WeData Data Development Platform FAQs

Copyright Notice

Trademark Notice

Service Notice

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

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

Contact Us

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

Contents

FAQs
Basics

Last updated: 2024-08-23 17:35:27

What is Tencent Cloud WeData?

Tencent Cloud WeData is a one-stop data collaboration development governance platform created by Tencent Cloud.

What are the common scenarios of WeData?

Common scenarios of WeData include enterprise data warehouse construction, collaborative data visualization development, data asset management, and data governance. For more details, see Scenarios.

What product capabilities does WeData provide?

WeData offers full-link data processing capabilities including ad-hoc analysis, workflow collaborative orchestration, data asset management, and data governance.

WeData Product Architecture and Advantages?

As a big data development and governance platform, WeData aims to provide open source, easy to use, and secure and stable product services. The overall architecture consists of an operational management system and development and operation tools. For details, please refer to Product Architecture.

What types of data sources does WeData support?

Supports various on-cloud and self-built data sources, such as:

- Relational Database: MySQL, PostgreSQL, Oracle, SQL Server, IBM DB2, Dameng DM, SAP HANA, etc.
- Big Data Storage: Hive, HDFS, HBase, Kudu, Clickhouse, DLC, Impala, Gbase, Tbase, etc.
- Semi-structured: FTP, SFTP, COS.
- NoSQL:Redis,Elasticsearch.
- Message Queue: Kafka, etc.

Does WeData support private deployment?

WeData supports private deployment. You can Submit a ticket to contact us.

Usage

Last updated: 2024-08-24 18:08:06

Which sub-accounts can authorize CAM roles?

Policy name	Policy Description	Function
AdministratorAccess	This policy allows you to manage all users and their permissions, financial-related information, and cloud service assets within the account	• Can manage sub-account permissions • Can log in to WeData Console • Can create WeData Project
QCloudResourceFullAccess	This policy allows you to manage all cloud service assets within the account	• Can log in to WeData Console • Can create WeData Project
ReadOnlyAccess	This policy allows you read-only access to all cloud service assets within the account that support interface-level or resource-level authentication	• Can log in to WeData Console • Cannot create WeData Project

How to obtain/view EMR VPC?

Go to Tencent Cloud EMR Console, select the region and ID corresponding to the cluster. Enter the **Cluster Detail Page > Basic Information > Instance information** page to view the cluster network information.

Note

The network information of EMR by default uses "--" to separate the ID and display name. You can copy the information before "--" (inside the red box) for search.

How to obtain/view the data source VPC?

Taking MySQL as an example, go to the corresponding Tencent Cloud MySQL Console, select the region and ID of the corresponding instance. Enter the **Instance Detail Page > Instance Details** page, and view the instance network information.

When synchronizing the whole MySQL database, click "Pause" immediately after running, the task status always shows "In operation"?

Cause Analysis: After the whole database synchronization task starts running, clicking **Pause** immediately will cause the task status to remain "In operation". Backend data is continuously synchronizing, far exceeding the actual size of the source table.
Solution: When the MySQL whole database synchronization task is in the snapshot phase, the "Pause" operation is not supported.

After clicking "Continue" on a "Paused" task, some data is duplicated?

Cause Analysis: In the current version, resuming a real-time task after it was paused will start from the last checkpoint, so partial data duplication is a normal occurrence.

During real-time synchronization, after deleting the data source or changing the data source structure, the task shows "Running" but is not actually synchronizing?

Cause Analysis: During real-time task operation, if the following scenarios occur, the task still shows as running but is not actually synchronizing any data.

- Deleting the data source in the MySQL source database. For example, during synchronization from MySQL to Hive, deleting the MySQL or Hive table in the MySQL source database.
- Changing the data source structure in the MySQL source database. For example, during synchronization from MySQL to Hive, deleting or modifying certain fields in the MySQL or Hive table in the MySQL source database.
  Both operations of deleting the data source and changing data structures are considered abnormal issues. WeData temporarily cannot obtain the abnormal status of the task.

Solution: After obtaining the abnormal logs, the solutions for the above two situations are as follows:

1. Delete the data source. Users need to reselect a new data source on the WeData real-time synchronization interface, configure the mapping relationship, and then click **submit**, choosing "Keep job status data, continue running".
2. Changing the data source structure. Users need to click **Refresh** in the WeData Real-time Synchronization Interface to obtain the latest data source structure, configure the mapping relationship, then click **Submit**, and choose "Retain Job Status Data, Continue Running".

When synchronizing MySQL sharded tables with different schemas, the task shows "Running" but the task is not successfully synchronized?

Cause Analysis: When synchronizing MySQL sharded tables, if the schemas of the configured tables are different and the fields establishing the mapping relationship do not exist in certain tables, the task status will show "Running" but will not actually execute successfully. Additionally, since Oceanus currently uses a strategy of continuous restarting when a task encounters an error, the task remains in a "Running" state.
Solution: For the unsuccessful scenario mentioned above, exclude non-compliant tables in the WeData Real-time Synchronization Interface, configure the mapping relationship, then click **Submit**, and choose "Retain Job Status Data, Continue Running".

When performing MySQL whole database synchronization, the "Oceanus job config length exceeds 100k limit" warning appears?

Problem Description: When performing MySQL whole database synchronization, the "Oceanus job config length exceeds 100k limit" warning appears.
Cause Analysis: To ensure the performance of the whole database migration task, the total length of the task configuration cannot exceed 100k in the current version. If too many data tables or the total number of table fields are too large, the system prompts a limit warning.
Solution: It is recommended that a single whole database migration task does not configure more than 20 tables. It can be divided into multiple whole database migrations or sharded tasks.

Does the Definition field configuration support Complex Types?

The current version does not support Complex Types such as Array, Map, and Struct, only Primitive Types.

An error occurs when submitting the task: "ache.calcite.sql.validate.SqlValidatorException: Cast function cannot convert value of type ..."?

Problem Description: When submitting a real-time synchronization task, the system reports the error "[TencentCloudSDKError] Code=FailedOperation, Message=ache.calcite.sql.validate.SqlValidatorException: Cast function cannot convert value of type ...".
Cause Analysis: The types between the source and target fields of real-time synchronization cannot be converted.
Solution:

1. Please check whether the field types configured between the source and target ends match. For specific type conversions between data sources and target ends, refer to Synchronized Data Source.
2. If the task is a MySQL-Ckafka single table synchronization task, check whether the mapping relationship for pkNames_wedata_di and sqlType_wedata_di with the target end is established. Since the current version does not support Complex Types such as Array, Map, and Struct in Ckafka fields, the pkNames_wedata_di and sqlType_wedata_di parameters cannot be written to the Ckafka target end.

Sub-account prompts "No PassRole" when redirected to Tencent Cloud Oceanus Console through operation and maintenance?

Solution: Please refer to the Oceanus Permission Management document.

Data Integration

Last updated: 2024-11-25 09:31:41

1. How to handle ClickHouse to DLC sync reporting network timeout

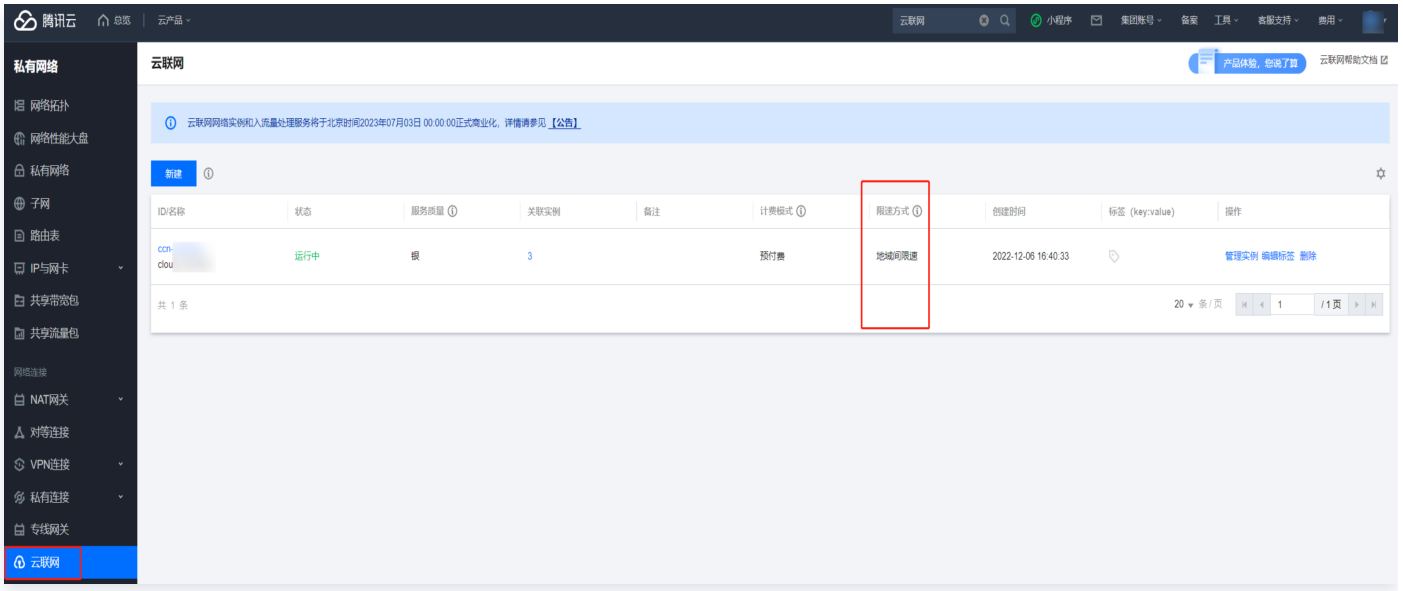
Problem description:
When users sync ClickHouse data to DLC, it succeeds with small datasets, but reports network timeout with larger tables.

```
at com.***.***.plugin.rdbms.util.DBUtil.query(DBUtil.java:470)
at com.***.***.plugin.rdbms.util.DBUtil.query(DBUtil.java:431)
at com.***.***.plugin.rdbms.reader.CommonRdbmsReader$Task.startRead(CommonRdbmsReader.java:196)
at com.***.***.plugin.reader.clickhousereader.ClickhouseReader$Task.startRead(ClickhouseReader.java:72)
at com.***.***.core.taskgroup.runner.ReaderRunner.run(ReaderRunner.java:59)
at java.lang.Thread.run(Thread.java:748)
[2022-04-14 21:16:08]-[INFO] Total 0 records, 0 bytes | Speed 0B/s, 0 records/s | Error 0 records, 0 bytes
[2022-04-14 21:16:08]-[INFO]

该任务最可能的错误原因是:
SocketTimeoutException: Read timed out
[2022-04-14 21:16:08]-[ERROR] Task execution failed
[2022-04-14 21:16:08]-[INFO] commitTask success.
[2022-04-14 21:16:08]-[INFO] taskId: <20[redacted]2>
```

Locating the Issue:

1. Logging into the execution environment POD, Telnet to the target ClickHouse datasource connection shows connectivity, and logging in is also successful, thus ruling out network issues.
2. It is suspected that large ClickHouse data volume or high concurrency leads to read timeout. Further reduce concurrency in CCN.



