

# **Data Lake Compute**

## **Operation Guide**

### **Product Documentation**



## Copyright Notice

©2013-2024 Tencent Cloud. All rights reserved.

Copyright in this document is exclusively owned by Tencent Cloud. You must not reproduce, modify, copy or distribute in any way, in whole or in part, the contents of this document without Tencent Cloud's the prior written consent.

## Trademark Notice



All trademarks associated with Tencent Cloud and its services are owned by Tencent Cloud Computing (Beijing) Company Limited and its affiliated companies. Trademarks of third parties referred to in this document are owned by their respective proprietors.

## Service Statement

This document is intended to provide users with general information about Tencent Cloud's products and services only and does not form part of Tencent Cloud's terms and conditions. Tencent Cloud's products or services are subject to change. Specific products and services and the standards applicable to them are exclusively provided for in Tencent Cloud's applicable terms and conditions.

# Contents

## Operation Guide

### Data Exploration

- SQL Editor

- Data Query Task

  - SELECT Task

  - Querying Partition Table

  - Querying JSON Data

  - Querying Data from Other Sources

  - Using View

  - INSERT INTO

  - Querying Script Parameters

- Query Script Analysis

### Data Management

- Data Catalogs and DMC

- Data Table Management

- Data View Management

- Function Management

- Partition Field Policy

### Data Job

- Overview

- Configuring Data Access Policy

- Creating Data Job

- Managing Data Job

### Task History

### Engine Management

- SuperSQL Engine

  - SuperSQL Engine Overview

  - Purchasing Private Data Engine

  - Renewing SuperSQL Engine

  - Managing Private Data Engine

  - Disaster Recovery Cluster

  - Engine Kernel Version

  - Engine Network Configuration

  - Associating Tag with Private Engine Resource

  - Engine Local Cache

Custom Task Scheduling Pool

Ops Management

Permission Management

CAM Service

Permission Overview

User and Work Group

Sub-Account Permission Management

Storage Configuration

Managed Storage Configuration

Binding a Metadata Acceleration Bucket

Audit Log

Monitoring and Alarms

Data Engine Monitoring

Data Job Monitoring

Access Point Gateway Engine Monitoring

Monitoring Alarm Configuration

Query Script Management

System Restraints

Metadata Information

Computing Task

# Operation Guide

## Data Exploration

### SQL Editor

Last updated : 2024-07-17 17:36:45

The SQL editor provided by Data Lake Compute (DLC) supports data querying using unified SQL statements, compatible with SparkSQL. You can complete data query tasks using standard SQL. For detailed syntax instructions, refer to [SQL Syntax](#).

You can access the SQL editor through data exploration, where you can perform simple data management, multi-session data queries, query record management, and download record management.

## Data Management

Data management supports adding data sources, managing databases, and managing data tables.

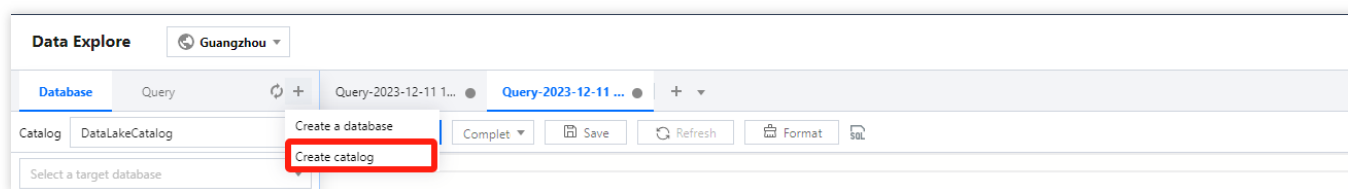
### Creating a data catalog

Currently, Data Lake Compute supports the management of COS and EMR Hive data catalogs. The directions are as follows:

1. Log in to the [Data Lake Compute console](#) and select the service region. You need to have the admin permission.
2. Select **Data Explore** on the left sidebar, hover over



on the **Database & table** tab, and click **Create catalog**.



For detailed directions, see [Querying Data from Other Sources](#).

### Managing a database

You can create, delete, and view the details of a database in the SQL editor.


### Managing a data table

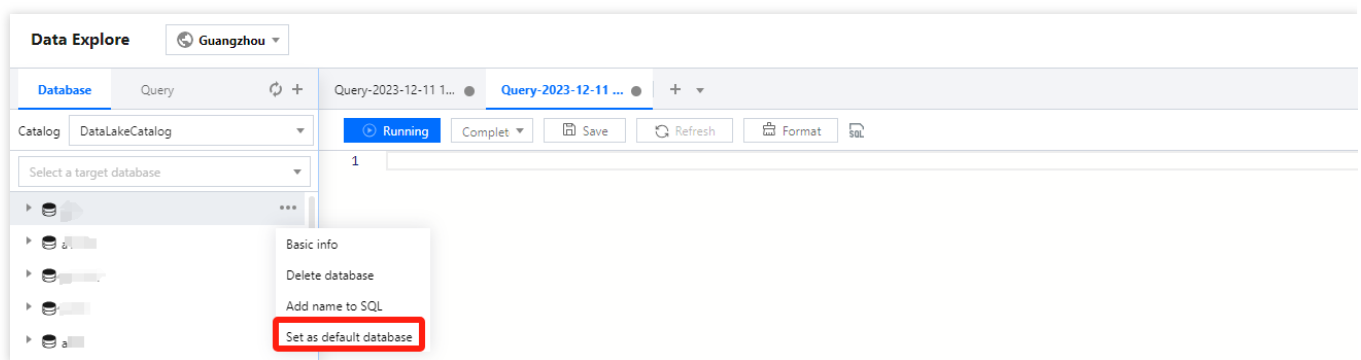
You can create, query, and view the details of a data table in the SQL editor.

## Changing the default database

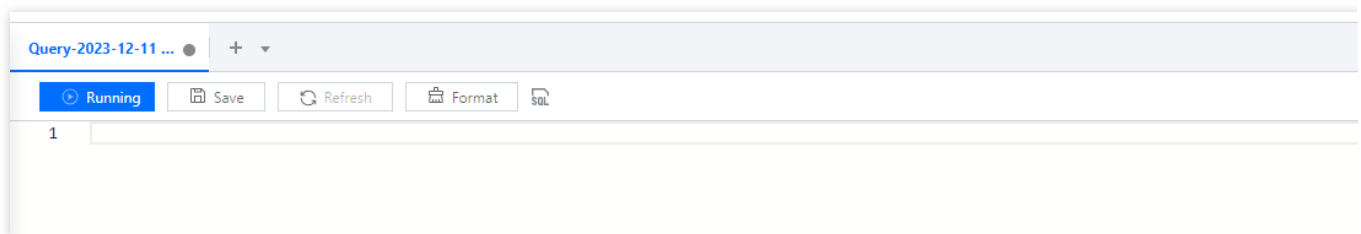
You can use the SQL editor to specify the default database for query tasks. If no database is specified in a query statement, the query will be executed in the default database.

1. Log in to the [Data Lake Compute console](#) and select the service region.
2. Select **Data Explore** on the left sidebar, hover over the target database name, click

, and click **Set as default database** to set the database as the default database.



3. You can also change the default database in the **Default database** selection box.



## Data Query

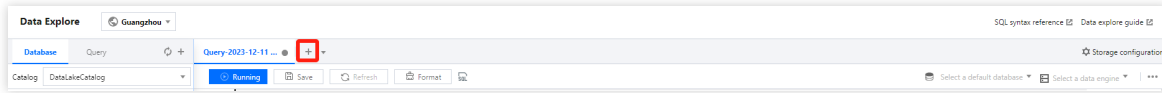
### Add Query Page

The SQL editor supports adding multiple pages for data querying, with each query page having independent configurations (default database, computation engine used, query records, etc.). This facilitates users in running and managing multiple tasks.

You can create a new query page by clicking on the



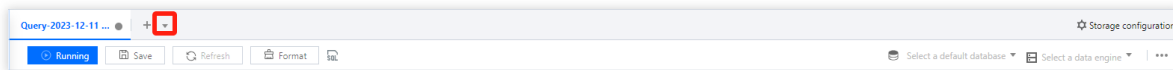
icon, and switch the editor interface by clicking on the tab bar.



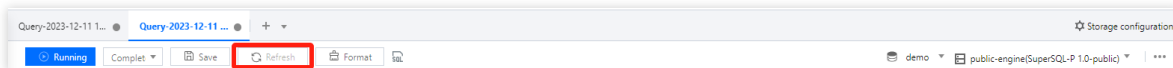
For your convenience, you can save frequently used query pages by clicking the **Save** button. You can also quickly open your saved pages by clicking the



icon.

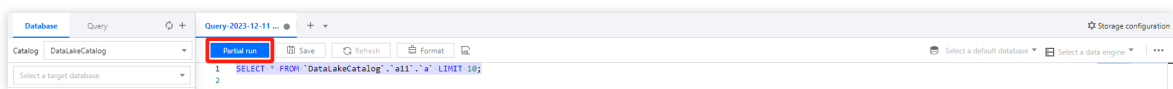


For saved query page information, you can click the **Refresh** button to update and synchronize the saved information, ensuring the accuracy of the query statement.



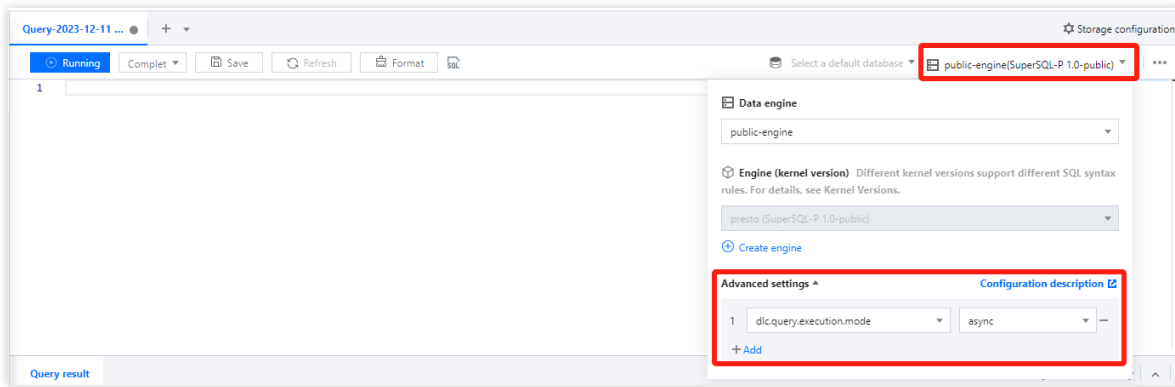
The editor supports running multiple different SQL statements simultaneously. Clicking the **Run** button will execute all SQL statements within the editor, simultaneously dividing them into multiple SQL tasks.

If you need to run a portion of the statement, select the required statement and click **Partial run**.



## Engine Parameter Configuration

After selecting the data engine, you can configure parameters for the data engine. After selecting the data engine, click **Add** in Advanced Settings to configure.



The currently supported configuration parameters are as follows:

Engine	Configuration name	Start Value	Configuration Notes
SparkSQL	spark.sql.files.maxRecordsPerFile	0	The maximum number of records that can be written to a single file. If this value is zero or negative, there are no restrictions.
	spark.sql.autoBroadcastJoinThreshold	10MB	Configure the maximum byte size of the table of all working nodes displayed when executing a connection. By setting this value to "-1", the display can be disabled.
	spark.sql.shuffle.partitions	200	Default Partition Count.
	spark.sql.sources.partitionOverwriteMode	static	When the value is set to static, all qualifying partitions will be deleted prior to executing the overwrite operation. For instance, in a partitioned table, there is a partition "2022-01". When using the INSERT OVERWRITE statement to write data to the "2022-02" partition, the data in the "2021-01" partition will also be overwritten. When the value is set to 'dynamic', partitions will not be deleted in advance, but will be overwritten during runtime for those partitions where data is written.



	spark.sql.files.maxPartitionBytes	128MB	The maximum number of bytes to be packaged into a single partition when reading a file.
Presto	use_mark_distinct	true	Determines whether the engine redistributes data when executing the distinct function. If the distinct function is called multiple times in a query, it is recommended to set this parameter to false.
	USEHIVEFUNCTION	true	Determines whether to use Hive functions when executing a query; if you need to use Presto native functions, please set the parameter to false.
	query_max_execution_time	-	This setting is used to establish a query timeout. If the execution time of a query exceeds the set time, the query will be terminated. The units supported are d-day, h-hour, m-minute, s-second, ms-millisecond (for example, 1d represents 1 day, 3m represents 3 minutes).
	dlc.query.execution.mode	async	The engine query execution mode is set to async mode by default. In this mode, the task will perform a complete query calculation, save the results to COS, and then return them to the user, allowing the user to download the query results after the query is completed. Users can also change this value to sync. In sync mode, queries may not necessarily perform full calculations. Once partial results are available, they will be directly returned to the user by the engine, without being saved to COS. Therefore, users can achieve lower query latency and duration, but the results are only saved in the system for 30 seconds. This mode is recommended for

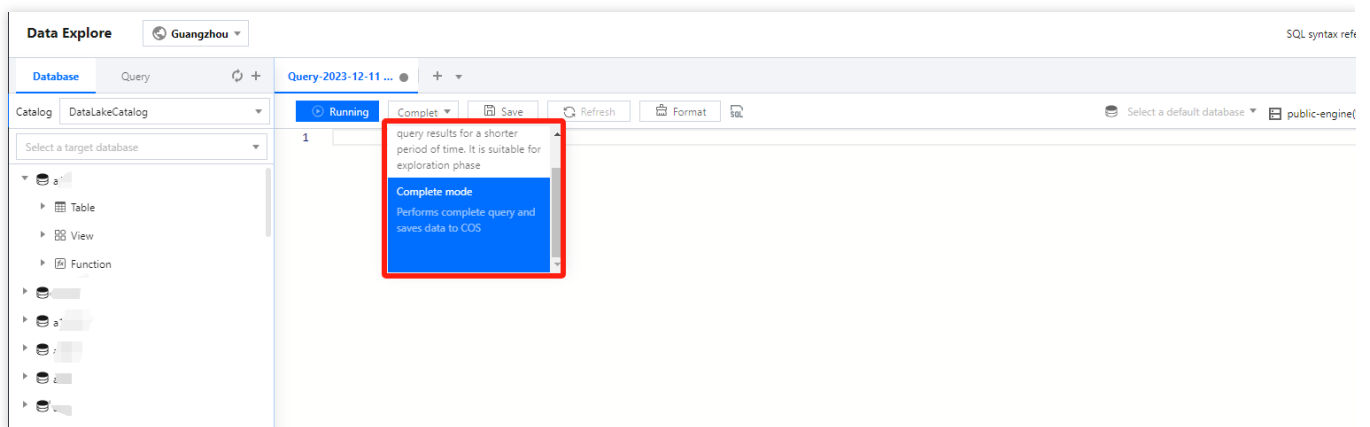
users who do not need to download the complete query results from COS, but expect lower query latency and duration, such as during the query exploration phase or BI result display.

### Presto Execution Mode

When the user selects the Presto engine, Data Exploration supports the user to choose to run in "Fast Mode" or "Full Mode".

Quick Query: This offers faster speed, but the query results cannot be persistently saved. It is suitable for the exploration phase.

Full Mode: Execute a full query and save the data to object storage.



### Search results

Through the SQL editor, you can directly view the query results. You can expand or collapse the display height of the query results by clicking the

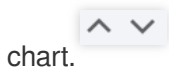
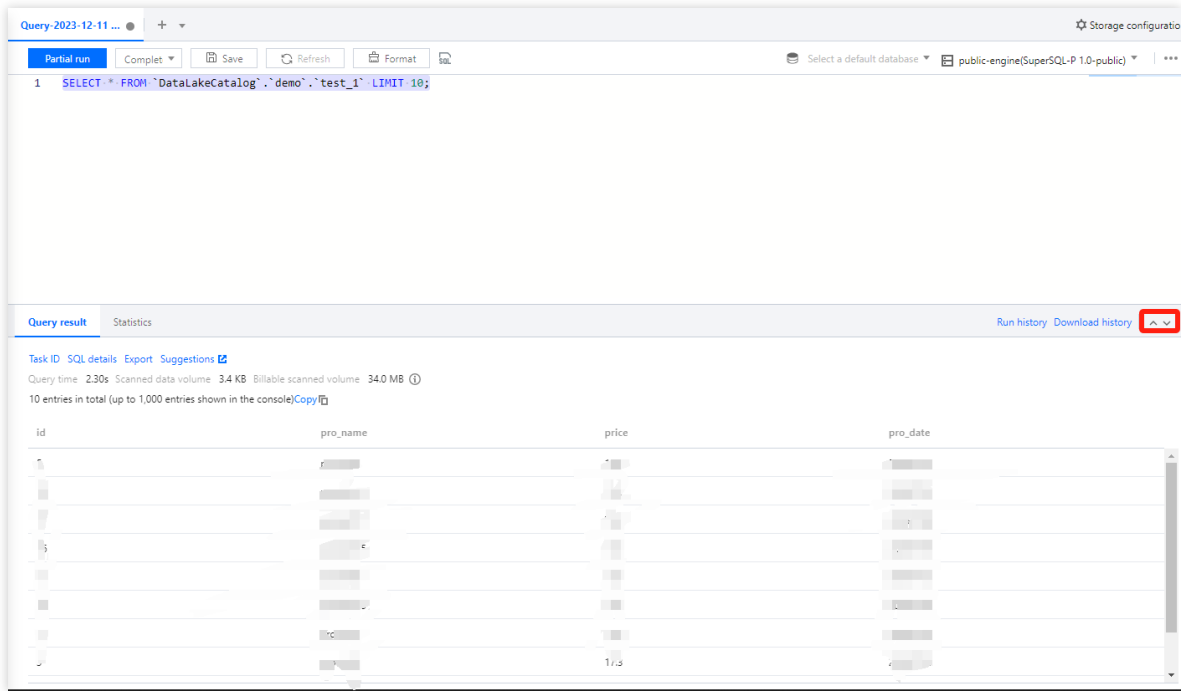
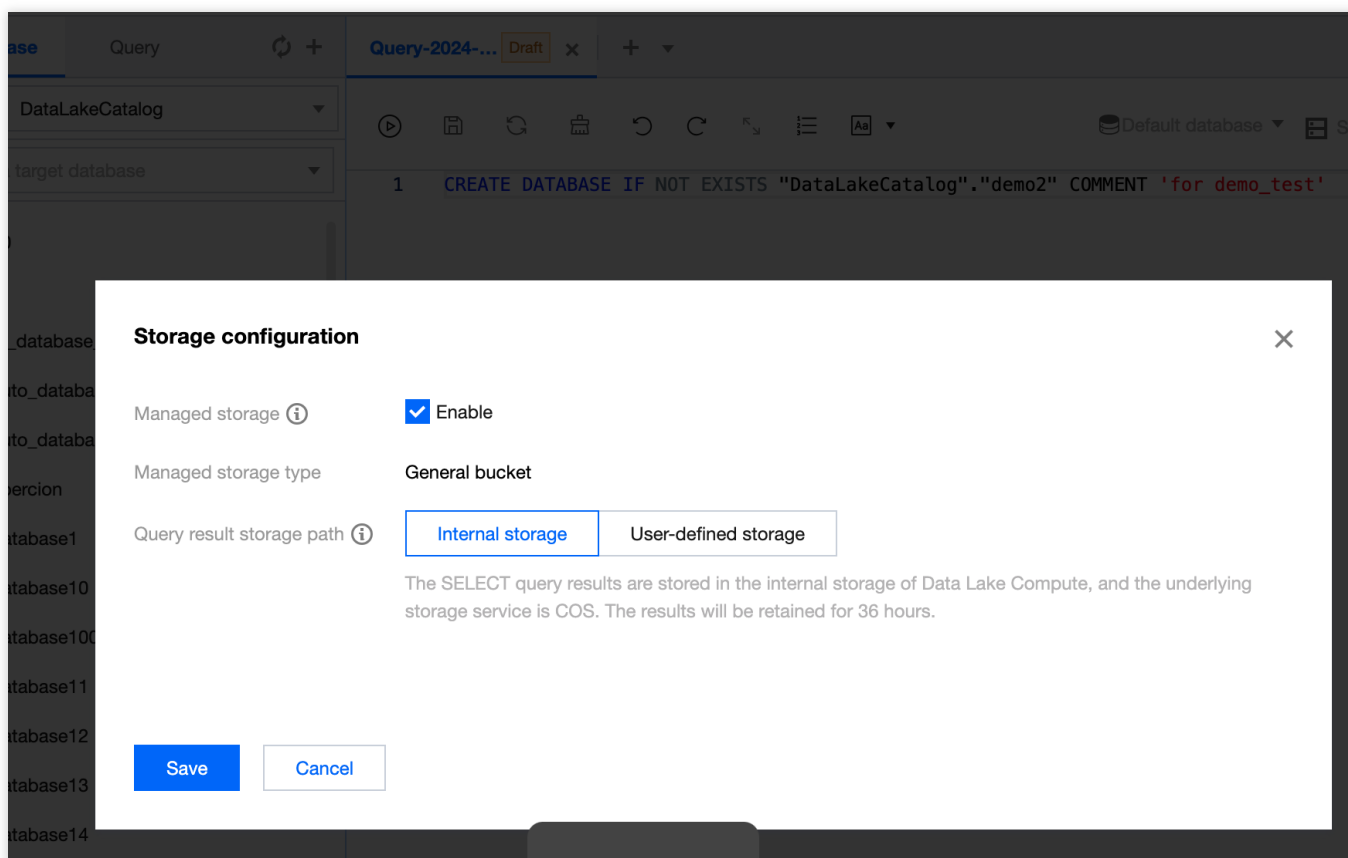


chart.



You can configure the query result storage directory through the configuration button in the upper right corner, supporting configuration to the COS path or built-in storage.



The console will return a maximum of 1000 results for a single task. If more results are needed, the API can be used. For instructions on API-related operations, refer to the [API Documentation](#).

Query results can be downloaded locally when no COS storage path is specified. For detailed instructions, refer to [Obtaining Task Results](#).

## Querying statistical data

The query results under the Presto engine and SparkSQL engine support the display of optimized quantification with different characteristics.

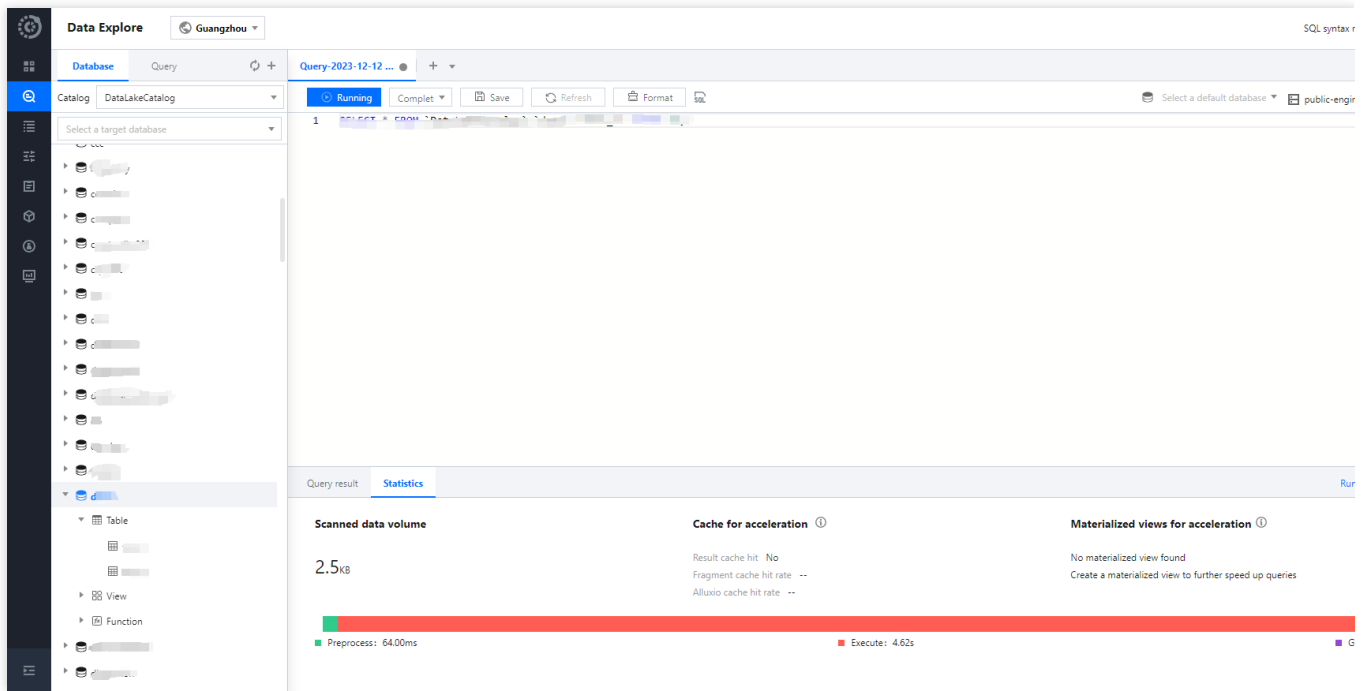
The SparkSQL engine supports viewing:

1. Data Scanning Volume
2. Cache Acceleration
3. Adaptive Shuffle
4. Materialized View Acceleration

The Presto engine supports viewing:

1. Data Scanning Volume
2. Cache Acceleration
3. Materialized View Acceleration

Click on the **Statistics** column to review the statistical data and optimization suggestions for the query results.

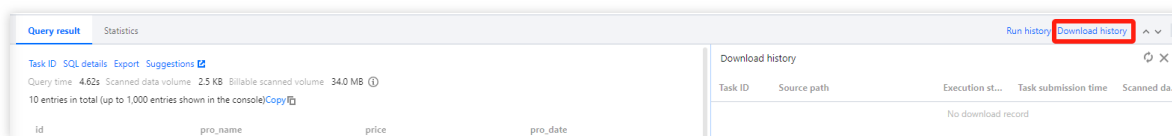


## Historical Queries

Each query page can save the running history of the past three months and supports viewing the query results of the past 24 hours. You can quickly find past task information through the running history. For detailed operations, refer to Task History Records.

## Download History Management

Each query result's download task can be viewed in the **Download history**, where you can check the status of the download task and related parameter information.



# Data Query Task

## SELECT Task

Last updated : 2024-07-17 16:04:41

You can query, analyze, and compute the data in a created database or data table with SQL statements.

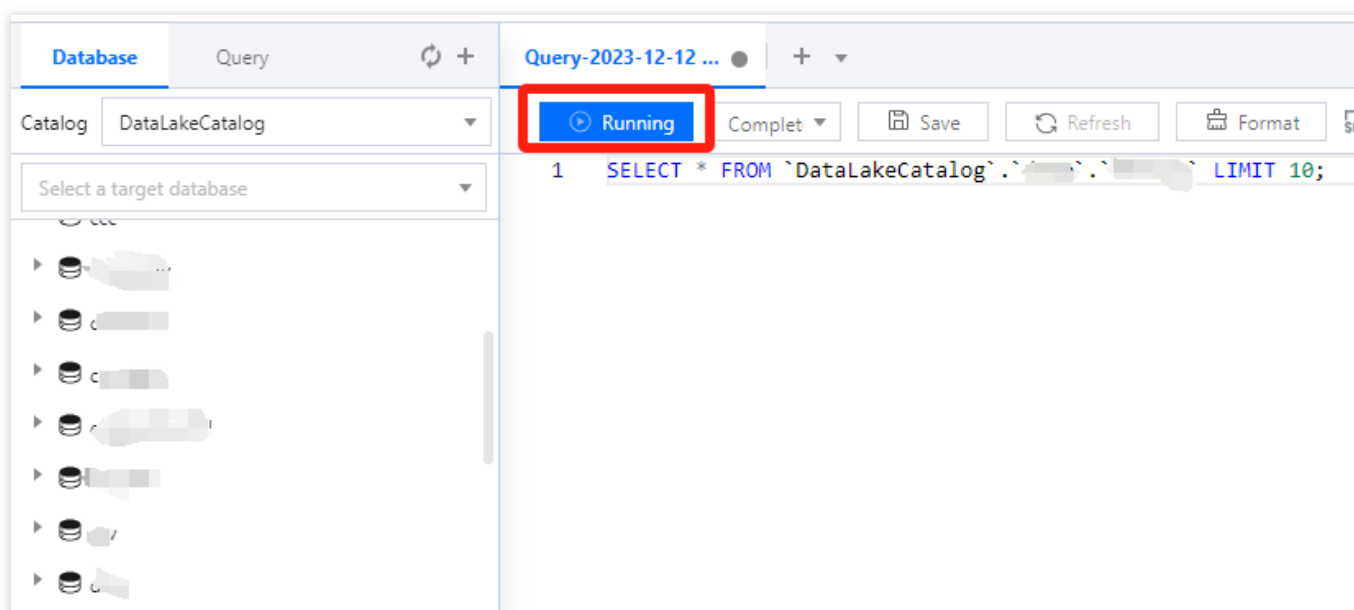
### Running a `SELECT` query task

1. Select the default database and compute resource.

You can select a default database. Then, when there is no database specified in a SQL statement, the statement will be executed in the default database.

You can select a public or private cluster as the compute resource.

2. Write a standard SQL statement and click **Running**.



In Data Lake Compute, a task can run for up to 30 minutes.

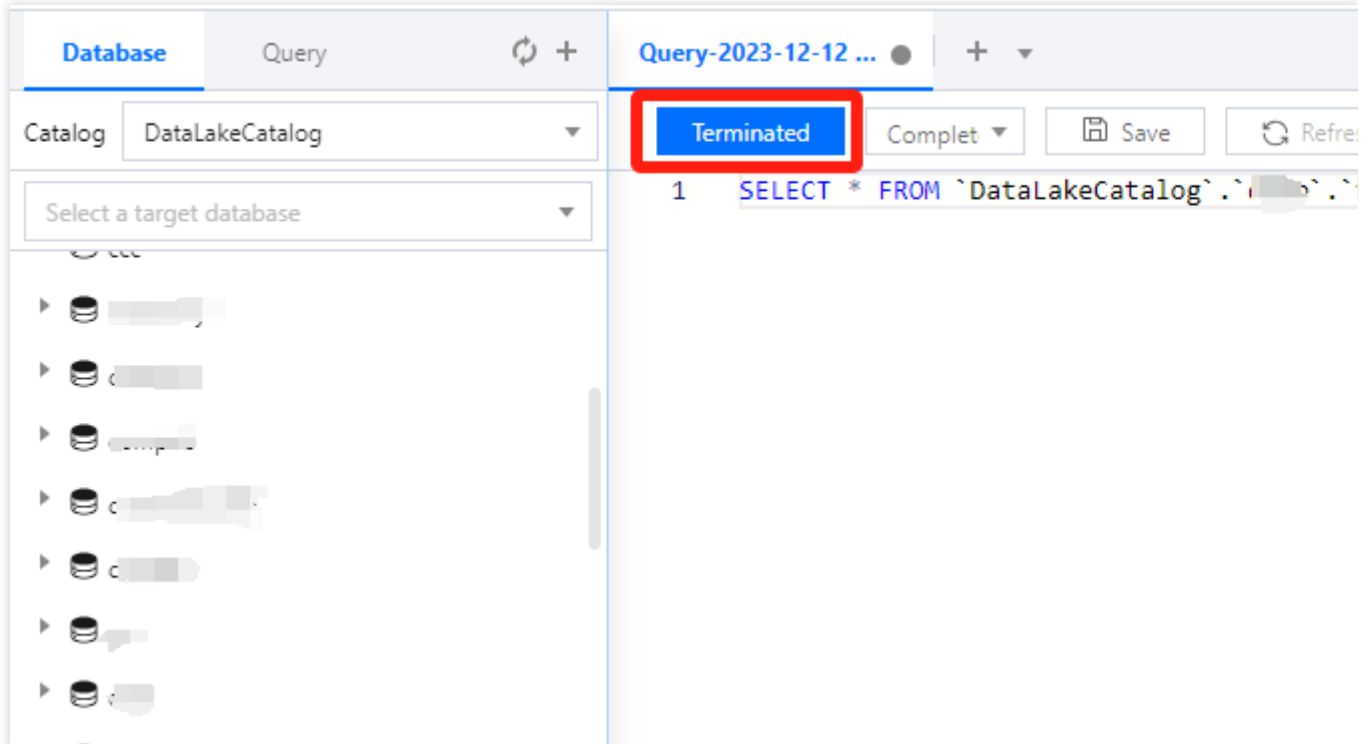
Data Lake Compute is serverless, so compute resources will be scheduled temporarily. It may take longer than usual to return the result of the first DML task.

3. The query result will be displayed in the console after the task is completed.

If you exit the console page, you cannot view the query result of a historical task there again. In this case, you can view the task result file in **Run history** or the query result COS bucket you configured.

## Canceling a running query task

During task running, the **Run** button becomes **Terminated**, which you can click to cancel the task. Then, Data Lake Compute will not return the query result but will calculate the scanned data volume. If you use the public engine, the scanned data volume will incur fees. For billing details, see [Billing Overview](#).



# Querying Partition Table

Last updated : 2024-07-17 16:17:22

Storing data in partition catalogs can greatly reduce the scanned data volume of a computing task in Data Lake Compute and thereby significantly enhance the computing performance. The general practice of data partitioning is to store data in different catalogs by time. For example, data generated on the same day can be stored in the same catalog, and catalogs can be organized in a "year-month-day" structure. In Data Lake Compute, a table and its partitions must adopt the same data format.

## Creating a Partition Table

To create a partition table, you need to specify the partition field in the table creation statement.

## Adding Partitioned Data

Specifying a partition during data table creation is only to configure the partition field and doesn't allow running a query statement immediately to get data. You need to add partitioned data to a data table. If new partitioned data is added to the data catalog, you also need to add the partition information to the data table.

### Manually adding a partition

Use the `ALTER TABLE ADD PARTITION` statement to add a specified partition catalog to a data table. If the partition catalog is compatible with the Hive partitioning rule (**partition column name=partition column value**), you don't need to specify the data path; otherwise, you need to.

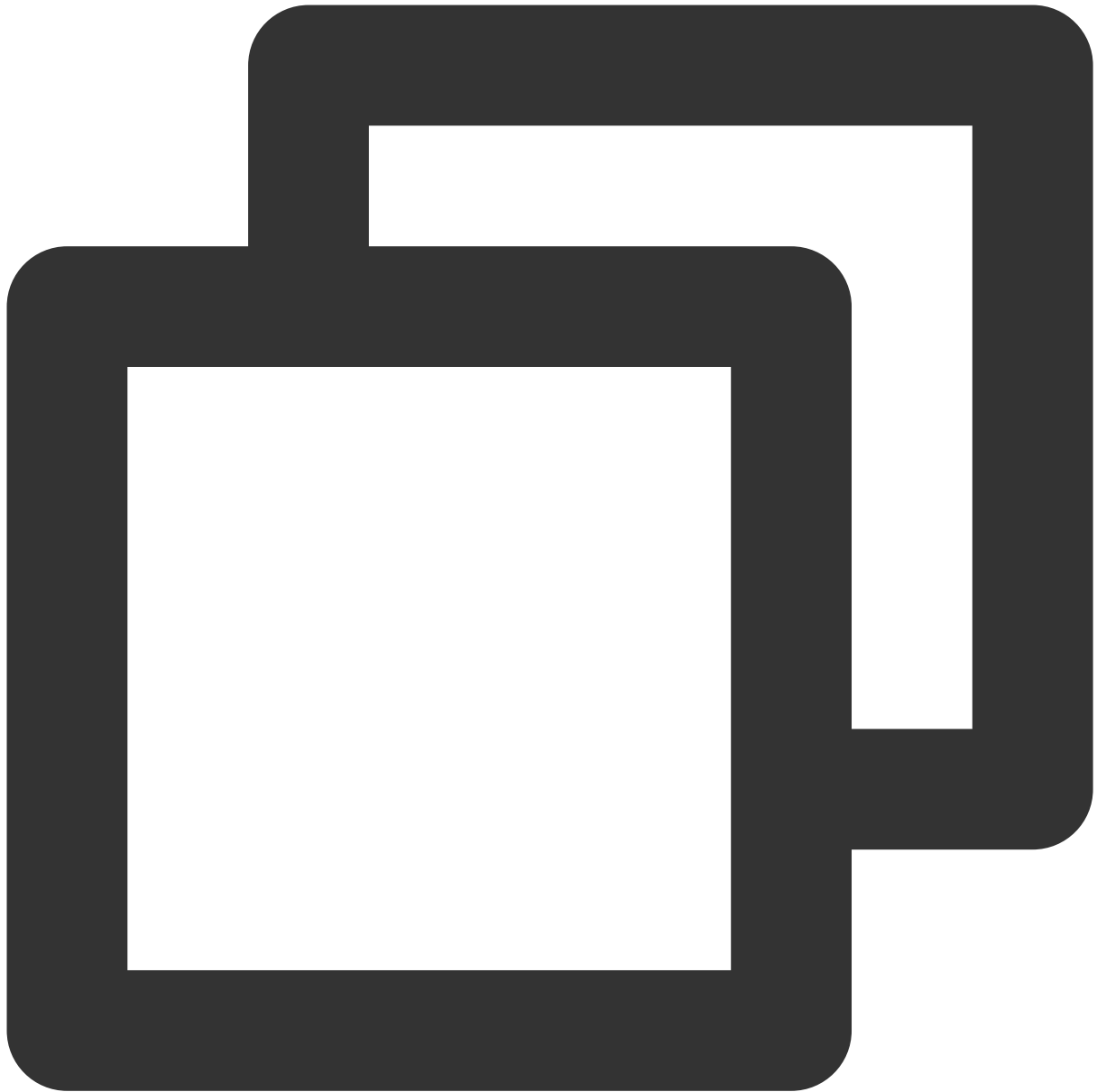
Sample 1: Adding a single partition catalog





```
ALTER TABLE tabel_demo ADD  
PARTITION (dt = '2021-01-01');
```

Sample 2: Adding multi-level nested partition catalogs



```
ALTER TABLE tabel_demo ADD  
PARTITION (year = '2021', month='01', day='01');
```

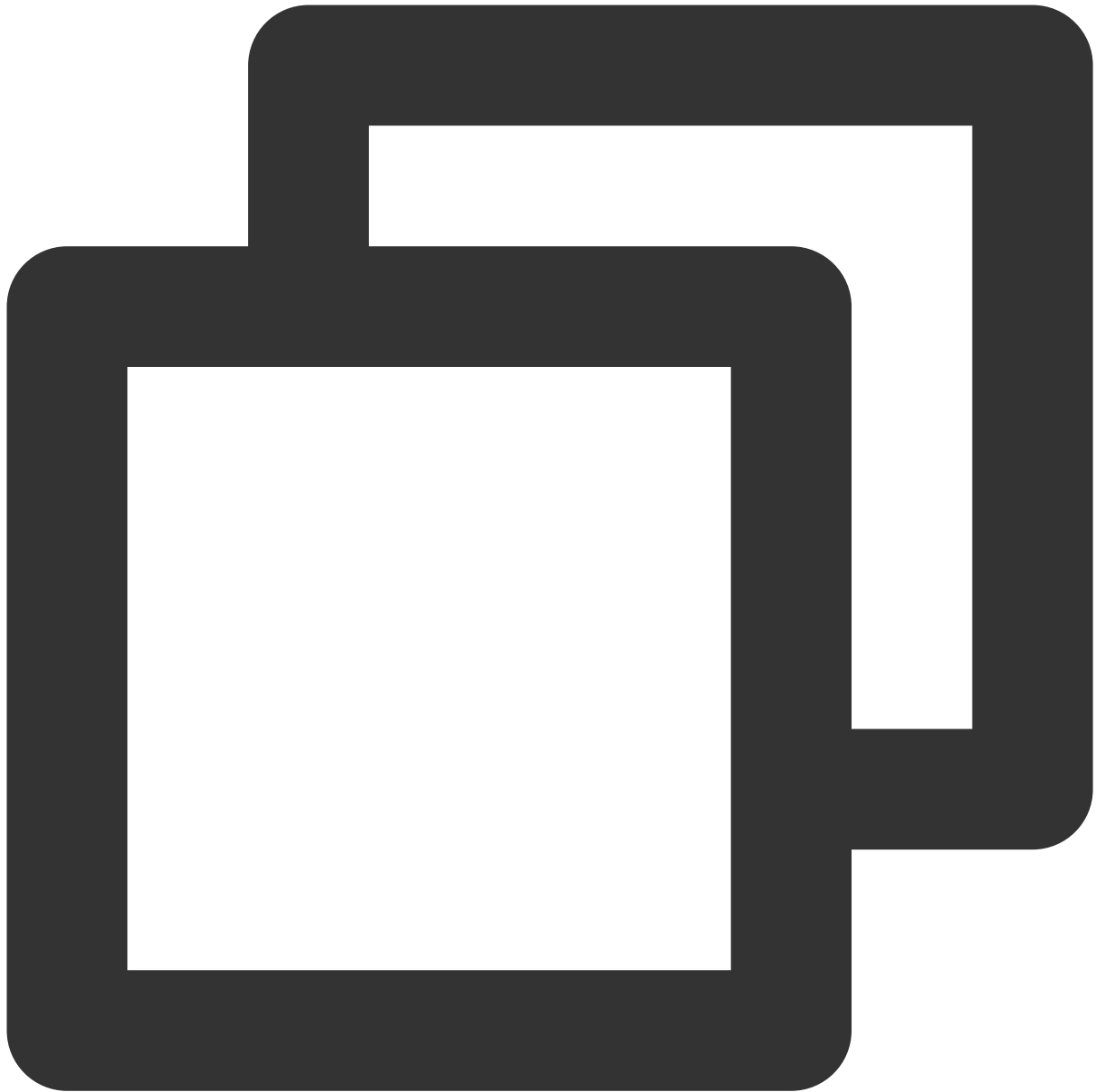
Sample 3: Displaying the specified partition path



```
ALTER TABLE tabel_demo ADD  
PARTITION (year = '2021', month='01', day='01') LOCATION 'cosn://tablea_demo' ;
```

### Automatically adding a partition

Use the `MSCK REPAIR TABLE` statement to scan the data catalog specified during table creation. If there is a new partition catalog, the system will automatically add the partitions to the metadata of the data table. Below is a sample:



```
MSCK REPAIR TABLE table_demo
```

## System Restraints

`MSCK REPAIR TABLE` only adds partitions to the metadata of the data table but does not delete them. To delete an added partition, run the `ALTER TABLE table-name DROP PARTITION` statement.

`MSCK REPAIR TABLE` is not recommended if the data volume is large, as the system will scan all the data, which may take a long time, cause the task to time out, and make the partition information of the data table incomplete.

A partition catalog must be compatible with the Hive partitioning rule of **partition column name=partition column value**; otherwise, use `ALTER TABLE ADD PARTITION` to load a partition.

Make sure that data of a table is stored in a separate folder. For example, if the `cosn://tablea_a` data in table A and the `s3://table_a/table_b` data in table B are stored in COS and both tables are partitioned by string, then `MSCK REPAIR TABLE` will add partitions of table B to table A. To avoid this, use separate folder structures, such as `cosn://tablea_a` and `cosn://tablea_b`.

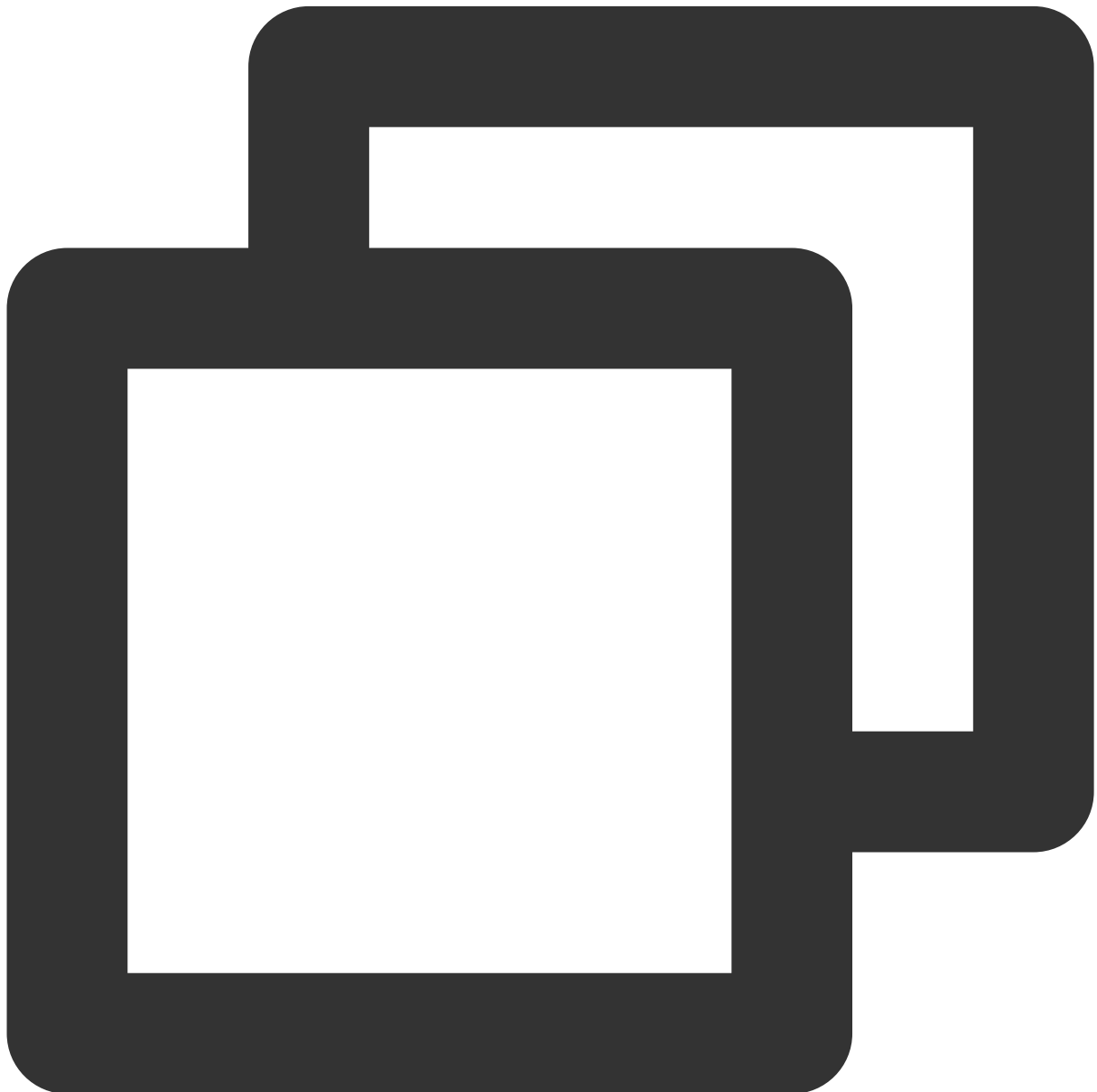
The statement may incur data read/write fees charged by COS. For more information, see [Billing Overview](#).

