

DataArts Studio

FAQs

Issue 01
Date 2022-09-30



Copyright © Huawei Technologies Co., Ltd. 2022. 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 Technologies Co., Ltd.

Trademarks and Permissions



HUAWEI and other Huawei trademarks are trademarks 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 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
1.1 Regions and AZs.....	1
1.2 Can DataArts Studio Be Deployed in a Local Data Center or on a Private Cloud?.....	3
1.3 What Should I Do If a User Cannot View Existing Workspaces After I Have Assigned the Required Policy to the User?.....	3
1.4 Can I Delete DataArts Studio Workspaces?.....	4
1.5 Can I Transfer a Purchased or Trial Instance to Another Account?.....	4
1.6 Does DataArts Studio Support Version Upgrade?.....	4
1.7 Does DataArts Studio Support Version Downgrade?.....	4
1.8 How Do I View the DataArts Studio Instance Version?.....	4
2 Management Center	6
2.1 What Are the Precautions for Creating Data Connections?.....	6
2.2 Why Do DWS/Hive/HBase Data Connections Fail to Obtain the Information About Database or Tables?.....	6
2.3 Why Are MRS Hive/HBase Clusters Not Displayed on the Page for Creating Data Connections?.....	7
2.4 What Should I Do If the Connection Test Fails When I Enable the SSL Connection During the Creation of a DWS Data Connection?.....	7
2.5 Can I Create Multiple Data Connections in a Workspace in Proxy Mode?.....	7
2.6 Should I Choose a Direct or a Proxy Connection When Creating a DWS Connection?.....	8
2.7 How Do I Migrate the Data Development Jobs and Data Connections from One Workspace to Another?.....	8
2.8 Can I Delete Workspaces?.....	8
3 DataArts Migration	9
3.1 General.....	9
3.1.1 What Are the Advantages of CDM?.....	9
3.1.2 What Are the Security Protection Mechanisms of CDM?.....	10
3.1.3 How Do I Reduce the Cost of Using CDM?.....	11
3.1.4 Can I Upgrade a CDM Cluster?.....	12
3.1.5 How Is the Migration Performance of CDM?.....	12
3.1.6 What Is the Number of Concurrent Jobs for Different CDM Cluster Versions?.....	12
3.2 Functions.....	14
3.2.1 Does CDM Support Incremental Data Migration?.....	14
3.2.2 Does CDM Support Field Conversion?.....	14

3.2.3 What Component Versions Are Recommended for Migrating Hadoop Data Sources?.....	22
3.2.4 What Data Formats Are Supported When the Data Source Is Hive?.....	23
3.2.5 Can I Synchronize Jobs to Other Clusters?.....	23
3.2.6 Can I Create Jobs in Batches?.....	23
3.2.7 Can I Schedule Jobs in Batches?.....	23
3.2.8 How Do I Back Up CDM Jobs?.....	24
3.2.9 How Do I Configure the Connection If Only Some Nodes in the HANA Cluster Can Communicate with the CDM Cluster?.....	24
3.2.10 How Do I Use Java to Invoke CDM RESTful APIs to Create Data Migration Jobs?.....	24
3.2.11 How Do I Connect the On-Premises Intranet or Third-Party Private Network to CDM?.....	30
3.2.12 How Do I Set the Number of Concurrent Extractors for a CDM Migration Job?.....	32
3.2.13 Does CDM Support Real-Time Migration of Dynamic Data?.....	34
3.2.14 How Do I Obtain the Current Time Using an Expression?.....	34
3.3 Troubleshooting.....	34
3.3.1 What Can I Do If Error Message "Unable to execute the SQL statement" Is Displayed When I Import Data from OBS to SQL Server?.....	34
3.3.2 What Should I Do If the MongoDB Connection Migration Fails?.....	35
3.3.3 What Should I Do If a Hive Migration Job Is Suspended for a Long Period of Time?.....	35
3.3.4 What Should I Do If an Error Is Reported Because the Field Type Mapping Does Not Match During Data Migration Using CDM?.....	35
3.3.5 What Should I Do If a JDBC Connection Timeout Error Is Reported During MySQL Migration?.....	36
3.3.6 What Should I Do If a CDM Migration Job Fails After a Link from Hive to DWS Is Created?.....	37
3.3.7 How Do I Use CDM to Export MySQL Data to an SQL File and Upload the File to an OBS Bucket?	37
3.3.8 What Should I Do If CDM Fails to Migrate Data from OBS to DLI?.....	38
3.3.9 What Should I Do If a CDM Connector Reports the Error "Configuration Item [linkConfig.iamAuth] Does Not Exist"?.....	38
3.3.10 What Should I Do If Error Message "Configuration Item [linkConfig.createBackendLinks] Does Not Exist" Is Displayed During Data Link Creation or Error Message "Configuration Item [throttlingConfig.concurrentSubJobs] Does Not Exist" Is Displayed During Job Creation?.....	38
3.3.11 What Should I Do If Message "CORE_0031:Connect time out. (Cdm.0523)" Is Displayed During the Creation of an MRS Hive Link?.....	38
3.3.12 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?.....	39
3.3.13 What Should I Do If I Cannot Obtain the Schema Name When Creating an Oracle Relational Database Migration Job?.....	39
3.3.14 What Should I Do If invalid input syntax for integer: "true" Is Displayed During MySQL Database Migration?.....	39
4 DataArts Architecture.....	41
4.1 What Is the Relationship Between Lookup Tables and Data Standards?.....	41
4.2 What Is the Difference Between ER Modeling and Dimensional Modeling?.....	41
4.3 What Data Modeling Methods Are Supported by DataArts Architecture?.....	41
4.4 How Can I Use Standardized Data?.....	42
4.5 Does DataArts Architecture Support Database Reverse?.....	42
4.6 What Are the Differences Between the Metrics in DataArts Architecture and DataArts Quality?.....	42

4.7 Why Does a Table Remain Unchanged When I Have Updated It in DataArts Architecture?..... 43

4.8 Can I Configure Lifecycle Management for Tables?..... 43

5 DataArts Factory.....44

5.1 How Many Jobs Can Be Created in DataArts Factory? Is There a Limit on the Number of Nodes in a Job?..... 44

5.2 Why Is There a Large Difference Between Job Execution Time and Start Time of a Job?.....44

5.3 Will Subsequent Jobs Be Affected If a Job Fails to Be Executed During Scheduling of Dependent Jobs? What Should I Do?..... 45

5.4 What Should I Pay Attention to When Using DataArts Studio to Schedule Big Data Services?.....45

5.5 What Are the Differences and Connections Among Environment Variables, Job Parameters, and Script Parameters?..... 45

5.6 What Do I Do If Node Error Logs Cannot Be Viewed When a Job Fails?.....47

5.7 What Should I Do If the Agency List Fails to Be Obtained During Agency Configuration?..... 48

5.8 How Do I Locate Job Scheduling Nodes with a Large Number?..... 49

5.9 Why Cannot Specified Peripheral Resources Be Selected When a Data Connection Is Created in Data Development?..... 49

5.10 Why Is There No Job Running Scheduling Log on the Monitor Instance Page After Periodic Scheduling Is Configured for a Job?..... 50

5.11 Why Does the GUI Display Only the Failure Result but Not the Specific Error Cause After Hive SQL and Spark SQL Scripts Fail to Be Executed?..... 50

5.12 What Do I Do If the Token Is Invalid During the Running of a Data Development Node?..... 50

5.13 How Do I View Run Logs After a Job Is Tested?.....51

5.14 Why Does a Job Scheduled by Month Start Running Before the Job Scheduled by Day Is Complete? 51

5.15 What Should I Do If Invalid Authentication Is Reported When I Run a DLI Script?..... 51

5.16 Why Cannot I Select the Desired CDM Cluster in Proxy Mode When Creating a Data Connection?... 52

5.17 Why Is There No Job Running Scheduling Record After Daily Scheduling Is Configured for the Job? 52

5.18 What Do I Do If No Content Is Displayed in Job Logs?..... 52

5.19 Why Do I Fail to Establish a Dependency Between Two Jobs?..... 53

5.20 What Should I Do If an Error Is Displayed During DataArts Studio Scheduling: The Job Does Not Have a Submitted Version?..... 53

5.21 What Do I Do If an Error Is Displayed During DataArts Studio Scheduling: The Script Associated with Node XXX in the Job Is Not Submitted?..... 54