3. If the same error persists after limiting concurrency and increasing the data read operation timeout in ClickHouse and still encounter errors, further investigation is required to ensure that the customer’s execution environment resource group and ClickHouse VPC are consistent. If not, the network must be connected via CCN (Cloud Connect Network).
4. Confirm the required public network bandwidth for cross-regional network connectivity with the CCN product, and verify whether the user has connected the two VPCs across regions via CCN but without enabling public network bandwidth.

5. For testing purposes, CCN defaults to 10Kb of traffic, explaining why small data syncs and network tests are normal, but large data volumes report network issues.

6. **Cause:**

   If the execution resource group and datasource are not in the same VPC and cross-region VPC connection via CCN is without enabling public network bandwidth, default CCN supports less than

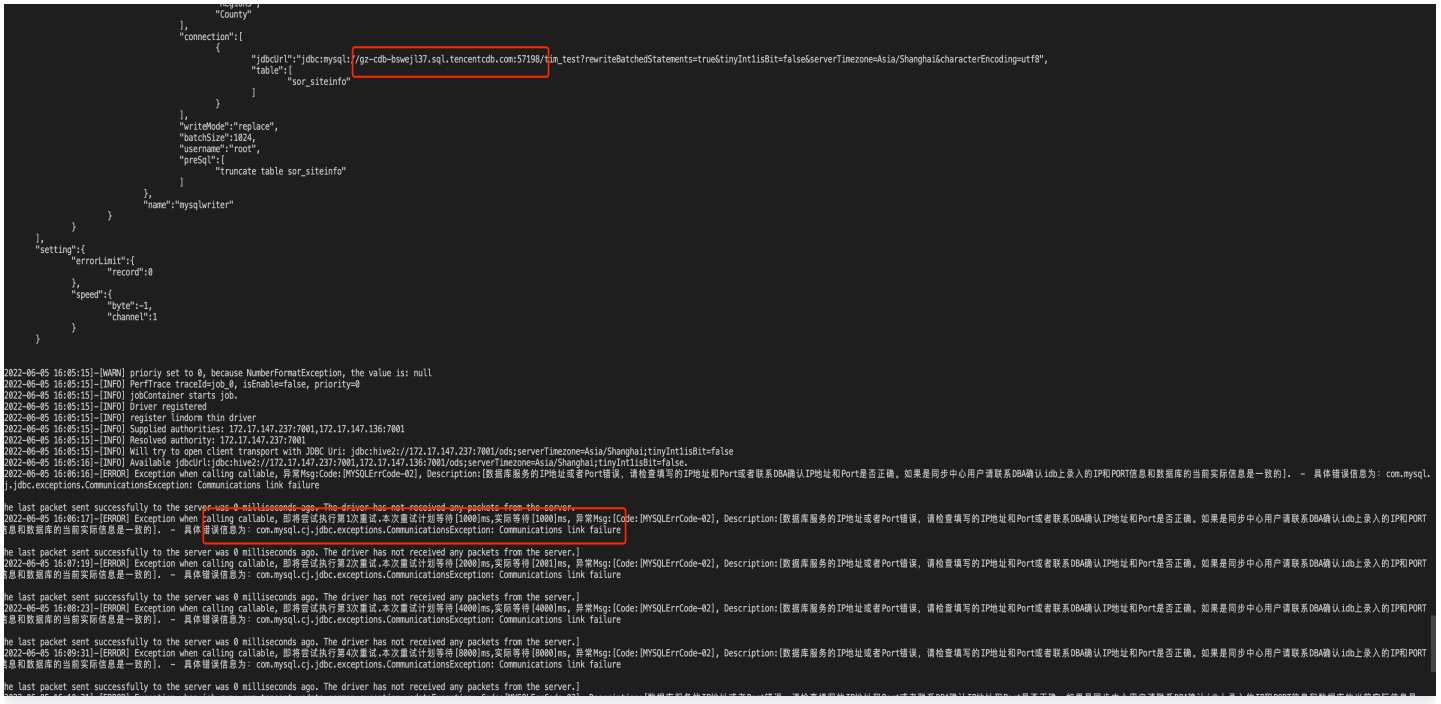
10Kb of traffic, exceeding which causes a disconnection, making small data volumes feasible but large data volumes fail.

7. **Solution:**

   Solve the issue by adding public network bandwidth via CCN.

2. How to handle issues when requiring public network access but network is disconnected

Problem description:



Locating the Issue:

Locate through logs.

Cause:

The data source uses a public network, and the resource group's network does not default to public.

Solution:

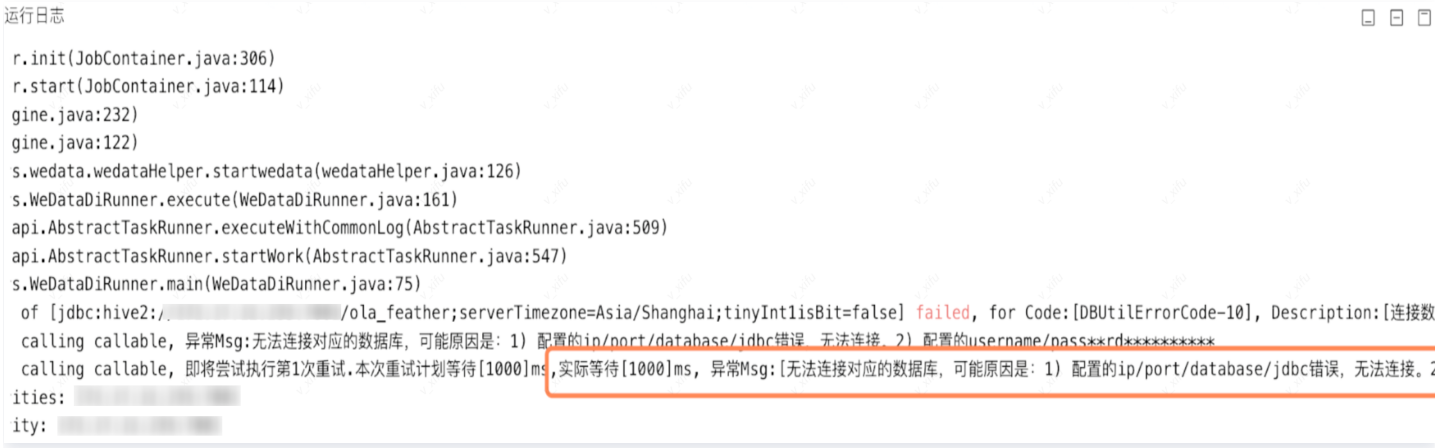
Add the subnet of the resource group to the NAT Gateway. For details, refer to NAT Gateway Configuration and Access the Internet through NAT Gateway .

> **Note**
>
> If it is the main account's resource group, you can replace the resource group's subnet routing policy with EKS.

3. How to handle data synchronization failure when it prompts that the data source is not accessible

Problem description:

Offline synchronization prompts that the database cannot be connected, but the database is actually accessible.



Locating the Issue:

The network of the WeData DataInLong resource group and CDB are not under the same VPC and are not interconnected.

Cause:

The DataInLong execution resource group and data source are under the user's VPC network. They need to be in the same VPC network, otherwise, the network will not be accessible, and the execution resource group cannot synchronize data correctly. If they are in different VPC networks.

Solution:

Use CCN or Peering Connection to enable cross-VPC network intercommunication. If the data source is a public network instance, configure the NAT Gateway.

4. How to handle Hive On COS table permission error during data write

Problem description:

Offline synchronization, hive data source, table data stored in COS, error during data write: java.lang.Exception: Retrieve the file metadata file failure.

Locating the Issue:

Check if the account is associated with the WeData_QCSRole role and if this role has the COS CAM policy configured.

Cause:

When integrating tasks for hive on cos, write data through the WeData_QCSRole role to obtain temporary credentials from CAM. If the COS-related CAM policy is not added under the WeData_QCSRole role, the obtained temporary credentials will not have permission to read/write the COS bucket.

Solution:

On the role page of CAM, search for "wedata", find WeData_QCSRole, and check if it includes COS in the associated policies. If not, add QcloudCOSFullAccess.

