies depending on the type of source data source, and the actual configuration interface of the data source shall prevail.



Step 4: Data Target Settings

In this step, you can define the target data source and related attributes such as database and table, set the write mode, database and table name mapping rules, and establish name mapping rules between target and source objects. Currently, the whole database target object matching strategies support **"Same as Source"** and **"Custom Definition"**

- **Duplicate name with the source database/source table**

©2013–2025 Tencent Cloud. All rights reserved.

By default, in a whole database synchronization task, the source database and data tables will be written into the target database or table with the same name. Under this policy, the system will match objects with the same name in the target data source when the task runs.



> **Description:**The task plan synchronizes TableA and TableB from the source database DB1 to the Doris data source. After configuring the **same name as the source database** and **same name as the source tables** strategies, the task will default to matching "DB1.TableA" and "DB1.TableB" within the Doris connection.

- **Custom Definition**

The Custom Definition rule supports setting special relationships between the source and target, such as adding a uniform prefix or suffix to the source database name or table name when writing to the target database or table during task execution. Under this policy, the system will by default match the target object based on the naming rules during task execution.



In Custom Definition mode, the system provides built-in system parameters for full database scenarios. The built-in parameters mainly cover the source data source name, source database name, source table name, etc. For Kafka types, dynamic matching of data field value within the message is supported. The following is an explanation of the built-in parameters in Custom Definition mode:

Parameter name	Parameter description
Source Data Source Name	${datasource_name_di_src}
Source Database Name	${db_name_di_src}
Source Table Name	${table_name_di_src}
Source Schema Name	${schema_name_di_src} **Description:**Applicable to data source types with a Schema attribute, such as PostgreSQL, Oracle, etc.
Source Topic Name	${topic_di_src} **Description:** Only applicable to Kafka types.

©2013–2025 Tencent Cloud. All rights reserved.

Data Field	${key} **Note** Only applicable to Kafka types. Please replace the key with the specific field name.

Note

- Example 1: If the source table name is table1 and the mapping rule is ${table_name_di_src}_inlong, the data from table1 will be finally mapped and written into table1_inlong.
- Example 2: If the source Kafka message format is {"name":"inlong";"age":12} and the mapping rule is ${name}_di, this message will be finally mapped and written into the inlong_di table.
- The configuration item here varies depending on the type of target data source, and the actual configuration interface of each target end shall prevail.

Step 5 (Optional): Batch manual creation of target tables

For the name matching strategy between source and target tables, the system will automatically convert the DDL structure between heterogeneous data sources, providing you with the ability to quickly modify and create target tables in batches.

Note

1. Batch table creation involves configuring the target database and table naming strategy. For operation steps, refer to Step Four of this document.
2. Batch table creation is only supported for certain target data sources and synchronization links. For specific supported links, refer to Automatic table creation capability in synchronization tasks.

1. Check the database and table name matching strategy, and click on **Batch Table Creation.**



2. Confirm table creation rules

In this step, the system will scan the target objects based on the database and table matching strategy and automatically generate target table DDL statements. You can check if the target tables match correctly, configure the table creation method, and edit the table creation statements. Key features and their corresponding descriptions are as follows:

©2013–2025 Tencent Cloud. All rights reserved.



Key Features		Feature Overview	Feature Examples
1	Target Database/Table Name	The system will by default generate the target database/table name based on the matching rules and also check if the database/table exists in the target data source: • Unmatched Databases/Tables: Databases/Tables that do not match the rules are not found in the target data source. The list will highlight unmatched databases/tables by default. • Successfully Matched Databases/Tables: The target data source contains database/table objects that meet the matching rules.	

2	Target Table Creation Method	For database/table objects in the target data source, the system provides multiple table creation strategies:
		• Unmatched Databases/Tables: Supports **Create New Table** and **Do Not Create New** options
		  ○ Create New Table: Automatically creates a new table based on the target table DDL generated from the conversion during this batch creation.
		  ○ Do Not Create New: Temporarily ignores this table in this operation.
		• Successfully Matched Databases/Tables: Supports **Use Existing Table** and **Delete Existing Table and Create New** options
		  ○ Use Existing Table: Temporarily ignores this table in this operation.
			1. Unmatched Databases/Tables Strategy Configuration:
			 
			2. Successfully Matched Databases/Tables Strategy Configuration:
			 

©2013–2025 Tencent Cloud. All rights reserved.

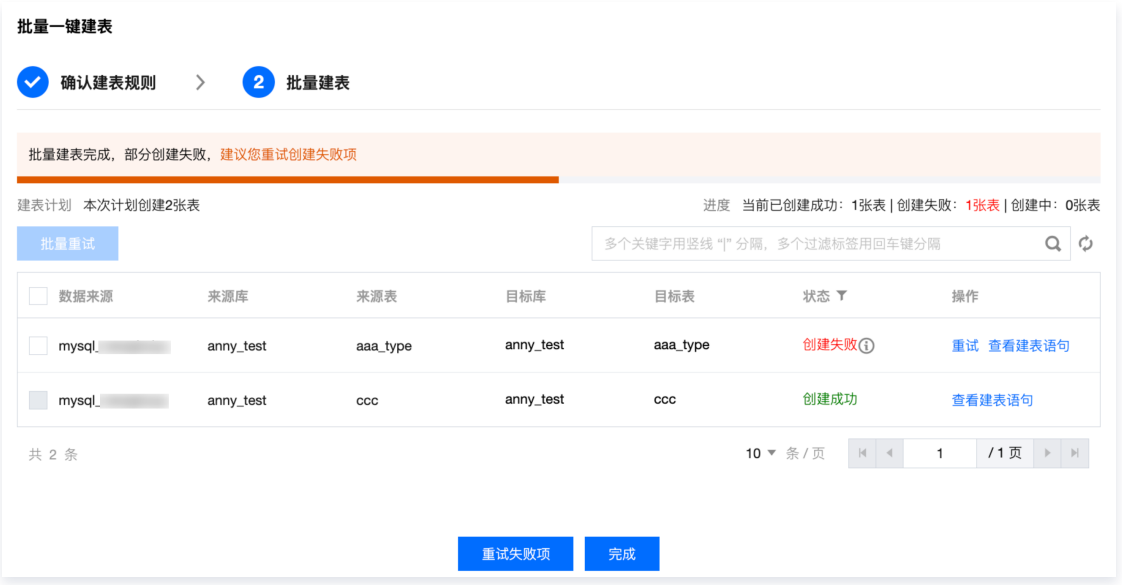
		○ Delete Existing Table and Create New: Deletes the existing table and recreates a table with the same name based on the target table DDL generated from the conversion in this operation.	
3	Preview/Edit Table Statements	For the specified target table creation method, the system will automatically generate DDL samples for you to view and edit the table creation statements: 1. The target table name is automatically generated based on the source table name and the target table matching strategy by default. 2. The target table fields are by default consistent with the source table name. 3. For target data sources with multiple data models (such as Doris), the system will specify the corresponding model based on the default strategy. You can manually modify the DDL statements to meet business requirements.	1. New Table DDL Statement:  2. Delete existing table and create DDL statement: 

3. Batch create target tables

   Once the target table creation rules are confirmed, you can create the target tables in bulk in this step.

   ○ After successful creation, you can click **complete** to close the pop-up window and continue configuring subsequent tasks.

   ○ If there are tables that failed to be created, you can view the reasons for the failures or click **retry** to recreate.



Step Six: Configure Runtime Resources and Strategies

This step mainly configures resources for the task, DDL and exception data processing strategies, and advanced runtime parameters.

1. **Integrated Resource Configuration**

   ○ Integrated resources support multiple allocation methods

   ○ Fixed Allocation: In this method, resource allocation remains constant throughout the full and incremental synchronization phases, preventing resource preemption among tasks. It is suitable for scenarios where data may

©2013–2025 Tencent Cloud. All rights reserved.

significantly change during task execution.

- Allocation by synchronization phase: Allocates planned resource usage based on different synchronization phases (full and incremental) to save overall resource usage.

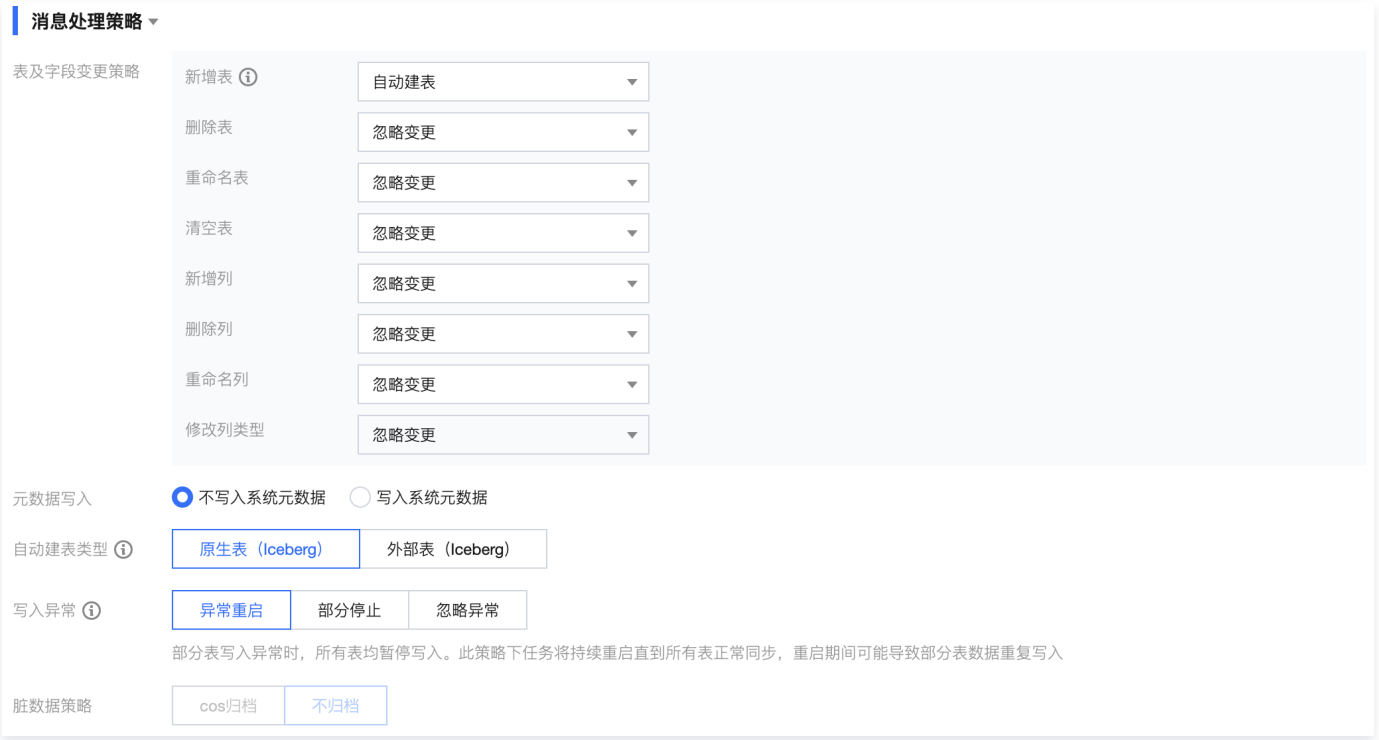
- Associate the current task with the corresponding integrated resource group and set the JM, TM specifications and task runtime parallelism. The actual CU (Compute Unit) consumption for the current task runtime = JobManager specification + TaskManager specification × parallelism.



2. **Message Processing Strategy:**

Configure the current task with settings including Write DDL Messages Response Policy, Metadata Write Strategy, Data Anomaly Write Strategy, and Dirty Data Archiving Strategy.

> **Note**
> The configuration options here vary according to the type of target data source. Please refer to the actual configuration interface for each link.



Parameter	Description
DDL Messages	This parameter is mainly used to set the process for passing the DDL change messages captured from the source to the downstream during task synchronization. The target end provides the following response policies for DDL messages:

©2013–2025 Tencent Cloud. All rights reserved.

1. Automatic Change: Under this strategy, the target end will automatically follow the structural changes of the source end, including automatic table creation, automatic column addition, etc.
2. Ignore Changes: Under this strategy, the target end will ignore DDL change messages and will not respond or send messages.
3. Log Alerts: Under this strategy, the target end will ignore DDL change messages, but the log will alert with details of the DDL change messages.
4. Task Error: Under this strategy, once a DDL change occurs in the source end, the entire task will experience continuous abnormal restarts and report exceptions.

> **Note**
> There are differences in DDL types and message handling supported by different sources and targets. Please refer to the configuration policies supported by different links. For details on DDL types supported by the data source, see Supported DML and DDL Operations.

Metadata Writing

When the option **Automatic Table Creation** is selected for the Add Table in the DDL policy, you can choose whether to write metadata. Checking the corresponding metadata fields will create the respective metadata fields in the target table, and the relevant system metadata will be written during synchronization.



Currently, only the following entire-database links support writing system metadata:

任务类型	数据来源	数据去向	支持的系统元数据
实时整库迁移任务	MySQL、TDSQL-C	Iceberg、TCHouse-P 、Doris、Kafka、DLC	database_wedata_di （管理参数database，表示数据库名称） table_wedata_di（管理参数table，表示数据表名） type_wedata_di （管理参数type，表示操作类型） ts_wedata_di（管理参数ts，表示日志时间） processing_time_wedata_di （处理时间，当前系统处理本条记录的机器本地时间）
	Oracle	Iceberg、TCHouse-P、Doris、Kafka、DLC	database_wedata_di：管理参数database，表示数据库名称 table_wedata_di：管理参数table，表示数据表名 ts_wedata_di：管理参数ts，表示日志时间 processing_time_wedata_di：处理时间，当前系统处理本条记录的机器本地时间
	PostgreSQL	Iceberg、TCHouse-P 、Doris、Kafka、DLC	database_wedata_di （管理参数database，表示数据库名称） table_wedata_di（管理参数table，表示数据表名） type_wedata_di （管理参数type，表示操作类型） ts_wedata_di（管理参数ts，表示日志时间） processing_time_wedata_di （处理时间，当前系统处理本条记录的机器本地时间）
	SQL Server	Iceberg、TCHouse-P、DLC	database_wedata_di：管理参数database，表示数据库名称 table_wedata_di：管理参数table，表示数据表名 ts_wedata_di：管理参数ts，表示日志时间 processing_time_wedata_di：处理时间，当前系统处理本条记录的机器本地时间

Write Exception

This parameter is used to set how to handle failed data writes due to various reasons such as unmatched table segment structures or field types during synchronization, and whether to interrupt the data flow. The overall write exception strategy includes:

1. Partial Stop: When a partial table write fails, only the data write for that table will stop, while other tables continue normal synchronization. The stopped table cannot resume writing during the current task run.
2. Abnormal restart: When a partial table write fails, all table writes are paused. Under this strategy, the task will continuously restart until all tables synchronize normally, which may lead to duplicated data writes for some tables during the restart period.

©2013-2025 Tencent Cloud. All rights reserved.

	3. Ignore Exceptions: Ignore data that cannot be written due to exceptions in the table and mark them as dirty data. The other data in the table and other tables within the task will synchronize normally. Two solutions are available for dirty data: COS Archiving and Non-Archiving.
Dirty Data	When the write exception configuration is set to **Ignore Exceptions**, you can choose whether to archive the ignored data: 1. COS Archiving: Uniformly archive the data with write exceptions into a COS file for storage. This method can prevent the loss of data with write exceptions, and allows for analysis and data supplementation in subsequent scenarios. 2. Do not archive: The task completely ignores and discards data with writing exceptions.

3. **Task Operation Strategy:**

Set the commit interval, maximum restart times, and task-level runtime parameters for the current task.

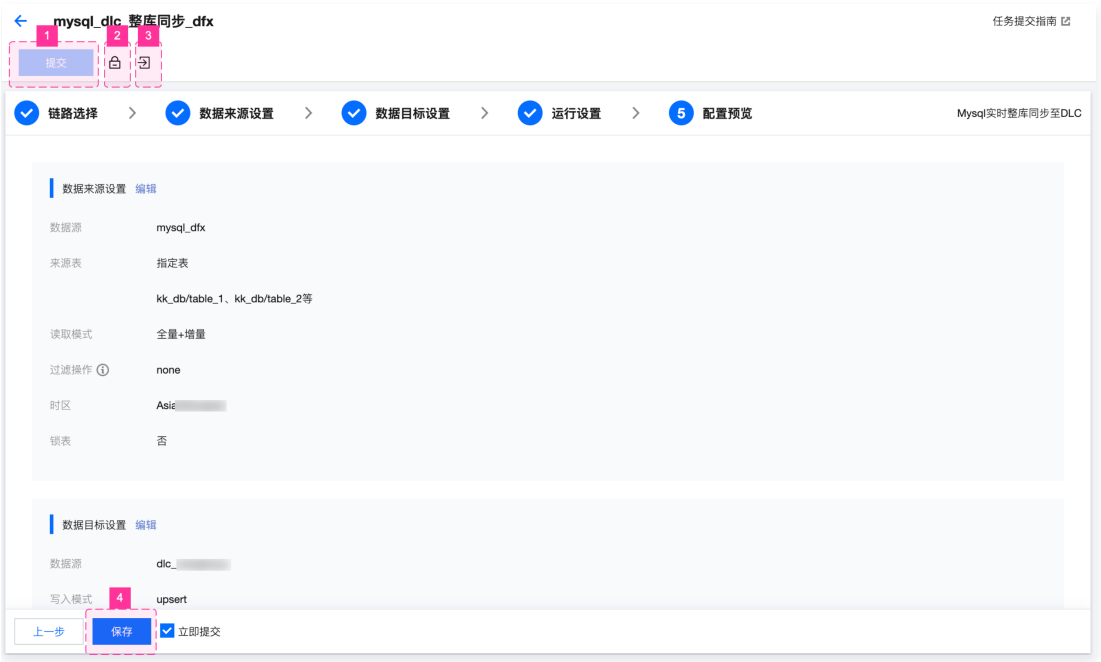


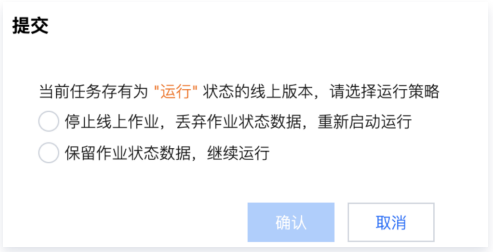
Parameter	Description
Checkpoint Interval	The maximum checkpoint interval for the current task submission.
Maximum Restart Times	Set the maximum restart threshold for failures during execution. If the restart count exceeds this threshold, the task status will be set to **Failed**. The setting range is [-1, 100], where • a value of 0 indicates no restart, • and -1 indicates no limit on maximum restart times.
Parameter	Set task-level runtime parameters. Different source and target ends support different task-level parameters. For details, please refer to Advanced Real-time Node Parameters.

Step 7: Configuration Preview and Task Submission

1. Configuration Preview

©2013-2025 Tencent Cloud. All rights reserved.



Serial number	Parameter	Description
1	Submit	Submit the current task to the production environment. On submission, different operation strategies can be selected based on the presence of production tasks • If there are no effective online tasks for the current task, i.e., it's the first submission or the online task is in a "failed" status, you can submit it directly. • If there are "Running" or "Paused" online tasks, you need to choose different strategies. Stopping the online job will abandon the previous task's running location, consuming data from the beginning. Retaining the job status will resume from the last consumed location after restarting.  **Note** Click **Start Now** to immediately run the task after submission, otherwise, it needs to be manually triggered to start running.
2	Lock/Unlock	The default creator is the first lock holder, and only the lock holder can edit task configurations and run tasks. If the lock holder does not perform any editing operations within 5 minutes, others can click the icon to acquire the lock. If successful, they can perform editing operations.
3	Go to Operations	Quickly jump to the task operations page based on the current task name.
4	Save	After previewing, you can click the **Save** button to save the whole database task configuration. The task will not be submitted to the **OPS Center** if only saved

2. Task Configuration Detection and Submission.

©2013–2025 Tencent Cloud. All rights reserved.

Steps	Step Description	
Task Configuration Detection	This step involves detection of the read end, write end, and resources in the task: • Detection passed: No configuration errors. • Detection failed: Configuration has issues that need to be fixed for further configuration. • Detection Alert: This is a system-recommended modification item. After modification, click **Retry** to recheck, or you can click **Ignore Exceptions** to proceed without blocking further configuration. The currently supported check items are listed in the table below.	

Submission Policy Selection	In this step, you can choose the submission policy for this task:
	• **First Submission: The first submission supports synchronizing data from the default or specified point.**
	  ○ **Start Now and synchronize from the default point position**: If the source-end configuration is set to "Full + Incremental" mode, it will first synchronize existing data (full phase) by default. After completion, binlog will be consumed to obtain changed data (incremental phase). If the source-end configuration is set to "Incremental Only", it will start reading from the latest binlog point position by default.
	  ○ **Start Now and synchronize from a designated time point**: The task will synchronize data according to the configured time and timezone. If the designated time point cannot be found, the task will start synchronizing from the earliest binlog point position by default. If the source-end configuration is "Full + Incremental", the task will skip the full phase and start synchronizing from the designated time point in the incremental phase by default.
	  ○ **Do not start now:** After submission, the task will not start running for the time being, and can be manually started later in the operation list.
	• **Non-First Submission: Supports starting or continuing tasks with running status.**
	  ○ **Continue running**: Under this strategy, after the new version of the task is submitted, it will continue running from the last sync point.
	  ○ **Restart and synchronize from the specified position:** Under this strategy, you can specify the position to start reading. The task will ignore the old version and restart reading from the specified position. If the specified time position is not found, the task will start synchronizing from the earliest binlog position by default.
	  ○ **Restart and run from the default point position**: Under this strategy, reading will start from the default point position as per the source-end configuration. If the source-end configuration is set to "Full + Incremental", it will first synchronize existing data (full phase) by default. After completion, binlog will be consumed to obtain changed data (incremental phase). If the source-
		1. First Submission
		 
		2.Non-First Submission
		 

©2013-2025 Tencent Cloud. All rights reserved.

	end configuration is set to "Incremental Only", it will start reading from the latest binlog point position by default. The supported submit run strategies vary with different task statuses. Details can be found in the subsequent tables. Each submission will generate a new version of the real-time task. You can configure the version description in the dialog box.	
Submitting the job	After successful submission, you can click **Go to Operations** to view the task's status.	

Task Configuration Detection supports the following check items:

Check Classification	Check Items	Description
Task Configuration Detection	Source Configuration	Check if any required fields are missing on the source side
	Target Configuration	Check if any required fields are missing on the target side
	Mapping Configuration	Check if field mappings are configured
	Resource Group Configuration	Check if the resource group is configured
Data Source Detection	Source end connectivity detection	Check if the data source at the source end is networked with the resource group in the task configuration. If the check fails, view diagnostic information. After resolving network issues, recheck; otherwise, the task is likely to fail.
	Target end connectivity detection	Check if the data source at the target end is networked with the resource group in the task configuration. If the check fails, view diagnostic information. After resolving network issues, recheck; otherwise, the task is likely to fail.
Resource Detection	Resource status detection	Check if the resource group is in an available status. If the resource status is unavailable, please switch the resource group in the task configuration; otherwise, the task is likely to fail.
	Resource Capacity Detection	Check if the remaining resources in the resource group meet the resource requirements of the task configuration. If the check fails, please reduce the task resource configuration or expand the resource group.

Submission and running strategies supported by different task running statuses:

Task Status	Submission and Running Strategies	Description
1. First Submission	Start immediately, synchronize from the default locus	Under this policy, reading will commence from the default position according to the configuration of the source. If the source is configured to the "full + incremental" reading mode, it

©2013–2025 Tencent Cloud. All rights reserved.

2. Stopped/ Anomaly Detected/ Initialization (Non-first submission)		will by default first synchronize the existing data (full phase), and then it can consume binlog to get the changed data (incremental phase); if the source is configured to "incremental only" reading, it will by default start reading from the latest binlog position.
	Start immediately, synchronize from the designated time point	With this strategy, you need to choose a specific start time and match the locus by time. 1. Start reading data from the designated time point. If the designated locus is not matched, the task will start synchronizing from the earliest binlog locus by default 2. If your source reading mode is 'Full + Incremental', choosing this strategy will skip the full phase and start synchronizing from the designated incremental locus
	Do not start now, manually start the task later in real-time task operation and maintenance	Under this strategy, the task will be submitted to real-time operation and maintenance without starting. The task can be started in bulk from the real-time operation and maintenance page later.
Running (Non-First Submission)	Continue running, retain job status data, continue running from the last synchronized position	Under this strategy, after the new version of the task is submitted, it will continue running from the last synchronized position.
	Restart, continue running from the specified time point	Under this strategy, you can specify the position to start reading. The task will ignore the old version and restart reading from the specified position. If the specified time position is not found, the task will start synchronizing from the earliest binlog position by default.
	Restart, stop the running task and discard the task status, start running from the default position	Under this strategy, the running task will be stopped and the task status will be discarded. The task will then start reading from the default position based on the source-end configuration. If the source-end configuration is "Full + Increment" reading mode, it will synchronize the stock data (Full phase) first, and after completion, it will consume the binlog to get the change data (Increment phase); if the source-end configuration is "Increment Only" reading mode, it will start reading from the latest binlog position by default.
Paused (Non-First Submission)	Continue running, retain job status data, continue running from the last synchronized position	Under this strategy, after the new version of the task is submitted, it will continue running from the last synchronized position. **Note:** • **Pausing** will create a snapshot, and upon resubmission, the task supports resuming from the last checkpoint. • When **forcibly pausing**, no snapshot is created, but upon resubmission, the task can run from the most recent snapshot created during task execution. This type of pause may result in partial data replay. If the target write operation is Append, duplicate data might occur; if the target write operation is Upsert, there won't be a duplication issue.
	Restart, continue running from the specified time point	Under this strategy, you can specify the position to start reading. The task will ignore the old version and restart reading from the specified position. If the specified time position is not found, the task will start synchronizing from the earliest binlog position by default.

©2013-2025 Tencent Cloud. All rights reserved.

	Restart, stop the running task and discard the task status, start running from the default position	Under this strategy, the running task will be stopped and the task status will be discarded. The task will then start reading from the default position based on the source-end configuration. If the source-end configuration is "Full + Increment" reading mode, it will synchronize the stock data (Full phase) first, and after completion, it will consume the binlog to get the change data (Increment phase); if the source-end configuration is "Increment Only" reading mode, it will start reading from the latest binlog position by default.
Failed (Non-First Submission)	Resume operation from the last execution failure (checkpoint) location	Under this policy, the task will continue running from the location of the last execution failure
	Restart, and start running from the default location based on the task reading configuration	Under this policy, it will start reading from the default location based on the source configuration. If the source configuration is set to "Full + Incremental" reading mode, it will first synchronize the stored data (Full phase), and upon completion, consume the binlog to acquire change data (Incremental phase); if the source configuration is set to "Incremental only," it will start reading using the latest binlog location by default.
Operating (Non-First Submission)	Not supported	When there is an existing task with the same name and its status is **Operating**, re-submitting the task is not supported

Subsequent Steps

After completing task configuration, you can perform operations and monitor alerts on the created tasks, such as configuring monitoring alerts and viewing key task performance indicators. For details, refer to Real-time Task Operation and Maintenance.

©2013-2025 Tencent Cloud. All rights reserved.

Supported DML and DDL Operations

Last updated: 2024-11-25 15:57:42

DDL Change Types and Response Policies

Real-time Whole Database Tasks support monitoring data source DDL changes and synchronizing changes from the source to the target in real time. The supported DDL scenarios and corresponding policies are shown in the table below:

DDL	Description
New Table	1. Automatic Table Creation: When a new table is created in the monitored database on the source end, the target end automatically creates a table and fields with the same structure. 2. Ignore Changes: The target end ignores the DDL change messages generated by the source end. The target end and logs do not respond or notify in any form. 3. Log Alerts: The target end only receives DDL change messages and logs the message content without triggering table creation. 4. Task Error: The target end receives DDL change messages and continuously restarts the task. Errors will appear in the task logs during the restart, and data writing exceptions will occur.
Add Column	1. Automatically Add Columns: When a new column appears in the monitored table at the source end, the target end will automatically create a column with the same name. 2. Ignore Changes: The target end ignores the DDL change messages generated by the source end. The target end and logs do not respond or notify in any form. 3. Log Alerts: The target end only receives DDL change messages and logs the message content. This policy does not trigger the addition of columns. 4. Task Error: The target end receives DDL change messages and continuously restarts the task. Errors will appear in the task logs during the restart, and data writing exceptions will occur.
Delete Column	1. Automatic Column Deletion: When a column is deleted in the monitored table at the source end, the target end will automatically delete the column with the same name. 2. Ignore Changes: The target end ignores the DDL change messages generated by the source end. The target end and logs do not respond or notify in any form. 3. Log Alerts: The target end only receives DDL change messages and logs the message content. This policy does not trigger the deletion of columns. 4. Task Error: The target end receives DDL change messages and continuously restarts the task. Errors will appear in the task logs during the restart, and data writing exceptions will occur.
Rename Column	1. Automatic Column Renaming: When a column name changes in the monitored table at the source end, the target end will automatically rename the column. 2. Ignore Changes: The target end ignores the DDL change messages generated by the source end. The target end and logs do not respond or notify in any form. 3. Log Alerts: The target end only receives DDL change messages and logs the message content. This policy does not trigger the renaming of columns. 4. Task Error: The target end receives DDL change messages and continuously restarts the task. Errors will appear in the task logs during the restart, and data writing exceptions will occur.
Modify Column Type	1. Auto Modify Column Type: When there is a change in the column type in the monitored database at the source end, such as a change in field length or adding a comment, the target end will automatically make the corresponding changes. 2. Ignore Changes: The target end ignores the DDL change messages generated by the source end. The target end and logs do not respond or notify in any form. 3. Log Alerts: The target end only receives DDL change messages and logs the message content. This policy does not trigger modification of column types. 4. Task Error: The target end receives DDL change messages and continuously restarts the task. Errors will appear in the task logs during the restart, and data writing exceptions will occur.
Deleting a table	1. Ignore Changes: The target end ignores the DDL change messages generated by the source end. The target end and logs do not respond or notify in any form.

©2013-2025 Tencent Cloud. All rights reserved.

Rename Table Clearing table data	2. Log Alerts: The target end only receives DDL change messages and logs the message content. This policy does not trigger drop table/rename table/clear table operations. 3. Task Error: The target end receives DDL change messages and continuously restarts the task. Errors will appear in the task logs during the restart, and data writing exceptions will occur.

Note
For source-side "high-risk" table-level operations, such as deleting tables, truncating tables, and renaming tables, the target side does not support automatic response to changes, meaning the target side will not automatically perform deletion, truncation, or renaming of tables.

Overview of DML and DDL Support

Whether the target end can automatically synchronize the DML and DDL changes of the source end depends on whether the source is able to capture these changes, and whether the target is able to respond to them. Only when the source end supports and the target end also supports, can the corresponding link support the respective DDL changes. For instance, if the MySQL source end supports **adding columns** and the DLC target end supports **adding columns**, then adding columns in the MySQL source table can automatically add columns in the DLC target end; if the MySQL source end supports **deleting tables**, but the Doris target end does not support **deleting tables**, then tables deleted in the MySQL data source will not be automatically deleted in the Doris target end.
Different source and target ends support varying DML, DDL types, and message processing. The support of some data sources is shown in the following table. For specific information, refer to the product page configuration.

Note
1. Read: "Yes" means the read end can perceive the corresponding DML, DDL operations and pass the changes downstream.
2. Write: "Yes" means the write end can follow and respond to the corresponding DML, DDL changes.
3. DDL change response currently only supports the Incremental Phase; the Full Volume Phase does not support DDL changes.

Type	Read (Source End)										
	DML			DDL							
	Insert	Delete	Update	Add Column	Delete Column	Rename Column	Modify Column Type	Add Table	Rename Table	Deleting a table	Clearing table data
MySQL	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
TDSQL-C MySQL	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
TDSQL MySQL	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Oracle	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
SQL Server	Yes	Yes	Yes	No	No	No	No	No	No	No	No
PostgreSQL	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes
OceanBase	Yes	Yes	Yes	No	No	No	No	No	No	No	No
Kafka	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Type	Write (Target End)	
	DML	DDL

©2013-2025 Tencent Cloud. All rights reserved.

	Insert	Delete	Update	Add Column	Delete Column	Rename Column	Modify Column Type	Add Table	Rename Table	Deleting a table	Clearing table data
DLC	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
Iceberg	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
Doris	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
StarRocks	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	No	No
TChouse-P	Yes	Yes	Yes	Yes	No	No	No	Yes	No	No	No
Elasticsearch	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	No
Hive	Yes	No	No	No	No	No	No	No	No	No	No
Kafka	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

DDL Resolution Syntax Requirements

> **Note:**
> Only SQL statements that meet DDL type and syntax requirements can be parsed and responded to by the source. Unsupported SQL statements may cause tasks to restart abnormally. Please evaluate whether the commonly used syntax for business table structure changes meets the requirements and configure DDL change responses and write exception handling policies that meet business requirements for the task.

MySQL

1. The following syntax is supported for renaming tables (rename table):

```
RENAME TABLE tbl_name TO new_tbl_name
```

2. The following syntax is supported for creating new tables (create table):

```
CREATE TABLE [IF NOT EXISTS] tbl_name (create_definition,...)
CREATE TABLE new_tbl LIKE orig_tbl;
CREATE TABLE new_tbl AS SELECT * FROM orig_tbl;
```

> **Warning:**
> CHECK constraints and Temporary Table are not supported

3. Modify Column and Rename Column operations support the following syntax:

```
ALTER TABLE t1 MODIFY b INT NOT NULL;
ALTER TABLE t1 CHANGE b a INT NOT NULL;
```

4. Add Column / Rename Column operations support the following syntax:

©2013-2025 Tencent Cloud. All rights reserved.

```
ALTER TABLE table_name
  ADD new_column_name column_definition
    [ FIRST | AFTER column_name ];
```

5. Delete Column operations support the following syntax:

```
Alter table t2 DROP [COLUMN] col_name
```

6. Clear Table operations support the following syntax:

```
truncate table a;
```

7. Drop Table operations support the following syntax:

```
drop table a;
```

©2013–2025 Tencent Cloud. All rights reserved.

Single table sync task configuration
Data sources supported by Single Table Synchronization

Last updated: 2024-11-25 16:03:22

The types of data sources supported for real-time single table synchronization are as follows:

Data Source		Real-Time Synchronization	
		Single Table Read	Single Table Write
Relational database	MySQL/TencentDB for MySQL	✓	✓
	TDSQL-C MySQL	✓	✓
	TDSQL MySQL	✓	✓
	PostgreSQL	✓	✓
	SQL Server	✓	✓
	Oracle	✓	✓
	OceanBase	✓	-
	Dameng DM	✓	✓
Big data	Hive	-	✓
	DLC	-	✓
	Doris/TCHouse-D	-	✓
	ClickHouse/TCHouse-C	-	✓
	Iceberg	-	✓
	HBase	-	✓
	HDFS	-	✓
	TBase	-	✓
	Greenplum	-	✓
	TCHouse-P	-	✓
NoSQL	Elasticsearch	-	✓
	MongoDB/Tencent Cloud MongoDB	✓	-
Message Queue	Kafka/CKafka	✓	✓
	TiDB-Kafka	✓	-
	DTS-kafka	✓	-

©2013-2025 Tencent Cloud. All rights reserved.

Overview of Single Table Task Configuration

Last updated: 2024-11-25 16:03:59

Background

Real-time single table synchronization monitors and synchronizes source data in real-time through binlog logs. Single table synchronization uses fixed field synchronization, synchronizing only the source fields specified in the task configuration to the target. Single table tasks support both Canvas and Form modes, covering data sources like MySQL, Kafka, MongoDB, SQL Server, Doris, and DLC.

Conditions and Restrictions

1. Data sources for the source and target ends have been completed for subsequent tasks. For details, see Data Source Management and Configuration Methods.
2. DataInLong resource group has been purchased. For details, see Configure Integrated Resource Group.
3. Network connectivity between the DataInLong resource group and data sources has been completed. For details, see Integrated Connectivity and Usage Plan.
4. Data source environment preparation has been completed. You can grant the database account configured for data sources the necessary permissions for corresponding operations before executing the sync task, based on your sync configuration requirements.
5. If the database account configured for the data source lacks read-write permissions, the task will fail. Please configure an account with the appropriate permissions according to the actual read-write scenario.

Operation step

Step 1: Create a Real-time Sync Task

After entering the **DataInLong > Real-time Sync** task page, click **New** Single Table Sync Task.
Configure the basic task information in the pop-up window, and after clicking **OK**, you will enter the task configuration page. The task creation parameters and descriptions are as follows:

Parameter	Description
Task Name	• Required Items • Naming rules: Supports only Chinese, English, digits, and underscores • Length: Cannot exceed 100 characters
Task Mode	• Form Mode: Suitable for single-table to single-table offline synchronization, applicable for data synchronization in the ODS layer without data cleaning steps. Default option. • Canvas Mode: Mainly adopts a visual drag-and-drop method, suitable for data pipelines that include cleaning steps. • Form Mode: Suitable for single-table to single-table offline synchronization, applicable for data synchronization in the ODS layer without data cleaning steps
Description	Optional items

Note
- In the current version, Canvas Mode only supports one write node and two read nodes; a single read node can only connect to one downstream node.
- The task must not contain any unconnected isolated nodes, otherwise the task submission will fail.
- The default naming convention for data nodes is: ${node type}_${code}.

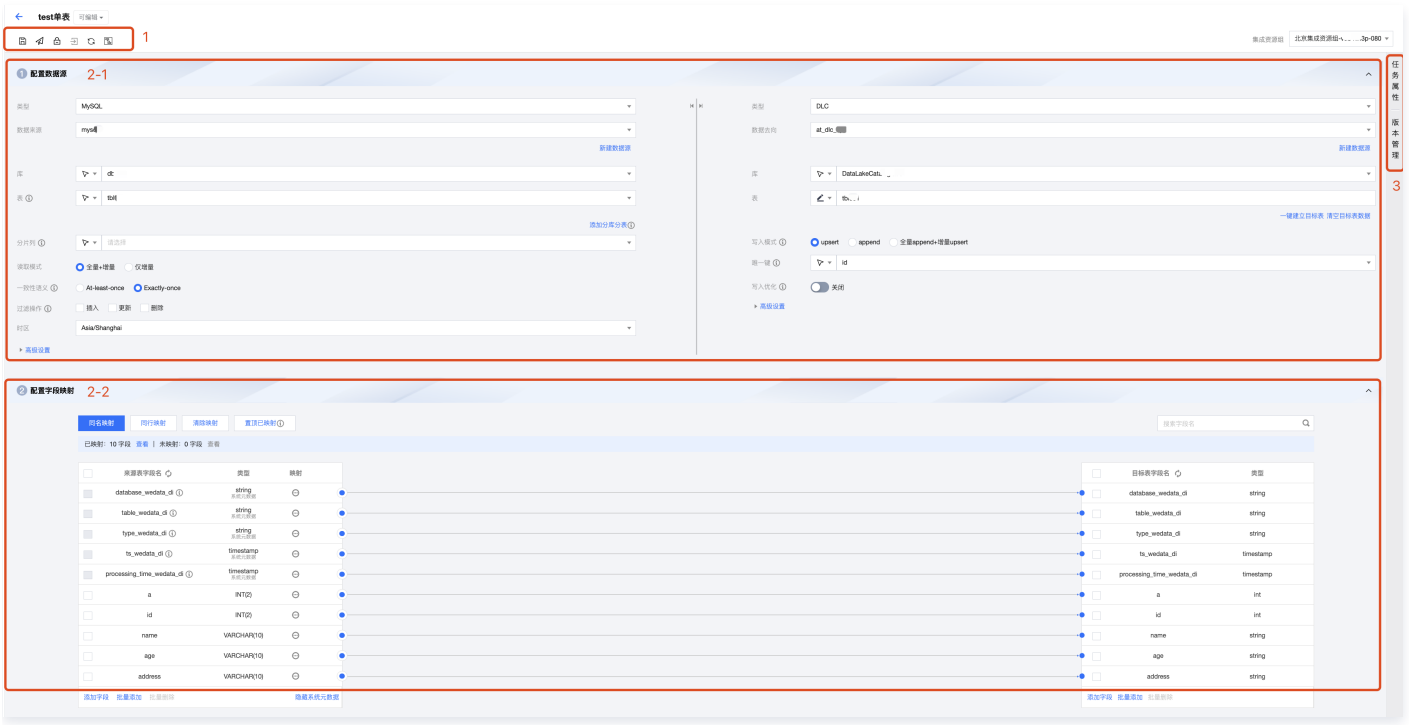
Step 2: Choose the Task Configuration Mode

Currently, task configuration offers two modes: Form and Canvas modes:

©2013-2025 Tencent Cloud. All rights reserved.

- Form Mode is suitable for raw data synchronization and supports data conversion using Flink built-in functions.
- Canvas Mode provides transition nodes and supports the addition of customized complex data transformations during the data synchronization process.

Configuration Mode 1: Form Configuration

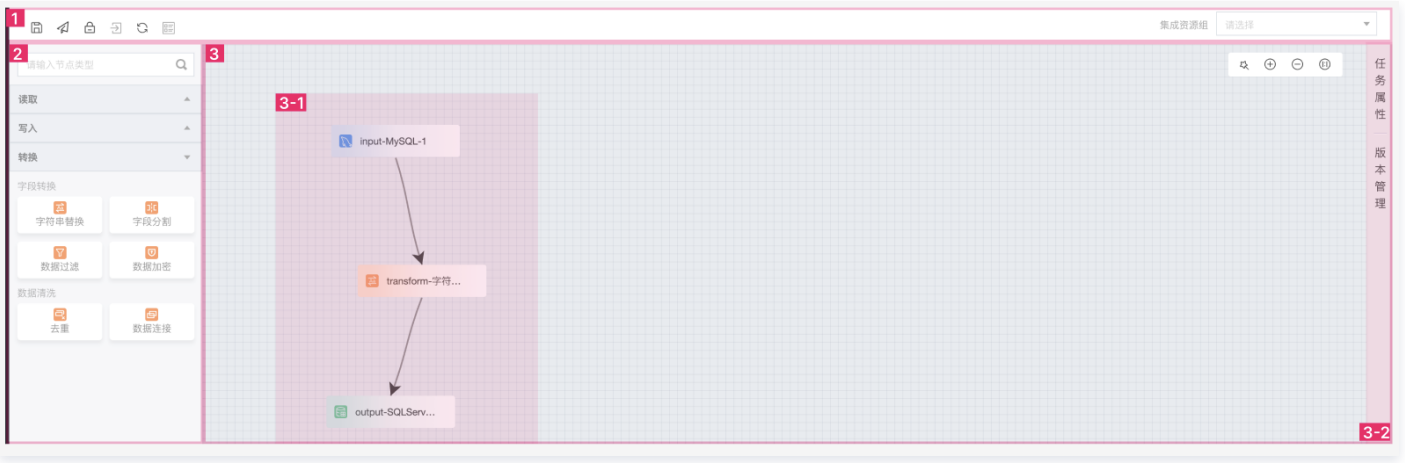


Serial number		Parameter description
1		Task Operation Bar. Operations affecting the entire task, including Save, Submit, Lock/Unlock, Go to Maintenance, Refresh, and Canvas Conversion.
2	2-1	Data source and target: Configure the data source, library, table, and read-write mode for the task.
	2-2	Field Mapping: Set the correspondence between source and target data. Subsequent tasks will only sync data between mapped fields. **Note** Some non-relational databases cannot automatically obtain field information, and fields need to be added individually or in batches manually.
3		Task Configuration: This configuration applies to the global task: • Real-time tasks include task attributes, version management configuration, providing task resource parallelism, CU usage configuration, etc.

Configuration Mode 2: Canvas Mode

In the list, click **Task Name** to enter the task configuration page, which generally includes three parts: Task Operation Bar, Data Node Menu, and Link Configuration Area:

©2013-2025 Tencent Cloud. All rights reserved.



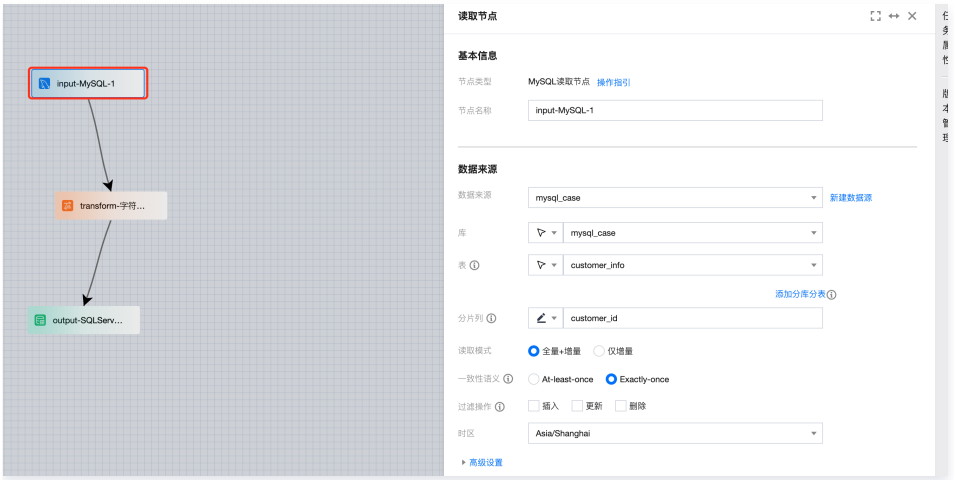
Serial number		Parameter description
1		Task Operation Bar: Operations that affect the entire task, including Save, Submit, Lock/Unlock, Go to Operation and Maintenance, Refresh, Form Conversion, etc.
2		Data Node Menu: Categorized by link object into read, write, and transition nodes. It supports adding nodes directly to the canvas via drag and drop.
3	3-1	Data Link: Composition of read, write, and transition nodes and the connections between them, representing the data flow direction within the synchronization task.
	3-2	Task Configuration: This configuration applies to the global task: • Real-time tasks include task attributes, version management configuration, providing task resource parallelism, CU usage configuration, etc.

Step 3: Data Node Configuration

After creating a new task, click **Task Name** in the task list to enter the Task Canvas or Form Interface. In Canvas Mode, you can drag data nodes and connections from the right-side node menu. The system will automatically create a Data Flow Chain based on node connections.

Configure Read Node

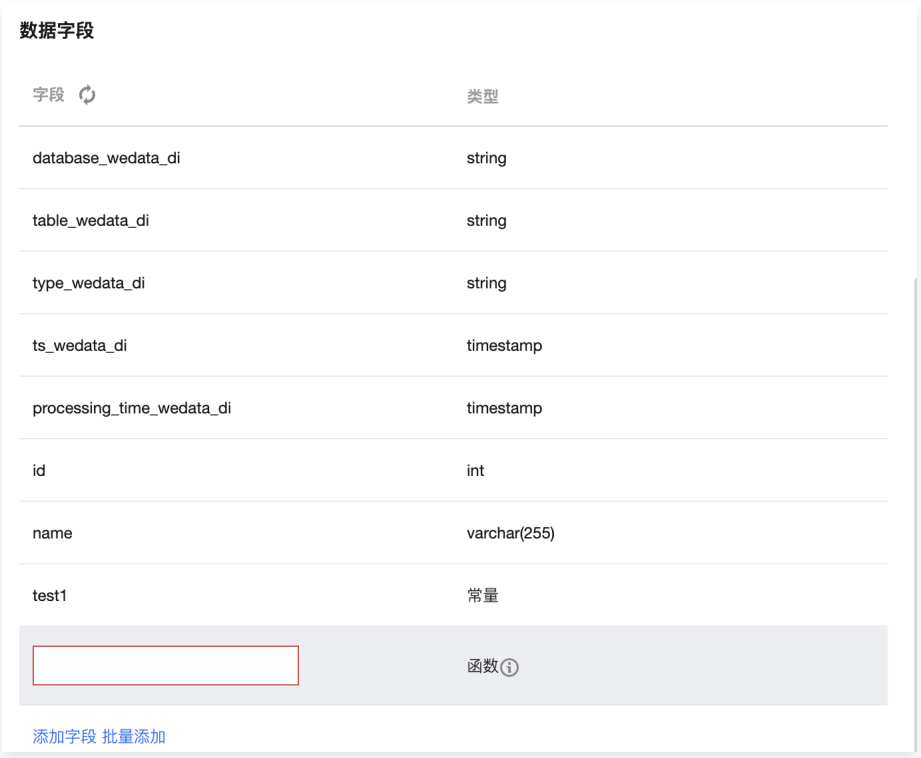
Read Node Configuration includes Basic Information, Data Source, and Data Fields.



- Basic Information
  The node name cannot be empty, and duplicate data nodes are not allowed within a single task.
- Data Source
  Configure the database table object to read and synchronization method information.

©2013-2025 Tencent Cloud. All rights reserved.

- Data Fields
  Based on the configured database table object, the system supports default metadata retrieval and manual field configuration.
  - Default Retrieval: For types such as MySQL, Hive, PostgreSQL, the system supports automatic retrieval of metadata fields and types based on their database information without manual editing.
  - Manual Configuration: For data sources like files (e.g., HDFS, COS) and columnar storage data sources (e.g., HBase, Mongo) that do not support automatic metadata retrieval, you can click Add Field to manually add field names and types. The reading node also supports constants and function types.



> **Note**
> - Constant Field: Only read nodes support configuring constant fields. Constant fields can be used to write a fixed constant value into the target table in case of a mismatch in the number of fields between the source and target tables.
> - Function Field: Only reading nodes support function constant fields. Function fields can generate new fields based on source fields, which are then written to the target table. For specific supported Flink functions, please refer to the Flink documentation.

Configure Transition Node (optional)

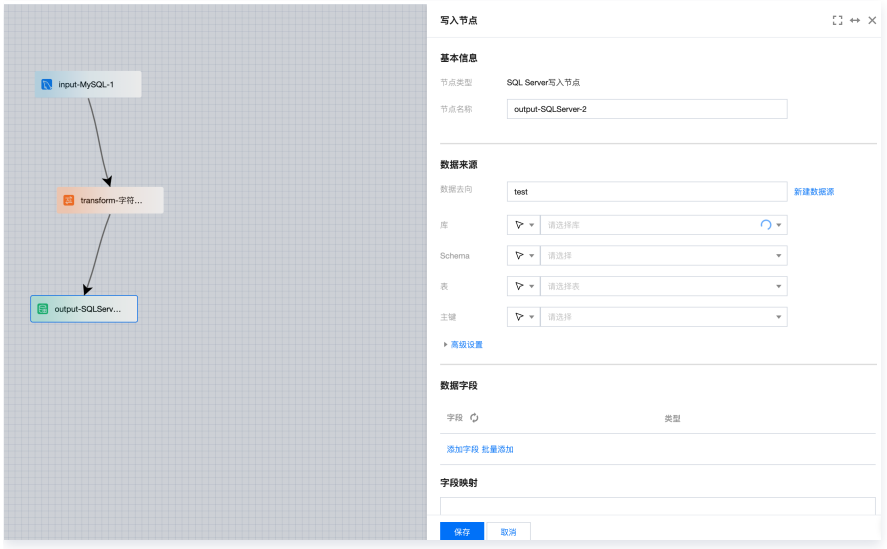
Transition node configuration includes basic information, conversion rules, and data fields. The transition node must be downstream of a read node. After connecting to a read node, the system will automatically retrieve field information from the upstream node and complete data conversion based on conversion rules.

- Basic Information
  Configure node name information. The node name cannot be empty, and duplicate data nodes are not allowed within a single task.
- Conversion Rules
  Configure field or data-level conversion rules. Field information is inherited from upstream nodes. After connecting to upstream nodes, the system automatically retrieves field information from the upstream nodes.
- Data Fields
  By default, all data fields from the upstream node are retrieved for subsequent write node mapping.

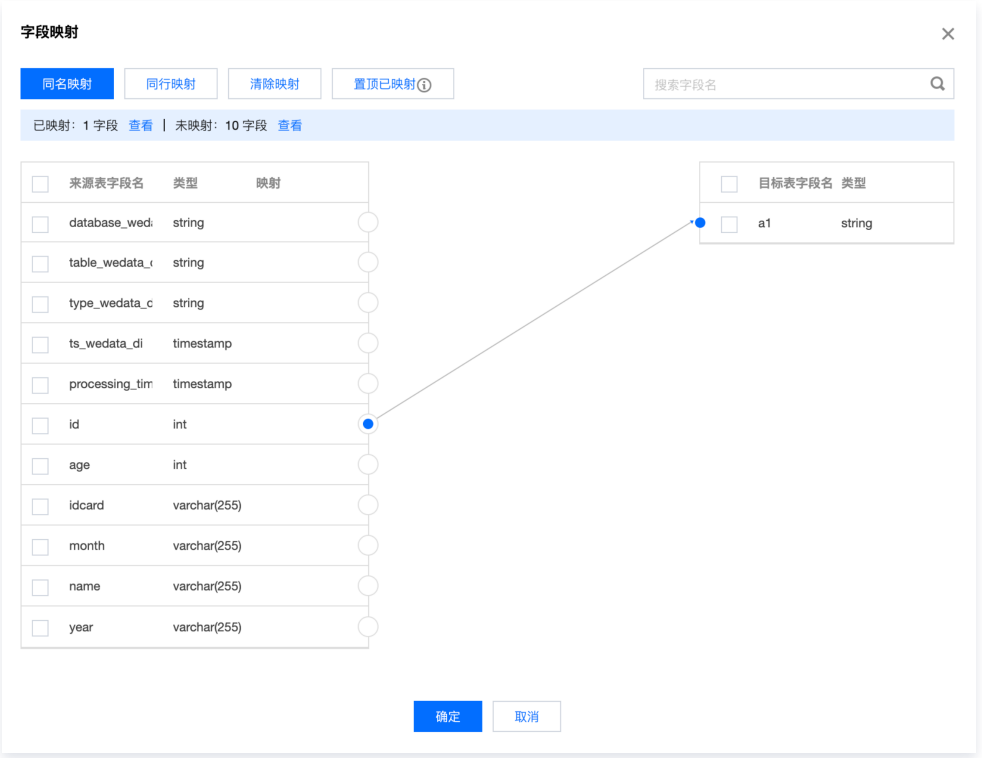
Configure Write Node

Write node configuration includes basic information, data source, data fields, and field mapping. The write node will write upstream data content to the target object according to the connection relationship.

©2013–2025 Tencent Cloud. All rights reserved.



- Basic Information
  The node name cannot be empty, and duplicate data nodes are not allowed within a single task.
- Data Source
  Configure the database table object to read and synchronization method information.
- Data Fields
  Based on the configured database table object, the system supports default metadata retrieval and manual field configuration.
  - Default Retrieval: For types such as MySQL, Hive, PostgreSQL, the system supports automatic retrieval of metadata fields and types based on their database information without manual editing.
  - Manual Configuration: For data sources like files (e.g., HDFS, COS) and columnar storage data sources (e.g., HBase, Mongo) that do not support automatic metadata retrieval, you can click **Field Configuration** to manually add field names and types.
- Field Mapping
  An additional field mapping relationship needs to be configured for the write node relative to the read node. The purpose of the field mapping relationship is to specify the source of the target field content through connections, and it supports same name mapping, peer mapping, and manual connection methods for configuring the relationship between the source and target nodes.



©2013–2025 Tencent Cloud. All rights reserved.

> **Note**
> - The prerequisite for configuring field mapping is that the current write node has a connected source (read node or transition node).
> - The content of target fields without a configured mapping relationship will be empty.
> - If there is no convertible relationship between the source field type and the target field type, it may cause task failure.

Step 4: Real-time Task Attribute Configuration

The real-time task attribute configuration includes **basic attributes** and **resource configuration** in two parts:



Category	Parameter	Description
Task Attributes	Task Name/Type	Displays basic information of the current task name and type.
	Responsible Person	The name of one or more space members responsible for this task, defaulting to the task creator.
	Description	Displays current task remark information.
Resource configuration	Integrated Resource Group	Specifies the name of the integrated resource group used by the current task; only one resource group can be bound to a task.

©2013–2025 Tencent Cloud. All rights reserved.

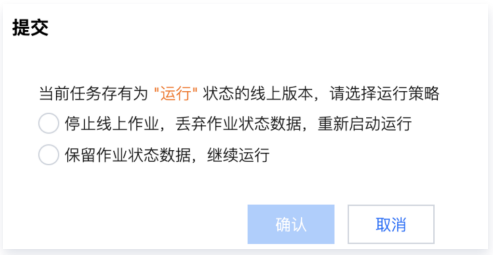
	Resource Allocation Method	Integrated resources support multiple allocation methods: • Fixed Allocation: In this method, resource allocation remains constant throughout the full and incremental synchronization phases, preventing resource preemption among tasks. It is suitable for scenarios where data may significantly change during task execution. • Phase-Based Allocation: Allocates resources based on different synchronization phases (full and incremental) to optimize overall resource usage. 
	JobManager	Supports 0.25, 0.5, 1, 2C. Once set, tasks will default to this specification. CU Task Actual Usage = JobManager Specification + TaskManager Specification × Concurrency.
	TaskManager	Supports 0.25, 0.5, 1, 2CU. Once set, tasks will default to this specification. CU Task Actual Usage = JobManager Specification + TaskManager Specification × Concurrency.
	Concurrency	Default concurrency for each operator.
Execution Policy	Checkpoint Interval	Set checkpoint interval time, supporting minute/second level.
	Maximum Restart Times	Set the maximum restart threshold for the task when a failure occurs during execution. If the number of restarts exceeds this threshold, the task status will change to **Failed**. The range is [-1,100], with 0 meaning no restart and -1 indicating no limit on the number of restarts.
	Associate alarm	Supports associating alarm rules with the current task

Step Five: Task Submission

Real-time synchronization tasks can configure operation strategies and be submitted to the production environment after configuration is completed. Currently, the task configuration page supports save, submit, lock/unlock, go to operations, and form/canvas conversion operations.

Serial number	Parameter	Description
1	Save	Save current task configuration information, including data node configuration, node connection, and task attribute configuration.
2	Submit	Submit the current task to the production environment. At the time of submission, different operation strategies can be selected based on whether there is an active production task associated with the current task. • If there are no effective online tasks for the current task, i.e., it's the first submission or the online task is in a "failed" status, you can submit it directly. • If there are "Running" or "Paused" online tasks, you need to choose different strategies. Stopping the online job will abandon the previous task's running location, consuming data from the beginning. Retaining the job status will resume from the last consumed location after restarting.

©2013-2025 Tencent Cloud. All rights reserved.

		 **Note** Clicking to start the task immediately will make it run immediately after submission. Otherwise, it will require manual trigger to start officially.
3	Lock/Unlock	The default creator is the first lock holder, and only the lock holder can edit task configurations and run tasks. If the lock holder does not perform any editing operations within 5 minutes, others can click the icon to acquire the lock. If successful, they can perform editing operations.
4	Go to Operations	Quickly jump to the real-time operations page based on the current task name.
5	Canvas conversion/Form conversion	Form Mode and Canvas Mode can be converted to each other. The form mode provides a canvas conversion feature, and the canvas mode provides a form conversion feature. **Note** • Currently, conversion is only allowed if the task is in a successfully saved state. • If the canvas mode contains transition nodes, conversion to form mode is not supported.

Task Configuration Detection

When submitting a task, basic environment and configuration checks will be performed. Pay close attention to the failed items and warning items. Failed items are likely to cause task execution failure, while warning items should be judged based on the business situation. The current check results will not block task submission, and you can choose to recheck or ignore the exceptions and continue submitting.



©2013–2025 Tencent Cloud. All rights reserved.

Currently supported inspection items:

Check Classification	Check Items	Description
Task Configuration Detection	Source Configuration	Check if any required fields are missing on the source side
	Target Configuration	Check if any required fields are missing on the target side
	Mapping Configuration	Check if field mappings are configured
	Resource Group Configuration	Check if the resource group is configured
	Same data table detection	Detect whether all real-time single-table tasks under the current project (including submitted and unsubmitted) have the same source table or target table or both the source and target tables are the same. The basis for judging the same table is whether it is a table with the same name under the same data source and the same DB. This inspection item is mainly used for scenarios where multiple tasks are not expected to repeatedly read the same table.
Data Source Detection	Source end connectivity detection	Check if the data source at the source end is networked with the resource group in the task configuration. If the check fails, view diagnostic information. After resolving network issues, recheck; otherwise, the task is likely to fail.
	Target end connectivity detection	Check if the data source at the target end is networked with the resource group in the task configuration. If the check fails, view diagnostic information. After resolving network issues, recheck; otherwise, the task is likely to fail.
Resource Detection	Resource status detection	Check if the resource group is in an available status. If the resource status is unavailable, please switch the resource group in the task configuration; otherwise, the task is likely to fail.
	Resource Capacity Detection	Check if the remaining resources in the resource group meet the resource requirements of the task configuration. If the check fails, please reduce the task resource configuration or expand the resource group.

Submission Policy Selection

After normal submission or ignoring exception submission, you will enter the submission policy selection page, where you can choose different operation strategies based on business needs:

©2013-2025 Tencent Cloud. All rights reserved.



The submission operation strategy supported by different task operation statuses varies:

Task Status	Submission and Running Strategies	Description
1. First Submission 2. Stopped/ Anomaly Detected/ Initialization (Non-first submission)	Start immediately, synchronize from the default locus	Under this policy, reading will commence from the default position according to the configuration of the source. If the source is configured to the "full + incremental" reading mode, it will by default first synchronize the existing data (full phase), and then it can consume binlog to get the changed data (incremental phase); if the source is configured to "incremental only" reading, it will by default start reading from the latest binlog position.
	Start immediately, synchronize from the designated time point	With this strategy, you need to choose a specific start time and match the locus by time. 1. Start reading data from the specified time point. If the specified position is not matched, the task will default to synchronizing from the earliest binlog position 2. If your source read mode is **Full + Increment**, selecting this strategy will default to skipping the full stage and starting synchronization from the specified increment time position
	Do not start now, manually start the task later in real-time task operation and maintenance	Under this strategy, the task will be submitted to real-time operation and maintenance without starting. The task can be started in bulk from the real-time operation and maintenance page later.
Running (Non-First Submission)	Continue running, retain job status data, continue running from the last synchronized position	Under this strategy, after the new version of the task is submitted, it will continue running from the last synchronized position.
	Restart, continue running from the specified time point	Under this strategy, you can specify the position to start reading. The task will ignore the old version and restart reading from the specified position. If the specified time position is not found, the task will start synchronizing from the earliest binlog position by default.
	Restart, stop the running task and discard the task status, start running from the default position	Under this strategy, the running task will be stopped and the task status will be discarded. The task will then start reading from the default position based on the source-end configuration. If the source-end configuration is "Full +

©2013-2025 Tencent Cloud. All rights reserved.

		Increment" reading mode, it will synchronize the stock data (Full phase) first, and after completion, it will consume the binlog to get the change data (Increment phase); if the source-end configuration is "Increment Only" reading mode, it will start reading from the latest binlog position by default.
Paused (Non-First Submission)	Continue running, retain job status data, continue running from the last synchronized position	Under this strategy, after the new version of the task is submitted, it will continue running from the last synchronized position. **Note:** • **Pausing** will create a snapshot, and upon resubmission, the task supports resuming from the last checkpoint. • When **forcibly pausing**, no snapshot is created, but upon resubmission, the task can run from the most recent snapshot created during task execution. This type of pause may result in partial data replay. If the target write operation is Append, duplicate data might occur; if the target write operation is Upsert, there won't be a duplication issue.
	Restart, continue running from the specified time point	Under this strategy, you can specify the position to start reading. The task will ignore the old version and restart reading from the specified position. If the specified time position is not found, the task will start synchronizing from the earliest binlog position by default.
	Restart, stop the running task and discard the task status, start running from the default position	Under this strategy, the running task will be stopped and the task status will be discarded. The task will then start reading from the default position based on the source-end configuration. If the source-end configuration is "Full + Increment" reading mode, it will synchronize the stock data (Full phase) first, and after completion, it will consume the binlog to get the change data (Increment phase); if the source-end configuration is "Increment Only" reading mode, it will start reading from the latest binlog position by default.
Failed (Non-First Submission)	Resume operation from the last execution failure (checkpoint) location	Under this policy, the task will continue running from the location of the last execution failure
	Restart, and start running from the default location based on the task reading configuration	Under this policy, reading will commence from the default position according to the configuration of the source. If the source is configured to the "full + incremental" reading mode, it will by default first synchronize the existing data (full phase), and then it can consume binlog to get the changed data (incremental phase); if the source is configured to "incremental only" reading, it will by default start reading from the latest binlog position.
In Operation (Non-First Submission)	Not supported	When there is an existing task with the same name and its status is **Operating**, re-submitting the task is not supported

Each submission will generate a new version of the real-time task. You can configure the version description in the dialog box.

Submission Result

After clicking the **Submit** button above, the submission progress and results will appear:

©2013-2025 Tencent Cloud. All rights reserved.



- Task Submitting:
  - Shows the submission progress percentage.
- Task Submission Result – Successful:
  - Show task submission successful result.
  - You can go to Operations for task operation and maintenance management or return to the Task List
- Task submission result – Failed:
  - Show reasons for task submission failure

Subsequent Steps

After completing the task configuration and submission, you can perform operation and maintenance as well as monitoring and alarm for the created task. This includes configuring monitoring alarms for the task and viewing key indicators of task operation. For details, please refer to Real-time Operation and Maintenance.

©2013–2025 Tencent Cloud. All rights reserved.

Conversion Node Configuration

Last updated: 2024-08-24 20:55:58

Single Table Real-time Synchronization supports three types of data nodes: input, output, and conversion. This article will focus on the configuration of each read, write, and conversion node and the underlying data type mapping.

Read-Write Node Configuration

Supported Data Sources		Real-Time Synchronization	
		Single Table Read Node Configuration	Single Table Write Node Configuration
Relational database	MySQL	✓	✓
	TDSQL-C MySQL	✓	✓
	PostgreSQL	✓	✓
	TCHouse-P	-	✓
	SQL Server	✓	✓
	Oracle	✓	✓
	Dameng DM	✓	-
Big data	Hive	-	✓
	DLC	-	✓
	Doris	-	✓
	ClickHouse	-	✓
	Iceberg	-	✓
	HBase	-	✓
	HDFS	-	✓
	TBase	-	✓
	Greenplum	-	✓
NoSQL	Elasticsearch	-	✓
	Mongo	✓	-
Message Queue	Kafka	✓	✓
	TiDB-Kafka	✓	-
	DTS-kafka	✓	-

Conversion Node Configuration

Field Conversion Related Nodes

String Replacement

The String Replacement Node primarily uses Java regular expressions to match and replace the contents of string fields. This node will perform the replacement operation when a string is partially or fully matched.
The conversion rule parameters are described as follows:

©2013-2025 Tencent Cloud. All rights reserved.

Parameter	Description
Field	Select the fields that need string matching. The fields processed by this node must be of string type. **Note** In this conversion node, multiple conversion rules can be configured for the same field. These rules will be executed serially based on the configuration order.
Regular expression	Java regular expression used to match the original field content.
New string	Replace the matched field content with the specified new string.
Replacement method	Supports replacing only the first match result or all matches.
Delete	Delete this rule.

Field splitting

The field split node is mainly used when the content of the original field is split using a specific symbol. The content after the split will be written into a new field, and the original content remains unchanged. This node is commonly used in scenarios where one-to-many mapping of field content is needed, such as splitting the field A containing "Tencent-cloud" with "-" and writing the results into two new fields B and C respectively.
The conversion rule parameters are explained as follows:

Parameter	Description
Field	Select the field that requires content splitting. **Note** In this conversion node, multiple conversion rules can be configured for the same field.
Separator	Enter the separator symbol. **Note** The separator can be filled through a drop-down menu or manually. You can select built-in separator symbols from the drop-down menu or manually fill in a custom separator.
Result field name	Enter the field name that will receive the split results. Multiple field names should be separated by commas. The split results will be filled into target fields sequentially by default. If more target field names are provided than split results, the extra fields will default to empty values. Conversely, any extra split content will be discarded. **Note** If field A (containing "Tencent-cloud") is split using "-" and the results are written into fields B, C, and D, then fields B and C will be filled with 'Tencent' and 'cloud', respectively, while the content of field D will be the default empty value.
Delete	Delete this rule.

Note
Currently, this node supports real-time sync tasks only.

Data cleaning related nodes

©2013-2025 Tencent Cloud. All rights reserved.

Data filtering

Each row of the table is filtered and matched based on the filtering rules. For matched data rows, you can choose to retain or remove them.
The parameters of the conversion rules are as follows:

Parameter	Description
Filter action	• Retain: Write the data that matches the filtering rules into the target table. • Remove: Do not write the data that matches the filtering rules into the target table.
Field	Select the field that requires content splitting. **Note** In this conversion node, multiple conversion rules can be configured for the same field.
Logical operators	Supports AND, OR
Operators	Currently supports >, <, <=, >=, =, !=, is null, and is not null
Type	Type of value for comparison with the field content: • Field: Use the content of the specified field to filter and compare. Typically, field content varies with different data rows. • Custom value: Use a fixed constant to compare with the filtering field content.
Comparison value	Select a field or enter a custom value.
Delete	Delete this rule.

Note
Currently, this node supports real-time sync tasks only.

Deduplication

Perform content deduplication on data within the window based on the processing time (process_time) of real-time data.
The parameter description for the conversion rules is as follows:

Parameter	Description
Deduplication action	• Retain the first entry: When duplicate data exists in the time window, retain the first entry in chronological order. • Retain the last one: When duplicate data exists in the time window, retain the last entry in chronological order.
Deduplication field	Specify the field used to determine if content is duplicated. By default, all fields are used if none are specified.
Delete	Delete this rule.

Note
Currently, this node supports real-time sync tasks only.

Data connection (join)

Real-time data stream join; currently supports regular join. Regular join is suitable for bounded input streams, retaining all states by default. The join node requires two input nodes by default.
The parameter description for the conversion rules is as follows:

©2013-2025 Tencent Cloud. All rights reserved.

Parameter	Description
Left/Right Table	Select the node name from the upstream input nodes in the join to be used as the left table object
Left/Right table associated key	The field names used for join association in the left and right tables
Connection methods	Supports left join, right join, and full join

©2013–2025 Tencent Cloud. All rights reserved.

Configuration of log collection task
Log Collection Configuration Overview

Last updated: 2024-08-24 20:56:31

Background

Log collection actively reports CVM cloud instance, self-built server, or TKE log file data to the external target end via Agent or SDK. The Agent is a lightweight log collector provided by InLong, which can be automatically installed and run within Tencent Cloud TKE, CVM, and other cloud cluster services, actively reporting specified file data and synchronizing them to the target end in real-time. Additionally, InLong provides Java and C++ SDKs for data reporting.

Conditions and Restrictions

1. Data source for both source and target ends has been completed for subsequent task usage. For details, see Data Source Management and Configuration Methods.
2. DataInLong resource group has been purchased. For details, see Configure Integrated Resource Group.
3. Network connectivity between DataInLong resource group and data source has been completed. For details, see Integration Connectivity and Usage Planning.
4. Data source environment preparation has been completed. You can grant the database account configured for data sources the necessary permissions for corresponding operations before executing the sync task, based on your sync configuration requirements.
5. If the database account configured for the data source lacks read-write permissions, the task will fail. Please configure an account with the appropriate permissions according to the actual read-write scenario.

Operation step

Step 1: Create a Collector

A collector is a prerequisite for TKE and CVM data source types. Before creating a collection task, you need to create an available collector in advance. Users can create and view it in **Settings Management** > **Collector Management**. For details, see Collector Management.



Step 2: Create a Sync Task

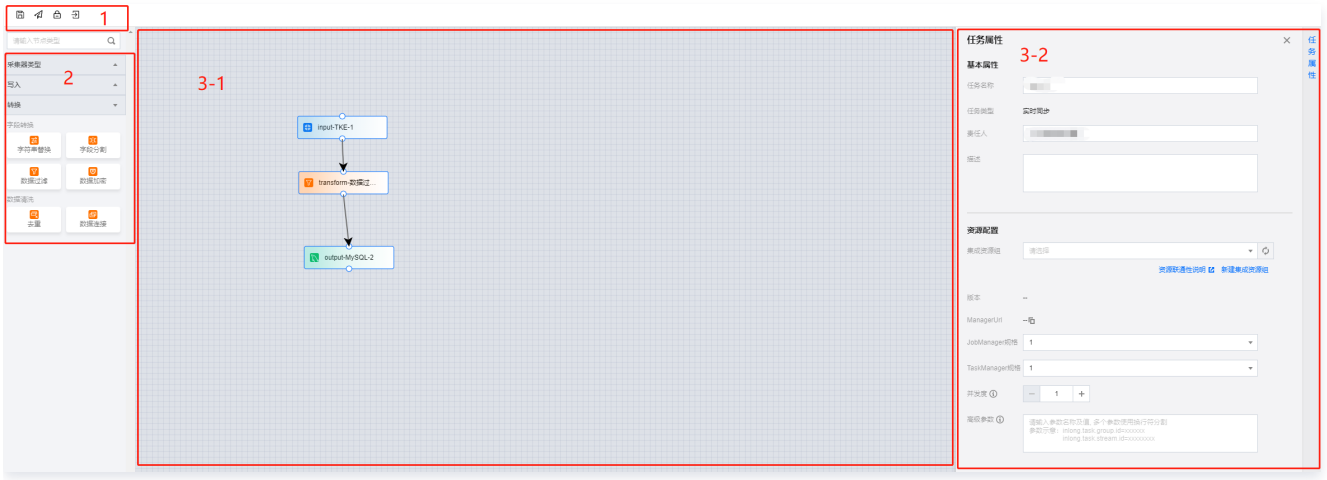
Go to **Configuration Center** > **Real-Time Sync Tasks** page, click **New** to create a new log collection task. Enter the task name and choose the configuration mode.
Currently, task configuration offers two modes: Form and Canvas modes:

- Form Mode is suitable for staging layer data synchronization and only supports using source functions for data transformation.
- Canvas Mode provides transition nodes and supports the addition of customized complex data transformations during the data synchronization process.

Configuration Mode 1: Canvas Mode

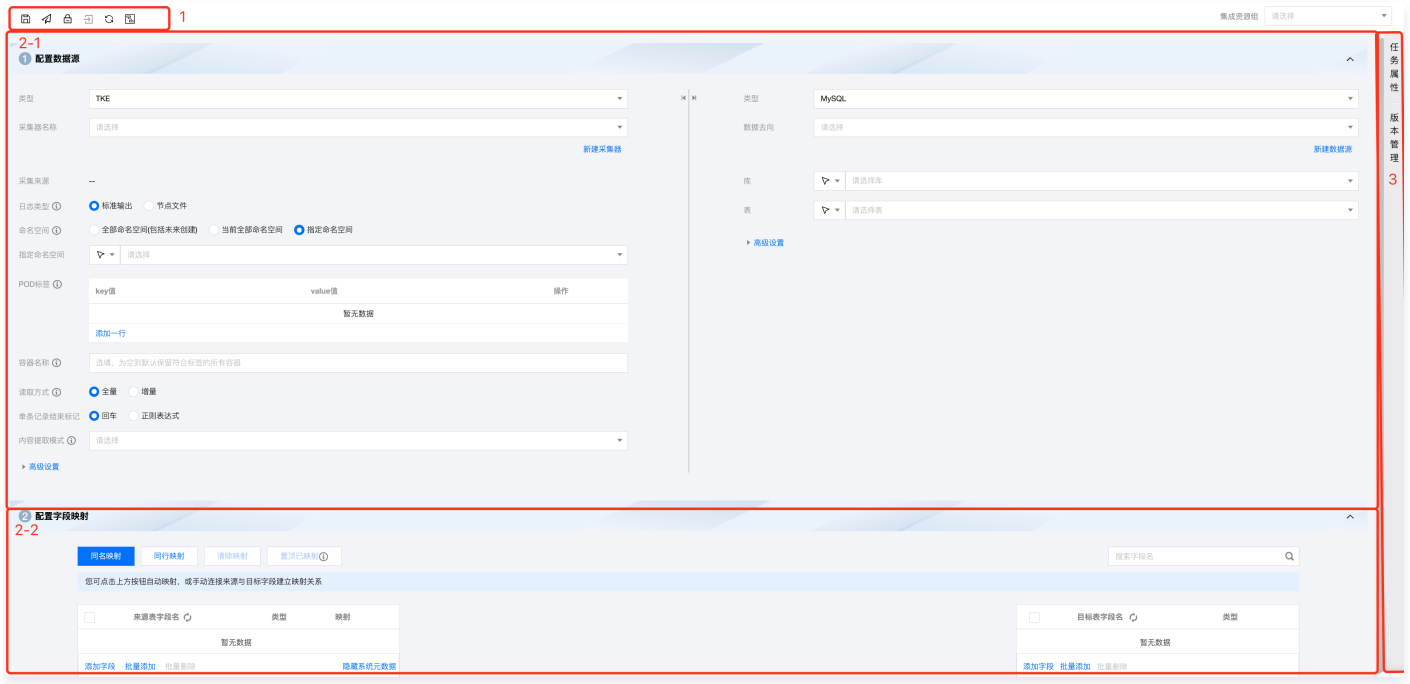
In the list, click **Task Name** to enter the task configuration page, which generally includes three parts: Task Operation Bar, Data Node Menu, and Link Configuration Area:

©2013-2025 Tencent Cloud. All rights reserved.



Serial number		Parameter description
1		Task Operation Bar: Operations that affect the entire task, including Save, Submit, Test Run, Stop, Unlock, and Go to Maintenance.
2		Data Node Menu: Categorized by link object into read, write, and transition nodes. It supports adding nodes directly to the canvas via drag and drop.
3	3–1	Data Link: Composition of read, write, and transition nodes and the connections between them, representing the data flow direction within the synchronization task.
	3–2	Task attribute configuration, which affects the global task, primarily includes basic attributes and resource configuration.

Configuration Mode 2: Form Configuration



Serial number	Parameter description
1	Task Operation Bar: Operations that affect the entire task, including Save, Submit, Test Run, Stop, Unlock, and Go to Maintenance.

©2013–2025 Tencent Cloud. All rights reserved.

2	2-1	Data source and target: Configure the data source, library, table, and read-write mode for the task.
	2-2	Field Mapping: Set the correspondence between source and target data. Subsequent tasks will only sync data between mapped fields.
3		Task attribute configuration, which affects the global task, primarily includes basic attributes and resource configuration.

Step 3: Configure Data Source

The steps for different data sources vary slightly, as detailed below:

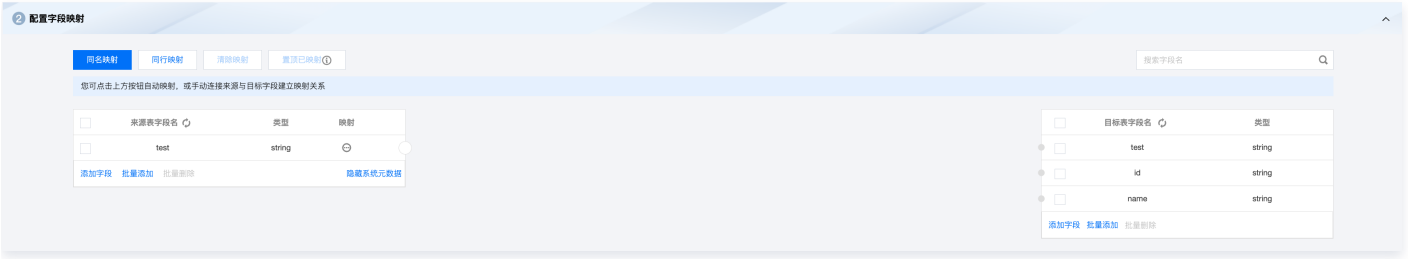
- TKE Source Collection Task Configuration
- SDK Source Collection Task Configuration
- CVM Source Collection Task Configuration

Step 4: Configure Data Target

Data Source	Supported data target
TKE / SDK / CVM	MySQL Write Node
	TDSQL-C for MySQL Write Node
	PostgreSQL Write Node
	SQL Server Write Node
	Oracle Write Node
	Hive Write Node
	ClickHouse Write Node
	Greenplum Write Node
	TBase Write Node
	DLC Write Node
	HBase Write Node
	Iceberg Write Node
	HDFS Write Node
	Doris Write Node
	Elasticsearch Write Node
	Kafka Write Node

Step 5: Configure Field Mapping

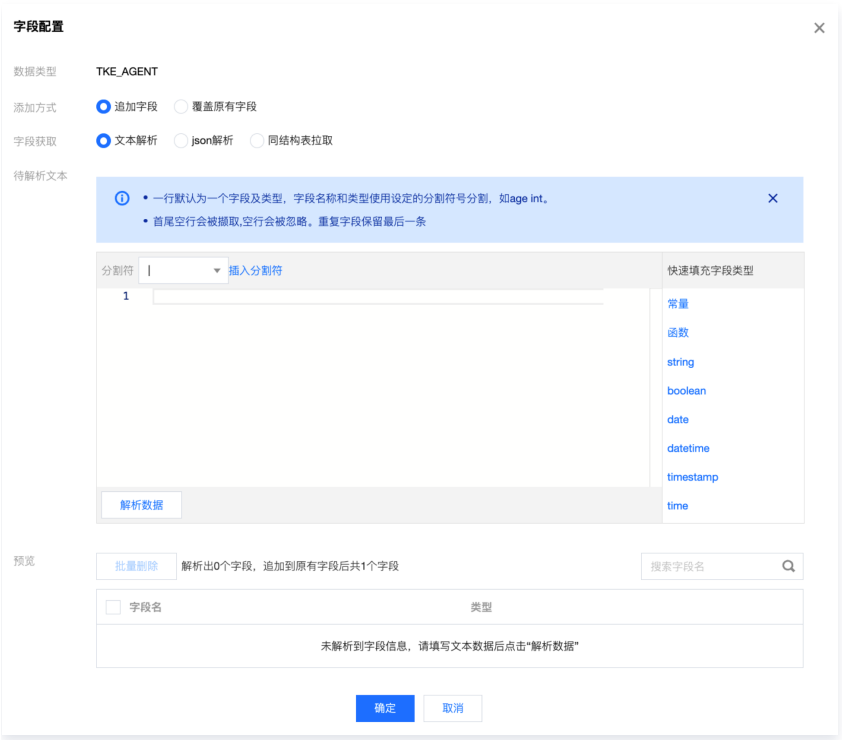
After configuring the data source and data target, the source table and target table field information will be displayed. We need to map the fields using Same Name Mapping and Peer Mapping. Fields can also be sorted and configured.



Click Add Field to configure the fields: You can change the field name and type, as well as delete and add new fields.

©2013-2025 Tencent Cloud. All rights reserved.

Click Bulk Add:



Parameter		Description
Data Type		The data source type of the current node.
Addition Method		• Append Field: Appends the new parsed fields after the original fields of the table. • Overwrite Original Field: The new parsed fields overwrite the original fields in the current source table.
Field Extraction		• Text Parsing: Parses based on the text content, with each line considered as a default field and type. Field names and types are separated by the specified delimiter, e.g., age int. Note: Leading and trailing empty lines will be captured; empty lines will be ignored. • JSON Parsing: Input JSON content and quickly parse the content based on key/value pairs, such as {"age":10,"name":"demo"}. The current system only supports parsing certain types. You can confirm and adjust the parsing results in Form Mode. Duplicate fields will retain the last entry. • Pull from Structural Table: Specify a table object of a data source and parse the fields of that table.
Text to be Parsed	Separator	Used to separate field name and type. Supports tab, |, space, such as "age|int".
	Quick Fill Field Type	Common field types, supporting constants, functions, variables, string, boolean, date, datetime, timestamp, time, double, float, tinyint, smallint, tinyint unsigned, int, mediumint, smallint unsigned, bigint, int unsigned, bigint unsigned, double precision, tinyint(1), char, varchar, text, varbinary, blob.
	Parse Data	Parse the input content.
Preview	Batch delete SNAT rules	After selecting the preview list, batch delete the parsing results.
	Field name	Field Name.

©2013–2025 Tencent Cloud. All rights reserved.

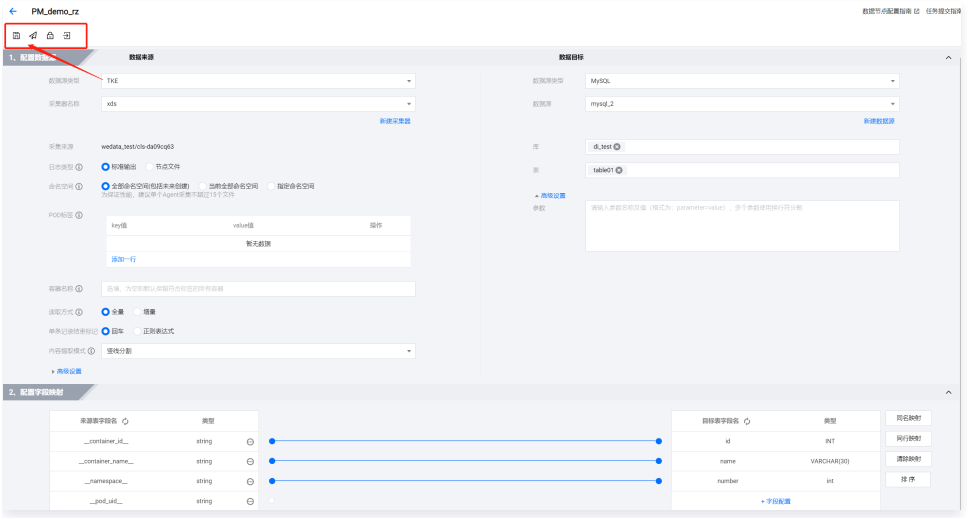
	Type	Field Type.

Step 6: Configure task attributes

Click the right side **Task Attributes** to enter, configure the correct Basic Attributes and Integrated Resource Group.

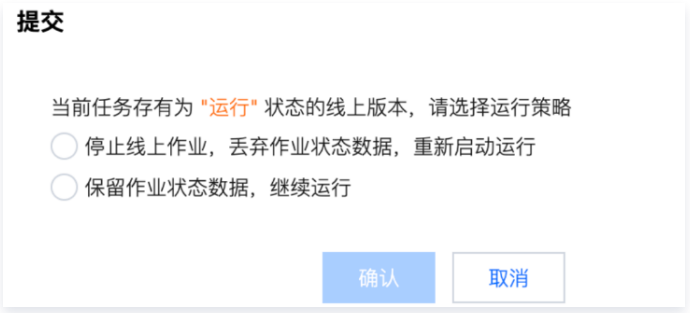


Step 7: Task submission



Serial nu	Parameter	Description

©2013–2025 Tencent Cloud. All rights reserved.

mber		
1	Submit	Submit the current task to the production environment. At the time of submission, different operation strategies can be selected based on whether there is an active production task associated with the current task. • If there are no effective online tasks for the current task, i.e., it's the first submission or the online task is in a "failed" status, you can submit it directly. • If there are "Running" or "Paused" online tasks, you need to choose different strategies. Stopping the online job will abandon the previous task's running location, consuming data from the beginning. Retaining the job status will resume from the last consumed location after restarting.  **Note** Click **Start Now** to immediately run the task after submission, otherwise, it needs to be manually triggered to start running.
2	Lock/Unlock	The default creator is the first lock holder, and only the lock holder can edit task configurations and run tasks. If the lock holder does not perform any editing operations within 5 minutes, others can click the icon to acquire the lock. If successful, they can perform editing operations.
3	Go to Operations	Quickly jump to the task operations page based on the current task name.
4	Save	After the preview is completed, click the **Save** button to save the entire database task configuration. If only saved, the task will not be submitted to the **Operations Center**.

Subsequent Steps

After completing the task configuration, you can perform operations and maintenance as well as monitoring and alarms on the created collection tasks, such as configuring task monitoring alarms, and viewing key indicators of task operation. For details, please refer to Real-time Task Operation and Maintenance.

©2013-2025 Tencent Cloud. All rights reserved.

TKE Source Collection Task Configuration

Last updated: 2024-08-24 20:56:59

Background

Agent is a lightweight log collector provided by DataInLong. It can be installed, managed, and operated in a no-code manner through the product interface for full lifecycle management. When deploying business using Tencent Cloud TKE, you can configure the Agent to collect standard output logs from each pod in the TKE cluster and deliver them to the downstream target.

Operation step

Step 1: Create a collection task

After entering the **Configuration Center** > **Real-time Synchronization Task** page, click **Create New Log Collection Task**. Enter the task name and select the configuration mode, supporting both **Form** and **Canvas** modes.



Step 2: Configure the data source

Select the TKE type in the data source type and configure the parameter information of this data source.

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Collector name	Select the collector available in the current project.
Collection Source	Select the configured TKE Agent name. View the TKE Agent configuration method
Log type	Two types are supported: Standard Output and Node File: • Standard Output: standard output logs of the container, STDOUT, STDERR. • Node File: text logs at a specified path inside the container. If Node File is selected, you need to manually enter the log file path.
Namespace	Select the namespace for this collection task: • All Namespaces: Includes all current and future namespaces. • All current namespaces: Select all current namespaces, excluding future created namespaces. • Specific namespace: Manually enter or select the appropriate namespace. **Note:** To ensure performance, it is recommended that a single Agent collects no more than 15 files per namespace. If more than 15 files need to be collected, multiple collectors can be created.
POD Tag	Filter the specified pod objects for data collection through Tag values. One key can correspond to multiple values, separated by commas. For example: "environment=production,qa", the system will filter out pod objects with environment = production or environment = qa.
Container name	Optional. If left blank, all containers that match the Tag will be retained by default.
Read mode	Supports two reading modes: Full and Increment • Full: Reads from the first line of the log file content. • Increment: Reads the latest content starting from the end of the log.
Single record end mark	Default is the Enter key. If you select a regular expression, you need to manually enter the correct regular expression.

©2013–2025 Tencent Cloud. All rights reserved.

Content extraction mode	Select the required content extraction mode. Three modes are supported: • Full content: Each log record content is parsed into a complete string with a key value of CONTENT. • JSON: Each log record content is parsed into JSON key-value pairs. The key-value must be within the data field definition. • Split: Parse log content based on the specified delimiter. Key values must be within the data fields Definition (supports vertical bar, comma, semicolon split).
Advanced Settings (Optional)	You can configure parameters based on business requirements.

Step 3. Configure the Data Target

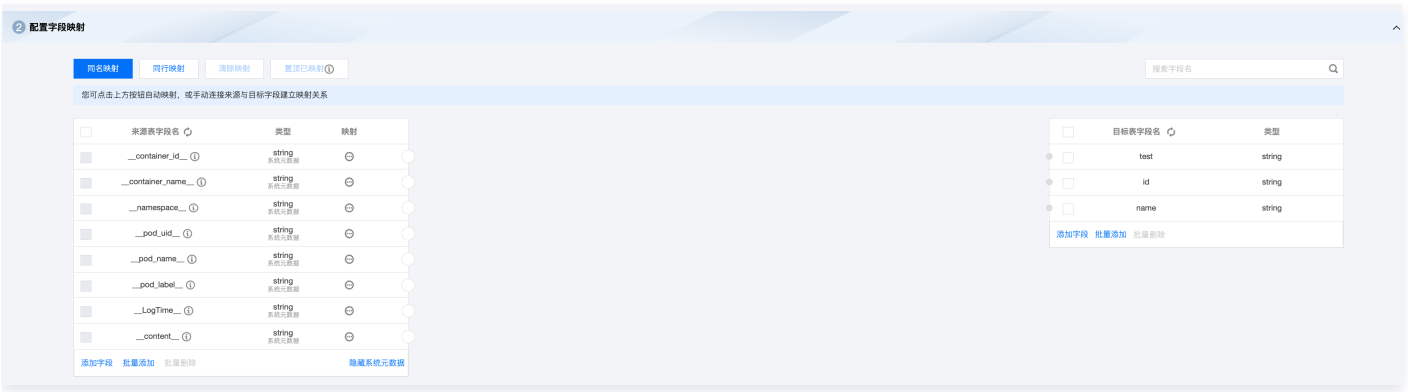
Log collection currently supports most mainstream database connections.

Data Source	Supported data target
TKE	MySQL Write Node
	TDSQL-C for MySQL Write Node
	PostgreSQL Write Node
	SQL Server Write Node
	Oracle Write Node
	Hive Write Node
	ClickHouse Write Node
	Greenplum Write Node
	TBase Write Node
	DLC Write Node
	HBase Write Node
	Iceberg Write Node
	HDFS Write Node
	Doris Write Node
	Elasticsearch Write Node
	Kafka Write Node

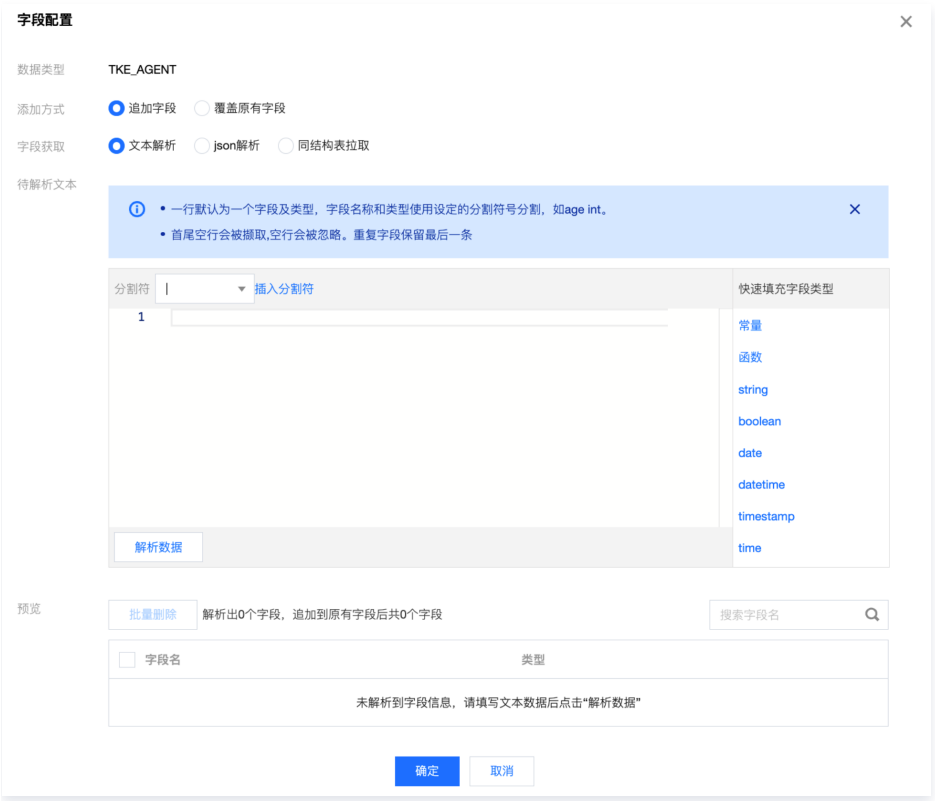
Step 4: Configure Field Mapping

After configuring the data source and data target, the source table and target table field information will be displayed. We need to map the fields using Same Name Mapping and Peer Mapping. Fields can also be sorted and configured.

©2013-2025 Tencent Cloud. All rights reserved.



Click Add Field to configure the fields: You can change the field name and type, as well as delete and add new fields.
Click Bulk Add:



Parameter	Description
Data Type	The data source type of the current node.
Addition Method	• Append Field: Appends the new parsed fields after the original fields of the table. • Overwrite Original Field: The new parsed fields overwrite the original fields in the current source table.
Field Extraction	• Text Parsing: Parses based on the text content, with each line considered as a default field and type. Field names and types are separated by the specified delimiter, e.g., age int. Note: Leading and trailing empty lines will be captured; empty lines will be ignored. • JSON Parsing: Input JSON content and quickly parse the content based on key/value pairs, such as {"age":10,"name":"demo"}. The current system only supports parsing certain types. You can confirm and adjust the parsing results in Form Mode. Duplicate fields will retain the last entry. • Pull from Structural Table: Specify a table object of a data source and parse the fields of that table.

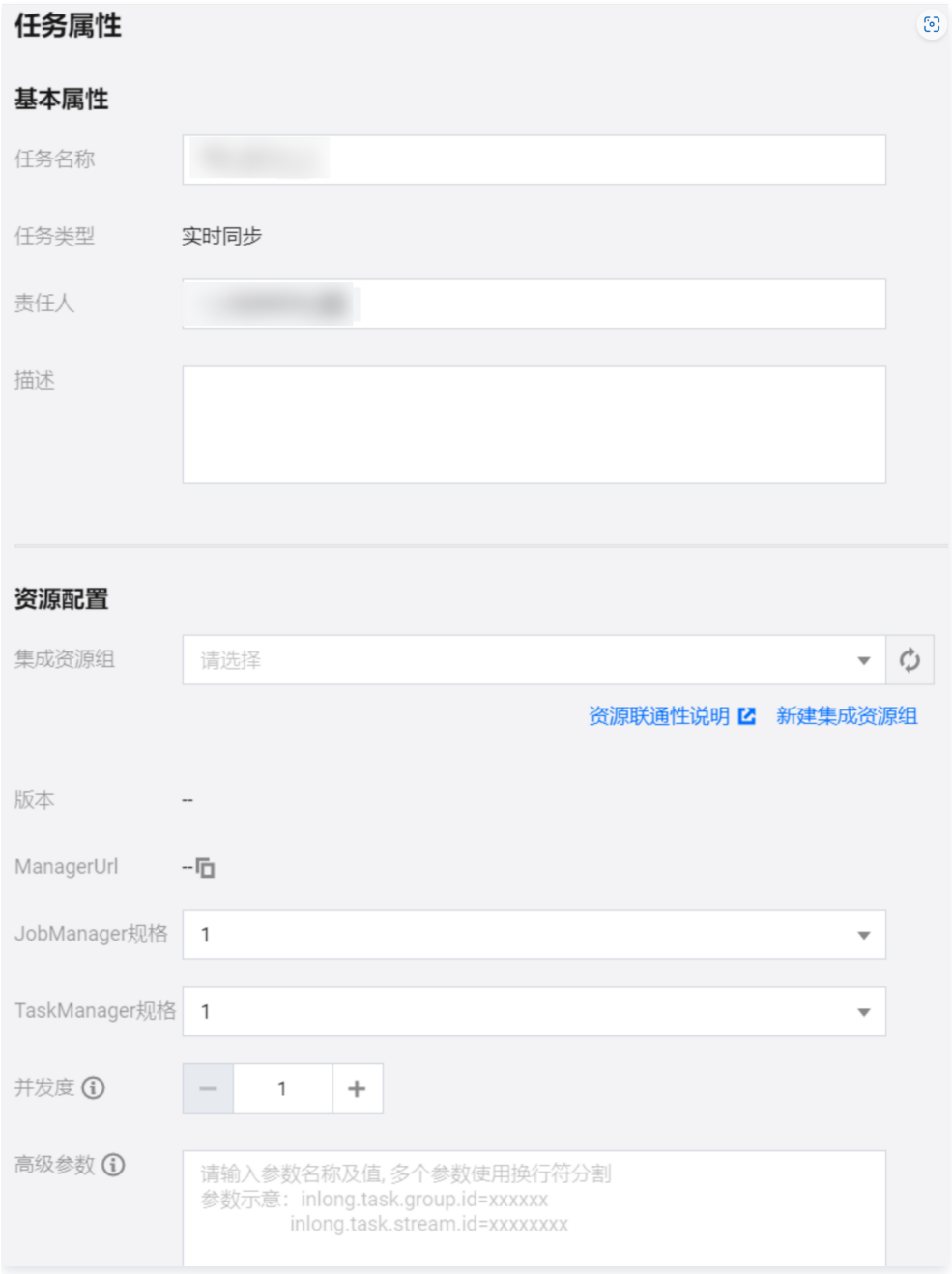
©2013–2025 Tencent Cloud. All rights reserved.

Text to be Parsed	Separator	Used to separate field name and type. Supports tab, |, space, such as "age|int".
	Quick Fill Field Type	Common field types, supporting constants, functions, variables, string, boolean, date, datetime, timestamp, time, double, float, tinyint, smallint, tinyint unsigned, int, mediumint, smallint unsigned, bigint, int unsigned, bigint unsigned, double precision, tinyint(1), char, varchar, text, varbinary, blob.
	Parse Data	Parse the input content.
Preview	Batch delete SNAT rules	After selecting the preview list, batch delete the parsing results.
	Field name	Field Name.
	Type	Field Type.

Step Five: Configure Task Attributes

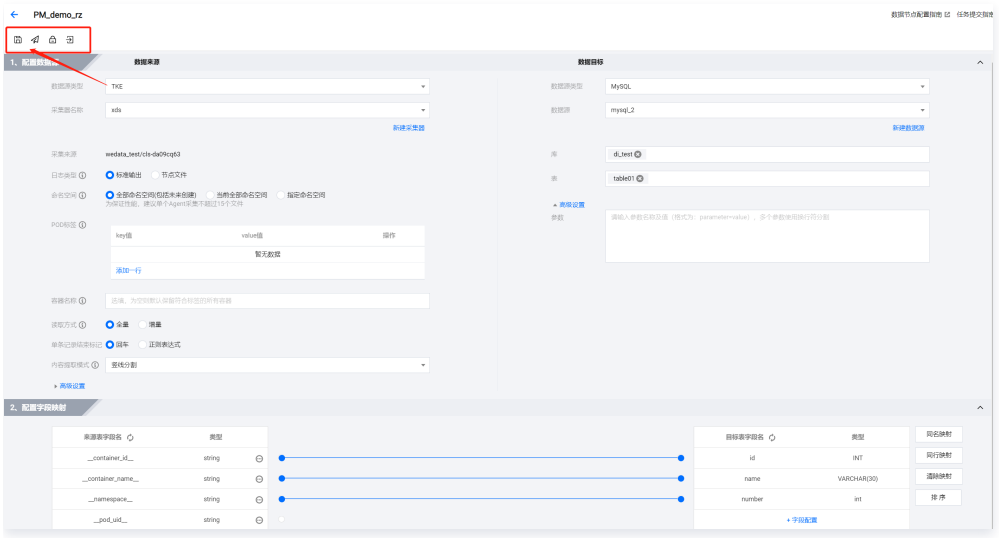
Click the right side **Task Attributes** to enter, configure the correct Basic Attributes and Integrated Resource Group.

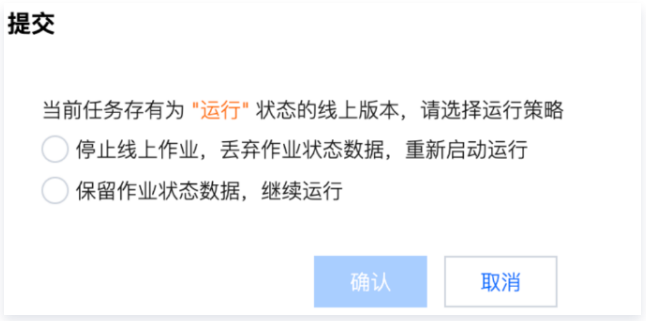
©2013–2025 Tencent Cloud. All rights reserved.



Step Six: Task Submission

©2013–2025 Tencent Cloud. All rights reserved.



Serial number	Parameter	Description
1	Submit	Submit the current task to the production environment. At the time of submission, different operation strategies can be selected based on whether there is an active production task associated with the current task. • If there are no effective online tasks for the current task, i.e., it's the first submission or the online task is in a "failed" status, you can submit it directly. • If there are "Running" or "Paused" online tasks, you need to choose different strategies. Stopping the online job will abandon the previous task's running location, consuming data from the beginning. Retaining the job status will resume from the last consumed location after restarting.  **Note** Click **Start Now** to immediately run the task after submission, otherwise, it needs to be manually triggered to start running.
2	Lock/Unlock	The default creator is the first lock holder, and only the lock holder can edit task configurations and run tasks. If the lock holder does not perform any editing operations within 5 minutes, others can click the icon to acquire the lock. If successful, they can perform editing operations.
3	Go to Operations	Quickly jump to the task operations page based on the current task name.
4	Save	After the preview is completed, click the **Save** button to save the entire database task configuration. If only saved, the task will not be submitted to the **Operations Center.**

Subsequent Steps

©2013–2025 Tencent Cloud. All rights reserved.

After completing the Task Configuration, you can perform Operation and Maintenance and monitoring and alarm on the created collection tasks. For example, configure monitoring and alarms for tasks and view key indicators of task operation. For more details, please refer to Real-time Task Operation and Maintenance.

SDK Source Collection Task Configuration

©2013-2025 Tencent Cloud. All rights reserved.

Last updated: 2024-08-24 20:57:30

Background

DataInLong provides Java, C++ SDKs, supporting reporting business data within the system through built-in SDKs.

Operation step

Step 1: Create a collection task

After entering **Configuration Center > Real-time Synchronization Task**, click **New** Log Collection Task. Enter a task name and select a configuration mode, supporting **Form** and **Canvas** modes.



Step 2: Configure the data source

Select SDK type in the data source type and configure the data source parameter information.



Parameter	Description
Content extraction mode	SDK supports three reading modes: • Full Content: Each log record content is parsed into a complete string with the key value __CONTENT__. • JSON: Each log record content is parsed into JSON key-value pairs. The keys need to be defined in the data field. • Symbol Segmentation: Log content is parsed based on specified delimiters. The keys need to be defined in the data field.

Step 3. Configure the Data Target

Log collection currently supports most mainstream database connections.

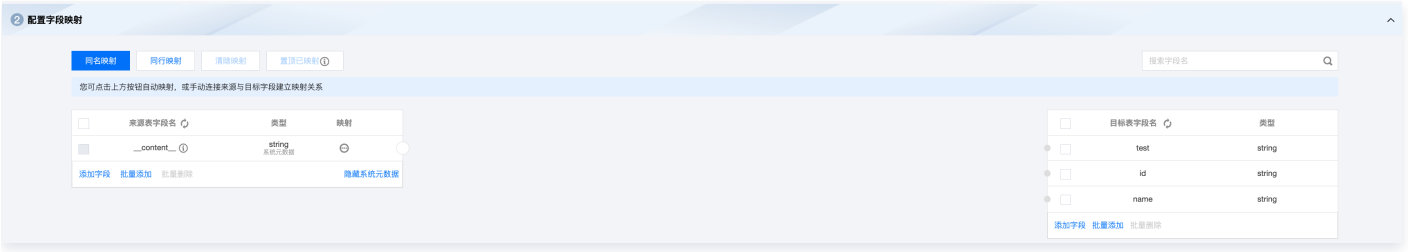
Data Source	Supported data target

©2013-2025 Tencent Cloud. All rights reserved.

SDK	MySQL Write Node
	TDSQL-C for MySQL Write Node
	PostgreSQL Write Node
	SQL Server Write Node
	Oracle Write Node
	Hive Write Node
	ClickHouse Write Node
	Greenplum Write Node
	TBase Write Node
	DLC Write Node
	HBase Write Node
	Iceberg Write Node
	HDFS Write Node
	Doris Write Node
	Elasticsearch Write Node
	Kafka Write Node

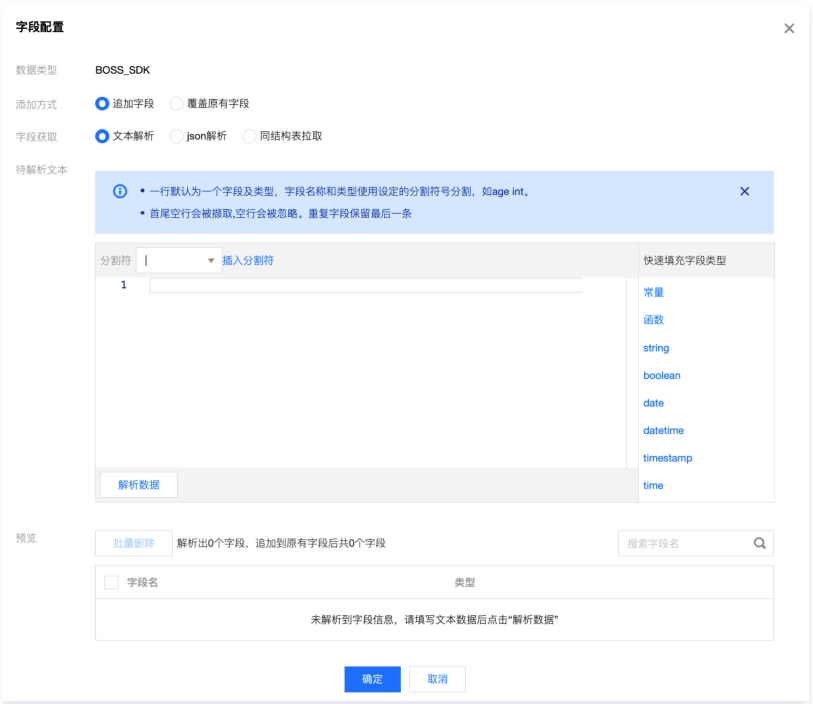
Step 4: Configure Field Mapping

After configuring the data source and data target, the source table and target table field information will be displayed. We need to map the fields using Same Name Mapping and Peer Mapping. Fields can also be sorted and configured.



Click Add Field to configure the fields: You can change the field name and type, as well as delete and add new fields.
Click Bulk Add:

©2013-2025 Tencent Cloud. All rights reserved.

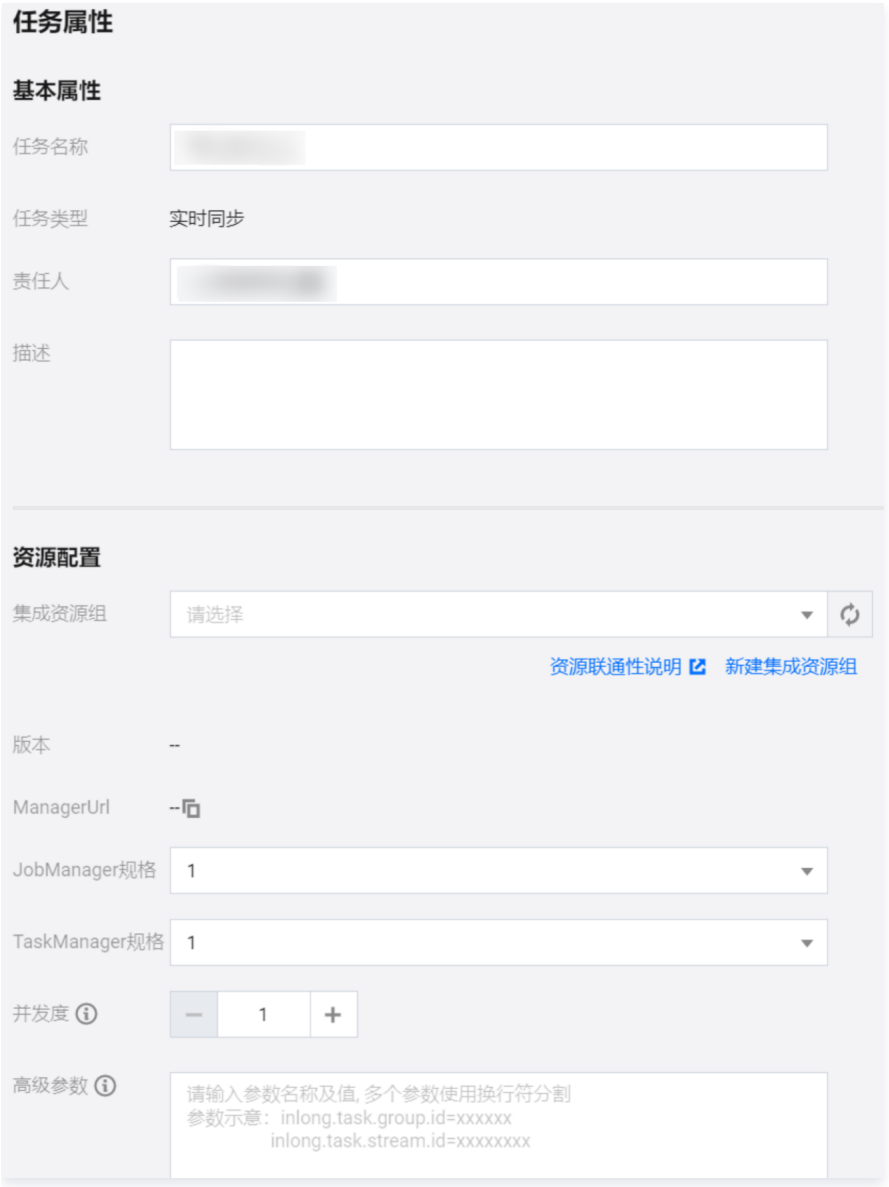


Parameter		Description
Data Type		The data source type of the current node.
Addition Method		• Append Field: Appends the new parsed fields after the original fields of the table. • Overwrite Original Field: The new parsed fields overwrite the original fields in the current source table.
Field Extraction		• Text Parsing: Parses based on the text content, with each line considered as a default field and type. Field names and types are separated by the specified delimiter, e.g., age int. Note: Leading and trailing empty lines will be captured; empty lines will be ignored. • JSON Parsing: Input JSON content and quickly parse the content based on key/value pairs, such as {"age":10,"name":"demo"}. The current system only supports parsing certain types. You can confirm and adjust the parsing results in Form Mode. Duplicate fields will retain the last entry. • Pull from Structural Table: Specify a table object of a data source and parse the fields of that table.
Text to be Parsed	Separator	Used to separate field name and type. Supports tab, |, space, such as "age|int".
	Quick Fill Field Type	Common field types, supporting constants, functions, variables, string, boolean, date, datetime, timestamp, time, double, float, tinyint, smallint, tinyint unsigned, int, mediumint, smallint unsigned, bigint, int unsigned, bigint unsigned, double precision, tinyint(1), char, varchar, text, varbinary, blob.
	Parse Data	Parse the input content.
Preview	Batch delete SNAT rules	After selecting the preview list, batch delete the parsing results.
	Field name	Field Name.
	Type	Field Type.

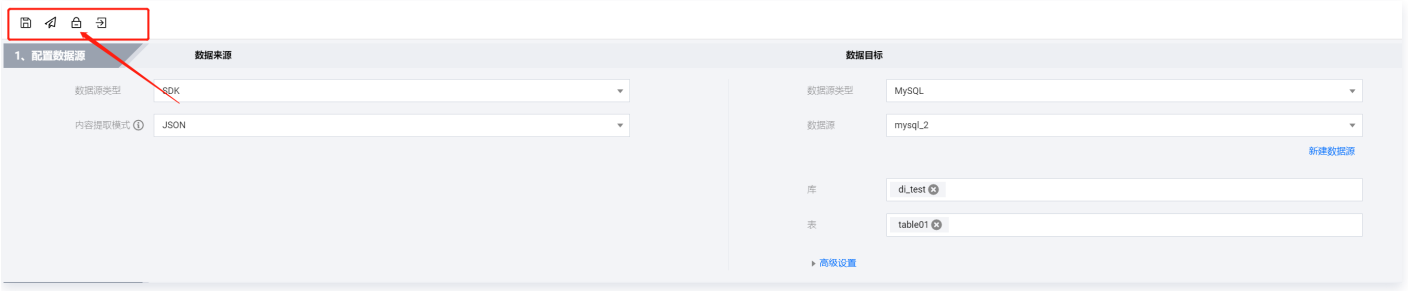
©2013–2025 Tencent Cloud. All rights reserved.

Step Five: Configure Task Attributes

Click the right side **Task Attributes** to enter, configure the correct Basic Attributes and Integrated Resource Group.

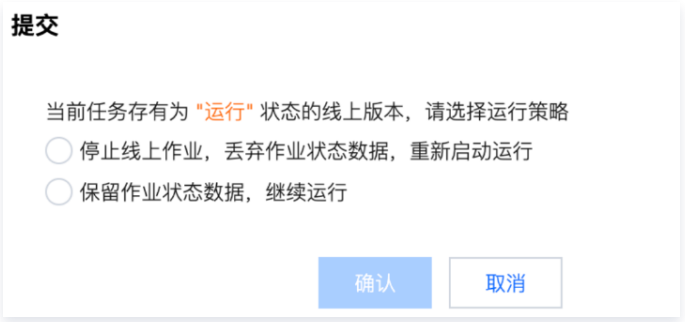


Step Six: Task Submission



Serial number	Parameter	Description

©2013–2025 Tencent Cloud. All rights reserved.

1	Submit	Submit the current task to the production environment. At the time of submission, different operation strategies can be selected based on whether there is an active production task associated with the current task. • If there are no effective online tasks for the current task, i.e., it's the first submission or the online task is in a "failed" status, you can submit it directly. • If there are "Running" or "Paused" online tasks, you need to choose different strategies. Stopping the online job will abandon the previous task's running location, consuming data from the beginning. Retaining the job status will resume from the last consumed location after restarting.  **Note** Click **Start Now** to immediately run the task after submission, otherwise, it needs to be manually triggered to start running.
2	Lock/Unlock	The default creator is the first lock holder, and only the lock holder can edit task configurations and run tasks. If the lock holder does not perform any editing operations within 5 minutes, others can click the icon to acquire the lock. If successful, they can perform editing operations.
3	Go to Operations	Quickly jump to the task operations page based on the current task name.
4	Save	After previewing, you can click the **Save** button to save the whole database task configuration. The task will not be submitted to the **OPS Center** if only saved

Subsequent Steps

After completing the Task Configuration, you can perform Operation and Maintenance and monitoring and alarm on the created collection tasks. For example, configure monitoring and alarms for tasks and view key indicators of task operation. For more details, please refer to Real-time Task Operation and Maintenance.

©2013-2025 Tencent Cloud. All rights reserved.

CVM Source Collection Task Configuration

Last updated：2024-08-24 20:58:01

Background

Agent is a lightweight log collector provided by DataInLong. It can complete installation, management, and operation through a product interface without any coding. When deploying business using Tencent Cloud CVM or self-built servers, you can configure Agent to collect logs and file information within the server and deliver them to downstream targets.

Operation step

Step 1: Create a collection task

After entering **Configuration Center > Real-time Synchronization Task**, click **New** Log Collection Task. Enter a task name and select a configuration mode, supporting **Form** and **Canvas** modes.



Step 2: Configure the data source

Select the CVM type in the data source type and configure the data source parameters.



Parameter	Description
Collector Group	Choose an available collector for the current project. If there isn't one, click **Create Collector** to

©2013-2025 Tencent Cloud. All rights reserved.

	create one.
Server categories	Select the server category it belongs to. The task will collect all servers under this category after selection.
File path	Manually enter the file path of the data source.
Blocklist	Disabled by default. When enabled, the configured blocklist file path will not be collected.
Read mode	CVM source supports two reading methods: • Full: Reads from the first line of the log file content. • Increment: Reads the latest content starting from the end of the log.
Single record end mark	Default is the Enter key. If you select a regular expression, you need to manually enter the correct regular expression.
Content extraction mode	Supports three content extraction modes: • Full Content: Each log record content is parsed into a complete string with the key value __CONTENT__. • JSON: Each log record content is parsed into JSON key-value pairs. The keys need to be defined in the data field. • Split: Parse log content based on the specified delimiter. Key values must be within the data fields Definition (supports vertical bar, comma, semicolon split).

Step 3. Configure the Data Target

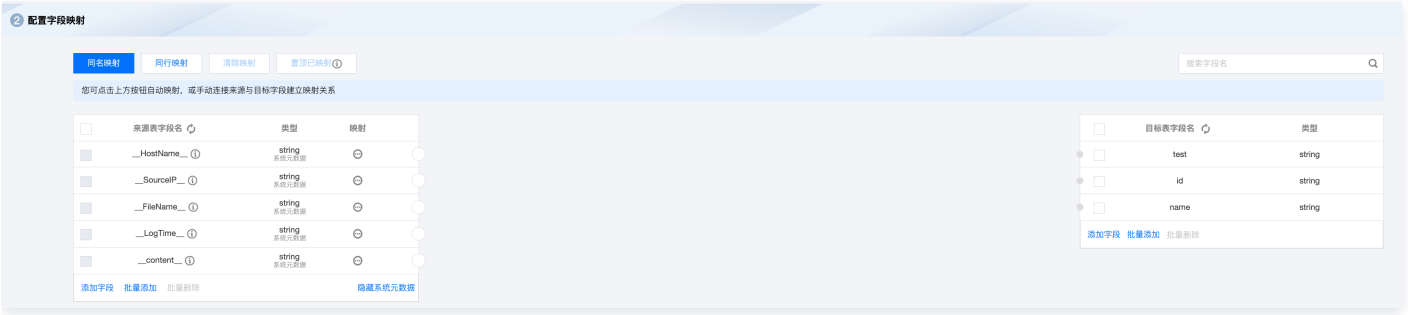
Log collection currently supports most mainstream database connections

Data Source	Supported data target
CVM	MySQL Write Node
	TDSQL-C for MySQL Write Node
	PostgreSQL Write Node
	SQL Server Write Node
	Oracle Write Node
	Hive Write Node
	ClickHouse Write Node
	Greenplum Write Node
	TBase Write Node
	DLC Write Node
	HBase Write Node
	Iceberg Write Node
	HDFS Write Node
	Doris Write Node
	Elasticsearch Write Node
	Kafka Write Node

Step 4: Configure Field Mapping

After configuring the data source and data target, the source table and target table field information will be displayed. We need to map the fields using Same Name Mapping and Peer Mapping. Fields can also be sorted and configured.

©2013-2025 Tencent Cloud. All rights reserved.



Click Add Field to configure the fields: You can change the field name and type, as well as delete and add new fields.
Click Bulk Add:



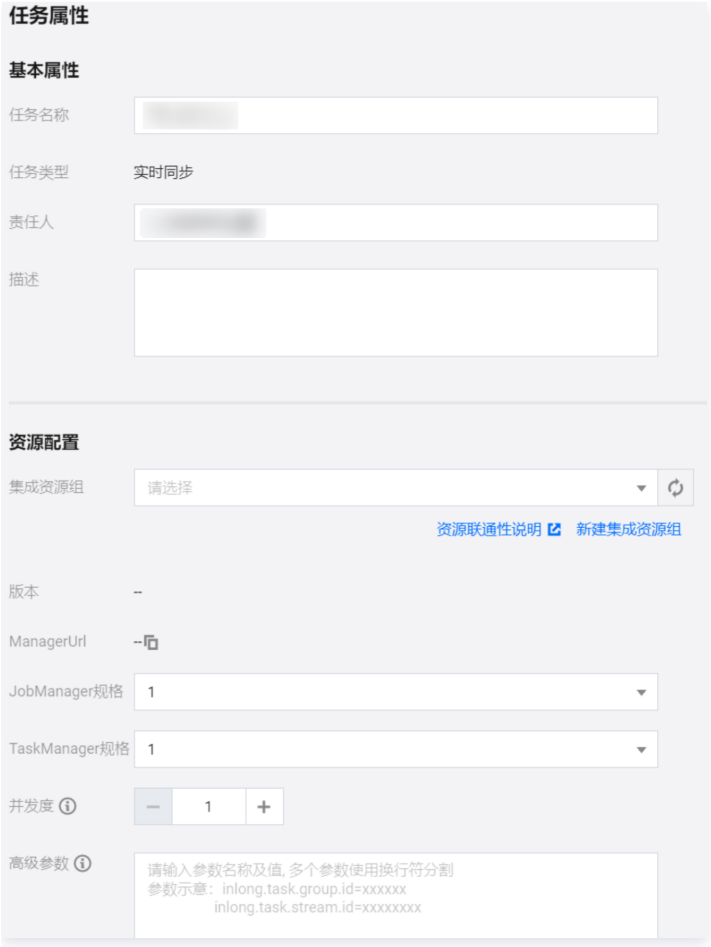
Parameter		Description
Data Type		The data source type of the current node.
Addition Method		• Append Field: Appends the new parsed fields after the original fields of the table. • Overwrite Original Field: The new parsed fields overwrite the original fields in the current source table.
Field Extraction		• Text Parsing: Parses based on the text content, with each line considered as a default field and type. Field names and types are separated by the specified delimiter, e.g., age int. Note: Leading and trailing empty lines will be captured; empty lines will be ignored. • JSON Parsing: Input JSON content and quickly parse the content based on key/value pairs, such as {"age":10,"name":"demo"}. The current system only supports parsing certain types. You can confirm and adjust the parsing results in Form Mode. Duplicate fields will retain the last entry. • Pull from Structural Table: Specify a table object of a data source and parse the fields of that table.
Text to be Parsed	Separator	Used to separate field name and type. Supports tab, |, space, such as "age|int".
	Quick Fill Field Type	Common field types, supporting constants, functions, variables, string, boolean, date, datetime, timestamp, time, double, float, tinyint, smallint, tinyint unsigned, int, mediumint, smallint unsigned, bigint, int unsigned, bigint unsigned, double precision, tinyint(1), char, varchar, text, varbinary, blob.

©2013–2025 Tencent Cloud. All rights reserved.

	Parse Data	Parse the input content.
Preview	Batch delete SNAT rules	After selecting the preview list, batch delete the parsing results.
	Field name	Field Name.
	Type	Field Type.

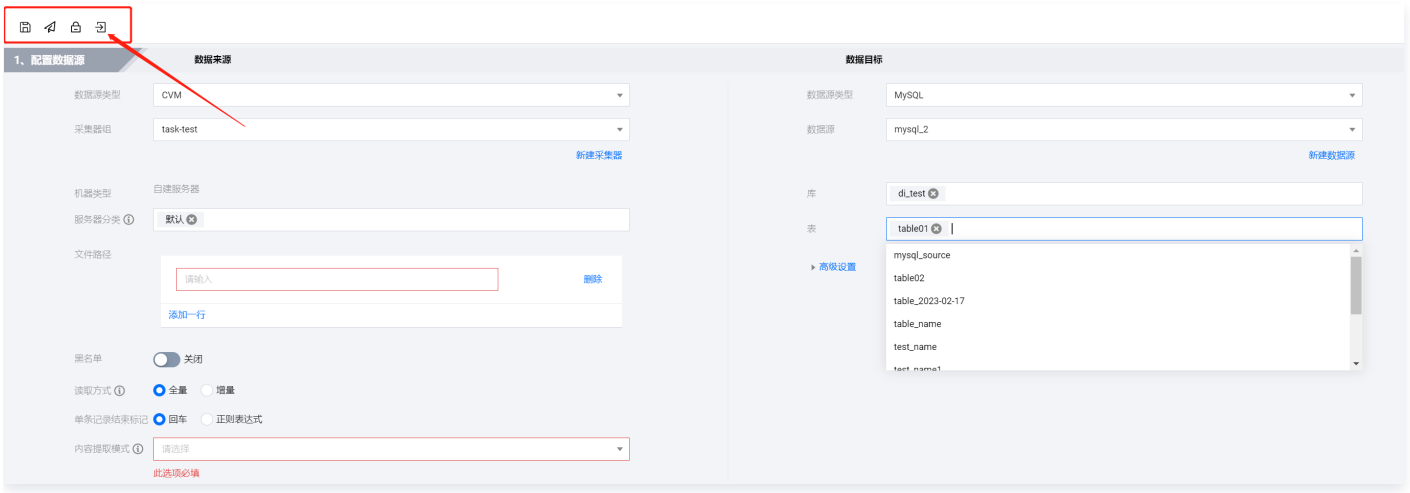
Step Five: Configure Task Attributes

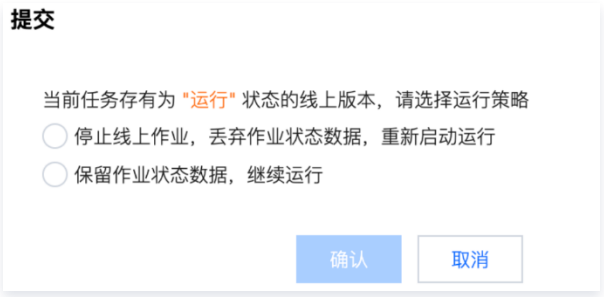
Click the right side **Task Attributes** to enter, configure the correct Basic Attributes and Integrated Resource Group.



Step Six: Task Submission

©2013–2025 Tencent Cloud. All rights reserved.



Serial number	Parameter	Description
1	Submit	Submit the current task to the production environment. Depending on whether there are production tasks, choose different operation strategies when submitting • If there are no active online tasks for the current task, i.e., it is the first submission or the online task is in "failed" status, it can be directly submitted • If there are "Running" or "Paused" online tasks, you need to choose different strategies. Stopping the online job will abandon the previous task's running location, consuming data from the beginning. Retaining the job status will resume from the last consumed location after restarting.  **Note** Click **Start Now** to immediately run the task after submission, otherwise, it needs to be manually triggered to start running.
2	Lock/Unlock	The default creator is the initial lock holder, and only the lock holder can edit task configuration and run tasks. If the lock holder does not perform any editing operations within 5 minutes, others can click the icon to acquire the lock and perform editing operations upon successful lock acquisition
3	Go to Operations	Quickly navigate to the task operation and maintenance page based on the current task name
4	Save	After previewing, click the **Save** button to save the full database task configuration. If only saved, the task will not be submitted to the **Operations Center**

Subsequent Steps

After completing the Task Configuration, you can perform Operation and Maintenance and monitoring and alarm on the created collection tasks. For example, configure monitoring and alarms for tasks and view key indicators of task operation. For more

©2013–2025 Tencent Cloud. All rights reserved.

details, please refer to Real-time Task Operation and Maintenance.

Collector Management

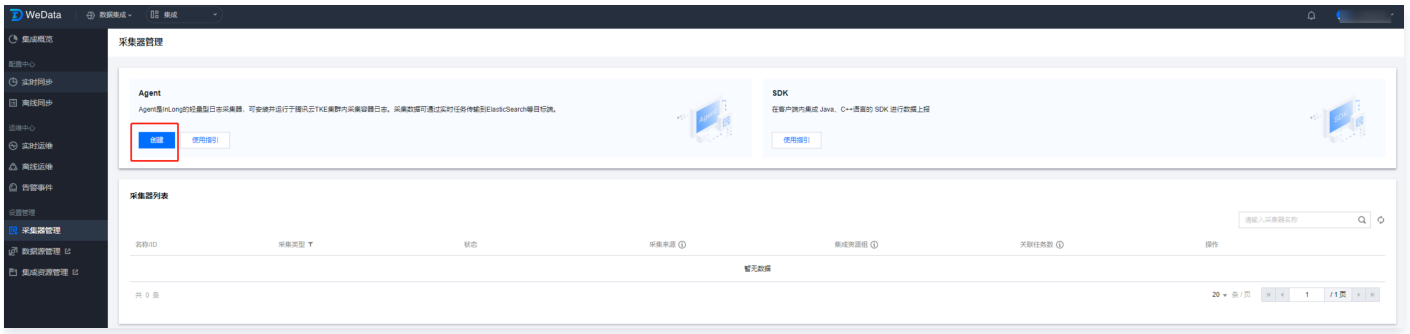
Last updated: 2024-08-23 18:20:26

©2013-2025 Tencent Cloud. All rights reserved.

Go to the **wedata DataInLong> Settings Management > Collector Management** interface. The page displays two options: Agent and SDK.
Agent is InLong's lightweight log collector that can be installed and run in Tencent Cloud's TKE cluster to collect container logs. The collected data can be transmitted to targets like ElasticSearch through real-time tasks.

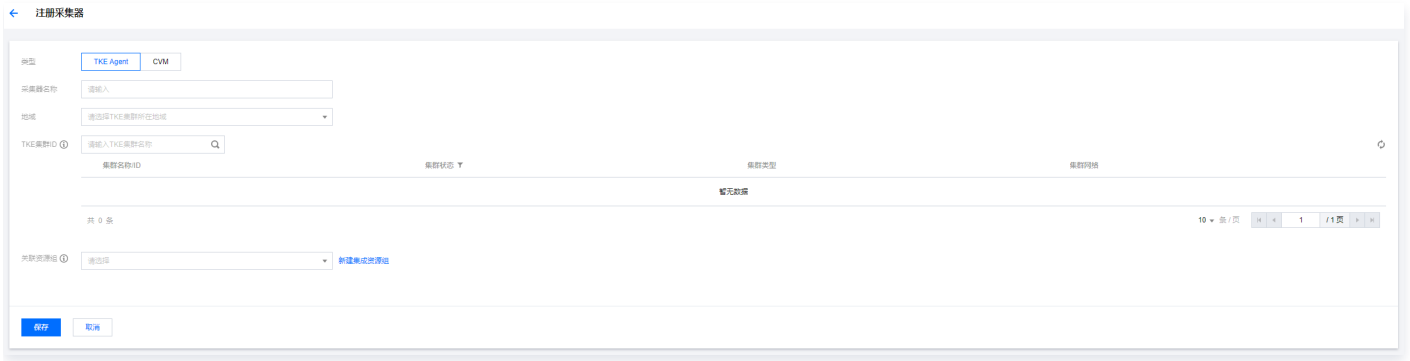
Create Agent Collector

1. Enter the DataInLong module, click **Collector Management > Agent > Create.**



2. In the Create Agent Collector pop-up window, configure the TKE Agent/CVM parameter information.

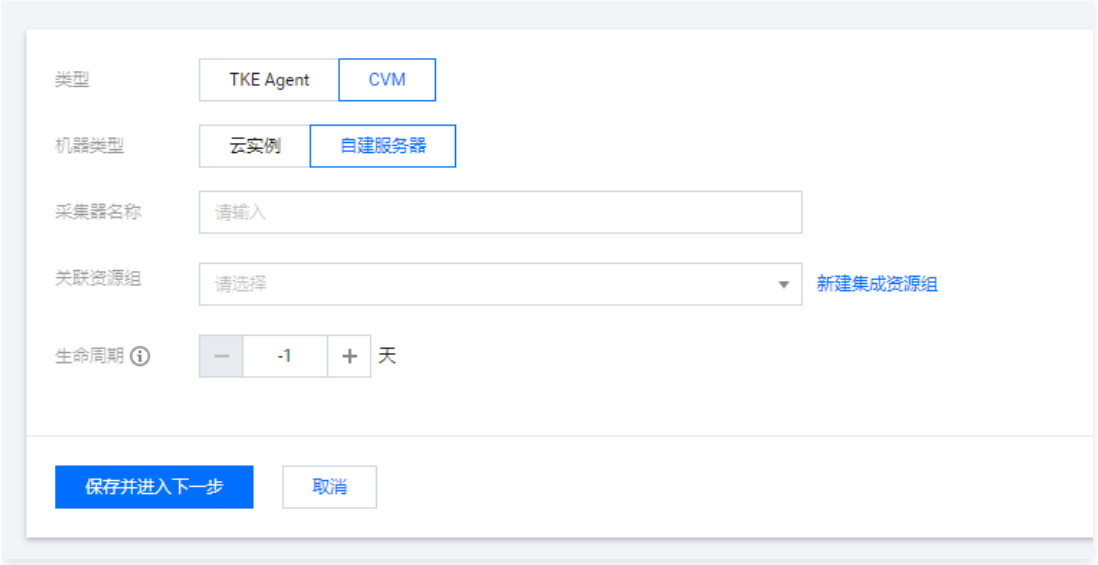
- **TKE Agent**



Parameter	Parameter description
Type	Please select TKE Agent.
Region	Select the region to which the TKE cluster belongs. You can log in to Tencent Cloud to view cluster information.
TKE Cluster ID	Select the TKE cluster information for installation. 1. Only TKE clusters in the "Running" state can have the Agent installed. 2. One agent will occupy 1C512M specifications of the cluster.
Associated Resource Group	Bind the Agent with a specific execution resource group. The Agent will use the manager URL within the resource group Report data. **Note:** TKE Cluster (Agent) must be in the same VPC as the integrated resource group, or the corresponding VPC must have public network configured to synchronize data. For the public network configuration process of the resource group, please refer to DataInLong Resource Group Public Network Configuration.

- **CVM**

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Parameter description
Machine type	Default self-built server.
Collector name	Custom collector name.
Associated Resource Group	Select the available resource group in the current project.
Lifecycle	Set the maximum heartbeat timeout for the Agent under the current collector group. If exceeded, the server IP corresponding to the Agent will be automatically reclaimed. The default is -1, which means not reclaimed. It can be edited and modified in the collector list.

3. Click **Save and Enter Next** to enter server management.



Note:

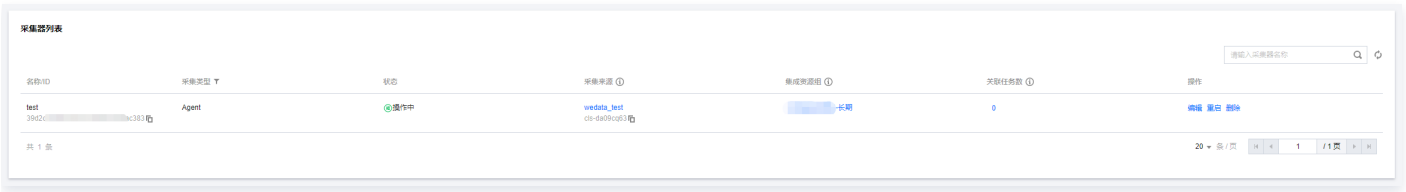
1. Self-built server Agent management requires manually running script commands. You can click **Install/Remove/Restart Agent** to view the steps.
2. The sync task configuration information is effective for all servers under the category by default. Adding or removing servers under the category will automatically increase or decrease the task collection scope. Editing the category name does not affect the original task operation.
3. The server ownership category can only be modified by command line or manually editing the configuration file.

4. After configuration, the Agent will act as a log collector on TKE, supporting the use of a single Agent for POD log extraction in multiple real-time tasks.

Collector list

The created collector can be managed and viewed in the collector list. The list displays content including: Name ID/Status/Collection Source/Integrated Resource Group/Number of Associated Tasks and related operations.

©2013-2025 Tencent Cloud. All rights reserved.



©2013–2025 Tencent Cloud. All rights reserved.

Real-time synchronization and operation and maintenance
Real-Time Operations and Maintenance
OPS Operations

Last updated: 2024-09-06 15:35:03

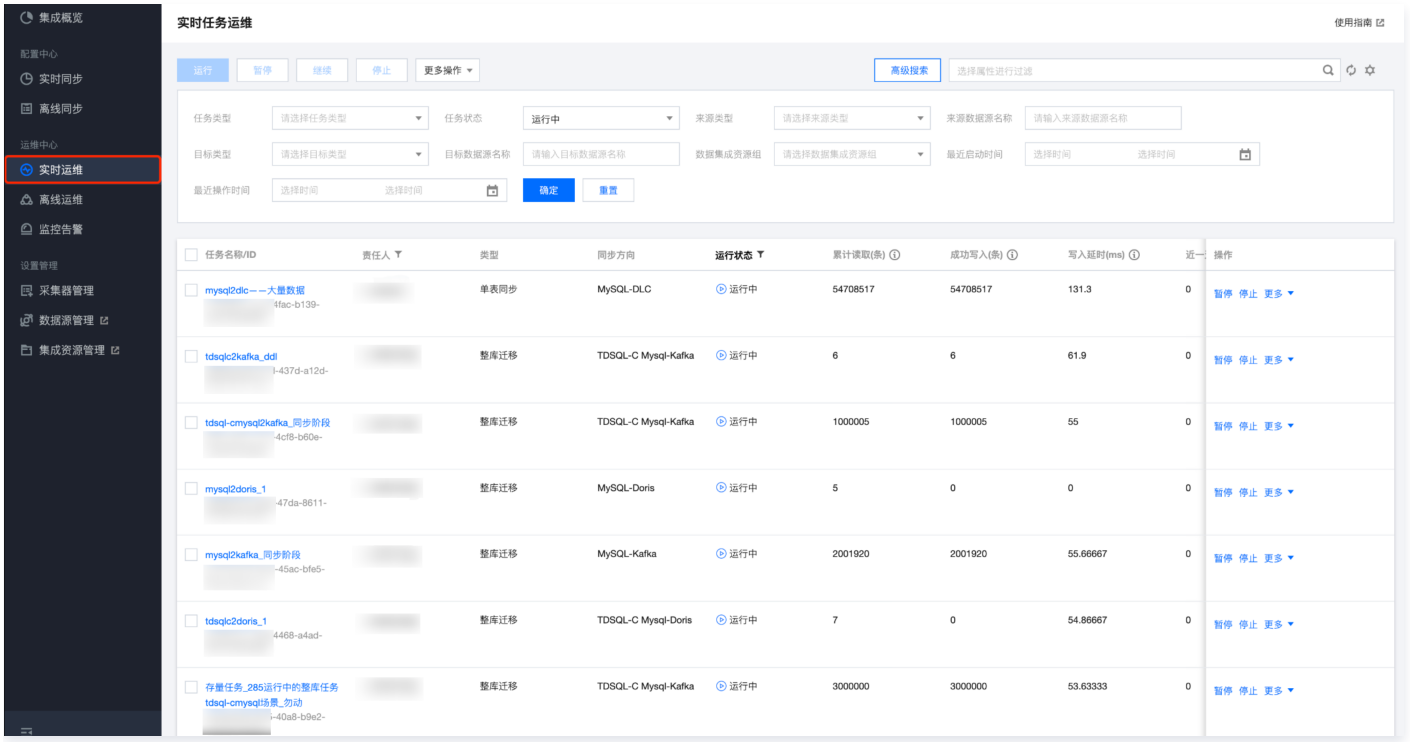
For real-time tasks that have been configured and submitted, you can enter the **Real-time Operations** section to view and manage the corresponding tasks. In the Real-time Operations section, you can check the basic running status of tasks, task metrics, running logs, configuration task abnormal alarms, and more. This document outlines common operation and maintenance activities for real-time synchronization tasks.

Prerequisites

The operations page only supports and displays real-time tasks that have been submitted. For details, please refer to Whole Database Task Configuration Overview, Single Table Task Configuration Overview, and Log Collection Configuration Overview.

Operations List

The operations list supports managing and controlling all submitted tasks, including start and stop operations.



Real-time Task Status and Meanings:

Status Types	Description
Initialization	The task has not yet started running after being submitted to the operation and maintenance for the first time.
Operating	Status is being changed. This status often appears shortly after completing task operations and the task is in the process of status changing.
Running	Task is running.
Suspended	The current task has been paused. The task status and read-end point are retained. However, there is a distinction between normal pause and forced pause. A normal pause generates a snapshot at the

©2013-2025 Tencent Cloud. All rights reserved.

	time of pausing, whereas a forced pause does not generate a snapshot during the operation. When resuming the task, it runs from the most recent snapshot created during task operation.
Stopped	Task has been manually stopped. **Note** If a user manually stops a task, it will switch to this status. Other non-exception cases of stopping.
Failed	Task execution failed.
Exception Detection	The configured resource group or other objects do not exist

Real-time task support operations and meanings:

Operation Type	Description
Running	Run the current task to start reading and writing
Pause	Pause the current task. After pausing, the task status and location are preserved. Subsequent operations can resume from the completed location.
Force Pause	A forced pause does not generate a snapshot. When resuming the task, it runs from the most recent snapshot created during task operation. This type of pause may cause some task data to replay. If the target write is append-based, there will be duplicate data; if the target write is upsert-based, there will be no duplication issue **Note** When a task cannot be paused normally, you can use Force Pause.
Resume	Resume from the last paused location. **Note** It is recommended not to pause tasks for too long, otherwise restarting may fail due to expired source end logs or lost locations.
Stop	Manually stop the current task. After stopping, the running status and location of the task will not be retained. **Note:** Please use the stop operation with caution. After stopping, restarting will discard the original running status and location of the task. Historical operation metrics and log data will also be unavailable.

Real-time task statuses and corresponding allowable operations:

Status Types	Operation			
	Running	Pause/Force Pause	Resume	Stop
Initialization	✓	–	–	–
Operating	–	–	–	–
Running	–	✓	–	✓

©2013-2025 Tencent Cloud. All rights reserved.

Suspended	–	–	✓	✓
Stopped	✓	–	–	–

©2013–2025 Tencent Cloud. All rights reserved.

Metrics Statistics

Last updated：2024-09-06 15:35:44

You can click **Real-time operations and maintenance > Actions > More >Statistical indicators** or **Real-time operations and maintenance >Task name >Statistical indicators** to enter the real-time synchronization statistical indicators interface to view the task synchronization progress, task resource utilization, and database table synchronization status. The statistical indicators are mainly divided into three dimensions: read/write indicators, resource indicators, and link details (specifically for whole database tasks).

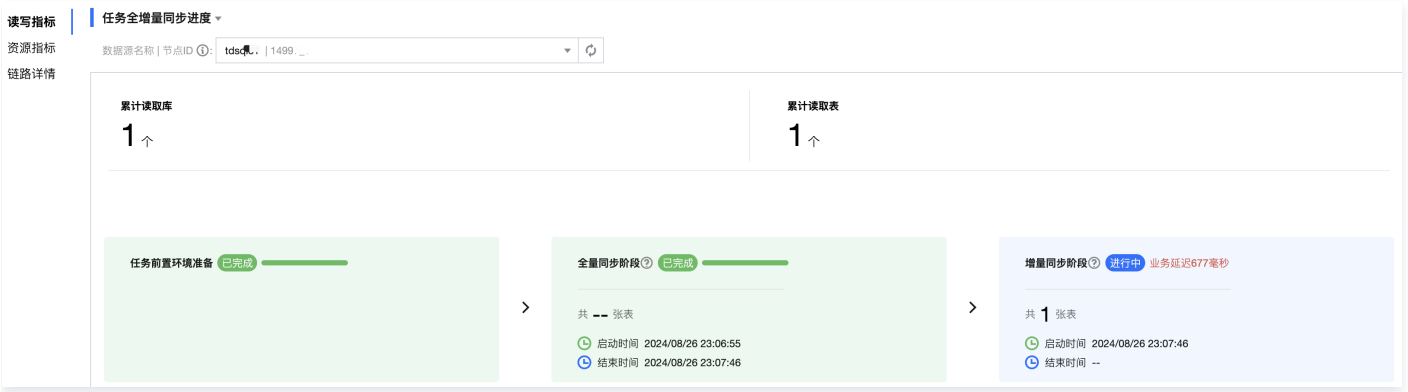
Read/Write Indicators

In the Read and Write Metrics page, you can view the full incremental synchronization progress of real-time synchronization tasks by data source dimension, view the amount of data synchronized by task and its key metrics such as read-write speed, throughput, and latency by time dimension.

Task full incremental synchronization progress

The full incremental synchronization progress of tasks can be viewed by data source dimension across three phases: Prepared environment, Full synchronization, and Incremental synchronization.

> **Note**
> Currently, task full incremental synchronization progress is only supported when the source end is MySQL, Oracle, or PostgreSQL.



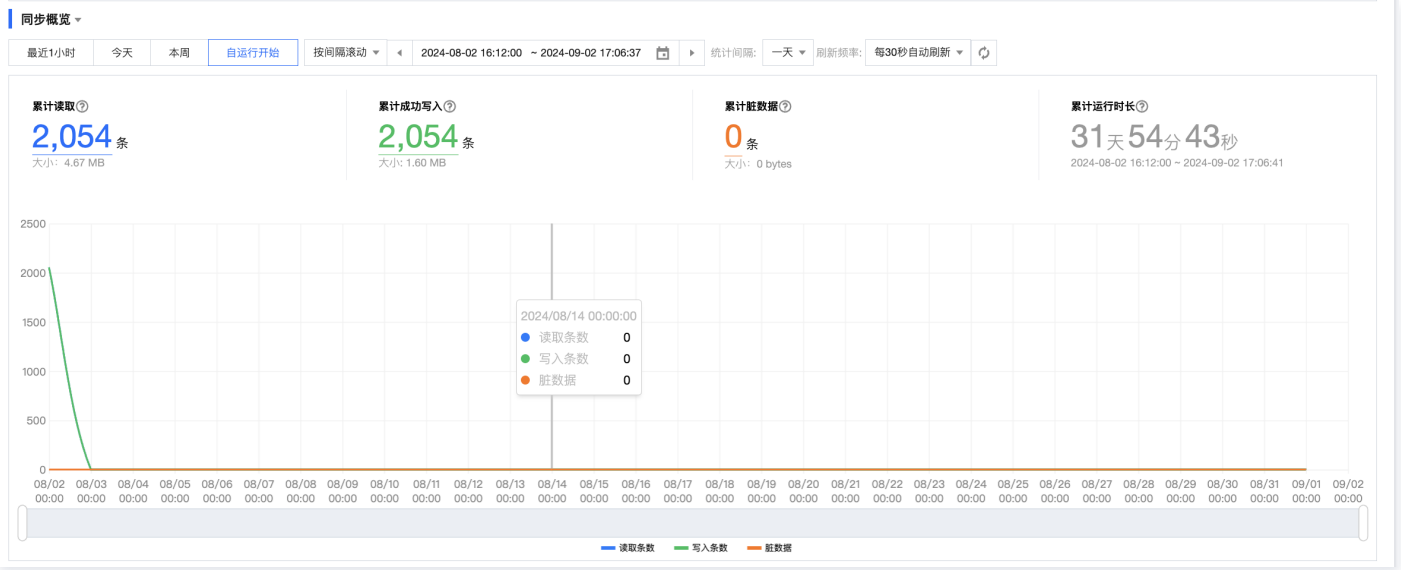
Synchronization Phase	Phase Overview	Display Metrics
Task Prepared Environment	This phase mainly initiates the corresponding task	Shows specific progress (percentage) and status (not started/in progress/completed)
Full Synchronization Phase	Reading all existing data from the source database tables	This phase only counts and displays the number of tables that have been successfully read and contain existing business data. It shows the full start time, the full end time, and the specific progress (percentage) and status (not started/in progress/completed) of this phase
Incremental Synchronization Phase	Consuming changed data from binlog and reading incremental data from the source database tables	• In this phase, only the number of tables that have been read successfully and have new business data are counted and displayed. It shows the incremental start time, incremental end time, and the synchronization status of this phase (Not Started/In Progress/Completed) • The task progresses to the incremental stage; under normal circumstances, the task will continuously remain in an **in progress** state, and does not have a specific end time. • The task business latency indicator is displayed on the right side of the status and is automatically refreshed every 5 seconds.

©2013-2025 Tencent Cloud. All rights reserved.

> **Note**
> There will be some business latency during the incremental phase, which is a normal phenomenon. However, if the latency is large, please check if it affects the business.

Synchronization Overview

Synchronization Overview allows you to view the total real-time task data synchronization for different time dimensions and intervals:

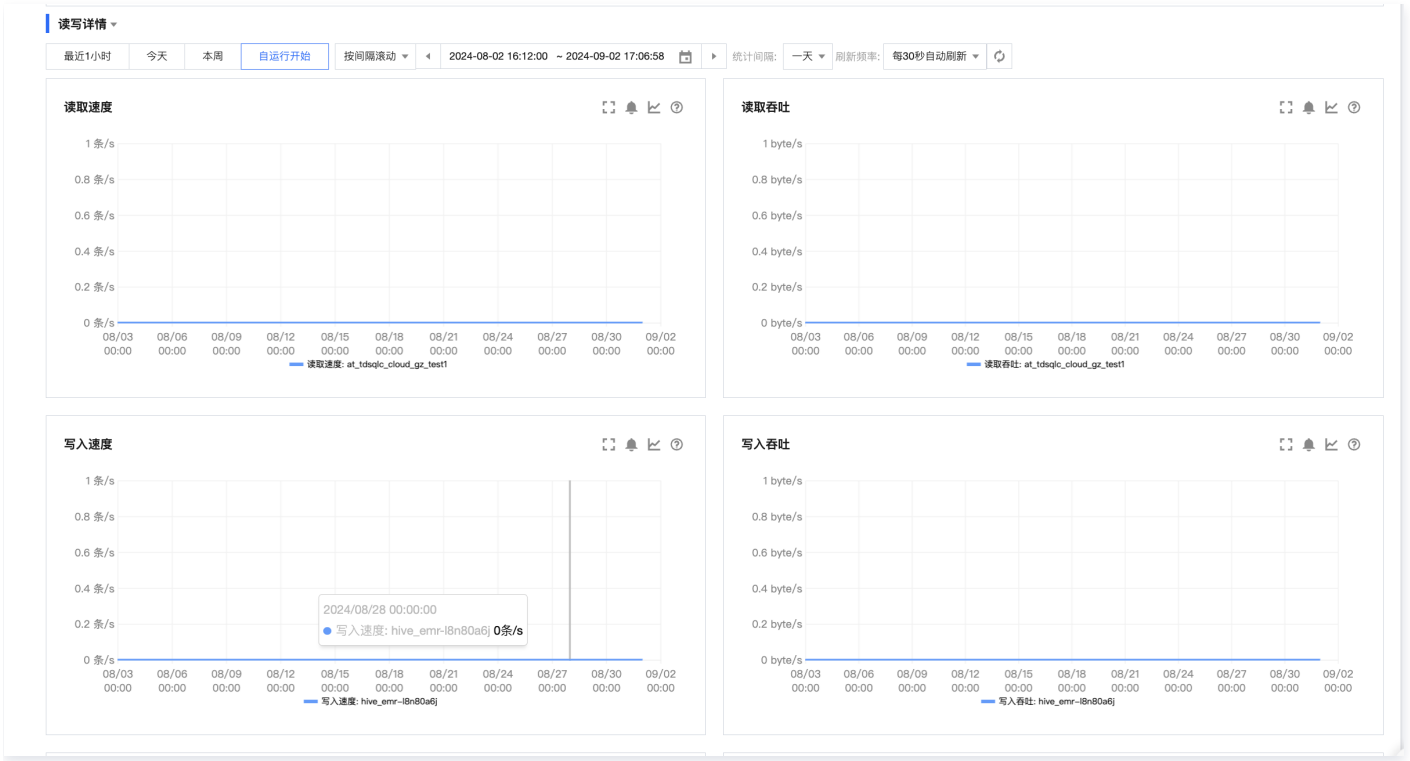


Parameter	Description
Time period	Support for selecting the Last Hour, Today, This Week, Since Start of Operation, and custom Definition time range
Accumulated Reads	During this task's runtime, the actual number of data entries read from the source. This metric does not include data that has been removed via filtering or other methods.
Accumulated Successful Writes	During this task's runtime, the total amount of data successfully written to the target from the data read.
Accumulated Dirty Data	During this task's runtime, the total amount of data that failed to write due to errors. This metric does not include data that was deliberately ignored/filtered due to task configuration, specified stoppages, abnormal restarts, or data filtering strategies.
Accumulated running duration	After this task starts, the accumulated total runtime duration (including pause time).

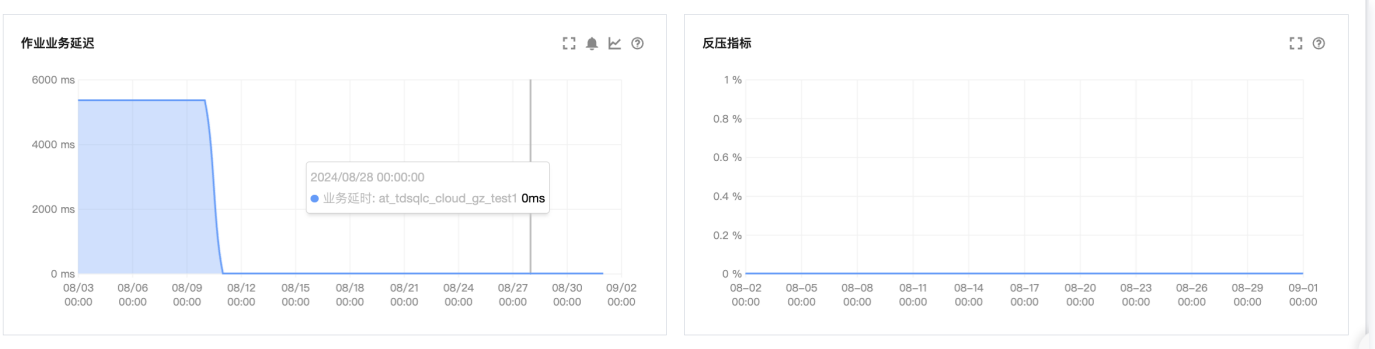
Read-write Details

Read and Write Details allow you to view real-time task read-write speeds and latency-related indicators for different time dimensions and intervals.

©2013-2025 Tencent Cloud. All rights reserved.





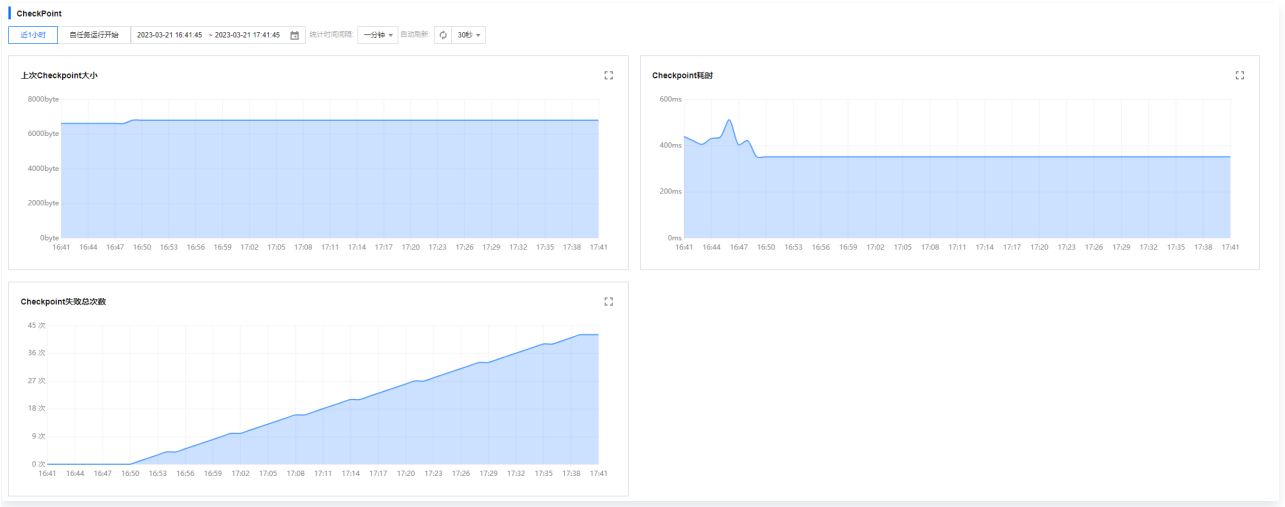


©2013–2025 Tencent Cloud. All rights reserved.

Parameter	Description
Time period	Support for selecting the Last Hour, Today, This Week, Since Start of Operation, and custom Definition time range
Read Speed	Total number of reads within the statistical interval/statistical interval
Read Throughput	Total read volume within the statistical interval/statistical interval
Write Speed	Total number of successful writes within the statistical interval/statistical interval
Write Throughput	Total successful write volume within the statistical interval/statistical interval
Write Latency	Link latency from the Source end to the Sink end: Write Latency = System Time − Record Read Time (Timestamp of LatencyMarker at the read end)
Number of Job Restarts	Number of restarts for the current job
Log Position Information	Real-time task CDC log position information
Job Transmission Latency	The time when the data is read by the Source (FetchTime) minus the data event time (EventTime).
Job Business Latency	The time when the data leaves the Source (EmitTime) minus the data event time (EventTime)
Backpressure Indicator	This indicator reflects the relative speed of the reading end and the writing end. Below 0.5% indicates the reading end is slower than the writing end; above 0.5% indicates the reading end is faster than the writing end

Resource Indicator

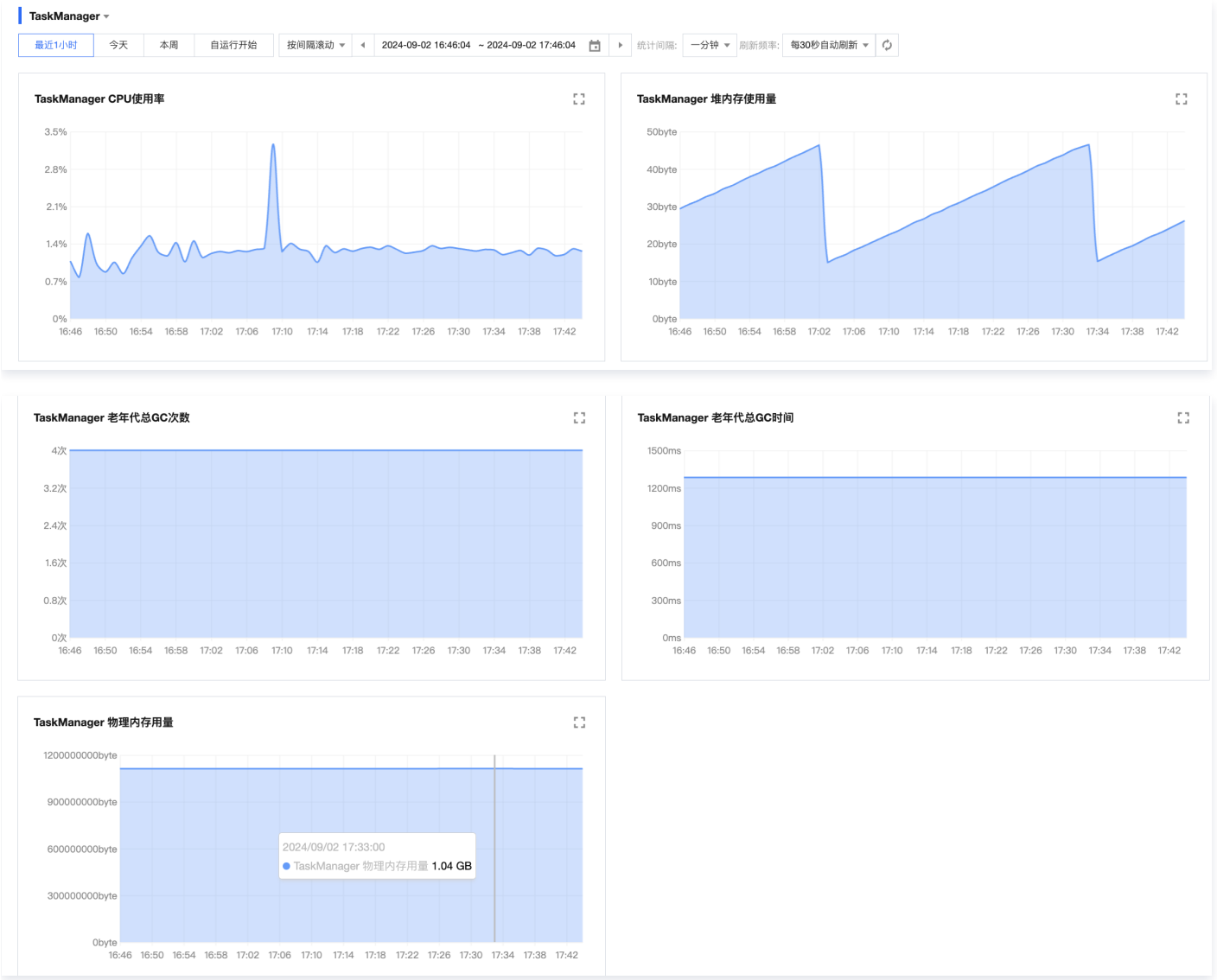
CheckPoint



Indicator Parameters	Description	Sample value
Last Checkpoint Size	Size of the most recent checkpoint for the current job	751321.00Bytes
Checkpoint Duration	Duration of the checkpoint for the current job	723.00ms
Total Checkpoint Failures	Total number of checkpoint failures for the current job	8 times

©2013–2025 Tencent Cloud. All rights reserved.

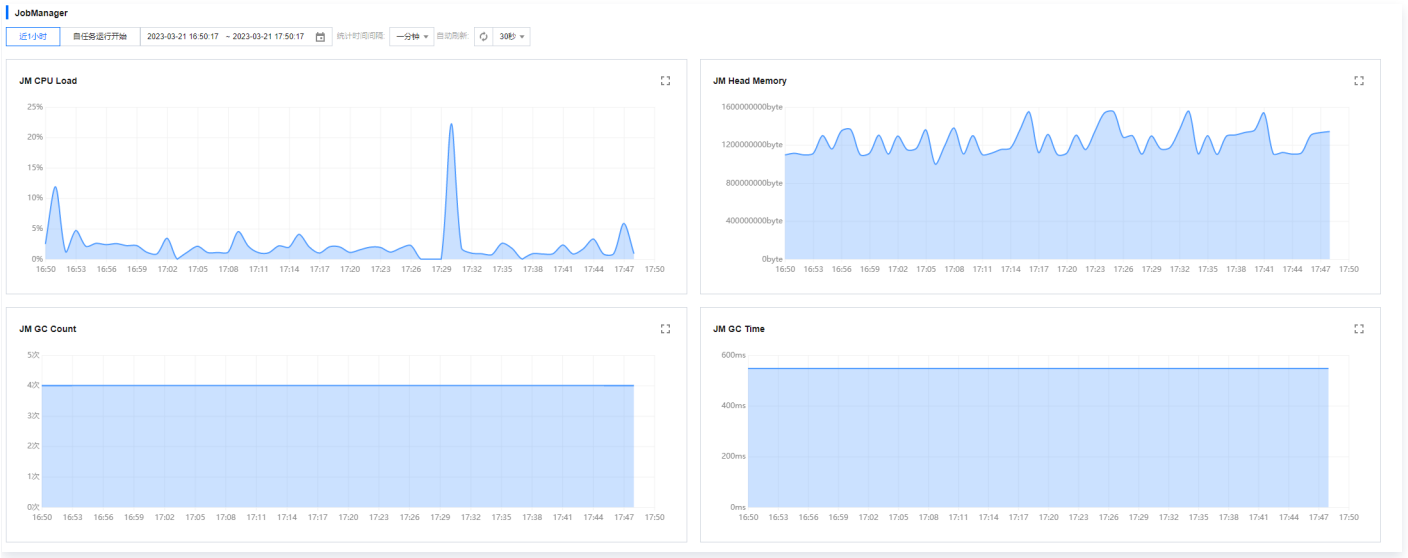
TaskManager



Indicator Parameters	Description	Sample value
TaskManager CPU utilization	CPU utilization of the current job's TaskManager.	7.12%
TaskManager Heap Memory Usage	Heap memory usage of the current job's TaskManager.	1040001560.00Bytes
TaskManager Total Elderly Generation GC Count	Current Job's TaskManager Elderly Generation GC Count.	3 times
TaskManager Total Elderly Generation GC Time	Current Job's TaskManager Elderly Generation GC Time.	701.00ms
TaskManager Physical Memory Usage	Physical Memory Usage (RSS) of the JVM running the current job's TaskManager, including all areas such as heap, off-heap, and native.	3597035110.00Bytes

JobManager

©2013–2025 Tencent Cloud. All rights reserved.



Indicator Parameters	Description	Sample value
JM CPU Load	Recent CPU Utilization Rate of the JVM at the TaskManager Dimension.	12%
JM Head Memory	Heap Memory Usage at the TaskManager Dimension.	1 time
JM GC Count	TaskManager Dimension Status.JVM.GarbageCollector.<GarbageCollector>.Count, number of GC (Garbage Collection) times.	5 times
JM GC Time	TaskManager Dimension Status.JVM.GarbageCollector.<GarbageCollector>.Time, GC (Garbage Collection) time.	65ms

Link Details

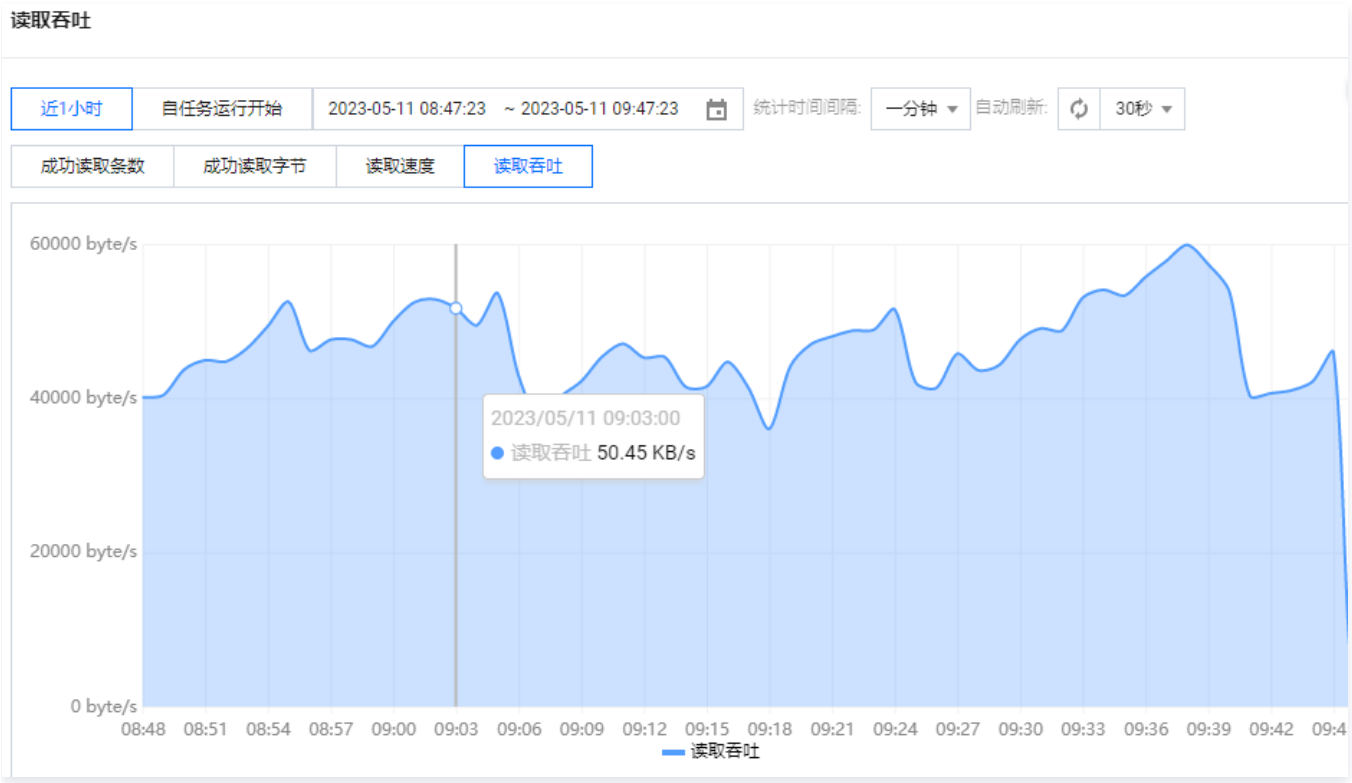
Whole Database Synchronization Task operation and maintenance supports viewing synchronization link details. Click **Real-time Operation and Maintenance > Task Name > Statistic Metrics > Link Details** to enter the link details operation and maintenance page. Link details support viewing data read/write metrics at table granularity.

Reading End



- Source Table Metrics: Display success details of tables read in the whole database synchronization task.
  - Including Database Name, Table Name, Number of Reads/Bytes, and Reading Speed/Throughput.
  - Trend chart supports viewing detailed metrics of a single table during the task running period.
  - Supports search by Table Name and Database Name.

©2013-2025 Tencent Cloud. All rights reserved.



- Trend
  - By default, expand the trend charts of each metric within the past hour. Manual adjustment of the display time range is supported.
  - Data in the pop-up window automatically refreshes every minute. Manual refresh is also supported.

Writing End



- Target Table Metrics: Displays the statistics of tables written in the whole database task.
  - Including Database Name, Table Name, Number of Writes/Bytes, Writing Speed/Throughput, and Dirty Data.
  - Trend chart supports viewing detailed metrics of a single table during the task running period.
  - Supports search by Table Name and Database Name.

©2013–2025 Tencent Cloud. All rights reserved.



- Trend
  - By default, expand the trend charts of each metric within the past hour. Manual adjustment of the display time range is supported.
  - Data in the pop-up window automatically refreshes every minute. Manual refresh is also supported.

©2013-2025 Tencent Cloud. All rights reserved.

Key Events

Last updated: 2024-09-06 15:36:10

Key events support viewing job start/stop events, job execution failed events, source table Schema changes, etc.
The real-time operation and maintenance interface supports accessing key events by clicking **Task Name> Key Events** from the task record.



Parameter	Description
Job Start/Stop	• When a user selects **Run** from the job operations menu, the system will attempt to start the job and automatically create a new Instance ID for this run; subsequently, a new start event can be seen in the event panel. • When the user stops, restarts the job, or the job crashes and exits, a stop event will be generated under this Instance ID. **The Event Occurrence Time refers to the time when the internal process is completed, not the time when the user clicked the operation on the interface.**
Job Execution Failed	When the job restarts during execution (changes from the `RUNNING` state to restarting `RESTARTING` or `FAILED` states), an "Execution Failed" event will be generated. If the job then re-enters the `RUNNING` state, a "Job failure recovered" event will be generated. Solutions corresponding to job execution failed events can also be viewed.
Snapshot Failure	If the job has periodic snapshots enabled and a snapshot fails, a "Snapshot Failure" event will be generated. If subsequent snapshots are successful, a "Snapshot Failure Recovered" event will be generated. Solutions corresponding to snapshot failure events can also be viewed.
Job Exceptions	Various job exception events will be continuously monitored and analyzed by the backend. When a job encounters a severe exception (e.g., prolonged TaskManager Full GC, high CPU occupancy rate for an extended period, abnormal Pod exits, etc.), a corresponding event will be pushed for users to view and subscribe to, allowing for the assessment of job health. Solutions corresponding to job exception events can also be viewed.
Source End Schema Changes	During the task synchronization process, if there is a Schema change on the source side, it is necessary for the business to be able to perceive it in a timely manner and to quickly judge and respond. When there is a Schema change on the source side, capturing and displaying change events is supported. In addition, it is possible to configure alerts for source-side Schema change events.

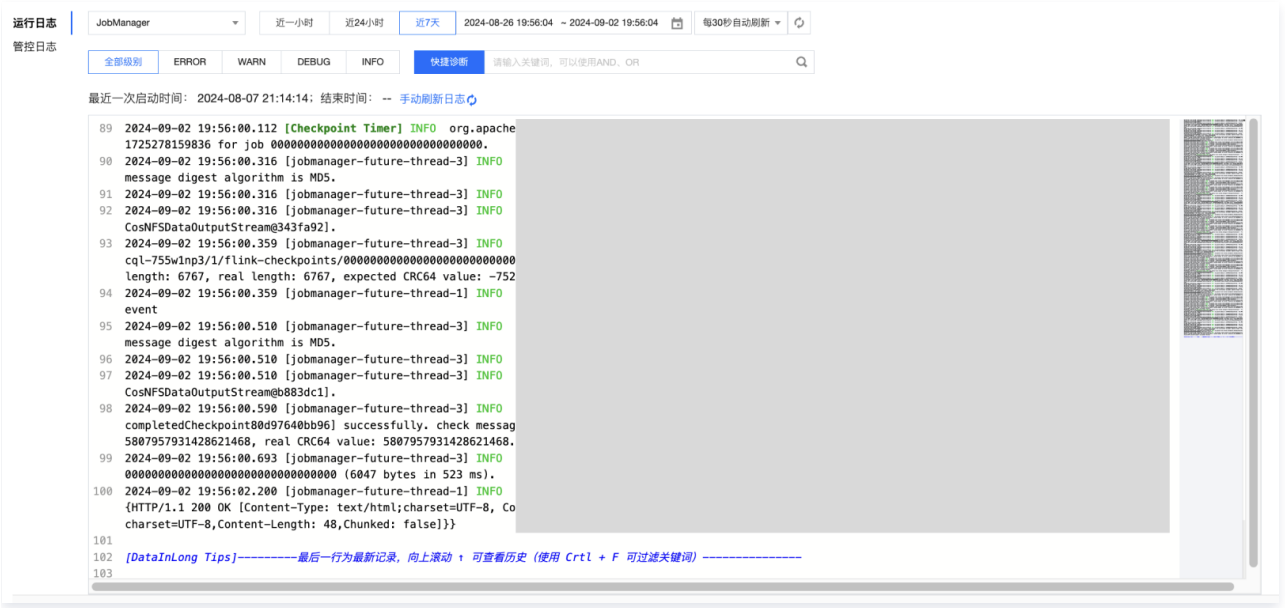
Note
To avoid returning too many events, we restrict the maximum consecutive time period for filtering to 7 Days, and you can view events from the past 90 days at most.

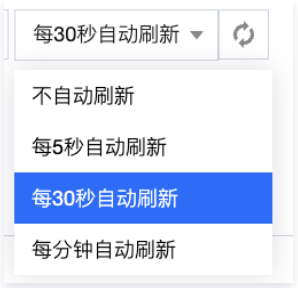
©2013-2025 Tencent Cloud. All rights reserved.

Viewing running logs

Last updated：2024-09-06 15:36:34

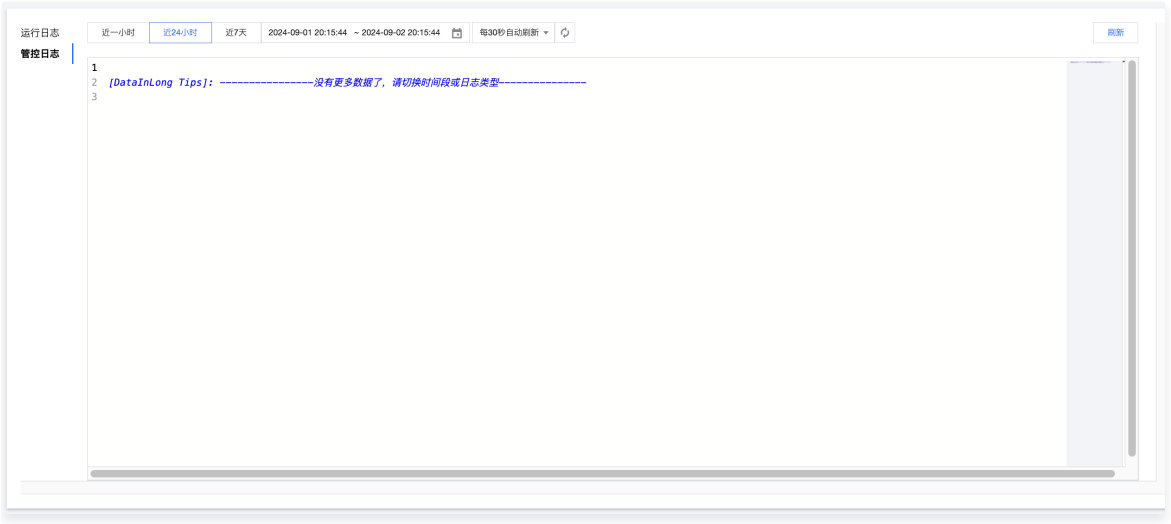
The running log interface supports viewing task running logs for the past 1 hour, 24 hours, and the past 7 days.
The real-time operations interface supports accessing the running log interface by clicking **task name > running log** from the task record list.

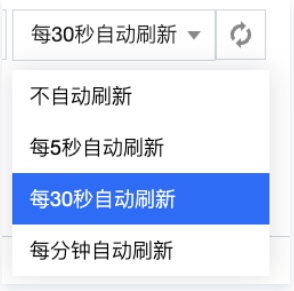


Parameter	Description
JobManager/Task Manager	Supports switching between Jobmanager and TaskManager instance information to view logs of different instances.
Time period	Supports filtering and retrieving log content by time. Quick clicks for logs within the past hour, day, and past 7 days are supported.
Refresh frequency	Supports setting the log refresh frequency. 
Log level	• All levels: Show logs of all levels. • ERROR: Show logs of ERROR level only. • WARN: Show logs of WARN level only. • DEBUG: Show logs of DEBUG level only. • INFO: Show logs of INFO level only.
Quick Diagnosis	Quickly add common problem keywords.

©2013-2025 Tencent Cloud. All rights reserved.

	
Keyword	User-defined keyword search. Supports keyword concatenation using AND and OR.
Log	Displays log content based on the selected time and level.



Parameter	Description
Time	Supports filtering and retrieving log content by time. Quick clicks for logs within the past hour, day, and past 7 days are supported
Refresh frequency	Supports setting the log refresh frequency 
Log	Displays log content based on the selected time and level

©2013-2025 Tencent Cloud. All rights reserved.

Alarm Subscription

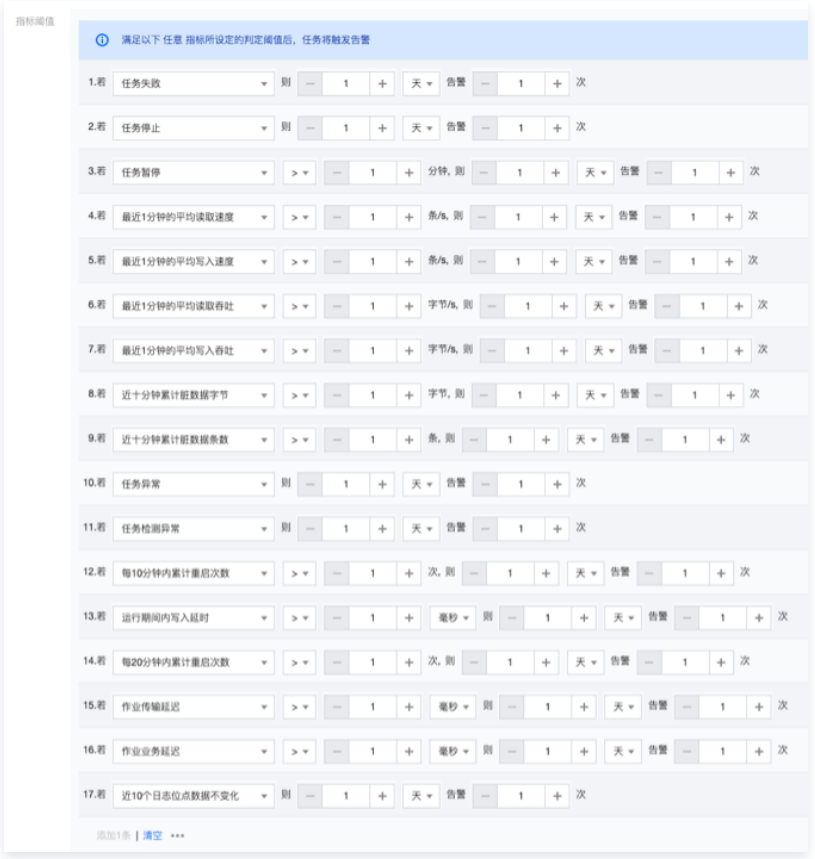
Last updated：2024-09-06 15:36:54

Alert Subscription provides the creation of task alert rules based on different indicators and alert thresholds for each sync task. A task can be associated with multiple alert levels and various alert rule monitoring.
The Real-Time Operation and Maintenance interface supports accessing Alert Subscription from the task record list by clicking **Task Name>Alert Subscription**.



Parameter	Description
Rule Name/ID	The user-specified alert rule name or the system-generated current alert rule ID.
Rule Status	• Enable: Enable the current alert rule. Alerts will be triggered during task execution based on the alert rule and threshold. • Disable: Disable alert validation.
Alarm Level	The importance of the currently triggered alert. Based on the alert levels of different metrics, distinguishes the content sent by the alarm. Currently, there are three alert types available: Normal, Important, and Urgent.
Indicator Threshold	Real-time sync tasks currently support configuring task exceptions, task detection exceptions, total restart times within every 10 minutes, write latency during the runtime, total restart times within every 20 minutes, and 17 other indicator alerts.

©2013-2025 Tencent Cloud. All rights reserved.

	
Last Alarm Time	Time of the most recent alarm
Operation	Support removing alarms, viewing alarm history, and editing rules. • Remove Alarm: Unlink the alarm rule from the current task without deleting the alarm rule or affecting its association with other tasks **Note** Remove Alarm: Only supports removing monitoring objects for specific tasks • Alarm History: View all alarm history information based on the current rule name • Edit Rule: Support editing the current alarm rule. The edited result will apply to all tasks associated with the rule

Batch Remove Alarm Rules

Support batch unbinding of alarm rules from the currently associated tasks, without deleting the alarm rules or breaking their association with other tasks.

Associate Alarm Rules

Support batch associating multiple existing alarm rules with the current task.

©2013–2025 Tencent Cloud. All rights reserved.



Parameter	Description
Rule Name/ID	The user-specified name of the alarm rule or the system-generated alarm rule ID.
Indicator Threshold	Real-time sync tasks currently support configuring task exceptions, task detection exceptions, total restart times within every 10 minutes, write latency during the runtime, total restart times within every 20 minutes, and 17 other indicator alerts.
Rule Creator	User who created the rule.

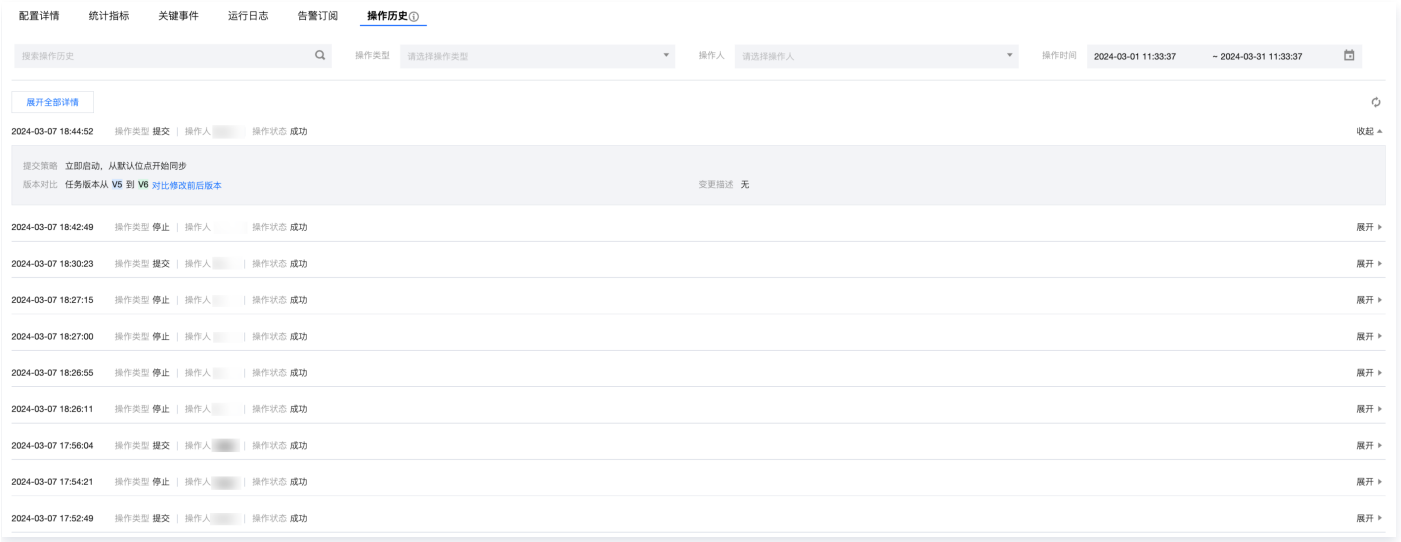
©2013-2025 Tencent Cloud. All rights reserved.

Operation History

Last updated：2024-09-06 15:37:20

In Operation History, view all non-query page operations on the Real-time Operation and Maintenance page. The Real-time Operation and Maintenance interface allows you to access the Operation History through the task record list. Click **Task Name> Operation History,** to enter the Operation History interface.

> **Note:**
> Only page operations are recorded, not OpenAPI operations.



Parameter	Description
Search	Supports keyword search for Operation History
Operation Type	Supports selecting All, Submit, Run, Stop, Pause, Force Pause, Continue, Delete, Add Alarm Rule, Remove Alarm Rule, and Change Responsible Person operations
Operator	Supports selecting all members under this project
Operation Time	Supports filtering operation records by time
Display List	• Operation Time: Displays the start time of the operation, accurate to the second. • Operator: Displays the login account of the user executing the operation. • Operational Status: Displays the result of the operation, Success/Failure • Expand: Click to display detailed content based on the operation type.

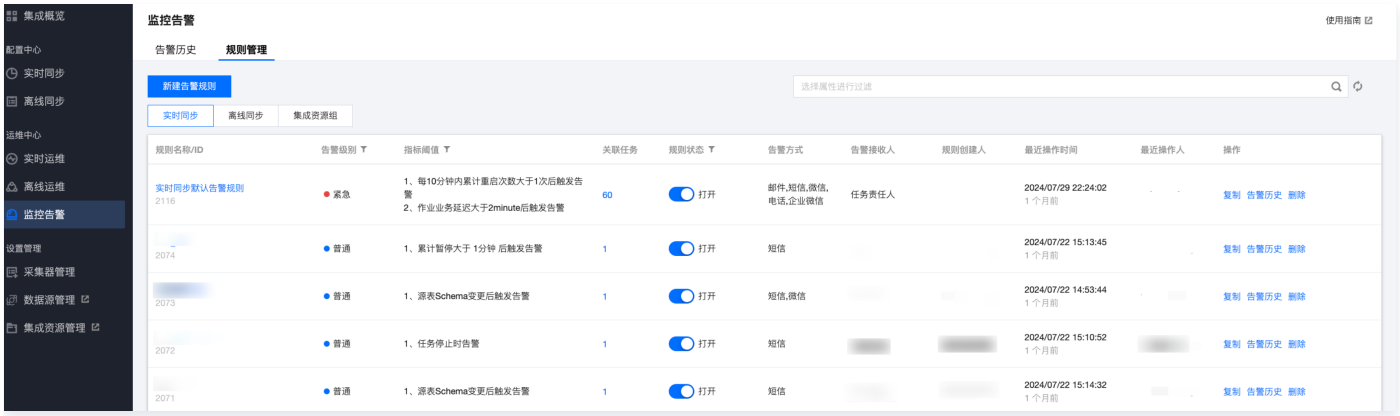
©2013-2025 Tencent Cloud. All rights reserved.

Monitoring alarm
Rule Management

Last updated: 2024-11-06 20:02:21

Rule Management List

After an alert rule is created, it will be displayed in the alert rule list. The list shows the rule name/ID, alert level, metric threshold, associated tasks, and provides features such as rule switch and alert history to help users manage and maintain alert rules.

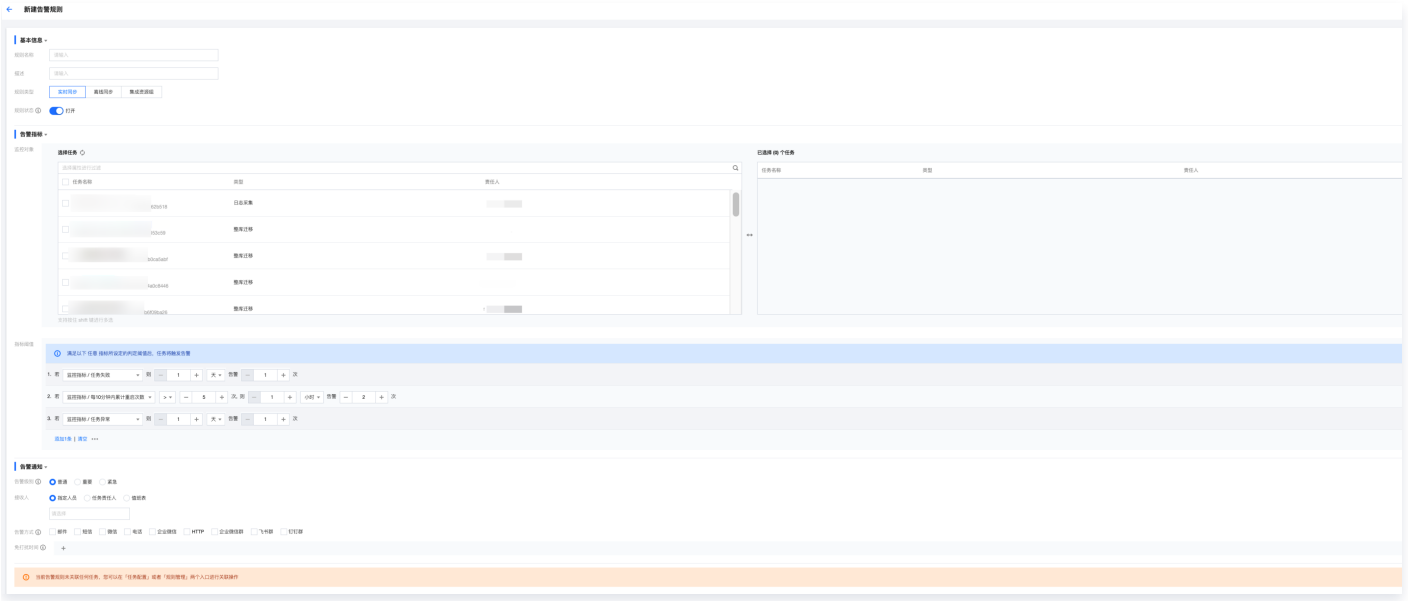


Parameter	Description
Rule Name/ID	The user-specified alert rule name or the system-generated current alert rule ID.
Alarm Level	The level corresponding to the triggered alarm rule.
Indicator Threshold	Real-time sync tasks currently support configuring task exceptions, task detection exceptions, total restart times within every 10 minutes, write latency during the runtime, total restart times within every 20 minutes, and 17 other indicator alerts.
Associated tasks	Displays the number of rules associated with the current task.
Rule Status	On: Activate the current alert rule. Alerts will be triggered based on the rule and threshold when tasks are running. Off: Deactivate the current alert rule.
Alarm Method	The method of alert notification configured in the current alert rule.
Alarm recipient	The recipients of alert notifications configured in the current alert rule.
Rule Creator	The name of the creator of the current rule.
Most recent operation time	Displays the most recent operation time for the rule, including all rule attributes such as creation, updates, activation, and deactivation.
Most recent operator	The name of the most recent operator for the current rule.
Operation	• Copy: Create a new rule with all attributes identical to the specified rule. • Alert History: View all alert history information based on the current rule name. • Delete: Delete the alert rule and stop alerts for all associated tasks. **Note** Copy: Modification of the **associated task type** in the newly copied rules is not supported.

Creating alarm rule

©2013-2025 Tencent Cloud. All rights reserved.

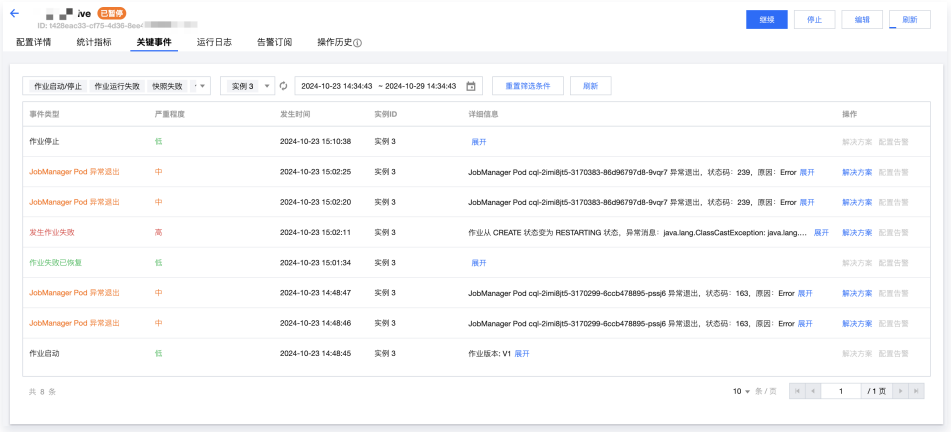
1. Log in to WeData Console.
2. Click on the project list in the left menu to find the target project for which you need to configure the alarm rule.
3. After selecting a project, click to enter the DataInLong module.
4. Click **Monitoring and Alarm > Rule Management** in the left menu to enter the alert rule management page. Click **Create New Alert Rule** and fill in the rule information.



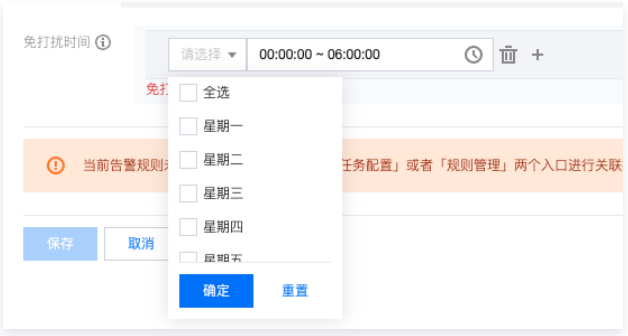
Description

Parameter		Description
Basic information	Rule name	The name of the alert rule must start with a letter or Chinese character and may contain letters, Chinese characters, numbers, and underscores.
	Description	A detailed description of the rule.
	Rule Category	Supports configuration of real-time synchronization, offline synchronization, and integrated resource group rule types.
	Rule Status	The configured rules will be effective/ineffective after being enabled/disabled.
Alarm Metrics	Monitoring object	Support selecting multiple tasks as the monitoring objects for the current rule.
	Indicator Threshold	Real-time synchronization tasks support two major categories of alert metrics: monitoring indicators and key events.  1. Monitoring indicators monitor the task's operational status and read/write conditions. They support configuring alerts for task failure, cumulative restart times within every 10 minutes, job business delay, and 17 other metrics. • State Indicators: Alerts are triggered when certain values are reached during task operations, such as failures, stops, exceptions, pauses, and restarts. For example: ○ Task Paused: Alert if the task is paused for a cumulative duration exceeding the set limit

©2013–2025 Tencent Cloud. All rights reserved.

		○ Task Stopped: Alert is triggered immediately after the task is manually or passively stopped ○ Task failed: Alert is triggered immediately when the task status is **failed** ○ Task Exception/Task Detection Exception: An alert is triggered immediately when an exception occurs during task execution ○ Task Restart: An alert is triggered immediately when the task restart exceeds the cumulative count **Note:** The task status here and the job status in key events are different concepts. For example, the occurrence of a **job failure** event in key events does not mean the task status is **failed**. This is because the task is configured with a restart policy. After a **job failure**, it will restart, keeping the task status as **running** until it reaches the limit of restart attempts, only then will the task status change to **failed**, which triggers the task failure alert. If there is no limit on the number of restarts for the task, it is suggested to configure an alert for the accumulated number of restarts within every 10 minutes, rather than configuring a task failure alert. • Read/Write Metrics: Alerts are triggered immediately when values exceed defined thresholds for read/write speed, task write latency, read/write throughput, and dirty data bytes/records. 2. Key events monitor core events during task execution. Currently, only the **Source Table Schema Change** event is supported, with other event types to be supported later. 
Alarm notification	Alarm Level	Distinguish the content of alert information to be sent according to the alert levels of different metrics. Currently, three types of alerts are available: ordinary, important, and emergency. 
	Recipient	After an alarm rule is triggered, it will send notification messages to the recipient. Currently, there are three ways to set up the recipient of alarm notifications after an alarm rule is triggered: "Specific Individuals", "Task Responsible Individuals", and "On-Call Schedules". • Designate Personnel: You can designate any one or multiple users as recipients of alarm information. • Task Owner: The task owner will receive the alarm information. • Duty Schedule: The duty schedule will be used as the recipient to send alarm information to the duty users.

©2013–2025 Tencent Cloud. All rights reserved.

	
Alarm Method	Currently, alert notifications can be sent via Email, SMS, WeChat, Telephone, Enterprise WeChat, HTTP, Enterprise WeChat group, FeiShu group, and DingTalk group. **Note** • Mobile, WeChat, Email, and other accounts can be configured in Tencent Cloud User Management. • Enterprise WeChat accounts can be configured in Tencent Cloud Federated Accounts.
Do Not Disturb Time	Supports configuring Do Not Disturb time. During this period, no messages will be sent to the alert recipient via alert notification. Alerts generated during Do Not Disturb time can be viewed in the Alarm History List. 

©2013–2025 Tencent Cloud. All rights reserved.

Alarm record

Last updated：2024-09-06 15:38:17

The Alert Page displays information of all triggered alert rules, including alert rules, alert levels, alert metrics, thresholds, and reach situations.
Supports quick retrieval by date (today/yesterday/last 7 days/last 30 days/from Definition date) and global search by task/rule/recipient name.



Parameter	Description
Alarm Time	The specific time of this triggered alarm.
Alarm Level	The level corresponding to the triggered alarm rule.
Task Name	Name of the task object that triggered the alarm. Click to jump to the Task Operation and Maintenance Page.
Rule name	The name of the rule that triggered the alert can be clicked to jump to the corresponding alert rule subscription page under the task.
Alarm Metrics	Alarm metric category configured in the current rule.
Cause of the Alarm	Displays specific metric information that triggered this alert.
Recipient	Default recipient information set by the rule.
Delivery Details	The default recipient's receipt of the alert information, including all sent successfully, partially sent successfully, and all failed to send. Click **View Details** to view the receipt status of each user in the pop-up window.  : Indicates that the notification was successfully sent to the user using this method. -- :Indicates that the current user has not configured the corresponding message notification method. Specific operations can be performed by associating the relevant category in Tencent Cloud Account Information > log in

©2013-2025 Tencent Cloud. All rights reserved.



to Method.

: Indicates that the notification was not sent successfully to the user using this method.

©2013–2025 Tencent Cloud. All rights reserved.

Real-time Node Advanced Parameters

Last updated: 2024-09-06 15:39:22

Parameter description

Type	Parameter Level	Read/Write	Applicable Scenario	Configuration	Description
Mysql / tdsql-c Mysql	Node Level	Read	Single Table + Whole Database	scan.newly-added-table.enabled=true	**Parameter Description:** Set this parameter to sense new tables after **Suspension** > **Continue**. The default is false 1. Use this parameter during full incremental synchronization. New tables will read the stock data first, then read the incremental data 2. Use this parameter during incremental synchronization. New tables will only read incremental data
		Read	Single Table + Whole Database	scan.incremental.snapshot.chunk.size=20000	**Parameter Description:** For tasks with uniform data distribution, this parameter represents the approximate number of records in one chunk. You can estimate the number of chunks by dividing the total data volume by the chunk size. The number of chunks affects whether jobmanager experiences an out-of-memory error. Currently, with 2CU, it can support more than 100,000 chunks. If the data volume is too large, you can increase the chunk size to reduce the number of chunks **Note:** For large data volume tasks (e.g., total data volume over 100 million, single record larger than 0.1M), it is generally recommended to set 20,000
		Read	Single Table + Whole Database	split-key.even-distribution.factor.upper-bound=10.0d	**Parameter Description:** During the MySQL existing data reading phase, if the data is quite scattered and the maximum value of the primary key field is extremely large, you can modify this parameter to use non-uniform splitting. This helps reduce the problem of having too many chunks due to the excessively large primary key value, which could lead to JM OOM issues **Note:** The default value is 10.0d and generally does not need to be modified
		Read	Single Table + Whole Database	debezium.query.fetch.size=0	**Parameter Description:** Represents the number of records fetched from the database each time. The default value is 0, which indicates the JDBC default fetch size **Note:** 1. For large tasks (e.g., total data volume over 100 million, individual record size over 0.1M) with only one reading instance, it is recommended to set this to 1024 records

©2013-2025 Tencent Cloud. All rights reserved.

					2. For tasks with multiple reading instances, it is recommended to lower this value to reduce memory consumption. Setting it to 512 records is advised
		Read	Single Table + Whole Database	debezium.max.queue.size=8192	**Parameter Description:** This attribute defines the maximum number of events stored in the internal queue. If this limit is reached, Debezium will pause reading new events until the unprocessed events are handled and committed. This helps prevent an excessive backlog of events in the queue, thus avoiding memory exhaustion and performance degradation. The default is 8192 **Note:** 1. For large tasks (e.g., total data volume over 100 million, individual record size over 0.1M) with only one reading instance, it is recommended to set this to 4096 2. For tasks with multiple reading instances, it is recommended to lower this value to reduce memory consumption. Setting it to 1024 is advised
	Job level	–	–	taskmanager.memory.managed.fraction=0.1	**Parameter Description:** Adjust the flink program taskmanager managed memory ratio
		–	–	table.exec.sink.upsert-materialize=NONE	**Parameter Description:** Due to the shuffle in distributed systems causing Changelog data disorder, the data received by the sink might be out of order in the global upsert. Therefore, an upsert materialization operator should be added before the upsert sink. This operator receives upstream Changelog data and generates an upsert view for downstream. This parameter controls the addition of the materialization operator **Note:** 1. By default, when unique keys encounter distributed disorder, the materialization operator will be added. It can also be set to not materialize (NONE) or force materialization (FORCE) 2. Optional values are: NONE, AUTO, FORCE
		–	–	table.exec.sink.not-null-enforcer=DROP	**Parameter Description:** Decide how to handle tasks when a NOT NULL field encounters a null value **Suggested values and functions:** 1. ERROR: A runtime exception is thrown when a NOT NULL field encounters a null value. 2. DROP: Data is directly discarded when a NOT NULL field encounters a null value
dlc	Node Level	Write	Single Table + Whole Database	write.distribution-mode=hash	**Parameter Description:** dlc concurrent writing, supports parameters none|hash (default) |range: 1. none: If there is a primary key, concurrent writing is based on the primary key; otherwise, single concurrent writing is used 2. hash: If a partition field exists, concurrent writing is based on the partition field;

©2013–2025 Tencent Cloud. All rights reserved.

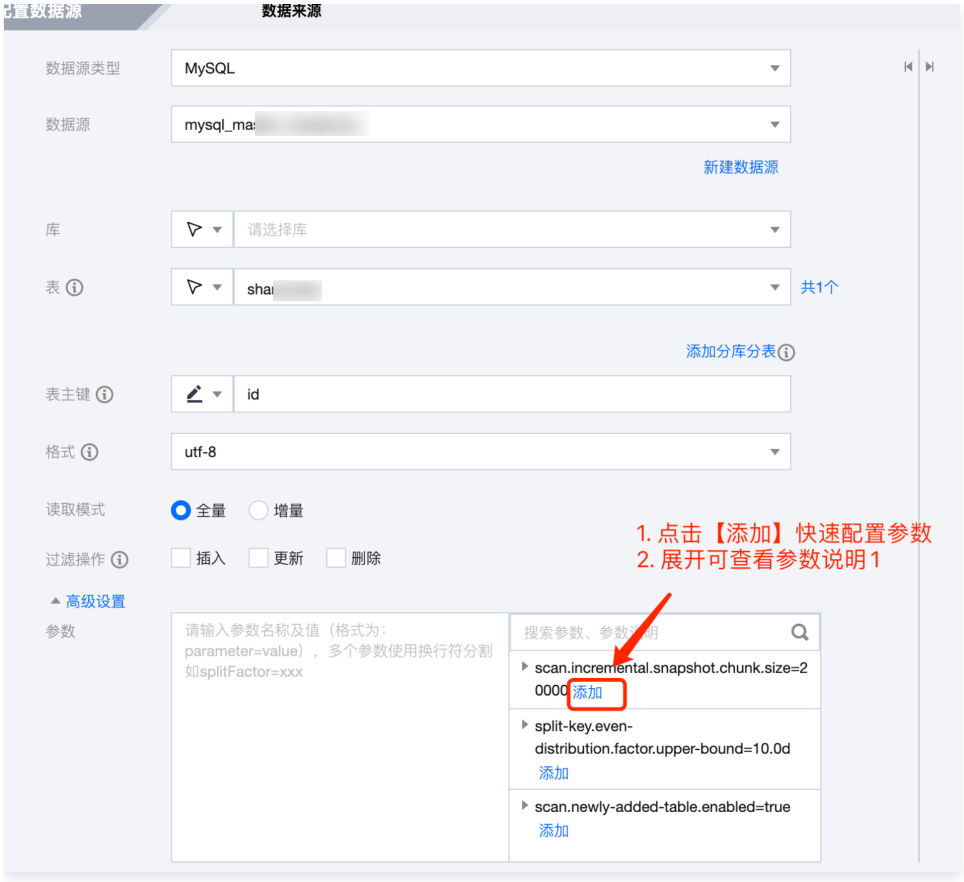
					otherwise, writing follows the 'none' parameter strategy 3. range: Currently not supported; strategy is the same as none
doris	Node Level	Write	Single Table Only	sink.properties.*=xxx	**Parameter Description:** Import Parameters for Stream Load. For example, 'sink.properties.column_separator' = ', ' **For detailed configuration, refer to:** Flink Doris Connector
	Node Level	Write	Single Table Only	sink.properties.columns=xxx	**Parameter Description:** Configure the function mapping relationship of columns. For example, 'sink.properties.columns' = 'dt,page,user_id,user_id=to_bitmap(user_id)' **Refer to:** Doris Stream Load
	Node Level	Write	Single Table + Whole Database	• sink.batch.size = 100000 • sink.batch.bytes= 83886080 • sink.batch.interval= 10s	**Parameter Description:** Improve the efficiency of writing to Doris **Note:** tm cu is recommended to be set to 2CU to avoid tm oom
oracle	Node Level	Read	Single Table + Whole Database	• 'debezium.log.mining.strategy' = 'online_catalog' • 'debezium.log.mining.continuous.mine' = 'true'	**Parameter Description:** Enabling this parameter can reduce data synchronization latency and reduce redo log storage. Suitable for single table synchronization + whole database synchronization (specified table) **Note:** 1. After setting, you cannot perceive new tables. If configured to synchronize all database tables / specified databases, new table data cannot be read 2. Not applicable to Oracle 19 version. This parameter is not supported; therefore, it needs to be set to false, otherwise the task will fail. **References:** General FAQ
	Node Level	Read	Single Table + Whole Database	debezium.lob.enabled=false	**Parameter Description:** Whether to synchronize blob type data, default is false **Note:** 1. Setting it to true may affect synchronization performance 2. Oracle default recommended configuration is false
mongodb	Node Level	Read	Single Table Only	scan.incremental.snapshot.enabled=true	**Parameter Description:** Enable concurrent reading, default is false **Notes:** Supported only in MongoDB version 4.0 and above
	Node Level	Read	Single Table Only	copy.existing=false	**Parameter Description:** Whether to copy existing data from the source collection: 1. Default is true, meaning data is read from full

©2013–2025 Tencent Cloud. All rights reserved.

					2. False means data is read from increment
	Node Level	Read	Single Table Only	poll.await.time.ms	**Parameter Description:** Change events pull interval, default is 1500ms **Note:** 1. For frequently changing collections, appropriately reduce the pull interval to enhance processing timeliness 2. For slowly changing collections, appropriately increase the pull interval to reduce database pressure
	Node Level	Read	Single Table Only	poll.max.batch.size	**Parameter Description:** The maximum number of change events per batch, default is 1,000 **Note:** Increasing this parameter will speed up pulling change events from the Cursor, but it will increase memory overhead
	Node Level	Read	Single Table Only	scan.incremental.snapshot.chunk.size.mb	**Parameter Description:** Incremental snapshot chunk size, unit is MB, default size is 64MB
	Node Level	Read	Single Table Only	changelog.normalize.enabled	**Parameter Description:** Whether to enable changelogNormalize operator, default is true, meaning enabled **Note:** MongoDB lacks -u Message. Enabling this operator will compensate for the -u Message but will consume some performance. Disabling this operator will increase transfer speed but delete operation cannot be synchronized downstream, other operations are unaffected

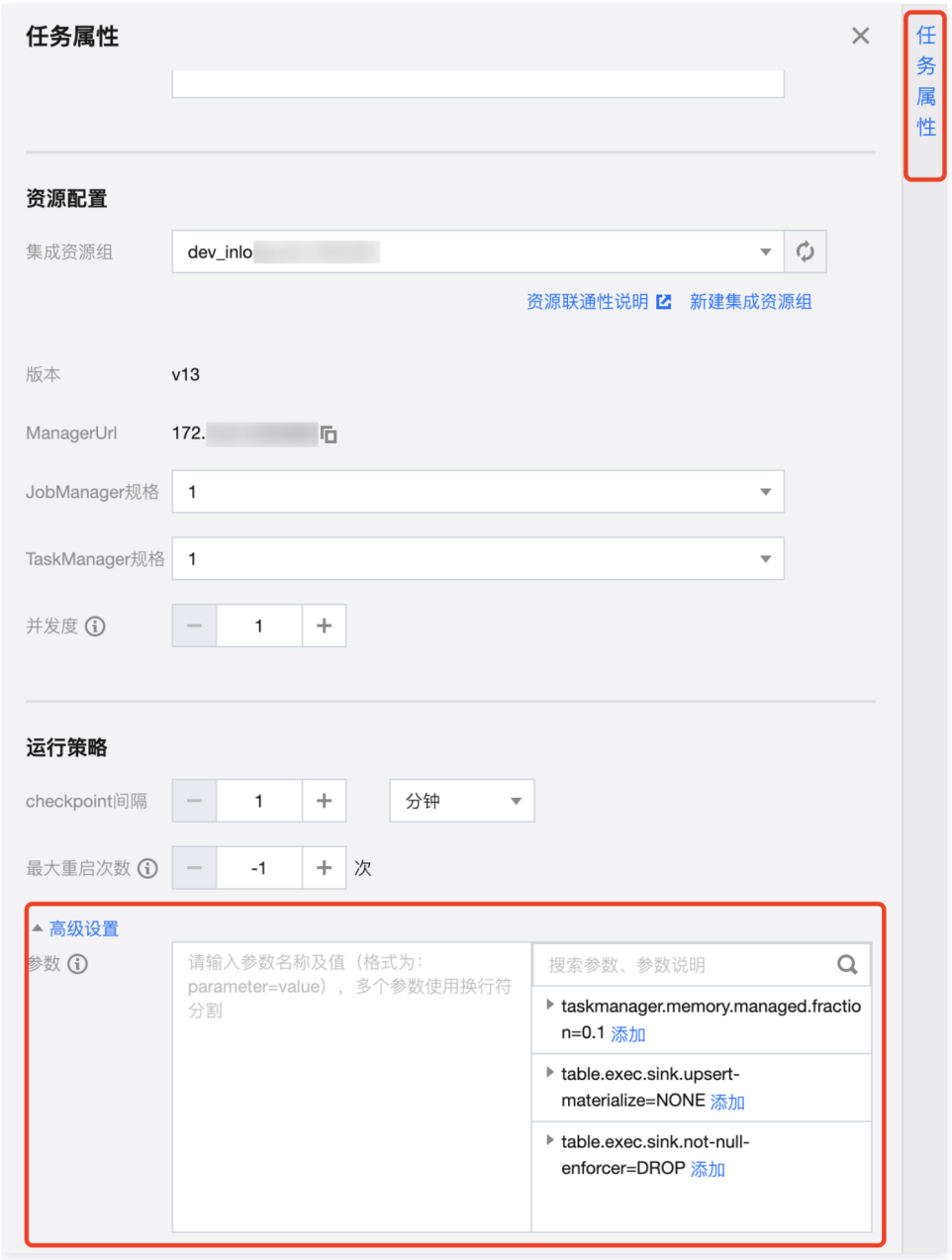
Node-level Configuration

©2013-2025 Tencent Cloud. All rights reserved.



Task-level Configuration

©2013-2025 Tencent Cloud. All rights reserved.



Note

1. One parameter per line; parameters that need to be used together should be written on the same line.
2. Each parameter comes with a default value.

©2013–2025 Tencent Cloud. All rights reserved.

Offline synchronization task configure and operation and maintenance
Supported Data Sources for Offline Synchronization
Data Source Overview

Last updated: 2024-12-12 17:54:43

DataInLong includes offline support for various data sources such as relational databases, big data storage, semi-structured data, NoSQL, and message queues. You can choose the corresponding feature modules for synchronization tasks based on the support of each module for the data sources.

Data Source		Offline Synchronization			Version
		Offline Read	Offline Write	Does it support batch task configuration?	
Relational database	MySQL	✓	✓	✓	5.6,5.7,8.0.x
	TDSQL-C MySQL	✓	✓	✓	-
	TDSQL MySQL	✓	✓	✓	InnoDB Type
	PostgreSQL	✓	✓	✓	9.6,10,11,12
	TDSQL for PostgreSQL	✓	✓	✓	5.21.6 Row and Column Storage Fusion Edition
	SQL Server	✓	✓	✓	2008,2009R2,2012,2014,2016,2017,2019
	Oracle	✓	✓	✓	11,12,19
	DB2	✓	✓	✓	-
	SAP HANA	✓	✓	✓	-
	SAP IQ(Sybase)	✓	-	✓	-
	DM	✓	✓	✓	7,8
	OceanBase (beta usage)	✓	✓	✓	3.1.4,4.2.1.3
Big data	TCHouse-P	✓	✓	✓	-
	Hive	✓	✓	✓	1.x,2.x,3.x
	DLC	✓	✓	✓	-
	Doris	✓	✓	✓	0.15,1.x
	StarRocks	✓	✓	✓	-
	ClickHouse	✓	✓	✓	20.7+

©2013-2025 Tencent Cloud. All rights reserved.

	Iceberg	✓	✓	✓	0.13.1+
	HBase	✓	✓	✓	2.2.x
	HDFS	✓	✓	✓	2.x,3.x
	Kudu	✓	✓	✓	–
	BigQuery	✓	–	–	–
	TBase	✓	✓	✓	–
	GaussDB	✓	✓	✓	–
	GBase	✓	✓	✓	–
	Impala	✓	✓	✓	–
	Greenplum	✓	✓	✓	4.x,5.x,6.x
Semi-structured	COS	✓	✓	✓	–
	Rest API	✓	–	–	–
	FTP	✓	✓	✓	–
	SFTP	✓	✓	✓	–
NoSQL	Redis	–	✓	–	–
	Elasticsearch	✓	✓	–	6.x,7.x
	CTSDB influxDB	✓	✓	✓	–
	TDengine	✓	✓	–	
	Mongo	✓	✓	–	4.x
	Alibaba Graph Database	✓	✓	–	Only Script Mode supported
Message Queue	Kafka	✓	✓	–	2.4.1,2.7.1,2.8.1,2.8.2 (including Kafka, Ckafka data sources)
	TDMQ Pulsar	–	✓	✓	Virtual Cluster Edition

Note
Data link read/write ends must meet the corresponding data source versions.

©2013–2025 Tencent Cloud. All rights reserved.

Supported Data Sources for Offline Synchronization MySQL Data Source

Last updated: 2024-12-12 17:55:43

MySQL Environment Preparation and Database Configuration

If you want to perform offline data synchronization using MySQL, you need to first confirm the supported MySQL data source versions and the supported data type conversions for reading.

Supported Editions

You need to follow these version limitations:

Type	Version	Driver
MySQL	5.6，5.7，8.0.x	JDBC Driver：8.0.21

Confirm MySQL Version

DataInLong has requirements for MySQL versions. Check whether the MySQL version to be synced meets these requirements. You can check the current MySQL database version with the following statement.

```
select version();
```

Setting MySQL Server Permissions

You can define a MySQL user with appropriate permissions.

1. **Create a MySQL User (Optional):**

```
mysql> CREATE USER 'user'@'localhost' IDENTIFIED BY 'password';
```

2. **Grant Necessary Permissions to the User:**

For offline data synchronization, this account must have the SELECT privilege for the database. Refer to the following example for the command:

```
mysql> GRANT SELECT ON *.* TO 'user'@'localhost';
```

3. **Refresh the User's Permissions:**

```
mysql>FLUSH PRIVILEGES;
```

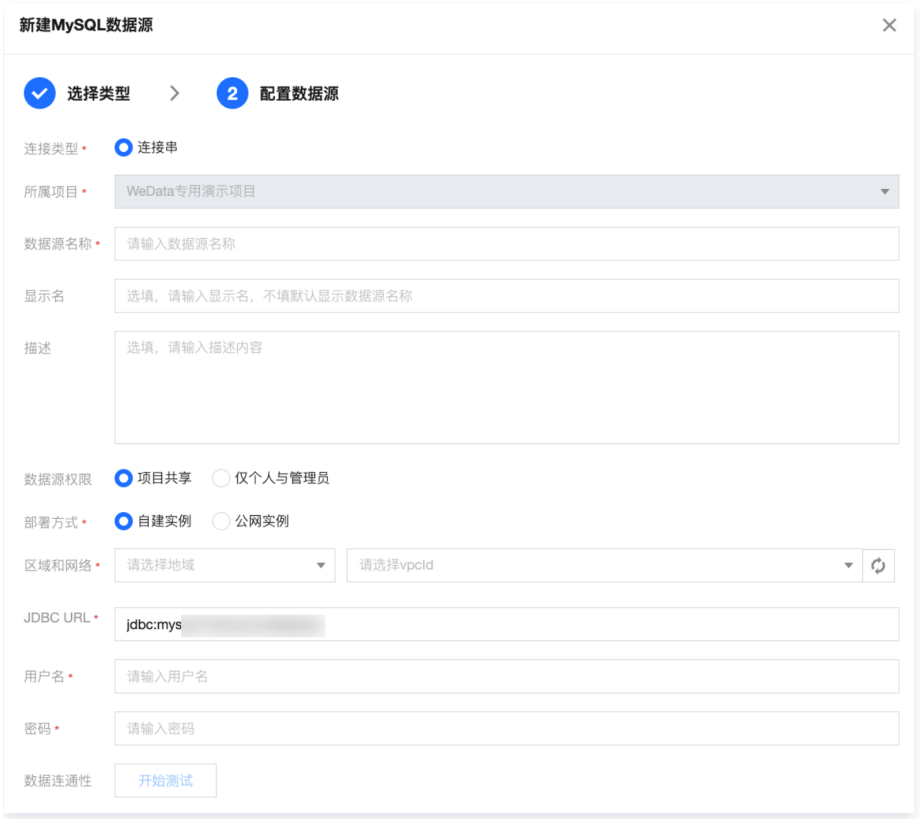
Data source configuration

Enter the data source configuration interface, MySQL data sources support two connection methods: cloud instance and connection string.

click **Project Management** > **Data Source Management** > **New Data Source** > **Choose MySQL Data Source**.

- Create a data source via connection string.

©2013-2025 Tencent Cloud. All rights reserved.



Parameters	Description
Data Source Name	The name of the created data source is user-defined and cannot be empty. It starts with a letter and can contain letters, digits, and underscores.
Description	Optional: Description of this data source
Data Source Permissions	Project Shared means that all members of the current data source project can use it, and Individual and Administrator Only means that the data source is only available to the creator and project administrators.
Deployment Methods	Supports two deployment methods: self-built instances and public network instances. Self-built instances are data source instances deployed on Tencent CVM, while public network instances are those in a customer's local IDC or other cloud resources, accessible via the public network.
Region and Network	When choosing a self-built instance, you need to select the region and vpcID where the data source instance is located.
JDBC URL	The connection string information used to connect to MySQL data source instances, including host IP, port, database name, etc.
Username	The username used to connect a database.
Password	The password used to connect a database.
Data Connectivity	Test whether the configured database can be connected. **Note** • If the connectivity test fails, the data source can still be saved. However, it cannot be used if the connectivity test fails and the data source is saved. • If the connectivity test fails, it may be because WeData is blocked by the network firewall where the database is located. You need to add Tencent Cloud MySQL Database Security Group.

MySQL Offline Single Table Read Node Configuration

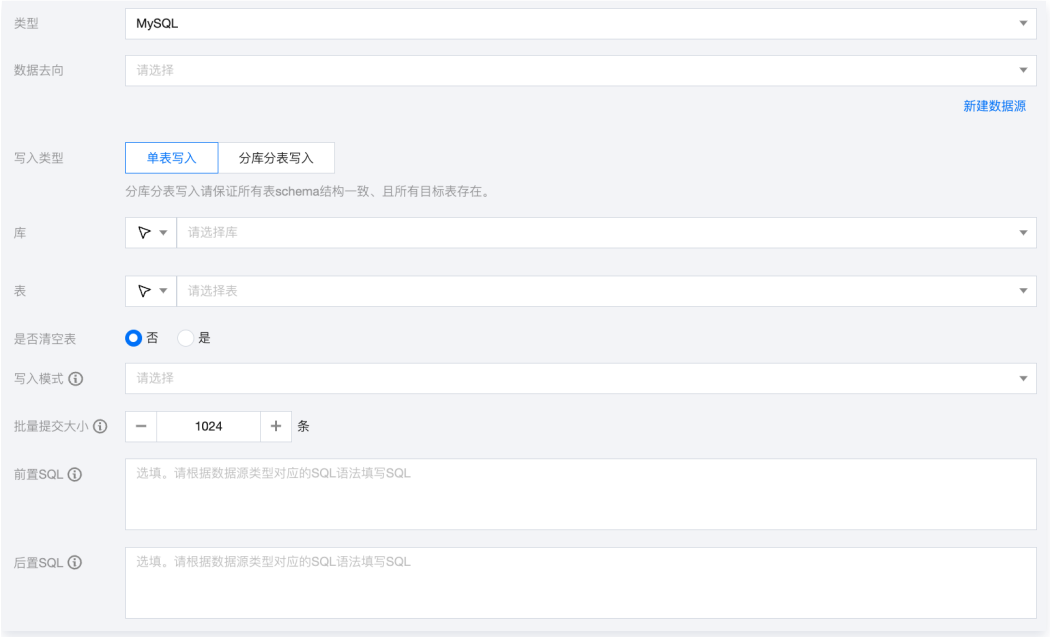
©2013-2025 Tencent Cloud. All rights reserved.



Parameters	Description
Data Source	Available MySQL data sources.
Database	Support selection or manual input of the library name to be read. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Support selecting or manually entering the Table Name to be read. • In the case of table partitioning, multiple table names can be selected or entered at the MySQL source end. The structure of multiple tables must be consistent. • In the case of table partitioning, supports configuring the table serial number range. For example, 'table_[0-99]' indicates reading 'table_0', 'table_1', 'table_2' up to 'table_99'; if the length of your table's numeric suffix is consistent, such as 'table_000', 'table_001', 'table_002' up to 'table_999', you can configure it as '"table": ["table_00[0-9]", "table_0[10-99]", "table_[100-999]"]'. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Adding Sharding	Suitable for sharding scenarios, click to configure multiple data sources, databases, and table information. In sharding and partitioning scenarios, ensure all table structures are consistent. The task configuration will default to display and use the first table structure for data acquisition.
Split Key	Specify the field used for data sharding. After specifying, concurrent tasks will be started for data synchronization to improve efficiency. You may use a column in the source data table as the split key; it is recommended to use a primary key or an indexed column as the split key. Only integer-type fields are supported.
Filter Condition (Optional)	In actual business scenarios, it is common to choose the current day's data for synchronization, specifying the where condition as gmt_create>$bizdate. The where condition can effectively perform business incremental synchronization. • If the where clause is not specified, including not providing the key or value for where, data synchronization will be considered as full data synchronization. • The where condition cannot be specified as limit 10, as this does not conform to MySQL WHERE clause constraints.
Advanced Settings (optional)	You can configure parameters based on business requirements.

©2013-2025 Tencent Cloud. All rights reserved.

MySQL Offline Single Table Write Node Configuration



Parameters	Description
Data Flow Direction	The MySQL data source that needs to be written to.
Write Type	• Single Table Write: Supports writing to a single table • Sharding write: Supports writing to sharded MySQL tables within the same data source at once. Ensure all table schemas are consistent and all target tables exist for sharding write.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Supports selection, or manual input of the table name to be written to • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. • Data synchronization can still be performed given network connection in DataInLong.
Clear Table	Before writing to the MySQL data table, you can manually choose whether to clear the data table.
Write mode	MySQL write supports three modes: • Append: If there is a primary key/unique index conflict, the conflicting rows cannot be written. • Overwrite: In case of a primary key/unique index conflict, the original row will be deleted, and a new row will be inserted. • On duplicate key: When a primary key/unique index conflict occurs, the new row will update the specified fields. The specified fields refer to the fields added in the field mapping configuration of the synchronization task.
Batch Submission Size	The number of records submitted in a batch at once. This value can greatly reduce the number of network interactions between the data synchronization system and MySQL, improving overall throughput. Setting this value too high may cause the data synchronization process to encounter an OOM (Out Of Memory) exception.
Pre-SQL (optional)	SQL command executed before the synchronization task. Fill in the SQL according to the correct SQL syntax corresponding to the data source type. For example, clear old data in the table before execution (truncate table tablename).

©2013-2025 Tencent Cloud. All rights reserved.

Post-SQL (optional)	SQL statements executed after the synchronization task are written according to the correct SQL syntax corresponding to the data source type, for example, adding a timestamp with `ALTER TABLE tablename ADD colname TIMESTAMP DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP`.

Data type conversion is supported

Read

The field types and type conversion mappings supported by MySQL read are as follows (When processing MySQL, the data types of the MySQL data source are mapped to the data types of the data processing engine):

MySQL Data Type	Internal Type
tinyint，smallint，mediumint，int，bigint，year	Long
float，double，decimal	Double
varchar，char，tinytext，text，mediumtext，longtext，set，json	String
date，datetime，timestamp，time	Date
bit，bool	Boolean
tinyblob，mediumblob，blob，longblob，varbinary	Bytes

Write

The field types and type conversion mappings supported by MySQL write are as follows:

Internal Type	MySQL Data Type
Long	tinyint，smallint，mediumint，int，bigint，year
Double	float，double，decimal
String	varchar，char，tinytext，text，mediumtext，longtext，set，json
Date	date，datetime，timestamp，time
Boolean	bit，bool
Bytes	tinyblob，mediumblob，blob，longblob，varbinary

FAQs

1. Using COLLATE to specify collation may cause some data synchronization duplicates

Cause:
COLLATE is used to specify field collation. COLLATE affects the order of the ORDER BY statement, the results of greater than and less than conditions in the WHERE clause, as well as DISTINCT, GROUP BY, HAVING, and other query results. Any character type comparison or sorting relates to COLLATE.
Solutions:
When setting the partition key for integration tasks, choose a non-COLLATE column.

MySQL Script Demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

Configure single database single table

```
"job": {
```

©2013-2025 Tencent Cloud. All rights reserved.

```
    "content": [
        {
            "reader": {
                "parameter": {
                    "password": "******",
                    "column": [  // Column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": [
                                "jdbc:mysql://ip:3306/database?
rewriteBatchedStatements=true&tinyInt1isBit=false&serverTimezone=Asia/Shanghai&characterEncoding=utf8&auto
Reconnect=true"
                            ],
                            "table": [ // Source table
                                "source_table"
                            ]
                        }
                    ],
                    "where": "id>10",  // Filters
                    "splitPk": "id",  // Split key
                    "username": "root"
                },
                "name": "mysqlreader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "postSql": [ // Post-SQL
                        ""
                    ],
                    "password": "******",
                    "session": [
                        "set session sql_mode='ANSI'"
                    ],
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": "jdbc:mysql://ip:3306/database?
rewriteBatchedStatements=true&tinyInt1isBit=false&serverTimezone=Asia/Shanghai&characterEncoding=utf8&auto
Reconnect=true",
                            "table": [ // Target table
                                "sink_table"
                            ]
                        }
                    ],
                    "writeMode": "replace", // Write mode
                    "batchSize": 1024, // Batch submission size
                    "username": "root",
                    "on": [ // Unique Index, different writing modes will have different writing
strategies when conflicts occur during the writing process
                        "id"
                    ],
                    "preSql": [  // Pre-SQL
                        "truncate table sink_table"
                    ]
```

©2013–2025 Tencent Cloud. All rights reserved.

```
                },
                "name": "mysqlwriter"
            }
        }
    ],
    "setting": {
        "errorLimit": {
            "record": 0, // Dirty Data Threshold
            "percentage": 0.02 // Dirty Data Percentage 1 represents 100%
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

Configure Sharding

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "password": "******", // Can be empty, one of the required parameters
                    "column": [ // Column name, must ensure all table structures are the same
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "password": "******", // Password for the data source of the first table
                            "jdbcUrl": [
                                "jdbc:mysql://ip:3306/database?
rewriteBatchedStatements=true&tinyInt1isBit=false&serverTimezone=Asia/Shanghai&characterEncoding=utf8&auto
Reconnect=true"
                            ],
                            "table": [
                                "source_table_1"
                            ],
                            "username": "root"
                        },
                        {
                            "password": "******", // Password for the data source of the second table
                            "jdbcUrl": [
                                "jdbc:mysql://ip:3306/database?
rewriteBatchedStatements=true&tinyInt1isBit=false&serverTimezone=Asia/Shanghai&characterEncoding=utf8&auto
Reconnect=true"
                            ],
                            "table": [
                                "source_table_2"
                            ],
                            "username": "root"
                        }
                    ],
                    "splitPk": "id", // Split key
                    "username": "root" // One of the required parameters
                },
                "name": "mysqlreader"
```

©2013–2025 Tencent Cloud. All rights reserved.

```
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "postSql": [ // Post-SQL
                        ""
                    ],
                    "password": "******",
                    "session": [
                        "set session sql_mode='ANSI'"
                    ],
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": "jdbc:mysql://ip:3306/database?
rewriteBatchedStatements=true&tinyInt1isBit=false&serverTimezone=Asia/Shanghai&characterEncoding=utf8&auto
Reconnect=true",
                            "table": [ // Target table
                                "sink_table"
                            ]
                        }
                    ],
                    "writeMode": "replace", // Write mode
                    "batchSize": 1024, // Batch submission size
                    "username": "root",
                    "on": [ // Unique Index, different writing modes will have different writing
strategies when conflicts occur during the writing process
                        "id"
                    ],
                    "preSql": [  // Pre-SQL
                        ""
                    ]
                },
                "name": "mysqlwriter"
            }
        }
    ],
    "setting": {
        "errorLimit": {
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

TDSQL-C for MySQL Data Source

Last updated: 2024-12-12 17:56:24

TDSQL-C for MySQL Offline Single Table Read Node Configuration



Parameters	Description
Data Source	Available TDSQL-C for MySQL data sources.
Database	Support selecting or manually entering the name of the database to be read • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Support selecting or manually entering the name of the table to be read • In the case of sharding, multiple table names can be selected or entered in the TDSQL-C for MySQL source. All tables must have consistent structures. • In the case of table partitioning, supports configuring the table serial number range. For example, 'table_[0-99]' indicates reading 'table_0', 'table_1', 'table_2' up to 'table_99'; if the length of your table's numeric suffix is consistent, such as 'table_000', 'table_001', 'table_002' up to 'table_999', you can configure it as '"table": ["table_00[0-9]", "table_0[10-99]", "table_[100-999]"]'. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Adding Sharding	You can create multiple data sources and add corresponding table objects. **Note:**

©2013-2025 Tencent Cloud. All rights reserved.

	When selecting multiple table objects in a sharded database, ensure that the Schema information is consistent (including field names and field types). The system's Data Field module will default to display metadata field information for the first table of the first data source. Inconsistencies in fields between multiple tables may cause execution to fail.
Split Key	Specify the field used for data sharding. After specifying, concurrent tasks will be started for data synchronization to improve efficiency. You may use a column in the source data table as the split key; it is recommended to use a primary key or an indexed column as the split key. Only integer-type fields are supported.
Filter Condition (Optional)	In actual business scenarios, it is common to choose the current day's data for synchronization, specifying the where condition as gmt_create>$bizdate. The where condition can effectively perform business incremental synchronization. • If the where clause is not specified, including not providing the key or value for where, data synchronization will be considered as full data synchronization. • You cannot specify the where condition as limit 10, which does not comply with TDSQL-C for MySQL WHERE clause constraints.
Advanced Settings (optional)	You can configure parameters based on business requirements.

TDSQL-C for MySQL Offline Single Table Write Node Configuration



Parameters	Description
Data Flow Direction	TDSQL-C for MySQL data source to be written.

©2013-2025 Tencent Cloud. All rights reserved.

Write Type	• Single Table Write: Supports writing to a single table • Sharding Write: Supports writing to sharded tables in the same data source in TDSQL-C for MySQL in one go. Ensure all table schemas are consistent, and all target tables exist.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Supports selection, or manual input of the table name to be written to • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Clear Table	You can manually choose whether to clear the TDSQL-C for MySQL data table before writing to it.
Write mode	TDSQL-C for MySQL supports three writing modes: • Append: If there is a primary key/unique index conflict, the conflicting rows cannot be written. • Overwrite: If there is a primary key/unique index conflict, the original rows will be deleted first, and then new rows will be inserted. • On duplicate key: If there is a primary key/unique index conflict, the new rows will replace the specified fields of the existing rows.
Batch Submission Size	The record count size for one-time batch submission can significantly reduce the number of network interactions between the data synchronization system and TDSQL and improve overall throughput. If this value is set too high, it can cause the data synchronization process to run into OOM exceptions.
Pre-SQL (optional)	SQL command executed before the synchronization task. Fill in the SQL according to the correct SQL syntax corresponding to the data source type. For example, clear old data in the table before execution (truncate table tablename).
Post-SQL (optional)	SQL statements executed after the synchronization task are written according to the correct SQL syntax corresponding to the data source type, for example, adding a timestamp with `ALTER TABLE tablename ADD colname TIMESTAMP DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP`.

Data type conversion is supported

Read

Supported data types and their corresponding conversions in TDSQL-C for MySQL read are as follows (when processing TDSQL-C for MySQL data, the data types from TDSQL-C for MySQL are mapped to the data types of the data processing engine):

MySQL Data Type	Internal Type
tinyint, smallint, mediumint, int, bigint, year	Long
float, double, decimal	Double
varchar, char, tinytext, text, mediumtext, longtext, set, json	String
date, datetime, timestamp, time	Date
bit, bool	Boolean
tinyblob, mediumblob, blob, longblob, varbinary	Bytes

Write

Supported data types and their corresponding conversions in TDSQL-C for MySQL write are as follows:

©2013-2025 Tencent Cloud. All rights reserved.

Internal Type	MySQL Data Type
Long	tinyint，smallint，mediumint，int，bigint，year
Double	float，double，decimal
String	varchar，char，tinytext，text，mediumtext，longtext，set，json
Date	date，datetime，timestamp，time
Boolean	bit，bool
Bytes	tinyblob，mediumblob，blob，longblob，varbinary

TDSQL-C for MySQL Script Demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "password": "******",
                    "column": [  // Column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": [
                                "jdbc:mysql://ip:3306/database?
rewriteBatchedStatements=true&tinyInt1isBit=false&serverTimezone=Asia/Shanghai&characterEncoding=utf8&auto
Reconnect=true"
                            ],
                            "table": [ // Source table
                                "source_table"
                            ]
                        }
                    ],
                    "where": "id>10",  // Filters
                    "splitPk": "id",  // Split key
                    "username": "root"
                },
                "name": "mysqlreader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "postSql": [ // Post-SQL
                        "alter table sink_table add time timestamp DEFAULT CURRENT_TIMESTAMP ON UPDATE
CURRENT_TIMESTAMP"
                    ],
                    "password": "******",
                    "session": [
                        "set session sql_mode='ANSI'"
                    ],
                    "column": [ //column name
                        "id",
                        "name"
                    ],
```

©2013-2025 Tencent Cloud. All rights reserved.

```
                    "connection": [
                        {
                            "jdbcUrl": "jdbc:mysql://ip:3306/database?
rewriteBatchedStatements=true&tinyInt1isBit=false&serverTimezone=Asia/Shanghai&characterEncoding=utf8&auto
Reconnect=true",
                            "table": [ // Target table
                                "sink_table"
                            ]
                        }
                    ],
                    "writeMode": "replace", // Write mode
                    "batchSize": 1024, // Batch submission size
                    "username": "root",
                    "on": [ // Unique Index, different writing modes will have different writing
strategies when conflicts occur during the writing process
                        "id"
                    ],
                    "preSql": [  // Pre-SQL
                        "truncate table abc1"
                    ]
                },
                "name": "mysqlwriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

Oracle Data Source

Last updated：2024-12-12 17:56:57

Supported Editions

The driver version used is ojdbc6, and it currently supports Oracle 11g, 12c, 18c, and 19c versions, with new features of 18c and 19c not supported.

Use Limits

When synchronizing data with DataInLong, only UTF8, AL32UTF8, AL16UTF16, and ZHS16GBK encoding formats are supported.

> **Note**
> Supports reading View Tables.

Oracle Offline Single Table Read Node Configuration

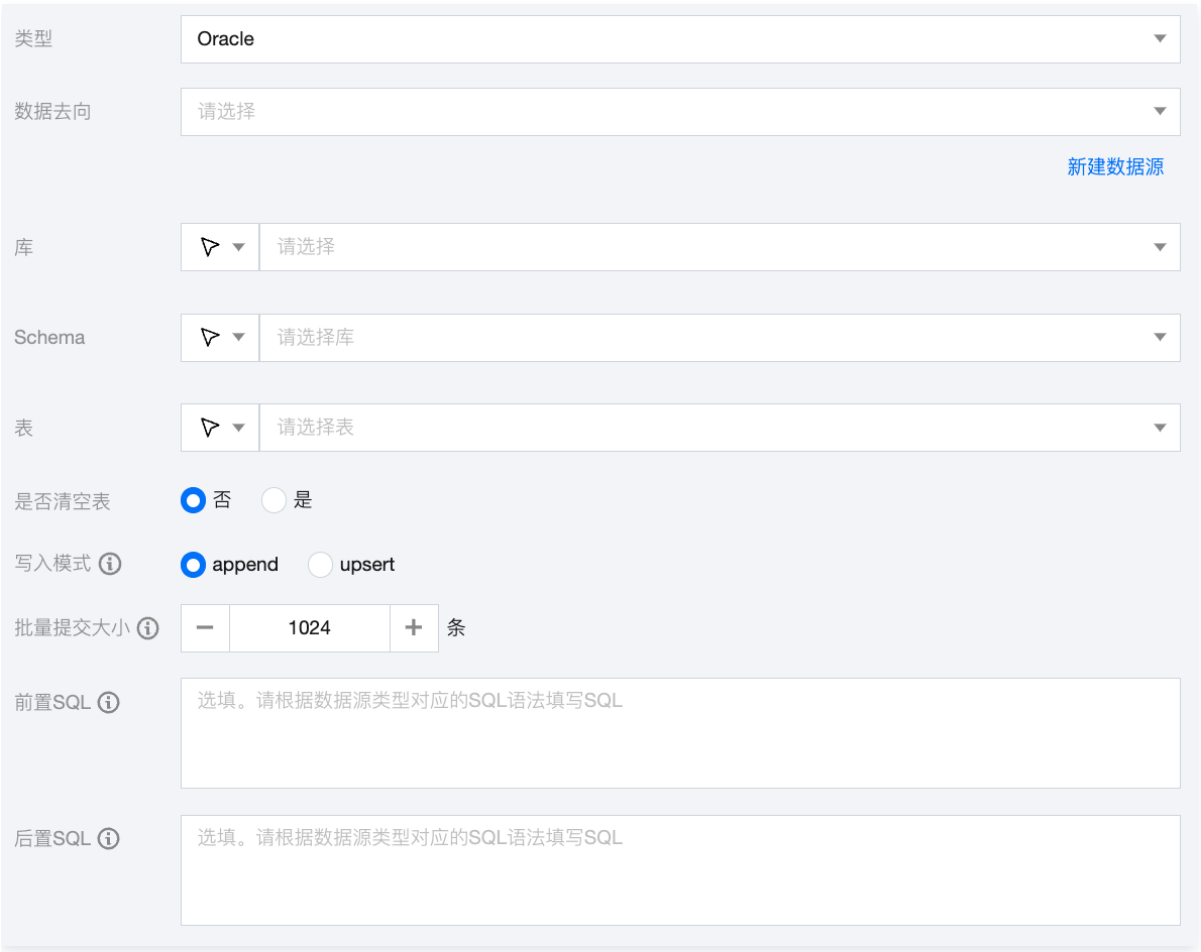


Parameters	Description
Data Source	Available Oracle data sources.
Database	Support selecting or manually entering the name of the database to be read • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Support selecting or manually entering the Schema to be read.
Table	Support selecting or manually entering the Table Name to be read.
Adding Sharding	You can create multiple data sources and add corresponding table objects.

©2013-2025 Tencent Cloud. All rights reserved.

	Note: When selecting multiple table objects in a sharded database, ensure that the Schema information is consistent (including field names and field types). The system's Data Field module will default to display metadata field information for the first table of the first data source. Inconsistencies in fields between multiple tables may cause execution to fail.
Split Key	• Specify the field for data sharding. Once specified, concurrent tasks will start for data synchronization, improving synchronization efficiency. • You can use a column in the source data table as the partition key. It is recommended to use the primary key or an indexed column as the partition key. If the field needs to be configured as a String, Floating Point, Date, or other types, please enter it manually.
Filter Condition (Optional)	Oracle constructs SQL based on the specified WHERE Condition and extracts data accordingly. For example, during testing, specify the WHERE Condition as row_number(). • WHERE Condition can effectively facilitate business incremental synchronization. • If the WHERE Condition is not configured or is empty, it will be treated as full table synchronization.

Oracle Offline Single Table Write Node Configuration



Parameters	Description
Data Flow Direction	The Oracle data source to be written to is required.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered.

©2013–2025 Tencent Cloud. All rights reserved.

	• When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Supports selection or manual input of the Oracle data pattern to be written.
Table	Supports selection, or manual input of the table name to be written to • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Clear Table	Before writing to the Oracle data table, you can manually choose whether to clear the data table.
Batch Submission Size	The record count size for one-time batch submission can greatly reduce the network interactions between the data synchronization system and Oracle, thereby enhancing overall throughput. If this value is set too large, it may cause the data synchronization running process to experience OOM exceptions.
Write mode	Oracle write supports two modes: append: Append Write. upsert: Data update and write based on the primary key field setting.
Pre-SQL (optional)	SQL command executed before the synchronization task. Fill in the SQL according to the correct SQL syntax corresponding to the data source type. For example, clear old data in the table before execution (truncate table tablename).
Post-SQL (optional)	SQL statements executed after the synchronization task are written according to the correct SQL syntax corresponding to the data source type, for example, adding a timestamp with `ALTER TABLE tablename ADD colname TIMESTAMP DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP`.

Data type conversion is supported

Read

The supported data types for reading in Oracle and their corresponding conversion relationships are as follows (When processing Oracle, it first maps the data types of the Oracle data source to the data types of the data processing engine):

Oracle Data Type	Internal Type
NUMBER，INTEGER，INT，SMALLINT	Long
NUMERIC，DECIMAL，FLOAT，DOUBLE PRECISION，REAL	Double
LONG，CHAR，NCHAR，VARCHAR，VARCHAR2，NVARCHAR2，CLOB，NCLOB，CHARACTER，CHARACTER VARYING，CHAR VARYING，NATIONAL CHARACTER，NATIONAL CHAR，NATIONAL CHARACTER VARYING，NATIONAL CHAR VARYING，NCHAR VARYING	String
TIMESTAMP，DATE	Date
bit，bool	Boolean
BLOB，BFILE，RAW，LONG RAW	Bytes

Write

The supported data types for writing in Oracle and their corresponding conversion relationships are as follows:

Internal Type	Oracle Data Type
Long	NUMBER，INTEGER，INT，SMALLINT
Double	NUMERIC，DECIMAL，FLOAT，DOUBLE PRECISION，REAL

©2013-2025 Tencent Cloud. All rights reserved.

String	LONG，CHAR，NCHAR，VARCHAR，VARCHAR2，NVARCHAR2，CLOB，NCLOB，CHARACTER，CHARACTER VARYING，CHAR VARYING，NATIONAL CHARACTER，NATIONAL CHAR，NATIONAL CHARACTER VARYING，NATIONAL CHAR VARYING，NCHAR VARYING
Date	TIMESTAMP，DATE
Boolean	bit，bool
Bytes	BLOB，BFILE，RAW，LONG RAW

Oracle Script Demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "password": "******",
                    "column": [ //column name
                        "ID",
                        "NAME"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": [
                                "jdbc:oracle:thin:@ip:1521:DATABASE"
                            ],
                            "table": [ // Source table
                                "SCHEMA.SOURCE_TABLE"
                            ]
                        }
                    ],
                    "where": "ID>10", // filters
                    "splitPk": "ID", // split key
                    "username": "root"
                },
                "name": "oraclereader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "postSql": [ // Post-SQL
                        ""
                    ],
                    "password": "******",
                    "column": [ //column name
                        "ID",
                        "NAME"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": "jdbc:oracle:thin:@ip:1521:DATABASE",
                            "table": [ // Target table
                                "SCHEMA.SINK_TABLE"
                            ]
                        }
                    ],
                    "writeMode": "upsert", // write mode
```

©2013–2025 Tencent Cloud. All rights reserved.

```
                    "batchSize": 1024, // Batch submission size
                    "username": "root",
                    "on": [ // unique index
                        "ID"
                    ],
                    "preSql": [ // Pre-SQL
                        ""
                    ]
                },
                "name": "oraclewriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

SQL Server Data Source

Last updated：2024-12-12 17:57:24

Supported Versions

Version	Driver
Azure SQL Database	sqljdbc4:4.0
SQL Server 2016	
SQL Server 2014	
SQL Server 2012	
PDW 2008R2 AU34	
SQL Server 2008R2	
SQL Server 2008	

Use Limits

Note
Supports reading View Tables.

SQL Server offline single table read node configuration

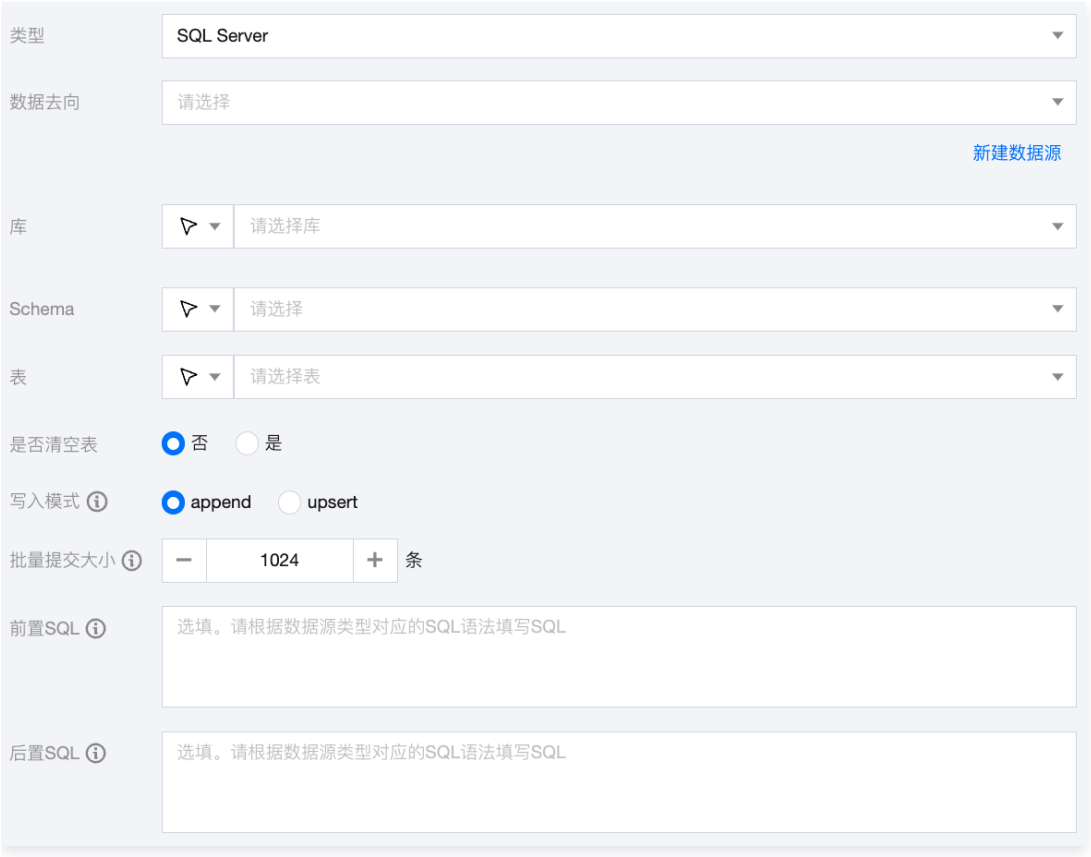


Parameters	Description
Data Source	Available SQL Server Data Source.
Database	Support selecting or manually entering the name of the database to be read • By default, the database bound to the data source is used as the default database; other databases need to be manually entered.

©2013-2025 Tencent Cloud. All rights reserved.

	• When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Support selecting or manually entering the Schema Name to be read.
Table	Supports selecting or manually entering the name of the table to read. Each task supports synchronizing only one table.
Split Key	Specify the field for data sharding. Once specified, concurrent tasks will be initiated for data synchronization. You can use a column from the source data table as the split key. It is recommended to use the primary key or an indexed column as the split key.
Filter Condition (Optional)	Fill in the corresponding filter statement based on the data type. This statement will serve as the filter condition for synchronizing data. SQL Server concatenates SQL based on the specified WHERE Condition and extracts data based on that SQL. For example, during testing, you can specify the WHERE Condition as limit 10. In actual business scenarios, you usually select the data for the current day for synchronization by specifying the WHERE Condition as gmt_create > $bizdate. • The WHERE condition can effectively perform business incremental synchronization. • If the WHERE condition is empty, it is considered to synchronize all information from the entire table.

SQL Server offline single table write node configuration



Parameters	Description
Data Flow Direction	SQL Server data source to be written to.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be

©2013–2025 Tencent Cloud. All rights reserved.

	performed given network connection in DataInLong.
Schema	Support selecting or manually entering the Schema Name to be read.
Table	Supports selection, or manual input of the table name to be written to • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Clear Table	Before writing to the SQL Server data table, you can manually choose whether to clear the data table.
Write mode	SQL Server supports two write modes: • append: Append write. • upsert: Update write according to the set primary key field.
Batch Submission Size	The size of the record count for one-time batch submission; this value can greatly reduce the number of network interactions between the data synchronization system and SQL Server, and improve overall throughput. If the value is set too high, it may cause an OOM exception in the data synchronization process.
Pre-SQL (optional)	SQL command executed before the synchronization task. Fill in the SQL according to the correct SQL syntax corresponding to the data source type. For example, clear old data in the table before execution (truncate table tablename).
Post-SQL (optional)	SQL statements executed after the synchronization task are written according to the correct SQL syntax corresponding to the data source type, for example, adding a timestamp with `ALTER TABLE tablename ADD colname TIMESTAMP DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP`.

Data type conversion is supported

Read

The supported data types for SQL Server reading and their corresponding conversion relationships are as follows (when processing SQL Server, the data types of the SQL Server data source will be mapped to the data processing engine's data types first):

SqlServer Data Type	Internal Type
bigint，int，smallint，tinyint	Long
float，decimal，real，numeric	Double
char，nchar，ntext，nvarchar，text，varchar，sysname	String
date，datetime，time，timestamp、datetimeoffset	Date
bit	Boolean
binary，varbinary	Bytes

Write

The supported data types for SQL Server writing and their corresponding conversion relationships are as follows:

Internal Type	SqlServer Data Type
Long	bigint，int，smallint，tinyint
Double	float，decimal，real，numeric
String	char，nchar，ntext，nvarchar，text，varchar，sysname
Date	date，datetime，time，timestamp

©2013-2025 Tencent Cloud. All rights reserved.

Boolean	bit
Bytes	binary，varbinary

SQL Server Script Demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": [
                                "jdbc:sqlserver://ip:1433;DatabaseName=database"
                            ],
                            "table": [ // Source table
                                "schema.source_table"
                            ]
                        }
                    ],
                    "where": "id>10", // Filters
                    "splitPk": "id", // Split key
                    "username": "root"
                },
                "name": "sqlserverreader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "postSql": [ // Post-SQL
                        ""
                    ],
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": "jdbc:sqlserver://ip:1433;DatabaseName=database",
                            "table": [ // Target table
                                "schema.sink_table"
                            ]
                        }
                    ],
                    "writeMode": "upsert", // write mode
                    "batchSize": 1024, // Batch submission size
                    "username": "root",
                    "on": [ // Unique Index
                        "id"
                    ],
```

©2013–2025 Tencent Cloud. All rights reserved.

```
                    "preSql": [ // Pre-SQL
                        ""
                    ]
                },
                "name": "sqlserverwriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

PostgreSQL Data Source

Last updated: 2024-12-12 17:57:59

Supported Editions

The driver version used by PostgreSQL is postgresql42.2.26, currently supporting PostgreSQL 9.6 and above.

Use Limits

> **Note**
> Supports reading View Tables.

1. The PostgreSQL data source supports the Password authentication method (including SCRAM-SHA-256). If the password or password authentication method changes on the PostgreSQL database side, you need to update the data source configuration, retest connectivity, and manually run related tasks for verification.
2. When table names or field names in PostgreSQL start with numbers or include uppercase and lowercase English letters, or dashes (-), you need to use double quotes ("") for escaping. Failure to do so will result in the PostgreSQL plugin failing to read or write data. However, in PostgreSQL Reader and Writer plugins, double quotes ("") are JSON keywords, so you need to use a backslash (\) to escape double quotes ("") again. For example, if the table name is 123Test, it should be escaped as \"123Test\".

> **Note**
> - In double quotes ("") both the beginning and ending quotes (") need to be escaped with a backslash (\).
> - Wizard Mode does not support escaping; you need to switch to Script Mode for escaping.

Sample code for escaping in Script Mode:

```
"parameter": {
    "datasource": "abc",
    "column": [
        "id",
        "\"123Test\"", // Add escape characters
],
"where": "",
"splitPk": "id",
"table": "public.wpw_test"
}
```

PostgreSQL Offline Single Table Read Node Configuration

©2013-2025 Tencent Cloud. All rights reserved.

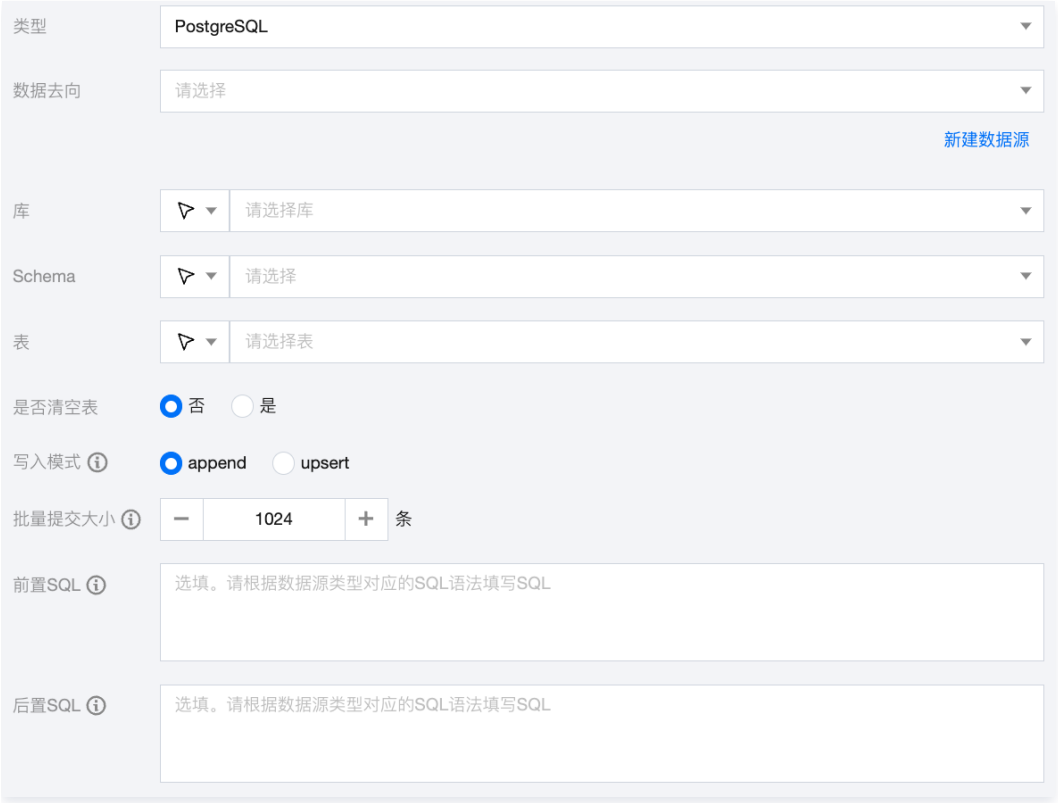


Parameters	Description
Data Source	Available PostgreSQL Data Source.
Database	Support selecting or manually entering the name of the database to be read • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Support selecting or manually entering the Schema Name to be read.
Table	Support selecting or manually entering the name of the table to be read • In the case of table partitioning, you can select or enter multiple table names at the PostgreSQL source, ensuring that all tables have consistent structures. • In the case of table partitioning, supports configuring the table serial number range. For example, 'table_[0-99]' indicates reading 'table_0', 'table_1', 'table_2' up to 'table_99'; if the length of your table's numeric suffix is consistent, such as 'table_000', 'table_001', 'table_002' up to 'table_999', you can configure it as '"table": ["table_00[0-9]", "table_0[10-99]", "table_[100-999]"]'. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Adding Sharding	Applicable to sharding scenarios, click to configure multiple data sources, libraries, and table information. **Note:** In sharding scenarios, ensure that the schema information (including field names and field types) is consistent among the selected table objects. The data field module defaults to displaying the metadata field information of the first table from the first data source. If there are inconsistencies among multiple tables, it may result in failures.
Split Key	You can use a column in the source data table as the split key. It is recommended to use the primary key or an indexed column as the split key. Only fields of type integer are supported. During data

©2013-2025 Tencent Cloud. All rights reserved.

	reading, data sharding is performed based on the configured field to achieve concurrent reading, which can improve data synchronization efficiency.
Filter Condition (Optional)	PostgreSQL concatenates the SQL based on the specified where condition and extracts data accordingly. During testing, you can specify actual business scenarios for the where condition, often choosing to sync data for the current day by setting the where condition to id>2 and sex=1.

PostgreSQL Offline Single Table Write Node Configuration



Parameters	Description
Data Flow Direction	PostgreSQL data source to write to.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Support selecting or manually entering the Schema Name to be read.
Table	Supports selection, or manual input of the table name to be written to • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Clear Table	You can manually choose whether to clear the PostgreSQL data table before writing to it.
Write mode	PostgreSQL write supports two modes: append: Append write. upsert: Update write based on the primary key field settings.
Batch Submission Size	The size of the record count for one-time batch submission. This value can greatly reduce the number of network interactions between the data synchronization system and PostgreSQL, and

©2013–2025 Tencent Cloud. All rights reserved.

	improve overall throughput. If this value is set too high, it may cause OOM exceptions in the data synchronization process.
Pre-SQL (optional)	SQL command executed before the synchronization task. Fill in the SQL according to the correct SQL syntax corresponding to the data source type. For example, clear old data in the table before execution (truncate table tablename).
Post-SQL (optional)	The SQL statement executed after the synchronization task should be written according to the correct SQL syntax for the data source type. For example, to add a timestamp with a default value of CURRENT_TIMESTAMP on update, you would use: alter table tablename add colname timestamp DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP.

Data type conversion is supported

Read

Supported data types and their conversion mappings for PostgreSQL read are as follows (when processing PostgreSQL, data types from the PostgreSQL data source are mapped to the data processing engine's data types):

PostgreSQL Data Type	Internal Type
smallint(int2)，integer(int4)，bigint(int8)， serial，bigserial	Long
real(float4)，double precision(float8)， money， numeric	Double
varchar(character varying)， char，uuid，json，array	String
date， time， timestamp	Date
bool(boolean)，bit	Boolean
bytea	Bytes

Write

The supported data types and their conversion mappings for PostgreSQL write are as follows:

Internal Type	PostgreSQL Data Type
Long	smallint(int2)，integer(int4)，bigint(int8)
Double	double precision，numeric， real
String	varchar(character varying)，char，bit，money
Date	date， time， timestamp
Boolean	bool(boolean)
Bytes	bytea

PostgreSQL script demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
```

©2013-2025 Tencent Cloud. All rights reserved.

```
                    ],
                    "connection": [
                        {
                            "jdbcUrl": [
                                "jdbc:postgresql://ip:5432/database?reWriteBatchedInserts=true"
                            ],
                            "table": [ // Source table
                                "schema.source_table"
                            ]
                        }
                    ],
                    "where": "id>10", // Filters
                    "splitPk": "id", // Split key
                    "username": "root"
                },
                "name": "postgresqlreader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "postSql": [ // Post-SQL
                        ""
                    ],
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": "jdbc:postgresql://ip:5432/database?reWriteBatchedInserts=true",
                            "table": [ // Target table
                                "schema.sink_table"
                            ]
                        }
                    ],
                    "writeMode": "upsert", // write mode
                    "batchSize": 1024, // Batch submission size
                    "username": "root",
                    "on": [ // unique index
                        "id"
                    ],
                    "preSql": [ // Pre-SQL
                        ""
                    ]
                },
                "name": "postgresqlwriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
```

©2013–2025 Tencent Cloud. All rights reserved.

