

**DataArts Studio**

# **User Guide**

**Date**      **2023-06-14**

---

# Contents

---

<b>1 Service Overview.....</b>	<b>1</b>
1.1 What Is DataArts Studio?.....	1
1.2 Basic Concepts.....	3
1.3 Functions.....	5
1.4 Advantages.....	7
1.5 Application Scenarios.....	8
1.6 DataArts Studio Permissions Management.....	8
1.7 DataArts Studio Permissions.....	10
1.8 Constraints.....	20
1.9 Related Services.....	21
<b>2 Preparations.....</b>	<b>23</b>
2.1 Preparations.....	23
2.2 Creating DataArts Studio Instances.....	23
2.2.1 Creating a DataArts Studio Basic Package.....	23
2.2.2 (Optional) Creating a DataArts Studio Incremental Package.....	25
2.3 Managing a Workspace.....	28
2.3.1 Creating and Managing a Workspace.....	28
2.3.2 (Optional) Changing the Job Log Storage Path.....	31
2.4 Authorizing Users to Use DataArts Studio.....	32
2.4.1 Creating an IAM User and Assigning DataArts Studio Permissions.....	33
2.4.2 Adding a Member and Assigning a Role.....	33
2.5 (Optional) Obtaining Authentication Information.....	35
<b>3 User Guide.....</b>	<b>37</b>
3.1 Preparations Before Using DataArts Studio.....	37
3.2 Management Center.....	38
3.2.1 Data Sources.....	39
3.2.2 Creating Data Connections.....	42
3.2.3 Migrating Resources.....	63
3.2.4 Tutorials.....	67
3.2.4.1 Creating an MRS Hive Connection.....	68
3.2.4.2 Creating a DWS Connection.....	73
3.2.4.3 Creating a MySQL Connection.....	79

3.3 DataArts Migration.....	83
3.3.1 Overview.....	83
3.3.2 Constraints.....	86
3.3.3 Supported Data Sources.....	91
3.3.4 Managing Clusters.....	116
3.3.4.1 Creating a CDM Cluster.....	116
3.3.4.2 Binding or Unbinding an EIP.....	116
3.3.4.3 Restarting a Cluster.....	117
3.3.4.4 Deleting a Cluster.....	118
3.3.4.5 Downloading Cluster Logs.....	119
3.3.4.6 Viewing Basic Cluster Information and Modifying Cluster Configurations.....	120
3.3.4.7 Viewing Metrics.....	123
3.3.4.7.1 CDM Metrics.....	123
3.3.4.7.2 Configuring Alarm Rules.....	125
3.3.4.7.3 Querying Metrics.....	126
3.3.5 Managing Links.....	127
3.3.5.1 Creating Links.....	127
3.3.5.2 Managing Drivers.....	132
3.3.5.3 Managing Agents.....	133
3.3.5.4 Managing Cluster Configurations.....	137
3.3.5.5 Link to a Common Relational Database.....	143
3.3.5.6 Link to a Database Shard.....	145
3.3.5.7 Link to MyCAT.....	146
3.3.5.8 Link to a Dameng Database.....	148
3.3.5.9 Link to a MySQL Database.....	149
3.3.5.10 Link to an Oracle Database.....	151
3.3.5.11 Link to DLI.....	153
3.3.5.12 Link to Hive.....	154
3.3.5.13 Link to HBase.....	161
3.3.5.14 Link to HDFS.....	168
3.3.5.15 Link to OBS.....	175
3.3.5.16 Link to an FTP or SFTP Server.....	175
3.3.5.17 Link to Redis/DCS.....	176
3.3.5.18 Link to DDS.....	177
3.3.5.19 Link to CloudTable.....	177
3.3.5.20 Link to CloudTable OpenTSDB.....	178
3.3.5.21 Link to MongoDB.....	179
3.3.5.22 Link to Cassandra.....	180
3.3.5.23 Link to Kafka.....	181
3.3.5.24 Link to DMS Kafka.....	182
3.3.5.25 Link to Elasticsearch/CSS.....	183
3.3.6 Managing Jobs.....	184

3.3.6.1 Table/File Migration Jobs.....	184
3.3.6.2 Creating an Entire Database Migration Job.....	194
3.3.6.3 Source Job Parameters.....	199
3.3.6.3.1 From OBS.....	200
3.3.6.3.2 From HDFS.....	206
3.3.6.3.3 From HBase/CloudTable.....	212
3.3.6.3.4 From Hive.....	213
3.3.6.3.5 From DLI.....	215
3.3.6.3.6 From FTP/SFTP.....	216
3.3.6.3.7 From HTTP.....	221
3.3.6.3.8 From a Common Relational Database.....	224
3.3.6.3.9 From MySQL.....	228
3.3.6.3.10 From Oracle.....	232
3.3.6.3.11 From a Database Shard.....	236
3.3.6.3.12 From MongoDB/DDS.....	238
3.3.6.3.13 From Redis.....	239
3.3.6.3.14 From Kafka/DMS Kafka.....	240
3.3.6.3.15 From Elasticsearch or CSS.....	241
3.3.6.3.16 From OpenTSDB.....	243
3.3.6.4 Destination Job Parameters.....	244
3.3.6.4.1 To OBS.....	244
3.3.6.4.2 To HDFS.....	249
3.3.6.4.3 To HBase/CloudTable.....	252
3.3.6.4.4 To Hive.....	254
3.3.6.4.5 To a Common Relational Database.....	256
3.3.6.4.6 To DWS.....	259
3.3.6.4.7 To DDS.....	264
3.3.6.4.8 To DCS.....	264
3.3.6.4.9 To CSS.....	264
3.3.6.4.10 To DLI.....	265
3.3.6.4.11 To OpenTSDB.....	266
3.3.6.5 Scheduling Job Execution.....	267
3.3.6.6 Job Configuration Management.....	269
3.3.6.7 Managing a Single Job.....	271
3.3.6.8 Managing Jobs in Batches.....	273
3.3.7 Auditing.....	275
3.3.7.1 Key CDM Operations Recorded by CTS.....	275
3.3.7.2 Viewing Traces.....	276
3.3.8 Tutorials.....	276
3.3.8.1 Creating an MRS Hive Link.....	276
3.3.8.2 Creating a MySQL Link.....	281
3.3.8.3 Migrating Data from MySQL to MRS Hive.....	285



3.3.8.4 Migrating Data from MySQL to OBS.....	294
3.3.8.5 Migrating Data from MySQL to DWS.....	298
3.3.8.6 Migrating an Entire MySQL Database to RDS.....	303
3.3.8.7 Migrating Data from Oracle to CSS.....	307
3.3.8.8 Migrating Data from Oracle to DWS.....	310
3.3.8.9 Migrating Data from OBS to CSS.....	318
3.3.8.10 Migrating Data from OBS to DLI.....	321
3.3.8.11 Migrating Data from MRS HDFS to OBS.....	325
3.3.8.12 Migrating the Entire Elasticsearch Database to CSS.....	329
3.3.9 Advanced Operations.....	332
3.3.9.1 Incremental Migration.....	332
3.3.9.1.1 Incremental File Migration.....	332
3.3.9.1.2 Incremental Migration of Relational Databases.....	334
3.3.9.1.3 Using Macro Variables of Date and Time.....	336
3.3.9.1.4 HBase/CloudTable Incremental Migration.....	340
3.3.9.2 Migration in Transaction Mode.....	341
3.3.9.3 Encryption and Decryption During File Migration.....	342
3.3.9.4 MD5 Verification.....	343
3.3.9.5 Field Conversion.....	344
3.3.9.6 Migrating Files with Specified Names.....	352
3.3.9.7 Regular Expressions for Separating Semi-structured Text.....	353
3.3.9.8 Recording the Time When Data Is Written to the Database.....	357
3.3.9.9 File Formats.....	360
3.4 DataArts Factory.....	369
3.4.1 Overview.....	369
3.4.2 Data Management.....	371
3.4.2.1 Data Management Process.....	371
3.4.2.2 Creating a Data Connection.....	372
3.4.2.3 Creating a Database.....	373
3.4.2.4 (Optional) Creating a Database Schema.....	376
3.4.2.5 Creating a Table.....	377
3.4.3 Script Development.....	385
3.4.3.1 Script Development Process.....	385
3.4.3.2 Creating a Script.....	386
3.4.3.3 Developing Scripts.....	388
3.4.3.3.1 Developing an SQL Script.....	388
3.4.3.3.2 Developing a Shell Script.....	394
3.4.3.3.3 Developing a Python Script.....	399
3.4.3.4 Submitting a Version and Unlocking the Script.....	401
3.4.3.5 (Optional) Managing Scripts.....	407
3.4.3.5.1 Copying a Script.....	407
3.4.3.5.2 Copying the Script Name and Renaming a Script.....	408

3.4.3.5.3 Moving a Script or Script Directory.....	410
3.4.3.5.4 Exporting and Importing a Script.....	411
3.4.3.5.5 Viewing Script References.....	413
3.4.3.5.6 Deleting a Script.....	413
3.4.3.5.7 Changing the Script Owner.....	415
3.4.3.5.8 Unlocking Scripts.....	416
3.4.4 Job Development.....	418
3.4.4.1 Job Development Process.....	418
3.4.4.2 Creating a Job.....	419
3.4.4.3 Developing a Job.....	423
3.4.4.4 Setting Up Scheduling for a Job.....	429
3.4.4.5 Submitting a Version and Unlocking the Script.....	434
3.4.4.6 (Optional) Managing Jobs.....	440
3.4.4.6.1 Copying a Job.....	440
3.4.4.6.2 Copying the Job Name and Renaming a Job.....	441
3.4.4.6.3 Moving a Job or Job Directory.....	443
3.4.4.6.4 Exporting and Importing a Job.....	444
3.4.4.6.5 Deleting a Job.....	448
3.4.4.6.6 Changing the Job Owner.....	449
3.4.4.6.7 Unlocking Jobs.....	450
3.4.5 Solution.....	452
3.4.6 Execution History.....	454
3.4.7 O&M and Scheduling.....	455
3.4.7.1 Overview.....	455
3.4.7.2 Monitoring a Job.....	456
3.4.7.2.1 Monitoring a Batch Job.....	456
3.4.7.2.2 Monitoring a Real-Time Job.....	461
3.4.7.3 Monitoring an Instance.....	467
3.4.7.4 Monitoring PatchData.....	471
3.4.7.5 Managing Notifications.....	471
3.4.7.5.1 Managing a Notification.....	471
3.4.7.5.2 Cycle Overview.....	475
3.4.7.6 Managing Backups.....	476
3.4.8 Configuration and Management.....	479
3.4.8.1 Configuring Resources.....	479
3.4.8.1.1 Configuring Environment Variables.....	479
3.4.8.1.2 Configuring an OBS Bucket.....	482
3.4.8.1.3 Managing Job Labels.....	483
3.4.8.1.4 Configuring Agencies.....	484
3.4.8.1.5 Configuring a Default Item.....	492
3.4.8.2 Managing Resources.....	494
3.4.9 Node Reference.....	500

3.4.9.1 Node Overview.....	500
3.4.9.2 CDM Job.....	501
3.4.9.3 Rest Client.....	507
3.4.9.4 Import GES.....	514
3.4.9.5 MRS Kafka.....	516
3.4.9.6 Kafka Client.....	518
3.4.9.7 ROMA FDI Job.....	520
3.4.9.8 DLI Flink Job.....	522
3.4.9.9 DLI SQL.....	526
3.4.9.10 DLI Spark.....	532
3.4.9.11 DWS SQL.....	539
3.4.9.12 MRS Spark SQL.....	544
3.4.9.13 MRS Hive SQL.....	549
3.4.9.14 MRS Presto SQL.....	554
3.4.9.15 MRS Spark.....	559
3.4.9.16 MRS Spark Python.....	564
3.4.9.17 MRS Flink Job.....	569
3.4.9.18 MRS MapReduce.....	571
3.4.9.19 CSS.....	573
3.4.9.20 Shell.....	575
3.4.9.21 RDS SQL.....	578
3.4.9.22 ETL Job.....	579
3.4.9.23 Python.....	584
3.4.9.24 Create OBS.....	586
3.4.9.25 Delete OBS.....	588
3.4.9.26 OBS Manager.....	590
3.4.9.27 Open/Close Resource.....	595
3.4.9.28 Subjob.....	597
3.4.9.29 For Each.....	598
3.4.9.30 SMN.....	601
3.4.9.31 Dummy.....	603
3.4.10 EL Expression Reference.....	604
3.4.10.1 Expression Overview.....	604
3.4.10.2 Basic Operators.....	607
3.4.10.3 Date and Time Mode.....	608
3.4.10.4 Env Embedded Objects.....	610
3.4.10.5 Job Embedded Objects.....	610
3.4.10.6 StringUtil Embedded Objects.....	612
3.4.10.7 DateUtil Embedded Objects.....	612
3.4.10.8 JSONUtil Embedded Objects.....	613
3.4.10.9 Loop Embedded Objects.....	614
3.4.10.10 OBSUtil Embedded Objects.....	615

3.4.10.11 Expression Use Example.....	615
3.4.11 Usage Guidance.....	617
3.4.11.1 Job Dependency.....	617
3.4.11.2 IF Statements.....	623
3.4.11.3 Obtaining the Return Value of a Rest Client Node.....	633
3.4.11.4 Using For Each Nodes.....	635
3.4.11.5 Developing a Python Script.....	642
3.4.11.6 Developing a DWS SQL Job.....	645
3.4.11.7 Developing a Hive SQL Job.....	649
3.4.11.8 Developing a DLI Spark Job.....	652
3.4.11.9 Developing an MRS Flink Job.....	656
3.4.11.10 Developing an MRS Spark Python Job.....	658
<b>4 FAQs.....</b>	<b>665</b>
4.1 Consultation.....	665
4.1.1 Regions.....	665
4.1.2 What Should I Do If a User Cannot View Existing Workspaces After I Have Assigned the Required Policy to the User?.....	665
4.1.3 Can I Delete DataArts Studio Workspaces?.....	666
4.1.4 Can I Transfer a Trial Instance to Another Account?.....	666
4.1.5 Does DataArts Studio Support Version Downgrade?.....	666
4.2 Management Center.....	666
4.2.1 What Are the Precautions for Creating Data Connections?.....	666
4.2.2 Why Do DWS/Hive/HBase Data Connections Fail to Obtain the Information About Database or Tables?.....	666
4.2.3 Why Are MRS Hive/HBase Clusters Not Displayed on the Page for Creating Data Connections?.....	667
4.2.4 What Should I Do If the Connection Test Fails When I Enable the SSL Connection During the Creation of a DWS Data Connection?.....	667
4.2.5 Can I Create Multiple Data Connections in a Workspace in Proxy Mode?.....	668
4.2.6 Should I Choose a Direct or a Proxy Connection When Creating a DWS Connection?.....	668
4.2.7 How Do I Migrate the Data Development Jobs and Data Connections from One Workspace to Another?.....	668
4.2.8 Can I Delete Workspaces?.....	668
4.3 DataArts Migration.....	668
4.3.1 General.....	668
4.3.1.1 What Are the Advantages of CDM?.....	668
4.3.1.2 What Are the Security Protection Mechanisms of CDM?.....	670
4.3.1.3 How Do I Reduce the Cost of Using CDM?.....	670
4.3.1.4 Can I Upgrade a CDM Cluster?.....	671
4.3.1.5 How Is the Migration Performance of CDM?.....	671
4.3.1.6 What Is the Number of Concurrent Jobs for Different CDM Cluster Versions?.....	671
4.3.2 Functions.....	672
4.3.2.1 Does CDM Support Incremental Data Migration?.....	672
4.3.2.2 Does CDM Support Field Conversion?.....	672

4.3.2.3 What Component Versions Are Recommended for Migrating Hadoop Data Sources?.....	679
4.3.2.4 What Data Formats Are Supported When the Data Source Is Hive?.....	680
4.3.2.5 Can I Synchronize Jobs to Other Clusters?.....	680
4.3.2.6 Can I Create Jobs in Batches?.....	680
4.3.2.7 Can I Schedule Jobs in Batches?.....	681
4.3.2.8 How Do I Back Up CDM Jobs?.....	681
4.3.2.9 How Do I Configure the Connection If Only Some Nodes in the HANA Cluster Can Communicate with the CDM Cluster?.....	681
4.3.2.10 How Do I Use Java to Invoke CDM RESTful APIs to Create Data Migration Jobs?.....	681
4.3.2.11 How Do I Connect the On-Premises Intranet or Third-Party Private Network to CDM?.....	687
4.3.2.12 How Do I Set the Number of Concurrent Extractors for a CDM Migration Job?.....	689
4.3.2.13 Does CDM Support Real-Time Migration of Dynamic Data?.....	689
4.3.3 Troubleshooting.....	690
4.3.3.1 What Can I Do If Error Message "Unable to execute the SQL statement" Is Displayed When I Import Data from OBS to SQL Server?.....	690
4.3.3.2 Why Is Error ORA-01555 Reported During Migration from Oracle to DWS?.....	690
4.3.3.3 What Should I Do If the MongoDB Connection Migration Fails?.....	691
4.3.3.4 What Should I Do If a Hive Migration Job Is Suspended for a Long Period of Time?.....	691
4.3.3.5 What Should I Do If an Error Is Reported Because the Field Type Mapping Does Not Match During Data Migration Using CDM?.....	691
4.3.3.6 What Should I Do If a JDBC Connection Timeout Error Is Reported During MySQL Migration?....	692
4.3.3.7 What Should I Do If a CDM Migration Job Fails After a Link from Hive to DWS Is Created?.....	693
4.3.3.8 How Do I Use CDM to Export MySQL Data to an SQL File and Upload the File to an OBS Bucket? .....	693
4.3.3.9 What Should I Do If CDM Fails to Migrate Data from OBS to DLI?.....	694
4.3.3.10 What Should I Do If a CDM Connector Reports the Error "Configuration Item [linkConfig.iamAuth] Does Not Exist"?.....	694
4.3.3.11 What Should I Do If Error Message "Configuration Item [linkConfig.createBackendLinks] Does Not Exist" Is Displayed During Data Link Creation or Error Message "Configuration Item [throttlingConfig.concurrentSubJobs] Does Not Exist" Is Displayed During Job Creation?.....	694
4.3.3.12 What Should I Do If Message "CORE_0031:Connect time out. (Cdm.0523)" Is Displayed During the Creation of an MRS Hive Link?.....	694
4.3.3.13 What Should I Do If Message "CDM Does Not Support Auto Creation of an Empty Table with No Column" Is Displayed When I Enable Auto Table Creation?.....	694
4.3.3.14 What Should I Do If I Cannot Obtain the Schema Name When Creating an Oracle Relational Database Migration Job?.....	695
4.4 DataArts Factory.....	695
4.4.1 How Many Jobs Can Be Created in DataArts Factory? Is There a Limit on the Number of Nodes in a Job?.....	695
4.4.2 Why Is There a Large Difference Between Job Execution Time and Start Time of a Job?.....	695
4.4.3 Will Subsequent Jobs Be Affected If a Job Fails to Be Executed During Scheduling of Dependent Jobs? What Should I Do?.....	695
4.4.4 What Should I Pay Attention to When Using DataArts Studio to Schedule Big Data Services?.....	696
4.4.5 What Are the Differences and Connections Among Environment Variables, Job Parameters, and Script Parameters?.....	696
4.4.6 What Do I Do If Node Error Logs Cannot Be Viewed When a Job Fails?.....	698

4.4.7 What Should I Do If the Agency List Fails to Be Obtained During Agency Configuration?.....	698
4.4.8 How Do I Locate Job Scheduling Nodes with a Large Number?.....	699
4.4.9 Why Cannot Specified Peripheral Resources Be Selected When a Data Connection Is Created in Data Development?.....	700
4.4.10 Why Is There No Job Running Scheduling Log on the Monitor Instance Page After Periodic Scheduling Is Configured for a Job?.....	700
4.4.11 Why Does the GUI Display Only the Failure Result but Not the Specific Error Cause After Hive SQL and Spark SQL Scripts Fail to Be Executed?.....	701
4.4.12 What Do I Do If the Token Is Invalid During the Running of a Data Development Node?.....	701
4.4.13 How Do I View Run Logs After a Job Is Tested?.....	701
4.4.14 Why Does a Job Scheduled by Month Start Running Before the Job Scheduled by Day Is Complete? .....	701
4.4.15 What Should I Do If Invalid Authentication Is Reported When I Run a DLI Script?.....	702
4.4.16 Why Cannot I Select the Desired CDM Cluster in Proxy Mode When Creating a Data Connection? .....	702
4.4.17 Why Is There No Job Running Scheduling Record After Daily Scheduling Is Configured for the Job? .....	702
4.4.18 What Do I Do If No Content Is Displayed in Job Logs?.....	703
4.4.19 Why Do I Fail to Establish a Dependency Between Two Jobs?.....	703
4.4.20 What Should I Do If an Error Is Displayed During DataArts Studio Scheduling: The Job Does Not Have a Submitted Version?.....	704
4.4.21 What Do I Do If an Error Is Displayed During DataArts Studio Scheduling: The Script Associated with Node XXX in the Job Is Not Submitted?.....	704
4.4.22 What Should I Do If a Job Fails to Be Executed After Being Submitted for Scheduling and an Error Displayed: Depend Job [XXX] Is Not Running Or Pause?.....	705
4.4.23 How Do I Create a Database And Data Table? Is the database a data connection?.....	705
4.4.24 Why Is No Result Displayed After an HIVE Task Is Executed?.....	705
4.4.25 Why Does the Last Instance Status On the Monitor Instance page Only Display Succeeded or Failed?.....	705
4.4.26 How Do I Create a Notification for All Jobs?.....	705
4.4.27 How Many Nodes Can Be Executed Concurrently in Each DataArts Studio Version?.....	706
4.4.28 What Is the Priority of the Startup User, Execution User, Workspace Agency, and Job Agency?.....	706

# 1 Service Overview

---

## 1.1 What Is DataArts Studio?

### Challenges to Enterprise Digital Transformation

Enterprises often face challenges in the following aspects when managing data:

- Governance
  - Inconsistent data system standards impact data exchange and sharing between different departments.
  - There are no great search tools to help service personnel locate the data they need when they need it.
  - If metadata fails to define data in business terms that are familiar to data consumers, the data is difficult to understand.
  - When there are no good methods to evaluate and control data quality, it makes the data hard to trust.
- Operations
  - Data analysts and decision makers require efficient data operations. There is no efficient data operations platform to address the growing and diversified demands for analytics and reporting.
  - Repeated development of the same data wastes time, slows down development, and results in too many data copies. Inconsistent data standards waste resources and drive up costs.
- Innovation
  - Data silos prevent data from being shared and circulated across departments in enterprises. As a result, cross-domain data analysis and data innovation fail to be stimulated.
  - Currently, most enterprises still utilize their data for analytics and reporting. There is a long way to go before enterprises have widespread, data-driven service innovation.

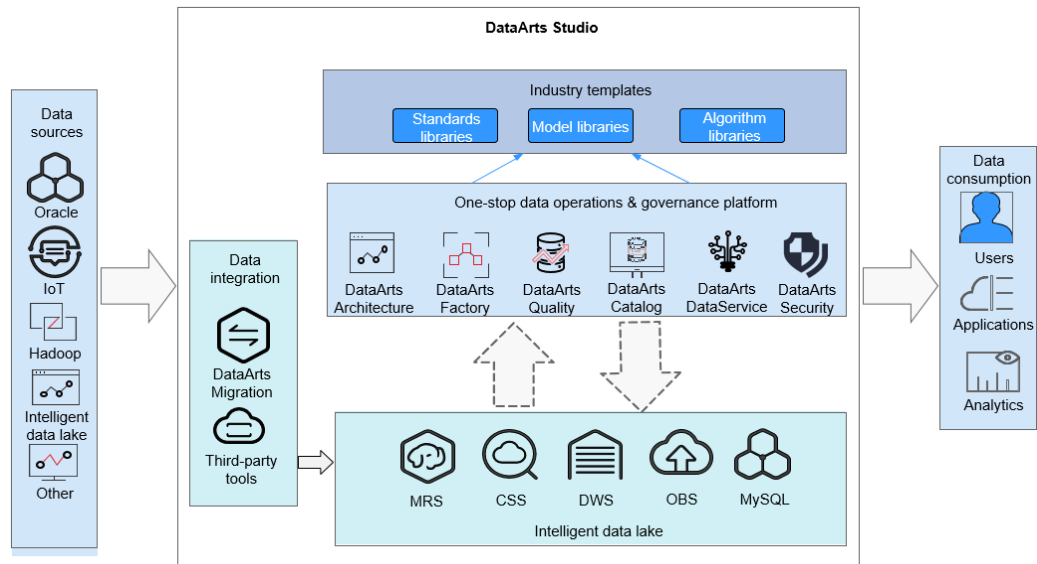
### What Is DataArts Studio?

DataArts Studio is a one-stop data operations platform that drives digital transformation. It allows you to perform many operations, such as integrating and

developing data. Incorporating big data storage, computing and analytical engines, it can also construct industry knowledge bases and help your enterprise build an intelligent end-to-end data system. This system can eliminate data silos, unify data standards, accelerate data monetization, and accelerate your enterprise's digital transformation.

Figure 1-1 shows the architecture.

Figure 1-1 Architecture



As shown in the figure, DataArts Studio is based on the data lake base and provides capabilities such as data integration, development, governance, and openness. DataArts Studio can connect to data lakes and cloud database services, such as Data Lake Insight (DLI), MRS Hive, and GaussDB(DWS). These data lakes and cloud database services are used as the data lake base. DataArts Studio can also connect to traditional enterprise data warehouses, such as Oracle and Greenplum.

DataArts Studio consists of the following functional modules:

- **Management Center**  
Management Center supports data connection management and connects to the data lake base for activities such as data development.
- **DataArts Migration**  
DataArts Migration supports data migration between 20+ data sources and integration of data sources into the data lake. It provides wizard-based configuration and management and supports single table, entire database, incremental, and periodic data integration.
- **DataArts Factory**  
DataArts Factory helps you build a big data processing center, create data models, integrate data, develop scripts, and orchestrate workflows.



## 1.2 Basic Concepts

### DataArts Studio Instance

A DataArts Studio instance is the minimum unit of compute resources provided for users. You can create, access, and manage multiple DataArts Studio instances at the same time. A DataArts Studio instance allows you to access seven modules: Management Center, DataArts Architecture, DataArts Migration, DataArts Factory, DataArts Quality, DataArts Catalog, and DataArts DataService. You can obtain DataArts Studio instances with specifications tailored to your service requirements.

### Workspace

A workspace enables admins to manage member permissions, resources, and configurations of the underlying compute engines.

The workspace is a basic unit for member management as well as role and permission assignment. Each team must have an independent workspace.

You can access the Management Center, DataArts Factory, and DataArts Migration modules, but only after your account is added to a workspace and assigned the permissions required to perform such operations.

### Member and Role

A member is an account that has been assigned the permissions required to access and use a workspace. As an admin, when you add a workspace member, you must set a role.

A role is a predefined combination of permissions. Different roles have different permission sets. After a role is assigned to a member, the member has all the permissions of that role. Each member must have at least one role, and they can have multiple roles at the same time.

### DataArts Migration

A DataArts Migration cluster is the smallest resource unit provided to users. DataArts Migration clusters run on ECSs. You can create data migration tasks in a CDM cluster and migrate data between homogeneous or heterogeneous data sources in the cloud and on-premises.

### Data Source

A data source is a medium for storing or processing data, such as a relational database, data warehouse, and data lake. Different data sources use different data storage, transmission, processing, and application modes, as well as different scenarios, technologies, and tools.

## Source Data

Source data is the data that is not processed after created. In data management, source data refers to the data directly from source files (such as service system databases, offline files, and IoT files) or copies of source files.

## Data Connection

A data connection is a collection of details required for accessing where data is stored, including the connection type, name, and login information.

## Concurrency

Concurrency refers to the maximum number of threads that can be concurrently read from the source in a data integration job.

## Dirty Data

Dirty data refers to the data meaningless to business or in invalid format. For example, if the source data of the VARCHAR type is not properly converted, it cannot be written to the destination column of the INT type.

## Job (DataArts Factory)

A job is composed of one or more nodes that run together to complete data operations.

## Node

A node is a definition for the actions to be performed on your data. For example, you can use the MRS Spark node to execute predefined Spark jobs in MRS.

## Solution

A solution is a series of convenient and systematic management operations that meet service requirements and objectives. Each solution can contain one or more business-related jobs, and each job can be reused by multiple solutions.

## Resource

A resource is the self-defined code or text file that you upload. It is invoked when nodes run.

## Expression Language (EL)

Node parameters in data development jobs can be dynamically generated based on the running environment using ELs. An EL often uses simple arithmetic and calculation logic and references embedded objects including job objects and tool objects.

## Environment Variable

An environmental variable is an object with a specific name in the operating system. It contains information to be used by one or more applications.

## PatchData

PatchData is an instance that was generated in the past by a repeatedly scheduled job.

# 1.3 Functions

## DataArts Migration: Efficient Ingestion of Multiple Heterogeneous Data Sources

DataArts Migration can help you seamlessly migrate batch data between 20+ homogeneous or heterogeneous data sources. You can use it to ingest data from both on-premises and cloud-based data sources, including file systems, relational databases, data warehouses, NoSQL databases, big data services, and object storage.

DataArts Migration uses a distributed compute framework and concurrent processing techniques to help you migrate enterprise data in batches without any downtime and rapidly build desired data structures.

You can manage data on the wizard-based task management page. You can easily create data migration tasks that meet your requirements. DataArts Migration provides the following functions:

- **Table/File/Entire DB migration**  
You can migrate tables or files in batches, and migrate an entire database between homogeneous and heterogeneous database systems. You can include hundreds of tables in a single job.
- **Incremental data migration**  
You can migrate files, relational databases, and HBase in an incremental manner. You can perform incremental data migration by using WHERE clauses and variables of date and time.
- **Migration in transaction mode**  
When a batch data migration job fails to be executed, data will be rolled back to the state before the job started and data in the destination table will be automatically deleted.
- **Field conversion**  
Field conversion includes anonymization, character string operations, and date operations.
- **File encryption**  
You can encrypt files that are migrated to a cloud-based file system in batches.
- **MD5 verification**  
MD5 is used to check file consistency from end to end.