Data Development

Last updated：2025-04-09 14:28:37

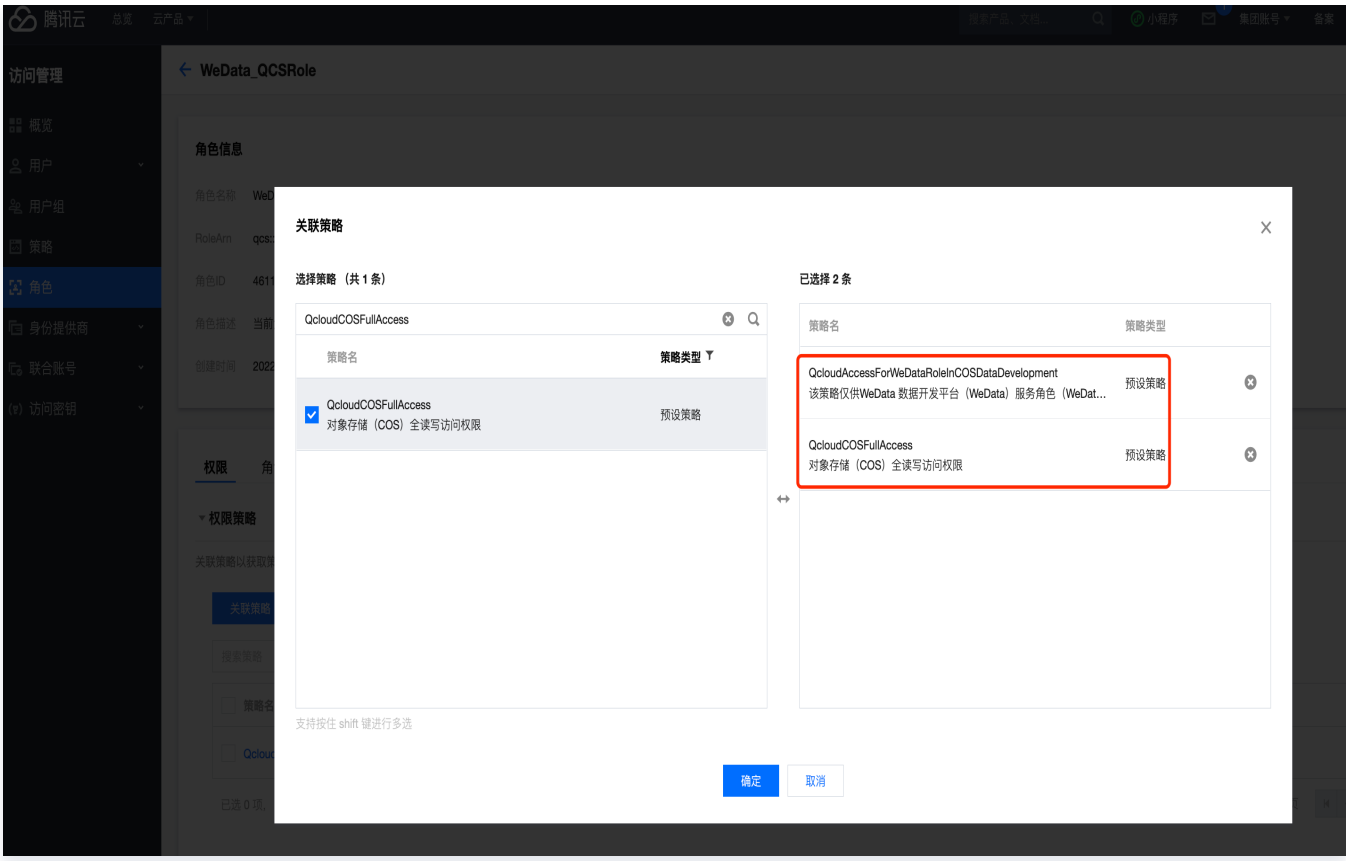
Unable to Select COS Bucket for Resource Upload

1. issue phenomenon

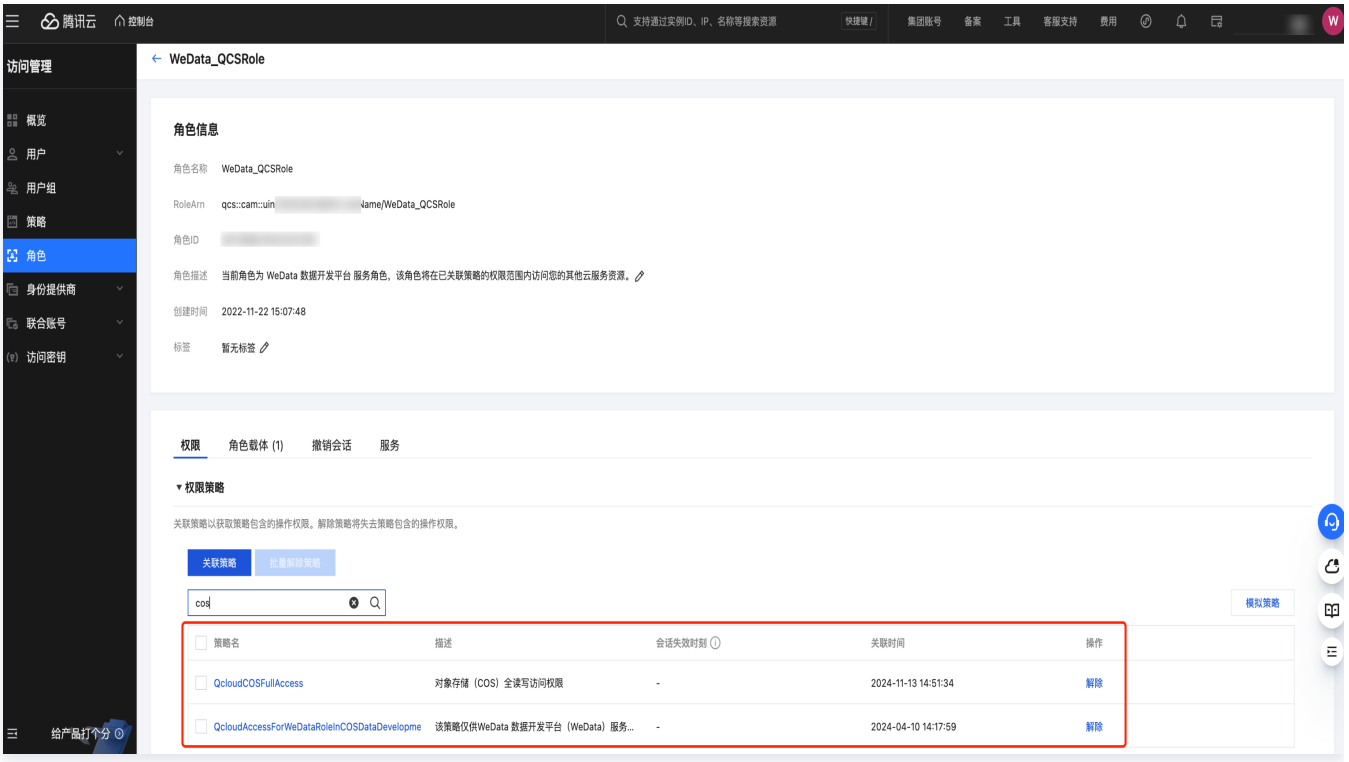
2. Solution

2.1 Enter Tencent Cloud Cloud Access Management, select the **role** menu, and search for WeData_QCSRole.

2.2 1. Click the role name **WeData_QCSRole,** select the associated policy, search for policies "QcloudAccessForWeDataRoleInCOSDataDevelopment" and "QcloudCOSFullAccess", and confirm the association.



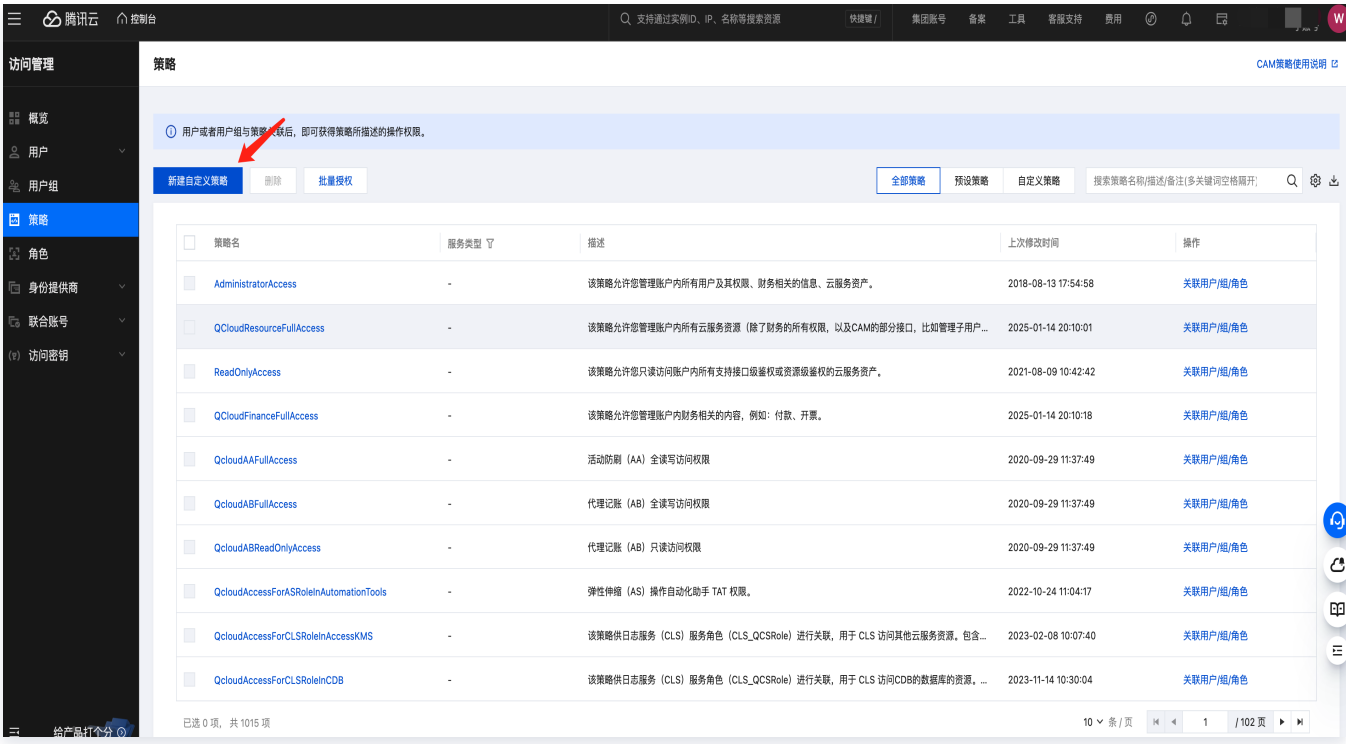
After successful association, this policy can be seen under the permission policy. Try uploading resources to the COS bucket again and it will succeed.