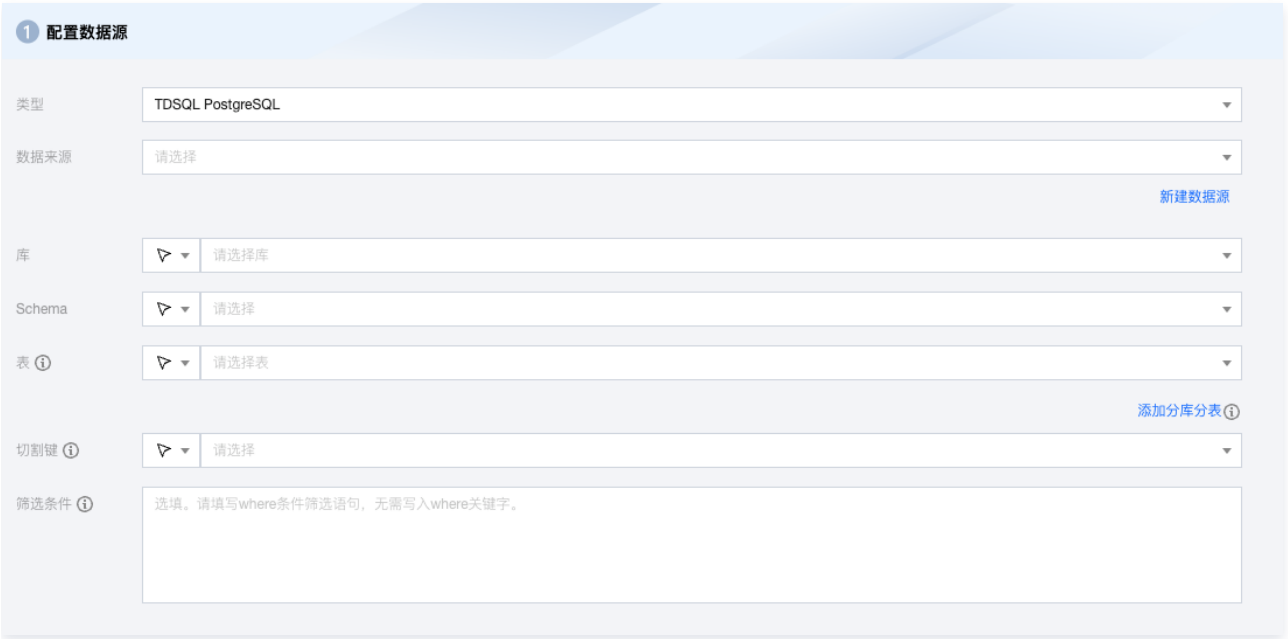
```
}
```

TDSQL for PostgreSQL Data Source

Last updated: 2024-12-12 17:58:33

©2013-2025 Tencent Cloud. All rights reserved.

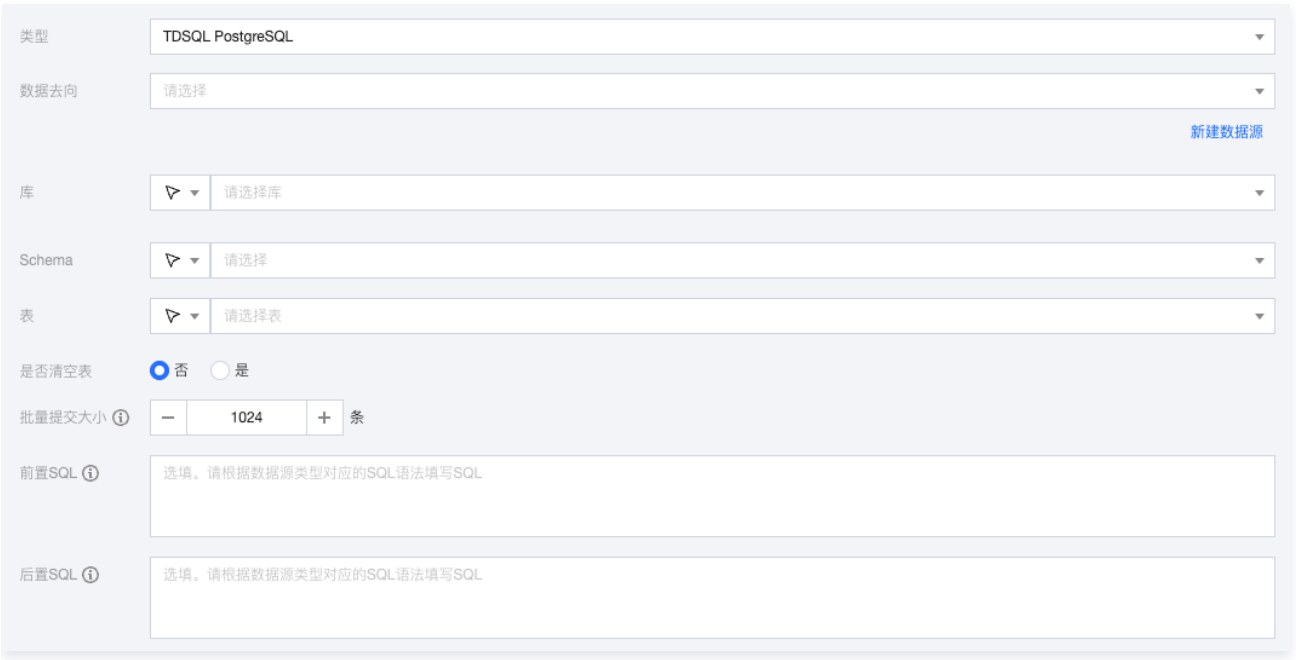
TDSQL for PostgreSQL Offline Single Table Read Node Configuration



Parameters	Description
Data Source	Select the configured TDSQL for PostgreSQL data source from the source end.
Database	Support selecting or manually entering the Library Name to be read. By default, the database bound to the data source is used as the default library, and other databases need to be manually entered. When the data source network is disconnected, resulting in the inability to directly pull library information, the database name can be manually entered. Data synchronization can still be performed when DataInLong network connectivity is available.
Schema	Support selecting or manually entering the Schema Name to be read.
Table	Support selecting or manually entering the Table Name to be read. When the data source network is disconnected, resulting in the inability to directly pull library information, the table name can be manually entered. Data synchronization can still be performed when DataInLong network connectivity is available.
Split Key	Specify the field for data sharding. Once specified, concurrent tasks will be initiated for data synchronization. You can use a column from the source data table as the split key. It is recommended to use the primary key or an indexed column as the split key.
Filter Condition (Optional)	Fill in the corresponding filter statement according to the data type. This statement will serve as the filter condition for the data to be synchronized.

TDSQL for PostgreSQL Offline Single Table Write Node Configuration

©2013–2025 Tencent Cloud. All rights reserved.



Parameters	Description
Data Flow Direction	Select the configured TDSQL for PostgreSQL data source from the target end.
Database	Support selecting or manually entering the Library Name to be read. By default, the database bound to the data source is used as the default library, and other databases need to be manually entered. When the data source network is disconnected, resulting in the inability to directly pull library information, the database name can be manually entered. Data synchronization can still be performed when DataInLong network connectivity is available.
Schema	Support selecting or manually entering the Schema Name to be read
Table	Support selecting or manually entering the Table Name to be read. When the data source network is disconnected, resulting in the inability to directly pull library information, the table name can be manually entered. Data synchronization can still be performed when DataInLong network connectivity is available.
Clear Table	You can manually choose whether to clear the TDSQL for PostgreSQL data table before writing to it.
Batch Submission Size	The size of the records submitted in bulk can significantly reduce the number of network interactions between the data synchronization system and TDSQL for PostgreSQL, and improve overall throughput. If this value is set too high, it could cause an OOM (Out Of Memory) error in the data synchronization process.
Pre-SQL	SQL statements executed before the synchronization task, filled in according to the correct SQL syntax corresponding to the data source type.
Post-SQL	SQL statements executed after the synchronization task, filled in according to the correct SQL syntax corresponding to the data source type.

Data type conversion is supported

Supported data types and conversion map for TDSQL for PostgreSQL read are as follows (when processing TDSQL for PostgreSQL, the data type of the TDSQL for PostgreSQL data source will first be mapped with the data type of the data processing engine):

Read

TDSQL-PostgreSQL Data Types	Internal Type
smallint，integer，bigint，serial，bigserial	Long

©2013-2025 Tencent Cloud. All rights reserved.

double precision，money，numeric，real	Double
varchar(character varying)，char(character)	String
date，time，timestamp	Date
bool(boolean)	Boolean
bytea	Bytes

Write

Internal Type	TDSQL-PostgreSQL Data Types
Long	smallint，integer，bigint
Double	double precision，numeric，real
String	varchar(character varying)，char(character)，bit，money
Date	date，time，timestamp
Boolean	bool
Bytes	bytea

TDSQL for PostgreSQL script demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": [
                                "jdbc:postgresql://ip:5432/database?reWriteBatchedInserts=true"
                            ],
                            "table": [ // Source table
                                "schema.source_table"
                            ]
                        }
                    ],
                    "where": "id>10", // Filters
                    "splitPk": "id", // Split key
                    "username": "root"
                },
                "name": "postgresqlreader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "postSql": [ // Post-SQL
                        ""
```

©2013-2025 Tencent Cloud. All rights reserved.

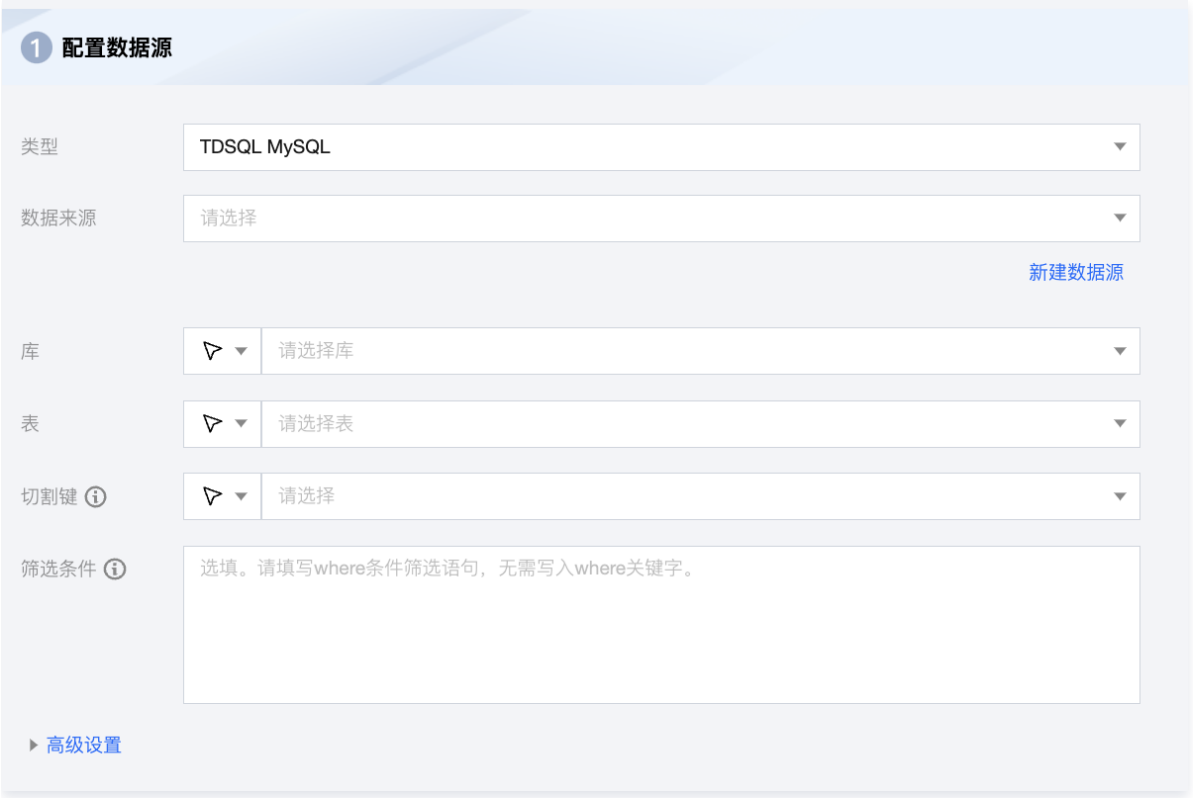
```
                    ],
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": "jdbc:postgresql://ip:5432/database?reWriteBatchedInserts=true",
                            "table": [ // Target table
                                "schema.sink_table"
                            ]
                        }
                    ],
                    "writeMode": "upsert", // write mode
                    "batchSize": 1024, // Batch submission size
                    "username": "root",
                    "on": [ // unique index
                        "id"
                    ],
                    "preSql": [ // Pre-SQL
                        ""
                    ]
                },
                "name": "postgresqlwriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

TDSQL for MySQL Data Source

Last updated：2024-12-12 17:59:07

TDSQL for MySQL Offline Single Table Read Node Configuration

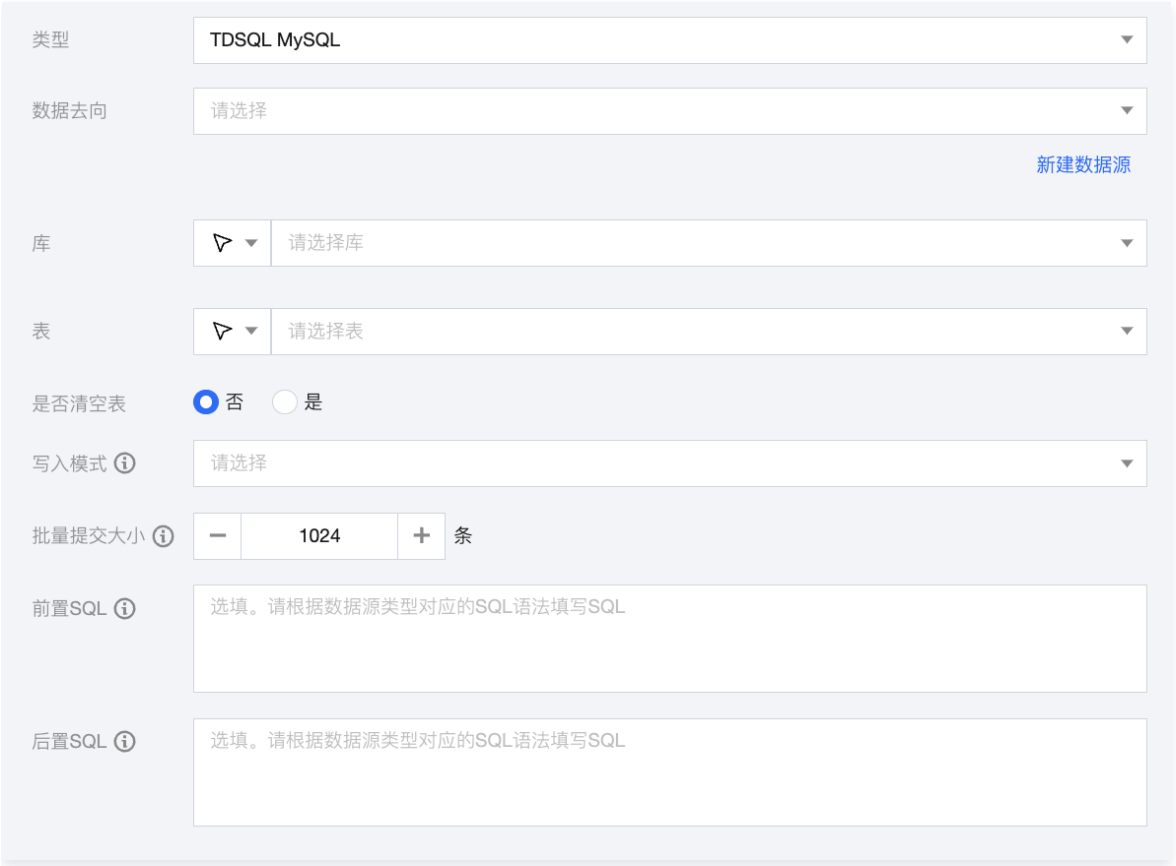


Parameters	Description
Data Source	Available TDSQL for MySQL Data Source.
Database	Support selecting or manually entering the name of the database to be read • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Support selecting or manually entering the name of the table to be read • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Split Key	Specify the field used for data sharding. After specifying, concurrent tasks will be started for data synchronization to improve efficiency. You may use a column in the source data table as the split key; it is recommended to use a primary key or an indexed column as the split key. Only integer-type fields are supported.
Filter Condition (Optional)	In actual business scenarios, it is common to choose the current day's data for synchronization, specifying the where condition as gmt_create>$bizdate. The where condition can effectively perform business incremental synchronization. • If the where clause is not specified, including not providing the key or value for where, data synchronization will be considered as full data synchronization. • You cannot specify the WHERE condition as LIMIT 10, which does not comply with TDSQL for MySQL WHERE clause constraints.

©2013-2025 Tencent Cloud. All rights reserved.

Advanced Settings (optional)	You can configure parameters based on business requirements.

TDSQL for MySQL Offline Single Table Write Node Configuration



Parameters	Description
Data Flow Direction	TDSQL for MySQL Data Source for writing.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Supports selection, or manual input of the table name to be written to • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Clear Table	Before writing to the TDSQL for MySQL Data Table, you can manually choose whether to clear the Data Table.
Write mode	TDSQL for MySQL supports three write modes: • Append: If there is a primary key/unique index conflict, the conflicting rows cannot be written. • Overwrite: If there is a primary key/unique index conflict, the original rows will be deleted first, and then the new rows will be inserted. • On duplicate key: If there is a primary key/unique index conflict, the new rows will replace the specified fields of the existing rows.

©2013–2025 Tencent Cloud. All rights reserved.

Batch Submission Size	The record count size for one-time batch submission can significantly reduce the number of network interactions between the data synchronization system and TDSQL and improve overall throughput. If this value is set too high, it can cause the data synchronization process to run into OOM exceptions.
Pre-SQL (optional)	SQL command executed before the synchronization task. Fill in the SQL according to the correct SQL syntax corresponding to the data source type. For example, clear old data in the table before execution (truncate table tablename).
Post-SQL (optional)	SQL statements executed after the synchronization task are written according to the correct SQL syntax corresponding to the data source type, for example, adding a timestamp with `ALTER TABLE tablename ADD colname TIMESTAMP DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP`.

TDSQL MySQL Script Demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": [
                                "jdbc:mysql://ip:3306/database?
rewriteBatchedStatements=true&tinyInt1isBit=false&serverTimezone=Asia/Shanghai&characterEncoding=utf8&auto
Reconnect=true"
                            ],
                            "table": [ // Source table
                                "source_table"
                            ]
                        }
                    ],
                    "where": "id>10", // Filters
                    "splitPk": "id", // Split key
                    "username": "root"
                },
                "name": "mysqlreader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "postSql": [ // Post-SQL
                        ""
                    ],
                    "password": "******",
                    "session": [
                        "set session sql_mode='ANSI'"
                    ],
                    "column": [ //column name
                        "id",
                        "name"
                    ],
```

©2013-2025 Tencent Cloud. All rights reserved.

```
                    "connection": [
                        {
                            "jdbcUrl": "jdbc:mysql://ip:3306/database?
rewriteBatchedStatements=true&tinyInt1isBit=false&serverTimezone=Asia/Shanghai&characterEncoding=utf8&auto
Reconnect=true",
                            "table": [ // Target table
                                "sink_table"
                            ]
                        }
                    ],
                    "writeMode": "replace", // write mode
                    "batchSize": 1024, // batch submission size
                    "username": "root",
                    "preSql": [ // Pre-SQL
                        ""
                    ]
                },
                "name": "mysqlwriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

DB2 Data Source

Last updated：2024-12-12 17:59:34

Use Limits

> **Note**
> Supports reading View Tables.

1. DB2 Writer obtains the protocol data generated by Reader through the Data Synchronization Framework and inserts the data into DB2 using the insert into statement (rows with primary key/unique index conflicts will not be written). For performance considerations, PreparedStatement + Batch is used, with rewriteBatchedStatements=true, buffering the data into the Thread Context Buffer. When the buffer reaches the set threshold, a write request is initiated.

> **Note**
> The entire task requires at least insert into permissions. Whether additional permissions are needed depends on the statements you specify in preSql and postSql when configuring the task.

2. Currently, the DB2 AES Encryption Algorithm is not supported.

DB2 Offline Single Table Read Node Configuration

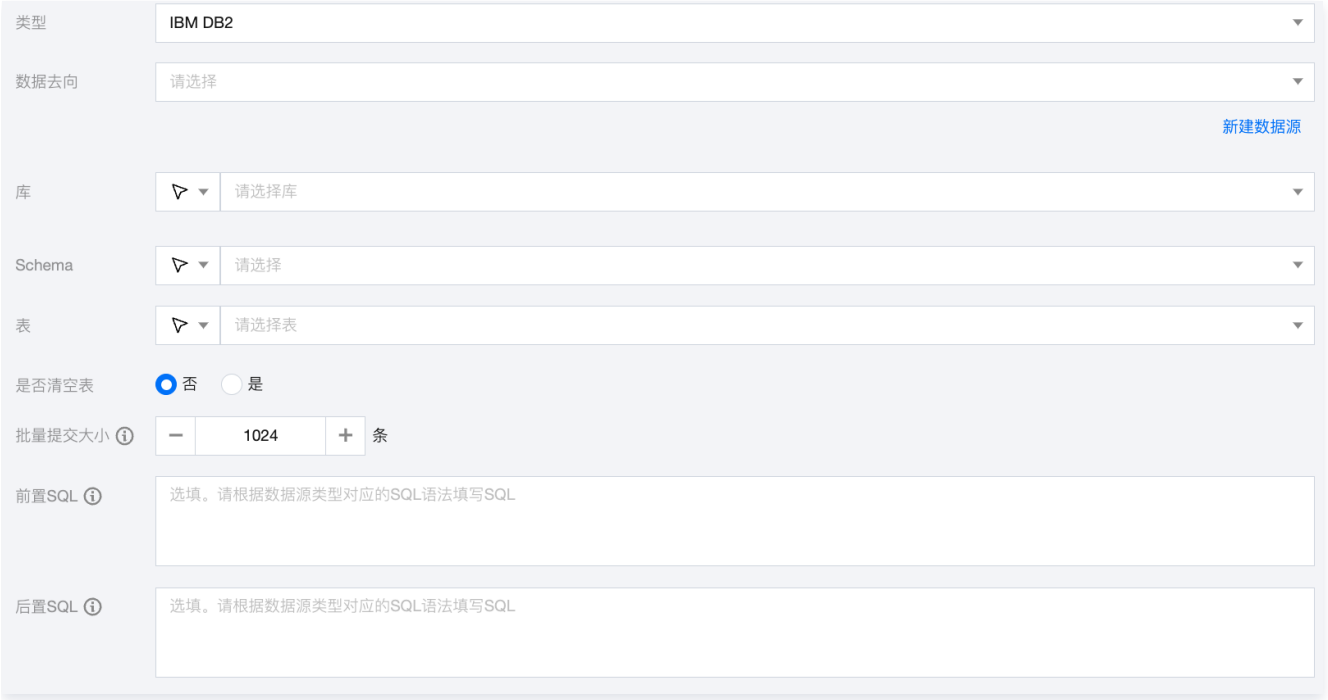


Parameters	Description
Data Source	Available DB2 Data Source.
Database	Support selecting or manually entering the name of the database to be read • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Support selecting or manually entering the Schema Name to be read.
Table	Support selecting or manually entering the Table Name to be read.

©2013-2025 Tencent Cloud. All rights reserved.

Split Key	Specify fields for Data Sharding. Once specified, concurrent tasks will be launched to perform Data Synchronization.
Filter Condition (Optional)	DB2 concatenates SQL based on the specified where condition and extracts data accordingly. In actual business scenarios, it is common to select data from the current day for synchronization, where the condition can be set as gmt_create>$bizdate. The where condition can effectively achieve business incremental synchronization. If this value is empty, it means synchronizing all information from the entire table.

DB2 Offline Single Table Write Node Configuration



Parameters	Description
Data Flow Direction	DB2 data source to be written to.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Support selecting or manually entering the Schema Name to be read.
Table	Supports selection, or manual input of the table name to be written to • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Clear Table	You can manually choose whether to clear the DB2 data table before writing to it.
Batch Submission Size	The size of the record count for one-time batch submission, which can greatly reduce the number of network interactions between the data synchronization system and DB2, and improve overall throughput. However, if this value is set too large, it may cause the data synchronization process to run into OOM exceptions.
Pre-SQL (optional)	SQL command executed before the synchronization task. Fill in the SQL according to the correct SQL syntax corresponding to the data source type. For example, clear old data in the table before execution (truncate table tablename).

©2013-2025 Tencent Cloud. All rights reserved.

Post-SQL (optional)	SQL statements executed after the synchronization task are written according to the correct SQL syntax corresponding to the data source type, for example, adding a timestamp with `ALTER TABLE tablename ADD colname TIMESTAMP DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP`.

Data type conversion is supported

Read

DB2 Data Type	Internal Type
BIGINT、INTEGER、SMALLINT	Long
NUMERIC、DECIMAL、FLOAT、REAL、DOUBLE	Double
CHAR、VARCHAR、LONGVARCHAR、NCHAR、CLOB	String
TIME、DATE、TIMESTAMP	Date
CHAR FOR BIT DATA、VARCHAR FOR BIT DATA	Bytes

Write

Internal Type	DB2 Data Type
Long	BIGINT、INTEGER、SMALLINT
Double	DECIMAL、NUMERIC、FLOAT、REAL、DOUBLE
String	CHAR、VARCHAR、NCHAR、CLOB、LONGVARCHAR
Date	DATE、TIME、TIMESTAMP
Bytes	CHAR FOR BIT DATA、VARCHAR FOR BIT DATA

DB2 Script Demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": [
                                "jdbc:db2://ip:50000/database"
                            ],
                            "table": [ // Source table
                                "schema.source_table"
                            ]
                        }
                    ],
                    "where": "id>10", // Filters
                    "username": "root"
                },
```

©2013-2025 Tencent Cloud. All rights reserved.

```
                "name": "rdbmsreader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "postSql": [ // Post-SQL
                        ""
                    ],
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": "jdbc:db2://ip:50000/database",
                            "table": [ // Target table
                                "schema.sink_table"
                            ]
                        }
                    ],
                    "batchSize": 1024, // Batch submission size
                    "username": "root",
                    "preSql": [ // Pre-SQL
                        ""
                    ]
                },
                "name": "rdbmswriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

Dameng DM Data Source

Last updated：2024-12-12 18:00:15

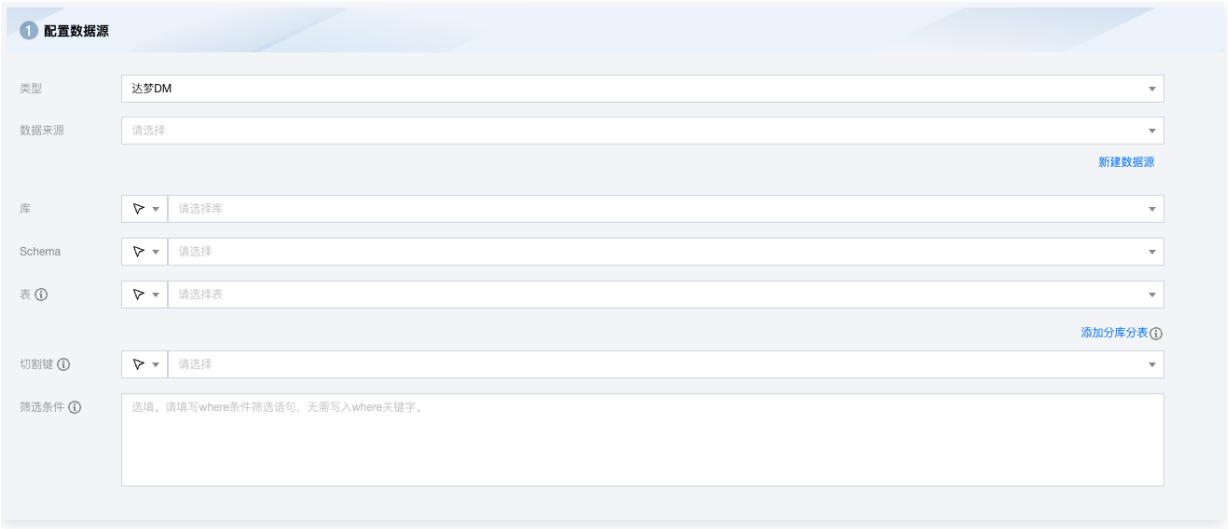
Supported Editions

Currently, DM versions 7 and 8 are supported.

Use Limits

> **Note**
> Supports reading View Tables.

DM Offline Single Table Read Node Configuration

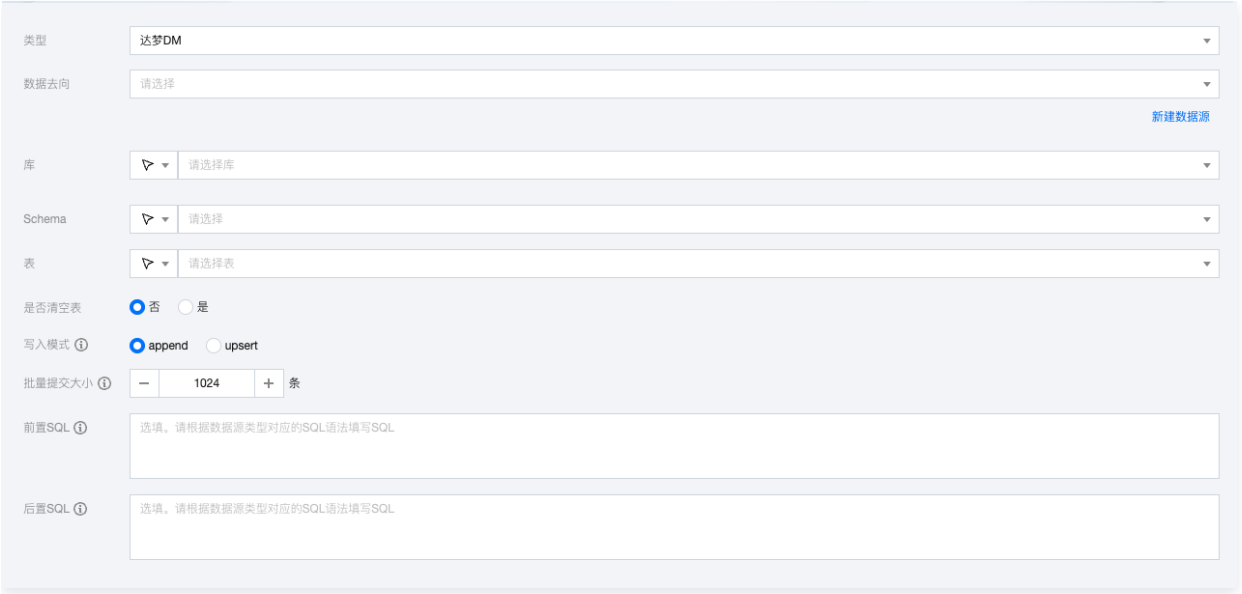


Parameters	Description
Data Source	Available DM Data Source.
Database	Support selecting or manually entering the name of the database to be read • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Support selecting or manually entering the Schema Name to be read.
Adding Sharding	You can create multiple data sources and add corresponding table objects. **Note:** When selecting multiple table objects in a sharded database, ensure that the Schema information is consistent (including field names and field types). The system's Data Field module will default to display metadata field information for the first table of the first data source. Inconsistencies in fields between multiple tables may cause execution to fail.
Table	Support selecting or manually entering the Table Name to be read.
Split Key	You can use a column from the source data table as the cutting key. It is recommended to use a primary key or an indexed column as the cutting key. Only fields of type Integer are supported.

©2013-2025 Tencent Cloud. All rights reserved.

	When reading data, the data is sharded based on the configured fields to achieve concurrent reading, which can improve data synchronization efficiency.
Filter Condition (Optional)	Fill in the corresponding filter statement according to the data type. This statement will be used as the filter condition for the data to be synchronized. DM constructs SQL based on the specified WHERE condition and extracts data according to that SQL. For example, during testing, you can specify the WHERE condition as LIMIT 10. In actual business scenarios, you usually specify the data for the current day for synchronization, with the WHERE condition specified as gmt_create > $bizdate. • The WHERE condition can effectively perform business incremental synchronization. • If the WHERE condition is empty, it is considered to synchronize all information from the entire table.

DM Offline Single Table Write Node Configuration



Parameters	Description
Data Flow Direction	DM data source to be written.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Support selecting or manually entering the Schema Name to be read.
Table	Supports selection, or manual input of the table name to be written to • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Clear Table	You can manually choose whether to clear the DM data table before writing to it.
Write mode	• append: Append write. • upsert: Update write according to the set primary key field.
Batch Submission Size	The size of the batch submission record. This value can significantly reduce the number of network interactions between the data synchronization system and DM, and improve overall throughput. If this value is set too large, it may cause an OOM exception in the data synchronization process.

©2013–2025 Tencent Cloud. All rights reserved.

Pre-SQL (optional)	SQL command executed before the synchronization task. Fill in the SQL according to the correct SQL syntax corresponding to the data source type. For example, clear old data in the table before execution (truncate table tablename).
Post-SQL (optional)	SQL statements executed after the synchronization task are written according to the correct SQL syntax corresponding to the data source type, for example, adding a timestamp with `ALTER TABLE tablename ADD colname TIMESTAMP DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP`.

Data type conversion is supported

Read

DM Data Type	Internal Type
BIGINT、INTEGER、SMALLINT、TINYINT	Long
NUMERIC、DECIMAL、DOUBLE、FLOAT、REAL	Double
CHAR、VARCHAR、LONGVARCHAR、CLOB	String
TIME、DATE、TIMESTAMP	Date
BIT、BOOLEAN	Boolean
BINARY、VARBINARY、BLOB、LONGVARBINARY	Bytes

Write

Internal Type	DM Data Type
Long	BIGINT、INTEGER、SMALLINT、TINYINT
Double	DECIMAL、NUMERIC、DOUBLE、FLOAT、REAL
String	CHAR、VARCHAR、CLOB、LONGVARCHAR
Date	DATE、TIME、TIMESTAMP
Boolean	BOOLEAN、BIT
Bytes	BINARY、VARBINARY、BLOB、LONGVARBINARY

Dameng DM Script Demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": [
                                "jdbc:dm://ip:5236/database"
                            ],
```

©2013-2025 Tencent Cloud. All rights reserved.

```
                            "table": [ // Source table
                                "schema.source_table"
                            ]
                        }
                    ],
                    "where": "id>10", // Filters
                    "username": "root"
                },
                "name": "rdbmsreader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "postSql": [ // Post-SQL
                        ""
                    ],
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": "jdbc:dm://ip:5236/database",
                            "table": [ // Target table
                                "schema.sink_table"
                            ]
                        }
                    ],
                    "writeMode": "upsert", // write mode
                    "batchSize": 1024, // Batch submission size
                    "username": "root",
                    "on": [ // unique index
                        "id"
                    ],
                    "preSql": [ // Pre-SQL
                        ""
                    ]
                },
                "name": "rdbmswriter"
            }
        }
    ],
    "setting": {
        "errorLimit": {
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

OceanBase Data Source

Last updated: 2024-12-12 18:00:43

> **Note**
> Note that OceanBase Data Source requires whitelisting before use.

OceanBase Offline Single Table Read Node Configuration

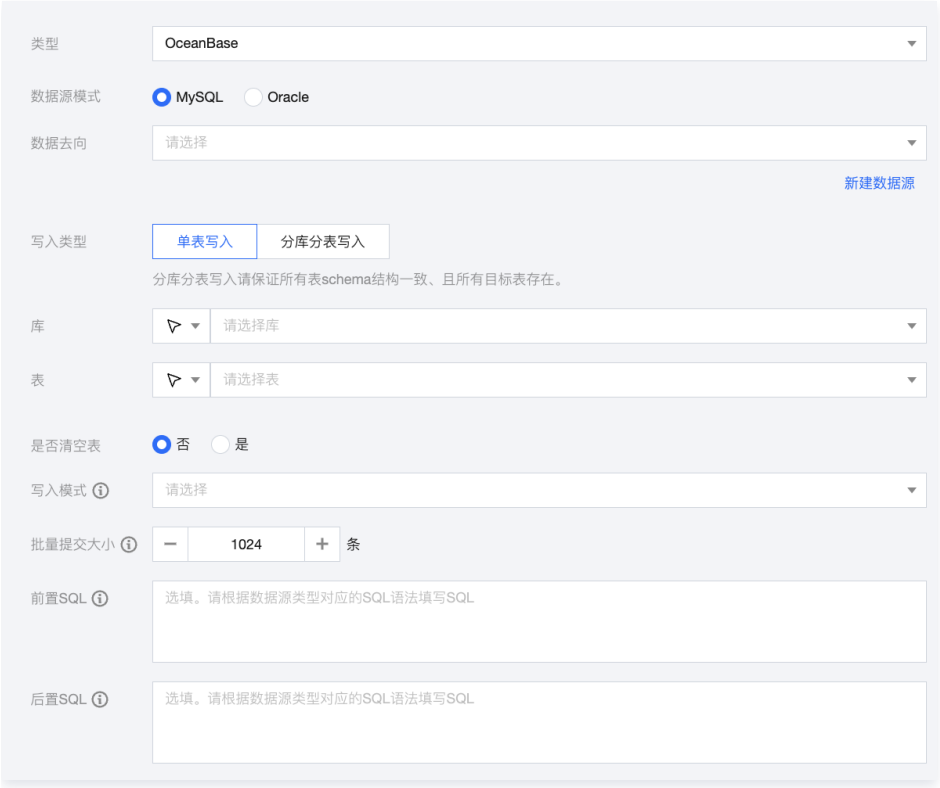


Parameters	Description
Data Source Mode	Supports MySQL Compatibility Mode and Oracle Compatibility Mode
Data Source	Available OceanBase Data Sources.
Database	Support selecting or manually entering the name of the database to be read • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Support selecting or manually entering the Table Name to be read.
Adding Sharding	Suitable for sharding scenarios, click to configure multiple data sources, databases, and table information. In sharding and partitioning scenarios, ensure all table structures are consistent. The task configuration will default to display and use the first table structure for data acquisition.

©2013-2025 Tencent Cloud. All rights reserved.

Split Key	• You can use a column from the source data table as the cutting key; it is recommended to use a primary key or an indexed column as the cutting key. Only fields of type Integer are supported. • When reading data, data sharding is performed based on the configured fields to achieve concurrent reading, which can improve data synchronization efficiency.
Filter Condition (Optional)	Fill in the corresponding filter statement according to the data type. This statement will be used as the filter condition for the data to be synchronized. Currently, the limit keyword filtering is not supported. The SQL syntax is consistent with the selected data source.

OceanBase Offline Single Table Write Node Configuration



Parameters	Description
Data Source Mode	Supports MySQL Compatibility Mode and Oracle Compatibility Mode
Data Flow Direction	Available OceanBase Data Sources.
Write Type (MySQL Compatibility Mode)	• Single Table Write: Supports writing to a single table • Sharding write: Supports writing to sharded MySQL tables within the same data source at once. Ensure all table schemas are consistent and all target tables exist for sharding write.
Database	Support selecting or manually entering the name of the database to be read • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Support selecting or manually entering the Table Name to be read.
Clear Table	Before writing to the MySQL data table, you can manually choose whether to clear the data table.
Write mode	• There are three writing modes supported under MySQL Compatibility Mode:

©2013–2025 Tencent Cloud. All rights reserved.

	○ Append: If there is a primary key/unique index conflict, the conflicting rows cannot be written. ○ Overwrite: In case of a primary key/unique index conflict, the original row will be deleted, and a new row will be inserted. ○ On duplicate key: In case of a primary key/unique index conflict, the new row will replace the specified fields. • There are two writing modes supported under Oracle Compatibility Mode: ○ Append: Append data. ○ Upsert: Update and write data according to the set primary key fields.
Batch Submission Size	The size of the record count for one-time batch submissions. This value can greatly reduce the number of network interactions between the data synchronization system and OceanBase, and improve overall throughput. If this value is set too high, it may cause the data synchronization process to run into an OOM (Out of Memory) exception.
Pre-SQL (optional)	SQL command executed before the synchronization task. Fill in the SQL according to the correct SQL syntax corresponding to the data source type. For example, clear old data in the table before execution (truncate table tablename).
Post-SQL (optional)	SQL statements executed after the synchronization task are written according to the correct SQL syntax corresponding to the data source type, for example, adding a timestamp with `ALTER TABLE tablename ADD colname TIMESTAMP DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP`.

OceanBase script demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "user_id"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": [
                                "jdbc:oceanbase://ip:9527/source_database"
                            ],
                            "table": [ // Source table
                                "source_database.source_table"
                            ]
                        }
                    ],
                    "where": "id>10", // Filters
                    "splitPk": "id", // Split key
                    "username": "root"
                },
                "name": "oceanbasev10reader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "postSql": [ // Post-SQL
                        ""
                    ],
                    "password": "******",
                    "session": [
```

©2013-2025 Tencent Cloud. All rights reserved.

```
                        "set session sql_mode='ANSI'"
                    ],
                    "column": [ //column name
                        "id",
                        "user_id"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": "jdbc:oceanbase://ip:9527/sink_database",
                            "table": [ // Target table
                                "sink_database.\"sink_table\""
                            ]
                        }
                    ],
                    "writeMode": "upsert", // write mode
                    "batchSize": 1024, // Batch submission size
                    "username": "root",
                    "on": [ // unique index
                        "id"
                    ],
                    "preSql": [ // Pre-SQL
                        ""
                    ]
                },
                "name": "oceanbasev10writer"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

SAP HANA Data Source

Last updated：2024-12-12 19:00:05

Use Limits

> **Note**
> Supports reading View Tables.

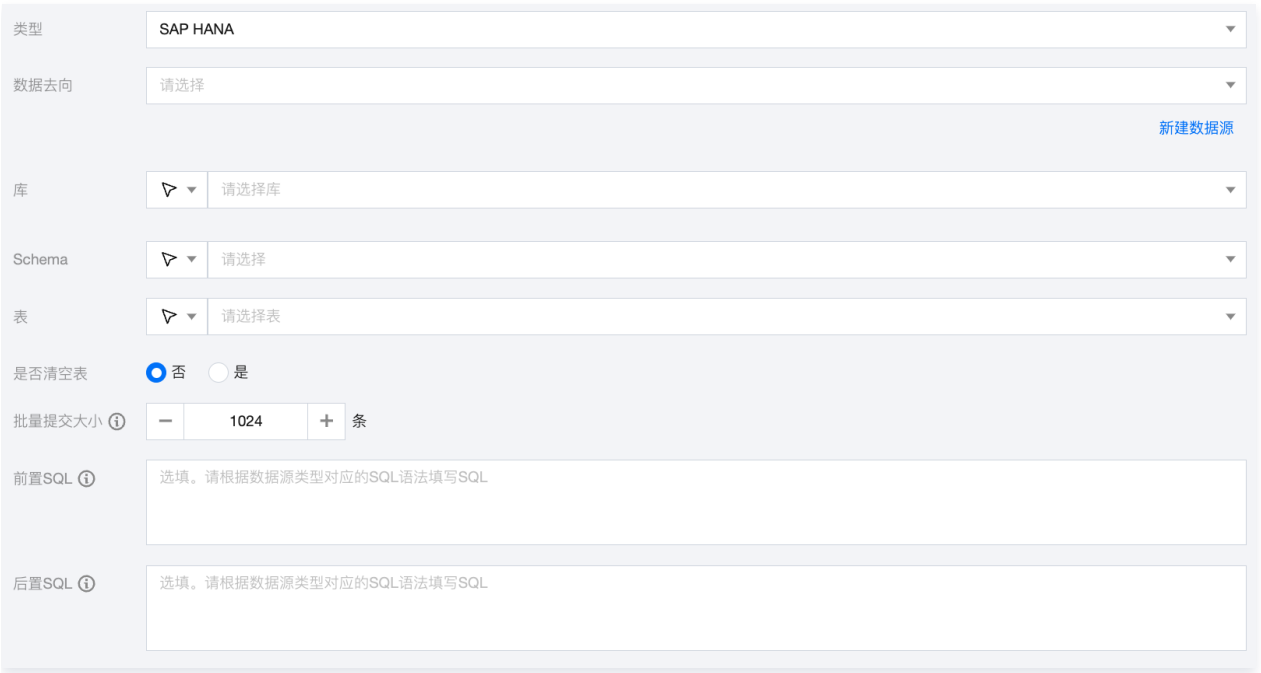
SAP HANA Offline Single Table Read Node Configuration



Parameter	Description
Data Source	Available SAP HANA Data Sources.
Database	Support selecting or manually entering the name of the database to be read • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Support selecting or manually entering the Schema Name to be read.
Table	Support selecting or manually entering the Table Name to be read.
Split Key	• You can use a column from the source data table as the cutting key; it is recommended to use a primary key or an indexed column as the cutting key. Only fields of type Integer are supported. • When reading data, data sharding is performed based on the configured fields to achieve concurrent reading, which can improve data synchronization efficiency.
Filter Condition (Optional)	Fill in the corresponding filter statement according to the data type. This statement will be used as the filter condition for the data to be synchronized. Currently, the limit keyword filtering is not supported. The SQL syntax is consistent with the selected data source.

SAP HANA Offline Single Table Write Node Configuration

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Flow Direction	SAP HANA data source to be written to.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Support selecting or manually entering the Schema Name to be read.
Table	Supports selection, or manual input of the table name to be written to • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Clear Table	You can manually choose whether to clear the SAP HANA data table before writing to it.
Batch Submission Size	Batch submission record count for one-time submission. This value can significantly reduce the number of network interactions between the data synchronization system and SAP HANA, and improve overall throughput. If this value is set too high, it can lead to data synchronization process OOM exceptions.
Pre-SQL (optional)	SQL command executed before the synchronization task. Fill in the SQL according to the correct SQL syntax corresponding to the data source type. For example, clear old data in the table before execution (truncate table tablename).
Post-SQL (optional)	SQL statements executed after the synchronization task are written according to the correct SQL syntax corresponding to the data source type, for example, adding a timestamp with `ALTER TABLE tablename ADD colname TIMESTAMP DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP`.

Data type conversion is supported

Read

SAP HANA Data Type	Internal Type

©2013-2025 Tencent Cloud. All rights reserved.

BIGINT,INTEGER,SMALLINT,TINYINT	Long
DECIMAL,FLOAT,REAL,DOUBLE	Double
CHAR,VARCHAR,NVARCHAR,CLOB,NCLOB	String
TIME,DATE,TIMESTAMP	Date
BLOB,VARBINARY	Bytes
BIT,BOOLEAN	Boolean

Write

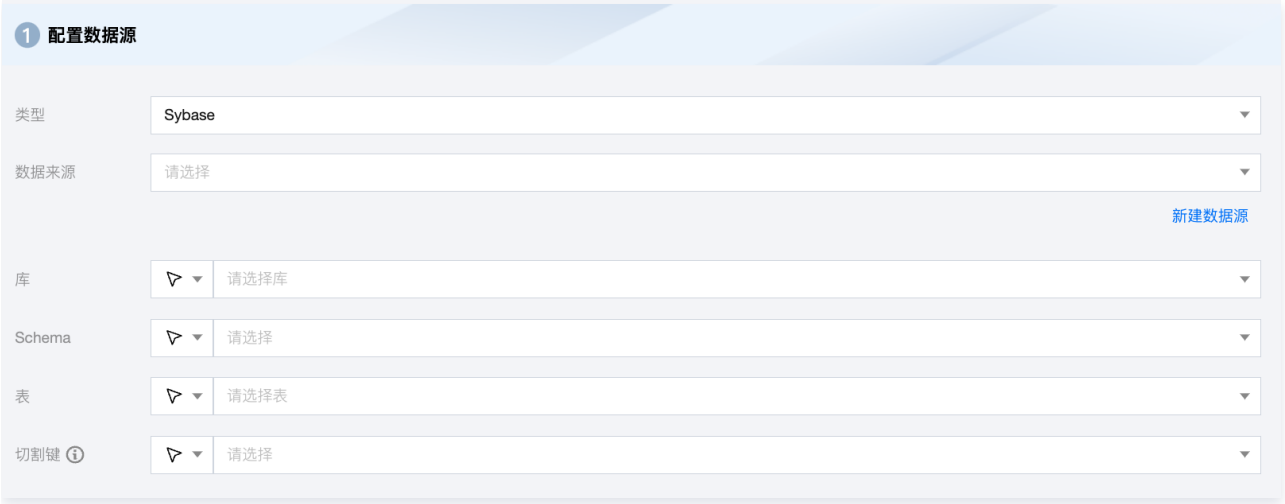
Internal Type	SAP HANA Data Type
Long	BIGINT,INTEGER,SMALLINT,TINYINT
Double	DECIMAL,FLOAT,REAL,DOUBLE
String	CHAR,VARCHAR,CLOB,NVARCHAR,NCLOB
Date	DATE,TIME,TIMESTAMP
Bytes	BLOB,VARBINARY
Boolean	BIT,BOOLEAN

©2013–2025 Tencent Cloud. All rights reserved.

SAP IQ (Sybase) Data Source

Last updated：2024-12-12 19:00:32

SAP IQ Offline Single Table Read Node Configuration



Parameters	Description
Data Source	Available SAP IQ Data Sources
Database	Supports selection or manual input of the library name to be read. By default, the database bound to the data source is used as the default library. Other databases require manual input of the library name. If the data source network is not connected and the library information cannot be directly retrieved, you can manually enter the database name. Data synchronization is still possible when the DataInLong network is connected.
Schema	Support selecting or manually entering the data pattern to be read
Table	Support selecting or manually entering the name of the table to be read
Split Key	• You can use a column from the source data table as the split key. It is recommended to use the primary key or an indexed column as the split key. Only fields of type Integer are supported • During data reading, data sharding is based on the configured fields to enable concurrent reading, which can improve data synchronization efficiency

Data type conversion is supported

Read

SAP IQ (Sybase) Data Types	Internal Type
INTEGER、SMALLINT	Long
DECIMAL、FLOAT、REAL	Double
CHAR、VARCHAR、TEXT、UNICHAR、UNIVARCHAR	String
DATETIME	Date
IMAGE	Bytes

Write

©2013-2025 Tencent Cloud. All rights reserved.

Internal Type	SAP IQ (Sybase) Data Types
Long	INTEGER、SMALLINT
Double	DECIMAL、FLOAT、REAL
String	CHAR、VARCHAR、TEXT、UNICHAR、UNIVARCHAR
Date	DATE、TIME、TIMESTAMP
Bytes	IMAGE

SAP IQ (Sybase) Script Demo

If you configure an offline task using **Script Mode**, you need to write the **reader parameters** in the task script according to the unified script format requirements (currently Sybase only supports reader).

```
"job": {
    "content": [
        {
            "reader": { // only supports reader
                "parameter": {
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": [
                                "jdbc:sybase:Tds:ip:5000/database?charset=utf8"
                            ],
                            "table": [ // Source table
                                "schema.source_table"
                            ]
                        }
                    ],
                    "where": "id>10", // Filters
                    "splitPk": "id", // Split key
                    "username": "root"
                },
                "name": "rdbmsreader"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

Doris/TCHouse-D Data Source

Last updated: 2024-12-12 19:01:11

The TCHouse-D data source configuration is the same as Doris. The following explanation uses the Doris data source as an example:

Supported Editions

Support Doris 0.x, 1.1.x, 1.2.x, 2.x versions.

Use Limits

1. Doris writes data through the Stream load HTTP interface, and it is necessary to ensure that the IP or port of the FE or BE in the data source is correctly filled in.
2. Since the principle of Stream load is that the import is initiated and distributed by BE, the recommended import data volume is between 1G and 10G. The default maximum Stream load import data volume is 10G.
   - To import files exceeding 10G, you need to modify the BE configuration `streaming_load_max_mb.`
   - For example: If the file to be imported is 15G, modify the BE configuration `streaming_load_max_mb` to 16000.
3. The default timeout for Stream Load is 600 seconds. Based on Doris' current maximum import speed, you need to modify the default timeout for import tasks for files larger than about 3G.
   - Import task timeout = Data Import Volume / 10M/s (the specific average import speed needs to be calculated by the user based on their cluster).
   - For example: If you are importing a 10G file, timeout = 1000s which equals 10G / 10M/s

Doris Offline Single Table Read Node Configuration

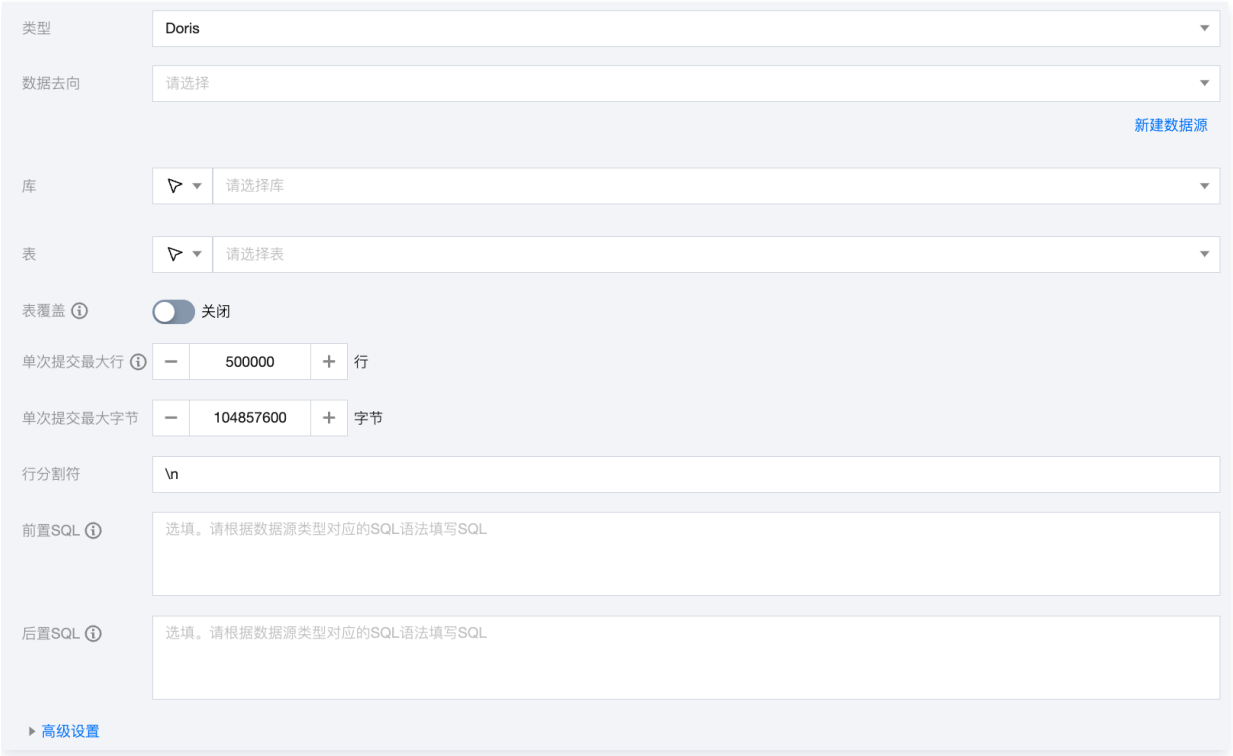


Parameters	Description
Data Source	Available Doris data sources to sync.
Database	Support selection or manual input of the library name to be read. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Support selecting or manually entering the Table Name to be read.

©2013-2025 Tencent Cloud. All rights reserved.

Split Key	Specify the field for data sharding. Once specified, concurrent tasks will be initiated for data synchronization. You can use a column from the source data table as the split key. It is recommended to use the primary key or an indexed column as the split key.
Filter Condition (Optional)	In actual business scenarios, the current day's data is usually selected for synchronization, with the where condition specified as gmt_create>$bizdate. The where condition can effectively perform incremental business synchronization. If the where statement is not filled, including not providing the key or value for where, data synchronization is considered as synchronizing all data.
Advanced Settings (optional)	You can configure parameters based on business requirements.

Doris Offline Single Table Write Node Configuration



Parameters	Description
Data Flow Direction	The Doris data source to be written.
Database	Supports selecting or manually entering the name of the database to be written to. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Supports selecting or manually entering the name of the table to be written to. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Table overwrite	Once enabled, Doris will support atomic overwrite operations at the table level. Before writing data, a new table with the same structure will be created using the CREATE TABLE LIKE statement. After importing new data into the new table, the old table will be atomically replaced via swap, achieving table overwrite.

©2013–2025 Tencent Cloud. All rights reserved.

Maximum rows per submission	The size of the record batch submitted at one time.
Maximum bytes per submission	The maximum data volume submitted in a single batch.
Row Separator (optional)	Doris key separator for writing, defaults to '\n'. Please ensure it matches the column separator of the created Doris table, otherwise the data will not be found in the Doris table.
Pre-SQL	SQL command executed before the synchronization task. Fill in the SQL according to the correct SQL syntax corresponding to the data source type. For example, clear old data in the table before execution (truncate table tablename).
Post-SQL	SQL statements executed after the synchronization task are written according to the correct SQL syntax corresponding to the data source type, for example, adding a timestamp with `ALTER TABLE tablename ADD colname TIMESTAMP DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP`.
Advanced Parameters	You can configure parameters based on business requirements.

Data type conversion is supported

Read

Doris Data Type	Internal Type
TINYINT、SMALLINT、INT、BIGINT	Long
FLOAT、DOUBULE、DECIMAL	Double
VARCHAR、CHAR、ARRAY、STRUCT、STRING	String
DATE、DATETIME	Date
BOOLEAN	Boolean

Write

Internal Type	Doris Data Type
Long	TINYINT、SMALLINT、INT、BIGINT
Double	DOUBLE、FLOAT、DECIMAL
String	STRING、VARCHAR、CHAR、ARRAY、STRUCT
Date	DATETIME、DATE
Boolean	BOOLEAN

FAQs

1. Error when partition does not exist: "no partition for this tuple"

Cause: Doris lacks the corresponding partition.
Solution: If using time partitions, it is recommended to enable dynamic partitioning.

```
For table tbl1, partition column k1 is of DATE type. Create a dynamic partition rule. Partition by days,
only retain the partitions of the last 7 days, and pre-create the partitions of the next 3 days.


CREATE TABLE tbl1
(
```

©2013-2025 Tencent Cloud. All rights reserved.

```
    k1 DATE,
    ...
)
PARTITION BY RANGE(k1) ()
DISTRIBUTED BY HASH(k1)
PROPERTIES
(
    "dynamic_partition.enable" = "true",
    "dynamic_partition.time_unit" = "DAY",
    "dynamic_partition.start" = "-7",
    "dynamic_partition.end" = "3",
    "dynamic_partition.prefix" = "p",
    "dynamic_partition.buckets" = "32"
);


Assume the current date is 2020-05-29. Then according to the above rules, tbl1 will generate the following
partitions:


p20200529: ["2020-05-29", "2020-05-30")
p20200530: ["2020-05-30", "2020-05-31")
p20200531: ["2020-05-31", "2020-06-01")
p20200601: ["2020-06-01", "2020-06-02")


The next day, on 2020-05-30, a new partition p20200602: ["2020-06-02", "2020-06-03") will be created


On 2020-06-06, since dynamic_partition.start is set to 7, the partition from 7 days ago, i.e., partition
p20200529, will be deleted.
```

Reference: For more details, please see Temporary Partition .

2. Synchronization data volume is too large causing:"The size of this batch exceed the max size of json type data"

Cause: The batch submission size is too large. It's recommended to ignore the JSON data volume check.

©2013–2025 Tencent Cloud. All rights reserved.



3. Import frequency is too fast causing:"tablet writer write failed, err=-235"

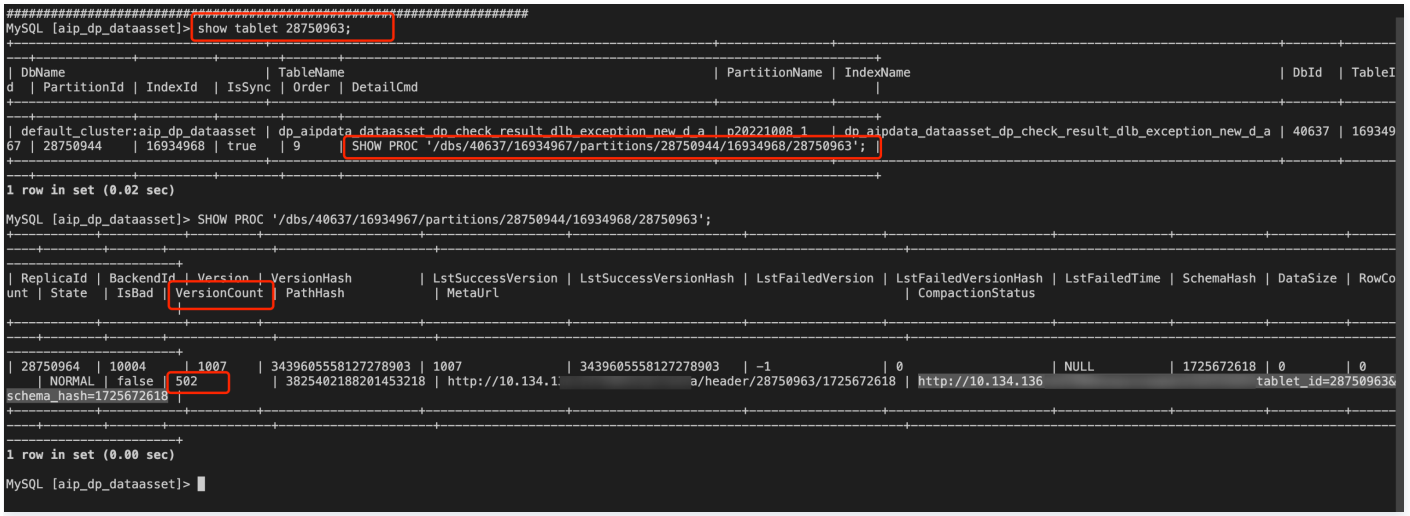
Cause: The fast import frequency causes the VersionCount to exceed the set size (default is 500, controlled by the BE parameter max_tablet_version_num). It is recommended to adjust the VersionCount.
Solutions:
1. Locate the error reported tablet_id.

```
show tablet 28750963;
```

2. Execute the DetailCmd show command.

```
SHOW PROC '/dbs/40637/16934967/partitions/28750944/16934968/28750963';
```

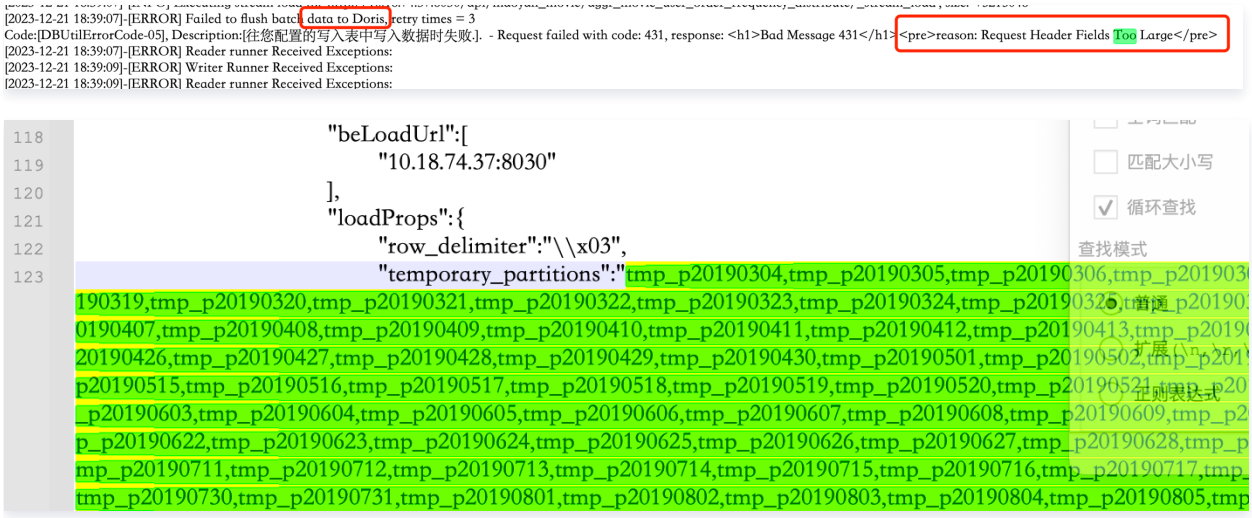
©2013-2025 Tencent Cloud. All rights reserved.



3. The versionCount in the result represents the number of versions. If a copy is found to have too many versions, the import frequency should be reduced or stopped. If the version count does not decrease after stopping imports, inspect the corresponding BE node's be.INFO logs, search for tablet id and compaction keywords, and check if the compaction is running normally.
4. Increase the max_tablet_version_num or optimize the tablet's compaction to improve compression efficiency and reduce the number of versions.

4. Header Temporary partition length exceeds limit: "Bad Message 431"

Cause: temporary_partition is too large, reduce the number of temporary partitions.



Reference: For more information, please see dynamic partition.

5. When DLC synchronizes with Doris, Doris reports an error: Field length exceeds schema length

Issue Information:

UUID length set to 200, in the synchronization record, UUID length is 232.

```
Reason: column_name[uuid], the length of input is too long than schema. first 32 bytes of input str:
[0000000000000BB4E595527BE******] schema length: 200; actual length: 232; . src line [];
```

Cause:

Spark engine of DLC calculates character length with each Chinese character occupying one length, while Doris calculates length treating Chinese as UTF-8 compilation, occupying 3 character lengths. The inconsistency in length calculation between the two ends may trigger an alert for strings near the critical length.

Solutions:

- Modify dirty data at the DLC source end

©2013-2025 Tencent Cloud. All rights reserved.

- Increase the length limit for UUID in the Doris table

Doris/TCHouse-D Script Demo

If you configure an offline task using **Script Mode**, you need to write the **reader parameter** and **writer parameter** in the task script according to the unified format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": [
                                "jdbc:mysql://ip:8030/source_database"
                            ],
                            "table": [ // Source table
                                "source_table"
                            ]
                        }
                    ],
                    "where": "id>10", // Filters
                    "splitPk": "id", // Split key
                    "username": "root"
                },
                "name": "dorisreader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "postSql": [ // Post-SQL
                        ""
                    ],
                    "swapTableName": "sink_table_timestamp", // When table overriding is enabled, a table
with the same structure is first created using 'create table like', and then replaced with the new table
using 'swap' method
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "swapTable": "true", // Whether to enable table overriding
                    "timeZone": "+08:00",
                    "database": "sink_database", // Target database
                    "password": "******",
                    "beLoadUrl": [
                        "ip:9030",
                        "ip:9030",
                        "ip:9030"
                    ],
                    "labelPrefix": "",
                    "maxBatchByteSize": 104857600, // Maximum bytes per submission
                    "lineDelimiter": "\n", //Row Separator
                    "jdbcUrl": "jdbc:mysql://ip:9030/sink_database",
                    "connectTimeout": -1,
                    "table": "sink_table", // Target table
```

©2013-2025 Tencent Cloud. All rights reserved.

```
                    "maxBatchRows": 500000, // Maximum rows per submission
                    "username": "root",
                    "preSql": [ // Pre-SQL
                        ""
                    ]
                },
                "name": "doriswriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

StarRocks Data Source

Last updated：2024-12-12 19:01:41

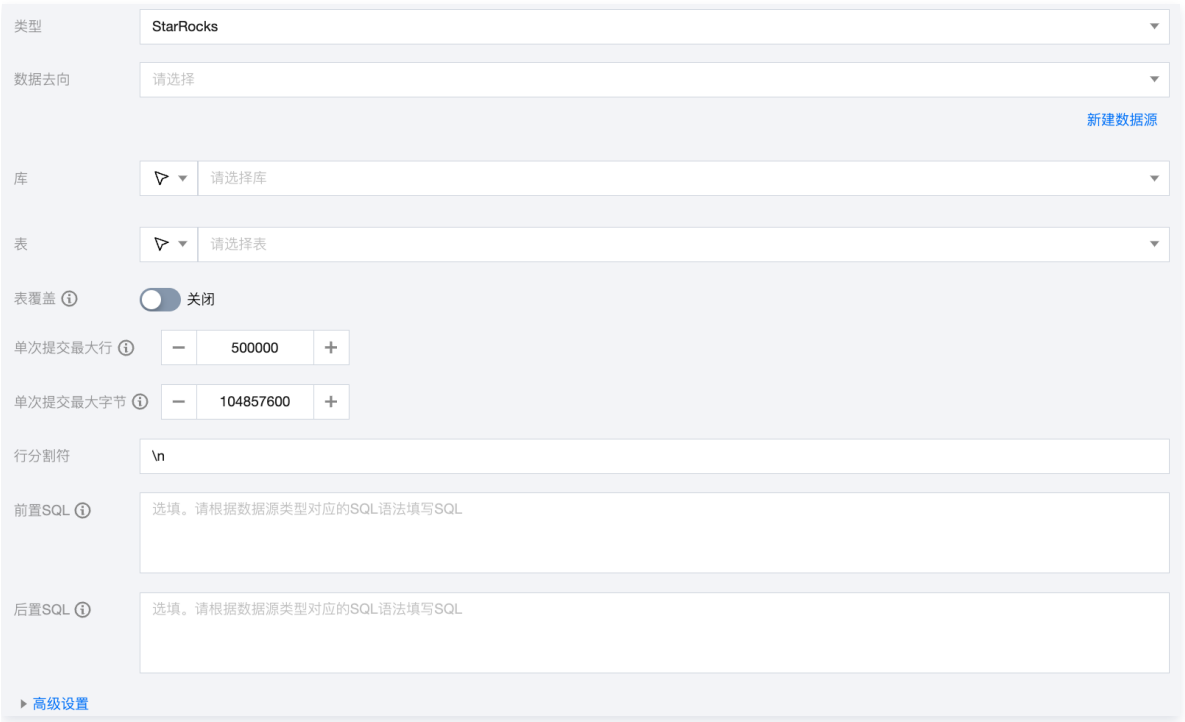
StarRocks Offline Single Table Read Node Configuration



Parameters	Description
Data Source	Select the configured StarRocks Data Source at the source end
Database	Support selecting or manually entering the Library Name to be read. By default, the database bound to the data source is used as the default library, and other databases need to be manually entered. When the data source network is disconnected, resulting in the inability to directly pull library information, the database name can be manually entered. Data synchronization can still be performed when DataInLong network connectivity is available.
Table	Support selecting or manually entering the Table Name to be read. When the data source network is disconnected, resulting in the inability to directly pull library information, the table name can be manually entered. Data synchronization can still be performed when DataInLong network connectivity is available.
Split Key	Specify the field for data sharding. Once specified, concurrent tasks will be initiated for data synchronization. You can use a column from the source data table as the split key. It is recommended to use the primary key or an indexed column as the split key.
Filter Condition (Optional)	Fill in the corresponding filter statement according to the data type. This statement will serve as the filter condition for the data to be synchronized.
Advanced Settings (optional)	You can configure parameters based on business requirements.

StarRocks Offline Single Table Write Node Configuration

©2013-2025 Tencent Cloud. All rights reserved.



Parameters	Description
Data Flow Direction	Select the configured StarRocks Data Source at the target end
Database	Support selecting or manually entering the Library Name to be read. By default, the database bound to the data source is used as the default library. Other databases need to manually enter the Library Name. When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Support selecting or manually entering the Table Name to be read. When the data source network is disconnected, resulting in the inability to directly pull library information, the table name can be manually entered. Data synchronization can still be performed when DataInLong network connectivity is available.
Table overwrite	After enabling, SR writing will support table-level Atomic Overwrite operations. Before writing data, a new table with the same structure will be created using the CREATE TABLE LIKE statement. After importing the new data into the new table, the old table will be atomically replaced with the swap method to achieve table overwrite.
Maximum rows per submission	The maximum number of records for a single StreamLoad import during task run. Default is 500,000.
Maximum bytes per submission	The maximum byte count for a single StreamLoad import during task run. Default is 100M.
Row Separator	Default is \n.
Pre-SQL	SQL statements executed before the synchronization task, filled in according to the correct SQL syntax corresponding to the data source type.
Post-SQL	SQL statements executed after the synchronization task, filled in according to the correct SQL syntax corresponding to the data source type.
Advanced Settings (optional)	You can configure parameters based on business requirements.

©2013-2025 Tencent Cloud. All rights reserved.

Data type conversion is supported

Read

StarRocks Data Type	Internal Type
TINYINT、SMALLINT、INT、BIGINT	Long
DECIMAL、DOUBLE、FLOAT	Double
CAHR、VARCHAR、ARRAY、STRUCT	String
DATE	Date
BINARY	Bytes
BOOLEAN	Boolean

Write

Internal Type	StarRocks Data Type
Long	BOOLEAN、INT
Double	DOUBLE
String	CAHR、VARCHAR、ARRAY、STRUCT
Date	DATE、DATETIME

StarRocks Script Demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": [
                                "jdbc:mysql://ip:8030/source_database?query_timeout=100"
                            ],
                            "table": [ // Source table
                                "source_table"
                            ]
                        }
                    ],
                    "where": "id>10", // Filters
                    "splitPk": "id", // Split key
                    "username": "root"
                },
                "name": "starrocksreader"
            },
            "transformer": [],
            "writer": {
```

©2013–2025 Tencent Cloud. All rights reserved.

```
                "parameter": {
                    "postSql": [ // Post-SQL
                        ""
                    ],
                    "swapTableName": "sink_table_timestamp", // After enabling table overwrite, first
create a table with the same structure using create table like, then replace the new table via swap
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "swapTable": "true", // Whether to enable table overwrite
                    "timeZone": "+08:00",
                    "database": "sink_database", // Target database
                    "password": "******",
                    "beLoadUrl": [
                        "ip:8030"
                    ],
                    "labelPrefix": "",
                    "maxBatchByteSize": 104857600, // Maximum bytes per submission
                    "lineDelimiter": "\n", //Row Separator
                    "jdbcUrl": "jdbc:mysql://ip:8030/sink_database?query_timeout=100",
                    "connectTimeout": -1,
                    "table": "sink_table", // Target table
                    "maxBatchRows": 500000, // Maximum lines per submission
                    "username": "root",
                    "preSql": [ // Pre-SQL
                        ""
                    ]
                },
                "name": "starrockswriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

DLC Data Source

Last updated：2024-12-12 19:02:26

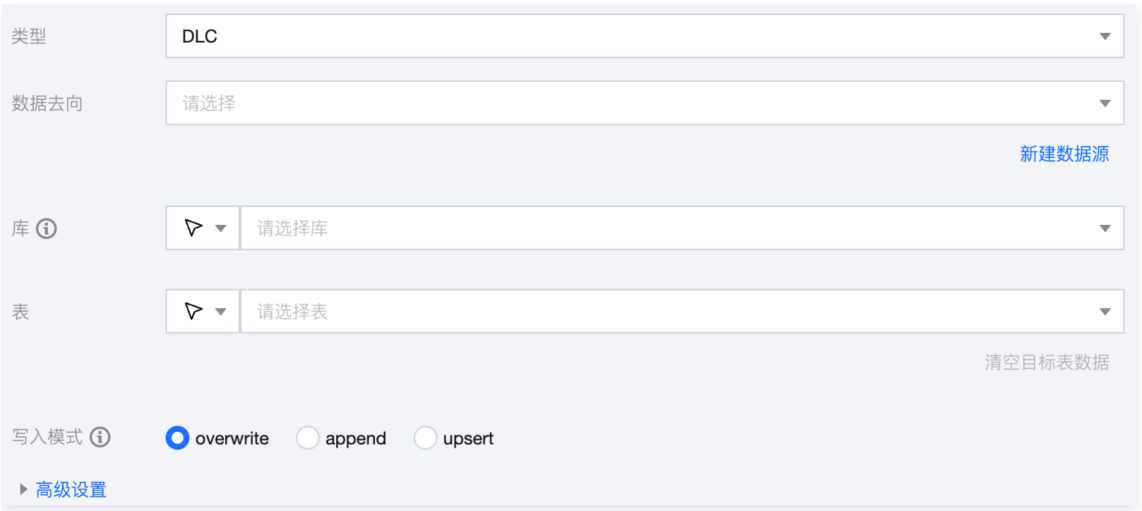
DLC Offline Single Table Read Node Configuration



Parameters	Description
Data Source	Available DLC data sources.
Database	Support selection or manual input of the library name to be read. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Support selecting or manually entering the Table Name to be read.
Split Key	Specify the field for data sharding. Once specified, concurrent tasks will be initiated for data synchronization. You can use a column from the source data table as the split key. It is recommended to use the primary key or an indexed column as the split key.
Filter Condition (Optional)	In actual business scenarios, it is common to synchronize the data of the current day by setting the where condition as gmt_create>$bizdate. • The WHERE condition can effectively perform business incremental synchronization. • If the where clause is not specified, including not providing the key or value for where, data synchronization will be considered as full data synchronization.
Advanced Settings (optional)	You can configure parameters based on business requirements.

DLC Offline Single Table Write Node Configuration

©2013-2025 Tencent Cloud. All rights reserved.



Parameters	Description
Data Flow Direction	DLC data source to be written to.
Database	Supports selecting or manually entering the name of the database to be written to. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Supports selecting or manually entering the name of the table to be written to. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong. • When the source table is MySQL, ES, or Kafka type, the target table can be created with one click.
Write mode	DLC write supports three modes: • overwrite: Overwrite existing data. • append: Append write, errors on primary key conflict. • upsert: Update write, updates data on primary key conflict.
Advanced Settings (optional)	You can configure parameters based on business requirements.

Data type conversion is supported

Read

DLC Data Type	Internal Type
INTEGER	Long
FLOAT、DOUBULE、DECIMAL	Double
STRING、VARCHAR、CHAR、ARRAY	String
TIMESTAMP、DATE、TIMESTAMP WITH TIME ZONE	Date
BOOLEAN	Boolean

Write

©2013–2025 Tencent Cloud. All rights reserved.

Internal Type	DLC Data Type
Long	INTEGER
Double	FLOAT、DOUBULE、DECIMAL
String	STRING、MAP、STRUCT
Date	TIMESTAMP、DATE、TIMESTAMP WITH TIME ZONE
Boolean	BOOLEAN
Bytes	BINARY

FAQs

1. The actual value of the DLC field is null, but the read value is an empty string '', causing dirty data when converting to LONG type

Cause:
Since the integrated DLC read task submits the Query SQL to the dlc, the dlc engine processes it and the integration directly reads the Result File. In the file, null is handled as an empty string, which the integration cannot process into a specified value.
Solutions:
Use the DLC function IFNULL to convert the value when configuring the task. This pre-processes the value in the dlc engine. Similar situations can be handled using other DLC functions.



DLC Script Demo

If you configure an offline task using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "password": "******",
                    "resultFormat": "CSV",
                    "hadoopConfig": {
                        "service.region": "ap-beijing",
                        "fs.cosn.userinfo.secretId": "***",
                        "service.engine.type": "BATCH",
                        "service.jwt.secret": "******",
                        "executor.secret.key": "***",
                        "owner.uin": "******",
                        "operator.uin": "******",
                        "fs.cosn.trsf.fs.ofs.tmp.cache.dir": "/tmp/",
                        "dlc.spark.executor.num": "2",
                        "dlc.spark.executor.max.num": "5",
                        "tmp.token.open": "true",
                        "fs.cosn.userinfo.secretKey": "***",
```

©2013-2025 Tencent Cloud. All rights reserved.

```
                        "fs.cosn.impl": "org.apache.hadoop.fs.CosFileSystem",
                        "fs.cosn.trsf.fs.ofs.user.appid": ******,
                        "user.appid": ******,
                        "executor.secret.id": "***",
                        "fs.cosn.tmp.dir": "/tmp/hadoop_cos",
                        "fs.cosn.credentials.provider":
"com.tencent.wedata.datax.plugin.reader.dlcreader.provider.WeDataCredentialsProvider",
                        "gateway.url": "***",
                        "fs.cosn.bucket.region": "ap-beijing",
                        "dlc.spark.driver.size": "large",
                        "fs.cosn.trsf.fs.ofs.bucket.region": "ap-beijing",
                        "dlc.spark.executor.size": "medium"
                    },
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": [ //JDBC contains information about the database and engine
                                "jdbc:dlc:dlc.internal.tencentcloudapi.com?
task_type=SparkSQLTask&database_name=source_database&datasource_connection_name=DataLakeCatalog&region=ap-
beijing&data_engine_name=source_engine"
                            ],
                            "table": [ // Source table
                                "source_table"
                            ]
                        }
                    ],
                    "where": "id>10", // Filters
                    "skipHeader": "true",
                    "splitPk": "id", // Split key
                    "username": "***"
                },
                "name": "dlcreader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "endpoint": "dlc.internal.tencentcloudapi.com",
                    "secretKey": "***",
                    "dbName": "sink_database",
                    "hadoopConfig": {
                        "fs.cosn.userinfo.region": "ap-beijing",
                        "service.region": "ap-beijing",
                        "qcloud.dlc.credential.provider.class":
"com.tencent.wedata.datax.plugin.writer.dlcwriter.provider.WeDataDLCCredentialProvider",
                        "service.engine.type": "BATCH",
                        "service.jwt.secret": "******",
                        "executor.secret.key": "***",
                        "owner.uin": "******",
                        "operator.uin": "******",
                        "fs.cosn.posix_bucket.fs.impl": "org.apache.hadoop.fs.CosNFileSystem",
                        "tmp.token.open": "true",
                        "dlc.spark.executor.num": "1",
                        "dlc.spark.executor.max.num": "3",
                        "request.identity.token": "******",
                        "fs.lakefs.impl": "org.apache.hadoop.fs.lakefs.LakeFileSystem",
                        "fs.cosn.impl": "org.apache.hadoop.fs.CosFileSystem",
                        "user.appid": "******",
                        "executor.secret.id": "***",
```

©2013–2025 Tencent Cloud. All rights reserved.

```
                        "service.secret.id": "***",
                        "fs.cosn.credentials.provider":
"com.tencent.wedata.datax.plugin.writer.dlcwriter.provider.WeDataFsCosCredentialsProvider",
                        "service.endpoint": "dlc.internal.tencentcloudapi.com",
                        "gateway.url": "***",
                        "dlc.spark.driver.size": "medium",
                        "service.secret.key": "***",
                        "dlc.spark.executor.size": "medium"
                    },
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    //JDBC contains information about the database and engine
                    "jdbcUrl": "jdbc:dlc:dlc.internal.tencentcloudapi.com?
task_type=SparkSQLTask&database_name=sink_database&datasource_connection_name=DataLakeCatalog&region=ap-
beijing&data_engine_name=sink_engine",
                    "secretId": "***",
                    "writeMode": "overwrite", // Write mode
                    "region": "ap-beijing",
                    "tableName": "sink_table", // Target table
                    "on": []
                },
                "name": "dlcwriter"
            }
        }
    ],
    "setting": {
        "errorLimit": {
            "record": 0 // Dirty data threshold
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

Iceberg Data Source

Last updated：2024-12-12 19:02:54

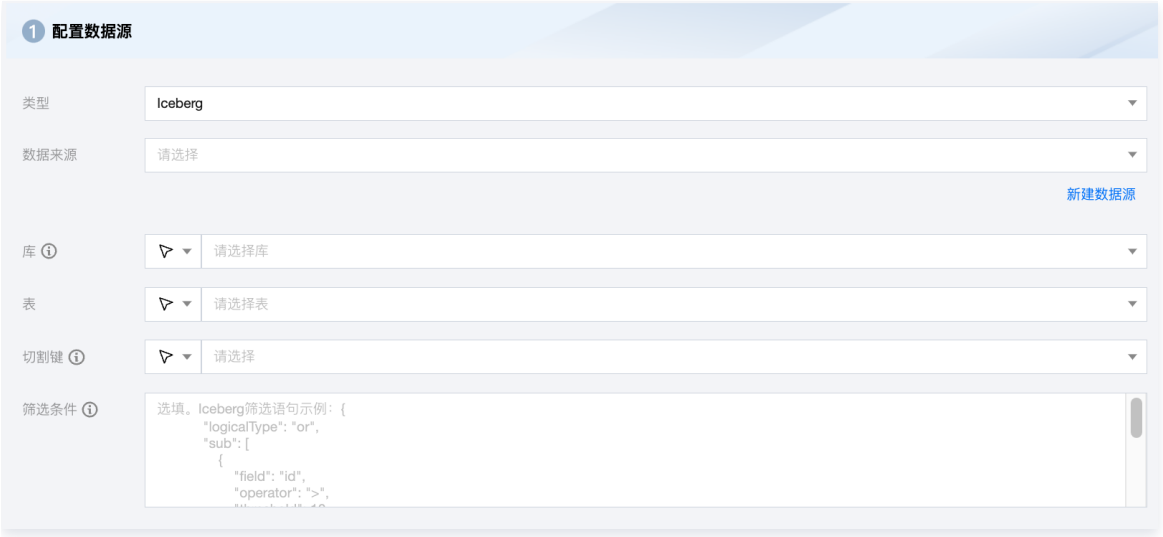
Supported Editions

Supports Iceberg 1.1.x version.

Use Limits

1. A connection to the Hive Metastore service is required. Please correctly configure the Metastore Thrift protocol's IP and port in the data source. If you use a self-defined Iceberg data source, you also need to upload hive-site.xml, core-site.xml, and hdfs-site.xml.
2. Currently only supports Hive catalog, Hadoop catalog is not supported.
3. The WHERE condition for data source reading currently only supports the Iceberg Java API, not the Spark SQL syntax. For details, please refer to Iceberg JavaAPI Expressions.

Iceberg Offline Single Table Read Node Configuration



Parameters	Description
Data Source	Available Iceberg Data Source.
Database	Support selection or manual input of the library name to be read. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Support selecting or manually entering the Table Name to be read.
Split Key	Specify the field for data sharding. Once specified, concurrent tasks will be initiated for data synchronization. You can use a column from the source data table as the split key. It is recommended to use the primary key or an indexed column as the split key. **Note:** If you want to start concurrent tasks for data synchronization, you must specify the cutting key, otherwise, it cannot be started.

©2013-2025 Tencent Cloud. All rights reserved.

Filter Condition (Optional)	In practical business scenarios, it is common to select the data of the current day for synchronization and set the WHERE condition as gmt_create>$bizdate. The WHERE condition can effectively perform incremental business synchronization. If you do not provide a WHERE statement, including the key or value, the data synchronization will be considered as full data synchronization.

Iceberg Offline Single Table Write Node Configuration



Parameters	Description
Data Flow Direction	Iceberg data source to be written.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Supports selection, or manual input of the table name to be written to • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Write mode	Iceberg write supports three modes: • overwrite: Overwrite. • append: Append write. • upsert: Update write according to the set primary key field.

Data type conversion is supported

Read

Iceberg Data Type	Internal Type
int、long	Long
float、double、decimal	Double
string、fixed、binary、struct、list、map	String
date、time、timestamp、timestamptz	Date
boolean	Boolean

Write

Internal Type	Iceberg Data Type
Long	int、long(bigint)

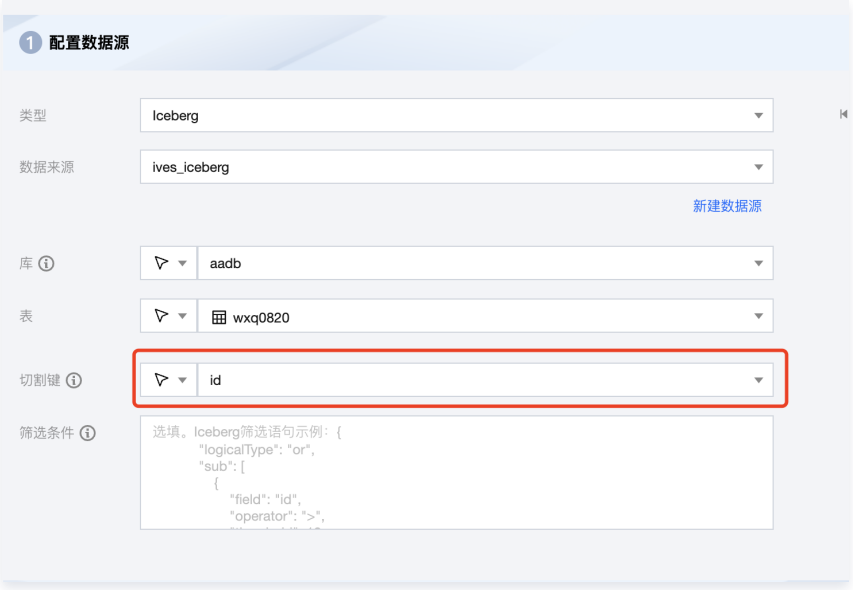
©2013–2025 Tencent Cloud. All rights reserved.

Double	float、double、decimal
String	string、struct、list、map
Date	date、time、timestamp、timestamptz
Bytes	binary
Boolean	boolean

Practice Tutorial

Optimize Iceberg Table Read Rate

1. Currently, Iceberg supports sharded concurrent reading, and the cutting key can be of types string, long, int, decimal, or timestamp.

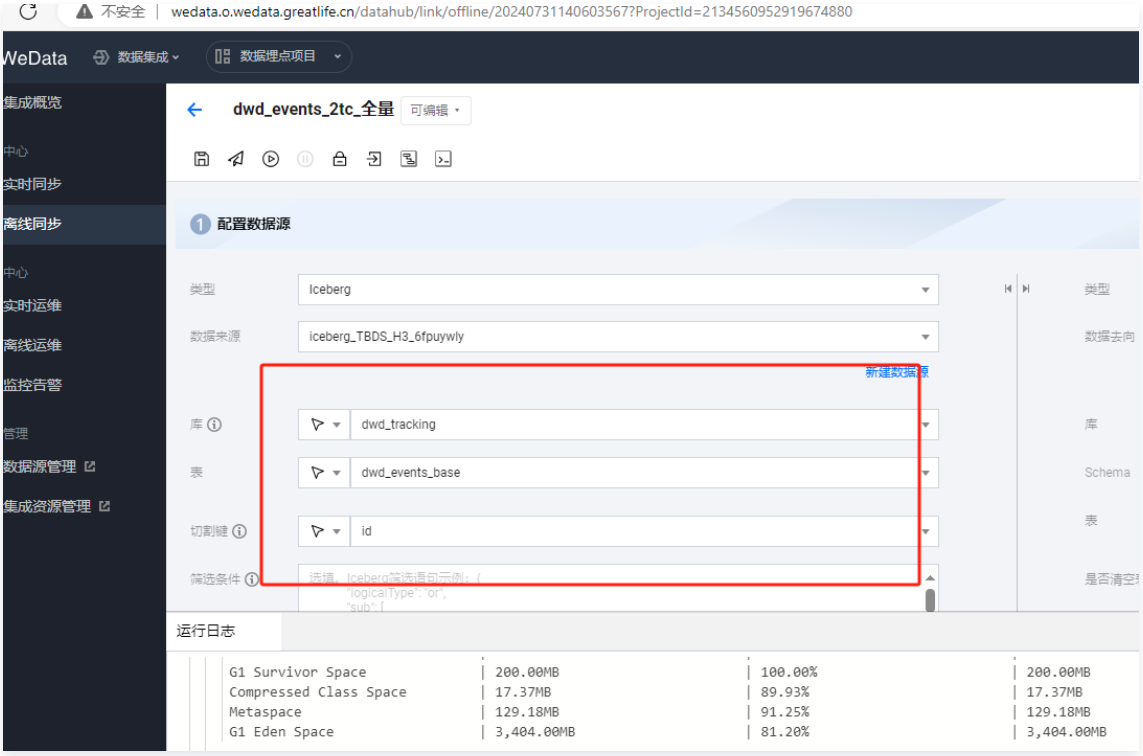


2. In practice, to achieve optimal read performance, it is best to set Iceberg tables as partition tables and choose the partition fields as cutting keys.

©2013–2025 Tencent Cloud. All rights reserved.



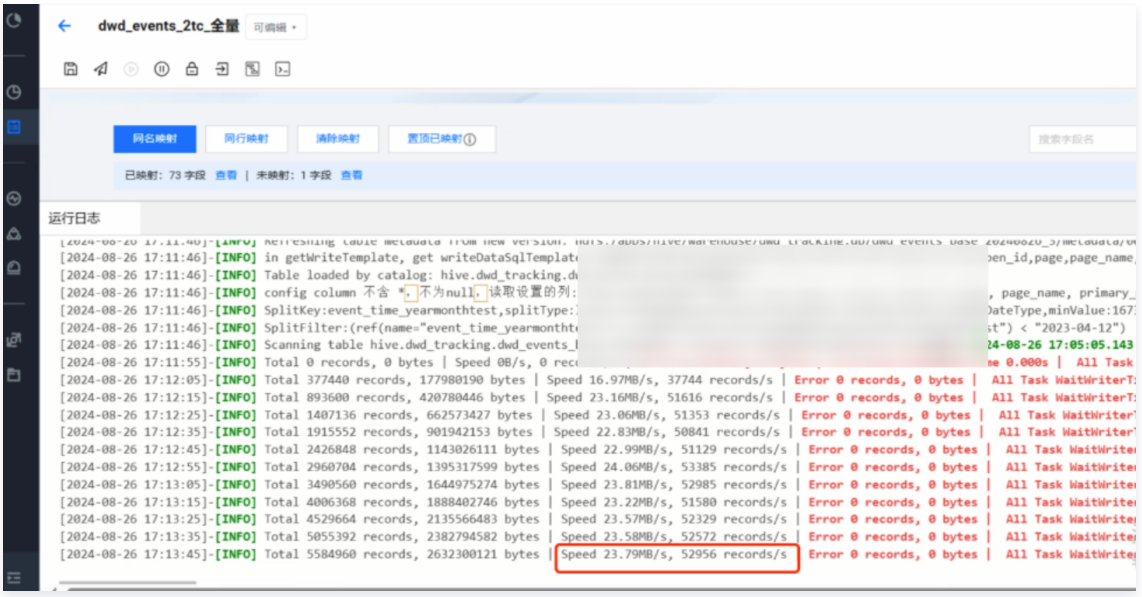
3. Example: Configuring an offline task to fully read an Iceberg table. The original table is a non-partitioned table with 74 columns and 200 million rows. By choosing a field id as the cutting key and setting 8 concurrent reads, the sync rate is only about 4M/s.



4. After adjusting the original table to a partition table and using 'year-month' as the partition key, 28 partitions were created. Using the same 8 concurrent reads and choosing the partition field event_time_yearmonthtest as the cutting key, the rate

©2013-2025 Tencent Cloud. All rights reserved.

increased to 24M/s, a 6-fold performance improvement.



Iceberg script demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "catalogWarehouseLocation": "/user/hive/warehouse",
                    "filter": { // Filters
                        "sub": [
                            {
                                "field": "id",
                                "threshold": 10,
                                "operator": ">"
                            },
                            {
                                "field": "id",
                                "threshold": 3,
                                "operator": "<"
                            }
                        ],
                        "logicalType": "or"
                    },
                    "catalogURI": "thrift://ip:7004",
                    "dbName": "source_database", // Source database
                    "hadoopConfig": {  // If Hadoop is a high availability cluster, you need to configure
NameNode high availability information
                        "dfs.namenode.https-address.mycluster.nn2": "ip:4009",
                        "dfs.namenode.https-address.mycluster.nn1": "ip:4009",
                        "dfs.ha.namenodes.mycluster": "nn1,nn2",
                        "dfs.namenode.servicerpc-address.mycluster.nn2": "ip:4010",
                        "dfs.namenode.servicerpc-address.mycluster.nn1": "ip:4010",
                        "dfs.namenode.rpc-address.mycluster.nn2": "ip:4007",
                        "dfs.namenode.rpc-address.mycluster.nn1": "ip:4007",
                        "dfs.namenode.http-address.mycluster.nn2": "ip:4008",
                        "dfs.namenode.http-address.mycluster.nn1": "ip:4008",
```

©2013-2025 Tencent Cloud. All rights reserved.

```
                        "dfs.client.failover.proxy.provider.mycluster":
"org.apache.hadoop.hdfs.server.namenode.ha.ConfiguredFailoverProxyProvider",
                        "dfs.nameservices": "mycluster",
                        "fs.defaultFS": "hdfs://mycluster",
                    },
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "splitPk": "id", // Split key
                    "tableName": "source_table" // Source table
                },
                "name": "icebergreader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "catalogWarehouseLocation": "/user/hive/warehouse",
                    "catalogURI": "thrift://ip:7004",
                    "dbName": "sink_database", // Target table
                    "hadoopConfig": { // If Hadoop is a high availability cluster, you need to configure
NameNode high availability information
                        "dfs.namenode.https-address.mycluster.nn2": "ip:4009",
                        "dfs.namenode.https-address.mycluster.nn1": "ip:4009",
                        "dfs.ha.namenodes.mycluster": "nn1,nn2",
                        "dfs.namenode.servicerpc-address.mycluster.nn2": "ip:4010",
                        "dfs.namenode.servicerpc-address.mycluster.nn1": "ip:4010",
                        "dfs.namenode.rpc-address.mycluster.nn2": "ip:4007",
                        "dfs.namenode.rpc-address.mycluster.nn1": "ip:4007",
                        "dfs.namenode.http-address.mycluster.nn2": "ip:4008",
                        "dfs.namenode.http-address.mycluster.nn1": "ip:4008",
                        "dfs.client.failover.proxy.provider.mycluster":
"org.apache.hadoop.hdfs.server.namenode.ha.ConfiguredFailoverProxyProvider",
                        "dfs.nameservices": "mycluster",
                        "fs.defaultFS": "hdfs://mycluster",
                        "spark.eventLog.dir": "hdfs://mycluster/spark-history",
                    },
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "writeMode": "overwrite", // Write mode
                    "tableName": "sink_table", // Target table
                    "on": []
                },
                "name": "icebergwriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

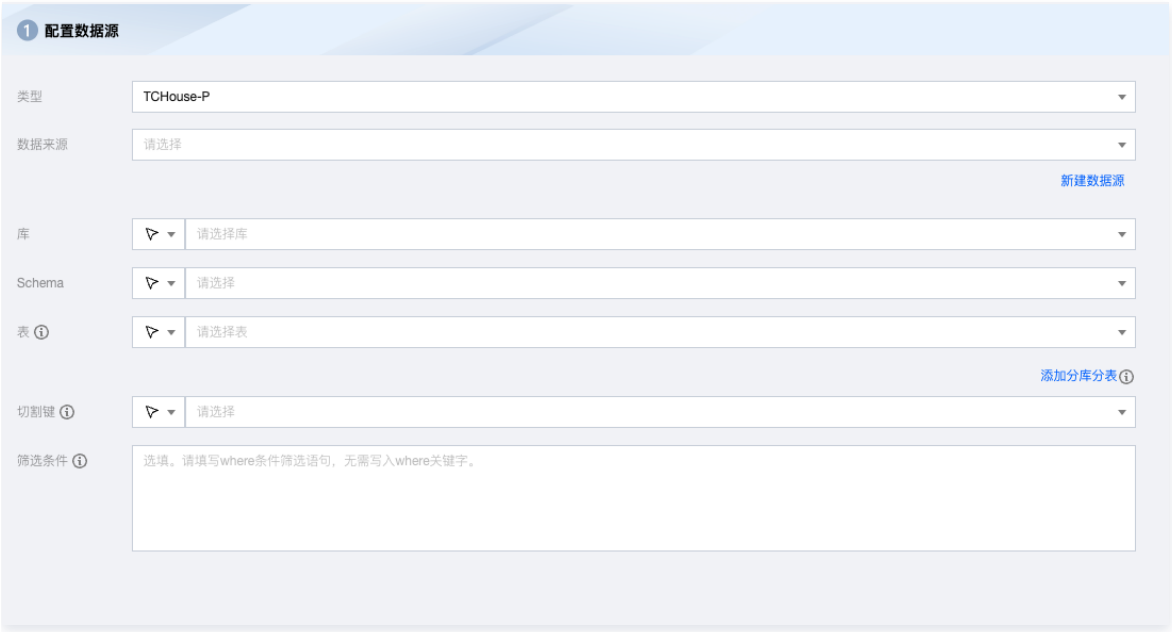
TCHouse-P Data Source

Last updated: 2024-12-12 19:03:41

Supported Editions

Currently supports TCHouse-P 3.16.x versions

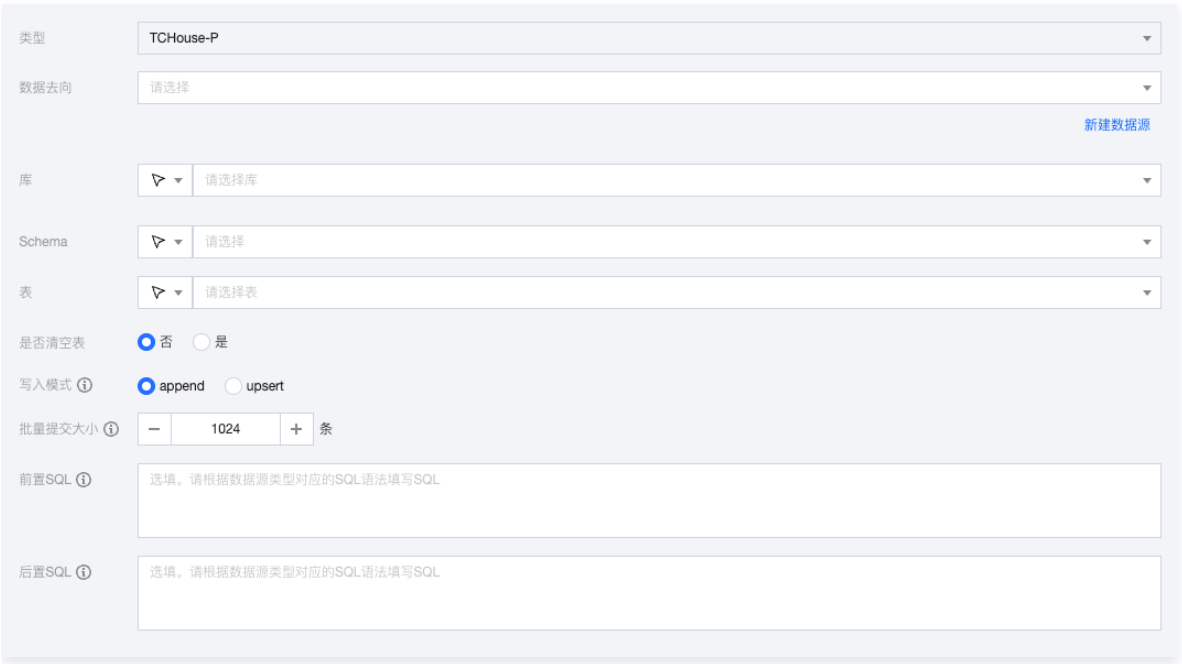
TCHouse-P Offline Single Table Read Node Configuration



Parameters	Description
Data Source	Select the configured TCHouse-P data source from the source terminal
Database	Support selecting or manually entering the Library Name to be read. By default, the database bound to the data source is used as the default library, and other databases need to be manually entered. When the data source network is disconnected, resulting in the inability to directly pull library information, the database name can be manually entered. Data synchronization can still be performed when DataInLong network connectivity is available.
Schema	Support selecting or manually entering the Schema Name to be read
Table	Support selecting or manually entering the Table Name to be read. When the data source network is disconnected, resulting in the inability to directly pull library information, the table name can be manually entered. Data synchronization can still be performed when DataInLong network connectivity is available.
Split Key	Specify the field for data sharding. Once specified, concurrent tasks will be initiated for data synchronization. You can use a column from the source data table as the split key. It is recommended to use the primary key or an indexed column as the split key.
Filter Condition (Optional)	Fill in the corresponding filter statement according to the data type. This statement will serve as the filter condition for the data to be synchronized.

TCHouse-P Offline Single Table Write Node Configuration

©2013-2025 Tencent Cloud. All rights reserved.



Parameters	Description
Data Flow Direction	Select the configured TCHouse-P data source from the target terminal.
Database	Support selecting or manually entering the Library Name to be read. By default, the database bound to the data source is used as the default library, and other databases need to be manually entered. When the data source network is disconnected, resulting in the inability to directly pull library information, the database name can be manually entered. Data synchronization can still be performed when DataInLong network connectivity is available.
Schema	Support selecting or manually entering the Schema Name to be read.
Table	Support selecting or manually entering the Table Name to be read. When the data source network is disconnected, resulting in the inability to directly pull library information, the table name can be manually entered. Data synchronization can still be performed when DataInLong network connectivity is available.
Clear Table	Before writing to this data table, you can manually choose whether to clear the data table.
Write mode	• append: When a primary key/unique index conflict occurs, conflicting rows cannot be written. • upsert: When there is no primary key/unique index conflict, new rows can be inserted; if a primary key/unique index conflict occurs, it will be updated.
Batch Submission Size	The size of the record count for one-time batch submission can greatly reduce the number of network interactions between the data synchronization system and TCHouse-P, and improve overall throughput. If this value is set too high, it will cause the data synchronization process to run into OOM exceptions.
Pre-SQL	SQL statements executed before the synchronization task, filled in according to the correct SQL syntax corresponding to the data source type.
Post-SQL	SQL statements executed after the synchronization task, filled in according to the correct SQL syntax corresponding to the data source type.

Data type conversion is supported

Supported data types and conversion mappings for TCHouse-P read operations are as follows (when processing TCHouse-P, the data type of TCHouse-P data source will first be mapped to the data type of the data processing engine):

Read

TCHouse-P Data Types	Internal Type

©2013-2025 Tencent Cloud. All rights reserved.

smallint，integer，bigint，serial，bigserial	Long
double precision，money，numeric，real	Double
varchar(character varying)，char(character)	String
date，time，timestamp	Date
bool(boolean)	Boolean
bytea	Bytes

Write

Internal Type	TCHouse-P Data Types
Long	smallint，integer，bigint
Double	double precision，numeric，real
String	varchar(character varying)，char(character)，bit，money
Date	date，time，timestamp
Boolean	bool
Bytes	bytea

TCHouse-P Script Demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "password": "*******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": [
                                "jdbc:postgresql://ip:9000/database?reWriteBatchedInserts=true"
                            ],
                            "table": [ // Source table
                                "schema.source_table"
                            ]
                        }
                    ],
                    "where": "id>10", // Filters
                    "splitPk": "id", // Split key
                    "username": "root"
                },
                "name": "postgresqlreader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "postSql": [ // Post-SQL
```

©2013-2025 Tencent Cloud. All rights reserved.

```
                        ""
                    ],
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": "jdbc:postgresql://ip:9000/postgres?reWriteBatchedInserts=true",
                            "table": [ // Target table
                                "schema.sink_table"
                            ]
                        }
                    ],
                    "batchSize": 1024, // Batch submission size
                    "username": "root",
                    "preSql": [ // Pre-SQL
                        ""
                    ]
                },
                "name": "tbasewriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

ClickHouse Data Source

Last updated：2024-12-12 19:04:15

Supported Editions

Supports ClickHouse 20.7+ versions

Use Limits

1. Supports connecting to ClickHouse using JDBC and only supports reading data using JDBC Statement.
2. Supports filtering certain columns, column reordering, and other features. You can fill in the columns yourself.

ClickHouse Offline Single Table Read Node Configuration

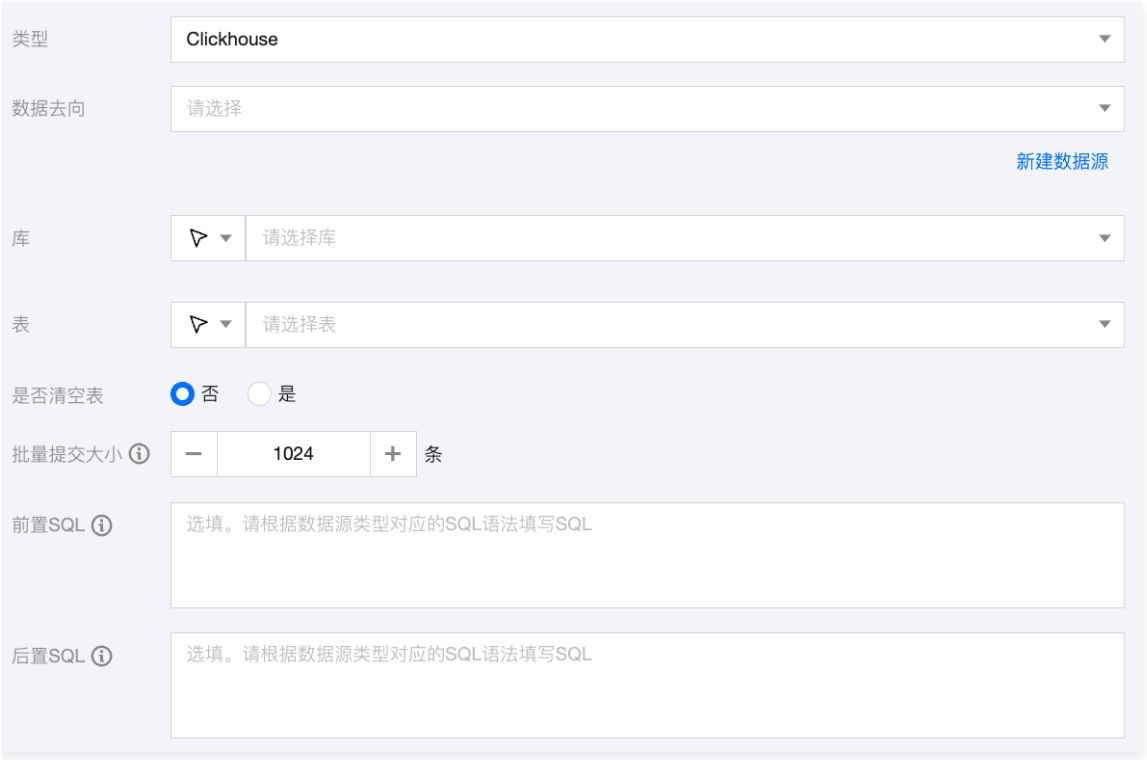


Parameters	Description
Data Source	Available ClickHouse Data Sources.
Database	Support selecting or manually entering the name of the database to be read • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Support selecting or manually entering the Table Name to be read.
Split Key	• When extracting data with ClickHouse, if splitPk is specified, it means you want to use the field represented by splitPk for data sharding, hence data synchronization will initiate concurrent tasks to improve data synchronization efficiency. • You can use a column from the source data table as the cutting key. It is recommended to use a primary key or an indexed column as the cutting key. Only fields of type Integer are supported. When reading data, the data is sharded based on the configured fields to achieve concurrent reading, which can improve data synchronization efficiency.
Filter Condition (Optional)	In actual business scenarios, it is common to synchronize the data of the current day by setting the where condition as gmt_create>$bizdate. • The WHERE condition can effectively perform business incremental synchronization.

©2013-2025 Tencent Cloud. All rights reserved.

	• If the where clause is not specified, including not providing the key or value for where, data synchronization will be considered as full data synchronization.

ClickHouse Offline Single Table Write Node Configuration



Parameters	Description
Data Flow Direction	Clickhouse data source to be written.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Supports selection, or manual input of the table name to be written to • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Clear Table	You can manually select whether to clear the Clickhouse data table before writing.
Batch Submission Size	The size of records submitted in batch at once. This value can significantly reduce the number of network interactions between the data synchronization system and Clickhouse and improve overall throughput. If this value is set too high, it might cause an Out Of Memory (OOM) exception in the data synchronization process.
Pre-SQL (optional)	SQL command executed before the synchronization task. Fill in the SQL according to the correct SQL syntax corresponding to the data source type. For example, clear old data in the table before execution (truncate table tablename).
Post-SQL (optional)	SQL statements executed after the synchronization task are written according to the correct SQL syntax corresponding to the data source type, for example, adding a timestamp with `ALTER TABLE tablename ADD colname TIMESTAMP DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP`.

©2013-2025 Tencent Cloud. All rights reserved.

Data type conversion is supported

Read

ClickHouse Data Types	Internal Type
Integer、Smallint、Tinyint、Bigint	Long
Float32、Flout64、Decimal	Double
String、Array、Null	String
Date32、Datetime64	Date

Write

Internal Type	ClickHouse Data Types
Long	Integer、Smallint、Tinyint、Bigint
Double	Float32、Flout64、Decimal
String	String、Array
Date	Date32、Datetime64

FAQs

Failed to insert the same data after deletion

Cause:
Data inserted into a replica table will be divided into data blocks. Each data block will generate a block_id stored in the block subdirectory of the corresponding table directory in ZooKeeper.
Data blocks are automatically deduplicated based on block_id. For the **same data block** (same size and rows in the same order) that is written multiple times, the block will only be written once.
Therefore, repeated inserts of the same data into a replica table will only write the data the first time. When the data is deleted, it is successfully removed, but the block_id in ZooKeeper is not deleted. This causes subsequent inserts of the same data to be considered duplicate data by the table engine and skipped, so the data does not get written to the table and cannot be queried.
Solution:
Set insert_deduplicate=0 to temporarily disable the deduplication mechanism, find the corresponding block_id in ZooKeeper, and manually delete this block_id.**It is recommended to use truncate table to delete data.**.

ClickHouse Script Demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": [
                                "jdbc:clickhouse://ip:8123/source_database"
                            ],
```

©2013–2025 Tencent Cloud. All rights reserved.

```
                            "table": [ // Source table
                                "source_database.source_table"
                            ]
                        }
                    ],
                    "where": "id>10", // Filters
                    "splitPk": "id", // Split key
                    "username": "root"
                },
                "name": "clickhousereader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "postSql": [ // Post-SQL
                        ""
                    ],
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": "jdbc:clickhouse://172.1.1.1:1433/sink_database?compress=111",
                            "table": [ // Target table
                                "sink_database.sink_table"
                            ]
                        }
                    ],
                    "batchSize": 1024, // Batch submission size
                    "username": "root",
                    "preSql": [ // Pre-SQL
                        ""
                    ]
                },
                "name": "clickhousewriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

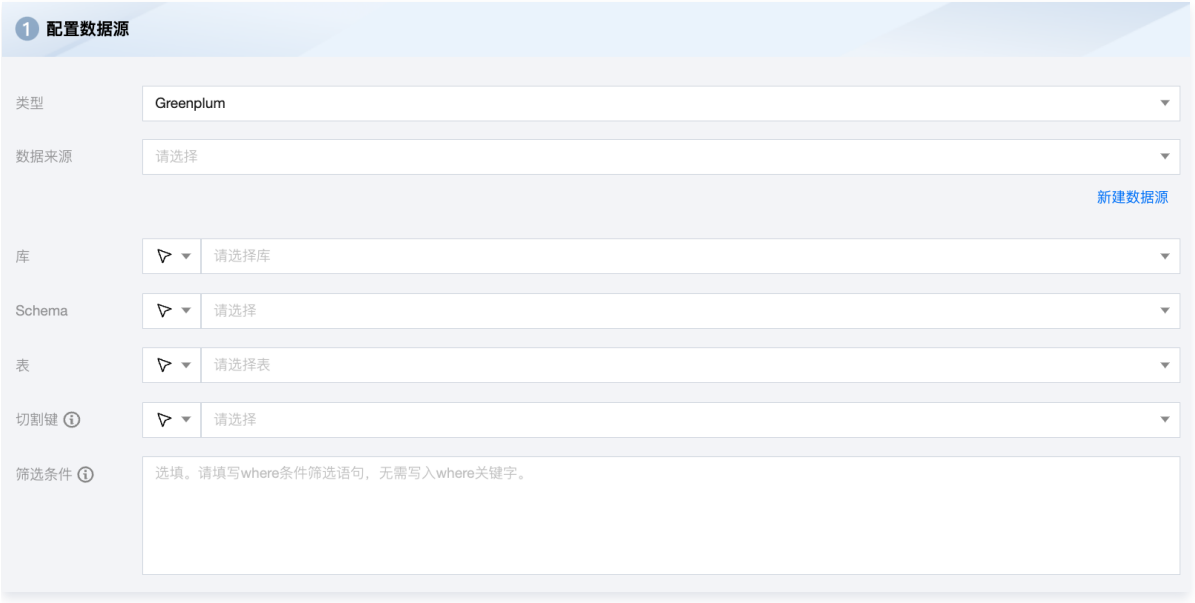
Greenplum Data Source

Last updated：2024-12-12 19:04:42

Supported Editions

Supports Greenplum 4.x, 5.x, 6.x versions.

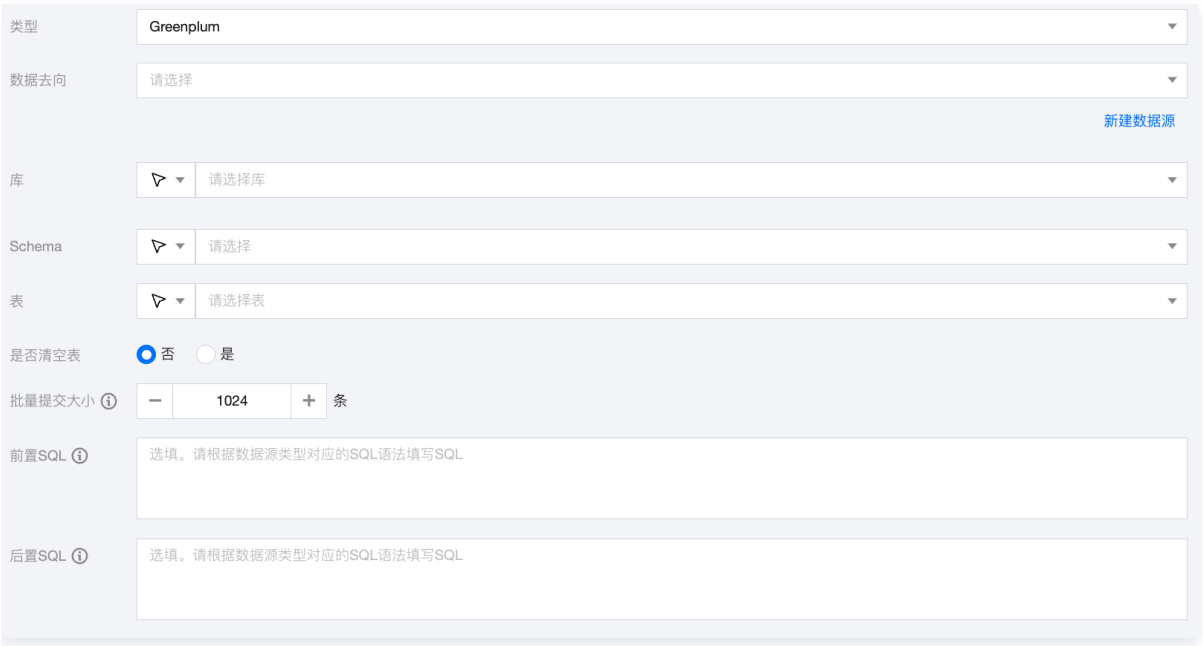
Greenplum Offline Single Table Read Node Configuration



Parameters	Description
Data Source	Available Greenplum Data Source.
Database	Support selecting or manually entering the name of the database to be read • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Support selecting or manually entering the Schema Name to be read.
Table	Support selecting or manually entering the Table Name to be read.
Split Key	Specify the field for data sharding. Once specified, concurrent tasks will be initiated for data synchronization. You can use a column from the source data table as the split key. It is recommended to use the primary key or an indexed column as the split key.
Filter Condition (Optional)	In actual business scenarios, it is common to synchronize the data of the current day by setting the where condition as gmt_create>$bizdate. • The WHERE condition can effectively perform business incremental synchronization. • If the where clause is not specified, including not providing the key or value for where, data synchronization will be considered as full data synchronization.

Greenplum Offline Single Table Write Node Configuration

©2013-2025 Tencent Cloud. All rights reserved.



Parameters	Description
Data Flow Direction	Greenplum data source to be written to.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Support selecting or manually entering the Schema Name to be read.
Table	Supports selection, or manual input of the table name to be written to • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Clear Table	Before writing to the Greenplum data table, you can manually select whether to clear the data table.
Batch Submission Size	The size of the batch submission record count. This value can greatly reduce the number of network interactions between the data synchronization system and Greenplum, thereby improving overall throughput. If this value is set too large, it may cause the data synchronization process to run into OOM exceptions.
Pre-SQL (optional)	SQL command executed before the synchronization task. Fill in the SQL according to the correct SQL syntax corresponding to the data source type. For example, clear old data in the table before execution (truncate table tablename).
Post-SQL (optional)	SQL statements executed after the synchronization task are written according to the correct SQL syntax corresponding to the data source type, for example, adding a timestamp with `ALTER TABLE tablename ADD colname TIMESTAMP DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP`.

Data type conversion is supported

Read

Greenplum Data Types	Internal Type

©2013-2025 Tencent Cloud. All rights reserved.

smallint、integer、bigint	Long
decimal、double precision、real	Double
character、character varying、	String
date	Date
boolean、bit	Boolean

Write

Internal Type	Greenplum Data Types
Long	smallint、integer、bigint
Double	decimal、double precision、real
String	character、character varying
Date	date
Boolean	bit、boolean

Greenplum Script Demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": [
                                "jdbc:postgresql://ip:5432/database?reWriteBatchedInserts=true"
                            ],
                            "table": [ // Source table
                                "schema.source_table"
                            ]
                        }
                    ],
                    "where": "id>10", // Filters
                    "splitPk": "id", // Split key
                    "username": "root"
                },
                "name": "postgresqlreader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "postSql": [ // Post-SQL
                        ""
                    ],
                    "password": "******",
```

©2013–2025 Tencent Cloud. All rights reserved.

```
                    "session": [
                        "set optimizer to 'off'",
                        "set enable_seqscan to 'off'"
                    ],
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": "jdbc:postgresql://ip:5436/database?reWriteBatchedInserts=true",
                            "table": [ // Target table
                                "schema.sink_table"
                            ]
                        }
                    ],
                    "batchSize": 1024, // Batch submission size
                    "username": "root",
                    "preSql": [ // Pre-SQL
                        ""
                    ]
                },
                "name": "postgresqlwriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

Hive Data Source

Last updated：2025-01-16 15:04:08

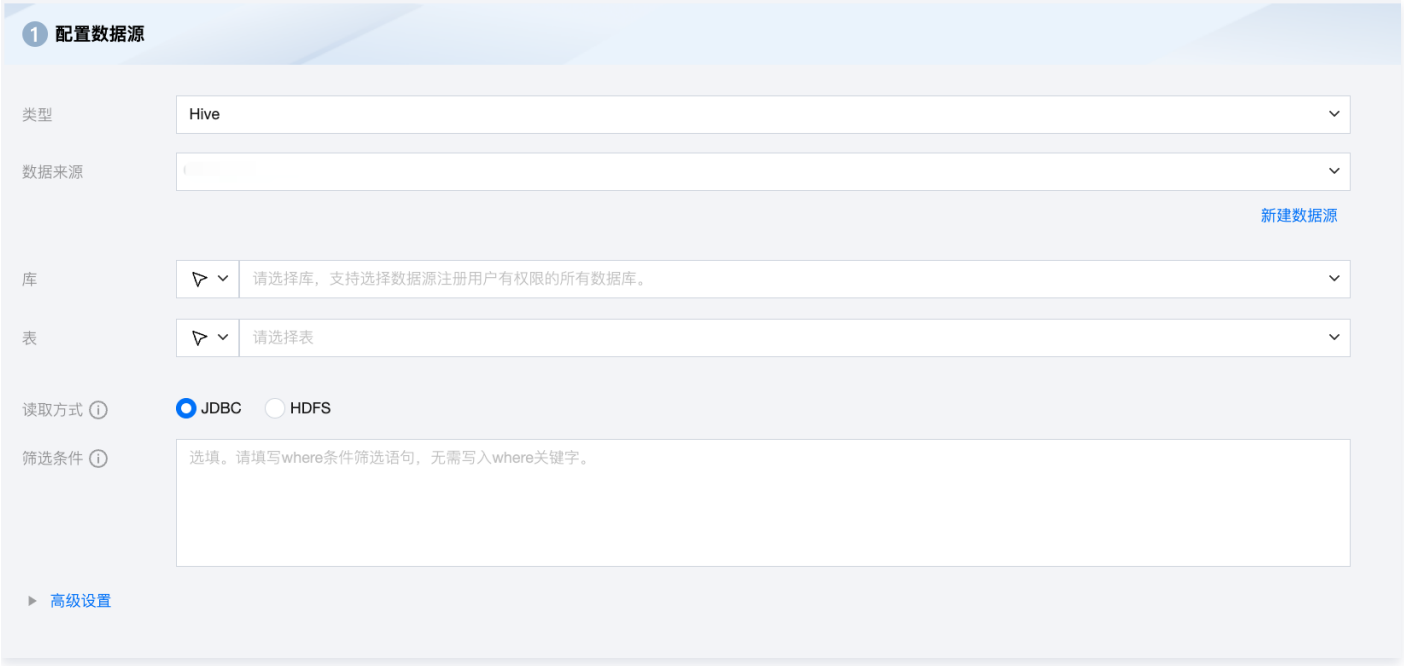
Supported Editions

Supports Hive 1.x, 2.x, 3.x versions.

Use Limits

1. Currently, only TextFile, ORCFile, and ParquetFile formats are supported for reading. It is recommended to use ORC or Parquet; currently, only TextFile, ORCFile, and ParquetFile formats are supported for writing. RCFile format is not supported yet.
2. Hive reading uses JDBC to connect to HiveServer2 to retrieve data. Make sure the HiveServer2 JDBC URL is correctly configured in the data source.
3. When the underlying storage of a hive table is in TXT format, it is not recommended to include characters used as delimiters in file data items for both JDBC and HDFS reading methods, as this will cause data parsing errors.
4. Hive writing requires a connection to the Hive Metastore service. Please correctly configure the Metastore Thrift protocol IP and port in the data source. If it's a self-defined Hive data source, you also need to upload hive-site.xml, core-site.xml, and hdfs-site.xml.
5. When rebuilding Hive tables with additional columns, the cascade keyword needs to be added to prevent Partition Metadata from not being updated, which can affect data queries.
6. In the process of offline synchronization to the Hive cluster using DataInLong, temporary files will be generated on the server-side and automatically deleted upon task completion. Pay attention to the file count limit of the HDFS directory on the server to avoid reaching the limit unexpectedly and causing the HDFS file system to become unavailable.
7. Hive's JDBC and HDFS reading methods do not support the Brotli compression format.
8. Usage limitations related to field types:
   - Hive will convert unsupported field types to STRING type when reading and synchronize to the target end.
   - Hive's JDBC and HDFS reading methods do not support reading UNION type.
   - Hive's JDBC reading method supports reading ARRAY, MAP, and STRUCT types.
   - Hive's HDFS reading mode
     - When the source-end is in Parquet format, reading ARRAY, MAP, and STRUCT types is supported.
     - When the source-end is in ORCFile format, reading ARRAY and MAP types is supported.
     - When the source-end is in TextFile format, reading ARRAY, MAP, STRUCT, and UNION types is not supported.

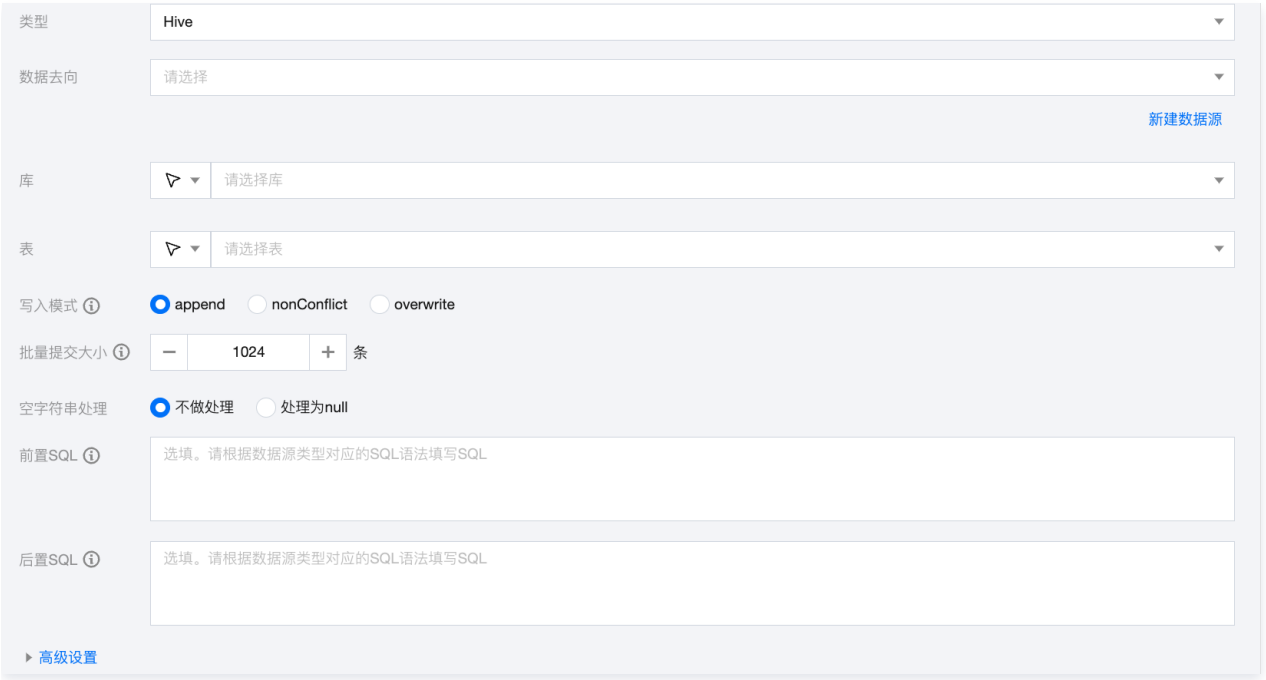
Hive offline single-table reading node configuration



©2013-2025 Tencent Cloud. All rights reserved.

Parameter	Description
Data Source	Available Hive Data Source
Database	Support selecting or manually entering the name of the database to be read • By default, the database bound to the data source will be used as the default database. For other databases, the database name must be entered manually • When the data source network is not connected, making it impossible to directly pull database information, manually enter the database name. Data synchronization can still be performed when the DataInLong network is connected
Table	Support selecting or manually entering the name of the table to be read
Read mode	• JDBC Read: Supports data filtering using Where conditions, but the read efficiency may be slow. • HDFS Read: Does not support data filtering using Where conditions, but directly accesses the underlying data files of the hive table for reading, resulting in relatively higher read efficiency. **Note:** This method does not support reading view tables.
Filter Condition (Optional)	When reading data using Hive JDBC, it supports filtering data with Where conditions. However, in this scenario, the Hive engine may generate MapReduce tasks at the underlying layer, resulting in slower efficiency
Advanced Settings (optional)	You can configure parameters based on business requirements.

Hive offline single-table writing node configuration



Parameter	Description
Data Flow Direction	The Hive data source to write into.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered.

©2013-2025 Tencent Cloud. All rights reserved.

	• When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Supports selection, or manual input of the table name to be written to • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Write mode	Hive writing supports three modes: • Append: Retain original data, new rows are appended • nonConflict: Reports an error on data conflict • Overwrite: Deletes original data and writes new data writeMode is a high-risk parameter. Please pay attention to the data output directory and write mode to avoid data deletion by mistake. The data loading behavior needs to be used with hiveConfig. Please be mindful of your configuration.
Batch Submission Size	The record count of a one-time batch submission can greatly reduce the number of network interactions between the data synchronization system and Hive, and improve overall throughput. If the value is set too high, it can lead to OOM exceptions during the data synchronization process.
Empty string processing	• No processing: Do not process empty strings during writing. • Processed as null: Empty strings are processed as null during writing.
Pre-SQL (optional)	SQL command executed before the synchronization task. Fill in the SQL according to the correct SQL syntax corresponding to the data source type. For example, clear old data in the table before execution (truncate table tablename).
Post-SQL (optional)	SQL statements executed after the synchronization task are written according to the correct SQL syntax corresponding to the data source type, for example, adding a timestamp with `ALTER TABLE tablename ADD colname TIMESTAMP DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP`.
Advanced Settings (optional)	You can configure parameters based on business requirements.

Data type conversion is supported

Read

The supported data types and corresponding conversions for Hive reading are as follows (when working with Hive, the data types from the Hive data source and the data processing engine will be mapped first):

Hive Data Type	Internal Type
TINYINT, SMALLINT, INT, BIGINT	LONG
FLOAT, DOUBLE	DOUBLE
STRING, CHAR, VARCHAR, STRUCT, MAP, ARRAY, UNION, BINARY	STRING
BOOLEAN	BOOLEAN
DATE, TIMESTAMP	DATE

Write

The supported data types and corresponding conversions for Hive reading are as follows:

Internal Type	Hive Data Type
LONG	TINYINT, SMALLINT, INT, BIGINT

©2013-2025 Tencent Cloud. All rights reserved.

DOUBLE	FLOAT, DOUBLE
STRING	STRING, CHAR, VARCHAR, STRUCT, MAP, ARRAY, UNION, BINARY
BOOLEAN	BOOLEAN
DATE	DATE, TIMESTAMP

FAQs

1.Hive on CHDFS table write error: Permission denied: No access rules matched

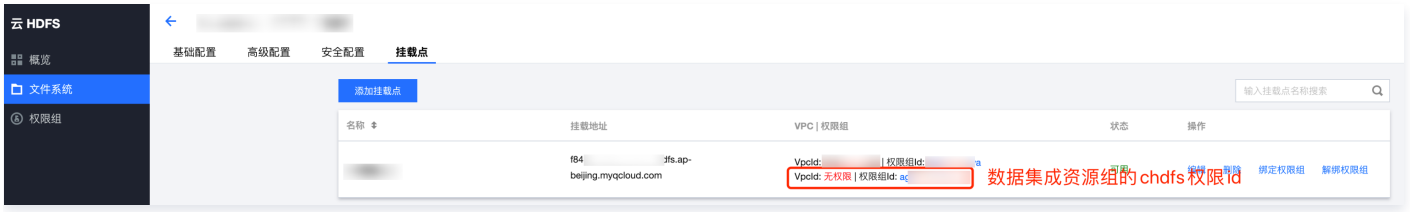
Issue Information:



Cause:

DataInLong resource group is a fully managed resource group. Its outbound IP range is not within the customer's VPC internal network segment, so you need to open security groups for the DataInLong resource group.

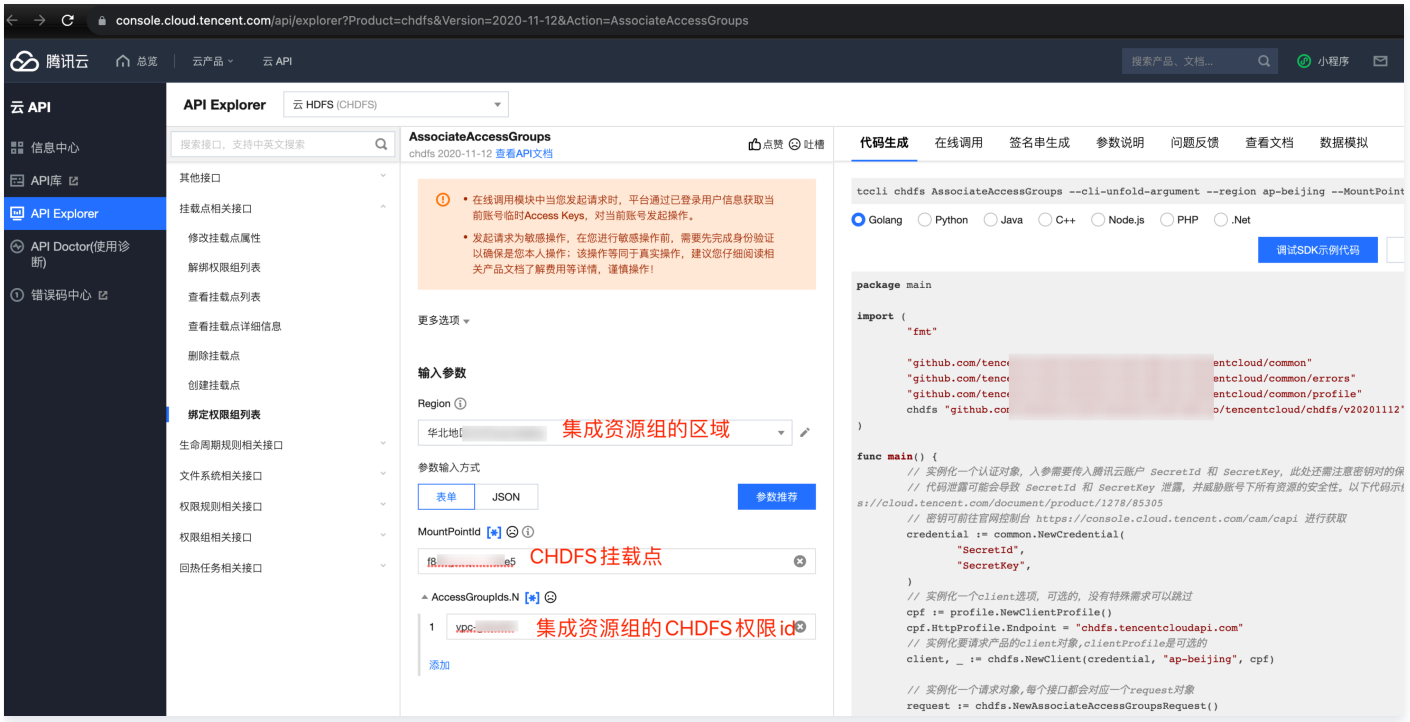
Please authorize the CHDFS permission ID for the integration resource group at the CHDFS mount point.



Solutions:

Open the page: API 3.0 Explorer, fill in the parameters, where **MountPointId** is the customer's CHDFS mount point, **AccessGroupIds.N** is the CHDFS permission ID for the DataInLong resource group (Note: permission IDs vary by region).

©2013–2025 Tencent Cloud. All rights reserved.



DataInLong resource group CHDFS permission ID and region mapping table:

Region	DataInLong resource group CHDFS permission ID
Beijing	ag-wgbku4no
Guangzhou	ag-x1bhppnr
Shanghai	ag-xnjfr9d3
Singapore	ag-phxtv0ah
USA Silicon Valley	ag-tgwl8bca
Virginia	ag-esxpwxjn

2. Some fields return NULL during synchronization

Cause:
After creating the Hive partition table, omitting the cascade keyword in subsequent ALTER statements results in unupdated Partition Metadata, affecting data queries.
Solutions:

```
alter table table_name add columns (time_interval  TIMESTAMP) cascade;
```

3. Hive write reports [fs.defaultFS] as a required parameter error

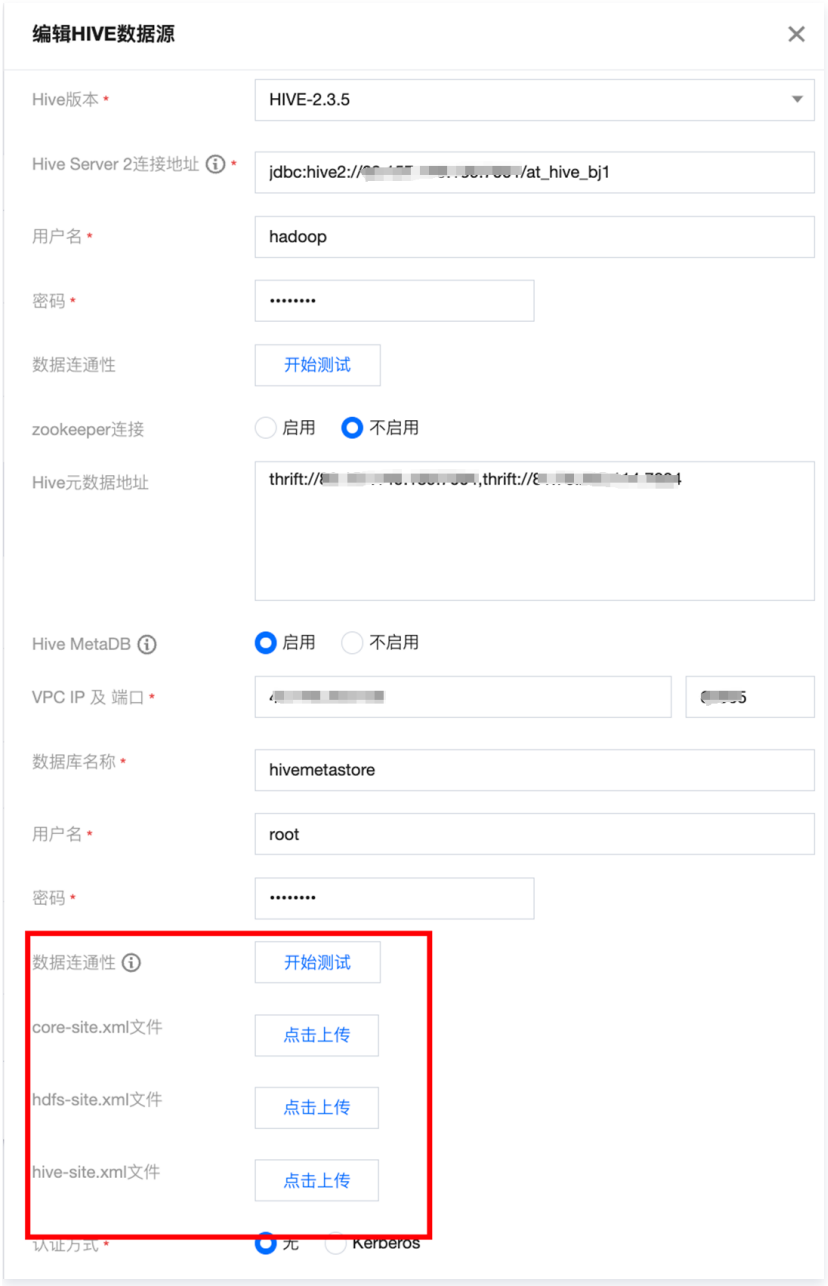
Issue Information:

```
[2023-09-06 17:29:34]-[ERROR] Exception when job run: com.tencent.wedata.common.exception.wedataException: Code:[HiveWriter-01], Description:[您缺失了必须填写的参数值.]. - 您提供配置文件有误，[fs.defaultFS]是必填参数，不允许为空或者留白 .
        at com.tencent.wedata.common.exception.wedataException.aswedataException(wedataException.java:28)
        at com.tencent.wedata.plugin.writer.hivewriter.HiveWriter$Job.init(HiveWriter.java:172)
        at com.tencent.wedata.core.job.JobContainer.initJobWriter(JobContainer.java:717)
        at com.tencent.wedata.core.job.JobContainer.init(JobContainer.java:307)
        at com.tencent.wedata.core.job.JobContainer.start(JobContainer.java:114)
        at com.tencent.wedata.core.Engine.start(Engine.java:232)
        at com.tencent.wedata.core.Engine.entry(Engine.java:122)
        at com.tencent.teg.scheduler.runner.runners.wedata.wedataHelper.startwedata(wedataHelper.java:126)
        at com.tencent.teg.scheduler.runner.runners.WeDataDiRunner.execute(WeDataDiRunner.java:161)
        at com.tencent.cloud.wedata.runner.common.api.AbstractTaskRunner.executeWithCommonLog(AbstractTaskRunner.java:508)
        at com.tencent.cloud.wedata.runner.common.api.AbstractTaskRunner.startWork(AbstractTaskRunner.java:546)
        at com.tencent.teg.scheduler.runner.runners.WeDataDiRunner.main(WeDataDiRunner.java:75)
```

Cause:

©2013–2025 Tencent Cloud. All rights reserved.

The configuration file was not uploaded when setting up the data source. fs.defaultFS is part of core-site.xml.
Solutions:




```
</configuration>
[root@172 ~]# cat /usr/local/service/hadoop/etc/hadoop/core-site.xml | grep "fs.defaultFS"
        <name>fs.defaultFS</name>
[root@172 ~]#
```

Hive Script Demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
```

©2013–2025 Tencent Cloud. All rights reserved.

```
                    "password": "******",
                    "hadoopConfig": { // If Hadoop is a high availability cluster, you need to configure
NameNode high availability information
                        "dfs.namenode.https-address.mycluster.nn2": "ip:4009",
                        "dfs.namenode.https-address.mycluster.nn1": "ip:4009",
                        "dfs.ha.namenodes.mycluster": "nn1,nn2",
                        "dfs.namenode.servicerpc-address.mycluster.nn2": "ip:4010",
                        "dfs.namenode.servicerpc-address.mycluster.nn1": "ip:4010",
                        "dfs.namenode.rpc-address.mycluster.nn2": "ip:4007",
                        "dfs.namenode.rpc-address.mycluster.nn1": "ip:4007",
                        "dfs.namenode.http-address.mycluster.nn2": "ip:4008",
                        "dfs.namenode.http-address.mycluster.nn1": "ip:4008",
                        "dfs.client.failover.proxy.provider.mycluster":
"org.apache.hadoop.hdfs.server.namenode.ha.ConfiguredFailoverProxyProvider",
                        "dfs.nameservices": "mycluster",
                        "fs.defaultFS": "hdfs://mycluster",
                        "spark.eventLog.dir": "hdfs://mycluster/spark-history",
                    },
                    "column": [
                        "id",
                        "name"
                    ],
                    "where": "id>10", // Filters
                    "connection": [
                        {
                            "jdbcUrl": [ // Supports multiple Hive sources

"jdbc:hive2://ip:2181,ip:2181,ip:2181/database;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserve
r2;?mapred.job.name=null_null"
                            ],
                            "table": [ // Source table
                                "source_table"
                            ]
                        }
                    ],
                    "username": "hadoop"
                },
                "name": "hivereader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "postSql": [ // Post-SQL
                        ""
                    ],
                    "database": "database",
                    "fetchSize": 1024, // Batch submission size
                    "hadoopConfig": { // If Hadoop is a high availability cluster, you need to configure
NameNode high availability information
                        "dfs.namenode.https-address.mycluster.nn2": "ip:4009",
                        "dfs.namenode.https-address.mycluster.nn1": "ip:4009",
                        "dfs.ha.namenodes.mycluster": "nn1,nn2",
                        "dfs.namenode.servicerpc-address.mycluster.nn2": "ip:4010",
                        "dfs.namenode.servicerpc-address.mycluster.nn1": "ip:4010",
                        "dfs.namenode.rpc-address.mycluster.nn2": "ip:4007",
                        "dfs.namenode.rpc-address.mycluster.nn1": "ip:4007",
                        "dfs.namenode.http-address.mycluster.nn2": "ip:4008",
                        "dfs.namenode.http-address.mycluster.nn1": "ip:4008",
                        "dfs.client.failover.proxy.provider.mycluster":
"org.apache.hadoop.hdfs.server.namenode.ha.ConfiguredFailoverProxyProvider",
                        "dfs.nameservices": "mycluster",
```

©2013–2025 Tencent Cloud. All rights reserved.

```
                        "fs.defaultFS": "hdfs://mycluster",
                        "spark.eventLog.dir": "hdfs://mycluster/spark-history",
                    },
                    "column": [ //column name
                        {
                            "name": "id",
                            "type": "int"
                        },{
                            "name": "name",
                            "type": "varchar"
                        }
                    ],
                    "emptyAsNull": true, // Empty string processing
                    "writeMode": "overwrite", // Write mode
                    "metastoreUris": "thrift://ip:7004",
                    "table": "sink_table", // Target table
                    "preSql": [ // Pre-SQL
                        ""
                    ]
                },
                "name": "hivewriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

HDFS Data Source

Last updated：2024-12-12 19:06:55

Supported Editions

Supports HDFS 2.x, 3.x versions.

Use Limits

Offline Read

HDFS Reader supports the following features:

- Supports files in TXT, ORC, PARQUET, CSV, and JSON formats, and requires the file content to be a logically meaningful two-dimensional table.
- Supports various data types (represented as String), column pruning, and column constants.
- Only supports * and ? as file wildcards. Specifying wildcards will traverse multiple file information.
- Multiple files can support concurrent reading.
- Currently, the plugin uses Hive version 2.3.7 and Hadoop version 3.2.3.

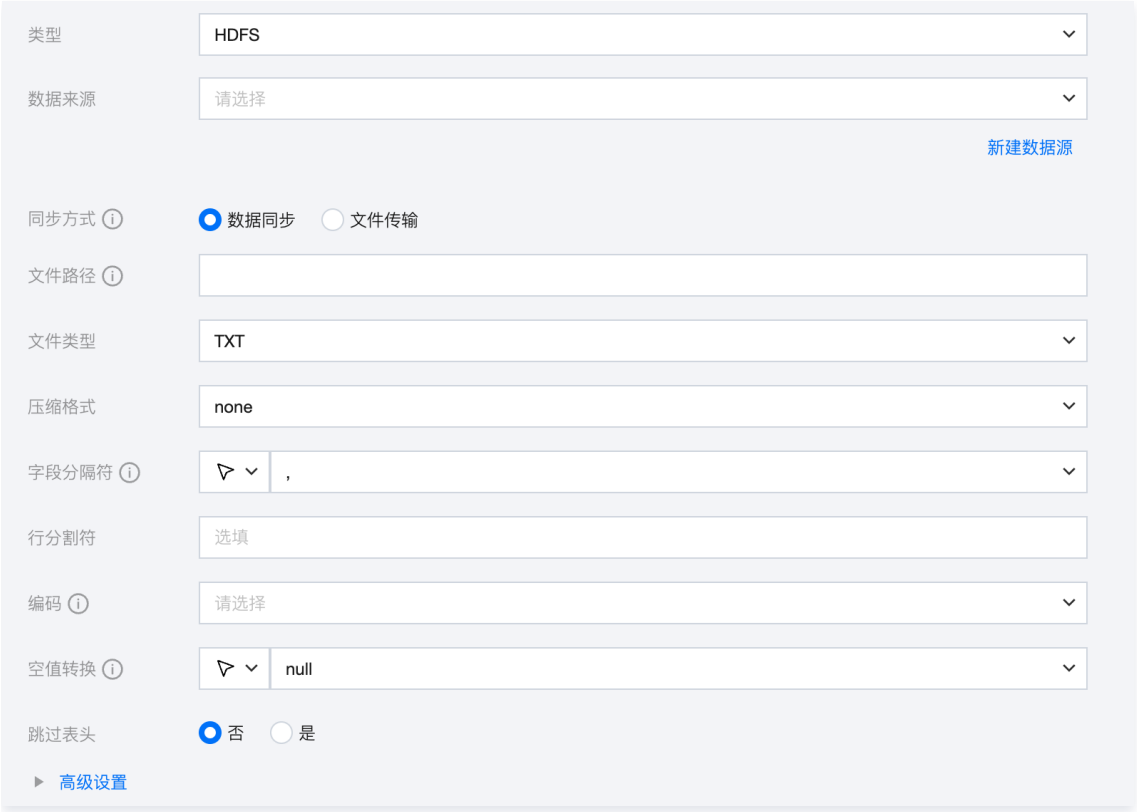
> **Note**
> HDFS Reader does not support multi-threaded concurrent reading of a single file due to internal file splitting algorithms.

Offline Write

When using HDFS Writer, please note the following:

1. Currently, HDFS Writer only supports four file formats: TXT, ORC, PARQUET, and CSV.
2. Since HDFS is a file system and does not have the concept of Schema, it does not support writing to partial columns.

HDFS Offline Single Table Read Node Configuration



Parameters	Description

©2013-2025 Tencent Cloud. All rights reserved.

Data Source	Select the available HDFS data source in the current project.
Synchronization Method	HDFS supports two synchronization methods: • Data Synchronization: Parses structured data content and maps and synchronizes data content based on field relationships. • File Transfer: Transfers the entire file without content parsing, suitable for unstructured data synchronization. **Note:** File transfer only supports data sources where both the source and destination are file types (COS/HDFS/SFTP/FTP), and the synchronization method for both must be file transfer.
File path	Path information of the file system. The path supports using '*' as a wildcard. Specifying the wildcard will traverse multiple files. For example, specifying / will read all files in the / directory, specifying /bazhen/ will read all files in the bazhen directory. HDFS currently supports * and ? as file wildcards, with syntax similar to typical Linux command line file wildcards.
File Type	HDFS supports five file types: TXT, ORC, PARQUET, CSV, and JSON. • TXT: Represents the TextFile file format. • ORC: Represents the ORCFile file format. • PARQUET: Represents the standard Parquet file format. • CSV: Represents the standard HDFS file format (logical two-dimensional table). • JSON: Represents the JSON file format.
Compression Format	Currently supported: none, deflate, gzip, bzip2, lz4, snappy. Since there is no unified stream format for snappy yet, DataInLong currently only supports the most mainstream hadoop-snappy (snappy stream format on hadoop) and framing-snappy (snappy stream format recommended by Google).
Field delimiter	Field delimiter for reading. When HDFS reads TXT or CVS data, a field delimiter needs to be specified. If not specified, the default is a comma (,). • Other available delimiters: '\t', '\u001', '|', 'space', ';', ',', '. • If you want to treat each line as a column at the destination, please use a character that does not appear in the line content as a separator. For example, an invisible character \u0001. **Note** Only TXT and CSV file formats are supported.
Row Separator (optional)	Linux default is \n, Windows default is \r\n. Manual entry supports a single character as the row separator. **Note** Only TXT and CSV file formats are supported.
Encoding	Encoding configuration for reading the file. Supports utf8 and gbk.
Null Value Conversion	• When reading, the specified string will be converted to NULL. • Support dropdown selection or manual input. Dropdown options support: empty string, space, \n, \0, null
Skip table header	No: The header is not included when reading. Yes: The header is included when reading. **Note** • Only TXT and CSV file formats are supported. • If the source includes a header and you choose not to skip it, it is recommended that the destination does not include the header to avoid duplicate header data.

©2013-2025 Tencent Cloud. All rights reserved.

HDFS Offline Single Table Write Node Configuration



Parameters	Description
Data Flow Direction	Select the available HDFS data source in the current project.
Synchronization Method	HDFS supports two synchronization methods: • Data Synchronization: Parses structured data content and maps and synchronizes data content based on field relationships. • File Transfer: Transfers the entire file without content parsing, suitable for unstructured data synchronization. Note: File transfer only supports data sources where both the source and destination are file types (COS/HDFS/SFTP/FTP), and the synchronization method for both must be file transfer.
File path	Path information of the file system. The path supports using ‘*’ as a wildcard, specifying the wildcard will traverse multiple file information.
File Name	Optional. The default file name is part.
Write mode	HDFS supports three write modes: • append: No processing before writing, directly use the filename to write, ensuring no file name conflict. • nonConflict: Error when file name conflicts. • overwrite: Clear all files with the same filename prefix before writing. For example, if "fileName": "abc", all files starting with abc in the corresponding directory will be cleared.

©2013–2025 Tencent Cloud. All rights reserved.

File Type	HDFS supports four file types: TXT, ORC, PARQUET, and CSV. • TXT: Represents the TextFile file format. • ORC: Represents the ORCFile file format. • PARQUET: Represents the standard Parquet file format. • CSV: Represents the standard HDFS file format (logical two-dimensional table).
Compression Format	Currently supported: none, deflate, gzip, bzip2, lz4, snappy. Since there is no unified stream format for snappy yet, DataInLong currently only supports the most mainstream hadoop-snappy (snappy stream format on hadoop) and framing-snappy (snappy stream format recommended by Google).
Field delimiter	Field separator for writing to HDFS, you need to ensure it matches the field separator of the created HDFS table. Otherwise, data cannot be found in the HDFS table. Options: '\t', '\u001', '|', ' space ', ',', ';', '\u005E\u0001\u005E'. **Note** Only TXT and CSV file formats are supported.
Row Separator (optional)	Linux default is \n, Windows default is \r\n. Manual entry supports a single character as the row separator. **Note** Only TXT and CSV file formats are supported.
Empty string processing	• No processing: Do not process empty strings during writing. • Process as NULL: When writing, empty strings are processed as NULL.
Null Value Conversion	When writing, convert NULL to a specified string. NULL represents unknown or not applicable values, different from 0, empty string, or other values. **Note** Only TXT and CSV file formats are supported for selection.
Include header	No: The header is not included when writing. Yes: The header is included when writing. **Note** • Only TXT and CSV file formats are supported. • If the source includes a header and you choose not to skip it, it is recommended that the destination does not include the header to avoid duplicate header data.
Mark the completion file.	A blank .ok file indicates the task transfer is complete. • No generation: After completion of the task, a blank .ok file is not generated (default selected). • Generation: After completion of the task, a blank .ok file is generated.
Mark completion filename	When filling in, include the complete file path, name, and suffix. By default, the .ok file is generated in the target path according to the target file name and suffix. Supports using time scheduling parameters such as ${yyyyMMdd} in data synchronization and file transfer modes. Support using built-in variable ${filetrans_di_src} in file transfer mode to represent the source file name. For example: /user/test/${filetrans_di_src}_${yyyyMMdd}.ok. **Note:** In file transfer mode, if the target file name **Target file name** uses ${filetrans_di_src} to reference the source file name, and if the completion marker files are generated according to the target file name

©2013-2025 Tencent Cloud. All rights reserved.

	and its suffix as .ok files, the number of generated completion marker files depends on the number of files at the source.
Advanced Settings (optional)	You can configure parameters based on business requirements.

Data type conversion is supported

Read

HDFS supported data types and conversion mappings are as follows (When processing HDFS, the data types from the HDFS data source will be first mapped to the data types in the data processing engine):

HDFS (Hive table) data types	Internal Type
TINYINT，SMALLINT，INT，BIGINT	LONG
FLOAT，DOUBLE	DOUBLE
STRING，CHAR，VARCHAR，STRUCT，MAP，ARRAY，UNION，BINARY	STRING
BOOLEAN	BOOLEAN
DATE，TIMESTAMP	DATE

Write

HDFS supported data types and conversion mappings are as follows:

Internal Type	HDFS (Hive table) data types
LONG	TINYINT，SMALLINT，INT，BIGINT
DOUBLE	FLOAT，DOUBLE
STRING	STRING，CHAR，VARCHAR，STRUCT，MAP，ARRAY，UNION，BINARY
BOOLEAN	BOOLEAN
DATE	DATE，TIMESTAMP

HDFS script demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "path": "/path/source.txt", // File path
                    "nullFormat": "null", // Null Value Conversion.
                    "hadoopConfig": { // If Hadoop is a high availability cluster, you need to configure
NameNode high availability information.
                        "dfs.namenode.https-address.mycluster.nn2": "ip:4009",
                        "dfs.namenode.https-address.mycluster.nn1": "ip:4009",
                        "dfs.nameservices": "mycluster",
                        "dfs.ha.namenodes.mycluster": "nn1,nn2",
                        "dfs.namenode.servicerpc-address.mycluster.nn2": "ip:4010",
                        "dfs.namenode.servicerpc-address.mycluster.nn1": "ip:4010",
```

©2013–2025 Tencent Cloud. All rights reserved.

```
                        "fs.defaultFS": "hdfs://mycluster",
                        "dfs.namenode.rpc-address.mycluster.nn2": "ip:4007",
                        "dfs.namenode.rpc-address.mycluster.nn1": "ip:4007",
                        "dfs.namenode.http-address.mycluster.nn2": "ip:4008",
                        "dfs.namenode.http-address.mycluster.nn1": "ip:4008",
                        "dfs.client.failover.proxy.provider.mycluster":
"org.apache.hadoop.hdfs.server.namenode.ha.ConfiguredFailoverProxyProvider",
                    },
                    "column": [
                        {
                            "name": "null",
                            "index": 0,
                            "type": "string"
                        }
                    ],
                    "defaultFS": "hdfs://ip:4007",
                    "skipHeader": "false", //Skip Header
                    "fieldDelimiter": ",", //Field Delimiter
                    "encoding": "utf-8", //Encoding method
                    "fileType": "text" //File type
                },
                "name": "hdfsreader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "path": "/path/sink", //File path
                    "fileName": "sink_table", //File name
                    "nullFormat": "null", // Null Value Conversion.
                    "hadoopConfig": { // If Hadoop is a high availability cluster, you need to configure
NameNode high availability information.
                        "dfs.namenode.https-address.mycluster.nn2": "ip:4009",
                        "dfs.namenode.https-address.mycluster.nn1": "ip:4009",
                        "dfs.nameservices": "mycluster",
                        "dfs.ha.namenodes.mycluster": "nn1,nn2",
                        "dfs.namenode.servicerpc-address.mycluster.nn2": "ip:4010",
                        "dfs.namenode.servicerpc-address.mycluster.nn1": "ip:4010",
                        "fs.defaultFS": "hdfs://mycluster",
                        "dfs.namenode.rpc-address.mycluster.nn2": "ip:4007",
                        "dfs.namenode.rpc-address.mycluster.nn1": "ip:4007",
                        "dfs.namenode.http-address.mycluster.nn2": "ip:4008",
                        "dfs.namenode.http-address.mycluster.nn1": "ip:4008",
                        "dfs.client.failover.proxy.provider.mycluster":
"org.apache.hadoop.hdfs.server.namenode.ha.ConfiguredFailoverProxyProvider",
                    },
                    "column": [ //column name
                        {
                            "name": "1",
                            "type": "string"
                        }
                    ],
                    "defaultFS": "hdfs://ip:4007",
                    "emptyAsNull": false, //Whether to process empty strings, if processed, empty strings
will be converted to null
                    "writeMode": "append", //Write mode
                    "suffix": "txt", //File suffix
                    "fieldDelimiter": ",", //Field Delimiter
                    "encoding": "utf-8", //Encoding method
                    "fileType": "text" //File type
                },
                "name": "hdfswriter"
```

©2013–2025 Tencent Cloud. All rights reserved.

```
            }
        }
    ],
    "setting": {
        "errorLimit": {
            "record": 0 // Dirty data threshold
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

HBase Data Source

Last updated：2024-12-12 19:07:29

Supported Editions

Supports HBase 2.2.x, 2.3.x versions.

Use Limits

1. HBase Reader does not support reading data written by Phoenix, as Phoenix has special processing.
2. HBase Writer supports concatenating multiple source fields as the HBase table rowkey.
3. Timestamp (version) support for writing to HBase:
   - Use the current time as the version.
   - Specify the source column as the version.
   - Specify a time as the version.

HBase offline single table reading node configuration

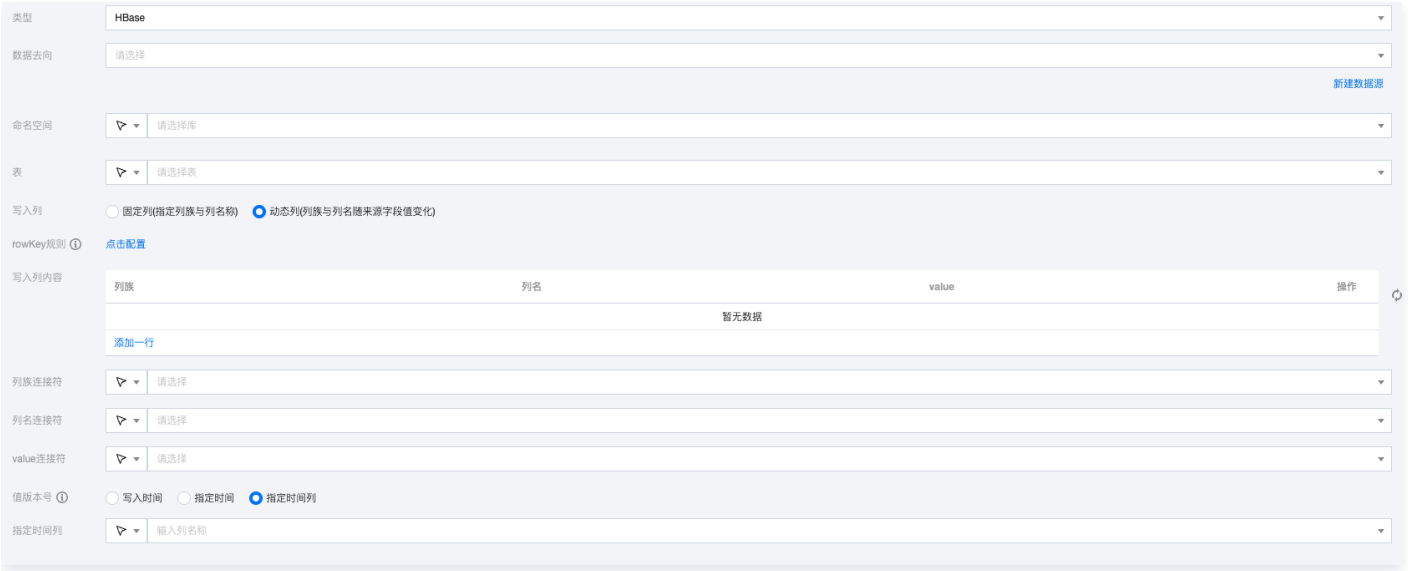


Parameters	Description
Data Source	Available HBase data source
Namespace	Select the available space under this data source
Table	Support selecting or manually entering the name of the table to be read
Read Mode	Supports horizontal table reading mode, treating the HBase table as a regular two-dimensional table (horizontal table) and reading the latest version of the data Supports vertical table reading mode, treating the HBase table as a vertical table
maxVersion	You need to specify the number of versions to read, which can only be -1 or a number greater than 1, where -1 means reading all versions. Default is -1 if not specified **Note** When the reading mode is vertical table, maxVersion needs to be configured
Encoding	Encoding configuration for reading the file. Supports utf8 and gbk.
Start RowKey	You need to specify the start RowKey.
Stop RowKey	You need to specify the end RowKey.

©2013-2025 Tencent Cloud. All rights reserved.

Scan Cache Size	Supports specifying the number of rows read from HBase each time, with a default of 256 if not specified.
Scan Batch Size	Supports specifying the number of columns read from HBase each time, with a default of 100 if not specified.

HBase offline single table write node configuration



Parameters	Description
Data Flow Direction	HBase data source to be written.
Namespace	Support selection or manual input of the space to be written.
Table	Support selection or manual input of the table name to be written: When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Write Column	HBase supports two ways to write columns: • Fixed Column (specify column family and column name). • Dynamic Column (column family and column name change with the source field value). You need to manually configure the column content, column family separator, column name separator, and value separator (optional: ' \u001 ', ' | ', ' : ', ', ').
rowKey rules	Configure HBase table rowKey generation rules
Write column content	When selecting a dynamic column (column family and column name change with the source field value), you need to manually configure the column content, column family separator, column name separator, and value separator (optional: ' \u001 ', ' | ', ' :', ', ').
Value Version Number	Specify the timestamp for writing to HBase. Support current time, specified time column, or specified time (choose one of the three). If not configured, the current time is used: • index: Specify the index corresponding to the Reader-side column, starting from 0. Ensure it can be converted to LONG. • type: If it is of Date type, it will try to parse using yyyy-MM-dd HH:mm:ss and yyyy-MM-dd HH:mm:ssSss formats. If it is a specified time, then the index is -1. • value: Value of the specified time, of LONG type.

Data type conversion is supported

Read

©2013-2025 Tencent Cloud. All rights reserved.

The supported data types and corresponding conversions for HBase reading are as follows (when processing HBase, the data types of the HBase data source will first be mapped to the data types of the data processing engine):

HBase Data Type	Internal Type
int，short，long	Long
float，double	Double
string，binary string	String
date	Date
boolean	Boolean

Write

The supported data types and corresponding conversions for HBase reading are as follows:

Internal Type	HBase Data Type
Long	int，short，long
Double	float，double
String	string，binary string
Date	date
Boolean	boolean

HBase script demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "mode": "normal", // Read mode, divided into horizontal table read and vertical table
read
                    "hadoopConfig": {
                        // Refer to high availability configuration information in the Hive data source
                    },
                    "column": [ //column name
                        {
                            "name": "info:id",
                            "type": "int"
                        },
                        {
                            "name": "info:name",
                            "type": "string"
                        }
                    ],
                    "range": { // Synchronization range, starting rowKey and ending rowKey
                        "endRowkey": "",
                        "startRowkey": ""
                    },
                    "encoding": "utf-8", //Encoding method
                    "table": "default:source_table", // Source table
                    "hbaseConfig": {
```

©2013–2025 Tencent Cloud. All rights reserved.

```
                        "hbase.zookeeper.quorum": "ip:2181,ip:2181,ip:2181",
                        "zookeeper.znode.parent": "/hbase-emr-******"
                    }
                },
                "name": "hbase11xreader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "mode": "normal",
                    "rowkeyColumn": [ // rowKey rules, refer to form mode rowKey generation rules
                        {
                            "func": "random",
                            "index": -1,
                            "type": "string",
                            "value": "2"
                        },
                        {
                            "index": 0,
                            "type": "string"
                        }
                    ],
                    "hadoopConfig": {
                        // Refer to high availability configuration information in the Hive data source
                    },
                    "column": [ //column name
                        {
                            "name": "info:id",
                            "index": "0",
                            "type": "string"
                        },
                        {
                            "name": "info:name",
                            "index": "1",
                            "type": "string"
                        }
                    ],
                    "encoding": "utf-8", //Encoding method
                    "table": "default:sink_table", // Target table
                    "hbaseConfig": {
                        "hbase.zookeeper.quorum": "ip:2181,ip:2181,ip:2181",
                        "zookeeper.znode.parent": "/hbase-emr-******"
                    }
                },
                "name": "hbase11xwriter"
            }
        }
    ],
    "setting": {
        "errorLimit": {
            "record": 0 // Dirty data threshold
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

TBase Data Source

Last updated: 2024-12-12 19:08:01

©2013-2025 Tencent Cloud. All rights reserved.

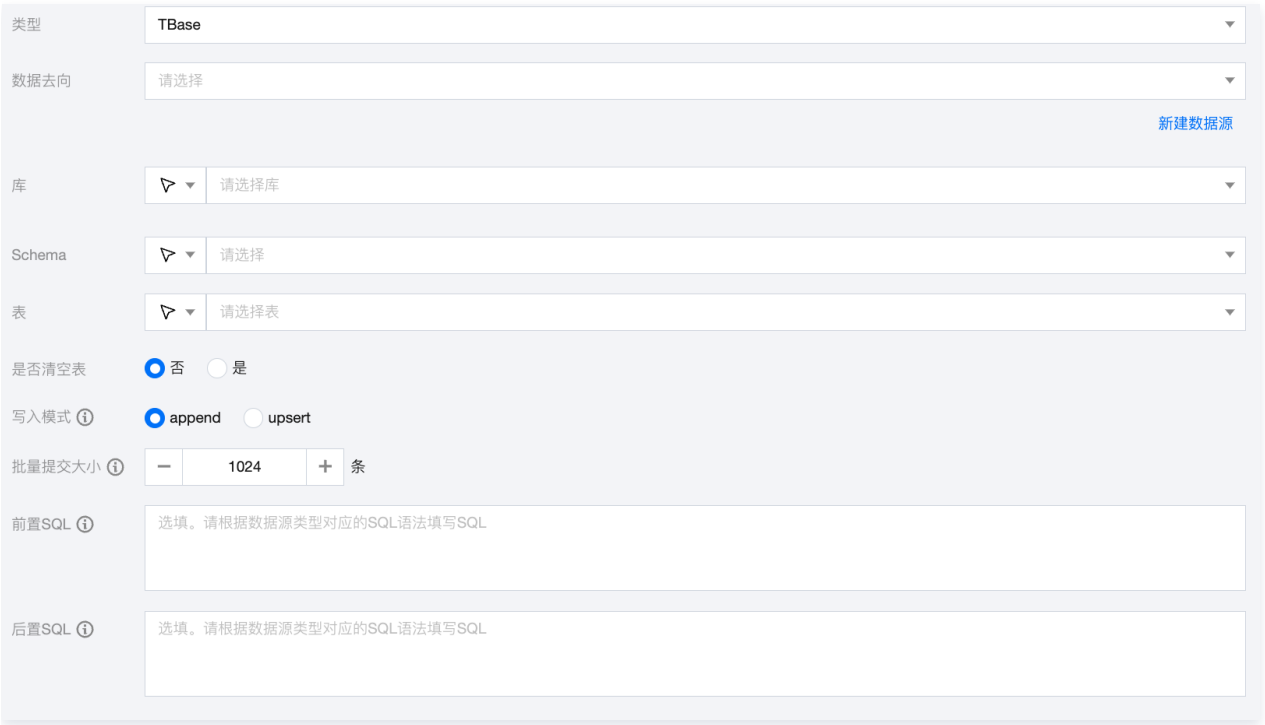
TBase Offline Single Table Read Node Configuration



Parameters	Description
Data Source	Available TBase Data Source.
Database	Support selecting or manually entering the name of the database to be read • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Support selecting or manually entering the Schema Name to be read.
Table	Supports selecting or manually inputting the table name to be read. Multiple tables can be read simultaneously. When configuring multiple tables, ensure the schema structure is consistent across tables.
Split Key	When TBase performs data extraction, specifying splitPk means you want to use the field represented by splitPk for data sharding. Consequently, data synchronization will start concurrent tasks to improve synchronization efficiency. • It is recommended to use the primary key as splitPk, because the primary key is usually evenly distributed, making the resulting shards less prone to data hotspots. • Currently, splitPk only supports integer data slicing; it does not support string, float, date, or other types. If you specify other unsupported types, the splitPk feature will be ignored, and synchronization will use a single channel. • If the splitPk value is set to empty, it will be considered that you do not allow sharding on a single table, so a single channel will be used for extraction.
Filter Condition (Optional)	Fill in the corresponding filter statement based on the data type. This statement will act as the filter condition for the data to be synchronized. TBase will concatenate the SQL based on the specified WHERE condition and extract data according to that SQL. For example, during testing, you can specify the WHERE condition as limit 10. In actual business scenarios, you would typically choose to synchronize the data for the current day, specifying the WHERE condition as gmt_create > $bizdate. • The WHERE condition can effectively perform business incremental synchronization. • If the WHERE condition is empty, it is considered to synchronize all information from the entire table.

TBase Offline Single Table Write Node Configuration

©2013–2025 Tencent Cloud. All rights reserved.



Parameters	Description
Data Source	TBase data source to be written to.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Support selecting or manually entering the Schema Name to be read.
Table	Supports selection, or manual input of the table name to be written to • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Clear Table	Before writing to this TBase data table, you can manually choose whether to clear the table.
Write mode	TBase supports two write modes: • append: Append write, errors on primary key conflict. • upsert: Update write, updates data on primary key conflict.
Batch Submission Size	The record size for one-time batch submission. This value can greatly reduce the number of network interactions between the data synchronization system and TBase, and improve overall throughput. If this value is set too high, it may cause OOM exceptions in the data synchronization process.
Pre-SQL (optional)	SQL command executed before the synchronization task. Fill in the SQL according to the correct SQL syntax corresponding to the data source type. For example, clear old data in the table before execution (truncate table tablename).
Post-SQL (optional)	SQL statements executed after the synchronization task are written according to the correct SQL syntax corresponding to the data source type, for example, adding a timestamp with `ALTER TABLE tablename ADD colname TIMESTAMP DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP`.

Data type conversion is supported

©2013-2025 Tencent Cloud. All rights reserved.

The supported data types and their corresponding conversions for TBase are as follows (when processing TBase, the data types between the TBase data source and the data processing engine will be mapped first):

Write

Internal Type	TBase Data Type
Long	INTEGER、SMALLINT、BIGINT
Double	DECIMAL、FLOAT、DOUBLE、REAL、NUMERIC
String	CHAR、VARCHAR、ARRAY
Date	DATE、TIMESTAMP、TIME
Boolean	BOOL

TBase script demo

If you configure an offline task using **Script Mode**, you need to write the **reader parameter** and **writer parameter** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "password": "******",
                    "column": [//column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": [
                                "jdbc:postgresql://ip:9000/database?reWriteBatchedInserts=true"
                            ],
                            "table": [ // Source table
                                "schema.source_table"
                            ]
                        }
                    ],
                    "where": "id>10", // Filters
                    "splitPk": "id", // Split key
                    "username": "root"
                },
                "name": "postgresqlreader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "postSql": [ // Post-SQL
                        ""
                    ],
                    "password": "******",
                    "pkColumn": "id", // when write mode is upsert, the primary key needs to be specified
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
```

©2013–2025 Tencent Cloud. All rights reserved.

```
                            "jdbcUrl": "jdbc:postgresql://ip:9000/database?reWriteBatchedInserts=true",
                            "table": [
                                "schema.sink_table"
                            ]
                        }
                    ],
                    "writeMode": "upsert", // write mode
                    "batchSize": 1024, // Batch submission size
                    "username": "root",
                    "preSql": [ // Pre-SQL
                        ""
                    ]
                },
                "name": "tbasewriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

GBase Data Source

Last updated：2024-12-12 19:08:28

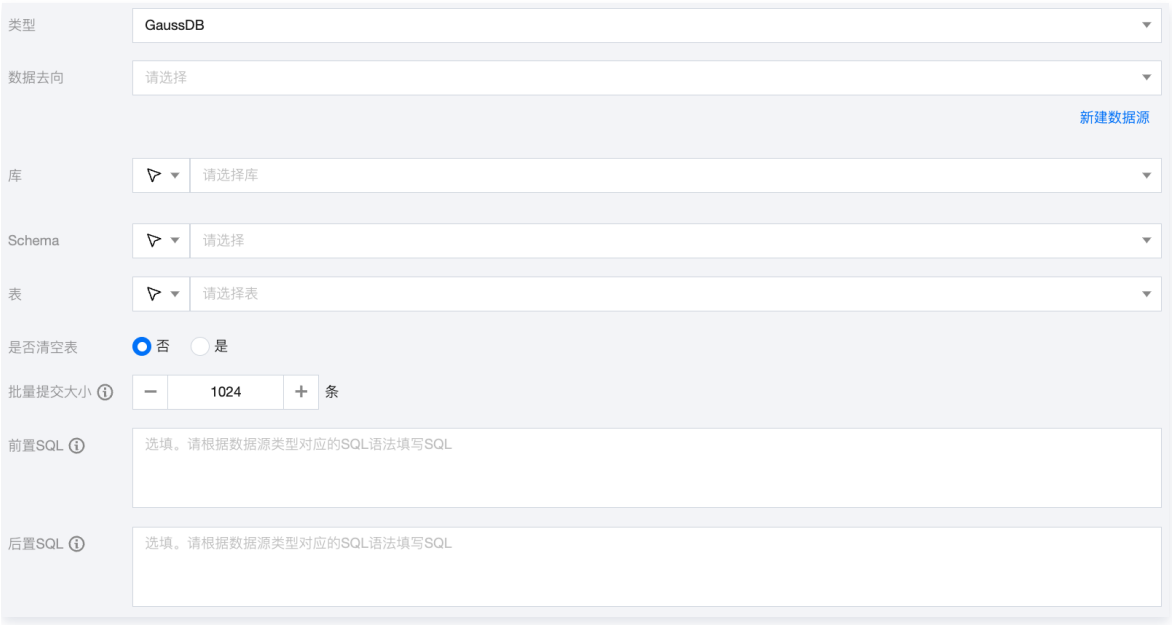
GBase Offline Single Table Read Node Configuration



Parameters	Description
Data Source	Available GBase Data Source.
Database	Support selection or manual input of the library name to be read. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Support selecting or manually entering the Schema Name to be read.
Table	Supports selection or manual input of the table name to be read. Multiple tables can be read simultaneously. When configured for multiple tables, you need to ensure the schema structures of the tables are consistent. GBase does not check the logical consistency of the tables.
Split Key	When extracting data from GBase, if splitPk is specified, it means you want to use the field represented by splitPk for data sharding. Data synchronization will therefore start concurrent tasks to improve the efficiency of data synchronization. • It is recommended to use the primary key as splitPk, because the primary key is usually evenly distributed, making the resulting shards less prone to data hotspots. • Currently, splitPk only supports integer data slicing; it does not support string, float, date, or other types. If you specify other unsupported types, the splitPk feature will be ignored, and synchronization will use a single channel. • If the splitPk value is set to empty, it will be considered that you do not allow sharding on a single table, so a single channel will be used for extraction.
Filter Condition (Optional)	Fill in the corresponding filter statement based on data types. This statement will serve as the filter condition for the data to be synchronized. GBase generates SQL by appending the specified WHERE condition and extracts data according to the SQL. For example, in testing, the WHERE condition can be specified as limit 10. In actual business scenarios, the typical choice is to synchronize the data of the current day by setting the WHERE condition to gmt_create > $bizdate. • The WHERE condition can effectively perform business incremental synchronization. • If the WHERE condition is empty, it is considered to synchronize all information from the entire table.

©2013-2025 Tencent Cloud. All rights reserved.

GBase Offline Single Table Write Node Configuration



Parameters	Description
Data Flow Direction	GBase Data Source to be written to.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Support selecting or manually entering the Schema Name to be read.
Table	Supports selection, or manual input of the table name to be written to • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong. • If the source table is of Oracle type, GBase supports one-click creation of the target table for writing purposes.
Clear Table	Before writing to the GBase data table, you can manually choose whether to clear the data table.
Batch Submission Size	The size of record count for batch submission. This value can significantly reduce the number of network interactions between the data synchronization system and GBase, improving overall throughput. If set too high, it may cause the data synchronization run process to encounter OOM Exception.
Pre-SQL (optional)	SQL command executed before the synchronization task. Fill in the SQL according to the correct SQL syntax corresponding to the data source type. For example, clear old data in the table before execution (truncate table tablename).
Post-SQL (optional)	SQL statements executed after the synchronization task are written according to the correct SQL syntax corresponding to the data source type, for example, adding a timestamp with `ALTER TABLE tablename ADD colname TIMESTAMP DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP`.

Data type conversion is supported

Read

©2013-2025 Tencent Cloud. All rights reserved.

GBase Data Types	Internal Type
TINYINT、SMALLINT、INTEGER、MEDIUMINT、BIGINT	Long
DECIMAL、FLOAT、DOUBLE、REAL、NUMERIC	Double
CHAR、VARCHAR、TINYTEXT、TEXT、MEDIUMTEXT、LONGTEXT	String
DATE、TEAR、DATETIME、TIMESTAMP	Date
TINYBLOB、MEDIUMBOLB、LONGBLOB	Bytes
BOOLEAN	Boolean

Write

Internal Type	GBase Data Types
Long	TINYINT、SMALLINT、INTEGER、MEDIUMINT、BIGINT
Double	DECIMAL、FLOAT、DOUBLE、NUMERIC、REAL
String	CHAR、VARCHAR、TINYTEXT、TEXT、MEDIUMTEXT、LONGTEXT
Date	DATE、YEAR、TIMESTAMP、DATETIME
Bytes	TINYBLOB、MEDIUMBOLB、LONGBLOB
BOOLEAN	Boolean

GBase Script Demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": [
                                "jdbc:gbase://ip:5258/database?rewriteBatchedStatements=true"
                            ],
                            "table": [ // Source table
                                "schema.source_table"
                            ]
                        }
                    ],
                    "where": "id>10", // Filters
                    "username": "root"
                },
                "name": "rdbmsreader"
            },
            "transformer": [],
```

©2013–2025 Tencent Cloud. All rights reserved.

```
            "writer": {
                "parameter": {
                    "postSql": [ // Post-SQL
                        ""
                    ],
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": "jdbc:gbase://ip:5258/database?rewriteBatchedStatements=true",
                            "table": [ // Target table
                                "schema.sink_table"
                            ]
                        }
                    ],
                    "batchSize": 1024, // Batch submission size
                    "username": "root",
                    "preSql": [ // Pre-SQL
                        ""
                    ]
                },
                "name": "rdbmswriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

GaussDB Data Source

Last updated：2024-12-12 19:08:54

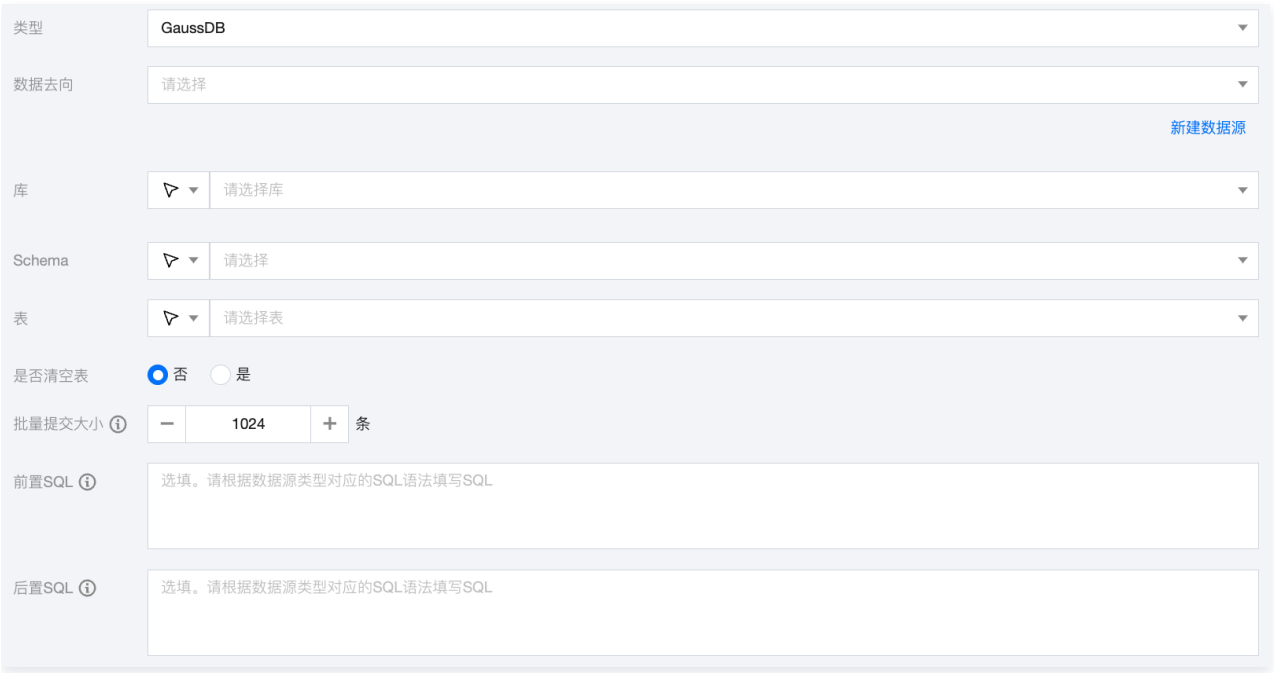
GaussDB Offline Single Table Read Node Configuration



Parameters	Description
Data Source	Available GaussDB Data Source.
Database	Support selecting or manually entering the name of the database to be read • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Support selecting or manually entering the Schema Name to be read.
Table	Support selecting or manually entering the Table Name to be read.
Split Key	Specify the field for data sharding. Once specified, concurrent tasks will be initiated for data synchronization. You can use a column from the source data table as the split key. It is recommended to use the primary key or an indexed column as the split key.
Filter Condition (Optional)	In actual business scenarios, it is common to synchronize the data of the current day by setting the where condition as gmt_create>$bizdate. • The WHERE condition can effectively perform business incremental synchronization. • If the where clause is not specified, including not providing the key or value for where, data synchronization will be considered as full data synchronization.

GaussDB Offline Single Table Write Node Configuration

©2013-2025 Tencent Cloud. All rights reserved.



Parameters	Description
Data Flow Direction	GaussDB data source to be written to.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Schema	Support selecting or manually entering the Schema Name to be read.
Table	Supports selection, or manual input of the table name to be written to • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Clear Table	You can manually choose whether to clear the GaussDB data table before writing to it.
Batch Submission Size	The batch submission record count can significantly reduce the number of network interactions between the data synchronization system and GaussDB, and improve overall throughput. If this value is set too high, it may cause an OOM (Out of Memory) exception in the data synchronization process.
Pre-SQL (optional)	SQL command executed before the synchronization task. Fill in the SQL according to the correct SQL syntax corresponding to the data source type. For example, clear old data in the table before execution (truncate table tablename).
Post-SQL (optional)	SQL statements executed after the synchronization task are written according to the correct SQL syntax corresponding to the data source type, for example, adding a timestamp with `ALTER TABLE tablename ADD colname TIMESTAMP DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP`.

Data type conversion is supported

The supported data types and conversion relationships for GaussDB are as follows (when processing GaussDB, the data types of the GaussDB data source and the data processing engine will first be mapped):

Read

©2013-2025 Tencent Cloud. All rights reserved.

GaussDB Data Type	Internal Type
bigint，bigserial，integer，smallint，serial	Long
double precision，money，numeric，real	Double
varchar，char，text，bit，inet	String
date，time，timestamp	Date
bool	Boolean
bytea	Bytes

Write

Internal Type	GaussDB Data Type
Long	bigint，bigserial，integer，smallint，serial
Double	double precision，money，numeric，real
String	varchar，char，text，bit
Date	date，time，timestamp
Boolean	bool
Bytes	bytes

GaussDB script demo

If you configure an offline task using **Script Mode**, you need to write the **reader parameter** and **writer parameter** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": [
                                "jdbc:postgresql://ip:5432/database?reWriteBatchedInserts=true"
                            ],
                            "table": [ // Source table
                                "schema.source_table"
                            ]
                        }
                    ],
                    "where": "id>10", // Filters
                    "splitPk": "id", // Split key
                    "username": "root"
                },
                "name": "postgresqlreader"
            },
            "transformer": [],
            "writer": {
```

©2013–2025 Tencent Cloud. All rights reserved.

```
                "parameter": {
                    "postSql": [ // Post-SQL
                        ""
                    ],
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": "jdbc:postgresql://ip:5432/database?reWriteBatchedInserts=true",
                            "table": [ // Target table
                                "schema.sink_table"
                            ]
                        }
                    ],
                    "writeMode": "upsert", // write mode
                    "batchSize": 1024, // Batch submission size
                    "username": "root",
                    "on": [ // unique index
                        "id"
                    ],
                    "preSql": [ // Pre-SQL
                        ""
                    ]
                },
                "name": "postgresqlwriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

Impala Data Source

Last updated：2024-12-12 19:09:22

Supported Editions

Supports Impala version 4.1.0.

Impala Offline Single Table Read Node Configuration



Parameters	Description
Data Source	Select the configured Impala data source from the source end.
Database	Support selection or manual input of the library name to be read. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Support selecting or manually entering the Table Name to be read. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Split Key	Specify the field for data sharding. Once specified, concurrent tasks will be initiated for data synchronization. You can use a column from the source data table as the split key. It is recommended to use the primary key or an indexed column as the split key.
Filter Condition (Optional)	Fill in the corresponding filter statement according to the data type. This statement will serve as the filter condition for the data to be synchronized.

Impala Offline Single Table Write Node Configuration

©2013-2025 Tencent Cloud. All rights reserved.



Parameters	Description
Data Flow Direction	Select the configured Impala data source from the target end.
Database	Support selection or manual input of the library name to be read. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Support selecting or manually entering the Table Name to be read. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Clear Table	Before writing to the Impala data table, you can manually choose whether to clear the data table.
Batch Submission Size	The size of the record count for batch submissions greatly reduces the number of network interactions between the data synchronization system and Impala, and enhances overall throughput. If set too large, it may cause OOM exceptions in the data synchronization runtime process.
Pre-SQL	SQL statements executed before the synchronization task, filled in according to the correct SQL syntax corresponding to the data source type.
Post-SQL	SQL statements executed after the synchronization task, filled in according to the correct SQL syntax corresponding to the data source type.

Data type conversion is supported

Read

Impala Data Type	Internal Type
BIGINT、INT、SMALLINT、TINYINT	Long
DECIMAL、DOUBLE、FLOAT、REAL	Double
CHAR、VARCHAR、ARRAY、STRUCT	String
TIMESTAMP	Date

©2013-2025 Tencent Cloud. All rights reserved.

BOOLEAN	Boolean

Write

Internal Type	Impala Data Type
Long	BIGINT、INT、SMALLINT、TINYINT
Double	DECIMAL、DOUBLE、FLOAT、REAL
String	CHAR、VARCHAR、ARRAY、STRUCT
Date	TIMESTAMP
Boolean	BOOLEAN

Impala Script Demo

If you configure an offline task using **Script Mode**, you need to write the **reader parameter** and **writer parameter** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "password": "******",
                    "column": [ //column name
                        "id",
                        "name"
                    ],
                    "connection": [
                        {
                            "jdbcUrl": [

"jdbc:impala://ip:21000/database;AuthMech=3;UID=hadoop;OptimizedInsert=0;UseSasl=0;PWD="
                            ],
                            "table": [ //source table
                                "source_table"
                            ]
                        }
                    ],
                    "where": "id>10", //filters
                    "splitFactor": 1,
                    "splitPk": "id", //split key
                    "username": "hadoop"
                },
                "name": "impalareader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "postSql": [ //post-SQL
                        ""
                    ],
                    "password": "******",
                    "column": [ //column name
                        "*"
                    ],
                    "connection": [
                        {
```

©2013–2025 Tencent Cloud. All rights reserved.

```
                            "jdbcUrl":
"jdbc:impala://ip:21000/database;AuthMech=3;UID=hadoop;OptimizedInsert=0;UseSasl=0;PWD=Wedata@2022",
                            "table": [ //target table
                                "sink_table"
                            ]
                        }
                    ],
                    "batchSize": 1024, // Batch submission size
                    "username": "hadoop",
                    "preSql": [ //pre-SQL
                        ""
                    ]
                },
                "name": "impalawriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

Kudu Data Source

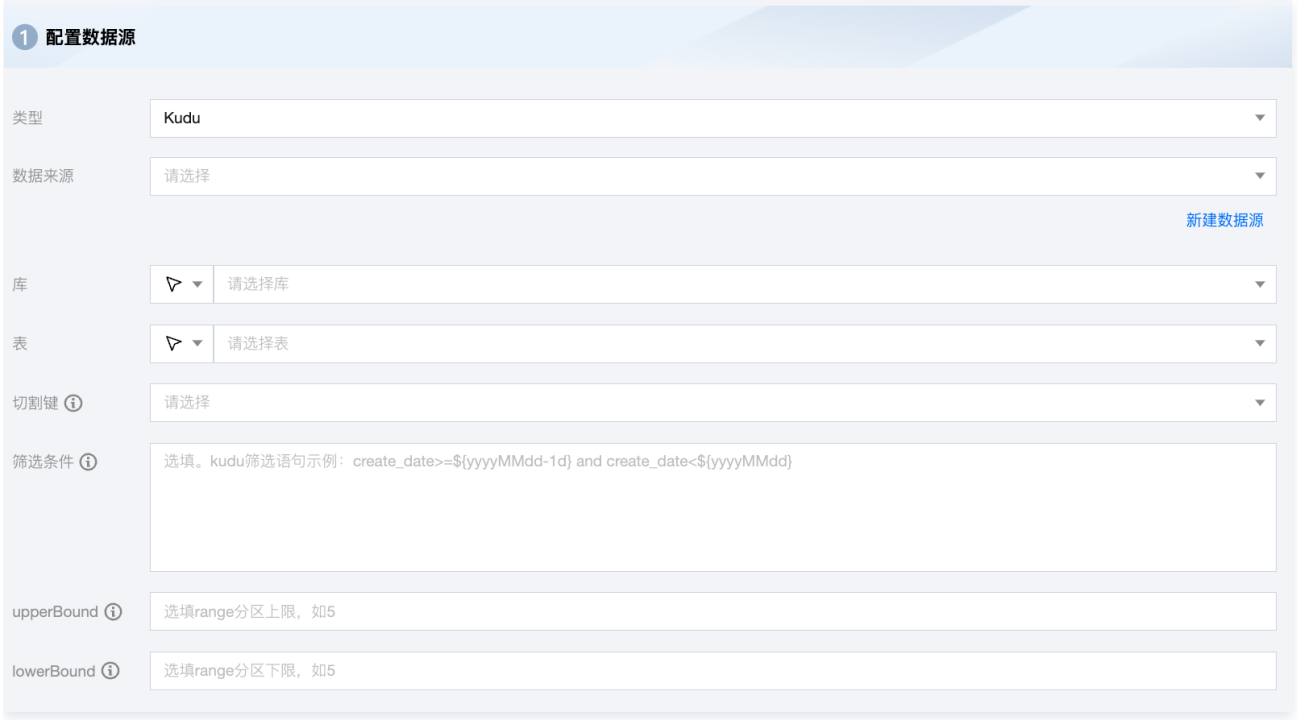
Last updated：2024-12-12 19:09:49

Use Limits

1. Kudu reader must configure upperBound and lowerBound for concurrency settings to be effective.
2. upperBound and lowerBound in kudu reader are of type long, so only time or integer fields support being set as Bound.
3. The reader uses kudu-client to directly connect to the Kudu server and read data. The where condition does not support Impala SQL syntax.
4. For incremental synchronization, the where condition syntax is:
   `create_time>= '${yyyy-MM-dd-1d HH:mm:ss}' and create_time < '${yyyy-MM-dd HH:mm:ss}'`
   Currently, Bound supports integer and date function configuration. The usage method for date function configuration:

```
// Convert to 13-bit timestamp (milliseconds)
TimestampMillis('yyyy-MM-ddTHH:mm:00+0800')
TimestampMillis('2023-07-10T00:00:00+0800')
TimestampMillis('2023-07-10 00:00:00')
TimestampMillis('2023-07-10')

// Convert to 10-bit timestamp (seconds)
TimestampSeconds('yyyy-MM-ddTHH:mm:00+0800')
TimestampSeconds('2023-07-10T00:00:00+0800')
TimestampSeconds('2023-07-10 00:00:00')
TimestampSeconds('2023-07-10')
```

Kudu offline single table read node configuration



Parameters	Description
Data Source	Available Kudu Data Source.
Database	Support selection or manual input of the library name to be read. • By default, the database bound to the data source is used as the default database; other databases need to be manually entered.

©2013-2025 Tencent Cloud. All rights reserved.

	• When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Support selecting or manually entering the Table Name to be read.
Split Key	Specify the field for data sharding. Once specified, concurrent tasks will be initiated for data synchronization. You can use a column from the source data table as the split key. It is recommended to use the primary key or an indexed column as the split key.
Filter Condition (Optional)	In actual business scenarios, the current day's data is usually selected for synchronization, with the where condition specified as gmt_create>$bizdate. The where condition can effectively perform incremental business synchronization. If the where statement is not filled, including not providing the key or value for where, data synchronization is considered as synchronizing all data.
upperBound	Partition upper limit. • If the sql create table statement partition "5"<= values <="10", then lowerBound is "5" and upperBound is "10"; • If the sql create table statement partition value ="x", then lowerBound is "x" and upperBound is "x\000";
lowerBound	Partition lower limit. • If the sql create table statement partition "5"<= values <="10", then lowerBound is "5" and upperBound is "10"; • If the sql create table statement partition value ="x", then lowerBound is "x" and upperBound is "x\000";

Kudu Offline Single-Table Write Node Configuration



Parameters	Description
Data Flow Direction	Kudu data source to be written.
Database	Supports selection, or manual input of the library name to be written to • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Table	Supports selection, or manual input of the table name to be written to • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.

©2013-2025 Tencent Cloud. All rights reserved.

Clear Table	Before writing to this Kudu data table, you can manually choose whether to clear the data table.
Write mode	Kudu write supports three modes: • Append: When a primary key/unique index conflict occurs, the conflicting row cannot be written • Overwrite: When a primary key/unique index conflict occurs, the original row will be deleted first, and then the new row will be inserted • On duplicate key: When a primary key/unique index conflict occurs, the new row will replace the specified fields of the statement
Batch Submission Size	The record count size for batch submissions can greatly reduce the number of network interactions between the data synchronization system and Kudu, thereby improving overall throughput. If the value is set too high, it will cause an OOM exception in the data synchronization process.

Data type conversion is supported

Read

Kudu Data Type	Internal Type
int8、int16、int32、int64	Long
float、double、decimal	Double
string、varchar	String
unixtime_micors、date	Date
binary	Bytes
bool	Boolean

Write

Internal Type	Kudu Data Type
Long	int8、int16、int32、int64
Double	float、double、decimal
String	string、date
Date	unixtime_micors、varchar
Bytes	binary
Boolean	bool

Kudu Script Demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "masterAddress": "ip:7051,ip:7051,ip:7051", //supports multiple ips
                    "upperBound": "10", //partition limit
                    "column": [ //column name
                        "id",
                        "name"
                    ],
```

©2013–2025 Tencent Cloud. All rights reserved.

```
                    "where": [ //filters
                        "id>10"
                    ],
                    "lowerBound": "5", //partition lower limit
                    "table": "database.source_table" //source table
                },
                "name": "kudureader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "kuduConfig": {
                        "kudu.master_addresses": "ip:7051,ip:7051,ip:7051" //supports multiple ips
                    },
                    "column": [ //column name
                        {
                            "name": "id",
                            "type": "int"
                        },
                        {
                            "name": "name",
                            "type": "string"
                        }
                    ],
                    "writeMode": "append", //Write mode
                    "batchSize": 1024, // Batch submission size
                    "encoding": "UTF-8", //encoding type
                    "table": "database.sink_table" //target table
                },
                "name": "kuduwriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

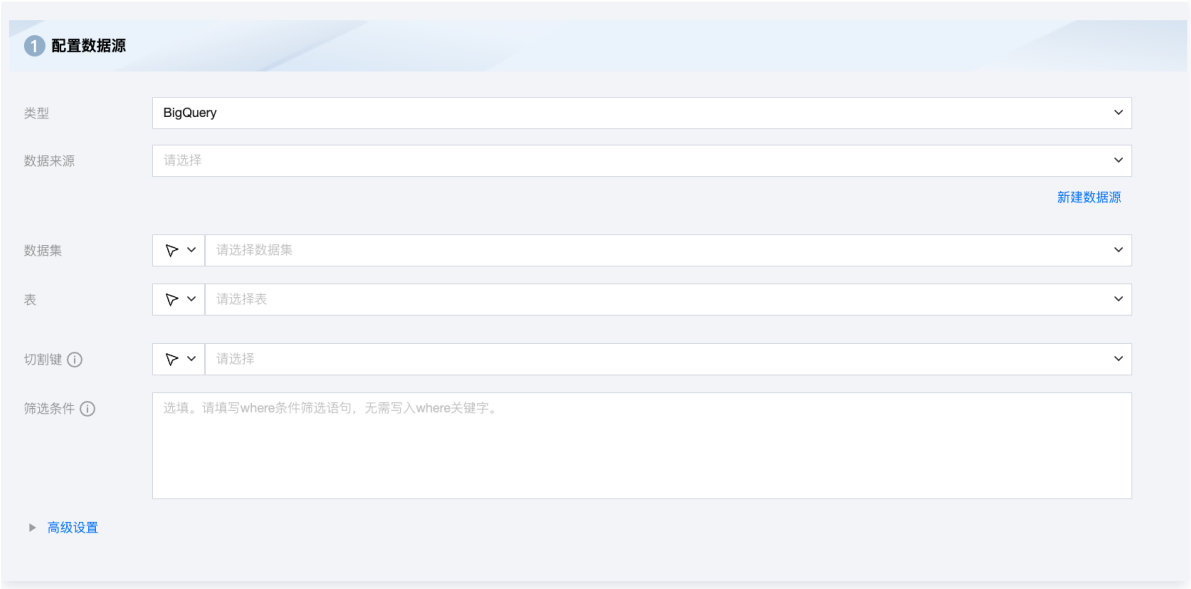
BigQuery Data Source

Last updated：2024-12-12 19:10:18

Use Limits

- BigQuery data source is only supported in Hong Kong (China), Singapore, usa (Virginia), and germany (Frankfurt). Other regions are not supported yet.
- The split key only supports INT64, FLOAT64, TIMESTAMP, BIGNUMERIC, NUMERIC, DATE, and DATETIME.

BigQuery offline single table read node configuration



Parameter	Description
Data Source	Select the configured BigQuery data source from the source end.
Data set	Support selecting or manually entering the dataset name to be read. • The database bound to the data source is set as the default library by default, other databases require manual input of the dataset name. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the dataset name. Data synchronization can still be performed given network connection in DataInLong.
Table	Support selecting or manually entering the Table Name to be read. • When the data source network is not connected, causing table information to be unavailable for direct fetching, you can manually enter the table name. Data synchronization can still be performed given network connection in DataInLong.
Split Key	Specify the field for data sharding. Once specified, concurrent tasks will be initiated for data synchronization. You can use a column from the source data table as the split key. It is recommended to use the primary key or an indexed column as the split key. ⚠ **Note:** • Supported split key types: INT64, FLOAT64, TIMESTAMP, BIGNUMERIC, NUMERIC, DATE, DATETIME. • If the number of concurrent tasks is set to greater than 1, this item is required; otherwise, the concurrency settings will be invalid.

©2013-2025 Tencent Cloud. All rights reserved.

Filter Condition (Optional)	Fill in the corresponding filter statement according to the data type. This statement will serve as the filter condition for the data to be synchronized.

Data type conversion is supported

Read

BigQuery data types.	Internal Type
BOOLEAN	BOOL
INTEGER	LONG
FLOAT,BIGNUMERIC,NUMERIC	DOUBLE
STRING,TIMESTAMP,DATE,TIME,GEOGRAPHY,JSON,INTERVAL,RECORD	STRING

BigQuery script demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** in the task script according to the unified script format requirements.

```
"job": {
        "content": [
            {
                "reader": {
                    "parameter": {
                        "databaseName": "source_database", //source database
                        "credential": { //bigquery data source configuration information
                            "token_uri": "",
                            "private_key_id": "",
                            "client_x509_cert_url": "",
                            "project_id": "",
                            "universe_domain": "",
                            "auth_uri": "",
                            "auth_provider_x509_cert_url": "",
                            "client_email": "",
                            "private_key": "",
                            "type": "service_account",
                            "client_id": ""
                        },
                        "column": [ //column name
                            "id",
                            "name"
                        ],
                        "projectId": "project_id",
                        "tableName": "source_table" // Source table
                    },
                    "name": "bigqueryreader"
                },
                "transformer": [],
            }
        ],
        "setting": {
            "memory": { // Memory settings
                "taskLimit": 335544320,
                "channelLimit": 67108864,
                "writerLimit": 134217728,
                "jobLimit": 872415232,
```

©2013–2025 Tencent Cloud. All rights reserved.

```
                "readerLimit": 134217728
            },
            "errorLimit": { // Dirty data threshold
                "record": 0
            },
            "speed": {
                "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed
in byte/s
                "channel": 1 // Concurrency limit
            }
        }
    }
```

©2013–2025 Tencent Cloud. All rights reserved.

FTP Data Source

Last updated: 2024-12-12 19:11:05

Use Limits

FTP Reader implements the feature of reading data from remote FTP files and converting it into the data synchronization protocol. Remote FTP files themselves are unstructured data storage. Currently, FTP Reader supports the following features:

Supported	Not supported.
• Supports, and only supports, reading TXT files, and requires the schema in TXT to be a two-dimensional table. • Supports CSV-like format files with custom delimiters. • Supports reading various types of data (represented as STRING), supports column pruning and column constants. • Supports recursive reading and filename filtering. • Supports text compression. Supports uncompressed, zip, gzip, bzip2. • Multiple files can support concurrent reading.	• Single file multi-threaded concurrent reading. This involves internal splitting algorithms for a single file. • Single file cannot technically support multi-threaded concurrent reading in a compressed state.

FTP Writer implements the feature of converting the DataInLong protocol into an FTP file. FTP files themselves are unstructured data storage. Currently, FTP Writer supports the following features:

Supported	Not supported.
• Supports, and only supports, writing text type (not BLOB, such as video data) files, and requires the schema in the text to be a two-dimensional table. • Supports CSV-like and TEXT format files with custom delimiters. • Text compression is not supported during writing. • Supports multi-threaded writing, with each thread writing to a different subfile.	• Single file cannot support concurrent writing. • FTP itself does not provide data types. FTP Writer writes all data to FTP files as STRING type.

FTP offline single-table read node configuration

©2013-2025 Tencent Cloud. All rights reserved.



Parameters	Description
Data Source	Select the available FTP data source in the current project.
Synchronization Method	FTP supports two synchronization methods: • Data Synchronization: Parses structured data content and maps and synchronizes data content based on field relationships. • File Transfer: Transfers the entire file without content parsing, suitable for unstructured data synchronization. **Note:** File transfer only supports data sources where both the source and destination are file types (COS/HDFS/SFTP/FTP), and the synchronization method for both must be file transfer.
File path	For the remote FTP file system path and filename information, you need to provide the complete file path and name with the path and file suffix included. Multiple paths can be specified here. • When specifying a single remote FTP file, FTP can currently use only a single thread for data extraction. In the future, multiple threads will be used to read single files concurrently in non-compressed file cases. • When specifying multiple remote FTP files, FTP supports using multiple threads for data extraction. The number of threads is specified by the number of channels. • When specifying wildcards, FTP attempts to list information for multiple files. For example, specifying / means reading all files in the directory. Specifying /bazhen/ means reading all files in the bazhen directory. Currently, FTP only supports the asterisk (*) as a file wildcard and supports using scheduling parameters for flexible file name and path configuration.
File Type	Supports two file types: TXT, CSV. • TXT: Represents the TextFile file format. • CSV: Represents the standard HDFS file format (logical two-dimensional table).

©2013-2025 Tencent Cloud. All rights reserved.

Field delimiter	Read Field Delimiter. When FTP reads data, a field delimiter needs to be specified. If not specified, it will default to (,), and the interface configuration will also default to (,).
Row Separator (optional)	Linux default is \n, Windows default is \r\n. Manual entry supports a single character as the row separator.
Encoding	Encoding configuration for reading the file. Supports utf8 and gbk.
Null Value Conversion	Convert the specified string to null during read.
Text compression type	Supports uncompressed, zip, gzip, bzip2
Skip table header	• No: Do not skip the header when reading. • Yes: Skip the header when reading.
Advanced Settings (optional)	You can configure parameters based on business requirements.

About the file path description:

- It is generally not recommended to use an asterisk (*), as it may easily cause JVM memory overflow errors during task execution.
- Data synchronization treats all Text Files synced under a job as a single data table. You must ensure that all Files can adapt to the same set of schema information.
- You must ensure that the files read are in a CSV-like format and grant readable permissions to the data synchronization system.
- If no matching files are found for extraction in the path specified by Path, the sync task will report an error.

FTP offline single-table write node configuration

©2013-2025 Tencent Cloud. All rights reserved.



Parameters	Description
Data Flow Direction	Select the available FTP data source in the current project.
Synchronization Method	FTP supports two synchronization methods: • Data Synchronization: Parses structured data content and maps and synchronizes data content based on field relationships. • File Transfer: Transfers the entire file without content parsing, suitable for unstructured data synchronization. **Note:** File transfer only supports data sources where both the source and destination are file types (COS/HDFS/SFTP/FTP), and the synchronization method for both must be file transfer.
File path	Path information of the file system. The path supports using ‘*’ as a wildcard, specifying the wildcard will traverse multiple file information.
File Name	The name of the written file. This file name will have a random suffix added as the actual write name.
Write mode	FTP supports three write modes: • append: Directly write using filename without any preprocessing to ensure no filename conflict. • nonConflict: Error when filename conflicts. • overwrite: Clears all files with the same filename prefix before writing.
File Type	Supports two file types: TXT, CSV. • TXT: Represents the TextFile file format. • CSV: Represents the standard HDFS file format (logical two-dimensional table).

©2013–2025 Tencent Cloud. All rights reserved.

Field delimiter	Field separator for writing. The field separator for FTP writing needs to be consistent with the field separator of the created FTP table; otherwise, data cannot be found in the FTP table. Options: '\t', '\u001', '|', 'space', ';', ','.
Row Separator (optional)	Linux default is \n, Windows default is \r\n. Manual entry supports a single character as the row separator.
Encoding	Encoding configuration for writing files. Supports utf8 and gbk encodings.
Null Value Conversion	When writing, convert null to the specified string.
Include header	• No: Do not skip the header when writing. • Yes: Skip the header when writing.
Mark the completion file.	A blank .ok file indicates task completion. • Do not generate: No blank .ok file will be created after task completion. (Selected by default) • Generation: After completion of the task, a blank .ok file is generated.
Mark completion filename	When filling in, include the complete file path, name, and suffix. By default, the .ok file is generated in the target path according to the target file name and suffix. • Supports the use of time scheduling parameters like ${yyyyMMdd} in data synchronization and file transfer modes. • Supports using the built-in variable ${filetrans_di_src} in the file transfer mode to represent the source file name. • For example, /user/test/${filetrans_di_src}_${yyyyMMdd}.ok. **Note:** In the file transfer mode **Target File Name**, if you reference the source file name using ${filetrans_di_src} and generate the .ok file according to the target file name and suffix when marking the completion file, the number of generated completion files depends on the number of files from the source.
Advanced Settings (optional)	You can configure parameters based on business requirements.

Data type conversion is supported

FTP implements a bidirectional channel feature for reading and writing FTP. Remote FTP files themselves are unstructured data storage, and the data processing engine automatically converts them to Bytes type during reading and writing.

FAQs

1. ftp write task error: Please ensure... has directory ls permissions, errorMessage: connect timed out

Cause:

- Integrated connection to FTP currently supports only passive mode. This might be because the passive mode is not enabled in the FTP service configuration.
- The client connects to the FTP server's port 21 and sends the username and password to log in. After successfully logging in, another port (above 1024) is used for listing directories or reading data. The data transfer port might not be open.

Solutions:

Check FTP server configuration:

1. Is passive mode enabled?

```
# Configuration file
pasv_enable=YES # Enable passive mode
pasv_min_port=${number}  # Minimum port for passive mode
pasv_max_port=${number}  # Maximum port for passive mode
```

©2013–2025 Tencent Cloud. All rights reserved.

2. Make sure the server has opened ports in the specified range.

Install lftp command in the pod:

```
# Copy the APK package to the pod:
cd /data/home/ryanrliao
kube ${resource_group}
kubectl cp ftp/lftp-4.8.3-r2.apk -n ${resource_group}/${pod_name}:/data/wedata/runner
# Install after entering the pod:
sudo su
apk add --allow-untrusted --no-network lftp-4.8.3-r2.apk
# Connect using the command
lftp -u ${username},'${password}' -p ${port} ${ip}
# Use active mode
lftp -e "set ftp:passive-mode 0" -u ${username},'${password}' -p ${port} ${ip}
# It can also be used to connect to SFTP
lftp sftp://${username}:${password}@${ip}:${port}
```

FTP Script Demo

If you configure an offline task using **Script Mode**, you need to write the **reader parameter** and **writer parameter** in the task script according to the unified script format requirements.

```
"job": {
    "syncType": "data", //synchronization method, supports data synchronization and file transfer
    "content": [
        {
            "reader": {
                "parameter": {
                    "nullFormat": "", //Null Value Conversion
                    "compress": "zip", //File compression type
                    "column": [
                        "*"
                    ],
                    "skipHeader": "false", //Skip Header
                    "fieldDelimiter": ",", //Field Delimiter
                    "encoding": "utf-8", //Encoding type
                    "path": "/path/source", //Source file path
                    "protocol": "ftp",
                    "password": "******",
                    "port": 21,
                    "lineDelimiter": "\n", //Row Separator
                    "host": "ip",
                    "fileType": "text", //File type
                    "username": "ftpuser"
                },
                "name": "ftpreader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "fileName": "sinkFile", //Target file name
                    "nullFormat": "null", // Null Value Conversion.
                    "writeMode": "truncate", //Write mode
                    "suffix": "txt", //File suffix
                    "encoding": "utf-8",
                    "fieldDelimiter": ",", //Field Delimiter
                    "path": "/path/sink", //Target file path
                    "protocol": "ftp",
                    "password": "******",
```

©2013–2025 Tencent Cloud. All rights reserved.

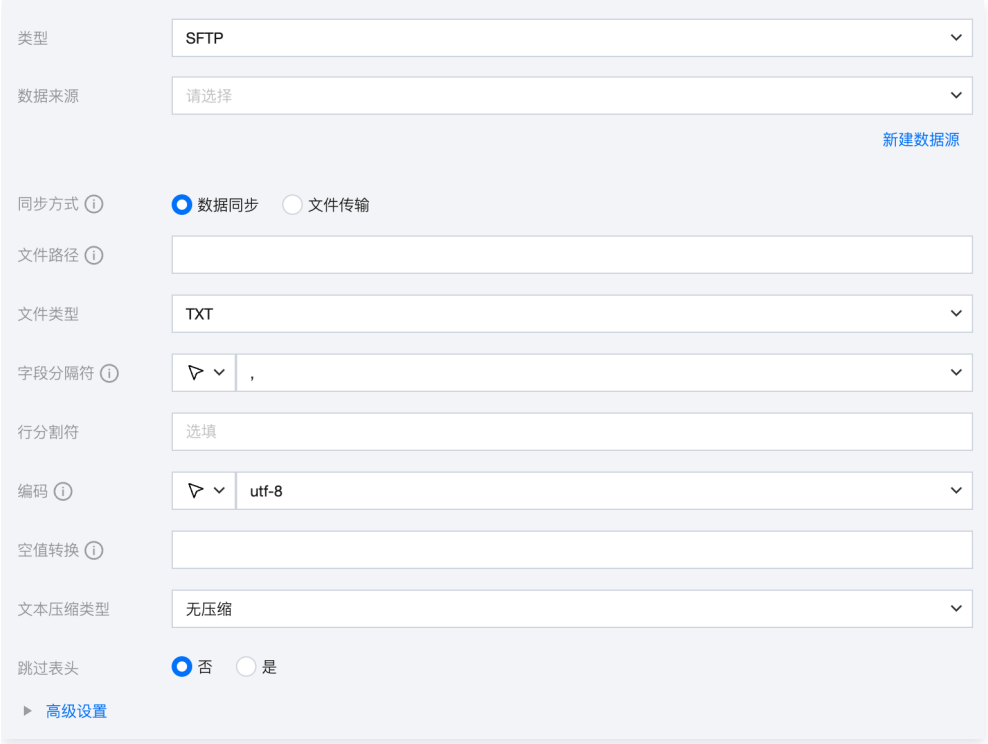
```
                    "port": 21,
                    "lineDelimiter": "\n", //Line Delimiter
                    "host": "ip",
                    "fileType": "text", //File type
                    "username": "ftpuser"
                },
                "name": "ftpwriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

SFTP Data Source

Last updated：2024-12-12 19:11:32

SFTP Offline Single Table Read Node Configuration



Parameter	Description
Data Source	Select available SFTP data sources in the current project.
Synchronization Method	SFTP supports two synchronization methods: • Data Synchronization: Parses structured data content and maps and synchronizes data content based on field relationships. • File Transfer: Transfers the entire file without content parsing, suitable for unstructured data synchronization. Note: File transfer only supports data sources where both the source and destination are file types (COS/HDFS/SFTP/FTP), and the synchronization method for both must be file transfer.
File path	For the path and file name information of the SFTP file system, you need to fill in the complete file path and file name, including the path and file suffix. Multiple paths can be specified here. • When specifying a single remote SFTP file, SFTP can currently use only single-threading for data extraction. In the future, multi-threading concurrency for a single file will be supported for non-compressed files. • When specifying multiple remote SFTP files, SFTP supports the use of multi-threading for data extraction. The number of concurrent threads is specified by the number of channels. • When specifying a wildcard, SFTP attempts to list multiple file information. For example, specifying / means reading all files in the / directory, and specifying /bazhen/ means reading all files in the bazhen directory. SFTP currently supports only the asterisk (*) as a file wildcard and supports using scheduling parameters to flexibly configure the file name and file path.
File Type	SFTP supports two file types: TXT, CSV. • TXT: Represents the TextFile file format. • CSV: Represents the standard HDFS file format (logical two-dimensional table).
Field delimiter	Field separator for reading. When reading data, SFTP requires specifying the field separator. If not specified, it defaults to a comma (,). The interface configuration will also default to a comma (,).

©2013-2025 Tencent Cloud. All rights reserved.

Row Separator (optional)	Linux default is \n, Windows default is \r\n, manually input supports a single character as a row separator.
Encoding	Encoding configuration for reading the file. Supports utf8 and gbk.
Null Value Conversion	Convert the specified string to null during read.
Text compression type	Supports uncompressed, zip, gzip, bzip2.
Skip table header	• No: Do not skip the header when reading. • Yes: Skip the header when reading.
Advanced Settings (optional)	You can configure parameters based on business requirements.

SFTP Offline Single Table Write Node Configuration



Parameter	Description
Data Source	Select available SFTP data sources in the current project.
Synchronization Method	SFTP supports two synchronization methods: • Data Synchronization: Parses structured data content and maps and synchronizes data content based on field relationships. • File Transfer: Transfers the entire file without content parsing, suitable for unstructured data synchronization.

©2013–2025 Tencent Cloud. All rights reserved.

	Note: File transfer only supports data sources where both the source and destination are file types (COS/HDFS/SFTP/FTP), and the synchronization method for both must be file transfer.
File path	Path information of the file system. The path supports using '*' as a wildcard, specifying the wildcard will traverse multiple file information.
File Name	The name of the written file. This file name will have a random suffix added as the actual write name.
Write mode	SFTP supports three write modes: • append: No processing before writing, ensuring no filename conflicts. • nonConflict: Error when filename conflicts. • overwrite: Clears all files with the same filename prefix before writing.
File Type	Supports two file types: TXT, CSV. • TXT: Represents the TextFile file format. • CSV: Represents the standard HDFS file format (logical two-dimensional table).
Field delimiter	Field delimiter for writing. When writing to SFTP, ensure the field delimiter matches the one in the created SFTP table; otherwise, data cannot be found in the SFTP table. Options: '\t', '\u001', '|', 'space', ';', ',', '.
Encoding	Encoding configuration for writing files. Supports utf8 and gbk encodings.
Null Value Conversion	When writing, convert null to the specified string.
Include header	• No: Do not skip the header when writing. • Yes: Skip the header when writing.
Mark the completion file.	An empty .ok file indicates the task transfer is complete. • Do not generate: After the task is completed, do not generate an empty .ok file. (Default selected) • Generate: After the task is completed, generate an empty .ok file
Mark completion filename	When filling out, you need to include the complete file path, name, and suffix. By default, the .ok file is generated according to the target file name and suffix in the target path. Support using time scheduling parameters ${yyyyMMdd} in data synchronization and file transfer modes. Support using built-in variable ${filetrans_di_src} in file transfer mode to represent the source file name. For example: /user/test/${filetrans_di_src}_${yyyyMMdd}.ok. **Note:** In file transfer mode, if the target file name **Target file name** uses ${filetrans_di_src} to reference the source file name, and if the completion marker files are generated according to the target file name and its suffix as .ok files, the number of generated completion marker files depends on the number of files at the source.
Advanced Settings (optional)	You can configure parameters based on business requirements.

Data type conversion is supported

SFTP implements a bidirectional channel feature for reading and writing SFTP, where remote SFTP files are unstructured data storage. The data processing engine automatically converts to Bytes type during reading and writing.

SFTP Script Demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "syncType": "data", //synchronization method, supports data synchronization and file transfer
    "content": [
```

©2013-2025 Tencent Cloud. All rights reserved.

```
        {
            "reader": {
                "parameter": {
                    "path": "/path/source_file/", //File path
                    "protocol": "sftp",
                    "password": "******",
                    "nullFormat": "null", // Null Value Conversion.
                    "port": 22,
                    "host": "ip",
                    "column": [ //column name
                        "*"
                    ],
                    "skipHeader": "false", //Skip Header
                    "fieldDelimiter": "|", //Field delimiter
                    "encoding": "utf-8", //Encoding method
                    "fileType": "text", //File type
                    "username": "wedata"
                },
                "name": "ftpreader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "fileName": "sink_file", //Target file name
                    "nullFormat": "null", // Null Value Conversion.
                    "writeMode": "truncate", //Write mode
                    "suffix": "txt", //File suffix
                    "encoding": "utf-8", //Encoding method
                    "fieldDelimiter": ",", //Field Delimiter
                    "path": "/path/sink", //File path
                    "protocol": "sftp",
                    "password": "******",
                    "port": 22,
                    "host": "ip",
                    "fileType": "text", //File type
                    "username": "wedata"
                },
                "name": "ftpwriter"
            }
        }
    ],
    "setting": {
        "errorLimit": {
            "record": 0 // Dirty data threshold
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

COS Data Source

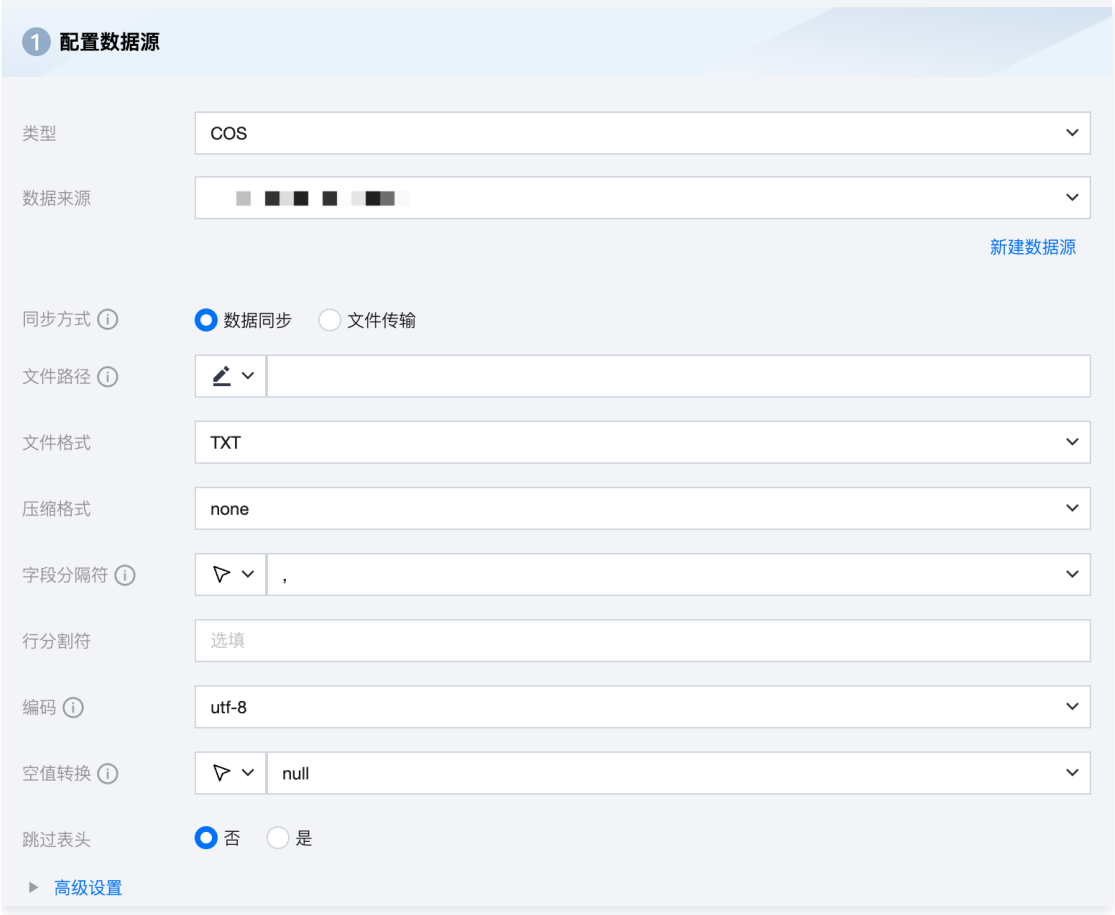
Last updated：2024-12-12 19:14:31

Use Limits

It is recommended to add QcloudCOSBucketConfigRead and QcloudCOSDataFullControl policies while configuring read and write permissions for sub-account authorization data. This is because Spark might require headBucket permission when operating COS files in certain scenarios.



COS Offline Single Table Reading Node Configuration



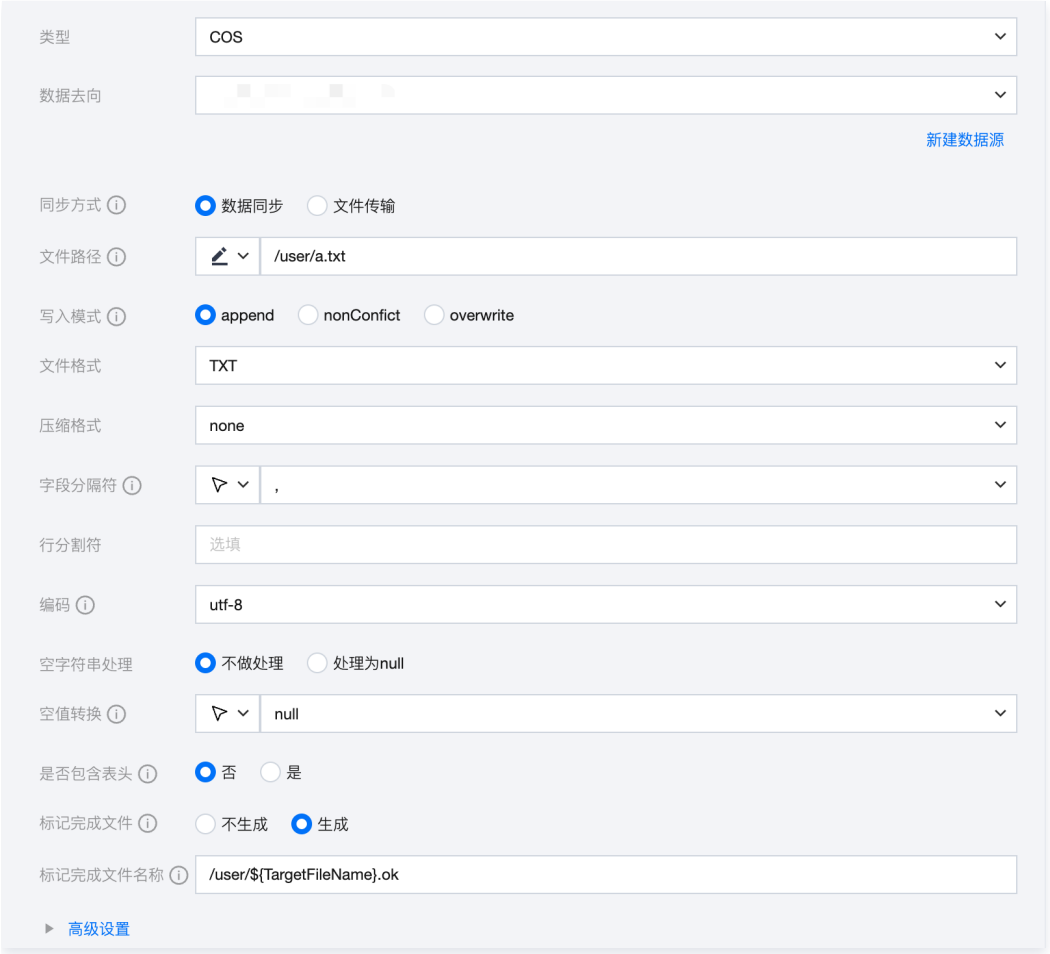
Parameters	Description
Data Source	Select the available COS data source in the current project.

©2013-2025 Tencent Cloud. All rights reserved.

