DataArts Studio

FAQs

Issue	01
Date	2025-02-20

HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD.

Copyright © Huawei Cloud Computing Technologies Co., Ltd. 2025. All rights reserved.

No part of this document may be reproduced or transmitted in any form or by any means without prior written consent of Huawei Cloud Computing Technologies Co., Ltd.

Trademarks and Permissions

HUAWEI and other Huawei trademarks are the property of Huawei Technologies Co., Ltd.
All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei Cloud and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Contents

1 Consultation and Billing

1.1 How Do I Select a Region and an AZ?

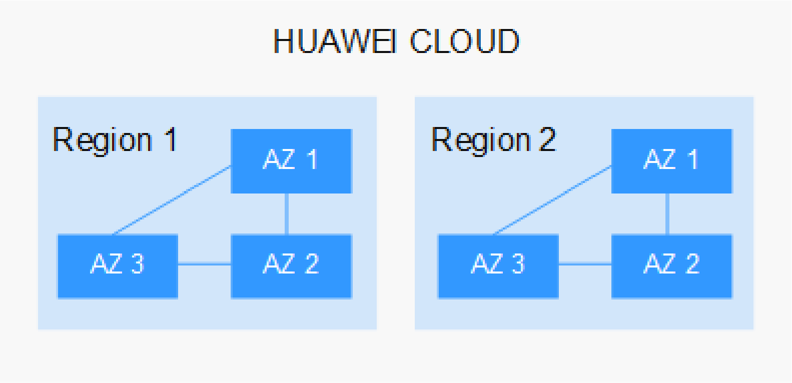
Concepts

We use a region and an availability zone (AZ) to identify the location of a data center. You can create resources in a specific region and AZ.

- Regions are divided from the dimensions of geographical location and network latency. Public services, such as Elastic Cloud Server (ECS), Elastic Volume Service (EVS), Object Storage Service (OBS), Virtual Private Cloud (VPC), Elastic IP (EIP), and Image Management Service (IMS), are shared within the same region. Regions are classified as universal regions and dedicated regions. A universal region provides universal cloud services for common tenants. A dedicated region provides services of the same type only or for specific tenants.
- An AZ is a physical location using independent power supplies and networks. Faults in an AZ do not affect other AZs. A region can contain multiple AZs, which are physically isolated but interconnected through internal networks. This ensures the independence of AZs and provides low-cost and low-latency network connections.

Figure 1-1 shows the relationship between the regions and AZs.

Figure 1-1 Regions and AZs



HUAWEI CLOUD provides services in many regions around the world. You can select a region and AZ as needed. For more information, see **Global Products and Services**.

Region Selection

When selecting a region, consider the following factors:

- Location

  Select a region close to you or your target users. This reduces network latency and accelerates access.

- Relationship between cloud services

  Cloud services in different regions cannot communicate with each other through an internal network.

  For example, if you want to enable communication between DataArts Studio (containing modules such as Management Center and CDM) and services in other regions (such as MRS and OBS), use a public network or Direct Connect. If DataArts Studio and the other services are in the same region, instances in the same subnet and security group can communicate with each other by default.

- Resource price

  Resource pricing may vary in different regions. For details, see **Product Pricing Details**.

AZ Selection

AZ to which the CDM cluster in the DataArts Studio instance belongs. The DataArts Studio instance communicates with other services through the CDM cluster.

When you buy your first DataArts Studio instance or incremental package, you can select any available AZ. When you buy another DataArts Studio instance or incremental package, determine whether to select the same AZ as that for the first instance or package based on your DR and network latency demands.

- If your application requires good DR capability, deploy resources in different AZs in the same region.
- If your application requires a low network latency between instances, deploy resources in the same AZ.

Changing the Region or AZ of an Instance

- During the validity period of your yearly/monthly DataArts Studio package, you can unsubscribe from the package in the current region and purchase a package in another region. **Unsubscribing from in-use resources** is available.
- You cannot change the region or AZ of an instance.

Regions and Endpoints

An endpoint is the **request address** for calling an API. Endpoints vary depending on services and regions. You can obtain the endpoints of the service from **Regions and Endpoints**.

1.2 What Is a Database, Data Warehouse, Data Lake, and Huawei FusionInsight Intelligent Data Lake? What Are the Differences and Relationships Between Them?

As the Internet and IoT continue to evolve, data management tools are developing rapidly to cope with massive data. As a result, concepts related to big data are emerging, such as database, data warehouse, data lake, and lakehouse. This section describes these concepts and their relationships, as well as the corresponding Huawei products and solutions.

Database

A database is a warehouse where data is organized, stored, and managed based on the data structure.

In a broad sense, databases have been used in computers since the 1960s. However, hierarchical and network models were prevailing at that time, and they lacked structural independency between data and applications. As a result, database application was limited.

Nowadays, a database usually refers to a relational database. A relational database organizes data based on relational models and stores data in a set of tables with columns and rows. Therefore, data is well-structured and independent with low redundancy. In 1970, relational databases were born to completely separate data from applications in software and since then have become an indispensable part of mainstream computer systems. Relational databases have become one of the most important database products. Almost all new database products from database vendors support relational databases. Some non-relational database products also have APIs that support relational databases.

Relational databases process basic and routine transactions using online transaction processing (OLTP), such as bank transactions.

Data Warehouse

The large-scale application of databases has facilitated the exponential growth of data. Online analytical processing (OLAP) is desired more than ever to explore the relationship between data and unveil the untapped data value. However, it is difficult to share data between different databases, and data integration and analysis also face great challenges.

To overcome these challenges, Bill Inmon, the father of the data warehouse, proposed the idea of data warehousing in 1990. The data warehouse runs on a unique data storage architecture to perform OLAP on a large amount of data accumulated over the years. In this way, enterprises can obtain valuable information from massive data quickly and effectively to make informed decisions. The appearance of data warehouses has prompted the information industry to develop from operational systems based on relational databases to decision support systems.

Unlike a database, a data warehouse has the following features:

- It is theme-oriented. It supports various services and operational data. Therefore, the required data needs to be extracted from multiple heterogeneous data sources, processed and integrated, and reorganized by the theme.
- A data warehouse mainly supports enterprise decision analysis and operations involved are mainly data queries. Therefore, it improves the query speed and cuts down the total cost of ownership (TCO) by optimizing the table structures and storage modes.

Table 1-1 Comparison between data warehouses and databases

Dimension	Data Warehouse	Database
Application scenarios	OLAP	OLTP
Data source	Multiple	Single
Data normalization	Denormalized schemas	Highly normalized static schemas
Data access	Optimized read operations	Optimized write operations

Data Lake

Data is an important asset for enterprises. As production and operations data piles up, enterprises hope to save the data for effective management and centralized governance and explore data values.

The data lake provides a good answer to these requirements. It is a large data warehouse that centrally stores structured and unstructured data. It can store raw data from multiple sources and of multiple types. The data can be accessed, processed, analyzed, and transmitted without being structured. The data lake helps enterprises quickly complete federated analysis of heterogeneous data sources and explore data value.

A data lake is in essence a solution that consists of a data storage architecture and data processing tools.

- The data storage architecture must be scalable and reliable enough to store massive structured, semi-structured, and unstructured data.
- Data processing tools are classified into two types:
  - First type: focuses on how to migrate data into the lake, including defining data sources, formulating data synchronization policies, moving data, and compiling data catalogs.
  - Second type: focuses on how to analyze, explore, and utilize data in the lake. The data lake must be equipped with wide-ranging capabilities, such as comprehensive data and data lifecycle management, diversified data analytics, and secure data acquisition and release. Without these data governance tools, the quality of data in the lake cannot be guaranteed due to the lack of metadata. As a result, the data lake may turn into a data swamp.

With the development of big data and AI, the value of data in the data lake is increasing gradually. The data lake enables enterprises to build more optimized operation models by realizing centralized data management. It also helps enterprise with prediction analysis and recommendation models, stimulating further growth of enterprise capabilities.

The difference between a data warehouse and a data lake is analogous to that between a warehouse and a lake: A warehouse stores goods from a specific source while the water (raw data) in a lake comes from rivers, streams and other sources.

Table 1-2 Comparison between data lakes and data warehouses

Dimension	Data Lake	Data Warehouse
Application scenarios	Exploratory analytics, such as machine learning, data discovery, profiling, and prediction	Data analytics based on historical **structured data**
Cost	Low initial cost and high subsequent cost	High initial cost and low subsequent cost
Data quality	Massive raw data to be cleaned and normalized before use	High-quality data that can be used as the basis of facts
Target user	Data scientists and data developers	Business analysts

Huawei FusionInsight Intelligent Data Lake

Huawei's DAYU is a data enablement solution that helps large government agencies and enterprises customize their own intelligent data resource management solutions. This solution can import all-domain data into the data lake, eliminating data silos, unleashing the value of data, and empowering data-driven digital transformation.

DAYU, with the FusionInsight Intelligent Data Lake as the core, contains computing engines such as the database, data warehouse, and data lake, as well as platforms such as DataArts Studio. DAYU provides comprehensive data enablement capabilities, covering data collection, aggregation, computing, asset management, and data openness.

The FusionInsight Intelligent Data Lake solution provides the following services:

- Database
  - Relational database: **RDS**, **TaurusDB**, **GaussDB**, and **PostgreSQL**.
  - Non-relational database: **Document Database Service (DDS)** and **GeminiDB** (compatible with Influx, Redis, Mongo, and Cassandra protocols)
- Data warehouse: **GaussDB (DWS)**
- Data lake: **MapReduce Service (MRS)** and **Data Lake Insight (DLI)**.

- Data governance platform: **DataArts Studio**

1.3 What Is the Relationship Between DataArts Studio and Huawei Horizon Digital Platform?

Huawei Horizon Digital Platform enables digital transformation for industry customers. Based on the cloud, optimize and integrate new ICT technologies and converged data to enable customers to achieve service collaboration and agile innovation

DataArts Studio is a data enablement module of Huawei Horizon Digital Platform. It helps you better manage and use data.

1.4 What Are the Differences Between DataArts Studio and ROMA?

In terms of the three-layer structure (data integration, data governance, and open data) of the data operations solution (data platform), DataArts Studio and ROMA differ mainly in data governance:

- ROMA connects multiple systems but does not provide data governance and planning functions.
- DataArts Studio allows you to analyze data and create unified models to break down data silos.

In practice, you may use DataArts Studio and ROMA together to achieve digital transformation.

1.5 Can DataArts Studio Be Deployed in a Local Data Center or on a Private Cloud?

DataArts Studio must be deployed based on HUAWEI CLOUD. If resources are isolated, DataArts Studio can be deployed in a full-stack DeC. In addition, DataArts Studio can be deployed on Huawei Cloud Stack or Huawei Cloud Stack Online.

For more information about the application scenarios and differences between the full-stack DeC, Huawei Cloud Stack, and Huawei Cloud Stack Online, **contact sales**.

1.6 How Do I Create a Fine-Grained Permission Policy in IAM?

Currently, DataArts Studio does not support the creation of fine-grained permission policies in IAM. You are advised to use DAYU policies and workspace roles to control permissions. .

DataArts Studio assigns permissions through **DAYU system roles** and **workspace roles**. To ensure that the IAM user permissions are normal, the user group to which the IAM user belongs must be assigned the DAYU User or DAYU Administrator role on the IAM console. In addition, ensure that the IAM user with the DAYU User role has been assigned the corresponding role in the DataArts Studio workspace.

A workspace role determines the permissions of a user in the workspace. Currently, four preset roles are available: admin, developer, operator, and viewer. For details about the permissions of the roles, see **DataArts Studio Permissions**.

- Admin: This role has all operation permissions in a workspace. You are advised to assign the admin role to the project owner, development owner, and O&M administrator.
- Developer: This role has permissions to create and manage resources in a workspace. You are advised to assign this role to users who develop and process tasks.
- Operator: This role has the operation permissions of services such as O&M and scheduling in a workspace, but cannot modify resources or configurations. You are advised to assign this role to users responsible for O&M management and status monitoring.
- Viewer: This role can view data in a workspace but cannot perform any other operation. You are advised to assign this role to users who only need to view data in a workspace but do not need to perform operations.
- Deployer: This role is unique to the enterprise mode and has permissions to release task packages in a workspace. In enterprise mode, when a developer submits a script or job version, the system generates a release task. After the developer confirms the release and the deployer approves the release request, the modified job is synchronized to the production environment.

1.7 How Do I Isolate Workspaces So That Users Cannot View Unauthorized Workspaces?

In DataArts Studio, **system roles** and **workspace roles** are used to assign permissions. By default, if a user is assigned the DAYU User system role but not a workspace role, the user cannot view the workspace.

If the user is also assigned the DAYU Administrator, Tenant Guest, or Tenant Administrator role, the user can view all workspaces.

1.8 What Should I Do If a User Cannot View Workspaces After I Have Assigned the Required Policy to the User?

Possible Causes

DataArts Studio assigns permissions through **DAYU system roles** and **workspace roles**. To ensure that the IAM user permissions are normal, the user group to which the IAM user belongs must be assigned the DAYU User or DAYU Administrator role on the IAM console. In addition, ensure that the IAM user with the DAYU User role has been assigned the corresponding role in the DataArts Studio workspace.

If you assign only the DAYU User system role but not a workspace role to a user, the user cannot access a workspace and an error message is displayed.

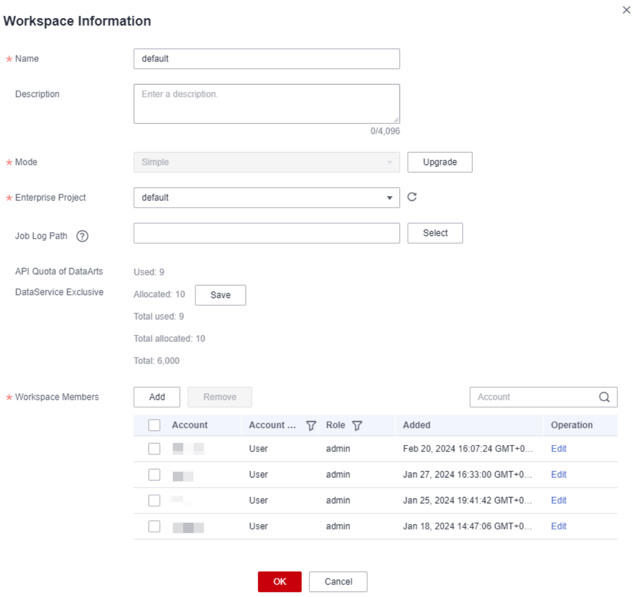
Solution

Check whether the user has been added to the workspace. If not, perform the following steps to add the user:

Step 1 Log in to the DataArts Studio console by following the instructions in **Accessing the DataArts Studio Instance Console**.

Step 2 On the **Workspaces** page, locate a workspace and click **Edit** in the **Operation** column.

Figure 1-2 Workspace Information dialog box



Workspace Information

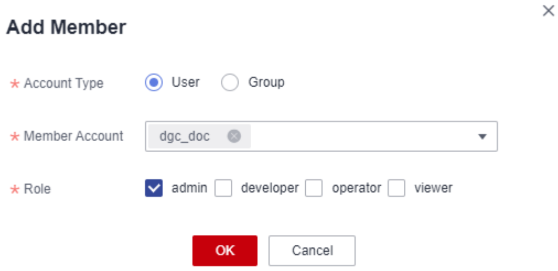
* Name	default
Description	Enter a description. 0/4,096
* Mode	Simple
* Enterprise Project	default
Job Log Path	
API Quota of DataArts DataService Exclusive	Used: 9 Allocated: 10 Total used: 9 Total allocated: 10 Total: 6,000

* Workspace Members

☐	Account	Account ...	Role	Added	Operation
☐	[redacted]	User	admin	Feb 20, 2024 16:07:24 GMT+0...	Edit
☐	[redacted]	User	admin	Jan 27, 2024 16:33:00 GMT+0...	Edit
☐	[redacted]	User	admin	Jan 25, 2024 19:41:42 GMT+0...	Edit
☐	[redacted]	User	admin	Jan 18, 2024 14:47:06 GMT+0...	Edit

Step 3 Click **Add** under **Workspace Members**. In the displayed **Add Member** dialog box, select **Add User** or **Add Group**, select a member account from the drop-down list, and select a role for it.

Figure 1-3 Adding a member



Add Member

* Account Type ● User ○ Group

* Member Account dgc_doc

* Role ☑ admin ☐ developer ☐ operator ☐ viewer

Step 4 Click **OK**. You can view or modify the members and roles in the member list, or remove members from the workspace.

----End

1.9 What Should I Do If Insufficient Permissions Are Prompted When I Am Trying to Perform an Operation as an IAM User?

Possible Causes

DataArts Studio assigns permissions through **DAYU system roles** and **workspace roles**. To ensure that the IAM user permissions are normal, the user group to which the IAM user belongs must be assigned the DAYU User or DAYU Administrator role on the IAM console. In addition, ensure that the IAM user with the DAYU User role has been assigned the corresponding role in the DataArts Studio workspace.

If you assign only the workspace role to the user, an error message is displayed, indicating that the user does not have permissions.

Solution

In this case, you need to check whether the user group to which the IAM user belongs has been assigned the DAYU User or DAYU Administrator role on the IAM console. To create an IAM user and assign a system role to the user, perform the following steps:

Step 1 Create a user group and assign a system role to the group.

Log in to the IAM console using a a Huawei account, create a user group and assign a DataArts Studio system role to the group. For example, the system role can be DAYU Administrator or DAYU User.

For details, see **Creating a User Group and Assigning Permissions**.

NOTE

- When configuring DataArts Studio permissions for a user group, enter **DAYU** in the search box to search for the permissions and select the permissions to be granted to the user group, for example, **DAYU User**.
- DataArts Studio is a project-level service deployed in specific physical regions. If you select **All resources** for **Scope**, the permission takes effect in all projects of all regions. If you select **Region-specific projects** for **Scope**, the permission takes effect only for a specified project. When accessing DataArts Studio, the IAM user must switch to the region where they have been assigned the required permissions.

Step 2 Create a user and add it to the user group.

Create users on the IAM console and add them to the group created in **Step 1**.

For details, see **Creating an IAM User**.

NOTE

An IAM user can pass the authentication and access DataArts Studio through an API or SDK only if **Programmatic access** is selected for **Access Type** during the creation of the IAM user.

Step 3 Create a custom workspace role for DAYU User, add it as a workspace member, and assign a role to the member.

DataArts Studio workspace roles determine the permissions of DAYU User in a workspace. There are five preset roles: admin, developer, deployer, operator, and viewer. For details about how to add a member and assign a role, see **Adding a Member and Assigning a Role**.

For details about the permissions of the roles, see **Permissions**.

Step 4 Log in to the console and verify permissions.

Log in to the console using the created user and verify permissions of the user.

- Choose **Service List** > **DataArts Studio**. Locate a DataArts Studio instance and click **Access**. Check whether the workspace list is displayed.
- Access a service module (for example, Management Center) to which your current user has been added and check whether you can perform the operations allowed for the workspace role assigned to you.

----End

1.10 Can I Delete DataArts Studio Workspaces?

Yes. The procedure is as follows.

NOTE

Mis-deletion may result in service loss. To delete a workspace, you must use the **DAYU Administrator** or **Tenant Administrator** account and ensure that the workspace does not contain any of the following resources:

- Management Center: data connections
- DataArts Migration: CDM clusters
- DataArts Architecture: subjects, logical models, standards, physical models, dimensional models, and metrics
- DataArts Factory: jobs, job directories, scripts, script directories, and resources
- DataArts Quality: quality jobs and comparison jobs
- DataArts Catalog: technical assets including tables and files, and metadata collection tasks
- DataArts DataService: clusters, APIs, and apps
- DataArts Security: sensitive data discovery tasks, masking policies, static masking tasks, and data watermarking tasks

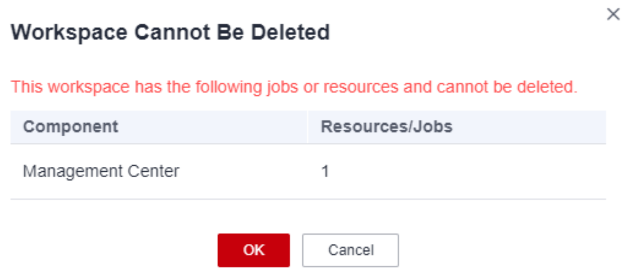
If any module has resources, a message is displayed, indicating that the workspace cannot be deleted.

1. Log in to the DataArts Studio console and go to the **Workspaces** page.
2. On the **Workspaces** page, locate the target workspace, click **More** in the **Operation** column, and select **Delete**.

3. In the **Delete Workspace** dialog box, click **OK**.

   If any module has resources, delete the resources as prompted and try again.

Figure 1-4 Message indicating that the workspace cannot be deleted



Workspace Cannot Be Deleted

This workspace has the following jobs or resources and cannot be deleted.

Component	Resources/Jobs
Management Center	1

1.11 Can I Transfer a Purchased or Trial Instance to Another Account?

No. Purchased or trial instances cannot be transferred to another account.

For how to authorize other users to use your instances, see **Authorizing Users to Use DataArts Studio**.

1.12 Does DataArts Studio Support Version Upgrade?

Yes. If your business volume keeps increasing and the purchased instance version cannot meet your requirements, we recommend that you upgrade the version.

You can log in to the DataArts Studio console, locate the instance to upgrade, click **Upgrade**, and buy a package with higher specifications.

- During the upgrade, the fees are settled each day.
- After the upgrade is complete, you will be billed based on the new package.
- After the package is upgraded, the system creates a CDM cluster. The CDM cluster in the original basic package will be reserved, but you will not be billed for it. You need to migrate data connections and jobs from the original cluster to the new one. For details, see **Can I Synchronize Jobs to Other Clusters?**

1.13 Does DataArts Studio Support Version Downgrade?

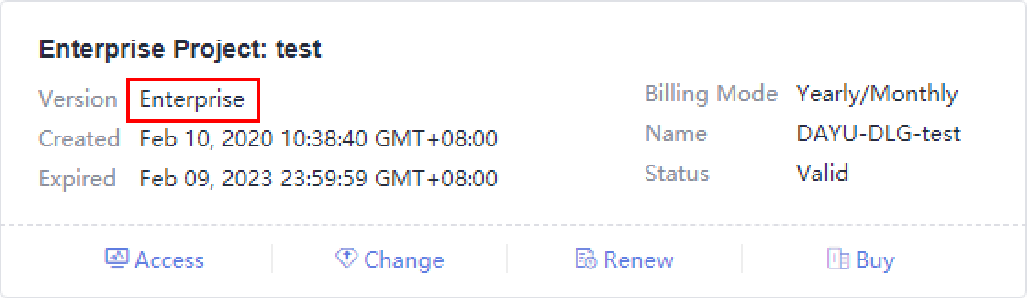
The version of a created DataArts Studio instance cannot be directly downgraded.

However, you can downgrade the version by creating an instance of an earlier version, migrating data to an instance of an earlier version, or unsubscribing from an instance of an earlier version.

1.14 How Do I View the DataArts Studio Instance Version?

You can view the version of a DataArts Studio instance on the instance card.

Figure 1-5 DataArts Studio instance card



Enterprise Project: test

Version Enterprise

Created Feb 10, 2020 10:38:40 GMT+08:00

Expired Feb 09, 2023 23:59:59 GMT+08:00

Billing Mode Yearly/Monthly

Name DAYU-DLG-test

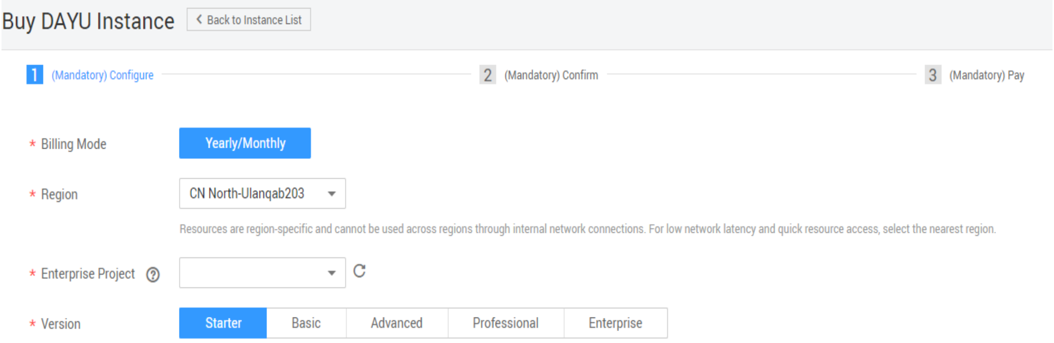
Status Valid

1.15 Why Can't I Select a Specified IAM Project When Purchasing a DataArts Studio Instance?

Check whether the current account has enabled the enterprise project function.

The enterprise project and IAM project cannot be enabled at the same time. If the enterprise project is enabled, you can buy only one instance in this enterprise project.

Figure 1-6 Buying a DataArts Studio instance



Buy DAYU Instance

1 (Mandatory) Configure — 2 (Mandatory) Confirm — 3 (Mandatory) Pay

* Billing Mode: Yearly/Monthly

* Region: CN North-Ulanqab203

Resources are region-specific and cannot be used across regions through internal network connections. For low network latency and quick resource access, select the nearest region.

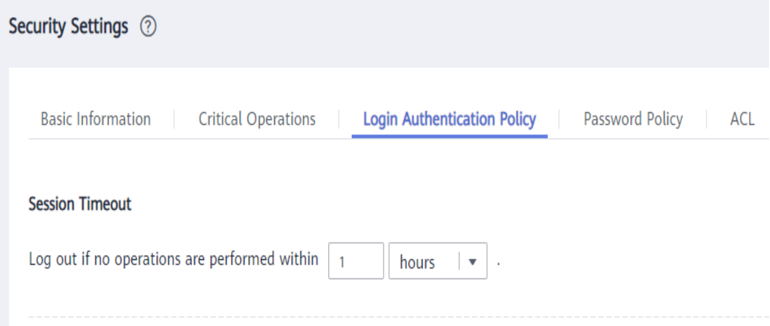
* Enterprise Project:

* Version: Starter

1.16 What Is the Session Timeout Period of DataArts Studio? Can the Session Timeout Period Be Modified?

If you do not perform any operations within the specified duration, the session becomes invalid, and you must log in the system again.

The session timeout policy can be set in the IAM service, as shown in the following figure.



Security Settings

Session Timeout

Log out if no operations are performed within 1 hours .

The session timeout policy is enabled by default and cannot be disabled. The administrator can set the session timeout period, which ranges from 15 minutes to 24 hours and the default value is 1 hour. This policy takes effect for the account and IAM users of the account.

1.17 Will My Data Be Retained If My Package Expires or My Pay-per-Use Resources Are in Arrears?

After a resource enters a grace period or retention period, HUAWEI CLOUD will notify you of this by email or text message. **If you still do not complete the renewal or top-up after the retention period has ended, your data stored in the DAS service will be deleted and the resources will be released.**

- Grace period: Once a monthly/yearly subscription has expired or a pay-per-use resource becomes in arrears, HUAWEI CLOUD provides a period of time during which you can renew the resource or top up your account. Within the grace period, you can still access and use your cloud service.
- Retention period: If you do not renew the yearly/monthly subscription or pay off the arrears within the grace period, the resource enters a retention period after the grace period has expired. During this period, your cloud services cannot be accessed or used, but your stored cloud data will be retained.

For details about how to set the grace period and retention period, see **Service Suspension and Resource Release**.

1.18 How Do I Check the Remaining Validity Period of a Package?

You can check it on the official website.

Log in at the Huawei Cloud official website, select **Billing & Costs** from the username drop-down list, and choose **Orders** > **Renewals** to view the remaining validity period of the package.

1.19 Why Isn't the CDM Cluster in a DataArts Studio Instance Billed?

When you purchase a DataArts Studio instance which is not free of charge, you will get a CDM cluster with 4 vCPUs and 8 GB memory for free.

You are advised to use the free CDM cluster provided by the DataArts Studio instance as an agent of data connections in Management Center. Do not use the CDM cluster to run data migration jobs.

1.20 Why Does the System Display a Message Indicating that the Number of Daily Executed Nodes Has Reached the Upper Limit? What Should I Do?

The number of daily executed nodes is the maximum scheduling times per day for job nodes in different versions of DataArts Studio instances. For details about the quota differences between the versions, see **DataArts Studio Versions**.

Possible Causes

This message is displayed when the sum of the **number of nodes that have been executed**, **that of nodes that are being executed**, and **that of nodes to be executed on a day** have reached the upper limit of the quota.

Solution

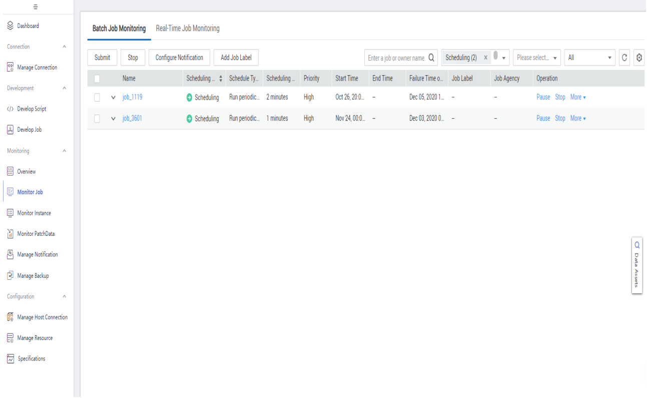
If the number of daily executed nodes exceeds the upper limit, it may be caused by frequent job scheduling.

1. Buy a job node scheduling times/day incremental package to increase the quota. For details, see **Buying a Job Node Scheduling Times/Day Incremental Package**.

2. You can perform the following operations to check for the job nodes that are scheduled more frequently than others, and adjust the scheduling cycles for these nodes or stop scheduling them.

1. In the left navigation tree of Data Development, choose **Monitoring** > **Monitor Instance**, select the current day, and view the jobs that are frequently scheduled.
2. In the left navigation tree of Data Development, choose **Monitoring** > **Monitor Job** to check whether the scheduling period of jobs that are frequently scheduled is set properly. If the scheduling period is inappropriate, adjust the scheduling period or stop the scheduling. Generally, the number of minute-level scheduling jobs executed every day exceeds the upper limit.

Figure 1-7 Viewing the scheduling period



Name	Scheduling ...	Schedule Ty...	Scheduling ...	Priority	Start Time	End Time	Failure Time o...	Job Label	Job Agency	Operation
job_1119	Scheduling	Run periodic...	2 minutes	High	Oct 26, 20:0...	--	Dec 05, 2020 1...	--	--	Pause Stop More ▾
job_3601	Scheduling	Run periodic...	1 minutes	High	Nov 24, 00:0...	--	Dec 03, 2020 0...	--	--	Pause Stop More ▾

2 Management Center

2.1 Which Data Sources Can DataArts Studio Connect To?

DataArts Studio can interconnect with cloud services such as GaussDB(DWS), DLI, and MRS Hive as well as traditional databases such as MySQL and Oracle. For details, see **Data Sources**.

NOTE

To interconnect DataArts Studio with a data source, create a data connection for the data source in Management Center. Data connections in Management Center are independent of the data links in DataArts Migration. They are used in different scenarios.

- The data connections in Management Center are used to connect to the data lake foundation. DataArts Studio provides one-stop data development, governance, and services based on the data lake foundation.
- Data links can be used only in DataArts Migration to integrate data from data sources into a data lake..

2.2 What Are the Precautions for Creating Data Connections?

- RDS data connections depend on OBS. If OBS is unavailable in the same region as DataArts Studio, RDS data connections are not supported.
- For host connections, only Linux hosts are supported.
- If changes occur in the connected data lake (for example, the MRS cluster capacity is expanded), you need to edit and save the connection.
- If the data lake authentication information in a data connection changes (for example, the password expires), the data connection becomes invalid. Ensure that the data lake authentication information is permanently valid to prevent any loss caused by connection failures.
- DataArts Studio does not support MRS clusters whose Kerberos encryption type is **aes256-sha2,aes128-sha2**, and only supports MRS clusters whose Kerberos encryption type is **aes256-sha1,aes128-sha1**.

- If a CDM cluster functions as the agent for a data connection in Management Center, the cluster supports a maximum of 200 concurrent active threads. If multiple data connections share an agent, a maximum of 200 SQL, Shell, and Python scripts submitted through the connections can run concurrently. Excess tasks will be queued. You are advised to plan multiple agents based on the workload.
- Before creating a data connection, ensure that you have obtained the required agent (CDM cluster) and that the CDM cluster can communicate with the data lake to be connected.
  - If the data lake is an on-premises database, you need the Internet or Direct Connect. Ensure that the host where the data source is located and the CDM cluster can access the Internet, and the connection port has been enabled in the firewall rule.
  - If the data lake is a cloud service (such as DWS and MRS), the following requirements must be met for network interconnection:
    - If the CDM cluster and the cloud service are in different regions, a public network or a dedicated connection is required for enabling communication between the CDM cluster and the cloud service.
    - If the CDM cluster and the cloud service are in the same region, VPC, subnet, and security group, they can communicate with each other by default. If they are in the same VPC but in different subnets or security groups, you must configure routing rules and security group rules. For details about how to configure routing rules, see **Configuring Routing Rules**. For details about how to configure security group rules, see **Configuring Security Group Rules**.
    - The cloud service instance and the DataArts Studio workspace belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

2.3 What Should I Do If Database or Table Information Cannot Be Obtained Through a GaussDB(DWS)/Hive/HBase Data Connection?

Possible Causes

The possible cause is that the CDM cluster is stopped or a concurrency conflict occurs. You can switch to another agent to temporarily avoid this issue.

Solution

To resolve this issue, perform the following steps:

Step 1 Check whether the CDM cluster is stopped.

- If yes, start the CDM cluster and check whether the data connection in Management Center recovers.
- If no, go to **step 2**.

Step 2 Check whether the CDM cluster is used as an agent for both a data migration job and a data connection in Management Center.

- If yes, do not use the data migration job and the data connection at the same time, or create another CDM cluster as an agent for the data migration job and the data connection.
- If no, go to **step 3**.

Step 3 Restart the CDM cluster to release resources and check whether the data connection recovers.

----End

2.4 Why Are MRS Hive/HBase Clusters Not Displayed on the Page for Creating Data Connections?

Possible causes are as follows:

- Hive/HBase components were not selected during MRS cluster creation.
- The enterprise project selected during MRS cluster creation is different from that in the workspace.
- The network between the CDM cluster and MRS cluster was disconnected when an MRS data connection is created.

  The CDM cluster functions as a network agent. MRS data connections that you are going to create need to communicate with CDM.

2.5 What Should I Do If a GaussDB(DWS) Connection Test Fails When SSL Is Enabled for the Connection?

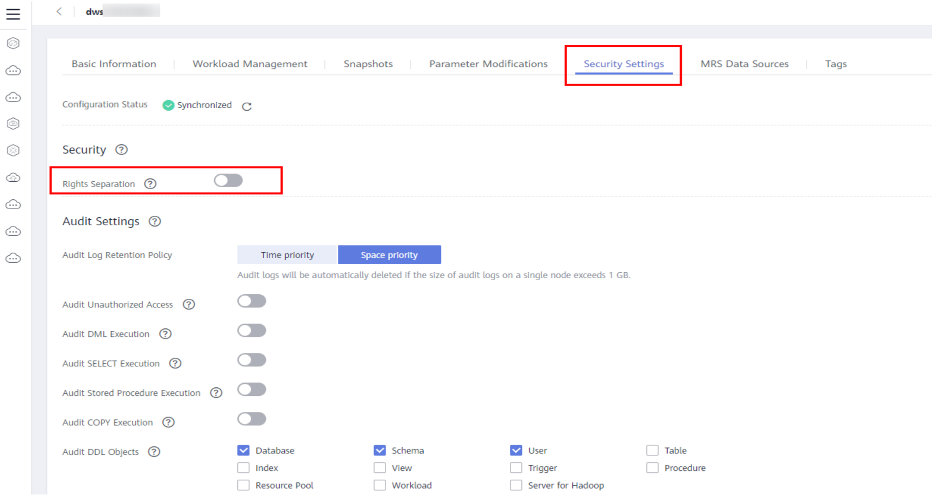
Possible Causes

The failure may be caused by the rights separation function of the GaussDB(DWS) cluster.

Solution

On the DWS console, click the corresponding cluster, choose **Security Settings**, and disable **Rights Separation**.

Figure 2-1 Disabling Rights Separation for the DWS cluster



Configuration Status ✓ Synchronized

Security

Rights Separation ○

Audit Settings

Audit Log Retention Policy: Time priority [✓ Space priority]

Audit logs will be automatically deleted if the size of audit logs on a single node exceeds 1 GB.

Audit Unauthorized Access ○

Audit DML Execution ○

Audit SELECT Execution ○

Audit Stored Procedure Execution ○

Audit COPY Execution ○

Audit DDL Objects

☑ Database ☑ Schema ☑ User □ Table
□ Index □ View □ Trigger □ Procedure
□ Resource Pool □ Workload □ Server for Hadoop

2.6 Can I Create Multiple Connections to the Same Data Source in a Workspace?

Yes, but the name of each connection must be unique.

To ensure that you can select the correct data connection in future development, you are advised to name the connections carefully to avoid confusion.

2.7 Should I Select the API or Proxy Connection Type When Creating a Data Connection in Management Center?

The API connection type is available only for DataArts Factory.

To ensure that a connection is available for other components such as DataArts Architecture, DataArts Quality, DataArts Catalog, and DataArts Service, you are advised to select the proxy connection type.

2.8 How Do I Migrate the Data Development Jobs and Data Connections from One Workspace to Another?

You can export data connections and jobs from one workspace and import them to another workspace.

- Export and import data connections in Management Center. For details, see **Migrating Resources**.
- Export and import jobs in DataArts Factory. For details, see **Exporting and Importing a Job**.

3 DataArts Migration (CDM Jobs)

3.1 What Are the Differences Between CDM and Other Data Migration Services?

HUAWEI CLOUD provides the following data migration services:

- **Cloud Data Migration (CDM)**
- **Object Storage Migration Service (OMS)**
- **Data Replication Service (DRS)**
- **Server Migration Service (SMS)**
- **Database and Application Migration UGO**
- **Data Express Service (DES)**

For details about the differences between the preceding services, see **Differences Between Data Migration Services**.

CDM

Cloud Data Migration (CDM) is an efficient and easy-to-use batch data integration service. Based on the big data migration to the cloud and intelligent data lake solution, CDM provides easy-to-use migration capabilities and capabilities of integrating multiple data sources to the data lake, reducing the complexity of data source migration and integration and effectively improving the data migration and integration efficiency. For details, see **What Is CDM?**.

CDM can migrate data from a database or object storage to a data lake or big data system.

Differences between CDM and DRS:

- If the destination is a big data system, CDM is recommended.
- If the destination is an OLTP database or DWS, DRS is recommended.

Differences between CDM and OMS:

- OMS is used to migrate data from the following clouds to HUAWEI CLOUD: Amazon Web Services (AWS), Alibaba Cloud, Microsoft Azure, Baidu Cloud, QingCloud, Qiniu Cloud, and Tencent Cloud.
- CDM is mainly used to migrate OBS data to a data lake or big data system for data development, cleaning, and governance. However, use OMS if you want to migrate an entire bucket.