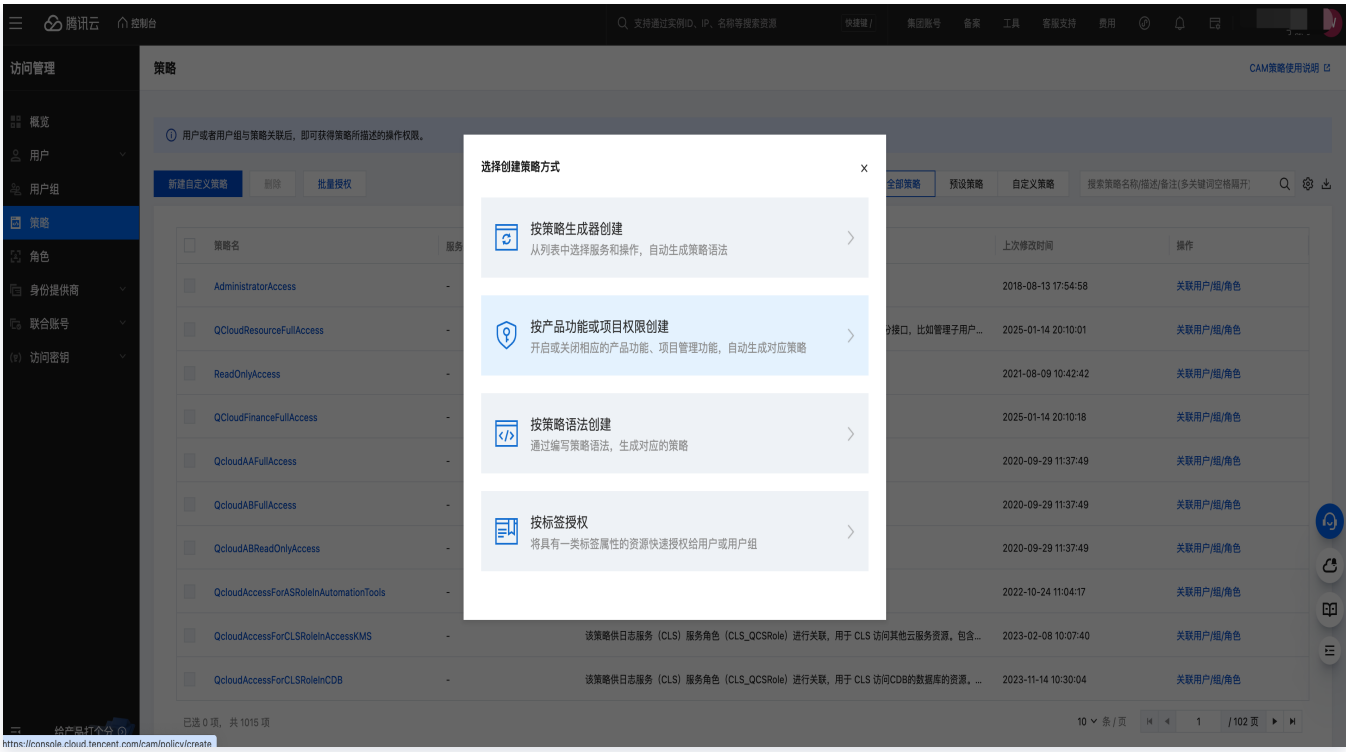
3. Supplementary solution

If you are concerned that the excessive permissions of COSFullAccess granted in the previous step, you can see the following operations to grant WeData only read access to a certain bucket specifically used for development.

3.1 Create a custom policy



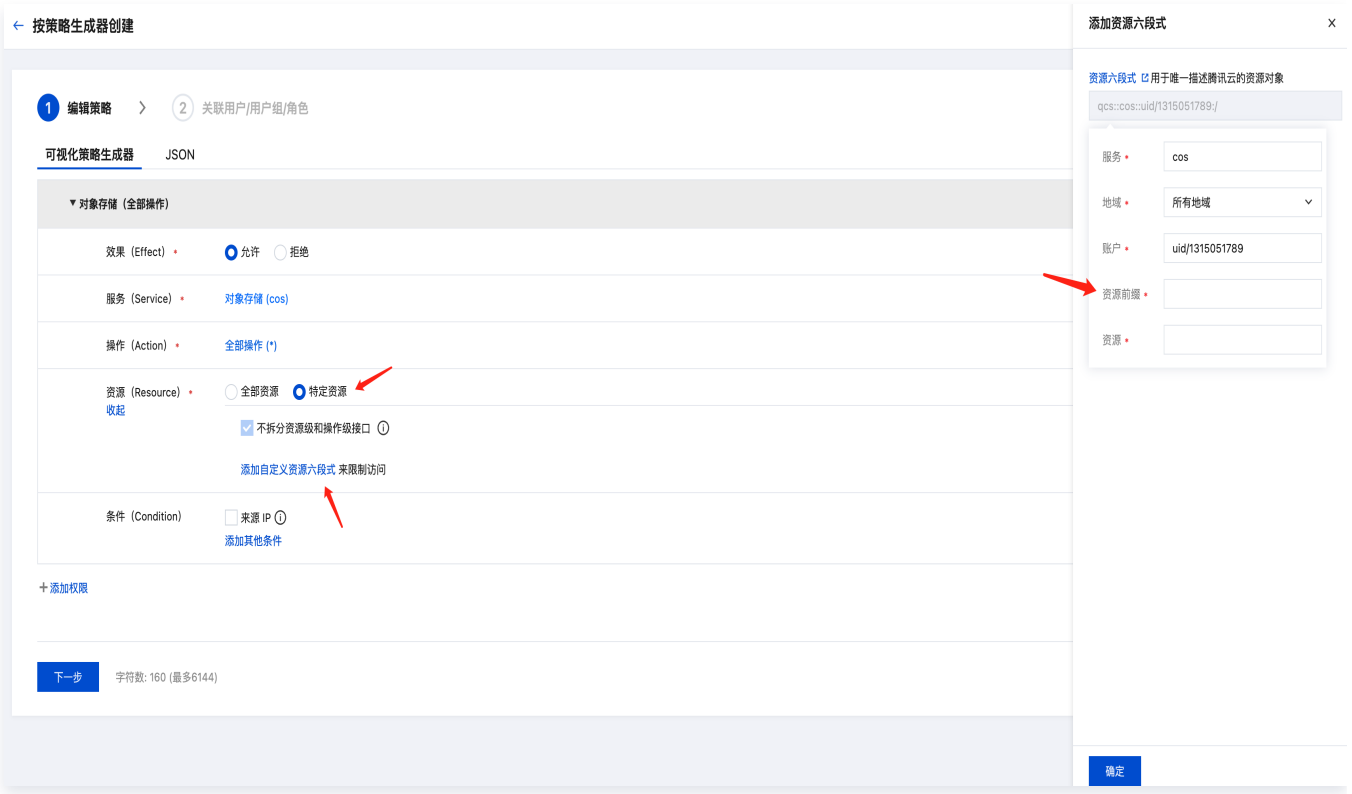
Select **create by policy generator.**



3.2 Generate a COS Policy

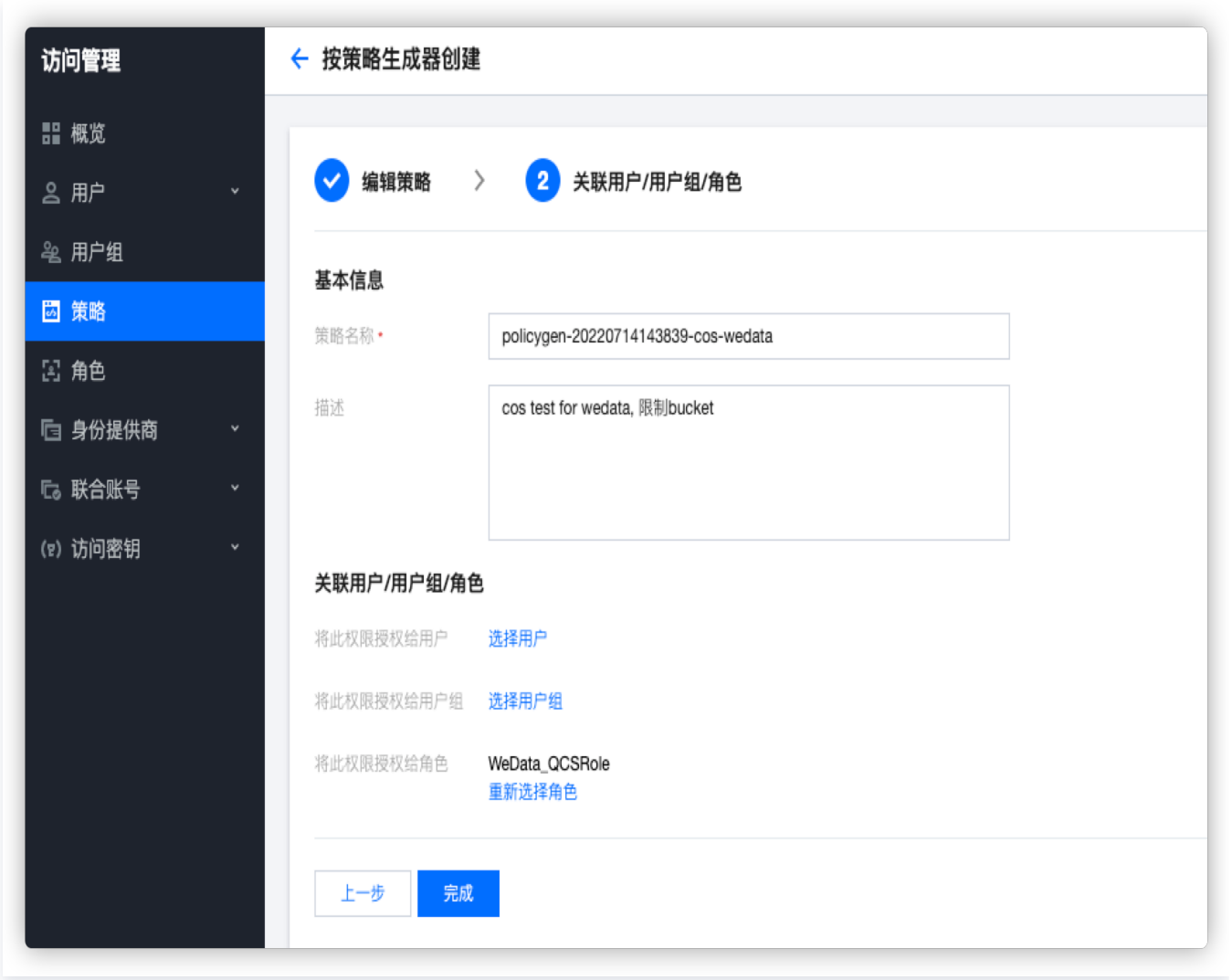
The focus is on limiting the specified cos resource, simultaneously expanding the permission of the drop-down list, the location of specific resources, and filling in the COS bucket provided to