Synchronization Method	COS supports two synchronization methods: • Data Synchronization: Parses structured data content and maps and synchronizes data content based on field relationships. • File Transfer: Transfers the entire file without content parsing, suitable for unstructured data synchronization. Note: File transfer only supports data sources where both the source and destination are file types (COS/HDFS/SFTP/FTP), and the synchronization method for both must be file transfer.
File path	The COS file path should include the bucket name, e.g., cosn://bucket_name. Note: File paths can achieve partial matching using the wildcard *, supporting matching folders/files that start with xxxx or end with xxxx. For example, cosn://bucket_name/*.txt
File format	COS supports five file types: TXT, ORC, PARQUET, CSV, JSON. • TXT: Represents the TextFile file format. • ORC: Represents the ORCFile file format. • PARQUET: Represents the standard Parquet file format. • CSV: Represents the standard HDFS file format (logical two-dimensional table). • JSON: Represents the JSON file format.
Compression Format	File compression method, currently supported: none, deflate, gzip, bzip2, lz4, snappy. **Note** • Since there is no unified stream format for snappy yet, DataInLong currently only supports the most mainstream hadoop-snappy (snappy stream format on hadoop) and framing-snappy (snappy stream format recommended by Google).
Field delimiter	Field delimiter for reading. When the file format is TXT or CSV, a field delimiter needs to be specified. If not specified, the default is a comma (,). • Other available delimiters: '\t', '\u001', '|', 'space', ';', ',', '. • If you want to treat each line as a column in the destination end, please use a character that does not exist in the line as the delimiter, such as the invisible character \u0001. **Note** Only TXT and CSV file formats support specifying a line delimiter.
Row Separator	When the file format is TXT or CSV, a line delimiter can be specified. The default for Linux is \n, and for Windows, it is \r\n. Custom single-character line delimiters are supported. **Note** Only TXT and CSV file formats support specifying a line delimiter.
Encoding	Encoding configuration for reading the file. Supports utf8 and gbk.
Null Value Conversion	When reading, the specified string will be converted to NULL.
Skip table header	No: The header is not included when reading. Yes: The header is included when reading. **Note** • Only TXT and CSV file formats are supported. • If the source includes a header and you choose not to skip it, it is recommended that the destination does not include the header to avoid duplicate header data.

©2013-2025 Tencent Cloud. All rights reserved.

COS Offline Single Table Writing Node Configuration



Parameters	Description
Data Flow Direction	Select the available COS data source in the current project.
Synchronization Method	COS supports two synchronization methods: • Data Synchronization: Parses structured data content and maps and synchronizes data content based on field relationships. • File Transfer: Transfers the entire file without content parsing, suitable for unstructured data synchronization. **Note:** File transfer only supports data sources where both the source and destination are file types (COS/HDFS/SFTP/FTP), and the synchronization method for both must be file transfer.
File path	The COS file path should include the bucket name, e.g., cosn://bucket_name. **Note** File paths can achieve partial matching using the wildcard *, supporting matching folders/files that start with xxxx or end with xxxx. For example, cosn://bucket_name/*.txt
Write mode	COS supports three write modes: • append: Directly write using filename without any preprocessing to ensure no filename conflict. • nonConflict: Error when filename conflicts.

©2013–2025 Tencent Cloud. All rights reserved.

	• overwrite: Clear all files with the same filename prefix before writing. For example, if "fileName": "abc", all files starting with abc in the corresponding directory will be cleared.
File format	COS supports four file types: TXT, CSV, ORC, PARQUET. • TXT: Represents the TextFile file format. • CSV: Represents the standard HDFS file format (logical two-dimensional table). • ORC: Represents the ORCFile file format. • PARQUET: Represents the standard Parquet file format.
Compression Format	File compression method, currently supported: none, deflate, gzip, bzip2, lz4, snappy. **Note** Since there is no unified stream format for snappy yet, DataInLong currently only supports the most mainstream hadoop-snappy (snappy stream format on hadoop) and framing-snappy (snappy stream format recommended by Google).
Field delimiter	When the file format is TXT or CSV, select or manually enter the field delimiter for reading. The field delimiter for writing to COS must match the field delimiter of the created COS table, otherwise, data cannot be queried in the COS table. Options: '\t', '\u001', '|', 'space', ';', ','. **Note** Only TXT and CSV file formats support specifying a line delimiter.
Row Separator	When the file format is TXT or CSV, a line delimiter can be specified. The default for Linux is \n, and for Windows, it is \r\n. Custom single-character line delimiters are supported. **Note** Only TXT and CSV file formats support specifying a line delimiter.
Encoding	Encoding configuration for writing files. Supports utf8 and gbk encodings.
Empty string processing	• No processing: Do not process empty strings during writing. • Process as NULL: When writing, empty strings are processed as NULL.
Null Value Conversion	When writing, convert NULL to the specified string. NULL represents an unknown or inapplicable value, different from 0, an empty string, or other values. **Note** Only TXT and CSV file formats are supported for selection.
Include header	• Yes: Include header when writing • No: Do not include header when writing **Note** Only TXT and CSV file formats support selecting whether to include the header.
Mark the completion file.	A blank .ok file indicates the task transfer is complete. • Do not generate: After the task is completed, a blank .ok file will not be generated. (Selected by default) • Generate: After the task is completed, a blank .ok file will be generated.
Mark completion filename	When filling in, include the full file path, name, and extension. By default, a .ok file is generated in the target path with the target file name and extension. Supports the use of time scheduling parameters like ${yyyyMMdd} in data synchronization and file transfer modes.

©2013-2025 Tencent Cloud. All rights reserved.

	Supports using the built-in variable ${filetrans_di_src} to represent the source file name in file transfer mode. For example: /user/test/${filetrans_di_src}_${yyyyMMdd}.ok. **Note:** In file transfer mode, if the "target file name" uses ${filetrans_di_src} to reference the source file name, and the completion marker files are generated according to the target file name and extension, the number of generated completion marker files depends on the number of files at the source end.
Advanced Settings (optional)	You can configure parameters based on business requirements.

Data type conversion supports (data synchronization mode).

COS Select determines the data type of your input through the CAST Function. Generally, if you do not specify a data type using the CAST Function, COS Select will treat the input data type as string.

Read

COS data type	Internal Type
INTEGER	LONG
FLOAT、DECIMAL	DOUBLE
STRING	STRING
TIMESTAMP	DATE
BOOL	BOOLEAN

Write

Internal Type	COS data type
LONG	INTEGER
Double	FLOAT、DECIMAL
STRING	STRING
DATE	TIMESTAMP
BOOLEAN	BOOL

COS script demo.

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "path": "/",
                    "nullFormat": "null", // Null Value Conversion.
                    "compress": "gzip", //Compression Format
                    "lineDelimiter": "\n", //Line Delimiter
                    "hadoopConfig": {
                        "fs.defaultFS": "cosn://source-bucket",
```

©2013–2025 Tencent Cloud. All rights reserved.

```
                        "fs.cosn.userinfo.secretId": "******",
                        "fs.cosn.tmp.dir": "/tmp/hadoop_cos",
                        "fs.cosn.trsf.fs.ofs.tmp.cache.dir": "/tmp/",
                        "fs.cosn.userinfo.secretKey": "******",
                        "fs.cosn.bucket.region": "ap-beijing",
                        "fs.cosn.impl": "org.apache.hadoop.fs.CosFileSystem",
                        "fs.cosn.trsf.fs.ofs.bucket.region": "ap-beijing",
                        "fs.cosn.trsf.fs.ofs.user.appid": "******"
                    },
                    "column": [
                        "*"
                    ],
                    "defaultFS": "cosn://source-bucket", //Source File Path
                    "skipHeader": "false", //Skip Header
                    "fieldDelimiter": ",", //Field Delimiter
                    "encoding": "utf-8", //Encoding method
                    "fileType": "text" //File type
                },
                "name": "hdfsreader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "path": "/",
                    "fileName": "part",
                    "nullFormat": "null", // Null Value Conversion.
                    "compress": "gzip", //Compression Format
                    "hadoopConfig": {
                        "fs.defaultFS": "cosn://sink-bucket",
                        "fs.cosn.userinfo.secretId": "******",
                        "fs.cosn.tmp.dir": "/tmp/hadoop_cos",
                        "fs.cosn.trsf.fs.ofs.tmp.cache.dir": "/tmp/",
                        "fs.cosn.userinfo.secretKey": "******",
                        "fs.cosn.bucket.region": "ap-beijing",
                        "fs.cosn.impl": "org.apache.hadoop.fs.CosFileSystem",
                        "fs.cosn.trsf.fs.ofs.bucket.region": "ap-beijing",
                        "fs.cosn.trsf.fs.ofs.user.appid": "******"
                    },
                    "defaultFS": "cosn://sink-bucket", //Target file path
                    "emptyAsNull": false,
                    "writeMode": "append", //Write mode
                    "suffix": "parquet", //File suffix
                    "fieldDelimiter": ",", //Field delimiter
                    "encoding": "utf-8", //Encoding method
                    "fileType": "parquet" //File type
                },
                "name": "hdfswriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
```

©2013–2025 Tencent Cloud. All rights reserved.

```
}
```

Rest API Data Source

©2013–2025 Tencent Cloud. All rights reserved.

Last updated: 2024-12-12 19:15:09

Rest API Offline Single Table Read Node Configuration



Parameters	Description
Data Source	Available Rest API Data Sources.
Synchronization Method	Supports two synchronization methods: data synchronization and file transfer • Data Synchronization: Parses structured data content and maps and synchronizes data content based on field relationships. • File Transfer: Transfers the entire file without content parsing, suitable for unstructured data synchronization.
Request method	Supports POST and GET request methods. • Use the POST method to fill in JSON type parameters. • Use the GET method with abc=1&def=1.
Returned Data Type	The format of the returned data, currently only JSON data is supported.
JSON Data Path (optional)	The path to query a single JSON object or JSON array from the returned result.
Returned Data Structure	Supports single JSON data and JSON array data.
Request Header (optional)	Header information to pass to the RESTful interface.

©2013-2025 Tencent Cloud. All rights reserved.

Request Parameters (optional)	• Use the GET method with abc=1&def=1. • Use the POST method to fill in JSON type parameters.
Request method	When running a sync task once, check if multiple requests are initiated. If so, configure the corresponding parameters and starting index value.
Skip table header	• No: Do not skip the table header during synchronization. • Yes: Skip the table header during synchronization. **Note** Only supported when Synchronization Method is **Data Synchronization**.

Data type conversion is supported

Type Classification	DataInLong column Configuration Type
Integer Type	LONG，INT
String Type	STRING
Floating Point Type	DOUBLE，FLOAT
Boolean Type	BOOLEAN
Date and Time Type	DATE

Rest API Usage Example

Scene 1: Single Read

Assume the API request parameters are as follows:

```
// Assume it is a GET request
paramA=a&paramB=b
```

The returned data structure of the API is as follows:

```
{
    "code": "success",
    "data": {
        "id": "100",
        "name": "name100"
    }
}
```

To read all returned fields, the task can be configured as:

©2013–2025 Tencent Cloud. All rights reserved.



Fields can be configured as:



Scenario 2: Paged read

Assume the API request parameters are as follows:

```
// Assume it's a post request, where pageNum is the page number
{
    "pageNum": 1,
    "paramA": "a",
    "paramB": "b"
}
```

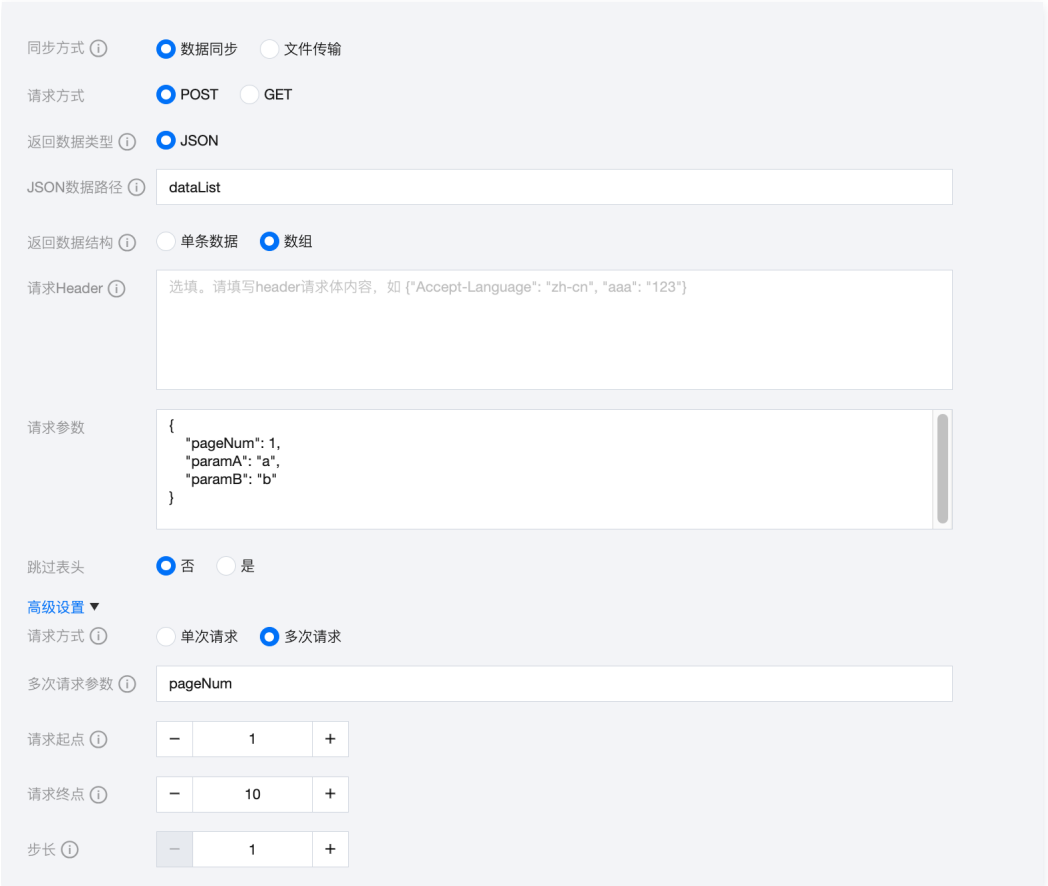
The returned data structure of the API is as follows:

```
// Where dataList is the business data in array format
{
    "total": "20",
```

©2013–2025 Tencent Cloud. All rights reserved.

```
    "code": "success",
    "dataList": [
        {
            "id": 1,
            "name": "name_1",
            "info": {
                "infoId": "info_1",
                "infoName": "infoName_1"
            }
        },
        {
            "id": 2,
            "name": "name_2",
            "info": {
                "infoId": "info_2",
                "infoName": "infoName_2"
            }
        },
        {
            "id": 3,
            "name": "name_3",
            "info": {
                "infoId": "info_3",
                "infoName": "infoName_3"
            }
        }
    ]
}
```


If you need to synchronize data from page 1 to page 10 and expect to read the content under dataList in the returned structure, the task can be configured as follows:



©2013–2025 Tencent Cloud. All rights reserved.

Fields can be configured as:



Rest API script demo

If you configure an offline task using **script mode**, you need to write the **reader parameters** in the task script according to the unified script format requirements. (Rest API only supports reader)

```
"job": {
    "syncType": "data", //synchronization method, supports data synchronization and file transfer
    "content": [
        {
            "reader": {
                "parameter": {
                    "headers": { //Request Header
                        "Accept-Language": "zh-cn "
                    },
                    "method": "POST", // Request method, supports GET and POST
                    "responseStructure": "object", // Type of returned data, supports single data and
array
                    "requestParams": { // Request parameters
                        "name": "wedata",
                        "pagenumber": 1
                    },
                    "resultKey": "", // JSON Path
                    "connection": [ // Source URL address
                        {
                            "url": "https://cloud.tencent.com/document/product/1267/105543"
                        }
                    ],
                    "skipHeader": "false"
                },
                "name": "httpreader"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
```

©2013–2025 Tencent Cloud. All rights reserved.

```
            }
        }
    }
```

©2013–2025 Tencent Cloud. All rights reserved.

Elasticsearch/Tencent Cloud Elasticsearch Data Source

Last updated：2024-12-12 19:15:48

The configuration method for Tencent Cloud Elasticsearch Data Source is the same as that for Elasticsearch Data Source. Here, the Elasticsearch Data Source is used as an example for explanation.

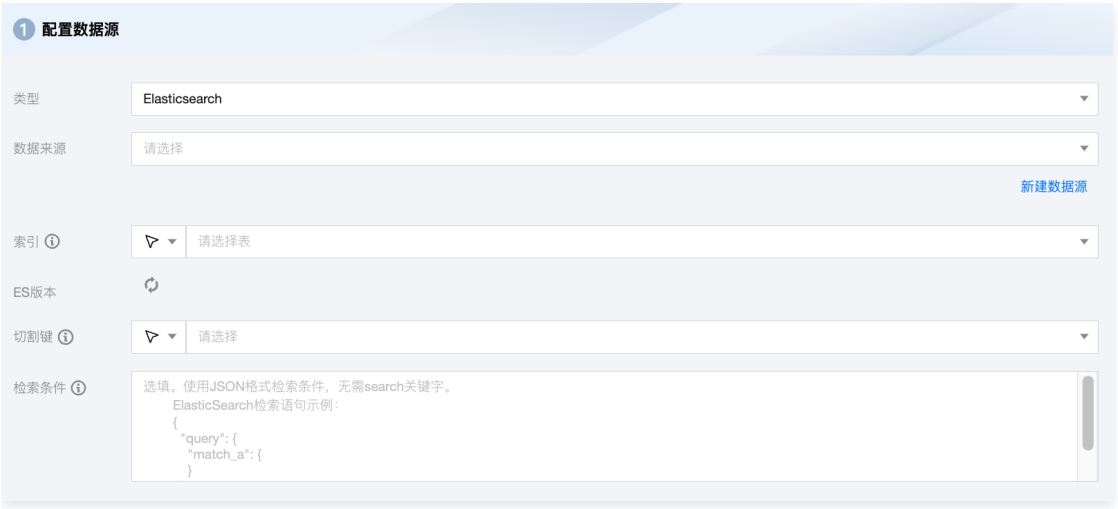
Supported Versions

Supports Elasticsearch 6.x, 7.x versions.

Use Limits

1. Elasticsearch Reader will fetch shard information from the server for data synchronization. Ensure the server's shards are alive during synchronization to avoid data inconsistency risks.
2. Does not support synchronizing fields of type scaled_float.
3. Does not support synchronizing indexes with fields containing the keyword $ref.

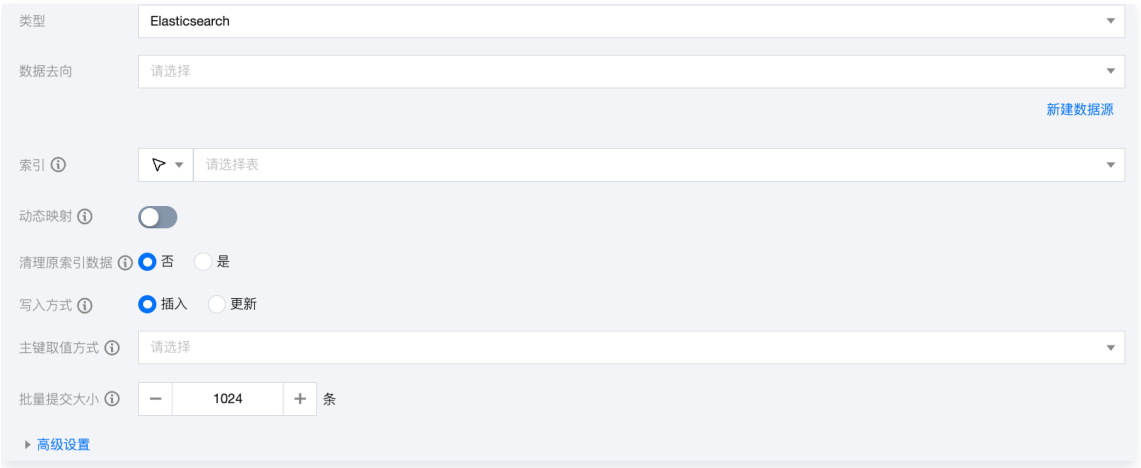
Elasticsearch Offline Single Table Read Node Configuration



Parameter	Description
Data Source	Select an available Elasticsearch data source in the current project.
Index	Supports multiple index names or regular expressions. For index name regular expressions, use wildcard (*), e.g., index_*.
ES Version	Determine ES version based on data source and index.
Split Key	Specify the field for data sharding. Once specified, concurrent tasks will be initiated for data synchronization. You can use a column from the source data table as the split key. It is recommended to use the primary key or an indexed column as the split key.
Search Criteria (Optional)	Use JSON format for search.

Elasticsearch Offline Single Table Write Node Configuration

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Flow Direction	Select an available Elasticsearch data source in the current project.
Index	Index Name in ElasticSearch.
Dynamic Mapping	Definition: When an unknown field is found in a document, should the synchronization task add mapping for the field using the Elasticsearch dynamic mapping mechanism. • On: Retain Elasticsearch's automatic mappings. • Off: Default off, generate and update Elasticsearch's mappings according to the configured columns of the synchronization task. • The default type for Elasticsearch 7.x is _doc. When using Elasticsearch's automatic mappings, please configure _doc with esVersion set to 7.
Clean the Original Index Data	Manually choose whether to clean the original index data: • No: Retain the existing data in the index before importing. • Yes: Delete the original index and recreate the same name index before importing data. This will delete the data under that index.
Write Mode	Support two write modes: Insert and Update • Insert: Insert all data directly. • Update: Update data if the same primary key exists; otherwise, insert.
Primary key value method	Three value methods are supported: • Source Table Primary Key: The document's id uses the primary key of the source table. • Composite Primary Key: The document's id is determined by multiple columns from the source table. • No Primary Key: A default _id value is generated.
Batch Submission Size	The size of the record count for one-time batch submission can significantly reduce the number of network interactions between the data synchronization system and ElasticSearch, and improve overall throughput. If this value is set too high, it can cause the data synchronization process to run into OOM (Out Of Memory) issues.
Advanced Settings (optional)	You can configure parameters based on business requirements.

Data type conversion is supported

Read

Elasticsearch Data Type	Internal Type
byte,short,integer,long,unsigned long	Long

©2013–2025 Tencent Cloud. All rights reserved.

float,double,half_float	Double
string,text,keyword,integer_range,long_range,float_range,double_range,date_range,array,object,nested,flattened,geo_point,geo_shape	String
date	Date
binary	Bytes
boolean	Boolean

Write

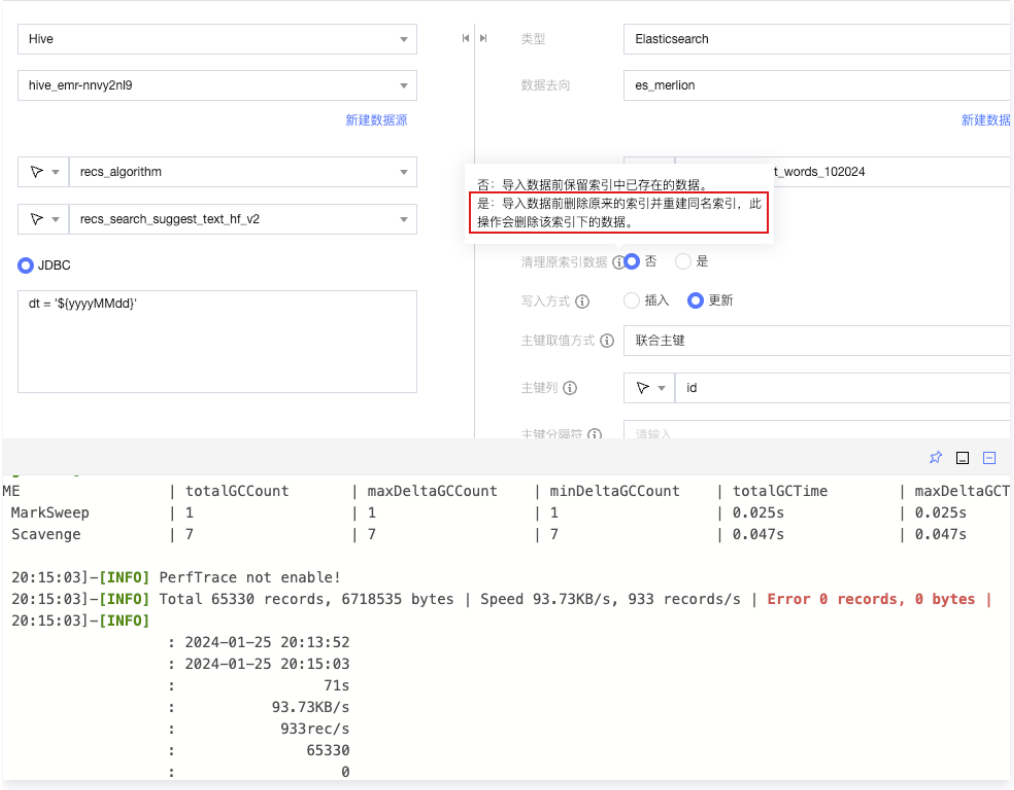
Internal Type	Elasticsearch Data Type
Long	byte,short,integer,long
Double	float,double
String	string,text,keyword,object,nested,geo_point,geo_shape,ip,binary,completion
Date	date
Boolean	boolean

FAQs

After clearing the index with ES writer, the automatically recreated index fields may not match the expected types

Issue Information:

ES write node configuration "Cleanup original index data", when selecting "Yes" to rerun the task, ES writer will not proactively create the index. When directly writing data, ES will infer and automatically create the index based on the data. The new index field type may be different from the old version.



Cause:

©2013–2025 Tencent Cloud. All rights reserved.

The method ES writer uses to clean up the original index data is to directly delete the index. There is no syntax similar to MySQL's truncate table. After the index is deleted, ES Server infers index field types, which may differ from the original index's field types.
Solution:
Define the index field types in advance through the ES template.
Create an example: (merlion_suggest_words_* indicates it applies to all indexes starting with merlion_suggest_words_).

```
PUT _template/merlion_suggest_words
{
  "template": "merlion_suggest_words_*",
  "order": 1,
  "mappings": {
    "properties": {
      "biz_owner": {
        "type": "keyword"
      },
      "create_time": {
        "type": "date",
        "format": "yyyy-MM-dd HH:mm:ss||date_hour_minute_second||strict_date_optional_time||epoch_millis"
      },
      "is_deleted": {
        "type": "short"
      },
      "sug_words": {
        "type": "keyword"
      },
      "language": {
        "type": "keyword"
      }
    }
  },
  "settings": {
    "refresh_interval": "5s",
    "number_of_replicas": 1,
    "number_of_shards": 1
  }
}
```

ElasticSearch Script Demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "accessId": "elastic",
                    "endpoint": "http://ip:9200",
                    "search": { // Search Criteria
                        "query": {
                            "mathc_a": {}
                        }
                    },
                    "accessKey": "******",
                    "scroll": "100m",
                    "column": [ //column name
                        {
                            "name": "_id",
```

©2013–2025 Tencent Cloud. All rights reserved.

```
                        "type": "string"
                    },
                    {
                        "name": "_index",
                        "type": "string"
                    }
                    {
                        "name": "_id",
                        "type": "string"
                    }
                ],
                "index": "source_index" // Index
            },
            "name": "elasticsearchreader"
        },
        "transformer": [],
        "writer": {
            "parameter": {
                "settings": {
                    "index": {
                        "number_of_shards": 5,
                        "number_of_replicas": 0
                    }
                },
                "column": [ //column name
                    {
                        "name": "a1",
                        "type": "string"
                    },
                    {
                        "name": "a2",
                        "type": "string"
                    },
                    {
                        "name": "_id",
                        "type": "id"
                    }
                ],
                "index": "sink_index", // Index
                "type": "_doc",
                "splitter": ",",
                "accessId": "elastic",
                "actionType": "index",
                "endpoint": "http://ip:9200",
                "accessKey": "******",
                "cleanup": false, // Whether to clean up original index data
                "discovery": false,
                "dynamic": true,
                "batchSize": 1024, // Batch submission size
                "esVersion": "5.6.4"
            },
            "name": "elasticsearchwriter"
        }
    }
],
"setting": {
    "errorLimit": { // Dirty data threshold
        "record": 0
    },
    "speed": {
```

©2013–2025 Tencent Cloud. All rights reserved.

```
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

Mongo/Tencent Cloud Mongo Data Source

Last updated: 2024-12-12 19:20:16

The configuration method for Tencent Cloud Mongo data source is the same as that for Mongo data source. Here, the Mongo data source is used as an example to explain.

Supported Versions

Supports MongoDB versions 3.6, 4.x, 5.x, and 6.x.

Use Limits

1. For incremental synchronization, the where condition syntax is:
   `create_time>= '${yyyy-MM-dd-1d HH:mm:ss}' and create_time < '${yyyy-MM-dd HH:mm:ss}'`
   Currently, Bound supports integer and date function configuration. The usage method for date function configuration:

```
// Convert to 13-bit timestamp (milliseconds)
TimestampMillis('yyyy-MM-ddTHH:mm:00+0800')
TimestampMillis('2023-07-10T00:00:00+0800')
TimestampMillis('2023-07-10 00:00:00')
TimestampMillis('2023-07-10')

// Convert to 10-bit timestamp (seconds)
TimestampSeconds('yyyy-MM-ddTHH:mm:00+0800')
TimestampSeconds('2023-07-10T00:00:00+0800')
TimestampSeconds('2023-07-10 00:00:00')
TimestampSeconds('2023-07-10')
```

2. Supports connections using the corresponding account for the MongoDB database. If you are using TencentDB for MongoDB, there will be a default admin account. For security policies, avoid using admin as the access account when adding a MongoDB data source.
3. If MongoDB is a Sharding Cluster, when configuring the data source, you need to configure the mongos address and avoid configuring the mongod/shard node address. Otherwise, during the data extraction from MongoDB for the synchronization task, it may result in querying only the specified shard data instead of the expected whole dataset. For details about mongos and mongod, please see mongos shards.
4. In the case of concurrency greater than 1, custom cutting keys are supported. The field chosen as the custom cutting key must: 1) have an index created, otherwise the concurrency settings will not take effect; 2) all values of the field must be of the same data type, otherwise the number of synchronized entries may be inaccurate.

> **Note**
> When concurrency is greater than 1, task splitting will use the cutting key field for division. Therefore, in this scenario, the cutting key field does not support mixed types. If the field set as the cutting key has multiple data types, the number of synchronized entries may be inaccurate. In this case, it is recommended to use single concurrency for data synchronization and either not configure splitFactor or set splitFactor to 1.

Mongo Offline Single Table Read Node Configuration

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Source	Available Mongo data source.
Database	Support selecting or manually entering the name of the database to be read • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Collection	Support selecting or manually entering the collection to be read.
Separator	When the field is of array type, use a delimiter to concatenate the array contents.
Cutting key (optional)	Specify fields for Data Sharding. Once specified, concurrent tasks will be launched to perform Data Synchronization. **⚠ Note:** • If the number of concurrent tasks is set to greater than 1, this item is required; otherwise, the concurrency settings will be invalid. • The field selected as the custom cutting key must: 1. Have an index created, otherwise the concurrency settings will not take effect; 2. All values of the field must be of the same data type, otherwise the number of entries in data synchronization will be inaccurate.
Filter Condition (Optional)	Fill in the corresponding filter statement according to the data type. This statement will serve as the filter condition for the data to be synchronized.
Advanced Settings (optional)	You can configure parameters based on business requirements.

Mongo Offline Single Table Write Node Configuration

©2013–2025 Tencent Cloud. All rights reserved.



Parameter	Description
Data Flow Direction	Select the available MongoDB data sources in the current project.
Database	Support selecting or manually entering the name of the database to be read • By default, the database bound to the data source is used as the default database; other databases need to be manually entered. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Collection	Support selecting or manually entering the collection to be read.
Write mode	Indicates whether to overwrite the data for the same primary key when writing. By default, the business primary key for each record uses the _id
Precondition Statement	Optional, indicating the pre-operation before writing data to MongoDB. Ensure the precondition statement meets JSON syntax requirements, such as {"type": "remove"}
Advanced Settings (optional)	You can configure parameters based on business requirements.

Data type conversion is supported

Read

The supported data types and conversion mappings for Mongo read are as follows (when processing Mongo, the data types of the Mongo data source will first be mapped to the data types of the data processing engine):

Mongo Data Type	Internal Type
INT32,INT64,BIGINT	LONG
DOUBLE,DECIMAL	DOUBLE
STRING, ARRAY,MAP,LIST	STRING
DATE,DATETIME	DATE
BOOLEAN	BOOLEAN
BYTES	BYTES

Write

The supported data types and conversion mappings for Mongo write are as follows:

©2013-2025 Tencent Cloud. All rights reserved.

Internal Type	Mongo Data Type
LONG	int, Long
DOUBLE	DOUBLE,DECIMAL
STRING	string, array
DATE	DATE,DATETIME
BOOLEAN	BOOLEAN
BYTES	BYTES

Mongo script demo.

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "userPassword": "******",
                    "address": [ // Source data source addresses
                        "ip:27017"
                    ],
                    "query": "{\"create_date\":{\"$gt\":ISODate(\"${yyyy-MM-dd-
1d}T00:00:00+0800\"),\"$lt\":ISODate(\"${yyyy-MM-dd}T00:00:00+0800\")}}", // Filtering condition
                    "dbName": "database", // Source database
                    "column": [
                        {
                            "name": "id",
                            "type": "string"
                        },
                        {
                            "name": "name",
                            "type": "string"
                        }
                    ],
                    "authDb": "admin",
                    "userName": "mongouser",
                    "splitter": "-", // Separator, used to concatenate array contents when the field is of
array type
                    "collectionName": "source_collection" // Source collection
                },
                "name": "mongodbreader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "userPassword": "******",
                    "address": [ // Supports multiple IPs
                        "ip:45400",
                        "ip:41748",
                        "ip:41748"
                    ],
                    "dbName": "database", // Target database
                    "column": [
                        {
                            "name": "id",
```

©2013–2025 Tencent Cloud. All rights reserved.

```
                            "type": "string"
                        },
                        {
                            "name": "name",
                            "type": "string"
                        }
                    ],
                    "authDb": "admin",
                    "userName": "mongouser",
                    "collectionName": "sink_collections", // Target collection
                    "preSql": { //precedence SQL
                        "type": "remove"
                    }
                },
                "name": "mongodbwriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

Redis Data Source

Last updated：2024-12-12 19:20:59

Use Limits

When using Redis Writer to write data to Redis, if the Value type is List, the synchronization results of rerunning the synchronization task are not idempotent. Therefore, if the Value type is List, the corresponding data on Redis needs to be manually cleared when rerunning the synchronization task.

Redis Offline Single Table Write Node Configuration

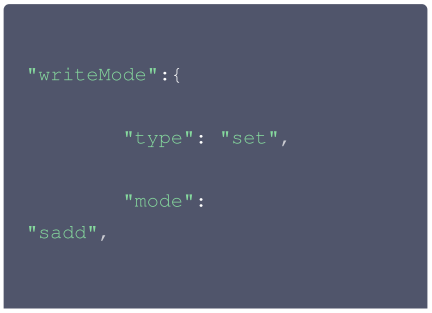


Parameters	Description
Data Flow Direction	Select an available Redis data source in the current project.
Database	Supports selection, or manual input of the library name to be written to • By default, Redis generates databases 0 – 15 as the default database. • When the data source network is not connected, causing database information to be unavailable for direct fetching, you can manually enter the database name. Data synchronization can still be performed given network connection in DataInLong.
Data Type	The value types for Redis writing include the following five types: • String (string) • String List (list) • String Set (set) • Ordered String Collection (zset) • Hashing (hash) Different value types will have slightly different data type configurations.
Write Mode	Automatically fill in the corresponding write mode based on the write type.
Key separator	The key separator for Redis writing must match the field separator of the created Redis table; otherwise, data cannot be retrieved from the Redis table. Options: '\t', '\u001', '|', 'space', ';' ', '.
Value separator	The value separator for Redis writing must match the value separator of the created Redis table; otherwise, data cannot be retrieved from the Redis table. Options: '\t', '\u001', '|', 'space', ';' ', '.
Cache expiration type	Three expiration types are supported: • Permanent Validity: The key value is set to permanent validity with no expiration time. • Fixed Failure Duration: Data expires a fixed period after the actual batch write-in time.

©2013-2025 Tencent Cloud. All rights reserved.

	• Uniform Expiry Time: All written data expires together at a specified time.
Batch submission count	The number of records submitted in a batch can significantly reduce the number of network interactions between the data synchronization system and Redis, and improve overall throughput. If this value is set too high, it can cause the data synchronization process to encounter an OOM Exception.
Advanced Settings (optional)	You can configure parameters based on business requirements.

Data Type

Value Type	type Parameter (Required)	mode Parameter (Required)	valueFieldDelimiter Parameter (Optional)	writeMode Configuration Sample
String (string)	type needs to be configured as string.	mode is the write mode parameter. When the value is a string (string), mode needs to be set to set. If the data to be stored already exists, it will overwrite the existing data.	valueFieldDelimiter The delimiter for between values is set to the default \u0001. This setting is mainly for cases when the source data has more than two columns per row, for example, when there are three columns, they are separated by the delimiter as in the example value1\u0001value2\u0001value3. If the source data has only two columns (i.e., key and value), then this configuration is not necessary.	
String List (list)	Type must be configured as list.	Mode is the writing mode parameter, value as a list of strings (list) can be configured as: • lpush, to store data on the leftmost side of the list. • rpush, to store data on the rightmost side of the list.		
String Set (set)	Type must be configured as set.	When mode is the writing mode parameter, and value is a set of strings (set): mode must be configured as sadd, indicating to store data in the set collection. If the data to be stored already exists, it will overwrite the existing data.		

©2013–2025 Tencent Cloud. All rights reserved.

				<pre>"valueFieldDelimiter": "\u0001"
				}</pre>

Ordered String Collection (zset)	Type must be configured as zset.	When mode is the writing mode parameter, and value is a sorted set of strings (zset): mode must be configured as zadd, indicating to store data in the zset sorted collection. If the data to be stored already exists, it will overwrite the existing data.	This parameter is not necessary.	<pre>"writeMode":{
				"type": "zset",
				"mode": "zadd"
				}</pre> • When the value type is zset, every row of the data source must follow the corresponding standards. That is, each row, in addition to the key, can only have one pair of score and value, and the score must be before the value for Redis Writer to parse whether the column corresponds to score or value.

Hashing (hash)	type needs to be configured as hash.	When mode is the writing mode parameter, and value is a hash, mode must be configured as hset, indicating to store data in the hash collection. If the data to be stored already exists, it will overwrite the existing data.	This parameter is not necessary.	<pre>"writeMode":{
				"type": "hash",
				"mode": "hset"
				}</pre> • When the value type is hash, every row of the data source must follow the corresponding standards. That is, each row, in addition to the key, can only have one pair of attribute and value, and the attribute must be before the value for Redis Writer to parse whether the column corresponds to attribute or value.

Redis Script Demo

If you configure offline tasks using **script mode**, you need to write the **writer parameter** in the task script according to the unified script format requirements. (Redis only supports writer)

```
"job": {
    "content": [
        {
            "writer": {
                "parameter": {
```

©2013–2025 Tencent Cloud. All rights reserved.

```
                    "valueIndexes": [
                        1
                    ],
                    "password": "******",
                    "database": "1", // Target slot
                    "expireTime": { // Cache expiration time
                        "unixtime": 0,
                        "seconds": 300
                    },
                    "port": 6379,
                    "keyFieldDelimiter": "|", // Key separator
                    "host": "******",
                    "writeMode": {
                        "mode": "set", // Write mode
                        "valueFieldDelimiter": ",", // Value separator
                        "type": "string" // Data type
                    },
                    "batchSize": 1000, // Batch submission size
                    "keyIndexes": [
                        0
                    ],
                    "username": ""
                },
                "name": "rediswriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

Kafka/CKafka data source

Last updated：2024-12-12 19:22:42

Supported Editions

1. Supports Kafka versions 2.4.x, 2.7.x, 2.8.x.
2. Both Kafka and CKafka can select Kafka data source and follow the Kafka steps for read/write configuration.

Use Limits

1. Kafka message formats currently only support JSON (no nested), CSV (single delimiter), and Avro. Other special formats of JSON such as Oracle ogg json are not supported.
2. When both parameter.groupId and parameter.kafkaConfig.group.id are configured, parameter.groupId takes precedence over the group.id in kafkaConfig configuration.

Kafka Offline Single Table Read Node Configuration



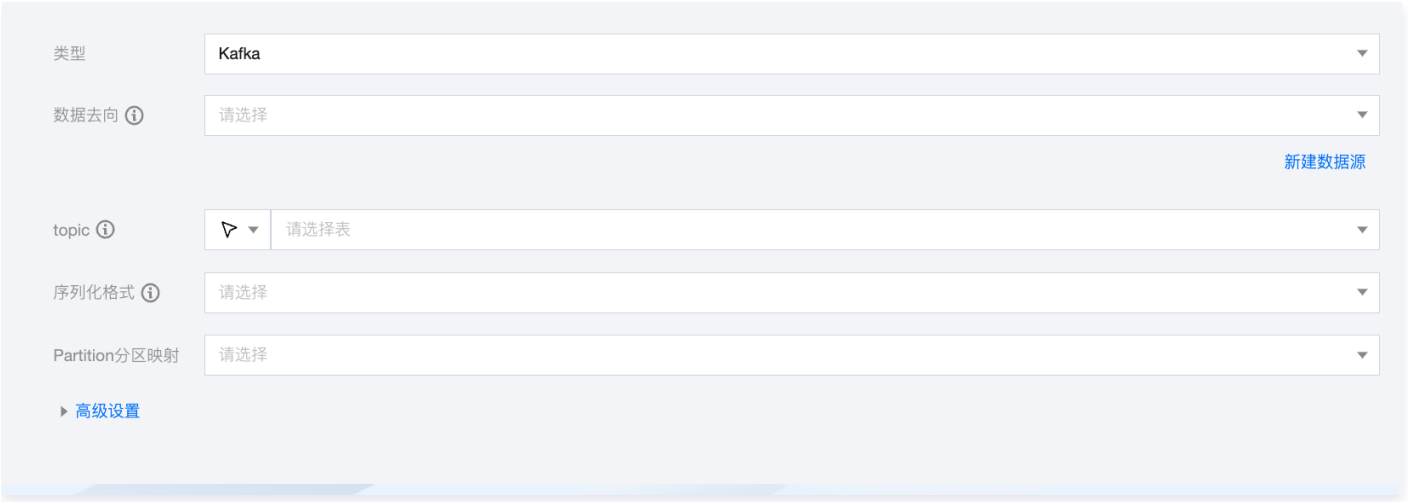
Parameters	Description
Data Source	The data source type supported for Kafka reading is Kafka and CKafka.
topic	Kafka's Topic is an aggregation of feeds of messages that Kafka processes.

Serialization Format	Kafka data to be read supports constant columns, data columns, and attribute columns:
	• Constant Column: A column wrapped in single quotes is a constant column, for example, ["'abc'", "'123'"]
	• Data column
	  ○ If your data is in JSON format, it supports fetching JSON attributes, e.g., ["event_id"]
	  ○ If your data is in JSON format, it supports fetching nested JSON attributes, e.g., ["tag.desc"]
	• Attribute Column
	  ○ __key__ represents the message key
	  ○ __value__ represents the complete content of the message
	  ○ __partition__ represents the partition of the current message
	  ○ __headers__ represents the headers of the current message
	  ○ __offset__ represents the offset of the current message
	  ○ __timestamp__ represents the timestamp of the current message

©2013-2025 Tencent Cloud. All rights reserved.

Consumer Group ID	Avoid this parameter conflicting with other consumption processes to ensure the correctness of the consumption location. If not specified, the default setting is group.id=WeData_group_${Task ID}.
Cycle Start Point	During task cycle runtime, each read of Kafka's start point. The default is the scheduled time of the last cycle, optional: Partition Start Point, Consumer Group Current Point, Specified Locus, Specified Time. • Specified Time: When Kafka writes data, it automatically generates a unixtime timestamp as the time record of the data. The synchronization task reads the corresponding data from Kafka by converting the user's configured yyyymmddhhmmss value to a unixtimestamp, e.g., "beginDateTime": "20210125000000". • Partition Start Point: Extract data from the smallest undeleted offset in each partition of the Kafka topic. • Consumer Group Current Point: Read data starting from the specified offset saved by the consumer group ID configured in the task, usually the last stopping point of the process using this consumer group ID (ensure that this consumer group ID is only used by the configured DataInLong task to avoid data loss due to sharing the group ID). If using the group current point, must configure the consumer group ID; otherwise, the DataInLong task will randomly generate a group ID, and due to reset policies differences of the offset will cause errors or read from the start or end point. Group offsets are periodically auto-committed to Kafka server-side, so after a task failure, rerunning the task may lead to data duplication or loss. In wizard mode, records beyond the end point are automatically discarded, and these discarded records' group offsets have already been committed to the server, making them unreadable in the next cycle.
Cycle End Point	During task cycle runtime, each read end point of Kafka. The default is the scheduled time of this cycle. When the keyType or valueType is configured as STRING, this option specifies the encoding for parsing the string.
Point Read Mode	The starting point for synchronizing data when the synchronization task is manually run. Provides two read modes: • latest: Read from the last offset position. • earliest: Read from the start point.

Kafka offline single table write node configuration



Parameters	Description
Data Flow Direction	The data source type supported for Kafka reading is Kafka and CKafka.
Topic	Kafka's Topic is an aggregation of feeds of messages that Kafka processes.
Serialization Format	Kafka data to be read supports constant columns, data columns, and attribute columns: • Constant Column: A column wrapped in single quotes is a constant column, for example, ["'abc'", "'123'"] • Data column: ○ If your data is in JSON format, it supports fetching JSON attributes, e.g., ["event_id"]

©2013-2025 Tencent Cloud. All rights reserved.

	○ If your data is in JSON format, it supports fetching nested JSON attributes, e.g., ["tag.desc"] • Attribute column: ○ __key__ represents the message key. ○ __value__ represents the full content of the message. ○ __partition__ represents the partition of the current message. ○ __headers__ represents the headers of the current message. ○ __offset__ represents the offset of the current message. ○ __timestamp__ represents the timestamp of the current message.
Partition Partition Mapping	Support for polling write to partition, hash write based on specified fields, and specify partition modes. • If you choose the hash write based on specified fields mode, you need to specify the field name. • If you choose the specify partition mode, you need to set the partition number.
Advanced Settings (optional)	You can configure parameters based on business requirements.

Data type conversion is supported

Read

Other data types are converted to string type by default.

Format of reading Kafka value	Source field type	Internal Type
JSON	Integer、Long、BigInteger	Long
	Float、Double	Double
	Boolean	Bool
	JSONArray、JSONObject	String
CSV	Integer、Long、BigInteger	Long
	Float、Double	Double
	Boolean	Bool
	JSONArray、JSONObject	String
AVRO	Integer、Long、BigInteger	Long
	Float、Double	Double
	Boolean	Bool
	JSONArray、JSONObject	String

Write

Other data types are converted to string type by default.

Format of writing Kafka value	Internal Type	Target end Field Type
JSON	Long	Integer、Bigint
	Double	Float、Double、Decimal
	Bool	Boolean
	String	String、Varchar、Array
	Date	Date、Timestamp

©2013–2025 Tencent Cloud. All rights reserved.

CSV	Long	Integer、Bigint
	Double	Float、Double、Decimal
	Bool	Boolean
	String	String、Varchar、Array
	Date	Date、Timestamp
AVRO	Long	Integer、Bigint
	Double	Float、Double、Decimal
	Bool	Boolean
	String	String、Varchar、Array
	Date	Date、Timestamp

Kafka/CKafka script demo

If you configure offline tasks using **script mode**, you need to write the **reader parameters** and **writer parameters** in the task script according to the unified script format requirements.

```
"job": {
    "content": [
        {
            "reader": {
                "parameter": {
                    "server": "ip:9092",
                    "endOffset": "9223372036854775807", //Cycle End Point
                    "kafkaConfig": {
                        "group.id": "WeData_group_1", //Consumer Group ID
                        "client.id": "WeData_client"
                    },
                    "messageFormat": "json", //serialization format
                    "column": [
                        "__key__",
                        "__value__",
                        "__partition__",
                        "__headers__",
                        "__offset__",
                        "__timestamp__"
                    ],
                    "encoding": "UTF-8",
                    "stopWhenPollEmpty": true,
                    "beginOffset": "seekToLast", //cycle starting point
                    "valueType": "STRING",
                    "skipExceedRecord": "true",
                    "topic": "source_topic", //source topic
                    "keyType": "STRING",
                    "waitTime": 120
                },
                "name": "kafkareader"
            },
            "transformer": [],
            "writer": {
                "parameter": {
                    "server": "ip:9092",
                    "messageFormat": "json", //serialization format
                    "kafkaConfig": {
```

©2013–2025 Tencent Cloud. All rights reserved.

```
                        "client.id": "WeData_client"
                    },
                    "partitioner": "roundRobin", // partition mapping method, supports polling, specify
field hash, and specify partition
                    "column": [
                        {
                            "name": "key",
                            "type": "string"
                        },
                        {
                            "name": "value",
                            "type": "string"
                        },
                        {
                            "name": "partition",
                            "type": "string"
                        },
                        {
                            "name": "headers",
                            "type": "string"
                        },
                        {
                            "name": "offset",
                            "type": "string"
                        },
                        {
                            "name": "timestamp",
                            "type": "string"
                        }
                    ],
                    "topic": "sink_topic" // target topic
                },
                "name": "kafkawriter"
            }
        }
    ],
    "setting": {
        "errorLimit": { // Dirty data threshold
            "record": 0
        },
        "speed": {
            "byte": -1, // No limit on sync speed, positive integer indicates maximum transfer speed in
byte/s
            "channel": 1 // Concurrency limit
        }
    }
}
```

©2013–2025 Tencent Cloud. All rights reserved.

Conversion Node Configuration

Last updated: 2024-12-20 11:06:39

Offline synchronization supports three types of data nodes: input, output, and conversion. This article will focus on the configuration of each read, write, and conversion node and the underlying data type mappings.

Read-Write Node Configuration

Database Type	Read node	Write node
Relational Database	MySQL Offline Node (Read)	MySQL Offline Node (Write)
	TDSQL-C MySQL Offline Node (Read)	TDSQL-C MySQL Offline Node (Write)
	PostgreSQL Offline Node (Read)	PostgreSQL Offline Node (Write)
	TCHouse-P Offline Node (Read)	TCHouse-P Offline Node (Write)
	TDSQL-PostgreSQL Offline Node (Read)	TDSQL-PostgreSQL Offline Node (Write)
	SQL Server Offline Node (Read)	SQL Server Offline Node (Write)
	Oracle Offline Node (Read)	Oracle Offline Node (Write)
	DB2 Offline Node (Read)	DB2 Offline Node (Write)
	DM Offline Node (Read)	DM Offline Node (Write)
	SAP HANA Offline Node (Read)	SAP HANA Offline Node (Write)
	SAP IQ (sybaseIQ) Offline Node (Read)	-
Big data	HIVE Offline Node (Read)	Hive Offline Node (Write)
	HBase Offline Node (Read)	HBase Offline Node (Write)
	Clickhouse Offline Node (Read)	Clickhouse Offline Node (Write)
	DLC Offline Node (Read)	DLC Offline Node (Write)
	Kudu Offline Node (Read)	Kudu Offline Node (Write)
	HDFS Offline Node (Read)	HDFS Offline Node (Write)
	Greenplum Offline Node (Read)	Greenplum Offline Node (Write)
	GaussDB Offline Node (Read)	GaussDB Offline Node (Write)
	Gbase Offline Node (Read)	Gbase Offline Node (Write)
	TBase Offline Node (Read)	TBase Offline Node (Write)
	Iceberg Offline Node (Read)	Iceberg Offline Node (Write)
	StarRocks Offline Node (Read)	StarRocks Offline Node (Write)
	Impala Offline Node (Read)	Impala Offline Node (Write)
	Doris Offline Node (Read)	Doris Offline Node (Write)
Message list	Kafka Offline Node (Read)	Kafka Offline Node (Write)
NoSQL	Mongo Offline Node (Read)	Mongo Offline Node (Write)
	Elasticsearch Offline Node (Read)	Elasticsearch Offline Node (Write)
	-	Redis Offline Node (Write)

©2013-2025 Tencent Cloud. All rights reserved.

Semi-structured	COS Offline Node (Read)	COS Offline Node (Write)
	FTP Offline Node (Read)	FTP Offline Node (Write)
	SFTP Offline Node (Read)	SFTP Offline Node (Write)
	RESTful APIs Offline Node (Read)	–

Conversion Node Configuration

String Replacement

The String Replacement Node primarily uses Java regular expressions to match and replace the contents of string fields. This node will perform the replacement operation when a string is partially or fully matched.
The conversion rule parameters are described as follows:

Parameters	Description
Field	Select the fields that need string matching. The fields processed by this node must be of string type. **Note** In this conversion node, multiple conversion rules can be configured for the same field. These rules will be executed serially based on the configuration order.
Regular Expression	Java regular expression used to match the original field content.
New string	Replace the matched field content with the specified new string.
Replacement method	Supports replacing only the first match result or all matches.
Delete	Delete this rule.

Time Format Conversion

The Time Format Conversion node is mainly used to convert time fields in the table to the target format. For example, convert the yyyymmdd time format to yyyymmdd hh:mm:ss format.
The conversion rules are explained as follows:

Parameters	Description
Field	Select the field that needs time format matching. The field should be of types like datetime, time, timestamp, date, etc. **Note** In this conversion node, multiple conversion rules can be configured for the same field.
Source Time Format	Specify the time format for this field. If the specified format does not match the actual format, it will not be converted and the data will be treated as dirty data.
Target Time Format	For each record, if the current field matches the specified source and target formats, the system will convert the source time format to the target format.
Delete	Delete this rule.

Note
Currently, this node only supports Offline Synchronization Tasks.

©2013–2025 Tencent Cloud. All rights reserved.

Value Conversion

The Value Conversion node is mainly used to standardize the content of fields. This node will execute the replacement operation when the field content is exactly equal to the specified content. After replacement, the original field content will be updated. For example, "CHINA" will be uniformly replaced with "CHN".
The parameters of the conversion rule are described as follows:

Parameters	Description
Field	Select the field that needs time format matching. The field is by default sourced from upstream fields and supports processing all upstream fields. **Note** In this conversion node, multiple conversion rules can be configured for the same field.
Matching Value	Used to determine if the original field is equal to this specified value. If equal, the replacement behavior will replace the value.
Replacement Value	The target for value conversion, cannot be empty.
Delete	Delete this rule.

Note
Currently, this node only supports Offline Synchronization Tasks.

©2013–2025 Tencent Cloud. All rights reserved.

Time Parameter Description

Last updated: 2024-08-24 13:30:39

For periodically running offline tasks, the system supports using time parameters to automatically obtain the data time of periodic task instances.
Time parameters are defined using ${…}, supporting formats such as ${yyyyMMdd}, ${yyyy-MM-dd}, ${HH:mm:ss}, and ${yyyyMMddHHmmss}. Here, yyyy represents the 4-digit year, yy represents the 2-digit year, MM represents the month, dd represents the day, HH represents the hour, mm represents the minute, and ss represents the second. Flexible combinations of these parts are supported.
The system's built-in parameter ${timestamp} corresponds to the 10-digit timestamp of the scheduling time, with second-level precision.
Below is a sample:
Using time 20210710080000 as a reference

Time	Time Parameter Format	Remarks
Next N Years	${yyyyMMdd+Ny}	• If referencing dt=${yyyyMMdd-1M}, it will be replaced with: dt=20210610 • If referencing dt=${yyyyMMdd-1d}, it will be replaced with: dt=20210709 • If referencing time=${yyyyMMddHHmmss-3h}, it will be replaced with: time=20210710050000 • If referencing ti=${yyyyMMddHHmmss-25m}, it will be replaced with: ti=20210710073500
Previous N Years	${yyyyMMdd-Ny}	
Next N Months	${yyyyMMdd+NM}	
Previous N Months	${yyyyMMdd-NM}	
Next N Weeks	${yyyyMMdd+Nw}	
Previous N Weeks	${yyyyMMdd-Nw}	
N days later	${yyyyMMdd+Nd}	
N days before	${yyyyMMdd-Nd}	
N hours later	${yyyyMMddHHmmss+NH}	
N hours before	${yyyyMMddHHmmss-NH}	
N minutes later	${yyyyMMddHHmmss+Nm}	
N minutes before	${yyyyMMddHHmmss-Nm}	
Scheduled Timestamp	${timestamp}	${timestamp}=1625875200 Supports using "+","-" and other operators
Scheduled Timestamp (13 digits)	${timestamp_ms}	${timestamp_ms}=1698508800000 Supports using "+","-" and other operators.

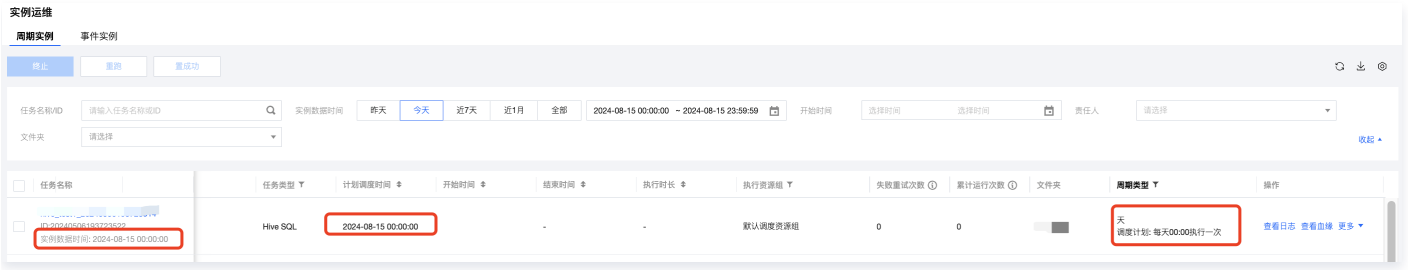
Provides quick conversion expressions for common dates, as follows:

Serial number	Time Parameter Format	Description
1	${yyyyMMdd+TE}	TENDAY END The data date corresponds to the end of ten days
2	${yyyyMMdd+ME}	MONTH END The data date corresponds to the end of the month

©2013-2025 Tencent Cloud. All rights reserved.

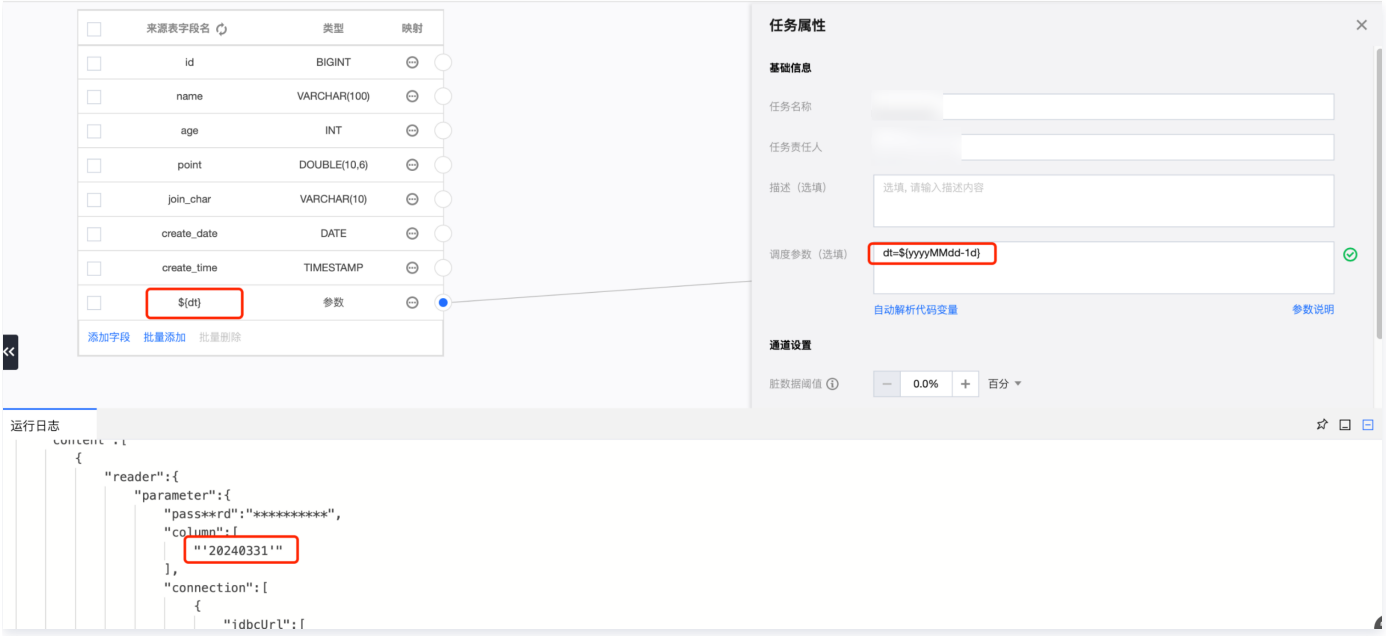
3	${yyyyMMdd+QE}	QUARTER END The data date corresponds to the end of the quarter
4	${yyyyMMdd+HYE}	HALF YEAR END The data date corresponds to the end of half a year
5	${yyyyMMdd+YE}	YEAR END The data date corresponds to the end of the year
6	${yyyyMMdd+TS}	TENDAY START The data date corresponds to the beginning of a ten-day period
7	${yyyyMMdd+MS}	MONTH START The data date corresponds to the beginning of the month
8	${yyyyMMdd+HYS}	HALFYEAR START The data date corresponds to the beginning of the half-year
9	${yyyyMMdd+YS}	YEAR START The data date corresponds to the beginning of the year
10	${yyyyMMdd+PME}	PRI MONTH END The data date corresponds to the end of last month
11	${yyyyMMdd+PYE}	PRI YEAR END The data date corresponds to the end of the previous year

Description:
wedata The time parameter here is automatically obtained from the periodic task "instance data time." Be careful not to confuse it with "planned scheduling time":
"Planned scheduling time" is the expected scheduling run time for the periodic instance. It can be configured in the **Development page** > **Scheduling Settings**. "Instance data time" corresponds to the periodic instance scheduling period. Once the scheduling period is determined, the "instance data time" is fixed. For example, for a daily periodic task configured, the instance data time on 20240815 would be 2024-08-15 00:00:00. The planned scheduling time can be configured to any time between 2024-08-15 00:00:00 and 2024-08-15 23:59:00 (accurate to the minute). Please refer to the figure below.



Use placeholders to replace time parameters

Use the placeholder `$dt` to replace time parameters. For example, `dt=${yyyyMMdd-1d}` denotes the time parameter.



©2013-2025 Tencent Cloud. All rights reserved.

Offline node advanced parameters

Last updated: 2024-12-12 20:15:43

©2013-2025 Tencent Cloud. All rights reserved.

Parameter Description

Offline type	Read/Write	Configuration contents	Applicable Scenario	Description
MySQL	Read	splitFactor=5	Single table	–
		ignoreRegex=true	Single table	After adding this parameter, the selected or entered table name containing special characters will be used as the actual table name for data synchronization without regex matching. If this parameter is not added, regex matching will be performed on table names containing special characters by default. Note: You need to manually enter tables with special characters and add backticks or double quotes to the table name according to the database's characteristics.
TDSQL MySQL	Read	splitFactor=5	Single table	–
		ignoreRegex=true	Single table	After adding this parameter, the selected or entered table name containing special characters will be used as the actual table name for data synchronization without regex matching. If this parameter is not added, regex matching will be performed on table names containing special characters by default. Note: You need to manually enter tables with special characters and add backticks or double quotes to the table name according to the database's characteristics.
TDSQL-C MySQL	Read	ignoreRegex=true	Single table	After adding this parameter, the selected or entered table name containing special characters will be used as the actual table name for data synchronization without regex matching. If this parameter is not added, regex matching will be performed on table names containing special characters by default. Note: You need to manually enter tables with special characters and add backticks or double quotes to the table name according to the database's characteristics.
Doris	Read	query_timeout=604800	Single table	Query timeout period in seconds
		exec_mem_limit=4294967296	Single table	Set execution memory limit to restrict memory usage during query execution
		parallel_fragment_exec_instance_num=8	Single table	Specify the number of instances executing parallel fragments
	Write	sameNameWildcardColumn=true	Single table	MySQL-Doris configuration* supports same-name field mapping
		loadProps={"format":"csv","column_separator":"\\x01","row_delimiter":"\\x03"}	Single table	Write in CSV format. It has higher performance compared to the default JSON format. It needs to be used with the row delimiter \x03. The delimiter does not support \r, \b, and hidden characters in hexadecimal need to be input in the form of \\.

©2013-2025 Tencent Cloud. All rights reserved.

		replaceDelimiter=Replacement character	Single table	If the data item contains characters used as delimiters, separating data by delimiters during import will cause field misalignment. Adding this parameter can replace characters in the data item that match the specified row or column delimiter with user-defined characters (the replacement operation will change the data), and then separate the data by row or column delimiter. Without this parameter, the data is directly separated by the delimiter.
Hive	Read	mapreduce.job.queuename=root.default	Single table	Specify which queue the job is submitted to for execution
		hive.execution.engine=mr	Single table	A configuration parameter in Hive to specify the Hive query execution engine. The current default is mr; no need to modify.
		Note: When setting advanced parameters, mapreduce.job.queuename=root.default and hive.execution.engine=mr must be used together; using a single parameter will not be effective.		
DLC	Read	fs.cosn.trsf.fs.ofs.data.transfer.thread.count=8	Single table	DLC concurrent writing supports the parameters none|hash|range. Parameter descriptions: 1. none: If there is a primary key, concurrent writing is based on the primary key; otherwise, single concurrent writing is used 2. hash: If there is a partition field, concurrent writing is based on the partition field; otherwise, it follows the none parameter strategy 3. range: Currently not supported; strategy is the same as none
		fs.cosn.trsf.fs.ofs.prev.read.block.count=4	Single table	DLC enables small file merging, which can also be enabled in the DLC console. This parameter is false by default. There is an option to enable it in the full database synchronization interface, and for single table synchronization, this parameter needs to be configured manually.
Mongodb	Read	batchSize=1000	Single table	Number of entries for batch reading
COS	Read	ignoreDirEmpty=true	Single table	After adding this parameter, the task will not report an error if there are no files or only empty files in the source file path in data synchronization mode. If this parameter is not added, the default is ignoreDirEmpty=false, and the task will report an error if there are no files or only empty files in the source file path.
		ignoreFileNotFound=true	Single table	After adding this parameter, the task will not report an error if the source file or file path does not exist in data synchronization or file transfer mode. If this parameter is not added, the default is ignoreFileNotFound=false, and the task will report an error if the source file or file path does not exist.
		transEmpFile=true	Single table	After adding this parameter, empty files can be synchronized to the FTP/SFTP/HDFS/S3/COS

©2013-2025 Tencent Cloud. All rights reserved.

				destination in file transfer mode. If this parameter is not added, the default is transEmpFile=false, and empty files will not be synchronized in file transfer mode.
	Write	splitFileSize=134217728	Single table	• Single file chunk size • Not effective for hive on COS • Supports text, ORC, Parquet types of files
		hadoopConfig={}	Single table	Supports adding configurations to hadoopConfig
		saveEmpFile=true	Single table	After adding this parameter, empty files generated at the destination will be retained if there is no data at the source in data synchronization mode. If this parameter is not added, the default is saveEmpFile=false, and empty files generated at the destination will not be retained if there is no data at the source in data synchronization mode.
HDFS	Read	ignoreDirEmpty=true	Single table	After adding this parameter, the task will not report an error if there are no files or only empty files in the source file path in data synchronization mode. If this parameter is not added, the default is ignoreDirEmpty=false, and the task will report an error if there are no files or only empty files in the source file path.
		ignoreFileNotFound=true	Single table	After adding this parameter, the task will not report an error if the source file or file path does not exist in data synchronization or file transfer mode. If this parameter is not added, the default is ignoreFileNotFound=false, and the task will report an error if the source file or file path does not exist.
		transEmpFile=true	Single table	After adding this parameter, empty files can be synchronized to the FTP/SFTP/HDFS/S3/COS destination in file transfer mode. If this parameter is not added, the default is transEmpFile=false, and empty files will not be synchronized in file transfer mode.
	Write	splitFileSize=134217728	Single table	• Single file chunk size • Not effective for hive on HDFS • Supports text, ORC, Parquet types of files
		writeMode=overwrite	Single table	This parameter applies to file transfer mode. After adding this parameter, files with the same name at the destination as the source will be cleaned before writing. If this parameter is not added, the default in file transfer mode is writeMode=nonconflict, which skips writing when files with the same name are encountered at the source and destination.
		saveEmpFile=true	Single table	After adding this parameter, empty files generated at the destination will be retained if there is no data at the source in data synchronization mode. If this parameter is not added, the default is saveEmpFile=false, and empty files generated at the destination will not be retained if there is no data at the source in data synchronization mode.
SFTP/FTP	Read	ignoreFileNotFound=true	Single table	After adding this parameter, the task will not report an error if the source file or file path does not exist

©2013–2025 Tencent Cloud. All rights reserved.

				in data synchronization or file transfer mode. If this parameter is not added, the default is ignoreFileNotFound=false, and the task will report an error if the source file or file path does not exist.
		transEmpFile=true	Single table	After adding this parameter, empty files can be synchronized to the FTP/SFTP/HDFS/S3/COS destination in file transfer mode. If this parameter is not added, the default is transEmpFile=false, and empty files will not be synchronized in file transfer mode.
	Write	deleteEmpFile=true	Single table	After adding this parameter, empty files generated at the destination will not be retained if there is no data at the source in data synchronization mode. If this parameter is not added, the default is deleteEmpFile=false, and empty files generated at the destination will be retained if there is no data at the source in data synchronization mode.
Hive	Write	compress=none/snappy/lz4/bzip2/gzip/deflate	Single table	Default is none. Effective only for textfile format, not for ORC/Parquet (ORC/Parquet requires specifying compression in the create table statement)
		format=orc/parquet	Single table	The format of HDFS temporary files defaults to ORC and is unrelated to the final Hive table format
		partition=static	Single table	Static partitioning mode. Suitable for single partition writing, saving more memory
Doris	Write	sameNameWildcardColumn=true	Single table	MySQL-Doris configuration* supports same-name field mapping.
	Write	loadProps={"format":"csv","column_separator":"\\x01","row_delimiter":"\\x03"}	Single table	CSV format writing. Higher performance compared to the default JSON format. Needs to be used with row separator \\x03.
DLC	Write	fs.cosn.trsf.fs.ofs.data.transfer.thread.count=8	Single table	DLC concurrent writing supports the parameters none|hash|range. Parameter descriptions: 1. none: If a primary key exists, concurrent writing is based on the primary key; otherwise, single concurrent writing 2. hash: If a partition field exists, concurrent writing is based on the partition field; otherwise, writing follows the 'none' parameter strategy 3. range: Currently not supported; strategy is the same as none
		fs.cosn.trsf.fs.ofs.prev.read.block.count=4	Single table	DLC enables small file merging, which can also be enabled in the DLC console. This parameter is false by default. There is an option to enable it in the full database synchronization interface, and for single table synchronization, this parameter needs to be configured manually.
StarRocks	Read	splitFactor=5	Single table	-
		query_timeout=604800	Single table	Query timeout period in seconds
		exec_mem_limit=4294967296	Single table	Set execution memory limit to restrict memory usage during query execution

©2013-2025 Tencent Cloud. All rights reserved.

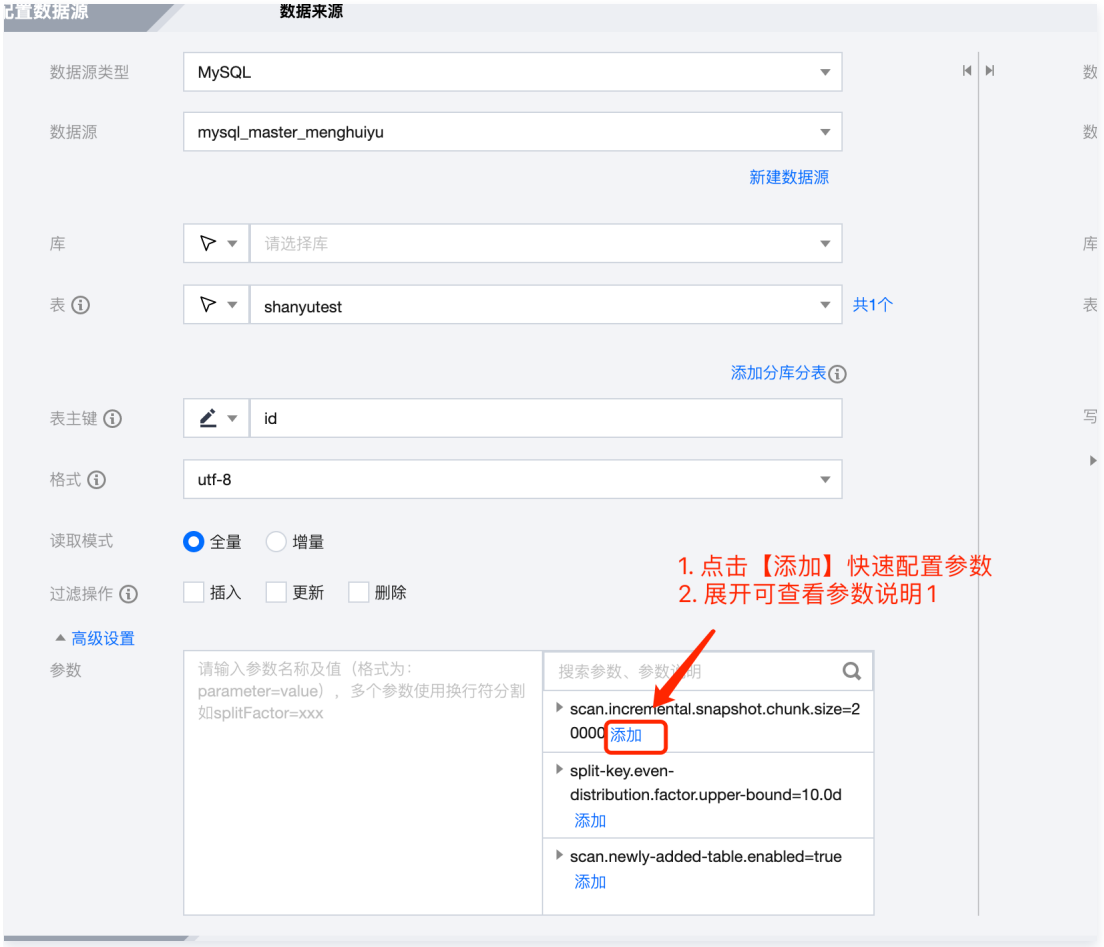
		parallel_fragment_exec_instance_num=8	Single table	Specify the number of instances executing parallel fragments
	Write	loadProps={"format":"csv","column_separator":"\\x01","row_delimiter":"\\x03"}	Single table	CSV format writing. Higher performance compared to the default JSON format. Needs to be used with row separator \x03. Separators do not support \r, \b, and hexadecimal hidden characters need to be input in the form of \\.
		replaceDelimiter=Replacement character	Single table	If the data item contains characters used as delimiters, separating data by delimiters during import will cause field misalignment. Adding this parameter can replace characters in the data item that match the specified row or column delimiter with user-defined characters (the replacement operation will change the data), and then separate the data by row or column delimiter. Without this parameter, the data is directly separated by the delimiter.
Mongodb	Write	replaceKey=id	Single table	When the write mode is overwrite, this is used as the business primary key for updates
		batchSize=2000	Single table	The number of entries for batch writes. If not set, the default is 1,000
Elasticsearch	Write	compression=true	Single table	HTTP request, enable compression
		multiThread=true	Single table	HTTP request, whether multi-threading is enabled
		ignoreWriterError	Single table	Ignore write errors, do not retry, continue writing
		ignoreParseError=false	Single table	Ignore data format parsing errors, continue writing
		alias	Single table	The alias in Elasticsearch is similar to the view mechanism in databases. Create an alias 'my_index_alias' for the index 'my_index'. Operations on 'my_index_alias' are the same as on 'my_index'. Configure 'alias' to create an alias for the specified index after data import completes.
		aliasMode=append	Single table	The mode of adding an alias after data import includes 'append' (add mode) and 'exclusive' (only keep this one). 'Append' adds the current index to the alias mapping (one alias corresponds to multiple indices), 'exclusive' first deletes the alias, then adds the current index to the alias mapping (one alias corresponds to one index). Subsequently, the alias will be converted to the actual index name. An alias can be used for index migration, unified queries across multiple indices, and can implement the view feature.
		nullToDate=null	Single table	Convert null values to date type, filling null
Kafka	Write	kafkaConfig={}	Single table	Supports Kafka Producer configuration item

Metadata Field	Read/Write	Configuration contents

©2013-2025 Tencent Cloud. All rights reserved.

Kafka	Read	• __key__ represents the key of the message • __value__ represents the full content of the message • __partition__ represents the partition of the current message • __headers__ represents the headers information of the current message • __offset__ represents the offset of the current message • __timestamp__ indicates the timestamp of the current message.
Elasticsearch	Read	• _id supports obtaining _id information

Configuration Method



©2013–2025 Tencent Cloud. All rights reserved.

Automatic Table Creation Capability in Synchronization Tasks

Last updated: 2024-12-12 20:12:57

DataInLong provides features such as One-Click Table Creation and Automatic Table Creation in both offline and real-time scenarios, helping you quickly construct target data tables before or during synchronization, thereby improving the efficiency of data link usage. This document mainly introduces the usage methods of the corresponding features, supported links, and the core type conversion relationships between heterogeneous data sources.

Feature Introduction and Usage Scenarios

DataInLong offers three table creation methods: one-click table creation, batch table creation, and automatic table creation based on DDL messages:

Feature Items	Feature Illustration	Feature Introduction	Use Cases
One-Click Table Creation Batch Table Creation	 	This feature can automatically complete the DDL conversion from the source to the target end based on multiple table objects specified by the source end. It supports users to manually modify the converted DDL content and create them in batches at once. **Feature Keywords: Automatic Heterogeneous DDL Conversion, Batch Table Creation, Manual Table Creation.**	This feature is suitable for quickly constructing target tables similar to the source table structure in the case of inventory at the source end. For complex table models that require business customization, users can edit the automatically generated DDL statements based on business characteristics, improving the efficiency of heterogeneous DDL writing.
Automatic Table Creation Based on Source End DDL Messages		This feature can monitor in real-time whether there are new table DDL messages at the source end. Once identifiable new table messages are generated at the source end, the target end can automatically generate target table objects based on the source table structure. **Feature Keywords: Automatic DDL Perception and Response, Automatic Table Creation.**	This feature is suitable for data synchronization scenarios where it is expected that the target end and the source end maintain real-time structural consistency during real-time database relocation.

Links supporting One-Click Table Creation

Offline Single Table Tasks and Real-time Single Table Tasks support one-click creation of target tables based on source tables, mainly for scenarios involving the addition of a small number of database tables for ingestion and warehousing.

Offline Single Table Sync Task

Target Index	Target	Supported source ends
1	Doris	MySQL, TencentDB for MySQL, TDSQL-C for MySQL, DLC, SQL Server, TDSQL for MySQL, PostgreSQL, TDSQL for PostgreSQL, Oracle, Oceanbase, Dameng
2	TChouse-D	MySQL, TencentDB for MySQL, TDSQL-C for MySQL, DLC, SQL Server, TDSQL for MySQL, PostgreSQL, TDSQL for PostgreSQL, Oracle, Oceanbase, Dameng

©2013-2025 Tencent Cloud. All rights reserved.

3	Hive	MySQL, TencentDB for MySQL, TDSQL for MySQL, TDSQL-C for MySQL, Oracle, SQL Server, PostgreSQL, Oceanbase, Dameng
4	Iceberg	MySQL, TencentDB for MySQL, TDSQL-C for MySQL, TDSQL for MySQL, Oracle, SQL Server, PostgreSQL, Oceanbase, Dameng
5	TCHouse-P	MySQL, TencentDB for MySQL, TDSQL-C for MySQL, TDSQL for MySQL, Oracle, SQL Server, PostgreSQL, Oceanbase, Dameng
6	DLC	MySQL, TencentDB for MySQL, Elasticsearch, TDSQL for MySQL, TDSQL-C for MySQL, TencentDB for Elasticsearch, Kafka, PostgreSQL, TDSQL for PostgreSQL, Oracle, SQL Server, Oceanbase, Dameng

Real-time single-table synchronization tasks

Target Index	Target	Supported source ends
1	Doris	MySQL, TencentDB for MySQL, TDSQL-C for MySQL, SQL Server, TDSQL for MySQL, PostgreSQL, Oracle, Oceanbase, Dameng
2	TChouse-D	MySQL, TencentDB for MySQL, TDSQL-C for MySQL, SQL Server, TDSQL for MySQL, PostgreSQL, Oracle, Oceanbase, Dameng
3	Hive	MySQL, TencentDB for MySQL, TDSQL for MySQL, TDSQL-C for MySQL, Oracle, SQL Server, PostgreSQL, Oceanbase, Dameng
4	Iceberg	MySQL, TencentDB for MySQL, TSQL-C for MySQL, TDSQL for MySQL, Oracle, PostgreSQL, SQL Server, Oceanbase, Dameng
5	TCHouse-P	MySQL, TencentDB for MySQL, TDSQL-C for MySQL, TDSQL for MySQL, Oracle, PostgreSQL, SQL Server, Oceanbase, Dameng
6	DLC	MySQL, TencentDB for MySQL, TDSQL-C for MySQL, TDSQL for MySQL, PostgreSQL, Oracle, SQL Server, Kafka, Oceanbase, Dameng

Batch Table Creation Supported Link

Offline Batch Creation of Single Table Tasks and Real-time Whole Database Tasks support batch creation of Target Tables based on Source Tables during the task configuration phase. This is mainly for scenarios involving the Ingestion and Warehousing of large numbers of Existing Database Tables.

Offline Batch Sync Task

Target Index	Target	Supported source ends
1	Doris	MySQL, TencentDB for MySQL, TDSQL-C for MySQL, DLC, SQL Server, TDSQL for MySQL, PostgreSQL, TDSQL for PostgreSQL, Oracle, Oceanbase, Dameng
2	TChouse-D	MySQL, TencentDB for MySQL, TDSQL-C for MySQL, DLC, SQL Server, TDSQL for MySQL, PostgreSQL, TDSQL for PostgreSQL, Oracle, Oceanbase, Dameng
3	Hive	MySQL, TencentDB for MySQL, TDSQL for MySQL, TDSQL-C for MySQL, Oracle, SQL Server, PostgreSQL, Oceanbase, Dameng
4	Iceberg	MySQL, TencentDB for MySQL, TDSQL-C for MySQL, TDSQL for MySQL, Oracle, SQL Server, PostgreSQL, Oceanbase, Dameng
5	TCHouse-P	MySQL, TencentDB for MySQL, TDSQL-C for MySQL, TDSQL for MySQL, Oracle, SQL Server, PostgreSQL, Oceanbase, Dameng
6	DLC	MySQL, TencentDB for MySQL, Elasticsearch, TDSQL for MySQL, TDSQL-C for MySQL, TencentDB for Elasticsearch, Kafka, PostgreSQL, TDSQL for PostgreSQL, Oracle, SQL Server, Oceanbase, Dameng

©2013-2025 Tencent Cloud. All rights reserved.

Real-time Full Database Migration Task

Target Index	Target	Supported source ends
1	Doris	MySQL、TDSQL-C MySQL、TDSQL MySQL
2	DLC	MySQL、TDSQL-C MySQL、TDSQL MySQL

Supports Link for Automatic Table Creation Based on Source End DDL Messages

Except for the following links, other real-time full database links support the creation of target databases and tables automatically during the task run phase based on source-side DDL messages:

No.	Source	Target	Remarks
1	MySQL	PostgreSQL	Only supports using existing databases, schemas, data tables
2	OceanBase	Doris、Iceberg	Only supports using existing databases, data tables
3	Kafka	Doris	Only supports using existing databases, data tables

Core heterogeneous data source table creation type conversion relationships

MySQL to Doris Table Creation Field Type Conversion:

	MySQL (Source End)		DORIS (Target End)		Additional Notes
	Data Type	Precision Description	Data Type	Precision Description	
Value Type	BOOLEAN	0 represents false, 1 represents true	BOOLEAN	0 represents false, 1 represents true	-
	TINYINT	Range [-128, 127]	TINYINT	Range [-128, 127]	-
	SMALLINT	Range [-32768, 32767]	SMALLINT	Range [-32768, 32767]	-
	MEDIUMINT	Range [-8388608, 8388607]	INT	Range [-2147483648, 2147483647]	-
	INT	Range [-2147483648, 2147483647]	INT	Range [-2147483648, 2147483647]	-
	BIGINT	Range [-9223372036854775808, 9223372036854775807]	BIGINT	Range [-9223372036854775808, 9223372036854775807]	-
	UNSIGNED TINYINT	Range [0, 255]	SMALLINT	Range [-32768, 32767]	-
	UNSIGNED MEDIUMINT	Range [0, 16777215]	INT	Range [-2147483648, 2147483647]	-

©2013-2025 Tencent Cloud. All rights reserved.

	UNSIGNED INT	Range [0, 4294967295]	BIGINT	Range [-9223372036854775808, 9223372036854775807]	-
	UNSIGNED BIGINT	Range [0, 18446744073709551615]	LARGEINT	Range [-2^127 + 1 ~ 2^127 - 1]	-
	FLOAT	4-byte floating point number	FLOAT	4-byte floating point number	-
	DOUBLE	8-byte floating point number	DOUBLE	8-byte floating point number	-
	DECIMAL	DECIMAL(M,D), M range [1, 65], D range [0, 30]	DECIMALV3	DECIMAL(M,D), M range [1, 38], D range [0, precision]	-
Datetime Type	YEAR	Range: 1901 to 2155 Display format: YYYY	SMALLINT	Range [-32768, 32767]	-
	TIME	Range: -838:59:59 to 838:59:59 Display format: hh:mm:ss or hh:mm:ss	STRING	Variable-length string, maximum (default) support 1048576 bytes (1MB)	-
	DATE	Range: 1000-01-01 to 9999-12-31 Display format: YYYY-MM-DD	DATEV2	Range: 0000-01-01 to 9999-12-31 Display format: YYYY-MM-DD	-
	DATETIME	1000-01-01 00:00:00 to 9999-12-31 23:59:59 Display format: YYYY-MM-DD HH:mm:ss	DATETIMEV2	0000-01-01 00:00:00 to 9999-12-31 23:59:59 Print format: YYYY-MM-dd HH:mm:ss.SSSSSS, time precision optional.	-
	TIMESTAMP	UTC 1970-01-01 00:00:01 to 2038-01-19 03:14:07 Display format: YYYY-MM-DD HH:mm:ss	DATETIMEV2	Display format: YYYY-MM-DD HH:mm:ss	TIMESTAMP field data changes with system time zone but DATETIME field data does not. It is recommended to perform time zone conversion according to business scenarios
String Type	CHAR	0 to 255 characters	CHAR	Fixed-length string, range is 1 - 255	-
	VARCHAR	0 to 65,535 characters	VARCHAR	Variable-length string, range is 1 - 65,533	If MySQL field length exceeds 65,533, it is recommended to convert to string

©2013-2025 Tencent Cloud. All rights reserved.

	TINYTEXT、TEXT	0 to 255 characters	STRING	Variable-length string, maximum (default) support 1048576 bytes (1MB)	-
	MEDIUMTEXT、LONGTEXT	0 to 65,535 characters	STRING	Variable-length string, maximum (default) support 1048576 bytes (1MB)	Precision loss may occur if MySQL field length exceeds 1,048,576 bytes
Binary Strings	TINYBLOB、BLOB	Binary string, 0 to 255 bytes	STRING	Variable-length string, maximum (default) support 1048576 bytes (1MB)	-
	MEDIUMBLOB、LONGBLOB	Binary string, 0 to 16,777,215 bytes, maximum 16M	STRING	Variable-length string, maximum (default) support 1048576 bytes (1MB)	Precision loss may occur if MySQL field length exceeds 1,048,576 bytes
	BINARY、VARBINARY	Fixed-length binary data, up to 255 bytes	STRING	Variable-length string, maximum (default) support 1048576 bytes (1MB)	-
Other	JSON	JSON Data, maximum storage size is 1GB	STRING	Variable-length string, maximum (default) support 1048576 bytes (1MB)	When the MySQL field size exceeds 1 MB, precision may be lost.
	SET、BIT	String collection, up to 64 members	STRING	Variable-length string, maximum (default) support 1048576 bytes (1MB)	-
	ENUM	Enumeration object, up to 65,535 members	UNSUPPORTED	-	Not supported.

MySQL to DLC Iceberg table field type conversion:

MySQL Type (Source)	DLC Iceberg Table (target end)	Description
tinyint(1)	int	-
smallint	smallint	-
int	int	-
mediumint	int	-
bigint	bigint	-
float	float	-
double	double	-
decimal	decimal	-
datetime	timestamp	-
timestamp	timestamp	-
date	date	-
time	time	-
tinytext	string	-

©2013-2025 Tencent Cloud. All rights reserved.

text	string	–
mediumtext	string	–
longtext	string	–
varchar	string	–
char	string	–
bool	boolean	–
tinyblob	binary	–
mediumblob	binary	–
blob	binary	–
longblob	binary	–
varbinary	binary	–
binary	binary	–
decimal unsigned(p,x)	decimal(p+1, x)	Description: decimal(p+1, x)/string (converted to string when over length)
decimal unsigned(p,x)	decimal(p+1, x)	Description: decimal(p+1, x)/string (converted to string when over length)
int unsigned	bigint	–
int unsigned zerofill	bigint	–
smallint unsigned	int	–
smallint unsigned zerofill	int	–
mediumint unsigned	bigint	–
mediumint unsigned zerofill	bigint	–
float unsigned	double	–
float unsigned zerofill	double	–
double unsigned	decimal(20,0)	–
double unsigned zerofill	decimal(20,0)	–
bigint unsigned	decimal(20,0)	–

©2013–2025 Tencent Cloud. All rights reserved.

Practice Tutorial
Sync MySQL Sharding to Hive

Last updated：2024-08-24 21:15:53

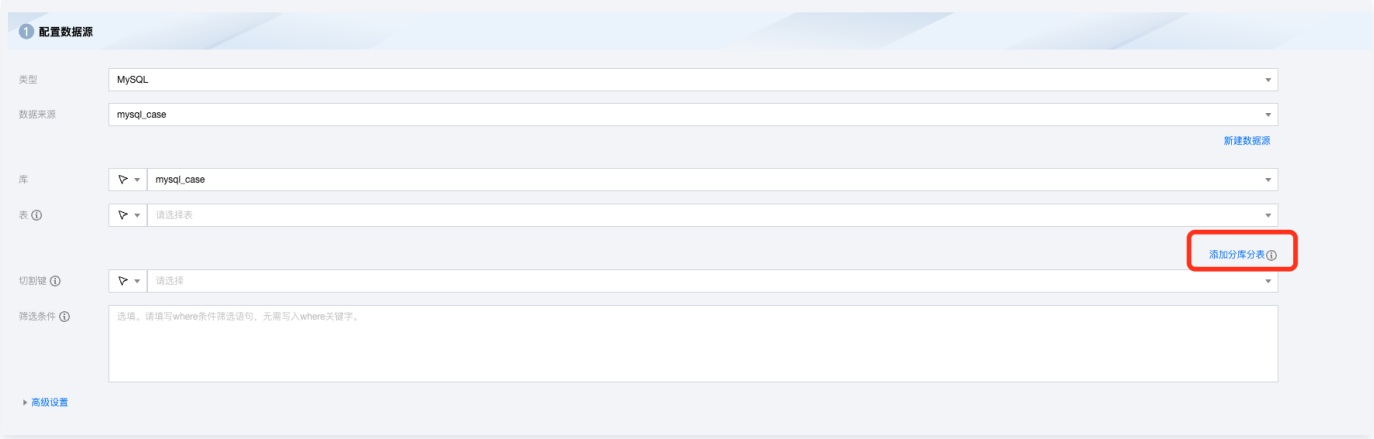
Currently, both offline and real-time data synchronization tasks support syncing MySQL sharding to the target data source. This practice introduces the offline synchronization of MySQL sharding to the Hive data source.

Applicable Scenario

The business layer stores basic structured business data using sharding in different MySQL database instances or different tables within the same instance. The application layer needs to synchronize tables distributed in different databases to the ODS layer in the data warehouse for unified storage.

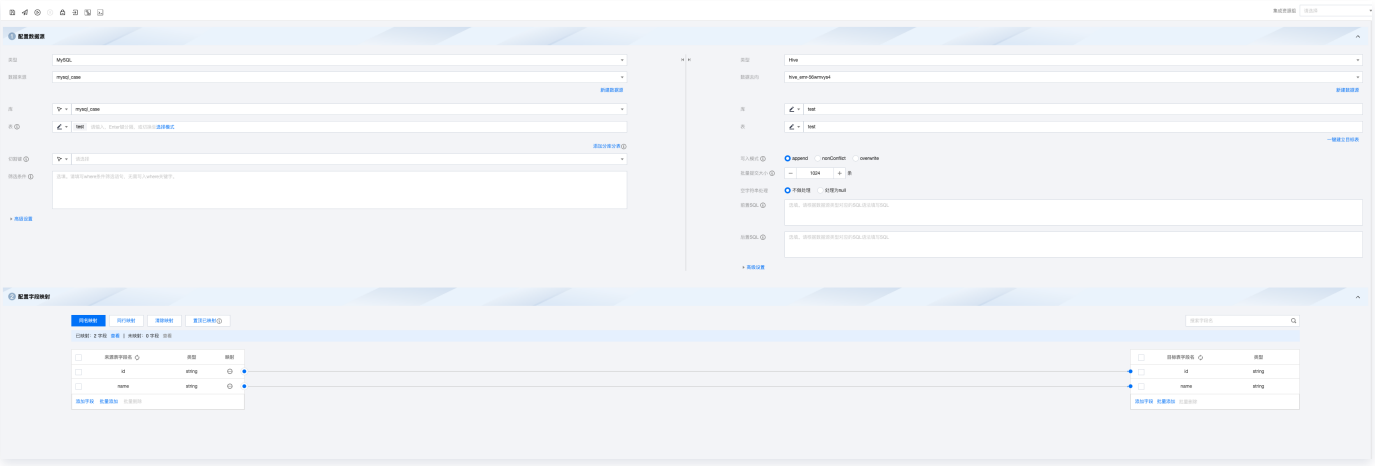
Operation step

1. Create an offline synchronization task, and click the corresponding task name from the offline synchronization list to enter the canvas configuration interface.
2. Drag and drop the MySQL read node. By default, the MySQL read node supports one data source. Multiple MySQL tables can be selected within a single data source (database). If there are sharding cases, you can click **Add Sharding** to create multiple data sources and add the corresponding table objects.



> **Note**
> The multiple table objects selected in sharding cases must ensure consistent Schema Information (including field names and field types). In the **Data Fields** module, the system defaults to display the Metadata Field Information of the first table in the first data source. Inconsistent fields across multiple tables may cause operation failure.

3. Drag and drop the Hive write node, and configure the Read and Write Field Mapping.



©2013-2025 Tencent Cloud. All rights reserved.

4. Save the **Task Information**, then perform a test run or submit it to the operations center.

SQL Server Offline Data Synchronization to Hive Partition

©2013–2025 Tencent Cloud. All rights reserved.

Last updated：2024-08-24 21:16:27

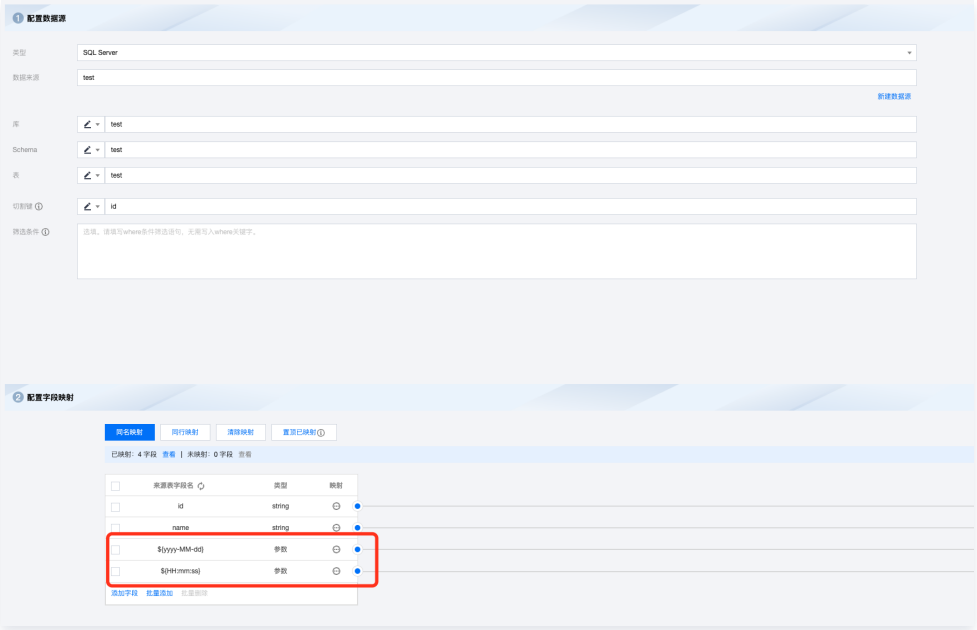
Periodic offline synchronization tasks support periodic updates of partition table data. This practice introduces SQL Server table data regular updates to Hive partitions.

Applicable Scenario

Periodically create or update Hive partition table data, and write the content of the data table into the corresponding first-level or multi-level partitions according to the periodic task instance schedule.

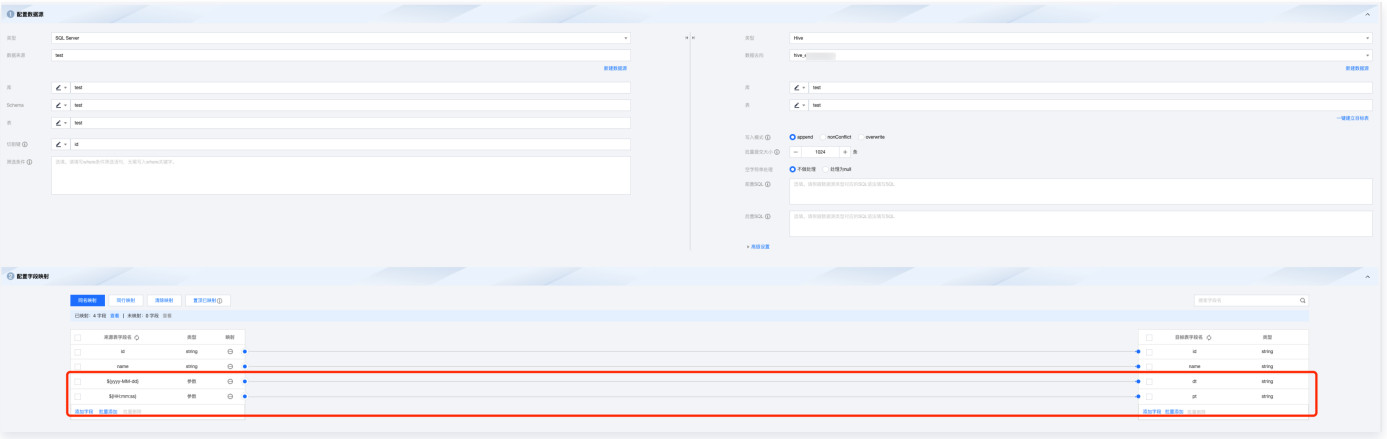
Operation step

1. Create an offline synchronization task and click on the corresponding task name from the offline synchronization list to enter the form configuration interface.
2. Configure the data source information and set the time parameter in the **Data Field** module.



> **Note**
> - For the definition method of time parameters, please refer to Time Parameter Description.
> - The time parameter type should be configured as "parameter".

3. Configure data source information and establish field mappings between time parameters and partition fields in the **Field Mapping** module.



4. Save the **Task Information**, then perform a test run or submit it to the operations center.

©2013-2025 Tencent Cloud. All rights reserved.

MySQL + gh-ost Real-time Synchronization to Kafka

Last updated: 2024-08-24 21:16:54

Business Scenario

- The business scenario for gh-ost is to perform online table structure changes in MySQL, specifically Online DDL, without affecting the normal operation of the business. It can resolve issues caused by traditional commands like alter table or create index, such as table locks, performance degradation, and synchronization delays. It is suitable for scenarios where table modifications are necessary, such as adding new columns, adding indexes, modifying column types, etc.
- This practice introduces using gh-ost to change a MySQL table and then synchronizing its DDL change records to Kafka in real-time.

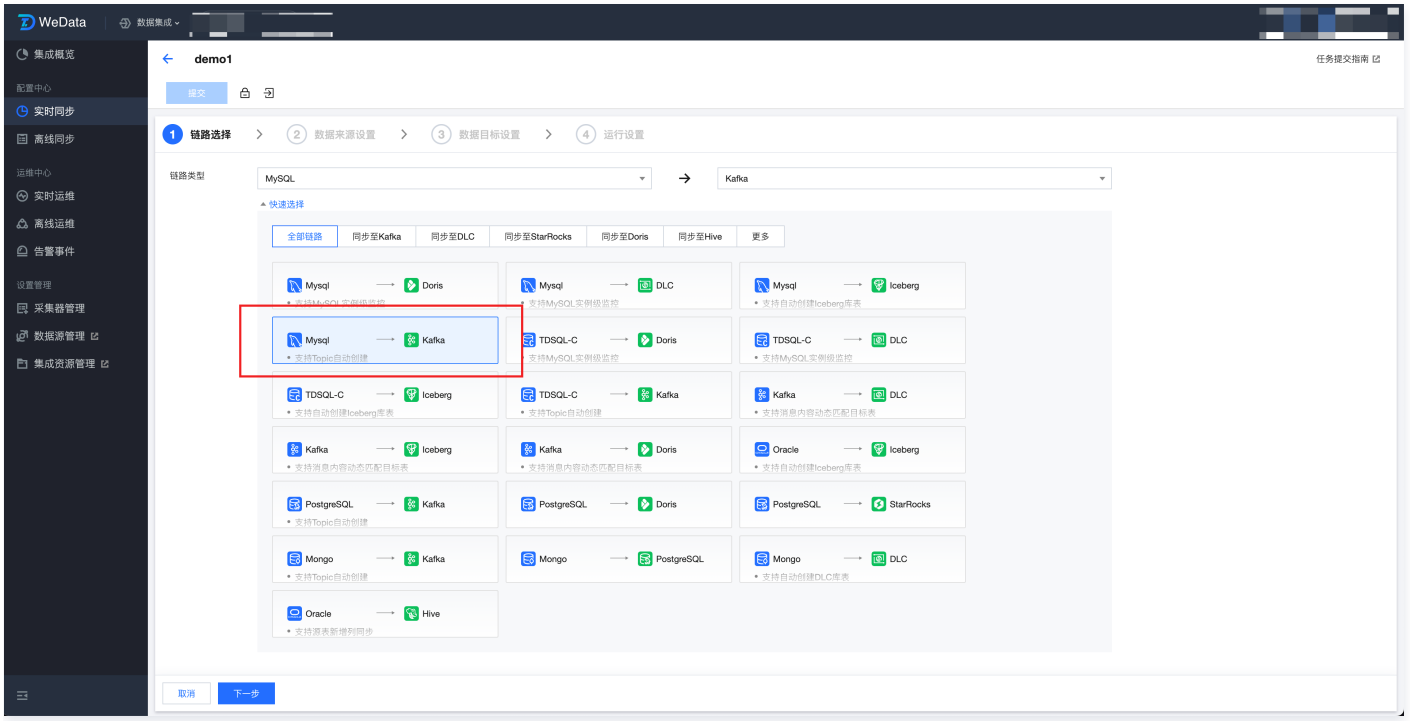
Prerequisites for using gh-ost

MySQL environment preparation: DataInLong MySQL environment preparation and database configuration.

1. gh-ost must be able to access MySQL.
2. If MySQL is Tencent Cloud's CDB, the `--aliyun-rds` parameter needs to be added to the gh-ost command execution.
3. During the execution of the gh-ost tool, the generated temporary table rules are `^_(.*)_(gho|ghc|del)$`, where (.*) is the name of the table being changed. Custom temporary table names are not supported.
4. For other gh-ost limitations, refer to: `gh-ost/requirements-and-limitations.md at master · github/gh-ost · GitHub`.

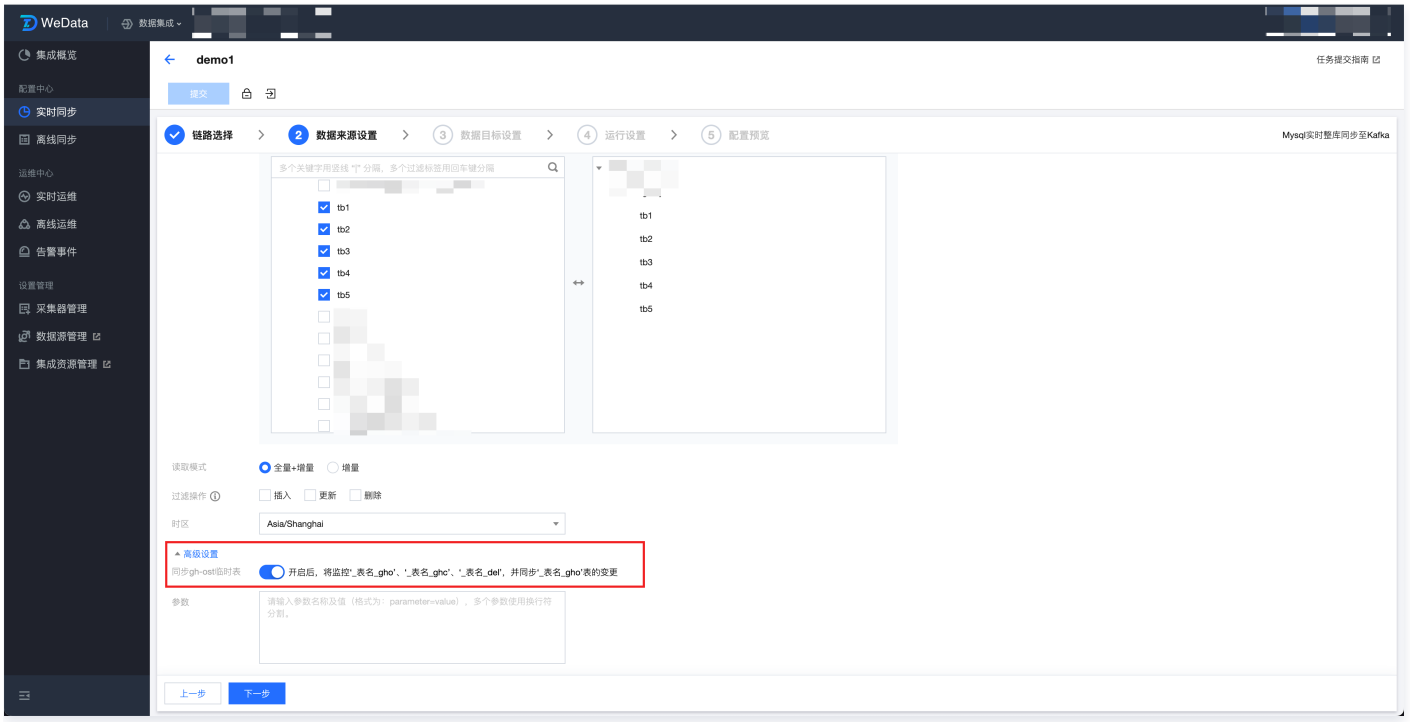
Operation step

1. Create a whole database synchronization task for real-time synchronization from MySQL to Kafka.

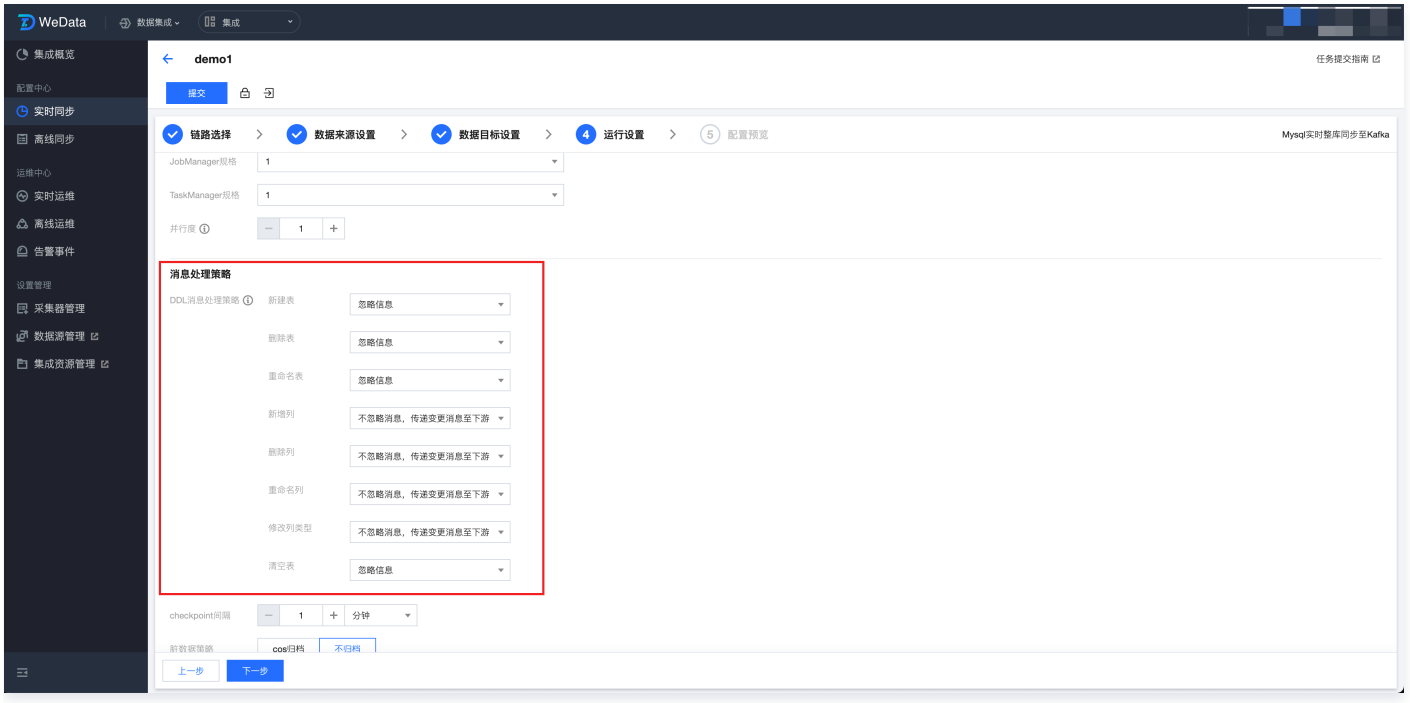


2. In the task configuration **Data Source Settings**, **Advanced Settings** enable the **Synchronize gh-ost Temporary Table** feature.

©2013-2025 Tencent Cloud. All rights reserved.



3. In the task configuration **Run Settings**, **Message Processing Strategy** select the required DDL Message Processing Strategy, here choose to add columns, delete columns, rename columns, and modify column types. After configuring the remaining options, submit the task.



4. Use gh-ost to change the original table.

   4.1 View the table structure.

```
mysql> desc tb1;
+-------+--------------+------+-----+---------+-------+
| Field | Type         | Null | Key | Default | Extra |
+-------+--------------+------+-----+---------+-------+
| id    | int          | NO   | PRI | NULL    |       |
| name  | varchar(255) | YES  |     | NULL    |       |
+-------+--------------+------+-----+---------+-------+
```

©2013-2025 Tencent Cloud. All rights reserved.

```
2 rows in set (0.00 sec)
```

4.2 Use the gh-ost tool to add a field to the source table.

- gh-ost Tool Download: Releases · github/gh-ost · GitHub
- After downloading, decompress and execute the following command to add the `c` field to the `tb1` table.

```
./gh-ost \
--max-load=Threads_running=25 \
--critical-load=Threads_running=1000 \
--chunk-size=1000 \
--throttle-control-replicas="" \
--max-lag-millis=1500 \
--user="root" \ // username
--password="test" \ // password
--host=127.0.0.1 \ // MySQL's IP
--allow-on-master \
--database="databaseName" \ // The name of the database where the table to be altered is located
--table="tb1" \ // The name of the table to be altered
--verbose \
--alter="engine=innodb" \
--switch-to-rbr \
--allow-master-master \
--exact-rowcount \
--concurrent-rowcount \
--default-retries=120 \
--panic-flag-file=/tmp/ghost.panic.flag \
--default-retries=120 \
--alter="add column c varchar(255);" \ // ALTER change statement
--approve-renamed-columns \
--initially-drop-ghost-table \
--initially-drop-old-table \
--ok-to-drop-table \
--aliyun-rds \ // When using Tencent Cloud CDB, this parameter needs to be added
--execute
```

4.3 View the modified table structure.

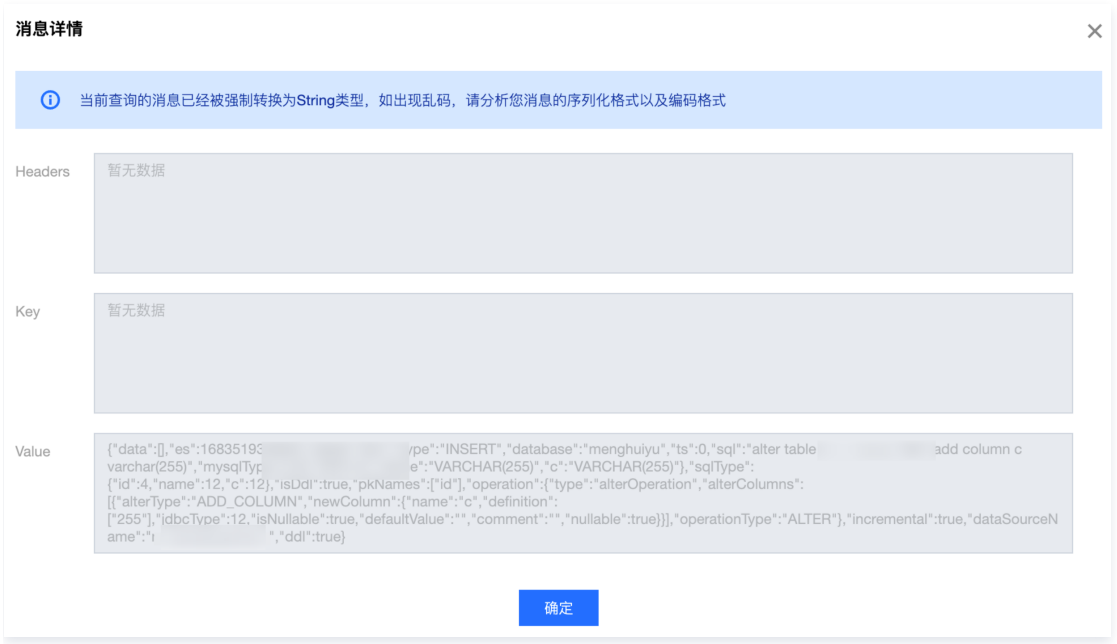
```
mysql> desc tb1;
+-------+--------------+------+-----+---------+-------+
| Field | Type         | Null | Key | Default | Extra |
+-------+--------------+------+-----+---------+-------+
| id    | int          | NO   | PRI | NULL    |       |
| name  | varchar(255) | YES  |     | NULL    |       |
| c     | varchar(255) | YES  |     | NULL    |       |
+-------+--------------+------+-----+---------+-------+
3 rows in set (0.00 sec)
```

5. View the results on Kafka.

```
{"data":[],"es":1683519346931,"table":"tb1","type":"INSERT","database":"databaseName","ts":0,"sql":"alter
table `databaseName`.`tb1` add column c varchar(255)","mysqlType":
{"id":"INT(-1)","name":"VARCHAR(255)","c":"VARCHAR(255)"},"sqlType":
{"id":4,"name":12,"c":12},"isDdl":true,"pkNames":["id"],"operation":
{"type":"alterOperation","alterColumns":[{"alterType":"ADD_COLUMN","newColumn":{"name":"c","definition":
["255"],"jdbcType":12,"isNullable":true,"defaultValue":"","comment":"","nullable":true}}],"operationType":
"ALTER"},"incremental":true,"dataSourceName":"databaseName","ddl":true}
```

Below is a sample:

©2013-2025 Tencent Cloud. All rights reserved.



©2013–2025 Tencent Cloud. All rights reserved.

FAQs
Common Issues with Real-Time Sync

Last updated: 2024-08-24 14:02:39

MySQL Class

1. MySQL serverid Conflict

Error message:

```
com.github.shyiko.mysql.binlog.network.ServerException: A slave with the same server_uuid/server_id as
this slave has connected to the master.
```

Solution: We have now optimized to randomly generate serverid. If in previous tasks the server-id was explicitly specified in MySQL advanced parameters, it is recommended to delete it, as multiple tasks may use the same data source with the same server-id, causing a conflict.

2. Binlog File Not Found Error:

Error message:

```
Caused by: org.apache.kafka.connect.errors.ConnectException: The connector is trying to read binlog
starting at GTIDs xxx and binlog file 'binlog.xxx', pos=xxx, skipping 4 events plus 1 rows, but this is no
longer available on the server. Reconfigure the connector to use a snapshot when needed.
```

Error Causes:
If the binlog file being read by the job has been purged from the MySQL server, an error will occur. This can happen due to various reasons, such as a short retention period of the binlog or the job processing speed not catching up with binlog generation speed, which exceeds the maximum retention period of MySQL binlog files, leading to binlog files being purged from the MySQL server and rendering the reading binlog position invalid.
Solution: If the job processing speed cannot catch up with the binlog generation speed, consider increasing the binlog retention period or optimizing the job to reduce backpressure and accelerate source consumption. If there are no abnormalities in the job status, the binlog might have been purged due to other operations on the database, hence making access impossible. It is necessary to check MySQL database-side information to determine why the binlog was purged.

3. MySQL Connection Reset

Error message:

```
EventDataDeserializationException: Failed to deserialize data of EventHeaderV4 .... Caused by:
java.net.SocketException: Connection reset.
```

Error Causes:

1. Network Issues.
2. The job experiences backpressure, preventing the source from reading data, making the binlog client idle. If the binlog connection remains idle after timeout, the MySQL server will disconnect the idle connection.

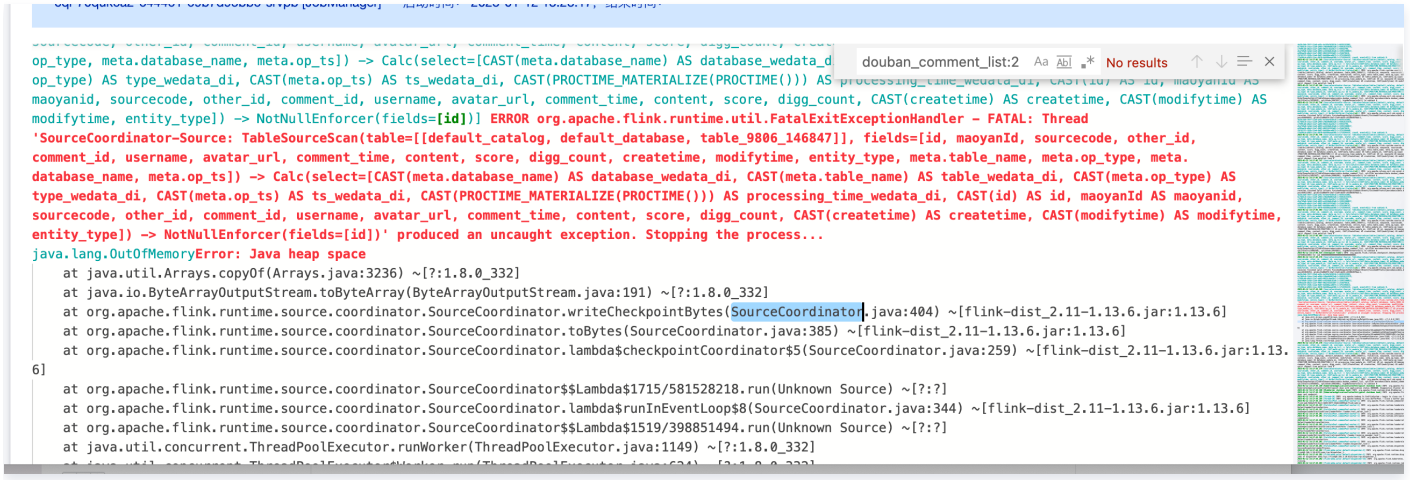
Solution:

1. If it's a network issue, you can increase MySQL network parameters set global slave_net_timeout = 120; (default 30s) set global thread_pool_idle_timeout = 120.
2. If backpressure is causing the issue, you can reduce it by adjusting the job, such as increasing parallelism, improving write speed, or enhancing task manager memory to reduce GC.

4. Mysql2dlc Task JobManager OOM

Error message:

©2013-2025 Tencent Cloud. All rights reserved.



Causes and Solutions:

1. The user's data volume is relatively large. You can increase the jobmanager CU count and use the MySQL advanced parameter scan.incremental.snapshot.chunk.size to increase the chunk size, default is 8096.
2. The user's data volume is not large, but there is a large difference between the maximum and minimum values of the primary key, resulting in many chunks being divided when using the equal chunk strategy. Modify the distribution factor so that user data follows an uneven data splitting logic, split-key.even-distribution.factor.upper-bound=5.0d. The default distribution factor has been modified to 10.0d.

5. The format of the user's binlog data is incorrect, causing a parsing exception in Debezium

Error message:

```
ERROR io.debezium.connector.mysql.MySqlStreamingChangeEventSource [] - Error during binlog processing.
Last offset stored = null, binlog reader near position = mysql-bin.000044/211839464.
2023-02-20 21:37:28.480 [blc-172.17.48.3:3306] ERROR io.debezium.pipeline.ErrorHandler [] - Producer
failure
io.debezium.DebeziumException: Error processing binlog event.
```

Solution:
After modifying binlog_row_image=full, you need to restart the database.

6. Is gh-ost supported?

Yes, it does not migrate the temporary table data generated by Online DDL changes. It only migrates the original DDL data executed using gh-ost in the source database. You can use the default or configure your own regular expressions for gh-ost shadow tables and unused tables.

DLC Type

1. Error when synchronizing incremental data (changelog data) to DLC

Error message:

©2013-2025 Tencent Cloud. All rights reserved.

```
2022-07-12 21:03:57.161 [IcebergFilesCommitter -> Sink: IcebergSink HYBRIS.xny_sl03_pro.subs_crm_clue (1/1)#34] WARN  org.apache.flink.runtime.taskmanager.Task [] -
IcebergFilesCommitter -> Sink: IcebergSink HYBRIS.xny_sl03_pro.subs_crm_clue (1/1)#34 (ced203d857ca39a5a1eb207be88110b5) switched from INITIALIZING to FAILED with failure cause:
java.lang.IllegalArgumentException: Cannot write delete files in a v1 table
    at org.apache.iceberg.ManifestFiles.writeDeleteManifest(ManifestFiles.java:154)
    at org.apache.iceberg.SnapshotProducer.newDeleteManifestWriter(SnapshotProducer.java:374)
    at org.apache.iceberg.MergingSnapshotProducer.lambda$newDeleteFilesAsManifests$10(MergingSnapshotProducer.java:681)
    at java.util.HashMap.forEach(HashMap.java:1289)
    at org.apache.iceberg.MergingSnapshotProducer.newDeleteFilesAsManifests(MergingSnapshotProducer.java:678)
    at org.apache.iceberg.MergingSnapshotProducer.prepareDeleteManifests(MergingSnapshotProducer.java:664)
    at org.apache.iceberg.MergingSnapshotProducer.apply(MergingSnapshotProducer.java:533)
    at org.apache.iceberg.SnapshotProducer.apply(SnapshotProducer.java:164)
    at org.apache.iceberg.SnapshotProducer.lambda$commit$2(SnapshotProducer.java:283)
    at org.apache.iceberg.util.Tasks$Builder.runTaskWithRetry(Tasks.java:404)
    at org.apache.iceberg.util.Tasks$Builder.runSingleThreaded(Tasks.java:214)
    at org.apache.iceberg.util.Tasks$Builder.run(Tasks.java:198)
    at org.apache.iceberg.util.Tasks$Builder.run(Tasks.java:190)
    at org.apache.iceberg.SnapshotProducer.commit(SnapshotProducer.java:282)
    at org.apache.iceberg.flink.sink.IcebergFilesCommitter.commitOperation(IcebergFilesCommitter.java:312)
    at org.apache.iceberg.flink.sink.IcebergFilesCommitter.commitDeltaTxn(IcebergFilesCommitter.java:299)
    at org.apache.iceberg.flink.sink.IcebergFilesCommitter.commitUpToCheckpoint(IcebergFilesCommitter.java:218)
    at org.apache.iceberg.flink.sink.IcebergFilesCommitter.initializeState(IcebergFilesCommitter.java:153)
    at org.apache.flink.streaming.api.operators.StreamOperatorStateHandler.initializeOperatorState(StreamOperatorStateHandler.java:118)
    at org.apache.flink.streaming.api.operators.AbstractStreamOperator.initializeState(AbstractStreamOperator.java:290)
    at org.apache.flink.streaming.runtime.tasks.OperatorChain.initializeStateAndOpenOperators(OperatorChain.java:441)
    at org.apache.flink.streaming.runtime.tasks.StreamTask.restoreGates(StreamTask.java:585)
    at org.apache.flink.streaming.runtime.tasks.StreamTaskActionExecutor$1.call(StreamTaskActionExecutor.java:55)
```

Cause:
The DLC table is a V1 table, which does not support changelog data.
Solution:
1. Modify the DLC table to V2 and enable upsert support.
2. To enable upsert support: ALTER TABLE tblname SET TBLPROPERTIES ('write.upsert.enabled'='true').
3. To modify the table to V2: ALTER TABLE tblname SET TBLPROPERTIES ('format-version'='2').

Check if the table properties are set successfully: SHOW TBLPROPERTIES tblname.

2. Cannot write incompatible dataset to table with schema

Error Details:

```
Caused by: java.lang.IllegalArgumentException: Cannot write incompatible dataset to table with schema:
* mobile should be required, but is optional
    at org.apache.iceberg.types.TypeUtil.checkSchemaCompatibility(TypeUtil.java:364)
    at org.apache.iceberg.types.TypeUtil.validateWriteSchema(TypeUtil.java:323)
```

Error Causes:
When creating a dlc table, the user set a NOT NULL constraint on the field.
Resolution:
Do not set a NOT NULL constraint when creating the table.

3. Synchronize mysql2dlc, an Array Out of Bounds error occurred

Problem Details:

```
java.lang.ArrayIndexOutOfBoundsException: 1
    at
org.apache.flink.table.data.binary.BinarySegmentUtils.getLongMultiSegments(BinarySegmentUtils.java:736) ~
[flink-table-blink_2.11-1.13.6.jar:1.13.6]
    at org.apache.flink.table.data.binary.BinarySegmentUtils.getLong(BinarySegmentUtils.java:726) ~[flink-
table-blink_2.11-1.13.6.jar:1.13.6]
    at
org.apache.flink.table.data.binary.BinarySegmentUtils.readTimestampData(BinarySegmentUtils.java:1022) ~
[flink-table-blink_2.11-1.13.6.jar:1.13.6]
    at org.apache.flink.table.data.binary.BinaryRowData.getTimestamp(BinaryRowData.java:356) ~[flink-
table-blink_2.11-1.13.6.jar:1.13.6]
    at org.apache.flink.table.data.RowData.lambda$createFieldGetter$39385f9c$1(RowData.java:260) ~[flink-
table-blink_2.11-1.13.6.jar:1.13.6]
```

Error Causes:
The issue caused by using the time field as the primary key.

©2013-2025 Tencent Cloud. All rights reserved.

Solution:

1. Do not use the time field as the primary key.
2. The user still needs the time field for uniqueness. It is recommended that the user add a redundant string field in dlc and map the upstream time field to the redundant field using a function.

4. DLC Task reported that it is not an Iceberg Table

Error message:

```
[flink-dist_2.11-1.13.6.jar:1.13.6]
    ... 10 more
Caused by: org.apache.iceberg.exceptions.NoSuchIcebergTableException: Not an iceberg table: HYBRIS.origindb.
movieunifiedorder01_unified_order_sharding__maoyan_order_v2 (type=null)
    at org.apache.iceberg.exceptions.NoSuchIcebergTableException.check(NoSuchIcebergTableException.java:33) ~
[pool-1-thread-1-1678868648050-prod-sort-connector-iceberg-dlc-2.6.29-1-1.4.0.jar:1.4.0]
```

Error Causes:

1. The user can run statements in dlc to check the type of the table.
2. Statement: desc formatted table name.
3. The table type can be viewed.

Resolution:

Choose the correct engine to create an Iceberg type table.

5. The order of Flink SQL fields is inconsistent with the DLC target table fields, causing an error

Solution:

Do not modify the field order of the task table.

Doris Type

1. How to select and optimize the Doris specifications?

Please refer to Doris Resource Specifications and Optimization Suggestions.

2. Too many import tasks, new import task submission error "current running txns on db xxx is xx, larger than limit xx"?

Adjust the FE parameter: max_running_txn_num_per_db, default is 100, can be appropriately increased, recommended to control within 500.

3. Error err=[E-235] occurs due to too fast import frequency?

- Parameter optimization suggestion: You can temporarily resolve it by appropriately increasing the max_tablet_version_num parameter. The default is 200, recommended to control within 2000.
- Business optimization suggestion: Reduce the import frequency to fundamentally solve this problem.

4. Import file is too large and restricted by parameters. Error "The size of this batch exceeds the max size"?

Adjust the BE parameter: streaming_load_max_mb, recommended to exceed the size of the file to be imported.

5. Data import error: "[-238]"?

- Reason: The -238 error usually occurs when the volume of data imported in the same batch is too large, resulting in too many Segment files in a single tablet.
- Parameter optimization suggestion: You can appropriately increase the BE parameter max_segment_num_per_rowset. The default value is 200, can be increased in multiples (such as 400, 800), recommended to control within 2000
- Business optimization suggestion: It is recommended to reduce the data volume of a single batch import.

6. Import failure, error: "too many filtered rows xxx, "ErrorURL":" or Insert has filtered data in strict mode, tracking url=xxxx."?

©2013-2025 Tencent Cloud. All rights reserved.

Reason: The schema, partition, etc. of the table does not match the imported data. You can execute doris commands in TCHouse-P Studio or the client to view the specific reason: show load warnings on `<tracking url>`, `<tracking url>` is the error URL returned in the error message.

Offline Synchronization Frequently Asked Questions

©2013-2025 Tencent Cloud. All rights reserved.

Last updated：2024-08-24 14:03:04

1. Confirm that connecting CCN products across regions requires public network bandwidth

DataInLong resource groups need to be in the same VPC network as the customer's database for internal network access. If the integration resource group and the database are in different regional VPCs, the network needs to be connected through CCN CCN. CCN CCN is set to allow 10Kb of traffic by default for convenience in testing. This is only suitable for small data synchronization and network testing. For performance testing and production tasks, please purchase bandwidth to avoid data synchronization being limited by network bandwidth.



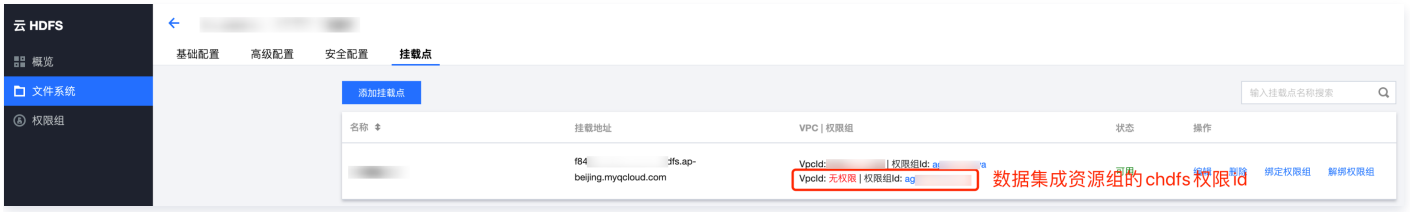
2. Hive On CHDFS Table Write Error: Permission denied: No access rules matched

Error message:

```
  at com.tencent.cloud.wedata.runner.common.api.AbstractTaskRunner.executeWithCommonLog(AbstractTaskRunner.java:508)
  at com.tencent.cloud.wedata.runner.common.api.AbstractTaskRunner.startWork(AbstractTaskRunner.java:546)
  at com.tencent.teg.scheduler.runner.runners.WeDataDiRunner.main(WeDataDiRunner.java:75)
[2022-12-29 01:00:52]-[ERROR] 获取FileSystem时发生网络IO异常,请检查您的网络是否正常!HDFS地址: [message:defaultFS =ofs://f4m[redacted]w.chdfs.ap-beijing.myqcloud.com]
[2022-12-29 01:00:52]-[ERROR] Exception when job run: com.tencent.wedata.common.exception.wedataException: Code:[HiveWriter-06], Description:[与HDFS建立连接时出现IO异常.]. - IOException: P
  at com.tencent.wedata.common.exception.wedataException.aswedataException(wedataException.java:26)
  at com.tencent.wedata.plugin.writer.hivewriter.HdfsHelper.getFileSystem(HdfsHelper.java:368)
  at com.tencent.wedata.plugin.writer.hivewriter.HiveWriter$Job.init(HiveWriter.java:194)
  at com.tencent.wedata.core.job.JobContainer.initJobWriter(JobContainer.java:717)
  at com.tencent.wedata.core.job.JobContainer.init(JobContainer.java:307)
  at com.tencent.wedata.core.job.JobContainer.start(JobContainer.java:114)
  at com.tencent.wedata.core.Engine.start(Engine.java:232)
  at com.tencent.wedata.core.Engine.entry(Engine.java:122)
  at com.tencent.teg.scheduler.runner.runners.wedata.wedataHelper.startwedata(wedataHelper.java:126)
  at com.tencent.teg.scheduler.runner.runners.WeDataDiRunner.execute(WeDataDiRunner.java:161)
  at com.tencent.cloud.wedata.runner.common.api.AbstractTaskRunner.executeWithCommonLog(AbstractTaskRunner.java:508)
  at com.tencent.cloud.wedata.runner.common.api.AbstractTaskRunner.startWork(AbstractTaskRunner.java:546)
  at com.tencent.teg.scheduler.runner.runners.WeDataDiRunner.main(WeDataDiRunner.java:75)

[2022-12-29 01:00:52]-[INFO] Total 0 records, 0 bytes | Speed 0B/s, 0 records/s | Error 0 records, 0 bytes |  All Task WaitWriterTime 0.000s |  All Task WaitReaderTime 0.000s | Percenta
[2022-12-29 01:00:52]-[INFO]

该任务最可能的错误原因是:
wedataException: Code:[HiveWriter-06], Description:[与HDFS建立连接时出现IO异常.]. - IOException: Permission denied: No access rules matched
[2022-12-29 01:00:52]-[ERROR] Task execution failed
[2022-12-29 01:00:52]-[INFO] commitTask success.
[2022-12-29 01:00:52]-[INFO] taskId: <2[redacted]37>
[2022-12-29 01:00:52]-[INFO] taskState: <FAILED>
```

Error Causes:

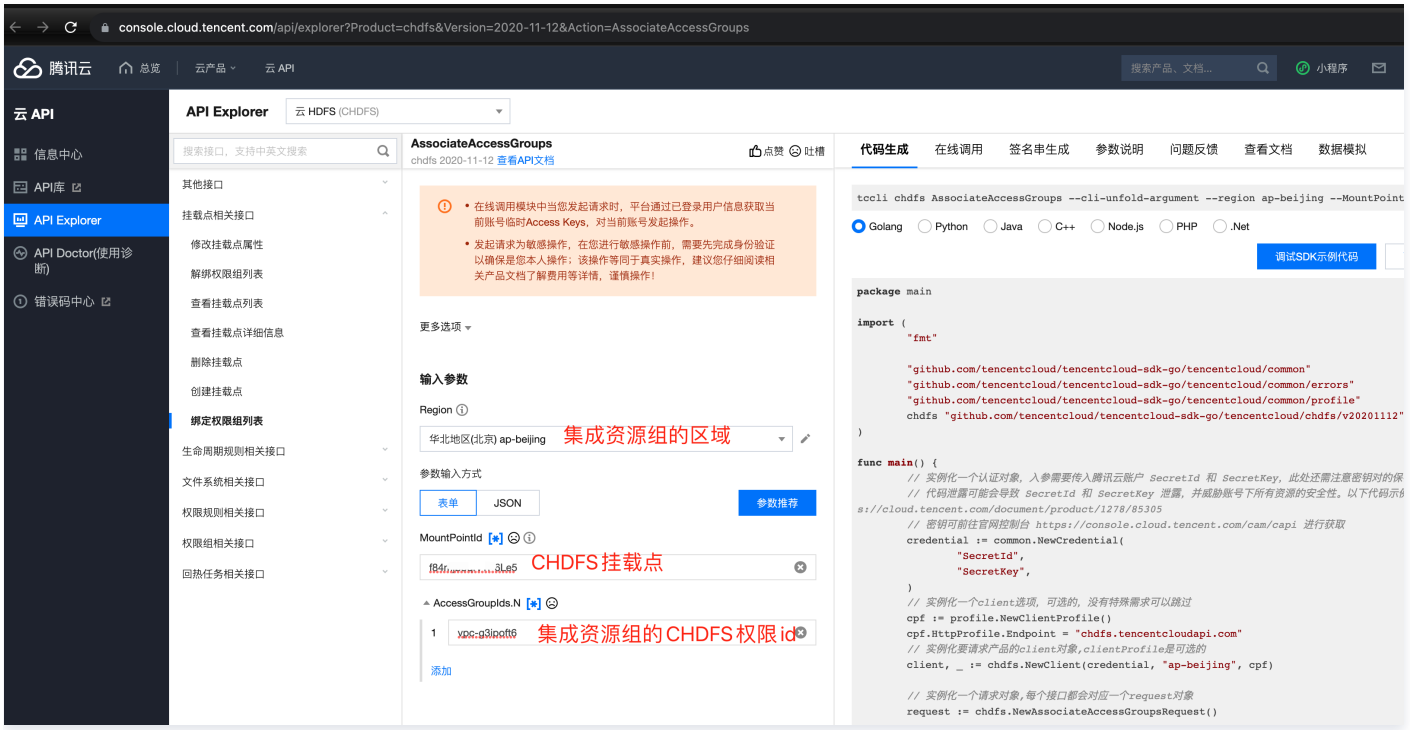
DataInLong resource group is a fully managed resource group. The outbound network segment of the resource group is not within the customer's VPC internal network segment. You need to open the security group for the DataInLong resource group.
Please authorize the CHDFS permission ID for the integration resource group at the CHDFS mount point.



Solution:

©2013-2025 Tencent Cloud. All rights reserved.

Open the page: API 3.0 Explorer and fill in the parameters, where **MountPointId** is the customer's CHDFS mount point and **AccessGroupIds.N** are the DataInLong resource group's CHDFS permission IDs (Note that different regions have different permission IDs).



DataInLong resource group CHDFS permission ID and region mapping table:

Region	DataInLong resource group CHDFS permission ID
Beijing	ag-wgbku4no
Guangzhou	ag-x1bhppnr
Shanghai	ag-xnjfr9d3
Singapore	ag-phxtv0ah
USA Silicon Valley	ag-tgwl8bca
Virginia	ag-esxpwxjn

©2013-2025 Tencent Cloud. All rights reserved.

Others

Last updated: 2024-08-24 14:03:25

1. What data sources does DataInLong support?

DataInLong supports various cloud and self-built data sources, such as:

- Relational databases: MySQL, PostgreSQL, Oracle, SQL Server, IBM DB2, DM, SAP HANA, etc.
- Big Data Storage: Hive, HDFS, HBase, Kudu, Clickhouse, DLC, Impala, Gbase, Tbase, etc.
- Semi-structured: FTP, SFTP, COS.
- NoSQL:Redis,Elasticsearch.
- Message Queue: Kafka, etc.

2. What synchronization scenarios does DataInLong support?

DataInLong primarily supports offline and real-time data synchronization scenarios. Offline scenarios support periodic data ingestion; real-time scenarios monitor data changes based on database logs or message queues and pull the latest data in real time.

3. What data synchronization directions are supported?

DataInLong uses a star schema data model. For supported data source types, sources and destinations can be flexibly paired.

4. What data object synchronization methods are supported?

DataInLong supports single database single table, sharded database sharded table, and batch multi-database multi-table synchronization methods.

©2013-2025 Tencent Cloud. All rights reserved.

Data Development
Data Development Overview

Last updated: 2024-11-22 16:51:24

WeData Data Development provides data engineers and related personnel with an efficient, collaborative, and intelligent data development platform. The WeData Data Development module supports various engines (EMR, Data Lake Compute (DLC), Tencent Cloud Data Warehouse TCHouse-P, etc.), offering capabilities such as script development, visual development, task orchestration, task release, and task operation and maintenance, helping organizations and enterprises efficiently build their data warehouses.

Use Limits

1. The type of compute tasks available in WeData Data Development depends on the compute engine bound to the project. Binding to an EMR engine requires starting Hive and Spark services, allowing the use of HiveSQL, SparkSQL, and Spark task types for development. Other tasks in WeData based on the EMR engine, such as Trino, also require the corresponding EMR services to be enabled. Binding to a TCHouse-P engine enables TCHouse-P tasks; binding to a DLC engine allows the use of DLC SQL and DLC Spark task types. Other types of relational databases, if needed for data development in WeData, require creating a data source within WeData to use JDBC tasks for creating SQL scripts. WeData Data Development is responsible for the execution of these scripts.
2. Each compute engine has its own data access control system. WeData does not perform specific data access control but instead interfaces with each engine's permissions through account interoperability. For instance, permissions for EMR data in WeData are managed via account mapping.

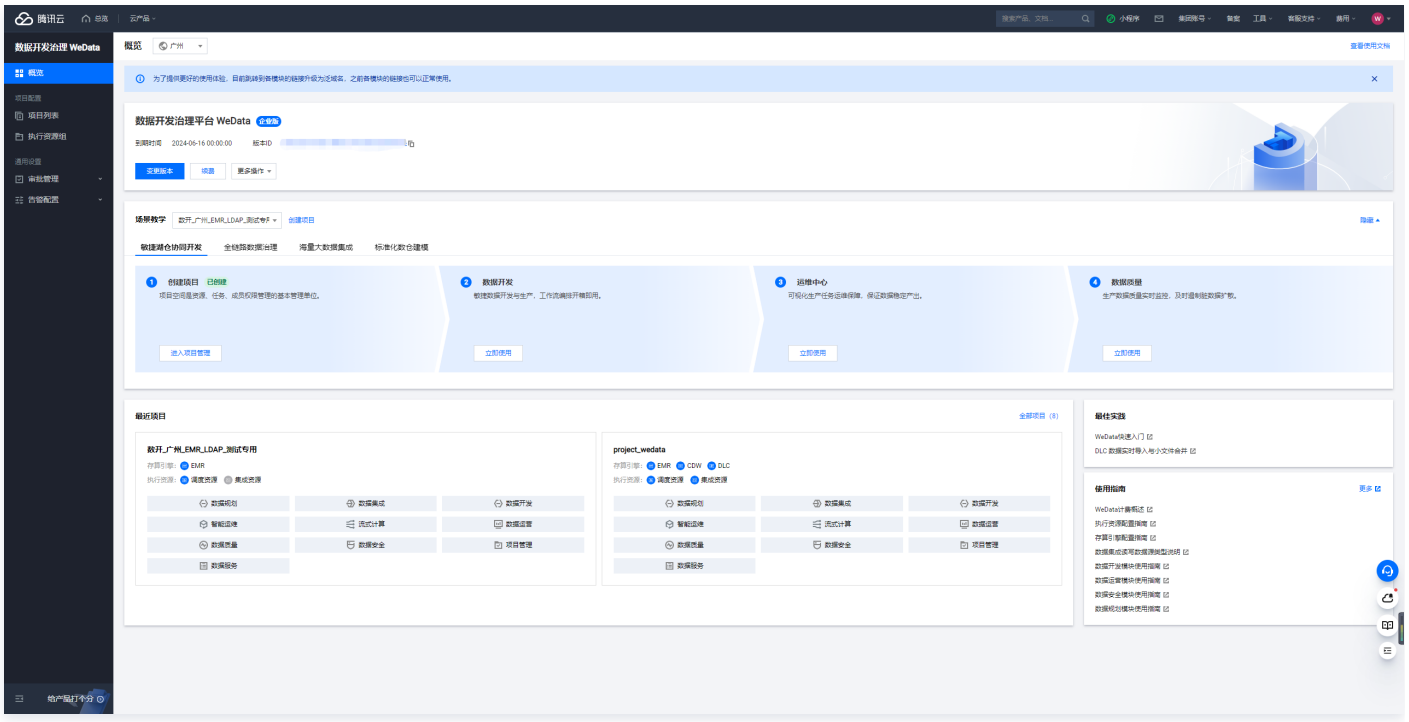
©2013-2025 Tencent Cloud. All rights reserved.

Data Development feature guidance
Feature Index
Overall Data Development Feature Index

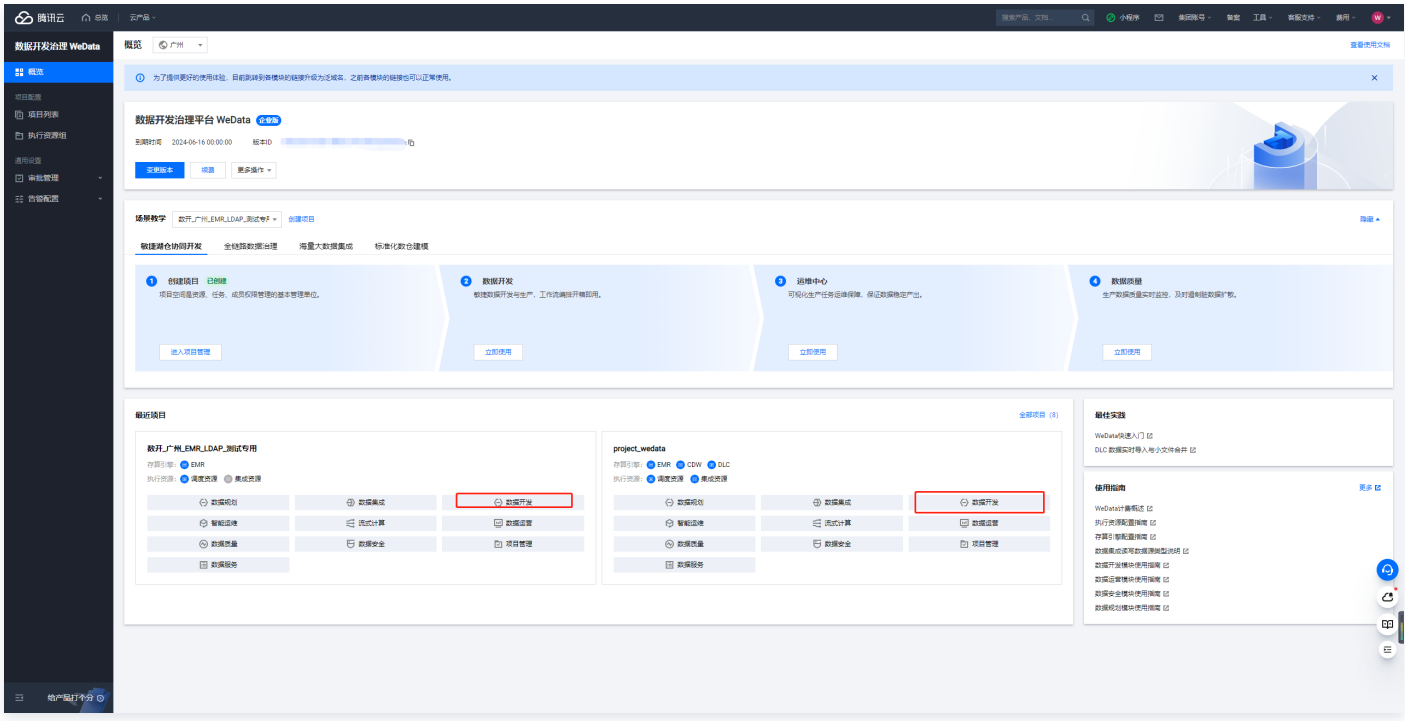
Last updated: 2024-11-22 16:51:51

This document introduces the WeData Data Development interface features to help users understand the overall layout and features of the Data Development module.

1. Log in to WeData Console.

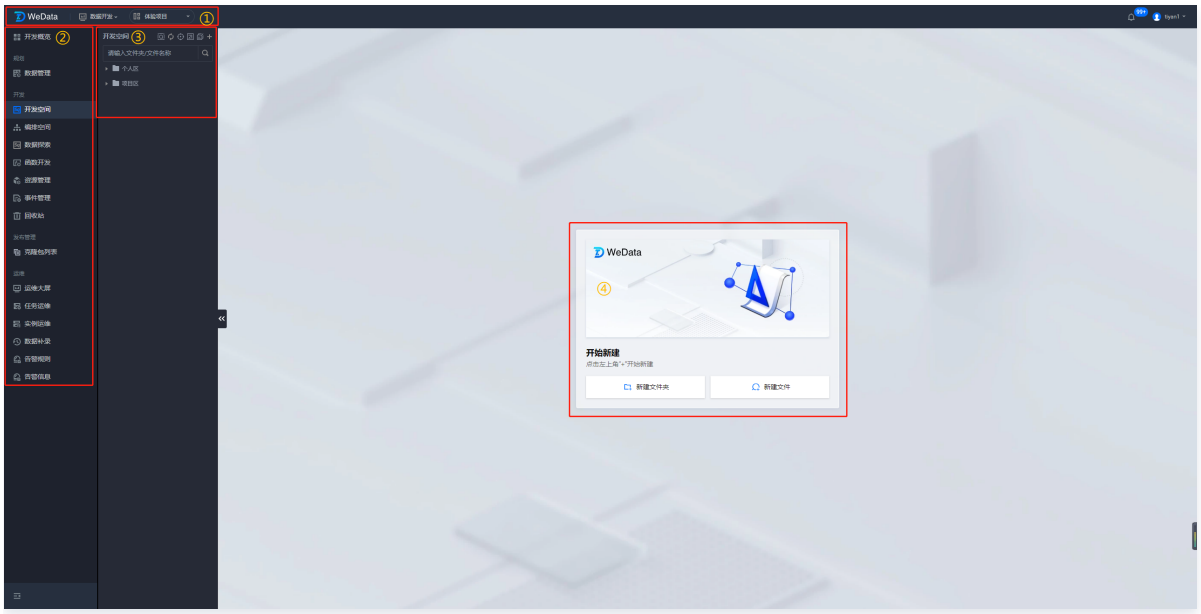


2. Click **Data Development** in the project list to enter the data development module interface.



©2013-2025 Tencent Cloud. All rights reserved.

Overall Feature Introduction



Feature Overview

Region	Description
1	• Return to WeData Console Click the WeData icon to return to the main interface of the WeData Console. • Switch WeData Feature Modules Click the 数据开发 icon to select other feature modules within the WeData products:  • Switch Projects Click the WeData产品演示项目for产品 icon to select other project spaces in the current region.
2	In this area, click the icon to display the names of the corresponding feature buttons (expanded by default). • Development Overview: Provides a full-chain data development process from database and table creation, data development, task submission, to task operations and maintenance. • Development

©2013–2025 Tencent Cloud. All rights reserved.

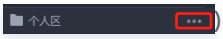
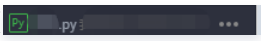
- Data Management: Database table management supports the creation and DDL conversion of database tables. Currently, the supported types on the public cloud include Hive and DLC.
- Development Space: Development Space is WeData's script development and debugging module, providing users with script management, debugging, and SQL script debugging result viewing features. It also offers quick query capabilities for SQL-type scripts.
- Orchestration Space: Orchestration Space is the task configuration and management module, providing users with visual workflow configuration, task management, task debugging, task scheduling configuration, task submission, and publishing features.
- **Data Exploration**: Data Exploration currently offers quick query and analysis capabilities for DLC data tables.
- Function Development: Supports the creation and management of user-defined functions for Hive SQL, Spark SQL, and DLC. Once created and submitted, these functions can be referenced in development scripts and workflow computation tasks through the function library.
- Resource Management: Resource Management provides users with the ability to manage additional resource files used in the data development process, including resource files needed for function creation and Jar files used by Spark tasks.
- Event Management: Event Management provides management features for scheduling dependency events, allowing tasks in WeData to be triggered and executed by events.
- Recycle Bin: The Recycle Bin provides management capabilities for deleted scripts, tasks, and files from Development Space, Orchestration Space, and Resource Management.

- Release management
  - Clone Package List: The Clone Package List is used to manage clone tasks from the current project to other target projects, providing features for creating, editing, and viewing clone package details.
- OPS
  - Operation and Maintenance Dashboard: The Operation and Maintenance Dashboard displays key operation and maintenance metrics in report form, including task and instance running status statistics, task status distribution, task completion trends, and scheduling resource changes.
  - Task Operation and Maintenance: Task Operation and Maintenance supports users in managing and maintaining tasks from both the workflow and task perspectives, including stopping, pausing, rerunning tasks, and viewing instance and historical records.
  - Instance Operation and Maintenance: Tasks generate corresponding instances based on periodic scheduling, and Instance Operation and Maintenance provides features like instance log viewing and rerunning.
  - Data Supplement: Data Supplement provides a unified interface for managing data supplement plans, allowing users to view detailed statuses of created data supplement plans and perform batch operations on instances. Data Supplement does not offer a direct feature for creating supplement plans, which must be initiated from the task list.
  - Alert Rules: The Alert Rules page provides alert task configuration features, pushing alert information to specified targets after setting alert conditions through alert notifications.
  - Alert Information: The Alert Information page provides a detailed list of all configured alert tasks, with filtering options by time, task name, and recipient.

3

- Global Search (): Supports global code search in Development Space and Orchestration Space.
- Refresh (): Refresh the directory tree of the current project's development space/orchestration space.
- Localization (): One-click localization of the currently open script in the left tree.
- Collapse (): One-click collapse of all expanded directories.
- Batch operations ():
  - Orchestration Space: Supports batch importing/exporting of Data Workflows and configuration information of computing tasks. Supports batch operations on computing tasks, including deletion, modifying execution resource groups, changing responsible persons, and altering task parameters. Allows hiding or showing categorized folders of computing tasks.
  - Development Space: Supports batch importing of development scripts via Excel. Supports batch deletion of script files.
- New (): Allows creating folders and files, and creating different types of scripts and tasks in Development Space and Orchestration Space.

©2013-2025 Tencent Cloud. All rights reserved.

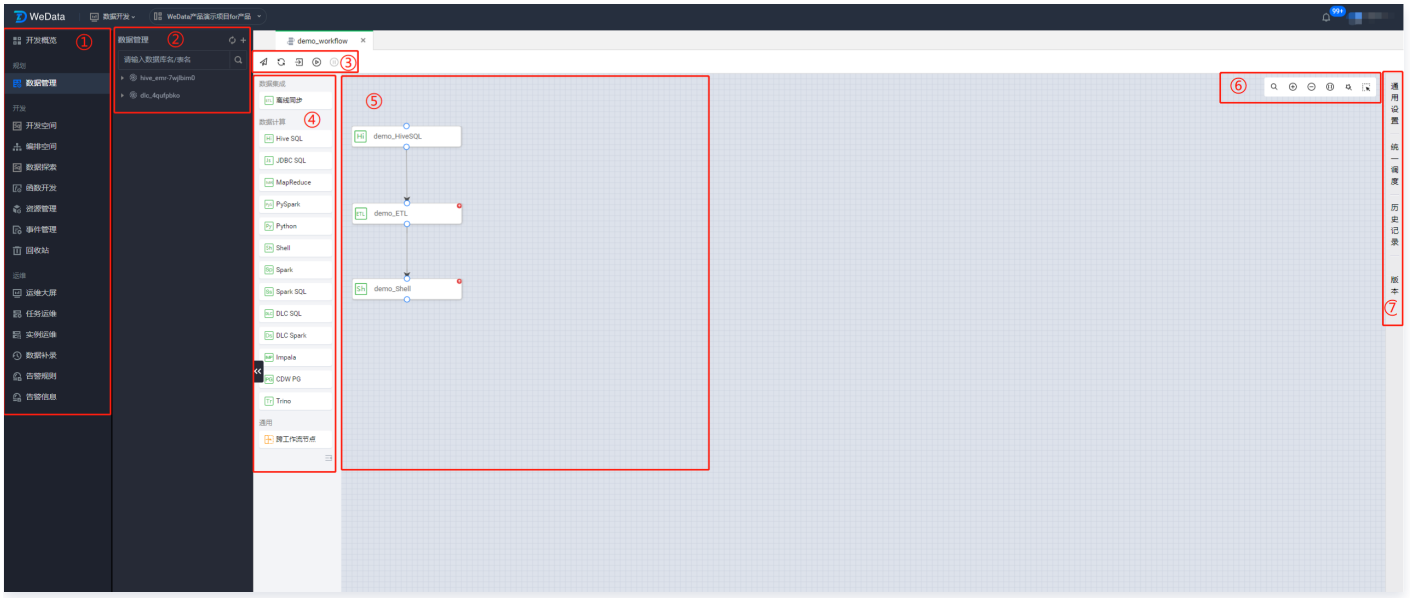
	• Directory Manipulation (): Enables creating new folders and files. Supports batch exporting/importing script files, all organized as ZIP files per directory structure. • File Operations (): Enables renaming, moving, downloading, and deleting development files, and viewing citation situations of development files in Orchestration Space.
4	New Folder/File: When entering data development for the first time, if no files are created yet, users can create their first folder or file here.

©2013–2025 Tencent Cloud. All rights reserved.

Data Development Workflow Feature Index

Last updated: 2024-11-22 16:52:12

The WeData data workflow provides a visual interface for creating and organizing workflows. A workflow is a logical collection of multiple tasks. It is recommended to organize a single complete data processing scenario in a workflow to avoid having too many tasks or tasks without logical connections, thereby reducing the maintenance costs of the workflow. The data workflow provides various types of task management and orchestration capabilities, such as Hive SQL, Spark SQL, Shell, etc.



Feature Overview

Region	Description
1	For details of all features included in the Data Development module, please refer to the overall Data Development feature index.
2	Orchestration Space Directory Tree Area: In this area, you can view created folders and workflow tasks. Double-click to enter the edit interface. You can also perform operations such as global search, refresh, localization, collapse, and new creation.
3	• Submit ( ): After editing the workflow, click the icon in the toolbar to submit the nodes in the workflow to the scheduling system. • Refresh ( ): Refresh all configuration information of the current workflow to ensure the parameters are up-to-date. • Go to Operation and Maintenance ( ): Click this icon to go to the Operation and Maintenance - Workflow List page. • Workflow Testing ( ): Click this icon to test the current workflow. • Stop Workflow Testing ( ): Click this icon to stop testing the current workflow.
4	Task Node List: This area displays all the node information that Data Development supports configuration for.
5	Canvas: Double-click the workflow name in the directory tree to edit the dependency running relationships between tasks in the expanded workflow canvas.

©2013-2025 Tencent Cloud. All rights reserved.

6	This area is for the format configuration items of the canvas. You can configure the current workflow canvas accordingly.
7	This area allows you to configure the scheduling of the current workflow tasks to ensure the workflow runs properly. • General Settings: General configurations take effect for the current workflow orchestration and the task nodes within the workflow. Specific configurations can also be set in the node attributes (the scheduling is ultimately based on the node attribute settings; information that has been individually set for a node will not change with modifications to the workflow configuration). • Unified Scheduling: Set unified scheduling configurations for all task nodes under the workflow. Supports conventional and crontab methods. • History: Supports viewing the historical change records of the workflow, including the operator, change time, and operation content. • Version: Displays the historical submission records of computational tasks. You can view the Node Historical Versions, Submitted by, Submission Time, Change Type, Status, Remarks, and other information in the version panel. You can view the information of a single version and support the comparison between two selected versions.

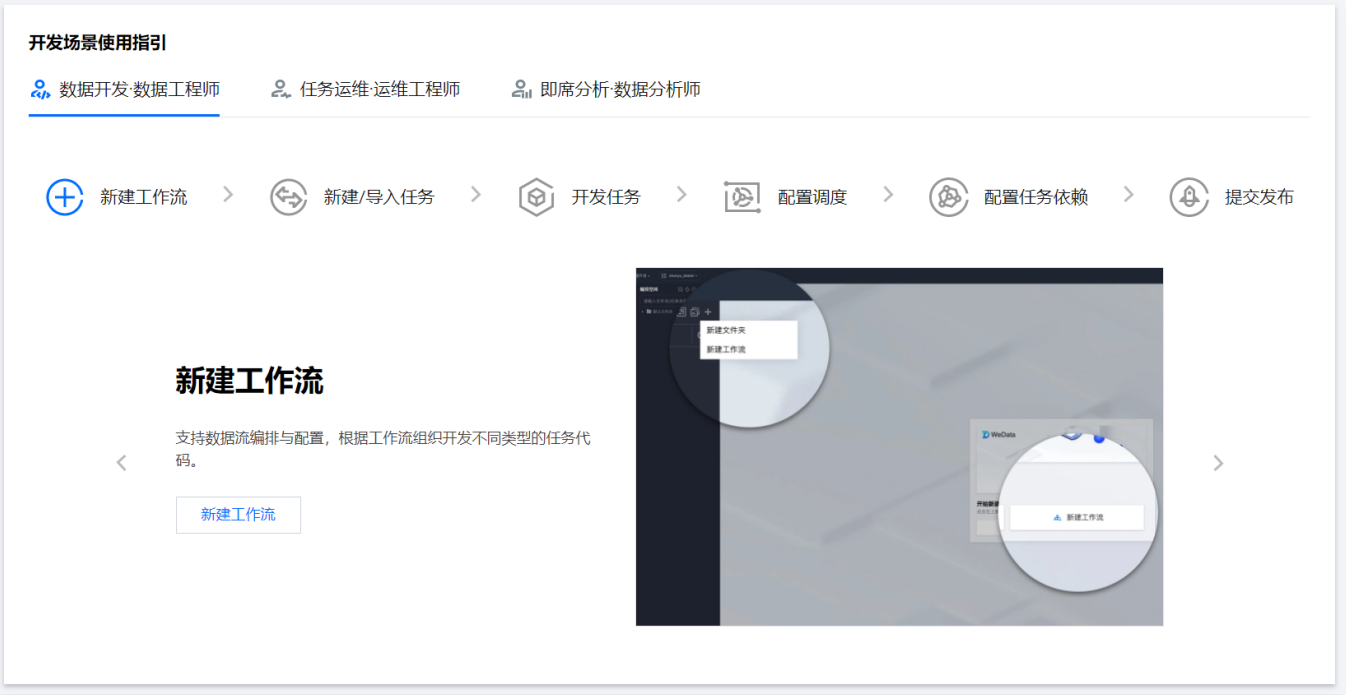
©2013–2025 Tencent Cloud. All rights reserved.

Quick Start on Data Development

Last updated：2024-11-22 16:55:36

Overview

This article uses EMR as the computing and storage engine, sets up an orchestration workflow on WeData, and utilizes task nodes such as Hive SQL and ETL to introduce how data engineers can use task development, workflow design, and scheduling policies to create data processing workflows, helping data engineers understand the basic usage logic of the data development module quickly.



- Step 1: Create a workflow

  Create a new workflow to organize data tasks and develop different types of code tasks, orchestrating the data workflow. You can set the name, description, responsible person, execution engine, parameters, and other attributes of the workflow as needed.
- Step 2: Create a task

  In the newly created workflow, you can create new data tasks and configure parameters as needed. Various task types can be selected to complete the data workflow construction.
- Step 3: Workflow Orchestration

  The data development process involves orchestrating computing tasks, sequencing different types of tasks using various task nodes to form a data workflow.
- Step 4: Configure task nodes

  In the constructed data task nodes, write the corresponding code, such as Python, Shell, Hive SQL, etc. You can use visual tools or manually write code for development.
- Step 5: Configure scheduling

  Configure scheduling execution policies for development tasks to achieve automatic execution of data tasks. You can set the scheduling time, frequency, parameter passing, execution method, event scheduling, etc., to execute data tasks as needed.
- Step 6: Configure task dependencies

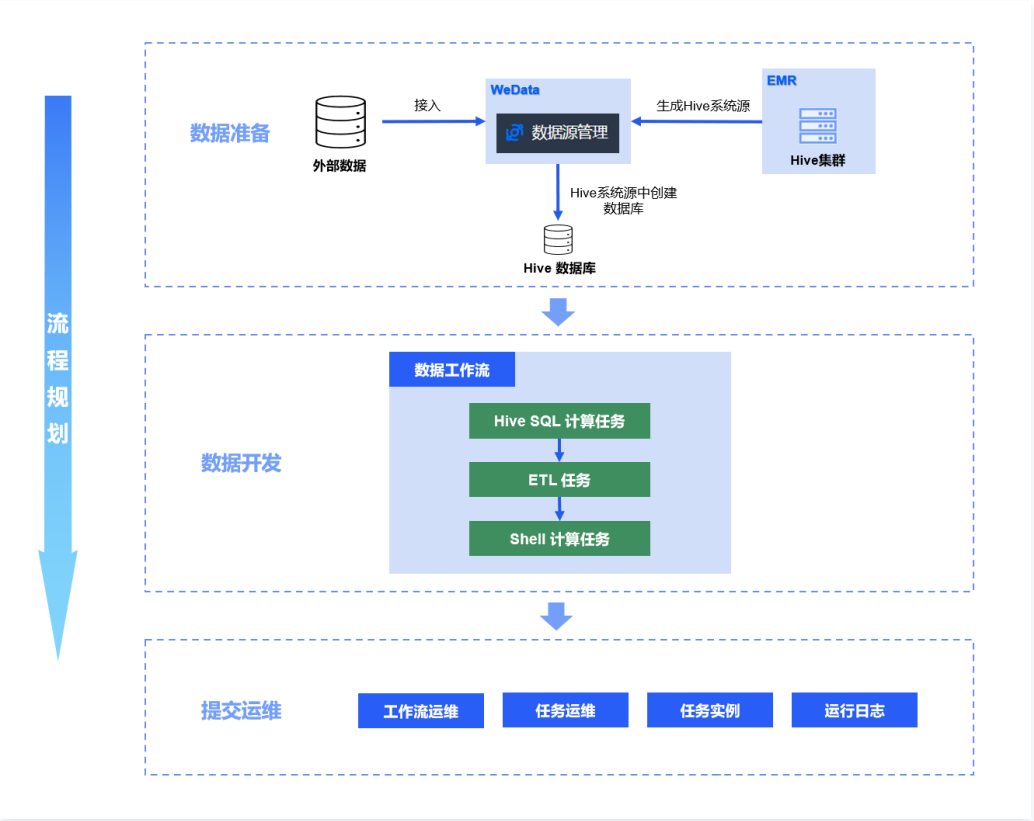
  Downstream data tasks in the data workflow need to rely on the results of upstream tasks for computation. Therefore, task dependencies need to be set. You can define the upstream and downstream dependencies between tasks, including self-dependency within the data workflow and self-dependency of data tasks.

The following demonstrates the overall process of data engineers in WeData data development according to the above steps.

Process Planning

©2013-2025 Tencent Cloud. All rights reserved.

Usage planning for quick start of the data development process is divided into three parts: **Data Preparations, Data Development, and Submit Operation and Maintenance.**



- Data Preparations: Prepare the data in a MySQL database and connect it to WeData; use the data management feature in data development to prepare the database under the Hive system data source connected via EMR. Example data and database configuration steps are provided in subsequent steps.
- Data Development: Create a data workflow. In the data workflow, use Hive SQL computation tasks to create new data tables under the Hive system source database. Then, use ETL computation tasks to synchronize the prepared MySQL data to the Hive data tables, and finally, use Shell computation tasks to record the completion time of the process.
- Submit Operation and Maintenance: After configuring and submitting the workflow, it enters the operation and maintenance state. The Hive SQL, ETL, and Shell tasks within will periodically run according to the scheduling policy configured in the data development process, generating instances. View the results of the data development process through SQL queries and log monitoring.

Data Preparations

Before officially proceeding with the data development process, prepare a sample data set first.

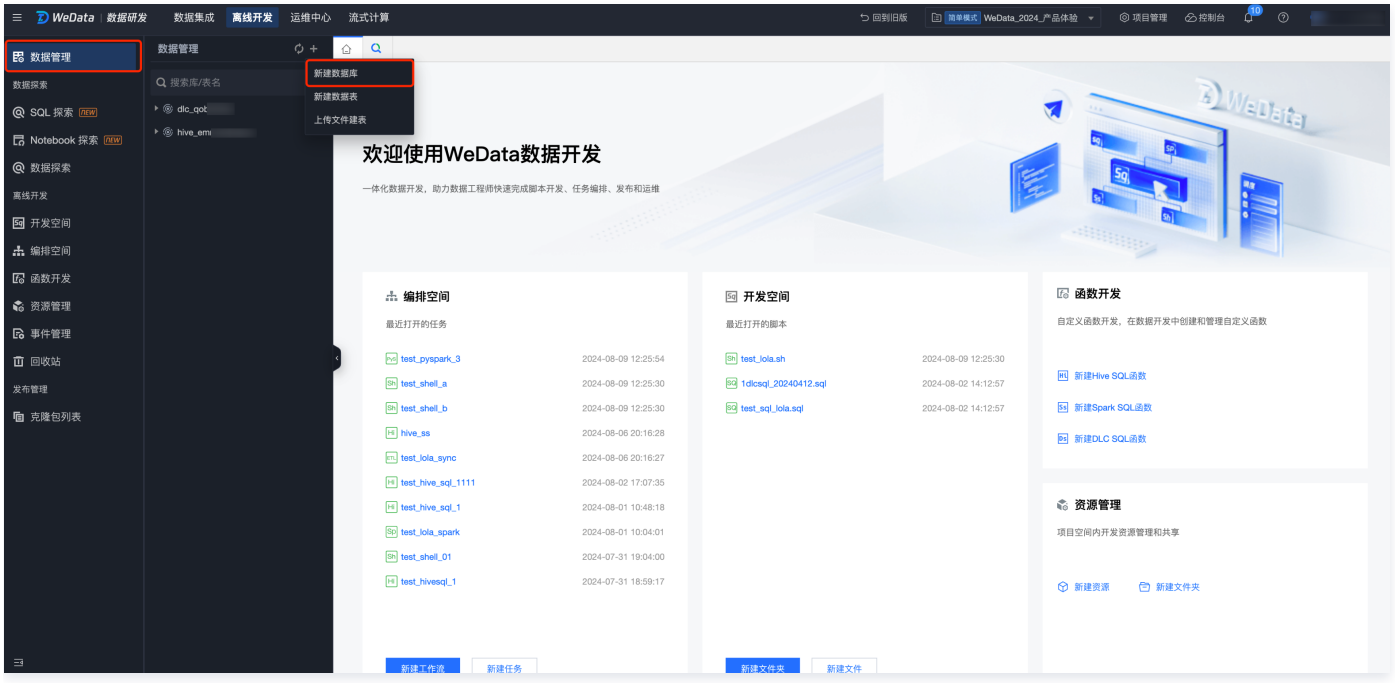
1. In Data Source Management, connect a MySQL type data source named demo_mysql, and connect a database named demo_mysql to it. Prepare a data table named mysql_etl_read and insert data into it. The related table creation and data insertion statements are as follows:

```
Create a data table named mysql_etl_read in the MySQL data source demo_mysql database
CREATE TABLE IF NOT EXISTS demo_mysql.mysql_etl_read (
    id VARCHAR(20),          #Serial Number
    order_id VARCHAR(20),    #Order Number
    user_id VARCHAR(20),     #User Number
    gender VARCHAR(20),      #Gender
    address VARCHAR(50),     #Address
    vip VARCHAR(50),         #User Level Identifier
    telephone VARCHAR(50),   #Telephone
    id_card VARCHAR(50)      #User ID Card
);
```

©2013–2025 Tencent Cloud. All rights reserved.

```
#Insert 10 records into the created mysql_etl_read data table
INSERT INTO demo_mysql.mysql_etl_read (id,order_id,user_id,gender,address,vip,telephone,id_card)
VALUES ('1','DD20210101','U0001','Male',' Shanghai Jing'an District','Regular
User','13800000001','310101200001011111');
INSERT INTO demo_mysql.mysql_etl_read (id,order_id,user_id,gender,address,vip,telephone,id_card)
VALUES ('2','DD20210102','U0002','Male',' Beijing Haidian District','Regular
User','13900000002','110101200001021122');
INSERT INTO demo_mysql.mysql_etl_read (id,order_id,user_id,gender,address,vip,telephone,id_card)
VALUES ('3','DD20210103','U0003','Female','Guangdong Province Shenzhen Nanshan District','Regular
User','13700000003','440301200001031133');
INSERT INTO demo_mysql.mysql_etl_read (id,order_id,user_id,gender,address,vip,telephone,id_card)
VALUES ('4','DD20210104','U0004','Female','Guangdong Province Guangzhou Yuexiu District','Gold
VIP','15800000004','440101200001041144');
INSERT INTO demo_mysql.mysql_etl_read (id,order_id,user_id,gender,address,vip,telephone,id_card)
VALUES ('5','DD20210105','U0005','Male','Jiangsu Province Nanjing Gulou District','Regular
User','13600000005','320106200001051155');
INSERT INTO demo_mysql.mysql_etl_read (id,order_id,user_id,gender,address,vip,telephone,id_card)
VALUES ('6','DD20210106','U0006','Female','Jiangsu Province Suzhou Gusu District','Silver
VIP','13500000006','320508200001061166');
INSERT INTO demo_mysql.mysql_etl_read (id,order_id,user_id,gender,address,vip,telephone,id_card)
VALUES ('7','DD20210107','U0007','Male','Zhejiang Province Hangzhou West Lake District','Regular
User','15900000007','330106200001071177');
INSERT INTO demo_mysql.mysql_etl_read (id,order_id,user_id,gender,address,vip,telephone,id_card)
VALUES ('8','DD20210108','U0008','Male','Sichuan Province Chengdu Jinjiang District','Regular
User','15600000008','510104200001081188');
INSERT INTO demo_mysql.mysql_etl_read (id,order_id,user_id,gender,address,vip,telephone,id_card)
VALUES ('9','DD20210109','U0009','Female',' Shanghai Xuhui District','Bronze
VIP','13800000009','310104200001091199');
INSERT INTO demo_mysql.mysql_etl_read (id,order_id,user_id,gender,address,vip,telephone,id_card)
VALUES ('10','DD20210110','U0010','Male',' Beijing Chaoyang District','Regular
User','13900000010','110105200001101100');
```

2. Then return to the Data Development module, on the **Data Management** page within the **Data Management Directory**, click the top right ➕ button, select **New Database**, and the basic information configuration dialogue box will pop up.



©2013–2025 Tencent Cloud. All rights reserved.

3. Select the **EMR** provided **Hive System Source**, and create a database named **demo_hive**. This system source is automatically generated based on the bound EMR engine. Configuration attributes are as follows:
- Data Source Type: Hive
- Data Source: (Hive System Source)
- Database name: demo_hive
- Description: None



4. After completing the above preparations, you can proceed with the data development process by referring to the subsequent steps.

Data Development

Step 1: Create a new workflow

1. Enter the data workflow.

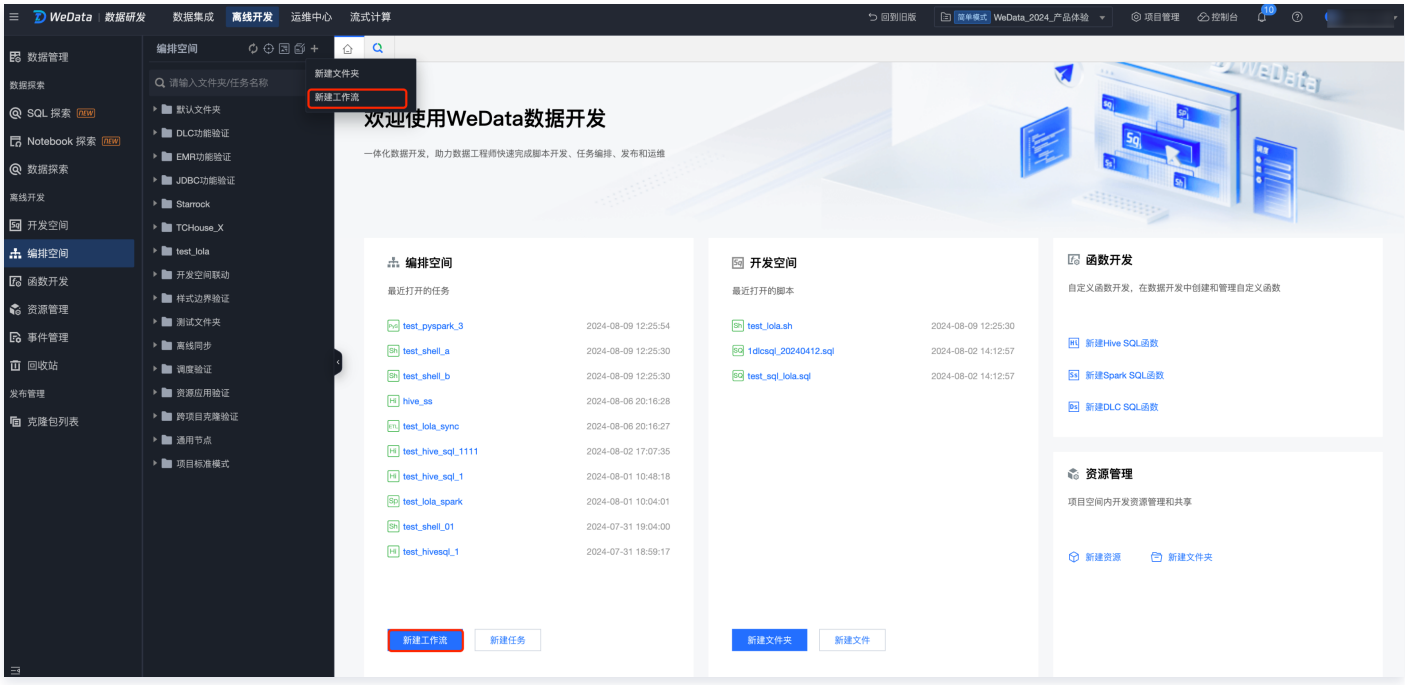
©2013–2025 Tencent Cloud. All rights reserved.



2. Create a workflow.

2.1 Workflow creation: In the Orchestration Space page, click **Create New Workflow**.

2.2 Orchestration directory creation: In the top-right corner of the orchestration space directory, click ➕, then from the expanded options, click **Create New Workflow**.



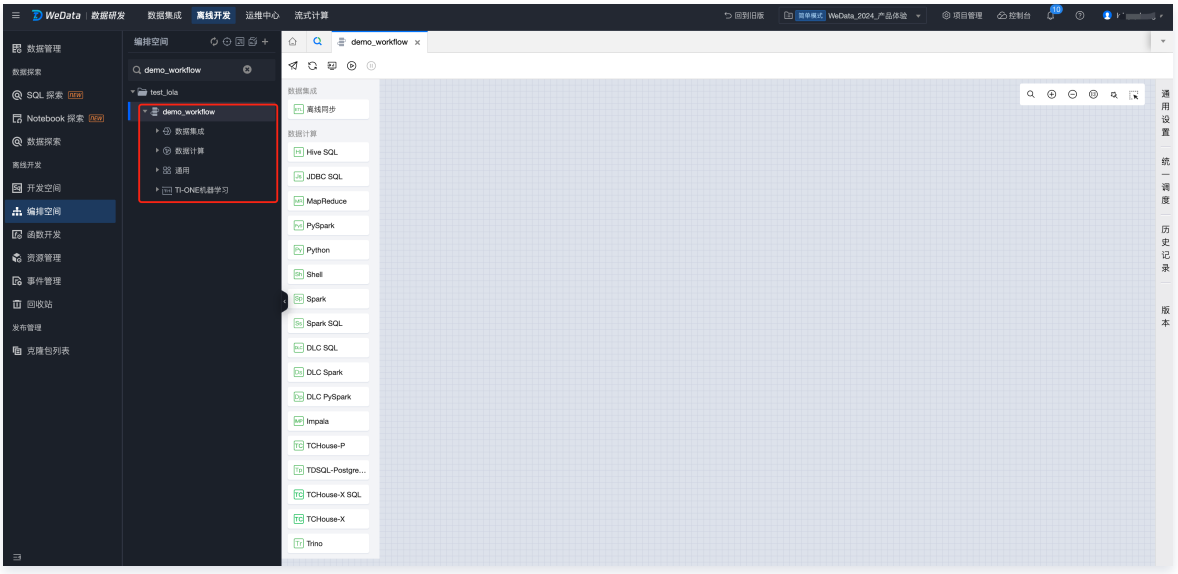
2.3 In the pop-up box, configure the basic information of the workflow, including the workflow name and the folder where the workflow is located. After confirming, the new step is completed. Here, a data workflow named **demo_workflow** is created in the **default folder**.

©2013-2025 Tencent Cloud. All rights reserved.



3. Open the workflow

   In the default folder of **Orchestration Space** on the left, find demo_workflow, move the left mouse button over it and double-click to open the **workflow canvas** of demo_workflow.



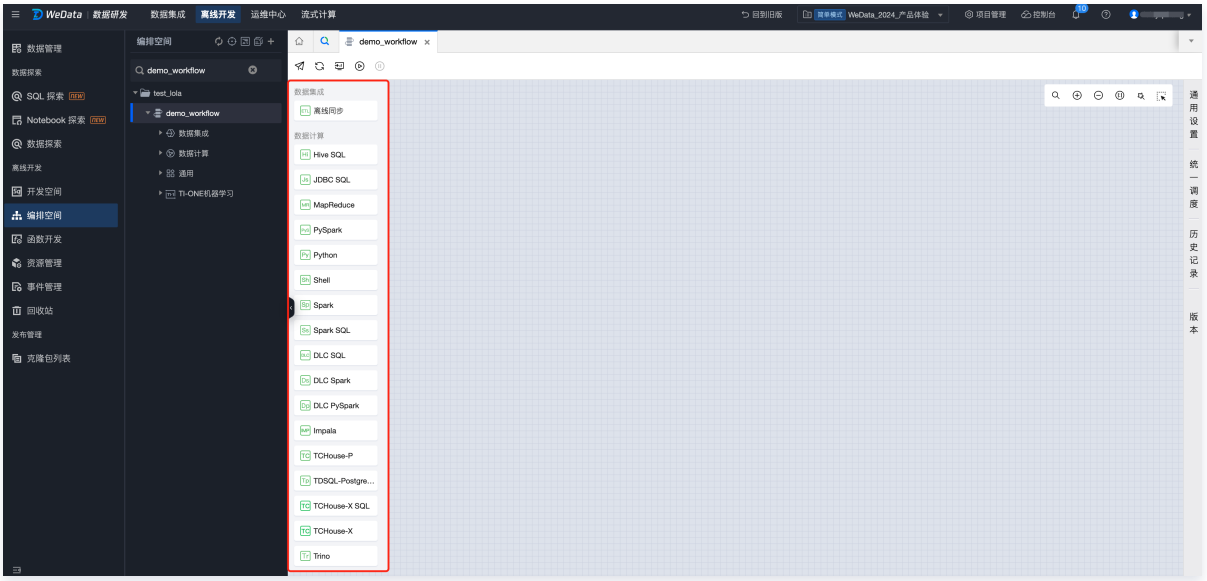
Step 2: Create a new task

WeData supports the creation of computing tasks manually by creating them on the canvas. Create a task named **demo_HiveSQL** of the type **Hive SQL** in the demo_workflow workflow.

- Canvas creation:

  1.1 In the demo_workflow workflow canvas, click the type of computing task you need to use.

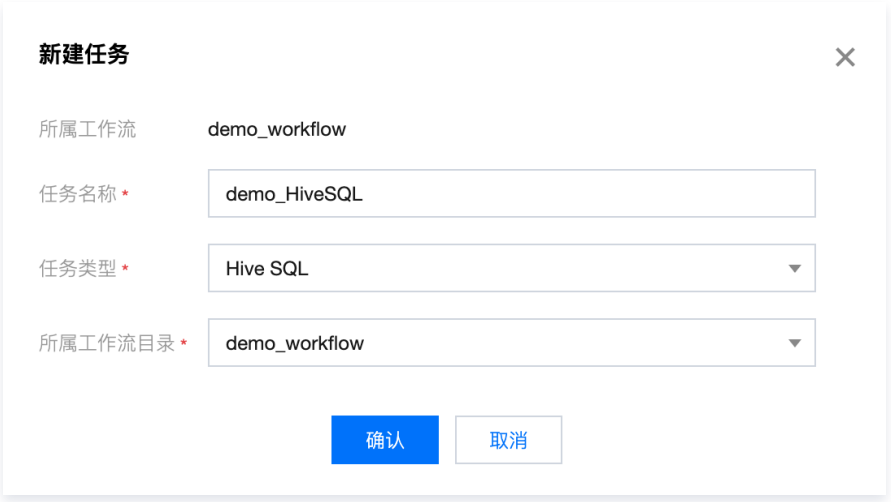
©2013-2025 Tencent Cloud. All rights reserved.



1.2 In the demo_workflow workflow, select the Hive SQL type task, and configure the basic information in the pop-up box, including the task name and configuration method.

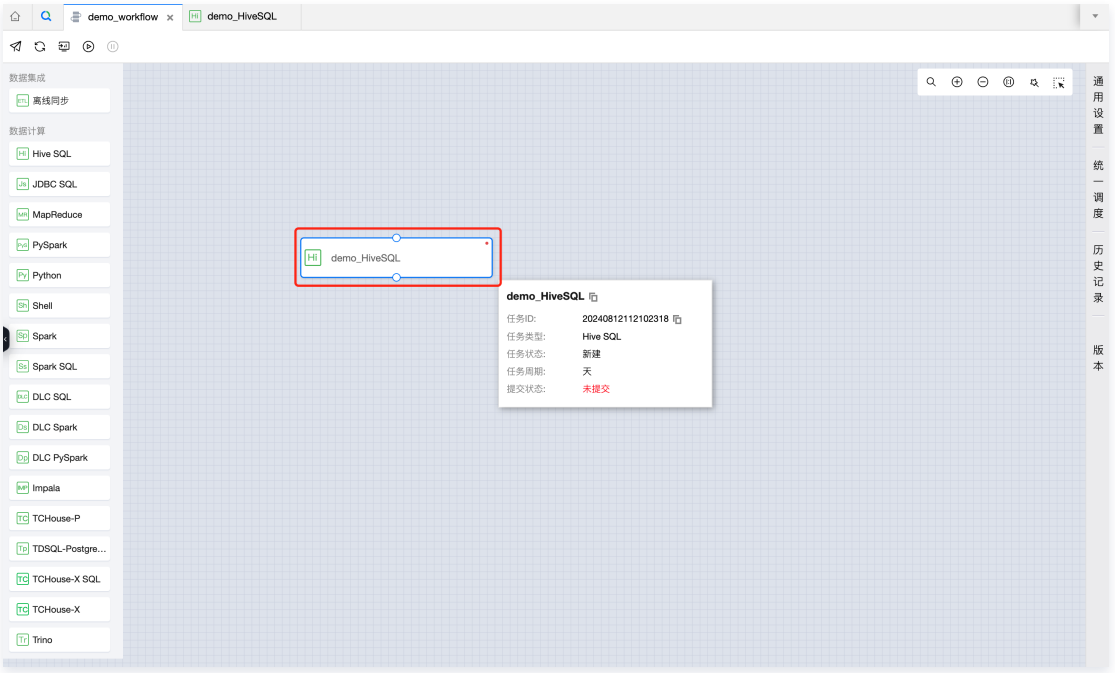
- Create a computing task named demo_HiveSQL with the default task type as Hive SQL.

1.3 Click Confirm to complete the creation of demo_HiveSQL.



1.4 Return to the demo_workflow canvas, and you will see demo_HiveSQL appear as a task node on the canvas.

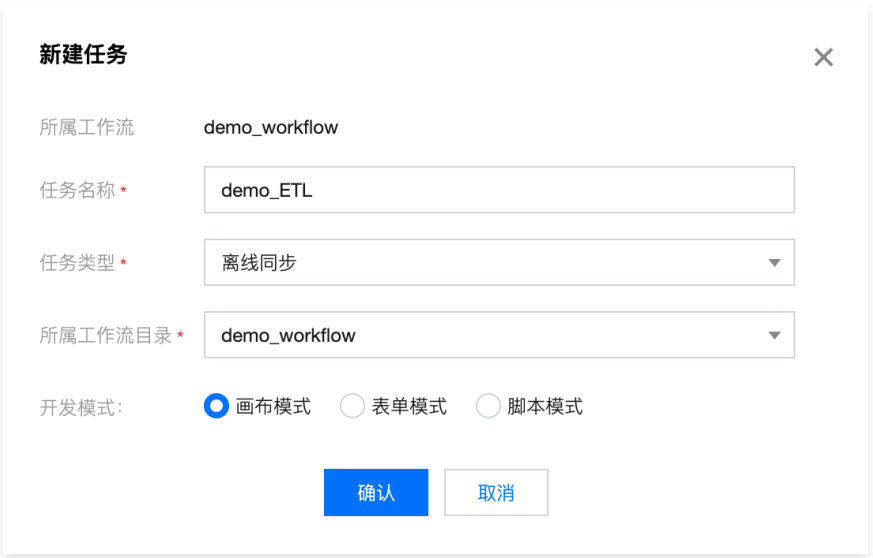
©2013-2025 Tencent Cloud. All rights reserved.



1.5 Next, follow the above steps to create two more computing tasks in the demo_workflow workflow by creating them on the canvas. These tasks are named **demo_ETL**, an **offline synchronization type task**, and **demo_Shell**, a **Shell** type task. The basic attribute configurations for both tasks are as follows:

demo_ETL:

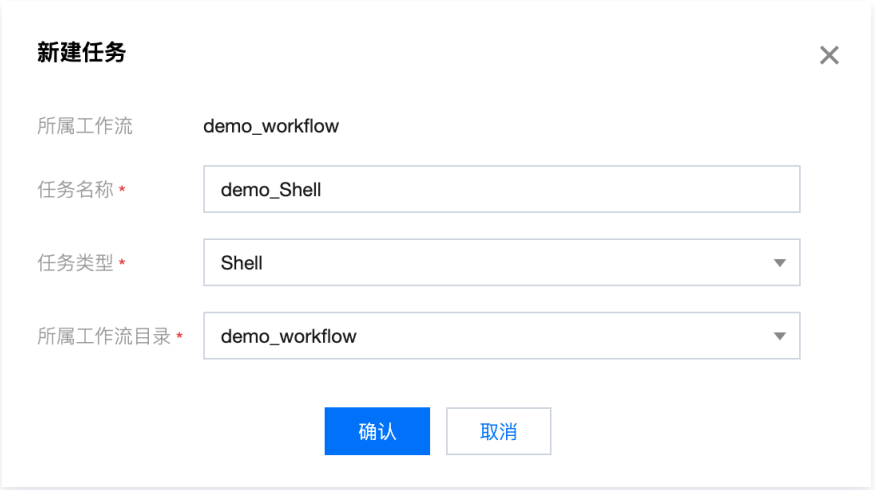
- Task name: demo_ETL
- Task type: Offline Synchronization (default)
- Development mode: Canvas mode



demo_Shell:

- Task name: demo_Shell
- Task type: Shell (default)

©2013–2025 Tencent Cloud. All rights reserved.



1.6 After creating demo_ETL and demo_Shell, you will have three computing tasks on the demo_workflow workflow canvas.



Step 3: Workflow Orchestration

Arrange the task nodes in order according to the data processing sequence.

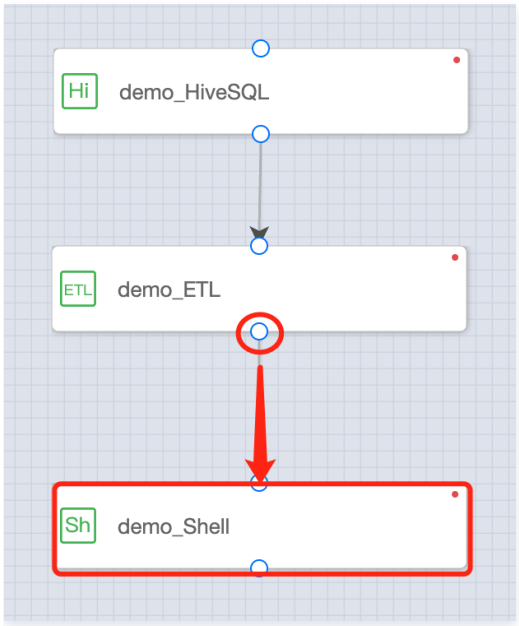
- The WeData data development process involves orchestrating computing tasks by using different types of task nodes in a certain sequence to form a data workflow.
- Arrange the three computing tasks on the canvas so that they execute in the order of demo_HiveSQL > demo_ETL > demo_Shell in the data processing flow. The specific orchestration method is as follows:
  - Drag the three task nodes on the canvas and arrange them in the desired execution order from top to bottom. Then, draw a connection cable from the connection point before one node to the next task node. Complete the workflow orchestration of the development task as shown in the figure below.

Step 4: Configure Task Nodes

- Node Configuration:
  - demo_HiveSQL: Use Hive SQL statements to create a Hive table named **hive_etl_write** in the demo_hive database.
  - demo_ETL: Use the mysql_etl_read table as the **data reading** node and the hive_etl_write table as the **data writing** node to perform data synchronization, syncing the data from the mysql_etl_read table to the hive_etl_write table.
  - demo_Shell: Use a Shell statement to output the completion time of the demo_workflow execution.

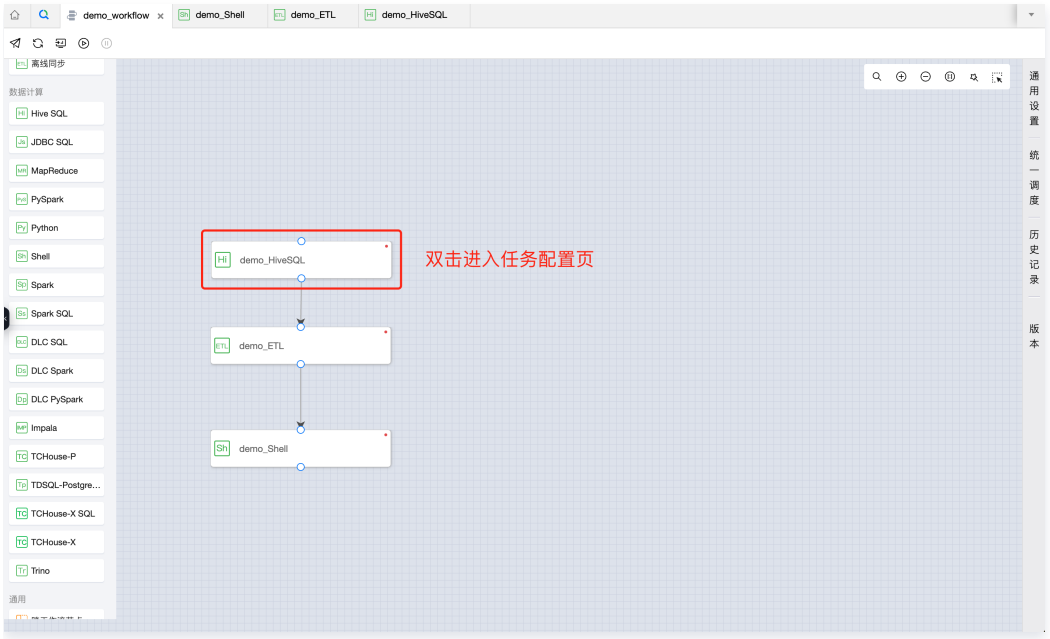
The specific configurations of the task nodes are as follows:

©2013–2025 Tencent Cloud. All rights reserved.



demo_HiveSQL Node

1. In the demo_workflow workflow canvas, double click on the demo_HiveSQL node to enter the corresponding task node configuration page.

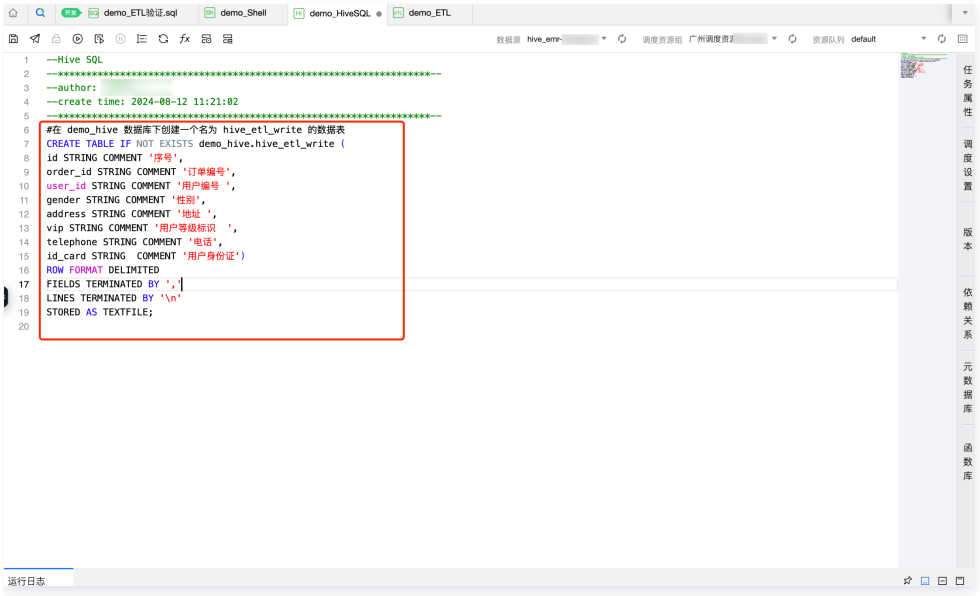


2. Write the Hive SQL statements into the demo_HiveSQL code editing area. The statements for creating the demo_hive database and the hive_etl_write table are as follows:

```
Create a table named hive_etl_write in the demo_hive database
CREATE TABLE IF NOT EXISTS demo_hive.hive_etl_write (
id STRING COMMENT 'ID',
order_id STRING COMMENT 'Order ID',
user_id STRING COMMENT 'User ID',
gender STRING COMMENT 'Gender',
address STRING COMMENT 'Address',
vip STRING COMMENT 'VIP Level Indicator',
telephone STRING COMMENT 'Telephone',
id_card STRING COMMENT 'User ID Card')
ROW FORMAT DELIMITED
```

©2013–2025 Tencent Cloud. All rights reserved.

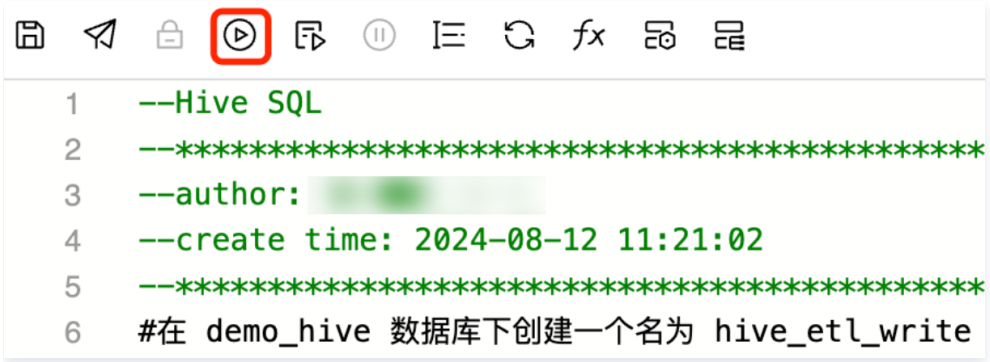
```
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n'
STORED AS TEXTFILE;
```



3. After the statement is written, ensure that the hive_etl_write data table is created under the Hive system source where the demo_hive database is located. Pay attention to whether the selected data source is appropriate.



4. After the statement is written and the data source is selected, click the button in the upper left corner to debug and run the task node.



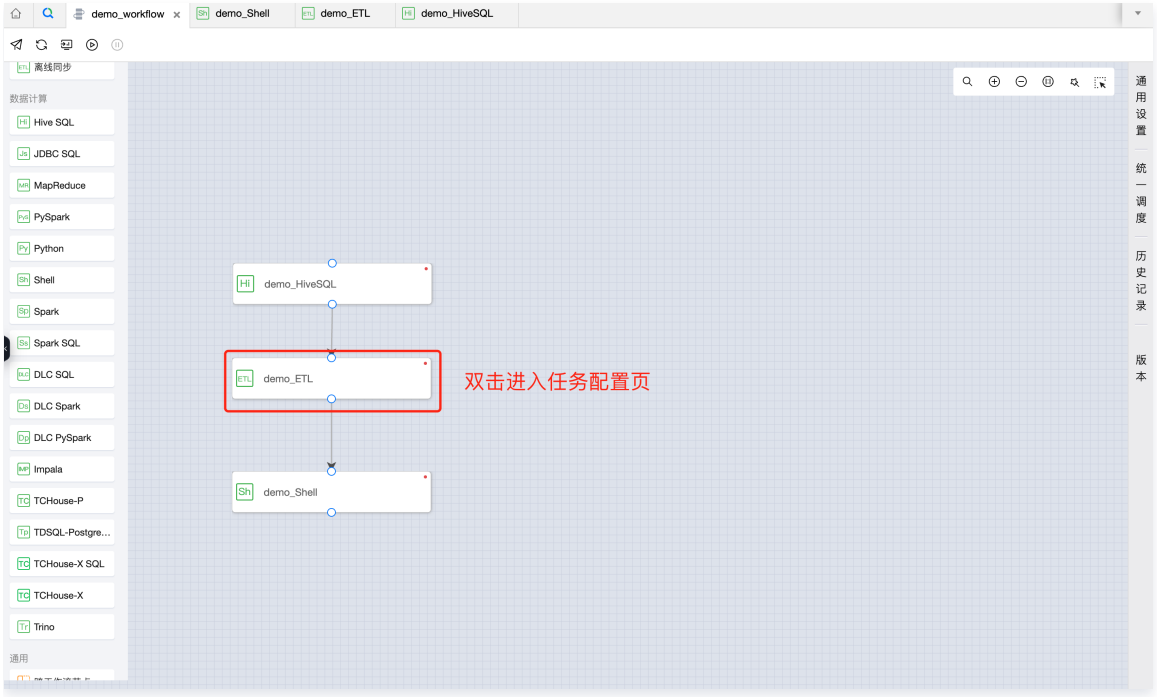
5. The running process and results can be viewed through the **Running Log**, making it convenient to monitor script execution progress and troubleshoot issues. Upon successful execution, the demo_hive database and the hive_etl_write data table will be created. At this point, the demo_HiveSQL task node configuration is complete.

©2013–2025 Tencent Cloud. All rights reserved.



demo_ETL Node

1. On the demo_workflow canvas, double-click the demo_ETL node to enter the configuration page of the corresponding task node.



2. **Read Node:** On the left side, select the MySQL type under the read types to create an ETL read node for a MySQL data table.

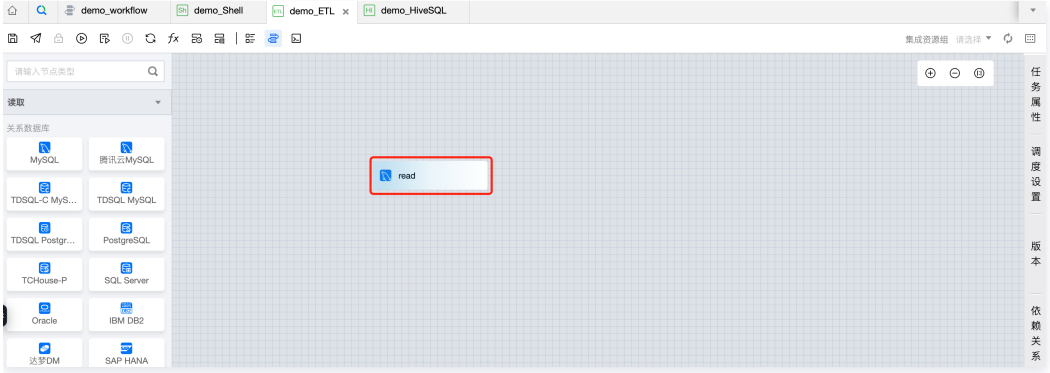
©2013–2025 Tencent Cloud. All rights reserved.



3. Click the MySQL type, and name it **read** in the new read node pop-up.

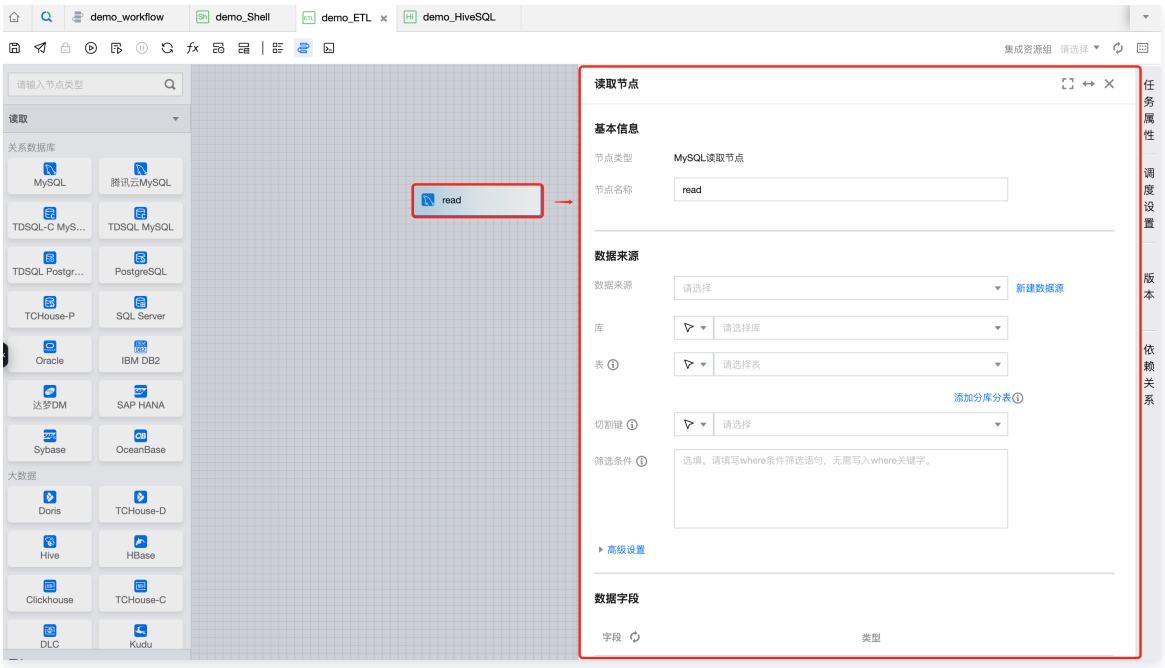


4. After the read node is created, you can see the read node on the demo_ETL canvas.



5. Double-click the read node to pop up the configuration items on the right side of the canvas.

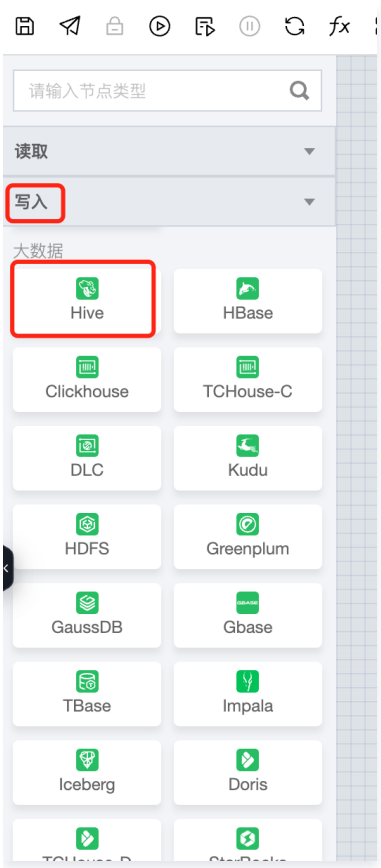
©2013-2025 Tencent Cloud. All rights reserved.



6. Fill in the information and configure it according to the following attributes:
   - **Basic Information**
     - Node Name: read (default)
   - **Data Source**
     - Data Source: demo_mysql
     - Library: demo_mysql
     - Table: mysql_etl_read
     - Cutting Key: None (default)
     - Filter Condition: None (default)
     - Advanced Settings: None (default)
   - **Data Fields (default)**
7. After completing the configuration of the read node, click the **Off** button in the upper right corner. In the confirmation pop-up to close the node, select **Save** to complete the read node configuration.
8. Next, return to the demo_ETL canvas page to configure the write node.

Write Node: Select Hive type from the **write** options on the left side to create an ETL write node that writes data into a Hive table.

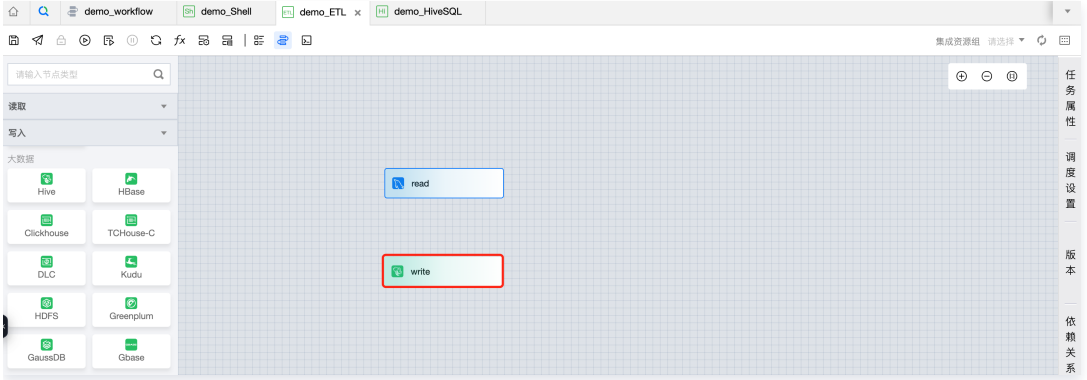
©2013–2025 Tencent Cloud. All rights reserved.



9. Click the Hive type, and name it **write** in the new read node pop-up.

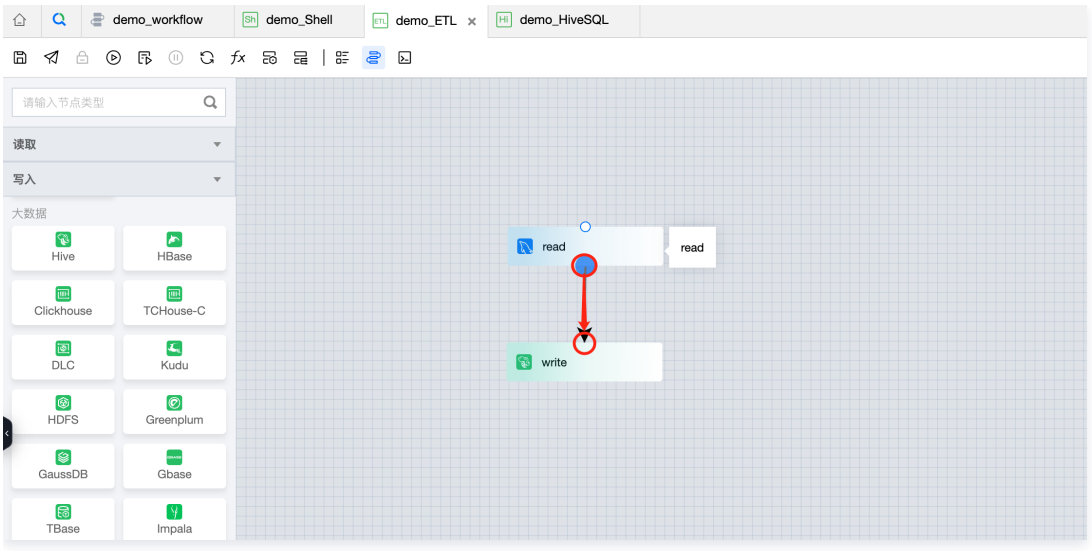


10. After the write node is created, you can see the write node in the demo_ETL canvas.

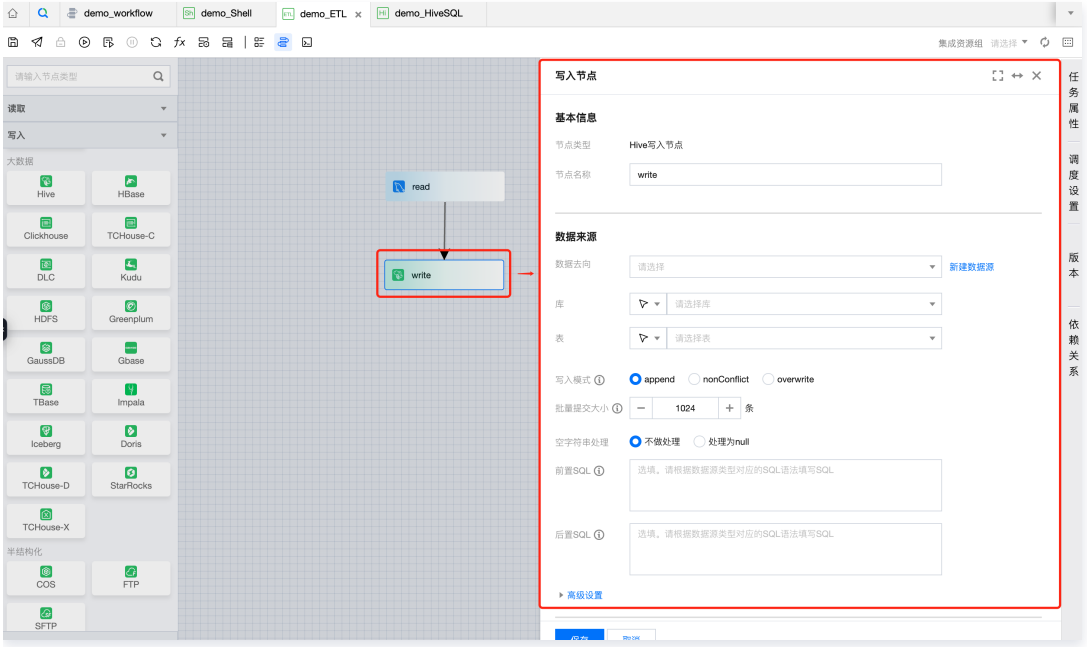


11. Drag a connection cable from the connection point below the read node and connect it to the connection point above the write node. This establishes the data synchronization link between the read node and the write node.

©2013–2025 Tencent Cloud. All rights reserved.



12. After connecting the read and write nodes, double-click the write node to pop up the configuration items on the right side of the canvas.



13. Configure the write node according to the following attributes:

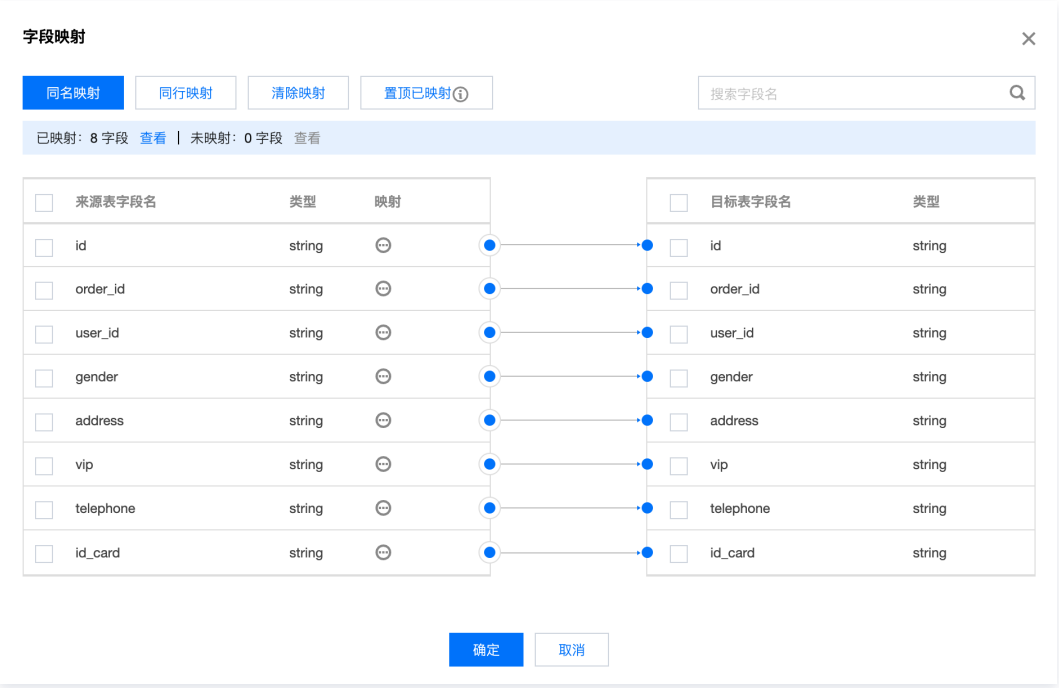
- **Basic Information**
  - Node Name: write (default)
- **Data Source**
  - Data Source: (EMR's Hive System Source)
  - Library: demo_hive
  - Table: hive_etl_write
  - Write Mode: append (default)
  - Batch Submission Size: 1024 (default)
  - Empty String Handling: No processing (default)
  - Pre SQL: None (default)
  - Post SQL: None (default)
- **Advanced Settings: None (default)**
  - Data Fields (default)
  - Field Mapping

©2013–2025 Tencent Cloud. All rights reserved.

14. Click **Configure Field Mapping** to pop up the data table field mapping configuration between the read node and the write node.



15. In the field mapping configuration, select **Same Name Mapping** to establish a field mapping relationship between the read node and the write node based on the same field names, thus matching the data synchronization relationship between read and write.

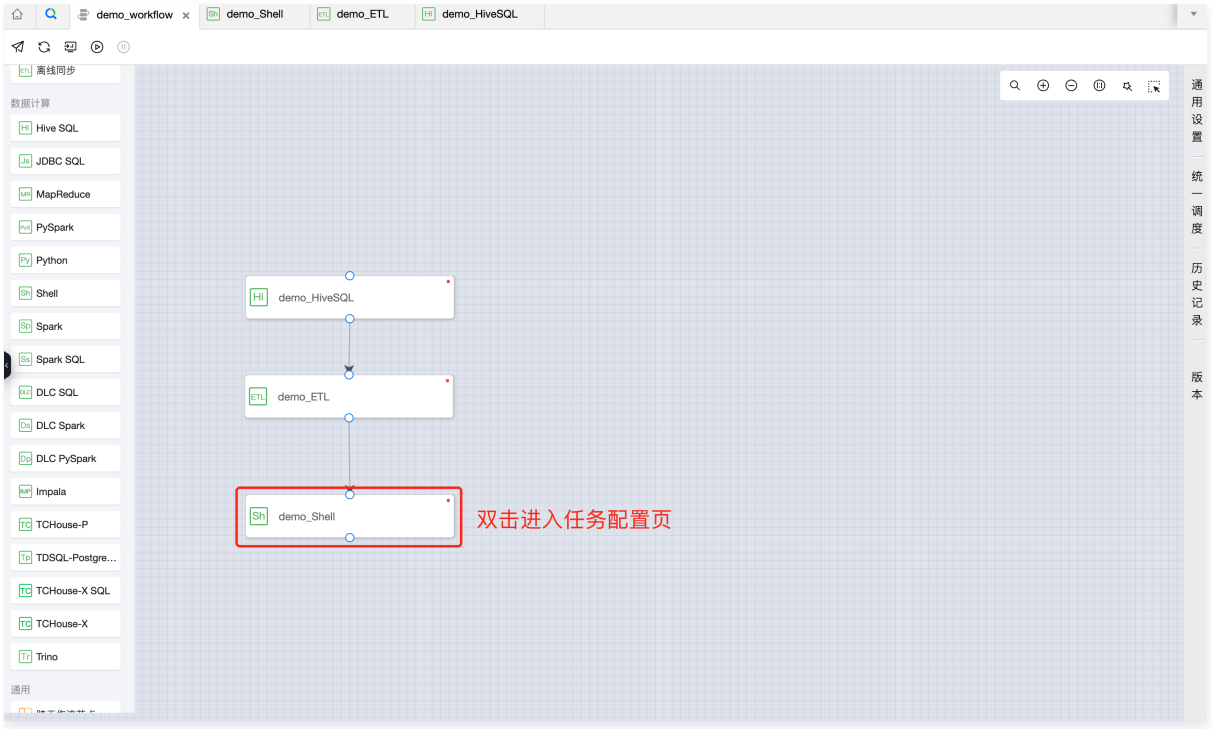


16. After completing the configuration of the write node, click the **Close** button in the upper right corner. In the confirmation pop-up to close the node, select **Save** to complete the write node configuration.
17. At this point, the demo_ETL task node configuration is complete.

demo_Shell Node

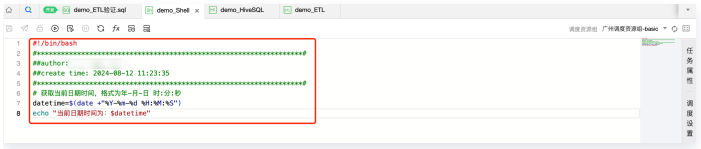
1. On the demo_workflow canvas, double-click the demo_Shell node to enter the configuration page of the corresponding task node.

©2013-2025 Tencent Cloud. All rights reserved.

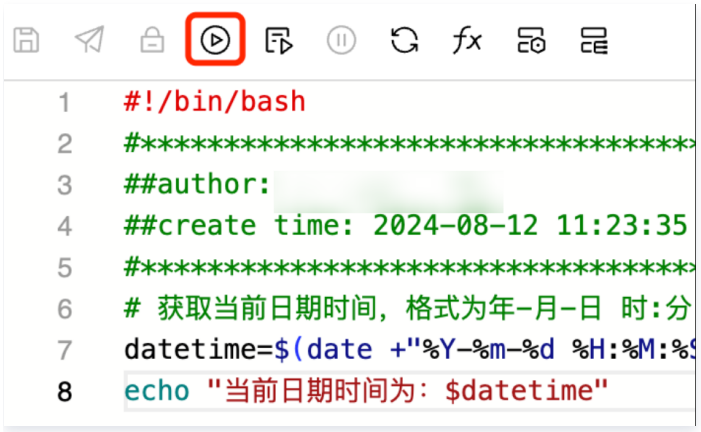


2. Write the Shell statement in the demo_Shell code editing area. An example statement to get the execution completion date and time of the demo_workflow is as follows:

```
# Get the current date and time in the format year-month-day hour:minute:second
datetime=$(date +"%Y-%m-%d %H:%M:%S")
echo "The current date and time is: $datetime"
```

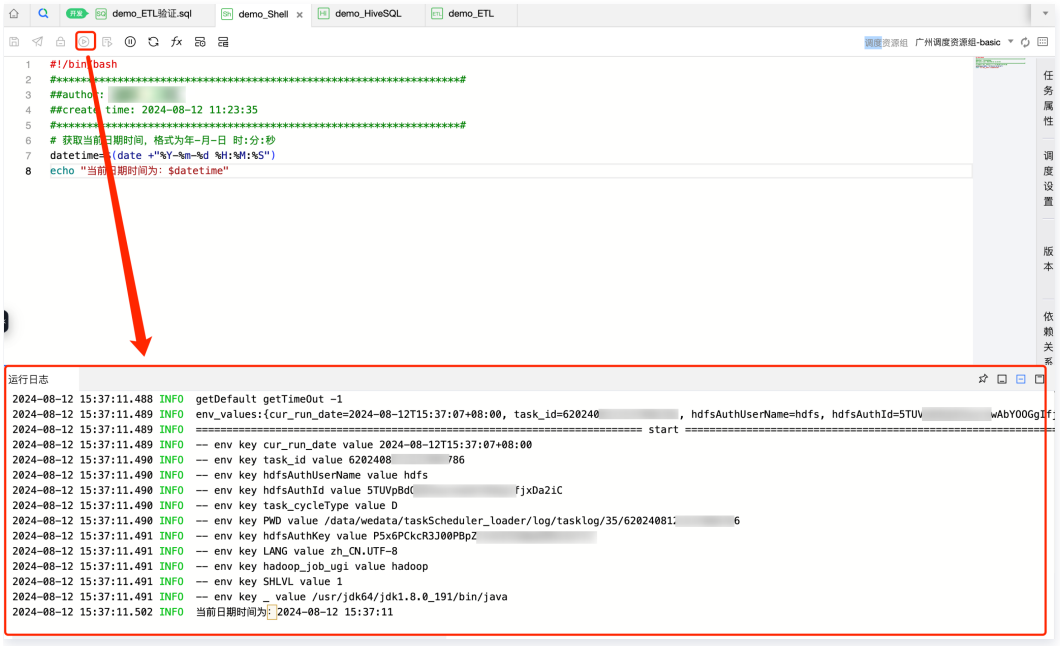


3. After the statement is written, click the  button in the upper left corner to debug and run the task node.



4. The running process and results can be viewed through the **Running Log**, making it convenient to monitor script execution progress and troubleshoot issues. Upon successful execution, the completion time of the current Shell task will be output in the running log. At this point, the demo_Shell task node configuration is complete.

©2013–2025 Tencent Cloud. All rights reserved.



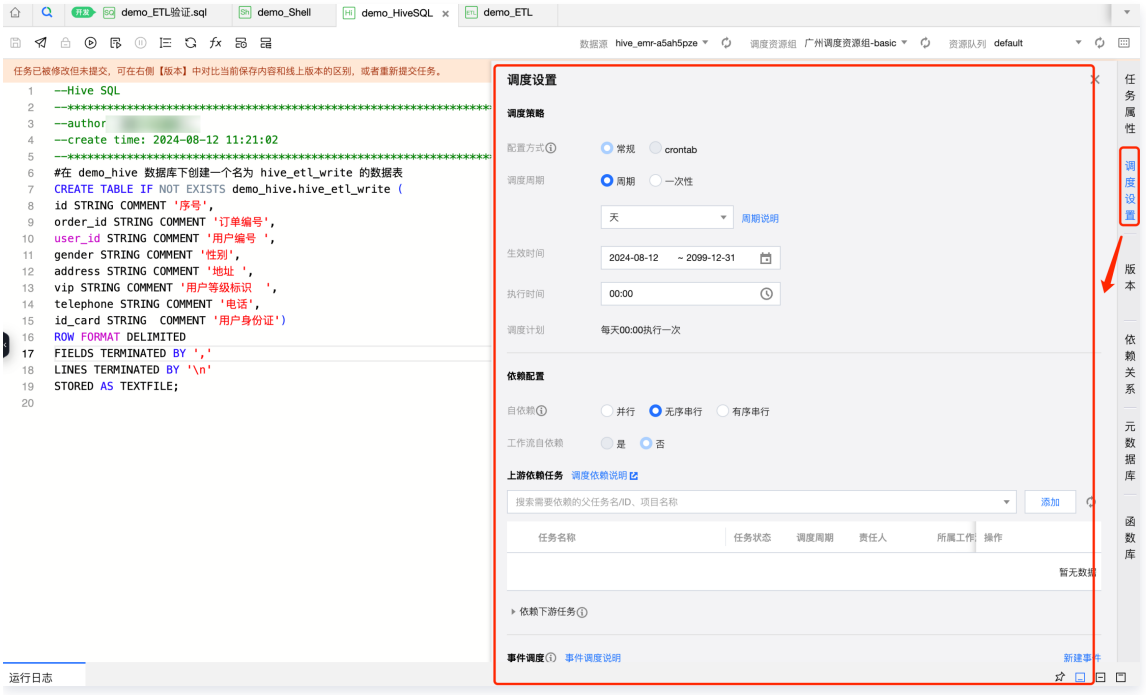
5. Now that the three task nodes have been configured according to the plan, the next step is to configure the scheduling policy for the workflow.

Step 5. Configure Scheduling

WeData provides **Task Scheduling** to configure scheduling policies individually for each task node. Here, the demo_workflow selects task scheduling for configuration. Assign a scheduling policy to each of the three task nodes.

demo_HiveSQL

1. Enter the configuration page of the demo_HiveSQL task node, click **Scheduling Settings** in the right sidebar to configure the scheduling policy for this task node.



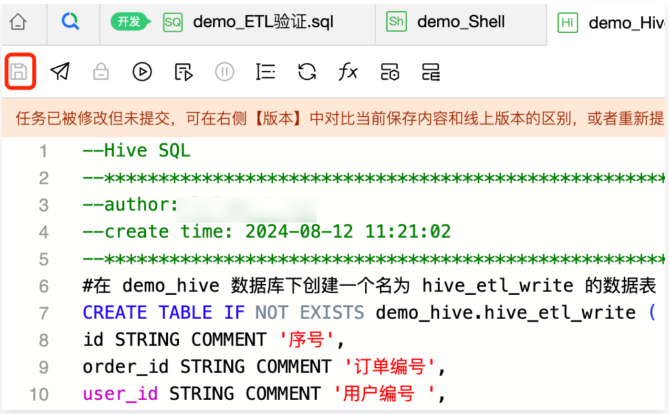
2. Configure the scheduling policy for the demo_HiveSQL execution with the following attributes:

- **Scheduling Policy**
  - Scheduling Period: Cycle – Day (default, executed once a day)
  - Effective Time: (default)

©2013–2025 Tencent Cloud. All rights reserved.

- Execution Time: 00:00 (default, execution time)
- **Event Scheduling**
  - Configure Listening Event: None (default)
- **Dependency Configuration**
  - Self Dependency: Unordered Serial (default)
  - Workflow Self Dependency: No (default)
- **Upstream Dependent Tasks (default)**
  - Parameter Passing (default)
- **Advanced Settings (default)**

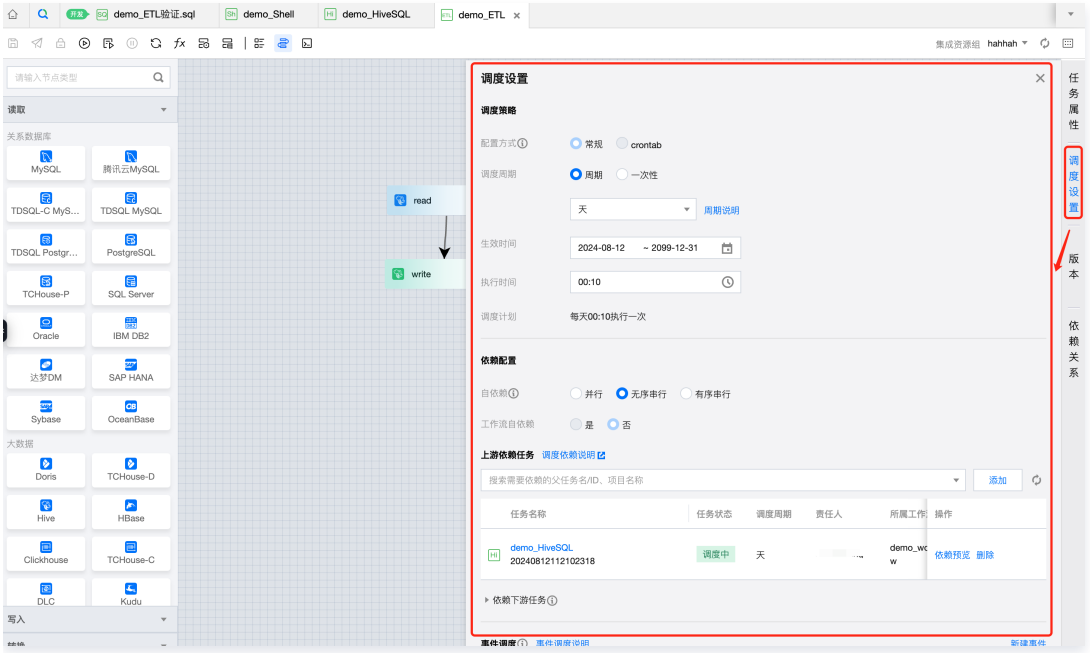
3. After configuration, click in the top left corner of the task node configuration page to save.



4. In subsequent steps, after publishing the demo_HiveSQL task, it will execute once a day at **00:00** according to the current scheduling policy.

demo_ETL

1. Enter the configuration page of the demo_ETL task node, click **Scheduling Settings** in the right sidebar to configure the scheduling policy for this task node.



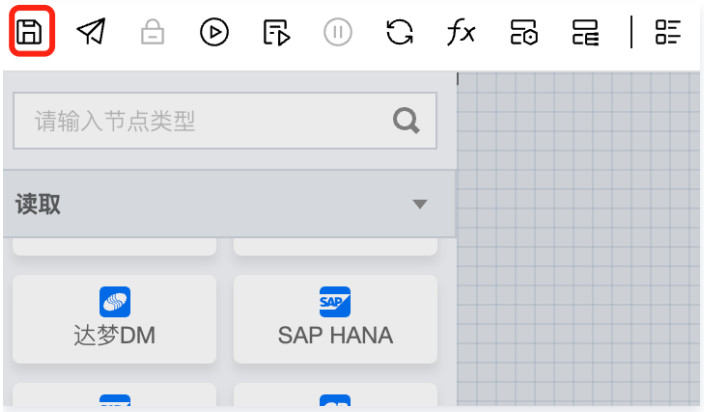
2. Configure the scheduling policy for the demo_ETL execution with the following attributes:

- **Scheduling Policy**
  - Scheduling Period: Cycle – Day (default, executed once a day)
  - Effective Time: (default)

©2013–2025 Tencent Cloud. All rights reserved.

- Execution Time: 00:10 (execution time)
- **Event Scheduling**
  - Configure Listening Event: None (default)
- **Dependency Configuration**
  - Self Dependency: Unordered Serial (default)
  - Workflow Self Dependency: No (default)
- **Upstream Dependent Tasks (default)**
- **Advanced Settings (default)**

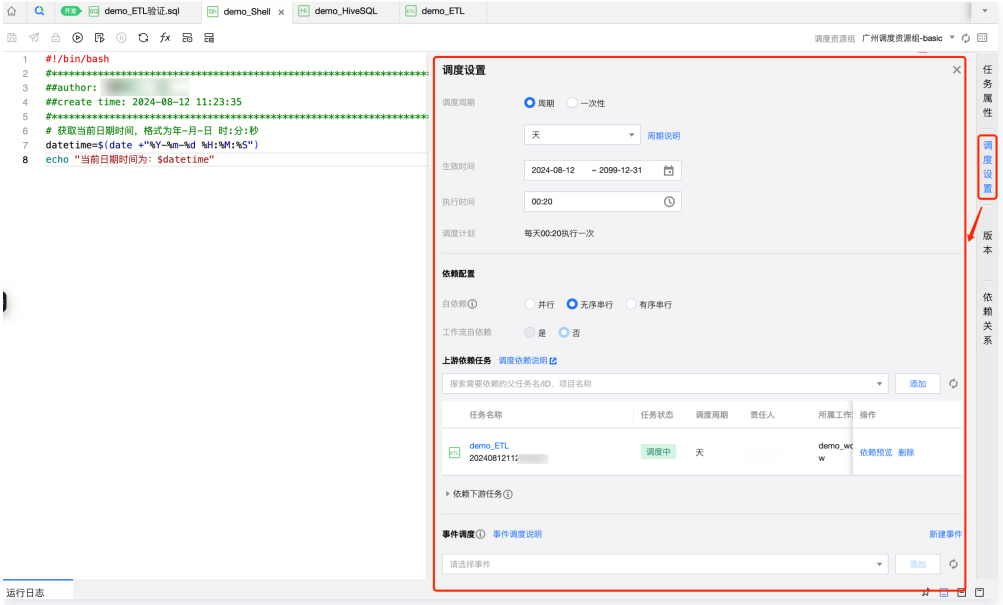
3. After configuration, click in the top left corner of the task node configuration page to save.



4. In subsequent steps, after publishing the demo_ETL task, it will execute once a day at **00:10** according to the current scheduling policy.

demo_Shell

1. Enter the configuration page of the demo_Shell task node, click **Scheduling Settings** in the right sidebar to configure the scheduling policy for this task node.



2. Configure the scheduling policy for the demo_Shell execution with the following attributes:

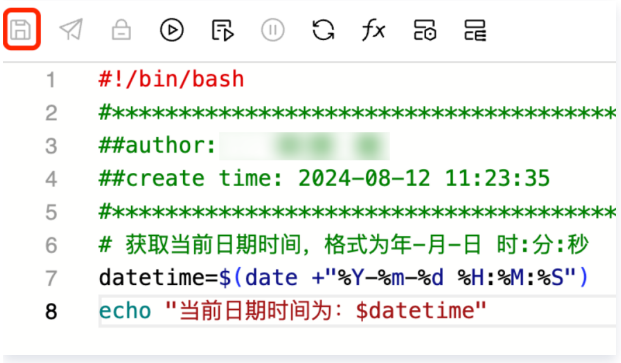
- **Scheduling Policy**
  - Scheduling Period: Cycle – Day (default, executed once a day)
  - Effective Time: (default)
  - Execution Time: 00:20 (execution time)
- **Event Scheduling**

©2013–2025 Tencent Cloud. All rights reserved.

- Configure Listening Event: None (default)
- Configure event-driven hooks (default)

- **Dependency Configuration**
  - Self Dependency: Unordered Serial (default)
  - Workflow Self Dependency: No (default)
- **Upstream Dependent Tasks (default)**
- **Parameter Passing (default)**
- **Advanced Settings (default)**

3. After configuration, click in the top left corner of the task node configuration page to save.



4. In subsequent steps, after publishing the demo_Shell task, it will execute once a day at **00:20** according to the current scheduling policy.
5. At this point, the scheduling policies for the demo_HiveSQL, demo_ETL, and demo_Shell task nodes have all been configured. Once the workflow is published in Step Six, the scheduling policies for the task nodes will take effect.

Step Six: Configure Task Dependencies

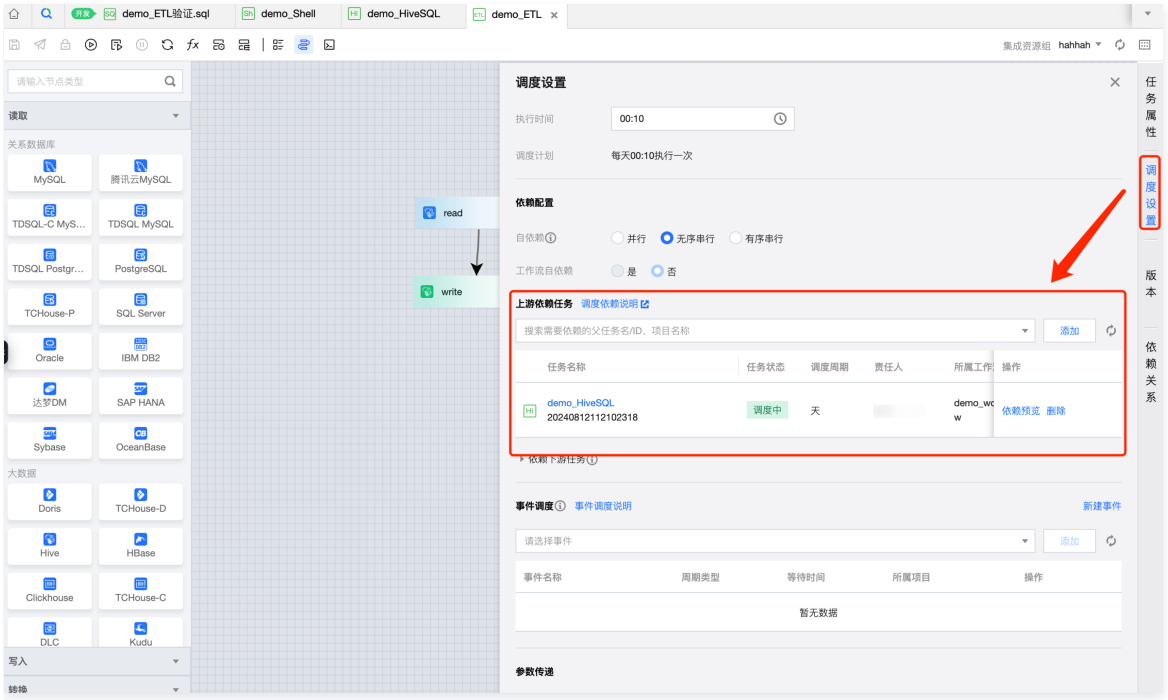
The demo_HiveSQL, demo_ETL, and demo_Shell task nodes use connections to orchestrate upstream and downstream dependencies in the workflow. In this step, you need to confirm whether the dependencies between task nodes are correct. According to the current orchestration, the upstream and downstream relationships between task nodes should be as follows:

Task Node	Upstream Dependent Task	Scheduling Time
demo_HiveSQL	No	Execute once daily at 00:00
demo_ETL	demo_HiveSQL	Execute once daily at 00:10
demo_Shell	demo_ETL	Execute once daily at 00:20

You can manage dependency relationships through the following steps:
Return to the canvas page of the demo_workflow workflow. The dependency configuration method for task nodes is the same. Here, we take the demo_ETL task node as an example. Select demo_ETL to enter the task node configuration page.

In the demo_ETL task node, click the **Scheduling Settings** feature on the right side. In the **Upstream Dependent Task** area, you can view the upstream dependent tasks of the current task node, and you can also manually add or remove dependencies. This feature allows for managing the upstream and downstream dependencies of task nodes in the workflow and supports the configuration of task node dependencies across workflows.

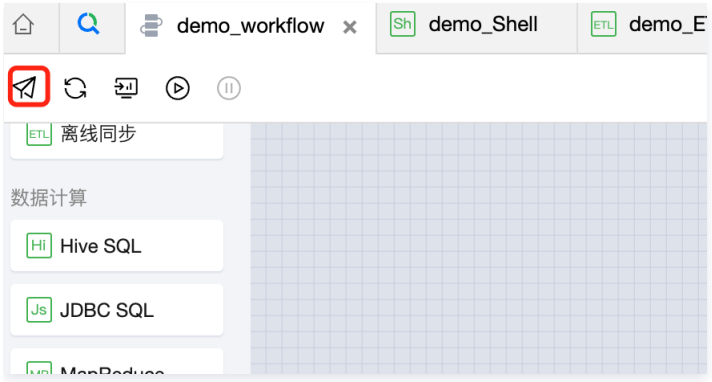
©2013-2025 Tencent Cloud. All rights reserved.



Submit for OPS

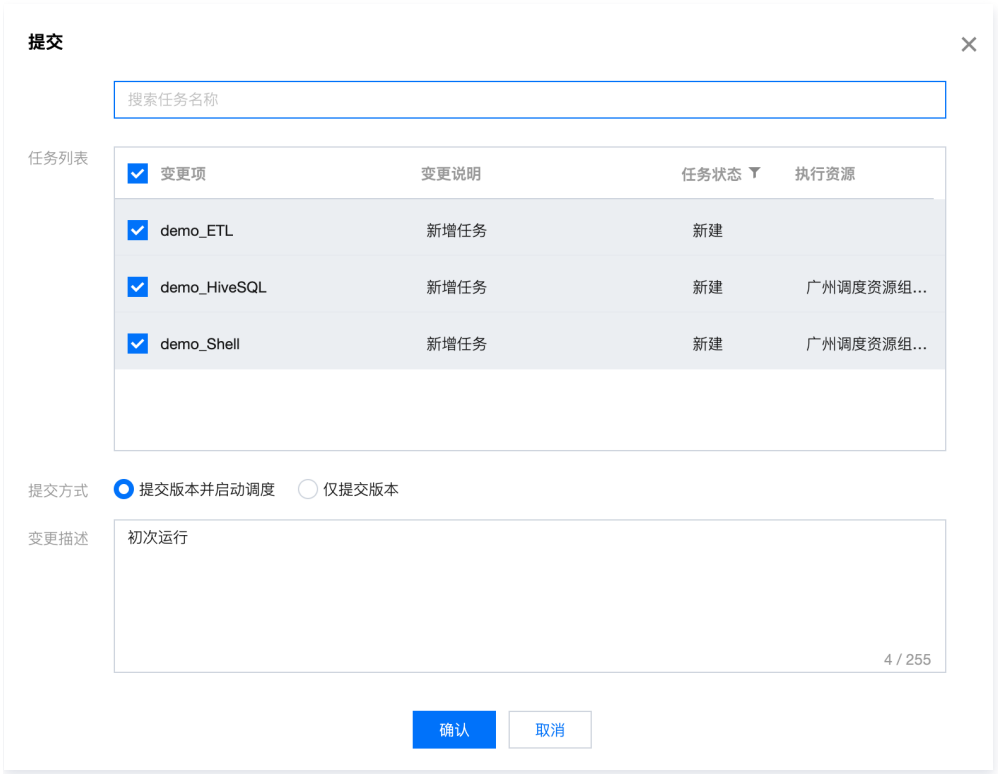
Following the above configuration steps, a simple data development workflow has been designed. Next, after submission and publication, this workflow can be managed in the OPS phase.

1. On the canvas page of demo_workflow, click the workflow in the upper left corner.

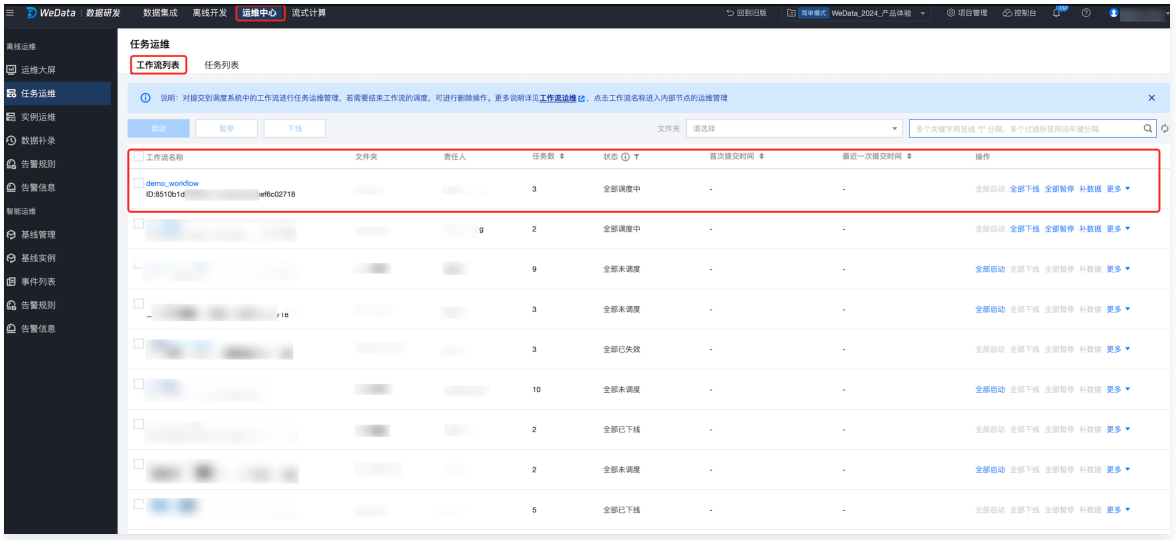


2. In the confirmation popup for submission, you can see the task nodes involved in OPS for this workflow. After confirming that everything is correct, choose **Submit Version and Start Scheduling** as the submission method, and fill in any content in the change description field to publish the demo_workflow workflow to the OPS phase.