OMS

OMS helps you migrate data from object storage on other clouds online to the OBS on HUAWEI CLOUD. For details, see **Object Storage Migration Service**.

OMS provides the following functions:

- Online data migration: It helps you easily and smoothly migrate object storage data from the public cloud of other cloud service providers to HUAWEI CLOUD.
- Cross-region replication: It enables you to replicate and back up data across regions of HUAWEI CLOUD.

Currently, OMS can migrate object storage data of the following clouds to HUAWEI CLOUD: Amazon Web Services (AWS), Alibaba Cloud, Microsoft Azure, Baidu Cloud, HUAWEI CLOUD, Kingsoft Cloud, QingCloud, Qiniu Cloud, and Tencent Cloud.

CDM can also migrate object storage data. However, CDM and OMS differ in the following way:

- OMS is used to migrate data from other clouds to HUAWEI CLOUD.
- CDM is mainly used to migrate OBS data to a data lake or big data system for data development, cleaning, and governance.

DRS

DRS is a stable, efficient, and easy-to-use cloud service for database online migration and synchronization in real time. DRS is used to migrate data from mainstream databases to other databases (including third-party databases), for example, from OLTP to OLTP or DWS. For details, see **Data Replication Service**.

Currently, the following database links are supported:

MySQL databases on HUAWEI CLOUD or other clouds to RDS for MySQL

PostgreSQL databases on HUAWEI CLOUD or other clouds to RDS for PostgreSQL

MongoDB databases on HUAWEI CLOUD or other clouds to DDS

Oracle->RDS for MySQL

......

Differences between DRS and CDM:

- DRS migrates data to a database, for example, a MySQL or a MongoDB database.
- CDM migrates data to a data lake or big data system, for example, MRS HDFS or FusionInsight HDFS.

Differences between DRS and UGO:

- DRS is used for full/incremental data migration or synchronization.
- UGO is used for the evaluation, structure migration, and syntax conversion before a heterogeneous database migration.

SMS

SMS is a P2V/V2V migration service that helps you migrate applications and data from on-premises x86 physical servers or VMs on private or public clouds to ECSs on HUAWEI CLOUD.

UGO

UGO is a professional cloud service that focuses on heterogeneous database structure migration. It automatically converts the syntax of the DDL in databases and the database SQL statements encapsulated in service programs into the SQL syntax of GaussDB or RDS on HUAWEI CLOUD. It uses pre-migration evaluation, structure migration, and automatic syntax conversion to identify possible reconstruction in advance, improve the conversion rate, and minimize the database migration cost. For details, see **Database and Application Migration UGO**.

In short, UGO is used for the evaluation, structure migration, and syntax conversion before a heterogeneous database migration.

DES

DES provides you with physical devices to make it easier to migrate terabytes, or even petabytes of data to HUAWEI CLOUD inexpensively and much faster than would be possible over a network connection. For details, see **Data Express Service**.

Differences Between Data Migration Services

Table 3-1 Differences between data migration services

Service Name	Functions	Differences with Other Services
CDM	● Migrates big data to the cloud. ● Migrates multiple data sources to the data lake.	**Differences with DRS:** DRS is used to migrate databases, while CDM is used to migrate data to big data systems.

Service Name	Functions	Differences with Other Services
OMS	Object storage migration ● Migrates object storage data from other clouds to HUAWEI CLOUD. ● Migrates data between different regions of HUAWEI CLOUD.	**Differences with CDM:** OMS is used to migrate data from other clouds to HUAWEI CLOUD, while CDM is mainly used to migrate OBS data to a data lake or big data system for data development, cleaning, and governance.
DRS	Migrates data between mainstream databases and HUAWEI CLOUD. ● Online database migration ● Real-time database synchronization	● **Differences with CDM:** DRS is used to migrate databases, while CDM is used to migrate data to big data systems. ● **Differences with UGO:** DRS is used to migrate and synchronize data between homogeneous and heterogeneous databases, while UGO is used to migrate the structure and syntax of heterogeneous databases and evaluate the migration.
SMS	Server migration Migrates data from physical servers or VMs on private or public clouds to HUAWEI CLOUD	-
UGO	● Migrates database structure. ● Evaluates database migration. ● Migrates syntax.	**Differences with DRS:** DRS is used to migrate and synchronize data between homogeneous and heterogeneous databases, while UGO is used to migrate the structure and syntax of heterogeneous databases and evaluate the migration.
DES	● Migrates terabytes or even petabytes of data to HUAWEI CLOUD. ● Uses physical media for migration.	-

3.2 What Are the Advantages of CDM?

CDM is developed based on a distributed computing framework and leverages the parallel data processing technology. **Table 3-2** details the advantages of CDM.

Table 3-2 CDM advantages

Item	User-Developed Script	CDM
Ease of use	You need to prepare server resources, and install and configure software, which is time-consuming. Because the data source types are different, the program uses different access interfaces, such as JDBC and native APIs, to read and write data. In this case, various libraries and SDKs are required when you write data migration scripts, resulting in high development and management costs.	CDM provides a web-based management console for enabling services on web pages in real time. You can migrate data by configuring data sources and migration jobs on the GUI and CDM will manage and maintain the data sources and migration jobs for you. In other words, you only need to focus on the data migration logic without worrying about the environment, which greatly reduces development and maintenance costs. CDM also provides RESTful APIs to support third-party system calling and integration.
Real-time monitoring	You need to select specific versions to develop as required.	You can use Cloud Eye to automatically monitor CDM clusters in real time and manage alarms and notifications, so that you can keep track of CDM cluster performance metrics.
O&M free	You need to develop and optimize O&M functions, especially alarm and notification functions, to ensure system availability. Otherwise, manual attendance is required.	With CDM, you do not need to maintain resources such as servers and VMs. CDM has the log, monitoring, and alarm functions, which send notifications to related personnel in a timely manner to avoid 24/7 hours of manual O&M.

Item	User-Developed Script	CDM
High efficiency	During data migration, the read and write process is completed in one job. Limited by available resources, the performance is poor and generally cannot meet the requirements of scenarios where massive sets of data need to be migrated.	Based on the distributed computing framework, CDM jobs are split into independent sub-jobs and executed concurrently, which drastically improves data migration efficiency. In addition, efficient data import interfaces are provided to import data from Hive, HBase, MySQL databases, and Data Warehouse Service (DWS).
Various data sources	Different tasks must be developed for different data sources, generating a number of scripts.	Data sources such as databases, Hadoop services, NoSQL databases, data warehouses, and files are supported.
Different network environments	As the cloud computing technology develops, user data may be stored in different environments, such as public clouds, on-premises or hosted Internet data centers (IDCs), and hybrid scenarios. In heterogeneous environments, data migration is subject to various factors, for example, network connectivity, which causes inconvenience for development and maintenance.	CDM helps you easily cope with various data migration scenarios, including data migration to the cloud, data exchange on the cloud, and data migration to on-premises service systems, regardless of whether the data is stored on on-premises IDCs, cloud services, third-party clouds, or self-built databases or file systems on ECSs.

3.3 What Are the Security Protection Mechanisms of CDM?

CDM is a fully hosted service that provides the following capabilities to protect user data security:

- Instance isolation: CDM users can use only their own instances. Instances are isolated from each other and cannot access each other.
- System hardening: System hardening for security has been performed on the operating system of the CDM instance, so attackers cannot access the operating system from the Internet.
- Key encryption: Keys of various data sources entered when users create links on CDM are stored in CDM databases using high-strength encryption algorithms.
- No intermediate storage: During data migration, CDM processes only data mapping and conversion without storing any user data or data fragments.

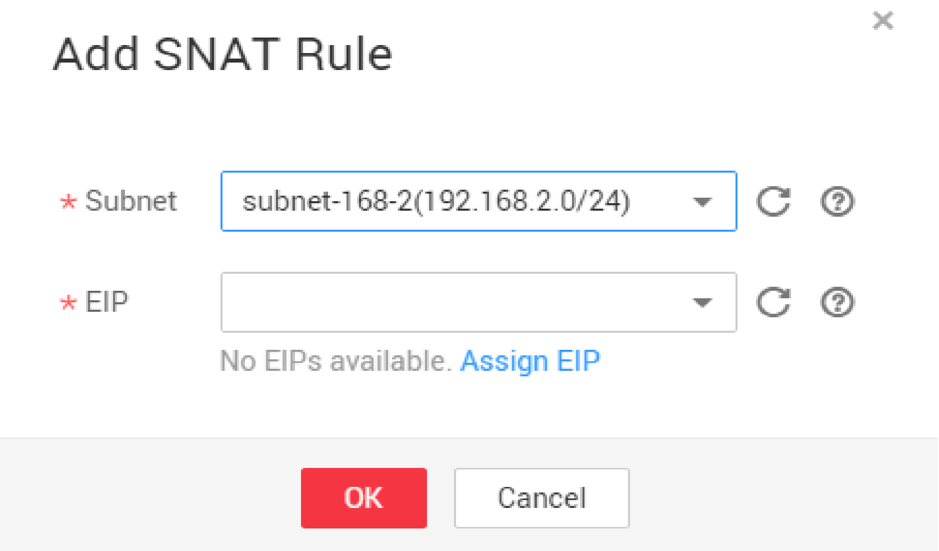
3.4 How Do I Reduce the Cost of Using CDM?

When migrating the data on the public network, use NAT Gateway to share the EIPs with other ECSs in the subnet. In this way, data on the on-premises data center or third-party cloud can be migrated in a more economical and convenient manner.

The following details the operations:

1. Suppose that you have created a CDM cluster (no dedicated EIP needs to be bound to the CDM cluster). Record the VPC and subnet where the CDM cluster is located.
2. Create a NAT gateway. Select the same VPC and subnet as the CDM cluster.
3. After the NAT gateway is created, return to the NAT gateway console list, click the created gateway name, and then click **Add SNAT Rule**.

Figure 3-1 Adding an SNAT rule



Add SNAT Rule

* Subnet: subnet-168-2(192.168.2.0/24)

* EIP:

No EIPs available. Assign EIP

4. Select a subnet and an EIP. If no EIP is available, apply for one.

   After that, access the CDM management console and migrate data from the public network to the cloud through the Internet. For example, migrate files from the FTP server in the on-premises data center to OBS and migrate relational databases from the third-party cloud to RDS.

3.5 Will I Be Billed If My CDM Cluster Does Not Use the Data Transmission Function?

A CDM cluster is billed when it is running, even if you are not using it.

You are advised to delete the clusters that you seldom use and create clusters when needed. For details about CDM cluster billing, see **Pricing Details**.

3.6 Why Am I Billed Pay per Use When I Have Purchased a Yearly/Monthly CDM Incremental Package?

Check whether the region and specifications of the package are the same as those of the CDM cluster. If not, the package cannot be used. To view the CDM cluster specifications and region, log in to the CDM console, choose **Cluster Management** in the navigation pane, and click the cluster name in the cluster list.

If the package and the CDM cluster have the same region and specifications, pay-per-use fees are generated in the following scenarios:

If you buy a pay-per-use incremental package and then a yearly/monthly incremental package, you will be billed in pay-per-use mode first, and then in yearly/monthly mode.

3.7 How Do I Check the Remaining Validity Period of a Package?

You can check it on the official website.

Log in at the Huawei Cloud official website, select **Billing & Costs** from the username drop-down list, and choose **Orders** > **Renewals** to view the remaining validity period of the package.

3.8 Can CDM Be Shared by Different Tenants?

CDM can be shared with IAM users of the same tenant through authorization.

To authorize an IAM user, perform the following steps:

1. **Create a user group and assign permissions**

   Create a user group on the IAM console, and attach the **CDM ReadOnlyAccess** policy to the group.

2. **Create an IAM user.**

   Create a user on the IAM console and add it to the user group created in **1**.

3. **Log in** and verify permissions.

   In the authorized region, perform the following operations:

   - Choose **Service List** > **Cloud Data Migration**. On the CDM console, view clusters. If no message appears indicating insufficient permissions to perform the operation, the **CDM ReadOnlyAccess** policy has already taken effect.
   - Choose any other service in **Service List**. If a message appears indicating that you have insufficient permissions to access the service, the **CDM ReadOnlyAccess** policy has already taken effect.

3.9 Can I Upgrade a CDM Cluster?

No. To use a later version cluster, you can create one.

3.10 How Is the Migration Performance of CDM?

Theoretically, a cdm.large CDM instance can migrate 1 TB to 8 TB data per day. The actual transmission rate is affected by factors such as the Internet bandwidth, cluster specifications, file read/write speed, number of concurrent jobs, and disk read/write performance. For details, see **Performance White Paper**.

3.11 What Is the Number of Concurrent Jobs for Different CDM Cluster Versions?

CDM migrates data through data migration jobs. It works in the following way:

1. When data migration jobs are submitted, CDM splits each job into multiple tasks based on the **Concurrent Extractors** parameter in the job configuration.

   NOTE

   Jobs for different data sources may be split based on different dimensions. Some jobs may not be split based on the **Concurrent Extractors** parameter.

2. CDM submits the tasks to the running pool in sequence. Tasks (defined by **Maximum Concurrent Extractors**) run concurrently. Excess tasks are queued.

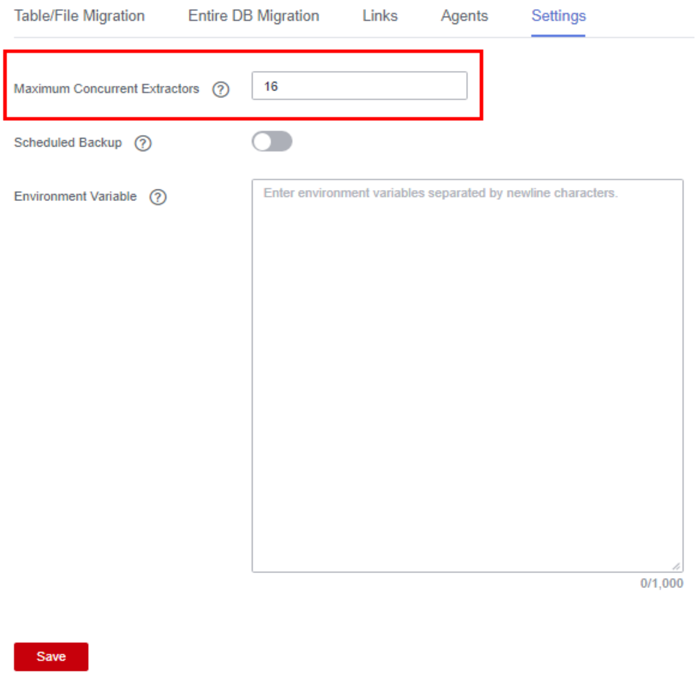
Changing Concurrent Extractors

1. The maximum number of concurrent extractors for a cluster varies depending on the CDM cluster flavor. You are advised to set the maximum number of concurrent extractors to twice the number of vCPUs of the CDM cluster.

Table 3-3 Maximum number of concurrent extractors for a CDM cluster

Flavor	vCPUs/Memory	Maximum Concurrent Extractors
cdm.large	8 vCPUs, 16 GB	16
cdm.xlarge	16 vCPUs, 32 GB	32
cdm.4xlarge	64 vCPUs, 128 GB	128

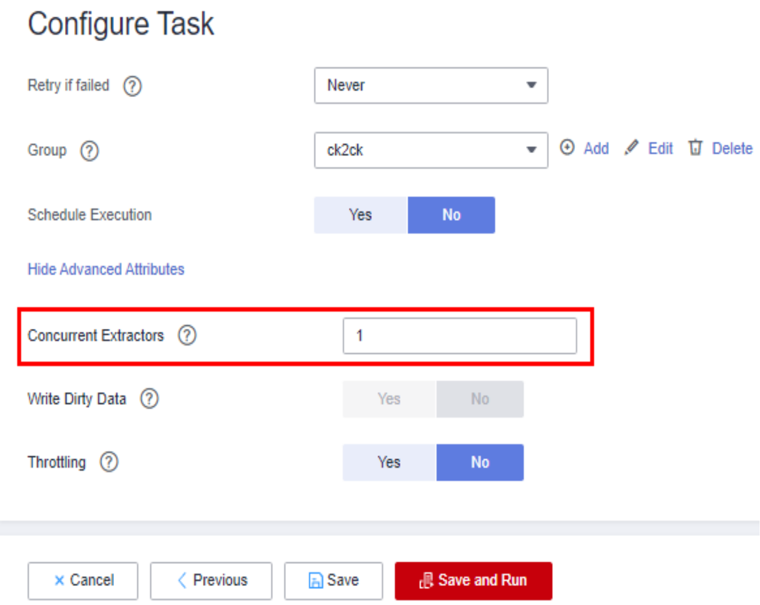
Figure 3-2 Setting Maximum Concurrent Extractors for a CDM cluster



Maximum Concurrent Extractors: 16

2. Configure the number of concurrent extractors based on the following rules:
   a. When data is to be migrated to files, CDM does not support multiple concurrent tasks. In this case, set a single process to extract data.
   b. If each row of the table contains less than or equal to 1 MB data, data can be extracted concurrently. If each row contains more than 1 MB data, it is recommended that data be extracted in a single thread.
   c. Set **Concurrent Extractors** for a job based on **Maximum Concurrent Extractors** for the cluster. It is recommended that **Concurrent Extractors** is less than **Maximum Concurrent Extractors**.
   d. If the destination is DLI, you are advised to set the number of concurrent extractors to 1. Otherwise, data may fail to be written.

Figure 3-3 Setting Concurrent Extractors for a job



Configure Task

Retry if failed: Never

Group: ck2ck

Schedule Execution: Yes [✓ No]

Hide Advanced Attributes

Concurrent Extractors: 1

Write Dirty Data: Yes No

Throttling: Yes [✓ No]

3.12 Does CDM Support Incremental Data Migration?

Symptom

Does CDM support incremental data migration?

Solution

CDM supports incremental data migration.

With scheduled jobs and macro variables of date and time, CDM provides incremental data migration in the following scenarios:

- Incremental file migration
- Incremental migration of relational databases
- HBase/CloudTable incremental migration

For details, see **Incremental Migration**.

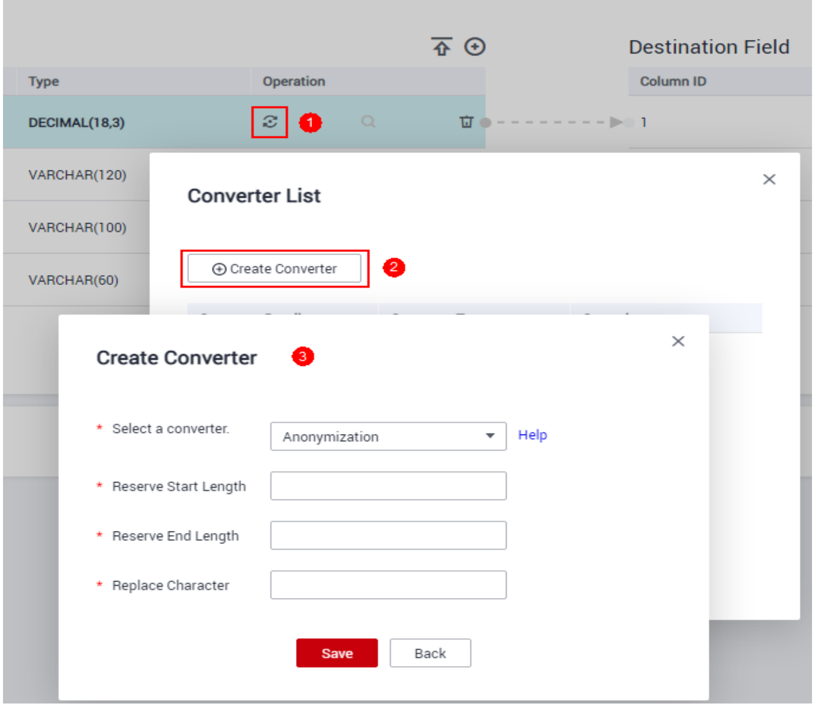
3.13 Does CDM Support Field Conversion?

Yes. CDM supports the following field converters:

- **Anonymization**
- **Trim**
- **Reverse String**
- **Replace String**
- **Expression Conversion**

You can create a field converter on the **Map Field** page when creating a table/file migration job.

Figure 3-4 Creating a field converter



Select a converter. Anonymization

Reserve Start Length

Reserve End Length

Replace Character

Anonymization

This converter is used to hide key information about the character string. For example, if you want to convert **12345678910** to **123****8910**, configure the parameters as follows:

- Set **Reserve Start Length** to **3**.
- Set **Reserve End Length** to **4**.
- Set **Replace Character** to *.

Trim

This converter is used to automatically delete the spaces before and after a string. No parameters need to be configured.

Reverse String

This converter is used to automatically reverse a string. For example, reverse **ABC** into **CBA**. No parameters need to be configured.

Replace String

This converter is used to replace a character string. You need to configure the object to be replaced and the new value.

Expression Conversion

This converter uses the JSP expression language (EL) to convert the current field or a row of data. The JSP EL is used to create arithmetic and logical expressions. In an expression, you can use integers, floating point numbers, strings, constants **true** and **false**, and **null**.

During data conversion, if the content to be replaced contains a special character, use a backslash (\) to escape the special character to a common one.

- The expression supports the following environment variables:
  - **value**: indicates the current field value.
  - **row**: indicates the current row, which is an array type.
- The expression supports the following Utils:
  a. If the field is of the string type, convert all character strings into lowercase letters, for example, convert **aBC** to **abc**.

     Expression: StringUtils.lowerCase(value)

  b. Convert all character strings of the current field to uppercase letters.

     Expression: StringUtils.upperCase(value)

  c. Convert the format of the first date field from 2018-01-05 15:15:05 to 20180105.

     Expression: DateUtils.format(DateUtils.parseDate(row[0],"yyyy-MM-dd HH:mm:ss"),"yyyyMMdd")

  d. Convert a timestamp to a date string in *yyyy-MM-dd hh:mm:ss* format, for example, convert **1701312046588** to **2023-11-30 10:40:46**.

     Expression: DateUtils.format(NumberUtils.toLong(value),"yyyy-MM-dd HH:mm:ss")

  e. Convert a date string in the yyyy-MM-dd hh:mm:ss format to a timestamp.

     Expression: DateUtils.getTime(DateUtils.parseDate(value,"yyyy-MM-dd hh:mm:ss"))

  f. If the field value is a date string in *yyyy-MM-dd* format, extract the year from the field value, for example, extract **2017** from **2017-12-01**.

     Expression: StringUtils.substringBefore(value,"-")

  g. If the field value is of the numeric type, convert the value to a new value which is two times greater than the original value:

     Expression: value*2

  h. Convert the field value **true** to **Y** and other field values to **N**.

     Expression: value=="true"?"Y":"N"

  i. If the field value is of the string type and is left empty, convert it to **Default**. Otherwise, the field value will not be converted.

     Expression: empty value? "Default":value

  j. Convert date format **2018/01/05 15:15:05** to **2018-01-05 15:15:05**:

     Expression: DateUtils.format(DateUtils.parseDate(value,"yyyy/MM/dd HH:mm:ss"),"yyyy-MM-dd HH:mm:ss")

  k. Obtain a 36-bit universally unique identifier (UUID):

Expression: CommonUtils.randomUUID()

l. If the field is of the string type, capitalize the first letter, for example, convert **cat** to **Cat**.

   Expression: StringUtils.capitalize(value)

m. If the field is of the string type, convert the first letter to a lowercase letter, for example, convert **Cat** to **cat**.

   Expression: StringUtils.uncapitalize(value)

n. If the field is of the string type, use a space to fill in the character string to the specified length and center the character string. If the length of the character string is not shorter than the specified length, do not convert the character string. For example, convert **ab** to meet the specified length 4.

   Expression: StringUtils.center(value,*4*)

o. Delete a newline (including **\n**, **\r**, and **\r\n**) at the end of a character string. For example, convert **abc\r\n\r\n** to **abc\r\n**.

   Expression: StringUtils.chomp(value)

p. If the string contains the specified string, **true** is returned; otherwise, **false** is returned. For example, **abc** contains **a** so that **true** is returned.

   Expression: StringUtils.contains(value,"*a*")

q. If the string contains any character of the specified string, **true** is returned; otherwise, **false** is returned. For example, **zzabyycdxx** contains either **z** or **a** so that **true** is returned.

   Expression: StringUtils.containsAny(value,"*za*")

r. If the string does not contain any one of the specified characters, **true** is returned. If any specified character is contained, **false** is returned. For example, **abz** contains one character of **xyz** so that **false** is returned.

   Expression: StringUtils.containsNone(value,"*xyz*")

s. If the string contains only the specified characters, **true** is returned. If any other character is contained, **false** is returned. For example, **abab** contains only characters among **abc** so that **true** is returned.

   Expression: StringUtils.containsOnly(value,"*abc*")

t. If the character string is empty or null, convert it to the specified character string. Otherwise, do not convert the character string. For example, convert the empty character string to null.

   Expression: StringUtils.defaultIfEmpty(value,*null*)

u. If the string ends with the specified suffix (case sensitive), **true** is returned; otherwise, **false** is returned. For example, if the suffix of **abcdef** is not null, **false** is returned.

   Expression: StringUtils.endsWith(value,*null*)

v. If the string is the same as the specified string (case sensitive), **true** is returned; otherwise, **false** is returned. For example, after strings **abc** and **ABC** are compared, **false** is returned.

   Expression: StringUtils.equals(value,"*ABC*")

w. Obtain the first index of the specified character string in a character string. If no index is found, **-1** is returned. For example, the first index of **ab** in **aabaabaa** is 1.

Expression: StringUtils.indexOf(value,"*ab*")

x. Obtain the last index of the specified character string in a character string. If no index is found, **-1** is returned. For example, the last index of **k** in **aFkyk** is 4.

   Expression: StringUtils.lastIndexOf(value,"*k*")

y. Obtain the first index of the specified character string from the position specified in the character string. If no index is found, **-1** is returned. For example, the first index of **b** obtained after the index 3 of **aabaabaa** is 5.

   Expression: StringUtils.indexOf(value,"*b*",*3*)

z. Obtain the first index of any specified character in a character string. If no index is found, **-1** is returned. For example, the first index of **z** or **a** in **zzabyycdxx.** is 0.

   Expression: StringUtils.indexOfAny(value,"*za*")

aa. If the string contains any Unicode character, **true** is returned; otherwise, **false** is returned. For example, **ab2c** contains only non-Unicode characters so that **false** is returned.

   Expression: StringUtils.isAlpha(value)

ab. If the string contains only Unicode characters and digits, **true** is returned; otherwise, **false** is returned. For example, **ab2c** contains only Unicode characters and digits, so that **true** is returned.

   Expression: StringUtils.isAlphanumeric(value)

ac. If the string contains only Unicode characters, digits, and spaces, **true** is returned; otherwise, **false** is returned. For example, **ab2c** contains only Unicode characters and digits, so that **true** is returned.

   Expression: StringUtils.isAlphanumericSpace(value)

ad. If the string contains only Unicode characters and spaces, **true** is returned; otherwise, **false** is returned. For example, **ab2c** contains Unicode characters and digits so that **false** is returned.

   Expression: StringUtils.isAlphaSpace(value)

ae. If the string contains only printable ASCII characters, **true** is returned; otherwise, **false** is returned. For example, for **!ab-c~**, **true** is returned.

   Expression: StringUtils.isAsciiPrintable(value)

af. If the string is empty or null, **true** is returned; otherwise, **false** is returned.

   Expression: StringUtils.isEmpty(value)

ag. If the string contains only Unicode digits, **true** is returned; otherwise, **false** is returned.

   Expression: StringUtils.isNumeric(value)

ah. Obtain the leftmost characters of the specified length. For example, obtain the leftmost two characters **ab** from **abc**.

   Expression: StringUtils.left(value,*2*)

ai. Obtain the rightmost characters of the specified length. For example, obtain the rightmost two characters **bc** from **abc**.

   Expression: StringUtils.right(value,*2*)

aj. Concatenate the specified character string to the left of the current character string and specify the length of the concatenated character string. If the length of the current character string is not shorter than the specified length, the character string will not be converted. For example, if **yz** is concatenated to the left of **bat** and the length must be 8 after concatenation, the character string is **yzyzybat** after conversion.

Expression: StringUtils.leftPad(value,*8*,"*yz*")

ak. Concatenate the specified character string to the right of the current character string and specify the length of the concatenated character string. If the length of the current character string is not shorter than the specified length, the character string will not be converted. For example, if **yz** is concatenated to the right of **bat** and the length must be 8 after concatenation, the character string is **batyzyzy** after conversion.

Expression: StringUtils.rightPad(value,*8*,"*yz*")

al. If the field is of the string type, obtain the length of the current character string. If the character string is null, **0** is returned.

Expression: StringUtils.length(value)

am. If the field is of the string type, delete all the specified character strings from it. For example, delete **ue** from **queued** to obtain **qd**.

Expression: StringUtils.remove(value,"*ue*")

an. If the field is of the string type, remove the substring at the end of the field. If the specified substring is not at the end of the field, no conversion is performed. For example, remove **.com** at the end of **www.domain.com**.

Expression: StringUtils.removeEnd(value,"*.com*")

ao. If the field is of the string type, delete the substring at the beginning of the field. If the specified substring is not at the beginning of the field, no conversion is performed. For example, delete **www.** at the beginning of **www.domain.com**.

Expression: StringUtils.removeStart(value,"*www.*")

ap. If the field is of the string type, replace all the specified character strings in the field. For example, replace **a** in **aba** with **z** to obtain **zbz**.

Expression: StringUtils.replace(value,"*a*","*z*")

If the content to be replaced contains a special character, the special character must be escaped to a common character. For example, if you want to delete **\t** from a string, use the following expression: StringUtils.replace(value,"\\t",""), which means escaping the backslash (\) again.

aq. If the field is of the string type, replace multiple characters in the character string at a time. For example, replace **h** in **hello** with **j** and **o** with **y** to obtain **jelly**.

Expression: StringUtils.replaceChars(value,"*ho*","*jy*")

ar. If the string starts with the specified prefix (case sensitive), **true** is returned; otherwise, **false** is returned. For example, **abcdef** starts with **abc**, so that **true** is returned.

Expression: StringUtils.startsWith(value,"*abc*")

as. If the field is of the string type, delete all the specified characters at the beginning and end of the field. the field. For example, delete all **x**, **y**, **z**, and **b** from **abcyx** to obtain **abc**.

Expression: StringUtils.strip(value,"*xyz*b")

at. If the field is of the string type, delete all the specified characters at the end of the field, for example, delete the "abc" string at the end of the field.

Expression: StringUtils.stripEnd(value,*"abc"*)

au. If the field is of the string type, delete all the specified characters at the beginning of the field, for example, delete all spaces at the beginning of the field.

Expression: StringUtils.stripStart(value,*null*)

av. If the field is of the string type, obtain the substring after the specified position (the index starts from 0, including the character at the specified position) of the character string. If the specified position is a negative number, calculate the position in the descending order. The first digit at the end is -1. For example, obtain the character whose index is 2 from **abcde** (that is, **c**) and the string after it, that is, **cde**.

Expression: StringUtils.substring(value,*2*)

aw. If the field is of the string type, obtain the substring in a specified range (the index starts from 0, including the character at the start and excluding the character at the end). If the range is a negative number, calculate the position in the descending order. The first digit at the end is -1. For example, obtain the string between the second character (c) and fourth character (e) of **abcde**, that is, **cd**.

Expression: StringUtils.substring(value,*2*,4)

ax. If the field is of the string type, obtain the substring after the first specified character. For example, obtain the substring after the first **b** in **abcba**, that is, **cba**.

Expression: StringUtils.substringAfter(value,"*b*")

ay. If the field is of the string type, obtain the substring after the last specified character. For example, obtain the substring after the last **b** in **abcba**, that is, **a**.

Expression: StringUtils.substringAfterLast(value,"*b*")

az. If the field is of the string type, obtain the substring before the first specified character. For example, obtain the substring before the first **b** in **abcba**, that is, **a**.

Expression: StringUtils.substringBefore(value,"*b*")

ba. If the field is of the string type, obtain the substring before the last specified character. For example, obtain the substring before the last **b** in **abcba**, that is, **abc**.

Expression: StringUtils.substringBeforeLast(value,"*b*")

bb. If the field is of the string type, obtain the substring nested within the specified string. If no substring is found, **null** is returned. For example, obtain the substring between **tag** in **tagabctag**, that is, **abc**.

Expression: StringUtils.substringBetween(value,"*tag*")

bc. If the field is of the string type, delete the control characters (char≤32) at both ends of the character string, for example, delete the spaces at both ends of the character string.

Expression: StringUtils.trim(value)

bd. Convert the character string to a value of the byte type. If the conversion fails, **0** is returned.

Expression: NumberUtils.toByte(value)

be. Convert the character string to a value of the byte type. If the conversion fails, the specified value, for example, **1**, is returned.

Expression: NumberUtils.toByte(value,*1*)

bf. Convert the character string to a value of the double type. If the conversion fails, **0.0d** is returned.

Expression: NumberUtils.toDouble(value)

bg. Convert the character string to a value of the double type. If the conversion fails, the specified value, for example, **1.1d**, is returned.

Expression: NumberUtils.toDouble(value,*1.1d*)

bh. Convert the character string to a value of the float type. If the conversion fails, **0.0f** is returned.

Expression: NumberUtils.toFloat(value)

bi. Convert the character string to a value of the float type. If the conversion fails, the specified value, for example, **1.1f**, is returned.

Expression: NumberUtils.toFloat(value,*1.1f*)

bj. Convert the character string to a value of the int type. If the conversion fails, **0** is returned.

Expression: NumberUtils.toInt(value)

bk. Convert the character string to a value of the int type. If the conversion fails, the specified value, for example, **1**, is returned.

Expression: NumberUtils.toInt(value,*1*)

bl. Convert the character string to a value of the long type. If the conversion fails, **0** is returned.

Expression: NumberUtils.toLong(value)

bm. Convert the character string to a value of the long type. If the conversion fails, the specified value, for example, **1L**, is returned.

Expression: NumberUtils.toLong(value,*1L*)

bn. Convert the character string to a value of the short type. If the conversion fails, **0** is returned.

Expression: NumberUtils.toShort(value)

bo. Convert the character string to a value of the short type. If the conversion fails, the specified value, for example, **1**, is returned.

Expression: NumberUtils.toShort(value,*1*)

bp. Convert the IP string to a value of the long type, for example, convert **10.78.124.0** to **172915712**.

Expression: CommonUtils.ipToLong(value)

bq. Read an IP address and physical address mapping file from the network, and download the mapping file to the map collection. *url* indicates the address for storing the IP mapping file, for example, **http://10.114.205.45:21203/sqoop/IpList.csv**.

Expression: HttpsUtils.downloadMap("*url*") br. Cache the IP address and physical address mappings and specify a key for retrieval, for example, **ipList**.

Expression: CommonUtils.setCache("*ipList*",HttpsUtils.downloadMap("*url*"))

bs. Obtain the cached IP address and physical address mappings.

Expression: CommonUtils.getCache("*ipList*")

bt. Check whether the IP address and physical address mappings are cached.

Expression: CommonUtils.cacheExists("*ipList*")

bu. Based on the specified offset type (month/day/hour/minute/second) and offset (positive number indicates increase and negative number indicates decrease), convert the time in the specified format to a new time, for example, add 8 hours to **2019-05-21 12:00:00**.

Expression: DateUtils.getCurrentTimeByZone("*yyyy-MM-dd HH:mm:ss*",value, "*hour*", *8*)

bv. If the value is empty or null, "aaa" is returned. Otherwise, **value** is returned.

Expression: StringUtils.defaultIfEmpty(value,*"aaa"*)

3.14 What Component Versions Are Recommended for Migrating Hadoop Data Sources?

The recommended component versions can be used as both the source and destination.

Table 3-4 Recommended component versions

Hadoop Type	Component	Description
MRS/Apache/ FusionInsight HD	Hive	2.*x* versions are not supported. The following versions are recommended: ● 1.2.X ● 3.1.X
	HDFS	Recommended versions: ● 2.8.X ● 3.1.X
	HBase	Recommended versions: ● 2.1.X ● 1.3.X

3.15 What Data Formats Are Supported When the Data Source Is Hive?

Symptom

Which data formats are supported when the data source is Hive?

Solution

CDM can read and write data in SequenceFile, TextFile, ORC, or Parquet format from the Hive data source.

3.16 Can I Synchronize Jobs to Other Clusters?

Symptom

Can I synchronize jobs to other CDM clusters?

Solution

CDM does not support direct job migration across clusters. However, you can use the batch job import and export function to indirectly implement cross-cluster migration as follows:

1. Export all jobs from CDM cluster 1 and save the jobs' JSON files to a local PC.

   For security purposes, no link password is exported when CDM exports jobs. All passwords are replaced by *Add password here*.
2. Edit each JSON file on the local PC by replacing *Add password here* with the actual password of the corresponding link.
3. Import the edited JSON files to CDM cluster 2 in batches to implement job migration between cluster 1 and cluster 2.

3.17 Can I Create Jobs in Batches?

Symptom

Can I create CDM jobs in batches?

Solution

CDM supports batch job creation with the help of the batch import function. You can create jobs in batches as follows:

1. Create a job manually.
2. Export the job and save the job's JSON file to a local PC.
3. Edit the JSON file and replicate more jobs in the JSON file according to the job configuration.

4. Import the JSON file to the CDM cluster to implement batch job creation.

You can also enable automatic job creation based on For Each operators. For details, see **Creating Table Migration Jobs in Batches Using CDM Nodes**.

3.18 Can I Schedule Jobs in Batches?

Symptom

Can I schedule CDM jobs in batches?

Solution

Yes.

1. Access the DataArts Factory module of the DataArts Studio service.
2. In the navigation pane of the DataArts Factory homepage, choose **Data Development** > **Develop Job** to create a job.
3. Drag multiple CDM Job nodes to the canvas and orchestrate the jobs.

3.19 How Do I Back Up CDM Jobs?

Symptom

How do I back up CDM jobs?

Solution

You can use the batch export function of CDM to save all job scripts to a local PC. Then, you can create a cluster and import the jobs again when necessary.

3.20 What Should I Do If Only Some Nodes in a HANA Cluster Can Communicate with the CDM Cluster?

Symptom

What should I do If only some nodes in a HANA cluster can communicate with the CDM cluster?

Solution

To ensure that CDM can communicate with the HANA cluster, perform the following operations:

1. Disable Statement Routing of the HANA cluster. Note that this will increase the pressure on configuration nodes.
2. When creating a HANA link, add the advanced attribute **distribution** and set its value to **off**.

After the preceding configurations are complete, CDM can communicate with the HANA cluster.

3.21 How Do I Use Java to Invoke CDM RESTful APIs to Create Data Migration Jobs?

CDM provides RESTful APIs to implement automatic job creation or execution control by program invocation.

The following describes how to use CDM to migrate data from table **city1** in the MySQL database to table **city2** on DWS, and how to use Java to invoke CDM RESTful APIs to create, start, query, and delete a CDM job.