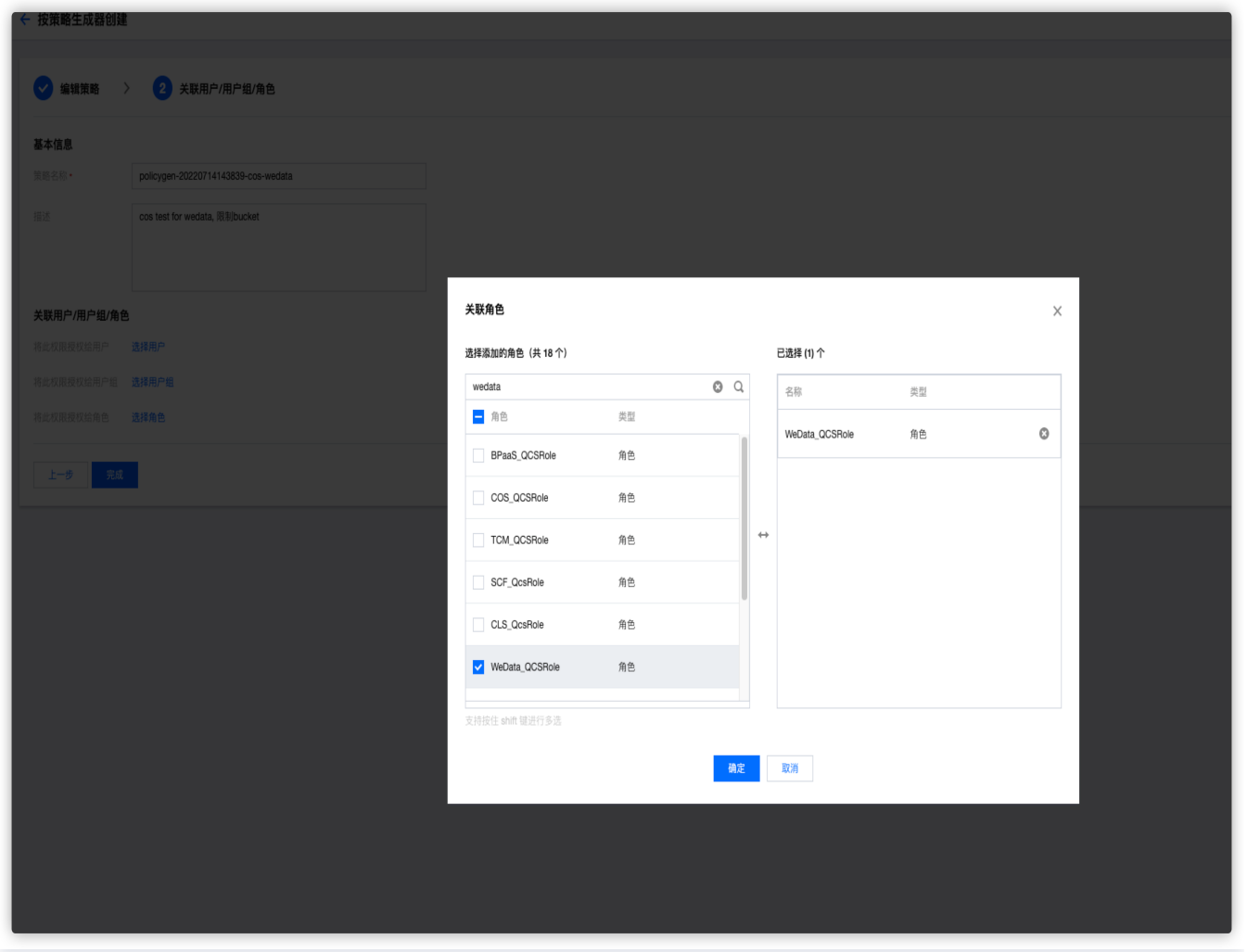
WeData Development Engineer. Fill in other content by referring to the screenshot.



The completed Summary Strategy Map is as follows:

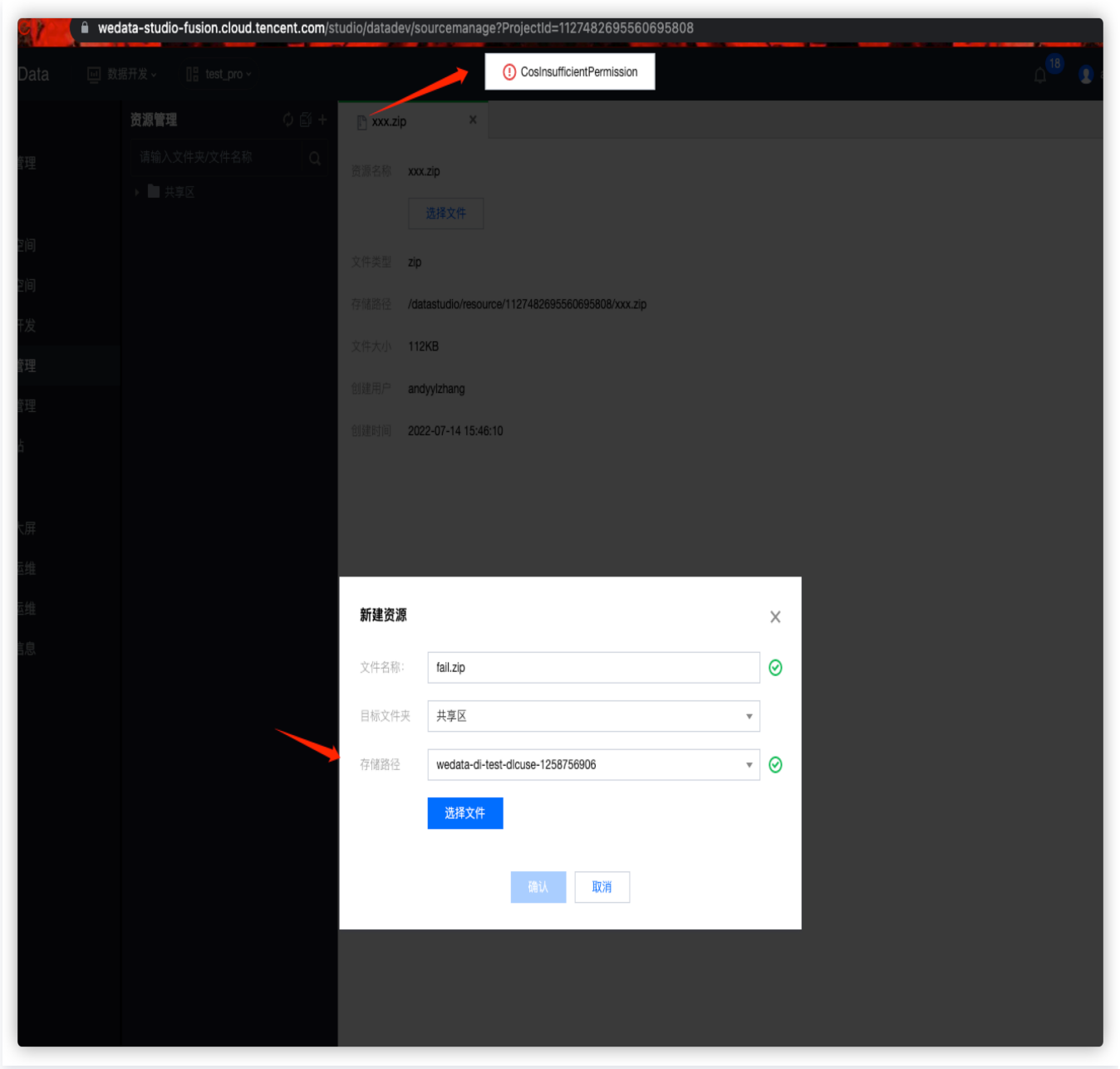
3.3 Save policy, associated with the role of WeData.



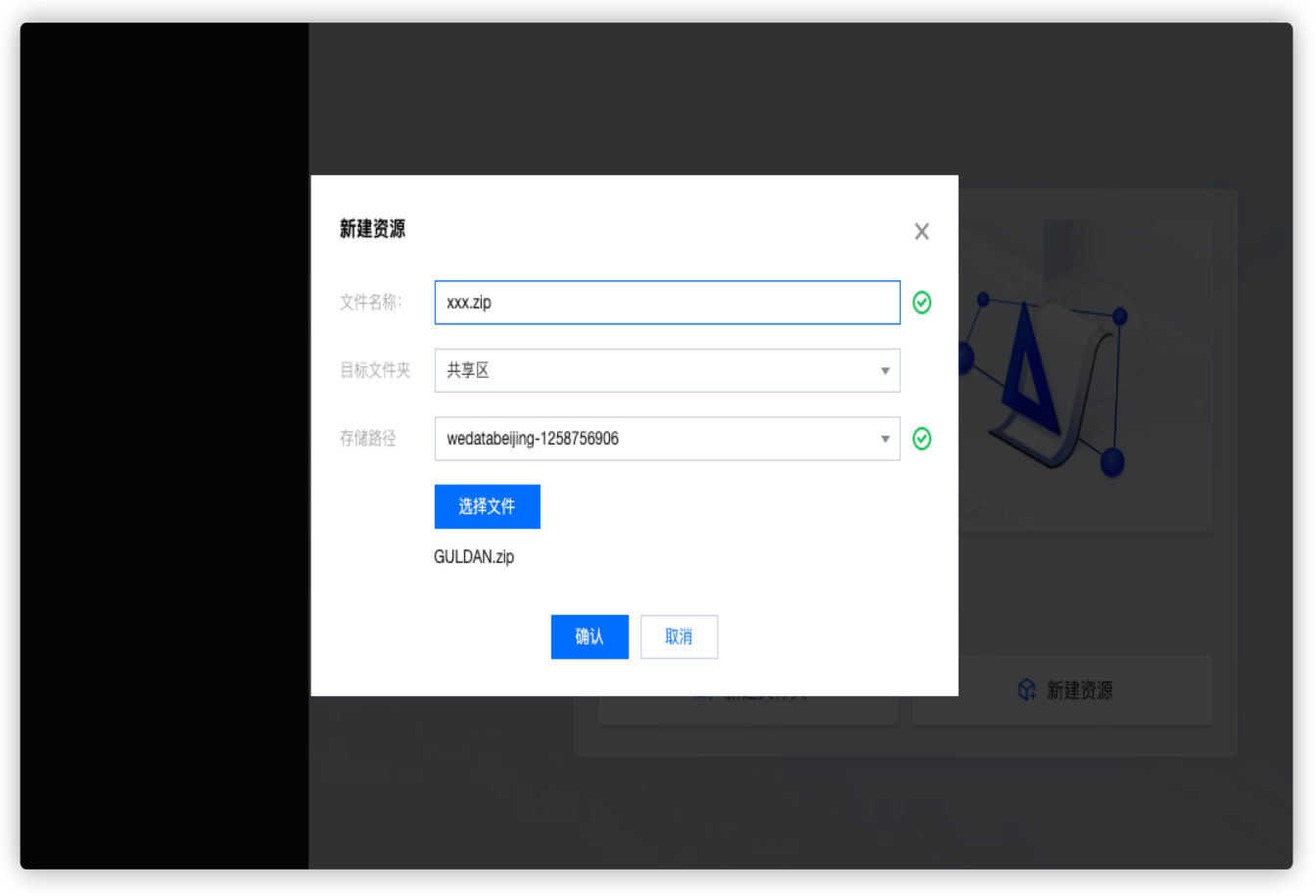


3.4 Test WeData

The bucket dropdown list still shows ALL buckets. However, if you select a bucket without authorization, an error message indicating no permission will appear.



