Cloud Data Migration

User Guide

Issue 1

Date 2022-09-30

HUAWEI

HUAWEI TECHNOLOGIES CO., LTD.

Trademarks and Permissions


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 Permissions Management

1.1 Creating a User and Granting CDM Permissions

This chapter describes how to use **Identity and Access Management (IAM)** to implement fine-grained permissions control for your CDM resources. With IAM, you can:

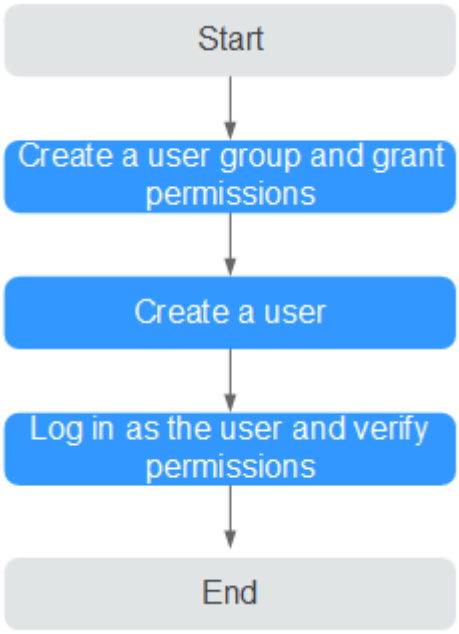
- Create IAM users for employees based on your enterprise's organizational structure. Each IAM user will have their own security credentials for accessing CDM resources.
- Grant only the permissions required for users to perform a specific task.
- Entrust a HUAWEI CLOUD account or cloud service to perform efficient O&M on your CDM resources.

If your HUAWEI CLOUD account does not require individual IAM users, skip this chapter.

This section describes the procedure for granting permissions (see **Figure 1-1**).

Process Flow

Figure 1-1 Process of granting CDM permissions



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

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

2. **Create an IAM user.**

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

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

   Log in to the CDM console by using the user created, and verify that the user only has read permissions for CDM.

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

1.2 Creating a Custom Policy

Custom policies can be created to supplement the system-defined policies of CDM. For the actions that can be added to custom policies, see **Permissions Policies and Supported Actions**.

You can create custom policies in either of the following ways:

- Visual editor: Select cloud services, actions, resources, and request conditions. This does not require knowledge of policy syntax.

- JSON: Edit JSON policies from scratch or based on an existing policy.

For details, see **Creating a Custom Policy**. The following section contains examples of common CDM custom policies.

Example Custom Policies

- Example 1: Allowing users to create a CDM cluster

```
{
   "Version": "1.1",
   "Statement": [
      {
         "Effect": "Allow",
         "Action": [
            "cdm:cluster:create"
         ]
      }
   ]
}
```

- Example 2: Denying CDM cluster deletion

  A policy with only "Deny" permissions must be used in conjunction with other policies to take effect. If the permissions assigned to a user contain both "Allow" and "Deny", the "Deny" permissions take precedence over the "Allow" permissions.

  The following method can be used if you need to assign permissions of the **CDM FullAccess** policy to a user but you want to prevent the user from deleting CDM clusters. Create a custom policy for denying CDM cluster deletion, and attach both policies to the group to which the user belongs. Then, the user can perform all operations on CDM resources except deleting CDM clusters. The following is an example of a deny policy:

```
{
   "Version": "1.1",
   "Statement": [
      {
         "Effect": "Deny",
         "Action": [
            "cdm:cluster:delete"
         ]
      }
   ]
}
```

- Example 3: Defining permissions for multiple services in a policy

  A custom policy can contain actions of multiple services that are of the global or project-level type. The following is an example policy containing actions of multiple services:

```
{
    "Version": "1.1",
    "Statement": [
        {
            "Action": [
                "cdm:cluster:list",
                "cdm:cluster:get",
                "ecs:*:get*",
                "ecs:*:list*",
                "vpc:*:get*",
                "vpc:*:list*",
                "evs:*:get*",
                "evs:*:list*",
                "bss:*:view*"
            ],
            "Effect": "Allow"
```



```
        }
    ]
}
```

2 Managing Clusters

2.1 Creating a CDM Cluster

Scenario

CDM provides isolated clusters to ensure secure and reliable data migration. Currently, a cluster supports only one server.

Prerequisites

You have applied for a VPC, subnet, and security group. If the CDM cluster tries to connect to another cloud service, ensure that the cluster and the cloud service are in the same VPC. Otherwise, an EIP is required.

NOTE

If VPC peering connection is configured, the peer VPC subnet may overlap with the CDM management network. As a result, data sources in the peer VPC cannot be accessed. You are advised to use the public network for cross-VPC data migration, or contact the administrator to add specific routes to the VPC peering connection in the CDM background.

Procedure

Step 1 Log in to the CDM management console.

Step 2 Click **Buy CDM Cluster**. The page for a CDM cluster is displayed.

Step 3 Configure the cluster parameters. **Table 2-1** describes the required parameters.

Table 2-1 Parameter description

Parameter	Example Value	Description
Region	EU-Dublin	Region where the CDM cluster resides. Resources in different regions cannot communicate with each other.
AZ	AZ2	For details, see **AZs**.

Parameter	Example Value	Description
Name	cdm-aff1	Custom CDM cluster name **NOTE** After a CDM cluster is created, its name cannot be changed.
Instance Type	cdm.large	Currently, the following flavors are available: • **cdm.large**: 8 vCPUs and 16 GB of memory. The maximum and assured bandwidths are 3 Gbit/s and 0.8 Gbit/s. Up to 16 jobs can be executed concurrently. • **cdm.xlarge**: 16 vCPUs and 32 GB of memory. The maximum and assured bandwidths are 10 Gbit/s and 4 Gbit/s. Up to 32 jobs can be executed concurrently. This flavor is suitable for migrating terabytes of data that requires a bandwidth of 10GE. • **cdm.4xlarge**: 64 vCPUs and 128 GB of memory. The maximum and assured bandwidths are 40 Gbit/s and 36 Gbit/s. Up to 64 jobs can be executed concurrently.
VPC	vpc1	VPC, subnet, and security group where the CDM cluster belongs to, which are used to communicate with the desired data source. They can be selected based on the migration source and destination. • If the CDM cluster and the data source to be connected belong to different VPCs or the data source is an on-premises one, the CDM cluster needs to be bound with an elastic IP address (EIP). • If the data source is a cloud service, you are advised to configure the network of the CDM cluster to be the same as that of the cloud service and the CDM cluster does not need to be bound with an EIP. • If the data source is a cloud service, and CDM and the cloud service are in the same VPC but in different subnets, configure security group rules to interconnect the CDM cluster with the cloud service. For details, see the **Virtual Private Cloud User Guide**.
Subnet	subnet-1	
Security Group	sg-1	
Enterprise Project	default	On the management console, click **Enterprise** in the upper right corner to access the enterprise project management page to create an enterprise project.

Parameter	Example Value	Description
Tags	cluster_owner:cdm	Tag parameters can be configured when Advanced Configuration is set to Custom. If you want to use the same tag to identify multiple types of cloud resources, you can customize the tag key and tag value. Then, you can filter cloud resources with the same tag in the TMS tag system. **NOTE** • A cluster can have a maximum of 10 tags. • A tag key and a tag value can contain a maximum of 36 and 43 characters, respectively.
Notification	No	After the function is enabled, configure a maximum of 20 mobile numbers or email addresses. You will be notified of job failures (only table/file migration jobs) and EIP exceptions by SMS message or email. **NOTE** The EIP exception notification takes effect only after the VPC policy agency of the corresponding region is created on the IAM management console. You can also choose **Authorize EIP Check** > **Create Agency** on the **Cluster Management** page to create an agency.

Step 4 Check the current configuration and click **Buy Now** to go to the page for confirming the order.

NOTE

You cannot modify the flavor of an existing cluster. If you require a higher flavor, create a cluster with your desired flavor.

Step 5 Click **Submit**. The system starts to create a CDM cluster. You can view the creation progress on the **Cluster Management** page.

----End

2.2 Binding or Unbinding an EIP

Scenario

After creating a CDM cluster, you can bind an EIP to or unbind an EIP from the cluster.

- If CDM needs to access a local or Internet data source, or a cloud service in another VPC, bind an EIP to the CDM cluster or use a NAT gateway to enable the CDM cluster to share the EIP with ECSs to access the Internet. For details, see **Adding a SNAT Rule**.
- To create an EIP exception notification, choose **Authorize EIP Check** > **Create Agency** on the **Cluster Management** page. The EIP exception notification takes effect only after the VPC policy agency of the corresponding region is created on the IAM management console.

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.

Prerequisites

- You have created a CDM cluster.
- Your EIP quota is sufficient.

Procedure

Step 1 Log in to the CDM console. In the left navigation pane, choose **Cluster Management**.

Figure 2-1 Cluster list

NOTE

The **Source** column is displayed only when you access the **DataArts Migration** page from the DataArts Studio console.

Step 2 Bind an EIP to or unbind an EIP from a cluster.

- Binding an EIP: In the **Operation** column, click **Bind EIP**. The **Bind EIP** dialog box is displayed.
- Unbinding an EIP: In the **Operation** column, choose **More** > **Unbind EIP**.

Step 3 Click **Yes**.

----End

2.3 Restarting a Cluster

Scenario

After modifying some configurations (for example, disabling user isolation), you must restart the cluster to make the modification take effect.

Prerequisites

You have created a CDM cluster.

Restarting a cluster

Step 1 Log in to the CDM console. In the left navigation pane, choose **Cluster Management**.

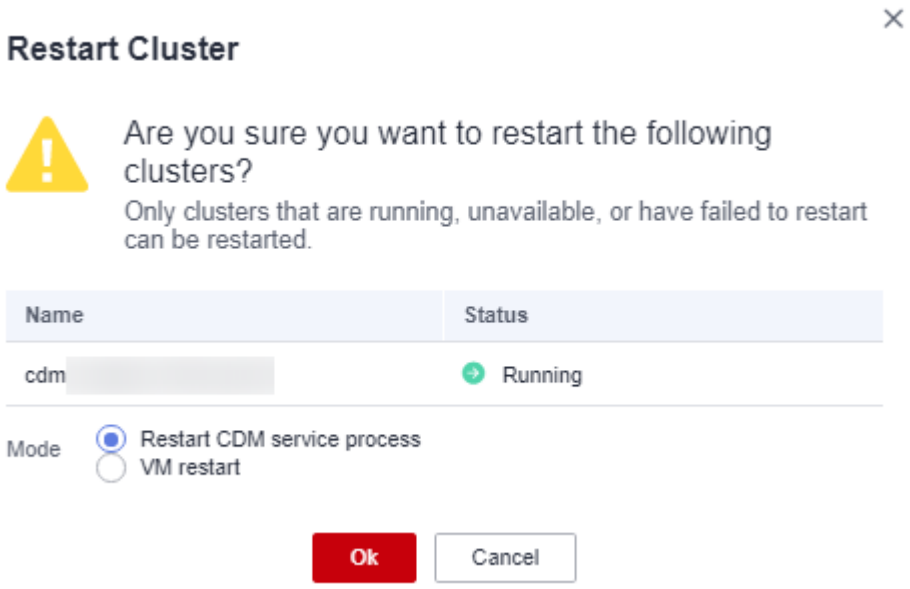
Figure 2-2 Cluster list

NOTE

The **Source** column is displayed only when you access the **DataArts Migration** page from the DataArts Studio console.

Step 2 Locate the row that contains the target cluster, click **More** in the **Operation** column, and select **Restart** from the drop-down list.

Figure 2-3 Restarting a cluster



Step 3 Select **Restart CDM service process** or **VM restart** and click **OK**.

- **Restart CDM service process**: Only the CDM service process is restarted. The cluster VM will not be restarted.
- **VM restart**: The service process will be interrupted and VMs in the cluster will be restarted.

----End

2.4 Deleting a Cluster

Scenario

You can delete a CDM cluster that you no longer use.

CAUTION

After a CDM cluster is deleted, the cluster and its data are destroyed and cannot be restored. Exercise caution when performing this operation.

Before deleting a cluster, note the following:

- Ensure that the cluster to be deleted is no longer used and that the link and job data in the cluster has been backed up through the job export function described in **Managing Jobs in Batches**.
- You are not advised to delete the CDM cluster which is free of charge. If you delete it, you can only purchase clusters.
- After a CDM cluster is deleted, it will not be billed in pay-per-use mode and the package duration will not be deducted. If you have purchased a CDM discount package or a yearly/monthly CDM incremental package for the CDM cluster to delete, unsubscribe from the package.

Prerequisites

You have created a CDM cluster.

Deleting a Cluster

Step 1 Log in to the CDM console. In the left navigation pane, choose **Cluster Management**.

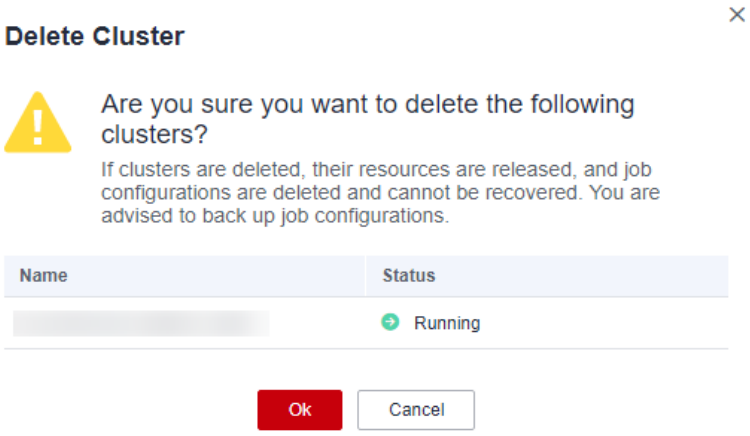
Figure 2-4 Cluster list

NOTE

The **Source** column is displayed only when you access the **DataArts Migration** page from the DataArts Studio console.

Step 2 Locate the row that contains the target cluster, click **More** in the **Operation** column, and select **Delete** from the drop-down list.

Figure 2-5 Deleting a cluster



Step 3 Click **OK** to start deleting the CDM cluster.

----End

2.5 Downloading Cluster Logs

Scenario

This section describes how to obtain cluster logs to view the job running history and locate job failure causes.

Prerequisites

You have created a CDM cluster.

Procedure

Step 1 Log in to the CDM console. In the left navigation pane, choose **Cluster Management**.

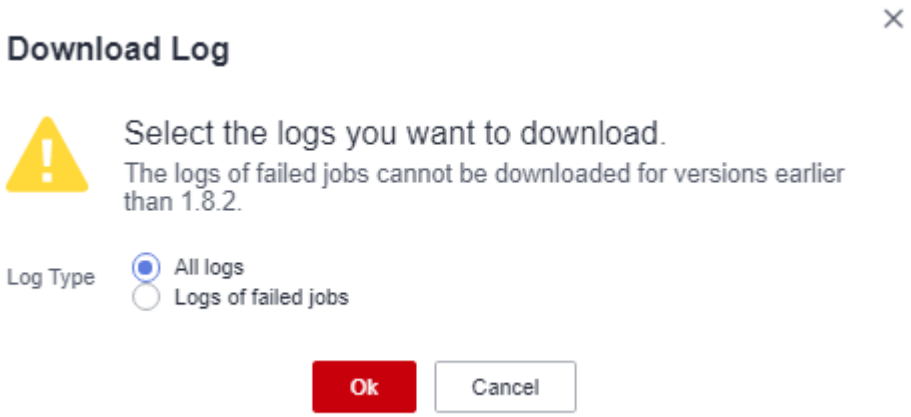
Figure 2-6 Cluster list

NOTE

The **Source** column is displayed only when you access the **DataArts Migration** page from the DataArts Studio console.

Step 2 Locate the row that contains a cluster, click **More** in the **Operation** column, and select **Download Log** from the drop-down list.

Figure 2-7 Download Log



Step 3 In the displayed dialog box, click **OK** to download logs to a local PC.

----End

2.6 Viewing Basic Cluster Information and Modifying Cluster Configurations

Scenario

After creating a CDM cluster, you can view its basic information and modify its configurations.

- You can view the following basic cluster information:
  - Cluster information: cluster version, creation time, project ID, instance ID, and cluster ID
  - Instance configuration: cluster flavor, CPU, and memory
  - Network configuration
- You can modify the following cluster configurations:
  - Notification: If a CDM migration job (only table/file migration) fails or the EIP is abnormal, CDM sends an SMS or email notification to the user. Notifications generated by this function will not be charged.
  - User isolation: determines whether other users can operate the migration jobs or links in the cluster.
    - If this function is enabled, migration jobs and links in the cluster are isolated. Other IAM users of the HUAWEI CLOUD account cannot operate the jobs and links.
    - If this function is disabled, migration jobs and links in the cluster can be shared by users. All IAM users with the required permission in the HUAWEI CLOUD account can view and perform operations on the jobs and links in the cluster.

      After disabling **User Isolation**, restart the cluster VM for the settings to take effect.
- Managing cluster tags

  You can add, modify, and delete CDM cluster tags. Tags can be used to identify multiple types of cloud resources. Cloud resources with the same tag can be filtered out in the TMS tag system.

  NOTE

  A maximum of 10 tags can be added to a CDM cluster.

Prerequisites

You have created a CDM cluster.

Viewing Basic Cluster Information

Step 1 Log in to the CDM console. In the left navigation pane, choose **Cluster Management**.

Figure 2-8 Cluster list

NOTE

The **Source** column is displayed only when you access the **DataArts Migration** page from the DataArts Studio console.

Step 2 Click the cluster name to view its basic information.

----End

Modifying Cluster Configurations

Step 1 Log in to the CDM console. In the left navigation pane, choose **Cluster Management**.

Figure 2-9 Cluster list

NOTE

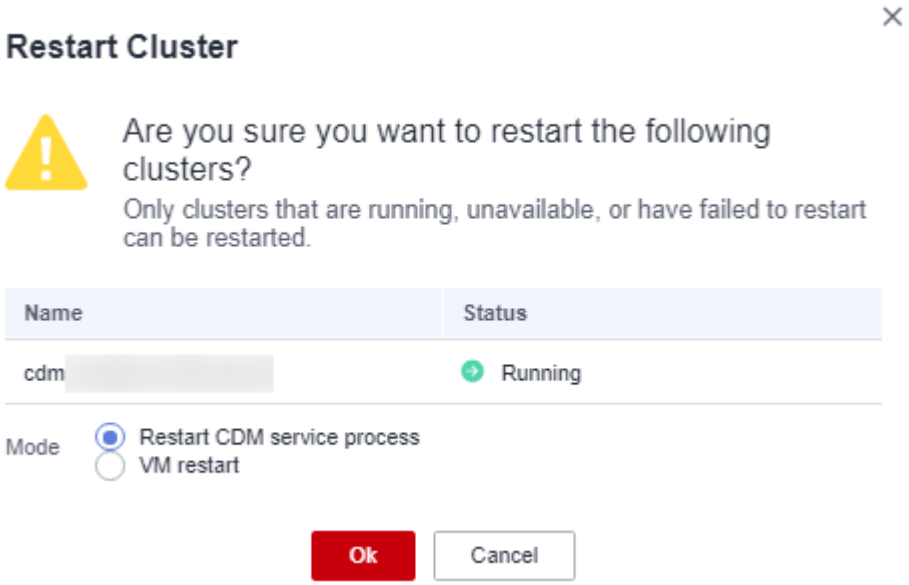
The **Source** column is displayed only when you access the **DataArts Migration** page from the DataArts Studio console.

Step 2 Click the name of a cluster and click the **Cluster Configuration** tab to modify **Notification** and **User Isolation** configuration.

Step 3 Click **Save**. The **Cluster Management** page is displayed.

Step 4 If **User Isolation** is disabled, choose **More** > **Restart** in the **Operation** column to restart the cluster VM for the settings to take effect.

Figure 2-10 Restarting a cluster



- **Restart CDM service process**: Only the CDM service process is restarted. The cluster VM will not be restarted.
- **VM restart**: The service process will be interrupted and VMs in the cluster will be restarted.

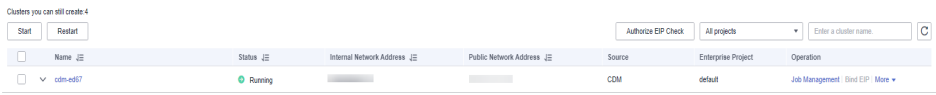
Step 5 Select **VM restart** and click **Yes**.

----End

Managing CDM Cluster Tags

Step 1 Log in to the CDM console. In the navigation pane, choose **Cluster Management**.

Figure 2-11 Cluster list

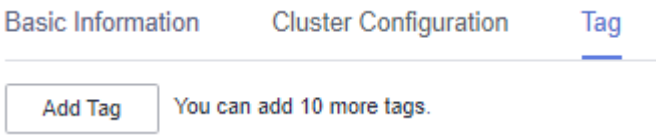


NOTE

The **Source** column is displayed only when you access the **DataArts Migration** page from the DataArts Studio console.

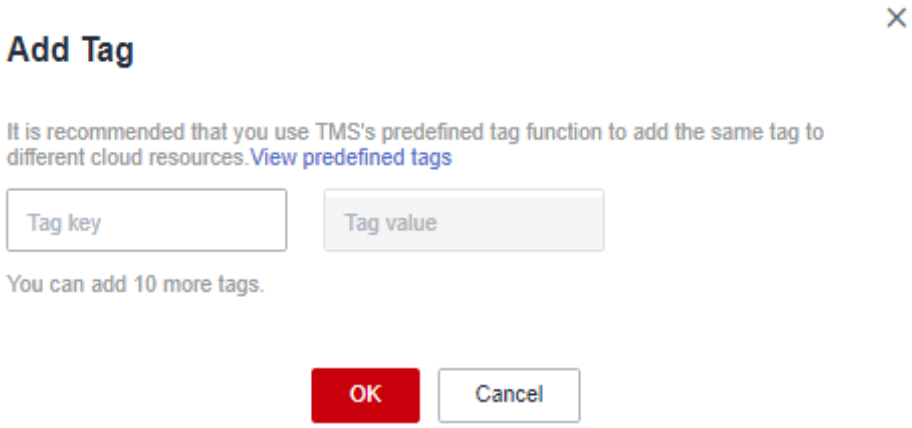
Step 2 Click a cluster name and then the **Tag** tab.

Figure 2-12 Modifying Cluster Configurations



Step 3 Click **Add Tag** and add tags to the CDM cluster.

Figure 2-13 Adding/Editing a tag

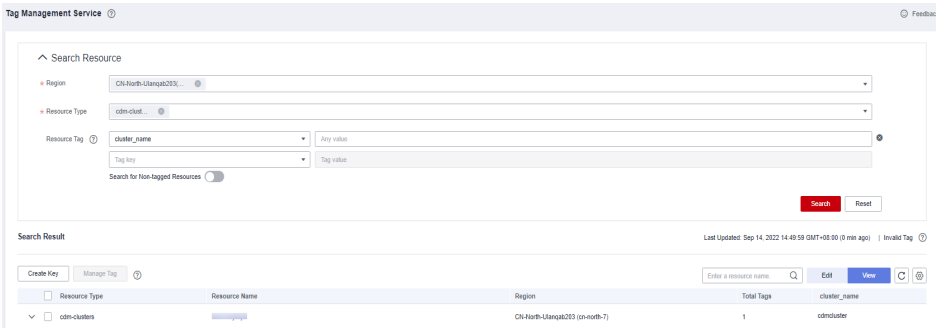


NOTE

- A cluster can have a maximum of 10 tags.
- A tag key and a tag value can contain a maximum of 36 and 43 characters, respectively.

Step 4 (Optional) In the tag list, click **Edit** or **Delete** in the **Operation** column to modify or delete tags.

Step 5 On the TMS console, set resource search criteria and click **Search** to search for the tags you added.



----End

2.7 Viewing Metrics

2.7.1 CDM Metrics

Prerequisites

You have obtained required Cloud Eye permissions.

Function

This section describes metrics reported by CDM to Cloud Eye as well as their namespaces and dimensions. You can use APIs provided by Cloud Eye to query metric information generated for CDM.

Namespace

SYS.CDM

Metrics

Table 2-2 lists the CDM metrics.

Table 2-2 CDM metrics

ID	Name	Description	Value Range	Monitored Object	Monitoring Period (Raw Data)
bytes_in	Bytes In	Measures the network inbound rate of the monitored object. Unit: byte/s	≥ 0 bytes/s	Cloud Data Migration	1 minute
bytes_out	Bytes Out	Measures the network outbound rate of the monitored object. Unit: byte/s	≥ 0 bytes/s	Cloud Data Migration	1 minute
cpu_usage	CPU Usage	Measures the CPU usage of the monitored object. Unit: %	0% to 100%	Cloud Data Migration	1 minute
mem_usage	Memory Usage	Measures the memory usage of the monitored object. Unit: %	0% to 100%	Cloud Data Migration	1 minute
disk_usage	Disk Usage	Measures the disk usage of the physical server accommodating the monitored ECS, which is not accurate as that obtained on the monitored ECS. Unit: %	0.001% to 90%	Cloud Data Migration	1 minute

ID	Name	Description	Value Range	Monitored Object	Monitoring Period (Raw Data)
disk_io	Disk I/O	Measures the bytes read from and written to a disk per second on the physical server accommodating the monitored ECS, which is not accurate as those obtained on the monitored ECS. Unit: Byte/s	0 GB to 10 GB	Cloud Data Migration	1 minute
tomcat_heap_usage	Heap Memory Usage	Measures the heap memory usage of the physical server accommodating the monitored ECS, which is not accurate as that obtained on the monitored ECS. Unit: %	0.001% to 90%	Cloud Data Migration	1 minute
tomcat_connect	Tomcat Concurrent Connections	Measures the number of Tomcat concurrent connections on the physical server.	0 to 2,147,483,647	Cloud Data Migration	1 minute
tomcat_thread_count	Tomat Threads	Measures the number of Tomcat threads on the physical server.	0 to 2,147,483,647	Cloud Data Migration	1 minute
pg_connect	Database Connections	Measures the number of Postgres database connections on the physical server.	0 to 2,147,483,647	Cloud Data Migration	1 minute
pg_submission_row	Rows	Measures the number of rows in the submission table of the Postgres database on the physical server.	0 to 2,147,483,647	Cloud Data Migration	1 minute
pg_failed_job_rate	Job Failure Rate	Measures the job failure rate of the sqoop process on the physical server. Unit: %	0.001% to 100%	Cloud Data Migration	1 minute

ID	Name	Description	Value Range	Monitored Object	Monitoring Period (Raw Data)
inodes_usage	Inodes Usage	Measures the disk inodes usage of the physical server accommodating the monitored ECS, which is not accurate as that obtained on the monitored ECS. Unit: %	0.001% to 0.9%	Cloud Data Migration	1 minute

Dimension

Key	Value
instance_id	CDM instance

2.7.2 Configuring Alarm Rules

Scenario

Set the alarm rules to customize the monitored objects and notification policies. Then, learn CDM running status in a timely manner.

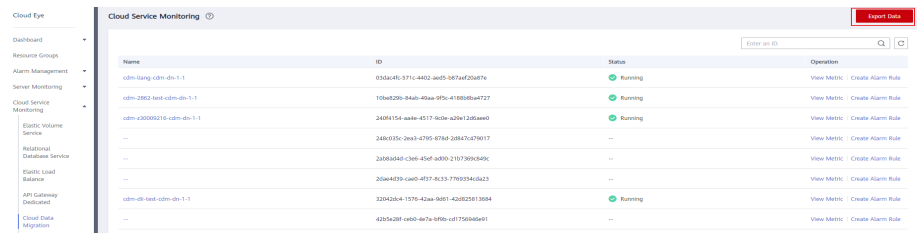
A CDM alarm rule includes the alarm rule name, monitored object, metric, threshold, monitoring interval, and whether to send a notification. This section describes how to set CDM alarm rules.

Procedure

Step 1 Access the CDM console, choose **Cluster Management**. Locate a cluster, click **More** in the **Operation** column, and select **View Metric** from the drop-down list.

Step 2 In the navigation pane, choose **Cloud Service Monitoring** > **Cloud Data Migration**. In the right pane, locate a CDM cluster and click **Create Alarm Rule** in the **Operation** column.

Figure 2-14 Monitored CDM clusters



Step 3 Set the alarm rule for the CDM cluster as prompted.

Step 4 After the setting is complete, click **Confirm**. When an alarm that meets the rule is generated, the system automatically sends a notification.

NOTE

For more information about monitoring and alarms, see the .

----End

2.7.3 Querying Metrics

Scenario

You can use Cloud Eye to monitor the running status of a CDM cluster. You can view the monitoring metrics on the Cloud Eye console.

Monitored data takes some time for transmission and display. The status displayed on the Cloud Eye console is the status obtained 5 to 10 minutes before. You can view the monitored data of a newly created CDM cluster 5 to 10 minutes later.

Prerequisites

- The CDM cluster is running properly.

  If a cluster fails to be restarted or is unavailable, its monitoring metrics are unavailable. You can view the monitored data only after the cluster is restarted or recovered.
- The cluster has been properly running for about 10 minutes.

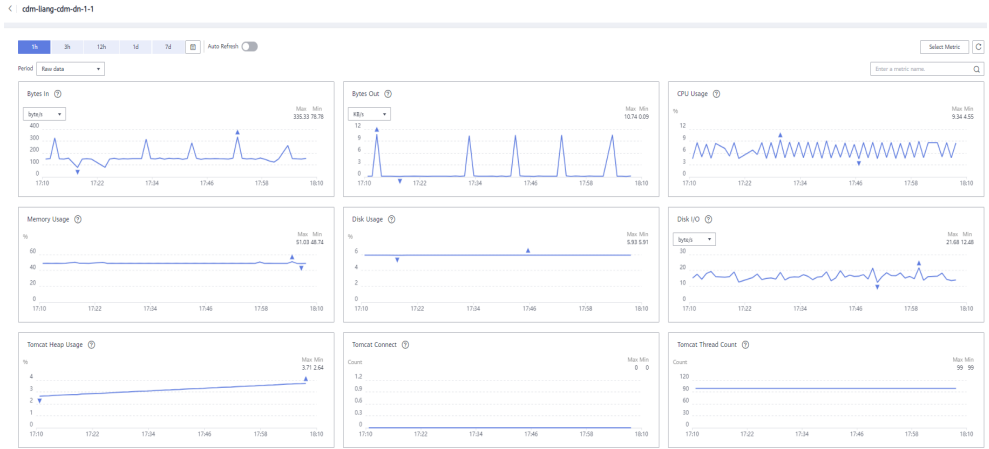
  The monitored data and graphs are available for a newly created cluster after the cluster runs for at least 10 minutes.

Procedure

Step 1 Access the CDM console, choose **Cluster Management**. Locate a cluster, click **More** in the **Operation** column, and select **View Metric** from the drop-down list.

Step 2 On the CDM monitoring page, you can view the graphs of all monitoring metrics.

Figure 2-15 Querying Metrics



Step 3 Click  in the upper right corner of the graphs to zoom in the graphs.

Step 4 You can select a time period in the upper left corner to view metric changes in this time period.

----End

3 Managing Links

3.1 Supported Data Sources

CDM provides the following migration modes which support different data sources:

- **Table/File migration** in the import of data into a data lake or migration of data to the cloud. For details, see **Data Sources Supported by Table/File Migration**.
- **Entire DB migration** in the import of data into a data lake or migration of data to the cloud. For details, see **Supported Data Sources in Entire DB Migration**.
- In addition, this section provides the data types supported in database migration. For details, see **Data Types Supported in Open-Source MySQL Database Migration**, **Data Types Supported in Oracle Database Migration**, and **Data Types Supported in SQL Server Database Migration**.

Data Sources Supported by Table/File Migration

Table/File migration can migrate data in tables or files.

Table 3-1 describes the supported data sources.

Table 3-1 Supported data sources during table/file migration

Category	Source	Destination	Description
Data warehouse	GaussDB(DWS)	• Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI) • Hadoop: MRS HDFS, MRS HBase, and MRS Hive • Object-based storage: Object Storage Service (OBS)	The DWS physical machine management mode is not supported.

Category	Source	Destination	Description
	Data Lake Insight (DLI)	• Relational database: RDS for MySQL, RDS for PostgreSQL, RDS for SQL Server, MySQL, PostgreSQL, Microsoft SQL Server, and Oracle • NoSQL: CloudTable • Search: Elasticsearch and Cloud Search Service (CSS)	-
Hadoop	MRS HDFS MRS HBase MRS Hive	• Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI) • Hadoop: MRS HDFS, MRS HBase, and MRS Hive • Object-based storage: Object Storage Service (OBS) • Relational database: RDS for MySQL, RDS for PostgreSQL, RDS for SQL Server, MySQL, PostgreSQL, Microsoft SQL Server, and Oracle • NoSQL: CloudTable • Search: Elasticsearch and Cloud Search Service (CSS)	• Supported by local storage. Only MRS Hive is supported in storage-compute decoupling scenarios. • Only MRS Hive is supported in Ranger scenarios. • Not supported if SSL is enabled for ZooKeeper • Recommended MRS HDFS versions: – 2.8.X – 3.1.X • Recommended MRS HBase versions: – 2.1.X – 1.3.X • MRS Hive 2.*x* versions are not supported. The following versions are recommended: – 1.2.X – 3.1.X
	FusionInsight HDFS	• Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI)	• FusionInsight cannot serve as the destination.

Category	Source	Destination	Description
	FusionInsight HBase FusionInsight Hive	• Hadoop: MRS HDFS, MRS HBase, and MRS Hive • Object-based storage: Object Storage Service (OBS) • NoSQL: CloudTable • Search: Elasticsearch and Cloud Search Service (CSS)	• Supported only by local storage and not in storage-compute decoupling scenarios • Not supported by Ranger • Not supported if SSL is enabled for ZooKeeper • Recommended FusionInsight HDFS versions: – 2.8.X – 3.1.X • Recommended FusionInsight HBase versions: – 2.1.X – 1.3.X • Recommended FusionInsight Hive versions: – 1.2.X – 3.1.X
	Apache HBase Apache Hive	• Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI) • Hadoop: MRS HDFS, MRS HBase, and MRS Hive • Object-based storage: Object Storage Service (OBS) • NoSQL: CloudTable • Search: Elasticsearch and Cloud Search Service (CSS)	• Apache cannot serve as the destination. • Supported only by local storage and not in storage-compute decoupling scenarios • Not supported by Ranger • Not supported if SSL is enabled for ZooKeeper • Recommended Apache HBase versions:

Category	Source	Destination	Description
	Apache HDFS		– 2.1.X – 1.3.X • Apache Hive 2.*x* versions are not supported. The following versions are recommended: – 1.2.X – 3.1.X • Recommended Apache HDFS versions: – 2.8.X – 3.1.X
Object storage	Object Storage Service (OBS)	• Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI) • Hadoop: MRS HDFS, MRS HBase, and MRS Hive • NoSQL: CloudTable • Search: Elasticsearch and Cloud Search Service (CSS)	Object Storage Migration Service (OMS) is recommended for migration between object storage services.
File system	FTP	• Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI) • Hadoop: MRS HDFS, MRS HBase, and MRS Hive • NoSQL: CloudTable • Search: Elasticsearch and Cloud Search Service (CSS)	• The file system cannot serve as the destination. • Only text files such as CSV files can be migrated from FTP or SFTP servers to search services. Binary files cannot. • obsutil is recommended for migrating data from file systems to OBS.
	SFTP		
	HTTP	Hadoop: MRS HDFS	
Relational database	RDS for MySQL	• Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI) • Hadoop: MRS HDFS, MRS HBase, and MRS Hive	• You are advised to use Data Replication Service (DRS) to migrate data
	RDS for PostgreSQL		

Category	Source	Destination	Description
	RDS for SQL Server	• Object-based storage: Object Storage Service (OBS) • NoSQL: CloudTable • Relational database: RDS for MySQL, RDS for PostgreSQL, and RDS for SQL Server • Search: Elasticsearch and Cloud Search Service (CSS)	between OLTP databases. • RDS for MySQL does not support the SSL mode. • Recommended Microsoft SQL Server version: 2005 or later
	MySQL	• Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI) • Hadoop: MRS HDFS, MRS HBase, and MRS Hive • Object-based storage: Object Storage Service (OBS) • NoSQL: CloudTable • Search: Elasticsearch and Cloud Search Service (CSS)	
	PostgreSQL		
	Microsoft SQL Server		
	Oracle		

Category	Source	Destination	Description
	SAP HANA	• Data warehouse: Data Lake Insight (DLI) • Hadoop: MRS Hive	SAP HANA data sources have the following restrictions: • SAP HANA cannot serve as the destination. • Only the 2.00.050.00.1592305219 version is supported. • Only the Generic Edition is supported. • BW/4 FOR HANA is not supported. • Only database names, table names, and column names consisting of English letters are supported. Special characters such as spaces and symbols are not allowed. • The following data types are supported: date, digit, Boolean, and character (except SHORTTEXT). Other data types such as binary are not supported. • During migration, tables cannot be automatically created at the destination.

Category	Source	Destination	Description
	Database sharding	• Data warehouse: Data Lake Insight (DLI) • Hadoop: MRS HBase and MRS Hive • Search: Elasticsearch and Cloud Search Service (CSS) • Object-based storage: Object Storage Service (OBS)	Database shards cannot serve as the destination.
NoSQL	Distributed Cache Service (DCS)	Hadoop: MRS HDFS, MRS HBase, and MRS Hive	NoSQL except CloudTable cannot serve as the destination. For how to migrate data from Redis to DCS, see **Migrating Data from Self-Hosted Redis to DCS**.
	Redis		
	Document Database Service (DDS)		
	MongoDB		
	CloudTable	• Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI) • Hadoop: MRS HDFS, MRS HBase, and MRS Hive • Object-based storage: Object Storage Service (OBS) • Relational database: RDS for MySQL, RDS for PostgreSQL, RDS for SQL Server, MySQL, PostgreSQL, Microsoft SQL Server, and Oracle • NoSQL: CloudTable • Search: Elasticsearch and Cloud Search Service (CSS)	
	Cassandra	• Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI) • Hadoop: MRS HDFS, MRS HBase, and MRS Hive • Object-based storage: Object Storage Service (OBS) • NoSQL: CloudTable • Search: Elasticsearch and Cloud Search Service (CSS)	

Category	Source	Destination	Description
Message system	Apache Kafka	Search: Cloud Search Service (CSS)	The message system cannot serve as the destination.
	DMS Kafka		
	MRS Kafka	• Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI) • Hadoop: MRS HDFS, MRS HBase, and MRS Hive • Object-based storage: Object Storage Service (OBS) • Relational database: RDS for MySQL, RDS for PostgreSQL, and RDS for SQL Server • NoSQL: CloudTable • Search: Elasticsearch and Cloud Search Service (CSS)	• MRS Kafka cannot serve as the destination. • Supported only by local storage and not in storage-compute decoupling scenarios • Not supported by Ranger • Not supported if SSL is enabled for ZooKeeper
Search	Elasticsearch	• Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI) • Hadoop: MRS HDFS, MRS HBase, and MRS Hive • Object-based storage: Object Storage Service (OBS) • Relational database: RDS for MySQL, RDS for PostgreSQL, and RDS for SQL Server • NoSQL: CloudTable • Search: Elasticsearch and Cloud Search Service (CSS)	Only the non-security mode is supported.
	Cloud Search Service (CSS)		N/A

NOTE

In the preceding table, the non-cloud data sources, such as MySQL, include on-premises MySQL, MySQL built on ECSs, or MySQL on the third-party cloud.

Supported Data Sources in Entire DB Migration

Entire DB migration is used when an on-premises data center or a database created on an ECS needs to be synchronized to a database service or big data service on the cloud. It is suitable for offline database migration but not online real-time migration.

Table 3-2 lists the data sources supporting entire DB migration using CDM.

Table 3-2 Supported data sources in entire DB migration

Category	Data Source	Read	Write	Description
Data warehouse	Data Warehouse Service (DWS)	Supported	Supported	-
	FusionInsight LibrA	Supported	Not supported	-
Hadoop (available only for local storage, and not for storage-compute decoupling, Ranger, or ZooKeeper for which SSL is enabled)	MRS HBase	Supported	Supported	Entire DB migration only to MRS HBase Recommended versions: • 2.1.X • 1.3.X
	MRS Hive	Supported	Supported	Entire DB migration only to a relational database 2.*x* versions are not supported. The following versions are recommended: • 1.2.X • 3.1.X
	FusionInsight HBase	Supported	Not supported	Recommended versions: • 2.1.X • 1.3.X
	FusionInsight Hive	Supported	Not supported	Entire DB migration only to a relational database 2.*x* versions are not supported. The following versions are recommended: • 1.2.X • 3.1.X

Category	Data Source	Read	Write	Description
	Apache HBase	Supported	Not supported	Recommended versions: • 2.1.X • 1.3.X
	Apache Hive	Supported	Not supported	Entire DB migration only to a relational database 2.*x* versions are not supported. The following versions are recommended: • 1.2.X • 3.1.X
Relational database	RDS for MySQL	Supported	Supported	Migration from OLTP to OLTP is not supported. In this scenario, you are advised to use the Data Replication Service (DRS).
	RDS for PostgreSQL	Supported	Supported	
	RDS for SQL Server	Supported	Supported	
	MySQL	Supported	Not supported	
	PostgreSQL	Supported	Not supported	
	Microsoft SQL Server	Supported	Not supported	
	Oracle	Supported	Not supported	

Category	Data Source	Read	Write	Description
	SAP HANA	Supported	Not supported	• Only the 2.00.050.00.1592305219 version is supported. • Only the Generic Edition is supported. • BW/4 FOR HANA is not supported. • Only database names, table names, and column names consisting of English letters are supported. Special characters such as spaces and symbols are not allowed. • The following data types are supported: date, digit, Boolean, and character (except SHORTTEXT). Other data types such as binary are not supported. • During migration, tables cannot be automatically created at the destination.
	MyCAT	Supported	Not supported	-

Category	Data Source	Read	Write	Description
	Dameng database	Supported	Not supported	Only to DWS and Hive
NoSQL	Distributed Cache Service (DCS)	Not supported	Supported	Only migration from MRS to DCS is supported.
	Document Database Service (DDS)	Supported	Supported	Only migration between DDS and MRS is supported.
	CloudTable Service (CloudTable)	Supported	Supported	-

Data Types Supported in Open-Source MySQL Database Migration

When the source end is an open-source MySQL database and the destination end is a Hive or DWS database, the following data types are supported:

Table 3-3 Data types supported by the open-source MySQL database functioning as the source end

Category	Type	Description	Storage Format Example	Hive	DWS
Character string	CHAR(M)	A fixed-length string of 1 to 255 characters, for example, CHAR(5). The length limit is not mandatory. It is set to 1 by default.	'a' or 'aaaaa'	CHAR	CHAR
	VARCHAR(M)	A variable-length string consists of 1 to 255 characters (more than 255 characters for MySQL of a later version). Example: VARCHAR(25). When creating a field of the VARCHAR type, you must define the length.	'a' or 'aaaaa'	VARCHAR	VARCHAR

Category	Type	Description	Storage Format Example	Hive	DWS
Value	DECIMAL(M,D)	Uncompressed floating-point numbers cannot be unsigned. In unpacking decimals, each decimal corresponds to a byte. Defining the number of display lengths (M) and decimals (D) is required. NUMERIC is the synonym of DECIMAL.	52.36	DECIMAL	When D is 0, it corresponds to BIGINT. When D is not 0, it corresponds to NUMBERIC.
	NUMBERIC	Same as DECIMAL	-	DECIMAL	NUMBERIC
	INTEGER	An integer of normal size that can be signed. If the value is signed, it ranges from -2147483648 to 2147483647. If the value is unsigned, the value ranges from 0 to 4294967295. Up to 11-bit width can be specified.	5236	INT	INTEGER
	INTEGER UNSIGNED	Unsigned form of INTEGER	-	BIGINT	INTEGER
	INT	Same as INTEGER	5236	INT	INTEGER
	INT UNSIGNED	Same as INTEGER UNSIGNED	-	BIGINT	INTEGER

Category	Type	Description	Storage Format Example	Hive	DWS
	BIGINT	A large integer that can be signed. If the value is signed, it ranges from -9223372036854775808 to 9223372036854775807. If the value is unsigned, the value ranges from 0 to 18446744073709551615. Up to 20-bit width can be specified.	5236	BIGINT	BIGINT
	BIGINT UNSIGNED	Unsigned form of BIGINT	-	BIGINT	BIGINT
	MEDIUMINT	A medium-sized integer that can be signed. If the value is signed, it ranges from -8388608 to 8388607. If the value is unsigned, it ranges from 0 to 16777215, and you can specify a maximum of 9-bit width.	-128, 127	INT	INTEGER
	MEDIUMINT UNSIGNED	Unsigned form of MEDIUMINT	-	BIGINT	INTEGER
	TINYINT	A very small integer that can be signed. If signed, the value ranges from -128 to 127. If unsigned, the value ranges from 0 to 255, and you can specify a maximum of 4-bit width.	100	TINYINT	SMALLINT

Category	Type	Description	Storage Format Example	Hive	DWS
	TINYINT UNSIGNED	Unsigned form of TINYINT	-	TINYINT	SMALLINT
	BOOL	The bool of MySQL is tinyint(1).	-128, 127	SMALLINT	BYTEA
	SMALLINT	A small integer that can be signed. If the value is signed, it ranges from -32768 to 32767. If unsigned, the value ranges from 0 to 65535, and you can specify a maximum of 5-bit width.	9999	SMALLINT	SMALLINT
	SMALLINT UNSIGNED	Unsigned form of SMALLINT	-	INT	SMALLINT
	REAL	Same as DOUBLE	-	DOUBLE	-
	FLOAT(M,D)	Unsigned floating-point numbers cannot be used. The display length (M) and number of decimal places (D) can be specified. This is not mandatory, and the default value is 10,2. In the preceding information, 2 indicates the number of decimal places and 10 indicates the total number of digits (including decimal places). The decimal precision can reach 24 floating points.	52.36	FLOAT	FLOAT4

Category	Type	Description	Storage Format Example	Hive	DWS
	DOUBLE(M, D)	Unsigned double-precision floating-point numbers cannot be used. The display length (M) and number of decimal places (D) can be specified. This is not mandatory. The default value is 16,4, where 4 is the number of decimal places. The decimal precision can reach 53-digit. REAL is a synonym of DOUBLE.	52.36	DOUBLE	FLOAT8
	DOUBLE PRECISION	Similar to DOUBLE	52.3	DOUBLE	FLOAT8
Bit	BIT(M)	Stored bit type value. BIT(M) can store up to *M* bits of values, and *M* ranges from 1 to 64.	B'1111100' B'1100'	TINYINT	BYTEA
Time and date	DATE	The value is in the *YYYY-MM-DD* format and ranges from **1000-01-01** to **9999-12-31**. For example, **December 30, 1973** will be stored as **1973-12-30**.	1999-10-01	DATE	TIMESTAMP
	TIME	Stores information about the hour, minute, and second.	'09:10:21' or '9:10:21'	Not supported (string)	TIME

Category	Type	Description	Storage Format Example	Hive	DWS
	DATETIME	The date and time are in the *YYYY-MM-DD HH:MM:SS* format and range from **1000-01-01 00:00:00** to **9999-12-31 23:59:59**. For example, **3:30 p.m. on December 30, 1973** will be stored as **1973-12-30 15:30:00**.	'1973-12-30 15:30:00'	TIMESTAMP	TIMESTAMP
	TIMESTAMP	Timestamp type. Timestamp between midnight on January 1, 1970 and a time point in 2037. Similar to the DATETIME format (YYYYMMDDHHMMSS), except that no hyphen is required. For example, **3:30 p.m. December 30, 1973** will be stored as **19731230153000**.	19731230153000	TIMESTAMP	TIMESTAMP
	YEAR(M)	The year is stored in 2-digit or 4-digit number format. If the length is specified as 2 (for example, YEAR(2)), the year ranges from 1970 to 2069 (70 to 69). If the length is specified as 4, the year ranges from 1901 to 2155. The default length is 4.	2000	Not supported (string)	Not supported
Multimedia (binary)	BINARY(M)	The number of bytes is *M*. The length of a variable-length binary string ranges from 0 to *M*. *M* is the value length plus 1.	0x2A3B4058 (binary data)	Not supported	BYTEA

Category	Type	Description	Storage Format Example	Hive	DWS
	VARBINARY(M)	The number of bytes is *M*. A fixed binary string with a length of 0 to *M*.	0x2A3B4059 (binary data)	Not supported	BYTEA
	TEXT	The maximum length of the field is 65535 characters. TEXT is a "binary large object" and is used to store large binary data, such as images or other types of files.	0x5236 (binary data)	Not supported	Not supported
	TINYTEXT	A binary string of 0 to 255 bytes in short text	-	-	Not supported
	MEDIUMTEXT	A binary string of 0 to 167772154 bytes in medium-length text	-	-	Not supported
	LONGTEXT	A binary string of 0 to 4294967295 bytes in large-length text	-	-	Not supported
	BLOB	The maximum length of the field is 65535 characters. BLOB is a "binary large object" and is used to store large binary data, such as images or other types of files. BLOB is case-sensitive.	0x5236 (binary data)	Not supported	BYTEA
	TINYBLOB	A binary string of 0 to 255 bytes in short text	-	-	BYTEA
	MEDIUMBLOB	A binary string of 0 to 167772154 bytes in medium-length text	-	-	BYTEA
	LONGBLOB	A binary string of 0 to 4294967295 bytes in large-length text	0x5236 (binary data)	Not supported	BYTEA

Category	Type	Description	Storage Format Example	Hive	DWS
Special type	SET	SET is a string object that can have no or multiple values. The values come from the allowed column of values specified when the table is created. When specifying the SET column values that contain multiple SET members, separate the members with commas (,). The SET member value cannot contain commas (,).	-	-	Not supported
	JSON	-	-	Not supported	Not supported (TEXT)
	ENUM	When an ENUM is defined, a list of its values is created, which are the items that must be used for selection (or NULL). For example, if you want a field to contain "A", "B", or "C", you can define an ENUM ("A", "B", or "C"). Only these values (or NULL) can be used to fill in the field.	-	Not supported	Not supported

Data Types Supported in Oracle Database Migration

When the source end is an Oracle database and the destination end is a Hive or DWS database, the following data sources are supported:

Table 3-4 Data types supported by the Oracle database

Category	Type	Description	Hive	DWS
Character string	char	Fixed-length character string, which is padded with spaces to reach the maximum length.	CHAR	CHAR
	nchar	Fixed-length character string contains data in Unicode format.	CHAR	CHAR
	varchar2	Synonym of VARCHAR. It is a variable-length string, unlike the CHAR type, which does not pad the field or variable to reach its maximum length with spaces.	VARCHAR	VARCHAR
	nvarchar2	Variable-length character string contains data in Unicode format.	VARCHAR	VARCHAR
Value	number	Stores numbers with a precision of up to 38 digits.	DECIMAL	NUMERIC
	binary_float	2-bit single-precision floating point number	FLOAT	FLOAT8
	binary_double	64-bit double-precision floating point number	DOUBLE	FLOAT8
	long	A maximum of 2 GB character data can be stored.	Not supported	Not supported
Time and date	date	7-byte date/time data type, including seven attributes: century, year in the century, month, day in the month, hour, minute, and second.	DATE	TIMESTAMP
	timestamp	7-byte or 11-byte fixed-width date/time data type that contains decimals (seconds)	TIMESTAMP	TIMESTAMP
	timestamp with time zone	3-byte timestamp, which supports the time zone.	TIMESTAMP	TIME WITH TIME ZONE

Category	Type	Description	Hive	DWS
	timestamp with local time zone	7-byte or 11-byte fixed-width date/time data type. Time zone conversion occurs when data is inserted or read.	TIMESTAMP	Not supported (TEXT)
	interval year to month	5-byte fixed-width data type, which is used to store a time segment.	Not supported	Not supported (TEXT)
	interval day to second	11-byte fixed-width data type, which is used to store a time segment. The time segment is stored in days/hours/minutes/seconds. The value can also contain nine decimal places (seconds).	Not supported	Not supported (TEXT)
Multimedia (binary)	raw	A variable-length binary data type. Character set conversion is not performed for data stored in this data type.	Not supported	Not supported
	long raw	Stores up to 2 GB binary information.	Not supported	Not supported
	blob	A maximum of 4 GB data can be stored.	Not supported	Not supported
	clob	In Oracle 10g and later versions, a maximum of (4 GB) x (database block size) bytes of data can be stored. CLOB contains the information for which character set conversion is to be performed. This data type is ideal for storing plain text information.	Not supported	Not supported
	nclob	This type can store a maximum of 4 GB data. When the character set is converted, this type is affected.	Not supported	Not supported
	bfile	An Oracle directory object and a file name can be stored in the database column, and the file can be read through the Oracle directory object and file name.	Not supported	Not supported

Category	Type	Description	Hive	DWS
Others	rowid	In fact, it is the address of a row in the database table. It is 10 bytes long.	Not supported	Not supported
	urowid	It is a common row ID and does not have a fixed rowid table.	Not supported	Not supported

Data Types Supported in SQL Server Database Migration

When the source end is a SQL Server database and the destination end is a Hive, Oracle or DWS database, the following data sources are supported:

Table 3-5 Data types supported by the SQL Server database functioning as the source end

Category	Type	Description	Hive	DWS	Oracle
String data type	char	Fixed-length character string, which is padded with spaces to reach the maximum length.	CHAR	CHAR	CHAR
	nchar	Fixed-length character string contains data in Unicode format.	CHAR	CHAR	CHAR
	varchar	A variable-length string consists of 1 to 255 characters (more than 255 characters for MySQL of a later version). Example: VARCHAR(25). When creating a field of the VARCHAR type, you must define the length.	VARCHAR	VARCHAR	VARCHAR
	nvarchar	Stores variable-length Unicode character data, similar to varchar.	VARCHAR	VARCHAR	VARCHAR
Numeric data type	int	int is stored in four bytes, where one binary bit represents a sign bit, and the other 31 binary bits represent a length and a size, and may represent all integers ranging from -2^{31} to $2^{31} - 1$.	INT	INTEGER	INT

Category	Type	Description	Hive	DWS	Oracle
	bigint	bigint is stored in eight bytes, where one binary bit represents a sign bit, and the other 63 binary bits represent a length and a size, and may represent all integers ranging from -2^{63} to $2^{63} - 1$.	BIGINT	BIGINT	NUMBER
	smallint	Data of the smallint type occupies two bytes of storage space. One binary bit indicates a positive or negative sign of an integer value, and the other 15 binary bits indicate a length and a size, and may represent all integers ranging from -2^{15} to 2^{15}.	SMALLINT	SMALLINT	NUMBER
	tinyint	Tinyint data occupies one byte of storage space and can represent all integers ranging from 0 to 255.	TINYINT	TINYINT	NUMBER
	real	The value can be a positive or negative decimal number.	DOUBLE	FLOAT4	NUMBER
	float	The number of digits (in scientific notation) of the mantissa of a float value, which determines the precision and storage size	FLOAT	FLOAT8	binary_float
	decimal	Numeric data type with fixed precision and scale	DECIMAL	NUMERIC	NUMBER
	numeric	Stores zero, positive, and negative fixed point numbers.	DECIMAL	NUMERIC	NUMBER
Date and time data type	date	Stores date data represented by strings.	DATE	TIMESTAMP	DATE
	time	Time of a day, which is recorded in the form of a character string.	Not supported (string)	TIME	Not supported
	datetime	Stores time and date data.	TIMESTAMP	TIMESTAMP	Not supported

Category	Type	Description	Hive	DWS	Oracle
	datetime2	Extended type of datetime, which has a larger data range. By default, the minimum precision is the highest, and the user-defined precision is optional.	TIMESTAMP	TIMESTAMP	Not supported
	smalldatetime	The smalldatetime type is similar to the datetime type. The difference is that the smalldatetime type stores data from January 1, 1900 to June 6, 2079. When the date and time precision is low, the smalldatetime type can be used. Data of this type occupies 4-byte storage space.	TIMESTAMP	TIMESTAMP	Not supported
	timestamp	Timestamp data type	TIMESTAMP	TIMESTAMP	TIMESTAMP
	datetimeoffset	A time that uses the 24-hour clock and combined with date and the time zone.	Not supported (string)	TIMESTAMP	Not supported
Multimedia data types (binary)	text	Stores text data.	Not supported (string)	Not supported (string)	Not supported
	netxt	The function of this type is the same as that of the text type. It is non-Unicode data with variable length.	Not supported (string)	Not supported (string)	Not supported
	image	Variable-length binary data used to store pictures, catalog pictures, or paintings.	Not supported (string)	Not supported (string)	Not supported
	binary	Binary data with a fixed length of *n* bytes, where *n* ranges from 1 to 8,000.	Not supported (string)	Not supported (string)	Not supported

Category	Type	Description	Hive	DWS	Oracle
	varbinary	Variable-length binary data	Not supported (string)	Not supported (string)	Not supported
Currency data type	money	Stores currency values.	Not supported (string)	Not supported (string)	Not supported
	smallmoney	Similar to the money type, a currency symbol is prefixed to the input data. For example, the currency symbol of CNY is ¥.	Not supported (string)	Not supported (string)	Not supported
Data type	bit	Bit data type. The value is 0 or 1. The length is 1 byte. A bit value is often used as a logical value to determine whether it is true(1) or false(0). If a non-zero value is entered, the system replaces it with 1.	Not supported	Not supported	Not supported
Other data types	rowversion	Each piece of data has a counter. The value of the counter increases when an insert or update operation is performed on a table that contains the **rowversion** column in the database.	Not supported	Not supported	Not supported
	uniqueidentifier	A 16-byte globally unique identifier (GUID) is a unique number generated by the SQL Server based on the network adapter address and host CPU clock. Each GUID is a hexadecimal number ranging from 0 to 9 or a to f.	Not supported	Not supported	Not supported
	cursor	Cursor data type	Not supported	Not supported	Not supported
	sql_variant	Stores any valid SQL Server data except the text, image, and timestamp data, which facilitates the development of the SQL Server.	Not supported	Not supported	Not supported

Category	Type	Description	Hive	DWS	Oracle
	table	Stores the result set after a table or view is processed.	Not supported	Not supported	Not supported
	xml	Data type of the XML data. XML instances can be stored in columns or variables of the XML type. The stored XML instance size cannot exceed 2 GB.	Not supported	Not supported	Not supported

3.2 Creating Links

Scenario

Before creating a data migration job, create a link to enable the CDM cluster to read data from and write data to a data source. A migration job requires a source link and a destination link. For details on the data sources that can be exported (source links) and imported (destination links) in different migration modes (table/file migration), see **Supported Data Sources**.

The link configurations depend on the data source. This section describes how to create these links.

Constraints

If changes occur in the connected data source (for example, the MRS cluster capacity is expanded), you need to edit and save the connection.

Prerequisites

- A CDM cluster is available.
- The CDM cluster can communicate with the destination data source.
  - 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.
  - If the destination data source is a cloud service (such as DWS, MRS, and ECS), the following requirements must be met for network interconnection:
    - If the CDM cluster and the cloud service are in different regions, a public network or a dedicated connection is required for enabling communication between the CDM cluster and the cloud service. If the Internet is used for communication, ensure that an EIP has been bound to the CDM cluster, the host where the data source is located

can access the Internet, and the port has been enabled in the firewall rules.

- If the CDM cluster and the cloud service are in the same region, VPC, subnet, and security group, they can communicate with each other by default. If they are in the same VPC but in different subnets or security groups, you must configure routing rules and security group rules. For details about how to configure routing rules, see **configuring routes**. For details about how to configure security group rules, see **configuring security group rules**.
- The cloud service instance and the CDM cluster belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

- You have obtained the URL and the account for accessing the data source. The account is granted with the read and write permissions for the data source.
- When using the Agent, you need to use the main account to grant the CDM operation permission to the sub-account.

Creating Links

Step 1 Log in to the management console and choose **Service List** > **Cloud Data Migration**. On the CDM console, choose **Cluster Management** in the left navigation pane. Locate the row that contains the target cluster and click **Job Management** in the **Operation** column. On the displayed **Links** page, click **Create Link**. On the displayed page shown in **Figure 3-1**, select a connector.

The connectors are classified based on the type of the data source to be connected. All supported data types are displayed.

Figure 3-1 Selecting a connector type

Step 2 Select a data source and click **Next**. The following describes how to create a MySQL link.

The link parameters of different data sources vary. **Table 3-6** describes the link parameters.

Table 3-6 Link parameters

Connector	Description
• Data Warehouse Service • RDS for MySQL • RDS for PostgreSQL • RDS for SQL Server • PostgreSQL • Microsoft SQL Server • SAP HANA	Because the JDBC drivers used to connect to these relational databases are the same, the parameters to be configured are also the same and are described in **Link to a Common Relational Database**.
MySQL	For details about the parameters, see **Link to an RDS for MySQL/MySQL Database**.
Oracle	For details about the parameters, see **Link to an Oracle Database**.
Database Sharding	For details about the parameters, see **Link to a Database Shard**.
HUAWEI CLOUD OBS	For details about the parameters, see **Link to OBS**.
• MRS HDFS • FusionInsight HDFS • Apache HDFS	If the data source is HDFS of MRS, Apache Hadoop, or FusionInsight HD, see **Link to HDFS**.
• MRS HBase • FusionInsight HBase • Apache HBase	If the data source is HBase of MRS, Apache Hadoop, or FusionInsight HD, see **Link to HBase**.
• MRS Hive • FusionInsight Hive • Apache Hive	If the data source is Hive on MRS, Apache Hadoop, or FusionInsight HD, see **Link to Hive**.
CloudTable Service	If the data source is CloudTable, see **Link to CloudTable**.
• FTP • SFTP	If the data source is an FTP or SFTP server, see **Link to an FTP or SFTP Server**.