# Querying JSON Data

Last updated : 2024-07-17 16:18:53

## Query steps

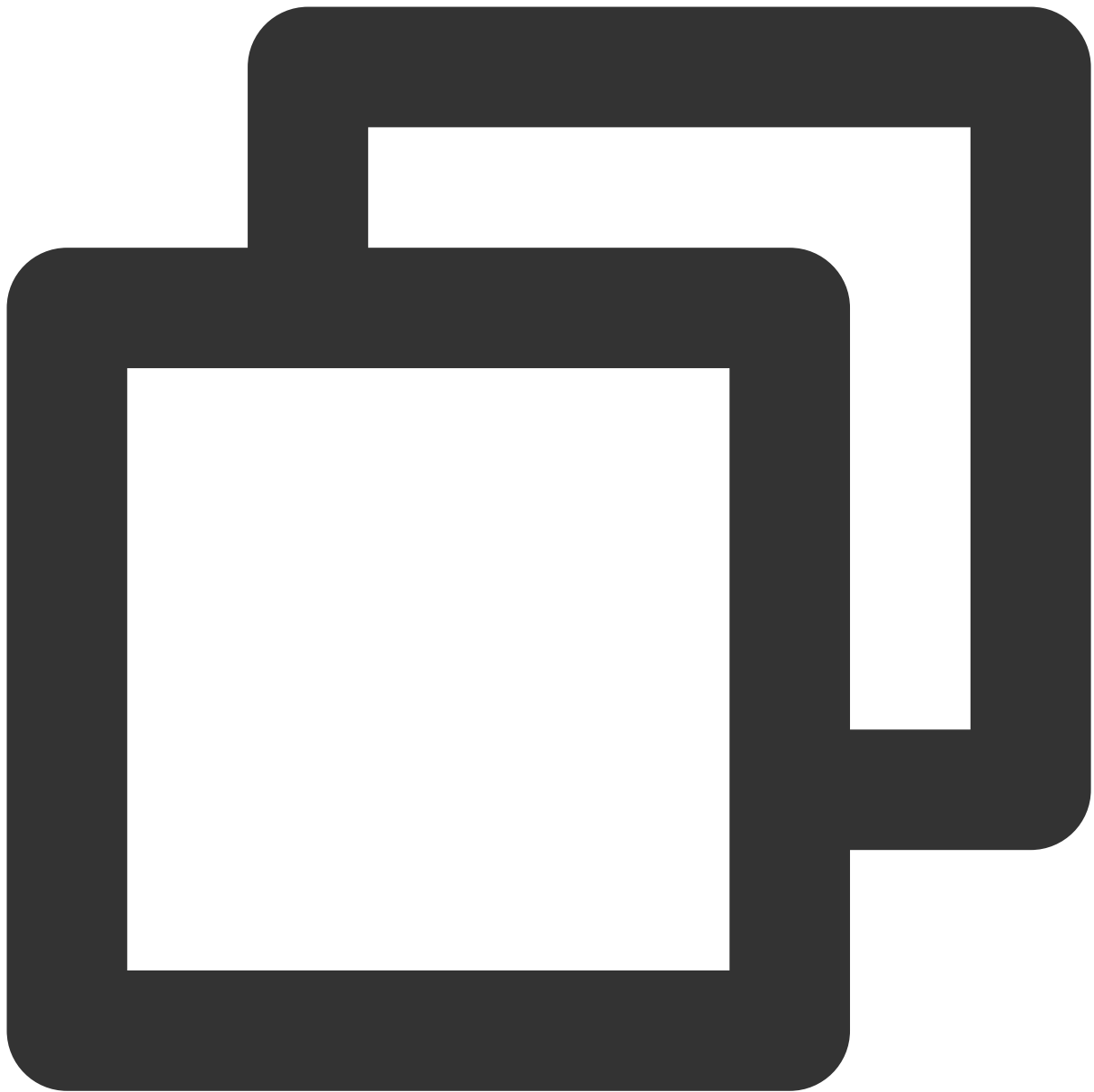
1. Create a data table and specify the JSON format for parsing.



```
CREATE EXTERNAL TABLE `order_demo` (
```

```
`docid` string COMMENT 'from deserializer',
`user` struct < id :int,
username :string,
name :string,
shippingaddress :struct < address1 :string,
address2 :string,
city :string,
state :string > > COMMENT 'from deserializer',
`children` array < string >
) ROW FORMAT SERDE 'org.apache.hive.hcatalog.data.JsonSerDe' LOCATION 'cosn://dlc-b
```

2. Run a query statement to query the JSON data. Data Lake Compute supports `json_parse()`, `json_extract_scalar()`, and `json_extract()` parsing functions.



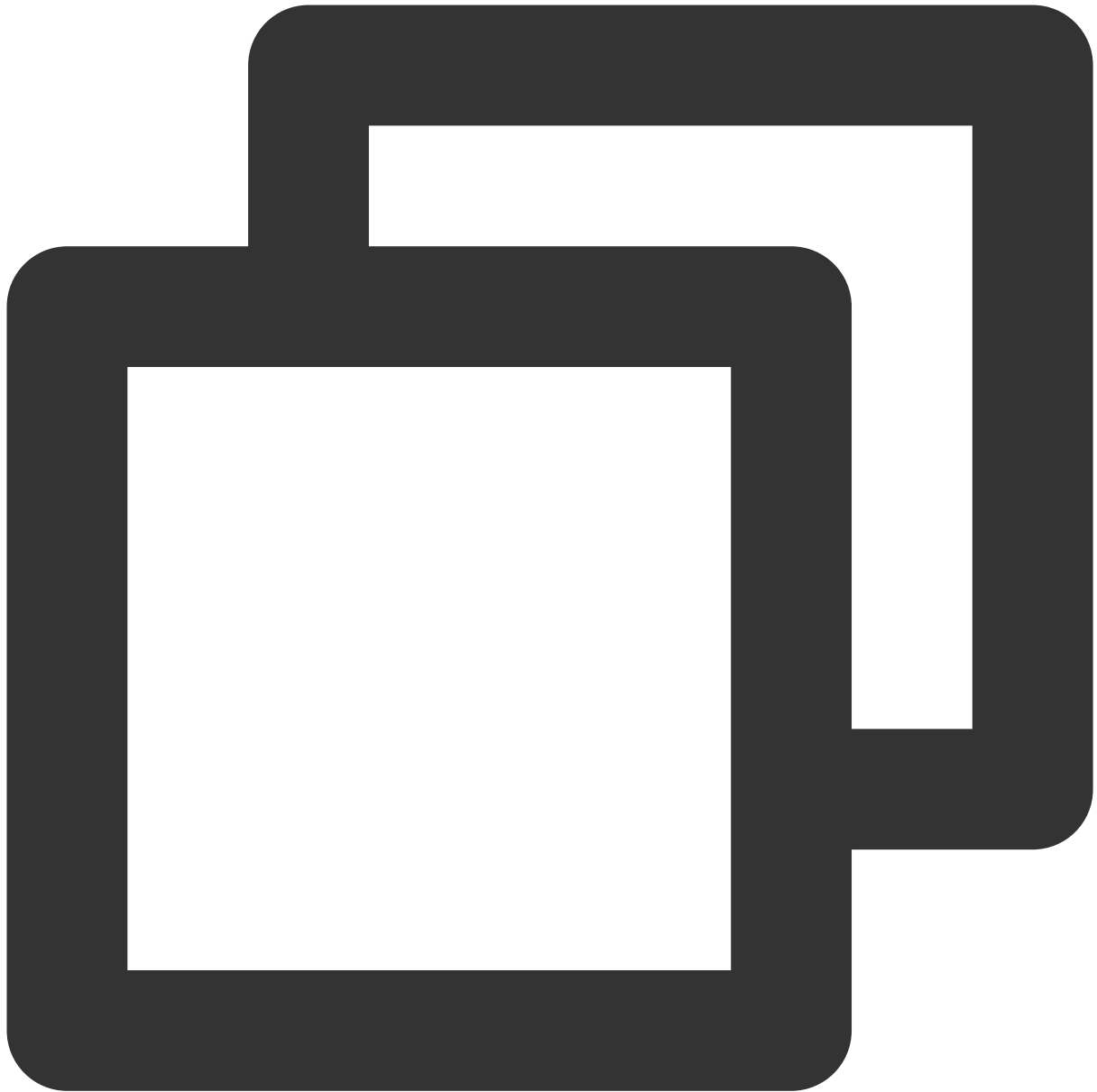
```
SELECT `user`.`shippingaddress`.`address1` FROM `order_demo` limit 10;
```

## System restraints

The data must be in complete JSON format; otherwise, Data Lake Compute cannot parse it.

A data row cannot contain a line break, and the JSON format cannot be optimized visually; for example:





```
{ "name": "Michael" }  
{ "name": "Andy", "age": 30 }  
{ "name": "Justin", "age": 19 }
```

Data Lake Compute will automatically recognize the first JSON level as the attribute column of a data table and recognize other nested structures as corresponding attribute values.

# Querying Data from Other Sources

Last updated : 2022-08-16 09:41:59

Data Lake Compute allows you to query and analyze data in an external table. Currently, data from MySQL and EMR Hive can be connected to it. You can add and manage other data sources in the Data Lake Compute console.

## Adding a data source

1. Log in to the [Data Lake Compute console](#) and select the service region. You need to have the permission to create data catalogs.
2. Select **Data Explore** on the left sidebar, hover over **+**, and click **Create data catalog**.
3. Select the data source type. Currently, MySQL and EMR Hive are supported. Before configuring MySQL, you need to add the Data Lake Compute subnet to the database's allowlist. Two configuration methods are supported: database instance and JDBC connection.
  - Supported EMR Hive versions are 2.0.1, 2.1.0, 2.2.0, 2.2.1, 2.3.0, 2.4.0, 2.5.0, 2.5.1, and 2.6.0. The configuration is performed through the EMR access address.
4. Enter the data source information and click **Create connection**.

## Managing Data

Currently, Data Lake Compute allows you to **view the database information of** and **preview data in** external tables.

### Viewing database information

1. Log in to the [Data Lake Compute console](#) and select the service region. You need to have the permission to view data tables.
2. Select **Data Explore** on the left sidebar, hover over **+**, and click **Basic info**. You can view the basic information of a data table in the pop-up window.

## Previewing data in a data table

1. Log in to the [Data Lake Compute console](#) and select the service region. You need to have the permission to view data tables.
2. Select **Data Explore** > **Data table**, hover over ..., and click **Preview data**. Then, you can run a SQL statement to query and display data in the data table.

# Using View

Last updated : 2024-07-17 16:22:09

In Data Lake Compute, a view is a logical table rather than a physical table. Whenever a view is referenced during a query, the query that defines the view will be executed. You can create a view through `SELECT` and reference it in future queries.

## System restraints

- A view name is case-insensitive and can contain up to 128 letters and underscores.
- Data Lake Compute doesn't support managing data access permissions through views.

# INSERT INTO

Last updated : 2024-07-17 16:23:11

The `INSERT INTO` statement can insert a `SELECT` query result in the source table to the target table as a new row.

# Querying Script Parameters

Last updated : 2024-07-17 16:23:47

Data Lake Compute allows you to configure date parameters to facilitate queries with scripts.

Data Lake Compute adopts the standard date format of `yyyymmddhh24miss` and uses the `${}` command to set a date as a variable consisting of the date and time.

**Date:** It can be in any date format or a predefined system variable, such as `yyyymmdd`, `yyyymm`, `yyyy-mm-dd`, `yy`, and `dataDate`.

**Time:** It can be +/-N cycles and supports `N/Nd`, `Nm`, `Nw`, `Nh`, and `Nmi`. It is compatible with various calculation formulas, such as `7*N` and `N/24`.

## Examples

+/- N Cycle	Method	Compatible Format	Example
N years later	<code>\${yyyymmdd+Ny}</code>	-	-
N years ago	<code>\${yyyymmdd-Ny}</code>	-	One year ago: <code>\${yyyymmdd-12m}</code> : 20190920
N months later	-	<code>\${yyyymmdd+Nm}</code>	-
N months ago	<code>\${yyyymmdd-Nm}</code>	<code>\$(add_months(yyyymmdd,-N))</code>	<code>\${yyyymmdd-1m}</code> : 20200820 <code>\${yyyymm}</code> : 202009 <code>\$(dataDate-1m)</code> : 20200820
N weeks later	<code>\${yyyymmdd+Nw}</code>	<code>\${yyyymmdd+7*N}</code>	-
N weeks ago	<code>\${yyyymmdd-Nw}</code>	<code>\${yyyymmdd-7*N}</code>	-
N days later	<code>\${yyyymmdd+N/Nd}</code>	-	-
N days ago	<code>\${yyyymmdd-N/Nd}</code>	-	<code>\${yyyymmdd-1}</code> , <code>\$(dataDate-1)</code>
N hours later	<code>\${yyyymmddhh24+Nh}</code>	<code>\${yyyymmddhh24+N/24}</code>	-
N hours ago	<code>\${yyyymmddhh24-Nh}</code>	<code>\${yyyymmddhh24-N/24}</code>	<code>\${yyyymmddhh24-1h}</code> : 2020092014

			<code>\${dataDate-1h}</code> : 2020092014
N minutes later	<code>\${yyyyymmddhh24mi+Nmi}</code>	<code>[\$[yyyyymmddhh24+N/24/60]</code>	-
N minutes ago	<code>\${yyyyymmddhh24mi-Nmi}</code>	<code>[\$[yyyyymmddhh24-N/24/60]</code>	<code>\${yyyyymmddhh24mi-10mi}</code> , <code>\${dataDate-10mi}</code>

**Note:**

Make sure that the variable or the part before `+/-` in the variable is in line with the standard date format; otherwise, the system cannot recognize and use it.

# Query Script Analysis

Last updated : 2024-08-07 17:08:48

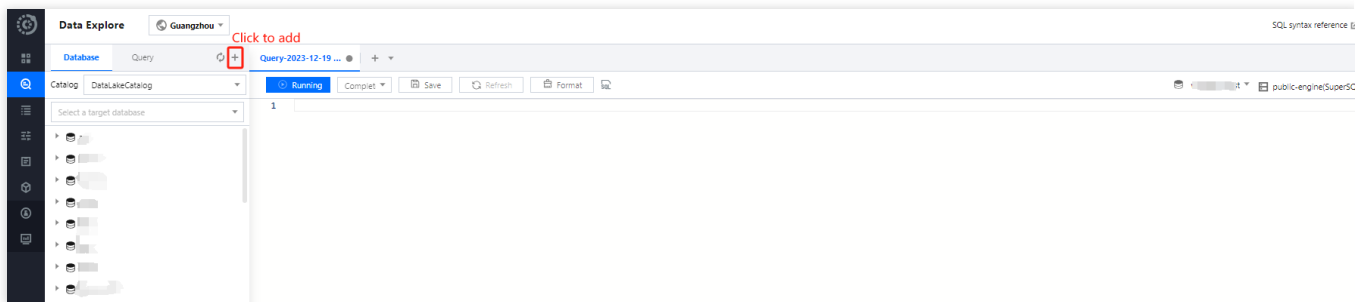
To facilitate users in quickly handling repetitive query tasks, DLC provides script file analysis.

## Note

The console allows saving up to 100 SQL scripts.

## Creating a New Query Directory

1. Log in to [DLC Console > Script Query Page](#).
2. On the query page, click Add Query Directory.



3. After filling in the directory configuration, you can save and complete the creation.



### Add query catalog ✕

Basic info

Catalog name

Permission settings An admin has all permissions by default and is not subject to the settings here

Available to

Work group

Add permissions for existing users in the work group and those to join later

User

Add permissions for individual users

**Confirm**

Directory name: Supports Chinese characters, letters, and underscores (\_), up to 25 characters.

Permission settings: You can set the visibility permissions for the script directory and the scripts within it based on the perspective of the workgroup or user.

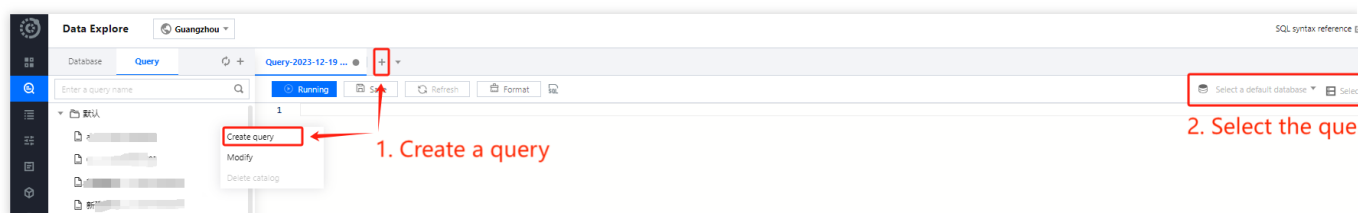
## Creating a New Query Script

1. Log in to [DLC Console > Script Query Page](#), You can click the library



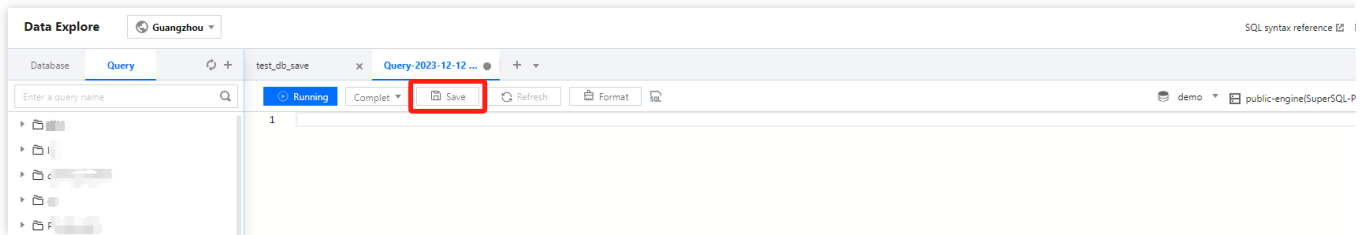
icon or directly add execution and save.

2. After the computation engine is selected, click Run to execute the script.



## Saving a Query Script

1. After the query is completed, click the Save button.
2. Queries created through the library will be saved under the directory of that library. Queries added through the tab bar can be saved directly in the root directory or an authorized library.



3. Query table permissions can be customized according to the public scope of the library, and table usage permissions can be specified for the public scope.

### Save query

Basic info

Query name:

Query catalog:

If you change the catalog, authorizations will be updated accordingly.

Permission settings: An admin has all permissions by default and is not subject to the settings here

Available to Work

group: Add permissions for existing users in the work group and those to join later

User:

Add permissions for individual users

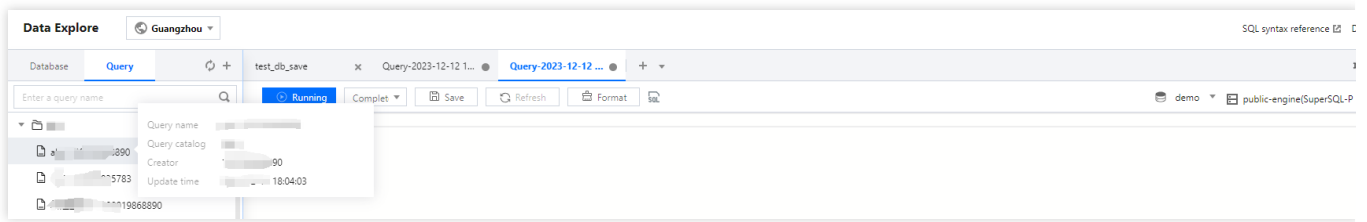
Permissions:  All  Read  Edit  Delete

Select permissions

[+Add](#)

## Viewing script information

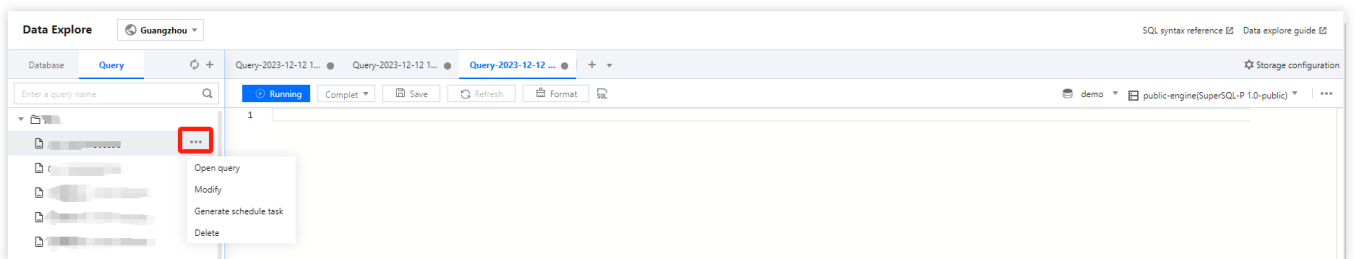
1. Hover the mouse pointer over the script name to view the script details.



2. Click the



icon next to the table you want to view, and select to open or query it.

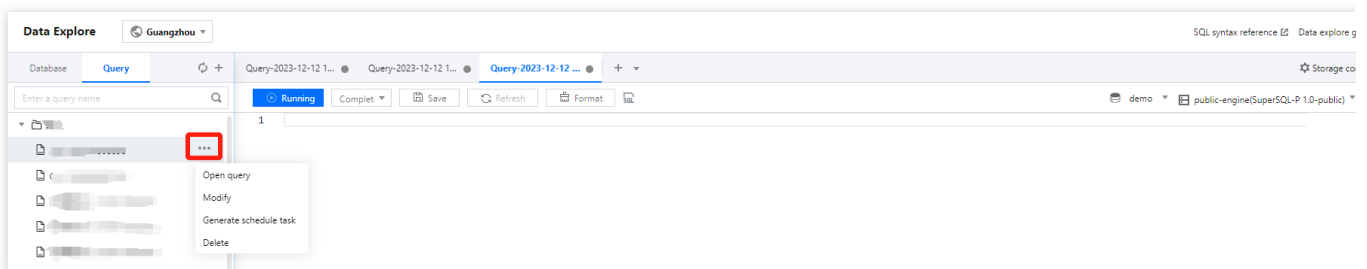


## Deleting a Query Script

Click the



icon next to the table you want to delete, and select to delete the script.



### Note:

Deleted scripts cannot be restored. Operate with caution.

# Data Management

## Data Catalogs and DMC

Last updated : 2024-07-31 17:27:26

External data and managed storage data in DLC can be managed through the Data Management Page by executing standard SQL statements and APIs. Through the Console Data Management Page, you can create, edit data catalogs, and create, query, delete databases and tables.

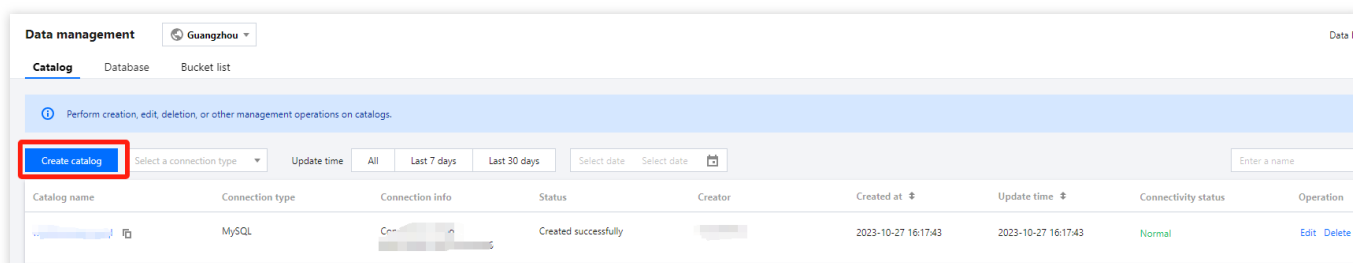
## Creating a data catalog

### Note:

The platform will automatically create a DataLakeCatalog for you for data management on the lake.

When you have external data sources and wish to perform federated analysis, you can follow the process below to create a data catalog for external data sources.

1. Log in to [DLC console](#), select the service region. The account used to log in must have the permission to create a catalog. For enabling sub-account permissions, refer to [Sub-account Permission Management](#).
2. Enter [Data Management](#), click **Create Catalog**.



3. Enter the data source creation visual interface. After filling in the connection information, complete the network configuration to connect the engine with the external data source.

**Create catalog**

1 Catalog configuration > 2 Network configuration

Connection type \* MySQL

Connection name \* te...

Description Up to 50 characters

Instance \* cd...75

Data source VPC \* vpc-7... subnet-... 253 IPs in total, 245 available

Username \* t...

Password \* ...

Back Next

**Create catalog**

1 Catalog configuration > 2 Network configuration

Use the bound data engine to query and analyze data from this data source. You can change the scope of the bound data engine via [Network configuration](#).

Data source VPC vpc-73vy8arx subnet-c93fr1js

253 IPs in total, 245 available  
You can configure a network for a data engine to access data sources over it. Enhanced network configuration offers faster data transmission and thus is suitable for accessing a large volume of data. Cross-source network configuration allows you to set several networks for one data engine for cross-source federated data query across several networks.

Network configuration type \* Enhanced Cross-source

Network configuration name \* It can contain up to 25 characters in letters, digits, and underscores (\_).

Available data engines \* Select a data engine  
Only the selected data engine can read data under this catalog. Only Presto private data engines are available for this selected catalog.

Configuration description Up to 50 characters

Back Confirm

4. After filling in the data source information, click **Confirm** to complete the creation of the data source.
5. In the Data Catalog List, view connection information, status, creator, and other information.

## Edit Data Catalog

1. Click **Data Catalog List > Operations > Edit** to modify the Data Catalog's description information, network configuration information, username, password, and running cluster, etc.

**Edit catalog**

Connection type \* MySQL

Connection name \* v-...

Description Up to 50 characters

JDBC \* jdbc:mysql

Example: jdbc:mysql://ip:port; database name is not required.

Data source VPC \* vpc-73... subnet-... 253 IPs in total, 245 available

Username \*

Password \*

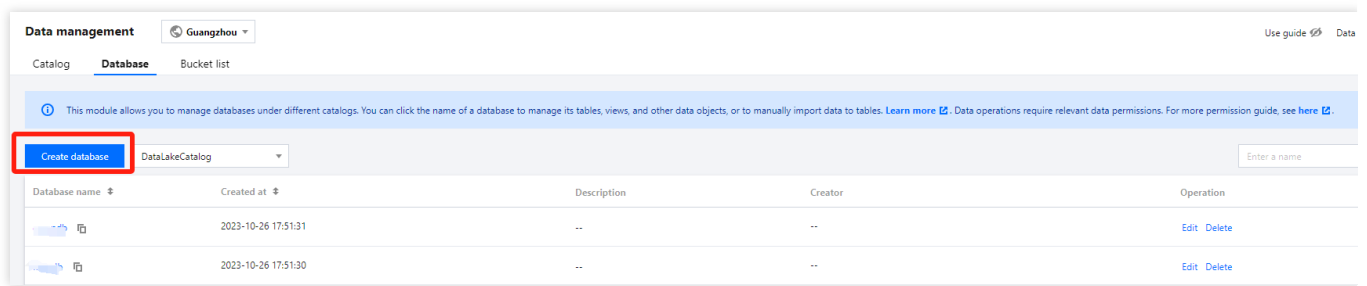
Use the bound data engine to query and analyze data from this data source. You can change the scope of the bound data engine via [Network configuration](#).  
Data engines bound to the data source VPC: --

Confirm Cancel

2. After modifications, click **Create** to reconstruct the Data Catalog.

## New database

1. Log in to [DLC Console](#), select the service region. The account used to log in must have database creation permissions.
2. Enter [Data Management](#), click on the directory name under the Data Catalog to view the databases within that directory.
3. Click **Create Data Catalog** to enter the Database Creation Visual Interface.



4. After filling in the relevant database information and saving, the database creation is complete. When creating a database, you can [enable data optimization](#) for the entire database.

The 'Create database' dialog box contains the following fields:

- Database name \***: A text input field with the placeholder 'Enter a database name'.
- Description**: A text input field with the placeholder 'Optional'.
- Data governance**: A toggle switch currently turned off.

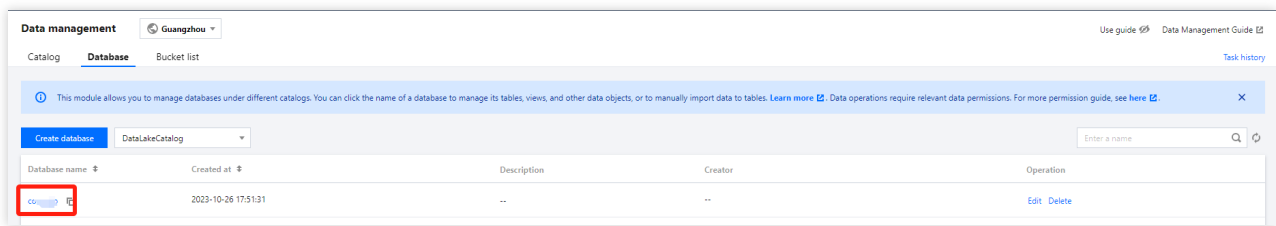
Database Name: Globally unique, supports English case-sensitive letters, numbers, "\_", cannot start with a number, up to 128 characters.

Description: Supports both Chinese and English, up to 2,048 characters.

A root account can create up to 100 databases.

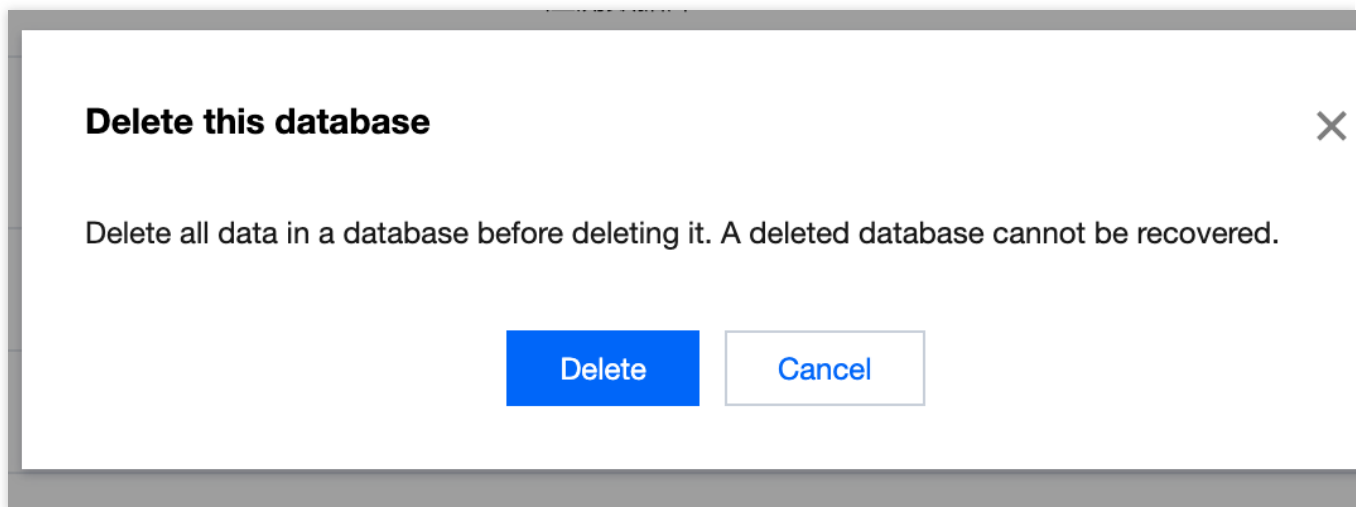
## View Database

1. Log in to [DLC Console](#), select the service region. The account used to log in must have database query permissions.
2. Enter [Data Management](#) > **Database**, select the data directory, click **Database Name** to access the database details, manage the database's tables. For a detailed operation guide, refer to [Data Table Management](#).



## Dropping a Database

1. Log in to [DLC Console](#), select the service region. The account used to log in must have database deletion permissions.
2. Enter [Data Management](#), click **Delete**. After confirming a second time, the database can be deleted.





# Data Table Management

Last updated : 2024-07-31 17:27:51

Users can use the DLC console or API to execute DDL statements to create a database.

## Creating Table

### Approach one: Create in Data Exploration

1. Log in to the [DLC console](#), select the service region, log in to users need to have the permission to create tables.
2. Enter the [Data Exploration](#) module, in the left list, click on an existing database, hover over the table row, then click the

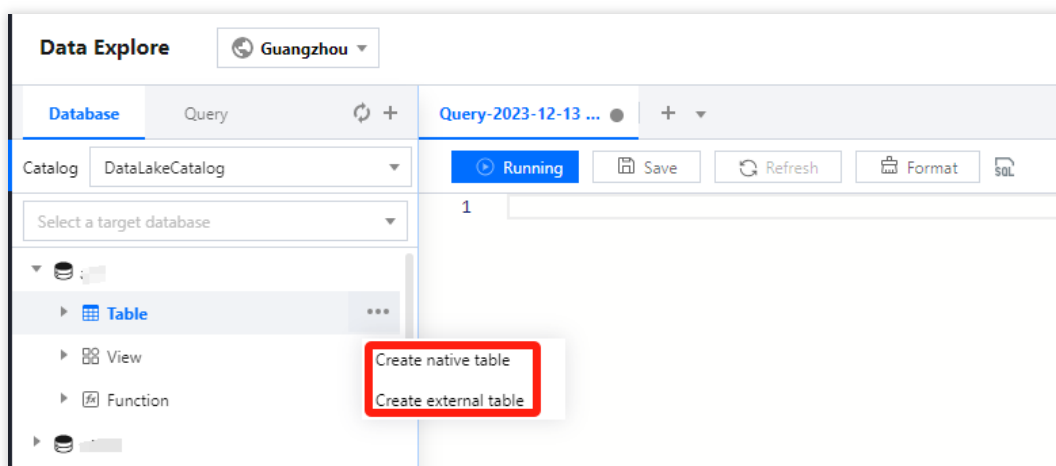


icon, click **Create Native Table** or **Create External Table**.

#### Note:

A native table refers to a table on the DLC managed storage. With a native table, you don't need to worry about the underlying Iceberg storage format, and it has capabilities like data optimization. To use a native table, you need to enable managed storage first, see [Managed Storage Configuration](#) for details.

The underlying data of the external table resides on your own COS. Creating an external table requires specifying the data path.



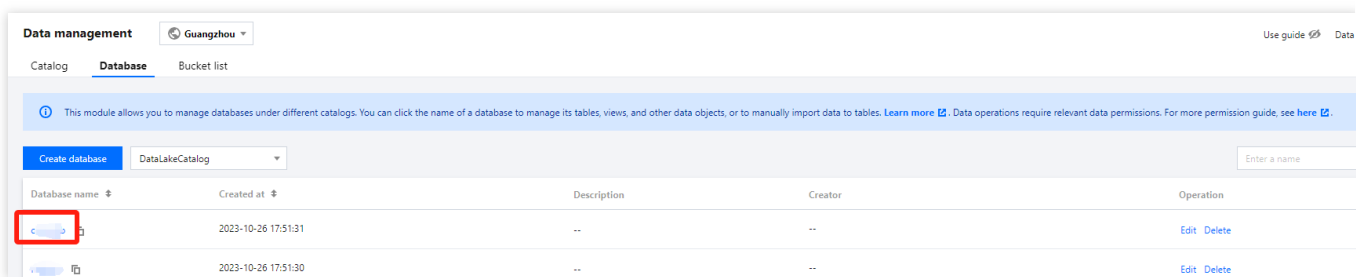
3. After clicking **Create Native Table/Create External Table**, the system will automatically generate an SQL template for creating a data table. Users can modify the SQL template to create a data table. After clicking **Run**, the SQL statement to create the data table is executed, completing the creation.



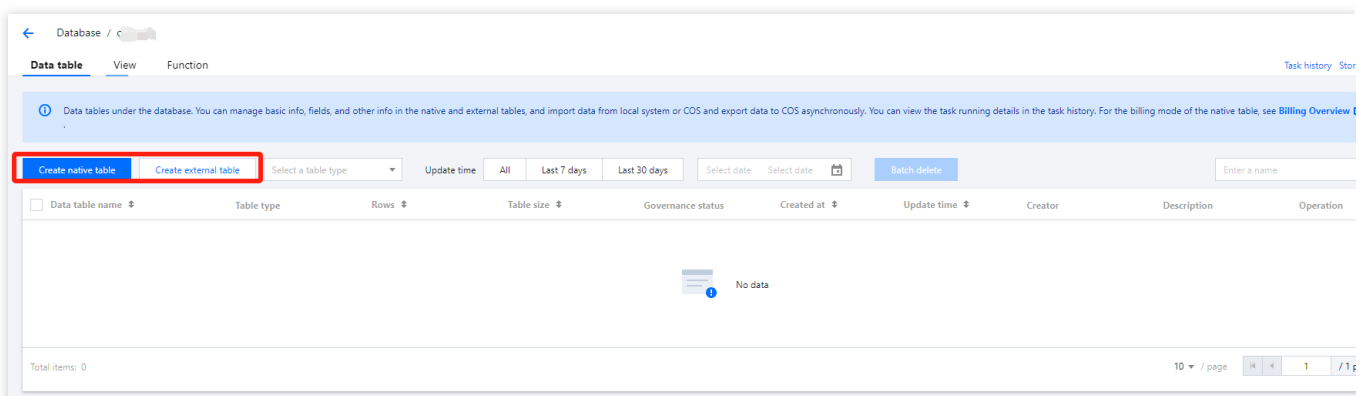
## Approach two: Create in Data Management

The Data Management module supports managing native tables and external tables stored in DLC.

1. Log in to the [DLC console](#), select the service region, log in to users need to have the permission to create tables.
2. Through the left menu, enter **Data Management**, enter **Database**, click the name of the database where the data table is located, enter the DMC page.



3. Click **Create Native Table** or **Create External Table** button to enter the data table configuration page.



Native table data sources support three different types: empty table, local upload, and COS COS. Choosing different data sources corresponds to different creation processes. Native tables support capabilities such as data optimization and can choose to inherit database governance rules or individually turn them on/off.

### 3.1 Create Empty Table: Create an empty table with no records.

Data Table Name: Cannot start with a number, supports uppercase and lowercase letters, numbers, and underscores —, with a maximum of 128 characters.

Support for entering data table description information.

Manually add and enter column names and field types. Supports the configuration of three complex type fields: array/map/struct.

**Create native table**
✕

Data table source Blank table

Create a table for specific data and import data, or directly create a blank table.

Data table name Enter a data table name

Data table version Select

Iceberg table version. v1: Analytic data tables; v2: Supports row-level updates and deletes.

Description Optional

Field name	Field type	Field configuration	Description	Operation
No data				

Add

Partitioning

Inherit database governance rules  Yes  No

The current data table inherits the governance rules of the database as follows:

Data governance

Attributes ▶

Confirm
Cancel
Show SQL

3.2 Local Upload: Upload local form files to DLC to create data tables, supports files up to 100MB.

CSV: Supports visual configuration of CSV parsing rules, including Compression Format, Column Splitting Symbol, Field Domain Symbol. Supports automatic inference of the data file's Schema and parsing the first row as Column Names.

Json: DLC only recognizes the first level of Json as columns, supports automatic inference of the Json file's Schema. The system will recognize the first level fields of Json as Column Names.

Supports common Big Data Format files like Parquet, ORC, AVRO, etc.

Manually add and enter Column Names and Field Types.

If the Automatic Structure Inference is selected, DLC will automatically fill in the detected columns, Column Names, and Field Types. If incorrect, please manually modify.

**Create native table**
✕

Data table source Upload ▾

Create a table for specific data and import data, or directly create a blank table.

Data path \* Select file

You can upload a file of up to 100 MB. For files larger than 100 MB, please use the COS mode or import them with API or other tools.

Data format Select a data format ▾

Data table name Enter a data table name

Data table version Select ▾

Iceberg table version. v1: Analytic data tables; v2: Supports row-level updates and deletes.

Description Optional

Field info Infer structure

Automatically infer the data structure based on the selected file. Please confirm the data structure info, or manually modify the data structure.

Field name	Field type	Field configuration	Description	Operation
No data				

Add

Partitioning  On  Off

Inherit database  Yes  No

Confirm
Cancel
Show SQL

### 3.3 Create a data table through COS COS.

Create a data table by reading the COS data buckets under the current account.

CSV: Supports visual configuration of CSV parsing rules, including Compression Format, Column Splitting Symbol, Field Domain Symbol. Supports automatic inference of the data file's Schema and parsing the first row as Column Names.

Json: DLC only recognizes the first level of Json as columns, supports automatic inference of the Json file's Schema.

The system will recognize the first level fields of Json as Column Names.

Supports common Big Data Format files like Parquet, ORC, AVRO, etc.

Manually add and enter Column Names and Field Types.

If the Automatic Structure Inference is selected, DLC will automatically fill in the detected columns, Column Names, and Field Types. If incorrect, please manually modify.

**Create native table**
✕

Data table source COS

Create a table for specific data and import data, or directly create a blank table.

Data path \* Select a data path [Select a COS path](#)

You can upload a file of up to 100 MB. For files larger than 100 MB, please use the COS mode or import them with API or other tools.

Data format Select a data format

Data table name Enter a data table name

Data table version Select

Iceberg table version. v1: Analytic data tables; v2: Supports row-level updates and deletes.

Description Optional

Field info Infer structure

Automatically infer the data structure based on the selected file. Please confirm the data structure info, or manually modify the data structure.

Field name	Field type	Field configuration	Description	Operation
No data				

Add

Partitioning

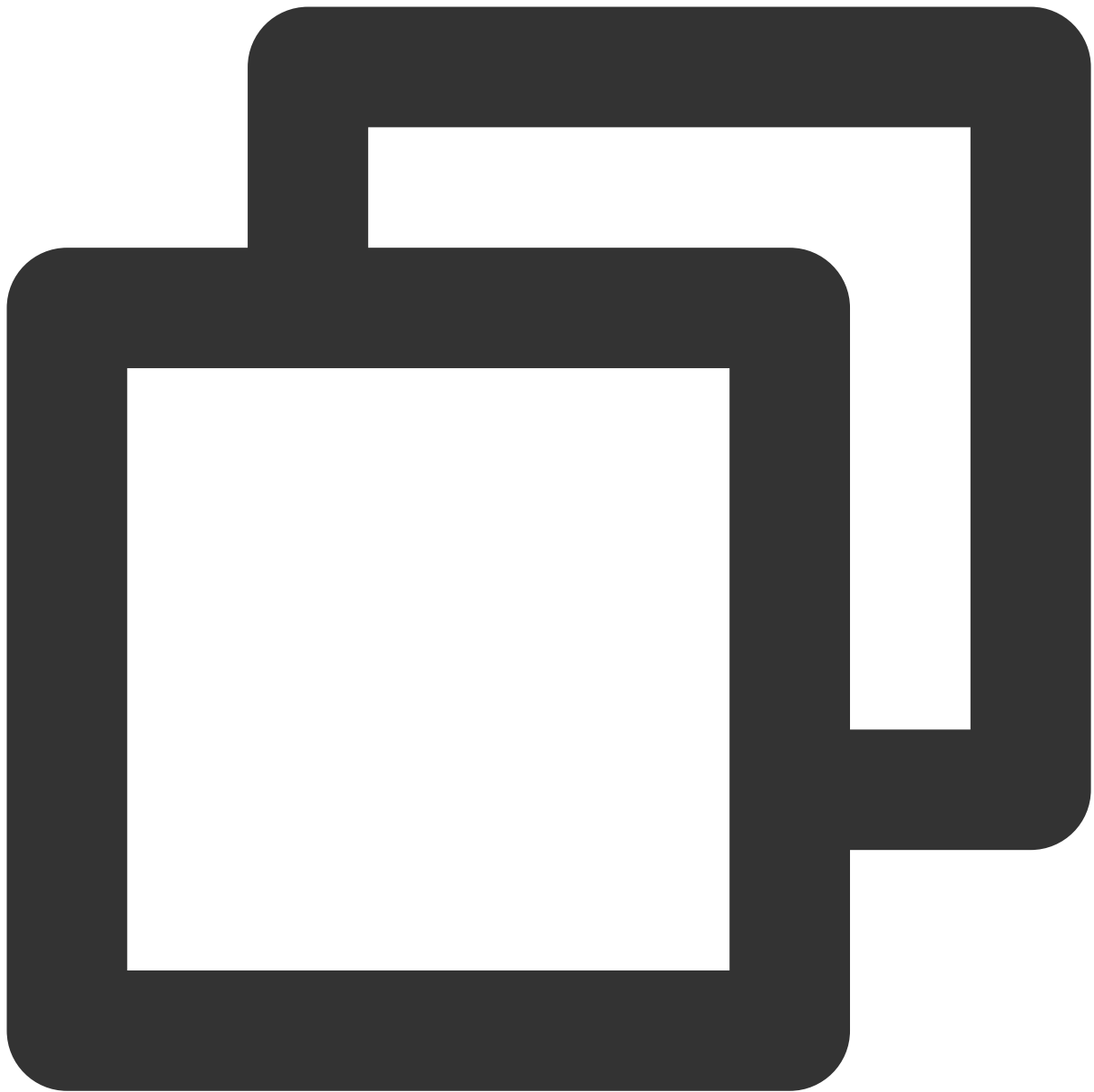
Inherit  Yes  No

Confirm
Cancel
Show SQL

4. Data Partitioning is often used to enhance Query Performance and is applied to large volume tables. DLC supports data querying by Data Partitioning. Users need to add partition information at this step. By partitioning your data, you can limit the amount of data scanned with each query, thereby improving Query Performance and reducing usage costs. DLC adheres to Apache Hive's partitioning rules.