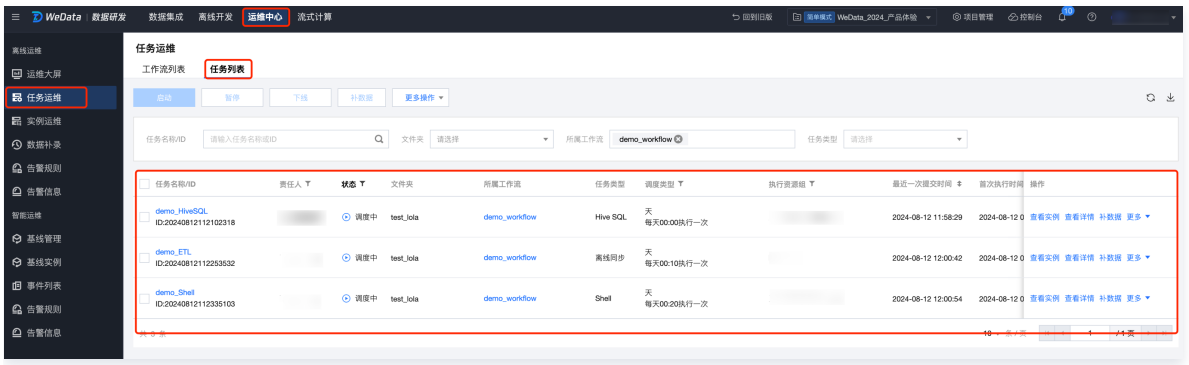
©2013–2025 Tencent Cloud. All rights reserved.



3. Subsequently, you can manage data workflows and task nodes through **Task OPS**.
4. In **Task OPS > Workflow List**, monitor the running status of the demo_workflow workflow.

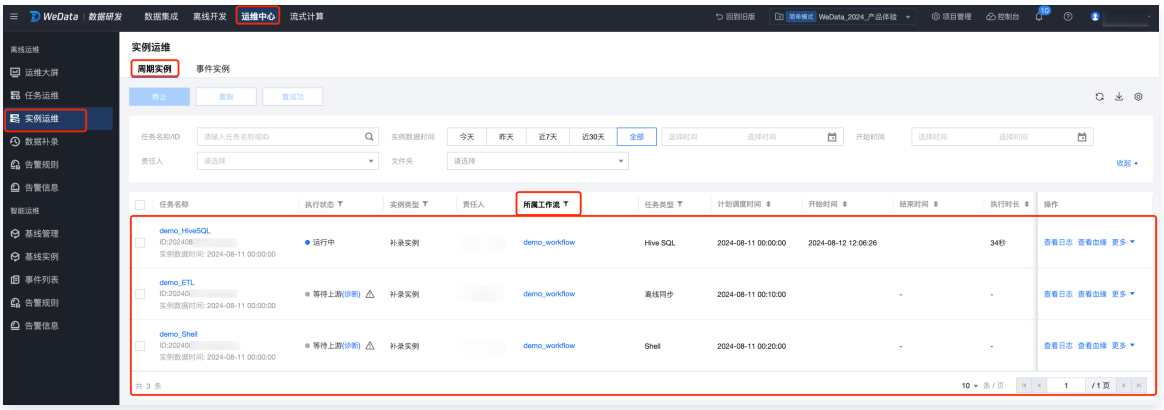


5. In **Task OPS > Task List**, monitor the running status of the demo_HiveSQL, demo_ETL, and demo_Shell task nodes.



©2013–2025 Tencent Cloud. All rights reserved.

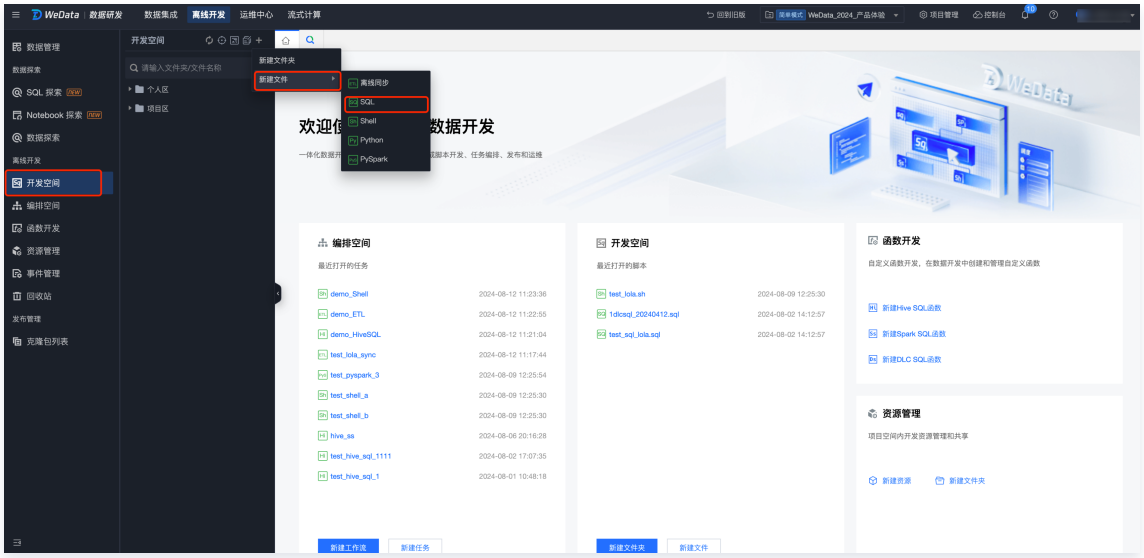
6. The demo_workflow workflow, according to the scheduling policy, will execute daily at 00:00. After execution, three task nodes will generate corresponding task instances. You can monitor the execution status of task instances through **Instance OPS > Cycle Instance**.
7. Use the filter to find instances generated by task nodes in the demo_workflow workflow. The filter content is as follows:
   - Planned scheduling time: (Select the date of the day after the workflow submission to OPS, because according to the demo_workflow scheduling policy, it will execute at 00:00 the next day to generate task instances)
   - Belonging workflow: demo_workflow



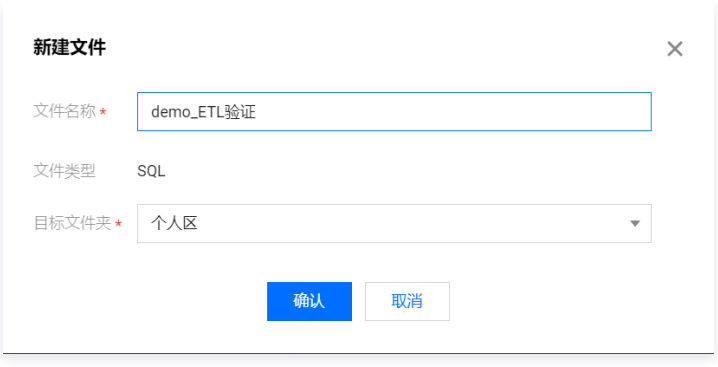
8. Once the instances run successfully, the data processing scripts within the task nodes will be effective. You can verify this:

   8.1 **Verification I: Verify the execution result of the demo_ETL task node. Check if the data in the mysql_etl_read data table has been synchronized to the hive_etl_write data table.**

   8.1.1 Enter the **Development Space**, click ➕ button above the **Development Space directory** on the left, select **Create New File** to create a **SQL** development script.

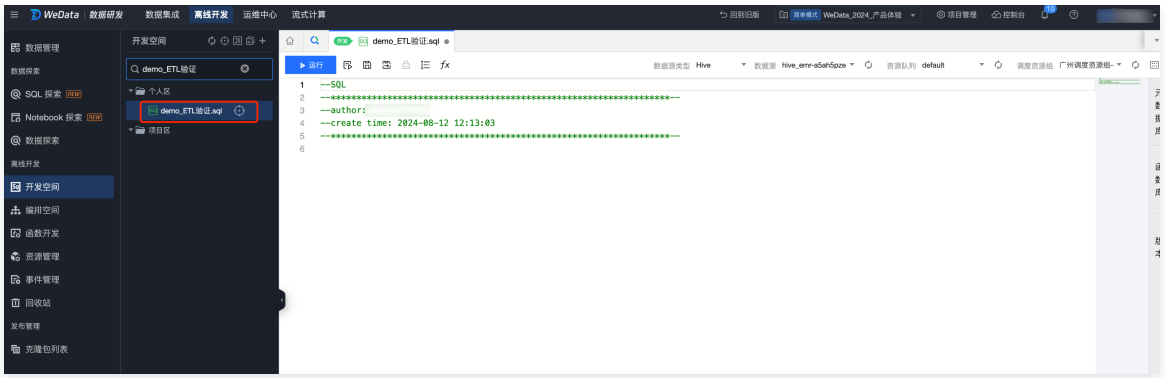


   8.1.2 Name the SQL development file **demo_ETL Verification** and create it in the **Personal Area**.



©2013–2025 Tencent Cloud. All rights reserved.

8.1.3 In the Personal Area of the development space directory, find the demo_ETL Verification SQL file, double-click to open it, and enter the query statement in the SQL editor to check if the data in the hive_etl_write data table meets the expectations.



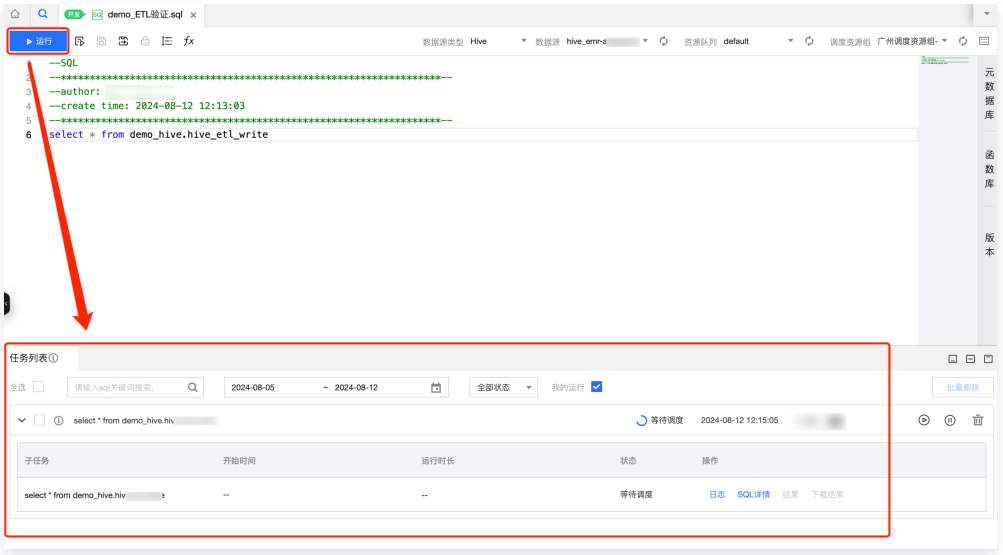
8.1.4 Query example statements are as follows:

```
select * from demo_hive.hive_etl_write
```

8.1.5 After writing the query statement, make sure that the query's **data source** is the same as the data source where the data table is located. For example, in this chapter's example, the hive_etl_write data table is created in the Hive system source provided by EMR, so select the system source for the data source. After confirming that the SQL statement and data source are correct, click **Run**.

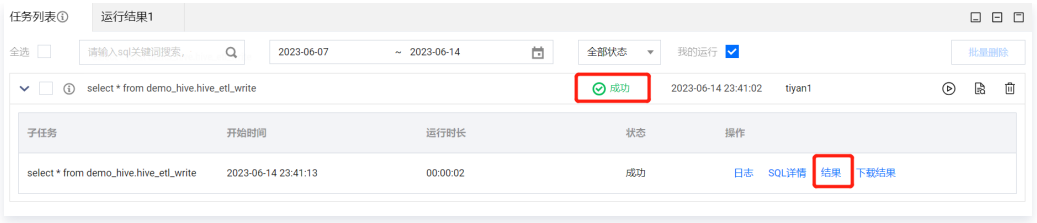


8.1.6 After running the SQL statement, the **SQL task list** will pop up at the bottom of the SQL file interface, showing the status of the SQL statement's execution.

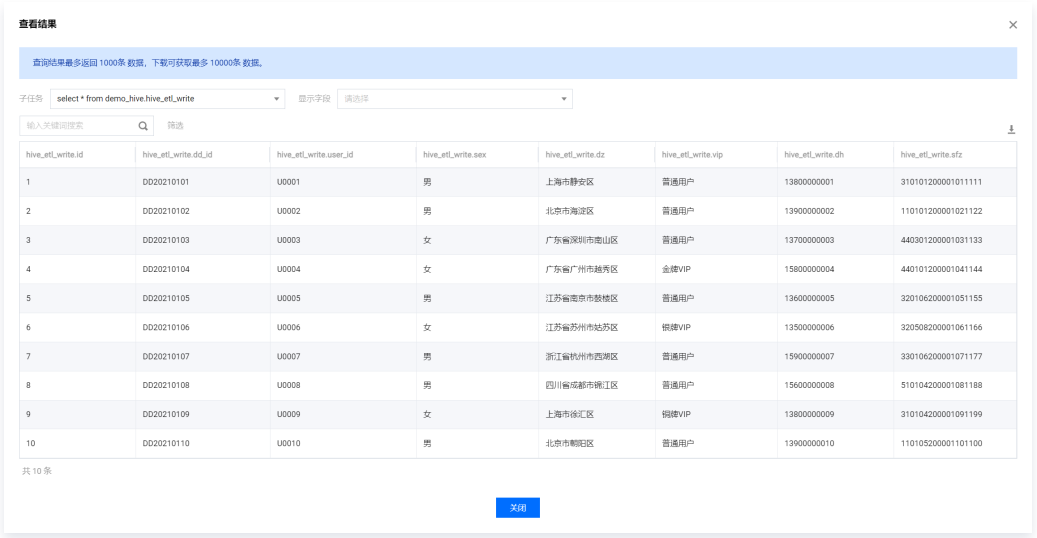


8.1.7 Wait for the SQL to **succeed**, then click the **Result** button under the SQL task operation column to view the query results.

©2013-2025 Tencent Cloud. All rights reserved.

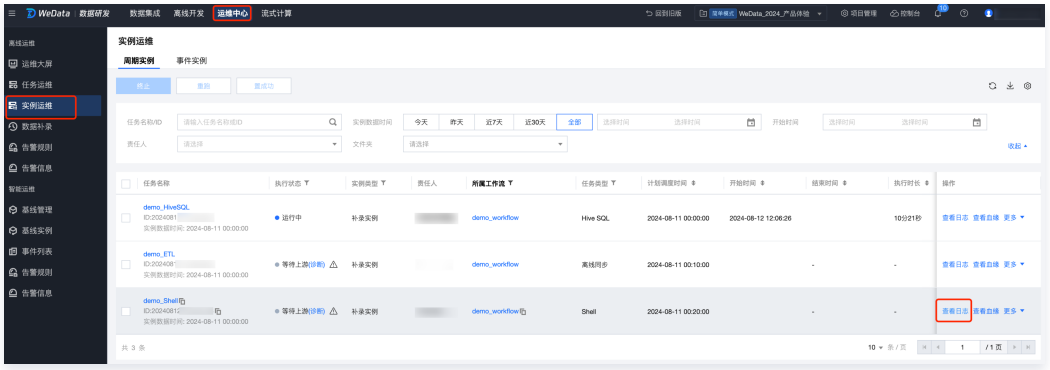


8.1.8 You will find that the data in the mysql_etl_read data table has been successfully synchronized to the hive_etl_write data table.



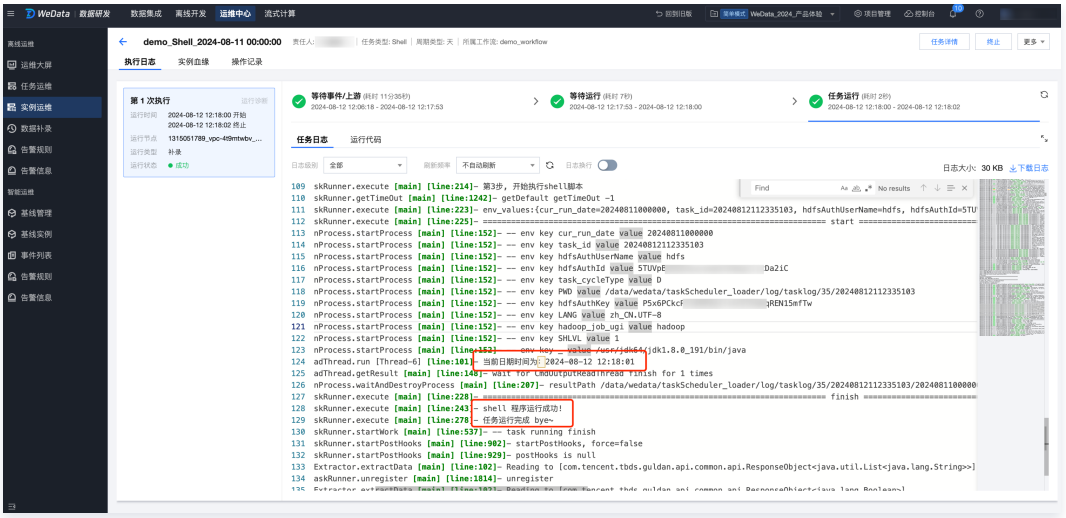
8.2 **Verification II: Verify the execution result of the demo_Shell task node. Obtain the date and time of workflow execution.**

8.2.1 Enter the **Instance OPS > Cycle Instance** page, filter to find the instance generated by the demo_Shell task node, click the log under the corresponding instance operation column to enter the instance run log page.



8.2.2 On the log page, scroll through the logs to view the Shell task's execution process and the acquired date and time of the workflow execution.

©2013–2025 Tencent Cloud. All rights reserved.



©2013–2025 Tencent Cloud. All rights reserved.

Engine Usage Instructions
Elastic MapReduce (EMR)

Last updated: 2025-06-11 18:04:33

WeData supports integration with the EMR computing engine to create tasks such as Hive, MR, Spark, Trino, and Kyuubi, and implements table management and task configuration in EMR. This article introduces the precautions for task development on WeData in EMR clusters.

Using Background

Elastic MapReduce (EMR) is a secure, low-cost, and highly reliable open-source big data platform based on cloud-native technology and pan-Hadoop ecosystem open-source technology. It provides easy-to-deploy and manage open-source big data components such as Hive, Spark, HBase, Flink, StarRocks, Iceberg, and Alluxio, helping customers efficiently build a cloud-based enterprise-level data lake technology architecture. Based on WeData and EMR, you can quickly build a data warehouse based on open-source big data. For details, see EMR Create and Manage Clusters.

Use Limits

Limit Type	Explanation
Elastic MapReduce (EMR) ClusterType	• Currently, WeData supports EMR on CVM and EMR on TKE clusters. • Currently, the versions supported for creation by EMR are as follows: "EMR-V2.0.1" "EMR-V2.2.0" "EMR-V2.3.0" "EMR-V2.5.0" "EMR-V2.6.0" "EMR-V2.7.0" "EMR-V3.0.0" "EMR-V3.1.0" "EMR-V3.2.0" "EMR-V3.2.1" "EMR-V3.3.0" "EMR-V3.4.0" "EMR-V3.5.0"
WeData features	The EMR task types supported in WeData data development include: HiveSQL, SparkSQL, Spark, MapReduce, PySpark, Shell, Impala, Trino. Among them, the Kyuubi Data Source requires the use of SparkSQL-type tasks to use.

Usage Process

The main process of using EMR in WeData includes the following steps:

Preparations

Preparation Type	Operation Instructions	Reference Link
Elastic MapReduce (EMR)	To ensure smooth use of EMR-related table creation and data development features in WeData, the EMR cluster must meet basic configurations. At least Hive and Spark services need to be installed in the EMR cluster. Other services, if used in WeData, must also be enabled accordingly in the EMR. For example, ranger, Kyuubi, etc.	Create and manage EMR clusters. See Creating a Cluster.
WeData	Bind the EMR cluster, obtain the latest cluster configuration from the EMR cluster. Configure the corresponding authentication method and account mapping.	-

©2013-2025 Tencent Cloud. All rights reserved.



Task Development

- Create a workflow

Task development is implemented through data workflow orchestration to achieve procedural execution of computing tasks. Before creating a computing task, it is necessary to create a data workflow, then orchestrate the computing task running process in the workflow.

- Create an EMR node

WeData performs task development based on the EMR engine. Different types of component services on the EMR cluster are integrated into WeData as system data sources by default (Hive, Trino, kyuubi, and Impala). Users can choose appropriate components to create EMR service resources according to business needs and bind the EMR cluster to projects in WeData. To create an EMR component service, for details, see new component of EMR.

- Task Development

Complete the binding between the EMR engine and the WeData project. In the created data workflow, create the computing task types supported by EMR. During the configuration process of the task node, use the system data source provided by EMR for task development and debugging.

- submit the task

After using the EMR system source data configuration and debugging without errors, save the corresponding computing task. Then submit the workflow where the computing task is located for publication, and it can be scheduled for execution in the Ops center.

Subsequent Operations

After completing the EMR task development, you can perform EMR data asset management, data quality monitoring, and data security management in WeData. This ensures normal output of EMR data and enables process control for data quality and data security.

©2013–2025 Tencent Cloud. All rights reserved.

Using Data Lake Compute (DLC)

Last updated: 2025-05-30 16:30:40

WeData supports integration with the DLC storage-compute separation computing engine, provides DLC database/table management functionality, and enables agile and efficient data lake analytics computation. Based on Spark and Presto capabilities, using standard SQL allows completing federated analysis and computing of Cloud Object Storage (COS) and multi-source databases.

Using Background

Tencent Cloud Data Lake Compute (DLC) provides agile and efficient managed data lake analytics and computing services. Users do not need to perform traditional data layered modeling, drastically reducing the preparation time for massive data analysis, effectively improving enterprise data agility.

Use Limits

Limit Type	Explanation

Data Lake Compute (DLC)	• Currently, WeData supports data management, analysis queries, and computing tasks for DLC type library tables.
	• Currently, the supported computing engine versions of DLC are as follows:
	  ○ standard engine
	  ○ Presto:Standard-P 1.0
	  ○ Spark:Standard-S 1.1,Standard-S 1.0,Standard-S 1.1(native)
	  ○ SuperSQL engine
	  ○ Spark SQL:SuperSQL-S 1.0,SuperSQL-S 3.5
	  ○ Spark job:Spark 2.4,Spark 3.2,Spark 3.5
	  ○ Presto:SuperSQL-P 1.0

WeData functionality	• task types supported in WeData data development include DLC SQL, DLC Spark, DLC PySpark. • WeData supports creating DLC tables and DLC functions.

Usage Process

The main process of using DLC in WeData includes the following steps:

Preparations

Preparation Type	Operation Instructions	Reference Link
Data Lake Compute (DLC)	To ensure smooth use of DLC-related table creation, data development, and data exploration features in WeData, the DLC cluster must meet basic configurations. For example, to use DLC's Spark Job Engine in WeData, you need to create a Spark Job Engine in DLC and grant the corresponding user permission to use the engine.	Create and manage DLC engine Please see Data Engine Description.
WeData	Bind the DLC cluster, obtain the latest cluster configuration from the DLC cluster.	By default, newly created projects will automatically use a dynamic key to interconnect with DLC. Data permission and engine permission management Please refer to DLC Permission Management.

Task Development

©2013-2025 Tencent Cloud. All rights reserved.

Creating a Workflow

Task development is implemented through procedural execution based on data workflow orchestration to create computing tasks. Before creating a computing task, you need to create a data workflow, then orchestrate the computing task running process within the workflow.

Create a DLC Node

WeData performs task development based on the DLC engine. After binding the DLC cluster with the project in WeData, the DLC system data source will be integrated into WeData. For details, see DLC Engine Access. Currently, the DLC SQL in the orchestration space only supports the DLC system source.

Task Development

Complete the creation of DLC-supported computing task types in the existing data workflow after binding the DLC engine to the WeData project. During the task node configuration process, use the system data source provided by DLC for task development and debugging.

Submit a Task

After configuring and debugging with the DLC system source data, save the corresponding computing task, then submit the workflow containing the computing task for publication. It can then be scheduled and executed in the Ops center.

Subsequent Operations

After DLC task development is completed, you can perform DLC metadata management, task operation and maintenance monitoring, and data quality monitoring in WeData to ensure normal DLC data output; you can perform multi-source joint queries and data analysis in the data exploration feature.

©2013-2025 Tencent Cloud. All rights reserved.

Tencent Cloud Data Warehouse TCHouse-P

Last updated: 2024-08-23 10:13:19

WeData supports integration with Tencent Cloud Data Warehouse TCHouse-P (Tencent Cloud House-P, TCHouse-P), providing PB-level data warehousing capabilities in data development, based on MPP large-scale parallel processing data capabilities.

Background

Tencent Cloud TCHouse-P distributed relational database is a general-purpose, high-performance database system designed for massive online real-time data analysis. For configuring the TCHouse-P cluster, see Project Configuration. For more information about TCHouse-P, see Tencent Cloud Data Warehouse TCHouse-P.

Use Limits

Limit	Restrictions
Tencent Cloud Data Warehouse TCHouse-P	Different types of clusters can be created in TCHouse-P, and different clusters have different connection types in WeData.
WeData feature	• WeData supports using SQL scripts with TCHouse-P data sources and arranging TCHouse-P as a computational task node in the data development workflow. • In WeData, it is currently recommended to use TCHouse-P clusters.

Getting Started

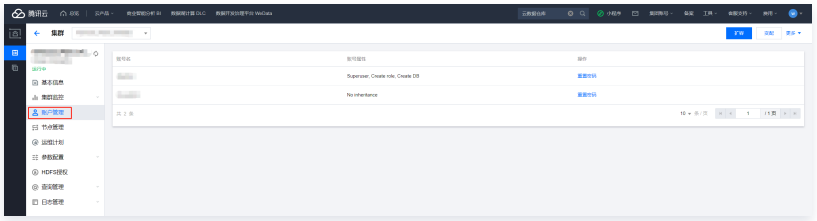
Preparations

Preparation Category	Operation Description	Reference Links
Tencent Cloud Data Warehouse TCHouse-P	To ensure the smooth use of TCHouse-P data in WeData for data development, it is necessary to ensure that the TCHouse-P cluster meets basic configuration requirements. Create TCHouse-P clusters in the Tencent Cloud Data Warehouse TCHouse-P product. 	Creating and managing TCHouse-P engine: TCHouse-P Cluster Management.
WeData	WeData requires binding TCHouse-P clusters in the storage and computation engine configuration.	TCHouse-P cluster resource configuration and permission policy: TCHouse-P CAM

©2013-2025 Tencent Cloud. All rights reserved.



When binding clusters, you need the cluster username and password, which can be obtained in the Account Management of the corresponding TCHouse-P cluster.



Task Development

- Creating workflow

Task Development achieves the procedural execution of computing tasks based on Data Workflow orchestration. You need to create a Data Workflow before creating a computing task, and then orchestrate the computing task execution process within the workflow.

- Create TCHouse-P Node

WeData develops tasks based on TCHouse-P clusters. After binding the TCHouse-P cluster to the project in WeData, the TCHouse-P system data source will be integrated into WeData. For details, please refer to TCHouse-P Engine Access. Currently, only TCHouse-P system sources are supported for TCHouse-P tasks in the orchestration space.

- Task Development

After binding the TCHouse-P engine to the WeData project, create a TCHouse-P type computing task in the created Data Workflow. During the task node configuration process, use the system data source provided by TCHouse-P for task development and debugging.

- Submitting the job

Once the task development and debugging with TCHouse-P system source data are completed, save the corresponding computing task. Then submit and publish the workflow containing the computing task, and it can be scheduled and executed in the Operations Center.

©2013-2025 Tencent Cloud. All rights reserved.

Data Management

Last updated: 2024-11-22 16:56:26

Enter the Data Management page

1. Log in to WeData Console.
2. Click **Project List** in the left menu to find the target project for data management features.
3. After selecting the project, click to enter the **Data Development** module.
4. Click **Data Management** in the left menu.

Data Management Overview

Currently, WeData Data Management supports the creation of Hive and DLC database tables in the EMR and DLC engine system sources.

> **Note:**
> Data sources can be displayed in the data management directory only after binding the computational storage engine on the project management page.

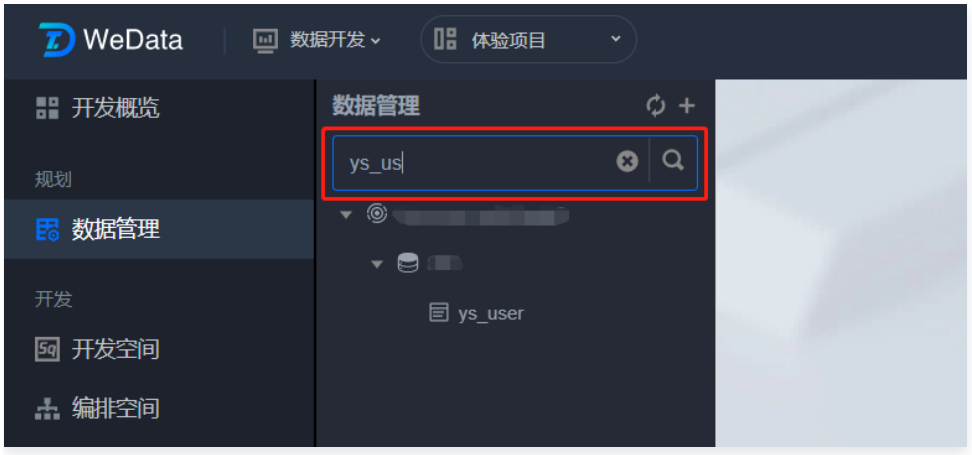
Data Management Directory

The directory tree is used to display the hierarchical structure and relationships of all database tables in the data sources. This feature enables the following functions:

- Quickly locate the target table. Using the directory tree feature, users can quickly locate the target table, improving operational efficiency and reducing operation time and error possibilities.
- Show relationships between database tables. The directory tree feature allows users to clearly see the hierarchical structure and relationships between database tables, making it easier to analyze and understand the associations and dependencies between them.
- Manage and maintain the data warehouse. Using the directory tree feature, users can classify and manage databases by data warehouse layers, facilitating database maintenance and adjustments, such as deleting or renaming tables, fields, and other operations.
- Convenient search feature. Using the search box feature of the directory tree, users can easily browse and search for database tables and navigate to the target table for operations.

Database and Table Search

The search feature helps users quickly locate and browse target database tables or datasets. It provides a clear hierarchical view and quick search capabilities, making it easy for users to find the data they need, thereby improving the efficiency of data management and queries.
Enter the database or data table name in the search window to search the corresponding structure from the Library Table Catalog. The search feature supports fuzzy search.



©2013-2025 Tencent Cloud. All rights reserved.

Refresh Directory

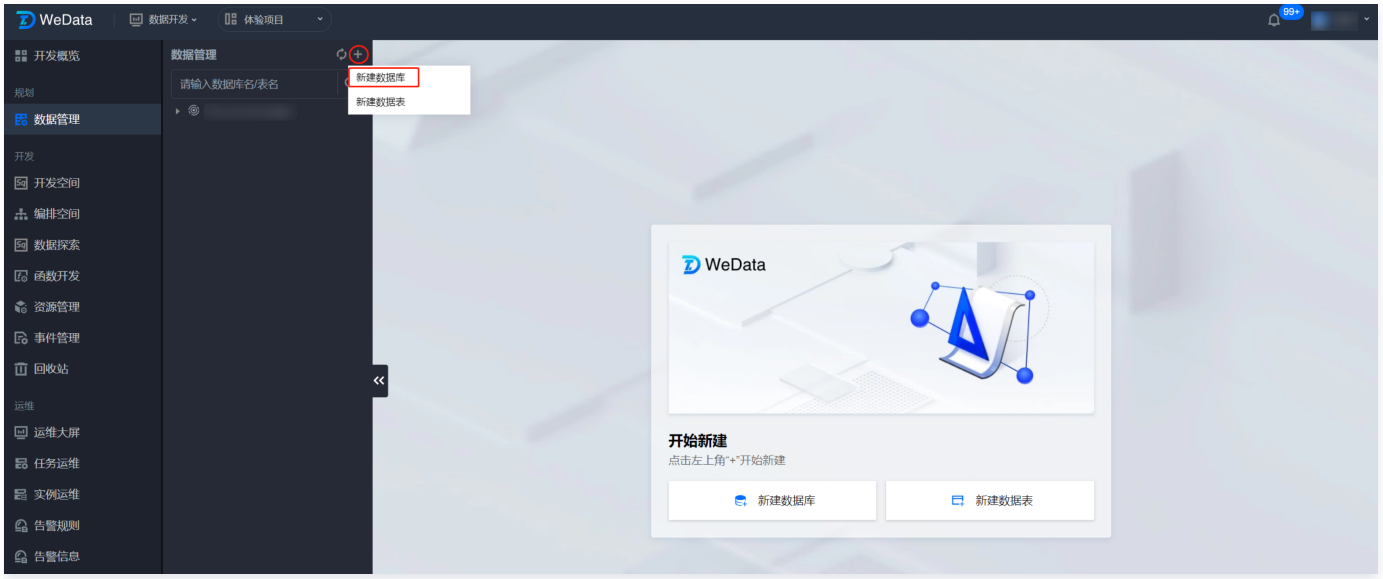
The refresh feature of the Directory Tree is used to reload the data source, database, and data table to update the displayed content in the Directory Tree. This helps users update and synchronize the latest data in the data sources, ensuring the latest data tables are available.



Database management

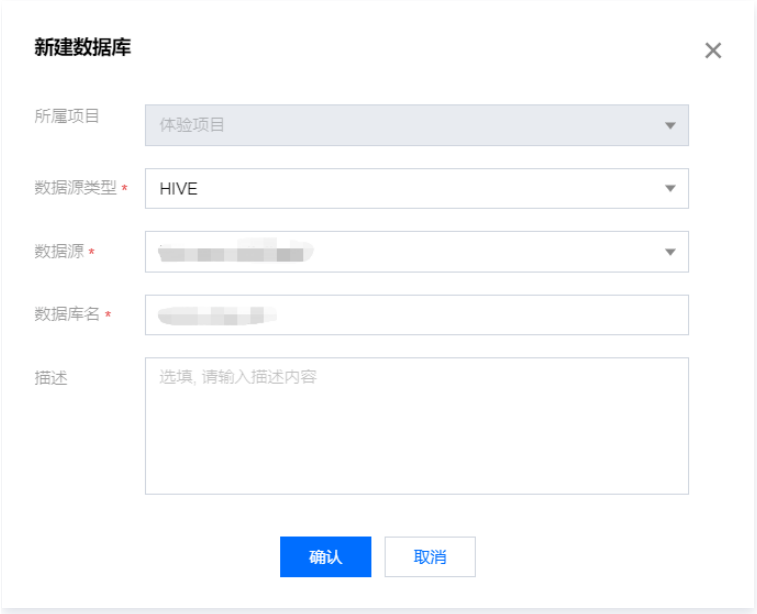
New database

Depending on the bound data source, you can create databases under Hive or DLC data sources.
In the Data Management Catalog, click **Create New Database**, follow the prompts to select data source type, data source, define the database name and description (optional). After configuration, the database can be created in the corresponding data source.



Hive Database

©2013–2025 Tencent Cloud. All rights reserved.



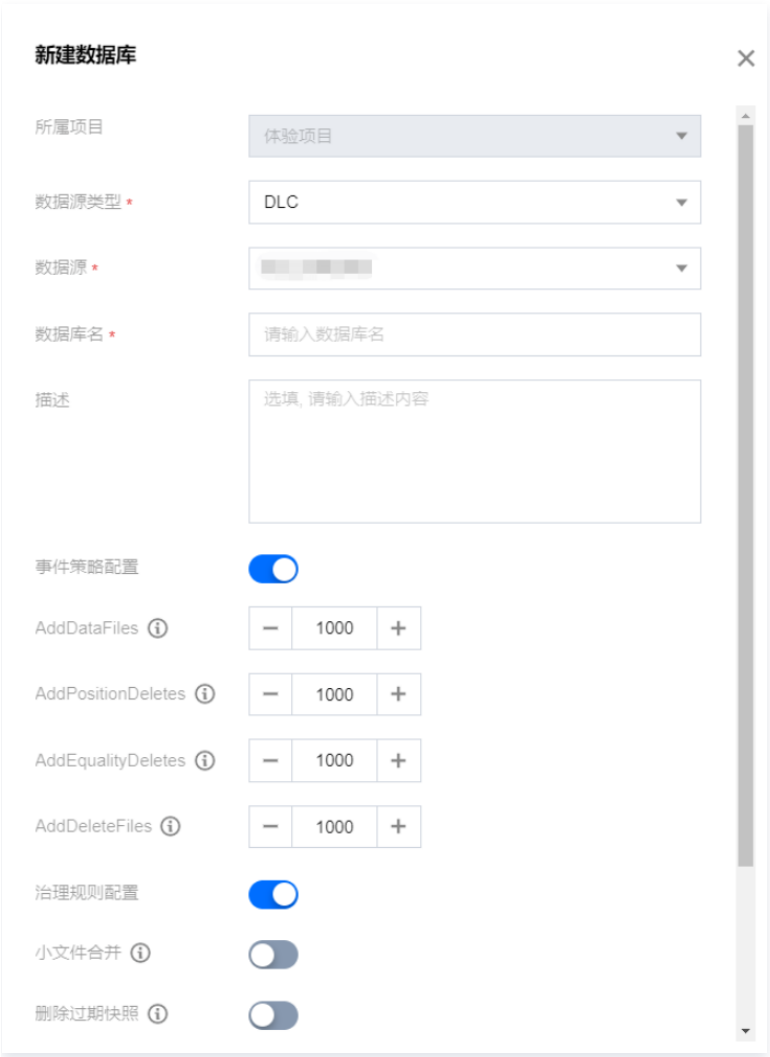
Hive database information:

Information	Description
Data source type	Select Hive type.
Data Source	Select Hive type data source.
Database name	Define Hive database name.
Description	Optional, define description.

DLC Database

When creating a database under a DLC data source, you can configure event strategies and governance rules for the database.

©2013–2025 Tencent Cloud. All rights reserved.



DLC database information:

Information		Description
Basic Information Configuration	Data source type	Select DLC type.
	Data Source	Select DLC type data source.
	Database name	Define DLC Database Name.
	Description	Optional, define description.
Event Strategy Configuration	AddDataFiles	Set the maximum number of added files. Exceeding this value will trigger small file merging.
	AddPositionDeletes	Set the maximum value of added Position Deletes. Exceeding this value will trigger small file merging.
	AddEqualityDeletes	Set the maximum value of added Equality Deletes. Exceeding this value will trigger small file merging.
	AddDeleteFiles	Set the number of added delete files. When the sum of expired snapshot AddDataFiles+AddDeleteFiles exceeds the threshold AddDataFiles+AddDeleteFiles, the snapshot will be deleted from that snapshot.
Governance Rules Configuration	Small File Merge	After enabling, a large number of data files smaller than the threshold will be merged into larger files, reducing the number of files and improving query performance.

©2013–2025 Tencent Cloud. All rights reserved.

	Delete Expired Snapshots	After enabling, expired historical snapshot information will be automatically cleaned, reducing the number of metadata/data files, saving storage space, and improving query speed.
	Delete Orphaned Files	After enabling, invalid data files will be automatically cleaned regularly, saving storage space.
	Metadata Merge	After enabling, metadata manifests files will be automatically merged, reducing the number of manifests files and improving data query efficiency.

Dropping a Database

In the Data Management directory tree, move the cursor to the database you want to delete, click ••• to expand the database operations menu, then click **Delete Database**. Confirm in the pop-up window to delete the corresponding database.

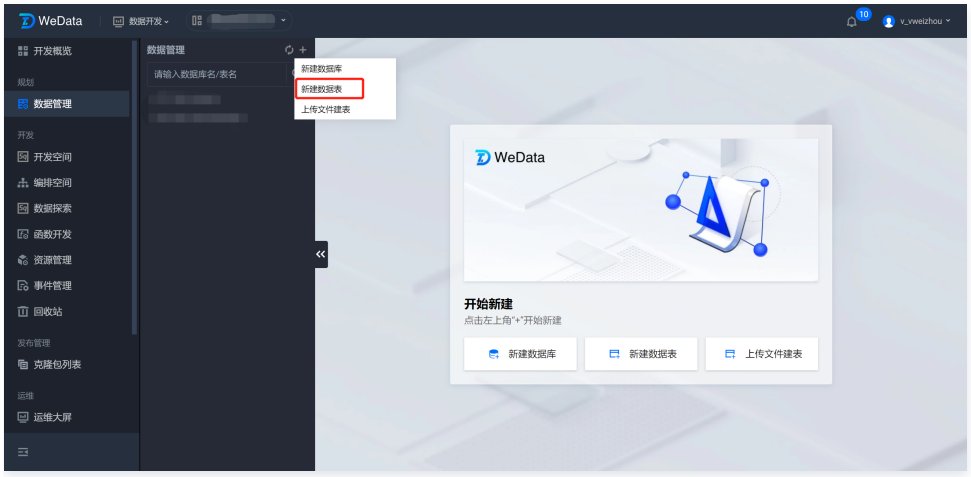


> **Note:**
> A deleted database cannot be recovered, please delete with caution.

Managing a data table

> **Note:**
> Before creating a data table, the database must be created first.

In the Data Management directory, click **Create New Table**. Follow the prompts in the pop-up window to select the data source type, data source, database, and define the table name. After configuration, click **OK** to enter the table**basic attributes** and**fields** design page.

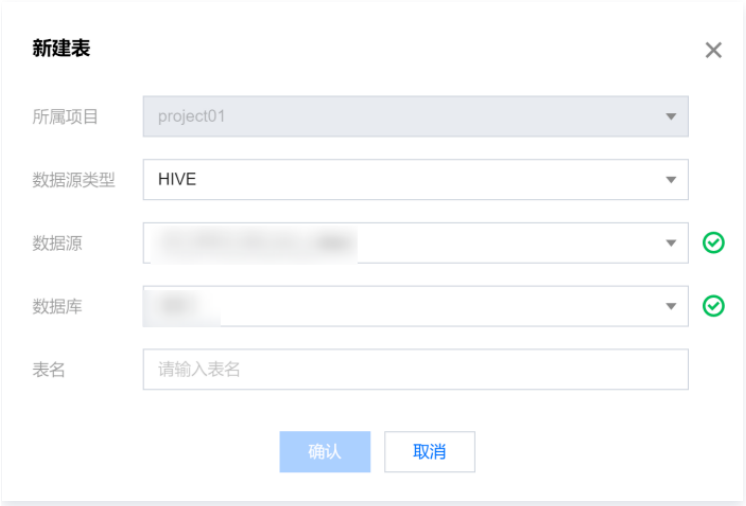


Hive Data Table

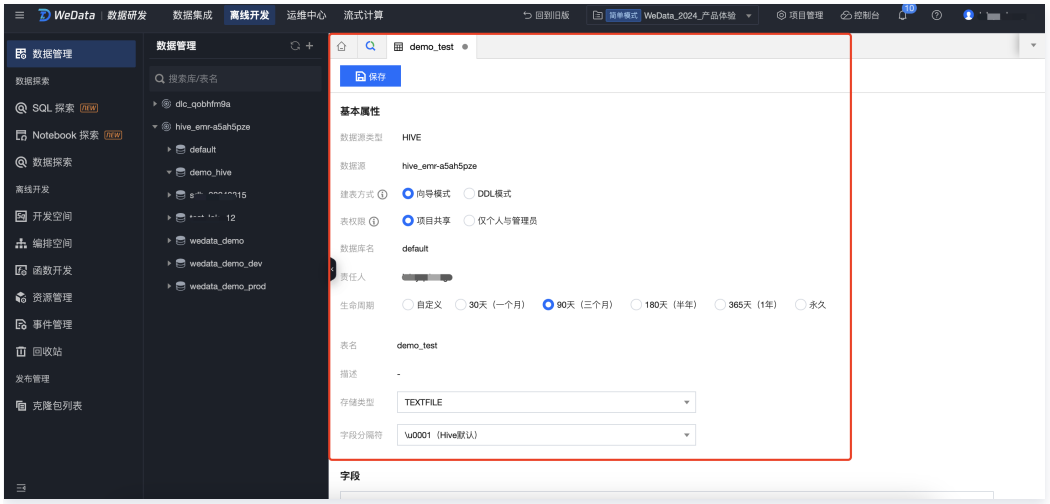
1. When using EMR as the computation engine, you can create Hive tables under the Hive data source in Data Management.

©2013–2025 Tencent Cloud. All rights reserved.

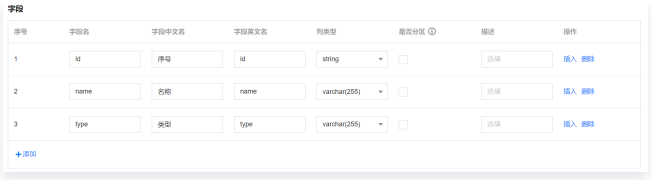
> **Note:**
> The Hive service needs to be started in the EMR cluster. If Ranger is enabled for Hive, ensure the Ranger username and password are correct. Currently, field modification and addition features are not provided.



2. After filling out the basic information in the new table pop-up window, you can enter the table design page to configure the table's basic attributes and field information.

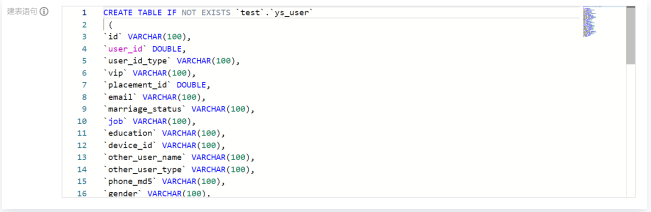


3. Hive Table Configuration:

Information	Description
Table Creation Method	• Wizard Mode Use the traditional manual method to add fields. After inserting fields, define the field name, Chinese name, English name, column type, whether it is partitioned, and description.  • DDL Mode Use SQL create table statements to create data tables. New table creation only supports the CREATE TABLE statement. Table editing only supports ALTER TABLE ADD/REPLACE COLUMNS statement. Example:

©2013–2025 Tencent Cloud. All rights reserved.

```
create table if not exists WeData_demo_db.user_info (
user_id string COMMENT 'User ID',
user_name string COMMENT 'Username',
user_age int COMMENT 'Age',
city string  COMMENT 'City'
) COMMENT 'User Information Table';
```

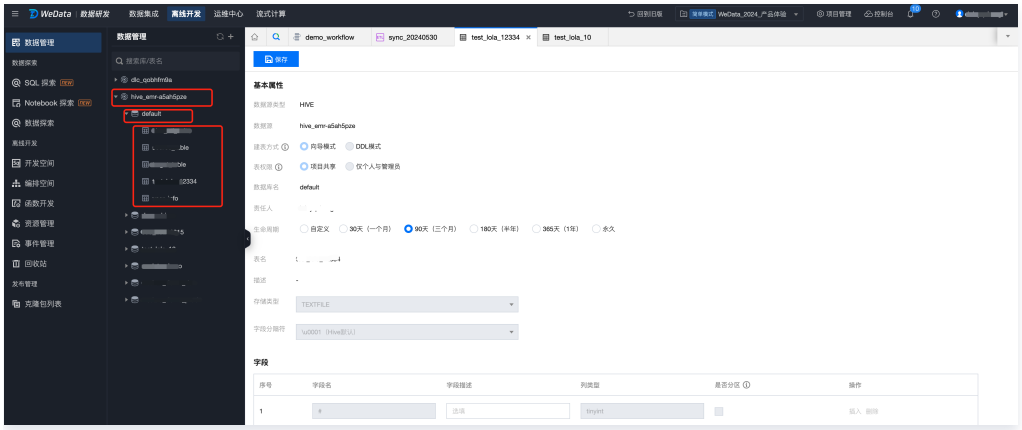


Note:
During the DDL table creation process, ensure that the table name part of the creation statement matches the name entered when creating a new data table.

Table permissions	• Project sharing Assign the data table permissions to the current project. All members within the project will gain permissions to the data table, including editing, inquiry, delete, etc. • For individual and administrator only Assign the data table permissions to the creator and the current project administrator. (Note: The data permissions effective time is expected to be about 30 seconds)
Lifecycle	EMR-Hive tables do not support lifecycle configuration. The current configuration is ineffective. Please note that this config item will be removed in future iterations.
Storage class	Supports four types of storage methods: • TEXTFILE: A text format storage type that stores pure text files with each line representing a record. • PARQUET: A columnar storage format that divides data into rows and columns and stores them by columns on the disk. It is faster than row-based storage in some scenarios and supports column-based compression. • ORC: An optimized columnar storage format used for storing and processing large-scale data. It uses advanced compression algorithms and indexing techniques to improve processing speed and query efficiency. • CSV: A common text format using a comma as the field delimiter, with field values enclosed in quotes.
Field delimiter	Separates each field in the data table for reading and processing in programs or systems. Supports five types of field delimiters: \u0001 (Hive default), | (vertical bar), (space), ; (semicolon), , (comma), \t (tab)
Field configuration	A field includes the field name, field description, column type, whether it is a partition, and other configuration information. Partition Field Description: Not all fields can be partition fields; at least one must be a non-partition field. Partition fields do not support array, map, or decimal types.

4. After configuring the **basic attributes** and **fields** of the data table, click **Save** in the top left corner to complete the creation of the data table. You can view the created data table in the data management directory on the left.

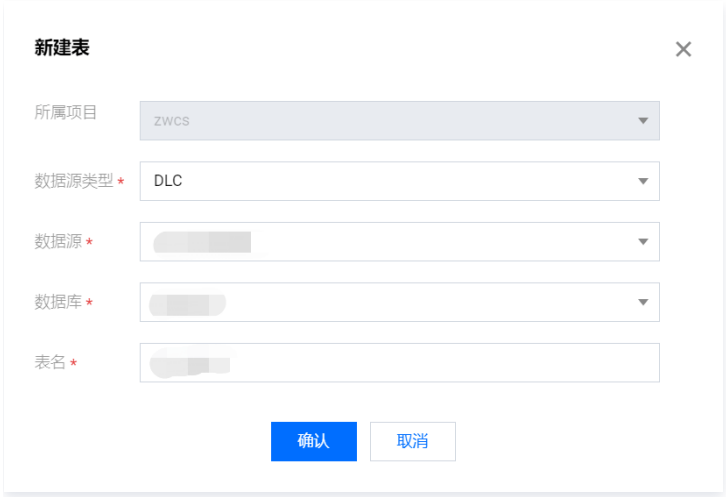
©2013-2025 Tencent Cloud. All rights reserved.



DLC Data Table

1. When using DLC as the computational storage engine, you can create DLC data tables under the DLC data source in Data Management.

> **Note:**
> Currently, DLC table creation only supports visual table creation. DDL table creation is not yet supported; please create using SQL statements directly in Data Development.



2. After completing the basic information in the new table popup, you can enter the Data Table Design Page to configure the data table format, field information, and parameter attributes.

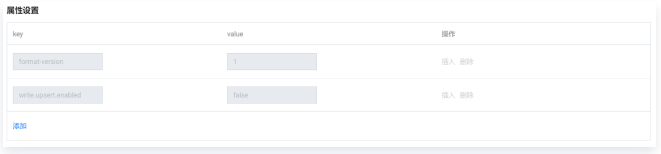
©2013–2025 Tencent Cloud. All rights reserved.



3. DLC Table Creation Configuration:

Information		Description
Data Table Format	Select the table creation type	You can choose to create an internal table or an external table.
	Table Source	When creating an internal table, specify to create an empty table or COS COS.
	Storage Path	For COS and External Table, fill in the full path of the location.
	Data Format	Data formats include: CSV, JSON, PARQUET, ORC, AVRO.
	Data Table Version	Select V1 or V2 data table version.
	upsert	When selecting the V2 data table version, you can choose whether to use upsert for writing.
Basic Attributes	Chinese name	Self-Definition Table Chinese Name.
	Description	Self-Definition Description Information.
Field Information	Field name	Design Table Field Names.
	Field Type	Supports DLC Data Table Field Types.
	Description	Self-Definition Field Description Information.

©2013-2025 Tencent Cloud. All rights reserved.

	Whether to use partition	Design partition, including partition field, conversion strategy, and strategy parameters.
	Event Strategy Configuration	AddDataFiles: Set the maximum number of added files. Exceeding this value will trigger small file merging.
		AddPositionDeletes: Set the maximum value of added Position Deletes. Exceeding this value will trigger small file merging.
		AddEqualityDeletes: Set the maximum value of added Equality Deletes. Exceeding this value will trigger small file merging.
		AddDeleteFiles: Set the number of added delete files. When the sum of expired snapshot AddDataFiles + AddDeleteFiles exceeds the threshold AddDataFiles + AddDeleteFiles, the snapshot will be deleted from that snapshot.
	Governance Rules Configuration	Support enabling data table governance rules. The governance rule configuration item can choose to inherit the governance rules of the database selected when the current data table is created, or define governance rules for the data table self definition. Including the following governance rules: • Small File Merge: After enabling, a large number of data files smaller than the threshold will be merged into larger files, reducing the number of files and improving query performance. • Delete Expired Snapshots: After enabling, expired historical snapshot information will be automatically cleaned, reducing the number of metadata/data files, saving storage space, and improving query speed. • Delete Orphaned Files: After enabling, invalid data files will be automatically cleaned regularly, saving storage space. • Metadata Consolidation: After enabling, metadata manifests files will be automatically merged, reducing the number of manifests files and improving data query efficiency.
Attribute settings	Parameter configuration	Support self definition data table parameter configuration, such as format-version, write.upsert.enabled. 

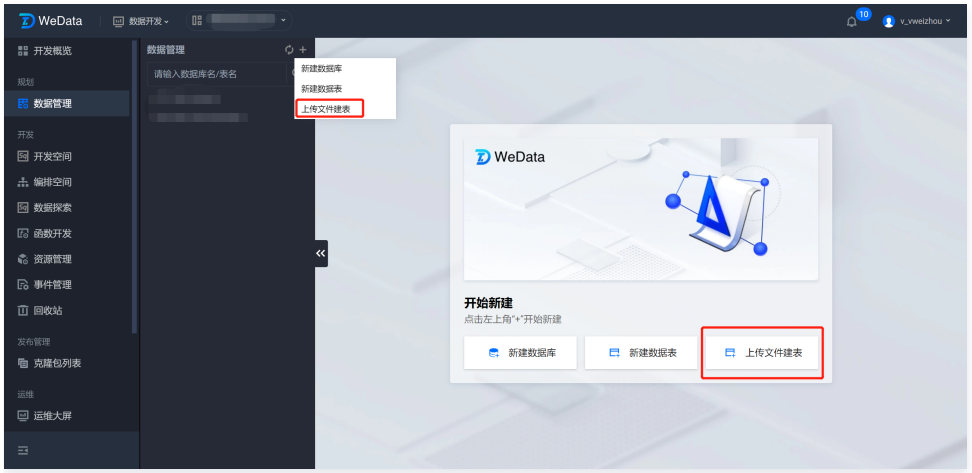
Upload Data Table

1. In the data management directory or data management main interface, click **Upload Files to Create Tables**. Currently, only Hive-type data tables are supported for upload and creation.

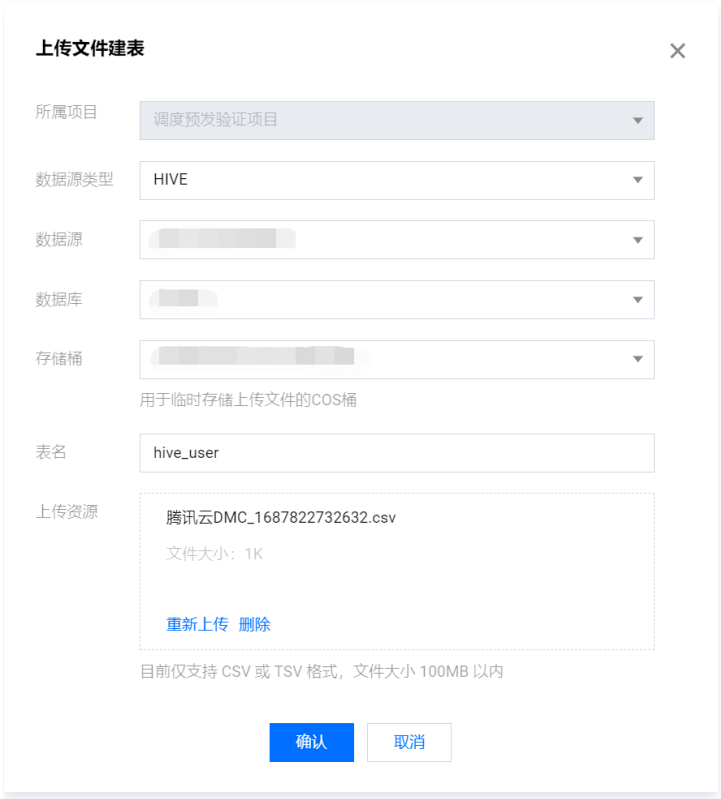
Upload Example:
Domestic Regional Population Statistics

> **Note:**
> 1. Currently, CSV and TSV files are supported for upload. The maximum file size is 100 MB.
> 2. You need to bind an EMR cluster in the WeData project, which has the corresponding Hive service.
> 3. If Ranger is configured in Project Management, the username and password for Ranger need to be correct.
> 4. The EMR_QCSRole role set in the COS bucket needs to have access permissions to COS, otherwise, data import will report an issue with the COS path.

©2013-2025 Tencent Cloud. All rights reserved.



2. In the popup window, select the data source type, data source, database, bucket, custom data table name, and upload table creation resources according to the prompts.



3. Upload file configuration:

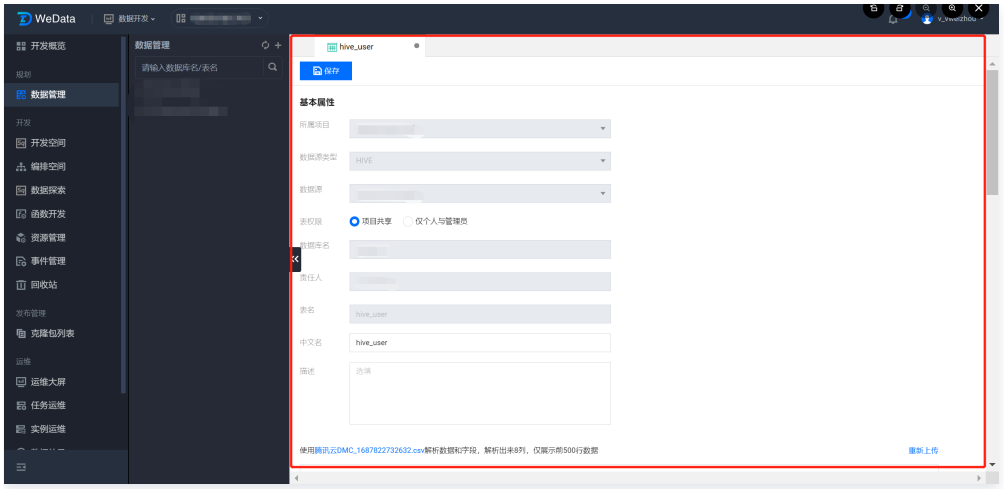
Information	Description
Data source type	Supports Hive type data sources.
Data Source	Select the corresponding data source type for WeData data source.
Database	Displays the Hive database bound to the current project, linked according to the data source type. Supports searching by library name.
Bucket	COS bucket for temporarily storing uploaded files.
Table name	Automatically fills in the uploaded filename by default, removing the suffix. Custom names are allowed.
Upload Resources	Click Upload or drag and drop to upload, providing an upload progress bar. Upload format: CSV or TSV format.

©2013–2025 Tencent Cloud. All rights reserved.

4. Here is an example of a CSV file, the data format is as follows:

	A	B	C	D	E	F	G
1	id	order_id	user_id	gender	address	vip	telephone
2	1	DD202101	U0001	男	上海	普通用户	13800000001
3	2	DD202101	U0002	男	北京	普通用户	13900000002
4	3	DD202101	U0003	女	广东	普通用户	13700000003
5	4	DD202101	U0004	女	广东	金牌VIP	15800000004
6	5	DD202101	U0005	男	内蒙古	普通用户	13600000005
7	6	DD202101	U0006	女	内蒙古	银牌VIP	13500000006
8	7	DD202101	U0007	男	浙江	普通用户	15900000007
9	8	DD202101	U0008	男	四川	普通用户	15600000008
10	9	DD202101	U0009	女	上海	铜牌VIP	13800000009
11	10	DD202101	U0010	男	北京	普通用户	13900000010

5. After configuring the popup message, click **OK** to enter the Table Creation Page.



6. On the table creation page, you can set data table permissions, Chinese table name, and table description information. The parsed uploaded file will provide fields, data preview, and support configuration of file format, column separator, column quotation marks, first row field confirmation, file encoding method, and field attributes.

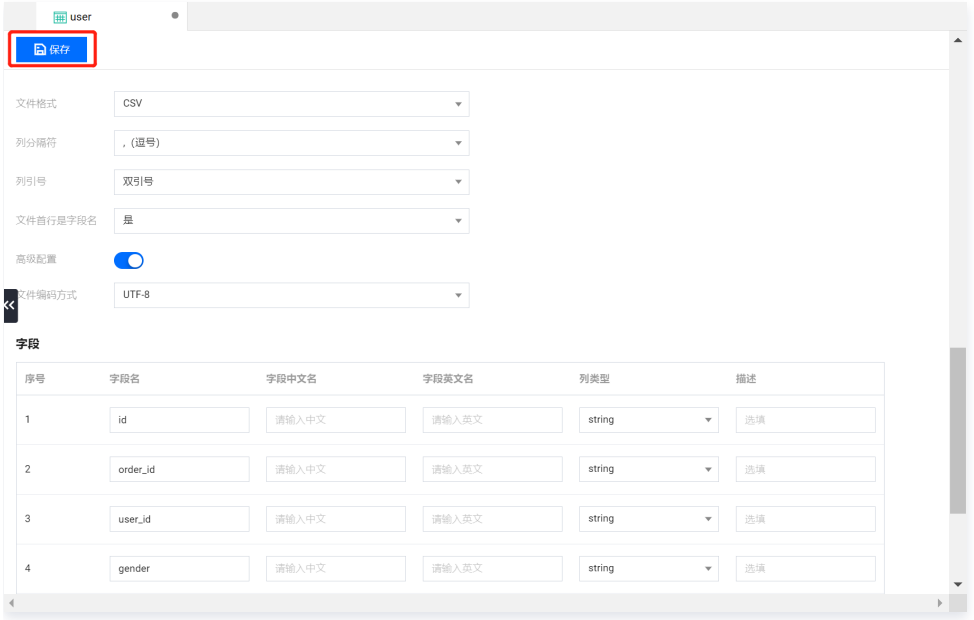
Information		Description
Basic Attributes	Table permissions	Select the permission attribution after the current data table is created, shared within the project or available only to personal and administrators.
	Chinese name	The default automatically carries the file name without the suffix. You can customize it.
	Description	Custom Data Table Description Information.
File Attributes	Data preview	Only the first 500 rows of data are displayed after file parsing, click **Re-upload** to open the file upload window for re-uploading the table creation file.



©2013–2025 Tencent Cloud. All rights reserved.

	File format	Dropdown selection, supports **CSV**, **TSV**.
	Column delimiter	Users can customize input, entering a single character or a Unicode escape sequence like \u0001. • CSV default uses: , (Comma) • TSV default uses: \t (Tab Character)
	Column Quotes	Default **Double quotes**, users can switch to **Single quotes**.
	The first row of the file is the column name	Default **No**, can switch to **Yes**.
	File Encoding	Default **UTF-8**, users can choose **UTF-8**, **GBK**, **ISO-8859-1**.
Field Attribute	Field name	Parse the field name according to the **first line of the file is the field name** attribute. When the first line of data in the file is not the field name, use column_1, column_2, column_3,.....column_x to fill the field names in order. Users can modify the field name with self-definition.
	Chinese Field Name	Self-Definition Chinese Field Name.
	English Field Name	Self-Definition English Field Name.
	Column Type	Select the corresponding field type supported by the data source according to the data source type.
	Description	Self-Definition Field Description Information.

7. After configuring the table creation information on the page, click **Save** in the upper left corner of the page to create the data table.



8. The corresponding data table generation progress can be viewed in the progress pop-up after saving. When the creation steps run successfully, the data table will be generated successfully.

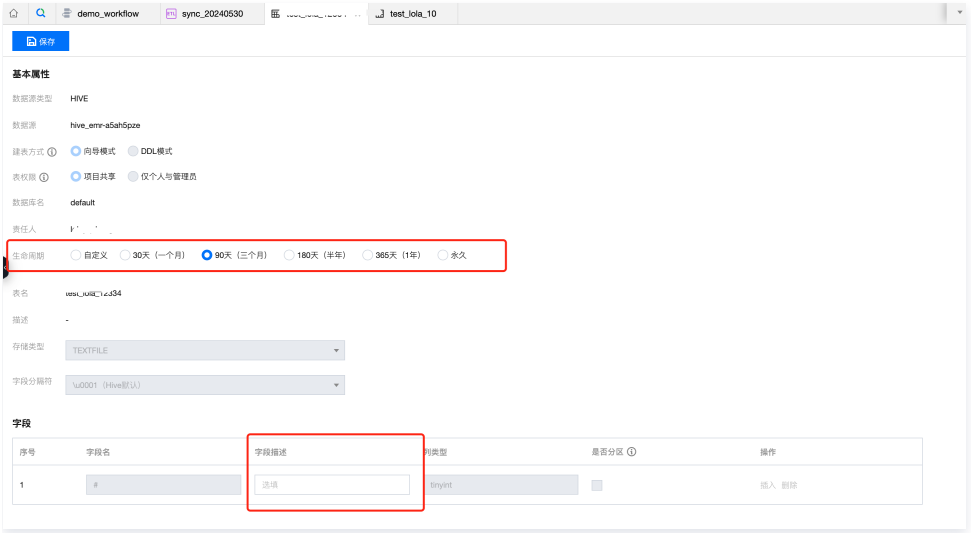
©2013-2025 Tencent Cloud. All rights reserved.



Editing Data Table

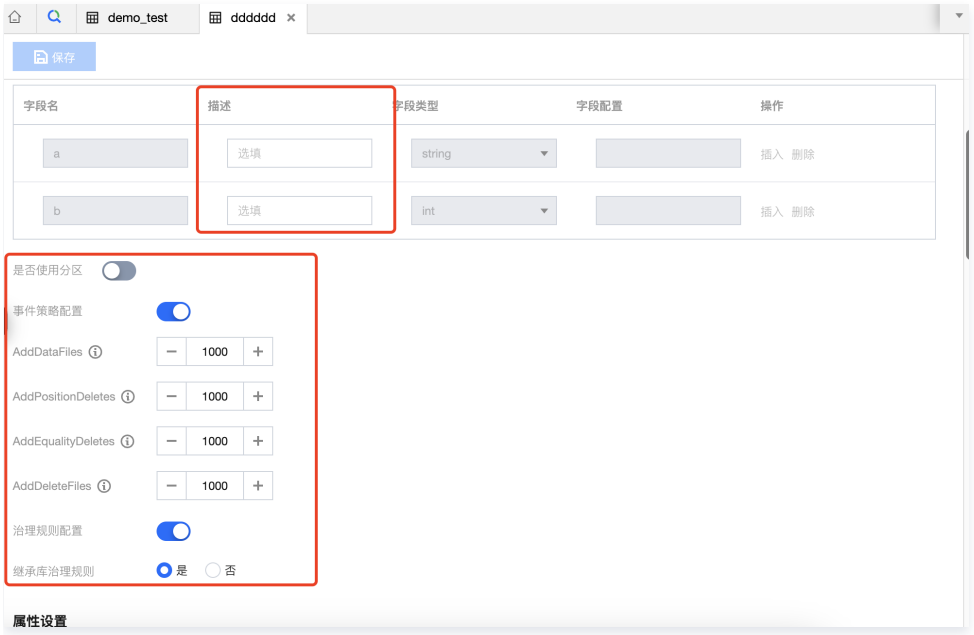
1. In the data management directory tree, move the cursor to the data table that needs to be edited, and double click the left mouse button to open the corresponding data table editing page. Some parameters of the data table can be edited on the page.

Editable content in Hive includes lifecycle and field description.

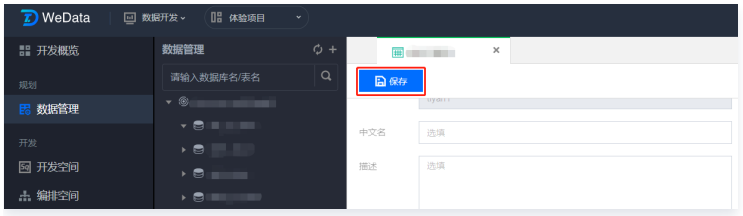


2. DLC editable content includes table field descriptions, event policy configurations, and governance rule configurations.

©2013–2025 Tencent Cloud. All rights reserved.



3. After editing the data table, click Save to complete the data table editing operation.



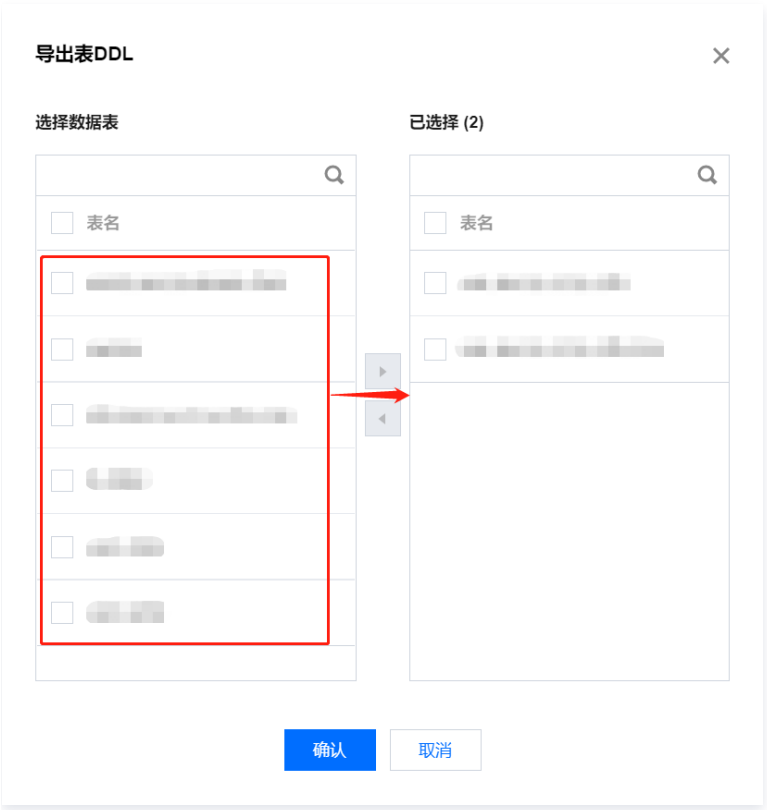
Export Table DDL

1. In the data management directory tree, move the cursor to the database where the data table to export the DDL is located, click ••• to expand the database operation menu, and then click **Export Table DDL**. In the pop-up window's left panel, select the data table in the current database that needs its DDL exported, add it to the right panel, and click Confirm to export the corresponding data table's DDL file.

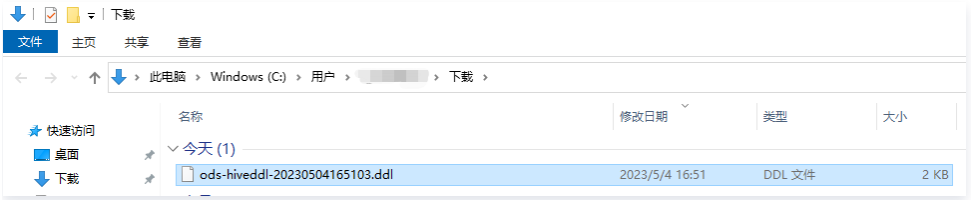


2. Select the data table that needs its DDL exported.

©2013–2025 Tencent Cloud. All rights reserved.



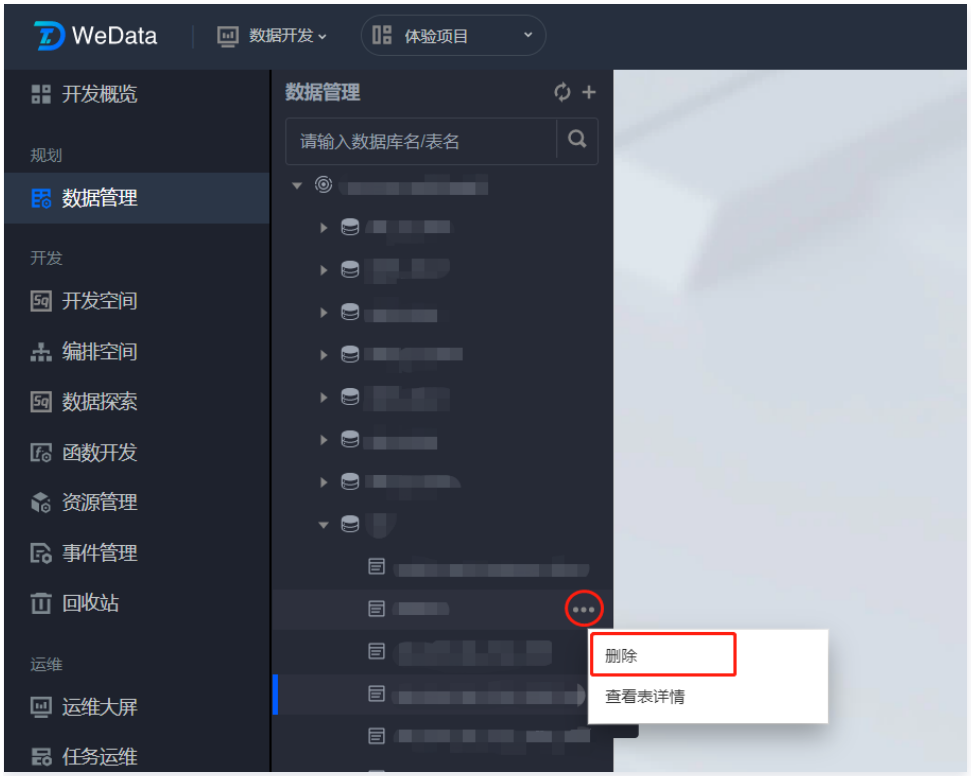
3. After exporting, you can find it in the system's **Download** directory.



Dropping a Table

1. In the data management directory tree, move the cursor to the data table that needs to be deleted, click ••• to expand the data table operation menu, and then click **Delete**. Confirm in the pop-up window to delete the corresponding data table.

©2013–2025 Tencent Cloud. All rights reserved.



Viewing table details

In the data management directory tree, move the cursor to the data table that needs details checked, click ••• to expand the data table operation menu, and then click **View Table Details**. You can check basic information, storage information, field information, data preview, and table DDL.

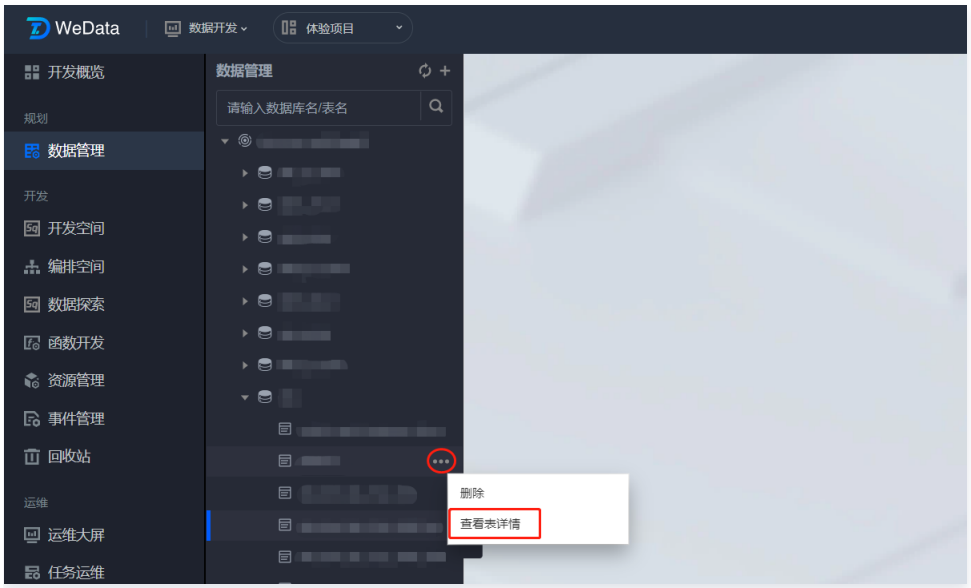


Table Information

©2013–2025 Tencent Cloud. All rights reserved.



Table Details Information:

Information		Description
Basic information	Data Type	The type of storage and compute engine that the data table belongs to.
	Database name	The name of the database that the data table belongs to.
	Table name	The identifier name of the data table.
	Responsible Person	The responsible person for the data table.
	Chinese name	The Chinese name of the data table.
	Description	User-defined description information.
Storage Information	Table Size	The data in the current data table has occupied the physical storage space.
	Lifecycle	The lifecycle of the current data table is used to control its effective usage time, enhancing overall security and saving storage and computing resources in the data governance process.
	Creation Time	Creation date and time of the current data table.

Field Information

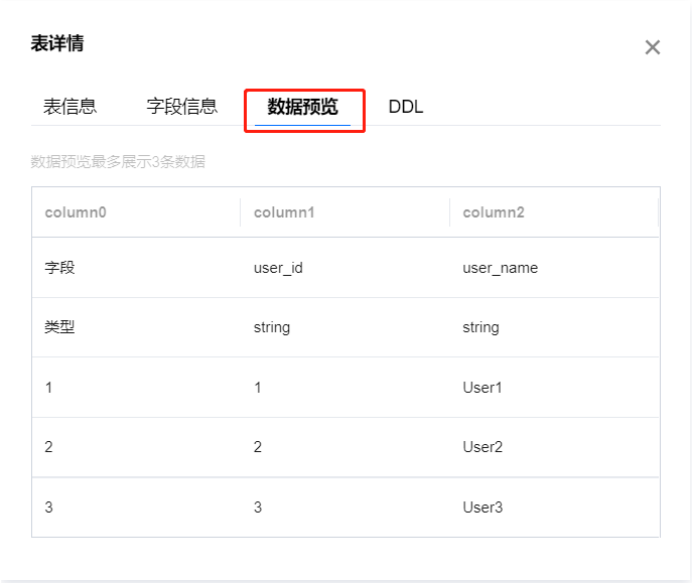
Displays the field metadata information of the current data table, including field sequence number, column name, field Chinese name, field English name, column type, partition status, description, and other information.



©2013-2025 Tencent Cloud. All rights reserved.

Data preview

Capture a portion of the actual data from the current data table as preview data for display, helping users quickly understand the data in the table, and providing a reference for data cleaning and data analysis.



DDL

By viewing the DDL of the data table, you can understand the table name, column names, data types, constraint conditions, and other important information, thus better understanding the table structure and characteristics.



©2013–2025 Tencent Cloud. All rights reserved.

Task Development
Task Development Introduction

Last updated: 2025-05-28 14:47:29

Enter the Task Development Page

1. Log in to the WeData Console.
2. Click **Project List** in the left menu to find the target project that requires the use of task development functionality.
3. After selecting a project, click to enter the **Data Development** module.
4. Click **Orchestration Space** in the left menu.

Task Development Overview

WeData's task development involves orchestrating computing tasks into data workflows for streamlined data processing. It supports flexible data development processes through features such as scheduling policies, event listeners, task parameters, self-dependencies, and function libraries. It can meet users' needs for data processing, processing, and conversion, and provides a visual configuration interface, enabling users to easily build and manage complex data processing workflows.

Streamlined Data Processing

Define data circulation and conversion rules between tasks, and perform data processing, cleaning, and conversion operations.

Workflow Orchestration of Data Workflows

Orchestrate and organize computing tasks as data processing nodes in the form of data workflows to form a complete data processing process.

Scheduling Policies

The scheduling policy is used to determine when to execute tasks. Based on periodic scheduling time and other conditions, workflow execution is automatically triggered to ensure tasks are processed in the scheduled order and time, meeting different business needs.

Listening for Events

Listening for events applies to scenarios where computing tasks depend on a specific event to trigger execution. It consists of a trigger program, trigger event, and listening task. First, define the trigger event based on business requirements under the project. Then, write a trigger program to send the event. The task runs after listening to the event reaching.

Task Parameters and Parameter Passing

Support using variable parameters in data workflow design and computing task configuration, and enable parameter transmission between tasks. Can set different input parameters for each computing task and transmit the task's output parameters to the next task, achieving data sharing and interaction between tasks.

Self-Dependent

Support self-dependency during computing task operations, meaning tasks can depend on the execution status of their previous period during scheduling execution.

Function Library

Provides library functionality, including common functions and algorithms for Hive SQL, Spark SQL, and DLC, such as mathematical functions, data transformation functions, and aggregate functions. It also supports UDF (User-Defined Function) custom functions to help users perform data processing and computational operations, providing users with more flexible and diverse data processing capabilities.

Data Development Collaboration

In WeData data development, create, write, and debug development scripts that collaborate with data workflows. Scripts configured in the development space for ad-hoc development can be directly integrated into the orchestration of data workflows as

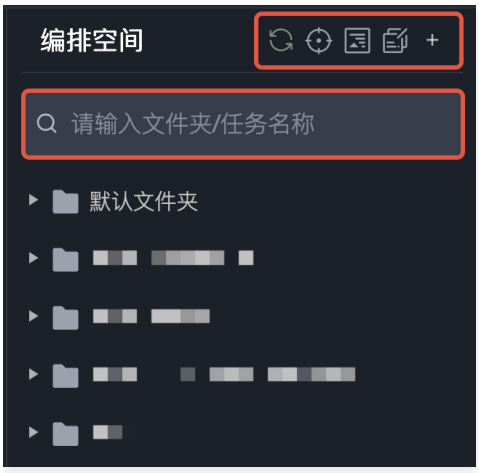
©2013-2025 Tencent Cloud. All rights reserved.

one of the task nodes, achieving code reuse and optimizing the overall process.

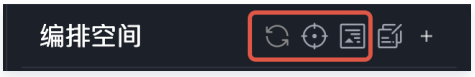
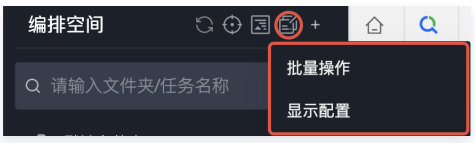
Workflow Introduction

The orchestration space provides configuration features for data workflows, supporting users to organize the development of different types of task code based on workflows and submit them to the scheduling system for periodic execution. A project can contain multiple workflows. WeData supports placing different workflows in the same folder for convenient and efficient management. A workflow is a collection of objects of various types, including data integration, computing tasks (Hive SQL, JDBC SQL, MapReduce, PySpark, Python, Shell, Spark, Spark SQL, DLC SQL, DLC Spark, Impala, TCHouse-P, Trino), and common tasks.

Workflow Directory



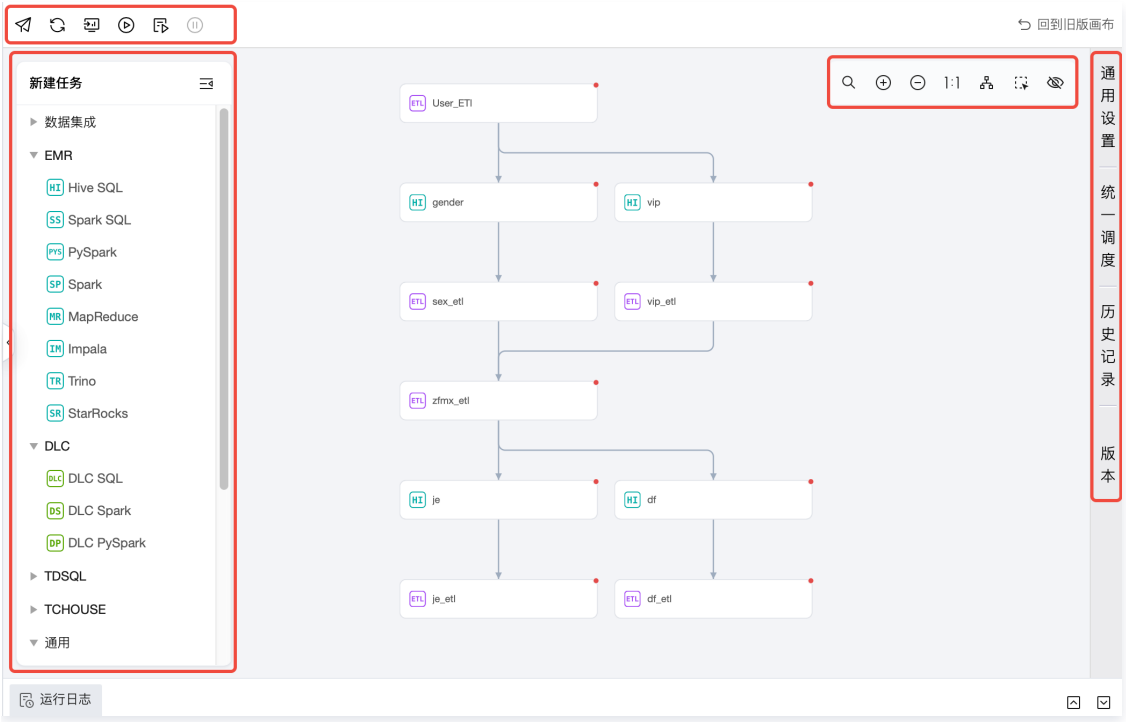
Directory feature

Feature	Description
Searching	Supports searching for search folders, workflows, and task names. 
Refreshing	• Refresh: Refresh the directory tree to obtain the latest state of the orchestrated directory. • Locate tree node: One-click locate to the current tree node. • Collapse the current tree node: One-click collapse all expanded directories. 
Batch	• Batch operation: Supports performing batch operations on all computing tasks in the orchestration directory. Operation items include submitting tasks (batch supported), deleting tasks, modifying resource groups, modifying accountable persons, modifying data sources, modifying task parameters, modifying scheduling cycles, modifying advanced scheduling settings, and modifying scheduling parameters. Batch operation records are viewable. • Display configuration: Supports the display and hiding of AI assistant, cross-workflow categories and nodes, highlighting code in brackets by clicking brackets, code snippets, etc. 
Creating	Supports creating folders and data workflows.

©2013-2025 Tencent Cloud. All rights reserved.



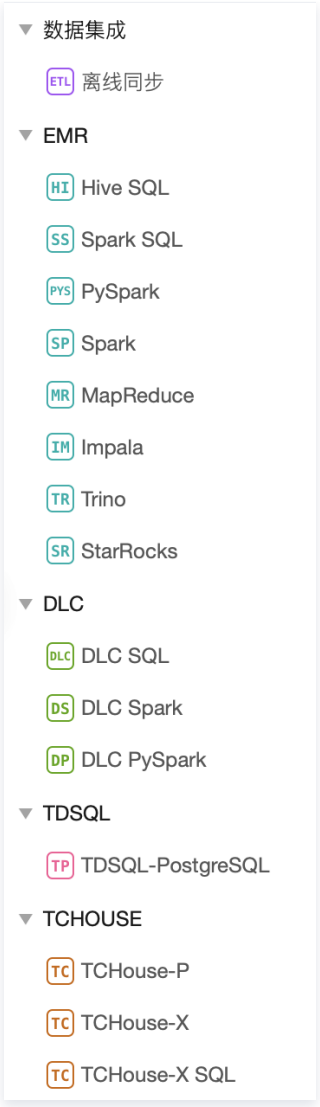
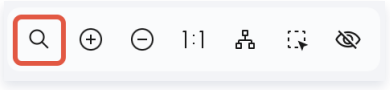
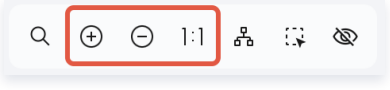
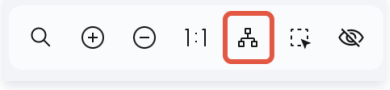
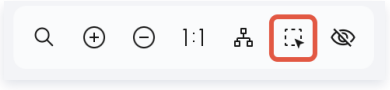
Workflow Canvas



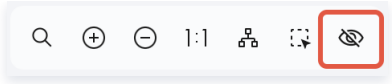
Canvas functionality

Feature	Description
Submit	Click the icon to submit the current workflow to the scheduling system (including node content, configuration properties, and dependency relationships), and generate a new version.
Refreshing	Click the icon to refresh the content on the current workflow canvas.
Go to Operation and Maintenance	Click the icon to go to Operation and Maintenance – Workflow List Page.
workflow testing	Click the icon to test the current workflow. During the test, click the icon to stop the test.
Advanced execution	Click the icon to add, modify, or delete project parameters, workflow parameters, etc. 
Task Type Directory	In the Task Type Directory, click a type of computing task to add a task node to the workflow canvas.

©2013–2025 Tencent Cloud. All rights reserved.

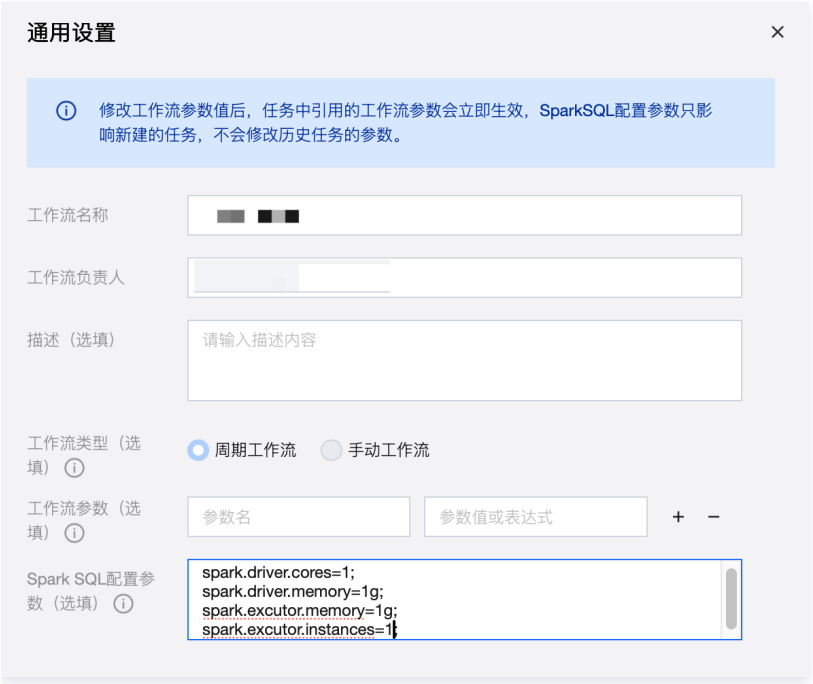
	
Locate	Click the icon, then freely select and locate the corresponding task in the pop-up filter criterion. 
Canvas zoom	Click the icon to zoom the workflow canvas. 
Format	Click the icon to standardize the layout format of tasks in the workflow. 
box select	Click the icon, the mouse becomes box selection mode, allowing you to simultaneously box select multiple tasks and execute batch operations. 
hide	Click the icon to hide cross-workflow nodes.

©2013–2025 Tencent Cloud. All rights reserved.



General Settings

Click the **General Settings** in the right panel to edit the current workflow's name, responsible person, add description information, workflow variables, and Spark SQL configuration parameters (optional). The Spark SQL configuration only takes effect for Spark SQL tasks within workflow tasks.



Feature Description

Feature	Description
Workflow Name	Custom workflow name.
Workflow Owner	Assign the workflow owner. For subsequent workflow submissions, changes, and related operations such as permissions, applications, and approvals, the owner will handle them.
Description (optional)	Custom workflow description.
Workflow type (optional)	Specify when creating a workflow. • Periodic workflow will generate instances periodically according to the scheduling configuration. • Manual workflow requires manual triggering to generate instances and does not execute periodically.
Workflow parameter (optional)	Workflow parameter (optional): The scope is the parameters of tasks within the current workflow. It is set through the universal setting parameter items of the workflow. The setting rule is: variable name = variable value. Multiple values can be separated by ";", for example, a=${yyyyMMdd};b=123;c=456. It supports filling in constants and scheduling date variables. For details, see Workflow-level variable usage process.
Spark SQL configuration parameters (optional)	Configure optimization parameters (threads, memory, CPU cores, etc.), only act on Spark SQL nodes. Separate multiple parameters with semicolons.

Unified Scheduling

Workflow scheduling supports two cycle scheduling configuration methods: general and crontab. For general configuration, see the one-time, minute, hour, day, week, month, and year scheduling configurations in scheduling settings. The crontab configuration method is more flexible and only supports configuration during unified workflow scheduling. Additionally, all task scheduling times

©2013-2025 Tencent Cloud. All rights reserved.

(crontab expressions) under the crontab configuration method must be the same. It does not support configuring cross-workflow dependency tasks or establishing dependencies with general configuration tasks.

> **Notes:**
> The operation of unified scheduling is similar to batch operation, which will change all task cycles under the current workflow to a unified scheduling cycle. It is advisable to use it when the scheduling cycles of tasks within the workflow are consistent.

- general configuration method



Configuration instructions

Feature	Description
Scheduling cycle	Execution cycle unit for task scheduling, supports minutes, hours, days, weeks, months, years, and one-time.
Effective Time	The effective time period of the scheduling time configuration. The system will automatically schedule within this time range according to the time configuration. It will not automatically schedule after the validity period is exceeded.
Execution Time	Users can manually set the execution interval duration of the task and the specific time for task execution to start. If the cycle interval is 10 minutes, the scheduling task will run every 10 minutes from 00:00 to 23:59 every day starting on May 27, 2025.
Calendar scheduling	Users can specifically select dates for executing and not executing scheduling in the scheduling calendar.
scheduling plan	Automatically generated based on the cycle time setting.

©2013-2025 Tencent Cloud. All rights reserved.

Self-Dependent	Configure the task self-dependency attribute for computing tasks in the current workflow. For the dependency feature, please see Task Self-Dependency.
Workflow Self-Dependency	Enabling this option means that computing tasks in the current workflow depend on all computing tasks from the previous cycle of the current workflow. The Workflow Self-Dependency feature takes effect only when tasks in the current workflow are in the same scheduling period and have a daily cycle.

- crontab configuration method

  crontab configuration supports fine-grained settings for year, month, week, day, hour, minute, and second. Once configured, it supports viewing the specific execution time.



Supports using crontab statements to configure scheduling cycles. Click **Configuration** to enter the configuration page.

©2013-2025 Tencent Cloud. All rights reserved.



History Records

Click the **History Records** in the right sidebar to view the historical operation information of the current workflow, including the operator (execution account), operation time, and specific operation content.



Version

Every time a data workflow is edited and submitted for operations, a workflow version will be generated accordingly. Click on the version in the right sidebar to view the historical version information of the current workflow, including the version name (version number), saver (version submitter), retention time (submission time), and change description.

> **Notes:**
> A workflow version will only be generated when a workflow is submitted. Submitting a single task will not generate a workflow version.

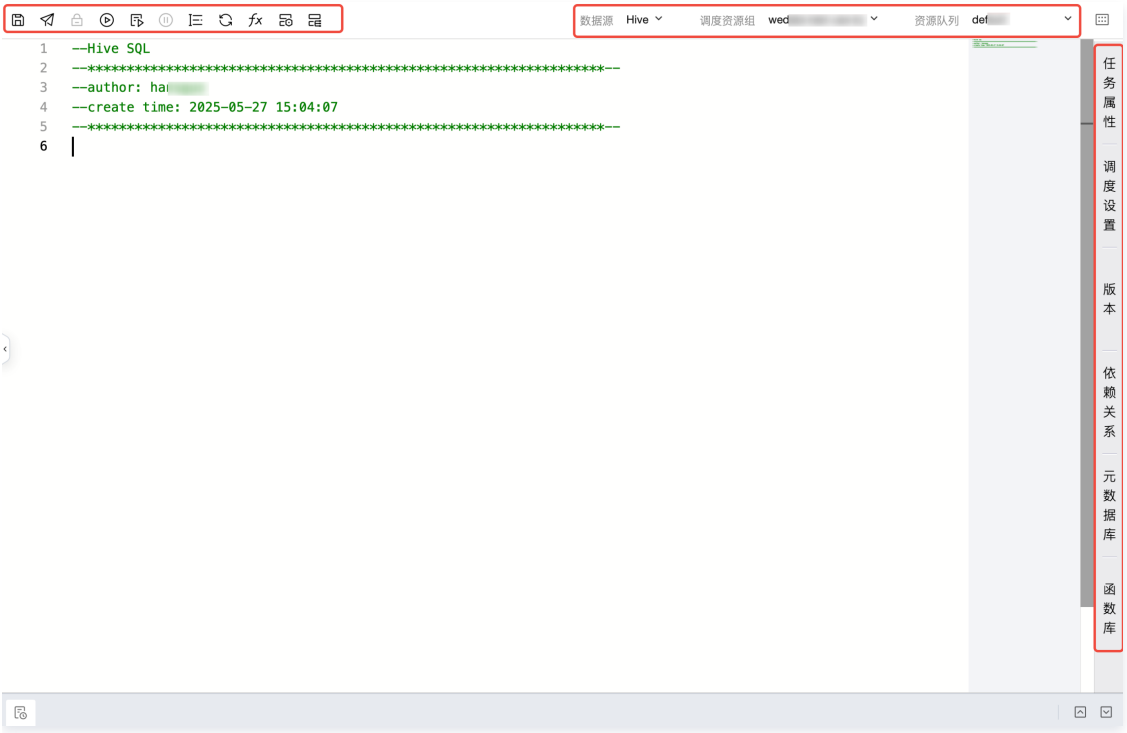
©2013–2025 Tencent Cloud. All rights reserved.



Use the **View** feature under the operation list to see the corresponding version's configuration message. The configuration message can be changed in the workflow's universal setting.



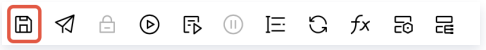
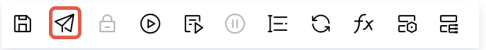
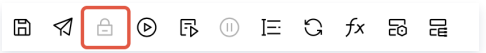
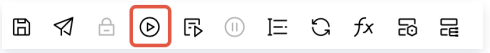
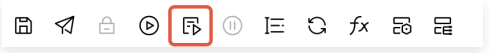
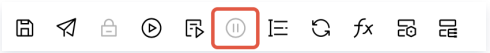
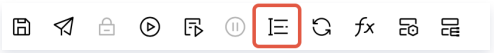
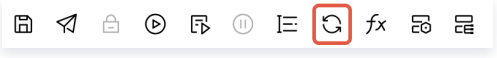
Computing Task Introduction



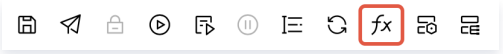
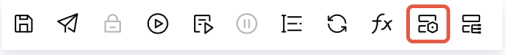
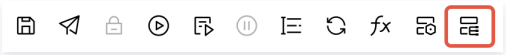
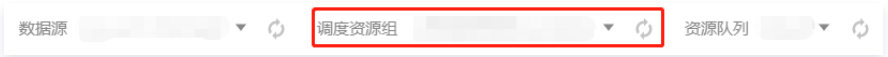
Canvas Functionality

Feature	Description

©2013–2025 Tencent Cloud. All rights reserved.

Save	Click the icon to save the current task node. 
Submit	Click the icon to submit the task node to the scheduling system (node basic content, scheduling configuration attributes) and generate a new version record. Functional limitation: The task's data source and scheduling conditions must be fully set before it can be properly committed. 
Lock/Unlock	Click the icon to lock/release the editing of the current file. If the task is locked by another user, it cannot be edited. 
Running	Click the icon to debug and run the current task node. 
Advanced execution	 Click the icon to run the current task node with variables. The system will automatically pop up the time parameters and custom parameters used in the code. 
Stop running	Click the icon to stop debugging and run the current task node. 
Format	Click the icon to standardize the format of code statements in the task. 
Refreshing	Click the icon to refresh the content of the current task node. 

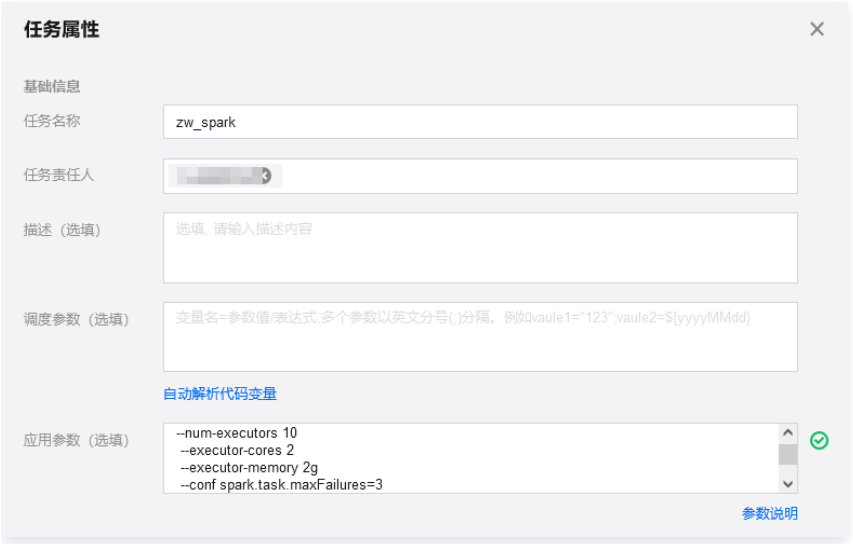
©2013–2025 Tencent Cloud. All rights reserved.

project variable	Click the icon to view the project global variable, which can be used in tasks. 
Task Ops	Click the icon to go to the Task Ops page and automatically filter the current task. 
Instance Ops	Click the icon to go to the Instance Ops page and automatically filter the current task. 
Data Source	Select the data source used for the current computing task. 
execution resource group	Select the execution resource group during task execution. 
resource queue	Select the resource queue used during task execution. 

Task Attributes

The current task's name, owner, and description information can be modified. Task scheduling parameters can be defined. Application parameters can be used. The function of automatic code variable parsing is provided. A parameter description document is provided to assist in the use of scheduling parameter functions.

- How to use scheduling parameters, please see task level variable usage process.
- How to use application parameters, please see application parameter usage process.



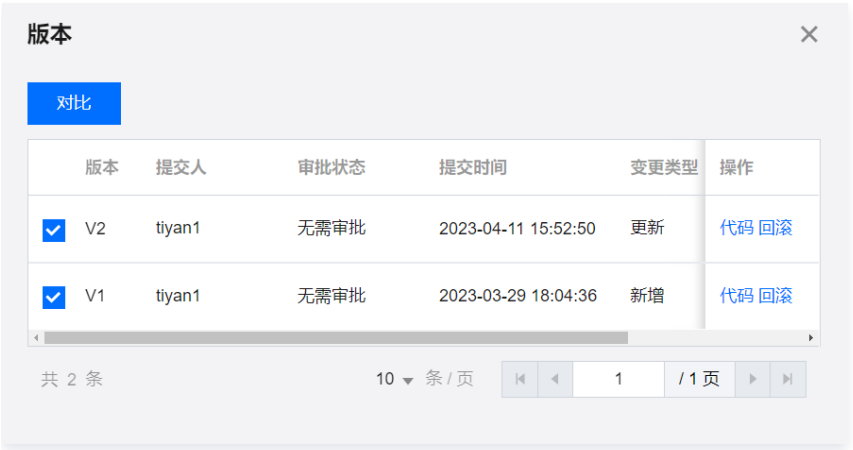
Scheduling Settings

Task scheduling includes scheduling policy, event scheduling, dependency configuration, upstream dependency task configuration, scheduling priority, failure strategy, and other configuration items. For details, see task scheduling.

Version

©2013–2025 Tencent Cloud. All rights reserved.

Show the historical submission records of computing tasks. You can view node historical versions, submitters, submission times, change types, statuses, remarks, and other information in the version panel. You can view the information of a single version and support selecting two versions for comparison.



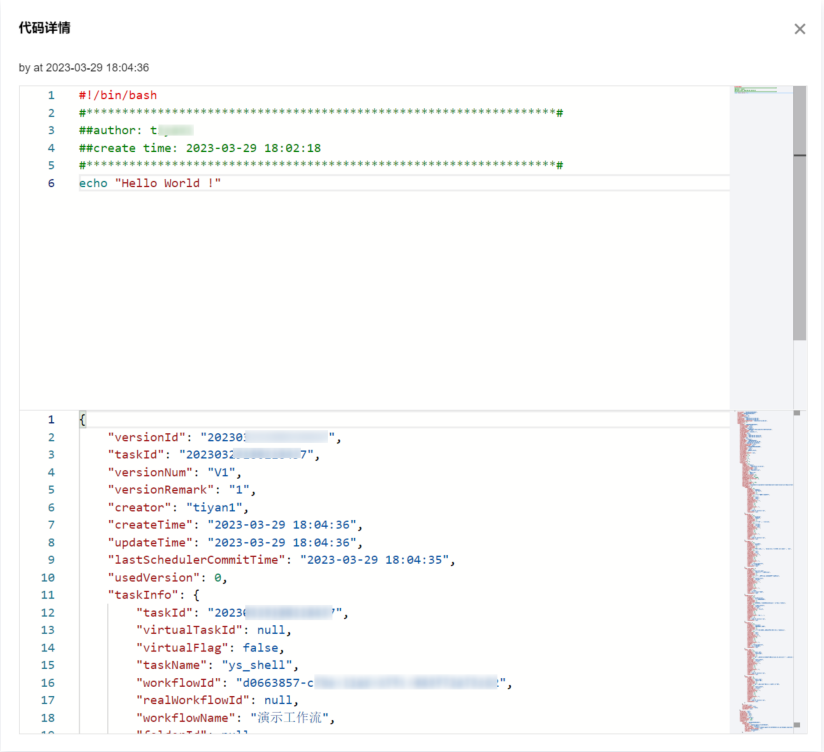
Only submitted task nodes have version information; otherwise, the version information is empty.
Each submission generates a new version and creates a new record in the **version** panel.

- Compare: Provide pairwise comparison between historical versions of computing tasks, displaying key information of both in the form of code and task configuration parameters.



- View the code and configuration parameters of a computing task in any version state.

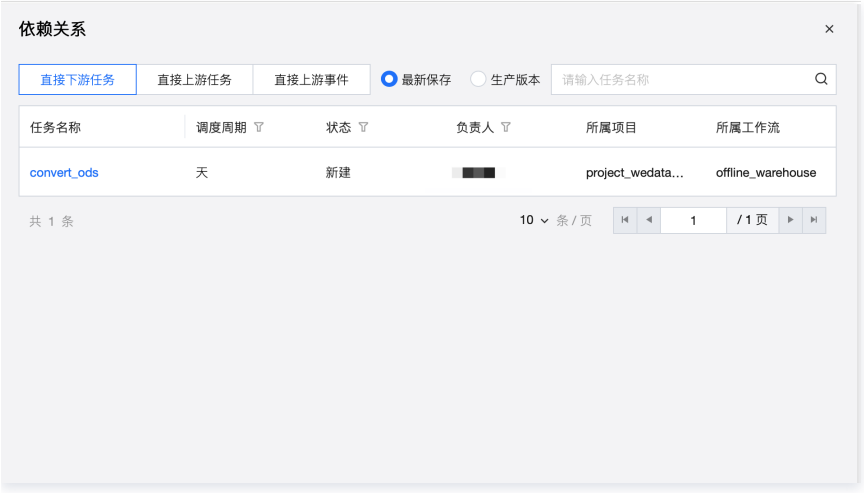
©2013–2025 Tencent Cloud. All rights reserved.



- Rollback: Only roll back the script content and configuration of the task, excluding dependency relationships. Submission after rollback is required for it to take effect. After rollback, changes (including code and task configurations) not submitted will be lost.

Dependency Relationship

Connect task nodes on the canvas or add event dependencies to nodes in scheduling settings, and the dependency relationships will display specific dependencies. Selecting the production version option shows dependencies from submitted runs, while selecting the latest saved option shows dependencies from the last save. Tasks support search and can be filtered by scheduling cycle, status, and owner. Events can be filtered by cycle type and validity time.



MetaDB

Display the metadata information of connected data sources under the current project. Library and table information can be retrieved by searching data sources, databases, and data tables, facilitating quick usage during task development. Quick functions for copying table query SQL, table DDL, and table names are provided.

> **Notes:**
> Note: The ability to currently copy table query SQL and table DDL only supports system data sources.

©2013–2025 Tencent Cloud. All rights reserved.



Function Library

Display functions that can be used in task development. Currently supports DLC SQL, Hive SQL, and Spark SQL functions. Select based on the engine targeted by the development task. The function library includes commonly used system functions, such as analysis functions corr, covar_pop, encryption functions hash, md5, and logical functions decode, nvl, etc. Additionally supports using custom functions. Function packages uploaded through resource management can be created via the function development feature and then displayed in this function library, allowing them to be called in development tasks. For details on creating custom functions, see Function Development.

©2013–2025 Tencent Cloud. All rights reserved.



©2013–2025 Tencent Cloud. All rights reserved.

Cycle Workflow Universal Development Process

Last updated: 2025-06-11 18:07:19

Entering the Orchestration Space Page

1. Log in to the WeData Console.
2. Click **Project List** in the left menu to find the target project requiring development process operation.
3. After selecting a project, click to enter the **Data Development** module.
4. Click **Orchestration Space** in the left menu.

Workflow Operation Steps

Create a Folder

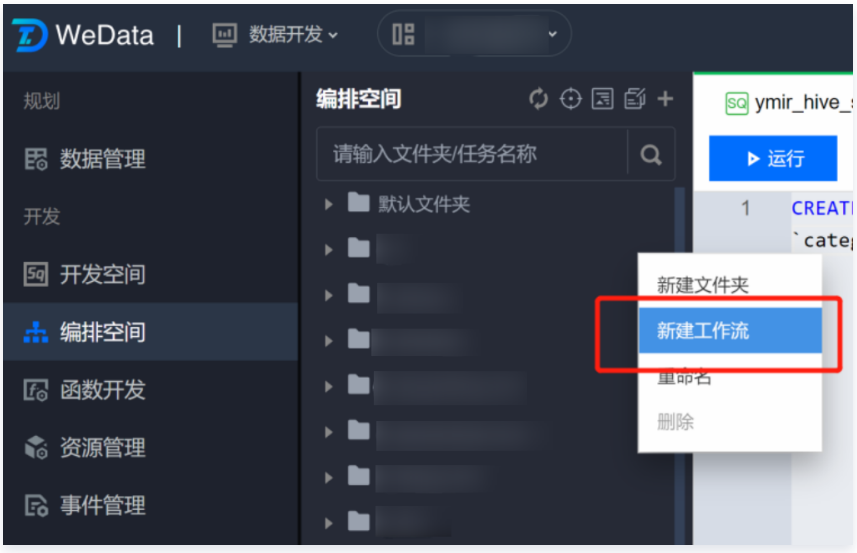
Hover the cursor over the ➕ button in the Orchestration Space Directory Tree, click and choose Create Folder, enter the folder name and select the target folder.



> **Note**
> <Note>WeData supports the creation of multi-level folder directories, and newly created folders can be saved to the root directory or other already created folders.</Note>

Create a New Workflow

1. In the left sidebar directory tree, select the target folder, click the right-click menu and choose **New Workflow**.



2. In the pop-up window for creating a new workflow, enter the workflow name, select the target folder, and click **Confirm**.

©2013-2025 Tencent Cloud. All rights reserved.

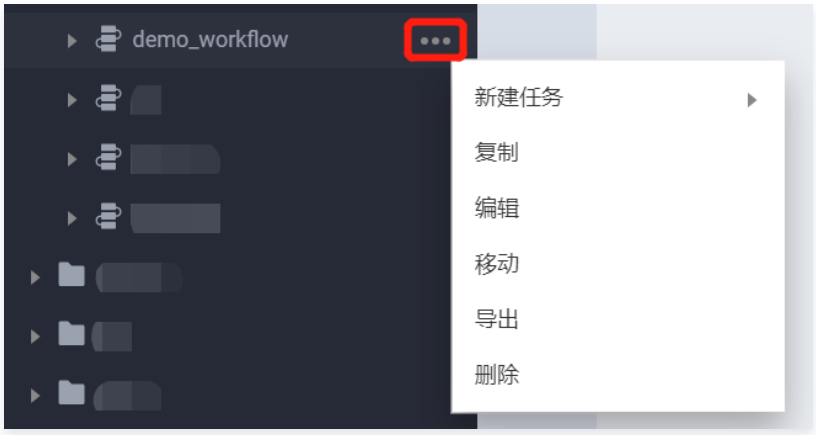


Note
Workflow only supports uppercase and lowercase letters, digits, and underscores, and can include up to 100 characters.

Designing a Workflow

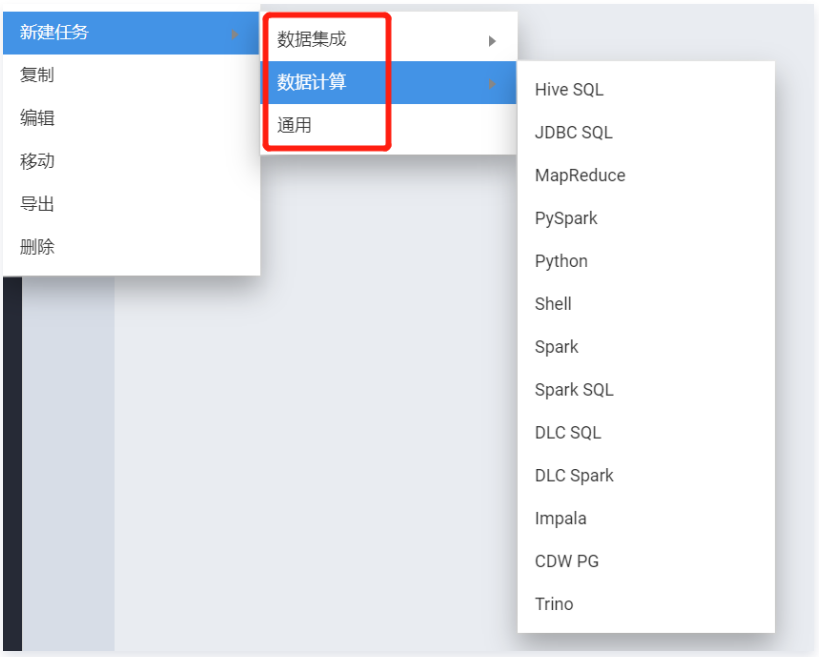
Creating and Editing Tasks

1. Select the newly created workflow in the left sidebar directory tree, click the ··· button at the end, and choose **Create a New Task** in the pop-up menu.

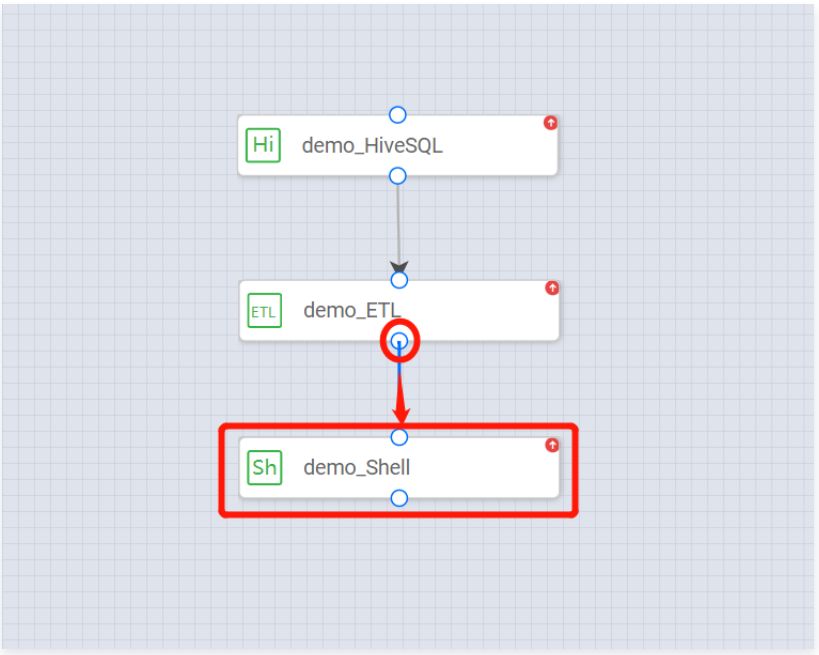


2. Select the task type to be configured in the workflow. Task scheduling supports multiple task types to meet various application development scenarios, including **Data Integration**, **Data Computing**, and **General** task types.

©2013–2025 Tencent Cloud. All rights reserved.



3. After successful creation (or double-clicking the task node name in the directory tree), perform task editing on the opened task Tab page.
4. Follow the Compute Task Operation Steps for task configuration. After configuring multiple compute tasks, orchestrate them on the workflow canvas to form upstream and downstream dependencies between task nodes.



Submit Workflow

1. After completing the workflow editing, click the icon in the toolbar to submit the nodes in the workflow to the scheduling system.
2. In the submission dialog box, view the details of the tasks to be submitted this time and enter the relevant change description.

©2013-2025 Tencent Cloud. All rights reserved.



3. Click **Confirm** to submit the workflow to the operations phase.
4. Nodes that have been submitted will not appear in the pop-up window if their content has not been modified, but changes to their properties will be properly committed. Click the icon to enter the **Operations** > **Task Operations** page.



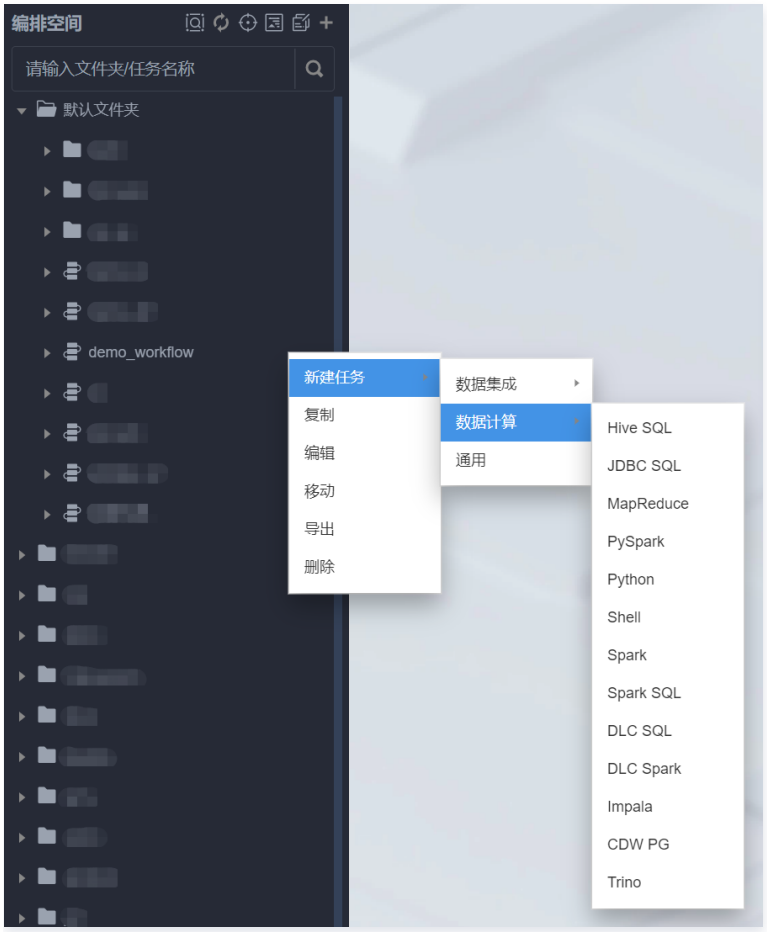
Calculating Task Operation Steps

Task node development supports multiple task types to meet various application development scenarios.

Creating a Task Node

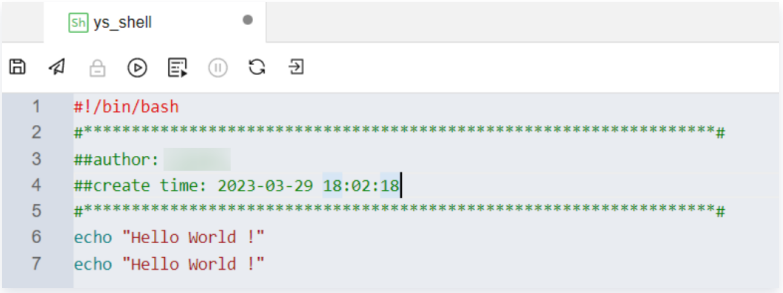
Select a workflow, then click **Create a New Task**, and select the type of computing task you need to create.

©2013–2025 Tencent Cloud. All rights reserved.



Configuring a Task Node

1. After successful creation, perform corresponding code development on the computational task canvas.

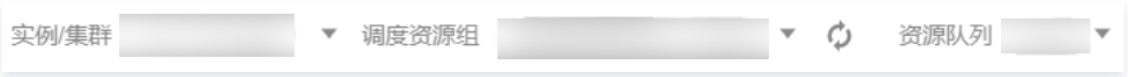


2. Configure task attributes and scheduling policies. Click the **Task Attributes** and **Scheduling Settings** on the right side of the task, and configure the basic information about the task and scheduling information in the panel.

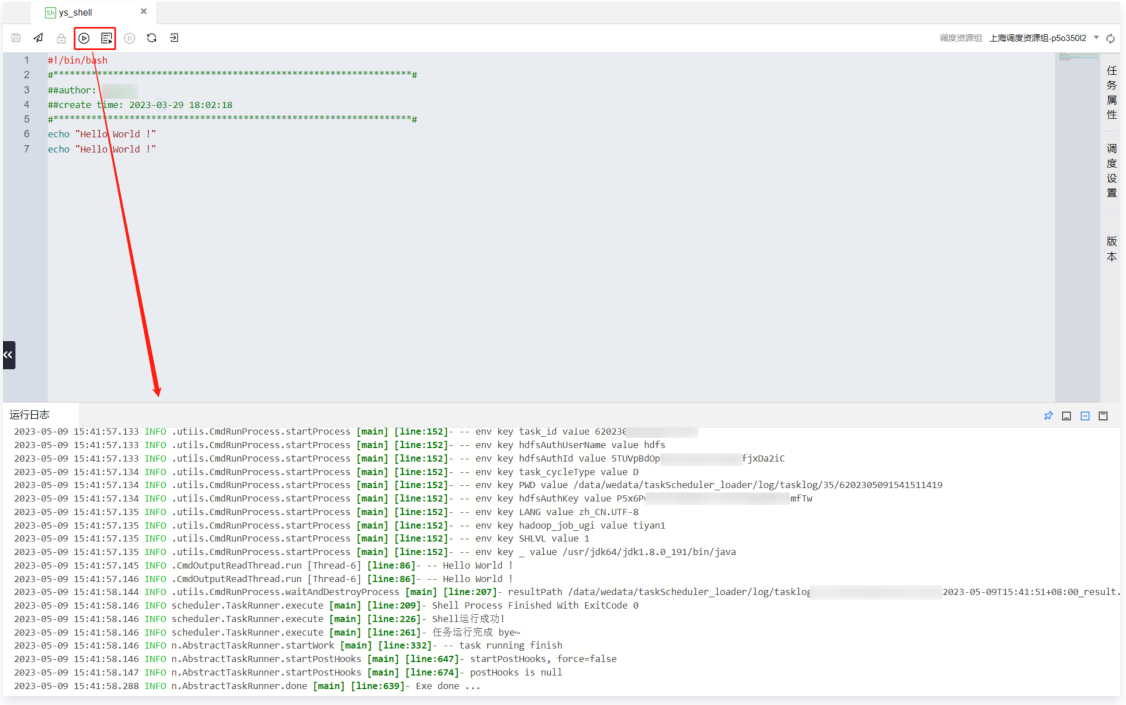
©2013–2025 Tencent Cloud. All rights reserved.



3. Select the data source, resource queue, and scheduling resource group required for executing the computing task.



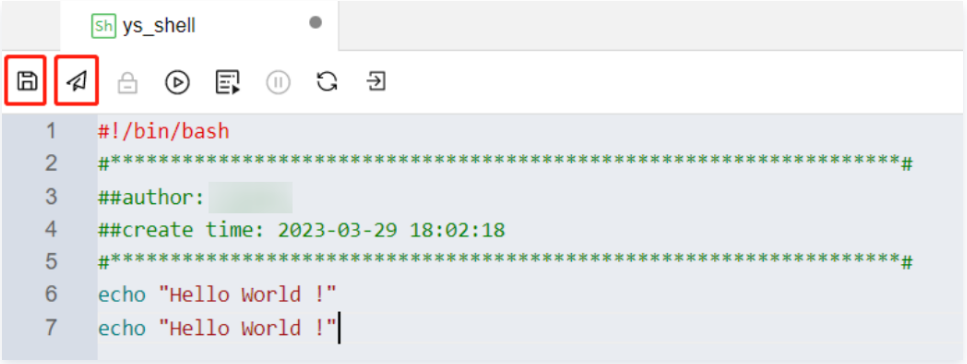
4. Execute the **Run** feature to test whether the computational task execution result meets expectations.



Submitting a Task Node

©2013–2025 Tencent Cloud. All rights reserved.

1. After configuring the computation code, task properties, scheduling policy, and computational resource, and ensuring the test task runs as expected, save and submit the compute task to complete the task node configuration.



2. Sequentially configure each computing task according to the steps. Then, based on **Workflow Operation Steps**, complete the orchestration configuration at the workflow level. You can then submit it for scheduling and running.

©2013–2025 Tencent Cloud. All rights reserved.

Task Node Type
Data Integration

Last updated: 2024-11-25 14:46:44

Offline Synchronization Task supports more than 40 types of write and read data sources. For details, please refer to Offline Synchronization Task under Data Integration.
For detailed information on supported data sources, please refer to Data Source Management.

©2013-2025 Tencent Cloud. All rights reserved.

EMR
Hive SQL

Last updated: 2024-11-25 14:47:06

Note:

You need to start the Hive component services in the EMR cluster.

1. The current user has permissions in the EMR cluster.
2. The corresponding database has been created in Hive, as in the example: wedata_demo_db.

Sample code

```
-- Create a user information table
create table if not exists wedata_demo_db.user_info (
user_id string COMMENT 'User ID',
user_name string COMMENT 'Username',
user_age int COMMENT 'Age',
city string COMMENT 'City'
) COMMENT 'User Information Table';

-- Insert data into the user information table
insert into wedata_demo_db.user_info values ('001', 'Zhang San', 28, 'Beijing');
insert into wedata_demo_db.user_info values ('002', 'Li Si', 35, 'Shanghai');
insert into wedata_demo_db.user_info values ('003', 'Wang Wu', 22, 'Shenzhen');
insert into wedata_demo_db.user_info values ('004', 'Zhao Liu', 45, 'Guangzhou');
insert into wedata_demo_db.user_info values ('005', 'Xiao Ming', 20, 'Beijing');
insert into wedata_demo_db.user_info values ('006', 'Xiao Hong', 30, 'Shanghai');
insert into wedata_demo_db.user_info values ('007', 'Xiao Gang', 25, 'Shenzhen');
insert into wedata_demo_db.user_info values ('008', 'Xiao Li', 40, 'Guangzhou');
insert into wedata_demo_db.user_info values ('009', 'Xiao Zhang', 23, 'Beijing');
insert into wedata_demo_db.user_info values ('010', 'Xiao Wang', 50, 'Shanghai');

select  * from  wedata_demo_db.user_info;
```

©2013-2025 Tencent Cloud. All rights reserved.

Spark SQL

Last updated: 2024-11-25 14:47:27

Note:

You need to start Hive and Spark component services in the EMR cluster.

1. The current user has permissions in the EMR cluster.
2. The corresponding database has been created in Hive, as in the example: wedata_demo_db.

Sample Code

```
-- Create a user information table
create table if not exists wedata_demo_db.user_info (
user_id string COMMENT 'User ID',
user_name string COMMENT 'Username',
user_age int COMMENT 'Age',
city string COMMENT 'City'
) COMMENT 'User Information Table';

-- Insert data into the user information table
insert into wedata_demo_db.user_info values ('001', 'Zhang San', 28, 'Beijing');
insert into wedata_demo_db.user_info values ('002', 'Li Si', 35, 'Shanghai');
insert into wedata_demo_db.user_info values ('003', 'Wang Wu', 22, 'Shenzhen');
insert into wedata_demo_db.user_info values ('004', 'Zhao Liu', 45, 'Guangzhou');
insert into wedata_demo_db.user_info values ('005', 'Xiao Ming', 20, 'Beijing');
insert into wedata_demo_db.user_info values ('006', 'Xiao Hong', 30, 'Shanghai');
insert into wedata_demo_db.user_info values ('007', 'Xiao Gang', 25, 'Shenzhen');
insert into wedata_demo_db.user_info values ('008', 'Xiao Li', 40, 'Guangzhou');
insert into wedata_demo_db.user_info values ('009', 'Xiao Zhang', 23, 'Beijing');
insert into wedata_demo_db.user_info values ('010', 'Xiao Wang', 50, 'Shanghai');

select  * from  wedata_demo_db.user_info;
```

©2013-2025 Tencent Cloud. All rights reserved.

PySpark

Last updated: 2024-11-25 14:47:52

Note:
You need to start Hive and Spark component services in the EMR cluster.
1. The current user has permissions in the EMR cluster.
2. Created corresponding databases and tables in Hive, as shown in the example: wedata_demo_db.
3. PySpark system automatically submits tasks using cluster mode.

Sample Code

```
from pyspark.sql import SparkSession
from pyspark.sql.types import StructType, StructField, IntegerType, StringType

spark = SparkSession.builder.appName("WeDataApp").getOrCreate()

schema = StructType([
    StructField("user_id", IntegerType(), True),
    StructField("user_name", StringType(), True),
    StructField("age", IntegerType(), True)
])

data = [(1, "Alice", 25), (2, "Bob", 30)]
df = spark.createDataFrame(data, schema=schema)

df.show()
```

```
from pyspark.sql import SparkSession
#
spark = SparkSession.builder.appName("WeDataApp").enableHiveSupport().getOrCreate()
#
df = spark.sql("SELECT * FROM WeData_demo_db.user_demo")
#
count = df.count()
#
print("The number of rows in the dataframe is:", count)
```

Parameter description

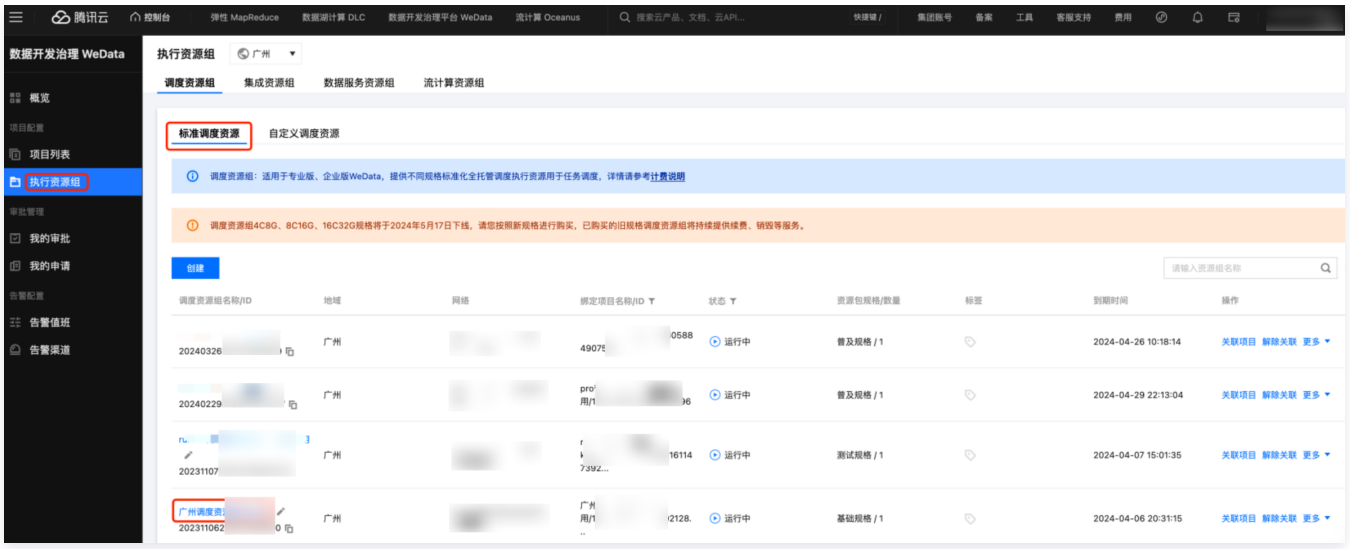
Parameter	Description
Python Version	Supports Python2, Python3.

Use the Python environment of the Schedule Resource Group in PySpark tasks

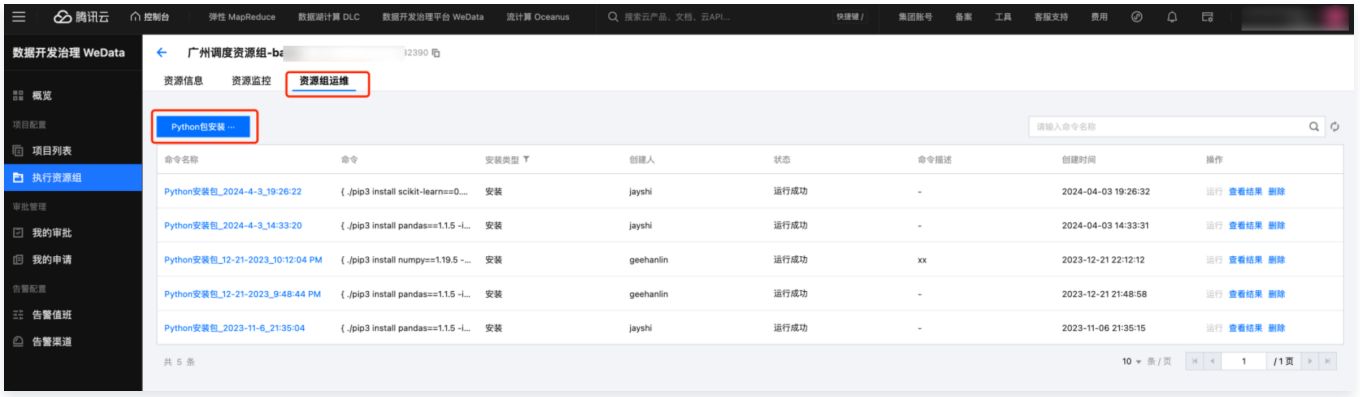
Install Python libraries in the Schedule Resource Group

1. Go to **Project Management** > **Execution Resource Group** > **Standard Scheduling Resource Group** page, click **Resource Details** to enter the resource operation and maintenance page.

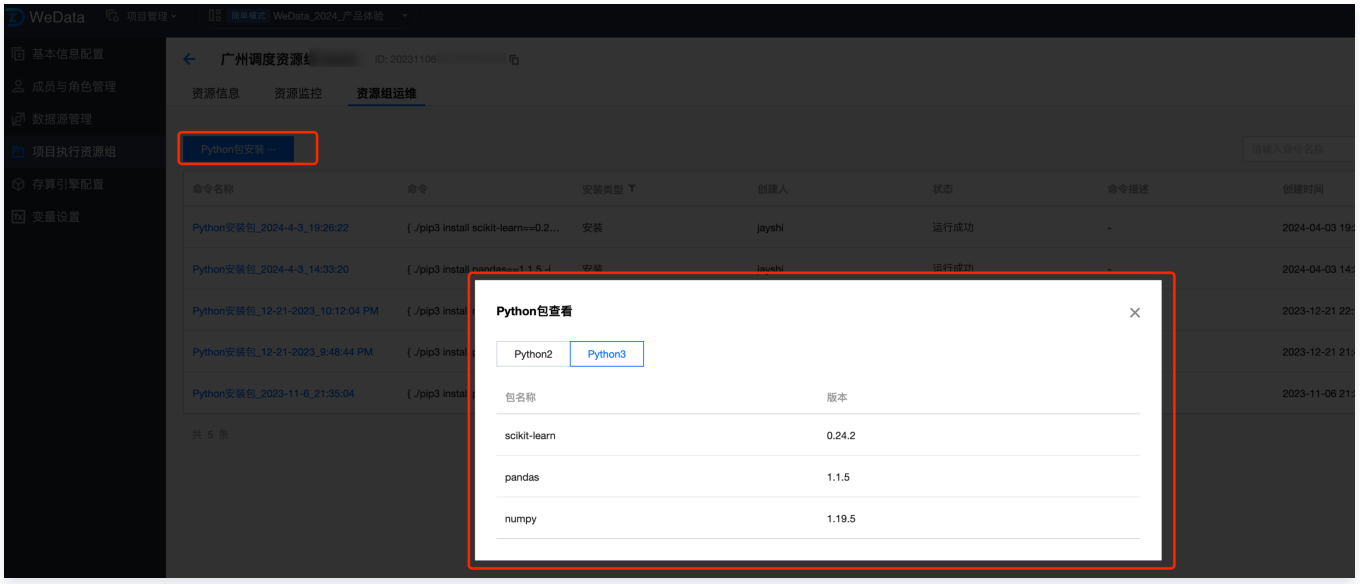
©2013-2025 Tencent Cloud. All rights reserved.



2. On the resource operation and maintenance page, click **Python Package Installation** to install built-in Python libraries. It is recommended to install the Python3 version. For specific installation instructions, refer to Scheduling Resources.



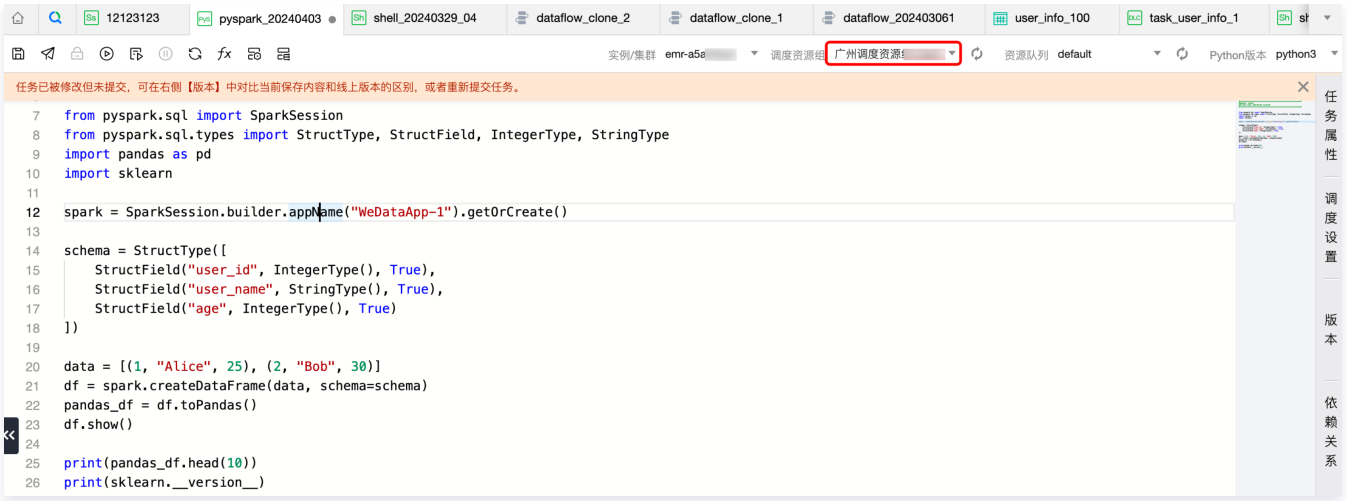
3. Currently, the platform only supports the installation of built-in libraries. Here, install the sklearn and pandas libraries. After installation, you can view the installed Python libraries using the **Python Package View** feature.



Edit PySpark Tasks

1. Create tasks, and select the Schedule Resource Group with the installed Python package.
2. Write PySpark code to use Python libraries, here using pandas and sklearn.

©2013–2025 Tencent Cloud. All rights reserved.




```
from pyspark.sql import SparkSession
from pyspark.sql.types import StructType, StructField, IntegerType, StringType
import pandas as pd
import sklearn

spark = SparkSession.builder.appName("WeDataApp-1").getOrCreate()

schema = StructType([
    StructField("user_id", IntegerType(), True),
    StructField("user_name", StringType(), True),
    StructField("age", IntegerType(), True)
])

data = [(1, "Alice", 25), (2, "Bob", 30)]
df = spark.createDataFrame(data, schema=schema)
pandas_df = df.toPandas()
df.show()

print(pandas_df.head(10))
print(sklearn.__version__)
```

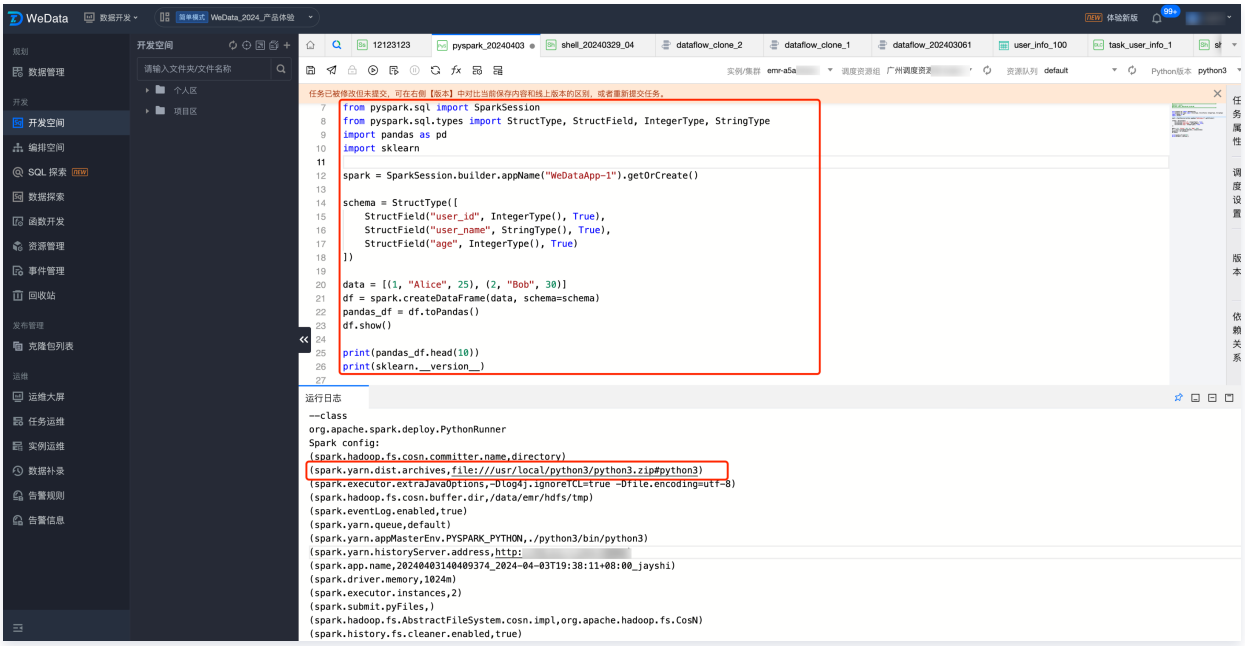
Debug PySpark Tasks

1. Click **Debug Runs** to view the logs and results of the debug run.

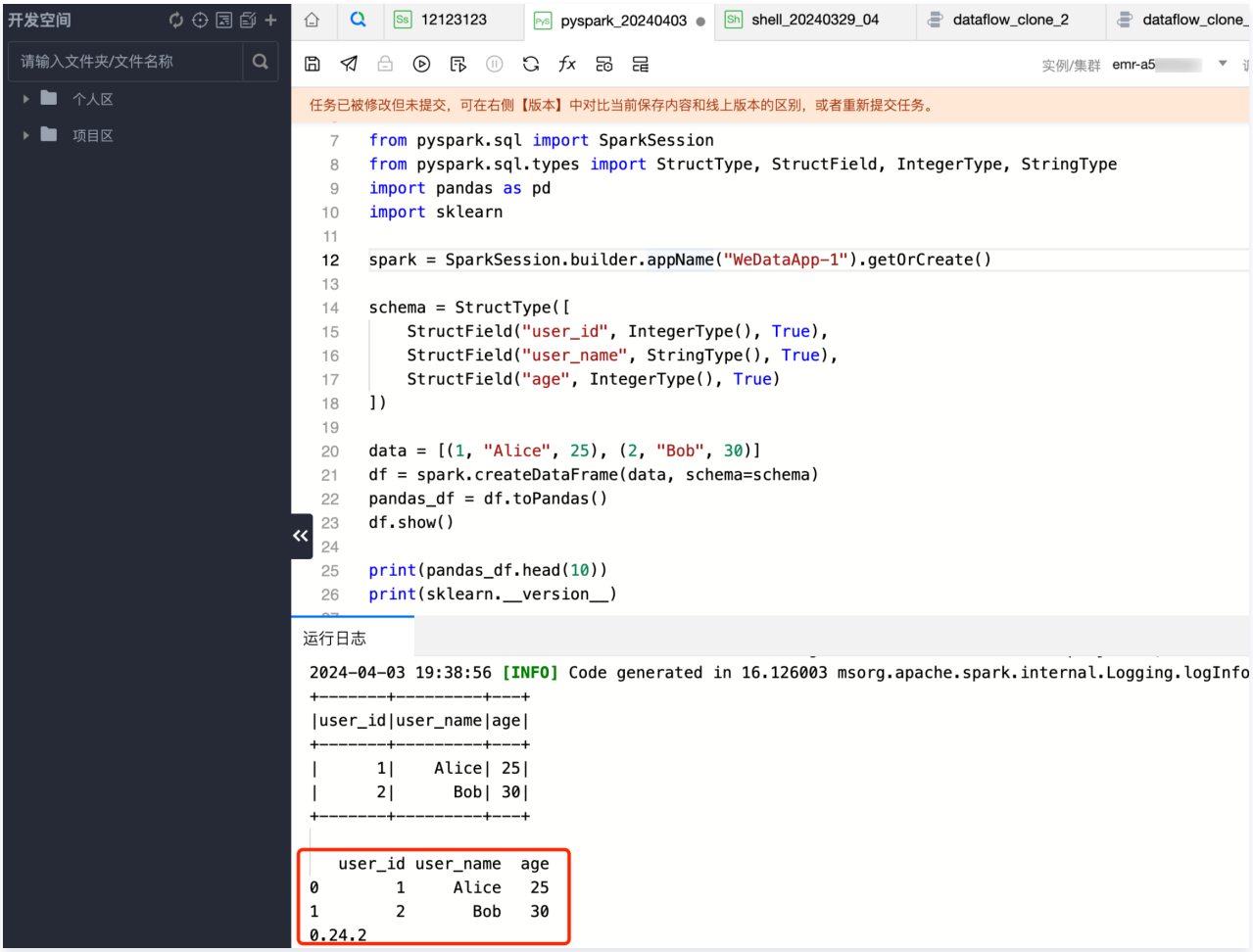
Example: You can view the logs to see that the Python environment of the Schedule Resource Group is used as the task running environment.

```
spark.yarn.dist.archives,file:///usr/local/python3/python3.zip#python3
```

©2013–2025 Tencent Cloud. All rights reserved.



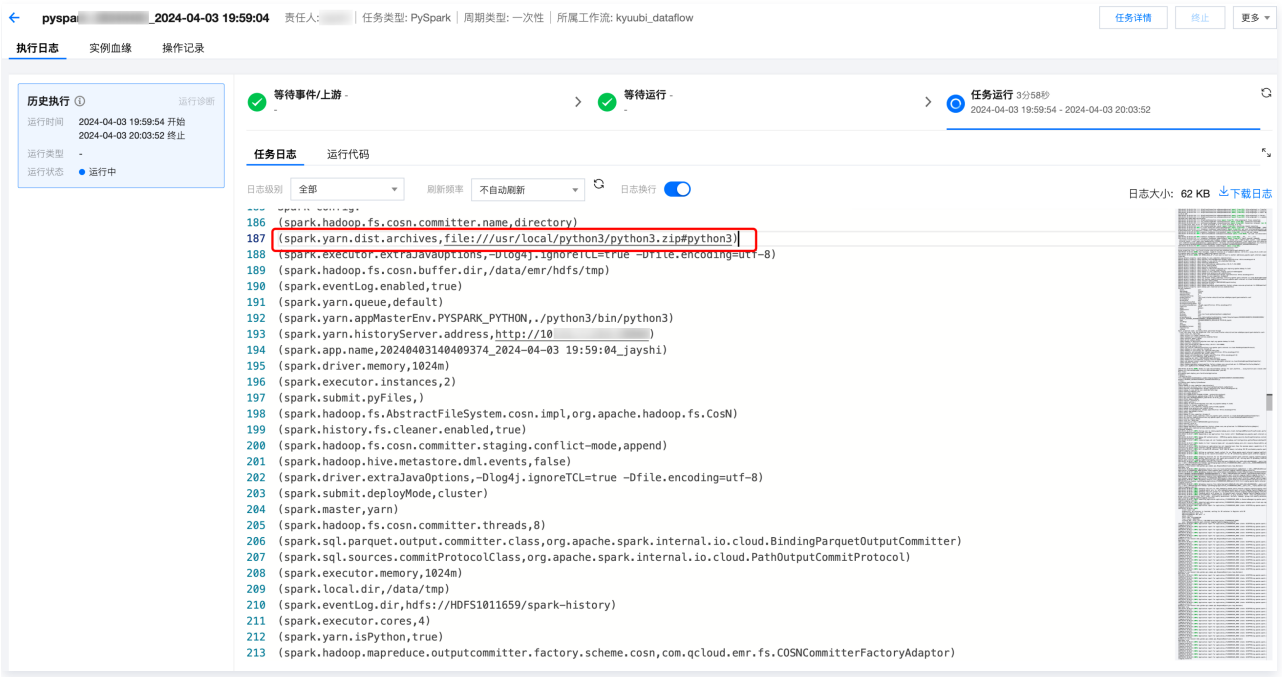
2. View the log results to verify that the installed pandas library was used and the installed sklearn library version was correctly printed.



Cron Scheduling PySpark Tasks

Cron scheduling runs to view the logs and results of the debug run. The logs show that the Python environment of the Schedule Resource Group is used as the task running environment.

©2013–2025 Tencent Cloud. All rights reserved.



©2013–2025 Tencent Cloud. All rights reserved.

Spark

Last updated: 2024-11-25 14:48:37

> **Note:**
> To start Hive and Spark component services in the EMR cluster, COS permissions are required, and resource files need to be stored in COS.
> Example: In the following example, the user has permissions in the EMR cluster.

Feature Overview

Submit a Spark task to the WeData workflow scheduling platform for execution.



Parameter description

Parameter	Description
spark program zip package	The user directly uploads the written Spark program code files, needs to package them into a jar, and then package all dependencies into a single zip file without including directories, just the files themselves.
Execution parameters	Execution parameters for the Spark program without requiring the user to write spark-submit, specify the submitting user, specify the submission queue, or specify the submission mode (default is yarn). Parameter format is: --class mainClass run.jar args or wordcount.py input output.
Application parameters	Application parameters for Spark.

SparkJar example:

Submit a word count task, which requires uploading the file to be counted to COS in advance.

Step 1: Locally write the Spark Jar task

Create engineering project

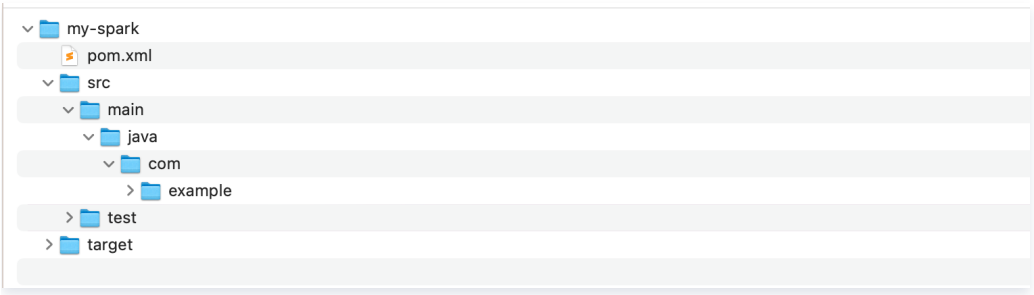
1. For example, use Maven to create a project and introduce Spark dependencies.

> **Description:**
> - Here, groupId and artifactId need to be replaced with the appropriate groupId and artifactId.
> - Here, the scope of the Spark dependency is set to scope, indicating that Spark dependencies need to be provided only during compilation and packaging. Runtime dependencies are provided by the platform.

```
# Generate a Maven project, or you can use an IDE
mvn archetype:generate -DgroupId=com.example -DartifactId=my-spark -DarchetypeArtifactId=maven-archetype-quickstart
```

©2013-2025 Tencent Cloud. All rights reserved.

2. The generated directory structure is as follows:



3. Import dependencies:

```
# Introduce Spark dependencies in pom.xml
<dependencies>
  <dependency>
    <groupId>org.apache.spark</groupId>
    <artifactId>spark-core_2.11</artifactId>
    <version>2.4.7</version>
    <scope>provided</scope>
  </dependency>
</dependencies>
```

Writing code

1. In the src/main/java/com/example directory, create a new Java Class. Enter the class name (e.g., WordCount) and add the sample code to the Class as follows:

```
package com.example;

import java.util.Arrays;
import org.apache.spark.SparkConf;
import org.apache.spark.api.java.JavaPairRDD;
import org.apache.spark.api.java.JavaRDD;
import org.apache.spark.api.java.JavaSparkContext;
import scala.Tuple2;

public class WordCount {
    public static void main(String[] args) {
        // create SparkConf object
        SparkConf conf = new SparkConf().setAppName("WordCount");
        // create JavaSparkContext object
        JavaSparkContext sc = new JavaSparkContext(conf);
        // read input file to RDD
        JavaRDD<String> lines = sc.textFile(args[0]);
        // split each line into words
        JavaRDD<String> words = lines.flatMap(line -> Arrays.asList(line.split(" ")).iterator());
        // count the occurrence of each word
        JavaPairRDD<String, Integer> wordCounts = words.mapToPair(word -> new Tuple2<>(word,
1)).reduceByKey((x, y) -> x + y);
        // save the word counts to output file
        wordCounts.saveAsTextFile(args[1]);
    }
}
```

Description:
Here, the scope of the Spark dependency is set to scope, indicating that Spark dependencies need to be provided only during compilation and packaging. Runtime dependencies are provided by the platform.

©2013–2025 Tencent Cloud. All rights reserved.

2. Package the code into a jar file and add the following packaging plugin in Maven:

```
<build>
    <plugins>
        <plugin>
            <groupId>org.apache.maven.plugins</groupId>
            <artifactId>maven-compiler-plugin</artifactId>
            <configuration>
                <source>1.8</source>
                <target>1.8</target>
                <encoding>utf-8</encoding>
            </configuration>
        </plugin>
        <plugin>
            <artifactId>maven-assembly-plugin</artifactId>
            <configuration>
                <descriptorRefs>
                <descriptorRef>jar-with-dependencies</descriptorRef>
                </descriptorRefs>
            </configuration>
            <executions>
                <execution>
                    <id>make-assembly</id>
                    <phase>package</phase>
                    <goals>
                        <goal>single</goal>
                    </goals>
                </execution>
           </executions>
         </plugin>
    </plugins>
</build>
```

3. Then, execute the following command in the project root directory:

```
mvn package
```

4. In the target directory, you can see the jar file with dependencies. Here, it is named my-spark-1.0-SNAPSHOT-jar-with-dependencies.jar.

Data preparation

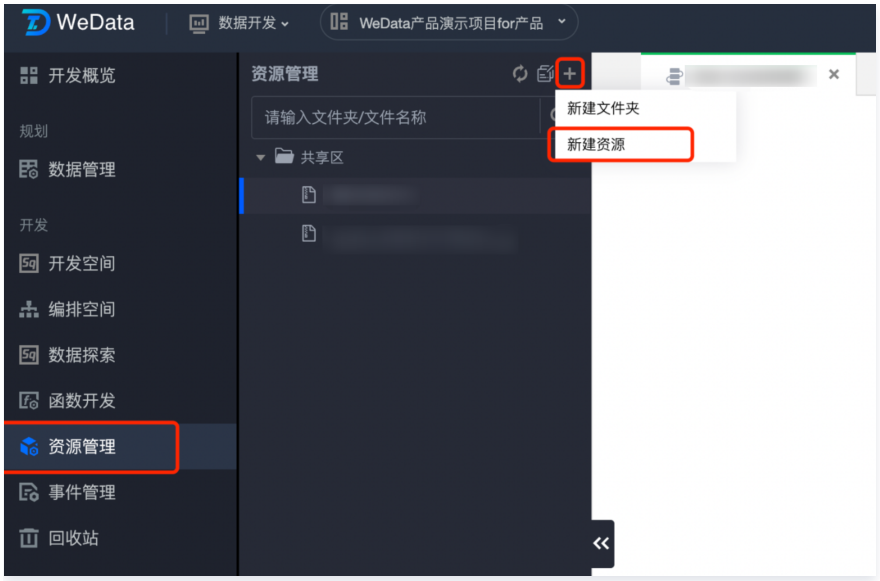
Since WeData data development only supports zip files, you first need to package the jar file into a zip file. You also need to perform the following operations to get the zip file. If there are other dependency configuration files, they can also be packaged into the zip file.

```
zip spark-wordcount.zip my-spark-1.0-SNAPSHOT-jar-with-dependencies.jar
```

Step 2: Upload the Spark Jar task package

1. Create a new resource file in Resource Management and upload the resource package.

©2013-2025 Tencent Cloud. All rights reserved.



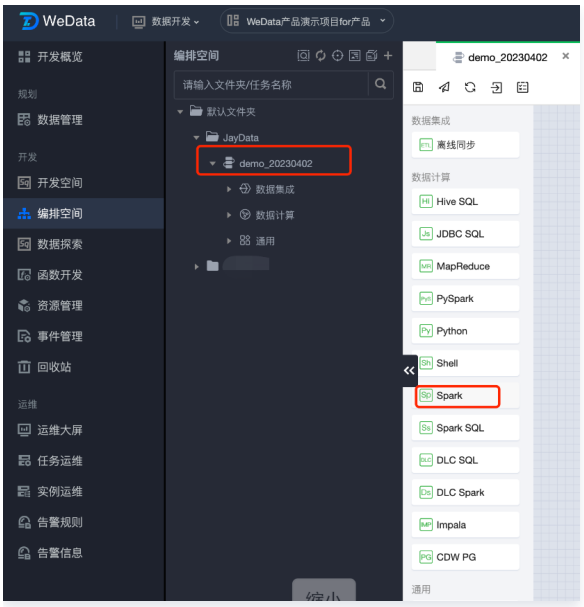
2. Create Resource Configuration:



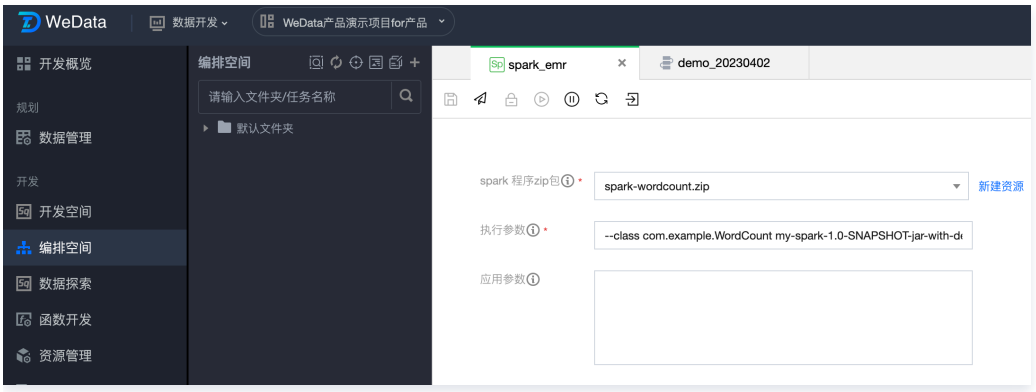
Step 3: Create Spark Jar task and configure scheduling

1. Create a new workflow in the orchestration space, and create a Spark task in the workflow.

©2013–2025 Tencent Cloud. All rights reserved.



2. Fill in task parameters.



3. Example format of execution parameters:

```
--class mainClass run.jar args or wordcount.py input output
```

4. A complete configuration sample is as follows:

```
--class com.example.WordCount  my-spark-1.0-SNAPSHOT-jar-with-dependencies.jar  cosn://wedata-demo-
1314991481/wordcount.txt
cosn://wedata-demo-1314991481/result/output
```

Note:

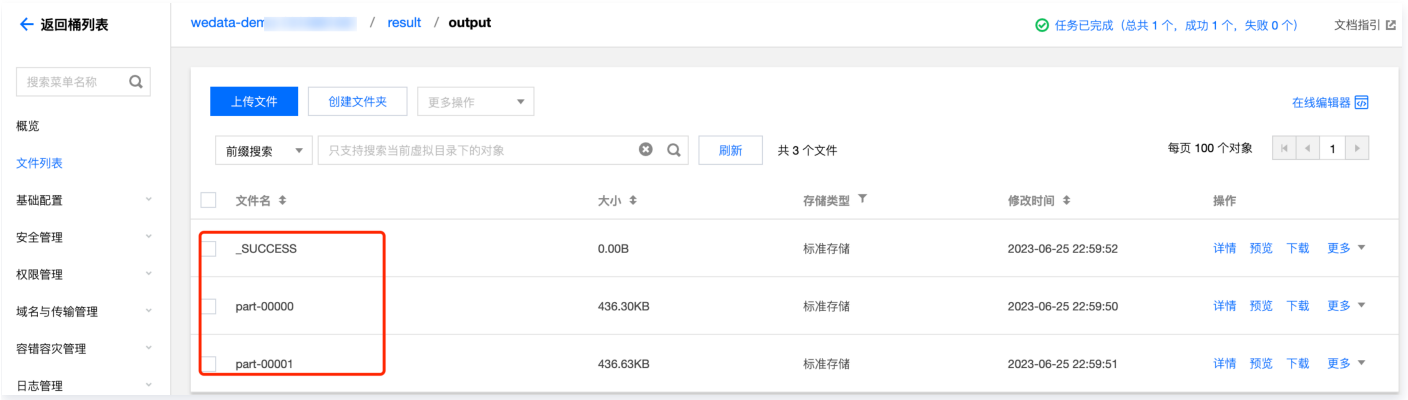
- cosn://wedata-demo-1314991481/wordcount.txt is the **COS Path of the file to be processed** .
- cosn://wedata-demo-1314991481/result/output is the **COS Path for the output calculation results** . This folder directory must not be pre-created, or the run will fail.

5. An example file of wordcount.txt is as follows:

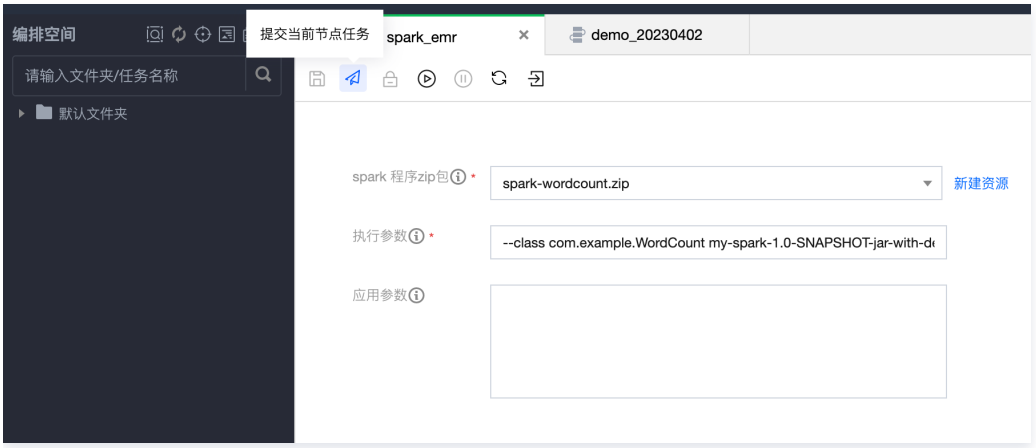
```
hello WeData
hello Spark
hello Scala
hello PySpark
hello Hive
```

©2013-2025 Tencent Cloud. All rights reserved.

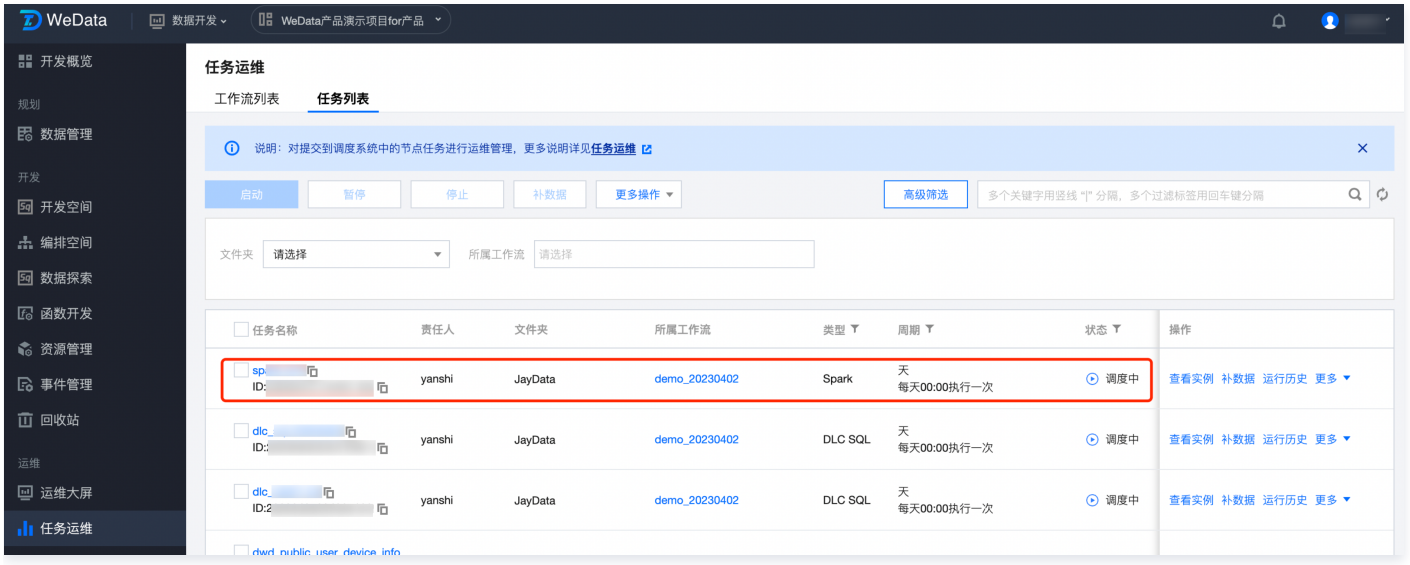
6. After debugging and running, check the calculation results as follows:



7. Publish the Spark task and start scheduling. Submit the Spark Jar task:



8. Spark Jar task operation and maintenance are shown below:



©2013–2025 Tencent Cloud. All rights reserved.

Impala

Last updated: 2024-11-25 14:49:07

Note:

You need to start the Impala component service in the EMR cluster.

1. The current user has permissions in the EMR cluster.
2. The corresponding database has been created in Impala, such as wedata_demo_db in the example.

Sample code

```
-- Create a user information table
create table if not exists wedata_demo_db.user_info (
user_id string COMMENT 'User ID',
user_name string COMMENT 'Username',
user_age int COMMENT 'Age',
city string COMMENT 'City'
) COMMENT 'User Information Table';

-- Insert data into the user information table
insert into wedata_demo_db.user_info values ('001', 'Zhang San', 28, 'Beijing');
insert into wedata_demo_db.user_info values ('002', 'Li Si', 35, 'Shanghai');
insert into wedata_demo_db.user_info values ('003', 'Wang Wu', 22, 'Shenzhen');
insert into wedata_demo_db.user_info values ('004', 'Zhao Liu', 45, 'Guangzhou');
insert into wedata_demo_db.user_info values ('005', 'Xiao Ming', 20, 'Beijing');
insert into wedata_demo_db.user_info values ('006', 'Xiao Hong', 30, 'Shanghai');
insert into wedata_demo_db.user_info values ('007', 'Xiao Gang', 25, 'Shenzhen');
insert into wedata_demo_db.user_info values ('008', 'Xiao Li', 40, 'Guangzhou');
insert into wedata_demo_db.user_info values ('009', 'Xiao Zhang', 23, 'Beijing');
insert into wedata_demo_db.user_info values ('010', 'Xiao Wang', 50, 'Shanghai');

select  * from  wedata_demo_db.user_info;
```

©2013-2025 Tencent Cloud. All rights reserved.

Trino

Last updated: 2024-11-25 14:49:30

> **Note:**
> Trino Component Services need to be started in the EMR clusters.

Sample code

```
SHOW CATALOGS;
SELECT current_catalog;

CREATE SCHEMA IF NOT EXISTS hive.wedata_demo_db;

-- Create a user information table
create table if not exists hive.wedata_demo_db.user_info (
user_id string COMMENT 'User ID',
user_name string COMMENT 'Username',
user_age int COMMENT 'Age',
city string COMMENT 'City'
) COMMENT 'User Information Table';

-- Insert data into the user information table
insert into hive.wedata_demo_db.user_info values ('001', 'Zhang San', 28, 'beijing');
insert into hive.wedata_demo_db.user_info values ('002', 'Li Si', 35, 'shanghai');
insert into hive.wedata_demo_db.user_info values ('003', 'Wang Wu', 22, 'shenzhen');
insert into hive.wedata_demo_db.user_info values ('004', 'Zhao Liu', 45, 'guangzhou');
insert into hive.wedata_demo_db.user_info values ('005', 'Xiao Ming', 20, 'beijing');
insert into hive.wedata_demo_db.user_info values ('006', 'Xiao Hong', 30, 'shanghai');
insert into hive.wedata_demo_db.user_info values ('007', 'Xiao Gang', 25, 'shenzhen');
insert into hive.wedata_demo_db.user_info values ('008', 'Xiao Li', 40, 'guangzhou');
insert into hive.wedata_demo_db.user_info values ('009', 'Xiao Zhang', 23, 'beijing');
insert into hive.wedata_demo_db.user_info values ('010', 'Xiao Wang', 50, 'shanghai');

-- Query data
select  * from  hive.wedata_demo_db.user_info;
```

©2013-2025 Tencent Cloud. All rights reserved.

DLC
DLC SQL

Last updated: 2024-11-25 14:49:59

> **Note:**
> You need to bind the DLC engine. Currently, three engines are supported: Spark SQL, Spark Job, and Presto. For engine kernel details, see DLC Engine Kernel Versions.
> 1. The current user needs corresponding DLC computing resources and database table permissions
> 2. The corresponding database table has already been created in DLC.

Feature Overview

Submit a DLC SQL task to the WeData Workflow Scheduling Platform for execution. When selecting the DLC data source type, "Advanced settings" are provided, supporting the configuration of Presto and Spark SQL parameters.



When using the Spark Job engine, you can configure the job resource specifications and parameters. The resource configuration cannot exceed the limits of the computing resources themselves.



Configuration instructions:

Configuration Item	Description
Resource configuration methods	There are two ways: Cluster Default Configuration and Custom Configuration 1. Use the Cluster Default Configuration Use the current task computing resource cluster configuration 2. Custom Custom configuration for Executor and Driver
Executor Resources	Enter the required number of resources. 1 CU is roughly equivalent to 1 CPU core and 4 GB memory. 1. Small: Single computing unit (1 CU) 2. Medium: Two computing units (2 CU) 3. Large: Four computing units (4 CU) 4. Xlarge (Extra Large): Eight computing units (8 CU)
Number of Executors	An Executor is a computing node or instance responsible for executing tasks and processing computational work. Each Executor uses the configured amount of resources.

©2013–2025 Tencent Cloud. All rights reserved.

Driver Resources	Enter the required number of Driver resources. 1 CU is roughly equivalent to 1 CPU core and 4 GB memory. 1. Small: Single computing unit (1 CU) 2. Medium: Two computing units (2 CU) 3. Large (Large): Four computing units (4 CU). 4. Xlarge (Extra Large): Eight computing units (8 CU).

Sample code

```
  -- Create a user information table
create table if not exists wedata_demo_db.user_info (
user_id string COMMENT 'User ID',
user_name string COMMENT 'Username',
user_age int COMMENT 'Age',
city string COMMENT 'City'
) COMMENT 'User Information Table';

-- Insert data into the user information table
insert into wedata_demo_db.user_info values ('001', 'Zhang San', 28, 'Beijing');
insert into wedata_demo_db.user_info values ('002', 'Li Si', 35, 'Shanghai');
insert into wedata_demo_db.user_info values ('003', 'Wang Wu', 22, 'Shenzhen');
insert into wedata_demo_db.user_info values ('004', 'Zhao Liu', 45, 'Guangzhou');
insert into wedata_demo_db.user_info values ('005', 'Xiao Ming', 20, 'Beijing');
insert into wedata_demo_db.user_info values ('006', 'Xiao Hong', 30, 'Shanghai');
insert into wedata_demo_db.user_info values ('007', 'Xiao Gang', 25, 'Shenzhen');
insert into wedata_demo_db.user_info values ('008', 'Xiao Li', 40, 'Guangzhou');
insert into wedata_demo_db.user_info values ('009', 'Xiao Zhang', 23, 'Beijing');
insert into wedata_demo_db.user_info values ('010', 'Xiao Wang', 50, 'Shanghai');

select  * from  wedata_demo_db.user_info;
```

Note:
When using Iceberg External Tables, there are differences in SQL syntax compared to Iceberg Native Tables. For details, see DLC Iceberg External Table vs. Native Table Syntax Differences.

Presto Engine Sample Code:

Applicable Table Types: Native Iceberg Tables, External Iceberg Tables.

```
  CREATE TABLE `cpt_demo`.`dempts` (
    id bigint COMMENT 'id number',
    num int,
    eno float,
    dno double,
    cno decimal(9,3),
    flag boolean,
    data string,
    ts_year timestamp,
    date_month date,
    bno binary,
    point struct<x: double, y: double>,
    points array<struct<x: double, y: double>>,
    pointmaps map<struct<x: int>, struct<a: int>>
)
COMMENT 'table documentation'
PARTITIONED BY (bucket(16, id), years(ts_year), months(date_month), identity(bno), bucket(3, num),
truncate(10, data));
```

©2013–2025 Tencent Cloud. All rights reserved.

SparkSQL Engine Sample Code

Applicable Table Types: Native Iceberg Tables, External Iceberg Tables.

```
CREATE TABLE `cpt_demo`.`dempts` (
    id bigint COMMENT 'id number',
    num int,
    eno float,
    dno double,
    cno decimal(9,3),
    flag boolean,
    data string,
    ts_year timestamp,
    date_month date,
    bno binary,
    point struct<x: double, y: double>,
    points array<struct<x: double, y: double>>,
    pointmaps map<struct<x: int>, struct<a: int>>
)
COMMENT 'table documentation'
PARTITIONED BY (bucket(16, id), years(ts_year), months(date_month), identity(bno), bucket(3, num),
truncate(10, data));
```

SparkSQL Job Engine Sample Code

Applicable Table Types: Native Iceberg Tables, External Iceberg Tables.

```
CREATE TABLE `cpt_demo`.`dempts` (
    id bigint COMMENT 'id number',
    num int,
    eno float,
    dno double,
    cno decimal(9,3),
    flag boolean,
    data string,
    ts_year timestamp,
    date_month date,
    bno binary,
    point struct<x: double, y: double>,
    points array<struct<x: double, y: double>>,
    pointmaps map<struct<x: int>, struct<a: int>>
)
COMMENT 'table documentation'
PARTITIONED BY (id, ts_year, date_month);
```

Note
More DLC Syntax References: DLC SQL Syntax Overview.

©2013–2025 Tencent Cloud. All rights reserved.

DLC PySpark

Last updated: 2024-11-25 14:50:32

> **Note:**
> Binding the DLC engine is required. Currently, DLC PySpark supports the Spark job engine. For details on the Engine Kernel, please refer to the DLC Engine Kernel Versions.

Feature Overview

Create a DLC PySpark task in WeData, submit it to the WeData Scheduling Platform, and execute it on the DLC engine.

Task parameter description

In the DLC PySpark task properties, you can add DLC PySpark task data access policies, entry parameters, dependency resources, Spark task conf parameters, and task images.

Parameter name	Parameter description
Data access policy	Required, a security policy for accessing COS data during task execution. For details, refer to DLC Configuring Data Access Policy.
Entry parameters	Optional, the entry parameters of the program. Multiple parameters can be specified, separated by "space".
Dependency resources	Optional, supports selecting --py-files, --files, --archives. Multiple COS paths for each resource can be entered, separated by commas (,).
Conf parameters	Optional, parameters starting with spark., written in k=v format. Multiple parameters should be written on separate lines. Example: spark.network.timeout=120s.
Task image	The image used for task execution. If a specific image is required for the task, you can choose between DLC built-in images and custom images. For details, refer to DLC Spark Images.
Resource configuration	• Use cluster resource configuration: Use the default resource configuration parameters of the cluster. • Custom Definition: Custom resource usage parameters for tasks, including executor size, driver size, and number of executors.

Sample code

```
from os.path import abspath

from pyspark.sql import SparkSession

if __name__ == "__main__":
    spark = SparkSession \
        .builder \
        .appName("Operate DB Example") \
        .getOrCreate()

    # 1. Create Database
    spark.sql("CREATE DATABASE IF NOT EXISTS DataLakeCatalog.dlc_db_test_py COMMENT 'demo test' ")
    # 2. Create Internal Table
    spark.sql("CREATE TABLE IF NOT EXISTS DataLakeCatalog.dlc_db_test_py.test(id int,name string,age
int) ")
    # 3. Write Internal Data
    spark.sql("INSERT INTO DataLakeCatalog.dlc_db_test_py.test VALUES (1,'Andy',12),(2,'Justin',3) ")
    # 4. Query Internal Data
```

©2013-2025 Tencent Cloud. All rights reserved.

```
    spark.sql("SELECT * FROM DataLakeCatalog.dlc_db_test_py.test ").show()

    # 5. Create External Table
    spark.sql("CREATE EXTERNAL TABLE IF NOT EXISTS DataLakeCatalog.dlc_db_test_py.ext_test(id int, name
string, age int) ROW FORMAT SERDE 'org.apache.hive.hcatalog.data.JsonSerDe' STORED AS TEXTFILE LOCATION
'cosn://cos-bucket-name/ext_test' ")
    # 6. Write External Data
    spark.sql("INSERT INTO DataLakeCatalog.dlc_db_test_py.ext_test VALUES (1,'Andy',12),(2,'Justin',3)
")
    # 7. Query External Data
    spark.sql("SELECT * FROM DataLakeCatalog.dlc_db_test_py.ext_test ").show()
    spark.stop()
```

©2013–2025 Tencent Cloud. All rights reserved.

TCHouse-P

Last updated: 2024-11-25 14:51:12

Note:

- Need to bind TCHouse-P engine.
- Permissions corresponding to TCHouse-P database tables need to be set in TCHouse-P.

Feature Overview

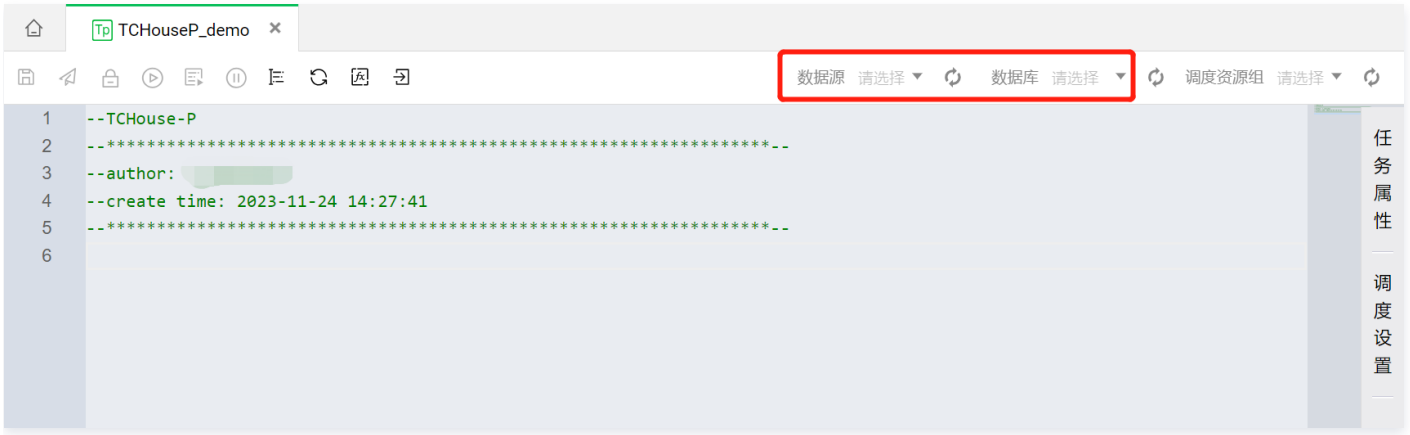
Submit a TCHouse-P task to the WeData Workflow Scheduling Platform.



Sample code:

```
-- Create a library
CREATE DATABASE wedata_demo_db;
```

Switch the database to the corresponding library in the TCHouse-P task.



Execute SQL on the task page:

```
-- Create schema
CREATE SCHEMA demo;

-- Create a table
CREATE TABLE demo.user_info (
  id SERIAL PRIMARY KEY,
  name VARCHAR(255) NOT NULL
);

-- Insert data
```

©2013-2025 Tencent Cloud. All rights reserved.

```
INSERT INTO demo.user_info (name) VALUES ('Alice'), ('Bob');


-- View data
select * from  demo.user_info;

-- View the schema of the corresponding library
SELECT schema_name
FROM information_schema.schemata
WHERE catalog_name = 'wedata_demo_db';

-- View tables in the corresponding schema
SELECT table_name
FROM information_schema.tables
WHERE table_schema = 'demo';
```

©2013–2025 Tencent Cloud. All rights reserved.

TDSQL for PostgreSQL

Last updated: 2024-11-25 14:51:31

> **Note:**
> The project needs to be linked to the TDSQL PostgreSQL data source.

Sample code

```
-- Create schema
create schema if not exists wedata_demo_schema ;

-- Create a table
CREATE TABLE wedata_demo_schema.user_info (
  id SERIAL PRIMARY KEY,
  name VARCHAR(255) NOT NULL
);

-- View tables in the corresponding schema
SELECT table_name
FROM information_schema.tables
WHERE table_schema = 'wedata_demo_schema';

-- Insert data
INSERT INTO wedata_demo_schema.user_info (name) VALUES ('Alice'), ('Bob');

-- View data
select * from  wedata_demo_schema.user_info;
```

©2013-2025 Tencent Cloud. All rights reserved.

General Node
Branch Node

Last updated: 2024-11-25 14:51:59

A Branch Node is a logic control node provided by WeData Data Development. Branch Nodes can execute different downstream tasks based on different branch logic defined in the system.

Use Limits

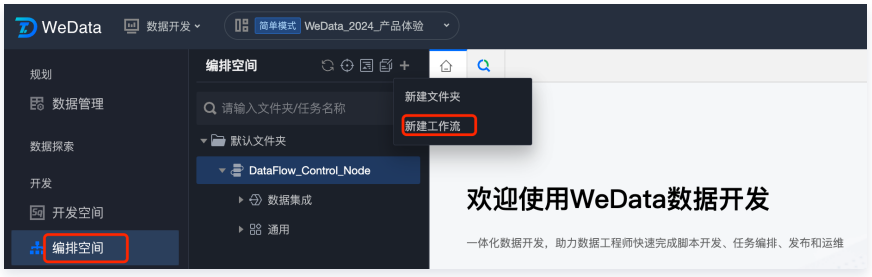
Note:

1. Parameter passing between tasks in a workflow during a test run is currently not supported, so you can specify the parameter values to be tested in the project parameters or task parameters when debugging a Branch Node.
2. A Branch Node supports up to 50 branch conditions.
3. Cross-project cloning is not currently supported for Branch Nodes.

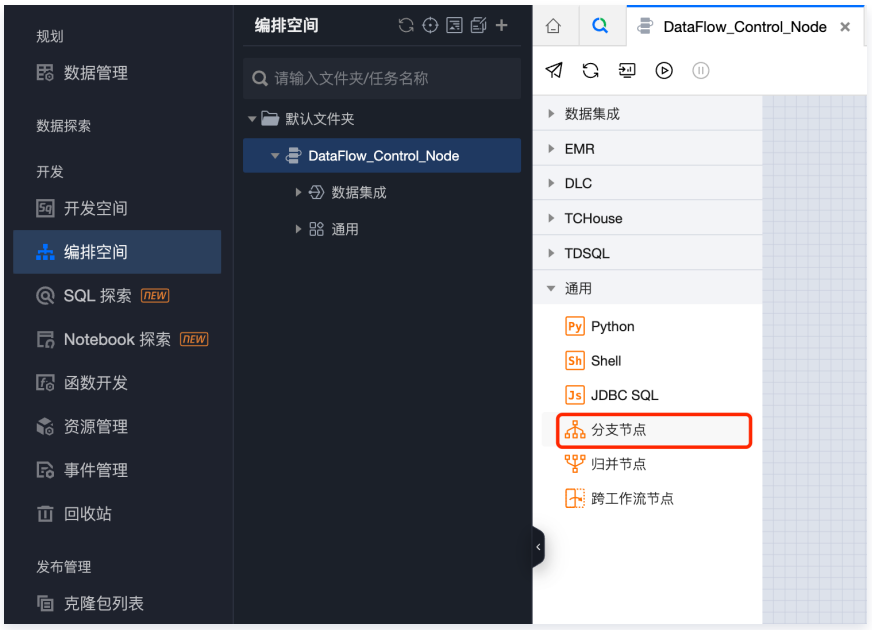
Branch Node Usage

Enter the WeData console and navigate to Project Management. In the project list, click **Data Development**. After entering the Data Development module, switch to the **Data Development** > **Orchestration Space** interface.

1. Create a new workflow in the upper left corner of the Orchestration Space.



2. Double-click the workflow. Under General, click Branch Node to create a Branch Node.



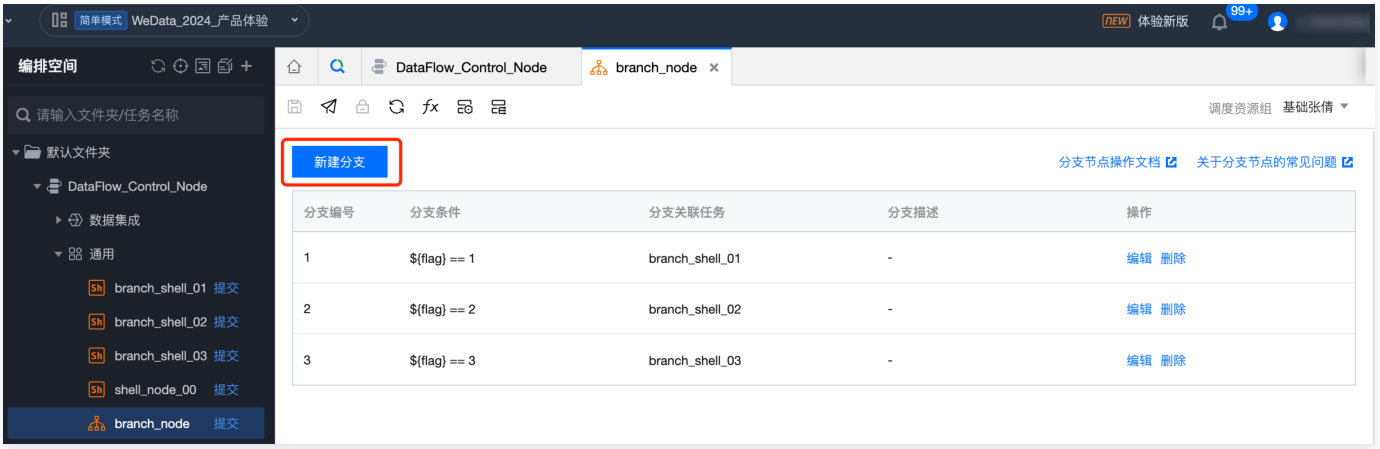
3. In the dialog box for creating a new Branch Node task, enter the task name, click **Confirm to Create Task**, and enter the Branch Node Configuration Interface.
4. Configure the Branch Node conditions. Click **Create Branch**, enter the branch conditions, associated tasks, and description information in the popup window.

©2013-2025 Tencent Cloud. All rights reserved.



Parameter name	Parameter description
Branch Conditions	Branch conditions only support logical judgment operations in Python. If the branch condition evaluates to true, it means the branch condition is met. Scheduling parameters and parameters referenced by the parent task can be used in branch conditions, refer to Parameter Usage.
Associated tasks	If a branch condition evaluates to true, the associated tasks will run automatically. Restriction on associated tasks: **Note:** • Repeated selection of already chosen associated tasks is not allowed. • Direct and indirect upstream tasks of a branch node are not permitted. If the branch condition evaluates to false, the associated task will fail during workflow debug runs In cron scheduling, the instance status will default to 'Skip Execution', and the directly downstream tasks will also automatically change to 'Skip Execution'.
Description	A simple description of branch conditions.

For example, in the following configuration, different scheduling parameter values result in different task executions. When ${flag} == 1, run the task branch_shell_01; when ${flag} == 2, run the task branch_shell_02.



5. Click **Save** to save the branch node configuration.
6. Click **Submit** in the toolbar to submit the task to the cron schedule.

Configuration Samples

©2013–2025 Tencent Cloud. All rights reserved.

1. Create a new workflow DataFlow_Control_Node in **Data Development** > **Orchestration Space**.
2. In the workflow DataFlow_Control_Node, create a shell task shell_node_00, with the task code as follows:

```
echo "hello"
echo "1,2,3"
```

3. Configure the output of the current task parameters in **Scheduling Settings** > **Parameter Passing**. The parameter name is flag. Upon task completion, the value of flag will be 1.



4. Create downstream nodes of the branch node. For example, the branch node's downstream has three branch tasks: branch_shell_01, branch_shell_02, and branch_shell_03. Print the corresponding task name in each task. The task codes for the three branches are as follows:

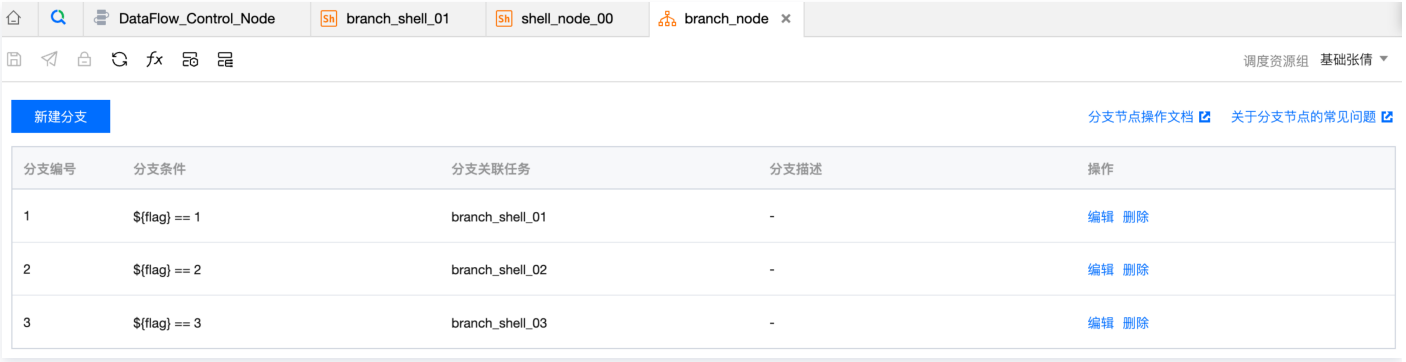
```
echo "hello - branch_shell_1"
```

```
echo "hello - branch_shell_2"
```

```
echo "hello - branch_shell_3"
```

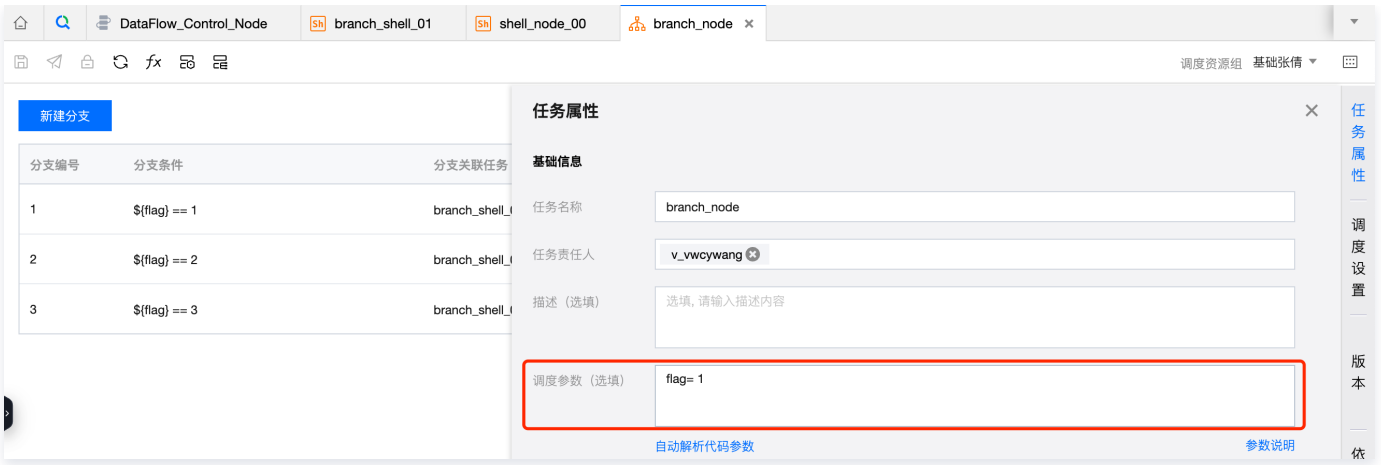
5. Create a branch node branch_node, configure branch conditions and associated tasks, and save the tasks. Configuration is as follows:

Branch Conditions	Associated Branch Tasks
${flag} == 1	branch_shell_01
${flag} == 2	branch_shell_02
${flag} == 3	branch_shell_03

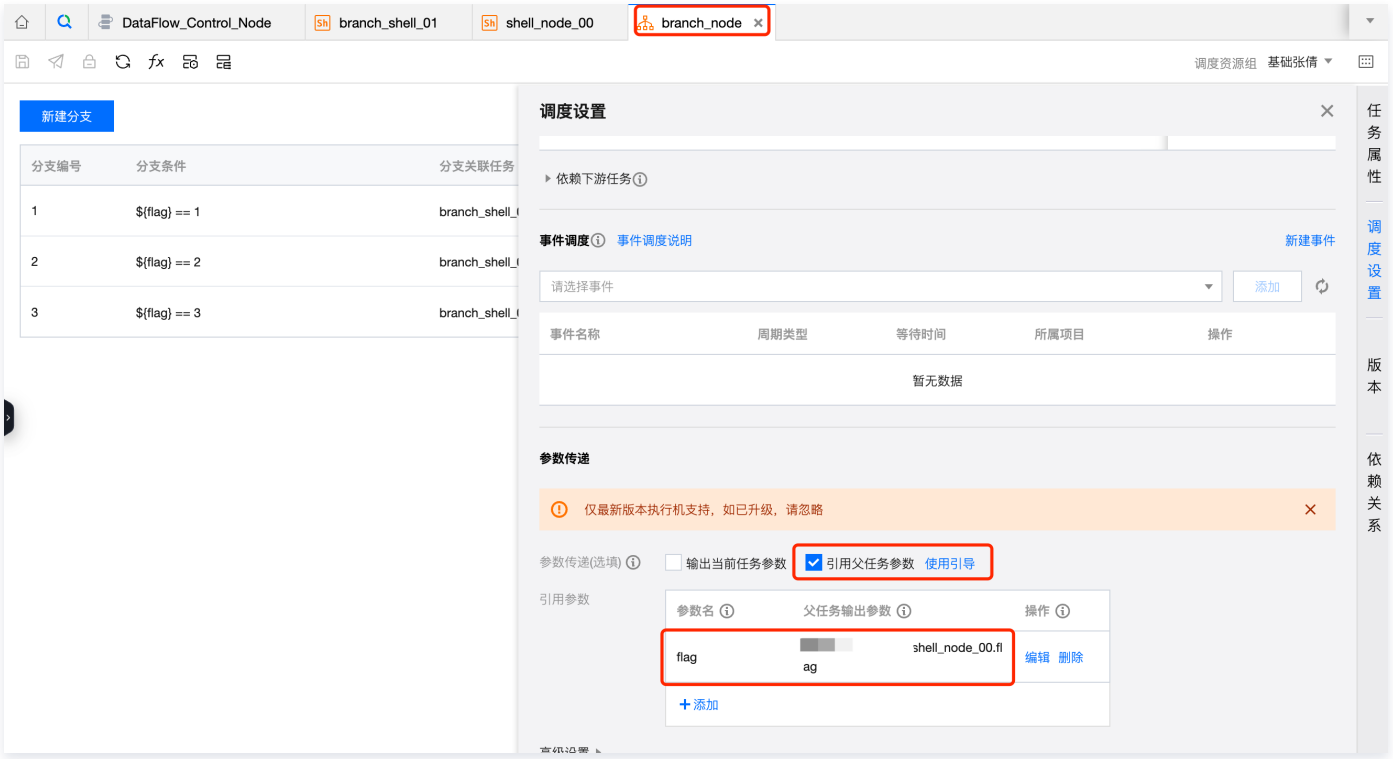


6. In **Task Attributes** > **Scheduling Parameters**, define the scheduling parameter value: flag = 1, then save the task.

©2013–2025 Tencent Cloud. All rights reserved.

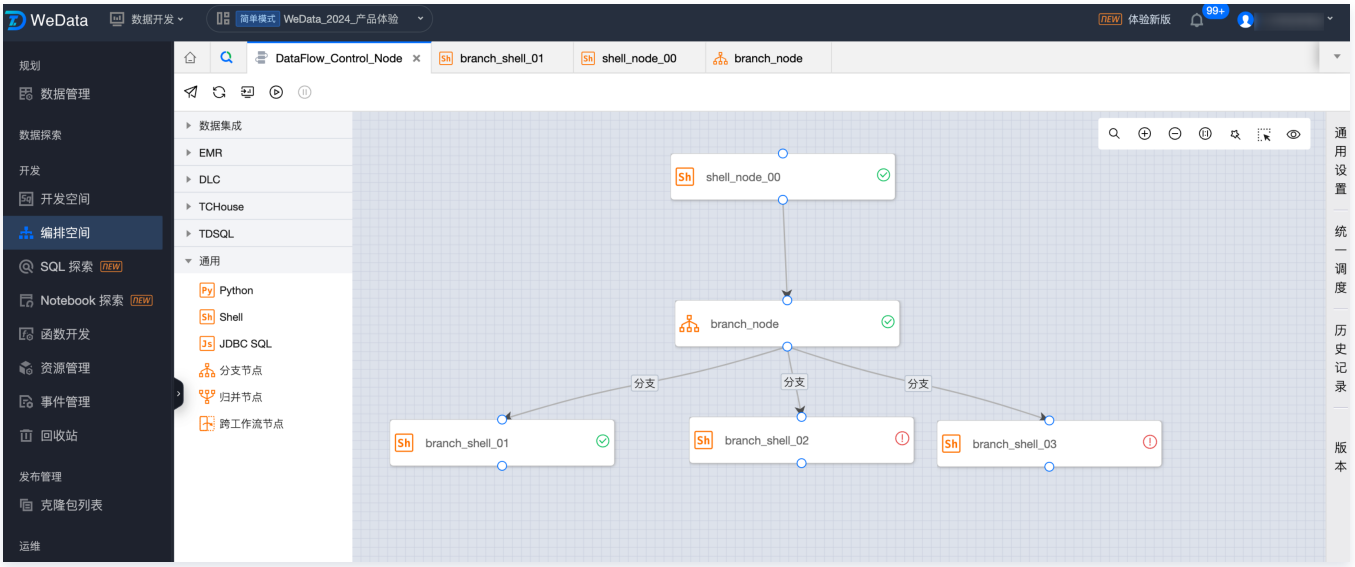


7. Configure shell_node_00 as the upstream of branch_node, and select to refer to parent task parameters in parameter passing, setting the parameter name as flag. The parent task output parameter is shell_node_00.flag.



8. Finally, go back to the workflow to view the configuration results as follows:

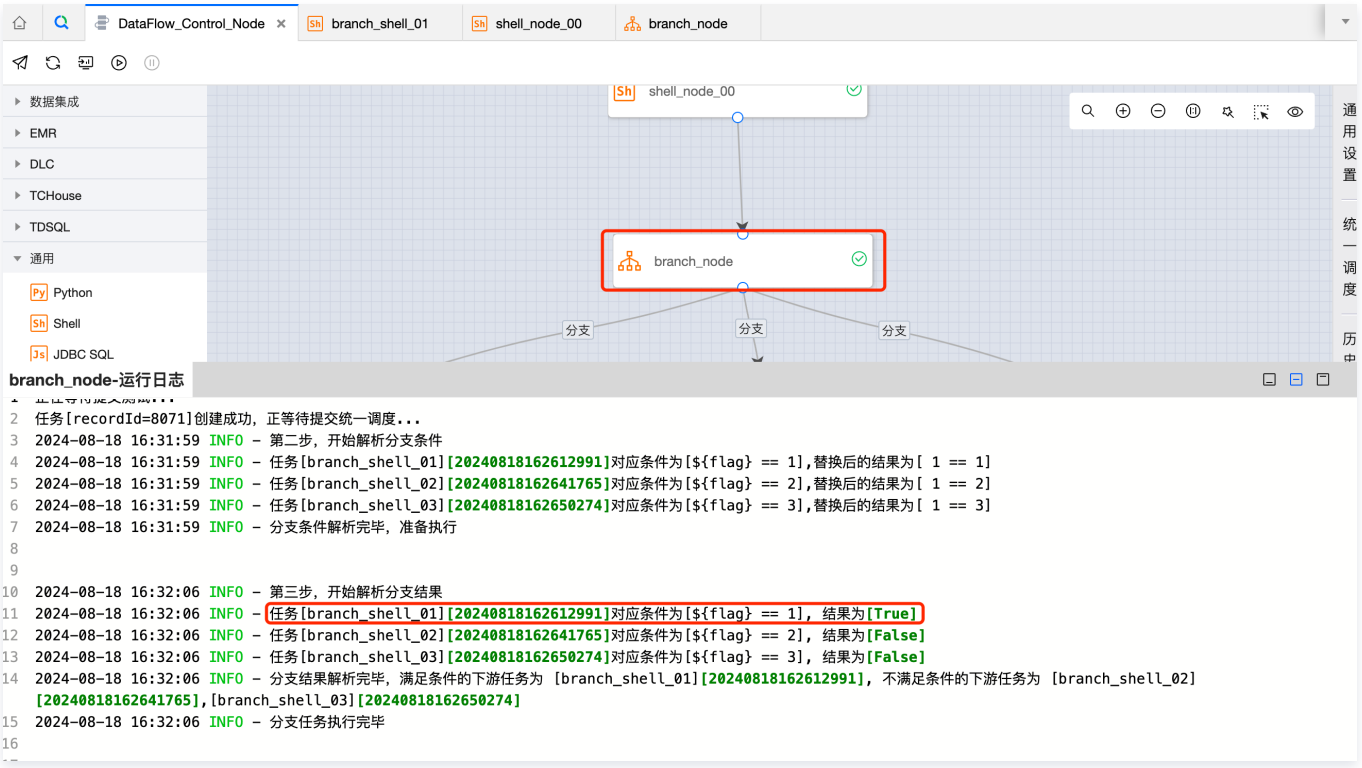
©2013–2025 Tencent Cloud. All rights reserved.



9. Debug Run

9.1 In the workflow toolbar, click **Test Run** to check the workflow test results:

9.2 Right-click to view the running log of branch node branch_node. Branch condition 1 is met, so branch_shell_01 is executed.



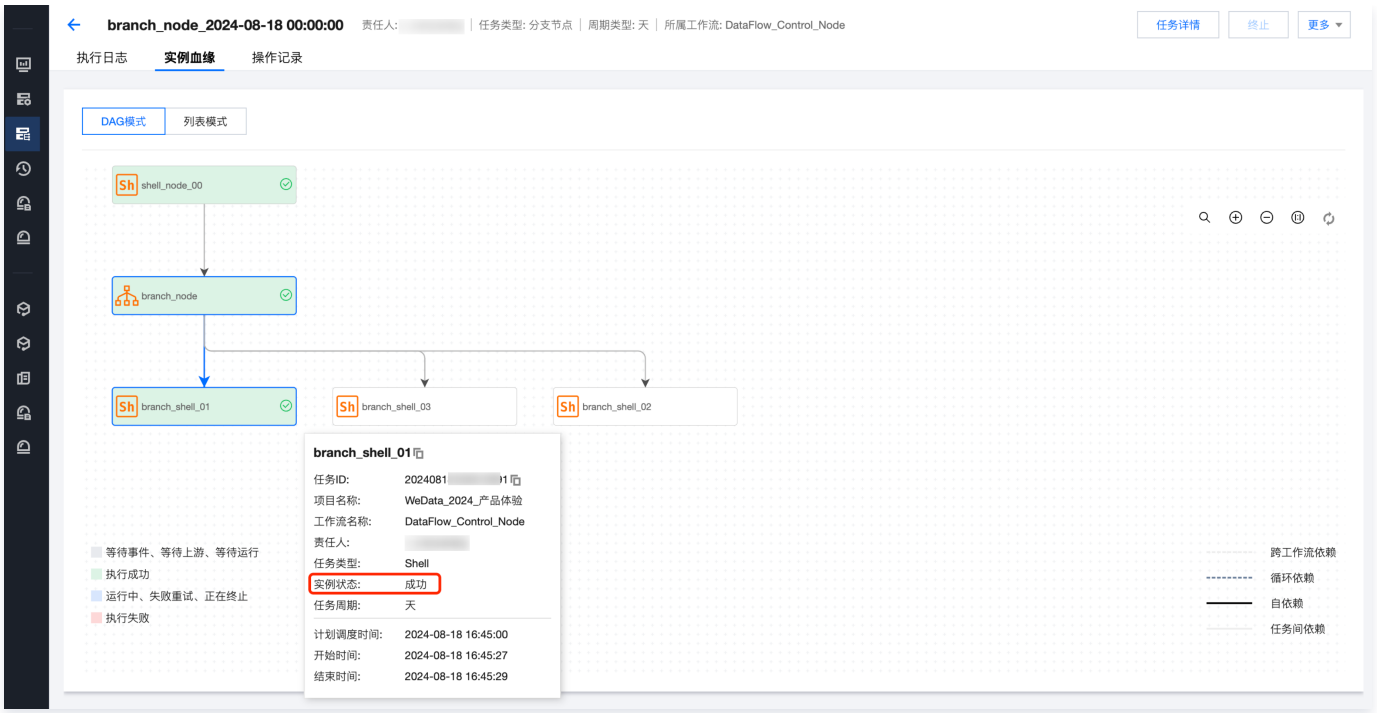
9.3 Right-click to view the execution log of task branch_shell_02. Since the condition is not met, branch_shell_02 is skipped. Similarly, task branch_shell_03 is also skipped due to unmet conditions.

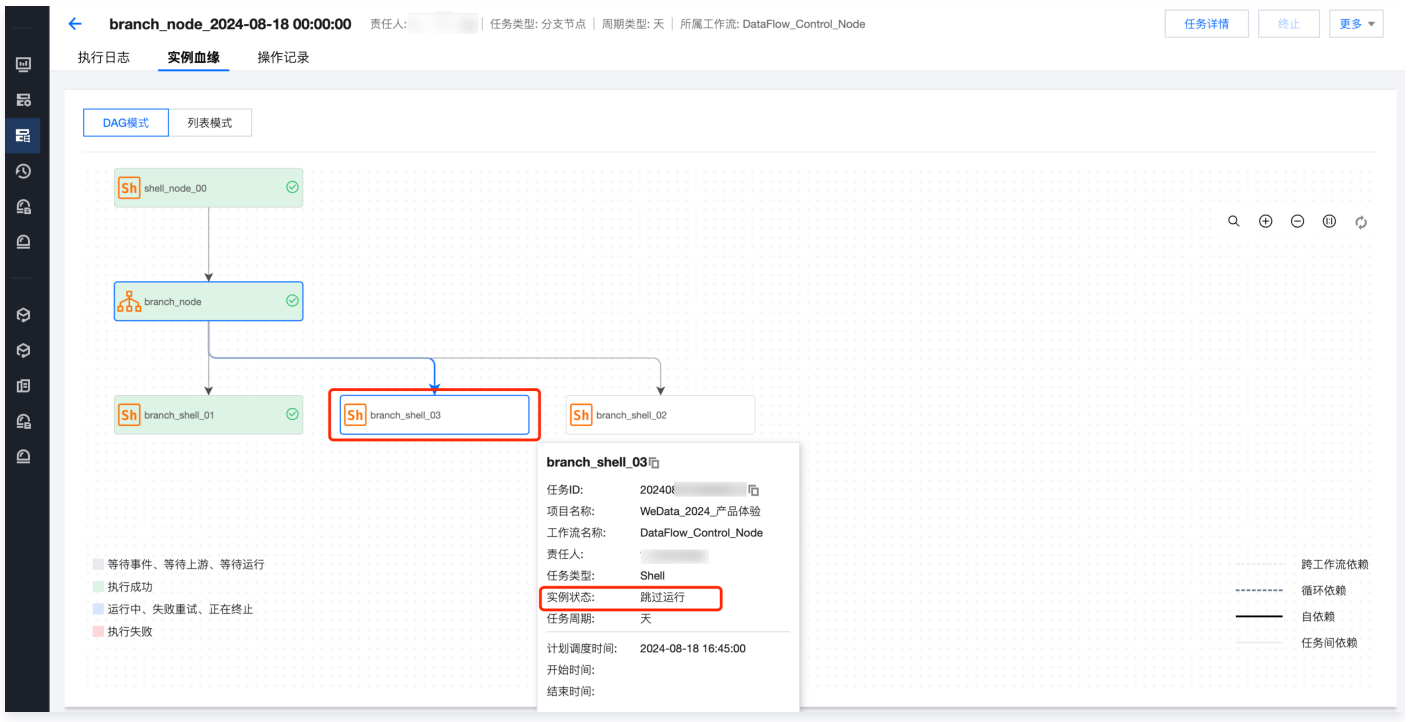
10. Submit for release, then check Cron Scheduling

10.1 Click the **Submit** button in the workflow toolbar to select and submit tasks in the current workflow. You can view the status of instances in the Cron Scheduling.

10.2 You can see branch_shell_01 is in a successful state, while branch_shell_03 is in a skipped state.

©2013-2025 Tencent Cloud. All rights reserved.





©2013–2025 Tencent Cloud. All rights reserved.

Merge Node

Last updated: 2024-11-25 14:52:32

A Merge Node is a logic control node provided by WeData Data Development. It can perform logical judgments on the status of upstream task instances and set merge conditions for the success of the Merge Node. It is mainly used to solve issues where multiple branch tasks of a Branch Node trigger an upstream task.

Use Limits

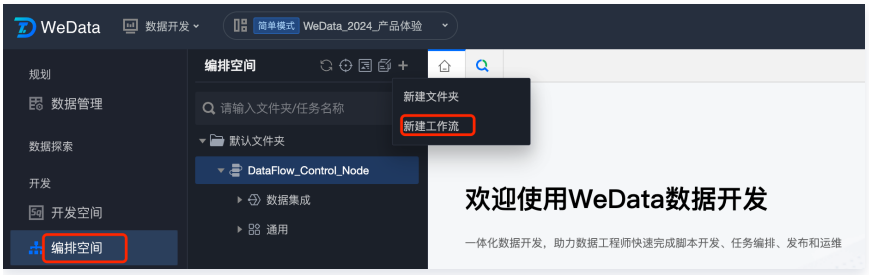
Note:

1. A Merge Node supports up to 50 merge conditions.
2. A Merge Node currently does not support cross-project cloning.
3. A Merge Node does not support the addition of upstream dependent tasks in the scheduling settings. It can only be configured on the Merge Node's details page.

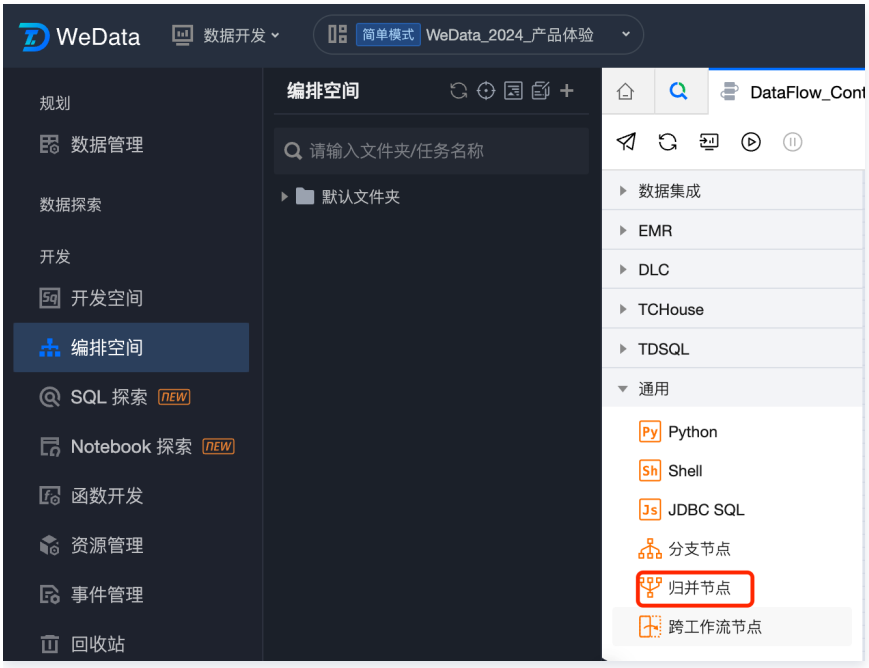
Using a Merge Node

Enter the WeData console's Project Management. In the project list, click **Data Development**. After entering the Data Development module, switch to the **Data Development** > **Orchestration Space** interface.

1. Create a new workflow in the upper-left corner of the Orchestration Space.

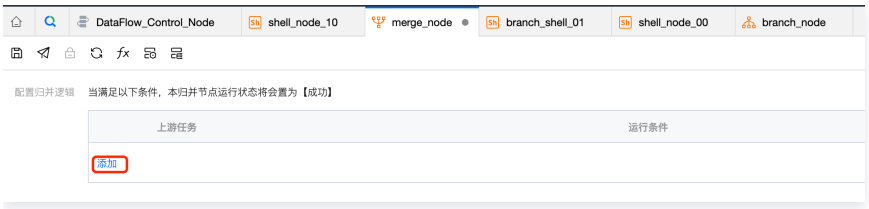


2. Double-click the workflow and then click Branch Node under Common to create a Merge Node.



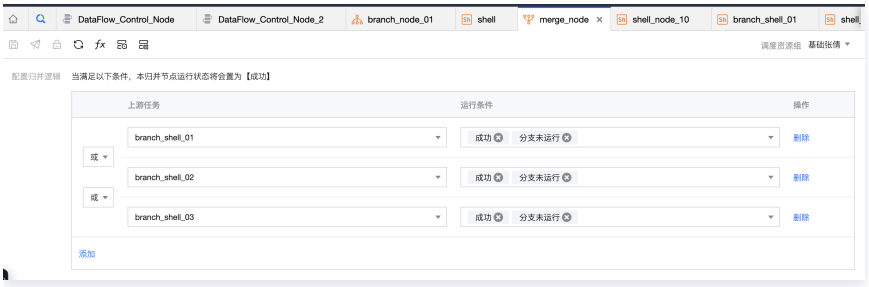
3. Enter the task name in the dialog box for creating a new Merge Node task, click **Confirm to create the task**, and enter the Merge Node configuration interface.
4. Merge Node configuration: click the **Add** button under the upstream tasks of the Merge Node to add merge conditions.

©2013-2025 Tencent Cloud. All rights reserved.



Parameter name	Parameter description
Upstream Task	The upstream tasks that need to be judged by the Merge Node. Selection Restriction: **Note:** • Duplicate selection of already chosen upstream tasks is not supported. • Direct and indirect downstream tasks of a merge node cannot be selected. • Branch nodes cannot be selected
Running Conditions	When a merge node is successfully set, the final status of the upstream task instance needs to be determined. The default is: Success or Branch Not Executed. • Success: The upstream task instance ran successfully. • Branch Not Executed: The upstream task instance did not meet the branch node condition and was skipped. • Failure: The upstream task instance failed to run.
Relationship Between Running Conditions	The relationship between multiple upstream task running state conditions defaults to OR, but can be set to AND.

For example, when one of the upstream task instances branch_shell_01, branch_shell_02, or branch_shell_03 runs to its final state and is in a success state, set the current merge node to a successful state.

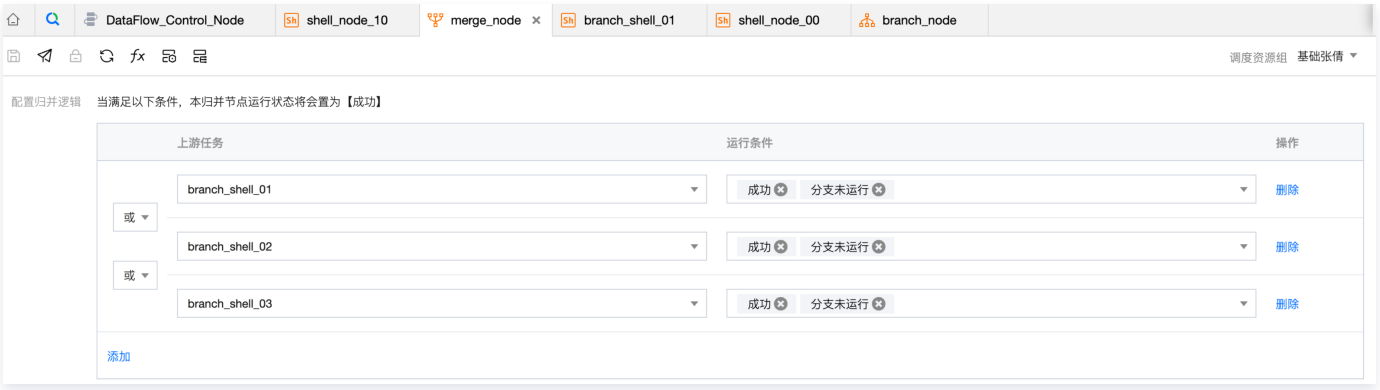


5. Other tasks can be configured downstream of the merge node to be triggered after a successful merge.
6. Click **Submit** on the workflow toolbar to select and submit tasks in the current workflow. You can check the instance status in the periodic scheduling.

Configuration Samples

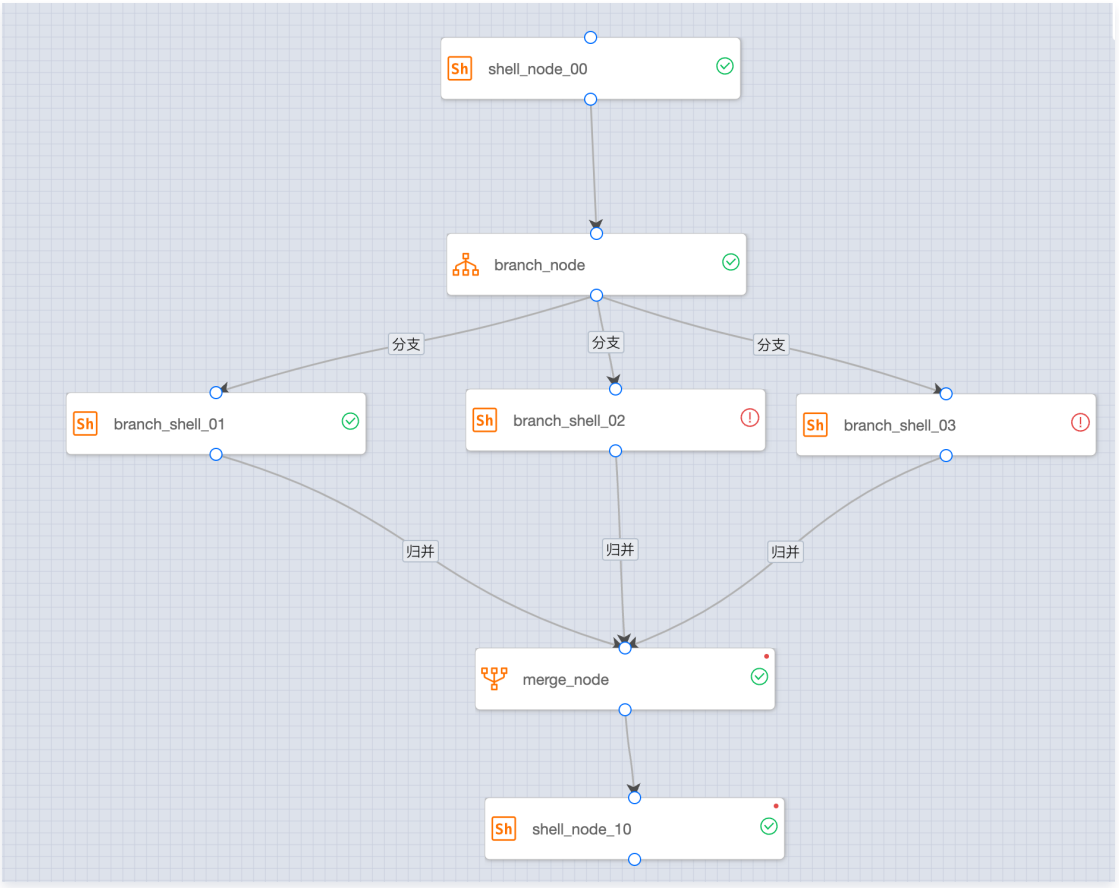
1. Continue configuring based on the Branch Node Configuration Sample in the DataFlow_Control_Node workflow.
2. In the DataFlow_Control_Node workflow, create a merge node: merge_node.
3. Select the upstream tasks for the merge node: branch_shell_01, branch_shell_02, branch_shell_03, and the running conditions as defaults: Success, Branch Not Executed. Set the relationship between multiple upstream task conditions to OR. After configuration, **Save** the tasks.

©2013–2025 Tencent Cloud. All rights reserved.



4. For easier testing, configure a shell task node downstream that depends on merge_node.

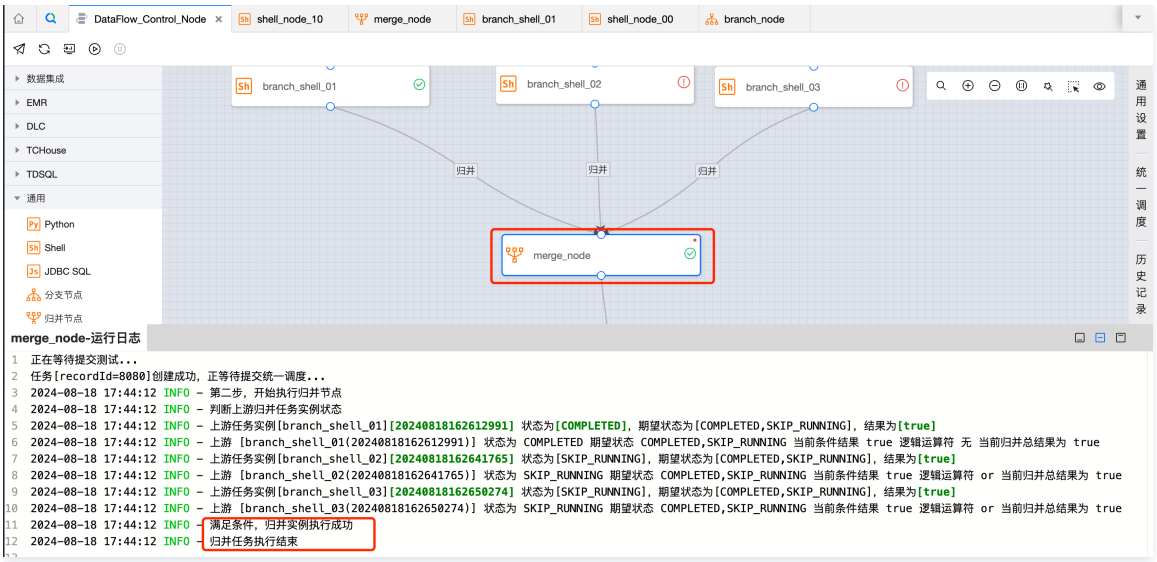
   The final configuration will look like this:



5. Workflow test run.

Click Workflow Test Run on the workflow toolbar to test the entire workflow. After the merge node has completed, check the merge node's run logs. If the merge node determines that the upstream tasks meet the conditions, the merge node is set to **Success**.

©2013–2025 Tencent Cloud. All rights reserved.



6. Submit the workflow to the schedule.

Click **Submit** on the workflow toolbar to select and submit the newly added merge node and downstream node tasks in the current workflow. You can check the instance status in the periodic scheduling.

©2013–2025 Tencent Cloud. All rights reserved.

JDBC SQL

Last updated：2024-11-25 14:52:54

Feature Overview

Submit SQL commands to the WeData Workflow Scheduling System for scheduling and execution.



Parameter description

Parameter	Description
Data source type	Currently, ORACLE, MYSQL, DORIS, CLICKHOUSE, TBASE, and STARROCKS types are supported.
Data Source	The data source information required for executing SQL tasks must be created first in **Data Management** > **Data Source Management**.
Scheduling Resource Group	The execution resources for running scheduled tasks.
SQL Editing	Edit SQL code, supporting save, submit, test run, version management, etc.

©2013-2025 Tencent Cloud. All rights reserved.

Shell

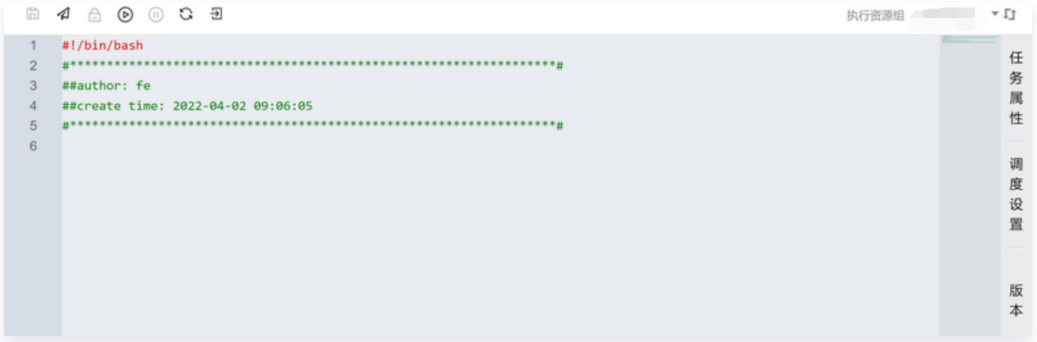
Last updated: 2024-08-19 10:49:11

> **Note:**
> Currently, to execute Hive and Spark commands in EMR using Shell scripts, you need to export corresponding cluster parameters first.

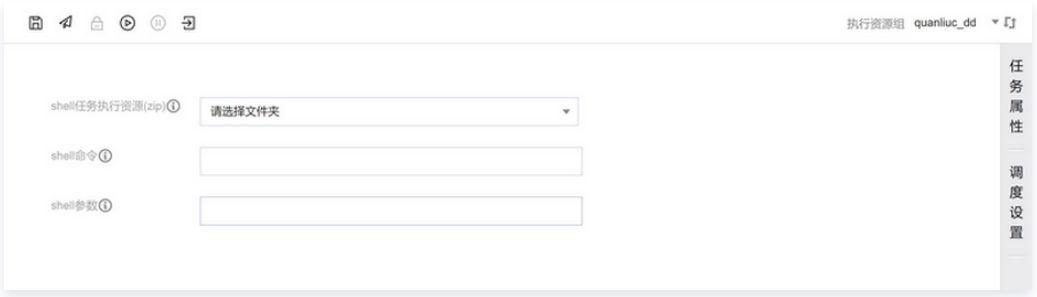
Development Mode

Shell tasks support both **Code Mode** and **Script Mode** for editing. You can switch between the two modes by clicking .

- Code Mode:



- Script Mode:



> **Note**
> Runtime Parameters: If there are variable parameters in the SQL, please assign values, such as key0=123. Separate multiple parameters with an English semicolon ";".

Examples

This article will introduce how to use a Shell script to determine if a specified path on Tencent Cloud COS (COS) contains a file. If it exists, it prints a message and returns exit code 0. If it does not exist, it returns exit code 1.

Step 1: Generate and Save the Key

1. Generate the access keys SecretID and SecretKey for Tencent Cloud COS storage space. For details, please refer to Tencent Cloud Certificate Authority M .
2. Save to a directory on the execution machine. Here it is saved as /etc/config/cos.conf.

```
SECRET_ID=AKID**************UQ8n
SECRET_KEY=kGYf**************rPkt
```

Step 2: Create a Shell Task

1. In the WeData Orchestration Space, create a Shell task and enter the following code, divided into the following modules:
   1.1 Definition of the required information to access COS, including Bucket name, object key, region, key ID, and key key:

©2013-2025 Tencent Cloud. All rights reserved.

```
bucket="anny-test-1315051789"
key="aa1/b.txt"
region="ap-beijing"
# Include /etc/config/cos.conf
. /etc/config/cos.conf
secretKey=$SECRET_ID
secretId=$SECRET_KEY
```

1.2 Key supports time variable substitution, for example:

```
key="aa1/b.txt${yyyyMMdd}"
```

1.3 Details click to view the built-in parameter description document, as shown below:



2. **Construct access link**

Use variables to construct the COS access link endpoint:

```
endpoint="https://$bucket.cos.$region"
```

3. **Generate authentication signature, no modification needed here**

Invoke the OpenSSL tool to generate an authentication signature based on access link, region, timestamp, and other information:

```
generate_auth() {
    start=$(date +%s)
    end=$((start + 600))
    qKeyTimeStr="$start;$end"
    formatStr="head\n$key\n\nhost=$endpoint\n"
    hashFormatStr=$(echo -n "$formatStr" | openssl dgst -sha1 -hex | awk '{print $2}')
    signKey=$(echo -n "$qKeyTimeStr" | openssl dgst -sha1 -hmac "$secretKey" -hex)
    stringToSign="sha1\nqKeyTimeStr\n$hashFormatStr\n"
    signature=$(echo -n "$stringToSign" | openssl dgst -sha1 -hmac "$signKey" -hex)
    auth=$(echo "q-sign-algorithm=sha1&q-ak=${secretId}&q-sign-time=${qKeyTimeStr}&q-key-
time=${qKeyTimeStr}&q-header-list=head&q-url-param-list=&q-signature=${signature}")
    echo $auth
}
```

4. **Send the request**

Use curl to send a GET request to check if the response code is 200:

```
response=$(curl -s -w "%{http_code}" -o /dev/null -H "Authorization: $auth" $endpoint/$key)

# Check response code
if [ $response -eq 200 ]; then
  echo "File exists"
  exit 0
else
  exit 1
fi
```

©2013–2025 Tencent Cloud. All rights reserved.

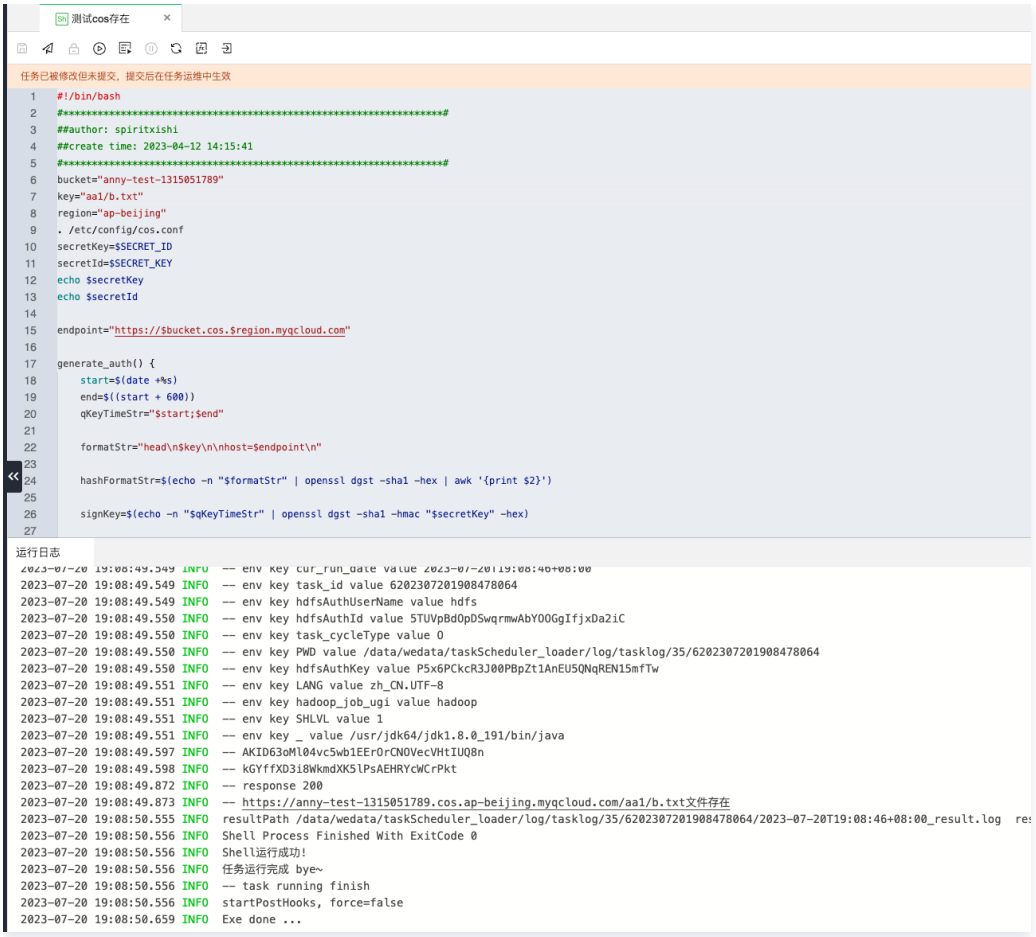
5. This way, when the file exists, it will print "File exists" and return 0. If it does not exist, it returns 1, completing the existence judgment.

Complete Shell script as follows:

```
# Define parameters
bucket="anny-test-1315051789"
key="aa1/b.txt${yyyyMMdd}"
region="ap-beijing"
. /etc/config/cos.conf
secretKey=$SECRET_ID
secretId=$SECRET_KEY
echo $secretKey
echo $secretId

# Concatenate URL
endpoint="https://$bucket.cos.$region.myqcloud.com"

# Define signature parameters and generate
generate_auth() {
    start=$(date +%s)
    end=$((start + 600))
    qKeyTimeStr="$start;$end"
    formatStr="head\n$key\n\nhost=$endpoint\n"
    hashFormatStr=$(echo -n "$formatStr" | openssl dgst -sha1 -hex | awk '{print $2}')
    signKey=$(echo -n "$qKeyTimeStr" | openssl dgst -sha1 -hmac "$secretKey" -hex)
    stringToSign="sha1\nqKeyTimeStr\n$hashFormatStr\n"
    signature=$(echo -n "$stringToSign" | openssl dgst -sha1 -hmac "$signKey" -hex)
    auth=$(echo "q-sign-algorithm=sha1&q-ak=${secretId}&q-sign-time=${qKeyTimeStr}&q-key-
time=${qKeyTimeStr}&q-header-list=head&q-url-param-list=&q-signature=${signature}")
    echo $auth
}
auth=$(generate_auth)

# Send the request
response=$(curl -s -w "%{http_code}" -o /dev/null -H "Authorization: $auth" $endpoint/$key)
echo response $response
if [ $response -eq 200 ]; then
  echo $endpoint/$key File exists
  exit 0
else
  echo $endpoint/$key File does not exist
  exit 1
fi
```

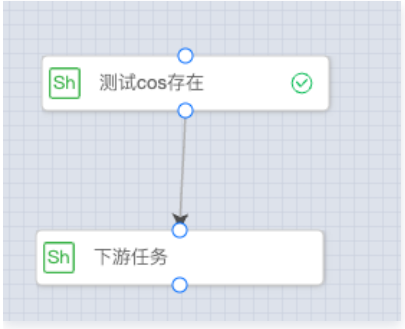
6. Running Log:

©2013–2025 Tencent Cloud. All rights reserved.



Step 3: Create downstream tasks and depend on Shell

Create downstream tasks and depend on Shell, because when the file exists, Shell returns 0, indicating success. Therefore, the downstream can start execution when the file exists.



Reference resources

Support referencing local resources in Shell tasks to execute commands, such as kjb, ktr files. For details, refer to Resource Management.

©2013–2025 Tencent Cloud. All rights reserved.

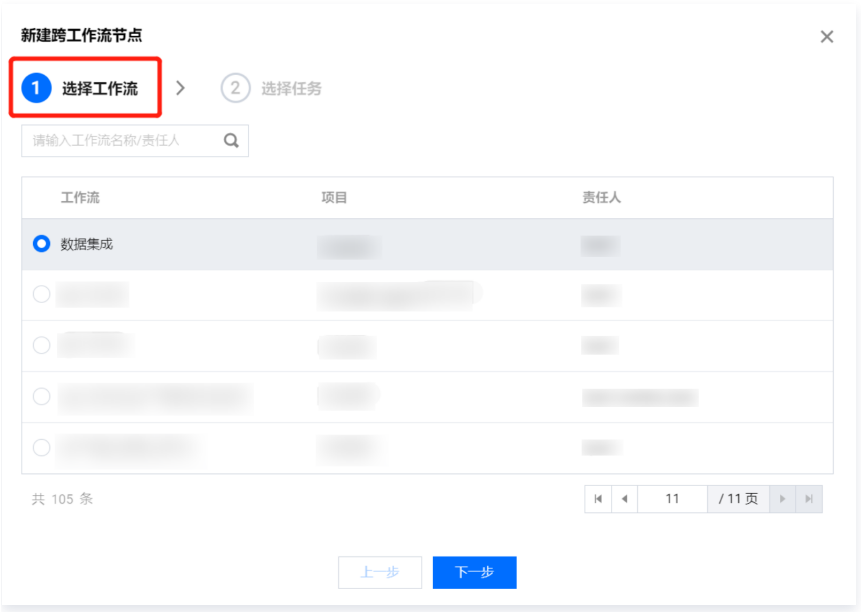
Other nodes
Cross-workflow node

Last updated: 2024-11-25 14:53:17

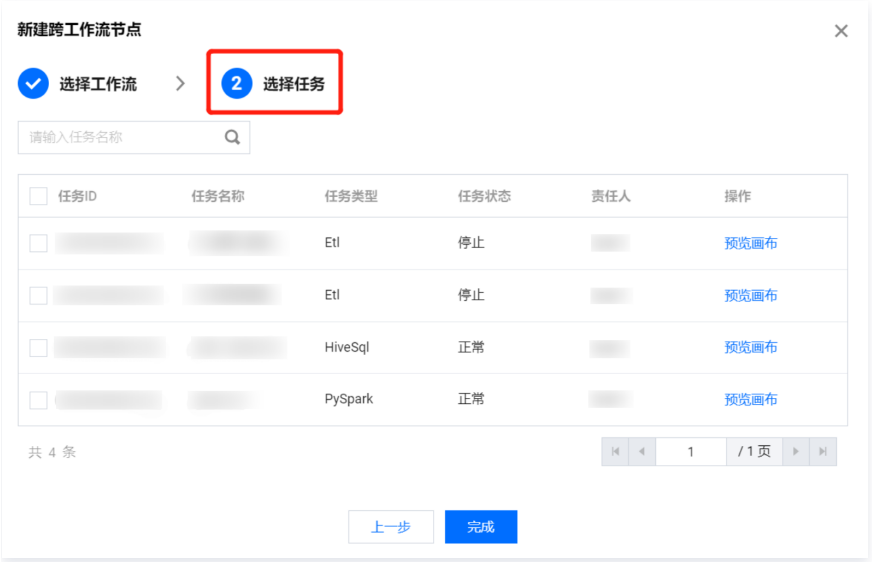
In the pop-up window, select the tasks from other workflows that need to be added to the current workflow. First, choose the other workflows, then select the development tasks within those workflows to be added as cross-workflow task nodes in the current workflow. Add them to the current workflow and finally click **Complete** to finish node configuration.

Operation step

- Select the workflow where the cross-workflow task node is located:



- Select the task node to be used as the cross-workflow task node:



Note
Cross-workflow tasks can only serve as parent nodes.

©2013-2025 Tencent Cloud. All rights reserved.

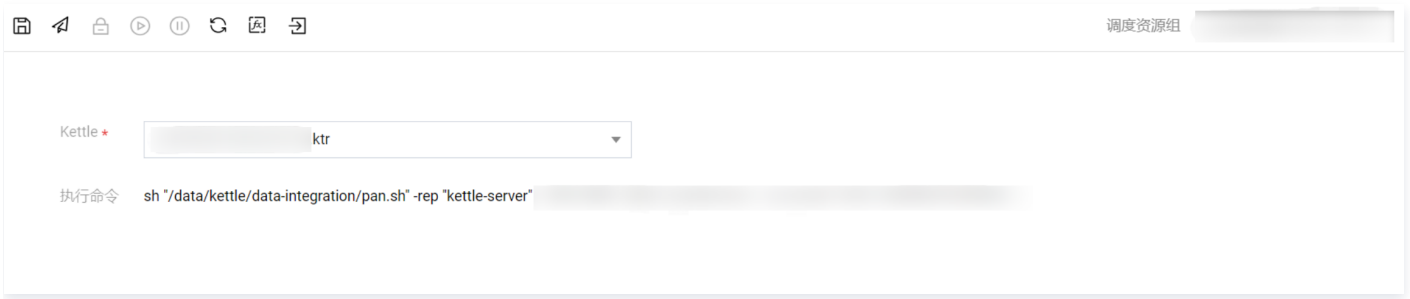
Kettle

Last updated: 2024-11-25 14:53:38

> **Note:**
> You need to enable Integrate Kettle in Advanced Configuration under Basic Information Configuration of Project Management and configure the resource library.

Feature Overview

Submit SQL commands to the WeData Workflow Scheduling System for scheduling and execution.



Parameter description

Parameter	Description
Kettle	Find the Kettle resource file in the resource path.
Command	Automatically generate the command based on the selected file type. Currently, this is only for display to help users understand the specific command. Modification is not supported.

©2013-2025 Tencent Cloud. All rights reserved.

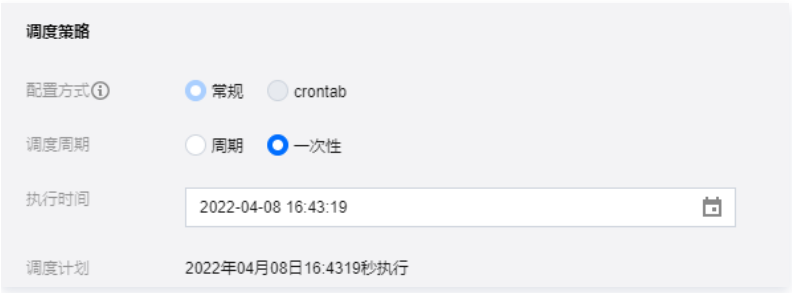
Task Scheduling Configuration
Basic Configuration

Last updated：2024-12-12 20:13:49

Scheduling Period Configuration

The scheduling policy supports regular periodic scheduling or one-time execution policies.

One-time scheduling



Parameter description:

Parameters	Description
Scheduling cycle	The execution cycle unit of task scheduling supports minutes, hours, days, weeks, months, years, and one-time scheduling.
Execution Time	One-time tasks can be specified down to the year, month, day, hour, minute, and second.
Scheduling Plan	Automatically generated according to the configured cycle and execution time.

Note
One-time tasks cannot generate new instances through resubmission after completion; a new task must be created for repeat execution.

Minute Scheduling



Parameter description:

Parameters	Description

Configuration Method	Single task configuration defaults to regular and cannot be modified.
Scheduling cycle	The execution cycle unit of task scheduling supports minutes, hours, days, weeks, months, years, and one-time scheduling.
Effective Time	The effective time period for scheduling configuration. The system will automatically schedule within this time range according to the time configuration. Beyond the validity period, it will not be automatically scheduled. The start date can only be selected from the current date onwards, defaulting to show the current date to 2099-12-31.
Execution Time	The execution time of minute tasks can only be chosen on the hour, with interval ranges of 5, 10, 15, 20, 30 minutes. Minute tasks are scheduled by generating instances based on execution time.
Scheduling Plan	Automatically generated according to the configured cycle and execution time.

Hourly Scheduling



Parameter description:

Parameters	Description
Configuration Method	Single task configuration defaults to regular and cannot be modified.
Scheduling cycle	The execution cycle unit of task scheduling supports minutes, hours, days, weeks, months, years, and one-time scheduling.
Effective Time	The effective time period for scheduling configuration. The system will automatically schedule tasks within this time range according to the time configuration. It will not schedule automatically after the valid period. The start date must be selected.
Execution time (choose either specified time or execution time)	• Execution Time For hourly tasks, the execution time can be set in hours and minutes with an interval range of 1 – 23 hours. Tasks generate instances based on the execution time for scheduling. • Interval time Controls the frequency of task instance generation within the specified execution time interval, using the set interval time as the unit.
Specified time (choose either execution time or specified time)	• Specified time Specify a time range from 0 – 23 hours, and multi-selection is supported. Hourly tasks generate and execute instances based on the specified time. • Specified minutes Specify the exact minute for instance generation after the specified time.

©2013-2025 Tencent Cloud. All rights reserved.

Scheduling Plan	Automatically generated according to the configured cycle and execution time.

Daily scheduling



Parameter description:

Parameters	Description
Configuration Method	Supports conventional and crontab configuration methods. A single task is configured by default to be conventional and cannot be modified.
Scheduling cycle	The execution cycle unit of task scheduling supports minutes, hours, days, weeks, months, years, and one-time scheduling.
Effective Time	The effective time period for scheduling configuration. The system will automatically schedule within this time range according to the time configuration. Beyond the validity period, it will not be automatically scheduled. The start date can only be selected from the current date onwards, defaulting to show the current date to 2099-12-31.
Execution Time	For daily tasks, the execution time can be set in hours and minutes. Instances begin execution at the set execution time.
Scheduling Plan	Automatically generated according to the configured cycle and execution time.

Weekly scheduling



Parameter description:

©2013-2025 Tencent Cloud. All rights reserved.

Parameters	Description
Configuration Method	Supports conventional and crontab configuration methods. A single task is configured by default to be conventional and cannot be modified.
Scheduling cycle	The execution cycle unit of task scheduling supports minutes, hours, days, weeks, months, years, and one-time scheduling.
Effective Time	The effective time period for scheduling configuration. The system will automatically schedule within this time range according to the time configuration. Beyond the validity period, it will not be automatically scheduled. The start date can only be selected from the current date onwards, defaulting to show the current date to 2099-12-31.
Specified Date	The specified date range for weekly tasks is Monday to Sunday, and multiple selections are supported.
Execution Time	For weekly tasks, the execution time can be set in hours and minutes. Instances begin execution at the specified time on the specified date.
Scheduling Plan	Automatically generated according to the configured cycle and execution time.

Monthly Scheduling



Parameter description:

Parameters	Description
Configuration Method	Supports conventional and crontab configuration methods. A single task is configured by default to be conventional and cannot be modified.
Scheduling cycle	The execution cycle unit of task scheduling supports minutes, hours, days, weeks, months, years, and one-time scheduling.
Effective Time	The effective time period for scheduling configuration. The system will automatically schedule within this time range according to the time configuration. Beyond the validity period, it will not be automatically scheduled. The start date can only be selected from the current date onwards, defaulting to show the current date to 2099-12-31.
Specified Date	The specified date range for monthly tasks is the 1st to the 31st of each month, and multiple selections are supported. Note: The end-of-month option is mutually exclusive with other dates.
Execution Time	For monthly tasks, the execution time can be set in hours and minutes. Instances begin execution at the specified time on the specified date.
Scheduling Plan	Automatically generated according to the configured cycle and execution time.