5.22 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?..... 54

5.23 How Do I Create a Database And Data Table? Is the database a data connection?..... 54

5.24 Why Is No Result Displayed After an HIVE Task Is Executed?..... 55

5.25 Why Does the Last Instance Status On the Monitor Instance page Only Display Succeeded or Failed? 55

5.26 How Do I Create a Notification for All Jobs?.....55

5.27 What Is the Maximum Number of Nodes That Can Be Executed Simultaneously?..... 55

5.28 What Is the Priority of the Startup User, Execution User, Workspace Agency, and Job Agency?.....56

6 DataArts Quality.....57

6.1 What Are the Differences Between Quality Jobs and Comparison Jobs?..... 57

6.2 How Can I Confirm that a Quality Job or Comparison Job Is Blocked?.....	57
6.3 How Do I Manually Restart a Blocked Quality Job or Comparison Job?.....	57
6.4 How Do I View Jobs Associated with a Quality Rule Template?.....	58
6.5 What Should I Do If the System Displays a Message Indicating that I Do Not Have the MRS Permission to Perform a Quality Job?.....	58
7 DataArts Catalog.....	62
7.1 What Are the Functions of the DataArts Catalog Module?.....	62
7.2 What Assets Can Be Collected by DataArts Catalog?.....	62
7.3 What Is Data Lineage?.....	62
7.4 How Do I Visualize Data Lineages in a Data Catalog?.....	63
8 DataArts DataService.....	64
8.1 What Languages Do Data Lake Mall SDKs Support?.....	64
8.2 What Can I Do If the System Displays a Message Indicating that the Proxy Fails to Be Invoked During API Creation?.....	64
8.3 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?.....	64
8.4 What Can I Do If an Error Is Reported When I Use an API?.....	64
8.5 Can Operators Be Transferred When API Parameters Are Transferred?.....	65
8.6 What Should I Do If the API Quota Provided by DataArts DataService Exclusive Has Been Used up?	65

1 Consultation and Billing

1.1 Regions and AZs

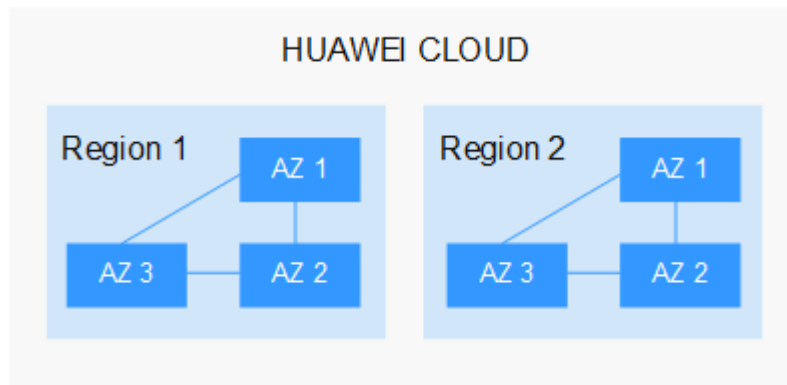
Concepts

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



Region Selection

When selecting a region, consider the following factors:

- Location

You are advised to select a region closest to your target users. This reduces network latency and improves access rate. However, Chinese mainland regions provide the same infrastructure, BGP network quality, and operations and configurations on resources. Therefore, if your target users are in the Chinese mainland, you do not need to consider the network latency differences when selecting a region.

The countries and regions outside the Chinese mainland, such as Bangkok, provide services for users outside the Chinese mainland. If your target users are in the Chinese mainland, these regions are not recommended because there may be a latency in accessing resources.

- 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.

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 a DataArts Studio instance or incremental package for the first time, there is no requirement on the AZ. When you buy a new DataArts Studio instance or incremental package, determine whether to select the same AZ as the existing one 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.
- 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 endpoints from [\(Optional\) Obtaining Authentication Information](#).

1.2 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.

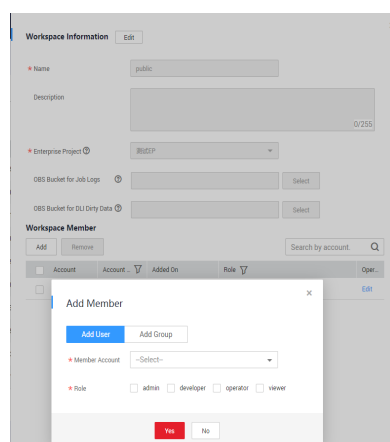
1.3 What Should I Do If a User Cannot View Existing Workspaces After I Have Assigned the Required Policy to the User?

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

Adding a Member and Assigning a Role

1. Log in to the DataArts Studio console and access the **Workspaces** page.
2. On the **Workspaces** page, locate the target workspace and click **Edit** in the **Operation** column.
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 1-2 Adding a member



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

1.4 Can I Delete DataArts Studio Workspaces?

After workspaces are created, they cannot be deleted. You can disable workspaces when they are no longer needed. You can enable them again when you need these workspaces.

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

No. the purchased or trial instance cannot be transferred to another account.

To add another account, see [Authorizing Users to Use DataArts Studio](#).

1.6 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?](#)

1.7 Does DataArts Studio Support Version Downgrade?

No. You cannot downgrade a purchased DataArts Studio instance.

1.8 How Do I View the DataArts Studio Instance Version?

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

Figure 1-3 DataArts Studio instance card

Enterprise Project: test

Version	Enterprise	Billing Mode	Yearly/Monthly
Created	Feb 10, 2020 10:38:40 GMT+08:00	Name	DAYU-DLG-test
Expired	Feb 09, 2023 23:59:59 GMT+08:00	Status	Valid

[Access](#) | [Change](#) | [Renew](#) | [Buy](#)

2 Management Center

2.1 What Are the Precautions for Creating Data Connections?

When creating a DWS, MRS Hive, RDS, and SparkSQL data connection, you must bind an agent provided by the CDM cluster. Currently, a version of the CDM cluster earlier than 1.8.6 is not supported.

2.2 Why Do DWS/Hive/HBase Data Connections Fail to Obtain the Information About Database or Tables?

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.

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

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

Possible causes are as follows:

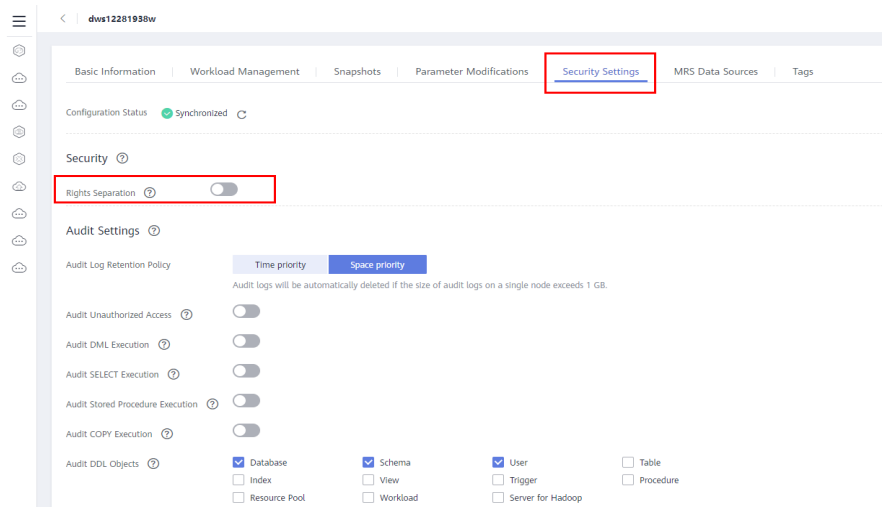
- Hive/HBase components were not selected during MRS cluster creation.
- The network between the CDM cluster and MRS cluster was disconnected when an MRS data connection is created.

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

2.4 What Should I Do If the Connection Test Fails When I Enable the SSL Connection During the Creation of a DWS Data Connection?

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



2.5 Can I Create Multiple Data Connections in a Workspace in Proxy Mode?

Multiple data connections of the same type or different types can be created in the same workspace, but their names must be unique.

2.6 Should I Choose a Direct or a Proxy Connection When Creating a DWS Connection?

You are advised to choose a proxy connection.

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

You can export the jobs in DataArts Factory and then import them to DataArts Factory in another workspace.

You can export data connections on the **Migrate Resources** page of Manager Center and then import them on the **Migrate Resources** page in another workspace.

2.8 Can I Delete Workspaces?

No, but you can change the names of workspaces.

3 DataArts Migration

3.1 General

3.1.1 What Are the Advantages of CDM?

CDM is developed based on a distributed computing framework and leverages the parallel data processing technology. [Table 3-1](#) details the advantages of CDM.

Table 3-1 CDM advantages

Item	User-Developed Script	CDM
Ease of use	<p>You need to prepare server resources, and install and configure software, which is time-consuming.</p> <p>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.</p>	<p>CDM provides a web-based management console for enabling services on web pages in real time.</p> <p>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.</p> <p>CDM also provides RESTful APIs to support third-party system calling and integration.</p>

Item	User-Developed Script	CDM
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.
High efficiency	During data migration, the read and write process is completed in one job. Limited by available resources, the performance is poor and 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.

3.1.2 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.

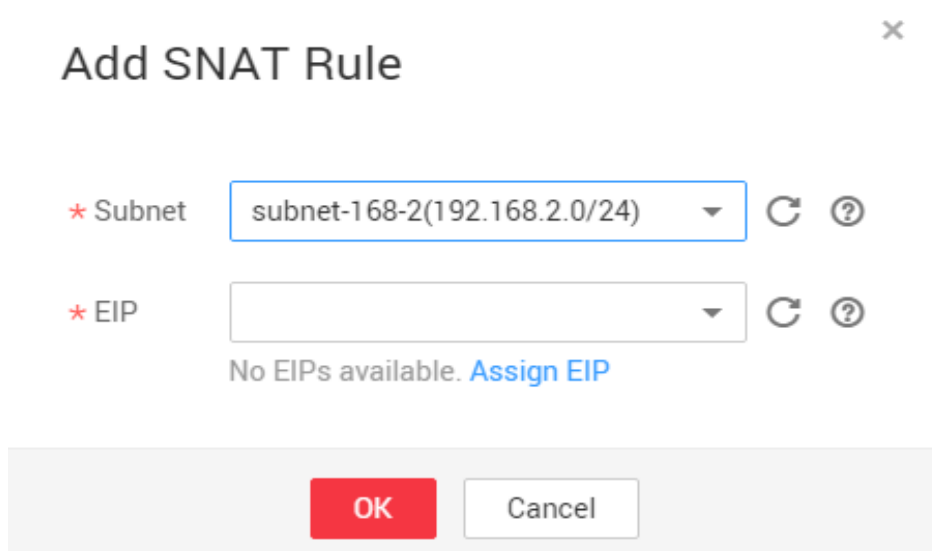
3.1.3 How Do I Reduce the Cost of Using CDM?