If you select an authorized bucket, you can operate.



3.5 Final Generated Strategy Text for reference

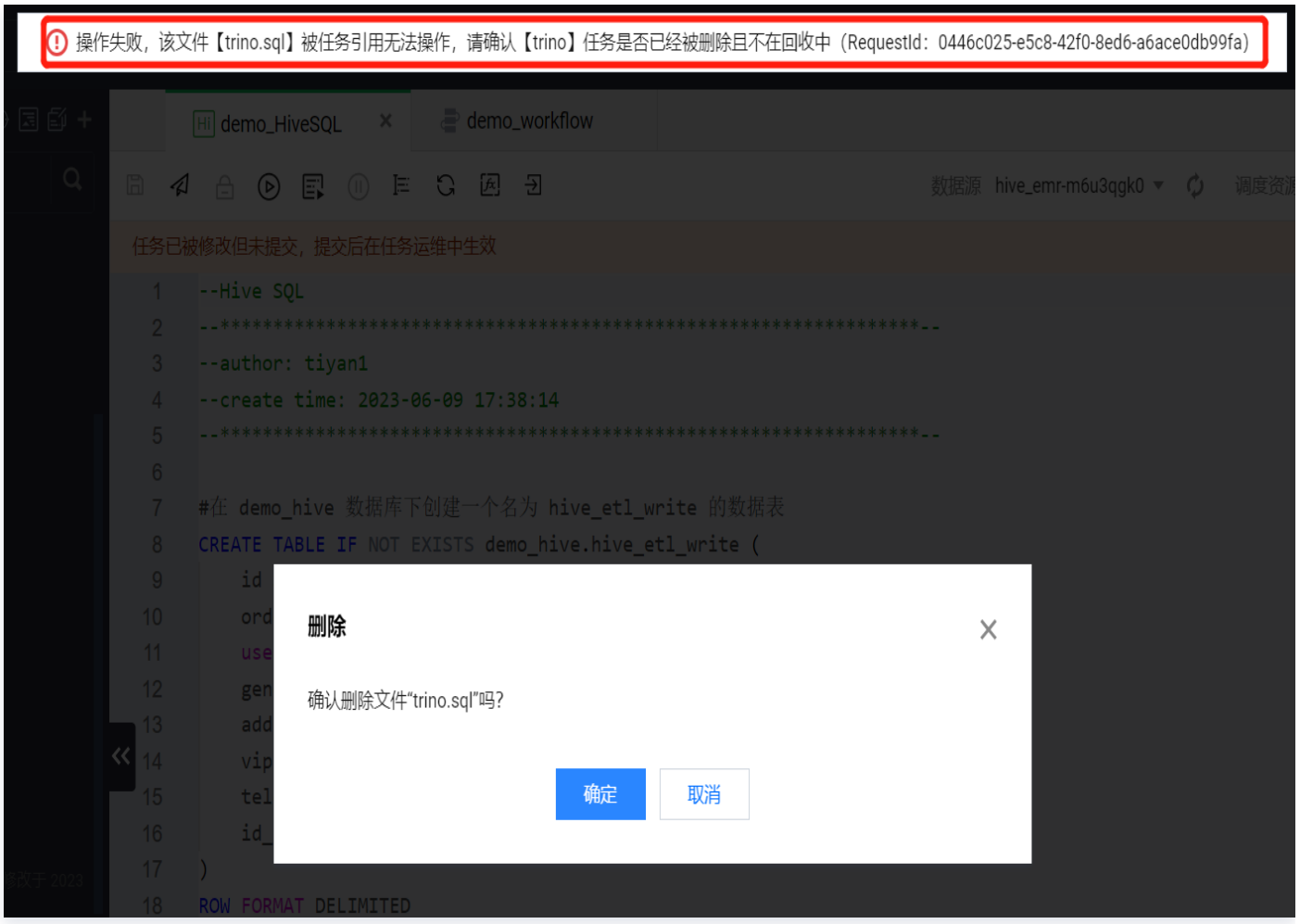
Directly create policies by filling in text. Without page operation, note that parameters in the resource part need to be modified.

```
{
    "version": "2.0",
    "statement": [
        {
            "effect": "allow",
            "action": [
                "cos:*"
            ],
            "resource": [
                "qcs::cos::uid/1258756906:wedatabeijing-1258756906/*"
            ]
        },
        {
            "effect": "allow",
            "action": [
                "cos:*"
            ],
            "resource": [
               "*"
```

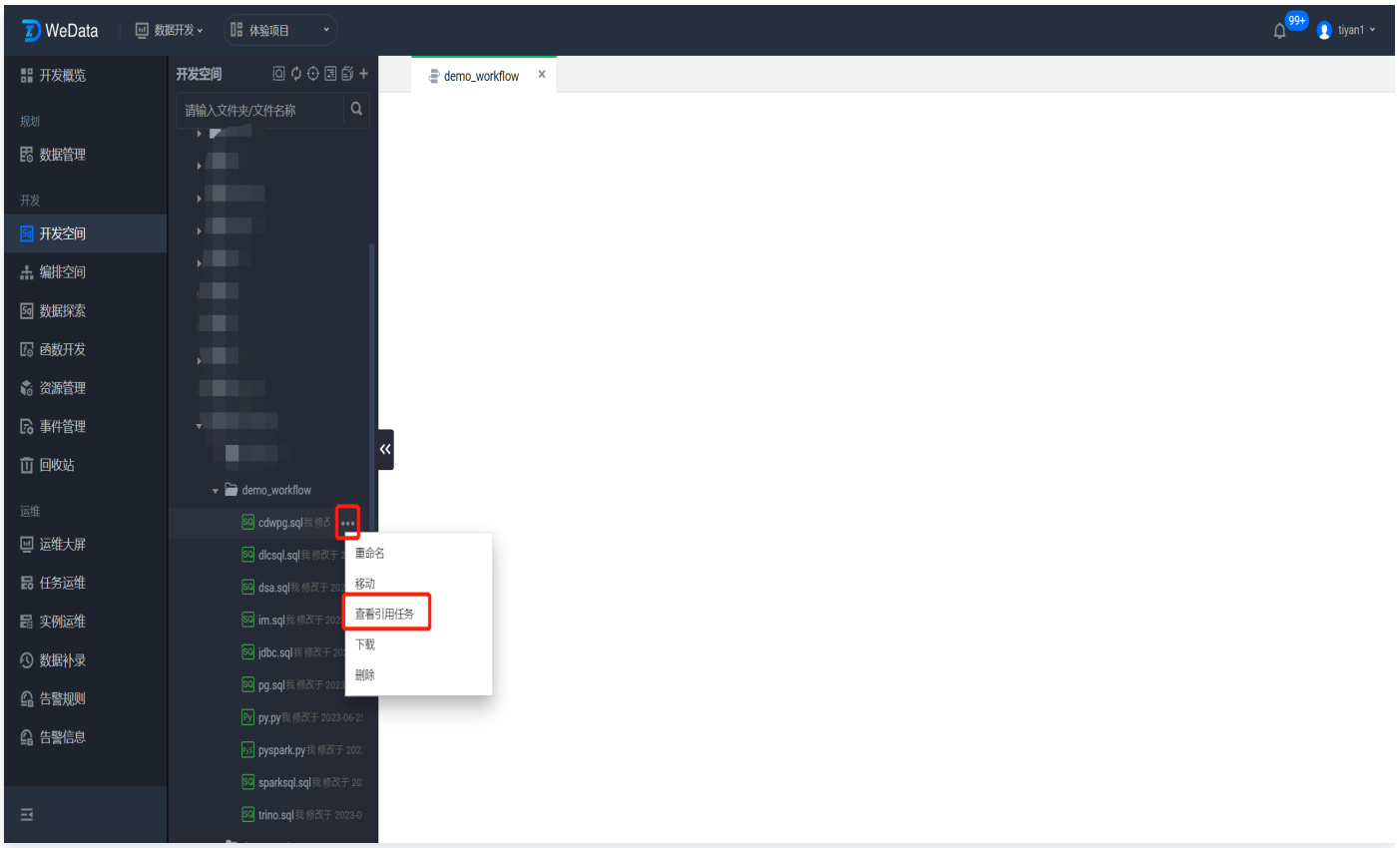
```
            ]
        }
    ]
}
```

Clicking on the Develop Space Task Cannot Be Deleted, It Is Referenced

1. issue phenomenon



2. Solution 1: Select the undeletable development task in the Develop Space directory, open the operation menu of the task, and click **View Referenced Tasks**. You can see the orchestration workflow information that refers to this task. Find the corresponding workflow according to the information. After corresponding deletion in the workflow, it can be normally deleted in the Develop Space.



3. Solution 2: According to the task name in the pop-up window, search and delete in the orchestration space and recycle bin.
   3.1 If the task is in the orchestration space, you can delete the task in the orchestration space at this point and check the delete script option.
   3.2 If queried in the recycle bin, delete the task in the recycle bin and then the development space script can be deleted.
   3.3 If it cannot be queried in the recycle bin either, you need to check whether the responsible person of this task is yourself. Currently, permission control has been implemented for task display in the recycle bin. You can only see tasks whose responsible person is the current logged-in account.