Connector	Description
HTTP	These connectors are used to read files with an HTTP/HTTPS URL, such as reading public files on the third-party object storage system and web disks. When creating an HTTP link, you only need to configure the link name. The URL is configured during job creation.
MongoDB	If the data source is a local MongoDB, see **Link to MongoDB**.
Document Database Service (DDS)	If the data source is DDS, see **Link to DDS**.
• Redis • Distributed Cache Service	If the data source is Redis or DCS, see **Link to Redis/DCS**.
• MRS Kafka • Apache Kafka	If the data source is MRS Kafka or Apache Kafka, see **Link to Kafka**.
Cloud Search Service (CSS) Elasticsearch	If the data source is CSS or Elasticsearch, see **Link to Elasticsearch/CSS**.
Data Lake Insight	If the data source is DLI, see **Link to DLI**.
DMS Kafka	If the data source is DMS Kafka, see **Link to DMS Kafka**.
Cassandra	If the data source is Cassandra, see **Link to Cassandra**.

NOTE

Currently, the following data sources are in the OBT phase: FunsionInsight HDFS, FunsionInsight HBase, FunsionInsight Hive, SAP HANA, Document Database Service, CloudTable Service, Cassandra, DMS Kafka, Cloud Search Service, and Sharding Database.

Step 3 After configuring the parameters of the link, click **Test** to check whether the link is available. Alternatively, click **Save**, and the system checks automatically.

If the network is poor or the data source is too large, the link test may take 30 to 60 seconds.

----End

Managing Links

CDM allows you to perform the following operations on created links:

- Deleting links: You can delete links that are not used by any job.

- Editing a link: You can modify link parameters but cannot reselect the connector. To modify a link, you need to re-enter the password needed to access the data source.
- Testing connectivity: You can test connectivity of a link that has been saved.
- Viewing the JSON file of a link: You can view parameters of a link in a JSON file.
- Editing the JSON file of a link: Modify parameters of a link in a JSON file.
- Viewing the backend link: You can view the backend link corresponding to a link. For example, you can query details about the backend link of a MyCAT link.

Before managing a link, ensure that the link is not used by any job to avoid affecting jobs. The procedure for managing connections is as follows:

Step 1 Log in to the management console and choose **Service List** > **Cloud Data Migration**. On the CDM console, choose **Cluster Management** in the left navigation pane. Locate the row that contains the target cluster and click **Job Management** in the **Operation** column. On the displayed page, click the **Links** tab.

Step 2 On the **Links** page, locate the link to be modified.

- Deleting a link: Click **Delete** in the **Operation** column to delete a link. Alternatively, select the links that are not used by any job and click **Delete Link** above the list to delete them.
- Editing the link: Click the link name or click **Edit** in the **Operation** column to access the page for modifying the link. When modifying the link, you need to enter the password for logging in to the data source again.
- Testing connectivity of the link: Click **Test Connectivity** in the **Operation** column.
- Viewing the JSON file of the link: In the **Operation** column, choose **More** > **View Link JSON** to view link parameters in JSON format.
- Editing the JSON file of the link: In the **Operation** column, choose **More** > **Edit Link JSON** to modify link parameters in JSON format.
- Viewing the backend link: Locate the row that contains a link and click **More** in the **Operation** column and select **View Backend Link** to view the backend link corresponding to the link.

----End

3.3 Managing Drivers

The Java Database Connectivity (JDBC) provides programmatic access to relational databases. Applications can execute SQL statements and retrieve data using the JDBC API.

Before connecting CDM to a relational database, you need to upload the JDK 8 .jar driver of the relational database.

Prerequisites

- A cluster has been created.

- You have downloaded one of the drivers listed in **Table 3-7**.
- (Optional) An SFTP link has been created by referring to **Link to an FTP or SFTP Server** and the corresponding driver has been uploaded to the offline file server.

How Do I Obtain a Driver?

Select a driver version that adapts to the database type. Note that the version of the uploaded driver does not need to match the version of the database to be connected. Obtain the JDK8 .jar driver of the recommended version by referring to **Table 3-7**.

Table 3-7 Drivers

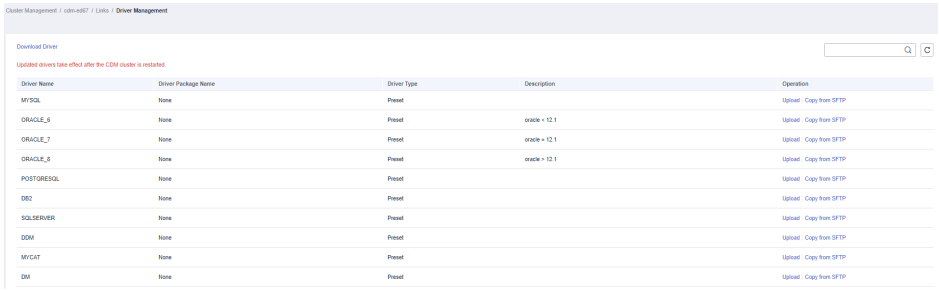
Relational Database Type	Driver Name	How to Obtain	Recommended Version
• RDS for MySQL • MySQL	MySQL MyCAT	**https://downloads.mysql.com/archives/c-j/**	mysql-connector-java-5.1.48.jar
Oracle	ORACLE_6 ORACLE_7 ORACLE_8	Driver packages: **https://www.oracle.com/database/technologies/appdev/jdbc-downloads.html** Driver packages of historical versions: **https://repo1.maven.org/maven2/com/oracle/database/jdbc/ojdbc8/12.2.0.1/**	ojdbc8.jar for version 12.2.0.1 **NOTE** New versions (for example, Oracle Database 21c (21.3) drivers) are not supported. If they are used, the schema name cannot be obtained during job creation.
• RDS for PostgreSQL • PostgreSQL	POSTGRESQL	**https://mvnrepository.com/artifact/org.postgresql/postgresql**	postgresql-42.1.4.jar for JDBC 4.2

Relational Database Type	Driver Name	How to Obtain	Recommended Version
• RDS for SQL Server • Microsoft SQL Server	SQLServer	Driver packages: **https://docs.microsoft.com/en-us/sql/connect/jdbc/download-microsoft-jdbc-driver-for-sql-server?view=sql-server-ver15** Driver packages of historical versions: **https://docs.microsoft.com/en-us/sql/connect/jdbc/release-notes-for-the-jdbc-driver?view=sql-server-ver15#previous-releases**	sqljdbc42.jar

Procedure

Step 1 Access the CDM console, choose **Cluster Management** in the navigation pane, locate the target cluster, and choose **Job Management** > **Link Management** > **Driver Management**. The **Driver Management** page is displayed.

Figure 3-2 Uploading a driver



Step 2 Click **Upload** in the **Operation** column and select a local driver.

Alternatively, click **Copy from SFTP** in the **Operation** column and configure the **SFTP Link** name and **Driver File Path**.

Step 3 (Optional) If you have uploaded an updated version of a driver, you must restart the CDM cluster for the new driver to take effect.

----End

3.4 Managing Agents

If your data is stored in HDFS or a relational database, you can deploy an agent on the source network. CDM pulls data from your internal data sources through an agent but cannot write data into the databases.

Figure 3-3 Scenario

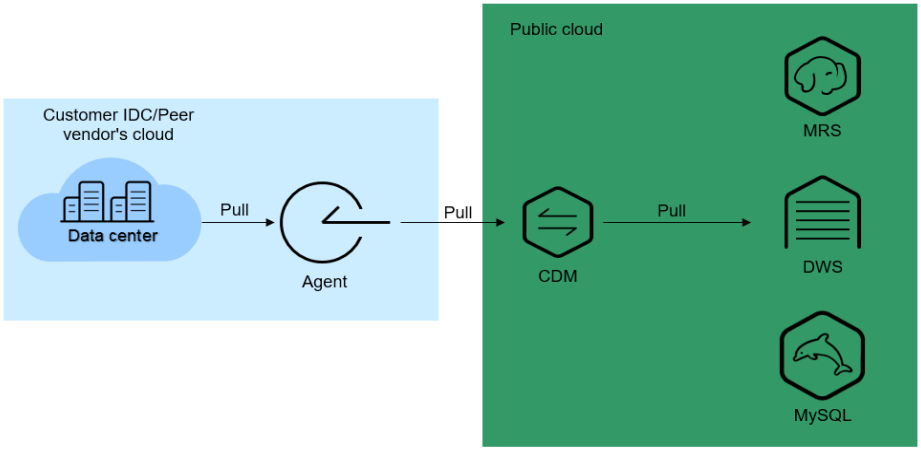
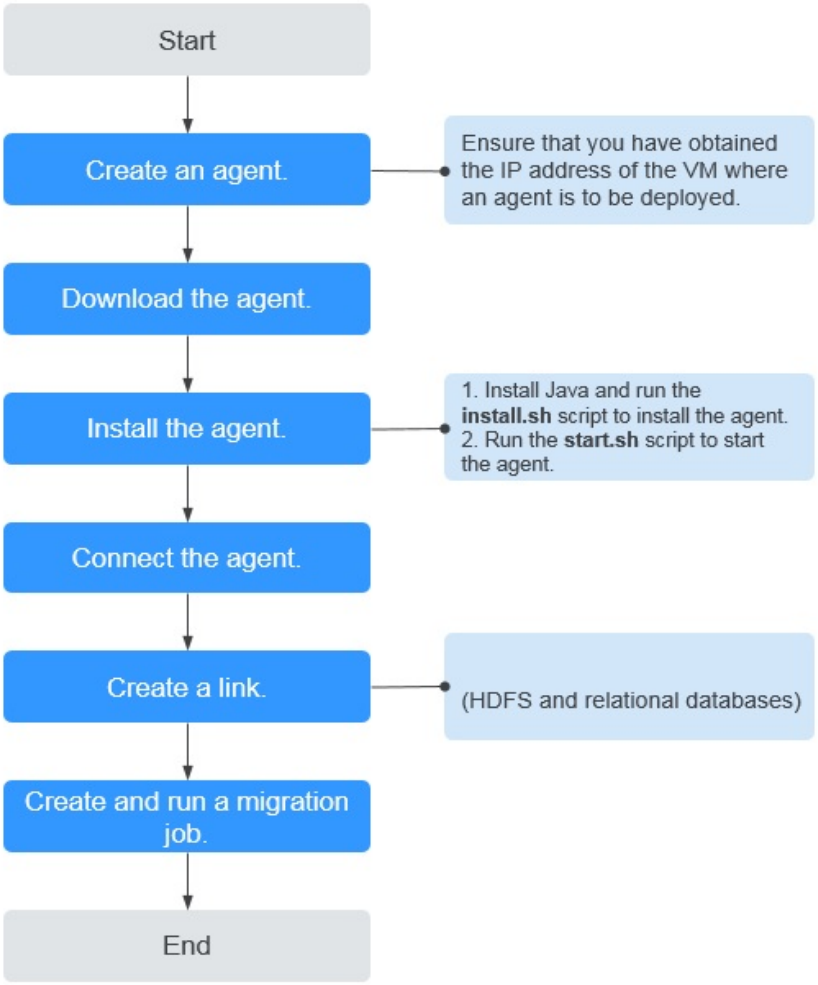


Figure 3-4 shows the process of using an agent.

Figure 3-4 Process



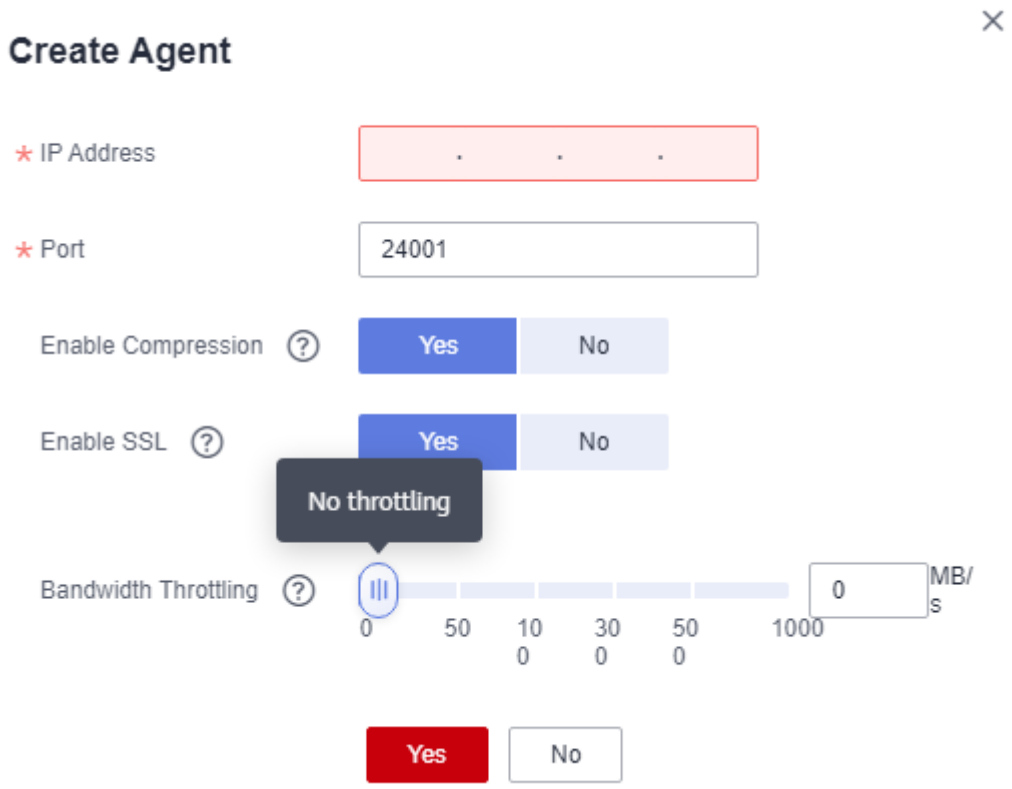
Prerequisites

A CDM cluster is available.

Creating an Agent

Step 1 Access the CDM console and choose **Cluster Management** in the left navigation pane. Locate the target cluster, choose **Job Management** > **Agent Management** > **Create Agent**, and configure agent parameters.

Figure 3-5 Creating an agent



- **IP Address**: Set this parameter to the IP address of the server where the agent is deployed on the source network.
- **Port**: custom port of the agent Recommended value range: 1024–65535.
- **Enable Compression**: whether to compress data using the gzip algorithm.
  - Enable this function for text data (data based on character encoding, such as MySQL INT data) because such data can be well compressed by the gzip algorithm. (For details about text data, see the related database documentation.)
  - Disable this function for binary data (data based on value encoding, such as MySQL BINARY data) because such data has been compressed, and compressing it again will increase the workload to decompress data and undermine the performance of the client. (For details about text data, see the related database documentation.)
- **Enable SSL**: whether to enable two-way SSL authentication Enable this function if security is of high priority.
- **Bandwidth Throttling**: set the maximum downstream rate of the agent. By default, there is no throttling.

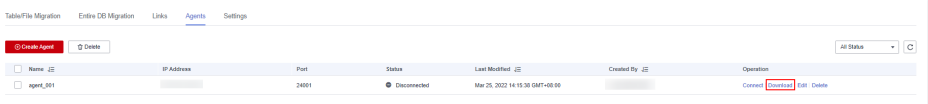
Step 2 Click **OK**. On the **Agent Management** page, view the created agent.

----End

Installing and Starting an Agent

Step 1 On the **Agent Management** page, locate the created agent and click **Download** in the **Operation** column.

Figure 3-6 Downloading an agent



Step 2 Prepare the server for installing the agent. The host has no special requirements for vCPUs, memory, and disks, but must meet the following requirements:

- Java 8 (64-bit) has been installed and Java environment variables have been configured.
- User **Ruby** must be granted the write permission of the **/tmp** directory. If there is no user **Ruby**, create one.

Step 3 Upload the downloaded agent package to the server.

Step 4 Decompress the package and run the following command to install the agent:

sh sbin/install.sh

Step 5 If you want to use the agent to connect to a relational database, you need to upload the corresponding drivers (see **Managing Drivers**) to the **/server/jdbc** directory in the agent installation directory and modify the version number of the corresponding database driver in the **properties** file in the same directory.

Step 6 Run the following command as user **root** to change the owner and group of the driver uploaded to the **/server/jdbc** directory to **Ruby**:

chown Ruby.Ruby * -R

Step 7 After the installation is complete, run the following commands to start the agent:

su Ruby

sh sbin/start.sh

Step 8 Run the following command to check whether the agent is started:

ps -ef | grep cdm

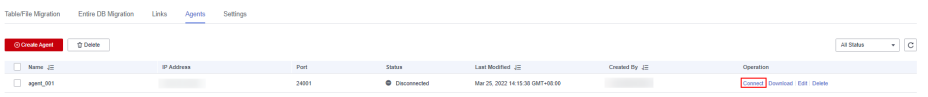
If the command output contains the running agent process, the agent process has been started.

----End

Connecting to an Agent

Step 1 On the **Agent Management** page, locate the created agent and click **Connect** in the **Operation** column.

Figure 3-7 Connecting to an agent



Step 2 After the agent is successfully connected, you can select it when creating a connection.

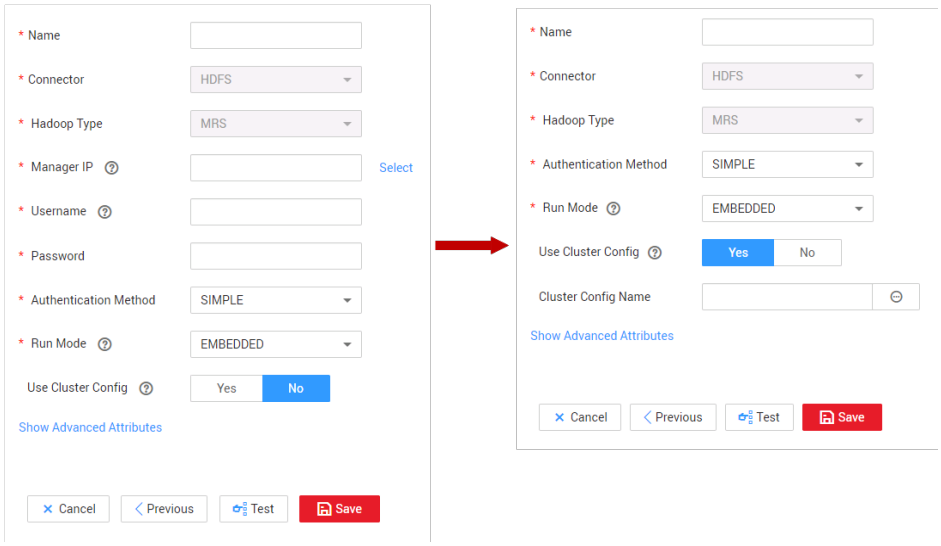
----End

3.5 Managing Cluster Configurations

On the **Cluster Configurations** page, you can create, edit, or delete Hadoop cluster configurations.

When creating a Hadoop link, the Hadoop cluster configurations can simplify the link creation. See **Figure 3-8** for details.

Figure 3-8 Comparison before and after using the cluster configurations



CDM supports the following types of Hadoop links:

- MRS clusters: MRS HDFS, MRS HBase, and MRS Hive
- FusionInsight clusters: FusionInsight HDFS, FusionInsight HBase, and FusionInsight Hive
- Apache clusters: Apache HDFS, Apache HBase, and Apache Hive

Scenario

Before creating a Hadoop link, you are advised to create cluster configurations to simplify the link parameter configurations.

Prerequisites

- A cluster has been created.
- You have obtained the Hadoop cluster configuration file and keytab file. See **Table 1** for details.

Obtaining the Cluster Configuration File and Keytab File

The methods for obtaining the Hadoop cluster configuration file and keytab file vary depending on the Hadoop cluster type. For details, see **Table 1**.

Table 3-8 Obtaining the cluster configuration file and keytab file

Hadoop Link	Obtaining the Cluster Configuration File	Obtaining the Keytab File
MRS cluster • MRS HDFS • MRS HBase • MRS Hive	For clusters of MRS 3.x: 1. Log in to FusionInsight Manager. 2. Choose **Cluster** > *Name of the desired cluster* > **Dashboard** > **More** > **Download Client**. 3. In the dialog box that is displayed, select **Configuration Files Only**. The platform type must be the same as that on the server. Click **OK** to download the configuration file to the local host. 4. Obtain the downloaded TAR package, which is the FusionInsight cluster configuration file. For clusters of MRS 2.x or earlier: 1. Log in to the MRS console. 2. Choose **Clusters** > **Active Clusters** and click a cluster name to go to the cluster details page. Click the **Components** tab. 3. Click **Download Client**. Set **Client Type** to **Only configuration files**, set **Download To** to **Server** or **Remote host**, customize the client path, and click **OK** to generate the client configuration file. 4. Save the generated configuration file to a local path. See MRS documentation for details.	For clusters of MRS 3.x: 1. Log in to FusionInsight Manager. 2. Choose **System** > **Permission** > **User**, locate the row that contains the target user, and choose **More** > **Download Authentication Credential** to download the authentication credential file. 3. Obtain the downloaded TAR package, which is the keytab file of the FusionInsight cluster. For clusters of MRS 2.x or earlier: 1. Log in to MRS Manager and click **System**. In the **Permission** area, click **Manage User**. 2. In the row of the user for whom you want to export the keytab file, choose **More** > **Download authentication credential** to download the authentication file. After the file is automatically generated, save it to a specified path and keep it properly. See MRS documentation for details.

Hadoop Link	Obtaining the Cluster Configuration File	Obtaining the Keytab File
FusionInsight clusters: • FusionInsight HDFS • FusionInsight HBase • FusionInsight Hive	1. Log in to FusionInsight Manager. 2. Choose **Cluster** > *Name of the desired cluster* > **Dashboard** > **More** > **Download Client**. 3. In the dialog box that is displayed, select **Configuration Files Only**. The platform type must be the same as that on the server. Click **OK** to download the configuration file to the local host. 4. Obtain the downloaded TAR package, which is the FusionInsight cluster configuration file. See the FusionInsight documentation for details.	1. Log in to FusionInsight Manager. 2. Choose **System** > **Permission** > **User**, locate the row that contains the target user, and choose **More** > **Download Authentication Credential** to download the authentication credential file. 3. Obtain the downloaded TAR package, which is the keytab file of the FusionInsight cluster. See the FusionInsight documentation for details.

Hadoop Link	Obtaining the Cluster Configuration File	Obtaining the Keytab File
Apache clusters: • Apache HDFS • Apache HBase • Apache Hive	In the Apache cluster scenario, only the required configuration files and packaging rules are described. For details about how to obtain each configuration file, see the corresponding documentation. • HDFS needs to compress the following files into a .zip package without the directory format: – hosts – core-site.xml – hdfs-site.xml – yarm-site.xml – mapred-site.xml – krb5.conf (optional, for clusters in security mode) • HBase needs to compress the following files into a .zip package without the directory format: – hosts – core-site.xml – hdfs-site.xml – yarm-site.xml – mapred-site.xml – hbase-site.xml – krb5.conf (optional, for clusters in security mode) • Hive needs to compress the following files into a .zip package without the directory format: – hosts – core-site.xml – hdfs-site.xml – yarm-site.xml	In the Apache cluster scenario, only the principles for packaging authentication credential files are required. For details about how to obtain the authentication credential files, see the corresponding documentation. 1. Rename the user's authentication credential file as **user.keytab**. 2. Compress the **user.keytab** file into a .zip package without the directory format: **user.keytab.zip**.

Hadoop Link	Obtaining the Cluster Configuration File	Obtaining the Keytab File
	– mapred-site.xml – hive-site.xml – hivemetastore-site.xml – krb5.conf (optional, for clusters in security mode)	

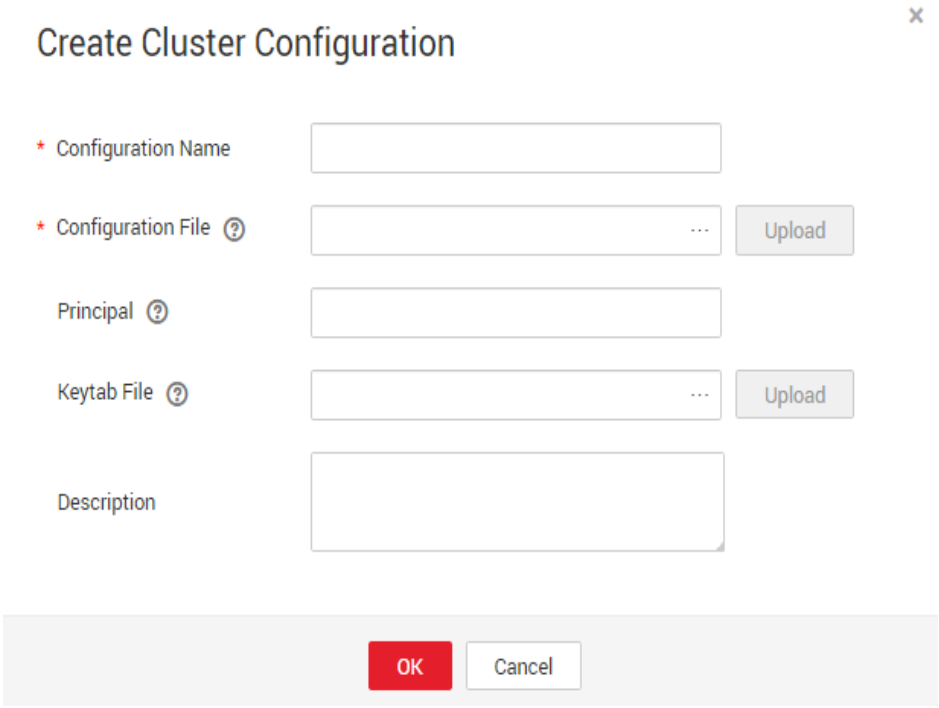
NOTE

- A cluster configuration file contains the configuration parameters of the cluster. If the cluster configuration parameters are modified, you need to obtain the configuration file again.
- The keytab file is the authentication credential file. Before obtaining the keytab file, you need to change the password of this user at least once in the cluster. Otherwise, the downloaded keytab file may be unavailable. After a user password is changed, the exported keytab file becomes invalid, and you need to export a keytab file again.
- The keytab file is used only in a cluster in security mode. In other cases, you do not need to prepare the keytab file.

Procedure

1. On the CDM console, choose **Cluster Management** in the left navigation pane. Locate the row that contains a cluster and choose **Job Management** > **Links** > **Cluster Configurations**.
2. On the **Cluster Configurations** page, click **Create Cluster Configuration** and set the parameters as prompt.

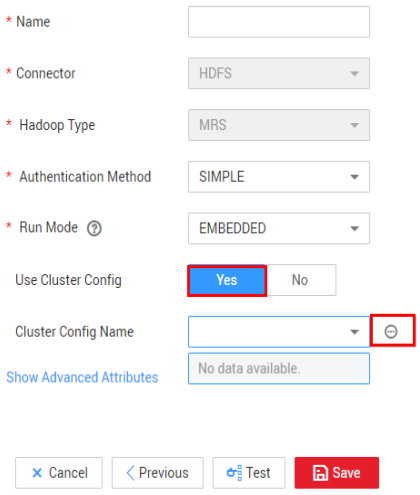
Figure 3-9 Creating cluster configurations



- **Configuration Name**: Enter a cluster configuration name that is easy to remember and distinguish based on the type of the data source to be connected.
- **Configuration File**: Click **Select File** to select a local cluster configuration file, and then click **Upload** on the right to upload the file.
- **Principal**: This parameter is required only for clusters in security mode. Principal is the username in Kerberos security mode and must be the same as that in the keytab file.
- **Keytab File**: Upload the keytab file only for clusters in security mode. Click **Select File** to select a local keytab file, and then click **Upload** on the right to upload the file.
- **Description**: Add a description to identify and distinguish the cluster configuration.

3. Click **OK**. When creating a Hadoop link, set **Authentication Method** as required, **Use Cluster Config** to **Yes**, and then select the corresponding cluster configuration name to quickly create a Hadoop link.

Figure 3-10 Use Cluster Config



3.6 Link to a Common Relational Database

Common relational databases include GaussDB(DWS), RDS for PostgreSQL, RDS for SQLServer, PostgreSQL, Microsoft SQL Server, and SAP HANA.

Prerequisites

You have uploaded required drivers by following the instructions in **Managing Drivers**.

Parameters for a link to a common relational database

Table 3-9 lists the link parameters.

Table 3-9 Parameters for a link to a common relational database

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	mysql_link
Database Server	IP address or domain name of the database to connect Click **Select** next to the text box and select a DWS or RDS DB instance in the displayed dialog box.	192.168.0.1

Parameter	Description	Example Value
Port	Port of the database to connect	The port number varies depending on the database. Examples: Default port of SQL Server: **1433** Default port of PostgreSQL: **5432**
Database Name	Name of the database to connect	dbname
Username	Username used for accessing the database This account must have the permissions required to read and write data tables and metadata.	cdm
Password	Password of the user	-
Use Agent	Whether to extract data from the data source through an agent	Yes
Agent	Click **Select** and select the agent created in **Managing Agents**.	-
Driver Version	Select a driver version that adapts to the database type.	-
Fetch Size	(Optional) Displayed when you click **Show Advanced Attributes**. Number of rows obtained by each request. Set this parameter based on the data source and the job's data size. If the value is either too large or too small, the job may run for a long time.	1000
SSL Encryption	(Optional) If you set this parameter to **Yes**, CDM can connect to the database (on-premises databases excluded) in SSL encryption mode. Security hardening has been performed on RDS for PostgreSQL. For this reason, when creating a link to RDS for PostgreSQL, set this parameter to **Yes**.	Yes

Parameter	Description	Example Value
Link Attributes	(Optional) Click **Add** to add the JDBC connector attributes of multiple specified data sources. For details, see the JDBC connector document of the corresponding database. The following are some examples: • **connectTimeout=360000** and **socketTimeout=360000**: When a large amount of data needs to be migrated or the entire table is retrieved using query statements, the migration fails due to connection timeout. In this case, you can customize the connection timeout interval (ms) and socket timeout interval (ms) to prevent failures caused by timeout. • **useCursorFetch=false**: By default, **useCursorFetch** is enabled, indicating that the JDBC connector communicates with relational databases using a binary protocol. Some third-party systems may have compatibility issues, causing migration time conversion errors. In this case, you can disable this function. Open-source MySQL databases support the **useCursorFetch** parameter, and you do not need to set this parameter.	sslmode=require
Reference Sign	(Optional) Delimiter between the names of the referenced tables or columns. For details, see the product documentation of the corresponding database.	'

3.7 Link to an RDS for MySQL/MySQL Database

Table 3-10 lists the parameters for a link to a MySQL database.

Table 3-10 Parameters for a link to a MySQL database

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	mysql_link
Database Server	IP address or domain name of the database to connect Click **Select** next to the text box and select a MySQL DB instance in the displayed dialog box.	192.168.0.1

Parameter	Description	Example Value
Port	Port of the database to connect	3306
Database Name	Name of the database to connect	dbname
Username	Username used for accessing the database This account must have the permissions required to read and write data tables and metadata.	cdm
Password	Password of the user	-
Use Local API	(Optional) Whether to use the local API of the database for acceleration. When you create a MySQL link, CDM automatically enables the **local_infile** system variable of the MySQL database to enable the LOAD DATA function, which accelerates data import to the MySQL database. If this parameter is enabled, the date type that does not meet the format requirements will be stored as 0000-00-00. For details, visit the official MySQL website. If CDM fails to enable this function, contact the database administrator to enable the **local_infile** system variable. Alternatively, set **Use Local API** to **No** to disable API acceleration. If data is imported to RDS for MySQL, the LOAD DATA function is disabled by default. In such a case, you need to modify the parameter group of the MySQL instance and set **local_infile** to **ON** to enable the LOAD DATA function. **NOTE** If **local_infile** on RDS is uneditable, it is the default parameter group. You need to create a parameter group, modify its values, and apply it to the RDS for MySQL instance. For details, see the *Relational Database Service User Guide*.	Yes
Use Agent	Whether to extract data from the data source through an agent	Yes
Agent	Click **Select** and select the agent created in **Managing Agents**.	-
local_infile Character Set	When using local_infile to import data to MySQL, you can configure the encoding format.	utf8
Driver Version	Select a driver version that adapts to the database type.	-

Parameter	Description	Example Value
Fetch Size	(Optional) Displayed when you click **Show Advanced Attributes**. Number of rows obtained by each request. Set this parameter based on the data source and the job's data size. If the value is either too large or too small, the job may run for a long time.	1000
Commit Size	(Optional) Displayed when you click **Show Advanced Attributes**. Number of records submitted each time. Set this parameter based on the data destination and the job's data size. If the value is either too large or too small, the job may run for a long time.	-

Parameter	Description	Example Value
Link Attributes	(Optional) Click **Add** to add the JDBC connector attributes of multiple specified data sources. For details, see the JDBC connector document of the corresponding database. The following are some examples: • **connectTimeout=360000** and **socketTimeout=360000**: When a large amount of data needs to be migrated or the entire table is retrieved using query statements, the migration fails due to connection timeout. In this case, you can customize the connection timeout interval (ms) and socket timeout interval (ms) to prevent failures caused by timeout. • **tinyInt1isBit=false** or **mysql.bool.type.transform=false**: By default, **tinyInt1isBit** is **true**, indicating that **TINYINT(1)** is processed as a bit, that is, **Types.BOOLEAN**, and **1** or **0** is read as **true** or **false**. As a result, the migration fails. In this case, you can set **tinyInt1isBit** to **false** to avoid migration failures. • **useCursorFetch=false**: By default, **useCursorFetch** is enabled, indicating that the JDBC connector communicates with relational databases using a binary protocol. Some third-party systems may have compatibility issues, causing migration time conversion errors. In this case, you can disable this function. Open-source MySQL databases support the **useCursorFetch** parameter, and you do not need to set this parameter. • **allowPublicKeyRetrieval=true**: By default, public key retrieval is disabled for MySQL databases. If TLS is unavailable and an RSA public key is used for encryption, connection to an MySQL database may fail. In this case, you can enable public key retrieval to avoid connection failures.	sslmode=require
Reference Sign	(Optional) Delimiter between the names of the referenced tables or columns. For details, see the product documentation of the corresponding database.	'

Parameter	Description	Example Value
Batch Size	Number of rows written each time. It should be less than Commit Size. When the number of rows written reaches the value of Commit Size, the rows will be committed to the database.	100

3.8 Link to an Oracle Database

Table 3-11 lists the parameters for a link to an Oracle database.

Table 3-11 Parameters for a link to an Oracle database

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	oracle_link
Database Server	IP address or domain name of the database to connect	192.168.0.1
Port	Port of the database to connect	Default port: 1521
Connection Type	Oracle database connection type. The following options are available: • **Service Name**: Use **SERVICE_NAME** to connect to the Oracle database. • **SID**: Use **SID** to connect to the Oracle database.	SID
Instance Name	Oracle instance ID, which is used to differentiate databases by instances. This parameter is available only when **Connection Type** is set to **SID**.	dbname
Database Name	Name of the database to connect This parameter is available only when **Connection Type** is set to **Service Name**.	dbname
Username	Username used for accessing the database This account must have the permissions required to read and write data tables and metadata.	cdm
Password	Password of the username	-
Use Agent	Whether to extract data from the data source through an agent	Yes
Agent	Click **Select** and select the agent created in **Managing Agents**.	-

Parameter	Description	Example Value
Oracle Version	Oracle database version. This parameter is available only for Oracle links. If **java.sql.SQLException: Protocol violation** is displayed, select another version.	Later than 12.1
Fetch Size	(Optional) Displayed when you click **Show Advanced Attributes**. Number of rows obtained by each request. Set this parameter based on the data source and the job's data size. If the value is either too large or too small, the job may run for a long time. A migration from the Oracle to DWS database may time out due to a long data write duration in the DWS database. In this case, reduce the value of **Fetch Size** for the Oracle database.	1000
Link Attributes	(Optional) Click **Add** to add the JDBC connector attributes of multiple specified data sources. For details, see the JDBC connector document of the corresponding database. The following are some examples: • **oracle.net.CONNECT_TIMEOUT=360000** and **oracle.jdbc.ReadTimeout=360000**: When a large amount of data needs to be migrated or the entire table is retrieved using query statements, the migration fails due to connection timeout. In this case, you can customize the connection timeout interval (ms) and read timeout interval (ms) to prevent failures caused by timeout.	sslmode=require
Reference Sign	(Optional) Delimiter between the names of the referenced tables or columns. For details, see the product documentation of the corresponding database.	'

3.9 Link to a Database Shard

Sharding refers to the link to multiple backend data sources at the same time. The link can be used as the job source to migrate data from multiple data sources to other data sources. **Table 3-12** lists the link parameters.

Table 3-12 Parameters for a link to a database shard

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	my_link
Username	Username used for accessing the database For a backend database A, this configuration takes effect only when no username and password are configured for A in the data source list. For a backend database B that has configured the username and password, this configuration does not takes effect.	cdm
Password	Password used for accessing the database. For a backend database A, this configuration takes effect only when no username and password are configured for A in the data source list. For a backend database B that has configured the username and password, this configuration does not takes effect.	-
Use Agent	Whether to extract data from the data source through an agent	Yes
Agent	Click **Select** and select the agent created in **Managing Agents**.	-
backendDatasource	Enter the type of the backend database. Currently, only MySQL is supported.	MySQL
Data Source List	Enter the IP address, port number, database name, account name, and password of the backend database, and separate them with colons (:). That is, ip:port:dbs:username:password. You can leave username:password empty. In this case, the username and password are used. If there are multiple backend databases, ensure that the table structures are the same and use vertical bars (|) to separate data sources. If the password contains a vertical bar (|) or colon (:), use a backslash (\) to escape the vertical bar. For example, **192.168.2.1:3306:cdm|192.168.2.2:3306:cdm:user:password** indicates that the IP address of the first backend database is **192.168.2.1**, the port number is **3306**, the database name is **cdm**, and the account name and password are configured in *user* and *password*. The IP address of the second backend database is **192.168.2.2**, the port number is **3306**, the database name is **cdm**, the account name is **user** and the password is **password**.	192.168.2.1:3306:cdm|192.168.2.2:3306:cdm:user:password

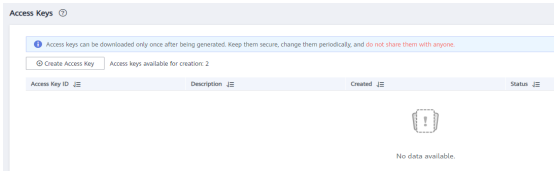
Parameter	Description	Example Value
Fetch Size	(Optional) Displayed when you click **Show Advanced Attributes**. Number of rows obtained by each request. Set this parameter based on the data source and the job's data size. If the value is either too large or too small, the job may run for a long time.	1000
Link Attributes	(Optional) Click **Add** to add the JDBC connector attributes of multiple specified data sources. For details, see the JDBC connector document of the corresponding database.	sslmode=require
Reference Sign	(Optional) Delimiter between the names of the referenced tables or columns. For details, see the product documentation of the corresponding database.	'

3.10 Link to DLI

When connecting CDM to DLI, configure the parameters as described in **Table 3-13**.

Table 3-13 Parameter description

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	dli_link
AK	AK/SK required for authentication during access to the DLI database. You need to create an access key for the current account and obtain an AK/SK pair. 1. Log in to the management console, move the cursor to the username in the upper right corner, and select **My Credentials** from the drop-down list. 2. On the **My Credentials** page, choose **Access Keys**, and click **Create Access Key**. See **Figure 3-11**.	-

Parameter	Description	Example Value
SK	**Figure 3-11** Clicking Create Access Key  3. Click **OK** and save the access key file as prompted. The access key file will be saved to your browser's configured download location. Open the **credentials.csv** file to view **Access Key Id** and **Secret Access Key**. **NOTE** • Only two access keys can be added for each user. • To ensure access key security, the access key is automatically downloaded only when it is generated for the first time and cannot be obtained from the management console later. Keep them properly.	-
Project ID	Project ID in the region where DLI resides You can obtain the project ID and account ID by performing the following steps: 1. Register with and log in to the management console. 2. Hover the cursor on the username in the upper right corner and select **My Credentials** from the drop-down list. 3. On the **My Credentials** page, obtain the account name and account ID, and obtain the project ID from the project list.	-

3.11 Link to Hive

CDM supports the following Hive data sources:

- **MRS Hive**
- **FusionInsight Hive**
- **Apache Hive**

MRS Hive

You can view a table during field mapping only when you have the permission to access the table connected to MRS Hive.

MRS Hive links apply to the MapReduce Service (MRS) on . **Table 3-14** describes related parameters.

NOTE

- Before creating an MRS link, you need to add an authenticated Kerberos user on MRS and log in to the MRS management page to change the initial password. Then use the new user to create an MRS link.
- To connect to an MRS 2.*x* cluster, create a CDM cluster of version 2.*x* first. CDM 1.8.*x* clusters cannot connect to MRS 2.*x* clusters.
- Currently, the Hive link obtains the **core-site.xml** configuration information from MRS HDFS. Therefore, if MRS Hive uses OBS as the underlying storage system, configure the AK/SK of OBS on MRS HDFS before creating the Hive link.
- Ensure that the MRS cluster and the DataArts Studio instance can communicate with each other. The following requirements must be met for network interconnection:
  - If the CDM cluster in the DataArts Studio instance and the MRS cluster are in different regions, a public network or a dedicated connection is required. If the Internet is used for communication, ensure that an EIP has been bound to the CDM cluster, and the MRS cluster can access the Internet and the port has been enabled in the firewall rule.
  - If the CDM cluster in the DataArts Studio instance and the MRS cluster are in the same region, VPC, subnet, and security group, they can communicate with each other by default. If they are in the same VPC but in different subnets or security groups, you must configure routing rules and security group rules. For details about how to configure routing rules, see **configuring routes**. For details about how to configure security group rules, see **configuring security group rules**.
  - The MRS cluster and the DataArts Studio workspace belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

Table 3-14 MRS Hive link parameters

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	hivelink
Manager IP	Floating IP address of MRS Manager. Click **Select** next to the **Manager IP** text box to select an MRS cluster. CDM automatically fills in the authentication information.	127.0.0.1
Authentication Method	Authentication method used for accessing MRS • **SIMPLE**: Select this for non-security mode. • **KERBEROS**: Select this for security mode.	SIMPLE
HIVE Version	Set this to the Hive version on the server.	HIVE_3_X

Parameter	Description	Example Value
Username	If **Authentication Method** is set to **KERBEROS**, you must provide the username and password used for logging in to MRS Manager. If you need to create a snapshot when exporting a directory from HDFS, the user configured here must have the administrator permission on HDFS. 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.	cdm
Password	Password used for logging in to MRS Manager	-
OBS storage support	The server must support OBS storage. When creating a Hive table, you can store the table in OBS.	No

Parameter	Description	Example Value
Run Mode	This parameter is used only when the Hive version is **HIVE_3_X**. Possible values are: • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **Standalone**: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select **STANDALONE** or configure different agents. Note: The STANDALONE mode is used to solve the version conflict problem. If the connector versions of the source and destination ends of the same link are different, a JAR file conflict occurs. In this case, you need to place the source or destination end in the STANDALONE process to prevent the migration failure caused by the conflict.	EMBEDDED
Check Hive JDBC Connectivity	Whether to check the Hive JDBC connectivity	No
Use Cluster Config	You can use the cluster configuration to simplify parameter settings for the Hadoop connection.	No
Cluster Config Name	This parameter is valid only when **Use Cluster Config** is set to **Yes**. Select a cluster configuration that has been created. For details, see **Managing Cluster Configurations**.	hive_01

Click **Show Advanced Attributes**, and then click **Add** to add configuration attributes of other clients. The name and value of each attribute must be configured. You can click **Delete** to delete no longer used attributes.

The following are some examples:

- **connectTimeout=360000** and **socketTimeout=360000**: When a large amount of data needs to be migrated or the entire table is retrieved using query statements, the migration fails due to connection timeout. In this case, you can customize the connection timeout interval (ms) and socket timeout interval (ms) to prevent failures caused by timeout.
- **hive.server2.idle.operation.timeout=360000**: To prevent Hive migration jobs from being suspended for a long time, you can customize the operation timeout period (ms).

- **hive.storeFormat=textfile**: During data migration from a relational database to Hive, tables in ORC format are automatically created by default. If you want textfile or parquet tables to be created, add **hive.storeFormat=textfile** or **hive.storeFormat=parquet**.
- **fs.defaultFS=obs://hivedb**: If the interconnected MRS Hive uses decoupled storage and compute, you can use this configuration to achieve better compatibility.

FusionInsight Hive

The FusionInsight Hive link is applicable to data migration of FusionInsight HD in the local data center. You must use Direct Connect to connect to FusionInsight HD.

Table 3-15 describes related parameters.

Table 3-15 FusionInsight Hive link parameters

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	hivelink
Manager IP	IP address of FusionInsight Manager	127.0.0.1
Manager Port	Port number of FusionInsight Manager	28443
CAS Server Port	Port number of the CAS server used to connect to FusionInsight	20009
Authentication Method	Authentication method used for accessing the cluster: • **SIMPLE**: Select this for non-security mode. • **KERBEROS**: Select this for security mode.	SIMPLE
HIVE Version	Hive version	HIVE_3_X
Username	Username used for logging in to FusionInsight Manager.	cdm
Password	Password used for logging in to FusionInsight Manager	-
OBS storage support	The server must support OBS storage. When creating a Hive table, you can store the table in OBS.	No

Parameter	Description	Example Value
Run Mode	This parameter is used only when the Hive version is **HIVE_3_X**. Possible values are: • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **Standalone**: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select **STANDALONE** or configure different agents. Note: The STANDALONE mode is used to solve the version conflict problem. If the connector versions of the source and destination ends of the same link are different, a JAR file conflict occurs. In this case, you need to place the source or destination end in the STANDALONE process to prevent the migration failure caused by the conflict.	EMBEDDED
Use Cluster Config	You can use the cluster configuration to simplify parameter settings for the Hadoop connection.	No
Cluster Config Name	This parameter is valid only when **Use Cluster Config** is set to **Yes**. Select a cluster configuration that has been created. For details, see **Managing Cluster Configurations**.	hive_01

Click **Show Advanced Attributes**, and then click **Add** to add configuration attributes of other clients. The name and value of each attribute must be configured. You can click **Delete** to delete no longer used attributes.

The following are some examples:

- **connectTimeout=360000** and **socketTimeout=360000**: When a large amount of data needs to be migrated or the entire table is retrieved using query statements, the migration fails due to connection timeout. In this case, you can customize the connection timeout interval (ms) and socket timeout interval (ms) to prevent failures caused by timeout.
- **hive.server2.idle.operation.timeout=360000**: To prevent Hive migration jobs from being suspended for a long time, you can customize the operation timeout period (ms).
- **hive.storeFormat=textfile**: During data migration from a relational database to Hive, tables in ORC format are automatically created by default. If you want textfile or parquet tables to be created, add **hive.storeFormat=textfile** or **hive.storeFormat=parquet**.

Apache Hive

The Apache Hive link is applicable to data migration of the third-party Hadoop in the local data center or ECS. You must use Direct Connect to connect to Hadoop in the local data center.

Table 3-16 describes related parameters.

Table 3-16 Apache Hive link parameters

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	hivelink
URI	NameNode URI	hdfs://hacluster
Hive Metastore	Hive metadata address. For details, see the **hive.metastore.uris** configuration item. Example: thrift://host-192-168-1-212:9083	-
Authentication Method	Authentication method used for accessing the cluster: • **SIMPLE**: Select this for non-security mode. • **KERBEROS**: Select this for security mode.	SIMPLE
HIVE Version	Hive version	HIVE_3_X
IP and Host Name Mapping	If the Hadoop configuration file uses the host name, configure the mapping between the IP address and host name. Separate the IP addresses and host names by spaces and mappings by semicolons (;), carriage returns, or line feeds.	-
OBS storage support	The server must support OBS storage. When creating a Hive table, you can store the table in OBS.	No
Principal	When **Authentication Method** is set to **KERBEROS**, this parameter is mandatory. It is the username in the Kerberos security mode and can be obtained from the Hadoop administrator. The value of this parameter must be the same as that in the Keytab file.	-

Parameter	Description	Example Value
Keytab File	When **Authentication Method** is set to **KERBEROS**, a Keytab file must be uploaded. The Keytab file is an authentication credential and can be obtained from the Hadoop administrator. Before obtaining the keytab file, you need to change the password of this user at least once in the cluster. Otherwise, the downloaded keytab file may be unavailable. After a user password is changed, the exported keytab file becomes invalid, and you need to export a keytab file again.	-
Run Mode	This parameter is used only when the Hive version is **HIVE_3_X**. Possible values are: • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **Standalone**: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select **STANDALONE** or configure different agents. Note: The STANDALONE mode is used to solve the version conflict problem. If the connector versions of the source and destination ends of the same link are different, a JAR file conflict occurs. In this case, you need to place the source or destination end in the STANDALONE process to prevent the migration failure caused by the conflict.	EMBEDDED
Use Cluster Config	You can use the cluster configuration to simplify parameter settings for the Hadoop connection.	No
Cluster Config Name	This parameter is valid only when **Use Cluster Config** is set to **Yes**. Select a cluster configuration that has been created. For details, see **Managing Cluster Configurations**.	hive_01
Hive JDBC URL	URL for connecting to Hive JDBC. By default, anonymous users are used.	-

Click **Show Advanced Attributes**, and then click **Add** to add configuration attributes of other clients. The name and value of each attribute must be configured. You can click **Delete** to delete no longer used attributes.

The following are some examples:

- **connectTimeout=360000** and **socketTimeout=360000**: When a large amount of data needs to be migrated or the entire table is retrieved using query statements, the migration fails due to connection timeout. In this case, you can customize the connection timeout interval (ms) and socket timeout interval (ms) to prevent failures caused by timeout.
- **hive.server2.idle.operation.timeout=360000**: To prevent Hive migration jobs from being suspended for a long time, you can customize the operation timeout period (ms).
- **hive.storeFormat=textfile**: During data migration from a relational database to Hive, tables in ORC format are automatically created by default. If you want textfile or parquet tables to be created, add **hive.storeFormat=textfile** or **hive.storeFormat=parquet**.

3.12 Link to HBase

CDM supports the following HBase data sources:

- **MRS HBase**
- **FusionInsight HBase**
- **Apache HBase**

MRS HBase

When connecting CDM to HBase of MRS, configure the parameters as described in **Table 3-17**.

NOTE

- Before creating an MRS link, you need to add an authenticated Kerberos user on MRS and log in to the MRS management page to change the initial password. Then use the new user to create an MRS link.
- To connect to an MRS 2.*x* cluster, create a CDM cluster of version 2.*x* first. CDM 1.8.*x* clusters cannot connect to MRS 2.*x* clusters.
- If the connection fails after you select a cluster, check whether the MRS cluster can communicate with the CDM instance which functions as the agent. They can communicate with each other in the following scenarios:
  - If the CDM cluster in the DataArts Studio instance and the MRS cluster are in different regions, a public network or a dedicated connection is required. If the Internet is used for communication, ensure that an EIP has been bound to the CDM cluster, and the MRS cluster can access the Internet and the port has been enabled in the firewall rule.
  - If the CDM cluster in the DataArts Studio instance and the MRS cluster are in the same region, VPC, subnet, and security group, they can communicate with each other by default. If they are in the same VPC but in different subnets or security groups, you must configure routing rules and security group rules. For details about how to configure routing rules, see **configuring routes**. For details about how to configure security group rules, see **configuring security group rules**.
  - The MRS cluster and the DataArts Studio workspace belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

Table 3-17 MRS HBase link parameters

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	mrs_hbase_link
Manager IP	Floating IP address of MRS Manager. Click **Select** next to the **Manager IP** text box to select an MRS cluster. CDM automatically fills in the authentication information.	127.0.0.1
Username	If **Authentication Method** is set to **KERBEROS**, you must provide the username and password used for logging in to MRS Manager. If you need to create a snapshot when exporting a directory from HDFS, the user configured here must have the administrator permission on HDFS. 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.	cdm
Password	Password used for logging in to MRS Manager	-

Parameter	Description	Example Value
Authentication Method	Authentication method used for accessing the cluster: • **SIMPLE**: Select this for non-security mode. • **KERBEROS**: Select this for security mode.	SIMPLE
HBase Version	HBase version	HBASE_2_X
Run Mode	Run mode of the HBase link. This parameter is used only when the HBase version is **HBASE_2_X**. • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **Standalone**: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select **STANDALONE** or configure different agents. Note: The STANDALONE mode is used to solve the version conflict problem. If the connector versions of the source and destination ends of the same link are different, a JAR file conflict occurs. In this case, you need to place the source or destination end in the STANDALONE process to prevent the migration failure caused by the conflict.	STANDALONE
Use Cluster Config	You can create cluster configurations on the **Links** page to simplify the configuration of Hadoop link parameters.	No
Cluster Config Name	This parameter is valid only when **Use Cluster Config** is set to **Yes**. Select a cluster configuration that has been created. For details, see **Managing Cluster Configurations**.	hbase_01

Click **Show Advanced Attributes**, and then click **Add** to add configuration attributes of other clients. The name and value of each attribute must be configured. You can click **Delete** to delete no longer used attributes.

FusionInsight HBase

When connecting CDM to HBase of FusionInsight HD, configure the parameters as described in **Table 3-18**.

Table 3-18 FusionInsight HBase link parameters

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	FI_hbase_link
Manager IP	IP address of FusionInsight Manager	127.0.0.1
Manager Port	Port number of FusionInsight Manager	28443
CAS Server Port	Port number of the CAS server used to connect to FusionInsight	20009
Username	Username used for logging in to FusionInsight Manager.	cdm
Password	Password used for logging in to FusionInsight Manager	-
Authentication Method	Authentication method used for accessing the cluster: • **SIMPLE**: Select this for non-security mode. • **KERBEROS**: Select this for security mode.	Kerberos
HBase Version	HBase version	HBASE_2_X
Run Mode	Run mode of the HBase link. This parameter is used only when the HBase version is **HBASE_2_X**. • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **Standalone**: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select **STANDALONE** or configure different agents. Note: The STANDALONE mode is used to solve the version conflict problem. If the connector versions of the source and destination ends of the same link are different, a JAR file conflict occurs. In this case, you need to place the source or destination end in the STANDALONE process to prevent the migration failure caused by the conflict.	STANDALONE
Use Cluster Config	You can use the cluster configuration to simplify parameter settings for the Hadoop connection.	No

Parameter	Description	Example Value
Cluster Config Name	This parameter is valid only when **Use Cluster Config** is set to **Yes**. Select a cluster configuration that has been created. For details, see **Managing Cluster Configurations**.	hbase_01

Click **Show Advanced Attributes**, and then click **Add** to add configuration attributes of other clients. The name and value of each attribute must be configured. You can click **Delete** to delete no longer used attributes.

Apache HBase

When connecting CDM to HBase of Apache Hadoop, configure the parameters as described in **Table 3-19**.

Table 3-19 Apache HBase link parameters

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	hadoop_hbase_link
ZK Link	ZooKeeper link of HBase Format: <host1>:<port>,<host2>:<port>,<host3>:<port>	zk1.example.com:2181,zk2.example.com:2181,zk3.example.com:2181
Authentication Method	Authentication method used for accessing the cluster: • **SIMPLE**: Select this for non-security mode. • **KERBEROS**: Select this for security mode.	Kerberos
Principal	When **Authentication Method** is set to **KERBEROS**, this parameter is mandatory. It is the username in the Kerberos security mode and can be obtained from the Hadoop administrator. The value of this parameter must be the same as that in the Keytab file.	-

Parameter	Description	Example Value
Keytab File	When **Authentication Method** is set to **KERBEROS**, a Keytab file must be uploaded. The Keytab file is an authentication credential and can be obtained from the Hadoop administrator. Before obtaining the keytab file, you need to change the password of this user at least once in the cluster. Otherwise, the downloaded keytab file may be unavailable. After a user password is changed, the exported keytab file becomes invalid, and you need to export a keytab file again.	-
IP and Host Name Mapping	If the configuration file uses the host name, configure the mapping between the IP address and host name. Separate the IP addresses and host names by spaces and mappings by semicolons (;), carriage returns, or line feeds.	10.3.6.9 hostname01 10.4.7.9 hostname02
HBase Version	HBase version	HBASE_2_X
Run Mode	Run mode of the HBase link. This parameter is used only when the HBase version is **HBASE_2_X**. • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **Standalone**: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select **STANDALONE** or configure different agents. Note: The STANDALONE mode is used to solve the version conflict problem. If the connector versions of the source and destination ends of the same link are different, a JAR file conflict occurs. In this case, you need to place the source or destination end in the STANDALONE process to prevent the migration failure caused by the conflict.	STANDALONE
Use Cluster Config	You can use the cluster configuration to simplify parameter settings for the Hadoop connection.	No

Parameter	Description	Example Value
Cluster Config Name	This parameter is valid only when **Use Cluster Config** is set to **Yes**. Select a cluster configuration that has been created. For details, see **Managing Cluster Configurations**.	hbase_01

Click **Show Advanced Attributes**, and then click **Add** to add configuration attributes of other clients. The name and value of each attribute must be configured. You can click **Delete** to delete no longer used attributes.

3.13 Link to HDFS

CDM supports the following HDFS data sources:

- **MRS HDFS**
- **FusionInsight HDFS**
- **Apache HDFS**

MRS HDFS

When connecting CDM to HDFS of MRS, configure the parameters as described in **Table 3-20**.

NOTE

- Before creating an MRS link, you need to add an authenticated Kerberos user on MRS and log in to the MRS management page to change the initial password. Then use the new user to create an MRS link.
- To connect to an MRS 2.*x* cluster, create a CDM cluster of version 2.*x* first. CDM 1.8.*x* clusters cannot connect to MRS 2.*x* clusters.
- If the connection fails after you select a cluster, check whether the MRS cluster can communicate with the CDM instance which functions as the agent. They can communicate with each other in the following scenarios:
  - If the CDM cluster in the DataArts Studio instance and the MRS cluster are in different regions, a public network or a dedicated connection is required. If the Internet is used for communication, ensure that an EIP has been bound to the CDM cluster, and the MRS cluster can access the Internet and the port has been enabled in the firewall rule.
  - If the CDM cluster in the DataArts Studio instance and the MRS cluster are in the same region, VPC, subnet, and security group, they can communicate with each other by default. If they are in the same VPC but in different subnets or security groups, you must configure routing rules and security group rules. For details about how to configure routing rules, see **configuring routes**. For details about how to configure security group rules, see **configuring security group rules**.
  - The MRS cluster and the DataArts Studio workspace belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

Table 3-20 MRS HDFS link parameters

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	mrs_hdfs_link
Manager IP	Floating IP address of MRS Manager. Click **Select** next to the **Manager IP** text box to select an MRS cluster. CDM automatically fills in the authentication information.	127.0.0.1
Username	If **Authentication Method** is set to **KERBEROS**, you must provide the username and password used for logging in to MRS Manager. If you need to create a snapshot when exporting a directory from HDFS, the user configured here must have the administrator permission on HDFS. 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.	cdm
Password	Password used for logging in to MRS Manager	-

Parameter	Description	Example Value
Authentication Method	Authentication method used for accessing MRS • **SIMPLE**: Select this for non-security mode. • **KERBEROS**: Select this for security mode.	SIMPLE
Run Mode	Run mode of the HDFS link. The options are as follows: • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **STANDALONE**: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select **STANDALONE** or configure different agents. Note: The STANDALONE mode is used to solve the version conflict problem. If the connector versions of the source and destination ends of the same link are different, a JAR file conflict occurs. In this case, you need to place the source or destination end in the STANDALONE process to prevent the migration failure caused by the conflict. • **Agent**: The link instance runs on an agent. If **Agent** is not used, and the CDM cluster connects to two or more clusters with Kerberos authentication enabled and the same realm, only one cluster can be connected in **EMBEDDED** mode, and the other clusters must be in **STANDALONE** mode.	STANDALONE
Agent	Click **Select** and select the agent created in **Connecting to an Agent**. This parameter is displayed when **Run Mode** is set to **Agent**.	-
Use Cluster Config	You can use the cluster configuration to simplify parameter settings for the Hadoop connection.	No
Cluster Config Name	This parameter is valid only when **Use Cluster Config** is set to **Yes**. Select a cluster configuration that has been created. For details, see **Managing Cluster Configurations**.	hdfs_01

Click **Show Advanced Attributes**, and then click **Add** to add configuration attributes of other clients. The name and value of each attribute must be configured. You can click **Delete** to delete no longer used attributes.

FusionInsight HDFS

When connecting CDM to HDFS of FusionInsight HD, configure the parameters as described in **Table 3-21**.

Table 3-21 FusionInsight HDFS link parameters

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	FI_hdfs_link
Manager IP	IP address of FusionInsight Manager	127.0.0.1
Manager Port	Port number of FusionInsight Manager	28443
CAS Server Port	Port number of the CAS server used to connect to FusionInsight	20009
Username	Username used for logging in to FusionInsight Manager. If you need to create a snapshot when exporting a directory from HDFS, the user configured here must have the administrator permission on HDFS.	cdm
Password	Password used for logging in to FusionInsight Manager	-
Authentication Method	Authentication method used for accessing the cluster: • **SIMPLE**: Select this for non-security mode. • **KERBEROS**: Select this for security mode.	KERBEROS

Parameter	Description	Example Value
Run Mode	Run mode of the HDFS link. The options are as follows: • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **STANDALONE**: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select **STANDALONE** or configure different agents. Note: The STANDALONE mode is used to solve the version conflict problem. If the connector versions of the source and destination ends of the same link are different, a JAR file conflict occurs. In this case, you need to place the source or destination end in the STANDALONE process to prevent the migration failure caused by the conflict. • **Agent**: The link instance runs on an agent.	STANDALONE
Agent	Click **Select** and select the agent created in **Connecting to an Agent**. This parameter is displayed when **Run Mode** is set to **Agent**.	-
Use Cluster Config	You can use the cluster configuration to simplify parameter settings for the Hadoop connection.	No
Cluster Config Name	This parameter is valid only when **Use Cluster Config** is set to **Yes**. Select a cluster configuration that has been created. For details, see **Managing Cluster Configurations**.	hdfs_01

Click **Show Advanced Attributes**, and then click **Add** to add configuration attributes of other clients. The name and value of each attribute must be configured. You can click **Delete** to delete no longer used attributes.