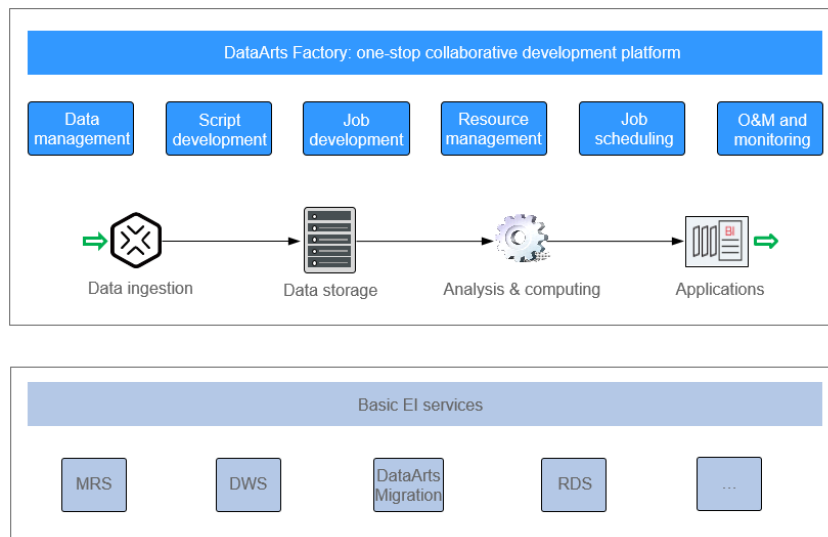
- **Dirty data archiving**

Data that fails to be processed during migration, is filtered out and is not compliant with conversion or cleansing rules is recorded in dirty data logs. You can easily analyze abnormal data. You can also set a threshold for the dirty data ratio to determine whether a task is successful.

## DataArts Factory: One-stop Collaborative Development

DataArts Factory provides an intuitive UI and built-in development methods for script and job development. DataArts Factory also supports fully hosted job scheduling, O&M, and monitoring, and incorporates industry data processing pipelines. You can create data development jobs in a few steps, and the entire process is visual. Online jobs can be jointly developed by multiple users. You can use DataArts Factory to manage big data cloud services and quickly build a big data processing center.

Figure 1-2 DataArts Factory architecture



DataArts Factory allows you to manage data, develop scripts, and schedule and monitor jobs. Data analysis and processing are easier than ever before.

- **Data management**

- You can manage multiple types of data warehouses, such as GaussDB (DWS), DLI, and MRS Hive.
- You can use the graphical interface and data definition language (DDL) to manage database tables.

- **Script development**

- Provides an online script editor that allows more than one operator to collaboratively develop and debug SQL, Python, and Shell scripts online.
- You can use Variables.

- **Job development**

- DataArts Factory provides a graphical designer that allows you to rapidly develop workflows through drag-and-drop and build data processing pipelines.

- DataArts Factory is preset with multiple task types such as data integration, SQL, and Shell. Data is processed and analyzed based on task dependencies.
- You can import and export jobs.
- **Resource management**  
You can centrally manage file, jar, and archive resources used during script and job development.
- **Job scheduling**
  - You can schedule jobs to run once or recursively and use events to trigger scheduling jobs.
  - Job scheduling supports a variety of hybrid orchestration tasks. The high-performance scheduling engine has been tested by hundreds of applications.
- **O&M and monitoring**
  - You can run, suspend, restore, or terminate a job.
  - You can view the operation details of each job and each node in the job.
  - You can use various methods to receive notifications when a job or task error occurs.

## 1.4 Advantages

### One-Stop Data Operations Platform

DataArts Studio is a one-stop data operations platform that allows you to perform many operations, including integrating data from every domain and connecting data from different data sources. In a word, it can help you build a comprehensive data governance solution.

### Diverse Data Development Types

DataArts Studio has a wide range of scheduling configuration policies and powerful job scheduling. It supports online collaborative development among multiple users, online editing and real-time query of SQL and shell scripts, and job development via data processing nodes such as CDM, SQL, MRS, Shell, MLS, and Spark.

### Unified Scheduling and O&M

Fully hosted scheduling is supported. Time- and event-based triggering mechanisms are available. You can schedule a task by minute, hour, day, week, or month.

The visualized task O&M center monitors all tasks and supports notification settings, enabling you to obtain real-time task status and ensuring normal running of services.

## 1.5 Application Scenarios

### Building Cloud-based Data Platforms with Speed

You can use DataArts Studio to migrate offline data to the cloud and integrate the data into big data services. On the DataArts Studio management console, you can use the integrated data to quickly start developing jobs and easily build enterprise data systems.

#### Advantages

- Quick data integration  
On the GUI, you can migrate offline or real-time data to cloud warehouses in just a few steps.
- Multiple warehouse services  
You can choose GaussDB (DWS), MRS, or any other warehouses to meet your service needs.
- Secure, stable, and cost-saving  
Data on the cloud is secure owing to one-stop data service capabilities and stable data warehouse services; you no longer need to build and maintain big data clusters, significantly reducing costs.

## 1.6 DataArts Studio Permissions Management

If you need to assign different permissions to employees in your enterprise to access your DataArts Studio resources, IAM is a good choice for fine-grained permissions management. IAM provides identity authentication, permissions management, and access control, helping you secure access to your resources.

With IAM, you can use your account to create IAM users for your employees, and assign permissions to the users to control their access to specific resource types. For example, if you want to allow some software developers in your enterprise to use DataArts Studio resources but disallow them to delete workspaces or perform any high-risk operations, you can create IAM users for the software developers and grant them only the permissions required for using DataArts Studio resources.

### DataArts Studio Permissions

By default, new IAM users do not have any permissions. To assign permissions to a user, add the user to one or more groups and assign permissions policies or roles to these groups. The user then inherits permissions from the groups it is a member of. After authorization, the users can perform specified operations.

DataArts Studio is a project-level service deployed in specific physical regions. To assign ServiceStage permissions to a user group, specify the scope as region-specific projects and select projects for the permissions to take effect. If **All projects** is selected, the permissions will take effect for the user group in all region-specific projects. When accessing DataArts Studio, users need to switch to a region where they are authorized to use cloud services.

- IAM Roles:** IAM initially provides a coarse-grained authorization mechanism to define permissions based on users' job responsibilities. This mechanism provides only a limited number of service-level roles for authorization. However, IAM roles are not an ideal choice for fine-grained authorization and secure access control.

Relying on IAM roles, DataArts Studio provides more flexible, fine-grained authorization based on workspace roles for specific operations.

The system-defined roles supported by DataArts Studio include **DAYU Administrator** and **DAYU User**. Workspace roles are based on the **DAYU User**. [DataArts Studio Permissions](#) describes the common operations supported by DataArts Studio and the permissions granted to each role. You can select roles as required.

**Table 1-1** DataArts Studio system-defined roles

Role	Description	Category
DAYU Administrator	User who has all permissions of DataArts Studio and workspaces  <b>NOTE</b> Users assigned the <b>Tenant Administrator</b> role have all permissions for all services except IAM. In other words, users with the <b>Tenant Administrator</b> role can perform all operations in DataArts Studio.	System-defined role

Role	Description	Category
DAYU User	<p>Common DataArts Studio user</p> <p>Users with the <b>DAYU User</b> role have the permissions of the role assigned to them in a workspace. Roles in a workspace can be admin, developer, operator, and viewer. For details about the operation permissions of each role, see <a href="#">DataArts Studio Permissions</a>.</p> <ul style="list-style-type: none"><li>• Admin: Users with this role have the permissions to perform all operations in a workspace. You are advised to assign this role to the project owner, development owner, and O&amp;M administrator.</li><li>• Developer: Users with this role have the permissions to create and manage work items, but cannot perform operations on workspaces, clusters, and reviewers. You are advised to assign this role to users who develop and process tasks.</li><li>• Operator: Users with this role have the permissions to perform operations such as O&amp;M and scheduling, but cannot modify work items or configurations. You are advised to assign this role to users for O&amp;M management and status monitoring.</li><li>• Viewer: Users with this role can only read data from DataArts Studio, but cannot perform operations on workspaces or modify work items or configurations. You are advised to assign this role to users who only want to view information in the workspace but do not perform any operation.</li></ul>	System-defined role

After a role is granted to you, you have all the permissions of the role. For details about how to grant permissions of a DataArts Studio role, see **Preparations > Creating IAM Users and Granting DataArts Studio Permissions** in *DataArts Studio User Guide*.

## 1.7 DataArts Studio Permissions

A workspace member can be assigned the role of admin, developer, operator, or viewer. This topic describes the permissions of each role.

- Admin: Users with this role have the permissions to perform all operations in a workspace. You are advised to assign this role to the project owner, development owner, and O&M administrator.
- Developer: Users with this role have the permissions to create and manage work items, but cannot perform operations on workspaces, clusters, and



reviewers. You are advised to assign this role to users who develop and process tasks.

- **Operator:** Users with this role have the permissions to perform operations such as O&M and scheduling, but cannot modify work items or configurations. You are advised to assign this role to users for O&M management and status monitoring.
- **Viewer:** Users with this role can only read data from DataArts Studio, but cannot perform operations on workspaces or modify work items or configurations. You are advised to assign this role to users who only want to view information in the workspace but do not perform any operation.

 **NOTE**

Accounts and users with the **DAYU Administrator** or **Tenant Administrator** role have all the permissions on account, including permissions to create DataArts Studio instances and DataArts Studio incremental packages. By default, other users do not have the permissions to create DataArts Studio instances. If they want to create DataArts Studio instances, they must obtain the required permissions.

Accounts and users with the **Tenant Administrator** role can perform all operations except IAM user management. For security purposes, you are not advised to assign this role to IAM users. Exercise caution when performing this operation.

## Workspace

Permission	Admin	Developer	Operator	Viewer
Creating Workspaces	Users with the <b>DAYU Administrator</b> or <b>Tenant Administrator</b> role can perform this operation.			
Modifying Workspaces	Yes	No	No	No
Disabling or Enabling Workspaces	Yes	No	No	No
Querying Workspaces	Yes	Yes	Yes	Yes
Adding Workspace Members	Yes	No	No	No
Modifying Workspace Members	Yes	No	No	No
Removing Workspace Members	Yes	No	No	No
Querying Workspace Members	Yes	Yes	Yes	Yes

## Management Center

Permission	Admin	Developer	Operator	Viewer
Creating Data Connections	Yes	Yes	No	No
Updating Data Connections	Yes	Yes	No	No
Deleting Data Connections	Yes	Yes	No	No
Obtaining Data Connections	Yes	Yes	Yes	Yes
Testing Data Connections	Yes	Yes	No	No
Obtaining the List of Data Source Types	Yes	Yes	Yes	Yes
Obtaining the List of Available Data Source Types of DataArts Catalog	Yes	Yes	Yes	Yes
Querying Hive Connection Information	Yes	Yes	Yes	Yes
Obtaining the List of Data Source Directories	Yes	Yes	Yes	Yes
Updating Data Source Extension Tables	Yes	Yes	No	No
Creating Data Collection Tasks	Yes	Yes	No	No
Obtaining the List of OBS Buckets	Yes	Yes	Yes	Yes
Obtaining the File List in an OBS Bucket	Yes	Yes	Yes	Yes
Importing Data Sources	Yes	Yes	No	No

Permission	Admin	Developer	Operator	Viewer
Exporting Data Sources	Yes	Yes	No	No
Obtaining the List of KMS Keys	Yes	Yes	Yes	Yes
Obtaining the List of CDM Clusters	Yes	Yes	Yes	Yes

## DataArts Migration

Permission	Admin	Developer	Operator	Viewer
Querying Data Connections	Yes	Yes	Yes	Yes
Testing Data Connections	Yes	Yes	Yes	No
Testing Connectivity	Yes	Yes	Yes	No
Creating Data Connections	Yes	Yes	Yes	No
Deleting Data Connections	Yes	Yes	Yes	No
Viewing Historical Jobs	Yes	Yes	Yes	Yes
Querying All Jobs in a Database	Yes	Yes	Yes	Yes
Querying Common Jobs	Yes	Yes	Yes	Yes
Checking Whether a Job Name Exists	Yes	Yes	Yes	Yes
Querying the Status of a Job	Yes	Yes	Yes	Yes
Obtaining Connection Metadata	Yes	Yes	Yes	Yes
Creating Connection Metadata	Yes	Yes	Yes	No

Permission	Admin	Developer	Operator	Viewer
Modifying Connection Metadata	Yes	Yes	Yes	No
Saving Jobs	Yes	Yes	Yes	No
Editing Jobs	Yes	Yes	Yes	No
Executing Jobs	Yes	Yes	Yes	No
Stopping Jobs	Yes	Yes	Yes	No
Querying the Statuses of Multiple Jobs	Yes	Yes	Yes	Yes
Querying Job Details or JSON	Yes	Yes	Yes	Yes
Querying Historical Job Execution Records	Yes	Yes	Yes	Yes
Viewing Job Logs	Yes	Yes	Yes	Yes
Deleting Jobs	Yes	Yes	Yes	No
Importing Jobs	Yes	Yes	Yes	No
Exporting Jobs	Yes	Yes	Yes	No
Backing Up Jobs	Yes	Yes	Yes	No
Querying Job Groups	Yes	Yes	Yes	Yes
Creating Job Groups	Yes	Yes	Yes	No
Modifying Job Groups	Yes	Yes	Yes	No
Deleting Job Groups	Yes	Yes	Yes	No
Querying Configuration Variables	Yes	Yes	Yes	No
Setting Configuration Variables	Yes	Yes	Yes	No
Isolating Users	Yes	Yes	Yes	No

Permission	Admin	Developer	Operator	Viewer
Authorizing EIP Check	Yes	No	No	No
Restarting Clusters	Yes	Yes	Yes	No
Binding EIPs	Yes	No	No	No
Unbinding EIPs	Yes	No	No	No
Modifying Clusters	Yes	Yes	No	No
Deleting Clusters	Yes	Yes	No	No
Creating Dynamic Clusters	Yes	Yes	No	No
Querying the List of Clusters	Yes	Yes	Yes	Yes
Querying Details About a Cluster	Yes	Yes	Yes	Yes
Querying Details About an Instance	Yes	Yes	Yes	Yes
Collecting Cluster Statistics	Yes	Yes	Yes	Yes
Cluster agent	Yes	Yes	Yes	No

## DataArts Factory

Permission	Admin	Developer	Operator	Viewer
Obtaining the List of Environment Variables	Yes	Yes	Yes	Yes
Updating Environment Variables	Yes	Yes	No	No
Importing Environment Variables	Yes	Yes	No	No
Exporting Environment Variables	Yes	Yes	No	No

Permission	Admin	Developer	Operator	Viewer
Obtaining the List of Data Tables	Yes	Yes	Yes	Yes
Viewing Table Details	Yes	Yes	Yes	Yes
Creating Data Tables	Yes	Yes	No	No
Updating Data Tables	Yes	Yes	No	No
Deleting Data Tables	Yes	Yes	No	No
Obtaining the List of Databases	Yes	Yes	Yes	Yes
Viewing Database Details	Yes	Yes	Yes	Yes
Creating Databases	Yes	Yes	No	No
Updating Databases	Yes	Yes	No	No
Deleting Databases	Yes	Yes	No	No
Obtaining the List of Schemas	Yes	Yes	Yes	Yes
Viewing Schema Details	Yes	Yes	Yes	Yes
Creating Schemas	Yes	Yes	No	No
Updating Schemas	Yes	Yes	No	No
Deleting Schemas	Yes	Yes	No	No
Obtaining Directory Trees	Yes	Yes	Yes	Yes
Creating Directories	Yes	Yes	No	No
Refreshing Directories	Yes	Yes	No	No
Deleting Directories	Yes	Yes	No	No

Permission	Admin	Developer	Operator	Viewer
Executing Scripts	Yes	Yes	Yes	No
Creating Scripts	Yes	Yes	No	No
Obtaining Script Details	Yes	Yes	Yes	Yes
Updating Scripts	Yes	Yes	No	No
Deleting Scripts	Yes	Yes	No	No
Obtaining the List of Scripts	Yes	Yes	Yes	Yes
Canceling Script Execution	Yes	Yes	Yes	No
Importing Scripts	Yes	Yes	No	No
Exporting Scripts/ Execution Results	Yes	Yes	Yes	No
Creating Solutions	Yes	Yes	No	No
Deleting Solutions	Yes	Yes	No	No
Updating Solutions	Yes	Yes	No	No
Viewing Solution Details	Yes	Yes	Yes	Yes
Obtaining the List of Solutions	Yes	Yes	Yes	Yes
Exporting Solutions	Yes	Yes	Yes	No
Importing Solutions	Yes	Yes	No	No
Obtaining the List of Jobs	Yes	Yes	Yes	Yes
Viewing Job Details	Yes	Yes	Yes	Yes
Creating Jobs	Yes	Yes	No	No
Renaming Jobs	Yes	Yes	No	No
Deleting Jobs	Yes	Yes	No	No
Updating Jobs	Yes	Yes	Yes	No

Permission	Admin	Developer	Operator	Viewer
Exporting Jobs	Yes	Yes	Yes	No
Importing Jobs	Yes	Yes	No	No
Verifying Parameters of Import Jobs	Yes	Yes	No	No
Performing Test Run	Yes	Yes	Yes	No
Suspending Jobs	Yes	Yes	Yes	No
Resuming Job Running	Yes	Yes	Yes	No
Running Jobs	Yes	Yes	Yes	No
Stopping Jobs	Yes	Yes	Yes	No
Obtaining the List of Instances	Yes	Yes	Yes	Yes
Rerunning Instances	Yes	Yes	Yes	No
Stopping Instances	Yes	Yes	Yes	No
Forcibly Succeed	Yes	Yes	Yes	No
Resuming Instance Running	Yes	Yes	Yes	No
Disabling Realtime Jobs	Yes	Yes	Yes	No
Recovering Realtime Jobs	Yes	Yes	Yes	No
Manually Retrying Job Nodes	Yes	Yes	Yes	No
Skipping Job Nodes	Yes	Yes	Yes	No
Suspending Job Nodes	Yes	Yes	Yes	No
Recovering Job Nodes	Yes	Yes	Yes	No
Forcibly Succeed	Yes	Yes	Yes	No



Permission	Admin	Developer	Operator	Viewer
Querying Data Connection Details	Yes	Yes	Yes	Yes
Obtaining the List of Data Connections	Yes	Yes	Yes	Yes
Creating Data Connections	Yes	Yes	No	No
Updating Data Connections	Yes	Yes	No	No
Deleting Data Connections	Yes	Yes	No	No
Testing Data Connections	Yes	Yes	No	No
Importing Data Connections	Yes	Yes	No	No
Exporting Data Connections	Yes	Yes	No	No
Obtaining the List of Resources	Yes	Yes	Yes	Yes
Viewing Resource Details	Yes	Yes	Yes	Yes
Creating Resources	Yes	Yes	No	No
Updating Resources	Yes	Yes	No	No
Deleting Resources	Yes	Yes	No	No
Importing Resources	Yes	Yes	No	No
Exporting Resources	Yes	Yes	Yes	No
Starting Daily Backups	Yes	Yes	Yes	No
Stopping Daily Backups	Yes	Yes	Yes	No
Obtaining the List of Backups	Yes	Yes	Yes	Yes

Permission	Admin	Developer	Operator	Viewer
Obtain the List of Notifications	Yes	Yes	Yes	Yes
Creating Notifications	Yes	Yes	No	No
Updating Notifications	Yes	Yes	No	No
Deleting Notifications	Yes	Yes	No	No
Creating Job Monitoring PatchData	Yes	Yes	No	No
Obtaining the List of PatchData Monitoring	Yes	Yes	Yes	Yes
Stopping Job PatchData	Yes	Yes	Yes	No

## 1.8 Constraints

### Browser Restrictions

The following table lists the recommended browser for logging in to DataArts Studio.

**Table 1-2** Browser compatibility

Browser Version	Description
Google Chrome 93.x or later	Recommended

### Use Restrictions

Before using DataArts Studio, you must read and understand the following restrictions:

1. DataArts Studio is a one-stop platform that provides data integration, development, and governance capabilities. DataArts Studio has no storage or computing capability and relies on the data lake base.
2. Different components of DataArts Studio support different data sources. You need to select a data lake base based on your service requirements. For details about the data lakes supported by DataArts Studio, see **Management Center > Data Sources Supported by DataArts Studio** in the *DataArts Studio User Guide*.

3. For details about the constraints on DataArts Migration, see **DataArts Migration > Constraints** in the *DataArts Studio User Guide*.

## Reliability Restrictions

To achieve high reliability in using DataArts Studio, you are advised to understand the following restrictions and measures:

1. The DataArts Migration cluster is deployed in standalone mode. A cluster fault may cause service and data loss. You are advised to use the CDM Job node of DataArts Factory to invoke CDM jobs and select two CDM clusters to improve reliability. For details, see **DataArts Factory > Nodes > CDM Job** in the *DataArts Studio User Guide*.
2. You can enable automatic backup and restoration of CDM jobs. Backups of CDM jobs are stored in OBS buckets. For details, see **DataArts Migration > Job Management > Job Configuration Management** in the *DataArts Studio User Guide*.
3. You can enable scheduled backup of assets such as scripts and jobs to OBS buckets, and restore assets from their backups. For details, see **DataArts Factory > O&M and Scheduling > Managing Backups** in the *DataArts Studio User Guide*.

## 1.9 Related Services

### IAM

DataArts Studio uses Identity and Access Management (IAM) for authentication and authorization.

### CTS

DataArts Studio uses Cloud Trace Service (CTS) to audit users' non-query operations on the management console to ensure that no invalid or unauthorized operations have been performed. CTS enhances security.

### ECS

CDM and DataArts DataService clusters of DataArts Studio consist of Elastic Cloud Servers (ECSs). In addition, DataArts Studio can use host connections to connect to ECSs and run Shell or Python scripts.

### VPC

Virtual Private Cloud (VPC) provides isolated network environments for DataArts Studio.

### EIP

Elastic IP (EIP) enables DataArts Studio to communicate with the Internet.

## **OBS**

DataArts Studio uses Object Storage Service (OBS) buckets to store logs.

## **SMN**

DataArts Studio uses Simple Message Notification (SMN) to send push notifications based on your subscription requirements, so that you can receive immediate notifications when specific events occur.

## **Direct Connect**

Direct Connect enables DataArts Studio to communicate with third-party data centers.

## **API Gateway**

API Gateway (APIG) enables DataArts Studio to provision the APIs of its modules.

## **DLI**

Data Lake Insight (DLI) can be used as the data lake for DataArts Studio and enables data integration, development, governance, and provisioning.

## **MRS**

MapReduce Service (MRS) can be used as the data lake for DataArts Studio and enables data integration, development, and governance.

## **GaussDB(DWS)**

GaussDB(DWS) can be used as the data lake for DataArts Studio and enables data integration, development, governance, and provisioning.

## **RDS**

Relational Database Service (RDS) provides data sources for DataArts Studio and enables data integration, development, and provisioning.

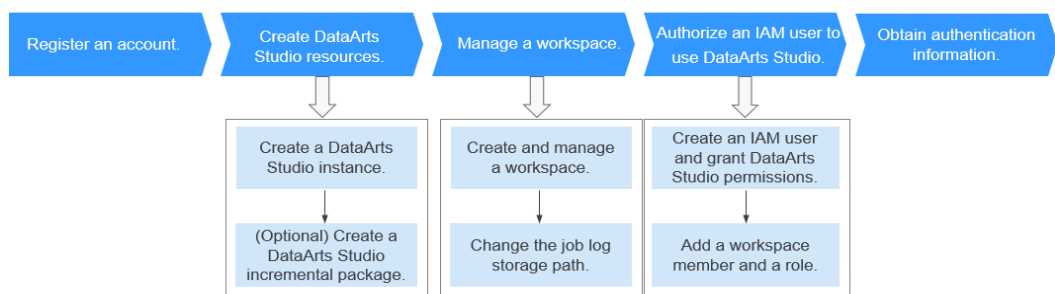
# 2 Preparations

## 2.1 Preparations

To use DataArts Studio, create a account, create a DataArts Studio instance, and authorize a user to use DataArts Studio.

For details about the preparations and operations, see later sections.

Figure 2-1 DataArts Studio preparation process



## 2.2 Creating DataArts Studio Instances

### 2.2.1 Creating a DataArts Studio Basic Package

#### Background

Only account users with the **DAYU Administrator** or **Tenant Administrator** permissions can create DataArts Studio instances or DataArts Studio incremental packages.


**NOTE**

Users with the **Tenant Administrator** permissions can perform all operations except IAM user management. For security purposes, you are not advised to grant the **Tenant Administrator** permissions to IAM users.

## Prerequisites

You have obtained a VPC, subnet, and security group. You can also apply for them when you create a DataArts Studio instance.

## Logging In to DataArts Studio Console

1. Log in to the cloud management console.
2. In the upper left corner of the console, click , and choose **DataArts Studio** to access the DataArts Studio console.

## Creating a DataArts Studio Basic Package

**Step 1** On the DataArts Studio console, click **Create Instance**.

**Step 2** On the displayed page, set parameters for the DataArts Studio instance. [Table 2-1](#) provides descriptions of the parameters.

**Table 2-1** DataArts Studio instance parameters

Parameter	Example	Description
Region	None	The region where the instance resides. Resources in different regions cannot communicate with each other.
Enterprise Project	default	The enterprise project associated with your DataArts Studio instance. This parameter is available only when an enterprise project has been created. If you want to connect the DataArts Studio instance to a cloud service (such as DWS, MRS, and RDS), ensure that the enterprise project of the DataArts Studio instance is the same as that of the cloud service instance. <ul style="list-style-type: none"><li>• You can create only one DataArts Studio instance for an enterprise project.</li><li>• If you want to enable communication between DataArts Studio and another cloud service, ensure that the enterprise project of DataArts Studio is the same as that of the cloud service.</li></ul>
Instance Name	DataArts Studio-test	Name of the DataArts Studio instance

**Step 3** (Optional) If you have set a tag key and a tag value, click **Add** to add the tag.

 **NOTE**

- A maximum of 20 tags can be added.
- Only one tag value can be added to a tag key.
- The key name must be unique in the same instance.

**Step 4** Confirm the settings and click **create Now**.

**Step 5** When you return to the DataArts Studio management console, the **Authorize Access** dialog box is displayed, prompting you to authorize the listed services. DataArts Studio interacts with the cloud services and needs to collaborate with them. Therefore, you need to create a cloud service agency to delegate permissions to DataArts Studio so that DataArts Studio can use these cloud services and perform task scheduling and resource O&M on behalf of you.

Cloud service agencies include permissions related to DWS, MRS, RDS, OBS, SMN, and KMS. You can access the IAM agency page to view the agency scope. You do not need to apply for an agency for users. The agency of the account is used.

Select all services and click **Authorize**. The platform automatically creates an agency.

- After the authorization is complete, the **Authorize Access** dialog box will not be displayed when you access the DataArts Studio console homepage next time.
- If you select only some services for authorization, the system still displays the dialog box next time you access the DataArts Studio console homepage, prompting you to authorize access to unauthorized cloud services.

**Step 6** In the list of instances, locate your instance and click **Access** to access the DataArts Studio console.

----End

## 2.2.2 (Optional) Creating a DataArts Studio Incremental Package

DataArts Studio provides basic and incremental packages. If the basic package cannot meet your requirements, you can create an incremental package. Before you create an incremental package, ensure that you have a DataArts Studio instance.

You can choose the following incremental packages:

- DataArts Migration (namely CDM) incremental package  
The DataArts Studio instance does not contain CDM clusters. To use the DataArts Migration function, create a CDM incremental package.

### Background

After you create an incremental package, the system automatically creates a cluster for your service based on your selected specifications.

## Creating a CDM Cluster

1. Locate an enabled instance and click **Create**.
2. On the displayed page, set parameters based on [Table 2-2](#).

**Table 2-2** Parameters for the CDM incremental package

Parameter	Description
Package	Select <b>CDM</b> .
AZ	<p>When you a DataArts Studio instance or incremental package for the first time, there is no requirement on the AZ.</p> <p>When you create a new DataArts Studio instance or incremental package, determine whether to select the same AZ as the existing one based on your DR and network latency demands.</p> <ul style="list-style-type: none"><li>• If your application requires good DR capability, deploy resources in different AZs in the same region.</li><li>• If your application requires a low network latency between instances, deploy resources in the same AZ.</li></ul>
Workspace	Select the workspace where the CDM incremental package is to be used. For example, if you want to create a CDM incremental package in workspace <b>A</b> of the <b>test</b> DataArts Studio instance, select <b>A</b> . After you create the CDM incremental package, a CDM cluster will be displayed in workspace <b>A</b> .
Cluster	Customize the cluster name.
Instance	<p>The following CDM cluster flavors are available:</p> <ul style="list-style-type: none"><li>• <b>cdm.large</b>: 8 vCPUs and 16 GB of memory The maximum and assured bandwidths are 3 Gbit/s and 0.8 Gbit/s. Up to 20 jobs can be executed concurrently.</li><li>• <b>cdm.xlarge</b>: 16 vCPUs and 32 GB of memory The maximum and assured bandwidths are 10 Gbit/s and 4 Gbit/s. Up to 100 jobs can be executed concurrently. This flavor is well suited to TB-level data migration requiring 10GE high-speed bandwidth.</li><li>• <b>cdm.4xlarge</b>: 64 vCPUs and 128 GB of memory The maximum and assured bandwidths are 40 and 36 Gbit/s, respectively. Up to 300 jobs can be executed concurrently.</li></ul>



Parameter	Description
VPC	<p>VPC to which the CDM cluster in the DataArts Studio instance belongs. A VPC is a secure, isolated, and logical network environment.</p> <p>If you want to connect the DataArts Studio instance or CDM cluster to a cloud service (such as DWS, MRS, and RDS), ensure that the CDM cluster can communicate with the cloud service. If the CDM cluster and the cloud service are in the same region, VPC, subnet, and security group, they can communicate with each other by default. If the CDM cluster and the cloud service are in the same VPC but in different subnets or security groups, you must configure routing rules and security group rules.</p> <p>For details, see <i>Virtual Private Cloud User Guide</i>.</p> <p><b>NOTE</b> After the CDM instance is created, the VPC cannot be changed.</p>
Subnet	<p>Subnet to which the CDM cluster in the DataArts Studio instance belongs. A subnet provides dedicated network resources that are isolated from other networks, improving network security.</p> <p>If you want to connect the DataArts Studio instance or CDM cluster to a cloud service (such as DWS, MRS, and RDS), ensure that the CDM cluster can communicate with the cloud service. If the CDM cluster and the cloud service are in the same region, VPC, subnet, and security group, they can communicate with each other by default. If the CDM cluster and the cloud service are in the same VPC but in different subnets or security groups, you must configure routing rules and security group rules.</p> <p>For details, see <i>Virtual Private Cloud User Guide</i>.</p> <p><b>NOTE</b> After the CDM instance is created, the VPC cannot be changed.</p>

Parameter	Description
Security Group	<p>Security group to which the CDM cluster in the DataArts Studio instance belongs. A security group is a set of ECS access rules. It provides access policies for ECSs that have the same security protection requirements and are mutually trusted in a VPC.</p> <p>If you want to connect the DataArts Studio instance or CDM cluster to a cloud service (such as DWS, MRS, and RDS), ensure that the CDM cluster can communicate with the cloud service. If the CDM cluster and the cloud service are in the same region, VPC, subnet, and security group, they can communicate with each other by default. If the CDM cluster and the cloud service are in the same VPC but in different subnets or security groups, you must configure routing rules and security group rules.</p> <p>For details, see <i>Virtual Private Cloud User Guide</i>.</p> <p><b>NOTE</b> After the CDM instance is created, the security group cannot be changed.</p>

---

**NOTICE**

You cannot modify the specifications of an existing cluster. If you need higher specifications, create another cluster.