Prepare the following data in advance:

1. Username, account name, and project ID of the cloud account
2. Create a CDM cluster and obtain the cluster ID.

   On the **Cluster Management** page, click the CDM cluster name to view the cluster ID, for example, **c110beff-0f11-4e75-8b10-da7cd882b0ef**.
3. Create a MySQL database and a DWS database, and create tables **city1** and **city2**. The statements for creating tables are as follows:
   ```
   MySQL:
   create table city1(code varchar(10),name varchar(32));
   insert into city1 values('NY','New York');
   DWS:
   create table city2(code varchar(10),name varchar(32));
   ```
4. In the CDM cluster, create a link to MySQL, such as a link named **mysqltestlink**. Create a link to DWS, such as a link named **dwstestlink**.
5. Run the following code. You are advised to use the HttpClient package of version 4.5. Maven configuration is as follows:
   ```
   <project>
   <modelVersion>4.0.0</modelVersion>
   <groupId>cdm</groupId>
   <artifactId>cdm-client</artifactId>
   <version>1</version>
   <dependencies>
   <dependency>
   <groupId>org.apache.httpcomponents</groupId>
   <artifactId>httpclient</artifactId>
   <version>4.5</version>
   </dependency>
   </dependencies>
   </project>
   ```

Sample Code

The code for using Java to invoke CDM RESTful APIs to create, start, query, and delete a CDM job is as follows:

```
package cdmclient;
import java.io.IOException;
import org.apache.http.Header;
import org.apache.http.HttpEntity;
import org.apache.http.HttpHost;
import org.apache.http.auth.AuthScope;
import org.apache.http.auth.UsernamePasswordCredentials;
import org.apache.http.client.CredentialsProvider;
import org.apache.http.client.config.RequestConfig;
```

```
import org.apache.http.client.methods.CloseableHttpResponse;
import org.apache.http.client.methods.HttpDelete;
import org.apache.http.client.methods.HttpGet;
import org.apache.http.client.methods.HttpPost;
import org.apache.http.client.methods.HttpPut;
import org.apache.http.entity.StringEntity;
import org.apache.http.impl.client.BasicCredentialsProvider;
import org.apache.http.impl.client.CloseableHttpClient;
import org.apache.http.impl.client.HttpClients;
import org.apache.http.util.EntityUtils;
public class CdmClient {
private final static String DOMAIN_NAME=" account name";
private final static String USER_NAME=" username";
Private final static String USER_PASSWORD= "Password of the cloud user";
private final static String PROJECT_ID="Project ID";
private final static String CLUSTER_ID="CDM cluster ID";
private final static String JOB_NAME="Job name";
private final static String FROM_LINKNAME="Source link name";
private final static String TO_LINKNAME="Destination link name";
Private final static String IAM_ENDPOINT= "IAM endpoint";
Private final static String CDM_ENDPOINT= "CDM endpoint";
private CloseableHttpClient httpclient;
private String token;

public CdmClient() {
this.httpclient = createHttpClient();
this.token = login();
}

private CloseableHttpClient createHttpClient() {
CloseableHttpClient httpclient =HttpClients.createDefault();
return httpclient;
}

private String login(){
HttpPost httpPost = new HttpPost("https://"+IAM_ENDPOINT+"/v3/auth/tokens");
String json =
"{\r\n"+
"\"auth\": {\r\n"+
"\"identity\": {\r\n"+
"\"methods\": [\"password\"],\r\n"+
"\"password\": {\r\n"+
"\"user\": {\r\n"+
"\"name\": \""+USER_NAME+"\",\r\n"+
"\"password\": \""+USER_PASSWORD+"\",\r\n"+
"\"domain\": {\r\n"+
"\"name\": \""+DOMAIN_NAME+"\"\r\n"+
"}\r\n"+
"}\r\n"+
"}\r\n"+
"},\r\n"+
"\"scope\": {\r\n"+
"\"project\": {\r\n"+
"\"name\": \"PROJECT_NAME\"\r\n"+
"}\r\n"+
"}\r\n"+
"}\r\n"+
"}\r\n";
try {
StringEntity s = new StringEntity(json);
s.setContentEncoding("UTF-8");
s.setContentType("application/json");
httpPost.setEntity(s);
CloseableHttpResponse response = httpclient.execute(httpPost);
Header tokenHeader = response.getFirstHeader("X-Subject-Token");
String token = tokenHeader.getValue();
System.out.println("Login successful");
return token;
} catch (Exception e) {
```

```
throw new RuntimeException("login failed.", e);
}
}
/*Create a job.*/

public void createJob(){
HttpPost httpPost = new HttpPost("https://"+CDM_ENDPOINT+"/cdm/v1.0/"+PROJECT_ID+"/
clusters/"+CLUSTER_ID+"/cdm/job");

/**The JSON information here is complex. You can create a job on the job management page, click Job
JSON Definition next to the job, copy the JSON content and convert it into a Java character string, and
paste it here.
*In the JSON message body, you only need to replace the link name, data import and export table names,
field list of the tables, and fields used for partitioning in the source table.**/

String json =
"{\r\n"+
"\"jobs\": [\r\n"+
"{\r\n"+
"\"from-connector-name\": \"generic-jdbc-connector\",\r\n"+
"\"name\": \""+JOB_NAME+"\",\r\n"+
"\"to-connector-name\": \"generic-jdbc-connector\",\r\n"+
"\"driver-config-values\": {\r\n"+
"\"configs\": [\r\n"+
"{\r\n"+
"\"inputs\": [\r\n"+
"{\r\n"+
"\"name\": \"throttlingConfig.numExtractors\",\r\n"+
"\"value\": \"1\"\r\n"+
"}\r\n"+
"],\r\n"+
"\"validators\": [],\r\n"+
"\"type\": \"JOB\",\r\n"+
"\"id\": 30,\r\n"+
"\"name\": \"throttlingConfig\"\r\n"+
"}\r\n"+
"]\r\n"+
"},\r\n"+
"\"from-link-name\": \""+FROM_LINKNAME+"\",\r\n"+
"\"from-config-values\": {\r\n"+
"\"configs\": [\r\n"+
"{\r\n"+
"\"inputs\": [\r\n"+
"{\r\n"+
"\"name\": \"fromJobConfig.schemaName\",\r\n"+
"\"value\": \"sqoop\"\r\n"+
"},\r\n"+
"{\r\n"+
"\"name\": \"fromJobConfig.tableName\",\r\n"+
"\"value\": \"city1\"\r\n"+
"},\r\n"+
"{\r\n"+
"\"name\": \"fromJobConfig.columnList\",\r\n"+
"\"value\": \"code&name\"\r\n"+
"},\r\n"+
"{\r\n"+
"\"name\": \"fromJobConfig.partitionColumn\",\r\n"+
"\"value\": \"code\"\r\n"+
"}\r\n"+
"],\r\n"+
"\"validators\": [],\r\n"+
"\"type\": \"JOB\",\r\n"+
"\"id\": 7,\r\n"+
"\"name\": \"fromJobConfig\"\r\n"+
"}\r\n"+
"]\r\n"+
"},\r\n"+
"\"to-link-name\": \""+TO_LINKNAME+"\",\r\n"+
"\"to-config-values\": {\r\n"+
"\"configs\": [\r\n"+
"{\r\n"+
"\"inputs\": [\r\n"+
"{\r\n"+
"\"name\": \"toJobConfig.schemaName\",\r\n"+
"\"value\": \"sqoop\"\r\n"+
"},\r\n"+
"{\r\n"+
"\"name\": \"toJobConfig.tableName\",\r\n"+
"\"value\": \"city2\"\r\n"+
"},\r\n"+
"{\r\n"+
"\"name\": \"toJobConfig.columnList\",\r\n"+
"\"value\": \"code&name\"\r\n"+
"}, \r\n"+
"{\r\n"+
"\"name\": \"toJobConfig.shouldClearTable\",\r\n"+
"\"value\": \"true\"\r\n"+
"}\r\n"+
"],\r\n"+
"\"validators\": [],\r\n"+
"\"type\": \"JOB\",\r\n"+
"\"id\": 9,\r\n"+
"\"name\": \"toJobConfig\"\r\n"+
"}\r\n"+
"]\r\n"+
"}\r\n"+
"}\r\n"+
"]\r\n"+
"}\r\n";
try {
StringEntity s = new StringEntity(json);
s.setContentEncoding("UTF-8");
s.setContentType("application/json");
httpPost.setEntity(s);
httpPost.addHeader("X-Auth-Token", this.token);
httpPost.addHeader("X-Language", "en-us");
CloseableHttpResponse response = httpclient.execute(httpPost);
int status = response.getStatusLine().getStatusCode();
if(status == 200){
System.out.println("Create job successful.");
}else{
System.out.println("Create job failed.");
HttpEntity entity = response.getEntity();
System.out.println(EntityUtils.toString(entity));
}
} catch (Exception e) {
e.printStackTrace();
throw new RuntimeException("Create job failed.", e);
}
}
/*Start the job.*/

public void startJob(){
HttpPut httpPut = new HttpPut("https://"+CDM_ENDPOINT+"/cdm/v1.0/"+PROJECT_ID+"/
clusters/"+CLUSTER_ID+"/cdm/job/"+JOB_NAME+"/start");
String json = "";
try {
StringEntity s = new StringEntity(json);
s.setContentEncoding("UTF-8");
s.setContentType("application/json");
httpPut.setEntity(s);
httpPut.addHeader("X-Auth-Token", this.token);
httpPut.addHeader("X-Language", "en-us");
CloseableHttpResponse response = httpclient.execute(httpPut);
int status = response.getStatusLine().getStatusCode();
if(status == 200){
System.out.println("Start job successful.");
}else{
```

```
System.out.println("Start job failed.");
HttpEntity entity = response.getEntity();
System.out.println(EntityUtils.toString(entity));
}
} catch (Exception e) {
e.printStackTrace();
throw new RuntimeException("Start job failed.", e);
}
}
/*Query the job running status cyclically until the job is complete.*/

public void getJobStatus(){
HttpGet httpGet = new HttpGet("https://"+CDM_ENDPOINT+"/cdm/v1.0/"+PROJECT_ID+"/
clusters/"+CLUSTER_ID+"/cdm/job/"+JOB_NAME+"/status");
try {
httpGet.addHeader("X-Auth-Token", this.token);
httpGet.addHeader("X-Language", "en-us");
boolean flag = true;
while(flag){
CloseableHttpResponse response = httpclient.execute(httpGet);
int status = response.getStatusLine().getStatusCode();
if(status == 200){
HttpEntity entity = response.getEntity();
String msg = EntityUtils.toString(entity);
if(msg.contains("\"status\":\"SUCCEEDED\"")){
System.out.println("Job succeeded");
break;
}else if (msg.contains("\"status\":\"FAILED\"")){
System.out.println("Job failed.");
break;
}else{
Thread.sleep(1000);
}

}else{
System.out.println("Get job status failed.");
HttpEntity entity = response.getEntity();
System.out.println(EntityUtils.toString(entity));
break;
}
}
} catch (Exception e) {
e.printStackTrace();
throw new RuntimeException("Get job status failed.", e);
}
}
/*Delete the job.*/

public void deleteJob(){
HttpDelete httpDelte = new HttpDelete("https://"+CDM_ENDPOINT+"/cdm/v1.0/"+PROJECT_ID+"/
clusters/"+CLUSTER_ID+"/cdm/job/"+JOB_NAME);
try {
httpDelte.addHeader("X-Auth-Token", this.token);
httpDelte.addHeader("X-Language", "en-us");
CloseableHttpResponse response = httpclient.execute(httpDelte);
int status = response.getStatusLine().getStatusCode();
if(status == 200){
System.out.println("Delete job successful.");
}else{
System.out.println("Delete job failed.");
HttpEntity entity = response.getEntity();
System.out.println(EntityUtils.toString(entity));
}
} catch (Exception e) {
e.printStackTrace();
throw new RuntimeException("Delete job failed.", e);
}
}
```

```
/*Close the process.*/

public void close(){
try {
httpclient.close();
} catch (IOException e) {
throw new RuntimeException("Close failed.", e);
}
}

public static void main(String[] args){
CdmClient cdmClient = new CdmClient();
cdmClient.createJob();
cdmClient.startJob();
cdmClient.getJobStatus();
cdmClient.deleteJob();
cdmClient.close();
}
}
```

3.22 How Do I Connect the On-Premises Intranet or Third-Party Private Network to CDM?

Many enterprises deploy key data sources on the intranet, such as databases and file servers. CDM runs on the cloud. To migrate the intranet data to the cloud using CDM, use any of the following methods to connect the intranet to the cloud:

- If the destination data source is an on-premises database, you need the Internet or Direct Connect. When using the Internet, ensure that an EIP has been bound to the CDM cluster, the security group of CDM allows outbound traffic from the host where the off-cloud data source is located, the host where the data source is located can access the Internet, and the connection port has been enabled in the firewall rules.
- Establish a VPN between the on-premises data center and the VPC where the service resides.
- Leverage Network Address Translation (NAT) or port forwarding to access the network in proxy mode.

The following describes how to use the port forwarding tool to access intranet data. The process is as follows:

1. Use a Windows computer as the gateway. The computer must be able to access both the Internet and the intranet.
2. Install the port mapping tool IPOP on the computer.
3. Configure port mapping using the tool.

NOTICE

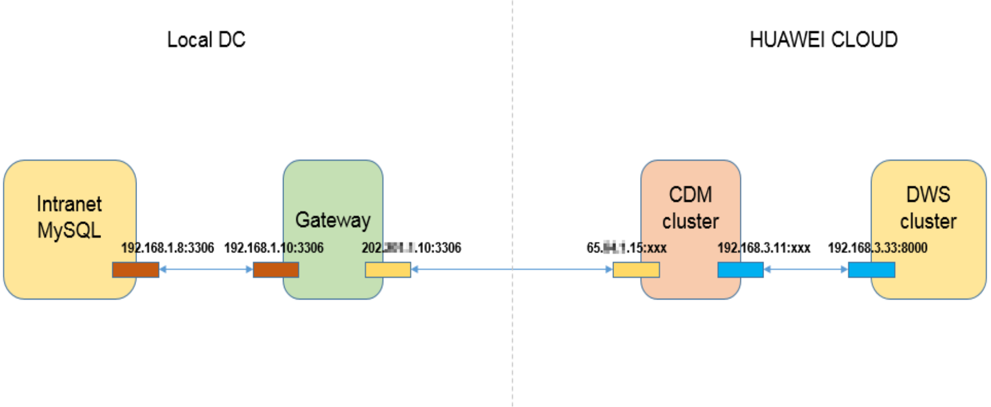
If the intranet database is exposed to the public network for a long time, security risks exist. Therefore, after data migration is complete, stop port mapping.

Scenario

Suppose that the MySQL database on the intranet is migrated to DWS. **Figure 3-5** shows the network topology.

In the figure, the intranet can be either an enterprise's data center or the intranet of the virtual data center on a third-party cloud.

Figure 3-5 Network topology example



Procedure

Step 1 Use a Windows computer as the gateway. Configure both the intranet and Internet IP addresses on the computer. Conduct the following test to check whether the gateway computer can fulfill service needs.

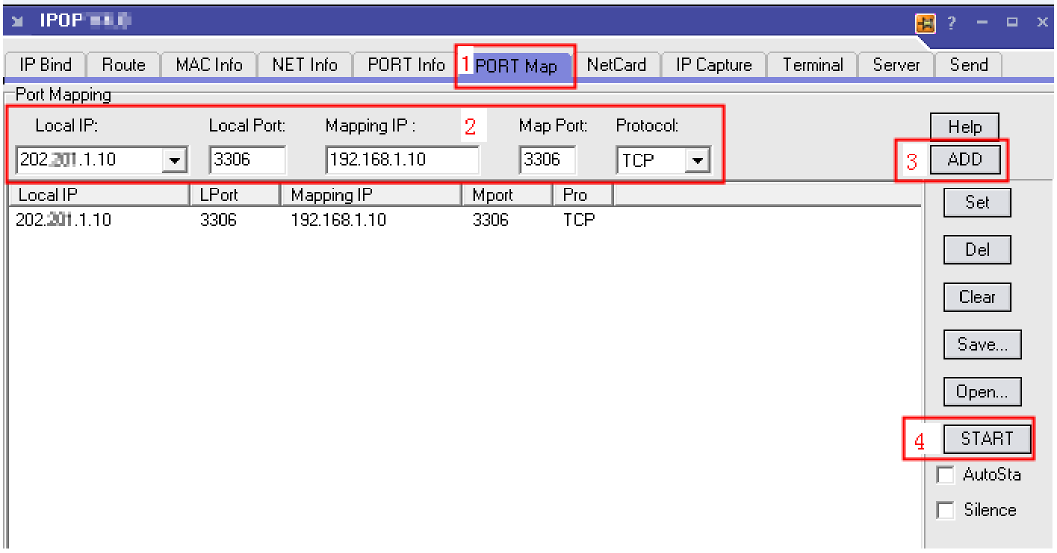
1. Run the **ping** command on the computer to check whether the intranet address of the MySQL database is pingable. For example, run **ping 192.168.1.8**.
2. Run the **ping** command on another computer that can access the Internet to check whether the public network address of the gateway computer is pingable. For example, run **ping 202.*xx*.*xx*.10**.

Step 2 Download the port mapping tool IPOP and install it on the gateway computer.

Step 3 Run the port mapping tool and select **PORT Map**. See **Figure 3-6**.

- **Local IP** and **Local Port**: Configure these two parameters to the public network address and port number of the gateway computer respectively, which must be entered when creating MySQL links on CDM.
- **Mapping IP** and **Map Port**: Configure these two parameters to the IP address and port number of the MySQL database on the intranet.

Figure 3-6 Configuring port mapping



Port Mapping

Local IP: 202.201.1.10
Local Port: 3306
Mapping IP : 192.168.1.10
Map Port: 3306
Protocol: TCP

Local IP	LPort	Mapping IP	Mport	Pro
202.201.1.10	3306	192.168.1.10	3306	TCP

Step 4 Click **ADD** to add a port mapping relationship.

Step 5 Click **START** to start mapping and receive data packets.

Then, you can use the EIP to read data from the MySQL database on the intranet on CDM and import the data to DWS.

NOTE

1. To access the on-premises data source, you must also bind an EIP to the CDM cluster.
2. Generally, DWS is accessible within the same VPC. When creating a CDM cluster, you must ensure that the VPC of the CDM cluster must be the same as that of DWS. In addition, it is recommended that CDM and DWS be in the same intranet and security group. If their security groups are different, you also need to enable data access between the security groups.
3. Port mapping can be used to migrate data between databases on the intranet or the SFTP servers.
4. For Linux computers, port mapping can also be implemented using IPTABLE.
5. When the FTP server on the intranet is mapped to the public network using port mapping, you need to check whether the PASV mode is enabled. In this case, the client and server are connected through a random port. Therefore, in addition to port 21 mapping, you also need to configure the port range mapping in PASV mode. For example, you can specify the **vsftp** port range by configuring **pasv_min_port** and **pasv_max_port**.

----End

3.23 Does CDM Support Parameters or Variables?

Symptom

Does CDM support parameters or variables?

Solution

Yes.

If a CDM job uses the **job parameters** or **variables** configured during data development, data can be indirectly migrated based on the parameters or variables during node scheduling in the DataArts Factory module.

3.24 How Do I Set the Number of Concurrent Extractors for a CDM Migration Job?

CDM migrates data through data migration jobs. It works in the following way:

1. When data migration jobs are submitted, CDM splits each job into multiple tasks based on the **Concurrent Extractors** parameter in the job configuration.

   NOTE

   Jobs for different data sources may be split based on different dimensions. Some jobs may not be split based on the **Concurrent Extractors** parameter.

2. CDM submits the tasks to the running pool in sequence. Tasks (defined by **Maximum Concurrent Extractors**) run concurrently. Excess tasks are queued.

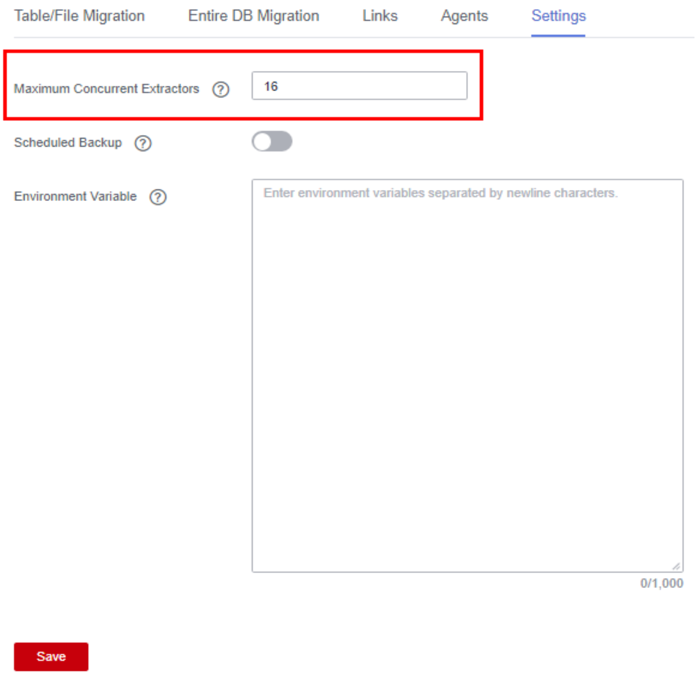
Changing Concurrent Extractors

1. The maximum number of concurrent extractors for a cluster varies depending on the CDM cluster flavor. You are advised to set the maximum number of concurrent extractors to twice the number of vCPUs of the CDM cluster.

Table 3-5 Maximum number of concurrent extractors for a CDM cluster

Flavor	vCPUs/Memory	Maximum Concurrent Extractors
cdm.large	8 vCPUs, 16 GB	16
cdm.xlarge	16 vCPUs, 32 GB	32
cdm.4xlarge	64 vCPUs, 128 GB	128

Figure 3-7 Setting Maximum Concurrent Extractors for a CDM cluster



Maximum Concurrent Extractors: 16

2. Configure the number of concurrent extractors based on the following rules:

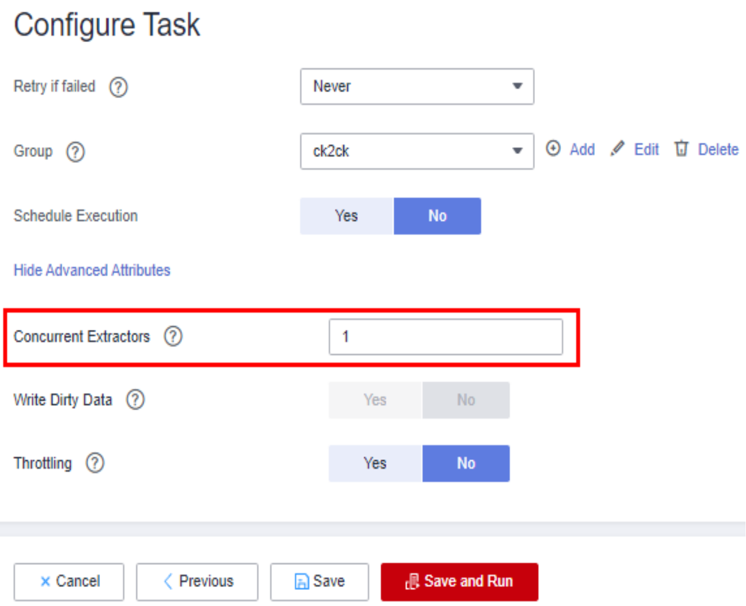
    a. When data is to be migrated to files, CDM does not support multiple concurrent tasks. In this case, set a single process to extract data.

    b. If each row of the table contains less than or equal to 1 MB data, data can be extracted concurrently. If each row contains more than 1 MB data, it is recommended that data be extracted in a single thread.

    c. Set **Concurrent Extractors** for a job based on **Maximum Concurrent Extractors** for the cluster. It is recommended that **Concurrent Extractors** is less than **Maximum Concurrent Extractors**.

    d. If the destination is DLI, you are advised to set the number of concurrent extractors to 1. Otherwise, data may fail to be written.

Figure 3-8 Setting Concurrent Extractors for a job



Configure Task

Retry if failed: Never

Group: ck2ck

Schedule Execution: Yes [✓ No]

Concurrent Extractors: 1

Write Dirty Data: Yes [✓ No]

Throttling: Yes [✓ No]

3.25 Does CDM Support Real-Time Migration of Dynamic Data?

Symptom

Does CDM support real-time migration of dynamic data?

Solution

No.

If data is written to the source during the migration, an error may occur.

3.26 Can I Stop CDM Clusters?

Symptom

Can I stop CDM clusters?

Solution

No.

No. From April 2022 on, CDM clusters cannot be stopped. If a CDM cluster is stopped, its resources may be occupied. As a result, the cluster cannot be started again.

3.27 How Do I Obtain the Current Time Using an Expression?

Symptom

How do I obtain the current time using an expression?

Solution

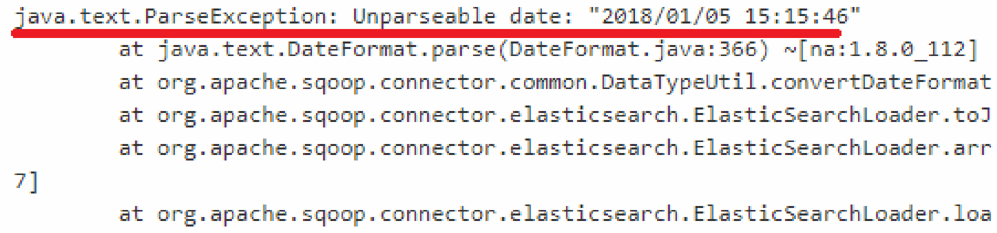
You can use the **DateUtils.format(${timestamp()},"yyyy-MM-dd HH:mm:ss")** expression on the **Map Field** page to obtain the current time. For details, see **Field Conversion**.

3.28 What Should I Do If the Log Prompts that the Date Format Fails to Be Parsed?

Symptom

When CDM is used to migrate other data sources to CSS, the job fails to be executed and the error message "Unparseable date" is displayed in the log. See **Figure 3-9**.

Figure 3-9 Log output



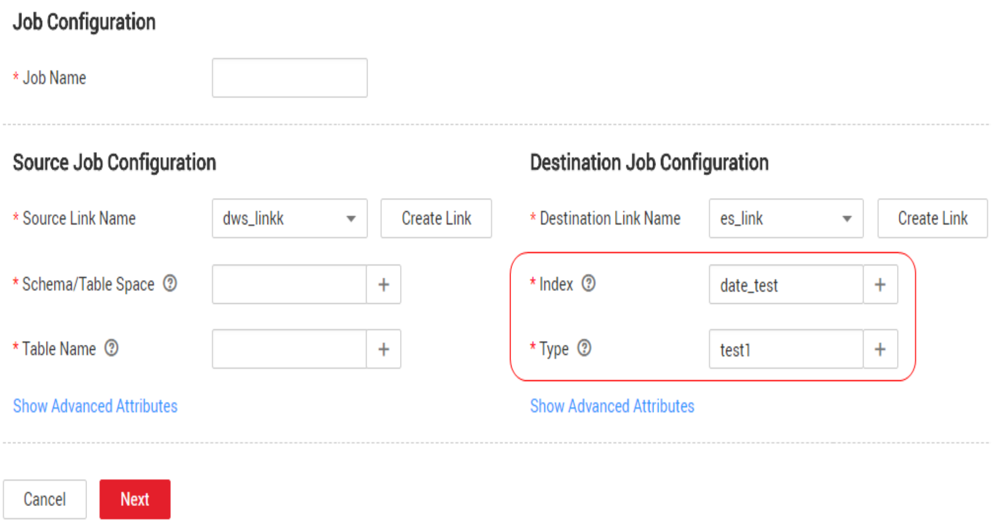
```
java.text.ParseException: Unparseable date: "2018/01/05 15:15:46"
        at java.text.DateFormat.parse(DateFormat.java:366) ~[na:1.8.0_112]
        at org.apache.sqoop.connector.common.DataTypeUtil.convertDateFormat
        at org.apache.sqoop.connector.elasticsearch.ElasticSearchLoader.toJ
        at org.apache.sqoop.connector.elasticsearch.ElasticSearchLoader.arr
7]
        at org.apache.sqoop.connector.elasticsearch.ElasticSearchLoader.loa
```

Possible Cause

CSS has a special processing mechanism on the time field. If the stored time data does not contain the time zone information, Kibana considers the time as the GMT.

The time displayed in the log may be different from the local time. For example, the time displayed in the log is eight hours earlier than the local time in the GMT +08:00 time zone. Therefore, when CDM migrates data to Cloud Search Service, if the index and type are automatically created by CDM (for example, if **date_test** and **test1** of the migration destination highlighted in **Figure 3-10** do not exist in Cloud Search Service, CDM automatically creates the index and type in Cloud Search Service), CDM, by default, sets the format of the time field to the standard format of *yyyy-MM-dd HH:mm:ss.SSS Z*, for example, **2018-01-08 08:08:08.666 +0800**.

Figure 3-10 Job configuration



Job Configuration

* Job Name

Source Job Configuration

* Source Link Name: dws_linkk
* Schema/Table Space
* Table Name

Destination Job Configuration

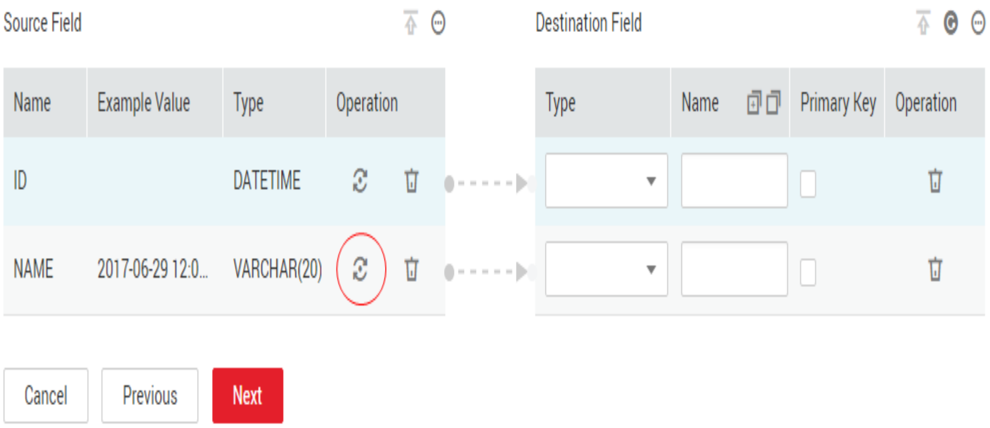
* Destination Link Name: es_link
* Index: date_test
* Type: test1

When data is imported from another data source to CSS, if the date format in the source data is not the standard format, for example, **2018/01/05 15:15:46**, the CDM job fails to be executed, and the log shows that the date format cannot be parsed. You need to configure a field converter on CDM to convert the format of the date field to the required format of CSS.

Solution

1. Edit the job and go to the **Map Field** tab page. Click the icon for creating a converter in the row of the source field to create a converter. See **Figure 3-11**.

   Figure 3-11 Creating a converter

   Source Field

Name	Example Value	Type	Operation
ID		DATETIME	
NAME	2017-06-29 12:0...	VARCHAR(20)	

   Destination Field

Type	Name	Primary Key	Operation
		☐	
		☐	

2. Select **Expression conversion** as the converter. Currently, expression conversion supports functions of the character string and date types. The syntax is similar to the Java character string and time functions. For details about how to compile the expression, see **Expression Conversion**.
3. In this example, the source time format is *yyyy/MM/dd HH:mm:ss*. To convert the source time format to *yyyy-MM-dd HH:mm:ss.SSS Z*, perform the following operations:

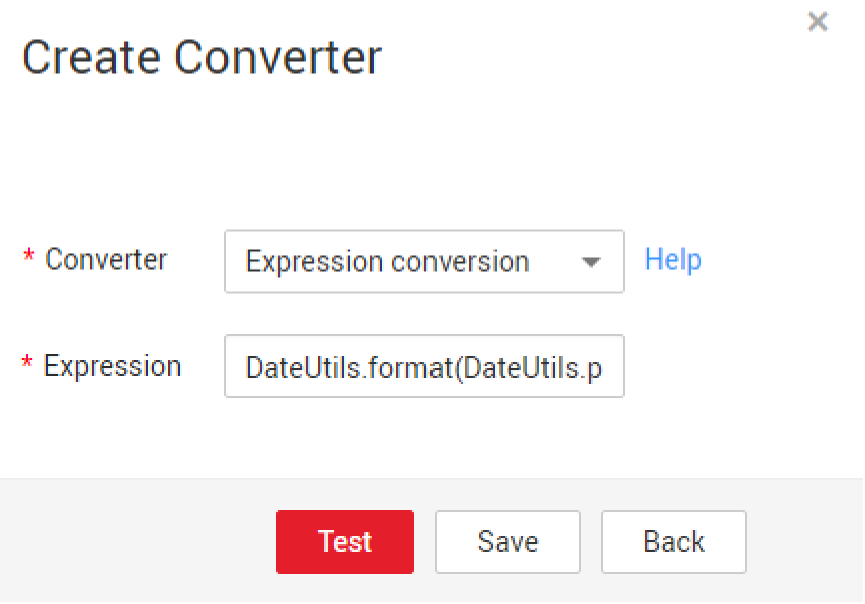
a. Add the time zone information **+0800** to the end of the original date character string. The corresponding expression is **value+" +0800"**.

b. Use the original date format to parse the string to a date object. You can use the DateUtils.parseDate function for parsing. The syntax is **DateUtils.parseDate(String value, String format)**.

c. Format the date object into a character string in target format by using the DateUtils.format function. The syntax is **DateUtils.format(Date date, String format)**.

In this example, the complete expression is **DateUtils.format(DateUtils.parseDate(value+" +0800","yyyy/MM/dd HH:mm:ss Z"),"yyyy-MM-dd HH:mm:ss.SSS Z")**. See **Figure 3-12**.

Figure 3-12 Configuring the expression



Create Converter

* Converter: Expression conversion

* Expression: DateUtils.format(DateUtils.p

4. Save the converter configuration and save and run the job to solve the problem that Cloud Search Service fails to parse the date format.

3.29 What Can I Do If the Map Field Tab Page Cannot Display All Columns?

Symptom

When data is exported from HBase/CloudTable using CDM, fields in the HBase/CloudTable table on the **Map Field** tab page occasionally cannot be displayed completely and cannot match the fields on the migration destination. As a result, the data imported to the migration destination is incomplete.

Possible Cause

HBase/CloudTable are schema-less, and the number of columns in each data is not fixed. On the **Map Field** page, there is a high probability that all columns cannot be obtained by obtaining example values. In this case, the data on the migration destination is incomplete after the job is executed.

To solve this problem, perform any of the following methods:

1. Add fields on the **Map Field** tab page.
2. Edit the JSON file of the job on the **Job Management** page (modify the **fromJobConfig.columns** and **toJobConfig.columnList** parameters).
3. Export the JSON file of the job to the local PC, modify the parameters in the JSON file (the principle is the same to that in 2), and then import the JSON file back to CDM.

You are advised to perform 1. The following uses data migration from HBase to DWS as an example.

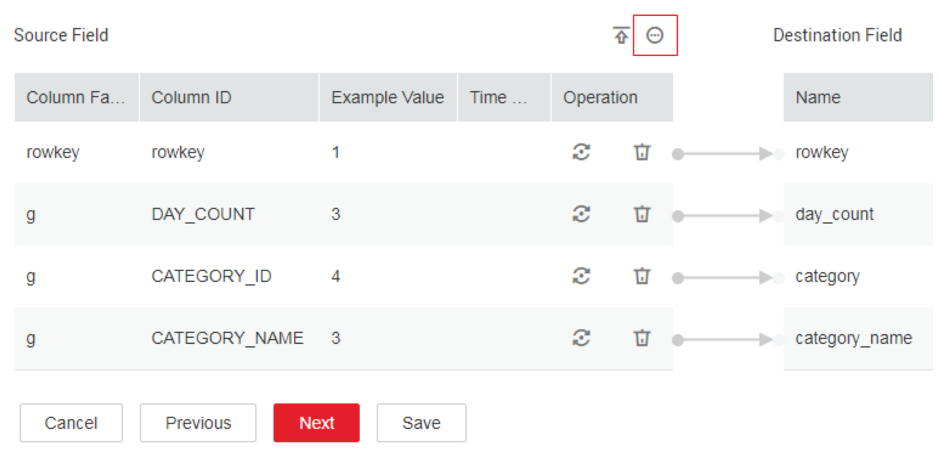
Solution 1: Adding Fields on the Map Field Tab Page

1. Obtain all fields in the tables to be migrated from source HBase. Use colons (:) to separate column families and columns. The following gives an example:

```
rowkey:rowkey
g:DAY_COUNT
g:CATEGORY_ID
g:CATEGORY_NAME
g:FIND_TIME
g:UPLOAD_PEOPLE
g:ID
g:INFOMATION_ID
g:TITLE
g:COORDINATE_X
g:COORDINATE_Y
g:COORDINATE_Z
g:CONTENT
g:IMAGES
g:STATE
```

2. On the **Job Management** page, locate the job for exporting data from HBase to DWS, click **Edit** in the row where the job resides, and go to the **Map Field** tab page.

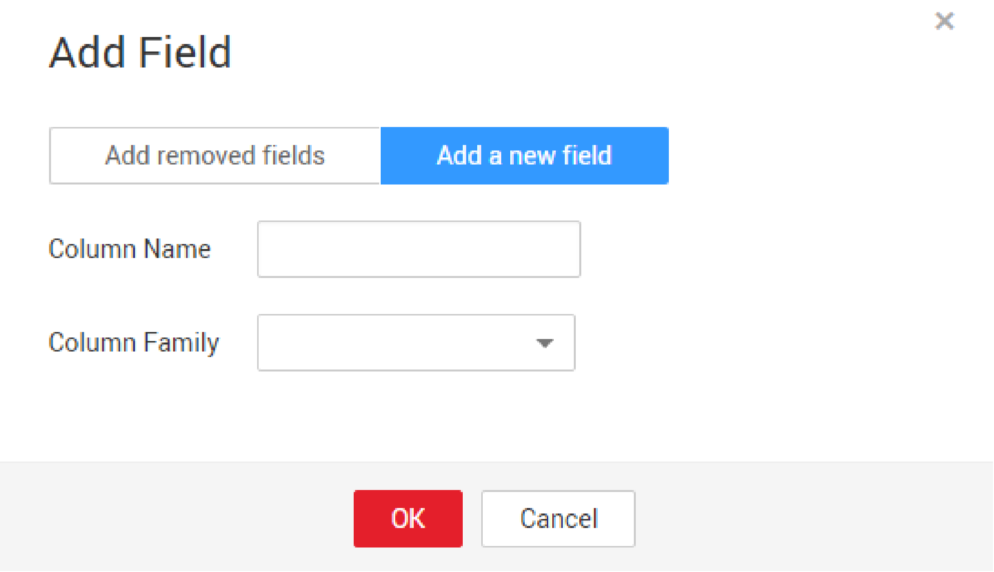
Figure 3-13 Field mapping 03



Column Fa...	Column ID	Example Value	Time ...	Operation	Name
rowkey	rowkey	1			rowkey
g	DAY_COUNT	3			day_count
g	CATEGORY_ID	4			category
g	CATEGORY_NAME	3			category_name

3. Click ⊕. In the dialog box that is displayed, select **Add a new field**.

Figure 3-14 Adding a field 04



NOTE

- After a field is added, the example value of the new field is not displayed on the console. This does not affect the transmission of field values. CDM directly writes the field values to the migration destination.
- To add new fields, the migration source must be MongoDB, HBase, relational databases, or Redis (data in Redis must be in the Hash format).