Apache HDFS

When connecting CDM to HDFS of Apache Hadoop, configure the parameters as described in **Table 3-22**.

Table 3-22 Apache HDFS link parameters

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	hadoop_hdfs_link
URI	NameNode URI You can enter **hdfs://*IP address of the NameNode instance*:8020**.	hdfs://*IP*:8020
Authentication Method	Authentication method used for accessing the cluster: ● **SIMPLE**: Select this for non-security mode. ● **KERBEROS**: Select this for security mode.	KERBEROS
Principal	When **Authentication Method** is set to **KERBEROS**, this parameter is mandatory. It is the username in the Kerberos security mode and can be obtained from the Hadoop administrator. The value of this parameter must be the same as that in the Keytab file.	-
Keytab File	When **Authentication Method** is set to **KERBEROS**, a Keytab file must be uploaded. The Keytab file is an authentication credential and can be obtained from the Hadoop administrator. Before obtaining the keytab file, you need to change the password of this user at least once in the cluster. Otherwise, the downloaded keytab file may be unavailable. After a user password is changed, the exported keytab file becomes invalid, and you need to export a keytab file again.	-

Parameter	Description	Example Value
Run Mode	Run mode of the HDFS link. The options are as follows: • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **STANDALONE**: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select **STANDALONE** or configure different agents. Note: The STANDALONE mode is used to solve the version conflict problem. If the connector versions of the source and destination ends of the same link are different, a JAR file conflict occurs. In this case, you need to place the source or destination end in the STANDALONE process to prevent the migration failure caused by the conflict. • **Agent**: The link instance runs on an agent. For Apache HDFS, you can select **Agent** only if **Authentication Method** is set to **SIMPLE**.	STANDALONE
IP and Host Name Mapping	This parameter is used only when **Run Mode** is set to **EMBEDDED** or **STANDALONE**. If the HDFS configuration file uses the host name, configure the mapping between the IP address and host name. Separate the IP addresses and host names by spaces and mappings by semicolons (;), carriage returns, or line feeds.	10.1.6.9 hostname01 10.2.7.9 hostname02
Agent	If **Run Mode** is set to **Agent**, click **Select** and select the agent created in **Connecting to an Agent**.	-
Use Cluster Config	You can use the cluster configuration to simplify parameter settings for the Hadoop connection.	No
Cluster Config Name	This parameter is valid only when **Use Cluster Config** is set to **Yes**. Select a cluster configuration that has been created. For details, see **Managing Cluster Configurations**.	hdfs_01

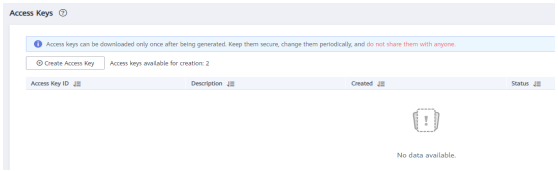
3.14 Link to OBS

When connecting CDM to the destination OBS bucket, you need to add the read and write permissions to the destination OBS bucket, and file authentication is not required.

When connecting CDM to OBS, configure the parameters as described in **Table 3-23**.

Table 3-23 Parameter description

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	obs_link
OBS Endpoint	You can obtain the endpoint by either of the following means: • To obtain the endpoint of an OBS bucket, go to the OBS console and click the bucket name to go to its details page. • 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**.	-
Port	Data transmission port. The HTTPS port number is 443 and the HTTP port number is 80.	443
OBS Bucket Type	Select a value from the drop-down list, generally, **Object Storage**.	Object Storage
AK	AK and SK are used to log in to the OBS server. You need to create an access key for the current account and obtain an AK/SK pair. To obtain an access key, perform the following steps: 1. Log in to the management console, move the cursor to the username in the upper right corner, and select **My Credentials** from the drop-down list. 2. On the **My Credentials** page, choose **Access Keys**, and click **Create Access Key**. See **Figure 3-12**.	-

Parameter	Description	Example Value
SK	**Figure 3-12** Clicking Create Access Key  3. Click **OK** and save the access key file as prompted. The access key file will be saved to your browser's configured download location. Open the **credentials.csv** file to view **Access Key Id** and **Secret Access Key**. **NOTE** • Only two access keys can be added for each user. • To ensure access key security, the access key is automatically downloaded only when it is generated for the first time and cannot be obtained from the management console later. Keep them properly.	-

3.15 Link to an FTP or SFTP Server

The FTP/SFTP link is used to migrate files from the on-premises file server or ECS to OBS or a database.

NOTE

Only FTP servers running Linux are supported.

When connecting CDM to an FTP or SFTP server, configure the parameters as described in **Table 3-24**.

Table 3-24 Parameter description

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	ftp_link
Host Name/IP Address	Host name or IP address of the FTP or SFTP server	ftp.apache.org
Port	Port number of the FTP or SFTP server, which is 21 by default	21

Parameter	Description	Example Value
Username	Username used for logging in to the FTP or SFTP server	cdm
Password	Password used for logging in to the FTP or SFTP server	-

3.16 Link to Redis/DCS

The Redis link is applicable to data migration of Redis created in the local data center or ECS. It is used to load data in the database or files to Redis.

The DCS link is used to load data from databases or files to Distributed Cache Service (DCS) on HUAWEI CLOUD. You are advised to use backup and restoration to migrate data from the third-party cloud Redis services to DCS.

When connecting CDM to an on-premises Redis database or DCS, configure the parameters as described in **Table 3-25**.

Table 3-25 Parameter description

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	redis_link
Redis Deployment Method	Two deployment methods are available: • **Single**: installation on a single-node system • **Cluster**: installation on a cluster • **Proxy**: installation using a proxy	Single
Redis Server List	List of MongoDB server addresses. Enter each address in the format of *IP address or domain name of the database server*:*port number*, and separate the entered addresses with semicolons (;).	192.168.0.1:7300;192.168.0.2:7301
Password	Password used for logging in to Redis	-
Redis Database Index	Index ID of a Redis database A Redis database is similar to a relational database. The total number of Redis databases can be set in the Redis configuration file. By default, there are 16 Redis databases. The database names are integers ranging from 0 to 15 instead of character strings.	0

3.17 Link to DDS

The DDS link is used to synchronize data from Document Database Service (DDS) on HUAWEI CLOUD to a big data platform.

When connecting CDM to DDS, configure the parameters as described in **Table 3-26**.

Table 3-26 DDS link parameters

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	dds_link
Server List	List of server addresses. Enter each address in the format of *IP address or domain name of the database server*:*port number*, and separate the entered addresses with semicolons (;).	192.168.0.1:7300;192.168.0.2:7301
Database Name	Name of the DDS database to be connected	DB_dds
Username	Username used for logging in to DDS	cdm
Password	Password used for logging in to DDS	-

3.18 Link to CloudTable

When connecting CDM to CloudTable, configure the parameters as described in **Table 3-27**.

Table 3-27 Parameter description

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	cloudtable_link
ZK Link	Obtain this parameter value from the cluster management page of CloudTable.	cloudtable-cdm-zk1.cloudtable.com:2181,cloudtable-cdm-zk2.cloudtable.com:2181

Parameter	Description	Example Value
IAM Authentication	If IAM authentication is enabled for the CloudTable cluster to be connected, set this parameter to **Yes**. Otherwise, set this to **No**. If you select **Yes**, enter the username, AK, and SK.	No
Username	Username used for accessing the CloudTable cluster	admin
AK	AK for accessing the CloudTable cluster. You need to create an access key for the current account and obtain an AK/SK pair.	-
SK	SK for accessing the CloudTable cluster. You need to create an access key for the current account and obtain an AK/SK pair.	-
Use Cluster Config	You can use the cluster configuration to simplify parameter settings for the Hadoop connection.	No
Cluster Config Name	This parameter is valid only when **Use Cluster Config** is set to **Yes**. Select a cluster configuration that has been created. For details, see **Managing Cluster Configurations**.	hadoop_01

Click **Show Advanced Attributes**, and then click **Add** to add configuration attributes of other clients. The name and value of each attribute must be configured. You can click **Delete** to delete no longer used attributes.

3.19 Link to MongoDB

This link is used to transfer data from a third-party cloud MongoDB service or MongoDB created in the on-premises data center or ECS to a big data platform.

When connecting CDM to an on-premises MongoDB database, configure the parameters as described in **Table 3-28**.

Table 3-28 Parameter description

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	mongodb_link

Parameter	Description	Example Value
Server List	List of MongoDB server addresses. Enter each address in the format of *IP address or domain name of the database server*:*port number*, and separate the entered addresses with semicolons (;).	192.168.0.1:7300;192.168.0.2:7301
Database Name	Name of the MongoDB database to be connected	DB_mongodb
Username	Username for logging in to MongoDB	cdm
Password	Password for logging in to MongoDB	-

3.20 Link to Cassandra

Table 3-29 Parameter description

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	mongodb_link
Service node	An address of one node or addresses of multiple nodes. Separate addresses with semicolons (;). You are advised to configure multiple nodes at a time.	192.168.0.1;192.168.0.2
Port	Port number of the Cassandra node to be connected.	9042
Username	User name for connecting to Cassandra.	cdm
Password	Password for connecting to Cassandra.	-
Connection timeout duration	(Optional) Displayed when you click **Show Advanced Attributes**. Connection timeout interval, in seconds.	5
Read timeout duration	(Optional) Displayed when you click **Show Advanced Attributes**. Read timeout interval, in seconds. If the value is less than or equal to 0, no timeout occurs.	12

3.21 Link to Kafka

MRS Kafka

When connecting CDM to Kafka of MRS, configure the parameters as described in **Table 3-30**.

Table 3-30 MRS Kafka link parameters

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	kafka_link
Manager IP	Floating IP address of MRS Manager. Click **Select** next to the **Manager IP** text box to select an MRS cluster. CDM automatically fills in the authentication information.	-
Username	Username used for logging in to MRS Manager 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.	-
Password	Password used for logging in to MRS Manager	-

Parameter	Description	Example Value
Authentication Method	Authentication method used for accessing MRS • **SIMPLE**: for non-security mode • **KERBEROS**: for security mode	Yes

Click **Show Advanced Attributes**, and then click **Add** to add configuration attributes of other clients. The name and value of each attribute must be configured. You can click **Delete** to delete no longer used attributes.

Apache Kafka

The Apache Kafka link is applicable to data migration of the third-party Hadoop in the local data center or ECS. You must use Direct Connect to connect to Hadoop in the local data center.

When connecting CDM to Kafka of Apache Hadoop, configure the parameters as described in **Table 3-31**.

Table 3-31 Parameter description

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	kafka_link
Kafka broker	IP address and port number of the Kafka broker	192.168.1.1:9092

Click **Show Advanced Attributes**, and then click **Add** to add configuration attributes of other clients. The name and value of each attribute must be configured. You can click **Delete** to delete no longer used attributes.

3.22 Link to DMS Kafka

When connecting CDM to DMS Kafka, configure the parameters as described in **Table 3-32**.

Table 3-32 Parameter description

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	dms_link
Service Type	DMS Kafka edition. Currently, only the Platinum edition is available.	Platinum
Kafka Broker	Address of a Kafka premium instance. The format is host:port.	-
Kafka SASL_SSL	Whether to enable SSL authentication when a client connects to a Kafka premium instance. If Kafka SASL_SSL is enabled, data will be encrypted before transmission for higher security, but performance will suffer.	Yes
Username	Username for connecting to DMS Kafka. This parameter is displayed when **Kafka SASL_SSL** is enabled.	-
Password	Password for connecting to DMS Kafka. This parameter is displayed when **Kafka SASL_SSL** is enabled.	-

3.23 Link to Elasticsearch/CSS

Elasticsearch

The Elasticsearch link is applicable to data migration of Elasticsearch services and Elasticsearch created in the local data center or ECS.

NOTE

The Elasticsearch connector supports only the non-security mode.

When connecting CDM to Elasticsearch, configure the parameters as described in **Table 3-33**.

Table 3-33 Parameter description

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	css_link

Parameter	Description	Example Value
Elasticsearch Server List	IP addresses or domain names (including the port numbers) of one or more Elasticsearch servers. The format is *ip:port*. Use semicolons (;) to separate multiple IP addresses.	192.168.0.1:9200 ; 192.168.0.2:9200

CSS

The Cloud Search Service (CSS) link is used to migrate log files or database records to the Elasticsearch engine for search and analysis.

When connecting CDM to CSS, configure the parameters as described in **Table 3-34**.

Table 3-34 Parameter description

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	css_link
Elasticsearch Server List	IP addresses or domain names (including the port numbers) of one or more Elasticsearch servers. The format is *ip:port*. Use semicolons (;) to separate multiple IP addresses.	192.168.0.1:9200 ; 192.168.0.2:9200
Security Mode Authentication	Whether to enable security mode. If **Security Mode** has been enabled for the CSS cluster to be connected, set this parameter to **Yes**. Otherwise, set this to **No**.	Yes
Username	This parameter is displayed when **Security Mode Authentication** is set to **Yes**. It indicates the username used for connecting to CSS.	admin
Password	This parameter is displayed when **Security Mode Authentication** is set to **Yes**. It indicates the password used for connecting to CSS.	-
HTTPS Access	This parameter is displayed when **Security Mode Authentication** is set to **Yes**. This parameter specifies whether to enable HTTPS access. HTTPS access is more secure than HTTP access.	Yes

4 Managing Jobs

4.1 Table/File Migration Jobs

Scenario

CDM supports table and file migration between homogeneous or heterogeneous data sources. For details about supported data sources, see **Data Sources Supported by Table/File Migration**.

Constraints

- The dirty data recording function depends on OBS.
- The JSON file of a job to be imported cannot exceed 1 MB.

Prerequisites

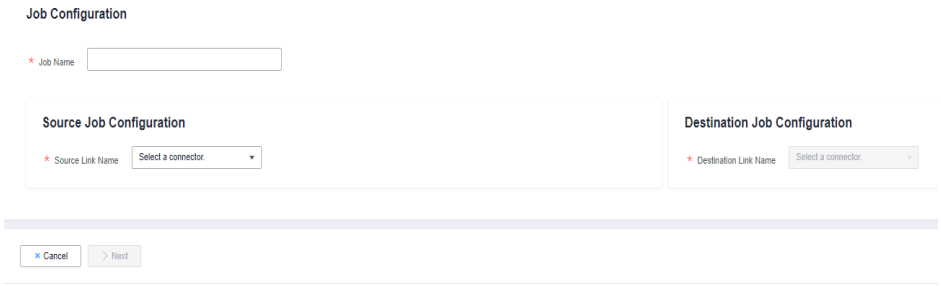
- You have created links based on the instructions in **Creating Links**.
- The CDM cluster can communicate with the data source.

Procedure

Step 1 Log in to the management console and choose **Service List** > **Cloud Data Migration**. In the left navigation pane, choose **Cluster Management**. Locate the target cluster and click **Job Management**.

Step 2 Choose **Table/File Migration** > **Create Job**. The page for configuring the job is displayed.

Figure 4-1 Creating a migration job

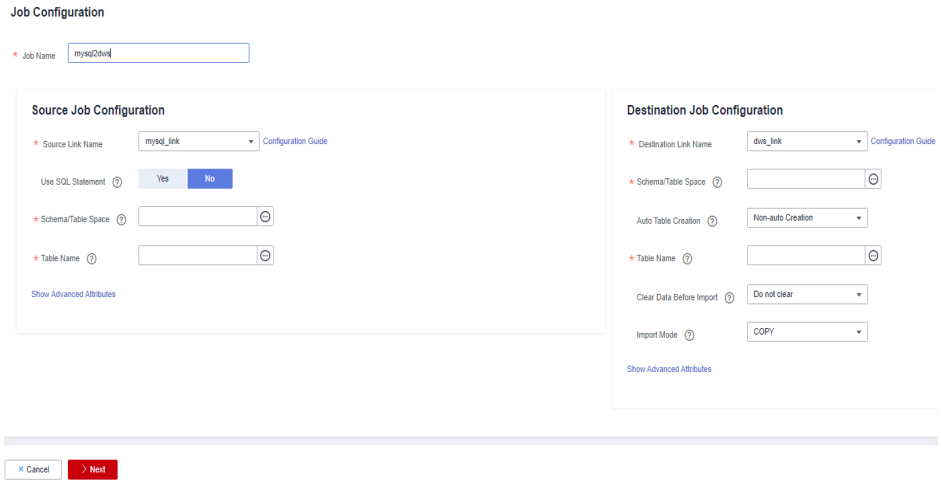


Step 3 Select the source and destination links.

- **Job Name**: Enter a string consisting of 1 to 240 characters. The name can contain digits, letters, hyphens (-), underscores (_), and periods (.), and cannot start with a hyphen (-) or period (.). An example value is **oracle2obs_t**.
- **Source Link Name**: Select the data source from which data will be exported.
- **Destination Link Name**: Select the data source to which data will be imported.

Step 4 Configure the source link parameters. **Figure 4-2** shows the job configurations for migrating MySQL to DWS.

Figure 4-2 Creating a job



The parameters vary with data sources. For details about the job parameters of other types of data sources, see **Table 4-1** and **Table 4-2**.

Table 4-1 Source link parameter description

Migration Source	Description	Parameter Settings
OBS	Data can be extracted in CSV, JSON, or binary format. Data extracted in binary format is free from file resolution, which ensures high performance and is more suitable for file migration.	For details, see **From OBS**.
• MRS HDFS • FusionInsight HDFS • Apache HDFS	HDFS data can be exported in CSV, Parquet, or binary format and can be compressed in multiple formats.	For details, see **From HDFS**.
• MRS HBase • FusionInsight HBase • Apache HBase • CloudTable Service	Data can be exported from MRS, FusionInsight HD, open source Apache Hadoop HBase, or CloudTable. You need to know all column families and field names of HBase tables.	For details, see **From HBase/CloudTable**.
• MRS Hive • FusionInsight Hive • Apache Hive	Data can be exported from Hive through the JDBC API. If the data source is Hive, CDM will automatically partition data using the Hive data partitioning file.	For details, see **From Hive**.
DLI	Data can be exported from DLI.	For details, see **From DLI**.
• FTP • SFTP	FTP and SFTP data can be exported in CSV, JSON, or binary format.	For details, see **From FTP/SFTP**.

Migration Source	Description	Parameter Settings
• HTTP	These connectors are used to read files with an HTTP/HTTPS URL, such as reading public files on the third-party object storage system and web disks. Currently, data can only be exported from the HTTP URLs.	For details, see **From HTTP**.
• Data Warehouse Service • RDS for MySQL • RDS for SQL Server • RDS for PostgreSQL	Data can be exported from the cloud database services.	When data is exported from these data sources, CDM uses the JDBC API to extract data. The job parameters for the migration source are the same. For details, see **From a Common Relational Database**.
• FusionInsight LibrA	Data can be exported from FusionInsight LibrA.	
• MySQL • PostgreSQL • Oracle • Microsoft SQL Server • SAP HANA • MyCAT • Database Sharding	The non-cloud databases can be those created in the on-premises data center or deployed on ECSs, or database services on the third-party clouds.	
• MongoDB • Document Database Service	Data can be exported from MongoDB or DDS.	For details, see **From MongoDB/DDS**.
Redis	Data can be exported from open source Redis.	For details, see **From Redis**.
• Apache Kafka • DMS Kafka • MRS Kafka	Data can only be exported to Cloud Search Service (CSS).	For details, see **From Kafka/DMS Kafka**.
• Cloud Search Service • Elasticsearch	Data can be exported from CSS or Elasticsearch.	For details, see **From Elasticsearch or CSS**.

Step 5 Configure job parameters for the migration destination based on **Table 4-2**.

Table 4-2 Parameter description

Migration Destination	Description	Parameter Settings
OBS	Files (even in a large volume) can be batch migrated to OBS in CSV or binary format.	For details, see **To OBS**.
MRS HDFS	You can select a compression format when importing data to HDFS.	For details, see **To HDFS**.
MRS HBase CloudTable Service	Data can be imported to HBase. The compression algorithm can be set when a new HBase table is created.	For details, see **To HBase/ CloudTable**.
MRS Hive	Data can be rapidly imported to MRS Hive.	For details, see **To Hive**.
DLI	Data can be imported to DLI.	For details, see **To DLI**.
• Data Warehouse Service • RDS for MySQL • RDS for SQL Server • RDS for PostgreSQL	Data can be imported to cloud database services.	For details about how to use the JDBC API to import data, see **To a Common Relational Database**.
Document Database Service	Data can be imported to the DDS but cannot be imported to the local MongoDB.	For details, see **To DDS**.
Distributed Cache Service	Data can be imported to DCS in the **String** or **Hashmap** value type. Data cannot be imported to the local Redis.	For details, see **To DCS**.
Cloud Search Service (CSS)	Data can be imported to CSS.	For details, see **To CSS**.

Step 6 After the parameters are configured, click **Next**. The **Map Field** tab page is displayed.

If files are migrated between FTP, SFTP, HDFS, and OBS and the migration source's **File Format** is set to **Binary**, files will be directly transferred, free from field mapping.

In other scenarios, CDM automatically maps fields of the source table and the destination table. You need to check whether the mapping and time format are correct. For example, check whether the source field type can be converted into the destination field type.

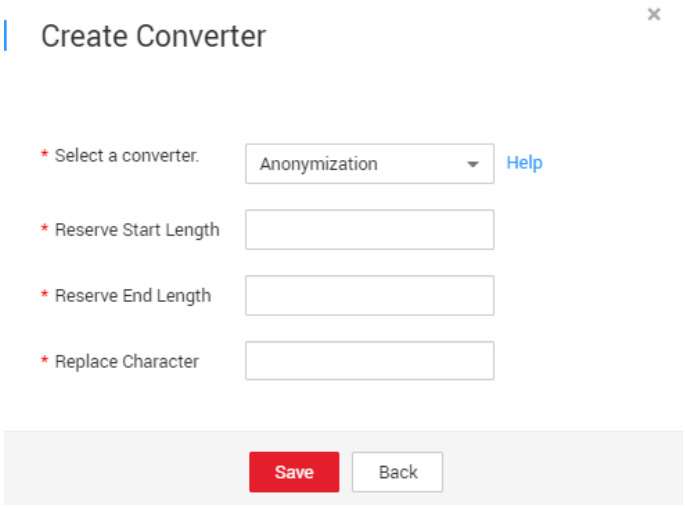
Figure 4-3 Field mapping

NOTE

- If the fields from the source and destination do not match, you can drag the fields to make adjustments.
- On the **Map Field** tab page, if CDM fails to obtain all columns by obtaining sample values (for example, when data is exported from HBase, CloudTable, or MongoDB, there is a high probability that CDM failed to obtain all columns), you can click and select **Add a new field** to add new fields to ensure that the data imported to the migration destination is complete.
- If the data is imported to DWS, you need to select the distribution columns in the destination fields. You are advised to select the distribution columns according to the following principles:
  1. Use the primary key as the distribution column.
  2. If multiple data segments are combined as primary keys, specify all primary keys as the distribution column.
  3. In the scenario where no primary key is available, if no distribution column is selected, DWS uses the first column as the distribution column by default. As a result, data skew risks exist.

Step 7 CDM supports field conversion. Click and then click **Create Converter**.

Figure 4-4 Creating a converter



CDM supports the following converters:

- **Anonymization**: hides key data in the character string.

  For example, if you want to convert **12345678910** to **123****8910**, configure the parameters as follows:

  – Set **Reserve Start Length** to **3**.
  – Set **Reserve End Length** to **4**.
  – Set **Replace Character** to *****.
- **Trim** automatically deletes the spaces before and after the character string.
- **Reverse string** automatically reverses a character string. For example, reverse **ABC** into **CBA**.
- **Replace string** replaces the specified character string.
- **Expression conversion** uses the JSP expression language (EL) to convert the current field or a row of data. For details, see **Converting Fields**.
- **Remove line break** deletes the newline characters, such as \n, \r, and \r\n from the field.

Step 8 Click **Next**, set job parameters, and click **Show Advanced Attributes** to display and configure optional parameters.

Figure 4-5 Task parameters

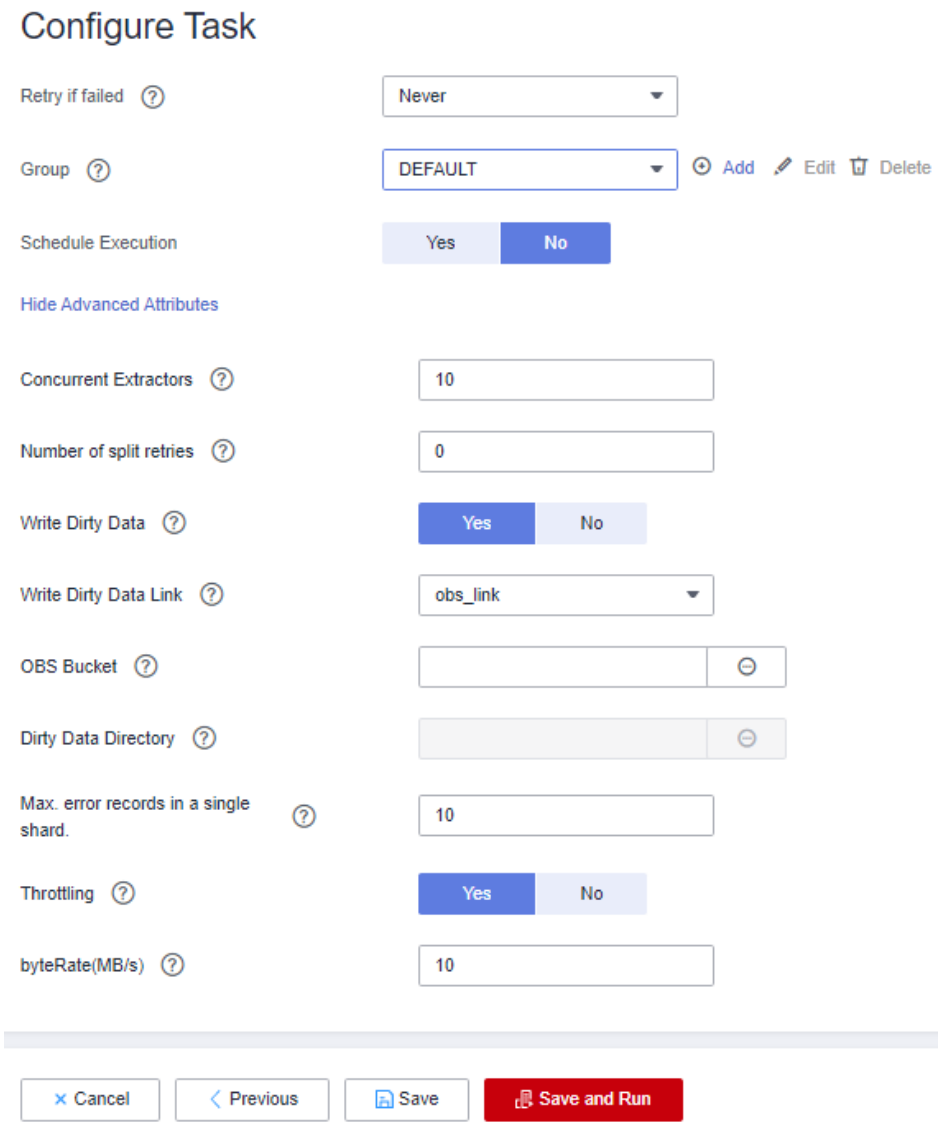


Table 4-3 describes related parameters.

Table 4-3 Parameter description

Parameter	Description	Example Value
Retry upon Failure	You can select **Retry 3 times** or **Never**. You are advised to configure automatic retry for only file migration jobs or database migration jobs with **Import to Staging Table** enabled to avoid data inconsistency caused by repeated data writes. **NOTE** If you want to set parameters in DataArts Studio DataArts Factory to schedule the CDM migration job, do not configure this parameter. Instead, set parameter **Retry upon Failure** for the CDM node in DataArts Factory.	Never
Job	Select a group where the job resides. The default group is **DEFAULT**. On the **Job Management** page, jobs can be displayed, started, or exported by group.	DEFAULT
Schedule Execution	If you select **Yes**, you can set the start time, cycle, and validity period of a job. For details, see **Scheduling Job Execution**. **NOTE** If you use DataArts Studio DataArts Factory to schedule the CDM migration job and configure this parameter, both configurations take effect. To ensure unified service logic and avoid scheduling conflicts, enable job scheduling in DataArts Factory and do not configure a scheduled task for the job in DataArts Migration.	No

Parameter	Description	Example Value
Concurrent Extractors	Configure the number of tasks to be split from a CDM 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. By setting appropriate values for this parameter and the **Maximum Concurrent Extractors** parameter, you can accelerate migration. Configure the number of concurrent extractors based on the following rules: 1. 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. 2. 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	1

Parameter	Description	Example Value
	that data be extracted in a single thread. 3. 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**.	
Concurrent Loaders	Number of Loaders to be concurrently executed This parameter is displayed only when HBase or Hive serves as the destination data source.	3
Number of split retries	Number of retries when a split fails to be executed. Value **0** indicates that no retry will be performed.	0
Write Dirty Data	Whether to record dirty data. By default, this parameter is set to **No**. Dirty data in CDM refers to the data in invalid format. If the source data contains dirty data, you are advised to enable this function. Otherwise, the migration job may fail.	Yes
Write Dirty Data Link	This parameter is displayed only when **Write Dirty Data** is set to **Yes**. Only links to OBS support dirty data writes.	obs_link
OBS Bucket	This parameter is displayed only when **Write Dirty Data Link** is a link to OBS. Name of the OBS bucket to which the dirty data will be written.	dirtydata

Parameter	Description	Example Value
Dirty Data Directory	This parameter is displayed only when **Write Dirty Data** is set to **Yes**. Dirty data is stored in the directory for storing dirty data on OBS. Dirty data is saved only when this parameter is configured. You can go to this directory to query data that fails to be processed or is filtered out during job execution, and check the source data that does not meet conversion or cleaning rules.	/user/dirtydir
Max. Error Records in a Single Shard	This parameter is displayed only when **Write Dirty Data** is set to **Yes**. When the number of error records of a single map exceeds the upper limit, the job will automatically terminate and the imported data cannot be rolled back. You are advised to use a temporary table as the destination table. After the data is imported, rename the table or combine it into the final data table.	0

Step 9 Click **Save** or **Save and Run**. On the page displayed, you can view the job status.

NOTE

The job status can be **New**, **Pending**, **Booting**, **Running**, **Failed**, or **Succeeded**.

Pending indicates that the job is waiting to be scheduled by the system, and **Booting** indicates that the data to be migrated is being analyzed.

----End

4.2 Creating an Entire Database Migration Job

Scenario

CDM supports entire DB migration between homogeneous and heterogeneous data sources. The migration principles are the same as those in **Table/File**

Migration Jobs. Each type of Elasticsearch, each key prefix of Redis, or each collection of MongoDB can be executed concurrently as a subtask.

Supported Data Sources in Entire DB Migration lists the data sources supporting entire database migration.

Field Mapping in Automatic Table Creation

CDM automatically creates tables at the destination during database migration. **Figure 4-6** describes the field mapping between the DWS tables created by CDM and source tables. For example, if you use CDM to migrate the Oracle database to DWS, CDM automatically creates a table on DWS and maps the **NUMBER(3,0)** field of the Oracle database to the **SMALLINT** field of DWS.

Figure 4-6 Field mapping in automatic table creation on DWS

Source Database							Destination Database
Oracle	MySQL	SQL Server	PostgreSQL	Db2	GaussDB	SAP HANA	DWS
NUMBER(p,0) (p=3 or p=5)	SMALLINT,TINYINT	SMALLINT,TINYINT	SMALLINT	DECIMAL	SMALLINT	SMALLINT,TINYINT	SMALLINT
NUMBER(10,0)	INT	INT	INTEGER	INT	INTEGER	INTEGER	INTEGER
NUMBER(19,0)	BIGINT	BIGINT	BIGINT	DECIMAL	BIGINT	BIGINT	BIGINT
None	None	None	OID	None	OID	CHAR(128)	OID
NUMBER(p,s) (0 < p <= 38)	DECIMAL(p,s) (0 < p <= 65)	DECIMAL(p,s) (0 < p <= 30)	NUMERIC(p,s) (p <= 1000)	DECIMAL	NUMERIC(p,s) (p <= 1000)	DECIMAL(p,s) (0 < p <= 38)	NUMERIC(p,s) (p <= 1000)
RAW	BINARY	BINARY	BYTEA	BINARY	BYTEA	BINARY	BYTEA
CHAR	CHAR	CHAR	CHAR	CHAR	CHAR	CHAR(p) (p <= 2000)	CHAR
NCHAR	NCHAR	NCHAR	NCHAR	NCHAR	NCHAR	NCHAR(p) (p <= 5000)	NCHAR
DATE	DATE	DATE	DATE	DATE	DATE	DATE	DATE
DATE	DATETIME	DATETIME2	TIMESTAMP	TIMESTAMP	TIMESTAMP	TIMESTAMP	TIMESTAMP
VARCHAR2(p) (p <= 4000)	VARCHAR	VARCHAR(p) (if p >= 8000 p=max)	VARCHAR(p) (p <= 10485760)	VARCHAR	VARCHAR(p) (p <= 10485760)	VARCHAR(p) (p <= 5000)	VARCHAR(p) (p <= 10485760)
FLOAT	DOUBLE	FLOAT	DOUBLE PRECISION	FLOAT	DOUBLE PRECISION	DOUBLE	DOUBLE PRECISION
FLOAT	REAL	FLOAT	REAL	FLOAT	REAL	REAL	REAL
CLOB	TEXT	TEXT	TEXT	TEXT	TEXT	CLOB	TEXT
DATE	None	TIME	TIME	TIME	TIME	TIME	TIME
BOOLEAN	None	TIME	BOOLEAN	BOOLEAN	BOOLEAN	BOOLEAN	BOOLEAN

Prerequisites

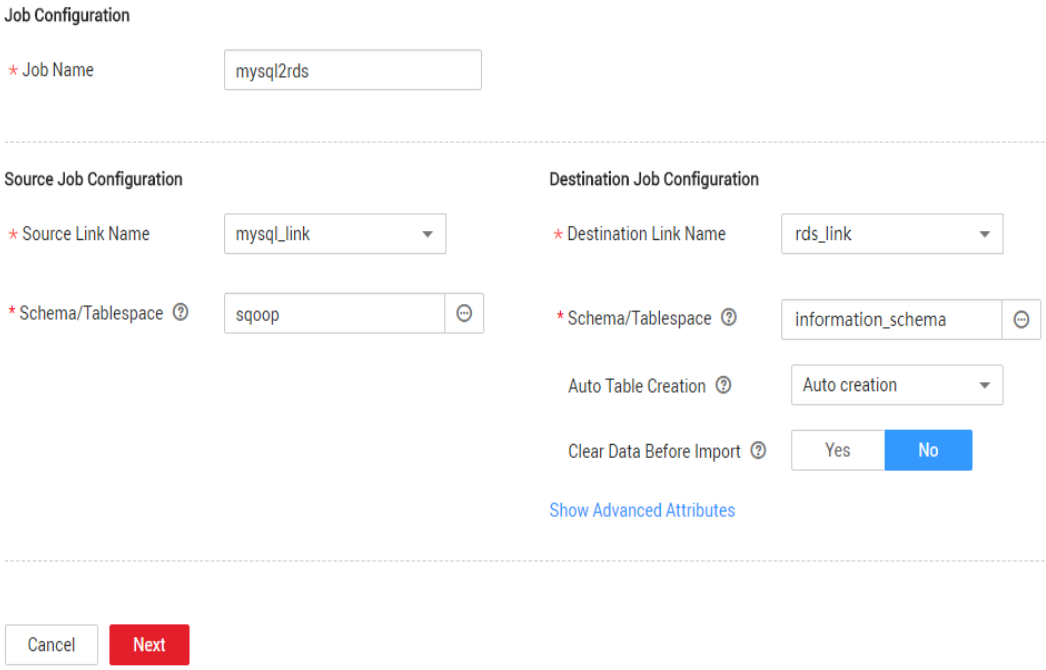
- You have created links according to **Creating Links**.
- The CDM cluster can communicate with the data source.

Procedure

Step 1 Log in to the management console and choose **Service List** > **Cloud Data Migration**. In the left navigation pane, choose **Cluster Management**. Locate the target cluster and click **Job Management**.

Step 2 Choose **Entire DB Migration** > **Create Job**. The page for configuring the job is displayed.

Figure 4-7 Creating an entire DB migration job



Step 3 Configure the related parameters of the source database according to **Table 4-4**.

Table 4-4 Parameter description

Source Database	Parameter	Description	Example Value
• DWS • FusionInsight LibrA • MySQL • PostgreSQL • SQL Server • Oracle • SAP HANA • MyCAT	Schema/ Tablespace	Name of the schema or tablespace from which data will be extracted. This parameter is displayed when **Use SQL Statement** is set to **No**. Click the icon next to the text box to go to the page for selecting a schema or directly enter a schema or tablespace. If the desired schema or tablespace is not displayed, check whether the login account has the permissions required to query metadata.	schema

Source Database	Parameter	Description	Example Value
	WHERE Clause	WHERE clause used to specify the tables to be extracted. This parameter applies to all subtables in the entire DB migration. If this parameter is not set, the entire table is extracted. If the table to be migrated does not contain the fields specified by the WHERE clause, the migration will fail. You can set a date macro variable to extract data generated on a specific date. For details, see **Incremental Migration of Relational Databases**.	age > 18 and age <= 60
	Null in Partition Column	Whether a partition field can be null	Yes
Hive	Database Name	Name of the database to be migrated. The user configured in the source link must have the permission to read the database.	hivedb
HBase CloudTable	Start Time	Start time (included). The format is *yyyy-MM-dd hh:mm:ss*. The dateformat time macro variable function is supported. Examples: **2017-12-31 20:00:00**, **${dateformat(yyyy-MM-dd, -1, DAY)} 02:00:00**, and **${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}**	-

Source Database	Parameter	Description	Example Value
	End Time	End time (excluded) The format is *yyyy-MM-dd hh:mm:ss*. The dateformat time macro variable function is supported. Examples: **2018-01-01 20:00:00**, **${dateformat(yyyy-MM-dd, -1, DAY)} 02:00:00**, and **${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}**	-
Redis	Key Filter Character	Filter character used to determine the keys to be migrated For example, if the value of this parameter is **a***, all asterisks (*) will be migrated.	-
DDS MongoDB	Database Name	Name of the database from which data is to be migrated. The user configured in the source link must have the permission to read the database.	mongodb
	Query Filter	Filter used to match documents. Example: **{HTTPStatusCode: {$gt:"400", $lt:"500"},HTTPMethod:"GET"}**	-

Source Database	Parameter	Description	Example Value
Elasticsearch CSS	Index	Index of the data to be extracted. The value can be a wildcard character. Multiple indexes that meet the wildcard condition can be migrated at a time. For example, if this parameter is set to **cdm***, CDM migrates all indexes starting with **cdm**, such as **cdm01**, **cdmB3**, **cdm_45** and so on. If multiple indexes are migrated at the same time, **Index** cannot be configured at the migration destination.	cdm*

Step 4 Configure the related parameters, from **Table 4-5**, for the destination cloud service.

Table 4-5 Destination job parameters

Source Database	Parameter	Description	Example Value
• DWS • FusionInsight LibrA • MySQL • PostgreSQL • SQL Server	-	For details about the destination job parameters required for entire DB migration to a relational database, see **To a Common Relational Database**.	schema
MRS HIVE	-	For details about the destination job parameters required for entire DB migration to MRS HIVE, see **To Hive**.	hivedb
MRS HBase CloudTable	-	For details about the destination job parameters required for entire DB migration to MRS HBase or CloudTable, see **To HBase/ CloudTable**.	Yes

Source Database	Parameter	Description	Example Value
MRS HDFS	-	For details about the destination job parameters required for entire DB migration to MRS HDFS, see **To HDFS**.	-
OBS	-	For details about the destination job parameters required for entire database migration to OBS, see **To OBS**.	-
DCS	-	For details about the destination job parameters required for entire database migration to DCS, see **To DCS**.	-
DDS	Database Name	Name of the database from which data is to be migrated. The user configured in the source link must have the permission to read the database.	mongodb
	Migration Behavior	Select **Add** or **Replace**.	-
CSS	Index	Index of the data to be extracted. The value can be a wildcard character. Multiple indexes that meet the wildcard condition can be migrated at a time. For example, if this parameter is set to **cdm***, CDM migrates all indexes starting with **cdm**, such as **cdm01**, **cdmB3**, **cdm_45** and so on. If multiple indexes are migrated at the same time, **Index** cannot be configured at the migration destination.	cdm*

Step 5 If a relational database is migrated, after job parameters are configured, click **Next** to access the page for selecting tables. You can select the tables to be migrated to the migration destination based on your requirements.

Step 6 Click **Next** and set job parameters.

Figure 4-8 Task parameters

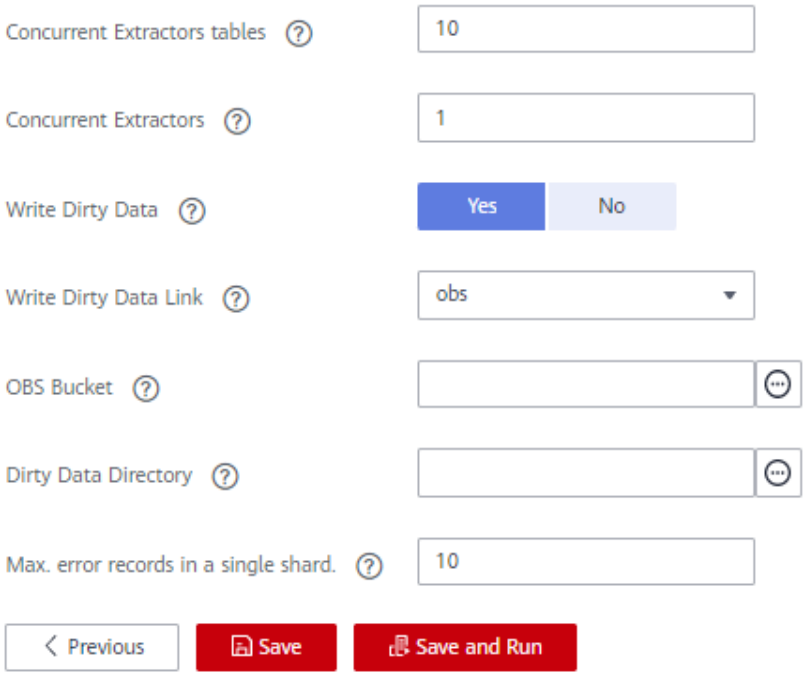


Table 4-6 describes related parameters.

Table 4-6 Task configuration parameters

Parameter	Description	Example Value
Concurrent Tables	Number of tables to be concurrently executed	3
Concurrent Extractors	Number of extractors to be concurrently executed. Generally, retain the default value.	1
Write Dirty Data	Whether to record dirty data. By default, this parameter is set to **No**.	Yes
Write Dirty Data Link	This parameter is only displayed when **Write Dirty Data** is set to **Yes**. Only links to OBS support dirty data writes.	obs_link
OBS Bucket	This parameter is only displayed when **Write Dirty Data Link** is a link to OBS. Name of the OBS bucket to which the dirty data will be written.	dirtydata

Parameter	Description	Example Value
Dirty Data Directory	This parameter is only displayed when **Write Dirty Data** is set to **Yes**. Directory for storing dirty data on OBS. Dirty data is saved only when this parameter is configured. You can go to this directory to query data that fails to be processed or is filtered out during job execution, and check the source data that does not meet conversion or cleaning rules.	/user/dirtydir
Max. Error Records in a Single Shard	This parameter is only displayed when **Write Dirty Data** is set to **Yes**. When the number of error records of a single map exceeds the upper limit, the job will automatically terminate and the imported data cannot be rolled back. You are advised to use a temporary table as the destination table. After the data is imported, rename the table or combine it into the final data table.	0

Step 7 Click **Save** or **Save and Run**.

When the job starts running, a sub-job will be generated for each table. You can click the job name to view the sub-job list.

----End

4.3 Source Job Parameters

4.3.1 From OBS

If the source link of a job is the **Link to OBS**, configure the source job parameters based on **Table 4-7**.

Advanced attributes are optional and not displayed by default. You can click **Show Advanced Attributes** to display them.

Table 4-7 Parameter description

Category	Parameter	Description	Example Value
Basic parameters	Bucket Name	Name of the bucket from which data will be migrated	BUCKET_2

Category	Parameter	Description	Example Value
	Source Directory/File	This parameter is available only when **Pull List File** is set to **No**. Directory or file path from which data will be extracted. You can enter a maximum of 50 file paths. By default, the file paths are separated by vertical bars (|). You can also customize a file separator. For details, see **Migration of a List of Files**. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job* – *Offset*) rather than (*Actual start time of the CDM job* – *Offset*).	FROM/example.csv
	File Format	Format in which CDM parses data. The options are as follows: • **CSV**: Source files will be migrated to tables after being converted to CSV format. • **Binary**: Files (even not in binary format) will be transferred directly. It is used for file copy. • **JSON**: Source files will be migrated to tables after being converted to JSON format.	CSV

Category	Parameter	Description	Example Value
	Pull List File	This parameter is displayed only when **File Format** is set to **Binary**. If the pull list file function is enabled, the content of a file (such as a .txt file) in an OBS bucket can be read as the list of files to be migrated. The content in the file must be the absolute path of the file to be migrated (rather than a directory). For example, the content is as follows: /052101/DAY20211110.data /052101/DAY20211111.data	Yes
	OBS Link of List File	This parameter is available only when **Pull List File** is set to **Yes**. You can select the OBS link where the list file is located.	OBS_test_link
	OBS Bucket of entries files	This parameter is available only when **Pull List File** is set to **Yes**. It indicates the name of the OBS bucket where the list file is located.	01
	Path/ Directory of entries files	This parameter is available only when **Pull List File** is set to **Yes**. It indicates the absolute path or directory of the list file in the OBS bucket. You are advised to select the absolute path of the file. If you select a directory, files in subdirectories can also be migrated. However, if the number of files in the directory is too large, the cluster memory may become insufficient.	/0521/ Lists.txt
	JSON Type	This parameter is displayed only when **File Format** is set to **JSON**. Type of a JSON object stored in a JSON file. The options are **JSON object** and **JSON array**.	JSON object
	JSON Reference Node	This parameter is used only when **File Format** is set to **JSON** and **JSON Type** is set to **JSON Object**. CDM parses the data under the JSON node. If the node's corresponding data is a JSON array, the system will extract data from the array in the same pattern. Use periods (.) to separate multi-layer nested JSON nodes.	data.list

Category	Parameter	Description	Example Value
Advanced attributes	Line Separator	Lind feed character in a file. By default, the system automatically identifies **\n**, **\r**, and **\r\n**. This parameter is displayed only when **File Format** is set to **CSV**.	\n
	Field Delimiter	Character used to separate fields in the file. To set the **Tab** key as the delimiter, set this parameter to **\t**. This parameter is displayed only when **File Format** is set to **CSV**.	,
	Use Quote Character	If you set this parameter to **Yes**, the field delimiters in the encircling symbol are regarded as a part of the string value. Currently, the default encircling symbol of CDM is **"**.	No
	Use RE to Separate Fields	Whether to use regular expressions to separate fields. If you set this parameter to **Yes**, **Field Delimiter** becomes invalid. This parameter is displayed only when **File Format** is set to **CSV**.	Yes
	Regular Expression	Regular expression used to separate fields. For details about regular expressions, see **Regular Expressions for Separating Semi-structured Text**.	^(\d.*\d) (\w*) \[(.*) \] ([\w\.]*) (\w.*).*
	Use First Row as Header	This parameter is displayed only when **File Format** is set to **CSV**. When you migrate a CSV file to a table, CDM writes all data to the table by default. If you set this parameter to **Yes**, CDM uses the first line of the CSV file as the heading line and does not write the line to the destination table.	No
	Encoding Type	Encoding type, for example, **UTF-8** or **GBK**. You can set the encoding type for text files only. This parameter is invalid when **File Format** is set to **Binary**.	GBK

Category	Parameter	Description	Example Value
	Compression Format	This parameter is displayed only when **File Format** is set to **CSV** or **JSON**. The options are as follows: • **NONE**: Files in all formats can be transferred. • **GZIP**: Only files in gzip format can be transferred. • **ZIP**: Only files in Zip format can be transferred. • **TAR.GZ**: Files in TAR.GZ format are transferred.	NONE
	Compressed File Suffix	This parameter is displayed when **Compression Format** is not **NONE**. This parameter specifies the extension of the files to be decompressed. The decompression operation is performed only when the file name extension is used in a batch of files. Otherwise, files are transferred in the original format. If you enter ***** or leave the parameter blank, all files are decompressed.	*
	Source File Processing Method	Operation performed on source files after the job completes. • No action • **Rename**: After the job completes, the source files are renamed by appending usernames and timestamps as suffixes to the file names. • **Delete**: After the job completes, the source files are deleted.	No action
	Start Job by Marker File	Whether to start a job by a marker file. A job is only started if there is a marker file for starting the job in the source path. If there is no marker file, the job will be suspended for a period of time specified by **Suspension Period**.	No

Category	Parameter	Description	Example Value
	Marker File	Name of the marker file for starting a job. If you specify a marker file, the migration job is executed only when the marker file exists in the source path. The marker file will not be migrated.	ok.txt
	Suspension Period	Waiting period for a marker file. If you set **Start Job by Marker File** to **Yes** but there is no marker file in the source path, the job fails when the suspension period times out. If you set this parameter to **0** and there is no marker file in the source path, the job will fail immediately. Unit: second	10
	File Separator	File separator. If you enter multiple file paths in **Source Directory/Files**, CDM uses the file separator to identify files. The default value is **|**.	|
	Filter Type	Only paths or files that meet the filtering conditions are transferred. The options are **None**, **Wildcard**, and **Regex**. For details, see **Incremental File Migration**.	Wildcard
	Directory Filter	If you set **Filter Type** to **Wildcard**, enter a wildcard character to filter paths. The paths that meet the filtering condition are migrated. You can configure multiple paths separated by commas (,). **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	*input

Category	Parameter	Description	Example Value
	File Filter	If you set **Filter Type** to **Wildcard**, you can enter a wildcard character to search for files in a specified path. The files that meet the search criteria are migrated. You can configure multiple files separated by commas (,). **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	*.csv,*.txt
	Time Filter	If you select **Yes**, files are transferred based on their modification time.	Yes
	Minimum Timestamp	If you set **Filter Type** to **Time Filter**, and specify a point in time for this parameter, only the files modified after the specified time are transferred. The time format must be *yyyy-MM-dd HH:mm:ss*. This parameter can be set to a macro variable of date and time. For example, **${timestamp(dateformat(yyyy-MM-dd HH:mm:ss,-90,DAY))}** indicates that only files generated within the latest 90 days are migrated. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	2019-06-01 00:00:00

Category	Parameter	Description	Example Value
	Maximum Timestamp	If you set **Filter Type** to **Time Filter**, and specify a point in time for this parameter, only the files modified before the specified time are transferred. The time format must be *yyyy-MM-dd HH:mm:ss*. This parameter can be set to a macro variable of date and time. For example, **${timestamp(dateformat(yyyy-MM-dd HH:mm:ss))}** indicates that only the files whose modification time is earlier than the current time are migrated. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	2019-07-01 00:00:00
	Encryption	If the source data is encrypted, CDM can decrypt the data before exporting it. Select whether to decrypt the source data and select a decryption algorithm. The options are as follows: • **NONE**: Export data without decrypting it. • **AES-256-GCM**: The AES 256-bit encryption algorithm is used to encrypt data. Currently, only the AES-256-GCM (NoPadding) encryption algorithm is supported. This parameter is used for encryption at the migration destination and decryption at the migration source. For details, see **Encryption and Decryption During File Migration**.	AES-256-GCM
	Disregard Non-existent Path or File	If this is set to **Yes**, the job can be successfully executed even if the source path does not exist.	No

Category	Parameter	Description	Example Value
	DEK	This parameter is displayed only when **Encryption** is set to **AES-256-GCM**. The key consists of 64 hexadecimal numbers and must be the same as the **DEK** configured during encryption. If the decryption and encryption keys are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	DD0AE00D FECD78BF0 51BCFDA2 5BD4E320 DB0A7AC7 5A1F3FC3D 3C56A457 DCDC1B
	IV	This parameter is displayed only when **Encryption** is set to **AES-256-GCM**. The initialization vector consists of 32 hexadecimal numbers and must be the same as the **IV** configured during encryption. If the initialization vectors are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	5C91687BA 886EDCD1 2ACBC3FF1 9A3C3F
	MD5 File Extension	This parameter is displayed only when **File Format** is set to **Binary**. This parameter is used to check whether the files extracted by CDM are consistent with source files. For details, see **MD5 Verification**.	.md5

NOTE

1. CDM supports incremental file migration (by skipping repeated files), but does not support resumable transfer.

   For example, if three files are to be migrated and the second file fails to be migrated due to the network fault. When the migration task is started again, the first file is skipped. The second file, however, cannot be migrated from the point where the fault occurs, but can only be migrated again.
2. During file migration, a single task supports millions of files. If there are too many files in the directory to be migrated, you are advised to split the files into different directories and create multiple tasks.

4.3.2 From HDFS

When the source link of a job is the **Link to HDFS**, that is, when data is exported from MRS HDFS, FusionInsight HDFS, or Apache HDFS, configure the source job parameters based on **Table 4-8**.

Table 4-8 Parameter description

Category	Parameter	Description	Example Value
Basic parameters	Source Link Name	Select a type from the drop-down list box.	hdfs_to_cdm
	Source Directory/File	This parameter is available only when **Pull List File** is set to **No**. Directory or file path from which data will be extracted. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	/user/cdm/
	File Format	File format used when transferring data. The options are as follows: • **CSV**: Source files will be migrated to tables after being converted to CSV format. • **Binary**: Files (even not in binary format) will be transferred directly. It is used for file copy. • **Parquet**: Source files will be migrated to tables after being converted to Parquet format.	CSV

Category	Parameter	Description	Example Value
	Pull List File	This parameter is displayed only when **File Format** is set to **Binary**. If the pull list file function is enabled, the content of a file (such as a .txt file) in an OBS bucket can be read as the list of files to be migrated. The content in the file must be the absolute path of the file to be migrated (rather than a directory). The following is example content: /mrs/job-properties/application_1634891604621_0014/job.properties /mrs/job-properties/application_1634891604621_0029/job.properties	Yes
	OBS Link of List File	This parameter is available only when **Pull List File** is set to **Yes**. You can select the OBS link where the list file is located.	OBS_test_link
	OBS Bucket of entries files	This parameter is available only when **Pull List File** is set to **Yes**. It indicates the name of the OBS bucket where the list file is located.	01
	Path/Directory of entries files	This parameter is available only when **Pull List File** is set to **Yes**. It indicates the absolute path or directory of the list file in the OBS bucket.	/0521/Lists.txt
Advanced attributes	Line Separator	Lind feed character in a file. By default, the system automatically identifies **\n**, **\r**, and **\r\n**. This parameter is displayed only when **File Format** is set to **CSV**.	\n
	Field Delimiter	Character used to separate fields in the file. To set the **Tab** key as the delimiter, set this parameter to **\t**. This parameter is displayed only when **File Format** is set to **CSV**.	,

Category	Parameter	Description	Example Value
	Use First Row as Header	This parameter is displayed only when **File Format** is set to **CSV**. When you migrate a CSV file to a table, CDM writes all data to the table by default. If you set this parameter to **Yes**, CDM uses the first line of the CSV file as the heading line and does not write the line to the destination table.	No
	Source File Processing Method	Operation performed on source files after the job completes. • No action • **Rename**: After the job completes, the source files are renamed by appending usernames and timestamps as suffixes to the file names. • **Delete**: After the job completes, the source files are deleted.	No action
	Start Job by Marker File	Whether to start a job by a marker file. A job is only started if there is a marker file for starting the job in the source path. If there is no marker file, the job will be suspended for a period of time specified by **Suspension Period**.	ok.txt
	Filter Type	Only paths or files that meet the filtering conditions are transferred. The options are **None**, **Wildcard**, and **Regex**. For details, see **Incremental File Migration**.	-

Category	Parameter	Description	Example Value
	Path Filter	If you set **Filter Type** to **Wildcard**, enter a wildcard character to filter paths. The paths that meet the filtering condition are migrated. You can configure multiple paths separated by commas (,). **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	*input
	File Filter	If you set **Filter Type** to **Wildcard**, you can enter a wildcard character to search for files in a specified path. The files that meet the search criteria are migrated. You can configure multiple files separated by commas (,). **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	*.csv
	Time Filter	If you select **Yes**, files are transferred based on their modification time.	Yes

Category	Parameter	Description	Example Value
	Minimum Timestamp	If you set **Filter Type** to **Time Filter**, and specify a point in time for this parameter, only the files modified after the specified time are transferred. The time format must be *yyyy-MM-dd HH:mm:ss*. This parameter can be set to a macro variable of date and time. For example, **${timestamp(dateformat(yyyy-MM-dd HH:mm:ss,-90,DAY))}** indicates that only files generated within the latest 90 days are migrated. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	2019-07-01 00:00:00

Category	Parameter	Description	Example Value
	Maximum Timestamp	If you set **Filter Type** to **Time Filter**, and specify a point in time for this parameter, only the files modified before the specified time are transferred. The time format must be *yyyy-MM-dd HH:mm:ss*. This parameter can be set to a macro variable of date and time. For example, **${timestamp(dateformat(yyyy-MM-dd HH:mm:ss))}** indicates that only the files whose modification time is earlier than the current time are migrated. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	2019-07-30 00:00:00
	Create Snapshot	If you set this parameter to **Yes**, CDM creates a snapshot for the source directory to be migrated (the snapshot cannot be created for a single file) before it reads files from HDFS. Then CDM migrates the data in the snapshot. Only the HDFS administrator can create a snapshot. After the CDM job is completed, the snapshot is deleted.	No

Category	Parameter	Description	Example Value
	Encryption	This parameter is displayed only when **File Format** is set to **Binary**. If the source data is encrypted, CDM can decrypt the data before exporting it. Select whether to decrypt the source data and select a decryption algorithm. The options are as follows: • **NONE**: Export data without decrypting it. • **AES-256-GCM**: The AES 256-bit encryption algorithm is used to encrypt data. Currently, only the AES-256-GCM (NoPadding) encryption algorithm is supported. This parameter is used for encryption at the migration destination and decryption at the migration source. For details, see **Encryption and Decryption During File Migration**.	AES-256-GCM
	DEK	This parameter is displayed only when **Encryption** is set to **AES-256-GCM**. The key consists of 64 hexadecimal numbers and must be the same as the **DEK** configured during encryption. If the decryption and encryption keys are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B

Category	Parameter	Description	Example Value
	IV	This parameter is displayed only when **Encryption** is set to **AES-256-GCM**. The initialization vector consists of 32 hexadecimal numbers and must be the same as the **IV** configured during encryption. If the initialization vectors are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	5C91687BA886EDCD12ACBC3FF19A3C3F
	MD5 File Extension	This parameter is displayed only when **File Format** is set to **Binary**. This parameter is used to check whether the files extracted by CDM are consistent with source files. For details, see **MD5 Verification**.	.md5

NOTE

HDFS supports the **UTF-8** encoding only. Retain the default value **UTF-8**.

4.3.3 From HBase/CloudTable

When the source link of a job is the **Link to HBase** or **Link to CloudTable**, that is, when data is exported from MRS HBase, FusionInsight HBase, CloudTable, or Apache HBase, configure the source job parameters based on **Table 4-9**.

NOTE

1. When you migrate data from CloudTable or HBase, CDM reads the first row of the table as an example of the field list. If the first row of data does not contain all fields of the table, you need to manually add fields.
2. Because HBase is schema-less, CDM cannot obtain the data types. If the data is stored in binary format, CDM cannot parse the data.
3. When data is exported from HBase or CloudTable, because HBase/CloudTable is schema-less storage systems, CDM requires that the source numeric fields be stored in regular decimal format rather than in binary format. For example, the value 100 needs to be stored as **100** rather than **01100100**.

Table 4-9 Parameter description

Parameter	Description	Example Value
Table Name	Name of the HBase table that data will be exported from This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	TBL_2
Column Families	(Optional) Column families to which the exported data belongs	CF1&CF2
Split Rowkey	(Optional) Whether to split a rowkey. The default value is **No**.	Yes
Rowkey Delimiter	(Optional) Delimiter used to split a rowkey. If this parameter is left empty, the rowkey will not be split.	|
Start Time	(Optional) Start time (including the value) for extracting data. The format is *yyyy-MM-dd HH:mm:ss*. Only the data generated at the specified time and later is extracted. This parameter can be set to a macro variable of date and time. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	2019-01-01 20:00:00

Parameter	Description	Example Value
End Time	(Optional) End time (excluding the value) for extracting data. The format is *yyyy-MM-dd HH:mm:ss*. Only the data generated before the time point is extracted. This parameter can be set to a macro variable of date and time. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	2019-02-01 20:00:00

4.3.4 From Hive

If the source link of a job is the **Link to Hive**, configure the source job parameters based on **Table 4-10**.