---

3. Click **Create Now**, confirm the specifications, and click **Next**.
4. View the CDM cluster in the corresponding workspace.

## 2.3 Managing a Workspace

### 2.3.1 Creating and Managing a Workspace

By default, a workspace will be automatically created after you create a DataArts Studio instance. You will be automatically assigned the admin role and can use the default workspace or create a new workspace.

A workspace in a DataArts Studio instance is the basic unit for member management and role and permission allocation. It provides all DataArts Studio functions. Workspaces are allocated by branch or subsidiary (such as the group, subsidiary, and department), business domain (such as the procurement, production, and sales), or implementation environment (such as the development, test, and production environment). There are no fixed rules.

Admins can manage user (member) permissions, resources, and compute engines in a workspace. To enable users to work together, admins can add users to a workspace and assign the preset roles of DataArts Studio (admin, developer, operator, and visitor) to the users. Users other than admins can access

Management Center, DataArts Migration, DataArts Factory, only after they are added to a workspace and assigned relevant roles.

## Constraints

The storage of job logs and dirty data depends on the OBS service. If the OBS service is unavailable, job logs and dirty data cannot be stored.

## Prerequisites

A DataArts Studio instance has been create. For details, see [Creating a DataArts Studio Basic Package](#).

## Background

- After you a DataArts Studio instance, you have the permission to create workspaces. DataArts Studio creates a default workspace for you and assign the admin role to you.
- In a DataArts Studio instance created by a master account, if an IAM user of the account needs to create a workspace, the IAM user must be assigned the **DAYU Administrator** or **Tenant Administrator** policy. By default, the master account has all the permissions for a DataArts Studio instance created by a sub-user.
- After workspaces are created, they cannot be deleted. You can disable workspaces when they are no longer needed. You can enable them again when you need these workspaces.
- Users assigned the **DAYU User** policy can access a workspace only after you add them as members of the workspace.

## Creating a Workspace

1. Log in to the DataArts Studio console using the **DAYU Administrator** account.
2. Click the **Workspaces** tab.
3. Click **Create**. On the **Workspace Information** page, set the parameters based on [Table 2-3](#). After the configuration is complete, click **OK**.

**Table 2-3** Parameters for creating a workspace

Parameter	Description
Name	Workspace names can contain only letters, numbers, underscores (_), and hyphens (-). They cannot exceed 32 characters. In a DataArts Studio instance, workspace names must be unique.
Description	A description of the workspace.

Parameter	Description
Enterprise Project	<p>The enterprise project associated with your DataArts Studio instance.</p> <p>This parameter is available only when an enterprise project has been created. If you want to connect the DataArts Studio instance to a cloud service (such as DWS, MRS, and RDS), ensure that the enterprise project of the DataArts Studio instance is the same as that of the cloud service instance.</p> <ul style="list-style-type: none"><li>You can create only one DataArts Studio instance for an enterprise project.</li><li>If you want to enable communication between DataArts Studio and another cloud service, ensure that the enterprise project of DataArts Studio is the same as that of the cloud service.</li></ul>
OBS Bucket for Job Logs	<p>The OBS bucket for storing logs of DataArts Studio data development jobs. To use the DataArts Factory module of DataArts Studio, workspace members must have the read and write permissions on the OBS bucket for storing job logs. Otherwise, the system cannot read or write job logs generated during data development.</p> <ul style="list-style-type: none"><li>Click <b>Select</b>. You can select a created OBS bucket. The selected OBS bucket is globally configured in the current workspace.</li><li>If this parameter is not set, job logs generated during data development are stored in the OBS bucket named <b>dlf-log-{projectId}</b> by default. <b>{projectId}</b> indicates the project ID, which can be obtained by referring to <a href="#">Obtaining a project ID and account ID</a>.</li></ul>
OBS Bucket for DLI Dirty Data	<p>The OBS bucket for storing dirty data generated during DLI SQL execution in DataArts Studio Data Development. To use DataArts Studio DataArts Factory to develop and execute DLI SQL statements, workspace members must have the read and write permissions on the OBS bucket where DLI dirty data is stored. Otherwise, the system cannot read or write the dirty data generated during DLI SQL execution.</p> <ul style="list-style-type: none"><li>Click <b>Select</b>. You can select a created OBS bucket. The selected OBS bucket is globally configured in the current workspace.</li><li>If this parameter is not set, dirty data generated during DLI SQL execution is stored in the OBS bucket named <b>dlf-log-{projectId}</b> by default.</li></ul>

## Editing a Workspace

1. Log in to the DataArts Studio console.


2. Select a DataArts Studio instance and click **Access**. Then, click the **Workspaces** tab.
3. On the **Workspaces** tab page, locate the target workspace and click **Edit** in the row that contains the workspace.
4. Click **Edit** at the top of the **Workspace Information** page to edit the workspace information and manage workspace members.
5. After the configuration is complete, click **Save** at the top of the **Workspace Information** page to save the settings.

## Disabling a Workspace


After a workspace is created, it is enabled by default. You can disable or enable it as required.

### NOTE

If you disable a workspace, you cannot access the workspace, edit the workspace content, or schedule jobs in the workspace.

1. Log in to the DataArts Studio console.
2. Select a DataArts Studio instance and click **Access**. Then, click the **Workspaces** tab.
3. On the **Workspaces** tab page, locate the target workspace, and turn off  in the **Status** column.
4. In the **Disable Workspace** dialog box displayed, read the impact of disabling a workspace. Click **Yes** to disable the workspace.

## Enabling a Workspace

1. Log in to the DataArts Studio console.
2. Select a DataArts Studio instance and click **Access**. Then, click the **Workspaces** tab.
3. On the **Workspaces** tab page, locate the target workspace, and turn on  in the **Status** column.
4. In the **Enable Workspace** dialog box displayed, read the impact of enabling a workspace. If you want to continue, click **Yes** to enable the workspace.

## 2.3.2 (Optional) Changing the Job Log Storage Path

By default, job logs and Data Lake Insight (DLI) dirty data are stored in an Object Storage Service (OBS) bucket named **dlf-log-*{Project ID}***. You can customize a log storage path. DataArts Factory allows you to configure an OBS bucket globally based on the workspace.

## Constraints

This function depends on the OBS service.

## Prerequisites

To change the job log storage path, either of the following conditions must be met:

- You have administrator rights.
- You have been assigned the **DAYU User** policy and you are the admin of the current workspace.

## Procedure

1. Log in to the DataArts Studio console using the **DAYU Administrator** or administrator account.
2. Click the **Workspaces** tab.
3. Click **Edit** in the **Operation** column.
4. On the **Workspace Information** page displayed, click **Edit** next to the workspace information.  
Click the **Select** button next to **OBS Bucket for Job Logs** to reselect a path.

**Figure 2-2** Changing the log path

**Workspace Information** Edit

\* Name: cassie

Description: 0/255

Job Log Path ⓘ: obs://dlf-log-687df4b2bf0d424abaf43ebd2e Select

\* API Quota of DLM Exclusive

Used:	0 USD	
Allocated:	0 USD	<span>Edit</span>
Total used:	12 USD	
Total allocated:	312 USD	
Total:	5,000 USD	

**Workspace Members**

Add Remove Account Q

<input type="checkbox"/>	Account	Account ... ▾	Added	Role ▾	Oper...
<input type="checkbox"/>	User		Apr 13, 2021 18:08:1...	Admin	<span>Edit</span>

5. Click **Save**.

## 2.4 Authorizing Users to Use DataArts Studio

## 2.4.1 Creating an IAM User and Assigning DataArts Studio Permissions

Identity and Access Management (IAM) can be used for fine-grained permissions management on your DataArts Studio 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 DataArts Studio resources.
- Assign users only the permissions required to perform a task.
- Entrust a account or cloud service to perform efficient O&M on your DataArts Studio resources.

If you do not require individual IAM users for permissions management, skip this section.

### Background

- Before assigning permissions to a user group, familiarize yourself with the DataArts Studio workspace role permissions that can be added to the user group and select permissions based on actual requirements.

### Procedure

1. Create a user group and assign permissions to it. Log in to the IAM console using a account, create a user group, and grant permissions of a common user (for example, **DAYU User**) to the group.

For details, see "User Groups and Authorization" > "Creating a User Group and Assigning Permissions" in *Identity and Access Management User Guide*.

#### NOTE

- When configuring DataArts Studio permissions for a user group, enter **DAYU** in the search box to search for the permissions and select the permissions to be granted to the user group, for example, **DAYU User**.
  - If an IAM user wants to create a workspace, you must assign the IAM user the **DAYU Administrator** policy. Users with the **DAYU Administrator** policy can perform all operations on DataArts Studio.
  - DataArts Studio is a project-level service deployed in specific physical regions. If you select **All resources** for **Scope**, the permission takes effect in all projects of all regions. If you select **Region-specific projects** for **Scope**, the permission takes effect only for a specified project. When accessing DataArts Studio, the IAM user must switch to the region where they have been assigned the required permissions.
2. Create a user and add the user to the user group. Create a user on the IAM console and add the user to the group created in [1](#).

For details, see "IAM Users" > "Creating an IAM User" in *Identity and Access Management User Guide*.

## 2.4.2 Adding a Member and Assigning a Role

If you want to allow another IAM user to use your DataArts Studio instance, create an IAM user by referring to [Creating an IAM User and Assigning DataArts Studio Permissions](#), and add the user as the workspace member and configure a role for the user by following the instructions in this section.

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

- **Admin:** Users with this role have the permissions to perform all operations in a workspace. You are advised to assign this role to the project owner, development owner, and O&M administrator.
- **Developer:** Users with this role have the permissions to create and manage work items, but cannot perform operations on workspaces, clusters, and reviewers. You are advised to assign this role to users who develop and process tasks.
- **Operator:** Users with this role have the permissions to perform operations such as O&M and scheduling, but cannot modify work items or configurations. You are advised to assign this role to users for O&M management and status monitoring.
- **Viewer:** Users with this role can only read data from DataArts Studio, but cannot perform operations on workspaces or modify work items or configurations. You are advised to assign this role to users who only want to view information in the workspace but do not perform any operation.

## Background

The **DAYU Administrator** account or the administrator can add members to the workspace.

## Adding a Member and Assigning a Role

1. Log in to the DataArts Studio console and access the **Workspaces** page.
2. On the **Workspaces** page, locate the target workspace and click **Edit** in the **Operation** column.
3. Click **Add** under **Workspace Members**. In the displayed **Add Member** dialog box, select **Add User** or **Add Group**, select a member account from the drop-down list, and select a role for it.
4. Click **OK**. You can view or modify the members and roles in the member list, or remove members from the workspace.

## Removing a Workspace Member

1. Log in to the DataArts Studio console and access the **Workspaces** page.
2. On the **Workspaces** page, locate the target workspace and click **Edit** in the **Operation** column.
3. On the **Workspace information** page, select the target member and click **Remove**.

### NOTE

The creator of a workspace cannot be removed.

4. In the **Remove Member** dialog box displayed, click **Yes**.



## 2.5 (Optional) Obtaining Authentication Information

When creating OBS links, making API calls, or locating issues, you may need to obtain information such as access keys, project IDs, and endpoints. This section describes how to obtain such information.

### Obtaining an Access Key

To obtain an access key, perform the following steps:

Users in the AP-Kuala Lumpur-OP6 region are virtual users authorized by federated authentication to access the cloud system, rather than real users in the cloud system. Therefore, you need to contact the administrator to obtain the AK/SK in the **AP-Kuala Lumpur-OP6** region.

### Obtaining a project ID and account ID

A project ID indicates a tenant's resources, and an account ID corresponds to the current account. You can view the project IDs and account IDs in different regions on the corresponding pages.

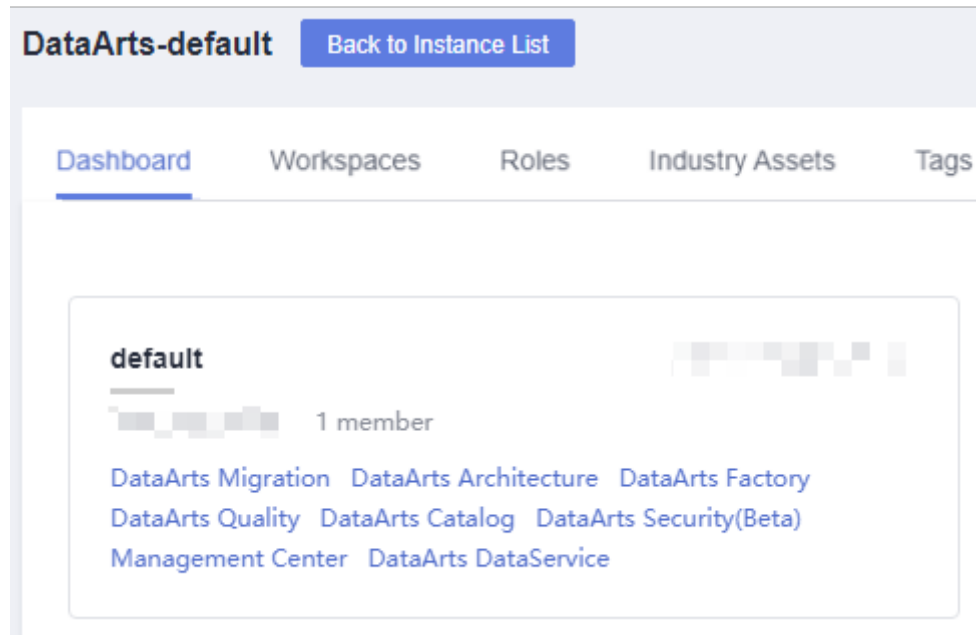
1. Register with and log in to the management console.
2. Hover the cursor on the username in the upper right corner and select **My Credentials** from the drop-down list.
3. On the **My Credentials** page, obtain the account name and account ID, and obtain the project ID from the project list.

### Obtaining the DataArts Studio Instance ID and Workspace ID

You can obtain the DataArts Studio instance ID and workspace ID from the URI of the DataArts Studio console.

1. On the homepage of the DataArts Studio console, locate a workspace and click a module, for example, **Management Center**.

**Figure 2-3** Management Center



2. On the **Management Center** page, obtain the values of **instanceId** and **workspace** in the browser address bar, which are the DataArts Studio instance ID and workspace ID, respectively.

As shown in **Figure 2-4**, the instance ID is **6b88...2688**, and the workspace ID is **1dd3bc...d93f0**.

**Figure 2-4** Obtaining the instance ID and workspace ID



## Obtaining an Endpoint

An endpoint is the **request address** for calling an API. Endpoints vary depending on services and regions.

**Table 2-4** DataArts Studio endpoint information

Region Name	Region Code	Component	Endpoint	Protocol
AP-Kuala Lumpur-OP6	my-kualalumpur-1	DataArts Migration	cdm.my-kualalumpur-1.alphaedge.tmone.com.my	HTTPS/HTTP
		DataArts Factory	dayu-dlf.my-kualalumpur-1.alphaedge.tmone.com.my	

# 3 User Guide

---

## 3.1 Preparations Before Using DataArts Studio

Before using DataArts Studio, you must conduct data and business surveys and select an appropriate data governance model.

Then, make the following preparations by referring to this topic:

- [DataArts Studio Preparations](#)
- [Preparing a Data Source](#)
- [Preparing a Data Lake](#)

### DataArts Studio Preparations

If you use DataArts Studio for the first time, create a DataArts Studio instance and create a workspace by following the instructions provided in section "Preparations" in the *DataArts Studio User Guide*. Then you can develop and operate data in the workspace.

### Preparing a Data Source

Many on-premises data sources are of MySQL, PostgreSQL, HBase, and Hive type. Therefore, you need to make the following preparations:

- The host where the data source is located can access the public network.
- Obtain the public network IP address, database port, and the administrator username and password for accessing the databases.
- Ensure that the database port is enabled in the outbound direction of the firewall rule to allow data to be migrated to the cloud.

After the data source is prepared, you can migrate the data source to the data lake by using data integration, and then perform data development, governance, and operations using DataArts Studio.

### Preparing a Data Lake

Before using DataArts Studio, select a cloud service as the data lake. The data lake stores raw data and data generated during data development and is used for

subsequent data development, services, and operations. For details on the data lake products supported by DataArts Studio, see [Data Sources](#).

After the data lake is prepared, you can [create a data connection](#) to connect DataArts Studio to the data lake and then perform [1](#) and [2](#). For details about the operations in [1](#) and [2](#), see "Step 2: Preparations" in *Getting Started*.

### 1. Creating a Database

Before using DataArts Migration to migrate your data to the cloud, create a destination database in the data lake. According to the implementation process of data lake governance, you are advised to create a database for each of the SDI layer, DWI layer, DWR layer, and DM layer in the data lake to implement hierarchical sharding. Data sharding is a concept involved in DataArts Architecture. You will know more about it during architecture design.

You can create a database in the data lake using either of the following methods:

- You can create a database on the DataArts Factory console of DataArts Studio. For details, see [DataArts Factory > Data Management > Creating a Database](#).
- You can also develop and execute a SQL script for creating a database in the SQL editor of DataArts Studio's DataArts Factory module or a data lake product, and then use the script to create a database. For details about how to develop a script, see [DataArts Factory > Script Development > Developing Scripts > Developing an SQL Script](#).

### 2. Creating a Data Table

Before using DataArts Migration to migrate your data to the cloud, create a destination table in the SDI layer database of the destination data lake to store raw data. During batch data migration, a destination table can be automatically created for the migration between relational databases and from a relational database to Hive. In this case, you do not need to create the destination table in the destination database in advance.

You can create a table in the data lake using either of the following methods: If a table contains a large number of fields, you are advised to create the table by compiling SQL scripts.

- You can create a table on the DataArts Factory console of DataArts Studio. For details, see [DataArts Factory > Data Management > Creating a Table](#).
- You can also develop and execute a SQL script for creating a table in the SQL editor of DataArts Studio's DataArts Factory module or a data lake product, and then use the script to create a table. For details about how to develop a script, see [DataArts Factory > Script Development > Developing Scripts > Developing an SQL Script](#).

## 3.2 Management Center

DataArts Studio Management Center provides a unified configuration and management entry for data connections and resource migration. Personalized entries and showcases can be customized as needed.

### 3.2.1 Data Sources

Before using DataArts Studio, select a cloud service or data warehouse as the data lake. The data lake stores raw data and data generated during data governance and serves for data development, data services, and data operations. DataArts Studio integrates a wealth of data engines and can connect to cloud data lakes and database cloud services, such as Data Warehouse Service (DWS), Data Lake Insight (DLI), and MapReduce Service (MRS) Hive. It can also connect to traditional enterprise databases, such as MySQL and PostgreSQL.

#### Data Sources Supported By DataArts Studio

Data sources supported by DataArts Studio are classified into data sources supported by DataArts Migration and those supported by other DataArts Studio components.

- The DataArts Migration component integrates data into the data lake and supports more types of data sources.

For details on data sources supported by DataArts Migration, see [Data Sources Supported by DataArts Migration](#). To use these data sources, you must create corresponding data links in DataArts Migration, which can only be used in the DataArts Migration module.

- The data sources supported by other DataArts Studio components form the data lake base of DataArts Studio.

[Table 3-1](#) lists the data sources supported by other components. For details, see [Overview](#). To use these data sources in other components, create data connections on the DataArts Studio Management Center console. These data connections cannot be used in the DataArts Migration module.

**Table 3-1** Data sources supported by other DataArts Studio components

Data Source Type	Management Center	DataArts Factory
DWS	Supported	Supported
DLI	Supported	Supported
MRS HBase	Supported	Not supported
MapReduce (MRS) Hive	Supported	Supported
MRS Kafka	Supported	Supported
MySQL	Supported	Not supported
MapReduce (MRS) Spark	Supported	Supported
RDS for MySQL	Supported	Supported
RDS for PostgreSQL	Supported	Supported
Host Connection	Supported	Supported
MapReduce (MRS) Presto	Supported	Supported

## Overview

**Table 3-2** Data source overview

Data Source Type	Description
DWS	DWS employs the shared-nothing architecture and massively parallel processing (MPP) engine. It is compatible with ANSI SQL 99, SQL 2003, and the PostgreSQL or Oracle database ecosystem, providing competitive solutions for analyzing petabytes of data in various industries.
DLI	DLI is a serverless big data compute and analysis service that is fully compatible with Apache Spark and Apache Flink ecosystems. With multi-model engines supported by DLI, enterprises can use SQL statements or programs to easily complete batch processing, stream processing, in-memory computing, and machine learning of heterogeneous data sources.
MRS HBase	<p>HBase undertakes data storage. It is an open-source, column-oriented, distributed storage system that is suitable for storing massive amounts of unstructured or semi-structured data. It features high reliability, high performance, and flexible scalability, and supports real-time data read/write.</p> <p>MRS HBase stores massive amount of data and supports data queries in milliseconds. MRS HBase can load and update logistics data in milliseconds, and query and analyze petabytes of time series data in seconds.</p>
MRS Hive	<p>Hive is a mechanism that can store, query, and analyze large-scale data stored in Hadoop. Hive defines simple SQL-like query language, which is known as HiveQL. It allows users familiar with SQL to query data.</p> <p>MRS Hive can be used to analyze terabytes or petabytes of data and quickly migrate on-premises Hadoop big data platforms (such as CDH and HDP) to the cloud without service interruption and service code modification.</p>

Data Source Type	Description
MRS Kafka	<p>MRS provides dedicated MRS Kafka clusters. Kafka is an open-source, distributed, partitioned, and replicated commit log service. Kafka is publish-subscribe messaging, rethought as a distributed commit log. It provides features similar to Java Message Service (JMS) but another design. It features message endurance, high throughput, distributed methods, multi-client support, and real time. It applies to both online and offline message consumption, such as regular message collection, website activeness tracking, aggregation of statistical system operation data (monitoring data), and log collection. These scenarios engage large amounts of data collection for Internet services.</p>
MySQL	<p>MySQL is one of the most popular open-source databases. It features excellent performance, uses mature and stable architecture, supports popular applications, adapts to multiple fields and industries, and supports various web applications. It is cost-effective and preferred by small- and medium-sized enterprises.</p>
MRS Spark	<p>Spark is an open-source parallel data processing framework. It helps users easily develop unified big data applications and perform cooperative processing, stream processing, and interactive analysis on data.</p> <p>Spark provides a framework featuring fast calculation, write, and interactive query. Spark has obvious advantages over Hadoop in terms of performance. Spark provides the Spark SQL language similar to SQL statements to process structured data.</p>
RDS	<p>RDS is an online, out-of-the-box relational database service that is based on the cloud computing platform. It is stable, reliable, scalable, and easy to manage.</p> <p>Currently, DataArts Studio supports only MySQL and PostgreSQL databases in RDS.</p>
Host Connection	<p>You can connect to a specified host during data development and execute shell or Python scripts on the host through script development and job development. If the host connection information changes, you only need to edit it on the <b>Host Connections</b> page, but do not need to edit it in scripts or jobs one by one.</p>

Data Source Type	Description
MRS Presto	<p>Presto is an open-source SQL query engine for running interactive analytic queries against data sources of all sizes. It applies to massive structured/semi-structured data analysis, massive multi-dimensional data aggregation/report, ETL, ad-hoc queries, and more scenarios.</p> <p>Presto allows querying data where it lives, including HDFS, Hive, HBase, Cassandra, relational databases, or even proprietary data stores. A Presto query can combine different data sources to perform data analysis across the data sources.</p>

## 3.2.2 Creating Data Connections

You can create data connections by configuring data sources. Based on the data connections of the Management Center, DataArts Studio performs data development, governance, services, and operations on the data lake base.

### Constraints

- RDS data connections depend on OBS. If OBS is unavailable in the same region as DataArts Studio, RDS data connections are not supported.
- If changes occur in the connected data lake (for example, the MRS cluster capacity is expanded), you need to edit and save the connection.

### Prerequisites

- You have created a data lake to connect, for example, a database or cloud service supported by DataArts Studio.
  - Before creating a DWS data connection, ensure that you have created a cluster in DWS and have the permissions required to view Key Management Service (KMS) keys.
  - Before creating an MRS HBase, MRS Hive, MRS Kafka, MRS Presto, or MRS Spark connection, ensure that you have created an MRS cluster and selected required components.
  - Before creating an RDS data connection, ensure that you have created an RDS DB instance. Currently, DataArts Studio supports only MySQL and PostgreSQL databases in RDS.
- The data lake to connect communicates with the DataArts Studio instance properly.
  - If the data lake is an on-premises database, a public network or a dedicated connection is required. Ensure that the host where the data source is located can access the public network and the port has been enabled in the firewall rule.
  - If the data lake is a cloud service (such as DWS and MRS), the following requirements must be met for network interconnection:

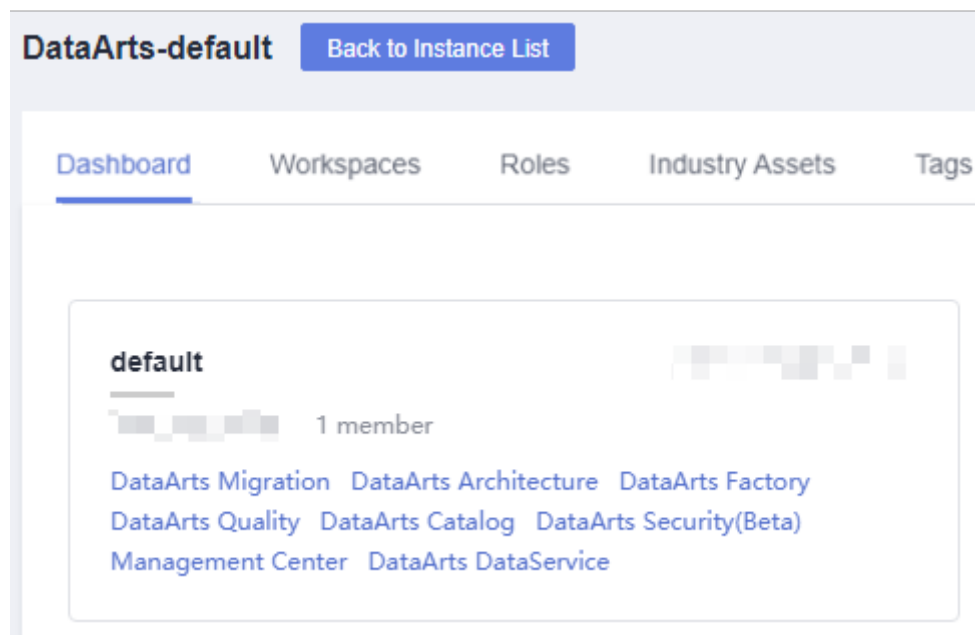


- If the CDM cluster in the DataArts Studio instance and the cloud service are in different regions, a public network or a dedicated connection is required.
- If the CDM cluster in the DataArts Studio instance and the cloud service are in the same region, VPC, subnet, and security group, they can communicate 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 **Adding Routes** in *Virtual Private Cloud (VPC) Usage Guide*. For details about how to configure security group rules, see **Security Group > Adding a Security Group Rule** in *Virtual Private Cloud (VPC) Usage Guide*.
- The cloud service instance and the DataArts Studio workspace belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

## Creating a Data Connection

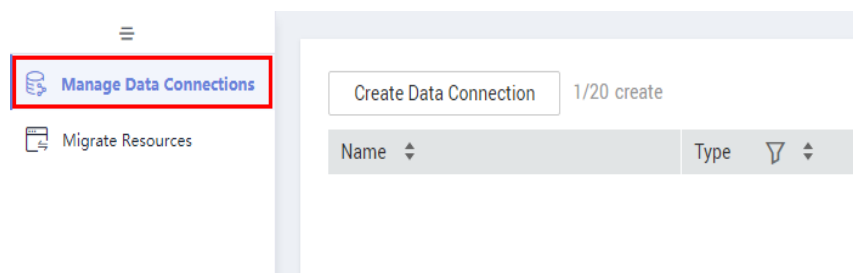
1. On the DataArts Studio console, locate a workspace and click **Management Center**.

Figure 3-1 Management Center



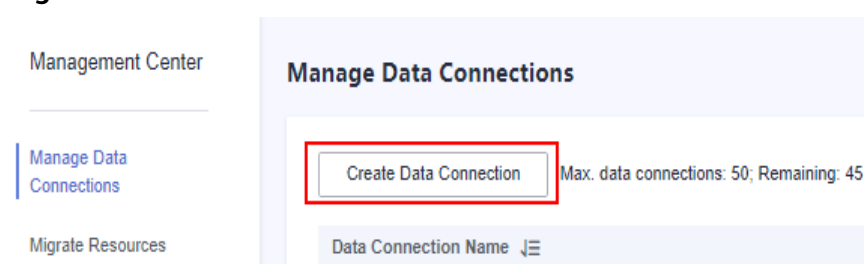
2. In the navigation pane, choose **Manage Data Connections**.

**Figure 3-2** Manage Data Connections



3. On the **Data Connection Management** page, click **Create Data Connection**. Select a data connection type and set the relevant parameters. See [Table 3-3](#).

**Figure 3-3** Create Data Connection



**Table 3-3** Data connections

Data Connection Type	Link
MRS Hive	<a href="#">Table 3-4</a>
MRS HBase	<a href="#">Table 3-5</a>
MRS Kafka	<a href="#">Table 3-6</a>
DWS	<a href="#">Table 3-9</a>
ORACLE	<a href="#">Table 3-10</a>
MRS Spark	<a href="#">Table 3-7</a>
RDS	See <a href="#">Table 3-8</a> . You can also create RDS connections to relational databases, such as MySQL, PostgreSQL, and Dameng databases.
Host Connection	See <a href="#">Table 3-11</a> .

4. Click **Test** to test connectivity of the data connection. If the test passes, the data connection is created.
5. After the test is successful, click **OK**. The system will create the data connection for you.