4. After all fields are added, check whether the mapping between the migration source and destination is correct. If the mapping is incorrect, drag the fields to adjust the field mapping.
5. Click **Next** and **Save**.

Solution 2: Modifying a JSON File

1. Obtain all fields in the tables to be migrated from source HBase. Use colons (:) to separate column families and columns. The following gives an example:

```
rowkey:rowkey
g:DAY_COUNT
g:CATEGORY_ID
g:CATEGORY_NAME
g:FIND_TIME
g:UPLOAD_PEOPLE
g:ID
g:INFOMATION_ID
g:TITLE
g:COORDINATE_X
g:COORDINATE_Y
g:COORDINATE_Z
g:CONTENT
g:IMAGES
g:STATE
```

2. In the DWS destination table, obtain the fields corresponding to the HBase table fields.

   If any field name corresponding to the HBase field does not exist in the DWS destination table, add it to the DWS table schema. Suppose that the fields in the DWS table are complete and are displayed as follows:

```
rowkey
day_count
category
category_name
find_time
upload_people
id
information_id
title
coordinate_x
coordinate_y
coordinate_z
content
images
state
```

3. On the **Job Management** page, locate the job for exporting data from HBase to DWS, and choose **More** > **Edit Job JSON** in the row where the job resides.
4. On the page that is displayed, edit the JSON file of the job.
   a. Modify the **fromJobConfig.columns** parameter of the migration source to the HBase fields obtained in 1. Use & to separate column numbers and colons (:) to separate column families and columns. The following gives an example:

```
"from-config-values": {
                "configs": [
                    {
                        "inputs": [
                            {
                                "name": "fromJobConfig.table",
                                "value": "HBase"
                            },
                            {
                                "name": "fromJobConfig.columns",
                                "value": "rowkey:rowkey&g:DAY_COUNT&g:CATEGORY_ID&g:CATEGORY_NAME&g:FIND_TIME&g:UPLOAD_PEOPLE&g:ID&g:INFOMATION_ID&g:TITLE&g:COORDINATE_X&g:COORDINATE_Y&g:COORDINATE_Z&g:CONTENT&g:IMAGES&g:STATE"
                            },
                            {
                                "name": "fromJobConfig.formats",
                                "value": {
                                    "2": "yyyy-MM-dd",
                                    "undefined": "yyyy-MM-dd"
                                }
                            }
                        ],
                        "name": "fromJobConfig"
                    }
                ]
            }
```

   b. Modify the **toJobConfig.columnList** parameter of the migration source to the field list of DWS obtained in 2.

   The sequence must be the same as that of HBase to ensure correct field mapping. Use & to separate field names. The following gives an example:

```
"to-config-values": {
                "configs": [
                    {
                        "inputs": [
                            {
                                "name": "toJobConfig.schemaName",
                                "value": "dbadmin"
                            },
                            {
                                "name": "toJobConfig.tablePreparation",
                                "value": "DO_NOTHING"
                            },
                            {
                                "name": "toJobConfig.tableName",
                                "value": "DWS "
                            },
                            {
                                "name": "toJobConfig.columnList",
                                "value":
"rowkey&day_count&category&category_name&find_time&upload_people&id&information
_id&title&coordinate_x&coordinate_y&coordinate_z&content&images&state"
                            },
                            {
                                "name": "toJobConfig.shouldClearTable",
                                "value": "true"
                            }
                        ],
                        "name": "toJobConfig"
                    }
                ]
            }
```

c. Retain the settings of other parameters, and then click **Save and Run**.

5. After the job is completed, check whether the data in the DWS table matches the data in HBase. If the mapping is incorrect, check whether the sequences of the HBase and DWS fields in the JSON file are the same.

3.30 How Do I Select Distribution Columns When Using CDM to Migrate Data to GaussDB(DWS)?

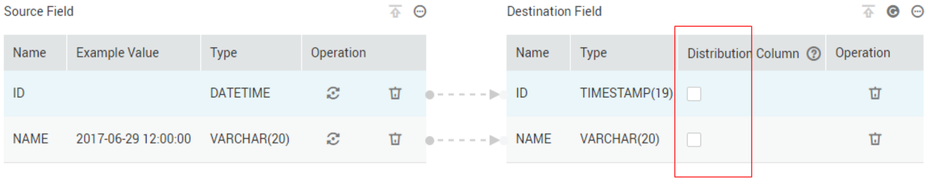
Symptom

How do I select distribution columns when using CDM to migrate data to GaussDB(DWS)?

Solution

When using CDM to migrate data to DWS or FusionInsight LibrA and create a table on DWS, select the distribution columns on the **Map Field** tab page.

Figure 3-15 Selecting distribution columns



Source Field

Name	Example Value	Type	Operation
ID		DATETIME	
NAME	2017-06-29 12:00:00	VARCHAR(20)	

Destination Field

Name	Type	Distribution Column	Operation
ID	TIMESTAMP(19)	□	
NAME	VARCHAR(20)	□	

Selecting the distribution column is very important for the running of DWS/FusionInsight LibrA. When migrating data to DWS/FusionInsight LibrA, you are advised to specify the distribution column according to the following principles:

1. Use the primary key as the distribution column.
2. If multiple data segments are combined as primary keys, specify all primary keys as the distribution column.

3. In the scenario where no primary key is available, if no distribution column is selected, DWS uses the first column as the distribution column by default. As a result, data skew risks exist.

Therefore, when a single table or entire database is imported to DWS/FusionInsight LibrA, you are advised to manually select a distribution column; otherwise, CDM automatically selects one. For more information about distribution columns, see **GaussDB(DWS)**.

If the DWS primary key or table contains only one field, the field type must be a common character string, value, or date. When data is migrated from another database to DWS, if automatic table creation is selected, the primary key must be of the following types. If no primary key is set, at least one of the following fields must be set. Otherwise, the table cannot be created and the CDM job fails.

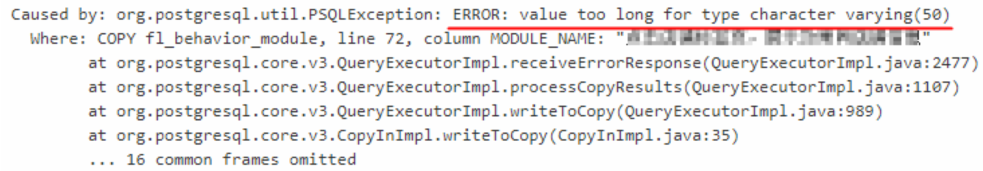
- INTEGER TYPES: TINYINT, SMALLINT, INT, BIGINT, NUMERIC/DECIMAL
- CHARACTER TYPES: CHAR, BPCHAR, VARCHAR, VARCHAR2, NVARCHAR2, TEXT
- DATA/TIME TYPES: DATE, TIME, TIMETZ, TIMESTAMP, TIMESTAMPTZ, INTERVAL, SMALLDATETIME

3.31 What Do I Do If the Error Message "value too long for type character varying" Is Displayed When I Migrate Data to DWS?

Symptom

When you use CDM to migrate data to DWS/FusionInsight LibrA, the migration fails and the error message "**value too long for type character varying**" is displayed in the log. See **Figure 3-16**.

Figure 3-16 Log output



```
Caused by: org.postgresql.util.PSQLException: ERROR: value too long for type character varying(50)
  Where: COPY fl_behavior_module, line 72, column MODULE_NAME: "[redacted]"
        at org.postgresql.core.v3.QueryExecutorImpl.receiveErrorResponse(QueryExecutorImpl.java:2477)
        at org.postgresql.core.v3.QueryExecutorImpl.processCopyResults(QueryExecutorImpl.java:1107)
        at org.postgresql.core.v3.QueryExecutorImpl.writeToCopy(QueryExecutorImpl.java:989)
        at org.postgresql.core.v3.CopyInImpl.writeToCopy(CopyInImpl.java:35)
        ... 16 common frames omitted
```

Possible Cause

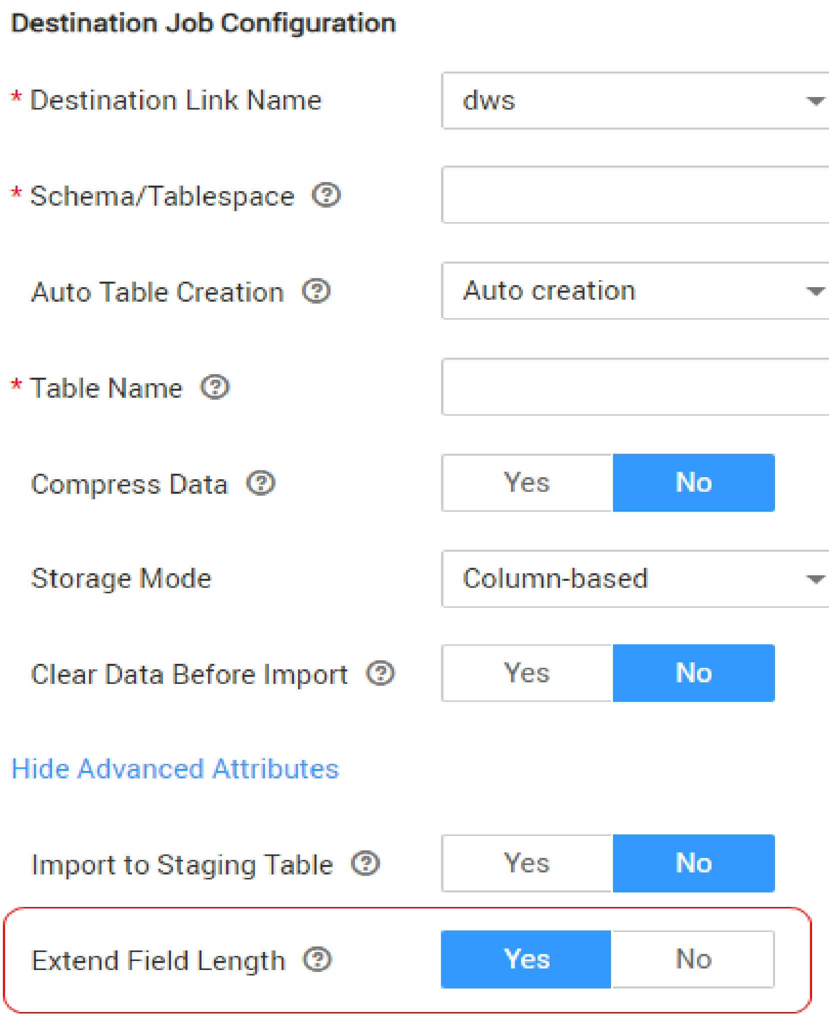
The data migrated to DWS is in Chinese, and the table is automatically created at the migration destination. The length of the varchar field of DWS is calculated by byte, and a Chinese character may occupy three bytes in UTF-8 encoding. If the length of a Chinese character exceeds that of the varchar field of DWS, an error occurs and the error message "**value too long for type character varying**" is displayed.

Solution

To solve this problem, you can select **Extend Field Length** to **Yes**, so that the length of the **varchar** field is automatically increased by three times when the destination table is created.

Edit the table/file migration job on CDM. In **Destination Job Configuration**, set **Auto Table Creation** to **Auto creation**, **Extend Field Length** is displayed in **Show Advanced Attributes**. Set **Extend Field Length** to **Yes**.

Figure 3-17 Extending field length



Destination Job Configuration

* Destination Link Name	dws
* Schema/Tablespace	
Auto Table Creation	Auto creation
* Table Name	
Compress Data	Yes [✓ No]
Storage Mode	Column-based
Clear Data Before Import	Yes [✓ No]

Hide Advanced Attributes

Import to Staging Table	Yes [✓ No]
Extend Field Length	[✓ Yes] No

3.32 What Can I Do If Error Message "Unable to execute the SQL statement" Is Displayed When I Import Data from OBS to SQL Server?

Symptom

When CDM is used to import data from OBS to SQL Server, the job fails to be executed and error message "Unable to execute the SQL statement. Cause: "String or binary data truncated" is displayed.

Possible Cause

The data in OBS exceeds the length limit of the SQL Server database.

Solution

When creating a table in the SQL Server database, increase the length of the database field. The length of the database field must be greater than that of the data in OBS.

3.33 What Should I Do If the Cluster List Is Empty, I Have No Access Permission, or My Operation Is Denied?

Symptom

When using CDM, you may encounter the following permission-related issues:

- The cluster list is empty on the CDM homepage.
- A message is displayed, indicating that you do not have the access permission.
- When you try to start a job or restart a cluster, an error message is displayed, indicating that your operation is denied by the current policy.

Possible Cause

The preceding issues are caused by incorrect permission configuration.

Solution

- If CDM is a module of DataArts Studio, perform the following operations:
  a. Check whether the DAYU Administrator or DAYU User role has been added. For details, see **DataArts Studio Permissions Management**.
  b. Check whether you have the permission (developer or viewer role) to access the workspace. For details, see **DataArts Studio Permissions**.
- If CDM is an independent service, perform the following operations:

a. Check whether IAM fine-grained authentication is enabled.

  - If it is disabled, check whether the **CDM Administrator** role has been added to the user group.
  - If it is enabled, go to 2.

b. Check whether a custom or preset policy has been added to enable you to access CDM, such as the **CDM FullAccess** and **CDM ReadOnlyAccess** policies. For details, see **Permissions Management**.

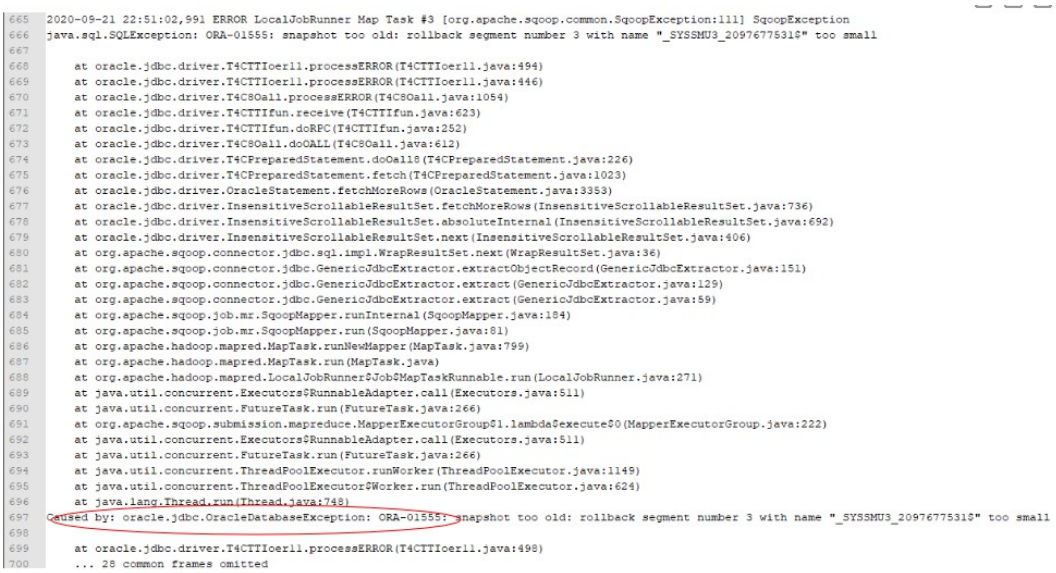
c. Check whether an access denial policy has been added to the enterprise project.

3.34 Why Is Error ORA-01555 Reported During Migration from Oracle to DWS?

Symptom

When CDM is used to migrate Oracle data to DWS, an error is reported, as shown in **Figure 3-18**.

Figure 3-18 Symptom



```
2020-09-21 22:51:02,991 ERROR LocalJobRunner Map Task #3 [org.apache.sqoop.common.SqoopException:111] SqoopException
java.sql.SQLException: ORA-01555: snapshot too old: rollback segment number 3 with name "_SYSSMU3_2097677531$" too small

    at oracle.jdbc.driver.T4CTTIoer11.processERROR(T4CTTIoer11.java:494)
    at oracle.jdbc.driver.T4CTTIoer11.processERROR(T4CTTIoer11.java:446)
    at oracle.jdbc.driver.T4C8Oall.processERROR(T4C8Oall.java:1054)
    at oracle.jdbc.driver.T4CTTIfun.receive(T4CTTIfun.java:623)
    at oracle.jdbc.driver.T4CTTIfun.doRPC(T4CTTIfun.java:252)
    at oracle.jdbc.driver.T4C8Oall.doOALL(T4C8Oall.java:612)
    at oracle.jdbc.driver.T4CPreparedStatement.doOall8(T4CPreparedStatement.java:226)
    at oracle.jdbc.driver.T4CPreparedStatement.fetch(T4CPreparedStatement.java:1023)
    at oracle.jdbc.driver.OracleStatement.fetchMoreRows(OracleStatement.java:3353)
    at oracle.jdbc.driver.InsensitiveScrollableResultSet.fetchMoreRows(InsensitiveScrollableResultSet.java:736)
    at oracle.jdbc.driver.InsensitiveScrollableResultSet.absoluteInternal(InsensitiveScrollableResultSet.java:692)
    at oracle.jdbc.driver.InsensitiveScrollableResultSet.next(InsensitiveScrollableResultSet.java:406)
    at org.apache.sqoop.connector.jdbc.sql.impl.WrapResultSet.next(WrapResultSet.java:36)
    at org.apache.sqoop.connector.jdbc.GenericJdbcExtractor.extractObjectRecord(GenericJdbcExtractor.java:151)
    at org.apache.sqoop.connector.jdbc.GenericJdbcExtractor.extract(GenericJdbcExtractor.java:129)
    at org.apache.sqoop.connector.jdbc.GenericJdbcExtractor.extract(GenericJdbcExtractor.java:59)
    at org.apache.sqoop.job.mr.SqoopMapper.runInternal(SqoopMapper.java:184)
    at org.apache.sqoop.job.mr.SqoopMapper.run(SqoopMapper.java:81)
    at org.apache.hadoop.mapred.MapTask.runNewMapper(MapTask.java:799)
    at org.apache.hadoop.mapred.MapTask.run(MapTask.java)
    at org.apache.hadoop.mapred.LocalJobRunner$Job$MapTaskRunnable.run(LocalJobRunner.java:271)
    at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
    at java.util.concurrent.FutureTask.run(FutureTask.java:266)
    at org.apache.sqoop.submission.mapreduce.MapperExecutorGroup$1.lambda$execute$0(MapperExecutorGroup.java:222)
    at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
    at java.util.concurrent.FutureTask.run(FutureTask.java:266)
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
    at java.lang.Thread.run(Thread.java:748)
Caused by: oracle.jdbc.OracleDatabaseException: ORA-01555: snapshot too old: rollback segment number 3 with name "_SYSSMU3_2097677531$" too small

    at oracle.jdbc.driver.T4CTTIoer11.processERROR(T4CTTIoer11.java:498)
    ... 28 common frames omitted
```

Cause Analysis

1. During data migration, if the entire table is queried and the table contains a large amount of data, the query takes a long time.
2. During the query, other users frequently perform the **commit** operation.
3. The RBS (the tablespace used for rollback) of Oracle is small. As a result, the migration task is not complete, the source database has been updated, and the rollback times out.

Summary and Suggestions

1. Reduce the data volume queried each time.
2. Modify the database configurations to increase the RBS of the Oracle database.

3.35 What Should I Do If the MongoDB Connection Migration Fails?

Symptom

What should I do If the MongoDB connection migration fails?

Solution

By default, the **userAdmin** role has only the permissions to manage roles and users and does not have the read and write permissions on a database.

If the MongoDB connection fails to be migrated, you need to view the user permission information in the MongoDB connection to ensure that the user has the read and write permissions on the specified database.

3.36 What Should I Do If a Hive Migration Job Is Suspended for a Long Period of Time?

Symptom

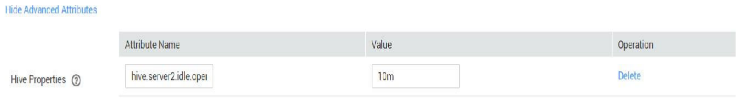
What should I do If a Hive migration job is suspended for a long period of time?

Solution

Manually stop the Hive migration job and add the following attribute settings to the Hive data connection:

- **Attribute Name**: **hive.server2.idle.operation.timeout**
- **Value**: **10m**

In the figure on the left:



Hive Properties

Attribute Name	Value	Operation
hive.server2.idle.oper	10m	Delete

3.37 What Should I Do If an Error Is Reported Because the Field Type Mapping Does Not Match During Data Migration Using CDM?

Symptom

When you use CDM to migrate data to DWS, the migration job fails and the error message "value too long for type character varying" is displayed in the execution log.

Possible Cause

The possible cause is that the type of the source table does not match that of the target table. For example, the **dli** field of the source is of the string type, and the **dws** field of the destination is of the varchar(50) type. As a result, the precision is default and the error message "value too long for type character varying" is reported. This issue also occurs for conversion from string to bigint and from bigint to int.

Solution

- Locate the field that is incorrectly mapped based on the error information and contact the DBA to modify the table structure.
- If this issue occurs only for a small amount of data, you can configure the dirty data policy to solve the issue.

3.38 What Should I Do If a JDBC Connection Timeout Error Is Reported During MySQL Migration?

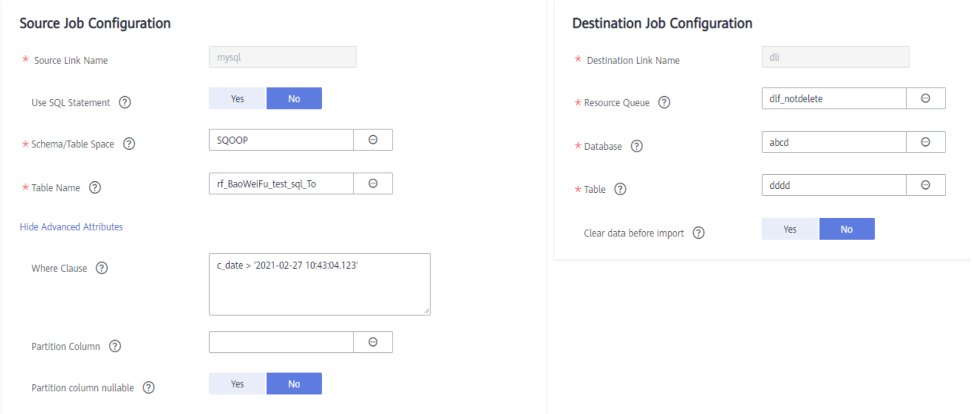
Symptom

The following error message is displayed during MySQL migration: "Unable to connect to the database server. Cause: connect timed out."

Possible Cause

The table has a large data volume, and the source end uses the where statement to filter data. However, the column is not an index column or the column values are not discrete. As a result, the entire table is scanned during the query, causing a JDBC connection timeout. The **c_date** field is not an index column, as shown in **Figure 3-19**.

Figure 3-19 Non-index column



Source Job Configuration

* Source Link Name	mysql
Use SQL Statement	Yes [✓ No]
* Schema/Table Space	SQOOP
* Table Name	rf_BaoWeiFu_test_sql_To
Hide Advanced Attributes	
Where Clause	c_date > '2021-02-27 10:43:04.123'
Partition Column	
Partition column nullable	Yes [✓ No]

Destination Job Configuration

* Destination Link Name	dli
* Resource Queue	dlf_notdelete
* Database	abcd
* Table	dddd
Clear data before import	Yes [✓ No]

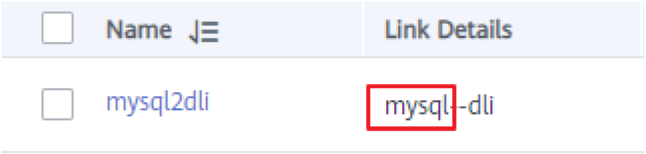
Solution

1. Contact the DBA to modify the table structure, set the columns to be filtered as index columns, and try again.

   If the failure persists because the data is not discrete, perform 2 to 4 and increase the JDBC timeout duration.

2. Locate the MySQL link name based on the job and obtain the link information.

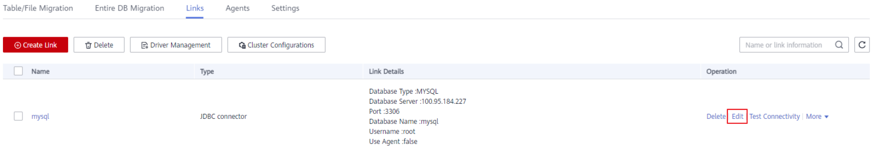
   Figure 3-20 Link information

□ Name	Link Details
□ mysql2dli	mysql--dli

3. Click the **Links** tab and click **Edit** to edit the link.

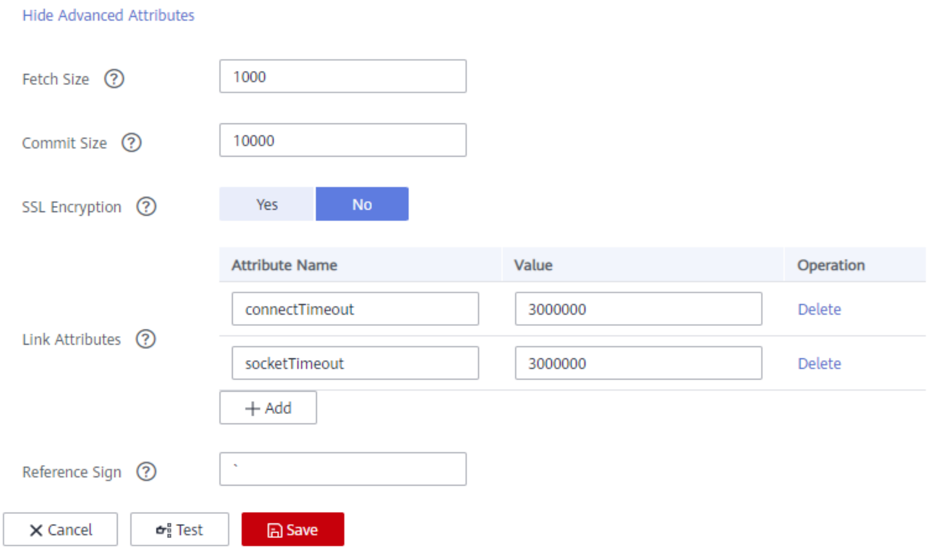
   Figure 3-21 Editing the link

□ Name	Type	Link Details	Operation
□ mysql	JDBC connector	Database Type :MYSQL Database Server :100.95.184.227 Port :3306 Database Name :mysql Username :root Use Agent :false	Delete | Edit | Test Connectivity | More ▾

4. Click **Show Advanced Attributes**, add parameters **connectTimeout** and **socketTimeout** and their values in **Link Attributes** , and click **Save**.

Figure 3-22 Editing advanced attributes



Fetch Size: 1000

Commit Size: 10000

SSL Encryption: Yes [✓ No]

Link Attributes:

Attribute Name	Value	Operation
connectTimeout	3000000	Delete
socketTimeout	3000000	Delete

Reference Sign: `

3.39 What Should I Do If a CDM Migration Job Fails After a Link from Hive to GaussDB(DWS) Is Created?

Symptom

A CDM migration job fails after a link from Hive to GaussDB(DWS) is created.

Solution

You are advised to clear historical data and try again. In addition, when creating a migration job, you are advised to enable the system to clear historical data. This greatly reduces the probability of failures.

3.40 How Do I Use CDM to Export MySQL Data to an SQL File and Upload the File to an OBS Bucket?

Symptom

How do I use CDM to export MySQL data to an SQL file and upload the file to an OBS bucket?

Solution

CDM does not support this operation. You are advised to manually export a MySQL data file, enable the SFTP service on the server, and create a CDM job with SFTP as the source and OBS as the destination. Then you can execute the created job to transfer the file.

3.41 What Should I Do If CDM Fails to Migrate Data from OBS to DLI?

Symptom

CDM fails to migrate data from OBS to DLI.

Solution

Dirty data writing is configured, but no dirty data exists. You need to decrease the number of concurrent tasks to avoid this issue.

3.42 What Should I Do If a CDM Connector Reports the Error "Configuration Item [linkConfig.iamAuth] Does Not Exist"?

This error is reported because the customer's certificate has expired. Update the certificate and reconfigure the connector.

3.43 What Should I Do If Error "Configuration Item [linkConfig.createBackendLinks] Does Not Exist" or "Configuration Item [throttlingConfig.concurrentSubJobs] Does Not Exist" Is Reported?

Symptom

Error message "Configuration Item [linkConfig.createBackendLinks] Does Not Exist" or "Configuration Item [throttlingConfig.concurrentSubJobs] Does Not Exist" is displayed during job creation.

Possible Cause

If you create a link or save a job in a CDM cluster of an earlier version, and then access a CDM cluster of a later version, this error occurs occasionally.

Solution

Manually clear the browser cache to avoid this error.

3.44 What Should I Do If Message "CORE_0031:Connect time out. (Cdm.0523)" Is Displayed During the Creation of an MRS Hive Link?

Symptom

Message "CORE_0031:Connect time out. (Cdm.0523)" is displayed during the creation of an MRS Hive link.

Solution

If a message is displayed indicating that the configuration file cannot be downloaded during the creation of an MRS Hive link, insufficient user permissions occur. You are advised to create another service user, grant required permissions to the user, and try again.

To create a data connection for an MRS security cluster, do not use user **admin**. The **admin** user is the default management page user and cannot be used as the authentication user of the security cluster. You can create an MRS user and set **Username** and **Password** to the username and password of the created MRS user when creating an MRS data connection.

NOTE

- If the CDM cluster version is 2.9.0 or later and the MRS cluster version is 3.1.0 or later, the created user must have the permissions of the **Manager_viewer** role to create links on CDM. To perform operations on databases, tables, and columns of an MRS component, you also need to add the database, table, and column permissions of the MRS component to the user by following the instructions in the MRS documentation.
- If the CDM cluster version is earlier than 2.9.0 or the MRS cluster version is earlier than 3.1.0, the created user must have the permissions of **Manager_administrator** or **System_administrator** to create links on CDM.
- A user with only the **Manager_tenant** or **Manager_auditor** permission cannot create connections.

3.45 What Should I Do If Message "CDM Does Not Support Auto Creation of an Empty Table with No Column" Is Displayed When I Enable Auto Table Creation?

Symptom

Message "CDM Does Not Support Auto Creation of an Empty Table with No Column" is displayed when I enable auto table creation.

Possible Cause

The cause is that the database table name contains special characters, resulting in incorrect syntax. You can resolve this issue by renaming the database table according to the naming rules for database objects.

For example, the name of a data table in the DWS data warehouse can contain a maximum of 63 characters and support letters, digits, underscores (_), dollar signs ($), and number signs (#), and must start with a letter or underscore (_).

Solution

Rename the table.

3.46 What Should I Do If I Cannot Obtain the Schema Name When Creating an Oracle Relational Database Migration Job?

Symptom

I cannot obtain the schema name when creating an Oracle relational database migration job.

Possible Cause

The latest ORACLE_8 driver, which is not supported, may have been uploaded, for example, the Oracle Database 21c (21.3) driver.

Solution

You are advised to use the ojdbc8.jar driver in Oracle Database 12c. You can download the driver from **https://www.oracle.com/database/technologies/jdbc-ucp-122-downloads.html**.

3.47 What Should I Do If invalid input syntax for integer: "true" Is Displayed During MySQL Database Migration?

Symptom

The MySQL database stores values **0** and **1**, rather than **true** and **false**. However, **true** or **false** is read during MySQL database migration, and the following error information is displayed: Unable to execute the SQL statement. Cause: ERROR: invalid input syntax for integer: "true" Where: COPY sd_mask_ext, line 1, column mask_type.

Possible Cause

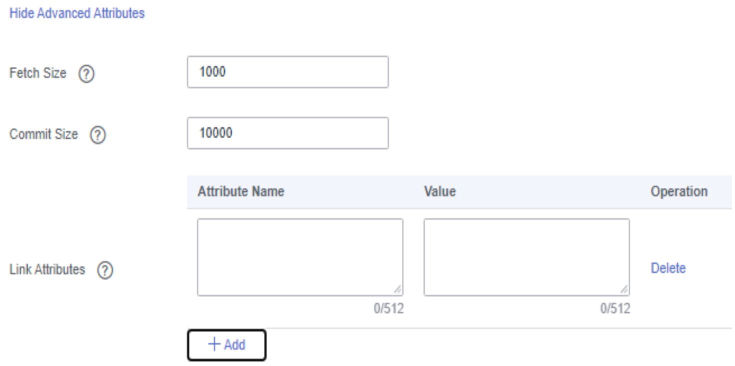
By default, **tinyInt1isBit** is set to **true** for MySQL databases. As a result, **TINYINT(1)** is processed as **BIT** (that is, **Types.BOOLEAN**), and **1** or **0** is read as **true** or **false**.

Solution

In the advanced attributes of the MySQL link, add either of the following parameters so that tables can be properly created at the destination:

- Parameter **tinyInt1isBit**, with its value set to **false**
- Parameter **mysql.bool.type.transform**, with its value set to **false**

Figure 3-23 Adding link attributes



Fetch Size: 1000

Commit Size: 10000

Link Attributes:

Attribute Name	Value	Operation
		Delete

4 DataArts Migration (Real-Time Jobs)

4.1 Overview

Table 4-1 Questions

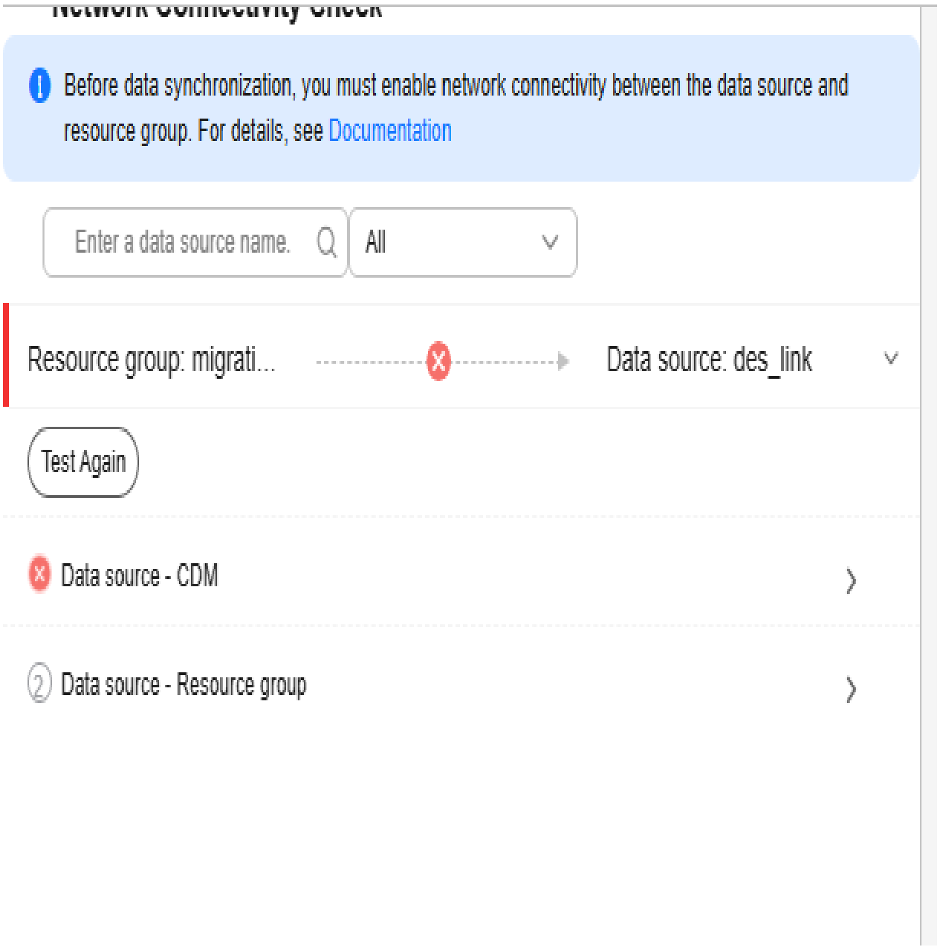
Type	Reference
Network connectivity	**How Do I Troubleshoot a Network Disconnection Between the Data Source and Resource Group?**
Network connectivity	**Which Ports Must Be Allowed by the Data Source Security Group So That DataArts Migration Can Access the Data Source?**
Hudi	**How Do I Configure a Spark Periodic Task for Hudi Compaction?**
GaussDB(DWS)	**What Should I Do If an Error Is Reported During DDL Synchronization of New Columns in a Real-Time MySQL-to-DWS Synchronization Job?**
	Why Does DWS Filter the Null Value of the Primary Key During Real-Time Synchronization from MySQL to DWS?
Kafka	**What Should I Do If a Job for Synchronizing Data from Kafka to DLI in Real Time Fails and "Array element access needs an index starting at 1 but was 0" Is Displayed?**
Oracle	**How Do I Grant the Log Archiving, Query, and Parsing Permissions of an Oracle Data Source?**
PostgreSQL	**How Do I Manually Delete Replication Slots from a PostgreSQL Data Source?**

4.2 How Do I Troubleshoot a Network Disconnection Between the Data Source and Resource Group?

Symptom

During the configuration of a real-time migration job, an exception is reported indicating that the data sources at the source and destination are disconnected from the resource group.

Figure 4-1 Abnormal connectivity



Before data synchronization, you must enable network connectivity between the data source and resource group. For details, see Documentation

Network Connectivity Check
Enter a data source name.
All
Resource group: migrati...
Data source: des_link
Test Again
Data source - CDM
Data source - Resource group

Solution

Troubleshoot this issue based on the following table.

Table 4-2 Troubleshooting network disconnection

Type	Exception	Method
Data source - CDM exception	Instance status	Check whether the cluster is running properly.
	Connectivity	1. If the CDM cluster and the data source are in the same VPC, ensure that the private IP address of the CDM cluster has been added to the inbound rule of the security group of the data source and that the IP address of the data source has been added to the outbound rule of the security group of the CDM cluster. 2. If the CDM cluster and the data source are not in the same VPC, create a VPC peering connection to connect them. Add the private IP address of the CDM cluster to the inbound security group rule of the data source and add the data source IP address to the outbound security group rule of the CDM cluster. For details, see **Creating a DataArts Studio Data Connection**.
Data source - resource group exception	Resource group status	Access a DataArts Studio instance and click the **Resources** tab. On the displayed **Real-Time Resources** tab page, check whether the resource group is running.
	Connectivity	1. Access the **Management Center** page, choose **Manage Data Connections**, and check whether the IP address or domain name of the data connection is the private IP address and whether an agent is associated with the connection. 2. Access a DataArts Studio instance and click the **Resources** tab and then the **Real-Time Network Connections** tab. Check whether a network connection to the VPC and subnet of the data source has been created and whether the network connection has been associated with a resource group. 3. Check whether the network segment of the resource group is allowed by the inbound rule of the security group of the data source instance. 4. Verify that all configurations are correct by referring to **Enabling Network Communications**.

4.3 Which Ports Must Be Allowed by the Data Source Security Group So That DataArts Migration Can Access the Data Source?

Symptom

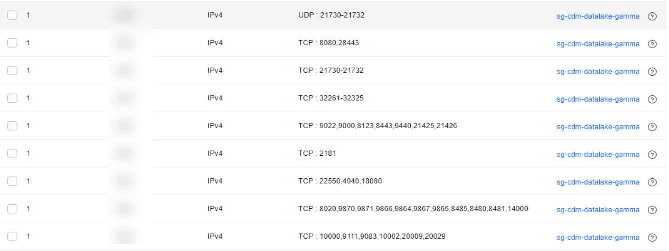
To enable communications between a resource group and a data source, you must allow the network segment of the resource group to access required ports in the security group to which the data source belongs.