©2013-2025 Tencent Cloud. All rights reserved.

Annual Scheduling



Parameter description:

Parameters	Description
Configuration Method	Supports conventional and crontab configuration methods. A single task is configured by default to be conventional and cannot be modified.
Scheduling cycle	The execution cycle unit of task scheduling supports minutes, hours, days, weeks, months, years, and one-time scheduling.
Effective Time	The effective time period for scheduling configuration. The system will automatically schedule within this time range according to the time configuration. Beyond the validity period, it will not be automatically scheduled. The start date can only be selected from the current date onwards, defaulting to show the current date to 2099-12-31.
Specified Month	The specified month range for annual tasks is January to December, and multiple selections are supported.
Specified Date	The specified date range for annual tasks is the 1st to the 31st, and multiple selections are supported. Note: The end-of-month option is mutually exclusive with other dates.
Execution Time	For annual tasks, the execution time can be set in hours and minutes. Instances begin execution at the specified time on the specified date.
Scheduling Plan	Automatically generated according to the configured cycle and execution time.

Scheduling Type

Support selecting normal scheduling and dry-run scheduling. Dry-run scheduling means that when the task is scheduled, the system will directly mark the instance as successful without actually executing the task. Normal scheduling executes the task code as usual.

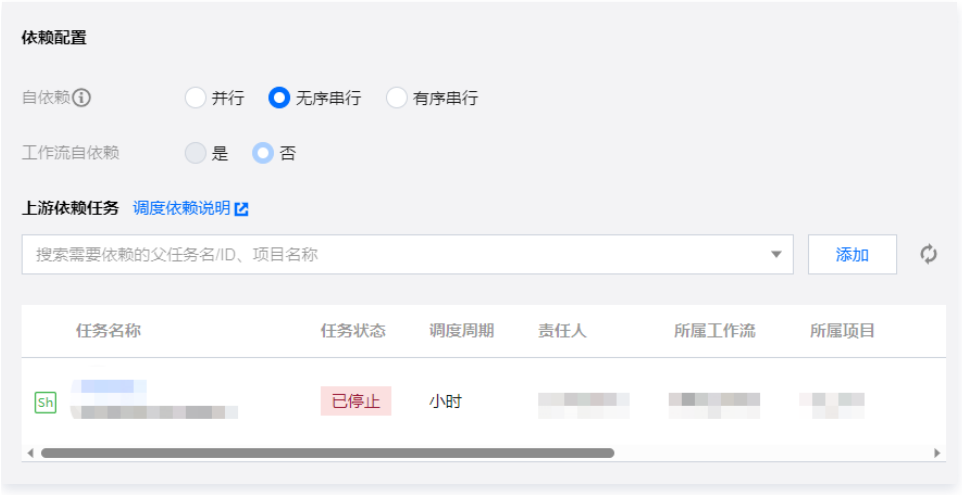
©2013-2025 Tencent Cloud. All rights reserved.



Scheduling Dependency Configuration

Dependency Configuration

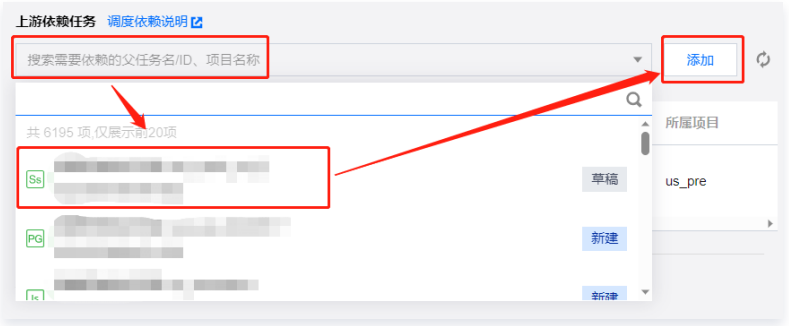
Dependency configuration supports setting upstream and downstream dependencies and execution order for task nodes and task instances.



Parameter description:

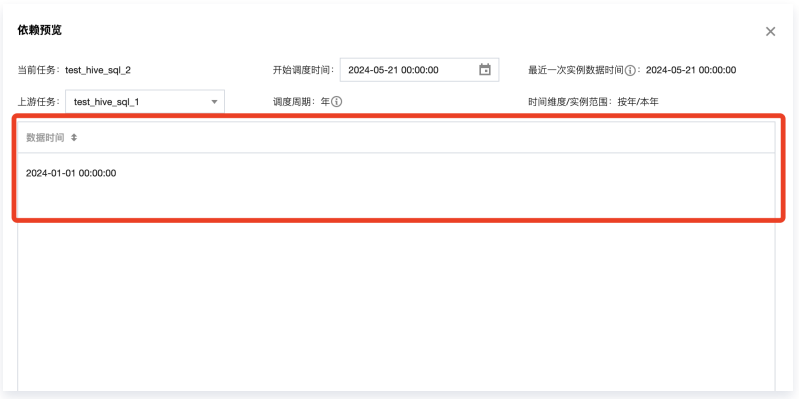
Information		Description
Self-dependency (Self-dependency refers to the dependency relationship between different instances of the same task)	Parallel	There is no dependency between the previous periodic instance and the next periodic instance. If a task has multiple instances, multiple instances will run simultaneously.
	Unordered Serial	There is no dependency between the current instance and the previous periodic instance. If a task has multiple instances, the system randomly selects one instance to run. Only one instance is in running status at a time.
	Ordered Serial	The current instance depends on the status of the previous periodic instance.
Workflow Self-dependency		There is no self-dependency configuration in task scheduling.

©2013-2025 Tencent Cloud. All rights reserved.

Upstream Dependency Task	Upstream Dependency Task Search	Supports finding required upstream dependency tasks by searching task name, task ID, or project name. 

	Upstream Dependency Task List	When you add a searched task as a dependency task, the corresponding task information, including task name, task status, scheduling period, person in charge, workflow, project, time dimension, instance range, execution strategy, and delete operation, will be displayed in the dependency task list.
		 
		The parameters are described as follows:
		• Task Name: The task name, task ID, and task type of the dependency task.
		• Task Status: The scheduling and running status of the dependency task.
		• Scheduling Cycle: Based on the dependent task's scheduling cycle.
		• Responsible Person: Based on the responsible person of the dependent task.
		• Belonging Workflow: The data workflow to which the dependent task belongs.
		• Belonging Project: The project to which the dependent task belongs.
		• Configuration Method: Recommended Strategy: Includes commonly used dependency configuration methods; Custom Definition: Users may define dependency configurations as needed.
		• Time Dimension/Instance Scope
		  ○ If the recommended strategy is chosen: select the dependency method and scope, available options vary with the task cycle. See the table below for detailed configuration items: Scheduling Dependency Explanation.
		  ○ If Custom Definition is chosen:
		  ○ Interval Mode: The input format is x,y. It indicates the range of data time offsets for dependent upstream task instances. For example, in the case of an interval (days), entering −10,−1 means depending on instances from the upstream task from 10 to 1 days prior, as a closed interval. x<=y, and the maximum difference can be 1000.
		  ○ List Mode: The input format is x,y,z. It represents specific data time offsets for dependent upstream task instances. For example, in the case of a list (days), entering −3,−2,−1 means depending on instances from the upstream task 3, 2, 1 days prior. x<y<z, and up to 50 specific time offsets can be entered.
		• Execution Strategy: Handling strategy when the dependent upstream task instance does not exist, the default is to wait for the upstream; waiting for the upstream means continuously waiting when the upstream task instance does not exist, normal execution means executing normally when the upstream task instance does not exist.
		• Operation
		  ○ Dependency Preview: Supports previewing the dependency results based on the configured dependency relationship to help users determine if the configured dependencies meet expectations. Users can select the start scheduling time to view the instance data time dependency instance list (instance data time).

©2013–2025 Tencent Cloud. All rights reserved.



- Delete: Deletion must be submitted to take effect in production.

Dependent Downstream Task	A dependent downstream task means that the current task can depend on any downstream task (including indirect downstream) to satisfy some unique dependency scenarios. For example: Task A is an upstream task of Task B. B1 cycle instance depends on A1 cycle instance, and A1 cycle instance wants to depend on B2 cycle instance, which means A depends on B, and B depends on A. You can configure Task B in Task A’s Scheduling Configuration – Dependencies on Upstream Tasks. Tasks can have cyclic dependency relationships, but instances cannot. Therefore, the system will regularly check the dependencies of instances. If there is a cyclic dependency issue after a task is launched, an alarm will be sent to you via email and SMS. 

Scheduling Dependency Explanation

Cycle	Upstream						
	Minute	Hour	Day	Week	Month	Year	One-time
Minute	(1) By Minute/Current Minute: Depends on instances with the same data timestamp. (2) By Minute/Previous Period: Depends on instances with the previous data timestamp.	(1) By Hour/Current Hour: All minute tasks within the hour need to wait for the hour task to be completed before running.	(1) By Day/Current: All minute tasks within the day need to wait for the daily task to be completed before running.	(1) By Week/This Week: All minute tasks within the week need to wait for the weekly task to be completed before running.	(1) By Month/This Month: All minute tasks within the month need to wait for the monthly task to be completed before running.	(1) By Year/This Year: All minute tasks within the year need to wait for the annual task to be completed before running.	All minute instances depend on one-time task instances and are independent of the data timestamp.
Hour	(1) By Hour/Previo	(1) By Hour/Recen	(1) By Day/Today:	(1) By Week/This	(1) By Month/This	(1) By Year/This	All hour instances

©2013–2025 Tencent Cloud. All rights reserved.

	us Hour (−60,0]: Depend on all upstream instances within the previous hour (−60,0]. (2) By Hour/Previous Hour: Depend on all upstream instances within the previous hour [−60,0). (3) By Hour/Current Hour: Depend on all upstream minute instances within the current hour.	t Instance: Depend on instances with the most recent data timestamp. Priority search within [0 AM today, current time], if not found then search within [current time, 0 AM next day).	All hourly tasks within the day need to wait for the daily task to be completed before running.	Week: All hourly tasks within the week need to wait for the weekly task to be completed before running.	Month: All hourly tasks within the month need to wait for the monthly task to be completed before running.	Year: All hourly tasks within the year need to wait for the annual task to be completed before running.	depend on one−time task instances and are independent of the data timestamp.
Day	(1) By Day/Previous Day (−24*60,0]: Depend on all upstream instances within the previous day (−24*60,0]. (2) By Day/Previous Day: Depend on all upstream instances within the previous day [−24*60,0). (3) By Day/Today: Depend on all upstream minute instances within the	(1) By Day/Previous Day (−24*60,0]: (2) By Day/Previous Day: (3) By Day/Today: Depend on all upstream hour instances within the current day.	(1) By Day/Today: Depend on upstream daily instances of the current day.	(1) By Week/This Week: All daily tasks within the week need to wait for the weekly task to be completed before running. (2) By Day/Most Recent Data Timestamp Instance: Depend on instances with the most recent data timestamp before the current time.	(1) By Month/This Month: All daily tasks within the month need to wait for the monthly task to be completed before running. (2) By Day/Most Recent Data Timestamp Instance: Depend on instances with the most recent data timestamp before the current time.	(1) By Year/This Year: All daily tasks within the year need to wait for the annual task to be completed before running.	All daily instances depend on one−time task instances and are independent of the data timestamp.

©2013−2025 Tencent Cloud. All rights reserved.

	current day.						
Week	(1) By Week/Last Week: Depend on all minute task instances from last week. (2) By Week/This Week: Depend on all minute task instances from this week.	(1) By Week/Last Week: Depend on all hourly task instances from last week. (2) By Week/This Week: Depend on all hourly task instances from this week.	(1) By Week/Last Week: Depend on all daily task instances from last week. (2) By Week/Last Friday: Depend on daily task instances from last Friday. (3) By Week/Last Sunday: Depend on daily task instances from last Sunday. (4) By Week/This Week: Depend on all daily task instances from this week. (5) By Day/Today: Depend on daily task instances from today. (6) By Day/Previous Day: Depend on daily task instances from the previous day.	(1) By Week/This Week: Depend on all daily task instances from this week. (2) By Day/Most Recent Data Timestamp Instance: Depend on instances with the most recent data timestamp before the current time.	(1) By Month/This Month: All weekly tasks within the month need to wait for the monthly task to be completed before running. (2) By Day/Most Recent Data Timestamp Instance: Depend on instances with the most recent data timestamp before the current time.	(1) By Year/This Year: All weekly tasks within the year need to wait for the annual task to be completed before running.	All weekly instances depend on one-time task instances and are independent of the data timestamp.
Month	(1) By Month/Last Month: Depend on all minute task instances from last month. (2) By Month/This	(1) By Month/Last Month: Depend on all hourly task instances from last month. (2) By Month/This	(1) By Month/Last Month: Depend on all daily task instances from last month. (2) By Month/End	(1) By Month/Last Month: Depend on all weekly task instances from last month. (2) By Month/This	(1) By Month/This Month: Depend on all monthly task instances from this month. (2) By Day/Most	(1) By Year/This Year: All tasks within the year need to wait for the annual task to be completed	All monthly instances depend on one-time task instances and are independent of the data timestamp.

©2013-2025 Tencent Cloud. All rights reserved.

	Month: Depend on all minute task instances from this month.	Month: Depend on all hourly task instances from this month.	of Last Month: Depend on the instances from the last day of last month. (3) By Month/This Month: Depend on all daily task instances from this month. (4) By Day/Today: Depend on daily task instances from today. (5) By Day/Previous Day: Depend on daily task instances from the previous day.	Month: Depend on all weekly task instances from this month. (3) By Day/Most Recent Data Timestamp Instance: Depend on instances with the most recent data timestamp before the current time.	Recent Data Timestamp Instance: Depend on instances with the most recent data timestamp before the current time.	before running.	
Year	(1) By Year/This Year: Depend on all minute task instances from this year.	(1) By Year/This Year: Depend on all hourly task instances from this year.	(1) By Day/All Days This Year: Depend on all daily task instances from this year. (2) By Day/Today: Depend on daily instances of the same data timestamp. (3) By Day/Previous Day: Depend on instances of the previous day's data timestamp.	(1) By Year/This Year: Depend on all weekly task instances from this year.	(1) By Month/All Months This Year: Depend on all monthly task instances from this year. (2) By Month/This Month: Depend on all monthly task instances from this month. (3) By Month/Last Month: Depend on all monthly task instances from last month. (4) By Month/End of Last	(1) By Year/This Year: Depend on all upstream annual task instances from this year.	All annual instances depend on one-time task instances and are independent of the data timestamp.

©2013–2025 Tencent Cloud. All rights reserved.

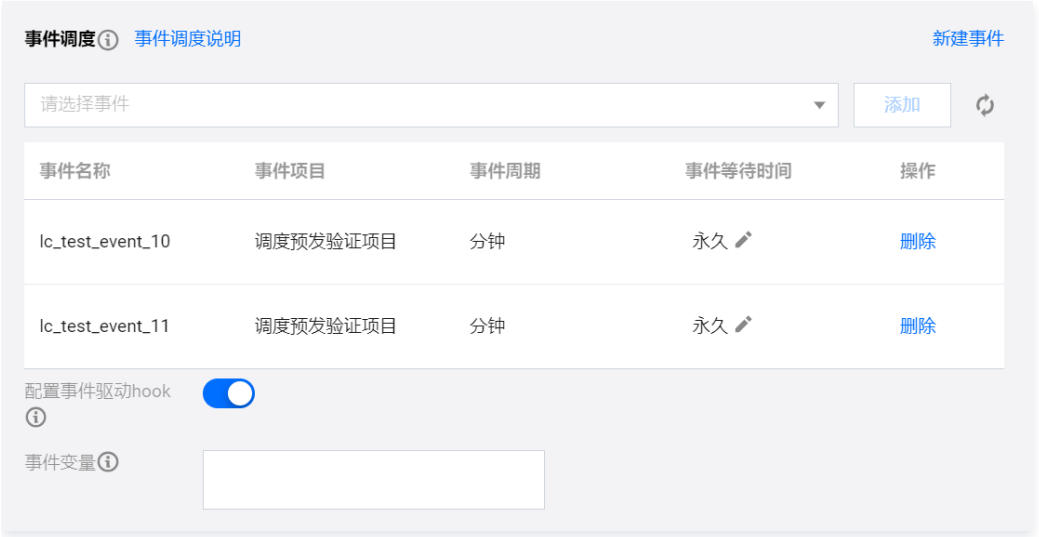
					Month: Depend on the instances from the last day of last month. (5) By Month/Beginning of Last Month: Depend on the instances from the first day of last month.		
One-time	(1) By Day/Today: Depend on all minute task instances within the date of the specified time.	(1) By Day/Today: Cross-cycle Default: Depend on all hourly task instances within the date of the specified time.	(1) By Day/Today: Depend on daily task instances within the date of the specified time.	(1) By Week/This Week: One-time tasks within the week need to wait for the weekly task to be completed before running.	(1) By Month/This Month: One-time tasks within the month need to wait for the monthly task to be completed before running.	(1) By Year/This Year: One-time tasks within the year need to wait for the annual task to be completed before running.	Depend on one-time task instances and are independent of the data timestamp.

©2013-2025 Tencent Cloud. All rights reserved.

Event Scheduling

Last updated：2024-11-25 14:54:48

Data development supports the event trigger scheduling feature, suitable for scenarios where tasks depend on a certain event to trigger execution. It consists of trigger programs, trigger events, and listening tasks. First, based on the business definition, trigger events need to be defined under the project, then write a trigger program to send events. Once the task listens to the event arrival, it can run.
Supports selecting cross-project events, provided the current user has permissions for the projects. Events can depend across projects. For instance, when a task is completed in Project A, it triggers Event 1. Tasks in Project B can listen to Event 1, and then tasks in Project B can run.



Parameter description:

Parameter	Description
Creating Events	Navigate to the Event Management Interface.
Add Event	Select the listening events that the computing task needs to depend on. Support for adding multiple listening events to a single task.
Event Wait Time	The waiting time for the task to listen to events. If the waiting time is exceeded, the task instance will be automatically killed.
Configure Event-driven Hook	The event configuration for Shell tasks can use the event-driven hook, effective only for Shell task types. After enabling, events can be triggered within Shell scripts using the echo command. The trigger command format: echo [event_variable]="[event_name],[data_time];[event_name],[data_time];", multiple events are separated by ";"
Event Variable	Effective when event-driven Hook configuration is enabled. The variable name for the definition trigger command. For example, when filled as events, input in the shell script: echo events="EventA,20220713;EventB,20220713;", after the task runs successfully, it can trigger events EventA and EventB.

Note:
If a task depends on multiple events, then when the task instance runs, it needs to wait for all its dependent events to be triggered before the instance can determine the status by waiting for events.

©2013-2025 Tencent Cloud. All rights reserved.

Parameter Transfer

Last updated: 2025-01-09 20:57:51

Summary of Features

Used to support parameter passing between upstream and downstream. For example, pass the current task's calculation result as a parameter for a child task. Multiple results can be specified for passing by column number.
Supported Scope: Hive SQL, JDBC SQL, Python, Shell, Spark SQL, DLC SQL, Impala, TCHouse-P.

Parameter Configuration Description

1. Output Current Task Parameters

If you need to pass parameters from the current task to downstream, check Output Current Task Parameters in parameter passing and configure the parameter information.
Fill in the output parameters produced by this task and the corresponding values (supports variables and constants), for example: Parent Task A calculates three columns in the code as results: 123, 234, 1234. In this configuration, define an output parameter name mark_id, and fill in mark_id = $[0], which indicates the value of the first column in the first row of the calculation result.
Example: As shown, define the parameter name a, taking the first variable in the code; define the parameter name b, taking the second variable in the code.

2. Cite Parent Task Parameters

If the current task needs to cite parameters defined in the parent task, check Cite Parent Task Parameters in parameter passing. Fill in the introduced parameters defined in the parent task, choose the parent task's output parameters for the value (no options if no parent node), for example: Child Task B, any defined parameter name mark_id, choose value node A.mark_id, using ${mark_id} in the code will replace ${mark_id} with the string 123.
Example: As shown, define the parameter name a, taking the parameter a defined in the upstream parent task hivesql_1; define the parameter name b, taking the parameter b defined in the upstream parent task hivesql_1.



Usage Example

Divided into three categories: SQL, Shell, Python. The following are examples of usage for each type (passed constants as an example below, also supports passing variables).

Type	Task type	As Output (Upstream)		As Input (Downstream)	
		Parameter Configuration	Code	Parameter Configuration	Code

©2013-2025 Tencent Cloud. All rights reserved.

SQL Type	Hive SQL JDBC SQL Spark SQL DLC SQL Impala TCHouse-P		SELECT "This is the value of the parameter"		SELECT '${parameter}' AS parameter_value;
Shell Type	Shell		echo "This is the value of the parameter"		expr ${parameter}
Python Type	Python		print("This is the value of the parameter")		print('${parameter}')

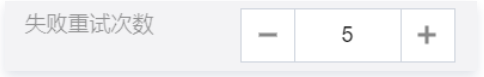
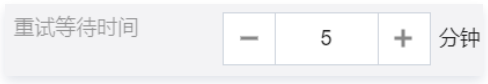
©2013-2025 Tencent Cloud. All rights reserved.

Advanced Settings

Last updated: 2024-11-25 14:55:33



Refer to the table below for information regarding advanced settings:

Information	Description
Task Priority	 Single selection. When multiple tasks exist, execute order is determined by priority. Default is low priority.
Retry Attempts	 Maximum number of retries if a task instance fails, up to 5 retries supported.
Retry Wait Time	 Wait time before retrying after a task instance fails.
Timeout Handling Policy	 Runtime duration timeout: Calculated from the start time of the task instance, if it exceeds the duration, the task will automatically terminate.

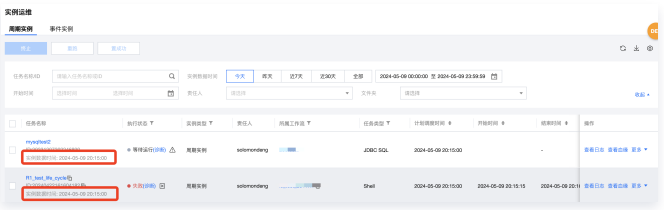
©2013-2025 Tencent Cloud. All rights reserved.

Scheduling Parameters

Last updated: 2024-11-25 14:55:58

1. Time Parameters

The scheduling parameters are used during task scheduling. They automatically replace values based on the business time of task scheduling and the format of the scheduling parameters, achieving dynamic value assignment within the task scheduling time. Currently, two types of scheduling parameters are supported: instance data time and scheduled execution time.
Applicable scope: DataInLong and data calculation tasks.

Scheduling Parameters	Parameter Values	Parameter Description
Instance Data Time	${…} is the time parameter format of the self-definition parameter. 'yyyy' represents the 4-digit year, 'yy' represents the 2-digit year, 'MM' represents the month, 'dd' represents the day, 'HH' represents the hour, 'mm' represents the minute, and 'ss' represents the second. You can freely combine parameters, such as ${yyyyMMdd}, ${yyyy-MM-dd}, ${HH:mm:ss}, and ${yyyyMMddHHmmss}	It refers to the time when the task processes business data, accurate to the second. The data time can be obtained through ${yyyyMMddHHmmss}. The default T + 0 method, data time rules are as follows: • Minute tasks, data time = scheduled execution time; • Hourly tasks, data time is yyyy-MM-dd HH:00:00 (yyyy-MM-dd HH is consistent with the scheduled execution time); • Day, week, year, month tasks, data time is yyyy-MM-dd 00:00:00 (yyyy-MM-dd is consistent with the scheduled execution time). • For test runs, the time is the submission execution time.  Common Parameters: • ${yyyyMMdd} • ${yyyy-MM-dd} • ${HH:mm:ss} • ${yyyyMMddHHmmss}
Scheduled Execution Time	$(…) is the time parameter format of the self-definition parameter.	It refers to the planned scheduling time of the instance, which is also the generation time of the instance, accurate to the second. The planned scheduling time can be obtained through $(yyyyMMddHHmmss). Common Parameters: $(yyyyMMdd)

The following are commonly used parameter formats for instance data time:

Time	Time Parameter Format	Remarks
Next N Years	${yyyyMMdd+Ny}	If referenced as dt=${yyyyMMdd-1M}, it will execute the replacement: dt=20210610 If referenced, it will execute the replacement: dt=20210709 If referenced as time=${yyyyMMddHHmmss-3h}, it will execute the replacement: time=20210710050000 If referenced as ti=${yyyyMMddHHmmss-25m}, it will execute the replacement: ti=20210710073500
Previous N Years	${yyyyMMdd-Ny}	
Next N Months	${yyyyMMdd+NM}	
Previous N Months	${yyyyMMdd-NM}	
Next N	${yyyyMMdd+Nw}	

©2013-2025 Tencent Cloud. All rights reserved.

Weeks		
Previous N Weeks	${yyyyMMdd-Nw}	
N days later	${yyyyMMdd+Nd}	
N days before	${yyyyMMdd-Nd}	
N hours later	${yyyyMMddHHmmss+NH}	
N hours before	${yyyyMMddHHmmss-NH}	
N minutes later	${yyyyMMddHHmmss+Nm}	
N minutes before	${yyyyMMddHHmmss-Nm}	
Scheduled Timestamp (10 digits)	${timestamp}	${timestamp}=1625875200 Supports using "+", "-", and other operators.
Scheduled Timestamp (13 digits)	${timestamp_ms}	${timestamp_ms}=1698508800000 Supports using "+", "-", and other operators.

Provides quick conversion expressions for common dates, as follows:

Serial number	Time Parameter Format	Description
1	${yyyyMMdd+TE}	TENDAY END The data date corresponds to the end of ten days
2	${yyyyMMdd+ME}	MONTH END The data date corresponds to the end of the month
3	${yyyyMMdd+QE}	QUARTER END The data date corresponds to the end of the quarter
4	${yyyyMMdd+HYE}	HALF YEAR END The data date corresponds to the end of half a year
5	${yyyyMMdd+YE}	YEAR END The data date corresponds to the end of the year
6	${yyyyMMdd+TS}	TENDAY START The data date corresponds to the beginning of a ten-day period
7	${yyyyMMdd+MS}	MONTH START The data date corresponds to the beginning of the month
8	${yyyyMMdd+HYS}	HALFYEAR START The data date corresponds to the beginning of the half-year
9	${yyyyMMdd+YS}	YEAR START The data date corresponds to the beginning of the year
10	${yyyyMMdd+PME}	PRI MONTH END The data date corresponds to the end of last month
11	${yyyyMMdd+PYE}	PRI YEAR END The data date corresponds to the end of the previous year

The following are commonly used parameter formats for planned scheduling time:

Time	Time Parameter Format	Remarks
Scheduling Time	$(yyyyMMddHHmmss)	If referenced as dt=$(yyyyMMdd-1M), it will execute the replacement: dt=20210610 If referenced, it will execute the replacement: dt=20210709

©2013-2025 Tencent Cloud. All rights reserved.

Next N Years	$(yyyyMMdd+Ny)	If referenced as time=$(yyyyMMddHHmmss-3h), it will execute the replacement: time=20210710050000 If referenced as ti=$(yyyyMMddHHmmss-25m), it will execute the replacement: ti=20210710073500
Previous N Years	$(yyyyMMdd-Ny)	
Next N Months	$(yyyyMMdd+NM)	
Previous N Months	$(yyyyMMdd-NM)	
Next N Weeks	$(yyyyMMdd+Nw)	
Previous N Weeks	$(yyyyMMdd-Nw)	
N days later	$(yyyyMMdd+Nd)	
N days before	$(yyyyMMdd-Nd)	
N hours later	$(yyyyMMddHHmmss+NH)	
N hours before	$(yyyyMMddHHmmss-NH)	
N minutes later	$(yyyyMMddHHmmss+Nm)	
N minutes before	$(yyyyMMddHHmmss-Nm)	
Scheduled Timestamp	$(timestamp)	$(timestamp)=1625875200 supports using "+", "-", and other operators.

Second, other system built-in parameters

Parameter	Meaning
${projectIdent}	Project Identifier
${workflowName}	Workflow name
${taskName}	Task Name
${taskId}	Task ID
${taskInCharge}	Responsible Person
${taskType}	Task Type

Use Case 1: Directly use in SQL to get information about the current project, workflow, etc.

```
select
  "${projectIdent}" as projectIdent,
  "${workflowName}" as workflowName,
  "${taskName}" as taskName,
  "${taskId}" as  taskId,
  "${taskInCharge}"  as taskInCharge,
```

©2013-2025 Tencent Cloud. All rights reserved.

```
  "${taskType}" as   taskType,
  user_id
from
  wedata_demo_db.user_info
limit
  10;
```

Use Case 2: Use --name + variable to specify the task name on yarn in SparkSQL, PySpark, and Spark tasks.

```
--name  ${projectIdent}-${workflowName}-${taskInCharge}
```

When connecting to the kyuubi data source in SparkSQL, the parameter name currently needs to be specified using spark.app.name.

```
spark.app.name=${projectIdent}-${workflowName}-${taskInCharge}
```



©2013–2025 Tencent Cloud. All rights reserved.

Function Development

Last updated：2024-11-25 14:57:56

1. Log in to WeData Console, and enter the function development page.
2. Click **Project List** in the left menu to find the target project for function development feature operations.
3. After selecting the project, click to enter the **Data Development** module.
4. Click **Function Development** in the left menu.

Function Overview

UDF functions uploaded in the Resource Management feature can be used in Function Development. Define the function's classification and class name, and you can use it during data development. Currently, Hive SQL, Spark SQL, and DLC SQL function creation are supported.

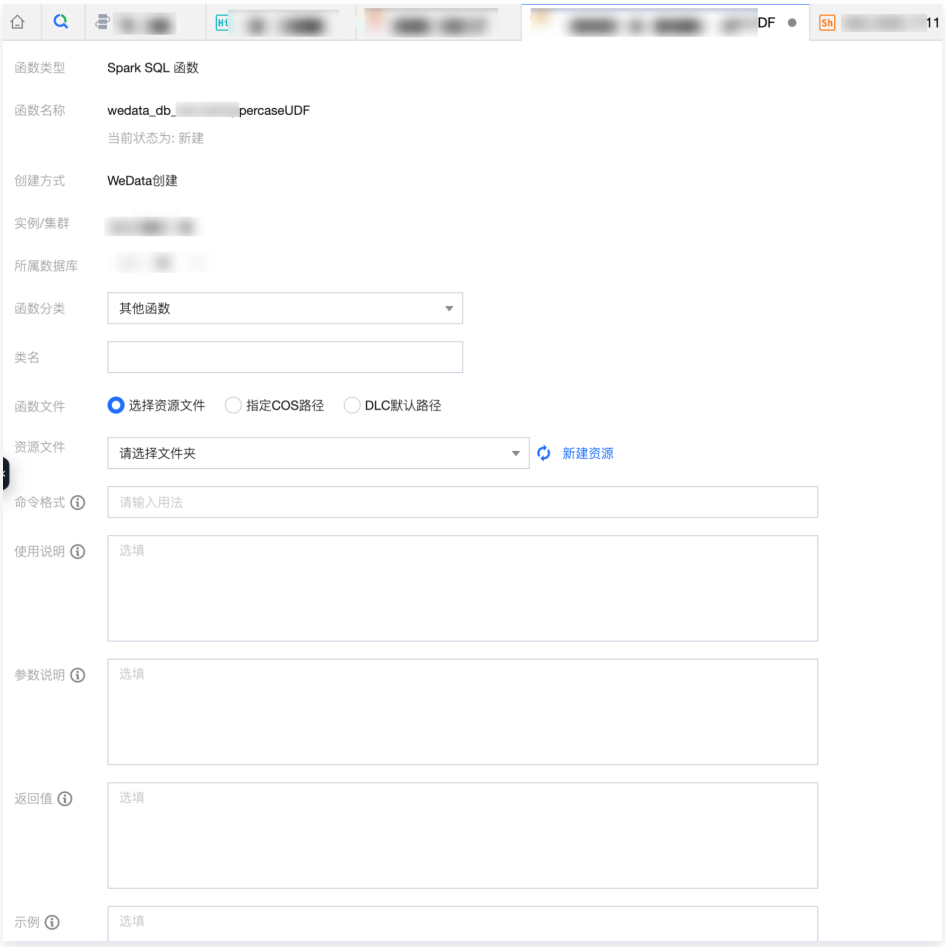
Creating function

1. On the function development page, click + to choose to create a new Hive SQL function, Spark SQL function, or DLC SQL function. You can also create the corresponding type of function by clicking the ••• button on the right side of the target path under the specific function type.



2. Configure the function in the pop-up window and click **Save** and submit to complete the function registration.

©2013-2025 Tencent Cloud. All rights reserved.



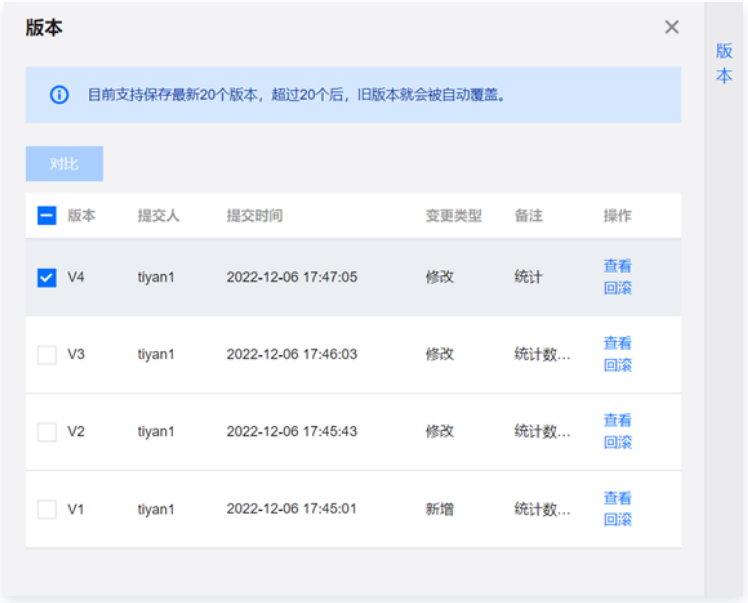
The configuration information is shown in the table below:

Information	Description
Function Classification	Create the function under the preset function classification according to its nature. Classifications include: Analytic Function, Encryption Function, Aggregation Function, Logical Functions, Date and Time Functions, Mathematical Functions, Conversion Functions, String Functions, IP and Domain Functions, Window functions, Other Functions.
Class	Enter the class name of the function.
Function File	Select the file path of the function source file: • Select the resource file: choose the function file from the jar or zip resources uploaded via the Resource Management feature. • Specify COS path: obtain the function file from the platform COS bucket path.
Resource File	The function file option is **Select Resource File** , you need to select the required function file in the resource management directory.
COS Path	The function file option is **Specify COS Path** , you need to enter the path where the function file is located in the platform COS bucket.
Command format	The format is: functionName(parameters). For example, the sum function command format is sum(col)
Use Instructions	Usage instructions for custom functions. For example: sum function usage instructions: calculate the total value.
Parameter description	Parameter description of the custom function. For example: sum function parameter description: col: mandatory. The column value can be of DOUBLE, DECIMAL, or BIGINT type. If the input is of STRING type, it will be implicitly converted to DOUBLE type before participating in computations.

©2013–2025 Tencent Cloud. All rights reserved.

Returned values	Return value description of the self-defined function. For example: sum function return value: returns DOUBLE type.
Sample code	Example description of the self-defined function. For example: sum function example: calculate the total sales (sales), command example: select sum(sales) from table.

3. After the function changes, you can save the history record with version features, including version number, submitter, submission time, change type, remarks, and support the operation of rolling back to the historical version.



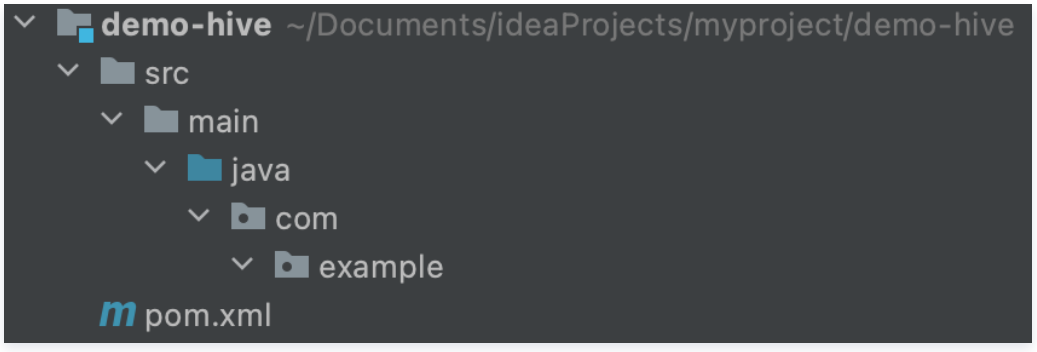
Function Example

Spark SQL Function Development Example

1. Create a project

   Create a Maven project and introduce hive-exec dependencies. Use the mvn command line to create the project, or you can create it through the IDEA tool, where groupId and artifactId are replaced with your own definition name.

```
mvn archetype:generate -DgroupId=com.example -DartifactId=demo-hive -Dveriosn=1.0-SNAPSHOT -
Dpackage=com.example
```



2. Writing code

   Introduce hive-exec and junit test pom dependencies.

```
<dependencies>
  <dependency>
      <groupId>org.apache.hive</groupId>
      <artifactId>hive-exec</artifactId>
      <version>2.3.8</version>
  </dependency>
```

©2013-2025 Tencent Cloud. All rights reserved.

```
    <dependency>
        <groupId>junit</groupId>
        <artifactId>junit</artifactId>
        <version>4.11</version>
        <scope>test</scope>
    </dependency>
</dependencies>
```

Create a Java class in the src/main/java/com/example directory, extend the org.apache.hadoop.hive.ql.exec.UDF class, write the evaluate method, and implement the specific behavior of the self-defined function, e.g., convert the input string to uppercase.

```
public class UppercaseUDF extends UDF {
    public String evaluate(String input) {
        return input.toUpperCase();
    }
}
```

3. Compiling and Packaging

Import the Maven packaging plugin. In the project root path, run the mvn package command for compilation and packaging. The generated package name is: demo-hive-1.0-SNAPSHOT.jar.

```
<build>
      <plugins>
          <plugin>
              <groupId>org.apache.maven.plugins</groupId>
              <artifactId>maven-compiler-plugin</artifactId>
              <version>3.8.1</version>
              <configuration>
                  <source>1.8</source>
                  <target>1.8</target>
              </configuration>
          </plugin>
          <!--(start) for package jar with dependencies -->
          <plugin>
              <groupId>org.apache.maven.plugins</groupId>
              <artifactId>maven-assembly-plugin</artifactId>
              <version>3.0.0</version>
              <configuration>
                  <archive>
                      <!--Specify the class where the main method is located-->
                      <manifest>
                          <mainClass>com.example.UppercaseUDF</mainClass>
                      </manifest>
                  </archive>
                  <!--Do not change jar-with-dependencies-->
                  <descriptorRefs>
                      <descriptorRef>jar-with-dependencies</descriptorRef>
                  </descriptorRefs>
                  <appendAssemblyId>false</appendAssemblyId>
              </configuration>
              <executions>
                  <execution>
                      <id>make-assembly</id> <!-- this is used for inheritance merges -->
                      <phase>package</phase> <!-- bind to the packaging phase -->
                      <goals>
                          <goal>single</goal>
                      </goals>
                  </execution>
```

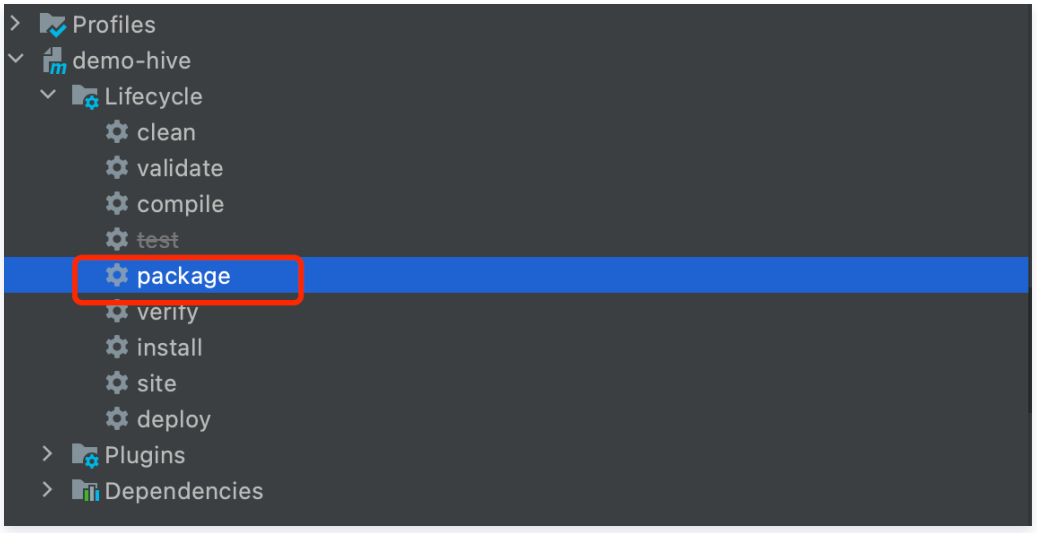
©2013-2025 Tencent Cloud. All rights reserved.

```
                </executions>
            </plugin>
            <!--(end) for package jar with dependencies -->
        </plugins>
    </build>

    <repositories>
        <repository>
            <id>alimaven</id>
            <name>aliyun maven</name>
            <url>http://maven.aliyun.com/nexus/content/groups/public/</url>
        </repository>
    </repositories>
```

Run the mvn package command:

```
mvn package -Dmaven.test.skip=true
```

You can also use the IDEA software tool to package and generate a jar file with dependencies.

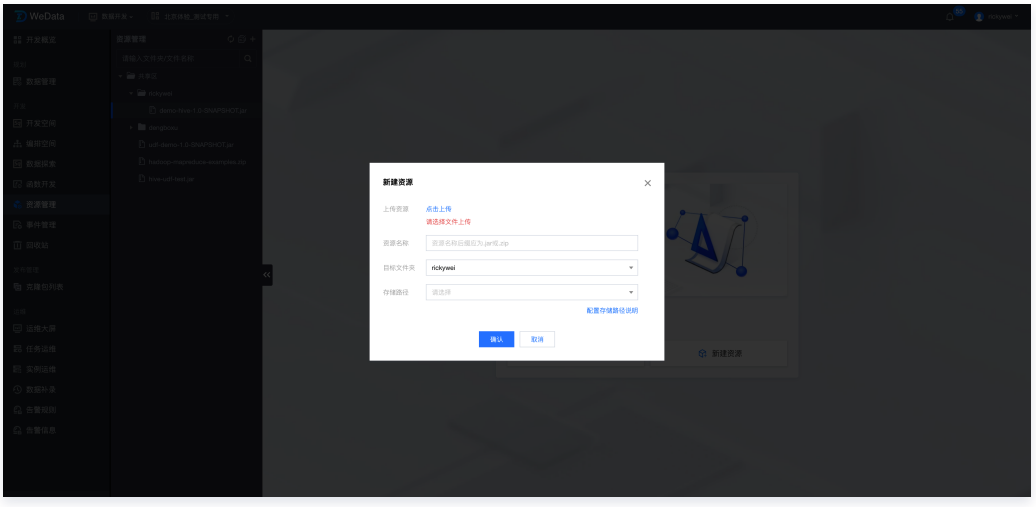


4. Function Operations

Enter the WeData function development page, create a custom function, fill in the fully qualified class path name: com.example.UppercaseUDF, select the corresponding resource file, which is the jar package implementing the custom function code. If there is no resource file, create the resource first.

4.1 Resource Upload:

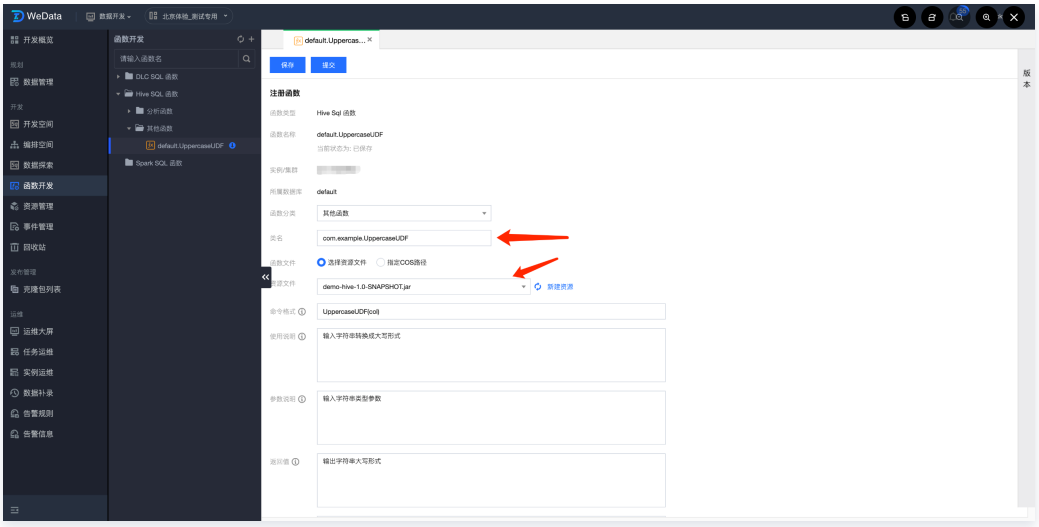
Upload the demo-hive-1.0-SNAPSHOT.jar function package through the resource management feature.



©2013-2025 Tencent Cloud. All rights reserved.

4.2 Function Creation:

Create a Spark SQL function through the function development feature.

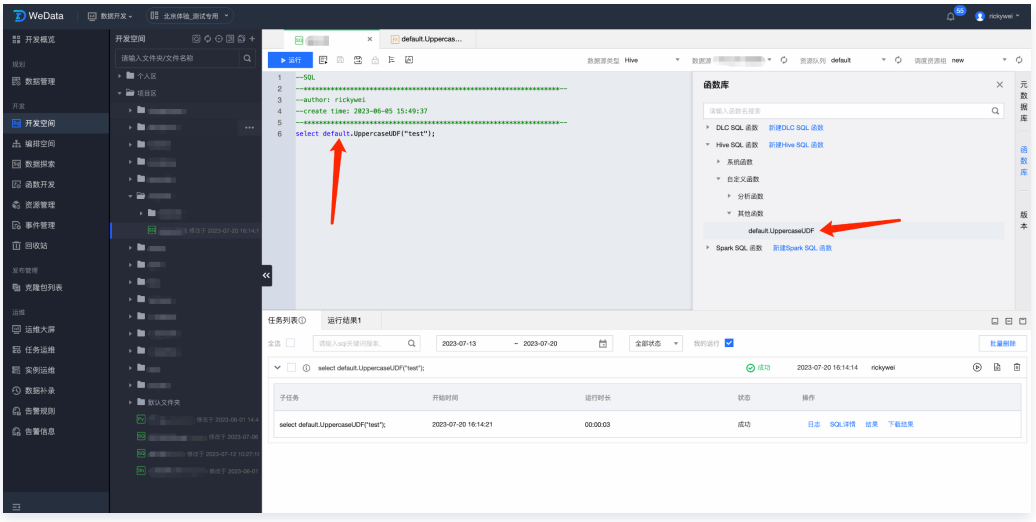


Example Function Information:

Information	Content
Function Classification	Other Functions
Class	com.example.UppercaseUDF
Function File	Select Resource File
Resource File	demo-hive-1.0-SNAPSHOT.jar
Command format	UppercaseUDF(col)
Use Instructions	Convert the input string to uppercase format
Parameter description	Input string type parameter
Returned values	Output string in uppercase

4.3 Function Usage:

In the Development Space, create a new SQL file and use the successfully created function to verify its feature.



DLC SQL Function Development Example

©2013-2025 Tencent Cloud. All rights reserved.

Users can write UDF functions and package them into jar files, then use them in Data Development – Function Management. Currently, DLC Lakehouse Computing UDFs are in Hive format, inheriting from org.apache.hadoop.hive.ql.exec.UDF and implementing the evaluate method. The steps to create functions on the function development page can refer to Hive SQL functions.
Java Code Example:

```
public class MyDiff extends UDF {
    public ArrayList<Integer> evaluate(ArrayList<Integer> input) {
        ArrayList<Integer> result = new ArrayList<Integer>();
        result.add(0, 0);
        for (int i = 1; i < input.size(); i++) {
        result.add(i, input.get(i) - input.get(i - 1));
        }
        return result;
    }
}
```

Sample POM file:

```
<dependencies>
    <dependency>
        <groupId>org.slf4j</groupId>
            <artifactId>slf4j-log4j12</artifactId>
            <version>1.7.16</version>
            <scope>test</scope>
        </dependency>
        <dependency>
            <groupId>org.apache.hive</groupId>
            <artifactId>hive-exec</artifactId>
            <version>1.2.1</version>
        </dependency>
</dependencies>
```

©2013–2025 Tencent Cloud. All rights reserved.

Resource Management

Last updated：2024-11-25 14:58:58

Feature Overview

The Resource Management feature supports uploading local .jar, .zip, .csv, .sh, .py, .kjb, and .ktr files to the platform's COS Bucket, which facilitates subsequent Script Development, Compute Task Development, Function Development, and other feature scenarios.

Create New Resource

Upload local resource files to WeData for management, providing the ability to create self-defined functions during the data development process.

1. Log in to WeData Console.
2. Click **Project List** in the left menu to find the target project for Resource Management.
3. After selecting the project, click to enter the **Data Development** module.
4. Click **Resource Management** in the left menu.
5. In the Resource Management directory, click ➕ to open the operation menu and select **Create New Resource**.

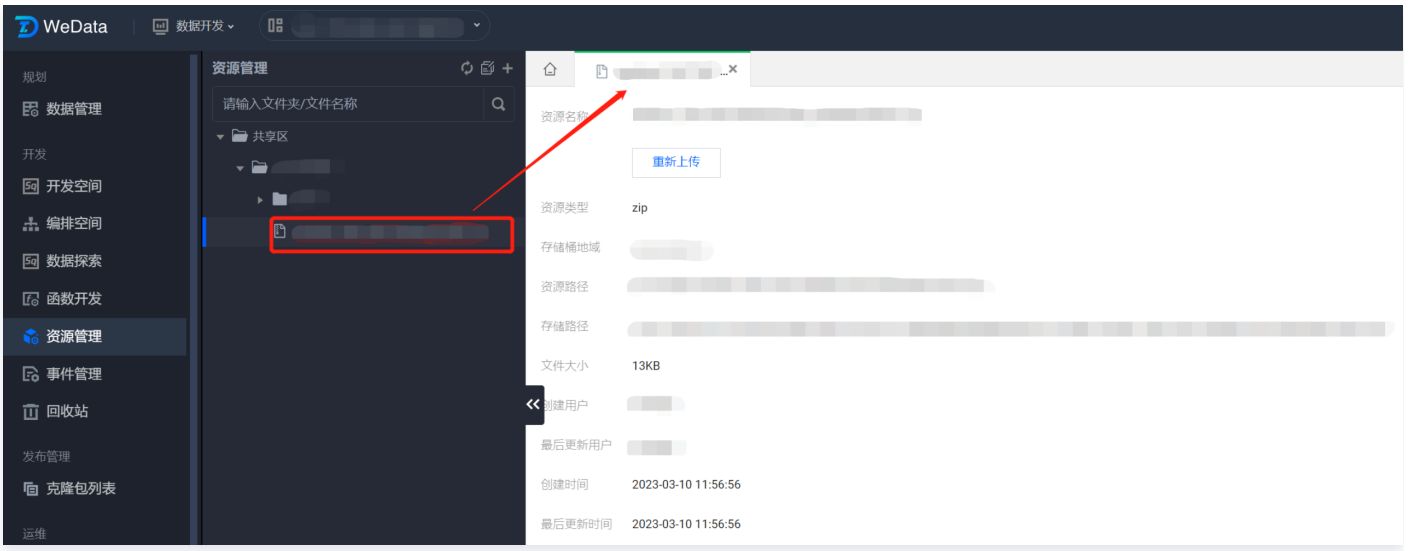


6. In the pop-up window, upload the local resource file, enter the resource name, select the resource directory folder, and storage path.



7. After configuration is complete, click **Confirm** to maintain the script resources within the Resource Management feature. You can also view detailed information on the respective resource page.

©2013-2025 Tencent Cloud. All rights reserved.



Resource Management Operations

1. Click the + icon to perform operations such as creating a new folder and creating a new resource item.
2. Click the icon to perform batch deletion operations.
3. Click the ··· icon to open, rename, delete, or download resources.
4. Double-click a created resource to view detailed information, including resource name, resource type, storage path, file size, creation user, creation time, and provide re-upload resource operation features.
5. Batch Deletion: Click **Batch Deletion**, select the resources to be deleted, and click **Confirm** to delete the resources.

> **Note**
> Resources that are referenced cannot be deleted.



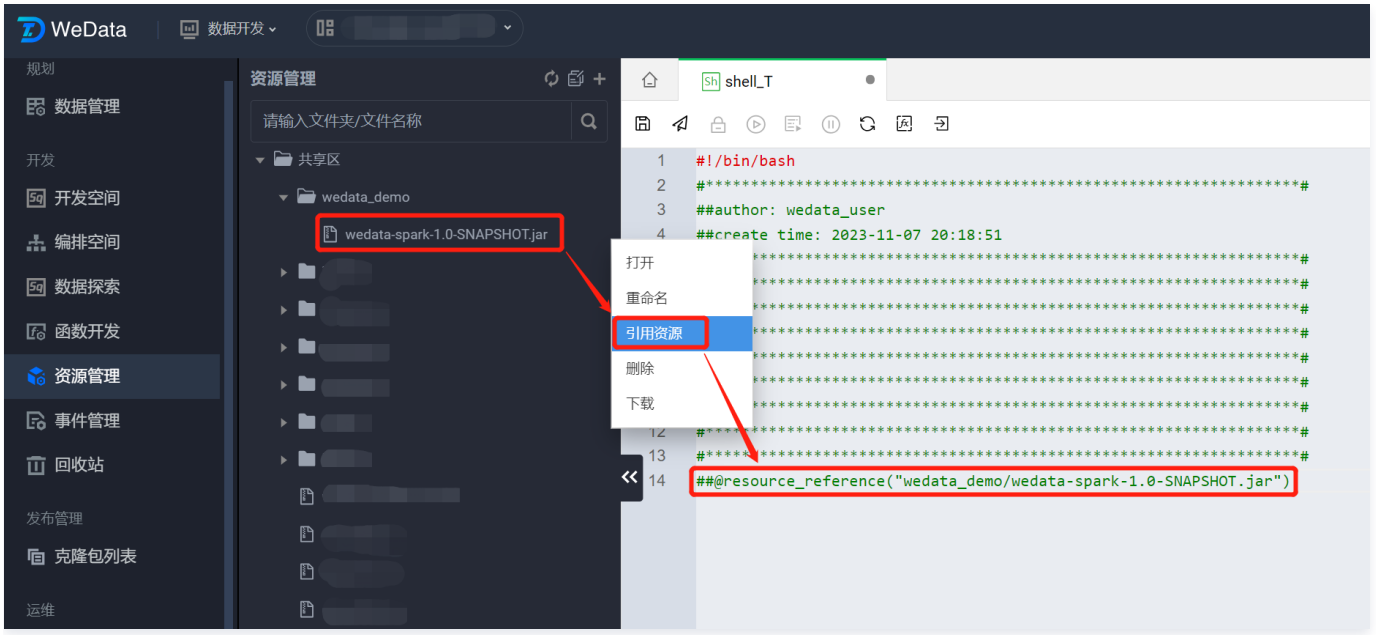
6. Create Resource: Click **Create Resource**, configure the file name and file address. After selecting the file to upload, click **Confirm** to complete the resource upload.

©2013-2025 Tencent Cloud. All rights reserved.



Reference resources

After selecting the file resource in the directory, click **Reference Resources**. If the active interface is an orchestration space computing task or a development space script, code will be automatically generated at the cursor's input location. Code format: `@resource_reference{"Resource Path/Resource Name"}` . For more code format information, see the Code Reference below.



> **Note:**
> Referencing resources is currently effective only for code type tasks and scripts. Before referencing, please open the code type tasks and scripts on the right.

Code Reference

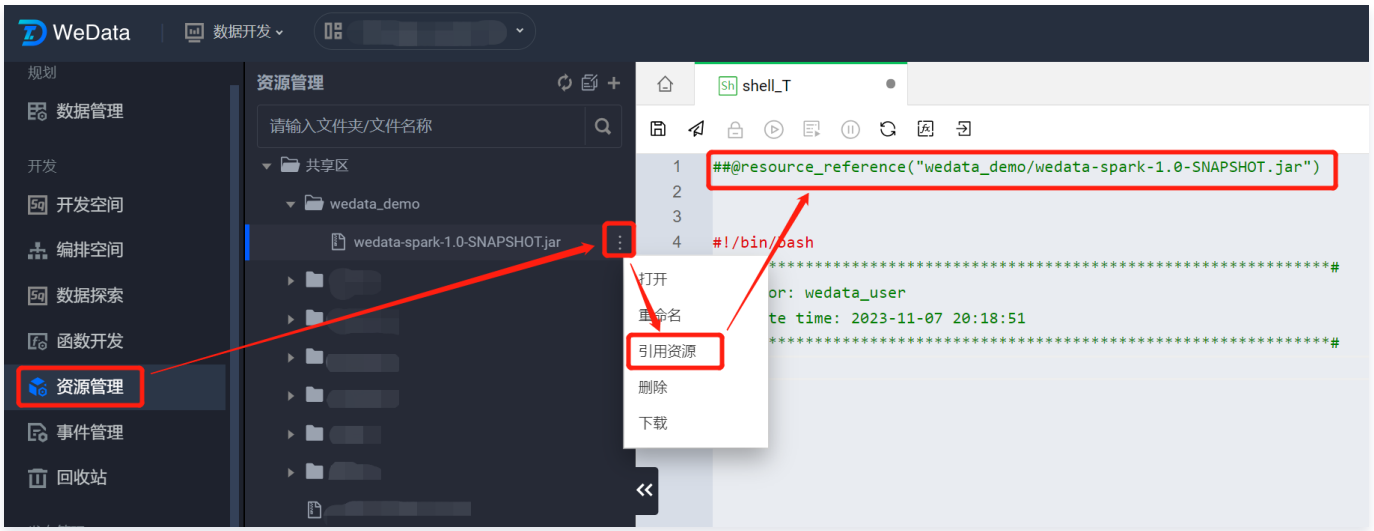
Task Type	Code Format
Development Space: SQL Orchestration Space: HiveSQL, JDBC SQL, SparkSQL, DLCSQL, Impala, TCHouse-P, Trino, TChouse-X	--@resource_reference("Resource Management Path/Resource Name")
Development Space: Shell, Python, PySpark Orchestration Space: Shell, Python, PySpark	##@resource_reference("Resource Management Path//Resource Name")

Sample Code

©2013-2025 Tencent Cloud. All rights reserved.

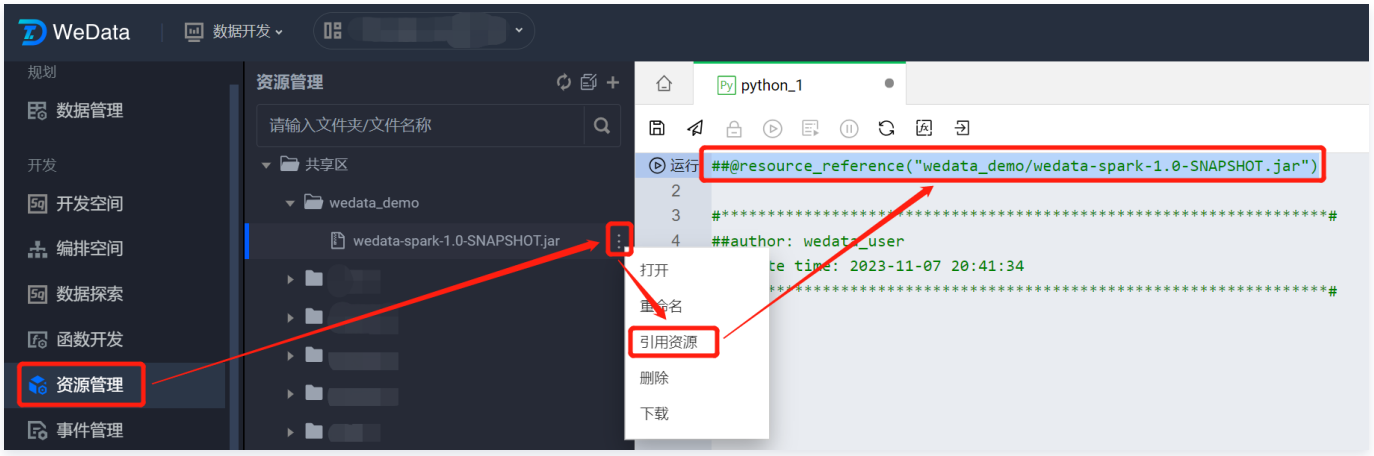
Referencing resources in development space Shell scripts

- Support for referencing local resources in Shell scripts to execute commands.
- For example: Insert resource references at the top of a Shell editor, resulting in the final file format:



Referencing resources in orchestration space Python tasks

- Support for referencing local resources in Python tasks to execute commands, such as py files.
- For example: Insert resource references at the top of a Python editor, resulting in the final file format:



©2013–2025 Tencent Cloud. All rights reserved.

Event Management

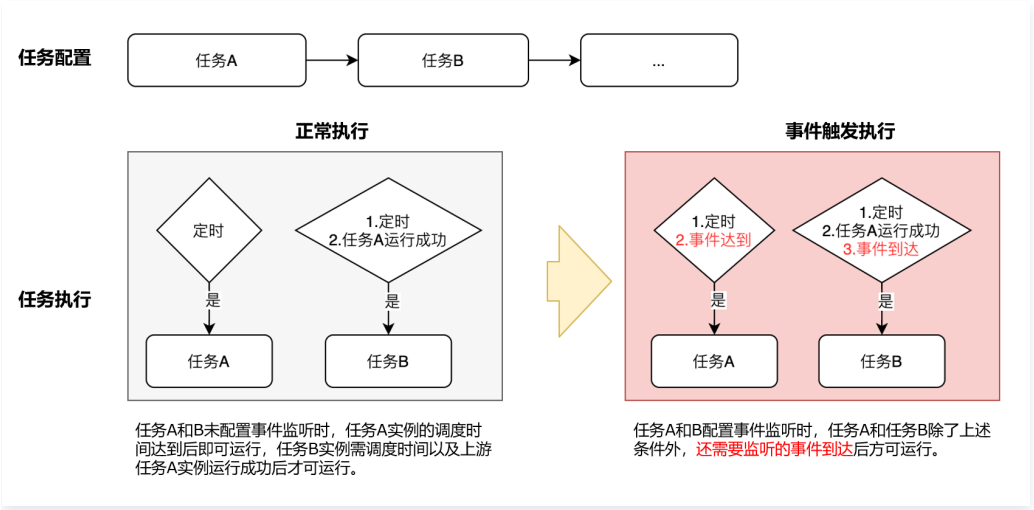
Last updated：2024-11-25 14:59:22

Event Overview

Orchestration Space supports event-triggered scheduling feature. After configuring event dependency listening in workflow tasks, the task needs to wait for the event to occur before execution.

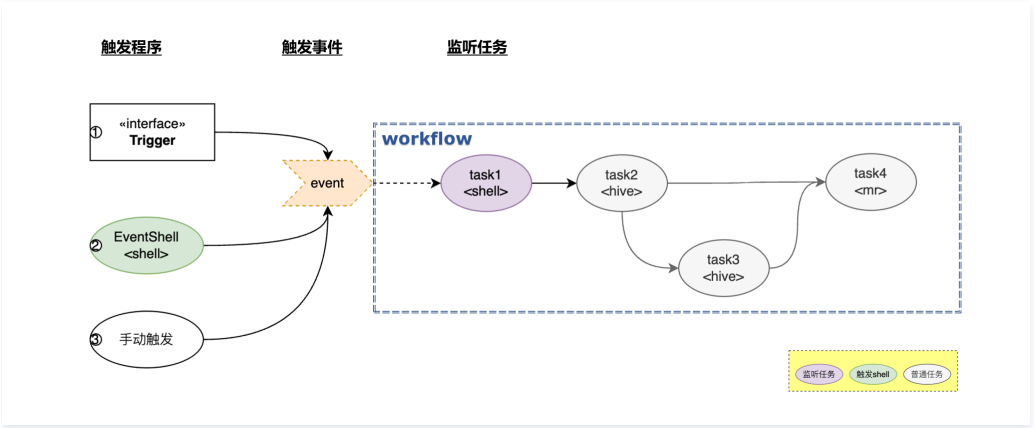
- Event: A message defined by the user based on business needs to trigger task instances. One event can support the triggering of multiple task instances.
- Event Trigger: The task scheduling system triggers task instances through external events, such as external scheduling systems and file events.

Event Trigger Principles



Event Trigger Process

Data development supports the event trigger scheduling feature, suitable for scenarios where tasks depend on a certain event to trigger execution. It consists of trigger programs, trigger events, and listening tasks. First, based on the business definition, trigger events need to be defined under the project, then write a trigger program to send events. Once the task listens to the event arrival, it can run.



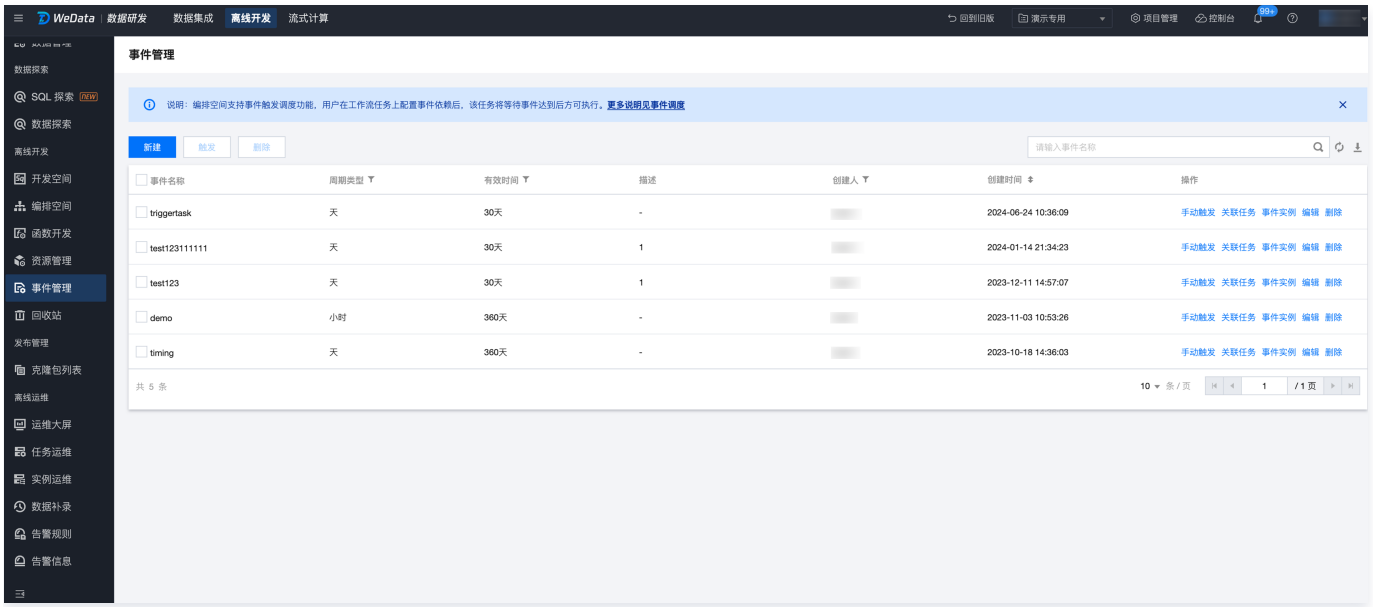
Operation step

1. Log in to WeData Console.
2. Click the **Project List** on the left.
3. Select the corresponding project and click to enter the **Data Development** module.
4. On the left menu, click **Event Management**.

©2013-2025 Tencent Cloud. All rights reserved.

Creating Events

1. The event management page can create events by adding or importing. Click the **New** button in the upper left corner and configure event information in the popup.



2. Enter the Create Event interface. The configuration items include Event Name, Cycle Type, Event Trigger Validity Period, and Description information.



Configuration information:

Information	Description
Event name	A message defined by the user based on business needs to trigger task instances. One event can support the triggering of multiple task instances.
Cycle Type	The event instance and task instance can be triggered only when their planned scheduling times match the cycle type of the listening event. The correspondence between the event cycle type and task cycle type is: • Minute > Minute (corresponding time format: yyyyMMddHHmm) • Hour > Hour (corresponding time format: yyyyMMddHH) • Day > Day, Week, Month, Year (corresponding time format: yyyyMMdd)

©2013–2025 Tencent Cloud. All rights reserved.

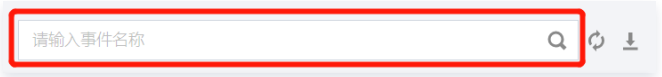
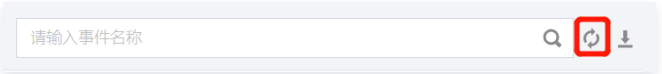
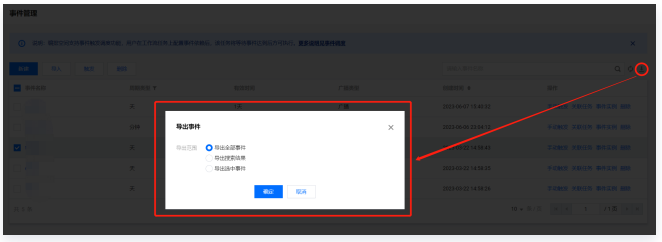
Event Trigger Valid Time	The maximum uptime of the event, after which it will be automatically terminated.

Event Operations

After the event is successfully created, you can manage it using various operation features.



Feature information:

Information	Description
New	Listen for event creation.
Trigger	Check the created listening events in the list. If the corresponding listening event has been associated with task scheduling, you can trigger events in batches through the trigger feature.
Delete	Delete the selected listening events in the list.
Edit	You can modify the event trigger valid time and description.
Manual trigger	Manually trigger a single listening event.
Associated tasks	View the scheduling tasks associated with the event.
Event instance	Navigate to the Instance Operation and Maintenance page to view the event instances generated after the event is triggered.
Delete	Delete a single instance.
Search	Search for listening events in the list by event name. 
Refresh	Refresh the current event list. 
Export	Export the event as a CSV file. You can choose to export all events, export search results, or export selected events. 

©2013–2025 Tencent Cloud. All rights reserved.

Development Space Overview

Last updated: 2024-11-25 14:59:52

Access the Script Development page

1. Log in to WeData Console.
2. Click **Project List** in the left menu and find the target project for the development space feature.
3. After selecting the project, click to enter the **Data Development** module.
4. Click **Development Space** in the left menu.

Development Space Overview

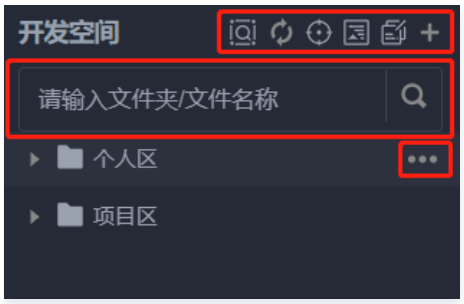
WeData Data Development's Development Space supports creating, editing, and debugging SQL, ETL, Shell, Python, PySpark development scripts. These can be used in collaboration with Data Workflows. Configured scripts from Ad-hoc Development can directly participate in Data Workflow orchestration, becoming a Task Node, achieving code reuse and overall process optimization.

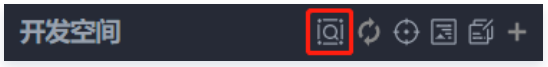
©2013-2025 Tencent Cloud. All rights reserved.

Development Space feature

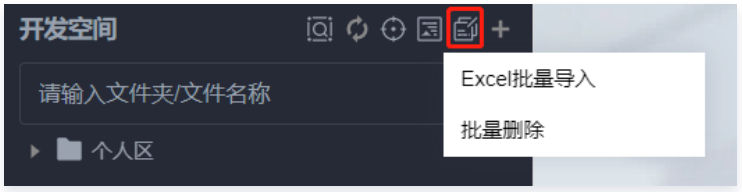
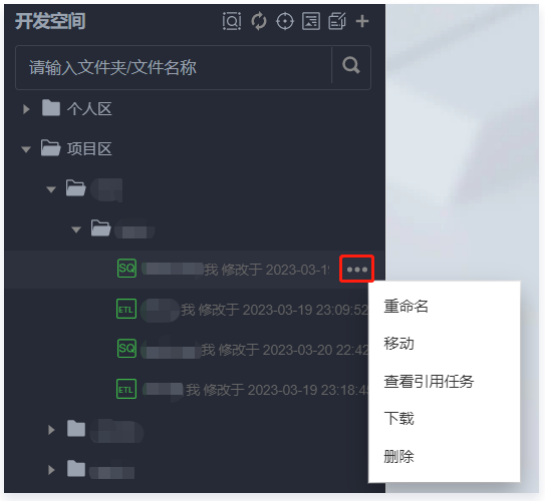
Last updated：2024-11-25 15:00:15

Feature Overview



Feature	Description
Search	Support search by folder and file name 
Global search	Support global directory search, providing rich search dimensions  Search page: 
Refresh	Refresh directory tree 
Locate current tree node	One-click positioning to the current tree node 
Collapse all tree nodes	One-click collapse of all expanded directories 
Batch operation	Support batch deletion of script files, batch import of SQL scripts in Excel format

©2013-2025 Tencent Cloud. All rights reserved.

	
New	Support creating new folders and files, including offline synchronization, SQL, Shell, Python file types 
Directory operations	Support creating new folders and files; support batch export and batch import of script files, both organized as zip files according to directory structure 
File Operations	Support renaming, moving, downloading, and deleting development files, and viewing the citation status of development files in the Orchestration Space 

©2013–2025 Tencent Cloud. All rights reserved.

Script Development

Last updated: 2024-11-25 15:00:50

Steps for Script Development

Step 1: Create a New Folder

1. Enter **Development Space**, in the **Development Space Directory Tree**, click + button, in the expanded dropdown menu, click **Create New Folder**.



2. Enter the folder name, choose the target folder, and click **Confirm**.



> **Note**
> WeData supports creating multi-level folder directories. You can save newly created folders in the root directory or other already existing folders.

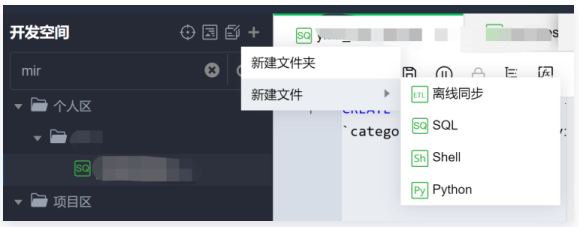
3. Users can freely choose to create and save folders and files in the Personal Area or Project Area under the Development Space.
   - **Personal Area:** Folders and files created or stored in the Personal Area can only be viewed by the current user. Other users cannot view, edit, or manage.
   - **Project Area:** Folders and files created or stored in the Project Area are shared project files, which can be viewed by all users under the current project. Non-project members cannot view them.



©2013-2025 Tencent Cloud. All rights reserved.

Step 2: Create a New File

1. Right-click the folder name on the Temporary Query Page or click the at the top of the directory tree, and select the type of file you want to create.

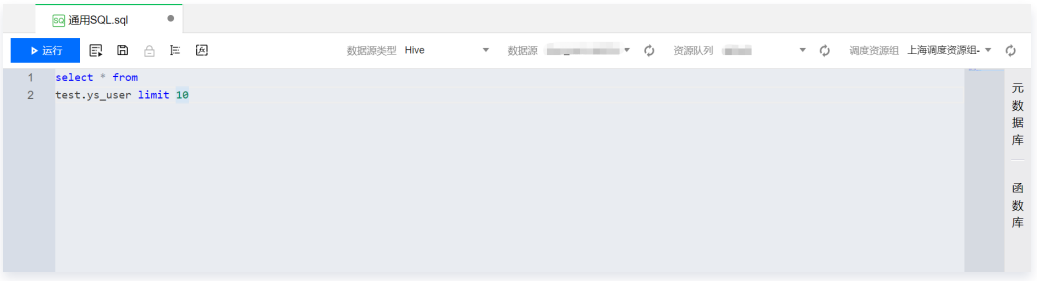


2. In the popup window, enter the script file name, choose the target folder, and click **Confirm**.

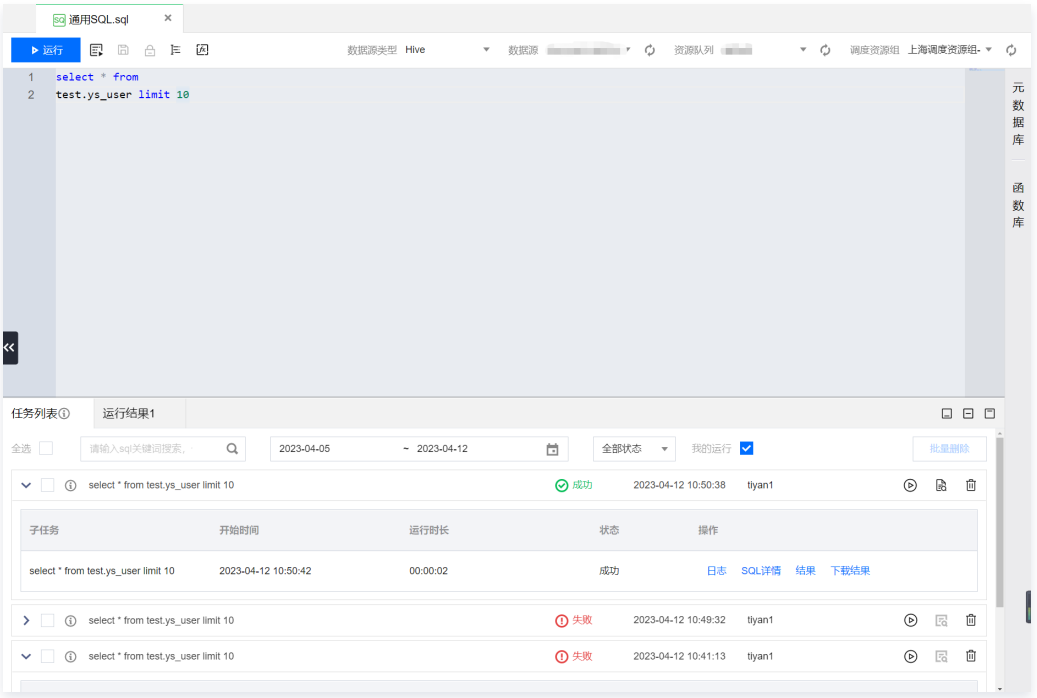
> **Note**
> File names can only include uppercase and lowercase letters, numbers, and underscores, and can be up to 100 characters long.

Step 3: Edit the file and run it

1. In the Script tab, enter the relevant code statements. For example, see the following screenshot for SQL:



2. click **Run** button to run the code and view the results.



Script File Type

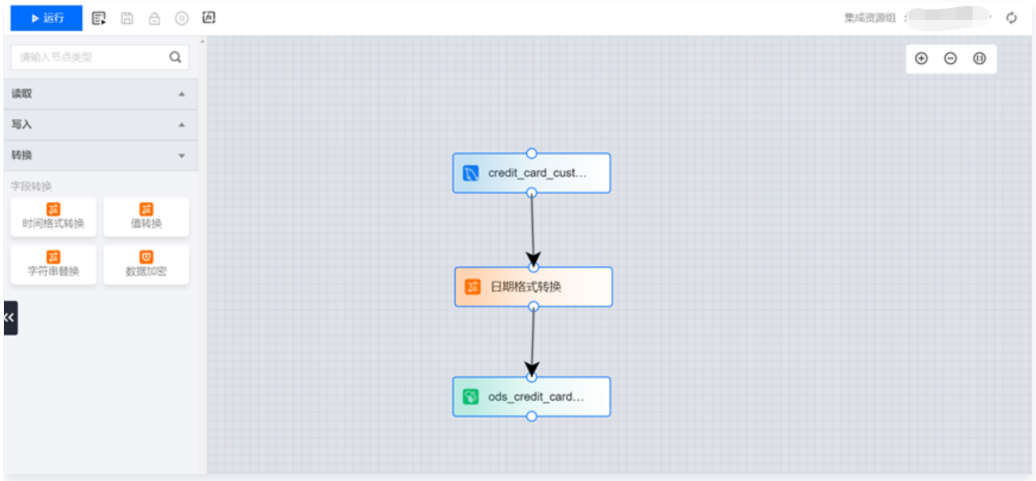
The file types supported by the development space include offline synchronization, SQL, Shell, Python, and PySpark. They support creating, editing, downloading, deleting, and more.

©2013–2025 Tencent Cloud. All rights reserved.

Synchronization File

> **Note**
> For configuration details, see the offline synchronization section in DataInLong Offline Synchronization.

Offline synchronization file suffix is ".dg", supporting creation and import. It supports offline synchronization process configuration, including reading data sources, field conversion, writing data source configuration, running, advanced running, saving, formatting, and variable viewing.



SQL File

SQL file suffix is ".sql", supporting creation and import. It supports multiple SQL edits and debugging, including Hive SQL, Spark SQL, JDBC SQL, DLC SQL, and others. It supports dozens of data sources including Hive, Oracle, and MySQL, as well as running, advanced running, saving, formatting, and project variable viewing.

For example: when the data source type is Spark, it provides "Advanced Settings" to support Spark SQL parameter configuration.



Currently supported SQL operation data source types are as follows:

Task Type	Supported Data Source Types	Data Source Origin
SQL Task	Hive	System Source&Custom Source
	SparkSQL	System Source
	Impala	System Source&Custom Source
	TCHouse-P	System Source
	Graph Database	Custom Definition Source
	TCHouse-X	Custom Definition Source
	DLC	System Source&Custom Source
	Mysql	Custom Definition Source
	PostgreSQL	Custom Definition Source

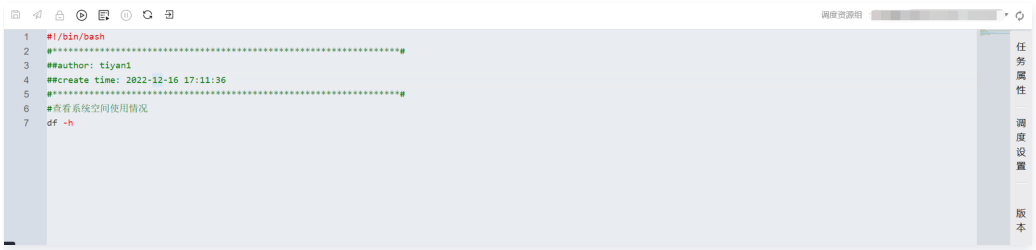
©2013-2025 Tencent Cloud. All rights reserved.

Oracle	Custom Definition Source
SQL Server	Custom Definition Source
Tbase	Custom Definition Source
IBM Db2	Custom Definition Source
DM Database	Custom Definition Source
Greenplum	Custom Definition Source
SAP HANA	Custom Definition Source
Clickhouse	Custom Definition Source
DorisDB	Custom Definition Source
TDSQL-C	Custom Definition Source
StarRocks	Custom Definition Source
Trino	System Source&Custom Source
Kyuubi	System Source&Custom Source

Shell File

The suffix of the Shell file is ".sh", supporting creation and import. Used for online development of Linux Shell files. Supports Shell file execution, advanced execution, viewing project variables, reference, editing, and saving.
Supports executing hdfs-related commands directly in the Shell script interface. You need to manually add the following environment variables in the shell script:

- export HADOOP_CONF_DIR=/usr/local/cluster-shim/v3/conf/emr-xxx/hdfs
- emr-xxx needs to be replaced with the user's own emr engine id.

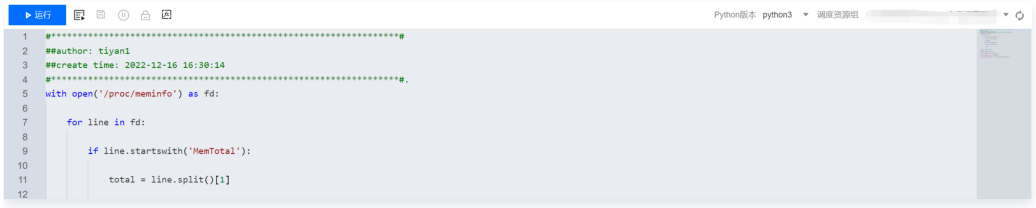


Reference resources

Supports referencing local resources in Shell scripts to execute commands, such as kjb, ktr files. For details, refer to Resource Management.

Python File

The suffix of the Python file is ".py", supporting creation and import. Used for online development of Python files. Supports Python file execution, advanced execution, viewing project variables, reference, editing, and saving. Versions support both Python 2 and Python 3.



Reference resources

Supports referencing local resources in Python scripts to execute commands, such as kjb, ktr files. For details, refer to Resource Management.

©2013-2025 Tencent Cloud. All rights reserved.

PySpark File

The suffix of the PySpark file is ".py", supporting creation and import. Used for online development of Python files. Supports writing Spark applications using Python. Supports execution, advanced execution, viewing project variables, reference, editing, and saving. Versions support both Python 2 and Python 3.



Reference resources

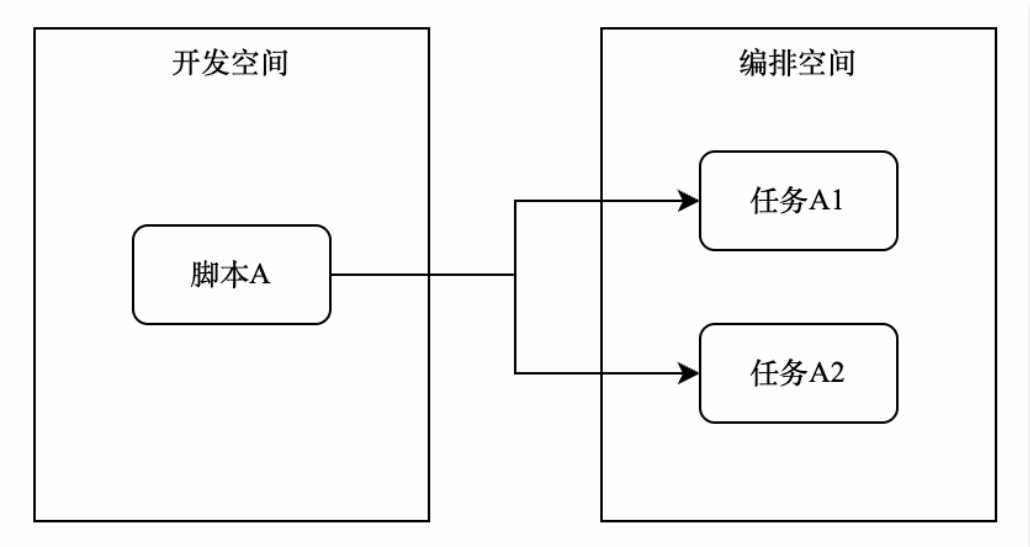
Supports referencing local resources in PySpark scripts to execute commands, such as kjb, ktr files. For details, see Resource Management.

©2013–2025 Tencent Cloud. All rights reserved.

Collaboration between Development Space and Orchestration Space

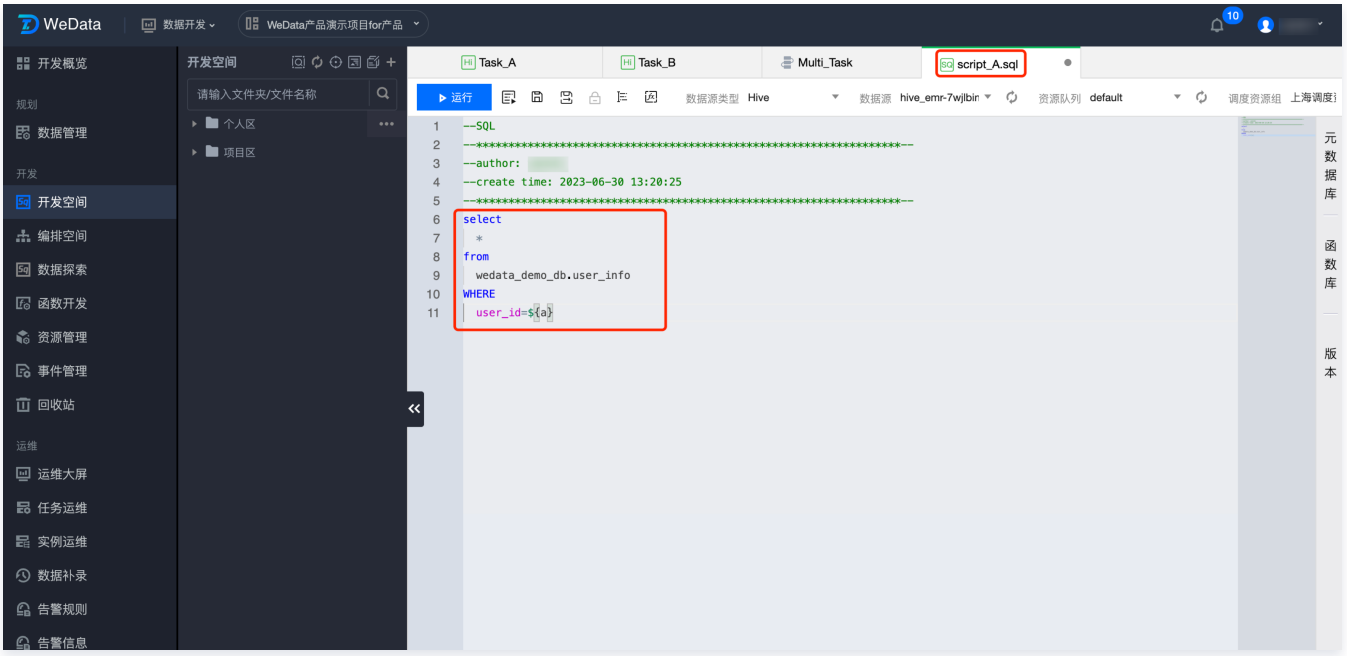
Last updated：2024-11-25 15:01:13

Debug scripts in the Development Space, view SQL execution results, and import them into the Orchestration Space. In the Development Space, the version of a single script can be controlled. When importing into the Orchestration Space, the latest Development Space script file will be imported.
In scenarios where Development Space and Orchestration Space are connected, a Development Space script can correspond to multiple task schedules in the Orchestration Space feature.



Collaboration Steps

1. Create Script A in the Development Space.

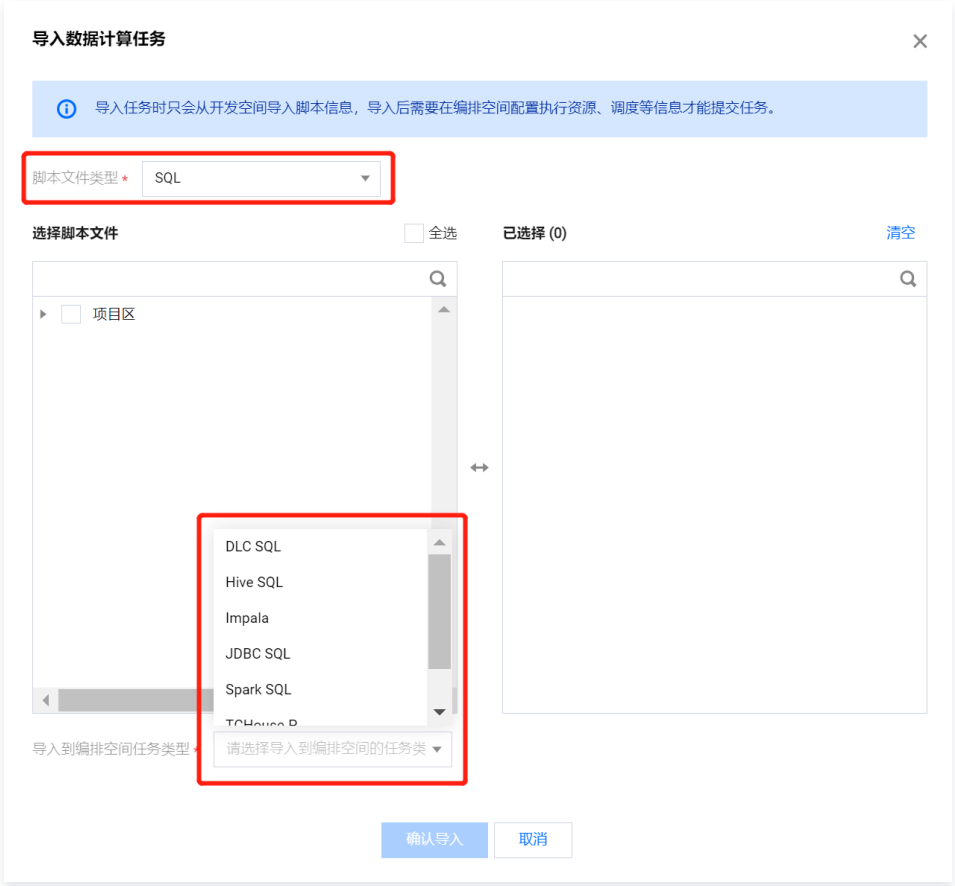


2. Import Script A into the Orchestration Space to generate Task A1. Import Script A again to generate Task A2.
   2.1 Click the Data Calculation directory in the workflow and select Import task.

©2013-2025 Tencent Cloud. All rights reserved.

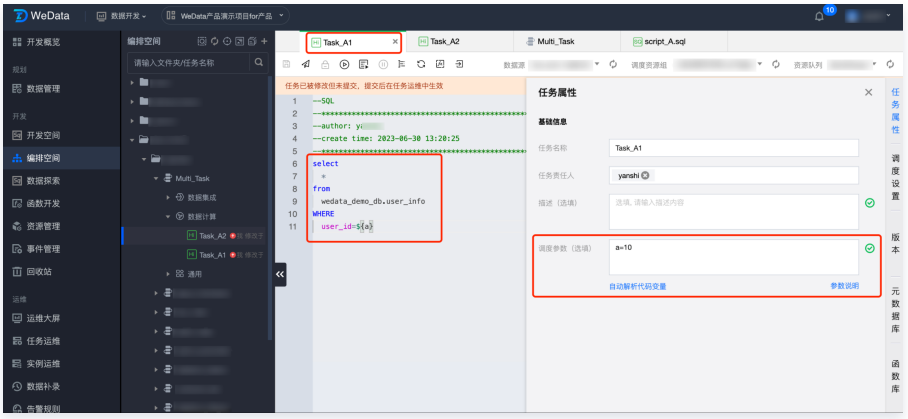


2.2 In the import configuration box, select the script file type from the Development Space, then select the script file, and finally choose the task type for the script to be imported into the Orchestration Space.

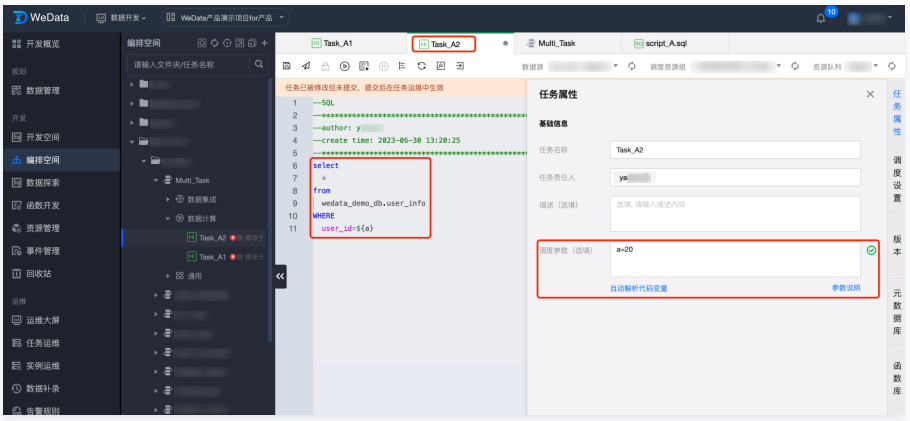


- Task A1:

©2013–2025 Tencent Cloud. All rights reserved.

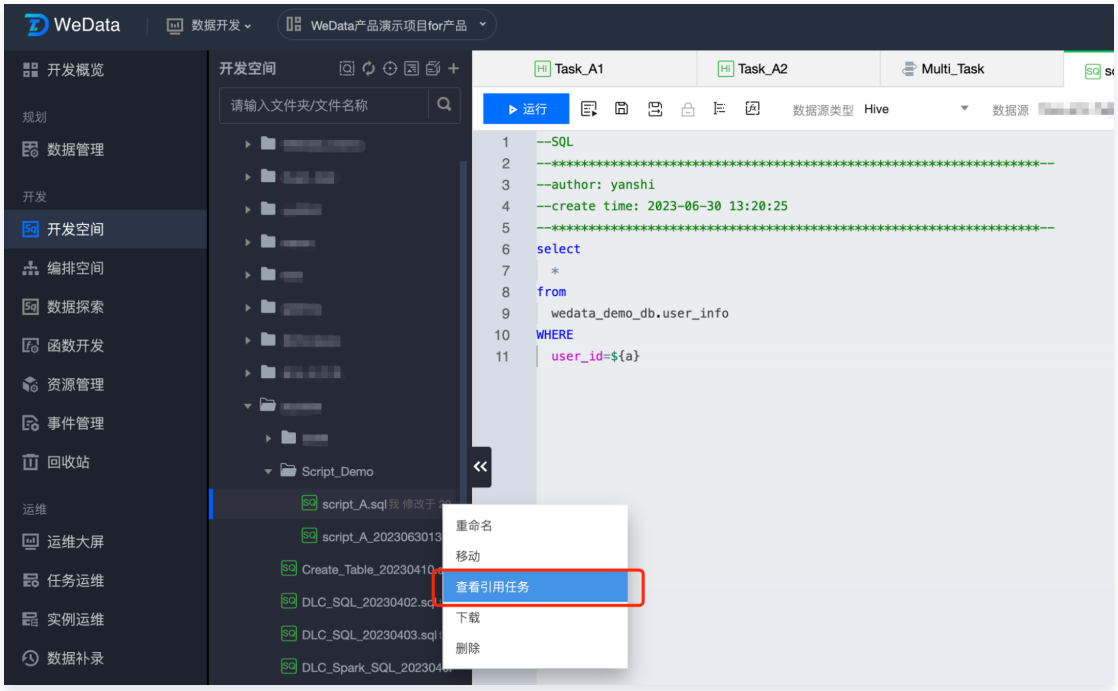


- Task A2:



3. Different scheduling variables can be set in the Orchestration Space's schedule variables to achieve different logic execution for the same script in different tasks.

3.1 View the feature of the Development Space script task to see the tasks referencing the current script.



3.2 Reference tasks:



Note

Modifying Script A in the Development Space will not affect the tasks already submitted for scheduling. If you need to modify the running code, you need to resubmit the tasks in the Orchestration Space.

©2013–2025 Tencent Cloud. All rights reserved.

Recycle Bin

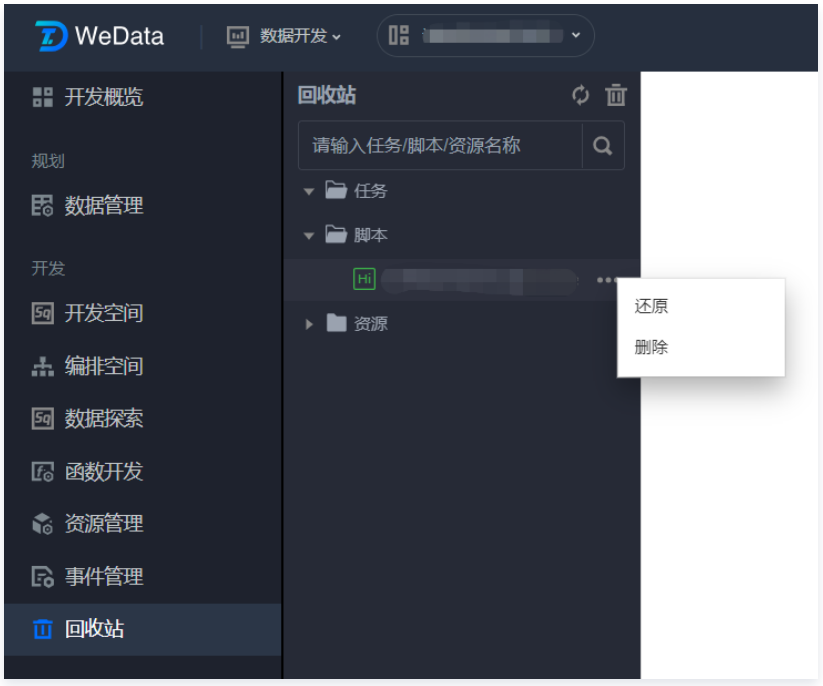
Last updated：2024-11-25 15:01:43

Description of the Feature

The Recycle Bin supports recovering deleted scripts, tasks, and resources from Development Space, Orchestration Space, and Resource Management, categorized into different directories.

Operation step

1. Log in to WeData Console.
2. Click **Project List** on the left side, select the corresponding project, and click to enter the **Data Development** module.
3. Click **Recycle Bin** in the left menu and then click Recycle Bin to view deleted tasks under the current project. You can view task details, restore tasks, or delete tasks as shown below:



4. By clicking the icon on the right side of the Recycle Bin, you can also empty the Recycle Bin directly.

> **Note:**
> Deleted resources can only be seen in the Recycle Bin by the corresponding owner.

©2013-2025 Tencent Cloud. All rights reserved.

Task Publishing
Cross-project Cloning

Last updated: 2024-11-25 15:02:22

WeData provides the ability to clone data development-related objects across projects. By using cross-project cloning, you can clone tasks, functions, and other objects created in one project to another project.

Application scenario

1. Development and production isolation scenarios: On public cloud, we highly recommend Project Standard Mode for development and production isolation. If it is necessary to isolate the development environment and production environment using different projects, cross-project cloning can be used.
2. After developing and testing in a project for a period of time, you may want to clone the existing test tasks to another project.

Use Limits

Note:

- The currently supported task types include: Offline Synchronization, HiveSQL, SparkSQL, PySpark, Spark, MapReduce, Impala, Trino, DLC SQL, DLC PySpark, TCHouse-P, TCHouse-X, TCHouse-X SQL, TDSQL-PostgreSQL, Python, Shell, JDBC SQL.
- Currently, only projects within the same Region can be cloned. Cross-Region cloning is not yet supported.
- Cross-project cloning does not support scenarios where tasks in the source project depend on tasks in the target project. If such dependencies exist, it may cause dependency chaos in the target project.
- For cross-project cloning, it is recommended to perform one-to-one cloning between the source and target projects. Cloning multiple source projects to a single target project may result in duplicate tasks from multiple source projects being overwritten in the target project.

Case Explanation

Taking the EMR engine as an example, this section introduces how to use cross-project cloning to clone objects from the source project to the target project. The principle of cross-project cloning involves selecting corresponding tasks and other objects in the source project and then mapping the attributes of those tasks and objects to the target project based on the chosen target project. Finally, tasks are created in the target project according to the mapping results.



©2013-2025 Tencent Cloud. All rights reserved.

Environment preparations

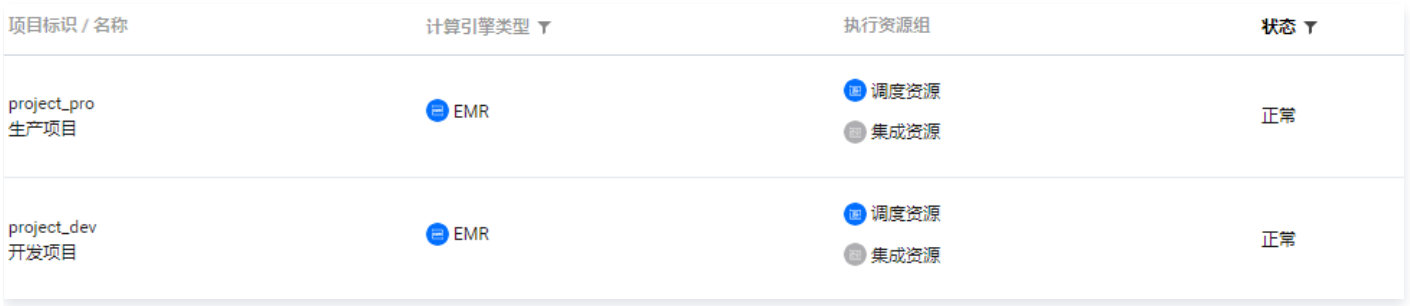
Compute Engine Side (EMR)

Create two clusters in EMR: Development Cluster (EMR-WeData-Dev) and Production Cluster (EMR-WeData-Pro).



WeData Side

Create two projects: Source project (project_dev) for development and Target project (project_pro) for production.

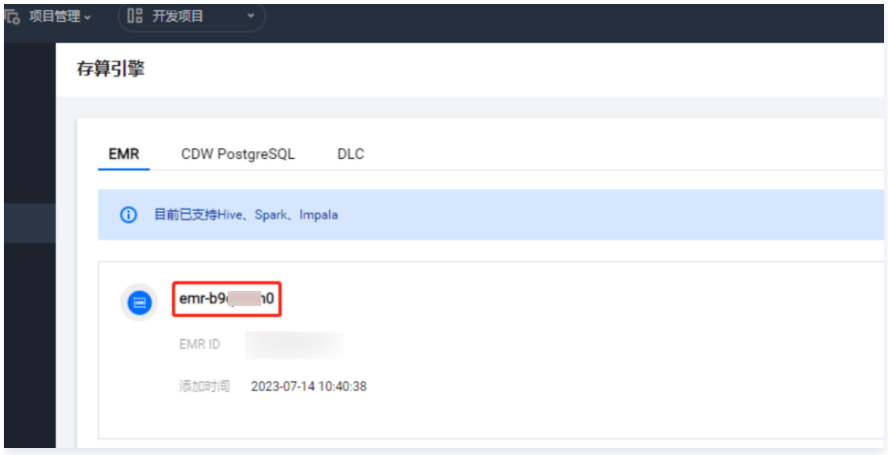


At the same time, create two scheduling resource groups: Resource Group for the development project (Guangzhou Schedule Resource Group -dev) and Resource Group for the production project (Guangzhou Schedule Resource Group -pro).

Source Project Configuration

- Storage and Computing Engine

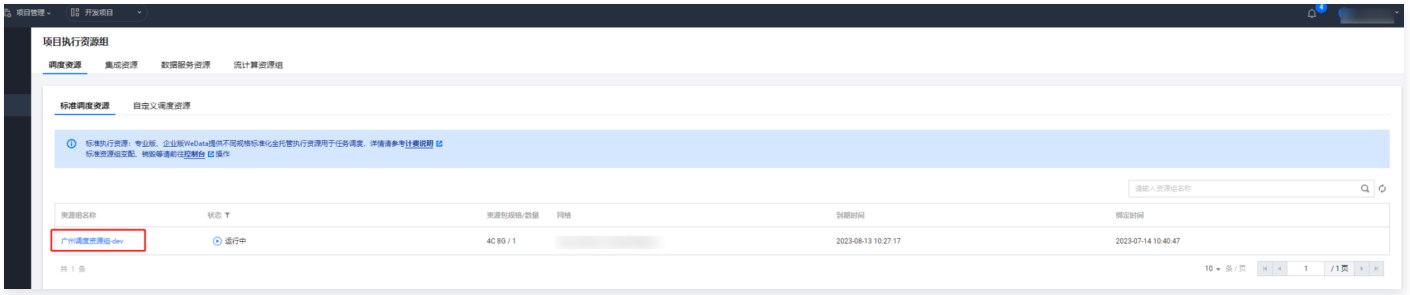
Bind EMR Cluster: EMR-WeData-Dev



- Execution Resource Group

Bind the scheduling resource group, name: Guangzhou Scheduling Resource Group -dev.

©2013-2025 Tencent Cloud. All rights reserved.



- Data Source

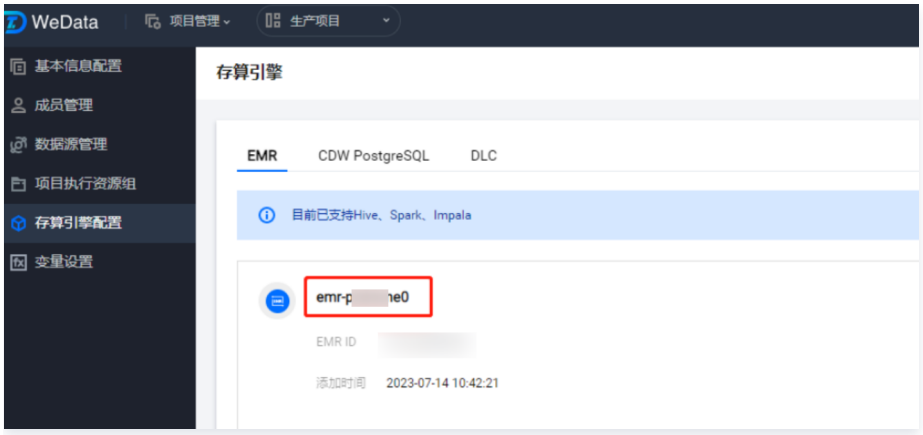
Automatically generate HIVE system data source, name: hive_emr-b9qs3kh0.



Target Project Configuration

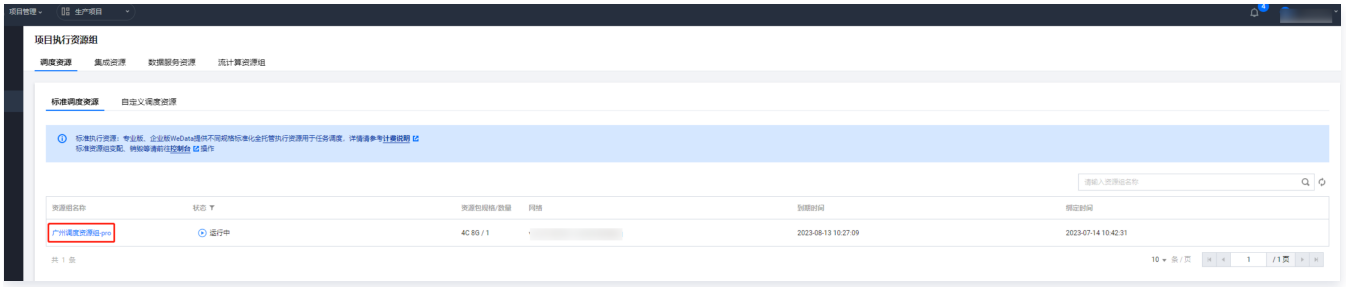
- Storage and Computing Engine

Bind EMR Cluster: EMR-WeData-Pro



- Execution Resource Group

Bind the standard scheduling resource group, name: Guangzhou Scheduling Resource Group -pro.



- Data Source

Automatically generate HIVE system data source, name: hive_emr-pb54xne0.



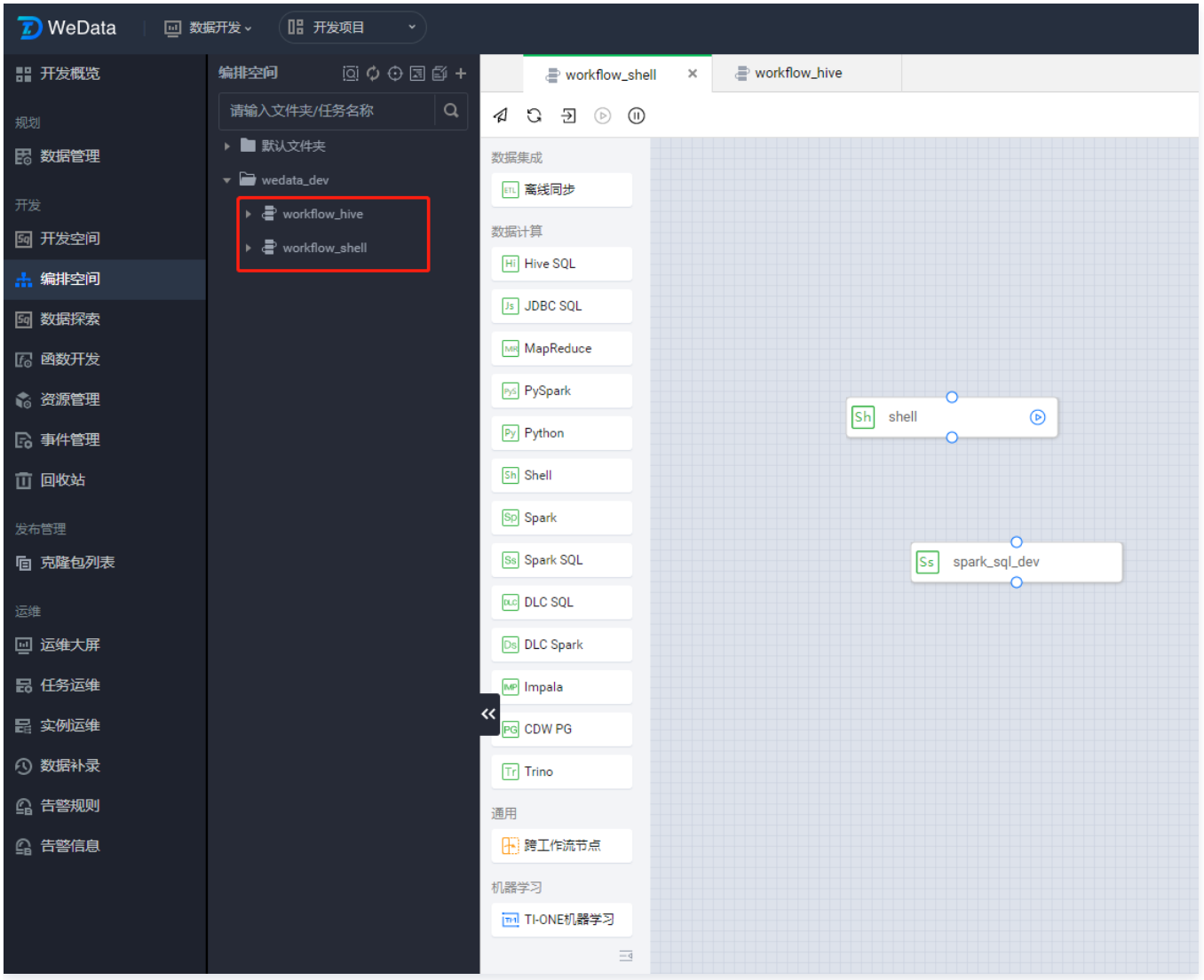
Data Development in Source Project

©2013-2025 Tencent Cloud. All rights reserved.

Upon completing the above environment setup, you can create workflows in the source project, develop tasks, debug tasks, configure scheduling, and submit tasks. For detailed workflows and task development, refer to WeData Task Development Documentation.

- Workflow

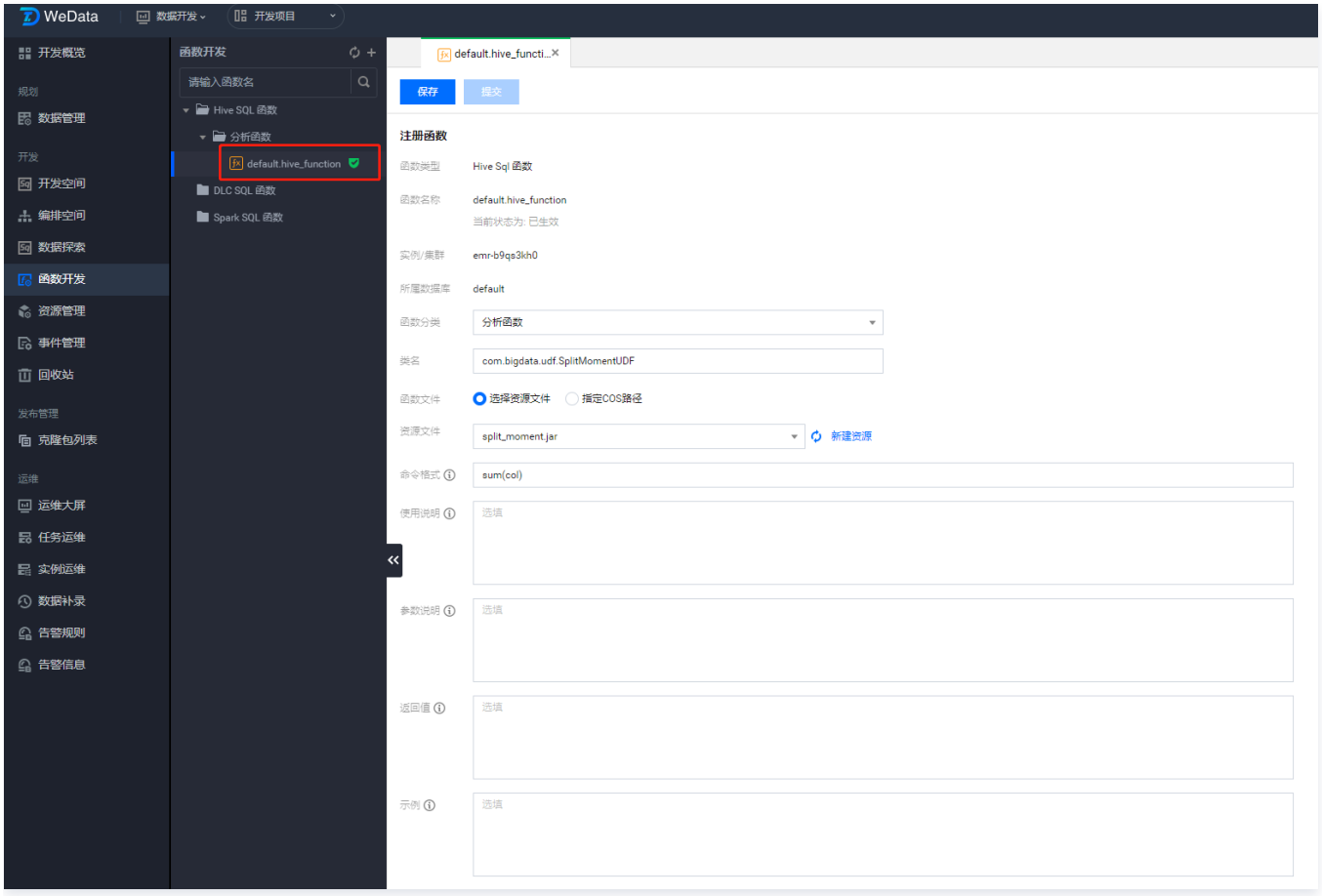
Create workflow_hive and workflow_shell workflows, and develop the corresponding tasks within the workflows. For example, in the figure below, tasks named shell and spark_sql_dev are developed in workflow_shell.



- Function

Create a function in Function Management, for instance, creating the hive_function in the default database of the source project.

©2013–2025 Tencent Cloud. All rights reserved.



- Project Variables

Create scheduling variables available for the project in the source project's project variables. For example, here a date variable yesterday with the value ${yyyy-MM-dd-1d} is created in the development environment.



- Event

Create the events needed for scheduling tasks in Event Management of the source project. For example, here, the event_1 event is created.



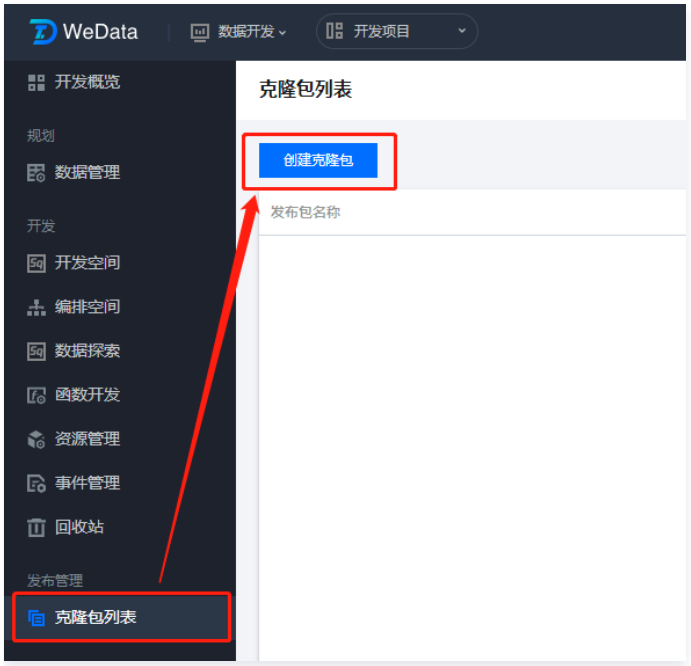
After completing task development in the development environment, use the cross-project cloning feature to clone the tasks from the development environment to the production environment.

Cross-project Cloning Operation Process

Create Cloning Package

1. Enter Development Project, click **Create Clone Package**, and go to the Create Clone Package page.

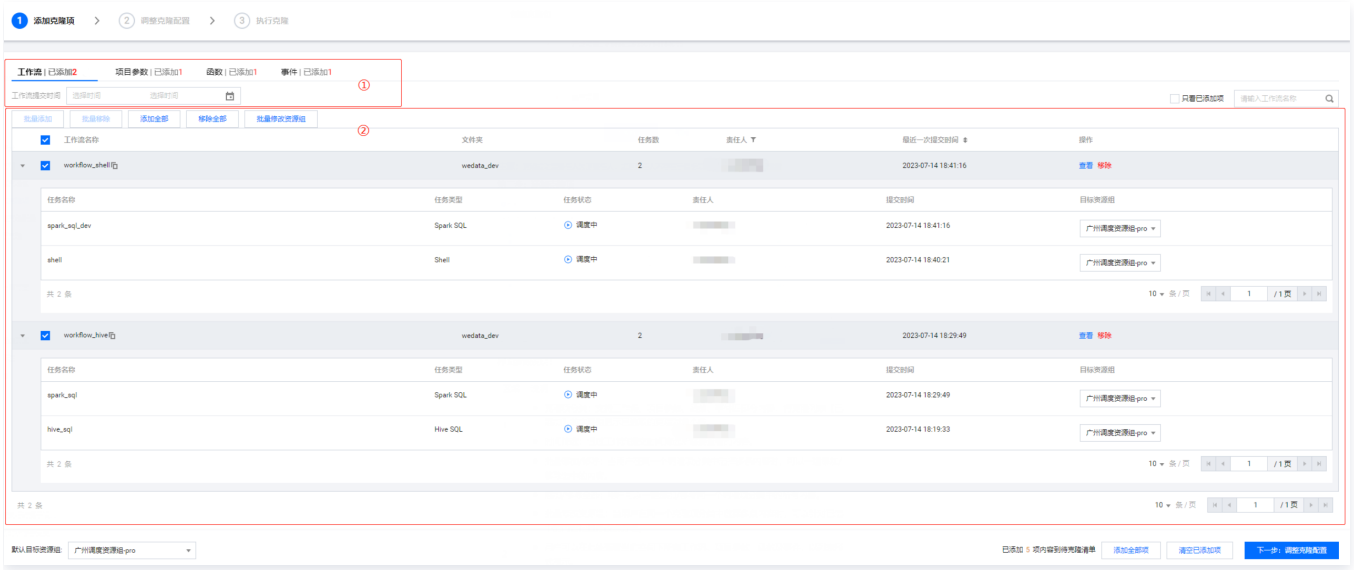
©2013-2025 Tencent Cloud. All rights reserved.



2. Fill in the Clone Package Name and Target Project (the source project defaults to the current project) and select Workflow Mode, click **Start Configuration**, to jump to this Project Parameter Details Page.



Step 1: Add Clone Items



Feature description:

©2013–2025 Tencent Cloud. All rights reserved.

Region	Description
1	• Clone Item Categories: Workflow/Task, Project Parameters, Functions, and Events. The List to be Cloned **will show the number** of selected clone objects. • Time Filtering: Filter content by workflow submission time to view the required content.

2	• Batch Add/Remove: When a user selects multiple items in the same clone item classification, they can add or remove the selected items with one click.
	• Add/Remove All: Users can add or remove all content within the same clone object classification in the project with one click.
	• Batch modify resource groups: Select multiple tasks to batch modify resource groups.
	• Two operations are supported: Viewing and To be added
	  ○ Viewing: Users can click **View** to navigate to the details of the corresponding workflow, project parameters, functions, and events.
	  ○ To be added/Remove: Users can individually add or remove workflows, project parameters, functions, and events from the list.

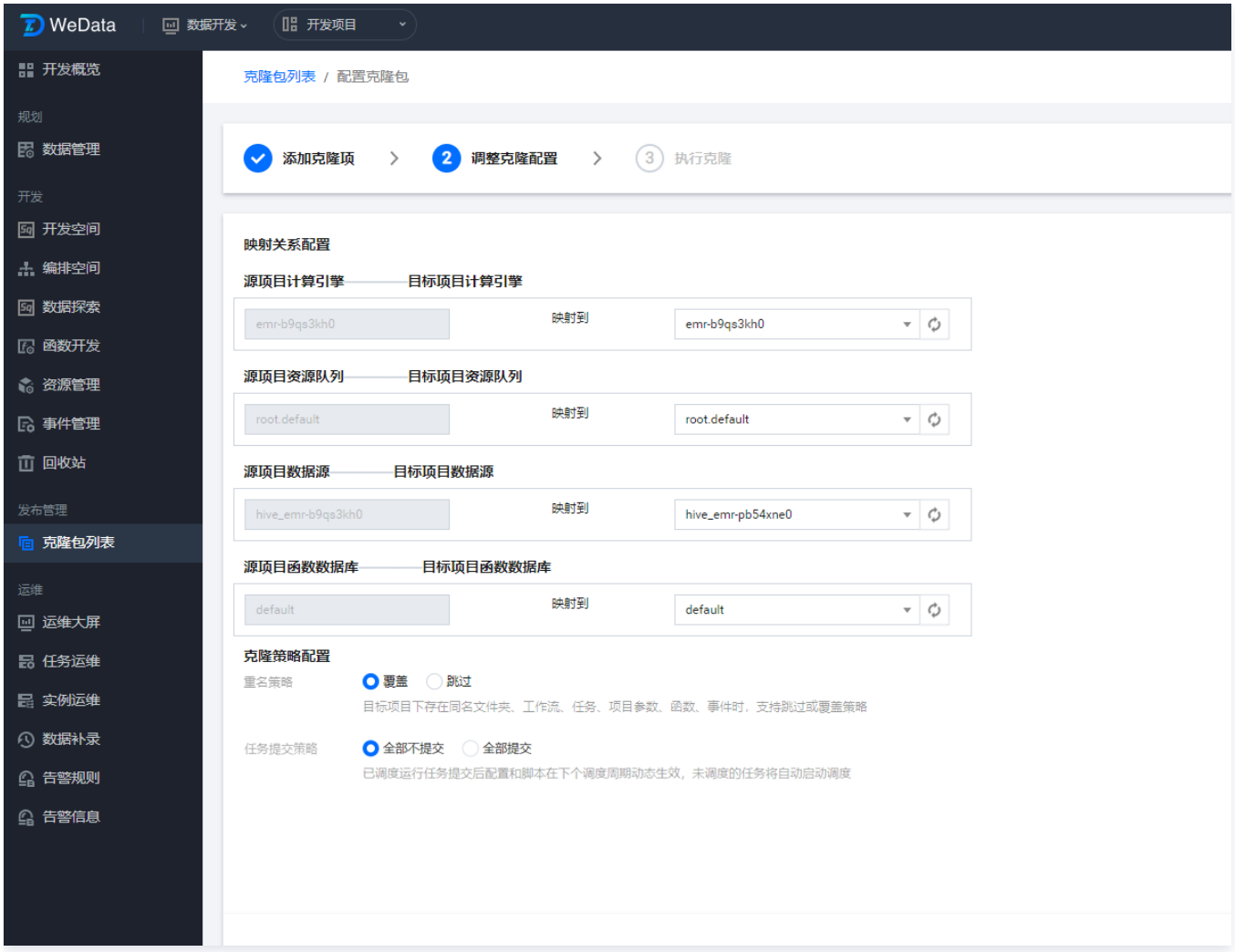
After adding the content to be cloned, select a default scheduling resource group and click **Next: Adjust Cloning Configuration**

Step 2: Adjust clone configuration

In the cloning configuration, you need to configure the corresponding mapping relationships, including: compute engine, resource queue, data source, database, etc.

- Compute Engine
  - The source project compute engine displays the compute engines used in the task, modifications are not supported.
  - The target project compute engine defaults to matching compute engines of the same type and name, but users can choose other engines of the same type.
- Resource Queue
  - The source project resource queue displays the resources used in the task, modifications are not supported.
  - The target resource queue defaults to matching resource queues of the same type and name, but users can choose other queues of the same type.
- Data Source
  - The source project data source displays all data sources used in the task, modifications are not supported.
  - The target data source defaults to matching data sources of the same type and name, but users can choose other data sources of the same type.
- Function Database
  - The source project function database displays all the databases used in the task, modifications are not supported.
  - The target project's Function Database defaults to matching function databases of the same type and name, but users can select other databases.

©2013–2025 Tencent Cloud. All rights reserved.



Clone Strategy

- Name Conflict Strategy: If there are folders, workflows, tasks, project parameters, functions, or events with the same name under the target project, the skip or overwrite strategy is supported.
- Automatic Task Submission:
  - Yes: When the user selects automatic task submission, the task is automatically submitted to the scheduling state, and the corresponding instance is generated in the next scheduling cycle.
  - No: When the user does not select automatic task submission, tasks that are not scheduled need to be manually submitted after cloning is completed.
- Auto-submit project parameters, functions, and events by default.

Step 3: Simulated Cloning

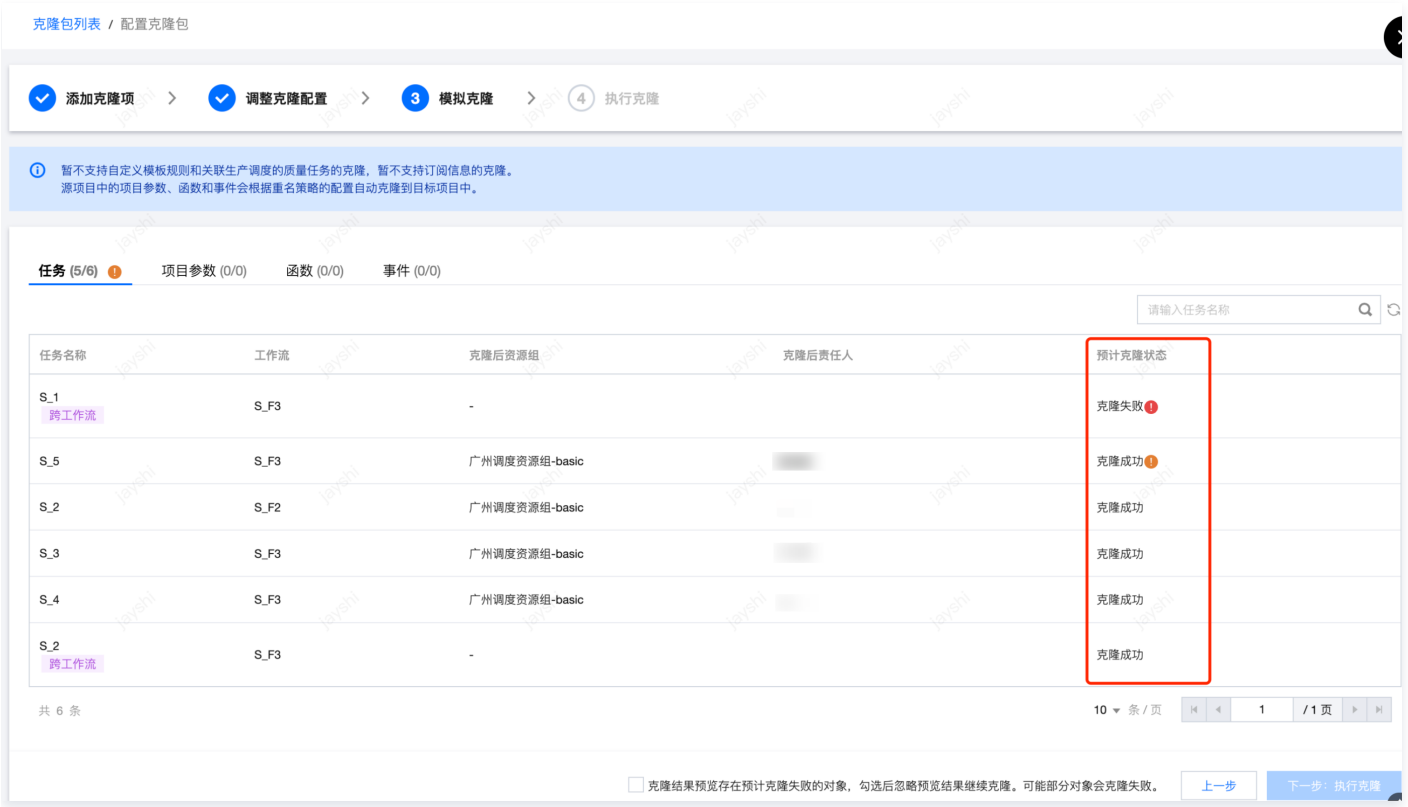
Simulated cloning performs a preliminary simulation cloning on the cloned objects selected by the user, previewing potential failure cases and missing dependencies that may occur after cloning, helping users to readjust the task attributes before cloning to avoid cloning risks.
Simulated cloning provides the expected cloning status:

- Cloning Failure: The reason for the current object's simulated cloning failure will be displayed, such as not finding the corresponding task of the cross-workflow node in the target project.
- Cloning Success (with Alerts): The task will be cloned successfully, but the upstream tasks and upstream events might not be found in the target project.
- Cloning Success (without Alerts): Cloning was successful, and simulated cloning was also successful.

1. Simulated cloning only performs a pre-check using the current task state. Since there might be changes in the source and target projects during the cloning process, it does not guarantee these issues will occur after cloning. Therefore, users can

©2013-2025 Tencent Cloud. All rights reserved.

continue the cloning process by clicking the checkbox below for cases with issues in simulated cloning.



2. click **Next: Execute Cloning** to proceed with the cloning.

Step 4: Execute Cloning

After executing cloning, wait for a while to view the final clone report.

- View Report: The clone report details the clone overview, clone results, and clone details.





- Clone Overview: By default, it displays information such as the clone package name, time, and mode.

©2013–2025 Tencent Cloud. All rights reserved.

- Clone Results: The total numbers of workflows/tasks, project parameters, functions, and events are displayed in table form, along with the numbers successfully cloned and those that failed.
- Clone Details: The specific contents and status in the clone package are displayed separately according to workflows, project parameters, functions, and events, along with the reasons for object cloning failures.

Dependency Explanation

Cross-project cloning tasks involve handling more complex dependency relationships. Here, we introduce the principles for handling dependencies in cross-project cloning.

Basic Assumptions

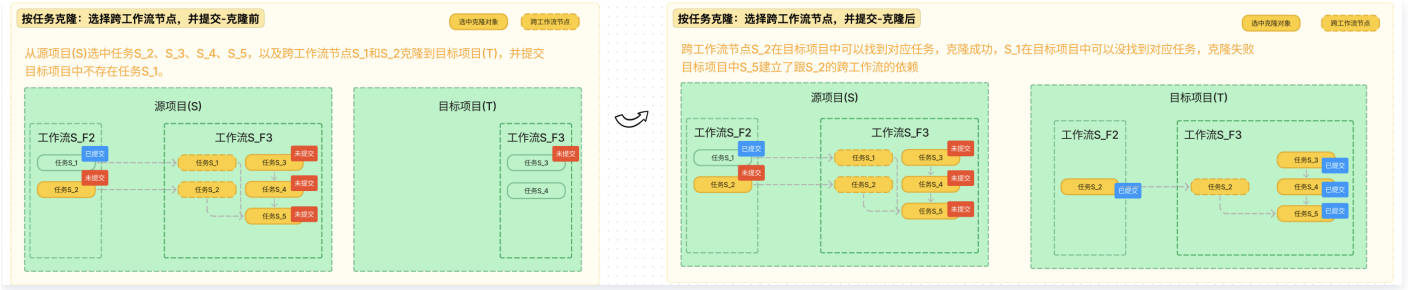
- Source Project Name: (S)Source
- Target Project Name: (T)Target
- Other Project Names: (O)Other
- Cross-Workflow Node: Cross-workflow nodes within the current workflow.

Task Cloning Guidelines

1. Whatever you select is cloned without adding extra information or changing the task's attributes.
   - If cross-workflow nodes are not selected, they will not be cloned, and cross-workflow dependencies will not be cloned.
   - Cloned cross-workflow nodes remain cross-workflow nodes, while non-cross-workflow nodes remain non-cross-workflow nodes. HiveSQL tasks remain HiveSQL tasks after cloning.
2. Dependencies on a third project's dependency relationship do not change with cloning.
   - Dependencies on Other Project (O) remain dependencies on Other Project (O) after cloning.
3. Within the same project, the retention of cross-workflow dependencies depends on whether the cross-workflow nodes and their corresponding tasks exist in the target project.
4. Within a workflow, the existence of dependencies depends on whether the upstream tasks exist.
5. Criteria for duplicate name determination: Task name + Workflow name must be identical to be considered the same task.
6. Cross-project cloning only copies tasks and cannot ensure that tasks successfully scheduled in the original project will be submitted successfully in the target project. For example, if the upstream task fails in the target project, the downstream task will also fail.

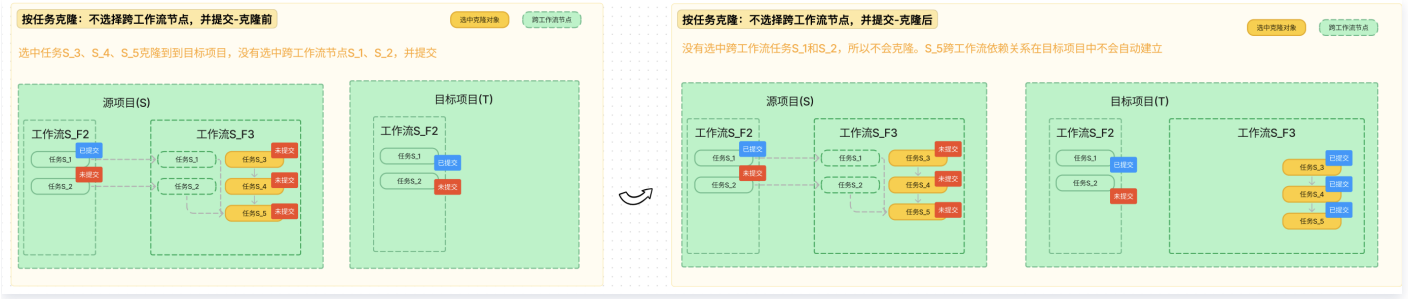
Common task dependency scenarios

Clone by task: Select cross-workflow cloud data disks

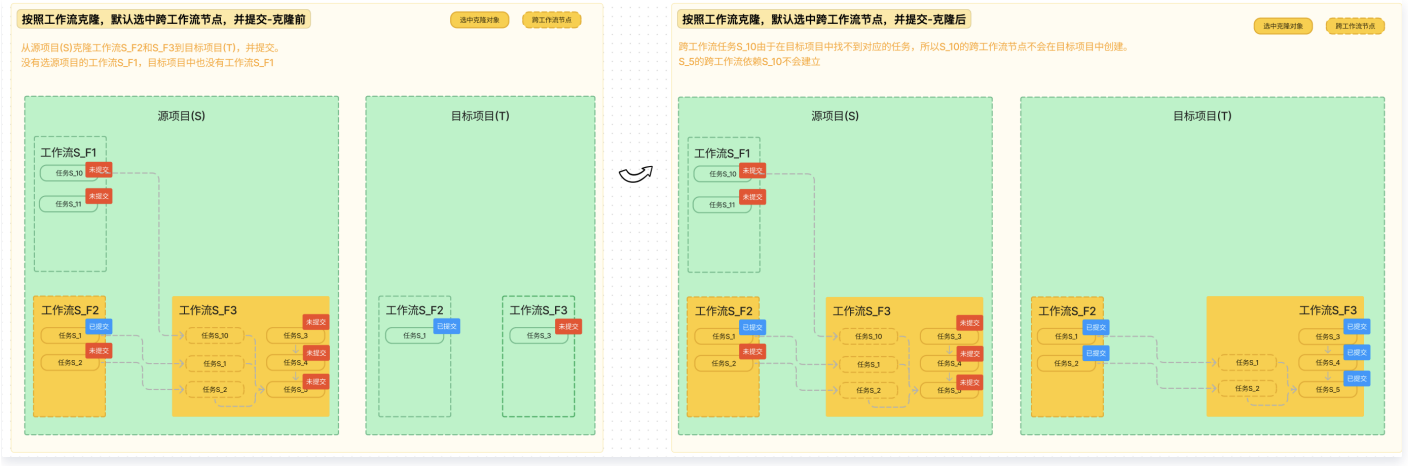


Clone by task: Do not select cross-workflow cloud data disks

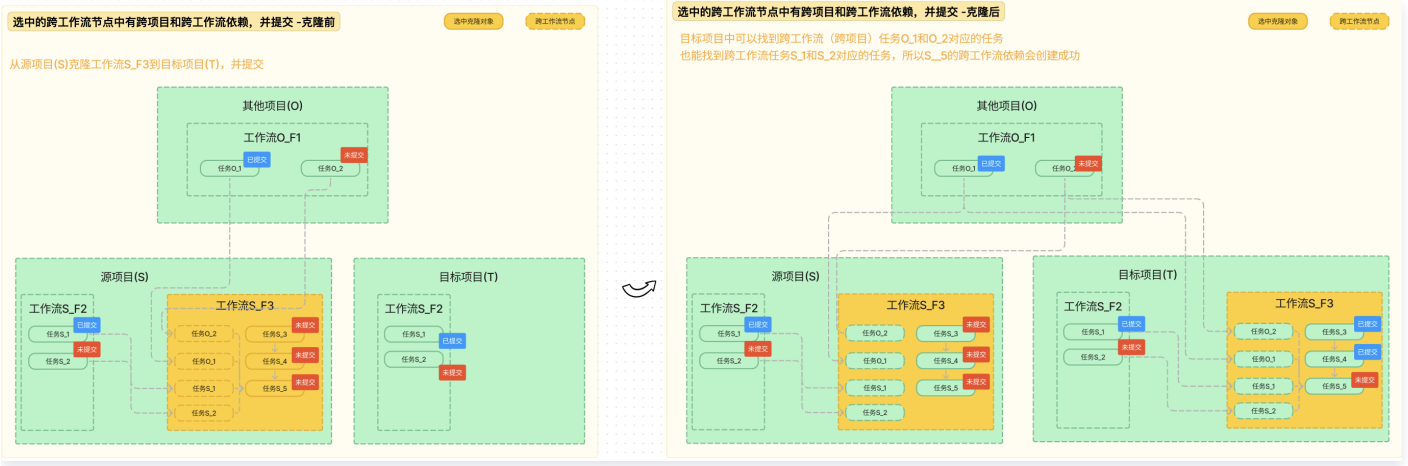
©2013-2025 Tencent Cloud. All rights reserved.



Clone by workflow: Cross-workflow cloud data disks in the workflow will be selected by default

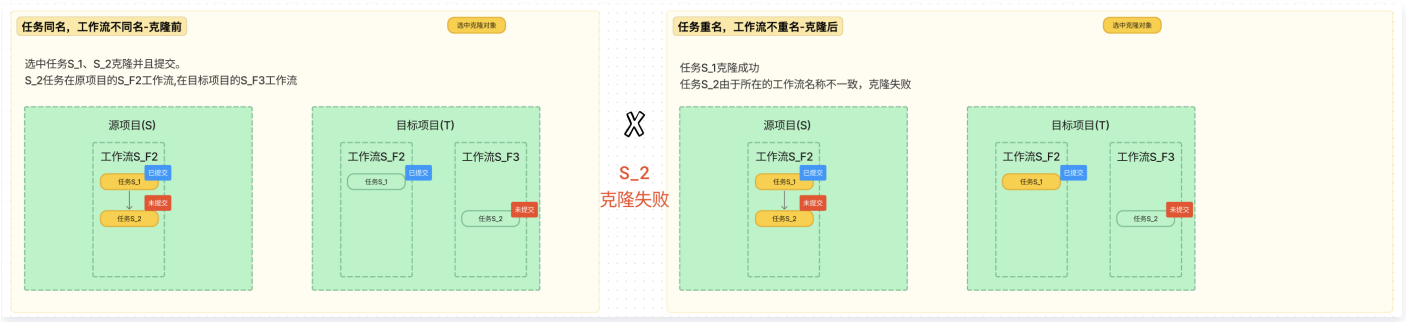


Selected tasks include cross-project and cross-workflow dependencies among cross-workflow cloud data disks

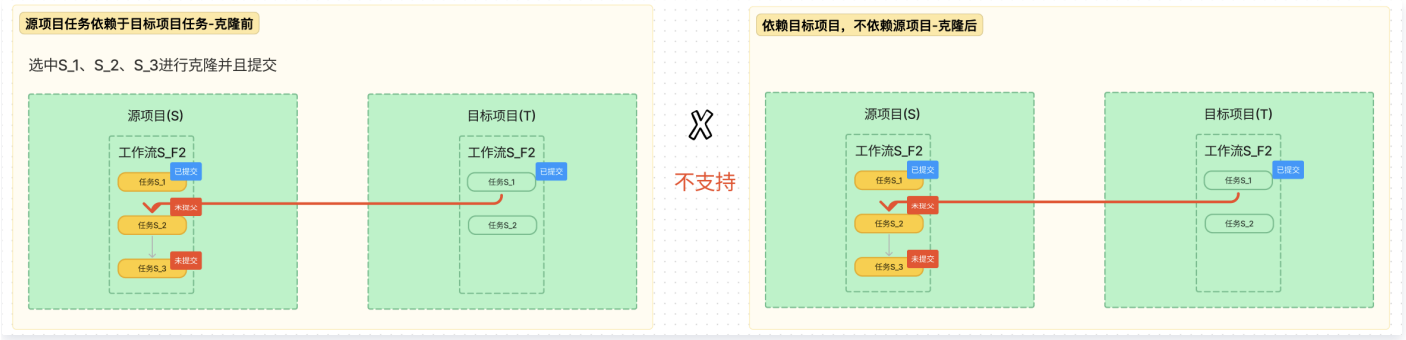


Same task name, different workflow name

©2013-2025 Tencent Cloud. All rights reserved.



Source project task depends on the target project task, this scenario is not supported



©2013–2025 Tencent Cloud. All rights reserved.

Release Center

Last updated：2024-11-25 14:45:06

The WeData Release Center provides the object release management feature after enabling Standard mode in the project space. The Release Center mainly offers **Pending Release Object Management** and **Release Package Management** features.
In WeData, objects related to data development will automatically enter the pending release object list after being edited, saved, and submitted. Create a release package from the pending release object list, and once the release package is approved, it will enter the production environment.

> **Note**
> The Release Center is currently in an invite-only testing phase. If you do not see the Release Center entry in your interface, you can click Applying for Trial to request access.

Submission and Release Process

After enabling Standard mode in the WeData project, related objects' submission and release processes are as follows:



Task

Tasks are completed and tested in data development. Tasks and workflows submitted enter the pending release list.

Function

After a function is created in the development library, it can be directly used in the development database. The function submission will enter the pending release list. Once published successfully, the function will be created in the production database.

Workflow

After a workflow is created, a corresponding workflow will be automatically generated in the production environment. After modifying the workflow parameters and application parameters, you can separately submit the workflow version to the pending release list. Once published successfully, the workflow parameters will take effect in the production environment.

Project Parameters

After creating and modifying project parameters, the development environment values of the parameters will take effect immediately after modification. For the production environment values, you need to submit the project parameters to the pending release list. Once published successfully, the production environment values will take effect.

©2013-2025 Tencent Cloud. All rights reserved.

Resources

After creating and re-uploading resources, you need to submit them. After submission, they will enter the pending release list. Once published successfully, the resources will take effect in the production environment.

Database Table

After a data table is created in the development library, it can be directly used in the development database. The data table submission will enter the pending release list. Once published successfully, the data table will be created in the production database.

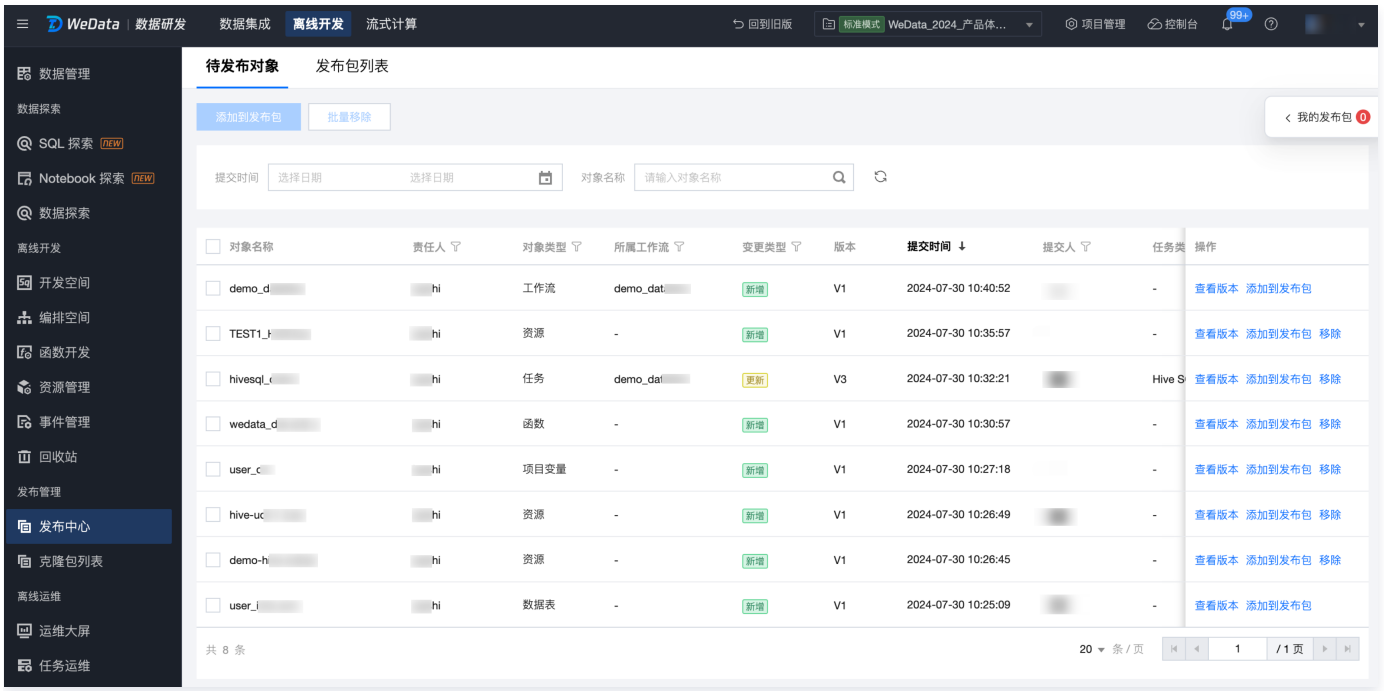
Pending Release Objects

After submitting, objects will enter the pending release list. The pending release object list includes tasks, functions, workflows, project parameters, resources, and data tables.

- When new objects are submitted to the pending release object list, the change type is **new**.
- If an object is modified continuously and new versions are continuously submitted, the new version will overwrite the previously submitted version in the pending release object list. The change type is **update**.
- If an object is deleted, the deleted object will enter the pending release object list. The change type is **delete**.

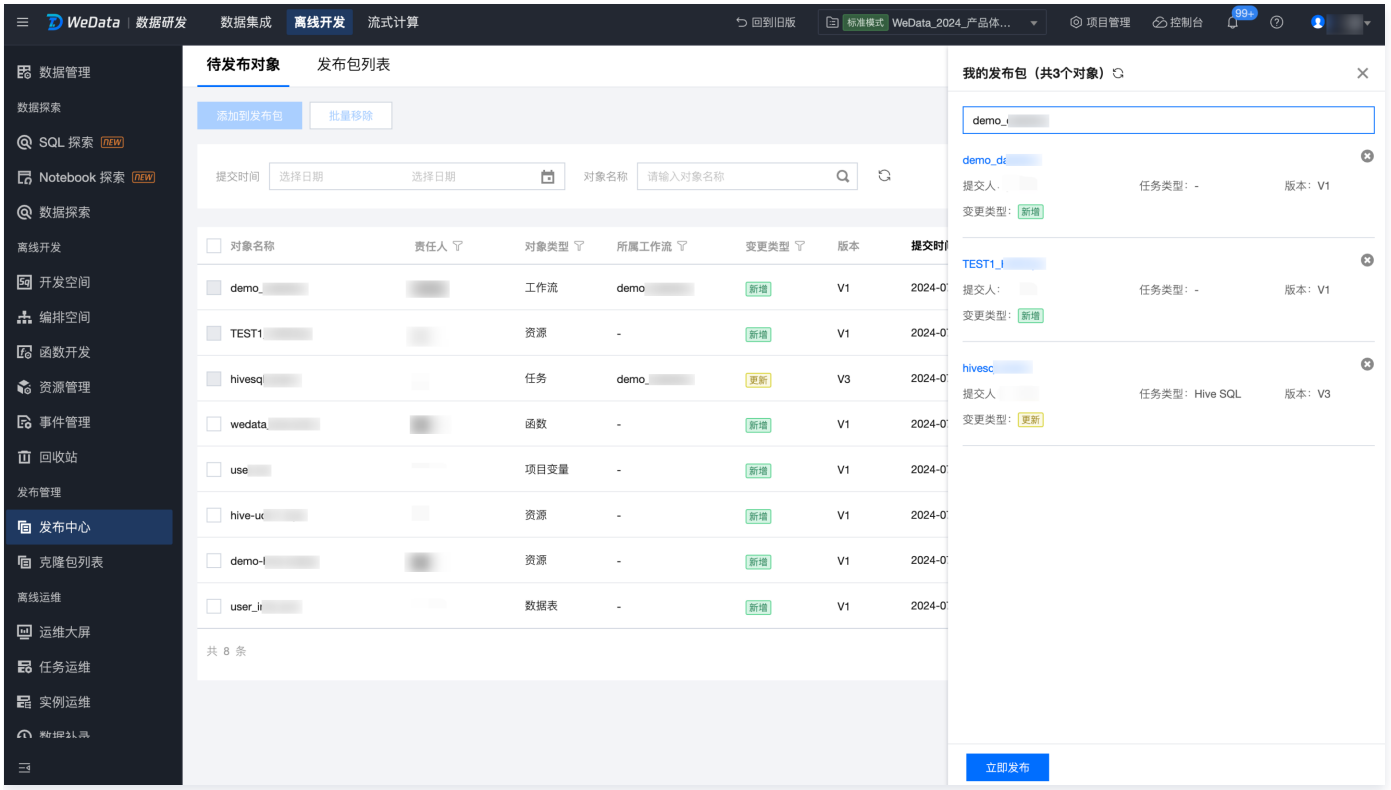
Operation step

1. Go to **Release Center** > **Pending Release Objects** list page, select the object, and then click **Add to Release Package**. The pending release object will be added to **My Release Packages** on the right. The default name of the release package is the first object in the package. Users can modify it according to their business scenario.



2. Click **Publish Now**, and the release package will automatically enter the release package list.
   - In the release package list, if approval is not required, it will automatically enter the release status. If approval is needed, it will enter the release status after approval.
   - During the release process in a release package, publication may fail due to issues like missing database tables or duplicate table names. If any object in the release package fails to release, the entire release package will fail. In this situation, you can choose to remove some objects or fix the issues and then release again.
   - After a successful release, you can view the details of the release package, which includes the version records of the objects at the time of release.

©2013-2025 Tencent Cloud. All rights reserved.

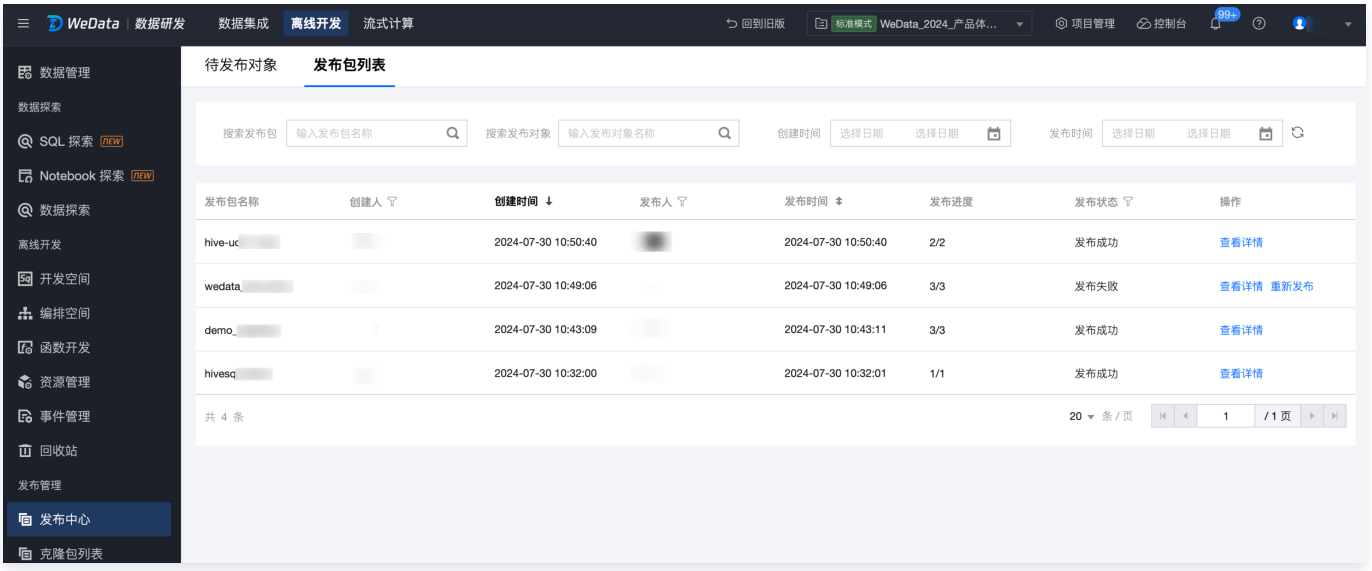


Release Package Management

The release package list provides features for release package management including search, status view, and details view of the release package.

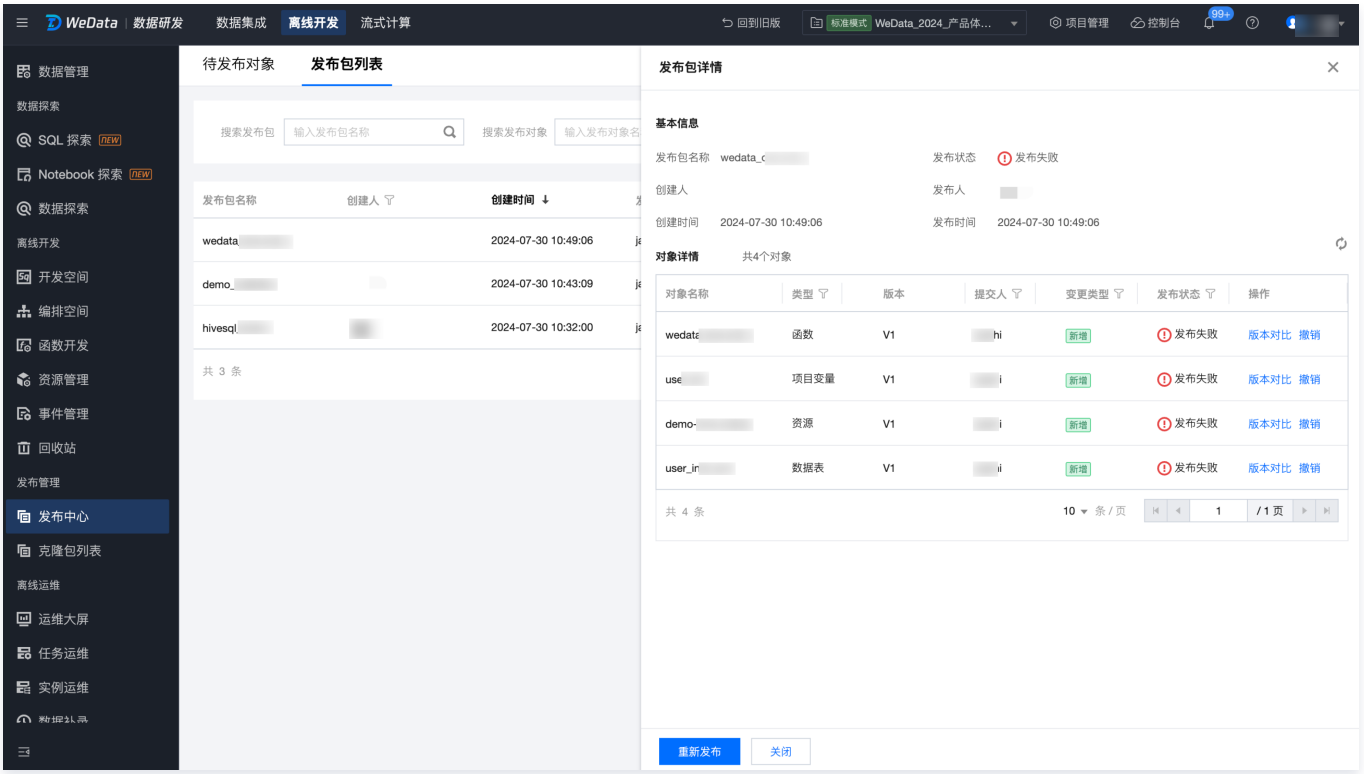
By default, the release package requires approval before it can be released. The approval process for the release package can be configured in Project Management – Basic Information Configuration – Approval Configuration. If the creator of the release package inputs the approver's name, it will directly enter the release state without needing approval.

- Go to the **Release Package List** page, in the operation column, click **View Details** to view the version and status of objects in the release package.



- Go to the **Release Package List** page, in the operation column, click **Revoke**. For a failed release package, you can re-release after fixing errors or revoke the failed objects and try to release again.

©2013–2025 Tencent Cloud. All rights reserved.



©2013–2025 Tencent Cloud. All rights reserved.

Task OPS
Operation Dashboard

Last updated: 2024-11-25 14:41:44

Enter the Operation Dashboard page

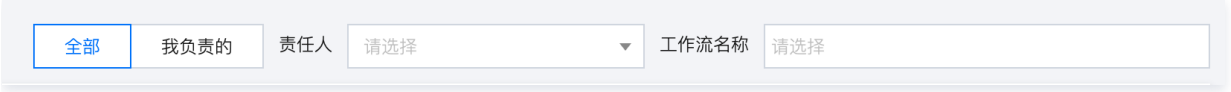
1. Log in to WeData Console.
2. Click **Project List** in the left menu to find the target project you want to **View Operation Dashboard**.
3. After selecting the project, click to enter the **Data Development** module.
4. Click **Operation Dashboard** in the left menu.

Operation Dashboard

The Operation Dashboard displays the overview information of data development tasks. It presents critical operation metrics in chart form, including Task Overview, Instance Overview, Instance Running Trends, Instance Duration Ranking, Scheduling Resources Usage, Scheduling Resources Usage Trends, Resource Group Instance Change Trends, Task Status Trends, Task Status Distribution, Task Cycle Distribution, and Instance Status Distribution.

Filtering

- All/My Tasks: Switch between statistics for all tasks under the project or tasks for which I am the Person in Charge.
- Person in Charge: Specify to view statistical information for tasks handled by certain individuals.
- Workflow name: Specify to view statistical information for certain workflows under the project.



Overview Information

The overview information summarizes key metrics for computing tasks and task instances.

- Task Overview: Statistics cover total task count, number of scheduled tasks, number of paused tasks, number of offline tasks, daily task count, hourly task count, and minute task count.



- Instance Overview: Statistics cover total instance count, number of running instances, number of instances awaiting execution, number of instances awaiting upstream, number of instances awaiting events, number of successfully run instances, and number of failed run instances.



Instance Running Trend

Supports viewing the trend chart of instance count changes over time. The trend can be viewed based on specific task types or instance statuses.

> **Note**
> For daily granularity data, the instance count at 12:00 of the day is displayed.

©2013-2025 Tencent Cloud. All rights reserved.



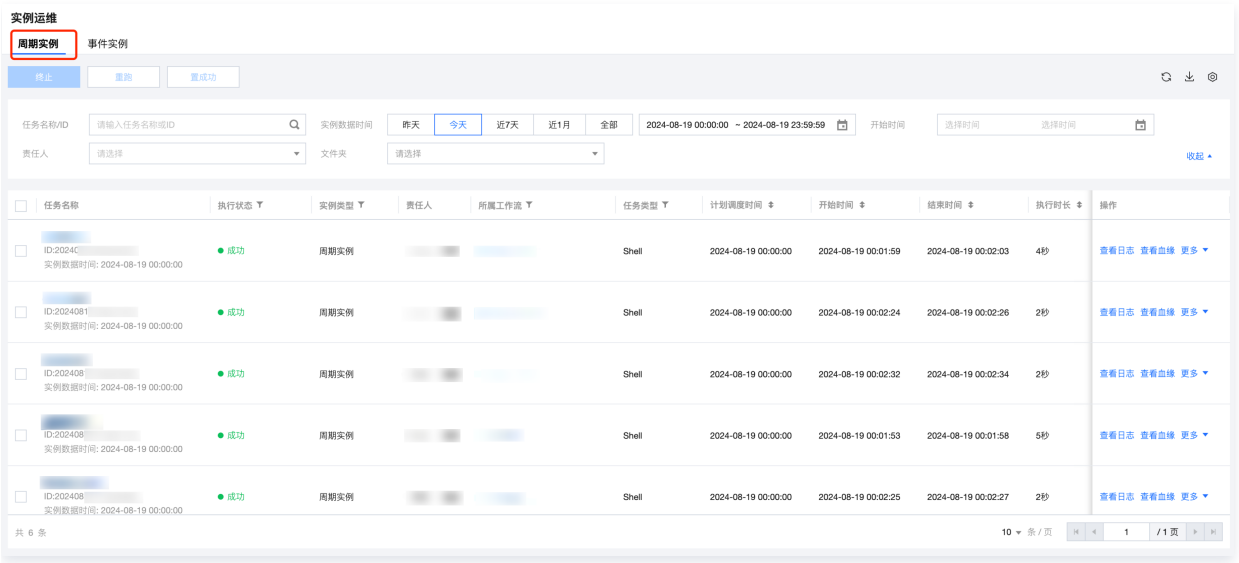
Instance Duration Ranking

- Instance Duration Ranking can be filtered by the last 12 hours, 1 day, 7 days, or 1 month time dimension to rank computing task instances based on waiting time and running time in descending order. It supports filtering by scheduling cycle type. Users can click **View More** to jump to a more detailed computing task instance page with comprehensive information.
- **Running Time Ranking:** By using the running duration ranking, you can identify the time consumption status of task instances, allowing for targeted performance optimization and overall efficiency improvement. You can locate resource-intensive computing tasks through the ranking, prioritize the execution of the most time-consuming tasks, or allocate more resources to enhance system computing efficiency. The ranking also serves as an indicator for evaluating the code quality of computing tasks, prompting redesign and optimization for tasks that frequently have long durations.
- **Scheduling Wait Time:** Scheduling wait time indicates the duration an instance spends in a waiting state before execution. It represents the time from when an instance is ready upstream to when it is dispatched to the execution machine. Instances may be in a waiting state due to insufficient scheduling resources or concurrency issues. You can sort by scheduling wait time.



Click **View More** to jump to the Instance Operation and Maintenance page. The Instance List is automatically sorted in descending order for easier access to detailed information.

©2013-2025 Tencent Cloud. All rights reserved.



Scheduling Resources Usage

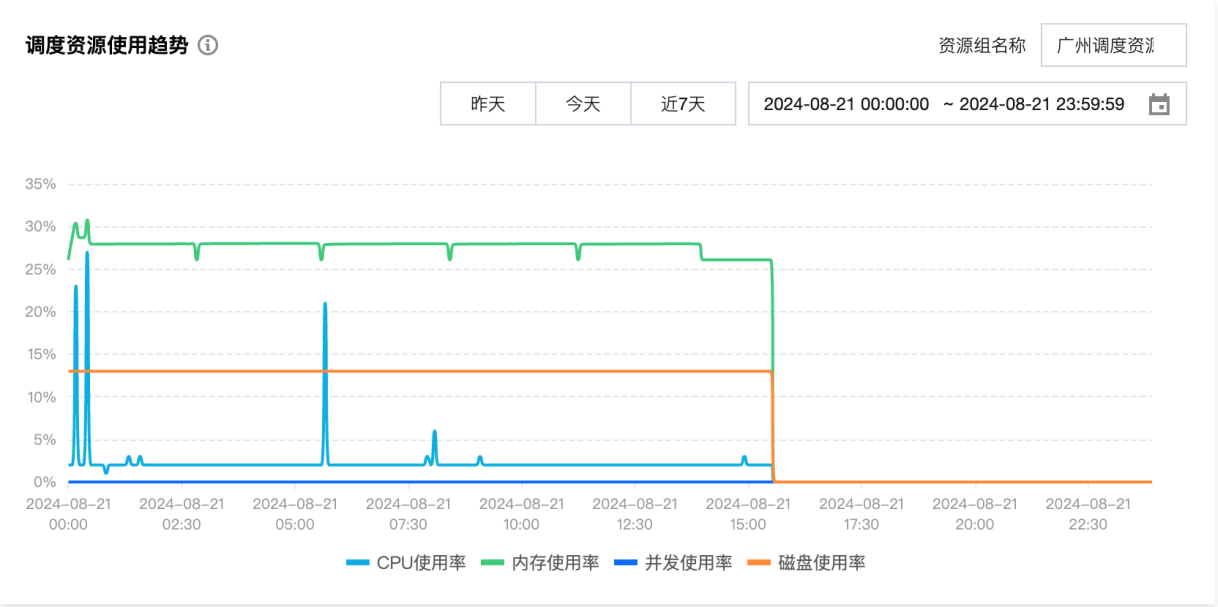
Displays the status and resource usage of scheduling resource groups bound to the current project. It supports viewing the status of scheduling resource groups, CPU usage, memory usage, concurrent usage rate, disk utilization, number of running instances, number of instances waiting to run, and number of tasks in scheduling. Clicking on the number of running instances, instances waiting to run, and tasks in scheduling allows you to view the corresponding task/instance list.



Scheduling Resources Usage Trends

Show the change trends of CPU utilization, Memory usage, Concurrency usage, and Disk usage of the Schedule Resource Group over time.

©2013–2025 Tencent Cloud. All rights reserved.



Instance Change Trend in Schedule Resource Group

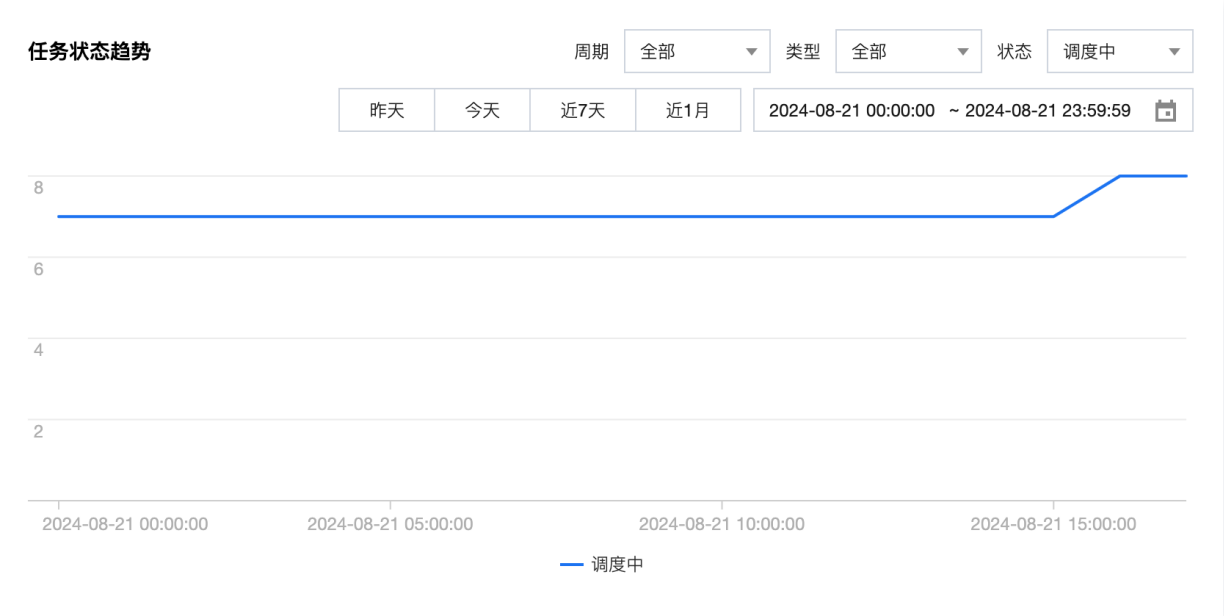
Show the change trends of the Running instance count and Number of Instances Awaiting Execution in the Schedule Resource Group over time.



Task Status Trend

Task Status Trend and Viewing shows the change trends of the number of tasks in different statuses over time. Supports filtering by Scheduling Period, Task Type, Task Status, and Time.

©2013–2025 Tencent Cloud. All rights reserved.



Task Status Distribution

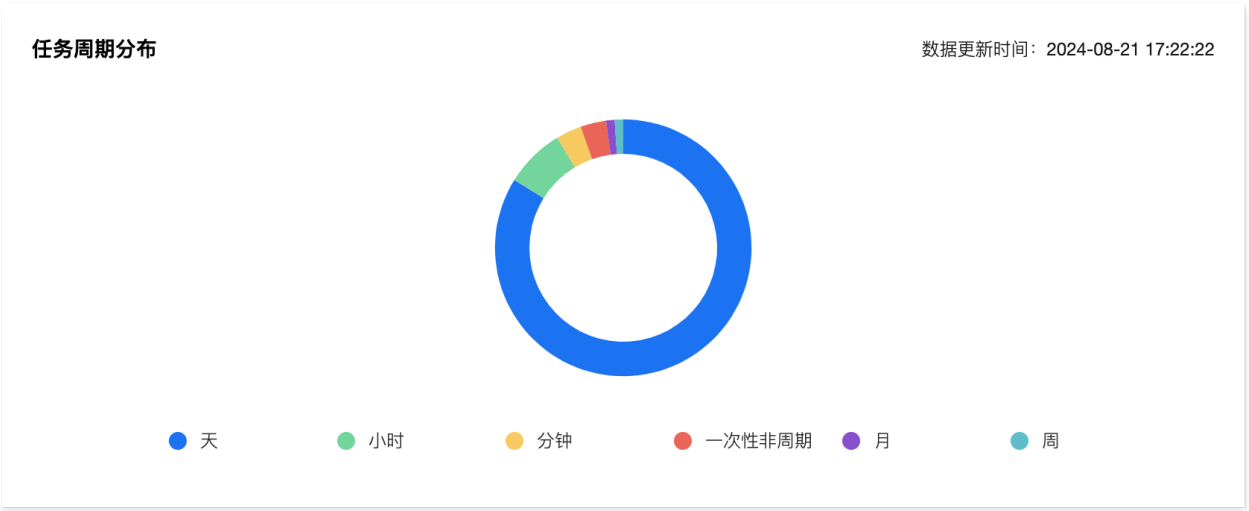
The Status Distribution Diagram uses a circular chart to show the proportion of current computing tasks in different statuses. Supports filtering by Task Type.



Task Cycle Distribution

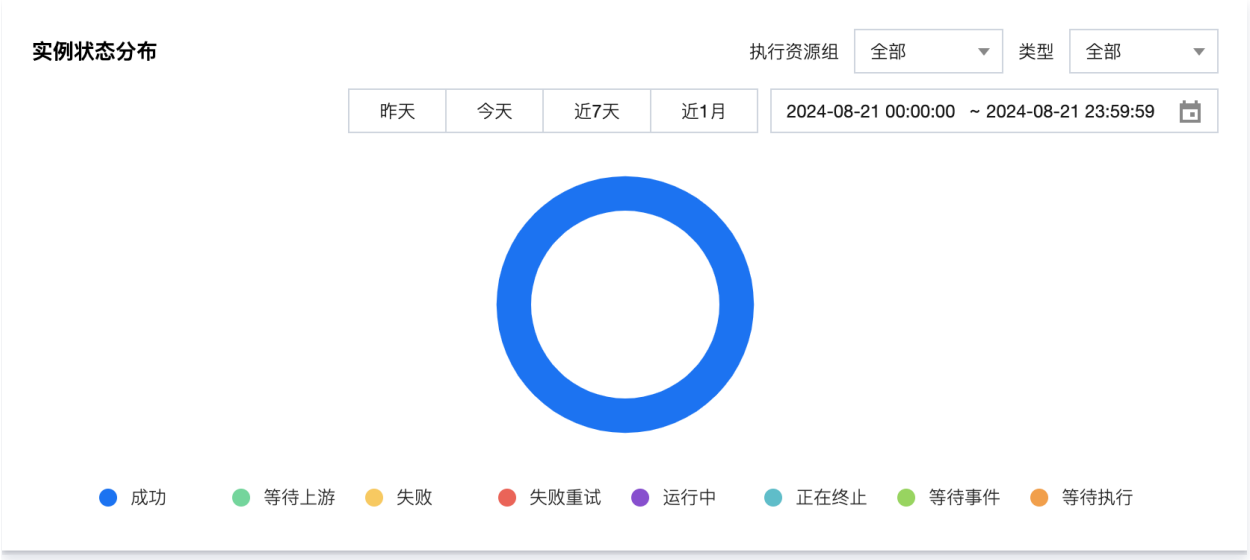
The Task Cycle Distribution Chart uses a circular chart to show the proportion of computing tasks with different scheduling cycles that are currently published, making it easier to observe the proportional relationship of different types of tasks and highlight key areas.

©2013–2025 Tencent Cloud. All rights reserved.



Instance Status Distribution

The Instance Status Distribution Chart uses a circular chart to show the proportion of current task instances in different running statuses. It supports filtering by execution resource group to view the running status distribution of task instances under different execution resources. By filtering through task types, you can see the running status distribution of task instances for different types of computing tasks. This allows for a quick understanding of task running statuses, and when instances are in a failed or other abnormal state, it helps to identify issues and promptly react for optimization and adjustment.



©2013–2025 Tencent Cloud. All rights reserved.

Task OPS

Last updated：2024-11-25 14:42:30

Enter the Task OPS page

1. Log in to WeData Console.
2. Click on **Project List** in the left-hand menu, and find the target project that requires **operational actions**.
3. After selecting the project, click to enter the Data Development module.
4. Click **Task OPS** from the left menu.

Task OPS Overview

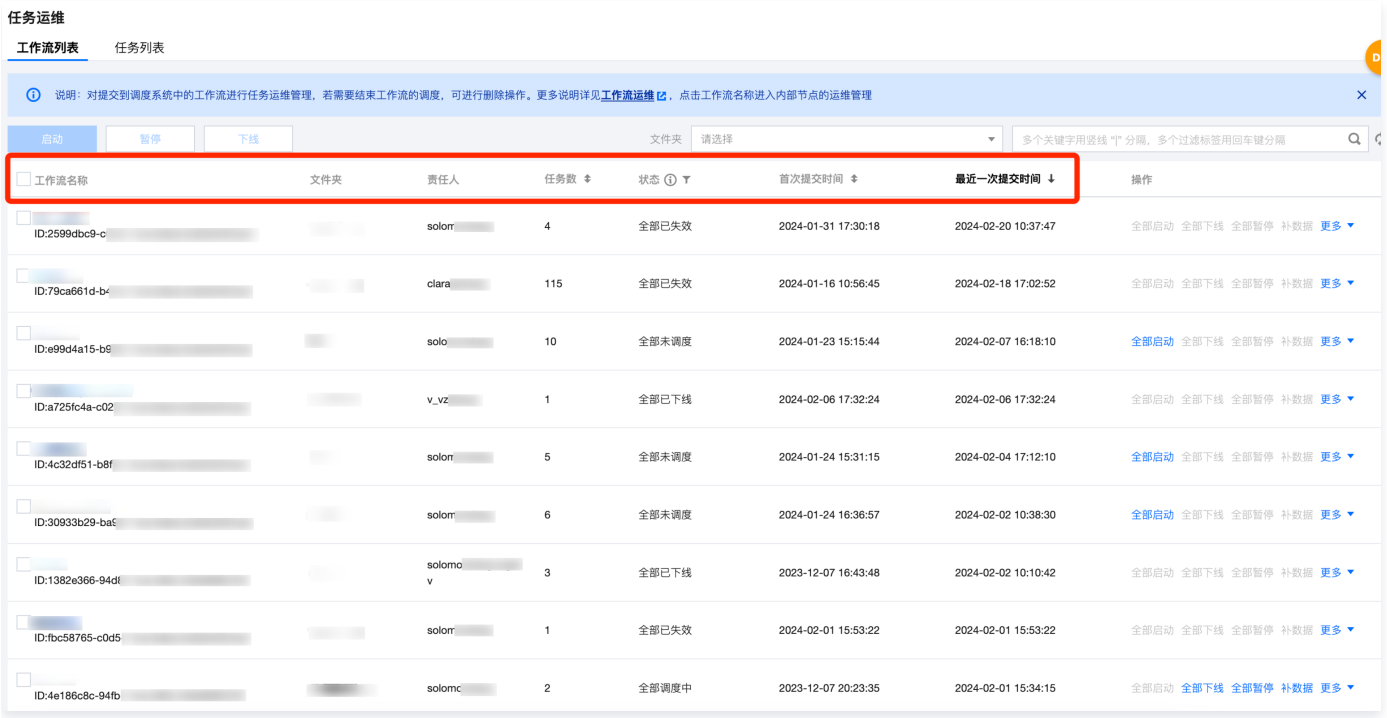
After users complete the configuration of computing tasks and the orchestration of task nodes in **Data Development > Orchestration Space** and submit the scheduling, they can manage and operate data workflows and computing tasks on the Task OPS interface, including operations such as starting, suspending, decommissioning, and deleting workflows and computing tasks. Other features include viewing workflow details, task execution history, upstream and downstream relationships between tasks, and task execution instances.

Workflow OPS

Workflow OPS list

Select the Workflow List tab on the Task OPS page. The Workflow List page displays data workflows that have been published and entered the OPS phase. In the list, you can view the current running status, scheduling period, and other information of workflows. A series of features allow you to maintain workflows, ensuring that data processing is efficient, stable, and reliable.

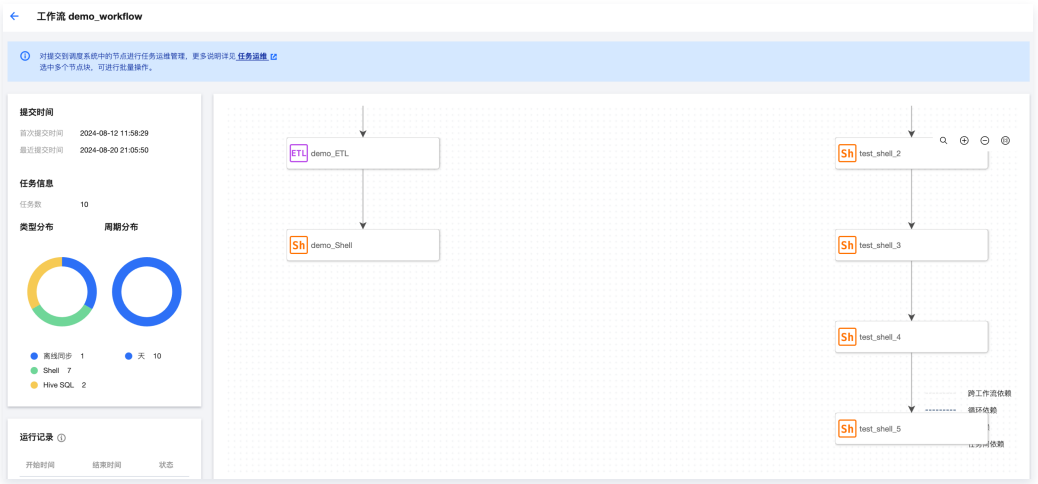
- Workflow list information:



Feature description:

Information	Description
Workflow name	Display the workflow name and ID, click to copy the corresponding workflow name or ID. The workflow ID is used for querying logs, troubleshooting, and other scenarios in the backend. Click on the workflow name to enter the detailed information page of the corresponding workflow.

©2013-2025 Tencent Cloud. All rights reserved.

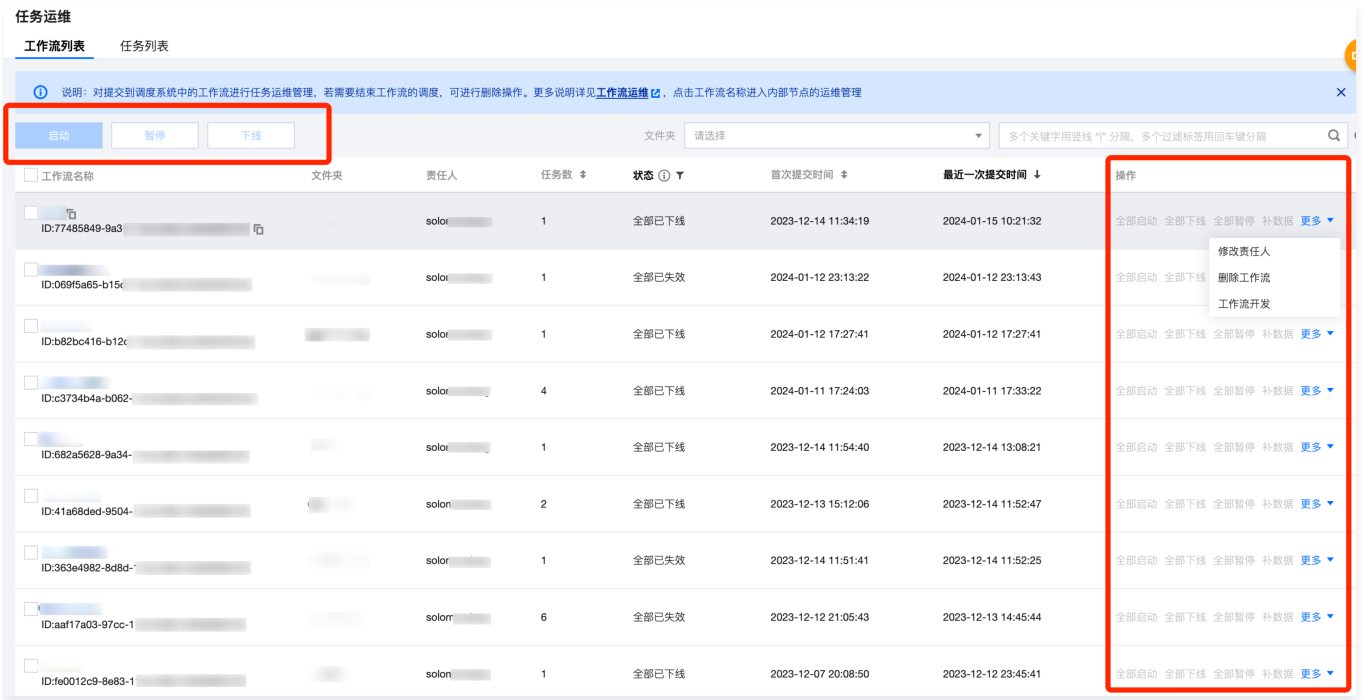


- Submission Time: Displays the date and time of the first OPS submission and the most recent OPS submission of the current workflow, in the format: yyyy-MM-dd HH:mm:ss.
- Task Information: Displays the number of computing task nodes involved in the orchestration of the current workflow.
- Type Distribution: Uses a pie chart to represent the proportion of different types of computing task nodes in the workflow.
- Cycle Distribution: Uses a pie chart to represent the proportion of different scheduling strategies for the various types of computing task nodes in the workflow.
- Run History: Shows the run history of the workflow (Only displayed if all tasks under the workflow have the same cycle).
- DAG Graph: Displays the current workflow's DAG computing node orchestration diagram. You can view the task type of each computational task node and the node orchestration logic. Hover over a task node to view the corresponding task node's Task ID, Project Name, Workflow Name, Person in Charge, Task Type, Task Status, Task Cycle, and Scheduling Plan.

Folder	Displays the workflow's belonging folder directory in the orchestration space for easy identification and unified management.
Responsible Person	Displays the name of the person responsible for the workflow, specifying the operation and maintenance responsibility, defaulting to the workflow creator.
Task Count	Displays the number of computational task nodes participating in the workflow orchestration.
Status	Workflow Running Status: Users can determine if the workflow is executing as expected based on its running status. There are six statuses: Scheduling In Progress, All Paused, Offline, Partially Scheduling, Undeployed, Expired. • Scheduling In Progress: All tasks in the workflow are in "Scheduling In Progress" • All Paused: All tasks in the workflow are "Paused" • Offline: All tasks in the workflow are "Offline" • Partially Scheduling: Some tasks in the workflow are "Scheduling In Progress", while others are "Paused" or "Offline" • Undeployed: All tasks in the workflow are in a non-"Scheduling In Progress" state, including "Offline", "Paused", and "Expired" • Expired: All tasks in the workflow are "Expired", and cannot be started or paused.
First Submission Time	Displays the date and time when the workflow was first submitted for operation and maintenance.
Most Recent Submission Time	Display the date and time when the workflow was last modified and submitted for O&M.

- Workflow O&M operations:

©2013-2025 Tencent Cloud. All rights reserved.



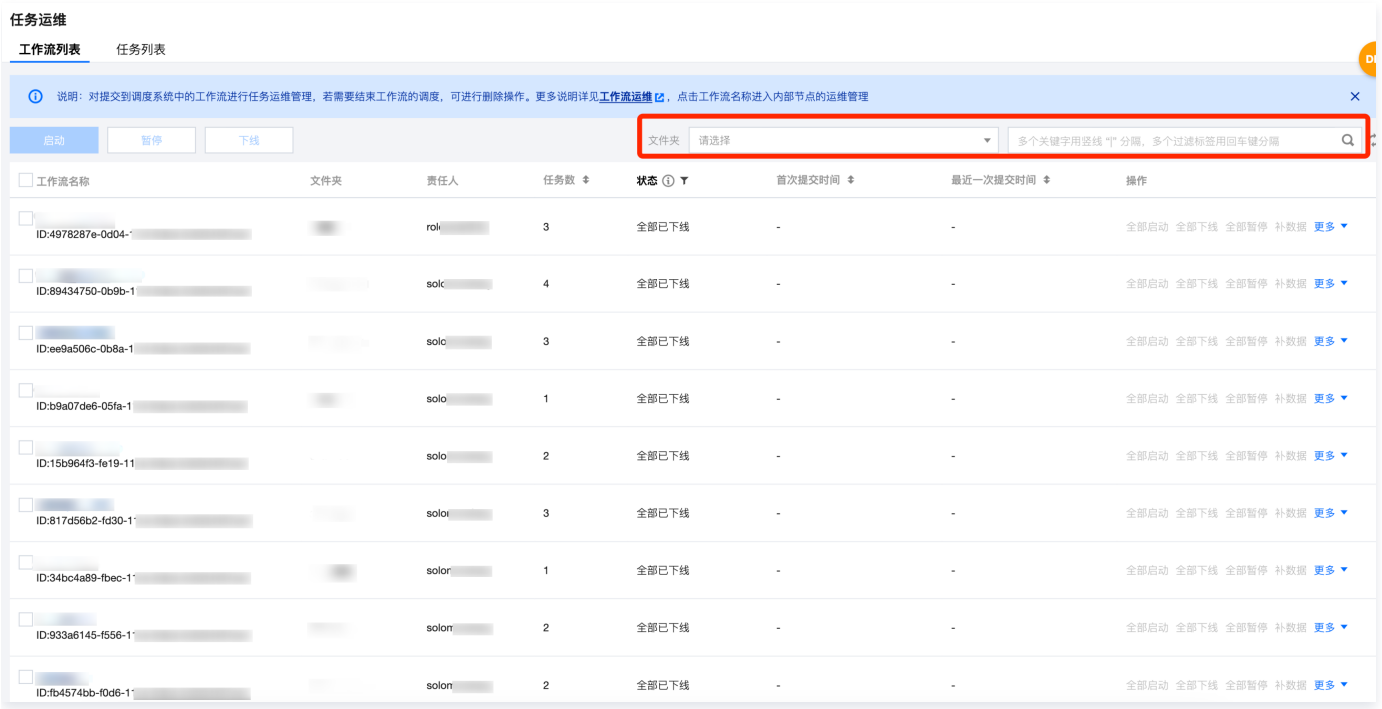
Feature description:

Operation	Description
Startup	Select the workflow by checking the box before its name, and then use the feature button above to batch execute scheduling operations like startup, pause, and offline for the selected workflows. • Start: Only valid for tasks with the status "Paused". After starting the workflow, all the tasks within this workflow will be in normal operation and instances will be generated according to the scheduling cycle. • Pause: Only valid for tasks with the status "Scheduling". After pausing the workflow, the tasks within this workflow will not generate new instances. • Offline: Only valid for tasks with the statuses "Scheduling", "Paused", and "Invalid". After going offline, the task status will be set to "Offline", and no more instances will be generated. If re-execution is needed, it must be resubmitted in the orchestration space.
Pause	
Offline	
Start All	For workflows with the running statuses "All Paused", "Partially Scheduling", and "All Unscheduled", all nodes with the status "Paused" belonging to them can be started.
Offline All	For workflows with the running statuses "All Scheduling", "All Paused", "Partially Scheduling", "All Unscheduled", "All Invalid", all nodes with the statuses "Scheduling", "Paused", "Invalid" belonging to them can be taken offline. After going offline, the task status will be set to "Offline", and no more instances will be generated. If re-execution is needed, it must be resubmitted in the orchestration space.

©2013-2025 Tencent Cloud. All rights reserved.

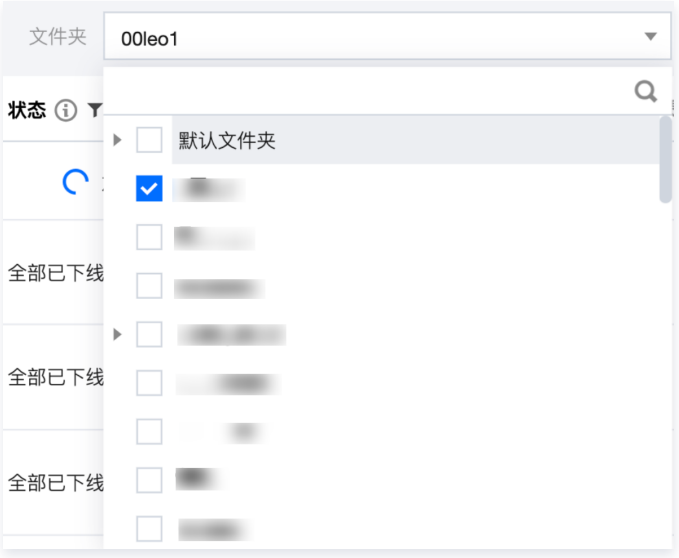
Pause All	For workflows with the running statuses "All Scheduling", "Partially Scheduling", all nodes with the status "Scheduling" belonging to them can be paused. After pausing the tasks, they will not generate new instances. Paused tasks can return to scheduling status through the start operation.
Backfill Data	For computing tasks with scheduling run failures or missing historical instances, perform supplementary executions to complete the data processing task cycle that is missing or has failed. For details, see Data Supplement.
Change Responsible Person	The default responsible person for the workflow is the user who created the workflow. You can change it to other users under the account, supporting multiple users as responsible persons.
Deletes a workflow	To delete a workflow in operation, it is necessary to take all the computing tasks in the workflow offline, ensuring the workflow status is "All Offline", then the workflow can be deleted. Workflows deleted from operation will not be removed from the orchestration space; they can be resubmitted for operation in the orchestration space later.
Workflow Development	Jumping to the corresponding workflow canvas in the orchestration space provides a shortcut for workflow operations and editing.

- **Workflow Filtering:**



Operation	Description
Filter by Folder	Support filtering workflows in operation and maintenance based on the folder directory in the orchestration space. Multi-level directories can be expanded, and multiple selections are supported.

©2013–2025 Tencent Cloud. All rights reserved.

	
Combined filtering by Person in Charge, Workflow name, Workflow ID	1. First, select the Filter Tag, then enter the filter keywords corresponding to the Tag. If multiple keywords need to be filtered under a single Tag, separate the keywords with a "|". 2. After configuring one Tag, click Enter to select the next Tag and then input the filtering keywords for the corresponding Tag. Repeat the above steps to achieve multi-condition combination filtering for operation and maintenance workflows. 
Workflow Refresh	You can view the latest status of the workflow list, usually used after performing operations on the workflow, to check the workflow's running status and help users get the latest operation and maintenance information of workflows.

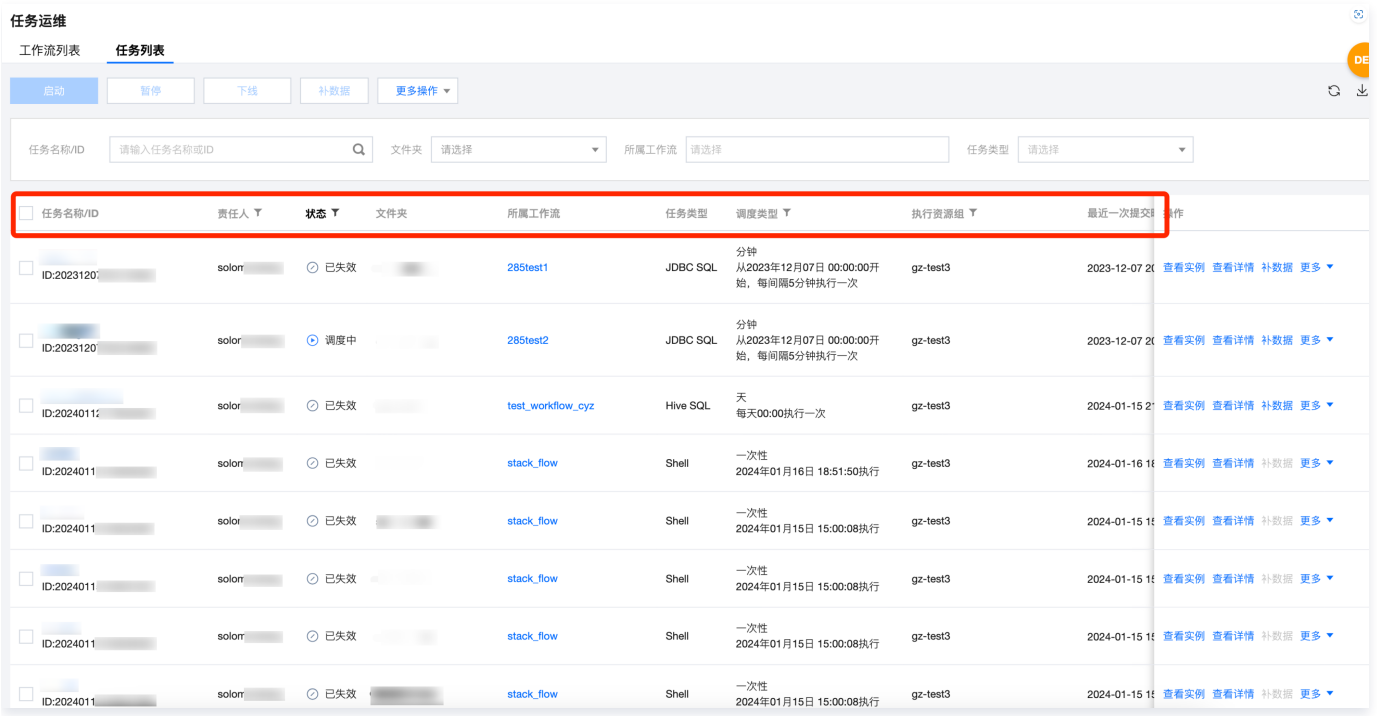
Task OPS

Task Operation and Maintenance List

Select the task list tab on the Task OPS page. The task list page displays computational task nodes from published workflows, showing the current scheduling status, scheduling period, and other information. A series of feature operations can be performed to maintain computational tasks, ensuring the stability, efficiency, and reliability of data processing nodes.

- Task List Information:

©2013–2025 Tencent Cloud. All rights reserved.



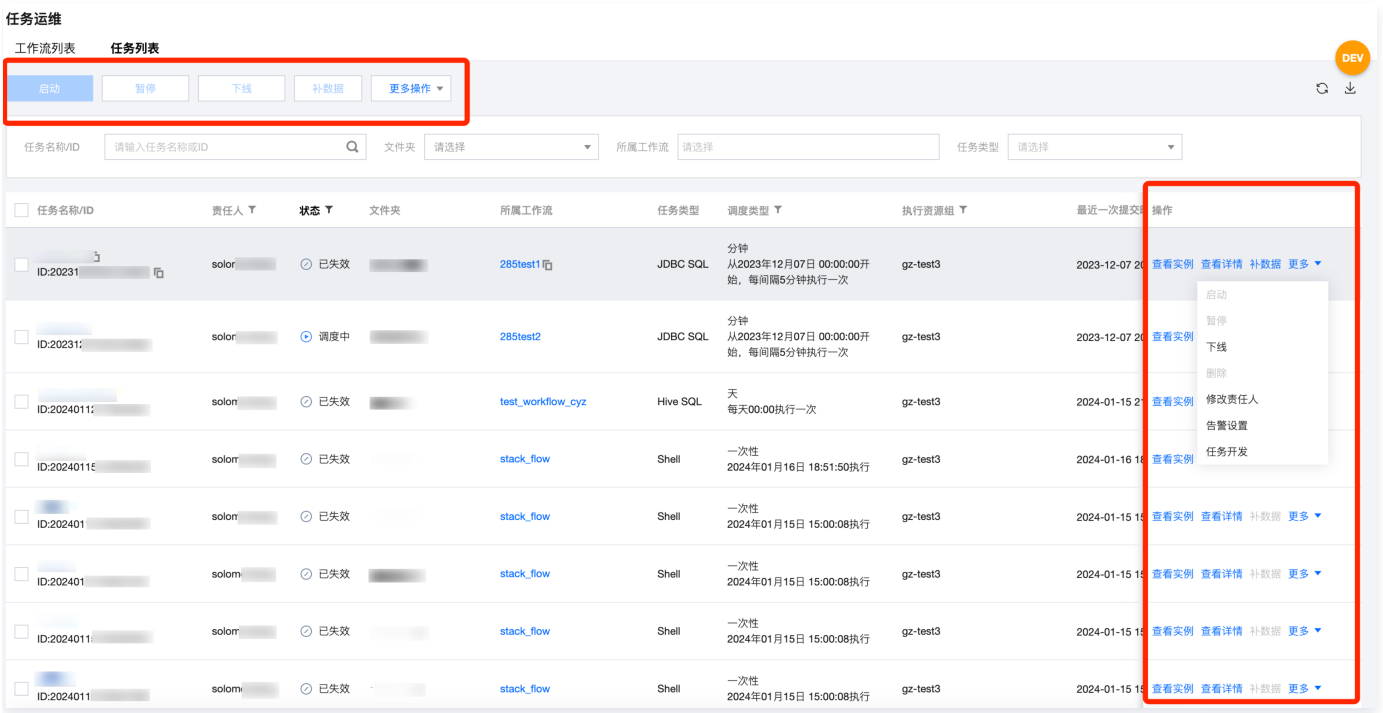
Feature description:

Information	Description
Task Name/ID	Supports copying the Task Name and ID. Clicking the Task Name navigates to **Task Details Page** > **Task Lineage**.
Responsible Person	The responsible person's name for the computing task, specifying the operation and maintenance responsibility attribution. By default, it is the creator of the computing task.
Status	Running status of the computing task. Users can determine whether the task is executed as expected based on its running status. There are 5 statuses: Scheduling In Progress, Paused, Offline, Expired. • Scheduling In Progress: The computing task is in a normal "scheduling" state and task instances will be generated according to the scheduling plan. • Paused: The compute task is paused and in the "Paused" state. New instances will not be generated, but existing instances will continue to run. • Offline: The compute task is offline and in the "Offline" state. Offline tasks will not generate new instances. To reschedule, you need to go to the orchestration space and resubmit. • Expired: The task is in an "expired" state, meaning the current time has exceeded the task's effective time. Expired tasks cannot be started or paused.
Folder	Folder directory of the computing task in the Orchestration Space, which helps to identify the task's attribution for unified management.
Associated Workflow	Name of the associated workflow of the computing task in the Orchestration Space, which helps to identify the task's attribution for unified management. Click the workflow name to jump to the specific workflow configuration page of the computing task.
Task Type	Task type of the computing task.
Scheduling Type	Scheduling period of the computing task.
Execution Resource Group	Configured execution resource group for the computing task.
Most Recent Submission Time	Displays the most recent modification and update submission date and time for the computing task. Format: yyyy-MM-dd HH:mm:ss.
First Execution Time	Displays the time when the computing task generates the first instance according to the scheduling period. Format: yyyy-MM-dd HH:mm:ss.

©2013-2025 Tencent Cloud. All rights reserved.

Creation Time	Displays the date and time of the first submission of the computing task for OPS. Format: yyyy-MM-dd HH:mm:ss.

- Task OPS operations:



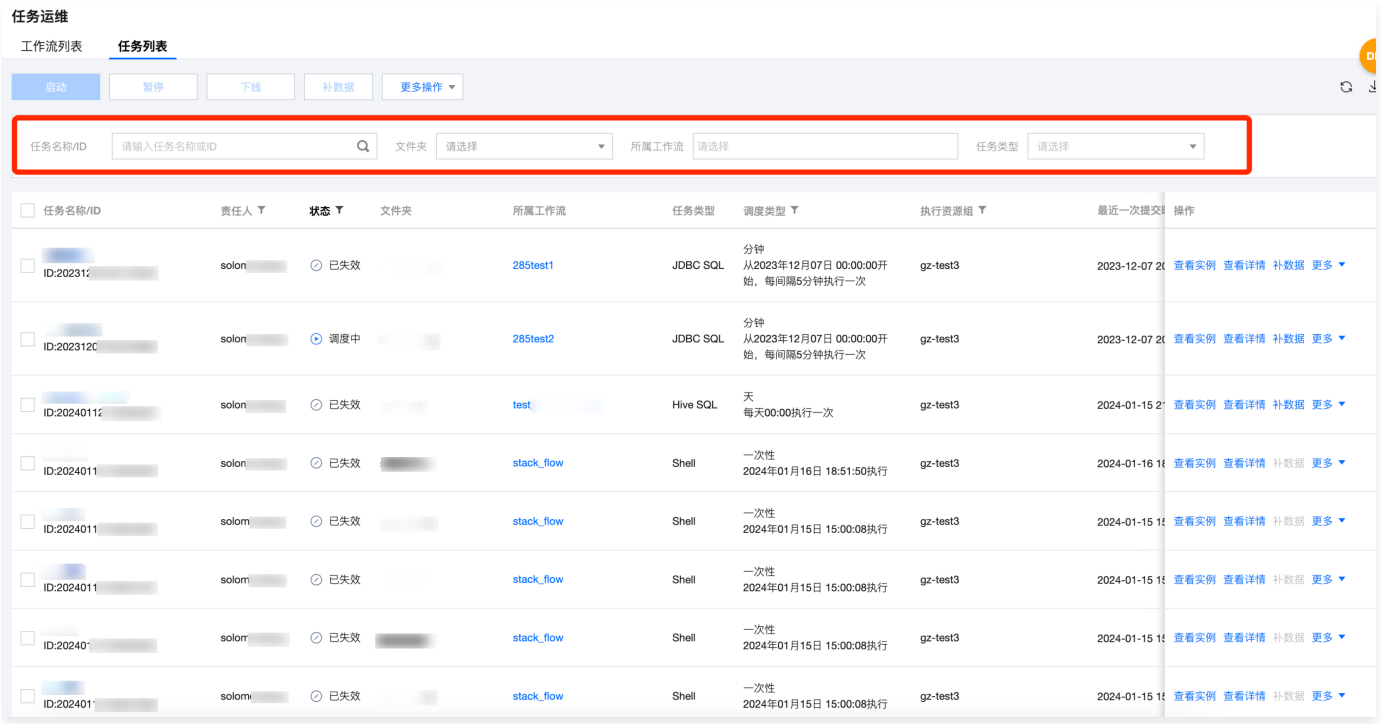
Feature description:

Operation	Description
Startup	Divided into batch operations and single-task operations. For batch operations, select the checkboxes in front of the task names, then use the feature button above; for single-task operations, click the operation item in the operation column of the task list. Start: Only valid for tasks with the "Paused" status. After starting the task, it will be in normal scheduling status and generate instances according to the scheduling period.
Pause	Suspension: Only valid for tasks with the "Scheduling" status. After suspending the task, no new instances will be generated. You can choose whether to terminate already generated instances during the suspension operation.
Offline	Decommission: Only valid for tasks with the statuses "Scheduling," "Paused," and "Invalid." After decommissioning the task, its status will be set to "Decommissioned," and no instances will be generated. To re-execute, go to the orchestration space and resubmit.
Backfill Data	Backfill Data: For tasks with failed scheduling instances, supplementing the historical missing instances, and handling the data processing cycle of missing or failed tasks. Refer to Data Supplement
Alarm settings	Alert settings: Set alert monitoring for the task's running status so that OPS personnel can quickly detect and respond to task anomalies. Refer to Task Alerts
Change Responsible Person	Modify Person in Charge: Change the person in charge of the task to clarify the responsibility for computing task OPS. Multiple persons can be added to the same task at once.
Delete	Delete: Only supports deleting computing tasks with the "Decommissioned" status. The deletion message can be sent to the person in charge of the task. If multiple computing tasks are selected, the tasks with other running statuses will not be deleted by default! There are two methods of deletion: After deletion, downstream tasks run normally. After deletion, downstream tasks fail.
Viewing	Navigate to the list interface of the instances of the corresponding computing task to view instance running

©2013-2025 Tencent Cloud. All rights reserved.

Instance	status, running time, and perform feature operations on the instances.
View details	Navigate to the task details page, supports basic task operations, and viewing basic task information. The task details page is introduced below.
Task Development	Navigate to the corresponding compute node configuration interface in the orchestration space, providing a shortcut for editing computing task nodes for OPS.

- Task Filter:



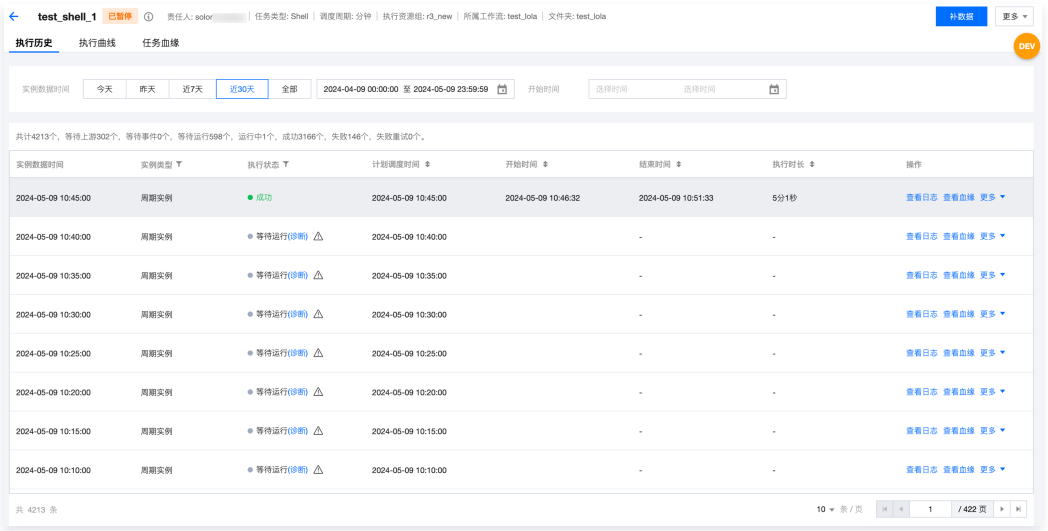
Operation	Description
Filter by Task Name/ID	Supports fuzzy matching by task name or task ID.
Filter by Folder	Supports filtering tasks under operation and maintenance based on the folder name in the orchestration space where the computing task is located. Multi-level directories can be expanded and multi-selection is supported.
Filter by Associated Workflow	Supports filtering by the workflow name to which the computing task belongs.
Filter by Task Type/Data Source	Supports filtering by task type. If you select an offline synchronization task, you can filter by source data table and target data table.

Task Details Page

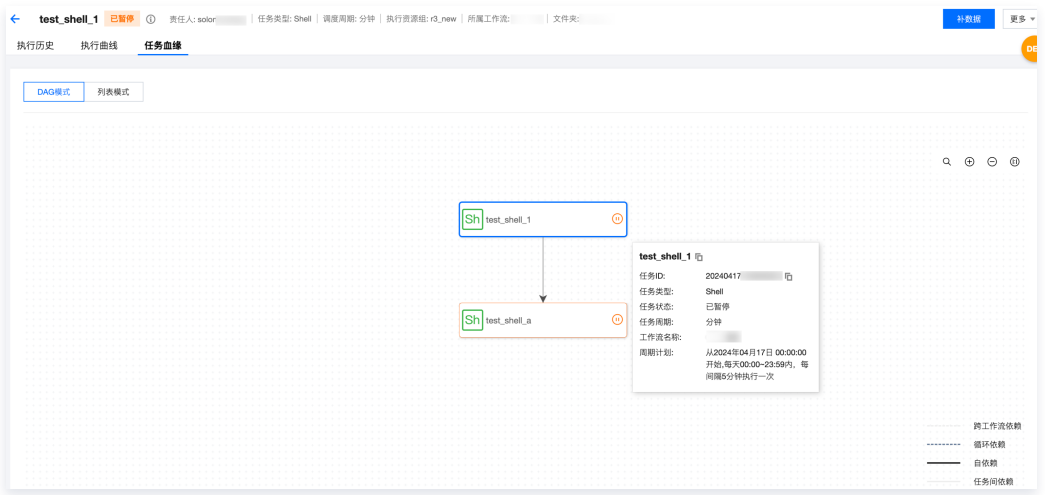
The task details page is a collection of task information and operations, including basic task information, basic operations, execution status of historical instances, and task lineage information. Task operation and maintenance can be performed on this page.
The task details page includes three tabs: Execution History, Execution Curve, and Task Lineage.

Content	Description
Execution History	Displays the basic situation of the historical instances of the task. • Filtering: Supports filtering instances by instance data time and start time. • Statistics: Counts the total number of instances within the filtered range, including the number of instances in each status (waiting for upstream, waiting for event, pending execution, running, success, failed).

©2013–2025 Tencent Cloud. All rights reserved.

	• Instance List: Displays the basic information and operations of instances. 
Execution Curve	Displays the execution time curve of the task's historical instances. • Execution Duration Change Trend: View the change trend of the execution duration for the last 25, 50, or 100 instances. Users can understand the historical instances' execution duration fluctuations to help determine if there are any anomalies. • Start/End Time Change Trend: View the change trend of the start and end times for the last 25, 50, or 100 instances. 
Task Lineage	Displays the lineage information of the current task. Divided into DAG Mode and List Mode. For complex lineage relationships with many nodes, it is recommended to use List Mode.

©2013–2025 Tencent Cloud. All rights reserved.



©2013–2025 Tencent Cloud. All rights reserved.

OPS of Instances

Last updated：2024-11-25 14:43:00

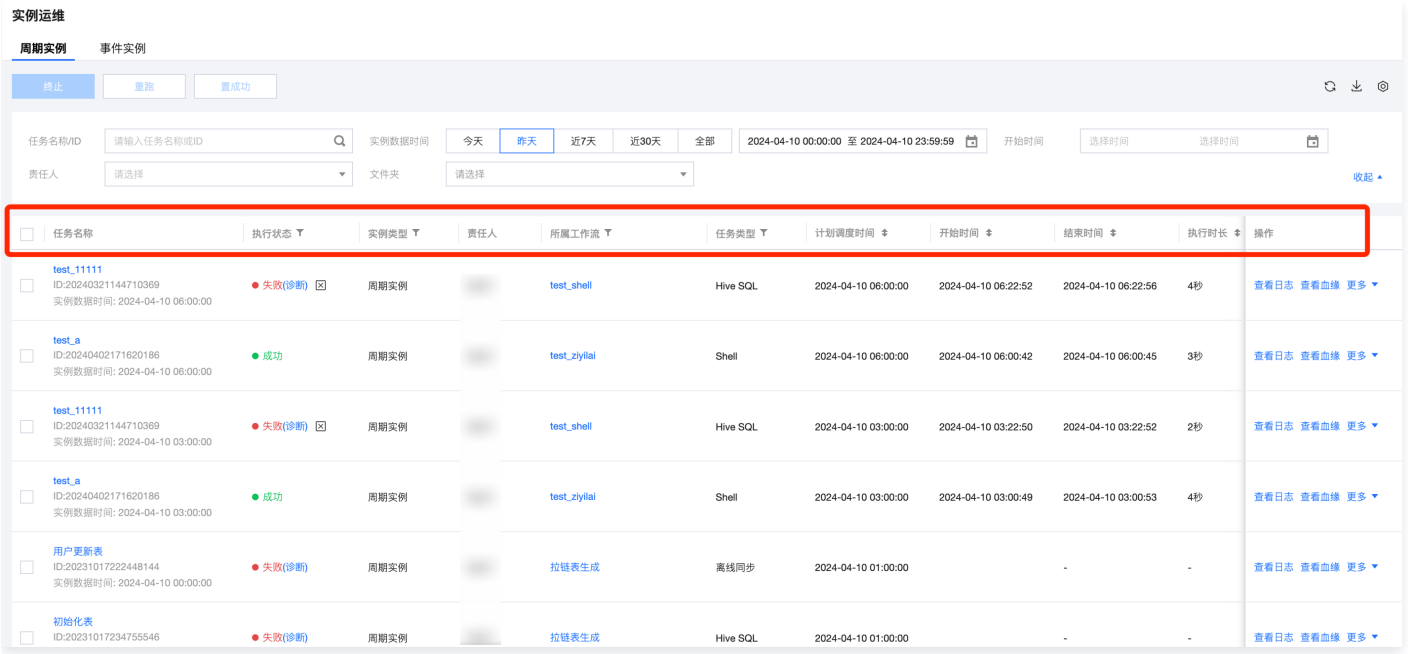
OPS of instances is the operations page for instance snapshots generated by computing tasks according to the scheduling policy and event listening. You can view and manage periodic instances and event instances through the page features. Perform a series of feature operations to maintain instances, ensuring the stability, effectiveness, and reliability of data processing and operation processes.

Enter the Ops page

1. Log in to Tencent Cloud WeData Data Development and Governance Platform.
2. Click the project list in the left-hand directory tree.
3. Select the corresponding item, and click to enter the Data Development Module.
4. In the left directory tree, click **Instance Operations**.

Periodic Instances List

The periodic instances page displays detailed information and operating features of instances generated by computing tasks through scheduling policies in a list form. You can view the instance running status, logs, and upstream and downstream relationships of parent and child instances. You can rerun, terminate, or mark instances as successful. **The system currently retains instances within three months; instances older than three months will be automatically cleaned up.**



Periodic Instances List Information

Feature description:

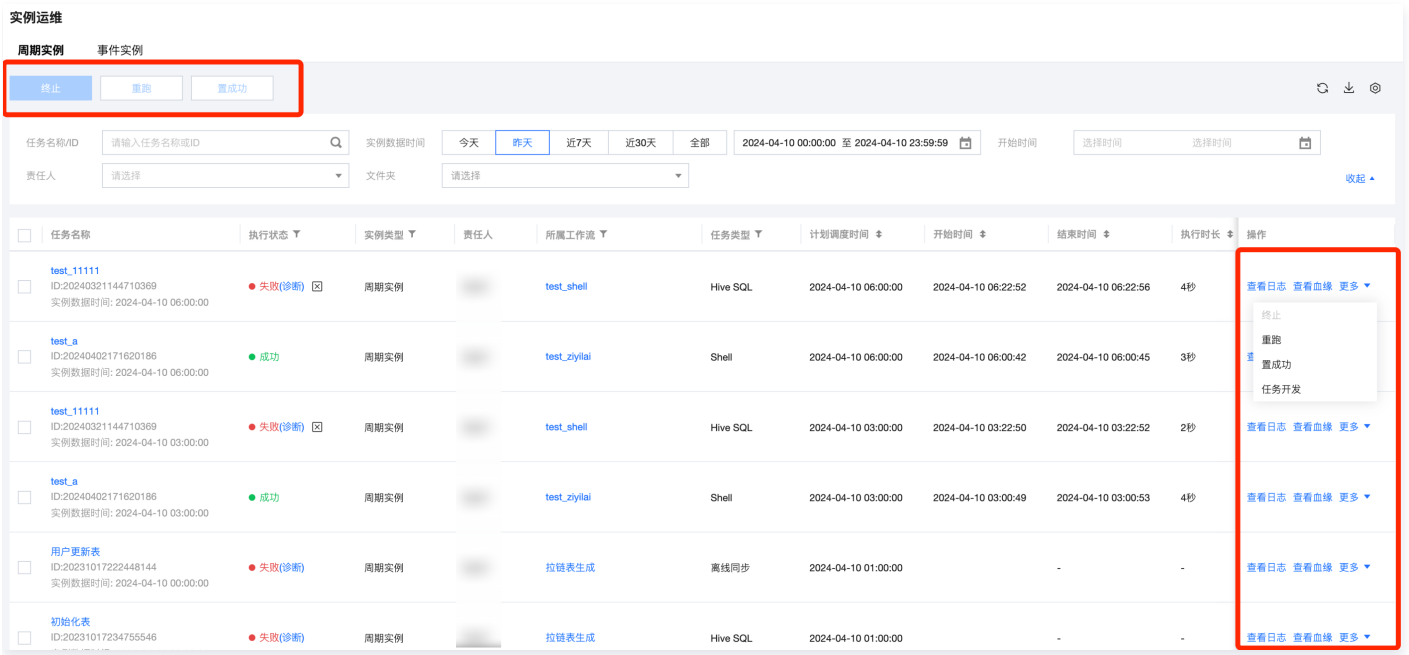
Information	Description
Task Name	Task Name: Click to jump to the instance details page – Instance Lineage, support copying. Task ID: Used for background log queries, troubleshooting scenarios, and supports copying. Instance Data Time: Displays the data time of the task. The task name + instance data time uniquely identify an instance.
Execution Status	The running status of periodic instances helps to identify faults, problems, and other exceptions. • Waiting for Event: The computing task to which the instance belongs is configured to listen for events. The instance can run normally after the event is triggered. • Wait for Upstream: The computing task to which the instance belongs has upstream dependent tasks in the workflow orchestration process. The current instance task will run normally only after the upstream computing task generates the instance and runs successfully.

©2013-2025 Tencent Cloud. All rights reserved.

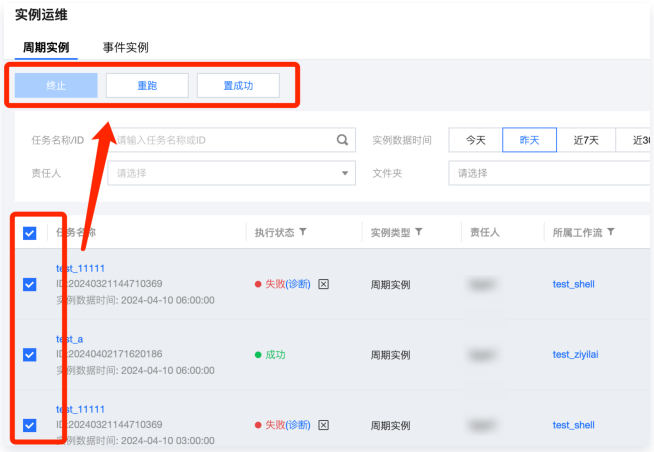
	• Pending Execution: Waiting for scheduling resources or exceeding the scheduling concurrency, waiting to be sent to the execution machine. The instance is in the ready-to-run stage shortly after being generated and will enter the next stage within a short period. • Running: The instance is in the running stage, and you can view the running process in real-time through the logs. • Terminating: Tasks in waiting for events, waiting for upstream, pending execution, and running stages can be terminated using the termination feature. • Failed Retry: When an instance execution fails, according to the failure retry configuration in the task scheduling policy, the instance will execute again after failing. • Failure: Task instance failed. • Success: Task instance succeeded. Diagnostic Information and Error Code information are provided in the execution status column • Diagnostic Information: Provides diagnostic information, possible causes, and operation guidance to help users quickly locate and resolve issues when an instance is abnormal. • Error Code: Provides error codes and error messages. If the problem cannot be located, you can copy the error code and seek assistance from the platform's support team.
Instance type	Displays the instance type according to the scheduling policy of generated periodic instances. • Periodic Instances: Compute task instances generated by periodic scheduling strategies, such as daily, weekly, monthly, hourly, and minute-based schedules. • Non-periodic Instance: Compute task instances generated for one-time execution. • Supplementary Instance: Task instances generated through data supplementation.
Responsible Person	Shows the responsible person's name for the periodic instance, clarifying the maintenance responsibility of the compute task instance. By default, the task creator is responsible.
Associated Workflow	Displays the name of the workflow that the periodic instance's compute task belongs to in the orchestration space, making it easy to identify and manage instances uniformly. Clicking the belonging workflow name navigates to the workflow configuration page of the corresponding compute task.
Task Type	Displays the type of compute task that the periodic instance belongs to.
Scheduled Execution Time	Scheduling operation cycle of periodic instances.
Start Time	The start time for the periodic instance to begin running after it is generated. The format is: yyyy-MM-dd HH:mm:ss.
End time	The end time for the periodic instance to stop running. The format is: yyyy-MM-dd HH:mm:ss.
Running Time Consumption	The execution duration of the periodic instance from start to end time after it starts running.
Execution Resource Group	The execution resource group used by the periodic instance for execution.
Retry Attempts	Periodic scheduling, backfill, or each rerun will count from 0. If the set retry attempts are exceeded without successful execution, the task fails. "Number of Retry Attempts on Failure" is set in the **Task Scheduling Settings** > **Advanced Settings** > **Number of Retry Attempts**, with a default value of 5.
Total run counts	Refers to the sum of execution counts for periodic scheduling, backfill, and each rerun of an instance.
Folder	Displays the directory of the folder that the periodic instance's compute task belongs to in the orchestration space, making it easy to identify and manage the compute task instances uniformly.
Cycle Type	Displays the scheduling period of the compute task that the periodic instance belongs to.

OPS of periodic instances

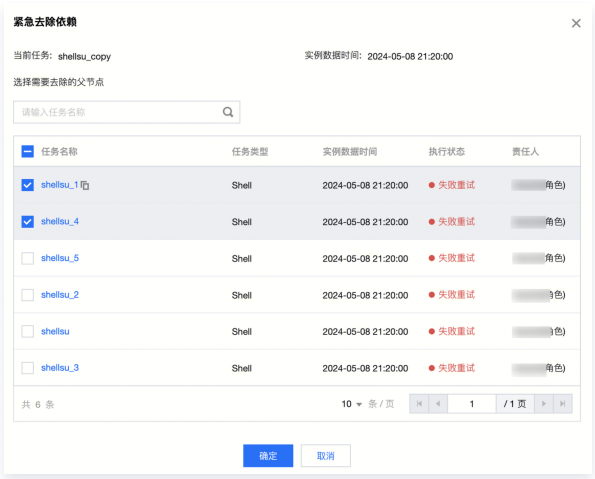
©2013-2025 Tencent Cloud. All rights reserved.



Feature description:

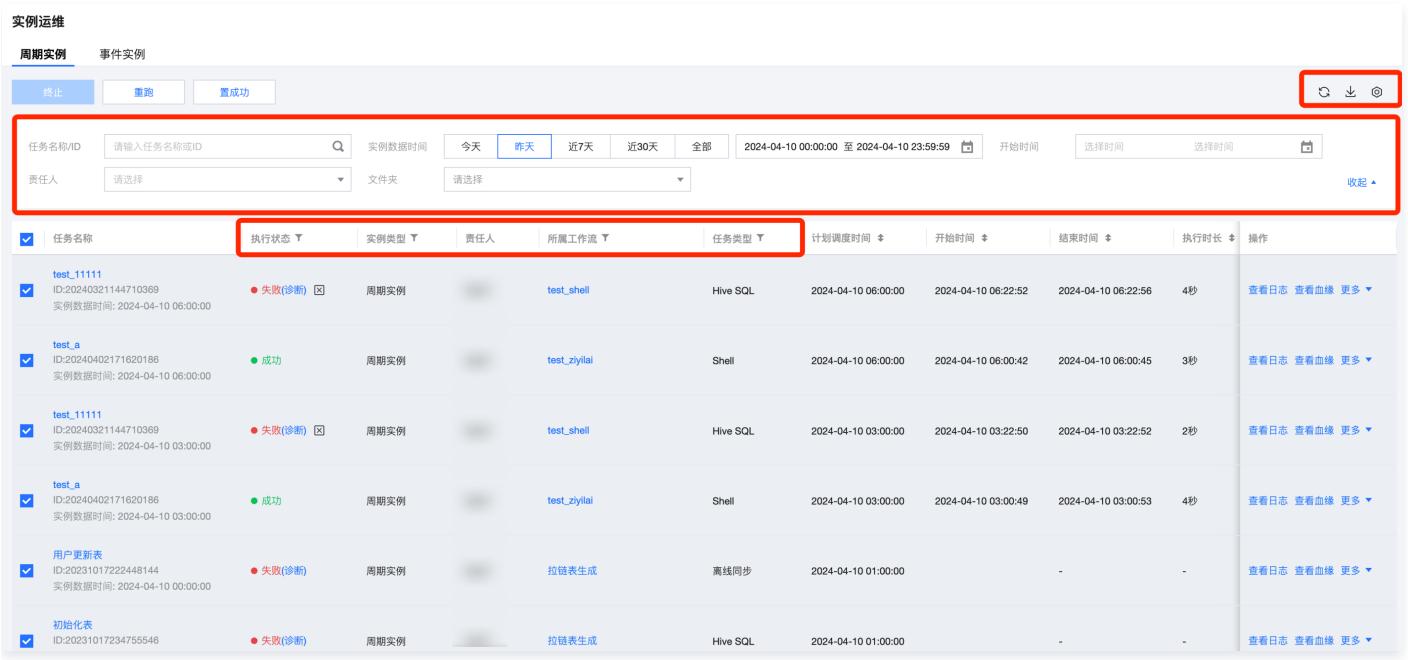
Operation	Description
Batch operation	Select the checkbox before the task name, then use the feature button above to batch terminate, rerun, or set the selected periodic instances as successfully executed. 
Viewing log	Navigate to **Instance Details Page** > **Execution Log** to view the full process of each execution after instance generation. This includes information such as Waiting for Event/Upstream, Pending Execution, Running Start Time, End Time, Time Consumption, Specific Circumstances, Execution Log, and Execute Code.
View Bloodline	Jump to **instance details page** > **Instance Lineage** , you can view the upstream and downstream relationships of the instance and perform simple information viewing and operations on the DAG Diagram.
Terminate	Valid only for Periodic Instances with a status of "Waiting for Event", "Waiting for Upstream", "Waiting to Run", or "Running". After terminating the instance, the corresponding instance will stop running and its execution status will be set to "Failed".
Re-run	Valid only for Periodic Instances with a status of "Failed", "Success", "Waiting to Run", or "Failed Retry". After rerunning the instance, the corresponding task instance will be executed again. Rerunning the instance may affect upstream and downstream dependencies, event listeners, and scheduling policies, so it's necessary to set relevant configuration items when rerunning. • Re-run Instance Range: Three rerun methods can be selected. ○ Current Instance: Only rerun the currently selected instance ○ This Instance and Downstream Instances: Rerun the currently selected instance and its downstream task instances

©2013-2025 Tencent Cloud. All rights reserved.

	○ Downstream Instances: Only rerun the downstream task instances of the currently selected instance • Downstream Instance Range: If the rerun instance range is set to "This Instance and Downstream Instances" or "Downstream Instances", you can specify the range of downstream instances. Includes: ○ Current Workflow: Downstream instances in the current workflow ○ Current Project: Downstream instances in the current project ○ All Projects: All downstream instances ○ Specified Instance: User-defined downstream instance range • Ignore Event: When an instance rerun involves compute tasks configurations with event listeners, you can choose whether to ignore the event trigger conditions during the rerun. • Check Parent Task: Check the directly upstream parent tasks of the selected instance. Choose whether to check instances from the previous cycle. Determined by settings in **Scheduling Configuration** > **Self-Dependency** .
Set Successful	Valid only for Periodic Instances with a status of "Running", "Failed", "Waiting for Event", "Waiting for Upstream", and "Waiting to Run". After marking the instance as successful, the execution status of the corresponding task instance will be set to "Success".
Emergency Dependency Removal	• This operation can remove the dependency relationship of instances. It is used in emergency scenarios when the upstream instance has not reached a successful status and affects the execution of downstream instances. Users can perform this operation if they determine that the upstream instance has little impact on the current instance. The dependency relationship between the corresponding upstream and current instance will be removed, ensuring the current instance can execute normally. • This operation only takes effect for the current instance and does not affect the dependency relationships of previous or subsequent instances. • This operation is irreversible, meaning subsequent reruns or supplementary operations for this instance will not check for the removed dependencies. • Dependencies can be partially removed as needed. • The removed dependency relationships will still be displayed on the instance DAG chart, marked specially for user differentiation. 
Task Development	Jump to the configuration interface of the compute node corresponding to the Periodic Instance in the Orchestration Space. It's a shortcut for instance maintenance to edit compute task nodes.

Periodic Instance Filtering & List Operations

©2013-2025 Tencent Cloud. All rights reserved.



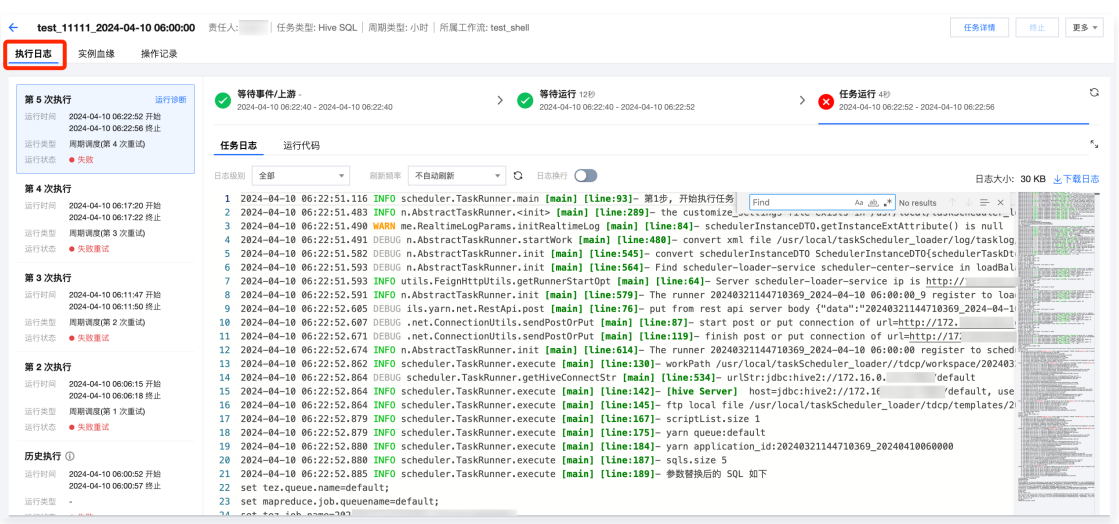
Operation	Description
Filter options – Top of the list	You can filter by Task Name/ID, Instance Data Time, Start Time, Person in Charge, and Folder.
Filter options – Inside the list	You can filter by Execution Status, Instance Type, Belongs to Workflow, Task Type, Execution Resource Group, and Cycle Type.
List Operation	Refresh: You can refresh the latest status of the periodic instance list, generally used to check the instance run status after operations, helping users easily obtain the latest maintenance information. Download: You can download the data of the current instance list. Currently, only instances generated before the current 0:00 can be downloaded. List Configuration: Supports adjusting the fields and the order of fields displayed in the current list.

Periodic Instance Details Page

The Instance Details Page aggregates all operation histories, operation statuses, instance lineage, operation records, and basic operations of instances, helping users view and maintain instances.

Tab Page	Description
Execution Log	View the entire process of each execution after the instance is generated, including Waiting for Events/Upstream, Waiting to Run, Start Time in Progress, End Time, Duration, Specifics, Execution Logs, Execution Code, etc. • Execution List: The far right is the record list of each execution of the instance. Each cycle schedule, rerun, and first execution or retry of backfill will be recorded as an execution. • Execution Information: Each execution can view the basic conditions of three stages: Waiting for Events/Upstream, Waiting to Run, and Task Running ○ Waiting for Events/Upstream: Displays the direct upstream instances/events list and statuses, making it easy to quickly identify blocking instances/events. ○ Waiting to run: Waiting to be assigned to the execution machine. ○ Task execution: Assigned to the execution machine and started executing. You can view the execution log and execute code.

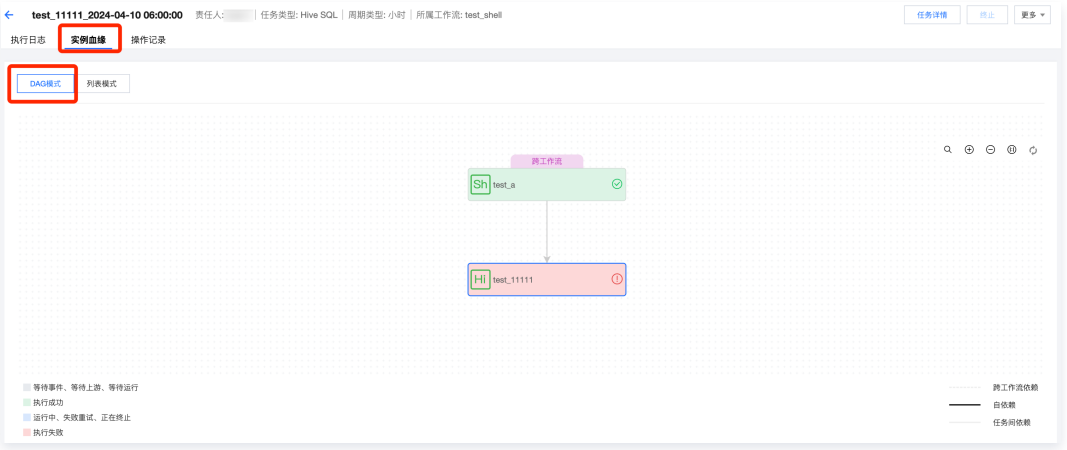
©2013–2025 Tencent Cloud. All rights reserved.



Instance lineage

View the upstream and downstream relationships of the current computing task instance within its workflow. The running status and detailed information for each compute node are displayed, with different colors marking the instance status.

Graph mode: View instance lineage in graph mode. It supports expanding upwards layer by layer and basic instance operations.



List mode: View instance lineage through list mode. You can choose to view upstream or downstream.



Operation Record

View the operation records for the instance, including actions like Set Successfully, Terminate, Rerun, among others. It lists the operator, operation time, and other relevant information.

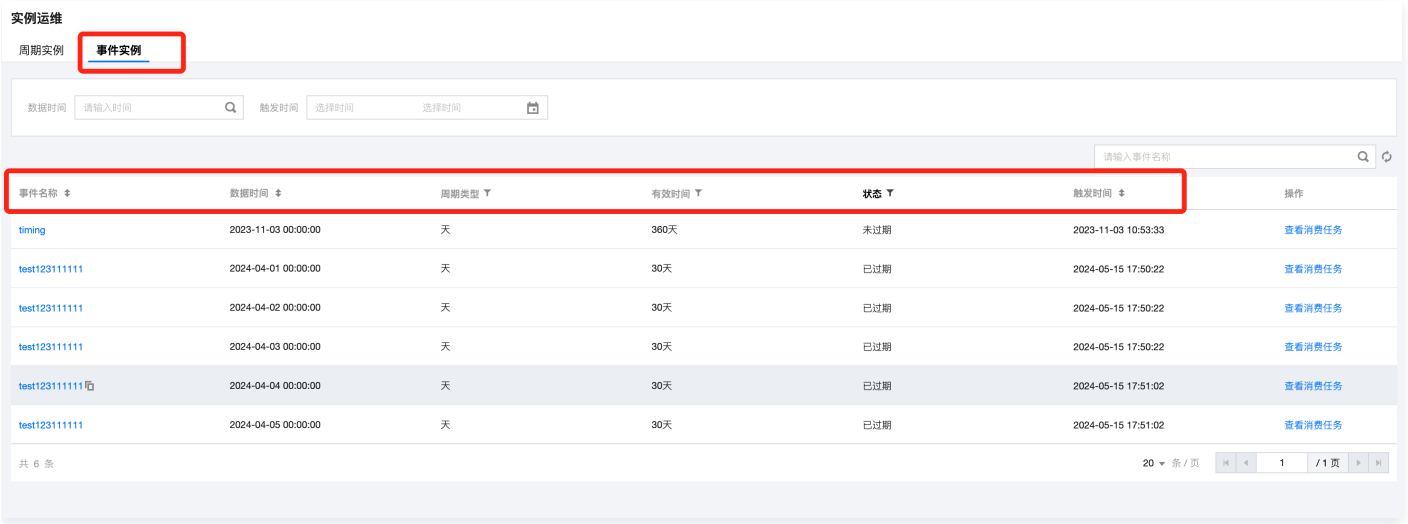


©2013–2025 Tencent Cloud. All rights reserved.

Event instance list

The event instance page displays the information and operation features of event instances generated after computing tasks trigger listening events as per the scheduling policy in a list format. You can view the consumption status of events, trigger time, consumption time, and more. Detailed information about the instance and consumption task can also be seen.

Event instance list information



Feature description:

Information	Description
Event name	Displays the name of the listening event that triggered the instance. Click **Event name** to view detailed information about the corresponding listening event, including cycle type, time format, validity period, and description information. 
Data time	Displays the planned scheduling time for the task instance that triggered this event instance.
Cycle Type	Displays the trigger cycle type of the event instance, including day, hour, minute, second.
Valid time	Displays the preset event waiting time in the scheduling configuration for the computing task that triggered the event instance. If the event is not triggered within this time, the task instance is automatically terminated.
Status	Displays the consumption status of the event instance, including Pending, In progress, Consumed, Expired.
Triggering time	Displays the date and time when the event instance was triggered.

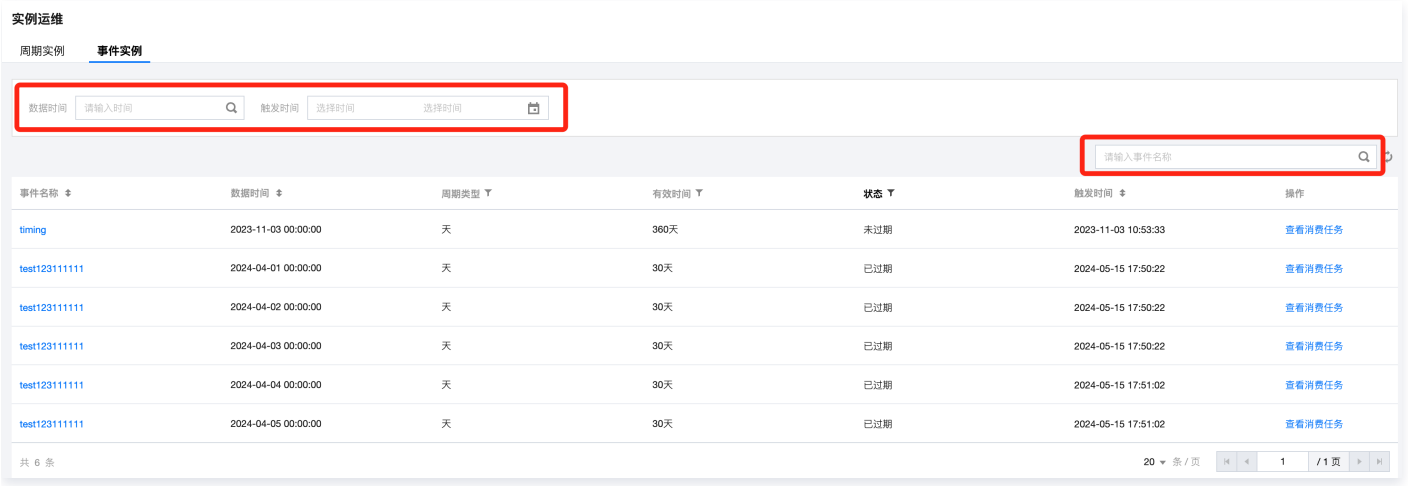
Event instance operations

©2013–2025 Tencent Cloud. All rights reserved.



Operation	Description
View consumption tasks	Click **View Consumption Tasks** under the event instance list to view the information of the triggered computing tasks.

Event instance filtering



Operation	Description
Filter options – Top of the list	You can filter by Data Timestamp, Trigger Time, and Event Name.
Filter options – Inside the list	You can filter by Cycle Type, Validity Period, and Status.

©2013–2025 Tencent Cloud. All rights reserved.