## Data Connection Parameter Description

**Table 3-4** MRS Hive data connection

Parameter	Mandatory	Description
Data Connection Name	Yes	The name of the data connection to create. Data connection names can contain 1 to 50 characters. They can include only letters, numbers, underscores (_), and hyphens (-).
Tag	No	<p>The attribute of the data connection to create. Tags make management easier.</p> <p><b>NOTE</b> The name of the tag. Only letters, numbers, and underscores (_) are allowed. Tag names cannot start with underscores (_). Enter up to 100 characters.</p>

Parameter	Mandatory	Description
Cluster Name	Yes	<p>The name of the MRS Hive cluster. Select an MRS cluster that Hive belongs to. If the MRS cluster is not displayed in the drop-down list, check whether the network connection between the MRS cluster and the DataArts Studio instance is normal.</p> <p>Ensure that the MRS cluster and the DataArts Studio instance can communicate with each other. The following requirements must be met for network interconnection:</p> <ul style="list-style-type: none"> <li>• 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.</li> <li>• 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 <b>Custom Route in Region Type I &gt; Adding Routes</b> in <i>Virtual Private Cloud (VPC) Usage Guide</i>. For details about how to configure security group rules, see <b>Security Group &gt; Adding a Security Group Rule</b> in <i>Virtual Private Cloud (VPC) Usage Guide</i>.</li> <li>• 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.</li> </ul>

Parameter	Mandatory	Description
Connection Type	Yes	<p>Connection type. <b>Proxy connection</b> is recommended.</p> <ul style="list-style-type: none"> <li>• <b>Proxy connection:</b> An agent (CDM cluster) is used to access MRS clusters. This method supports all versions of MRS clusters.</li> <li>• <b>MRS API connection:</b> MRS APIs are used to access MRS clusters. This method supports only MRS clusters of the 2.X or a later version. When you select <b>MRS API connection</b>, pay attention to the following restrictions: <ol style="list-style-type: none"> <li>1. Tables and fields cannot be viewed.</li> <li>2. When the SQL editor is used to run SQL statements, the execution results can be displayed only in logs.</li> <li>3. This method is not supported by data governance functions such as DataArts Architecture, DataArts Quality, and DataArts Catalog.</li> </ol> </li> </ul>
Username	No	<p>The username of the MRS cluster. This parameter is mandatory when <b>Connection Type</b> is set to <b>Proxy connection</b>. If a new MRS user is used for connection, you need to log in to Manager and change the initial password.</p> <p>You cannot create a data connection for an MRS security cluster as user <b>admin</b>. User <b>admin</b> is the management page user by default and cannot be used as the authentication user of a security cluster. You can create an MRS user by referring to <a href="#">Creating a Kerberos Authentication User for an MRS Security Cluster</a>.</p> <p>When creating an MRS data connection, set <b>Username</b> and <b>Password</b> to the new MRS username and password.</p> <p><b>NOTE</b></p> <ul style="list-style-type: none"> <li>• For clusters of MRS 3.1.0 or later, the user must at least have permissions of the <b>Manager_viewer</b> role to create data connections in Management Center. To perform database, table, and data operations on components, the user must also have user group permissions of the components.</li> <li>• For clusters earlier than MRS 3.1.0, the user must have permissions of the <b>Manager_administrator</b> or <b>System_administrator</b> role to create data connections in Management Center.</li> <li>• A user with only the <b>Manager_tenant</b> or <b>Manager_auditor</b> permission cannot create connections.</li> </ul>
Password	No	<p>The password for accessing the MRS cluster. This parameter is mandatory when <b>Connection Type</b> is set to <b>Proxy connection</b>.</p>

Parameter	Mandatory	Description
KMS Key	No	The name of the KMS key. This parameter is mandatory when <b>Connection Type</b> is set to <b>Proxy connection</b> .
Agent	No	<p>This parameter is mandatory when <b>Connection Type</b> is set to <b>Proxy connection</b>.</p> <p>MRS is not a fully managed service and cannot be directly connected to DataArts Studio. A CDM cluster can provide an agent for DataArts Studio to communicate with non-fully-managed services. Therefore, you need to select a CDM cluster when creating an MRS data connection. If no CDM cluster is available, create one through the DataArts Migration incremental package.</p> <p>As a network proxy, the CDM cluster must be able to communicate with the MRS cluster. To ensure network connectivity, the CDM cluster must be in the same region, AZ, VPC, and subnet as the MRS cluster. The security group rule must also allow the CDM cluster communicate with the MRS cluster.</p>

**Table 3-5** MRS HBase data connection

Parameter	Mandatory	Description
Data Connection Name	Yes	The name of the data connection to create. Data connection names can contain 1 to 50 characters. They can include only letters, numbers, underscores (_), and hyphens (-).
Tag	No	<p>The attribute of the data connection to create. Tags make management easier.</p> <p><b>NOTE</b> The name of the tag. Only letters, numbers, and underscores (_) are allowed. Tag names cannot start with underscores (_). Enter up to 100 characters.</p>

Parameter	Mandatory	Description
Cluster Name	Yes	<p>The name of the MRS HBase cluster. Select an MRS cluster that HBase belongs to. If the MRS cluster is not displayed in the drop-down list, check whether the network connection between the MRS cluster and the DataArts Studio instance is normal.</p> <p>Ensure that the MRS cluster and the DataArts Studio instance can communicate with each other. The following requirements must be met for network interconnection:</p> <ul style="list-style-type: none"> <li>• 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.</li> <li>• 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 <b>Custom Route in Region Type I &gt; Adding Routes</b> in <i>Virtual Private Cloud (VPC) Usage Guide</i>. For details about how to configure security group rules, see <b>Security Group &gt; Adding a Security Group Rule</b> in <i>Virtual Private Cloud (VPC) Usage Guide</i>.</li> <li>• 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.</li> </ul>

Parameter	Mandatory	Description
Username	Yes	<p>Username of the MRS cluster</p> <p>You cannot create a data connection for an MRS security cluster as user <b>admin</b>. User <b>admin</b> is the management page user by default and cannot be used as the authentication user of a security cluster. You can create an MRS user by referring to <a href="#">Creating a Kerberos Authentication User for an MRS Security Cluster</a>. When creating an MRS data connection, set <b>Username</b> and <b>Password</b> to the new MRS username and password.</p> <p><b>NOTE</b></p> <ul style="list-style-type: none"> <li>For clusters of MRS 3.1.0 or later, the user must at least have permissions of the <b>Manager_viewer</b> role to create data connections in Management Center. To perform database, table, and data operations on components, the user must also have user group permissions of the components.</li> <li>For clusters earlier than MRS 3.1.0, the user must have permissions of the <b>Manager_administrator</b> or <b>System_administrator</b> role to create data connections in Management Center.</li> <li>A user with only the <b>Manager_tenant</b> or <b>Manager_auditor</b> permission cannot create connections.</li> </ul>
Password	Yes	Password for accessing the MRS cluster.
KMS Key	Yes	Name of the KMS key.
Agent	Yes	<p>MRS is not a fully managed service and cannot be directly connected to DataArts Studio. A CDM cluster can provide an agent for DataArts Studio to communicate with non-fully-managed services. Therefore, you need to select a CDM cluster when creating an MRS data connection. If no CDM cluster is available, create one through the DataArts Migration incremental package.</p> <p>As a network proxy, the CDM cluster must be able to communicate with the MRS cluster. To ensure network connectivity, the CDM cluster must be in the same region, AZ, VPC, and subnet as the MRS cluster. The security group rule must also allow the CDM cluster communicate with the MRS cluster.</p>



**Table 3-6** MRS Kafka data connection

Parameter	Man dato ry	Description
Data Connection Name	Yes	The name of the data connection to create. Data connection names can contain 1 to 50 characters. They can include only letters, numbers, underscores (_), and hyphens (-).
Tag	No	The attribute of the data connection to create. Tags make management easier. <b>NOTE</b> The name of the tag. Only letters, numbers, and underscores (_) are allowed. Tag names cannot start with underscores (_). Enter up to 100 characters.
Cluster Name	Yes	The name of the MRS Kafka cluster. Select an MRS cluster that Kafka belongs to. If the MRS cluster is not displayed in the drop-down list, check whether the network connection between the MRS cluster and the DataArts Studio instance is normal. Ensure that the MRS cluster and the DataArts Studio instance can communicate with each other. The following requirements must be met for network interconnection: <ul style="list-style-type: none"> <li>• 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.</li> <li>• 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 <b>Custom Route in Region Type I &gt; Adding Routes</b> in <i>Virtual Private Cloud (VPC) Usage Guide</i>. For details about how to configure security group rules, see <b>Security Group &gt; Adding a Security Group Rule</b> in <i>Virtual Private Cloud (VPC) Usage Guide</i>.</li> <li>• 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.</li> </ul>

Parameter	Mandatory	Description
Username	Yes	<p>Username of the MRS cluster</p> <p>You cannot create a data connection for an MRS security cluster as user <b>admin</b>. User <b>admin</b> is the management page user by default and cannot be used as the authentication user of a security cluster. You can create an MRS user by referring to <a href="#">Creating a Kerberos Authentication User for an MRS Security Cluster</a>. When creating an MRS data connection, set <b>Username</b> and <b>Password</b> to the new MRS username and password.</p> <p><b>NOTE</b></p> <ul style="list-style-type: none"> <li>For clusters of MRS 3.1.0 or later, the user must at least have permissions of the <b>Manager_viewer</b> role to create data connections in Management Center. To perform database, table, and data operations on components, the user must also have user group permissions of the components.</li> <li>For clusters earlier than MRS 3.1.0, the user must have permissions of the <b>Manager_administrator</b> or <b>System_administrator</b> role to create data connections in Management Center.</li> <li>A user with only the <b>Manager_tenant</b> or <b>Manager_auditor</b> permission cannot create connections.</li> </ul>
Password	Yes	Password for accessing the MRS cluster.
KMS Key	Yes	Name of the KMS key.
Agent	Yes	<p>MRS is not a fully managed service and cannot be directly connected to DataArts Studio. A CDM cluster can provide an agent for DataArts Studio to communicate with non-fully-managed services. Therefore, you need to select a CDM cluster when creating an MRS data connection. If no CDM cluster is available, create one through the DataArts Migration incremental package.</p> <p>As a network proxy, the CDM cluster must be able to communicate with the MRS cluster. To ensure network connectivity, the CDM cluster must be in the same region, AZ, VPC, and subnet as the MRS cluster. The security group rule must also allow the CDM cluster communicate with the MRS cluster.</p>

**Table 3-7** MRS Spark data connection

Parameter	Mandatory	Description
Data Connection Name	Yes	The name of the data connection to create. Data connection names can contain 1 to 50 characters. They can include only letters, numbers, underscores (_), and hyphens (-).
Tag	No	The attribute of the data connection to create. Tags make management easier. <b>NOTE</b> The name of the tag. Only letters, numbers, and underscores (_) are allowed. Tag names cannot start with underscores (_). Enter up to 100 characters.
Cluster Name	Yes	The name of the MRS Spark cluster. Select an MRS cluster that Spark belongs to. If the MRS cluster is not displayed in the drop-down list, check whether the network connection between the MRS cluster and the DataArts Studio instance is normal. Ensure that the MRS cluster and the DataArts Studio instance can communicate with each other. The following requirements must be met for network interconnection: <ul style="list-style-type: none"> <li>• 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.</li> <li>• 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 <b>Custom Route in Region Type I &gt; Adding Routes</b> in <i>Virtual Private Cloud (VPC) Usage Guide</i>. For details about how to configure security group rules, see <b>Security Group &gt; Adding a Security Group Rule</b> in <i>Virtual Private Cloud (VPC) Usage Guide</i>.</li> <li>• 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.</li> </ul>

Parameter	Mandatory	Description
Connection Type	Yes	<p>Connection type. <b>Proxy connection</b> is recommended.</p> <ul style="list-style-type: none"> <li>• <b>Proxy connection:</b> An agent (CDM cluster) is used to access MRS clusters. This method supports all versions of MRS clusters.</li> <li>• <b>MRS API connection:</b> MRS APIs are used to access MRS clusters. This method supports only MRS clusters of the 2.X or a later version. When you select <b>MRS API connection</b>, pay attention to the following restrictions: <ol style="list-style-type: none"> <li>1. Tables and fields cannot be viewed.</li> <li>2. When the SQL editor is used to run SQL statements, the execution results can be displayed only in logs.</li> <li>3. This method is not supported by data governance functions such as DataArts Architecture, DataArts Quality, and DataArts Catalog.</li> </ol> </li> </ul>
Username	No	<p>The username of the MRS cluster. This parameter is mandatory when <b>Connection Type</b> is set to <b>Proxy connection</b>. If a new MRS user is used for connection, you need to log in to Manager and change the initial password.</p> <p>You cannot create a data connection for an MRS security cluster as user <b>admin</b>. User <b>admin</b> is the management page user by default and cannot be used as the authentication user of a security cluster. You can create an MRS user by referring to <a href="#">Creating a Kerberos Authentication User for an MRS Security Cluster</a>. When creating an MRS data connection, set <b>Username</b> and <b>Password</b> to the new MRS username and password.</p> <p><b>NOTE</b></p> <ul style="list-style-type: none"> <li>• For clusters of MRS 3.1.0 or later, the user must at least have permissions of the <b>Manager_viewer</b> role to create data connections in Management Center. To perform database, table, and data operations on components, the user must also have user group permissions of the components.</li> <li>• For clusters earlier than MRS 3.1.0, the user must have permissions of the <b>Manager_administrator</b> or <b>System_administrator</b> role to create data connections in Management Center.</li> <li>• A user with only the <b>Manager_tenant</b> or <b>Manager_auditor</b> permission cannot create connections.</li> </ul>

Parameter	Mandatory	Description
Password	No	The password for accessing the MRS cluster. This parameter is mandatory when <b>Connection Type</b> is set to <b>Proxy connection</b> .
KMS Key	No	The name of the KMS key. This parameter is mandatory when <b>Connection Type</b> is set to <b>Proxy connection</b> .
Agent	No	<p>This parameter is mandatory when <b>Connection Type</b> is set to <b>Proxy connection</b>.</p> <p>MRS is not a fully managed service and cannot be directly connected to DataArts Studio. A CDM cluster can provide an agent for DataArts Studio to communicate with non-fully-managed services. Therefore, you need to select a CDM cluster when creating an MRS data connection. If no CDM cluster is available, create one through the DataArts Migration incremental package.</p> <p>As a network proxy, the CDM cluster must be able to communicate with the MRS cluster. To ensure network connectivity, the CDM cluster must be in the same region, AZ, VPC, and subnet as the MRS cluster. The security group rule must also allow the CDM cluster to communicate with the MRS cluster.</p>

**Table 3-8** RDS data connection

Parameter	Mandatory	Description
Data Connection Name	Yes	The name of the data connection to create. Data connection names can contain 1 to 50 characters. They can include only letters, numbers, underscores (_), and hyphens (-).
Tag	No	<p>The attribute of the data connection to create. Tags make management easier.</p> <p><b>NOTE</b> The name of the tag. Only letters, numbers, and underscores (_) are allowed. Tag names cannot start with underscores (_). Enter up to 100 characters.</p>

Parameter	Mandatory	Description
IP Address	Yes	<p>The address for accessing RDS.</p> <p>If the data source is RDS, you can obtain the address from the RDS console.</p> <ol style="list-style-type: none"> <li>1. Log in to the management console using the created account.</li> <li>2. In the <b>Service List</b>, choose <b>Relational Database Service</b>. In the left navigation pane, choose <b>Instances</b>.</li> <li>3. Click the name of an instance. The basic information page of the instance is displayed.</li> </ol> <p>You can obtain the IP address on the <b>Connection Information</b> tab.</p>
Port	Yes	<p>The port for accessing RDS.</p> <p>If the data source is RDS, you can obtain the port from the RDS console.</p> <ol style="list-style-type: none"> <li>1. Log in to the management console using the account.</li> <li>2. In the <b>Service List</b>, choose <b>Relational Database Service</b>. In the left navigation pane, choose <b>Instances</b>.</li> <li>3. Click the name of an instance. The basic information page of the instance is displayed.</li> </ol> <p>You can obtain the database port on the <b>Connection Information</b> tab.</p>
Driver Name	Yes	<p>The name of the driver. The following values are available:</p> <ul style="list-style-type: none"> <li>• com.mysql.jdbc.Driver</li> <li>• org.postgresql.Driver</li> </ul>
Driver File Path	Yes	<p>Path of the driver file in the OBS bucket. You need to download the .jar driver file from the corresponding official website and upload it to the OBS bucket.</p> <ul style="list-style-type: none"> <li>• MySQL driver: Download it from <a href="https://downloads.mysql.com/archives/c-j/">https://downloads.mysql.com/archives/c-j/</a>. The 5.1.48 version is recommended.</li> <li>• PostgreSQL driver: Download it from <a href="https://jdbc.postgresql.org/download">https://jdbc.postgresql.org/download</a>. The 42.1.4 version is recommended.</li> </ul> <p><b>NOTE</b> To update the driver, you must restart the CDM cluster in DataArts Migration and then edit the data connection to upload the driver.</p>

Parameter	Man dato ry	Description
Username	Yes	The username of the database. The username is required for creating a cluster.
Password	Yes	The password for accessing the database. The password is required for creating a cluster.
KMS Key	Yes	The name of the KMS key. To obtain the key: 1. Log in to the management console using the account. 2. Click <b>Key Management Service</b> and select <b>Key Management Service</b> from the list on the left. You can obtain the key name from the key list.
Agent	Yes	RDS is not a fully managed service and cannot be directly connected to DataArts Studio. A CDM cluster can provide an agent for DataArts Studio to communicate with non-fully-managed services. Therefore, you need to select a CDM cluster when creating an RDS data connection. If no CDM cluster is available, create one through the DataArts Migration incremental package.  As a network proxy, the CDM cluster must be able to communicate with RDS. To ensure network connectivity, the CDM cluster must be in the same region, AZ, VPC, and subnet as RDS. The security group rule must also allow the CDM cluster to communicate with RDS.

**Table 3-9** DWS data connection

Parameter	Man dator y	Description
Data Connection Name	Yes	The name of the data connection to create. Data connection names can contain 1 to 50 characters. They can include only letters, numbers, underscores (_), and hyphens (-).
Tag	No	The attribute of the data connection to create. Tags make management easier.  <b>NOTE</b> The name of the tag. Only letters, numbers, and underscores (_) are allowed. Tag names cannot start with underscores (_). Enter up to 100 characters.

Parameter	Mandatory	Description
Manual	Yes	You can turn off or turn on to disable or enable the <b>Manual</b> function. <ul style="list-style-type: none"> <li>When <b>Manual</b> is disabled, you do not need to enter the IP address and port.</li> <li>When <b>Manual</b> is enabled, you must enter the IP address and port.</li> </ul>
IP Address	No	The IP address for accessing the cluster database through the internal network. This parameter is mandatory when <b>Manual</b> is enabled. The private network address is automatically generated when you create a cluster.
Port	No	The database port specified during DWS cluster creation. This parameter is mandatory when <b>Manual</b> is enabled. Ensure that you have enabled this port in the security group rule so that the DataArts Studio instance can connect to the database in the DWS cluster through this port.
SSL Connection	Yes	DWS supports SSL encryption and certificate authentication for communication between the client and server. You can use <b>SSL Connection</b> to set the communication mode. If <b>SSL Connection</b> is enabled, only SSL encryption can be used. If <b>SSL Connection</b> is disabled, both modes can be used. <b>SSL Connection</b> is disabled by default.
Cluster Name	Yes	The name of the selected DWS cluster.
Username	Yes	The database username, which is specified when the DWS cluster is created.
Password	Yes	The password for accessing the database, which is specified when the DWS cluster is created.
KMS Key	Yes	The name of the KMS key.
Connection Type	Yes	Connection type. <b>Proxy connection</b> is recommended. <ul style="list-style-type: none"> <li><b>Proxy connection:</b> An agent (CDM cluster) is used to access DWS clusters.</li> <li><b>Direct connection:</b> You can access DWS clusters directly.</li> </ul>



Parameter	Mandatory	Description
Agent	No	<p>This parameter is mandatory when <b>Connection Type</b> is set to <b>Proxy connection</b>.</p> <p>Data Warehouse Service (DWS) is not a fully managed service and thus cannot be directly connected to DataArts Studio. A CDM cluster can provide an agent for DataArts Studio to communicate with non-fully-managed services. Therefore, you need to select a CDM cluster when creating a DWS data connection. If no CDM cluster is available, create one through the DataArts Migration incremental package.</p> <p>As a network proxy, the CDM cluster must be able to communicate with the DWS cluster. To ensure network connectivity, the CDM cluster must be in the same region, AZ, VPC, and subnet as the DWS cluster. The security group rule must also allow the CDM cluster communicate with the DWS cluster.</p>

**Table 3-10** Oracle data connection

Parameter	Mandatory	Description
Data Connection Name	Yes	The name of the data connection to create. Data connection names can contain 1 to 50 characters. They can include only letters, numbers, underscores (_), and hyphens (-).
Tag	No	<p>The attribute of the data connection to create. Tags make management easier.</p> <p><b>NOTE</b> The name of the tag. Only letters, numbers, and underscores (_) are allowed. Tag names cannot start with underscores (_). Enter up to 100 characters.</p>
ip	Yes	The IP address of the database to connect. Both public and private IP addresses are supported.
Port	Yes	The port of the database to connect.

Parameter	Mandatory	Description
Username	Yes	<p>The username of the account for accessing the database. This account must have the permissions required to read and write data tables and metadata.</p> <p><b>NOTE</b> If you have the CONNECT permission (read-only permission) and are trying to create a connection, a message is displayed indicating that the table or schema does not exist. In this case, perform the following operations to grant permissions:</p> <ol style="list-style-type: none"> <li>1. Log in to the Oracle node as user <b>root</b>.</li> <li>2. Run the following command to switch to user <b>oracle</b>: <b>su oracle</b></li> <li>3. Run the following command to log in to the database: <b>sqlplus /nolog</b></li> <li>4. Run the following command to log in as user <b>sys</b>: <b>connect sys as sysdba;</b> Enter the password of user <b>sys</b> .</li> <li>5. Run the following SQL statement to grant permissions: <b>GRANT SELECT ON GV_\$INSTANCE to xxx;</b> In the preceding command, <i>xxx</i> indicates the name of the user to which the permissions will be granted.</li> </ol>
Password	Yes	The user password.
sid	Yes	The unique identifier of the Oracle database.
KMS Key	Yes	<p>The name of the KMS key.</p> <p>To obtain the key:</p> <ol style="list-style-type: none"> <li>1. Log in to the management console using the created account.</li> <li>2. Click <b>Key Management Service</b> and select <b>Key Management Service</b> from the list on the left.</li> </ol> <p>You can obtain the key name from the key list.</p>
Agent	Yes	<p>Oracle is not a fully managed service and cannot be directly connected to DataArts Studio. A CDM cluster can provide an agent for DataArts Studio to communicate with non-fully-managed services. Therefore, you need to select a CDM cluster when creating an Oracle data connection. If no CDM cluster is available, create one through the DataArts Migration incremental package.</p> <p>As a network proxy, the CDM cluster must be able to communicate with Oracle.</p>

**Table 3-11** Host Connection

Parameter	Mandatory	Description
Data Connection Name	Yes	Name of the host connection. The value can contain only letters, digits, hyphens (-), and underscores (_).
Host Address	Yes	IP address of the host For details, see section "Viewing Details About an ECS" in <i>Elastic Cloud Server User Guide</i> .
Agent	Yes	Agents provided by the CDM cluster.
Port	Yes	SSH port number of the host
Username	Yes	Username of the host
Login Mode	Yes	Mode for logging in to the host <ul style="list-style-type: none"> <li>• Key pair</li> <li>• Password</li> </ul>
Key Pair	Yes	If you select <b>Key pair</b> for <b>Login Mode</b> , you need to obtain the private key file, upload it to OBS, and select the OBS path. This parameter is available only when <b>Login Mode</b> is set to <b>Key pair</b> . <b>NOTE</b> The uploaded private key file must be in PEM format, and the uploaded private key file and the public key configured on the host must be in the same key pair.
Key Pair Password	No	If no password is set for the key pair, you do not need to set this parameter.
Password	Yes	Password for logging in to the host.
Host Connection Description	No	Description of the host connection

## Creating a Kerberos Authentication User for an MRS Security Cluster

You cannot create a data connection for an MRS security cluster as user **admin**. User **admin** is the management page user by default and cannot be used as the authentication user of a security cluster. To create an MRS user, perform the following steps:

For clusters of MRS 3.x:

1. Log in to **MRS Manager** as user **admin**.
2. Choose **System > Permission > User**. On the page displayed, click **Create** to add a dedicated user as the Kerberos authentication user. Select the user group **superGroup** for the user, and assign all roles to the user.

 NOTE

- For clusters of MRS 3.1.0 or later, the user must at least have permissions of the **Manager\_viewer** role to create data connections in Management Center. To perform database, table, and data operations on components, the user must also have user group permissions of the components.
  - For clusters earlier than MRS 3.1.0, the user must have permissions of the **Manager\_administrator** or **System\_administrator** role to create data connections in Management Center.
  - A user with only the **Manager\_tenant** or **Manager\_auditor** permission cannot create connections.
3. Log in to Manager as the new user and change the initial password. Otherwise, the connection fails to be created.
  4. Synchronize IAM users.
    - a. Log in to the MRS management console.
    - b. Choose **Clusters > Active Clusters**, select a running cluster, and click its name to go to its details page.
    - c. In the **Basic Information** area of the **Dashboard** page, click **Synchronize** on the right side of **IAM User Sync** to synchronize IAM users.

 NOTE

- When the policy of the user group to which the IAM user belongs changes from **MRS ReadOnlyAccess** to **MRS CommonOperations**, **MRS FullAccess**, or **MRS Administrator**, wait for 5 minutes until the new policy takes effect after the synchronization is complete because the **SSSD** (System Security Services Daemon) cache of cluster nodes needs time to be updated. Then, submit a job. Otherwise, the job may fail to be submitted.
- When the policy of the user group to which the IAM user belongs changes from **MRS CommonOperations**, **MRS FullAccess**, or **MRS Administrator** to **MRS ReadOnlyAccess**, wait for 5 minutes until the new policy takes effect after the synchronization is complete because the **SSSD** cache of cluster nodes needs time to be updated.

For clusters of MRS 2.x or earlier:

1. Log in to **MRS Manager** as user **admin**.
2. Choose **System > Manage User**. On the page displayed, add a dedicated user as the Kerberos authentication user. Select the user group **superGroup** for the user, and assign all roles to the user.

 NOTE

- For clusters of MRS 2.x or earlier, the user must have permissions of the **Manager\_administrator** or **System\_administrator** role to create data connections in Management Center.
  - A user with only the **Manager\_tenant** or **Manager\_auditor** permission cannot create connections.
3. Log in to MRS Manager as the new user and change the initial password. Otherwise, the connection fails to be created.
  4. Synchronize IAM users.
    - a. Log in to the MRS management console.
    - b. Choose **Clusters > Active Clusters**, select a running cluster, and click its name to go to its details page.

- c. In the **Basic Information** area of the **Dashboard** page, click **Synchronize** on the right side of **IAM User Sync** to synchronize IAM users.

 **NOTE**

- When the policy of the user group to which the IAM user belongs changes from **MRS ReadOnlyAccess** to **MRS CommonOperations**, **MRS FullAccess**, or **MRS Administrator**, wait for 5 minutes until the new policy takes effect after the synchronization is complete because the **SSSD** (System Security Services Daemon) cache of cluster nodes needs time to be updated. Then, submit a job. Otherwise, the job may fail to be submitted.
- When the policy of the user group to which the IAM user belongs changes from **MRS CommonOperations**, **MRS FullAccess**, or **MRS Administrator** to **MRS ReadOnlyAccess**, wait for 5 minutes until the new policy takes effect after the synchronization is complete because the **SSSD** cache of cluster nodes needs time to be updated.

## Editing a Data Connection

- Step 1** Log in to Management Center and click **Data Connection Management**.
- Step 2** In the data connection list, locate the data connection you want to edit and click **Edit** in the **Operation** column.
- Step 3** In the **Edit Data Connection** dialog box, modify connection parameters as required. For parameter details, see [Data Connection Parameter Description](#).
- Step 4** Click **Test** to test whether the data connection is valid. If the connection is normal, click **Yes**.

If the test connection is invalid, the data connection cannot be created. Modify the connection parameters as prompted and try again.

----End

## Deleting a Data Connection

If a data connection is deleted, the data table information of the data connection will also be deleted. Exercise caution when performing this operation. If the data connection you want to delete has been referenced, it cannot be deleted.

- Step 1** Log in to Management Center and click **Data Connection Management**.
- Step 2** In the data connection list, locate the data connection you want to delete and click **Delete** in the **Operation** column.
- Step 3** In the dialog box displayed, confirm the data connection information, and click **Yes**.

----End

## 3.2.3 Migrating Resources

To migrate resources in one workspace to another, you can use the resource migration function provided by DataArts Studio.

The resources that can be migrated include the data connections created in Management Center.

## Prerequisites

- Resource import and export depend on the OBS service.
- There are resources that can be migrated. For details on how to create data connections, see [Creating Data Connections](#).

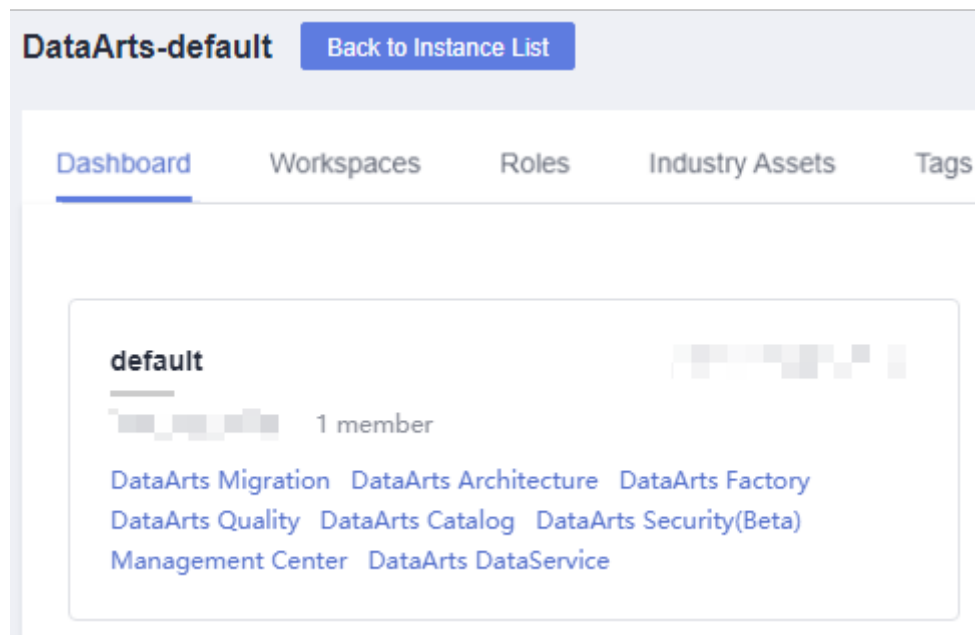
## Constraints

- Imported and exported resources are stored in JSON format.
- For security concerns, passwords of connections are not exported when the connections are exported. You need to enter the passwords when importing the connections.