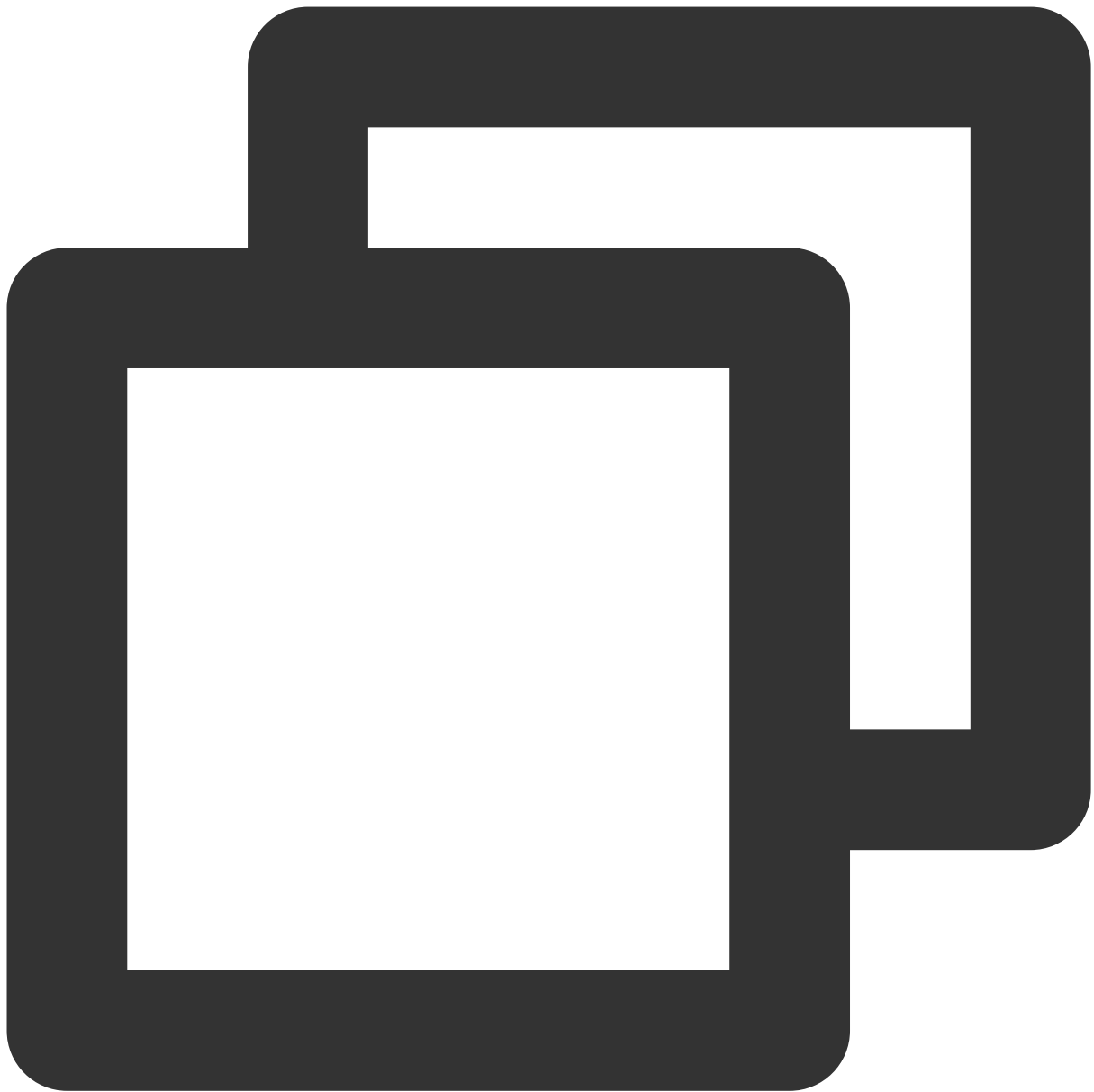
The partition column corresponds to a subdirectory under the COS path of the table, with the directory naming convention being **Partition Column Name=Partition Column Value**.

Example:



```
cosn://nanjin-bucket/CSV/year=2021/month=10/day=10/demo1.csv  
cosn://nanjin-bucket/CSV/year=2021/month=10/day=11/demo2.csv
```

If there are multiple partition columns, they need to be nested in the order specified in the create table statement.



```
CREATE EXTERNAL TABLE IF NOT EXISTS `COSDataCatalog`.`dlc_demo`.`table_demo` (  
  `_c0` string,  
  `_c1` string,  
  `_c2` string,  
  `_c3` string  
) PARTITIONED BY (`year` string, `month` string, `day` string)  
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'  
WITH SERDEPROPERTIES ('separatorChar' = ',', 'quoteChar' = '"')  
STORED AS TEXTFILE  
LOCATION 'cosn://bucket_name/folder_name/';
```

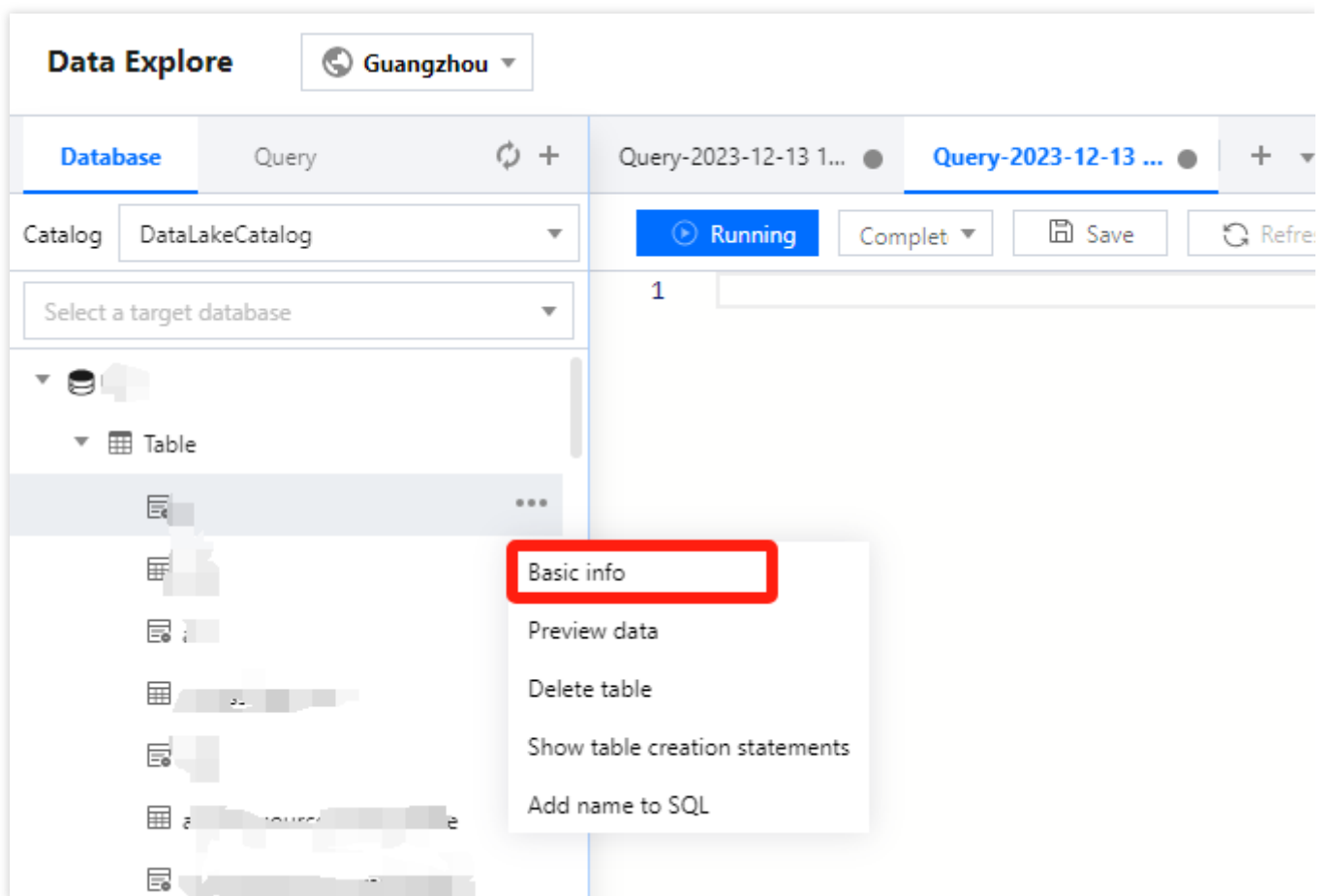
## Query basic information of the data table

### Approach one: Query in Data Exploration

In the Data Table Item, mouse hover over the **Data Table Name** row, then click the

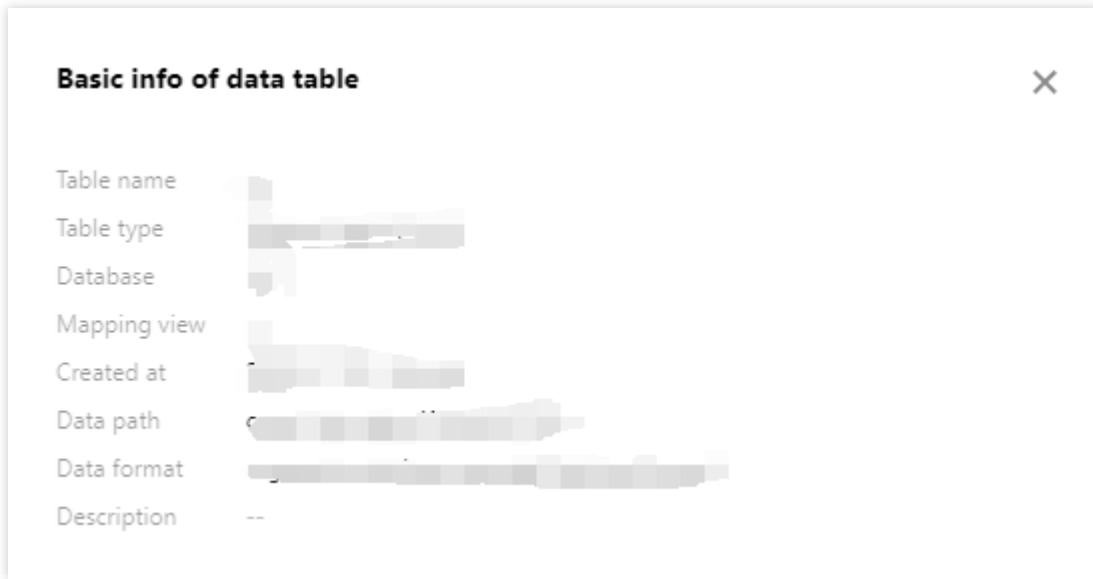


icon, in the Dropdown Menu click **Basic info** to view the basic information of the created data table.



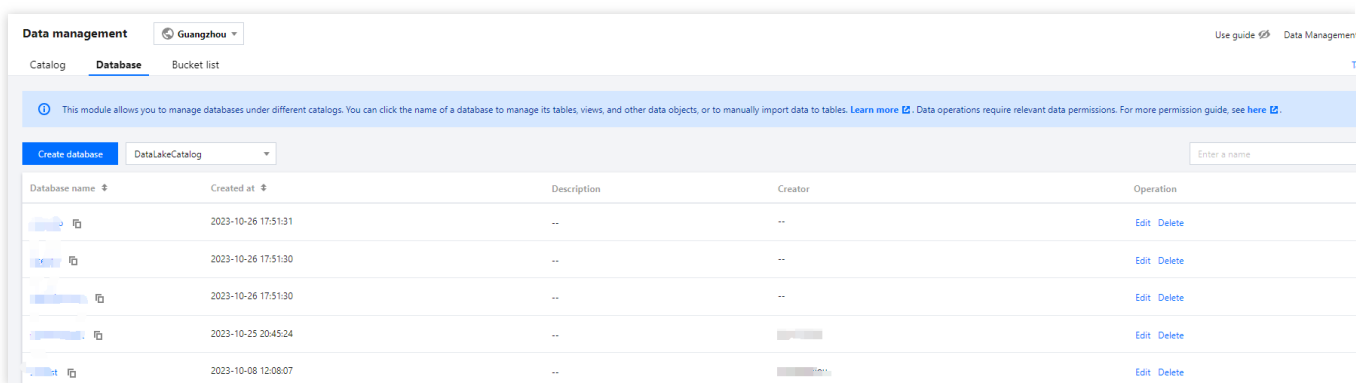
The basic information of the data table is as follows:

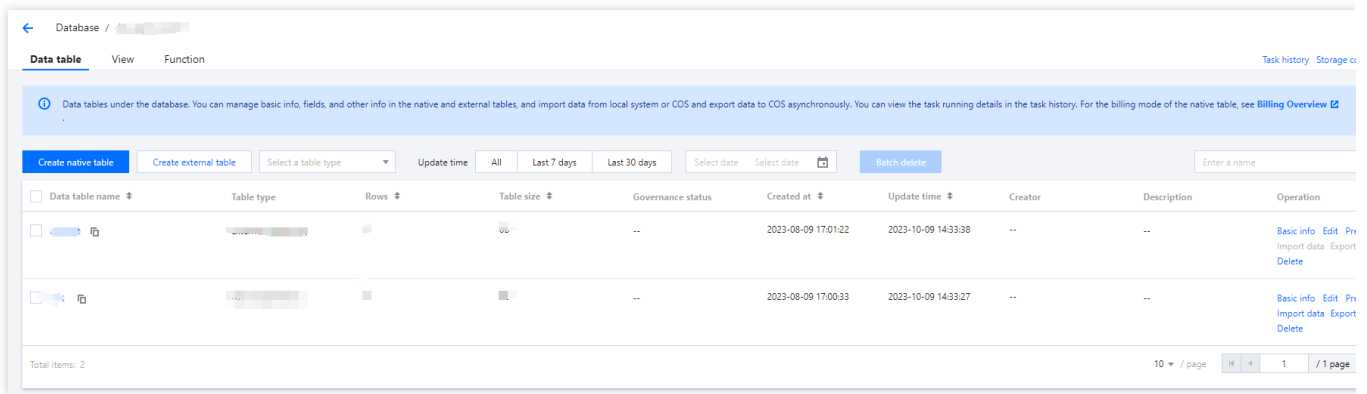




### Approach two: View in Data Management

1. Log in to the [DLC Console](#), select the service region. Users need to have the permission to view data tables.
2. Through the left menu, enter the **Data Management** page, click the name of the database where the data table is located, enter the DMC page. It supports querying information such as the number of rows, storage space, creator, fields, partitions, etc.



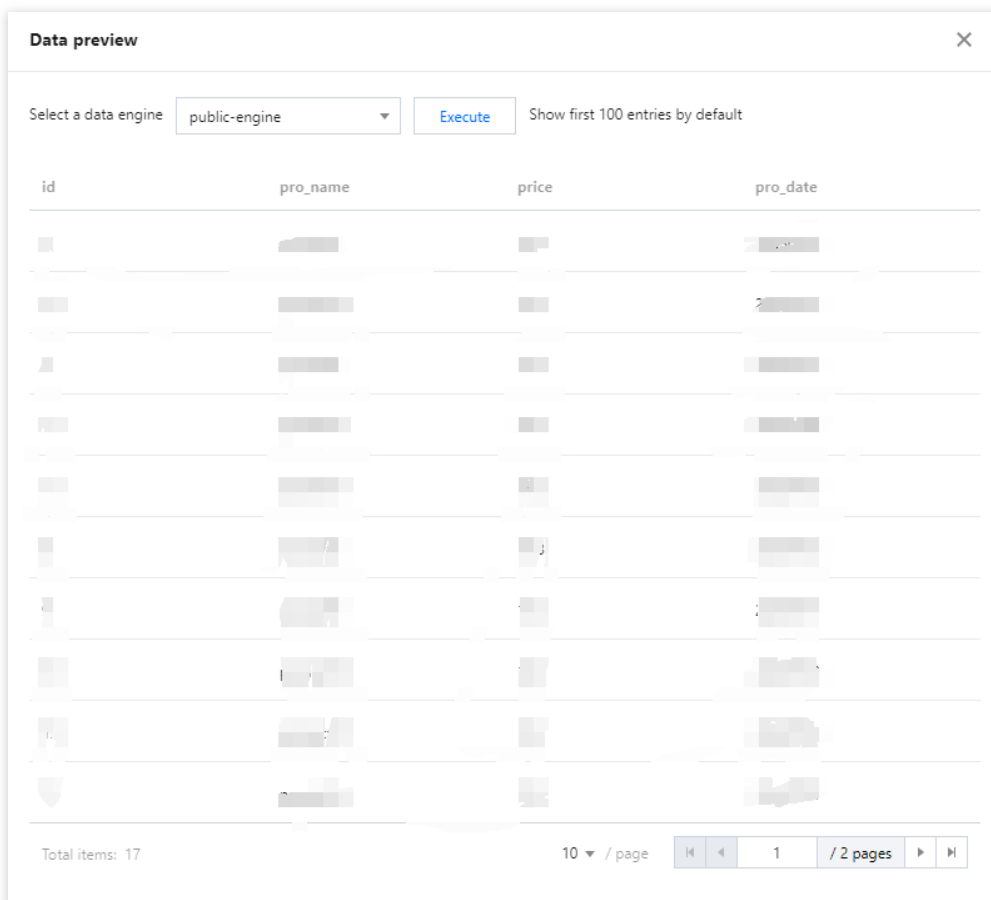


## Preview Data Table Data

In the Data Table Item, hover the mouse over the **Data Table Name** row, then click the



icon, in the Dropdown Menu click **Preview Data**. DLC will automatically generate a SQL Statements to preview the first 10 rows of data, executing the SQL Statements to query the top 10 rows of the data table.



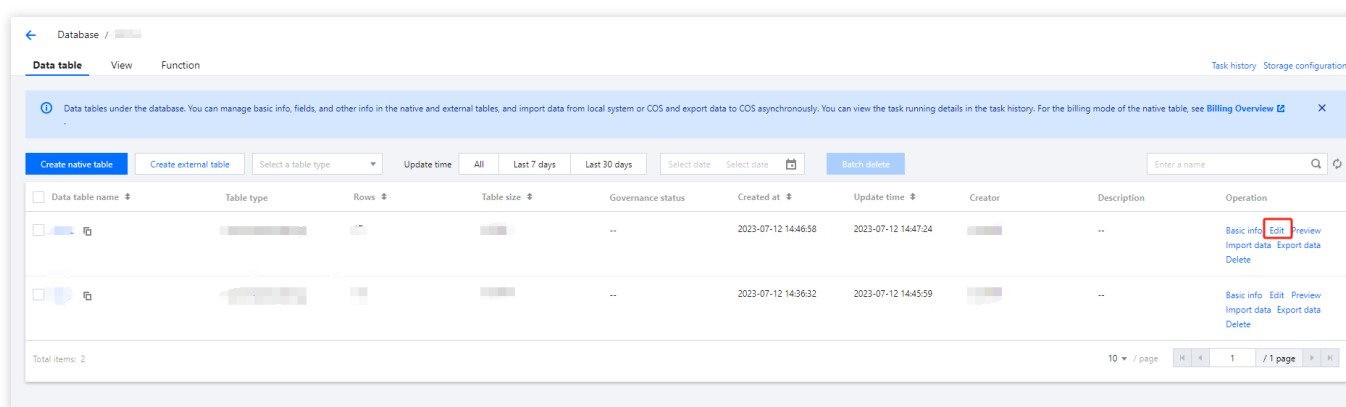
Support for previewing data in **Data Management > Database > Data Table > Data Table List**.

The Data Preview Function by default displays the first 100 rows of data.

## Editing Data Table Information

Support editing the Description information of the data table in the Data Management module.

1. Log in to the [DLC Console](#), select the Service Region. Users need to have the permission to edit data tables.
2. Through the left menu, enter the **Data Management > Database** page, click the name of the database where the data table is located, enter the DMC page.
3. Find the data you need to edit, click the **Edit** button on the right to edit.



4. After modification, click the **Confirm** button to complete the editing.

**Edit data table**

Data table name: [Input field]

Data table version: V1  
Iceberg table version. v1: Analytic data tables; v2: Supports row-level updates and deletes.

Upsert:

Created at: 2023-07-12 14:46:58

Update time: 2023-07-12 14:47:24

Description: [Input field]

Inherit database governance rules:  Yes  No  
The current data table inherits the governance rules of the database as follows:

Data governance:

Confirm Cancel

## Dropping a Table

### Approach one: Delete in Data Exploration

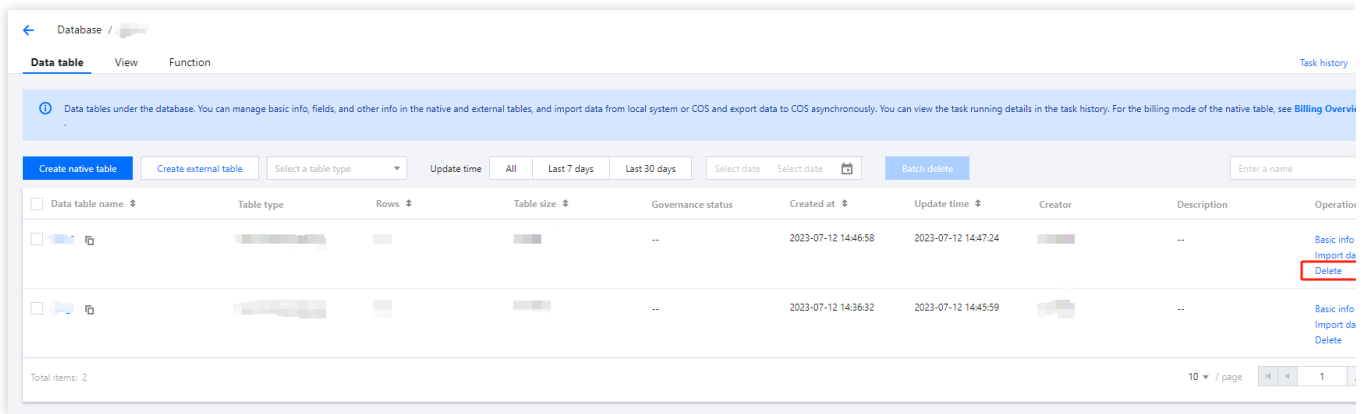
In the Data Table Items, hover the mouse over the **Data Table Name** row, then click the



icon, in the dropdown menu click **Delete**. DLC will automatically generate the SQL statement to drop the data table, execute the SQL statement to drop the table.

Dropping an external table, dropping a data table only removes the metadata stored in DLC, it does not affect the data source file.

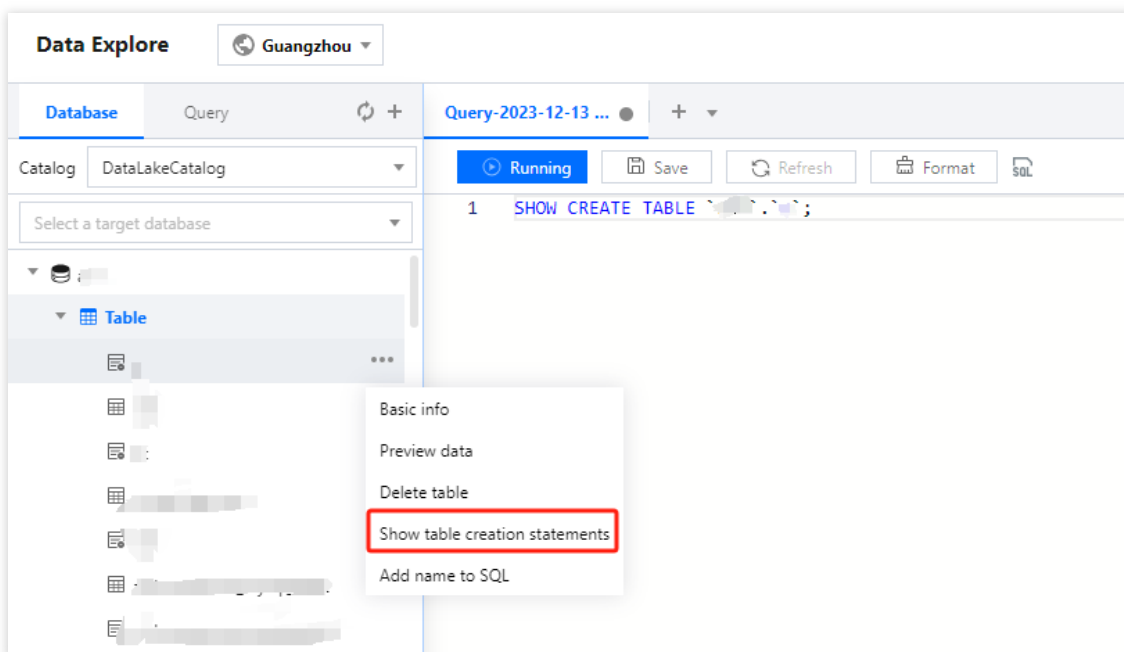
Deleting tables under the DataLakerCatalog directory will clear all data of that table, proceed with caution.



## Approach two: Delete in Data Management

Currently, Data Management only supports the management of databases and tables hosted in DLC. For external tables, please use approach one for deletion.

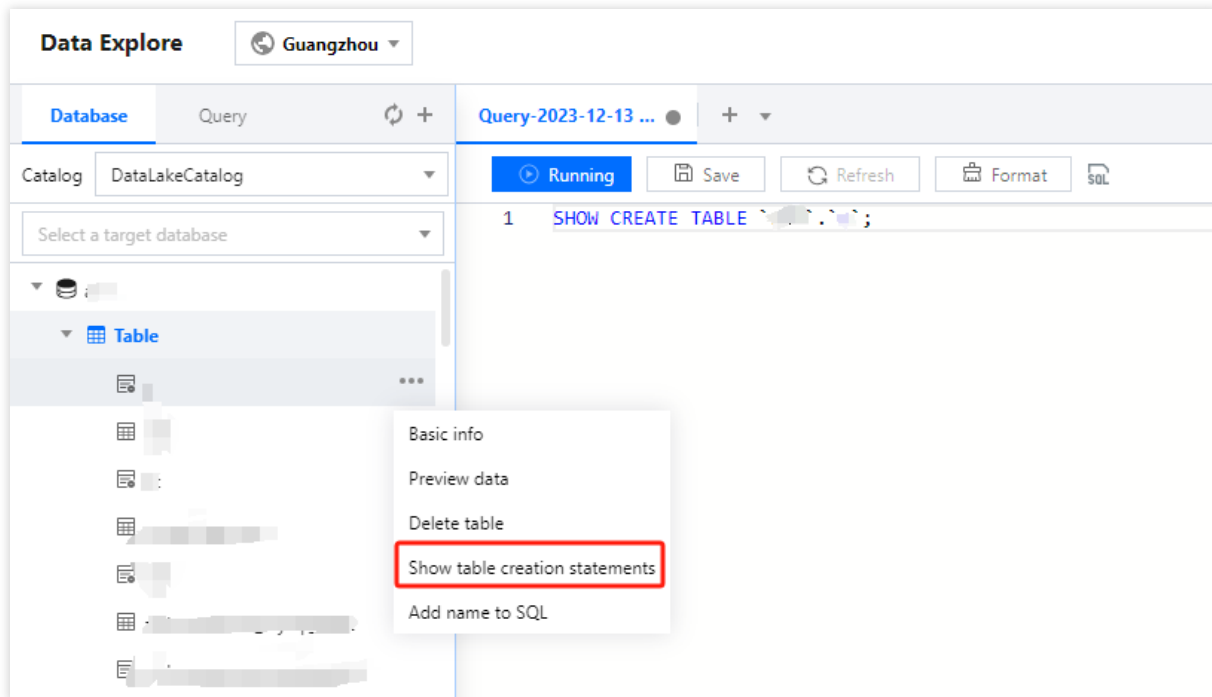
1. log in to the [DLC Console](#), select the service region, users need to have the permission to delete data tables.
2. Through the left menu, enter **Data Management > Database**, click the name of the database where the data table is located, to enter the DMC page.
3. Click the **Delete** button after the data table you wish to delete, after confirmation, the corresponding data table can be deleted and its data cleared.



## Show create table statement

In the Data Table Item, hover the mouse over the **Data Table Name** row, then click the

icon, in the dropdown menu click **Show table creation statements**. DLC will automatically generate the SQL statement to view the create table statement for that data table, execute the SQL statement to query the create table statement.



## System constraints

DLC allows up to 4096 data tables under each database, supports a maximum of 100,000 partitions per data table, and the maximum number of attribute columns per data table is 4096.

DLC will recognize data files under the same COS path as data from the same table, please ensure data for separate tables is kept in separate folder hierarchies.

DLC does not support multi-version data in COS; it can only query the latest version of data in a COS bucket.

All tables created on DLC are external tables, and the SQL statement to create the table must include the EXTERNAL keyword.

Table names must be unique within the same database.

Table names are case-insensitive and only support letters, numbers, and underscores (\_), with a maximum length of 128 characters.

If the table is a partitioned table, you must manually execute the ADD PARTITION statement or the MSCK statement to add partition information before you can query the partition data. For more details, see [Query partitioned table](#).

When creating a table with CSV, DLC will by default convert all field types to string, but this does not affect the computation and querying of raw data fields.

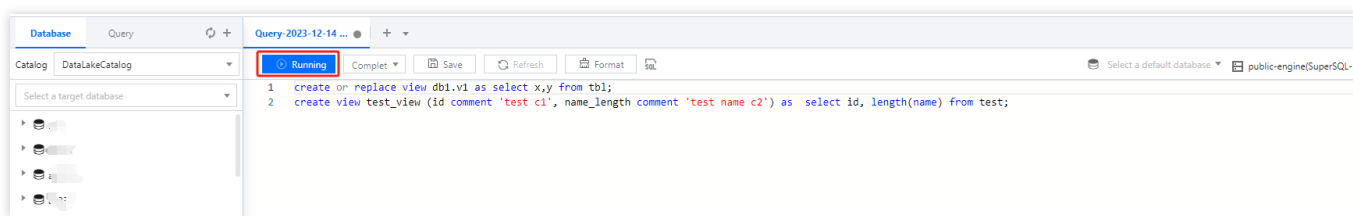
# Data View Management

Last updated : 2024-07-31 17:28:41

DLC provides data view query capabilities, allowing users to quickly and easily perform data queries and use through the management of data views.

## Create View

1. log in to [DLC console](#), select the service region, log in users must have the permission to create views.
2. Enter the **Data Exploration page**, you can create views using SQL statements. For details of the statement, see [SQL Syntax](#).
3. Select the computing resource, click the **Running** button to complete view creation.

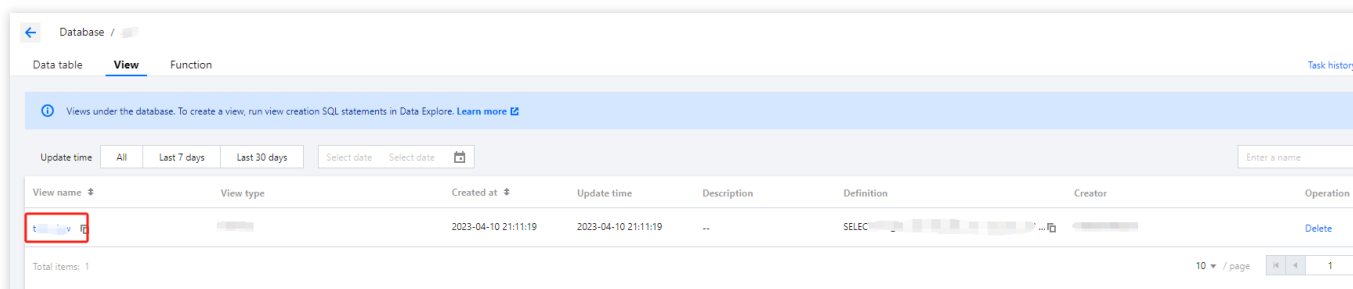


## View Views

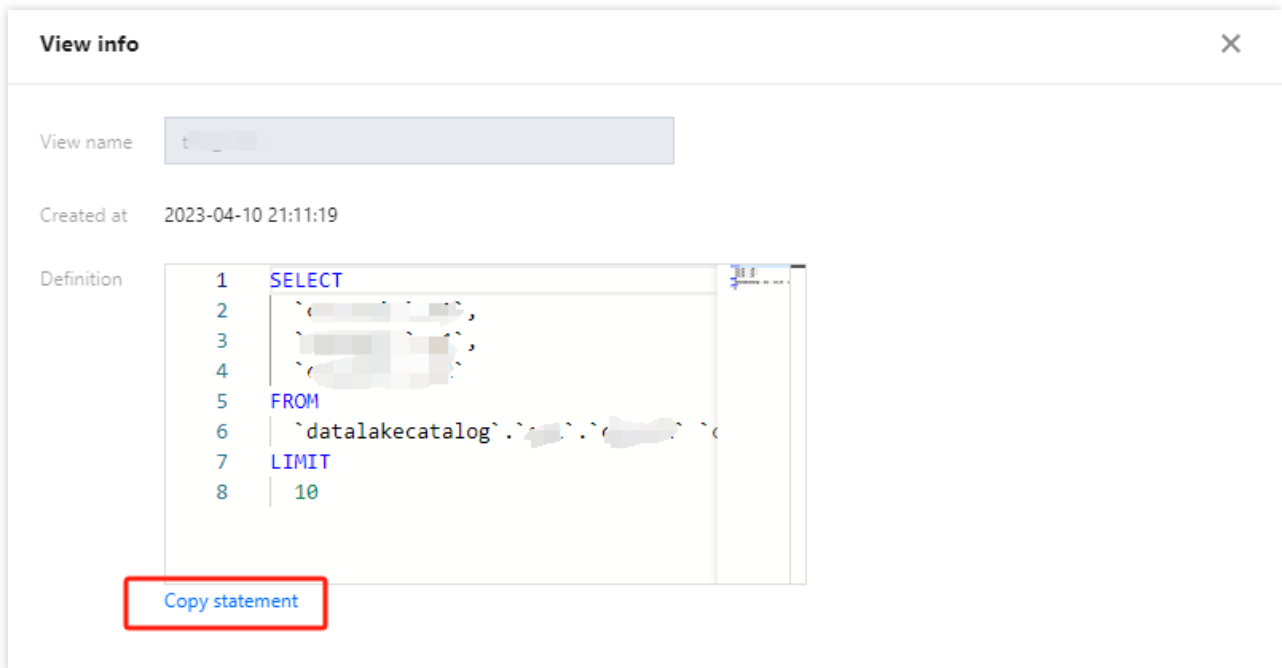
You can view the view using SQL statements through Data Exploration, see [SQL Syntax](#) for specific syntax.

Meanwhile, DLC also offers a Visual Interface for managing views, with the following operations.

1. Log in to the [DLC console](#), select the service region, log in users must have the permission to query views.
2. Enter the Data Management page, click on the **Database Name** where the view is located to enter the DMC page.
3. Click **View** to enter View Management.



4. Click the **View Name** you want to inspect to view its information. You can copy the SQL statement.

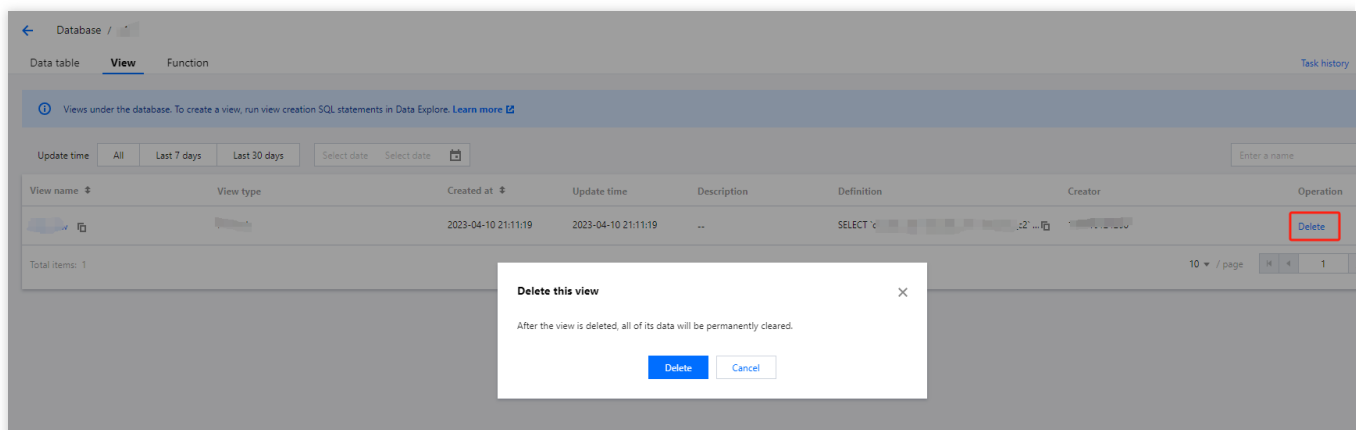


## Delete View

You can view the view using SQL statements through Data Exploration, see [SQL Syntax](#) for specific syntax.

Meanwhile, DLC also offers a Visual Interface for managing views, with the following operations.

1. Log in to [DLC Console](#), select the service region, users must have view deletion permissions.
2. Enter the Data Management page, click on the **Database Name** where the view is located to enter the DMC page.
3. Click **View** to enter View Management, then click the **Delete** button to delete the view.



### Caution

Deleting a view will clear all data under the view and cannot be recovered. Please proceed with caution.



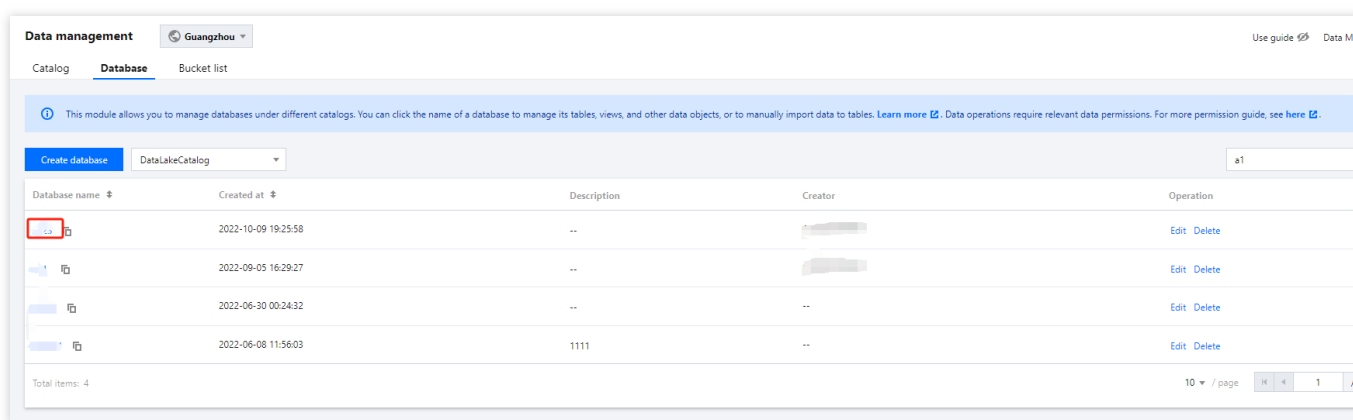
# Function Management

Last updated : 2024-07-31 17:28:59

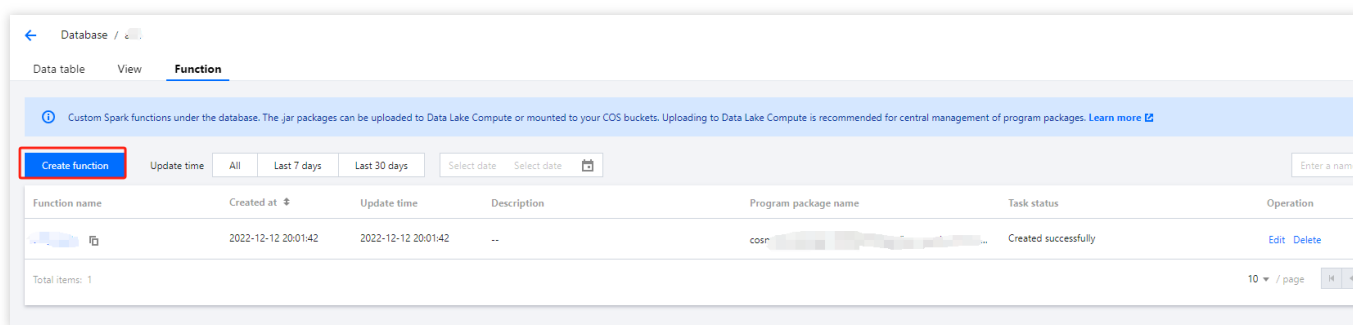
In DLC, you can use **User-Defined Functions** to process and construct data, and it supports function management.

## Creating function

1. Log in to the [DLC Console](#), select the service region. The account must have database operation permissions.
2. Go to the **Data Management Page**, click on the **database name** for which you need to build the function.



3. Select **Function**, then click on the **Create Function** button to enter the function creation menu.



### Create function

Function name \*

Description

Storage mode  Save on system  Mount on a specified COS path  
The storage mode of the function package. You can upload and save the function package to the system (recommended), or directly save it at a specified COS path.

Program package source  Upload  COS

File path \*   
Only a .jar package of up to 5 MB is supported

Function class name \*

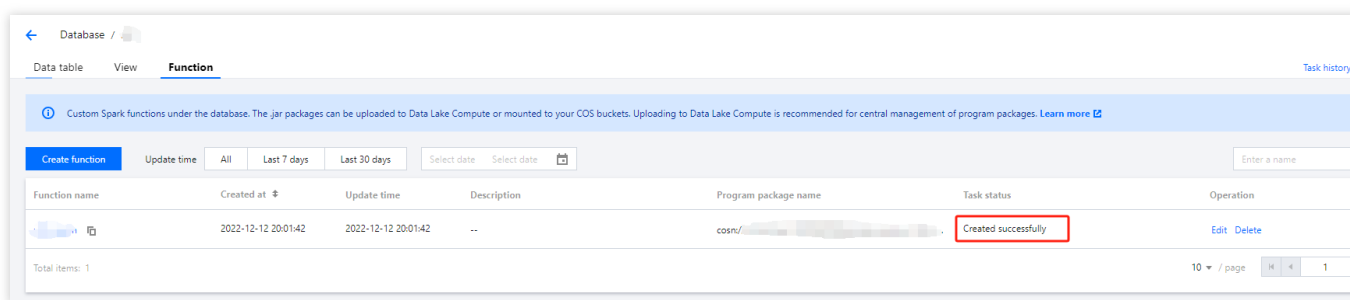
The function package can be uploaded locally or use an existing JAR package on COS. Local uploads only support the JAR format, with a maximum size of 5MB.

You need to select a Spark cluster for execution, which will not incur any costs during the operation.