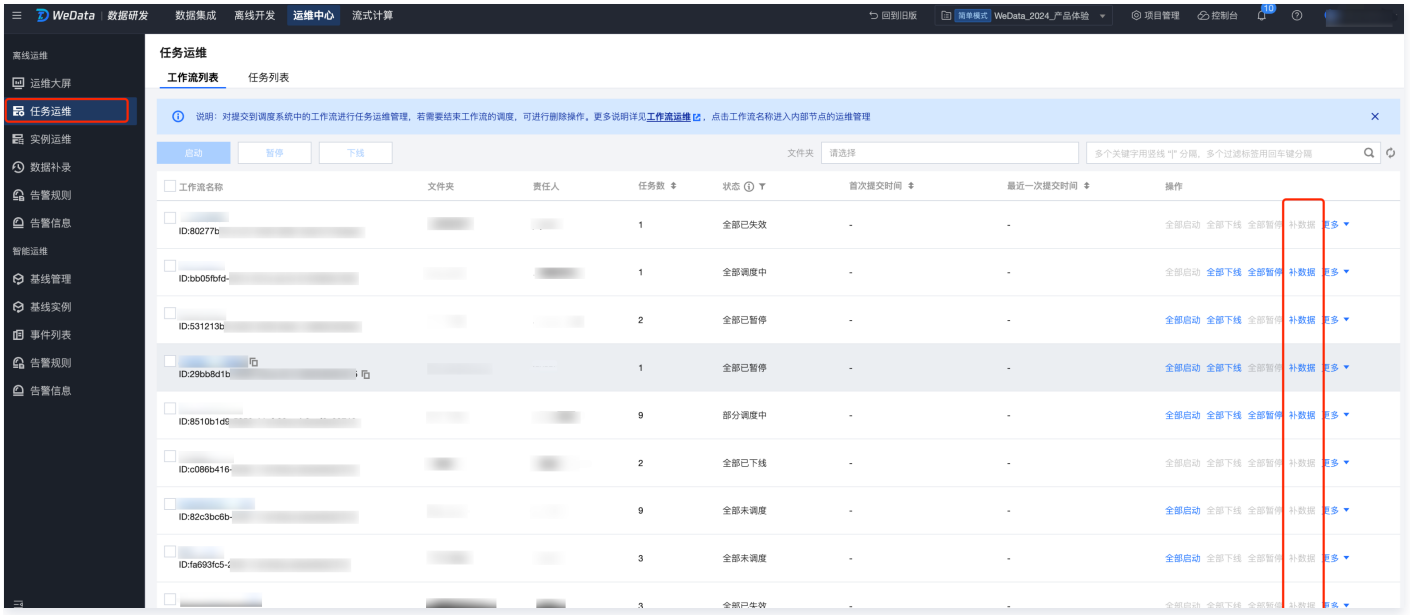
Data Supplement

Last updated：2024-12-12 20:14:30

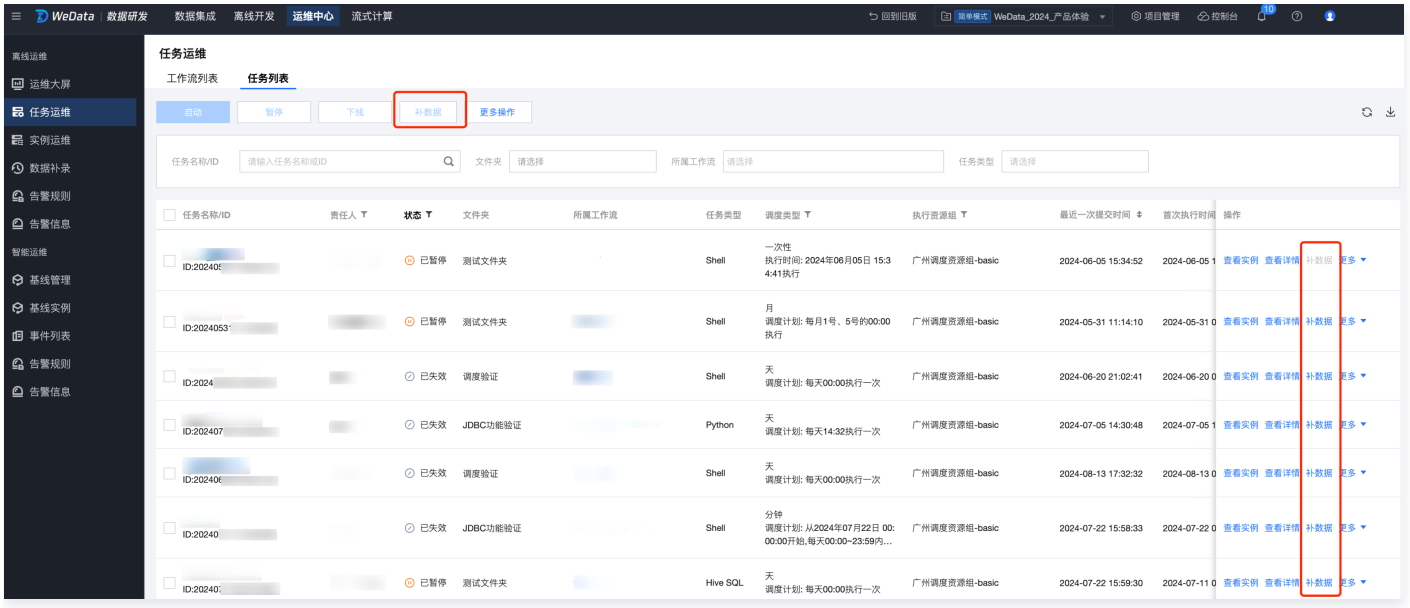
Data Supplement Entry

Log in to WeData Console > **Data Development** > **Task Operation and Maintenance** interface, where you can view the workflows and tasks created by the user. Filter to find the task that needs data supplementation, click **Data Supplement.**

> **Note**
> One-time tasks cannot supplement data



> **Note**
> The task list supports multi-selection for batch data supplementation.



Steps to Create a Supplementary Recording Task

Step 1: Select Task

©2013-2025 Tencent Cloud. All rights reserved.





Description

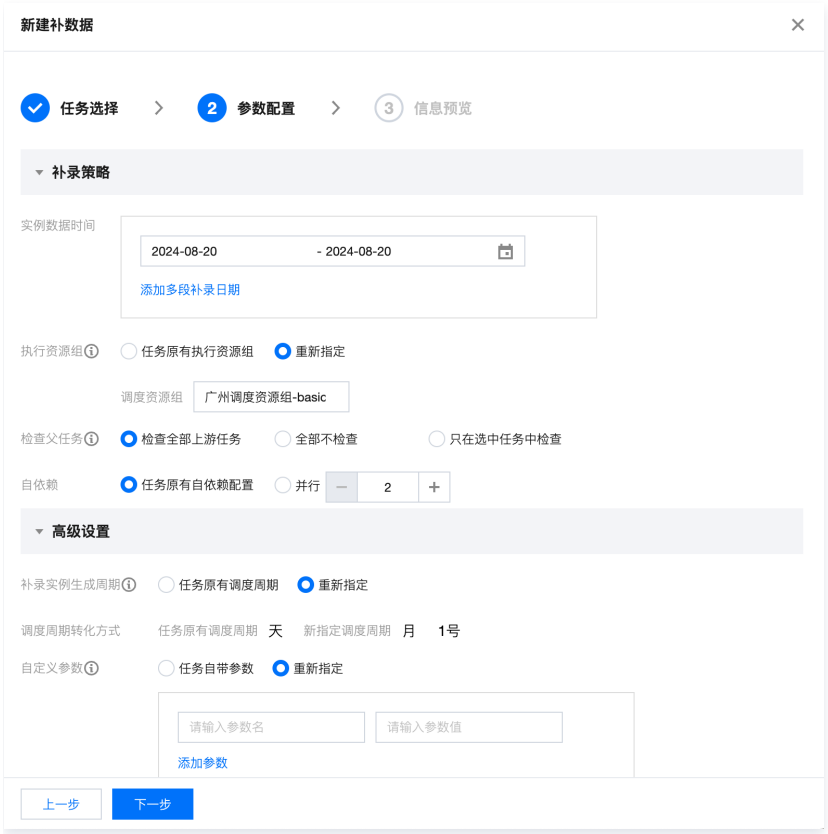
Parameters	Description
Supplementary Task Name	By default, it uses patch_YYYYMMDDHHmmss_unique number. For example: patch_20230624195944_548. Users can define their own names.
Description	Users can define the Supplemental Plan Description, up to 256 characters, optional.
Select Root Task	Select the root node that needs supplementation from the tasks in the workflow or the selected task.
Scope of Supplementary Task	• Selected Task: Can only be selected from the workflow or selected task. • Selected Task and Downstream Tasks: Can be selected from the workflow, selected task, and its downstream tasks. • Only include downstream tasks: Can only be selected from downstream tasks in the workflow or selected task.

©2013–2025 Tencent Cloud. All rights reserved.

	Note: Only tasks with the status "Scheduling In Progress", "Paused", or "Expired" support supplemental entries. The number of root nodes per supplemental entry cannot exceed 500. If this limit is exceeded, please narrow the selection range.
Select Downstream Task	When the range of supplemental tasks is selected as "Selected Task and Downstream Tasks" or "Only Include Downstream Tasks", the configuration for selecting downstream tasks will be displayed.

After selecting the tasks, click **Next**.

Step 2: Parameter Configuration



Description

Parameters	Description
Supplement Date	Users can customize the data date and select any time (including future dates), with yesterday selected by default (up to 10 time periods can be selected).
Hour Range	Default 00:00 – 23:59, effective only for tasks with an hourly or smaller cycle. Combined with the Data Date, it will generate supplemental instances. For example, if 01:00 – 02:00 is selected, the scheduled task instance at 01:00 will be supplemented in the hourly task, but the instance at 03:00 will not be supplemented.
Execution Resource Group	The original resource group of the task is default; users can reassign the scheduling resource group and integration resource group for the supplement instances without affecting the configuration of execution resources for periodic tasks.
Check Parent Task	• Check all upstream tasks: The selected supplement tasks will run according to the dependency relationship set in the task configuration. • Do not check any: The selected supplement tasks will not run based on task dependency and will run in parallel.

©2013–2025 Tencent Cloud. All rights reserved.

	• Only check within the selected tasks: The selected supplementary tasks only check dependencies within the selected tasks, not from unselected tasks.
Self-dependency	Default task's original self-dependency configuration (the self-dependency configuration of the original task). Users can define it themselves, and can select the parallelism degree (minimum 2, default value 2, maximum 10). Once set, the original self-dependency relationship will not be used.

Advanced Settings	• Instance Generation Cycle for Supplementary Recording: Allows re-specification of the instance generation cycle for supplementary tasks. Currently, only day instances can be converted to instances generated on the 1st of each month.
	• Custom Parameters:
	  ○ Default task parameters are provided, and users can redefine the task parameters to facilitate new logic execution for the supplementary instances. New parameters will override the previous task parameter values.
	  ○ Custom parameters are input in key-value format.
	• Offline Synchronization Rate Limiting: If the supplemental task includes offline synchronization tasks, the synchronization rate can be specified. Limit the synchronization rate by traffic or the number of records to protect the read and write pressure of the data source or destination. Default -1 means no rate limit.

After configuring the time and parameters, click **Next.**

Step 3: Information Preview

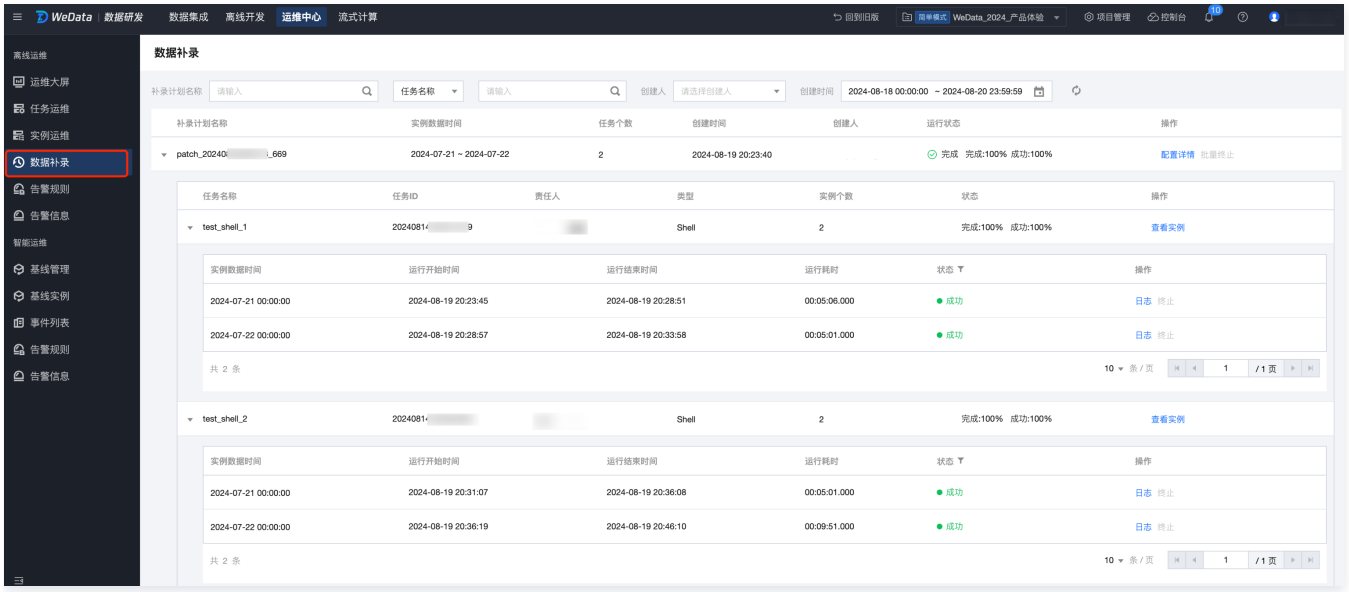


The preview of supplemental entries includes configuration preview, advanced configuration preview, and instance preview. The instance preview can show the estimated number of supplemental tasks and instances, as well as specific supplemental instances.

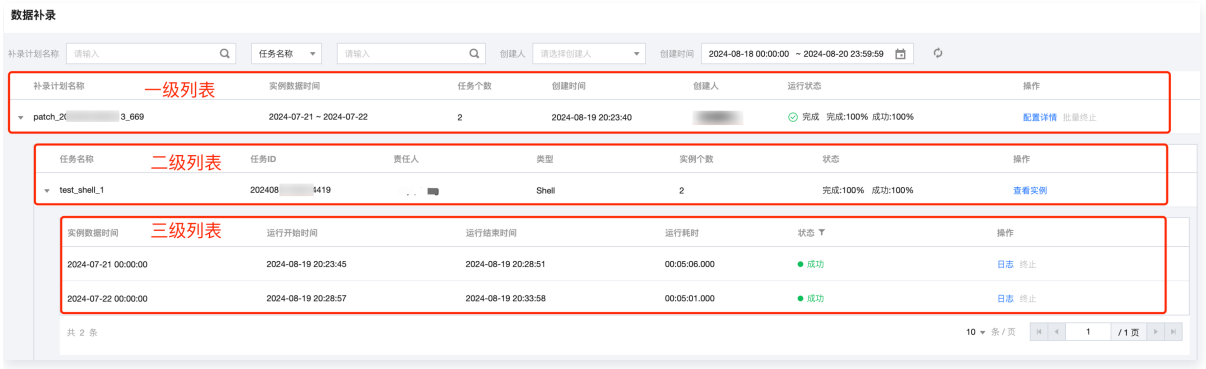
1. After previewing and confirming the data supplement, if the project requires approval, the supplement plan will be generated and executed only after approval.
2. After the supplement plan is created, users can click on the left **Data Supplement** to enter the supplement list and view details and progress.

Supplementary List

©2013-2025 Tencent Cloud. All rights reserved.



Data Supplement provides a three-level list view for the viewing and operation of supplementary tasks.



First-level list

- Search
  - Supplemental Task Name: Search for supplemental tasks by the task name.
  - Node Task Name: Search for supplemental tasks by the task node name included in the supplemental task.
  - Creator: Dropdown filter. Users can choose from all current project members.
  - Creation Time: By default, it shows the current day. Users can filter by creation time up to the current day.
- List Display: By default, display the supplemental task name, instance data time, task count, creation time, creator, running status, and operations.
- Running Status
  - Running: There are instances in the current supplemental task that are running or not yet running.
  - Complete: All instances in the current supplemental task are in the completed, terminated, or failed state.
  - Completion Rate: Total number of instances in (Success, Terminated, Failed) status / Total number of instances.
  - Success Rate: Total number of instances in (Success) status / Total number of instances.
- Operation
  - Configuration Details: Click to view the configuration details of the current supplemental task.
  - Batch Termination: Forcefully terminate instances that are "Waiting to Run" or "Running", other state instances are ignored by default! Batch operations are asynchronous, so there may be a delay in the status of instances.

Secondary List

- Display the current task name, Task ID, person in charge, type, number of instances, status, and operations.
- Operation-View Instance: If only one time period is selected for the supplemental recording, it supports clicking to jump to the instance list page and selecting the instances of the task in this supplemental recording.

©2013-2025 Tencent Cloud. All rights reserved.

Three-level List

- Display the current instance data time, start time of operation, end time of operation, runtime duration, status, and operations.
  - Diagnosis: When a task fails, diagnosis can be used to determine the cause of failure, and it supports jumping to the corresponding task details for viewing.
  - Operations: Support viewing log and terminating instance operations.

©2013–2025 Tencent Cloud. All rights reserved.

Task Alert

Last updated：2025-06-11 18:08:02

Alarm Rule

The Alarm Rule page provides configuration features for alarm tasks, supporting the configuration of alarm conditions and alarm notifications for the operating status of projects, tasks, and workflows.

Add Rule

1. Log in to the WeData Console.
2. Click the project list in the left menu to find the target project that requires configuring alarm rules.
3. Select a project, then click to enter the data development module.
4. Click **Alarm Rules** in the left menu to enter the Alarm Rule Management page, then click **Add Rule** and fill in the rule information.

Function Description

Information		Description
Basic Information	Rule Name	Alarm rule name, 1 – 128 characters, only Chinese, English, digits, and underscores.
	Monitoring Object	Select the monitoring object to set up rule alarms. Currently supported monitoring objects are tasks and projects. • When selecting a task as the configuration object, you can configure alarm rules for the task through three methods: computing tasks, workflows, and projects. • When selecting a project as the configuration object, configure alarm rules by checking the project. When selecting a task as the monitored object, the way to configure alarm rules is: • Configure by task: Click the **Add** button. In the pop-up dialog box, you can check the computational task nodes submitted for operations in the orchestration space and configure alarm rules for them.  • Configure by workflow: Click the **Add** button. In the pop-up dialog box, you can select workflows submitted for operations in the orchestration space and configure alarm rules for computational tasks within the workflows. Provide allowlist capability. After computational tasks in the workflow are added to the allowlist, the allowlisted tasks will not be monitored by alarm rules.  • Configure by project: Click the **Add** button. In the pop-up dialog box, you can select computational tasks submitted for operations in the orchestration space of the current project and set alarm rules for them. Provide allowlist capability. After computational tasks are added to the allowlist, the allowlisted tasks will not be monitored by alarm rules.

©2013–2025 Tencent Cloud. All rights reserved.

<table>
<tr><td></td><td></td><td> 

When selecting a project as the monitored object, the way to configure alarm rules is:
• Configure by project: Click the Add button. In the pop-up dialog box, you can select projects that the current user has joined (excluding visitors) and set alarm rules for them. Users can simultaneously select multiple projects for monitoring. When determining subsequent alarm conditions, the monitor values corresponding to the selected projects will be summed up for judgment.</td></tr>
<tr><td rowspan="2">Alarm condition (monitored object is project)</td><td>Upward fluctuation rate of failed instance count on the day (compared to the average of the last 7 days)</td><td>Use fluctuation rate and cumulative failed instance count on the day as conditions. Determine whether to trigger an alarm based on the number of times the conditions are met consecutively or cumulatively within the input number of periods.
 

Calculation formula for fluctuation rate:
(Current moment's failed instance count on the day – Unified moment's average failed instance count in the last 7 days) / Same time's failed instance count in the last 7 days
If this value is simultaneously in the following two statuses, it meets the alarm condition.
1. The average value is a positive number.
2. Exceed the set threshold</td></tr>
<tr><td>Downward fluctuation rate of failed instance count on the day (compared to the average of the last 7 days)</td><td>Use fluctuation rate and cumulative failed instance count on the day as conditions. Determine whether to trigger an alarm based on the number of times the conditions are met consecutively or cumulatively within the input number of periods.
 

Calculation formula for fluctuation rate:
(Current moment's failed instance count on the day – Unified moment's average failed instance count in the last 7 days) / Same time's failed instance count in the last 7 days
If this value is simultaneously in the following two statuses, it meets the alarm condition.
1. The average is negative.
2. Absolute value exceeds the set threshold</td></tr>
<tr><td>Alarm condition Monitored object is a task.</td><td>Execution failed</td><td>Trigger an alarm when the monitored task fails to generate an instance run. Supports configuration for periodic execution, backfilling, or rerun execution. You can choose the rule trigger condition of "failure after all retries complete" or "first run failure".
• Failure after all retries complete: According to the computing task scheduling strategy, in the case of task execution failure, if there are failures and related configuration for retries, proceed to execute according to the number of retries. Trigger the alarm rule when all instances have failed to run.
• First run failure: Trigger an alarm rule when the first generated instance fails to run according to the computing task scheduling strategy.</td></tr>
</table>

©2013–2025 Tencent Cloud. All rights reserved.

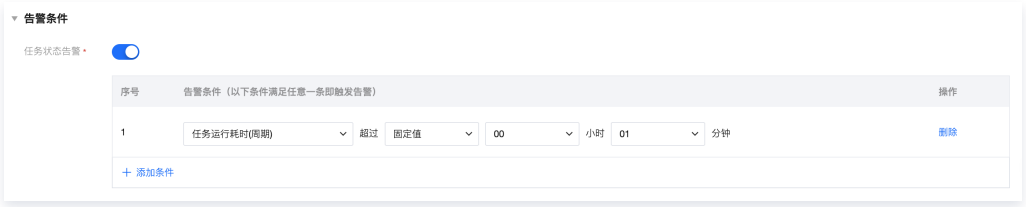




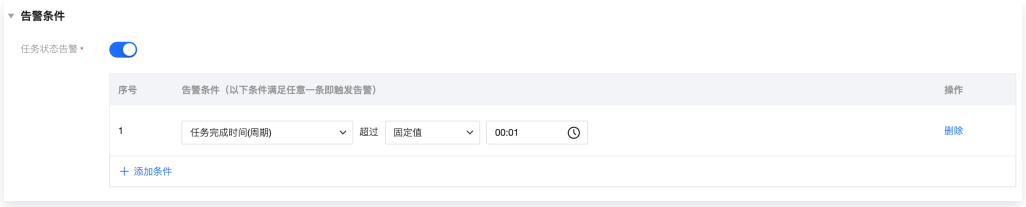
Execution timeout

Trigger an alarm when the monitored task's instance scheduling or running exceeds the preset time. Supports configuration for periodic execution, backfilling, or rerun execution. You can set rule trigger conditions for the four key time requirements: "task running time", "task completion time", "total time consumed for task waiting", and "task not completed within a period".

- Task running time: Select task running time (period) to indicate that the alarm is triggered if not completed beyond the threshold, calculated from the start time of the task period. Select task running time (rerun, backfill) to indicate that the alarm is triggered if not completed beyond the threshold, calculated from the start time of the task backfill or rerun. You can use a "fixed value" or "historical average" as the alarm condition.
  - Fixed value: If the instance does not complete execution within the specified hour minutes time requirement, the alarm rule is triggered.
  - Historical average: Calculate the average of the instance running times of the most recent 10 successful executions of a computing task, after removing the maximum and minimum values. The setting is invalid if there are fewer than 10 executions.

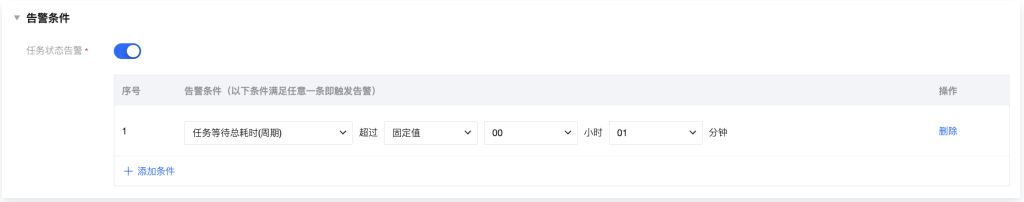
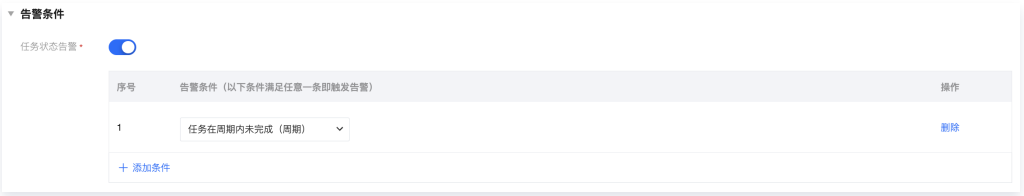
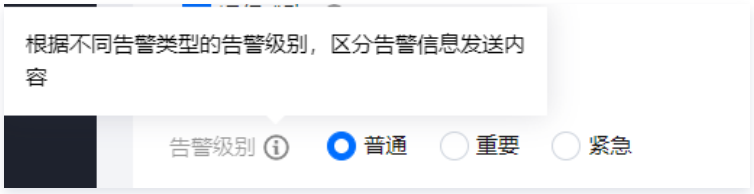


- Task completion time (period): Calculated from the task instance run start time. An alarm is triggered if not completed by the prescribed time point. A "fixed value" or "historical average" can be used as the alarm condition.
  - Fixed value: If the instance does not complete execution before the designated hour/minute time point, the alarm rule is triggered.
  - Historical average: Calculate the average of the instance running times of the most recent 10 successful executions of a computing task, after removing the maximum and minimum values. The setting is invalid if there are fewer than 10 executions.

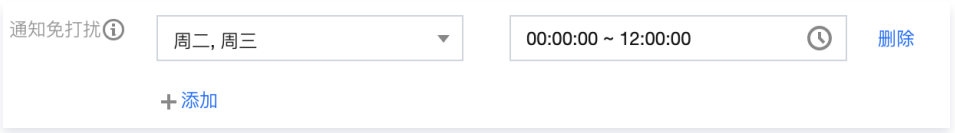


- Task waiting total time: Select task waiting total time (period) to indicate the interval time from the scheduled task time to the start of scheduling run time. If it exceeds the threshold and is not running, an alarm will be triggered. Select task waiting total time (rerun, supplementary entry) to indicate the interval time from the submission time to the start of scheduling run time for rerun and supplementary entry instances. If it exceeds the threshold and is not running, an alarm will be triggered. You can use "fixed value" or "historical average" as the alarm condition.
  - Fixed value: If the instance does not start within the specified hour minutes time requirement, the alarm rule is triggered.

©2013–2025 Tencent Cloud. All rights reserved.

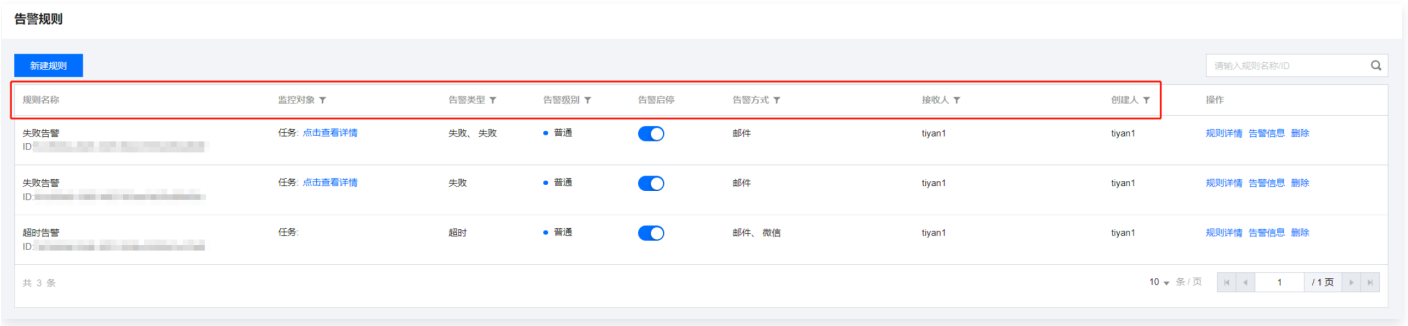
		○ Historical average: Calculate the average of the waiting durations of the most recent 10 successful executions of a computing task, after removing the maximum and minimum values. The setting is invalid if there are fewer than 10 executions.  • Task not completed within a period (period): An alarm is triggered when a task instance fails to complete within its current runtime cycle. Period = interval * period unit. For example: ○ Minute-based task: A 15-minute interval task has a period of 15 minutes. If the task runs for more than 15 minutes without completion, an alarm is triggered. ○ Hourly task: If the specified hour or interval is 1, the period is 1 hour; if the interval is 2, the period is 2 hours, and so on. ○ Day, week, month, and year tasks: The period is 1 day for all. 
	Operation successful	Trigger an alarm when the monitored task's generated instance runs successfully. Supports configuration for periodic execution, backfilling, or rerun execution.
Alarm notification	Alarm Level	Alarm severity of different types, differentiate alarm message content. Currently provides three alarm type selections: ordinary, important, and urgent. 
	Recipient	After an alarm rule is triggered, it sends alarm messages to the recipient. Currently supports three ways to set the alarm message recipient after the alarm rule is triggered: "designated personnel", "task owner", and "duty schedule". • Designated personnel: You can specify any one or more users as alert message recipients. • Task owner: Use the responsible person of the computing task as the alert message recipient. • Duty schedule: Use the scheduled duty schedule as the recipient to send alarm messages to duty users. After a user in the duty schedule clicks **Confirm Alarm** for an alarm message, users under the same duty schedule will no longer receive messages for this alarm. 
	Alert mode	After an alarm rule is triggered, the notification channels for sending alarm messages. Currently supports email, SMS, WeChat, phone, WeCom, HTTP, WeCom group, Lark Group, and DingTalk group push methods. Mobile phone, WeChat, email, and other accounts can be configured in the **Tencent Cloud Personal Center > Cloud Access Management > User** module. WeCom accounts are configured in the **Tencent Cloud Personal Center > Cloud Access Management > Associated Account**. HTTP is configured in the alarm channel.

©2013–2025 Tencent Cloud. All rights reserved.

		
	Notification frequency	Support defining the number of times an alarm is sent and the interval time between each sending. 
	Notification muting	Support setting the do not disturb period for notifications. Alarms will not be sent during the do not disturb period. Users can view alarm records in the alarm information. Do not disturb supports configuration based on week and time, and supports configuring multiple do not disturb periods. 

View Alarm Rule List

After an alarm rule is created, it will be displayed in the alarm rule list. The list shows the rule name, alarm type, alert mode, recipient, and other information. It also provides features such as rule switch and rule detail to help users manage and maintain alarm rules.



Feature Description

Information	Description
Rule Name	Display alarm rule name and ID.
Monitoring Object	Display tasks, workflows, and projects that have active alarm rules, and can view computing tasks involved in the alarm rules under the monitored objects.
Alarm Type	Display monitoring types for alarm rules: failure, timeout, success.
Alarm Level	Display alarm levels for alarm rules: ordinary, important, urgent.
Alarm On-Off	Display the current startup status of the alarm rule. It can be manually enabled or disabled. When in a stopped status, the alarm rule will not take effect, and no alarm information will be generated.
Alert mode	Display the notification channels for the alarm rule's alarm information.
Recipient	Display the alert message recipients of the alarm rule configuration.

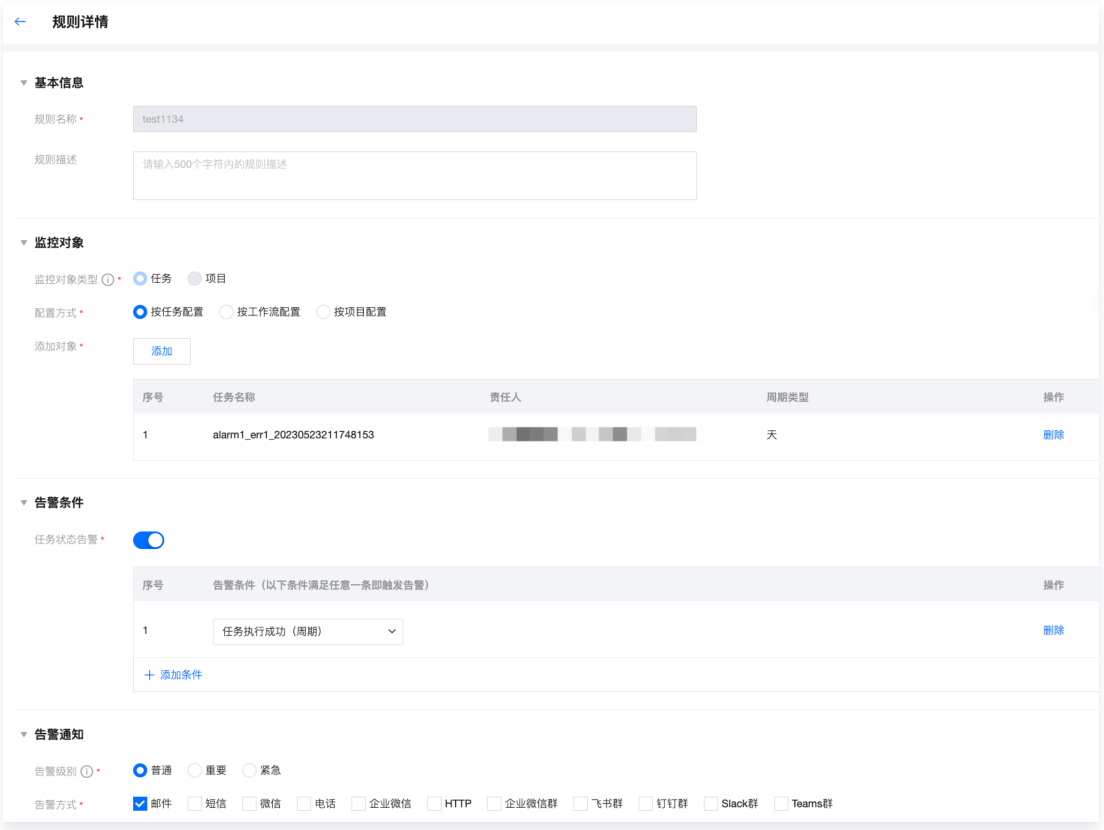
©2013-2025 Tencent Cloud. All rights reserved.

Creator	Display the creator of the current alarm rule.

Performing Alarm Rule Operations



Feature Description

Information	Description
Rule Detail	You can view various parameters set during alarm rule configuration through the rule detail, including the rule name, monitored object, monitoring task, alarm condition, and alarm notification. 
Alarm information	Navigate to the alarm information list page generated after the corresponding alarm rule is triggered, where you can view the alarm information details generated each time the alarm rule is triggered. 
Delete	Display monitoring types for alarm rules: failure, timeout, success.

©2013–2025 Tencent Cloud. All rights reserved.



Filter Alarm Rules

Enter the alarm rule name or ID in the search box to filter the list.

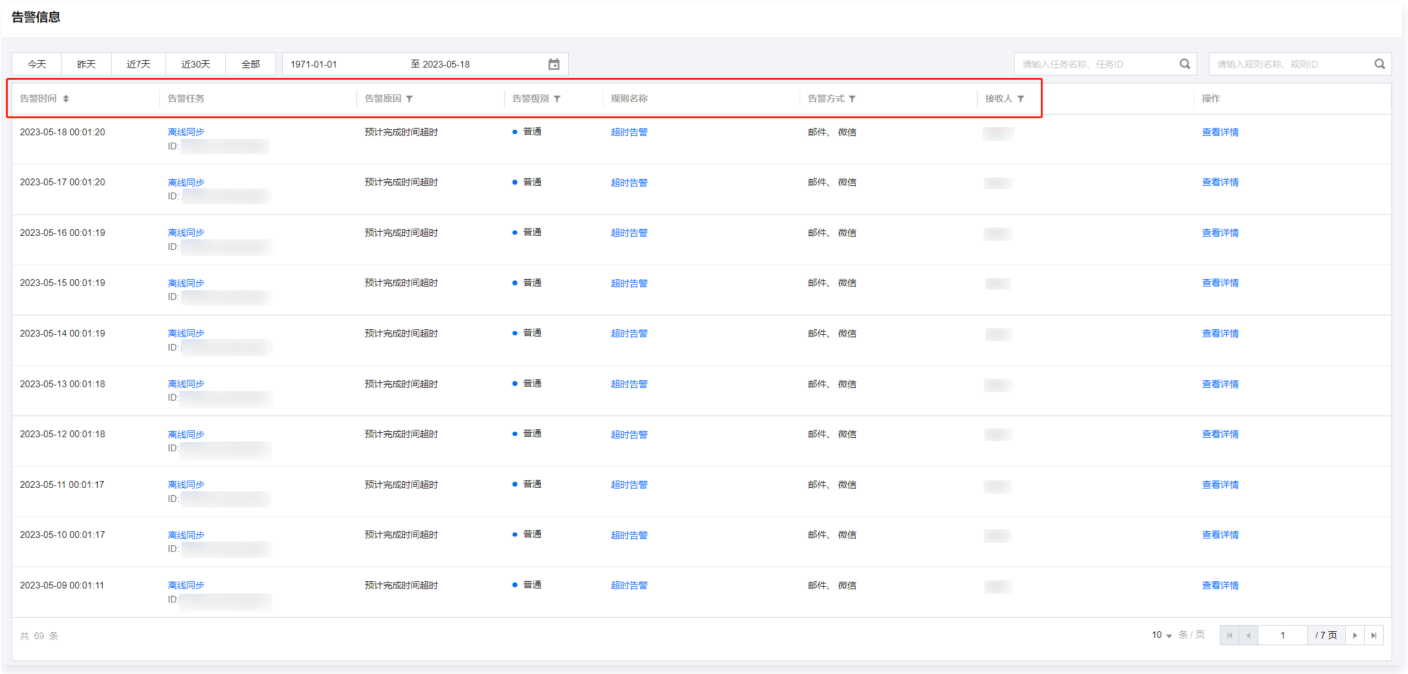


Alarm Information

Alarm information generated after an alarm rule is triggered against a monitored object will be displayed in the alarm information list. The list provides details and running logs of the alarm information, as well as basic information viewing functionality.

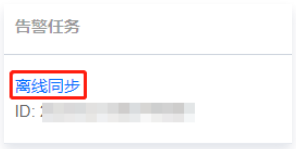
View Alarm Information List

1. Log in to the WeData Console.
2. Click the project list in the left menu to find the target project that requires data management feature operation.
3. Select a project, then click to enter the data development module.
4. Click **Alarm Information** in the left menu to enter the Alarm Information Management page.



Feature Description

©2013–2025 Tencent Cloud. All rights reserved.

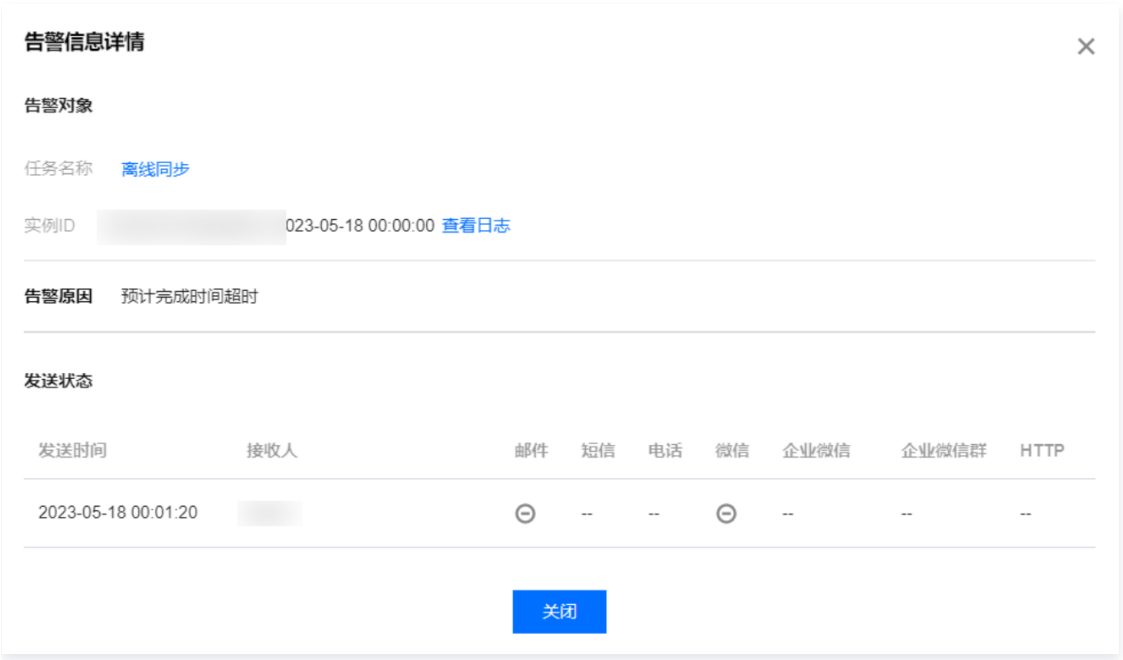
Information	Description
Alarm Time	Display the generation time of alarm information.
Alarm Task	Display the task instance name and instance ID that triggered the alarm information. Click the instance name to navigate to the corresponding instance's management page. 
Alarm Cause	Display the triggering cause of the current alarm information.
Alarm Level	Display the alarm level of alarm information: ordinary, important, urgent.
Rule Name	Display the alarm rule that triggered this item's alarm information. Click the rule name to navigate to the management page of the corresponding alarm rule. 
Alert mode	Display the notification channels for the alarm information.
Recipient	Display the recipients of the alarm information.

Perform Operations on Alarm Information

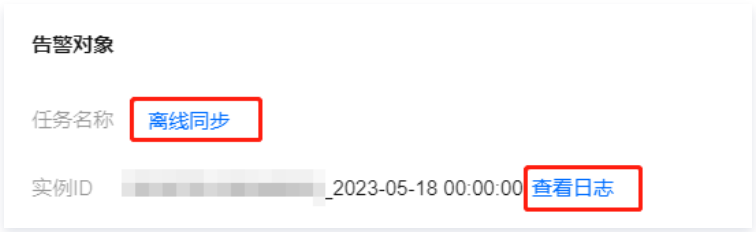


Click **View Detail** in the operation column. In the pop-up window, you can see the alarm object, alarm reason, and message sending status of the alarm information.

©2013–2025 Tencent Cloud. All rights reserved.



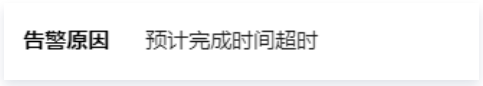
Function Description

Information	Description
Alarm object	Display the task instance name and instance ID that triggered the alarm information.  • Task name: Display the task name of the computing task that triggered the alarm information. Click the task name to navigate to the page where the instance that triggered the alarm is located. • Instance ID: Display the instance ID that triggered the alarm information. Click to view logs to navigate to the log information page of the corresponding instance.
Alarm Cause	Display the triggering cause of the current alarm information based on the trigger conditions of the configured alarm rules. For example, alarm condition selection: execution timeout > estimated completion time.

©2013–2025 Tencent Cloud. All rights reserved.



The alarm reason displayed after the rule is triggered is: expected completion time timeout.



Sending Status

Display the send time, recipient, and sending channels of the current alarm information. The sending status of each channel can be seen to determine whether it was sent successfully.



- Send Time: Displays the time when the alert information is sent to the recipient after the rule is triggered.
- Recipient: Display the recipients of the alarm information.
- Notification channels: Use different icons to display the alarm information sending status of each channel.

Filter Alarm Information

- Filter by task name and task ID: Support filtering the generated alarm information based on the calculation task name and task ID that triggered the alarm.



- Filter by rule name and rule ID: Support filtering the generated alarm information based on the alarm rule name and rule ID that triggered the alarm.

©2013–2025 Tencent Cloud. All rights reserved.



©2013–2025 Tencent Cloud. All rights reserved.

SQL Tag Statement

Last updated: 2024-11-25 14:41:16

WeData's development and orchestration space supports the SQL Tag statement feature, allowing users to use if-else, for loops, and other logical judgments within SQL to implement more complex business logic with simpler code.

Special Identifier at the Beginning of the Script

```
--SQL/VMSCRIPT
#set( $cnt = 10 )
select  $cnt;
```

Assignment Statements and Operators

```
--SQL/VMSCRIPT
#set( $cnt = 10 )
#set( $cnt = $cnt + 1  )
select $cnt;
```

Operator

Operator	Operators	Sample code
Object Equivalence	==	#if($foo == $bar)
Not Equals	!=	#if($foo != $bar)
Less Than or Equal To	<=	#if($foo <= 42)
Less Than	<	#if($foo < 42)
Greater Than or Equal To	>=	#if($foo >= 42)
Greater Than	>	#if($foo > 42)
Equals String	==	#if($foo == "bar")
Equals Number	==	#if($foo == 42)
Boolean NOT	!	#if(!$foo)

for-each Statement

Integer Format

```
--SQL/VMSCRIPT
#set( $cnt = 10 )
#set( $cnt = $cnt + 1  )
#foreach($i in $_range.of(1, $cnt, 1))
select $cnt,$i;
#end
```

List mode

```
--SQL/VMSCRIPT
#foreach($table in ["test_user_info_table", "test_issue_info_table"])
    select '$table' table_name, count(1) cnt from $table where substring(create_time, 1, 7) = '2022-11'
    #if($foreach.last != true)
    union all
```

©2013-2025 Tencent Cloud. All rights reserved.

```
    #end
#end
```

Conditional Judgment Statements

Using Assignment Variables

```
--SQL/VMSCRIPT
--This statement executes the branch where age > 2
#set( $flag_1 = 0 )
#set( $flag_2 = 0 )
select * from  default.demo_hive
where
#if($flag_1 == 1)
id = 1;
#elseif( $flag_2 == 1)
id = 2;
#else
id >2;
#end
```

Using System Parameters, system parameters must follow the ${variable_name} format.
Here, the variables flag_1 and flag_2 are both set to 0 in workflow task scheduling.

```
--SQL/VMSCRIPT
--This statement executes the branch where age > 2
select * from  default.demo_hive
where
#if(${flag_1} == 1)
age = 1;
#elseif(${flag_2} == 1)
age = 2;
#else
age >2;
#end
```

Use system variables + time variables, including dt, dt_pre_7d, and dt_pre_1d as the created time parameter variables.

```
--SQL/VMSCRIPT
--This statement executes the else branch
select * from  default.user_info
where
#if(${is_reload} == "1")
create_time <= ${dt};
#elseif(${is_deltareload} == "1")
create_time  between ${dt_pre_7d} and  ${dt};
#else
create_time  between ${dt_pre_1d} and  ${dt};
#end
```

Scope of Application Explanation

SQL Tags can be used in WeData's Development Space and Orchestration Space.

Conditional Judgment Statements

Implementation uses Velocity, related documentation can be found in Apache Velocity Engine VTL Reference .

©2013–2025 Tencent Cloud. All rights reserved.

> **Note:**
> The SQL Tags feature will automatically take effect for newly purchased Schedule Resource Groups. For older Schedule Resource Groups, assistance from the R&D team is needed for a restart upgrade. It is recommended to purchase new Schedule Resource Groups and then bind the project.

©2013–2025 Tencent Cloud. All rights reserved.

Variable Usage

Last updated: 2024-11-25 14:40:50

The use of variables in data development is divided into three levels: Project Level, Workflow Level, and Task Level. The coverage area of variables differs at each level:

- Project Level: Development scripts, data workflows, and task nodes within the project can be used.
- Workflow Level: Data workflows and the task nodes within their workflows can be used.
- Task Level: Only development scripts and task nodes in the Orchestration space can be created and used.
- Priority: For variables with the same name, Upstream Task Transfer>Task Level>Workflow Level>Project Level.

Project Level

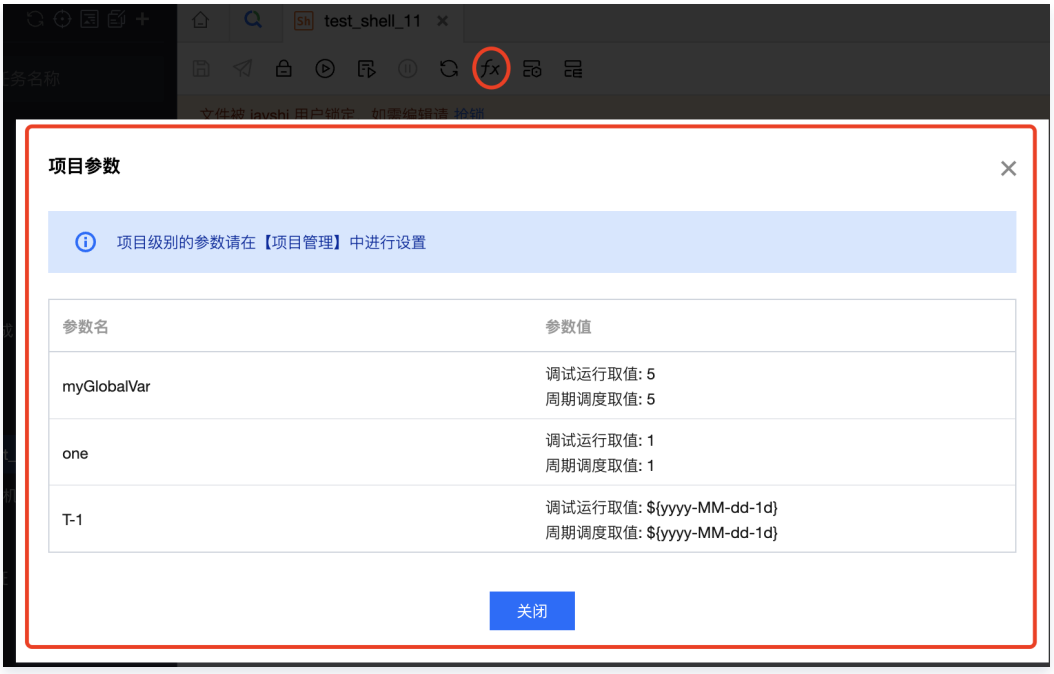
Variable Configuration

Project variables are applied to global data development tasks within the project. The specific configuration process can be referred to in Project Management-Variable Settings.

Variable Usage

Script files created through the development space and computing tasks created in the Orchestration space can all use project variables. Here is an example using SQL Script:

1. In the SQL script, click to view the configured project variables.

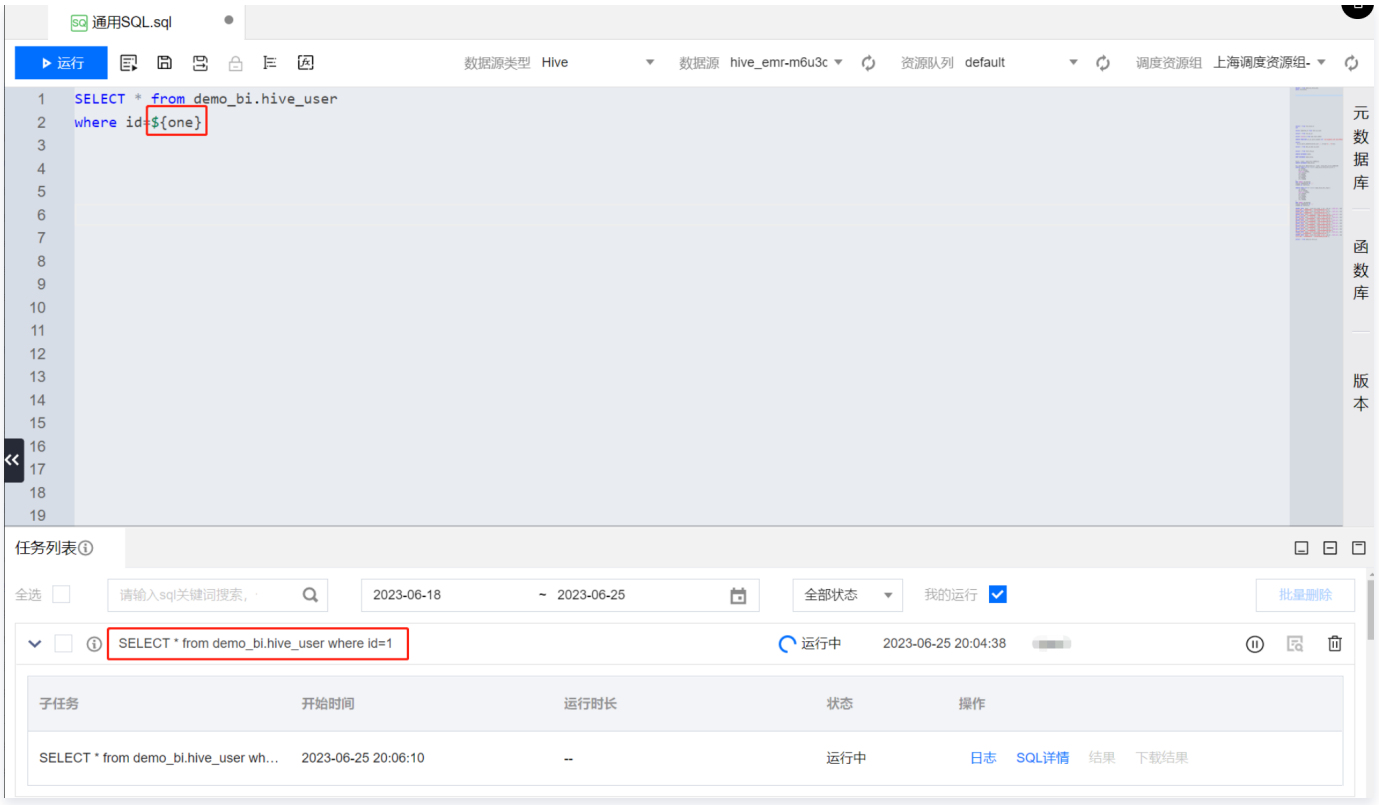


2. In the code, reference the variable name from the Definition of variables with the following format: ${ key }

```
SELECT * from demo_bi.hive_user
where id=${one}          #Here, ${one} references the 'one' global variable
```

3. From the diagram, you can see that after referencing the global variable 'one', the value of the variable is retrieved during SQL execution.

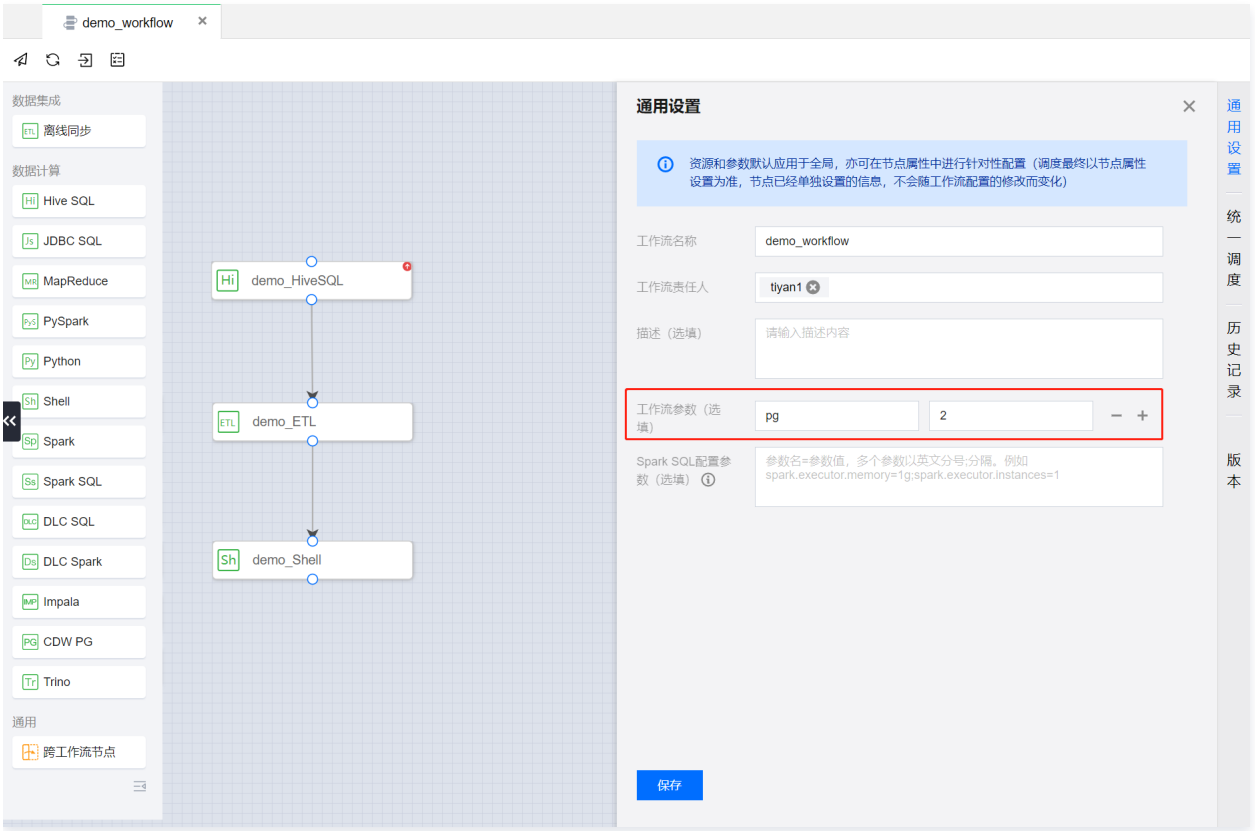
©2013-2025 Tencent Cloud. All rights reserved.



Workflow Level

Variable Configuration

Workflow level variables are applied to computational task nodes created within the corresponding workflow. Through the general settings on the right side of the workflow canvas page, you can define workflow variables.

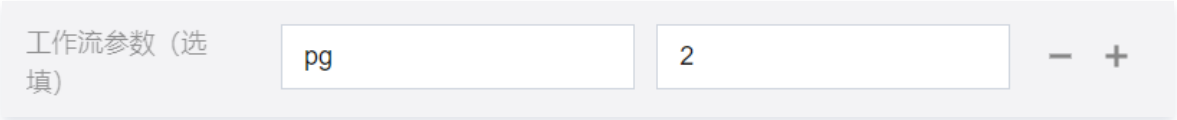


Variable Usage

©2013–2025 Tencent Cloud. All rights reserved.

In a data workflow, after creating a computational task node and entering the task node configuration page, you can use workflow variables during the configuration process. The following example configures variables in the demo_workflow workflow and uses them in a Hive SQL computational task within the workflow:

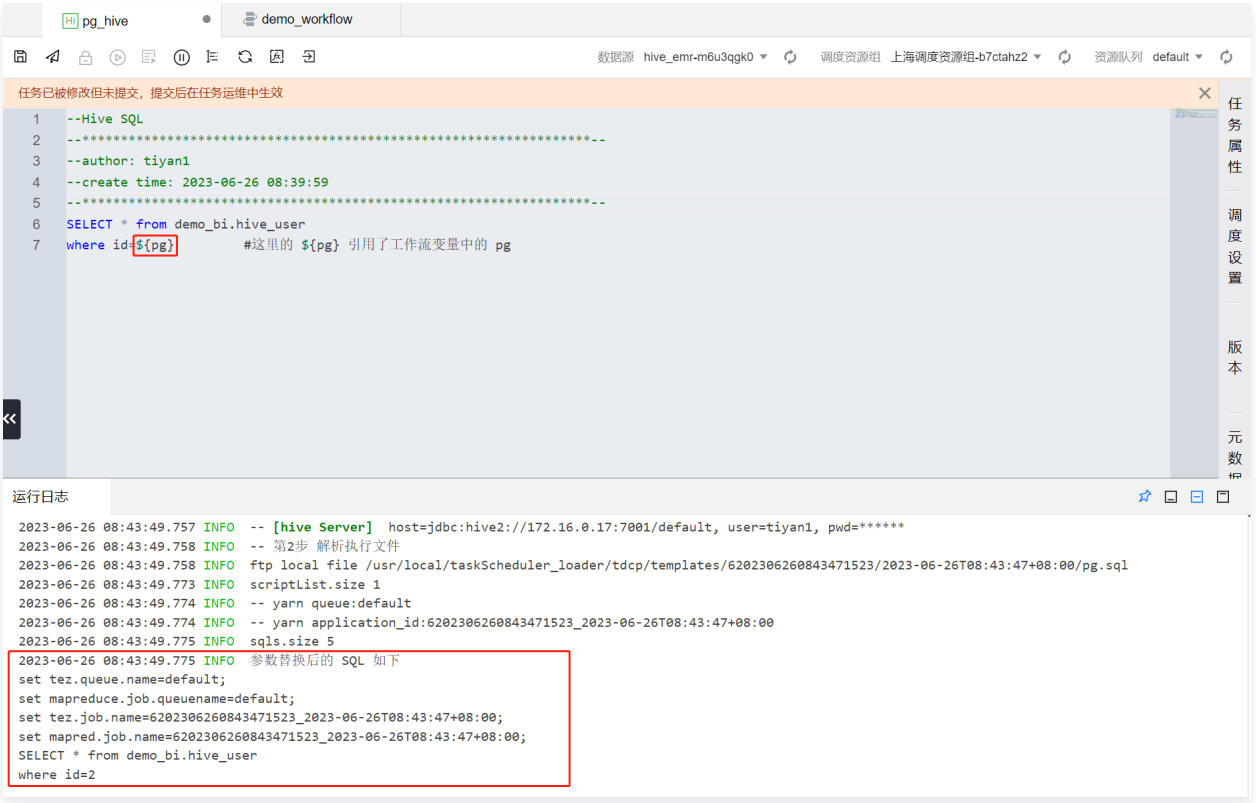
1. Configure the following workflow parameters in the general settings of the demo_workflow workflow.



2. Then enter the Hive SQL computational task within the workflow and reference the variable name of the workflow parameters in the code, formatted as: ${key}

```
SELECT * from demo_bi.hive_user
where id=${pg} # Here, ${pg} references the pg variable from the global variables
```

3. As shown in the diagram, when the pg workflow variable is referenced in the computational task, the corresponding code undergoes parameter substitution at runtime.

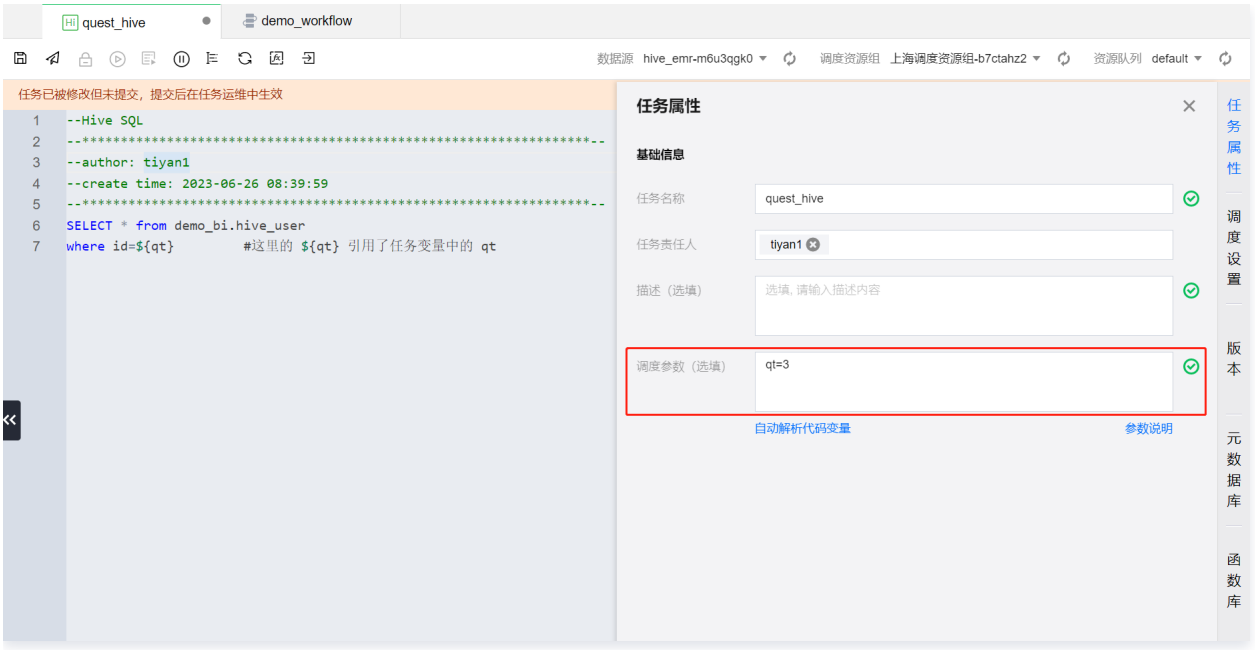


Task Level

Variable Configuration

Task level variables apply to computational task nodes in a data workflow. Each computational task can independently configure task variables suitable for itself. Through the task attributes on the right side of the computational task configuration page, you can define task variables for the corresponding task.

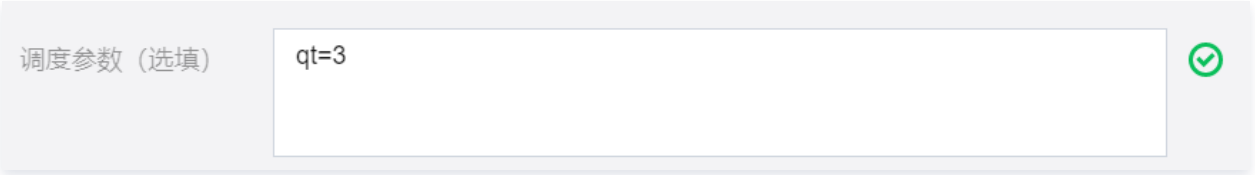
©2013–2025 Tencent Cloud. All rights reserved.



Variable Usage

Enter the computational task node and configure the task variables in the scheduling parameters section of the task attributes. The following uses a Hive SQL task as an example:

1. Configure the following task variables in the scheduling parameters of the quest_hive computational task.

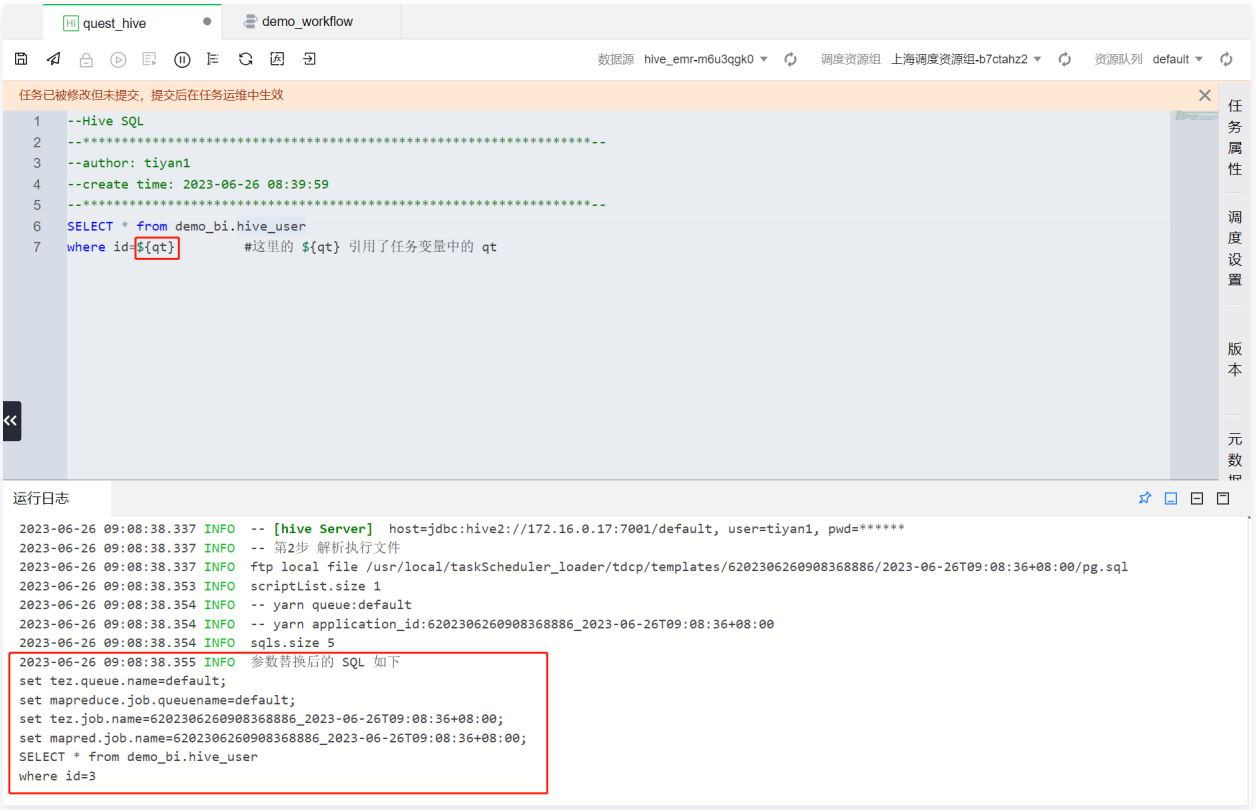


2. Then, when developing code in quest_hive, apply the variable name formatted as: ${VariableName}.

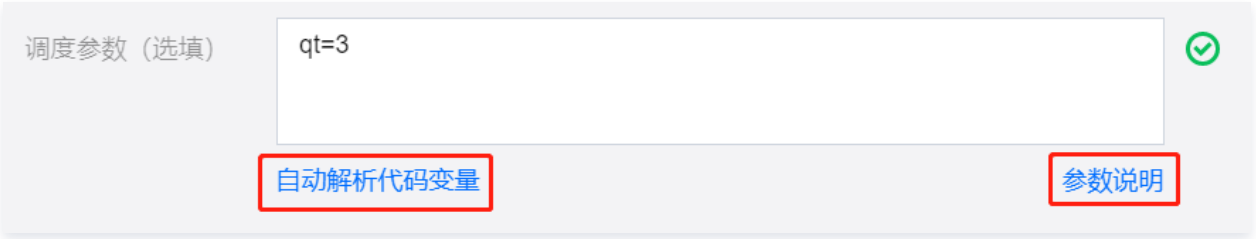
```
SELECT * from demo_bi.hive_user
where id=${qt} # Here, ${qt} references the qt variable from the task variables
```

3. As shown in the diagram, when the qt task variable is referenced in the computational task, the corresponding code undergoes parameter substitution at runtime.

©2013–2025 Tencent Cloud. All rights reserved.



4. In addition, providing detailed parameter descriptions during task variable configuration can help users quickly understand and use the feature; when clicking on the automatic code variable parsing feature during task variable configuration, a variable list will pop up, allowing you to check whether the variable configuration meets expectations.



©2013–2025 Tencent Cloud. All rights reserved.



5. You can view the parameter variable information available for the current task in the variable list, including project variables, workflow variables, and task variables. Modifying the task variables of the current task configuration and adding task variables for the current task are supported.



©2013–2025 Tencent Cloud. All rights reserved.

6. Different task variables configured within multiple task nodes in the data workflow can facilitate interoperability by providing parameter passing capabilities according to dependency relationships.

Application Variables

Project, workflow, and computing task information can be obtained in the form of application variables. View computing task-related information in SQL tasks and YARN.
Currently supported built-in application variables include:

1. Project Identification: ${projectIdent}
2. Workflow Name: ${workflowName}
3. Task Name: ${taskName}
4. Task ID: ${taskId}
5. Person in Charge: ${taskInCharge}
6. Task Type: ${taskType}

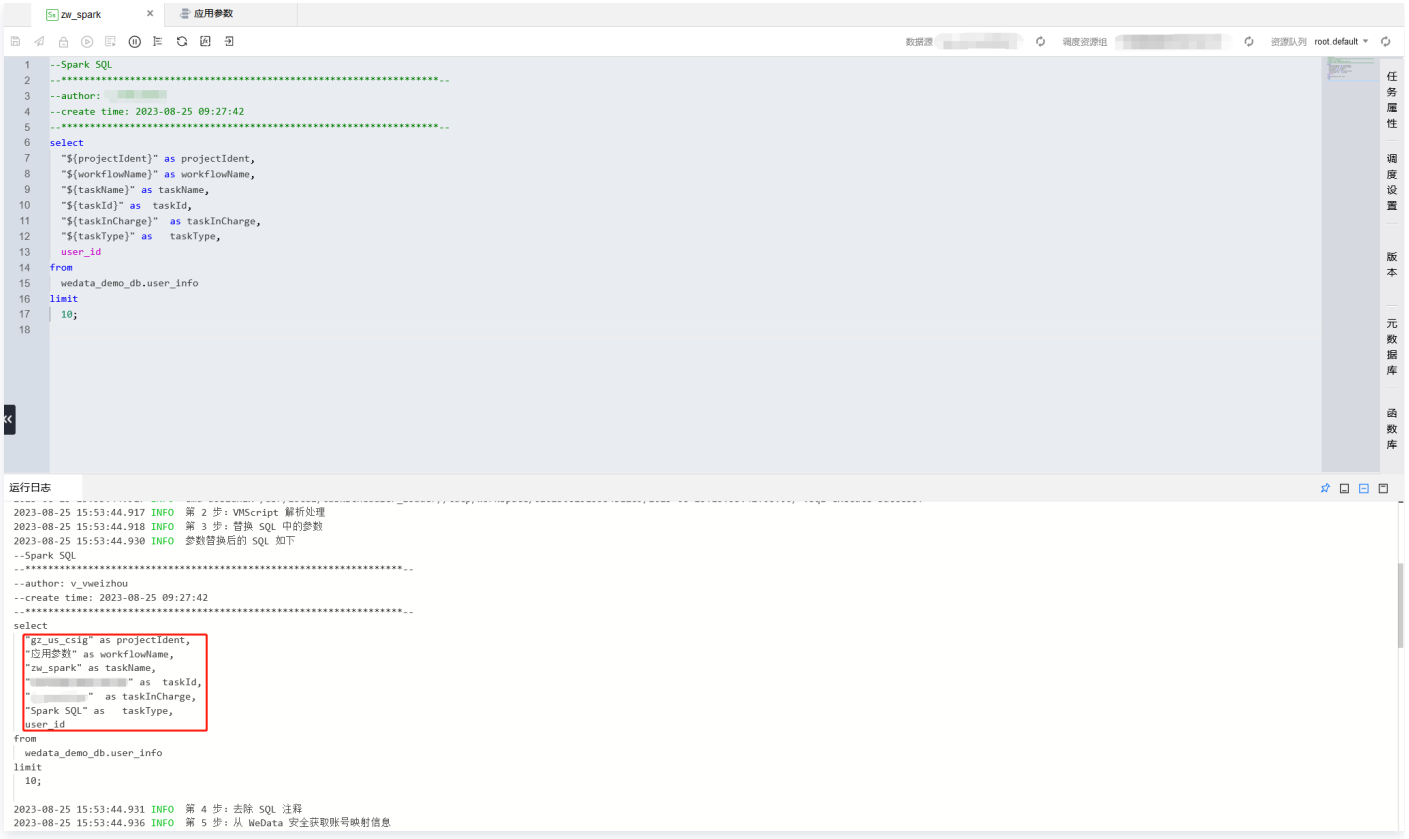
Application scenario

- Retrieve task information in SQL.

  Application variables can be queried in SQL statements to get current project, workflow, and other information.
  Sample SQL:

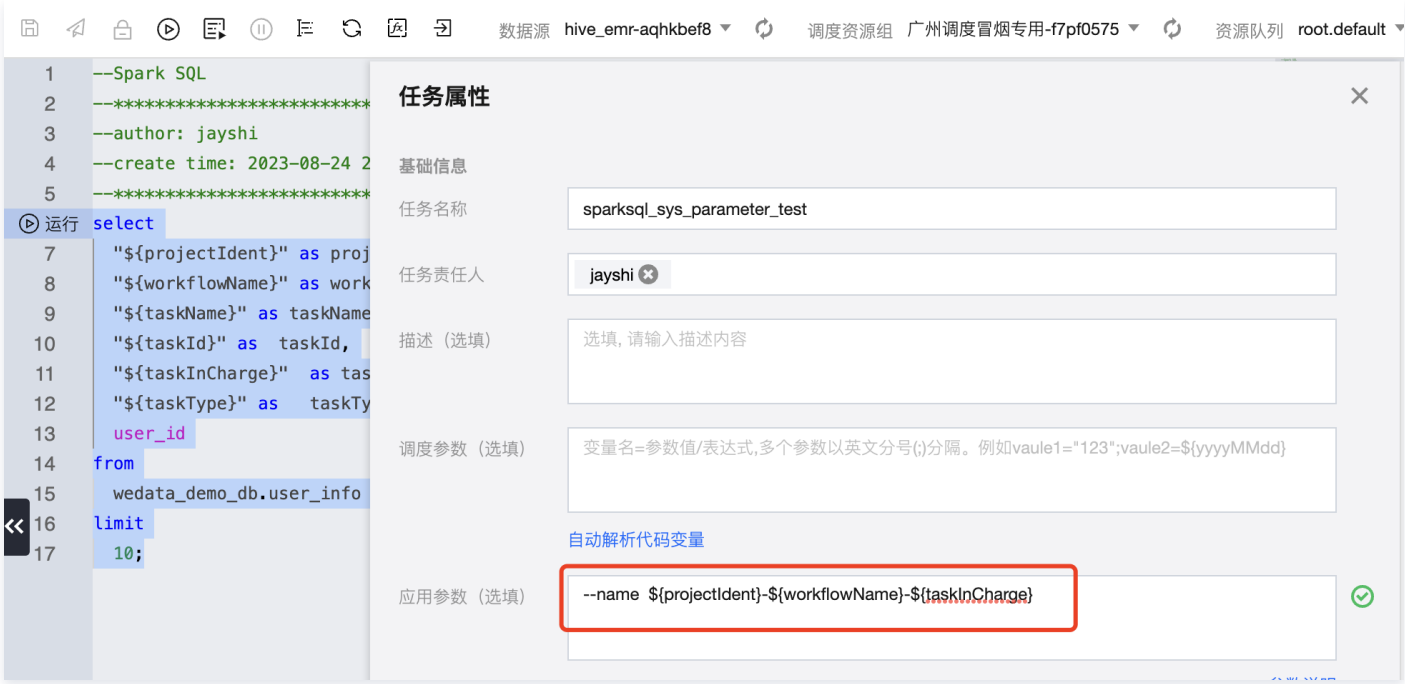
```
select
  "${projectIdent}" as projectIdent,
  "${workflowName}" as workflowName,
  "${taskName}" as taskName,
  "${taskId}" as  taskId,
  "${taskInCharge}"  as taskInCharge,
  "${taskType}" as   taskType,
  user_id
from
  wedata_demo_db.user_info
limit
  10;
```

©2013-2025 Tencent Cloud. All rights reserved.



- Specify the task name on YARN.

  Use --name + variable to specify the task name on YARN for SparkSQL, PySpark, and Spark task types.



Parameter Example:

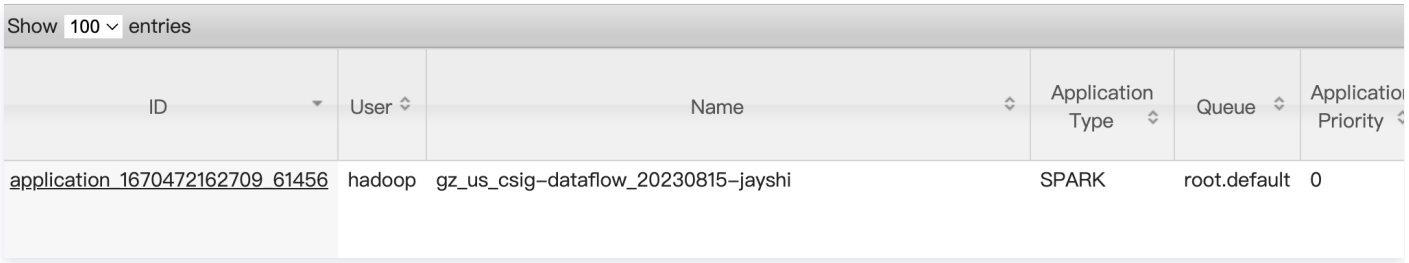
```
--name ${projectIdent}-${workflowName}-${taskInCharge}
```

> **Note:**
> When connecting to the kyuubi data source with SparkSQL, currently the parameter name needs to be specified using spark.app.name.

©2013–2025 Tencent Cloud. All rights reserved.

```
spark.app.name=${projectIdent}-${workflowName}-${taskInCharge}
```

Final effect on YARN:


Show 100 entries

ID	User	Name	Application Type	Queue	Applicatio Priority
application_1670472162709_61456	hadoop	gz_us_csig-dataflow_20230815-jayshi	SPARK	root.default	0

Other Explanations

If a kyuubi job has been submitted before, subsequent kyuubi jobs may reuse the previous YARN application ID, causing the application name to duplicate. This issue is currently being investigated and resolved by the developers.

©2013-2025 Tencent Cloud. All rights reserved.

Batch operation

Last updated：2024-11-25 14:40:22

Operation scenarios

WeData provides batch operations for tasks, allowing users to quickly modify related task properties, submit tasks, and delete tasks.

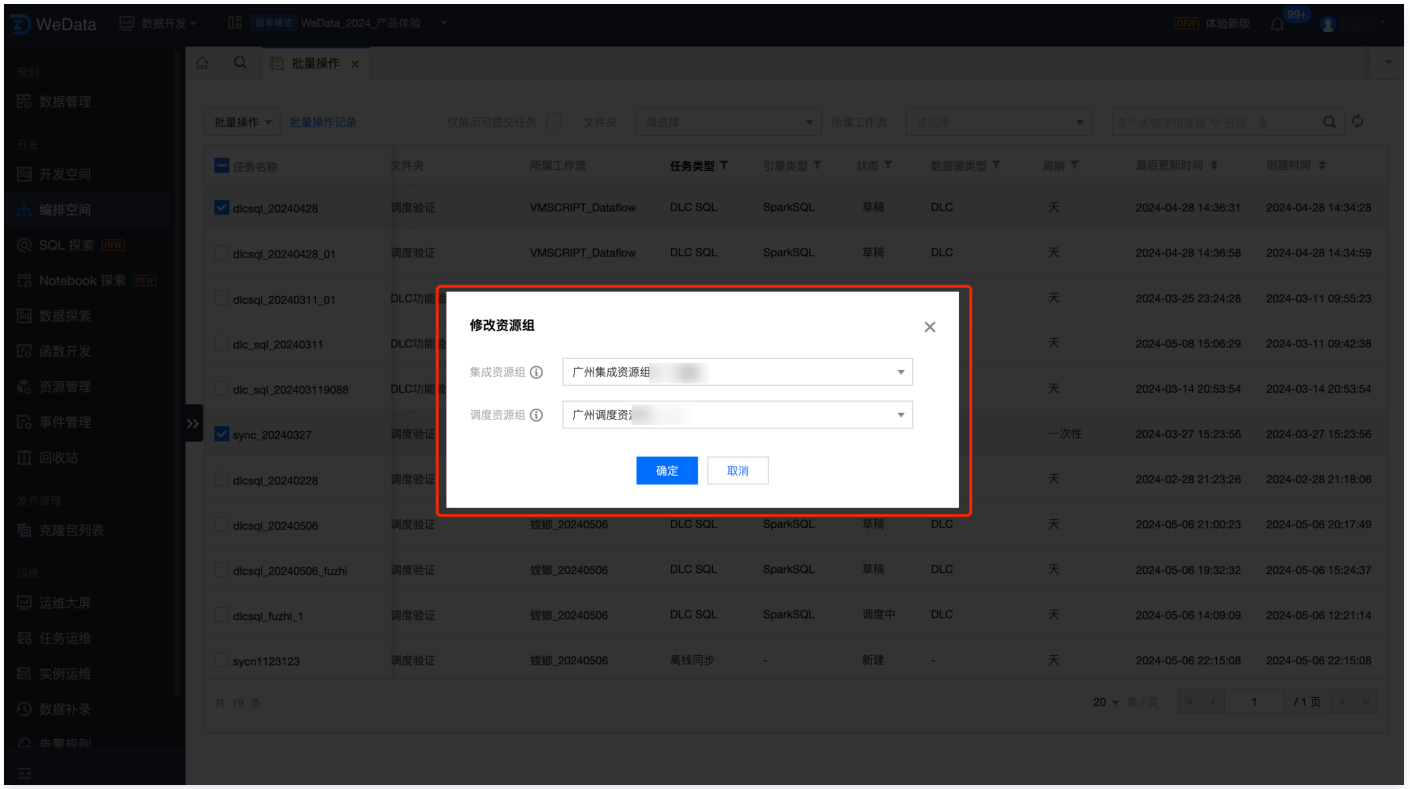
Log in to the WeData console, and navigate to **Data Development** > **Orchestration Space** from the left menu to access batch operations for tasks.



Operation step

Modify Resource Group

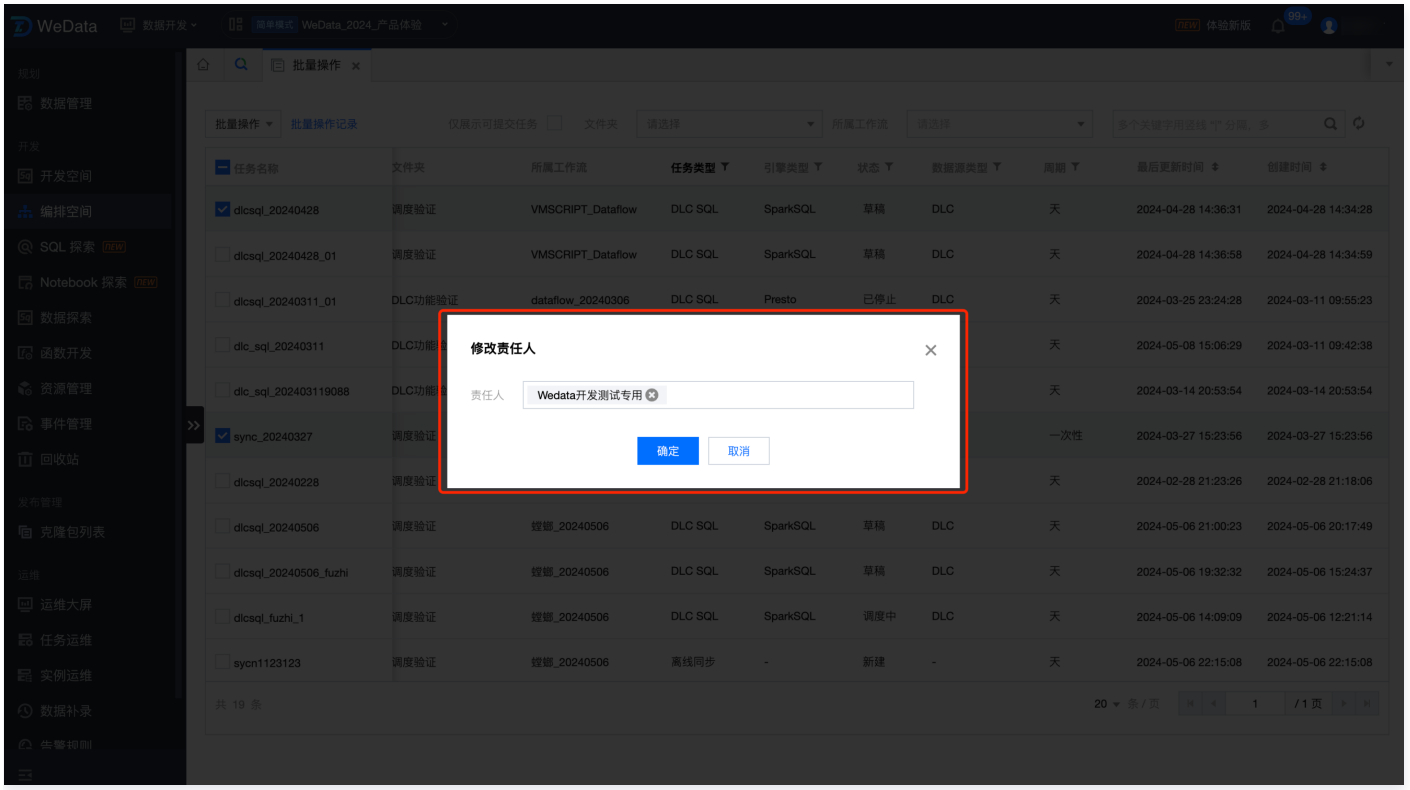
You can batch modify the Integrated Resource Group and Schedule Resource Group for selected tasks. The system will automatically identify if there are data calculation tasks and integration tasks among the selected tasks. If there are corresponding types of tasks, the options for Schedule Resource Group and Integrated Resource Group will automatically pop up, and the selected tasks' resource groups will be updated accordingly.



Change Responsible Person

You can batch modify the responsible person for selected tasks. Batch operations will change the selected tasks' responsible person to the chosen person.

©2013-2025 Tencent Cloud. All rights reserved.

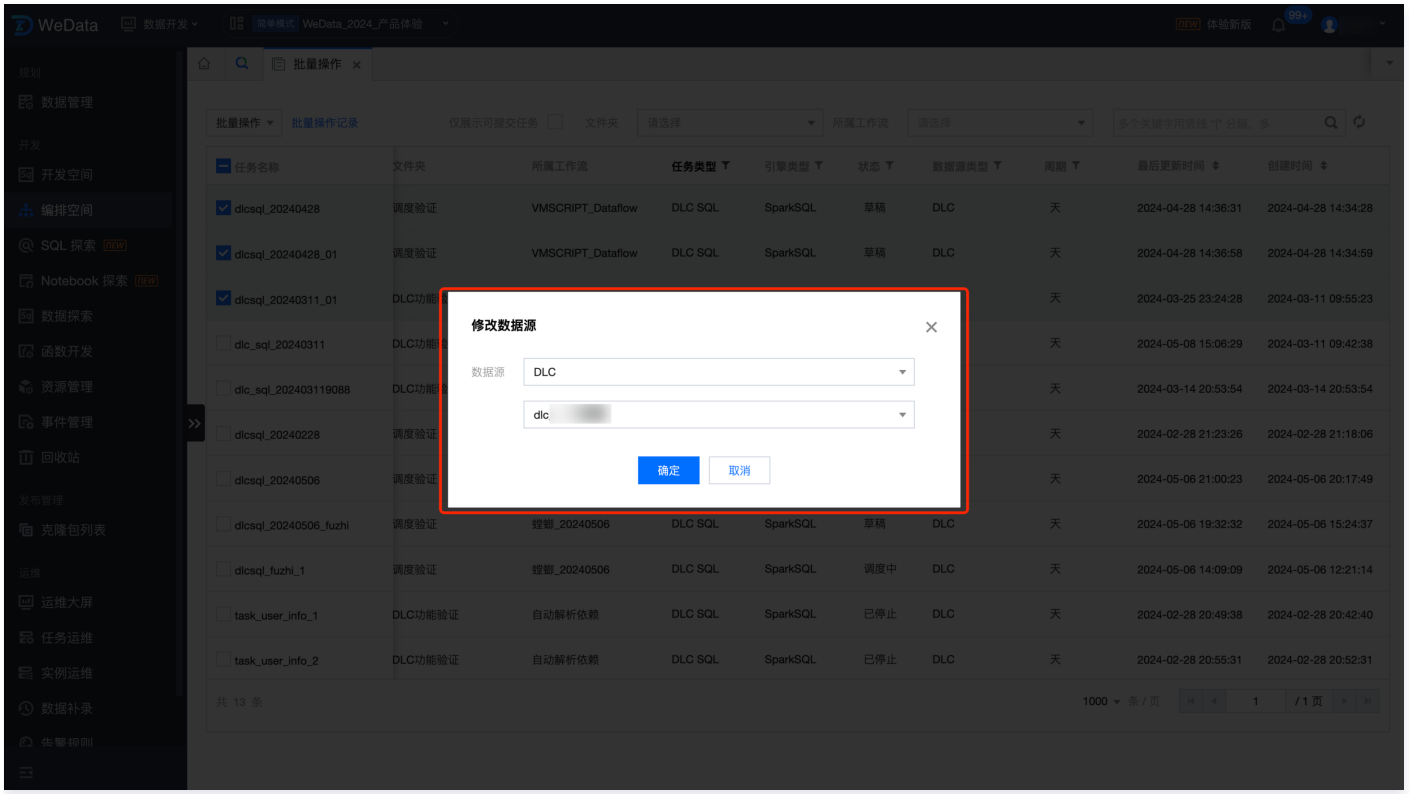


Modify Data Source

> **Note:**
> Currently, batch modification of data sources is only supported for selected tasks of the same task type, specifically Hive SQL, JDBC SQL, DLC SQL, Spark SQL, TCHouse-P, and Trino.

You can batch modify the data source for selected tasks. Batch operations will change the selected tasks' data source to the chosen data source from the dropdown list.

©2013-2025 Tencent Cloud. All rights reserved.



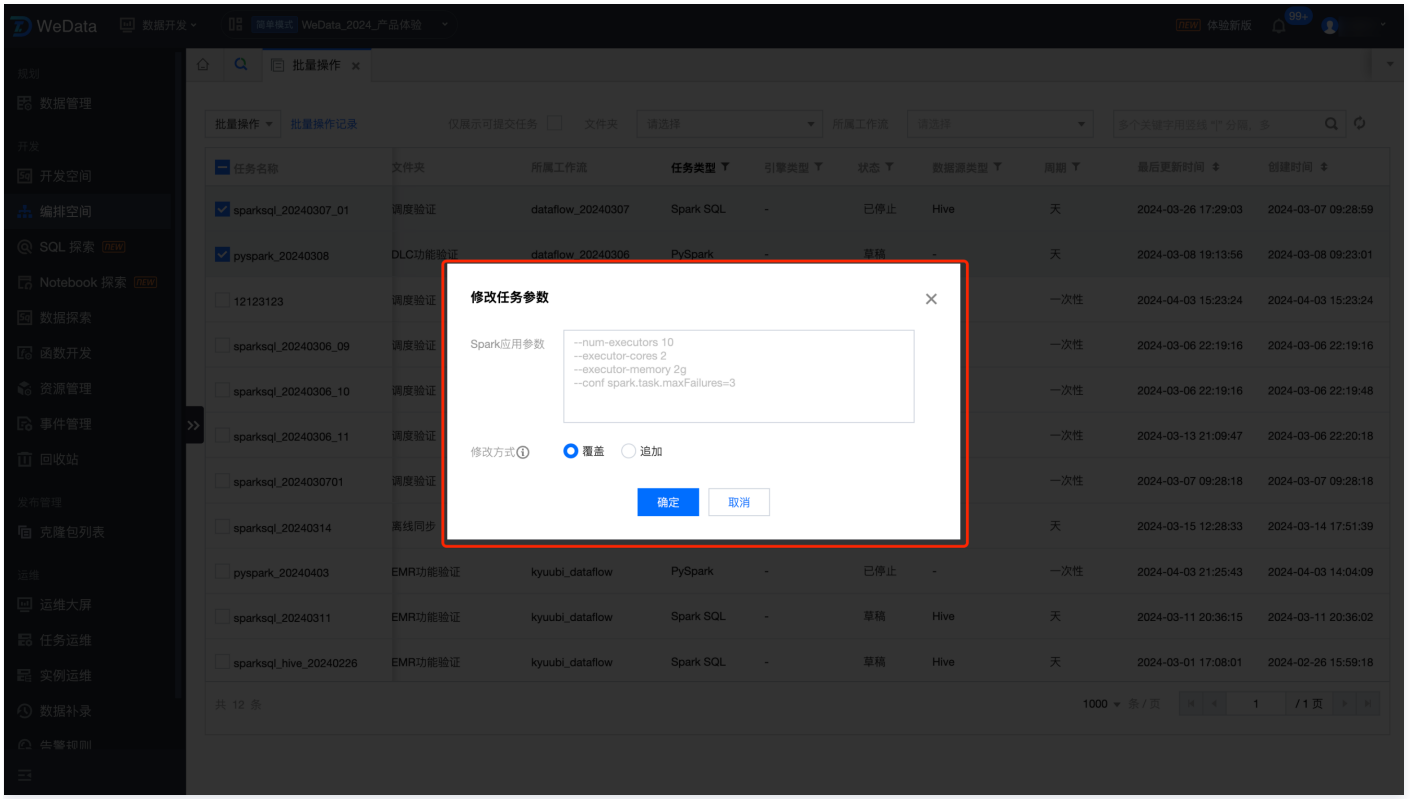
Modify Task Parameters

> **Note:**
> Task parameter modifications must meet one of the following conditions:
> - Tasks with the same engine type for DLC SQL, and the engine type cannot be empty.
> - All selected tasks are of the TCHouse-X type
> - All selected tasks are Spark, SparkSQL, or PySpark tasks.

You can batch modify the parameters of selected tasks. The specific input format for parameters will vary depending on the task type.
For SparkSQL, PySpark, and Spark tasks using the EMR engine, the parameter input content is as follows. Two modification methods are provided: **Overwrite and Append**.

- **Overwrite**: This will delete the existing application parameters of the task and use the application parameters configured in bulk.
- **Append**: This will not delete the existing application parameters of the task, but will only add the application parameters configured in bulk to the task’s existing parameters.

©2013–2025 Tencent Cloud. All rights reserved.

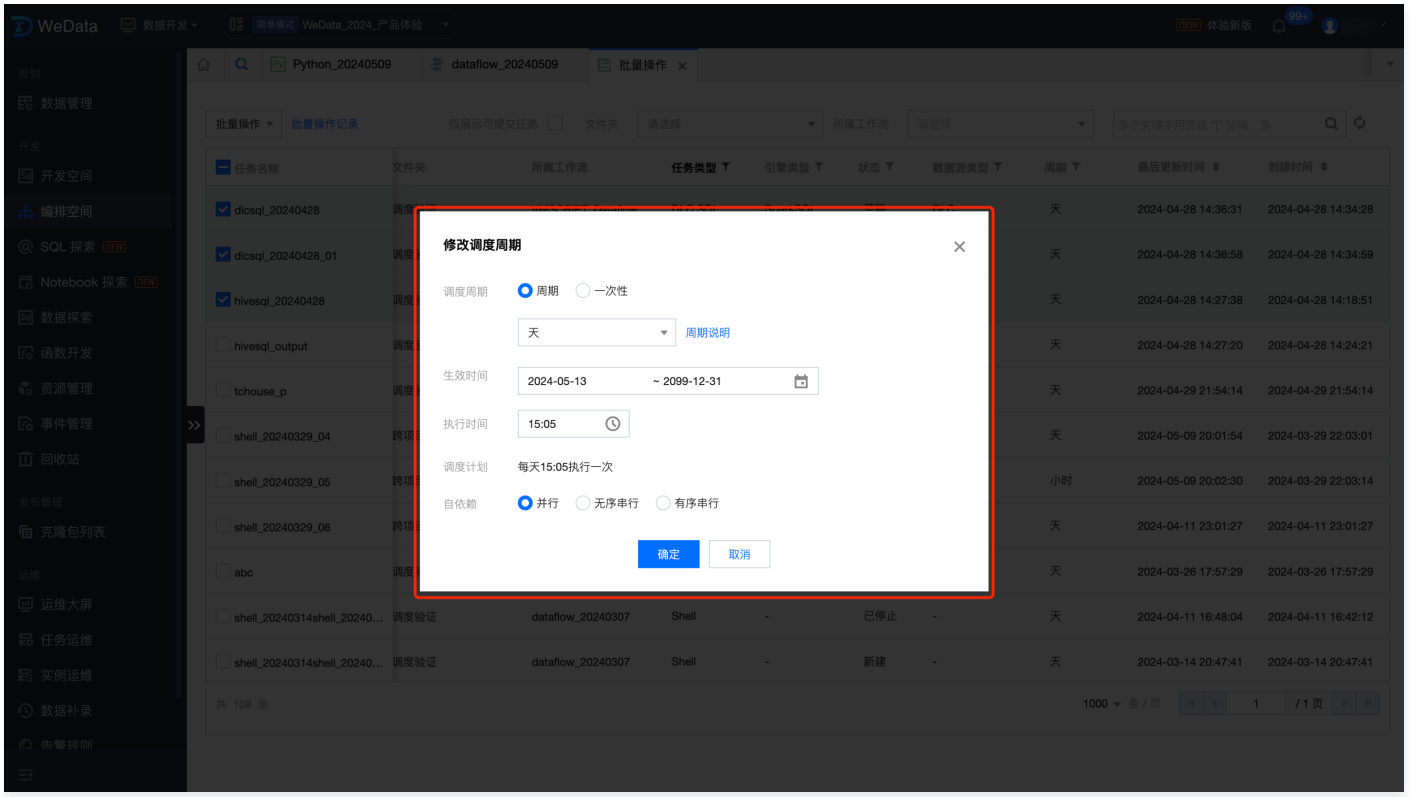


Modify scheduling period

> **Note:**
> Batch modification of the scheduling period will only modify the scheduling period of the selected tasks and will not change the dependencies of upstream and downstream tasks. Batch modification may lead to incorrect upstream and downstream dependencies, causing task submission failures.
> For example, if Task B downstream depends on hourly Task A upstream using a self-defined range, the range will not change automatically after modifications, which may lead to task submission failures.

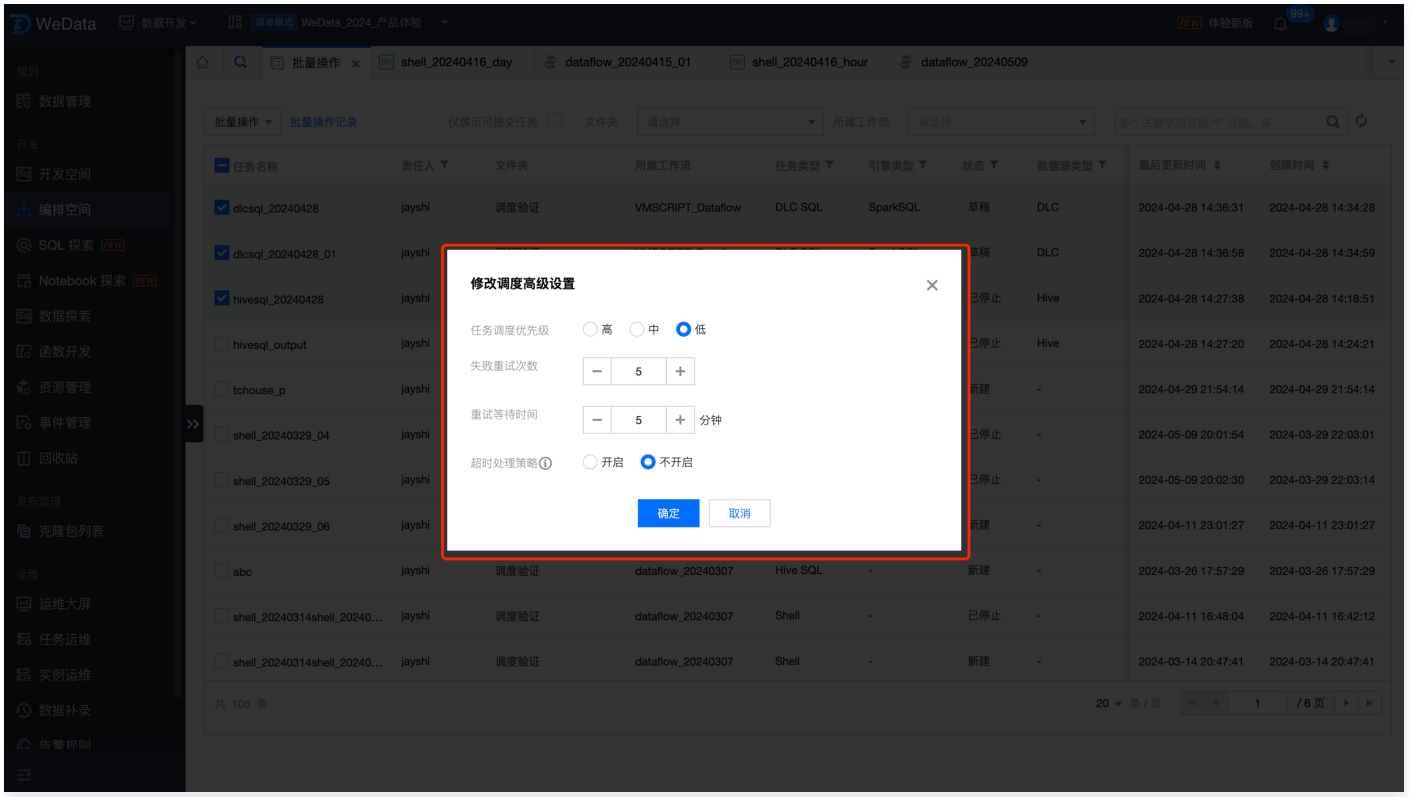
You can batch modify the scheduling period of the selected tasks. The modifiable content is consistent with the scheduling policy of the task's scheduling settings. Batch operations will change the scheduling policy of the selected tasks to the configured scheduling policy.

©2013–2025 Tencent Cloud. All rights reserved.



Modify advanced scheduling settings

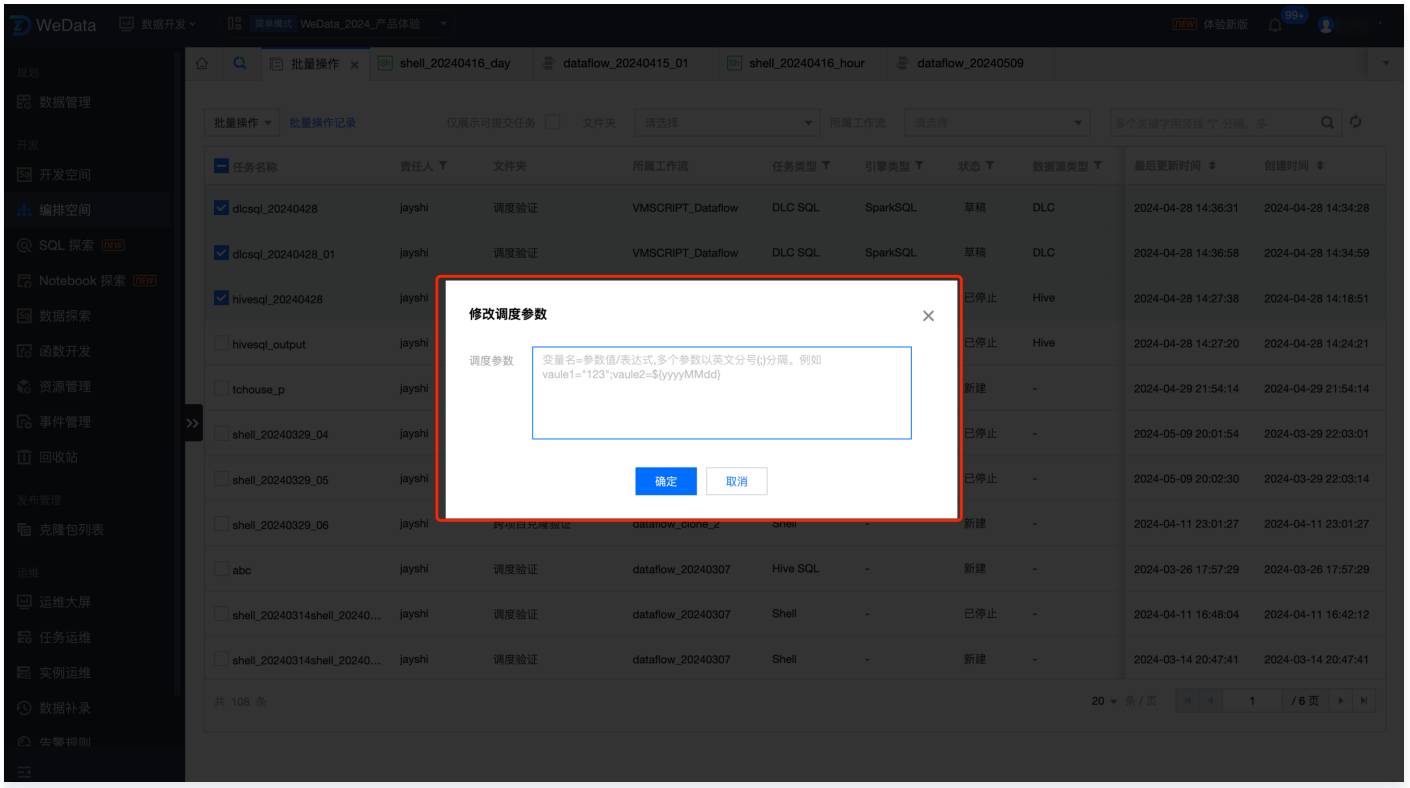
You can batch modify the advanced scheduling configuration of the selected tasks. The modifiable content is consistent with the advanced settings of the task's scheduling settings. Batch operations will change the advanced scheduling settings of the selected tasks to the configured content.



Modify scheduling parameters

©2013–2025 Tencent Cloud. All rights reserved.

You can batch modify the scheduling parameters of the selected tasks. The modifiable content is consistent with the content in the task's **Task Properties** > **Scheduling Parameters**. Batch operations will set the scheduling parameters of the selected tasks to the input scheduling parameters.



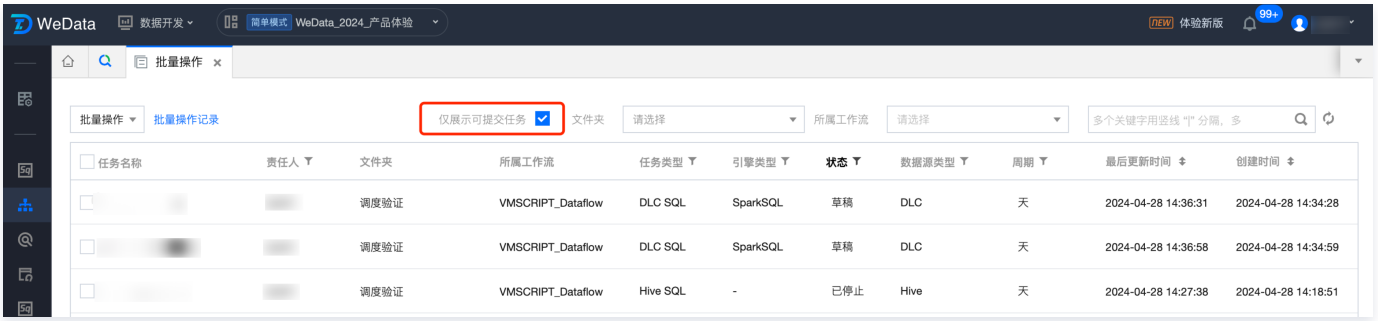
Submit tasks

> **Note:**
> When task submission approval is enabled in **Project Management** > **Basic Information Configuration** > **Approval Configuration**, task submissions will first go to the approval center. Only after approval will the tasks be truly submitted to task operations and maintenance.
> Whether the tasks are successfully submitted depends on the configuration of the tasks themselves and the status of upstream tasks.

You can batch submit the selected tasks, which currently include:

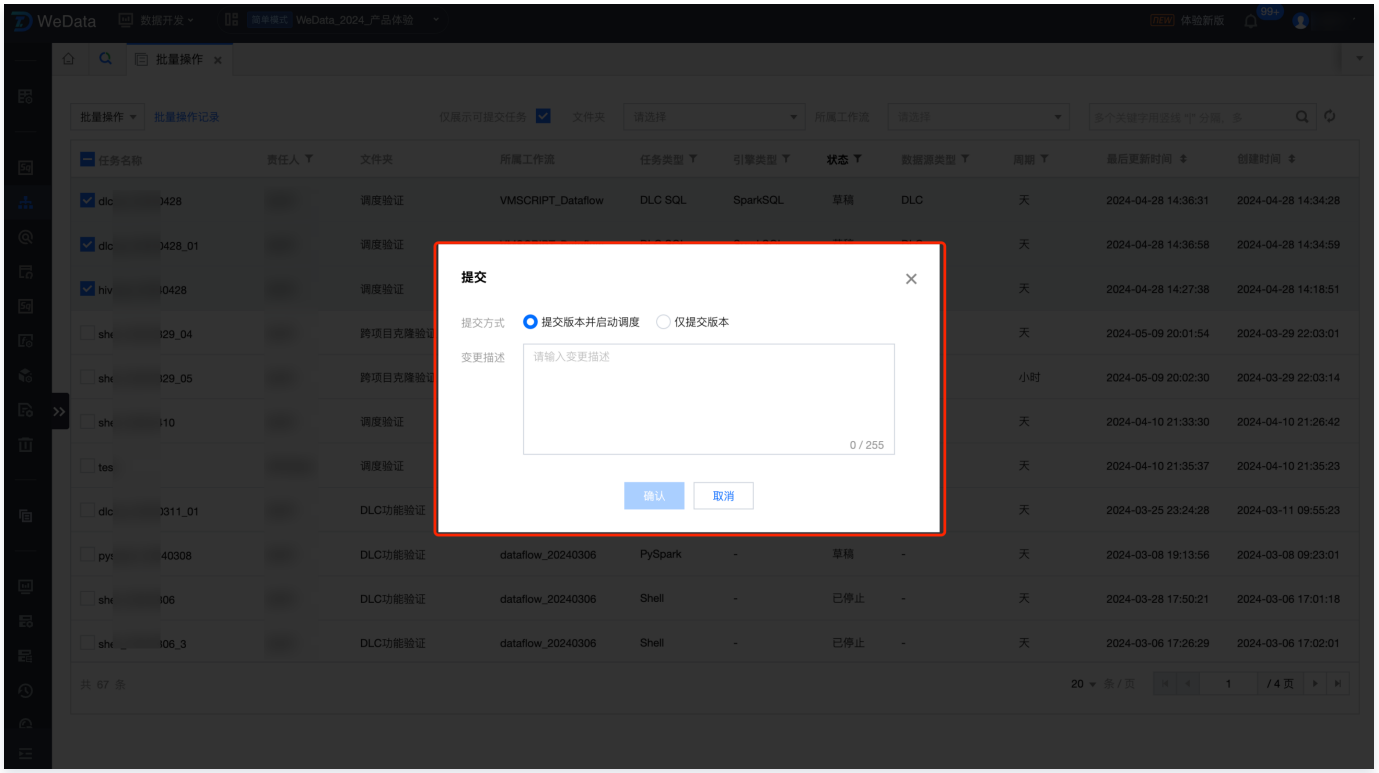
- Draft state tasks.
- Tasks submitted for scheduling but still have modifications.

1. Enter the bulk operation interface, and filter by selecting **Only show submittable tasks** at the top of the list.



2. Select the task you need to submit. Choose the submission method **Submit Version and Start Scheduling**, fill in the change description, and click **Confirm**.

©2013–2025 Tencent Cloud. All rights reserved.



Deleting task

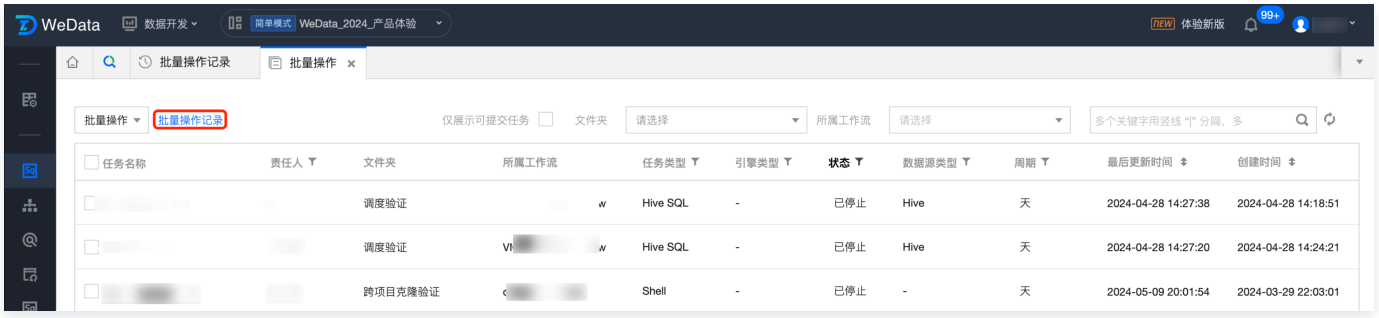
> **Note:**
> Deletion is only applicable to tasks with the status of "Stopped, Draft, New, Pending Review, Approval Rejected"; it will not be executed for other tasks.

Bulk deletion will delete the selected tasks, which will then move to the recycle bin by default. The system will automatically skip tasks that cannot be deleted.

Batch operation record

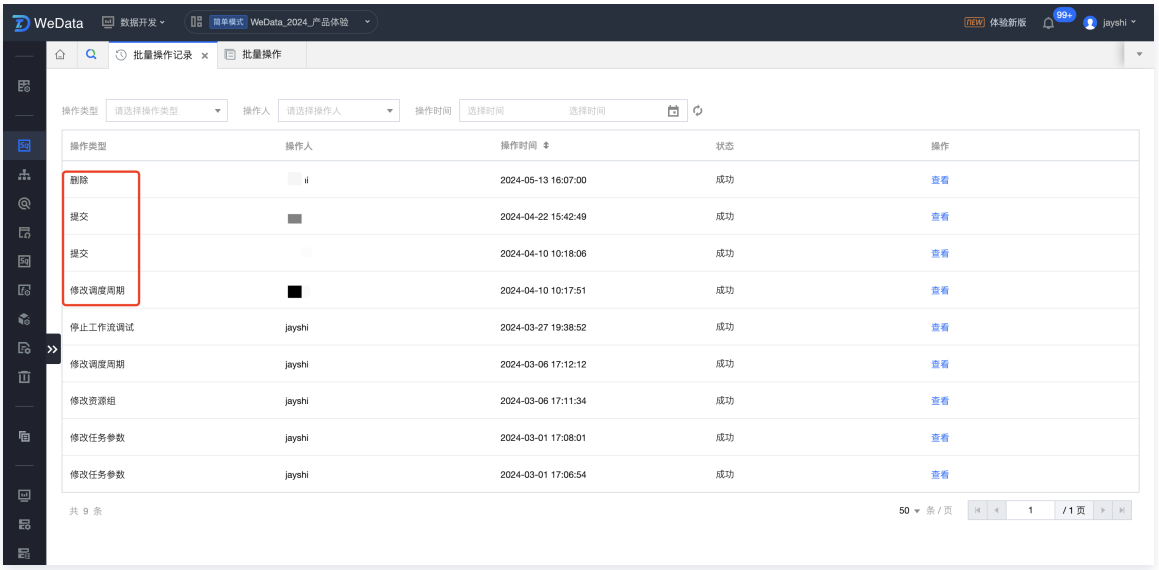
All types of batch operations will be recorded in **Batch Operation Record**, and you can view the progress and results in **Batch Operation Record**.

1. In the bulk operation interface, click **Batch Operation Record**.

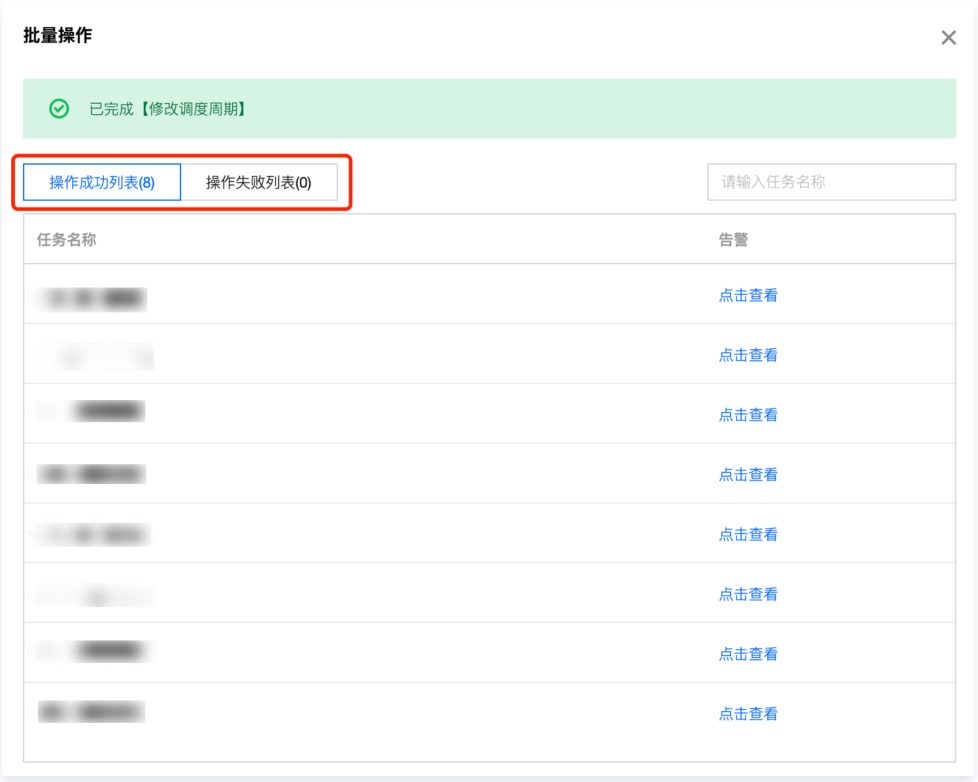


2. Enter the **Batch Operation Record** interface to view the progress and results of all types of batch operations.

©2013–2025 Tencent Cloud. All rights reserved.



3. Additionally, you can switch between the success list and the failure list to see which types of batch operations failed.



©2013–2025 Tencent Cloud. All rights reserved.

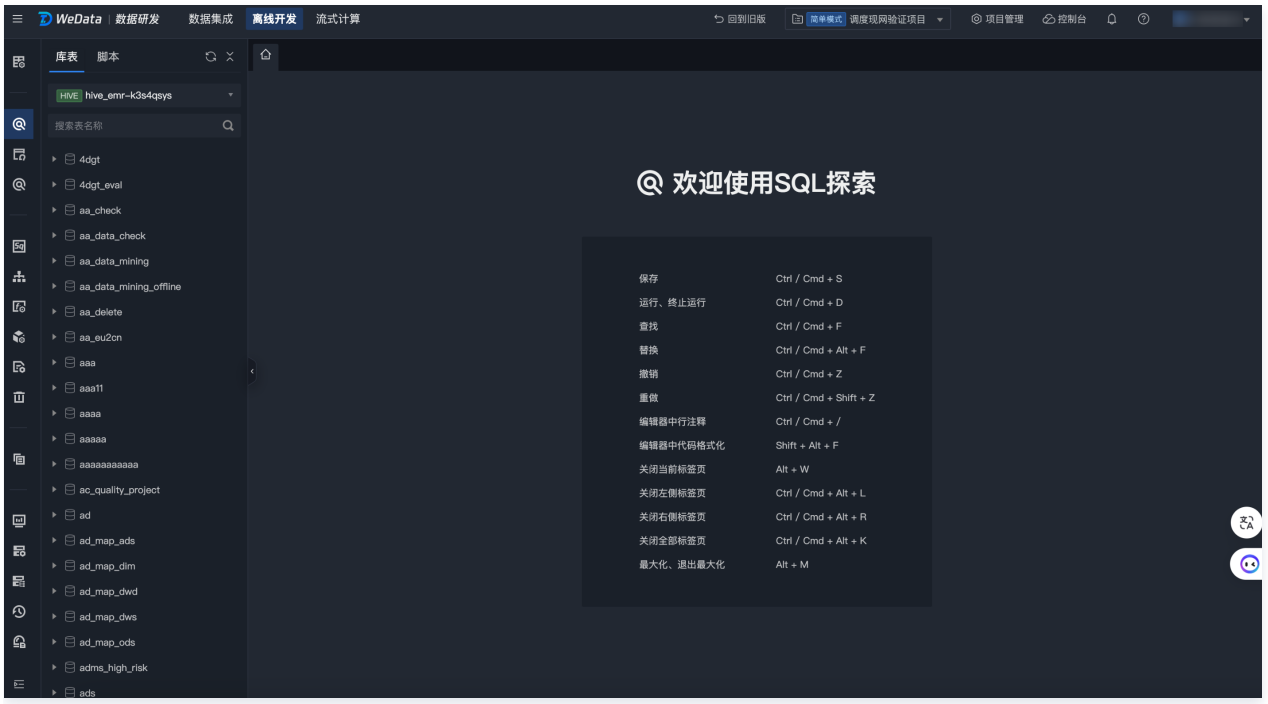
Data Analysis
SQL Explore

Last updated：2024-08-24 19:03:08

Feature Overview

WeData SQL Explore feature is aimed at data developers and business analysts, providing a convenient and fast data query and analysis IDE. It supports connecting to big data engines such as EMR, DLC for data exploration. The currently covered data sources include: EMR-Hive, EMR-Kyuubi, EMR-Trino, DLC, TCHouse-X.

> **Note**
> SQL exploration is currently in public beta and can be experienced for free with the "Data Exploration Experience Resource Group". Charges will be based on usage volume after the public beta.



Operation step

Enter the SQL Explore page

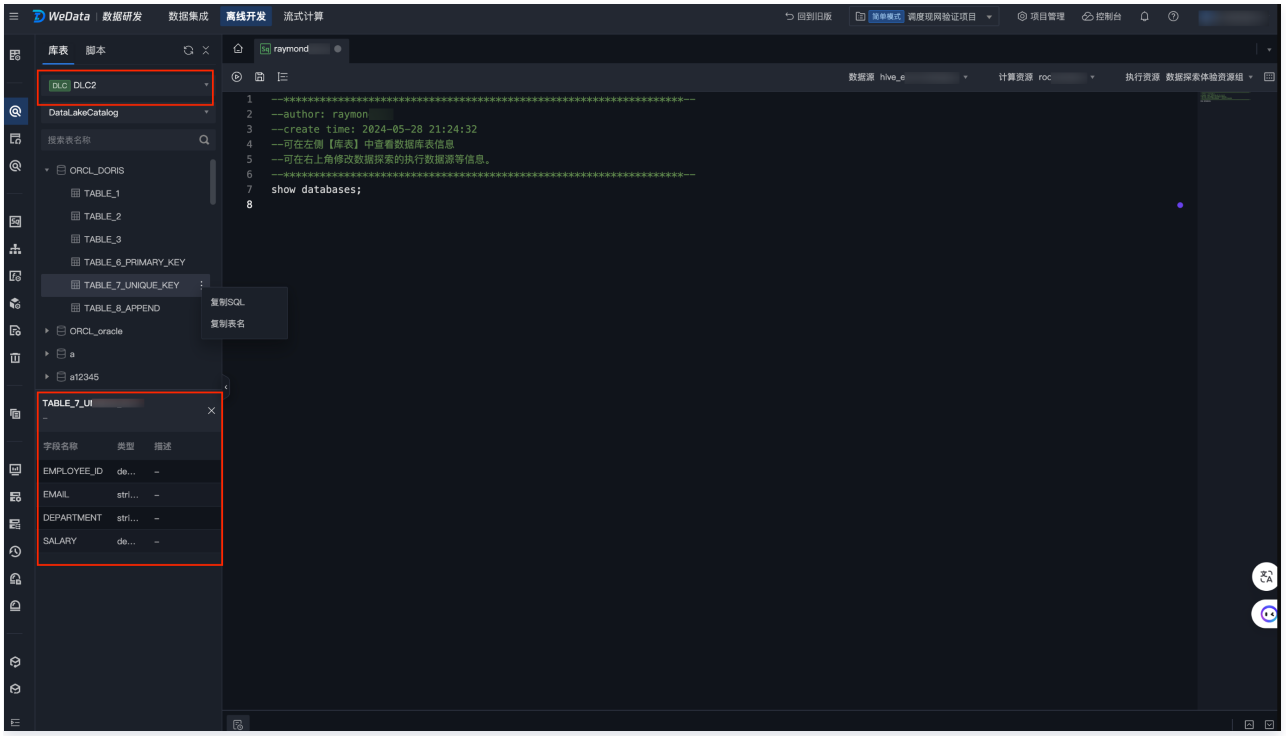
1. Log in to WeData Console.
2. Click **Project List** in the left menu to find the target project for the Notebook Exploration feature.
3. After selecting the project, click to enter the **Data Development** module.
4. Click the **SQL Explore** in the left menu.

Database and Table

Database and Table Query

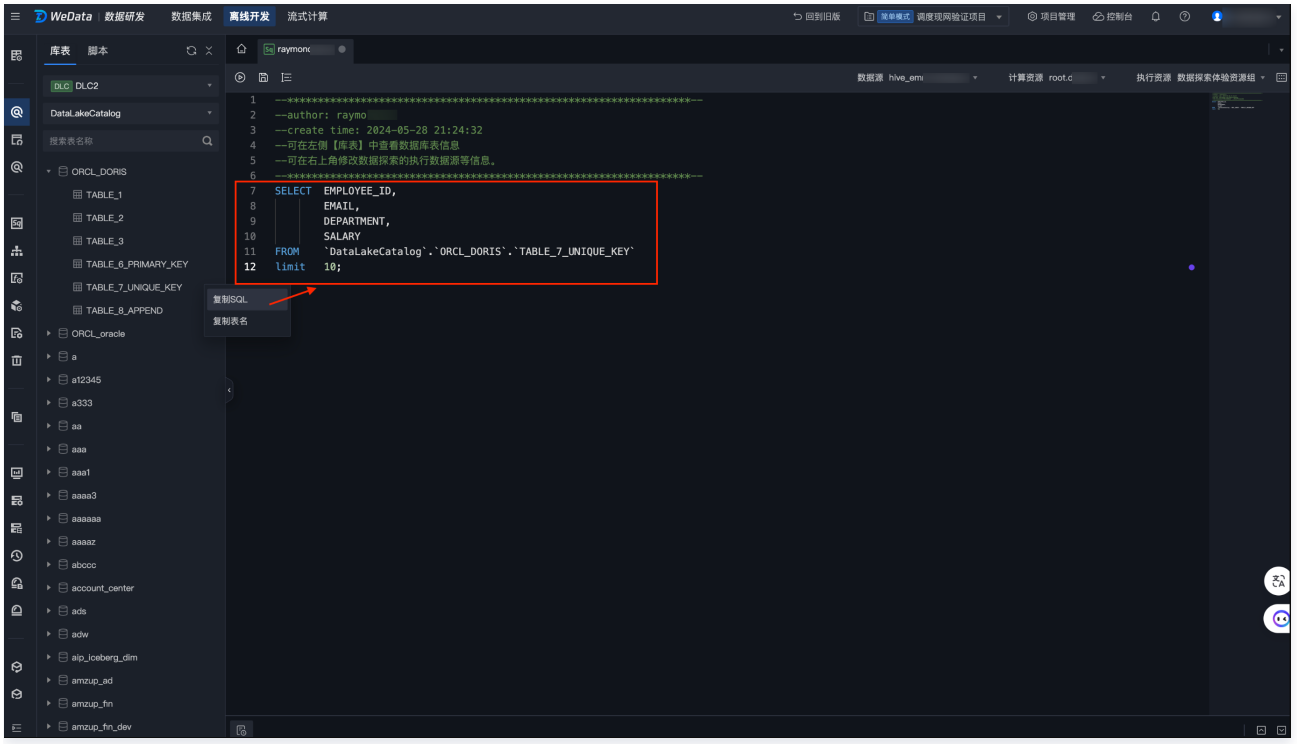
1. Select a data source, and all databases and table information under that data source will be displayed below.
2. Supports searching by table name.
3. Click the **Data Table Name** to quickly query the table's field name, field type, and description information.

©2013-2025 Tencent Cloud. All rights reserved.



Quick Operations

1. In the left directory tree, select a data table and click **More > Copy SQL > Copy Table Name** to copy the query statement and table name of that data table.
2. In the right script editing area, you can quickly paste the copied content.



Scripting

Create New Folder

Method 1: Switch to the "Script" tab, click "+", and select Create New Folder.

©2013–2025 Tencent Cloud. All rights reserved.



Method 2: Click the "**More**" button on the directory tree, and select Create New Folder.



Enter the folder name, select the parent folder, and click **Confirm** to create it successfully.



Create New Query Script

Method 1: Click "+" and select Create New Query Script.



Method 2: Click the "**More**" button on the directory tree, and select Create New Query Script.

©2013–2025 Tencent Cloud. All rights reserved.



Enter the script name, select the script type and parent folder, and click **Confirm** to create it successfully.



Run Query Script

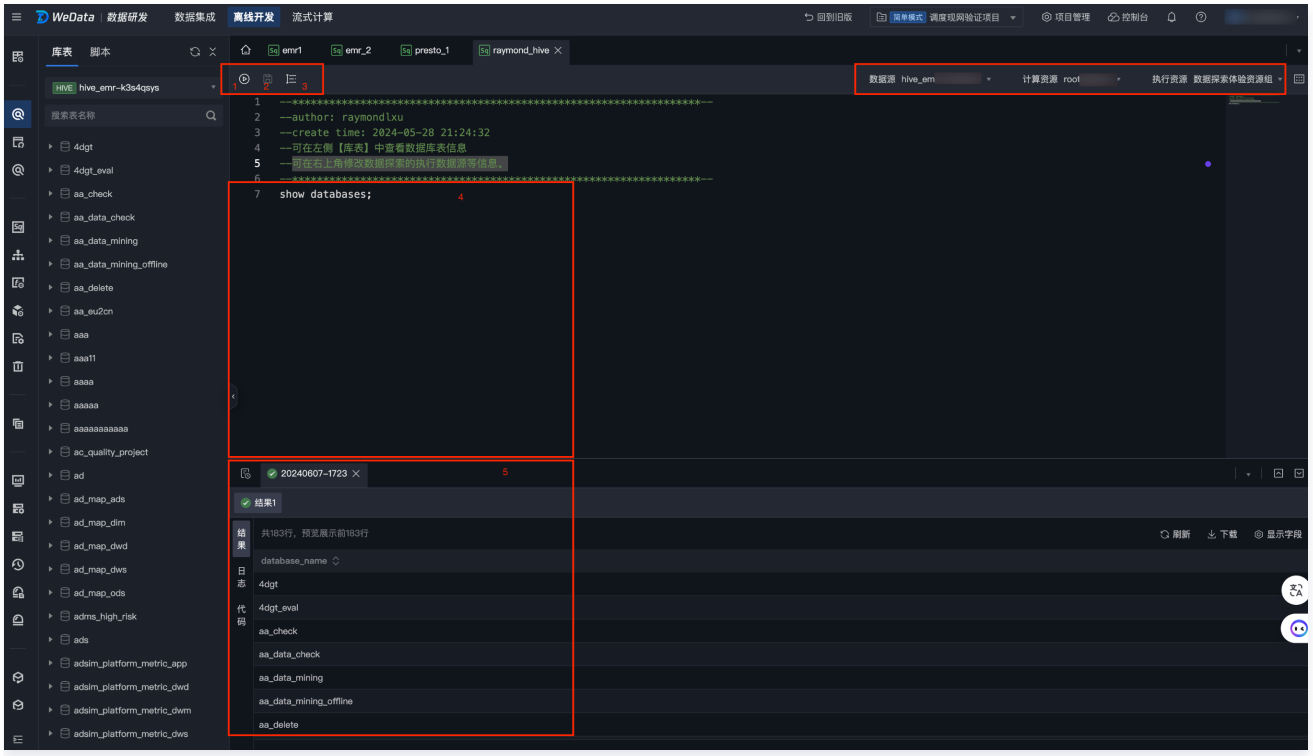
1. Configure data sources, computing resources, and execution resources for the current query script.

Configuration Item	Configuration Items
Data Source	Currently supported data source types include EMR-Hive, EMR-Kyuubi, EMR-Trino, DLC, and TCHouse-X. **Note** EMR-Hive, EMR-Kyuubi, and EMR-Trino data sources currently only support system data sources; self-defined data sources are not supported.
Computing Resources	Computing resources for specific data reading and writing engines: • When EMR-Hive or EMR-Kyuubi data sources are selected, the dropdown option shows queue information of the corresponding EMR cluster • When EMR-Trino data sources are selected, computing resources do not need to be selected • When DLC data sources are selected, the dropdown option shows the computing resources bound to the DLC engine.
Execution Resources	The computing resources used to execute the current query script currently only support the "Data Exploration Experience Resource Group". **Note**

©2013-2025 Tencent Cloud. All rights reserved.

> During the trial phase, users can try for free the "Data Exploration Experience Resource Group" with supported performance parameters of 20 concurrency per root account. After the trial, commercial billing will apply.

2. Enter SQL statements in the code editing area (marked as 4). Currently supported SQL statements include DDL, DML, DQL, etc.
3. Click the "**Formatting**" button (marked as 3) to format the current code segment.



4. Click the "**Run**" button (marked as 1) to execute the current script. Results will be returned in the area below (marked as 4). If there are multiple SQL statements, they will be returned as multiple results.
5. Click the "**Save**" button (marked as 2) to save the current script content for quick access next time.

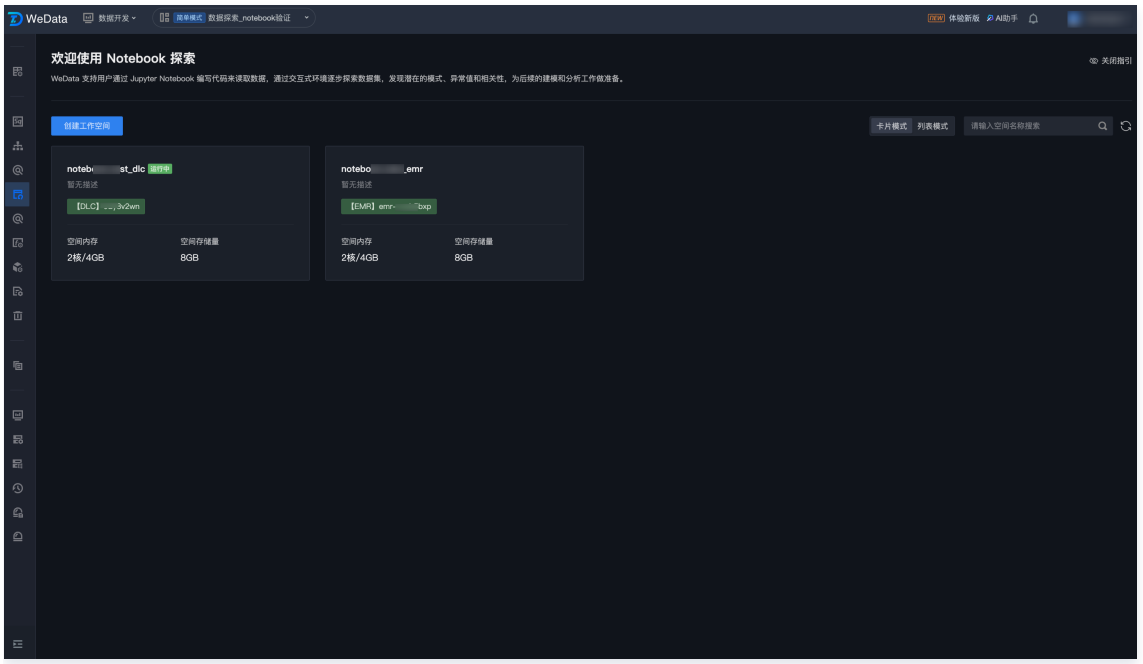
©2013–2025 Tencent Cloud. All rights reserved.

Notebook Exploration

Last updated：2025-01-25 20:06:35

Feature Overview

WeData has newly launched the Notebook Exploration feature, which supports reading data from Tencent Cloud Big Data Engines, EMR and DLC, through Jupyter Notebook. With interactive data analysis, users can explore data and perform Machine Learning. Currently, Notebook exploration is available on the Chinese site in Beijing, Shanghai, Guangzhou, Singapore, and Silicon Valley regions, as well as on the international site in Singapore and Frankfurt regions. You can contact a Ticket to activate the allowlist for trial use.



Features

One-Click workspace creation

No need to manually install Python environments or configure environmental dependencies. One-click creation of Notebook workspaces is supported, including a complete Jupyter Notebook environment and commonly used dependency packages.

User and resource isolation

Each user enjoys an exclusive workspace under different projects. The storage and computing resources of each workspace are isolated from each other. Users' tasks and file resources do not interfere with each other.

Integrated with big data engine foundation

Supports binding with EMR and DLC Big Data engines, allowing direct reading of data from big data storage and computing engines for interactive exploration, algorithm model training, and predictive data analysis.

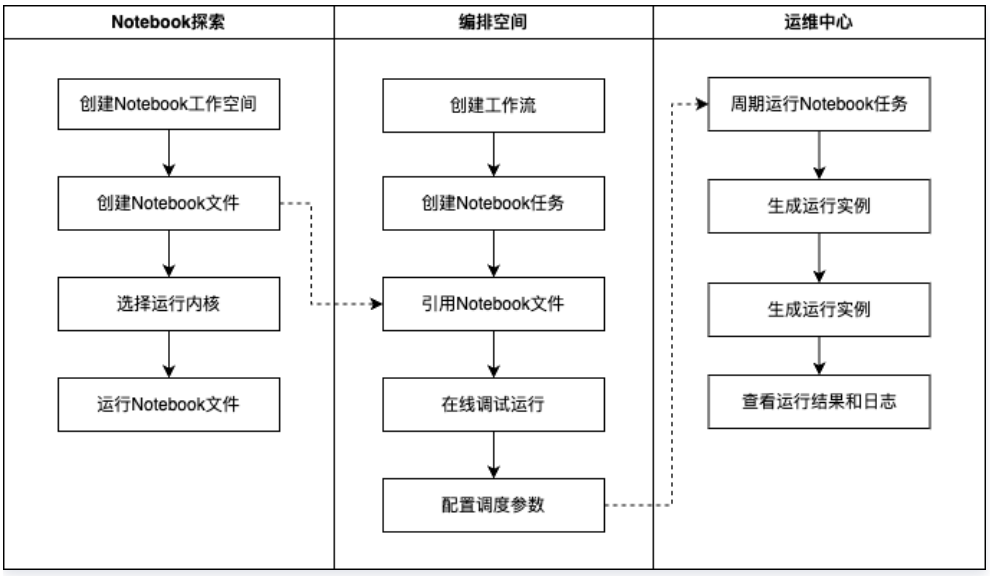
Built-In practical tutorial

The Notebook workspace comes with built-in Big Data tutorials, allowing users to get started quickly and easily.

Overall Usage Process

The full process for users to use Notebook in WeData is shown below:

©2013-2025 Tencent Cloud. All rights reserved.



Operation Steps

Create a Notebook Workspace

1. Log in to WeData Console.
2. Click **Project List** in the left menu to find the target project for the Notebook Exploration feature.
3. After selecting the project, click to enter the **Data Analysis** > **Notebook Exploration** module.

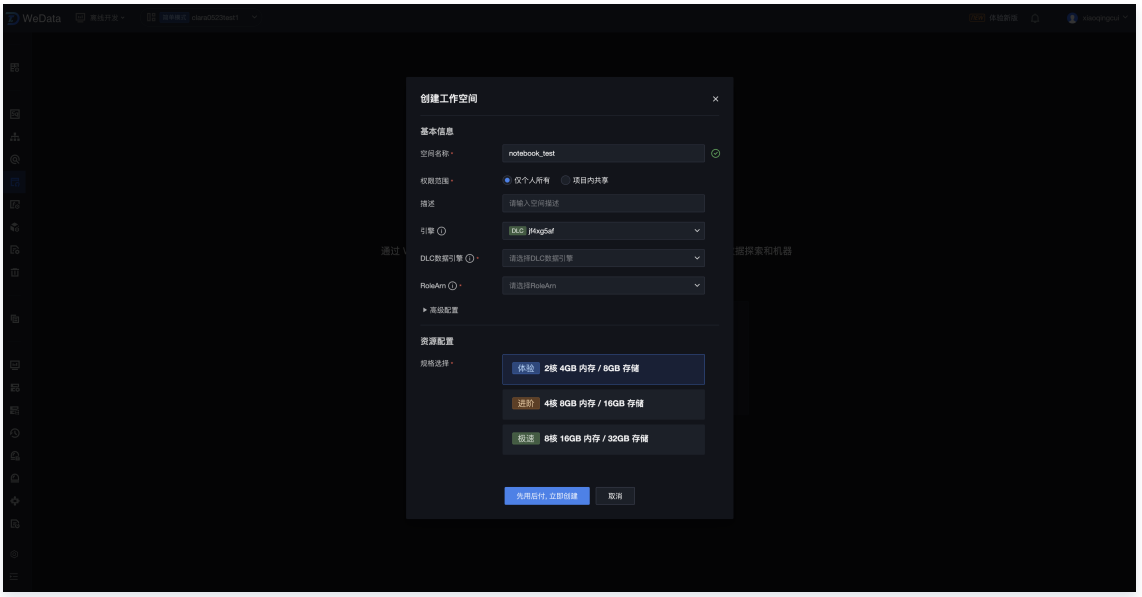


4. Enter the Notebook Exploration list page and click **Create workspace.**



5. Enter the workspace configuration page to set basic information and resource configuration information.

©2013–2025 Tencent Cloud. All rights reserved.



Attribute Name	Attribute Description	Required
Basic information	Configure the basic information of the Notebook workspace to create a Notebook workspace instance	
Workspace Name	Notebook workspace name can include Chinese characters, English letters, digits, underscores, and hyphens, and must not exceed 32 characters	Yes
Permission Scope	If "Individual Use Only" is selected, only the current user can access the workspace; if "Project Share" is selected, all members of the project can access the space for collaborative development.	Yes
Description	Notebook workspace description can include Chinese characters, English letters, digits, special characters, etc., and must not exceed 255 characters	No
Engines	Supports selecting the EMR or DLC storage engine bound to the current project. After selection, it will be pre-connected with the storage engine, and PySpark can be used to access it in Notebook tasks	No
Network	When selecting the EMR engine, further network configuration is required to establish network connectivity. By default, the VPC and subnet where the EMR engine is located are used	Yes
DLC Data Engine	When selecting the DLC engine, you need to further choose a DLC data engine bound to the project to execute DLC PySpark tasks. **Note** Only computing resources of the DLC Spark job type are supported.	Yes
Machine Learning	• If the DLC data engine you select contains a "machine learning" type resource group, this option will appear and be selected by default. • If the DLC data engine you select does not contain a "machine learning" type resource group, this option will not appear. To use it, please go to DLC to create one.	No
RoleArn	When selecting the DLC engine, further RoleArn needs to be selected to authorize access to data storage COS **Note**	Yes

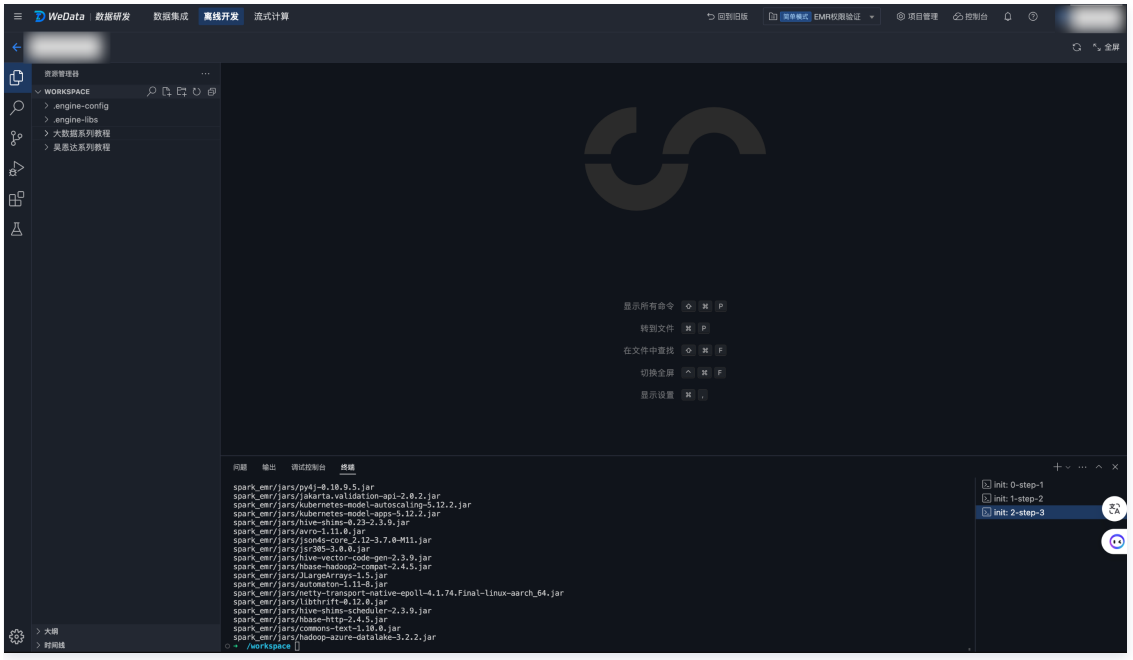
©2013–2025 Tencent Cloud. All rights reserved.

	RoleArn is the data access policy (CAM role arn) for the DLC engine to access Cloud Object Storage (COS), which needs to be configured by the user in DLC.	
Advanced configuration	You can choose to use Mlflow to manage experiments, data, and models in Notebook exploration. This feature currently requires allowlist access.	
MLflow service	After checking, the creation of experiments and machine learning using MLflow functions in Notebook tasks will be reported to the MLflow service. You can later view them in Machine Learning – Experiment Management and Model Management.	No
Resource Configuration	Configure the storage and computing resources of the workspace for executing Notebook task operations, CFS	
Specification Selection	Supported specifications include: • 2 CPU cores, 4 GB memory, and 8 GB storage (Experience Edition) • 4 CPU cores, 8 GB memory, and 16 GB storage (Advanced Edition) • 8 CPU cores, 16 GB memory, and 32 GB storage (Ultimate Edition)	Yes

Start/Stop Workspace Management

Launch workspace

1. click **Create Now** to enter the Notebook workspace launch page.
2. During the startup process, the PySpark environment will be configured, and common Python packages like numpy, pandas, scikit-learn, etc., will be installed. The installation process will take some time, so please wait patiently until it completes.
3. If the Notebook workspace enters the following page, it indicates that the startup is successful, and you can start creating Notebook tasks.

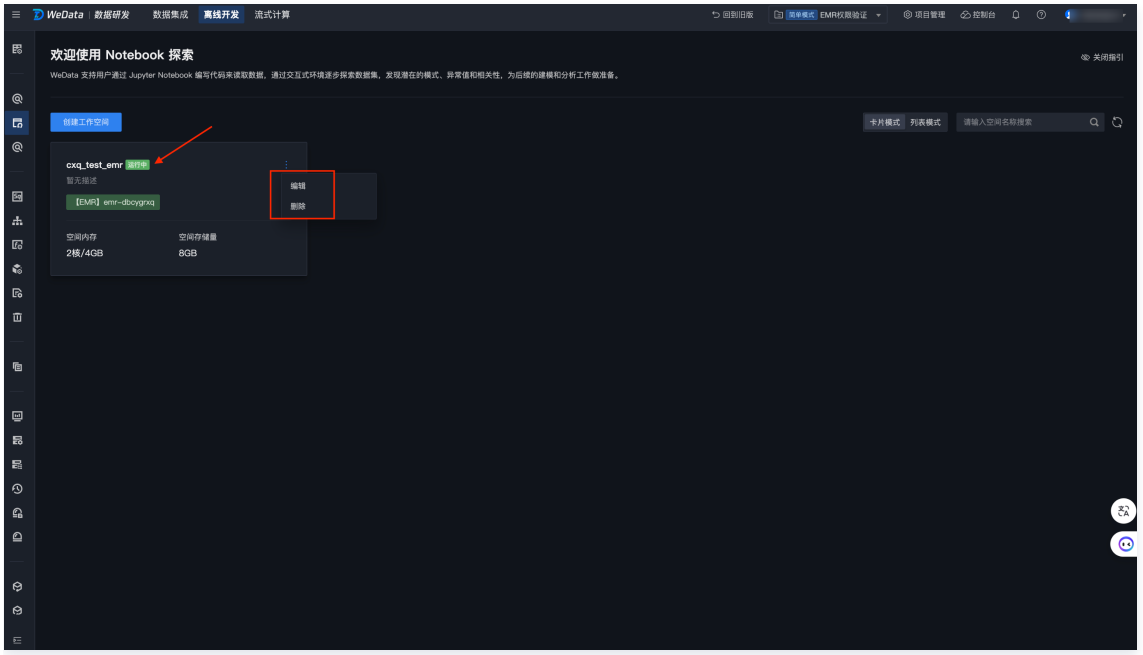


Log out of the workspace

1. Click the **log out** button at the top left to exit the current workspace and return to the list page.

The workspace will automatically stop ten minutes after exiting. Restarting a stopped workspace will restore the development environment and data.

©2013–2025 Tencent Cloud. All rights reserved.



Editing workspace

click the **Edit** button on the list page to modify the configuration information of the current workspace. Configurable items include: Space Name, Description, and Resource Configuration.



Deleting workspace

click the **Delete** button on the list page to delete the current workspace.

Create and Run a Notebook File

©2013–2025 Tencent Cloud. All rights reserved.

1. Create a Notebook File

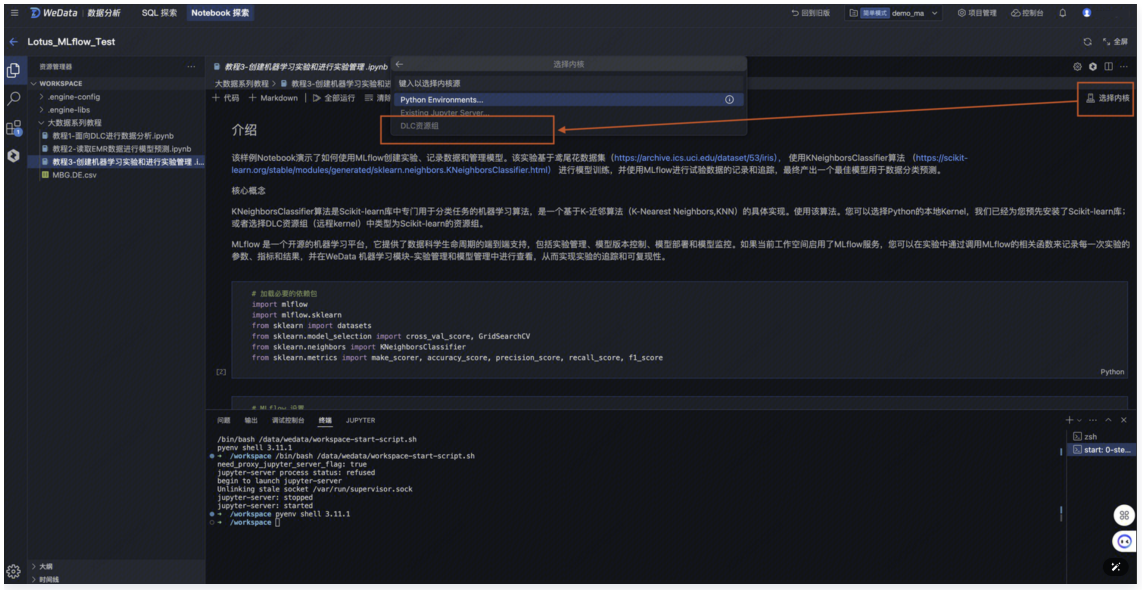
You can create folders and Notebook files in the left resource manager.

> **Note:**
> Notebook files need to end with (.ipynb).

2. Select a running kernel

Enter the Notebook file, click the top left **select kernel**, and choose a **kernel** from the dropdown options.

> **Note**
> In Jupyter Notebook, the kernel is the backend application that executes code, handles the execution of code cells, returns calculation results, and interacts with the user interface.
> WeData Notebook currently supports two types of kernels:
> - Python Environment: The default IPython kernel in Jupyter Notebook, supports Python code execution.
> - DLC resource group: A remote kernel provided by Tencent Cloud Big Data, allowing Python tasks to be submitted to the DLC resource group for execution.



If you select the DLC resource group, choose a machine learning resource group instance from the DLC data engine in the next level of options.



3. Run a Notebook File

Click **Run** to generate a Notebook kernel instance and start running the code. The results will be displayed below each cell.

Periodic Scheduling Of Notebook Tasks

Create a notebook task

1. Enter the project and open the menu **Data Research and Development** > **Offline Development**.

©2013–2025 Tencent Cloud. All rights reserved.



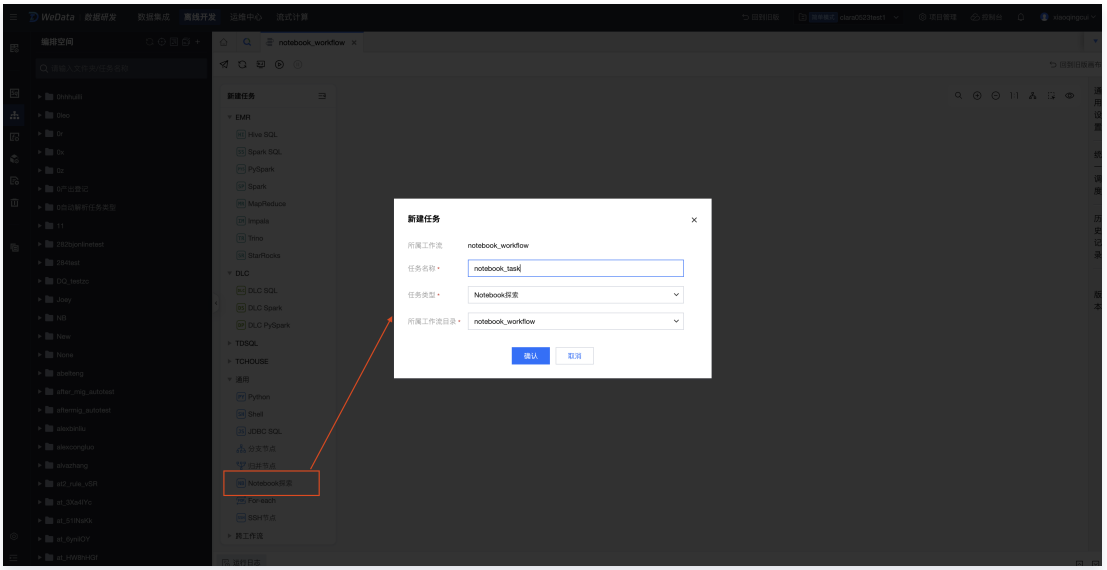
2. In the left directory, click **Create Workflow** and configure the workflow properties, including the workflow name, folder, etc.





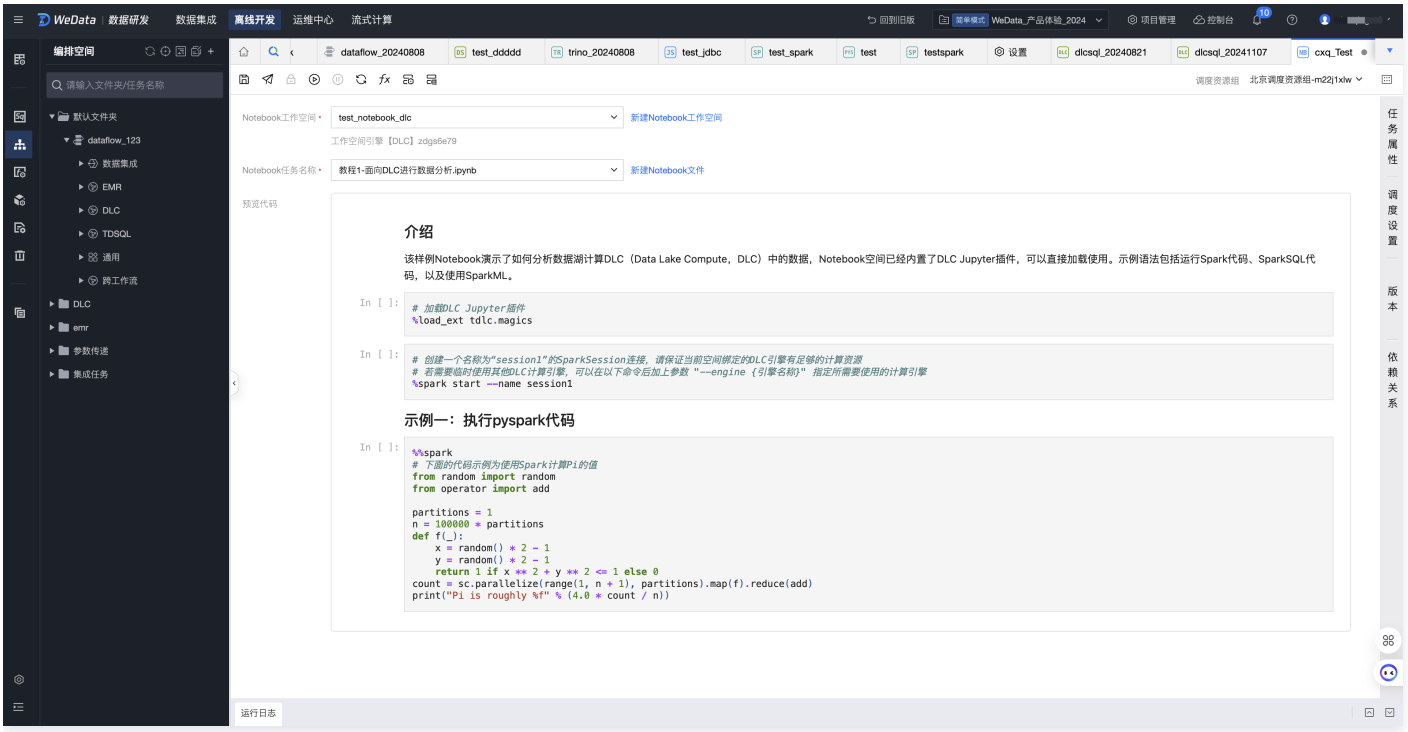
3. Create a task in the workflow, with the task type as **General–Notebook Exploration**. Configure the task's basic attributes on the Create New Task page, including the task name, task type, etc.

©2013–2025 Tencent Cloud. All rights reserved.



Configure and run a notebook task

On the Notebook task configuration page, refer to a file in a Notebook workspace.



1. Select a Notebook workspace

You can drop down to select all Notebook workspaces in the current project.

2. Select a Notebook file

You can drop down to select all files in the current Notebook workspace. Note: If the current user does not have permission for the Notebook workspace, they cannot access it for operations.

3. Preview code

After selecting a Notebook file, you can preview the specific content of the Notebook file below.

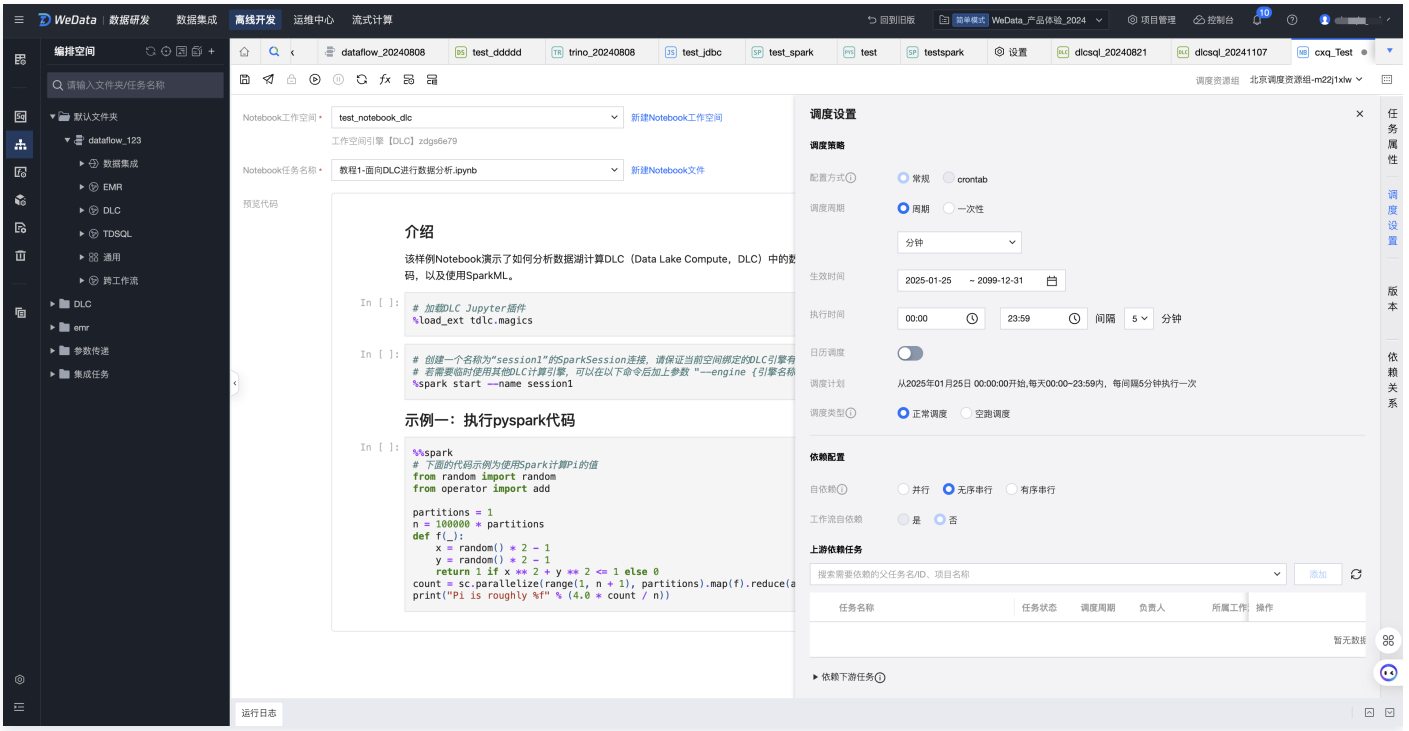
4. Run a notebook task

In the upper right corner, select the scheduling resource group and click **Run** to run the current Notebook file online. You can view the running logs, running code, and execution results below.

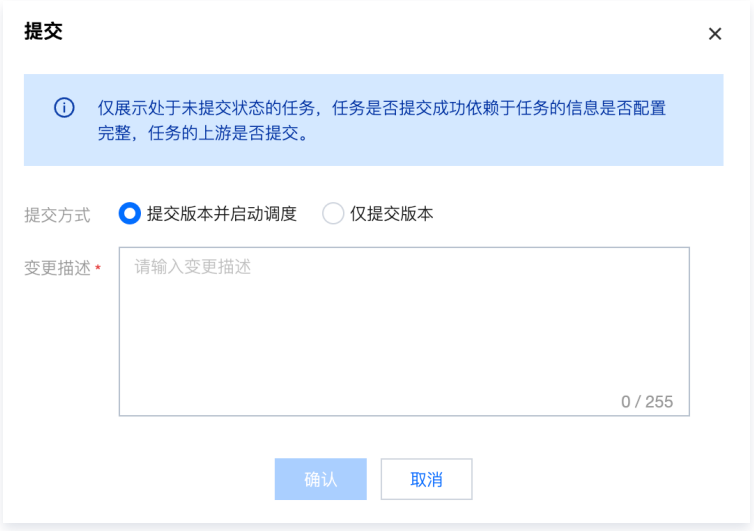
Configuring scheduling

©2013–2025 Tencent Cloud. All rights reserved.

1. Click on the right side **Scheduling Configuration** to set the scheduling cycle for the current Notebook task. For example, the figure below sets it to run every 5 minutes.

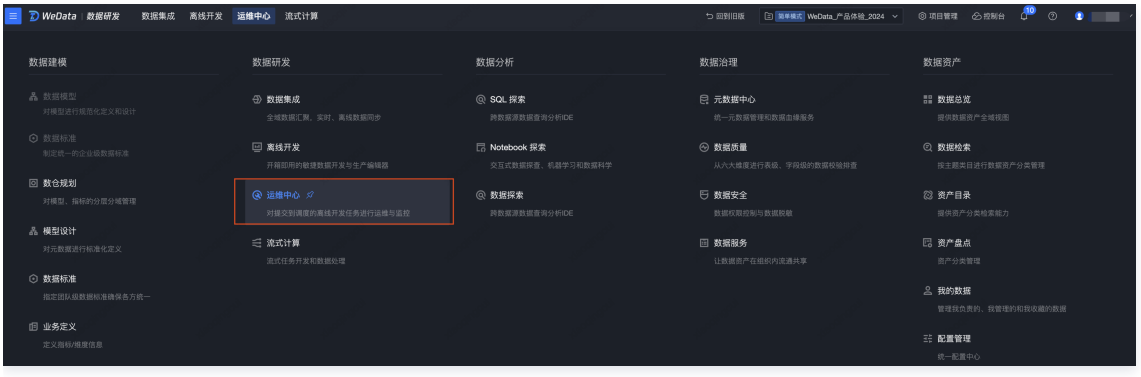


2. Click the **Submit** button to submit the current task to periodic scheduling.



Task ops

1. Go to **Data Research and Development** > **Ops Center**.



©2013–2025 Tencent Cloud. All rights reserved.

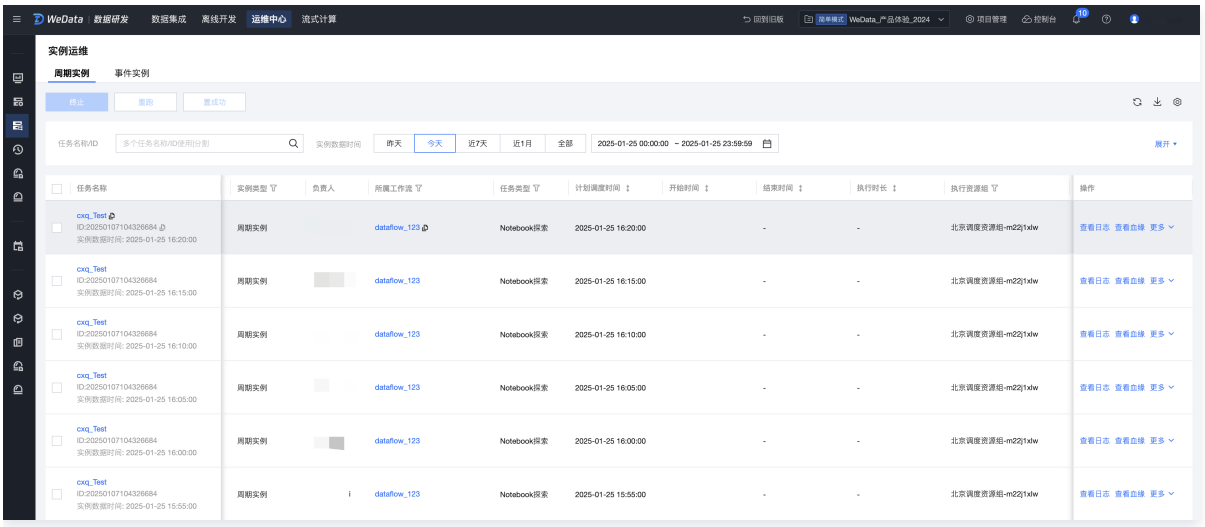
2. Task Ops

Click **Task Ops** to see the workflows submitted to the scheduler and the task nodes within the workflows.



3. OPS of Instances

Click **Instance Ops** to view each period instance generated by the workflow.



4. Enter the instance detail to view the running logs and results.

Practice Tutorial

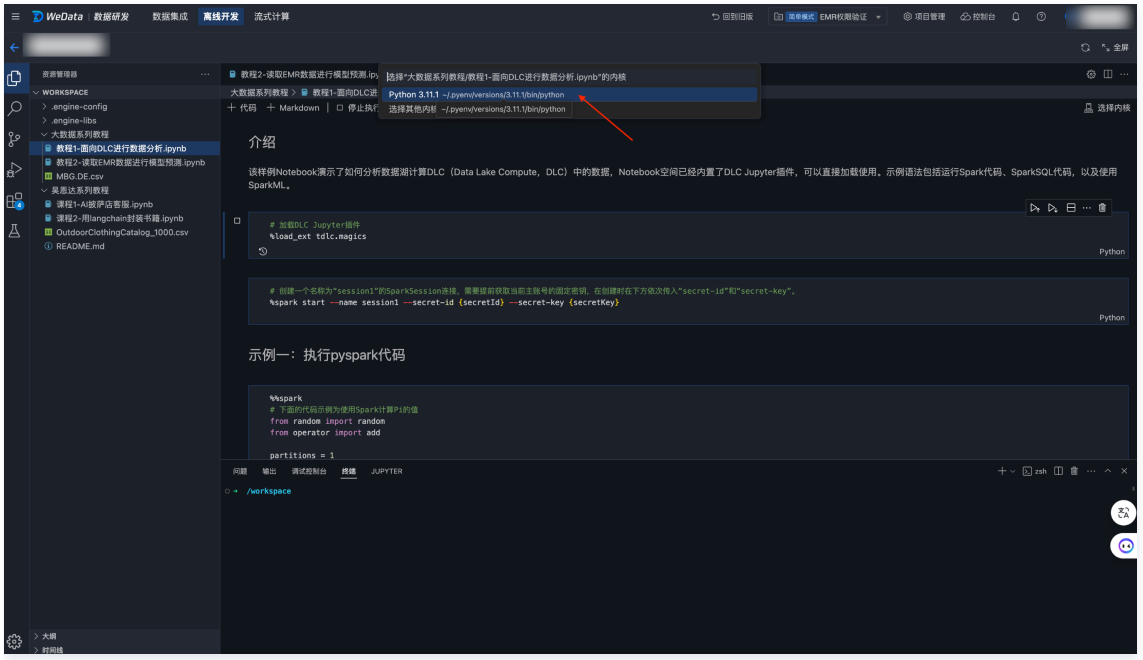
The Notebook workspace comes with built-in Big Data tutorials, allowing users to get started quickly and easily.

Tutorial 1: Data Analysis Using DLC Jupyter Plug-In

This sample Notebook demonstrates how to analyze data in Data Lake Compute (DLC). The Notebook workspace has the DLC Jupyter Plugin built-in, which can be loaded directly. The example syntax includes running Spark code, SparkSQL code, and using SparkML.

> **Note:**
> Using this tutorial, the Notebook workspace needs to bind the DLC engine and uncheck "Use Machine Learning Resource Group". The kernel should be set to Python Environment, and WeData Notebook will interact with DLC using the Jupyter Plugin.

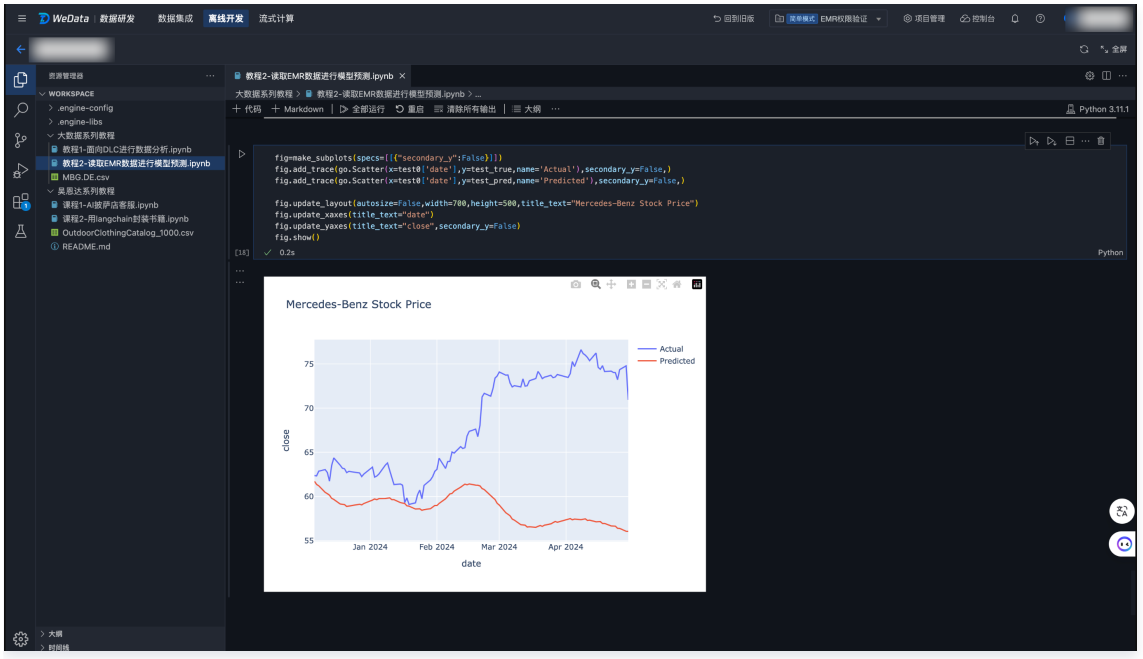
©2013-2025 Tencent Cloud. All rights reserved.



Tutorial 2 Reading EMR Data For Model Prediction

1. This sample Notebook demonstrates how to create an EMR-Hive Table and import local data into the EMR-Hive Table. Then, data is read from the EMR-Hive Table and converted into a pandas DataFrame for data preparation.
2. After completing data preparation, you can use the Prophet Time Series Algorithm to train a prediction model and finally evaluate the model accuracy and forecasts.

> **Note:**
> Using this tutorial, the Notebook workspace needs to bind the EMR engine.



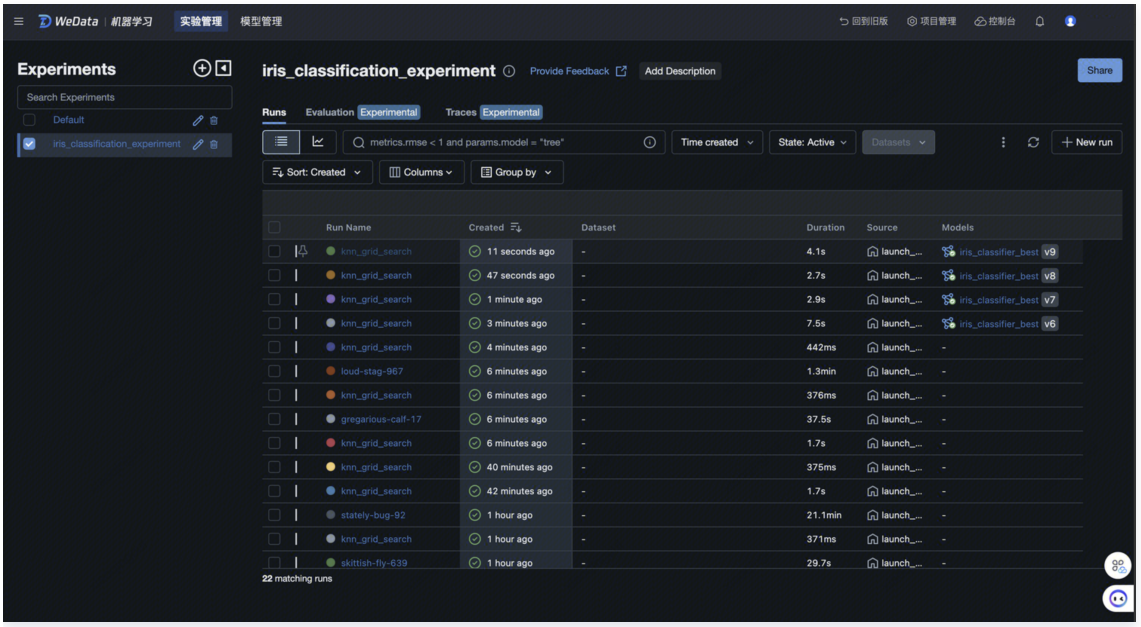
Tutorial 3 Creating Machine Learning Experiments and Managing Experiments

This sample Notebook demonstrates how to use MLflow to create experiments, record data, and manage models. The experiment is based on the Iris dataset, using the KNeighborsClassifier algorithm for model training, and uses MLflow to record and trace experimental data, ultimately producing an optimal model for data classification and prediction.

©2013-2025 Tencent Cloud. All rights reserved.

> **Note:**
> Using this tutorial, the Notebook workspace needs to bind the DLC engine and check "Use Machine Learning Resource Group". The kernel should be set to DLC Resource Group, and WeData Notebook will remotely submit the Notebook file to DLC for execution.

MLflow is an open-source machine learning platform that provides end-to-end support for the data science lifecycle, including experiment management, model versioning, model deployment, and model monitoring. If the current workspace has MLflow service enabled, you can record each experiment's parameters, indicators, and results by calling MLflow's related functions in the experiment, and view them in WeData Machine Learning Module > **Experiment Management** and **Model Management**, thus achieving experiment traceability and reproducibility.



©2013-2025 Tencent Cloud. All rights reserved.

Data Assets
Data Assets Overview

Last updated: 2024-08-22 17:32:42

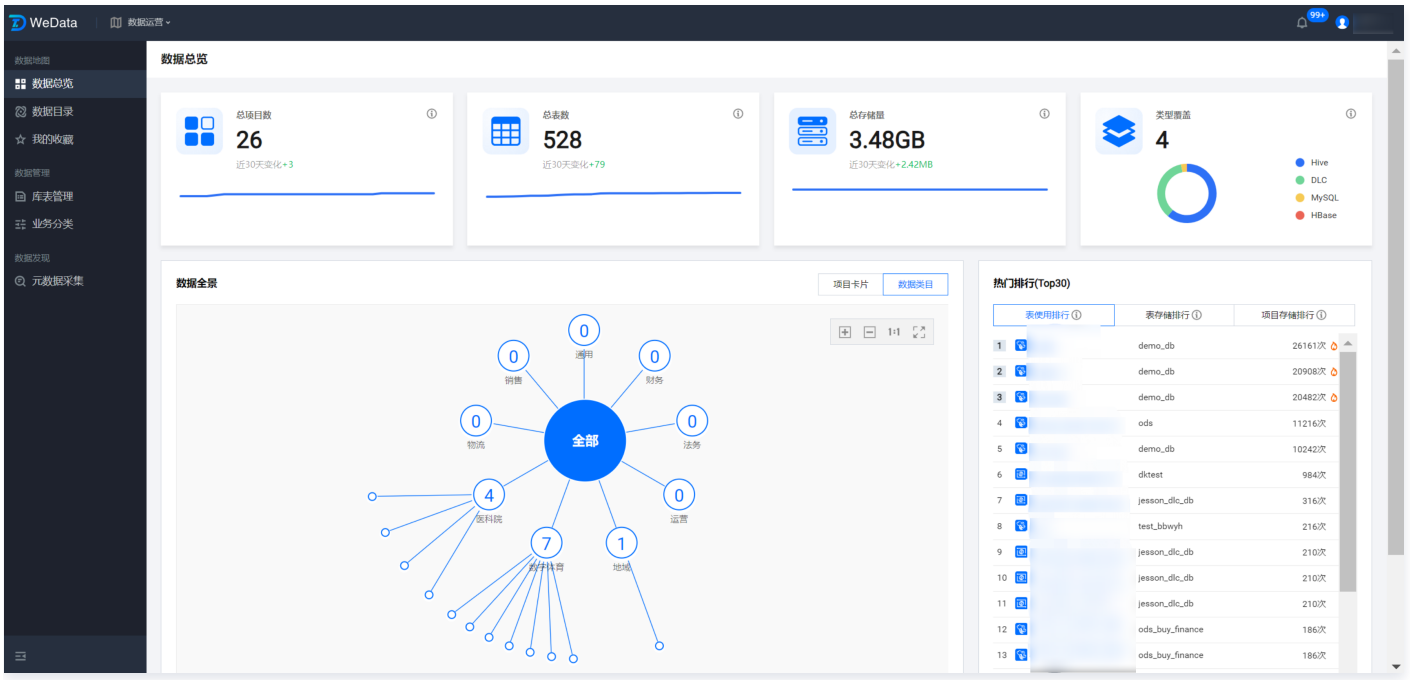
Enable data assets to be better shared and discovered within the organization (a main account, which can be a company or department), thereby enhancing data usage, reducing data repetition rate, aiding in data value mining, and achieving cost savings. The data assets module, based on metadata, provides a global data asset management feature, including data retrieval, asset inventory, asset directory, asset details view, business metadata management, lineage tracking, and temperature statistics.

©2013-2025 Tencent Cloud. All rights reserved.

Data map
Data Overview

Last updated: 2024-11-25 14:38:04

The Data Overview feature provides a global view of data assets, including total data quantity, storage size, total number of projects, access and storage ranking, helping you quickly understand data distribution and usage.



Parameter	Description
Total number of projects/Trend	Displays the total number of projects under the root account and the quantity change trend over the last 30 days.
Total number of tables/Trend	Displays the total number of data tables in all projects under the root account and the quantity change trend over the last 30 days.
Total storage volume/Trend	Displays the total storage volume of all data tables in all projects under the root account and the quantity change trend over the last 30 days.
Type coverage/Proportion	Displays the number of types of data tables covered under the root account, showing the proportion of Hive tables, MySQL tables, Kafka tables, HBase tables, and other types of tables.
Data Panorama	Centralized visualization of data information under the root account. • Project cards: Displays the number of databases, tables, and members for each project. • Data categories: Displays the category structure of data assets in the form of business themes.
Hot Ranking (top 30)	You can click a specific project name to navigate to the project details page. • Table Usage Ranking: Displays the real-time ranking of the usage of all tables under the main account. • Table Storage Ranking: Displays the offline ranking of the total storage of all tables under the main account. • Project Storage Ranking: Displays the offline ranking of the storage of all projects under the main account.

©2013-2025 Tencent Cloud. All rights reserved.

Data Extraction Search

Last updated：2024-11-25 14:38:33

Basic search

The data retrieval feature helps you quickly search all the asset information under the main account using keywords and other conditions. Click **Search** or **Enter key** to enter advanced search.



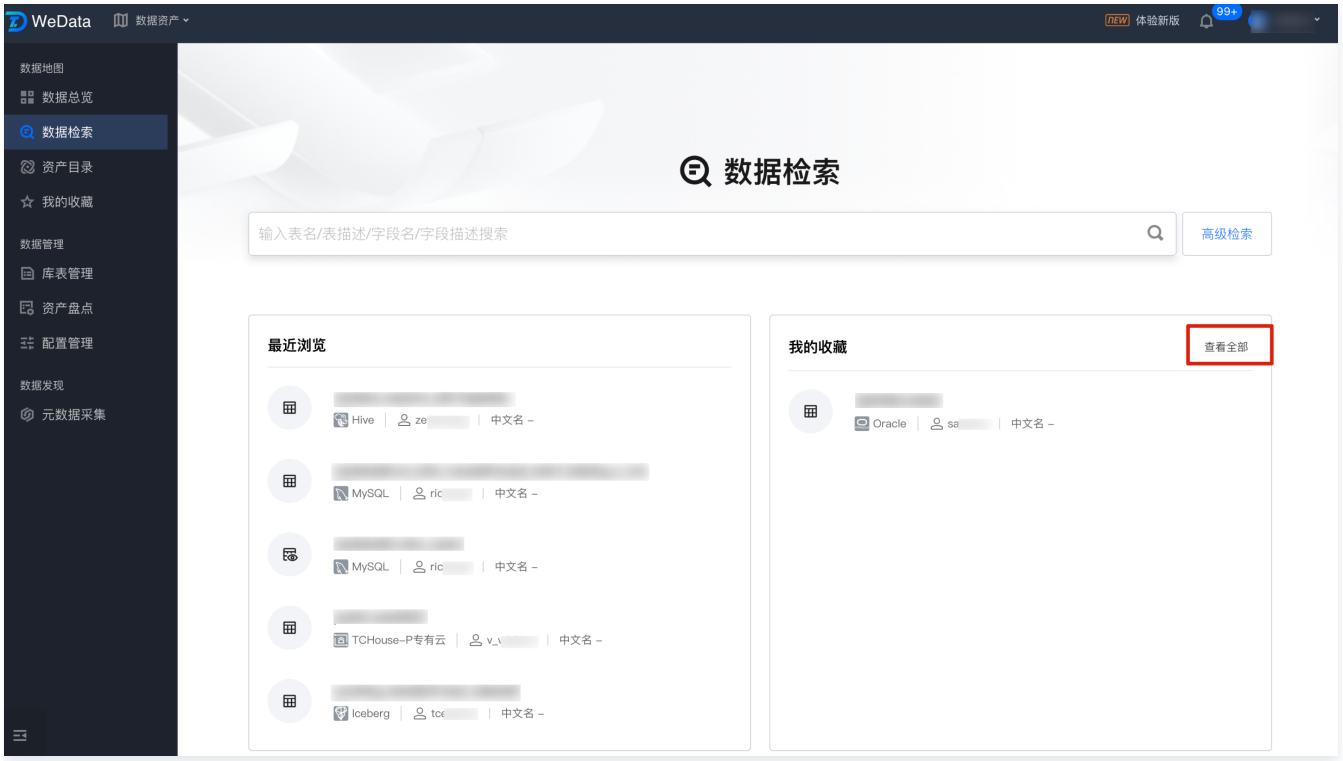
Recently Viewed

Recently viewed includes the tables you have browsed recently, displayed in reverse chronological order of the last 5 records. The format displayed is **Database Name** and **Table name**.

My Favorites

My favorites will display the tables that you have recently favorited, in reverse chronological order by the time of favoriting. Click **View All** to jump to the **My Favorites** module to view all the tables you have favorited.

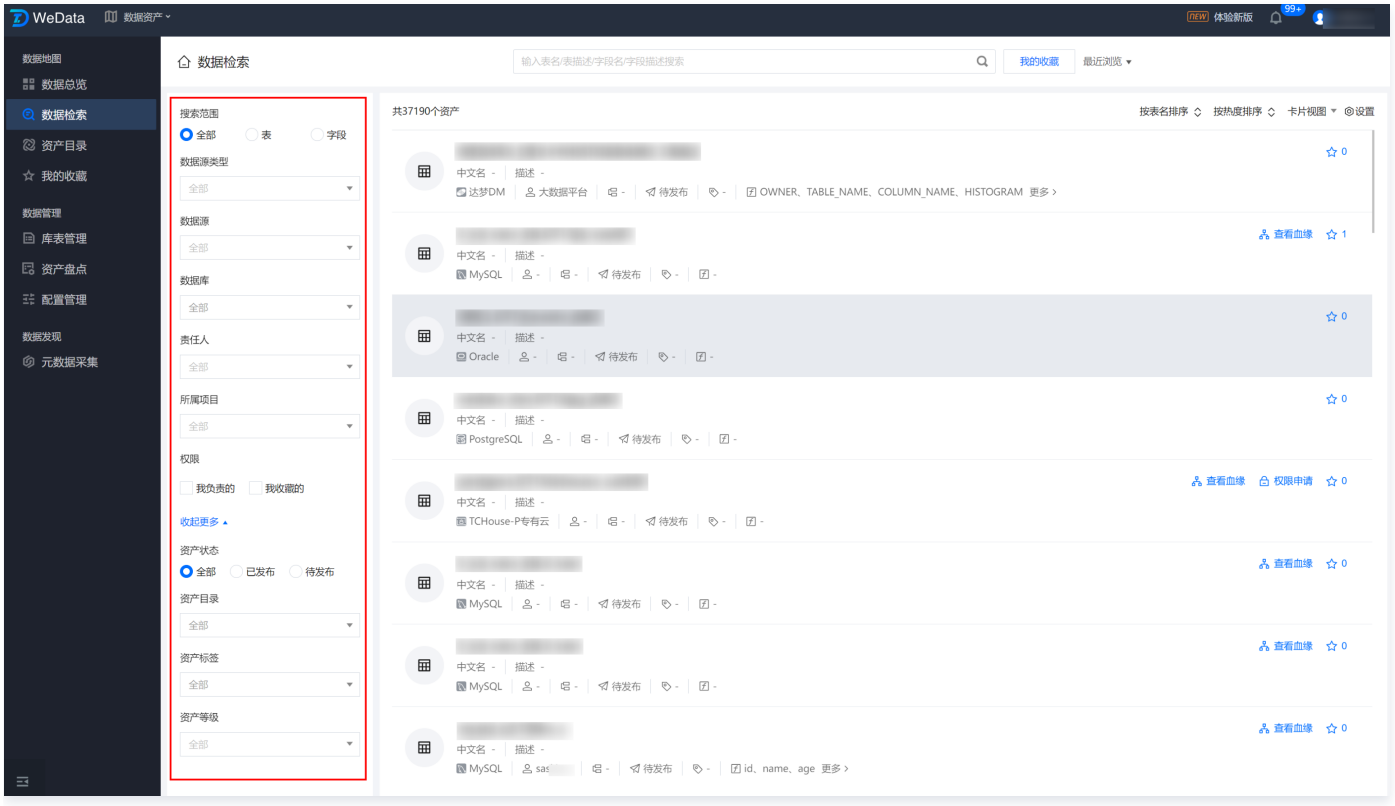
©2013-2025 Tencent Cloud. All rights reserved.



Advanced search

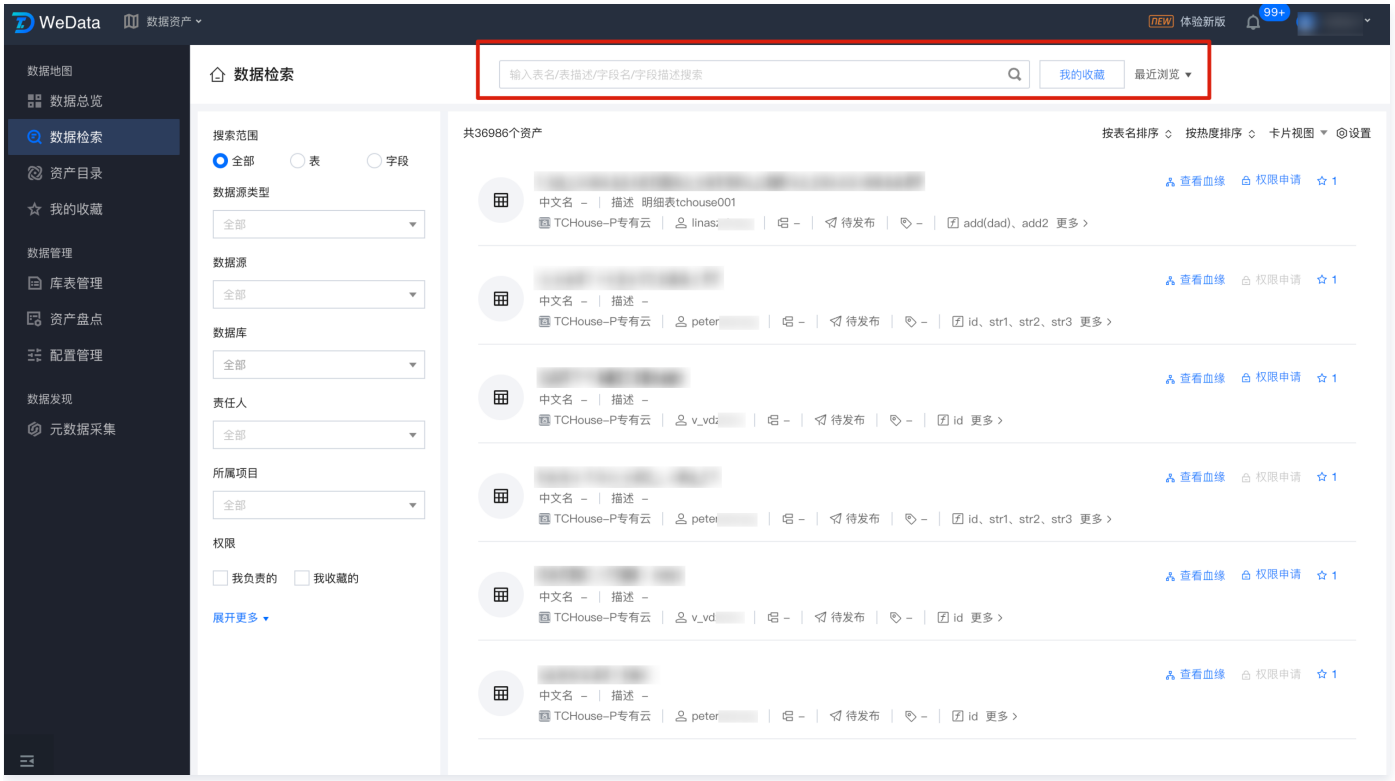
Search Scope

- The left side of the data retrieval page is the search area where you can refine the search results based on data source type, data source, database, and other conditions.
- In **Show More**, you can also search by asset status, directory, Tag, level, and other business attributes to quickly find the required table.



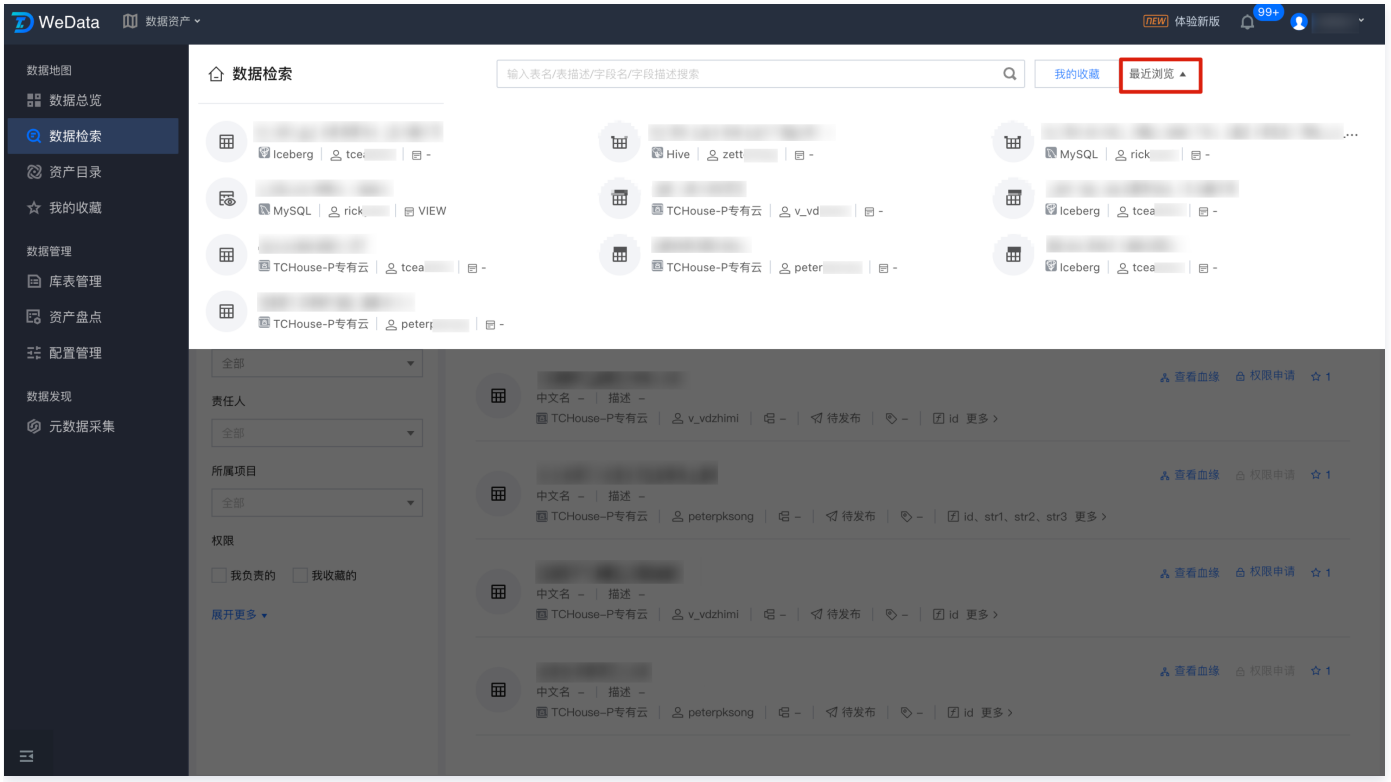
Search box

The top of the data retrieval page is the search box where you can reset keyword searches according to your needs. Next to the search box, you can click **My Favorites** to enter the **My Favorites** module.



Recently Viewed

Recently viewed shows the last 10 browsing records for your convenience.



Search Results

©2013–2025 Tencent Cloud. All rights reserved.

Search results will sort related data assets by matching degree, search terms, and business attributes. The data result list includes information such as data source type, owner, asset catalog, Tag, and fields, providing features like bloodline viewing, permission application, favorite/unfavorite. You can also sort by table name or popularity.

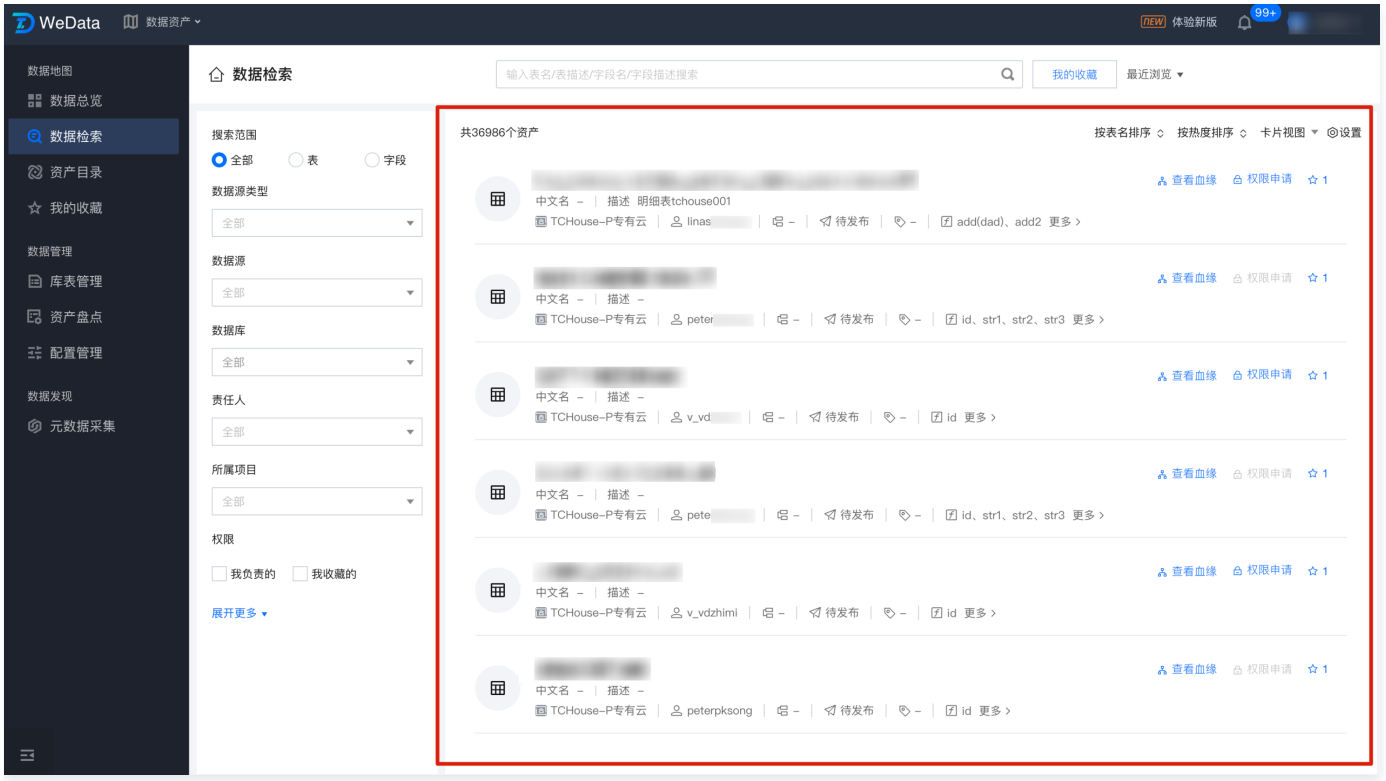
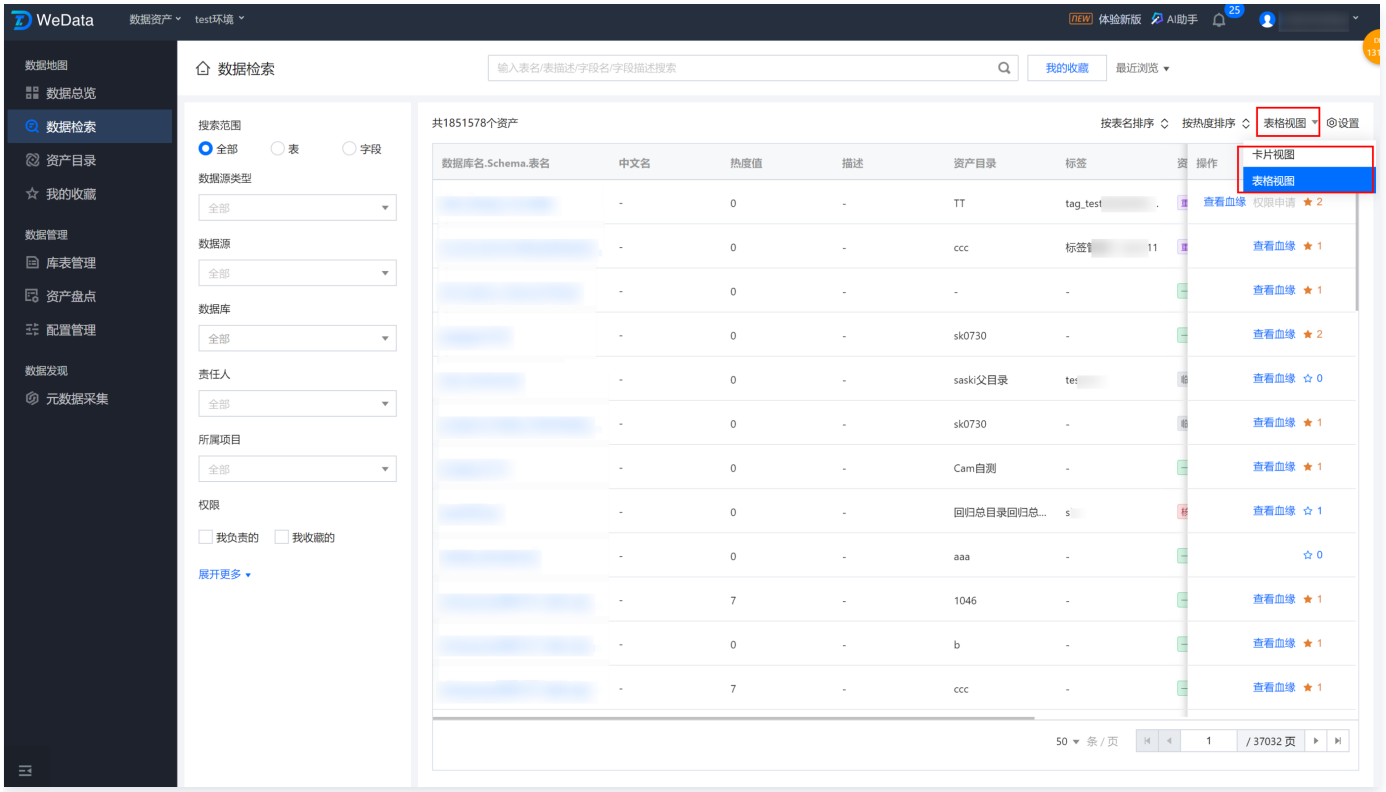


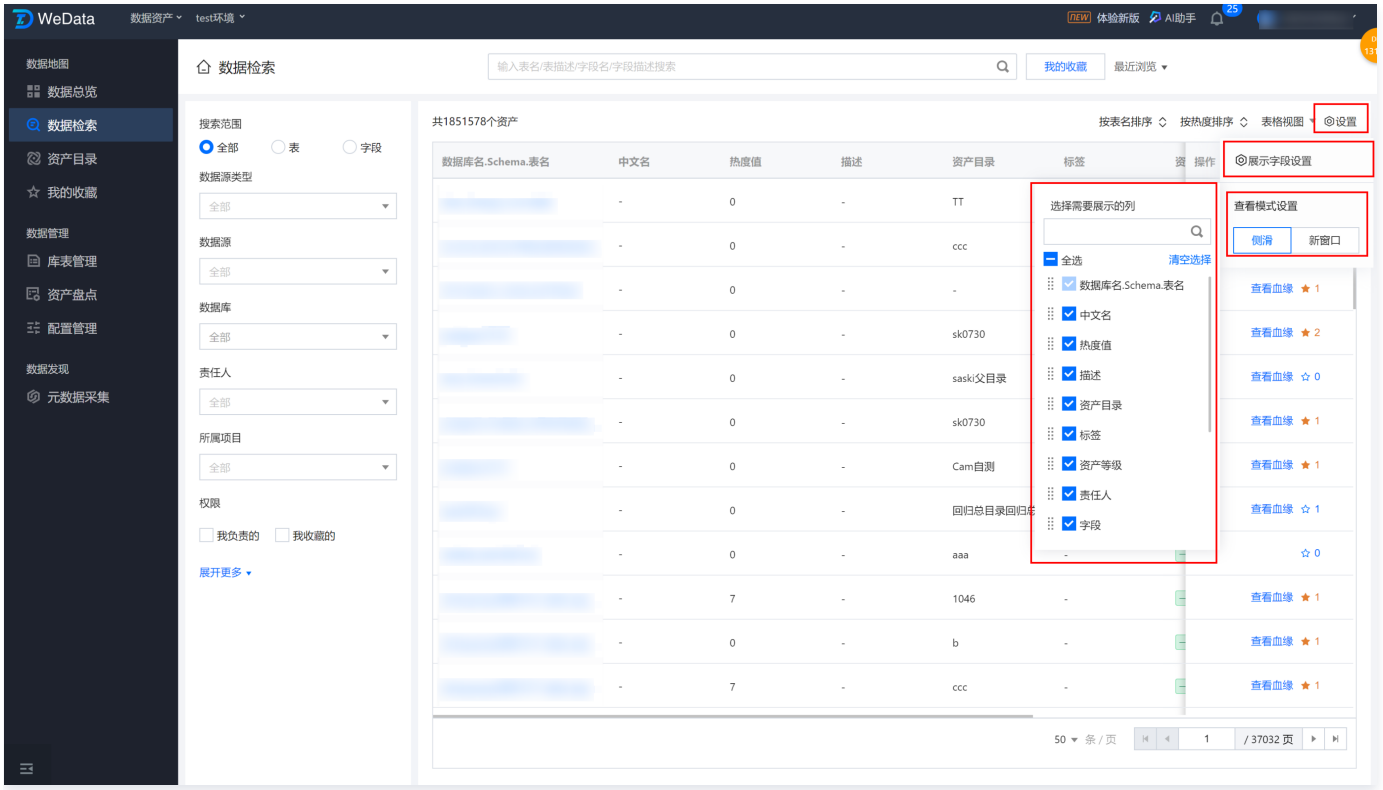
Table View

- In Card View, you can also browse search results in Table format (default is **Card View**). Adjust the displayed fields in Settings to facilitate Search and Inspect. See the Table View below.



©2013–2025 Tencent Cloud. All rights reserved.

- Additionally, you can adjust your preferences for Swipe or New Window Navigation in the **Settings** button. We also support field adjustment in Table View.



©2013–2025 Tencent Cloud. All rights reserved.

Table Details

Last updated：2024-11-25 14:39:01

> **Note**
> Due to differing levels of support for various data source types, not all types include the following features. Please refer to the actual display results on the page.

In the respective pages of various features in data management, click the table name you need to view to enter the table details page. Depending on the data source, the table details page includes business information, technical information, asset rating, basic information, data preview, output and changes, data lineage, data temperature, partition information, data quality, and access logs.

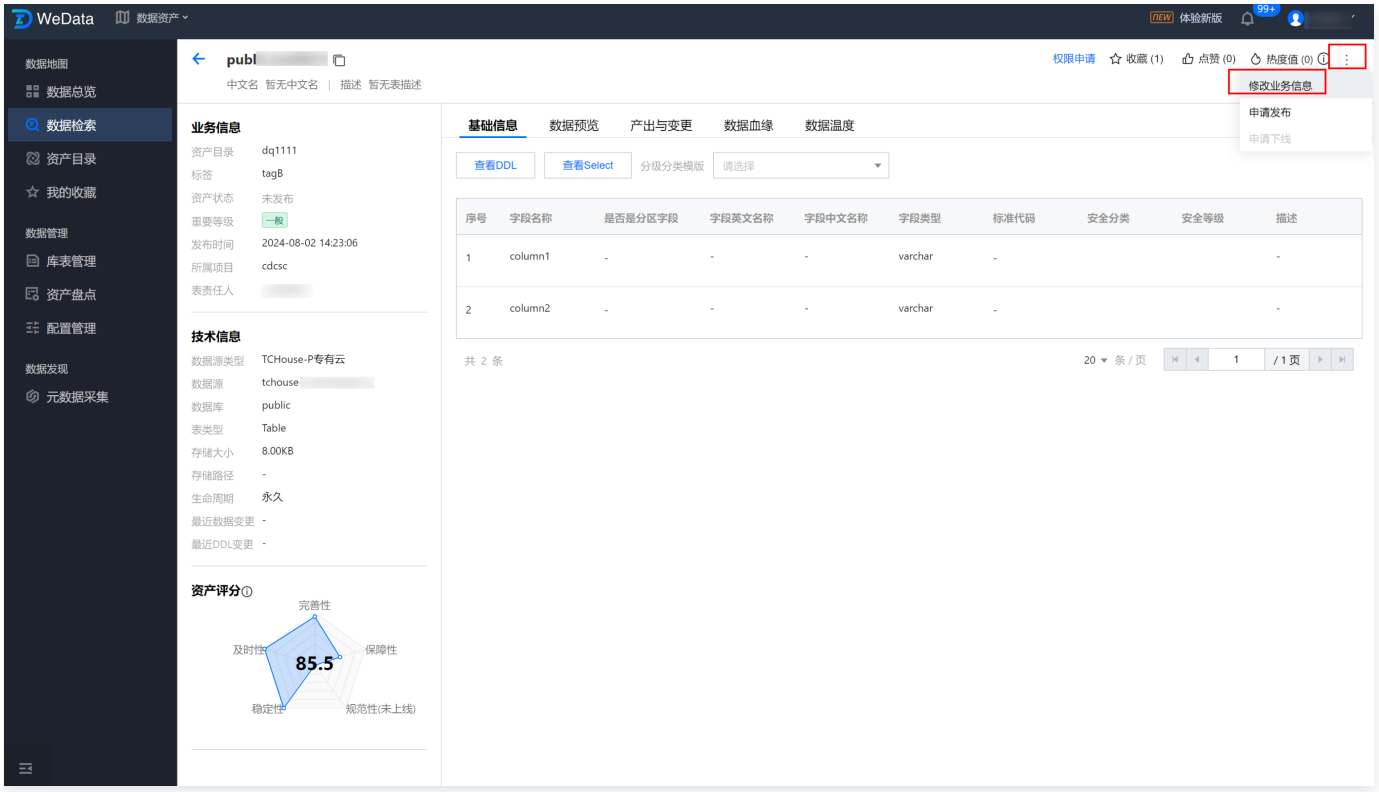
Business Information

- Displays the asset catalog, tag, asset status, importance level, release time, project to which the cluster belongs, and table owner of the current data table.



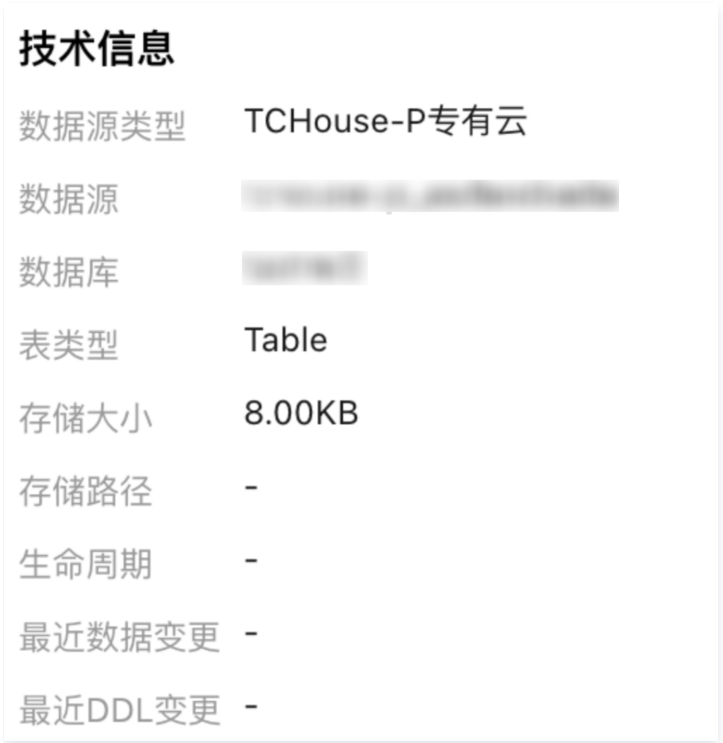
- Business information can be modified and adjusted by opening a pop-up window from the top right corner of the interface.

©2013-2025 Tencent Cloud. All rights reserved.



Technical Information

Displays the current data table's data source type, data source, database, table type, storage size, path, lifecycle, recent data, and DDL changes.



Asset Rating

The overall average score of various indicators, with a perfect score of 100, updated daily.

©2013–2025 Tencent Cloud. All rights reserved.



- Completeness: The current status of technical and business information completeness.
- Assurance: The current status of quality monitoring and access control.
- Timeliness: The status of data output on schedule over the past 30 days.
- Stability: The status of table structure changes over the last 30 days.
- Standardization:**Not Online**.

Basic Information

Provides view DDL and view select features, and displays field name, field English name, field Chinese name, field type, and field description.



Partition Information

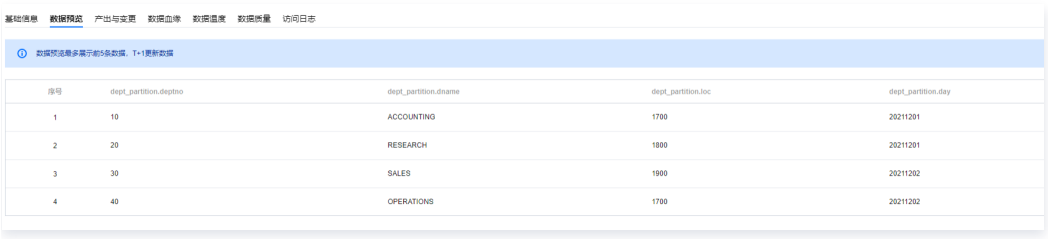
Partition information includes partition field details and specific content of the partition.



Data preview

Preview the data table content. A maximum of the first 5 records is supported. Data is updated on a T+1 basis.

©2013–2025 Tencent Cloud. All rights reserved.



Output and Change

- **Output Information:** Filter by output task/instance time, displaying task ID, execution count, scheduled time, start time, output time, execution duration, and output duration.
- **Change Information:** Display the change records of the table within the last 30 days, including change time, change type, change log, operator, and number of affected tables. Support downloading the change information locally via Excel.



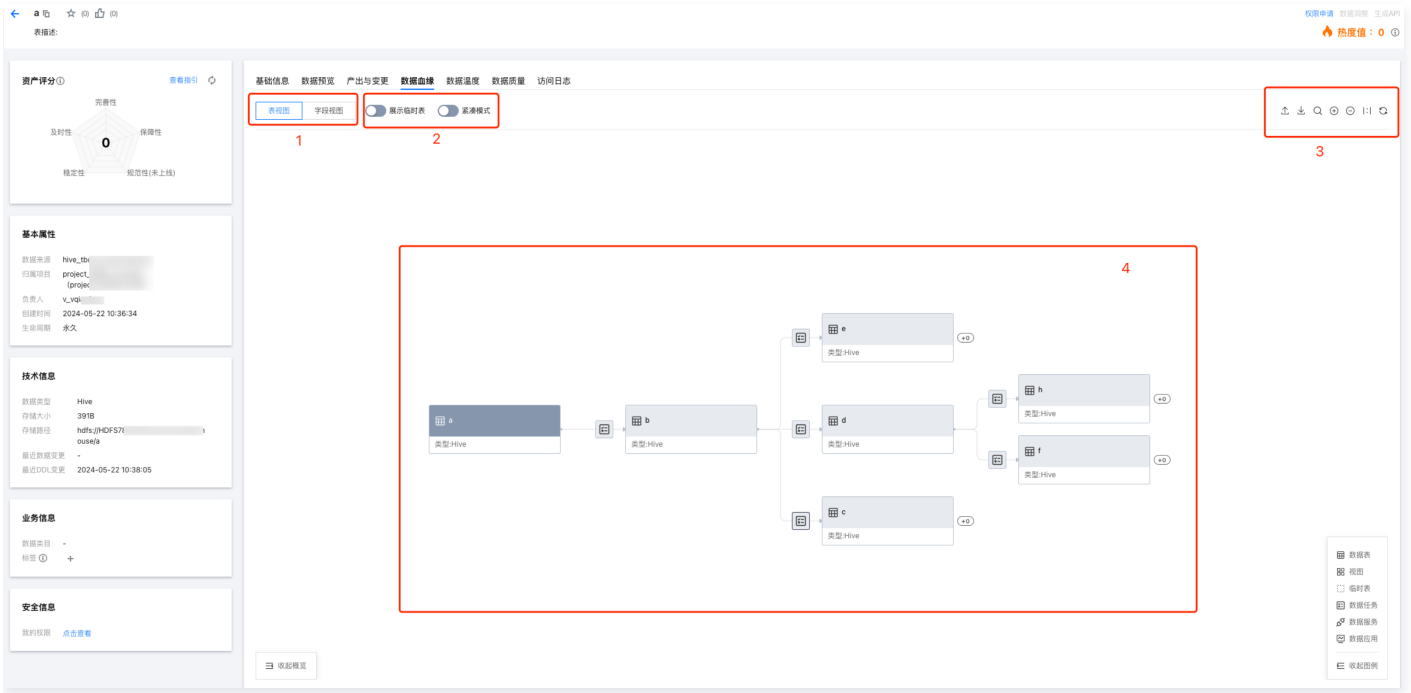
Lineage Relationship

WeData lineage relationship shows the full-chain data flow within all projects under the main account, including data sources, destinations, and associated tasks. The lineage relationship feature provides table/field-level lineage, including lineage between formal data tables and temporary tables used in tasks. The current version of the lineage relationship feature supports sync tasks and Hive SQL task lineage parsing, mainly covering MySQL and Hive table lineage.

Basic Operations

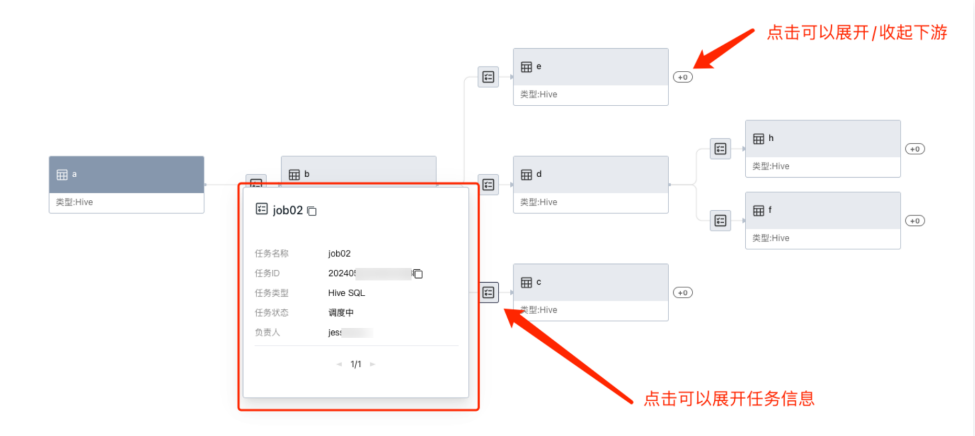
Lineage Relationship feature mainly shows the upstream and downstream data flow of center tables/fields, defaulting to display only the direct first-level upstream and downstream table lineage of the center table. Lineage tracing, switching display object granularity, and other operations can be performed on the canvas. Main features and operations include:

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Table Lineage / Field Lineage	Supporting switching between table/field dimensions to display lineage relationships • **Table Lineage:** Display the upstream and downstream relationships between tables at a table granularity. A node on the canvas represents a table. By default, it shows direct first-level upstream and downstream formal table lineage of the center table. • **Field Lineage:** Display the upstream and downstream related fields between tables at a field granularity. A node on the canvas represents a field. Field lineage defaults to the first field of the table as the center.
Display Temporary Table/Compact Mode	• Temporary Table Display Switch: By default, the bloodline of temporary tables is hidden, and only the bloodline relationship between formal tables involved in tasks is expanded. You can click the icon to switch the settings. After switching, the canvas will initialize to display the upstream and downstream level 1 formal and temporary tables centered on the central table. **Note** A temporary table refers to a table temporarily generated during the task SQL computation process. • Compact mode displays only table names
Import/Export	Supports importing and exporting data bloodlines using Excel files.
Map/Hierarchy mode	Table bloodlines support two modes: full table tracing and single link tracing. The default is map mode tracing.
Search	Supports searching for existing tables/fields on the canvas. After searching, the object will be displayed centered on the canvas.
Canvas tools: Zoom In/Zoom Out/Restore/Full Screen	Set the bloodline canvas and node size.
Bloodline Canvas	Display bloodline between tables/fields:

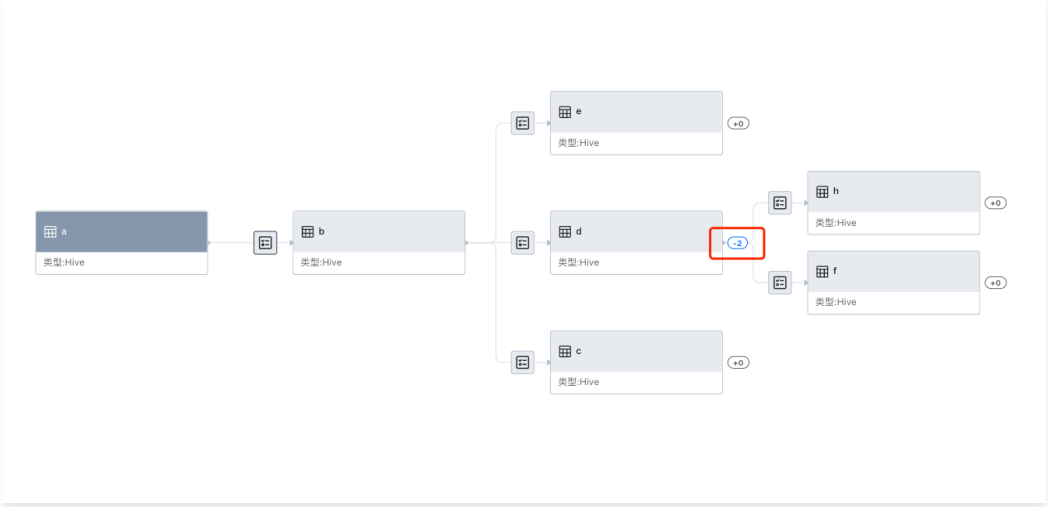
©2013–2025 Tencent Cloud. All rights reserved.



- **Tables/Fields:** A node on the canvas represents a table/field. By default, the table that enters the detail page is used as the central table on the canvas. The tables/fields associated upstream and downstream are displayed on the left and right sides of the central table.
  - Name: Table/Field name. For non-central tables/fields, you can click the link above the node to quickly enter the detail page of that table.
  - Number of Upstream Objects: The number of upstream level 1 tables/fields. If it is 0, it means there are no upstream objects.
  - Number of Downstream Objects: The number of downstream level 1 tables/fields. If it is 0, it means there are no downstream objects.
- **Data Flow Direction:** The arrow direction represents the data flow direction. The left side is the source data, and the right side is the destination data.
- **Associated Task:** Click the arrow to view synchronization/SQL task information associated with this data lineage.
- **Expand/Collapse:** Click the upstream/downstream numbers on the nodes in the canvas to expand/collapse the upstream/downstream objects. If the table/field is downstream of the central node, clicking will only expand its downstream objects; otherwise, it will expand its upstream objects.
- **Quick Expansion: Table Lineage > Map Mode** supports right-click on the target object to quickly expand multi-level upstream/downstream lineage.

Table Lineage

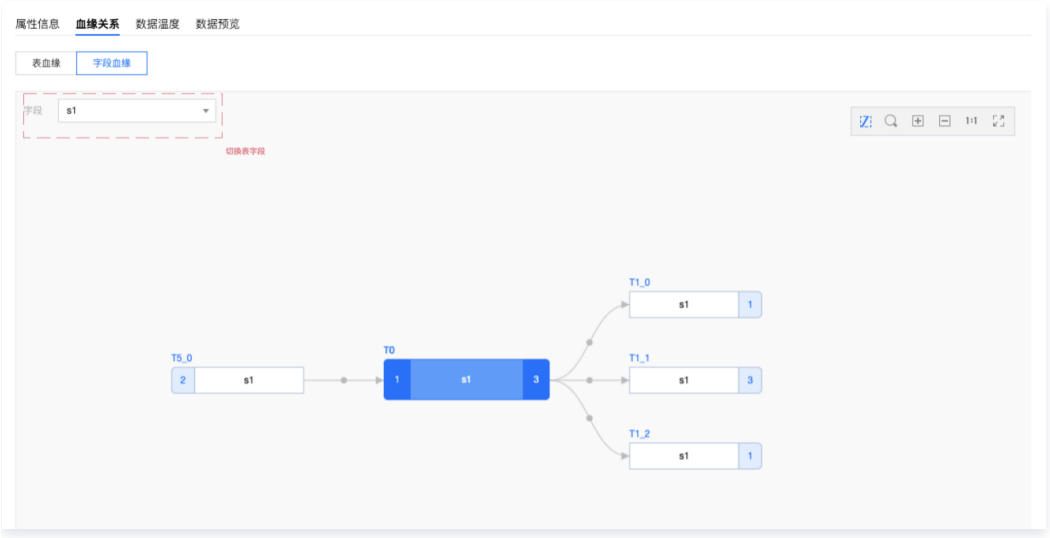
Table Lineage by default displays the direct level 1 upstream/downstream associated tables of the central table. It supports selecting a target table for upstream/downstream lineage tracing, expanding all direct level 1 upstream/downstream tables of the selected table in one go, while keeping the lineage of other tables at the same level unchanged.



Field Lineage

©2013-2025 Tencent Cloud. All rights reserved.

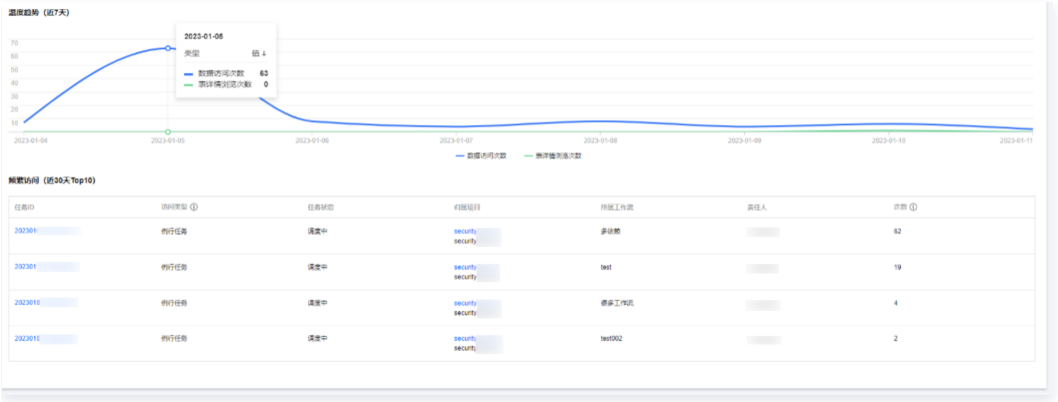
Field Lineage initializes with the first field of the central table. By default, it expands the direct level 1 upstream/downstream associations of that field. It supports selecting a target field for upstream/downstream lineage tracing. You can click the field selector on the upper left side of the canvas to switch displayed fields.



Data Temperature

Data Temperature provides temperature trends and frequent access task information.

- **Temperature Trend:** The number of data accesses and table detail viewings in the past seven days.
- **Frequent Access:** The most frequently accessed tasks in the past 30 days (task ID, access type, task status, project ownership, workflow, owner, and count).



Data Quality

Data Quality provides an overview of the quality control rules configured for data tables and their data quality inspection results.

©2013–2025 Tencent Cloud. All rights reserved.



Access Log

The Access Log provides a statistical overview of data table access, including access date, access account, task ID, access type, number of executions, and other information.

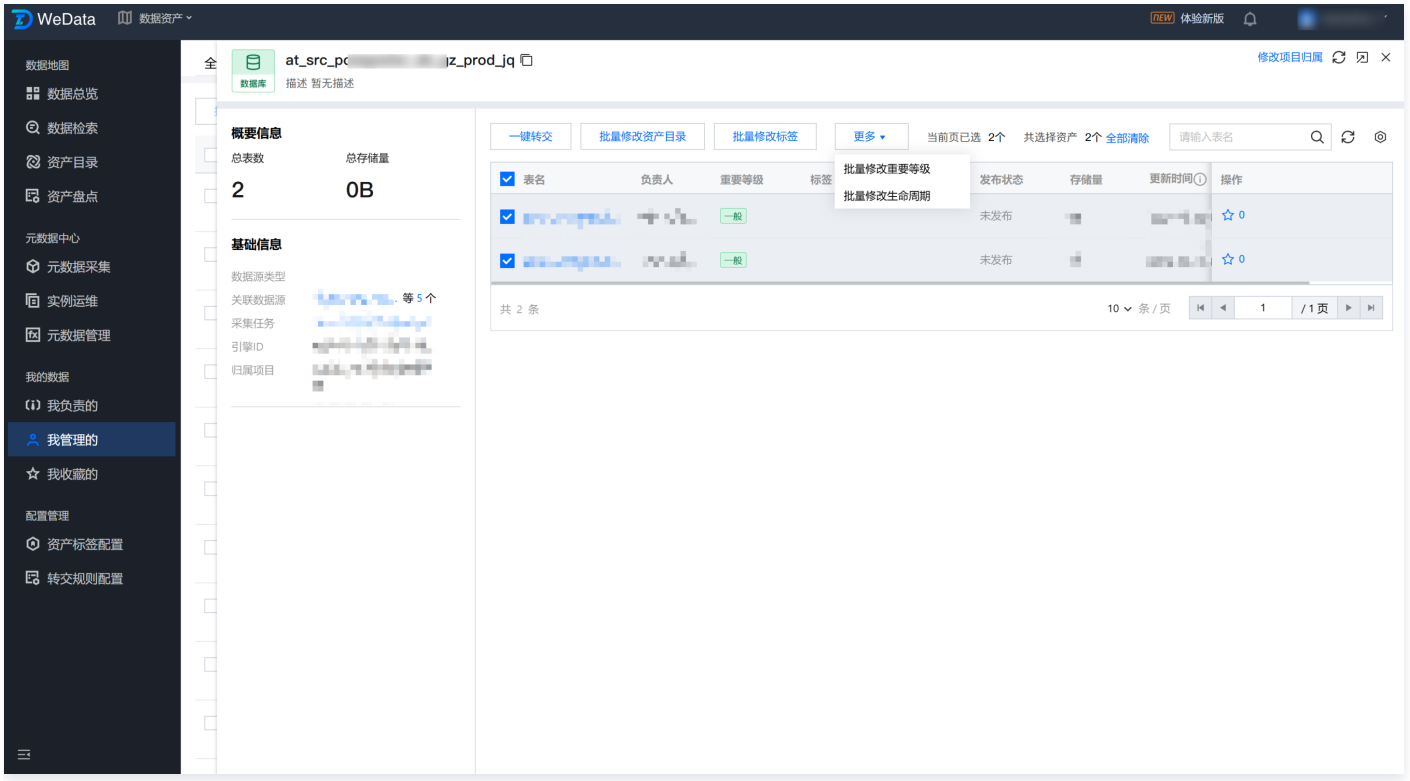


©2013–2025 Tencent Cloud. All rights reserved.

Database Details

Last updated：2024-12-16 16:25:46

Enter the My Management interface, where you can find the **Database List** and **Database Details** you manage. Meanwhile, on the **Asset Inventory** > **Table Details** page, the technical information of the table will display the database of this asset, and you can click to view the database details.



Database Overview

Database Overview shows the total number of tables and the total storage capacity under the database.



Threshold

Basic Information mainly displays the data source type, associated data source, collection task, engine ID, and project ownership. Click the link of the collection task to jump to the details of the corresponding collection task.

©2013-2025 Tencent Cloud. All rights reserved.



Table List

- **Table List** mainly displays the names of all data tables under the database, responsible persons, importance level, tags, asset directory, release status, storage capacity, and update time. Clicking the table name will jump to the table details page.
- **Table List** also supports batch operations in asset inventory, such as one-click transfer, batch modification of asset directory, batch modification of tags, batch modification of importance level, batch modification of lifecycle, and collection.

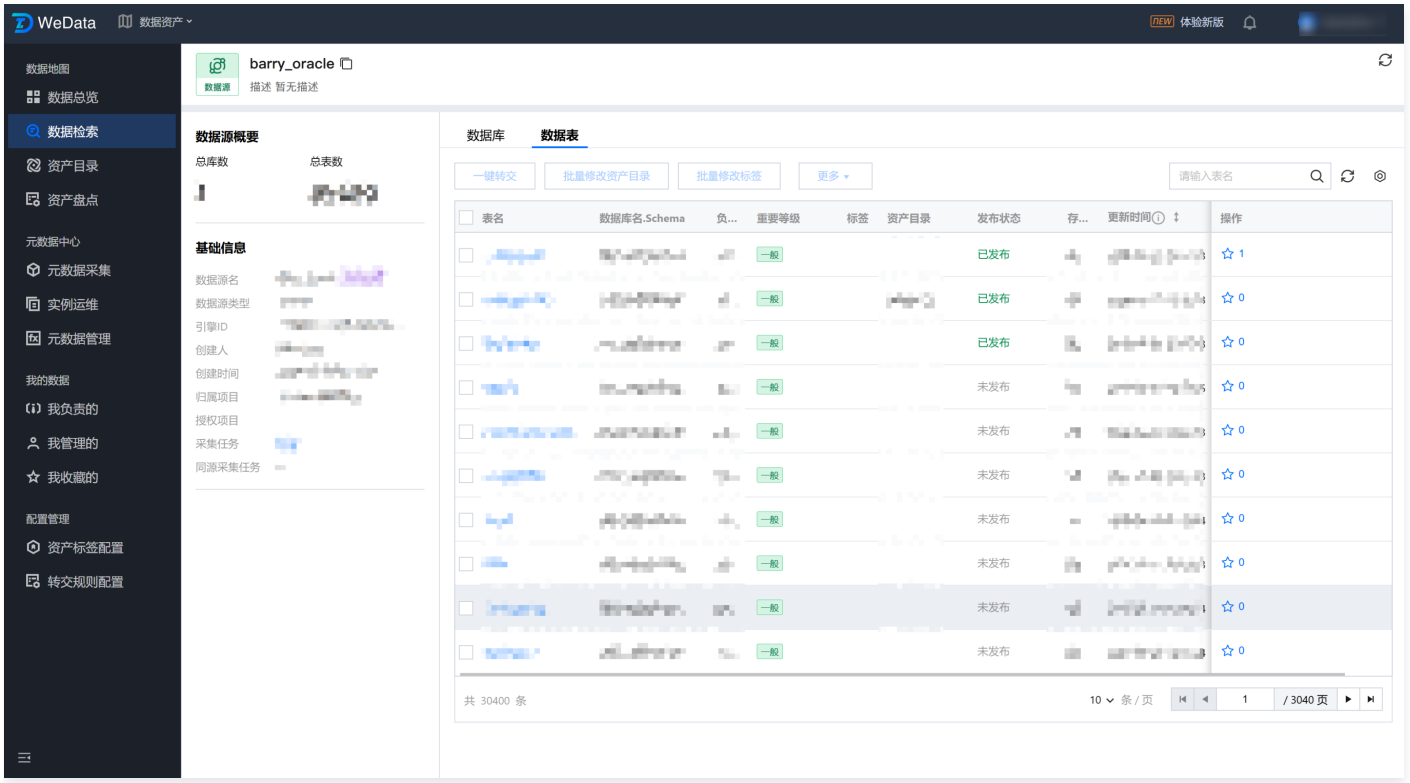


©2013–2025 Tencent Cloud. All rights reserved.

Data Source Details

Last updated：2024-12-16 16:26:09

On the **Asset Inventory** > **Table Details** page, the technical information of the table will display the data source of this asset and the associated data sources. Click to view the details of the data source. The data source details page includes an overview, basic information, and all databases and tables under the source.



Data Source Overview

Data Source Overview shows the total number of databases and tables under the data source.



Threshold

Basic Information mainly displays the data source name, data source type, engine ID, creator, creation time, affiliated project, authorized project, collection tasks, and same-source collection tasks. Additionally, clicking the link of the collection task will redirect to the details of the corresponding collection task.

©2013-2025 Tencent Cloud. All rights reserved.



List of Databases

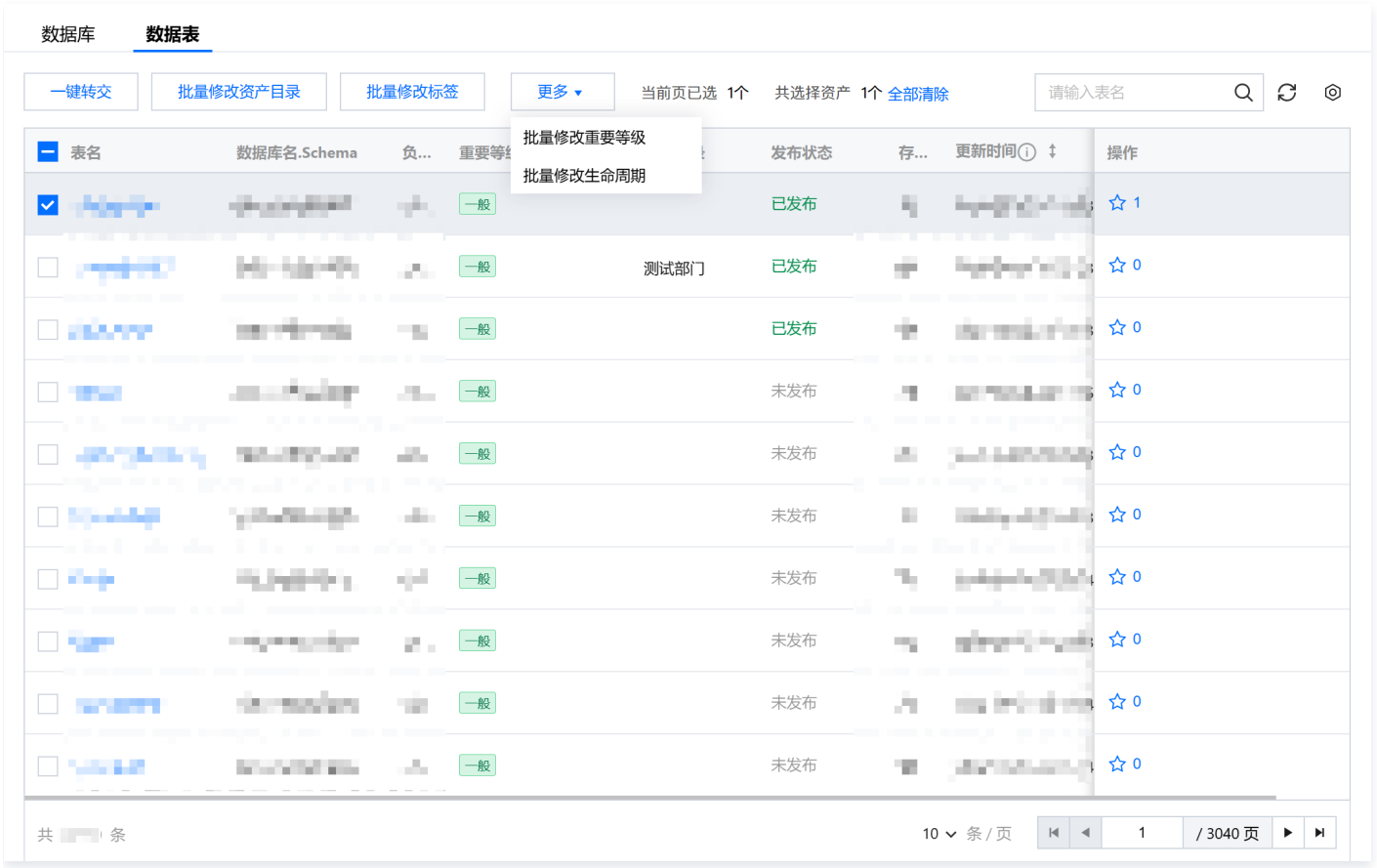
List of Databases mainly displays the names of all databases under the data source, storage capacity, and their affiliated projects. You can batch modify the affiliated projects of the databases under this data source here. Clicking the database name will redirect to the database details page.



Table List

- **Table List** mainly displays the names of all tables under the data source, database name, schema, person in charge, importance level, tags, asset directory, release status, storage capacity, and update time. Clicking the table name will redirect to the table details page.
- **Table List** also supports batch operations in asset inventory, such as one-click transfer, batch modification of asset directory, batch modification of tags, batch modification of importance level, batch modification of lifecycle, and collection.

©2013–2025 Tencent Cloud. All rights reserved.



©2013–2025 Tencent Cloud. All rights reserved.

Collection Task Details

Last updated：2024-12-16 16:26:33

On the Metadata Collection page, click the task name to enter the collection task details page. In the asset details page of the asset inventory (including table, database, and data source details), clicking the collection task will also redirect to the collection task details.



Collection Task Summary

Collection Task Summary shows the total number of databases and tables under this data source.



Threshold

Threshold mainly displays the task name, task owner, task status, creation method, and creation time.

©2013–2025 Tencent Cloud. All rights reserved.



Collection Range

Collection Range mainly displays the data source type, associated data source, and collection range. Click the associated data source to view the data source details page.



Collection Policies

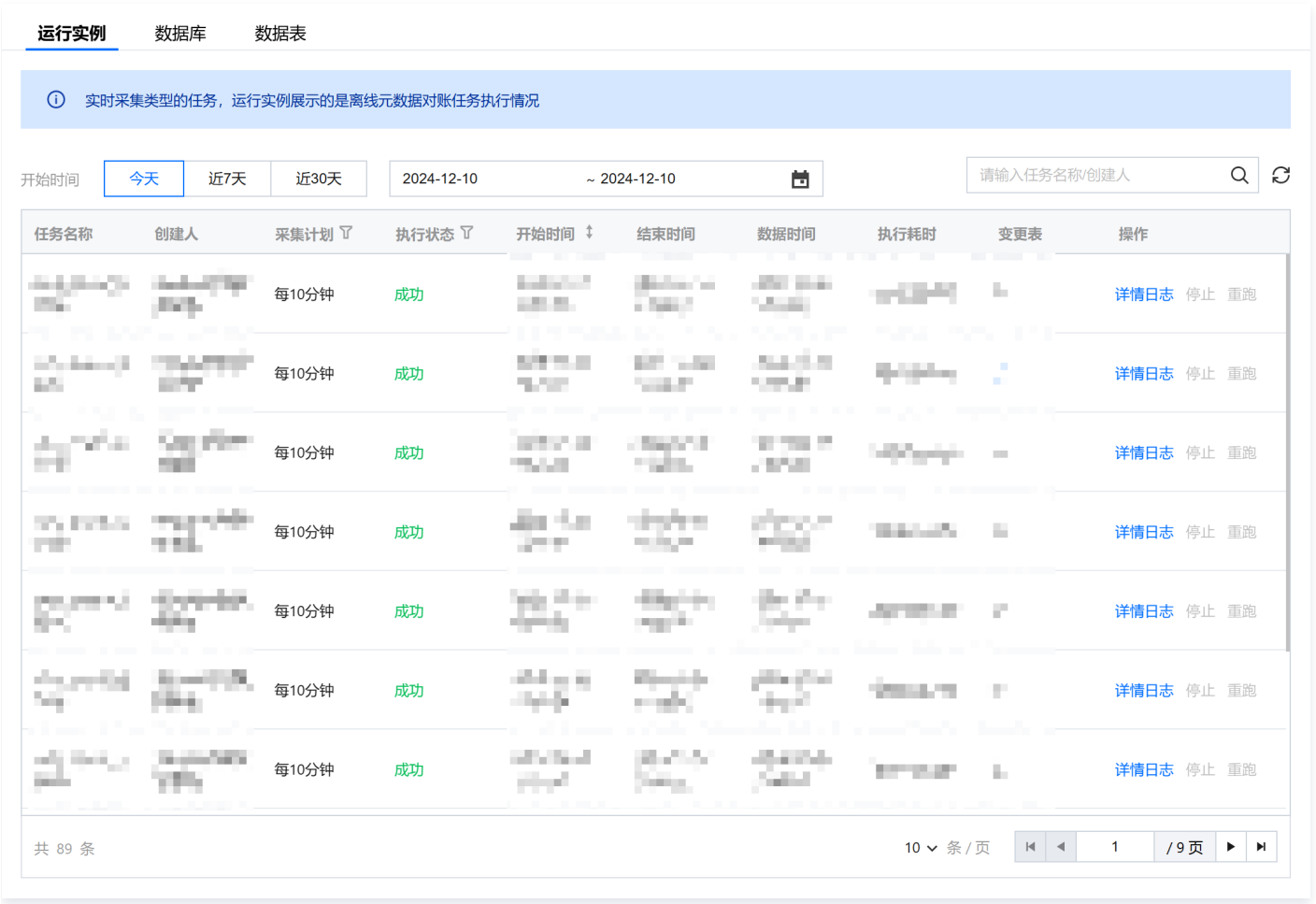
Collection Policies mainly displays the execution resources, execution plan, and the most recent execution time used by this collection task.



Running Instances

Under Running Instances, you can view the details of all running instances of the task, including task name, creator, collection plan, execution status, start time, end time, data time, execution duration, changed tables, and perform stop or rerun operations on running instances.