Solution

The ports used for different data sources vary. For details, see the official document of each data source.

The following table provides the ports of some data sources.

Table 4-3 Ports used for data sources

Data Source	Port
MySQL	3306
GaussDB(DWS)	8000
PostgreSQL	5432
Oracle	1521
Kafka	Non-security mode: 9092/9094 Security mode: 9093/9095
MRS Hudi	The ports used for MRS Hudi are complex. For details, see **Common Ports for MRS Cluster Services**. **Figure 4-2** Example security group rule for MRS Hudi  □ 1 [redacted] IPv4 UDP : 21730-21732 sg-cdm-datalake-gamma ⓘ □ 1 [redacted] IPv4 TCP : 8080,28443 sg-cdm-datalake-gamma ⓘ □ 1 [redacted] IPv4 TCP : 21730-21732 sg-cdm-datalake-gamma ⓘ □ 1 [redacted] IPv4 TCP : 32261-32325 sg-cdm-datalake-gamma ⓘ □ 1 [redacted] IPv4 TCP : 9022,9000,8123,8443,9440,21425,21426 sg-cdm-datalake-gamma ⓘ □ 1 [redacted] IPv4 TCP : 2181 sg-cdm-datalake-gamma ⓘ □ 1 [redacted] IPv4 TCP : 22550,4040,18080 sg-cdm-datalake-gamma ⓘ □ 1 [redacted] IPv4 TCP : 8020,9870,9871,9866,9864,9867,9865,8485,8480,8481,14000 sg-cdm-datalake-gamma ⓘ □ 1 [redacted] IPv4 TCP : 10000,9111,9083,10002,20009,20029 sg-cdm-datalake-gamma ⓘ

4.4 How Do I Configure a Spark Periodic Task for Hudi Compaction?

Symptom

When writing data to Hudi, DataArts Migration splits compaction tasks into Spark jobs and sends the Spark jobs to MRS for execution.

Solution

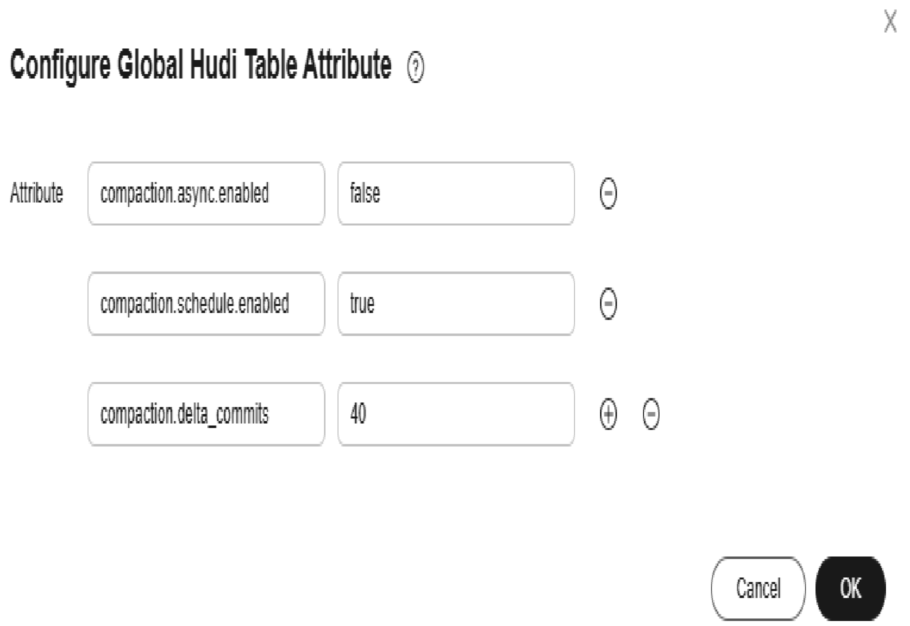
Step 1 Modify the configuration of the real-time migration job.

Disable asynchronous compaction, deletion of historical data files, and archiving for the migration job by configuring the parameters in **Global Configuration of Hudi Table Attributes** or **Edit Table Attribute**.

Table 4-4 Hudi table parameters

Parameter	Value	Description
compaction.schedule.enabled	true	Compaction plan generation is enabled.
compaction.delta_commits	60	Period for triggering compaction generated by the compaction plan
compaction.async.enabled	false	Asynchronous compaction is disabled.
clean.async.enabled	false	Data files of historical versions will be deleted.
hoodie.archive.automatic	false	Aging of Hudi commit files is enabled.

Figure 4-3 Disabling migration compaction



Configure Global Hudi Table Attribute

Attribute: compaction.async.enabled = false

compaction.schedule.enabled = true

compaction.delta_commits = 40

After the preceding parameters are set, no compaction task is executed after the job is started. Compaction plans are generated periodically. You can run the **run compaction on** command to execute compaction plans.

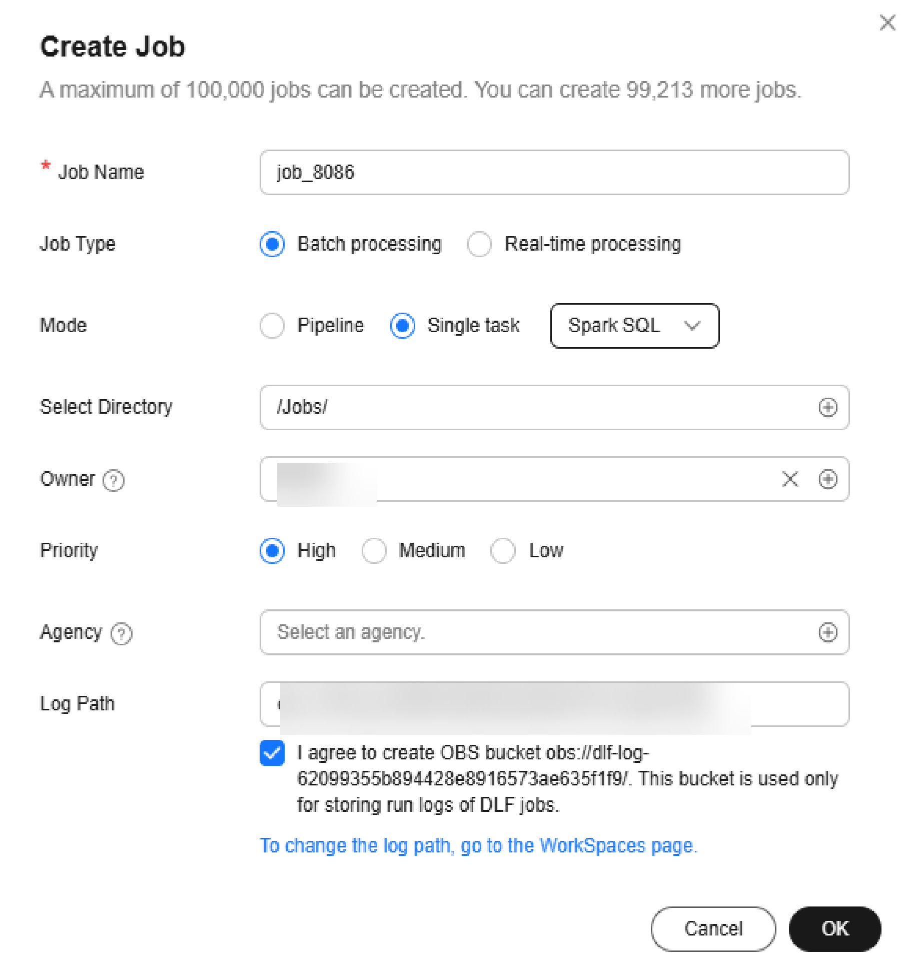
NOTE

Compaction plans must be generated by migration tasks and then executed by Spark. Otherwise, a Hudi timeline conflict will occur, causing the Spark compaction job to fail.

Step 2 Create a periodic compaction task for Spark SQL.

1. Go to the DataArts Factory console and create a Spark SQL job by following the instructions in **Developing a Batch Processing Single-Task SQL Job**.

Figure 4-4 Creating a single-task Spark SQL job



Create Job

A maximum of 100,000 jobs can be created. You can create 99,213 more jobs.

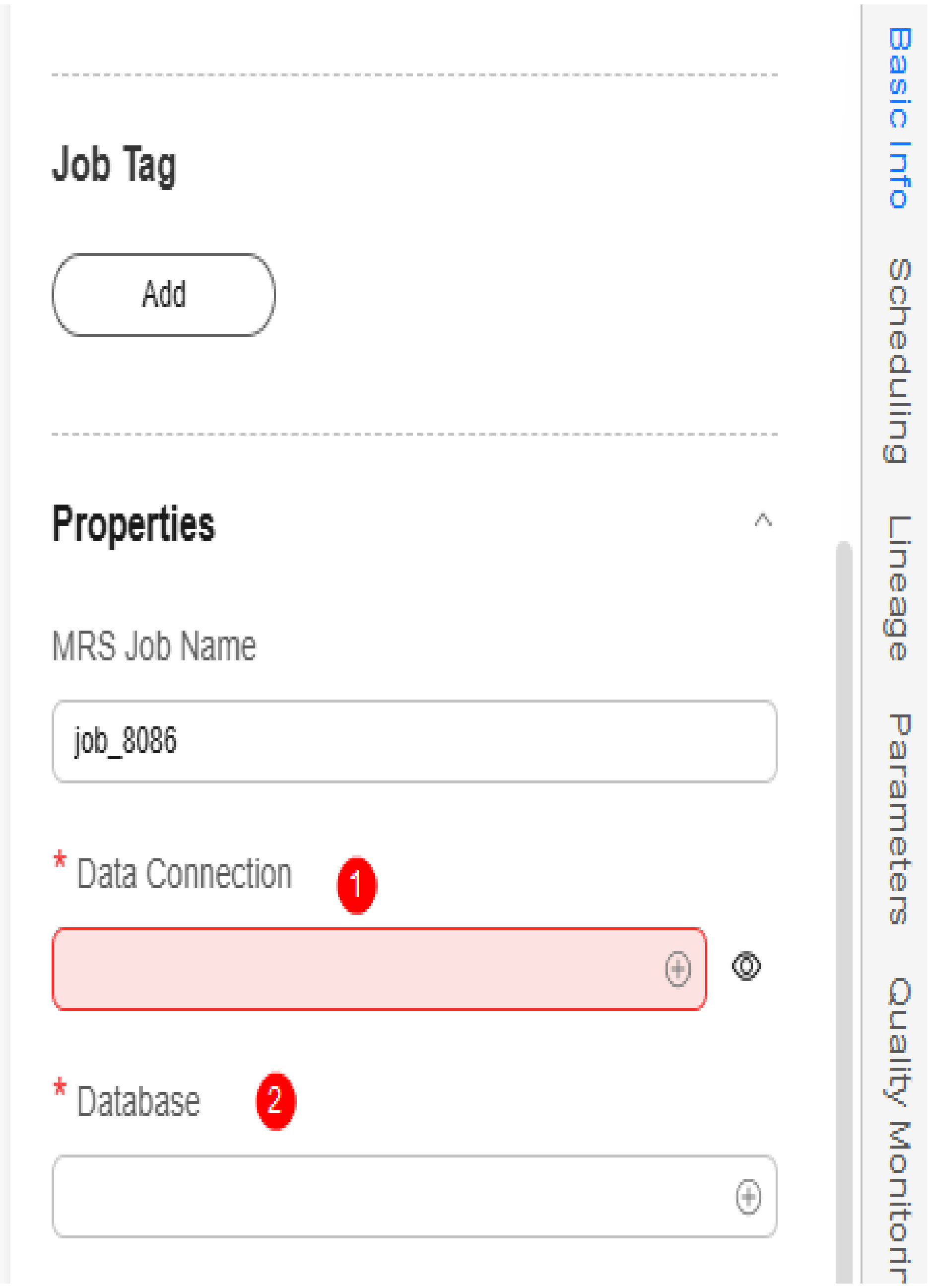
* Job Name	job_8086
Job Type	● Batch processing ○ Real-time processing
Mode	○ Pipeline ● Single task Spark SQL
Select Directory	/Jobs/
Owner	[redacted]
Priority	● High ○ Medium ○ Low
Agency	Select an agency.
Log Path	[redacted]

☑ I agree to create OBS bucket obs://dlf-log-62099355b894428e8916573ae635f1f9/. This bucket is used only for storing run logs of DLF jobs.

To change the log path, go to the WorkSpaces page.

2. Select the Spark data connection corresponding to Hudi and select the database to which the Hudi table belongs.

Figure 4-5 Configuring the connection and database



Job Tag

Add

Properties

MRS Job Name

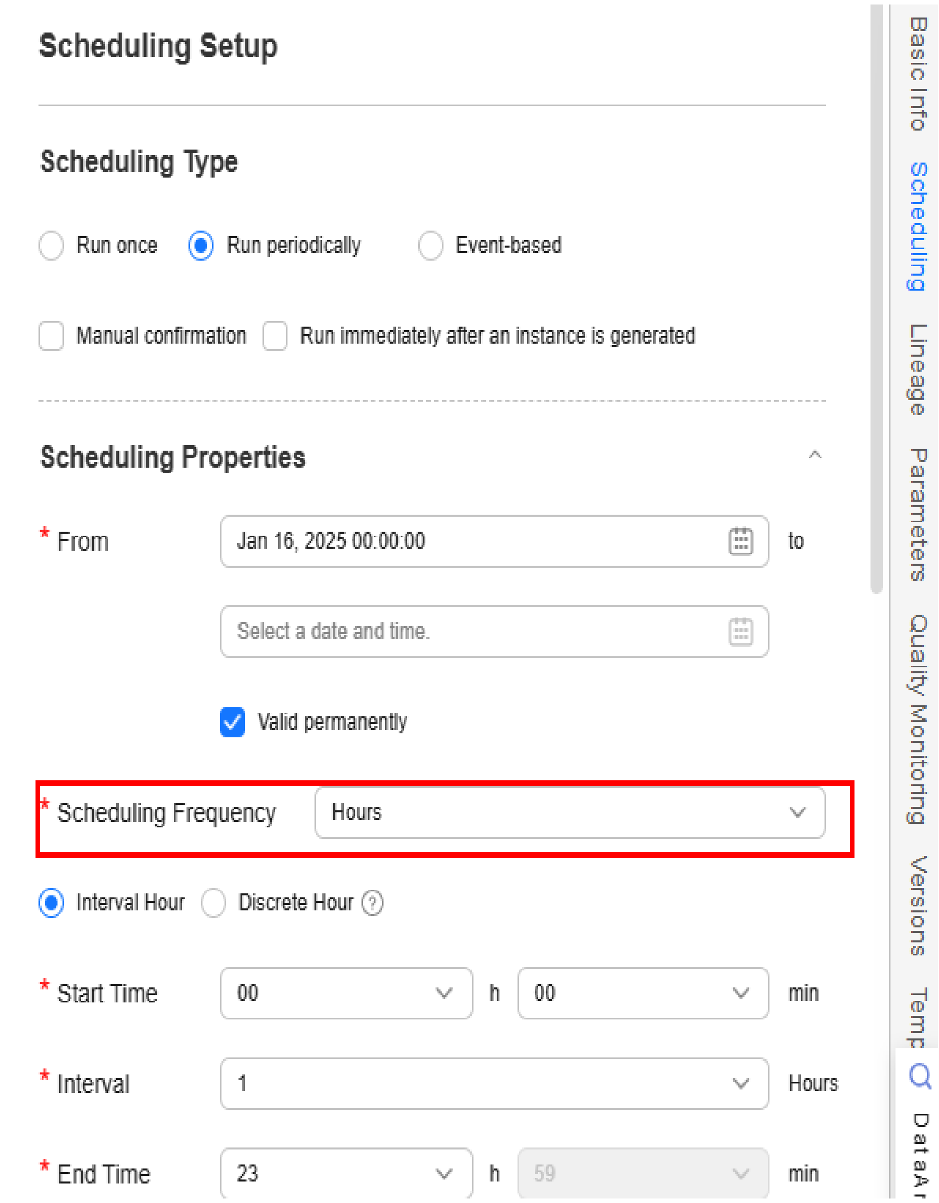
job_8086

* Data Connection

* Database

3. Configure the compaction scheduling period.

Figure 4-6 Configuring the scheduling period



Scheduling Setup

Scheduling Type

○ Run once ● Run periodically ○ Event-based

□ Manual confirmation □ Run immediately after an instance is generated

Scheduling Properties

* From: Jan 16, 2025 00:00:00 to

Select a date and time.

☑ Valid permanently

* Scheduling Frequency: Hours

● Interval Hour ○ Discrete Hour

* Start Time: 00 h 00 min

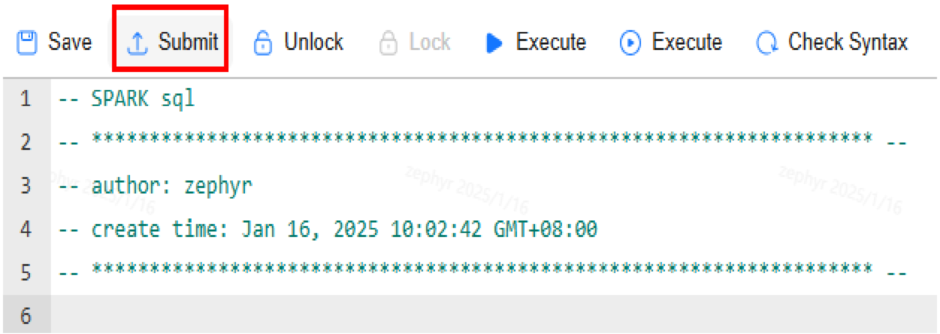
* Interval: 1 Hours

* End Time: 23 h 59 min

4. Enter the compaction statements of Spark SQL, and submit and run the job.

```
set hoodie.compact.inline = true;
set hoodie.run.compact.only.inline = true;
set hoodie.clean.automatic = false;
set hoodie.cleaner.commits.retained = 120;
set hoodie.keep.min.commits = 121;
set hoodie.keep.max.commits = 141;
run compaction on `db_name`.`table_name`;
run clean on `db_name`.`table_name`;
run archivelog on `db_name`.`table_name`;
```

Figure 4-7 Submitting and running the job



```
-- SPARK sql
-- ******************************************************************** --
-- author: zephyr
-- create time: Jan 16, 2025 10:02:42 GMT+08:00
-- ******************************************************************** --

```

----End

4.5 What Should I Do If an Error Is Reported During DDL Synchronization of New Columns in a Real-Time MySQL-to-DWS Synchronization Job?

Symptom

1. Run the real-time synchronization job of the migration mysql2dws link. During DDL synchronization, set the column adding operation to normal processing.

2. If the destination DWS table contains data, run the following DDL statement in the source MySQL database to add a column with a non-null constraint (the default value is an empty string): alter table test add column t_col varchar(30) not null default.

3. The migration job is abnormal, and an error message is displayed, indicating that the DDL statement fails to be executed. The failure cause is: column "t_col" contains null values.

Possible Cause

If the DWS database is compatible with Oracle, an empty string is processed as null. If there is data, a non-null column whose default value is an empty string cannot be added.

Solution

1. Modify the source DDL statement and set the default value of the new column to a non-null string.

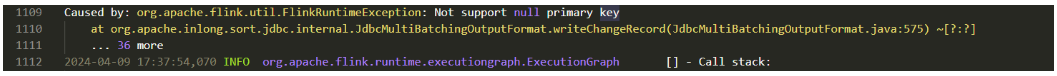
2. If the DDL cannot be modified, replace the DWS database with a MySQL-compatible database, which can be created using the following statement: create database bigdata with encoding 'UTF-8' dbcompatibility 'mysql' template template0;

4.6 Why Does DWS Filter the Null Value of the Primary Key During Real-Time Synchronization from MySQL to DWS?

Symptom

In a link from MySQL to DWS. The primary key of a manually created DWS table is different from that of the MySQL database. If the primary key field of DWS is a non-primary key field in MySQL and the field value is null in MySQL, an error is reported when data is written to DWS. The following figure shows the error message.

Figure 4-8 Error message



```
Caused by: org.apache.flink.util.FlinkRuntimeException: Not support null primary key
    at org.apache.inlong.sort.jdbc.internal.JdbcMultiBatchingOutputFormat.writeChangeRecord(JdbcMultiBatchingOutputFormat.java:575) ~[?:?]
    ... 36 more
2024-04-09 17:37:54,070 INFO  org.apache.flink.runtime.executiongraph.ExecutionGraph       [] - Call stack:
```

Possible Cause

In a link from MySQL to DWS. The primary key of a manually created DWS table is different from that of the MySQL database. If the primary key field of DWS is a non-primary key field in MySQL and the field value is null in MySQL, an error is reported.

Solution

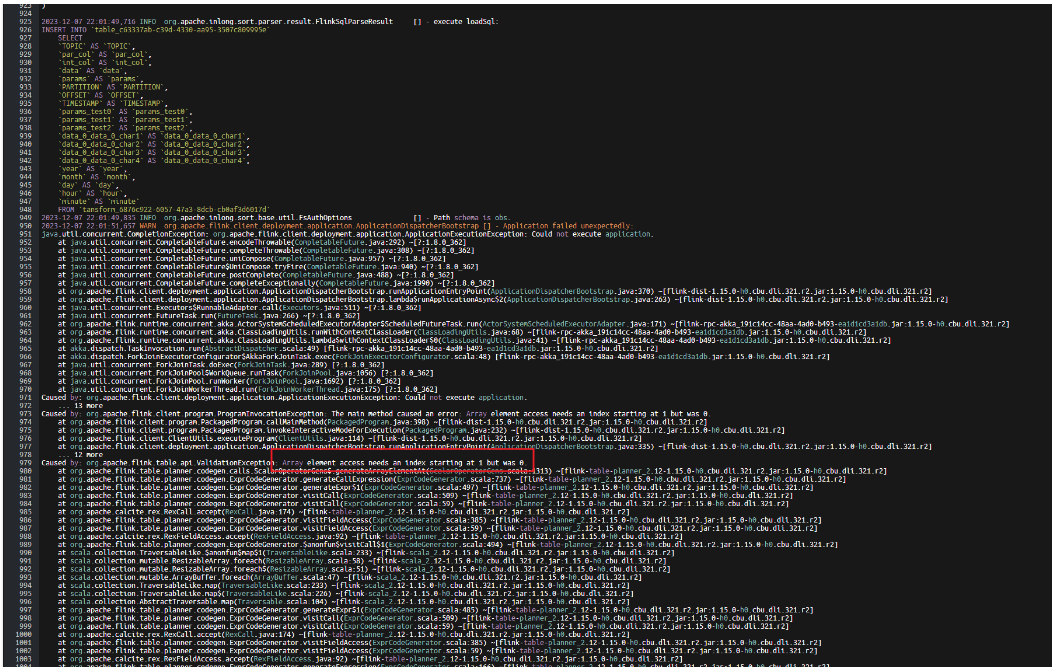
Before writing data to DWS, filter the primary key field of DWS. If the field is null, print a warning log and do not write the data.

4.7 What Should I Do If a Job for Synchronizing Data from Kafka to DLI in Real Time Fails and "Array element access needs an index starting at 1 but was 0" Is Displayed?

Symptom

A job for synchronizing data from Kafka to DLI in real time fails and "Array element access needs an index starting at 1 but was 0" is displayed.

Figure 4-9 Error message



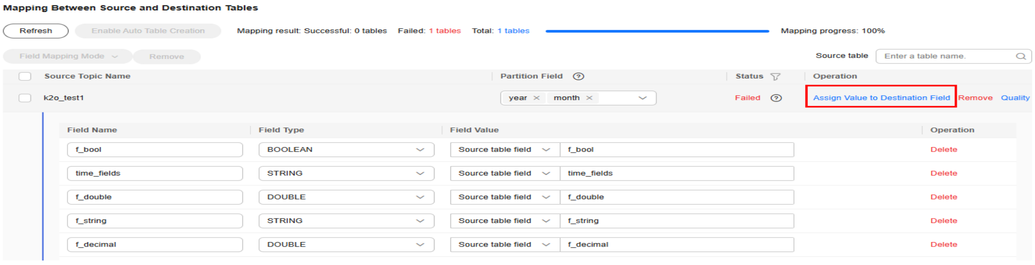
```
)

2023-12-07 22:01:49,716 INFO  org.apache.inlong.sort.parser.result.FlinkSqlParseResult     [] - execute loadSql:
INSERT INTO `table_c63337ab-c39d-4330-aa95-3507c809995e`
    SELECT
    `TOPIC` AS `TOPIC`,
    `par_col` AS `par_col`,
    `int_col` AS `int_col`,
    `data` AS `data`,
    `params` AS `params`,
    `PARTITION` AS `PARTITION`,
    `OFFSET` AS `OFFSET`,
    `TIMESTAMP` AS `TIMESTAMP`,
    `params_test0` AS `params_test0`,
    `params_test1` AS `params_test1`,
    `params_test2` AS `params_test2`,
    `data_0_data_0_char1` AS `data_0_data_0_char1`,
    `data_0_data_0_char2` AS `data_0_data_0_char2`,
    `data_0_data_0_char3` AS `data_0_data_0_char3`,
    `data_0_data_0_char4` AS `data_0_data_0_char4`,
    `year` AS `year`,
    `month` AS `month`,
    `day` AS `day`,
    `hour` AS `hour`,
    `minute` AS `minute`
    FROM `tansform_6876c922-6057-47a3-8dcb-cb0af3d6017d`
2023-12-07 22:01:49,835 INFO  org.apache.inlong.sort.base.util.FsAuthOptions               [] - Path schema is obs.
2023-12-07 22:01:51,657 WARN  org.apache.flink.client.deployment.application.ApplicationDispatcherBootstrap [] - Application failed unexpectedly:
java.util.concurrent.CompletionException: org.apache.flink.client.deployment.application.ApplicationExecutionException: Could not execute application.
    at java.util.concurrent.CompletableFuture.encodeThrowable(CompletableFuture.java:292) ~[?:1.8.0_362]
    at java.util.concurrent.CompletableFuture.completeThrowable(CompletableFuture.java:308) ~[?:1.8.0_362]
    at java.util.concurrent.CompletableFuture.uniCompose(CompletableFuture.java:957) ~[?:1.8.0_362]
    at java.util.concurrent.CompletableFuture$UniCompose.tryFire(CompletableFuture.java:940) ~[?:1.8.0_362]
    at java.util.concurrent.CompletableFuture.postComplete(CompletableFuture.java:488) ~[?:1.8.0_362]
    at java.util.concurrent.CompletableFuture.completeExceptionally(CompletableFuture.java:1990) ~[?:1.8.0_362]
    at org.apache.flink.client.deployment.application.ApplicationDispatcherBootstrap.runApplicationEntryPoint(ApplicationDispatcherBootstrap.java:370) ~[flink-dist-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.flink.client.deployment.application.ApplicationDispatcherBootstrap.lambda$runApplicationAsync$2(ApplicationDispatcherBootstrap.java:263) ~[flink-dist-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511) ~[?:1.8.0_362]
    at java.util.concurrent.FutureTask.run(FutureTask.java:266) ~[?:1.8.0_362]
    at org.apache.flink.runtime.concurrent.akka.ActorSystemScheduledExecutorAdapter$ScheduledFutureTask.run(ActorSystemScheduledExecutorAdapter.java:171) ~[flink-rpc-akka_191c14cc-48aa-4ad0-b493-ea1d1cd3a1db.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.flink.runtime.concurrent.akka.ClassLoadingUtils.runWithContextClassLoader(ClassLoadingUtils.java:68) ~[flink-rpc-akka_191c14cc-48aa-4ad0-b493-ea1d1cd3a1db.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.flink.runtime.concurrent.akka.ClassLoadingUtils.lambda$withContextClassLoader$0(ClassLoadingUtils.java:41) ~[flink-rpc-akka_191c14cc-48aa-4ad0-b493-ea1d1cd3a1db.jar:1.15.0-h0.cbu.dli.321.r2]
    at akka.dispatch.TaskInvocation.run(AbstractDispatcher.scala:49) [flink-rpc-akka_191c14cc-48aa-4ad0-b493-ea1d1cd3a1db.jar:1.15.0-h0.cbu.dli.321.r2]
    at akka.dispatch.ForkJoinExecutorConfigurator$AkkaForkJoinTask.exec(ForkJoinExecutorConfigurator.scala:48) [flink-rpc-akka_191c14cc-48aa-4ad0-b493-ea1d1cd3a1db.jar:1.15.0-h0.cbu.dli.321.r2]
    at java.util.concurrent.ForkJoinTask.doExec(ForkJoinTask.java:289) [?:1.8.0_362]
    at java.util.concurrent.ForkJoinPool$WorkQueue.runTask(ForkJoinPool.java:1056) [?:1.8.0_362]
    at java.util.concurrent.ForkJoinPool.runWorker(ForkJoinPool.java:1692) [?:1.8.0_362]
    at java.util.concurrent.ForkJoinWorkerThread.run(ForkJoinWorkerThread.java:175) [?:1.8.0_362]
Caused by: org.apache.flink.client.deployment.application.ApplicationExecutionException: Could not execute application.
    ... 13 more
Caused by: org.apache.flink.client.program.ProgramInvocationException: The main method caused an error: Array element access needs an index starting at 1 but was 0.
    at org.apache.flink.client.program.PackagedProgram.callMainMethod(PackagedProgram.java:398) ~[flink-dist-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.flink.client.program.PackagedProgram.invokeInteractiveModeForExecution(PackagedProgram.java:232) ~[flink-dist-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.flink.client.ClientUtils.executeProgram(ClientUtils.java:114) ~[flink-dist-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.flink.client.deployment.application.ApplicationDispatcherBootstrap.runApplicationEntryPoint(ApplicationDispatcherBootstrap.java:335) ~[flink-dist-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    ... 12 more
Caused by: org.apache.flink.table.api.ValidationException: Array element access needs an index starting at 1 but was 0.
    at org.apache.flink.table.planner.codegen.calls.ScalarOperatorGens$.generateArrayElementAt(ScalarOperatorGens.scala:313) ~[flink-table-planner_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.flink.table.planner.codegen.ExprCodeGenerator.generateCallExpression(ExprCodeGenerator.scala:737) ~[flink-table-planner_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.flink.table.planner.codegen.ExprCodeGenerator.generateExpr$1(ExprCodeGenerator.scala:497) ~[flink-table-planner_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.flink.table.planner.codegen.ExprCodeGenerator.visitCall(ExprCodeGenerator.scala:509) ~[flink-table-planner_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.flink.table.planner.codegen.ExprCodeGenerator.visitCall(ExprCodeGenerator.scala:59) ~[flink-table-planner_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.calcite.rex.RexCall.accept(RexCall.java:174) ~[flink-table-planner_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.flink.table.planner.codegen.ExprCodeGenerator.visitFieldAccess(ExprCodeGenerator.scala:385) ~[flink-table-planner_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.flink.table.planner.codegen.ExprCodeGenerator.visitFieldAccess(ExprCodeGenerator.scala:59) ~[flink-table-planner_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.calcite.rex.RexFieldAccess.accept(RexFieldAccess.java:92) ~[flink-table-planner_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.flink.table.planner.codegen.ExprCodeGenerator.$anonfun$visitCall$1(ExprCodeGenerator.scala:494) ~[flink-table-planner_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at scala.collection.TraversableLike.$anonfun$map$1(TraversableLike.scala:233) ~[flink-scala_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at scala.collection.mutable.ResizableArray.foreach(ResizableArray.scala:58) ~[flink-scala_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at scala.collection.mutable.ResizableArray.foreach$(ResizableArray.scala:51) ~[flink-scala_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at scala.collection.mutable.ArrayBuffer.foreach(ArrayBuffer.scala:47) ~[flink-scala_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at scala.collection.TraversableLike.map(TraversableLike.scala:233) ~[flink-scala_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at scala.collection.TraversableLike.map$(TraversableLike.scala:226) ~[flink-scala_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at scala.collection.AbstractTraversable.map(Traversable.scala:104) ~[flink-scala_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.flink.table.planner.codegen.ExprCodeGenerator.generateExpr$1(ExprCodeGenerator.scala:485) ~[flink-table-planner_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.flink.table.planner.codegen.ExprCodeGenerator.visitCall(ExprCodeGenerator.scala:509) ~[flink-table-planner_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.flink.table.planner.codegen.ExprCodeGenerator.visitCall(ExprCodeGenerator.scala:59) ~[flink-table-planner_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.calcite.rex.RexCall.accept(RexCall.java:174) ~[flink-table-planner_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.flink.table.planner.codegen.ExprCodeGenerator.visitFieldAccess(ExprCodeGenerator.scala:385) ~[flink-table-planner_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.flink.table.planner.codegen.ExprCodeGenerator.visitFieldAccess(ExprCodeGenerator.scala:59) ~[flink-table-planner_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.calcite.rex.RexFieldAccess.accept(RexFieldAccess.java:92) ~[flink-table-planner_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
    at org.apache.flink.table.planner.codegen.ExprCodeGenerator.generateExpression(ExprCodeGenerator.scala:166) ~[flink-table-planner_2.12-1.15.0-h0.cbu.dli.321.r2.jar:1.15.0-h0.cbu.dli.321.r2]
```

Possible Cause

The error message indicates that the array subscript must start from 1, and a[1] indicates the first element in the array.

Check the CDM real-time job. The migration source is Kafka, destination is the mapped DLI field, value source is the source table field, and value is **a[0]** (the nested JSON array in the Kafka message). When the source table field is used to assign a value to a destination column during field mapping, the array subscript must start from 1 (indicating the first element of the array). If subscript 0 is used by mistake, the job will fail. The error was caused by the incorrect subscript of the assignment array. You need to reset it.

Figure 4-10 Original parameter settings



Mapping Between Source and Destination Tables

Mapping result: Successful: 0 tables Failed: 1 tables Total: 1 tables Mapping progress: 100%

Source table: Enter a table name.

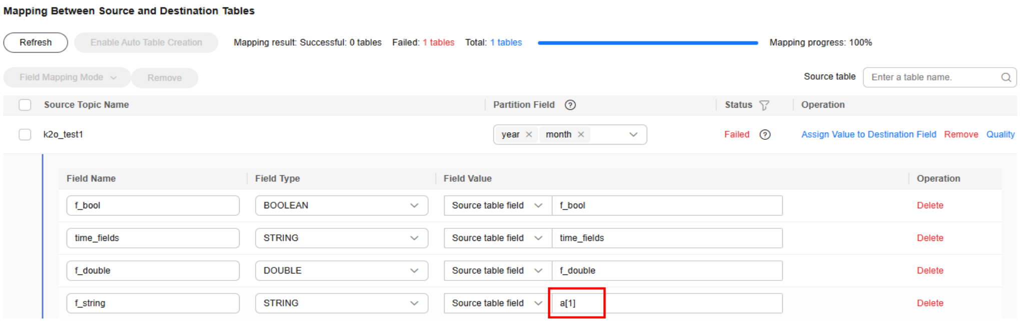
	Source Topic Name	Partition Field	Status	Operation
□	k2o_test1	year × month ×	Failed	Assign Value to Destination Field Remove Quality

Field Name	Field Type	Field Value		Operation
f_bool	BOOLEAN	Source table field	f_bool	Delete
time_fields	STRING	Source table field	time_fields	Delete
f_double	DOUBLE	Source table field	f_double	Delete
f_string	STRING	Source table field	f_string	Delete
f_decimal	DOUBLE	Source table field	f_decimal	Delete

Solution

Change the value of the field to **a[1]**, submit the job version, and restart the job.

Figure 4-11 Setting the value assignment parameters of the destination table



Mapping Between Source and Destination Tables

Mapping result: Successful: 0 tables Failed: 1 tables Total: 1 tables Mapping progress: 100%

Source table: Enter a table name.

Source Topic Name	Partition Field	Status	Operation
k2o_test1	year, month	Failed	Assign Value to Destination Field Remove Quality

Field Name	Field Type	Field Value		Operation
f_bool	BOOLEAN	Source table field	f_bool	Delete
time_fields	STRING	Source table field	time_fields	Delete
f_double	DOUBLE	Source table field	f_double	Delete
f_string	STRING	Source table field	a[1]	Delete

4.8 How Do I Grant the Log Archiving, Query, and Parsing Permissions of an Oracle Data Source?

Symptom

The default permissions of the Oracle data source are insufficient for executing real-time processing migration jobs.

Possible Cause

- The log archiving function must be enabled for the Oracle database. It is recommended that archived logs be retained for at least three days.
- The permissions for querying Oracle tables or parsing logs are missing.

Solution

1. Enable log archiving.
   a. Log in to the Oracle database as user **sysdba**.
   b. Run the SQL command **ARCHIVE LOG LIST** to query the archiving status of the current database. The following information indicates that log archiving is disabled:
   ```
   Database log mode No Archive Mode #Log archiving is disabled.
   Automatic archival Disabled
   Archive destination USE_DB_RECOVERY_FILE_DEST
   Oldest online log sequence 1
   Current log sequence 2
   ```
   c. Run the SQL command **SHUTDOWN IMMEDIATE** to shut down the database.
   d. Run the SQL command **STARTUP MOUNT** to start the database in MOUNT state.
   e. Run the SQL command **ALTER DATABASE ARCHIVELOG** to enable archiving mode.
   f. Run the SQL command **ARCHIVE LOG LIST** to query the archiving status. The following information indicates that log archiving is enabled:
   ```
   Databaselogmode Archive Mode #Log archiving is enabled.
   Automatic archival Enabled
   Archive destination USE_DB_RECOVERY_FILE_DEST
   Oldest online log sequence 1
   Next log sequence to archive 2
   Currentlogsequence 2
   ```

g. Run the SQL command **ALTER DATABASE OPEN** to start the database.

2. Enable supplementary logs for the Oracle database and tables to be migrated.

   a. Run the following SQL statement to enable supplemental logs for the database:

   ```
   ALTER DATABASE ADD SUPPLEMENTAL LOG DATA;
   ```

   b. Enable supplemental logs for the tables to be synchronized in real time.