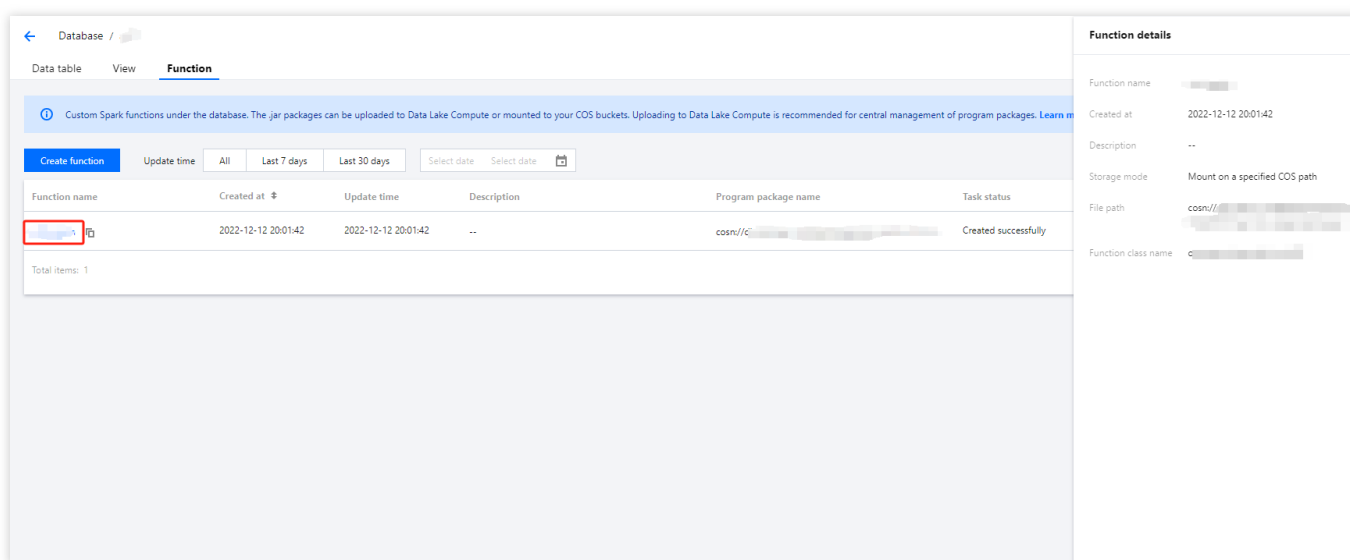
It is recommended to save the function package to the system for easy management and use. It also supports mounting to a specified COS path.

## View function information

1. Log in to the [DLC Console](#). The account must have database operation permissions.
2. Go to the **Data Management Page**, click on the **database name** of the function you want to view.
3. Select the function to view its Build Status. If the build fails, you can **edit** and resubmit.

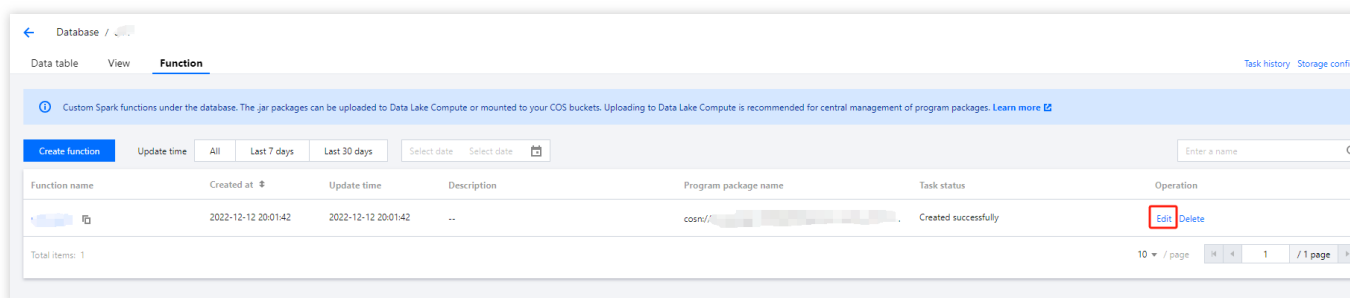


4. Click on the **Function Name** to directly view the function details.



## Editing Function Information

1. Log in to the [DLC Console](#), select the service region. The account must have database operation permissions.
2. Go to the **Data Management Page**, click on the **database name** of the function you want to view.
3. Select a **Function** and click the **Edit** button to enter the function information editing page.



### Edit function ✕

Function name \*

Created at 2022-12-12 20:01:42

Description

Storage mode  Save on system  Mount on a specified COS path  
The storage mode of the function package. You can upload and save the function package to the system (recommended), or directly save it at a specified COS path.

File path \*  [Select a COS path](#)  
Only a .jar package of up to 100 MB is supported.

Function class name \*

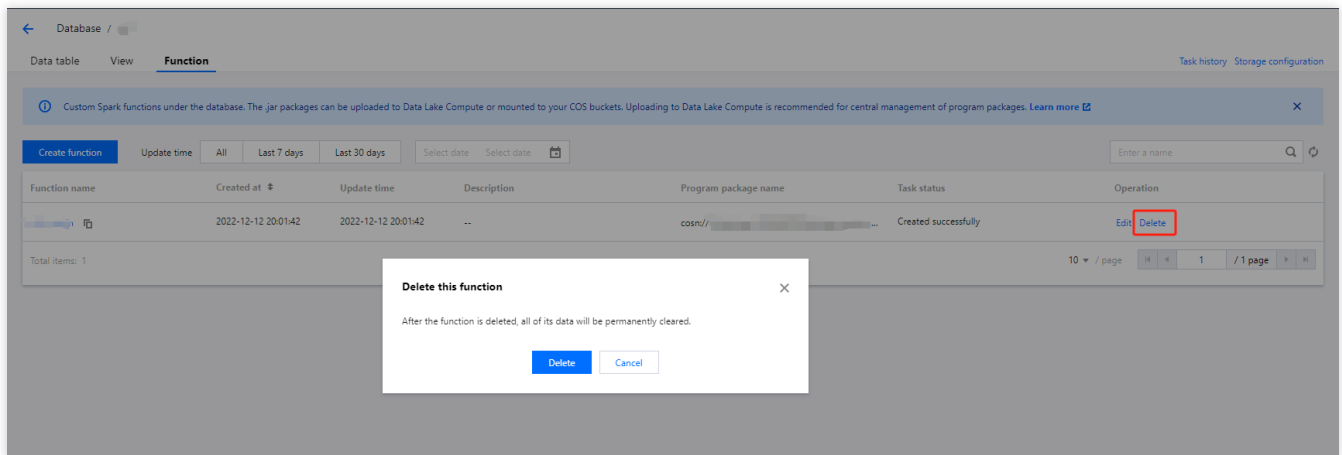
The function name, storage method, and upload method cannot be modified at this time. If you need to change this information, please recreate the function.

After modifying the function information, it will be rebuilt. Please proceed with caution.

## Deleting function

For functions that no longer need to be managed, you can delete them.

1. Go to the [DLC Console](#), select the service region, log in to an account with database operation permissions.
2. Go to the **Data Management Page**, click on the **database name** of the function you want to view.
3. Select the **function** and click the **delete** button to delete the function that is no longer needed.



## Caution

After deletion, the data under this function will be cleared and cannot be recovered. Please proceed with caution.

# Partition Field Policy

Last updated : 2024-07-31 17:29:14

In Hive, partition information appears in the form of directories. In Iceberg, partition information is recorded in the underlying data files, making Iceberg's partitions more flexible and allowing the partitioning strategy to evolve with changes in data volume. In DLC, you can create Iceberg tables to utilize features such as hidden partitions.

## Note:

By default, native tables are Iceberg tables. External tables, depending on the file format, can choose between Hive or Iceberg tables. For detailed syntax, refer to the document [CREATE TABLE](#).

With hidden partitions, when inserting and querying data, you do not need to specify partition information additionally as required in Hive.

Iceberg partition strategy supports the use of the following functions, with different fields and corresponding partition transformation strategies as shown in the table:

Partitioning Strategy	Field Type	Result Type
identity	Any	Source Type
bucket	int, long, decimal, date, time, timestamp, timestamptz, string, uuid, fixed, binary	int
truncate	int, long, decimal, string	Source Type
year	date, timestamp, timestamptz	int
month	date, timestamp, timestamptz	int
day	date, timestamp, timestamptz	date
hour	timestamp, timestamptz	int

# Data Job

## Overview

Last updated : 2024-07-17 16:36:54

Data Lake Compute provides Spark-based batch and flow computing capabilities for you to perform complex data processing and ETL operations through data jobs.

Currently, data jobs support the following versions:

Scala 2.12

Spark 3.1.2

## Preparations

Before starting a data job, you need to create a data access policy to ensure data security as instructed in [Configuring Data Access Policy](#).

Currently, only CKafka data source is supported for data job configuration, with more data sources to come in the future.

## Billing mode

A data job is billed by the data engine usage. Currently, pay-as-you-go and monthly subscription billing modes are supported. For more information, see [Data Engine Overview](#).

**Pay-as-you-go:** It is applicable to scenarios with a small number of data jobs or periodic usage. A data job is started after creation and automatically suspended after successful execution, after which no fees will be incurred.

**Monthly subscription:** It is applicable to scenarios where a large number of data jobs are regularly executed.

Resources are reserved in this mode, so you don't need to wait for data engine start.

### Note:

As a data job differs from a SQL job in terms of the compute engine type, you need to purchase a separate data engine for Spark jobs; otherwise, you can't run data jobs on a SparkSQL data engine.

## Job management

On the **Data job** management page, you can create, start, modify, and delete a data job.

1. Log in to the [Data Lake Compute console](#) and select **Data job** on the left sidebar.
2. Click **Create job**. For detailed directions, see [Creating Data Job](#).

3. In the list, you can view the current task status of the data job. You can also manage the job as instructed in [Managing Data Job](#).



# Configuring Data Access Policy

Last updated : 2024-07-17 17:44:52

## Data Access Policy (CAM role arn) Overview

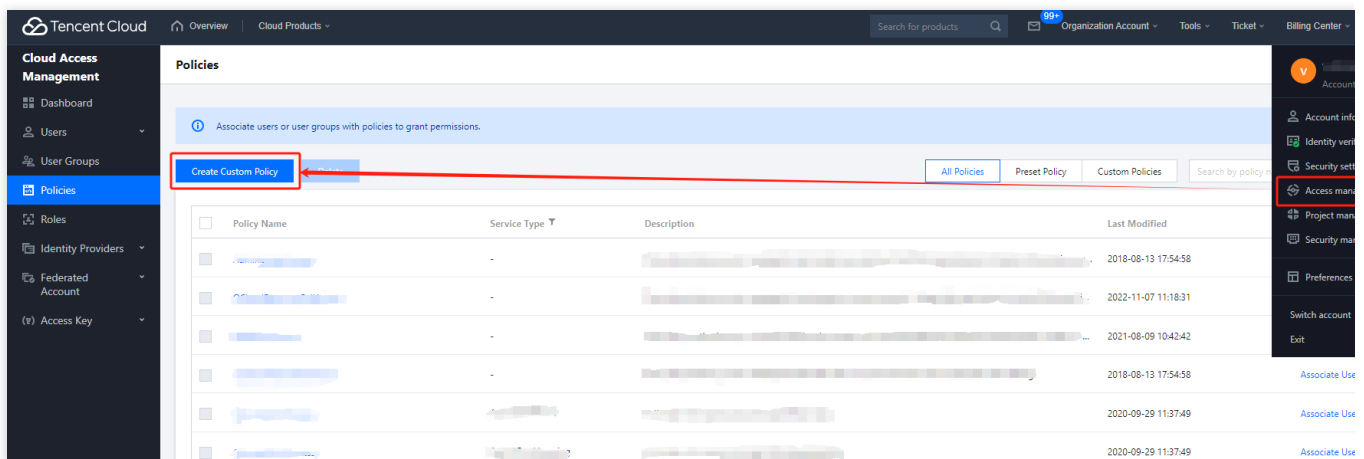
A data access policy (CAM role arn) allows you to configure permissions in CAM for accessing data in data sources and COS during data job execution.

When configuring a data job in Data Lake Compute, you need to specify the data access policy to protect data security.

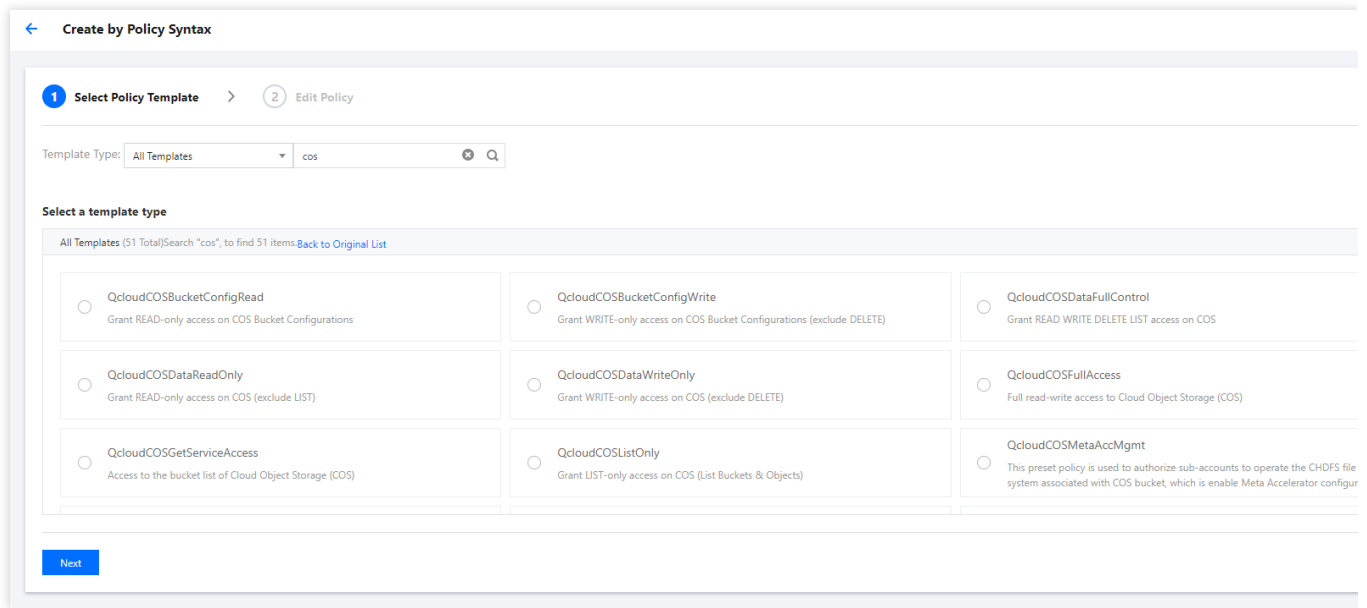
## Directions

### Step 1. Create a policy in CAM

1. Log in to the Tencent Cloud console and select **Cloud Access Management**. The logged-in account needs to have permissions to configure CAM; therefore, we recommend you use a root account or admin account.
2. Select **Policies** on the left sidebar to enter the policy management page. Click **Create Custom Policy** and select **Create by Policy Syntax**.



3. Search for COS in the policy template and select **COS permission templates**.

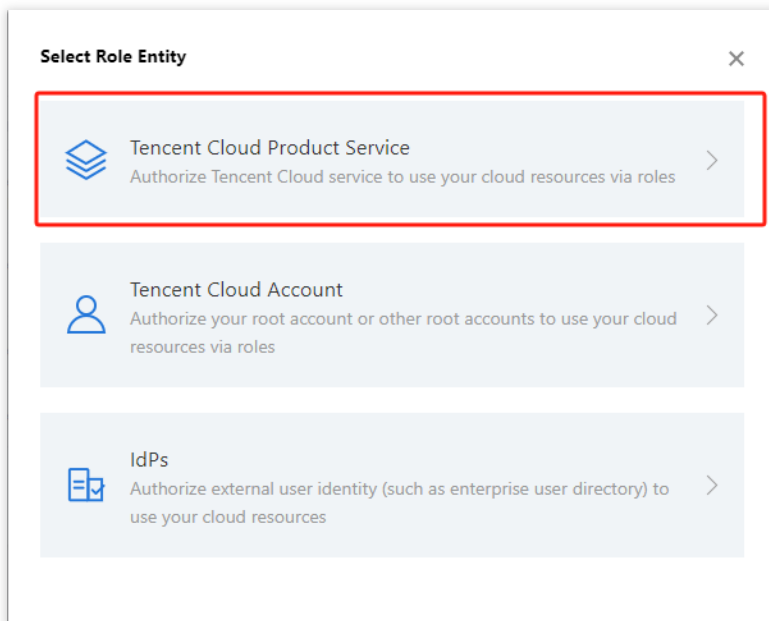


The preset templates define read-only and read/write permission policies. If they don't meet your needs, create a custom policy template as instructed in [Appendix](#).

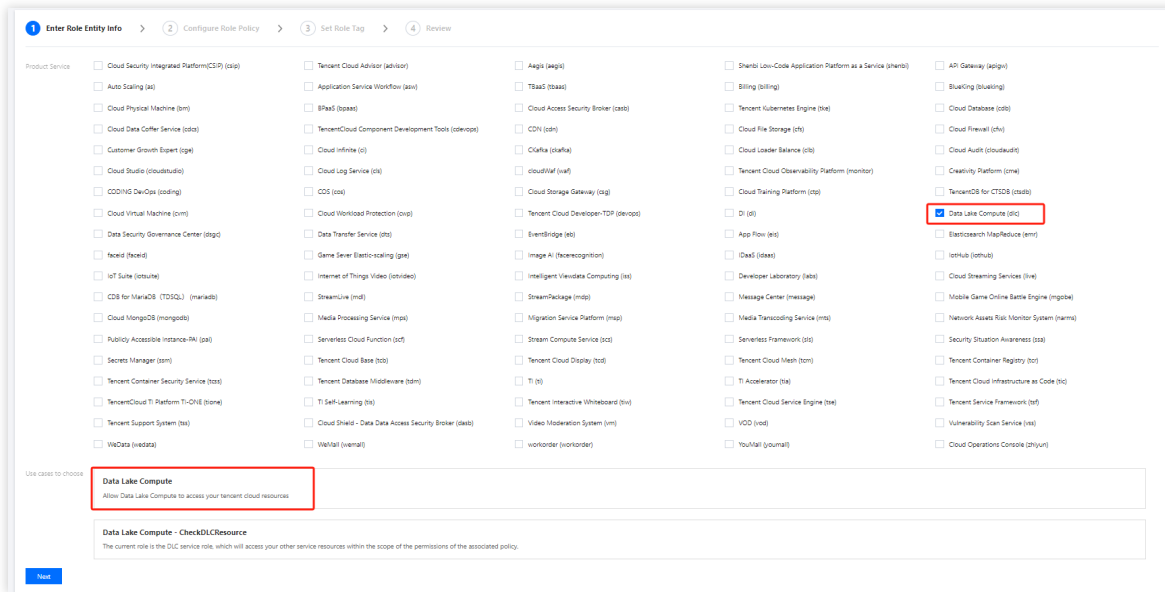
4. Select the template, set a name for the policy, and click **Save**.

## Step 2. Create a service role

1. Log in to the Tencent Cloud console and select **Cloud Access Management**. The logged-in account needs to have permissions to configure CAM; therefore, we recommend you use a root account or admin account.
2. Select **Role** on the left sidebar to enter the role management page. Click **Create Role** and select **Tencent Cloud Product Service**.



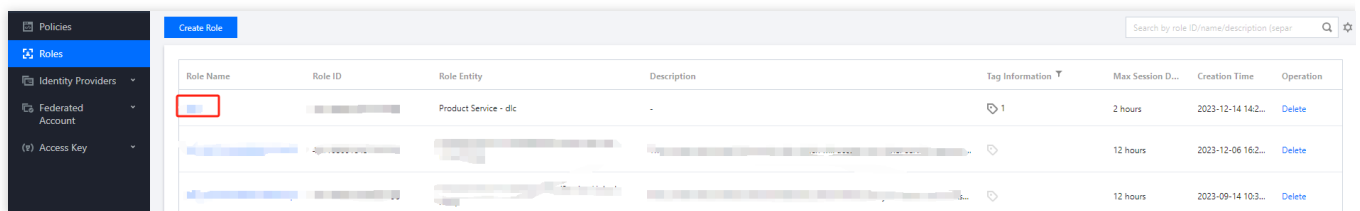
3. In the **Role Entity** service list, find and select **Data Lake Compute** and click **Next**.



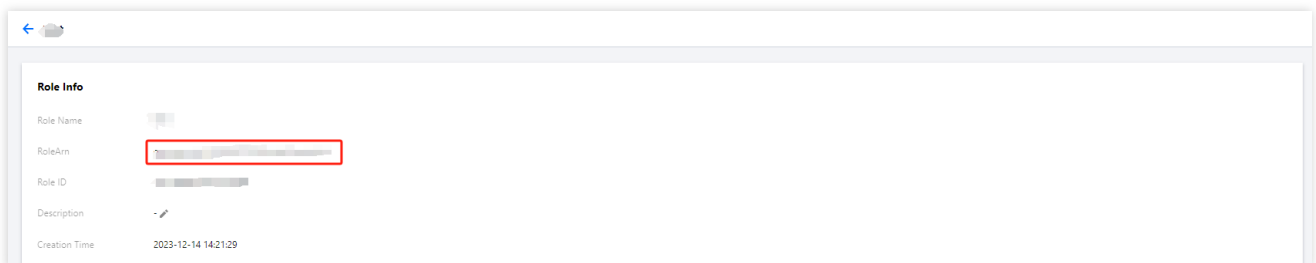
4. In the policy configuration, find and select the policy created in Step 1 and click **Next**.
5. Set a name for the role and click **Save**.

### Step 3. Get the role arn information

1. After creating the role in Step 2, return to the role list and find the created role.
2. Click **Role Name** to enter the role details page.



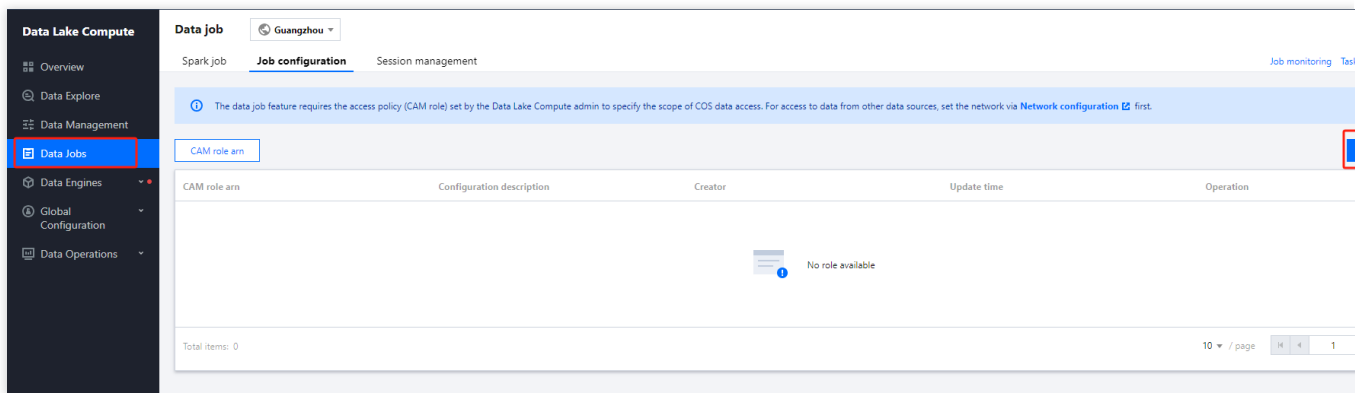
3. Find and copy the role arn information.



### Step 4. Configure the role arn in Data Lake Compute

1. Log in to the [Data Lake Compute console](#) with an admin account.

2. Select **Data job** on the left sidebar to enter the data job management page. Click **Job configuration** and select **CAM role arn**.
3. Click **Create role arn**.

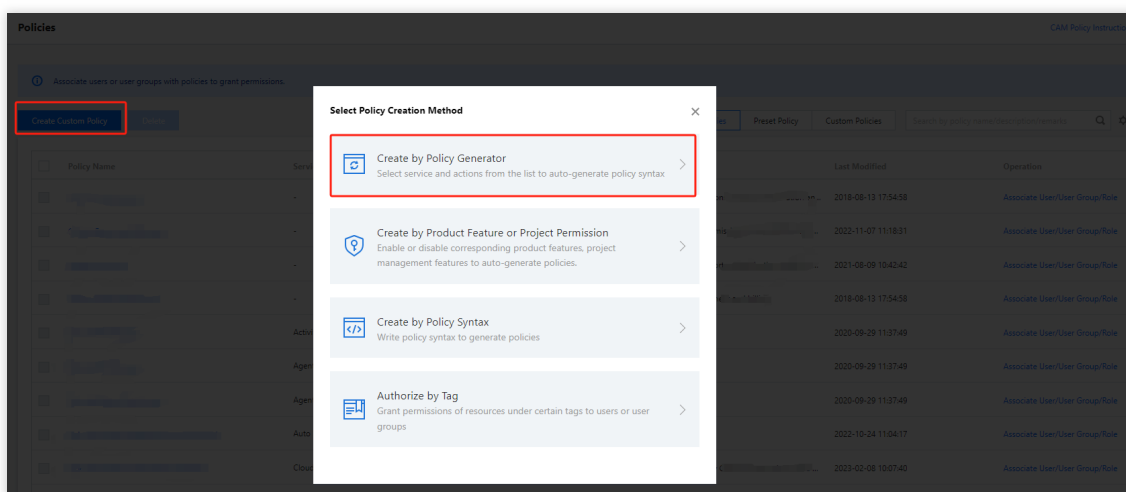


4. Paste the role arn information obtained in Step 3 in the input box and click **Save**.

## Appendix: Custom Policy Template

If the preset templates cannot meet your data management needs, you can configure a custom template in the following steps.

1. Log in to the [Tencent Cloud console](#) and select **Cloud Access Management**. The logged-in account needs to have permissions to configure CAM; therefore, we recommend you use a root account or admin account.
2. Select **Policies** on the left sidebar to enter the policy management page. Click **Create Custom Policy** and select **Create by Policy Generator**.



3. Select **Allow** as **Effect** and **COS** as **Service**. Select the resource scope as needed.

**Cloud Access Management**

- Dashboard
- Users
- User Groups
- Policies**
- Roles
- Identity Providers
- Federated Account
- Access Key

**Create by Policy Generator**

1 Edit Policy > 2 Associate User/User Group/Role

**Visual Policy Generator**    JSON

▼ COS(0 actions)

Effect \*     Allow     Deny

Service \*    COS (cos)

Action \*    **Select actions**  
Collapse     All actions (cos:\*)    Show More  
Add Custom Action

**Action Type**

- Read Show More
- Write Show More
- List Show More

Resource \*    Select resource

Condition     Source IP ⓘ  
Add other conditions

+ Add Permissions

**Next**    Characters: 114 (up to 6,144)

If you need to manage specific resources, click **Add a six-segment resource description** to add resources. You can use \* to indicate all the resources. For more information, see [Resource Description Method](#).

4. After completing the configuration, set a name for the policy and click **Save**. You can also select **Authorized Users** to authorize the policy to existing users.

# Creating Data Job

Last updated : 2024-07-17 17:45:32

## Preparations

Before creating a data job, you need to configure the CAM role arn to secure the data access from the data job. For detailed directions, see [Configuring Data Access Policy](#).

## Directions

1. Log in to the [Data Lake Compute console](#) and select **Data job** on the left sidebar.
2. Click **Create job**.

**Create job**
✕

---

**Basic info** ▲

Job name \*   
It can contain up to 100 characters in Chinese characters, letters, digits, and underscores ( ).

Job type \* Batch processing Stream processing SQL job

Data engine \*  ▼  
The billing mode of the selected data engine prevails. For more info, see [Data engine](#) . For network configuration of the data engine, see [Network configuration](#) .

Program package \*  COS  Upload

[Select a COS path](#)  
COS permissions are required, and .jar/.py files are supported.

Main class \*

Program entry parameter

Job parameter (--config)   
--config info, the parameter info started with "spark:", one entry per line.

CAM role arn \*  ↻  
It determines the data access scope of a Spark job. For configurations, see [Configure CAM role arn](#) .

**Network configuration** ▲

Create job
Cancel

Configure parameters as follows:

Parameter	Description
Job name	It can contain up to 40 letters, digits, and underscores.
Job type	<b>In batch:</b> Batch data jobs based on Spark JAR <b>In flow:</b> Flow data jobs based on Spark Streaming
Data source connection	Data source for <b>In batch</b> data jobs. Currently, it can only be CKafka, which needs to be configured in advanced in <b>Job configuration</b> .
Data engine	It can be a Spark job data engine for which you have the permission. If you select <b>Data source</b> , you can only select a data engine connected to the data source.
Program package	The JAR format is supported.

	You can select a local file of up to 5 MB in size or a file in COS. If the local file exceeds 5 MB, upload it to COS for use. You can directly enter a COS path.
Dependency JAR resource	The JAR format is supported. You can select multiple resources. You can select a local file of up to 5 MB in size or a file in COS. If the local file exceeds 5 MB, upload it to COS for use. You can directly enter multiple COS paths and separate them by semicolon.
Dependency file resource	You can select a local file of up to 5 MB in size or a file in COS. If the local file exceeds 5 MB, upload it to COS for use. You can directly enter multiple COS paths and separate them by semicolon.
CAM role arn	The data access policy configured in <b>Job configuration</b> , which specifies the scope of data accessible to a data job. For more information, see <a href="#">Configuring Data Access Policy</a> .
Main class	JAR package parameter in the main class. Separate multiple parameters by space.
Job parameter	<code>-config</code> information of the job, which starts with <code>spark.</code> in the format of <code>k=v</code> . Separate multiple parameters by line break. Example: <code>spark.network.timeout=120s</code>
Resource configuration	The engine resources that can be configured with the data job, the number of which cannot exceed the specifications of the selected data engine. Resource description: 1 CU ≈ 1-core 4 GB MEM Billable CUs = executor resource * executor quantity + driver resource Pay-as-you-go data engines are billed by the billable CUs.

3. After configuring the parameters, click **Save**.



# Managing Data Job

Last updated : 2024-07-17 17:47:58

This document describes how to manage a data job.

Edit a data job.

Start and stop a data job task.

View the data job and task details.

Delete a data job.

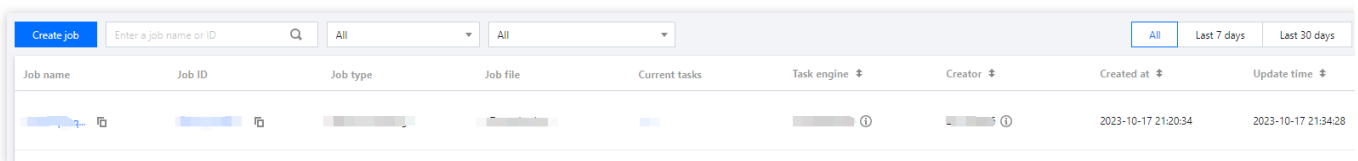
## Editing a data job

### Note:

A running data job cannot be edited.

The type of a data job cannot be changed. To change it, create a new data job as instructed in [Creating Data Job](#).

1. Log in to the [Data Lake Compute console](#), select the service region, and select **Data job** on the left sidebar.
2. Find the target data job and click **Edit**.



3. Edit the content and click **Save**.

## Starting and stopping a data job task

You can start and stop a created data job to generate corresponding tasks. A data job can generate multiple task instances and be executed multiple times.

Data task statuses are as follows:

Status	Description
Not started	Initial status after creation.
Running	The data task is running, during which the data job cannot be edited or deleted.
Successful	The task is executed successfully.
Failed	Failed to run the task. You can query the error message through the log or SparkUI.

Canceled

The task is manually canceled.

You can start and stop a data job task in the following steps:

1. Log in to the [Data Lake Compute console](#), select the service region, and select **Data job** on the left sidebar.
2. Find the target data job and click **Start** or **Stop** to change the task status.

### Note:

Starting a task instance will use compute engine resources. If the usage exceeds the configured upper limit, the task will be put into a queue.

Job name	Job ID	Job type	Job file	Current tasks	Task engine	Creator
[Redacted]	[Redacted]	[Redacted]	[Redacted]	[Redacted]	[Redacted]	[Redacted]

## Viewing the Data Job and Task Details

1. Log in to the [Data Lake Compute console](#), select the service region, and select **Data job** on the left sidebar.
2. Click **Job name** to enter the data job details page.

**Data job** Guangzhou

**Spark job** Job configuration Session management

[Create job](#) Enter a job name or ID   All

Job name	Job ID	Job type	Job file	Current tasks	Task engine
[Redacted]	[Redacted]	Batch processing	[Redacted]	[Redacted]	[Redacted]
[Redacted]	[Redacted]	Batch processing	[Redacted]	[Redacted]	[Redacted]
[Redacted]	[Redacted]	Batch processing	[Redacted]	[Redacted]	[Redacted]
[Redacted]	[Redacted]	Batch processing	[Redacted]	[Redacted]	[Redacted]
[Redacted]	[Redacted]	Batch processing	[Redacted]	[Redacted]	[Redacted]
[Redacted]	[Redacted]	Batch processing	[Redacted]	[Redacted]	[Redacted]
[Redacted]	[Redacted]	Batch processing	[Redacted]	[Redacted]	[Redacted]
[Redacted]	[Redacted]	Batch processing	[Redacted]	[Redacted]	[Redacted]
[Redacted]	[Redacted]	Batch processing	[Redacted]	[Redacted]	[Redacted]
[Redacted]	[Redacted]	Batch processing	[Redacted]	[Redacted]	[Redacted]

**Spark job details**

**Job info** Task history Monitoring and alerting

**Basic info**

Job name [Redacted]

Job ID [Redacted]

Current task ID [Redacted]

Current tasks [Redacted]

Task type **Batch processing**

Data engine [Redacted]

Job file [Redacted]

Main class [Redacted]

Program entry parameter [Redacted]

[Copy statement](#)

Job parameter [Redacted]

CAM role arn [Redacted]

Creator [Redacted]

Created at 2023-10-17 21:20:34

Update time 2023-10-17 21:34:28

**Network configuration**

Enhanced network [Redacted]

On the details page, you can view the basic information and task list of the data job. The task list contains the data

task information of the data job. You can view the task run log and SparkUI.

### Spark job details

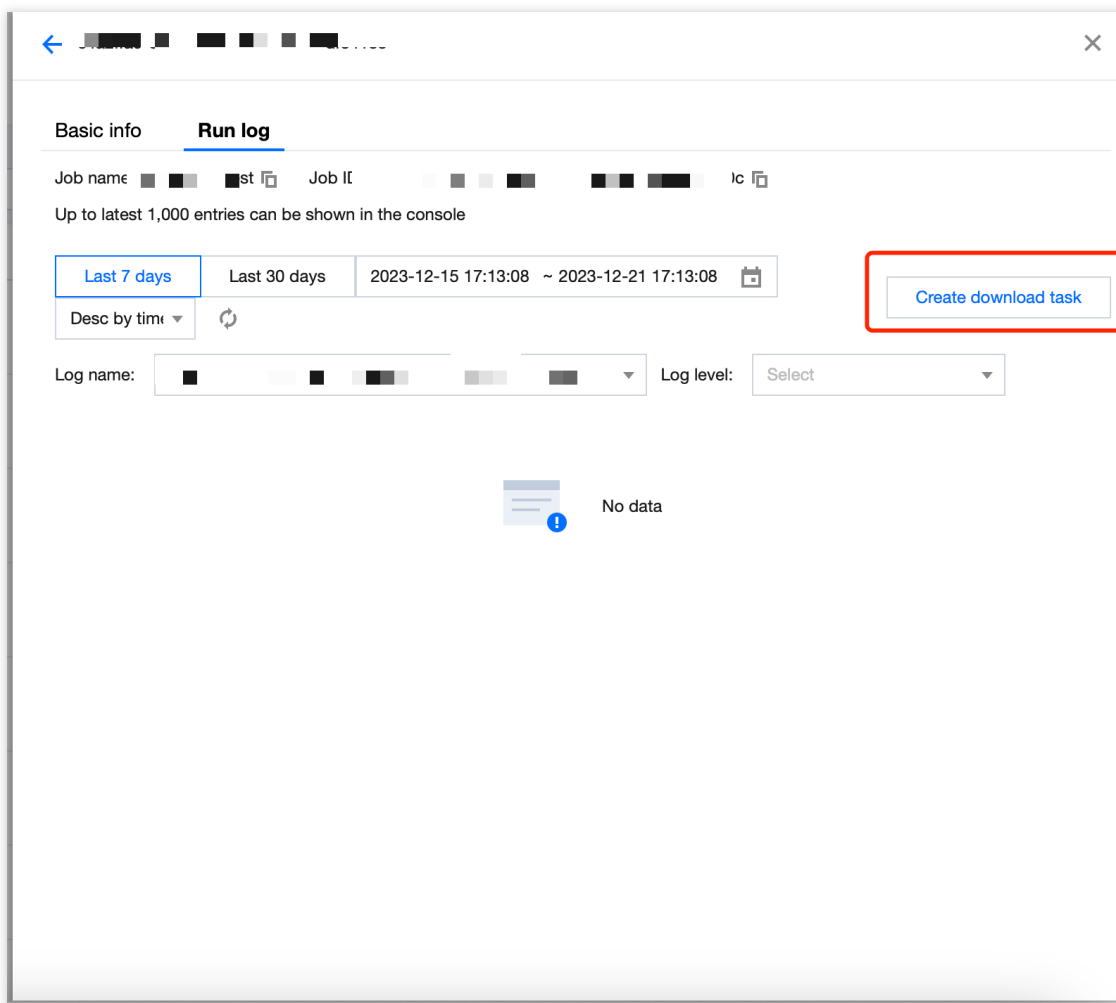
Job info **Task history** Monitoring and alerting

Select an executi... [Last 7 days](#) Last 30 days 2023-12-14 ~ 2023-12-20 [Refre](#)

Task ID	Executi...	Task submissi... ↕	Comput...	Operation
	Successful	2023-12-12 20:53:42	47.8s	<a href="#">Learn more Spark UI</a>

Total items: 1 10 ▾ / page 1 / 1 page

Click **Learn more** or **Task ID** to view the task details, which include the basic information and run log of the task. Currently, the run log allows you to view the last 1,000 data entries.



You can click **Create download task** to download the full log and click **Log download** to save the log locally.

#### Note:

The download record will be saved for three days, after which you cannot save the log locally and need to create a new download task.

## Deleting a data job

#### Note:

A data job with a running data task cannot be deleted.

1. Log in to the [Data Lake Compute console](#), select the service region, and select **Data job** on the left sidebar.
2. Find the target data job, click **Delete** > **OK**.

Job name	Job ID	Job type	Job file	Current tasks	Task engine ↕	Creator ↕
[REDACTED]	[REDACTED]	Batch processing	[REDACTED]	[REDACTED]	[REDACTED] ⓘ	[REDACTED] ⓘ

**Note:**

Note that deleting a data job will delete its data task information. Proceed with caution.

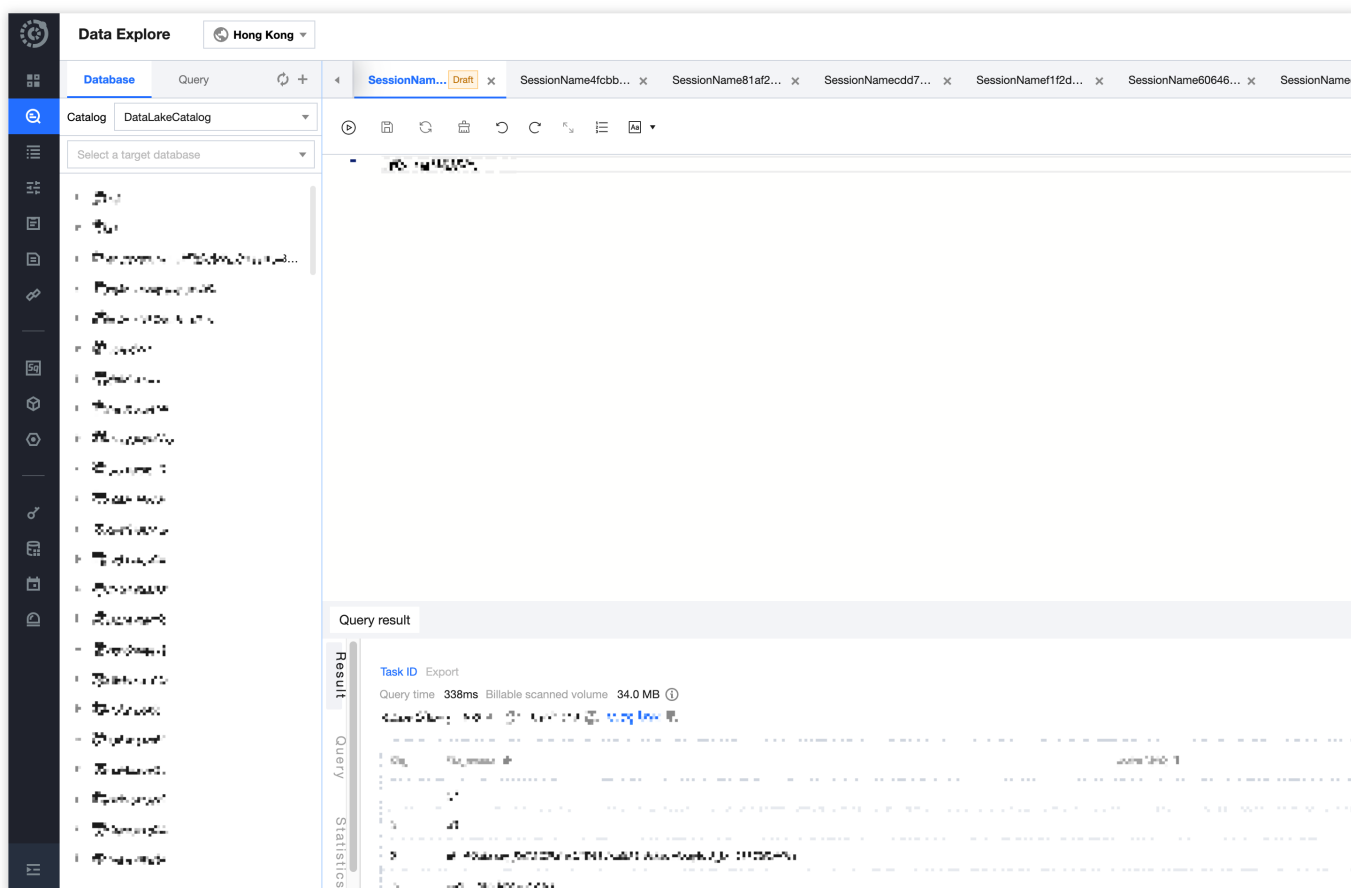
# Task History

Last updated : 2024-07-31 17:29:35

To facilitate users in querying historical task records, DLC provides three methods to search and process historical tasks.

## View historical tasks run in the Query Editor

1. Log in to [DLC console](#), select the service region.
2. Enter the **Data Exploration Page**, click on **Run History** within a single Session to view the task run history for that Session.
3. Click on the history record **Batch ID** to view the corresponding execution results on the left



Each Session's run history is independent, and a maximum of 45 days of run history is kept.

Historical task result data is saved for 24 hours. To view task results beyond 24 hours, the task must be rerun.

## View data import history in the Data Management feature

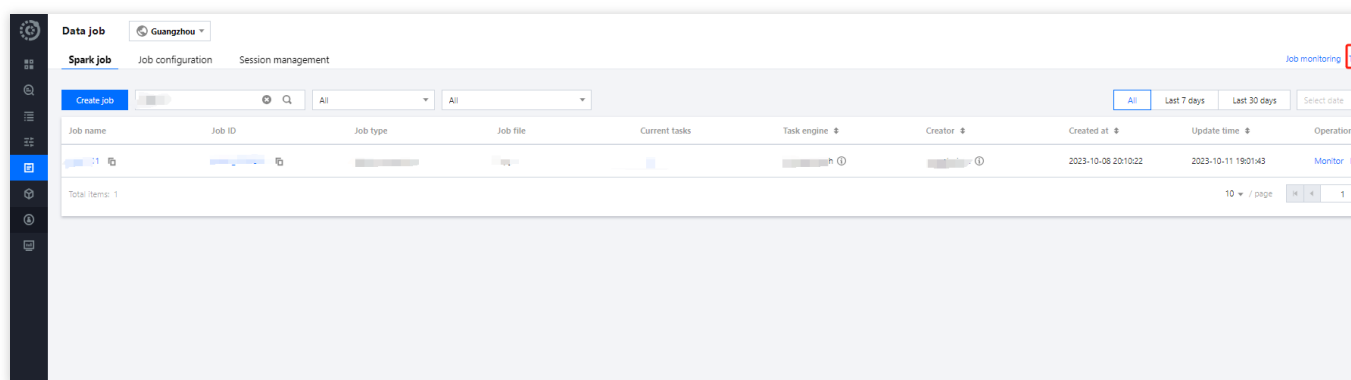
1. log in to [DLC Console > Data Management](#), select the service region.

### Note:

Log in to the account requires database-related permissions.