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



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.

 NOTE

If SSL encryption is configured for the access channel of a local data source, CDM cannot connect to the data source using the EIP.

3.1.4 Can I Upgrade a CDM Cluster?

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

3.1.5 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](#).

3.1.6 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. The maximum number of tasks (defined by **Maximum Concurrent Extractors**) run concurrently. Excess tasks are queued.

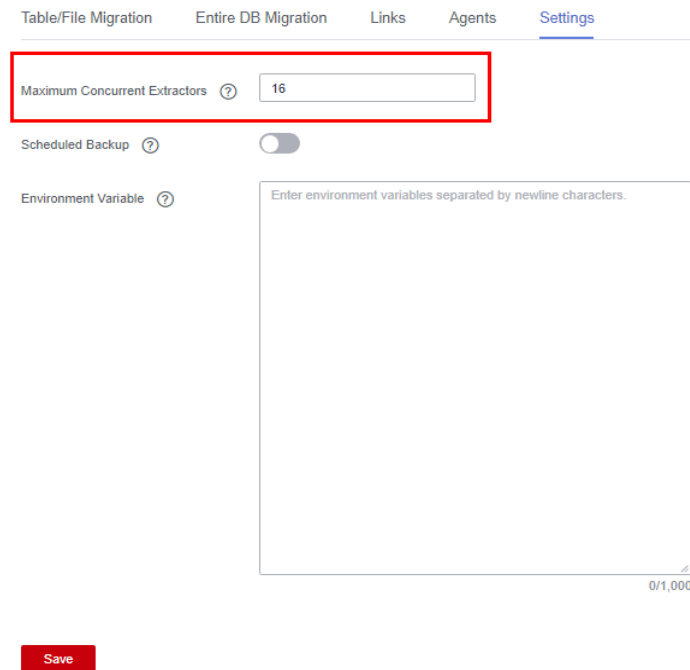
Changing Concurrent Extractors

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 3-2 Maximum number of concurrent extractors for a CDM cluster

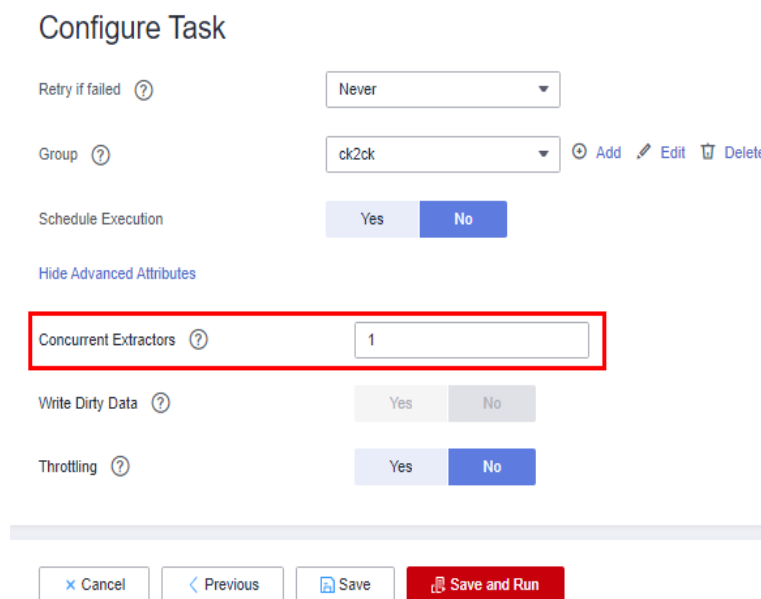
Flavor	vCPUs/Memory	Maximum Concurrent Extractors
<code>cdm.large</code>	8 vCPUs, 16 GB	16
<code>cdm.xlarge</code>	16 vCPUs, 32 GB	32
<code>cdm.4xlarge</code>	32 vCPUs, 64 GB	64

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



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**.

Figure 3-3 Setting Concurrent Extractors for a job



3.2 Functions

3.2.1 Does CDM Support Incremental Data Migration?

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

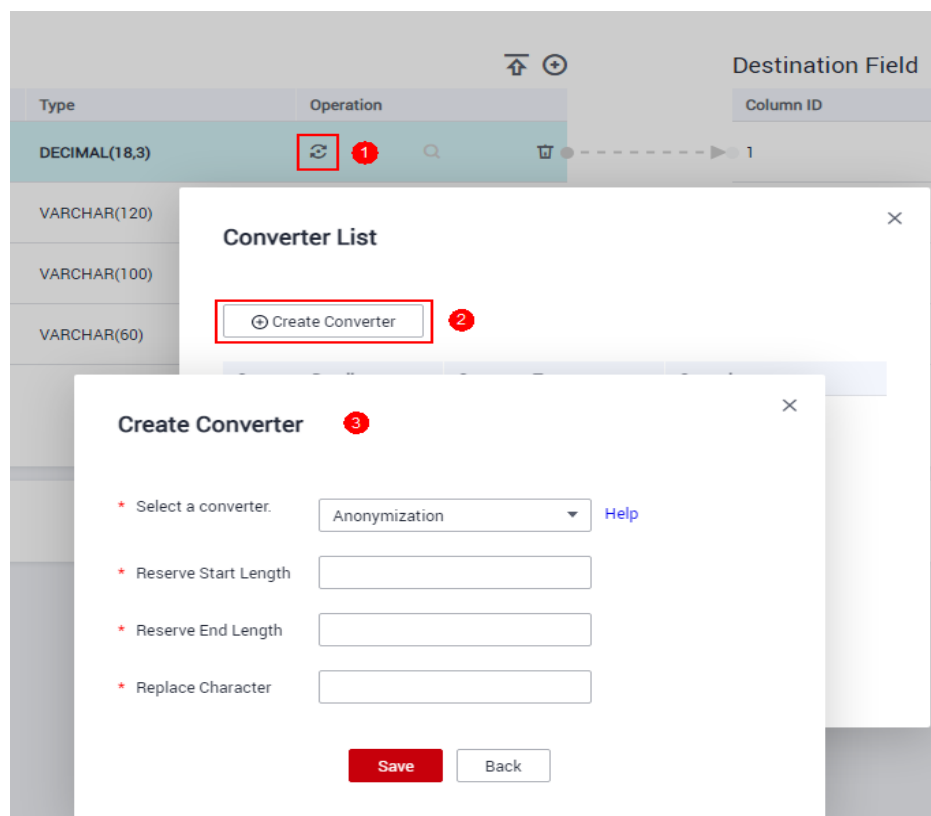
For details, see [Incremental Migration](#).

3.2.2 Does CDM Support Field Conversion?

Yes. CDM supports the following field converters:

- [Anonymization](#)
- [Trim](#)
- [Reverse String](#)
- [Replace String](#)
- [Expression Conversion](#)

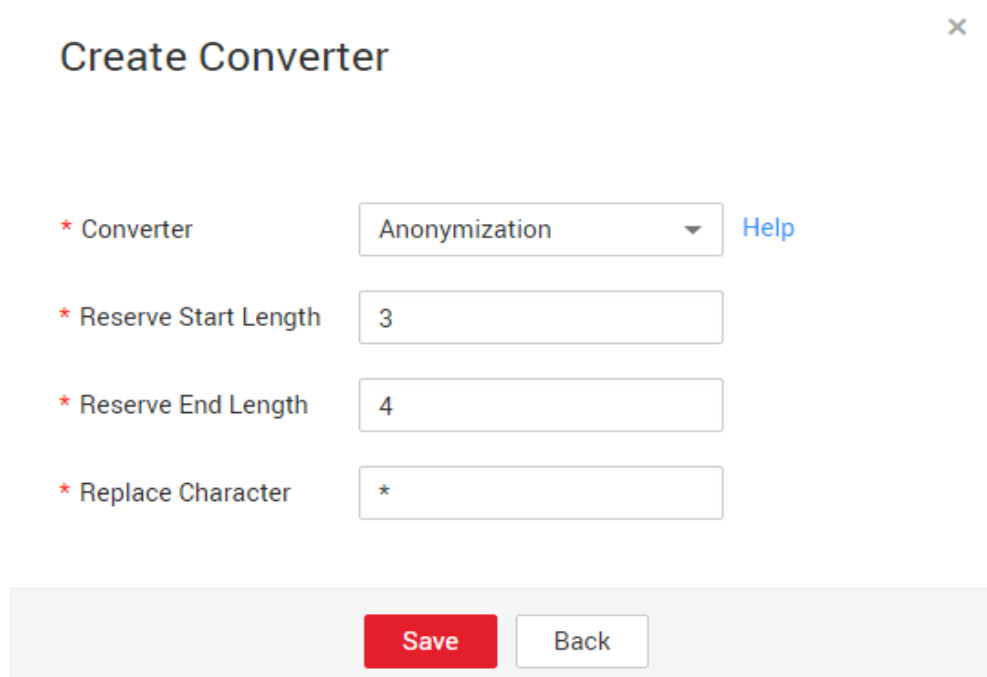
You can create a field converter on the **Map Field** page when creating a table/file migration job.