Table 4-10 Parameter description

Parameter	Description	Example Value
Database Name	Database name. Click the icon next to the text box. The dialog box for selecting the database is displayed.	default
Table Name	Hive table name. Click the icon next to the text box. The dialog box for selecting the table is displayed. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	TBL_E

Parameter	Description	Example Value
Read Mode	Two read modes are available: HDFS and JDBC. By default, the HDFS mode is used. If you do not need to use the WHERE condition to filter data or add new fields on the field mapping page, select the HDFS mode. • The HDFS mode shows good performance, but in this mode, you cannot use the WHERE condition to filter data or add new fields on the field mapping page. • The HDFS mode allows you to use the WHERE condition to filter data or add new fields on the field mapping page.	HDFS
Partition Filter Criteria	This parameter is displayed when you select the HDFS read mode and click **Show Advanced Attributes**. You can configure multiple values (separated by spaces) or a field value range. The time macro function is supported. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	• Single/ Multi-value filtering: "${dateformat(yyyyMMdd, -1, DAY)} ${dateformat(yyyyMMdd)}" • Filter by range: "${value} >= ${dateformat(yyyyMMdd, -7, DAY)} && ${value} < ${dateformat(yyyyMMdd)}"

Parameter	Description	Example Value
WHERE Clause	This parameter is displayed when you select the JDBC read mode and click **Show Advanced Attributes**. This parameter indicates the WHERE clause to be extracted. If this parameter is not set, the entire table is extracted. If the table to be migrated does not contain the fields specified by the WHERE clause, the migration will fail. You can set a date macro variable to extract data generated on a specific date. For details, see **Incremental Migration of Relational Databases**. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	age > 18 and age <= 60

NOTE

If the data source is Hive, CDM will automatically partition data using the Hive data partitioning file.

4.3.5 From DLI

If the source link of a job is the **Link to DLI**, configure the source job parameters based on **Table 4-11**.

Table 4-11 Parameter description

Parameter	Description	Example Value
Resource Queue	Resource queue to which the destination table belongs The default queue of DLI cannot be used for migration jobs. You need to create a SQL queue in DLI.	cdm
Database Name	Name of the database to which data will be written	dli
Table Name	Name of the table to which data will be written	car_detail

Parameter	Description	Example Value
Partition	Partition information. This parameter is available if **Clear Data Before Import** is set to **true**.	year=2020,location=sun

4.3.6 From FTP/SFTP

If the source link of a job is the **Link to an FTP or SFTP Server**, configure the source job parameters based on **Table 4-12**.

Advanced attributes are optional and not displayed by default. You can click **Show Advanced Attributes** to display them.

Table 4-12 Parameter description

Category	Parameter	Description	Example Value
Basic parameters	Source Directory/File	Directory or file path from which data will be extracted. You can enter a maximum of 50 file paths. By default, the file paths are separated by vertical bars (|). You can also customize a file separator. For details, see **Migration of a List of Files**. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	/ftp/a.csv|/ftp/b.txt

Category	Parameter	Description	Example Value
	File Format	Format in which CDM parses data. The options are as follows: • **CSV**: Source files will be migrated to tables after being converted to CSV format. • **Binary**: Files (even not in binary format) will be transferred directly. It is used for file copy. • **JSON**: Source files will be migrated to tables after being converted to JSON format.	CSV
	JSON Type	This parameter is displayed only when **File Format** is set to **JSON**. Type of a JSON object stored in a JSON file. The options are **JSON object** and **JSON array**.	JSON object
	JSON Reference Node	This parameter is used only when **File Format** is set to **JSON** and **JSON Type** is set to **JSON Object**. CDM parses the data under the JSON node. If the node's corresponding data is a JSON array, the system will extract data from the array in the same pattern. Use periods (.) to separate multi-layer nested JSON nodes.	data.list
Advanced attributes	Line Separator	Lind feed character in a file. By default, the system automatically identifies **\n**, **\r**, and **\r\n**. This parameter is displayed only when **File Format** is set to **CSV**.	\n
	Field Delimiter	Character used to separate fields in the file. To set the **Tab** key as the delimiter, set this parameter to **\t**. This parameter is displayed only when **File Format** is set to **CSV**.	,
	Use Quote Character	If you set this parameter to **Yes**, the field delimiters in the encircling symbol are regarded as a part of the string value. Currently, the default encircling symbol of CDM is **"**.	No
	Use RE to Separate Fields	Whether to use regular expressions to separate fields. If you set this parameter to **Yes**, **Field Delimiter** becomes invalid. This parameter is displayed only when **File Format** is set to **CSV**.	Yes

Category	Parameter	Description	Example Value
	Regular Expression	Regular expression used to separate fields. For details about regular expressions, see **Regular Expressions for Separating Semi-structured Text**.	^(\d.*\d) (\w*) \[(.*) \] ([\w\.]*) (\w.*).*
	Use First Row as Header	This parameter is displayed only when **File Format** is set to **CSV**. When you migrate a CSV file to a table, CDM writes all data to the table by default. If you set this parameter to **Yes**, CDM uses the first line of the CSV file as the heading line and does not write the line to the destination table.	Yes
	Encoding Type	Encoding type, for example, **UTF-8** or **GBK**. You can set the encoding type for text files only. This parameter is invalid when **File Format** is set to **Binary**.	UTF-8
	Compression Format	This parameter is displayed only when **File Format** is set to **CSV** or **JSON**. The options are as follows: • **NONE**: Files in all formats can be transferred. • **GZIP**: Only files in gzip format can be transferred. • **ZIP**: Only files in Zip format can be transferred. • **TAR.GZ**: Files in TAR.GZ format are transferred.	NONE
	Compressed File Suffix	This parameter is displayed when **Compression Format** is not **NONE**. This parameter specifies the extension of the files to be decompressed. The decompression operation is performed only when the file name extension is used in a batch of files. Otherwise, files are transferred in the original format. If you enter ***** or leave the parameter blank, all files are decompressed.	*

Category	Parameter	Description	Example Value
	Source File Processing Method	Operation performed on source files after the job completes. • No action • **Rename**: After the job completes, the source files are renamed by appending usernames and timestamps as suffixes to the file names. • **Delete**: After the job completes, the source files are deleted.	No action
	Start Job by Marker File	Whether to start a job by a marker file. A job is only started if there is a marker file for starting the job in the source path. If there is no marker file, the job will be suspended for a period of time specified by **Suspension Period**.	Yes
	Marker File	Name of the marker file for starting a job. If you specify a marker file, the migration job is executed only when the marker file exists in the source path. The marker file will not be migrated.	ok.txt
	Suspension Period	Waiting period for a marker file. If you set **Start Job by Marker File** to **Yes** but there is no marker file in the source path, the job fails when the suspension period times out. If you set this parameter to **0** and there is no marker file in the source path, the job will fail immediately. Unit: second	10
	File Separator	File separator. If you enter multiple file paths in **Source Directory/Files**, CDM uses the file separator to identify files. The default value is |.	|
	Filter Type	Only paths or files that meet the filtering conditions are transferred. The options are **None**, **Wildcard**, and **Regex**. For details, see **Incremental File Migration**.	None

Category	Parameter	Description	Example Value
	Directory Filter	If you set **Filter Type** to **Wildcard**, enter a wildcard character to filter paths. The paths that meet the filtering condition are migrated. You can configure multiple paths separated by commas (,). **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	*input,*out
	File Filter	If you set **Filter Type** to **Wildcard**, you can enter a wildcard character to search for files in a specified path. The files that meet the search criteria are migrated. You can configure multiple files separated by commas (,). **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	*.csv
	Time Filter	If you select **Yes**, files are transferred based on their modification time.	Yes
	Minimum Timestamp	If you set **Filter Type** to **Time Filter**, and specify a point in time for this parameter, only the files modified after the specified time are transferred. The time format must be *yyyy-MM-dd HH:mm:ss*. This parameter can be set to a macro variable of date and time. For example, **${timestamp(dateformat(yyyy-MM-dd HH:mm:ss,-90,DAY))}** indicates that only files generated within the latest 90 days are migrated. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	2019-07-01 00:00:00

Category	Parameter	Description	Example Value
	Maximum Timestamp	If you set **Filter Type** to **Time Filter**, and specify a point in time for this parameter, only the files modified before the specified time are transferred. The time format must be *yyyy-MM-dd HH:mm:ss*. This parameter can be set to a macro variable of date and time. For example, **${timestamp(dateformat(yyyy-MM-dd HH:mm:ss))}** indicates that only the files whose modification time is earlier than the current time are migrated. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	2019-07-30 00:00:00
	Encryption	If the source data is encrypted, CDM can decrypt the data before exporting it. Select whether to decrypt the source data and select a decryption algorithm. The options are as follows: • **NONE**: Export data without decrypting it. • **AES-256-GCM**: The AES 256-bit encryption algorithm is used to encrypt data. Currently, only the AES-256-GCM (NoPadding) encryption algorithm is supported. This parameter is used for encryption at the migration destination and decryption at the migration source. For details, see **Encryption and Decryption During File Migration**.	AES-256-GCM
	Disregard Non-existent Path or File	If this is set to **Yes**, the job can be successfully executed even if the source path does not exist.	No

Category	Parameter	Description	Example Value
	DEK	This parameter is displayed only when **Encryption** is set to **AES-256-GCM**. The key consists of 64 hexadecimal numbers and must be the same as the **DEK** configured during encryption. If the decryption and encryption keys are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B
	IV	This parameter is displayed only when **Encryption** is set to **AES-256-GCM**. The initialization vector consists of 32 hexadecimal numbers and must be the same as the **IV** configured during encryption. If the initialization vectors are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	5C91687BA886EDCD12ACBC3FF19A3C3F
	MD5 File Extension	This parameter is displayed only when **File Format** is set to **Binary**. This parameter is used to check whether the files extracted by CDM are consistent with source files. For details, see **MD5 Verification**.	.md5

4.3.7 From HTTP

When the source link of a job is the HTTP link, configure the source job parameters based on **Table 4-13**. Currently, data can only be exported from the HTTP URLs.

Table 4-13 Parameter description

Parameter	Description	Example Value
File URL	Use the GET method to obtain data from the HTTP/HTTPS URL. These connectors are used to read files with an HTTP/HTTPS URL, such as reading public files on the third-party object storage system and web disks.	https://bucket.obs.myhuaweicloud.com/object-key

Parameter	Description	Example Value
Pull List File	If this parameter is set to **Yes**, the system pulls the files corresponding to the URLs in the text file to be uploaded and stores them on OBS. The text file records the file paths on HDFS.	Yes
OBS Link of List File	Select an existing OBS link.	obs_link
OBS Bucket of entries files	Name of the OBS bucket that stores the text file	obs-cdm
Path/ Directory of entries files	Custom OBS directories that store the text file. Use slashes (/) to separate different directories.	test1
File Format	CDM supports **Binary** only, which indicates that files (even not in binary format) will be directly transferred.	Binary
Compression Format	Compression format of the source files. The options are as follows: ● **NONE**: Files in all formats can be transferred. ● **GZIP**: Only files in gzip format can be transferred. ● **ZIP**: Only files in Zip format can be transferred. ● **TAR.GZ**: Files in TAR.GZ format are transferred.	NONE
Compressed File Suffix	This parameter is displayed when **Compression Format** is not **NONE**. This parameter specifies the extension of the files to be decompressed. The decompression operation is performed only when the file name extension is used in a batch of files. Otherwise, files are transferred in the original format. If you enter ***** or leave the parameter blank, all files are decompressed.	*
File Separator	File separator. When multiple files are transferred, CDM uses the file separator to identify files. The default value is **|**. This parameter is not displayed if **Pull List File** is set to **Yes**.	|

Parameter	Description	Example Value
Query Parameter	• If you set this parameter to **Yes**, the name of the objects uploaded to OBS does not include the **query** parameter. • If you set this parameter to **No**, the name of the objects uploaded to OBS includes the **query** parameter.	No
Encryption	If the source data is encrypted, CDM can decrypt the data before exporting it. Select whether to decrypt the source data and select a decryption algorithm. The options are as follows: • **NONE**: Export data without decrypting it. • **AES-256-GCM**: The AES 256-bit encryption algorithm is used to encrypt data. Currently, only the AES-256-GCM (NoPadding) encryption algorithm is supported. This parameter is used for encryption at the migration destination and decryption at the migration source. For details, see **Encryption and Decryption During File Migration**.	AES-256-GCM
Disregard Non-existent Path or File	If this is set to **Yes**, the job can be successfully executed even if the source path does not exist.	No
DEK	This parameter is displayed only when **Encryption** is set to **AES-256-GCM**. The key consists of 64 hexadecimal numbers and must be the same as the **DEK** configured during encryption. If the decryption and encryption keys are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B
IV	This parameter is displayed only when **Encryption** is set to **AES-256-GCM**. The initialization vector consists of 32 hexadecimal numbers and must be the same as the **IV** configured during encryption. If the initialization vectors are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	5C91687BA886EDCD12ACBC3FF19A3C3F
MD5 File Extension	This parameter is used to check whether the files extracted by CDM are consistent with source files. For details, see **MD5 Verification**.	.md5

4.3.8 From a Common Relational Database

Common relational databases that can serve as the source include GaussDB(DWS), RDS for MySQL, RDS for PostgreSQL, RDS for SQL Server, Dameng, FusionInsight LibrA, PostgreSQL, Microsoft SQL Server, SAP HANA, and MyCAT.

To export data from the preceding databases, configure the source job parameters listed in **Table 4-14**.

Table 4-14 Parameter description

Category	Parameter	Description	Example Value
Basic parameters	Use SQL Statement	Whether you can use SQL statements to export data from a relational database	No

	SQL Statement	When **Use SQL Statement** is set to **Yes**, enter an SQL statement here. CDM exports data based on the SQL statement.
		NOTE
		• SQL statements can only be used to query data. Join and nesting are supported, but multiple query statements are not allowed, for example, **select * from table a; select * from table b**.
		• With statements are not supported.
		• Comments, such as **--** and **/***, are not supported.
		• Addition, deletion, and modification operations are not supported, including but not limited to the following:
		  • load data
		  • delete from
		  • alter table
		  • create table
		  • drop table
		  • into outfile
			select id,name from sqoop.user;

Category	Parameter	Description	Example Value
	Schema/Tablespace	Name of the schema or tablespace from which data will be extracted. This parameter is displayed when **Use SQL Statement** is set to **No**. Click the icon next to the text box to go to the page for selecting a schema or directly enter a schema or tablespace. If the desired schema or tablespace is not displayed, check whether the login account has the permissions required to query metadata. **NOTE** The parameter value can contain wildcard characters (*), which is used to export all databases whose names start with a certain prefix or end with a certain suffix. The examples are as follows: • **SCHEMA*** indicates that all databases whose names starting with **SCHEMA** are exported. • ***SCHEMA** indicates that all databases whose names ending with **SCHEMA** are exported. • ***SCHEMA*** indicates that all databases whose names containing **SCHEMA** are exported.	SCHEMA_E

Category	Parameter	Description	Example Value
	Table Name	Name of the table from which data will be extracted. This parameter is displayed when **Use SQL Statement** is set to **No**. Click the icon next to the text box to go to the page for selecting the table or directly enter a table name. If the desired table is not displayed, confirm that the table exists or that the login account has the permissions required to query metadata. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*). **NOTE** The table name can contain wildcard characters (*), which is used to export all tables whose names start with a certain prefix or end with a certain suffix. The number and types of fields in the tables must be the same. The examples are as follows: • **table*** indicates that all tables whose names starting with **table** are exported. • ***table** indicates that all tables whose names ending with **table** are exported. • ***table*** indicates that all tables whose names containing **table** are exported.	table

Category	Parameter	Description	Example Value
Advanced attributes	Partition Column	This parameter is displayed when **Use SQL Statement** is set to **No**, indicating that a field used to split data during data extraction. CDM splits a job into multiple tasks based on this field and executes the tasks concurrently. Fields with data distributed evenly are used, such as the sequential number field. Click the icon next to the text box to go to the page for selecting a field or directly enter a field. **NOTE** • The following types of partition columns are supported: CHAR, VARCHAR, LONGVARCHAR, TINYINT, SMALLINT, INTEGER, BIGINT, REAL, FLOAT, DOUBLE, NUMERIC, DECIMAL, BIT, BOOLEAN, DATE, TIME, and TIMESTAMP. It is recommended that the partition column have an index. • If the partition column type is CHAR, VARCHAR, or LONGVARCHAR, the column value cannot contain characters other than those in the ASCII character code table or Chinese characters.	id
	Where Clause	WHERE clause used to specify the data extraction range. This parameter is displayed when **Use SQL Statement** is set to **No**. If this parameter is not set, the entire table is extracted. You can set a date macro variable to extract data generated on a specific date. For details, see **Incremental Migration of Relational Databases**. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
	Null in Partition Column	Whether the partition column can contain null values	Yes
	Job Split Field	Used to split a job into multiple subjobs for concurrent execution.	-

Category	Parameter	Description	Example Value
	Minimum value of a split field	Specifies the minimum value of **Job Split Field** during data extraction.	-
	Maximum Split Field Value	Specifies the maximum value of **Job Split Field** during data extraction.	-
	Number of subjobs	Specifies the number of subjobs split from a job based on the data range specified by the minimum and maximum values of **Job Split Field**.	-
	Extract by Partition	When data is exported from an MySQL database, data can be extracted from each partition in the partitioned table. If this function is enabled, you can configure **Table Partition** to specify specific MySQL table partitions from which data is extracted. • This function does not support non-partitioned tables. • This parameter is available only for RDS for PostgreSQL and RDS for MySQL. • The database user must have the **SELECT** permission on the system views **dba_tab_partitions** and **dba_tab_subpartitions**.	No

4.3.9 From MySQL

If the source link of a job is the **Link to an RDS for MySQL/MySQL Database**, configure the source job parameters based on **Table 4-15**.

Table 4-15 Parameter description

Category	Parameter	Description	Example Value
Basic parameters	Use SQL Statement	Whether you can use SQL statements to export data from a relational database	No

Category	Parameter	Description	Example Value

	SQL Statement	When **Use SQL Statement** is set to **Yes**, enter an SQL statement here. CDM exports data based on the SQL statement.
		NOTE
		• SQL statements can only be used to query data. Join and nesting are supported, but multiple query statements are not allowed, for example, **select * from table a; select * from table b**.
		• With statements are not supported.
		• Comments, such as **--** and **/***, are not supported.
		• Addition, deletion, and modification operations are not supported, including but not limited to the following:
		  • load data
		  • delete from
		  • alter table
		  • create table
		  • drop table
		  • into outfile
			select id,name from sqoop.user;

	Schema/ Tablespace	Name of the schema or tablespace from which data will be extracted. This parameter is displayed when **Use SQL Statement** is set to **No**. Click the icon next to the text box to go to the page for selecting a schema or directly enter a schema or tablespace. If the desired schema or tablespace is not displayed, check whether the login account has the permissions required to query metadata. This parameter can be set to a regular expression to export all databases that meet the rule.	SCHEMA_E

Category	Parameter	Description	Example Value
	Table Name	Name of the table from which data will be extracted. This parameter is displayed when **Use SQL Statement** is set to **No**. Click the icon next to the text box to go to the page for selecting the table or directly enter a table name. If the desired table is not displayed, confirm that the table exists or that the login account has the permissions required to query metadata. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**. This parameter can be set to a regular expression to export all databases that meet the rule. For example, if **Table Name** is set to *user_[0-9]{1,2}*, tables from **user_0** to **user_9** and from **user_00** to **user_99** are matched. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	table

Category	Parameter	Description	Example Value
Advanced attributes	Partition Column	This parameter is displayed when **Use SQL Statement** is set to **No**, indicating that a field used to split data during data extraction. CDM splits a job into multiple tasks based on this field and executes the tasks concurrently. Fields with data distributed evenly are used, such as the sequential number field. Click the icon next to the text box to go to the page for selecting a field or directly enter a field. **NOTE** • The following types of partition columns are supported: CHAR, VARCHAR, LONGVARCHAR, TINYINT, SMALLINT, INTEGER, BIGINT, REAL, FLOAT, DOUBLE, NUMERIC, DECIMAL, BIT, BOOLEAN, DATE, TIME, and TIMESTAMP. It is recommended that the partition column have an index. • If the partition column type is CHAR, VARCHAR, or LONGVARCHAR, the column value cannot contain characters other than those in the ASCII character code table or Chinese characters.	id
	WHERE Clause	WHERE clause used to specify the data extraction range. This parameter is displayed when **Use SQL Statement** is set to **No**. If this parameter is not set, the entire table is extracted. You can set a date macro variable to extract data generated on a specific date. For details, see **Incremental Migration of Relational Databases**. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
	Null in Partition Column	Whether the partition column can contain null values	Yes
	Job Split Field	Used to split a job into multiple subjobs for concurrent execution.	-

Category	Parameter	Description	Example Value
	Minimum value of a split field	Specifies the minimum value of **Job Split Field** during data extraction.	-
	Maximum Split Field Value	Specifies the maximum value of **Job Split Field** during data extraction.	-
	Number of subjobs	Specifies the number of subjobs split from a job based on the data range specified by the minimum and maximum values of **Job Split Field**.	-
	Extract by Partition	When data is exported from a MySQL database, data can be extracted from each partition in the partitioned table. If this function is enabled, you can configure **Table Partition** to specify specific MySQL table partitions from which data is extracted. ● This function does not support non-partitioned tables. ● The database user must have the **SELECT** permission on the system views **dba_tab_partitions** and **dba_tab_subpartitions**.	No

NOTE

- In a migration from MySQL to DWS, the constraints on the incremental data migration function in MySQL Binlog mode are as follows:
  1. A single cluster supports only one incremental migration job in MySQL Binlog mode in the current version.
  2. In the current version, you are not allowed to delete or update 10,000 data records at a time.
  3. Entire DB migration is not supported.
  4. Data Definition Language (DDL) operations are not supported.
  5. Event migration is not supported.
  6. If you set **Migrate Incremental Data** to **Yes**, **binlog_format** in the source MySQL database must be set to **ROW**.
  7. If you set **Migrate Incremental Data** to **Yes** and binlog file ID disorder occurs on the source MySQL instance due to cross-machine migration or rebuilding during incremental data migration, incremental data may be lost.
  8. If a primary key exists in the destination table and incremental data is generated during the restart of the CDM cluster or full migration, duplicate data may exist in the primary key. As a result, the migration fails.
  9. If the destination DWS database is restarted, the migration will fail. In this case, restart the CDM cluster and the migration job.
- The recommended MySQL configuration is as follows:

```
# Enable the bin-log function.
log-bin=mysql-bin
# Row mode
binlog-format=ROW
# gtid mode. The recommended version is 5.6.10 or later.
gtid-mode=ON
enforce_gtid_consistency = ON
```

4.3.10 From Oracle

If the source link of a job is the **Link to an Oracle Database**, configure the source job parameters based on **Table 4-16**.

Table 4-16 Parameter description

Category	Parameter	Description	Example Value
Basic parameters	Use SQL Statement	Whether you can use SQL statements to export data from a relational database	No

Category	Parameter	Description	Example Value

	SQL Statement	When **Use SQL Statement** is set to **Yes**, enter an SQL statement here. CDM exports data based on the SQL statement.
		NOTE
		• SQL statements can only be used to query data. Join and nesting are supported, but multiple query statements are not allowed, for example, **select * from table a; select * from table b**.
		• With statements are not supported.
		• Comments, such as **--** and **/***, are not supported.
		• Addition, deletion, and modification operations are not supported, including but not limited to the following:
		  • load data
		  • delete from
		  • alter table
		  • create table
		  • drop table
		  • into outfile
			select id,name from sqoop.user;

	Schema/ Tablespace	Name of the schema or tablespace from which data will be extracted. This parameter is displayed when **Use SQL Statement** is set to **No**. Click the icon next to the text box to go to the page for selecting a schema or directly enter a schema or tablespace. If the desired schema or tablespace is not displayed, check whether the login account has the permissions required to query metadata. **NOTE** The parameter value can contain wildcard characters (*), which is used to export all databases whose names start with a certain prefix or end with a certain suffix. For example: • **SCHEMA*** indicates that all databases whose names starting with **SCHEMA** are exported. • ***SCHEMA** indicates that all databases whose names ending with **SCHEMA** are exported. • ***SCHEMA*** indicates that all databases whose names containing **SCHEMA** are exported.	SCHEMA_E

Category	Parameter	Description	Example Value
	Table Name	Name of the table from which data will be extracted. This parameter is displayed when **Use SQL Statement** is set to **No**. Click the icon next to the text box to go to the page for selecting the table or directly enter a table name. If the desired table is not displayed, confirm that the table exists or that the login account has the permissions required to query metadata. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*). **NOTE** The table name can contain wildcard characters (*), which is used to export all tables whose names start with a certain prefix or end with a certain suffix. The number and types of fields in the tables must be the same. The examples are as follows: • **table*** indicates that all tables whose names starting with **table** are exported. • ***table** indicates that all tables whose names ending with **table** are exported. • ***table*** indicates that all tables whose names containing **table** are exported.	table

Category	Parameter	Description	Example Value
Advanced attributes	Partition Column	This parameter is displayed when **Use SQL Statement** is set to **No**, indicating that a field used to split data during data extraction. CDM splits a job into multiple tasks based on this field and executes the tasks concurrently. Fields with data distributed evenly are used, such as the sequential number field. Click the icon next to the text box to go to the page for selecting a field or directly enter a field. **NOTE** • The following types of partition columns are supported: CHAR, VARCHAR, LONGVARCHAR, TINYINT, SMALLINT, INTEGER, BIGINT, REAL, FLOAT, DOUBLE, NUMERIC, DECIMAL, BIT, BOOLEAN, DATE, TIME, and TIMESTAMP. It is recommended that the partition column have an index. • If the partition column type is CHAR, VARCHAR, or LONGVARCHAR, the column value cannot contain characters other than those in the ASCII character code table or Chinese characters.	id
	WHERE Clause	WHERE clause used to specify the data extraction range. This parameter is displayed when **Use SQL Statement** is set to **No**. If this parameter is not set, the entire table is extracted. You can set a date macro variable to extract data generated on a specific date. For details, see **Incremental Migration of Relational Databases**. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
	Null in Partition Column	Whether the partition column can contain null values	Yes

Category	Parameter	Description	Example Value
	Extract by Partition	When data is exported from an Oracle database, data can be extracted from each partition in the partitioned table. If this function is enabled, you can configure **Table Partition** to specify specific Oracle table partitions from which data is extracted. • This function does not support non-partitioned tables. • The database user must have the **SELECT** permission on the system views **dba_tab_partitions** and **dba_tab_subpartitions**.	No
	Table Partition	Oracle table partition from which data is migrated. Separate multiple partitions with ampersands (&). If you do not set this parameter, all partitions will be migrated. If there is a subpartition, enter the partition in the *Partition.Subpartition* format, for example, **P2.SUBP1**.	P0&P1&P2.SUBP1&P2.SUBP3
	Job Split Field	Used to split a job into multiple subjobs for concurrent execution.	-
	Minimum value of a split field	Specifies the minimum value of **Job Split Field** during data extraction.	-
	Maximum Split Field Value	Specifies the maximum value of **Job Split Field** during data extraction.	-
	Number of subjobs	Specifies the number of subjobs split from a job based on the data range specified by the minimum and maximum values of **Job Split Field**.	-

NOTE

When an Oracle database is the migration source, if **Partitioning Field** or **Extract by Partition** is not configured, CDM automatically uses the ROWIDs to partition data.

4.3.11 From a Database Shard

If the source link of a job is the **Link to a Database Shard**, configure the source job parameters based on **Table 4-17**.

Table 4-17 Parameter description

Category	Parameter	Description	Example Value
Basic parameters	Schema/ Tablespace	Indicates the name of the schema or tablespace from which data is to be extracted. Click the icon next to the text box to go to the page for selecting a schema or tablespace. During a sharded link job, the tablespace corresponding to the first backend link is displayed by default. You can also enter a schema or tablespace name. If the desired schema or tablespace is not displayed, check whether the login account has the permissions required to query metadata. This parameter can be set to a regular expression to export all databases that meet the rule. For example, if **Table Name** is set to *user_[0-9]{1,2}*, tables from **user_0** to **user_9** and from **user_00** to **user_99** are matched.	SCHEMA_E

Category	Parameter	Description	Example Value
	Table Name	Indicates the name of the table from which data is to be extracted. Click the icon next to the text box to go to the page for selecting the table or directly enter a table name. If the desired table is not displayed, confirm that the table exists or that the login account has the permissions required to query metadata. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**. This parameter can be set to a regular expression to export all databases that meet the rule. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	table
Advanced attributes	WHERE Clause	Specifies the data extraction range. If this parameter is not set, the entire table is extracted. You can set a date macro variable to extract data generated on a specific date. For details, see **Incremental Migration of Relational Databases**. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'

NOTE

- If the **Source Link Name** is the backend link of the sharded link, the job is a common MySQL job.
- When creating a job whose source end is a sharded link, you can add a custom field with the sample value of **${custom(host)}** to the source field during field mapping. This field is used to view the data source of the table after the data of multiple tables across databases is migrated to the same table. The following sample values are supported:
  - ${custom(host)}
  - ${custom(database)}
  - ${custom(fromLinkName)}
  - ${custom(schemaName)}
  - ${custom(tableName)}

4.3.12 From MongoDB/DDS

When you migrate MongoDB or DDS data, CDM reads the first row of the collection as an example of the field list. If the first row of data does not contain all fields of the collection, you need to manually add fields.

When the source link of a job is the **Link to MongoDB**, that is, when data is exported from an on-premises MongoDB or DDS, configure the source job parameters based on **Table 4-18**.

Table 4-18 Parameter description

Parameter	Description	Example Value
Database Name	Name of the database from which data will be migrated	mongodb
Collection Name	Collection name, similar to the table name of a relational database. Click the icon next to the text box to go to the page for selecting the collection or directly enter a collection name. If the desired table is not displayed, check whether the table exists or whether the login account has the permission to query metadata.	COLLECTION

Parameter	Description	Example Value
Filter Condition	Conditions for filtering documents. CDM migrates only the data that meets the filter conditions. The examples are as follows: 1. Filter by expression: **{'last_name': 'Smith'}** indicates that all files whose **last_name** value is **Smith** are queried. 2. Filter by parameter: **{ x : "john" }, { z : 1 }** indicates that all **z** fields whose **x** is **john** are queried. 3. Filter by condition: **{ "field" : { $gt: 5 } }** indicates that the **field** values greater than 5 are queried. 4. Filter by time macro: {"ts":{$gte:ISODate("${dateformat(yyyy-MM-dd'T'HH:mm:ss.SSS'Z',-1,HOUR)}")}} indicates that the values greater than those after time macro conversion in the **ts** field are queried.	{'last_name': 'Smith'}

4.3.13 From Redis

Because DCS restricts the commands for obtaining keys, it cannot serve as the migration source but can be the migration destination. The Redis service of the third-party cloud cannot serve as the migration source. However, the Redis set up in the on-premises data center or on the ECS can be the migration source and destination.

When data is exported from an on-premises Redis, configure source job parameters as described in **Table 4-19**.

Table 4-19 Parameter description

Parameter	Description	Example Value
Redis Key Prefix	Key prefix, which is similar to the table name of a relational database	TABLE
Value Storage Type	The options are as follows: • **String**: without column name, such as **value1,value2** • **Hash**: with column name, such as **column1=value1,column2=value2**	String
Key Delimiter	Character used to separate table names and column names of a relational database	_

Parameter	Description	Example Value
Value Delimiter	Character used to separate columns when the storage type is string	;
Same Field	This parameter is displayed when **Value Storage Type** is set to **Hash**. The hash key contains the same field.	Yes

4.3.14 From Kafka/DMS Kafka

If the source link of a job is the **Link to Kafka** or **Link to DMS Kafka**, configure the source job parameters based on **Table 4-20**.

Table 4-20 Parameter description

Parameter	Description	Example Value
Topics	One or more topics can be entered.	est1,est2
Offset	Initial offset parameter • **Latest**: Maximum offset, indicating that the latest data will be extracted. • **Earliest**: Minimum offset, indicating that the earliest data will be extracted. • **Submitted**: data that has been submitted • **Time Range**: data within a specified time range	Latest
Permanent Running	Whether a job runs permanently.	Yes
Consumer Group ID	Consumer group ID If you export data from DMS Kafka, enter any value for Kafka Platinum but a valid consumer group ID for Kafka Basic.	sumer-group

Parameter	Description	Example Value
Data Format	Format used for parsing data. The options are as follows: • **Binary**: Data is transferred directly. It is not converted to another format. This setting is suitable for file migration. • **CSV**: Source data will be migrated after being converted in CSV format. • **JSON**: Source data will be migrated after being converted in JSON format. • **CDC (DRS_JSON)**: Source data will be migrated after being converted in DRS_JSON format.	Binary
Field Delimiter	The default value is space. To set the **Tab** key as the delimiter, set this parameter to **\t**.	,
Max. Poll Records	(Optional) Maximum number of records per poll	100
Max. Poll Interval	(Optional) Maximum interval between polls (seconds)	100

4.3.15 From Elasticsearch or CSS

If the source link of a job is the **Link to Elasticsearch/CSS**, configure the source job parameters based on **Table 4-21**.

Table 4-21 Job parameters when Elasticsearch or CSS is the source

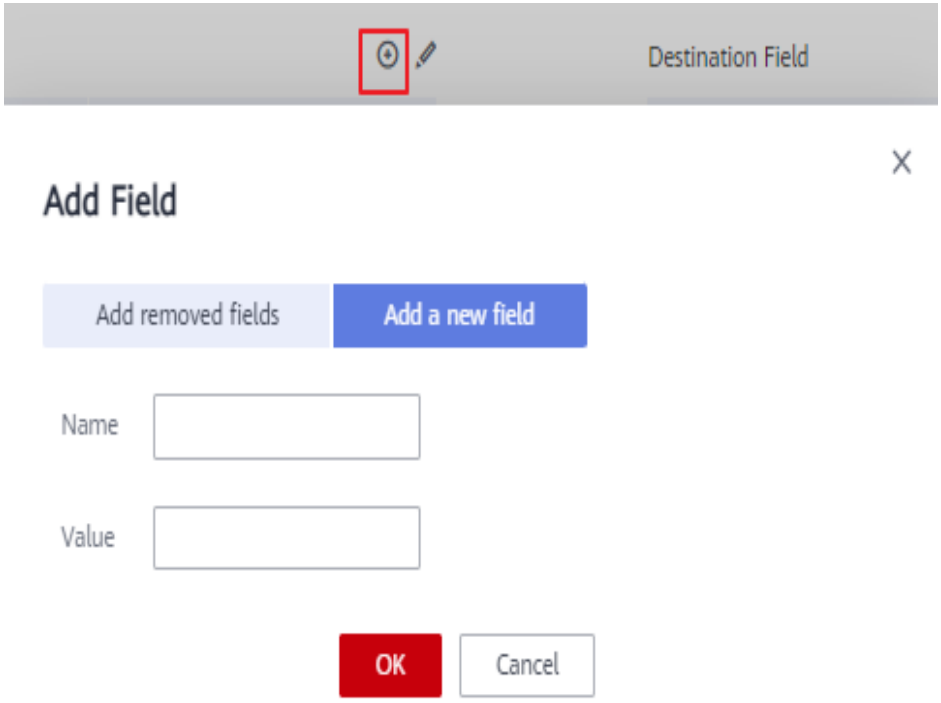
Parameter	Description	Example Value
Index	Elasticsearch index, which is similar to the name of a relational database. The index name can contain only lowercase letters.	index
Type	Elasticsearch type, which is similar to the table name of a relational database. The type name can contain only lowercase letters. **NOTE** Elasticsearch 7.x and later versions do not support custom types. Instead, only the **_doc** type can be used. In this case, this parameter does not take effect even if it is set.	_doc
Split Nested Field	(Optional) Whether to split the JSON content of the nested fields. For example, **a:{ b:{ c:1, d:{ e:2, f:3 } } }** can be split into **a.b.c**, **a.b.d.e**, and **a.b.d.f**.	No

Parameter	Description	Example Value
Filter Conditions	(Optional) CDM migrates only the data that meets the filter conditions. • Currently, only the query string (q syntax) of Elasticsearch can be used to filter source data. The q syntax is used in the following way: – In exact match, the ***column*:*data*** format is used to match and filter data. ***column*** indicates the field name, and ***data*** indicates the query condition, for example, **last_name:Smith**. In addition, if ***data*** is a string containing spaces, it must be enclosed in double quotation marks. If ***column*** is not specified, all fields will be matched by ***data***. – Multiple query conditions can be combined with connection words. The format is ***column1*:*data1* AND *column2*:*data2***. The connection words can be **AND**, **OR**, or **NOT**. They must be in uppercase, and there must be a space before and after each connection word. Example: **last_name:Smith AND last_name:John** – In range matching, you can directly use a condition expression to filter data. The expression is in ***column*:>*data*** format. The operator can be >, >=, <, or <=. An example is **time:>=1636905600000 AND time:<1637078400000**. It can also be used together with a macro variable of date and time, for example, **createTime:>=${timestamp(dateformat(yyyyMMdd,-1,DAY))} AND createTime:< ${timestamp(dateformat(yyyyMMdd))}**. – In range matching, you can also use the range syntax to filter data. The format is ***column*:{*data1* TO *data2*}**. **{** and **}** indicate that a value is not included. **[** and **]** indicate that a value is included. **TO** must be capitalized, and there must be a space before and after it. ***** indicates all data. For example, **time:{1636992000000 TO *]** filters out all the data greater than 1636992000000 in the **time** field. It can also be used together with a macro variable of date and time, for example, **createTime:[$**	last_name:Smith

Parameter	Description	Example Value
	{timestamp(dateformat(yyyyMMdd,-1,DAY))} TO ${timestamp(dateformat(yyyyMMdd))}}. • Source data cannot be filtered using the query domain-specific language (DSL) of Elasticsearch.	
Extract Meta-field	Whether to extract index meta-fields. For example, _index, _type, _id, and _score.	Yes

On the **Map Field** page, you can set custom fields for the source and destination.

Figure 4-9 Setting custom fields



4.4 Destination Job Parameters

4.4.1 To OBS

If the destination link of a job is the **Link to OBS**, configure the destination job parameters based on **Table 4-22**.

Advanced attributes are optional and not displayed by default. You can click **Show Advanced Attributes** to display them.

Table 4-22 Parameter description

Category	Parameter	Description	Example Value
Basic parameters	Bucket Name	Name of the OBS bucket that data will be written to	bucket_2
	Write Directory	OBS directory to which data will be written. Do not add **/** in front of the directory name. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	directory/
	File Format	Format in which data is written. The options are as follows: • **CSV**: Data is written in CSV format, which is used for migrating data tables to files. • **Binary**: Files will be transferred directly. CDM writes the files without changing their format. This setting is suitable for file migration. If data is migrated between file-related data sources, such as FTP, SFTP, HDFS, and OBS, the value of **File Format** must the same as the source file format.	CSV
	Duplicate File Processing Method	Files with the same name and size are identified as duplicate files. If there are duplicate files during data writing, the following methods are available: • Replace • Skip • Stop job For details, see **Incremental File Migration**.	Skip

Category	Parameter	Description	Example Value
Advanced attributes	Encryption	Whether to encrypt the uploaded data and the encryption mode. The options are as follows: • **None**: Data is written without encryption. • **KMS**: KMS in Data Encryption Workshop (DEW) is used for encryption. If KMS encryption is enabled, MD5 verification for data cannot be performed. • **AES-256-GCM**: The AES 256-bit encryption algorithm is used to encrypt data. Currently, only the AES-256-GCM (NoPadding) encryption algorithm is supported. This parameter is used for encryption at the migration destination and decryption at the migration source. For details, see **Encryption and Decryption During File Migration**.	KMS
	Key ID	Data encryption key. This parameter is displayed when **Encryption** is set to **KMS**. Click next to the text box to select the KMS key that was created in DEW. • If the KMS key of the same project as that of the CDM cluster is used, you do not need to modify **Project ID**. • If the KMS key of another project is used, you need to modify **Project ID**.	53440ccb-3e73-4700-98b5-71ff5476e621
	Project ID	ID of the project to which KMS ID belongs. The default value is the ID of the project to which the current CDM cluster belongs. • If KMS and the CDM cluster are in the same project, retain the default value of **Project ID**. • If KMS of another project is used, set this parameter to the ID of the project to which KMS belongs.	9bd7c4bd54e5417198f9591bef07ae67

Category	Parameter	Description	Example Value
	DEK	This parameter is displayed only when **Encryption** is set to **AES-256-GCM**. The key consists of 64 hexadecimal numbers. Remember the key configured here because the decryption key must be the same as that configured here. If the encryption and decryption keys are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B
	IV	This parameter is displayed only when **Encryption** is set to **AES-256-GCM**. The initialization vector consists of 32 hexadecimal numbers. Remember the initialization vector configured here because the initialization vector used for decryption must be the same as that configured here. If the initialization vectors are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	5C91687BA886EDCD12ACBC3FF19A3C3F
	Copy Content-Type	This parameter is displayed only when **File Format** is **Binary**, and both the migration source and destination are object storage. If you set this parameter to **Yes**, the Content-Type attribute of the source file is copied during object file migration. This function is mainly used for static website migration. The Content-Type attribute cannot be written to Archive buckets. Therefore, if you set this parameter to **Yes**, the migration destination must be a non-Archive bucket.	No
	Line Separator	Lind feed character in a file. By default, the system automatically identifies **\n**, **\r**, and **\r\n**. This parameter is not used when **File Format** is set to **Binary**.	\n
	Field Delimiter	Field delimiter in the file. This parameter is not used when **File Format** is set to **Binary**.	,

Category	Parameter	Description	Example Value
	File Size	This parameter is displayed only when the migration source is a database. Files are partitioned as multiple files by size so that they can be exported in proper size. The unit is MB.	1024
	Validate MD5 Value	The MD5 value can be verified only when files are transferred in **Binary** format. KMS encryption cannot be used if the MD5 value needs to be verified. Calculate the MD5 value of the source files and verify it with the MD5 value returned by OBS. If an MD5 file exists on the migration source, the system directly reads the MD5 file from the migration source and verifies it with the MD5 value returned by OBS. For details, see **MD5 Verification**.	Yes
	Record MD5 Verification Result	Whether to record the MD5 verification result when **Validate MD5 Value** is set to **Yes**	Yes
	Record MD5 Link	OBS link to which the MD5 verification result will be written	obslink
	Record MD5 Bucket	OBS bucket to which the MD5 verification result will be written	cdm05
	Record MD5 Directory	Directory to which the MD5 verification result will be written	/md5/
	Encoding Type	Encoding type, for example, **UTF-8** or **GBK**. This parameter is not used when **File Format** is set to **Binary**.	GBK

Category	Parameter	Description	Example Value
	Use Quote Character	This parameter is displayed only when **File Format** is **CSV**. It is used when database tables are migrated to file systems. If you set this parameter to **Yes** and a field in the source data table contains a field delimiter or line separator, CDM uses double quotation marks (") as the quote character to quote the field content as a whole to prevent a field delimiter from dividing a field into two fields, or a line separator from dividing a field into different lines. For example, if the **hello,world** field in the database is quoted, it will be exported to the CSV file as a whole.	No
	Use First Row as Header	This parameter is displayed only when data is exported from a relational database to OBS and **File Format** is set to **CSV**. When a table is migrated to a CSV file, CDM does not migrate the heading line of the table by default. If you set this parameter to **Yes**, CDM writes the heading line of the table to the file.	No
	Job Success Marker File	Whether to generate a marker file with a custom name in the destination directory after a job is executed successfully. If you do not specify a file name, this function is disabled by default.	finish.txt
	Customize Hierarchical Directory	If this parameter is set to **Yes**, the files after migration can be stored in a custom directory. That is, only files are migrated. The directories to which the files belong are not migrated.	Yes
	Hierarchical Directory	Custom storage directory for files after migration. The time macro variable is supported.	${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}

Category	Parameter	Description	Example Value
	Customize File Name	This parameter is displayed only when data is exported from a relational database to OBS and **File Format** is set to **CSV**. This parameter specifies the name of the file generated by OBS. The options are as follows: • **Character string**: Special characters are allowed. For example, if this parameter is set to **cdm#**, the name of the generated file is **cdm#.csv**. • Macro variable of time: If this parameter is set to **${timestamp()}**, the name of the generated file is **1554108737.csv**. • Macro variable of table name: If this parameter is set to **${tableName}**, the name of the generated file is **sqltabname.csv**. • Macro variable of version number: If this parameter is set to **${version}**, the name of the generated file is **v1.csv**. • Any combination of the character string and macro variable (macro variable of time, table name, or version number). For example, if this parameter is set to **cdm#${timestamp()}_${version}**, the name of the generated file is **cdm#1554108737_v1.csv**.	cdm

4.4.2 To HDFS

If the destination link of a job is one of them listed in **Link to HDFS**, configure the destination job parameters based on **Table 4-23**.

Table 4-23 Parameter description

Parameter	Description	Example Value
Write Directory	HDFS directory to which data will be written. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	/user/output
File Format	Format in which data is written. The options are as follows: • **CSV**: Data is written in CSV format, which is used for migrating data tables to files. • **Binary**: Files will be transferred directly. CDM writes the files without changing their format. This setting is suitable for file migration. If data is migrated between file-related data sources, such as FTP, SFTP, HDFS, and OBS, the value of **File Format** must the same as the source file format.	CSV
Duplicate File Processing Method	Files with the same name and size are identified as duplicate files. If there are duplicate files during data writing, the following methods are available: • Replace • Skip • Stop job	Stop job

Parameter	Description	Example Value
Compression Format	File compression format after data writing. The following compression formats are supported: • **None**: The files are not compressed. • **DEFLATE**: The files are compressed in DEFLATE format. • **gzip**: The files are compressed in gzip format. • **bzip2**: The files are compressed in bzip2 format. • **LZ4**: The files are compressed in LZ4 format. • **Snappy**: The files are compressed in snappy format.	Snappy
Line Separator	Lind feed character in a file. By default, the system automatically identifies **\n**, **\r**, and **\r\n**. This parameter is not used when **File Format** is set to **Binary**.	\n
Field Delimiter	Field delimiter in the file. This parameter is not used when **File Format** is set to **Binary**.	,
Use Quote Character	This parameter is displayed only when **File Format** is **CSV**. It is used when database tables are migrated to file systems. If you set this parameter to **Yes** and a field in the source data table contains a field delimiter or line separator, CDM uses double quotation marks (") as the quote character to quote the field content as a whole to prevent a field delimiter from dividing a field into two fields, or a line separator from dividing a field into different lines. For example, if the **hello,world** field in the database is quoted, it will be exported to the CSV file as a whole.	No
Use First Row as Header	When a table is migrated to a CSV file, CDM does not migrate the heading line of the table by default. If you set this parameter to **Yes**, CDM writes the heading line of the table to the file.	No
Write to Temporary File	Whether to write the binary file to a **.tmp** file first. After the migration is successful, run the **rename** or **move** command at the migration destination to restore the file.	No
Job Success Marker File	Whether to generate a marker file with a custom name in the destination directory after a job is executed successfully. If you do not specify a file name, this function is disabled by default.	finish.txt

Parameter	Description	Example Value
Customize Hierarchical Directory	Users can customize the directory hierarchy of files. Example: [Table name]/[Year]/[Month]/[Day]/[Data file name]. csv	-
Hierarchical Directory	Used to specify the directory level of a file, with time macro supported (the time format is yyyy/MM/dd). If this parameter is left blank, the directory does not have a hierarchical structure. Example: ${dateformat(yyyy/MM/dd, -1, DAY)}	-
Encryption	This parameter is displayed only when **File Format** is set to **Binary**. Whether to encrypt the uploaded data. The options are as follows: • **None**: Data is written without encryption. • **AES-256-GCM**: The AES 256-bit encryption algorithm is used to encrypt data. Currently, only the AES-256-GCM (NoPadding) encryption algorithm is supported. This parameter is used for encryption at the migration destination and decryption at the migration source. For details, see **Encryption and Decryption During File Migration**.	AES-256-GCM
DEK	This parameter is displayed only when **Encryption** is set to **AES-256-GCM**. The key consists of 64 hexadecimal numbers. Remember the key configured here because the decryption key must be the same as that configured here. If the encryption and decryption keys are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B
IV	This parameter is displayed only when **Encryption** is set to **AES-256-GCM**. The initialization vector consists of 32 hexadecimal numbers. Remember the initialization vector configured here because the initialization vector used for decryption must be the same as that configured here. If the initialization vectors are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	5C91687BA886EDCD12ACBC3FF19A3C3F

NOTE

HDFS supports the **UTF-8** encoding only. Retain the default value **UTF-8**.

4.4.3 To HBase/CloudTable

If the destination link of a job is one of them listed in **Link to HBase** or **Link to CloudTable**, configure the destination job parameters based on **Table 4-24**.

Table 4-24 Parameter description

Parameter	Description	Example Value
Table Name	Name of the HBase table to which data will be written. If you want to create an HBase table, you can copy the field names from the migration source. Click the icon next to the text box. The dialog box for selecting the table is displayed. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	TBL_2
Clear Data Before Import	Whether the data in the destination table is cleared before data import. The options are as follows: • **Yes**: The data is cleared. • **No**: The data is not cleared. Instead, it will be added to the existing table.	Yes
Rowkey Delimiter	(Optional) Used to combine multiple columns as a rowkey. Spaces are used by default.	,
Rowkey Data Redundancy	(Optional) Whether to write the rowkey data into HBase columns. The default value is **No**.	No
Compression Format	(Optional) Compression format used in creating an HBase table. The default value is **None**. • **None**: The files are not compressed. • **Snappy**: The files are compressed in snappy format. • **gzip**: The files are compressed in gzip format.	None

Parameter	Description	Example Value
Write WAL	Whether to enable Write Ahead Log (WAL) of HBase. The options are as follows: • **Yes**: If the HBase server breaks down after the function is enabled, you can replay the operations that have not been performed in WAL. • **No**: If you set this parameter to **No**, the write performance is improved. However, if the HBase server breaks down, data may be lost.	No
Match Data Type	• **Yes**: Data of the Short, Int, Long, Float, Double, and Decimal columns in the source database is converted into Byte[] arrays (binary) and written into HBase. Other types of data are written as character strings. If several types of data mentioned above are combined as rowkeys, they will be written as character strings. This function saves storage space. In specific scenarios, the rowkey distribution is evener. • **No**: All types of data in the source database are written into HBase as character strings.	No

4.4.4 To Hive

If the destination link of a job is the **Link to Hive**, configure the destination job parameters based on **Table 4-25**.

Table 4-25 Parameter description

Parameter	Description	Example Value
Database Name	Database name. Click the icon next to the text box. The dialog box for selecting the database is displayed.	default

Parameter	Description	Example Value
Auto Table Creation	This parameter is displayed only when the source is a relational database. The options are as follows: • **Non-auto creation**: CDM will not automatically create a table. • **Auto creation**: If the destination database does not contain the table specified by **Table Name**, CDM will automatically create the table. If the table specified by **Table Name** already exists, no table is created and data is written to the existing table. • **Deletion before creation**: CDM deletes the table specified by **Table Name**, and then creates the table again.	Non-auto creation
Table Name	Destination table name. Click the icon next to the text box. The dialog box for selecting the table is displayed. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	TBL_X
Clear Data Before Import	Whether the data in the destination table is cleared before data import. The options are as follows: • **Yes**: The data is cleared. • **No**: The data is not cleared. Instead, it will be added to the existing table.	Yes

Parameter	Description	Example Value
Partition to Clear	This parameter is available when **Clear Data Before Import** is set to **Yes**. When you enter the information about the partitions to be cleared, the data in the partitions will be cleared.	Single partition: **year=2020,location=sun** Multiple partitions: **['year=2020,location=sun', 'year=2021,location=earth']**

NOTE

1. When Hive serves as the destination end, a table whose storage format is ORC is automatically created.
2. When Hive serves as the migration destination, if the storage format is TEXTFILE, delimiters must be explicitly specified in the statement for creating Hive tables. The following gives an example:

```
CREATE TABLE csv_tbl(
smallint_value smallint,
tinyint_value tinyint,
int_value int,
bigint_value bigint,
float_value float,
double_value double,
decimal_value decimal(9, 7),
timestmamp_value timestamp,
date_value date,
varchar_value varchar(100),
string_value string,
char_value char(20),
boolean_value boolean,
binary_value binary,
varchar_null varchar(100),
string_null string,
char_null char(20),
int_null int
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
"separatorChar" = "\t",
"quoteChar"     = "'",
"escapeChar"    = "\\"
)
STORED AS TEXTFILE;
```

4.4.5 To a Common Relational Database

Common relational databases serving as the destination include RDS for MySQL, RDS for SQL Server, and RDS for PostgreSQL.

To import data to the preceding data sources, configure the destination job parameters listed in **Table 4-26**.

Table 4-26 Parameter description

Category	Parameter	Description	Example Value
Basic parameters	Schema/ Tablespace	Name of the database to which data will be written. The schema can be automatically created. Click the icon next to the text box to select a schema or tablespace.	schema
	Auto Table Creation	This parameter is displayed only when the source is a relational database. The options are as follows: • **Non-auto creation**: CDM will not automatically create a table. • **Auto creation**: If the destination database does not contain the table specified by **Table Name**, CDM will automatically create the table. If the table specified by **Table Name** already exists, no table is created and data is written to the existing table. • **Deletion before creation**: CDM deletes the table specified by **Table Name**, and then creates the table again.	Non-auto creation
	Table Name	Name of the table to which data will be written. Click the icon next to the text box. The dialog box for selecting the table is displayed. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	table

Category	Parameter	Description	Example Value
	Clear Data Before Import	Whether to clear the data in the destination table before data import. The options are as follows: • **Do not clear**: The data in the destination table is not cleared before data import. The imported data is just added to the table. • **Clear all data**: All data is cleared from the destination table before data import. • **Clear part of data**: Part of the data in the destination table is cleared before data import. If you select **Clear part of data**, you must configure **WHERE Clause** to specify which part will be deleted.	Clear part of data
	WHERE Clause	If **Clear Data Before Import** is set to **Clear part of data**, data in the destination table will be deleted based on the WHERE clause after the configuration is complete and before the import starts.	age > 18 and age <= 60
	Constraint Conflict Handling	Mode for handling conflicts in data migration • **insert into**: When a primary key or unique index conflict occurs, data cannot be written and will become dirty data. • **replace into**: When a primary key or unique index conflict occurs, the original row is deleted and a new row is inserted to replace all the fields in the original row. • **on duplicate key update**: When a primary key or unique index conflict occurs in a row in the destination table, the data columns except the unique constraint column in this row are updated.	insert into
	Loader Threads	Number of threads started in each loader. A larger number allows more concurrent write operations. **NOTE** This parameter is unavailable if **Constraint Conflict Handling** is set to **replace into** or **on duplicate key update**.	1

Category	Parameter	Description	Example Value
Advanced parameters	Import to Staging Table	If you set this parameter to **Yes**, the transaction mode is enabled. CDM automatically creates a temporary table and imports data to the temporary table. After the data is imported successfully, it is migrated to the destination table in transaction mode. If the import fails, the destination table is rolled back to the state before the job starts. For details, see **Migration in Transaction Mode**. The default value is **No**, indicating that CDM directly imports the data to the destination table. In this case, if the job fails to be executed, the data that has been imported to the destination table will not be rolled back automatically. **NOTE** If you select **Clear part of data** or **Clear all data** for **Clear Data Before Import**, CDM does not roll back the deleted data in transaction mode.	No
	Extend Field Length	When **Auto creation** is selected, the length of the character fields can be extended to three times the original length and then written to the destination table. If the encoding types of the source and destination databases are different, but the character fields in the source and destination tables are the same, errors may occur during data migration due to character length difference. **NOTE** When this function is enabled, some fields consume three times the storage space of the user.	No
	Use NOT NULL Constraint	If you choose to create a target table automatically and specify the NOT NULL constraint, keep the NOT NULL constraints of the source and target tables consistent.	Yes
	Prepare for Data Import	The SQL statement that is first executed before a task is executed. Currently, only one SQL statement can be executed in wizard mode.	create temp table
	Complete Statement After Data Import	The SQL statement that is executed after a task is executed. Currently, only one SQL statement can be executed.	merge into

4.4.6 To DWS

If the destination link of a job is a **DWS link**, configure the destination job parameters based on **Table 4-27**.

Table 4-27 Parameter description

Category	Parameter	Description	Example Value
Basic parameters	Schema/ Tablespace	Name of the database to which data will be written. The schema can be automatically created. Click the icon next to the text box to select a schema or tablespace.	schema
	Auto Table Creation	This parameter is displayed only when the source is a relational database. The options are as follows: • **Non-auto creation**: CDM will not automatically create a table. • **Auto creation**: If the destination database does not contain the table specified by **Table Name**, CDM will automatically create the table. If the table specified by **Table Name** already exists, no table is created and data is written to the existing table. • **Deletion before creation**: CDM deletes the table specified by **Table Name**, and then creates the table again. **Field Mapping in Automatic Table Creation on DWS** describes the field mapping between the DWS tables created by CDM and source tables.	Non-auto creation

Category	Parameter	Description	Example Value
	Table Name	Name of the table to which data will be written. Click the icon next to the text box. The dialog box for selecting the table is displayed. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**. **NOTE** If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).	table
	Compress Data	Whether to compress data when data is imported to DWS and **Auto creation** is selected	No
	Storage Mode	When data is imported to DWS and **Auto Creation** is selected, you can specify the data storage mode: • **Row-based**: Row-based storage. It is used for point queries (index-based simple queries with fewer return records), or the scenario that requires a large number of addition, deletion, and modification operations. • **Column-based**: Column-based storage. It is used for statistical analysis queries (group and join scenarios) or ad hoc queries (query conditions are uncertain and indexes can hardly be used to scan row-based tables).	Row-based
	Import Mode	Mode for importing data to DWS • In COPY mode, the source data is copied to the DataNode of DWS after passing through the management node. • In UPSERT mode, if a primary key or unique constraint conflict occurs, other data columns, except the primary key and unique constraint column, are updated.	COPY

Category	Parameter	Description	Example Value
	Clear Data Before Import	Whether to clear the data in the destination table before data import. The options are as follows: • **Do not clear**: The data in the destination table is not cleared before data import. The imported data is just added to the table. • **Clear all data**: All data is cleared from the destination table before data import. • **Clear part of data**: Part of the data in the destination table is cleared before data import. If you select **Clear part of data**, you must configure **WHERE Clause** to specify which part will be deleted.	Clear part of data
	WHERE Clause	If **Clear Data Before Import** is set to **Clear part of data**, data in the destination table will be deleted based on the WHERE clause after the configuration is complete and before the import starts.	age > 18 and age <= 60
	Constraint Conflict Handling	Mode for handling conflicts in data migration • **insert into**: When a primary key or unique index conflict occurs, data cannot be written and will become dirty data. • **replace into**: When a primary key or unique index conflict occurs, the original row is deleted and a new row is inserted to replace all the fields in the original row. • **on duplicate key update**: When a primary key or unique index conflict occurs in a row in the destination table, the data columns except the unique constraint column in this row are updated.	insert into
	Loader Threads	Number of threads started in each loader. A larger number allows more concurrent write operations. **NOTE** This parameter is unavailable if **Constraint Conflict Handling** is set to **replace into** or **on duplicate key update**.	1

Category	Parameter	Description	Example Value
Advanced parameters	Import to Staging Table	If you set this parameter to **Yes**, the transaction mode is enabled. CDM automatically creates a temporary table and imports data to the temporary table. After the data is imported successfully, it is migrated to the destination table in transaction mode. If the import fails, the destination table is rolled back to the state before the job starts. . The default value is **No**, indicating that CDM directly imports the data to the destination table. In this case, if the job fails to be executed, the data that has been imported to the destination table will not be rolled back automatically. **NOTE** If you select **Clear part of data** or **Clear all data** for **Clear Data Before Import**, CDM does not roll back the deleted data in transaction mode.	No
	Extending field length	When **Auto creation** is selected, the length of the character fields can be extended to three times the original length and then written to the destination table. If the encoding types of the source and destination databases are different, but the character fields in the source and destination tables are the same, errors may occur during data migration due to character length difference. When a character field containing Chinese characters is imported to DWS, the length of the character field must be automatically increased by three times. If a job fails to be executed and an error message similar to **value too long for type character varying** exists in the log when you import Chinese characters to DWS, you can enable this function to solve the problem. **NOTE** When this function is enabled, some fields consume three times the storage space of the user.	No
	Use NOT NULL Constraint	If you choose to create a target table automatically and specify the NOT NULL constraint, keep the NOT NULL constraints of the source and target tables consistent.	Yes

Category	Parameter	Description	Example Value
	Prepare for Data Import	The SQL statement that is first executed before a task is executed. Currently, only one SQL statement can be executed in wizard mode.	create temp table
	Complete Statement After Data Import	The SQL statement that is executed after a task is executed. Currently, only one SQL statement can be executed.	merge into

Field Mapping in Automatic Table Creation on DWS

Figure 4-10 describes the field mapping between DWS tables created by CDM and source tables. For example, if you use CDM to migrate the Oracle database to DWS, CDM automatically creates a table on DWS and maps the **NUMBER(3,0)** field of the Oracle database to the **SMALLINT** field of DWS.