2. Click on **Task History** in the top right corner to query data import history tasks.

3. Supports viewing historical tasks from the past 45 days

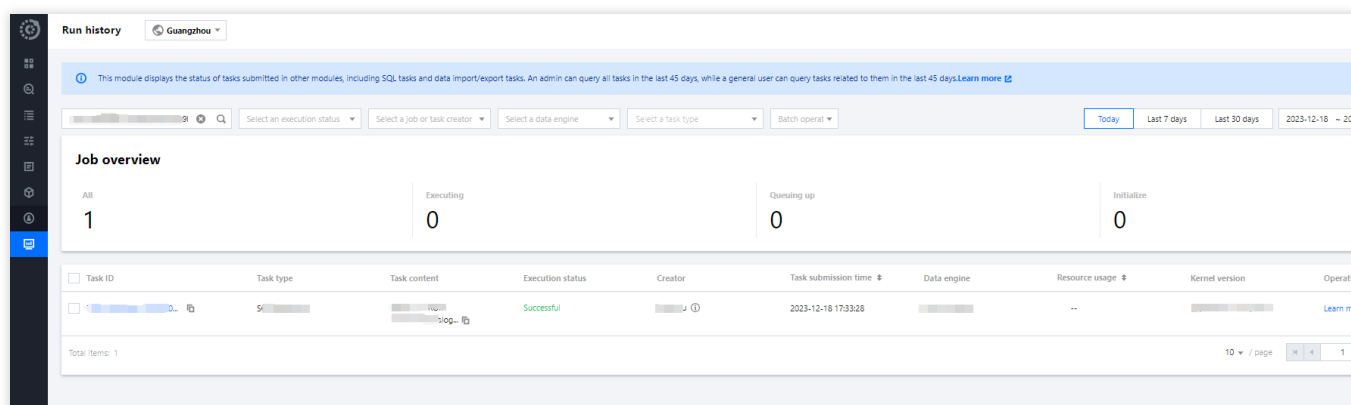


## View historical tasks in the Historical Operation feature

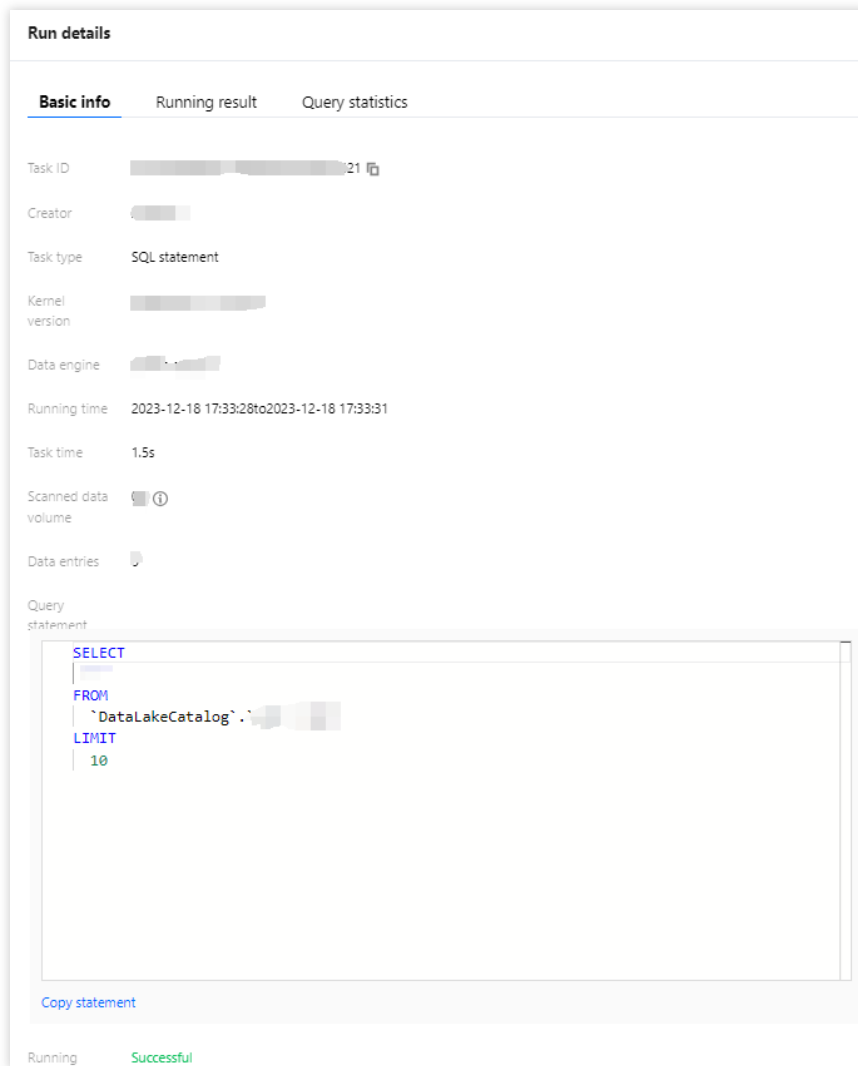
1. log in to [DLC Console > Historical Operation](#), select the service region.

2. Enter the Historical Operation page, where administrators can view all historical operation tasks from the past 45 days, and ordinary users can query tasks related to themselves from the past 45 days.

3. Supports filtering by task type, execution status, creator, data type, etc.



4. click **View Details** to see the task execution details and results.



**Run details**

**Basic info**   Running result   Query statistics

Task ID: [redacted] 21

Creator: [redacted]

Task type: SQL statement

Kernel version: [redacted]

Data engine: [redacted]

Running time: 2023-12-18 17:33:28 to 2023-12-18 17:33:31

Task time: 1.5s

Scanned data volume: [redacted]

Data entries: [redacted]

Query statement

```
SELECT
FROM
`DataLakeCatalog`.`[redacted]`
LIMIT
10
```

[Copy statement](#)

Running Successful

Historical task result data is saved for 24 hours. To view task results beyond 24 hours, the task must be rerun. You can directly **Copy Statement** to Data Exploration to execute the task.

You can directly click **Task ID** to quickly switch and view the task execution details.

For tasks that are running, you can **Cancel** them.



# Engine Management

## SuperSQL Engine

### SuperSQL Engine Overview

Last updated : 2024-07-31 17:51:33

Data engines empower the data analysis and computing service in Data Lake Compute. They are used in all computing operations and can be public or private based on your needs.

## Public engine

The Data Lake Compute service comes with the shared public engine, which is applicable to low-frequency analysis use cases with small data volumes. With this highly flexible and available engine, you don't need to configure or manage resources. Fees are charged by the scanned data volume of running tasks. For billing details, see [Billing Overview](#).

Since Data Lake Compute adopts serverless architecture, it needs to schedule the data engine for task execution for the first time over a period of time, which may take a longer time.

## Private engine

A private engine is a dedicated data engine that you purchase on a pay-as-you-go basis. For billing details, see [Billing Overview](#).

**Pay-as-you-go:** This billing mode is highly flexible and stable, where fees are charged by the CU usage. It is applicable to use cases where data is analyzed regularly, with compute resources elastically scaled based on the business load.

**Monthly subscription:** This billing mode is applicable to use cases where large amounts of data require long-term and stable analysis, with compute resources elastically scaled based on the business load. It guarantees always available resources with no need to wait for resource startup. Fees are charged by month based on the cluster specification (elastic clusters are billed by CU usage though).

## Compute engine types

A private engine can work with different compute engines in different use cases.

**SparkSQL:** It is suitable for stable and efficient offline SQL tasks.

**Spark job:** It is suitable for native Spark stream/batch data job processing.


Presto: It is suitable for agile and fast interactive query and analysis.

**Note:**

The compute engine type does not affect the unit price of a private engine.

## Engine scaling rules

Engine scaling rules can be configured in the [create resource](#) or in the spec configuration of the [SuperSQL Engine](#).

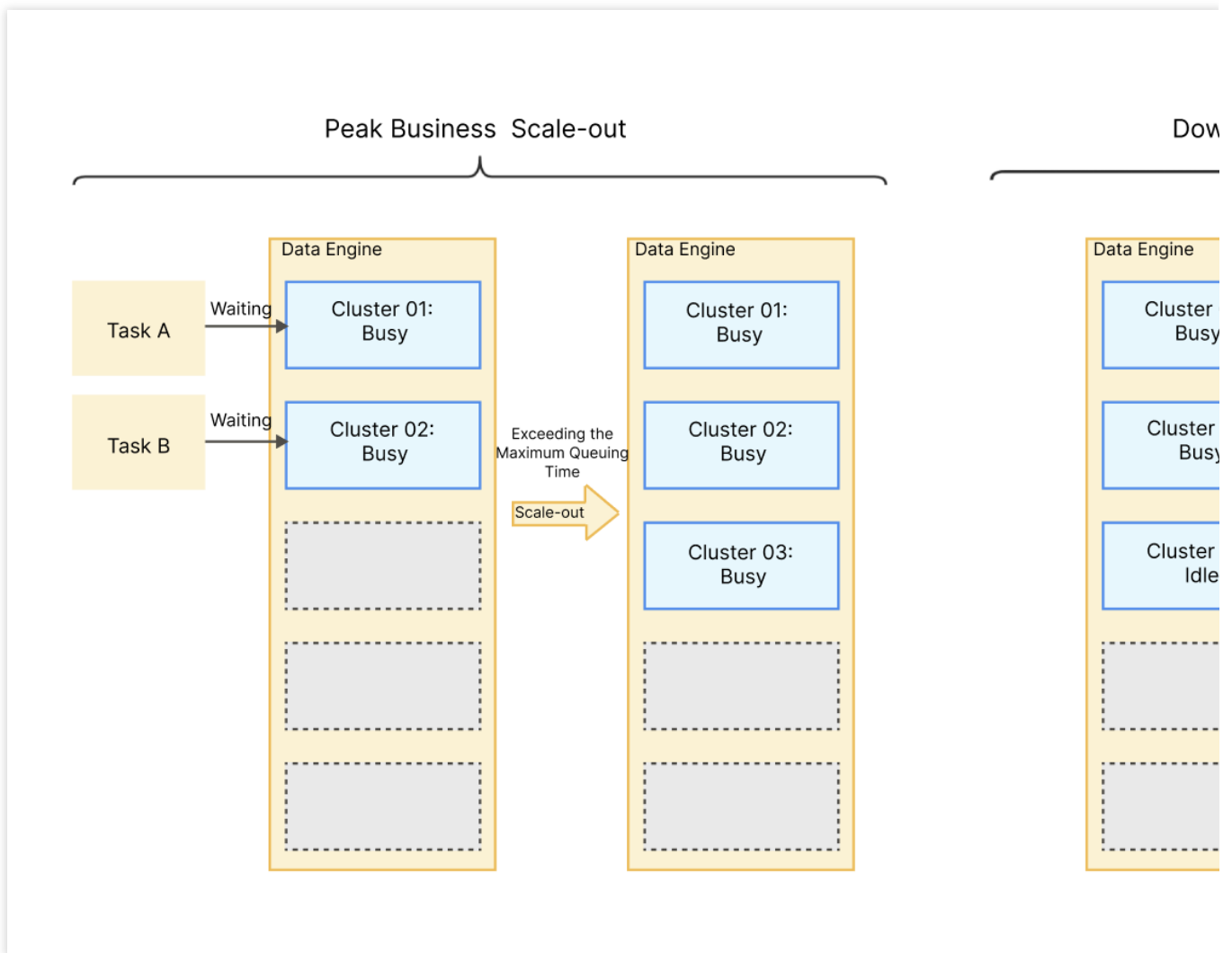
Cluster count	<input type="button" value="-"/> <input type="text" value="1"/> <input type="button" value="+"/>	Multiple clusters with fixed specs can be configured in a da
Max task concurrency of a cluster	<input type="button" value="-"/> <input type="text" value="20"/> <input type="button" value="+"/>	The max number of concurrent tasks that a cluster can proc longer compute time. When the concurrency reaches the cc be queued up.
Cluster scaling rules	<input checked="" type="radio"/> Yes <input type="radio"/> No <a href="#">Scaling rules</a> 	
Elastic clusters	<input type="button" value="-"/> <input type="text" value="1"/> <input type="button" value="+"/>	For elastic clusters, resources concurrency and queue time
Task queue-up time limit	<input type="button" value="-"/> <input type="text" value="1"/> <input type="button" value="+"/> Minute	The max task queue-up time. triggered immediately after a time exceeds this value, the c resources are made available resources required).

Number of clusters refers to the number of permanent clusters in the engine. Number of clusters + number of elastic clusters = the maximum number of clusters that can be reached when the engine is elastic.

Basic rule: Data engine scaling may occur only if the configured maximum cluster count is larger than the minimum cluster count.

Scale-out rule: The system will scale out a data engine as configured when queuing tasks cannot be accommodated by the idle concurrency and no clusters are being initialized.

Scale-in rule: If the number of clusters in the data engine is greater than the number of resident clusters, the average cluster load is lower than 20%, and some clusters are idle, the system shrinks the data engine.



**Note:** The cluster count of a data engine cannot be smaller than the minimum cluster count. A pay-as-you-go cluster can be suspended if it is not needed.

## Engine running status

A cluster may be in one of the following eight statuses: Starting, Running, Suspended, Suspending, Changing configuration, Isolated, Isolating, Recovering.

**Starting:** The cluster is being started. In this case, a pay-as-you-go private engine is not billed. A starting cluster cannot be selected for data computing.

**Running:** The cluster is running and can be selected for data computing.

**Suspended:** The cluster is suspended and cannot be selected for data computing.

**Suspending:** The cluster is being suspended and cannot be selected for data computing. This will affect running tasks.

**Changing configuration:** The cluster is undergoing a configuration change and cannot be selected for data computing.

**Isolated:** The cluster is isolated due to overdue payments and cannot be selected for data computing.

**Isolating:** The cluster is being isolated due to overdue payments and cannot be selected for data computing. This will affect running tasks.

**Recovering:** The cluster is being recovered from the **Isolated** status to the **Running** status after the account is topped up. It cannot be selected for data computing.

# Purchasing Private Data Engine

Last updated : 2024-07-17 17:55:49

A private data engine in Data Lake Compute supports pay-as-you-go and monthly subscription billing modes. For billing details, see [Billing Overview](#).

## Private engine purchase

You can purchase on the Data Lake Compute purchase page or in the console as instructed below:

1. Log in to the [Data Lake Compute console](#) and select the service region. You need to have the Tencent Cloud admin or financial collaborator permission.
2. Click **Data engine** on the left sidebar to enter the data engine management page.
3. Click **Create resource** in the top-left corner to enter the **Resource configuration** page. Configure the resource as needed and view the estimated price.

i Data Lake Compute offers both public and private data engines. A public data engine is managed by Data Lake Compute and billed by scanned required; a private data engine can be billed on a pay-as-you-go basis or subscribed monthly. For more billing info, see [Billing Overview](#). A private data engine supports the auto-suspension or scheduled suspension policy, with no fees charged on it after suspension. For operations and notes, see [Managing Private Data Engine](#).

Create resource

Bill query ↗

Renewal management ↗

Select a resource

Engine Name/ID	Engine type	Engine Status	Kernel version <span style="font-size: 0.8em;">i</span>	Billing mode	Auto-renewal
DataEngine-iwxhwnud <span style="font-size: 0.8em;">🗑️</span>	SparkSQL	Running	SuperSQL-S 3.5	Expire	No
DataEngine-p3d2xfq1 <span style="font-size: 0.8em;">🗑️</span>	Presto	Starting <span style="font-size: 0.8em;">i</span>	SuperSQL-P 1.0		--
DataEngine-public-1313074... <span style="font-size: 0.8em;">🗑️</span>	Presto	Running	SuperSQL-P 1.0-public	volume	--

Total items: 3

4. Confirm the price and make the purchase.

## Data Lake Compute [Back](#)

Engine edition Beta

SuperSQL engine
  Standard engine

Billing mode [Detailed comparison](#)

Pay-as-you-go
  Monthly subscription

In this mode, a cluster is billed based on the CUs used and can be suspended when no task is in progress. A suspended cluster incurs no cost loads and irregular task cycles.

Region

Hong Kong/Macao/TaiWan (China Region)
  Southeast Asia
  Eastern U.S.
  Europe
  Southeast Asia Pa

Hong Kong
  Singapore
  Virginia
  Frankfurt
  Jakarta

Cloud products in different regions are not interconnected over private networks and the region cannot be changed after you purchase the service. Please select the region nearest to your customers to reduce access latency.

### Cluster configuration

#### Basic configuration

Compute engine type

SparkSQL
  Spark job
  Presto

This is a memory engine for distributed SQL query. It supports real-time data write to SQL and real-time result return in Data Explore. It is suited for SparkSQL engine.

Kernel version

SuperSQL-P 1.0

SuperSQL-P is a Tencent-developed Presto-based engine kernel for interactive query and analytics. Syntax rules supported by different kernels see [Kernel Versions](#)

### Configuration parameter description:

**Region:** Cloud products in different regions are not interconnected over private networks and the region cannot be changed after you purchase the service. Proceed with caution.

**Compute engine:** Presto and Spark engines are supported. Note that the engine cannot be changed once purchased. Presto is suitable for faster interactive query and analysis and multi-source federated query, while Spark is suitable for more stable offline tasks with large data volumes.

**Cluster spec:** Cluster specification is measured in CU. 1 CU equals to 1 CPU core and 4 GB memory of compute resources. The specification determines the amount of compute resources during task execution and can be purchased as needed.

#### Note:

If you need more than 152 CUs, submit a ticket for assistance.

**Min cluster count:** Set the minimum number of clusters during cluster start or resident resources in a monthly subscribed cluster. Multiple clusters can deliver a higher concurrency.

**Max cluster count:** Set the maximum number of clusters for elastic scaling. If it is the same as the minimum cluster count, elastic scaling is not enabled for the cluster.

**Auto-start:** If it is enabled, a suspended data engine will be automatically started after receiving a task request.

**Note:**

As pay-as-you-go resources are not reserved, it is possible that they cannot be started right away. If you need resident and stable compute resources, purchase a monthly subscribed data engine instead.

**Suspension policy:** Configure the suspension method of a pay-as-you-go data engine. Automatic suspension and scheduled suspension are supported. A suspended pay-as-you-go data engine will not incur fees.

**Auto-suspension:** The data engine will be automatically switched to the **Suspended** status after it has been idle for a certain period of time.

**Timing policy:** You can configure scheduled start and suspension policies by week. The system will start or suspend clusters regularly as configured.

**Suspension after task end:** After the specified time elapses, if a task is running, the system will automatically suspend the data engine within five minutes after the task ends.

**Suspension after task pause:** After the specified time elapses, if a task is running, the system will pause the task and suspend the data engine immediately.

**Advanced configuration:** If you need to use federated query, configure the IP range in the advanced configuration.

**Tag:** Set tags to categorize purchased resources and allocate costs. For more information, see [Associating Tag with Private Engine Resource](#).




## Bill query

You can query bills in the Data Lake Compute console in the following steps:

1. Log in to the [Data Lake Compute console](#) and select the service region. You need to have the Tencent Cloud admin or financial collaborator permission.
2. Click **Data engine** on the left sidebar to enter the data engine management page.
3. Click **Bill query** to view the detailed bill and settlement information (the financial collaborator permission is required).

**i** Data Lake Compute offers both public and private data engines. A public data engine is managed by Data Lake Compute and billed by scanned data required; a private data engine can be billed on a pay-as-you-go basis or subscribed monthly. For more billing info, see [Billing Overview](#). A pay-as-you-go engine supports the auto-suspension or scheduled suspension policy, with no fees charged on it after suspension. For operations and notes, see [Managing Private](#).

[Create resource](#) [Bill query](#) [Renewal management](#)

Engine Name/ID	Engine type	Engine Status	Kernel version	Billing mode	Auto-renewal	
 DataEngine-iwxhwnu01	SparkSQL	Running	SuperSQL-S 3.5	Monthly subscription 2024-08-02 11:37:06 Expire	No	
 DataEngine-p3d2xtq1	Presto	Starting <b>i</b>	SuperSQL-P 1.0	Pay-as-you-go	--	
 DataEngine-public-1313074...	Presto	Running	SuperSQL-P 1.0-public	Pay by scanned data volume	--	

Total items: 3

## Renewal management

For a monthly subscribed private data engine, you can perform renewal and other operations in the Data Lake Compute console > Renewal management > Resource management in the following steps:

1. Log in to the [Data Lake Compute console](#) and select the service region. You need to have the Tencent Cloud admin or financial collaborator permission.
2. Click **Data engine** on the left sidebar to enter the data engine management page.
3. Click **Renewal management** to enter the resource list and renew resources (the financial collaborator permission is required).



**i** Data Lake Compute offers both public and private data engines. A public data engine is managed by Data Lake Compute and billed by scanned data required; a private data engine can be billed on a pay-as-you-go basis or subscribed monthly. For more billing info, see [Billing Overview](#). A pay-as-you-go engine supports the auto-suspension or scheduled suspension policy, with no fees charged on it after suspension. For operations and notes, see [Managing Private](#).

Create resource

Bill query

**Renewal management**

Select a resource

Engine Name/ID	Engine type	Engine Status	Kernel version <b>i</b>	Billing mode	Auto-renewal
	SparkSQL	Running	SuperSQL-S 3.5	Monthly subscription	No
	Presto	Starting <b>i</b>	SuperSQL-P 1.0		--
	Presto	Running	SuperSQL-P 1.0-public		--

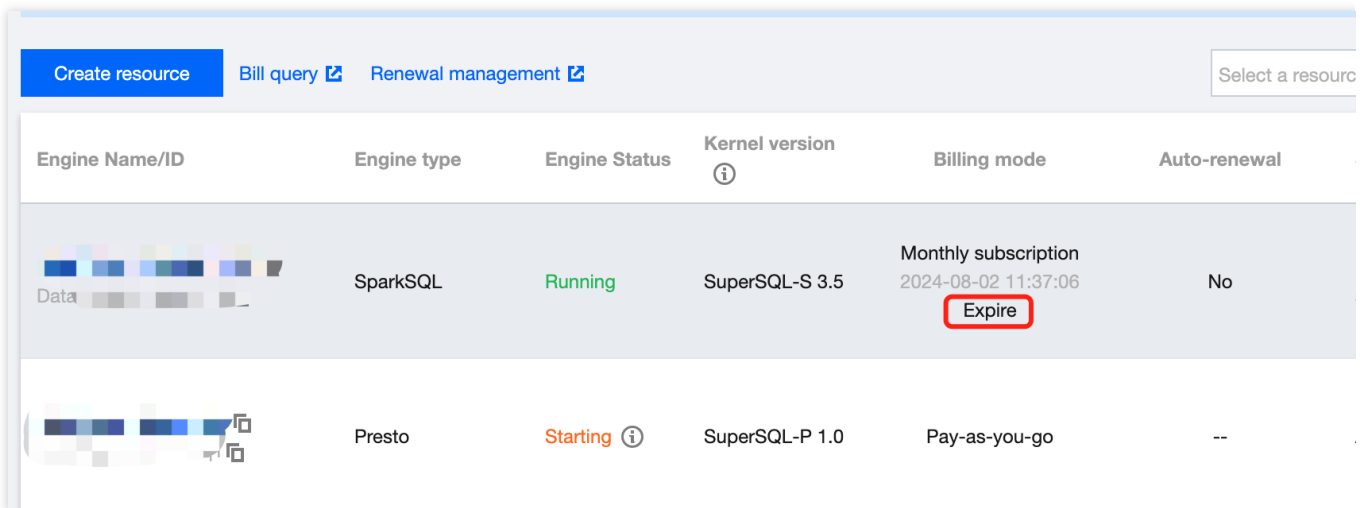
Total items: 3

# Renewing SuperSQL Engine

Last updated : 2024-07-31 17:55:25

You can renew a monthly subscribed data engine that has not expired or is isolated in the Data Lake Compute console.

1. Log in to the [Data Lake Compute console](#) and select the service region. You need to have the Tencent Cloud admin or financial collaborator permission.
2. Click **Data engine** on the left sidebar to enter the data engine management page.
3. Find the target data engine and click **More > Renew**. You can also renew resources that will expire soon (in seven days) by clicking **Renew** next to the expiration time.



Engine Name/ID	Engine type	Engine Status	Kernel version	Billing mode	Auto-renewal
Data [blurred]	SparkSQL	Running	SuperSQL-S 3.5	Monthly subscription 2024-08-02 11:37:06 <b>Expire</b>	No
[blurred]	Presto	Starting	SuperSQL-P 1.0	Pay-as-you-go	--

4. Check the renewal term and price and click **Confirm**. The renewal will be completed after the order is confirmed and paid.

## Note:

The billing cycle of a data engine that is renewed from the isolated status will start from the expiration date of the previous cycle.

# Managing Private Data Engine

Last updated : 2024-07-17 18:02:09

## Note:

You don't need to manage the public engine, as it is managed by Data Lake Compute in a unified manner.

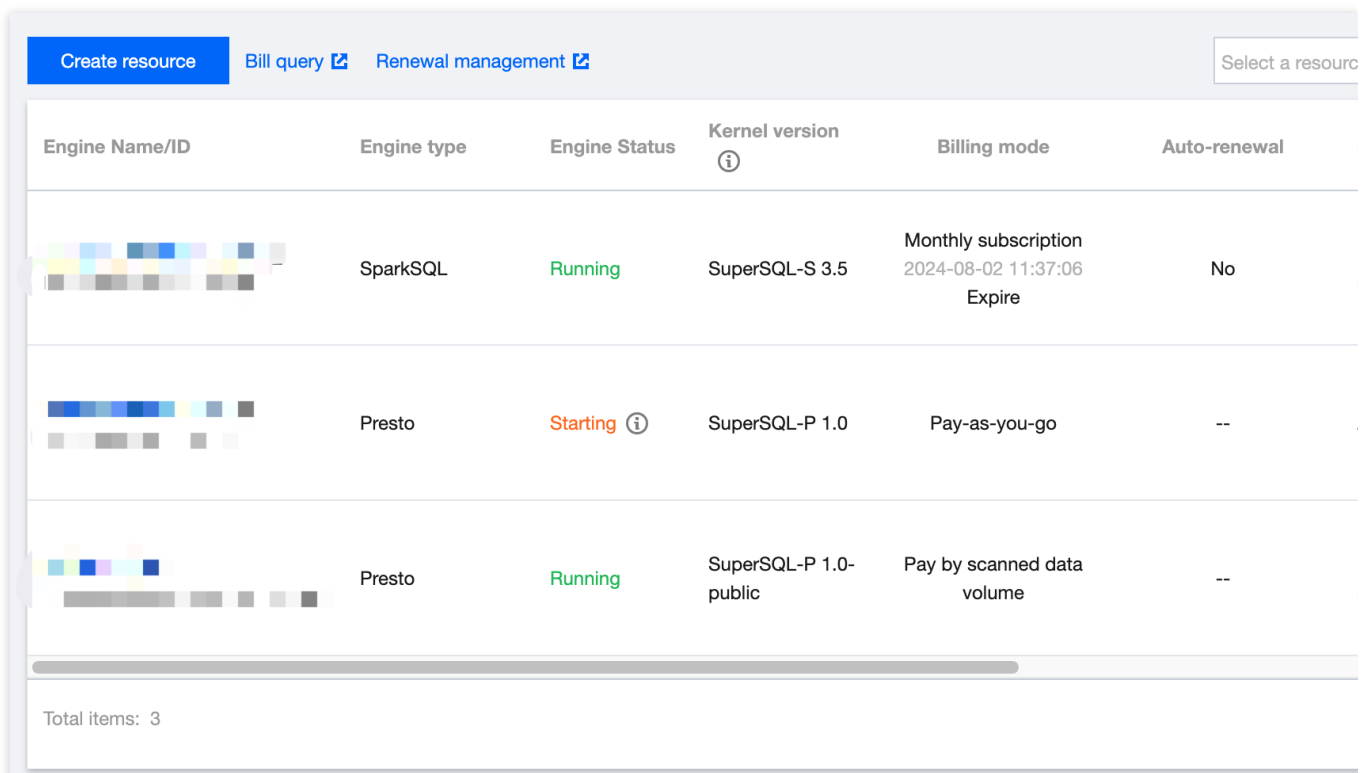
## Modifying the private engine configuration

### Note:

Fees may change as the private engine configuration changes. For more information, see [Configuration Adjustment Fees Description](#).

### Option 1. Data engine list

1. Log in to the [Data Lake Compute console](#) and select the service region. You need to have the Tencent Cloud admin or financial collaborator permission.
2. Click **Data engine** on the left sidebar to enter the data engine management page.
3. Find the target private engine and click **Spec configuration** on the right to enter the configuration modification page, where you can modify the cluster specification and elastic scaling policy.
4. After making changes, click **Save** to submit the order and make the payment.



Engine Name/ID	Engine type	Engine Status	Kernel version	Billing mode	Auto-renewal
	SparkSQL	Running	SuperSQL-S 3.5	Monthly subscription 2024-08-02 11:37:06 Expire	No
	Presto	Starting <span>i</span>	SuperSQL-P 1.0	Pay-as-you-go	--
	Presto	Running	SuperSQL-P 1.0- public	Pay by scanned data volume	--

Total items: 3

### Option 2. Data engine details

1. Log in to the [Data Lake Compute console](#) and select the service region. You need to have the Tencent Cloud admin or financial collaborator permission.
2. Click **Data engine** on the left sidebar to enter the data engine management page.
3. Find the target private engine and click the cluster name to enter the cluster details page, where you can modify the cluster specification and elastic scaling policy.
4. Adjust the parameters as needed and click **Save**.

The screenshot shows the 'Basic configuration' page for a SuperSQL engine. The page has a breadcrumb trail: 'SuperSQL engine' followed by several redacted names. There are two tabs: 'Basic configuration' (active) and 'Cluster monitoring'. The 'Basic info' section includes:

- Engine name: [Redacted]
- Resource ID: DataEngine-p3d2xfq1
- Description: [Redacted]
- Region: Hong Kong/Macao/TaiWan (China Region)-Hong Kong
- Engine Status: Starting (with a refresh icon)
- Billing mode: Pay-as-you-go
- Tag: No tag (with an edit icon)

A note below the tag field states: 'Tags are used to categorize resources. To learn more, see [Tag Documentation](#)'.

The 'Configuration info' section includes:

- Engine type: Presto
- Kernel version: [Redacted]
- Cluster count: 1
- Auto-scaling: Yes
- Elastic cluster co: [Redacted]
- Task queue-up time limit: 0 minute(s)
- Auto-start: Yes
- Auto-suspension: [Redacted]
- IP range of cluster: 10.255.252.0/22
- Network configuration: --

## Modifying the private engine information

1. Log in to the [Data Lake Compute console](#) and select the service region. You need to have the Tencent Cloud admin permission.
2. Click **Data engine** on the left sidebar to enter the data engine management page.
3. Find the target private engine and click the cluster name to enter the cluster details page, **where you can modify the cluster description, automatic start policy, and suspension policy**.
4. Adjust the parameters as needed and click **Save**.

The screenshot displays the configuration page for a SuperSQL engine. The breadcrumb navigation shows 'SuperSQL engine' followed by several redacted identifiers. The main content is split into two columns:

- Basic info:**
  - Engine name: [Redacted]
  - Resource ID: DataEngine-p3d2xfq1
  - Description: [Redacted]
  - Region: Hong Kong/Macao/TaiWan (China Region)-Hong Kong
  - Engine Status: Starting (with a refresh icon)
  - Billing mode: Pay-as-you-go
  - Tag: No tag (with an edit icon)
  - Footnote: Tags are used to categorize resources. To learn more, see [Tag Documentation](#)
- Configuration info:**
  - Engine type: Presto
  - Kernel version: [Redacted]
  - Cluster count: 1
  - Auto-scaling: Yes
  - Elastic cluster co: [Redacted]
  - Task queue-up time limit: 0 minute(s)
  - Auto-start: Yes
  - Auto-suspension: [Redacted]
  - IP range of cluster: 10.255.252.0/22
  - Network configuration: --

**Suspension policy:** Configure the suspension method of a pay-as-you-go data engine. Automatic suspension and scheduled suspension are supported. A suspended pay-as-you-go data engine will not incur fees.

**Auto-suspension:** The data engine will be automatically switched to the **Suspended** status after it has been idle for 15 minutes.

**Timing policy:** You can configure scheduled start and suspension policies by week. The system will start or suspend clusters regularly as configured.

**Suspension after task end:** After the specified time elapses, if a task is running, the system will automatically suspend the data engine within five minutes after the task ends.

**Suspension after task pause:** After the specified time elapses, if a task is running, the system will pause the task and suspend the data engine immediately.

## Enable suspension policy management

It supports the configuration of start & suspend policies for the exclusive data engine of billing by volume, which facilitates management and cost control.

### Note :

If the pay-as-you-go data engine is not suspended, charges will be generated. If the data engine is not needed, suspend it in time.

**Startup policy:** Supports automatic start, manual start, and scheduled start of the data engine.

**Automatic start:** After the configuration, if the data engine is in the suspended state and a task is submitted to the data engine, the data engine will automatically start.

**Manual start:** After the configuration, if the data engine is in the suspended state, you need to manually start the data engine before processing data tasks.

**Periodic startup:** You can configure a weekly periodic startup policy. The system periodically starts the cluster based on the configuration rules.

Timing policy

Scheduled start

Scheduled suspension

Suspension option  Suspension after task  Suspend after task pause

The suspension rules that can be set after the scheduled suspension feature is enabled. "Suspension after task" means that resources will be suspended at the specified time after the last task is ended. "Suspension after task pause" means that resources will be suspended at the specified suspension time after the task is paused.

**Suspension policy:** Supports the suspension mode of the data engine for charging by volume, including automatic suspension and scheduled suspension. Pay-as-you-go data engines do not incur any costs when suspended.

**Automatic suspension:** After the configuration, the data engine automatically switches to the suspended state 10 minutes after there is no task, and the triggering time can be configured.

Auto-suspension

If this option is enabled, the data engine is automatically suspended after the set trigger time of no task.

Auto-trigger time    min

Valid range: 1–999 min, which will affect the time waiting for suspending the data engine.

**Periodic policy -** You can configure weekly periodic start and suspension policies. The system starts and suspends the cluster periodically based on the configuration rules.

**Suspend after Completion:** If a task is being executed by the data engine within the specified time, the data engine automatically suspends the task within 5 minutes after the task is completed.

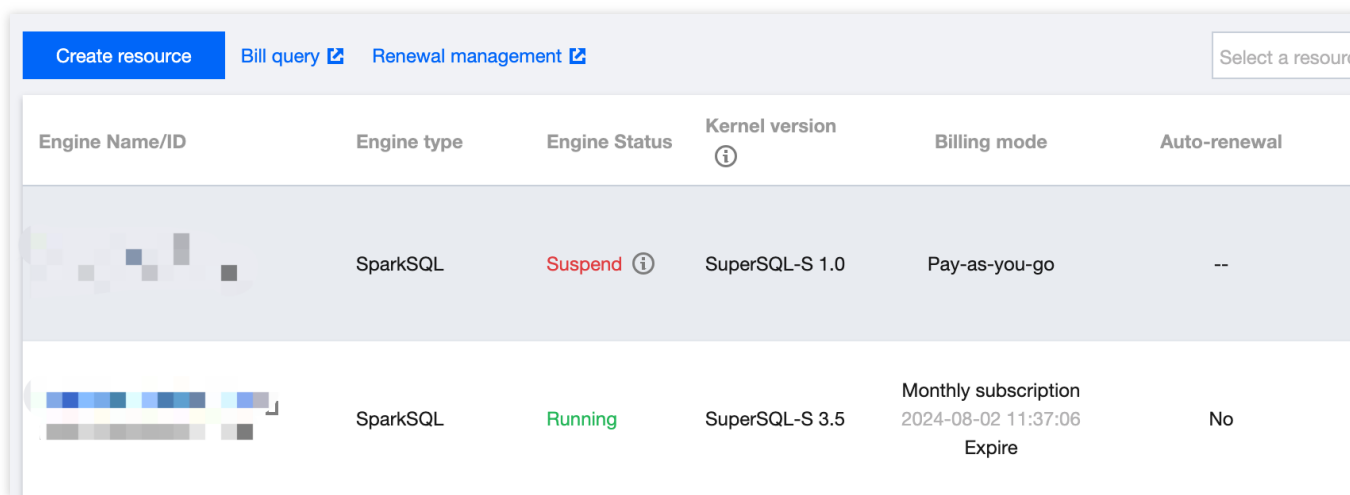
**Suspend after Automatic pause:** If a task is being executed on the data engine within the specified time, the system suspends the task and immediately suspends the data engine.

## Manually suspending/starting a private engine

### Note:

Monthly subscribed resources are resident and cannot be suspended.

1. Log in to the [Data Lake Compute console](#) and select the service region. You need to have the Tencent Cloud admin permission.
2. Click **Data engine** on the left sidebar to enter the data engine management page.
3. Find the target private engine, click **More**, and select **Start** or **Suspend** in the drop-down list.



Engine Name/ID	Engine type	Engine Status	Kernel version	Billing mode	Auto-renewal
[Blurred]	SparkSQL	Suspend ⓘ	SuperSQL-S 1.0	Pay-as-you-go	--
[Blurred]	SparkSQL	Running	SuperSQL-S 3.5	Monthly subscription 2024-08-02 11:37:06 Expire	No

## Terminating a private engine

You can terminate a data engine that is no longer needed. A monthly subscribed data engine will be returned automatically after termination. For more information, see [Refund](#).

### Note:

Note that a pay-as-you-go data engine cannot be recovered once terminated. Proceed with caution.

1. Log in to the [Data Lake Compute console](#) and select the service region. You need to have the Tencent Cloud admin permission.
2. Click **Data engine** on the left sidebar to enter the data engine management page.
3. Find the target private engine (only suspended clusters can be terminated), click **More**, and select **Terminate** in the drop-down list.
4. Confirm the termination.

Engine Name/ID	Engine type	Engine Status	Kernel version	Billing mode	Auto-renewal
[Redacted]	SparkSQL	Suspend ⓘ	SuperSQL-S 1.0	Pay-as-you-go	--
[Redacted]	SparkSQL	Running	SuperSQL-S 3.5	Monthly subscription 2024-08-02 11:37:06 Expire	No

## Cluster running logs

Data Lake Compute provides running logs within 14 days for private engines to help you stay informed of the start, suspension, and scaling of clusters. Cluster logs mainly include the following content:

Start time: The time when the cluster starts working.

Suspension time: The time when the cluster stops working.

Scale-out record: The time of the cluster scale-out and the number of added clusters.

Scale-in record: The time of the cluster scale-in and the number of removed clusters.

Startup and stop logs		Kernel version management
<b>Log info</b>		
Time	Action	Details
[Redacted]	Cluster scali...	Before expansion: number of clusters is 1, cluster size is 16CU, after expansion: number of clusters is
[Redacted]	Cluster susp...	Cluster suspended
[Redacted]	Scaling out ...	Before expansion: number of clusters is 0, cluster size is 16CU, after expansion: number of clusters is
[Redacted]	Scaling out ...	Before expansion: number of clusters is 0, cluster size is 16CU, after expansion: number of clusters is
Total items: 4		



# Disaster Recovery Cluster

Last updated : 2024-07-31 17:47:09

To ensure the stable operation of the compute engine under extreme scenarios, DLC provides an efficient and agile disaster recovery cluster capability. When you need a disaster recovery cluster, you can quickly switch to it to ensure normal service operation. The disaster recovery cluster is only charged during operation, for more details, please see [Cost Description](#).

## Operation step

1. Enter the DLC Console, click Data Engine to access the Data Engine Page.
2. Click on the Data Engine Resource Name to enter the Data Engine Detail Page.

The screenshot shows the Tencent Cloud Data Lake Compute console interface. The main content area displays the 'SuperSQL engine' details for the 'Hong Kong' region. A table lists the engines, with the 'document\_test' engine highlighted by a red box. The table columns include Engine Name/ID, Engine type, Engine Status, Kernel version, Billing mode, Auto-renewal, Start and stop policy, and Operation. The 'document\_test' engine is of type SparkSQL, has a status of 'Suspend', and is billed on a 'Pay-as-you-go' basis. The 'Running' engine is of type SparkSQL, has a status of 'Running', and is billed on a 'Monthly subscription' basis.

Engine Name/ID	Engine type	Engine Status	Kernel version	Billing mode	Auto-renewal	Start and stop policy	Operation
document_test DataEngine-44ncio7nlf	SparkSQL	Suspend	SuperSQL-S 1.0	Pay-as-you-go	--	Auto-start, Manual suspensi	Monitor Spec configuration Parameter Configuration More
	SparkSQL	Running	SuperSQL-S 3.5	Monthly subscription 2024-08-02 11:37:06 Expire	No	Manual start, Manual suspension	Monitor Spec configuration Parameter Configuration More

3. Click **Enable Disaster Recovery Cluster** and wait for the disaster recovery cluster to initialize.

**Basic configuration** Cluster monitoring Alarm configuration

**Basic info**

Engine name `document_test` Resource ID `DataEngine-44nfc07n`

Description None

Region Hong Kong/Macao/TaiWan (China Region)-Hong Kong

Engine Status **Suspend**

Billing mode Pay-as-you-go

Tag No tag

Tags are used to categorize resources. To learn more, see [Tag Documentation](#)

**Configuration info** [Set start and stop policy](#) [Change spec configuration](#)

Engine type **SparkSQL** Kernel version **SuperSQL-S 1.0** Engine Size **16 CU** Cluster count **1**

Auto-scaling **Yes** Elastic cluster count **1** Max task concurrency of a cluster **5**

Task queue-up time limit **0 minute(s)**

Auto-start **Yes** Auto-suspension **No** Timing policy **None**

IP range of cluster **10.255.0.0/16**

Network configuration --

**Failover cluster**

**Not enabled** [Enable now](#)

4. After the disaster recovery cluster is enabled, in the disaster recovery cluster information, click **Switch to Disaster Recovery Cluster** to adjust the running cluster to the disaster recovery cluster. Subsequently, jobs directed to this data engine will be submitted to the disaster recovery cluster. The disaster recovery cluster serves as a transition during extreme failures of the data engine.

**Failover cluster**

Backup cluster name `document_test_backup` Backup resource ID `DataEngine-cof240j5`

Engine Status **Starting** [Switch to failover cluster](#)

Billing mode **Pay-as-you-go**

**Failover cluster configuration**

Engine type **SparkSQL** Kernel version **SuperSQL-S 1.0** Engine Size **16 CU** Cluster count **1**

Auto-scaling **Yes** Elastic cluster count **1** Max task concurrency of a cluster **5**

Task queue-up time limit **0 minute(s)**

Auto-start **Yes** Auto-suspension **No** Timing policy **None**

5. Once the extreme failure is resolved, in the basic information of the data engine, click **Switch to Primary Cluster**, and the disaster recovery cluster will be suspended. Subsequently, jobs directed to this data engine will be submitted to the primary cluster.

### Basic info

Engine name	document_test	Resource ID	DataEngine-44nfc7n
Description	None		
Region	Hong Kong/Macao/TaiWan (China Region)-Hong Kong		
Engine Status	Suspend	<a href="#">Switch to primary cluster</a>	
Billing mode	Pay-as-you-go		
Tag	No tag		

Tags are used to categorize resources. To learn more, see [Tag Documentation](#)

## Disaster Recovery Cluster Specifications

The disaster recovery cluster always tries to match the specifications of the data engine itself to ensure that the original tasks can transition and run normally. When AS is enabled on the data engine itself, the AS rules of the disaster recovery cluster will be consistent with the data engine. At the same time, to save costs, the disaster recovery cluster always operates on a pay-as-you-go basis.

## Note on Fees

There is no charge for enabling the disaster recovery cluster. When switching to the disaster recovery cluster and it is running, charges will be applied according to the pay-as-you-go rates for the same specifications as the data engine.

Example:

1. When the data engine itself is a 16 CU SparkSQL engine with an annual and monthly subscription. After enabling the disaster recovery cluster, it becomes a 16 CU SparkSQL engine on a pay-as-you-go basis, and there is no charge while the disaster recovery cluster is suspended. When users switch to the disaster recovery cluster and it is running, additional charges for the disaster recovery cluster's use of CU duration will apply. For specific fees, please refer to [Billing Overview](#).
2. When the data engine itself is a 16 CU SparkSQL engine on a pay-as-you-go basis. After enabling the disaster recovery cluster, it remains a 16 CU SparkSQL engine on a pay-as-you-go basis, and there is no charge while the disaster recovery cluster is suspended. When users switch to the disaster recovery cluster and it is running, with the primary cluster suspended, only the fees for the disaster recovery cluster's use of CU duration will be charged.

# Engine Kernel Version

Last updated : 2024-07-31 17:47:29

DLC provides different kernel versions optimized for various use cases, with numerous features and performance enhancements. The available kernel versions are listed below.

If your scenario primarily involves interactive queries, it is recommended to use the Presto engine and SparkSQL engine with the latest kernel versions.

If your scenario primarily involves batch jobs, it is recommended to use the Spark job engine with the Spark 3.2 kernel version.

Engine Type	Kernel Version	Description
Presto	SuperSQL-P 1.0	Based on the native Presto 0.242 version, this implementation supports dynamic data source loading, enhanced Dynamic Filter, Iceberg V2 tables, INSERT OVERWRITE for non-partitioned tables, and execution of Hive UDFs.
SparkSQL	SuperSQL-S 1.0	Based on the native Spark 3.2 version, this implementation supports Iceberg 1.1.0, Hudi 0.12.0, and Adaptive Shuffle Manager.
	SuperSQL-S 3.5	Based on the native Spark3.5 version, this implementation supports Iceberg 1.5.0 and Adaptive Shuffle Manager. The current beta version is backward compatible with various SQL and data governance tasks of SuperSQL-S 1.0, providing a performance improvement of more than 33% over the S1.0 version.
SparkBatch	Spark 3.5	Based on the native Spark3.5 version, this implementation supports Iceberg 1.5.0, Python3 and Adaptive Shuffle Manager. The current beta version is backward compatible with various SQL, jar, pyspark and data governance tasks of Spark 3.2, with a performance improvement of more than 33% over Spark 3.2.
	Spark 3.2	Based on the original Spark3.2 version, this implementation supports Iceberg 1.1.0, Hudi 0.12.0, Python3, and Adaptive Shuffle Manager.
	Spark 2.4	Based on the native Spark2.4 version, this implementation supports Iceberg 0.13.1, Python2, and Python3.