## Exporting a Resource

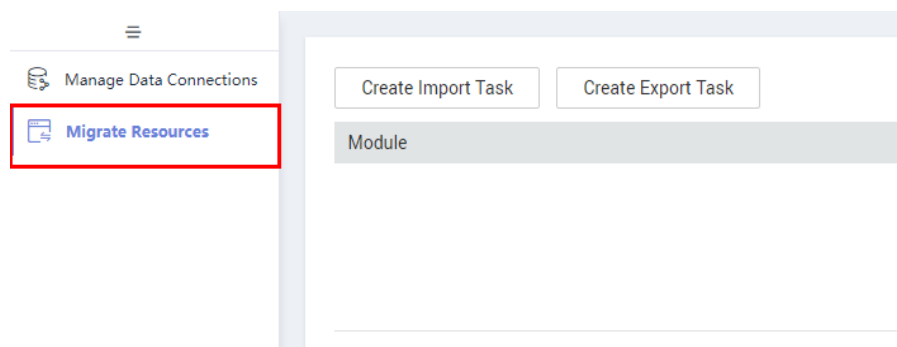
1. On the DataArts Studio console, locate a workspace and click **Management Center**.

Figure 3-4 Management Center



2. In the navigation pane, choose **Migrate Resources**.

Figure 3-5 Migrating Resources



3. Click **Create Export Task** to configure the file name and the OBS path for saving the file. If OBS is unavailable, you only need to set the file name.

**Figure 3-6** Export Task

**Export Task** X

① Select File ————— ② Select Template ————— ③ View Result

\* File Name

\* OBS Bucket


\* OBS Path

4. Click **Next** and select the resource to export.
5. Click **Next** and wait until the export is complete. The resource package is exported to the OBS path set in 3. If OBS is unavailable, you can click **Download** in the row of the corresponding migration task to download the exported resource package.

**Figure 3-7** Export completed

**Export Task** X

① Select File ————— ② Select Template ————— ③ View Result

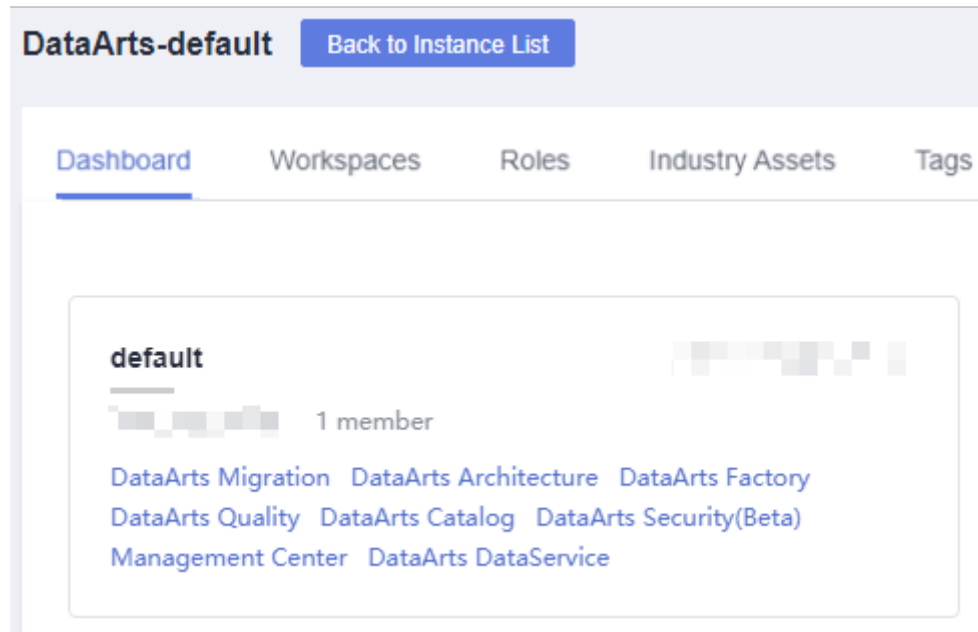
  
Task completed.

If no result is displayed in 1 minute, the export fails. Try again. If the failure persists, contact or technical support.

## Importing a Resource

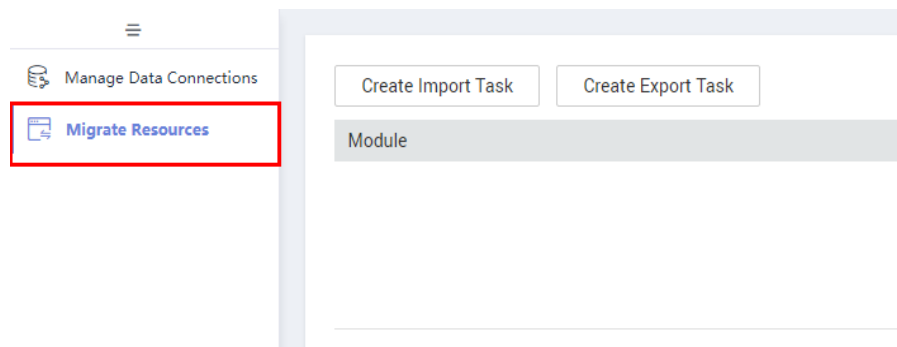
1. On the DataArts Studio console, locate a workspace and click **Management Center**.

Figure 3-8 Management Center



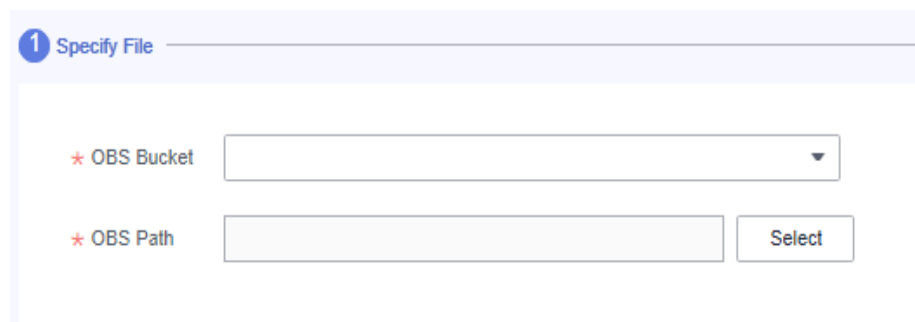
2. In the navigation pane, choose **Migrate Resources**.

Figure 3-9 Migrating Resources



3. Click **Create Import Task** and configure the path for saving the resources to import. If no OBS bucket is available, select the resource package to be uploaded from a local path.

Figure 3-10 Configuring the path for saving the resources to import



4. Click **Next** and select the resource to import.

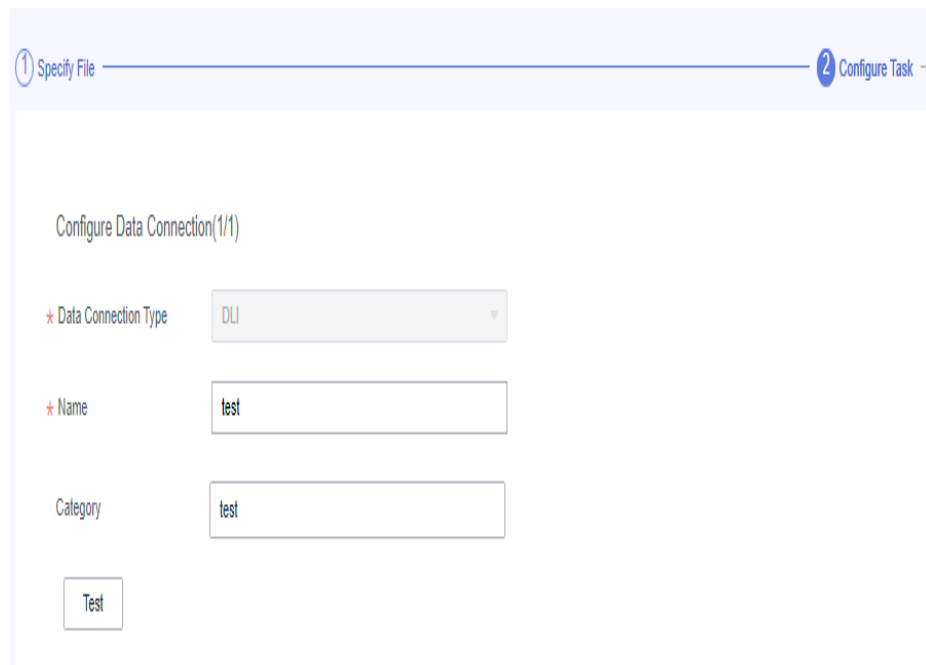


**Figure 3-11** Selecting the resource to import



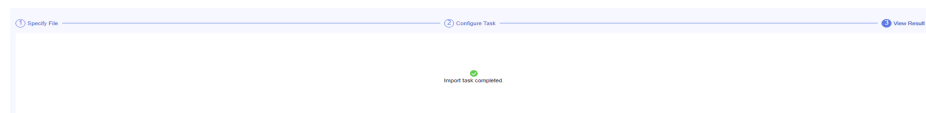
5. If you select **DataSource**, click **Next** to configure a data connection. The number of data connections required is determined by the number of data sources. Each data connection requires a password.

**Figure 3-12** Configuring a data connection



6. Click **Next** and wait until the import is complete.

**Figure 3-13** Import completed



If no result is displayed in 1 minute, the import fails. Try again. If the failure persists, contact or technical support.

## 3.2.4 Tutorials

### 3.2.4.1 Creating an MRS Hive Connection

This section describes how to create an MRS Hive connection between DataArts Studio and the data lake base.

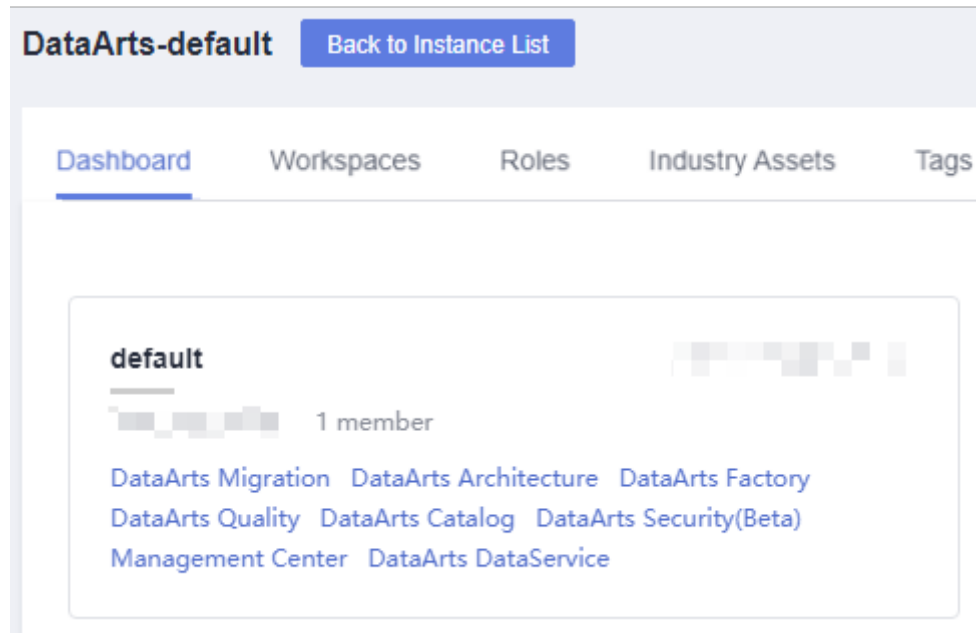
#### Prerequisites

- You have created a data lake to connect, for example, a database or cloud service supported by DataArts Studio.
  - Before creating a DWS data connection, ensure that you have created a cluster in DWS and have the permissions required to view Key Management Service (KMS) keys.
  - Before creating an MRS HBase, MRS Hive, MRS Kafka, MRS Presto, or MRS Spark connection, ensure that you have created an MRS cluster and selected required components.
  - Before creating an RDS data connection, ensure that you have created an RDS DB instance. Currently, DataArts Studio supports only MySQL and PostgreSQL databases in RDS.
- The data lake to connect communicates with the DataArts Studio instance properly.
  - If the data lake is an on-premises database, a public network or a dedicated connection is required. Ensure that the host where the data source is located can access the public network and the port has been enabled in the firewall rule.
  - If the data lake is a cloud service (such as DWS and MRS), the following requirements must be met for network interconnection:
    - If the CDM cluster in the DataArts Studio instance and the cloud service are in different regions, a public network or a dedicated connection is required.
    - If the CDM cluster in the DataArts Studio instance 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 **Adding Routes** in *Virtual Private Cloud (VPC) Usage Guide*. For details about how to configure security group rules, see **Security Group > Adding a Security Group Rule** in *Virtual Private Cloud (VPC) Usage Guide*.
    - The cloud service instance and the DataArts Studio workspace belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

#### Creating a Data Connection

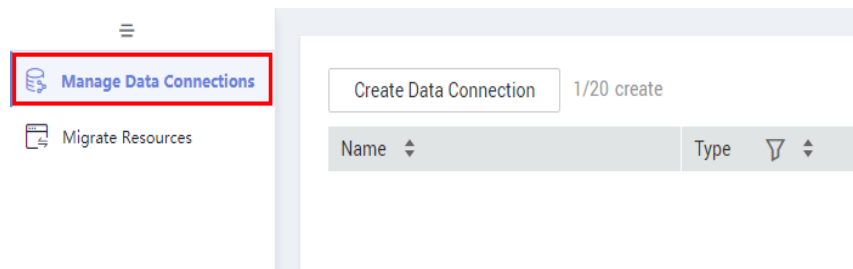
1. On the DataArts Studio console, locate a workspace and click **Management Center**.

Figure 3-14 Management Center



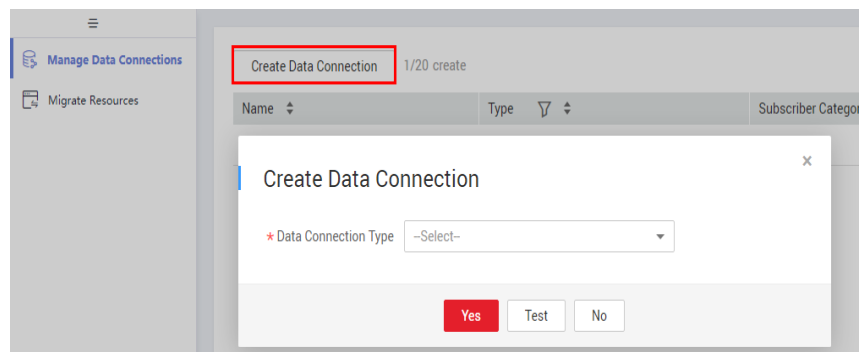
2. In the navigation pane, choose **Manage Data Connections**.

Figure 3-15 Creating a Data Connection



3. On the **Manage Data Connections** page, click **Create Data Connection**. In the displayed dialog box, select **MRS Hive** for **Data Connection Type** and set other parameters based on the descriptions in [Table 3-12](#).

Figure 3-16 Create Data Connection



**Figure 3-17** MRS Hive connection parameters

\* Data Connection Type

\* Name

Category

\* Cluster Name  [Manage Cluster](#)

\* Username

\* Password

\* KMS Key  [Access KMS](#)

\* Connection Type  Proxy connection  MRS API connection

\* Agent  [Manage CDM Clusters](#)

**Table 3-12** MRS Hive data connection

Parameter	Mandatory	Description
Data Connection Name	Yes	The name of the data connection to create. Data connection names can contain 1 to 50 characters. They can include only letters, numbers, underscores (_), and hyphens (-).
Tag	No	The attribute of the data connection to create. Tags make management easier. <b>NOTE</b> The name of the tag. Only letters, numbers, and underscores (_) are allowed. Tag names cannot start with underscores (_). Enter up to 100 characters.

Parameter	Mandatory	Description
Cluster Name	Yes	<p>The name of the MRS Hive cluster. Select an MRS cluster that Hive belongs to. If the MRS cluster is not displayed in the drop-down list, check whether the network connection between the MRS cluster and the DataArts Studio instance is normal.</p> <p>Ensure that the MRS cluster and the DataArts Studio instance can communicate with each other. The following requirements must be met for network interconnection:</p> <ul style="list-style-type: none"> <li>• 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.</li> <li>• 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 <b>Custom Route in Region Type I &gt; Adding Routes</b> in <i>Virtual Private Cloud (VPC) Usage Guide</i>. For details about how to configure security group rules, see <b>Security Group &gt; Adding a Security Group Rule</b> in <i>Virtual Private Cloud (VPC) Usage Guide</i>.</li> <li>• 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.</li> </ul>

Parameter	Mandatory	Description
Connection Type	Yes	<p>Connection type. <b>Proxy connection</b> is recommended.</p> <ul style="list-style-type: none"> <li>• <b>Proxy connection:</b> An agent (CDM cluster) is used to access MRS clusters. This method supports all versions of MRS clusters.</li> <li>• <b>MRS API connection:</b> MRS APIs are used to access MRS clusters. This method supports only MRS clusters of the 2.X or a later version. When you select <b>MRS API connection</b>, pay attention to the following restrictions: <ol style="list-style-type: none"> <li>1. Tables and fields cannot be viewed.</li> <li>2. When the SQL editor is used to run SQL statements, the execution results can be displayed only in logs.</li> <li>3. This method is not supported by data governance functions such as DataArts Architecture, DataArts Quality, and DataArts Catalog.</li> </ol> </li> </ul>
Username	No	<p>The username of the MRS cluster. This parameter is mandatory when <b>Connection Type</b> is set to <b>Proxy connection</b>. If a new MRS user is used for connection, you need to log in to Manager and change the initial password.</p> <p>To create a data connection for an MRS security cluster, do not use user <b>admin</b>. The <b>admin</b> 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 based on <a href="#">Creating a Kerberos Authentication User for an MRS Security Cluster</a> and set <b>Username</b> and <b>Password</b> to the username and password of the created MRS user when creating an MRS data connection.</p> <p><b>NOTE</b></p> <ul style="list-style-type: none"> <li>• For clusters of MRS 3.1.0 or later, the user must at least have permissions of the <b>Manager_viewer</b> role to create data connections in Management Center. To perform database, table, and data operations on components, the user must also have user group permissions of the components.</li> <li>• For clusters earlier than MRS 3.1.0, the user must have permissions of the <b>Manager_administrator</b> or <b>System_administrator</b> role to create data connections in Management Center.</li> <li>• A user with only the <b>Manager_tenant</b> or <b>Manager_auditor</b> permission cannot create connections.</li> </ul>

Parameter	Mandatory	Description
Password	No	The password for accessing the MRS cluster. This parameter is mandatory when <b>Connection Type</b> is set to <b>Proxy connection</b> .
KMS Key	No	Name of the KMS key. This parameter is mandatory when <b>Connection Type</b> is set to <b>Proxy connection</b> .
Agent	No	This parameter is mandatory when <b>Connection Type</b> is set to <b>Proxy connection</b> . MRS is not a fully managed service and cannot be directly connected to DataArts Studio. A CDM cluster can provide an agent for DataArts Studio to communicate with non-fully-managed services. Therefore, you need to select a CDM cluster when creating an MRS data connection. If no CDM cluster is available, create one through the DataArts Migration incremental package. As a network proxy, the CDM cluster must be able to communicate with the MRS cluster. To ensure network connectivity, the CDM cluster must be in the same region, AZ, VPC, and subnet as the MRS cluster. The security group rule must also allow the CDM cluster communicate with the MRS cluster.

4. Click **Test** to test connectivity of the data connection. If the test fails, the data connection fails to be created.
5. After the test is successful, click **OK** to create the data connection.

## Reference

1. Why is no MRS Hive cluster available in the dialog box for creating a data connection?  
Possible causes are as follows:
  - Hive/HBase components were not selected during MRS cluster creation.
  - The network between the CDM cluster and MRS cluster was disconnected when an MRS data connection is created.  
The CDM cluster functions as a network agent. MRS data connections that you are going to create need to communicate with CDM.
2. Why does a Hive data connection fail to obtain information about databases or tables?  
The possible cause is that the CDM cluster is stopped or a concurrency conflict occurs. You can switch to another agent to temporarily avoid this issue.

### 3.2.4.2 Creating a DWS Connection

This section describes how to create a DWS connection between DataArts Studio and the data lake base.

## Prerequisites

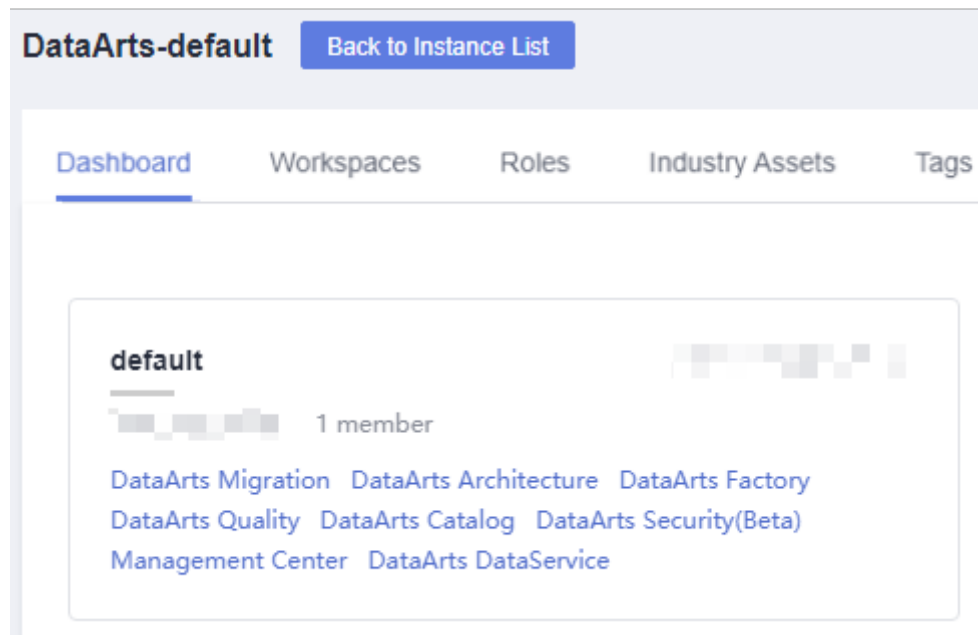
- You have created a data lake to connect, for example, a database or cloud service supported by DataArts Studio.
  - Before creating a DWS data connection, ensure that you have created a cluster in DWS and have the permissions required to view Key Management Service (KMS) keys.
  - Before creating an MRS HBase, MRS Hive, MRS Kafka, MRS Presto, or MRS Spark connection, ensure that you have created an MRS cluster and selected required components.
  - Before creating an RDS data connection, ensure that you have created an RDS DB instance. Currently, DataArts Studio supports only MySQL and PostgreSQL databases in RDS.
- The data lake to connect communicates with the DataArts Studio instance properly.
  - If the data lake is an on-premises database, a public network or a dedicated connection is required. Ensure that the host where the data source is located can access the public network and the port has been enabled in the firewall rule.
  - If the data lake is a cloud service (such as DWS and MRS), the following requirements must be met for network interconnection:
    - If the CDM cluster in the DataArts Studio instance and the cloud service are in different regions, a public network or a dedicated connection is required.
    - If the CDM cluster in the DataArts Studio instance 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 **Adding Routes** in *Virtual Private Cloud (VPC) Usage Guide*. For details about how to configure security group rules, see **Security Group > Adding a Security Group Rule** in *Virtual Private Cloud (VPC) Usage Guide*.
    - The cloud service instance and the DataArts Studio workspace belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

## Creating a Data Connection

1. On the DataArts Studio console, locate a workspace and click **Management Center**.

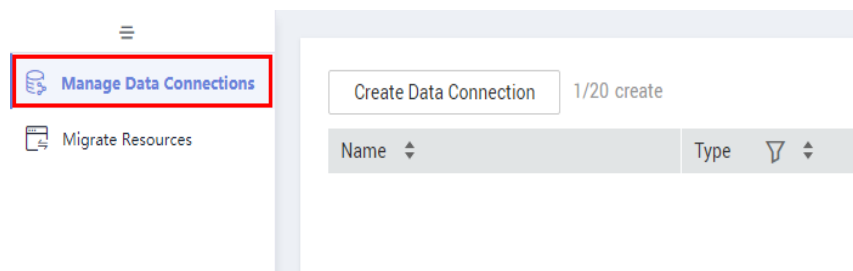


Figure 3-18 Management Center



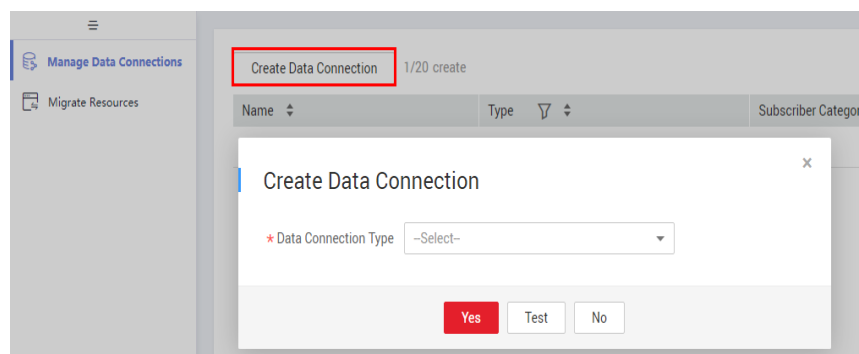
2. In the navigation pane, choose **Manage Data Connections**.

Figure 3-19 Creating a Data Connection



3. On the **Manage Data Connections** page, click **Create Data Connection**. In the displayed dialog box, select **DWS** for **Data Connection Type** and set other parameters based on the descriptions in [Table 3-13](#).

Figure 3-20 Create Data Connection



**Figure 3-21** DWS connection parameters

\* Data Connection Type

\* Name

Category

\* Manual

\* SSL Connection

\* Cluster Name  [Manage Cluster](#)

\* Username

\* Password

\* KMS Key  [Access KMS](#)

\* Connection Type  Proxy connection  Direct connection

\* Agent  [Manage CDM Clusters](#)

**Table 3-13** DWS data connection

Parameter	Mandatory	Description
Data Connection Name	Yes	The name of the data connection to create. Data connection names can contain 1 to 50 characters. They can include only letters, numbers, underscores (_), and hyphens (-).
Tag	No	The attribute of the data connection to create. Tags make management easier. <b>NOTE</b> The name of the tag. Only letters, numbers, and underscores (_) are allowed. Tag names cannot start with underscores (_). Enter up to 100 characters.
Manual	Yes	You can turn off or turn on to disable or enable the <b>Manual</b> function. <ul style="list-style-type: none"> <li>When <b>Manual</b> is disabled, you do not need to enter the IP address and port.</li> <li>When <b>Manual</b> is enabled, you must enter the IP address and port.</li> </ul>

Parameter	Mandatory	Description
IP	No	The IP address for accessing the cluster database through the internal network. This parameter is mandatory when <b>Manual</b> is enabled. The private network address is automatically generated when you create a cluster.
Port	No	The database port specified during DWS cluster creation. This parameter is mandatory when <b>Manual</b> is enabled. Ensure that you have enabled this port in the security group rule so that the DataArts Studio instance can connect to the database in the DWS cluster through this port.
SSL Connection	Yes	DWS supports SSL encryption and certificate authentication for communication between the client and server. You can use <b>SSL Connection</b> to set the communication mode. If <b>SSL Connection</b> is enabled, only SSL encryption can be used. If <b>SSL Connection</b> is disabled, both modes can be used. This function is disabled by default.
Cluster Name	Yes	The name of the selected DWS cluster.
Username	Yes	The database username, which is specified when the DWS cluster is created.
Password	Yes	The password for accessing the database, which is specified when the DWS cluster is created.
KMS Key	Yes	Name of the KMS key.
Connection Type	Yes	<p>Connection type. <b>Proxy connection</b> is recommended.</p> <ul style="list-style-type: none"> <li>• <b>Proxy connection:</b> An agent (CDM cluster) is used to access DWS clusters.</li> <li>• <b>Direct connection:</b> You can access DWS clusters directly.</li> </ul>

Parameter	Mandatory	Description
Agent	No	<p>This parameter is mandatory when <b>Connection Type</b> is set to <b>Proxy connection</b>.</p> <p>DWS is not a fully managed service and cannot be directly connected to DataArts Studio. A CDM cluster can provide an agent for DataArts Studio to communicate with non-fully-managed services. Therefore, you need to select a CDM cluster when creating a DWS data connection. If no CDM cluster is available, create one through the DataArts Migration incremental package.</p> <p>As a network proxy, the CDM cluster must be able to communicate with the DWS cluster. To ensure network connectivity, the CDM cluster must be in the same region, AZ, VPC, and subnet as the DWS cluster. The security group rule must also allow the CDM cluster communicate with the DWS cluster.</p>

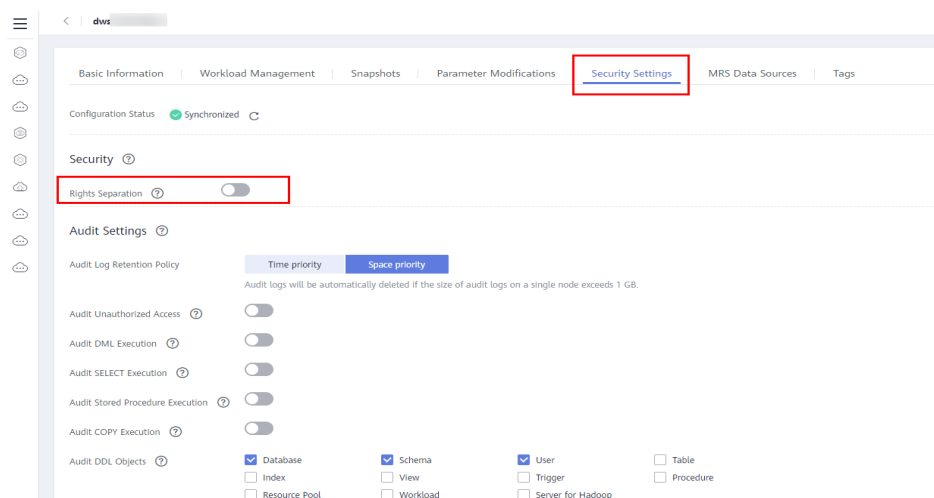
4. Click **Test** to test connectivity of the data connection. If the test fails, the data connection fails to be created.
5. After the test is successful, click **OK** to create the data connection.

## Reference

1. What should I do if the connection test fails when I enable the SSL connection during the creation of a DWS data connection?

The failure may be caused by the rights separation function of the DWS cluster. On the DWS console, click the corresponding cluster, choose **Security Settings**, and disable **Rights Separation**.

**Figure 3-22** Disabling Rights Separation for the DWS cluster



2. Why does a DWS data connection fail to obtain information about databases or tables?

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

### 3.2.4.3 Creating a MySQL Connection

This section describes how to create a MySQL connection between DataArts Studio and the data lake base.