Figure 3-4 Creating a field converter

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 *****.

Figure 3-5 Anonymization

Create Converter ×

* Converter [Help](#)

* Reserve Start Length

* Reserve End Length

* Replace Character

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. Within a JSP EL expression, you can use integers, floating point numbers, strings, the built-in constants **true** and **false** for boolean values, and **null**.

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 tool classes:

- **StringUtils**: string processing tool class. For details, see **org.apache.commons.lang.StringUtils** of the Java SDK code.

- DateUtils: date tool class
- CommonUtils: common tool class
- NumberUtils: string-to-value conversion class
- HttpsUtils: network file read class

Application examples:

1. 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)`
2. Convert all character strings of the current field to uppercase letters.
Expression: `StringUtils.upperCase(value)`
3. 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,"-")`
4. 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`
5. Convert the field value **true** to **Y** and other field values to **N**.
Expression: `value=="true"?"Y":"N"`
6. 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`
7. 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")`
8. Obtain a 36-bit universally unique identifier (UUID):
Expression: `CommonUtils.randomUUID()`
9. If the field is of the string type, capitalize the first letter, for example, convert **cat** to **Cat**.
Expression: `StringUtils.capitalize(value)`
10. 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)`
11. 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)`
12. 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)`
13. 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")`

14. 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")`
15. 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")`
16. 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")`
17. 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)`
18. 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)`
19. 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")`
20. 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")`
21. 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")`
22. 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)`
23. 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")`
24. 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)`
25. 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)`

26. 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)`

27. 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)`

28. 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)`

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

Expression: `StringUtils.isEmpty(value)`

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

Expression: `StringUtils.isNumeric(value)`

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

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

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

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

33. 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 **zyzybat** after conversion.

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

34. 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 **batzyzy** after conversion.

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

35. 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)`

36. 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")`

37. 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")`

38. 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.")`

39. 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")`

40. 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")`

41. 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")`

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

Expression: `StringUtils.strip(value, "xyz")`

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

Expression: `StringUtils.stripEnd(value, null)`

44. 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)`

45. If the field is of the string type, obtain the substring after the specified position (excluding 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. For example, obtain the character string after the second character of **abcde**, that is, **cde**.

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

46. If the field is of the string type, obtain the substring within the specified range of the character string. If the specified range is a negative number, calculate the range in the descending order. For example, obtain the character string between the second and fifth characters of **abcde**, that is, **cd**.

Expression: `StringUtils.substring(value, 2, 5)`

47. 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")`

48. 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")`

49. 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")`
50. 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")`
51. 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")`
52. 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)`
53. Convert the character string to a value of the byte type. If the conversion fails, **0** is returned.
Expression: `NumberUtils.toByte(value)`
54. 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)`
55. Convert the character string to a value of the double type. If the conversion fails, **0.0d** is returned.
Expression: `NumberUtils.toDouble(value)`
56. 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)`
57. Convert the character string to a value of the float type. If the conversion fails, **0.0f** is returned.
Expression: `NumberUtils.toFloat(value)`
58. 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)`
59. Convert the character string to a value of the int type. If the conversion fails, **0** is returned.
Expression: `NumberUtils.toInt(value)`
60. 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)`
61. Convert the character string to a value of the long type. If the conversion fails, **0** is returned.
Expression: `NumberUtils.toLong(value)`
62. 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)`
63. Convert the character string to a value of the short type. If the conversion fails, **0** is returned.
- Expression: `NumberUtils.toShort(value)`
64. 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)`
65. Convert the IP string to a value of the long type, for example, convert **10.78.124.0** to **172915712**.
- Expression: `CommonUtils.ipToLong(value)`
66. 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")`
67. Cache the IP address and physical address mappings and specify a key for retrieval, for example, **ipList**.
- Expression: `CommonUtils.setCache("ipList", HttpsUtils.downloadMap("url"))`
68. Obtain the cached IP address and physical address mappings.
- Expression: `CommonUtils.getCache("ipList")`
69. Check whether the IP address and physical address mappings are cached.
- Expression: `CommonUtils.cacheExists("ipList")`
70. 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)`

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

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

Table 3-3 Recommended component versions

Hadoop Type	Component	Description
MRS/Apache/ FusionInsight HD	Hive	2.x versions are not supported. The following versions are recommended: <ul style="list-style-type: none"> • 1.2.X • 3.1.X

Hadoop Type	Component	Description
	HDFS	Recommended versions: <ul style="list-style-type: none">• 2.8.X• 3.1.X
	HBase	Recommended versions: <ul style="list-style-type: none">• 2.1.X• 1.3.X

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

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

3.2.5 Can I Synchronize Jobs to Other Clusters?

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.

3.2.6 Can I Create Jobs in Batches?

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](#).

3.2.7 Can I Schedule Jobs in Batches?

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.

3.2.8 How Do I Back Up CDM Jobs?

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.

3.2.9 How Do I Configure the Connection If Only Some Nodes in the HANA Cluster Can Communicate with the CDM Cluster?

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.