Figure 4-10 Field mapping in automatic table creation

Source Database Type							Destination Database Type
Oracle	MySQL	SQL Server	PostgreSQL	Db2	GaussDB	SAP HANA	DWS
NUMBER(p,0) (p=3 or p=5)	SMALLINT,TINYINT	SMALLINT,TINYINT	SMALLINT	DECIMAL	SMALLINT	SMALLINT,TINYINT	SMALLINT
NUMBER(10,0)	INT	INT	INTEGER	INT	INTEGER	INTEGER	INTEGER
NUMBER(19,0)	BIGINT	BIGINT	BIGINT	DECIMAL	BIGINT	BIGINT	BIGINT
None	None	None	OID	None	OID	CHAR(128)	OID
NUMBER(p,s) (0 < p <= 38)	DECIMAL(p,s) (0 < p <= 65)	DECIMAL(p,s) (0 < p <= 30)	NUMERIC(p,s) (p <= 1000)	DECIMAL	NUMERIC(p,s) (p <= 1000)	DECIMAL(p,s) (0 < p <= 38)	NUMERIC(p,s) (p <= 1000)
RAW	BINARY	BINARY	BYTEA	BINARY	BYTEA	BINARY	BYTEA
CHAR	CHAR	CHAR	CHAR	CHAR	CHAR	CHAR(p) (p <= 2000)	CHAR
NCHAR	NCHAR	NCHAR	NCHAR	NCHAR	NCHAR	NCHAR(p) (p <= 5000)	NCHAR
DATE	DATE	DATE	DATE	DATE	DATE	DATE	DATE
DATE	DATETIME	DATETIME2	TIMESTAMP	TIMESTAMP	TIMESTAMP	TIMESTAMP	TIMESTAMP
VARCHAR2(p) (p <= 4000)	VARCHAR	VARCHAR(p) (if p >= 8000 p=max)	VARCHAR(p) (p <= 10485760)	VARCHAR	VARCHAR(p) (p <= 10485760)	VARCHAR(p) (p <= 5000)	VARCHAR(p) (p <= 10485760)
FLOAT	DOUBLE	FLOAT	DOUBLE PRECISION	FLOAT	DOUBLE PRECISION	DOUBLE	DOUBLE PRECISION
FLOAT	REAL	FLOAT	REAL	FLOAT	REAL	REAL	REAL
CLOB	TEXT	TEXT	TEXT	TEXT	TEXT	CLOB	TEXT
DATE	None	TIME	TIME	TIME	TIME	TIME	TIME
BOOLEAN	None	None	BOOLEAN	BOOLEAN	BOOLEAN	BOOLEAN	BOOLEAN

4.4.7 To DDS

If the destination link of a job is the **Link to DDS**, configure the destination job parameters based on **Table 4-28**.

Table 4-28 Parameter description

Parameter	Description	Example Value
Database Name	Database to which data is to be imported	mongodb
Collection Name	Collection of data to be imported, which is similar to the table name of a relational database. Click the icon next to the text box to go to the page for selecting the table or directly enter a table name. If the desired table is not displayed, check whether the table exists or whether the login account has the permission to query metadata.	COLLECTION

4.4.8 To DCS

If the data is imported to DCS, configure the destination job parameters based on **Table 4-29**.

Table 4-29 Parameter description

Parameter	Description	Example Value
Redis Key Prefix	Key prefix, which is similar to the table name of a relational database	TABLE
Value Storage Type	The options are as follows: • **String**: without column name, such as **value1,value2** • **Hash**: with column name, such as **column1=value1,column2=value2**	String
Key Delimiter	Character used to separate table names and column names of a relational database	_
Value Delimiter	Character used to separate columns when the storage type is string	;

4.4.9 To CSS

If the destination link of a job is the **Link to Elasticsearch/CSS**, that is, when data is imported to CSS, configure the destination job parameters based on **Table 4-30**.

Table 4-30 Parameter description

Parameter	Description	Example Value
Index	Elasticsearch index, which is similar to the name of a relational database. CDM supports automatic creation of indexes and field types. The index and field type names can contain only lowercase letters.	index
Type	Elasticsearch type, which is similar to the table name of a relational database. The type name can contain only lowercase letters. **NOTE** Elasticsearch 7.x and later versions do not support custom types. Instead, only the **_doc** type can be used. In this case, this parameter does not take effect even if it is set.	type
Pipeline ID	Pipeline used to convert the data format after data is transferred to Elasticsearch. Pipeline IDs are ready for use after being created in Kibana.	pipeline_id
Periodically Create Index	For streaming jobs that continuously write data to Elasticsearch, CDM periodically creates indexes and writes data to the indexes, which helps you delete expired data. The indexes can be created based on the following periods: • **Every hour**: CDM creates indexes on the hour. The new indexes are named in the format of *Index name+Year+Month+Day+Hour*, for example, **index2018121709**. • **Every day**: CDM creates indexes at 00:00 every day. The new indexes are named in the format of *Index name+Year+Month+Day*, for example, **index20181217**. • **Every week**: CDM creates indexes at 00:00 every Monday. The new indexes are named in the format of *Index name+Year+Week*, for example, **index201842**. • **Every month**: CDM creates indexes at 00:00 on the first day of each month. The new indexes are named in the format of *Index name+Year +Month*, for example, **index201812**. • **Do not create**: Do not create indexes periodically. When extracting data from a file, you must configure a single extractor, which means setting **Concurrent Extractors** to **1**. Otherwise, this parameter is invalid.	Every hour

4.4.10 To DLI

If the destination link of a job is the **Link to DLI**, configure the destination job parameters based on **Table 4-31**.

NOTE

When you use CDM to migrate data to DLI, DLI generates data files in the *dli-trans** temporary OBS bucket. Therefore, you need to grant the account corresponding to the AK/SK the permissions to read and write the *dli-trans** bucket and create directories. For details about how to add OBS permission policies, see **Adding an OBS Bucket Policy**.

Table 4-31 Parameter description

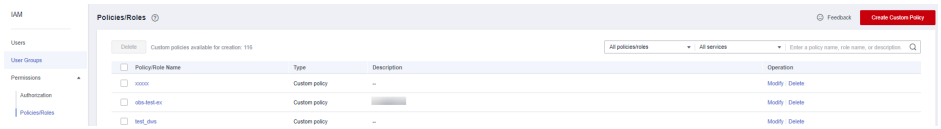
Parameter	Description	Example Value
Resource Queue	Resource queue to which the destination table belongs The default queue of DLI cannot be used for migration jobs. You need to create a SQL queue in DLI.	cdm
Database Name	Name of the database to which data will be written	dli
Table Name	Name of the table to which data will be written	car_detail
Clear Data Before Import	Whether to clear data in the destination table before data import If this parameter is set to **Yes**, data in the destination table will be cleared before the task is started.	No
Data Clearing Mode	This parameter is available when **Clear Data Before Import** is set to **Yes**. **TRUNCATE**: deletes standard data. **INSERT_OVERWRITE**: overwrites existing data with inserted data.	TRUNCATE
Partition	This parameter is available when **Clear Data Before Import** is set to **Yes**. When you enter partitions, data in these partitions will be cleared.	year=2020,location=sun

Adding an OBS Bucket Policy

Step 1 Log in to the IAM console.

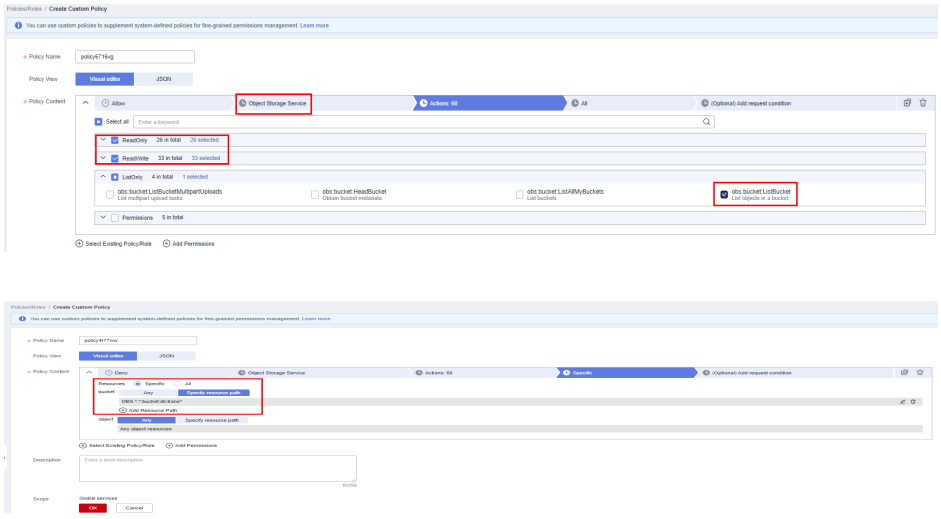
Step 2 In the navigation pane, choose **Permissions** > **Policies/Roles** and click **Create Custom Policy** in the upper right corner.

Figure 4-11 Creating a custom policy



Step 3 Enter a policy name and set **Policy Content**.

Figure 4-12 Configuring the policy



Step 4 Enter the policy description and click **OK**.

----End

4.5 Scheduling Job Execution

CDM supports scheduled execution of table/file migration jobs by minute, hour, day, week, and month. This section describes how to configure scheduled job parameters.

NOTE

- When configuring scheduled jobs, do not set the same scheduled time for different jobs. Instead, set different times to avoid exceptions.
- If you use DataArts Studio DataArts Factory to schedule the CDM migration job and configure this parameter, both configurations take effect. To ensure unified service logic and avoid scheduling conflicts, enable job scheduling in DataArts Factory and do not configure a scheduled task for the job in DataArts Migration.

Scheduling Job Execution by Minute

CDM allows jobs to be executed every several minutes. It is recommended that the cycle be at least 5 minutes.

- **Start Time**: indicates the time when the scheduled configuration takes effect, or the first time when the job is automatically executed.
- **Cycle (minutes)**: indicates the interval when a job is executed starting from the start time.

- **End Time**: This parameter is optional. If it is not set, the scheduled job keeps being automatically executed. If it is set, the scheduled job will be automatically stopped at the end time.

Figure 4-13 Scheduling job execution by minute

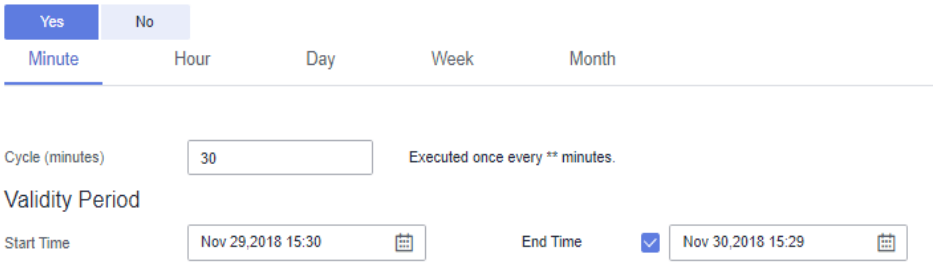


Figure 4-13 shows that the job will be automatically executed at 15:30:30 on November 29, 2018 for the first time at a cycle of 30 minutes, and will be automatically stopped at 15:29:00 on November 30, 2018.

Scheduling Job Execution by Hour

CDM allows jobs to be executed every several hours.

- **Cycle (hours)**: indicates the interval when a job is automatically executed.
- **Trigger Time (minute)**: indicates the exact time in each hour when a scheduled task is triggered. The value ranges from 0 to 59. You can set a maximum of 60 values and use commas (,) to separate these values. However, the values must be unique.

  If the trigger time is not within the validity period, the system selects a trigger time closest to the validity period for the scheduled job to be automatically executed at the first time. The following gives an example:

  - **Start Time**: **1:20:00**
  - **Cycle (hours)**: **3**
  - **Trigger Time (minute)**: **10**
- **Validity Period**: includes **Start Time** and **End Time**.
  - **Start Time**: indicates the time when the scheduled configuration takes effect.
  - **End Time**: This parameter is optional, which indicates the time when the scheduled job is automatically stopped. If this parameter is not set, the scheduled job keeps being automatically executed.

Figure 4-14 Scheduling job execution by hour

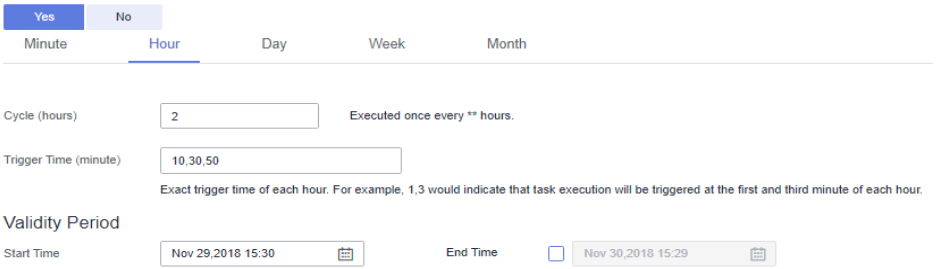


Figure 4-14 shows that the scheduled configuration will take effect at 15:30:00 on November 30, 2018. The job is automatically executed for the first time upon the scheduled configuration takes effect, at 15:50:00 for the second time, and at 17:10:00 for the third time. The job is triggered for three times every 2 hours and the configuration is always valid.

Scheduling Job Execution by Day

CDM allows jobs to be executed every several days.

- **Cycle (days)**: indicates the interval when a job is executed starting from the start time.
- **Validity Period**: includes **Start Time** and **End Time**.
  - **Start Time**: indicates the time when the scheduled configuration takes effect, or the first time when the job is automatically executed.
  - **End Time**: This parameter is optional, which indicates the time when the scheduled job is automatically stopped. If this parameter is not set, the scheduled job keeps being automatically executed.

Figure 4-15 Scheduling job execution by day

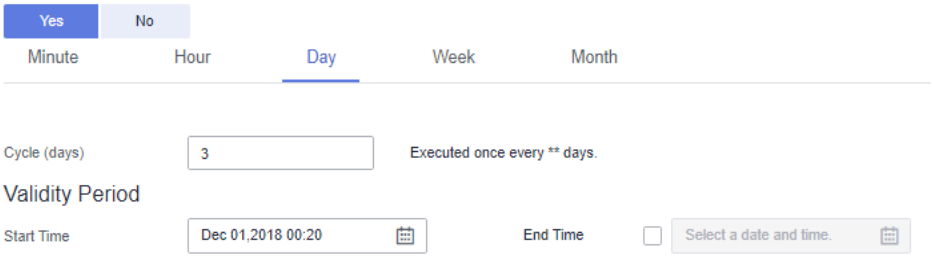


Figure 4-15 shows that the scheduled job will be automatically executed at 00:20:00 on December 1, 2018, and is executed once every three days. The configuration is always valid.

Scheduling Job Execution by Week

CDM allows jobs to be executed every several weeks.

- **Cycle (weeks)**: indicates the interval when a scheduled job is executed starting from the start time.
- **Trigger Time (day)**: You can specify the day of each week when the job is automatically executed. One or more days can be selected at a time.
- **Validity Period**: includes **Start Time** and **End Time**.
  - **Start Time**: indicates the time when the scheduled configuration takes effect.
  - **End Time**: This parameter is optional, which indicates the time when the scheduled job is automatically stopped. If this parameter is not set, the scheduled job keeps being automatically executed.

Figure 4-16 Scheduling job execution by week

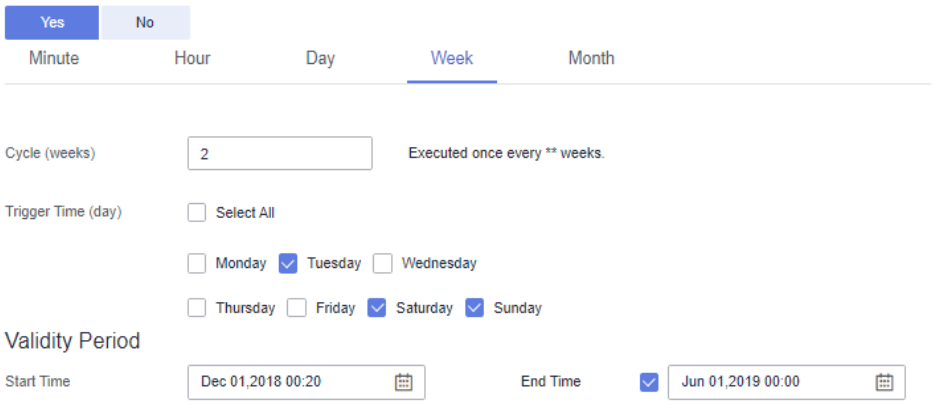


Figure 4-16 shows that the job will be automatically executed at 00:20:00 every Tuesday, Saturday, and Sunday every two weeks starting from 00:20:00 on December 1, 2018, and the job will be automatically stopped at 00:00:00 on June 1, 2019.

Scheduling Job Execution by Month

CDM allows jobs to be executed every several months.

- **Cycle (months)**: indicates the interval when a scheduled job is executed starting from the start time.
- **Trigger Time (day)**: indicates the day of each month when the job is executed. The value ranges from 1 to 31. You can set multiple values and use commas (,) to separate these values. However, the values must be unique.
- **Validity Period**: includes **Start Time** and **End Time**.
  - **Start Time**: indicates the time when the scheduled configuration takes effect. The automatic execution time is accurate to hour, minute, and second.
  - **End Time**: This parameter is optional, which indicates the time when the scheduled job is automatically stopped. If this parameter is not set, the scheduled job keeps being automatically executed.

Figure 4-17 Scheduling job execution by month

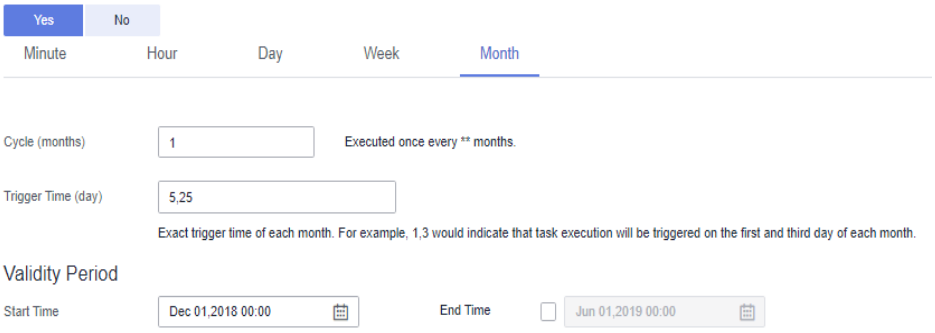


Figure 4-17 shows that the job will be automatically executed at 00:00:00 on every fifth and twenty-fifth day of each month starting from 00:00:00 on December 1, 2018. The configuration is always valid.

4.6 Job Configuration Management

On the **Settings** tab page, you can perform the following operations:

- **Maximum Concurrent Extractors**
- **Scheduled Backup/Restoration**
- **Environment Variables of Job Parameters**

Maximum Concurrent Extractors

Maximum number of concurrent extraction tasks in a cluster

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.

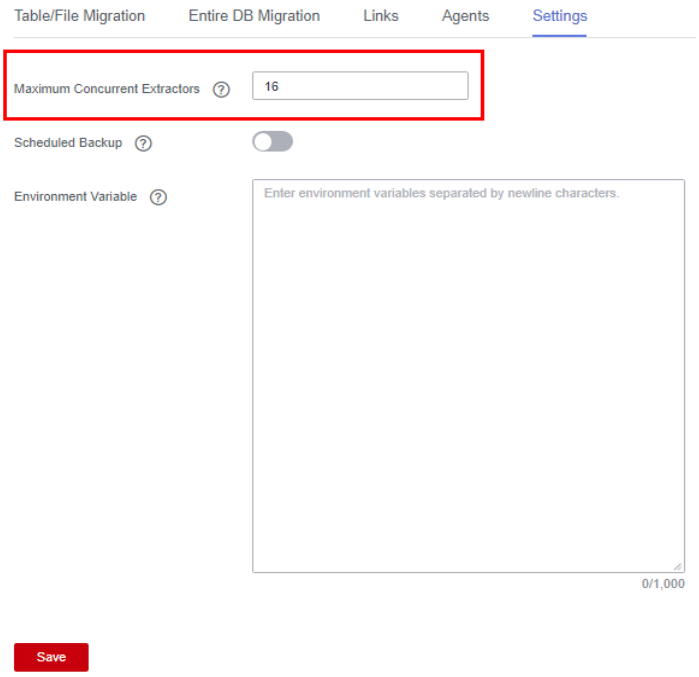
By setting an appropriate number of concurrent extractors for a job and the maximum number of concurrent extractors for the cluster, you can accelerate migration. You can configure the number of concurrent extractors as follows:

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 4-32 Maximum number of concurrent extractors for a CDM cluster

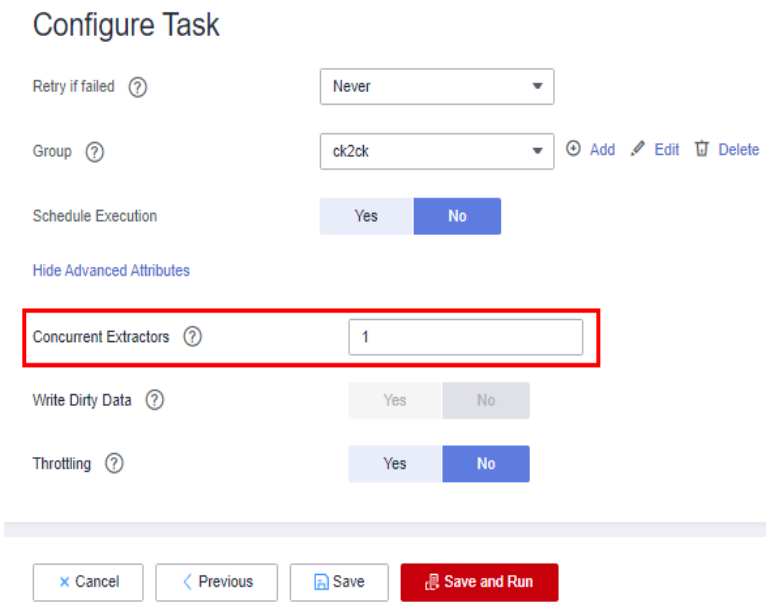
Flavor	vCPUs/Memory	Maximum Concurrent Extractors
cdm.large	8 vCPUs, 16 GB	16
cdm.xlarge	16 vCPUs, 32 GB	32
cdm.4xlarge	32 vCPUs, 64 GB	64

Figure 4-18 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 4-19 Setting Concurrent Extractors for a job



Scheduled Backup/Restoration

This function depends on the OBS service.

- Prerequisites

  You have created the **Link to OBS**.

- Scheduled backup

  On the **Job Management** page, click **Settings** and configure **Scheduled Backup** and its related parameters.

  Table 4-33 Scheduled backup parameters

Parameter	Description	Example Value
Scheduled Backup	Whether to enable automatic backup. This function is used to back up jobs but not links.	Enable
Backup Policy	• **All jobs**: CDM backs up all table/file migration jobs and entire DB migration jobs regardless of the job statuses. However, historical jobs are not backed up. • **All jobs by groups**: You select one or more job groups to back up.	All jobs
Backup Cycle	Select the backup cycle. • **Day**: The backup is performed daily at 00:00:00. • **Week**: The backup is performed at 00:00:00 every Monday. • **Month**: The backup is performed at 00:00:00 on the first day of each month.	Day
OBS Link for Writing Backups	Link used to back up jobs to OBS buckets. Select a link you have created on the **Links** page.	obslink
OBS Bucket	OBS bucket where backup files are stored	cdm
Backup Data Directory	Directory where backup files are stored	/cdm-bk/

- Restoring jobs

  If automatic backup has been performed, the backup list is displayed on the **Configuration Management** tab page. The OBS buckets where the backup files reside, backup paths, and backup time are displayed.

  You can click **Restore Backup** in the **Operation** column of the backup list to restore the CDM jobs.

Environment Variables of Job Parameters

When creating a migration job on CDM, the parameter (such as the OBS bucket name or file path) that can be manually configured, a field in a parameter, or a character in a field can be configured as a global variable, so that you can change parameter values in batches, or batch replace certain characters after jobs are exported or imported.

The following describes how to batch replace the OBS bucket name in a migration job.

1. On the **Job Management** page, click the **Configuration Management** tab and configure environment variables.

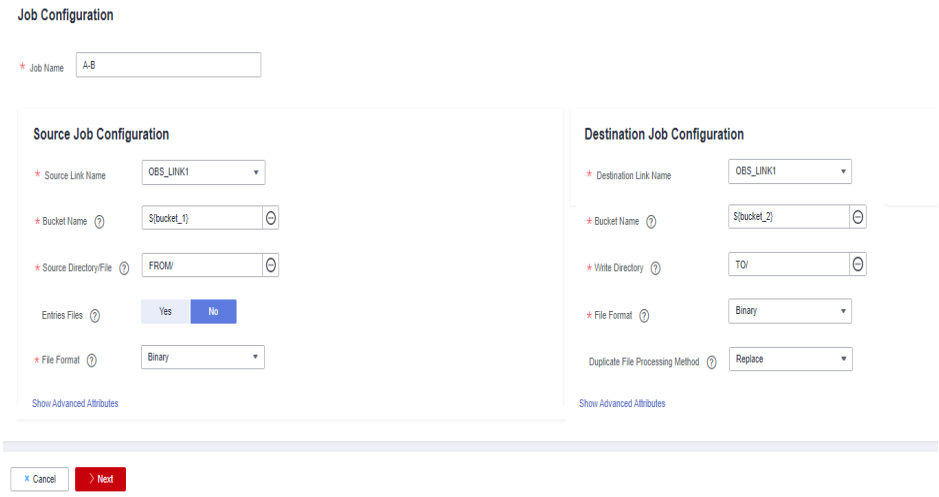
   ```
   bucket_1=A
   bucket_2=B
   ```

   Variable **bucket_1** indicates bucket A, and variable **bucket_2** indicates bucket B.

2. On the page for creating a CDM migration job, migrate data from bucket A to bucket B.

   Set the source bucket name to **${bucket_1}** and destination bucket name to **${bucket_2}**.

   Figure 4-20 Setting the bucket names to environment variables

3. If you want to migrate data from bucket C to bucket D, you do not need to change the job parameters. You only need to change the environment variables on the **Configuration Management** tab page as follows:

   ```
   bucket_1=C
   bucket_2=D
   ```

4.7 Managing a Single Job

Existing CDM jobs can be viewed, modified, deleted, started, and stopped. This section describes how to view and modify a job.

Viewing a Job

- **Viewing job status**

The job status can be **New**, **Pending**, **Booting**, **Running**, **Failed**, or **Succeeded**.

Pending indicates that the job is waiting to be scheduled by the system, and **Booting** indicates that the data to be migrated is being analyzed.

- **Viewing the historical records**

  On the **Historical Record** page, you can view job execution records, read/write statistics, and job execution logs.

- **Viewing job logs**

  On the **Historical Record** page, you can view all logs of a job.

  Alternatively, in the **Operation** column, choose **More** > **Log** to view the latest logs of the job.

- **Viewing the JSON file of a job**

  You can directly edit the JSON file of a job, which is equivalent to modifying the parameter settings of the job.

- **Querying the job statistics**

  You can open the preview window of a configured database job and view up to 1,000 pieces of data. By comparing the number of data records of the migration source and destination, you can check whether the migration was successful and whether data was lost.

- **Viewing historical jobs**

  CDM stores the jobs executed in the last month, including one-time jobs (jobs that are automatically deleted after execution) and jobs that are executed periodically. You can view and re-execute the jobs on the **Historical Jobs** tab page.

  For a job that is executed periodically, a historical job is generated on the **Historical Jobs** tab page each time when the job is executed, regardless of whether the job is executed successfully. The names of historical jobs will be the same as the original job but with a random character string appended.

Modifying a Job

- **Modifying the job parameters**

  You can reconfigure job parameters, but you cannot reselect source and destination links.

- **Editing the JSON file of a job**

  You can directly edit the JSON file of a job, which is equivalent to modifying the parameter settings of the job.

Procedure

Step 1 Log in to the management console and choose **Service List** > **Cloud Data Migration**. In the left navigation pane, choose **Cluster Management**. Locate the target cluster and click **Job Management**.

Step 2 Click **Historical Jobs** to view all historical jobs executed in the latest month.

CDM stores the jobs executed in the last month, including one-time jobs (jobs that are automatically deleted after execution) and jobs that are executed periodically. You can view and re-execute the jobs on the **Historical Jobs** tab page.

For a job that is executed periodically, a historical job is generated on the **Historical Jobs** tab page each time when the job is executed, regardless of whether the job is executed successfully. The names of historical jobs will be the same as the original job but with a random character string appended.

Step 3 Click **Table/File Migration**. The job list is displayed. You can perform the following operations on a single job:

- Modify the job parameters: Click **Edit** in the **Operation** column to modify the job parameters.
- Run the job: Click **Run** in the **Operation** column to manually start the job.
- View the historical records: Click **Historical Record** in the **Operation** column. On the **Historical Record** page that is displayed, view the job's historical execution records and read/write statistics. Click **Log** to view the job logs.
- Delete the job: Choose **More** > **Delete** in the **Operation** column to delete the job.
- Stop the job: Choose **More** > **Stop** in the **Operation** column to stop the job.
- View the job JSON: Choose **More** > **View Job JSON** in the **Operation** column to view the job JSON.
- Edit the job JSON: Choose **More** > **Edit Job JSON** in the **Operation** column to edit the job JSON files, which is similar to modify the job parameters.
- Configure a scheduled job: Locate a job and choose **More** > **Configure Scheduled Execution**. You can set the cycle for periodically executing the job. For details, see **Scheduling Job Execution**.

Step 4 After the modification, click **Save** or **Save and Run**.

----End

4.8 Managing Jobs in Batches

Scenario

This section describes how to manage CDM table/file migration jobs in batches. The following operations are involved:

- Manage jobs by group.
- Run jobs in batches.
- Delete jobs in batches.
- Export jobs in batches.
- Import jobs in batches.

You can export and import jobs in batches in the following scenarios:

- Job migration between CDM clusters: You can migrate jobs from a cluster of an earlier version to a new version.
- Job backup: You can stop or delete CDM clusters to reduce costs. In this case, you can export the job scripts in batches and save them, and create a cluster and import the job scripts if necessary.
- Batch job creation: You can manually create a job and export the job configuration file in JSON format. Copy the content in the JSON file to the

same file or new files, and then import the file/files to CDM to create jobs in batches.

Procedure

Step 1 Log in to the management console and choose **Service List** > **Cloud Data Migration**. In the left navigation pane, choose **Cluster Management**. Locate the target cluster and click **Job Management**.

Step 2 Click **Table/File Migration**. The job list is displayed. You can perform the following batch operations:

- **Manage jobs by group.**

  CDM allows users to add, modify, search for, and delete job groups. When a group is deleted, all jobs in the group are deleted.

  In the third step of creating a job, if jobs have been assigned to different groups, you can display, start, or export jobs by group.

- **Run jobs in batches.**

  After selecting one or more jobs, click **Run** to start these jobs in batches.

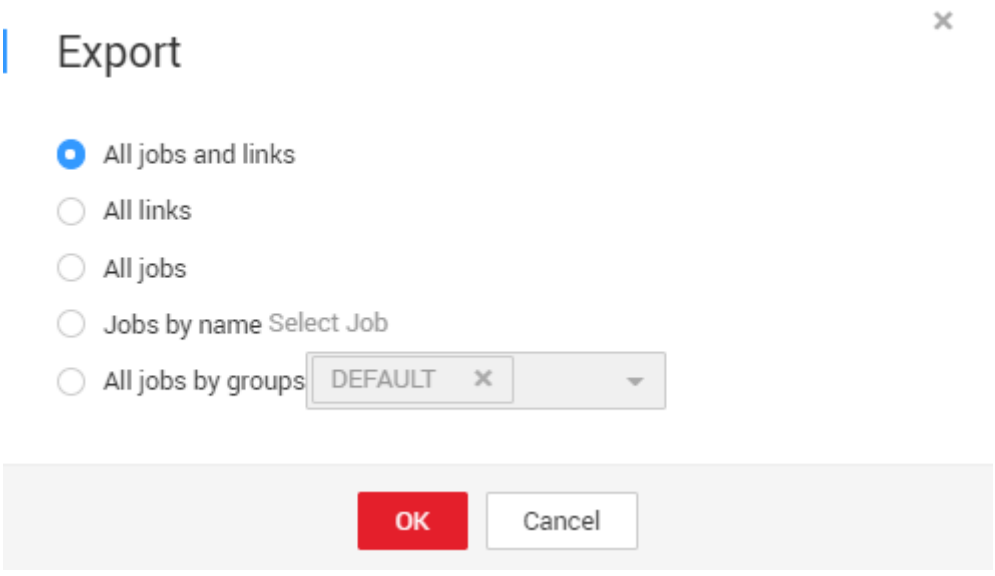
- **Delete jobs in batches.**

  After selecting one or more jobs, click **Delete** to delete these jobs in batches.

- **Export jobs in batches.**

  Click **Export**.

  Figure 4-21 Export

  - **All jobs and links**: Export all jobs and links at a time.
  - **All jobs**: Export all jobs at a time.
  - **All links**: Export all links at a time.
  - **Jobs by name**: Select the jobs to export and click **OK**.
  - **All jobs by groups**: Select the group to export and click **OK**.

Exported jobs are stored in JSON files, which can be used as backups or imported to other clusters.

NOTE

For security purposes, no link password is exported when jobs are exported. All passwords are replaced by *Add password here*.

- **Import jobs in batches.**

  Click **Import** and select the import format (text file or JSON).

  - **By JSON string**: Job files to be imported must be in JSON format and the file size cannot exceed 1 MB. If the job files to be imported are exported from CDM, edit the JSON files before importing them to CDM. Replace *Add password here* with the correct link passwords.
  - **By text file**: This mode can be used when the local JSON files cannot be uploaded properly. Paste the JSON strings for the jobs into the text box.

----End

5 Auditing

5.1 Key CDM Operations Recorded by CTS

CTS provides records of operations on cloud service resources. With CTS, you can query, audit, and backtrack those operations.

Table 5-1 CDM operations recorded by CTS

Operation	Resource Type	Trace Name
Creating a cluster	cluster	createCluster
Deleting a cluster	cluster	deleteCluster
Modifying cluster configurations	cluster	modifyCluster
Starting a cluster	cluster	startCluster
Restarting a cluster	cluster	startStopCluster
Importing a job	cluster	clusterImportJob
Binding an EIP	cluster	bindEip
Unbinding an EIP	cluster	unbindEip
Creating a link	link	createLink
Modifying a link	link	modifyLink
Deleting a link	link	deleteLink
Creating a job	job	createJob
Modifying a job	job	modifyJob
Deleting a job	job	deleteJob
Starting a job	job	startJob

Operation	Resource Type	Trace Name
Stopping a job	job	stopJob

5.2 Viewing Traces

Scenario

After you enable CTS, the system starts to record the CDM operations. The management console of CTS stores the traces of the latest seven days.

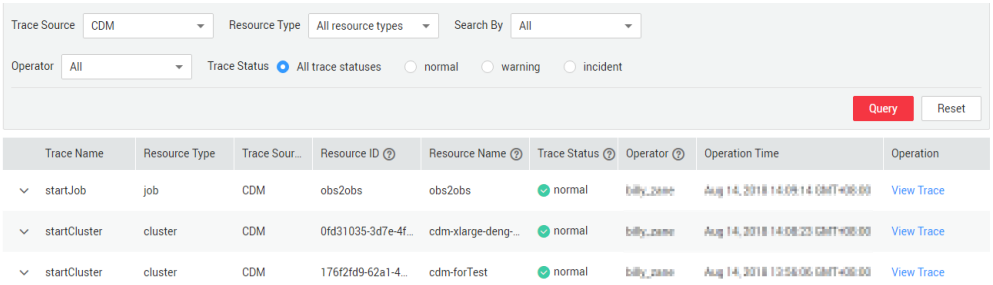
This section describes how to query these traces.

Procedure

1. Log in to the management console.
2. Click **Service List**, and choose **Management & Deployment** > **Cloud Trace Service**.
3. In the left navigation pane, click **Trace List**.

   Click **Filter** and specify filter criteria as needed.

   Figure 5-1 CDM traces



4. Unfold the target trace to view its details.
5. Click **View Trace** in the **Operation** column to view the trace structure details.

   For more information about CTS, see **Cloud Trace Service User Guide**.

6 Tutorials

6.1 Creating an MRS Hive Link

MRS Hive links are applicable to the MapReduce Service (MRS). This tutorial describes how to create an MRS Hive link.

Prerequisites

- You have created a CDM cluster.
- You have obtained the Manager IP address, and administrator account and password of the MRS cluster, and the account has the permissions to import and export data.
- The MRS cluster and the CDM cluster can communicate with each other. The following requirements must be met for network interconnection:
  - If the CDM cluster and the cloud service are in different regions, a public network or a dedicated connection is required for enabling communication between the CDM cluster and the cloud service. If the Internet is used for communication, ensure that an EIP has been bound to the CDM cluster, the host where the data source is located can access the Internet, and the port has been enabled in the firewall rules.
  - If the CDM cluster and the cloud service are in the same region, VPC, subnet, and security group, they can communicate with each other by default. If they are in the same VPC but in different subnets or security groups, you must configure routing rules and security group rules. For details about how to configure routing rules, see **configuring routes**. For details about how to configure security group rules, see **configuring security group rules**.
  - The cloud service instance and the CDM cluster belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

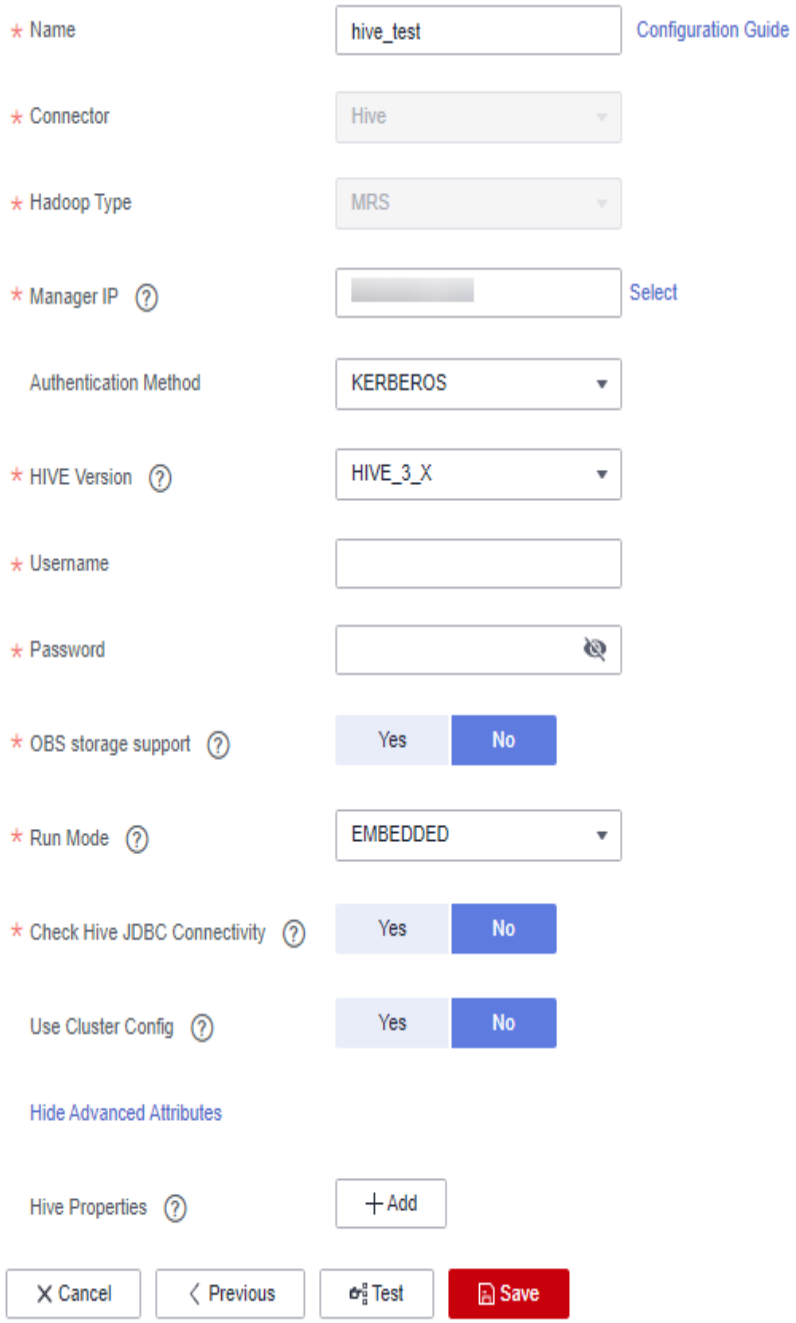
Creating an MRS Hive Link

Step 1 On the **Cluster Management** page, locate a cluster and click **Job Management** in the **Operation** column. On the displayed page, click the **Links** tab and then **Create Link**.

Figure 6-1 Selecting a connector type

Step 2 Select **MRS Hive** and click **Next** to configure parameters for the MRS Hive link.

Figure 6-2 Creating an MRS Hive link



Step 3 Click **Show Advanced Attributes** and set optional parameters. For details, see **Link to a Common Relational Database**. Retain the default values for the optional parameters and configure the mandatory parameters according to **Table 6-1**.

Table 6-1 MRS Hive link parameters

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	mrs-link
Manager IP	Floating IP address of MRS Manager. Click **Select** next to the **Manager IP** text box to select an MRS cluster. CDM automatically fills in the authentication information.	127.0.0.1
Authentication Method	Authentication method used for accessing MRS • **SIMPLE**: Select this for non-security mode. • **KERBEROS**: Select this for security mode.	SIMPLE
HIVE Version	Set this to the Hive version on the server.	HIVE_3_X
Username	If **Authentication Method** is set to **KERBEROS**, you must provide the username and password used for logging in to MRS Manager. If you need to create a snapshot when exporting a directory from HDFS, the user configured here must have the administrator permission on HDFS. 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.	cdm
Password	Password used for logging in to MRS Manager	-

Parameter	Description	Example Value
OBS storage support	The server must support OBS storage. When creating a Hive table, you can store the table in OBS.	No
Run Mode	This parameter is used only when the Hive version is **HIVE_3_X**. Possible values are: • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **Standalone**: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select **STANDALONE** or configure different agents. Note: The STANDALONE mode is used to solve the version conflict problem. If the connector versions of the source and destination ends of the same link are different, a JAR file conflict occurs. In this case, you need to place the source or destination end in the STANDALONE process to prevent the migration failure caused by the conflict.	EMBEDDED
Check Hive JDBC Connectivity	Whether to check the Hive JDBC connectivity	Disabled
Use Cluster Config	You can create cluster configurations on the **Links** page to simplify the configuration of Hadoop link parameters.	No
Hive Properties	Other parameters for the Hive client	-

NOTE

Click **Show Advanced Attributes**, and then click **Add** to add configuration attributes of other clients. The name and value of each attribute must be configured. You can click **Delete** to delete no longer used attributes.

Step 4 Click **Save** to return to the **Links**page.

----End

6.2 Creating a MySQL Link

MySQL links are applicable to third-party cloud MySQL services and MySQL created in a local data center or ECS. This tutorial describes how to create a MySQL link.

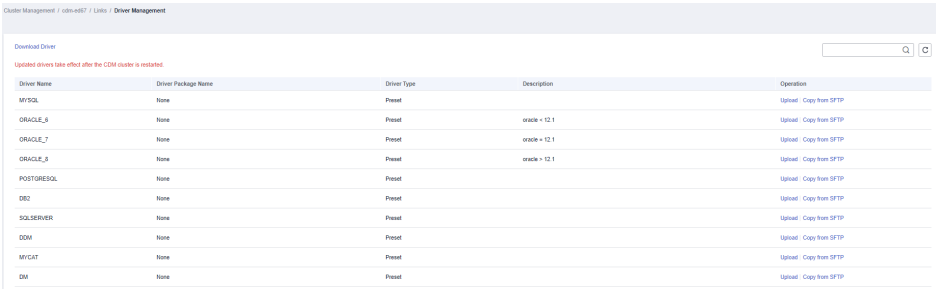
Prerequisites

- You have obtained the IP address, port, database name, username, and password for connecting to the MySQL database. In addition, the user must have the read and write permissions on the MySQL database.
- The on-premises MySQL database can be accessed through the public network. If the MySQL database is deployed on an on-premises data center or a third-party cloud, ensure that an IP address that can be accessed from the public network has been configured for the MySQL database, or the VPN or Direct Connect between the on-premises data center and the cloud service platform has been established.
- You have created a CDM cluster.

Creating a MySQL Link

Step 1 Access the CDM console, choose **Cluster Management** in the navigation pane, locate the target cluster, and choose **Job Management** > **Link Management** > **Driver Management**. The **Driver Management** page is displayed.

Figure 6-3 Uploading a driver



Step 2 In the upper left corner of the **Driver Management** page, click **Download Driver** to download the MySQL driver. For details, see **How Do I Obtain a Driver?**.

Step 3 On the **Driver Management** page, upload the MySQL driver using either of the following methods:

Click **Upload** in the **Operation** column and select a local driver.

Alternatively, click **Copy from SFTP** in the **Operation** column and configure the **SFTP Link** name and **Driver File Path**.

Step 4 On the **Cluster Management** page, click **Job Management** of the cluster and choose **Links** > **Create Link** to enter the page for selecting the connector, as shown in **Figure 6-4**.

Figure 6-4 Selecting a connector type

Step 5 Select **MySQL** and click **Next** to configure parameters for the MySQL link.

Figure 6-5 Creating a MySQL link

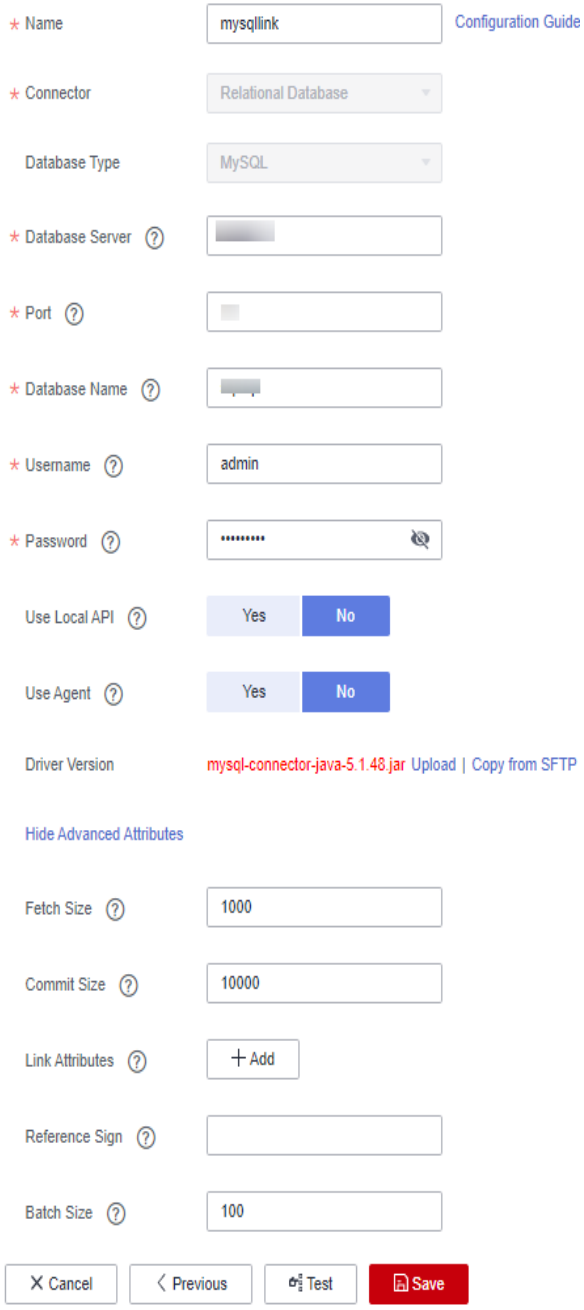


Table 6-2 MySQL link parameters

Parameter	Description	Example Value
Name	Enter a unique link name.	mysqllink
Database Server	IP address or domain name of the MySQL database	192.168.1.110
Port	MySQL database port	3306

Parameter	Description	Example Value
Database Name	Name of the MySQL database	sqoop
Username	User who has the read, write, and delete permissions on the MySQL database	admin
Password	Password of the user	-
Use Local API	Whether to use the local API of the database for acceleration. (The system attempts to enable the **local_infile** system variable of the MySQL database.)	Yes
Use Agent	Whether to extract data from the data source through an agent	Yes
local_infile Character Set	When using local_infile to import data to MySQL, you can configure the encoding format.	utf8
Driver Version	A driver version that adapts to MySQL	-
Agent	Click **Select** and select the agent created in **Connecting to an Agent**.	-
Fetch Size	Number of rows obtained by each request	1000
Commit Size	Obtaining data from the source through the agent	1000
Link Attributes	Custom attributes of the link	useCompression=true
Reference Sign	Delimiter used to separate referenced table names or column names This parameter is left blank by default.	'
Batch Size	Number of rows written each time. It should be less than Commit Size. When the number of rows written reaches the value of Commit Size, the rows will be committed to the database.	100

Step 6 Click **Save** to return to the **Links** page.

NOTE

If an error occurs during the saving, the security settings of the MySQL database are incorrect. In this case, you need to enable the EIP of the CDM cluster to access the MySQL database.

----End

6.3 Migrating Data from MySQL to MRS Hive

MRS provides enterprise-level big data clusters on the cloud. It contains HDFS, Hive, and Spark components and is applicable to massive data analysis of enterprises.

Hive supports SQL to help users perform extraction, transformation, and loading (ETL) operations on large-scale data sets. Query on large-scale data sets takes a long time. In many scenarios, you can create Hive partitions to reduce the total amount of data to be scanned each time. This significantly improves query performance.

Hive partitions are implemented by using the HDFS subdirectory function. Each subdirectory contains the column names and values of each partition. If there are multiple partitions, many HDFS subdirectories exist. It is not easy to load external data to each partition of the Hive table without relying on tools. With CDM, you can easily load data of the external data sources (relational databases, object storage services, and file system services) to Hive partition tables.

This section describes how to migrate data from the MySQL database to the MRS Hive partition table.

Scenario

Suppose that there is a **trip_data** table in the MySQL database. The table stores cycling records such as the start time, end time, start sites, end sites, and rider IDs. For details about the fields in the **trip_data** table, see **Figure 6-6**.

Figure 6-6 MySQL table fields

Column Name	#	Data Type
TripID	1	int(11)
Duration	2	int(11)
StartDate	3	timestamp
StartStation	4	varchar(64)
StartTerminal	5	int(11)
EndDate	6	timestamp
EndStation	7	varchar(64)
EndTerminal	8	int(11)
Bike	9	int(11)
SubscriberType	10	varchar(32)
ZipCodev	11	varchar(10)

The following describes how to use CDM to import the **trip_data** table in the MySQL database to the MRS Hive partition table. The procedure is as follows:

1. **Creating a Hive Partition Table on MRS Hive**
2. **Creating a CDM Cluster and Binding an EIP to the Cluster**
3. **Creating a MySQL Link**
4. **Creating a Hive Link**
5. **Creating a Migration Job**

Prerequisites

- MRS is available.
- You have obtained the IP address, port, database name, username, and password for connecting to the MySQL database. In addition, the user must have the read and write permissions on the MySQL database.
- You have uploaded a MySQL database driver by following the instructions provided in **Managing Drivers**.

Creating a Hive Partition Table on MRS Hive

On MRS Hive, run the following SQL statement to create a Hive partition table named **trip_data** with three new fields **y**, **ym**, and **ymd** used as partition fields. The SQL statement is as follows:

```
create table trip_data(TripID int,Duration int,StartDate timestamp,StartStation varchar(64),StartTerminal int,EndDate timestamp,EndStation varchar(64),EndTerminal int,Bike int,SubscriberType varchar(32),ZipCodev varchar(10))partitioned by (y int,ym int,ymd int);
```

NOTE

The **trip_data** partition table has three partition fields: year, year and month, and year, month, and date of the start time of a ride. For example, if the start time of a ride is **2018/5/11 9:40**, the record is saved in the **trip_data/2018/201805/20180511** partition. When the records in the **trip_data** table are summarized, only part of the data needs to be scanned, greatly improving the performance.

Creating a CDM Cluster and Binding an EIP to the Cluster

Step 1 If is an independent CDM service, create a CDM cluster by following the instructions provided in **Creating a Cluster**. If is used as a CDM component of DataArts Studio, create a CDM cluster by following the instructions provided in **Creating a Cluster**.

The key configurations are as follows:

- The flavor of the CDM cluster is selected based on the amount of data to be migrated. Generally, cdm.medium meets the requirements for most migration scenarios.
- The CDM and MRS clusters must be in the same VPC, subnet, and security group.

Step 2 After the CDM cluster is created, on the **Cluster Management** page, click **Bind EIP** in the **Operation** column to bind an EIP to the cluster. The CDM cluster uses the EIP to access MySQL.

Figure 6-7 Cluster list

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.

----End

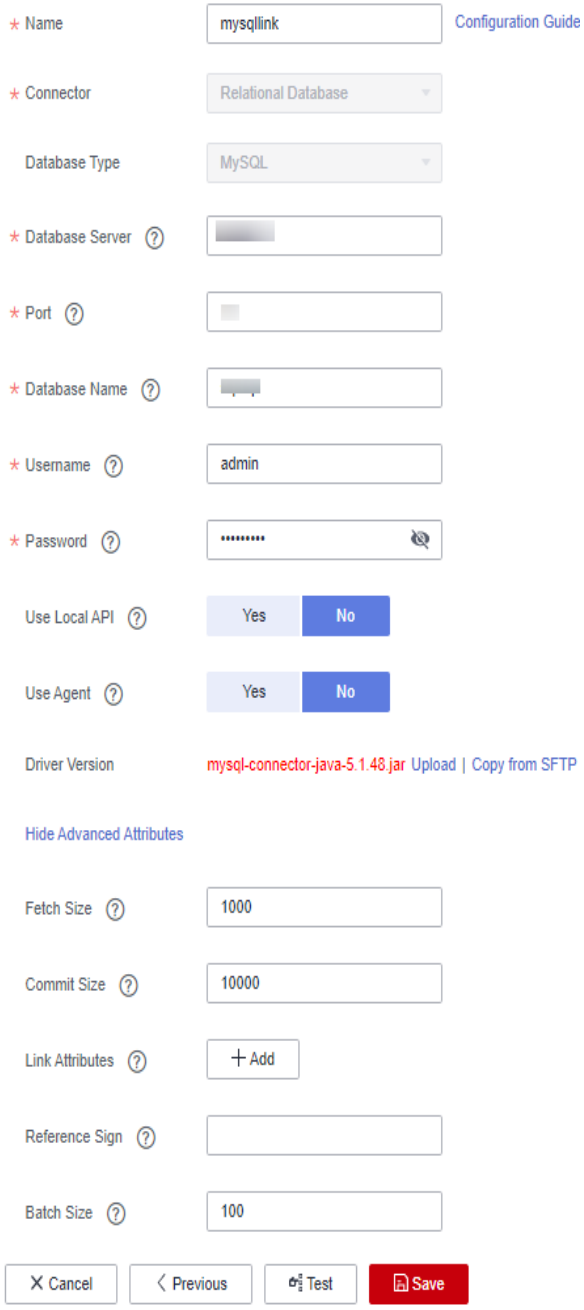
Creating a MySQL Link

Step 1 On the **Cluster Management** page, locate a cluster and click **Job Management** in the **Operation** column. On the displayed page, click the **Links** tab and then **Create Link**.

Figure 6-8 Selecting a connector

Step 2 Select **MySQL** and click **Next**. On the page that is displayed, configure MySQL link parameters.

Figure 6-9 Creating a MySQL link



Click **Show Advanced Attributes** and set optional parameters. For details, see **Link to Relational Databases**. Retain the default values of the optional parameters and configure the mandatory parameters according to **Table 6-3**.

Table 6-3 MySQL link parameters

Parameter	Description	Example Value
Name	Unique link name	mysqllink

Parameter	Description	Example Value
Database Server	IP address or domain name of the MySQL database server	192.168.1.110
Port	MySQL database port	3306
Database Name	Name of the MySQL database	sqoop
Username	User who has the read, write, and delete permissions on the MySQL database	admin
Password	Password of the user	-
Use Local API	Whether to use the local API of the database for acceleration. (The system attempts to enable the **local_infile** system variable of the MySQL database.)	Yes
Use Agent	Whether to extract data from the data source through an agent	Yes
local_infile Character Set	When using local_infile to import data to MySQL, you can configure the encoding format.	utf8
Driver Version	A driver version that adapts to MySQL	-
Agent	Click **Select** to select the agent created in **Connecting to an Agent**.	-
Fetch Size	(Optional) Displayed when you click **Show Advanced Attributes**. Number of rows obtained by each request. Set this parameter based on the data source and the job's data size. If the value is either too large or too small, the job may run for a long time.	1000
Commit Size	(Optional) Displayed when you click **Show Advanced Attributes**. Number of records submitted each time. Set this parameter based on the data destination and the job's data size. If the value is either too large or too small, the job may run for a long time.	1000
Link Attributes	Custom attributes of the link	useCompression=true

Parameter	Description	Example Value
Reference Sign	Delimiter used to separate referenced table names or column names This parameter is left blank by default.	'
Batch Size	Number of rows written each time. It should be less than Commit Size. When the number of rows written reaches the value of Commit Size, the rows will be committed to the database.	100

Step 3 Click **Save**. The **Link Management** page is displayed.

NOTE

If an error occurs during the saving, the security settings of the MySQL database are incorrect. In this case, you need to enable the EIP of the CDM cluster to access the MySQL database.

----End

Creating a Hive Link

Step 1 Click **Job Management** in the **Operation** column of the CDM cluster. On the displayed page, click the **Links** tab and then **Create Link**. The **Select Connector** page is displayed.

Figure 6-10 Selecting a connector type

Step 2 Select **MRS Hive** and click **Next** to configure parameters for the MRS Hive link.

Figure 6-11 Creating an MRS Hive link

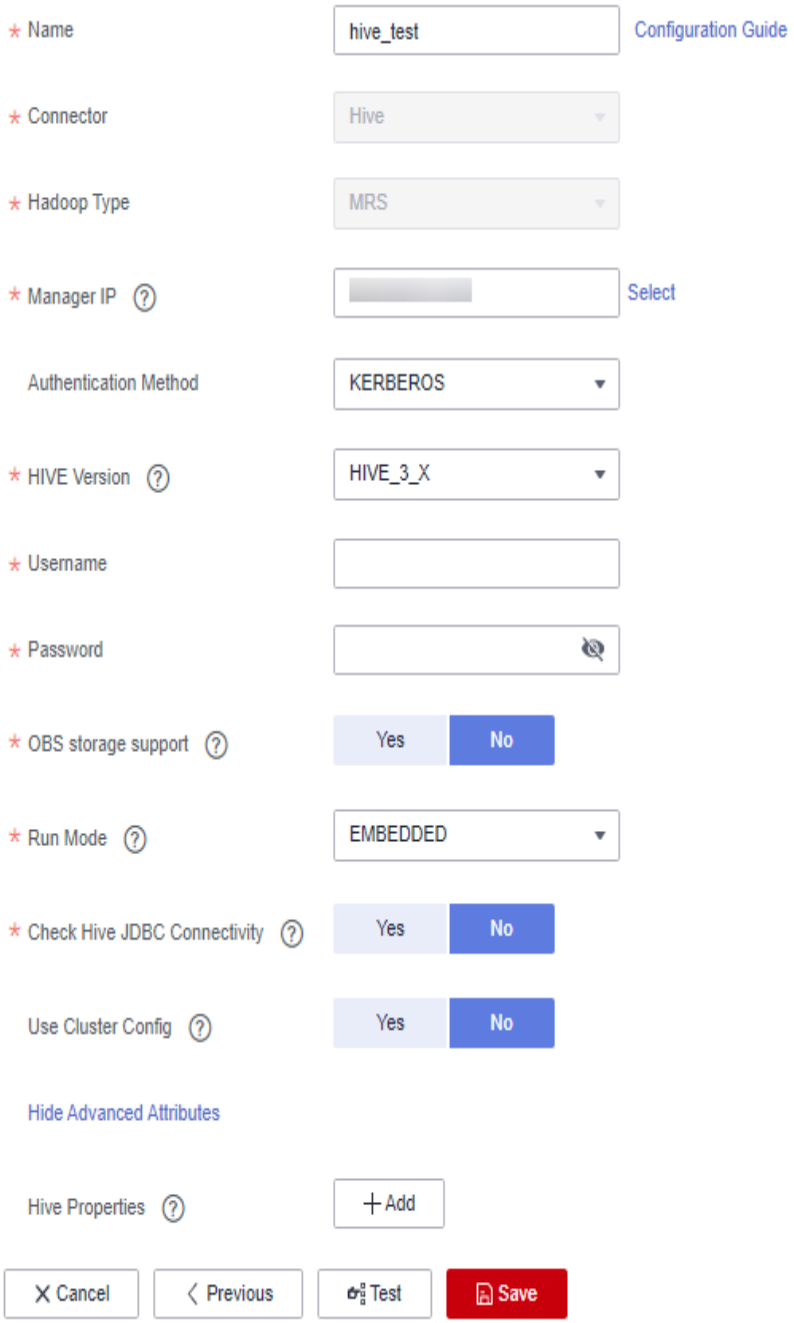


Table 6-4 describes the parameters. You can configure the parameters according to the actual situation.