©2013–2025 Tencent Cloud. All rights reserved.



Database

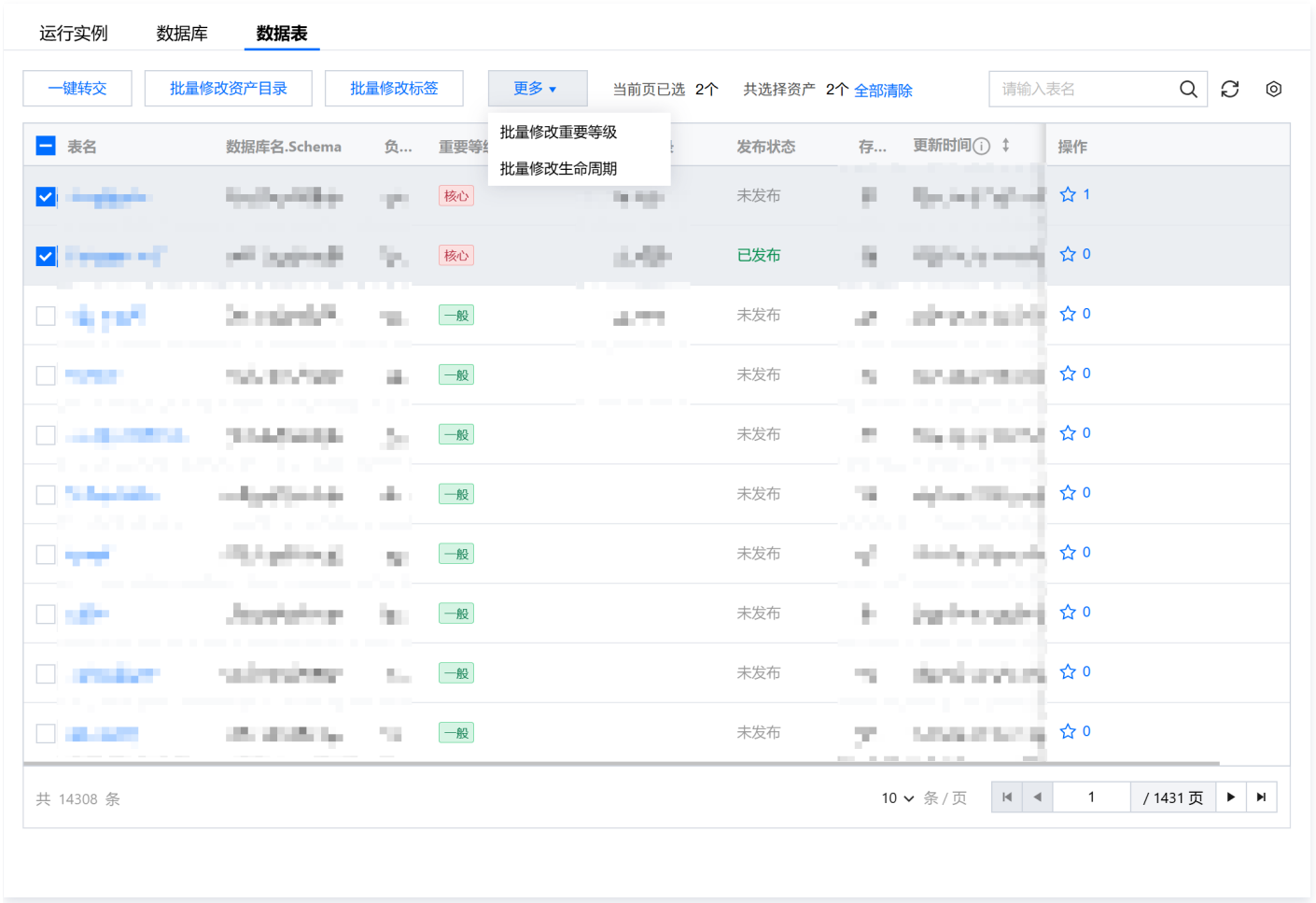
Database mainly displays the databases associated with this collection task. Here you can click the database name to view the database details or modify the project ownership of the databases under the collection task.



Data Table

- **Table List** mainly displays the names of all data tables under this data source, database name, schema, owner, importance level, tags, asset directory, release status, storage volume, and update time. Clicking the table name will redirect to the table details page.
- **Table List** also supports batch operations in asset inventory, such as one-click transfer, batch modification of asset directory, batch modification of tags, batch modification of importance level, batch modification of lifecycle, and collection.

©2013–2025 Tencent Cloud. All rights reserved.



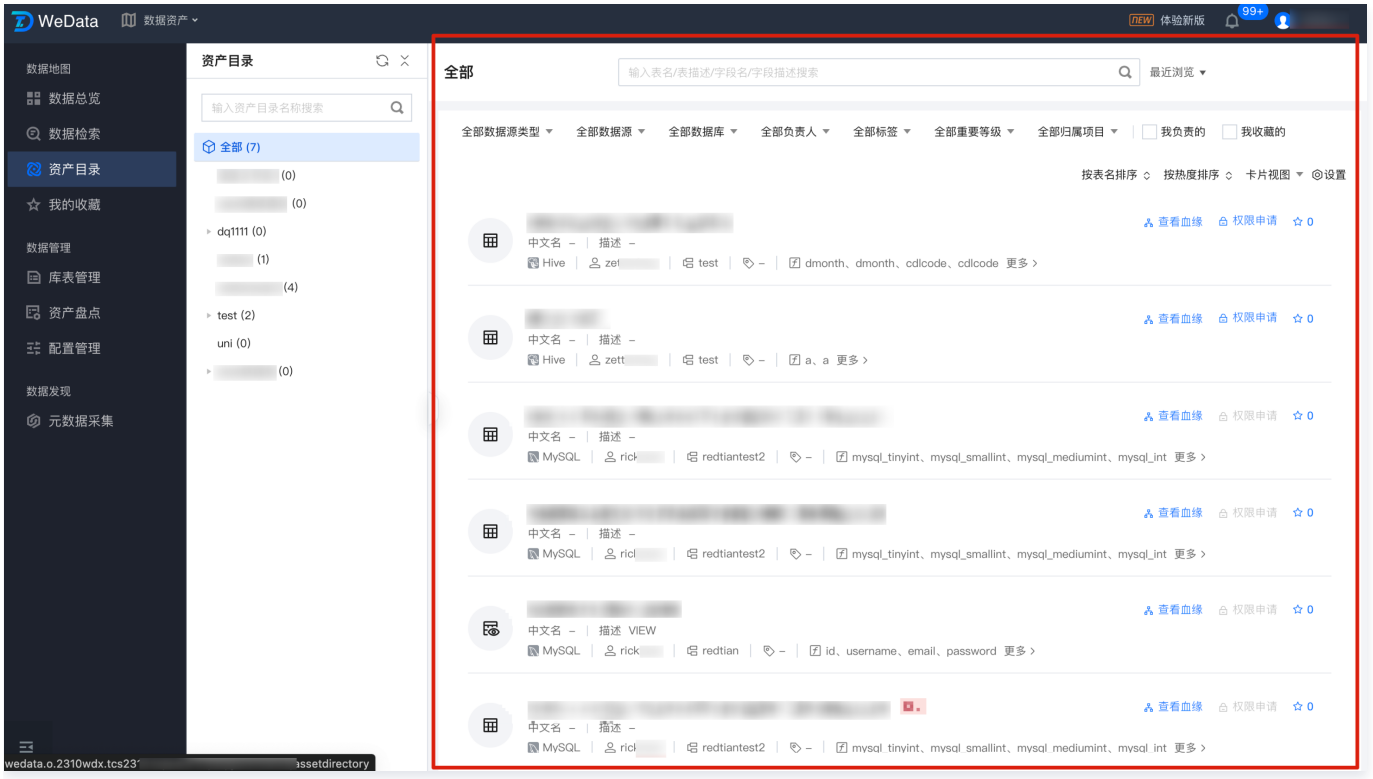
©2013–2025 Tencent Cloud. All rights reserved.

Asset Catalog

Last updated: 2024-12-16 16:27:28

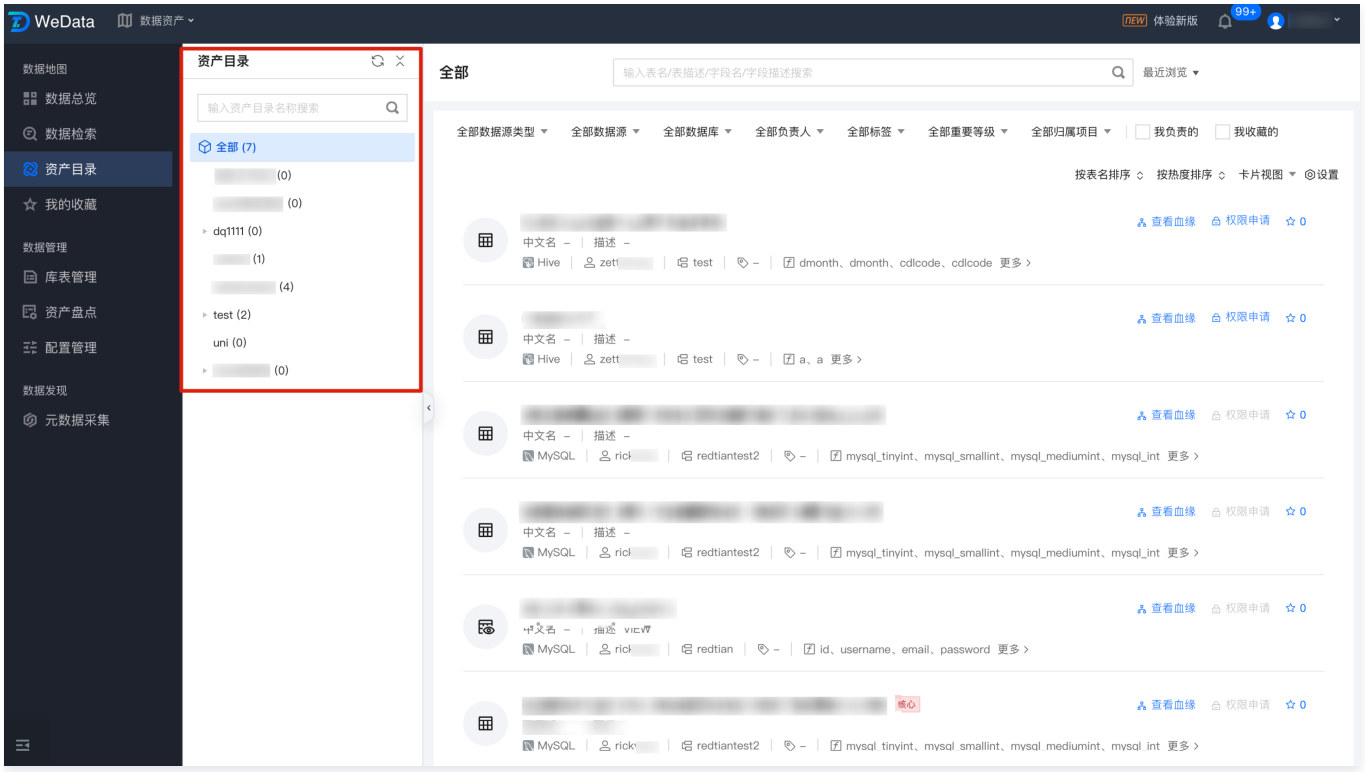
The Asset Catalog feature provides a view of all database table asset information published after inventory under the master account. It supports viewing table details from different perspectives. This page includes **global search**, **data table list** categorized by data source type, data source, database, and person in charge, and the **Asset Catalog**, which facilitates organizing and categorizing your data.

- **Global Search Box:** Provides a quick search entry to help you quickly locate target data tables within the root account.
- **Data Table List:** Includes information such as data source type, owner, asset catalog, Tag source, and fields. It provides features like lineage view, permission application, and favorite/unfavorite.
  - **View Bloodline** shows the bloodline of the selected table.
  - **Permission Application** allows you to apply for permissions for the selected table from the table owner.
  - **Favorite/Unfavorite** allows you to add the selected table to My Favorites for viewing.



- **Asset Catalog:** Helps you search the classification catalog of data, providing a clear directory structure. To create or modify directory nodes, please go to Asset Inventory for organization.

©2013-2025 Tencent Cloud. All rights reserved.



To see table details, click the table name to enter the Table Details page.
Data asset information is periodically updated through data acquisition tasks, see Metadata Collection for details.

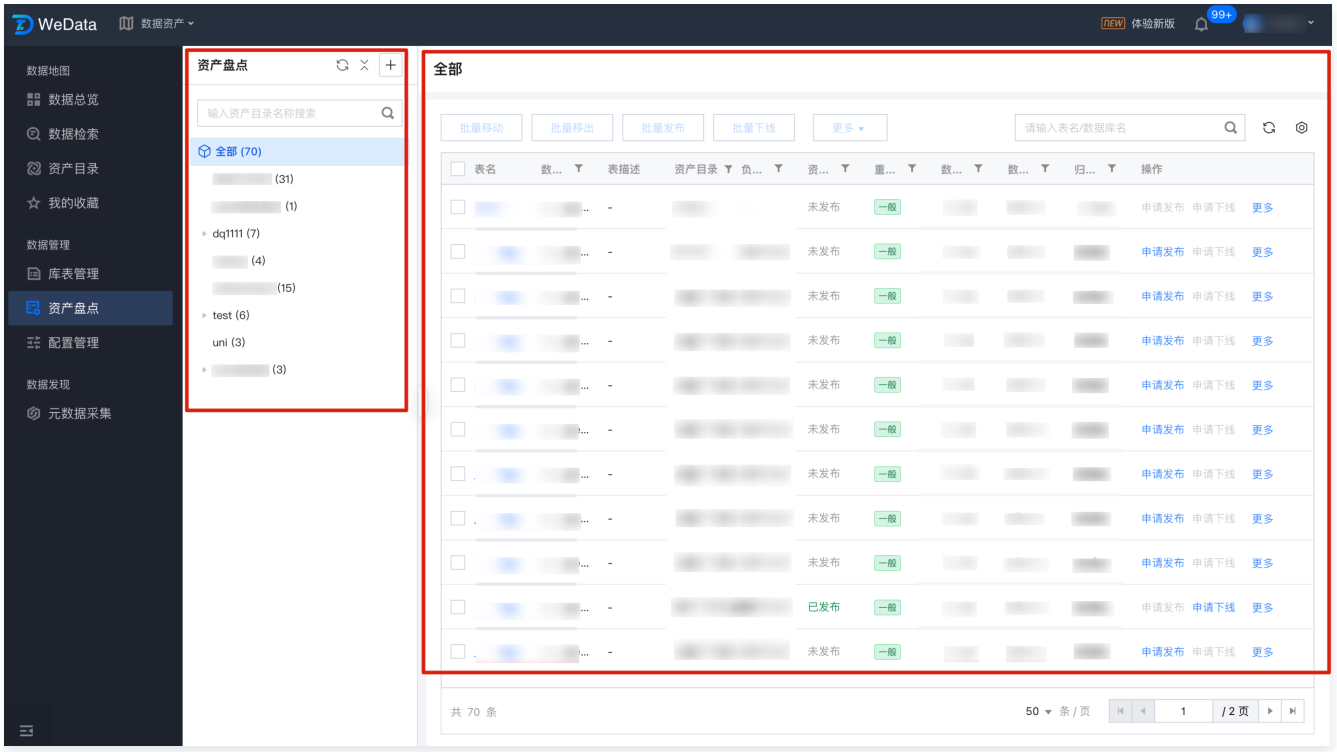
©2013–2025 Tencent Cloud. All rights reserved.

Asset Inventory
Asset Catalog Self-Definition

Last updated: 2024-12-16 16:27:59

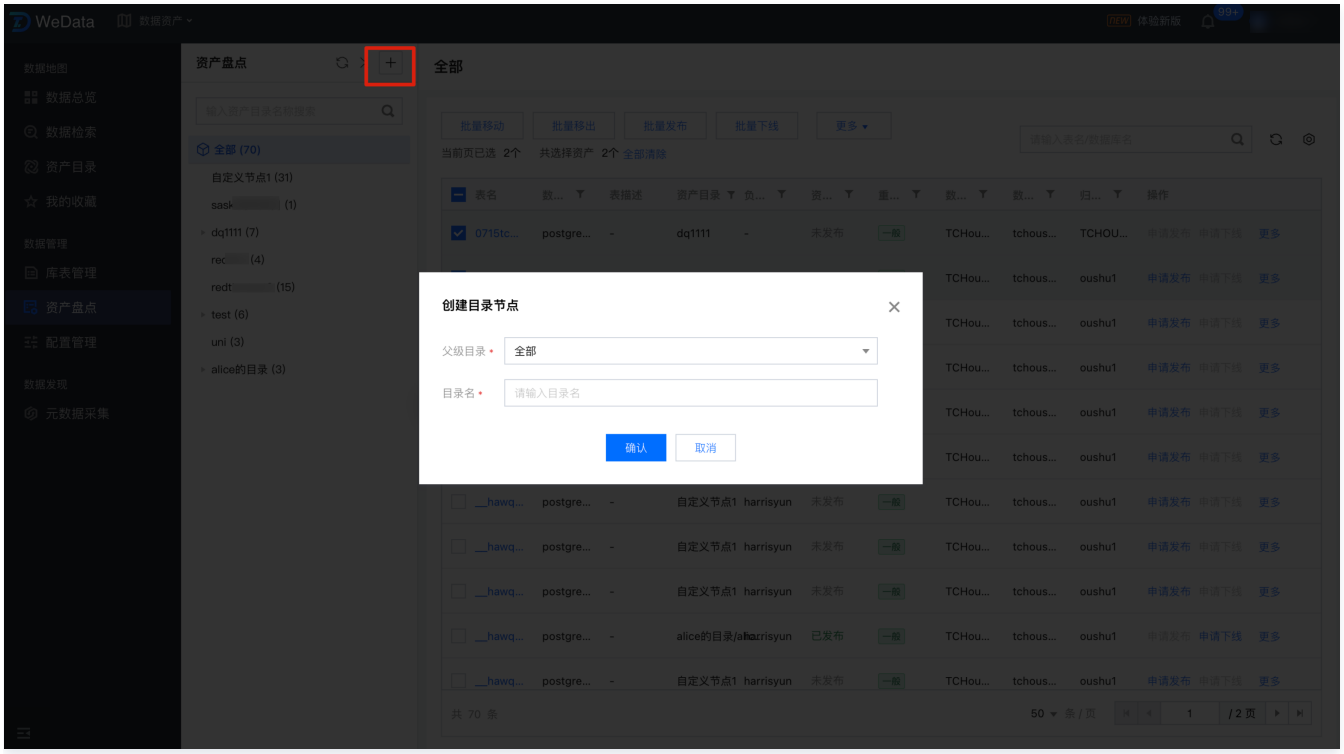
Asset inventory mainly provides configuration and management for the root account's business directory. Based on the enterprise's business scenarios and needs, you can design and classify data into topic categories and data layers under the root account to manage and locate tables more efficiently.
Asset Catalog Self-Definition helps you quickly create, find, and manage directory nodes. You can also batch manage assets through **Add Asset/Move Assets**, **Asset Publishing/Decommission**, and **Modify Business Information**.

In the data list, you can also click the table name to browse the Table Details.



1. Enter the **Asset Inventory** page, you can adjust the directory structure. Click the + button in the top left corner, you can create directory nodes by selecting the parent directory according to business needs.

©2013-2025 Tencent Cloud. All rights reserved.



2. You can also directly click the + or ... button next to the parent directory you want to edit to create a subdirectory or directly edit the parent directory.

- Method 1:



- Method 2:

©2013–2025 Tencent Cloud. All rights reserved.



©2013–2025 Tencent Cloud. All rights reserved.

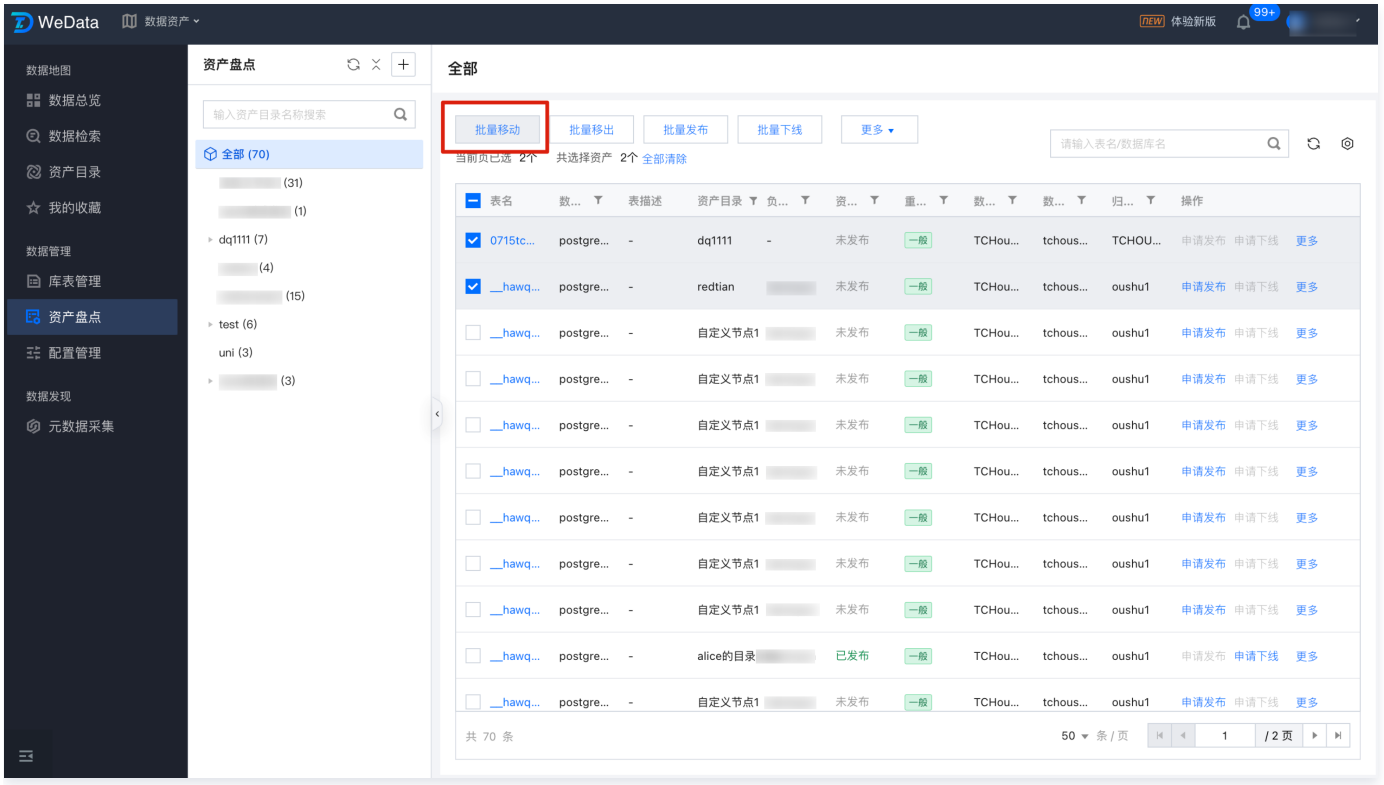
Batch Inventory Assets

Last updated：2024-12-16 16:28:24

After selecting multiple assets, you can perform batch operations such as batch moving, batch removal, batch publishing, and batch decommissioning.

Batch Move

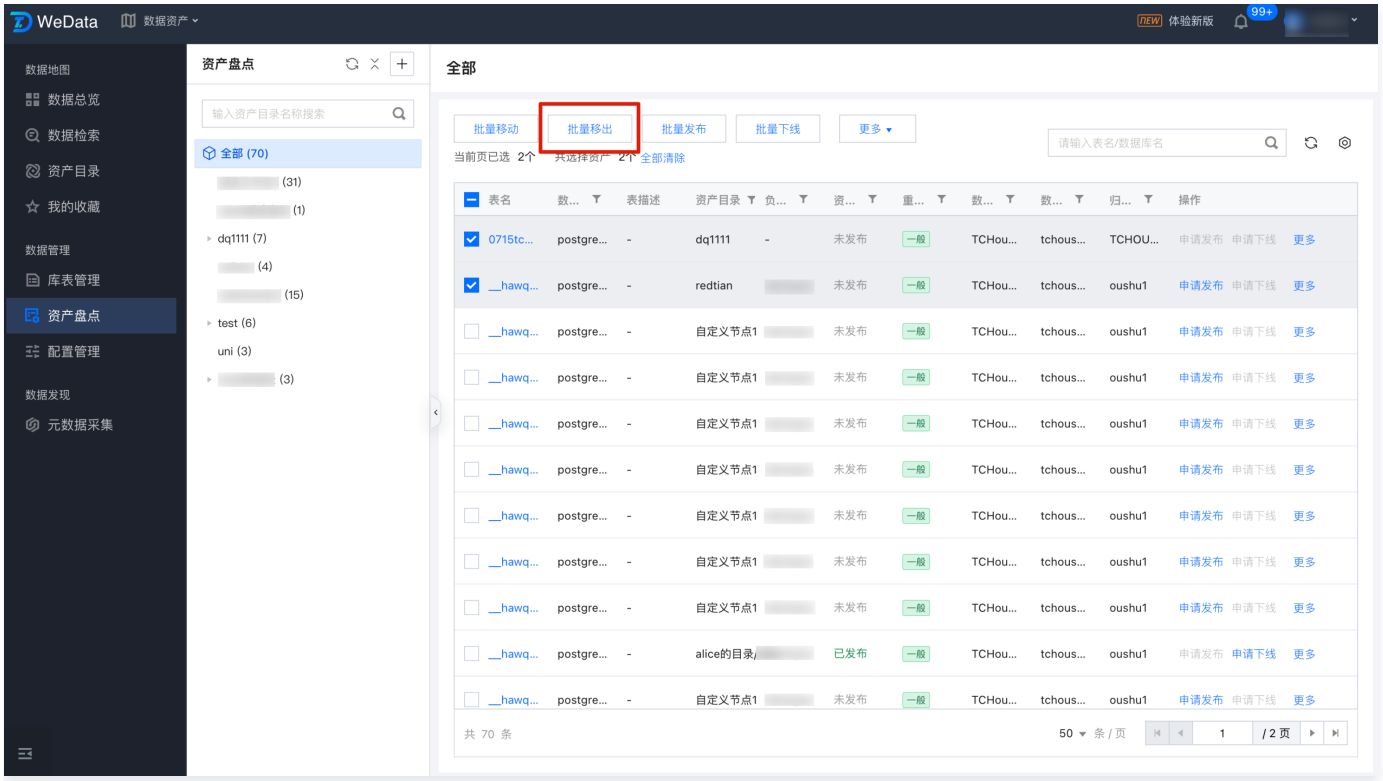
Go to the **Asset Inventory** page, select the assets you want to move, click the **Batch Move** at the top, and you can batch move the assets to other directory nodes.



Batch Removal

Go to the **Asset Inventory** page, select the assets you want to remove, click the **Batch Remove** at the top, and you can batch remove the assets from the current directory node.

©2013-2025 Tencent Cloud. All rights reserved.



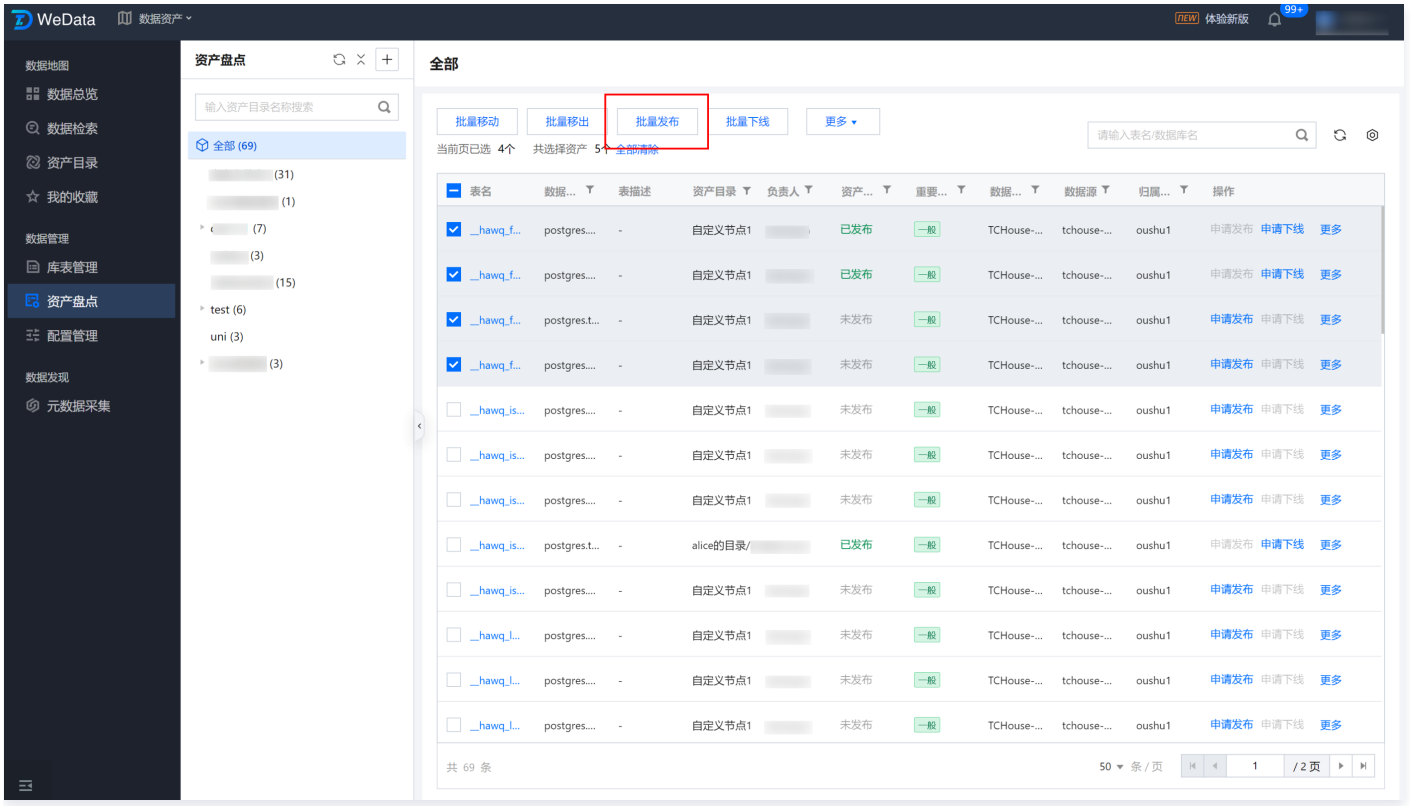
Batch Publish

The feature supports batch publishing, you can select multiple assets to publish at once. Additionally, you can also fill in the reason for applying for publication for better publication management.

> **Note**
> Only published assets can be browsed in the asset directory.

Go to the **Asset Inventory** page, select the assets you want to publish, click the **Batch Publish** at the top, and you can batch publish the assets at once.

©2013–2025 Tencent Cloud. All rights reserved.

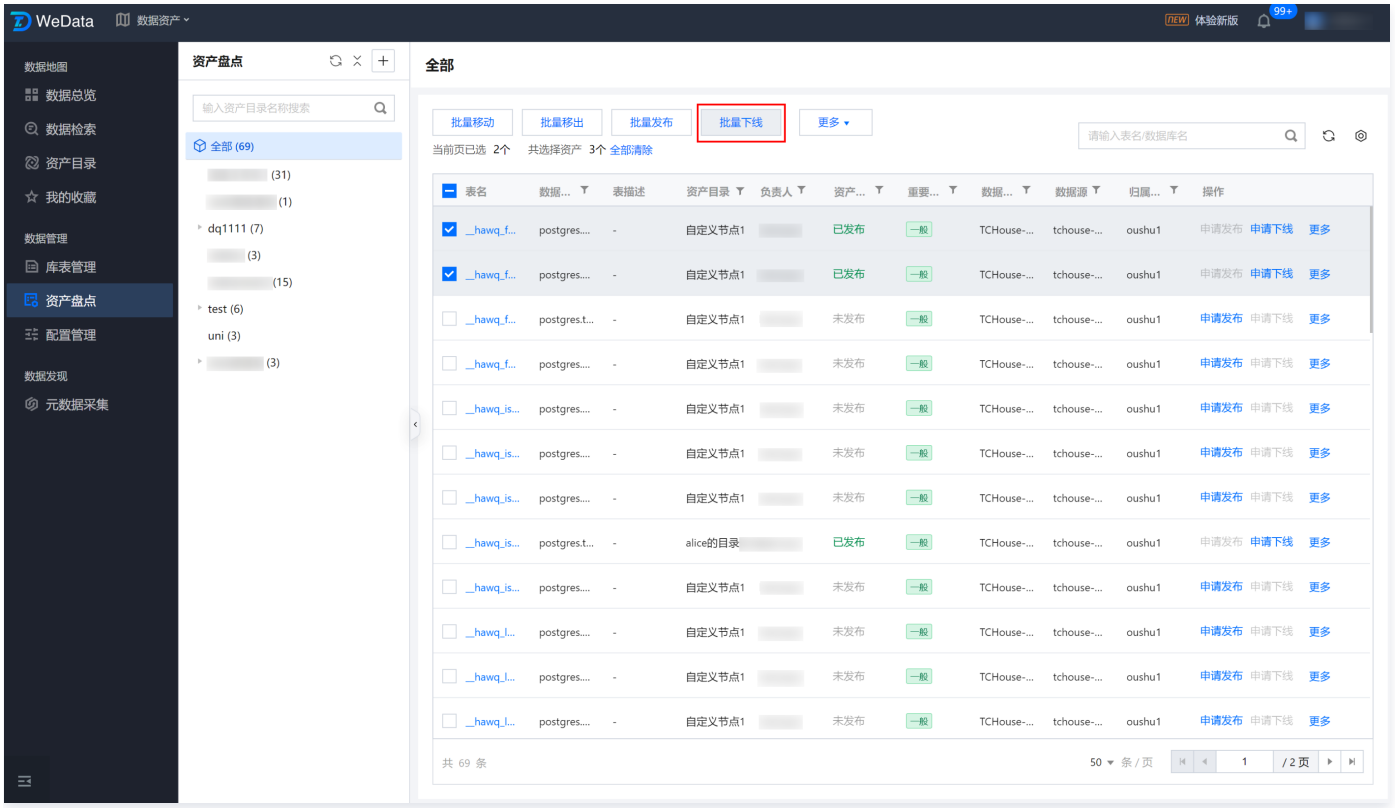


Batch Decommission

You can also batch decommission the selected assets.
Go to the **Asset Inventory** page, select the assets you want to decommission, click the **Batch Decommission** at the top, and you can batch decommission the assets at once.

> **Note:**
> After decommissioning, assets will no longer be retrievable in the asset directory; however, users who have already obtained usage rights can still continue to use them.

©2013–2025 Tencent Cloud. All rights reserved.





©2013–2025 Tencent Cloud. All rights reserved.

Modify Business Information

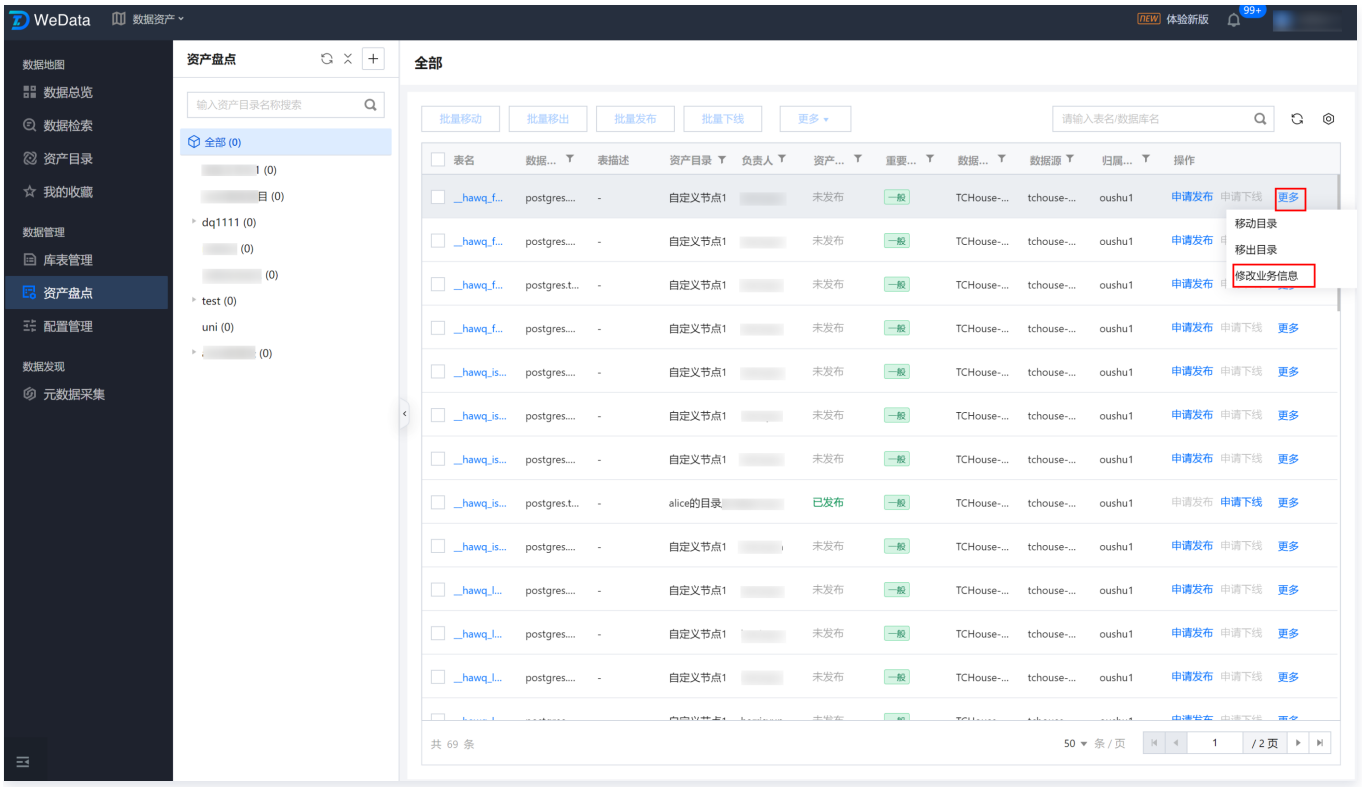
Last updated：2024-12-16 16:28:46

You can manage your business information through asset inventory. It supports single table modification of business information and batch download/batch modification of importance level.

Single Table Modification of Business Information

You can modify the business information of the asset, including asset catalog, asset tags, responsible person, importance level, lifecycle, and other important information.

1. Enter the **Asset Inventory** page, find the operation you need to modify, and click **More** > **Modify Operation Information**.



2. Enter the **Modify Operation Information** page, you can modify the **Operation Information** and **Technical Information**. After the modification, click **Save**.

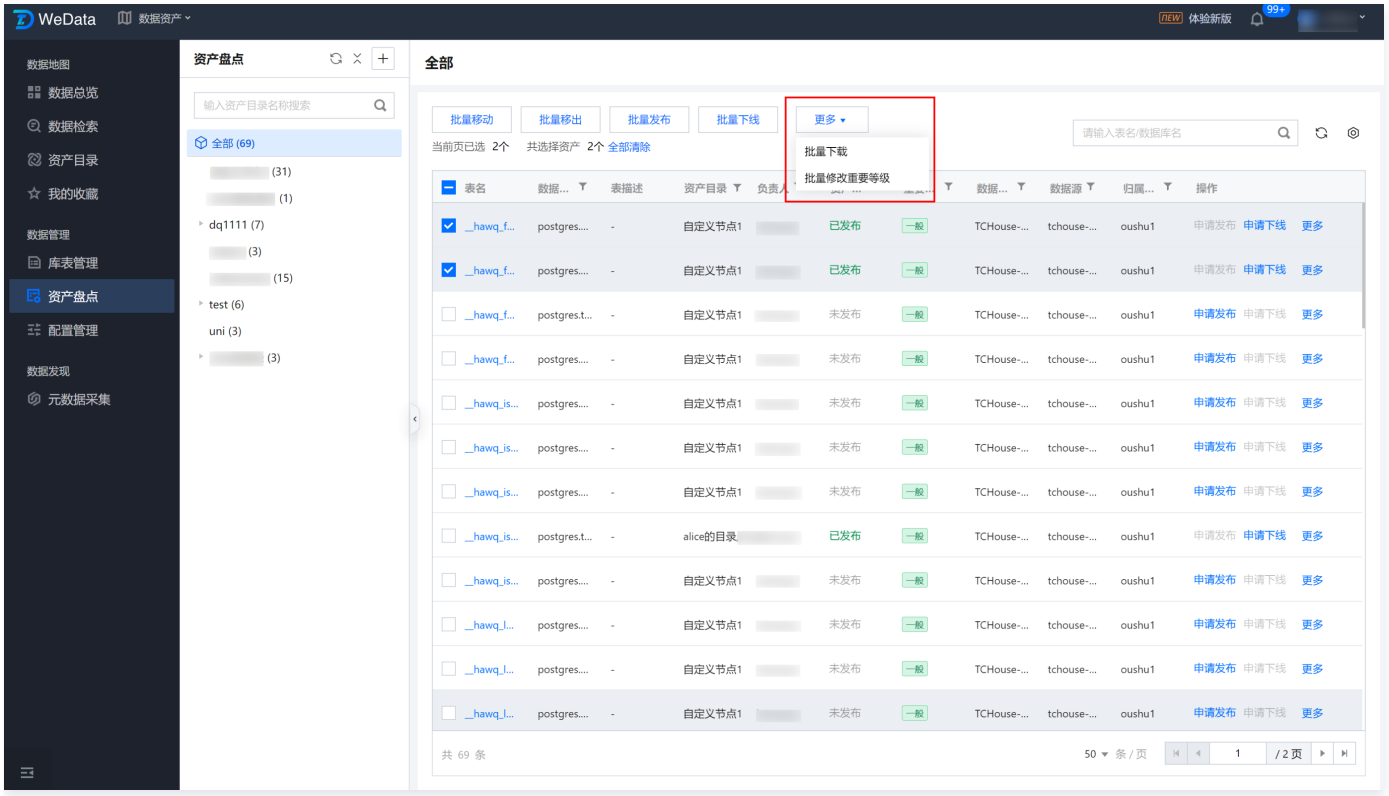
©2013-2025 Tencent Cloud. All rights reserved.



Batch Download/Batch Modification of Importance Level

Additionally, you can batch download selected assets or batch modify the importance level according to business needs.

1. Enter the Asset Inventory page, click the top **More** > **Batch Modify Importance Level**.



2. Enter the **Modify Importance Level** page, we recommend classifying according to business logic, modifying the importance level for easier retrieval and management later. After modification, click **Confirm**.

©2013–2025 Tencent Cloud. All rights reserved.



©2013–2025 Tencent Cloud. All rights reserved.

Data Discovery
Metadata collection

Last updated: 2024-12-16 16:29:12

WeData provides users with a metadata collection task feature from Definition. Administrators can perform metadata collection on the data source before enabling visual management. The collection granularity supports databases, and only one collection task can be created per database. The collection task will run and update metadata information based on the configured cycle, and manual execution and task editing management operations are also supported.

Supported Data Source Types

The currently supported data source types for metadata collection are as follows:

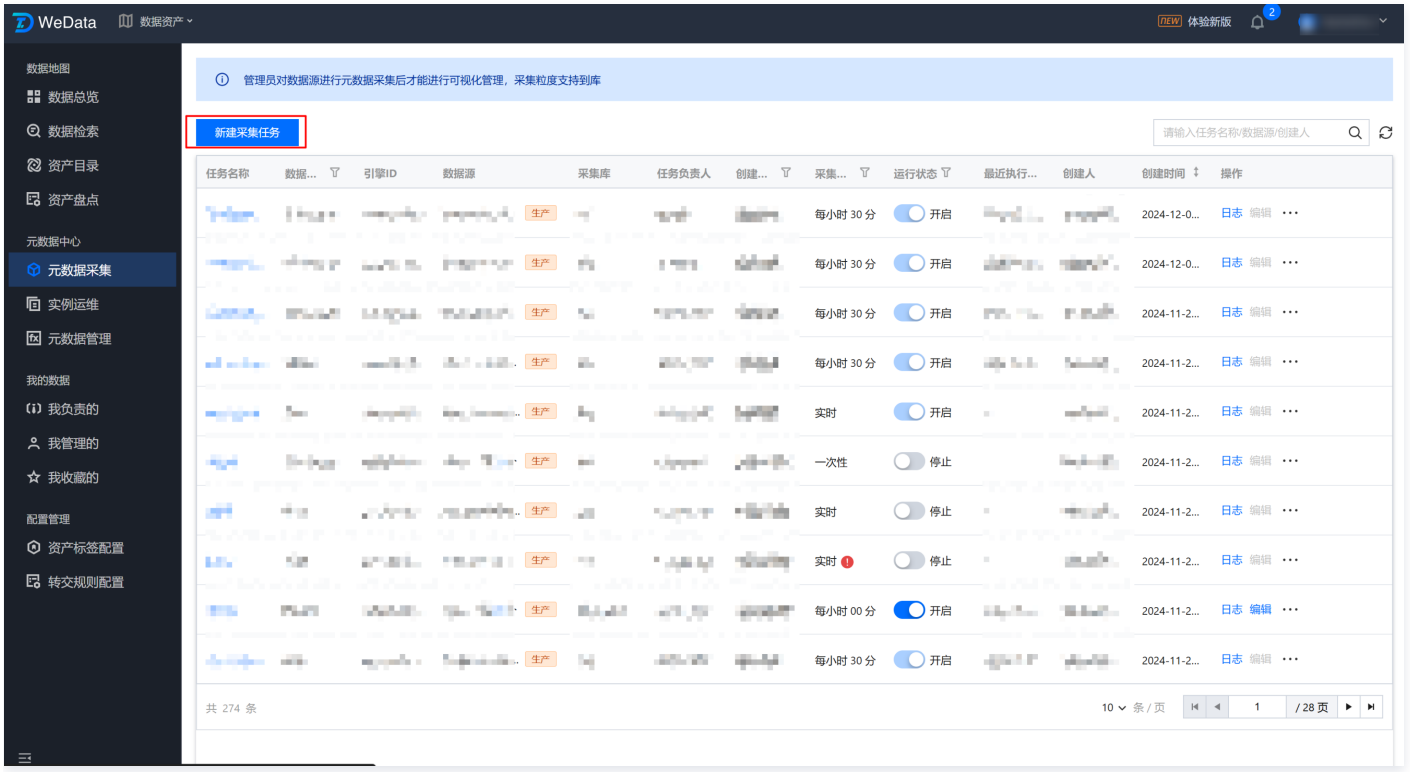
Information	Type
Big data	Hive
	HBase
	DLC
	ClickHouse
	TCHouse-C
	Iceberg
	Greenplum
	Doris
	StarRocks
	TCHouse-D
Relational Database	MySQL
	TencentDB for MySQL
	PostgreSQL
	Oracle
	SQL Server
	TCHouse-P
	TDSQL-PostgreSQL

Collection Task Configuration

Configuring a Collection Task

1. Enter the Cloud Data Collection interface, click **Create Collection Task**, and select the data source type.

©2013-2025 Tencent Cloud. All rights reserved.



2. Set the collection object.

©2013–2025 Tencent Cloud. All rights reserved.



Each collection task supports binding to a single data source under the WeData project. Data sources cannot be repeatedly bound to collection tasks.

Information	Details
Task Name	The name of the collection task cannot be empty. The name can start with letters or Chinese characters and may contain letters, Chinese characters, numbers, hyphens (-), and underscores (_)
Description	Optional, provide a description of the collection task
Associated Project	Specify the project to which the data source will belong and bind the data source management permissions to the project
Data Source	The name of the data source corresponding to the collection task. This data source can be viewed in the Project Management module
Database	A database can only correspond to one collection task. Databases that have already been collected cannot be selected
Data Table	Data Table to be Collected
Designate Table Owner	Person in Charge of Table-Level Permission Management
Task Owner	The person in charge has the authority to view, stop, start, view logs, view details, rerun, and modify task information

©2013-2025 Tencent Cloud. All rights reserved.

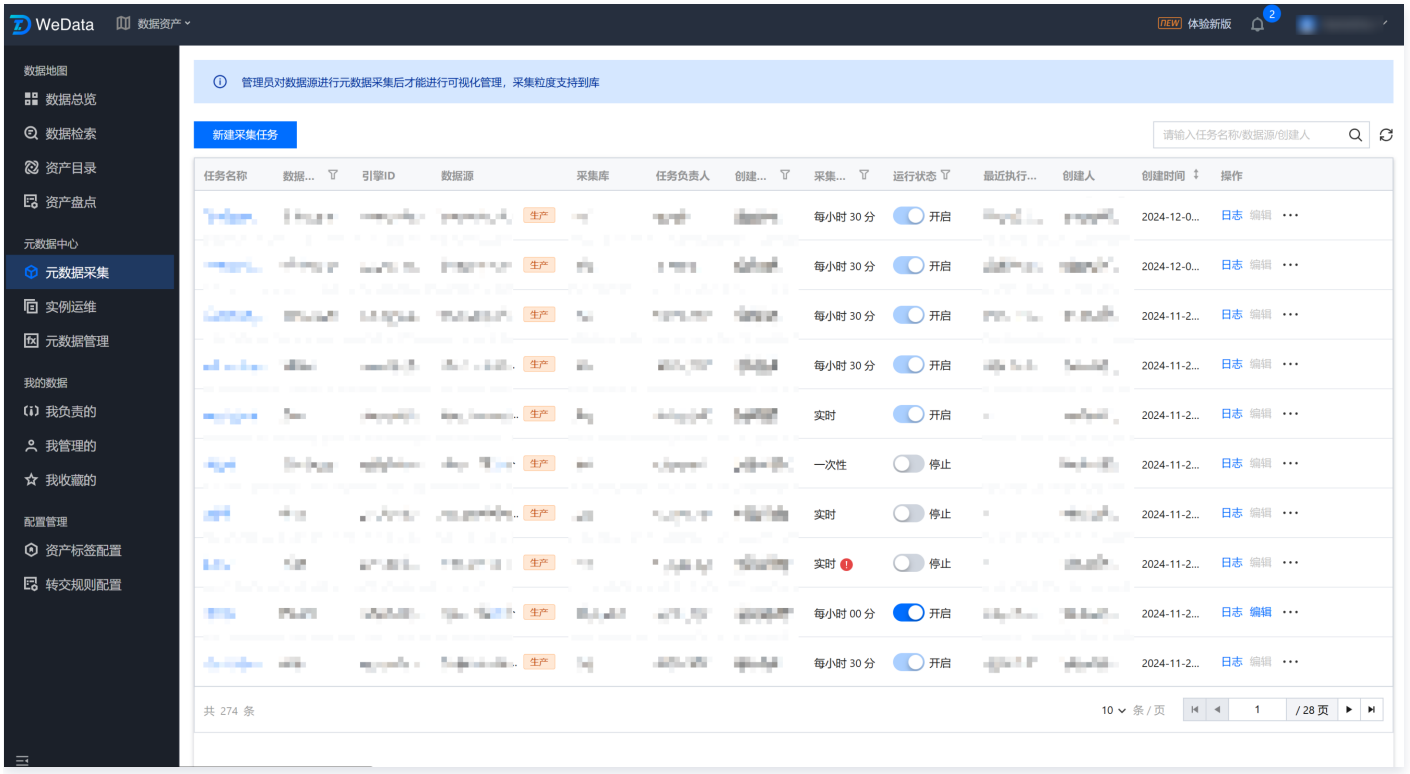
3. configure the collection schedule

Configure the cycle, specific date, and time information for the metadata collection task.

- Collection cycle: The current version supports hourly, daily, weekly, monthly, and one-time collection tasks.
- Collection Date: For weekly and monthly tasks, specific collection dates can be specified. Once configured, tasks will run as scheduled on these dates. For example, if the collection dates are set to the 1st, 5th, and 31st, the metadata collection task will run on the 1st, 5th, and 31st of each month.
- Execution Time: The specific time when the task is executed
- Run Immediately: After configuration, the collection task will be triggered immediately.

Collection Task List

The collection task list provides information on all collection tasks under the current user, including task name, collection object, technical type, project, task owner, and creator. It offers operations such as viewing task details, logs, editing, deleting, and transferring.



Information	Details
Task Name	Collection Task Name
Data Source Type	Data Source Type
Data Source	Data Source of the Collection Task
Collection Database	Collection Database of the Collection Task
Task Owner	Current Task Owner's Account Name
Creator	Account Name of the User who Created the Collection Task
Creation Time	Creation Time of the Collection Task
Collection Schedule	Collection Task Operation Cycle
Running Status	Running Status of the Collection Task
Most Recent Execution Time	Date and Time Information of the Most Recent Execution of the Collection Task

©2013–2025 Tencent Cloud. All rights reserved.

Operation	Provide operations including viewing collection task details, logs, editing, deleting, and transferring.

Run Collection Task

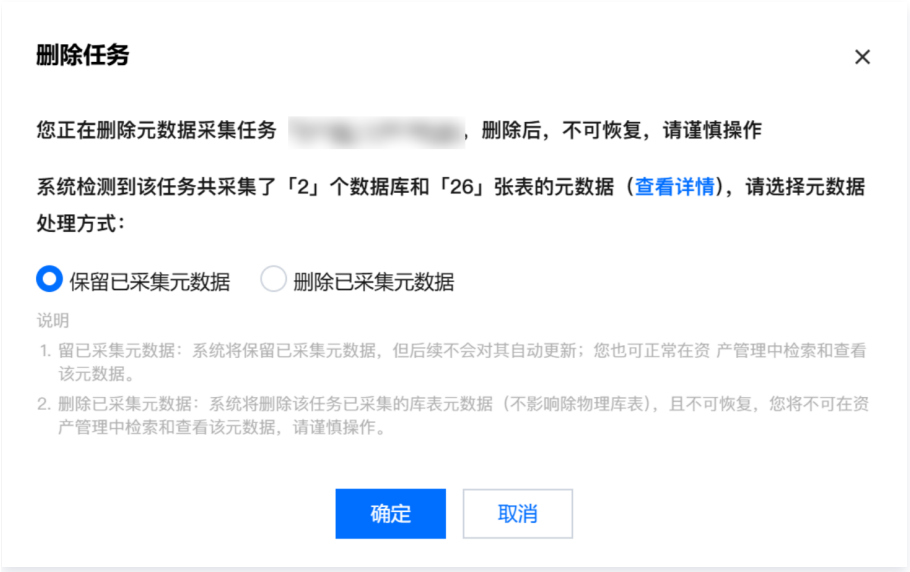
Manually run one-time/periodic tasks, tasks not in progress can be manually run.

Edit Collection Task

In a non-executing state, you can edit items, data sources, update methods, delete methods, and collection plans; support changes to the same type of data source, and the collection task collects based on the newly bound collection objects.

Delete Collection Task

After the collection task is deleted, the collection of the data source will stop, and the collected metadata can be deleted simultaneously.



Transfer Collection Task

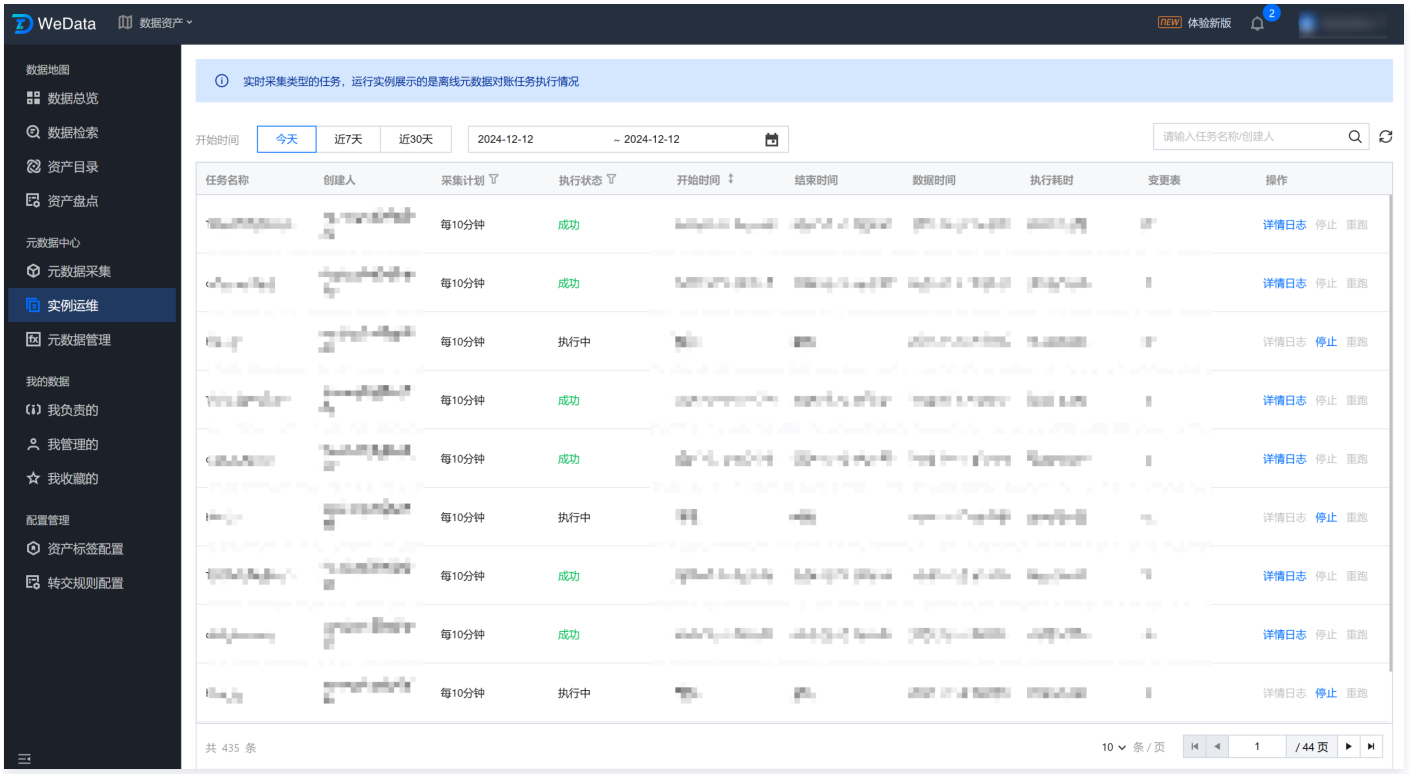
In the collection task list, you can transfer the collection task to another task owner, and the original owner will no longer have management rights for the task.

©2013-2025 Tencent Cloud. All rights reserved.

OPS of Instances

Last updated：2024-12-16 16:29:34

Through the OPS of Instances feature, you can manage instances after the collection tasks are executed, including viewing task name, creator, collection plan, execution status, start time, end time, data time, execution duration, change table, and viewing instance run logs, stopping, and rerunning.



Information	Details
Task Name	Collection Task Name
Creator	Account Name of the User who Created the Collection Task
Collection Schedule	Instance operation cycle
Execution Status	Instance execution status
Start Time	Instance start time
End Time	Instance end time
Data time	Instance data collection time
Execution duration	Instance running duration
Change table	The instance task collects data on table changes. Click to view change details and export change information.

©2013–2025 Tencent Cloud. All rights reserved.



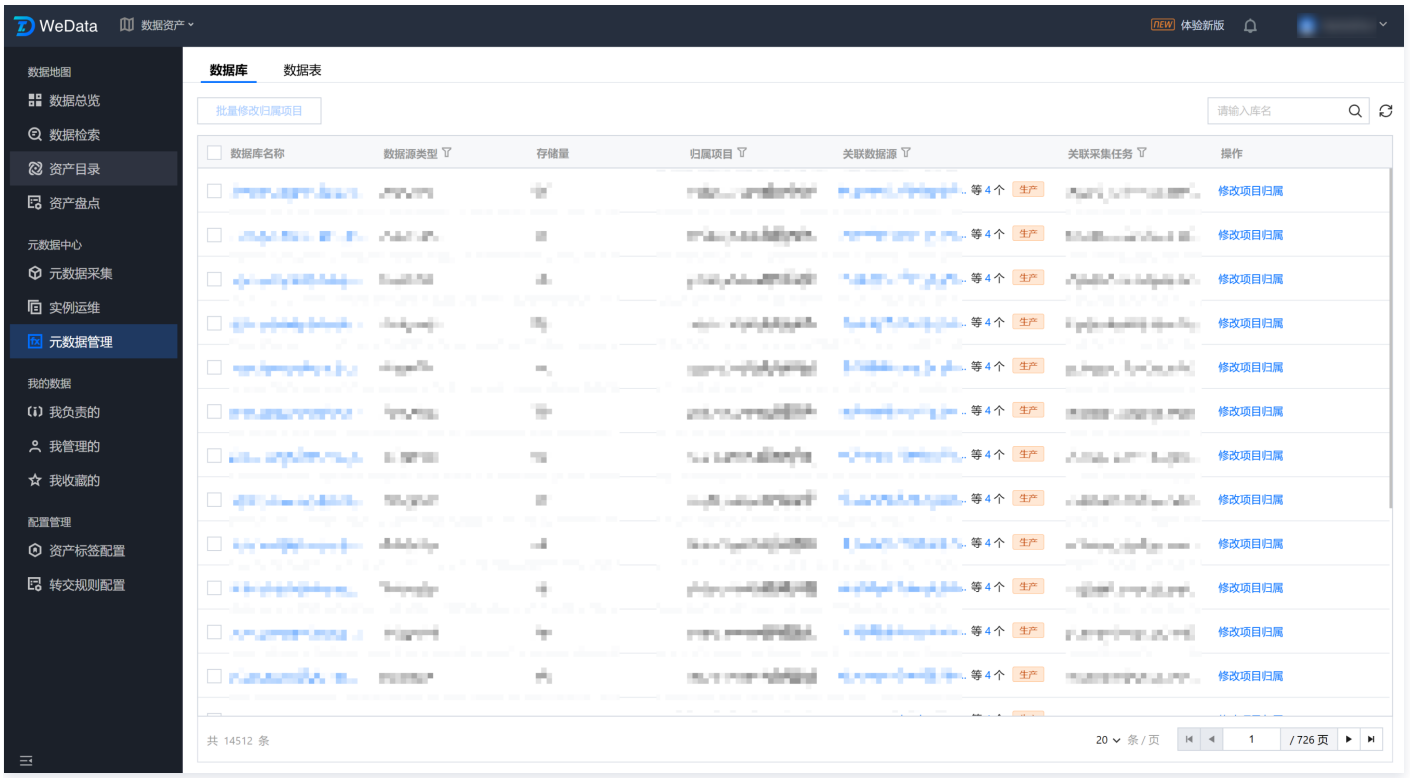
Operation	Provides features to view instance details, logs, stop, and rerun.

©2013–2025 Tencent Cloud. All rights reserved.

Ownership Management Database

Last updated：2024-12-16 16:29:55

Go to **Metadata Management** > **Database,** this page displays a list of databases collected by full metadata, including database name, data source type, associated project, storage volume, related data source, collection tasks, etc., and supports modifying the associated project.

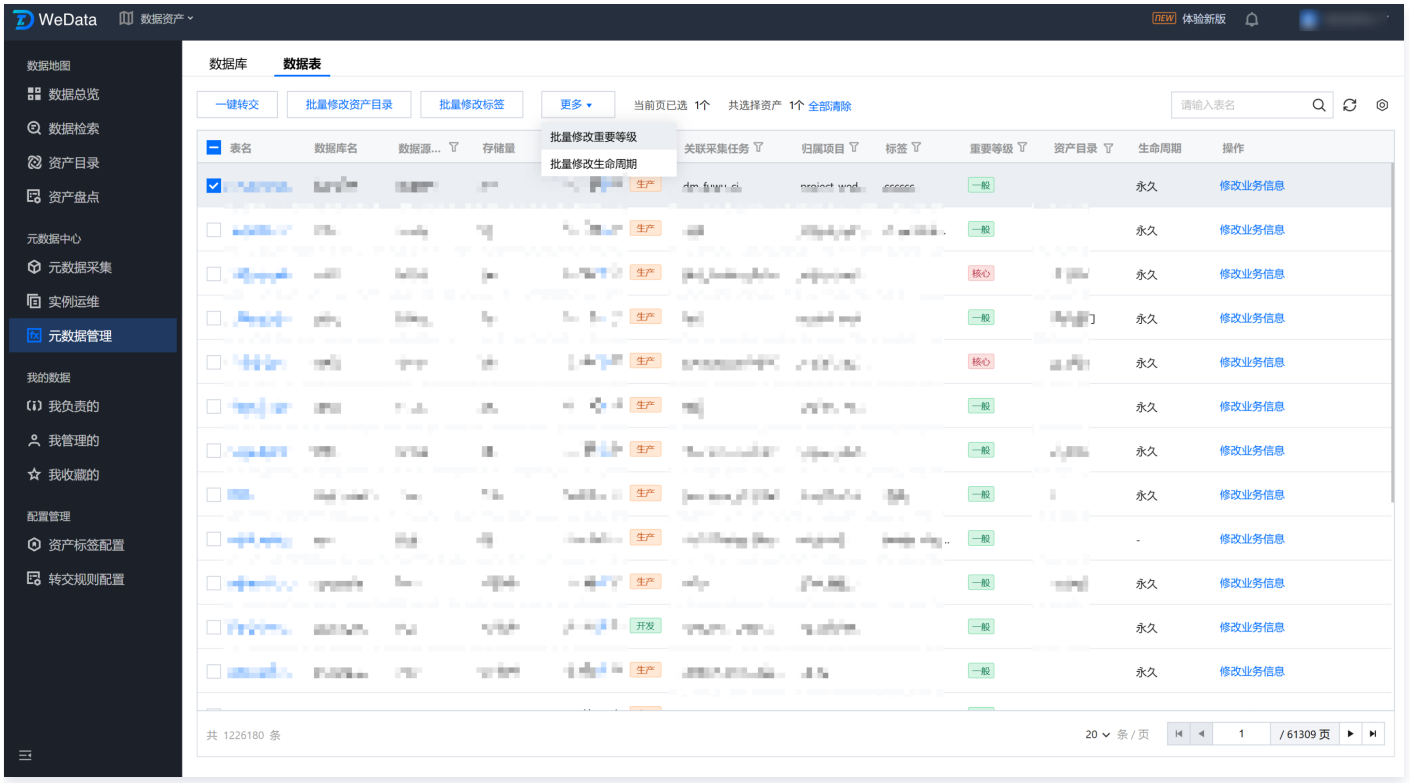


©2013-2025 Tencent Cloud. All rights reserved.

Data Table

Last updated：2024-12-16 16:30:14

Go to **Metadata Management** > **Data Table**. This page displays a list of tables with full metadata collection, including table name, database name, tag, importance level, asset directory, data source type, data source, and associated collection tasks. It supports one-click transfer and modification of business attributes.



©2013-2025 Tencent Cloud. All rights reserved.

Data Quality
Data Quality Overview

Last updated: 2025-04-25 15:18:32

Data quality is one of the core links in data governance. It aims to help users detect dirty data generated during data integration and data development immediately, automatically intercept exception tasks, block the spread and transmission of dirty data to downstream, and reduce user issue handling cost and resource waste.
Applicable Roles: Data Development Engineer, Data Warehouse Table Owner.

Fee Description

The costs generated by data quality task execution mainly include the following three parts:

1. WeData product feature version cost (premise).
2. **WeData Execution Resource Cost:** Charge based on the volume of scheduling resources consumed by Quality Task Instances.
3. Non-WeData direct costs: Quality Task Verification requires the cooperation of engines and data source services (for example, EMR, DLC, TCHouse-D, TCHouse-P, etc.) to execute. This will generate engine fees, which are charged by the engine side and not included in the WeData billing statement. For specific charging standards of each engine, please refer to the billing description in the engine product documentation on the Tencent Cloud official website.

For details, see the Purchase Guide.

Core Capabilities

The Quality Module mainly includes the following core features:

1. Support various Tencent Cloud Big Data Storage Engines (such as EMR, DLC, TCHouse - P, TCHouse - D) and open - source big data storage engines (such as Doris).
2. Configure data quality inspection rules at the table level and field level.
3. Configure execution policies based on actual business scenarios.
4. Set the rule strength to determine whether to block downstream tasks.
5. Support various user reach methods (WeCom group, WeChat, call, SMS, mail, Lark Group, DingTalk group).
6. Quality scoring can be obtained from six dimensions (accuracy, timeliness, integrity, uniqueness, consistency, and validity), and a library and table dimension quality report can be formed.

Features of Each Module

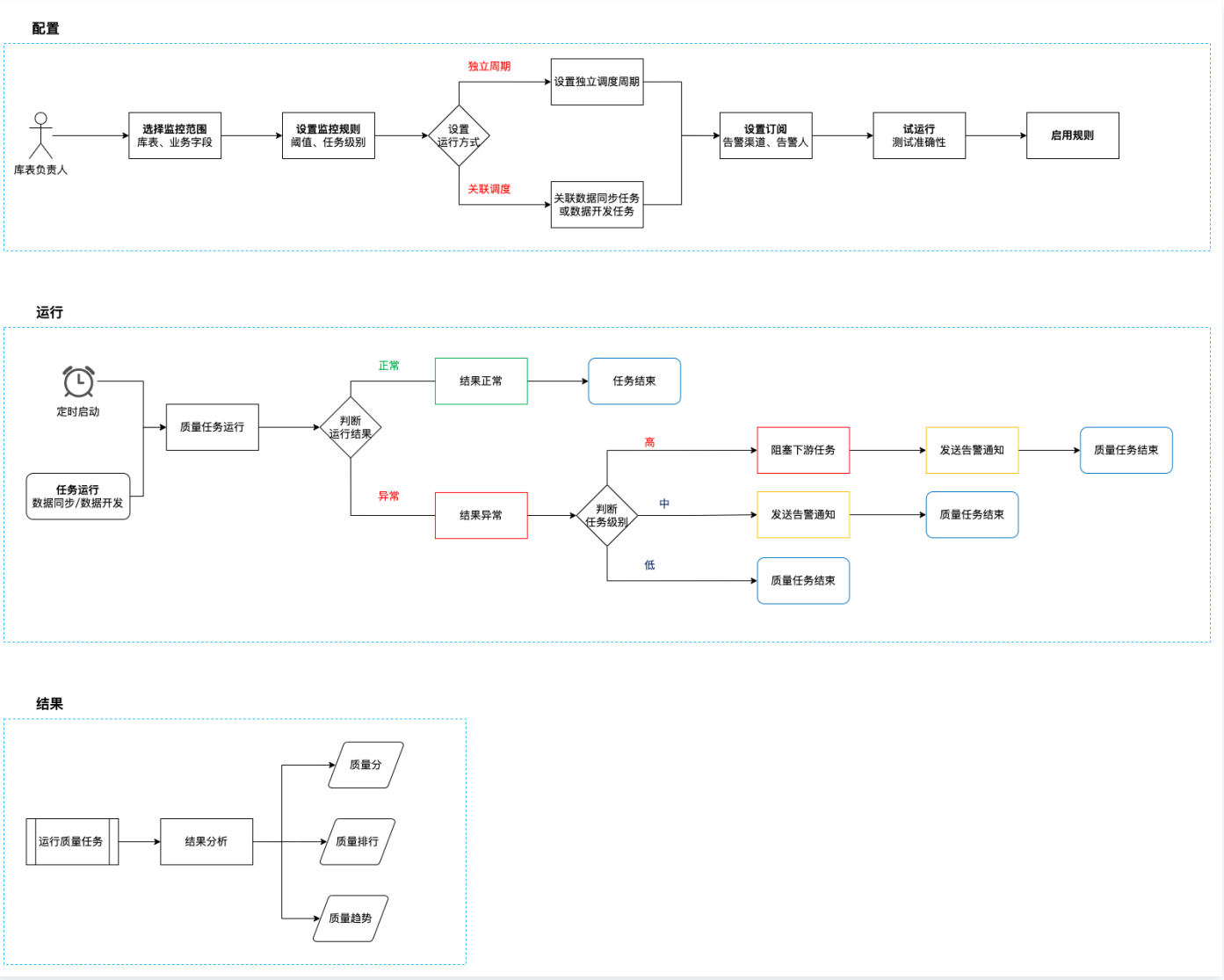
The introduction to the features of each Data Quality module is as follows:

Function	Overview
Quality Overview	**Quality Result Overview:** • view detection status; rule operation status • View alarm status Table alarm ranking
Rule Template	**Unified management of rule templates for unified reuse:** • 56+ system built-in templates: can only be viewed; • Custom rule template: supports CRUD operations.
Data monitoring	**Create detection rules:** • Support various Tencent Cloud big data engines: EMR, DLC, TCHouse - P, TCHouse - D, Doris; • Support various creation methods: single table addition, multiple tables addition, batch upload. **View detection rules:** • Support various viewing methods: view all, table dimension, rule dimension; • Support viewing the rule list of a table and performing rule management.
Ops management	**Execution instance and results:** • Support viewing the task running results of quality; and can view the historical running status of each rule;

©2013-2025 Tencent Cloud. All rights reserved.

	• Support exporting execution results and view historical export logs. **Quality Task:** • Support viewing the generated quality inspection tasks; • Support configuring alarm information for quality tasks. **Alarm Information:** • Support viewing historical alarm situations.
Quality Report	**Quality Report:** • Support counting historical operation results into quality scores in multiple dimensions: database and table, rule dimension. • Support viewing quality scores in multiple dimensions: comprehensive quality score, dimension quality score, Quality Detail Breakdown.

Core Process



Key Term Explanation:

Noun	Explanation
Independent Cycle	Set up periodic quality inspection for selected database tables and core business fields at custom frequencies such as daily, hourly, or by the minute. Quality tasks will be executed on a scheduled basis according to the set period. If an exception is detected, subscribers will be notified immediately.
Associated Scheduling	Associate quality tasks with production tasks (data sync tasks or data development tasks). After the production task execution is complete, insert a quality rule task execution. If an exception is detected,

©2013–2025 Tencent Cloud. All rights reserved.

	notify the handler to handle it immediately. Block downstream task execution according to the task level to avoid problem data expansion.

Must-Knows

- After configuring the configuration tables for EMR, DLC, TCHouse-P, and TCHouse-D, as well as the field data quality rules, the scheduling node for output data requires the use of a scheduling resource group with established network connectivity for execution. It must ensure that the executor is stable and its version has been updated to the latest version before data quality rule validation can be normally triggered.
- Each table can configure multiple table-level and field-level data quality rules and perform verifications simultaneously.

©2013-2025 Tencent Cloud. All rights reserved.

Getting Started with Data Quality

Last updated: 2025-04-25 15:16:52

> **Note**
> This document will take you through a complete configuration process to quickly understand the core capabilities of the data quality module.

Preparations

1. Purchase and activate the WeData product. For details, see WeData Purchase Guide.
2. In the WeData Console, create a project and bind data sources such as EMR, DLC, TCHouse-P, TCHouse-D, Doris (this document uses EMR as an example). For details, see WeData Project Management.
3. In the data development module, create a database and a data table (this document uses EMR as an example). For details, see WeData Data Management.
4. In the data development module, create a data development task. For details, see WeData Task Development Introduction.

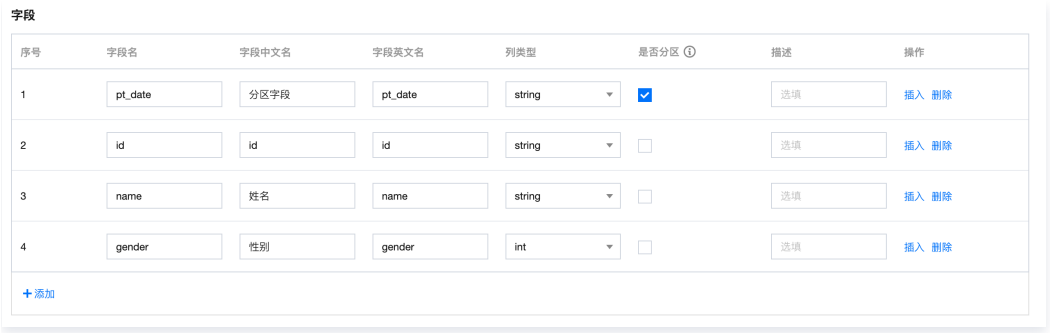
> **Note**
> When creating a project, purchase and bind an EMR engine (enable metadata scanning during binding) and bind scheduling resources.

Data sheet preparation

1. Create a database named "emall" in the emr-hive data source. Create a data table named "dq_test".
2. Add the following fields to the table:

Field Name	Chinese Name	Column Type	Partitioned
pt_date	partition field	string	Yes
id	id	int	No
name	Name	string	No
gender	Gender	int	No

Operation Screenshot:



Data Development Task Preparation

1. In the **Data Development** > **Orchestration Space** interface, click **Create Workflow**.
2. Create a task node in the workflow. Task type: Hive SQL.

   The Hive SQL is as follows:

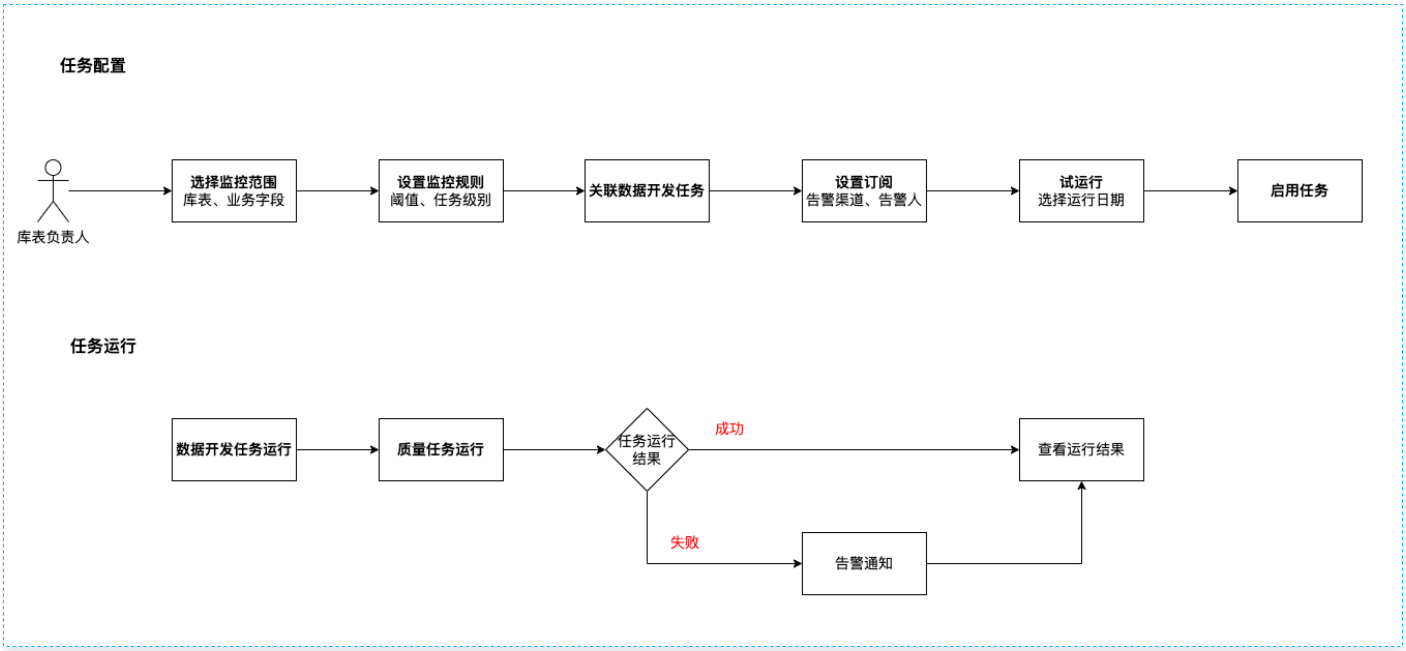
```
INSERT INTO TABLE email.dq_test PARTITION (pt_date = '2024-05-01')
```

©2013-2025 Tencent Cloud. All rights reserved.

```
VALUES ('1', 'Zhang San', '1');
```

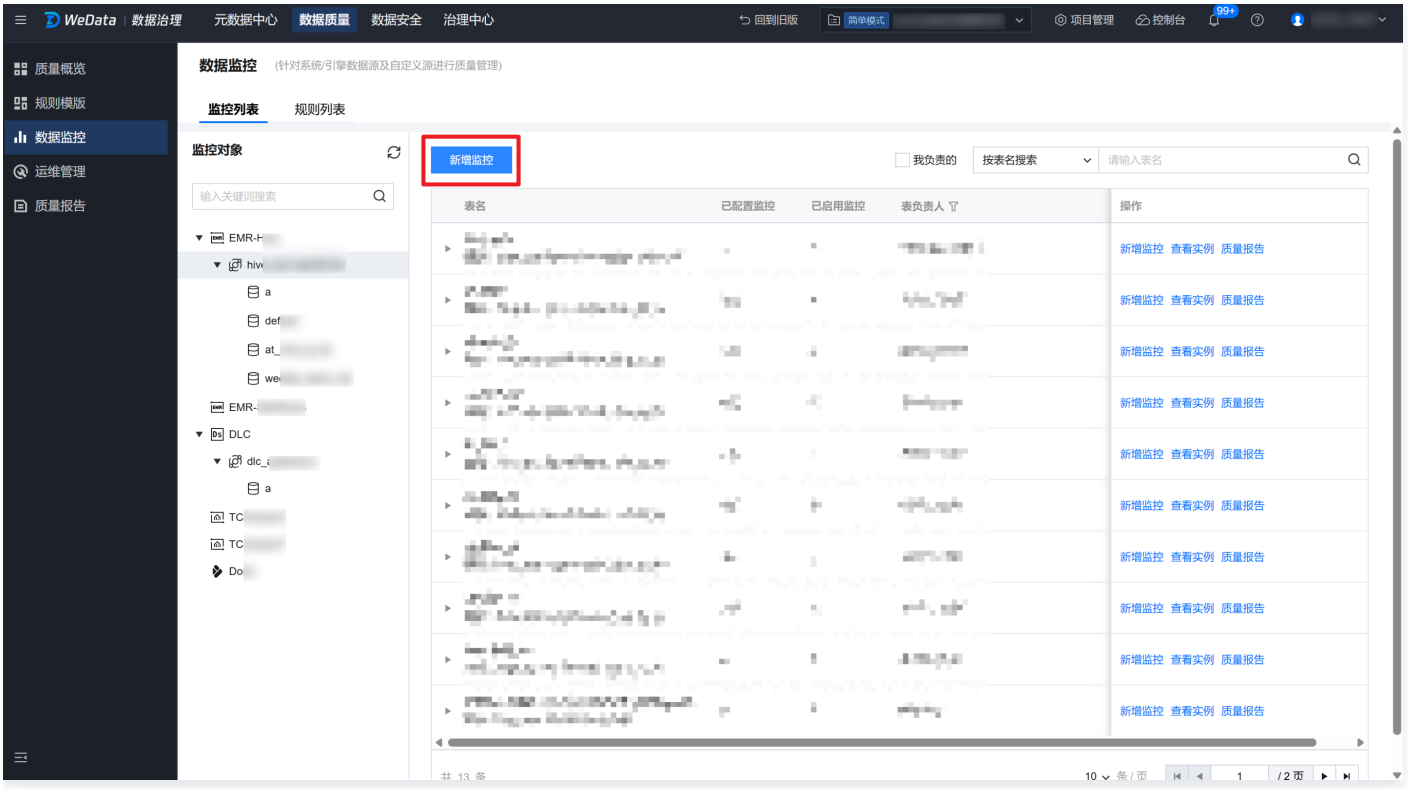
Configure Data Quality Monitoring Task

Configuration Process



Creating a Quality Monitoring Task

Enter the Data Quality > Data Monitoring > EMR interface, click Add Monitoring, and start creating a new quality monitoring.

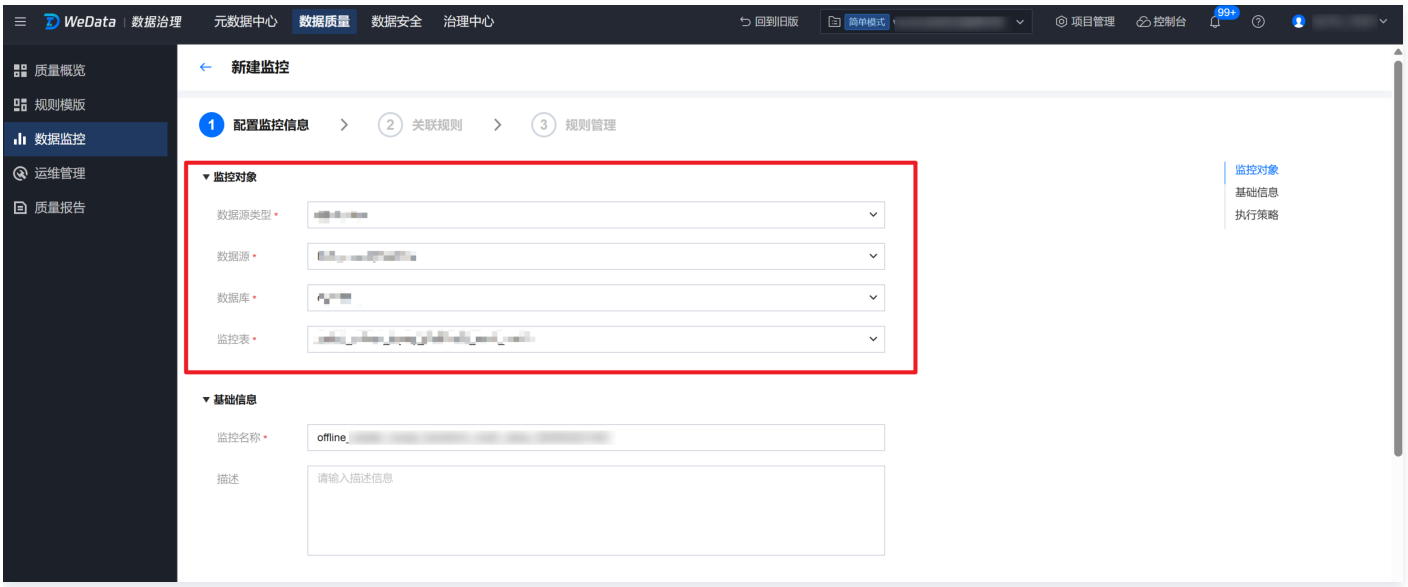


Step 1: Configure Monitoring Information

Monitoring Object

©2013–2025 Tencent Cloud. All rights reserved.

Enter the Create Monitoring interface, sequentially select the data source type, data source, database, and monitoring table of the monitoring object.

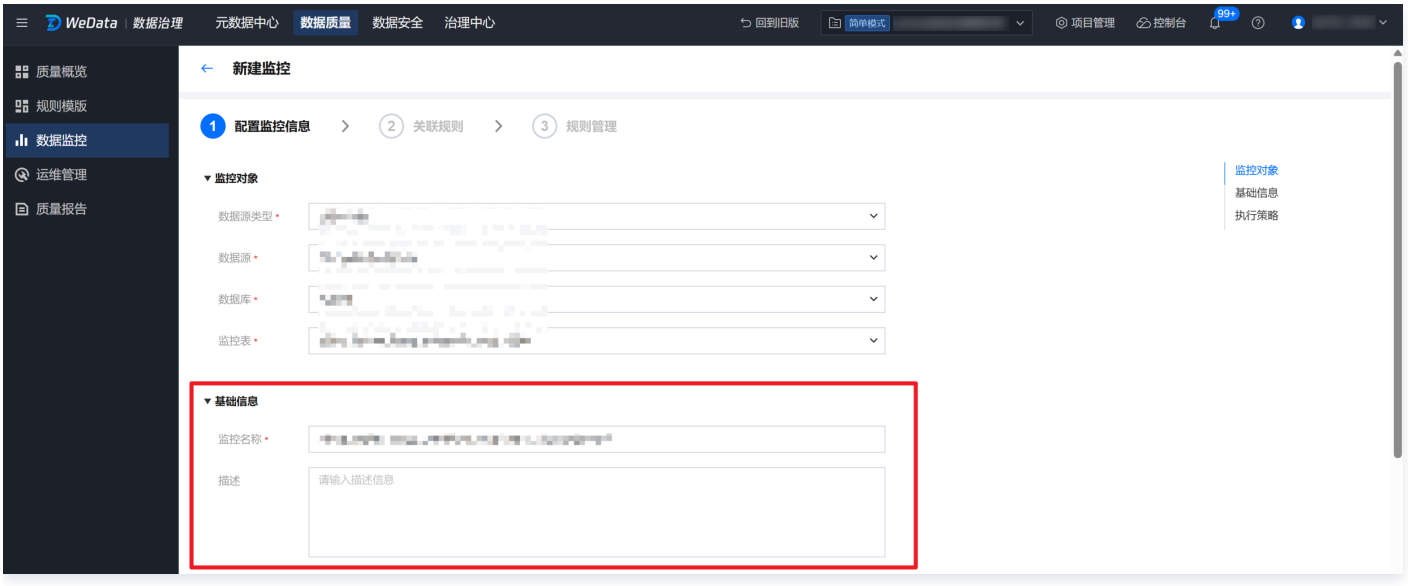


Basic information

Fill in the monitoring name. The default monitoring name is **monitoring table name_Creation monitoring timestamp**.

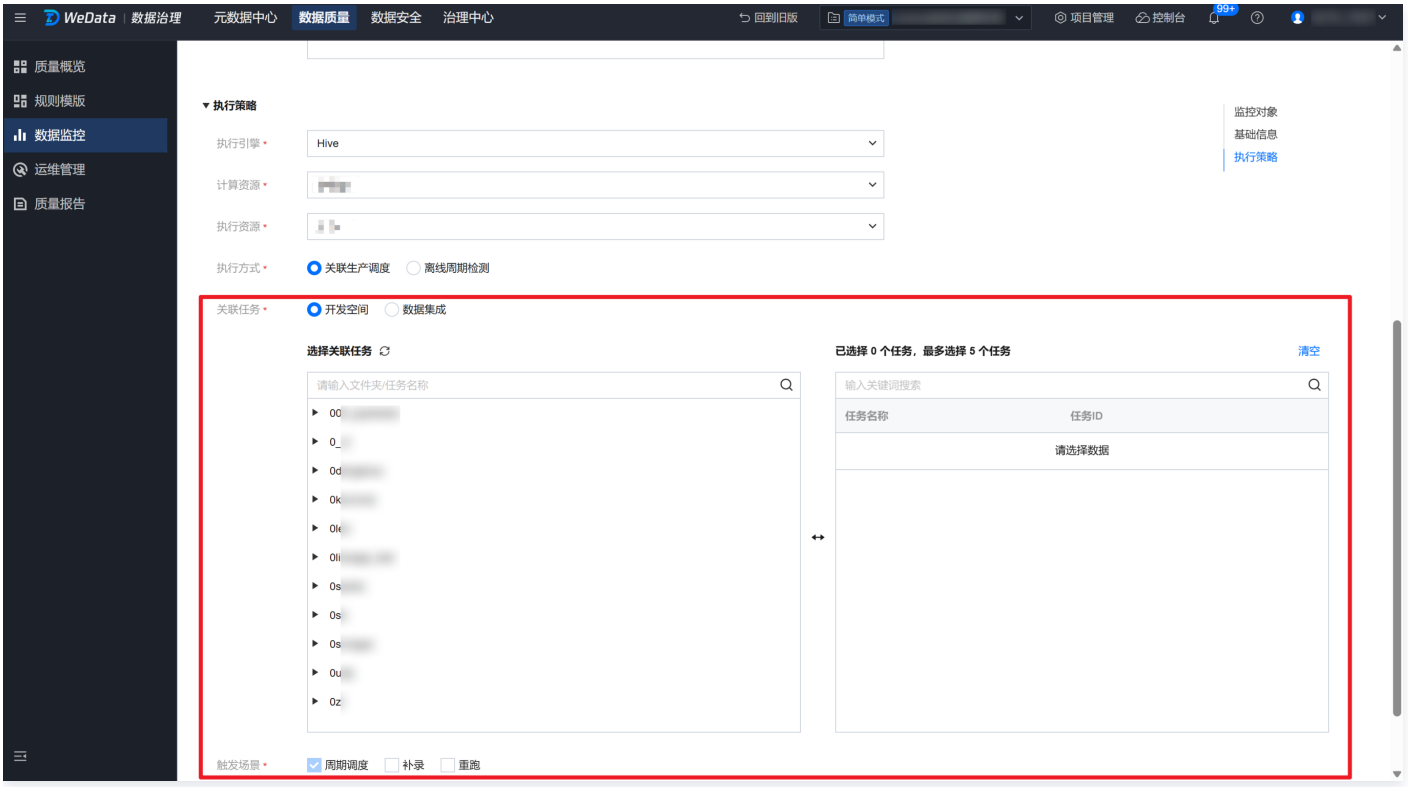
> **Note**
> Note: It is advisable to fill in an easily identifiable monitoring name to facilitate understanding of the business significance of the detection results in the monitoring task instance.
> For example, when the monitoring name is "North China region sales table monthly routine detection", the Ops management will show whether the detection of "North China region sales table monthly routine detection" succeeds or fails each time it runs.

Meanwhile, to facilitate interpretation, you can also add a business description.



Execution Policy

©2013–2025 Tencent Cloud. All rights reserved.



Element	Description
Execution Engine	Here you can select Hive and Spark, which are related to the purchased EMR resources. Generally, Hive tables can directly select the Hive engine.
computational resource	Select default Here you can select the resource group in the EMR cluster. Generally, you can directly select default.
execution resource	The execution resource here is the scheduling resource group that has been bound to the project.
Execution Method	Here you can select **associated production scheduling** and **offline cycle detection**: **Associated Production Scheduling**: Associate quality tasks with production tasks (data sync tasks or data development tasks). After the production task finishes running, insert a quality monitoring task. If an exception is detected, notify the handler immediately to handle it and block downstream task execution according to the rule level to avoid problem data expansion. **Note** The same quality inspection task can be associated with multiple production tasks; the same production task can also be associated with multiple quality inspection tasks. **Offline Cycle Detection**: Refers to independent scheduling. Set up periodic quality inspections for selected database tables and core business fields at self-defined frequencies such as daily, hourly, or by minute. Quality tasks are executed on a scheduled basis according to the set cycle. If an exception is detected, subscribers will be notified immediately.
associated task(s)	Here you can select the module to which the production task that needs to be associated belongs, either the Development Space or Data Integration. You can associate up to 5 tasks: • **Development Space:** Show the directory tree structure in the development space, support searching for folders or task names. • Data Integration: Display the task list in data integration.

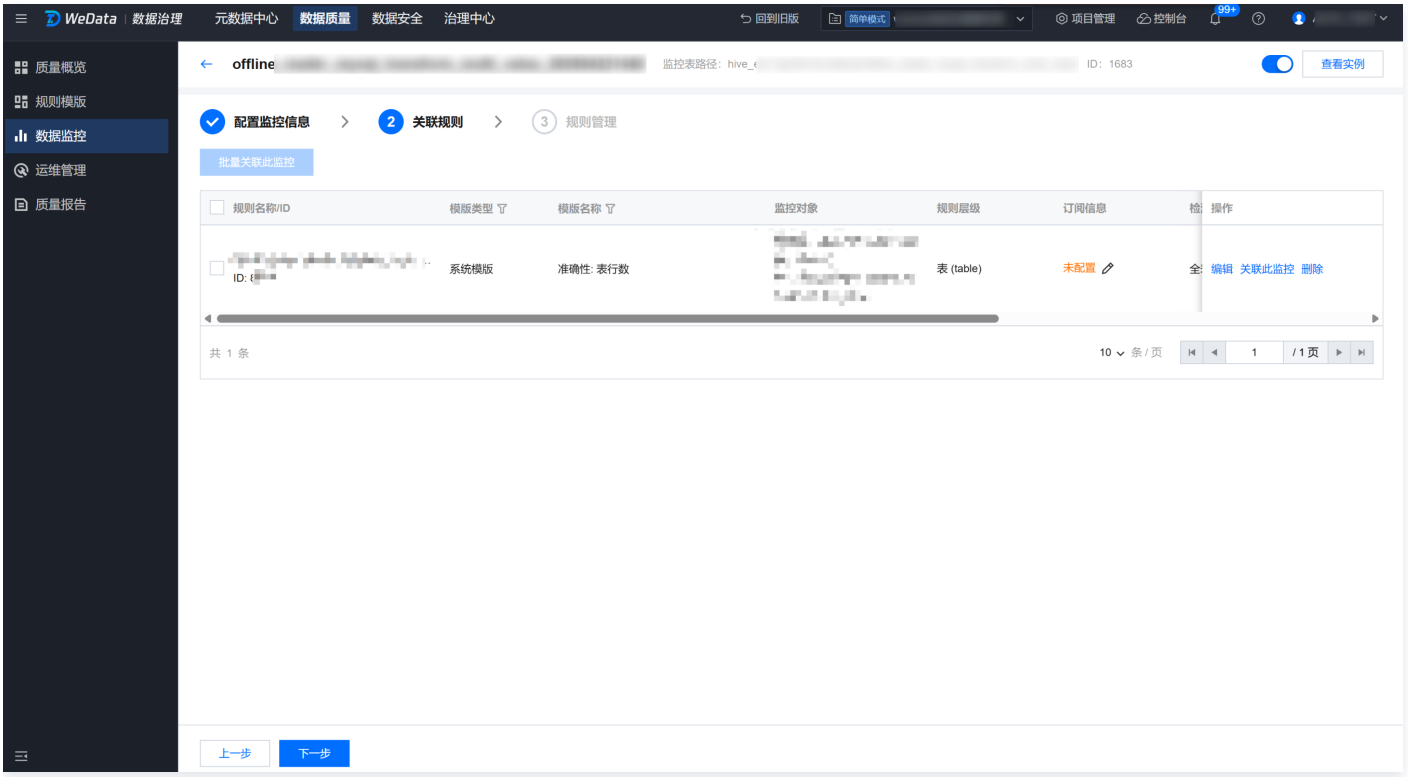
Click Next. After the monitoring is created, go to the Association Rule interface.

Step Two: Association Rule

©2013–2025 Tencent Cloud. All rights reserved.

Rules associated in one table are displayed, which are unassociated with other monitoring tasks. Quality rules can be edited or associated (optional).

> **Notes:**
> Quality rules that are unassociated with monitoring will not execute.

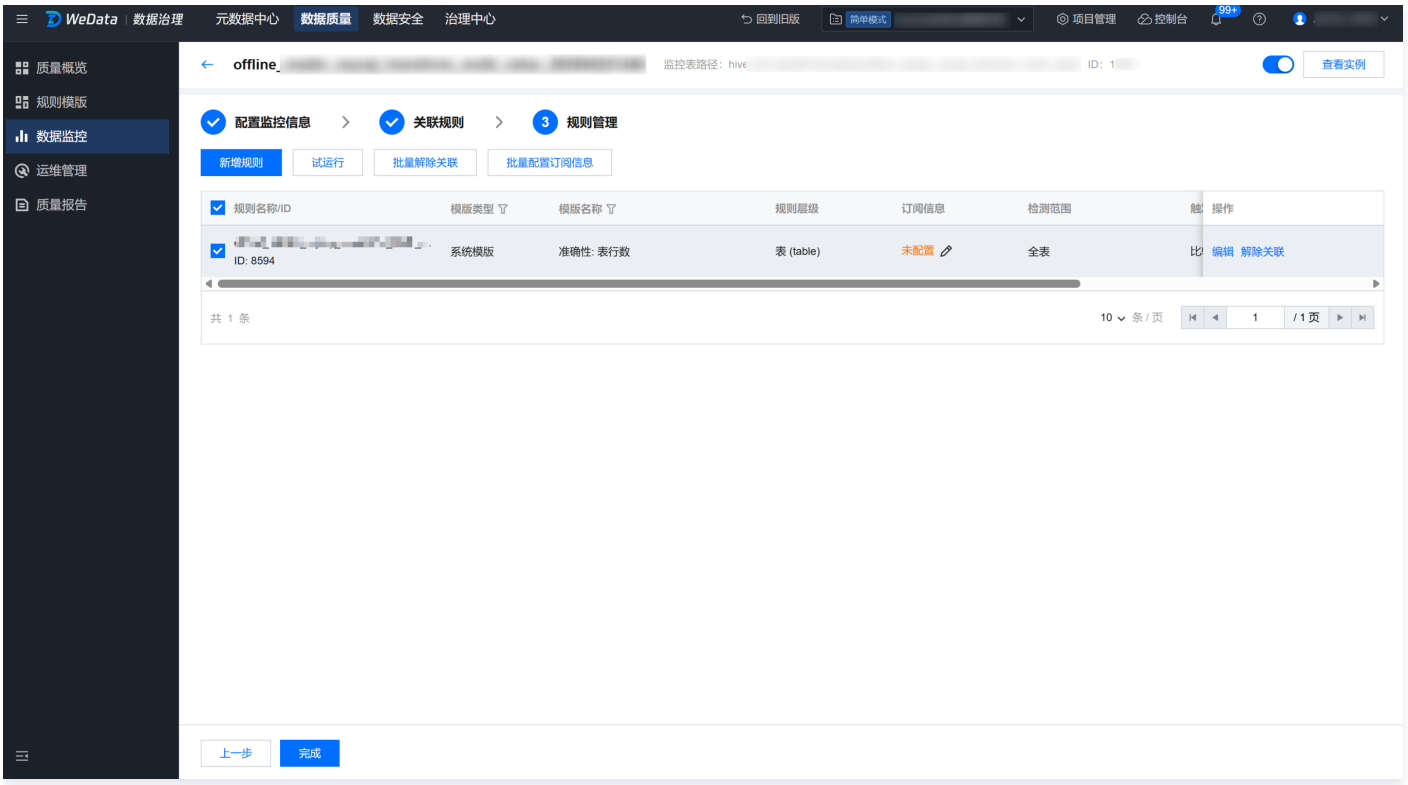


If existing rules cannot meet business needs, click Next to access the rule management interface.

Step Three: Rule Management

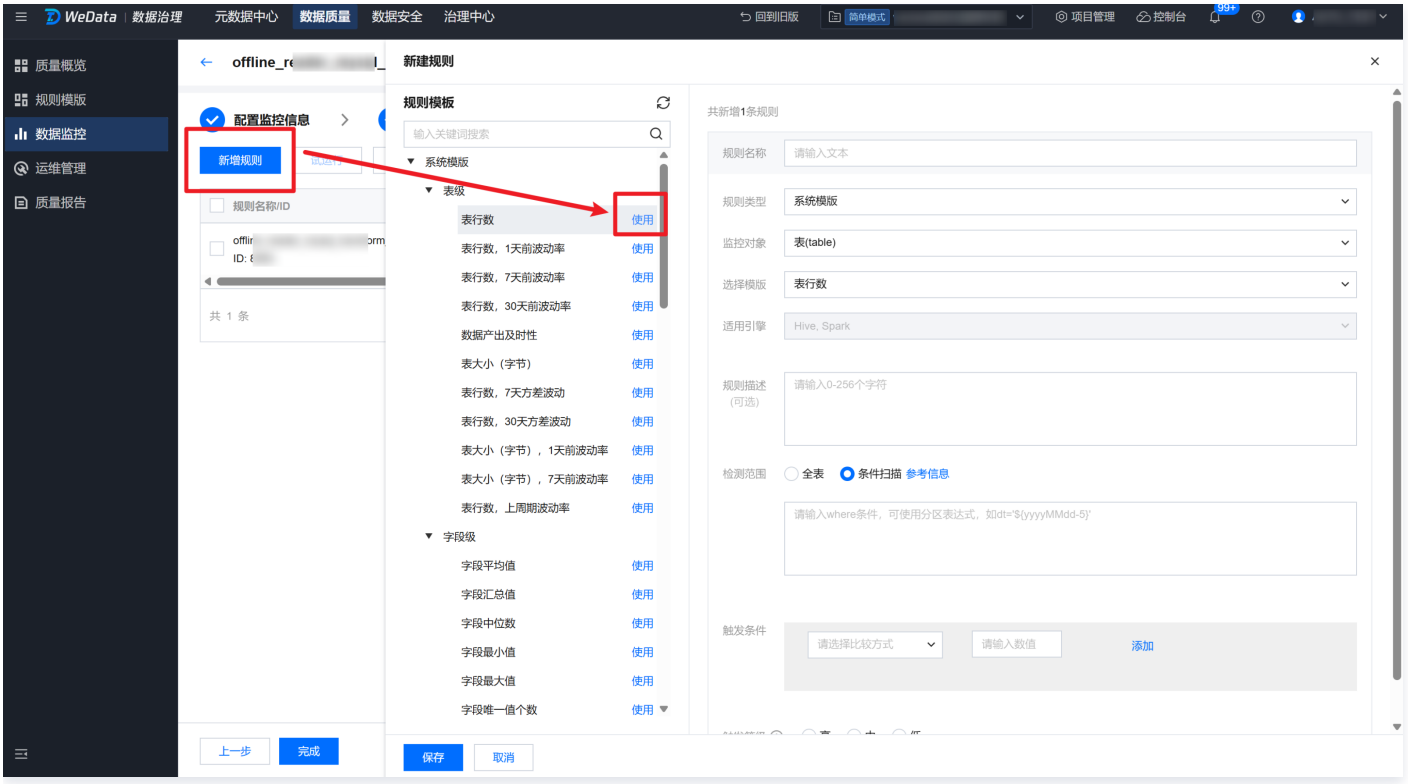
Rule management shows all rules associated with this monitoring. All previously associated quality rules will be echoed here.

©2013–2025 Tencent Cloud. All rights reserved.



Creating New Rule

Click Create New Rule in the rule list. A pop-up for creating a new rule will appear, supporting selection of 1. System Template, 2. Custom Template, 3. Custom SQL for quality verification.
One-time addition of multiple rules is supported here. Newly added rules will be directly associated with this monitoring.



Description:

Element	Note:

©2013-2025 Tencent Cloud. All rights reserved.

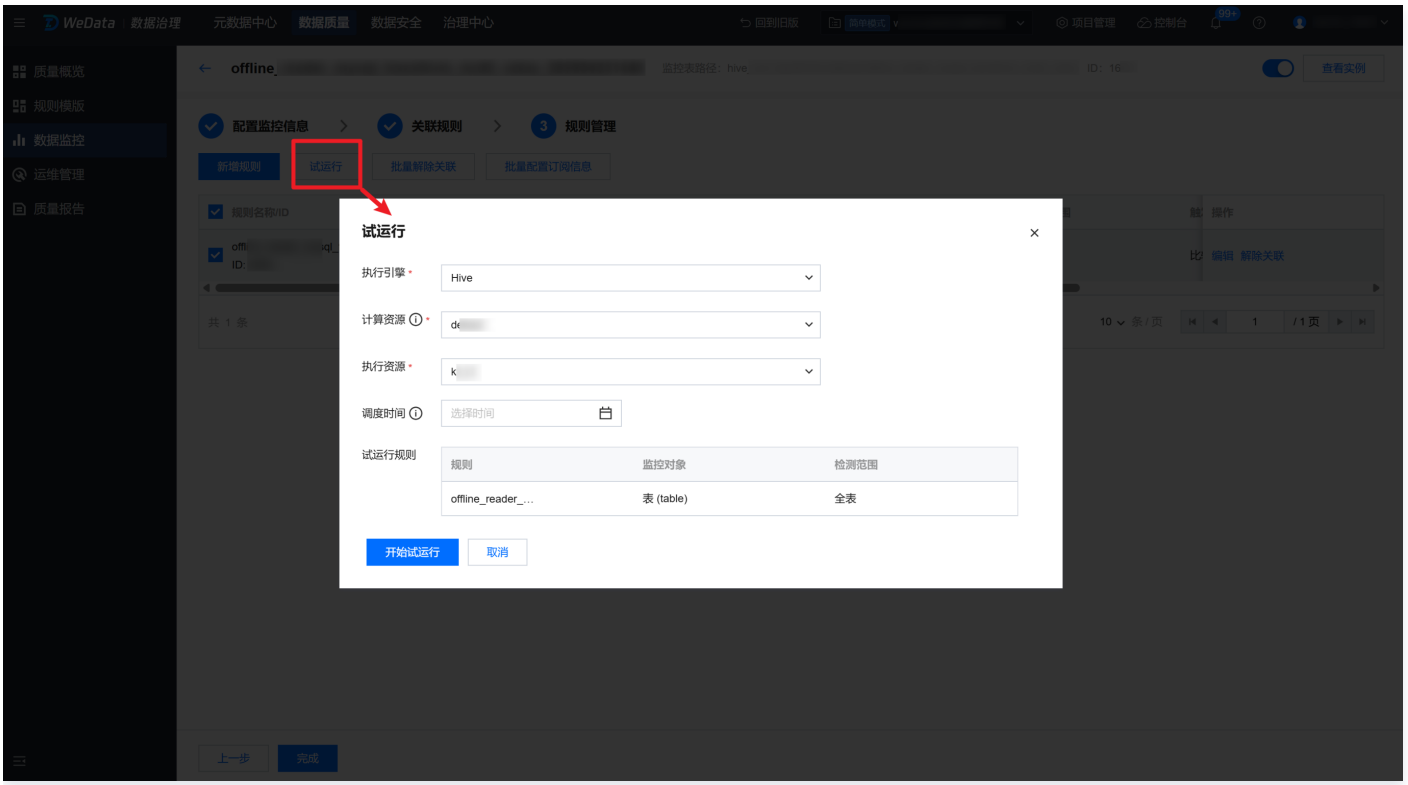
Rule Type	Selectable system template, custom template, and custom SQL. (If a rule template is selected from the left tree, the selected template parameters will be displayed by default here.) • System Template: WeData has 56 built-in rule templates that can be used for free. For more details on each template, refer to system template description. • Custom template: You can add rules applicable to your own business in the rule template menu for easy reuse. For detailed operation guidance, see Custom Template Description. • Custom SQL: You can directly fill in SQL statements as detection rules. For detailed operation guide, refer to Adding Quality Rules.
Monitoring Object	Monitored objects can be divided into table-level and field-level. (If a rule template is selected from the left-side dendrogram, the selected template parameters will be displayed by default here.) • Table level: You can monitor the number of rows and the table size (only supported for Hive tables). • Field level: You can monitor whether the field is empty, whether to repeat, average value, maximum value, minimum value, etc.
Template selection	WeData has 56 built-in rule templates that can be used for free. For more details on each template, refer to system template description. If a rule template is selected from the left dendrogram, the selected template parameters will be displayed by default here.
Detection scope	You can choose **full table** or **scanning with conditions**. • **Full table:** Quality rules will validate the full data under the table. • **Condition scan:** Quality rules will only validate the detection range filled in here. Example: `pt_date='${yyyy-MM-dd-1d}'` **Note** Note: Generally fill in the partition field here to avoid a full table scan during each quality task, which would result in computational resource wastage. In SQL, ***${yyyy-MM-dd-1d}*** is a date variable representing one day before the execution date and will be replaced with a specific date during Quality Task Execution. For example, when the quality task is executed at 2024-05-02 00:00:00, ${yyyy-MM-dd-1d} will be replaced with 2024-05-01. The replacement logic of specific time variables can be found in time parameter explanation.
Trigger Condition	• Comparison Operator: Select **less than.** • Compare value: Fill in 1. • - The number of table rows is less than 1. Combine with the time variable filled in combination with the detection range, which indicates that when there is no new data added yesterday, an alarm is triggered. **Note** Note: The trigger condition filled here is **abnormal value**, that is, the conditions for triggering alarms.
trigger level	Select **Medium**. Trigger levels can be divided into: High, Medium, Low. • High: Immediately block downstream task execution when an alarm is triggered (valid only when associated with production tasks). • Medium: Trigger alarm only. • Low: Do not trigger an alarm, only show abnormal result display.

Trial Run

1. Select the rules you wish to trial run, and click **Trial Run**.
2. Modify scheduling time, and click **Start Trial Run**. After the trial run ends, you can click **Click to View Operation Results**.

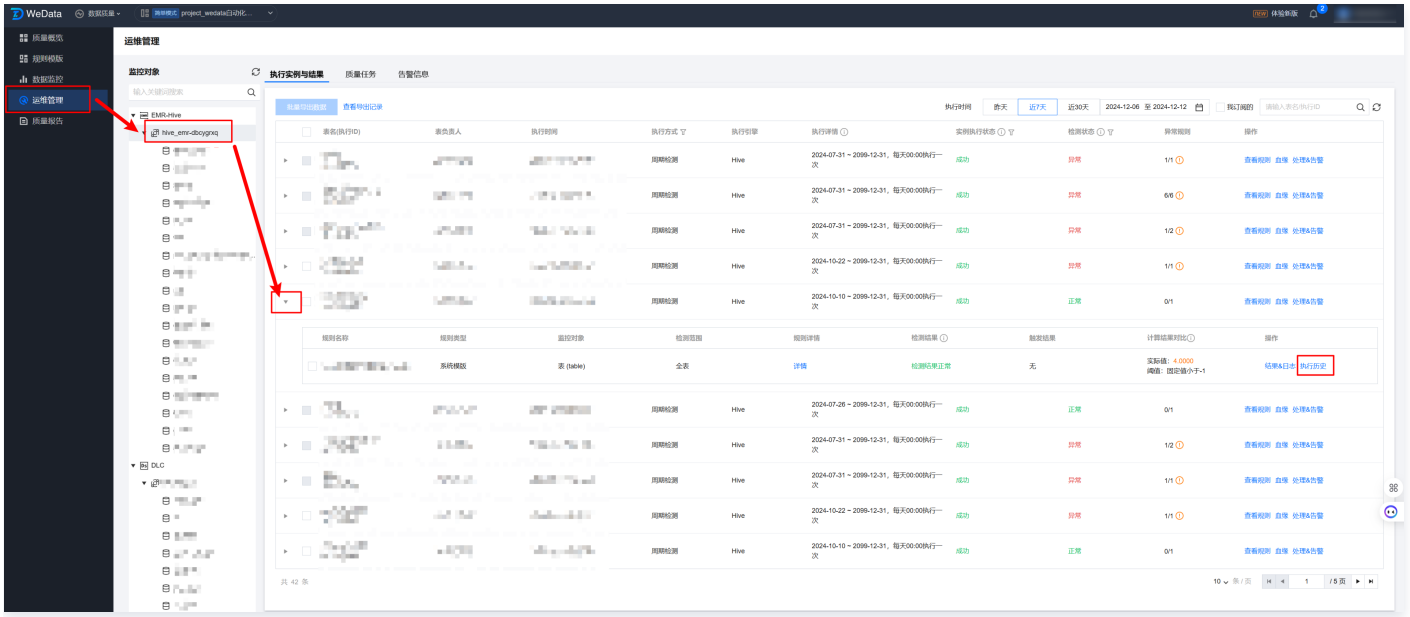
©2013-2025 Tencent Cloud. All rights reserved.

> **Note**
> Since data partition is filled as 2024-05-01 when inserting data, 2024-05-02 can be filled here.



View Trial Run Results

1. After entering the execution instance and result page, you can click on the left dropdown triangle to expand rule details.
2. Click the **Execution History** of a certain monitoring item to view historical run results.



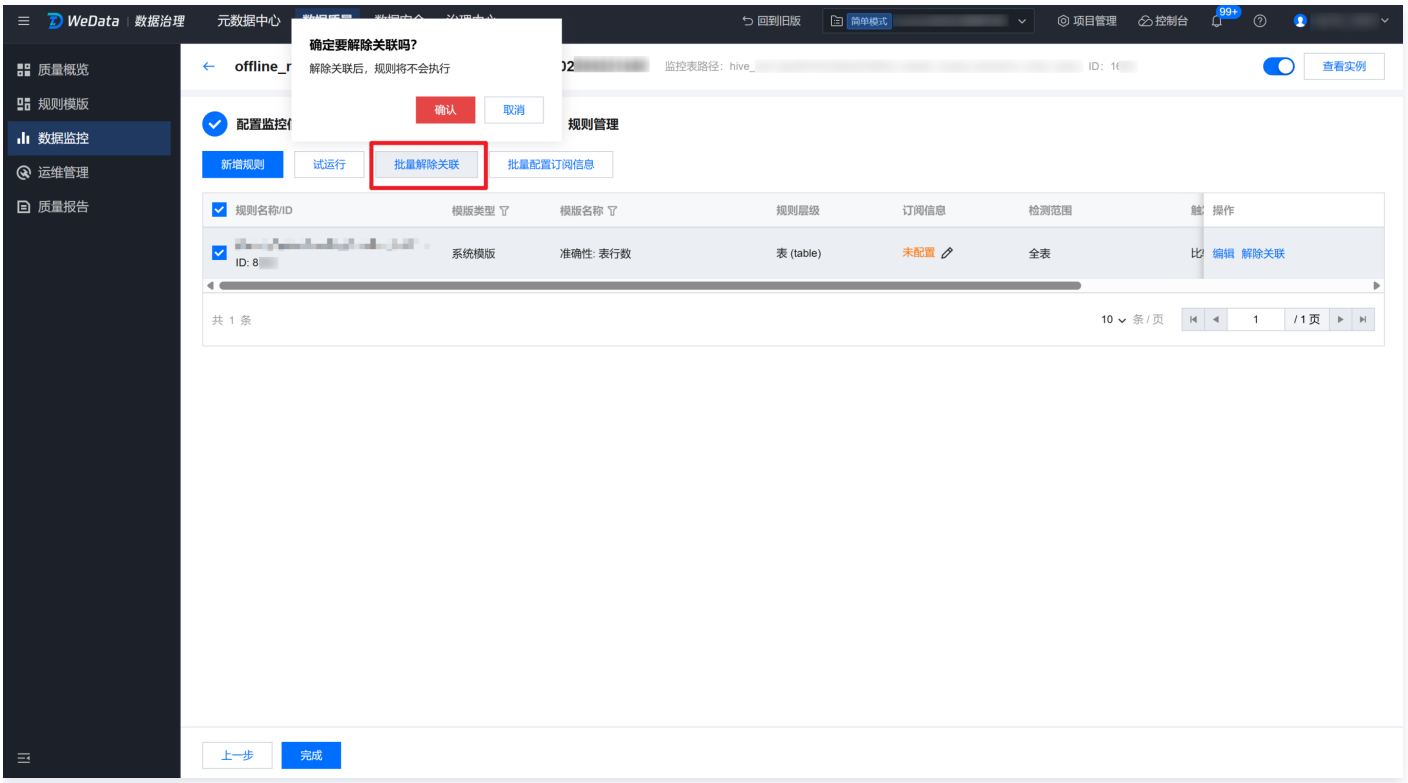
Batch Disassociation

Batch unselect the correlation between the selected rules and the monitor.

> **Notes:**
> - Quality rules that are unassociated with monitoring will not execute.

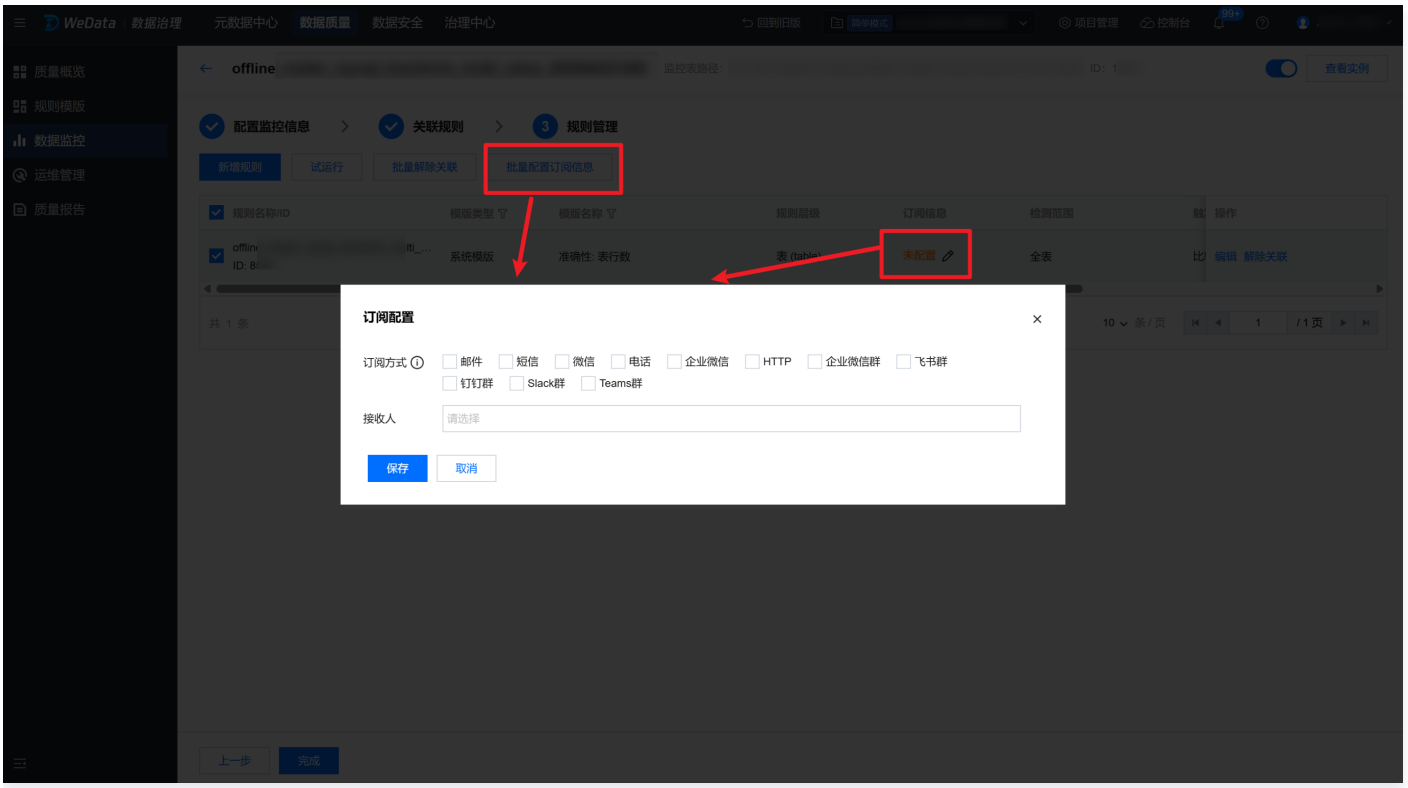
©2013-2025 Tencent Cloud. All rights reserved.

- To reassociate this rule, you can perform operations on the **Association Rule** page.



Configure Subscription

When the task running result of quality detection meets trigger conditions (i.e., the result is not as expected and the detection result is failed), what method to use to send notifications and whom to notify.
Enter the Subscription Settings interface, select the Subscription Method, set the Recipient, and click Save.

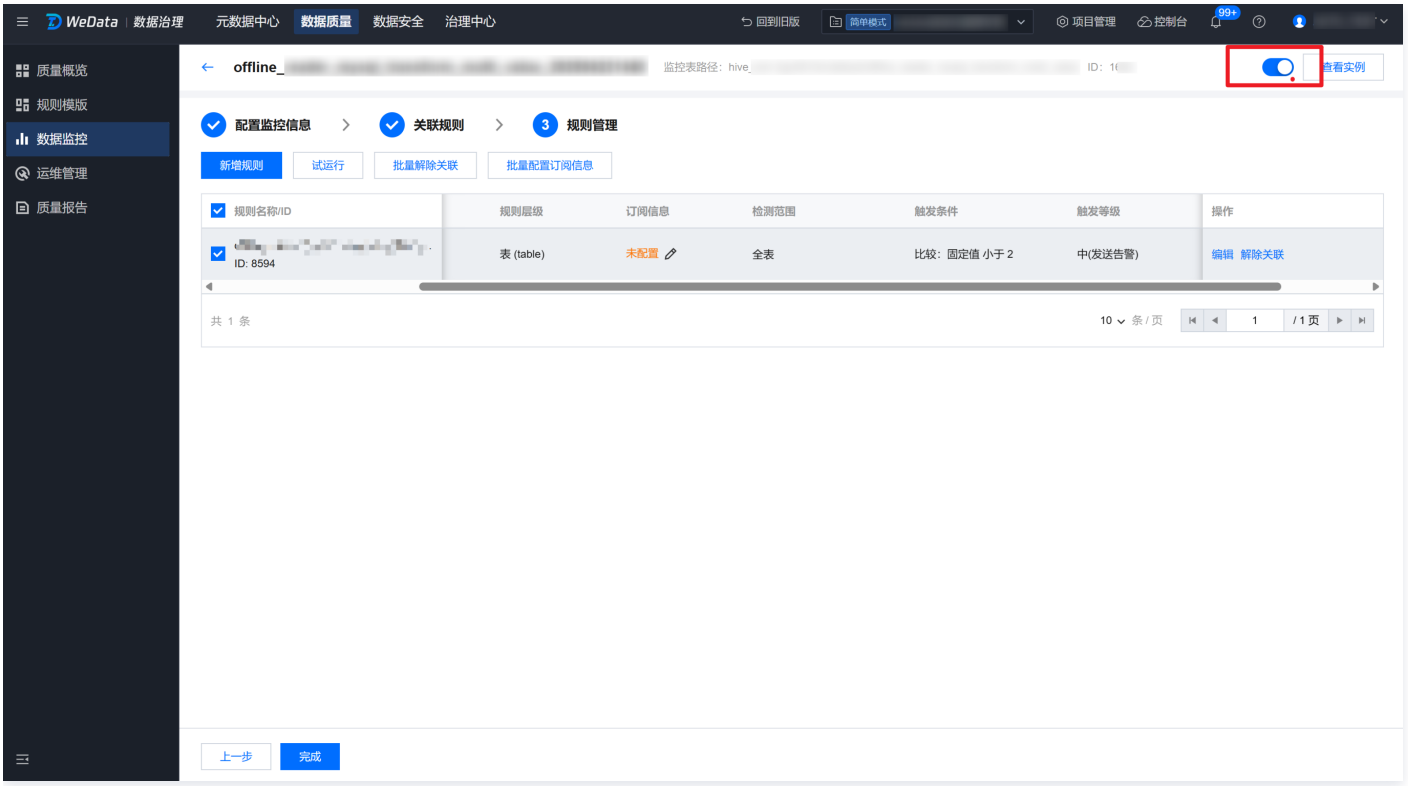


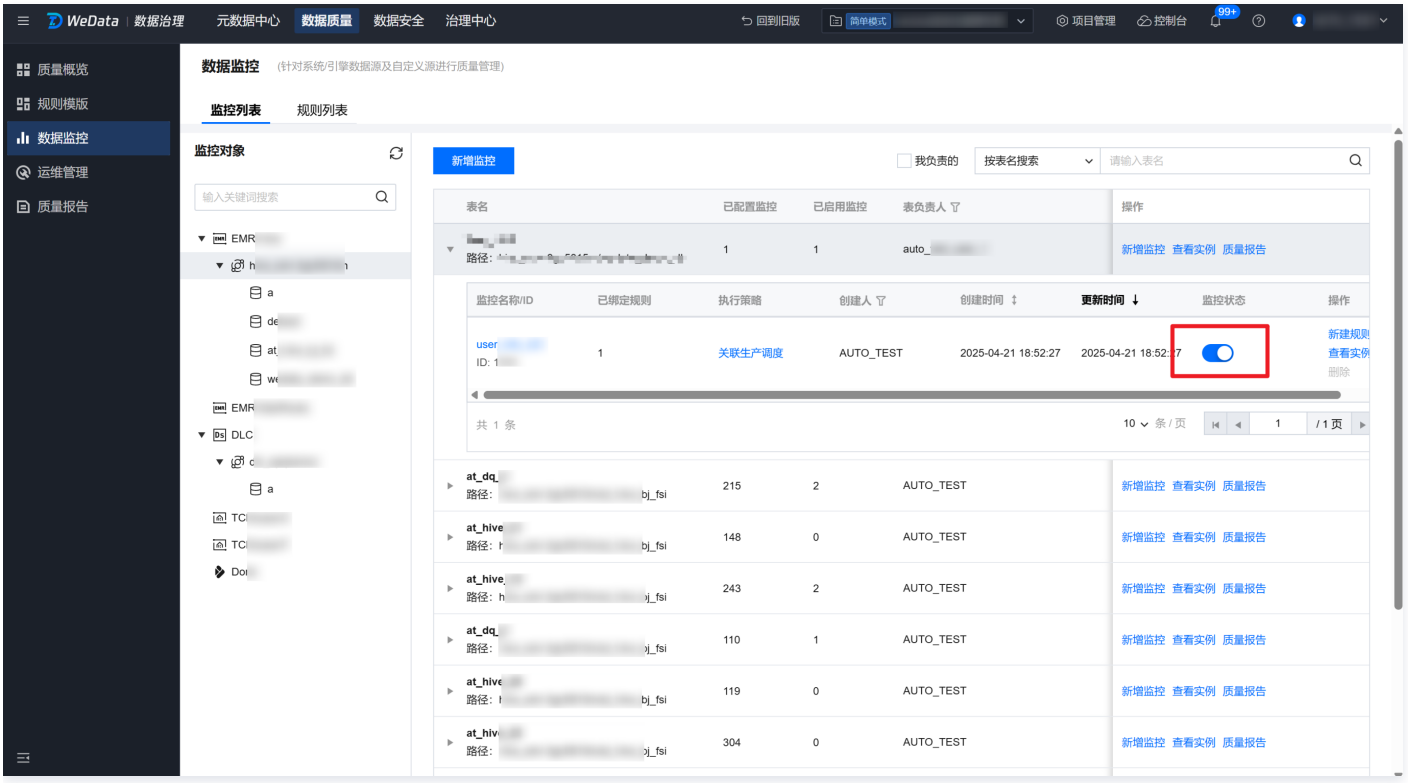
Enable Monitoring

©2013–2025 Tencent Cloud. All rights reserved.

According to the trial run result of monitoring, you can enable or disable monitoring in the monitoring status column **enable** or **disable monitoring** (monitoring is enabled by default after creation) .

> **Note**
> Only monitoring tasks in the **enabled** state will automatically run.





View Task Execution Result

©2013–2025 Tencent Cloud. All rights reserved.

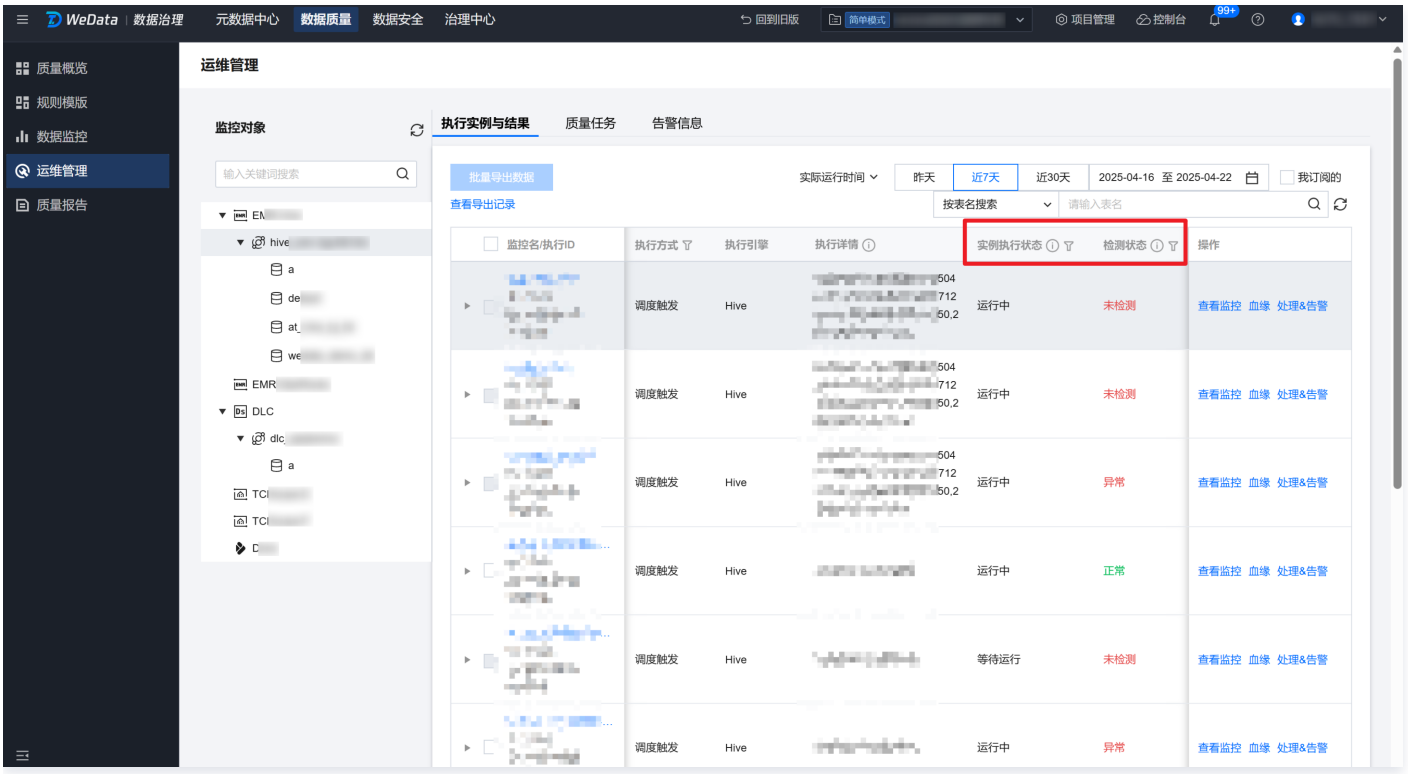
When the quality task execution is complete, if the detection result is abnormal, the system will automatically send an alarm and block downstream tasks for high-trigger-level tasks.
For database table owners and data development engineers, promptly troubleshoot exceptions and solve problems.

1. On the Data Quality > Ops Management > Execution Instance and Results page, select data sources in sequence and then the database to view task execution results.
2. On this page, you can view the task detection status. Expand through the left dropdown triangle to view the results and logs of each rule.

Note

Detection status is divided into two categories:

- **– Instance execution status:**
  - Note: Waiting to run: The scheduling task instance is waiting to run on the scheduling side.
  - – Executing: The scheduling task instance is in execution.
  - – Succeed: The scheduling task instance is distributed successfully.
  - Failure: Scheduling task instance distribution failed. Possible causes: impact of the execution machine environment, manual user termination, business exception reporting.
- **– Detection status**
  - – Normal: After all rules under the current instance are executed successfully, the results of rule detection methods are all in normal status.
  - – Abnormal: If any rule under the current instance is triggered, it is deemed abnormal.
  - – In-progress detection: Rules under the instance are being executed.
  - Note: Not detected: The rule has not yet started execution or execution was unsuccessful.



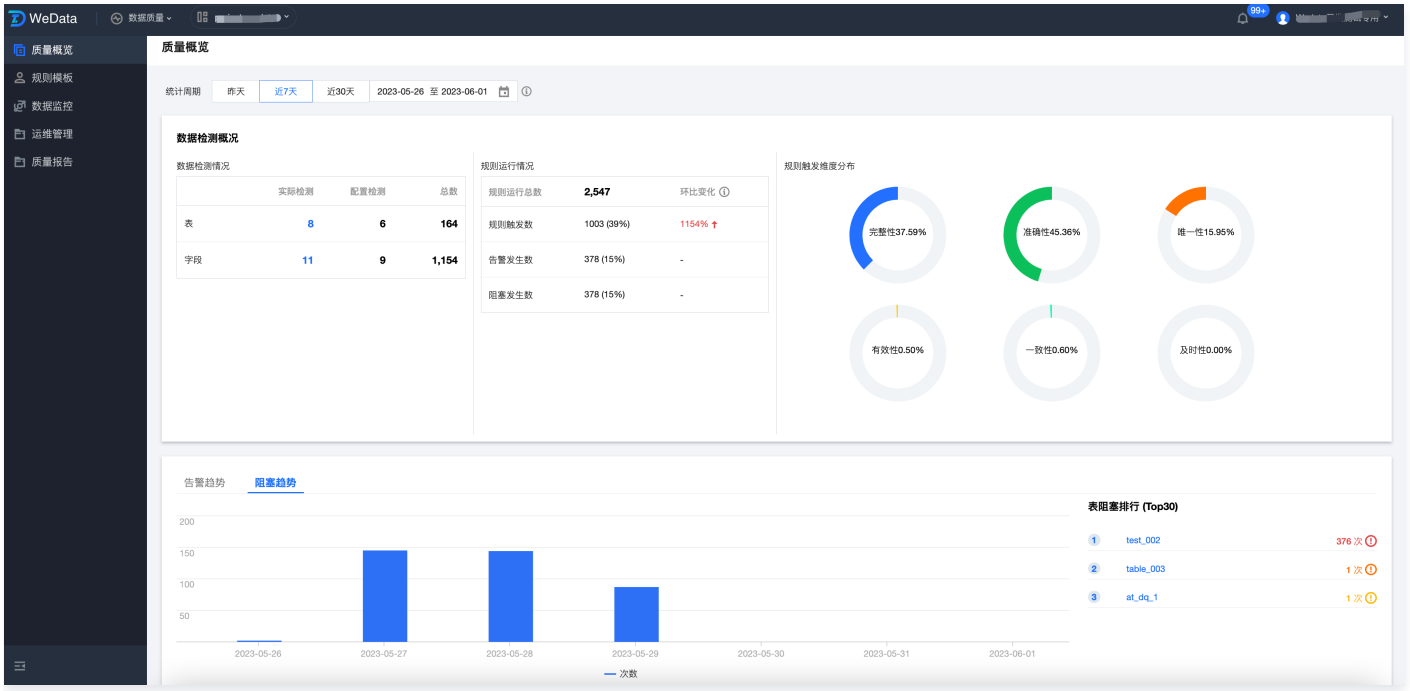
©2013–2025 Tencent Cloud. All rights reserved.

Viewing Quality Overview

Last updated：2024-12-16 16:36:41

Feature Details

The quality overview primarily displays various data indicators and alarms within the selected statistical period for quality inspection results, blocking trends, helping users to grasp the overall data quality situation, gain insight into periodic changes in data quality, and quickly pinpoint issues.



Functional Details	Details
Data Inspection Status	Displays the actual number of detections for tables (deduplicated) and fields (deduplicated), the number of configured detections, total number of all tables, and total number of fields within the statistical period.
Rule Operating Status	Displays the number of rule executions, rule triggers, alarm occurrences, and blocking times within the statistical period and their proportion and month-over-month change rate.
Rule Trigger Dimensional Distribution	Among all tasks with anomalies detected within the period, displays the proportion of each dimension (number of various anomaly rules/total number of anomaly rules).
Alarm Trend	Trend graph of alarm occurrences within the period (by day).
Table Alarm Ranking	Top 30 tables with the most alarm occurrences within the period.
Blocking Trend	Trend graph of blocking occurrences within the period (by day).
Table Blocking Ranking	Top 30 tables with the most blocking occurrences within the period.

Notes

The current statistical period only supports a range of two months before/after the selected date.

©2013-2025 Tencent Cloud. All rights reserved.

Manage Rule Templates
Feature Overview

Last updated: 2024-12-16 16:52:56

Feature Overview

The Quality Rule Template is an important means to improve the efficiency of creating and maintaining quality rules. In this module, you can:

1. View the system **built-in templates**.
2. Manage **custom Definition templates**.

Core Capabilities

View system built-in templates

View 56+ system built-in templates.

Manage templates from custom definitions

1. Users can set commonly used detection logic as templates and use these templates to create rules.
2. Subsequently, by modifying the details of the template, you can batch modify all quality rules that use this template.

©2013-2025 Tencent Cloud. All rights reserved.

View system built-in templates

Last updated: 2024-12-16 16:53:40

The system has built-in 56 rule templates that can be used directly. Please read the use cases of each template carefully before use.

View Template List

You can view the system template list on the rule template management page.
Users can filter queries based on template names, description keywords, types, dimensions, and applicable engines. Users can also create and manage templates in bulk on the Custom Template interface.



Field	Details
Template Type	Currently, two types of templates are supported: table-level and field-level. Filtering is supported
Template Name	Template Naming
Template Description	A detailed description of the specific execution logic and formulas of the template rules
Dimension	Accuracy, Uniqueness, Integrity, Consistency, Timeliness, Validity, filtering supported
Applicable Engines	The engine types applicable to this template. Currently, Hive, Spark, DLC, TCHouse-D, and Doris types are supported. Filtering is supported
Reference count	The number of rules that the current template is referenced and associated with. Filtering is supported

Template Distribution

Monitoring obj	Rule dimension	Compute Item	Calculation Subitem	Description	Numeric Type		Numeric - Volatility Type				Numeric - Standard Score Type		Others			
					Fi	Va	Pr	1d	7d	30	7d	30	Em	For	En	Val

©2013-2025 Tencent Cloud. All rights reserved.

ect					xedValue	lueRange	eviousCycle	ayago	aysago	daysago	ays	days	pty/Uniqueue/Duplicate	matmatching	umeratedrange	uesize
Table-level	Accuracy	Number of table rows		Calculate the number of data rows	✓	–	✓	✓	✓	✓	✓	✓	–	–	–	–
		Table size (bytes)		Calculate the size of the data table (only supports Hive tables)	✓	–	–	✓	✓	–	–	–	–	–	–	–
	Timeliness	Data output timeliness		Calculate the number of data rows. If the row count is 0, it is considered no data output	✓ = 0	–	–	–	–	–	–	–	–	–	–	–
Field-level	Accuracy	Field value	Average value	Calculate the average value	✓	–	–	✓	✓	✓	✓	✓	–	–	–	–
			Summed value	Calculate the summed value	✓	–	–	✓	✓	✓	✓	✓	–	–	–	–
			Median	Calculate the median value	✓	–	–	✓	✓	✓	✓	✓	–	–	–	–
			Minimum value	Calculate the minimum value	✓	–	–	✓	✓	✓	✓	✓	–	–	–	–
			Maximum value	Calculate the maximum value	✓	–	–	✓	✓	✓	✓	✓	–	–	–	–
	Uniqueness	Field's unique value	Count of unique values	Validate unique values	–	–	–	–	–	–	–	–	✓	–	–	–
			Count of unique values / Total Rows		–	–	–	–	–	–	–	–	✓	–	–	–
		Field's duplicate	Count of duplica	Validate duplicate values	–	–	–	–	–	–	–	–	✓	–	–	–

©2013–2025 Tencent Cloud. All rights reserved.

	values	te values													
		Count of duplicate values / Total Rows		–	–	–	–	–	–	–	–	✓	–	–	–
Completeness	Field null values	Count of null values	Validation control	–	–	–	–	–	–	–	–	✓	–	–	–
		Count of null values / Total Rows		–	–	–	–	–	–	–	–	✓	–	–	–
Validity	Mobile number format	Count of invalid entries	Regular expression validation, conforms to Mainland China mobile number format	–	–	–	–	–	–	–	–	–	✓	–	–
		Count of invalid entries / Total Rows		–	–	–	–	–	–	–	–	–	✓	–	–
	Email format	Count of invalid entries	Regular expression validation, conforms to email format	–	–	–	–	–	–	–	–	–	✓	–	–
		Count of invalid entries / Total Rows		–	–	–	–	–	–	–	–	–	✓	–	–
	ID card format	Count of invalid entries	Regular Expression Validation, conforms to the Chinese mainland ID Card format	–	–	–	–	–	–	–	–	–	✓	–	–
		Count of invalid entries / Total Rows		–	–	–	–	–	–	–	–	–	✓	–	–
Consistency	Field data range	Value Range	Check if the numerical value is within the value range	–	✓	–	–	–	–	–	–	–	–	–	–

©2013–2025 Tencent Cloud. All rights reserved.

			Enumerated range	Check if the character value is within the enumerated values	–	–	–	–	–	–	–	–	–	–	✓	–
		Field data relevance		Compare the size with a certain field in another database table	–	–	–	–	–	–	–	–	–	–	–	✓

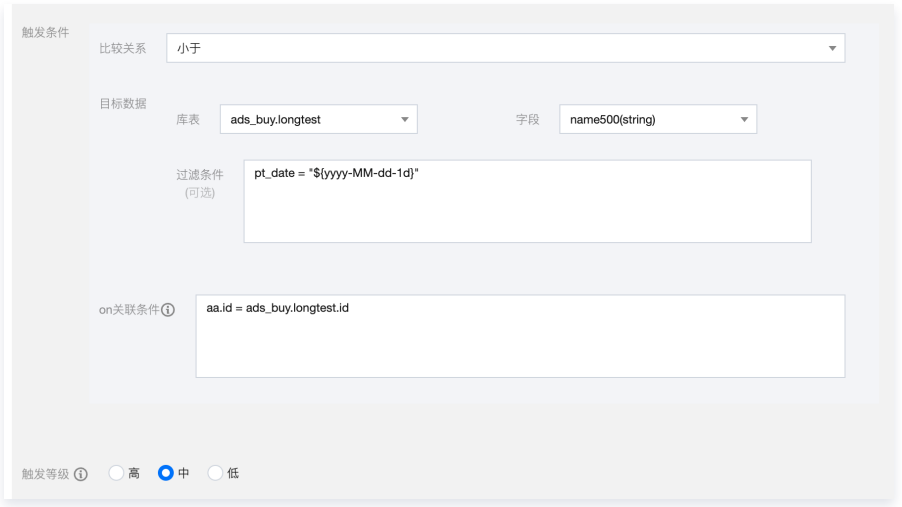
Use Instructions

Terminology		Explanation
Monitoring object	Table-level	When the monitored object is at the table level, it can monitor the number of rows, table size, and timeliness of table data output (equivalent to the number of rows).
	Field-level	When the monitored object is at the field level, it can monitor the field's numerical value (including average value, maximum value, minimum value, median, aggregate value), field value format (Phone Number, Email, ID Card Number), and whether the field is empty.
Rule dimension	–	The rule dimension is to calculate the quality score, reflecting the quality proportion of different types of rules. The system has six built-in rule dimensions: Accuracy, Uniqueness, Integrity, Consistency, Timeliness, and Validity.
Validation Method	Numeric Type	Mainly includes numerical comparison and numerical range comparison.
	Volatility-based	**Definitions:** Volatility-based is used to reflect numerical fluctuation, i.e., the rise/fall compared to a certain point in time. **Calculation Formula:** Volatility = Current scan result / Scan result at a certain point in time * 100%. **Note** The calculation result of volatility is a percentage. When using the volatility template, you must **Specify partition**. **Example 1: 7-day Cyclical Fluctuation** When a partition is specified and the baseline value is chosen from data 7 days ago, if the calculation result is 100%, It means that the data in this partition has doubled compared to the partition data 7 days ago. **Example 2: Previous Cycle Fluctuation:** When a partition is specified and the baseline value is chosen from the last run cycle, and the rule is associated with a production scheduling task (e.g., an offline development task), if the calculation result is 100%, It means that the statistical data after the end of this offline development task has doubled compared to the statistical data after the end of the last run. **Example 3: Cyclical Volatility + Default Cycle:** When setting quality rules using the cycle volatility template, a default period such as 7 days ago is set. If this rule is not associated with a production scheduling task, and the calculated result is: 100%. It means that the data in this partition has doubled compared to the partition data 7 days ago. That is, the current data is compared with the data of 7 days ago.
	Standard Score Type	**Definitions:** The standard score is an important statistical concept that indicates whether a value falls within a credible range.

©2013-2025 Tencent Cloud. All rights reserved.

(Variance Volatility)	If the calculation result is too large or too small, there is a high probability that this data is an abnormal value. **Calculation Formula:** $$z = \frac{x - \mu}{\sigma}$$ **Note** The calculation result of the standard score is a unitless decimal, indicating whether the data is outlier in the dataset. Generally, it is considered an abnormal value if the absolute value of the standard score is greater than 3, at which point the likelihood of being normal is only 0.28% • [−1,1]: Normal likelihood: 68.26% • [−2,2]: Normal likelihood: 95.44% • [−3,3]: Normal likelihood: 99.72% • Not within [−3,3]: Normal likelihood: 0.28%
Others	No limit on the verification of field types. • Null/Unique/Repeat: Count or ratio of null values/unique values/repeat values; • Format Matching: Count or ratio of non−conforming formats; • Enumeration Range: Count of values not within the enumerated values; **Note** Fill in the expected value here. When the field is out of range, an alarm will be triggered.  • Field Relevance: Count if the value matches another table's field value. ○ Comparative Relationship: Greater than, less than, equal to; ○ Target Data: Database table, field, filter conditions; ○ Associated Conditions: Associated fields of the two tables. **Note** Comparison Table needs to correspond with the Detection Table data one−to−one.

©2013−2025 Tencent Cloud. All rights reserved.



©2013–2025 Tencent Cloud. All rights reserved.

Add and manage custom Definition templates

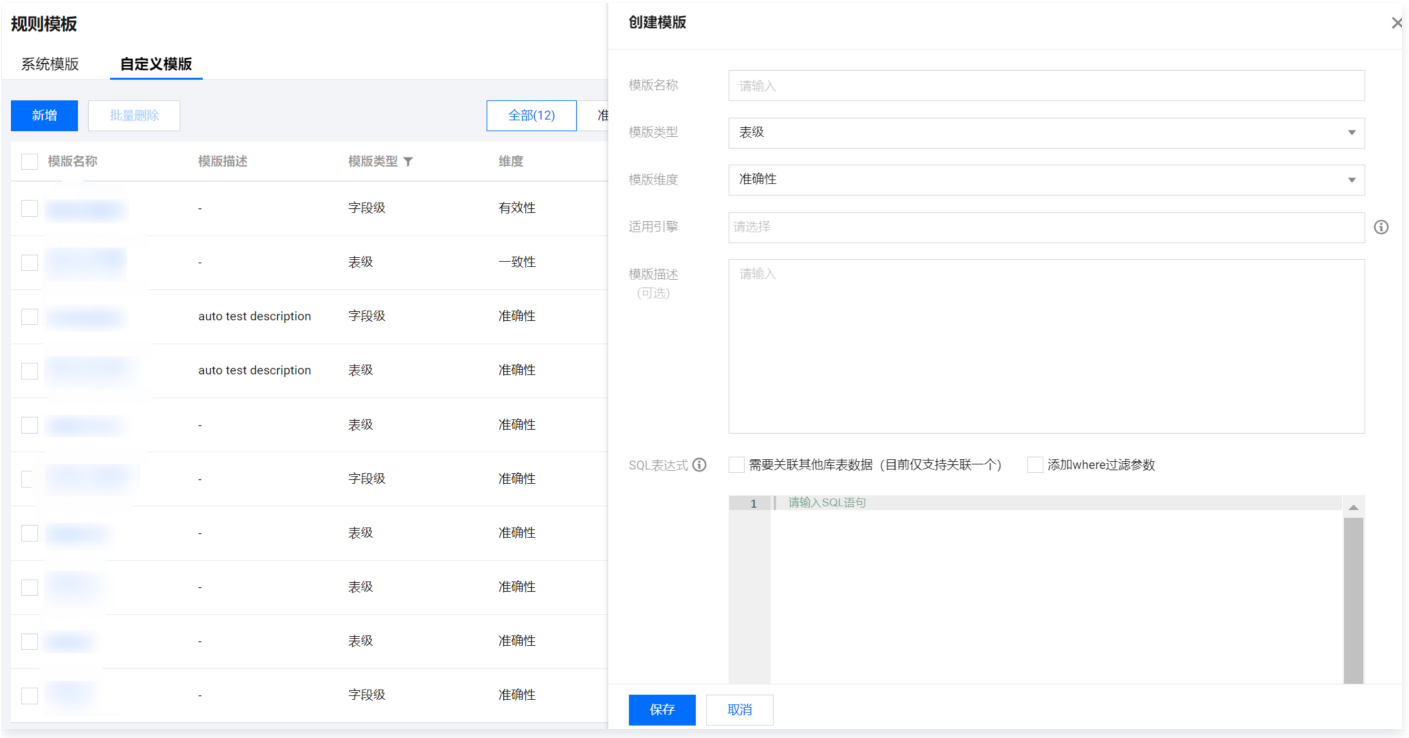
Last updated: 2024-12-16 16:54:05

When the system template cannot meet the actual project needs, users can define frequently used SQL detection statements as templates and reuse them in projects.

Create custom Definition templates

Support users to create custom Definition monitoring rule templates using SQL expressions.

1. Go to **Rule Templates** > **Custom Definition Templates**, click the **Add** button, and the template creation window will pop up.
2. In the template creation interface, configure template dimensions, applicable targets, template description, and SQL expressions to complete the template creation.



Information	Details
Template Name	Custom Definition Rule Template Name
Template Type	Template Inspection Targets, Table or Field
Template Dimensions	6 Dimensions: Accuracy, Uniqueness, Integrity, Consistency, Timeliness, Validity
Applicable Engines	Support for Hive, Spark, DLC, TCHouse-D, and Doris.
Template Description	Add description information for the custom Definition template
SQL Expression	Fill in an SQL statement, and preset parameter variables in SQL to meet more usage scenarios. **Note:** The result must be one row and one column, i.e., the result must be a fixed value (could be a character or a value); If the result is multiple rows, the system will automatically take the first one.

©2013-2025 Tencent Cloud. All rights reserved.

You can preset parameter variables in SQL, and actual users can fill in the values themselves.

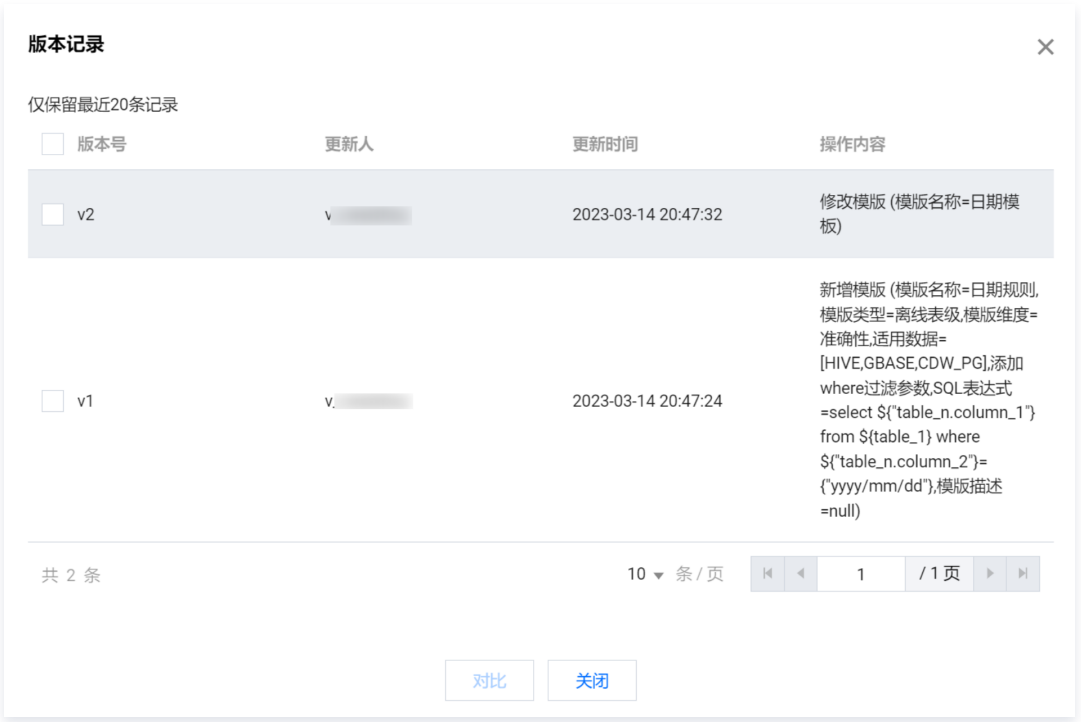
1. Current detection rule associated Main Table: ${table_1};
2. Fields in the current detection rule associated Main Table: ${table_1.column_1}, ${table_1.column_2}...${table_1.column_n};
3. Support associating one other table under the same data source and library: ${table_2} or ${table_3}...${table_n}, note that only one table is supported;
4. Fields of other associated tables: ${table_2.column_1}, ${table_2.column_2}...${table_2.column_n};
5. Set where filter parameters: use ${param_1}...${param_n};
6. Use time parameters, e.g., ${yyyy-MM-dd-1d}, for details refer to Time Parameter Description.

Custom Definition Template List



Information		Details
Template Name		6 Dimensions: Accuracy, Uniqueness, Integrity, Consistency, Timeliness, Validity
Template Description		Table types applicable to this template
Template Type		Add description information for the custom Definition template
Dimension		Check dimensions of custom definition rules. 6 Dimensions: Accuracy, Uniqueness, Integrity, Consistency, Timeliness, Validity
Applicable objects		Applicable objects of rule templates: Table, Field
Reference count		Number of times the rule template is referenced by detection tasks
Update time		Last updated time of the rule template
Creator		Creator of the rule template
Operation	Edit	Edit Rule Template. Editable content is the same as the configuration items for creating a template
	Delete	Single delete or bulk delete from Definition Rule Template
	Version history	View the historical version update record of the rule template. Comparative feature supported

©2013-2025 Tencent Cloud. All rights reserved.



©2013–2025 Tencent Cloud. All rights reserved.

Configure Quality Task Features Overview

Last updated: 2025-04-25 15:19:29

Feature Description

Configuring a quality monitoring task is the starting point of the data quality module. In this module, you can:

1. newly-added quality monitoring task
2. Manage quality monitoring tasks.

Core Capabilities

Adding New Quality Monitoring Task

- support various Tencent Cloud Big Data Storage Engines (EMR, DLC, TCHouse-P, TCHouse-D and Doris).
- Configure data quality inspection rules at the table level or field level.
- Configure execution policies based on actual business scenes.
- Set the strength of rules to determine whether to block downstream tasks.
- Support various user reach methods (WeCom Group, WeChat, call, SMS, mail, Lark Group, DingTalk Group).
- Support trial runs.

Managing Quality Monitoring Tasks

- Support viewing the list of monitoring tasks under a data table based on the data table, and support enabling and disabling.
- Support viewing the Quality Rule List, and support editing and managing.

Prerequisites

- The Tencent Cloud big data engines EMR, DLC, TCHouse-P, TCHouse-D and Doris have been bound in the project.
- Bind a scheduling resource group in the project.

©2013-2025 Tencent Cloud. All rights reserved.

Add Rule under Monitoring Task

Last updated: 2025-04-25 15:29:18

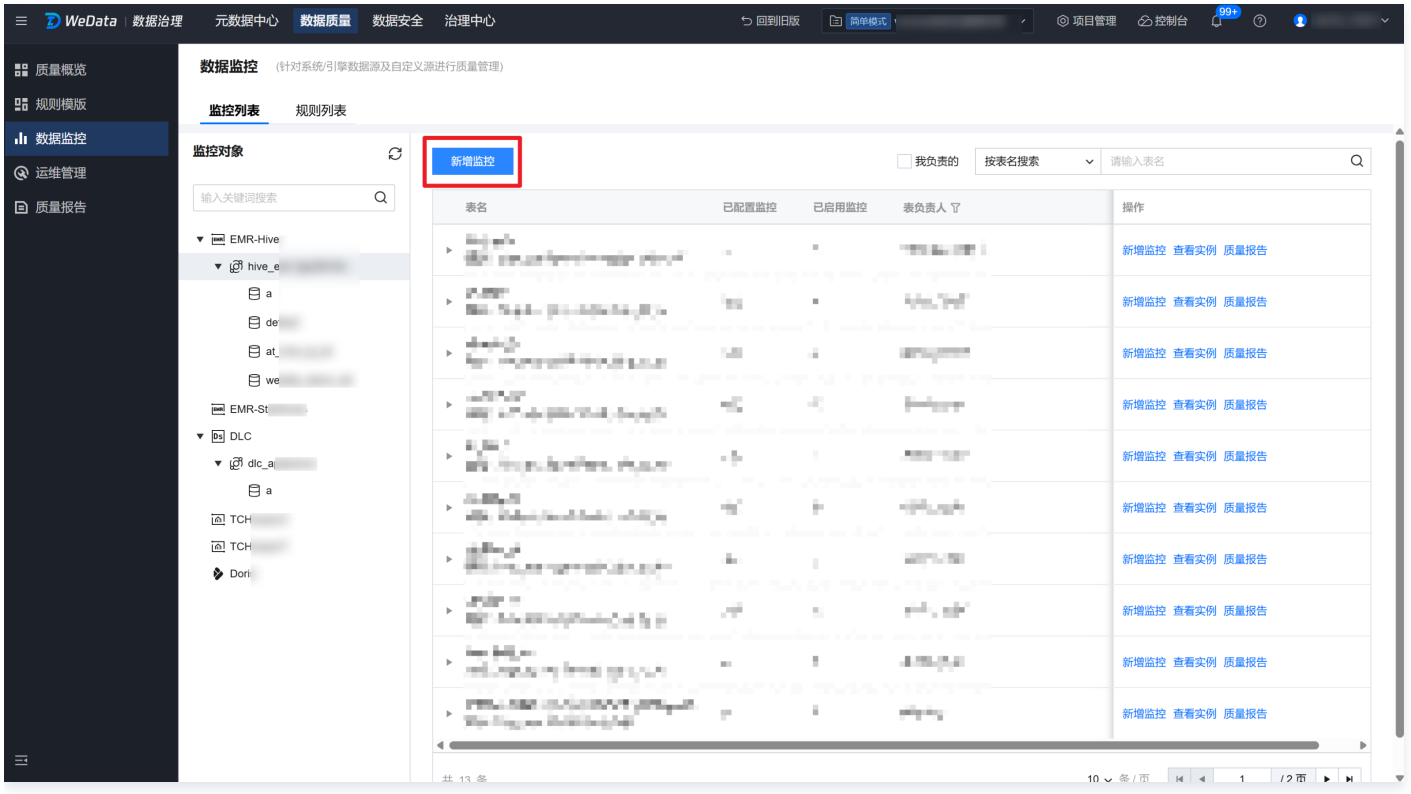
Pre-Explanation

Type	Note:
Data Source Type	Currently, WeData supports the following **data source types:** • EMR ○ EMR-Hive • DLC • TCHouse-P • TCHouse-D • Doris

The following three methods	Currently, WeData supports the following **three new methods:**
	• Create Quality Rule under Monitoring
	  ○ Can only add rules under one table (monitoring tasks can only monitor one table).
	  ○ Multiple rules can be added at once.
	• Multiple Tables Addition: Batch create monitoring rules for multiple fields in multiple tables from the same data source.
	  ○ You can select multiple tables and multiple fields at once.
	  ○ You can only select one monitoring rule at a time.
	• **Batch upload:** Upload an Excel template and import in batches.
	  ○ You can only target the same type of data source at a time.
	  ○ Can only support custom SQL (no support for system-built-in templates and custom templates).
	  ○ Only 100 items can be uploaded at a time.

Create a Quality Monitoring Task

Enter the Data Quality > Data Monitoring > EMR interface, click Add Monitoring, and start adding a quality monitoring.

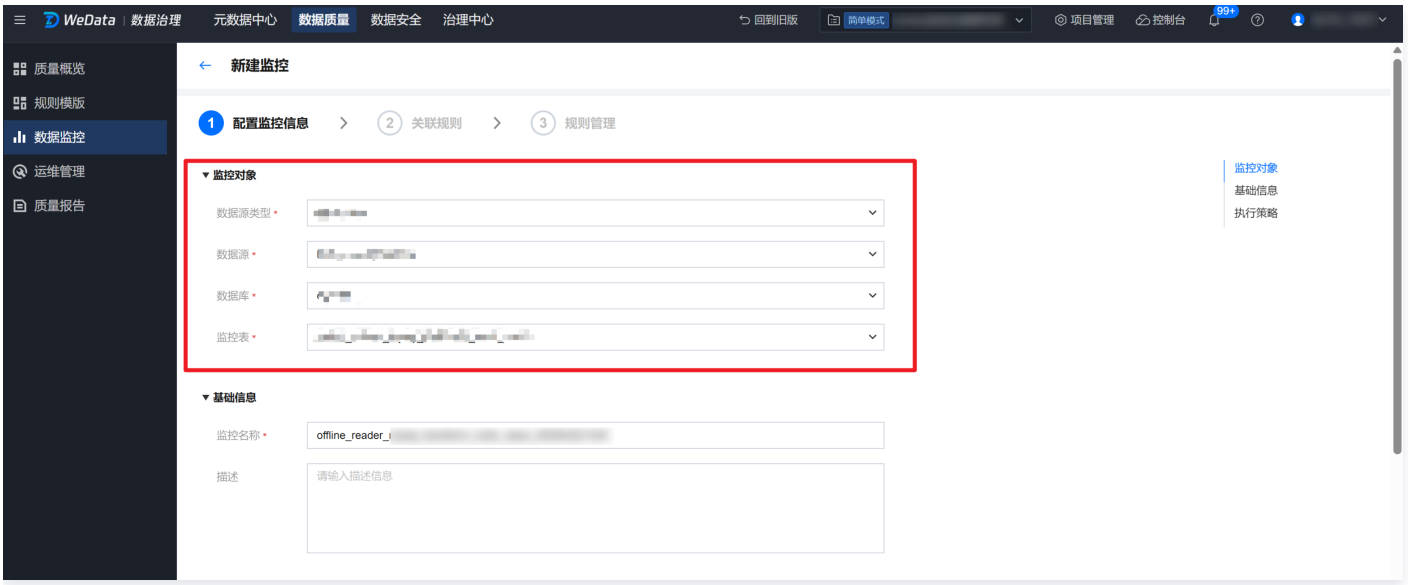


©2013-2025 Tencent Cloud. All rights reserved.

Step 1: Configuring Monitoring Information

Monitoring Object

Enter the Create Monitoring interface, sequentially select the data source type, data source, database, and monitoring table of the monitoring object.

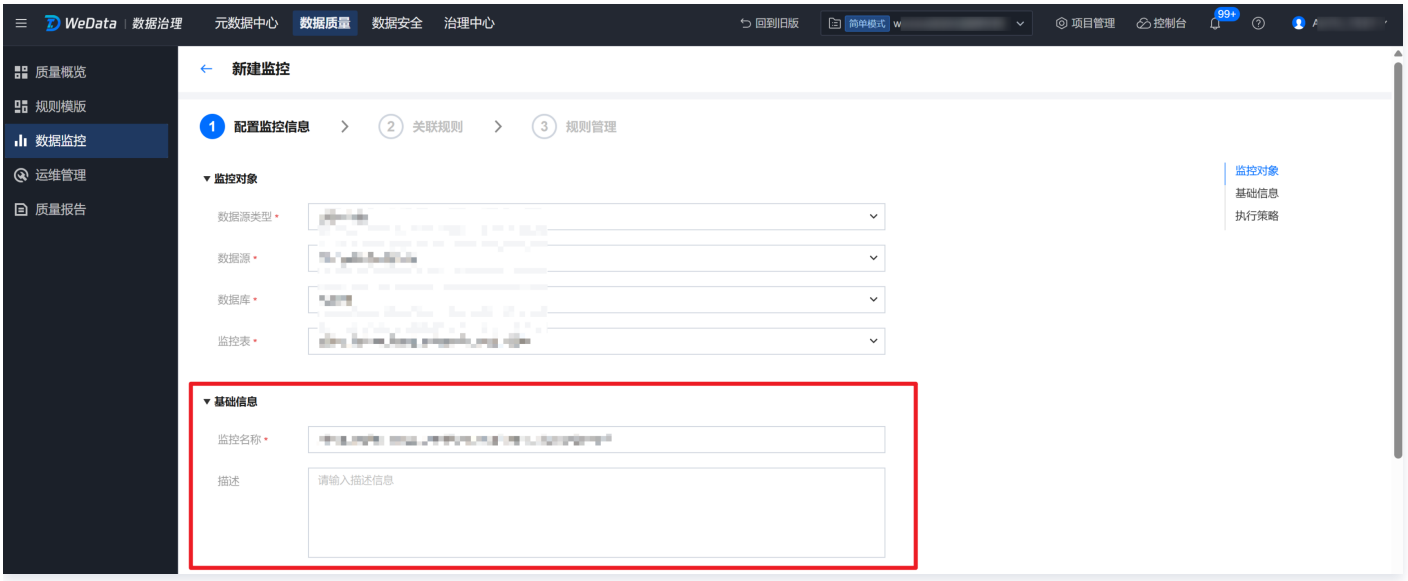


Basic information

Fill in the monitoring name. The default monitoring name is **monitoring table name_Creation monitoring timestamp**.

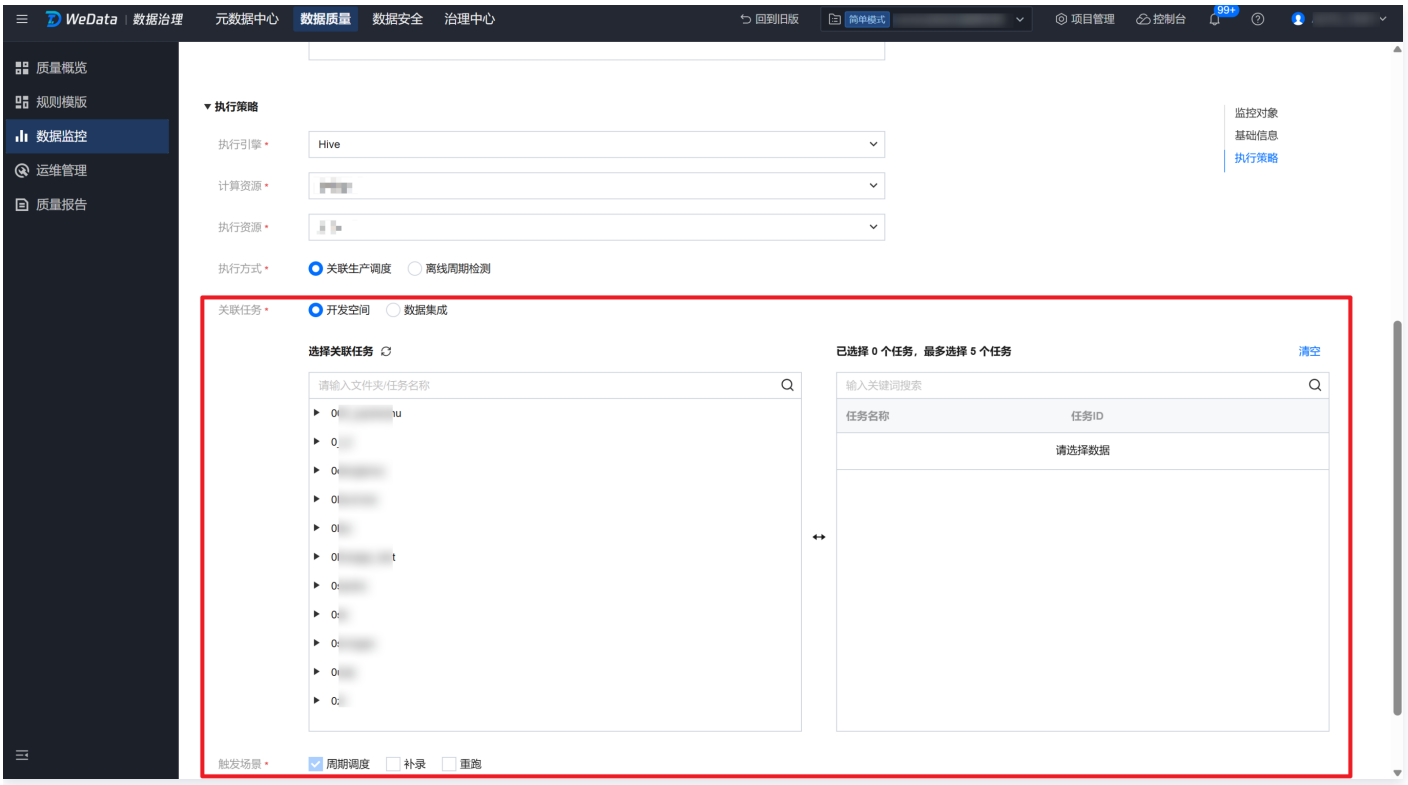
> **Note**
> Suggest filling in an easily identifiable monitoring name to facilitate understanding of the business significance of detection results in monitoring task instances.
> Note: When the monitoring name is "North China region sales table monthly routine detection", the Ops management will show whether the detection of "North China region sales table monthly routine detection" succeeds or fails every time it runs.

Meanwhile, for convenience in interpretation, you can also add a business description.



Execution Policy

©2013–2025 Tencent Cloud. All rights reserved.



Element	Description
Execution Engine	Here you can select Hive and Spark, which are related to the purchased EMR resources. Generally, Hive tables can directly select the Hive engine.
computational resource	select default Here you can select the resource group in the EMR cluster. Generally, you can directly select default.
execution resource	The execution resource here refers to the scheduling resource group that has been bound to the project.
Execution Method	Here you can select **associated production scheduling** and **offline cycle detection**: **Associate Production Scheduling**: Link quality tasks with production tasks (data sync tasks or data development tasks). After the production task completes, insert a quality monitoring task. If an exception is detected, notify the handler immediately for processing. Based on the rule level in monitoring, block downstream task execution to prevent problem data expansion. **Note** The same quality inspection task can be associated with multiple production tasks; the same production task can also be associated with multiple quality inspection tasks. **Offline Cycle Detection**: Refers to independent scheduling. Conduct periodic quality inspections on selected database tables and core business fields at customizable frequencies such as daily, hourly, or by minute. Quality tasks are executed on a scheduled basis according to the set period. If an exception is detected, subscribers will be notified immediately.
Associated task(s)	Here you can select the module to which the production task that needs to be associated belongs, either the development space or Data Integration. A maximum of 5 tasks can be associated: • Development space: Show the directory tree structure in the development space. Support searching for folders or task names. • Data Integration: Display the task list in data integration.

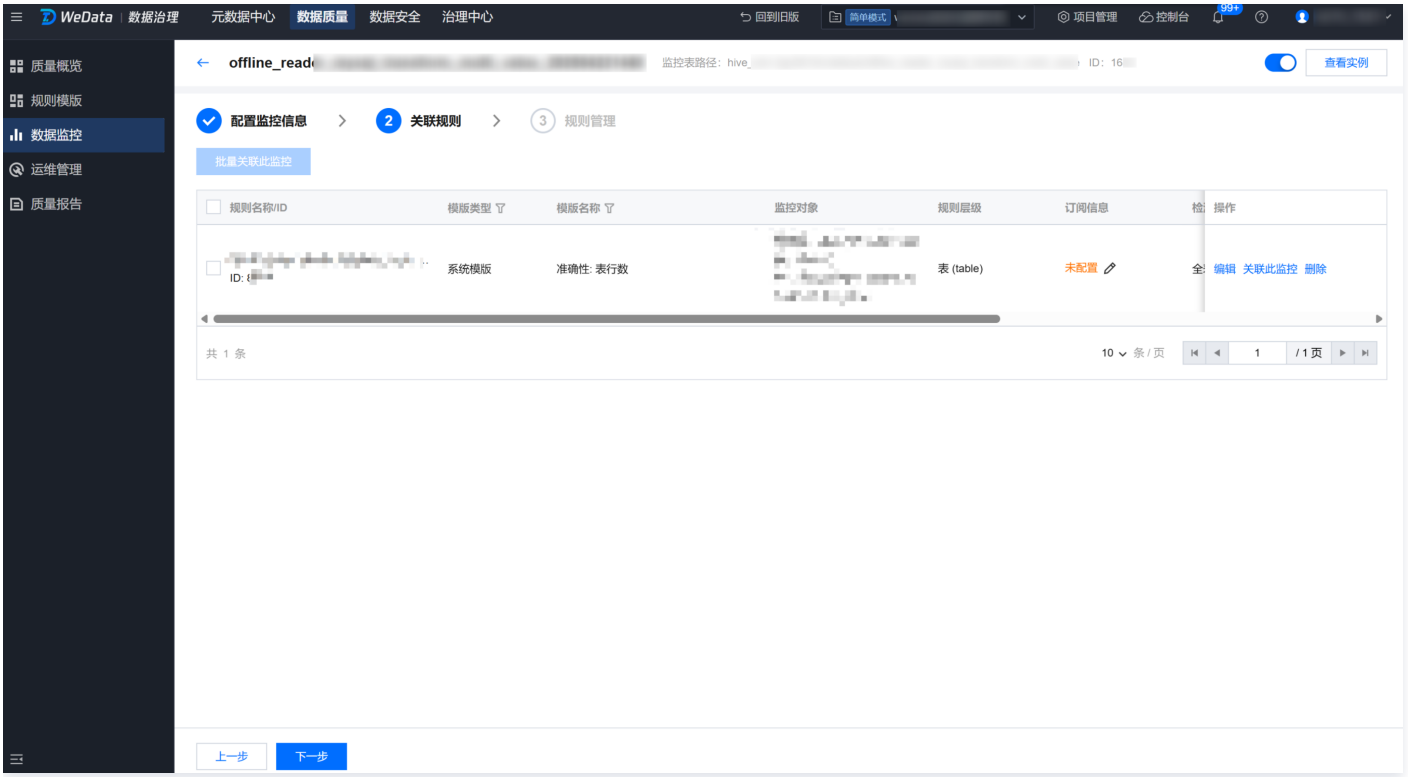
Click Next. After the monitoring is created, enter the association rule interface.

Step 2: Association Rule

©2013–2025 Tencent Cloud. All rights reserved.

Association rules are displayed under the same table, for rules not associated with other monitoring tasks. Support editing or associating quality rules (optional).

> **Notes:**
> Quality rules that are unassociated with monitoring will not execute.

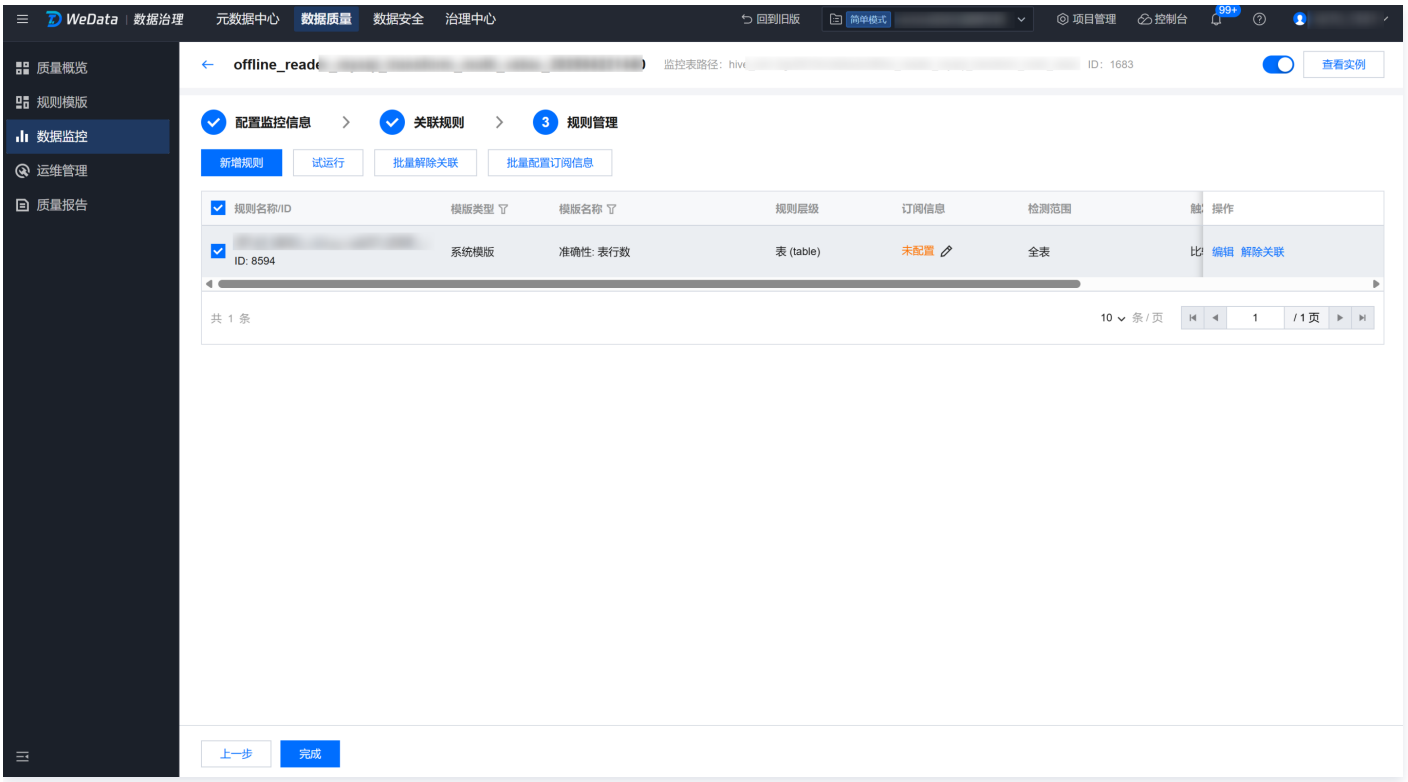


If existing rules cannot meet business needs, click Next to access the rule management interface.

Step 3: Rule Management

Rule management displays all rules associated with this monitor. All previously associated quality rules will be echoed here.

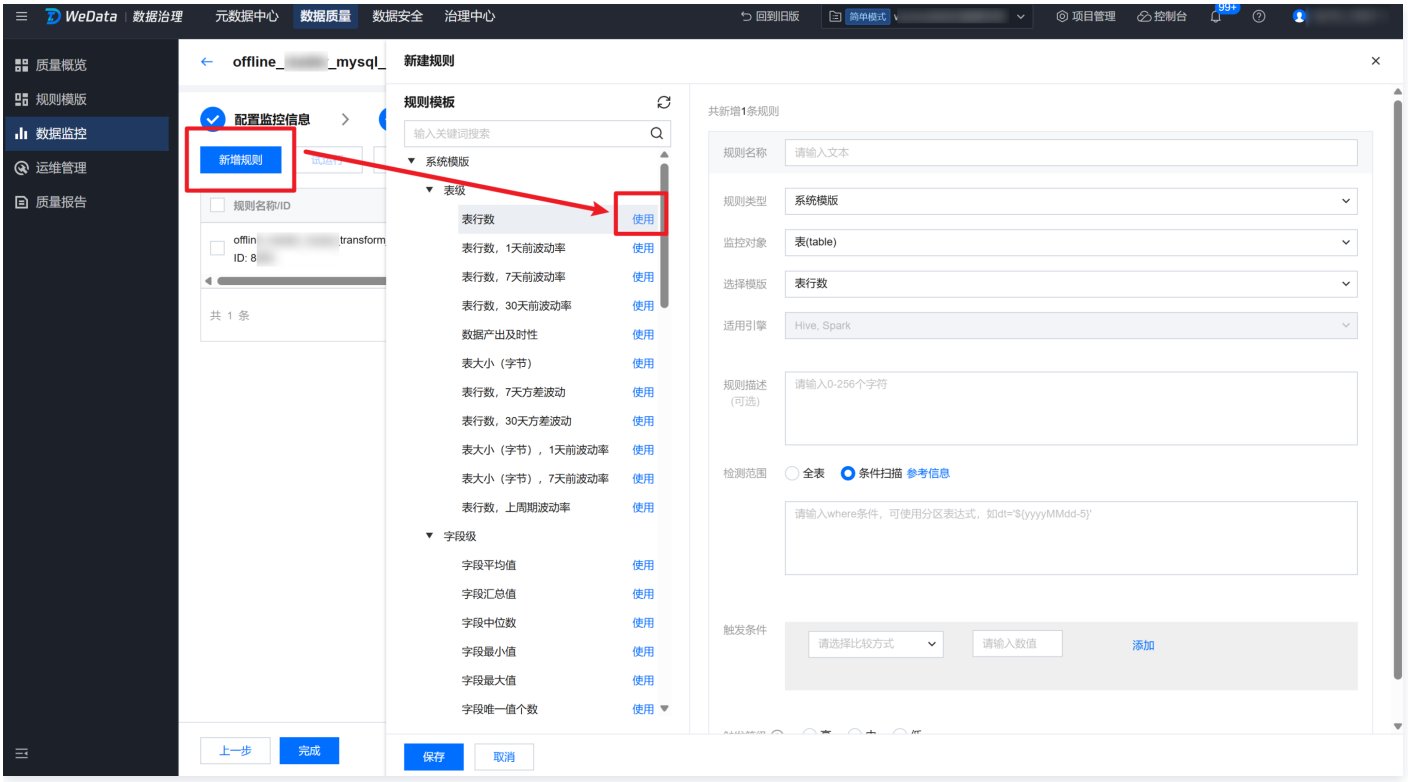
©2013–2025 Tencent Cloud. All rights reserved.



Creating New Rule

Click Create New Rule in the rule list. A pop-up for creating a new rule will appear, supporting selection to perform quality verification.

Support creating multiple rules at once. Newly added rules will be directly associated with this monitor.



Explanation:

Element	Note:

©2013–2025 Tencent Cloud. All rights reserved.

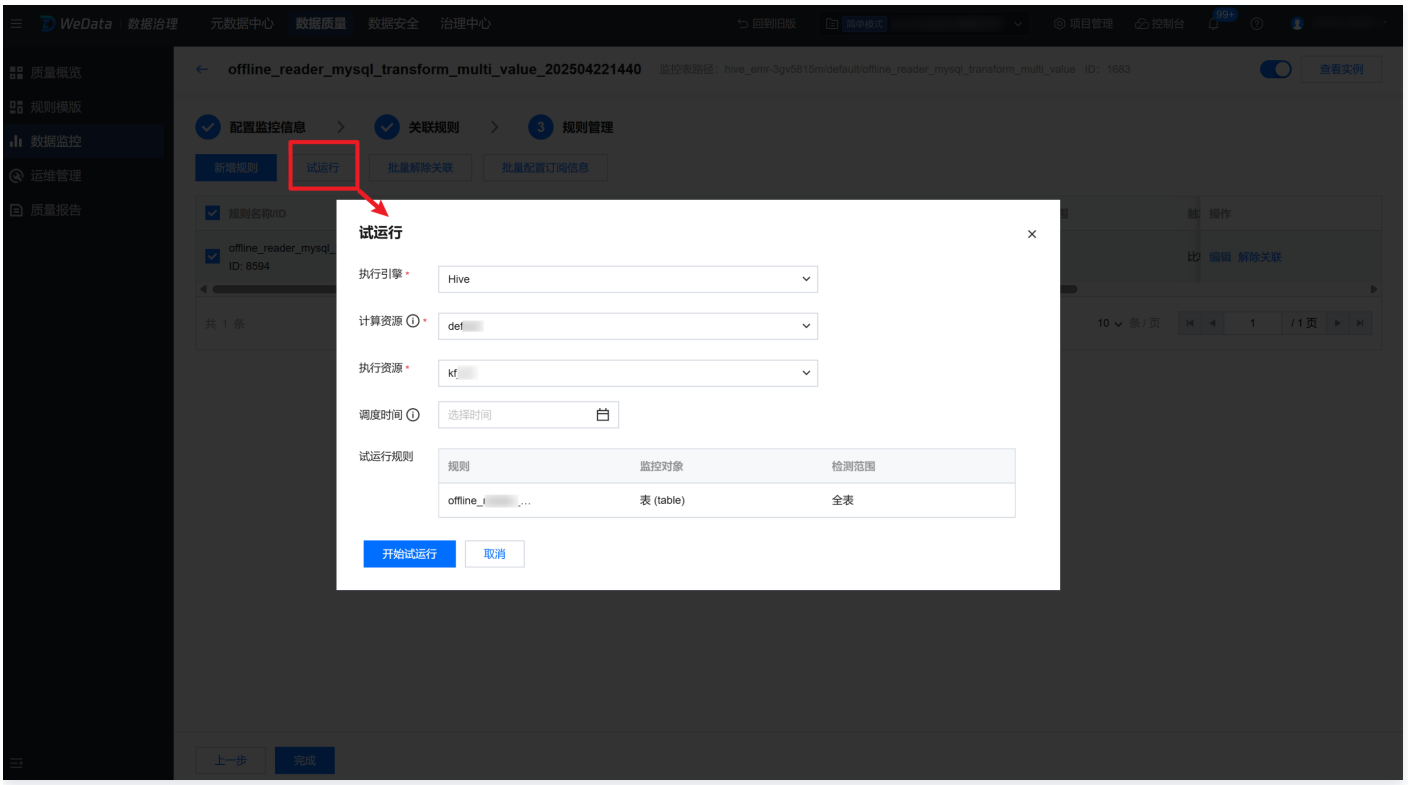
Rule Type	Select a system template, custom template, or custom SQL. (If a rule template is selected from the dendrogram on the left, the selected template parameters will be displayed by default here.) • System Template: WeData has 56 built-in rule templates that can be used for free. For more details on each template, refer to system template description. • Custom template: You can add rules applicable to your own business in the rule template menu for easy reuse. For detailed operation instructions, see Custom Template Description. • Custom SQL: You can directly fill in SQL statements as detection rules. For detailed operation instructions, refer to Creating New Quality Rules.
Monitoring Object	Monitored objects can be divided into: table-level and field-level. (If selection rules templates are chosen from the left-side dendrogram, selected template parameters will be displayed by default here.) • Table level: You can monitor the number of rows and the table size (only supported for Hive tables). • Field level: You can monitor whether the field is empty, whether it is repeated, the average value, the maximum value, the minimum value, etc.
Template selection	WeData has 56 built-in rule templates that can be used for free. For more details on each template, refer to system template description. If a rule template is selected from the dendrogram on the left, the selected template parameters will be displayed by default here.
Detection scope	You can choose **full table** or **scanning with conditions**. • Full table: The quality rule will validate the full data under the table. • Condition scan: Quality rules will only validate the detection range filled in here. Example: `pt_date='${yyyy-MM-dd-1d}'` **Note** Note: Generally, fill in the partition field here to avoid a full table scan during each quality task, which would waste computational resources. In SQL, ***${yyyy-MM-dd-1d}*** is a date variable representing one day before the execution date and will be replaced with a specific date during Quality Task Execution. For example, when the quality task is executed at 2024-05-02 00:00:00, ${yyyy-MM-dd-1d} will be replaced with 2024-05-01. The replacement logic of specific time variables can be found in time parameter explanation.
Trigger Condition	• Comparison operator: Select **less than.** • Compare value: Fill in 1. • - The number of table rows is less than 1. Combine the time variable filled in combination with the detection range, which indicates that when there is no new data added yesterday, an alarm is triggered. **Note** Note: The trigger condition filled here is **abnormal value**, that is, the conditions for triggering alarms.
trigger level	Select **Medium**. Trigger levels can be divided into: High, Medium, Low. • High: Immediately block downstream task execution when an alarm is triggered (valid only when associated with production tasks). • Medium: Trigger alarm only. • Low: Do not trigger an alarm, only show abnormal result display.

Trial Run

1. Select the rules you wish to trial run, and click **Trial Run**.
2. Modify the scheduling time, and click **Start Trial Run**. After the trial run ends, you can click **Click to View Operation Results**.

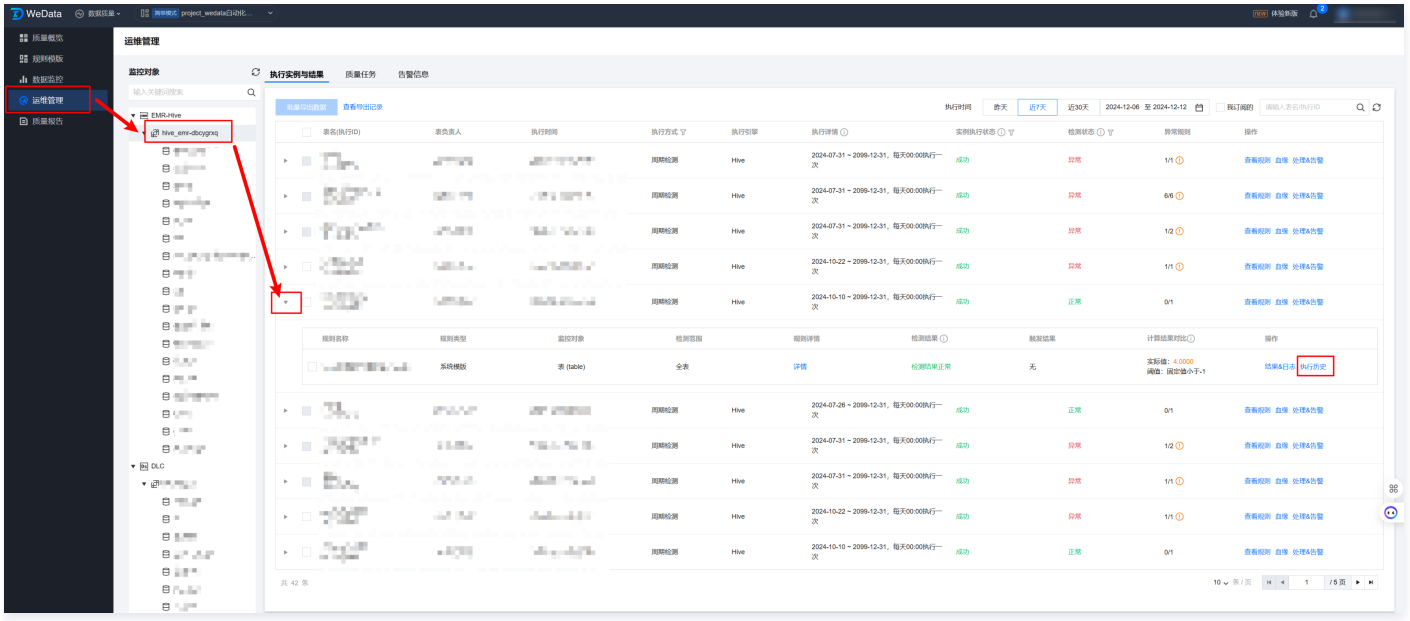
©2013-2025 Tencent Cloud. All rights reserved.

> **Note**
> Note: When inserting data, if the filled data partition is 2024-05-01, you can fill in 2024-05-02 here.



View Trial Run Results

1. After entering the **Execution Instance and Result Page**, you can click on the left **dropdown triangle** to expand the rule details.
2. Click the **execution history** of a specific monitoring item to view historical run results.



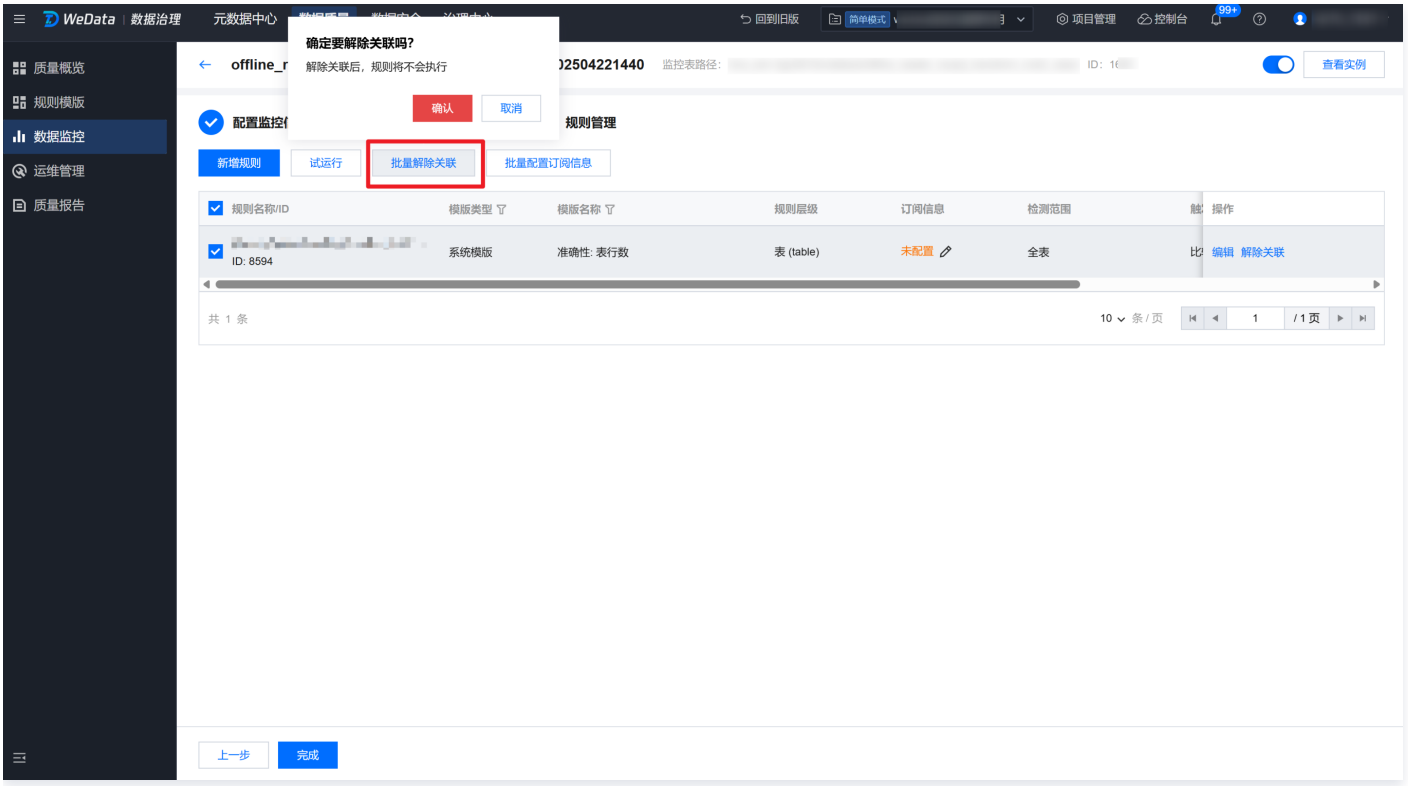
Batch Disassociation

Bulk disassociation of the selected rules from the monitor.

> **Notes:**
> Quality rules that are unassociated with monitoring will not execute.

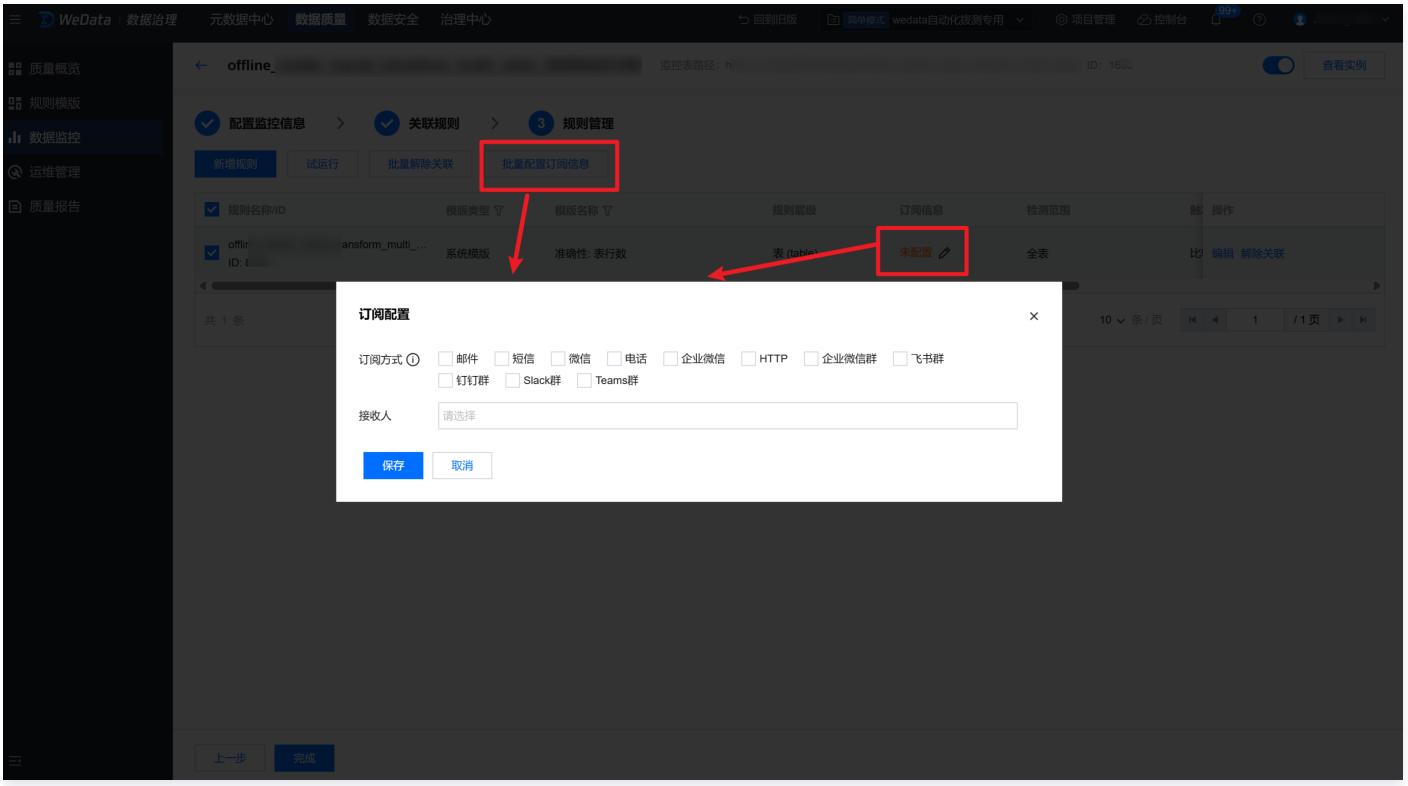
©2013-2025 Tencent Cloud. All rights reserved.

To reassociate this rule, you can perform operations on the **Association Rule** page.



Configure Subscription

When the running result of the quality detection task meets the trigger conditions (i.e., the result is not as expected and the detection result is failed), what method to use for sending notifications and whom to notify.
Enter the Subscription Settings interface, select the Subscription Method, set the Recipient, and click Save.

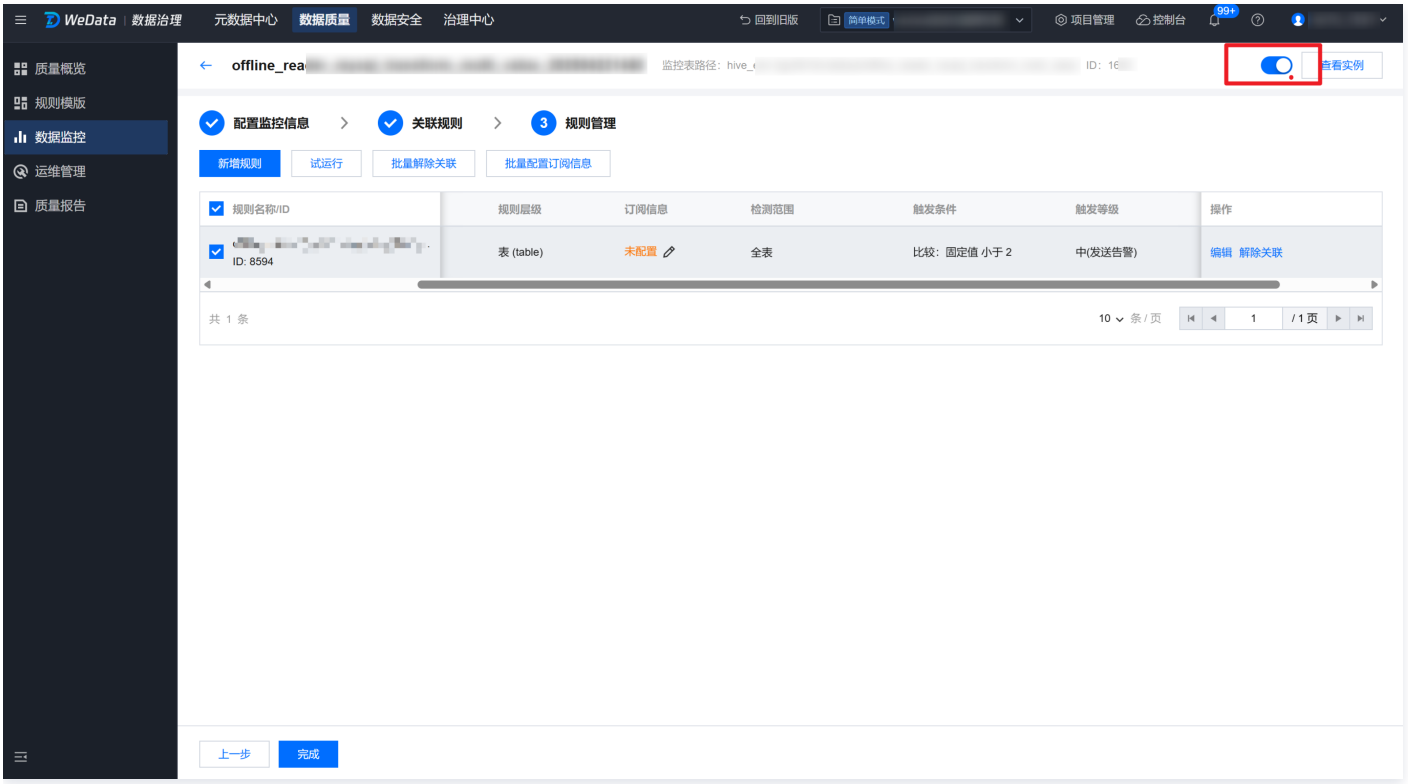


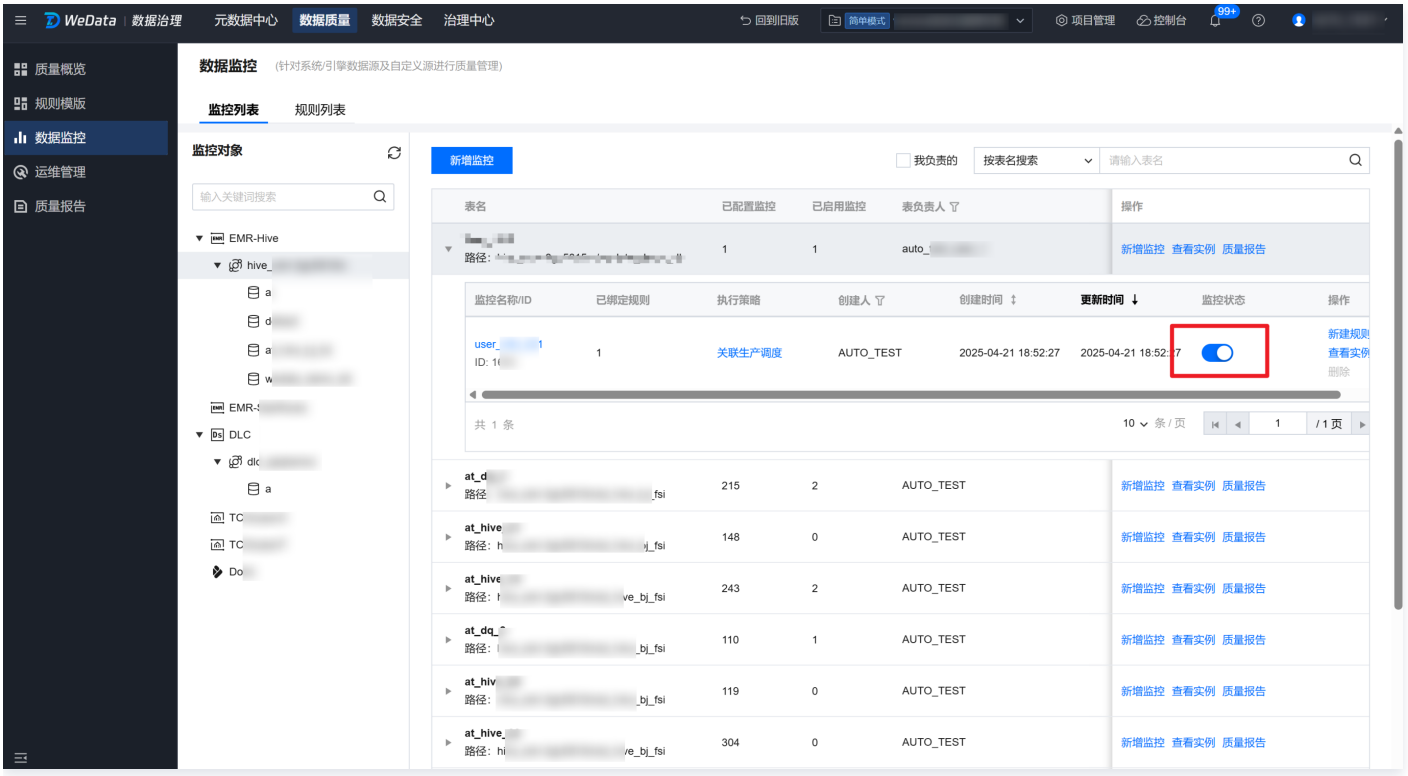
Enable Monitoring

©2013–2025 Tencent Cloud. All rights reserved.

According to the trial run result of monitoring, you can enable or disable monitoring in the monitoring status column **enable** or **disable monitoring** (monitoring is enabled by default after creation) .

> **Note**
> Only monitoring tasks in the **enablement** state will automatically run.





©2013–2025 Tencent Cloud. All rights reserved.

Add Rules in Batches

Last updated: 2025-04-25 15:25:48

In the **Rule List**, you can use **Add Rules to Multiple Tables** and **Batch Upload Rule** to batch configure quality rules for multiple tables.

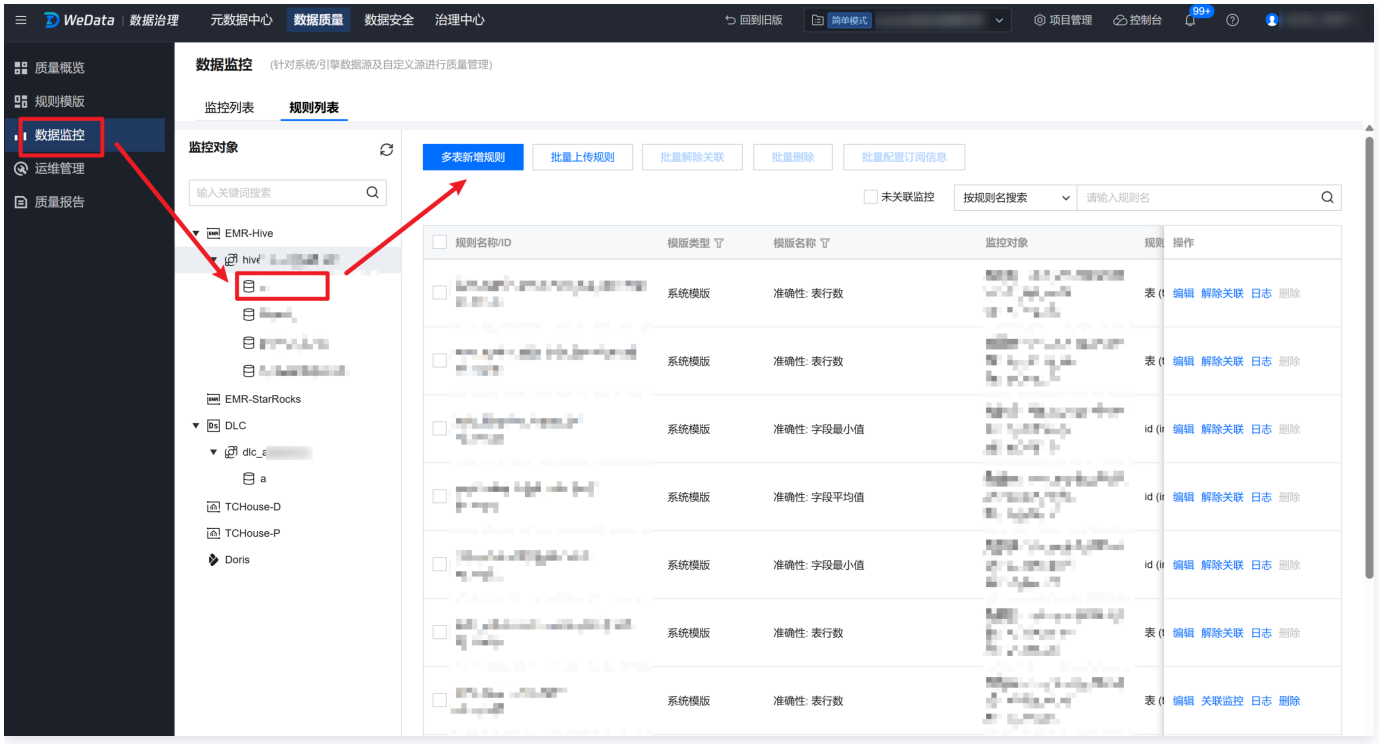
Add Rules to Multiple Tables

Support setting monitoring rules for multiple tables.
Use case: Add the same monitoring rule to multiple databases and tables in one-time operation.

Step 1: Enter Creation Page

Enter the Data Quality > Data Monitoring > EMR interface, click Add Rules to Multiple Tables, and start creating multiple quality rules.

> **Note**
> Currently, WeData supports the following supported data source types: EMR-Hive, DLC, TCHouse-P, TCHouse-D, and Doris.



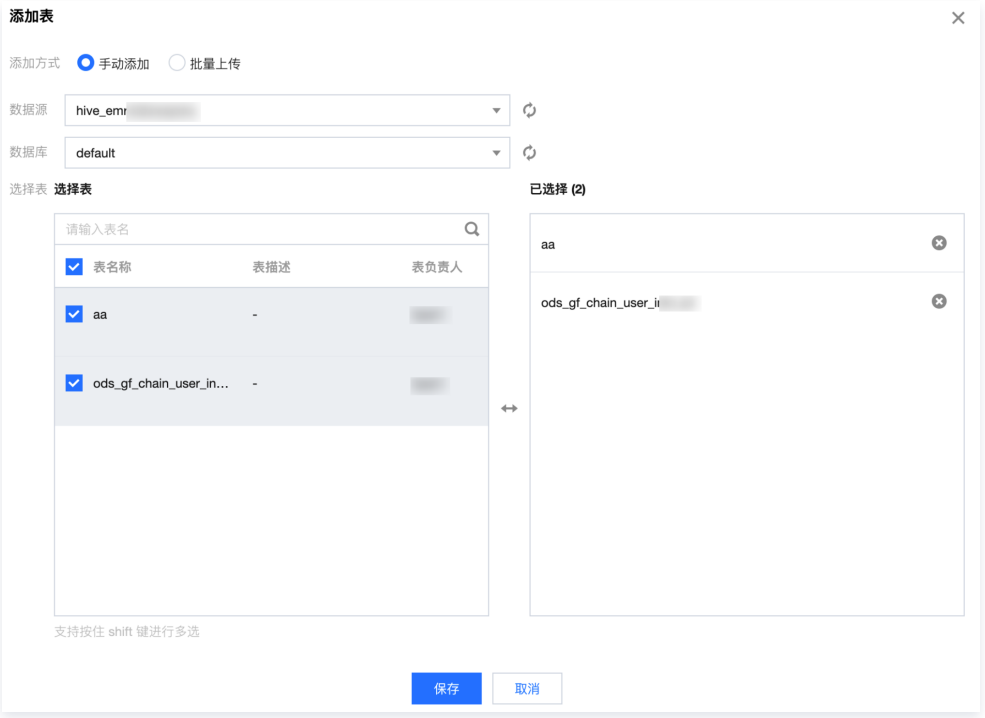
Step 2: Set Monitoring Scope

1. Click Add Rules in Batches to enter the Add Rules in Batches interface. You can choose to monitor multiple tables or monitor multiple fields.
   - Multi-table: Add multiple tables from the same data source, which can span across databases.
   - multi-field: Add multiple tables from the same data source, which can span across databases and tables.

©2013-2025 Tencent Cloud. All rights reserved.



2. Click **Add Table** to add a table/field.
   - Manually add: Just check.
   - Batch upload: You need to upload an Excel file. The templates for monitoring multiple tables and multi-field are different.
     - Monitor multiple tables: The Excel file contains three columns: data source name, database name, and table name.
     - Monitor multiple fields: The Excel file contains four columns: data source name, database name, table name, and field name.



3. Click Next to enter the rule configuration page.

Step 3: Configuring Monitoring Rules

1. Select Rule Template: Only supports **system templates** and **custom templates**.

- WeData has 56 built-in rule templates that can be used for free. For more details on each template, see system template description.
- Custom template: You can add rules applicable to your own business in the rule template menu for easy reuse. For detailed operation guidance, see Custom Template Description.
  - Database and table parameters: Render the page and let users select based on the SQL statements filled in according to the custom template.

©2013–2025 Tencent Cloud. All rights reserved.

- table_1 represents the currently selected table; table_2...table_n represent other tables, and the determined table is required (currently only 1 is supported).
- ${table_1.column_1}...${table_1.column_n} refer to the fields in the table, and the determined fields need to be selected.

- where parameter: Render the page and let users select based on the SQL statements filled in according to the custom template.
  - ${param_1}...${param_n} represent the parameters in the where condition, and specific values need to be filled in.

2. Select a template: Filtering will be performed here based on rule type and monitored object.

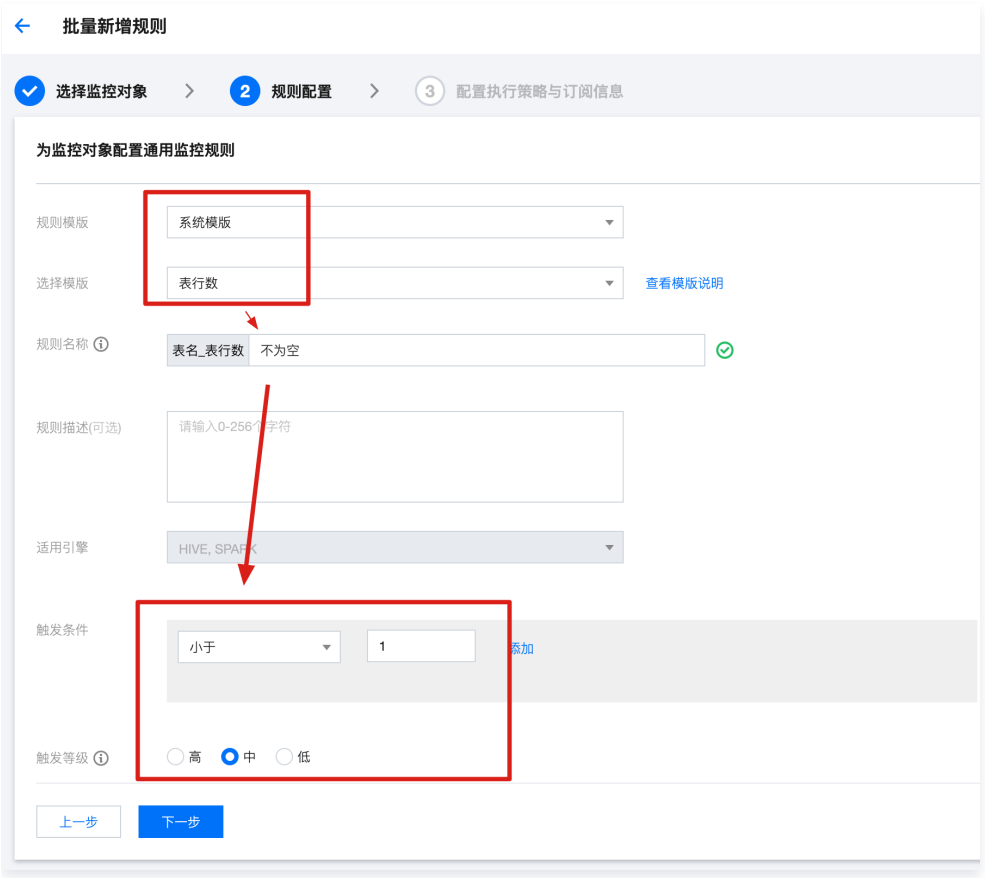
   For example, select **system template**, set the monitoring object to **table level**, and here you can only select **number of rows**, **table size**, etc.
3. Trigger condition: The comparison operator can select interval values and size values.

   Example: When the number of table rows is less than 1, combined with the time variable filled in combination with the detection range, it indicates that when there is no new data added yesterday, an alarm is triggered.
   - Comparison operator: Select `;`
   - Compare value: Fill in 1.

   Trigger conditions vary for different templates. See System Template Usage Instructions . for detailed configuration logic.

> **Note**
> Note: The trigger condition filled here is an **abnormal value**, which is the condition for triggering alarms.

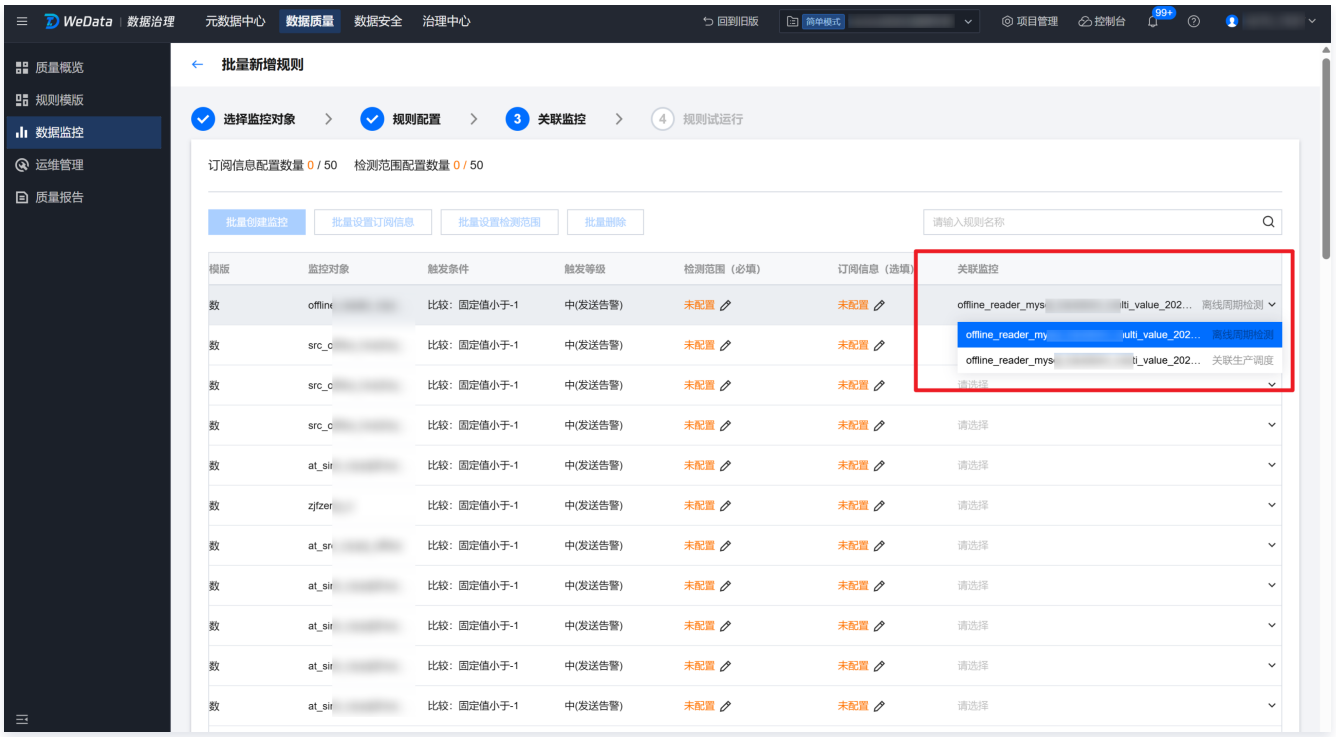
4. Click **Next** to enter the strategy configuration execution page.



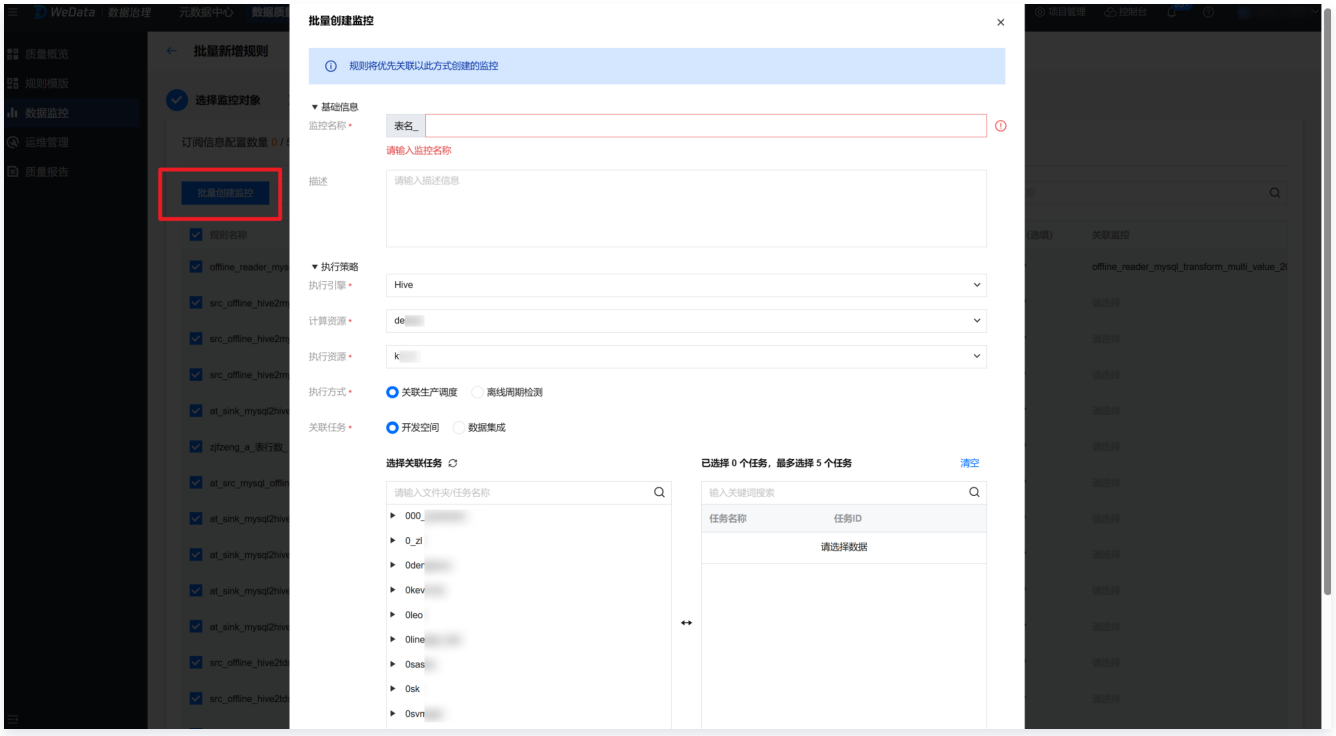
Step 4: Associate Monitoring

1. In the **Associate Monitoring** field, click the dropdown list and select a monitoring task from the table corresponding to this rule to associate.

©2013–2025 Tencent Cloud. All rights reserved.



2. If necessary, to bulk create and associate quality monitoring tasks, select the corresponding rule and click **Bulk Create Monitoring**.



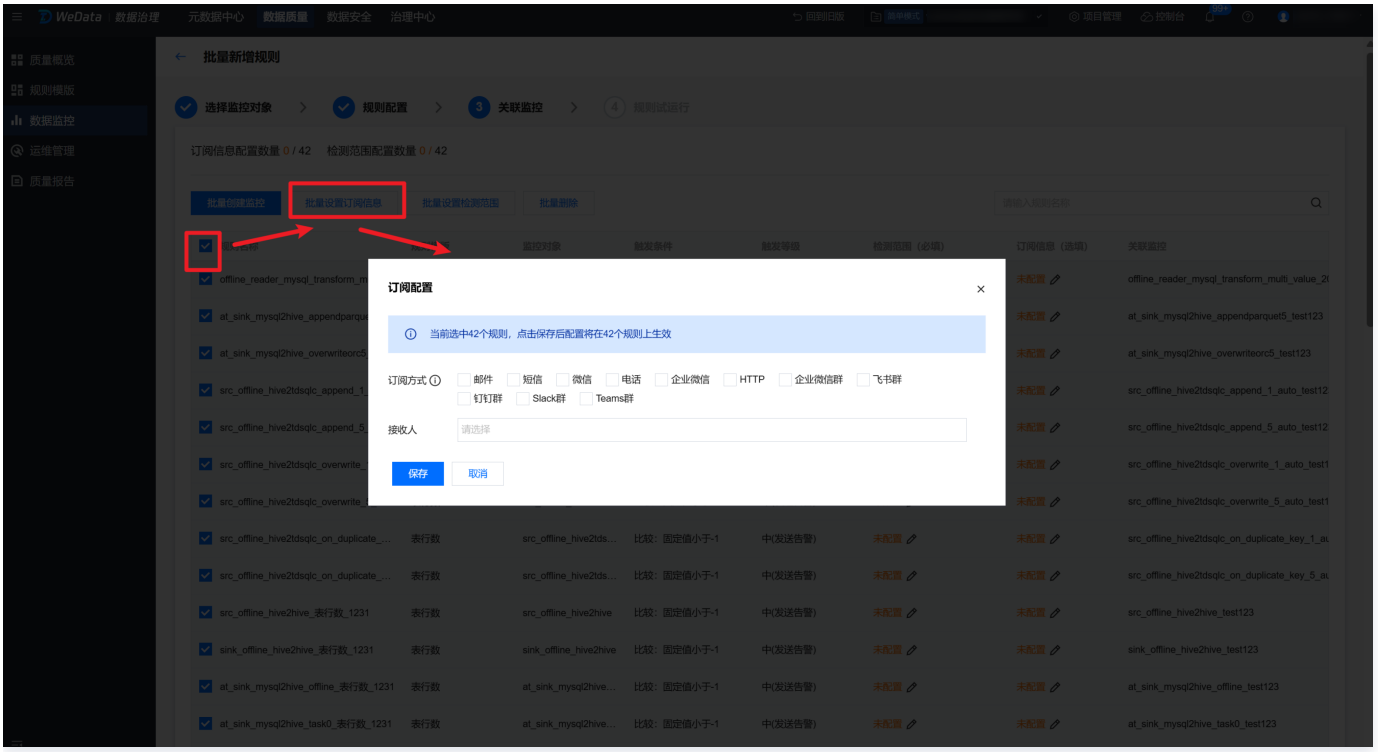
3. After creating a monitor, the selected rule will be automatically associated with the created monitor. The created monitor will be displayed in the **monitoring list**.

Step 5: Set Up a Subscription

Set up a subscription notification. When the running result of a quality detection task meets trigger conditions (i.e., the result is not as expected and the detection result is failed), determine what method to use to send notifications and whom to notify.

1. Enter the **Batch Set Subscription Information** interface to batch set subscription information.
2. On the pop-up subscription settings page, select the **subscription method** you need, set the **recipient**. Click **Save**.

©2013–2025 Tencent Cloud. All rights reserved.



Step 6: Set Detection Scope

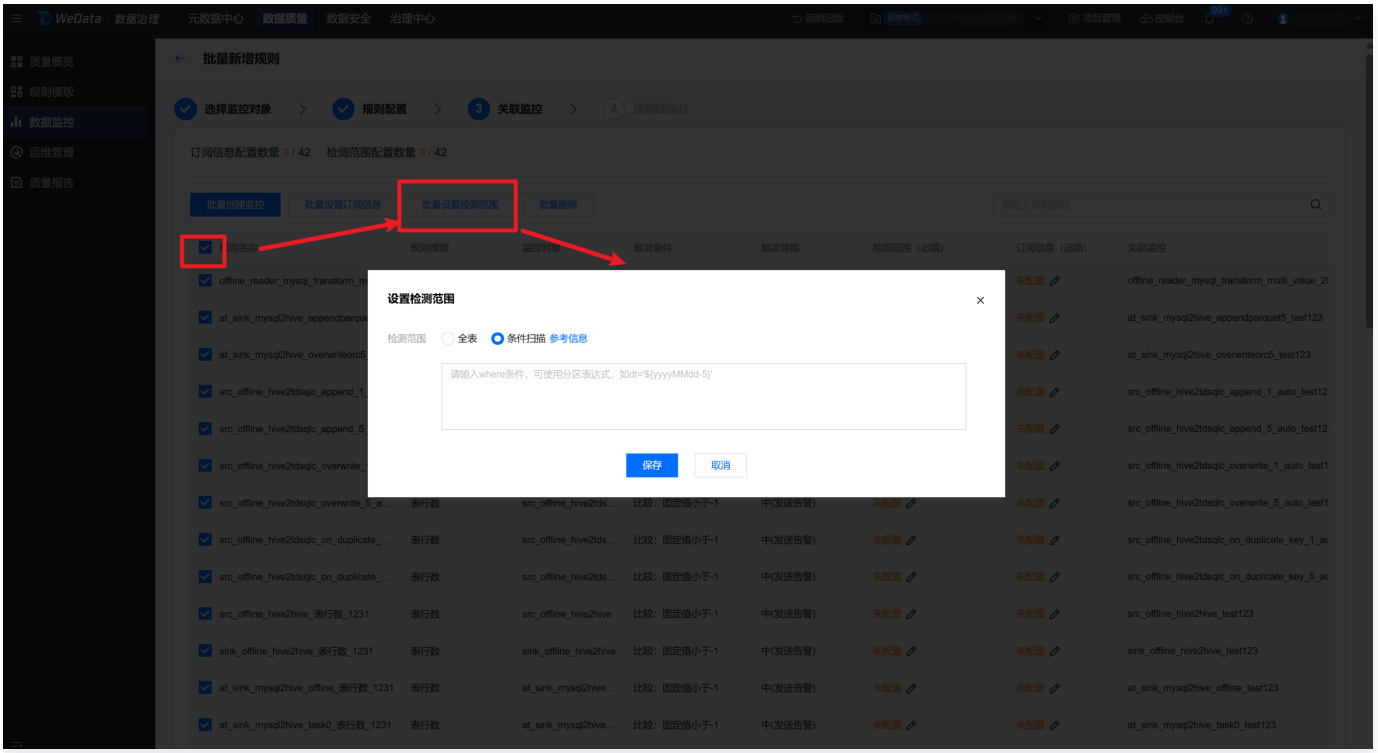
1. Click **Batch Set Detection Scope** to enter the **Set Detection Scope** interface and just batch set the detection scope.
2. Add the following information on the pop-up page.
   - Here you can select conditional scan and full table scan. It is recommended to choose **conditional scan**.
   - Partition where conditions can be filled in, for example:

```
pt_date='${yyyy-MM-dd-1d}'
```

> **Note**
> Here, generally fill in the partition field to avoid a full table scan during each quality task, which would result in a waste of computational resources.

In SQL, ***${yyyy-MM-dd-1d}*** is a date variable representing one day before the execution date and will be replaced with a specific date during Quality Task Execution.
For example: When the quality task is executed at 2024-05-02 00:00:00, ${yyyy-MM-dd-1d} will be replaced with 2024-05-01.
The replacement logic of specific time variables can be found in time parameter explanation.

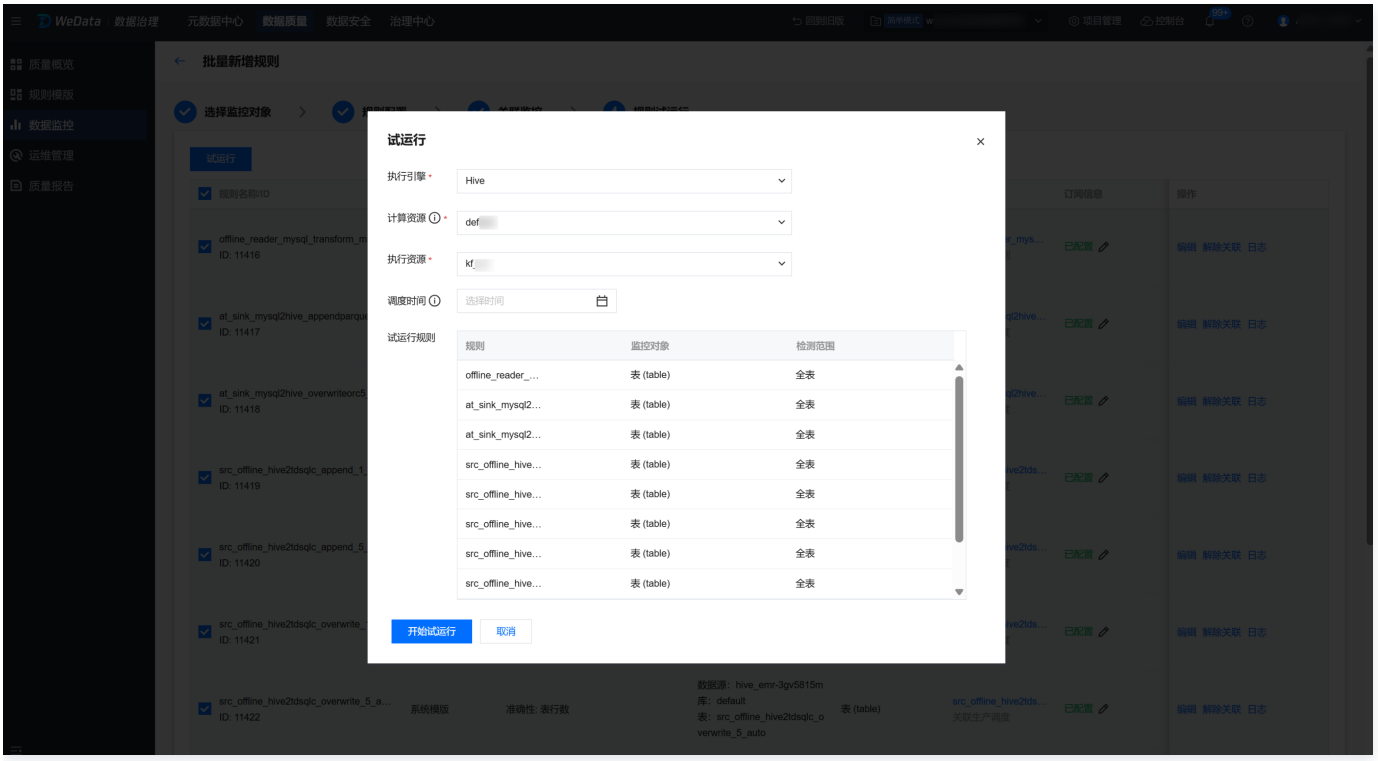
©2013-2025 Tencent Cloud. All rights reserved.



Step 7: Trial Run

1. Selection rules. Click **Trial Run**.
2. Modify the scheduling time and click **Start Trial Run**. After the trial run ends, you can click **Click to View Operation Results**.

> **Note**
> The scheduling time filled in here will modify the partition time variable.
> For example, if you fill in 2024-05-02 here, ${yyyy-MM-dd} represents 2024-05-02, and ${yyyy-MM-dd-1d} represents 2024-05-01.



Batch Upload Rule

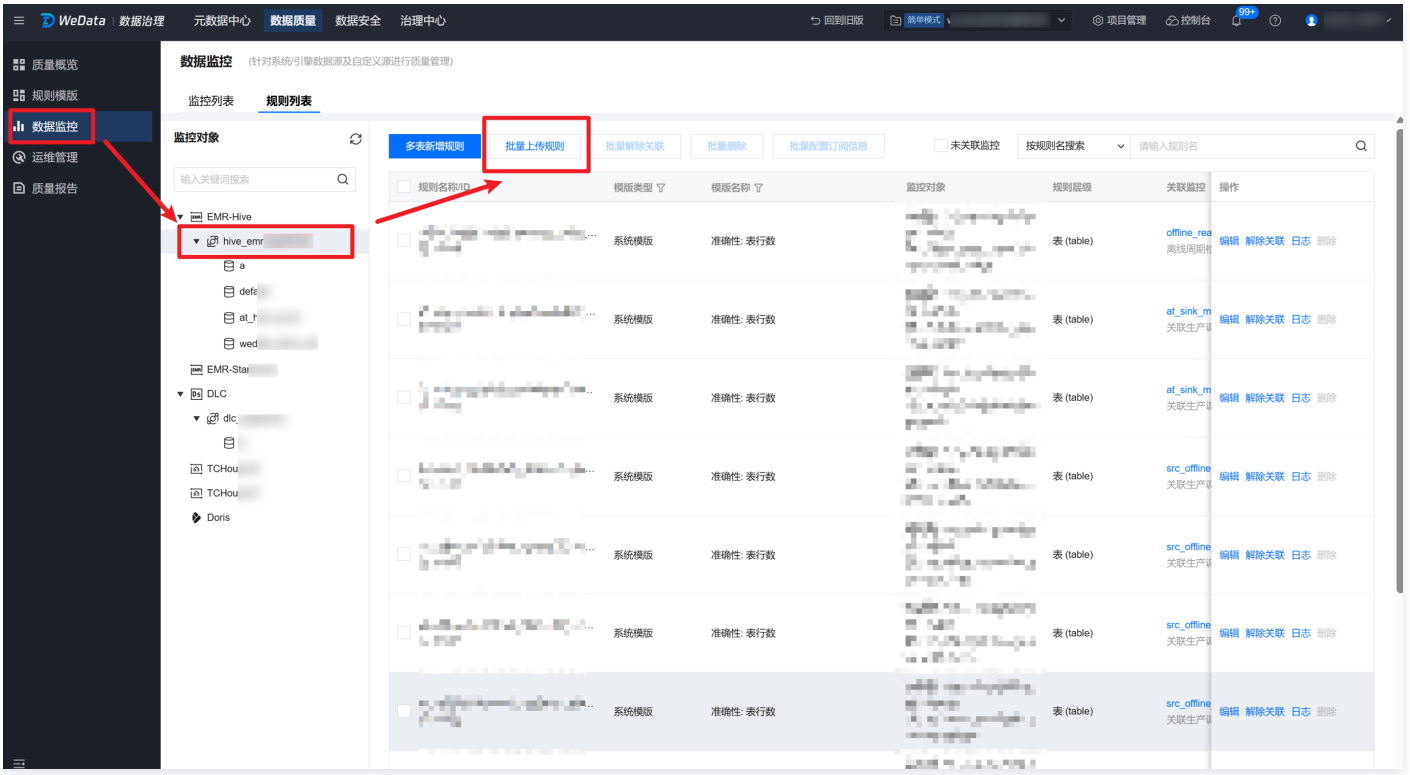
©2013-2025 Tencent Cloud. All rights reserved.

Upload attachments by uploading Excel files. Add new quality monitoring rules. Support setting monitoring rules for multiple tables. Use case: Add multiple monitoring rules to multiple databases and tables in one-time operation.

Step 1: Enter Creation Page

Enter the Data Quality > Data Monitoring > EMR interface, click Batch Upload Rule, and you can add multiple quality rules.

> **Note**
> Currently, WeData supports the following supported data source types: EMR-Hive, DLC, TCHouse-P, TCHouse-D, and Doris.



Step 2: Upload Template

1. Enter the Data Monitoring > Batch Upload Rule interface. On the upload template interface, click Download Template. After local modification, click Click Upload to upload the template.
   - SQL statement: You need to fill in a SQL statement here. The requirements are as follows:
     - The result must be in one row and one column, that is, a fixed value.
     - Only partition variables are allowed, such as ${yyyy-MM-dd}.
     - not allowed to use column name variables
   - Trigger condition: The comparison operator and the compare value are separated by a colon.
     - Comparison operator: Just fill in Chinese characters.
     - Compare value: Just fill in a numeric value.
2. Click Next to continue configuration.

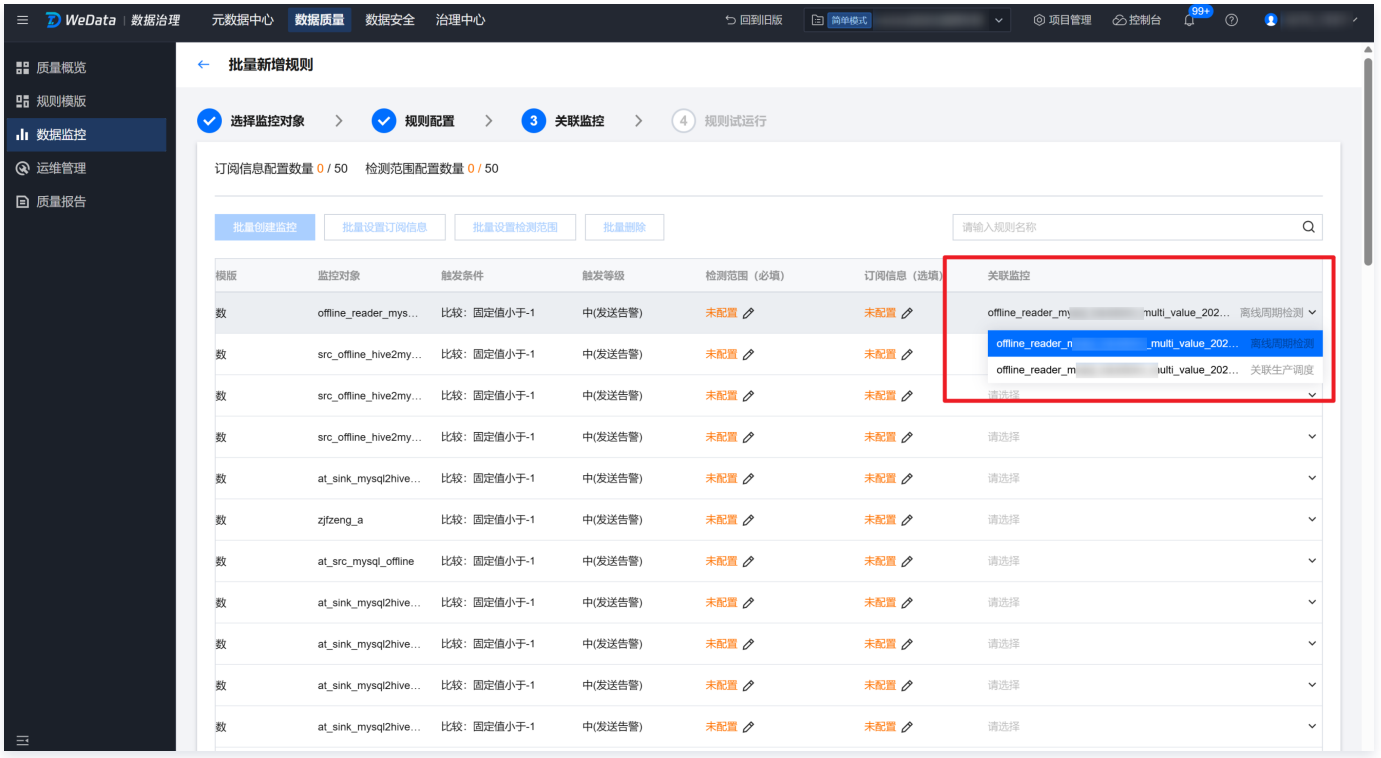
©2013-2025 Tencent Cloud. All rights reserved.