3.2.10 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 **dwestestlink**.
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";
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();
}
}
```

3.2.11 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 IPOPOP on the computer.
3. Configure port mapping using the tool.

NOTICE

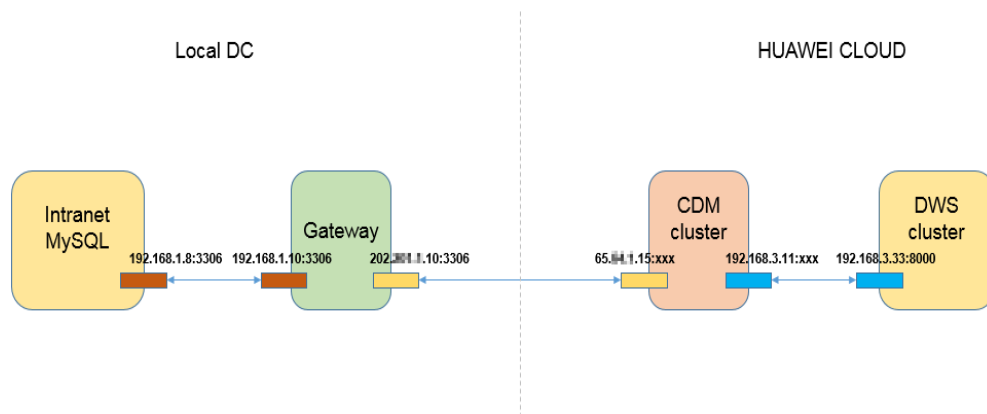
If the intranet database is exposed to the public network for a long time, security risks exist. Therefore, after data migration is complete, stop port mapping.

Scenario

Suppose that the MySQL database on the intranet is migrated to DWS. [Figure 3-6](#) 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.

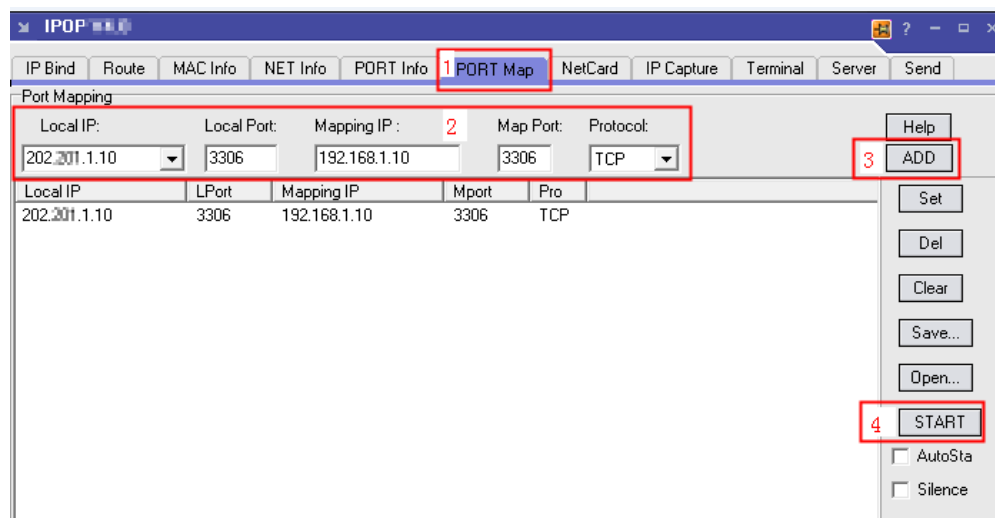
Figure 3-6 Network topology example



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 3-7](#).
- **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 3-7 Configuring port mapping



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

3.2.12 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. The maximum number of tasks (defined by **Maximum Concurrent Extractors**) run concurrently. Excess tasks are queued.

Changing Concurrent Extractors

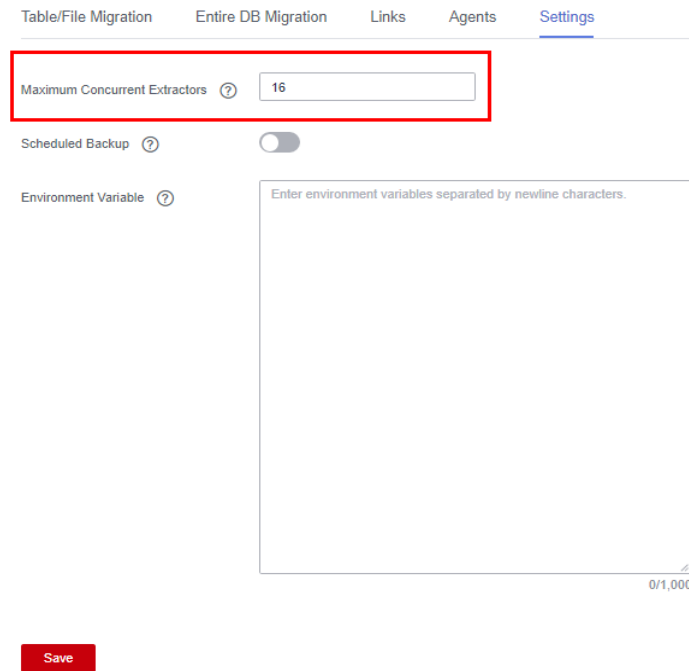
1. The maximum number of concurrent extractors for a cluster varies depending on the CDM cluster flavor. You are advised to set the maximum number of concurrent extractors to twice the number of vCPUs of the CDM cluster.

Table 3-4 Maximum number of concurrent extractors for a CDM cluster

Flavor	vCPUs/Memory	Maximum Concurrent Extractors
cdm.large	8 vCPUs, 16 GB	16

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

Figure 3-8 Setting Maximum Concurrent Extractors for a CDM cluster



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**.

Figure 3-9 Setting Concurrent Extractors for a job

Configure Task

Retry if failed ?

Group ? + Add ✎ Edit ✖ Delete

Schedule Execution Yes No

[Hide Advanced Attributes](#)

Concurrent Extractors ?

Write Dirty Data ? Yes No

Throttling ? Yes No

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

No. If data is written to the source during the migration, an error may occur.

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

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](#).

3.3 Troubleshooting

3.3.1 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.

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

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.

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

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:



3.3.4 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.

3.3.5 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. As shown in [Figure 3-10](#), the `c_date` field is not an index column.

Figure 3-10 Non-index column

The image shows two side-by-side configuration panels for a data migration job. The left panel is titled 'Source Job Configuration' and contains the following fields: 'Source Link Name' (mysql), 'Use SQL Statement' (Yes/No), 'Schema/Table Space' (SQOOP), 'Table Name' (rf_BaoWeiFu_test_sql_To), 'Where Clause' (c_date > '2021-02-27 10:43:04.123'), 'Partition Column' (empty), and 'Partition column nullable' (Yes/No). The right panel is titled 'Destination Job Configuration' and contains: 'Destination Link Name' (dli), 'Resource Queue' (dlf_notdelete), 'Database' (abcd), 'Table' (dddd), and 'Clear data before import' (Yes/No).

Solution

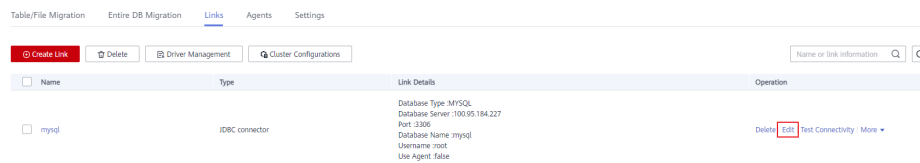
1. Contact the DBA to modify the table structure, set the columns to be filtered as index columns, and try again.
If the failure persists because the data is not discrete, perform [2](#) to [4](#) and increase the JDBC timeout duration.
2. Locate the MySQL link name based on the job and obtain the link information.

Figure 3-11 Link information



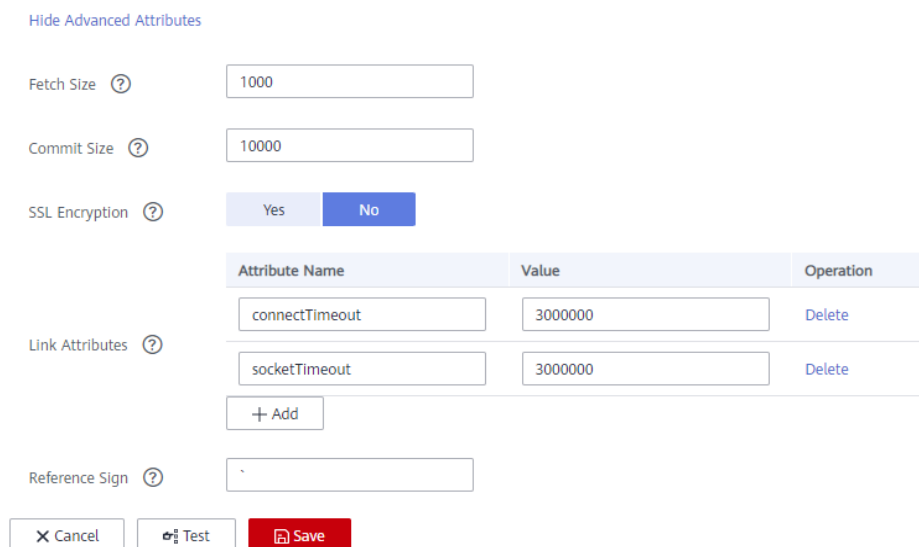
3. Click the **Links** tab and click **Edit** to edit the link.

Figure 3-12 Editing the link



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

Figure 3-13 Editing advanced attributes



3.3.6 What Should I Do If a CDM Migration Job Fails After a Link from Hive to DWS Is Created?

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.

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

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.

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

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

3.3.9 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.

3.3.10 What Should I Do If Error Message "Configuration Item [linkConfig.createBackendLinks] Does Not Exist" Is Displayed During Data Link Creation or Error Message "Configuration Item [throttlingConfig.concurrentSubJobs] Does Not Exist" Is Displayed During Job Creation?

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.

Manually clear the browser cache to avoid this error.

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

This failure occurs because you do not have the required permissions. Create another service user, grant the required permissions to it, 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.

3.3.12 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?

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 (_).

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

This may be because you have uploaded the latest ORACLE_8 driver (for example, Oracle Database 21c (21.3) driver), which is not supported yet. You are advised to use the ojdbc8.jar driver in Oracle Database 12c. You can download it from <https://www.oracle.com/database/technologies/jdbc-ucp-122-downloads.html>.

3.3.14 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 3-14 Adding link attributes

[Hide Advanced Attributes](#)

Fetch Size [?](#)

Commit Size [?](#)

Link Attributes [?](#)

Attribute Name	Value	Operation
<input type="text"/>	<input type="text"/>	Delete

0/512 0/512

4 DataArts Architecture

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

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.

4.2 What Is the Difference Between ER Modeling and Dimensional Modeling?

- 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.

4.3 What Data Modeling Methods Are Supported by DataArts Architecture?