Task Creation Failure in the Orchestration Space, with a Notification That the Task Has the Same Name, but the Task Cannot Be Queried in the Orchestration Space

Solution: Query the script files in the development space and check whether there are existing files with the same name under the directory.

Development Task Debugging and Execution: Notification That No Available Nodes

1. issue phenomenon

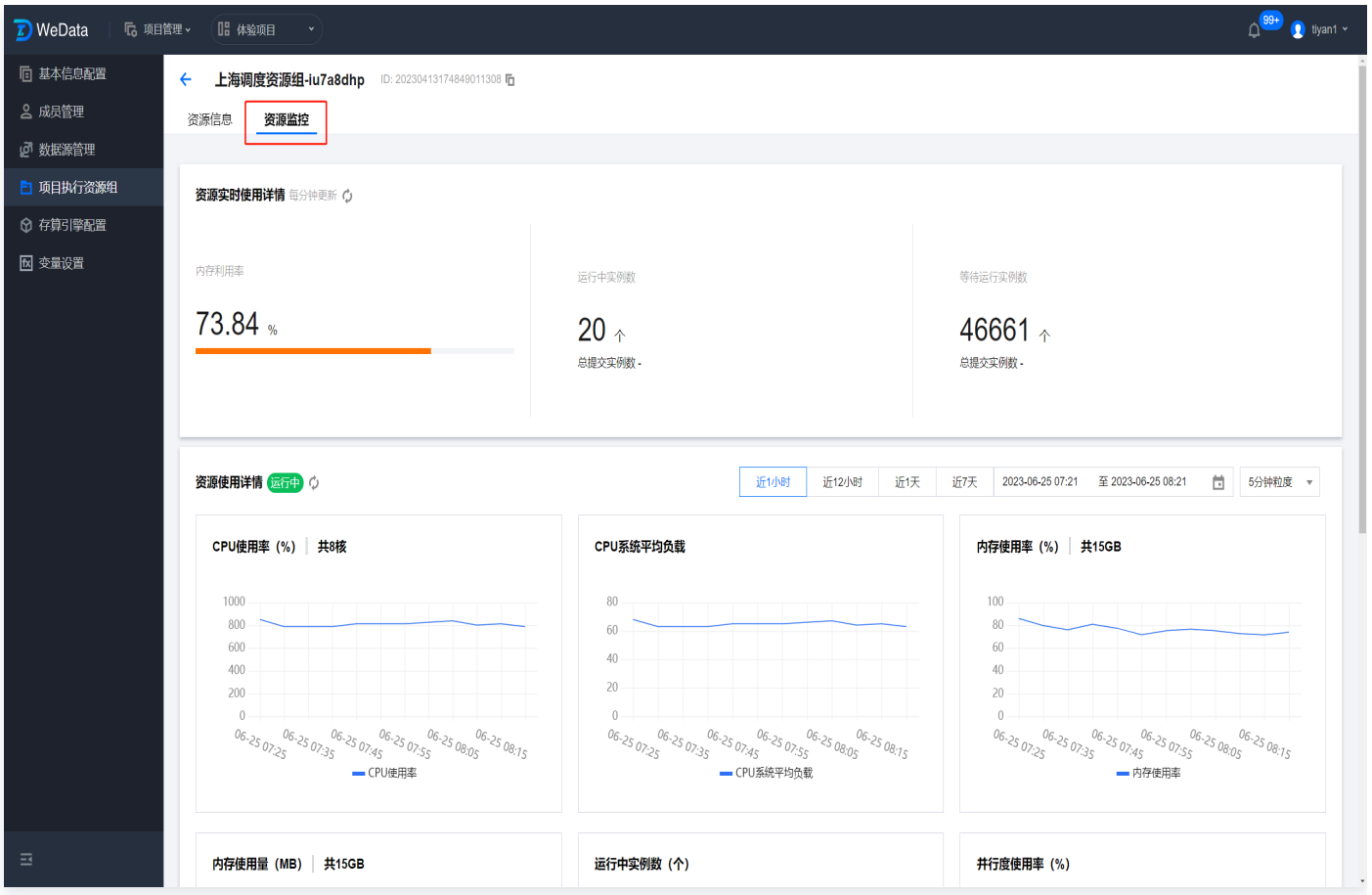
   When users are performing debugging and running of development scripts and task nodes, they see from the logs the task diagnostic information notification: No available nodes. At this point, it is

necessary to consider whether the scheduling resource group has reached saturation. Currently, the development task is unable to apply for scheduling resources, leading to a running failure.

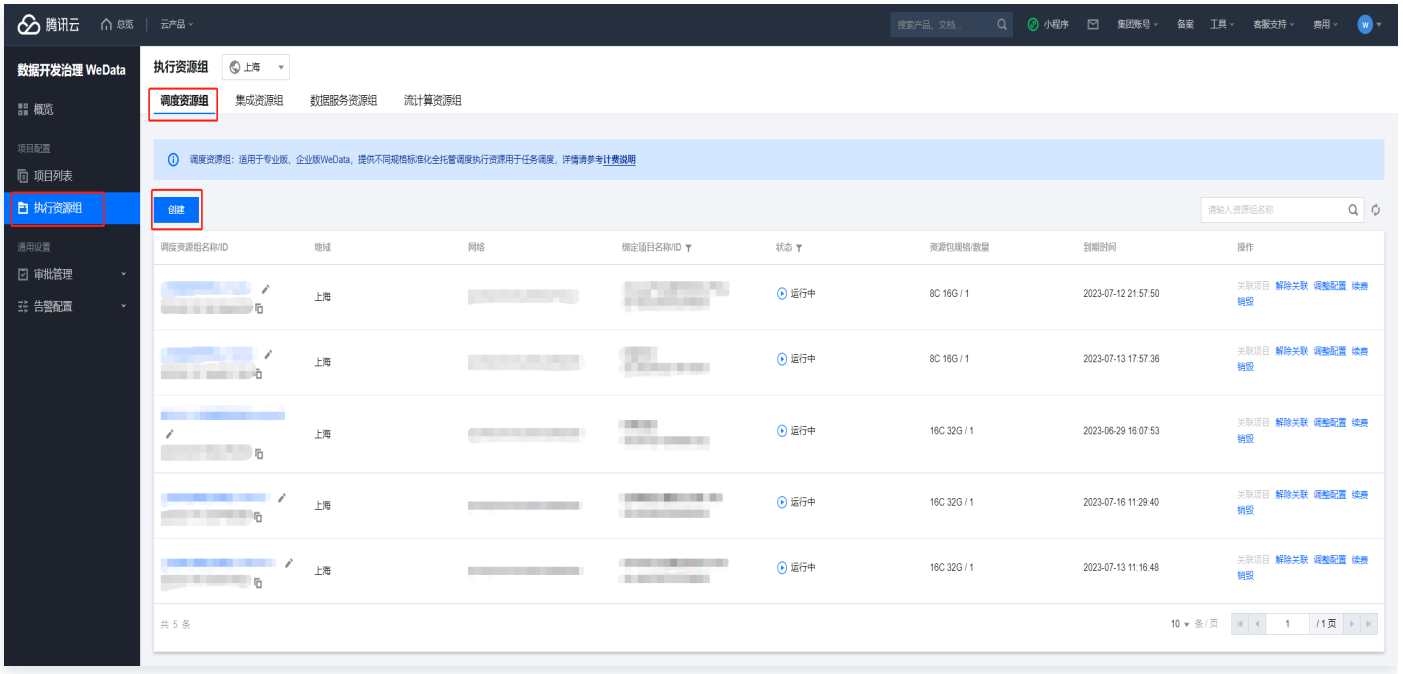
2. Solution

Enter **Project Management > Project Execution Resource Group**, find the selected scheduling resource group when the task is running, and click the resource group name to view the resource group usage status.

Through resource monitoring, you can see the current load information of the resource group. Based on monitoring contents such as the frequency of use of the resource group and high-concurrency time points, reasonably plan the scheduling cycle of development tasks, and you can reduce the occurrence frequency of node occupancy issues.