Table 6-4 MRS Hive link parameters

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	hivelink
Manager IP	Floating IP address of MRS Manager. Click **Select** next to the **Manager IP** text box to select an MRS cluster. CDM automatically fills in the authentication information.	127.0.0.1
Authentication Method	Authentication method used for accessing MRS • **SIMPLE**: Select this for non-security mode. • **KERBEROS**: Select this for security mode.	SIMPLE
HIVE Version	Set this to the Hive version on the server.	HIVE_3_X
Username	If **Authentication Method** is set to **KERBEROS**, you must provide the username and password used for logging in to MRS Manager. If you need to create a snapshot when exporting a directory from HDFS, the user configured here must have the administrator permission on HDFS. 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.	cdm
Password	Password used for logging in to MRS Manager	-

Parameter	Description	Example Value
OBS storage support	The server must support OBS storage. When creating a Hive table, you can store the table in OBS.	No
Run Mode	This parameter is used only when the Hive version is **HIVE_3_X**. Possible values are: • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **Standalone**: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select **STANDALONE** or configure different agents. Note: The STANDALONE mode is used to solve the version conflict problem. If the connector versions of the source and destination ends of the same link are different, a JAR file conflict occurs. In this case, you need to place the source or destination end in the STANDALONE process to prevent the migration failure caused by the conflict.	EMBEDDED
Check Hive JDBC Connectivity	Whether to check the Hive JDBC connectivity	No
Use Cluster Config	You can use the cluster configuration to simplify parameter settings for the Hadoop connection.	No
Cluster Config Name	This parameter is valid only when **Use Cluster Config** is set to **Yes**. Select a cluster configuration that has been created. For details, see **Managing Cluster Configurations**.	hive_01

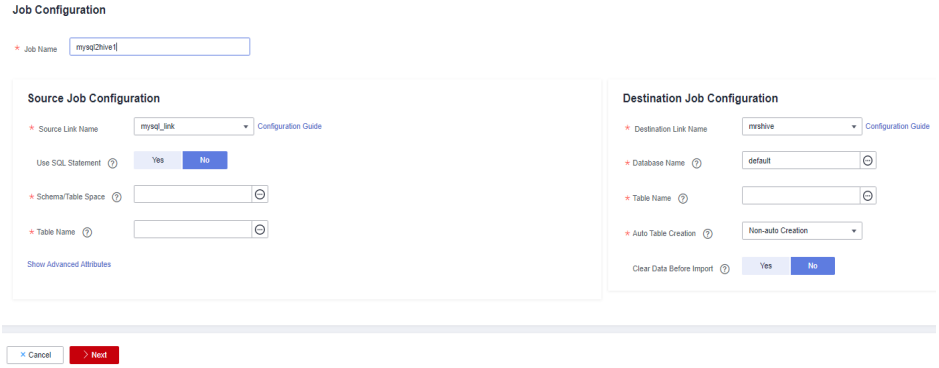
Step 3 Click **Save**. The **Link Management** page is displayed.

----End

Creating a Migration Job

Step 1 Choose **Table/File Migration** > **Create Job** to create a data migration job. **Figure 6-12** illustrates how to create a migration job.

Figure 6-12 Creating a job for migrating data from MySQL to Hive



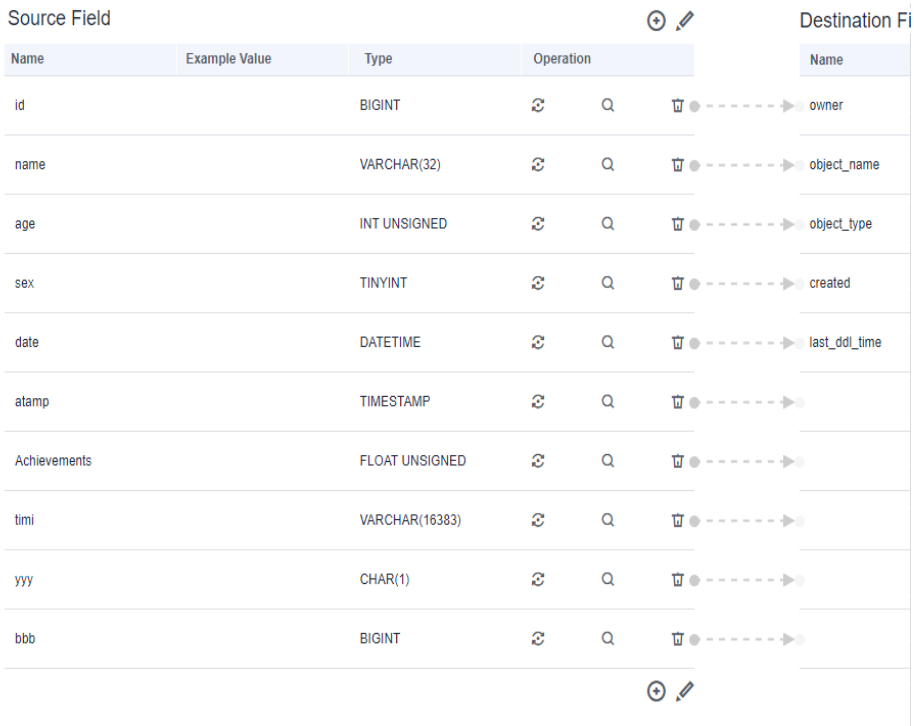
NOTE

Set **Clear Data Before Import** to **Yes**, so that the data in the Hive table will be cleared before data import.

Step 2 After the parameters are configured, click **Next**. The **Map Field** tab page is displayed. See **Figure 6-13**.

Map the fields of the MySQL table and Hive table. The Hive table has three more fields **y**, **ym**, and **ymd** than the MySQL table, which are the Hive partition fields. Because the fields of the source table cannot be directly mapped to the destination table, you need to configure an expression to extract data from the **StartDate** field in the source table.

Figure 6-13 Hive field mapping



Step 3 Click to display the **Converter List** dialog box, and then choose **Create Converter** > **Expression conversion**. See **Figure 6-14**.

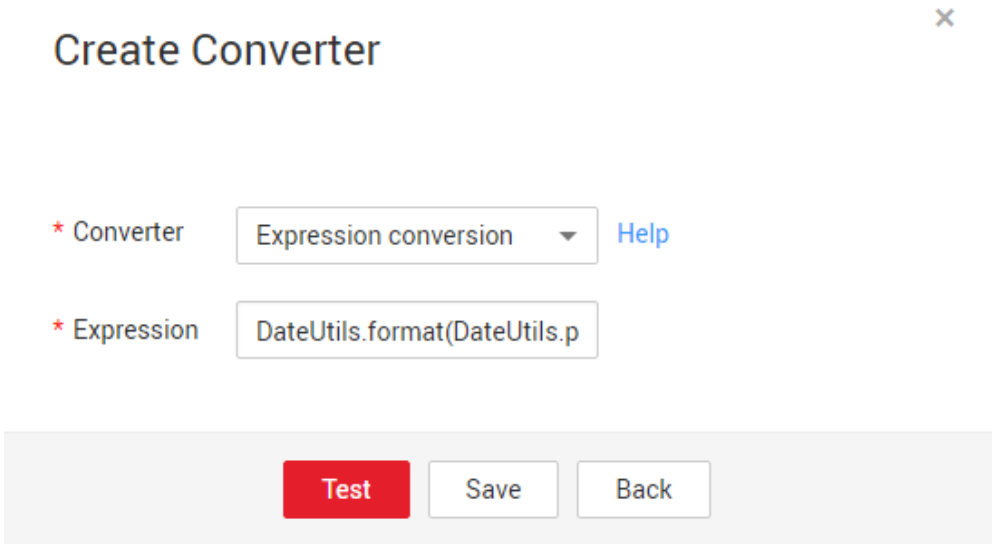
The expressions for the **y**, **ym**, and **ymd** fields are as follows:

DateUtils.format(DateUtils.parseDate(row[2],"yyyy-MM-dd HH:mm:ss.SSS"),"yyyy")

DateUtils.format(DateUtils.parseDate(row[2],"yyyy-MM-dd HH:mm:ss.SSS"),"yyyyMM")

DateUtils.format(DateUtils.parseDate(row[2],"yyyy-MM-dd HH:mm:ss.SSS"),"yyyyMMdd")

Figure 6-14 Configuring the expression



NOTE

The expressions in CDM support field conversion of common character strings, dates, and values.

Step 4 Click **Next** and set task parameters. Generally, retain the default values of all parameters.

In this step, you can configure the following optional functions:

- **Retry Upon Failure**: If the job fails to be executed, you can determine whether to automatically retry. Retain the default value **Never**.
- **Group**: Select the group to which the job belongs. The default group is **DEFAULT**. On the **Job Management** page, jobs can be displayed, started, or exported by group.
- **Schedule Execution**: To configure scheduled jobs, see **Scheduling Job Execution**. Retain the default value **No**.
- **Concurrent Extractors**: Enter the number of extractors to be concurrently executed. Retain the default value **1**.
- **Write Dirty Data**: Specify this parameter if data that fails to be processed or filtered out during job execution needs to be written to OBS for future viewing. Before writing dirty data, create an OBS link. Retain the default value **No** so that dirty data is not recorded.

- **Delete Job After Completion**: Retain the default value **Do not delete**.

Step 5 Click **Save and Run**. The **Job Management** page is displayed, on which you can view the job execution progress and result.

Step 6 After the job is successfully executed, in the **Operation** column of the job, click **Historical Record** to view the job's historical execution records and read/write statistics.

On the **Historical Record** page, click **Log** to view the job logs.

----End

6.4 Migrating Data from MySQL to OBS

Scenario

CDM supports table-to-OBS data migration. This section describes how to migrate tables from a MySQL database to OBS. The process is as follows:

1. **Creating a CDM Cluster and Binding an EIP to the Cluster**
2. **Creating a MySQL Link**
3. **Creating an OBS Link**
4. **Creating a Migration Job**

Prerequisites

- You have obtained the domain name, port number, AK, and SK for accessing OBS.
- You have obtained the IP address, port, database name, username, and password for connecting to the MySQL database. In addition, the user must have the read and write permissions on the MySQL database.
- You have uploaded a MySQL database driver by following the instructions provided in **Managing Drivers**.

Creating a CDM Cluster and Binding an EIP to the Cluster

Step 1 If is an independent CDM service, create a CDM cluster by following the instructions provided in **Creating a Cluster**. If is used as a CDM component of DataArts Studio, create a CDM cluster by following the instructions provided in **Creating a Cluster**.

The key configurations are as follows:

The flavor of the CDM cluster is selected based on the amount of data to be migrated. Generally, cdm.medium meets the requirements for most migration scenarios.

Step 2 After the CDM cluster is created, on the **Cluster Management** page, click **Bind EIP** in the **Operation** column to bind an EIP to the cluster. The CDM cluster uses the EIP to access MySQL.

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.

----End

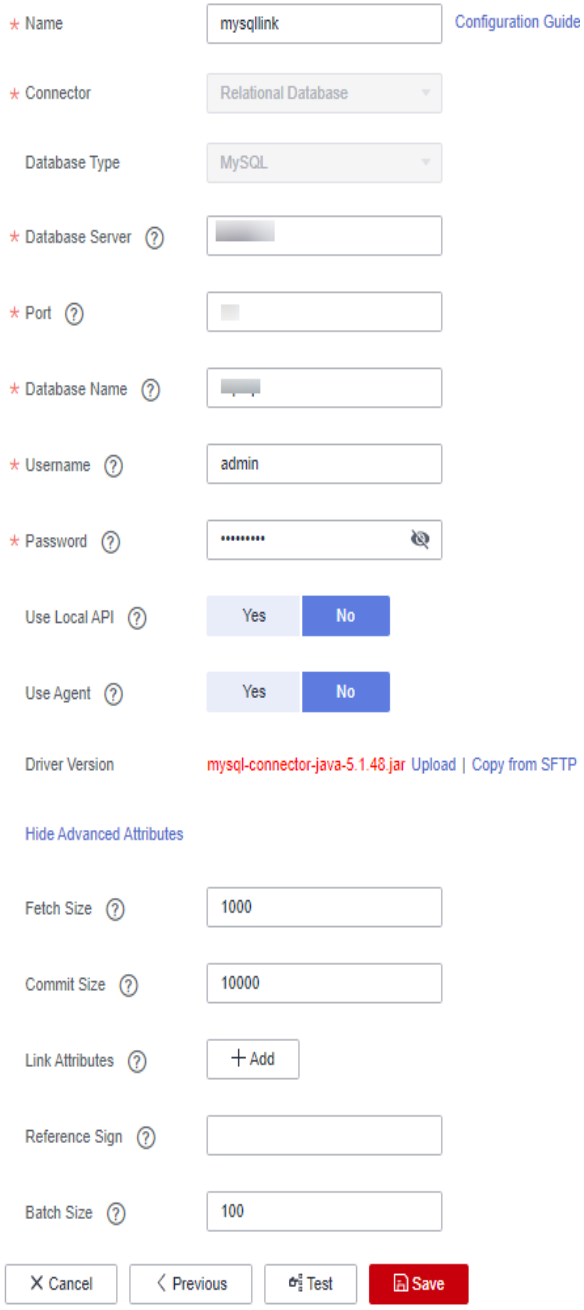
Creating a MySQL Link

Step 1 On the **Cluster Management** page, locate a cluster and click **Job Management** in the **Operation** column. On the displayed page, click the **Links** tab and then **Create Link**.

Figure 6-15 Selecting a connector

Step 2 Select **MySQL** and click **Next**. On the page that is displayed, configure MySQL link parameters.

Figure 6-16 Creating a MySQL link



Click **Show Advanced Attributes** and set optional parameters. For details, see **Link to Relational Databases**. Retain the default values of the optional parameters and configure the mandatory parameters according to **Table 6-5**.

Table 6-5 MySQL link parameters

Parameter	Description	Example Value
Name	Unique link name	mysqllink

Parameter	Description	Example Value
Database Server	IP address or domain name of the MySQL database server	192.168.1.110
Port	MySQL database port	3306
Database Name	Name of the MySQL database	sqoop
Username	User who has the read, write, and delete permissions on the MySQL database	admin
Password	Password of the user	-
Use Local API	Whether to use the local API of the database for acceleration. (The system attempts to enable the **local_infile** system variable of the MySQL database.)	Yes
Use Agent	Whether to extract data from the data source through an agent	Yes
local_infile Character Set	When using local_infile to import data to MySQL, you can configure the encoding format.	utf8
Driver Version	A driver version that adapts to MySQL	-
Agent	Click **Select** to select the agent created in **Connecting to an Agent**.	-
Fetch Size	(Optional) Displayed when you click **Show Advanced Attributes**. Number of rows obtained by each request. Set this parameter based on the data source and the job's data size. If the value is either too large or too small, the job may run for a long time.	1000
Commit Size	(Optional) Displayed when you click **Show Advanced Attributes**. Number of records submitted each time. Set this parameter based on the data destination and the job's data size. If the value is either too large or too small, the job may run for a long time.	1000
Link Attributes	Custom attributes of the link	useCompression=true

Parameter	Description	Example Value
Reference Sign	Delimiter used to separate referenced table names or column names This parameter is left blank by default.	'
Batch Size	Number of rows written each time. It should be less than Commit Size. When the number of rows written reaches the value of Commit Size, the rows will be committed to the database.	100

Step 3 Click **Save**. The **Link Management** page is displayed.

NOTE

If an error occurs during the saving, the security settings of the MySQL database are incorrect. In this case, you need to enable the EIP of the CDM cluster to access the MySQL database.

----End

Creating an OBS Link

Step 1 Click **Job Management** in the **Operation** column of the CDM cluster. On the displayed page, click the **Links** tab and then **Create Link**. The **Select Connector** page is displayed.

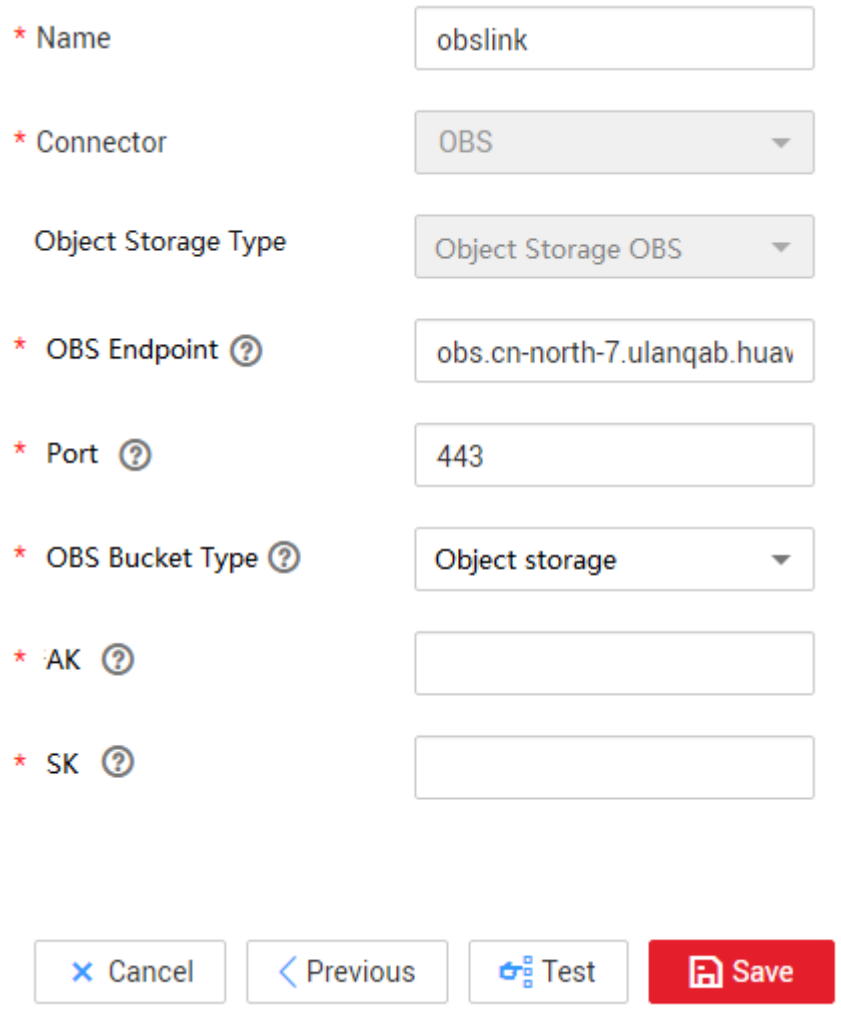
Figure 6-17 Selecting a connector type

Step 2 Select **Object Storage Service (OBS)** and click **Next** to configure parameters for the OBS link.

- **Name**: Enter a custom link name, for example, **obslink**.

- **OBS Server** and **Port**: Enter the actual OBS address information.
- **AK** and **SK**: Enter the AK and SK used for logging in to OBS.

Figure 6-18 Creating an OBS link



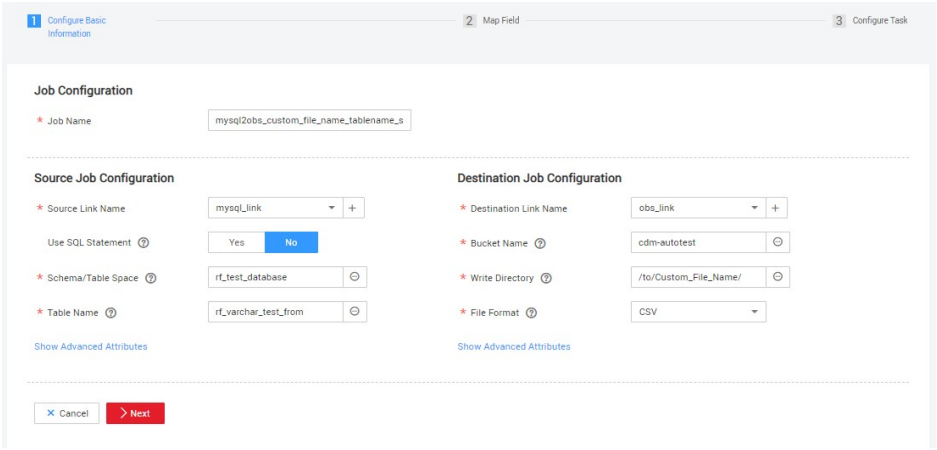
Step 3 Click **Save**. The **Link Management** page is displayed.

----End

Creating a Migration Job

Step 1 Choose **Table/File Migration** > **Create Job** to create a job for exporting data from the MySQL database to OBS.

Figure 6-19 Creating a job for migrating data from MySQL to OBS

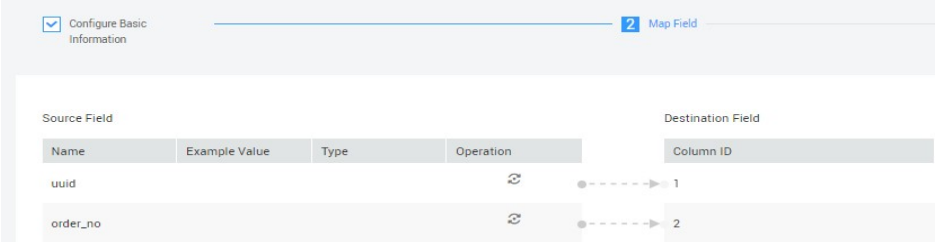


- **Job Name**: Enter a unique name.
- **Source Job Configuration**
  - **Source Link Name**: Select the **mysqllink** created in **Creating a MySQL Link**.
  - **Use SQL Statement**: Select **No**.
  - **Schema/Tablespace**: name of the schema or tablespace from which data is to be extracted
  - **Table Name**: name of the table from which data is to be extracted
  - Retain the default values of other optional parameters. For details, see **From a Common Relational Database**.
- **Destination Job Configuration**
  - **Destination Link Name**: Select the **obslink** created in **Creating an OBS Link**.
  - **Bucket Name**: Select the bucket from which the data will be migrated.
  - **Write Directory**: Enter the directory to which data is to be written on the OBS server.
  - **File Format**: Select **CSV**.
  - Retain the default values of the optional parameters in **Show Advanced Attributes**. For details, see **To OBS**.

Step 2 Click **Next**. The **Map Field** page is displayed. CDM automatically matches the source and destination fields, as shown in **Figure 6-20**.

- If the field mapping is incorrect, you can drag the fields to adjust the mapping.
- The expressions in CDM support field conversion of common character strings, dates, and values. For details, see **Converting Fields**.

Figure 6-20 Table-to-file field mapping



Step 3 Click **Next** and set task parameters. Generally, retain the default values of all parameters.

In this step, you can configure the following optional functions:

- **Retry Upon Failure**: If the job fails to be executed, you can determine whether to automatically retry. Retain the default value **Never**.
- **Group**: Select the group to which the job belongs. The default group is **DEFAULT**. On the **Job Management** page, jobs can be displayed, started, or exported by group.
- **Schedule Execution**: To configure scheduled jobs, see **Scheduling Job Execution**. Retain the default value **No**.
- **Concurrent Extractors**: Enter the number of extractors to be concurrently executed. CDM supports concurrent extraction of MySQL data. If indexes are configured for the source table, you can increase the number of concurrent extractors to accelerate the migration.
- **Write Dirty Data**: Specify this parameter if data that fails to be processed or filtered out during job execution needs to be written to OBS for future viewing. Before writing dirty data, create an OBS link. For file-to-table data migration, you are advised to write dirty data.
- **Delete Job After Completion**: Retain the default value **Do not delete**. You can also set this parameter to **Delete** to prevent an accumulation of too many migration jobs.

Step 4 Click **Save and Run**. The **Job Management** page is displayed, on which you can view the job execution progress and result.

Step 5 After the job is successfully executed, in the **Operation** column of the job, click **Historical Record** to view the job's historical execution records and read/write statistics.

On the **Historical Record** page, click **Log** to view the job logs.

----End

6.5 Migrating Data from MySQL to DWS

Scenario

CDM supports table-to-table data migration. This section describes how to migrate data from MySQL to DWS. The process is as follows:

1. **Creating a CDM Cluster and Binding an EIP to the Cluster**

2. **Creating a MySQL Link**
3. **Creating a DWS Link**
4. **Creating a Migration Job**

Prerequisites

- You have obtained the IP address, port number, database name, username, and password for connecting to DWS. In addition, you must have the read, write, and delete permissions on the DWS database.
- You have obtained the IP address, port, database name, username, and password for connecting to the MySQL database. In addition, the user must have the read and write permissions on the MySQL database.
- You have uploaded a MySQL database driver by following the instructions provided in **Managing Drivers**.

Creating a CDM Cluster and Binding an EIP to the Cluster

Step 1 If is an independent CDM service, create a CDM cluster by following the instructions provided in **Creating a Cluster**. If is used as a CDM component of DataArts Studio, create a CDM cluster by following the instructions provided in **Creating a Cluster**.

The key configurations are as follows:

- The flavor of the CDM cluster is selected based on the amount of data to be migrated. Generally, cdm.medium meets the requirements for most migration scenarios.
- The VPC, subnet, and security group of the CDM cluster must be the same as those of the DWS cluster.

Step 2 After the CDM cluster is created, on the **Cluster Management** page, click **Bind EIP** in the **Operation** column to bind an EIP to the cluster. The CDM cluster uses the EIP to access MySQL.

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.

----End

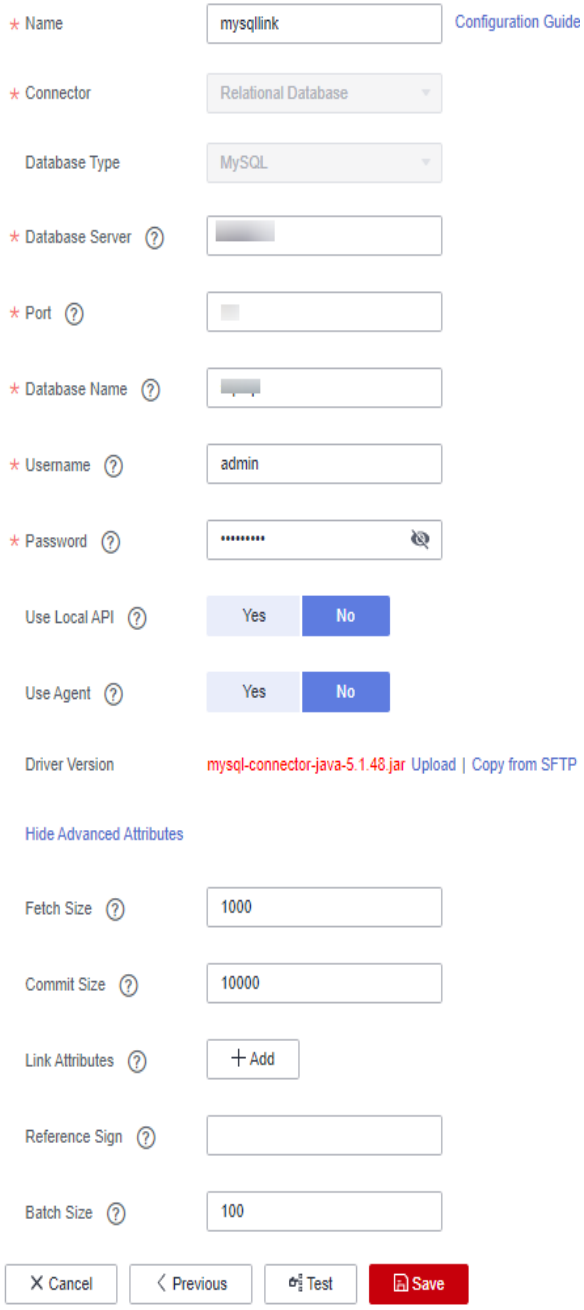
Creating a MySQL Link

Step 1 On the **Cluster Management** page, locate a cluster and click **Job Management** in the **Operation** column. On the displayed page, click the **Links** tab and then **Create Link**.

Figure 6-21 Selecting a connector

Step 2 Select **MySQL** and click **Next**. On the page that is displayed, configure MySQL link parameters.

Figure 6-22 Creating a MySQL link



Click **Show Advanced Attributes** and set optional parameters. For details, see **Link to Relational Databases**. Retain the default values of the optional parameters and configure the mandatory parameters according to **Table 6-6**.

Table 6-6 MySQL link parameters

Parameter	Description	Example Value
Name	Unique link name	mysqllink

Parameter	Description	Example Value
Database Server	IP address or domain name of the MySQL database server	192.168.1.110
Port	MySQL database port	3306
Database Name	Name of the MySQL database	sqoop
Username	User who has the read, write, and delete permissions on the MySQL database	admin
Password	Password of the user	-
Use Local API	Whether to use the local API of the database for acceleration. (The system attempts to enable the **local_infile** system variable of the MySQL database.)	Yes
Use Agent	Whether to extract data from the data source through an agent	Yes
local_infile Character Set	When using local_infile to import data to MySQL, you can configure the encoding format.	utf8
Driver Version	A driver version that adapts to MySQL	-
Agent	Click **Select** to select the agent created in **Connecting to an Agent**.	-
Fetch Size	(Optional) Displayed when you click **Show Advanced Attributes**. Number of rows obtained by each request. Set this parameter based on the data source and the job's data size. If the value is either too large or too small, the job may run for a long time.	1000
Commit Size	(Optional) Displayed when you click **Show Advanced Attributes**. Number of records submitted each time. Set this parameter based on the data destination and the job's data size. If the value is either too large or too small, the job may run for a long time.	1000
Link Attributes	Custom attributes of the link	useCompression=true

Parameter	Description	Example Value
Reference Sign	Delimiter used to separate referenced table names or column names This parameter is left blank by default.	'
Batch Size	Number of rows written each time. It should be less than Commit Size. When the number of rows written reaches the value of Commit Size, the rows will be committed to the database.	100

Step 3 Click **Save**. The **Link Management** page is displayed.

NOTE

If an error occurs during the saving, the security settings of the MySQL database are incorrect. In this case, you need to enable the EIP of the CDM cluster to access the MySQL database.

----End

Creating a DWS Link

Step 1 On the **Cluster Management** page, locate a cluster and click **Job Management** in the **Operation** column. On the displayed page, click the **Links** tab and then **Create Link**.

Figure 6-23 Selecting a connector type

Step 2 Select **Data Warehouse Service** and click **Next** to configure the DWS link parameters. Set the mandatory parameters listed in **Table 6-7** and retain the default values for the optional parameters.

Table 6-7 DWS link parameters

Parameter	Description	Example Value
Name	Enter a unique link name.	dwslink
Database Server	IP address or domain name of the DWS database	192.168.0.3
Port	DWS database port	8000
Database Name	Name of the DWS database	db_demo
Username	User who has the read, write, and delete permissions on the DWS database	dbadmin
Password	Password of the user	-
Use Agent	Whether to extract data from the data source through an agent	Yes
Agent	Click **Select** and select the agent created in **Connecting to an Agent**.	-
Import Mode	**COPY**: Migrate the source data to the DWS management node and then copy the data to DataNodes. To access DWS through the Internet, select **COPY**.	COPY

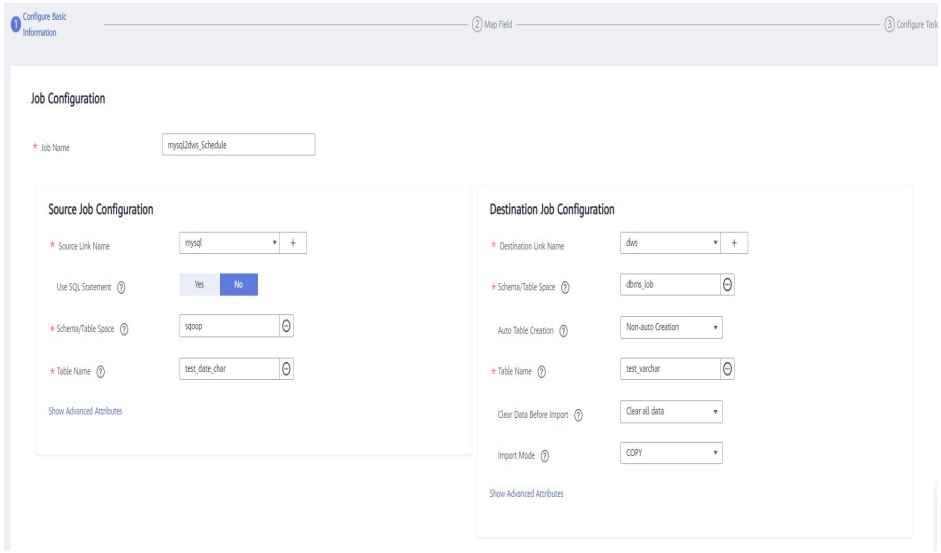
Step 3 Click **Save**.

----End

Creating a Migration Job

Step 1 Choose **Table/File Migration** > **Create Job** to create a job for exporting data from the MySQL database to DWS.

Figure 6-24 Creating a job for migrating data from MySQL to DWS



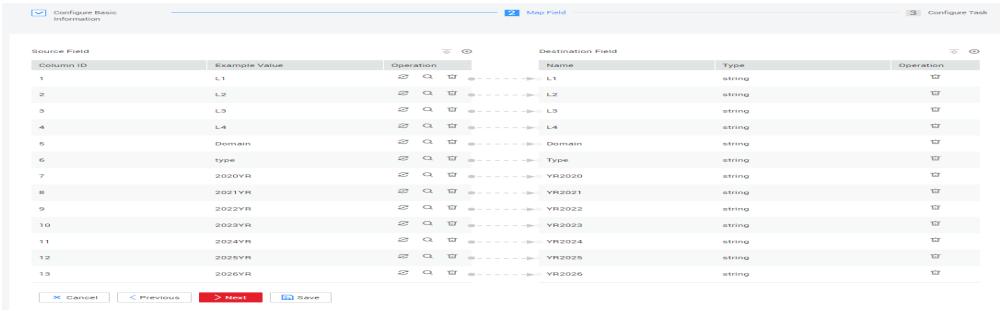
- **Job Name**: Enter a unique name.
- **Source Job Configuration**
  - **Source Link Name**: Select the **mysqllink** created in **Creating a MySQL Link**.
  - **Use SQL Statement**: Select **No**.
  - **Schema/Tablespace**: name of the schema or tablespace from which data is to be extracted
  - **Table Name**: name of the table from which data is to be extracted
  - Retain the default values of other optional parameters. For details, see **From a Common Relational Database**.
- **Destination Job Configuration**
  - **Destination Link Name**: Select the **dwslink** created in **Creating a DWS Link**.
  - **Schema/Tablespace**: Select the DWS database to which data is to be written.
  - **Auto Table Creation**: This parameter is displayed only when both the migration source and destination are relational databases.
  - **Table Name**: Name of the table to which data is to be written. You can enter a table name that does not exist. CDM automatically creates the table in DWS.
  - **isCompress**: whether to compress data. If you select **Yes**, high-level compression will be performed. CDM applies to compression scenarios where the I/O read/write volume is large and the CPU is sufficient (the computing load is relatively low). For more compression levels, see **Compression Levels**.
  - **Orientation**: You can create row- or column-store tables as needed. Generally, if a table contains many columns (called a wide table) and its query involves only a few columns, column storage is recommended. If a table contains only a few columns and a query includes most of the fields, row storage is recommended.

  - **Extend char length**: If the data encoding formats of the migration source and destination are different, the character length of the automatic table creation may be insufficient. If you select **Yes** for this parameter, the character length will be increased by three times during automatic table creation.
  - **Clear Data Before Import**: whether to clear data in the destination table before the migration task starts.

Step 2 Click **Next**. The **Map Field** page is displayed. CDM automatically matches the source and destination fields, as shown in **Figure 6-25**.

- If the field mapping is incorrect, you can drag the fields to adjust the mapping.
- You can map fields in batches.
- The expressions in CDM support field conversion of common character strings, dates, and values. For details, see **Converting Fields**.

Figure 6-25 Table-to-table field mapping



Step 3 Click **Next** and set task parameters. Generally, retain the default values of all parameters.

In this step, you can configure the following optional functions:

- **Retry Upon Failure**: If the job fails to be executed, you can determine whether to automatically retry. Retain the default value **Never**.
- **Group**: Select the group to which the job belongs. The default group is **DEFAULT**. On the **Job Management** page, jobs can be displayed, started, or exported by group.
- **Schedule Execution**: To configure scheduled jobs, see **Scheduling Job Execution**. Retain the default value **No**.
- **Concurrent Extractors**: Enter the number of extractors to be concurrently executed. You can increase the value of this parameter to improve migration efficiency.
- **Write Dirty Data**: Dirty data may be generated during data migration between tables. You are advised to select **Yes**.
- **Delete Job After Completion**: Retain the default value **Do not delete**.

Step 4 Click **Save and Run**. The **Job Management** page is displayed, on which you can view the job execution progress and result.

Step 5 After the job is successfully executed, in the **Operation** column of the job, click **Historical Record** to view the job's historical execution records and read/write statistics.

On the **Historical Record** page, click **Log** to view the job logs.

----End

6.6 Migrating an Entire MySQL Database to RDS

Scenario

This section describes how to migrate the entire on-premises MySQL database to RDS using the CDM's entire DB migration function.

Currently, CDM can migrate the entire on-premises MySQL database to RDS for MySQL, RDS for PostgreSQL, or RDS for SQL Server. The following describes how to migrate the entire database to RDS. The procedure is as follows:

1. **Creating a CDM Cluster and Binding an EIP to the Cluster**
2. **Creating a MySQL Link**
3. **Creating an RDS Link**
4. **Creating an Entire DB Migration Job**

Prerequisites

- You have sufficient EIP quota.
- You have obtained an RDS database instance and the database engine of this instance is MySQL.
- The on-premises MySQL database can be accessed through the public network. If the MySQL database is deployed on an on-premises data center or a third-party cloud, ensure that an IP address that can be accessed from the public network has been configured for the MySQL database, or the VPN or Direct Connect between the on-premises data center and the cloud service platform has been established.
- You have obtained the IP addresses, names, usernames, and passwords of the on-premises MySQL database and RDS for MySQL.
- You have uploaded a MySQL database driver by following the instructions provided in **Managing Drivers**.

Creating a CDM Cluster and Binding an EIP to the Cluster

Step 1 If is an independent CDM service, create a CDM cluster by following the instructions provided in **Creating a Cluster**. If is used as a CDM component of DataArts Studio, create a CDM cluster by following the instructions provided in **Creating a Cluster**.

The key configurations are as follows:

- The flavor of the CDM cluster is selected based on the amount of data to be migrated. Generally, cdm.medium meets the requirements for most migration scenarios.
- The CDM cluster and the RDS for MySQL instance must be in the same VPC. In addition, it is recommended that the CDM cluster be in the same subnet and security group as the RDS for MySQL instance.

- If the same subnet and security group cannot be used for security purposes, ensure that a security group rule has been configured to allow the CDM cluster to access the RDS for MySQL instance.

Step 2 After the CDM cluster is created, on the **Cluster Management** page, click **Bind EIP** in the **Operation** column to bind an EIP to the cluster. The CDM cluster uses the EIP to access the on-premises MySQL database.

Figure 6-26 Cluster list

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.

----End

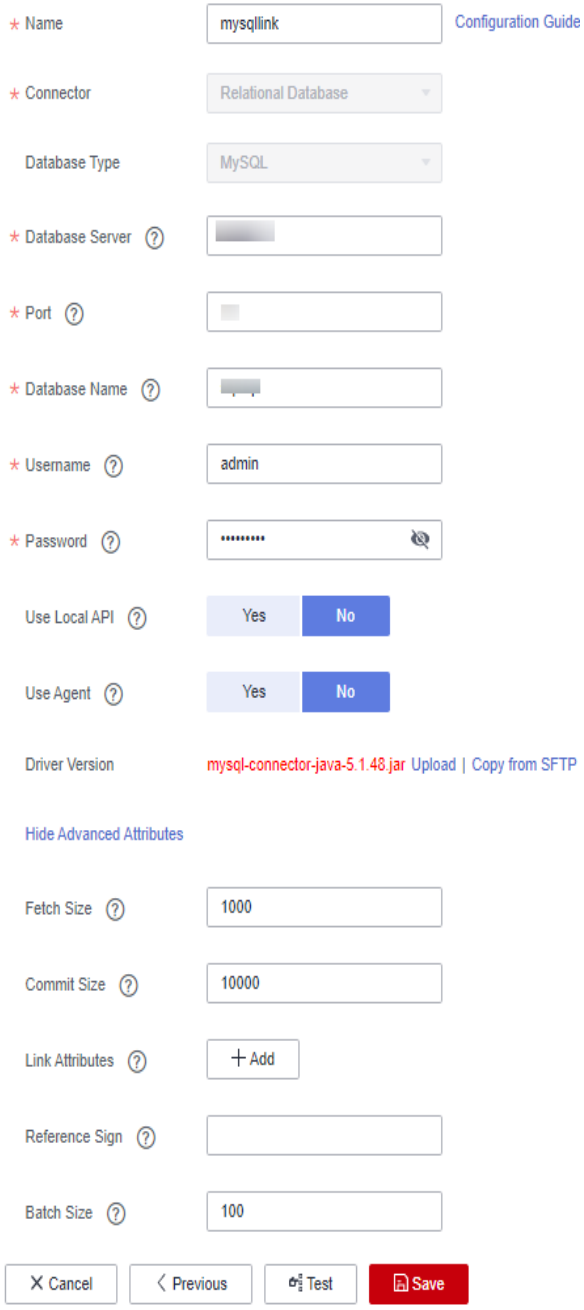
Creating a MySQL Link

Step 1 On the **Cluster Management** page, locate a cluster and click **Job Management** in the **Operation** column. On the displayed page, click the **Links** tab and then **Create Link**.

Figure 6-27 Selecting a connector

Step 2 Select **MySQL** and click **Next**. On the page that is displayed, configure MySQL link parameters.

Figure 6-28 Creating a MySQL link



Click **Show Advanced Attributes** and set optional parameters. For details, see **Link to Relational Databases**. Retain the default values of the optional parameters and configure the mandatory parameters according to **Table 6-8**.

Table 6-8 MySQL link parameters

Parameter	Description	Example Value
Name	Unique link name	mysqllink

Parameter	Description	Example Value
Database Server	IP address or domain name of the MySQL database server	192.168.1.110
Port	MySQL database port	3306
Database Name	Name of the MySQL database	sqoop
Username	User who has the read, write, and delete permissions on the MySQL database	admin
Password	Password of the user	-
Use Local API	Whether to use the local API of the database for acceleration. (The system attempts to enable the **local_infile** system variable of the MySQL database.)	Yes
Use Agent	Whether to extract data from the data source through an agent	Yes
local_infile Character Set	When using local_infile to import data to MySQL, you can configure the encoding format.	utf8
Driver Version	A driver version that adapts to MySQL	-
Agent	Click **Select** to select the agent created in **Connecting to an Agent**.	-
Fetch Size	(Optional) Displayed when you click **Show Advanced Attributes**. Number of rows obtained by each request. Set this parameter based on the data source and the job's data size. If the value is either too large or too small, the job may run for a long time.	1000
Commit Size	(Optional) Displayed when you click **Show Advanced Attributes**. Number of records submitted each time. Set this parameter based on the data destination and the job's data size. If the value is either too large or too small, the job may run for a long time.	1000
Link Attributes	Custom attributes of the link	useCompression=true

Parameter	Description	Example Value
Reference Sign	Delimiter used to separate referenced table names or column names This parameter is left blank by default.	'
Batch Size	Number of rows written each time. It should be less than Commit Size. When the number of rows written reaches the value of Commit Size, the rows will be committed to the database.	100

Step 3 Click **Save**. The **Link Management** page is displayed.

NOTE

If an error occurs during the saving, the security settings of the MySQL database are incorrect. In this case, you need to enable the EIP of the CDM cluster to access the MySQL database.

----End

Creating an RDS Link

Step 1 Click **Job Management** in the **Operation** column of the CDM cluster. On the displayed page, click the **Links** tab and then **Create Link**. The **Select Connector** page is displayed.

Figure 6-29 Selecting a connector type

Step 2 Select **RDS for MySQL** and click **Next** to configure parameters for the RDS for MySQL link.

- **Name**: Enter a custom link name, for example, **rds_link**.
- **Database Server** and **Port**: Enter the address information about the RDS for MySQL database.
- **Database Name**: Enter the name of the RDS for MySQL database.

- **Username** and **Password**: Enter the username and password used for logging in to the database.

  NOTE

  - During RDS link creation, if **Use Local API** in **Show Advanced Attributes** is set to **Yes**, you can use the LOAD DATA function provided by MySQL to speed up data import.
  - The LOAD DATA function is disabled by default on RDS for MySQL, so you need to modify the parameter group of the MySQL instance and set **local_infile** to **ON** to enable this function.
  - If the **local_infile** parameter group cannot be edited, it is the default parameter group. You need to create a parameter group and modify its value, and apply it to the MySQL instance of RDS.

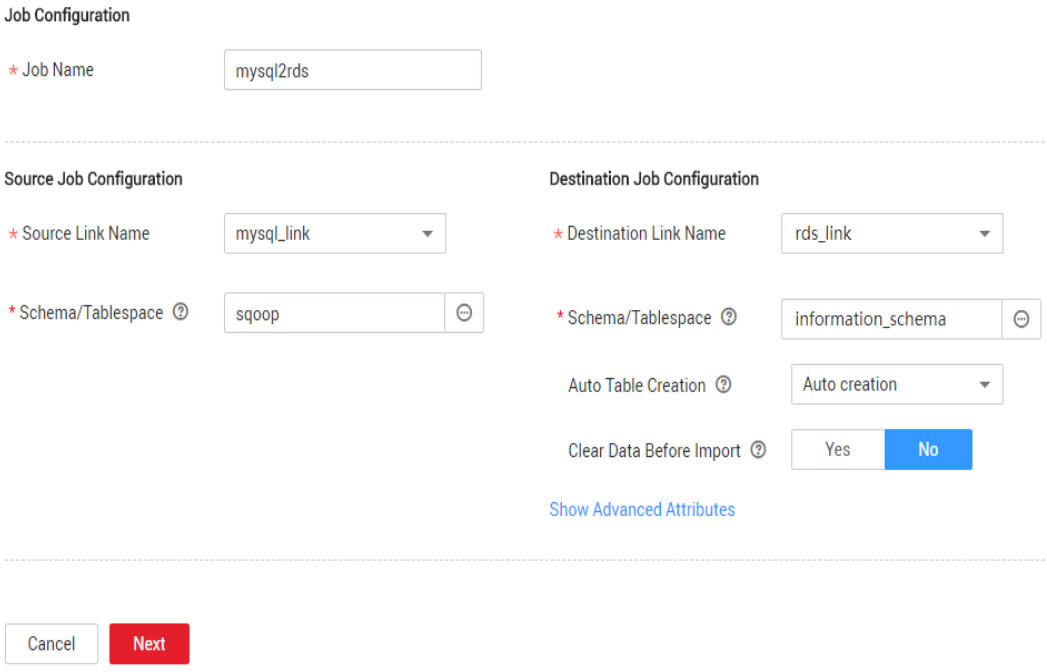
Step 3 Click **Save**. The **Link Management** page is displayed.

----End

Creating an Entire DB Migration Job

Step 1 After the two links are created, choose **Entire DB Migration** > **Create Job** to create a migration job. See **Figure 6-30**.

Figure 6-30 Creating an entire DB migration job



- **Job Name**: Enter a name for the entire DB migration job.
- **Source Job Configuration**
  - **Source Link Name**: Select the **mysql_link** link created in **Creating a MySQL Link**.
  - **Schema/Tablespace**: Select the on-premises MySQL database from which data is to be exported.
- **Destination Job Configuration**

- **Destination Link Name**: Select the **rds_link** link created in **Creating an RDS Link**.
- **Schema/Tablespace**: Select the name of the RDS database to which data is to be imported.
- **Auto Table Creation**: Select **Auto creation**, which indicates that CDM automatically creates tables in the RDS database when tables of the on-premises MySQL database do not exist in the RDS database.
- **Clear Data Before Import**: Select **Yes**, which indicates that when a table with the same name as the table in the on-premises MySQL database exists in the RDS database, CDM clears data in the table on RDS.
- Retain the default values of the optional parameters in **Show Advanced Attributes**.

Step 2 Click **Next**. The page for selecting tables to be migrated is displayed. You can select all or part of tables to migrate.

Step 3 Click **Save and Run** and CDM immediately starts the entire DB migration job.

When the job starts running, a sub-job will be generated for each table. You can click the job name to view the sub-job list.

Step 4 In the **Operation** column of the job, click **Historical Record** to view the job's historical execution records and read/write statistics.

There are no logs for the entire DB migration job. However, the sub-jobs have logs. On the **Historical Record** page of the sub-jobs, click **Log** to view the job logs.

----End

6.7 Migrating Data from Oracle to CSS

Scenario

Cloud Search Service provides users with structured and unstructured data search, statistics, and report capabilities. This section describes how to use CDM to migrate data from the Oracle database to Cloud Search Service. The procedure is as follows:

1. **Creating a CDM Cluster and Binding an EIP to the Cluster**
2. **Creating a Cloud Search Service Link**
3. **Creating an Oracle Link**
4. **Creating a Migration Job**

Prerequisites

- You have subscribed to Cloud Search Service and obtained the IP address and port number of the Cloud Search Service cluster.
- You have obtained the IP address, name, username, and password of the Oracle database.
- If the Oracle database is deployed on an on-premises data center or a third-party cloud, ensure that an IP address that can be accessed from the public

network has been configured for the Oracle database, or the VPN or Direct Connect between the on-premises data center and has been established.

- You have uploaded an Oracle database driver by following the instructions provided in **Managing Drivers**.

Creating a CDM Cluster and Binding an EIP to the Cluster

Step 1 If is an independent CDM service, create a CDM cluster by following the instructions provided in **Creating a Cluster**. If is used as a CDM component of DataArts Studio, create a CDM cluster by following the instructions provided in **Creating a Cluster**.

The key configurations are as follows:

- The flavor of the CDM cluster is selected based on the amount of data to be migrated. Generally, cdm.medium meets the requirements for most migration scenarios.
- The CDM and Cloud Search Service clusters must be in the same VPC. In addition, it is recommended that the CDM cluster be in the same subnet and security group as the Cloud Search Service cluster.
- If the same subnet and security group cannot be used for security purposes, ensure that a security group rule has been configured to allow the CDM cluster to access the Cloud Search Service cluster.

Step 2 After the CDM cluster is created, on the **Cluster Management** page, click **Bind EIP** in the **Operation** column to bind an EIP to the cluster. The CDM cluster uses the EIP to access the Oracle data source.

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.

----End

Creating a Cloud Search Service Link

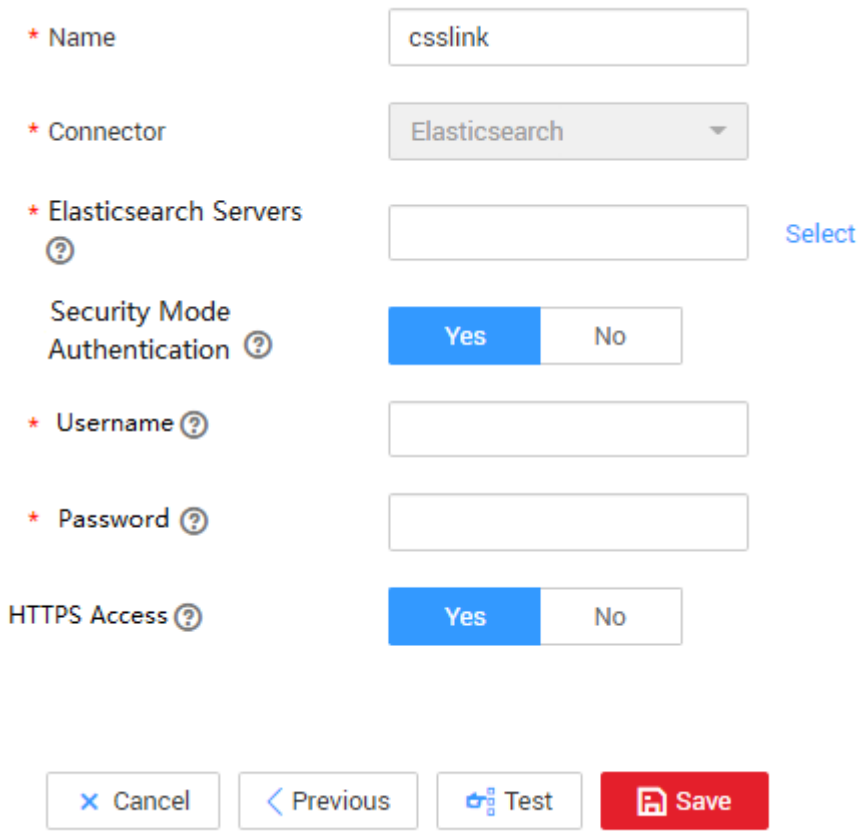
Step 1 Click **Job Management** in the **Operation** column of the CDM cluster. On the displayed page, click the **Links** tab and then **Create Link**. The **Select Connector** page is displayed.

Figure 6-31 Selecting a connector

Step 2 Select **Cloud Search Service** and click **Next**. On the page that is displayed, configure the CSS link parameters.

- **Name**: Enter a custom link name, for example, **csslink**.
- **Elasticsearch Server List**: Enter the IP address and port number of the Cloud Search Service cluster (cluster later than 5.*x*). The format is *ip:port*. Use semicolons to separate multiple addresses. For example, **192.168.0.1:9200;192.168.0.2:9200**.
- **Username** and **Password**: Enter the username and password used for logging in to the Cloud Search Service cluster. The user must have the read and write permissions on the database.

Figure 6-32 Creating a CSS link



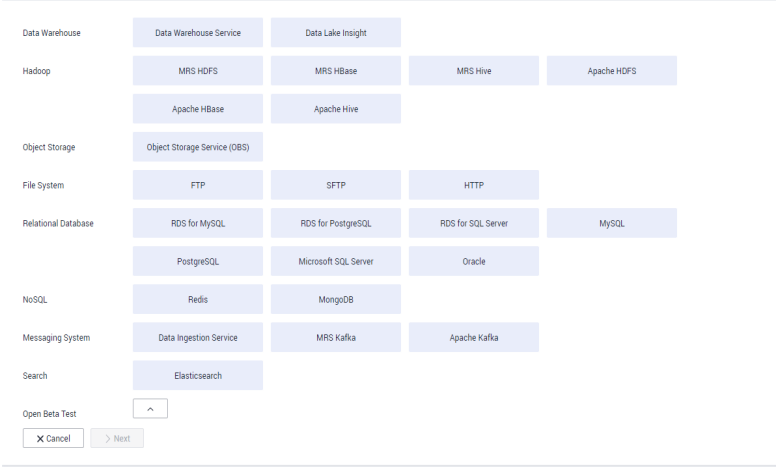
Step 3 Click **Save**. The **Link Management** page is displayed.

----End

Creating an Oracle Link

Step 1 Click **Job Management** in the **Operation** column of the CDM cluster. On the displayed page, click the **Links** tab and then **Create Link**. The **Select Connector** page is displayed.

Figure 6-33 Selecting a connector type



Step 2 Select **Oracle** and click **Next** to configure parameters for the Oracle link.

- **Name**: Enter a custom link name, for example, **oracle_link**.
- **Database Server** and **Port**: Enter the address and port number of the Oracle server.
- **Database Name**: Enter the name of the Oracle database whose data is to be exported.
- **Username** and **Password**: Enter the username and password used for logging in to the Oracle database. The user must have the permission to read the Oracle metadata.

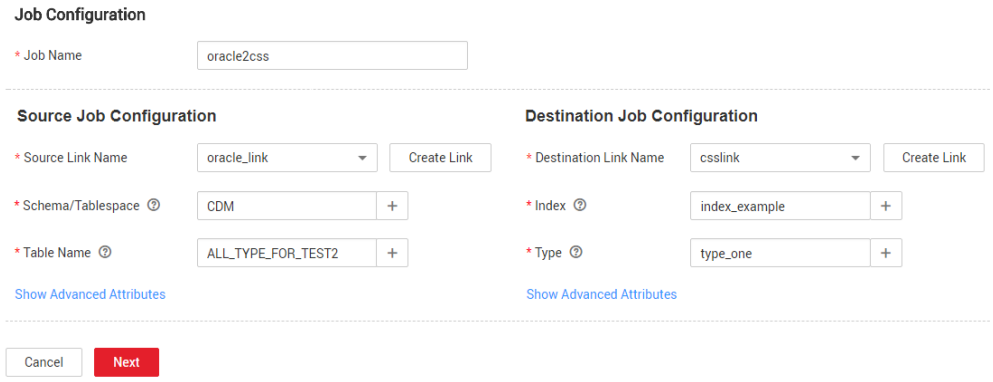
Step 3 Click **Save**. The **Link Management** page is displayed.

----End

Creating a Migration Job

Step 1 Choose **Table/File Migration** > **Create Job** to create a job for exporting data from the Oracle database to Cloud Search Service.

Figure 6-34 Creating a job for migrating data from Oracle to Cloud Search Service



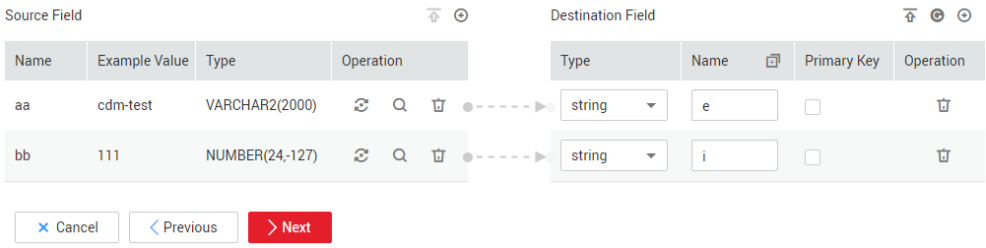
- **Job Name**: Enter a unique name.
- **Source Job Configuration**

  - **Source Link Name**: Select the **oracle_link** link created in **Creating an Oracle Link**.
  - **Schema/Tablespace**: Enter the name of the database whose data is to be migrated.
  - **Table Name**: Enter the name of the table to be migrated.
  - Retain the default values of the optional parameters in **Show Advanced Attributes**. For details, see **From a Common Relational Database**.
- **Destination Job Configuration**
  - **Destination Link Name**: Select the **csslink** link created in **Creating a Cloud Search Service Link**.
  - **Index**: Select the Elasticsearch index of the data to be written. You can also enter a new index. CDM automatically creates the index on Cloud Search Service.
  - **Type**: Select the Elasticsearch type of the data to be written. You can enter a new type. CDM automatically creates a type at the migration destination.
  - Retain the default values of the optional parameters in **Show Advanced Attributes**. For details, see **To CSS**.

Step 2 Click **Next**. The **Map Field** page is displayed. CDM automatically matches the source and destination fields. See **Figure 6-35**.

- If the field mapping is incorrect, you can drag the fields to adjust the mapping.
- If the type is automatically created at the migration destination, you need to configure the type and name of each field.
- CDM supports field conversion during the migration. For details, see **Converting Fields**.

Figure 6-35 Field mapping of Cloud Search Service



Step 3 Click **Next** and set task parameters. Generally, retain the default values of all parameters.

In this step, you can configure the following optional functions:

- **Retry Upon Failure**: If the job fails to be executed, you can determine whether to automatically retry. Retain the default value **Never**.
- **Group**: Select the group to which the job belongs. The default group is **DEFAULT**. On the **Job Management** page, jobs can be displayed, started, or exported by group.
- **Schedule Execution**: To configure scheduled jobs, see **Scheduling Job Execution**. Retain the default value **No**.

- **Concurrent Extractors**: Enter the number of extractors to be concurrently executed. Retain the default value **1**.
- **Write Dirty Data**: Specify this parameter if data that fails to be processed or filtered out during job execution needs to be written to OBS for future viewing. Before writing dirty data, create an OBS link. Retain the default value **No** so that dirty data is not recorded.
- **Delete Job After Completion**: Retain the default value **Do not delete**.

Step 4 Click **Save and Run**. The **Job Management** page is displayed, on which you can view the job execution progress and result.

Step 5 After the job is successfully executed, in the **Operation** column of the job, click **Historical Record** to view the job's historical execution records and read/write statistics.

On the **Historical Record** page, click **Log** to view the job logs.

----End

6.8 Migrating Data from Oracle to DWS

Scenario

CDM supports table-to-table migration. This section describes how to use CDM to migrate data from Oracle to Data Warehouse Service (DWS). The procedure is as follows:

1. **Creating a CDM Cluster and Binding an EIP to the Cluster**
2. **Creating an Oracle Link**
3. **Creating a DWS Link**
4. **Creating a Migration Job**

Prerequisites

- You have obtained a DWS cluster and the IP address, port number, database name, username, and password for connecting to the DWS database. In addition, you must have the read, write, and delete permissions on the DWS database.
- You have obtained the IP address, name, username, and password of the Oracle database.
- If the Oracle database is deployed on an on-premises data center or a third-party cloud, ensure that an IP address that can be accessed from the public network has been configured for the Oracle database, or the VPN or Direct Connect between the on-premises data center and has been established.
- You have uploaded an Oracle database driver by following the instructions provided in **Managing Drivers**.

Creating a CDM Cluster and Binding an EIP to the Cluster

Step 1 If is an independent CDM service, create a CDM cluster by following the instructions provided in **Creating a Cluster**. If is used as a CDM component of

DataArts Studio, create a CDM cluster by following the instructions provided in **Creating a Cluster**.

The key configurations are as follows:

- The flavor of the CDM cluster is selected based on the amount of data to be migrated. Generally, cdm.medium meets the requirements for most migration scenarios.
- The VPC, subnet, and security group of the CDM cluster must be the same as those of the DWS cluster.
- If the same subnet and security group cannot be used, for security reasons, ensure that a security group rule has been configured to allow the CDM cluster to access the CSS cluster.

Step 2 After the CDM cluster is created, locate the row that contains the cluster and click **Bind EIP** in the **Operation** column. (CDM uses an EIP to access the Oracle data source.)

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.

----End

Creating an Oracle Link

Step 1 Click **Job Management** in the **Operation** column of the CDM cluster. On the displayed page, click the **Links** tab and then **Create Link**. The **Select Connector** page is displayed.

Figure 6-36 Selecting a connector type

Step 2 Select **Oracle** and click **Next** to configure parameters for the link.