	自定义SQL	适用引擎	触发条件	触发等级
1	select count(*) from emall.dws_trade_order	HIVE;SPARK	(大于:1)或(在区间内:[0,1))	高

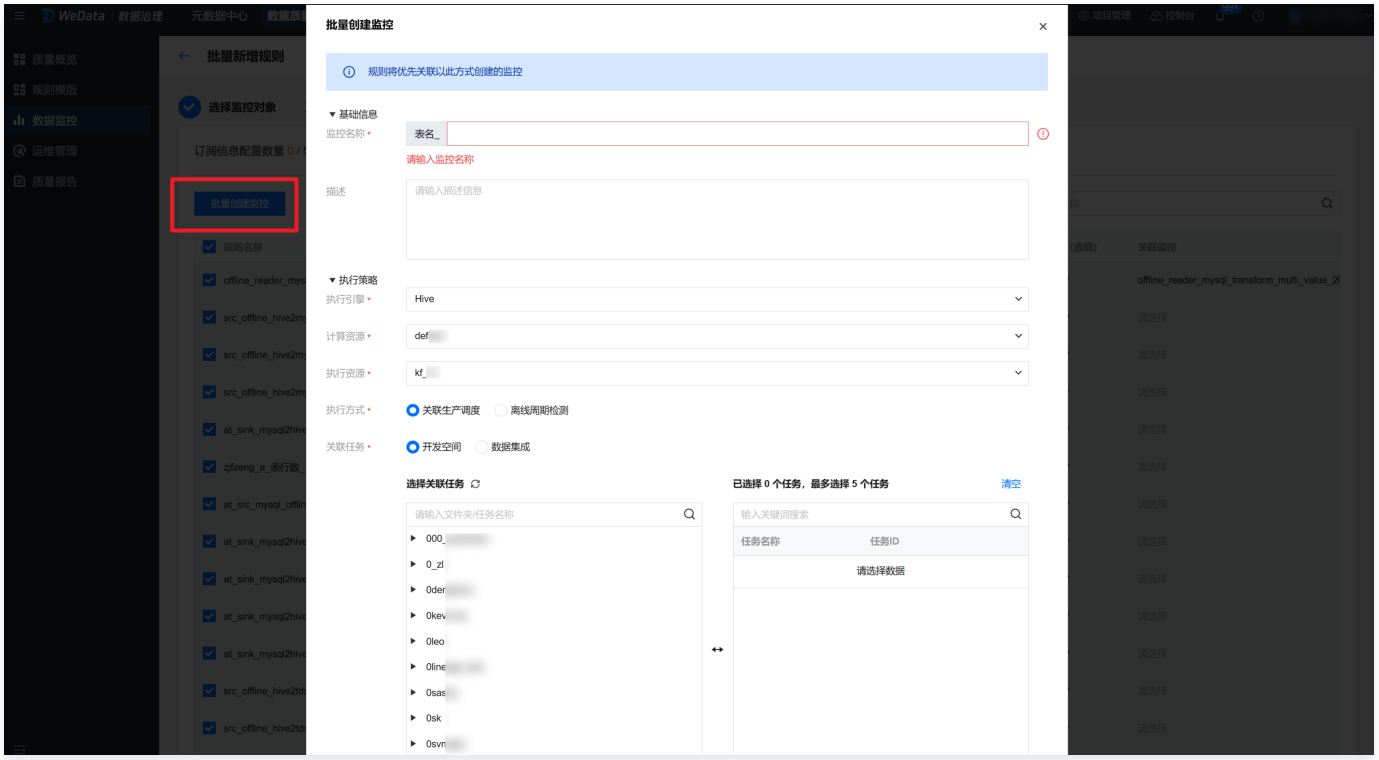
Step 3: Associate Monitoring

1. In the **Associate Monitoring** field, click the dropdown list and select a monitoring task from the table corresponding to this rule to associate.



2. If necessary, to bulk create and associate quality monitoring tasks, select the corresponding rule and click **Bulk Create Monitoring**.

©2013–2025 Tencent Cloud. All rights reserved.

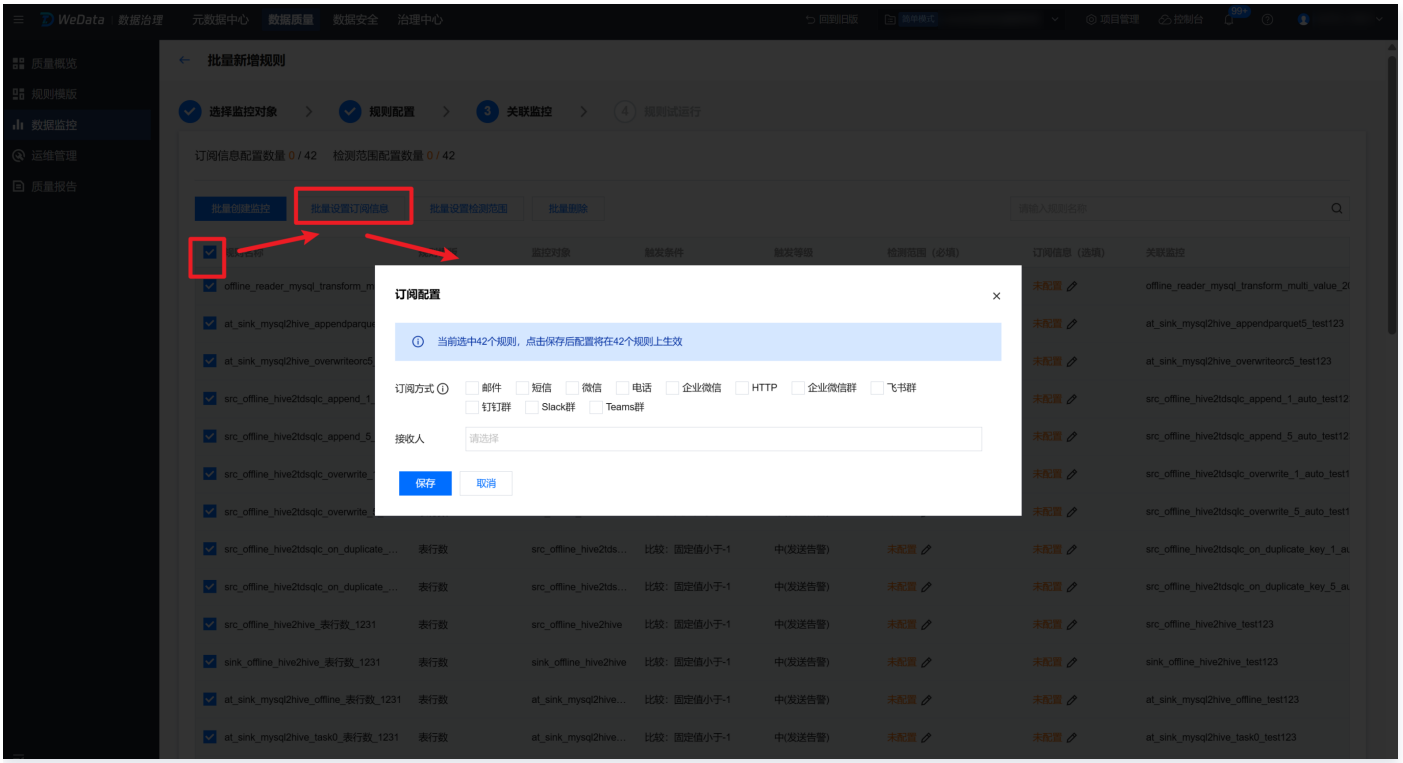


3. After creating a monitor, the selected rule will be automatically associated with the created monitor. The created monitor will be displayed in the **monitoring list**.

Step 4: Set Up a Subscription

Set up a subscription notification. When the running result of a quality detection task meets trigger conditions (i.e., the result is not as expected and the detection result is failed), determine what method to use to send notifications and whom to notify.

1. Enter the **Batch Set Subscription Information** interface to batch set subscription information.
2. On the pop-up subscription settings page, select the **subscription method** you need, set the **recipient**. Click **Save**.



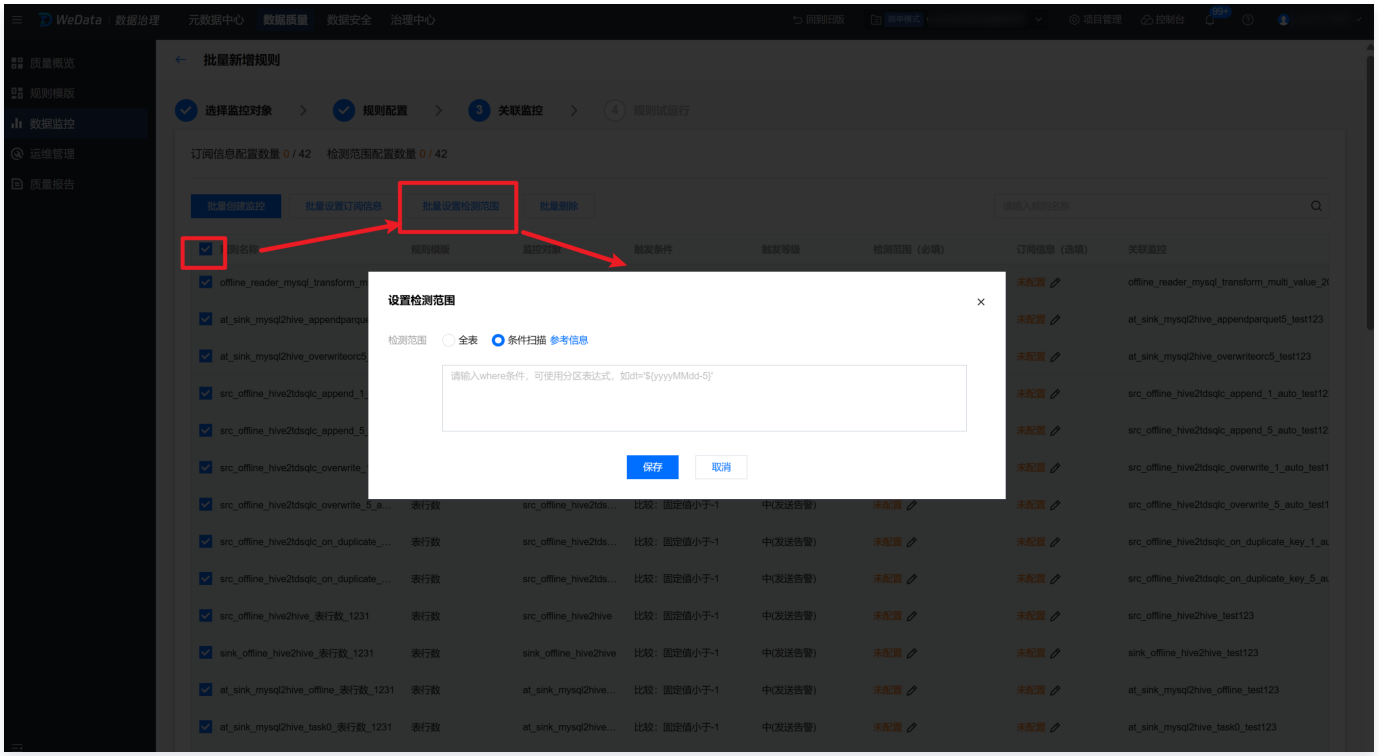
Step 5: Set Detection Scope

1. Click **Batch Set Detection Scope** to enter the **Set Detection Scope** interface and just batch set the detection scope.
2. Add the following information on the pop-up page.
   - Here you can select conditional scan and full table scan. It is recommended to choose **conditional scan**.
   - Partition where conditions can be filled in, for example:

```
pt_date='${yyyy-MM-dd-1d}'
```

> **Note**
> Here, generally fill in the partition field to avoid a full table scan during each quality task, which would result in a waste of computational resources.

In SQL, ***${yyyy-MM-dd-1d}*** is a date variable representing one day before the execution date and will be replaced with a specific date during Quality Task Execution.
For example: When the quality task is executed at 2024-05-02 00:00:00, ${yyyy-MM-dd-1d} will be replaced with 2024-05-01.



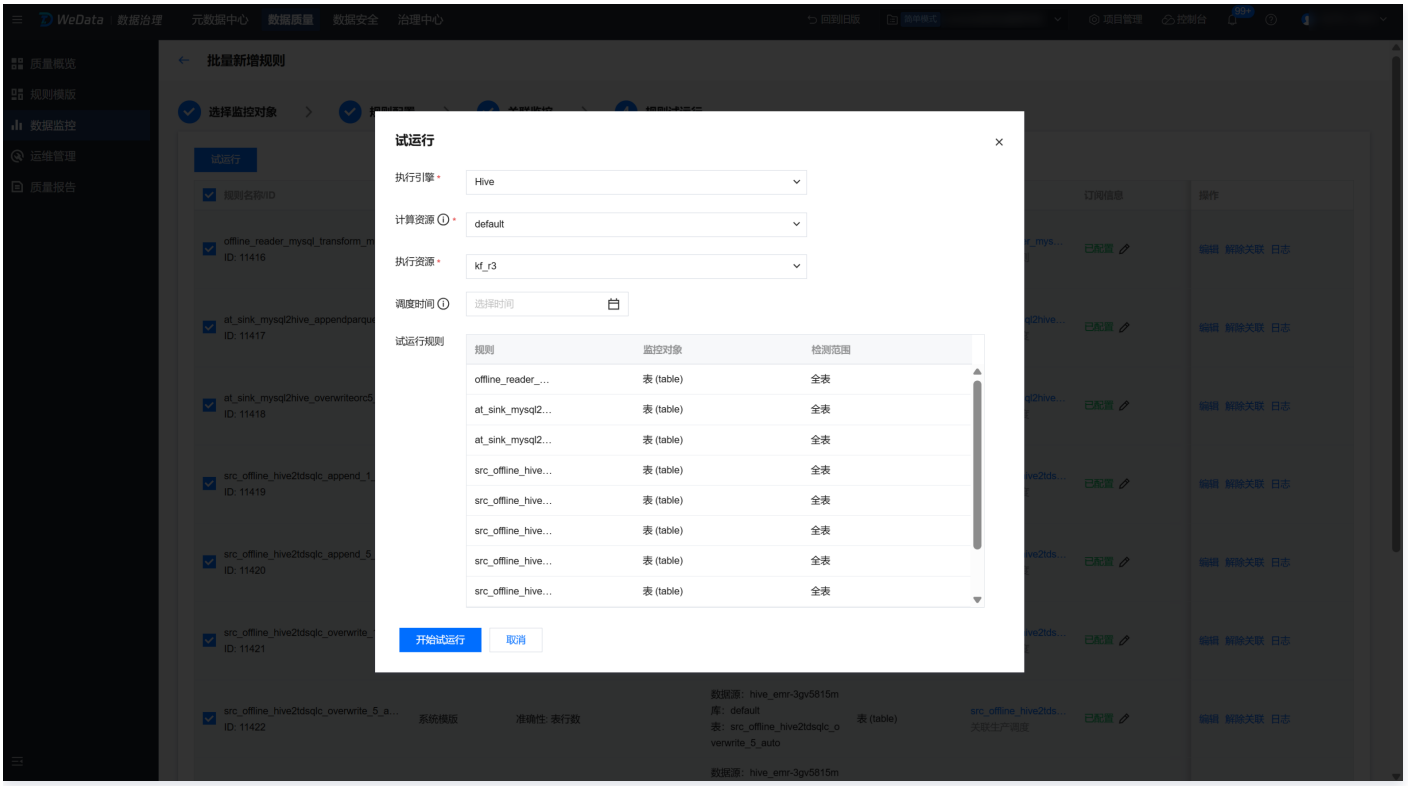
The replacement logic of specific time variables can be found in time parameter explanation.

Step 6: Trial Run

1. Selection rules. Click **Trial Run**.
2. Modify the scheduling time and click **Start Trial Run**. After the trial run ends, you can click **Click to View Operation Results**.

> **Note**
> The scheduling time filled in here will modify the partition time variable.
> For example, if you fill in 2024-05-02 here, ${yyyy-MM-dd} represents 2024-05-02, and ${yyyy-MM-dd-1d} represents 2024-05-01.

©2013-2025 Tencent Cloud. All rights reserved.



©2013–2025 Tencent Cloud. All rights reserved.

Investigate quality issues
Feature Overview

Last updated: 2024-12-16 16:39:00

Feature Overview

Identifying quality issues is an important value embodiment of the data quality module. In this module, you can:

1. View quality task execution results.
2. View historical alarm information.

Core Capabilities

View quality task execution results

- Support viewing the running status and results of quality tasks.
- Support quickly viewing data lineage to trace data sources and destinations.
- Support exporting instance operation results to find problematic data.

View alarm information

- Support viewing sent alarm records.
- Support tracking the quality rules that triggered the alarms.

Preconditions

Quality detection tasks have been configured and triggered successfully.

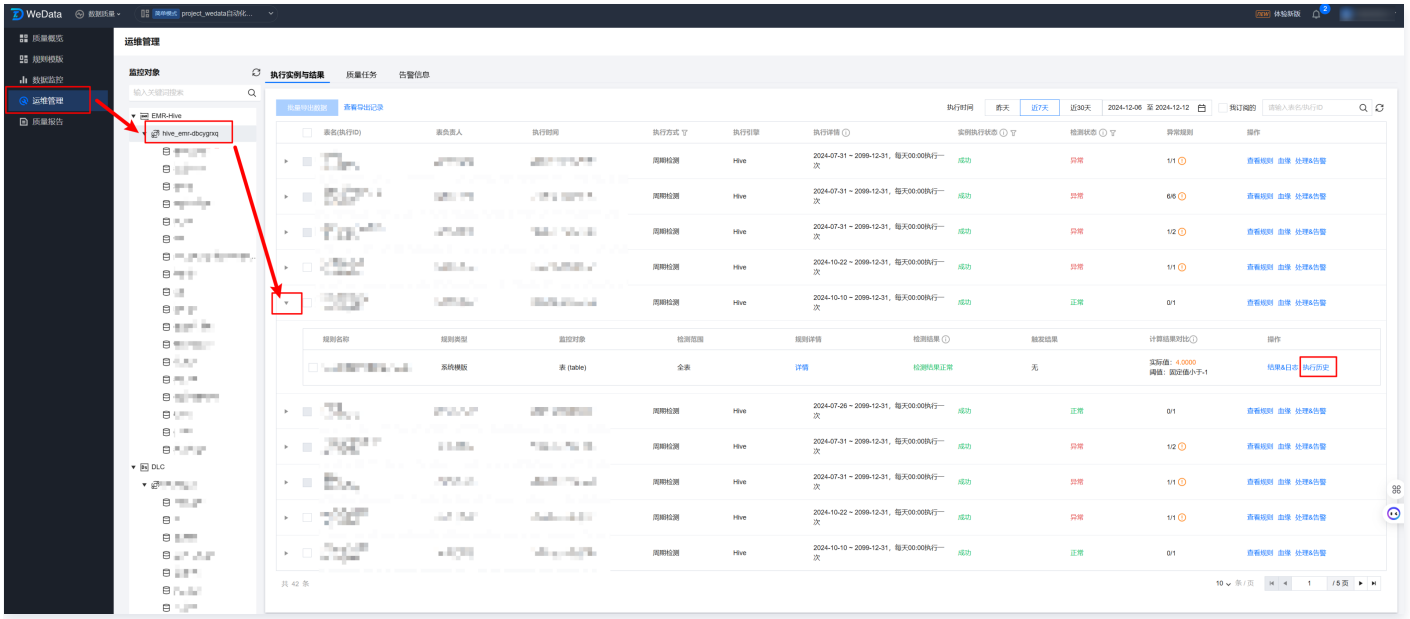
©2013-2025 Tencent Cloud. All rights reserved.

View Monitoring Results

Last updated: 2025-04-27 15:11:30

Execution Instance and Results

Users can view the operation status of quality inspection tasks on the Operations Management – Execution Instance and Results page.
Users can filter quality inspection tasks by data source, database, execution time, and subscription status, and search tasks using table name, execute task ID, table owner, and other information.



Note:
In the operations list, a detection task is uniquely determined by the table name + execution time + execution method. If the same table is detected multiple times, the latest detection status will be displayed.

Field	Details
Table name Execution ID	Table name Quality Inspection Task ID
Table Owner	Owner of the Quality Inspection Table
Execution Time	Quality Inspection Task Trigger Execution Time
Execution Method	Quality Inspection Task Trigger Method, including Trial Run, Periodic Detection, and Scheduling Detection (related to production scheduling tasks).
Execution Engine	Quality Inspection Task Execution Engine, currently supporting Hive, Spark, DLC, TCHouse-P, and Doris.
Execution Details	If the quality task execution method is periodic inspection, the set periodic inspection information (time range, execution frequency) will be displayed. If it is schedule-triggered, the associated schedule task ID will be displayed.

©2013-2025 Tencent Cloud. All rights reserved.

Instance Execution Status	The instance execution status is the execution status of the quality task instance issued to the scheduling side, including: 1. Waiting to Run: The scheduled task instance is waiting to run on the scheduling side. 2. Running: The scheduled task instance is running. 3. Successful: The scheduled task instance was successfully issued. 4. Failed: The scheduled task instance failed to issue (possible reasons: execution environment impact, user manually killed, business exception reported).
Detection Status	The detection status is a collection of all rule detection results within the current instance: 1. Normal: After all rules under the current instance are successfully executed, the rule detection results are normal. 2. Abnormal: The instance is considered abnormal if any rule is triggered. 3. Detecting: Rules under the instance are being executed. 4. Not Detected: Rules have not started execution or have not been executed successfully.
Abnormal Rules	Number of Abnormal Rules/Total Rules
Operation	• Supports redirecting to **Data monitoring** to view the table's rule configuration. • Supports redirecting to **Data Operation** to view lineage information in the table details for troubleshooting. • Supports handling & alarm, manually sending abnormal reminders.

View Rule Execution Situation

The operations management list supports expansion to view the triggering conditions of each rule:

View Rule Execution Results

Users can select and click the dropdown icon on the left side of the list to view rule detection task details (secondary list).

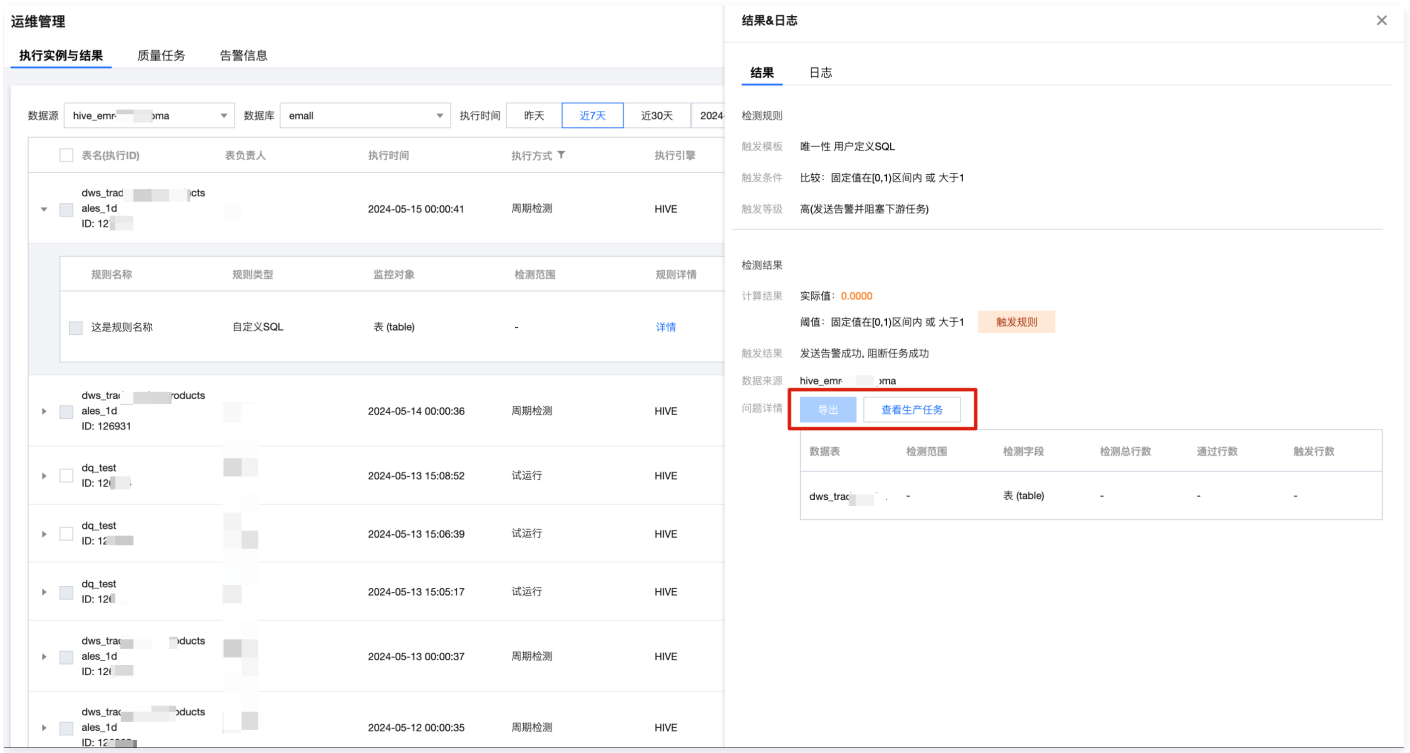


Field	Details
Rule Name	Configured Rule Name
Rule Type	System Template/Custom Template/Custom SQL
Data Object	Table-level rules (table) or table field information (Field Type)
Detection scope	Actual detection range (full table or set partitions)
Rule details	Rule configuration information, including trigger template, trigger conditions, and trigger levels.
Detection result	Divided into three statuses: detection passed, trigger rules, detection failed (due to other reasons causing the detection task to fail, not the data's quality).
Trigger result	Divided into three statuses: no alarm sent, alarm sent successfully, block task successfully (subscribed alarm recipients will receive alarm information through the configured channel).
Calculation result comparison	Detected result data/configured threshold, with anomalous result data highlighted in different colors. Multiple detection results are separated by commas.
Operation	Supports viewing rule task execution results & logs and execution history

Viewing results & logs

©2013-2025 Tencent Cloud. All rights reserved.

Click the **Results & Logs** in the secondary task list operation column to view the specific detection task information, supports exporting problem details, and jumps to the **; Data Operation** ; table details to **; Production and Changes** ; view production tasks.



Export single rule execution result

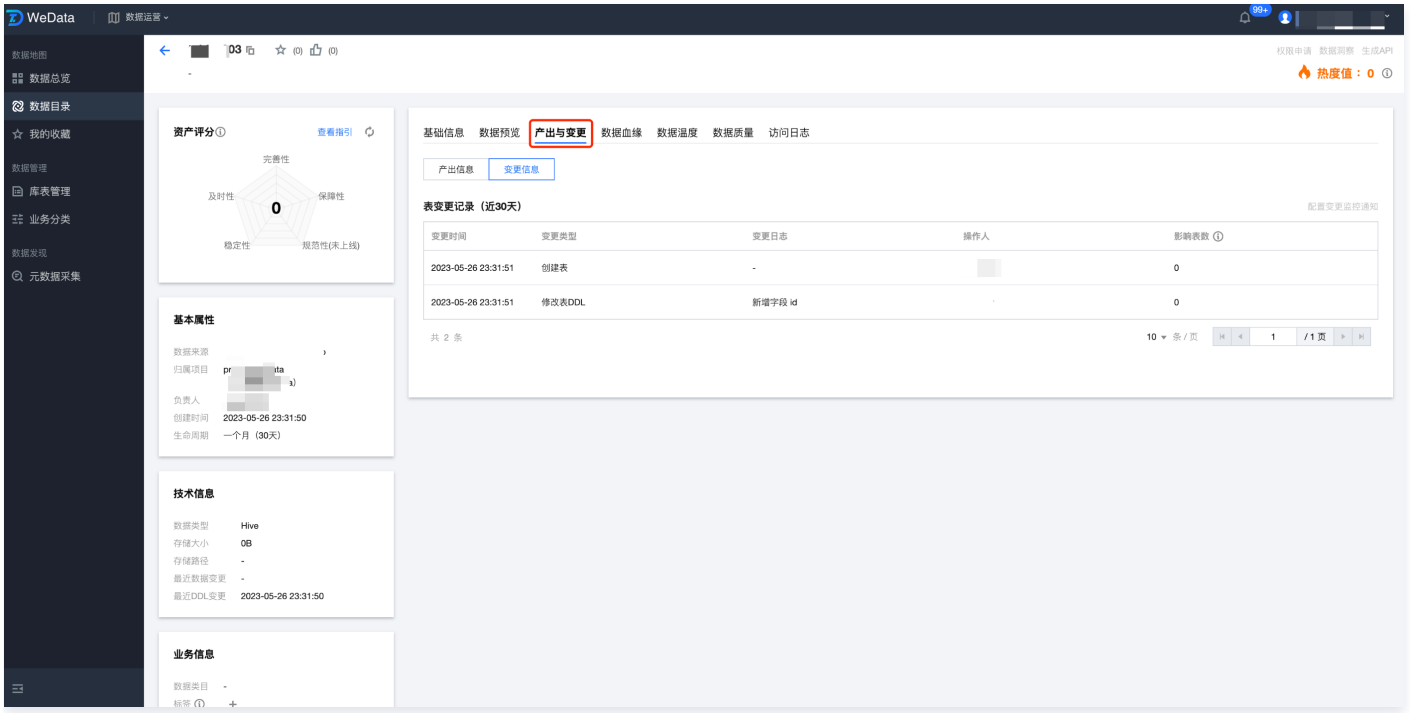
Enter the **Results and Logs > Results page,** go to the Export Data page, click **Export**.



Viewing production tasks

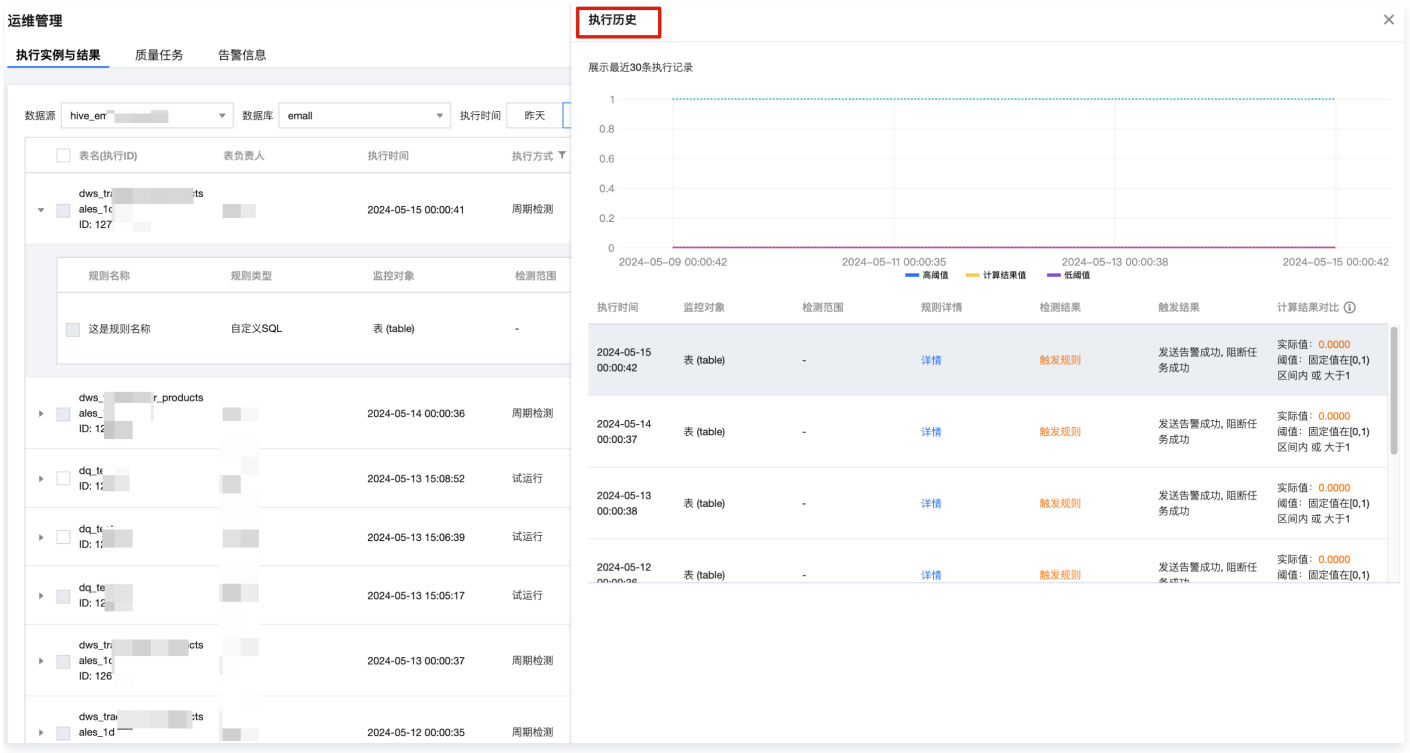
Enter **Results & Logs > Results page,** click **View production tasks.**

©2013–2025 Tencent Cloud. All rights reserved.



Viewing execution logs

Supports viewing execution logs, click **logs** to view the latest 30 task execution records

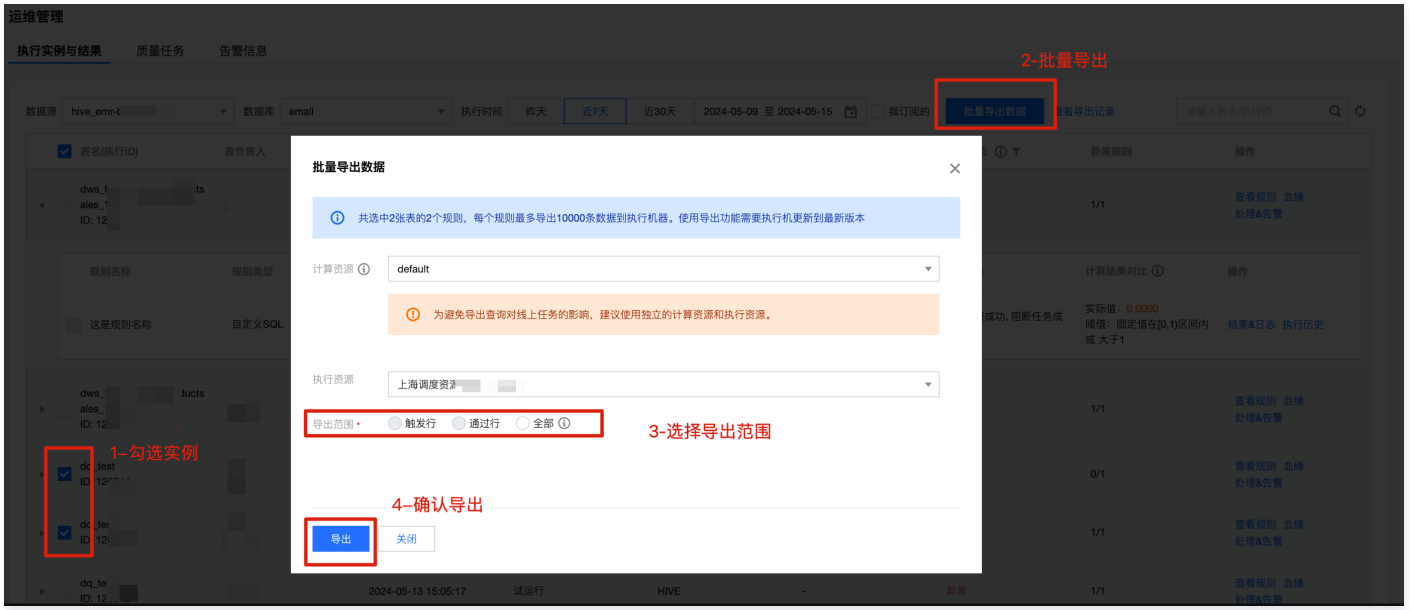


Bulk export execution results

Step One: Submit Export Task

Users can select execute instances and submit export tasks.

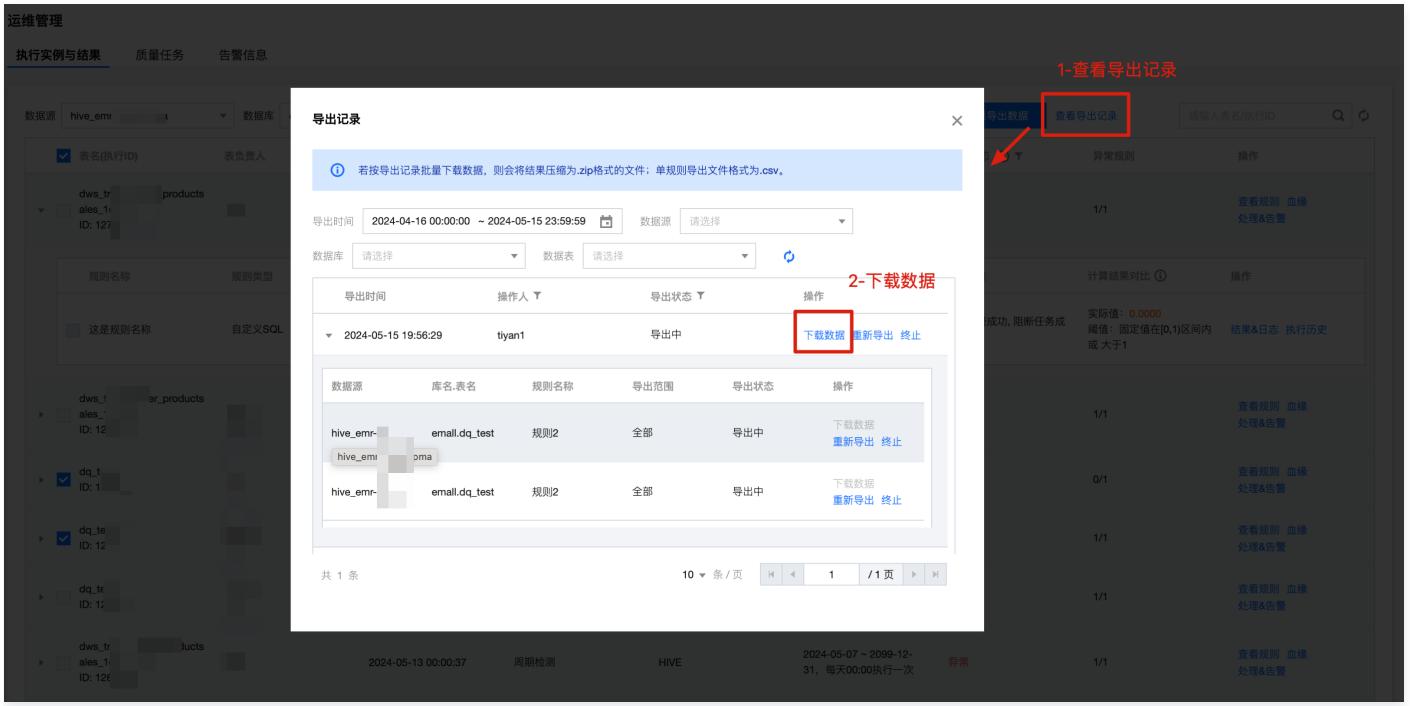
©2013–2025 Tencent Cloud. All rights reserved.



Step Two: Download Export Data

After submitting the export task and upon task completion (Export Status: **Exporting** > **Export Successful**), data can be downloaded.

> **Note**
> File Generation Logic: Each quality rule generates one file.



Configure Task Alerts

Last updated：2024-12-16 16:38:39

> **Note**
> During the operation of quality monitoring tasks, anomalies such as network, resource issues, link interruptions, or execution timeouts may occur.
> You can use this feature to configure alerts for quality tasks. When a quality task encounters an anomaly, the system will send notifications immediately.

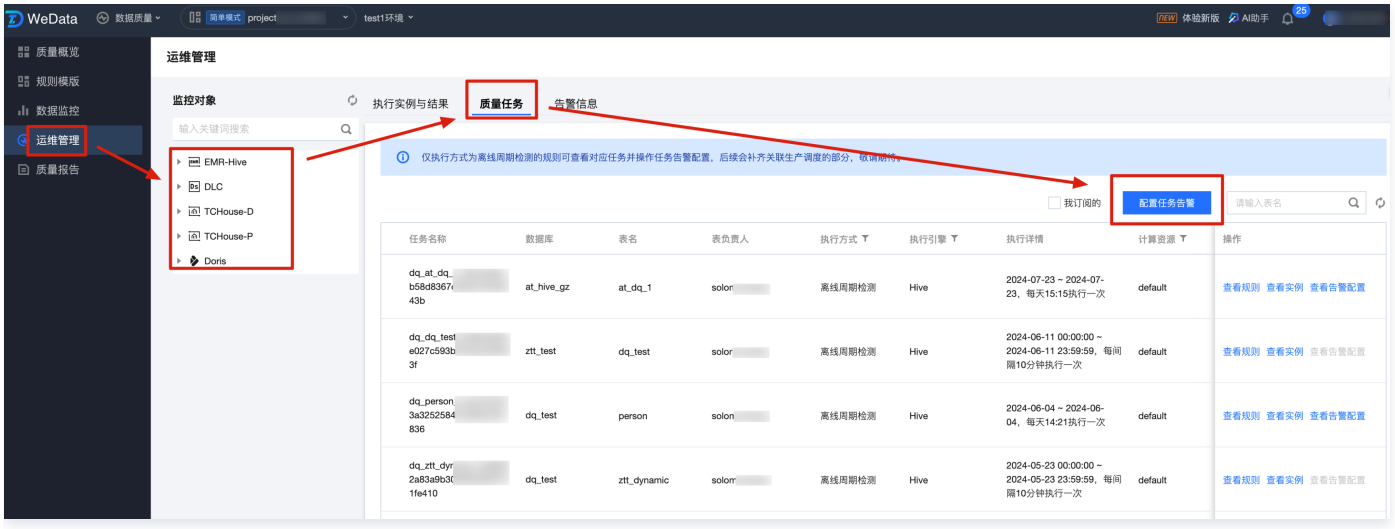
Configure Task Alerts

When an anomaly occurs during the operation of a quality task, such as a link interruption or execution timeout, an alert is sent to the person in charge, allowing them to promptly identify and address the issue.

> **Note**
> The states of the task operation include the following conditions:
> - Task Execution Failed;
> - Task Execution Successful;
> - Task Operation Timeout;
> - Task Completion Timeout;
> - Task Waiting for Scheduling Timeout.

Step 1: Enter the Quality Task List

Enter **Operations Management** > **Quality Task Interface,** click **Configure Task Alerts**, to access the Add New Alert page.

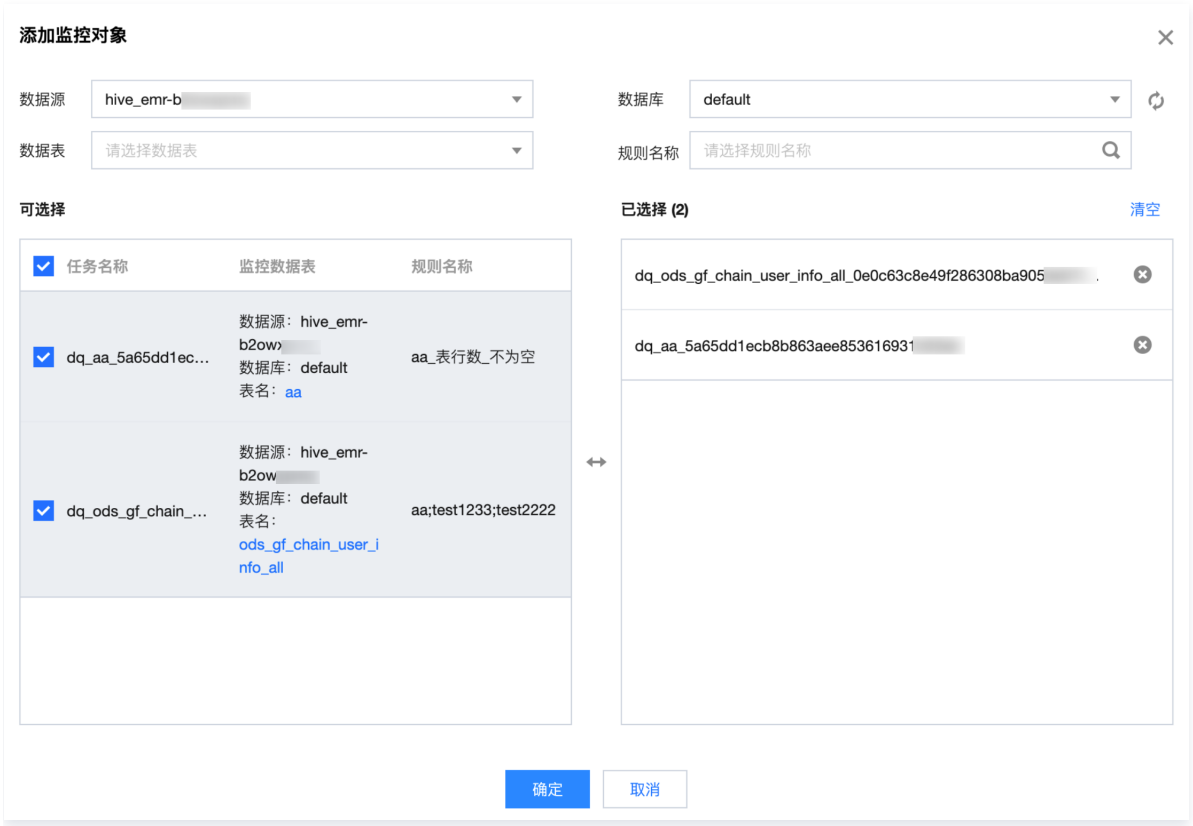


Step 2: Select Monitoring Objects

You can select monitoring tasks as monitoring objects. Check the tasks you need to monitor and click **OK**.

> **Note**
> You can select multiple monitoring tasks at once.

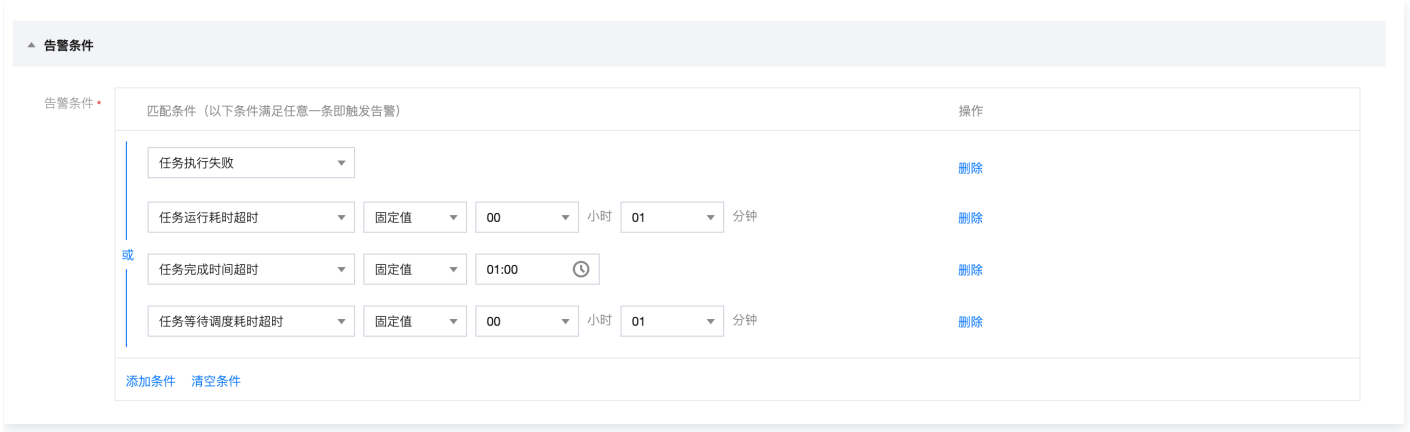
©2013–2025 Tencent Cloud. All rights reserved.



Step 3: Set Alarm Conditions

You can set alarm conditions: task execution failure, task running time timeout, task completion time timeout, task waiting for scheduling timeout.

> **Note**
> - Task running time and task waiting for scheduling time both indicate an alarm when the duration exceeds the threshold
> - Task completion time timeout triggers an alarm when the completion time point exceeds the threshold.



Step 4: Set Alarm Notification Recipients and Notification Methods

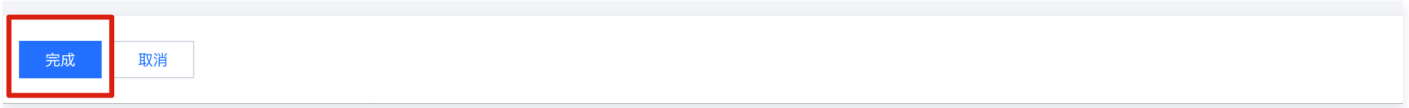
Here you can set the alert level, method, recipient, frequency, and do not disturb time.

©2013–2025 Tencent Cloud. All rights reserved.



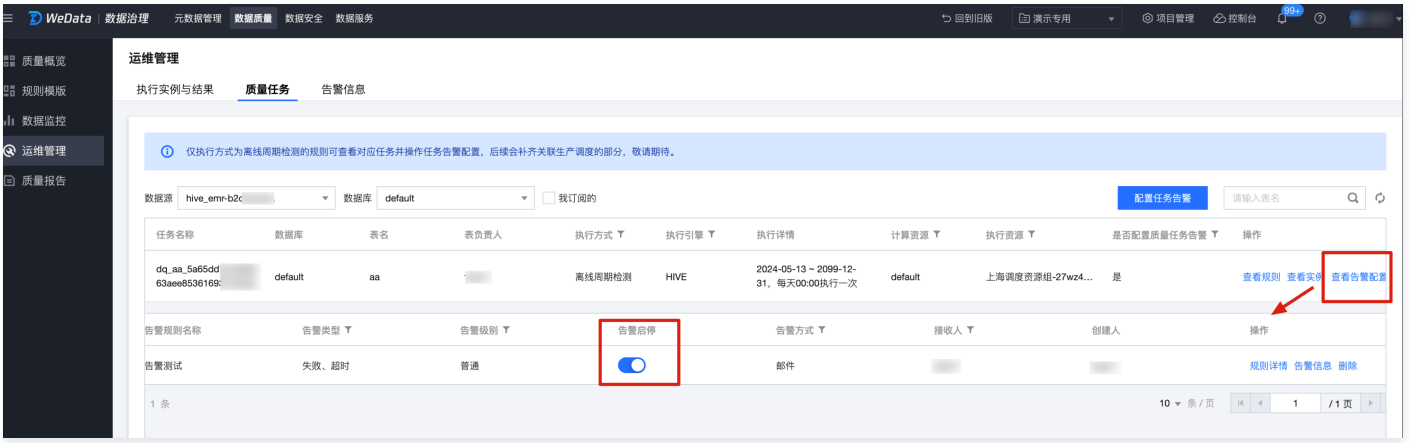
Step 5: Completion

After all content is set, click the **Done** button to submit the information.



Step 6: Alarm On-Off

1. Go to **O&M Management** > **Quality Task** interface. In the quality task list, click **Viewing Alarms** to view **Rule details** and **Viewing Alarms** in the operation column.
2. At the same time, you can view the status of alarm on-off and also enable or disable the **Alarm On-Off** button with one click.



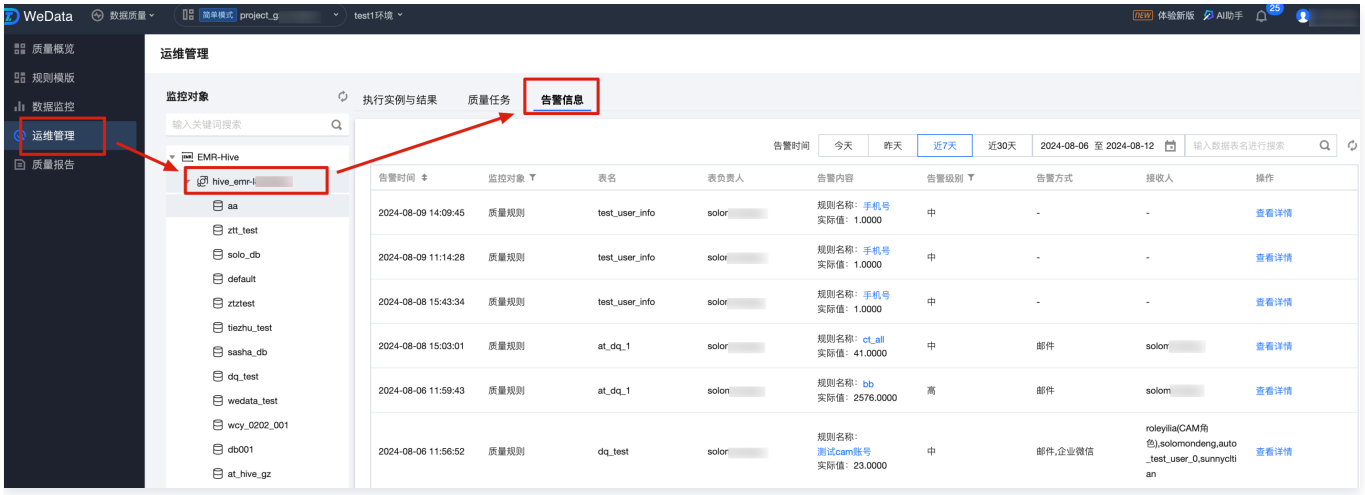
©2013-2025 Tencent Cloud. All rights reserved.

Viewing Alarms

Last updated：2024-12-16 16:47:33

View alarm list

Users can view alarm information on the **Operations Management** > **Alarm Information** page.
Users can filter alarm information by data source, database, data table, and alarm time.



View alert details

Click the **Alarm Details** button to view the alarm details. On the details page, users can see alarm reasons, sending status, and other detailed information.



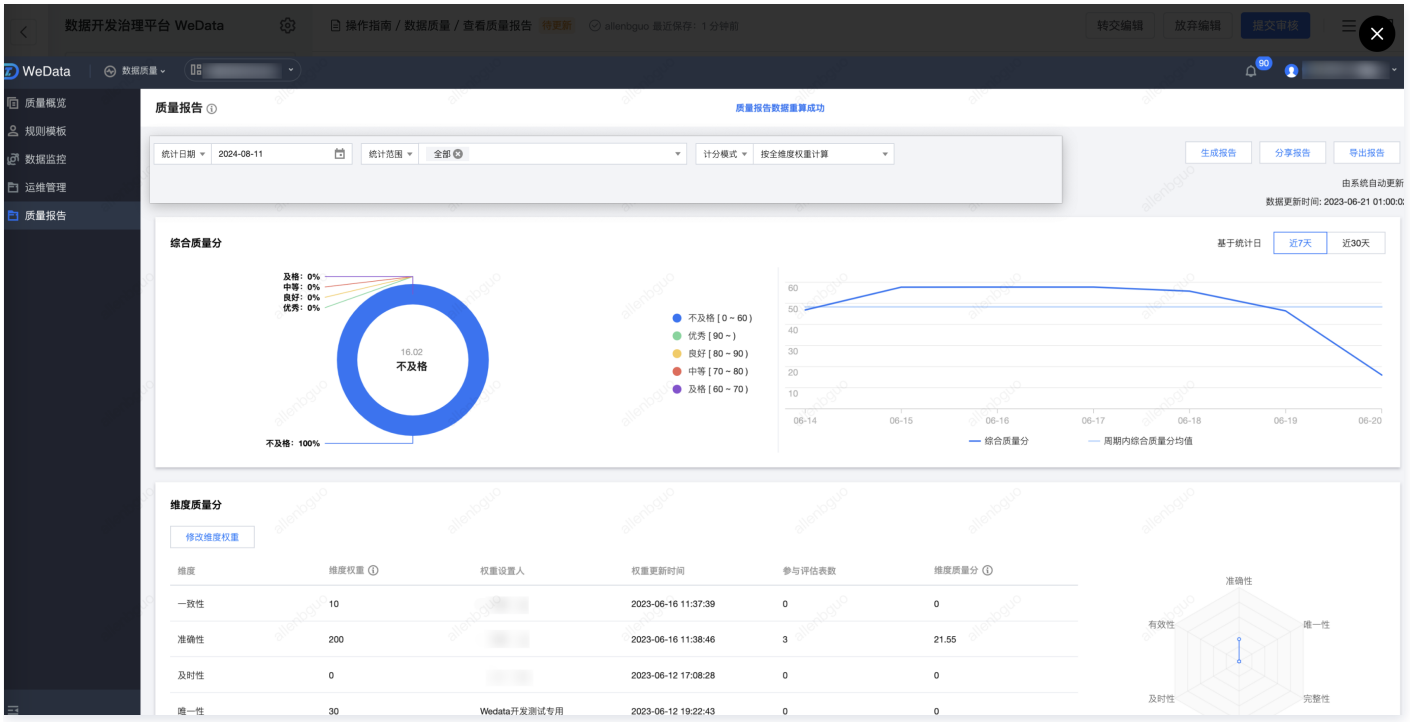
©2013–2025 Tencent Cloud. All rights reserved.

View Quality Report

Last updated: 2024-11-25 14:28:03

Based on a comprehensive quality scoring system, the Quality Report provides detailed scores across various dimensions, including Comprehensive Quality Score, Dimension Quality Score, and Table Quality Score details. Users can define the statistical scope (statistical date, data source, database, data table) and select the scoring model that meets their business needs to generate the Quality Report. It also supports one-click sharing and report export.

> **Note:**
> The Quality Report data supports automatic system updates (daily at 1:00 AM) and manual recalculations triggered by users after modifying dimension weights. The specific update methods and times will be notified.



Scoring Logic

All quality scores in the Quality Report are calculated based on the selected scoring model, using the **Rule Quality Score** as the basis. Calculations are made according to the user-defined filtering scope (statistical date, data source, database, data table) and the rules of dimension delineation, resulting in Table Quality Score, Database Quality Score, Dimension Quality Score, and Comprehensive Quality Score.

Dimension Quality Score	Scoring Model	Calculation Formula
Rule Quality Score	-	The pass rate of this rule (number of successful executions / total number of executions) x 100
Table Quality Score	Calculated based on full dimensional weights	∑(AVG(the same dimension rule quality scores associated with the table) * dimension weight) / ∑dimension weights
	Calculated based on configured dimension weights	∑(AVG(the same dimension rule quality scores associated with the table) * dimension weight) / ∑configured dimension weights
	Do not calculate based on dimension	AVG(quality scores of all rules associated with the table)

©2013-2025 Tencent Cloud. All rights reserved.

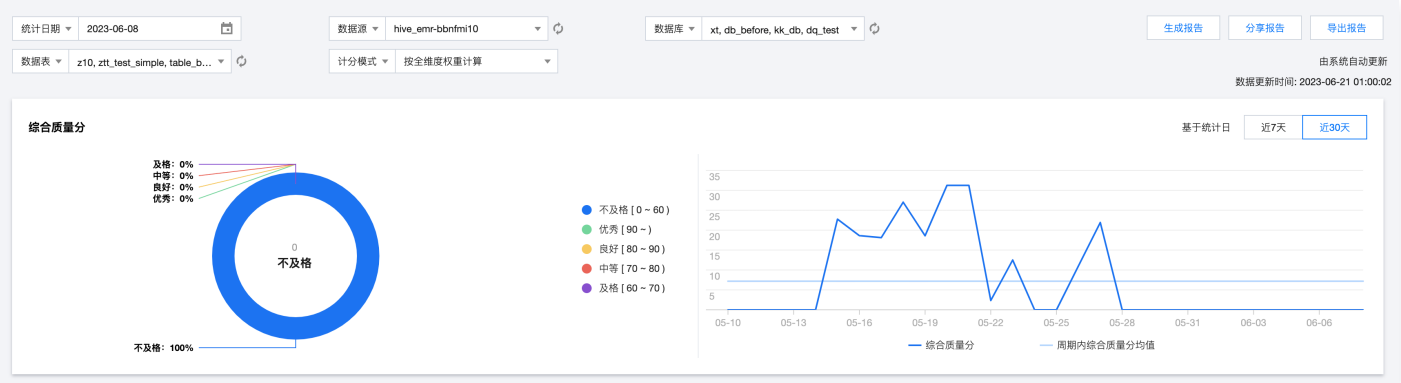
	weight	
Library quality scores	–	∑(table quality scores) / number of tables detected within the selected range
Dimension quality scores	–	∑(quality scores of tables within the same dimension) / number of tables detected in that dimension
Overall quality score	–	∑(table quality scores) / tables detected within the selected range

Note:
The rules associated with the table include both table-level and field-level configurations within the table. For library, dimension, and overall quality scores, the calculation formulas are the same under different scoring modes, but the table quality scores used for the calculations differ.

Specific details

- **Overall Quality Score**

Provides global quality score data, allowing users to view trend charts for the past 7 days and 30 days, as well as the average overall quality score within the period.



- **Dimension Quality Score**

Displays scores for six dimensions: accuracy, uniqueness, completeness, consistency, timeliness, and effectiveness. Supports users in defining each dimension's scoring weight. Information such as weight settings, weight setters, weight update times, and the number of tables involved in the assessment will be recorded and displayed.



©2013–2025 Tencent Cloud. All rights reserved.



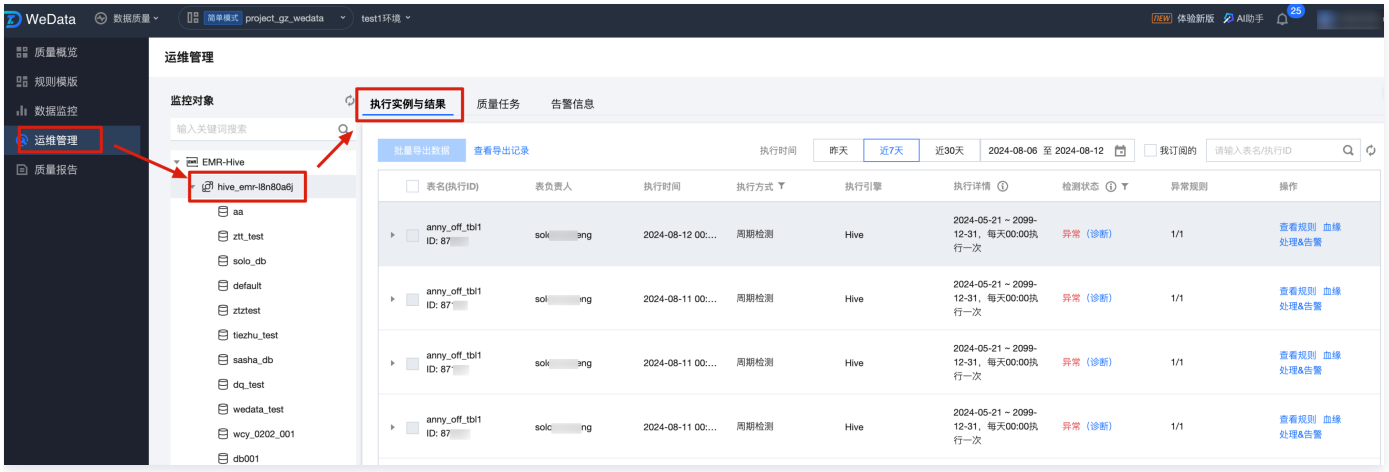
Note:

1. Only project administrators have permission to modify dimension weights. Modifications will take effect the following day by default. Users can also choose to manually trigger recalculations for immediate effect. Supports recalculating historical scores within the past 60 days using the modified dimension weights.
2. Dimension weights are global. When selecting the first two scoring models (all dimensions, configured rule dimensions), changes will impact Comprehensive Quality Scores, Dimension Quality Scores, Database Quality Scores, and Table Quality Scores after taking effect.
3. The report recalculation status will be synchronized. When the quality scores are being recalculated, the entry for modifying dimension weights will be grayed out until the recalculation is completed, supporting the modification operation again.

- **Table Quality Score Details**

Supports viewing detailed information on database and table quality scores.

Click on a specific table name to jump to the operations management interface for that table, where users can view the execution status of related quality tasks.



Click on a specific table quality score to view details of the associated rules, rule quality scores, detection pass situations, etc., allowing users to quickly locate issues.

©2013–2025 Tencent Cloud. All rights reserved.

规则名称	规则质量分	规则类型	规则模版	数据对象	检测情况（通过...
系统规则2_汇总值	0	系统模版	准确性:字段汇总值	age (int)	0 / 114
系统规则1_最大值	0	系统模版	准确性:字段最大值	id (int)	0 / 114
系统2_2	0	系统模版	准确性:字段平均值	id (int)	0 / 114
100测试	88.6	系统模版	准确性:字段最小值	id (int)	101 / 114

Also supports viewing the historical trend of table quality scores.



©2013–2025 Tencent Cloud. All rights reserved.

Data security
Introduction to COS Data Security Solution

Last updated: 2024-12-16 16:59:51

In today's digital age, data security has become one of the most concerned issues for enterprises and individuals. As a leading cloud computing service provider, Tencent Cloud offers a series of data security features to protect customer data from threats. This document will introduce the data security features provided by Tencent Cloud's WeData, offering a systematic data security system from asset inventory, risk assessment, and data protection to risk monitoring and operation, helping users understand and use these features to protect enterprise data.

Overall architecture

The overall architecture of WeData data security is as follows:



Note:

Dear customer, some features are not yet available. We greatly appreciate your attention and support to our product. Our team is working hard to ensure every detail is meticulously crafted and optimized. We will launch these features as soon as possible to bring a more comprehensive enhancement to your product experience. We look forward to witnessing the growth and progress of the product with you!

Feature Name	Feature Overview
Asset Inventory	WeData offers data asset security classification and grading features, including classification and grading configuration, classification and grading tasks, and sensitive data distribution features. With various built-in industry classification and grading templates, it ensures that enterprises can quickly identify and understand their data security status.
Risk Assessment (Not yet available)	WeData provides data risk assessment features, identifying, analyzing, and evaluating potential risks in data assets. The aim is to understand data security risks in advance to intervene and reduce the adverse impact of data on the business.
Data Protection	WeData offers various data security protection features, including data access control, static/dynamic data masking features, ensuring the security control of data usage and access on the platform.
Risk Monitoring (Not yet available)	WeData provides end-to-end risk monitoring, including platform operation log auditing, data access auditing, and user behavior analytics features. Through post-analysis and tracing, potential data

©2013-2025 Tencent Cloud. All rights reserved.

	security risks can be identified.
Continuous Operation (Not yet available)	WeData provides risk handling and alarm mechanisms, including risk handling notifications, risk handling alarms, and risk handling reports. Through continuous monitoring and addressing of risk issues, we continuously improve enterprise data security issues.

©2013–2025 Tencent Cloud. All rights reserved.

Safe Mode

Last updated：2024-12-16 17:00:16

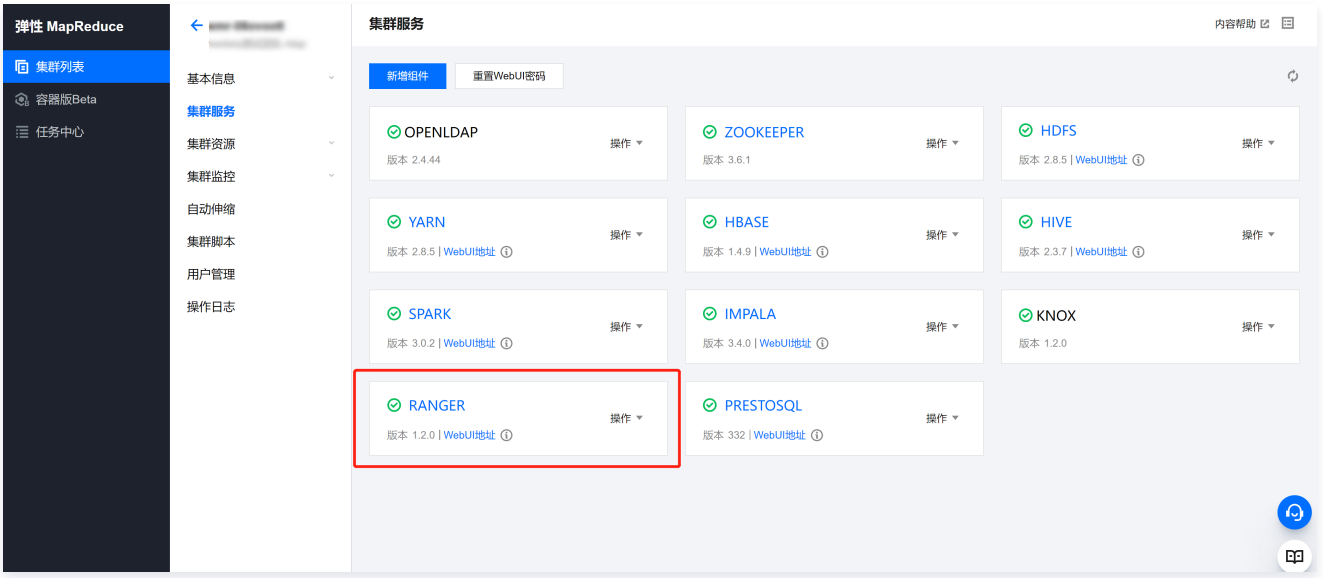
Users can choose to use WeData in either safe or non-safe mode.
In safe mode, Tencent Cloud WeData integrates with the Ranger permission policy system, achieving accountability and table-level permission control capabilities. It also provides channels for permission application and approval, secure data access control, and full-process user data security from pre-approval, real-time interception, to post-event auditing, building a security barrier. However, if users choose not to enable safe mode, they will not enjoy the aforementioned data security services.

Enable Safe Mode

Steps for enabling WeData safe mode are as follows:

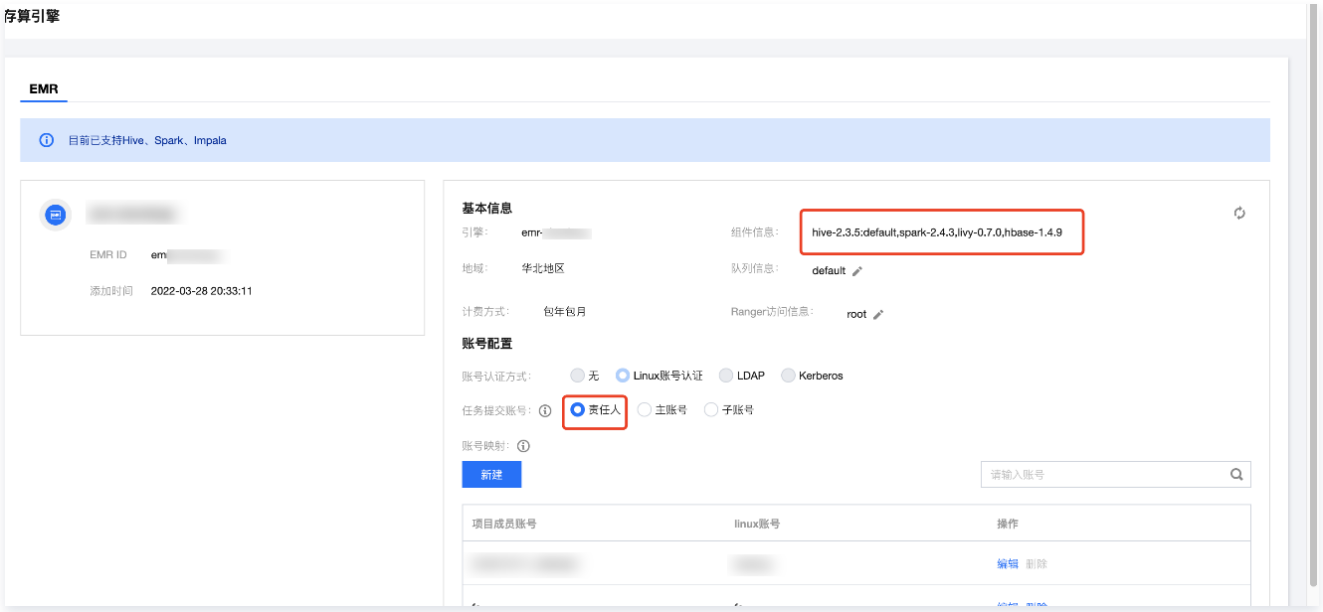
1. Enable the Ranger component in the EMR engine.
   Users need to select the Ranger component when purchasing the EMR engine and enable the Ranger service in the EMR Console.

> **Note**
> The Ranger component is not mandatory by default and requires the user to manually select and add it.



2. When configuring the EMR computational storage engine, select the responsible person for the task submission account. For details, please refer to Computational Storage Engine Configuration.

©2013-2025 Tencent Cloud. All rights reserved.



As shown in the figure, only when the EMR engine configured for the current project is purchased and the Ranger component is enabled, and the task submission account selects the responsible person, the project is in safe mode and can use data security services.

Non-safe Mode

Contrary to enabling safe mode, Tencent Cloud WeData is in non-safe mode if any of the following conditions are not met:

1. The Ranger service is enabled in the EMR engine.
2. The task submission account selects the responsible person.

©2013-2025 Tencent Cloud. All rights reserved.

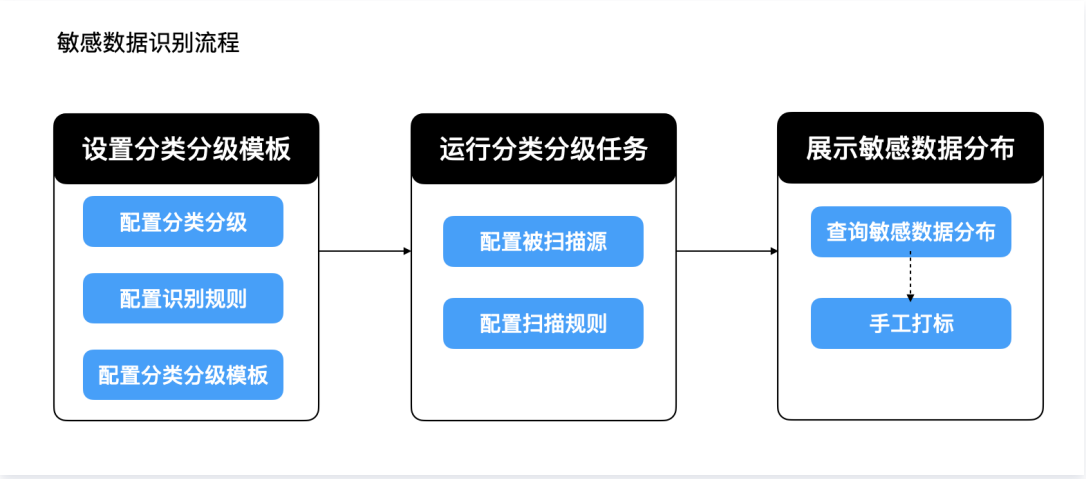
Asset Inventory Check
Classification and Grading Configuration

Last updated: 2024-12-16 17:00:46

Data classification and grading is achieved through a comprehensive analysis of the data. By categorizing data and applying different control strategies for classified data and sensitive data, fine-tuned management is realized. The combination of data security levels and sensitive data recognition rules allows for periodic checks of the data in the form of sensitive data recognition tasks, and outputs corresponding data sensitivity classifications and gradings. Manual marking of sensitive data is also supported.

Flowchart

The overall process is as follows:



Sensitive Data Grading

Provides users the Definition and management features for data security grading, establishing up to ten levels of data security.

1. Log in to Tencent Cloud and go to the WeData Console. Select your project and navigate to **Data Security** > **Classification and Grading Configuration** > **Sensitive Data Grading** page. Click **Create New Grade**.



2. Go to the **Create New Grade** interface, set the grading information, and click **OK**.
   - Grade Name: The name of the sensitivity level can be defined by the user.
   - Grade Description: Briefly describe the grade.
   - Grade Details: Sensitivity levels can have up to ten levels and can be defined as needed.

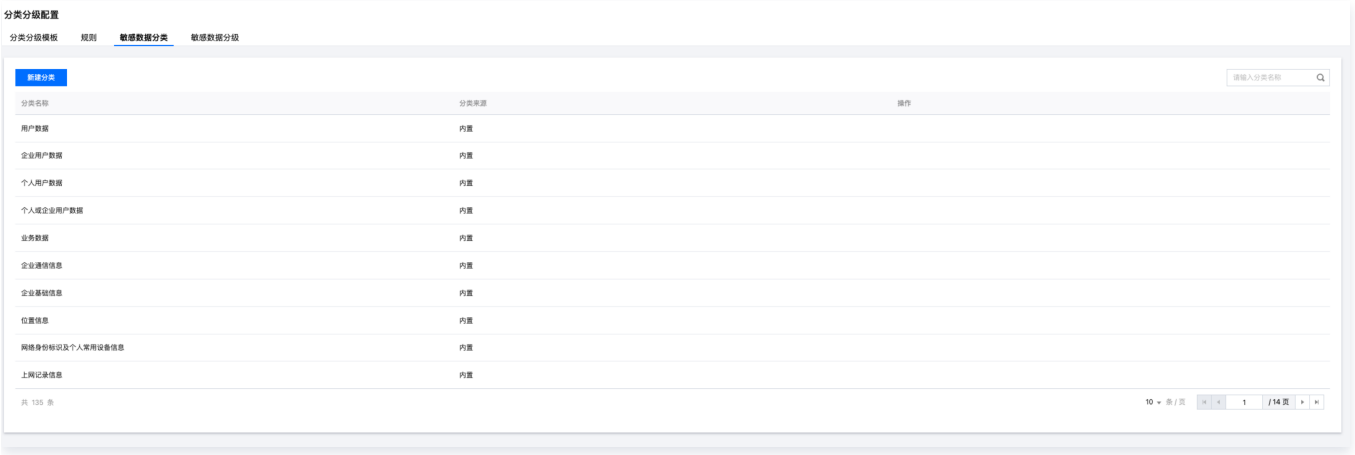
©2013-2025 Tencent Cloud. All rights reserved.



Sensitive Data Classification

Provides users the Definition and management features for data security classification, establishing classification dimensions and subcategories to adapt to different types and scenarios of business-sensitive data management.

1. Go to **Classification and Grading Configuration** > **Sensitive Data Classification** page and click **Create New Category**.



2. Go to the Create New Category interface, fill in the classification name, and click **OK**.

©2013-2025 Tencent Cloud. All rights reserved.



3. Each classification exists independently and can be combined into different multi-level classification hierarchies. You can freely use different templates and reuse classifications. In the new/edit template interface, you can freely combine different classifications in the full list to form multi-level categories.

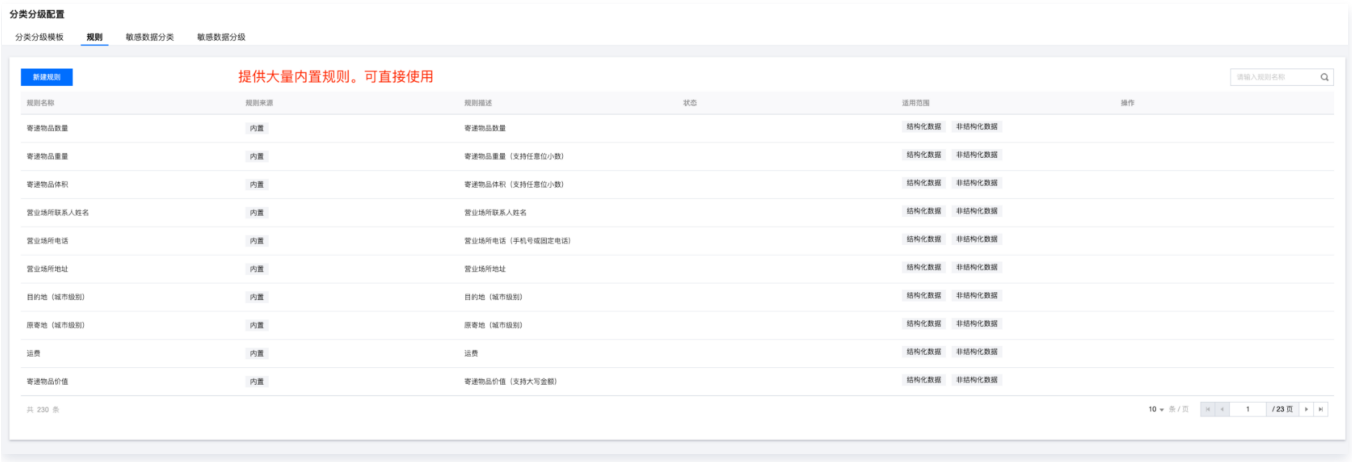


Recognition Rules Configuration

Identify sensitive data in data table fields, field descriptions, and data content by creating sensitive data recognition rules, mainly using regular expressions to establish rules. Combining data security classification with recognition rules allows classification and grading of identified sensitive data.

1. We provide system rule templates that can be directly used in data recognition tasks. Enter **Classification and Grading Configuration** > **Rule** page, click **Create Rule**.

©2013–2025 Tencent Cloud. All rights reserved.



2. Enter **Create Rule** interface, fill in the basic rule information, click **Confirm**.



Information	Details
Status	Controls whether it is available in the template. Built-in rules are enabled by default and cannot be disabled.

©2013–2025 Tencent Cloud. All rights reserved.

Field Rules	Supports two recognition methods: keyword and regular expression: • **Keyword:** Identifies content by matching entered keywords. Supports full-text matching and case-insensitivity configuration:  • **Regex:** Recognizes content through regular expressions. After configuration is complete, you can enter test text to verify the recognition effect: 
Unstructured Data	Recognition rules for unstructured data support regex, keyword, and ignore characters recognition methods.

Classification and Grading Template

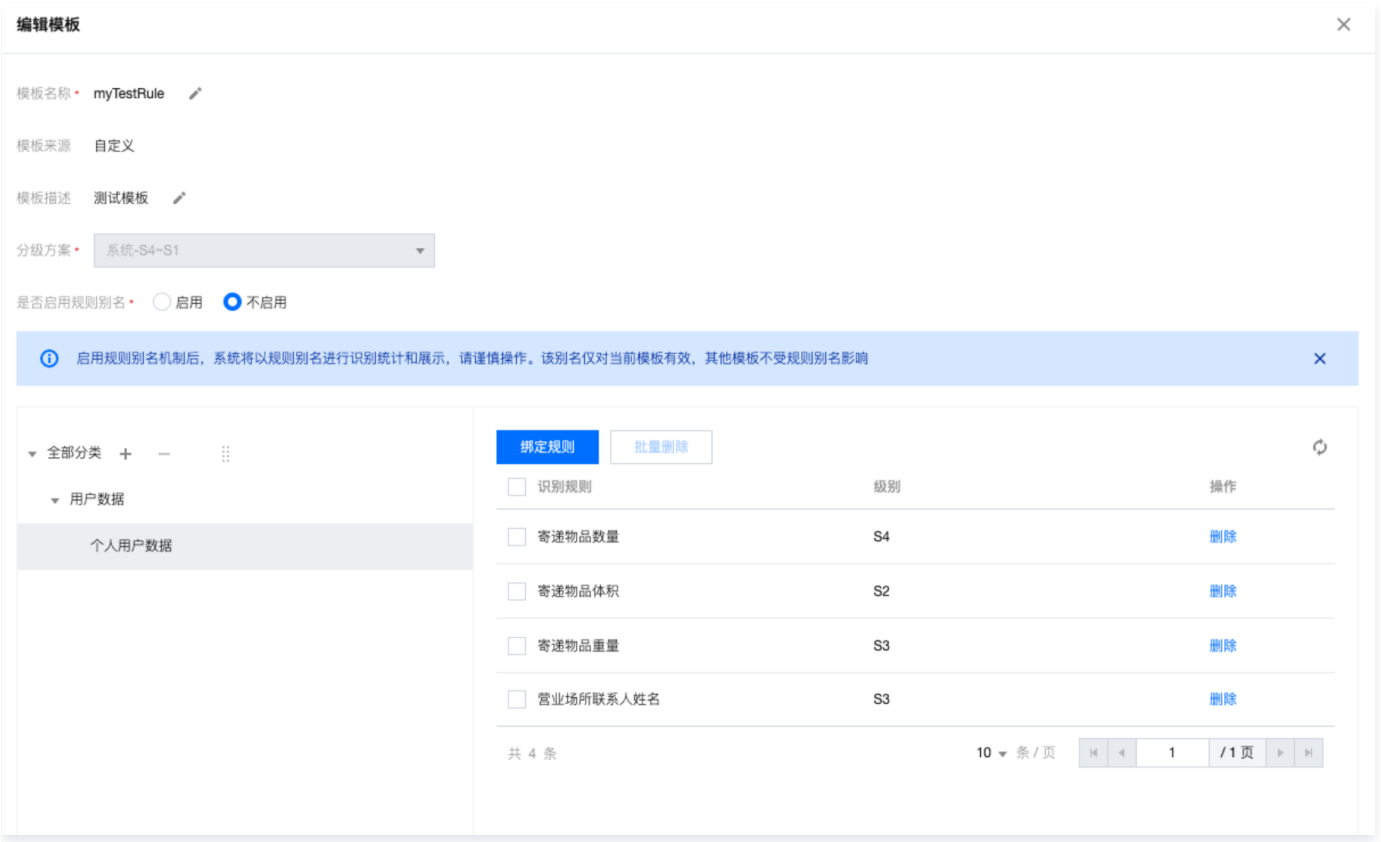
The classification and grading templates integrate classifications, gradings, and recognition rules to implement a sensitive data identification rule set for specific business scenarios. In WeData, data sensitivity identification is ultimately executed at the granularity of the classification and grading template. For convenience, five built-in classification and grading templates are provided, such as: JR/T 0197-2020 <Financial Data Security Data Security Grading Guide>.

1. Enter **Classification and Grading Configuration** > **Classification and Grading Template** page, click **Import Template**.



2. Enter **Edit Template** interface, you can define and edit the template yourself.

©2013-2025 Tencent Cloud. All rights reserved.



Information	Details
Template Name	Enter a template name, you can define it yourself.
Template Description	Briefly describe the template
Grading Scheme	Select the grading scheme.
Enable Rule Alias	Once the rule alias mechanism is enabled, the system will use the rule alias for recognition, statistics, and display. Please operate with caution. This alias is only effective for the current template and will not affect other templates. Core Scenarios: Mainly used for built-in rules, users can modify the name as needed

©2013–2025 Tencent Cloud. All rights reserved.

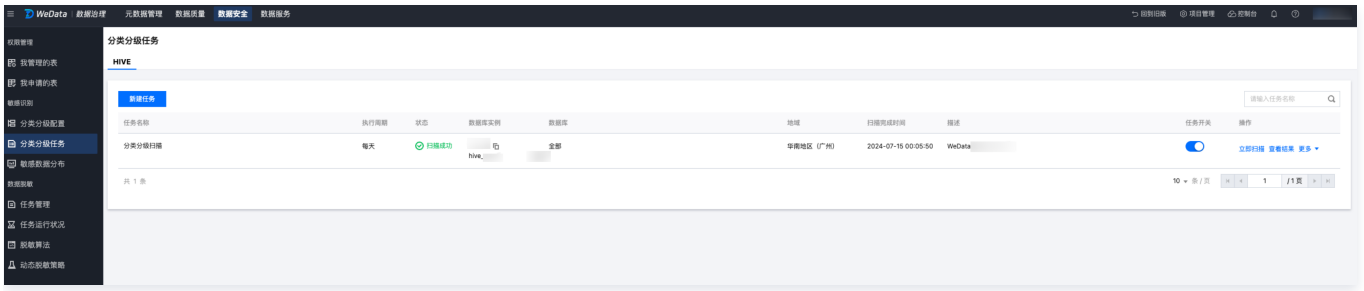
Sensitive data identification

Last updated：2024-11-25 14:32:15

Classification and Grading Task

Administrators can create and manage sensitive data identification tasks for data tables using pre-defined identification rules. When a rule is matched, the corresponding data table fields will be tagged for classification and grading.

1. Log in to Tencent Cloud and enter the WeData console, select the project and go to the **Data Security** > **Classification and Grading Task** page. Click the **Create New Task** button to create a new task.



2. In the **Create New Task** interface, fill in the basic information, configure the scan target and scanning schedule in the pop-up window, and click **OK**.

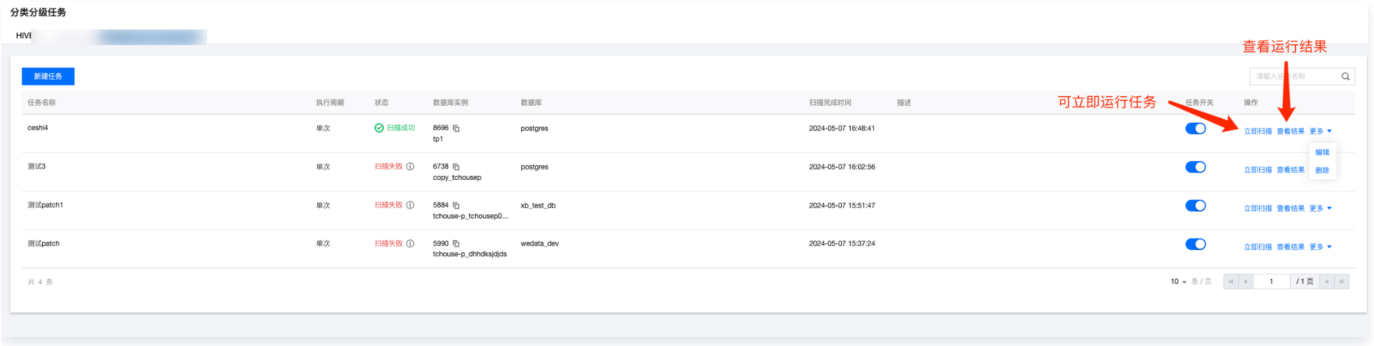


©2013-2025 Tencent Cloud. All rights reserved.

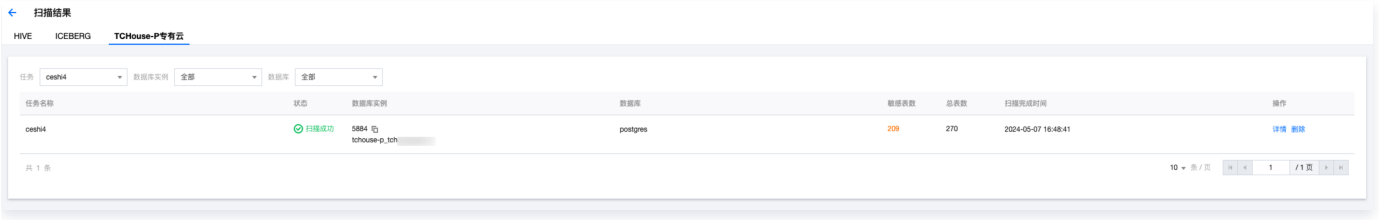
Information	Details
Task Switch	Used to control whether the classification and grading task is executed.
Select Data Source	Choose the data source to be scanned from WeData's data sources.
Database	If selecting a database not bound to sensitive data identification, it will occupy identification resources. If the limit is reached, new bindings cannot be made. You can extend resources or release resources to enable new bindings.
General Rule Set	When the switch is on, the "General Rule Set" template and the selected category grading template are used for scanning together. "General Rule Set" is the default rule, which can be modified in the category grading configuration.
Classification and grading template	Select the pre-configured classification and grading template to be used as the scanning rule.
Execution plan	• Immediate: Run the task scan immediately after configuration. The task runs only once. • Timer: Set the scheduled cycle for running the scan task. It supports execution cycles of single, daily, weekly, and monthly.

Viewing the Result

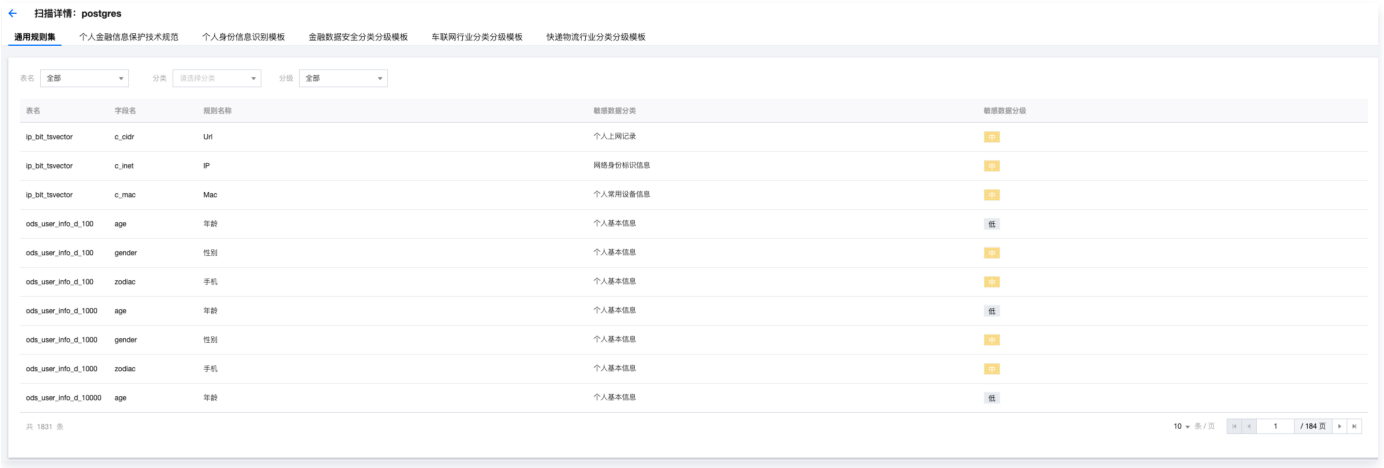
1. In the operation column, click **Scan Now** to run the task immediately. Click **View Results** to check the task execution results.



2. Entering the Result Details page, you can view the scan target and results.



3. Click **Details** to display the sensitive data identified by the recognition template.



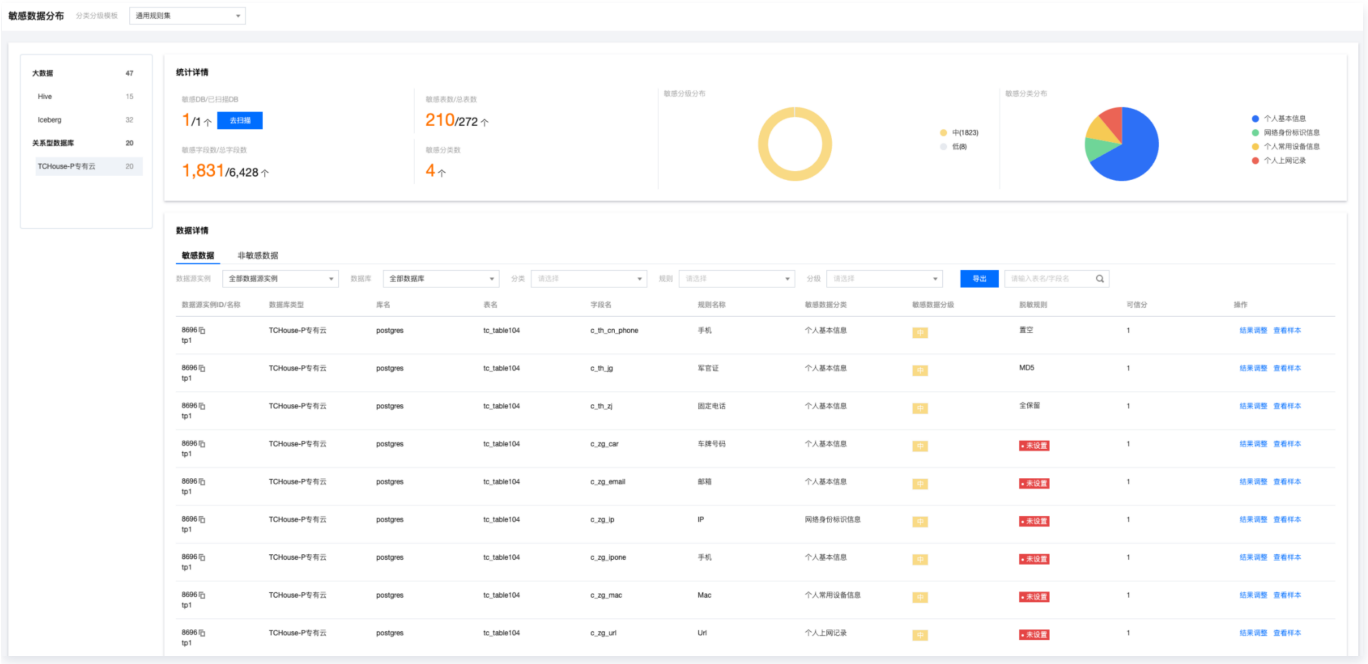
©2013-2025 Tencent Cloud. All rights reserved.

Sensitive Data Distribution

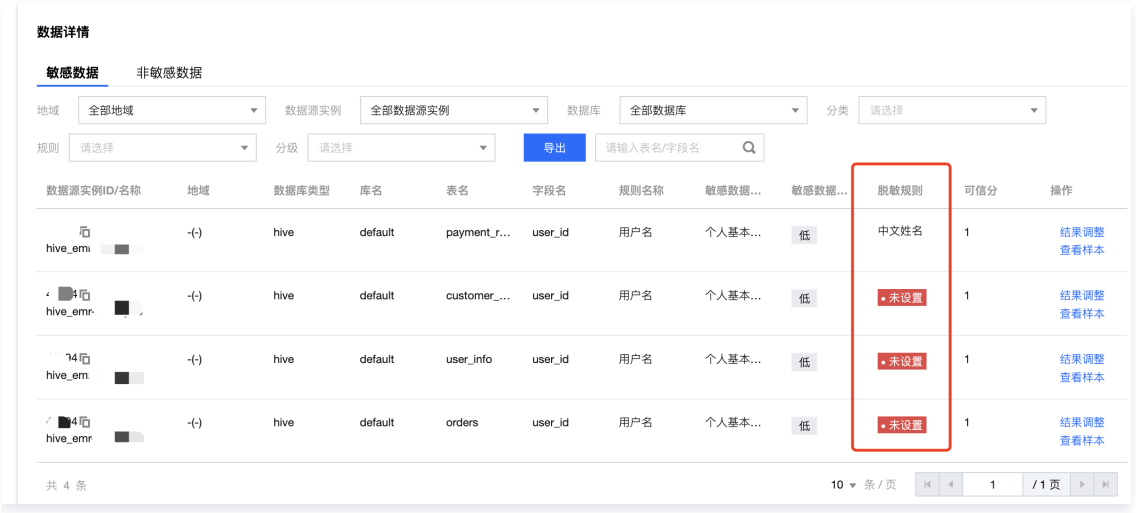
Displaying the distribution statistics and detailed information of all classified and graded sensitive data under the recognition task scan results, supporting manual correction.

Sensitive data

The sensitive data list shows the sensitive data discovered by the recognition task, including the database, table name, field name, recognition rules, security classification, security level, data masking rules, and operations for manual correction and viewing samples.



Click on the **Desensitization Rules** in the list 未设置 or the already configured rule name to enter the **Desensitization Rule Configuration** interface. Choose the **Release Method** to configure the desensitization rules.



©2013–2025 Tencent Cloud. All rights reserved.

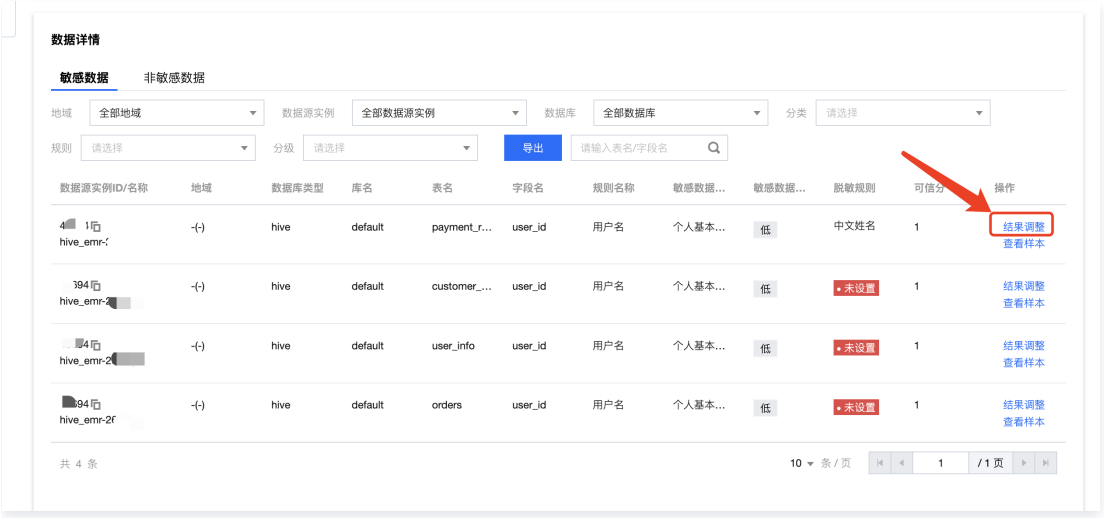


Information	Details
Publishing method	Strategies for processing historical data, for instance, if a running static data masking task uses Algorithm A for field columnA, and the algorithm for this field changes to Algorithm B, the following processing strategies apply: • "Incremental Release": The data masking task will continue to use Algorithm A for masking. If this field was not previously configured with a data masking algorithm, it will be changed to use Algorithm A after the release. • "Coverage Release": The field in this data masking task will use Algorithm B for masking, regardless of whether it was previously configured with a data masking algorithm. • "No Processing": The field in this data masking task will not undergo any processing, regardless of whether a data masking algorithm was previously configured.

Manual Tagging

Support manual classification and grading operations to adjust the results of sensitive data identification. Sensitive data can be changed to non-sensitive data and vice versa.

> **Note**
> Scanning the same template again will default to the manually corrected version.



Click the **Result Adjustment** operation to enter the "Manual Labeling" configuration page.

©2013–2025 Tencent Cloud. All rights reserved.



Information	Details
Set as Non-sensitive Field	Sensitive data can be changed to non-sensitive data and vice versa.

©2013-2025 Tencent Cloud. All rights reserved.

Data protection
Dynamic Masking

Last updated: 2024-11-25 14:32:42

Dynamic DMask

Scope of application

Support user-level data dynamic masking policy configuration. This masking policy is currently only for the WeData platform.
Applicable scope:

- After enabling dynamic masking, data query performance will decline.
- Dynamic masking currently supports the following scenarios:

Scenario name	Description
Data Development > Development Space	Display and download of task execution results.
Data Development > Orchestration Space	Display of task execution results.
Data Development > Data Management	Data preview.
Data Service > Service Development	Display of API call return results
Data Operations > Data Directory	Data preview

Enable Dynamic Masking

Log in to Tencent Cloud, go to the Data Development Governance Platform WeData Console, then navigate to **Project Management > Basic Information Configuration > Data Security Configuration** page. Manually enable the dynamic masking mode switch; the dynamic masking strategy needs to be enabled to take effect.

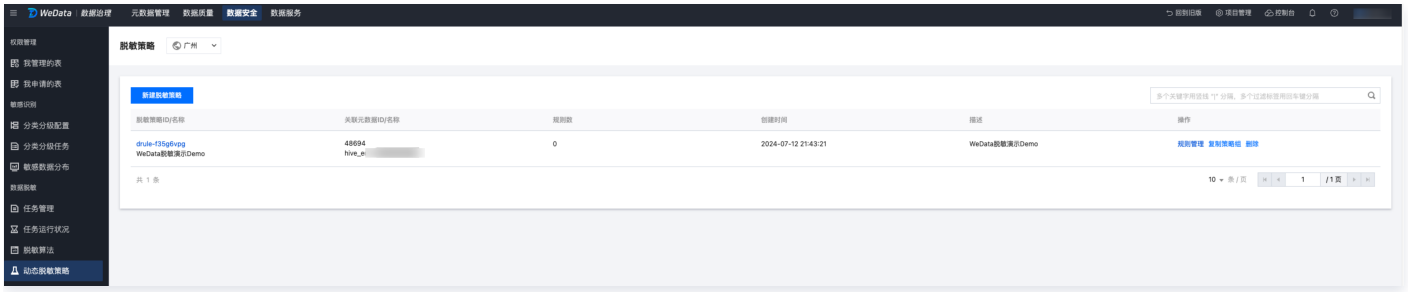
©2013-2025 Tencent Cloud. All rights reserved.



Desensitization Strategy Configuration

Step One: Create a New Desensitization Strategy

1. Select **Data Security** > **DMask** > **Dynamic Masking Strategy** to enter the masking strategy page, then click **Create Masking Strategy**.



2. Enter the Create New Desensitization Strategy interface, fill in the following related information, click **OK**.
   - Desensitization Strategy Name: Define the name of the desensitization strategy.
   - Associated Metadata: Used to specify the data source that needs dynamic desensitization. When a data source that is not bound is selected, it will occupy dynamic desensitization resources. When the limit is reached, no more bindings can be made. New bindings can be achieved through resource expansion or releasing resources.
   - Description: Briefly describe the desensitization strategy.

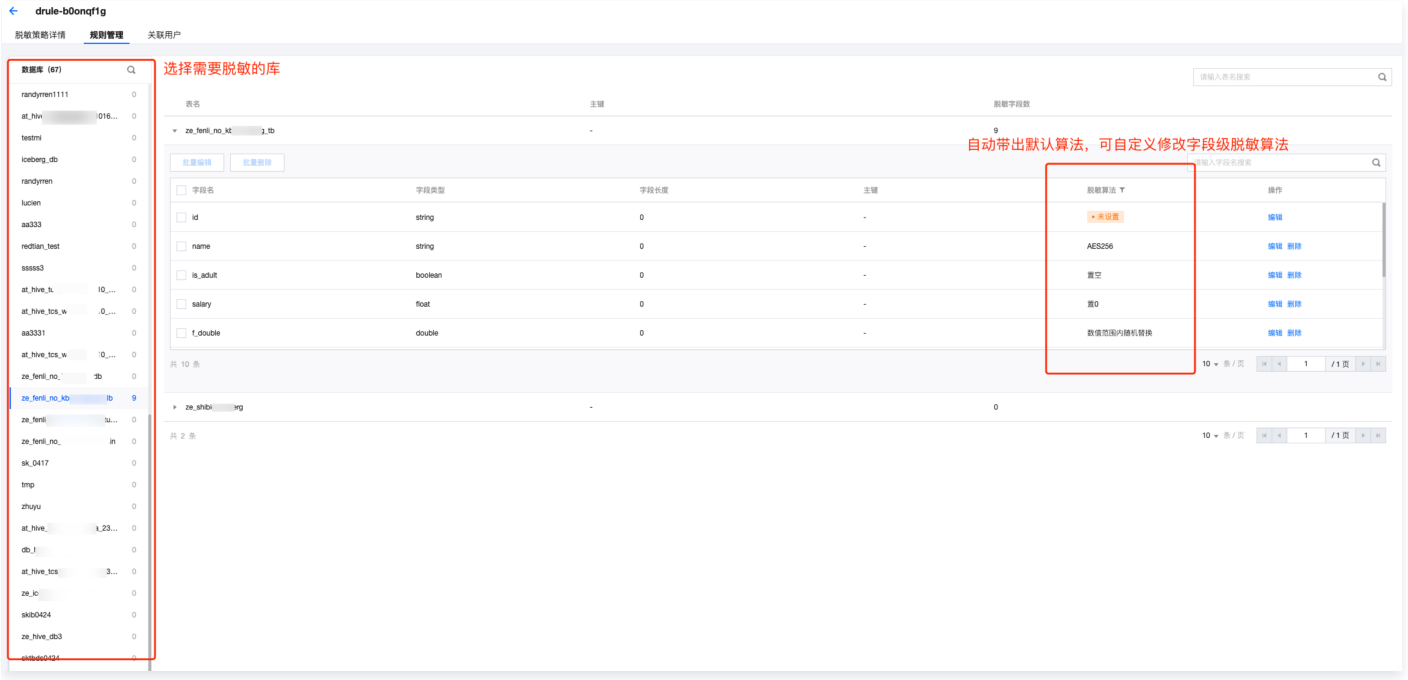
©2013–2025 Tencent Cloud. All rights reserved.



Step Two: Manage Desensitization Algorithm

Go to the rule management page, you can manage the desensitization algorithm in the operation column. You can edit and delete the current field's desensitization algorithm.

- click **Edit** to modify the current field's desensitization algorithm. By default, the desensitization algorithm configured on the sensitive data distribution page will be automatically displayed, and modifications are supported.
- click **Delete** to clear the current field's desensitization algorithm.



Step Three: Configure Associated Users

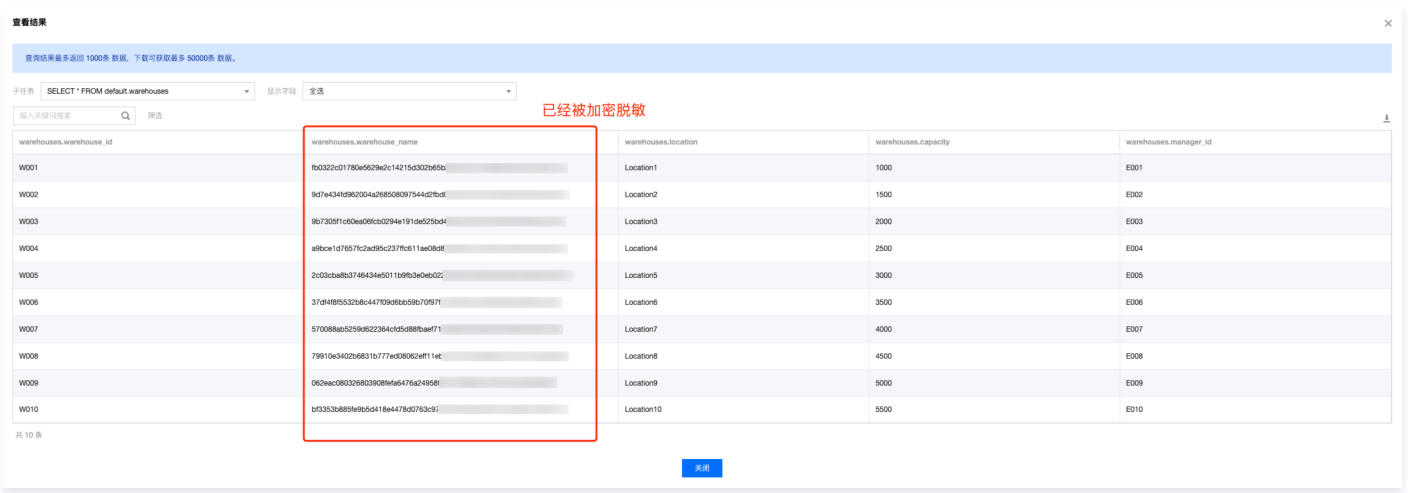
Go to the associated users page, click the **Add** button to associate users.



Step Four: Verification

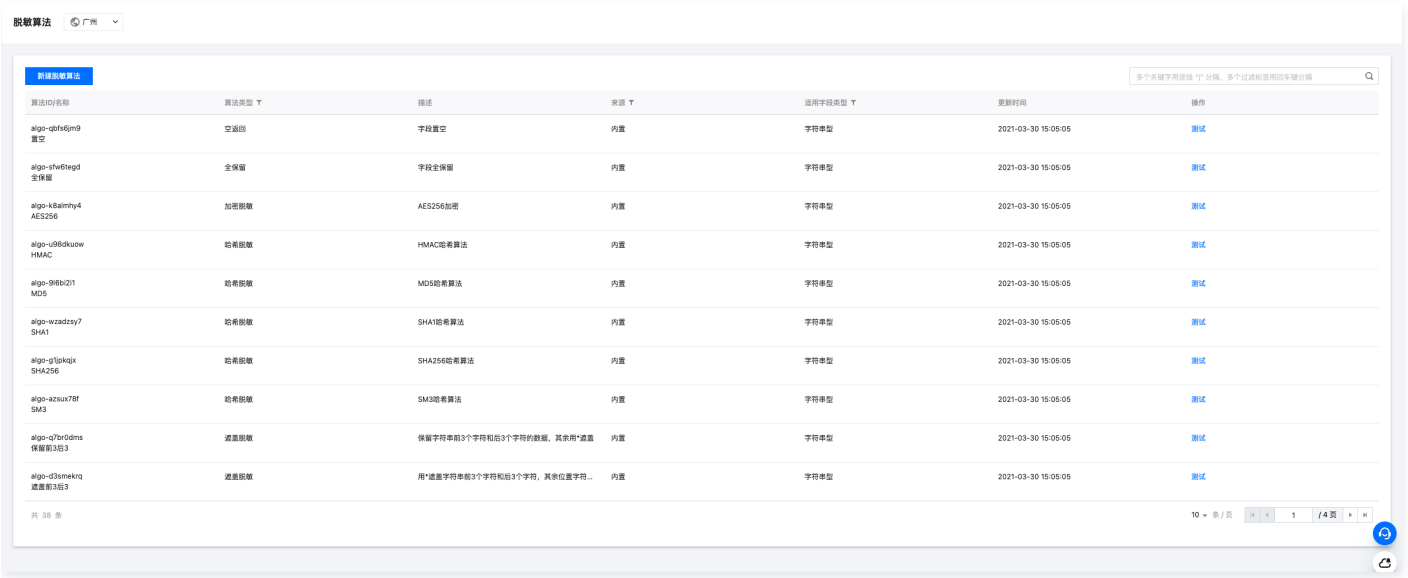
©2013–2025 Tencent Cloud. All rights reserved.

Enter the view result page to test and query data.



Desensitization Algorithm Configuration

1. Provide the configuration and management of desensitization algorithms, which can be used for static/dynamic desensitization strategies. Various built-in algorithms can be used directly (some national cryptographic algorithms are supported, such as SM1, SM3, etc.)



2. Enter the **Desensitization Algorithm** interface, click **Add Algorithm** to add a new algorithm by definition.

©2013–2025 Tencent Cloud. All rights reserved.



Configuration Item	Description
Basic Information	• Algorithm Name: Enter the algorithm name, can be defined. • Source: Defined. • Description: Briefly describe the algorithm.
Algorithm Configuration	Algorithm Type: Currently added algorithms support types including Encryption Masking, Hash Masking, Cover Masking, Transformation Masking, Replacement Masking; select an algorithm and adjust its parameters to define according to its characteristics.
Testing	Select an algorithm and adjust its parameters to define according to its characteristics. 1. Enter the data to be tested and verified in the raw data, click the Test button to display the algorithm running result in the desensitization result. 2. Repeat the above actions to adjust the algorithm until it meets the requirements.

3. After completing the test, click **OK** to save the algorithm.

©2013–2025 Tencent Cloud. All rights reserved.

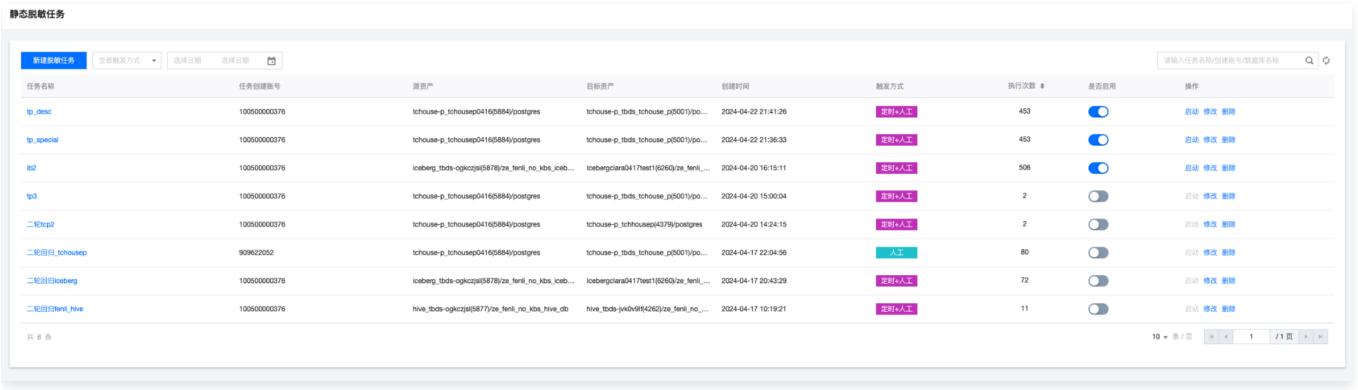
Static Masking

Last updated：2024-11-25 14:33:22

Static DMask uses offline tasks to complete DMask according to task policies, protecting sensitive information from being leaked.

Task Management

1. Log in to Tencent Cloud and go to the WeData Console. Select the project and navigate to **Data Security** > **Static Masking Task** page. Click **Create Masking Task** to configure the static masking task.



Step 1: Configure Basic Task Information

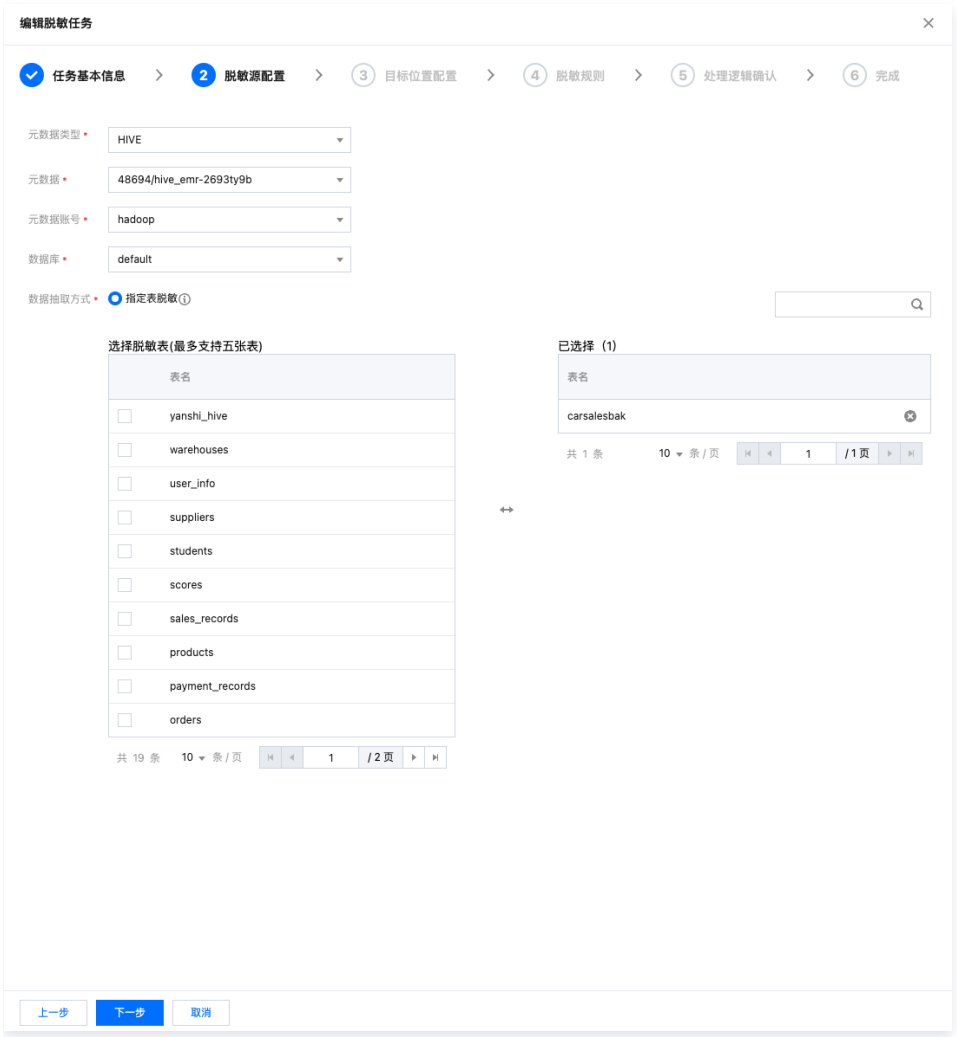
Go to the Basic Task Information page, fill in the **Task Name,** and click Next.



Step 2: Configure Masking Source

Go to the Masking Source configuration page, select the **Masking Table** (up to five tables can be selected).

©2013–2025 Tencent Cloud. All rights reserved.



Information	Details
Metadata Types	Currently supports EMR-Hive type data sources.
Metadata	Select the data source that needs to be masked, which comes from the static masking resource binding's data source.
Metadata Account	Select the account to execute the masking task, provided by the data source configuration.
Database	Select the Database that needs to be masked.
Data Extraction Method	Specify Table Masking: You can select up to 5 tables in a single session.

Step 3: Target Location Configuration

Go to the **Target Location Configuration** page, fill in the relevant information, and click Next.

©2013-2025 Tencent Cloud. All rights reserved.

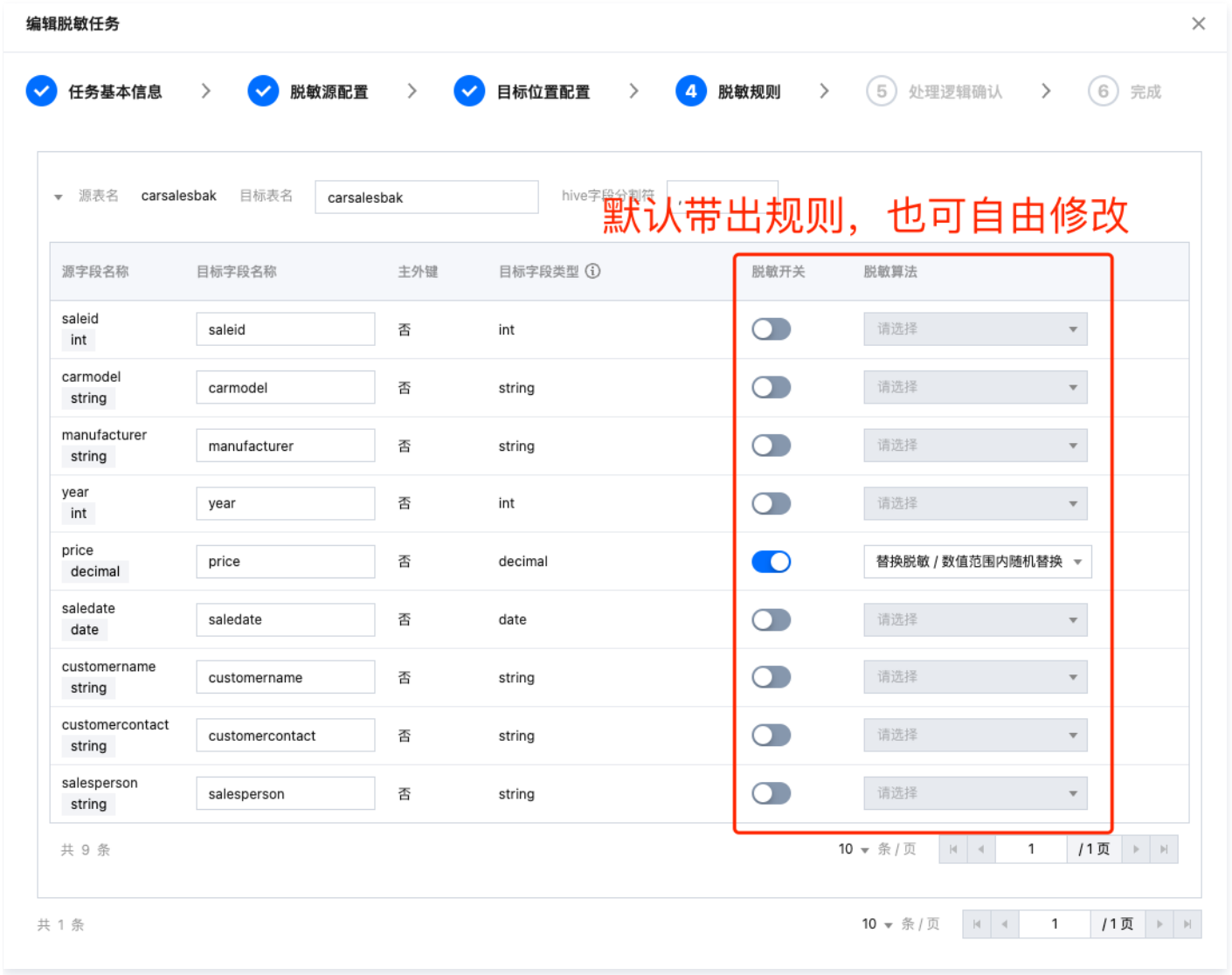


Information	Details
Metadata Types	Currently supports EMR-Hive type data sources.
Metadata	Select the data source that needs to be masked, which comes from the static masking resource binding's data source.
Metadata Account	Select the account to execute the masking task, provided by the data source configuration.
Database	Select the Database that needs to be masked.

Step 4: Data Masking Rules

Go to the **Data Masking Rules** page, you can manually enable the **Masking Switch**. Once enabled, you can freely modify the masking algorithm.

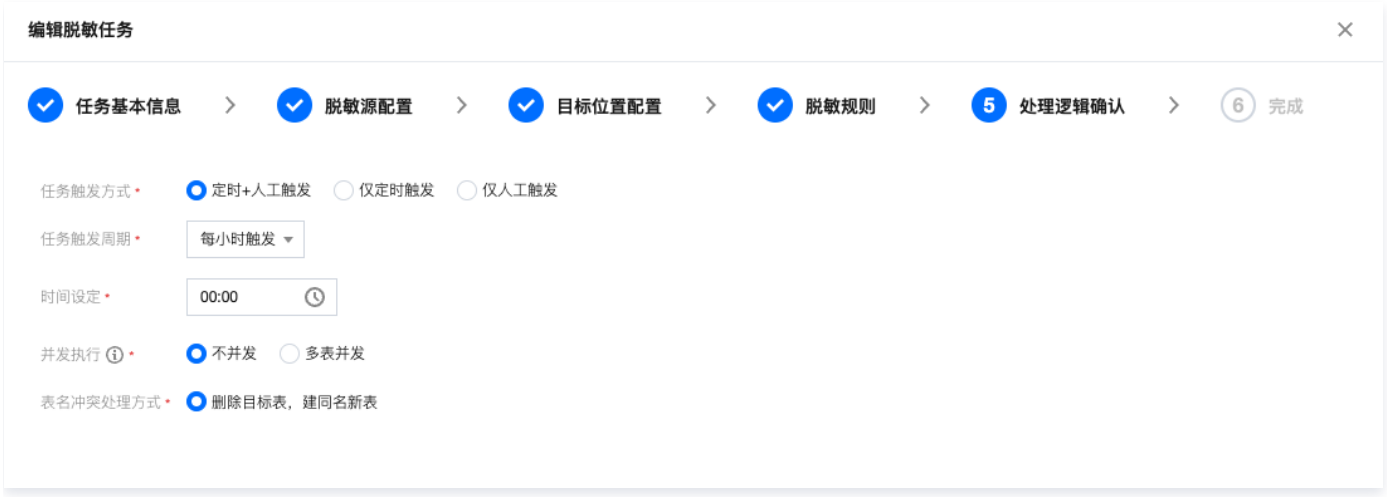
©2013-2025 Tencent Cloud. All rights reserved.



Information	Details
Masking Switch	Used to control whether a field is masked. When enabled, you can configure the masking algorithm.

Step 5: Configure Masking Task Trigger Rules

Go to the **Processing Logic Confirmation** page, configure the relevant information, and click Next.

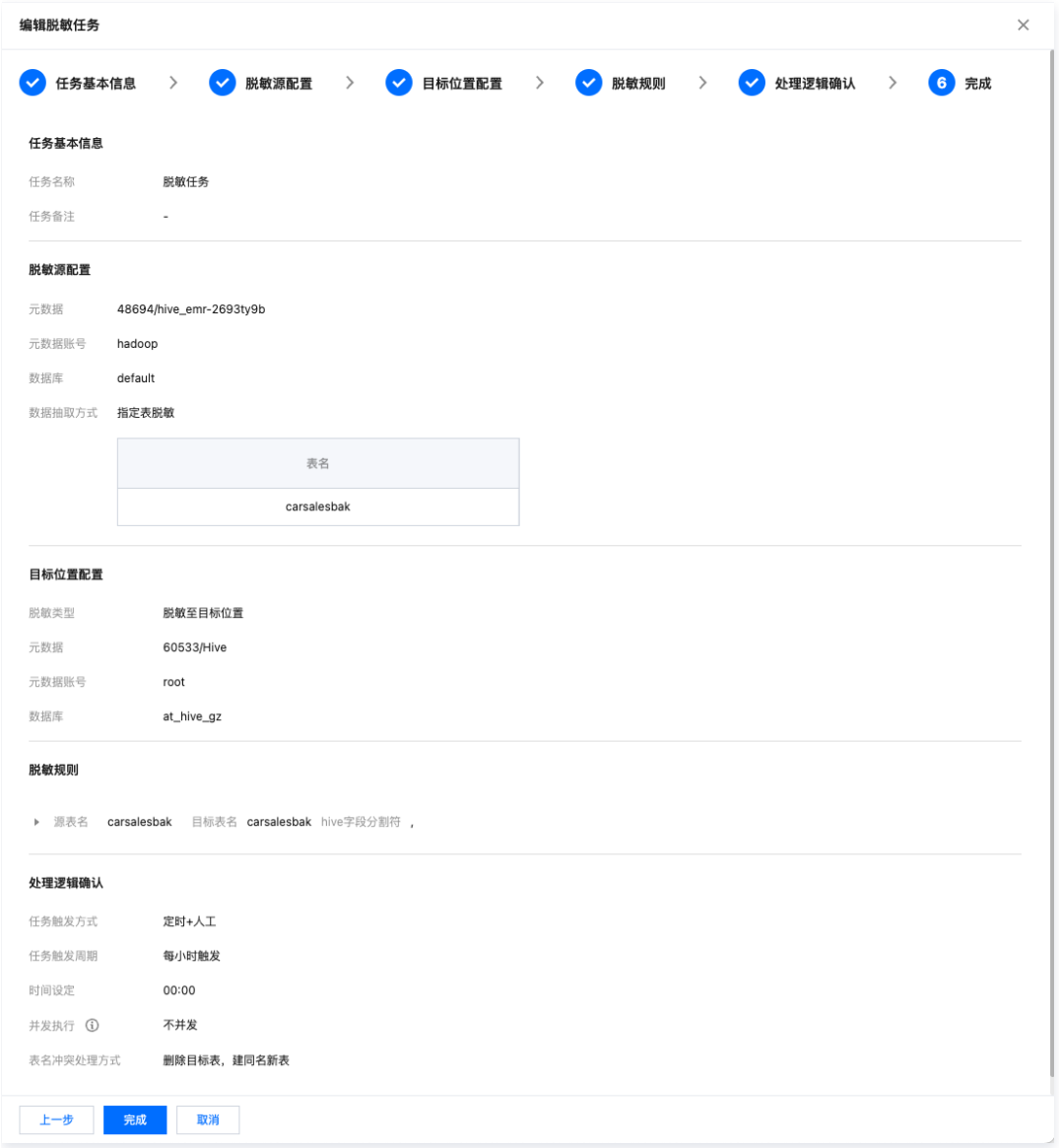


©2013–2025 Tencent Cloud. All rights reserved.

Information	Details
Task Trigger Method	• Scheduled + Manual Trigger: The task supports both scheduled execution and manual trigger. • Scheduled Only: The task can only be triggered on a schedule. • Manual Only: The task can only be triggered manually.
Concurrent Execution	• Non-Concurrent: When masking involves multiple tables, it will execute the masking tasks in series per table. • Concurrent Execution for Multiple Tables: When masking involves multiple tables, execute the masking tasks in parallel per table.
Table Name Conflict Handling	Delete Target Table, Create New Table with Same Name: When there is a table name conflict, delete the target table and create a new table with the same name.

Step 6: Summarize Configuration Information and Secondary Confirmation

Go to the completion page, confirm the configuration information again, and make sure there are no errors, then click **Complete**.

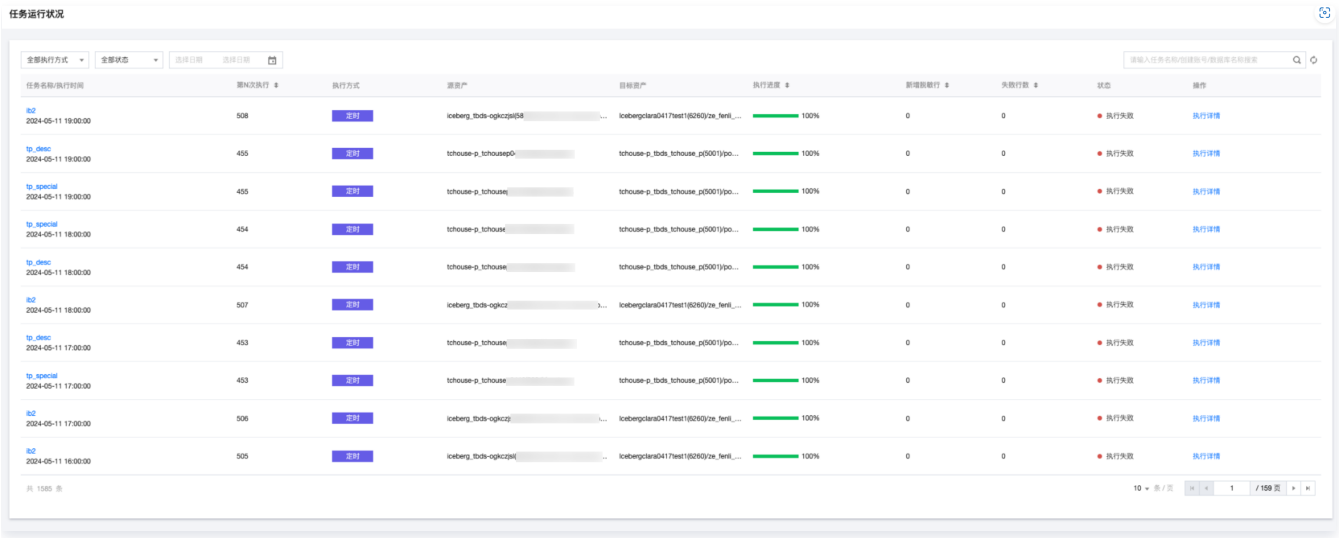


Task Running Status

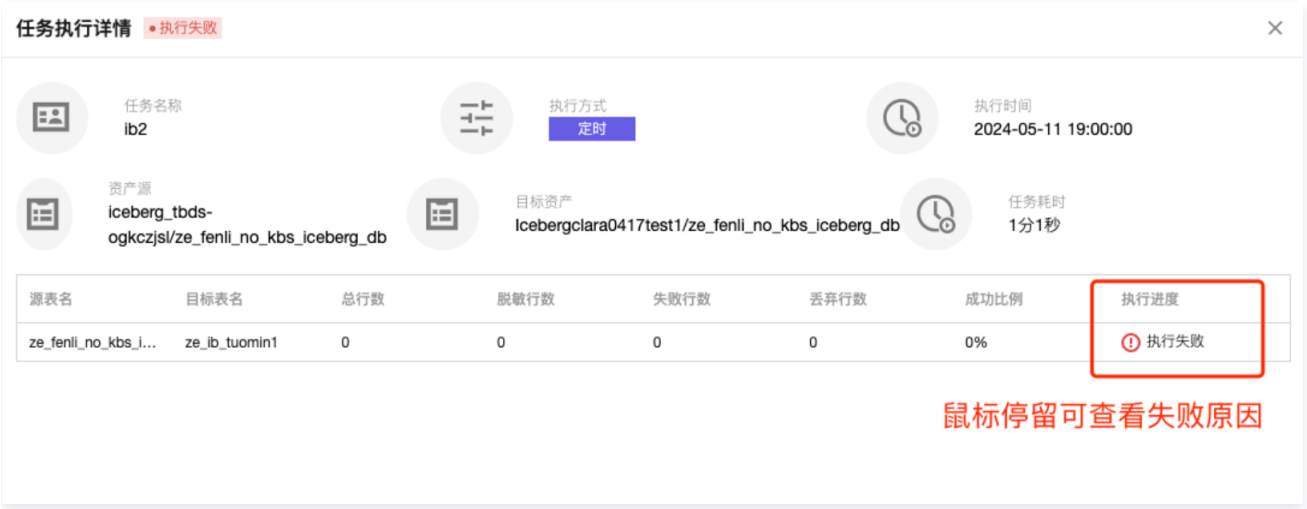
Summarize and display the static masking task running status.

1. You can view the running status of static masking tasks on the Task Running Status page. Click **Execution Details** to see the detailed situation of specific tasks.

©2013-2025 Tencent Cloud. All rights reserved.



2. Go to the **Execution Details** page. You can hover over the execution failure area to see the cause of the failure.



©2013–2025 Tencent Cloud. All rights reserved.

Tables I Manage

Last updated: 2024-11-25 14:33:40

Log in to Tencent Cloud, then enter the WeData Console, select a project, and go to the **Data Security** > **Tables I Manage** page.

> **Note:**
> User permissions configured through policies have a higher priority than permission auditing. The granting of library table permissions in the 'My Permissions' feature, as well as application approval, are subject to being overridden by policy management due to changes at the Ranger engine layer, where policy privileges exceed individual user grants. Specifically, when a project administrator role configures a user's library table access permissions via policy management, any library table permissions that the corresponding user has acquired through authorization by others or application approval in 'Tables I've Applied for' and 'Tables I Manage' will be overridden and no longer displayed.

EMR

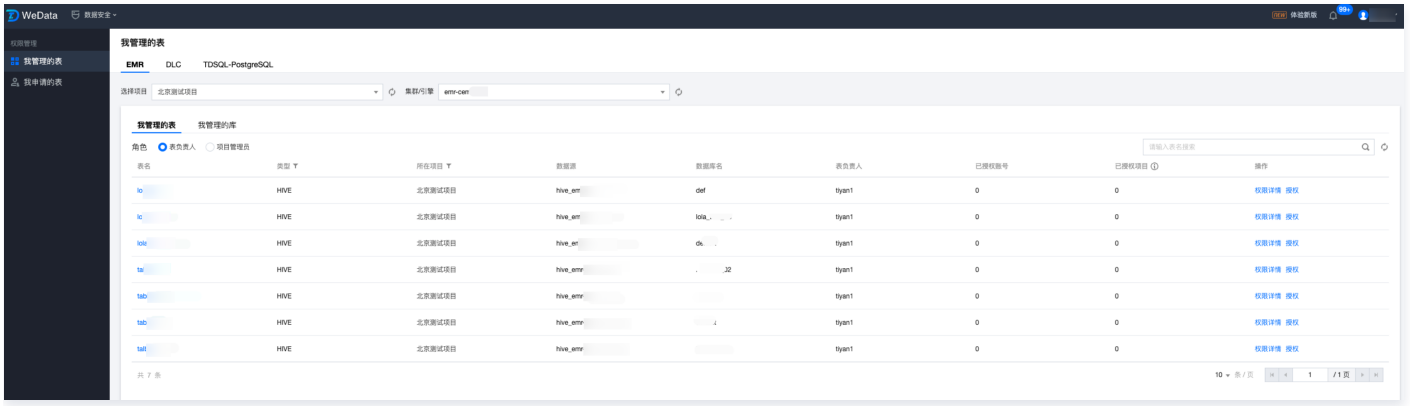
Implemented based on the underlying engine's Ranger policy configuration, only main accounts/CAM authorized sub-accounts can operate and manage.

Tables/Databases I Manage

Displays the library/table information that the current logged-in user has the authority to grant to others, and allows the execution of 'Grant' permissions and the viewing of 'Permission Details' operations.

Roles

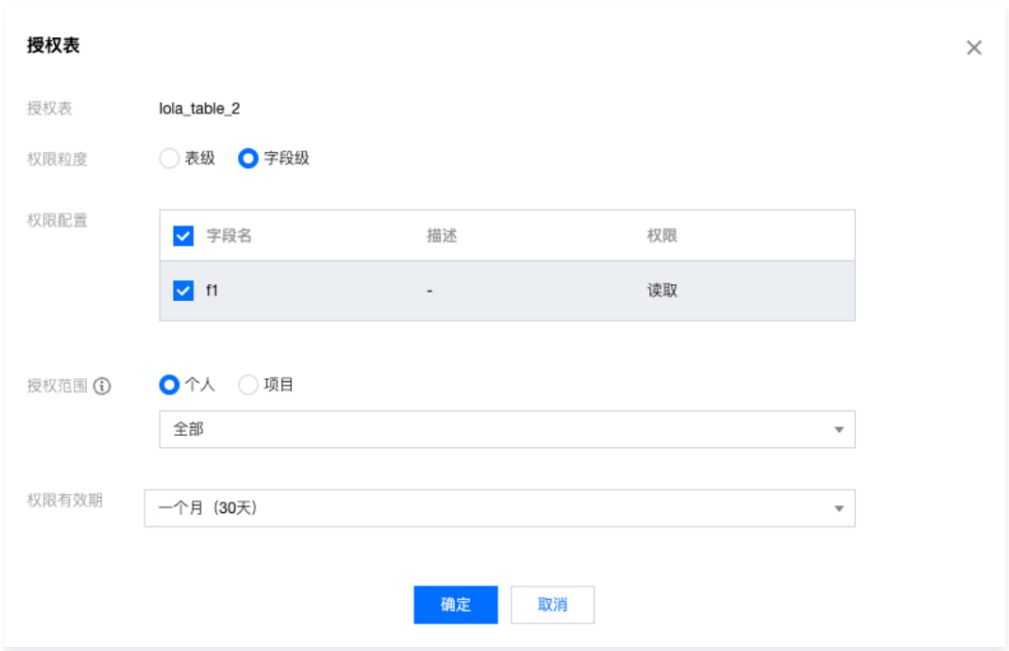
- Table Owner: Displays the database/table information for which the currently logged-in user is the Table Owner.
- Project Administrator: Displays the database/table information owned by the Project Administrator role.



Authorization

Grant individual or project permissions to tables or fields.

©2013-2025 Tencent Cloud. All rights reserved.



©2013–2025 Tencent Cloud. All rights reserved.

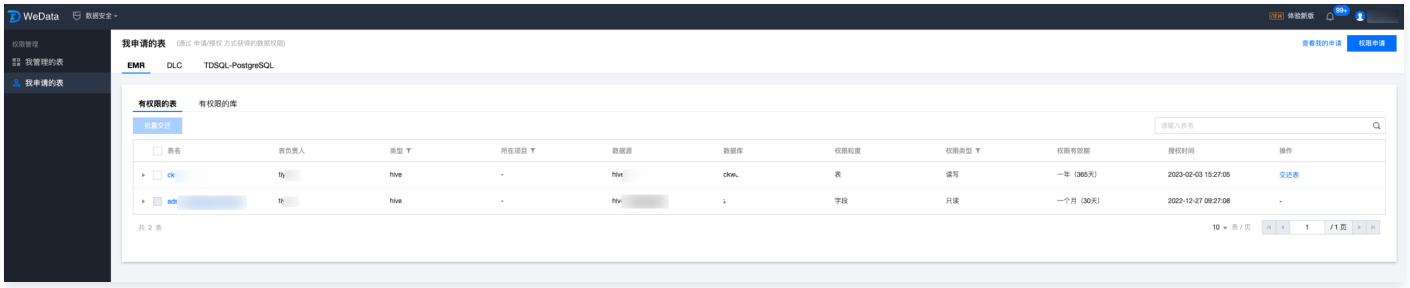
My Applied Tables

Last updated：2024-08-24 18:36:04

Log in to Tencent Cloud, enter the WeData Console, select a project and navigate to **Data Security** > **My Applied Tables** page.

Tables/Databases with Permissions

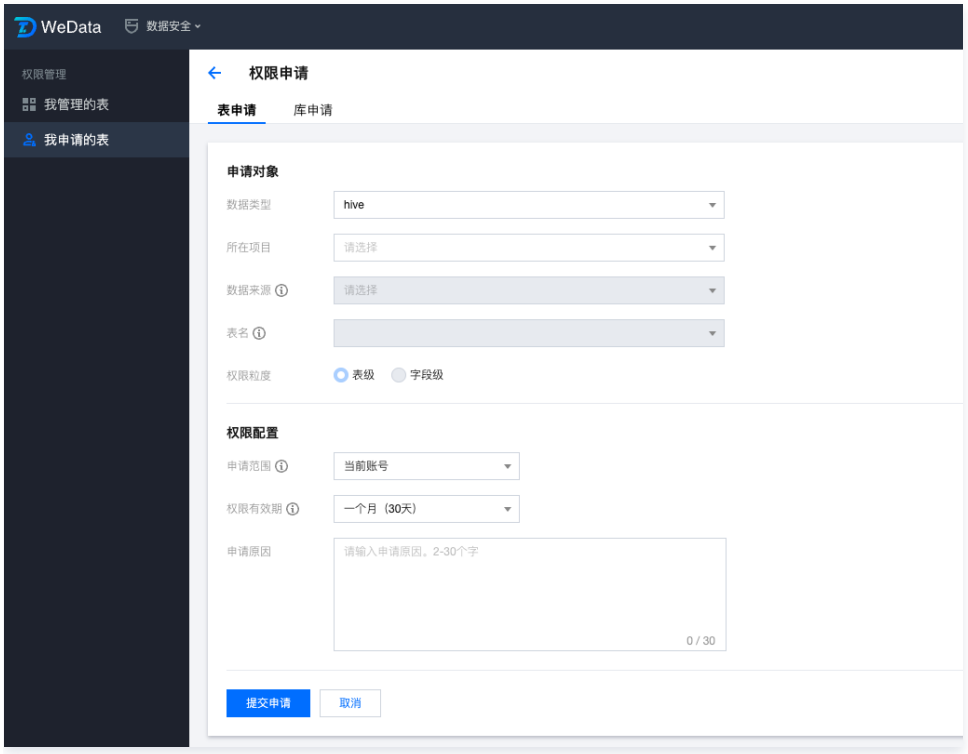
For different users, you can view the data tables and databases they have permissions for, and provide a return permission feature, specific to returning table permissions and field permissions.



Through the permission application feature, users can apply for the relevant permissions of data tables and databases.

Table Application

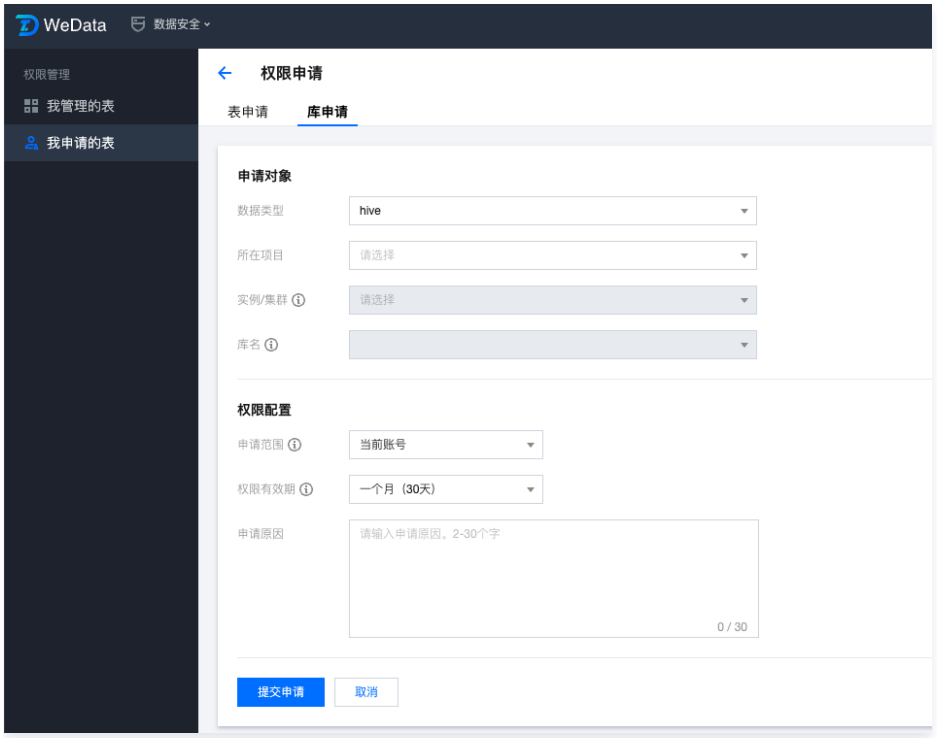
To apply for data table permissions, you need to select the application object, specifying the data type, project, data source, table name, and permission granularity. In terms of permissions, you need to configure the application scope, the validity period of the permissions, and fill in the application reason.



Database Application

To apply for database permissions, you need to select the application object, specifying the data type, project, instance/cluster, database name, and read-only or read-write. In terms of permissions, you need to configure the application scope, the validity period of permissions, and fill in the application reason.

©2013-2025 Tencent Cloud. All rights reserved.

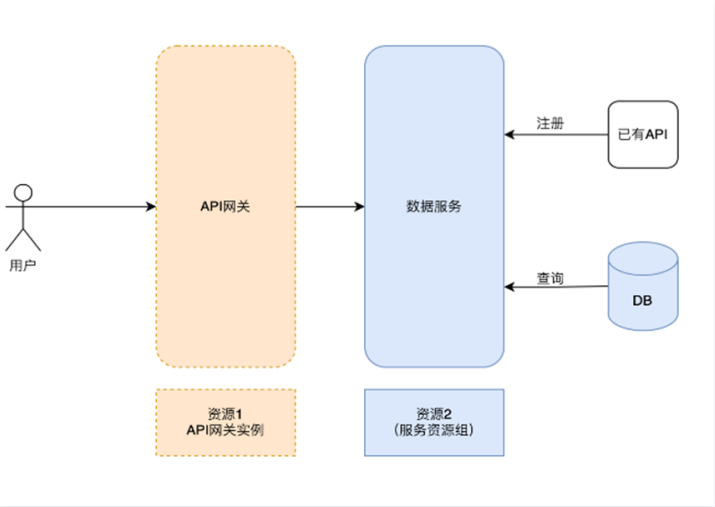


©2013–2025 Tencent Cloud. All rights reserved.

Data services
Data Service Overview

Last updated：2024-08-22 21:01:15

Data service is a component of the WeData platform's data application, aimed at providing developers and enterprises with an easy-to-use, reliable, secure, and highly scalable data service solution. WeData Data Service offers a range of APIs and development tools, enabling developers to easily integrate APIs into their applications. Features include API production, API management, and API market, covering the full lifecycle of APIs, helping enterprises manage both internal and external API services and build a unified data service bus.



Scenarios

The data service is positioned to provide an easy-to-use, reliable, secure, and highly scalable data service solution for developers and enterprises. Use cases for these platforms include DataInLong, data analysis, data visualization, machine learning, and artificial intelligence.

Data Application

The data service API platform can help enterprises collect, organize, and analyze data from multiple sources, and provide the results to applications. This can help enterprises better understand their business and customers, and make more informed decisions.

Data Analysis

The data service API platform can help enterprises analyze their business data and provide insights into customer behavior, trends, and patterns. This can help enterprises better understand their customers and develop more effective marketing and business strategies.

Data visualization

The data service API platform can convert data into easy-to-understand charts and graphs, enabling enterprises to better understand their business and customer data. This can help enterprises better understand their customers and develop more effective marketing and business strategies.

Getting Started

The usage of data service features can be divided into three stages: preparation, development, and operational. These steps help users easily integrate data services into their applications and ensure the performance, reliability, and security of the data services. As follows:

Preparation Stage

©2013-2025 Tencent Cloud. All rights reserved.

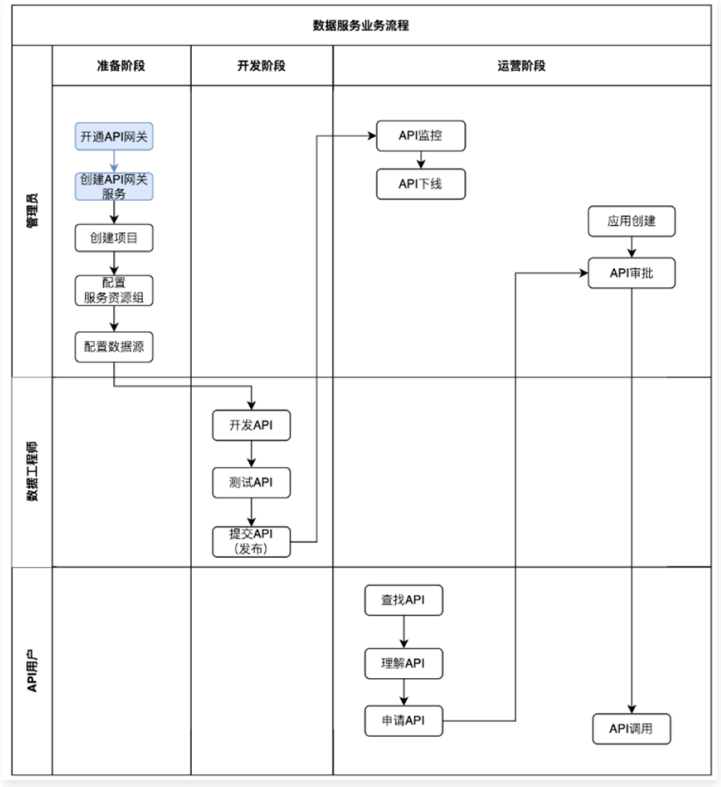
Preparation Stage: The authentication, monitoring, and other features of the data service API rely on the API Gateway. Before using, you need to complete the activation of API Gateway services and the creation of API services. For more information about API Gateway, refer to Obtaining Access Permission.

Development Stage

In the data service – service development module, complete the development, registration, testing, and release of APIs. This involves using API documentation, understanding the features and parameters of APIs, and selecting the most suitable APIs for their needs. Users also need to test the APIs and ensure they operate as expected. Once the APIs are tested and deemed to work as expected, users can integrate the APIs into their applications.

Management Stage

After the API is published to the API Gateway, enterprise users can view the API through the API Market and apply for scheduling as needed. During the management phase, users need to monitor and maintain the API. This helps them ensure the performance, reliability, and security of the API and handle any issues or failures promptly. Users also need to manage aspects such as API access control, data security, and API Keys to ensure the security and protection of user data.



Strengths

©2013–2025 Tencent Cloud. All rights reserved.



Quick Production

Data service API platforms typically provide easy-to-use APIs and development tools, enabling developers to quickly integrate data services into their applications. This means developers can focus on the core features of their applications without spending a lot of time and effort building and maintaining their own data services.

Flexibility

Data service platforms can adapt to various data sources and data formats, easily integrating into any existing system. The platform also supports various data analysis and visualization tools, which can be customized according to user needs.

Scalability

Data service API platforms usually have high scalability, capable of handling a large number of data requests and supporting multiple users accessing simultaneously. Additionally, these platforms often support cloud deployment, allowing computing resources to be automatically scaled up according to demand.

Security

Data service platforms offer multi-layer security protection, including data encryption and access control, ensuring the security of data services and protecting user data from unauthorized access and abuse. The platform also supports multi-tenant deployment, ensuring the isolation and protection of user data.

©2013-2025 Tencent Cloud. All rights reserved.

Preparations

Last updated：2024-08-24 18:24:00

To use WeData Data Services, you need to buy and create an API Gateway instance on TencentCloud API Gateway, and also purchase service resources. This article is aimed at guiding you through the preparations for using WeData Data Services.

Data Service Resource Group Preparation

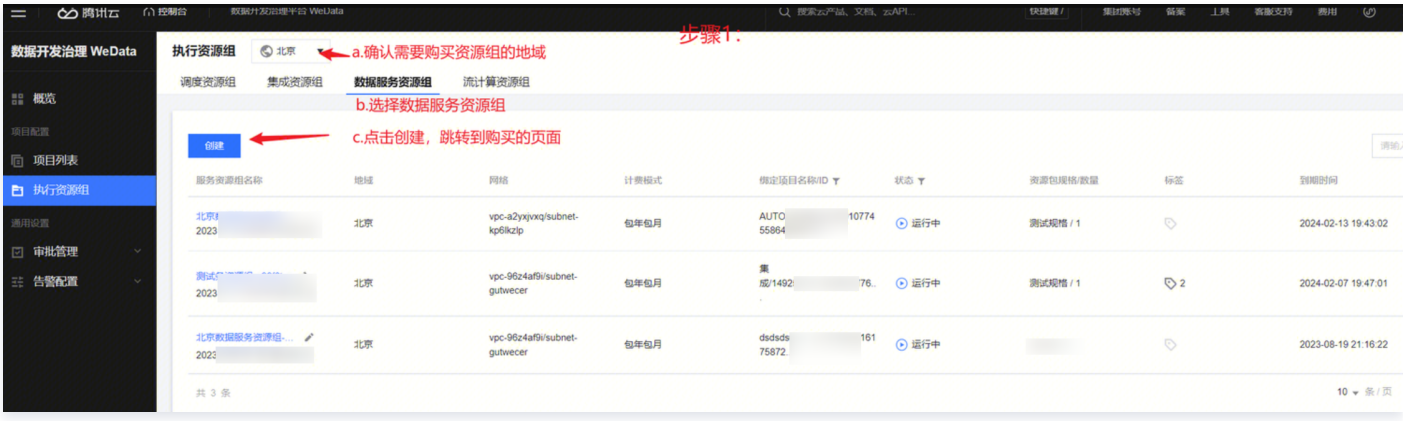
Users need to purchase and prepare a Data Service Resource Group.

1. Introduction to Data Service Resource Group Configuration Data Service Resource Group.
2. Specifications and Billing of Data Service Resource Group Explanation of Data Service Resource Billing.
3. Resource Group Configuration:
   3.1 The fully managed WeData Data Service Resource Group includes three parts: CLB + EKS + Peering Connection (there is also a cost for connecting two VPCs).
   3.2 The specific resource package configuration is related to CLB and EKS.
   3.3 Consider the relevant billing rules and quota for Peering Connection (refer to the VPC connection part in the scheduling and integration section).
   3.4 For EKS nodes, design the approximate resource combination based on the API Gateway's quota situation (e.g., xx cores, xx GB, refer to the EKS usage in the scheduling resource).
   3.5 It is recommended to directly purchase Intranet Instances for CLB. When creating an API in API Gateway services, you can directly bind the intranet CLB to connect.

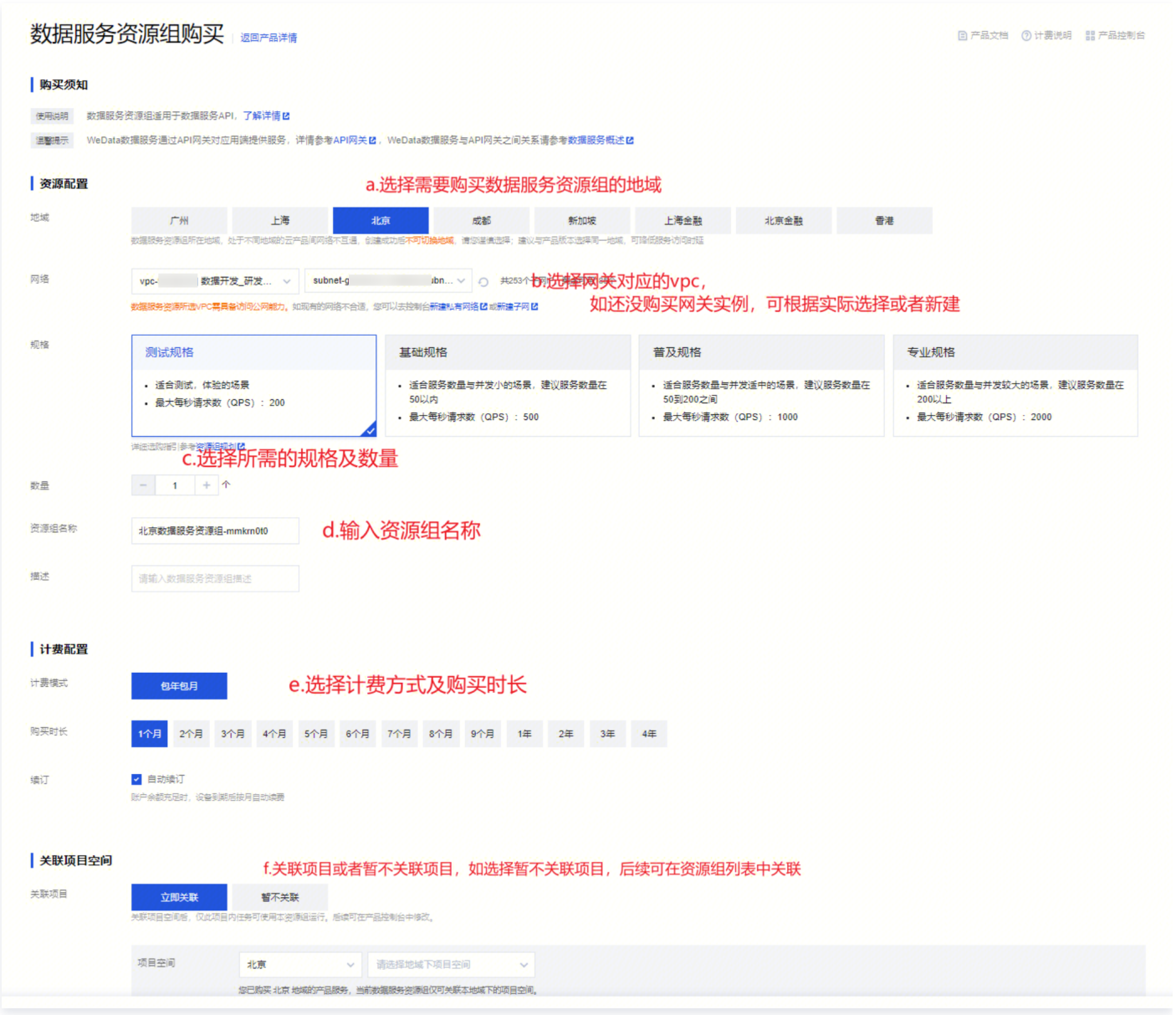
Operation step

Purchase Data Service Resource Group

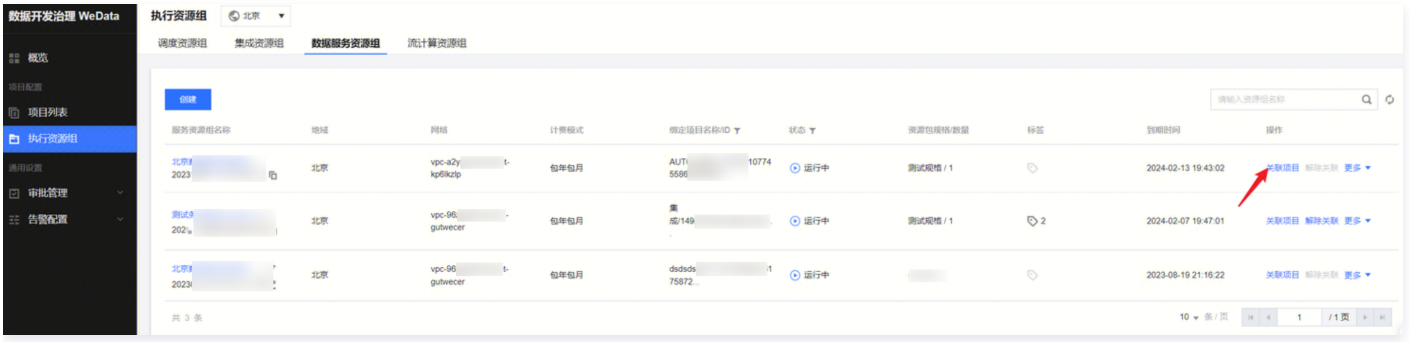
1. Log in to WeData Console, determine the region where WeData is located, for example, Beijing.
2. After creating the project and configuring the related configuration items, purchase the Data Service Resource Group.
   2.1 View or Purchase Data Service Resource Group



   2.2 Purchase Data Service Resource Group



3. Adjust or associate the data service resource group with a project. After linking the resource group, you can select the corresponding resource group when creating an API in the data service. Enter the resource group list, select the data service resource group tab, click on the data service resource group that needs to be associated with a project, and click **Associate Project** to proceed.



4. Enter the Associate Project interface, select the project name to be associated, and click **OK**.

©2013–2025 Tencent Cloud. All rights reserved.



API Gateway Purchase Service Configuration

Overview

To use WeData data services, you need to purchase Tencent Cloud Native Gateway. For more details, refer to the Cloud Native Gateway Overview.

Directions

1. Go to the Purchase Page to buy resources consistent with the WeData service resource group.

> **Note**
> When purchasing a gateway instance, make sure the gateway network VPC is consistent with the data service resource group VPC.

2. If you don't have a data service resource group yet, you can purchase the resource group first and then the API Gateway instance; or buy the API Gateway instance first and then the data service resource group. The purchase interface is shown as follows:



3. After successfully purchasing the gateway instance, you will see the name of the instance you just bought on the Instance List. As shown below:



4. Enter the Create API interface. You can select the newly created gateway service and choose the corresponding data service resource group.

> **Note:**
> Here you need to select **Tencent Cloud Native Gateway**.

©2013–2025 Tencent Cloud. All rights reserved.



At this point, you have completed all the preliminary work for using WeData data services.

©2013–2025 Tencent Cloud. All rights reserved.

Getting Started
Test API

Last updated：2024-08-22 21:04:11

Operation scenarios

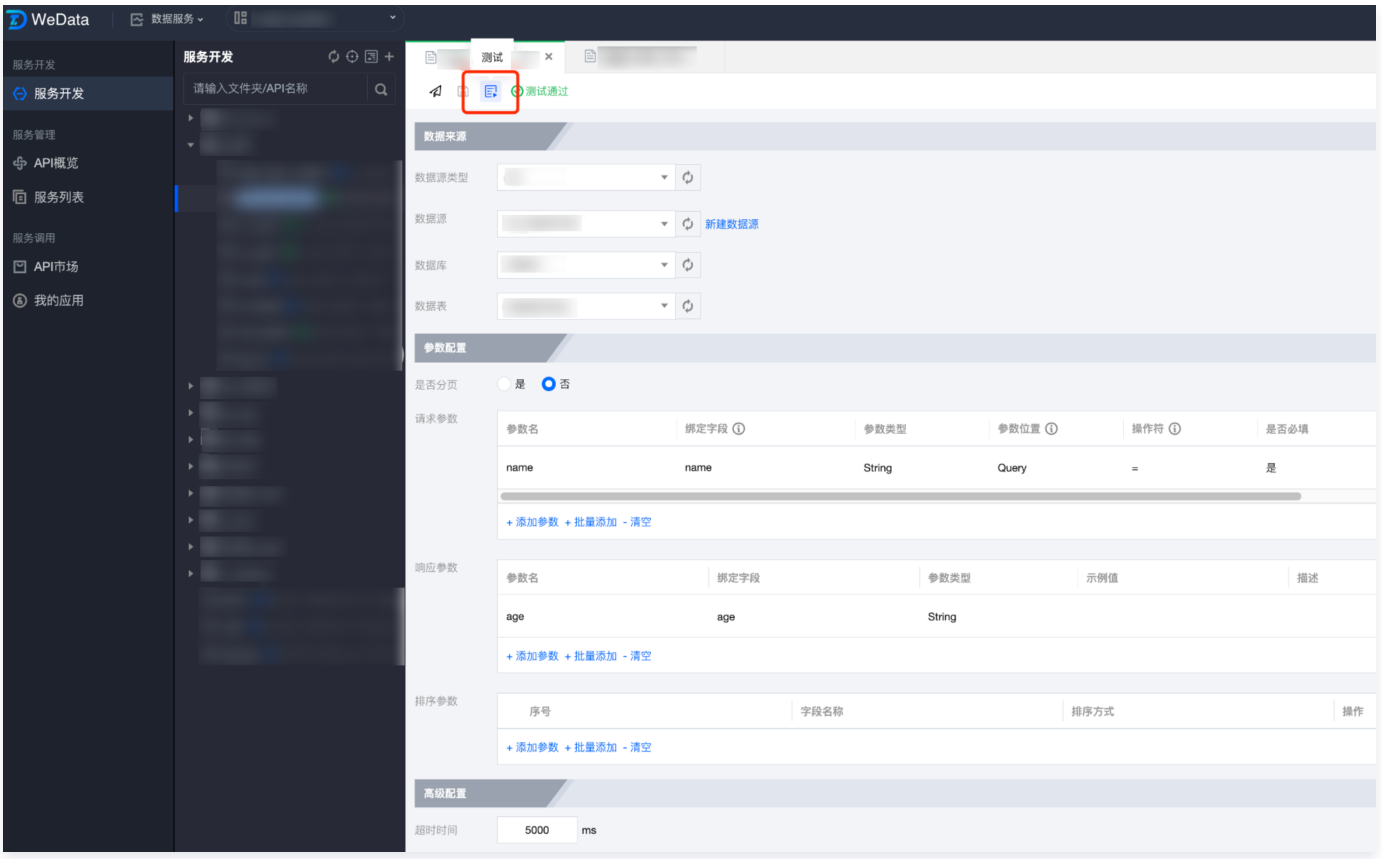
Conduct feature testing on the configured and developed API to ensure it is available.

Prerequisites

Complete the creation and saving of the API.

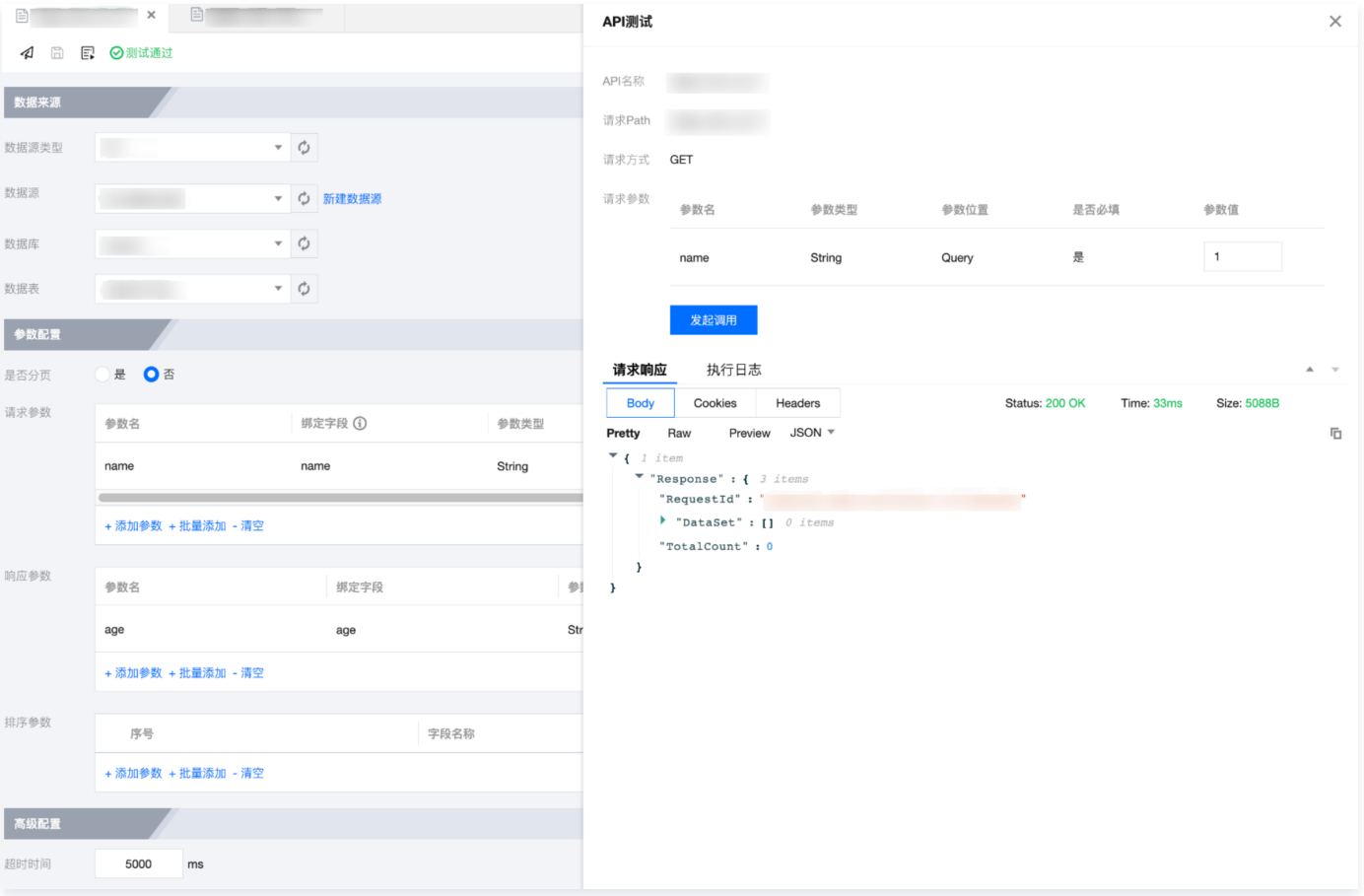
Operation step

1. Log in to WeData Console , click the project list on the left, select a project, and click **Data Service,** to enter the data service homepage.
2. Click on the left **Service Development > Service Development** , and in the middle API directory tree, double-click to open the newly created API.
3. Click the toolbar **Test** to test the API configuration.

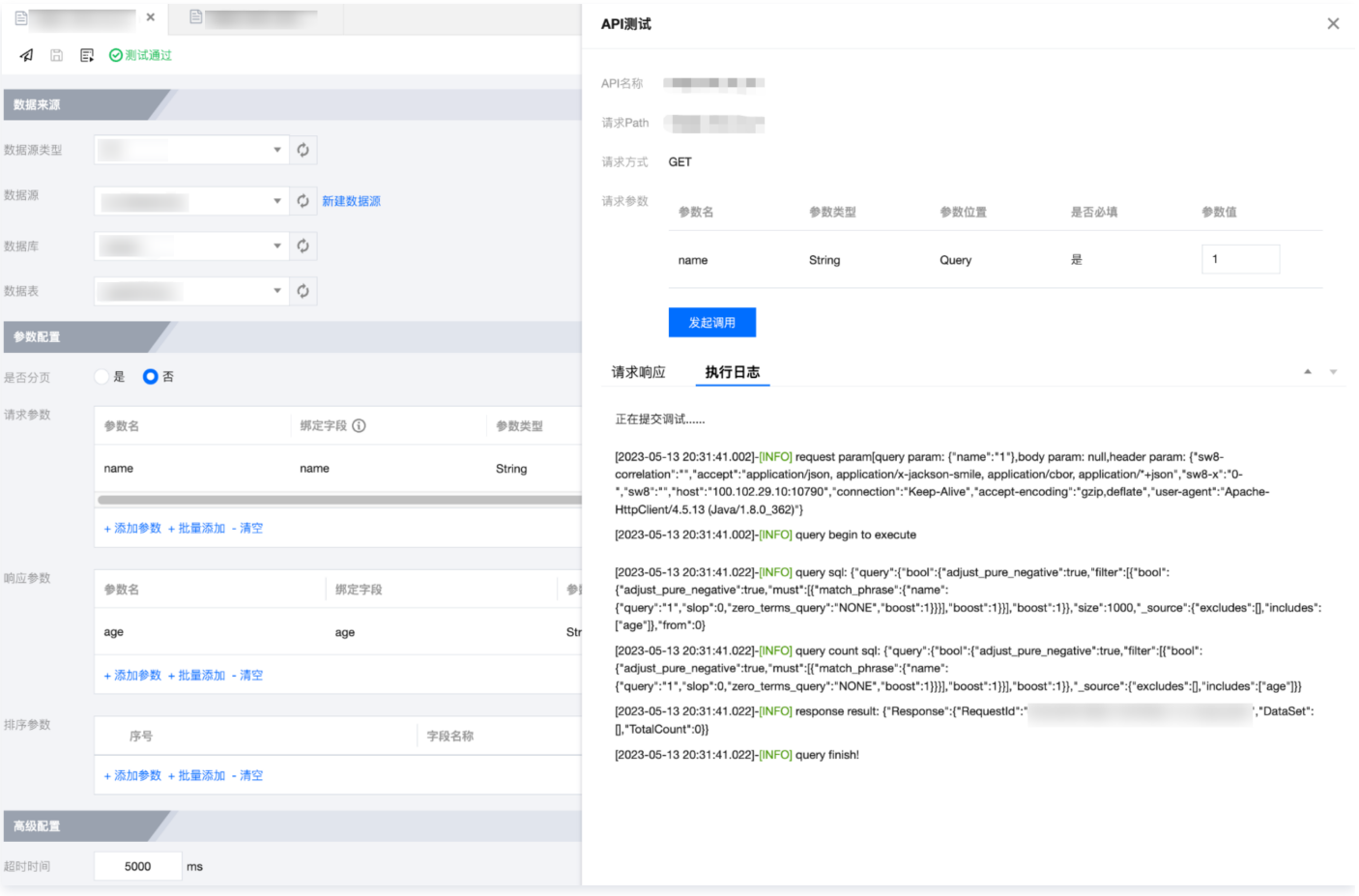


4. In the popped-up API test sidebar, enter the request parameters, click **Initiate the call,** and the test results will be displayed below.

©2013-2025 Tencent Cloud. All rights reserved.



5. In the test results, switch to execution log to view the test run logs.



©2013–2025 Tencent Cloud. All rights reserved.

Release API

Last updated：2024-08-22 21:04:36

Operation scenarios

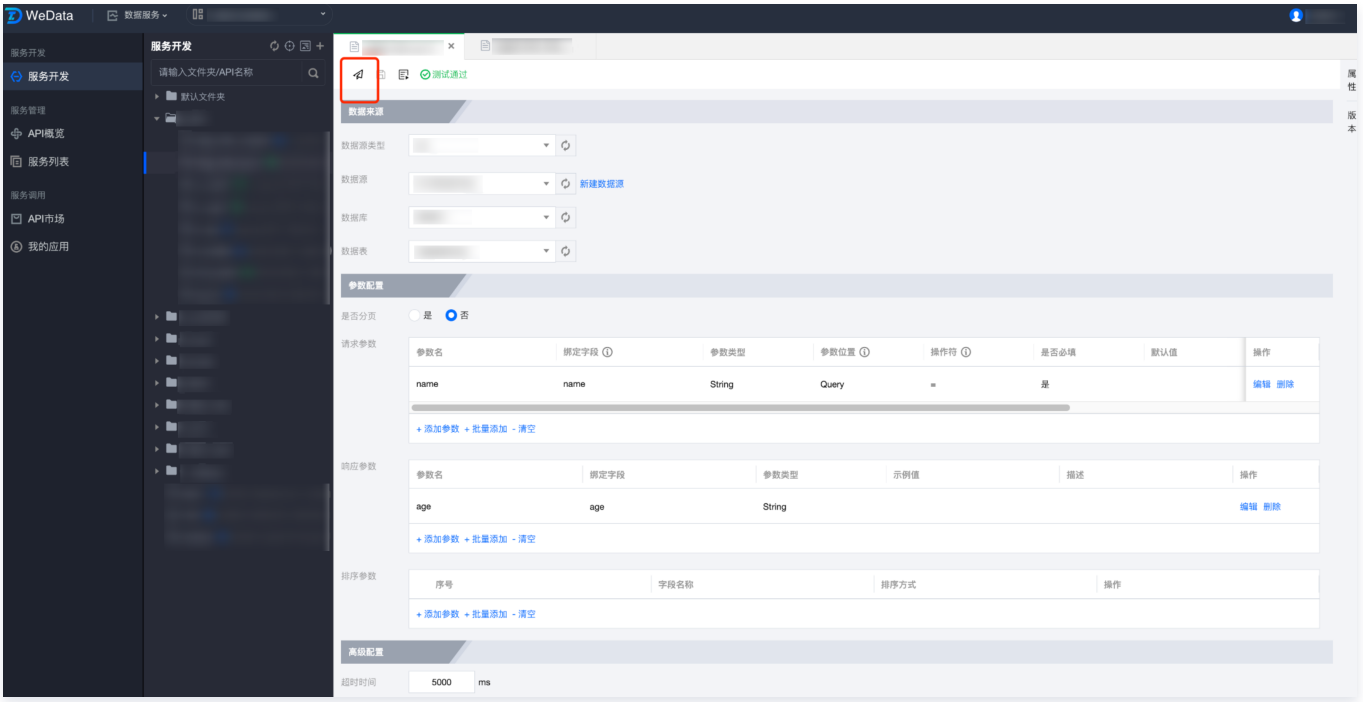
Publish the API to the API Gateway for user invocation.

Prerequisites

Complete API creation and testing.

Operation step

1. Log in to WeData console, click on the project list on the left, select a project and click **Data Service** to enter the homepage of Data Service.

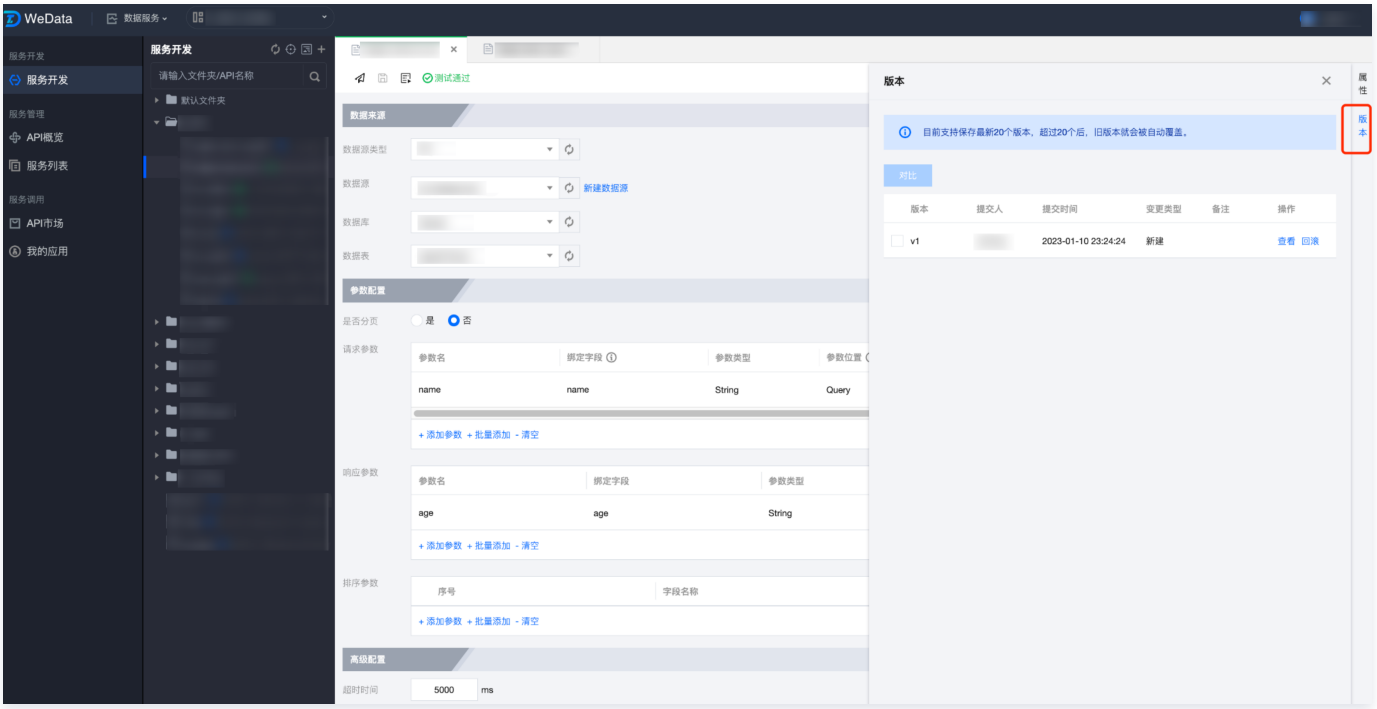


2. Click on **Service Development > Service Development** on the left, then in the API directory tree in the middle, double-click to open the newly created API.
3. Click **Confirm** to publish the current API to the API Gateway. You can then navigate to **Service Management > Service List** to view the published API.



4. Published API Service List.

©2013-2025 Tencent Cloud. All rights reserved.



©2013–2025 Tencent Cloud. All rights reserved.

API authorization

Last updated：2024-08-22 21:05:01

Operation scenarios

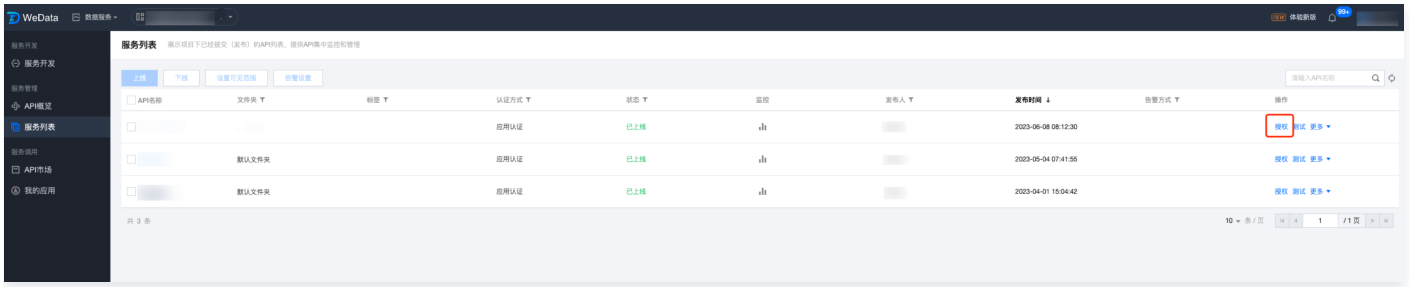
Users authorize their managed APIs to other users or revoke authorization, supporting application-level authorization.

Prerequisites

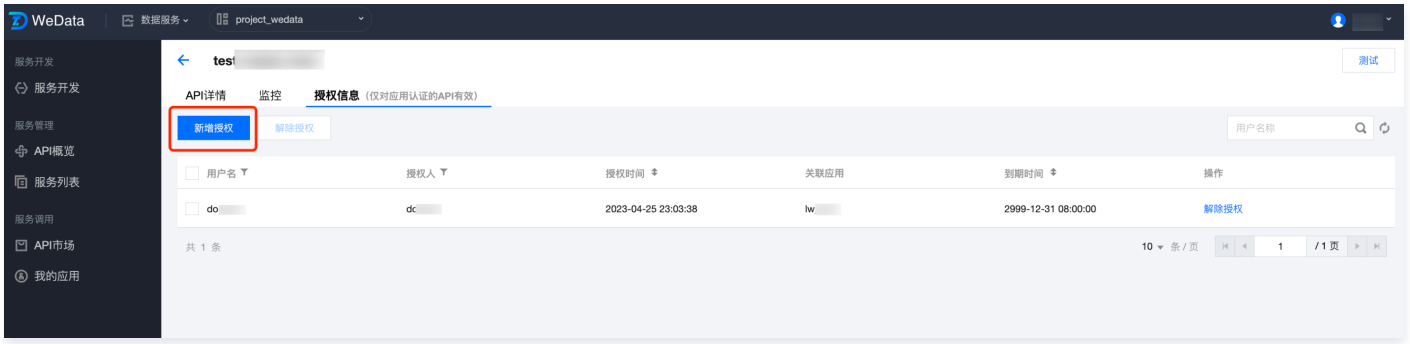
Operate among WeData project members already added by the user, and they must have the project administrator or data engineer role.

Operation step

1. Log in to the WeData Console, click the project list on the left, select a project, and click **Service Management** to enter the data service management page.
2. After entering **Service Development > Service List**, click **Authorization** under the operation column to authorize the selected API.



3. After entering the authorization page, click **Add Authorization**.



4. Enter the relevant information for the new authorization.



Parameter description:

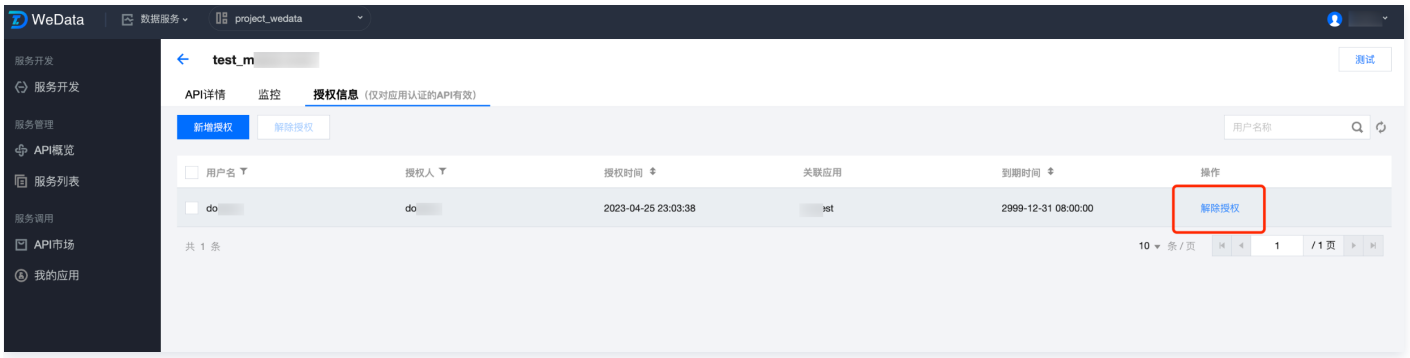
Parameter	Description

©2013-2025 Tencent Cloud. All rights reserved.

Authorization Method	Currently supports individual user authorization
User	Supports authorization for all WeData users under the tenant, with multi-selection supported. When multiple users are selected, the application identity only supports selecting all
Application	Users use the application identity to call APIs
Valid till	Set an authorization period. Calls will fail after expiration

Revoke Authorization

After completing authorization configuration, you can click **Revoke Authorization** on the API authorization information page to revoke a user's permissions.



Search API

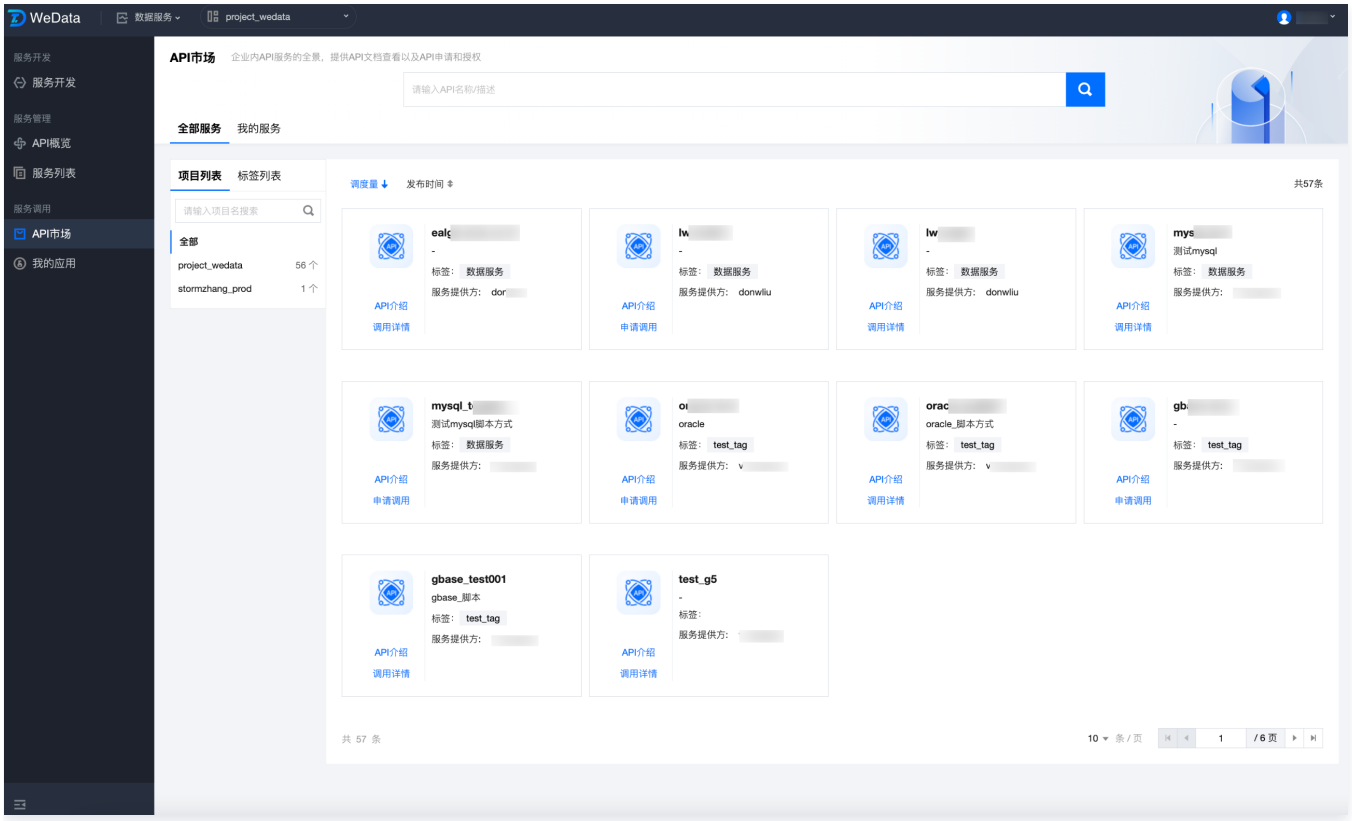
Operation scenarios

1. Users search for the required API by keywords or other means.
2. Users select the required API from the All Services or My Services list.
3. Users find the required API through the Project List or Tag List.

Operation step

1. Log in to the WeData Console, click the project list on the left, select a project, and click **API Marketplace,** to enter the API Marketplace homepage.
2. On the API Marketplace page, enter keywords to search for the required API. You can click the project list and Tag List on the left to filter APIs.
3. Select **All Services** to view the online services.

©2013–2025 Tencent Cloud. All rights reserved.


WeData
数据服务
project_wedata
服务开发
服务开发
服务管理
API概览
服务列表
服务调用
API市场
我的应用
API市场
企业内API服务的全景，提供API文档查看以及API申请和授权
请输入API名称/描述
全部服务
我的服务
项目列表
标签列表
请输入项目名搜索
全部
project_wedata 56 个
stormzhang_prod 1 个
调度量
发布时间
共57条
API介绍
调用详情
标签： 数据服务
服务提供方： donwliu
测试mysql
申请调用
oracle
标签： test_tag
oracle_脚本方式
gbase_脚本
共 57 条
10 条 / 页
1
/ 6 页

Calling APIs

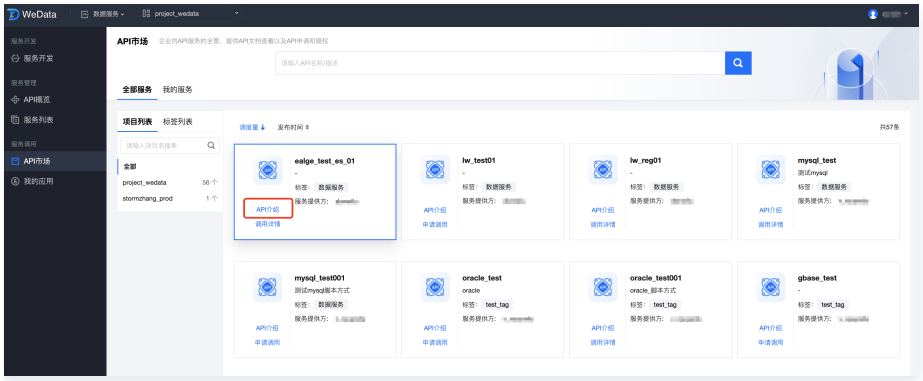
Last updated：2024-08-22 21:05:28

Operation scenarios

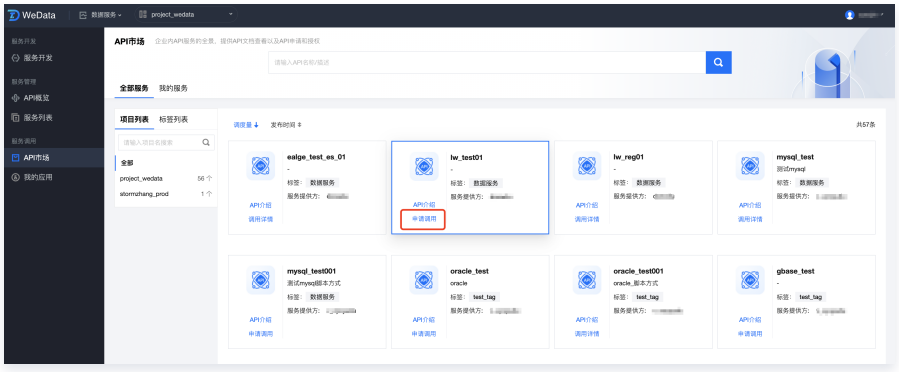
To apply for and invoke the identified API, an application must be submitted when the API authentication method is "application authentication".

Operation step

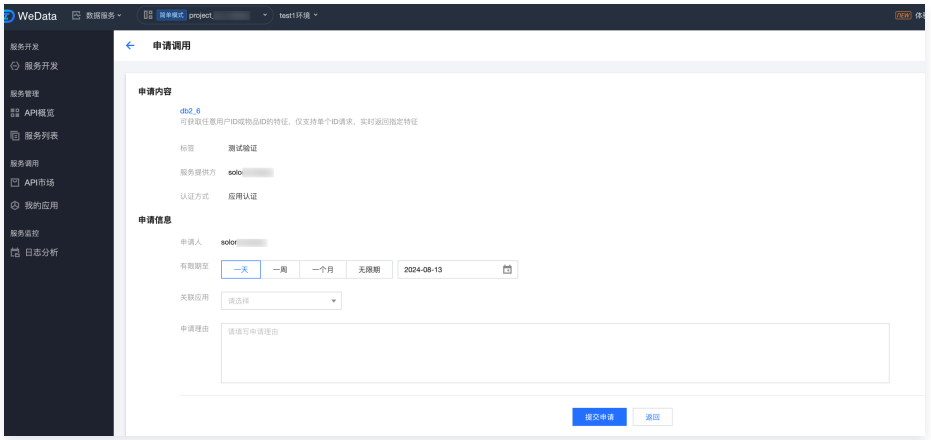
1. After locating the API, you can click **API Introduction** to view the API documentation.



2. Click API **Application for Invocation**, which includes both application authentication and non-authentication methods.
   2.1 **Method 1:** For "application authentication" API application, follow these steps:
      2.1.1 Click **Application for Invocation** under the required data service and configure the application information on the page.



      2.1.2 Details of the application for invocation information.
      - Application Content: Confirm the API details, tags, service provider, and appropriateness of the authentication method.
      - Application Information: Confirm the applicant, configure the maximum number of requests, validity period, associated applications, and application reason.



©2013-2025 Tencent Cloud. All rights reserved.

2.1.3 After configuration, click **Submit Application**. For "application authentication" APIs, Wedata currently uses the application authentication method of the API Gateway itself.



2.1.4 Application voucher: After the applicant initiates an application to the API provider, the API provider retrieves the App Voucher Information (including Application ID, ApiAppKey, and ApiAppSecret) in **My Applications** > **New Application** > **Application Configuration Completed**.

- Calculate the signature. The method to calculate the signature varies by gateway type: API Gateway or Cloud Native Gateway . Use the signing key to sign the request content.
- Initiate the invocation using the API signature. For specific invocation information, refer to API Details .

2.2 **Method 2:** For "authentication-free" APIs, no application is needed; you can directly invoke them.



API key

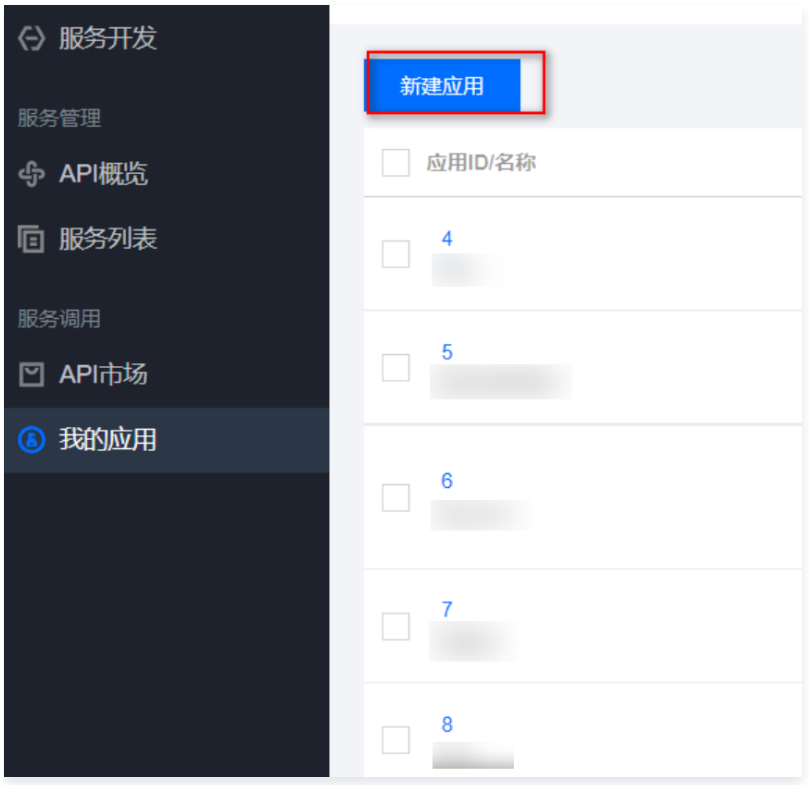
Operation scenarios

Users create keys to provide AppKey, AppSecret, and other access tokens for the API service.

©2013-2025 Tencent Cloud. All rights reserved.

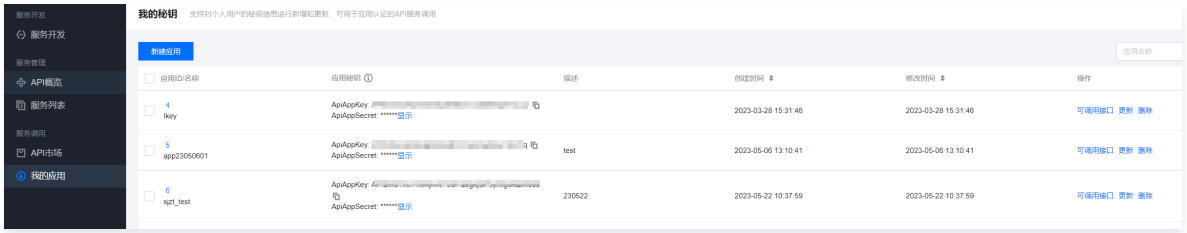
Operation step

1. Click **My Applications** on the left, then click **New Application**, configure the application name and description, this completes the application creation and automatically generates the AppKey and AppSecret.



Application List

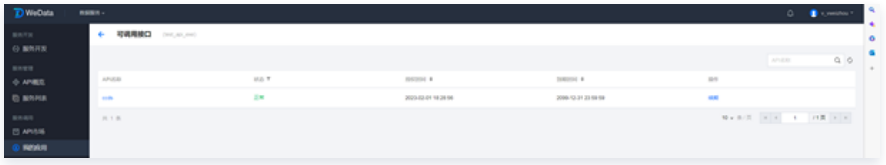
The application list displays information of the created applications, including Application ID, Application Key, Description, Creation Time, Modification Time. It provides features to view available API interfaces, update applications, and delete applications.



Related parameter description:

Parameter	Description
Application ID/Name	Application ID number and English name
Application Key	ApiAppKey and ApiAppSecret need to be included as parameters in the Header during requests. The method for calculating signatures varies across different gateway types: API Gateway, Cloud Native Gateway.
Description	Application Description
Creation Time	Application Creation Time
Modification Time	Last Modification Time of Application Configurations
Callable APIs	Click to view the callable APIs under the application

©2013–2025 Tencent Cloud. All rights reserved.

	
Update	Updating the application AppSecret will affect the APIs called by the application, please proceed with caution
Delete	Delete the selected application

©2013–2025 Tencent Cloud. All rights reserved.

Other Features
Data Service Alarm Feature Description

Last updated: 2024-08-22 21:06:14

Feature Overview

The data service module of the WeData platform supports monitoring and alarming for API calls. Through defined alarm rules, it collects monitoring data, analyzes it, and sends alarm notifications, thereby achieving API alarm features. Common features of the API alarm feature include flexible alarm rule definitions, various alarm notification methods, historical alarm record queries, alarm blocking, and custom alarm notification templates.

Product Features

Users can set alarms for APIs, and alarms will be triggered when the alarm rules are met, enabling proactive operation and maintenance.
The product features of the API alarm feature generally include the following steps:

Flexible Alarm Rules

Users can define monitoring conditions for APIs. For example, an alarm can be triggered when the error rate of an API exceeds a certain threshold. The API alarm feature should support flexible alarm rule definitions based on indicators like request count, response time, internal network outgoing traffic, effective call count, etc.

Multiple Alarm Notification Methods

Supports various alarm methods such as Email, SMS, WeChat, Telephone, WeCom, and HTTP.

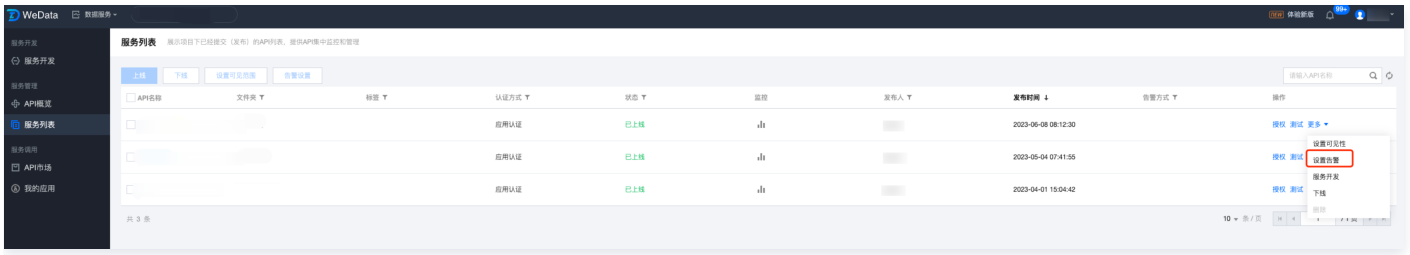
Operation step

Prerequisites

API release completed.

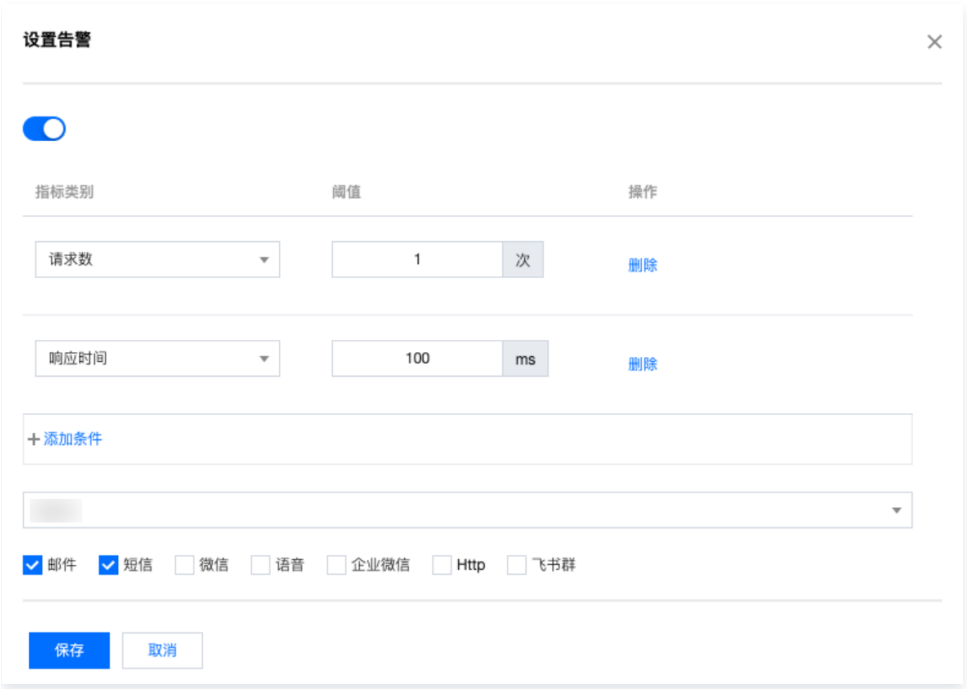
Operation step

1. Log in to WeData Console, click the project list on the left, select a project and click **Service Management** to enter the data service management page.
2. Find the target API in the service list, click **Alarm Settings** in the operation column to set alarm conditions for the selected service, and alarms can be triggered when the service calls exceed the set threshold.



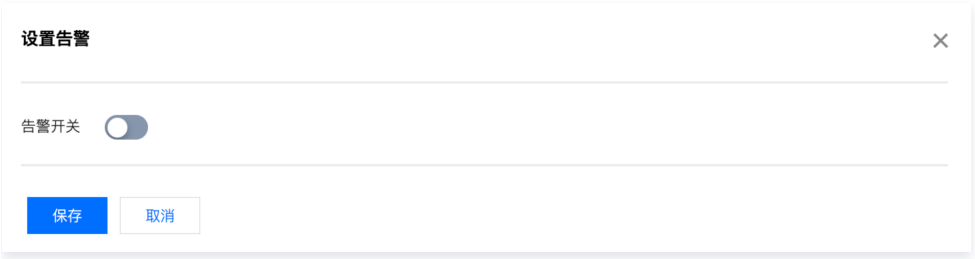
3. Enter the alarm setting interface, set alarm parameters, and you can refer to the following table to set them:

©2013-2025 Tencent Cloud. All rights reserved.



Parameter	Description
Alarm Switch	Enable or disable alarm settings. Alarms will not be triggered when it is turned off.
Alarm Condition	Set alarm conditions for the service. When multiple conditions are selected, an alarm will be triggered if any one of them is met. Alert metrics support call count and error rate.
Recipient	Information recipients when alarm conditions are triggered.
Alarm Method	Email, SMS, WeChat, Voice, WeCom, HTTP, FeiShu Group.

4. Click **Alarm Settings** again, turn off the alert switch to cancel the alarm for this API.



©2013–2025 Tencent Cloud. All rights reserved.

Data Service Monitoring feature Description

Last updated：2024-08-22 21:06:38

Feature Overview

The API monitoring feature is a tool for monitoring and analyzing the performance of application interfaces. It helps developers identify and resolve performance issues in applications, thereby enhancing the reliability and availability of applications. This document will introduce the basic principles, features, and advantages of the API monitoring feature. The API monitoring feature is an essential tool that helps developers promptly identify and resolve performance issues in applications, thereby enhancing the reliability and availability of applications. When using the API monitoring feature, it's necessary to configure the monitoring appropriately based on actual needs and business scenarios. To ensure the normal operation of the application, these APIs need to be monitored and managed. This document will provide a detailed introduction to the implementation principles, common features, and usage methods of the API monitoring feature.

Application Scenario

The application scenarios of the API monitoring feature include the following aspects:

Real-Time monitoring

API monitoring tools can monitor the usage of APIs in real-time, including metrics such as request volume, response time, and error rate. This helps developers and administrators promptly identify and resolve issues.

Performance Analysis

API monitoring tools can provide performance data of APIs, such as response time distribution and throughput. This data can help developers understand performance bottlenecks and optimization opportunities of the API.

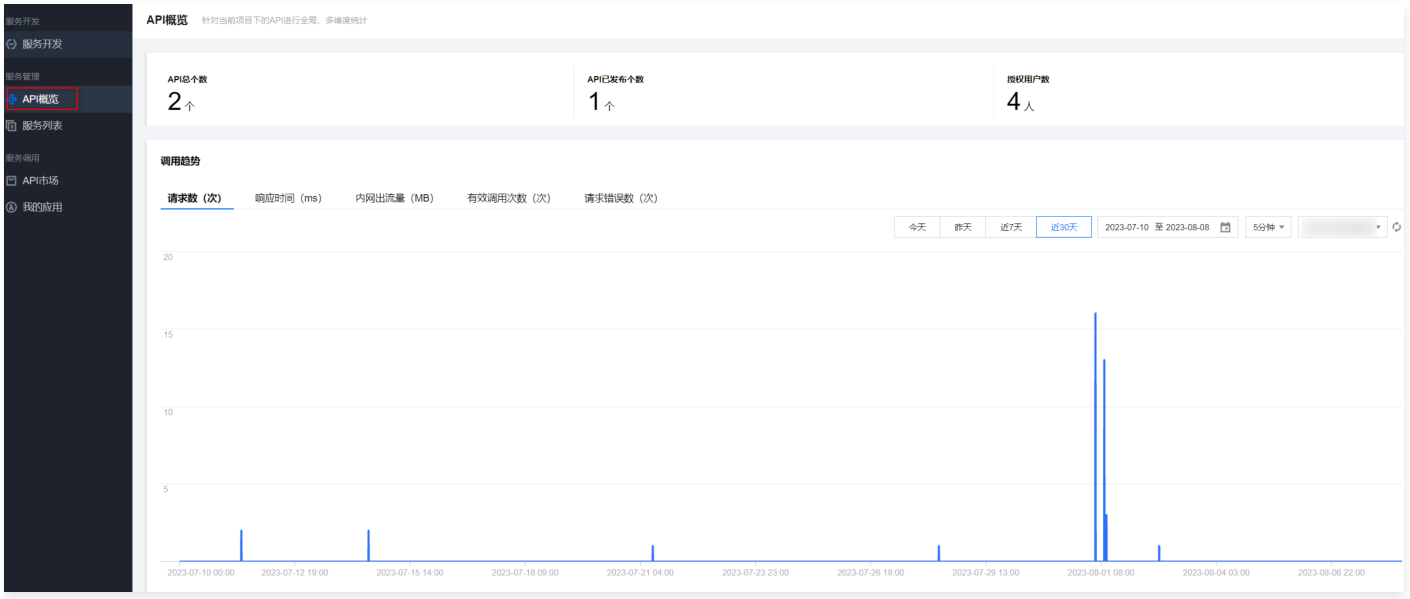
Alarms and Notifications

API monitoring tools can set alarm rules and notification methods, such as sending email or SMS notifications when the API error rate exceeds a certain threshold. This helps developers and administrators respond to issues in a timely manner.

Usage Steps

API Overview

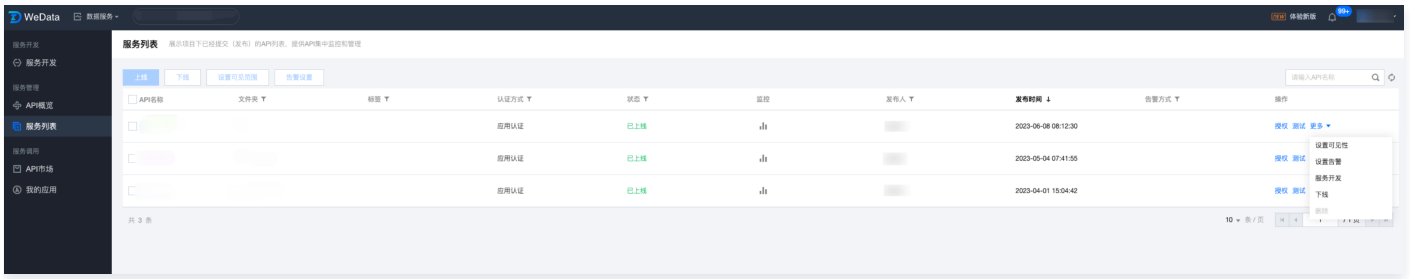
Go to **Data Service > API Overview** to view the overall release status and call trends of the API. Monitoring is supported for published APIs, including total API count, number of published APIs, and authorized users. Call trends support viewing the overall trend of request count, response time, intranet outbound traffic, effective invocation count, and request error count.



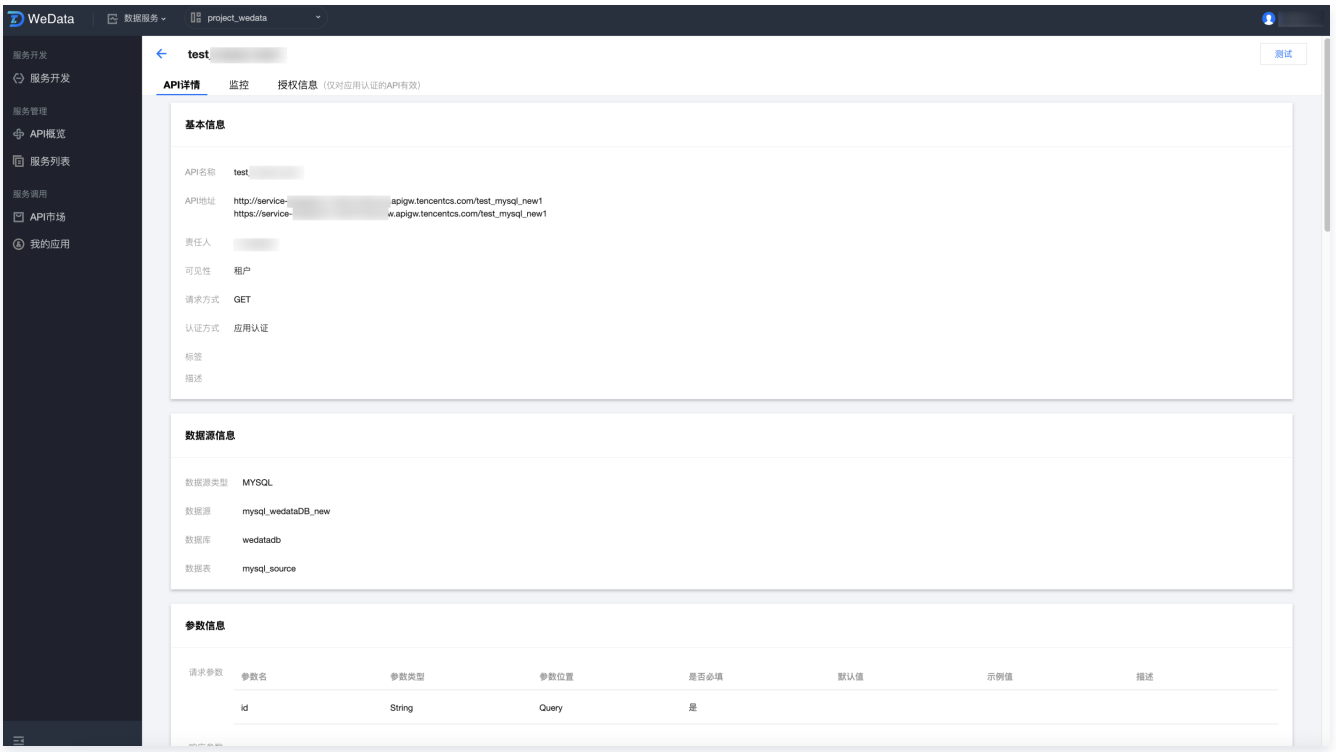
API Details and Monitoring

©2013-2025 Tencent Cloud. All rights reserved.

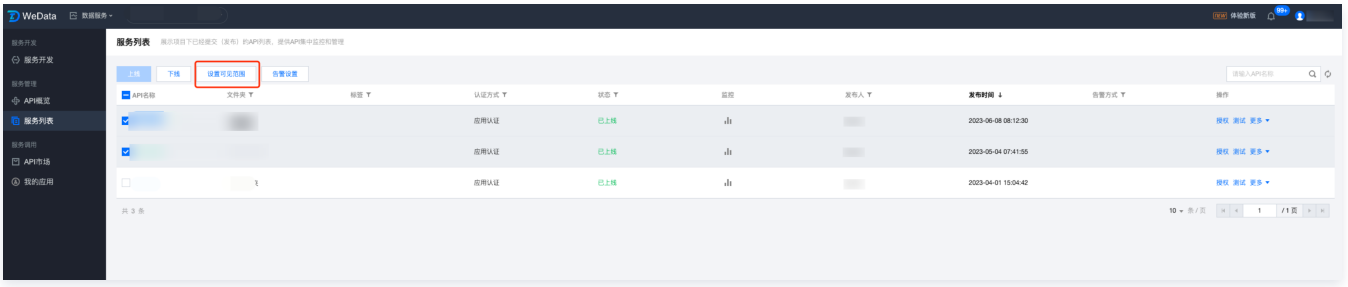
Go to **Data Service > Service List** to view all published service lists. It supports decommissioning, testing, monitoring, setting alerts, setting visibility, service development, authorization, and delete operations on the services.



- Click **API name** to view the API documentation. API details include basic information, data source information, parameter information, invocation examples, and error codes.



- In the service list, select the service and click **Set visibility range.**

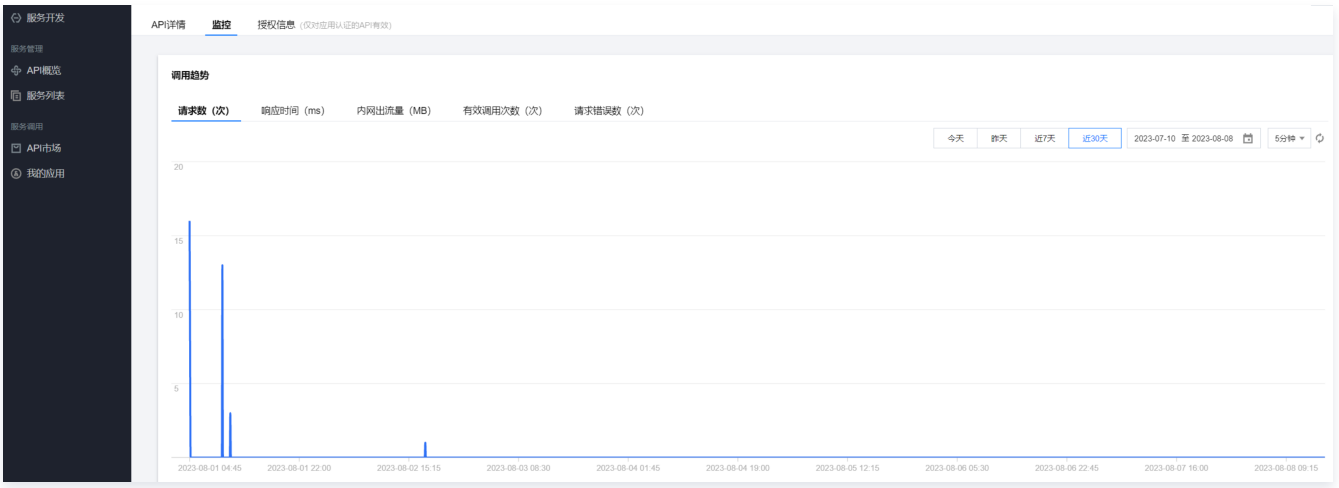


- Setting the API visibility range can choose tenant, project, or private.

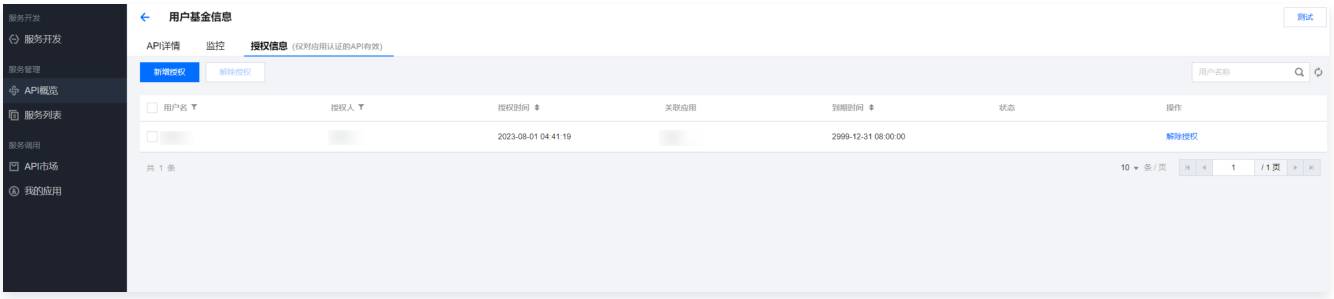
©2013–2025 Tencent Cloud. All rights reserved.



- In **Service List > API Name**, enter API details, click **Monitoring** to view the current API call status, including the number of requests, response time, intranet outbound traffic, effective invocation count, and trend of request error count.



- In **Service List > API Name**, enter API details, click **Authorization** to view the current API authorization status, including username, authorized person, authorization time, associated application, expiration time, and revoke authorization operation.



©2013–2025 Tencent Cloud. All rights reserved.

Data Service Authentication Feature Description

Last updated：2024-08-22 21:07:02

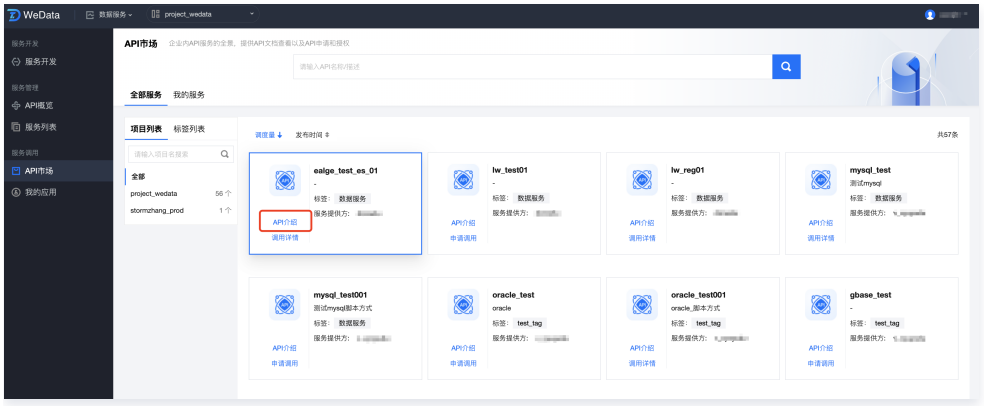
Feature Overview

API authentication refers to the support for user identity and permission verification when using APIs. It is an important feature to ensure the security and stability of APIs. The data service platform supports identity verification by using AppKey and AppSecret as access tokens to sign API requests.

Operation step

To apply for and invoke the identified API, an application must be submitted when the API authentication method is "application authentication".

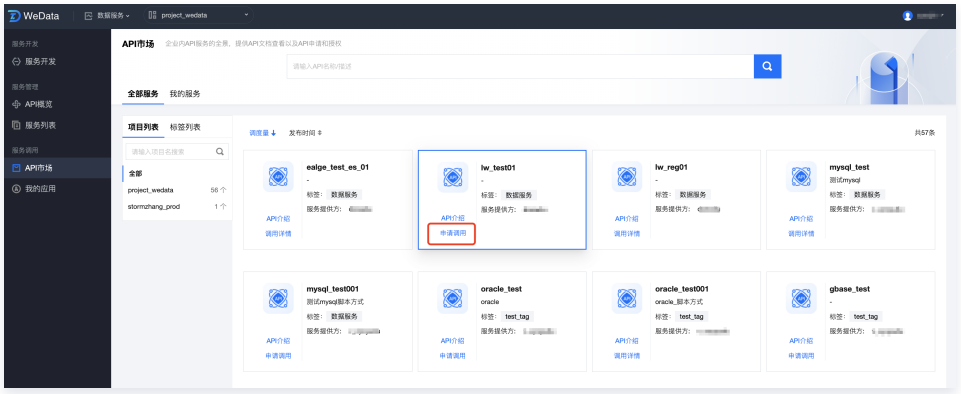
1. After identifying the API, click **API Introduction** to view the API documentation.



2. Click **API Application for Invocation**. There are two authentication methods: **application authentication** and **authentication-free**.

2.1 **Method 1:** For "application authentication" API application, follow these steps:

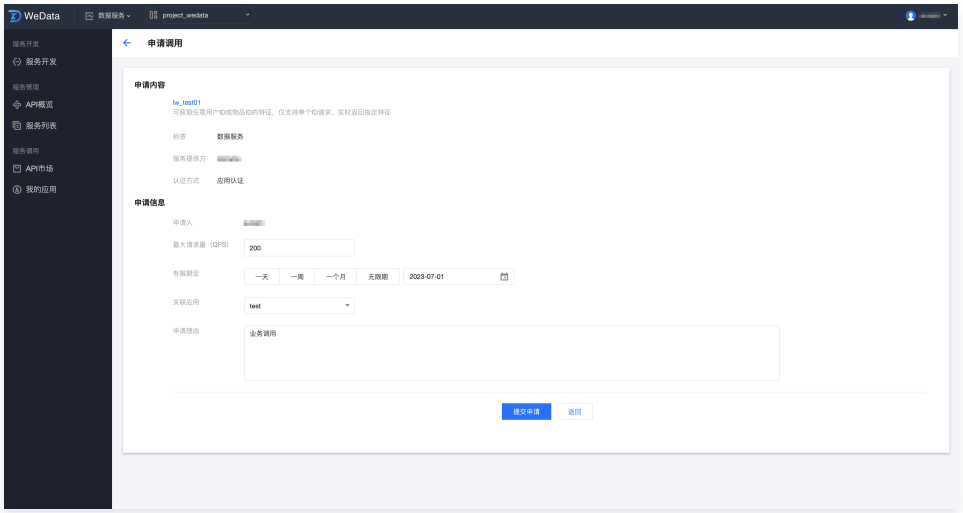
2.1.1 Click the application for invocation under the required data service.



2.1.2 Configure the application information on the page.

- Application Content: Confirm the API details, tags, service provider, and appropriateness of the authentication method.
- Application Information: Confirm the applicant, configure the maximum number of requests, validity period, associated applications, and application reason.

©2013-2025 Tencent Cloud. All rights reserved.



2.1.3 After configuration, click **Submit Application**. For "application authentication" APIs, Wedata currently uses the application authentication method of the API Gateway itself.



2.1.4 Application Credentials: After the applicant submits a request to the API provider, the API provider will issue App credential information (including Application ID, ApiAppKey, and ApiAppSecret) once the application is configured via **My Applications** > **Create New Application** > **Application Configuration Completion**.

- Calculate the signature. The method to calculate the signature varies by gateway type: API Gateway or Cloud Native Gateway. Use the signing key to sign the request content.

- Initiate the invocation using the API signature. For specific invocation information, refer to API Details.

2.2 **Method 2:** For "authentication-free" APIs, no application is needed; you can directly invoke them.

©2013-2025 Tencent Cloud. All rights reserved.



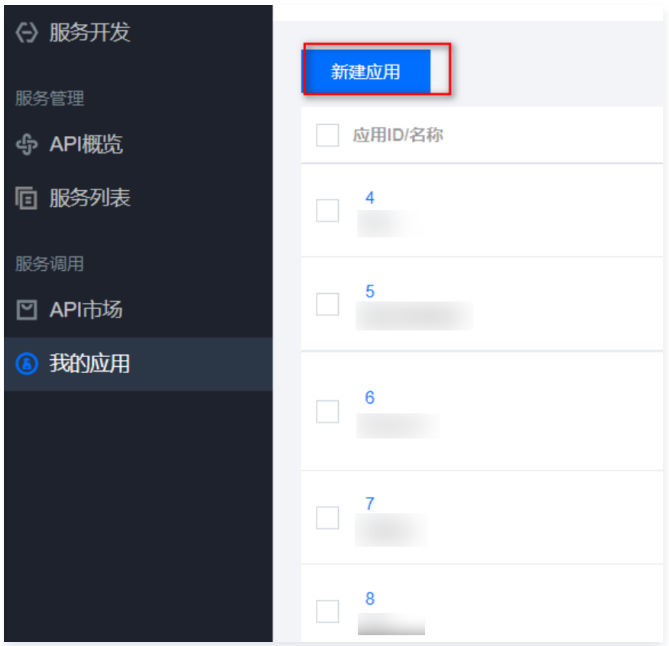
API key

Operation scenarios

Users create keys to provide AppKey, AppSecret, and other access tokens for the API service.

Operation step

1. Click on the left side **Service Invocation > My Applications**, then click **Create Application,** configure the application name and description, to complete the creation of the application, automatically generating AppKey and AppSecret.



2. Enter the application name and description.

©2013–2025 Tencent Cloud. All rights reserved.



3. Application list.

- The application list displays the information of the created applications, including application ID, application key, description, creation time, modification time, and provides functions to view callable APIs, update applications, and delete applications.
- Parameter description:

Parameter	Description
Application ID/Name	Application ID number and English name
Application Key	ApiAppKey and ApiAppSecret need to be included as parameters in the Header during requests. The method for calculating signatures varies across different gateway types: API Gateway, Cloud Native Gateway.
Description	Application Description
Creation Time	Application Creation Time
Modification Time	Last Modification Time of Application Configurations
Callable APIs	Click to view the callable APIs under the application
Update	Updating the application AppSecret will affect the APIs called by the application, please operate with caution.
Delete	Delete the selected application

©2013–2025 Tencent Cloud. All rights reserved.

Data Service Log Analysis Feature Description

Last updated: 2024-08-22 21:07:23

> **Note:**
> Scope: This feature is only applicable to Tencent Cloud Native Gateway and is not used for Legacy API Gateway (soon to be discontinued).

Operation step

Before using the log analysis feature, you need to install the log plugin on the API Gateway instance to report call logs to WeData CLS.

Install the plugin

> **Note:**
> Each gateway instance requires installing the plugin. For example, if you have purchased three gateway instances, you need to install it three times.

1. Go to the Tencent Cloud Microservice Engine console and select **Cloud Native Gateway** > Plugin Management page.
2. Select the gateway instance that needs the plugin installation, switch to the Custom Plugin tab, and click the **Upload Plugin** button.

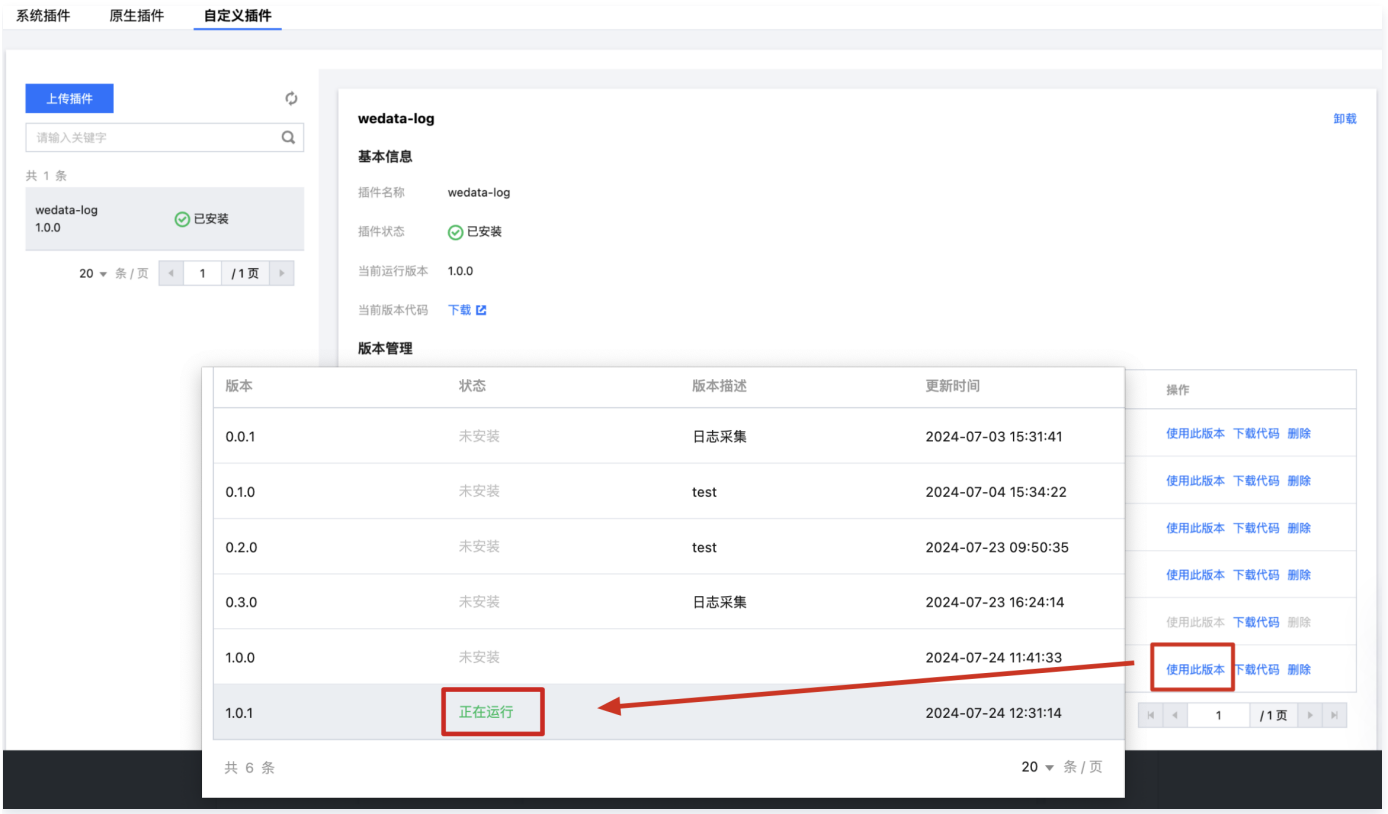


3. On the upload plugin page, fill in the relevant information:
   - **Upload zip package:** Upload the plugin installation package, click here to download the plugin package .
   - **Version number:** You can input any version number, suggested: 1.0.0.
   - **Version description:** Provide a brief description of the version.

©2013-2025 Tencent Cloud. All rights reserved.



4. Click **Use this version** , the plugin will start installing. Once installed, the status will change to **Running** .



5. After the plugin installation is complete, in the **Service Monitoring** > **Log Analysis** list, click **Update Log Plugin** to establish log delivery connection.



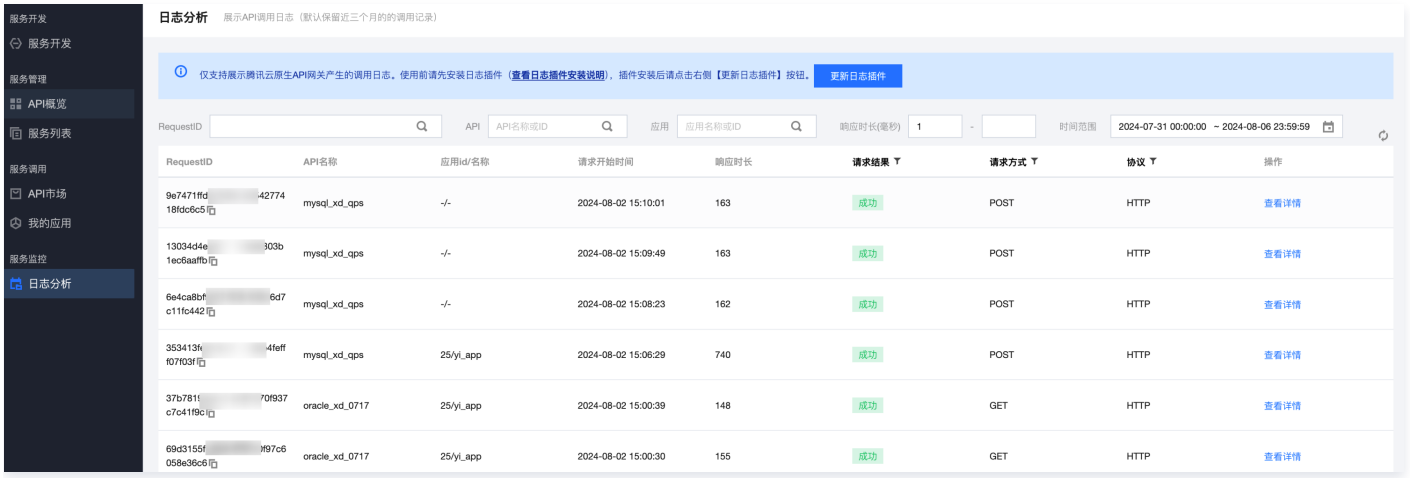
©2013–2025 Tencent Cloud. All rights reserved.

> **Note:**
> Each gateway instance requires installing the plugin. For example, if you have purchased three gateway instances, you need to install it three times.

View log list

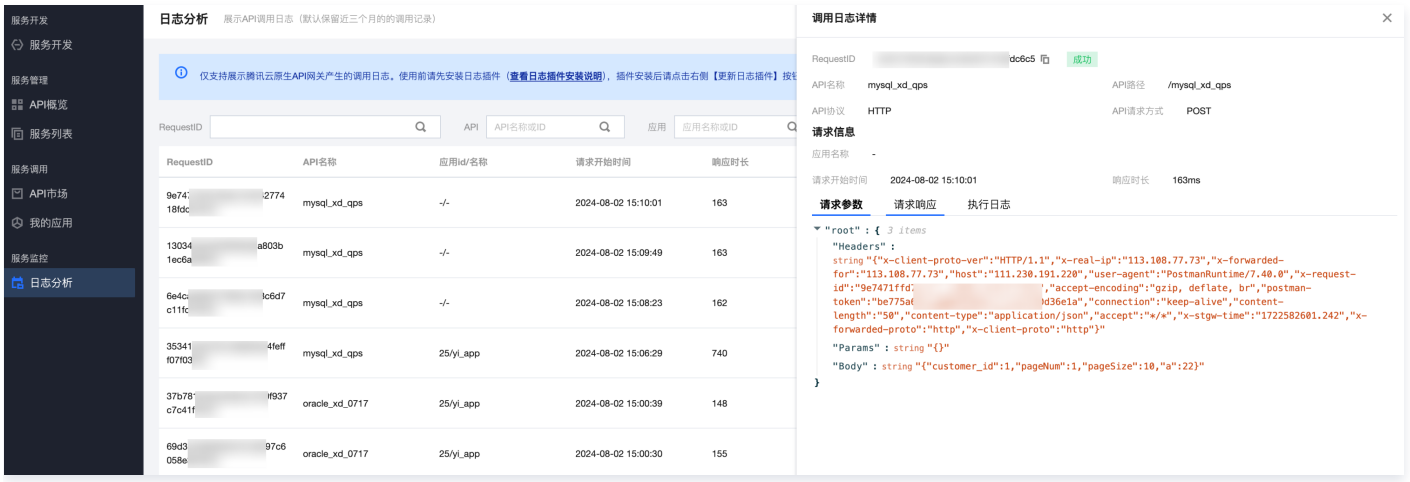
> **Warning:**
> API call logs are stored for three months by default and will be automatically cleaned after three months.

After calling the API, you can view the API call records on the **Data Service** > **Log Analysis** page.



View log details

You can view log details in the log list, including request parameters and request response.



©2013–2025 Tencent Cloud. All rights reserved.

Practice Tutorial
Script Mode Advanced SQL Usage Guide

Last updated: 2024-08-24 18:24:36

Advanced SQL (Mybatis Syntax) Examples

Data services support generating APIs through script mode. In script mode, both Basic SQL and Advanced SQL are supported. Advanced SQL includes common Mybatis tag types such as if, choose, when, otherwise, trim, foreach, and where, allowing you to flexibly implement null checks, multi-value iteration, dynamic table lookups, dynamic sorting, and complex query logic such as aggregation.
Additionally, Advanced SQL is compatible with some Mybatis syntax. Therefore, when using Mybatis syntax, pay attention to the use of "$" and "#" symbols. This article introduces code examples for common scenarios.

Basic SQL Example:

```
select
country, //select field as the API return field
name as username, //When there is a field alias, the alias serves as the return field
max(age) as max_age //Database functions are supported
from
 db_name.table_name   //The database table to be queried
where
hobby=#{hobby} //hobby corresponds to the API request parameter, formatted as #{request parameter}
group by country,name
```

Note

1. Please refer to the example specifications to input SQL. After inputting, the SQL will automatically parse into API request and response parameters.
2. Once the parameters are parsed, please complete the parameter definition to generate a complete API document for the user.
3. SQL supports multi-table joins. When writing, pay attention to the syntax of the chosen database.

Advanced SQL Examples:

Example 1: Sort the results by different table fields based on conditions

```
SELECT
col_a AS col_a,
col_b AS col_b,
col_c AS col_c
FROM     db_name.table_name
<choose>
    <when test='var == 1'>
      ORDER BY col_a
    </when>
    <when test='var == 2'>
      ORDER BY col_b
    </when>
    <when test='var == 3'>
      ORDER BY col_a, col_c
    </when>
    <when test='var == 4'>
      ORDER BY col_c, col_a
    </when>
```

©2013-2025 Tencent Cloud. All rights reserved.

```
</choose>
```

Request parameters:

Parameter Name	Parameter Type	Parameter Position	Operator	Required	Default Value	Sample Value	Description
var	String	Query	=	Yes	1	1	Sort Parameter

Return Parameters:

Parameter Name	Binding Field	Parameter Type	Sample Value	Description
col_a	col_a	String	abc	Field a
col_b	col_b	String	abc	Field b
col_c	col_c	String	abc	Field c

Example 2: Control queries to different data tables based on conditions

```
select
col_a as col_a,
col_b as col_b,
col_c as col_c
from
<choose>
    <when test='var == 1'>
      db_name.table_name_01
    </when>
    <when test='var == 2'>
      db_name.table_name_02
    </when>
</choose>
```

Request parameters:

Parameter Name	Parameter Type	Parameter Position	Operator	Required	Default Value	Sample Value	Description
var	String	Query	=	Yes	1	1	Query table parameters: 1,db_name.table_name_01 2,db_name.table_name_02

Return Parameters:

Parameter Name	Binding Field	Parameter Type	Sample Value	Description
col_a	col_a	String	abc	Field a
col_b	col_b	String	abc	Field b
col_c	col_c	String	abc	Field c

Example 3: Use field value checks to control whether to retain where query conditions, i.e., query logic for non-mandatory request parameters

```
select
col_a as col_a,
col_b as col_b,
col_c as col_c
```

©2013–2025 Tencent Cloud. All rights reserved.

```
from db_name.table_name_01
<where>
    <if test='list!=null'>
    col_a in
      <foreach collection="list" open="(" close=")" separator="," item="col_a_item">
        #{col_a_item}
      </foreach>
    </if>
</where>
```

Request parameters:

Parameter Name	Parameter Type	Parameter Position	Operator	Required	Default Value	Sample Value	Description
list	StringList	Query	=	No	a,b,c	a,b,c	col_a input parameters set

Response parameters:

Parameter Name	Binding Field	Parameter Type	Sample Value	Description
col_a	col_a	String	abc	Field a
col_b	col_b	String	abc	Field b
col_c	col_c	String	abc	Field c

Special fields require escape processing

In mybatis, SQL is written in xml files. Parsing < ,>,<=,>= in xml will cause errors. Use escape characters in such cases.

<	<=	>	>=	&	'	"
<	<=	>	>=	&	'	"

©2013–2025 Tencent Cloud. All rights reserved.

Appendix
API Error Code Description

Last updated: 2024-08-22 21:08:50

WeData Data Service returns different request results during calls. When an error occurs, an error code is returned. Through the error code information, users can quickly identify the issue and troubleshoot errors independently. Below are the descriptions and handling suggestions for the error codes.

Error code	Description	Suggestion
ActionOffline	API has been decommissioned	No
AuthFailure.InvalidAuthorization	The Authorization in the request header does not meet Tencent Cloud standards.	No
AuthFailure.InvalidSecretId	Invalid key (not TencentCloud API key type)	Third-party service error
AuthFailure.MFAFailure	MFA error	No
AuthFailure.SecretIdNotFound	The key does not exist	Check if the key has been deleted or disabled. If the status is normal, please check if the key is entered correctly, ensuring there are no spaces at the beginning or end
AuthFailure.SignatureExpire	The signature expired	Timestamp and server time difference must not exceed five minutes. Please check if the local time is synchronized with standard time
AuthFailure.SignatureFailure	Signature error	Signature calculation error. Please refer to the signature method documentation in the call method to check the signature calculation process
AuthFailure.TokenFailure	Token error	No
AuthFailure.UnauthorizedOperation	Unauthorized request	No
DryRunOperation	DryRun operation indicates the request would be successful, but the DryRun parameter was included	No
FailedOperation	Operation failed	No
InternalError	An internal error occurs	No
InvalidAction	API does not exist	No
InvalidParameter	Parameter error (including parameter format, type, etc.)	No
InvalidParameterValue	Parameter value error	No
InvalidRequest	The multipart format of the request body is incorrect	No
IpInBlacklist	IP address in blocklist	No
IpNotInWhitelist	IP address not in allowlist	No
LimitExceeded	Exceeded quota limit	No

©2013-2025 Tencent Cloud. All rights reserved.

MissingParameter	Missing parameters	No
NoSuchProduct	Product does not exist	No
NoSuchVersion	API version does not exist	No
RequestLimitExceeded	Number of requests exceeds frequency limit	No
RequestLimitExceeded.GlobalRegionUinLimitExceeded	Root account exceeds frequency limit	No
RequestLimitExceeded.IPLimitExceeded	IP frequency limit	No
RequestLimitExceeded.UinLimitExceeded	Root account frequency limit	No
RequestSizeLimitExceeded	Request packet exceeds size limit	No
ResourceInUse	Resource is occupied	No
ResourceInsufficient	Insufficient resources	No
ResourceNotFound	The resource does not exist	No
ResourceUnavailable	Resource unavailable	No
ResponseSizeLimitExceeded	Response packet exceeds size limit	No
ServiceUnavailable	The current service is temporarily unavailable	No
UnauthorizedOperation	Unauthorized operation	No
UnknownParameter	Unknown parameter error. Multiple undefined parameters from the user will cause an error	No
UnsupportedOperation	Operation is not supported	No
UnsupportedProtocol	HTTP(S) request protocol error. Only GET and POST requests are supported	No
UnsupportedRegion	The interface does not support the specified region	No

©2013–2025 Tencent Cloud. All rights reserved.