# Engine Network Configuration

Last updated : 2024-07-31 17:47:50

DLC supports configuring the network (VPC) for the data engine, facilitating the management of data engine access to different data source networks.

## Network Configuration Type

Based on different business scenarios, Data Lake Computing offers two types of network configurations.

**Enhanced Network Configuration:** Suitable for situations requiring high-speed, stable access to data within a single VPC.

### Caution

Data engines of non-Spark job types can only be bound to one Enhanced Network Configuration.

**Cross-origin Network Configuration:** Suitable for cross-origin federated data queries requiring access to multiple VPCs. A data engine can be bound to multiple Cross-origin Network Configurations.

## Network Configuration Status

**Initial:** The network configuration is being initialized, and the network is not yet effective.

**Success:** The network configuration is effective for the bound engine.

**Failure:** Network configuration failed, it can be deleted and reconfigured.

## Network Configuration Security Policies

If you have configured a Security Group Policy for the VPC, inbound rules need to be added for different types of network configurations.

**Enhanced Network:** In the Security Group, add inbound rules for the IP range of the VPC where the data source is located.

**Cross-origin Network:** In the Security Group, add inbound rules for the IP range where the network configuration's bound engine is located.

## Create Network Configuration

1. log in to [DLC console](#), select the service region.

2. Access **Engine Management > Engine Network Configuration** through the left navigation menu.
3. Click the **Create Network Configuration** button to enter the creation page.

**Create network configuration** ✕

The enhanced type is suitable for the scenario where a fast and stable VPC is required for data access. Only a set of enhanced network configuration can be bound to a data engine.  
 The cross-source type is suitable for cross-source federated data query across several VPCs. A data engine can be bound with several sets of cross-source network configurations.

Network configuration type \*  Enhanced  Cross-source

Configuration name \*

Instance source  Data Lake Compute-hosted catalog  New network configuration

Catalog \*

Data source VPC   🔄 0 IPs in total, 0 available

The data engine network will connect all subnets in the VPC. If existing networks do not meet your needs, you can [create a VPC](#) in the console.

Bound data engines \*

Configuration description

Configure parameters as follows:

Configuration	Required	Filling Instructions
Network Configuration Type	Yes	Select based on use case: Enhanced Network Configuration: Suitable for scenarios requiring high-speed, stable access to data within a single VPC Cross-origin Network Configuration: Suitable for scenarios involving cross-origin federated query analysis requiring access to data across multiple VPCs
Configuration Name	Yes	Supports Chinese, English, and _, with a maximum of 35 characters
Instance Source	Yes	Supports two sources: DLC data directory: You can select the data directory that has been created under DLC's Data Management New Network Configuration: Choose a new data source to create a network connection. Currently, supported data sources include MySQL, Kafka, EMR HDFS (COS, HDFS, Chdfs), PostgreSQL, SQLServer, and ClickHouse. If the data source required for the network configuration is not yet supported, select Other and manually specify the VPC
Data directory	Yes	Based on the selected instance source, choose the corresponding data directory. The range of available data directories will be related to your account

		permissions
Bind data engine	Yes	Select the data engine associated with this network configuration. If the data engine is in an isolated or initializing status, it cannot be selected
Configuration description	No	No more than 100 characters

4. Fill out and save to create a network configuration.

#### Caution

After creation, the network will be in an initialization state, and its status can be viewed in the list afterward.

## Delete network configuration

You can manage and delete network configurations that are no longer needed or have failed to configure by deleting them. The steps are as follows:

1. [DLC Console](#), select the service region.
2. Access **Engine Management > Engine Network Configuration** through the left navigation menu.
3. Find the network configuration you wish to delete. You can filter search results, but be sure to select the correct Network Configuration Type.
4. Click the **Delete** button. After a secondary confirmation, the deletion will be complete.

#### Caution

After deletion, the data engine will not be able to use this network configuration. If access is required, it must be reconfigured. Please proceed with caution.

## Modifying description information

You can modify the description of an existing network configuration by following these steps:

1. [DLC Console](#), select the service region.
2. Access **Engine Management > Engine Network Configuration** through the left navigation menu.
3. Find the network configuration you wish to delete. You can filter search results, but be sure to select the correct Network Configuration Type.
4. Click the **Modify description information** button to edit and modify.

# Associating Tag with Private Engine Resource

Last updated : 2024-07-17 18:06:24

## Overview

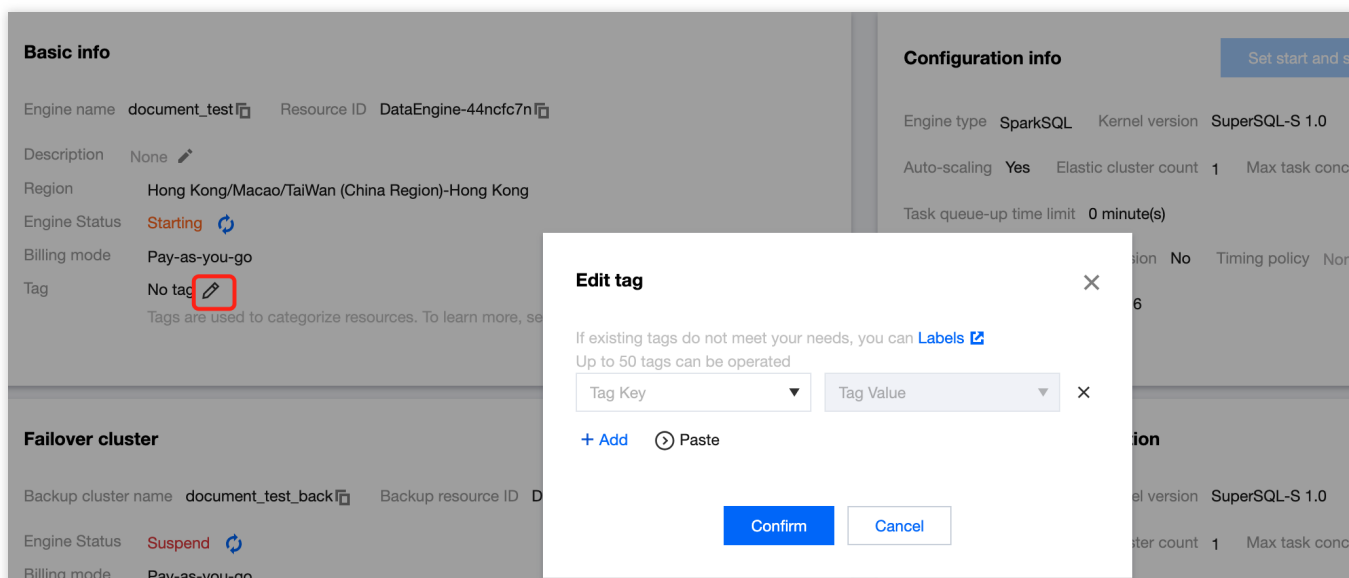
A tag is used to categorize and manage resources. It consists of a tag key and a tag value. A tag key can correspond to multiple values. You can create tags and bind them to cloud resources for easier management. Data Lake Compute supports binding tags to private engines in the console or on the purchase page, thereby enabling multidimensional category management and bill breakdown for private engine resources.

## Creating a Tag and Binding a Resource

Create a tag and bind it to a private engine for resource categorization and unified management.

### Directions

1. Log in to the [Tag console](#) to create a tag as instructed in [Creating Tags and Binding Resources](#).
2. Log in to the [Data Lake Compute console](#).
3. Click **SuperSQL Engine** on the left sidebar to enter the **Data engine list** page.
4. Click a resource name to enter the resource details page. Click **Edit** to pop up the tag edit window and select a tag for binding.



5. Click **Confirm** to bind the tag to the private engine. You can click **Edit** again to unbind or modify the tag.



**Basic info**

Engine name `at_data_engine_presto` Resource ID `DataEngine-p3d2xfq1`

Description `autotest_presto_engine`

Region `Hong Kong/Macao/TaiWan (China Region)-Hong Kong`

Engine Status Starting

Billing mode `Pay-as-you-go`

Tag `test:123`

Tags are used to categorize resources. To learn more, see [Tag Documentation](#)

**Configuration info** Set start and stop p...

Engine type `Presto` Kernel version `SuperSQL-P 1.0`

Cluster count `1`

Auto-scaling `Yes` Elastic cluster count `4` Max ta...

Task queue-up time limit `0 minute(s)`

Auto-start `Yes` Auto-suspension `No` Timing poli...

IP range of cluster `10.255.252.0/22`

Network configuration `--`

## Binding a Tag on the Purchase Page

You can bind a tag when purchasing a private engine resource in both monthly subscription and pay-as-you-go billing modes.

**Info configuration**

Resource name   
It can contain up to 100 Chinese characters, letters, digits, hyphens (-) and underscores (\_) only. A duplicate name is not allowed.

Description   
Optional, up to 250 characters.

Tag

Tag Key	Tag Value	<a href="#">Delete</a>
<span style="border: 1px dashed #ccc; display: inline-block; width: 100%; height: 20px;"></span> <span style="color: blue; font-weight: bold; margin-left: 100px;">+ Add</span>		
<span>📄 Paste</span>		
<a href="#">OK</a>	<a href="#">Cancel</a>	

Tags are used to categorize resources. To learn more, see [Tag Documentation](#)

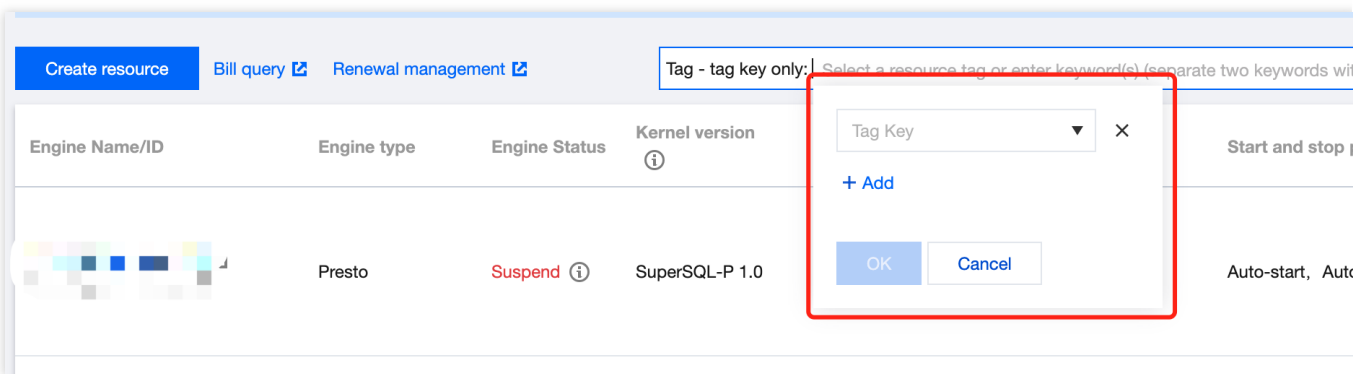
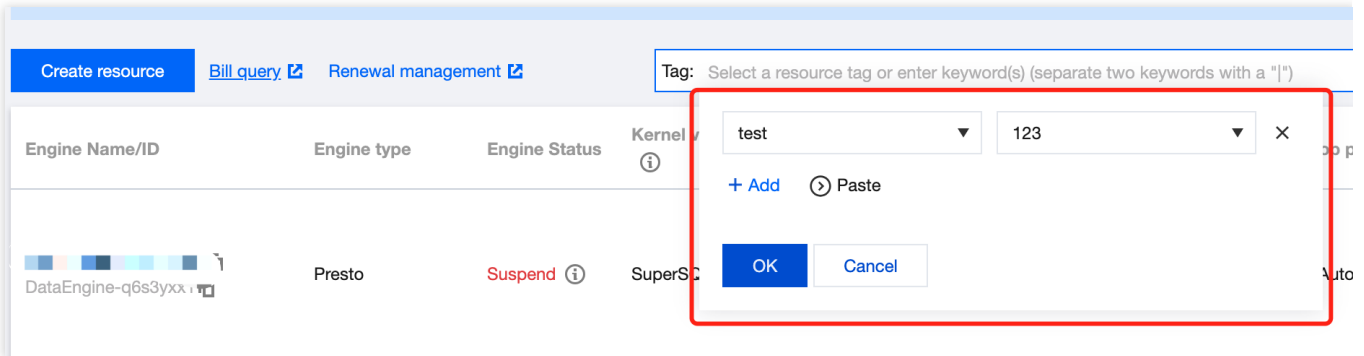
Terms of agreement  I have read and agree to the [Service Level Agreement for Data Lake Compute](#) and [Refund Policy](#)

# Filtering Resources by Tag

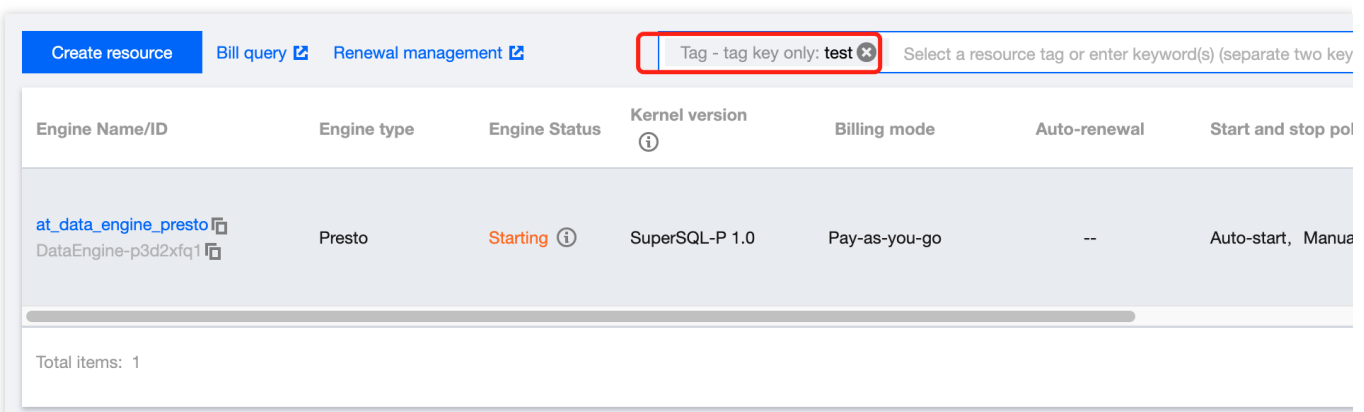
You can filter resources by tag on the **SuperSQL Engine** page in the Data Lake Compute console.

## Directions

1. Log in to the Data Lake Compute console and select **SuperSQL Engine**.
2. Select a tag in the tag search box. You can filter resources by tag key or tag key-value.



3. Click the search icon to get the list of engines with that tag.



## Allocating Costs by Tag

You can bind tags in the organization or business dimension for cost allocation by department, project team, region, etc.

### Directions

1. Log in to the [Tag console](#) and create a tag.
2. Bind the tag to an engine resource in the tag console, on the **SuperSQL Engine** page in the Data Lake Compute console, or on the purchase page.
3. Go to the [Billing Center](#) to set a cost allocation tag. For more information, see [Cost Allocation Tags](#).
4. Go to the [Bill Overview](#) page, select the aggregation by tag tab, and view the column chart and list of resources aggregated by tag key.

# Engine Local Cache

Last updated : 2024-07-31 17:48:05

To ensure stable operation of Spark engine query analysis when network bandwidth is limited (e.g. during storage system throttling), the DLC Spark engine provides a local cache capability. When you need to cache table data, you can quickly enable caching by adding engine configuration.

## Directions

1. Create a Spark Engine: For details, see [Purchase Exclusive Data Engine](#).
2. Add Cache Configuration: Go to the [DLC Console > Data Engine](#). Select the engine created in Step 1, click **Parameter Configuration**, and add the configuration items from [Cache Configuration Item Explanation](#).

### Spark SQL Engine Configuration:

The screenshot shows the Tencent Cloud console interface for managing SuperSQL engines. The left pane displays a table of engine instances, and the right pane shows the configuration change interface for a selected instance.

资源名称/ID	引擎类型	内核版本	运行状态	付费类型
givy	SparkSQL	SuperSQL-S 1.0	运行	按量计费
qb901960	SparkSQL	SuperSQL-S 1.0	运行	按量计费
.8km	Presto	SuperSQL-P 0.1	运行	按量计费
	Spark作业	--	运行	--
	SparkSQL	--	运行	--

The right pane shows the configuration change interface for a selected instance. It includes a warning message: "修改引擎参数配置将需要重启集群." (Modifying engine parameter configuration will require restarting the cluster). There is a toggle for "数据加密" (Data Encryption) which is currently turned off. Below, there is a "参数配置" (Parameter Configuration) section with a table for adding configurations:

序号	配置项	配置值
1	spark.hadoop.fs.cosn.impl	alluxio.hadoop.ShimFileS

A "+ 添加" (Add) button is located below the configuration table.

### Note:

After the configuration is added, the engine cluster will restart. It is recommended to enable the cache when no tasks are running to avoid affecting ongoing tasks.

3. To use the engine cache, go to Data Exploration, write the query SQL in the SQL interface, select the engine with the cache enabled, and execute the SQL. Once executed, the engine will cache the DLC external table data locally. When the SQL is executed again, the data will be fetched from the local cache, improving query efficiency.

### Spark SQL Engine Query:

⏪ 📄 ↻ 🗑️ ⏪ ⏩ ⏏️ ☰
请选择默认数据库 ▾

```

1 select test1.id,test1.name,test2.age from DataLakeCatalog.test_cry.h_test1 test1
2 left join DataLakeCatalog.test_cry.h_test2 test2 on test1.id = test2.id
            
```

查询结果

统计数据

[Task ID](#) [SQL详情](#) [导出结果](#) [优化建议](#)

查询耗时 10.69s

### Spark Batch Engine Query:

⏪ 📄 ↻ 🗑️ ⏪ ⏩ ⏏️ ☰

```

1 set spark.hadoop.fs.cosn.impl=alluxio.hadoop.ShimFileSystem;
2 select test1.id,test1.name,test2.age from DataLakeCatalog.test_cry.h_test1 test1
3 left join DataLakeCatalog.test_cry.h_test2 test2 on test1.id = test2.id
4
5
6
            
```

查询结果

TaskID: fdd1f66b-10f6-402a-b5a2-8e7af8c618c0 [🗑️](#)

[点击查看集群日志](#)

ExecuteSQL: select test1.id,test1.name,test2.age from DataLakeCatalog.test\_cry.h\_test1 test1 left join DataLakeCatalog.test\_cry.h\_test2 test2

2023-11-28 15:05:27 当前任务状态: available... 请等待...

2023-11-28 15:05:27 当前任务运行成功, [点击查看运行结果](#)

2023-11-28 15:05:29 任务运行结束

Task ID	SQL	开始时间	运行时长 <span style="font-size: 0.7em;">ⓘ</span>
1 fdd1f66b-10f6-402a-b5a2-8e7af8c61... <a href="#">🗑️</a>	select test1.id,test1.name,test2.age from ...	2023-11-28 15:05:04	20.00s

## Cache Description

### Cache Configuration Items Description

Configuration Items	Configuration Values	Configuration Items Description

spark.hadoop.fs.cosn.impl	alluxio.hadoop.ShimFileSystem	Fixed value; the configuration value is the cache implementation class. Configure this value to enable the cache feature. If the cache feature is enabled, configuring a value other than this will result in the engine not being able to access COS data. Please follow the instructions carefully. If you need to disable the cache after enabling it, please delete this configuration item.
---------------------------	-------------------------------	---

## Cache Usage Instructions

### 1. Engine Type Description

SparkSQL Engine: When the engine restarts, the cached data becomes invalid because it is a local cache.

SparkBatch Engine: The SparkBatch engine runs tasks at the session level. Once the task execution is complete, the cached data becomes invalid.

### 2. Table Type Description

Currently, only DLC external tables are cached.

# Custom Task Scheduling Pool

Last updated : 2024-07-31 17:48:18

## Application scenario

Applicable Engine: Spark SQL Engine.

When you submit multiple tasks to the engine, for example, submitting multiple SQL tasks to the Spark SQL cluster simultaneously, the tasks submitted by the business may have dependencies, so the engine will default to scheduling these tasks in a FIFO manner when scheduling and executing.

However, in some special cases, you may need to define the priorities of certain tasks yourself, for example in the following scenario:

The submitted task has a high priority and needs to be executed with the highest priority, not wanting it to queue for cluster resources.

The submitted task has a low priority, hoping that it will not preempt resources from other tasks as much as possible. It will be executed when resources are available, and it will queue when resources are not.

## Customize Scheduling Rules

In the Spark SQL Engine, each executed SQL task Job is split into a collection of multiple tasks, TaskSet, and our scheduling is based on TaskSet. Whenever the cluster has idle resources, it takes a Task from all Job's TaskSet according to the scheduling algorithm for dispatch execution.

Our scheduling algorithm is to define multiple scheduling pools, placing Job/TaskSet in the corresponding scheduling pool, and obtaining the Task that needs to be dispatched for execution according to the scheduling pool.

### Scheduling Pool and Its Attributes

You can define multiple scheduling pools, each with four attributes:

**name:** The name of the scheduling pool, which you can name yourself. It can be named default, indicating the default scheduling pool.

**schedulingMode:** The scheduling rule, supporting two modes: FIFO and FAIR. The scheduling algorithm when there are multiple TaskSets within a scheduling pool.

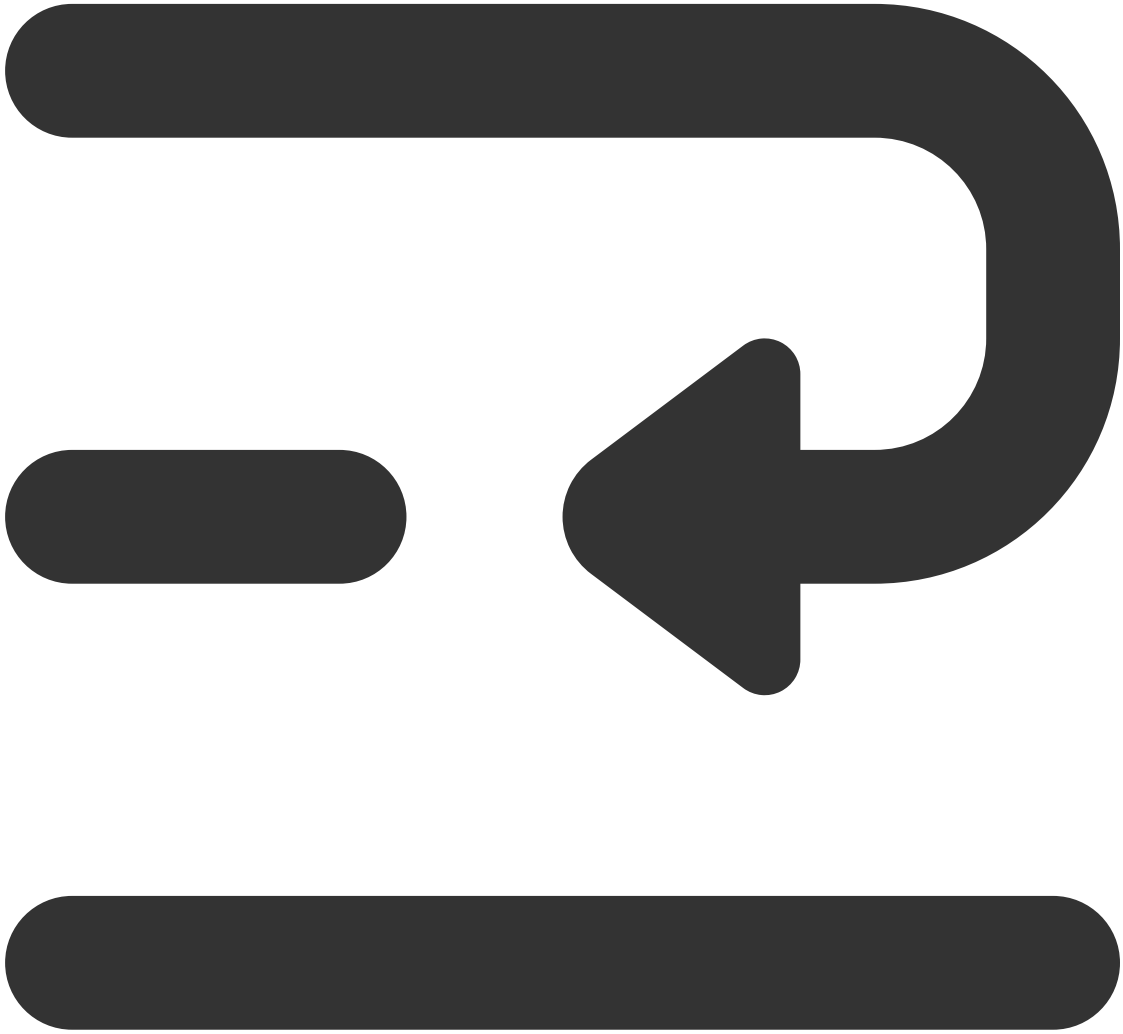
**FIFO:** Tasks are dispatched in the order that TaskSets are submitted.

**FAIR:** Tasks from multiple TaskSets are dispatched fairly. The specific dispatch rules are related to the minShare and weight attributes of the scheduling pool.

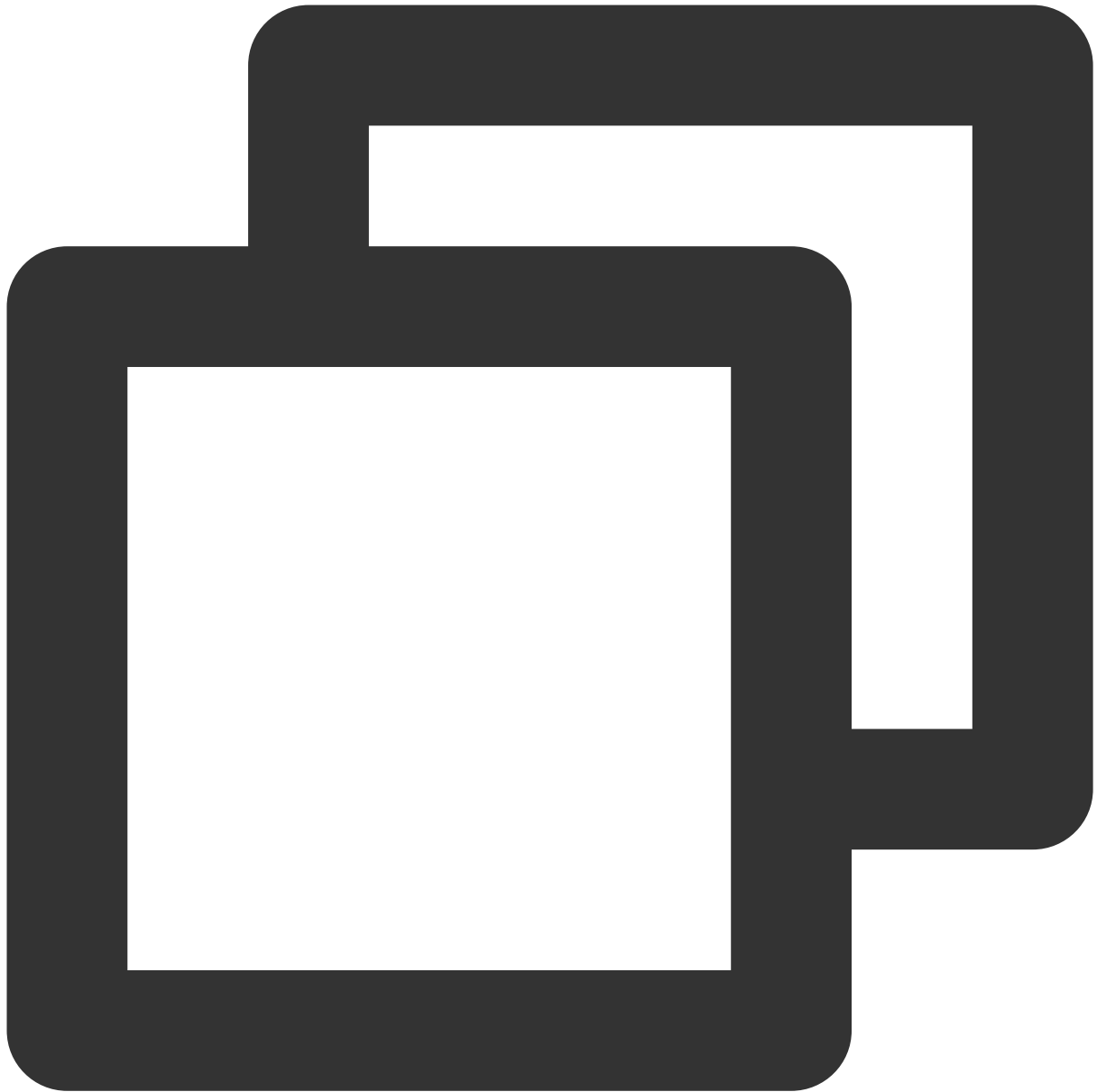
**minShare:** The minimum number of cores required, must be greater than 0, that is, the minimum number of Tasks that can run. During scheduling, priority is given to the number of Tasks running in the scheduling pool reaching minShare.

weight: The weight. Scheduling pools with a higher weight will have their Tasks prioritized. Weight comparison will only occur after minShare is met.

The scheduling configuration requires you to write an xml file, in the following formats:







```
<?xml version="1.0"?>
<allocations>
  <pool name="production">
    <schedulingMode>FAIR</schedulingMode>
    <weight>1</weight>
    <minShare>2</minShare>
  </pool>
  <pool name="test">
    <schedulingMode>FIFO</schedulingMode>
    <weight>2</weight>
    <minShare>3</minShare>
  </pool>
</allocations>
```

```
</pool>  
</allocations>
```

## Scheduling Configuration Reference Example

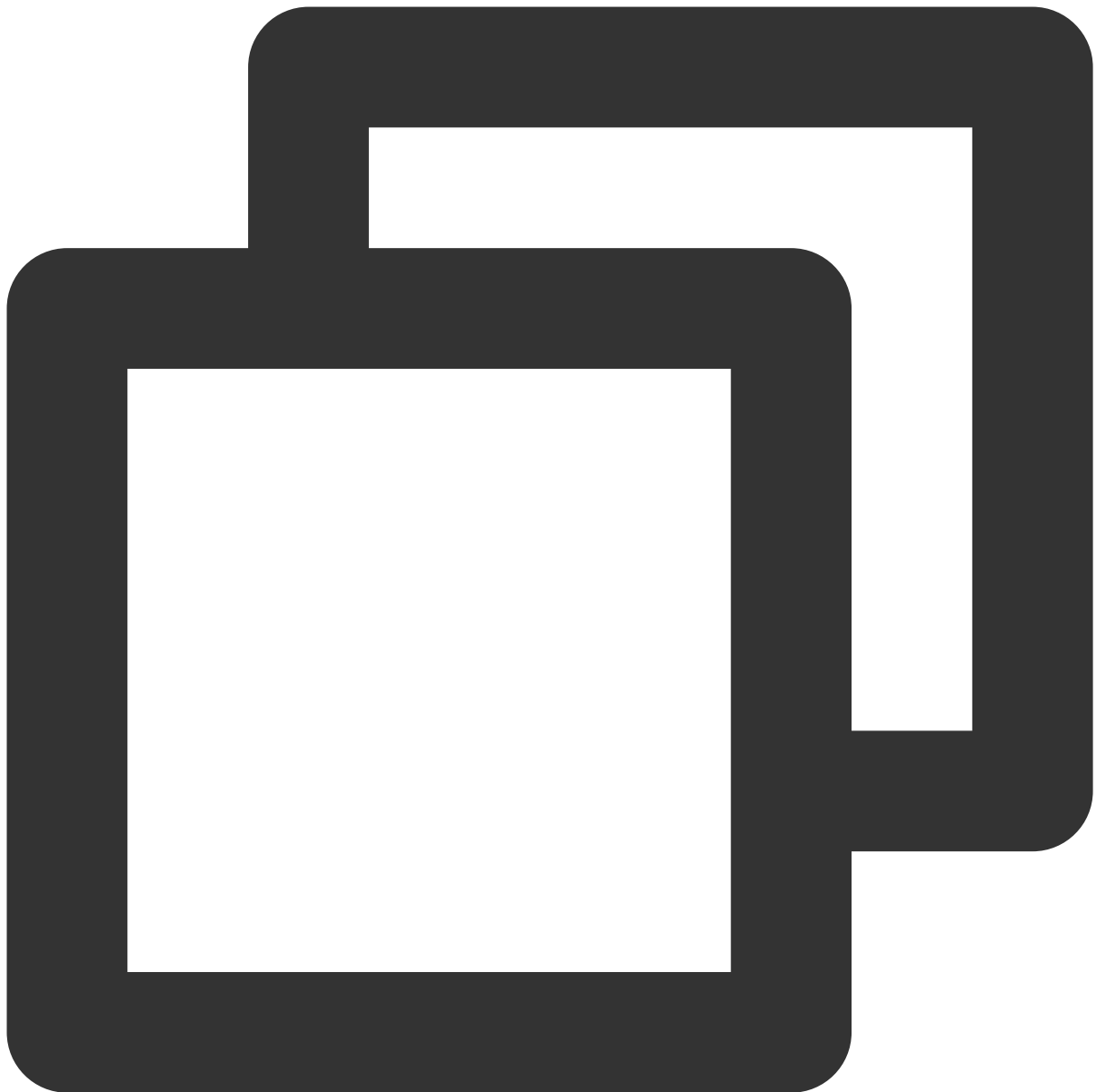
You can refer to the settings for three scheduling pools:

Default Scheduling Pool `default:schedulingMode = FIFO`, `weight = 1`, `minShare = (Cluster Cores - Driver Cores)`. This scheduling pool is the default submission pool for tasks, with ordinary priority. Execution is in sequential order, and it can utilize all of the cluster's computing resources.

Slow Task Scheduling Pool `straggler:schedulingMode = FAIR`, `weight = 1`, `minShare = 1`. This scheduling pool is dedicated to slow task submissions, with ordinary priority. Since `minShare = 1`, it does not preempt resources from tasks submitted to the default pool. Tasks in the straggler scheduling pool are executed when the cluster has more available resources.

High Priority Scheduling Pool `special:schedulingMode = FIFO`, `weight = 1000`, `minShare = (Cluster Cores - Driver Cores)`. This scheduling pool is for tasks that need priority execution in special circumstances. However, due to the presence of `minShare`, this pool does not monopolize all cluster resources. Tasks in both the default and special pools continue to be executed, typically dispatching an equal number of Tasks from each pool.

Taking a 16CU cluster (with the driver being 4CU) as an example, the configuration for this reference example is as follows:



```
<?xml version="1.0"?>
<allocations>
  <pool name="default">
    <schedulingMode>FIFO</schedulingMode>
    <weight>1</weight>
    <minShare>12</minShare>
  </pool>
  <pool name="straggler">
    <schedulingMode>FAIR</schedulingMode>
    <weight>1</weight>
    <minShare>1</minShare>
  </pool>
</allocations>
```

```

</pool>
<pool name="special">
  <schedulingMode>FIFO</schedulingMode>
  <weight>1000</weight>
  <minShare>12</minShare>
</pool>
</allocations>
    
```

## Operation method

1. After preparing the xml file for the scheduling pool, place it in a path on cos, for example cosn://bucket-appid/fairscheduler.xml.
2. Add the following configuration in the engine settings.

The screenshot shows the 'SuperSQL engine' management page in the Tencent Cloud console. The left sidebar contains navigation options like 'Overview', 'Data Explore', 'Data Scheduling', and 'SuperSQL Engine'. The main area displays a table of engine instances with columns for Engine Name/ID, Engine type, Engine Status, Kernel version, Billing mode, Auto-renewal, Start and stop policy, and Operation. The 'Operation' column for the Presto engine instance has a red box around the 'Parameter Configuration' link.

Engine Name/ID	Engine type	Engine Status	Kernel version	Billing mode	Auto-renewal	Start and stop policy	Operation
DataEngine-ksyfgonl	Spark job	Starting	Spark 3.2	Pay-as-you-go	--	Manual start, Manual suspension	Monitor Spec config Parameter Configura More
[Blurred]	Presto	Suspend	SuperSQL-P 1.0	Pay-as-you-go	--	Auto-start, Auto-suspens	Monitor Presto UI Spec configuration Parameter Configura More
[Blurred]	Spark job	Running	Spark 3.2	Pay-as-you-go	--	Auto-start, Manual suspe	Monitor Spec config Parameter Configura More
[Blurred]	SparkSQL	Suspend	SuperSQL-S 1.0	Pay-as-you-go	--	Auto-start, Auto-suspens	Monitor Spec config Parameter Configura More
[Blurred]	SparkSQL	Running	SuperSQL-S 3.5	Monthly subscription 2024-08-02 11:37:06 Expire	No	Manual start, Manual suspension	Monitor Spec config Parameter Configura More

Parameter configuration spark.scheduler.allocation.file, set to the path of your scheduling pool xml file cosn://bucket-appid/fairscheduler.xml.

### Configuration change

**!** If engine parameter configurations are changed, you must restart the cluster to apply the new configurations.

Data encryption i

Parameter Configuration

1	sqark.scheduler.allocation.file	cc ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ 424723/fa	-
---	---------------------------------	----------------------------------	---

[+ Add](#)

This operation requires restarting the cluster.

3. When submitting a task, specify the following parameters as task parameters: spark.scheduler.pool = the name of the scheduling pool to submit to. If it is the default scheduling pool, it does not need to be specified.

The screenshot shows the 'Data engine' configuration page. The 'Advanced settings' section is highlighted with a red box and contains the following configuration items:

Advanced settings <span style="float: right;">Configuration description <a href="#">?</a></span>	
1	<div style="display: flex; justify-content: space-between; align-items: center;"> <div style="flex-grow: 1;"> <input type="text" value="sqark.scheduler.allocation.file"/> </div> <div style="flex-grow: 1;"> <input type="text" value="cc ■ ■ ■ ■ ■ ■ ■ ■ ■ ■ 424723/fa"/> </div> <div style="text-align: right;">-</div> </div>

[+ Select configuration.](#) [More](#) v

## Notes

Scheduling occurs at the time node when: the cluster has idle resources and there is a task that needs scheduling. Therefore, if the cluster is already fully occupied by a task, for example, a slow task, it must wait for one Task of that task to be completed before beginning to schedule other tasks with higher priority. Therefore, it is important to note

that the time consumption of a single Task of a slow task should be relatively reasonable; otherwise, it might still lead to long periods of occupying cluster resources.

# Ops Management

## Permission Management

### CAM Service

Last updated : 2024-07-17 15:29:49

Data Lake Compute has a complete data access control mechanism and divides permissions into operation permissions and data permissions. The former is managed by CAM, while the latter is managed by the permission module of Data Lake Compute.

- A root account has all the operation and data permissions of Data Lake Compute by default.
- If a sub-user is granted the operation permissions of Data Lake Compute, the sub-user can grant the data permissions to other sub-users and can be regarded as an "admin" of this type of sub-users.
- If a sub-user is granted the data read/write permissions, the sub-user can query data as permitted. The data permissions are granted by an "admin".
- The data permissions of all sub-users other than root accounts are granted by an "admin". They cannot query data which they don't have permissions on.

A root account has all the operation permissions of Data Lake Compute by default and can grant sub-users the access permissions of Data Lake Compute through CAM, so that the sub-users can have corresponding operation permissions of Data Lake Compute.

## Directions

1. Create and authorize a sub-user.

In the CAM console, create a sub-user and grant permissions as instructed in [Sub-user authorization](#).

- Preset policy `QcloudDLCFullAccess` : All the operation permissions in Data Lake Compute.
- Custom policy: Specified operation permissions of Data Lake Compute.

2. Log in to the Data Lake Compute console with a sub-user account and verify the permissions.

If the operation succeeds, the authorization has taken effect.

## Operation permission category

Data Lake Compute operation permissions are categorized by API as follows.

Permission Type	Description
Metadata management	Manipulate the metadata information of databases and data tables managed in Data Lake Compute.
Task management	Submit and view tasks in Data Lake Compute.
Permission management	Manage users' data access permissions.
System configuration	Perform basic configurations of the Data Lake Compute service.

## Sub-user authorization

If you access Data Lake Compute as a root account, skip this step.

1. Create a sub-account as instructed in [Creating and Authorizing Sub-account](#).
2. Create a custom policy.
  - On the [Policies](#) page in the CAM console, click **Create Custom Policy**.
  - In the pop-up window, click **Create by Policy Syntax**.
  - On the **Create by Policy Syntax** page, select **Blank Template** and click **Next**.
  - In the template, enter the **Policy Name** (e.g., `DLCDataAccess`) and **Description**, copy the following policy, paste it into **Policy Content**, and click **Complete**. A sub-user bound to the custom policy can log in to the Data Lake Compute console to run SQL tasks but cannot manage data permissions. For more information, see [Sub-Account Permission Management](#).

```
{
  "version": "2.0",
  "statement": [
    {
      "effect": "allow",
      "action": [
        "dlc:DescribeStoreLocation",
        "dlc:DescribeTable",
        "dlc:DescribeViews",
        "dlc:CancelTask",
        "dlc:CreateDatabase",
        "dlc:CreateScript",
        "dlc:CreateTable",
        "dlc:CreateTask",
        "dlc>DeleteScript",
        "dlc:DescribeDatabases",

```



```
"dlc:DescribeScripts",
"dlc:DescribeTables",
"dlc:DescribeTasks",
"dlc:DescribeQueue"
],
"resource": [
  "*"
]
}
```

5. Bind the preset or custom policy to a sub-account, and the sub-account can log in to and access Data Lake Compute. For more information, see [Setting Sub-user Permissions](#).

- Preset policy: `QcloudDLCFullAccess` .
- Custom policy: The policy customized in the above steps for accessing Data Lake Compute.

# Permission Overview

Last updated : 2024-07-17 15:42:58

Data Lake Compute permissions include data permissions and data engine permissions. If you have the admin permission, you can log in to the Data Lake Compute console or use an API to grant a sub-user data and data engine permissions. Sub-users cannot use, modify, or delete data or data engines before they are authorized.

## User and work group

Data Lake Compute provides the user mode and work group mode for personnel permission management.

User: You can select users in CAM, including sub-accounts and collaborator accounts.

Work group: It is a group of users with the same permissions managed in the product.

### Note:

If users are granted different permissions from those granted in their work groups, all the granted permissions will take effect.

A work group allows you to quickly grant permissions to a batch of users, so it is recommended for batch user authorization. For detailed directions, see [User and User Group](#).

## User type

In Data Lake Compute, **User type** can be **Admin** or **General user**.

Admin: An admin have all the data, engine, and task permissions and can add, authorize, and remove users and work groups in Data Lake Compute.

General user: A general user is added by an admin, has no Data Lake Compute permissions by default, and needs to be authorized. Only data and engine permissions that can be **regranted** can be granted to general users.

Permission and Operation	Admin	General User
Data permissions	All	None by default (to be authorized by an admin)
Data engine permissions	All	None by default (to be authorized by an admin)
User management	Yes	No
Work group management	Yes	No
Authorization scope	All	Permissions that <b>can be regranted</b>

**Note:**

The above permissions only include those defined in Data Lake Compute. To perform purchase, configuration adjustment, and refund operations that involve billing, log in to the CAM console and get the financial collaborator permission `QCloudFinanceFullAccess` (for detailed directions, see [Creating and Authorizing Sub-account](#)).

## Data permissions

Data Lake Compute data permissions allow operations on data catalogs, databases, and data tables. To facilitate your management and configuration, permissions can be granted in the standard or advanced mode.

In standard mode, you can grant roles while ignoring the specific permission configuration (for more information on roles and permissions, see [Sub-Account Permission Management](#)). The authorization granularity can be data catalog, database, or data table. This mode is suitable for quick authorization with no complex permission management involved.

In advanced mode, you can grant permissions at the database, data table, view, or function level. It is suitable for refined permission management.

SQL statements for permission operations are as follows:

Action	CREATE	ALTER	DROP	SELECT	INSERT	DELETE	Target
CREATE DATABASE	✓	-	-	-	-	-	Cataglog
ALTER DATABASE	-	✓	-	-	-	-	Database
DROP DATABASE	-	-	✓	-	-	-	Database
CREATE TABLE	✓	-	-	-	-	-	Database
CREATE TABLE AS SELECT	✓	-	-	✓	✓	-	Database/Table
DROP TABLE	-	-	✓	-	-	-	Table
ALTER TABLE LOCATION	-	✓	-	-	-	-	Table
ALTER PARTITION LOCATION	-	✓	-	-	-	-	Table

ALTER TABLE ADD PARTITION	-	✓	-	-	-	-	Table
ALTER TABLE DROP PARTITION	-	✓	-	-	-	-	Table
ALTER TABLE	-	✓	-	-	-	-	Table
CREATE VIEW	✓	-	-	-	-	-	Database
ALTER VIEW PROPERTIES	-	✓	-	-	-	-	View
ALTER VIEW RENAME	-	✓	-	-	-	-	View
DROP VIEW PROPERTIES	-	✓	✓	-	-	-	View
DROP VIEW	-	-	✓	-	-	-	View
SELECT TABLE	-	-	-	✓	-	-	Table
INSERT	-	-	-	-	✓	-	Table
INSERT OVERWRITE	-	-	-	-	✓	✓	Table
CREATE FUNCTION	✓	-	-	-	-	-	Database
DROP FUNCTION	-	-	✓	-	-	-	Function
SELECT VIEW	-	-	-	✓	-	-	View
SELECT FUNCTION	-	-	-	✓	-	-	Function

## Data engine permissions

Data Lake Compute data engine permissions allow using, modifying, manipulating, monitoring, and deleting data engines as detailed below:

Use: The permission to use engines to perform tasks.