Figure 6-37 Creating an Oracle link

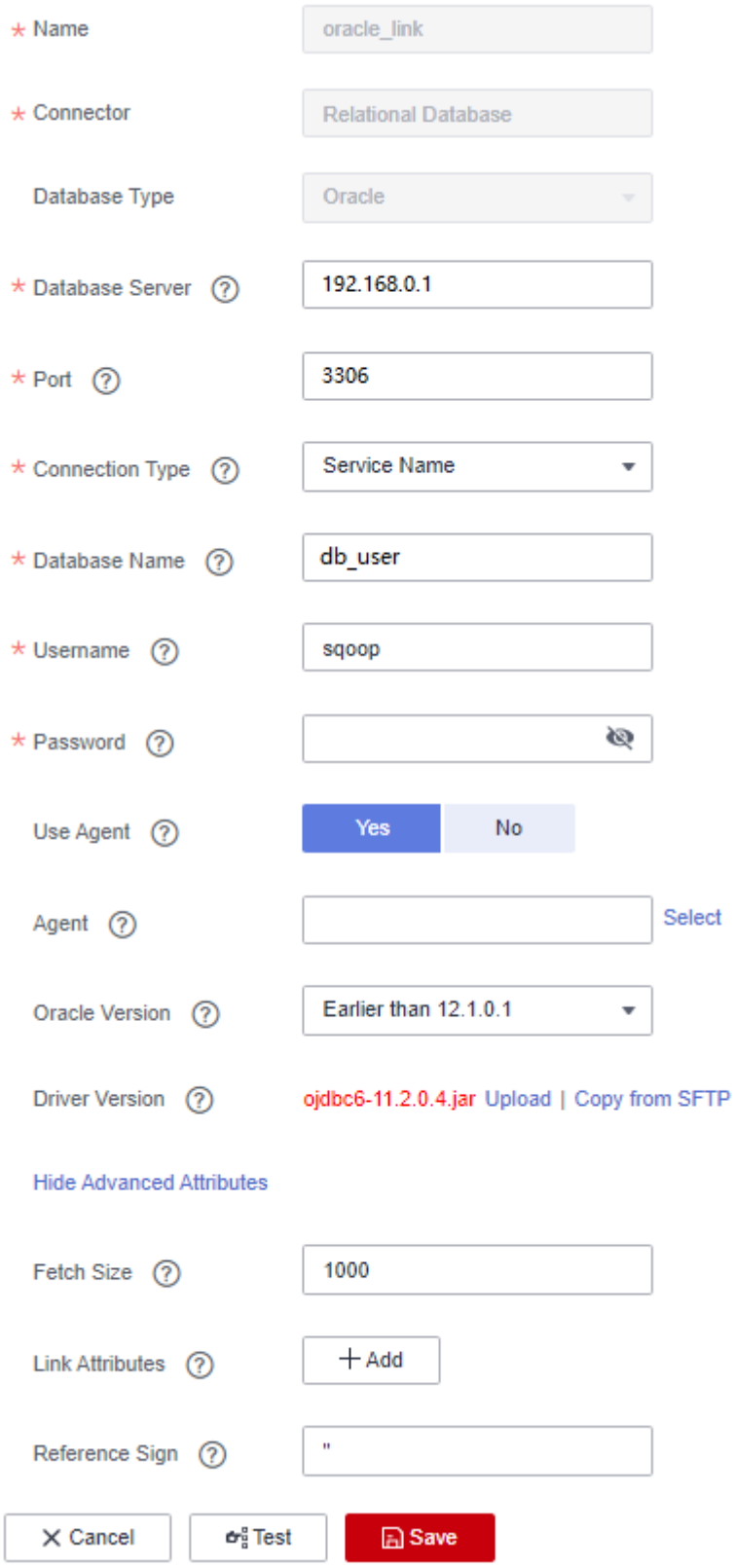


Table 6-9 Oracle link parameters

Parameter	Description	Example Value
Name	Enter a unique link name.	oracle_link
Database Server	Database server domain name or IP address	192.168.0.1
Port	Oracle database port	3306
Connection Type	Type of the Oracle database link	Service Name
Database Name	Name of the database to be connected	db_user
Username	User who has the read permission of the Oracle database	admin
Password	Password used for logging in to the Oracle database	-
Use Agent	Whether to extract data from the data source through an agent	Yes
Agent	Click **Select** and select the agent created in **Connecting to an Agent**.	-
Oracle Version	The latest version is used by default. If the version is incompatible, select another version.	Later than 12.1
Driver Version	A driver version that adapts to the Oracle database	-
Fetch Size	Number of rows obtained by each request	1000
Link Attributes	Custom attributes of the link	useCompression=true
Reference Sign	Delimiter used to separate referenced table names or column names This parameter is left blank by default.	'

Step 3 Click **Save**. The **Links** page is displayed.

----End

Creating a DWS Link

Step 1 Click **Job Management** in the **Operation** column of the CDM cluster. On the displayed page, click the **Links** tab and then **Create Link**. The **Select Connector** page is displayed.

Figure 6-38 Selecting a connector type

Step 2 Select **Data Warehouse Service** and click **Next** to configure the DWS link parameters. Set the mandatory parameters listed in **Table 6-10** and retain the default values for the optional parameters.

Table 6-10 DWS link parameters

Parameter	Description	Example Value
Name	Enter a unique link name.	dwslink
Database Server	IP address or domain name of the DWS database	192.168.0.3
Port	DWS database port	8000
Database Name	Name of the DWS database	db_demo
Username	User who has the read, write, and delete permissions on the DWS database	dbadmin
Password	Password of the user	-
Use Agent	Whether to extract data from the data source through an agent	Yes
Agent	Click **Select** and select the agent created in **Connecting to an Agent**.	-
Import Mode	**COPY**: Migrate the source data to the DWS management node and then copy the data to DataNodes. To access DWS through the Internet, select **COPY**.	COPY

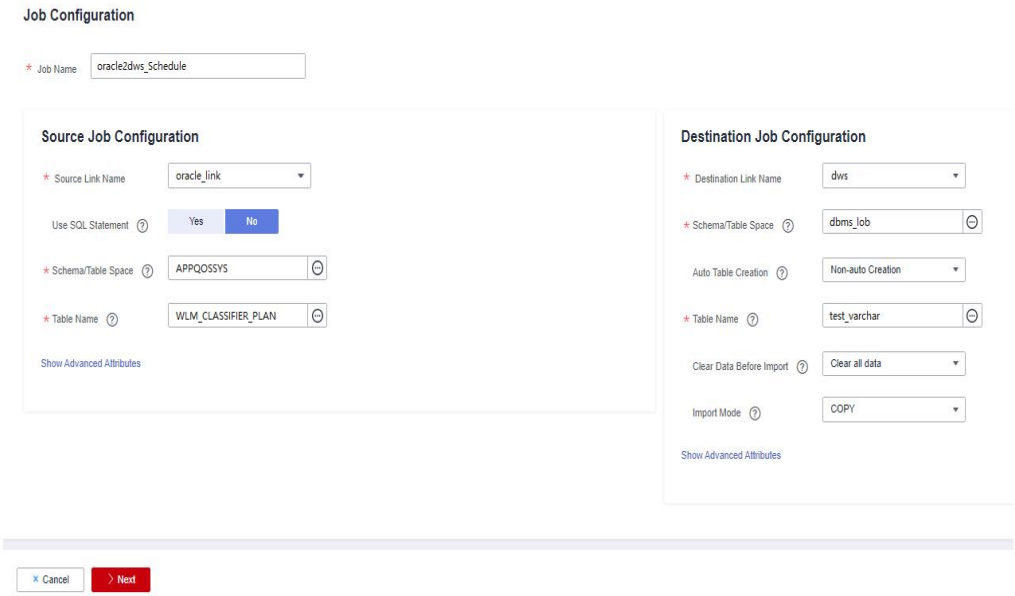
Step 3 Click **Save**.

----End

Creating a Migration Job

Step 1 Choose **Table/File Migration** > **Create Job** to create a job for exporting data from the Oracle database to DWS.

Figure 6-39 Creating a job for migrating data from Oracle to DWS



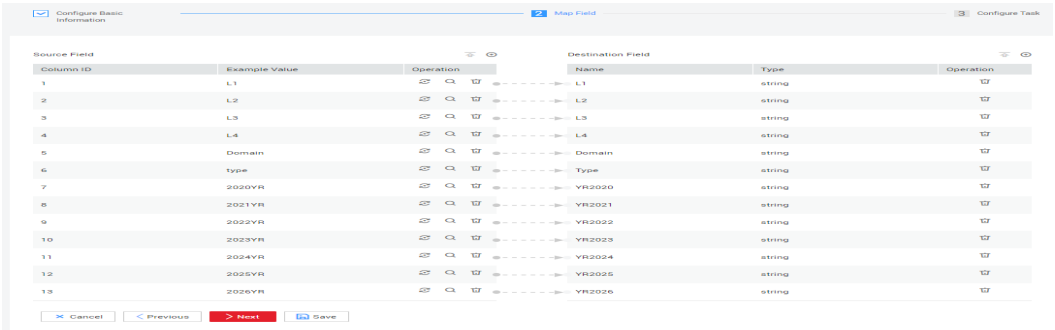
- **Job Name**: Enter a unique name.
- **Source Job Configuration**
  - **Source Link Name**: Select the **oracle_link** created in **Creating an Oracle Link**.
  - **Schema/Tablespace**: Enter the name of the database whose data is to be migrated.
  - **Table Name**: Enter the name of the table whose data is to be migrated.
  - Retain the default values of the optional parameters in **Show Advanced Attributes**. For details, see **From a Common Relational Database**.
- **Destination Job Configuration**
  - **Destination Link Name**: Select the **dwslink** created in **Creating a DWS Link**.
  - **Schema/Tablespace**: Select the DWS database to which data is to be written.
  - **Auto Table Creation**: This parameter is displayed only when both the migration source and destination are relational databases.
  - **Table Name**: Name of the table to which data is to be written. You can enter a table name that does not exist. CDM automatically creates the table in DWS.
  - **Orientation**: You can create row- or column-store tables as needed. Generally, if a table contains many columns (called a wide table) and its query involves only a few columns, column storage is recommended. If a table contains only a few columns and a query includes most of the fields, row storage is recommended.

  - **Extend char length**: If the data encoding formats of the migration source and destination are different, the character length of the automatic table creation may be insufficient. If you select **Yes** for this parameter, the character length will be increased by three times during automatic table creation.
  - **Clear Data Before Import**: whether to clear data in the destination table before the migration task starts.

Step 2 Click **Next**. The **Map Field** page is displayed. CDM automatically matches the source and destination fields, as shown in **Figure 6-40**.

- If the field mapping is incorrect, you can drag the fields to adjust the mapping.
- You can map fields in batches.
- The expressions in CDM support field conversion of common character strings, dates, and values. For details, see **Converting Fields**.

Figure 6-40 Table-to-table field mapping



Step 3 Click **Next** and set task parameters. Generally, retain the default values of all parameters.

In this step, you can configure the following optional functions:

- **Retry Upon Failure**: If the job fails to be executed, you can determine whether to automatically retry. Retain the default value **Never**.
- **Group**: Select the group to which the job belongs. The default group is **DEFAULT**. On the **Job Management** page, jobs can be displayed, started, or exported by group.
- **Schedule Execution**: To configure scheduled jobs, see **Scheduling Job Execution**. Retain the default value **No**.
- **Concurrent Extractors**: Enter the number of extractors to be concurrently executed. You can increase the value of this parameter to improve migration efficiency.
- **Write Dirty Data**: Dirty data may be generated during data migration between tables. You are advised to select **Yes**.
- **Delete Job After Completion**: Retain the default value **Do not delete**.

Step 4 Click **Save and Run**. The **Job Management** page is displayed, on which you can view the job execution progress and result.

Step 5 After the job is successfully executed, in the **Operation** column of the job, click **Historical Record** to view the job's historical execution records and read/write statistics.

On the **Historical Record** page, click **Log** to view the job logs.

----End

NOTE

If the migration times out because writing data to the destination costs a long time, reduce the value of the **Fetch Size** parameter.

6.9 Migrating Data from OBS to CSS

Scenario

CDM supports data migration between cloud services. This section describes how to use CDM to migrate data from OBS to CSS. The procedure is as follows:

1. **Creating a CDM Cluster**
2. **Creating a Cloud Search Service Link**
3. **Creating an OBS Link**
4. **Creating a Migration Job**

Prerequisites

- You have obtained the domain name, port number, AK, and SK for accessing OBS.
- You have subscribed to Cloud Search Service and obtained the IP address and port number of the Cloud Search Service cluster.

Creating a CDM Cluster

If is an independent CDM service, create a CDM cluster by following the instructions provided in **Creating a Cluster**. If is used as a CDM component of DataArts Studio, create a CDM cluster by following the instructions provided in **Creating a Cluster**.

The key configurations are as follows:

- The flavor of the CDM cluster is selected based on the amount of data to be migrated. Generally, cdm.medium meets the requirements for most migration scenarios.
- The CDM and Cloud Search Service clusters must be in the same VPC. In addition, it is recommended that the CDM cluster be in the same subnet and security group as the Cloud Search Service cluster.
- If the same subnet and security group cannot be used for security purposes, ensure that a security group rule has been configured to allow the CDM cluster to access the Cloud Search Service cluster.

Creating a Cloud Search Service Link

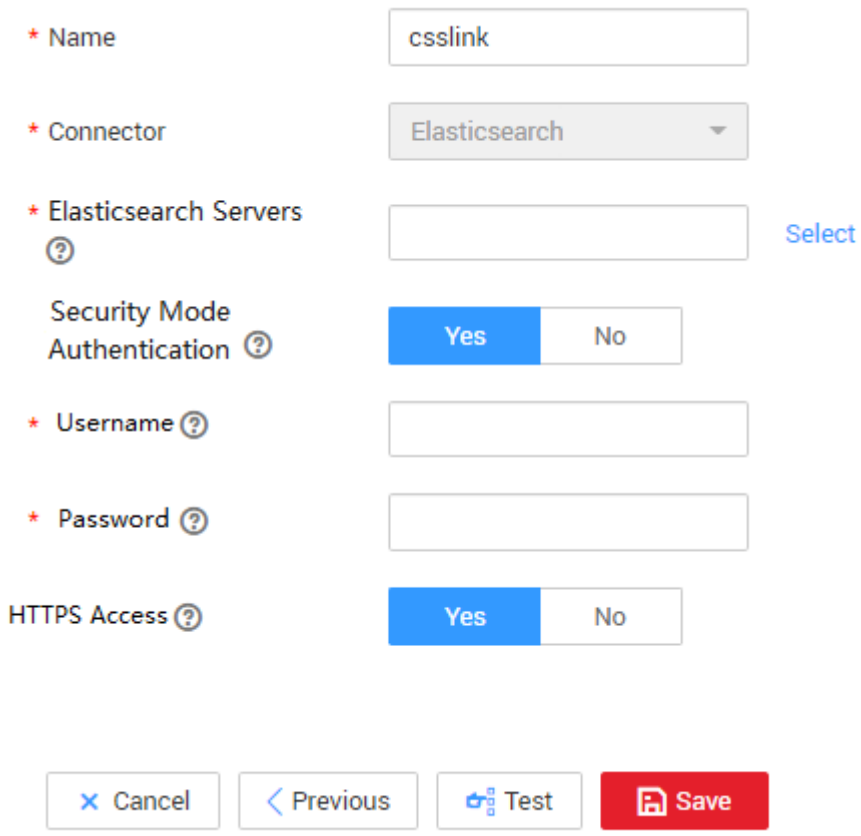
Step 1 Click **Job Management** in the **Operation** column of the CDM cluster. On the displayed page, click the **Links** tab and then **Create Link**. The **Select Connector** page is displayed.

Figure 6-41 Selecting a connector

Step 2 Select **Cloud Search Service** and click **Next**. On the page that is displayed, configure the CSS link parameters.

- **Name**: Enter a custom link name, for example, **csslink**.
- **Elasticsearch Server List**: Enter the IP address and port number of the Cloud Search Service cluster (cluster later than 5.*x*). The format is *ip:port*. Use semicolons to separate multiple addresses. For example, **192.168.0.1:9200;192.168.0.2:9200**.
- **Username** and **Password**: Enter the username and password used for logging in to the Cloud Search Service cluster. The user must have the read and write permissions on the database.

Figure 6-42 Creating a CSS link



Step 3 Click **Save**. The **Link Management** page is displayed.

----End

Creating an OBS Link

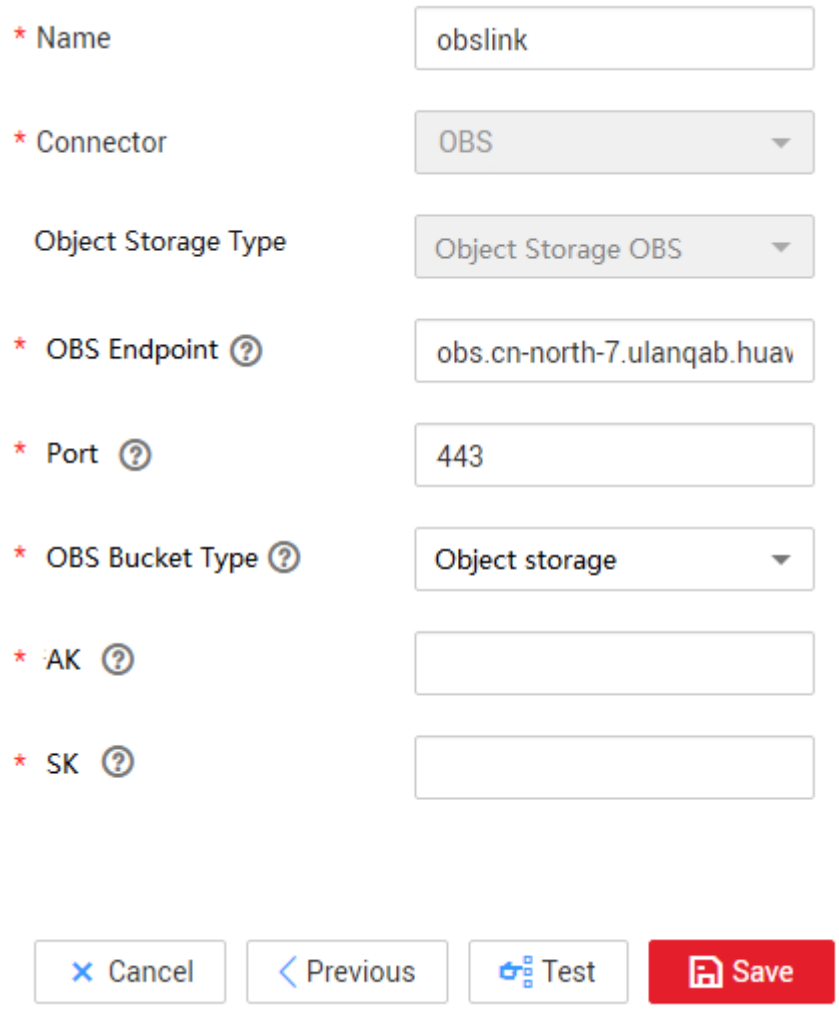
Step 1 Click **Job Management** in the **Operation** column of the CDM cluster. On the displayed page, click the **Links** tab and then **Create Link**. The **Select Connector** page is displayed.

Figure 6-43 Selecting a connector type

Step 2 Select **Object Storage Service (OBS)** and click **Next** to configure parameters for the OBS link.

- **Name**: Enter a custom link name, for example, **obslink**.
- **OBS Server** and **Port**: Enter the actual OBS address information.
- **AK** and **SK**: Enter the AK and SK used for logging in to OBS.

Figure 6-44 Creating an OBS link



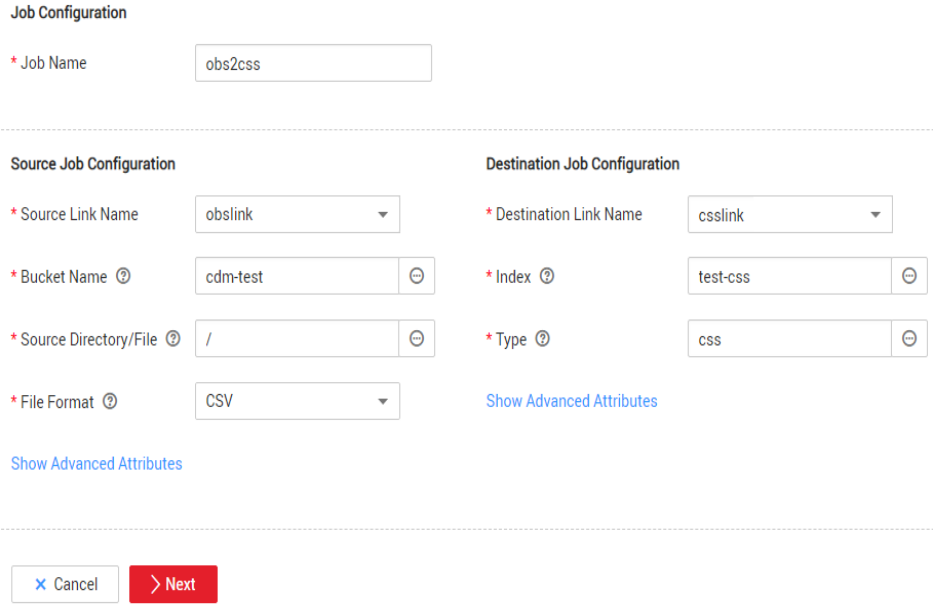
Step 3 Click **Save**. The **Link Management** page is displayed.

----End

Creating a Migration Job

Step 1 Choose **Table/File Migration** > **Create Job** to create a job for exporting data from OBS to Cloud Search Service.

Figure 6-45 Creating a job for migrating data from OBS to Cloud Search Service



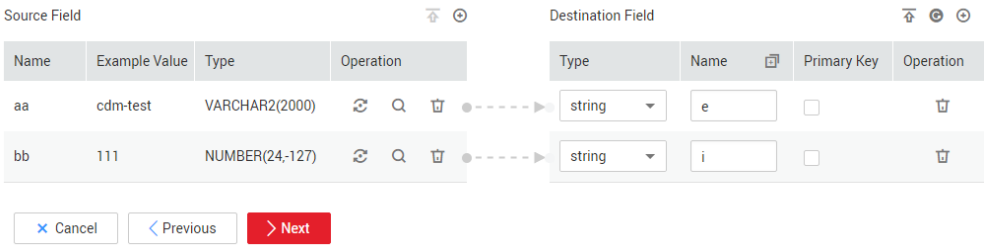
- **Job Name**: Enter a unique name.
- **Source Job Configuration**
  - **Source Link Name**: Select the **obslink** link created in **Creating an OBS Link**.
  - **Bucket Name**: Select the bucket from which the data will be migrated.
  - **Source Directory/File**: Set this parameter to the path of the data to be migrated. You can migrate all directories and files in the bucket.
  - **File Format**: Select **CSV** for migrating files to a data table.
  - Retain the default values of the optional parameters in **Show Advanced Attributes**. For details, see **From OBS**.
- **Destination Job Configuration**
  - **Destination Link Name**: Select the **csslink** link created in **Creating a Cloud Search Service Link**.
  - **Index**: Select the Elasticsearch index of the data to be written. You can also enter a new index. CDM automatically creates the index on Cloud Search Service.
  - **Type**: Select the Elasticsearch type of the data to be written. You can enter a new type. CDM automatically creates a type at the migration destination.
  - Retain the default values of the optional parameters in **Show Advanced Attributes**. For details, see **To CSS**.

Step 2 Click **Next**. The **Map Field** page is displayed. CDM automatically matches the source and destination fields. See **Figure 6-46**.

- If the field mapping is incorrect, you can drag the fields to adjust the mapping.
- If the type is automatically created at the migration destination, you need to configure the type and name of each field.

- CDM supports field conversion during the migration. For details, see **Converting Fields**.

Figure 6-46 Field mapping of Cloud Search Service



Step 3 Click **Next** and set task parameters. Generally, retain the default values of all parameters.

In this step, you can configure the following optional functions:

- **Retry Upon Failure**: If the job fails to be executed, you can determine whether to automatically retry. Retain the default value **Never**.
- **Group**: Select the group to which the job belongs. The default group is **DEFAULT**. On the **Job Management** page, jobs can be displayed, started, or exported by group.
- **Schedule Execution**: To configure scheduled jobs, see **Scheduling Job Execution**. Retain the default value **No**.
- **Concurrent Extractors**: Enter the number of extractors to be concurrently executed. Retain the default value **1**.
- **Write Dirty Data**: Specify this parameter if data that fails to be processed or filtered out during job execution needs to be written to OBS for future viewing. Before writing dirty data, create an OBS link. Retain the default value **No** so that dirty data is not recorded.
- **Delete Job After Completion**: Retain the default value **Do not delete**.

Step 4 Click **Save and Run**. The **Job Management** page is displayed, on which you can view the job execution progress and result.

Step 5 After the job is successfully executed, in the **Operation** column of the job, click **Historical Record** to view the job's historical execution records and read/write statistics.

On the **Historical Record** page, click **Log** to view the job logs.

----End

6.10 Migrating Data from OBS to DLI

Scenario

DLI is a fully hosted big data query service. This section describes how to use CDM to migrate data from OBS to DLI. The procedure includes four steps:

1. **Creating a CDM Cluster**

2. **Creating a DLI Link**
3. **Creating an OBS Link**
4. **Creating a Migration Job**

Prerequisites

- You have enabled OBS and DLI and have the permissions to read data from OBS.
- You have created resource queues, databases, and tables on DLI.

Creating a CDM Cluster

If is an independent CDM service, create a CDM cluster by following the instructions provided in **Creating a Cluster**. If is used as a CDM component of DataArts Studio, create a CDM cluster by following the instructions provided in **Creating a Cluster**.

In this scenario, if the CDM cluster is used only to migrate data from OBS to DLI and does not need to migrate data of other data sources, there is no special requirements on the VPC, subnet, and security group of the CDM cluster. You can specify them based on your needs. CDM accesses DLI and OBS through the intranet. The flavor of the CDM cluster is selected based on the amount of data to be migrated. Generally, cdm.medium meets the requirements for most migration scenarios.

Creating a DLI Link

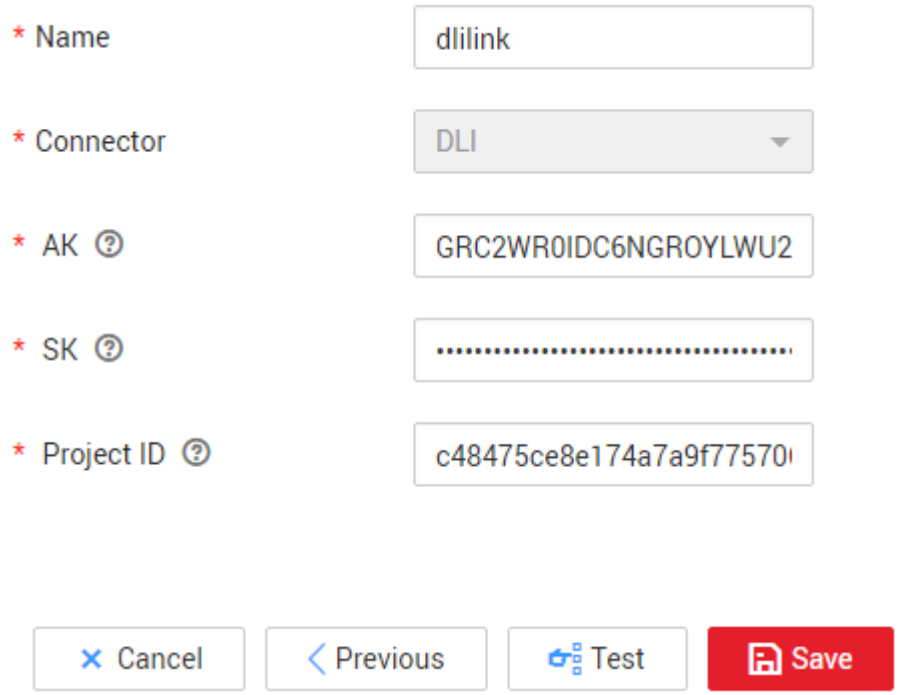
Step 1 Click **Job Management** in the **Operation** column of the CDM cluster. On the displayed page, click the **Links** tab and then **Create Link**. The **Select Connector** page is displayed.

Figure 6-47 Selecting a connector

Step 2 Select **Data Lake Insight**, click **Next**, and configure the DLI link parameters. See **Figure 6-48**.

- **Name**: Enter a custom link name, for example, **dlilink**.
- **AK** and **SK**: Enter the AK and SK used for accessing the DLI database.
- **Project ID**: Enter the project ID of the region to which DLI belongs.

Figure 6-48 Creating a DLI link



Step 3 Click **Save**. The **Link Management** page is displayed.

----End

Creating an OBS Link

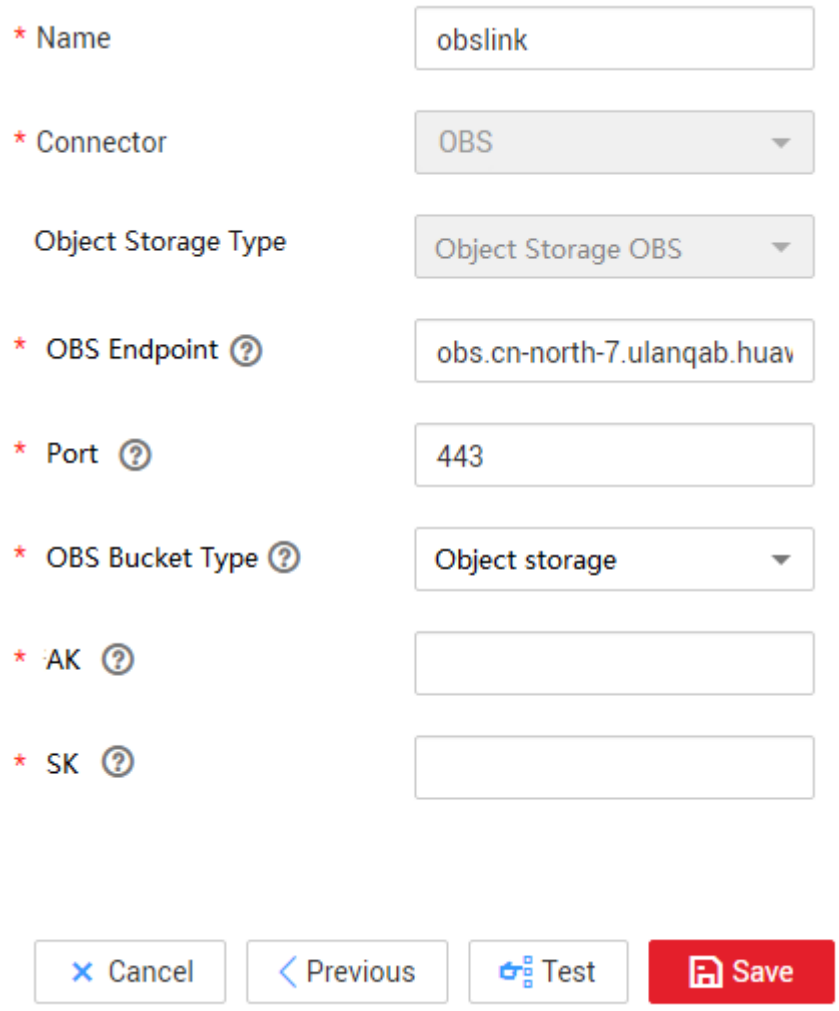
Step 1 Click **Job Management** in the **Operation** column of the CDM cluster. On the displayed page, click the **Links** tab and then **Create Link**. The **Select Connector** page is displayed.

Figure 6-49 Selecting a connector type

Step 2 Select **Object Storage Service (OBS)** and click **Next** to configure parameters for the OBS link.

- **Name**: Enter a custom link name, for example, **obslink**.
- **OBS Server** and **Port**: Enter the actual OBS address information.
- **AK** and **SK**: Enter the AK and SK used for logging in to OBS.

Figure 6-50 Creating an OBS link



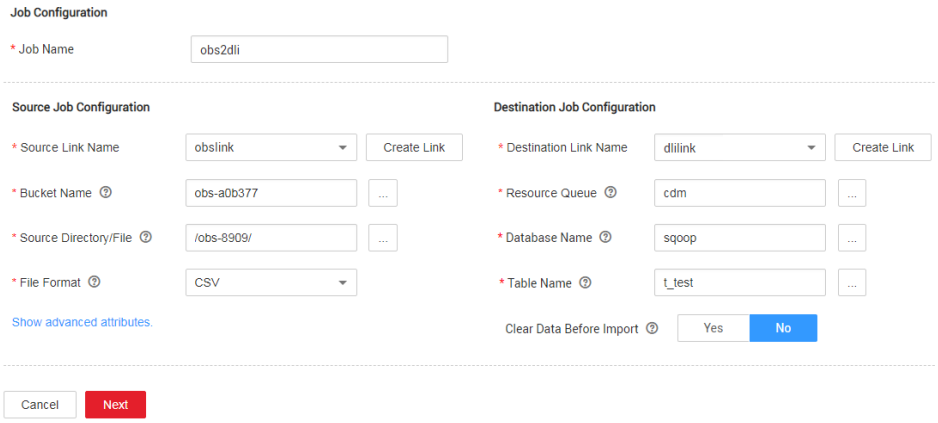
Step 3 Click **Save**. The **Link Management** page is displayed.

----End

Creating a Migration Job

Step 1 Choose **Table/File Migration** > **Create Job** to create a job for migrating data from OBS to DLI. See **Figure 6-51**.

Figure 6-51 Creating a job for migrating data from OBS to DLI



- **Job Name**: Enter a custom job name.
- **Source Link Name**: Select the **obslink** link created in **Creating an OBS Link**.
  - **Bucket Name**: Select the bucket from which the data is to be migrated.
  - **Source Directory/File**: Set this parameter to the path of the data to be migrated.
  - **File Format**: Select **CSV** or **JSON** for transferring files to a data table.
  - Retain the default values of the optional parameters in **Show Advanced Attributes**. For details, see **From OBS**.
- **Destination Link Name**: Select the **dlilink** link created in **Creating a DLI Link**.
  - **Resource Queue**: Enter the resource queue to which the destination table belongs.
  - **Database Name**: Enter the name of the database to which data is to be written.
  - **Table Name**: Enter the name of the table to which data is to be written. CDM cannot automatically create tables on DLI. The table must be created on DLI in advance, and the field types and formats of the table must be consistent with those of the data to be migrated.
  - **Clear Before Importing Data**: Choose whether to clear data in the destination table before data import. In this example, retain the default value.

Step 2 Click **Next**. The **Map Field** page is displayed. CDM automatically matches the source and destination fields.

- If the field mapping is incorrect, you can drag the fields to adjust the mapping.
- CDM supports field conversion during the migration. For details, see **Converting Fields**.

Step 3 Click **Next** and set task parameters. Generally, retain the default values of all parameters.

In this step, you can configure the following optional functions:

- **Retry Upon Failure**: If the job fails to be executed, you can determine whether to automatically retry. Retain the default value **Never**.

- **Group**: Select the group to which the job belongs. The default group is **DEFAULT**. On the **Job Management** page, jobs can be displayed, started, or exported by group.
- **Schedule Execution**: To configure scheduled jobs, see **Scheduling Job Execution**. Retain the default value **No**.
- **Concurrent Extractors**: Enter the number of extractors to be concurrently executed. Retain the default value **1**.
- **Write Dirty Data**: Specify this parameter if data that fails to be processed or filtered out during job execution needs to be written to OBS for future viewing. Before writing dirty data, create an OBS link. Retain the default value **No** so that dirty data is not recorded.
- **Delete Job After Completion**: Retain the default value **Do not delete**.

Step 4 Click **Save and Run**. The **Job Management** page is displayed, on which you can view the job execution progress and result.

Step 5 After the job is successfully executed, in the **Operation** column of the job, click **Historical Record** to view the job's historical execution records and read/write statistics.

On the **Historical Record** page, click **Log** to view the job logs.

----End

6.11 Migrating Data from MRS HDFS to OBS

Scenario

CDM supports file-to-file data migration. This section describes how to migrate data from MRS HDFS to OBS. The process is as follows:

1. **Creating a CDM Cluster and Binding an EIP to the Cluster**
2. **Creating an MRS HDFS Link**
3. **Creating an OBS Link**
4. **Creating a Migration Job**

Prerequisites

- You have obtained the domain name, port number, AK, and SK for accessing OBS.
- MRS is available.
- Your EIP quota is sufficient.

Creating a CDM Cluster and Binding an EIP to the Cluster

Step 1 If is an independent CDM service, create a CDM cluster by following the instructions provided in **Creating a Cluster**. If is used as a CDM component of DataArts Studio, create a CDM cluster by following the instructions provided in **Creating a Cluster**.

The key configurations are as follows:

- The flavor of the CDM cluster is selected based on the amount of data to be migrated. Generally, cdm.medium meets the requirements for most migration scenarios.
- The VPC, subnet, and security group of the CDM cluster must be the same as those of the MRS cluster.

Step 2 After the CDM cluster is created, on the **Cluster Management** page, click **Bind EIP** in the **Operation** column to bind an EIP to the cluster. The CDM cluster uses the EIP to access MRS HDFS.

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.

----End

Creating an MRS HDFS Link

Step 1 On the **Cluster Management** page, locate a cluster and click **Job Management** in the **Operation** column. On the displayed page, click the **Links** tab and then **Create Link**.

Figure 6-52 Selecting a connector type

Step 2 Select **MRS HDFS** and click **Next** to configure parameters for the MRS HDFS link.

- **Name**: Enter a custom link name, for example, **mrs_hdfs_link**.
- **Manager IP**: IP address of MRS Manager. Click **Select** next to the **Manager IP** text box to select a created MRS cluster. CDM automatically fills in the authentication information.
- **Username**: If **Authentication Method** is set to **KERBEROS**, set the username and password for logging in to MRS Manager.

  If you need to create a snapshot when exporting a directory from HDFS, the user configured here must have the administrator permission on HDFS.
- **Password**: password for logging in to MRS Manager
- **Authentication Method**: authentication method for accessing MRS

- **Run Mode**: Select the running mode of the HDFS link.

----End

Creating an OBS Link

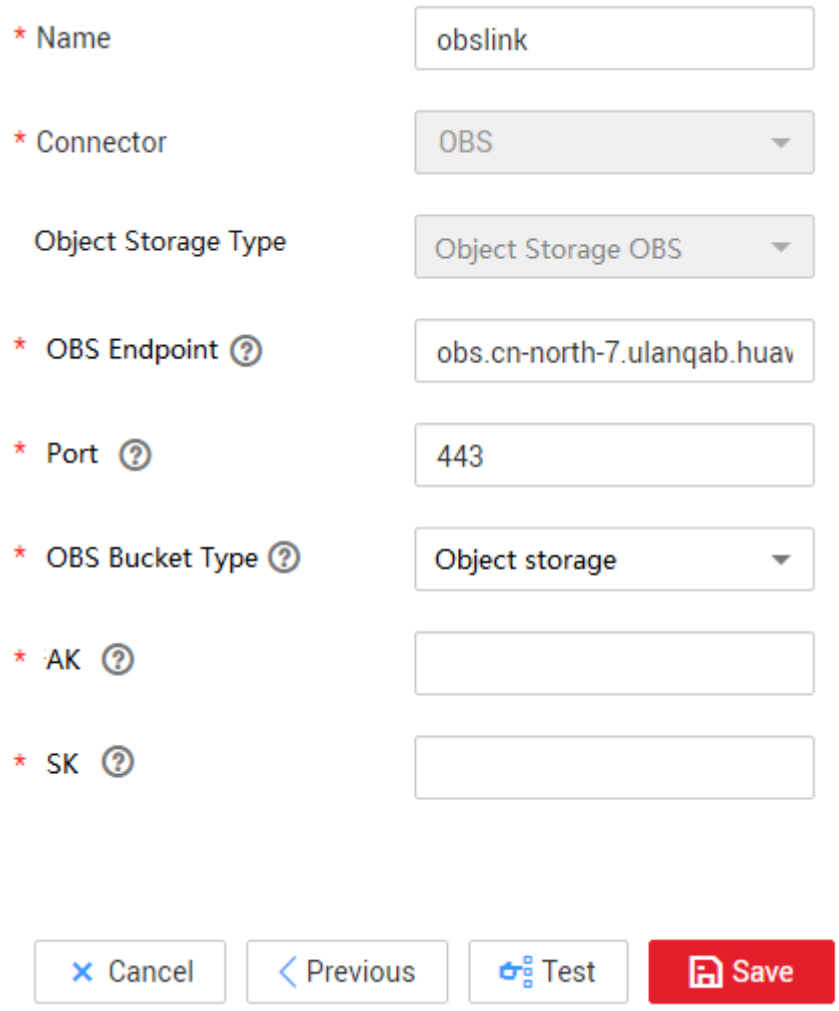
Step 1 Click **Job Management** in the **Operation** column of the CDM cluster. On the displayed page, click the **Links** tab and then **Create Link**. The **Select Connector** page is displayed.

Figure 6-53 Selecting a connector type

Step 2 Select **Object Storage Service (OBS)** and click **Next** to configure parameters for the OBS link.

- **Name**: Enter a custom link name, for example, **obslink**.
- **OBS Server** and **Port**: Enter the actual OBS address information.
- **AK** and **SK**: Enter the AK and SK used for logging in to OBS.

Figure 6-54 Creating an OBS link



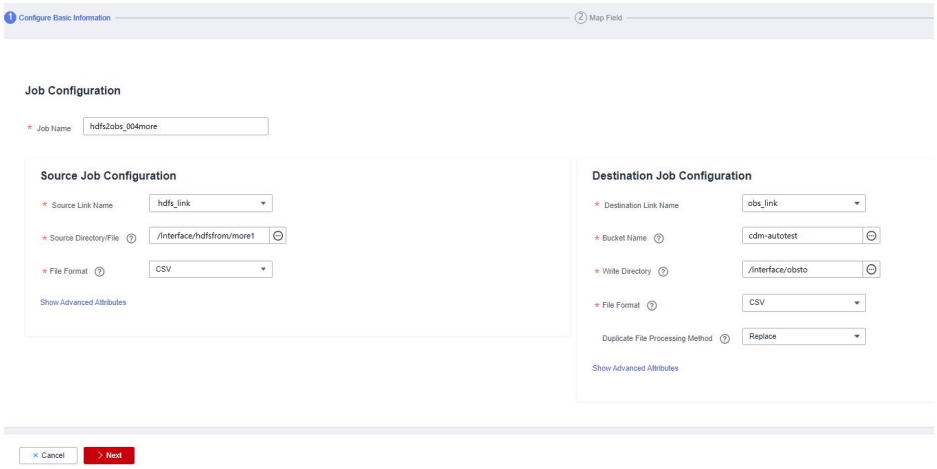
Step 3 Click **Save**. The **Link Management** page is displayed.

----End

Creating a Migration Job

Step 1 Choose **Table/File Migration** > **Create Job** to create a job for exporting data from the MRS HDFS database to OBS.

Figure 6-55 Creating a job for migrating data from MRS HDFS to OBS



- **Job Name**: Enter a unique name.
- **Source Job Configuration**
  - **Source Link Name**: Select the **hdfs_llink** created in **Creating an MRS HDFS Link**.
  - **Source Directory/File**: Enter the directory or file path of the data to be migrated.
  - **File Format**: Select the file format used for data transmission. Select **Binary**. If files are transferred without being parsed, the file format does not have to be **Binary**. This applies to file copy.
  - Retain the default values of other optional parameters. For details, see **From HDFS**.
- **Destination Job Configuration**
  - **Destination Link Name**: Select the **obs_link** created in **Creating an OBS Link**.
  - **Bucket Name**: Select the bucket from which the data will be migrated.
  - **Write Directory**: Enter the directory to which data is to be written on the OBS server.
  - **File Format**: Select **Binary**.
  - Retain the default values of the optional parameters in **Show Advanced Attributes**. For details, see **To OBS**.

Step 2 Click **Next**. The **Map Field** page is displayed. CDM automatically matches the source and destination fields.

- If the field mapping is incorrect, you can drag the fields to adjust the mapping.
- The expressions in CDM support field conversion of common character strings, dates, and values. For details, see **Converting Fields**.

Step 3 Click **Next** and set task parameters. Generally, retain the default values of all parameters.

In this step, you can configure the following optional functions:

- **Retry Upon Failure**: If the job fails to be executed, you can determine whether to automatically retry. Retain the default value **Never**.

- **Group**: Select the group to which the job belongs. The default group is **DEFAULT**. On the **Job Management** page, jobs can be displayed, started, or exported by group.
- **Schedule Execution**: To configure scheduled jobs, see **Scheduling Job Execution**. Retain the default value **No**.
- **Concurrent Extractors**: Enter the number of extractors to be concurrently executed. CDM supports concurrent extraction of multiple files. Increasing the value of this parameter can improve migration efficiency.
- **Write Dirty Data**: Select **No**. The file-to-file migration is binary, and no dirty data will be generated.
- **Delete Job After Completion**: Retain the default value **Do not delete**. You can also set this parameter to **Delete** to prevent an accumulation of too many migration jobs.

Step 4 Click **Save and Run**. The **Job Management** page is displayed, on which you can view the job execution progress and result.

Step 5 After the job is successfully executed, in the **Operation** column of the job, click **Historical Record** to view the job's historical execution records and read/write statistics.

On the **Historical Record** page, click **Log** to view the job logs.

----End

6.12 Migrating the Entire Elasticsearch Database to CSS

Scenario

CSS provides users with structured and unstructured data search, statistics, and report capabilities. This section describes how to use CDM to migrate the entire Elasticsearch database to Cloud Search Service. The procedure is as follows:

1. **Creating a CDM Cluster and Binding an EIP to the Cluster**
2. **Creating a Cloud Search Service Link**
3. **Creating an Elasticsearch Link**
4. **Creating an Entire DB Migration Job**

Prerequisites

- You have sufficient EIP quota.
- You have subscribed to CSS and obtained the IP address and port number of the CSS cluster.
- You have obtained the IP address, port number, username, and password of the on-premises Elasticsearch database server.

  If the Elasticsearch server is deployed on an on-premises data center or a third-party cloud, ensure that an IP address that can be accessed from the public network has been configured for the Elasticsearch server, or the VPN or Direct Connect between the on-premises data center and HUAWEI CLOUD has been established.

Creating a CDM Cluster and Binding an EIP to the Cluster

Step 1 If is an independent CDM service, create a CDM cluster by following the instructions provided in **Creating a Cluster**. If is used as a CDM component of DataArts Studio, create a CDM cluster by following the instructions provided in **Creating a Cluster**.

The key configurations are as follows:

- The flavor of the CDM cluster is selected based on the amount of data to be migrated. Generally, cdm.medium meets the requirements for most migration scenarios.
- The CDM and Cloud Search Service clusters must be in the same VPC. In addition, it is recommended that the CDM cluster be in the same subnet and security group as the Cloud Search Service cluster.
- If the same subnet and security group cannot be used for security purposes, ensure that a security group rule has been configured to allow the CDM cluster to access the Cloud Search Service cluster.

Step 2 After the CDM cluster is created, on the **Cluster Management** page, click **Bind EIP** in the **Operation** column to bind an EIP to the cluster. The CDM cluster uses the EIP to access the on-premises Elasticsearch.

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.

----End

Creating a Cloud Search Service Link

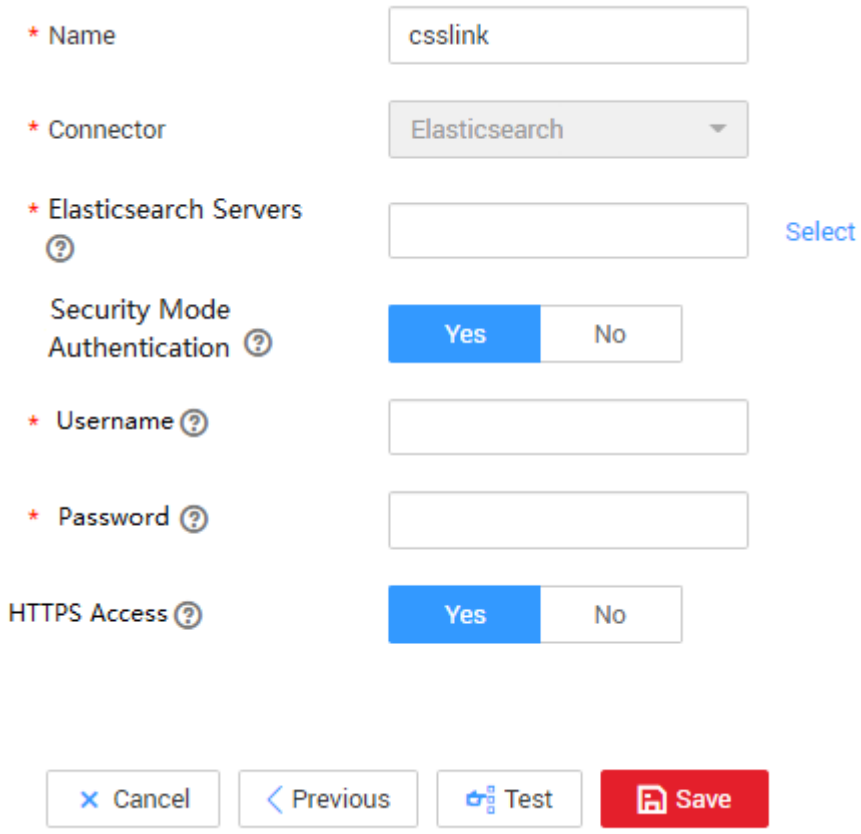
Step 1 Click **Job Management** in the **Operation** column of the CDM cluster. On the displayed page, click the **Links** tab and then **Create Link**. The **Select Connector** page is displayed.

Figure 6-56 Selecting a connector

Step 2 Select **Cloud Search Service** and click **Next**. On the page that is displayed, configure the CSS link parameters.

- **Name**: Enter a custom link name, for example, **csslink**.
- **Elasticsearch Server List**: Enter the IP address and port number of the Cloud Search Service cluster (cluster later than 5.*x*). The format is *ip:port*. Use semicolons to separate multiple addresses. For example, **192.168.0.1:9200;192.168.0.2:9200**.
- **Username** and **Password**: Enter the username and password used for logging in to the Cloud Search Service cluster. The user must have the read and write permissions on the database.

Figure 6-57 Creating a CSS link



Step 3 Click **Save**. The **Link Management** page is displayed.

----End

Creating an Elasticsearch Link

Step 1 On the **Cluster Management** page, locate a cluster and click **Job Management** in the **Operation** column. On the displayed page, click the **Links** tab and then **Create Link**.

Figure 6-58 Selecting a connector type

Step 2 Select **Elasticsearch** and click **Next** to configure parameters for the Elasticsearch link. The parameters are the same as those for the CSS link.

- **Name**: Enter a custom link name, for example, **es_link**.
- **Elasticsearch Server List**: Enter the IP address and port number of the on-premises Elasticsearch database. Use semicolons to separate multiple addresses.

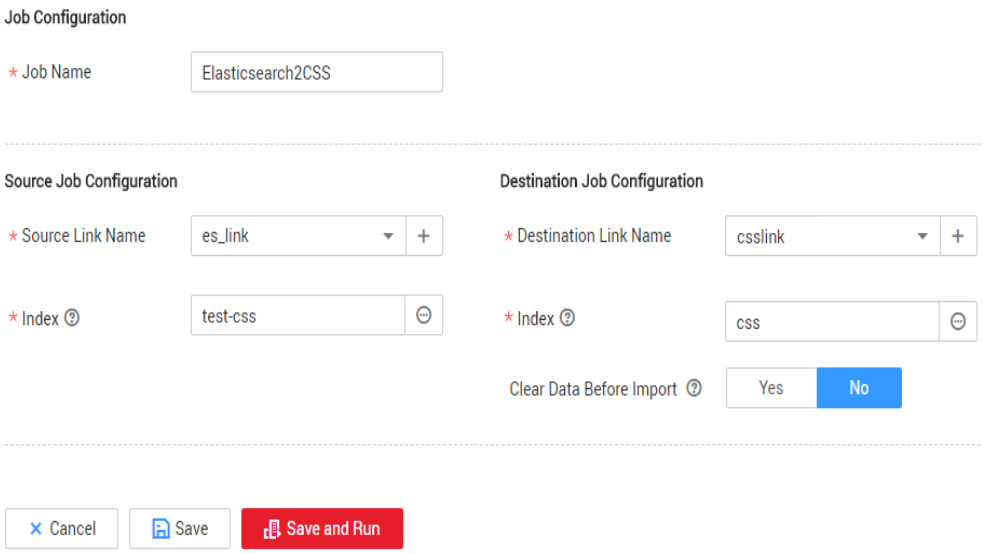
Step 3 Click **Save**. The **Link Management** page is displayed.

----End

Creating an Entire DB Migration Job

Step 1 Choose **Entire DB Migration** > **Create Job** to create an entire DB migration job.

Figure 6-59 Creating an entire DB migration job



- **Job Name**: Enter a unique name.

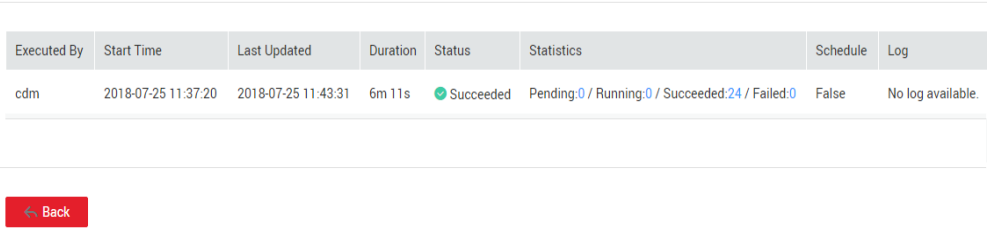
- **Source Job Configuration**
  - **Source Link Name**: Select the **es_link** link created in **Creating an Elasticsearch Link**.
  - **Index**: Click the icon next to the text box to select an index in the on-premises Elasticsearch database or manually enter an index name. The name can contain only lowercase letters. If multiple indexes need to be migrated at a time, set this parameter to a wildcard character. CDM migrates all indexes that meet the wildcard condition. For example, if this parameter is set to **cdm***, CDM migrates all indexes starting with **cdm**, such as **cdm01**, **cdmB3**, **cdm_45** and so on.
- **Destination Job Configuration**
  - **Destination Link Name**: Select the **csslink** link created in **Creating a Cloud Search Service Link**.
  - **Index**: Enter the index of the data to be written. You can select an existing index in Cloud Search Service or manually enter an index name that does not exist. The name can contain only lowercase letters. CDM automatically creates the index in Cloud Search Service. If multiple indexes are migrated at a time, this parameter cannot be configured. CDM automatically creates indexes at the migration destination.
  - **Clear Data Before Import**: If the selected index already exists in Cloud Search Service, you can choose whether to clear the data in the index before importing data. If you select **No**, the data is added to the index.

Step 2 Click **Save and Run**. The **Job Management** page is displayed, on which you can view the job execution progress and result.

A sub-job will be generated for each type in the on-premises Elasticsearch index for concurrent execution. You can click the job name to view the sub-job progress.

Step 3 After the job is successfully executed, in the **Operation** column of the job, click **Historical Record** to view the job's historical execution records, read/write statistics, and job logs (only the sub-jobs have job logs).

Figure 6-60 Historical Record



----End

6.13 More Cases and Practices

For more advanced guidance and cases of DataArts Migration, see **Best Practices**.

7 Advanced Data Migration Guidance

7.1 Incremental Migration

7.1.1 Incremental File Migration

CDM supports incremental migration of file systems. After full migration is complete, all new files or only specified directories or files can be exported.

Currently, CDM supports the following incremental migration modes:

1. **Exporting the files in a specified directory**
   - Application scenarios: The migration source is a file system (OBS/HDFS/FTP/SFTP). In incremental migration, only the specified files are written to the migration destination. The existing records are not updated or deleted.
   - Key configurations: **File/Path Filter** and Schedule Execution
   - Prerequisites: The source directory or file name contains the time field.
2. **Exporting the files modified after the specified time point**
   - Application scenarios: The migration source is a file system (OBS/HDFS/FTP/SFTP). The specified time point refers to the time when the file is modified. CDM migrates the files modified after the specified time point.
   - Key configurations: **Time Filter** and Schedule Execution
   - Prerequisites: None

NOTE

If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).

File/Path Filter

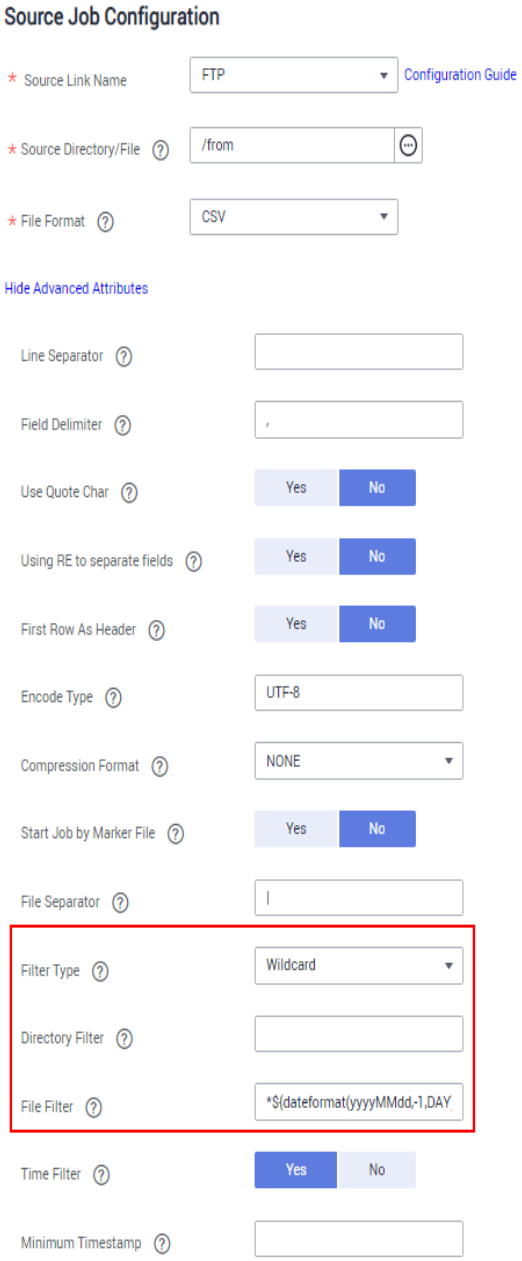
- Parameter position: When creating a table/file migration job, if the migration source is a file system, set **Filter Type** in advanced attributes of **Source Job Configuration** to **Wildcard** or **Regular expression**.
- Parameter principle: If you select **Wildcard** for **Filter Type**, CDM filters files or paths based on the configured wildcard character and migrates only files or paths that meet the specified condition.
- Example configurations:

  Suppose that the source file name contains the date and time field, such as **2017-10-15 20:25:26**, the **/opt/data/file_20171015202526.data** file is generated. Set the parameters as follows:

  a. **Filter Type**: Select **Wildcard**.

  b. **File Filter**: Enter **"*${dateformat(yyyyMMdd,-1,DAY)}*"**, which is the format of the macro variables of date and time supported by CDM. For details, see **Using Macro Variables of Date and Time**.

Figure 7-1 Filtering files



c. Schedule Execution: Set **Cycle (days)** to **1**.

In this way, you can import the files generated in the previous day to the destination directory every day to implement incremental synchronization.

In incremental file migration, **Path Filter** is used in the same way as **File Filter**. The path name must contain the time field. In this case, all files in the specified path can be synchronized periodically.

Time Filter

- Parameter position: When creating a table/file migration job, if the migration source is a file system, set select **Yes** for **Time Filter**.
- Parameter principle: Only files generated from the **Minimum Timestamp** to the **Maximum Timestamp** will be migrated by CDM.

- Example configurations:

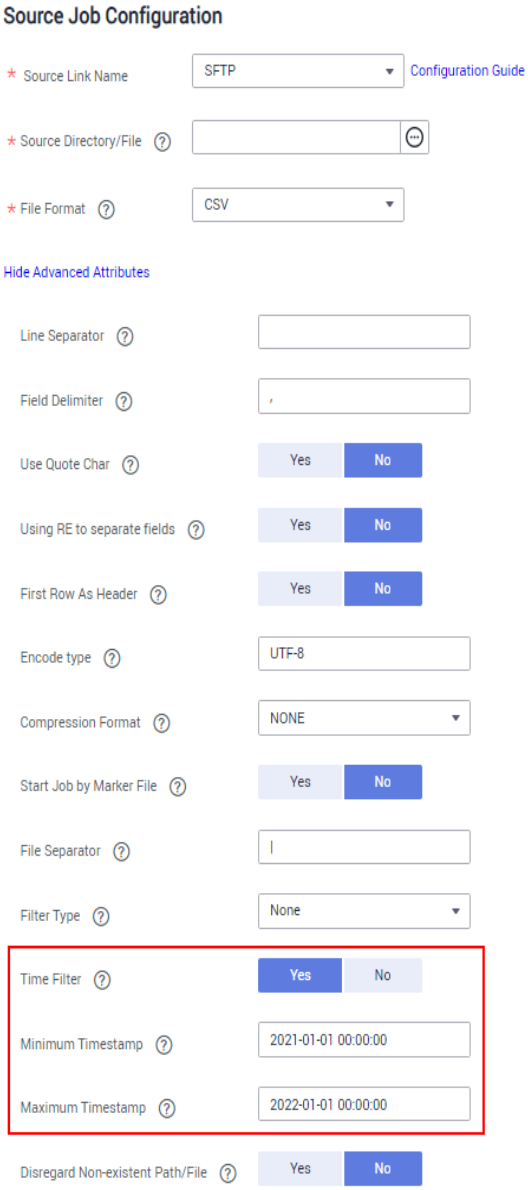
  For example, if you want CDM to synchronize only the files generated from January 1, 2021 to January 1, 2022 to the destination, configure the following parameters:

  a. **Time Filter**: select **Yes**.

  b. **Minimum Timestamp**: Enter a value in the format of *yyyy-MM-dd HH:mm:ss*, such as **2021-01-01 00:00:00**.

  c. **Maximum Timestamp**: Enter a value in the format of *yyyy-MM-dd HH:mm:ss*, such as **2022-01-01 00:00:00**.

Figure 7-2 Time Filter



In this way, the CDM job migrates only the files generated from January 1, 2021 to January 1, 2022, and performs incremental synchronization next time it is started.