   ```
   ALTER TABLE "schema_name"."table_name" ADD SUPPLEMENTAL LOG DATA (ALL) COLUMNS;
   ```

   After the setting is successful, run the following SQL statement. If **ALL_COLUMN_LOGGING** is returned, supplemental logs are enabled for the tables:

   ```
   SELECT 'KEY', LOG_GROUP_TYPE FROM ALL_LOG_GROUPS WHERE OWNER = 'schema_name'
   AND TABLE_NAME = 'table_name';
   "KEY" LOG_GROUP_TYPE
   KEY ALL_COLUMN_LOGGING
   ```

3. Grant required permissions to the Oracle user.

   – Reference commands for granting permissions to Oracle 19 users:

   ```
   sqlplus sys/password@//localhost:1521/ORCLCDB as sysdba
   CREATE USER mgrationuser IDENTIFIED BY mgrationuserPWD DEFAULT TABLESPACE
   logminer_tbs QUOTA UNLIMITED ON logminer_tbs CONTAINER=ALL;
   GRANT CREATE SESSION TO mgrationuser CONTAINER=ALL;
   GRANT SET CONTAINER TO mgrationuser CONTAINER=ALL;
   GRANT SELECT ON V_$DATABASE to mgrationuser CONTAINER=ALL;
   GRANT FLASHBACK ANY TABLE TO mrationuser CONTAINER=ALL;
   GRANT SELECT ANY TABLE TO mgrationuser CONTAINER=ALL;
   GRANT SELECT_CATALOG_ROLE TO mgrationuser CONTAINER=ALL;
   GRANT EXECUTE_CATALOG_ROLE TO mgrationuser CONTAINER=ALL;
   GRANT SELECT ANY TRANSACTION TO mgrationuser CONTAINER=ALL;
   GRANT LOGMINING TO mgrationuser CONTAINER=ALL;
   GRANT CREATE TABLE TO mgrationuser CONTAINER=ALL;
   -- Don't need to execute this statement, If you set 'scan.incremental.snapshot.enabled=true'
   (default).
   GRANT LOCK ANY TABLE TO mgrationuser CONTAINER=ALL;
   GRANT CREATE SEQUENCE TO mgrationuser CONTAINER=ALL;
   GRANT EXECUTE ON DBMS_LOGMNR TO mgrationuser CONTAINER=ALL;
   GRANT EXECUTE ON DBMS_LOGMNR_D TO mgrationuser CONTAINER=ALL;
   GRANT SELECT ON V_$LOG TO mgrationuser CONTAINER=ALL;
   GRANT SELECT ON V_$LOG_HISTORY TO mgrationuser CONTAINER=ALL;
   GRANT SELECT ON V_$LOGMNR_LOGS TO mgrationuser CONTAINER=ALL;
   GRANT SELECT ON V_$LOGMNR_CONTENTS TO mgrationuser CONTAINER=ALL;
   GRANT SELECT ON V_$LOGMNR_PARAMETERS TO mgrationuser CONTAINER=ALL;
   GRANT SELECT ON V_$LOGFILE TO mgrationuser CONTAINER=ALL;
   GRANT SELECT ON V_$ARCHIVED_LOG TO mgrationuser CONTAINER=ALL;
   GRANT SELECT ON V_$ARCHIVE_DEST_STATUS TO mgrationuser CONTAINER=ALL;
   exit;
   ```

   – Reference commands for granting permissions to Oracle 11g users:

   ```
   sqlplus sys/password@host:port/SID AS SYSDBA;
     CREATE USER mgrationuser IDENTIFIED BY mgrationuserPDW DEFAULT TABLESPACE
   LOGMINER_TBS QUOTA UNLIMITED ON LOGMINER_TBS;
     GRANT CREATE SESSION TO mgrationuser;
     GRANT SELECT ON V_$DATABASE to mgrationuser;
     GRANT FLASHBACK ANY TABLE TO mgrationuser;
     GRANT SELECT ANY TABLE TO mgrationuser;
     GRANT SELECT_CATALOG_ROLE TO mgrationuser;
     GRANT EXECUTE_CATALOG_ROLE TO mgrationuser;
     GRANT SELECT ANY TRANSACTION TO mgrationuser;
     GRANT CREATE TABLE TO mgrationuser;
     GRANT LOCK ANY TABLE TO mgrationuser;
     GRANT ALTER ANY TABLE TO mgrationuser;
     GRANT CREATE SEQUENCE TO mgrationuser;
     GRANT EXECUTE ON DBMS_LOGMNR TO mgrationuser;
     GRANT EXECUTE ON DBMS_LOGMNR_D TO mgrationuser;
     GRANT SELECT ON V_$LOG TO mgrationuser;
     GRANT SELECT ON V_$LOG_HISTORY TO mgrationuser;
   ```

```
GRANT SELECT ON V_$LOGMNR_LOGS TO mgrationuser;
GRANT SELECT ON V_$LOGMNR_CONTENTS TO mgrationuser;
GRANT SELECT ON V_$LOGMNR_PARAMETERS TO mgrationuser;
GRANT SELECT ON V_$LOGFILE TO mgrationuser;
GRANT SELECT ON V_$ARCHIVED_LOG TO mgrationuser;
GRANT SELECT ON V_$ARCHIVE_DEST_STATUS TO mgrationuser;
exit
```

4.9 How Do I Manually Delete Replication Slots from a PostgreSQL Data Source?

Symptom

The replication slots of the PostgreSQL data source cannot be automatically deleted. When the number of replication slots reaches the upper limit, new jobs cannot be executed. You must manually delete replication slots.

Possible Cause

The replication slots of the PostgreSQL data source cannot be automatically deleted.

Solution

1. Log in to the source database used by the job.
2. Query the name of the streaming replication slot corresponding to the database object selected for the synchronization task.
   ```
   select slot_name from pg_replication_slots where database = 'database';
   ```
3. Run the following statement to delete the corresponding streaming replication slot:
   ```
   select * from pg_drop_replication_slot('slot_name');
   ```
4. Run the following statement to check whether the streaming replication slot is deleted:
   ```
   select slot_name from pg_replication_slots where slot_name = 'slot_name';
   ```

5 DataArts Architecture

5.1 What Is the Relationship Between Lookup Tables and Data Standards?

Symptom

Relationship between lookup tables and data standards

Solution

A lookup table consists of the names, codes, and data types of multiple table fields. The table fields in a code table can be associated with a data standard, and the data standard is applied to the fields in a model table.

5.2 What Are the Differences Between ER Modeling and Dimensional Modeling?

Symptom

Differences between ER modeling and dimensional modeling

Solution

- ER modeling is transactional and complies with 3NF modeling.
- Dimensional modeling mainly refers to the design of fact tables and dimension tables. Dimensional modeling is mainly used to implement multi-angle and multi-layer data query and analysis.

DataArts Studio is a data lake operations platform. Dimensional modeling is used more frequent.

5.3 What Data Modeling Methods Are Supported by DataArts Architecture?

Symptom

Data modeling methods supported by DataArts Architecture

Solution

DataArts Studio DataArts Architecture supports the following three types of modeling methods:

- **ER modeling**

  ER modeling describes the business activities within an enterprise. Compliant with the third normal form (3NF), ER modeling is designed for data integration. It is used for combining and merging data with similarities by subject. ER modeling results cannot be used directly for decision-making, but they are a useful tool.

  You can divide ER modeling into three levels of abstraction: design conceptual models, logical models, and physical models.

  - **Physical model**: A physical model is based on logical models and is used to design the database architecture for data storage with a range of technical factors all considered. For example, the selected data warehouse could be defined as DWS or DLI.

- **Dimensional modeling**

  Dimensional modeling is the construction of models based on analysis and decision-making requirements. It is mainly used for data analysis. Dimensional modeling is focused on how to quickly analyze user requirements and respond rapidly to complicated large-scale queries.

  A multidimensional model is a fact table that consists of numeric measurement metrics. The fact table is associated with a group of dimensional tables that contain description attributes through primary or foreign keys.

  Typical dimensional models include star models and snowflake models used in some special scenarios.

  In the DataArts Architecture module of DataArts Studio, dimensional modeling involves constructing bus matrices to extract business facts and dimensions for model creation. You need to sort out business requirements for constructing metric systems and creating summary models.

- **Data mart**

  A data mart (DM) aggregates data from multiple layers and consists of a specific analysis object and its related metrics. The DM provides all statistical data by subject.

5.4 How Can I Use Standardized Data?

Symptom

Application scenarios of standardized data

Solution

Standardized data can be used as basic BI information, source data of upper-layer applications, and visualized reports of various data.

5.5 Does DataArts Architecture Support Database Reversing?

Symptom

Whether DataArts Architecture supports database reversing

Solution

DataArts Architecture supports database reversing.

5.6 What Are the Differences Between the Metrics in DataArts Architecture and DataArts Quality?

Symptom

Differences between the metrics in DataArts Architecture and DataArts Quality

Solution

The metrics in DataArts Architecture focus on business and are used to measure the overall characteristics of objects. The metrics in DataArts Quality focus on monitoring and are used to manage all business metrics, including their sources and definitions.

Metrics in DataArts Quality are independent of business metrics and technical metrics in DataArts Architecture. Metrics in DataArts Quality will be unavailable soon. You are advised to use the metrics in DataArts Architecture.

5.7 Why Doesn't the Table in the Database Change After I Have Modified Fields in an ER or Dimensional Model?

Possible Causes

The table in the database does not change after the fields in an ER or dimensional model are modified.

Solution

This is because the **Data Table Update Mode** parameter is not configured. Its default value is **No update**.

To configure the table update mode, perform the following steps:

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.
2. On the DataArts Architecture page, choose **Configuration Center** in the left navigation pane.
3. Click the **Functions** tab.
4. Set **Data Table Update Mode** to **DDL-based update** or **Drop and create**.
   - **No update**: The system does not update tables in a database.
   - **DDL-based update**: The system updates tables in the database based on the DDL update template configured in DDL template management. The underlying data warehouse engine determines whether the update is successful. Different types of data warehouses support different table update modes. If the data warehouse does not support table update operations on the DataArts Architecture page, the tables in the database may be inconsistent with those in DataArts Architecture. For example, table fields cannot be deleted when DLI tables are updated. If table fields are deleted from the tables in DataArts Architecture, the corresponding table fields cannot be deleted from the database.

     If the offline database supports the syntax for updating the table architecture, you can configure the syntax in the DDL template. Then, the update operation can be managed on DataArts Studio. Otherwise, update the table by rebuilding it.
   - **Drop and create**: The system deletes an existing table in a database and then creates a table. This option ensures that the tables in the database are the same as those in DataArts Architecture. However, since the table is deleted first, you are advised to select this option only in the development and design phase or test phase. After the product is brought online, you are not advised to select this option.
5. Click **OK**.

5.8 Can I Configure Lifecycle Management for Tables?

Symptom

Whether lifecycle management can be configured for tables

Solution

No. This function is unavailable now.

5.9 How Should I Select a Subject When a Public Dimension (Date, Region, Supplier, or Product) Is Shared by Multiple Subject Areas?

DataArts Architecture does not provide public dimensions. Each dimension must belong to a subject. If a public dimension is used, you are advised to create a public subject for the public dimension.

In addition, permissions of dimensions are classified by model instead of by subject. Therefore, subjects do not affect permission control or query.

6 DataArts Factory

6.1 How Many Jobs Can Be Created in DataArts Factory? Is There a Limit on the Number of Nodes in a Job?

Symptom

Whether the number of jobs and the number of nodes in a job are limited during data development

Solution

By default, each user can create a maximum of 10,000 jobs, and each job can contain a maximum of 200 nodes.

In addition, the system allows you to adjust the maximum quota as required. If you want to do so, submit a service ticket.

6.2 Does DataArts Studio Support Custom Python Scripts?

Symptom

Whether DataArts Studio supports custom Python scripts

Solution

Yes.

6.3 How Can I Quickly Rectify a Deleted CDM Cluster Associated with a Job?

Possible Causes

The CDM cluster associated with the job has been deleted.

Solution

After the CDM cluster is deleted, the association information in the data development job remains intact. You only need to create a cluster and job with the same names on CDM. The data development job will remind you that the original CDM cluster and job will be replaced before using the newly created ones.

6.4 Why Is There a Large Difference Between Job Execution Time and Start Time of a Job?

Symptom

On the **Running History** page, there is a large difference between **Job Execution Time** and **Start Time**. **Job Execution Time** is the time when the job is expected to be executed. **Start Time** is the time when the job starts to be executed.

Possible Causes

In Data Development, a maximum of five instances can be concurrently executed in a job. If **Start Time** of a job is later than **Job Execution Time**, the job instances in the subsequent batch will be queued.

Solution

If you find that the difference between **Job Execution Time** and **Start Time** becomes large, adjust **Job Execution Time** accordingly.

6.5 Will Subsequent Jobs Be Affected If a Job Fails to Be Executed During Scheduling of Dependent Jobs? What Should I Do?

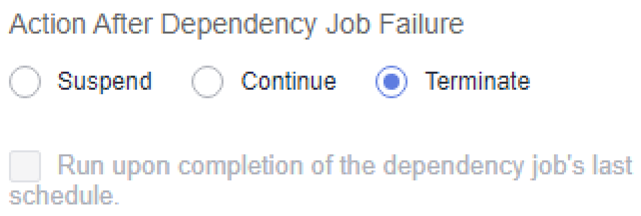
Possible Causes

One of the jobs that depend on each other fails during scheduling.

Solution

The subsequent jobs may be suspended, continued, or canceled, depending on the configuration.

Figure 6-1 Job dependencies



Action After Dependency Job Failure

○ Suspend ○ Continue ● Terminate

□ Run upon completion of the dependency job's last schedule.

In this case, do not stop the job. You can rerun the failed job instance or stop the abnormal instance and then run it again. After the instance failure is removed, the subsequent operations will continue. If you manually process the failure not in DataArts Factory but in other ways, you can force the job instance to succeed after the failure is removed and then subsequent jobs will continue to run properly.

6.6 What Should I Pay Attention to When Using DataArts Studio to Schedule Big Data Services?

Symptom

Notes for scheduling big data services using DataArts Studio

Solution

Lock management is unavailable for DLI and MRS. Therefore, if you perform read and write operations on the tables simultaneously, data conflict will occur and the operations will fail.

If you want to perform read and write operations on the data tables of big data services, use either of the following methods to perform serial operations:

- Create a job with two nodes, one for the read operation and the other for the write operation, and execute the nodes in sequence to avoid conflicts.
- Create a job for the read operation and another job for the write operation, and configure a dependency relationship between the two jobs to avoid conflicts.

6.7 What Are the Differences and Relationships Between Environment Variables, Job Parameters, and Script Parameters?

Symptom

Differences and relationships between environment variables, job parameters, and script parameters

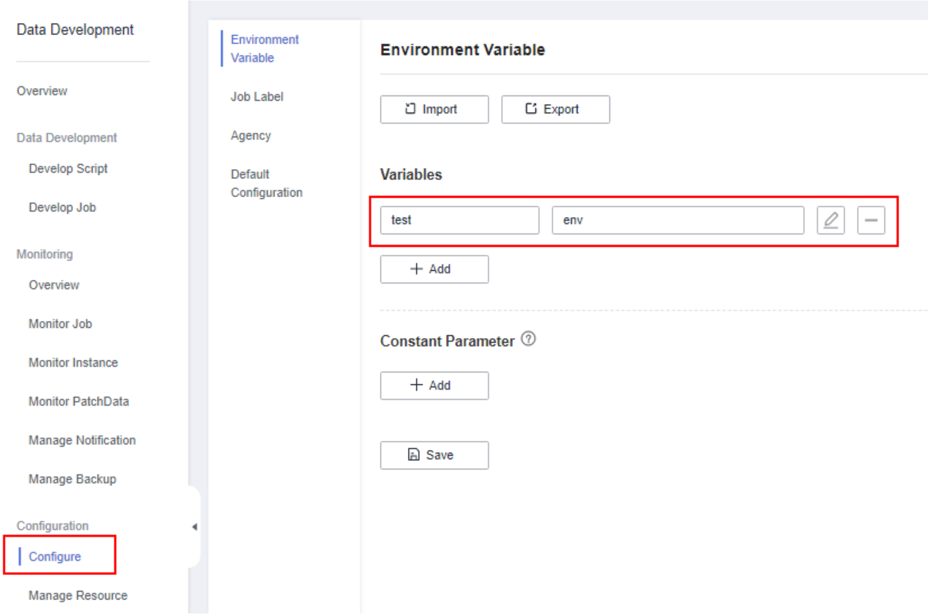
Solution

The application scopes of workspace environment variables, job parameters, and script parameters are different. If a workspace environment variable, a job parameter, and a script parameter have the same name, their priorities are as follows: **job parameter > workspace environment variable > script parameter**.

Introduction and usage of environment variables, job parameters, and script parameters are as follows:

- Variables and constants can be defined in environment variables. Environment variables take effect in current workspace.
  - The value of a variable (such as **workspace name**) varies depending on the workspace. When exporting a variable from a workspace and import it to another workspace, you must reconfigure its value.
  - The value of a constant in different workspaces is the same. When importing a constant to another workspace, you do not need to reconfigure its value.

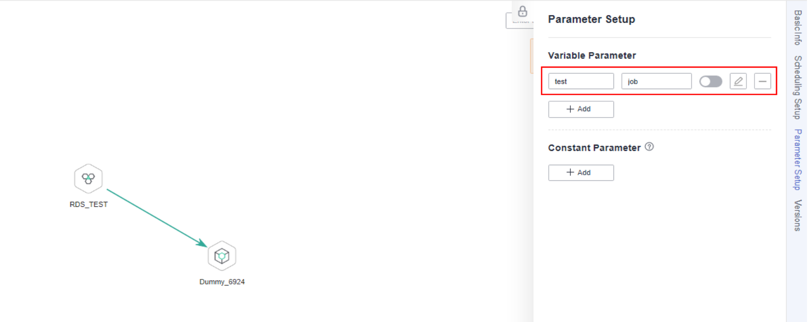
Figure 6-2 Environment variable



test | env

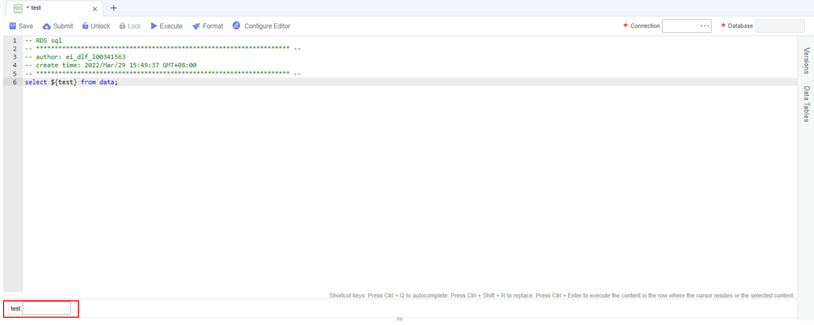
- Parameters and constants can be defined in job parameters. Job parameters take effect in current job.
  - The value of a parameter varies depending on jobs. When exporting a parameter from a workspace and import it to another workspace, you must reconfigure its value.
  - The value of a constant in different jobs is the same. When importing a constant to another job, you do not need to reconfigure its value.

Figure 6-3 Job parameter.



- Script parameters take effect in current script and it can be used in the following ways.
  - Enter SQL script parameters in the script editor (Flink SQL is not supported). If the script is executed independently, you can configure the parameters in the lower part of the editor, as shown in **Figure 6-4**. If the script is executed by job scheduling, you can assign values to the parameters based on node attributes, as shown in **Figure 6-5**.
  - For Shell scripts, you can enter a parameter and an interactive parameter to transfer the parameters.
  - For Python scripts, you can enter a parameter and an interactive parameter to transfer the parameters.

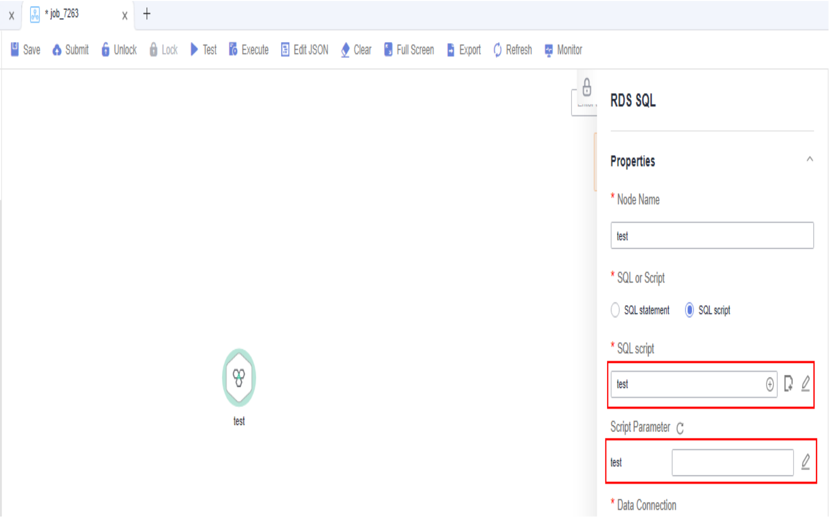
Figure 6-4 Configuring script parameters when the script is executed independently



```
-- RDS sql
-- ********************************************************************* --
-- author: ei_dlf_l00341563
-- create time: 2022/Mar/29 15:49:37 GMT+08:00
-- ********************************************************************* --
select ${test} from data;
```

Shortcut keys: Press Ctrl + Q to autocomplete. Press Ctrl + Shift + R to replace. Press Ctrl + Enter to execute the content in the row where the cursor resides or the selected content.

Figure 6-5 Configuring script parameters when the script is executed by job scheduling



RDS SQL

Properties

* Node Name: test

* SQL or Script: ○ SQL statement ● SQL script

* SQL script: test

Script Parameter: test

* Data Connection

6.8 What Should I Do If a Job Log Cannot Be Opened and Error 404 Is Reported?

Possible Causes

You do not have the required permissions.

Solution

Job logs are stored in OBS buckets. You must configure the bucket directory for storing job logs in the workspace and check whether your account has the OBS read permission by checking the OBS permission and OBS bucket policy in IAM.

NOTE

The OBS path is only supported for OBS buckets and not for parallel file systems.

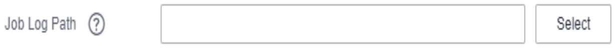
You are using either of the following accounts:

- **DAYU Administrator** or **Tenant Administrator**
- **DAYU User**, which is the administrator of the current workspace

To configure the bucket directory for storing job logs, perform the following steps:

1. Log in to the DataArts Studio console.
2. On the **Workspaces** page, locate a workspace and click **Edit** in the **Operation** column.
3. In the displayed **Workspace Information** dialog box, click the **Select** button next to **Job Log Path** and select a path.

Figure 6-6 Changing the job log path



4. Click **OK**.

NOTE

When you create a job, a bucket named **dlf-log-{projectID}** will be created by default. If the bucket exists, you do not need to create a bucket again.

6.9 What Should I Do If the Agency List Fails to Be Obtained During Agency Configuration?

Possible Causes

If error message "Policy doesn't allow iam:agencies:listAgencies to be performed." is displayed when you are creating a workspace-level or job-level agency, you may lack the required permissions.

Solution

Add the **View Agency List** policy for the current user.

You can create a custom policy (query the agency list based on specified conditions) and assign it to a user group for refined access control.

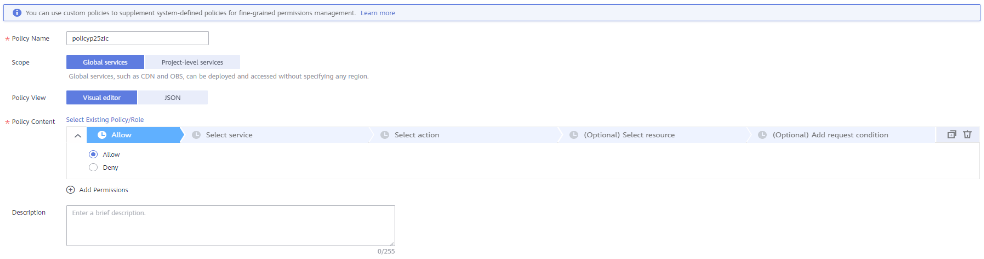
Step 1 Log in to the Huawei Cloud management console.

Step 2 On the management console, hover the mouse pointer over the username in the upper right corner, and choose **Identity and Access Management** from the drop-down list.

Step 3 In the navigation pane, choose **Permissions** > **Roles**. Then, click **Create Custom Policy**.

Step 4 Enter a policy name.

Figure 6-7 Policy name



You can use custom policies to supplement system-defined policies for fine-grained permissions management. Learn more

* Policy Name: policyp25zic

Scope: Global services | Project-level services

Global services, such as CDN and OBS, can be deployed and accessed without specifying any region.

Policy View: Visual editor | JSON

* Policy Content: Select Existing Policy/Role

Allow > Select service > Select action > (Optional) Select resource > (Optional) Add request condition

● Allow
○ Deny

Add Permissions

Description: Enter a brief description.

0/255

Step 5 Set **Scope** based on the region where the service is deployed. In this example, you need to grant IAM the permission to query the agency list based on specified conditions. As IAM is a global service, select **Global services** for **Scope**.

Step 6 Select **Visual editor** for **Policy View**.

Step 7 Configure a policy in **Policy Content**.

1. Select **Allow**.
2. Select **Identity and Access Management (IAM)** for **Select service**.
3. Select **iam:agencies:listAgencies** for **Select action**.

Step 8 Click **OK**.

Step 9 Add the policy defined in **Step 7** to the group to which the current user belongs. For details, see **Creating a User Group and Granting Permissions**.

Step 10 In the navigation pane on the left, choose **Agencies**. Locate the target agency, click **Authorize** in the **Operation** column, add the created custom policy to the agency, and click **OK**.

The current user can log out of the system and then log in again to obtain the agency list.

----End

6.10 Why Can't I Select Specified Peripheral Resources When Creating a Data Connection in DataArts Factory?

Possible Causes

The specified peripheral resources may not be in the same region as the instance.

Solution

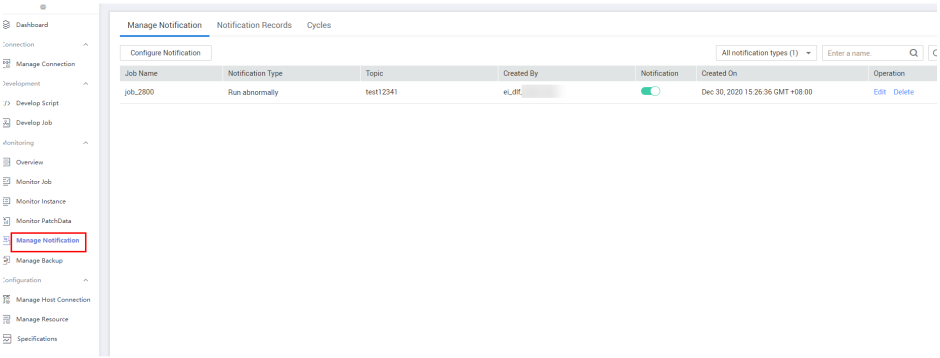
Ensure that the current DataArts Studio instance and peripheral resources are in the same region and IAM project. If the enterprise project function is enabled for your account, the current instance and peripheral resources must be in the same enterprise project.

6.11 Why Can't I Receive Job Failure Alarm Notifications After I Have Configured SMN Notifications?

Symptom

No job failure alarm notifications are received after SMN notifications for job exceptions or failures have been configured in **Monitoring** > **Manage Notification**.

Figure 6-8 Manage Notification



Job Name	Notification Type	Topic	Created By	Notification	Created On	Operation
job_2800	Run abnormally	test12341	ei_dlf_[redacted]	●	Dec 30, 2020 15:26:36 GMT +08:00	Edit Delete

Solution

To solve the problem, perform the following steps:

Step 1 Check whether the failed job is being scheduled. No notification is sent for jobs in the test running state. SMN notifications are sent only for jobs in the scheduling state.

Step 2 On the **Data Development** page, choose **Monitoring** > **Manage Notification** to check whether the notification function is enabled.

Step 3 Log in to the SMN console and check whether the SMN topic has been subscribed to.

Step 4 Check whether the subscription endpoint of the SMN topic has its own name and whether the subscription is confirmed.

Step 5 Check whether the SMN channel is normal. You can send messages to your topic on the SMN console to check whether you can receive notifications from SMN.

----End

6.12 Why Is There No Job Running Scheduling Log on the Monitor Instance Page After Periodic Scheduling Is Configured for a Job?

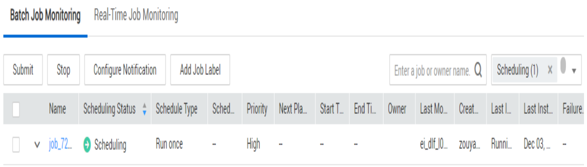
Possible Causes

Why Is There No Job Running Scheduling Log on the Monitor Instance Page After Periodic Scheduling Is Configured for a Job? This job has not been started or the job on which this depends has not been executed.

Solution

1. On the DataArts Factory page, choose **Monitoring** > **Job Monitoring** to check whether the target job is being scheduled. A job can be scheduled only within the scheduling period.

Figure 6-9 Viewing the job scheduling status



Batch Job Monitoring | Real-Time Job Monitoring

Submit | Stop | Configure Notification | Add Job Label | Enter a job or owner name. | Scheduling (1) ×

□	Name	Scheduling Status	Schedule Type	Sched...	Priority	Next Pla...	Start T...	End Ti...	Owner	Last Mo...	Creat...	Last I...	Last Inst...	Failure...
□	job_72...	Scheduling	Run once	--	High	--	--	--		ei_dlf_l0...	zouya...	Runni...	Dec 03, ...	--

2. If a job depends on other jobs, choose **Monitoring** > **Monitor Instance** to view the running status of the dependent jobs. If the job is self-dependent, expand the search time to check whether the job is waiting for running due to the failure of a historical job instance.

6.13 Why Isn't the Error Cause Displayed on the Console When a Hive SQL or Spark SQL Scripts Fails?

Possible Causes

This issue may be caused by the connection mode.

Solution

Check whether the data connection used by the Hive SQL and Spark SQL scripts is an MRS API connection or proxy connection.

In MRS API connection mode, DataArts Studio users submit scripts to MRS through APIs and then check whether the scripts are executed successfully. MRS does not send the specific error cause to DataArts Studio. Therefore, the GUI displays only the execution result (success or failure) but does not display the error cause.

In proxy connection mode, DataArts Studio users submit and run scripts, and check whether the scripts are executed successfully. In addition, the error information and script execution results are displayed in the logs on the DataArts Factory script execution page.

If you want to view the error cause, go to the job management page on the MRS console.

6.14 What Should I Do If the Token Is Invalid During the Execution of a Data Development Node?

Symptom

During the execution of a data development node, a message is displayed indicating that the token is invalid.

Solution

Check whether the permissions of the current user in IAM are changed, whether the user is removed from the user group, or whether the permission policy of the user group to which the user belongs is changed.

If they are indeed changed, log in to the system again.

6.15 How Do I View Run Logs After a Job Is Tested?

Symptom

How to view run logs after testing a job

Solution

Method 1: After the node test is complete, right-click the current node and choose **View Log** from the shortcut menu.

Method 2: Click **Monitor** in the upper part of the canvas, expand the job instance on the **Monitor Instance** page, and view node logs.

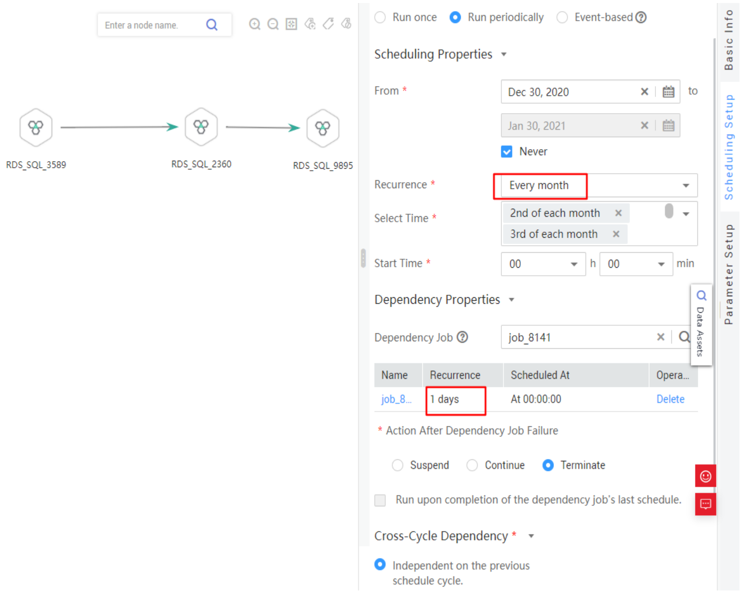
6.16 Why Does a Job Scheduled by Month Start Running Before the Job Scheduled by Day Is Complete?

Possible Causes

The monthly job depends on daily jobs of the previous month rather than the current month.

Jobs scheduled by month depend on jobs scheduled by day. Why does a job scheduled by month start running before the job scheduled by day is complete?

Figure 6-10 Viewing the job scheduling period and dependency attributes



○ Run once ● Run periodically ○ Event-based

Scheduling Properties

From * Dec 30, 2020 to Jan 30, 2021

☑ Never

Recurrence * Every month

Select Time * 2nd of each month, 3rd of each month

Start Time * 00 h 00 min

Dependency Properties

Dependency Job job_8141

Name	Recurrence	Scheduled At	Opera...
job_8...	1 days	At 00:00:00	Delete

* Action After Dependency Job Failure

○ Suspend ○ Continue ● Terminate

□ Run upon completion of the dependency job's last schedule.

Cross-Cycle Dependency *

● Independent on the previous schedule cycle.

Solution

Although jobs scheduled by month depend on jobs scheduled by day, whether jobs scheduled by month in the current month are executed depends on whether all jobs scheduled by day in the previous month are complete, not the jobs scheduled by day in the current month.

For example, whether the monthly scheduled jobs run in November depends on whether the daily scheduled jobs were complete in October.

6.17 What Should I Do If Invalid Authentication Is Reported When I Run a DLI Script?

Possible Causes

This issue may be caused by insufficient permissions.

Solution

Check whether the current user has the DLI Service User or DLI Service Admin permissions in IAM.

6.18 Why Cannot I Select a Desired CDM Cluster in Proxy Mode When Creating a Data Connection?

Possible Causes

The CDM cluster may be stopped.

Solution

Check whether the CDM cluster is stopped. If it is stopped, restart it.

6.19 Why Is There No Job Running Scheduling Record After Daily Scheduling Is Configured for the Job?

Symptom

Daily scheduling is configured for the job, but there is no job scheduling record in the instance.

Cause Analysis

Cause 1: Check whether the job scheduling is started. If not, the job will not be scheduled.

Cause 2: The instance query time range is too long. If a dependent or self-dependent job is configured, check whether the historical job instance is waiting for running due to the dependency failure. As a result, no new job instance is generated.

Solutions

Configure Job exception alarms and instance timeout duration. When the waiting time exceeds the instance timeout duration, the system sends an alarm notification.

6.20 What Do I Do If No Content Is Displayed in Job Logs?

Symptom

There is no content contained in the job log.

Cause Analysis

If the bucket directory for storing job logs has been configured in the workspace, check whether you have the global permission of OBS so that you can create and operate buckets.

Solutions

Method 1: Create a bucket named dlf-log-{projectID} in OBS and grant the operation permission to the scheduling user.

NOTE

The OBS path is only supported for OBS buckets and not for parallel file systems.

Method 2: Add global OBS administrator permission in IAM user permissions.

6.21 Why Do I Fail to Establish a Dependency Between Two Jobs?

Symptom

Two jobs are created, but the dependency relationship cannot be established.

Cause Analysis

Check whether the two jobs' recurrence are both every week or every month. Currently, if the two jobs' recurrence are both every week or every month, the dependency relationship cannot be established..

Solutions

You can place the two jobs whose recurrence are both every week or every month in the same canvas before running them.

6.22 What Should I Do If an Error Is Reported During Job Scheduling in DataArts Studio, Indicating that the Job Has Not Been Submitted?

Symptom

An error is reported during job scheduling in DataArts Studio, indicating that the job has not been submitted.

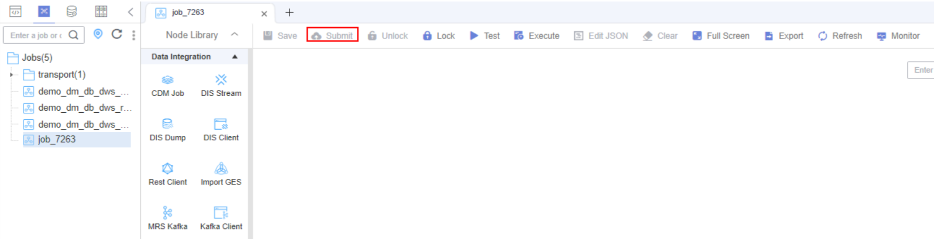
Cause Analysis

Job scheduling process begins before the version is submitted. As a result, an error is reported during scheduling. Ensure that the job has a submitted version before it is scheduled.

Solutions

1. Step 1: Submit a job version (not a script).
2. Step 2: Schedule the job.

Figure 6-11 Submitting a version



6.23 What Should I Do If an Error Is Reported During Job Scheduling in DataArts Studio, Indicating that the Script Associated with Node XXX in the Job Has Not Been Submitted?

Symptom

An error is reported when DataArts Studio executes scheduling: The script associated with node XXX in the job is not submitted.

Cause Analysis

Job scheduling process begins before the script version is submitted. As a result, an error is reported during scheduling. Ensure that the job has a submitted script version before the job is scheduled.

Solutions

1. Step 1: Switch to the script development page and find the corresponding script.
2. Step 2: Submit the script version.
3. Step 3: Schedule the job.

6.24 What Should I Do If a Job Fails to Be Executed After Being Submitted for Scheduling and an Error Displayed: Depend Job [XXX] Is Not Running Or Pause?

Symptom

After a job is submitted for scheduling, the job fails to be executed and the following error is displayed "depend job [XXX] is not running or pause".

Cause Analysis

The upstream dependency job is not in the running state.

Solutions

Check the upstream dependency jobs. If the upstream dependency jobs are not in the running state, re-schedule these jobs.

6.25 How Do I Create Databases and Data Tables? Do Databases Correspond to Data Connections?

Symptom

This section describes how to create databases and data tables, and the relationship between databases and data connections.

Solution

You can create databases and data tables in DataArts Studio.

Databases do not correspond to data connections. Data connections connect DataArts Studio and other data services.

6.26 Why Is No Result Displayed After a Hive Task Is Executed?

Possible Causes

The connection mode may be incorrect.

Solution

Clear the cache data and use the direct connection to display the data.

6.27 Why Is the Last Instance Status On the Monitor Instance Page Either Successful or Failed?

Symptom

The **Last Instance Status** on the **Monitor Instance** page is either successful or failed.

Solution

The last instance status indicates a job has been executed, and the status can only be successful or failed. The Monitor Instance page displays all statuses of the job, including canceled and suspended. In addition, job running exceptions and errors are all job failure statuses.

6.28 How Do I Configure Notifications for All Jobs?

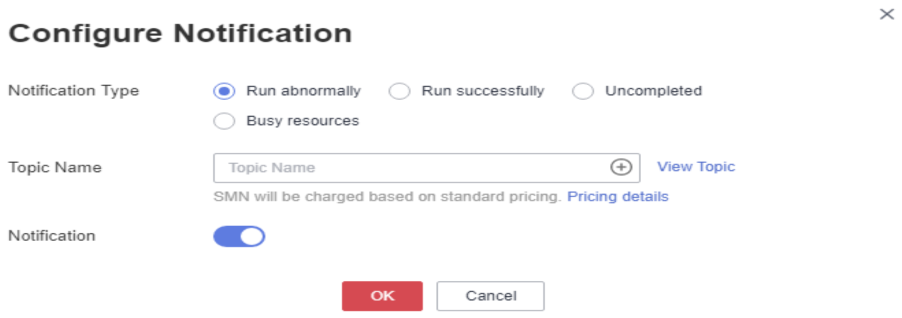
Symptom

How to configure notifications for all jobs

Solution

1. On the DataArts Studio console, locate a workspace and click **DataArts Factory**.
2. In the navigation pane on the left, choose **Monitoring** > **Job Monitoring**. Then click the **Batch Jobs** tab.
3. Select jobs and click **Configure Notification**.