Modify: The permission to modify the basic information and configuration information of engines (modifying the configuration information requires the CAM financial collaborator permission).

Manipulate: The permission to suspend and restart engines.

Monitor: The permission to view the running tasks and monitoring information of engines.

Delete: The permission to return engines.

## Permission granting

A single user can be granted multiple permissions. For detailed directions, see [Sub-Account Permission Management](#).

# User and Work Group

Last updated : 2024-07-17 15:44:57

Data Lake Compute provides the user mode and work group mode for personnel permission management. For more information on permissions, see [Permission Overview](#).

## Description

**User:** You can select users in CAM, including sub-accounts and collaborator accounts.

**Work group:** It is a group of users with the same permissions managed in the product.

### Note:

If users are granted different permissions from those granted in their work groups, all the granted permissions will take effect.

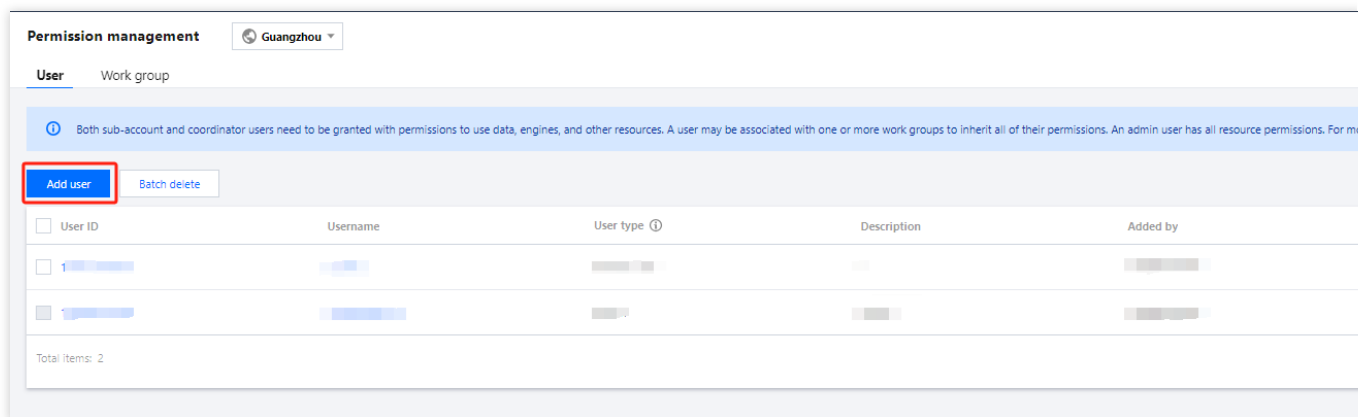
A work group allows you to quickly grant permissions to a batch of users, so it is recommended for batch user authorization.

## User Management

User management requires Data Lake Compute operation permissions. For more information, see [CAM Service](#).

### Adding a user

1. Log in to the [Data Lake Compute console](#), select the service region, and go to the **Permission management** page.
2. Click **Add user** to add an account with a specified user ID to Data Lake Compute for management.



3. After entering the **User ID**, bind the user to a work group (which requires the admin permission). If binding is not needed, directly click **Complete**.

## Viewing user information

A Data Lake Compute admin can modify the basic information and permissions of a user.

1. Log in to the [Data Lake Compute console](#), select the service region, and go to the **Permission management** page.
2. Search for the target **User ID** and click the **Username** to view the user information and permissions.

## Editing user information

You can edit the description and work group of a user. For detailed directions, see [Sub-Account Data Authorization](#).

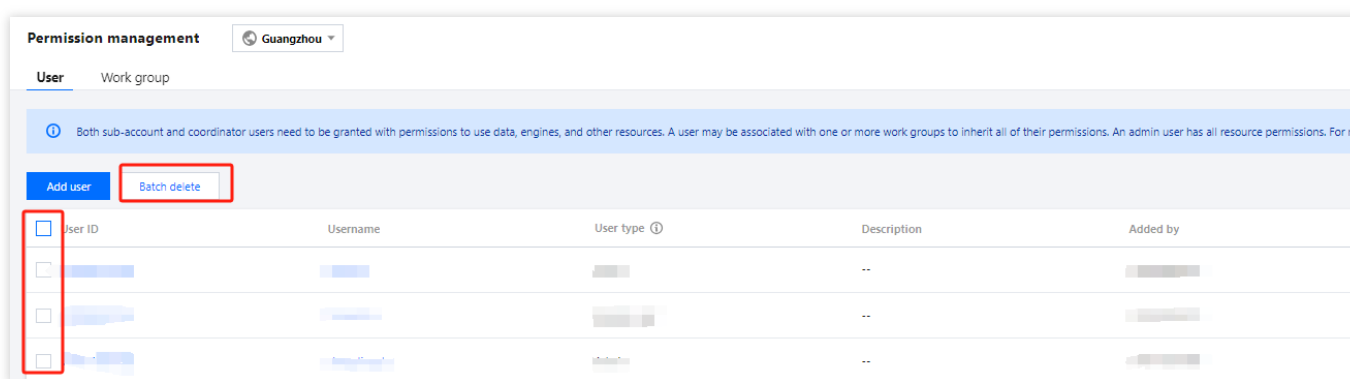
1. Log in to the [Data Lake Compute console](#), select the service region, and go to the **Permission management** page.

2. Search for the target user account ID and click **Edit** in the **Operation** column to enter the edit page.

## Removing a user

If you don't want a user to use Data Lake Compute any more, you can use an admin account to remove the user. Then, the Data Lake Compute permission granted to the user will be revoked.

1. Log in to the [Data Lake Compute console](#), select the service region, and go to the **Permission management** page.
2. Search for and select one or multiple target user account IDs and click **Batch remove** to remove them from Data Lake Compute.



## Work Group Management

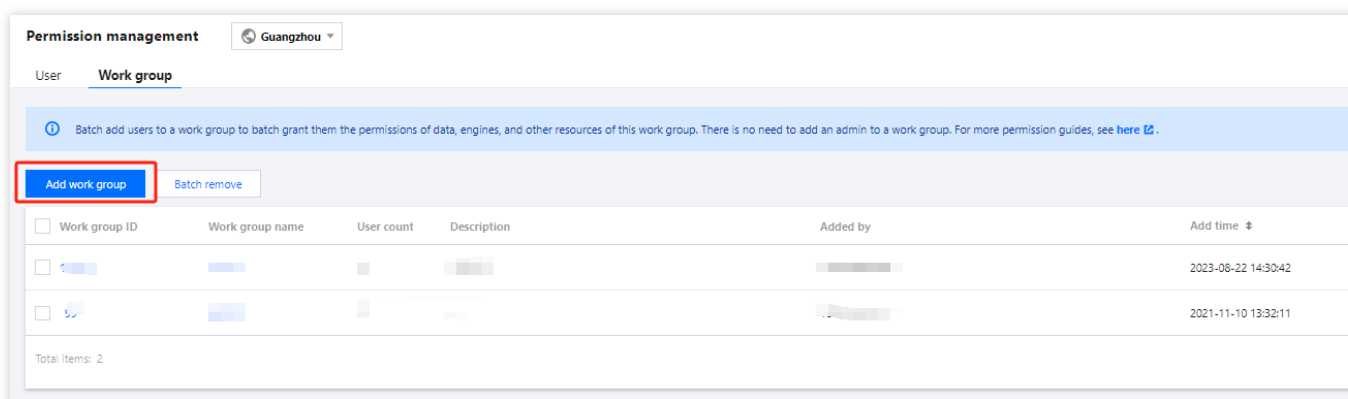
Work group management requires Data Lake Compute operation permissions. For more information, see [CAM Service](#).

### Adding a work group

You can manage permissions that need to be repeatedly granted to users through a work group. The following describes how to add a work group.

1. Log in to the [Data Lake Compute console](#), select the service region, and go to the **Permission management** page.
2. Click **Work group** to enter the work group management page.
3. Click **Add work group**, enter relevant information, and click **Confirm**.

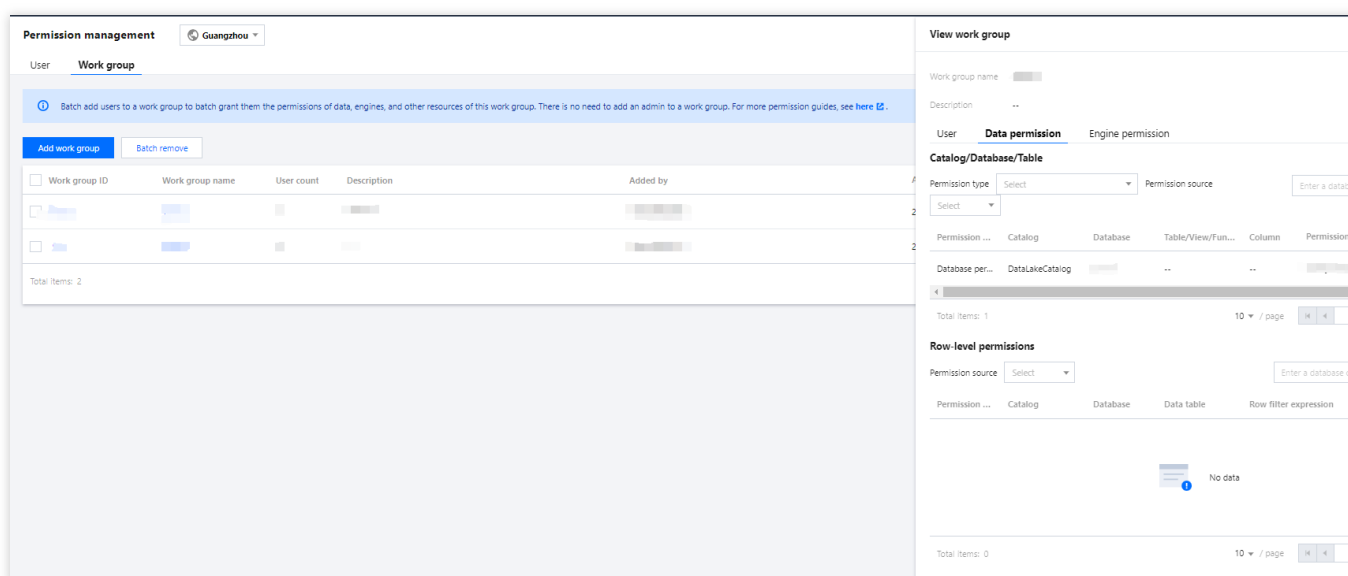




## Viewing work group information

You can view the information of a work group in the following steps:

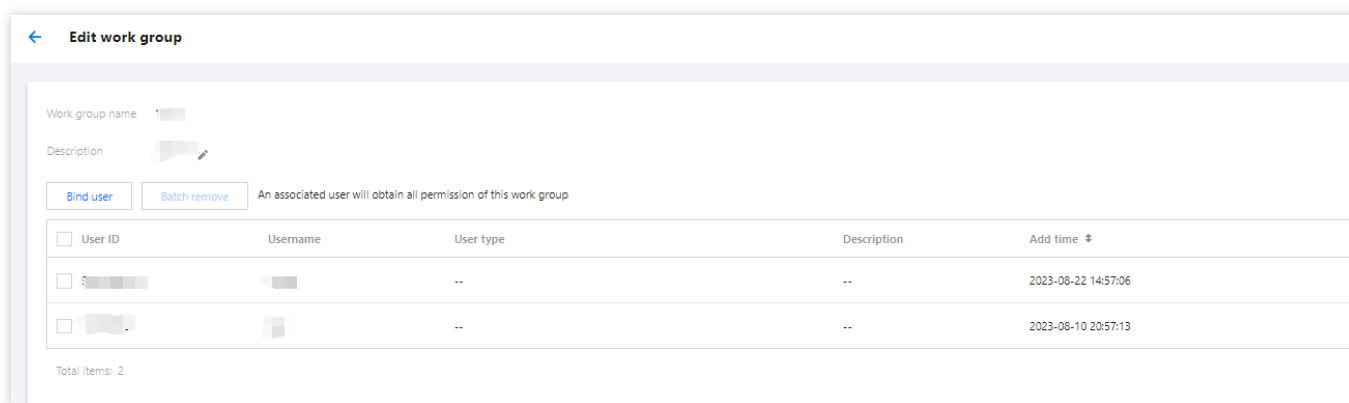
1. Log in to the [Data Lake Compute console](#), select the service region, and go to the **Permission management** page.
2. Click **Work group** to enter the work group management page.
3. Search for the target work group and click **Work group ID** or **Work group name** to view the work group information.



## Editing work group information

You can modify the description and users of a work group in the following steps:

1. Log in to the [Data Lake Compute console](#), select the service region, and go to the **Permission management** page.
2. Click **Work group** to enter the work group management page.
3. Find the target **Work group name** and click **Edit** in the **Operation** column.



To edit the description, click



You can click **Bind user** to add Data Lake Compute users to the work group.

Select multiple target users and click **Batch remove**, or click **Remove** in the **Operation** column of a specific target user. Removed users will no longer have the permissions of the work group, which does not affect other permissions granted to them though.

## Deleting a work group

A Data Lake Compute admin can remove work groups.

### Note:

After a work group is removed, all its permissions granted to users in it will be revoked. Note that a removed work group cannot be recovered. Proceed with caution.

1. Log in to the [Data Lake Compute console](#), select the service region, and go to the **Permission management** page.
2. Click **Work group** to enter the work group management page.
3. Select multiple target work groups and click **Batch remove**, or click **Remove** in the **Operation** column of a specific target work group.

**Permission management** Guangzhou

User **Work group**

Batch add users to a work group to batch grant them the permissions of data, engines, and other resources of this work group. There is no need to add an admin to a work group. For more permission guides, see [here](#).

Add work group Batch remove

<input type="checkbox"/>	Work group ID	Work group name	User count	Description	Added by	Add time #
<input type="checkbox"/>	[blurred]	[blurred]	[blurred]	[blurred]	[blurred]	2023-08-22 14:30:42
<input type="checkbox"/>	[blurred]	[blurred]	[blurred]	[blurred]	[blurred]	2021-11-10 13:32:11

Total items: 2

# Sub-Account Permission Management

Last updated : 2024-07-17 15:46:12

## User permission

User permissions include data permissions and engine permissions (for more information on permissions, see [Permission Overview](#)). The former is required to access data in Data Lake Compute, while the latter is used for resource management. Data Lake Compute enables permission management at the database, table, and column levels, so that you can authorize a user or work group for refined data permission management in different use cases.

## User and work group

You can authorize a user or create and authorize a work group of users. For detailed directions, see [User and Work Group](#).

**User:** You can select users in CAM, including sub-accounts and collaborator accounts.

**Work group:** It is a group of users with the same permissions managed in the product.

### **Note:**

If users are granted different permissions from those granted in their work groups, all the granted permissions will take effect.

A work group allows you to quickly grant permissions to a batch of users, so it is recommended for batch user authorization.

## Granting a user a permission

Grant permissions to the specified user.

1. Set a user to **Admin** or **General user**. Admins have the permissions of all the data and engines by default with no need to be bound to a work group. They can also manage admin users other than the root account. **Set an admin with caution.**

**Add user**

1 Basic info > 2 Bind work group

User ID:

Username:

User type:

An admin has all permissions for all resources (including data and engines), and can manage other admins except the root account user. A general user needs to be granted with relevant permissions or associated with a work group to access corresponding resources.

Description:

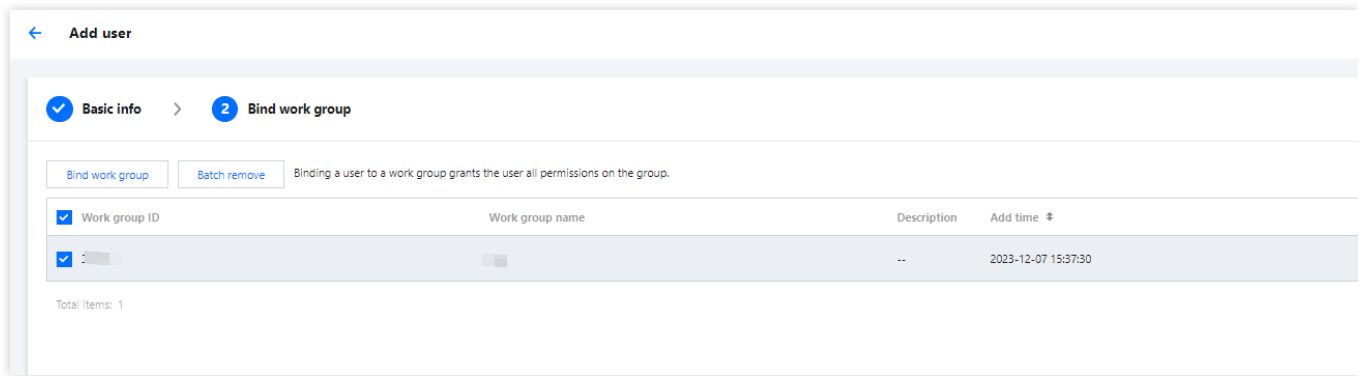
2. Bind a work group: General users need to be granted permissions or bound to a work group before they can access resources.

**Bind work group**  Binding a user to a work group grants the user all permissions on the group.

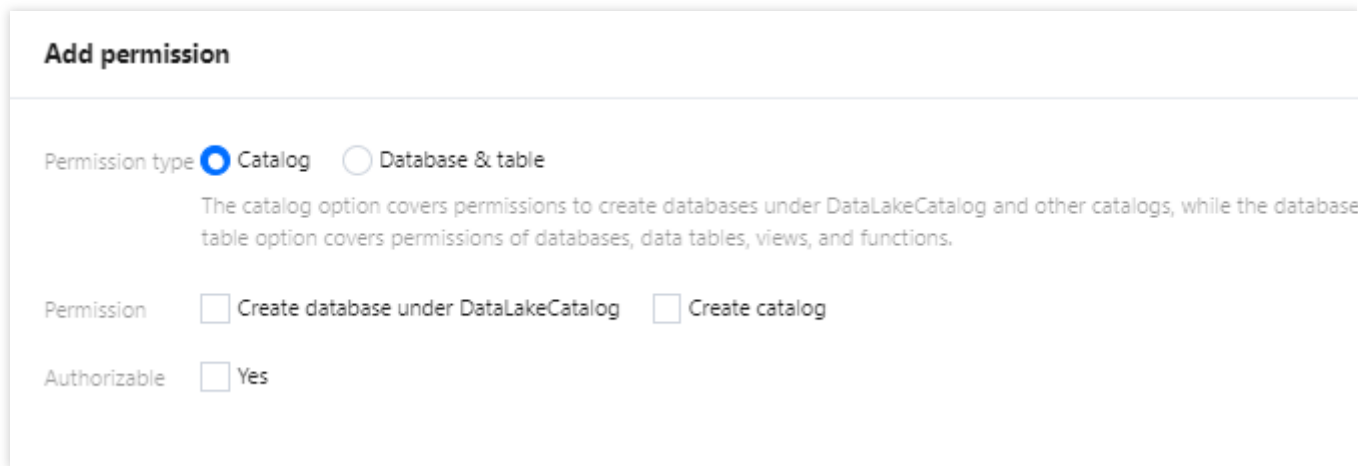
<input type="checkbox"/>	Work group ID	Work group name	Description	Add time ↕	Added by	Operation

Total items: 0 10 / page   1 / 1 page

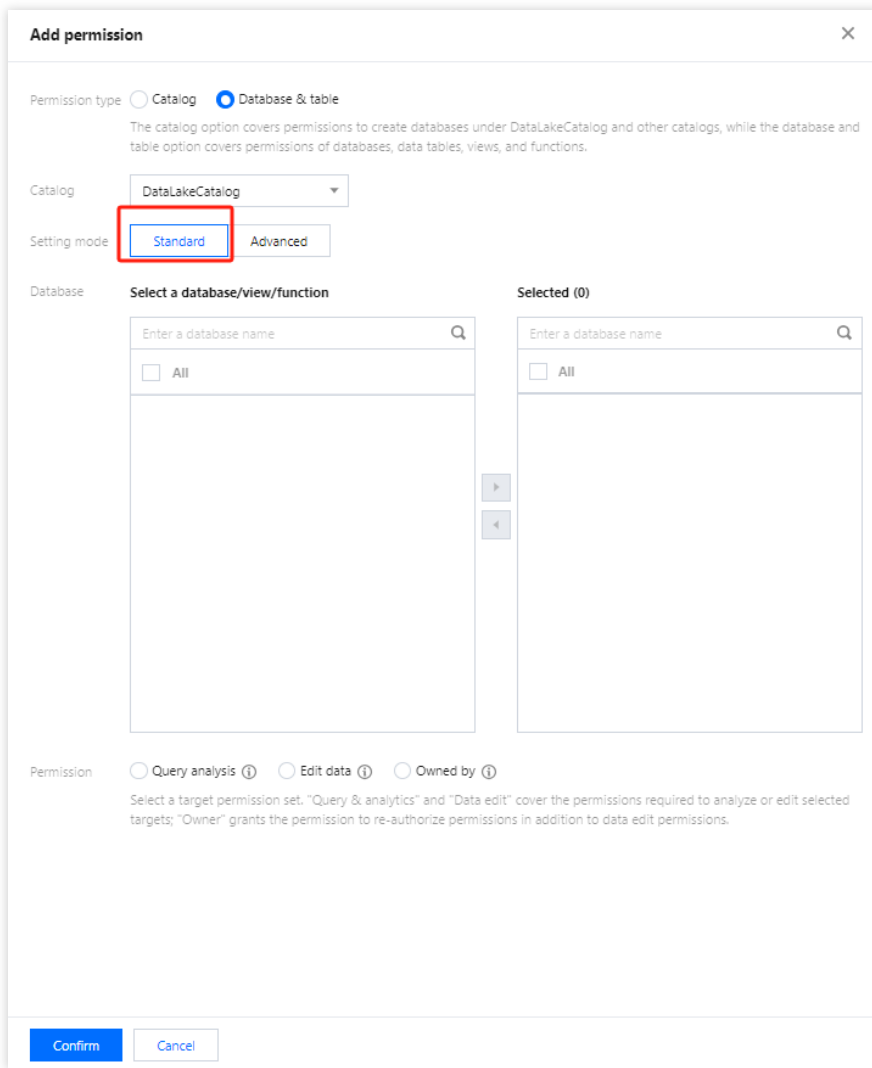
3. Add a data permission: In the **User list**, click **Authorize** in the **Operation** column and select **Data permission** to grant permissions at the data catalog or database/table level.



Add a data catalog permission. You can grant permissions to create databases under DataLakeCatalog and create other data catalogs.



Add a database/table permission: You can grant permissions in **Standard** or **Advanced** mode. In standard mode, you can grant database/table permissions in the specified catalog and set **Query & analytics**, **Data edit**, and **Owner** permissions.



Specific permissions are as follows:

Permission Type	Database	Data Table	View and Function
Query & analytics	<ul style="list-style-type: none"> <li>Query all the tables, views, and functions in databases.</li> <li>Create data tables.</li> </ul>	Query	Query
Data edit	<ul style="list-style-type: none"> <li>Modify and delete databases and create tables.</li> <li>Permissions of all the tables, views, and functions.</li> </ul>	<ul style="list-style-type: none"> <li>Query, insert, update, and delete data.</li> <li>Modify and delete tables.</li> </ul>	Query, create, modify, and delete.
Owner (grants the permission to re-authorize permissions in addition to data edit permissions)	<ul style="list-style-type: none"> <li>Modify and delete databases and create tables.</li> <li>Permissions of all the tables, views, and functions.</li> </ul>	<ul style="list-style-type: none"> <li>Query, insert, update, and delete data.</li> <li>Modify and delete tables.</li> </ul>	Query, create, modify,

and delete.

**Advanced permission settings:** When selecting a single database, you can further set the permissions to query, insert, update, and delete tables, views, and functions; when selecting multiple databases, you can only set permissions at the database level.

In advanced mode, you can set permissions at the column level. When selecting a single data table, you can add the permission to query columns. You can select one or more columns or all of them for authorization.

**Add permission** [X]

Permission type  Catalog  Database & table  
The catalog option covers permissions to create databases under DataLakeCatalog and other catalogs, while the database and table option covers permissions of databases, data tables, views, and functions.

Catalog DataLakeCatalog

Setting mode Standard **Advanced**

Database st

When selecting a single database, you can continue to set permissions for tables, views, functions, and columns; but when selecting more than one databases, you can only set permissions at the database level.

Name Data table in

Column col1

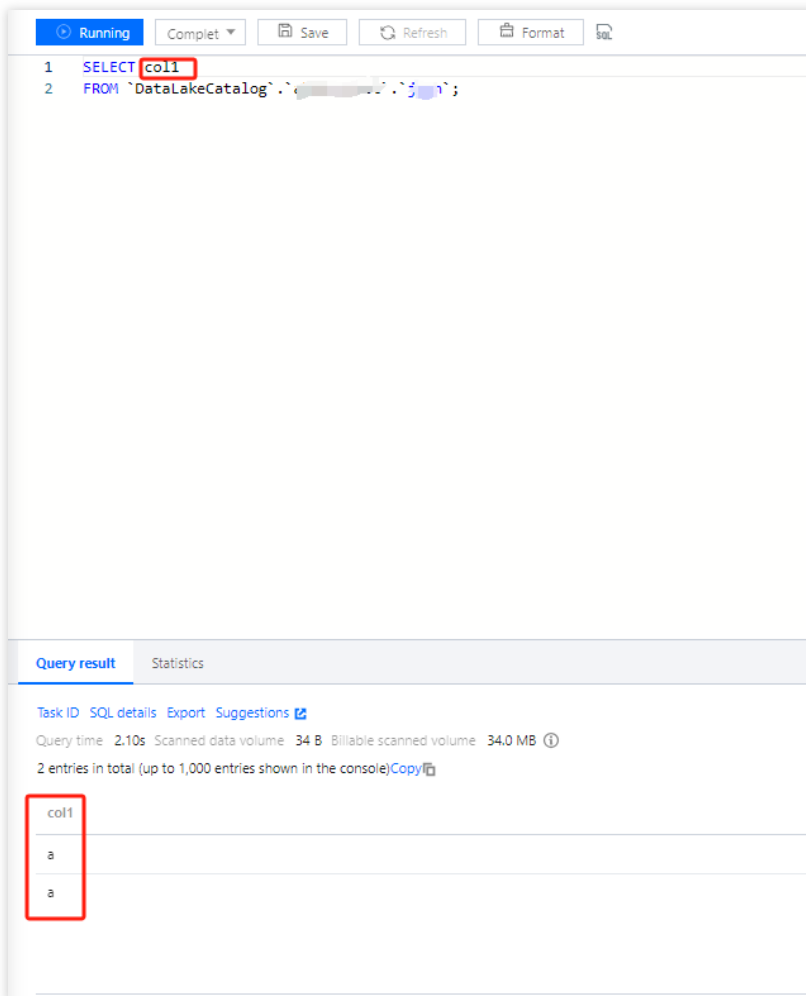
Column permission  SELECT ⓘ

Authorizable  Yes

Confirm Cancel

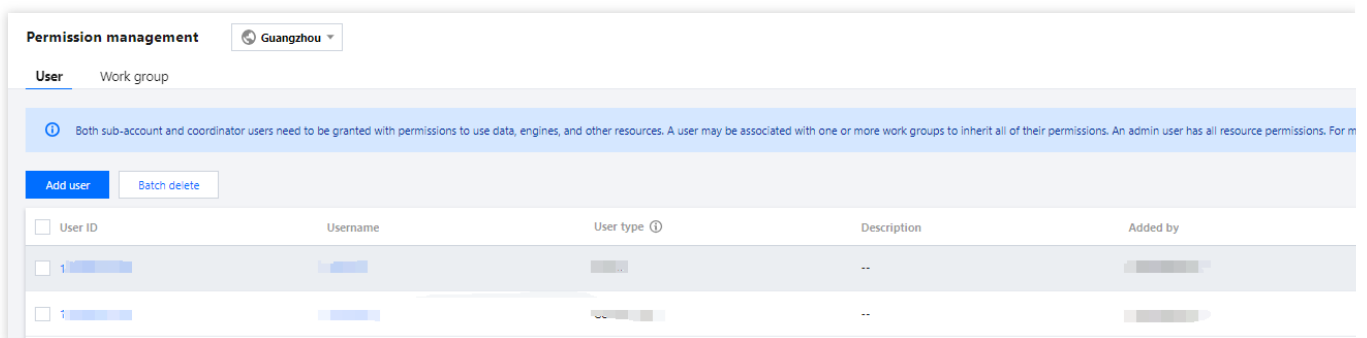
Click **Confirm** and perform queries in the **Data Explore** module. Enter the following SQL statement to preview the information of **col1** and run the statement to view the preview result of the column.





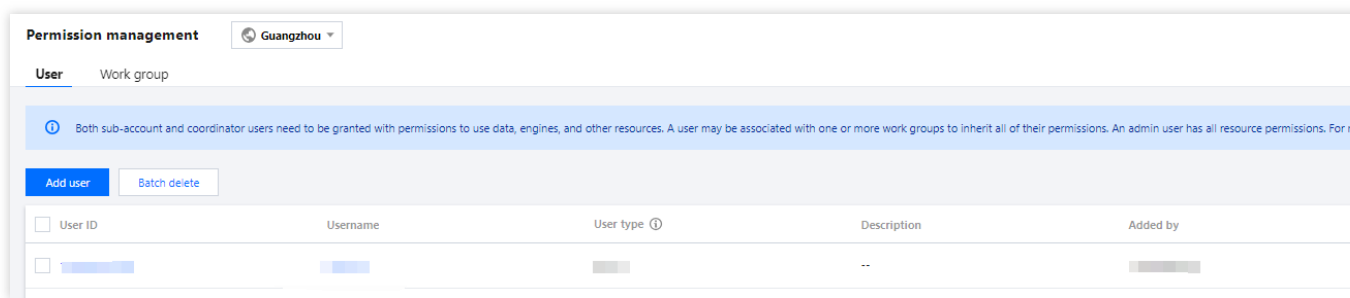
The permission is not granted for data column **b** in the data table. If you enter the SQL statement to view the information of **b**, the query cannot be performed due to lack of permission.

4. Add an engine permission: In the **User list**, click **Authorize** in the **Operation** column and select **Engine permission** to grant permissions to use, modify, manipulate, monitor, and delete specified resources.

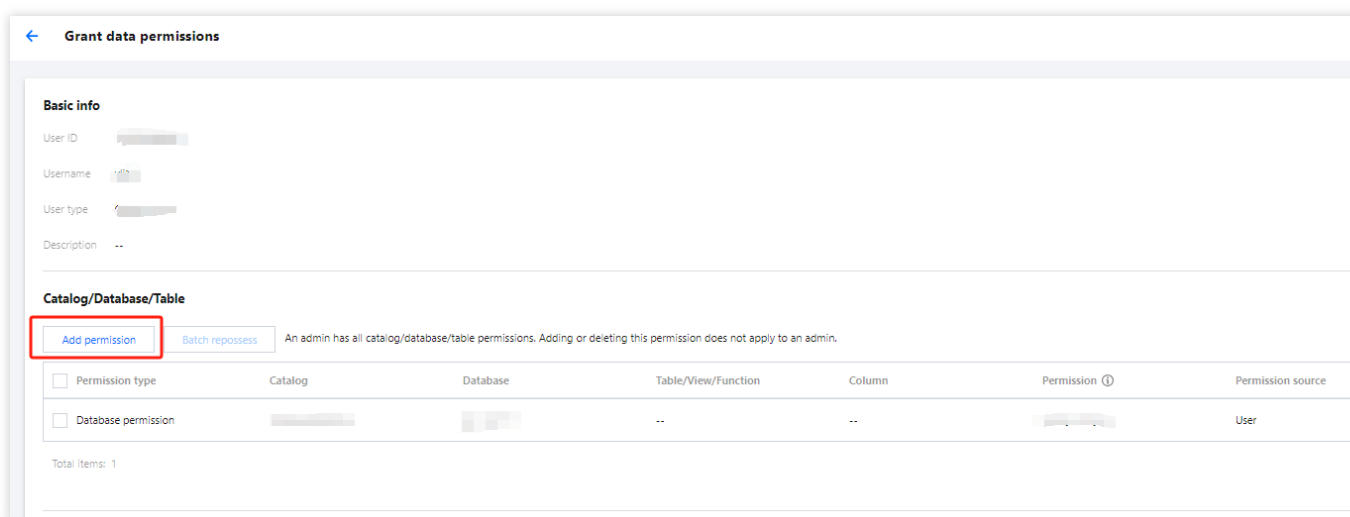


## Modifying a user permission

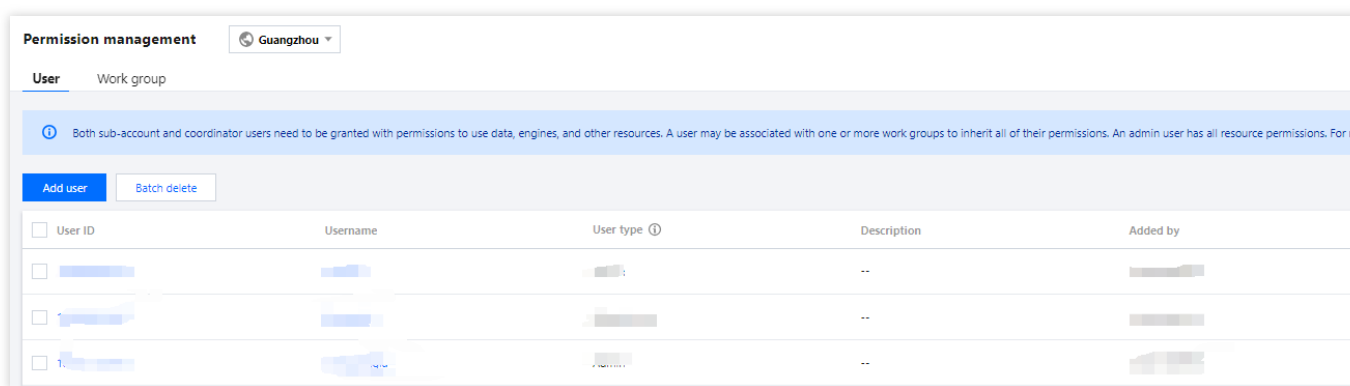
1. In the **User list**, click **Authorize** and select **Data permission** or **Engine permission**.



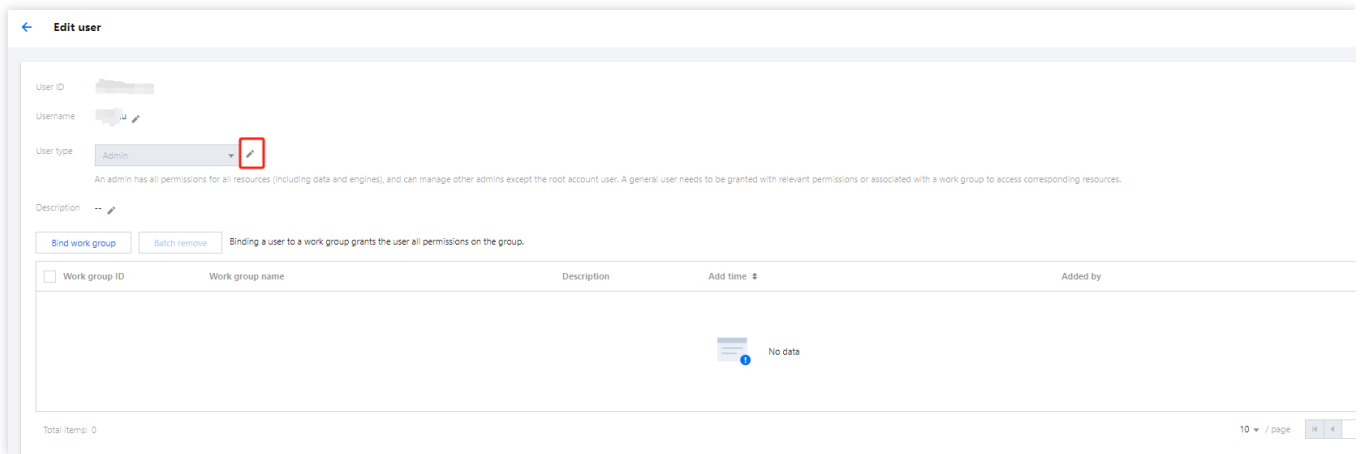
The following takes data permission as an example. On the **Data permission authorization** page, click **Add permission** or **Remove** to modify a permission. The steps for engine permission modification are similar.



2. Modify **Work group** or **User type**. Click **Operation** > **Edit** to enter the **Edit user** page, where you can modify the **Username**, **User type**, and **Description**. You can also add/remove general users to/from a work group.

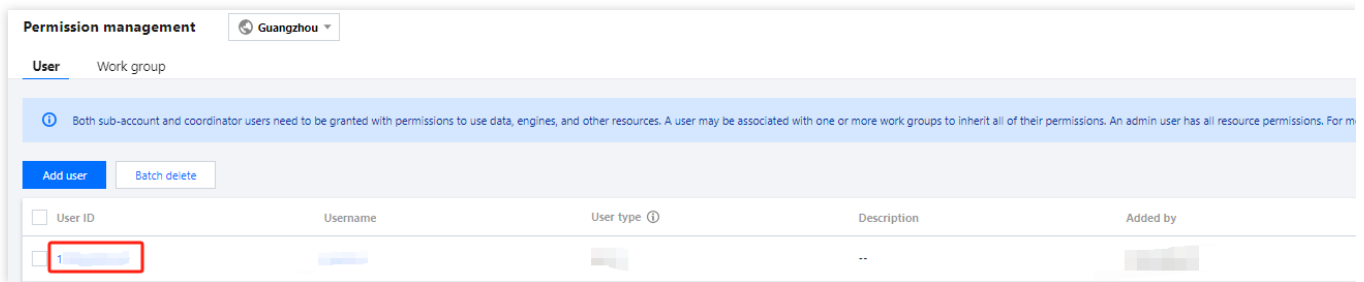


Click **Edit** to modify **User type**.



## Viewing a user's permissions

1. Click a user ID in the user list to enter the user details page.



2. View the user's work group, data permission, and engine permission information

**View user**

User ID [redacted]  
 Username sh[redacted]  
 User type A[redacted]  
 Description --

Work group    **Data permission**    Engine permission

**Catalog/Database/Table**

Include the user's data permissions and those inherited from a work group

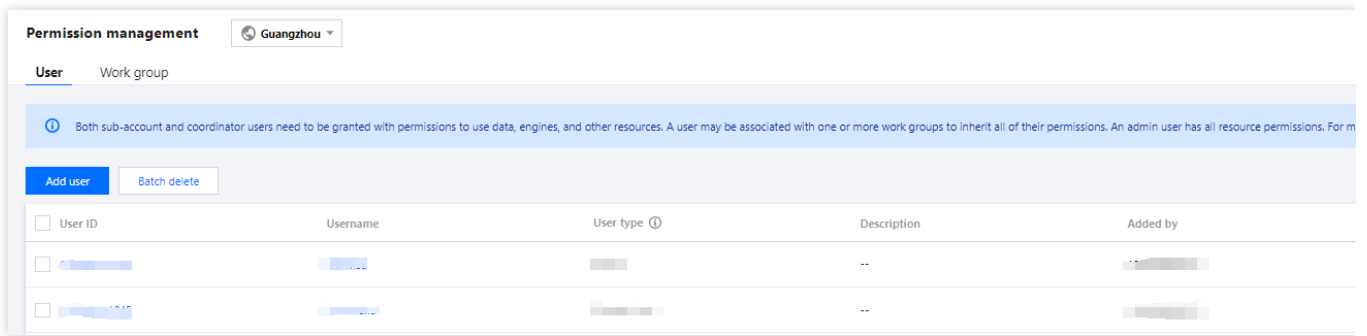
Permission type     Permission source

Permission ...	Catalog	Database	Table/View/Fun...	Column	Permission ⓘ	Permissi...
Function per...	[redacted]	[redacted]	[redacted]	--	[redacted]	[redacted]
Function per...	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Admin permi...	[redacted]	[redacted]	--	[redacted]	[redacted]	[redacted]

Total items: 3    10 / page    1 / 1 page

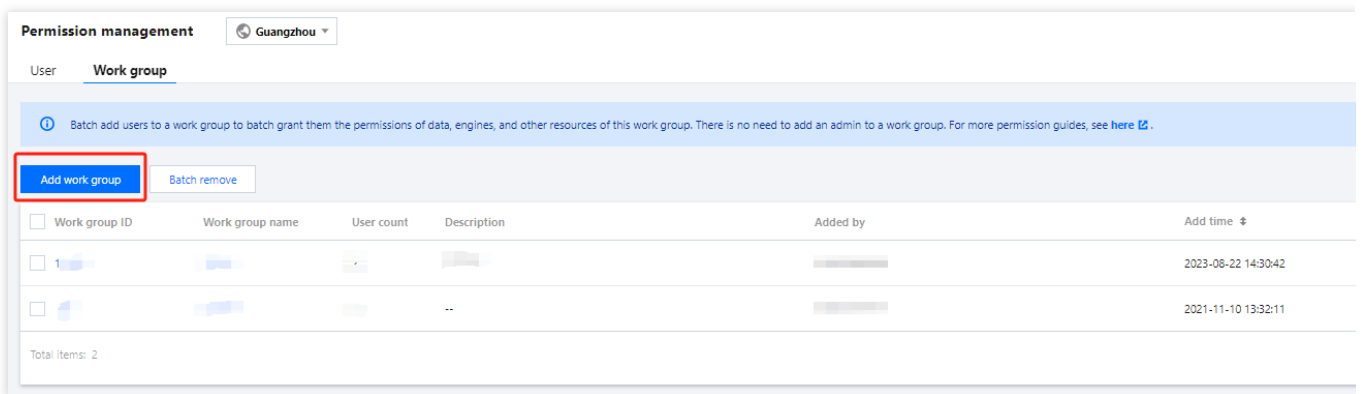
## Revoking a user's permissions

Remove permissions to be revoked from the permission list of a user. This operation requires the admin permission.



## Adding and removing a work group permission

Only admins can add or remove work group permissions in a similar way to manipulate data permissions. Users in a work group have all the permissions of the group, so you can bind users to a work group to grant them the data and engine permissions of the work group. Admins don't need to be bound to a work group.



# Storage Configuration

## Managed Storage Configuration

Last updated : 2024-07-31 17:30:11

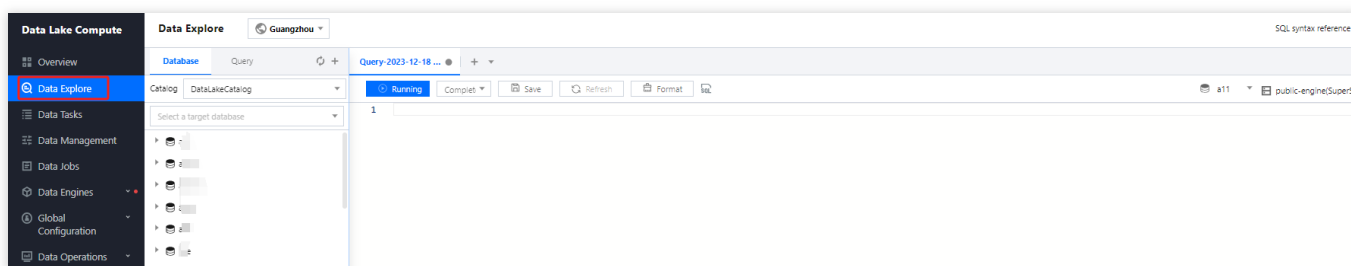
Managed storage refers to the storage space hosted on the Data Lake product, with COS as the underlying storage. Managed storage contains data such as native tables, user program packages, and query results. Therefore, to utilize the capabilities of native tables and data optimization, it is necessary to enable managed storage first. The native tables on managed storage are by default in the Iceberg format, so you don't need to manage the underlying file contents. For details on managed storage billing, please refer to [Billing Overview](#).

This document introduces how to enable and configure managed storage.

## Enable Managed Storage

### Step 1: Enter Managed Storage Configuration

You can enter the managed storage configuration in the [Data Exploration](#) module or the [Global Configuration > Storage Configuration](#) module.



### Step 2: Open Managed Storage

1. Check to enable managed storage and save.

Here, you can specify the managed storage type as either a Metadata Acceleration Bucket or an Ordinary Bucket. The billing for both is consistent, but it is necessary to separately configure engine access permissions for the Metadata Acceleration Bucket. For details, please refer to [Binding of Metadata Acceleration Bucket](#).

2. The query result path is used to temporarily store SQL query results, Spark Job Shuffle data, etc. You need to specify a path to ensure the normal operation of jobs and tasks. If you have enabled managed storage, it is recommended to configure the query result path as **Managed Storage**. You can also configure the query result path to your own account's [COS bucket](#) path.