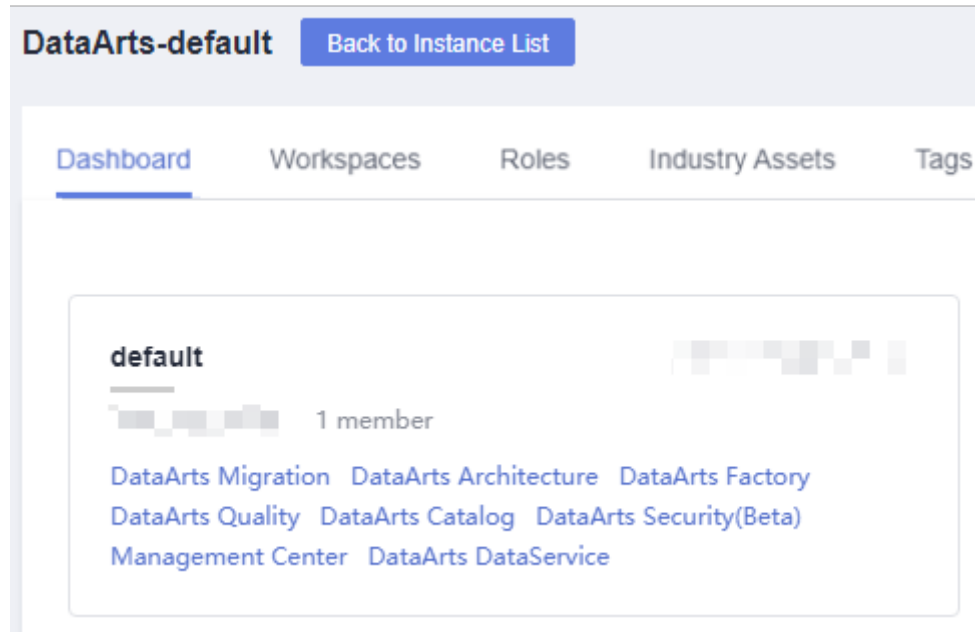
#### Prerequisites

- You have created a data lake to connect, for example, a database or cloud service supported by DataArts Studio.
  - Before creating a DWS data connection, ensure that you have created a cluster in DWS and have the permissions required to view Key Management Service (KMS) keys.
  - Before creating an MRS HBase, MRS Hive, MRS Kafka, MRS Presto, or MRS Spark connection, ensure that you have created an MRS cluster and selected required components.
  - Before creating an RDS data connection, ensure that you have created an RDS DB instance. Currently, DataArts Studio supports only MySQL and PostgreSQL databases in RDS.
- The data lake to connect communicates with the DataArts Studio instance properly.
  - If the data lake is an on-premises database, a public network or a dedicated connection is required. Ensure that the host where the data source is located can access the public network and the port has been enabled in the firewall rule.
  - If the data lake is a cloud service (such as DWS and MRS), the following requirements must be met for network interconnection:
    - If the CDM cluster in the DataArts Studio instance and the cloud service are in different regions, a public network or a dedicated connection is required.
    - If the CDM cluster in the DataArts Studio instance 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 **Adding Routes** in *Virtual Private Cloud (VPC) Usage Guide*. For details about how to configure security group rules, see **Security Group > Adding a Security Group Rule** in *Virtual Private Cloud (VPC) Usage Guide*.
    - The cloud service instance and the DataArts Studio workspace belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

## Creating a Data Connection

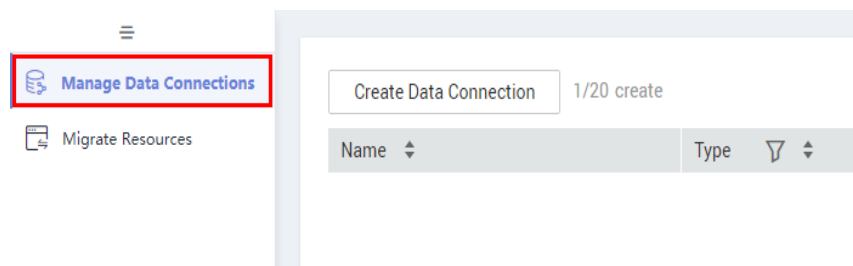
1. On the DataArts Studio console, locate a workspace and click **Management Center**.

Figure 3-23 Management Center



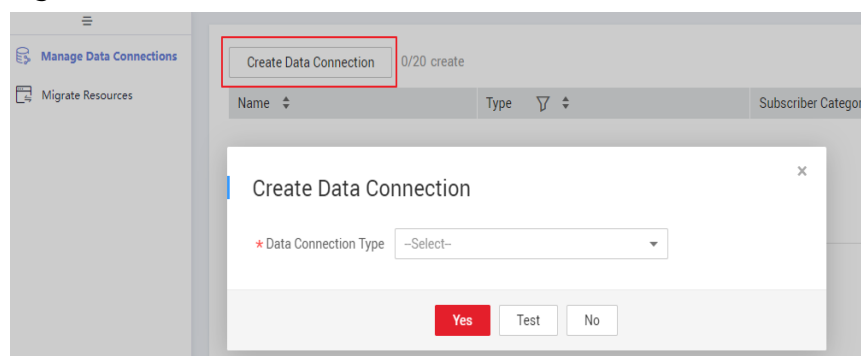
2. In the navigation pane, choose **Manage Data Connections**.

Figure 3-24 Creating a Data Connection



3. On the **Manage Data Connections** page, click **Create Data Connection**. In the displayed dialog box, select **RDS** for **Data Connection Type** and set other parameters based on the descriptions in [Table 3-14](#).

Figure 3-25 Create Data Connections



 **NOTE**

- You are not advised to select **MySQL (pending offline)** for **Data Connection Type**. Instead, You are advised to select **RDS**.
- RDS data connections depend on OBS. If OBS is unavailable in the same region as DataArts Studio, RDS data connections are not supported.

**Figure 3-26** RDS connection parameters

The screenshot shows a 'Create Data Connection' dialog box with the following fields and values:

- Data Connection Type:** RDS
- Name:** mysql
- Tag:** Enter a keyword.
- IP Address:** 114 . 116 . 231 . 174
- Port:** 3306
- Driver Name:** com.mysql.jdbc.Driver
- Driver File Path:** obs://obs-dayu-lgh/mysql-connector-java8-5.1. (with a 'Select' button)
- Username:** root
- Password:** .....
- KMS Key:** dlf/default (with a 'Access KMS' link)
- Agent:** cdm-dayu (with a 'Manage CDM' link)

At the bottom, there are three buttons: 'Yes' (highlighted in red), 'Test', and 'No'.

**Table 3-14** RDS data connection

Parameter	Mandatory	Description
Data Connection Name	Yes	The name of the data connection to create. Data connection names can contain 1 to 50 characters. They can include only letters, numbers, underscores (_), and hyphens (-).
Tag	No	The attribute of the data connection to create. Tags make management easier. <b>NOTE</b> The name of the tag. Only letters, numbers, and underscores (_) are allowed. Tag names cannot start with underscores (_). Enter up to 100 characters.

Parameter	Mandatory	Description
IP	Yes	<p>The address for accessing RDS.</p> <p>If the data source is RDS, you can obtain the address from the RDS console.</p> <ol style="list-style-type: none"> <li>1. Log in to the management console using the created account.</li> <li>2. In the <b>Service List</b>, choose <b>Relational Database Service</b>. In the left navigation pane, choose <b>Instances</b>.</li> <li>3. Click the name of an instance. The basic information page of the instance is displayed.</li> </ol> <p>You can obtain the IP address on the <b>Connection Information</b> tab.</p>
Port	Yes	<p>The port for accessing RDS.</p> <p>If the data source is RDS, you can obtain the port from the RDS console.</p> <ol style="list-style-type: none"> <li>1. Log in to the management console using the account.</li> <li>2. In the <b>Service List</b>, choose <b>Relational Database Service</b>. In the left navigation pane, choose <b>Instances</b>.</li> <li>3. Click the name of an instance. The basic information page of the instance is displayed.</li> </ol> <p>You can obtain the database port on the <b>Connection Information</b> tab.</p>
Driver Name	Yes	<p>The name of the driver. The following values are available:</p> <ul style="list-style-type: none"> <li>• com.mysql.jdbc.Driver</li> <li>• org.postgresql.Driver</li> </ul>
Driver File Path	Yes	<p>Path of the driver file in the OBS bucket. You need to download the .jar driver file from the corresponding official website and upload it to the OBS bucket.</p> <ul style="list-style-type: none"> <li>• MySQL driver: Download it from <a href="https://downloads.mysql.com/archives/c-j/">https://downloads.mysql.com/archives/c-j/</a>. The 5.1.48 version is recommended.</li> <li>• PostgreSQL driver: Download it from <a href="https://jdbc.postgresql.org/download">https://jdbc.postgresql.org/download</a>. The 42.1.4 version is recommended.</li> </ul> <p><b>NOTE</b> To update the driver, you must restart the CDM cluster in DataArts Migration and then edit the data connection to upload the driver.</p>



Parameter	Mandatory	Description
Username	Yes	The username of the database. The username is required for creating a cluster.
Password	Yes	The password for accessing the database. The password is required for creating a cluster.
KMS Key	Yes	Name of the KMS key. To obtain the key: 1. Log in to the management console using the account. 2. Click <b>Key Management Service</b> and select <b>Key Management Service</b> from the list on the left. You can obtain the key name from the key list.
Agent	Yes	RDS is not a fully managed service and cannot be directly connected to DataArts Studio. A CDM cluster can provide an agent for DataArts Studio to communicate with non-fully-managed services. Therefore, you need to select a CDM cluster when creating an RDS data connection. If no CDM cluster is available, create one through the DataArts Migration incremental package.  As a network proxy, the CDM cluster must be able to communicate with RDS. To ensure network connectivity, the CDM cluster must be in the same region, AZ, VPC, and subnet as RDS. The security group rule must also allow the CDM cluster to communicate with RDS.

4. Click **Test** to test connectivity of the data connection. If the test fails, the data connection fails to be created.
5. After the test is successful, click **OK** to create the data connection.

## Reference

1. What Are the Precautions for Creating an RDS Data Connection?  
When creating an RDS data connection, you need to bind an agent provided by the CDM cluster. Currently, a version of the CDM cluster earlier than 1.8.6 is not supported.

## 3.3 DataArts Migration

### 3.3.1 Overview

DataArts Migration is an efficient and easy-to-use data integration service. Based on the big data migration to the cloud and intelligent data lake solutions, CDM

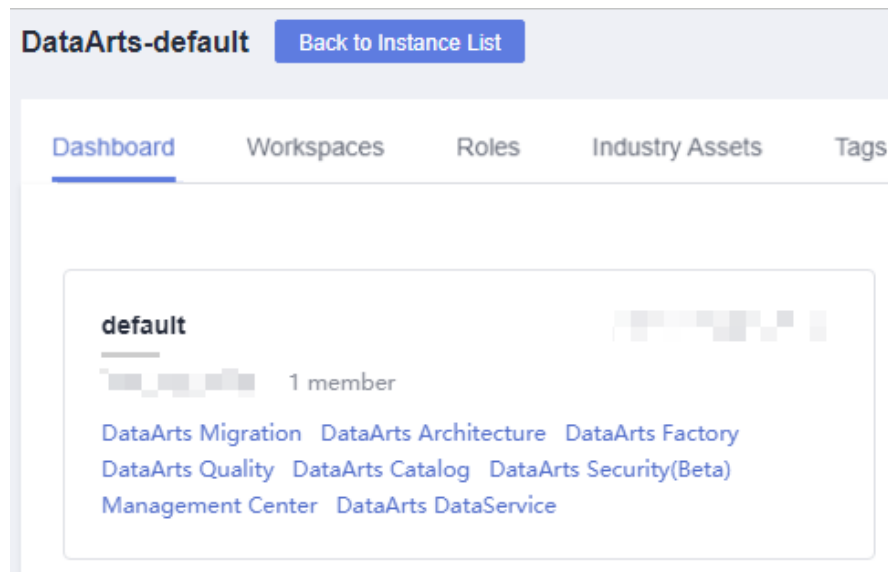
provides easy-to-use migration capabilities and can integrate various types of data sources into the data lake, which simplifies data source migration and integration and improves efficiency for you.

In this document, DataArts Migration refers to Cloud Data Migration (CDM).

You can access the CDM console using either of the following methods:

- Log in to the CDM console and choose **Cluster Management** in the navigation pane.
- Log in to the DataArts Studio console. Locate a workspace and click **DataArts Migration**.

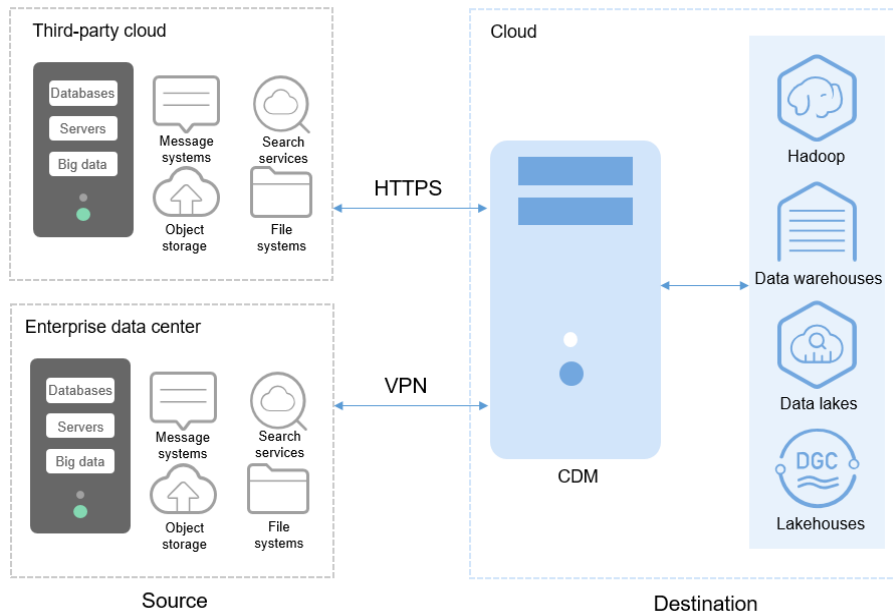
**Figure 3-27** DataArts Migration



## Introduction to CDM

CDM uses a distributed compute framework and concurrent processing techniques to help you migrate enterprise data in batches without any downtime and rapidly build desired data structures.

Figure 3-28 CDM



## Functions

- **Table/file/entire DB migration**

Tables or files can be migrated in batches. An entire database can be migrated between homogeneous and heterogeneous databases. A job can migrate hundreds of tables.
- **Incremental data migration**

CDM supports incremental migration of files, relational databases, and HBase/CloudTable, as well as with WHERE clauses and macro variables of date and time.
- **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.
- **Field conversion**

CDM supports field conversion functions, such as anonymization, character string operations, and date operations.
- **File encryption**

When files are migrated to a file system, CDM can encrypt the files written to the cloud.
- **MD5 verification**

MD5 verification is supported to check the file consistency from end to end and output verification result.
- **Dirty data archiving**

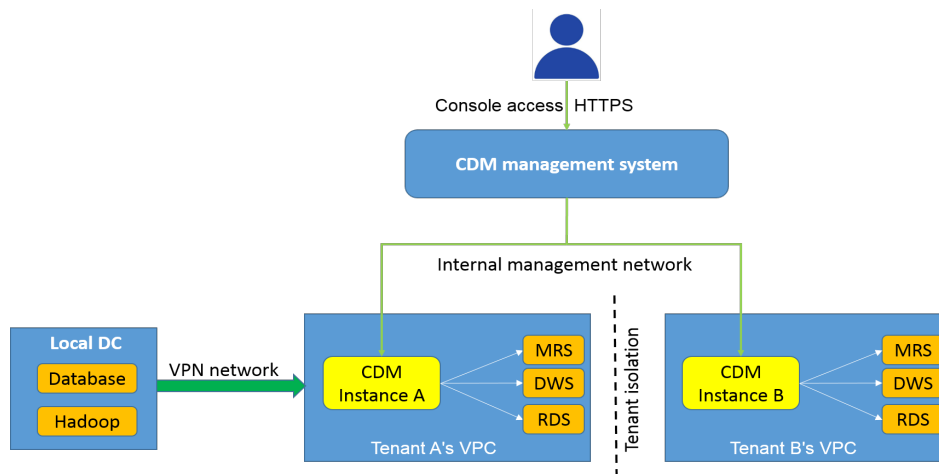
CDM can archive the data that fails to be processed during migration, has been filtered out, or is not compliant with conversion or cleaning rules to dirty data logs. The threshold for dirty data ratio can be set to determine whether a task is successful.

## Migration Principles

When a tenant uses CDM, the CDM system provisions a fully-managed CDM instance in the tenant's VPC. The instance allows only console and RESTful API access. Therefore the tenant cannot access the instance through other interfaces (such as SSH). This ensures data isolation between CDM tenants, prevents data leakage, and ensures transmission security during data migration between different cloud services in a VPC. Tenants can also use the VPN to migrate data from the on-premises data center to cloud services to ensure migration security.

CDM works in push-pull mode. CDM pulls data from the migration source and pushes the data to the migration destination. Data access operations are initiated by CDM. SSL will be used if the data source (such as RDS) supports it. During the migration, the usernames and passwords of the migration source and destination are required. Such information is stored in the database of the CDM instance. Protecting such information is critical to ensure CDM security.

Figure 3-29 Migration principles



### 3.3.2 Constraints

#### CDM System Constraints

1. You cannot modify the flavor of an existing cluster. If you require a higher flavor, create a cluster with your desired flavor.
2. Arm CDM clusters do not support agents. The CDM cluster version (Arm or x86) is determined by the architecture of underlying resources.
3. CDM does not support the function of controlling the data migration speed. Therefore, do not perform data migration during peak hours.
4. The baseline and maximum bandwidths of the NIC of the `cdm.large` CDM instance is 0.8 Gbit/s and 3 Gbit/s, respectively. The theoretical maximum volume of data that can be transmitted per instance per day is about 8 TB. Similarly, the baseline and maximum bandwidths of the NIC of the `cdm.xlarge` instance are 4 Gbit/s and 10 Gbit/s, respectively, and the theoretical maximum volume of data that can be transmitted per instance per day is about 40 TB. The baseline and maximum bandwidths of the NIC of the `cdm.4xlarge` instance is 36 Gbit/s and 40 Gbit/s, respectively, and the theoretical maximum

volume of data that can be transmitted per instance per day is about 360 TB. You can use multiple CDM instances if you want faster data transfer.

The actual amount of data that can be migrated in a day depends on the data source type, the read and write performance of the source and destination, and the actual available bandwidth. Typically you can migrate as much as 8 TB per day (large file migration to OBS) using the `cdm.large` instance. It is recommended that you test the speed with a small amount of data before migration.

5. 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.
6. 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. You can export links and jobs configured on CDM to a local directory. To ensure password security, CDM does not export the link password of the corresponding data source. Therefore, before importing job configurations to CDM, you need to manually input the password in the exported JSON file or configure the password in the import dialog box.
8. The cluster cannot automatically upgrade to a new version. You need to use the job export and import functions to upgrade the cluster to the new version.
9. If OBS is unavailable, CDM does not automatically back up users' job configurations. You need to export and back up configuration data using the export function.
10. 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.
11. If the destination of a CDM job is a DWS or NewSQL database, constraints of the source end, such as the primary key and unique index, cannot be migrated together.
12. When performing a CDM job, ensure that the JSON file formats of the two clusters are the same so that jobs can be imported from the source cluster to the destination cluster.

## General Constraints on Database Migration

1. CDM is mainly used for batch migration. It supports only limited incremental migration but does not support real-time incremental migration.
2. The entire DB migration of CDM supports only data table migration but not migration of database objects such as stored procedures, triggers, functions, and views.

CDM applies only to scenarios where databases are migrated to the cloud at a time, including homogeneous and heterogeneous database migrations. CDM is not applicable to data synchronization, for example, disaster recovery and real-time synchronization.

3. If CDM fails to migrate an entire database or table, the data that has been imported to the target table will not be rolled back automatically. If you want to perform migration in transaction mode, configure the **Import to Staging Table** parameter to enable a rollback upon a migration failure.  
In extreme cases, the created stage table or temporary table cannot be automatically deleted. You need to manually clear the table (the table name of the stage table ends with **\_cdm\_stage**), for example, **cdmtet\_cdm\_stage**).
4. If CDM needs to access data sources in the on-premises data center (for example, the on-premises MySQL database), the data sources must support Internet access and the CDM instances must be bound with elastic IP addresses. In this case, the security practice is to configure the firewall or security policies to allow only the EIPs of the CDM instances to access the local data sources.
5. Only common data types are supported, including character strings, digits, and dates. Object types are limited. If objects are too large, migration cannot be performed.
6. Only the GBK and UTF-8 character sets are supported.
7. A field name cannot contain & and %.

## Permissions Configuration for Relational Database Migration

Common minimum permissions required by relational database migration:

- MySQL: You need to have the read permission on the **INFORMATION\_SCHEMA** database and data tables.
- Oracle: You need to have the **resource** role and have the **select** permissions on the data table in the tablespace.
- Dameng: You need to have the **select any table** permission in the schema.
- DWS: You need to have the **schema usage** permission and the query permission on the data tables.
- SQL Server: You need to have the **sysadmin** permission.
- PostgreSQL: You need to have the **select** permission on schema tables in the database.

## Constraints on FusionInsight HD and Apache Hadoop

If the FusionInsight HD and Apache Hadoop data sources are deployed in the on-premises data center, CDM must access all nodes in the cluster for reading and writing the Hadoop files. Therefore, the network access must be enabled for each node.

## Constraints on DWS and FusionInsight Libra

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

- CHARACTER TYPES: CHAR, BPCHAR, VARCHAR, VARCHAR2, NVARCHAR2, TEXT
  - DATA/TIME TYPES: DATE, TIME, TIMETZ, TIMESTAMP, TIMESTAMPTZ, INTERVAL, SMALLDATETIME
2. In DWS, the character string '' is null. A null character string cannot be inserted into a field with non-null constraints. This is inconsistent with the MySQL behavior. MySQL does not consider that '' is null. Migration from MySQL to DWS may fail due to the preceding reason.
  3. When the Gauss Data Service (GDS) mode is used to quickly import data to DWS, you need to configure a security group or firewall policy to allow DataNodes of DWS or FusionInsight LibrA to access port 25000 of the CDM IP address.
  4. When data is imported to DWS in GDS mode, CDM automatically creates a foreign table for data import. The table name ends with a universally unique identifier (UUID), for example, **cdmtest\_aecf3f8n0z73dsl72d0d1dk4lcir8cd**. If a job fails, it will be automatically deleted. In extreme cases, you may need to manually delete it.

## Constraints on OBS

1. During file migration, the system automatically transfers the files concurrently. In this case, **Concurrent Extractors** in the task configuration is invalid.
2. Resumable transfer is not supported. If CDM fails to transfer files, OBS fragments are generated. You need to clear fragments on the OBS console to prevent space occupation.
3. CDM does not support the versioning control function of OBS.
4. During incremental migration, the number of files or objects in the source directory of a single job depends on the CDM cluster flavor. A **cdm.large** cluster supports a maximum of 300,000 files; a **cdm.medium** cluster supports a maximum of 200,000 files; and a **cdm.small** cluster supports a maximum of 100,000 files.

If the number of files or objects in a single directory exceeds the upper limit, split the files or objects into multiple migration jobs based on subdirectories.

## Constraints on DLI

To use CDM to migrate data to DLI, you must have the read permissions of OBS.

## Constraints on Oracle

Real-time incremental data synchronization is not supported for Oracle databases.

## Constraints on DCS and Redis

1. 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.

2. Only the hash and string data formats are supported.

## Constraints on DDS and MongoDB

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.

## Constraints on CSS and Elasticsearch

1. CDM supports automatic creation of indexes and field types. The index and field type names can contain only lowercase letters.
2. You cannot modify the field type under an index after it is created, but only create another field.

If you need to modify the field type, you need to create an index or run the Elasticsearch command on Kibana to delete the existing index and create another index (the data is also deleted).

3. When the field type of the index created by CDM is date, the data format must be *yyyy-MM-dd HH:mm:ss.SSS Z*. For example, **2018-08-08 08:08:08.888 +08:00**.

During data migration to CSS, if the original data of the **date** field does not meet the format requirements, you can use the expression conversion function of CDM to convert the data to the preceding format.

## Constraints on Kafka

1. The data in the message body is a record in CSV format that supports multiple delimiters. Messages cannot be parsed in binary or other formats.

## Constraints on CloudTable and HBase

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.