DataArts Studio DataArts Architecture supports Entity Relationship (ER) modeling and dimensional modeling:

- **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.

- **Conceptual model:** A conceptual model is a representation of business processes and business data involved in different activities. It can be used to represent the relationships between business entities.
- **Logical model:** A logical model is more detailed than a conceptual model. It is used to outline the entities, attributes, and relationships of a business. It enables communication between IT and business staff. A logical model is a set of standardized logic table structures. Determined by business rules, a logical model outlines business objects, data items of the business objects, and relationships between business objects.
- **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.
- **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.

4.4 How Can I Use Standardized Data?

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

4.5 Does DataArts Architecture Support Database Reverse?

Yes. Currently, database reverse can be performed on Data Warehouse Service (DWS), Data Lake Insight (DLI), and MapReduce Service (MRS Hive).

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

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.

4.7 Why Does a Table Remain Unchanged When I Have Updated It in DataArts Architecture?

This is because you did not configure the table update mode before updating the table. To configure the table update mode, perform the following steps:

1. Access the Data Architecture console and choose **Configuration Center** in the left navigation pane.
2. Click **Functions**.
3. Set **Data Table Update Mode** to **Drop and create**.
4. Click **OK**.

4.8 Can I Configure Lifecycle Management for Tables?

No. This function is unavailable now.

5 DataArts Factory

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

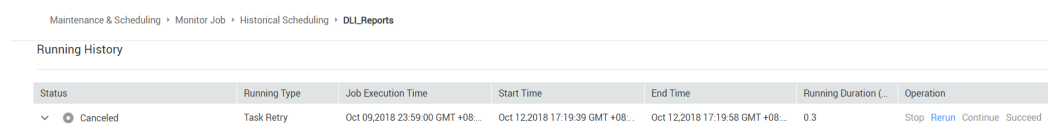
By default, each user can create a maximum of 10,000 jobs, and each job can contain a maximum of 200 nodes.

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

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

On the **Running History** page, there is a large difference between **Job Execution Time** and **Start Time**, as shown in the figure below. **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.

Figure 5-1 Running History page



Status	Running Type	Job Execution Time	Start Time	End Time	Running Duration (...)	Operation
▼ Canceled	Task Retry	Oct 09,2018 23:59:00 GMT +08:...	Oct 12,2018 17:19:39 GMT +08:...	Oct 12,2018 17:19:58 GMT +08:...	0:3	Stop Rerun Continue Succeed

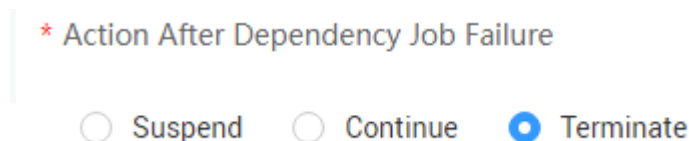
In Data Development, a maximum of five instances can be concurrently executed in a job. If **Start Time** of a job is later than **Job Execution Time**, the job instances in the subsequent batch will be queued.

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

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

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

Figure 5-2 Job dependencies



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.

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

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.

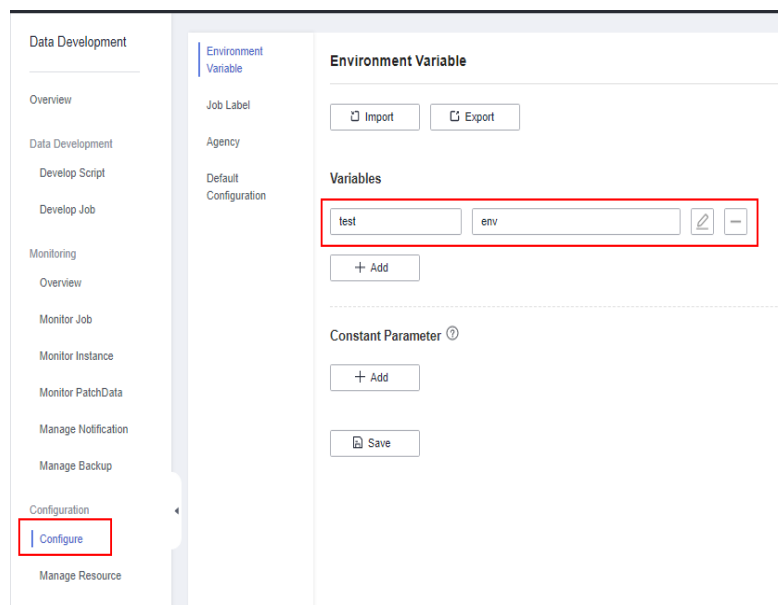
5.5 What Are the Differences and Connections Among Environment Variables, Job Parameters, and Script Parameters?

Parameters can be set in environment variables, job parameters, and script parameters, but their application scopes are different. If there is a conflict when parameters in environment variables, job parameters, and script parameters of the same name, the calling priority is: **job parameters > environment variables > script parameters**.

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

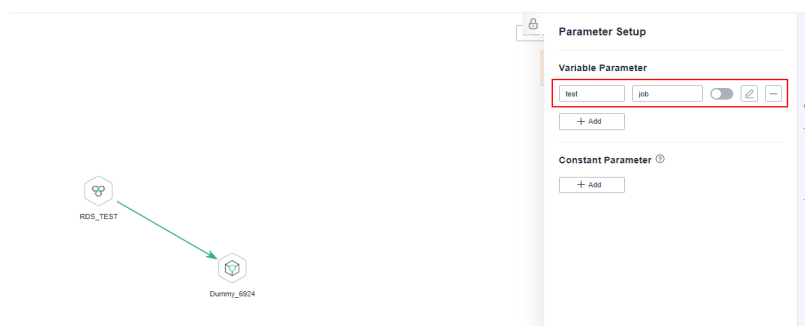
- Variables and constants can be defined in environment variables. Environment variables take effect in current workspace.
 - The value of a variable (such as **workspace name**) varies depending on the workspace. When exporting a variable from a workspace and import it to another workspace, you must 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 5-3 Environment variable



- 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 5-4 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 5-5](#). If the script is executed by job scheduling, you can assign values to the parameters based on node attributes, as shown in [Figure 5-6](#).
 - For Shell scripts, you can enter a parameter and an interactive parameter in the upper part of the editor to transfer the parameters.
 - Python scripts do not support parameter transfer.

Figure 5-5 Configuring script parameters when the script is executed independently

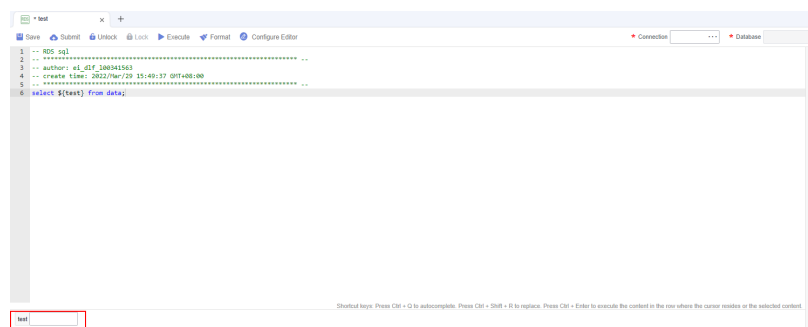
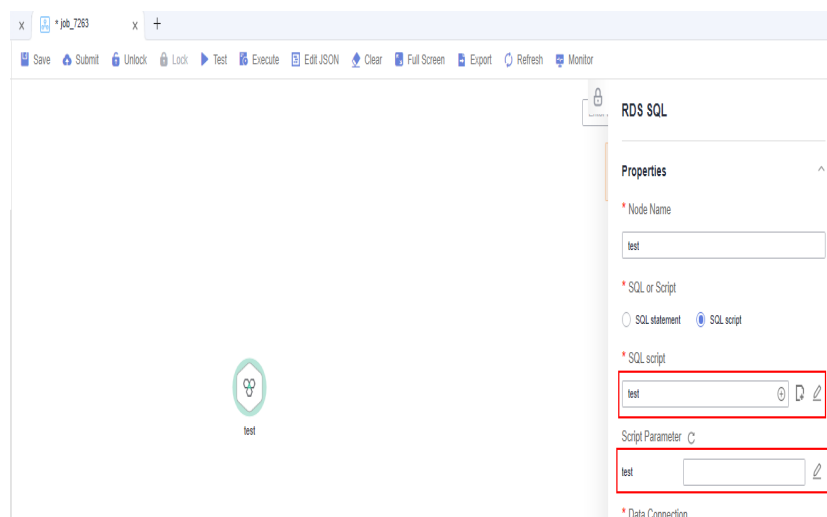


Figure 5-6 Configuring script parameters when the script is executed by job scheduling



5.6 What Do I Do If Node Error Logs Cannot Be Viewed When a Job Fails?

Error logs are stored in OBS. The current account must have the OBS read permissions to view logs. You can check the OBS permissions and OBS bucket policies in IAM.

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.

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

When a workspace- or job-level agency is configured, the following error is reported when the agency list is viewed:

Policy doesn't allow iam:agencies:listAgencies to be performed.

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

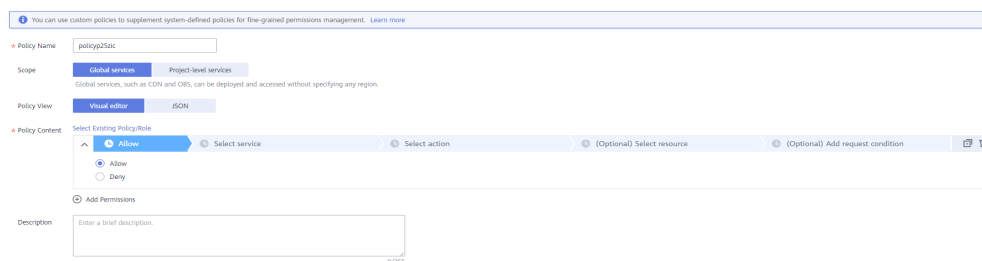
You can create a custom policy (query the agency list based on specified conditions) and assign it to a user group for refined access control.

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**. Then, click **Create Custom Policy**.

Step 4 Enter a policy name.



Step 5 Set **Scope** to **Global services**. The scope you set is where the custom policy takes effect. In this example, the custom policy has the permissions required to view the agency lists based on specified conditions.

Step 6 Set **Policy View** to **Visual editor**.

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](#).

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

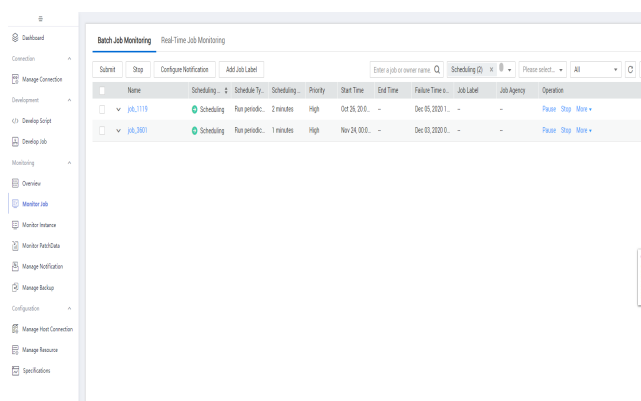