### Storage configuration

Managed storage ⓘ  Enable

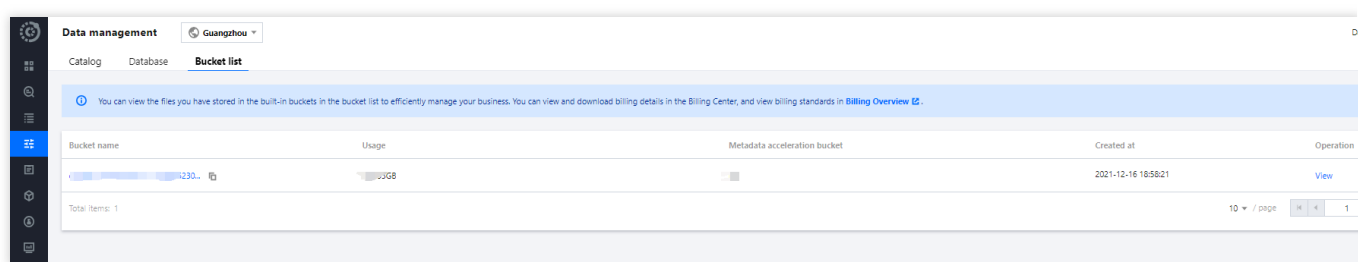
Managed storage type ⓘ General bucket

Query result storage path ⓘ Managed storage User-defined storage

Save Cancel

## View managed bucket

After enabling managed storage, a bucket will be created, and you can view the buckets and data on managed storage in the [Data Management](#) module.



## Destroy Managed Storage

Destroying data is a high-risk action; only after all database table data has been deleted, can you proceed to destroy managed storage. Destroying managed storage requires administrator privileges.

### Step one: Delete database table data

To destroy managed storage, you must first delete all database table data on the managed storage.

You can refer to the [Data Catalog and DMC](#) and [Data Table Management](#) documents to delete the database table data, or you can run the [DROP Syntax](#) in the [Data Exploration](#) module to delete the database table data.

## **Step two: Destroy Managed Storage**

After deleting the database table data, you can destroy managed storage on the managed storage configuration tab under the [Storage Configuration](#) module.

Destroying managed storage will delete all DLC managed buckets, so please proceed with caution.



# Binding a Metadata Acceleration Bucket

Last updated : 2024-07-31 17:30:27

DLC supports the binding of Fusion Bucket to accelerate Query Analysis Performance. To use this feature, you need to create a Metadata Acceleration Bucket. DLC Managed Storage provides Metadata Acceleration Bucket. Use COS Bucket under the user's account. For details, please see [COS>Metadata Acceleration](#).

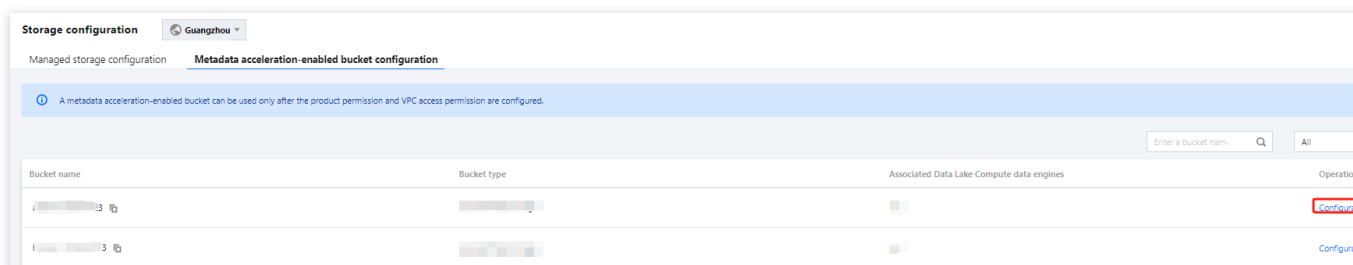
When accessing the DLC Metadata Acceleration Bucket, binding of permissions is necessary. The Permission Binding Process is as follows.

## Bind Data Engine and Metadata Acceleration Bucket

1. log in to [Data Lake Computing Console](#), enter Common Management > [Storage Configuration](#).
2. Enter the **Metadata Acceleration Bucket Configuration Page**, select the bucket you want to bind, and click **Configure**.

### Note:

Only Metadata Acceleration Buckets are displayed on the Metadata Acceleration Bucket page; ordinary buckets (buckets without the metadata acceleration feature enabled) will not be shown.



3. Click **Bind** to bind the data engine that needs to access this bucket to the Metadata Acceleration Bucket.

**Edit access to metadata acceleration-enabled bucket** ✕

Metadata acceleration-enabled bucket name

Metadata acceleration-enabled bucket type

**Bind data engine**

🔄

Data engine name	Operation
<input type="text" value=""/>	<a href="#">Bind</a>
ri- <input type="text" value=""/>	<a href="#">Unbind</a>

Total items: 2 10 / page ⏪ ⏩ 1 / 1 page

**Associate Tencent Cloud products**

Product	Resource	Operation
No data yet		
<a href="#">Add product</a>		

**Set HDFS user** [Edit](#)

Superuser

Note: This section enables you to manage the tenant information of compute nodes.

**Set access to HDFS metadata**

VPC name/ID	Node IP	Operation
<input type="text" value=""/>	<input type="text" value=""/>	<a href="#">Edit</a> <a href="#">Delete</a>

## Bind computing resources of SCS

If you use SCS to stream data into the lake, and the storage written to is a Metadata Acceleration Bucket, then you need to configure access permissions for the Metadata Acceleration Bucket under [Storage Configuration](#). Under the Tencent Cloud Product Binding section, create a new product, select Stream Computing Oceanus and the corresponding resources, then click save.

**Edit access to metadata acceleration-enabled bucket**
✕

Metadata acceleration-enabled bucket name

Metadata acceleration-enabled bucket type

**Bind data engine**

Data engine name	Operation
<input type="text"/>	<a href="#">Bind</a>
<input type="text"/>	<a href="#">Unbind</a>

Total items: 2      10 / page        1 / 1 page

**Associate Tencent Cloud products**

Product	Resource	Operation
<input type="text" value="Stream Compute Service"/>	<input type="text" value="Select"/>	<a href="#">Save</a> <a href="#">Cancel</a>

**Set HDFS user** [Edit](#)

Superuser

Note: This section enables you to manage the tenant information of compute nodes.

**Set access to HDFS metadata**

VPC name/ID	Node IP	Operation
<input type="text"/>	<input type="text"/>	

## Bind computing resources of non-DLC data engines

Sometimes, the computing resources you need to access the Metadata Acceleration Bucket are not from a DLC data engine. In this case, you can configure access permissions for the Metadata Acceleration Bucket under [Storage Configuration](#).

HDFS User Configuration is used to configure the super user of your computing resources accessing DLC, usually root/hadoop/presto/flink.

HDFS Metadata Permissions Configuration is used to configure the VPC Network Environment you allow to access DLC, usually the VPC where the computing resources of the above mentioned non-DLC data engines are located.

**Set HDFS user** [Edit](#)

Superuser



Note: This section enables you to manage the tenant information of compute nodes.

**Set access to HDFS metadata**

VPC name/ID	Node IP	Operation
[Redacted]	[Redacted]	<a href="#">Edit</a> <a href="#">Delete</a>
[Redacted]	[Redacted]	<a href="#">Edit</a> <a href="#">Delete</a>
[Redacted]	[Redacted]	<a href="#">Edit</a> <a href="#">Delete</a>
<a href="#">Add</a>		

Note: This section enables you to allow/forbid specified compute nodes in a specified VPC to operate the current bucket.

# Audit Log

Last updated : 2024-07-31 17:30:53

DLC provides an operation log audit service based on Tencent Cloud's CloudAudit service, ensuring you can understand the system operation records in real time and check the operation information.

## Notes

Before using the audit CLS of DLC, you need to activate Tencent Cloud's [CloudAudit service](#). If the service is not yet activated, you can activate it with the primary account.

## Use Instructions

The Data Lake Computing Console currently displays up to 3 months of log information. To view older log information, you can go to CloudAudit.

The audit logs contain console operations and API call operations. Currently, it supports viewing log information for engine management, task management, data source management, workgroup management, user management, scheduled task instance management, scheduled task management, and scheduling plan management.

## Operation Guide

1. log in to [Data Lake Computing Console](#), select **Service Region**.
2. Through the left menu **Data Operation and Maintenance**, select the Audit Log feature.
3. Supports log queries based on user UIN or request ID.
4. Detailed log information can be viewed by clicking **Query Details**.

**Run history** Guangzhou

This module displays the status of tasks submitted in other modules, including SQL tasks and data import/export tasks. An admin can query all tasks in the last 45 days, while a general user can query tasks related to them in the last 45 days. [Learn more](#)

Select an execution status | Select a job or task creator | Select a data engine | Select a task type | Batch operat | Today | Last 7 days | Last 30 days | 2023-12-18 ~ 202

### Job overview

All	Executing	Queueing up	Initialize
1	0	0	0

Task ID	Task type	Task content	Execution status	Creator	Task submission time	Data engine	Resource usage	Kernel version	Operatio
<input type="checkbox"/>			Successful		2023-12-18 17:33:28		--		<a href="#">Learn mo</a>

Total items: 1 10 / page 1

# Monitoring and Alarms

## Data Engine Monitoring

Last updated : 2024-07-31 17:31:18

Data Lake Compute (DLC) provides monitoring services for data engines based on the Tencent Cloud Observability Platform (TCOP), ensuring you can understand the real-time status of data engines and configure data alarms. For alarm configuration methods, see [Monitoring Alarm Configuration](#).

## Usage Notice

Before using the Data Lake Compute (DLC) monitoring service, you need to activate the TCOP service. If this service is not yet activated, you can use the root account to activate it.

The use of the TCOP service may incur related charges. For detailed pricing information, see [Billing Overview](#).

## Monitoring Access

### Access Point I: Data Lake Compute (DLC) Console

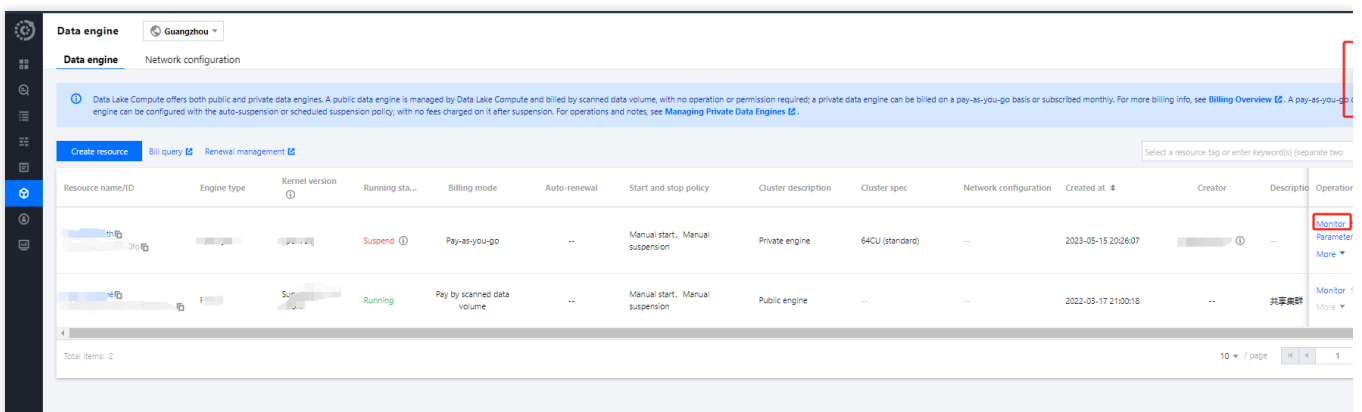
#### Note:

The account must have monitoring permissions for the data engine.

1. Log in to the [DLC console](#) and select the service region.
2. Navigate to the **SuperSQL engine** page from the left menu.
3. Viewing methods supported:

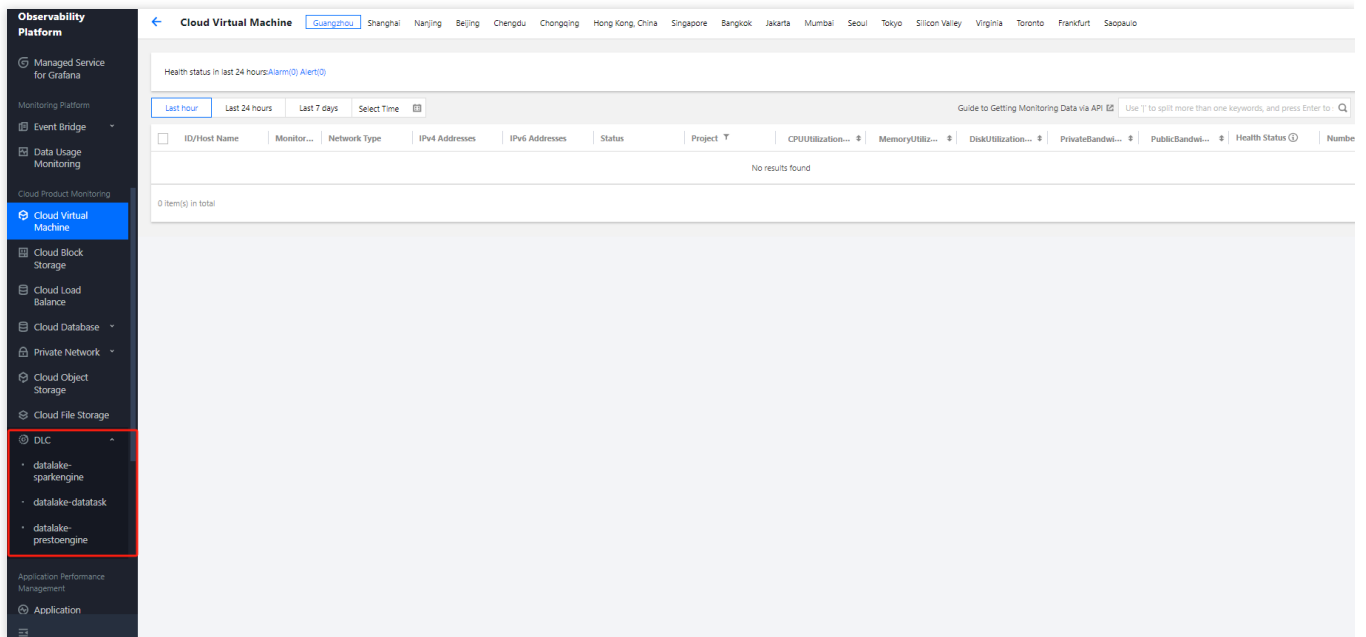
Method 1: Select the engine type to enter the matching engine monitoring list.

Method 2: Select the target engine from the engine list and click **Monitoring** to view the target engine monitoring.

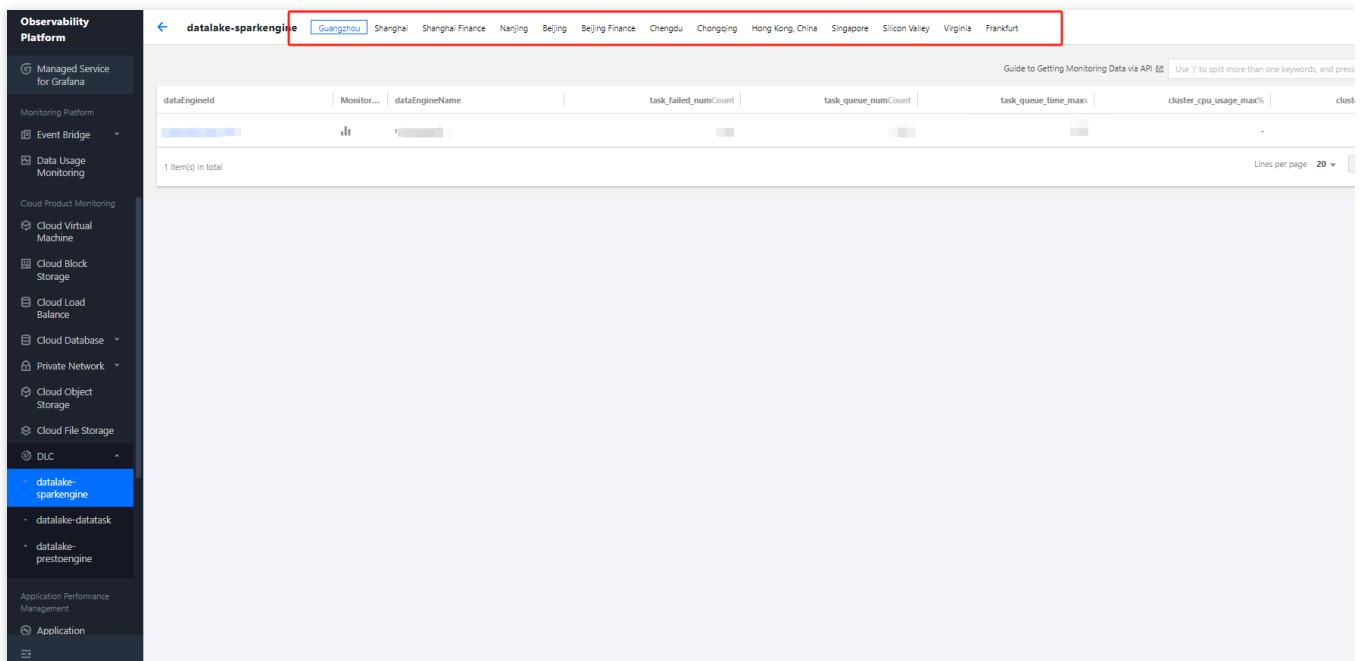


### Access Point Two: TCOP

1. Log in to the **TCOP** with an account that has the necessary permissions.
2. Select **Cloud Product Monitoring** from the left menu, find Data Lake Compute DLC, and choose the type of monitoring you need to view.



3. After selecting the monitoring type, you will enter the monitoring page. Select the corresponding region to view the monitoring resource information for that region.

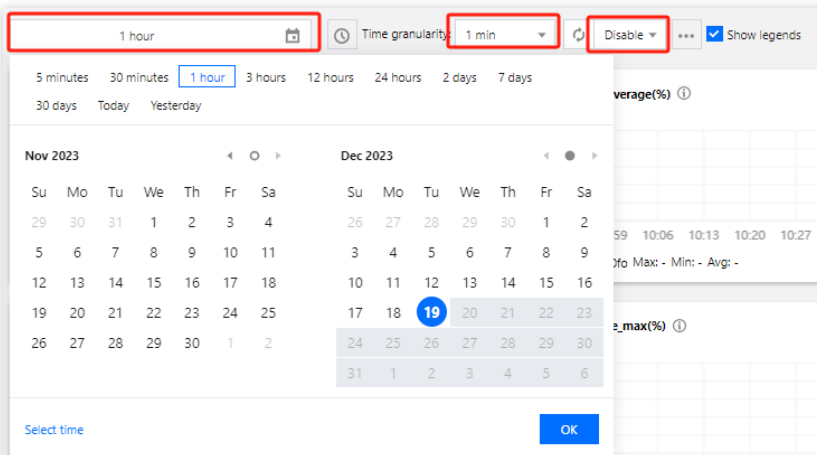


4. Click the **Engine ID** to enter the detailed monitoring page.

## Monitoring Granularity Configuration



You can configure the monitoring data time range, time granularity, and auto-update interval at the top of the monitoring page.



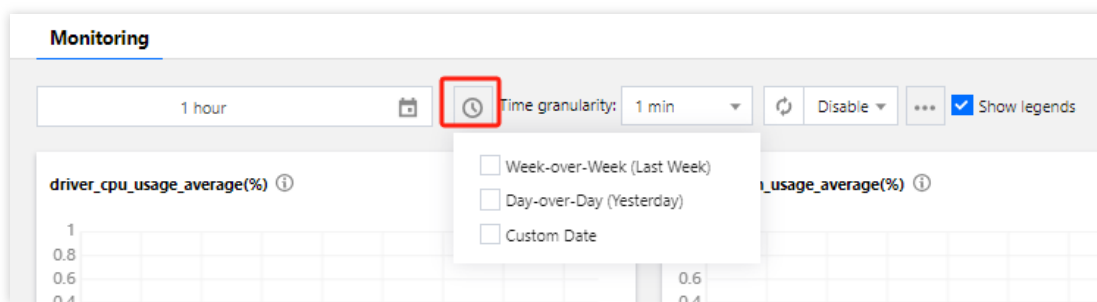
Monitoring data time range: Accurate to the minute, supports selecting data for a specific time period.

Time granularity: Interval between monitoring points, configurable to 1 minute or 5 minutes.

Auto-update data: Configures the automatic refresh interval for page data, with options to set it to off, 30 seconds, 5 minutes, 30 minutes, or 1 hour.

## Monitoring Data Comparison

You can select a time period for data comparison. After selecting the comparison time range through one click, you can view the comparison data in the data compass below.



## Monitoring Metrics

Monitoring Type	Monitoring Metrics
CPU	Maximum CPU utilization of all Driver nodes
	Maximum CPU utilization of all Executor nodes
	Average CPU utilization of all Driver nodes

	Average CPU utilization of all Executor nodes
	Maximum CPU utilization of all clusters
	Average CPU utilization of all clusters
Memory	Maximum memory utilization of all Driver nodes
	Maximum memory utilization of all Executor nodes
	Average memory utilization of all Driver nodes
	Average memory utilization of all Executor nodes
	Maximum memory utilization of all clusters
	Average memory utilization of all clusters
Tasks	Number of canceled tasks
	Number of failed tasks
	Number of initialized tasks
	Average task initialization time
	Maximum task initialization time
	Number of queued tasks
	Average task queue time
	Maximum task queue time
	Number of running tasks
	Number of successful tasks
Network	Maximum inbound bandwidth of all Driver nodes network
	Maximum inbound bandwidth of all Executor nodes network
	Average inbound bandwidth of all Driver nodes network
	Average inbound bandwidth of all Executor nodes network
	Maximum outbound bandwidth of all Driver nodes network
	Maximum outbound bandwidth of all Executor nodes network

	Average outbound bandwidth of all Driver nodes network
	Average outbound bandwidth of all Executor nodes network
Cloud Disk	Maximum cloud disk utilization of all Driver nodes
	Maximum cloud disk utilization of all Executor nodes
	Average cloud disk utilization of all Driver nodes
	Average cloud disk utilization of all Executor nodes
CU	Job Engine CU Count
	CU Utilization

# Data Job Monitoring

Last updated : 2024-07-31 17:31:39

DLC provides monitoring services for data jobs based on TCOP service, ensuring that you can understand the operation of data jobs in real time and configure data alarms.

## Notes

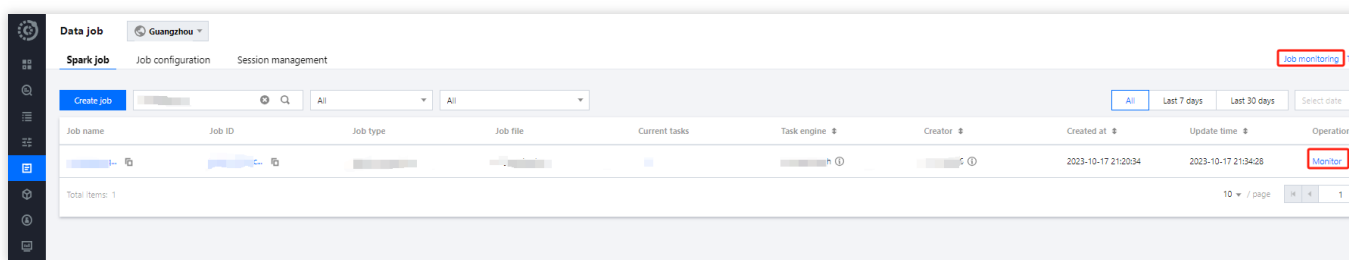
Before using the monitoring service of DLC, you need to activate the TCOP service (for usage details, refer to [TCOP Documentation](#)). If the service has not been activated, it can be done using the root account.

Fees may be incurred during the use of TCOP service; for detailed fee information, refer to [TCOP Billing Overview](#).

## Monitoring Entrance

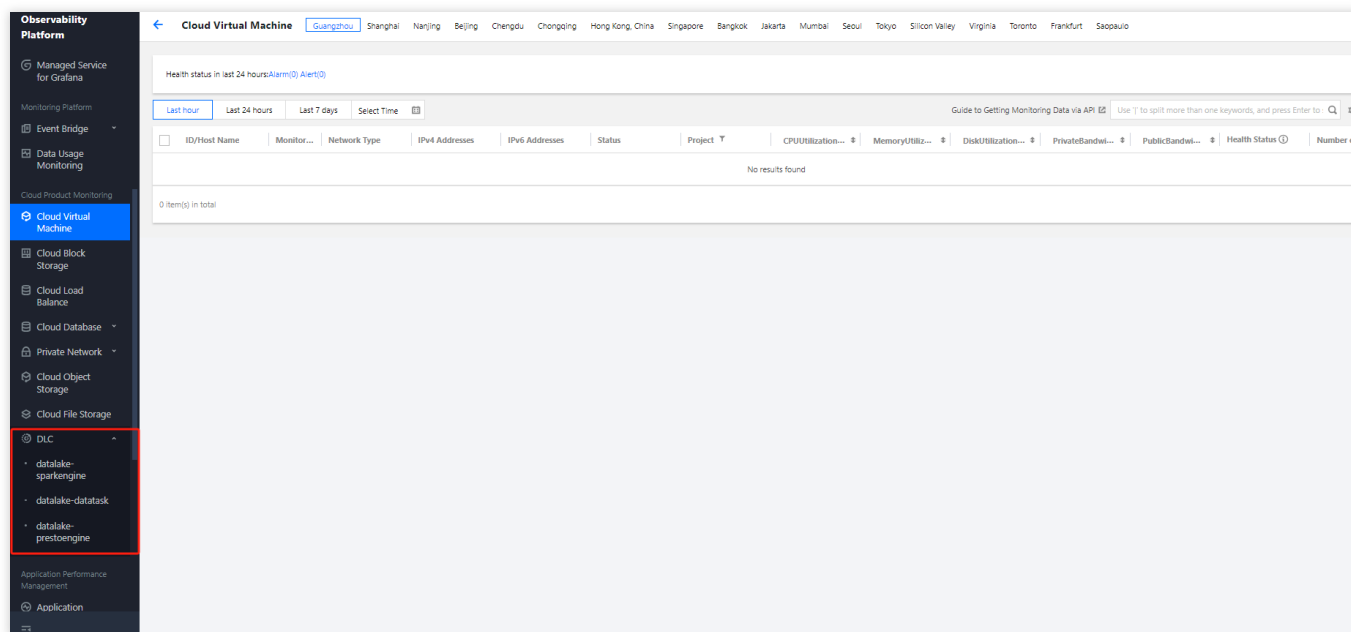
### Entrance one: DLC Console

1. Log in to [DLC Console > Data Job](#), and select the service region.
2. Or enter the Data Job page from the left sidebar.
3. In the top right corner, click **Job Monitoring** to go to the monitoring page. Or click the **Monitoring** feature of the target job to enter its monitoring page.

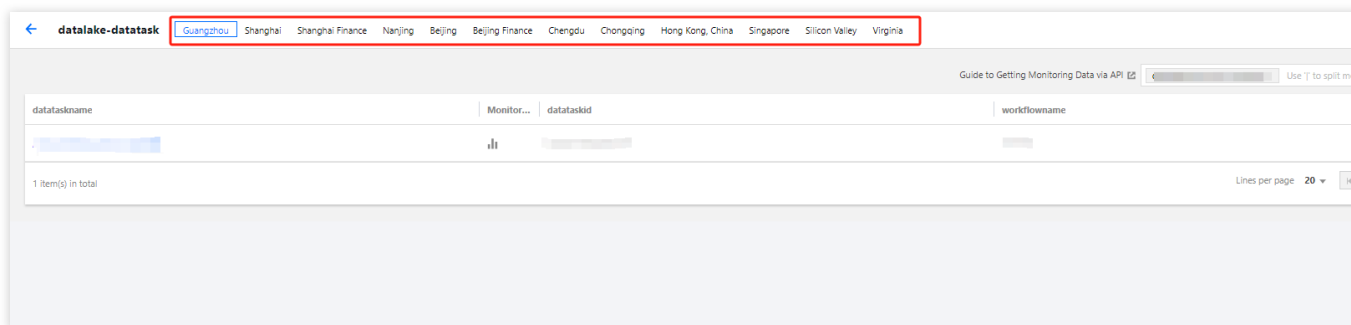


### Entrance two: TCOP

1. Log in to [TCOP Console](#). Account must have the required permissions.
2. In the left menu, select Cloud Product Monitoring, find DLC, and choose the type of monitoring you wish to view.



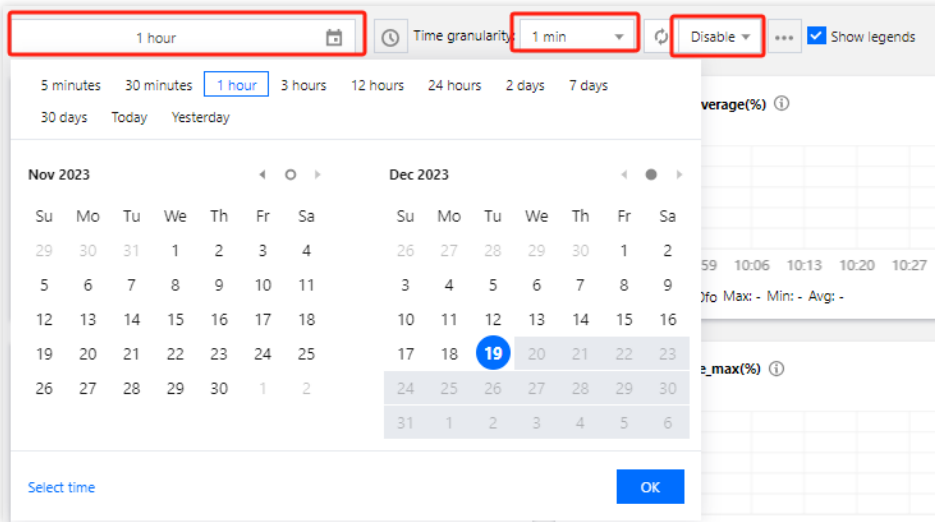
3. After selecting the monitoring type, enter the monitoring page and select the respective region to view the monitoring job information for that region.



4. Click **Job ID** to enter the monitoring details.

## Monitoring Granularity Configuration

Supports configuring the monitoring data time period, time granularity, and automatic update time range through the monitoring settings at the top.



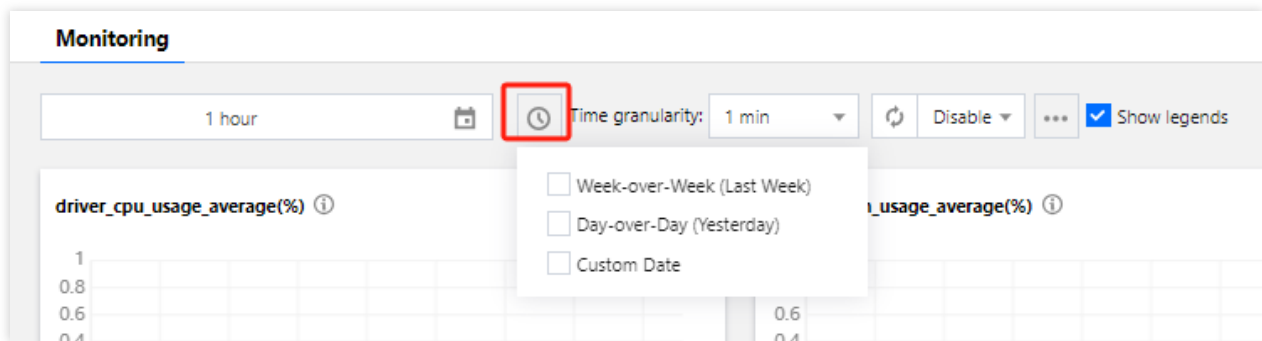
Monitoring Data Time Range: Precise to minutes, supports selecting data for a specific period.

Time Granularity: Monitoring point interval time, supports configuring for 1 minute or 5 minutes.

Automatic Data Update: Page data auto-refresh configuration, supports configuring off, 30s, 5min, 30min, 1h.

## Monitoring Data Comparison

Supports selecting data for a specific period to compare monitoring data. After clicking to select the comparison time range, you can view the comparison data in the data compass below.



## Monitoring Metric

Monitoring Type	Monitoring Metric
Job	Job error Log Count
	Job warn Log Count



# Access Point Gateway Engine Monitoring

Last updated : 2024-07-31 17:31:54

DLC provides monitoring services for the access point gateway engine based on TCOP service, ensuring you can understand the gateway status in real time.

## Notes

Before using DLC's monitoring service, you need to activate the TCOP service (for usage details, see [TCOP Documentation](#)). If the service has not been activated yet, it can be activated using the root account.

TCOP service usage may incur related tariffs, for detailed tariff information, see [TCOP Billing Overview](#).

## Monitoring Entrance

### Entrance one: DLC Console

1. Log in to the <1>Standard Engine</1> page, and select the Service Region.
2. Select the Standard Engine, and click on **Monitoring** at the access point to enter the monitoring data display interface.

### Configuration Entrance: TCOP

1. Log in to the [TCOP Console](#), the account must have the relevant permissions.
2. From the left menu, select Cloud Product Monitoring, enter the [Policy Management](#) page under Alarm Management, select Data Lake Computing, and choose the corresponding Access Point Gateway Engine.

## Access Point Gateway Engine Monitoring Configuration Type

### Creating alarm policy

1. DLC Access Point Gateway supports alarm capabilities. Log in to [TCOP](#), click **Alarm Management**, and select the [Policy Management page](#).
2. Click **New Policy**, for policy type choose "Data Lake Computing". Access Point Gateway supports alarms for three dimensions, including:

"Gateway" alarm dimension is: appid/gatewayid.

"Gateway (Multi-dimensional)" alarm dimension is: appid/gatewayid/instanceid.

"Gateway Engine (Multi-dimensional)" alarm dimension is: appid/gatewayid/engineid/processid.



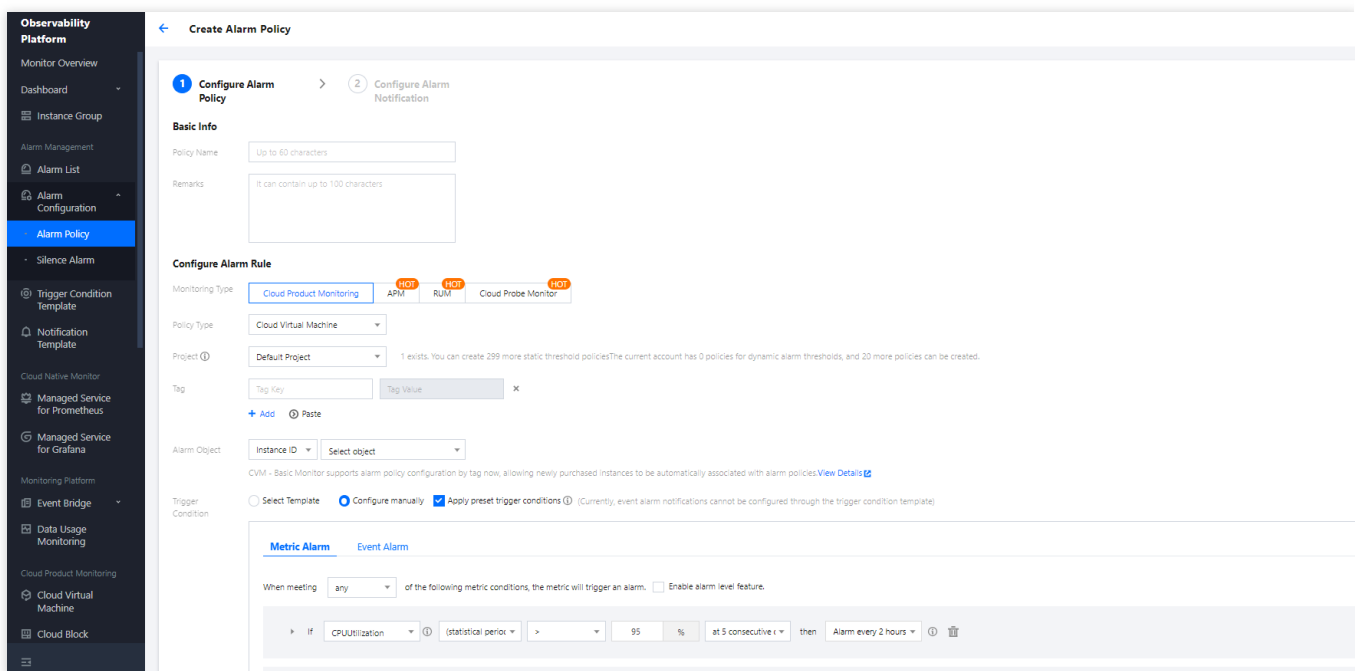
Name	Supported Dimensions	Advantages and Use Cases
Gateway (Multi-dimensional)	<p>Supports: CPU, Memory, Disk, Network Fine-grained Alerting.</p> <p>For example, to configure an alert for the CPU utilization of an Access Point Gateway, you can choose to configure one, several instances under a specific Access Point Gateway, or any instance node triggering the threshold to alert.</p>	<p>Alert supports more dimensions, and the alert method is more flexible. Basic Metrics are recommended to use this approach.</p>
API Gateway	<p>Mainly aimed at monitoring the overall load situation of the current gateway, aggregating basic metrics according to Access Point Gateway Nodes, and supporting Service-level Metric Alerts.</p> <p>For example: <code>execute_statement_num</code> (number of statements executed), <code>opened_operation_num</code> (number of operations opened), <code>launch_engine_num</code> (number of engines started), <code>engine_process_thread_num</code> (number of threads started by the engine).</p>	<p>Supports Dashboard. Suitable for Single-node access point gateway or service metric alert.</p>
Gateway Engine (Multidimensional)	<p>The Gateway Engine refers to the monitoring and alarm of the process of starting the DLC engine by the Access Point Gateway.</p> <p>For example: <code>engine_process_thread_num</code> (number of threads started by the engine), mainly aimed at monitoring the process information of the engine started by the current Access Point Gateway</p>	<p>Supports fine-grained alerting, for example: commonly configure any engine's process count under a specific Access Point Gateway ID to reach the threshold to trigger an alert. Suitable for alerting on process metrics started by the Access Point Gateway.</p>

# Monitoring Alarm Configuration

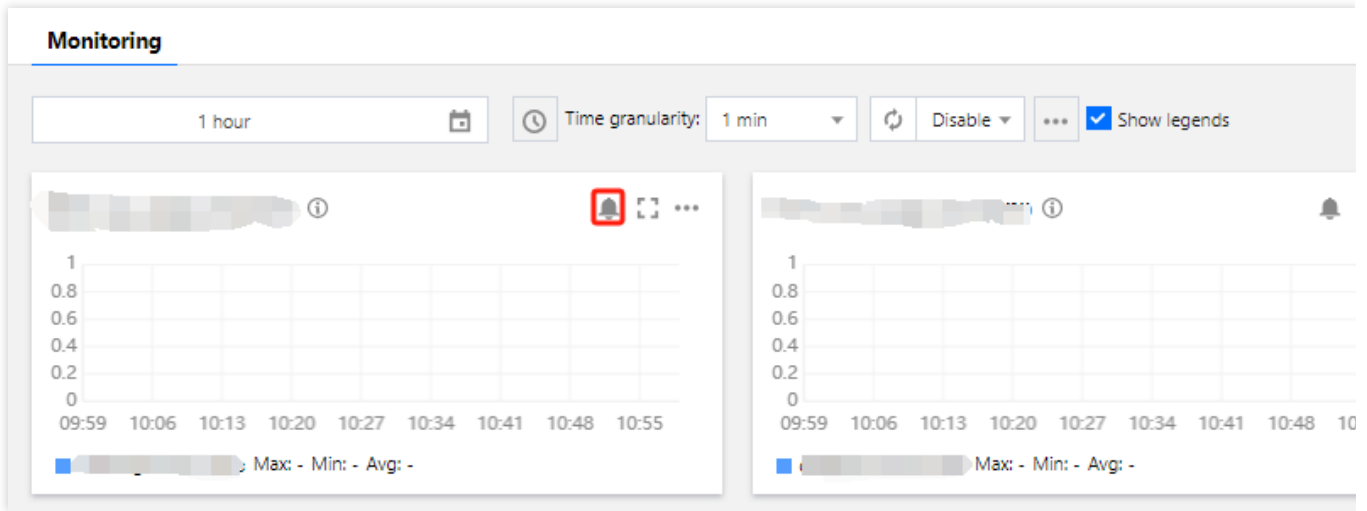
Last updated : 2024-07-31 17:32:15

## Configuring New Alarm Policy

Supports configuring monitoring alarms for specific metrics. You can go to [Creating Alarm Policy](#) to configure the content of the alarm.



Or click the monitoring content for which you need to configure an alarm to enter the configuration page, where you can configure the content of the alarm.



## Managing an alarm policy

To manage configured alarm policies, you can perform configuration management through the [Policy Management](#) page.

The screenshot shows the 'Alarm Management' section of the Tencent Cloud console, specifically the 'Policy Management' tab. The left sidebar contains a navigation menu with 'Alarm Policy' highlighted. The main content area features a table of alarm policies. At the top, there are buttons for 'Create Policy', 'Delete', and 'More', along with an 'Advanced Filter' and a search box for 'Policy Name/ID'. The table has columns for 'Policy Name', 'Monitoring Type', 'Policy Type', 'Alarm Rule', 'Project', 'Associated Instances', 'Notification Template', 'Last Modified', 'Alarm On-Off', and 'Oper'. One policy is listed with the following details: Policy Name (redacted), Monitoring Type 'Tencent Cloud services', Policy Type 'datalake-gateway-engine-md', Alarm Rule (redacted), Project (redacted), Associated Instances (redacted), Notification Template (redacted), Last Modified '2023/11/14 20:56:14', and Alarm On-Off 'On'. A 'Copy Alarm' link is visible at the end of the row. Below the table, it indicates 'Total Items: 1' and a pagination control showing '20 / page'.

## Configuration Instructions

Configuration Item	Configuration Instructions
Policy name	Name of the alarm policy, up to 60 characters
Remarks	Remarks for the alarm policy, up to 100 characters
Monitoring Type	Please select Cloud Product Monitoring
Policy Type	Please select DLC
Policy Tag	Support for managing policy content via Tag requires relevant permissions to operate
Alarm Object	You can configure alarms for Instance ID (supports multiple selections), grouped instances, and all instances
Alert Configuration Template	You can choose a template or configure manually. Administrators need to create the template in advance, and it supports configuring multiple alert rules
Notification Template	Supports creating or selecting existing notification templates, with support for configuring up to 3 templates

# Query Script Management

Last updated : 2022-08-16 09:41:59

Data Lake Compute provides a script file management feature to facilitate and accelerate repeated query tasks.

Note :

Up to 100 SQL scripts can be saved in the console.

## Creating a script

1. Log in to the [Data Lake Compute console](#), go to the **Query analysis** page, and click **Resource package**.

2. On the **Resource package** tab, hover over the **Script file** row and click  > **Create script**.

3. Enter and save the script content.

- Script name: It can contain up to 25 letters and underscores.
- Script description: It can contain up to 2,048 characters.
- SQL statement: It must be a standard SQL statement containing up to 1,000 characters within 2 MB in size.

## Running a script

1. Log in to the [Data Lake Compute console](#), go to the **Query analysis** page, and click **Resource package**.

2. On the **Resource package** tab, hover over the target script name and click  > **Copy script to SQL**.

3. Select the compute engine and click **Run**.

## Viewing script information

1. Log in to the [Data Lake Compute console](#), go to the **Query analysis** page, and click **Resource package**.



2. On the **Resource package** tab, hover over the target script name and click **> Learn more** to view the script details.

## Deleting a script

1. Log in to the [Data Lake Compute console](#), go to the **Query analysis** page, and click **Resource package**.



2. On the **Resource package** tab, hover over the target script name and click **> Delete** to delete the script.

Note :

Note that a deleted script cannot be recovered. Proceed with caution.

# System Restraints

## Metadata Information

Last updated : 2024-07-17 18:17:23

The following lists the limits on the numbers of databases, data tables, attribute columns, and partitions.

Item	Maximum Number
Databases per account	1,000
Data tables per account	10,000
Tables per database	4,096
Columns per data table	4,096
Partitions per table	10,000
Partitions per root account	1,000,000
Fields per table	4,096
Custom functions per account	100
Catalogs that can be created	20

### Database

Name: It can contain up to 127 characters and must be unique in a data link.

Description: It can contain up to 2,048 characters.

Data address of an external table (COS address): It can contain up to 888 characters (COS path length limit).

Parameter: It is in the `Map<string:string>` format and can contain up to 127 characters. All parameters can contain up to 3,000 characters in total.

### Data table/View

Name: It can contain up to 127 characters and must be unique in a database.

Description: It can contain up to 1,000 characters.

Data address of an external table (COS address): It can contain up to 888 characters (COS path length limit).

Parameter: It is in the `Map<string:string>` format and can contain up to 127 characters. All parameters can contain up to 512,000 characters in total.

### Attribute column

Name: It can contain up to 127 characters and must be unique in a data table.

Description: It can contain up to 256 characters.

Field value: It can contain up to 131,072 characters. Longer values cannot be created.

## **Partition**

Partition field name: It can contain up to 127 characters.



# Computing Task

Last updated : 2022-08-16 09:42:00

- A single SQL statement cannot exceed 2 MB in size.