7.1.2 Incremental Migration of Relational Databases

CDM supports incremental migration of relational databases. After a full migration is complete, data in a specified period can be incrementally migrated. For example, data added on the previous day can be exported at 00:00:00 every day.

- **Migrating incremental data within a specified period of time**
  - Application scenarios: The source end is a relational database. The destination end can be of any type.
  - Key configurations: **WHERE Clause** and Schedule Execution
  - Prerequisites: The data table contains a date and time field or timestamp field.

In incremental migration, only the specified data is written to the data table. The existing records are not updated or deleted.

NOTE

If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).

WHERE Clause

- Parameter position: When creating a table/file migration job, if the source end is a relational database, the **Where Clause** parameter is available in the advanced attributes of **Source Job Configuration**.
- Parameter principle: Set **WHERE Clause** to an SQL statement, for example, **age > 18 and age <= 60**, CDM exports only the data that meets the SQL statement requirement. If **WHERE Clause** is not specified, the entire table is exported.

  Where Clause can be set to **macro variables of date and time**. When the data table contains the **date** or **timestamp** field, **Where Clause** and Schedule Execution can be used together to extract data of a specified date.
- Example configurations:

  Suppose that the database table contains column **DS** indicating the time, the value type of the column is **varchar(30)**, and the inserted time format is similar to *2017-xx-xx*. See **Figure 7-3**. Set the parameters as follows:

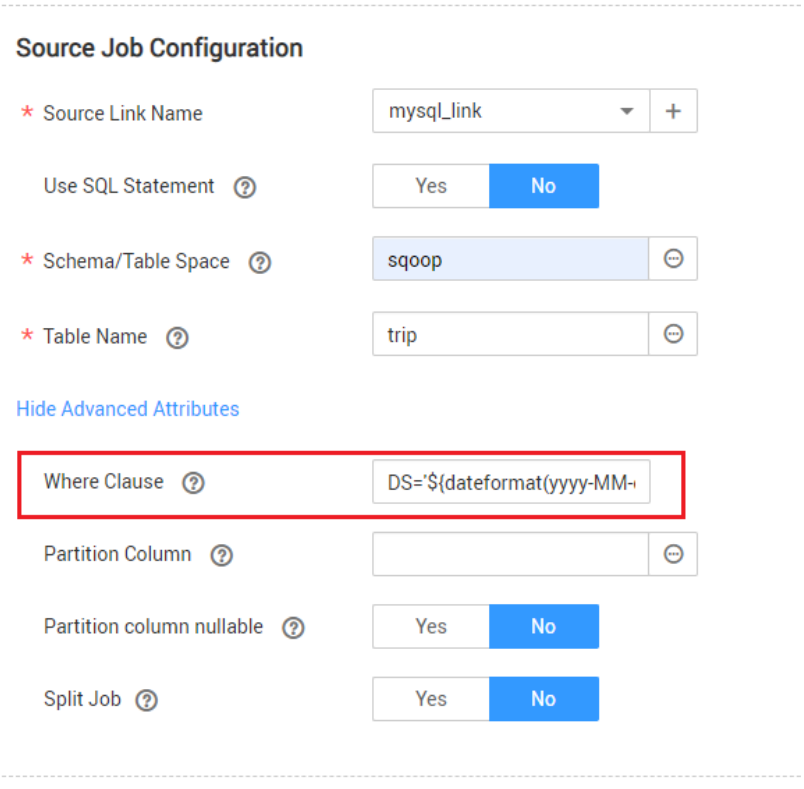
Figure 7-3 Table data

SELECT * FROM SQOOP.CDM_20171016

	FOO	BAR	DS
1	5	snap	2017-05-01
2	5	snap	2017-05-01
3	1	google	2017-05-02
4	4	oracle	2017-05-02
5	6	amd	2017-05-02
6	7	nvda	2017-05-02
7	1	google	2017-05-02
8	4	oracle	2017-05-02
9	6	amd	2017-05-02
10	7	nvda	2017-05-02
11	2	facebook	2017-10-15
12	3	tesla	2017-10-15
13	2	facebook	2017-10-15
14	3	tesla	2017-10-15

a. **WHERE Clause**: Set this parameter to **DS='${dateformat(yyyy-MM-dd,-1,DAY)}'**.

Figure 7-4 WHERE Clause



b. Scheduling job execution: Set **Cycle (days)** to **1** and **Start Time** to **00:00:00**.

In this way, all data generated on the previous day can be exported at 00:00:00 every day. **WHERE Clause** can be configured to various **macro variables of date and time**. You can use the macro variables of date and time and scheduled jobs with specified cycle of minutes, hours, days, weeks, or months together to automatically export data at a specific time.

7.1.3 HBase/CloudTable Incremental Migration

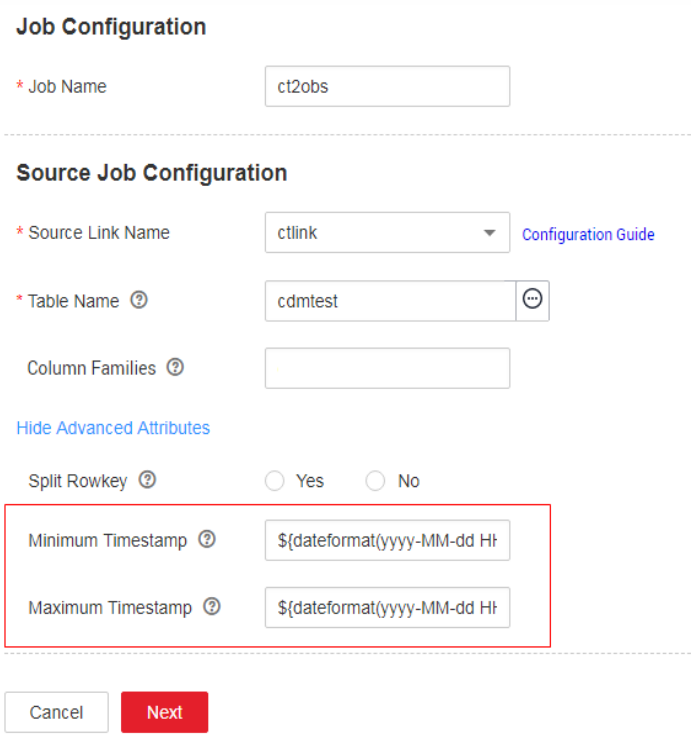
You can use CDM to export data in a specified period of time from HBase (including MRS HBase, FusionInsight HBase, and Apache HBase) and CloudTable. The CDM scheduled jobs can be used together to implement incremental migration of HBase and CloudTable.

NOTE

If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).

When creating a table/file migration job and selecting the link to HBase or CloudTable as the source link, you can set the time range in advanced attributes.

Figure 7-5 Time range



- Start time (including the value) for extracting data. The format is *yyyy-MM-dd HH:mm:ss*. Only the data generated at the specified time and later is extracted.

- End time (excluding the value) for extracting data. The format is *yyyy-MM-dd HH:mm:ss*. Only the data generated before the time point is extracted.

The two parameters can be set to **macro variables of date and time**. Examples are as follows:

- If **Minimum Timestamp** is set to **${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}**, only the data generated after the day before is exported.
- If **Maximum Timestamp** is set to **${dateformat(yyyy-MM-dd HH:mm:ss)}**, only the data generated before the specified time point is exported.

If both parameters are configured, CDM exports only the data generated on the previous day. In addition, if the job is configured to execute at 00:00:00 every day, the data generated every day can be incrementally synchronized.

7.2 Using Macro Variables of Date and Time

During the creation of table/file migration jobs, CDM supports the macro variables of date and time in the following parameters of the source and destination links:

- Source directory or file
- Source table name
- Directory filter and file filter of the **wildcard** type
- Start time and end time of the **time filter** type
- Partition filter criteria and where clause
- Write directory
- Destination table name

You can use the **${}** macro variable definition identifier to define the macros of the time type. currently, dateformat and timestamp are supported.

By using the macro variables of date and time and scheduled job, you can implement incremental synchronization of databases and files.

NOTE

If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (*Planned start time of the data development job – Offset*) rather than (*Actual start time of the CDM job – Offset*).

dateformat

dateformat supports two types of parameters:

- **dateformat(format)**

  format indicates the date and time format. For details about the format definition, see the definition in **java.text.SimpleDateFormat.java**.

  For example, if the current date is **2017-10-16 09:00:00**, **yyyy-MM-dd HH:mm:ss** indicates **2017-10-16 09:00:00**.

- dateformat(format, dateOffset, dateType)
  - **format** indicates the format of the returned date.

  - **dateOffset** indicates the date offset.
  - **dateType** indicates the type of the date offset.

    Currently, **dateType** supports SECOND, MINUTE, HOUR, and DAY.

  For example, if the current date is **2017-10-16 09:00:00**, then:

  - **dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)** indicates the day before the current day, that is, **2017-10-15 09:00:00**.
  - **dateformat(yyyy-MM-dd HH:mm:ss, -1, HOUR)** indicates one hour before the current time, that is, **2017-10-16 08:00:00**.
  - **dateformat(yyyy-MM-dd HH:mm:ss, -1, MINUTE)** indicates one minute before the current time, that is, **2017-10-16 08:59:00**.
  - **dateformat(yyyy-MM-dd HH:mm:ss, -1, SECOND)** indicates one second before the current time, that is, **2017-10-16 08:59:59**.

timestamp

timestamp supports two types of parameters:

- **timestamp()**

  Indicates the returned timestamp of the current time, that is, the number of milliseconds that have elapsed since 00:00:00 on January 1, 1970 (1970-01-01 00:00:00 GMT). For example, 1508078516286.

- **timestamp(dateOffset, dateType)**

  Indicates the timestamp returned after time offset. **dateOffset** and **dateType** indicate the date offset and the offset type, respectively.

  For example, if the current date is **2017-10-16 09:00:00**, **timestamp(-10, MINUTE)** indicates that the timestamp generated 10 minutes before the current time point is returned, that is, **1508115000000**.

Macro Variable Definition of Time and Date

Suppose that the current time is **2017-10-16 09:00:00**, then **Table 7-1** describes the macro variable definitions of time and date.

Table 7-1 Macro variable definition of time and date

Macro Variable	Description	Display Effect
${dateformat(yyyy-MM-dd)}	Returns the current date in **yyyy-MM-dd** format.	2017-10-16
${dateformat(yyyy/MM/dd)}	Returns the current date in **yyyy/MM/dd** format.	2017/10/16
${dateformat(yyyy_MM_dd HH:mm:ss)}	Returns the current time in **yyyy_MM_dd HH:mm:ss** format.	2017_10_16 09:00:00
${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}	Returns the current time in **yyyy-MM-dd HH:mm:ss** format. The date is one day before the current day.	2017-10-15 09:00:00

Macro Variable	Description	Display Effect
${timestamp()}	Returns the timestamp of the current time, that is, the number of milliseconds that have elapsed since 00:00:00 on January 1, 1970.	1508115600000
${timestamp(-10, MINUTE)}	Returns the timestamp generated 10 minutes before the current time point.	1508115000000
${timestamp(dateformat(yyyyMMdd))}	Returns the timestamp of 00:00:00 of the current day.	1508083200000
${timestamp(dateformat(yyyyMMdd,-1,DAY))}	Returns the timestamp of 00:00:00 of the previous day.	1507996800000
${timestamp(dateformat(yyyyMMddHH))}	Returns the timestamp of the current hour.	1508115600000

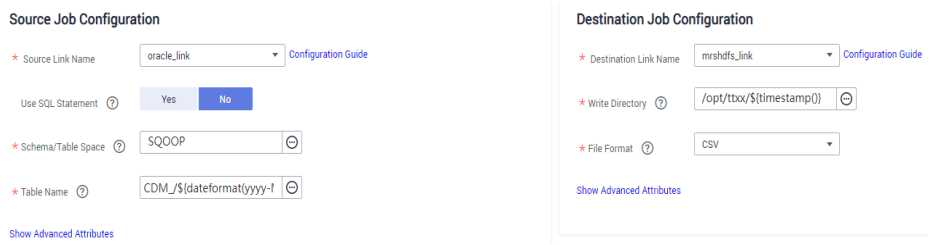
Time and Date Macro Variables of Paths and Table Names

Figure 7-6 shows an example. If:

- **Table Name** under **Source Link Configuration** is set to **CDM_/${dateformat(yyyy-MM-dd)}**.
- **Write Directory** under **Destination Link Configuration** is set to **/opt/ttxx/${timestamp()}**.

After the macro definition conversion, this job indicates that data in table **SQOOP.CDM_20171016** in the Oracle database is migrated to the **/opt/ttxx/1508115701746** directory of the HDFS server.

Figure 7-6 Setting **Table Name** and **Write Directory** to a time and date macro variable



Currently, a table name or path name can contain multiple macro variables. For example, **/opt/ttxx/${dateformat(yyyy-MM-dd)}/${timestamp()}** is converted to **/opt/ttxx/2017-10-16/1508115701746**.

Time and Date Macro Variables in the Where Clause

Figure 7-7 uses table **SQOOP.CDM_20171016** as an example. The table contains column **DS**, which indicates the time.

Figure 7-7 Table data

SELECT * FROM SQOOP.CDM_20171016

	FOO	BAR	DS
1	5	snap	2017-05-01
2	5	snap	2017-05-01
3	1	google	2017-05-02
4	4	oracle	2017-05-02
5	6	amd	2017-05-02
6	7	nvda	2017-05-02
7	1	google	2017-05-02
8	4	oracle	2017-05-02
9	6	amd	2017-05-02
10	7	nvda	2017-05-02
11	2	facebook	2017-10-15
12	3	tesla	2017-10-15
13	2	facebook	2017-10-15
14	3	tesla	2017-10-15

Suppose that the current date is **2017-10-16** and you want to export data generated the day before the current day (DS = 2017-10-15), then you can set the value of **Where Clause** to **DS='${dateformat(yyyy-MM-dd,-1,DAY)}'** when creating a job. In this way, you can export all data that complies with the DS = 2017-10-15 condition.

Implementing Incremental Synchronization by Configuring the Macro Variables of Date and Time and Scheduled Jobs

Two simple application scenarios are as follows:

- The database table contains column **DS** that indicates the time, the value type of the column is **varchar(30)**, and the inserted time format is similar to **2017-xx-xx**.

  In a scheduled job, the cycle is one day, and the scheduled job is executed at 00:00:00 every day. Set the value of **Where Clause** to **DS='${dateformat(yyyy-MM-dd,-1,DAY)}'**, and then data generated in the previous day will be exported at 00:00:00 every day.
- The database table contains column **time** that indicates the time, the type is **Number**, and the inserted time format is timestamp.

  In a scheduled job, the cycle is one day, and the scheduled job is executed at 00:00:00 every day. Set the value of **Where Clause** to **time between ${timestamp(-1,DAY)} and ${timestamp()}**, and then data generated on the previous day will be exported at 00:00:00 every day.

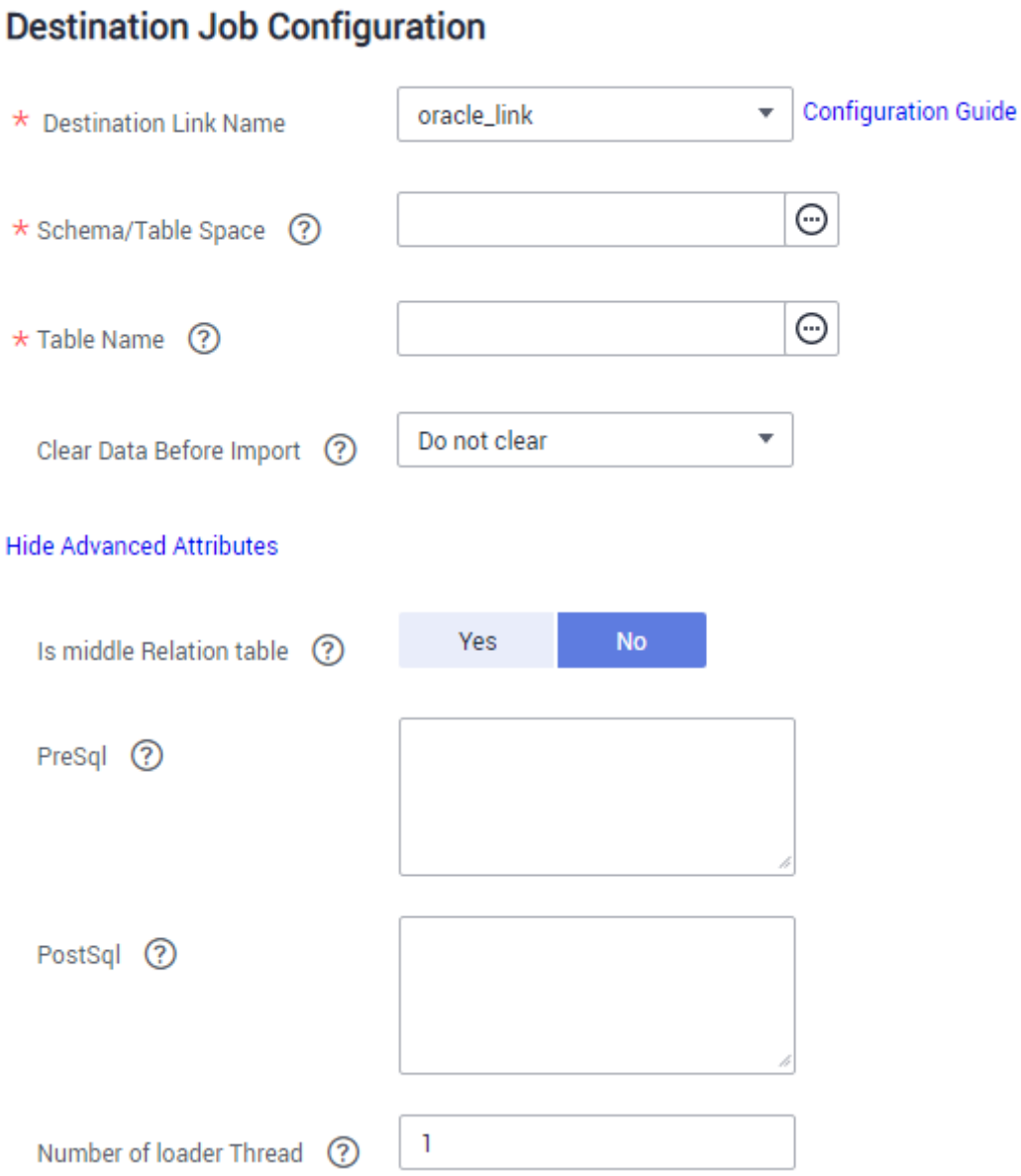
Configuration principles of other application scenarios are the same.

7.3 Migration in Transaction Mode

When a CDM job fails to be executed, CDM rolls back the data to the state before the job starts and automatically deletes data from the destination table.

- Parameter position: When creating a table/file migration job, if the migration source is a relational database, set **Import to Staging Table** in the advanced attributes of **Destination Job Configuration** to determine whether to enable the transaction mode.
- Parameter principle: If you set this parameter to **Yes**, CDM automatically creates a temporary table and imports the data to the temporary table. After the data is imported successfully, CDM migrates the data to the destination table in transaction mode of the database. If the import fails, the destination table is rolled back to the state before the job starts.

Figure 7-8 Migration in transaction mode



NOTE

If you select **Clear part of data** or **Clear all data** for **Clear Data Before Import**, CDM does not roll back the deleted data in transaction mode.

7.4 Encryption and Decryption During File Migration

When you migrate files to a file system, CDM can encrypt and decrypt those files. Currently, CDM supports the following encryption modes:

- **AES-256-GCM**
- **KMS Encryption**

AES-256-GCM

Currently, only AES-256-GCM (NoPadding) is supported. This algorithm is used for encryption at the migration destination and decryption at the migration source. The supported source and destination data sources are as follows:

- Data sources supported by the migration source: OBS, FTP, SFTP, HDFS (supported in the binary format), and HTTP (applicable to scenarios where OBS shared files are downloaded)
- Data sources supported by the migration destination: OBS, FTP, SFTP, and HDFS (supported in the binary format)

The following part describes how to use AES-256-GCM to decrypt the encrypted files to be exported from OBS and encrypt the files to be imported to OBS. The methods for using the algorithm on other data sources are the same.

- **Configure decryption at the migration source.**

  When you use CDM to create a job for exporting files from OBS, set the migration source to OBS and set the following parameters in the advanced settings of **Source Job Configuration**:

  a. **Encryption**: Select **AES-256-GCM**.

  b. **DEK**: The key must be the same as that configured in **Encryption**. Otherwise, the decrypted data is incorrect and the system does not display an error message.

  c. **IV**: The initialization vector must be the same as that configured in **Encryption**. Otherwise, the decrypted data is incorrect and the system does not display an error message.

  In this way, after CDM exports encrypted files from OBS, the files written to the migration destination are decrypted plaintext files.

- **Configure encryption at the migration destination.**

  When you use CDM to create a job for importing files to OBS, set the migration destination to OBS and set the following parameters in the advanced settings of **Destination Job Configuration**:

  a. **Encryption**: Select **AES-256-GCM**.

  b. **DEK**: custom encryption key. The key consists of 64 hexadecimal numbers. It is case-insensitive but must contain 64 characters. For example, **DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B**.

  c. **IV**: custom initialization vector. The initialization vector consists of 32 hexadecimal numbers. It is case-insensitive but must contain 32 characters. For example, **5C91687BA886EDCD12ACBC3FF19A3C3F**.

  In this way, after CDM imports files to OBS, the files on the migration destination are encrypted using the AES-256-GCM algorithm.

KMS Encryption

NOTE

The migration source does not support KMS encryption.

CDM supports KMS encryption if tables, files, or a whole database is migrated to OBS. In the **Advanced Attributes** area of the **Destination Job Configuration** page, set the parameters.

A key must be created in KMS of DEW in advance. For details, see the *Data Encryption Workshop User Guide*.

After KMS encryption is enabled, objects to be uploaded will be encrypted and stored on OBS. When you download the encrypted objects, the encrypted data will be decrypted on the server and displayed in plaintext to users.

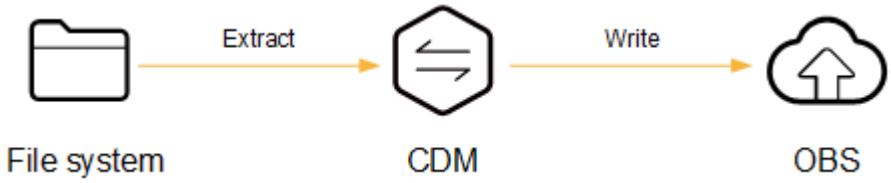
NOTE

- If KMS encryption is enabled, **MD5 verification** cannot be used.
- If the KMS ID of another project is used, change **Project ID** to the ID of the project to which KMS belongs. If KMS and CDM are in the same project, retain the default value of **Project ID**.
- After KMS encryption is performed, the encryption status of the objects on OBS cannot be changed.
- A key in use cannot be deleted. Otherwise, the object encrypted with this key cannot be downloaded.

7.5 MD5 Verification

CDM extracts data from the migration source and writes the data to the migration destination. **Figure 7-9** shows the migration mode when files are migrated to OBS.

Figure 7-9 Migrating files to OBS



During the process, CDM uses MD5 to verify file consistency.

- **Extract**
  - The migration source can be OBS, HDFS, FTP, SFTP, or HTTP. It can check whether the files extracted by CDM are consistent with source files.
  - This function is controlled by the **MD5 File Extension** parameter (available when **File Format** is set to **Binary**) in **Source Job Configuration**. Set this parameter to the file name extension of the MD5 file in the source file system.
  - If a source file **build.sh** and a file for saving MD5 value **build.sh.md5** are located in the same directory, and **MD5 File Extension** is configured, only the file **build.sh.md5** is migrated to the destination. Files without the MD5 value or whose MD5 values do not match fail to be migrated, and the MD5 file is not migrated.
  - If **MD5 File Extension** is not configured, all files are migrated.

- **Write**
  - Currently, this function can be used only when OBS serves as the migration destination. It can check whether the files written to OBS are consistent with those extracted from CDM.
  - This function is controlled by the **Validate MD5 Value** parameter in **Destination Job Configuration**. After the files are read and written to OBS, the MD5 value in the HTTP header is used to verify the files on OBS and the verification result is written to an OBS bucket (the bucket can be the one that does not store migration files). If the migration source does not have the MD5 file, the verification will not be performed.

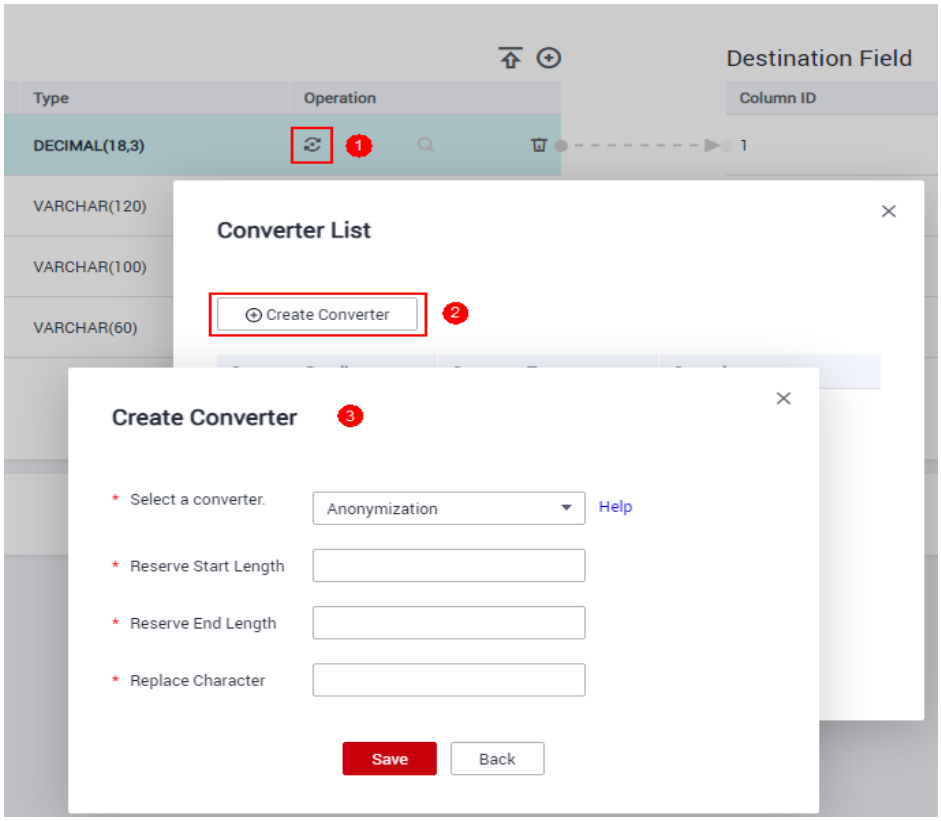
NOTE

- When files are migrated to a file system, only the extracted files are verified.
- When files are migrated to OBS, both the extracted files and files written to OBS are verified.
- If MD5 verification is used, **KMS encryption** cannot be used.

7.6 Field Conversion

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

Figure 7-10 Creating a field converter



NOTE

Field mapping is not involved when the binary format is used to migrate files to files.

CDM can convert fields during migration. Currently, the following field converters are supported:

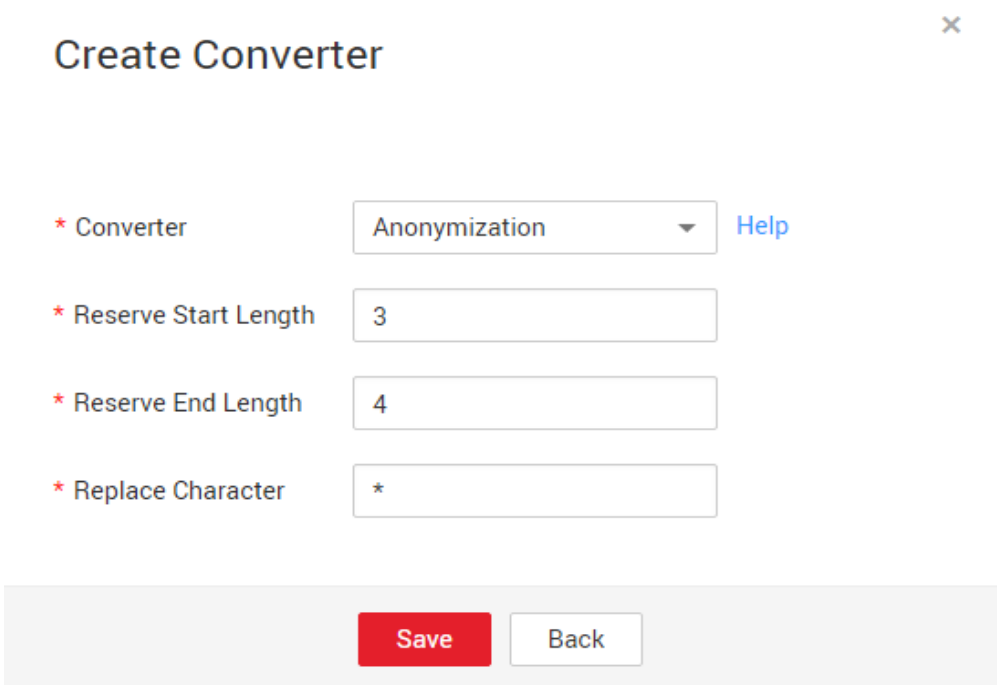
- **Anonymization**
- **Trim**
- **Reverse String**
- **Replace String**
- **Remove line break**
- **Expression Conversion**

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 7-11 Anonymization



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.

Remove line break

This converter is used to delete the newline characters, such as \n, \r, and \r\n from the field.

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 **yzyzybat** 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 **batyzyzy** 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*)

7.7 Migrating Files with Specified Names

You can migrate files (a maximum of 50) with specified names from FTP, SFTP, or OBS at a time. The exported files can only be written to the same directory on the migration destination.

When creating a table/file migration job, if the migration source is FTP, SFTP, or OBS, **Source Directory/File** can contain a maximum of 50 file names, which are separated by vertical bars (|). You can also customize a file separator.

NOTE

1. CDM supports incremental file migration (by skipping repeated files), but does not support resumable transfer.

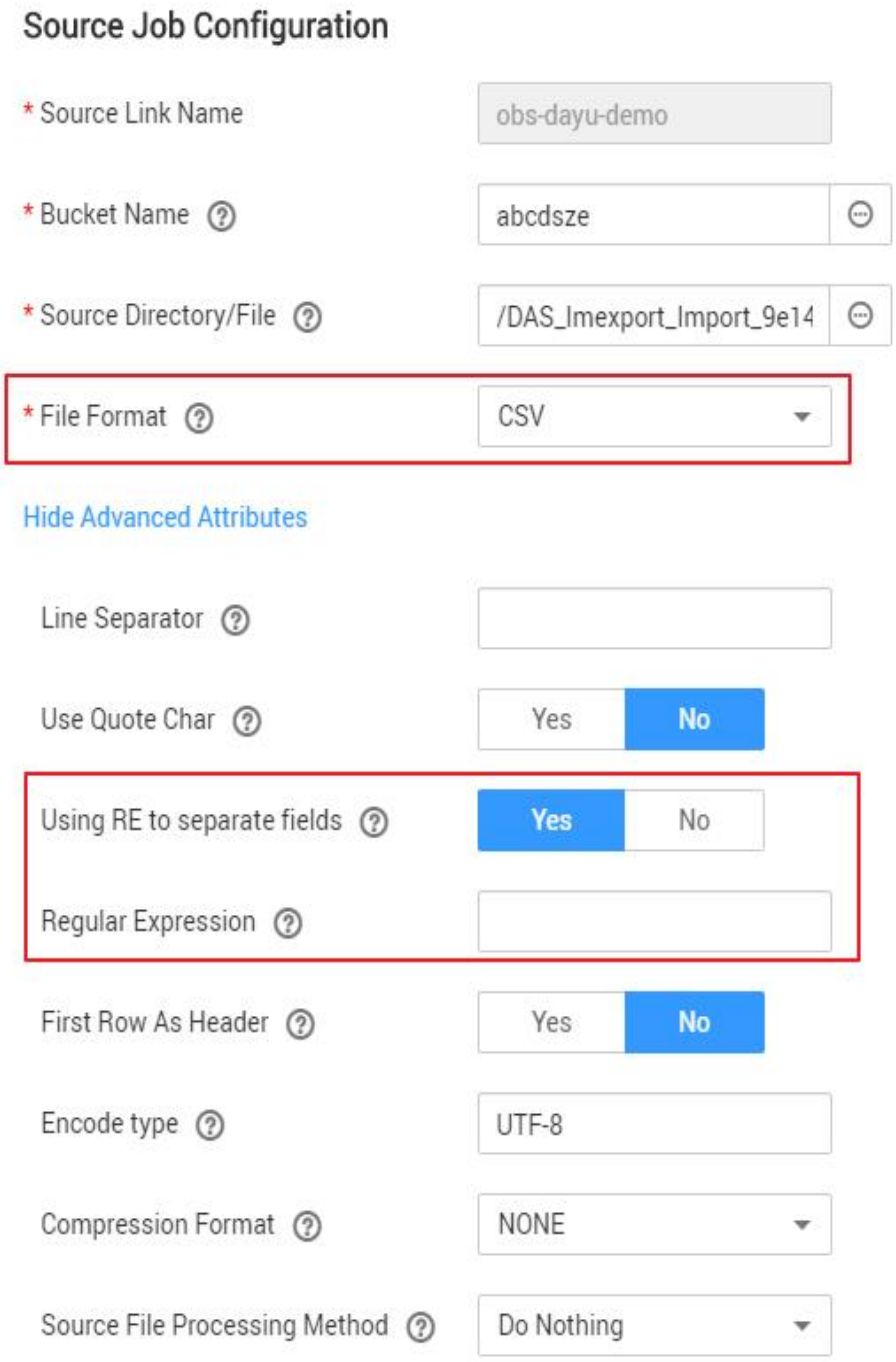
   For example, if three files are to be migrated and the second file fails to be migrated due to the network fault. When the migration task is started again, the first file is skipped. The second file, however, cannot be migrated from the point where the fault occurs, but can only be migrated again.
2. During file migration, a single task supports millions of files. If there are too many files in the directory to be migrated, you are advised to split the files into different directories and create multiple tasks.

7.8 Regular Expressions for Separating Semi-structured Text

During table/file migration, CDM uses delimiters to separate fields in CSV files. However, delimiters cannot be used in complex semi-structured data because the field values also contain delimiters. In this case, the regular expression can be used to separate the fields.

The regular expression is configured in **Source Job Configuration**. The migration source must be an object storage or file system, and **File Format** must be **CSV**.

Figure 7-12 Setting regular expression parameters



During the migration of CSV files, CDM can use regular expressions to separate fields and write parsed results to the migration destination. For details about the syntax of the regular expression, refer to the related documents. This section describes the regular expressions of the following log files:

- **Log4J Log**
- **Log4J Audit Log**
- **Tomcat Log**
- **Django Log**

- **Apache Server Log**

Log4J Log

- Log sample:
  ```
  2018-01-11 08:50:59,001 INFO
  [org.apache.sqoop.core.SqoopConfiguration.configureClassLoader(SqoopConfiguration.java:251)]
  Adding jars to current classloader from property: org.apache.sqoop.classpath.extra
  ```
- Regular expression:
  ```
  ^(\d.*\d) (\w*)  \[(.*)\] (\w.*).*
  ```
- Parsing result:

Table 7-2 Log4J log parsing result

Column Number	Example Value
1	2018-01-11 08:50:59,001
2	INFO
3	org.apache.sqoop.core.SqoopConfiguration.configureClassLoader(SqoopConfiguration.java:251)
4	Adding jars to current classloader from property: org.apache.sqoop.classpath.extra

Log4J Audit Log

- Log sample:
  ```
  2018-01-11 08:51:06,156 INFO
  [org.apache.sqoop.audit.FileAuditLogger.logAuditEvent(FileAuditLogger.java:61)]
  user=sqoop.anonymous.user    ip=189.xxx.xxx.75    op=show    obj=version    objId=x
  ```
- Regular expression:
  ```
  ^(\d.*\d) (\w*)  \[(.*)\] user=(\w.*)    ip=(\w.*)    op=(\w.*)    obj=(\w.*)    objId=(.*).*
  ```
- Parsing result:

Table 7-3 Log4J audit log parsing result

Column Number	Example Value
1	2018-01-11 08:51:06,156
2	INFO
3	org.apache.sqoop.audit.FileAuditLogger.logAuditEvent(FileAuditLogger.java:61)
4	sqoop.anonymous.user

Column Number	Example Value
5	189.xxx.xxx.75
6	show
7	version
8	x

Tomcat Log

- Log sample:
  ```
  11-Jan-2018 09:00:06.907 INFO [main] org.apache.catalina.startup.VersionLoggerListener.log OS Name:            Linux
  ```
- Regular expression:
  ```
  ^(\d.*\d) (\w*) \[(.*)\] ([\w\.]*) (\w.*).*
  ```
- Parsing result:

Table 7-4 Tomcat log parsing result

Column Number	Example Value
1	11-Jan-2018 09:00:06.907
2	INFO
3	main
4	org.apache.catalina.startup.VersionLoggerListener.log
5	OS Name:Linux

Django Log

- Log sample:
  ```
  [08/Jan/2018 20:59:07 ] settings     INFO     Welcome to Hue 3.9.0
  ```
- Regular expression:
  ```
  ^\[(.*)\] (\w*)     (\w*)     (.*).*
  ```
- Parsing result:

Table 7-5 Django log parsing result

Column Number	Example Value
1	08/Jan/2018 20:59:07
2	settings
3	INFO
4	Welcome to Hue 3.9.0

Apache Server Log

- Log sample:
  ```
  [Mon Jan 08 20:43:51.854334 2018] [mpm_event:notice] [pid 36465:tid 140557517657856] AH00489: Apache/2.4.12 (Unix) OpenSSL/1.0.1t configured -- resuming normal operations
  ```
- Regular expression:
  ```
  ^\[(.*)\] \[(.*)\] \[(.*)\] (.*).*
  ```
- Parsing result:

Table 7-6 Apache server log parsing result

Column Number	Example Value
1	Mon Jan 08 20:43:51.854334 2018
2	mpm_event:notice
3	pid 36465:tid 140557517657856
4	AH00489: Apache/2.4.12 (Unix) OpenSSL/1.0.1t configured -- resuming normal operations

7.9 Recording the Time When Data Is Written to the Database

When you create a job on the CDM console to migrate tables or files of a relational database, you can add a field to record the time when they were written to the database.

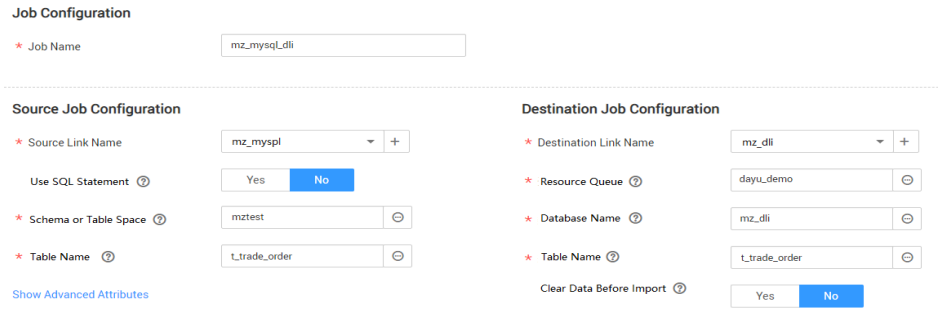
Prerequisites

A link has been created, and the source end of the connector is a relational database.

Creating a Table/File Migration Job

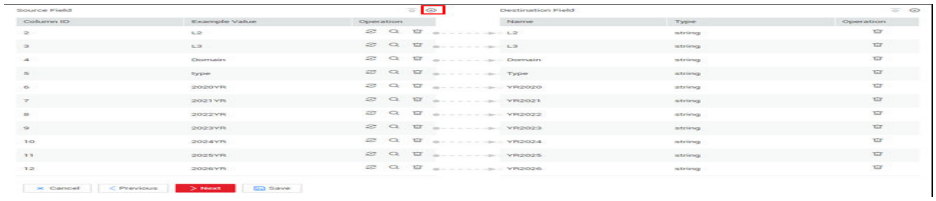
Step 1 Create a table/file migration job, and select the created source connector and destination connector.

Figure 7-13 Configuring the job



Step 2 Click **Next** to go to the **Map Field** page and click .

Figure 7-14 Configuring field mapping



Step 3 Click the **Custom Fields** tab, set the field name and value, and click **OK**.

Name: Enter **InputTime**.

Value: Enter **${timestamp()}**. For more time macro variables, see **Table 7-7**.

Figure 7-15 Add Field

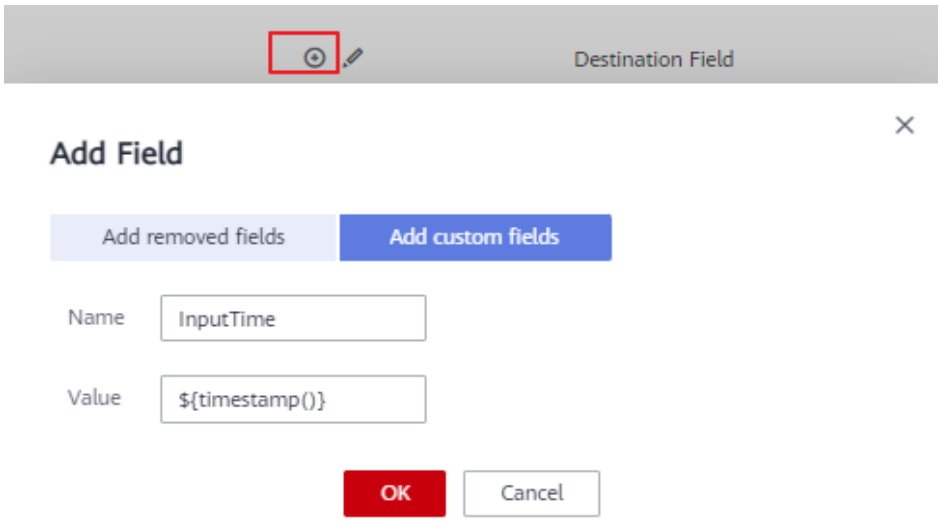


Table 7-7 Macro variable definition of time and date

Macro Variable	Description	Display Effect
${dateformat(yyyy-MM-dd)}	Returns the current date in **yyyy-MM-dd** format.	2017-10-16
${dateformat(yyyy/MM/dd)}	Returns the current date in **yyyy/MM/dd** format.	2017/10/16
${dateformat(yyyy_MM_dd HH:mm:ss)}	Returns the current time in **yyyy_MM_dd HH:mm:ss** format.	2017_10_16 09:00:00
${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}	Returns the current time in **yyyy-MM-dd HH:mm:ss** format. The date is one day before the current day.	2017-10-15 09:00:00
${timestamp()}	Returns the timestamp of the current time, that is, the number of milliseconds that have elapsed since 00:00:00 on January 1, 1970.	1508115600000
${timestamp(-10, MINUTE)}	Returns the timestamp generated 10 minutes before the current time point.	1508115000000
${timestamp(dateformat(yyyyMMdd))}	Returns the timestamp of 00:00:00 of the current day.	1508083200000
${timestamp(dateformat(yyyyMMdd,-1,DAY))}	Returns the timestamp of 00:00:00 of the previous day.	1507996800000
${timestamp(dateformat(yyyyMMddHH))}	Returns the timestamp of the current hour.	1508115600000

NOTE

- After a field is added, its sample value is not displayed on the console. This does not affect the field value transmission. CDM directly writes the field value to the destination end.
- The **Custom Fields** tab is available only when the source connector is JDBC, HBase, MongoDB, Elasticsearch, or Kafka, or the destination connector is HBase.

Step 4 Click **Next** and set task parameters. Generally, retain the default values of all parameters.

In this step, you can configure the following optional functions:

- **Retry Upon Failure**: If the job fails to be executed, you can determine whether to automatically retry. Retain the default value **Never**.
- **Group**: Select the group to which the job belongs. The default group is **DEFAULT**. On the **Job Management** page, jobs can be displayed, started, or exported by group.
- **Schedule Execution**: If you want the job to be automatically executed at a scheduled time, retain the default value **No**.
- **Concurrent Extractors**: Enter the number of extractors to be concurrently executed. Retain the default value **1**.
- **Write Dirty Data**: Specify this parameter if data that fails to be processed or filtered out during job execution needs to be written to OBS for future viewing. Before writing dirty data, create an OBS link. Retain the default value **No** so that dirty data is not recorded.
- **Delete Job After Completion**: Retain the default value **Do not delete**.

Step 5 Click **Save and Run**. On the **Table/File Migration** page, you can view the job execution progress and result.

Step 6 After the job is successfully executed, in the **Operation** column of the job, click **Historical Record** to view the job's historical execution records and read/write statistics.

On the **Historical Record** page, click **Log** to view the job log.

----End

7.10 File Formats

When creating a CDM job, you need to specify **File Format** in the job parameters of the migration source and destination in some scenarios. This section describes the application scenarios, subparameters, common parameters, and usage examples of the supported file formats.

- **CSV**
- **JSON**
- **Binary**
- **Common parameters**
- **Solutions to File Format Problems**

CSV

To read or write a CSV file, set **File Format** to **CSV**. The CSV format can be used in the following scenarios:

- Import files to a database or NoSQL.
- Export data from a database or NoSQL to files.

After selecting the CSV format, you can also configure the following optional sub-parameters:

1. Line Separator

2. Field Delimiter

3. Encoding Type

4. Use Quote Character

5. Use RE to Separate Fields

6. Use First Row as Header

7. File Size

1. **Line Separator**

   Character used to separate lines in a CSV file. The value can be a single character, multiple characters, or special characters. Special characters can be entered using the URL encoded characters. The following table lists the URL encoded characters of commonly used special characters.

   Table 7-8 URL encoded characters of special characters

Special Character	URL Encoded Character
Space	%20
Tab	%09
%	%25
Enter	%0d
Newline character	%0a
Start of heading\u0001 (SOH)	%01

2. **Field Delimiter**

   Character used to separate columns in a CSV file. The value can be a single character, multiple characters, or special characters. For details, see **Table 7-8**.

3. **Encoding Type**

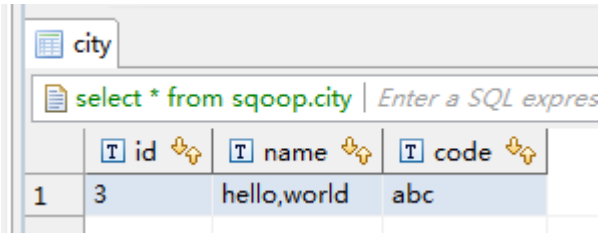
   Encoding type of a CSV file. The default value is **UTF-8**.

   If this parameter is specified at the migration source, the specified encoding type is used to parse the file. If this parameter is specified at the migration destination, the specified encoding type is used to write data to the file.

4. **Use Quote Character**

   - Exporting data from a database or NoSQL to CSV files (configuring **Use Quote Character** at the migration destination): If a field delimiter appears in the character string of a column of data at the migration source, set **Use Quote Character** to **Yes** at the migration destination to quote the character string as a whole and write it into the CSV file. Currently, CDM uses double quotation marks ("") as the quote character only. **Figure 7-16** shows that the value of the **name** field in the database contains a comma (,).

Figure 7-16 Field value containing the field delimiter



If you do not use the quote character, the exported CSV file is displayed as follows:

```
3.hello,world,abc
```

If you use the quote character, the exported CSV file is displayed as follows:

```
3,"hello,world",abc
```

If the data in the database contains double quotation marks ("") and you set **Use Quote Character** to **Yes**, the quote character in the exported CSV file is displayed as three double quotation marks ("""). For example, if the value of a field is **a"hello,world"c**, the exported data is as follows:

```
"""a"hello,world"c"""
```

- Exporting CSV files to a database or NoSQL (configuring **Use Quote Character** at the migration source): If you want to import the CSV files with quoted values to a database correctly, set **Use Quote Character** to **Yes** at the migration source to write the quoted values as a whole.

5. **Use RE to Separate Fields**

   This function is used to parse complex semi-structured text, such as log files. For details, see **Using Regular Expressions to Separate Semi-structured Text**.

6. **Use First Row as Header**

   This parameter is used when CSV files are exported to other locations. If this parameter is specified at the migration source, CDM uses the first row as the header when extracting data. When the CSV files are transferred, the headers are skipped. The number of rows extracted from the migration source is more than the number of rows written to the migration destination. The log files will output the information that the header is skipped during the migration.

7. **File Size**

   This parameter is used when data is exported from the database to a CSV file. If a table contains a large amount of data, a large CSV file is generated after migration, which is inconvenient to download or view. In this case, you can specify this parameter at the migration destination so that multiple CSV files with the specified size can be generated. The value of this parameter is an integer. The unit is MB.

JSON

The following describes information about the JSON format:

- **JSON Types Supported by CDM**
- **JSON Reference Node**

- **Copying Data from a JSON File**

1. **JSON types supported by CDM: JSON object and JSON array**
   - JSON object: A JSON file contains a single object or multiple objects separated/merged by rows.
     i. The following is a single JSON object:

```
{
    "took" : 190,
    "timed_out" : false,
    "total" : 1000001,
    "max_score" : 1.0
 }
```

     ii. The following are JSON objects separated by rows:

```
{"took" : 188, "timed_out" : false, "total" : 1000003, "max_score" : 1.0 }
{"took" : 189, "timed_out" : false, "total" : 1000004, "max_score" : 1.0 }
```

     iii. The following are merged JSON objects:

```
{
   "took": 190,
   "timed_out": false,
   "total": 1000001,
   "max_score": 1.0
 }
 {
   "took": 191,
   "timed_out": false,
   "total": 1000002,
   "max_score": 1.0
 }
```

   - JSON array: A JSON file is a JSON array consisting of multiple JSON objects.

```
[{
    "took" : 190,
    "timed_out" : false,
    "total" : 1000001,
    "max_score" : 1.0
 },
 {
    "took" : 191,
    "timed_out" : false,
    "total" : 1000001,
    "max_score" : 1.0
 }]
```

2. **JSON Reference Node**

   Root node that records data. The data corresponding to the node is a JSON array. CDM extracts data from the array in the same mode. Use periods (.) to separate multi-layer nested JSON nodes.

3. **Copying Data from a JSON File**

   a. Example 1: Extract data from multiple objects that are separated or merged. A JSON file contains multiple JSON objects. The following gives an example:

```
{
   "took": 190,
   "timed_out": false,
   "total": 1000001,
   "max_score": 1.0
}
{
   "took": 191,
   "timed_out": false,
   "total": 1000002,
   "max_score": 1.0
```



```
}
{
  "took": 192,
  "timed_out": false,
  "total": 1000003,
  "max_score": 1.0
}
```

To extract data from the JSON object and write data to the database in the following formats, set **File Format** to **JSON** and **JSON Type** to **JSON object**, and then map fields.

took	timedOut	total	maxScore
190	false	1000001	1.0
191	false	1000002	1.0
192	false	1000003	1.0

b. Example 2: Extract data from the reference node. A JSON file contains a single JSON object, but the valid data is on a data node. The following gives an example:

```
{
  "took": 190,
  "timed_out": false,
  "hits": {
     "total": 1000001,
     "max_score": 1.0,
     "hits":
      [{
        "_id": "650612",
        "_source": {
           "name": "tom",
           "books": ["book1","book2","book3"]
        }
      },
      {
         "_id": "650616",
         "_source": {
            "name": "tom",
            "books": ["book1","book2","book3"]
         }
      },
    {
         "_id": "650618",
         "_source": {
            "name": "tom",
            "books": ["book1","book2","book3"]
         }
      }]
  }
}
```

To write data to the database in the following formats, set **File Format** to **JSON**, **JSON Type** to **JSON object**, and **JSON Reference Node** to **hits.hits**, and then map fields.

ID	SourceName	SourceBooks
650612	tom	["book1","book2","book3"]
650616	tom	["book1","book2","book3"]

ID	SourceName	SourceBooks
650618	tom	["book1","book2","book3"]

c. Example 3: Extract data from the JSON array. A JSON file is a JSON array consisting of multiple JSON objects. The following gives an example:

```
[{
    "took" : 190,
    "timed_out" : false,
    "total" : 1000001,
    "max_score" : 1.0
 },
 {
    "took" : 191,
    "timed_out" : false,
    "total" : 1000002,
    "max_score" : 1.0
 }]
```

To write data to the database in the following formats, set **File Format** to **JSON** and **JSON Type** to **JSON array**, and then map fields.

took	timedOut	total	maxScore
190	false	1000001	1.0
191	false	1000002	1.0

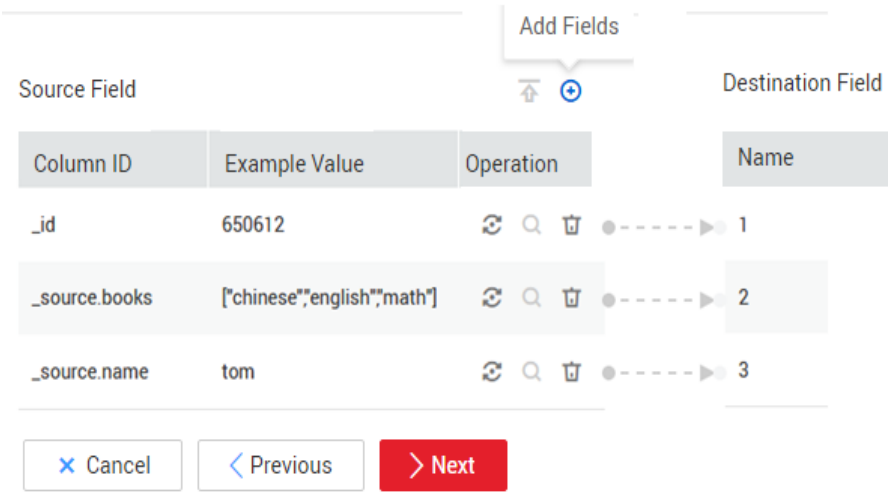
d. Example 4: Configure a converter when parsing the JSON file. On the premise of **example 2**, to add the **hits.max_score** field to all records, that is, to write the data to the database in the following formats, perform the following operations:

ID	SourceName	SourceBooks	MaxScore
650612	tom	["book1","book2","book3"]	1.0
650616	tom	["book1","book2","book3"]	1.0
650618	tom	["book1","book2","book3"]	1.0

Set **File Format** to **JSON**, **JSON Type** to **JSON object**, and **JSON Reference Node** to **hits.hits**, and then create a converter.

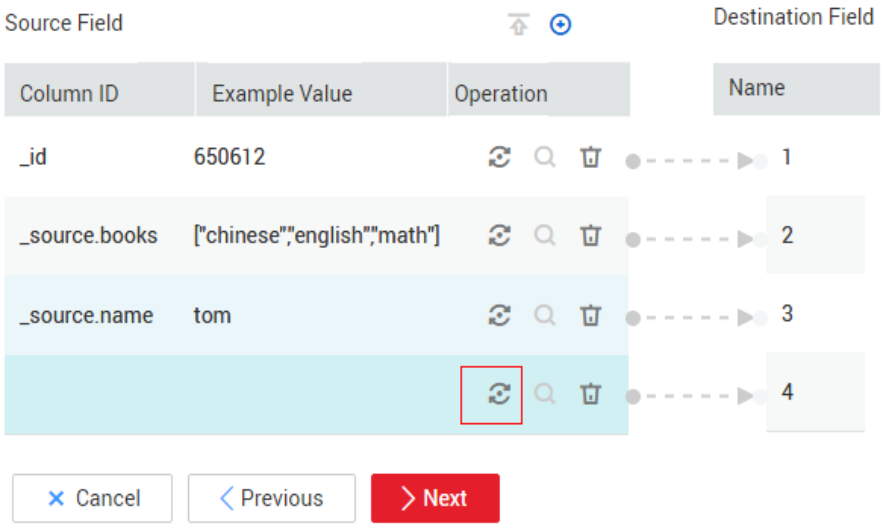
i. Click  to add a field.

Figure 7-17 Adding a field



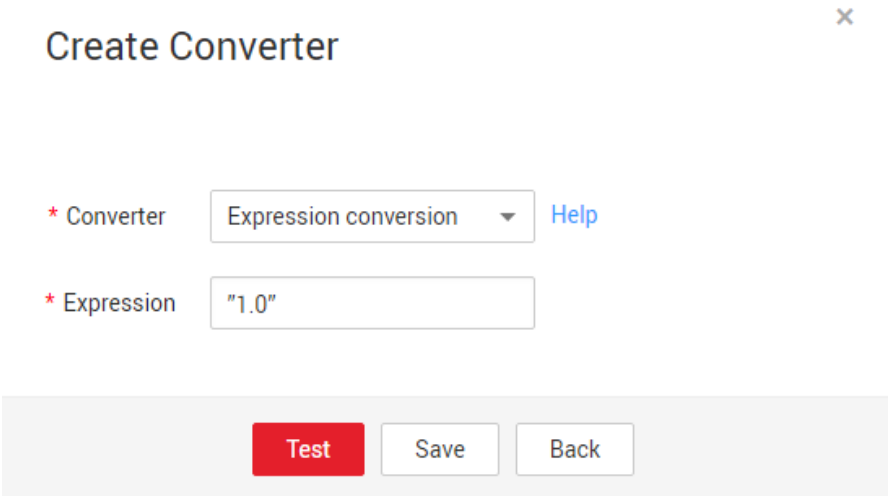
ii. Click to create a converter for the new field.

Figure 7-18 Creating a field converter



iii. Set **Converter** to **Expression conversion**, enter **"1.0"** in the **Expression** text box, and click **Save**.

Figure 7-19 Configuring a field converter



Binary

If you want to copy files between file systems, you can select the binary format. The binary format delivers the optimal rate and performance in file transfer, and does not require field mapping.

- **Directory structure for file transfer**

  CDM can transfer a single file or all files in a directory at a time. After the files are transferred to the migration destination, the directory structure remains unchanged.

- **Migrating incremental files**

  When you use CDM to transfer files in binary format, configure **Duplicate File Processing Method** at the migration destination for incremental file migration. For details, see **Incremental File Migration**.

  During incremental file migration, set **Duplicate File Processing Method** to **Skip**. If new files exist at the migration source or a failure occurs during the migration, run the job again, so that the migrated files will not be migrated repeatedly.

- **Write to Temporary File**

  When migrating files in binary format, you can specify whether to write the files to a temporary file at the migration destination. If this parameter is specified, the file is written to a temporary file during file replication. After the file is successfully migrated, run the **rename** or **move** command to restore the file at the migration destination.

- **Generate MD5 Hash Value**

  An MD5 hash value is generated for each transferred file, and the value is recorded in a new **.md5** file. You can specify the directory where the MD5 value is generated.

Common parameters

- **Source File Processing Method**

After a file is copied successfully, CDM can perform operations on the source file, including renaming the file, deleting the file, and performing no operation on the file.

- **Start Job by Marker File**

  In automation scenarios, a scheduled task is configured on CDM to periodically read files from the migration source. However, files are being generated at the migration source. As a result, CDM reads data repeatedly or fails to read data from the migration source. You can specify the marker file for starting a job as **ok.txt** in the job parameters of the migration source. After the file is successfully generated at the migration source, the **ok.txt** file is generated in the file directory. In this way, CDM can read the complete file.

  In addition, you can set the suspension period. Within the suspension period, CDM periodically queries whether the marker file exists. If the file does not exist after the suspension period expires, the job fails.

  The marker file will not be migrated.

- **Job Success Marker File**

  After data is successfully migrated to a file system, an empty file is generated in the destination directory. You can specify the file name. Generally, this parameter is used together with **Start Job by Marker File**.

  Note that the file cannot be confused with the file to be transferred. For example, if the file to be transferred is **finish.txt** and the job success marker file is set to **finish.txt**, the two files will overwrite each other.

- **Filter**

  When using CDM to migrate files, you can specify a filter to filter files. Files can be filtered by wildcard character or time filter.

  - If you select **Wildcard**, CDM migrates only the paths or files that meet the filter condition.
  - If you select **Time Filter**, CDM migrates only the files modified after the specified time point.

  For example, the **/table/** directory stores a large number of data table directories divided by day. **DRIVING_BEHAVIOR_20180101** to **DRIVING_BEHAVIOR_20180630** store all data of **DRIVING_BEHAVIOR** from January to June. To migrate only the table data of **DRIVING_BEHAVIOR** in March, set **Source Directory/File** to **/table**, **Filter Type** to **Wildcard**, and **Path Filter** to **DRIVING_BEHAVIOR_201803***.

Solutions to File Format Problems

1. When data in a database is exported to a CSV file, if the data contains commas (,), the data in the exported CSV file is disordered.

   The following solutions are available:

   a. Specify a field delimiter.

      Use a character that does not exist in the database or a rare non-printable character as the field delimiter. For example, set **Field Delimiter** at the migration destination to **%01**. In this way, the exported field delimiter is **\u0001**. For details, see **Table 7-8**.

   b. Use the quote character.

      Set **Use Quote Character** to **Yes** at the migration destination. In this way, if the field in the database contains the field delimiter, CDM quotes the

field using the quote character and write the field as a whole to the CSV file.

2. The data in the database contains line separators.

   Scenario: When you use CDM to export a table in the MySQL database (a field value contains the line separator **\n**) to a CSV file, and then use CDM to import the exported CSV file to MRS HBase, data in the exported CSV file is truncated.

   Solution: Specify a line separator.

   When you use CDM to export MySQL table data to a CSV file, set **Line Separator** at the migration destination to **%01** (ensure that the value does not appear in the field value). In this way, the line separator in the exported CSV file is **%01**. Then use CDM to import the CSV file to MRS HBase. Set **Line Separator** at the migration source to **%01**. This avoids data truncation.