----End

5.8 How Do I Locate Job Scheduling Nodes with a Large Number?

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

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 5-7 Viewing the scheduling period



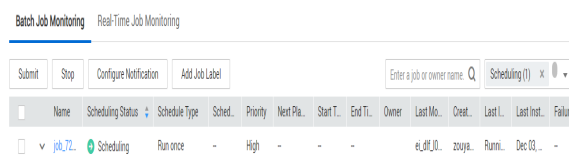
5.9 Why Cannot Specified Peripheral Resources Be Selected When a Data Connection Is Created in Data Development?

Ensure that the current 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.

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

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

Figure 5-8 Viewing the job scheduling status



Name	Scheduling Status	Schedule Type	Sched.	Priority	Next Pla.	Start T.	End T.	Owner	Last Mo.	Creat.	Last L.	Last Inst.	Failure
job_72	Scheduling	Runonce	-	High	-	-	-	ei_08_00_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.

5.11 Why Does the GUI Display Only the Failure Result but Not the Specific Error Cause After Hive SQL and Spark SQL Scripts Fail to Be Executed?

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

In direct connection mode, DataArts Studio users submit the 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.

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

5.12 What Do I Do If the Token Is Invalid During the Running of a Data Development Node?

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.

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

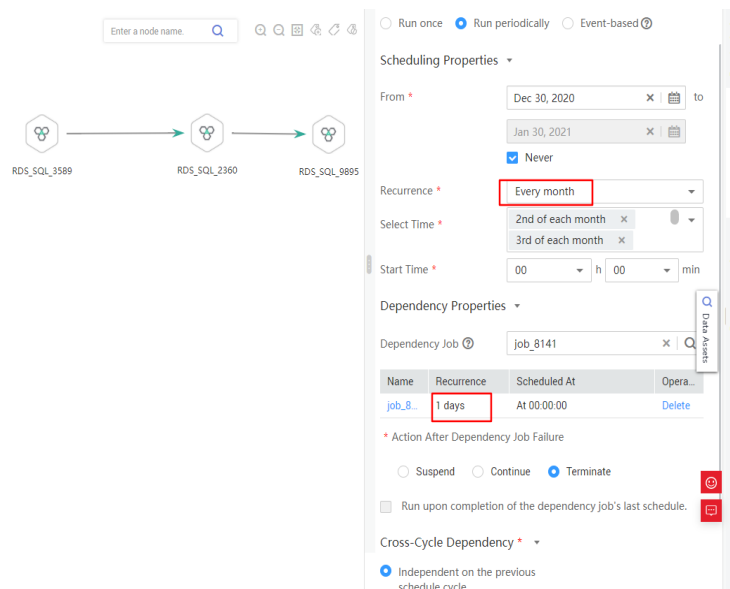
Method 1: After the node test is complete, right-click the current node and choose **View Log** from the shortcut menu.

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

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

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 5-9 Viewing the job scheduling period and dependency attributes



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.

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

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

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

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

5.17 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.

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

Symptom

There is no content contained in the job log.

Cause Analysis

Check whether the user has the global permission of the object storage service (OBS) in IAM to ensure that the user 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.

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

5.19 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.

5.20 What Should I Do If an Error Is Displayed During DataArts Studio Scheduling: The Job Does Not Have a Submitted Version?

Symptom

An error is reported when DataArts Studio executes scheduling: The job does not have a submitted version. Submit the job version first.

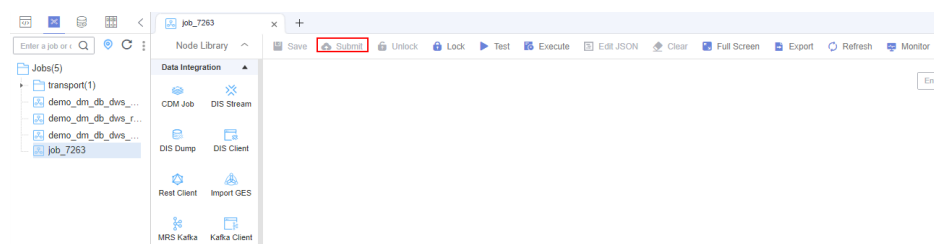
Cause Analysis

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 5-10 Submitting a version



5.21 What Do I Do If an Error Is Displayed During DataArts Studio Scheduling: The Script Associated with Node XXX in the Job Is Not 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.

5.22 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.

5.23 How Do I Create a Database And Data Table? Is the database a data connection?

Databases and data tables can be created in DLI.

A database does not correspond to a data connection. A data connection is a connection channel for creating DataArts Studio and other data services.

5.24 Why Is No Result Displayed After an HIVE Task Is Executed?

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

5.25 Why Does the Last Instance Status On the Monitor Instance page Only Display Succeeded or Failed?

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.

5.26 How Do I Create a Notification for All Jobs?

1. Choose **Monitoring > Monitor Job** and click the **Batch Job Monitoring** tab.
2. Select the jobs to be configured and click **Configure Notification**.

Figure 5-11 Creating a notification

Configure Notification ×

Notification Type Run abnormally Run successfully Uncompleted
 Busy resources

Topic Name [View Topic](#)
SMN will be charged based on standard pricing. [Pricing details](#)

Notification

OK Cancel

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

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

The following table lists the number of nodes that can be executed concurrently in each DataArts Studio version.

Table 5-1 Number of nodes that can be executed concurrently in each DataArts Studio version

Version	Number of Nodes Executed per Day	Number of Nodes Executed Concurrently
Starter	5,000	50
Basic	20,000	100
Advanced	40,000	200
Professional	80,000	300
Enterprise	200,000	400

5.28 What Is the Priority of the Startup User, Execution User, Workspace Agency, and Job Agency?

The system obtains permissions for the job agency, workspace agency, and execution user in sequence, and then executes jobs with the permissions.

By default, a job is executed by the user who starts the job. If a job is started by a user with low permissions, the job will fail to be executed due to insufficient permissions. To resolve this issue, you can configure an agency or an execution user.

- After an agency is configured for a job, the job interacts with other services through the agency, preventing job execution failures caused by permission issues. There are two types of agencies, workspace agencies and job agencies. Workspace agencies have a higher priority than job agencies.
 - Workspace agency: applies to all the jobs in a workspace. You can choose **Configuration > Configure > Agency** to configure a workspace agency.
 - Job agency: applies to a single job. You can configure a job agency in the basic information of a job.
- After an execution user is configured, the user will be used to start the job. You can configure an execution user in the basic information of a job.

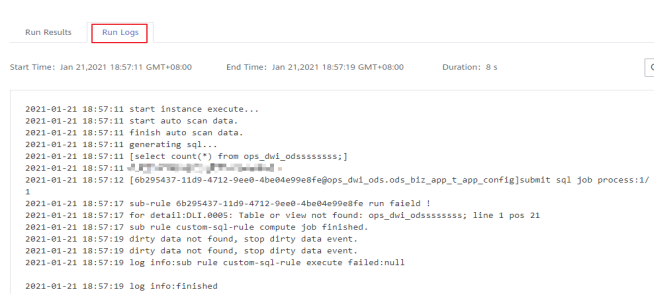
6 DataArts Quality

6.1 What Are the Differences Between Quality Jobs and Comparison Jobs?

- 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.

6.2 How Can I Confirm that a Quality Job or Comparison Job Is Blocked?

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.