Figure 6-12 Configuring notifications



Configure Notification

Notification Type: ● Run abnormally ○ Run successfully ○ Uncompleted ○ Busy resources

Topic Name: Topic Name

SMN will be charged based on standard pricing. Pricing details

Notification: on

4. Set notification parameters and click **OK**.

6.29 What Is the Maximum Number of Nodes That Can Be Executed Simultaneously?

The maximum number of nodes that can be executed simultaneously is related to the job node scheduling times per day. The following table lists the mapping.

To view the quota of the job node scheduling times per day, click **More** of a DataArts Studio instance and select **Quota Usage**.

Table 6-1 Maximum number of nodes that can run concurrently in a DataArts Studio instance

Job Node Scheduling Times/Day of a DataArts Studio Instance	Maximum Number of Nodes That Can Run Concurrently in a DataArts Studio Instance
<=500	10
<=5000	50
<=20000	100
<=40000	200
<=80000	300
> 80000	400

You can configure the maximum number of nodes that can be executed simultaneously in a workspace. The procedure is as follows.

Procedure

Step 1 Log in to the DataArts Studio console by following the instructions in **Accessing the DataArts Studio Instance Console**.

Step 2 On the DataArts Studio console, locate a workspace and click **DataArts Factory**.

Step 3 In the navigation pane, choose **Configuration** > **Configure**.

Step 4 Choose **Nodes Concurrently Running**.

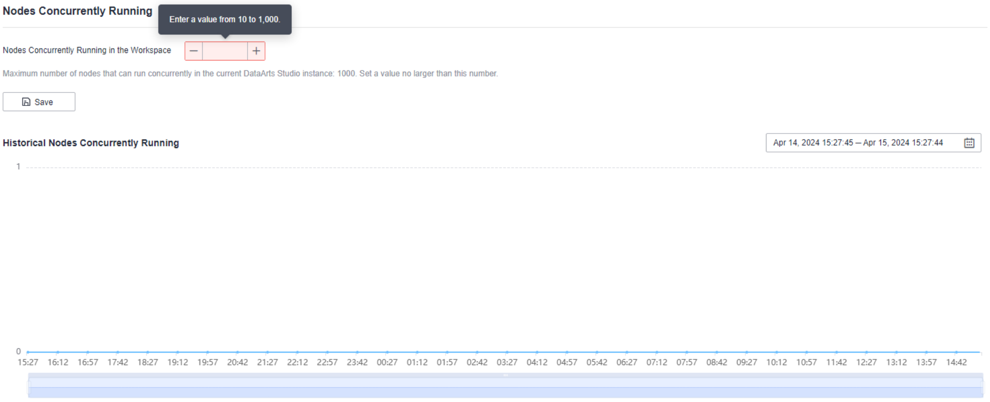
Step 5 Set **Nodes Concurrently Running in the Workspace**. Ensure that the value is less than or equal to the maximum number of nodes that can run concurrently in the DataArts Studio instance.

Table 6-2 lists the maximum number of nodes that can run concurrently in the DataArts Studio instance. To view the quota of the job node scheduling times per day, click **More** of a DataArts Studio instance and select **Quota Usage**.

Table 6-2 Maximum number of nodes that can run concurrently in a DataArts Studio instance

Job Node Scheduling Times/Day of a DataArts Studio Instance	Maximum Number of Nodes That Can Run Concurrently in a DataArts Studio Instance
<=500	10
<=5000	50
<=20000	100
<=40000	200
<=80000	300
> 80000	400

Figure 6-13 Configuring the number of concurrently running nodes



Enter a value from 10 to 1,000.

Nodes Concurrently Running in the Workspace

Maximum number of nodes that can run concurrently in the current DataArts Studio instance: 1000. Set a value no larger than this number.

Apr 14, 2024 15:27:45 — Apr 15, 2024 15:27:44

Step 6 Click **Save**.

----End

Viewing the Number of Historical Nodes Concurrently Running

Step 1 In the navigation pane, choose **Configuration** > **Configure**.

Step 2 Choose **Nodes Concurrently Running**.

Step 3 In the **Historical Nodes Concurrently Running** area, set the time range.

Step 4 Click **OK**.

NOTE

The maximum time range is 24 hours.

----End

6.30 Can I Change the Time Zone of a DataArts Studio Instance?

Symptom

Whether the time zone of a DataArts Studio instance can be changed. If not, how can the time of a DataArts Studio instance be adapted to the local time.

Solution

Currently, the time zone of a DataArts Studio instance cannot be changed.

During the scheduling of data development jobs, an EL expression can be used to adapt to the local time. The following is an example EL expression:

```
#{DateUtil.format(DateUtil.addHours(Job.planTime,-7),"yyyy-MM-dd")}
```

6.31 How Do I Synchronize the Changed Names of CDM Jobs to DataArts Factory?

Symptom

Updated names of CDM jobs cannot be synchronized to DataArts Factory.

Solution

After renaming a CDM job, you need to select the renamed CDM job in the CDM node attributes on the DataArts Factory console.

6.32 Why Does the Execution of an RDS SQL Statement Fail and an Error Is Reported Indicating That hll Does Not Exist?

Symptom

An RDS SQL statement fails to be executed and an error is reported indicating that hll does not exist.

Solution

By default, the hll plug-in is created in the public schema. The SQL statement must contain the schema to which the hll belongs.

6.33 What Should I Do If Error Message "The account has been locked" Is Displayed When I Am Creating a DWS Data Connection?

Cause Analysis

If the number of incorrect password attempts for connecting to the DWS cluster reaches the value of **failed_login_attempts** (10 by default), the account is automatically locked.

Solution

For details about how to unlock the account, see **How Do I Unlock an Account?**

6.34 What Should I Do If a Job Instance Is Canceled and Message "The node start execute failed, so the current node status is set to cancel." Is Displayed?

Symptom

A job instance is canceled and message "The node start execute failed, so the current node status is set to cancel." is displayed.

Solution

Some jobs on which this job depends failed. You can click the question mark (?) on the right of the canceled instance status to view the failed jobs.

6.35 What Should I Do If Error Message "Workspace does not exists" Is Displayed When I Call a DataArts Factory API?

Symptom

Error message "Workspace does not exists" is displayed when I call a DataArts Factory API.

Solution

Add the project ID to the request header in the code, that is, header.add("X-Project-Id",*Project ID*).

6.36 Why Don't the URL Parameters for Calling an API Take Effect in the Test Environment When the API Can Be Called Properly Using Postman?

Symptom

The URL parameters for calling an API take effect in the test environment when the API can be called properly using Postman.

Solution

Escape the connector (&) for URL parameters.

6.37 What Should I Do If Error Message "Agent need to be updated?" Is Displayed When I Run a Python Script?

Possible Causes

The version of the CDM cluster selected during host connection creation may be too early.

Solution

Use a CDM cluster of version 2.8.6 or a later version when creating a host connection.

6.38 Why Is an Execution Failure Displayed for a Node in the Log When the Node Status Is Successful?

Symptom

An execution failure is displayed for a node whose status is successful.

Solution

The **Succeed** operation was performed on the node, which changed the job instance (and node) status to successful.

6.39 What Should I Do If an Unknown Exception Occurs When I Call a DataArts Factory API?

Symptom

An unknown exception occurs when I call a DataArts Factory API.

Solution

DataArts Studio is a project-level service. Choose a project-level token.

6.40 Why Is an Error Message Indicating an Invalid Resource Name Is Displayed When I Call a Resource Creation API?

Symptom

An error message indicating an invalid resource name is displayed when I call a resource creation API?

Solution

The resource name can contain a maximum of 32 characters, including only letters, digits, underscores (_), and hyphens (-).

6.41 Why Does a PatchData Task Fail When All PatchData Job Instances Are Successful?

Symptom

A PatchData task fails when all PatchData job instances are successful.

Solution

The PatchData task may contain jobs in other workspaces. You can check the statuses of the job instance of the PatchData task in other workspaces.

6.42 Why Is a Table Unavailable When an Error Message Indicating that the Table Already Exists Is Displayed During Table Creation from a DWS Data Connection?

Cause Analysis

Role permissions are insufficient.

Solution

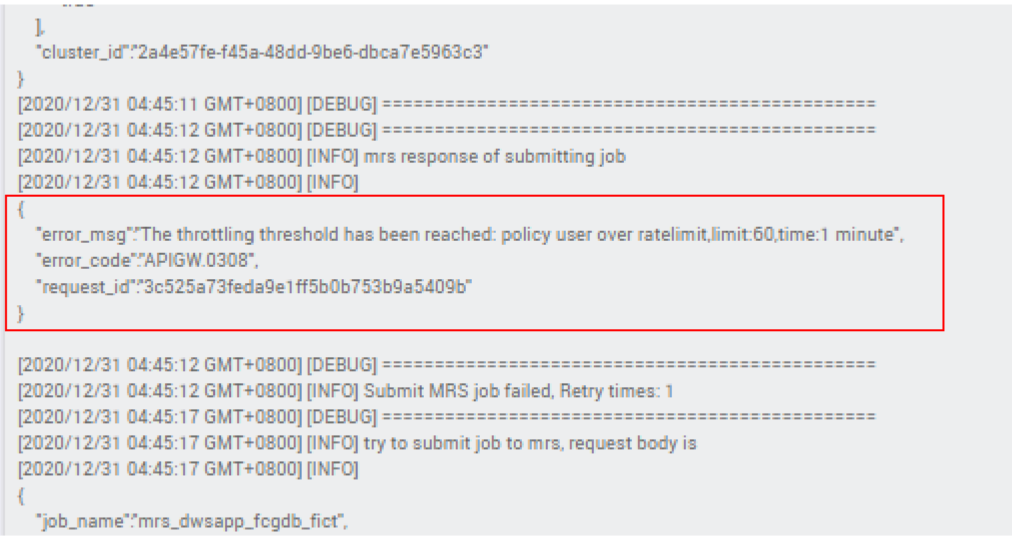
The table already exists, but you do not have permissions to view and edit it.

6.43 What Should I Do If Error Message "The throttling threshold has been reached: policy user over ratelimit,limit:60,time:1 minute." Is Displayed When I Schedule an MRS Spark Job?

Symptom

Error message "The throttling threshold has been reached: policy user over ratelimit,limit:60,time:1 minute" is displayed when I schedule an MRS Spark job.

Figure 6-14 Error message



```
   ],
   "cluster_id":"2a4e57fe-f45a-48dd-9be6-dbca7e5963c3"
}
[2020/12/31 04:45:11 GMT+0800] [DEBUG] ==============================================
[2020/12/31 04:45:12 GMT+0800] [DEBUG] ==============================================
[2020/12/31 04:45:12 GMT+0800] [INFO] mrs response of submitting job
[2020/12/31 04:45:12 GMT+0800] [INFO]
{
   "error_msg":"The throttling threshold has been reached: policy user over ratelimit,limit:60,time:1 minute",
   "error_code":"APIGW.0308",
   "request_id":"3c525a73feda9e1ff5b0b753b9a5409b"
}

[2020/12/31 04:45:12 GMT+0800] [DEBUG] ==============================================
[2020/12/31 04:45:12 GMT+0800] [INFO] Submit MRS job failed, Retry times: 1
[2020/12/31 04:45:17 GMT+0800] [DEBUG] ==============================================
[2020/12/31 04:45:17 GMT+0800] [INFO] try to submit job to mrs, request body is
[2020/12/31 04:45:17 GMT+0800] [INFO]
{
   "job_name":"mrs_dwsapp_fcgdb_fict",
```

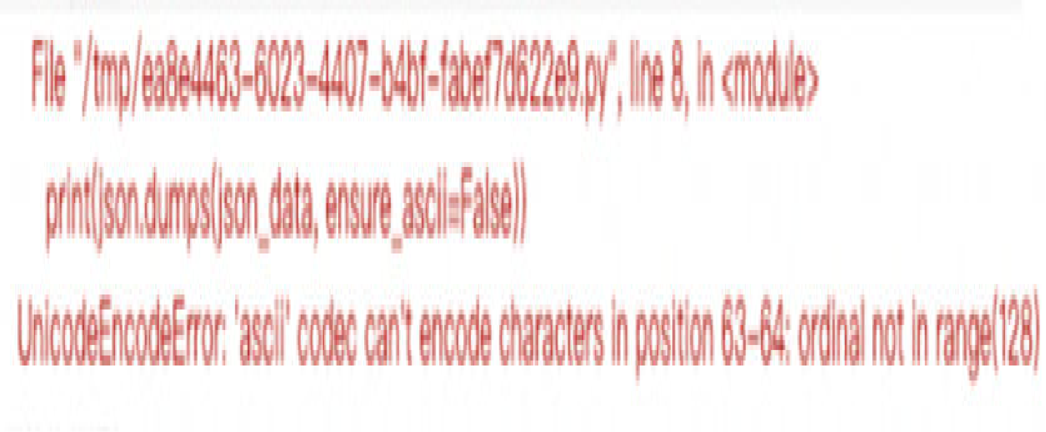
Solution

MRS APIs can be called by a single user for a maximum of 60 times per minute. Therefore, the solution is to call the API less frequently.

6.44 What Should I Do If Error Message "UnicodeEncodeError: 'ascii' codec can't encode characters in position 63-64: ordinal not in range(128)" Is Displayed When I Run a Python Script?

This error occurs when **json.dumps(json_data, ensure_ascii=False)** is configured in the Python script. The following figure shows the error.

Figure 6-15 Error message



```
File "/tmp/ea8e4463-6023-4407-b4bf-fabef7d622e9.py", line 8, in <module>
  print(json.dumps(json_data, ensure_ascii=False))
UnicodeEncodeError: 'ascii' codec can't encode characters in position 63-64: ordinal not in range(128)
```

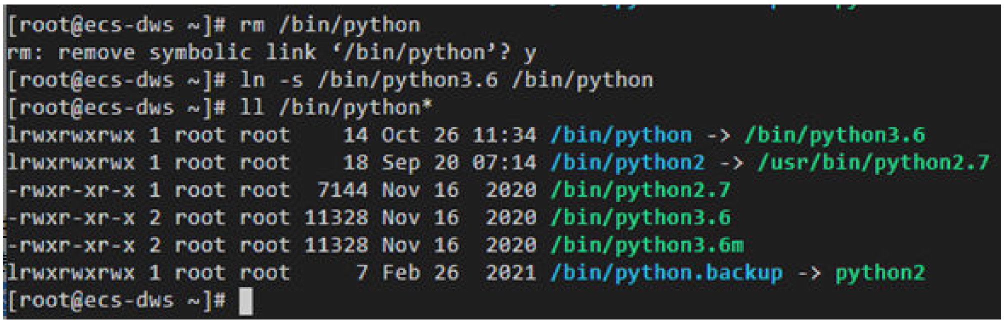
Possible Cause

By default, DataArts Studio uses the Python2 interpreter. This interpreter uses the ASCII encoding format by default and cannot encode Chinese characters. As a result, the error occurs. Therefore, you need to convert the encoding format to UTF-8.

Solution

1. Use the Python3 interpreter to create a soft connection on the host.

   Figure 6-16 Creating a soft connection on the host

```
[root@ecs-dws ~]# rm /bin/python
rm: remove symbolic link '/bin/python'? y
[root@ecs-dws ~]# ln -s /bin/python3.6 /bin/python
[root@ecs-dws ~]# ll /bin/python*
lrwxrwxrwx 1 root root    14 Oct 26 11:34 /bin/python -> /bin/python3.6
lrwxrwxrwx 1 root root    18 Sep 20 07:14 /bin/python2 -> /usr/bin/python2.7
-rwxr-xr-x 1 root root  7144 Nov 16  2020 /bin/python2.7
-rwxr-xr-x 2 root root 11328 Nov 16  2020 /bin/python3.6
-rwxr-xr-x 2 root root 11328 Nov 16  2020 /bin/python3.6m
lrwxrwxrwx 1 root root     7 Feb 26  2021 /bin/python.backup -> python2
[root@ecs-dws ~]#
```

2. Set the standard encoding mode in the file.

   # -*- coding: utf-8 -*-; Alternatively, set the encoding format for the host, that is, create a **sitecustomize.py** file in the **Lib\site-packages** folder in the Python installation directory and write the following information in the file:

```
# encoding=utf8
#import sys
#reload(sys)
#sys.setdefaultencoding('utf8')
```

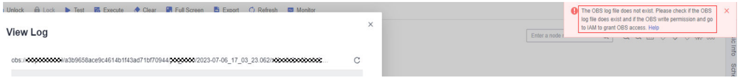
3. Restart Python and run **sys.getdefaultencoding()** to view the default encoding format, which is **utf8**.

6.45 What Should I Do If an Error Message Is Displayed When I View Logs?

Symptom

When you view the logs of a data development node, the error message shown in the following figure is displayed.

Figure 6-17 Message displayed



obs://[redacted]/a3b9658ace9c4614b1f43ad71bf70944/[redacted]/2023-07-06_17_03_23.062/[redacted]...

The OBS log file does not exist. Please check if the OBS log file does exist and if the OBS write permission and go to IAM to grant OBS access. Help

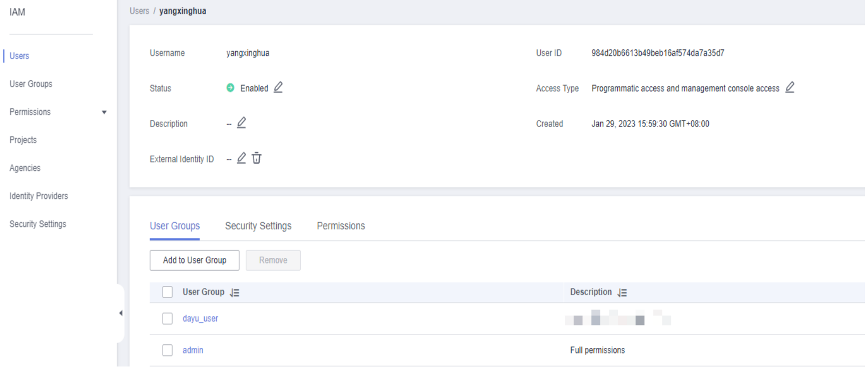
Possible Causes

Logs of data development jobs are stored in OBS buckets. This message is displayed if the user group to which you belong does not have the OBS operation permission, or no OBS log file is available.

Solution

1. Log in to the IAM management console as the administrator.
2. In the navigation pane, choose **Users**. Then click your username to go to the user information page.
3. Obtain the user group to which your user belongs.

Figure 6-18 User group to which your user belongs

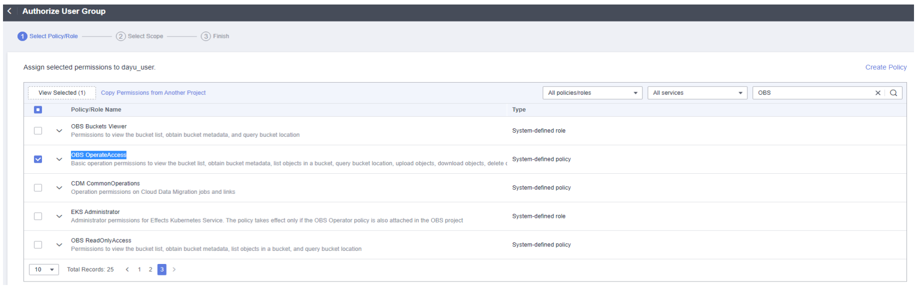


Username yangxinghua
User ID 984d20b6613b49beb16af574da7a35d7
Status Enabled
Access Type Programmatic access and management console access
Description --
Created Jan 29, 2023 15:59:30 GMT+08:00
External Identity ID --

User Group	Description
dayu_user	[redacted]
admin	Full permissions

4. In the navigation pane, choose **User Groups**. Locate the row that contains the user group obtained in the previous step and click **Authorize** in the **Operation** column.
5. On the displayed **Authorize User Group** page, search for and select the **OBS OperateAccess** or **OBS Administrator** permission.

Figure 6-19 Assigning permissions to the user group



Authorize User Group

1 Select Policy/Role — 2 Select Scope — 3 Finish

Assign selected permissions to dayu_user.

Create Policy

View Selected (1) Copy Permissions from Another Project

All policies/roles ▾ All services ▾ OBS

■		Policy/Role Name	Type
□	∨	OBS Buckets Viewer Permissions to view the bucket list, obtain bucket metadata, and query bucket location	System-defined role
☑	∨	OBS OperateAccess Basic operation permissions to view the bucket list, obtain bucket metadata, list objects in a bucket, query bucket location, upload objects, download objects, delete	System-defined policy
□	∨	CDM CommonOperations Operation permissions on Cloud Data Migration jobs and links	System-defined policy
□	∨	EKS Administrator Administrator permissions for Effects Kubernetes Service. The policy takes effect only if the OBS Operator policy is also attached in the OBS project	System-defined role
□	∨	OBS ReadOnlyAccess Permissions to view the bucket list, obtain bucket metadata, list objects in a bucket, and query bucket location	System-defined policy

10 ▾ Total Records: 25 < 1 2 3 >

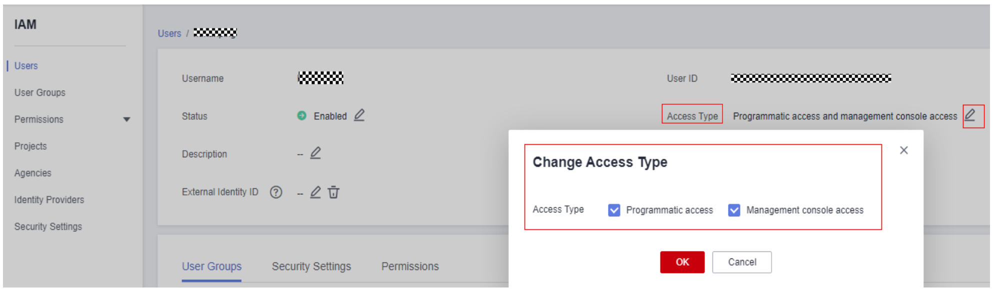
6. Click **Next** and select the scope requiring minimum authorization. The **All resources** option is selected by default.
7. Click **OK**.

   If the permissions are correct, check whether a log file exists in the OBS bucket.

Handling the Error After Running a Job

1. Log in to the IAM management console as the administrator.
2. In the navigation pane, choose **Users**. Then click your username to go to the user information page.
3. Click ✎ next to **Access Type** to change the access type.
4. Select **Programmatic access** and **Management console access**.

Figure 6-20 Configuring the access type



Users / [redacted]

Username [redacted]

User ID [redacted]

Status Enabled

Access Type Programmatic access and management console access

Description --

External Identity ID --

Change Access Type

Access Type ☑ Programmatic access ☑ Management console access

5. Click **OK**.

NOTICE

- When creating a workspace on the management console, you can set the OBS path for storing job logs only to an OBS object bucket rather than a parallel file system. If you do not set the OBS path for storing job logs, DataArts Factory writes logs to the **dlf-log-***{projectId}* bucket, and DataArts DataService writes logs to the **dlm-log-***{projectId}* bucket by default.
- If you do not select an existing OBS bucket for **Job Log Path**, the default **dlf-log-***{projectId}* bucket cannot be created when you run the job for the first time. As a result, logs cannot be written. To ensure that job logs can be properly written to the OBS bucket, select an existing OBS path when creating a workspace.

6.46 What Should I Do If a Shell/Python Node Fails and Error "session is down" Is Reported?

This section uses the Shell node as an example.

Symptom

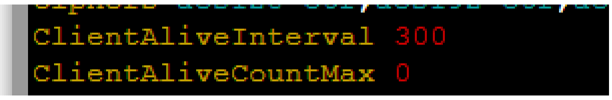
The Shell node fails to be executed, but the Shell script is executed successfully.

Possible Causes

1. Obtain the run logs of the Shell node.

```
[2021/11/17 02:00:36 GMT+0800] [INFO] No job-level agency is set, Workspace-level agency is
dlg_agency, Execute job use agency dlg_agency, job id is
07572F197E4642E5BE549C2B656F157Ctm7cHkHd
[2021/11/17 02:00:36 GMT+0800] [DEBUG]
================================================
[2021/11/17 02:00:36 GMT+0800] [INFO] Get response from agent when try to submit shell running
job :
[2021/11/17 02:00:36 GMT+0800] [INFO]
{
"jobResultList":[
{
"jobId":"a567f7f5-3c9e-4dfc-a464-bd477ac5b1ea",
"status":"created",
"errorCode":0,
"failCount":0,
"result":[

]
}
],
"agentId":"614853ee-c1c6-456d-9aa6-fc84ad1281ed"
}
[2021/11/17 02:00:36 GMT+0800] [DEBUG]
================================================
[2021/11/17 02:05:56 GMT+0800] [DEBUG]
================================================
[2021/11/17 02:05:56 GMT+0800] [INFO] Job Run finish , the raw output is :
[2021/11/17 02:05:56 GMT+0800] [INFO]
{
"jobId":"a567f7f5-3c9e-4dfc-a464-bd477ac5b1ea",
"status":"failed",
"errorCode":3427,
"errorMessage":"Shell script job execute failed.",
"failCount":0,
"result":[
{
"is_success":false,
"exeTime":300.609
}
]
}
[2021/11/17 02:05:56 GMT+0800] [DEBUG]
==============================================
[2021/11/17 02:05:56 GMT+0800] [DEBUG]
==============================================
[2021/11/17 02:05:56 GMT+0800] [INFO] The return code is : [-1].
[2021/11/17 02:05:56 GMT+0800] [DEBUG]
==============================================
[2021/11/17 02:05:56 GMT+0800] [INFO] Execute shell script job finished.
[2021/11/17 02:05:56 GMT+0800] [ERROR] Shell exit code is not 0
[2021/11/17 02:05:56 GMT+0800] [DEBUG]
==============================================
[2021/11/17 02:05:56 GMT+0800] [ERROR] Shell script job execute failed. Please contact ECS
Service.
[2021/11/17 02:05:56 GMT+0800] [ERROR] Exception message: RuntimeException: Shell script job
execute failed. Please contact ECS Service.
[2021/11/17 02:05:56 GMT+0800] [ERROR] Root Cause message:RuntimeException: Shell script job
execute failed. Please contact ECS Service.
```

2. Ensure that the values of the following parameters in the **sshd_config** file are as follows.

   ```
   ClientAliveInterval 300
   ClientAliveCountMax 0
   ```

   Cause: The SSH session times out and is disconnected. As a result, the Shell node fails.

Solution

1. Open the **/etc/ssh/sshd_config** file of the ECS and add or update the following parameter values:

   ClientAliveInterval 300

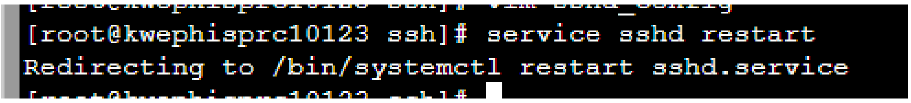
   ClientAliveCountMax 3

   NOTE

   The ClientAliveInterval parameter specifies the interval for the server to send requests to a client. The default value is **0**, indicating that the server does not send requests to the client. Value **300** indicates that the server sends a request every five minutes and the client sends a response accordingly. In this process, a persistent connection is maintained. The default value of **ClientAliveCountMax** is **3**. If the number of times that the client does not respond to requests sent by the server reaches the value of this parameter, the server disconnects the connection to the client. Normally, the client sends responses.

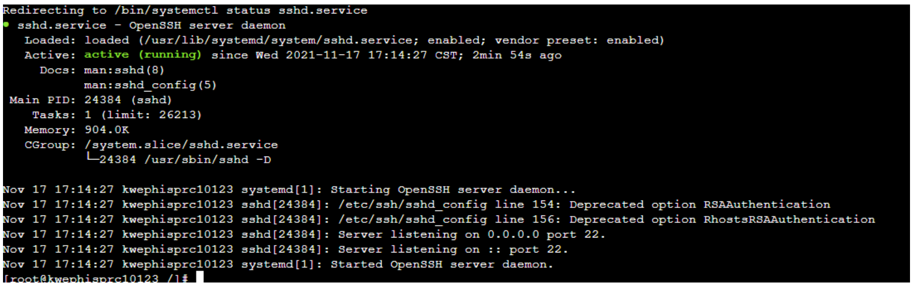
2. After the modification, restart the sshd of the ECS and run the following command:

   ```
   restart sshd.service
   ```

   ```
   [root@kwephisprc10123 ssh]# service sshd restart
   Redirecting to /bin/systemctl restart sshd.service
   ```

3. Check whether sshd is started successfully. (The following figure shows that sshd is started successfully.)



```
Redirecting to /bin/systemctl status sshd.service
● sshd.service - OpenSSH server daemon
   Loaded: loaded (/usr/lib/systemd/system/sshd.service; enabled; vendor preset: enabled)
   Active: active (running) since Wed 2021-11-17 17:14:27 CST; 2min 54s ago
     Docs: man:sshd(8)
           man:sshd_config(5)
 Main PID: 24384 (sshd)
    Tasks: 1 (limit: 26213)
   Memory: 904.0K
   CGroup: /system.slice/sshd.service
           └─24384 /usr/sbin/sshd -D

Nov 17 17:14:27 kwephisprc10123 systemd[1]: Starting OpenSSH server daemon...
Nov 17 17:14:27 kwephisprc10123 sshd[24384]: /etc/ssh/sshd_config line 154: Deprecated option RSAAuthentication
Nov 17 17:14:27 kwephisprc10123 sshd[24384]: /etc/ssh/sshd_config line 156: Deprecated option RhostsRSAAuthentication
Nov 17 17:14:27 kwephisprc10123 sshd[24384]: Server listening on 0.0.0.0 port 22.
Nov 17 17:14:27 kwephisprc10123 sshd[24384]: Server listening on :: port 22.
Nov 17 17:14:27 kwephisprc10123 systemd[1]: Started OpenSSH server daemon.
[root@kwephisprc10123 /]# █
```

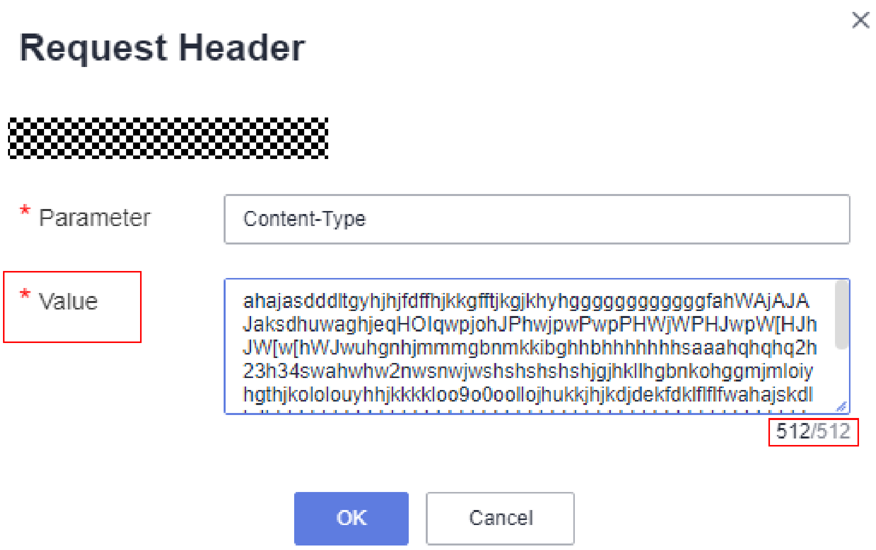
6.47 What Should I Do If a Parameter Value in a Request Header Contains More Than 512 Characters?

The Rest Client operator is used as an example.

Symptom

When configuring a parameter for a job operator, you need to enter the parameter name and value. If you already enter 512 characters for the parameter value, you cannot enter more characters.

Figure 6-21 Configuring a request header parameter



Request Header

[redacted]

* Parameter: Content-Type

* Value: ahajasdddltgyhjhjfdffhjkkgfftjkgjkhyhgggggggggggfahWAjAJA JaksdhuwaghjeqHOIqwpjohJPhwjpwPwpPHWjWPHJwpW[HJh JW[w[hWJwuhgnhjmmmgbnmkkibghhbhhhhhhhsaaahqhqhq2h 23h34swahwhw2nwsnwjwshshshshshjgjhkllhgbnkohggmjmloiy hgthjkololouyhhjkkkkloo9o0oollojhukkjhjkdjdekfdklflflfwahajskdl

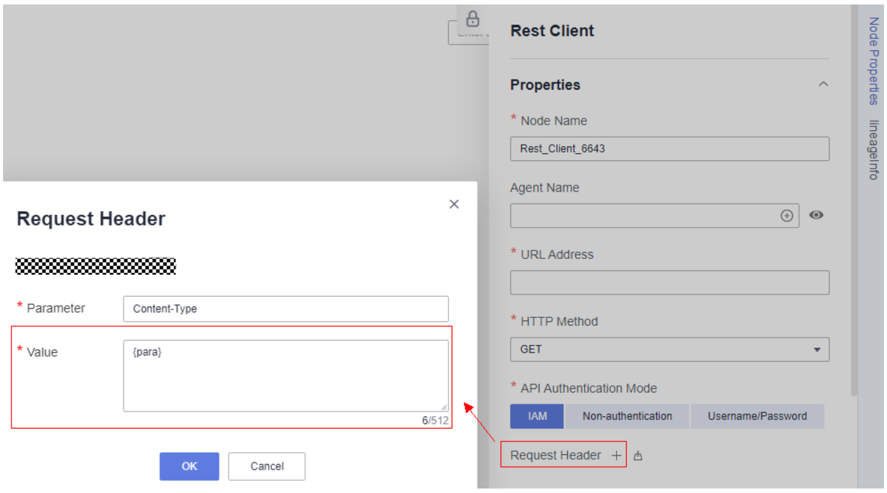
512/512

Solution

1. Configure a request header parameter for the job node.

   Enter a variable name for **Value**, for example, **{para}**.

Figure 6-22 Configuring a request header parameter



Rest Client

Properties

* Node Name: Rest_Client_6643

Agent Name

* URL Address

* HTTP Method: GET

* API Authentication Mode: IAM | Non-authentication | Username/Password

Request Header

[redacted]

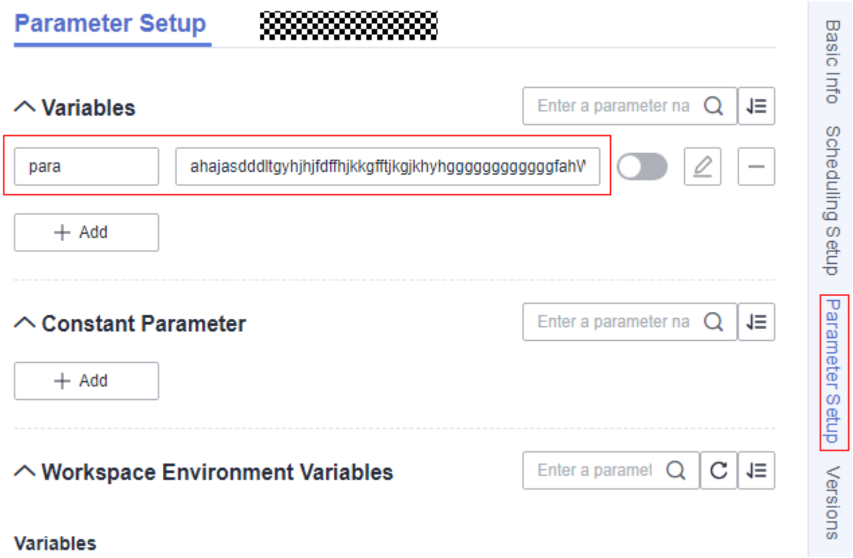
* Parameter: Content-Type

* Value: {para}

6/512

2. Configure job parameters.
   a. Click **Parameter Setup**.
   b. Enter variable **para** and its value. The value can contain a maximum of 1,024 characters.

Figure 6-23 Configuring job parameters



Parameter Setup [redacted]

Variables

para: ahajasdddltgyhjhjfdffhjkkgfftjkgjkhyhggggggggggggfahV

Constant Parameter

Workspace Environment Variables

Variables

The preceding method resolves the length issue of the request header parameter value.

6.48 What Should I Do If a Message Is Displayed Indicating that the ID Does Not Exist During the Execution of a DWS SQL Script?

Possible Causes

This issue is caused by the case of the ID.

Solution

During the execution of a DWS SQL script, the system uses lowercase letters by default. If a field is in upper case, add "".

Example: select * from *table1* order by "ID";

```
select * from table order by "ID";
```

6.49 How Do I Check Which Jobs Invoke a CDM Job?

Symptom

Jobs that invoke a CDM job

Solution

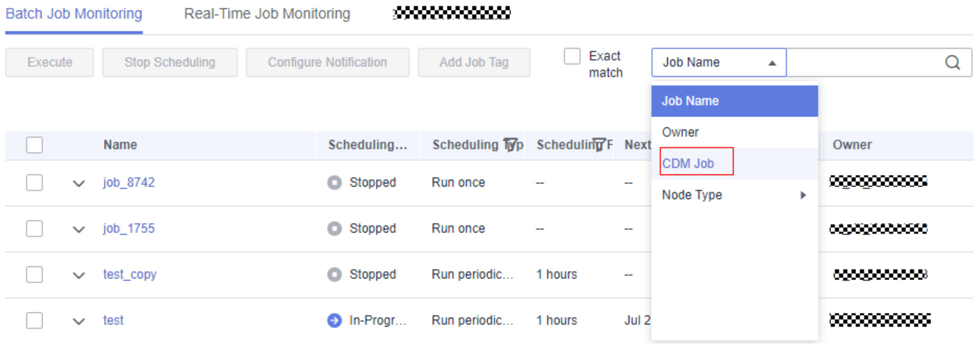
1. In the left navigation pane of DataArts Factory, choose **Monitoring** > **Job Monitoring**.
2. Click the **Batch Job Monitoring** tab.
3. Query the scheduling and execution of the CDM job by filter criteria.

   NOTE

   You can query the scheduling and execution information about CDM jobs by **CDM Job**.

   You can query the scheduling and execution information of CDMJob operators by **CDMJob** in **Node Type**.

Figure 6-24 Batch Job Monitoring



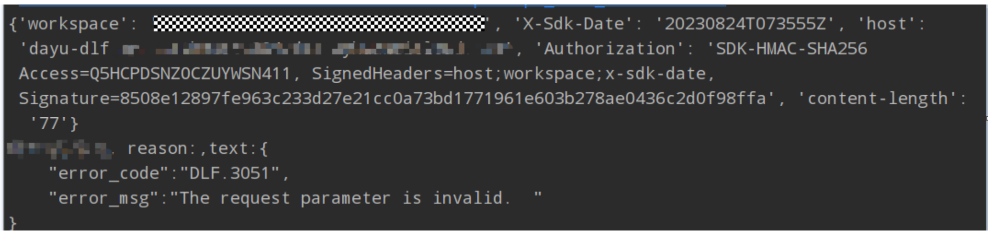
□		Name	Scheduling...	Scheduling Typ	Schedulin F	Next	Owner
□	∨	job_8742	Stopped	Run once	--	--	[redacted]
□	∨	job_1755	Stopped	Run once	--	--	[redacted]
□	∨	test_copy	Stopped	Run periodic...	1 hours	--	[redacted]
□	∨	test	In-Progr...	Run periodic...	1 hours	Jul 2	[redacted]

6.50 What Should I Do If Error Message "The request parameter invalid" Is Displayed When I Use Python to Call the API for Executing Scripts?