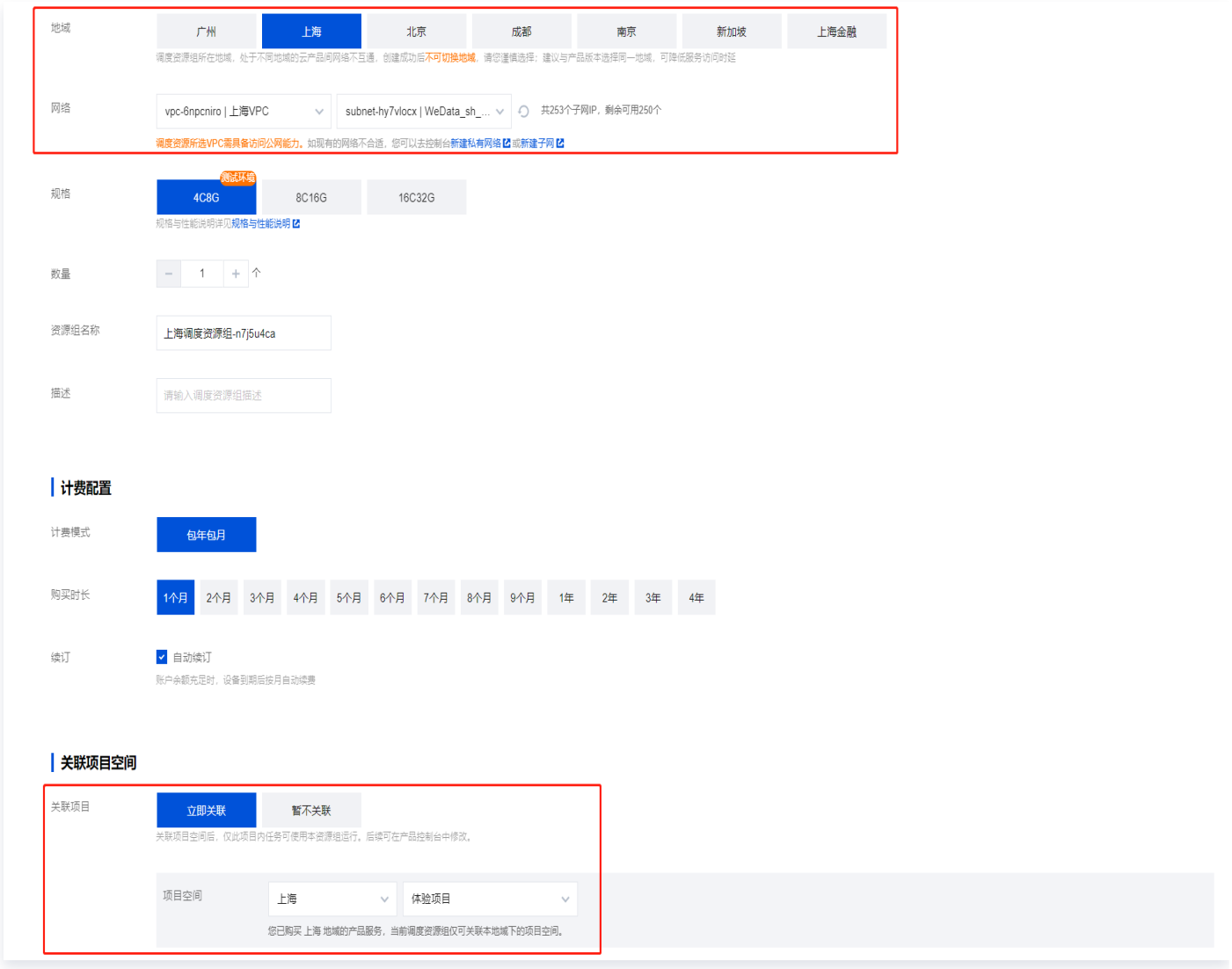


Additionally, you can also add new scheduling resource groups to alleviate load pressure. Enter console -> go to the Execution Resource Group page, and click Create under the Scheduling Resource Group tab to add new scheduling resource groups. For configuration methods, see Scheduling Resource Group Configuration.



Notably, during the configuration process of creating a scheduling resource group, it should be ensured that the region, network, associated project are consistent with the project where the

development task is located.

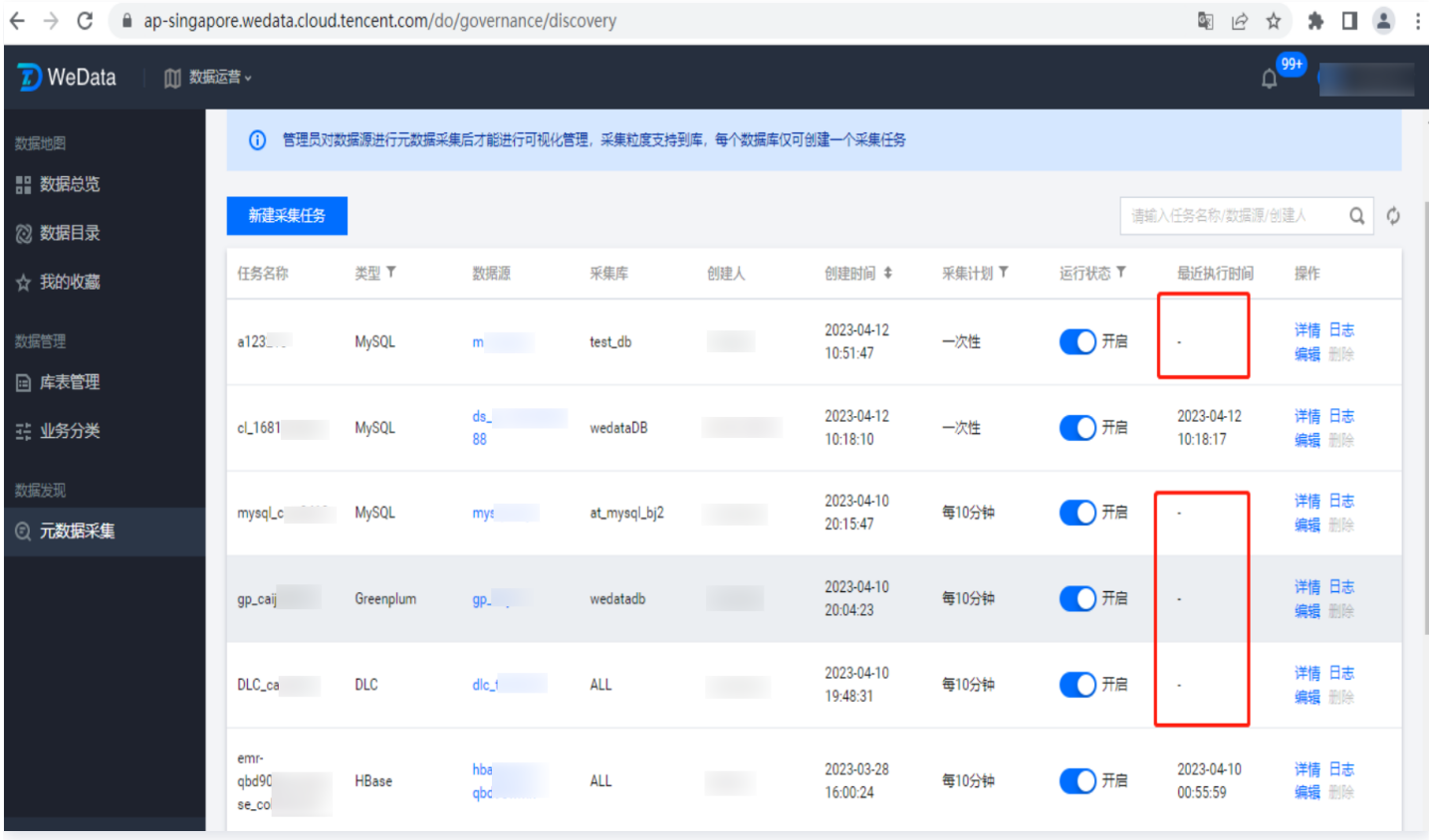


Data Operations

Last updated：2024-08-23 17:51:41

The newly created collection task did not execute when scheduled

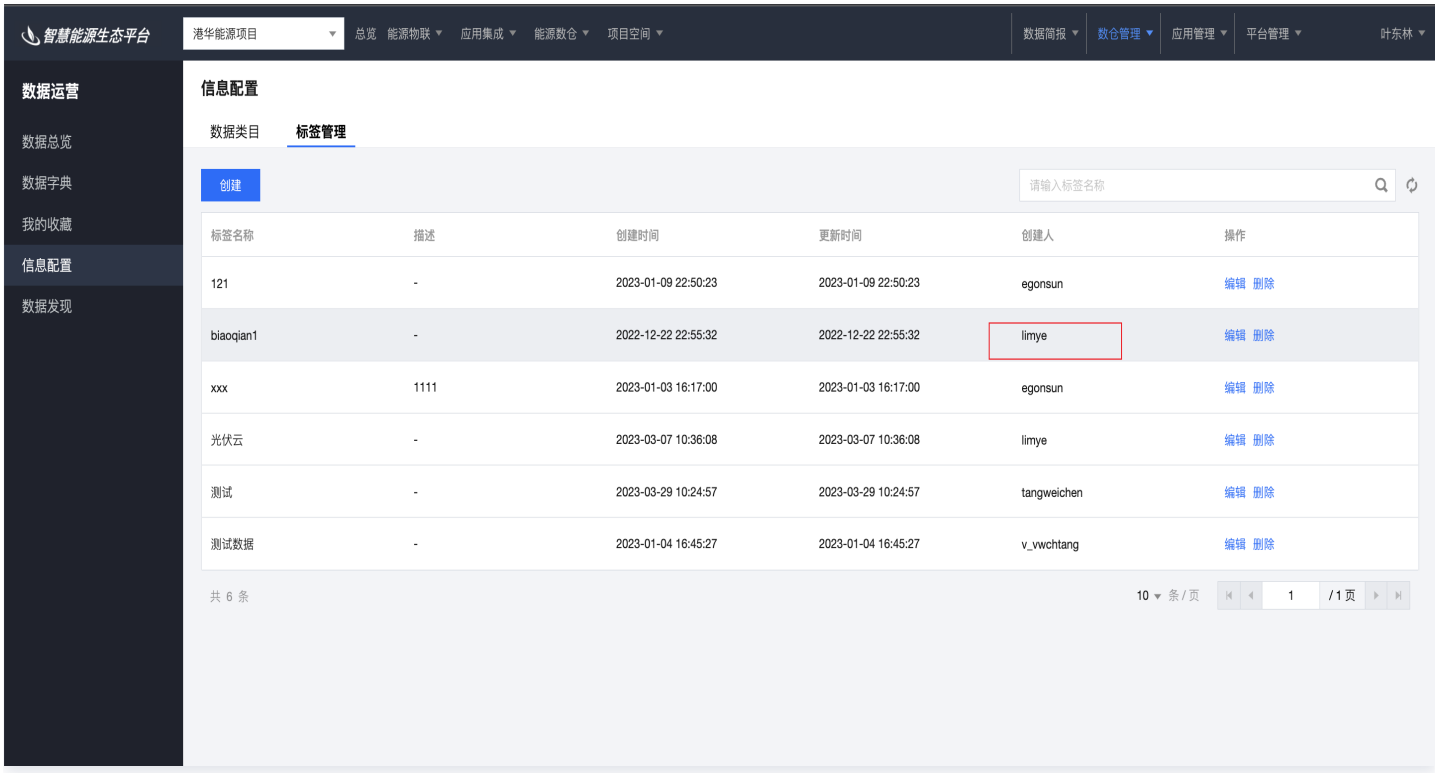
Symptom: The scheduled collection task did not execute.



Cause: Execution Resource Group configuration error.
Solution: Change the Execution Resource Group and try again.

The Business Information Tag cannot be deleted properly

Symptom: The Tag does not have a delete key and cannot be deleted.



Cause: Business logic issue.

Solution:

- When User A accesses the table, a box Tag is added to the table.
- When User B accesses the table again, they cannot delete the box Tag added by User A. They can only choose +1 (up arrow) or do nothing.
- After User B chooses +1, they can then choose −1 (down arrow).

Misinterpretation of information display in Data Panorama

The meaning of the number "0" in the circle node indicates that there are no more sub-directories under this directory. It is not related to the number of tables present in that directory.