```
Run Results 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_dwl_odssssssss;]
2021-01-21 18:57:11 [6b295437-1109-4712-9ee0-4be04e99e8fe@ops_dwl_ods_biz_app_t_app_config]submit sql job process:1/1
2021-01-21 18:57:12 sub-rule eb295437-1109-4712-9ee0-4be04e99e8fe run failed !
2021-01-21 18:57:17 For detail:DLI.0005: Table or view not found: ops_dwl_odssssssss; 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
```

6.3 How Do I Manually Restart a Blocked Quality Job or Comparison Job?

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.

Instance Name	Type	Running Status	Notification	Start Time	Instance Search Duration	Operation
	Quality job	Failed	Successfully	Jan 21, 2021 19:00:40 G...	00:00:09	Rerun Details Rectify

6.4 How Do I View Jobs Associated with a Quality Rule Template?

Step 1 Click **Publish History** in the **Operation** column of the target rule template.

Figure 6-1 Viewing publish history

Start Date - End Date	Enter the version name
Mar 09, 2021 19:41:54 GMT+08:00	V2.0 Published
Mar 08, 2021 15:05:15 GMT+08:00	V1.0 Published

Step 2 Click **Suspend** on the right of a historical version. You can view the jobs associated with the rule template.

Figure 6-2 Viewing associated jobs

Name	Type	Template	Version
			V1.0

----End

6.5 What Should I Do If the System Displays a Message Indicating that I Do Not Have the MRS Permission to Perform a Quality Job?

An error is reported when a user executes 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. !"

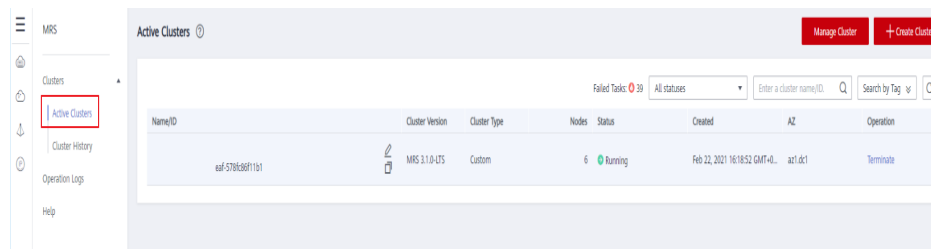
This problem occurs because the user does not have the operation permission on the MRS cluster.

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:

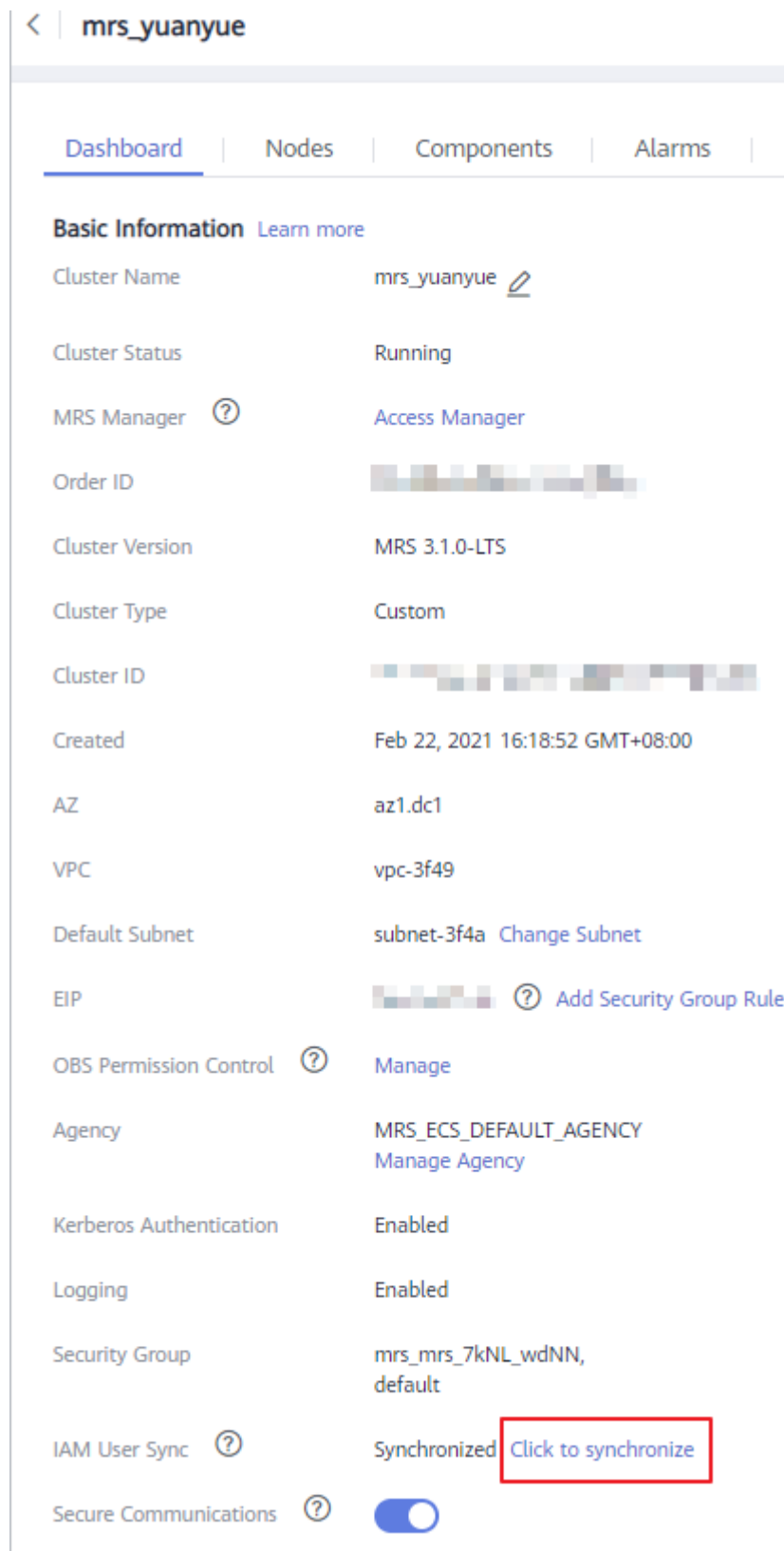
Step 1 Log in to the MRS console, view the existing clusters, and click a cluster name to access the cluster overview page.

Figure 6-3 MRS cluster instance



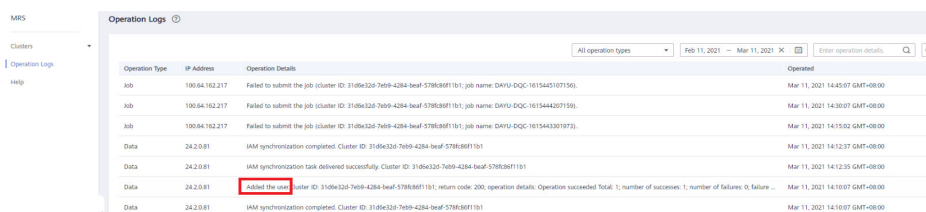
Step 2 In the IAM User Sync area, click **Click to synchronize**.

Figure 6-4 Click to synchronize



Step 3 View the operation result in the **Operation Logs** area.

Figure 6-5 Operation log



Operation Type	IP Address	Operation Details	Operated
job	100.64.162.217	Failed to submit the job (cluster ID: 3109e326-7699-4284-beaf-5786b6f11b1; job name: DAIJU-DQC-1615442107156)	Mar 11, 2021 14:45:07 GMT+08:00
job	100.64.162.217	Failed to submit the job (cluster ID: 3109e326-7699-4284-beaf-5786b6f11b1; job name: DAIJU-DQC-1615444207156)	Mar 11, 2021 14:30:07 GMT+08:00
job	100.64.162.217	Failed to submit the job (cluster ID: 3109e326-7699-4284-beaf-5786b6f11b1; job name: DAIJU-DQC-16154433071973)	Mar 11, 2021 14:15:02 GMT+08:00
Data	242.0.81	IAM synchronization completed. Cluster ID: 3109e326-7699-4284-beaf-5786b6f11b1	Mar 11, 2021 14:12:37 GMT+08:00
Data	242.0.81	IAM synchronization task delivered successfully. Cluster ID: 3109e326-7699-4284-beaf-5786b6f11b1	Mar 11, 2021 14:12:35 GMT+08:00
Data	242.0.81	Added the user (cluster ID: 3109e326-7699-4284-beaf-5786b6f11b1; 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	242.0.81	IAM synchronization completed. Cluster ID: 3109e326-7699-4284-beaf-5786b6f11b1	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.

CAUTION

You need to create an account with the same name as the current primary account.

----End

7 DataArts Catalog

7.1 What Are the Functions of the DataArts Catalog Module?

The DataArts Catalog module collects metadata and displays a data asset map of an enterprise, which contains all the metadata and data lineages.

7.2 What Assets Can Be Collected by DataArts Catalog?

For details, see [Data Sources](#).

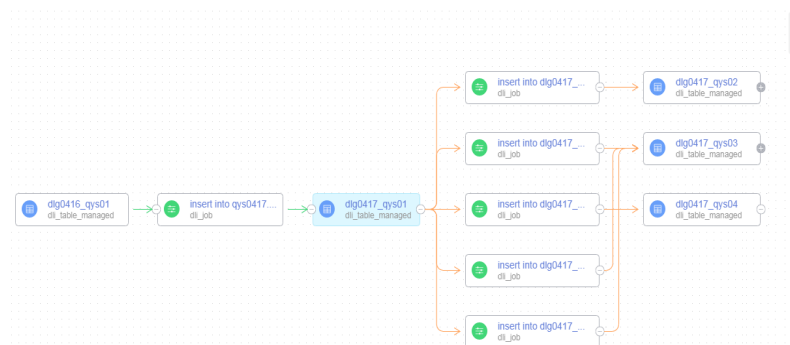
7.3 What Is Data Lineage?

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 7-1 Data lineage example



7.4 How Do I Visualize Data Lineages in a Data Catalog?

To display data lineages in a data catalog, you must schedule related jobs and collect metadata.

For details about the data lineage scheme, see [Node Lineages](#).

8 DataArts DataService

8.1 What Languages Do Data Lake Mall SDKs Support?

Data Lake Mall SDKs support C#, Python, Go, JavaScript, PHP, C++, C, Android, and Java.

8.2 What Can I Do If the System Displays a Message Indicating that the Proxy Fails to Be Invoked During API Creation?

Restart the CDM cluster during off-peak hours to release memory.

8.3 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?

Set the header parameter when invoking the API.

header parameter: x-Authorization, nvalid ___ parameter: ___,

8.4 What Can I Do If an Error Is Reported When I Use an API?

Note that each subdomain name can be accessed a maximum of 1000 times every day.

8.5 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.

8.6 What Should I Do If the API Quota Provided by DataArts DataService Exclusive Has Been Used up?

You can change the API quota. For details, see [Setting the Allocated API Quota](#).