Symptom

Error message "The request parameter invalid" is displayed when Python is used to call the API for executing scripts.

Call the script execution API by following the instructions in **Executing a Script**.



```
{'workspace': [redacted]', 'X-Sdk-Date': '20230824T073555Z', 'host':
 'dayu-dlf [redacted]', 'Authorization': 'SDK-HMAC-SHA256
 Access=Q5HCPDSNZOCZUYWSN411, SignedHeaders=host;workspace;x-sdk-date,
 Signature=8508e12897fe963c233d27e21cc0a73bd1771961e603b278ae0436c2d0f98ffa', 'content-length':
  '77'}
[redacted] reason:,text:{
    "error_code":"DLF.3051",
    "error_msg":"The request parameter is invalid.  "
}
```

View logs.

Error: Content type 'application/octet-stream' not supported

Possible Causes

The value of the **content-type** parameter is **application/json**.

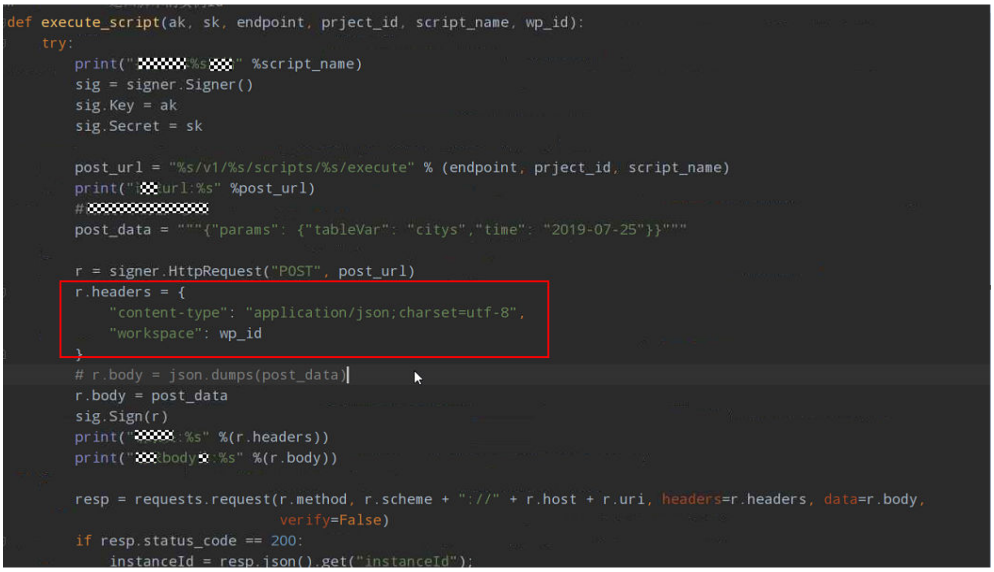
NOTE

content-type indicates the request body type or format. Its default value is **application/json**.

If the request body contains Chinese characters, use **charset=utf8** to specify the Chinese character set.

Solution

Change the value of **content-type**.



```
def execute_script(ak, sk, endpoint, prject_id, script_name, wp_id):
    try:
        print("[redacted]%s[redacted]" %script_name)
        sig = signer.Signer()
        sig.Key = ak
        sig.Secret = sk

        post_url = "%s/v1/%s/scripts/%s/execute" % (endpoint, prject_id, script_name)
        print("[redacted]url:%s" %post_url)
        #[redacted]
        post_data = """{"params": {"tableVar": "citys","time": "2019-07-25"}}"""

        r = signer.HttpRequest("POST", post_url)
        r.headers = {
            "content-type": "application/json;charset=utf-8",
            "workspace": wp_id
        }
        # r.body = json.dumps(post_data)
        r.body = post_data
        sig.Sign(r)
        print("[redacted]:%s" %(r.headers))
        print("[redacted]body[redacted]:%s" %(r.body))

        resp = requests.request(r.method, r.scheme + "://" + r.host + r.uri, headers=r.headers, data=r.body,
                                verify=False)
        if resp.status_code == 200:
            instanceId = resp.json().get("instanceId");
```

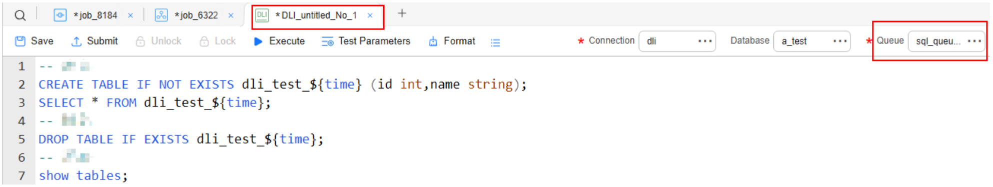
After the value of **content-type** is changed, the API can be called successfully.

6.51 What Should I Do If the Default Queue of a New DLI SQL Script in DataArts Factory Has Been Deleted?

Symptom

The default queue of a new DLI SQL script has been deleted.

Figure 6-25 DLI SQL script



* Connection dli

Database a_test

* Queue sql_queu...

```
-- [redacted]
CREATE TABLE IF NOT EXISTS dli_test_${time} (id int,name string);
SELECT * FROM dli_test_${time};
-- [redacted]
DROP TABLE IF EXISTS dli_test_${time};
-- [redacted]
show tables;
```

Possible Cause

When a DLI SQL script is used or opened, the queue selected for the script is stored in the cache. This queue is automatically selected for another DLI SQL script created in the workspace.

Solution

When creating a DLI SQL script in the workspace, select a valid DLI resource queue. When you create DLI SQL scripts in the future, this valid DLI resource queue will be used by default.

6.52 Does the Event-based Scheduling Type in DataArts Factory Support Offline Kafka?

Symptom

During the configuration of a job in DataArts Factory, offline Kafka cannot be selected when **Scheduling Type** is set to **Event-based** and **Event Type** is set to **KAFKA**.

Solution

Only MRS Kafka is supported in this case.

7 DataArts Quality

7.1 What Are the Differences Between Quality Jobs and Comparison Jobs?

Possible Causes

Differences between quality jobs and comparison jobs

Solution

- You can create quality jobs to apply the created rules to existing tables.
- Comparison jobs support cross-source data comparison. You can apply created rules to two tables for quality monitoring and output the comparison result.

  Data comparison is critical to ensure data consistency in data development and migration. The cross-source data comparison capability is the key to checking consistency of the data before and after migration or processing.

7.2 How Can I Confirm that a Quality Job or Comparison Job Is Blocked?

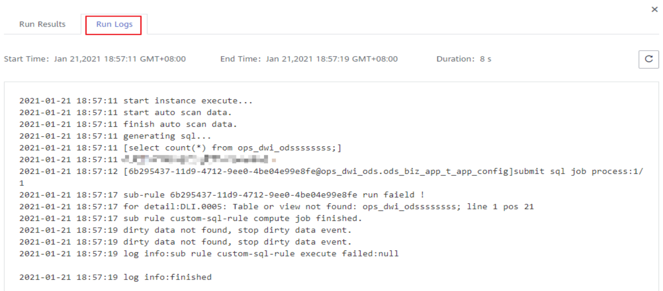
Possible Causes

How to check whether a quality job or comparison job is blocked

Solution

If a job is in the running state for a long period of time, choose **Quality Monitoring** > **O&M Management** in the navigation pane of the Data Quality Control page, click **Details** in the **Operation** column, and then click **Run Logs**. If the run log is not updated, the job is blocked.

Figure 7-1 Job run logs



Start Time: Jan 21,2021 18:57:11 GMT+08:00 End Time: Jan 21,2021 18:57:19 GMT+08:00 Duration: 8 s

```
2021-01-21 18:57:11 start instance execute...
2021-01-21 18:57:11 start auto scan data.
2021-01-21 18:57:11 finish auto scan data.
2021-01-21 18:57:11 generating sql...
2021-01-21 18:57:11 [select count(*) from ops_dwi_odsssssssss;]
2021-01-21 18:57:11 [redacted]
2021-01-21 18:57:12 [6b295437-11d9-4712-9ee0-4be04e99e8fe@ops_dwi_ods.ods_biz_app_t_app_config]submit sql job process:1/
1
2021-01-21 18:57:17 sub-rule 6b295437-11d9-4712-9ee0-4be04e99e8fe run faield !
2021-01-21 18:57:17 for detail:DLI.0005: Table or view not found: ops_dwi_odsssssssss; line 1 pos 21
2021-01-21 18:57:17 sub rule custom-sql-rule compute job finished.
2021-01-21 18:57:19 dirty data not found, stop dirty data event.
2021-01-21 18:57:19 dirty data not found, stop dirty data event.
2021-01-21 18:57:19 log info:sub rule custom-sql-rule execute failed:null

2021-01-21 18:57:19 log info:finished
```

7.3 How Do I Manually Restart a Blocked Quality Job or Comparison Job?

Possible Causes

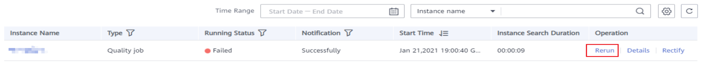
How to restart a blocked quality job or comparison job

Solution

A blocked job will be automatically terminated if it is not started within one day.

To manually restart a blocked job, choose **Quality Monitoring** > **O&M Management** in the navigation pane of the Data Quality Control page, and click **Cancel** in the **Operation** column of the job. After the job status changes to **Failed**, click **Rerun** in the **Operation** column to restart the job.

Figure 7-2 Running a job



Instance Name	Type	Running Status	Notification	Start Time	Instance Search Duration	Operation
[redacted]	Quality job	● Failed	Successfully	Jan 21,2021 19:00:40 G...	00:00:09	Rerun | Details | Rectify

7.4 How Do I View Jobs Associated with a Quality Rule Template?

Possible Causes

How to view the jobs associated with a quality rule template

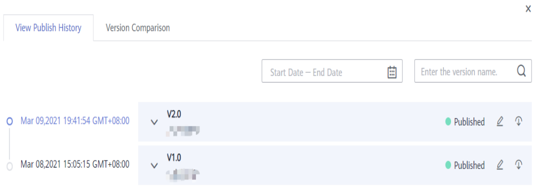
Solution

Step 1 On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Quality**.

Step 2 Choose **Quality Monitoring** > **Rule Templates**.

Step 3 Click **View History** in the **Operation** column of the target rule template.

Figure 7-3 Viewing publish history

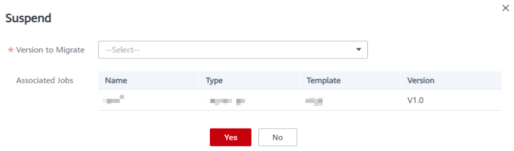


Mar 09,2021 19:41:54 GMT+08:00 V2.0 [redacted]

Mar 08,2021 15:05:15 GMT+08:00 V1.0 [redacted]

Step 4 Click **Suspend** on the right of a historical version. You can view the jobs associated with the rule template.

Figure 7-4 Viewing associated jobs



Suspend

* Version to Migrate: --Select--

Associated Jobs

Name	Type	Template	Version
[redacted]	[redacted]	[redacted]	V1.0

----End

7.5 What Should I Do If the System Displays a Message Indicating that I Do Not Have the MRS Permission to Perform a Quality Job?

Possible Causes

An error is reported when you execute a quality job. The following information is recorded in the job log: "The current user does not exist on MRS Manager. Grant the user sufficient permissions on IAM and then perform IAM user synchronization on the Dashboard tab page!".

Solution

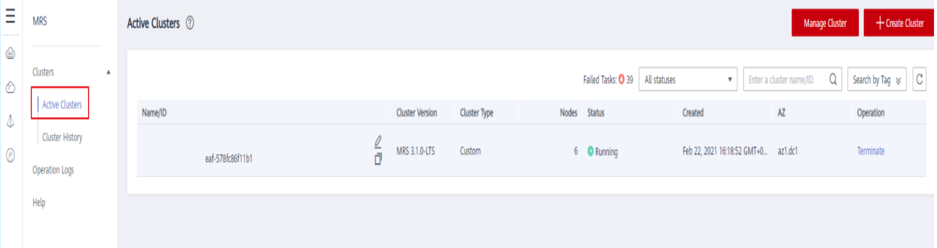
This problem occurs because the user does not have the operation permission on the MRS cluster.

If the user is newly added to a tenant, find the corresponding MRS cluster instance on the MRS cluster list page and click **Synchronize**.

The procedure is as follows:

Step 1 Log in to the MRS console, view the existing clusters, and click a cluster name to access the cluster overview page.

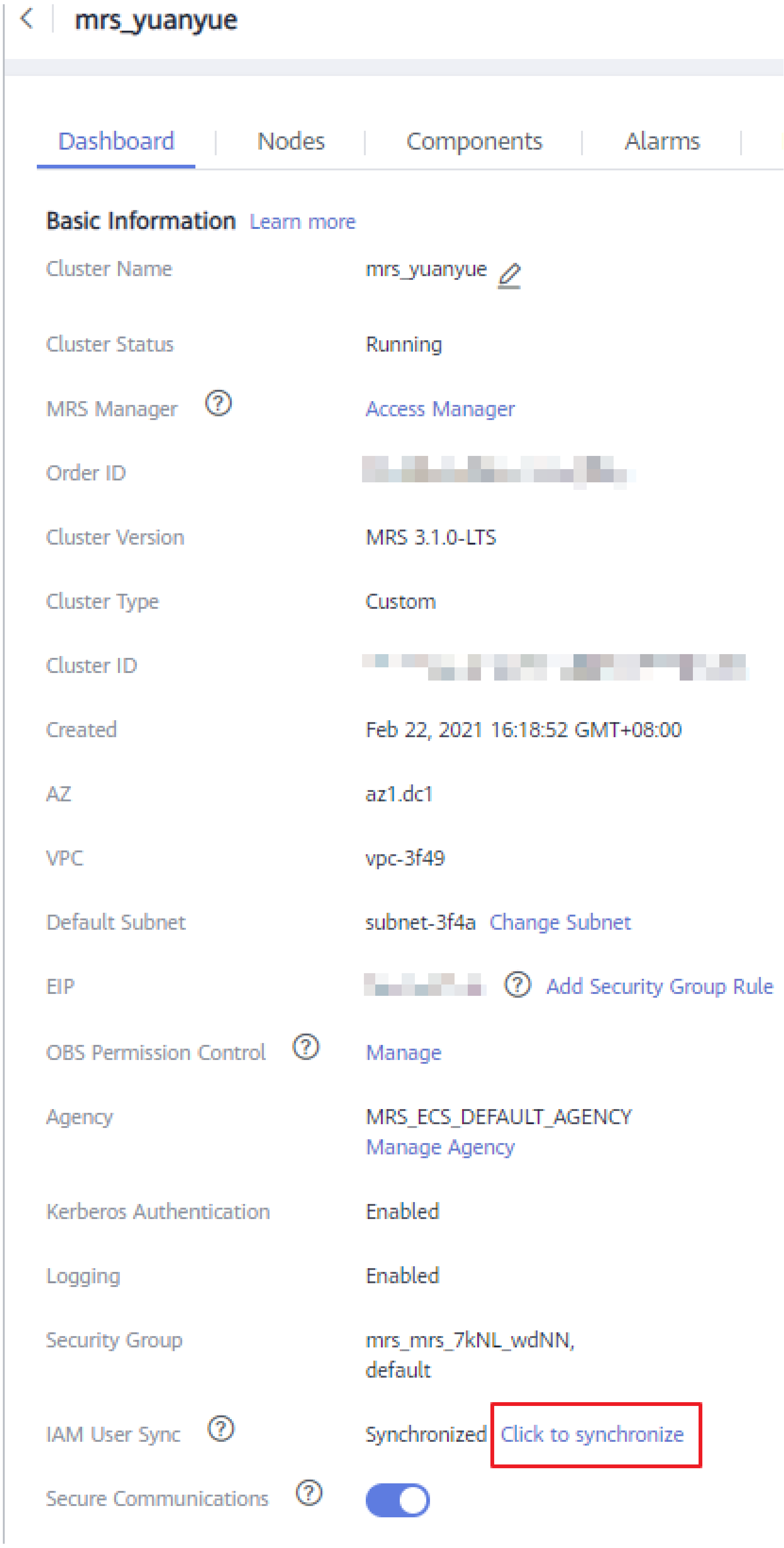
Figure 7-5 MRS cluster instance



Name/ID	Cluster Version	Cluster Type	Nodes	Status	Created	AZ	Operation
eaf-578fc86f11b1	MRS 3.1.0-LTS	Custom	6	Running	Feb 22, 2021 16:18:52 GMT+0...	az1.dc1	Terminate

Step 2 In the **IAM User Sync** area, click **Click to synchronize**.

Figure 7-6 Click to synchronize



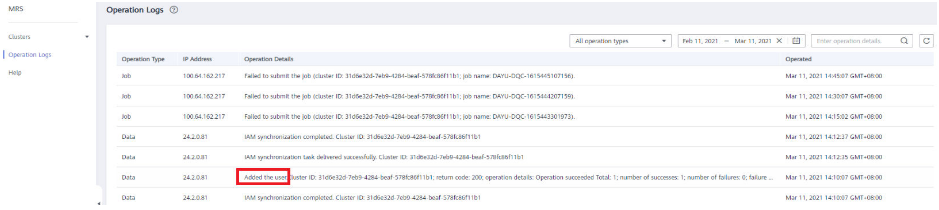
mrs_yuanyue

Basic Information Learn more

Cluster Name	mrs_yuanyue
Cluster Status	Running
MRS Manager	Access Manager
Order ID	[redacted]
Cluster Version	MRS 3.1.0-LTS
Cluster Type	Custom
Cluster ID	[redacted]
Created	Feb 22, 2021 16:18:52 GMT+08:00
AZ	az1.dc1
VPC	vpc-3f49
Default Subnet	subnet-3f4a Change Subnet
EIP	[redacted] Add Security Group Rule
OBS Permission Control	Manage
Agency	MRS_ECS_DEFAULT_AGENCY Manage Agency
Kerberos Authentication	Enabled
Logging	Enabled
Security Group	mrs_mrs_7kNL_wdNN, default
IAM User Sync	Synchronized Click to synchronize
Secure Communications	

Step 3 View the operation result in the **Operation Logs** area.

Figure 7-7 Operation log



Operation Type	IP Address	Operation Details	Operated
Job	100.64.162.217	Failed to submit the job (cluster ID: 31d6e32d-7eb9-4284-beaf-578fc86f11b1; job name: DAYU-DQC-1615445107156).	Mar 11, 2021 14:45:07 GMT+08:00
Job	100.64.162.217	Failed to submit the job (cluster ID: 31d6e32d-7eb9-4284-beaf-578fc86f11b1; job name: DAYU-DQC-1615444207159).	Mar 11, 2021 14:30:07 GMT+08:00
Job	100.64.162.217	Failed to submit the job (cluster ID: 31d6e32d-7eb9-4284-beaf-578fc86f11b1; job name: DAYU-DQC-1615443301973).	Mar 11, 2021 14:15:02 GMT+08:00
Data	24.2.0.81	IAM synchronization completed. Cluster ID: 31d6e32d-7eb9-4284-beaf-578fc86f11b1	Mar 11, 2021 14:12:37 GMT+08:00
Data	24.2.0.81	IAM synchronization task delivered successfully. Cluster ID: 31d6e32d-7eb9-4284-beaf-578fc86f11b1	Mar 11, 2021 14:12:35 GMT+08:00
Data	24.2.0.81	Added the user cluster ID: 31d6e32d-7eb9-4284-beaf-578fc86f11b1; return code: 200; operation details: Operation succeeded Total: 1; number of successes: 1; number of failures: 0; failure ...	Mar 11, 2021 14:10:07 GMT+08:00
Data	24.2.0.81	IAM synchronization completed. Cluster ID: 31d6e32d-7eb9-4284-beaf-578fc86f11b1	Mar 11, 2021 14:10:07 GMT+08:00

Step 4 After the preceding steps are complete, the account has been synchronized. If the system still displays a message indicating that you lack the MRS permission, log in to the Manager and create an account with the same name as the current primary account.

NOTE

You need to create an account with the same name as the current primary account.

----End

8 DataArts Catalog

8.1 What Are the Functions of the DataArts Catalog Module?

The DataArts Catalog module collects metadata and displays a data asset map in a workspace of an enterprise, which contains metadata and data lineages.

- Metadata management

  The metadata management module is the cornerstone of data lake governance. It allows users to create tasks with custom policies for collecting technical metadata from data sources. and customize a business metamodel to batch import business metadata, associate business metadata with the technical metadata, and manage and apply linkages throughout the entire chain.

- Data map

  Data maps facilitate data search, analysis, development, mining, and operations. They provide lineage information and impact analysis. Data maps make data search easier and faster than before.

  - Keyword search and fuzzy search are supported, helping you quickly locate the data you need.
  - Data maps can be also used to query table details by table names, helping users quickly master usage rules. After obtaining the detailed data information, users can add additional description.
  - Data maps display the source, destination, and processing logic of each table and field.
  - You can classify and tag data assets as required.

8.2 What Assets Can Be Collected by DataArts Catalog?

DataArts Catalog can collect data assets from data lakes, such as MRS Hive, DLI, and GaussDB(DWS), and also the metadata of the following data sources:

1. Relational databases, such as MySQL and PostgreSQL databases (You can use RDS connections to collect the metadata of these databases.)

2. Cloud Search Service (CSS)
3. Graph Engine Service (GES)
4. Object Storage Service (OBS)
5. MRS Hudi (MRS Hudi is a data format. The metadata is stored in Hive, and operations are performed using Spark.) You can enable synchronization of the Hive table configuration for Hudi tables, and then you can collect the metadata of Hudi tables by collecting the MRS Hive metadata.

For details, see **Data Sources**.

8.3 What Is Data Lineage?

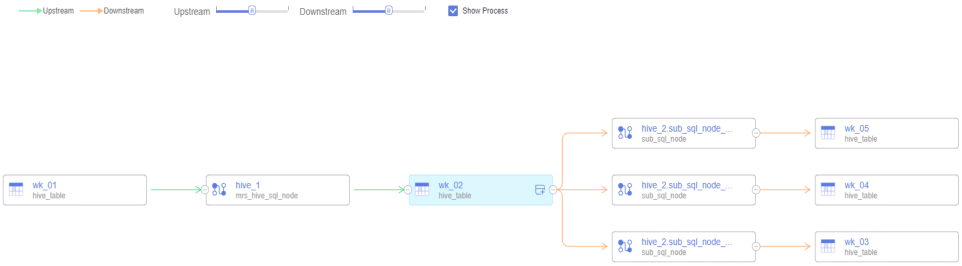
In the era of big data, various types of data are rapidly generated due to explosive data growth. The massive and complex data information is converged, transformed, and transferred to generate new data and aggregate into an ocean of data.

During this process, a relationship is formed between the data, and these relationships are their lineages. They are analogous to the genetic relationships between people. However, in contrast from our human lineages, data lineages have the following distinct features:

- **Belongingness**: Specific data belongs to a specific organization or individual.
- **Multi-source**: One piece of data can have multiple sources. One piece of data may be generated by processing multiple pieces of data, and there may be multiple such processes.
- **Traceability**: The data lineage is traceable. It reflects the data lifecycle and the entire process from data generation to data disappearance.
- **Hierarchy**: The data lineage is hierarchical. Data classification and summary form new data, and different levels of description result in data layers.

Figure 8-1 shows the lineage relationship graph for DataArts Studio. ▦ indicates a data table, and ⑂ indicates a job node. They are orchestrated using arrows. As shown in the graph, the data in table **wk_01** is processed on the **hive_1** job node and then written to table **wk_02**. The data in table **wk_02** is processed on the **hive_2** job node and written to tables **wk_03**, **wk_04**, and **wk_05**, respectively.

Figure 8-1 Data lineage example



8.4 How Do I Visualize Data Lineages in a Data Catalog?

To display data lineages in DataArts Catalog, you must complete metadata collection tasks first. In addition, ensure that data development jobs contain the node types and scenarios that support automatic lineage parsing, or that input and output tables of lineages have been customized in job nodes. If there is a successful data development job scheduling task, the system generates the lineage in the job and displays it in DataArts Catalog.

For details about the generation and display of data lineages, see **Node Lineages**.

9 DataArts Security

9.1 Why Isn't Data Masked Based on a Specified Rule After a Data Masking Task Is Executed?

Possible Causes

Static masking tasks depend on sensitive data discovery tasks. If the statuses of sensitive data fields were not changed to **Valid** on the **Sensitive Data Distribution** page, the system considers that there is no sensitive field, so does not mask data based on a rule.

Solution

Before creating a static masking task, you must create a sensitive data discovery task. After sensitive fields are discovered, change the statuses of the sensitive fields to **Valid** on the **Sensitive Data Distribution** page.

9.2 What Should I Do If a Message Is Displayed Indicating that Necessary Request Parameters Are Missing When I Approve a GaussDB(DWS) Permission Application?

Possible Causes

The object to be authorized is not synchronized to the GaussDB(DWS) data source.

Solution

You can synchronize the object to be authorized to the GaussDB(DWS) data source by synchronizing users, and try again.

9.3 What Should I Do If Error Message "FATAL: Invalid username/password,login denied" Is Displayed During the GaussDB(DWS) Connectivity Check When Fine-grained Authentication Is Enabled?

Possible Causes

The current user is not synchronized to the GaussDB(DWS) data source or does not have th GaussDB(DWS) Database Access permission.

Solution

Synchronize the current login user to the GaussDB(DWS) data source, grant the DWS Database Access permissions to the user, and test the connectivity again.

9.4 What Should I Do If Error Message "Failed to obtain the database" Is Displayed When I Select a Database in DataArts Factory After Fine-grained Authentication Is Enabled?

Possible Causes

The data development user does not have the DWS Database Access permission.

Solution

Grant the DWS Database Access permission to the data development user and try selecting a database again.

9.5 Why Does the System Display a Message Indicating Insufficient Permissions During Permission Synchronization to DLI?

The tasks of synchronizing permissions to DLI are completed through the cloud service agency (**dlg_agency**). The agency must have the permissions listed in **Table 9-1**.

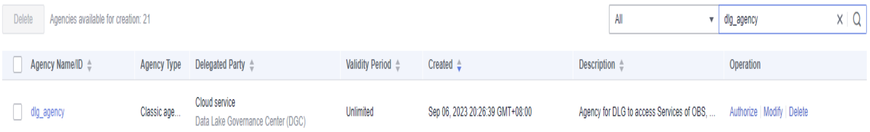
Table 9-1 Required permissions

Permission	Purpose	Mandatory	Authorization Item/System Permission (Configure Either of Them)	
IAM permission	This permission is required for the system to obtain users or user groups, or create roles. For example, user synchronization fails if this permission is missing.	Yes	• iam:users:listUsers • iam:groups:listGroups • iam:users:listUsersForGroup • iam:roles:createRole • iam:roles:deleteRole • iam:roles:updateRole • iam:permissions:grantRoleToGroup • iam:permissions:listRoleAssignments • iam:permissions:revokeRoleFromGroup	Security Administrator
Permission for synchronizing permissions to DLI	This permission is required for DLI permission synchronization. For example, if this permission is missing, DLI permission synchronization fails and the system displays a message indicating insufficient permissions.	Mandatory for DLI permission management	• dli:database:grantPrivilege • dli:table:grantPrivilege • dli:column:grantPrivilege • dli:queue:grantPrivilege	DLI FullAccess

If this message is displayed, perform the following operations to grant permissions (system permissions in this example) to **dlg_agency**:

1. Log in to the IAM console.
2. In the navigation pane, choose **Agencies**.
3. Search for **dlg_agency** and click **Authorize** in the **Operation** column.

Figure 9-1 Granting permissions to dlg_agency

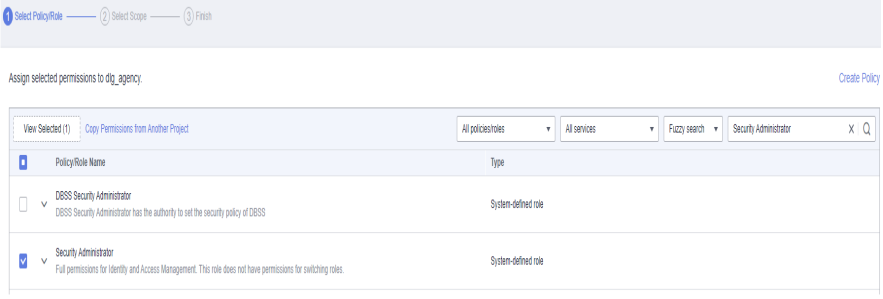


Agencies available for creation: 21

☐ Agency Name/ID	Agency Type	Delegated Party	Validity Period	Created	Description	Operation
☐ dlg_agency	Classic age...	Cloud service Data Lake Governance Center (DGC)	Unlimited	Sep 06, 2023 20:26:39 GMT+08:00	Agency for DLG to access Services of OBS, ...	Authorize | Modify | Delete

4. On the displayed page, search for and select **Security Administrator** and **DLI FullAccess**, and click **Next**.

Figure 9-2 Selecting Security Administrator



Assign selected permissions to dlg_agency.

■	Policy/Role Name	Type
☐	DBSS Security Administrator DBSS Security Administrator has the authority to set the security policy of DBSS	System-defined role
☑	Security Administrator Full permissions for Identity and Access Management. This role does not have permissions for switching roles.	System-defined role

5. Click **OK**. Wait for 15 to 30 minutes. The permissions will be synchronized to DLI.

10 DataArts DataService

10.1 What Languages Do DataArts DataService SDKs Support?

This type of SDK is encapsulated based on the data APIs created in DataArts DataService of DataArts Studio. By invoking the sample code provided by the SDK, you can call the data APIs in DataArts DataService to obtain open data easily and quickly.

DataArts DataService SDKs support C#, Python, Go, JavaScript, PHP, C++, C, Android, Java, and other languages. For details, see **SDK Reference**.

10.2 What Can I Do If the System Displays a Message Indicating that the Proxy Fails to Be Invoked During API Creation?

Possible Causes

The CDM agent in the data connection is abnormal, for example, the memory usage is too high.

Solution

In a short term, you are advised to restart the CDM cluster during offpeak hours. In a long term, you need to reduce the workload of the CDM cluster.

10.3 What Should I Do If the Background Reports an Error When I Access the Test App Through the Data Service API and Set Related Parameters?

Possible Causes

The header parameter is not set.

Solution

Set the header parameter when invoking the API.

```
header parameter: x-Authorization, nvalid ___ parameter: ___,
```

10.4 How Many Times Can a Subdomain Name Be Accessed Using APIs Every Day?

It depends on the target to which the API is published.

- In DataArts DataService Exclusive, the API is published to a DataArts DataService Exclusive cluster by default and can be published by version. After the API is published, it can be called through the intranet or Internet.

For details, see **Publishing an API**.

10.5 Can Operators Be Transferred When API Parameters Are Transferred?

No.

Only parameters are transferred. Operators are fixed. To transfer multiple parameters, use the **in(${})** method.

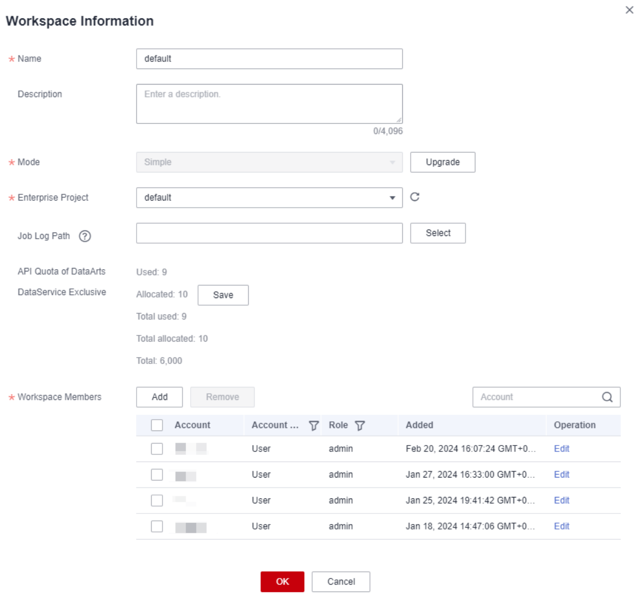
10.6 What Should I Do If No More APIs Can Be Created When the API Quota in the Workspace Is Used Up?

By default, the total API quota for a DataArts DataService Exclusive cluster in a DataArts Studio instance is 5,000 by default. If the API quota for a workspace is not used up, you can allocate more quotas to the current workspace.

Step 1 Log in to the DataArts Studio console.

Step 2 On the **Workspaces** page, locate the target workspace and click **Edit** in the **Operation** column.

Figure 10-1 Workspace Information dialog box



Workspace Information

* Name	default
Description	Enter a description. 0/4,096
* Mode	Simple Upgrade
* Enterprise Project	default
Job Log Path	Select
API Quota of DataArts DataService Exclusive	Used: 9 Allocated: 10 Save Total used: 9 Total allocated: 10 Total: 6,000

* Workspace Members Add Remove Account

Account	Account ...	Role	Added	Operation
[redacted]	User	admin	Feb 20, 2024 16:07:24 GMT+0...	Edit
[redacted]	User	admin	Jan 27, 2024 16:33:00 GMT+0...	Edit
[redacted]	User	admin	Jan 25, 2024 19:41:42 GMT+0...	Edit
[redacted]	User	admin	Jan 18, 2024 14:47:06 GMT+0...	Edit

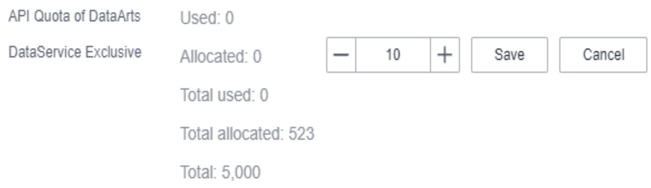
Step 3 Locate **API Quota of DataArts DataService Exclusive** and click **Edit** in the **Operation** column to set it. Click **OK** to save the change.

The allocated quota indicates the quota that can be used in the current workspace. It cannot be less than the used quota or greater than the unallocated quota (total quota minus total allocated quota).

NOTE

You can create 10 DataArts DataService Exclusive APIs for free in each DataArts Studio instance, and you will be charged for each extra API.

Figure 10-2 Setting the allocated quota



API Quota of DataArts DataService Exclusive

Used: 0

Allocated: 0 — 10 + Save Cancel

Total used: 0

Total allocated: 523

Total: 5,000

Step 4 In the **Workspace Information** dialog box, click **OK**.

----End

10.7 How Can I Access APIs of DataArts DataService Exclusive from the Internet?

The APIs can be accessed from the Internet only if the DataArts DataService Exclusive cluster can be accessed from the Internet.

To enable access to the DataArts DataService Exclusive cluster from the Internet, select **Enable public Access** when creating the cluster. After a DataArts DataService Exclusive cluster is created, you cannot bind an EIP to it if **Enable public Access** was not selected during the cluster creation, that is, the cluster cannot be accessed from the Internet.

In this case, you can export APIs of the current cluster, create another DataArts DataService Exclusive cluster and select **Enable public Access**, and import APIs from the old cluster to the new cluster to enable public network access.

In addition, if you already have an APIG dedicated instance or ROMA Connect instance for which Internet access is enabled, you can publish APIs to the APIG dedicated instance or ROMA Connect instance to provide a public access stream.

10.8 How Can I Access APIs of DataArts DataService Exclusive Using Domain Names?

The APIs can be accessed using domain names which are bound to the DataArts DataService Exclusive cluster.

- Binding a private domain name: A private domain name takes effect in a VPC. When a private domain name is bound, it is associated with a private IP address. Then you can call APIs using the private domain name in the same VPC in the private network.

  On the **Clusters** page, locate a cluster, click **More** in the **Operation** column, select **Bind Private Zone**, and enter a custom private domain name. DataArts DataService invokes the DNS service to associate the private domain name with the private IP address. Each tenant can add up to 50 private domain names in all projects.

  The private domain name supports various domain name levels and must comply with domain name naming rules.

  – Domain name labels are separated by dot (.), and each label does not exceed 63 characters.
  – A domain name label can contain letters, digits, and hyphens (-) and cannot start or end with a hyphen.
  – The total length of the domain name cannot exceed 254 characters.

- Binding a public domain name: A public domain name is resolved on the Internet. When a public domain name is bound, it is associated with a public IP address. Then you can call APIs using the public domain name on the Internet. On the **Clusters** page, locate a cluster, click **More** in the **Operation** column, select **Bind Public Zone**, and enter a registered domain name. DataArts DataService invokes the DNS service to associate the public domain name with the public IP address. To bind a public domain name, ensure that **Public Address** has been enabled during cluster creation and an EIP has been bound to the cluster. Otherwise, the public domain name cannot be bound to the cluster. In addition, each tenant can have up to 50 public domain names.

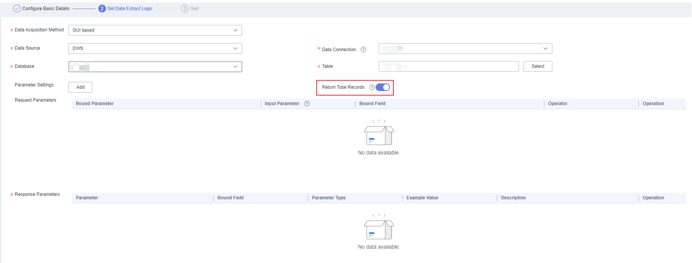
  The public domain name can include a primary domain name and its subdomain name, for example, abc.example.com.

10.9 What Should I Do If It Takes a Long Time to Obtain the Total Number of Data Records of a Table Through an API If the Table Contains a Large Amount of Data?

Scenario

During API creation, **Return Total Records** is enabled. If a table contains a large amount of data, obtaining the total number of data records of the table through an API is time-consuming.

Figure 10-3 Return Total Records



* Data Acquisition Method: GUI based

* Data Source: DWS

* Data Connection: [redacted]

* Database: [redacted]

* Table: [redacted]

Parameter Settings

Return Total Records: enabled

Request Parameters

Bound Parameter	Input Parameter	Bound Field	Operator	Operation

* Response Parameters

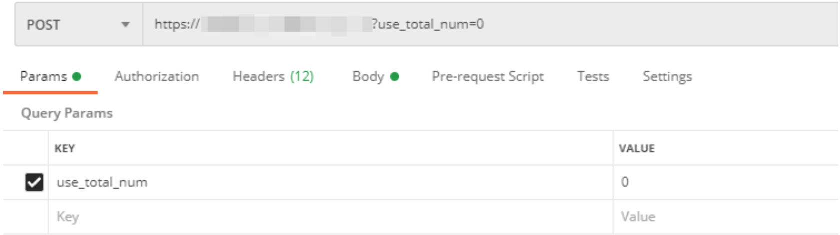
Parameter	Bound Field	Parameter Type	Example Value	Description	Operation

Solution

During a pagination query, you can use the **use_total_num** parameter to determine whether the total number of data records is calculated and returned.

For example,, add parameter **use_total_num** and set its value to **1** to obtain the total number of data records. In subsequent requests, add parameter **use_total_num** and set its value to **0** so that the total number of data records is not returned.

Figure 10-4 Total number of data records



POST https://[redacted]?use_total_num=0

Query Params

	KEY	VALUE
☑	use_total_num	0
	Key	Value