## Constraints on Hive

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;

```

### 3.3.3 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 3-15](#) describes the supported data sources.

**Table 3-15** Supported data sources during table/file migration

Category	Source	Destination	Description
Data warehouse	GaussDB(DWS)	<ul style="list-style-type: none"> <li>• Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI)</li> <li>• Hadoop: MRS HDFS, MRS HBase, and MRS Hive</li> <li>• Object-based storage: Object Storage Service (OBS)</li> <li>• Relational database: RDS for MySQL, RDS for PostgreSQL, RDS for SQL Server, MySQL, PostgreSQL, Microsoft SQL Server, and Oracle</li> <li>• NoSQL: CloudTable</li> </ul>	The DWS physical machine management mode is not supported.

Category	Source	Destination	Description
	Data Lake Insight (DLI)	<ul style="list-style-type: none"> <li>Search: Elasticsearch and Cloud Search Service (CSS)</li> </ul>	-
Hadoop	MRS HDFS	<ul style="list-style-type: none"> <li>Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI)</li> </ul>	<ul style="list-style-type: none"> <li>Supported by local storage. Only MRS Hive is supported in storage-compute decoupling scenarios.</li> <li>Only MRS Hive is supported in Ranger scenarios.</li> <li>Not supported if SSL is enabled for ZooKeeper</li> <li>Recommended MRS HDFS versions:                             <ul style="list-style-type: none"> <li>- 2.8.X</li> <li>- 3.1.X</li> </ul> </li> <li>Recommended MRS HBase versions:                             <ul style="list-style-type: none"> <li>- 2.1.X</li> <li>- 1.3.X</li> </ul> </li> <li>MRS Hive 2.x versions are not supported. The following versions are recommended:                             <ul style="list-style-type: none"> <li>- 1.2.X</li> <li>- 3.1.X</li> </ul> </li> </ul>
	MRS HBase	<ul style="list-style-type: none"> <li>Hadoop: MRS HDFS, MRS HBase, and MRS Hive</li> </ul>	
	MRS Hive	<ul style="list-style-type: none"> <li>Object-based storage: Object Storage Service (OBS)</li> <li>Relational database: RDS for MySQL, RDS for PostgreSQL, RDS for SQL Server, MySQL, PostgreSQL, Microsoft SQL Server, and Oracle</li> <li>NoSQL: CloudTable</li> <li>Search: Elasticsearch and Cloud Search Service (CSS)</li> </ul>	
	FusionInsight HDFS	<ul style="list-style-type: none"> <li>Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI)</li> </ul>	<ul style="list-style-type: none"> <li>FusionInsight cannot serve as the destination.</li> <li>Supported only by local storage and not in storage-compute</li> </ul>
	FusionInsight HBase	<ul style="list-style-type: none"> <li>Hadoop: MRS HDFS, MRS HBase, and MRS Hive</li> <li>Object-based storage: Object Storage Service (OBS)</li> <li>NoSQL: CloudTable</li> </ul>	

Category	Source	Destination	Description
	FusionInsight Hive	<ul style="list-style-type: none"> <li>Search: Elasticsearch and Cloud Search Service (CSS)</li> </ul>	<p>decoupling scenarios</p> <ul style="list-style-type: none"> <li>Not supported by Ranger</li> <li>Not supported if SSL is enabled for ZooKeeper</li> <li>Recommended FusionInsight HDFS versions:                             <ul style="list-style-type: none"> <li>2.8.X</li> <li>3.1.X</li> </ul> </li> <li>Recommended FusionInsight HBase versions:                             <ul style="list-style-type: none"> <li>2.1.X</li> <li>1.3.X</li> </ul> </li> <li>Recommended FusionInsight Hive versions:                             <ul style="list-style-type: none"> <li>1.2.X</li> <li>3.1.X</li> </ul> </li> </ul>
	Apache HBase Apache Hive	<ul style="list-style-type: none"> <li>Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI)</li> <li>Hadoop: MRS HDFS, MRS HBase, and MRS Hive</li> <li>Object-based storage: Object Storage Service (OBS)</li> <li>NoSQL: CloudTable</li> <li>Search: Elasticsearch and Cloud Search Service (CSS)</li> </ul>	<ul style="list-style-type: none"> <li>Apache cannot serve as the destination.</li> <li>Supported only by local storage and not in storage-compute decoupling scenarios</li> <li>Not supported by Ranger</li> <li>Not supported if SSL is enabled for ZooKeeper</li> <li>Recommended Apache HBase versions:                             <ul style="list-style-type: none"> <li>2.1.X</li> <li>1.3.X</li> </ul> </li> <li>Apache Hive 2.x versions are not</li> </ul>

Category	Source	Destination	Description
	Apache HDFS		<p>supported. The following versions are recommended:</p> <ul style="list-style-type: none"> <li>- 1.2.X</li> <li>- 3.1.X</li> </ul> <ul style="list-style-type: none"> <li>● Recommended Apache HDFS versions: <ul style="list-style-type: none"> <li>- 2.8.X</li> <li>- 3.1.X</li> </ul> </li> </ul>
Object Storage	Object Storage Service (OBS)	<ul style="list-style-type: none"> <li>● Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI)</li> <li>● Hadoop: MRS HDFS, MRS HBase, and MRS Hive</li> <li>● NoSQL: CloudTable</li> <li>● Search: Elasticsearch and Cloud Search Service (CSS)</li> </ul>	Object Storage Migration Service (OMS) is recommended for migration between object storage services.
File system	FTP	<ul style="list-style-type: none"> <li>● Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI)</li> <li>● Hadoop: MRS HDFS, MRS HBase, and MRS Hive</li> <li>● NoSQL: CloudTable</li> <li>● Search: Elasticsearch and Cloud Search Service (CSS)</li> </ul>	<ul style="list-style-type: none"> <li>● The file system cannot serve as the destination.</li> <li>● Only text files such as CSV files can be migrated from FTP or SFTP servers to search services. Binary files cannot.</li> <li>● obsutil is recommended for migrating data from file systems to OBS. For details, see .</li> </ul>
	SFTP		
	HTTP	Hadoop: MRS HDFS	
Relational database	RDS for MySQL	<ul style="list-style-type: none"> <li>● Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI)</li> <li>● Hadoop: MRS HDFS, MRS HBase, and MRS Hive</li> <li>● Object-based storage: Object Storage Service (OBS)</li> <li>● NoSQL: CloudTable</li> </ul>	<ul style="list-style-type: none"> <li>● You are advised to use Data Replication Service (DRS) to migrate data between OLTP databases.</li> <li>● RDS for MySQL does not</li> </ul>
	RDS for PostgreSQL		

Category	Source	Destination	Description
	RDS for SQL Server	<ul style="list-style-type: none"> <li>Relational database: RDS for MySQL, RDS for PostgreSQL, and RDS for SQL Server</li> <li>Search: Elasticsearch and Cloud Search Service (CSS)</li> </ul>	<ul style="list-style-type: none"> <li>support the SSL mode.</li> <li>Recommended Microsoft SQL Server version: 2005 or later</li> </ul>
	MySQL	<ul style="list-style-type: none"> <li>Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI)</li> </ul>	
	PostgreSQL	<ul style="list-style-type: none"> <li>Hadoop: MRS HDFS, MRS HBase, and MRS Hive</li> </ul>	
	Microsoft SQL Server	<ul style="list-style-type: none"> <li>Object-based storage: Object Storage Service (OBS)</li> </ul>	
	Oracle	<ul style="list-style-type: none"> <li>NoSQL: CloudTable</li> <li>Search: Elasticsearch and Cloud Search Service (CSS)</li> </ul>	

Category	Source	Destination	Description
	SAP HANA	<ul style="list-style-type: none"> <li>• Data warehouse: Data Lake Insight (DLI)</li> <li>• Hadoop: MRS Hive</li> </ul>	<p>SAP HANA data sources have the following restrictions:</p> <ul style="list-style-type: none"> <li>• SAP HANA cannot serve as the destination.</li> <li>• Only the 2.00.050.00.159 2305219 version is supported.</li> <li>• Only the Generic Edition is supported.</li> <li>• BW/4 FOR HANA is not supported.</li> <li>• Only database names, table names, and column names consisting of English letters are supported. Special characters such as spaces and symbols are not allowed.</li> <li>• The following data types are supported: date, digit, Boolean, and character (except SHORTTEXT). Other data types such as binary are not supported.</li> <li>• During migration, tables cannot be automatically created at the destination.</li> </ul>

Category	Source	Destination	Description
	Database sharding	<ul style="list-style-type: none"> <li>• Data warehouse: Data Lake Insight (DLI)</li> <li>• Hadoop: MRS HBase and MRS Hive</li> <li>• Search: Elasticsearch and Cloud Search Service (CSS)</li> <li>• Object-based storage: Object Storage Service (OBS)</li> </ul>	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.
	Redis		
	Document Database Service (DDS)		
	MongoDB		
	CloudTable	<ul style="list-style-type: none"> <li>• Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI)</li> <li>• Hadoop: MRS HDFS, MRS HBase, and MRS Hive</li> <li>• Object-based storage: Object Storage Service (OBS)</li> <li>• Relational database: RDS for MySQL, RDS for PostgreSQL, RDS for SQL Server, MySQL, PostgreSQL, Microsoft SQL Server, and Oracle</li> <li>• NoSQL: CloudTable</li> <li>• Search: Elasticsearch and Cloud Search Service (CSS)</li> </ul>	
Cassandra	<ul style="list-style-type: none"> <li>• Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI)</li> <li>• Hadoop: MRS HDFS, MRS HBase, and MRS Hive</li> <li>• Object-based storage: Object Storage Service (OBS)</li> <li>• NoSQL: CloudTable</li> <li>• Search: Elasticsearch and Cloud Search Service (CSS)</li> </ul>		

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	<ul style="list-style-type: none"> <li>Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI)</li> <li>Hadoop: MRS HDFS, MRS HBase, and MRS Hive</li> <li>Object-based storage: Object Storage Service (OBS)</li> <li>Relational database: RDS for MySQL, RDS for PostgreSQL, and RDS for SQL Server</li> <li>NoSQL: CloudTable</li> <li>Search: Elasticsearch and Cloud Search Service (CSS)</li> </ul>	<ul style="list-style-type: none"> <li>MRS Kafka cannot serve as the destination.</li> <li>Supported only by local storage and not in storage-compute decoupling scenarios</li> <li>Not supported by Ranger</li> <li>Not supported if SSL is enabled for ZooKeeper</li> </ul>
Search	Elasticsearch	<ul style="list-style-type: none"> <li>Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI)</li> <li>Hadoop: MRS HDFS, MRS HBase, and MRS Hive</li> </ul>	Only the non-security mode is supported.
	Cloud Search Service (CSS)	<ul style="list-style-type: none"> <li>Object-based storage: Object Storage Service (OBS)</li> <li>Relational database: RDS for MySQL, RDS for PostgreSQL, and RDS for SQL Server</li> <li>NoSQL: CloudTable</li> <li>Search: Elasticsearch and Cloud Search Service (CSS)</li> </ul>	You are advised to use Logstash to import data to CSS.

 **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 3-16** lists the data sources supporting entire DB migration using CDM.



**Table 3-16** 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: <ul style="list-style-type: none"> <li>• 2.1.X</li> <li>• 1.3.X</li> </ul>
	MRS Hive	Supported	Supported	Entire DB migration only to a relational database 2.x versions are not supported. The following versions are recommended: <ul style="list-style-type: none"> <li>• 1.2.X</li> <li>• 3.1.X</li> </ul>
	FusionInsight HBase	Supported	Not supported	Recommended versions: <ul style="list-style-type: none"> <li>• 2.1.X</li> <li>• 1.3.X</li> </ul>
	FusionInsight Hive	Supported	Not supported	Entire DB migration only to a relational database 2.x versions are not supported. The following versions are recommended: <ul style="list-style-type: none"> <li>• 1.2.X</li> <li>• 3.1.X</li> </ul>

Category	Data Source	Read	Write	Description
	Apache HBase	Supported	Not supported	Recommended versions: <ul style="list-style-type: none"> <li>• 2.1.X</li> <li>• 1.3.X</li> </ul>
	Apache Hive	Supported	Not supported	Entire DB migration only to a relational database 2.x versions are not supported. The following versions are recommended: <ul style="list-style-type: none"> <li>• 1.2.X</li> <li>• 3.1.X</li> </ul>
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	<ul style="list-style-type: none"> <li>• Only the 2.00.050.00.15 92305219 version is supported.</li> <li>• Only the Generic Edition is supported.</li> <li>• BW/4 FOR HANA is not supported.</li> <li>• Only database names, table names, and column names consisting of English letters are supported. Special characters such as spaces and symbols are not allowed.</li> <li>• The following data types are supported: date, digit, Boolean, and character (except SHORTTEXT). Other data types such as binary are not supported.</li> <li>• During migration, tables cannot be automatically created at the destination.</li> </ul>
	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 3-17** 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 NUMERIC.
	NUMERIC	Same as DECIMAL	-	DECIMAL	NUMERIC
	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 <i>M</i> bits of values, and <i>M</i> ranges from 1 to 64.	B'1111100' B'1100'	TINYINT	BYTEA
Time and date	DATE	The value is in the <i>YYYY-MM-DD</i> format and ranges from <b>1000-01-01</b> to <b>9999-12-31</b> . For example, <b>December 30, 1973</b> will be stored as <b>1973-12-30</b> .	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
	DATE TIME	The date and time are in the <i>YYYY-MM-DD HH:MM:SS</i> format and range from <b>1000-01-01 00:00:00</b> to <b>9999-12-31 23:59:59</b> . For example, <b>3:30 p.m. on December 30, 1973</b> will be stored as <b>1973-12-30 15:30:00</b> .	'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, <b>3:30 p.m. December 30, 1973</b> will be stored as <b>19731230153000</b> .	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
Multi media (binary)	BINARY(M)	The number of bytes is <i>M</i> . The length of a variable-length binary string ranges from 0 to <i>M</i> . <i>M</i> 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 <i>M</i> . A fixed binary string with a length of 0 to <i>M</i> .	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 3-18** 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 3-19** 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.	TIMES TAMP	TIMES TAMP	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.	TIMES TAMP	TIMES TAMP	Not supported
	timestamp	Timestamp data type	TIMES TAMP	TIMES TAMP	TIMES TAMP
	datetimeoffset	A time that uses the 24-hour clock and combined with date and the time zone.	Not supported (string)	TIMES TAMP	Not supported
Multimedia data types (binary)	text	Stores text data.	Not supported (string)	Not supported (string)	Not supported
	nettext	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 <i>n</i> bytes, where <i>n</i> 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
	small money	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 <b>rowversion</b> column in the database.	Not supported	Not supported	Not supported
	unique identifier	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

## 3.3.4 Managing Clusters

### 3.3.4.1 Creating a CDM Cluster

CDM provides independent clusters for secure and reliable data migration. Clusters are isolated from each other and cannot access each other.

CDM clusters can be used in the following scenarios:

- They can be used to create and run data migration jobs.
- They can function as agents for connecting Management Center to a data lake.

If you want to use DataArts Migration but the DataArts Studio instance has no CDM cluster, create one by following the instructions in section "(Optional) Creating an Incremental Package" in "Preparations" in the *DataArts Studio User Guide*.

### 3.3.4.2 Binding or Unbinding an EIP

#### Scenario

After a CDM cluster is created, you can bind an EIP to or unbind it from the CDM 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..
- 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 navigation pane, choose **Cluster Management**.

**Figure 3-30** Cluster list

Name	Status	Internal Network Address	Public Network Address	Source	Enterprise Project	Operation
	Running		--	CDM	default	Job Management Bind EIP More

### 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

### 3.3.4.3 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** Access the CDM console and choose **Cluster Management** in the left navigation pane.

**Figure 3-31** Cluster list

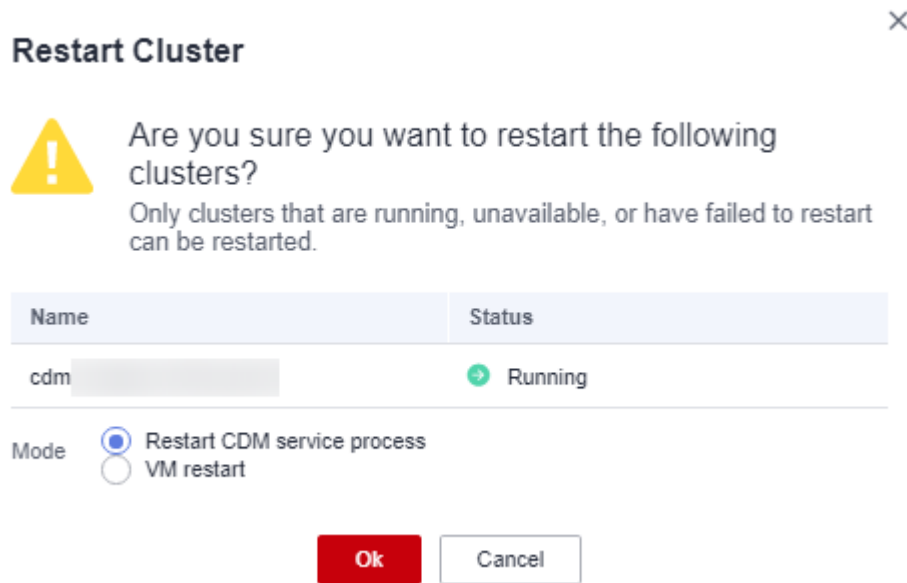
Name	Status	Internal Network Address	Public Network Address	Source	Enterprise Project	Operation
	Running		--	CDM	default	Job Management Bind EIP More

 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 3-32** 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

### 3.3.4.4 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](#).

## Prerequisites

You have created a CDM cluster.

## Deleting a Cluster

- Step 1** Access the CDM console and choose **Cluster Management** in the left navigation pane.

**Figure 3-33** Cluster list

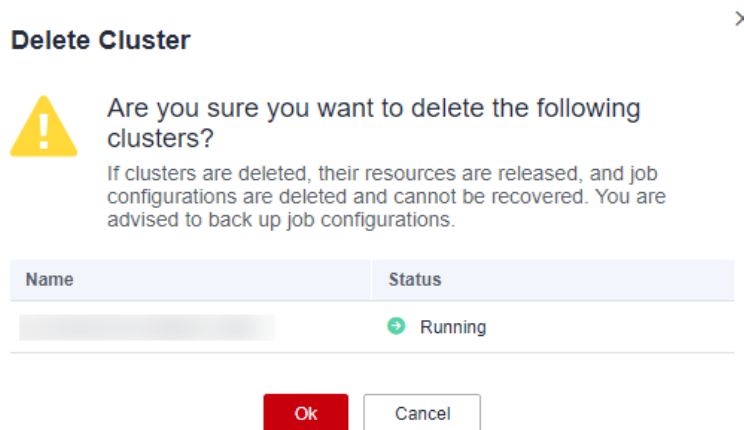
Name	Status	Internal Network Address	Public Network Address	Source	Enterprise Project	Operation
	Running			CDM	default	Job Management Bind EIP More

### 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 3-34** Deleting a cluster



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

----End

### 3.3.4.5 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** Access the CDM console and choose **Cluster Management** in the left navigation pane.

**Figure 3-35** Cluster list



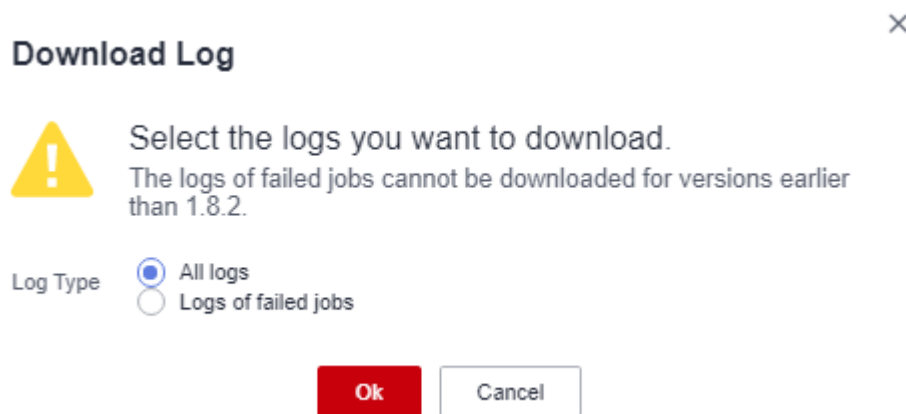
The screenshot shows a table with columns: Name, Status, Internal Network Address, Public Network Address, Source, Enterprise Project, and Operation. A single row is visible with a 'Running' status and 'CDM' as the source. The 'Operation' column has a dropdown menu with 'Job Management', 'Bind EIP', and 'More' options.

### 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 3-36** Download Log



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

----End

### 3.3.4.6 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.
    - 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 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 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 navigation pane, choose **Cluster Management**.

**Figure 3-37** Cluster list



Name	Status	Internal Network Address	Public Network Address	Source	Enterprise Project	Operation
	Running			CDM	default	Job Management   Bind EIP   More

 **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 navigation pane, choose **Cluster Management**.

**Figure 3-38** Cluster list



### NOTE

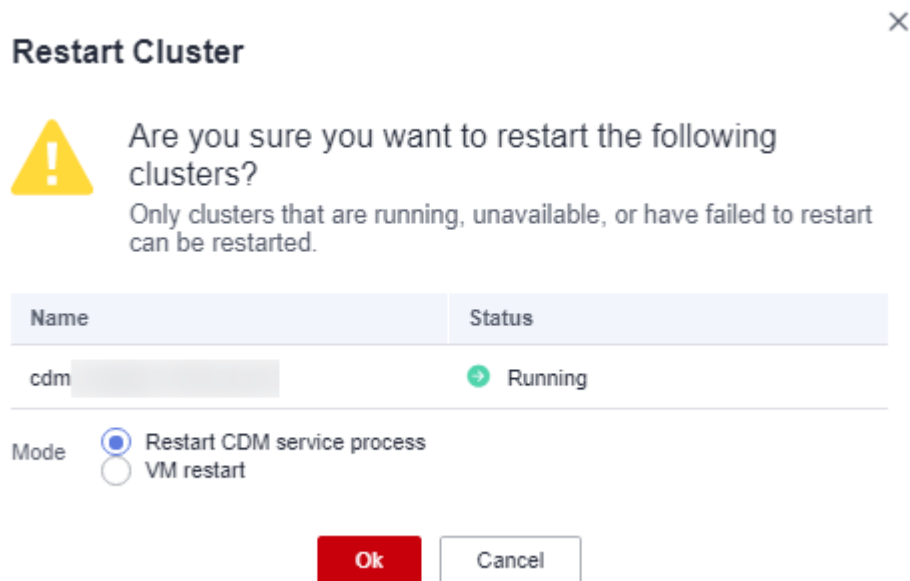
The **Source** column is displayed only when you access the **DataArts Migration** page from the DataArts Studio console.

**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 3-39** 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



### 3.3.4.7 Viewing Metrics

#### 3.3.4.7.1 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 3-20](#) lists the CDM metrics.

**Table 3-20** 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

ID	Name	Description	Value Range	Monitored Object	Monitoring Period (Raw Data)
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
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_connet	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	Tomcat 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

ID	Name	Description	Value Range	Monitored Object	Monitoring Period (Raw Data)
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
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

### 3.3.4.7.2 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 3-40** Monitored CDM clusters

Name	ID	Status	Operation
cdm-bang-cdm-dm-1-1	035a4d1-371c-4812-ae05-987af226a79	Running	View Metric   Create Alarm Rule
cdm-2002-test-cdm-dm-1-1	150a8299-844b-49aa-9f5c-418903ba727	Running	View Metric   Create Alarm Rule
cdm-a00000210-cdm-dm-1-1	24091154-aa4e-4517-9c0e-a20a12a6aa90	Running	View Metric   Create Alarm Rule
---	---	---	---
cdm-035c-3ea3-4795-678a-338a7a479a17	246033c-3ea3-4795-678a-338a7a479a17	---	View Metric   Create Alarm Rule
---	---	---	---
cdm-0a0d-c9a8-43af-a800-2107399c9a9c	2a95a0d-c9a8-43af-a800-2107399c9a9c	---	View Metric   Create Alarm Rule
---	---	---	---
cdm-4839-ca0b-4937-8c33-7789334d0a23	20aa4839-ca0b-4937-8c33-7789334d0a23	---	View Metric   Create Alarm Rule
---	---	---	---
cdm-d6-fa6d-cdm-dm-1-1	320a26a-1576-42aa-9a81-42682813884	Running	View Metric   Create Alarm Rule
---	---	---	---
cdm-425a-20f1-a80c-6a7a-9f80-d117569a6e91	425a20f1-a80c-6a7a-9f80-d117569a6e91	---	View Metric   Create Alarm Rule

**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 *Cloud Eye User Guide*.

----End

### 3.3.4.7.3 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

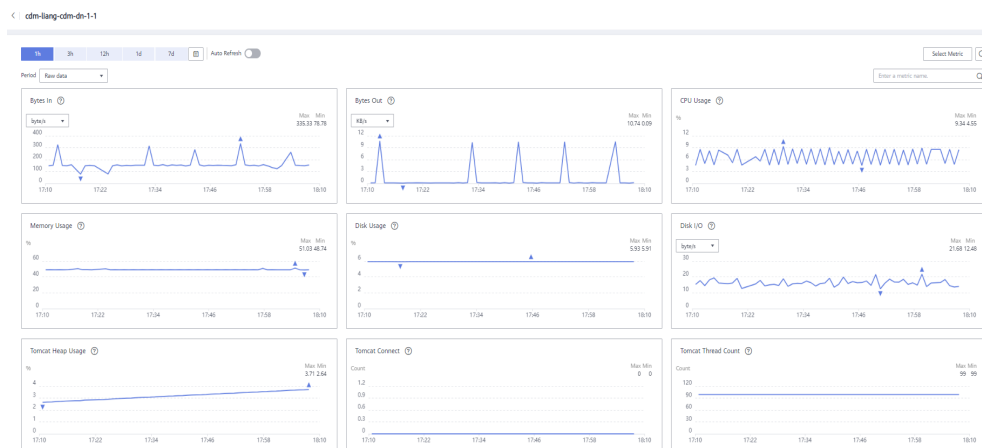
- The CDM cluster is running properly.  
If a cluster fails to be restarted or is unavailable, its monitoring metrics are unavailable. You can view the monitored data only after the cluster is restarted or recovered.
- The cluster has been properly running for about 10 minutes.  
The monitored data and 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 3-41** 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

## 3.3.5 Managing Links

### 3.3.5.1 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 **Adding Routes** in *Virtual Private Cloud (VPC) Usage Guide*. For details about how to configure security group rules, see **Security Group > Adding a Security Group Rule** in *Virtual Private Cloud (VPC) Usage Guide*.
  - 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, 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 3-42** 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 3-21](#) describes the link parameters.

**Table 3-21** Link parameters

Connector	Description
<ul style="list-style-type: none"> <li>Data Warehouse Service</li> <li>RDS for MySQL</li> <li>RDS for PostgreSQL</li> <li>RDS for SQL Server</li> <li>PostgreSQL</li> <li>Microsoft SQL Server</li> <li>SAP HANA</li> </ul>	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 <a href="#">Supported Data Sources</a> .
MySQL	For details about the parameters, see <a href="#">Link to a MySQL Database</a> .
Oracle	For details about the parameters, see <a href="#">Link to an Oracle Database</a> .
Database Sharding	For details about the parameters, see <a href="#">Link to a Database Shard</a> .
Object Storage Service (OBS)	For details about the parameters, see <a href="#">Link to OBS</a> .
<ul style="list-style-type: none"> <li>MRS HDFS</li> <li>FusionInsight HDFS</li> <li>Apache HDFS</li> </ul>	If the data source is HDFS of MRS, Apache Hadoop, or FusionInsight HD, see <a href="#">Link to HDFS</a> .

Connector	Description
<ul style="list-style-type: none"> <li>• MRS HBase</li> <li>• FusionInsight HBase</li> <li>• Apache HBase</li> </ul>	If the data source is HBase of MRS, Apache Hadoop, or FusionInsight HD, see <a href="#">Link to HBase</a> .
<ul style="list-style-type: none"> <li>• MRS Hive</li> <li>• FusionInsight Hive</li> <li>• Apache Hive</li> </ul>	If the data source is Hive on MRS, Apache Hadoop, or FusionInsight HD, see <a href="#">Link to Hive</a> .
CloudTable Service	If the data source is CloudTable, see <a href="#">Link to CloudTable</a> .
<ul style="list-style-type: none"> <li>• FTP</li> <li>• SFTP</li> </ul>	If the data source is an FTP or SFTP server, see <a href="#">Link to an FTP or SFTP Server</a> .
HTTP	<p>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.</p> <p>When creating an HTTP link, you only need to configure the link name. The URL is configured during job creation.</p>
MongoDB	If the data source is a local MongoDB, see <a href="#">Link to MongoDB</a> .
Document Database Service (DDS)	If the data source is DDS, see <a href="#">Link to DDS</a> .
<ul style="list-style-type: none"> <li>• Redis</li> <li>• Distributed Cache Service</li> </ul>	If the data source is Redis or DCS, see <a href="#">Link to Redis/DCS</a> .
<ul style="list-style-type: none"> <li>• MRS Kafka</li> <li>• Apache Kafka</li> </ul>	If the data source is MRS Kafka or Apache Kafka, see <a href="#">Link to Kafka</a> .
Cloud Search Service (CSS) Elasticsearch	If the data source is CSS or Elasticsearch, see <a href="#">Link to Elasticsearch/CSS</a> .
Data Lake Insight	If the data source is DLI, see <a href="#">Link to DLI</a> .
DMS Kafka	If the data source is DMS Kafka, see <a href="#">Link to DMS Kafka</a> .
Cassandra	If the data source is Cassandra, see <a href="#">Link to Cassandra</a> .

#### NOTE

Currently, the following data sources are in the OBT phase: FusionInsight HDFS, FusionInsight HBase, FusionInsight 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

### 3.3.5.2 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 3-22](#).
- (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 3-22](#).

**Table 3-22 Drivers**

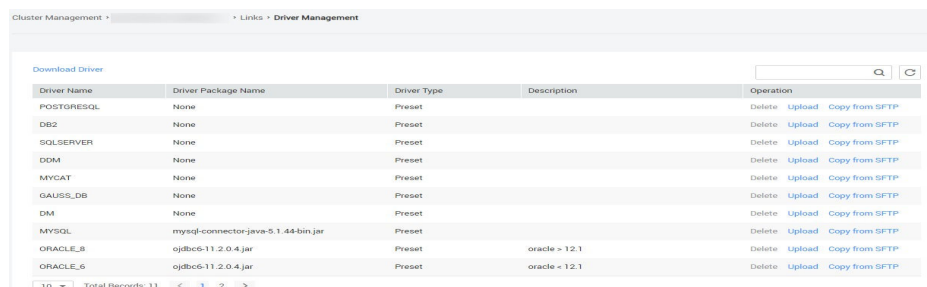
Relational Database Type	Driver Name	How to Obtain	Recommended Version
<ul style="list-style-type: none"> <li>• RDS for MySQL</li> <li>• MySQL</li> </ul>	MySQL MyCAT	<a href="https://downloads.mysql.com/archives/c-j/">https://downloads.mysql.com/archives/c-j/</a>	mysql-connector-java-5.1.48.jar
Oracle	ORACLE_6 ORACLE_7 ORACLE_8	Driver packages: <a href="https://www.oracle.com/database/technologies/appdev/jdbc-downloads.html">https://www.oracle.com/database/technologies/appdev/jdbc-downloads.html</a>  Driver packages of historical versions: <a href="https://repo1.maven.org/maven2/com/oracle/database/jdbc/ojdbc8/12.2.0.1/">https://repo1.maven.org/maven2/com/oracle/database/jdbc/ojdbc8/12.2.0.1/</a>	ojdbc8.jar for version 12.2.0.1  <b>NOTE</b> 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.
<ul style="list-style-type: none"> <li>• RDS for PostgreSQL</li> <li>• PostgreSQL</li> </ul>	POSTGRES RESQL	<a href="https://jdbc.postgresql.org/download">https://jdbc.postgresql.org/download</a>	postgresql-42.1.4.jar for JDBC 4.2

Relational Database Type	Driver Name	How to Obtain	Recommended Version
<ul style="list-style-type: none"> <li>RDS for SQL Server</li> <li>Microsoft SQL Server</li> </ul>	SQLServer	<p>Driver packages:</p> <p><a href="https://docs.microsoft.com/en-us/sql/connect/jdbc/download-microsoft-jdbc-driver-for-sql-server?view=sql-server-ver15">https://docs.microsoft.com/en-us/sql/connect/jdbc/download-microsoft-jdbc-driver-for-sql-server?view=sql-server-ver15</a></p> <p>Driver packages of historical versions:</p> <p><a href="https://docs.microsoft.com/en-us/sql/connect/jdbc/release-notes-for-the-jdbc-driver?view=sql-server-ver15#previous-releases">https://docs.microsoft.com/en-us/sql/connect/jdbc/release-notes-for-the-jdbc-driver?view=sql-server-ver15#previous-releases</a></p>	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 3-43** 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.

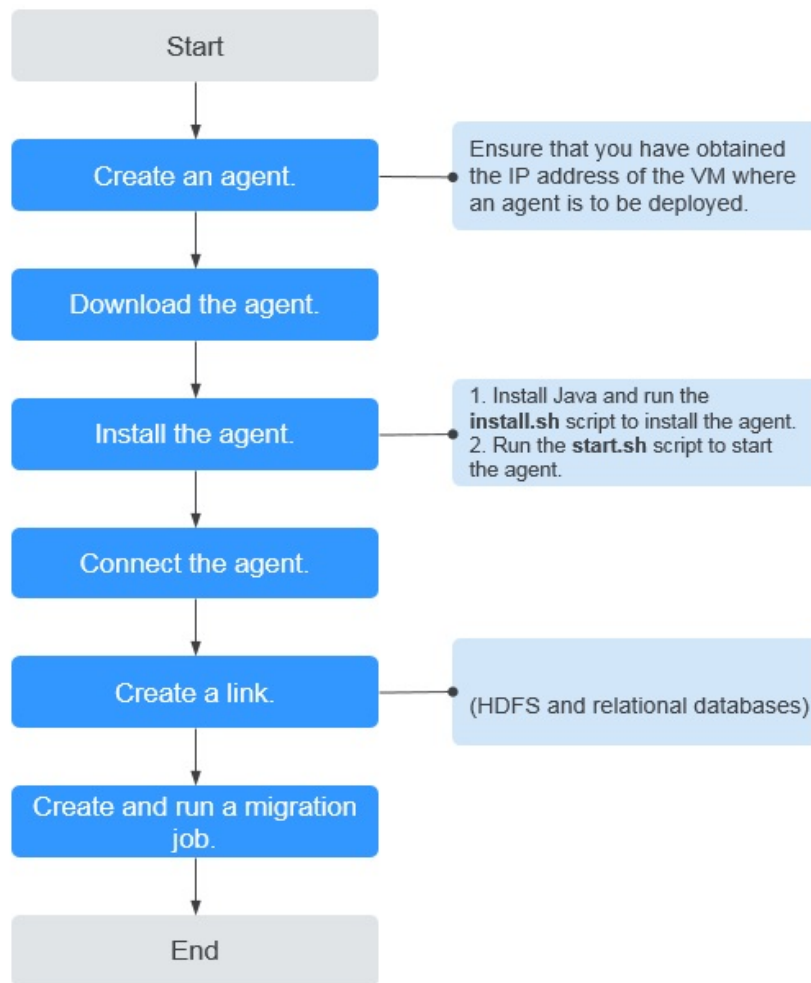
----End

### 3.3.5.3 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 3-44 shows the process of using an agent.

Figure 3-44 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 3-45 Creating an agent

The screenshot shows the 'Create Agent' dialog box. It includes the following elements:

- IP Address:** A text input field with a red asterisk and a dotted pattern, indicating a required field for an IP address.
- Port:** A text input field with a red asterisk, containing the value '24001'.
- Enable Compression:** A radio button group with 'Yes' selected.
- Enable SSL:** A radio button group with 'Yes' selected.
- Bandwidth Throttling:** A slider control ranging from 0 to 1000 MB/s. The current value is 0, and a tooltip indicates 'No throttling'.
- Buttons:** 'Yes' and 'No' buttons at the bottom.

- **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.

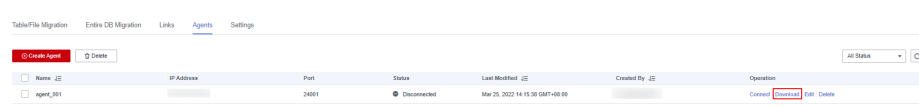
**Step 2** Click **OK**. On the **Agent Management** page, view the created agent.

----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 3-46** Downloading an agent



Name	IP Address	Port	Status	Last Modified	Created By	Operation
agent_001		2401	Disconnected	Mar 25, 2022 14:15:38 GMT+08:00		Connect Download Edit Delete

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

```
su Ruby
```

```
sh sbin/start.sh
```

**Step 7** Run the following command to check whether the agent is started:

```
ps -ef | grep agent
```

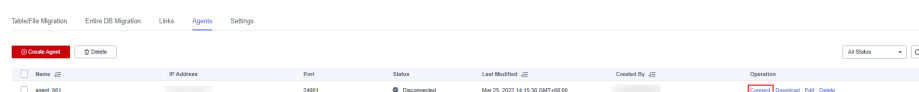
If the command output contains the running agent process, the agent process has been started.

----End

## Connecting to an Agent

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

**Figure 3-47** Connecting to an agent



Name	IP Address	Port	Status	Last Modified	Created By	Operation
agent_001		2401	Disconnected	Mar 25, 2022 14:15:38 GMT+08:00		Connect Download Edit Delete

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

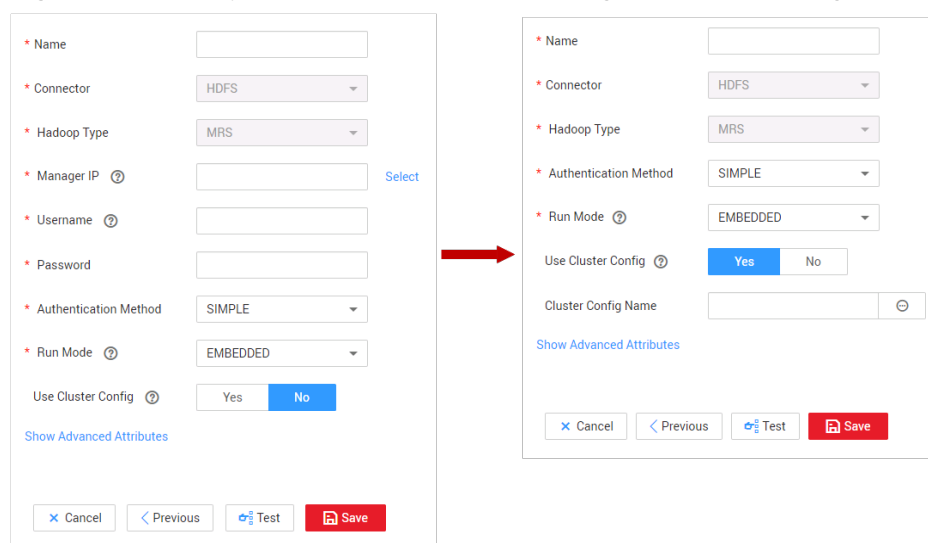
----End

### 3.3.5.4 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 3-48](#) for details.

**Figure 3-48** 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 3-23** Obtaining the cluster configuration file and keytab file

Hadoop Link	Obtaining the Cluster Configuration File	Obtaining the Keytab File
<p>MRS cluster</p> <ul style="list-style-type: none"> <li>• MRS HDFS</li> <li>• MRS HBase</li> <li>• MRS Hive</li> </ul>	<p>For clusters of MRS 3.x:</p> <ol style="list-style-type: none"> <li>1. Log in to FusionInsight Manager.</li> <li>2. Choose <b>Cluster &gt; Name of the desired cluster &gt; Dashboard &gt; More &gt; Download Client</b>.</li> <li>3. In the dialog box that is displayed, select <b>Configuration Files Only</b>. The platform type must be the same as that on the server. Click <b>OK</b> to download the configuration file to the local host.</li> <li>4. Obtain the downloaded TAR package, which is the FusionInsight cluster configuration file.</li> </ol> <p>For clusters of MRS 2.x or earlier:</p> <ol style="list-style-type: none"> <li>1. Log in to the MRS console.</li> <li>2. Choose <b>Clusters &gt; Active Clusters</b> and click a cluster name to go to the cluster details page. Click the <b>Components</b> tab.</li> <li>3. Click <b>Download Client</b>. Set <b>Client Type</b> to <b>Only configuration files</b>, set <b>Download To</b> to <b>Server</b> or <b>Remote host</b>, customize the client path, and click <b>OK</b> to generate the client configuration file.</li> <li>4. Save the generated configuration file to a local path.</li> </ol> <p>See MRS documentation for details.</p>	<p>For clusters of MRS 3.x:</p> <ol style="list-style-type: none"> <li>1. Log in to FusionInsight Manager.</li> <li>2. Choose <b>System &gt; Permission &gt; User</b>, locate the row that contains the target user, and choose <b>More &gt; Download Authentication Credential</b> to download the authentication credential file.</li> <li>3. Obtain the downloaded TAR package, which is the keytab file of the FusionInsight cluster.</li> </ol> <p>For clusters of MRS 2.x or earlier:</p> <ol style="list-style-type: none"> <li>1. Log in to MRS Manager and click <b>System</b>. In the <b>Permission</b> area, click <b>Manage User</b>.</li> <li>2. In the row of the user for whom you want to export the keytab file, choose <b>More &gt; Download authentication credential</b> to download the authentication file. After the file is automatically generated, save it to a specified path and keep it properly.</li> </ol> <p>See MRS documentation for details.</p>



Hadoop Link	Obtaining the Cluster Configuration File	Obtaining the Keytab File
<p>FusionInsight clusters:</p> <ul style="list-style-type: none"> <li>• FusionInsight HDFS</li> <li>• FusionInsight HBase</li> <li>• FusionInsight Hive</li> </ul>	<ol style="list-style-type: none"> <li>1. Log in to FusionInsight Manager.</li> <li>2. Choose <b>Cluster</b> &gt; <i>Name of the desired cluster</i> &gt; <b>Dashboard</b> &gt; <b>More</b> &gt; <b>Download Client</b>.</li> <li>3. In the dialog box that is displayed, select <b>Configuration Files Only</b>. The platform type must be the same as that on the server. Click <b>OK</b> to download the configuration file to the local host.</li> <li>4. Obtain the downloaded TAR package, which is the FusionInsight cluster configuration file.</li> </ol> <p>See the FusionInsight documentation for details.</p>	<ol style="list-style-type: none"> <li>1. Log in to FusionInsight Manager.</li> <li>2. Choose <b>System</b> &gt; <b>Permission</b> &gt; <b>User</b>, locate the row that contains the target user, and choose <b>More</b> &gt; <b>Download Authentication Credential</b> to download the authentication credential file.</li> <li>3. Obtain the downloaded TAR package, which is the keytab file of the FusionInsight cluster.</li> </ol> <p>See the FusionInsight documentation for details.</p>

Hadoop Link	Obtaining the Cluster Configuration File	Obtaining the Keytab File
<p>Apache clusters:</p> <ul style="list-style-type: none"> <li>• Apache HDFS</li> <li>• Apache HBase</li> <li>• Apache Hive</li> </ul>	<p>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.</p> <ul style="list-style-type: none"> <li>• HDFS needs to compress the following files into a .zip package without the directory format: <ul style="list-style-type: none"> <li>- hosts</li> <li>- core-site.xml</li> <li>- hdfs-site.xml</li> <li>- yarm-site.xml</li> <li>- mapred-site.xml</li> <li>- krb5.conf (optional, for clusters in security mode)</li> </ul> </li> <li>• HBase needs to compress the following files into a .zip package without the directory format: <ul style="list-style-type: none"> <li>- hosts</li> <li>- core-site.xml</li> <li>- hdfs-site.xml</li> <li>- yarm-site.xml</li> <li>- mapred-site.xml</li> <li>- hbase-site.xml</li> <li>- krb5.conf (optional, for clusters in security mode)</li> </ul> </li> <li>• Hive needs to compress the following files into a .zip package without the directory format: <ul style="list-style-type: none"> <li>- hosts</li> <li>- core-site.xml</li> <li>- hdfs-site.xml</li> <li>- yarm-site.xml</li> </ul> </li> </ul>	<p>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.</p> <ol style="list-style-type: none"> <li>1. Rename the user's authentication credential file as <b>user.keytab</b>.</li> <li>2. Compress the <b>user.keytab</b> file into a .zip package without the directory format: <b>user.keytab.zip</b>.</li> </ol>

Hadoop Link	Obtaining the Cluster Configuration File	Obtaining the Keytab File
	<ul style="list-style-type: none"><li>- mapred-site.xml</li><li>- hive-site.xml</li><li>- hivemetastore-site.xml</li><li>- krb5.conf (optional, for clusters in security mode)</li></ul>	

 **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 3-49** Creating cluster configurations

The screenshot shows a dialog box titled "Create Cluster Configuration" with a close button (X) in the top right corner. The dialog contains the following fields and controls:

- Configuration Name:** A text input field with a red asterisk (\*) indicating it is required.
- Configuration File:** A text input field with a help icon (?), a file selection icon (...), and an "Upload" button.
- Principal:** A text input field with a help icon (?).
- Keytab File:** A text input field with a help icon (?), a file selection icon (...), and an "Upload" button.
- Description:** A larger text input field.

At the bottom of the dialog, there are two buttons: "OK" (highlighted in red) and "Cancel".

- **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 3-50 Use Cluster Config**

### 3.3.5.5 Link to a Common Relational Database

Common relational databases include GaussDB(DWS), RDS for MySQL, RDS for PostgreSQL, RDS for SQLServer, PostgreSQL, Microsoft SQL Server, IBM Db2, 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 3-24](#) lists the link parameters.

**Table 3-24** 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 <b>Select</b> 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: <b>1433</b> Default port of PostgreSQL: <b>5432</b>
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 <b>Select</b> and select the agent created in <a href="#">Managing Agents</a> .	-
Driver Version	Select a driver version that adapts to the database type.	-
Fetch Size	(Optional) Displayed when you click <b>Show Advanced Attributes</b> . 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 <b>Yes</b> , 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 <b>Yes</b> .	Yes

Parameter	Description	Example Value
Link Attributes	(Optional) Click <b>Add</b> to add the JDBC connector attributes of multiple specified data sources. For details, see the JDBC connector document of the corresponding database.  <b>NOTE</b> By default, <b>useCursorFetch</b> is enabled, indicating that the JDBC connector communicates with relational databases using the binary protocol.	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.	'

### 3.3.5.6 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 3-25](#) lists the link parameters.

**Table 3-25** 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 take 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 take effect.	-
Use Agent	Whether to extract data from the data source through an agent	Yes
Agent	Click <b>Select</b> and select the agent created in <a href="#">Managing Agents</a> .	-

Parameter	Description	Example Value
backendData source	Enter the type of the backend database. Currently, only MySQL is supported.	MySQL
Data Source List	<p>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:db:username:password. You can leave username:password empty. In this case, the username and password are used.</p> <p>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.</p> <p>For example, <b>192.168.2.1:3306:cdm 192.168.2.2:3306:cdm:user:password</b> indicates that the IP address of the first backend database is <b>192.168.2.1</b>, the port number is <b>3306</b>, the database name is <b>cdm</b>, and the account name and password are configured in <i>user</i> and <i>password</i>. The IP address of the second backend database is <b>192.168.2.2</b>, the port number is <b>3306</b>, the database name is <b>cdm</b>, the account name is <b>user</b> and the password is <b>password</b>.</p>	192.168.2.1:3306:cdm 192.168.2.2:3306:cdm:user:password
Fetch Size	<p>(Optional) Displayed when you click <b>Show Advanced Attributes</b>.</p> <p>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.</p>	1000
Link Attributes	(Optional) Click <b>Add</b> 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.	'

### 3.3.5.7 Link to MyCAT

MyCAT is an open-source distributed database system. Its core function is to split a large table into multiple small tables and store them in the backend MySQL or other databases. [Table 3-26](#) lists the parameters for a MyCAT link.



**Table 3-26** MyCAT 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	mycat_link
Database Server	IP address or domain name of the database to connect  Click <b>Select</b> next to the text box and select a DWS or RDS DB instance in the displayed dialog box.	192.168.0.1
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 username	-
Use Local API	(Optional) Whether to use the local API of the database for acceleration.  When you create a link, CDM automatically enables the <b>local_infile</b> system variable of the MySQL database to enable the LOAD DATA function, which accelerates data import to the MySQL database.  If CDM fails to enable this function, contact the database administrator to enable the <b>local_infile</b> system variable. Alternatively, set <b>Use Local API</b> to <b>No</b> to disable API acceleration.	Yes
Create Backend Links	Whether to create backend links	Yes
managerUsername	MyCAT management username	root
managerPassword	MyCAT management password	123456
managerPort	MyCAT management port	9066
Backend Data Source	Type of the MyCAT backend database	MySQL
backendUsername	Username of the MyCAT backend database	cdm

Parameter	Description	Example Value
backendPassword	Password of the MyCAT backend database	-
Fetch Size	(Optional) Displayed when you click <b>Show Advanced Attributes</b> . 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 <b>Add</b> 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.	'

### 3.3.5.8 Link to a Dameng Database

When connecting CDM to a Dameng database, configure the parameters as described in [Table 3-27](#).

**Table 3-27** Parameters for a link to a Dameng 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	dm_link
Database Server	IP address or domain name of the database to connect Click <b>Select</b> next to the text box and select a DWS or RDS DB instance in the displayed dialog box.	192.168.0.1
Port	Port of the database to connect	The port number varies depending on the database.
Database Name	Name of the database to connect	dbname

Parameter	Description	Example Value
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	-
Agent	Click <b>Select</b> and select the agent created in <a href="#">Managing Agents</a> .	-
Driver Version	Select a driver version that adapts to the database type.	-
Fetch Size	(Optional) Displayed when you click <b>Show Advanced Attributes</b> . 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 <b>Yes</b> , 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 <b>Yes</b> .	Yes
Link Attributes	(Optional) Click <b>Add</b> to add the JDBC connector attributes of multiple specified data sources. For details, see the JDBC connector document of the corresponding database. <b>NOTE</b> By default, <b>useCursorFetch</b> is enabled, indicating that the JDBC connector communicates with relational databases using the binary protocol.	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.	'

### 3.3.5.9 Link to a MySQL Database

[Table 3-28](#) lists the parameters for a link to a MySQL database.

**Table 3-28** 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 <b>Select</b> next to the text box and select a MySQL DB instance in the displayed dialog box.	192.168.0.1
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	<p>(Optional) Whether to use the local API of the database for acceleration.</p> <p>When you create a MySQL link, CDM automatically enables the <b>local_infile</b> system variable of the MySQL database to enable the LOAD DATA function, which accelerates data import to the MySQL database.</p> <p>If CDM fails to enable this function, contact the database administrator to enable the <b>local_infile</b> system variable. Alternatively, set <b>Use Local API</b> to <b>No</b> to disable API acceleration.</p> <p>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 <b>local_infile</b> to <b>ON</b> to enable the LOAD DATA function.</p> <p><b>NOTE</b> If <b>local_infile</b> 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 <i>Relational Database Service User Guide</i>.</p>	Yes
Use Agent	Whether to extract data from the data source through an agent	Yes
Agent	Click <b>Select</b> and select the agent created in <a href="#">Managing Agents</a> .	-

Parameter	Description	Example Value
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.	-
Fetch Size	(Optional) Displayed when you click <b>Show Advanced Attributes</b> . 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 <b>Show Advanced Attributes</b> . 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.	-
Link Attributes	(Optional) Click <b>Add</b> to add the JDBC connector attributes of multiple specified data sources. For details, see the JDBC connector document of the corresponding database. <b>NOTE</b> By default, <b>useCursorFetch</b> is enabled, indicating that the JDBC connector communicates with relational databases using the binary protocol. The open-source MySQL database supports the <b>useCursorFetch</b> parameter. 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.	'
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

### 3.3.5.10 Link to an Oracle Database

**Table 3-29** lists the parameters for a link to an Oracle database.

**Table 3-29** 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: <ul style="list-style-type: none"> <li>• <b>Service Name:</b> Use <b>SERVICE_NAME</b> to connect to the Oracle database.</li> <li>• <b>SID:</b> Use <b>SID</b> to connect to the Oracle database.</li> </ul>	SID
Instance Name	Oracle instance ID, which is used to differentiate databases by instances. This parameter is available only when <b>Connection Type</b> is set to <b>SID</b> .	dbname
Database Name	Name of the database to connect This parameter is available only when <b>Connection Type</b> is set to <b>Service Name</b> .	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 <b>Select</b> and select the agent created in <a href="#">Managing Agents</a> .	-
Oracle Version	Oracle database version. This parameter is available only for Oracle links. If <b>java.sql.SQLException: Protocol violation</b> is displayed, select another version.	Later than 12.1

Parameter	Description	Example Value
Fetch Size	<p>(Optional) Displayed when you click <b>Show Advanced Attributes</b>.</p> <p>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.</p> <p>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 <b>Fetch Size</b> for the Oracle database.</p>	1000
Link Attributes	<p>(Optional) Click <b>Add</b> to add the JDBC connector attributes of multiple specified data sources. For details, see the JDBC connector document of the corresponding database.</p> <p><b>NOTE</b></p> <p>By default, <b>useCursorFetch</b> is enabled, indicating that the JDBC connector communicates with relational databases using the binary protocol.</p> <ul style="list-style-type: none"> <li>The open-source MySQL database supports the <b>useCursorFetch</b> parameter. You do not need to set this parameter.</li> </ul>	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.	'

### 3.3.5.11 Link to DLI

When connecting CDM to DLI, configure the parameters as described in [Table 3-30](#).

**Table 3-30** 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.	-
SK	You need to create an access key for the current account and obtain an AK/SK pair.	-

Parameter	Description	Example Value
Project ID	<p>Project ID in the region where DLI resides</p> <p>You can obtain the project ID and account ID by performing the following steps:</p> <ol style="list-style-type: none"> <li>1. Register with and log in to the management console.</li> <li>2. Hover the cursor on the username in the upper right corner and select <b>My Credentials</b> from the drop-down list.</li> <li>3. On the <b>My Credentials</b> page, obtain the account name and account ID, and obtain the project ID from the project list.</li> </ol>	-

### 3.3.5.12 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 cloud. [Table 3-31](#) 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 **Custom Route in Region Type I > Adding Routes** in *Virtual Private Cloud (VPC) Usage Guide*. For details about how to configure security group rules, see **Security Group > Adding a Security Group Rule** in *Virtual Private Cloud (VPC) Usage Guide*.
  - 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 3-31** 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 <b>Select</b> next to the <b>Manager IP</b> 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 <ul style="list-style-type: none"> <li>• <b>SIMPLE</b>: Select this for non-security mode.</li> <li>• <b>KERBEROS</b>: Select this for security mode.</li> </ul>	SIMPLE
HIVE Version	Hive version Set it to the Hive version on the server.	HIVE_3_X

Parameter	Description	Example Value
Username	<p>If <b>Authentication Method</b> is set to <b>KERBEROS</b>, 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.</p> <p>To create a data connection for an MRS security cluster, do not use user <b>admin</b>. The <b>admin</b> 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 <b>Username</b> and <b>Password</b> to the username and password of the created MRS user when creating an MRS data connection.</p> <p><b>NOTE</b></p> <ul style="list-style-type: none"> <li>• 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 <b>Manager_viewer</b> role to create links on CDM. To perform operations on databases, tables, and data of a component, you also need to add the user group permissions of the component to the user.</li> <li>• 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 <b>Manager_administrator</b> or <b>System_administrator</b> to create links on CDM.</li> <li>• A user with only the <b>Manager_tenant</b> or <b>Manager_auditor</b> permission cannot create connections.</li> </ul>	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	<p>This parameter is used only when the Hive version is <b>HIVE_3_X</b>. Possible values are:</p> <ul style="list-style-type: none"> <li>• <b>EMBEDDED</b>: The link instance runs with CDM. This mode delivers better performance.</li> <li>• <b>Standalone</b>: 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 <b>STANDALONE</b> or configure different agents.</li> </ul> <p>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.</p>	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	<p>This parameter is valid only when <b>Use Cluster Config</b> is set to <b>Yes</b>. Select a cluster configuration that has been created.</p> <p>For details, see <a href="#">Managing Cluster Configurations</a>.</p>	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.

## 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 3-32](#) describes related parameters.

**Table 3-32** 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: <ul style="list-style-type: none"> <li>● <b>SIMPLE</b>: Select this for non-security mode.</li> <li>● <b>KERBEROS</b>: Select this for security mode.</li> </ul>	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
Run Mode	This parameter is used only when the Hive version is <b>HIVE_3_X</b> . Possible values are: <ul style="list-style-type: none"> <li>● <b>EMBEDDED</b>: The link instance runs with CDM. This mode delivers better performance.</li> <li>● <b>Standalone</b>: 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 <b>STANDALONE</b> or configure different agents.</li> </ul> <p>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.</p>	EMBEDDED

Parameter	Description	Example Value
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 <b>Use Cluster Config</b> is set to <b>Yes</b> . Select a cluster configuration that has been created. For details, see <a href="#">Managing Cluster Configurations</a> .	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.

## 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 3-33](#) describes related parameters.

**Table 3-33** 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 <b>hive.metastore.uris</b> configuration item. Example: thrift://host-192-168-1-212:9083	-
Authentication Method	Authentication method used for accessing the cluster: <ul style="list-style-type: none"><li>• <b>SIMPLE</b>: Select this for non-security mode.</li><li>• <b>KERBEROS</b>: Select this for security mode.</li></ul>	SIMPLE
HIVE Version	Hive version	HIVE_3_X

Parameter	Description	Example Value
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 <b>Authentication Method</b> is set to <b>KERBEROS</b> , 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 <b>Authentication Method</b> is set to <b>KERBEROS</b> , 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	<p>This parameter is used only when the Hive version is <b>HIVE_3_X</b>. Possible values are:</p> <ul style="list-style-type: none"> <li>• <b>EMBEDDED</b>: The link instance runs with CDM. This mode delivers better performance.</li> <li>• <b>Standalone</b>: 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 <b>STANDALONE</b> or configure different agents.</li> </ul> <p>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.</p>	EMBEDDED

Parameter	Description	Example Value
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 <b>Use Cluster Config</b> is set to <b>Yes</b> . Select a cluster configuration that has been created. For details, see <a href="#">Managing Cluster Configurations</a> .	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.

### 3.3.5.13 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 3-34](#).

 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.
- 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 **Custom Route in Region Type I > Adding Routes** in *Virtual Private Cloud (VPC) Usage Guide*. For details about how to configure security group rules, see **Security Group > Adding a Security Group Rule** in *Virtual Private Cloud (VPC) Usage Guide*.
  - 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 3-34** 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 <b>Select</b> next to the <b>Manager IP</b> text box to select an MRS cluster. CDM automatically fills in the authentication information.	127.0.0.1



Parameter	Description	Example Value
Username	<p>If <b>Authentication Method</b> is set to <b>KERBEROS</b>, 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.</p> <p>To create a data connection for an MRS security cluster, do not use user <b>admin</b>. The <b>admin</b> 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 <b>Username</b> and <b>Password</b> to the username and password of the created MRS user when creating an MRS data connection.</p> <p><b>NOTE</b></p> <ul style="list-style-type: none"><li>• 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 <b>Manager_viewer</b> role to create links on CDM. To perform operations on databases, tables, and data of a component, you also need to add the user group permissions of the component to the user.</li><li>• 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 <b>Manager_administrator</b> or <b>System_administrator</b> to create links on CDM.</li><li>• A user with only the <b>Manager_tenant</b> or <b>Manager_auditor</b> permission cannot create connections.</li></ul>	cdm
Password	Password used for logging in to MRS Manager	-
Authentication Method	Authentication method used for accessing the cluster: <ul style="list-style-type: none"><li>• <b>SIMPLE</b>: for non-security mode</li><li>• <b>KERBEROS</b>: for security mode</li></ul>	SIMPLE
HBase Version	HBase version	HBASE_2_X

Parameter	Description	Example Value
Run Mode	<p>This parameter is used only when the HBase version is <b>HBASE_2_X</b>. Running mode of the HBase link. The options are as follows:</p> <ul style="list-style-type: none"><li>• <b>EMBEDDED</b>: The link instance runs with CDM. This mode delivers better performance.</li><li>• <b>STANDALONE</b>: 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 <b>STANDALONE</b> or configure different agents.</li></ul> <p>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.</p>	STANDALONE
Use Cluster Config	You can create cluster configurations on the <b>Links</b> page to simplify the configuration of Hadoop link parameters.	No
Cluster Config Name	<p>This parameter is valid only when <b>Use Cluster Config</b> is set to <b>Yes</b>. Select a cluster configuration that has been created.</p> <p>For details, see <a href="#">Managing Cluster Configurations</a>.</p>	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 3-35](#).

**Table 3-35** 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: <ul style="list-style-type: none"> <li>● <b>SIMPLE</b>: for non-security mode</li> <li>● <b>KERBEROS</b>: for security mode</li> </ul>	Kerberos
HBase Version	HBase version	HBASE_2_X
Run Mode	This parameter is used only when the HBase version is <b>HBASE_2_X</b> . Running mode of the HBase link. The options are as follows: <ul style="list-style-type: none"> <li>● <b>EMBEDDED</b>: The link instance runs with CDM. This mode delivers better performance.</li> <li>● <b>STANDALONE</b>: 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 <b>STANDALONE</b> or configure different agents.</li> </ul> <p>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.</p>	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 <b>Use Cluster Config</b> is set to <b>Yes</b> . Select a cluster configuration that has been created. For details, see <a href="#">Managing Cluster Configurations</a> .	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 3-36](#).

**Table 3-36** 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: <ul style="list-style-type: none"> <li>● <b>SIMPLE</b>: for non-security mode</li> <li>● <b>KERBEROS</b>: for security mode</li> </ul>	Kerberos
Principal	When <b>Authentication Method</b> is set to <b>KERBEROS</b> , 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 <b>Authentication Method</b> is set to <b>KERBEROS</b> , 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	This parameter is used only when the HBase version is <b>HBASE_2_X</b> . Running mode of the HBase link. The options are as follows: <ul style="list-style-type: none"> <li>• <b>EMBEDDED</b>: The link instance runs with CDM. This mode delivers better performance.</li> <li>• <b>STANDALONE</b>: 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 <b>STANDALONE</b> or configure different agents.</li> </ul> <p>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.</p>	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 <b>Use Cluster Config</b> is set to <b>Yes</b> . Select a cluster configuration that has been created. For details, see <a href="#">Managing Cluster Configurations</a> .	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.

### 3.3.5.14 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 3-37](#).

### 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.
- 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 **Custom Route in Region Type I > Adding Routes** in *Virtual Private Cloud (VPC) Usage Guide*. For details about how to configure security group rules, see **Security Group > Adding a Security Group Rule** in *Virtual Private Cloud (VPC) Usage Guide*.
  - 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 3-37** 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 <b>Select</b> next to the <b>Manager IP</b> text box to select an MRS cluster. CDM automatically fills in the authentication information.	127.0.0.1
Username	<p>If <b>Authentication Method</b> is set to <b>KERBEROS</b>, 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.</p> <p>To create a data connection for an MRS security cluster, do not use user <b>admin</b>. The <b>admin</b> 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 <b>Username</b> and <b>Password</b> to the username and password of the created MRS user when creating an MRS data connection.</p> <p><b>NOTE</b></p> <ul style="list-style-type: none"> <li>• 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 <b>Manager_viewer</b> role to create links on CDM. To perform operations on databases, tables, and data of a component, you also need to add the user group permissions of the component to the user.</li> <li>• 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 <b>Manager_administrator</b> or <b>System_administrator</b> to create links on CDM.</li> <li>• A user with only the <b>Manager_tenant</b> or <b>Manager_auditor</b> permission cannot create connections.</li> </ul>	cdm
Password	Password used for logging in to MRS Manager	-
Authentication Method	<p>Authentication method used for accessing MRS</p> <ul style="list-style-type: none"> <li>• <b>SIMPLE</b>: Select this for non-security mode.</li> <li>• <b>KERBEROS</b>: Select this for security mode.</li> </ul>	SIMPLE

Parameter	Description	Example Value
Run Mode	<p>Run mode of the HDFS link. The options are as follows:</p> <ul style="list-style-type: none"> <li>• <b>EMBEDDED</b>: The link instance runs with CDM. This mode delivers better performance.</li> <li>• <b>STANDALONE</b>: 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 <b>STANDALONE</b> or configure different agents. Note: The <b>STANDALONE</b> 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 <b>STANDALONE</b> process to prevent the migration failure caused by the conflict.</li> <li>• <b>Agent</b>: The link instance runs on an agent. If <b>Agent</b> 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 <b>EMBEDDED</b> mode, and the other clusters must be in <b>STANDALONE</b> mode.</li> </ul>	STANDALONE
Agent	Click <b>Select</b> and select the agent created in <a href="#">Connecting to an Agent</a> . This parameter is displayed when <b>Run Mode</b> is set to <b>Agent</b> .	-
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 <b>Use Cluster Config</b> is set to <b>Yes</b> . Select a cluster configuration that has been created. For details, see <a href="#">Managing Cluster Configurations</a> .	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 3-38](#).

**Table 3-38** 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: <ul style="list-style-type: none"> <li>● <b>SIMPLE</b>: Select this for non-security mode.</li> <li>● <b>KERBEROS</b>: Select this for security mode.</li> </ul>	KERBEROS

Parameter	Description	Example Value
Run Mode	<p>Run mode of the HDFS link. The options are as follows:</p> <ul style="list-style-type: none"><li>• <b>EMBEDDED</b>: The link instance runs with CDM. This mode delivers better performance.</li><li>• <b>STANDALONE</b>: 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 <b>STANDALONE</b> or configure different agents. Note: The <b>STANDALONE</b> 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 <b>STANDALONE</b> process to prevent the migration failure caused by the conflict.</li><li>• <b>Agent</b>: The link instance runs on an agent.</li></ul>	STANDALONE
Agent	Click <b>Select</b> and select the agent created in <a href="#">Connecting to an Agent</a> . This parameter is displayed when <b>Run Mode</b> is set to <b>Agent</b> .	-
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 <b>Use Cluster Config</b> is set to <b>Yes</b> . Select a cluster configuration that has been created. For details, see <a href="#">Managing Cluster Configurations</a> .	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 3-39](#).

**Table 3-39** 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 <b>hdfs://IP address of the NameNode instance:8020</b> .	hdfs:// <b>IP</b> :8020
Authentication Method	Authentication method used for accessing the cluster: <ul style="list-style-type: none"> <li>• <b>SIMPLE</b>: Select this for non-security mode.</li> <li>• <b>KERBEROS</b>: Select this for security mode.</li> </ul>	KERBEROS
Principal	When <b>Authentication Method</b> is set to <b>KERBEROS</b> , 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 <b>Authentication Method</b> is set to <b>KERBEROS</b> , 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	<p>Run mode of the HDFS link. The options are as follows:</p> <ul style="list-style-type: none"> <li>• <b>EMBEDDED</b>: The link instance runs with CDM. This mode delivers better performance.</li> <li>• <b>STANDALONE</b>: 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 <b>STANDALONE</b> 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.</li> <li>• <b>Agent</b>: The link instance runs on an agent.</li> </ul>	STANDALONE
IP and Host Name Mapping	<p>This parameter is used only when <b>Run Mode</b> is set to <b>EMBEDDED</b> or <b>STANDALONE</b>.</p> <p>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.</p>	<p>10.1.6.9 hostname01</p> <p>10.2.7.9 hostname02</p>
Agent	<p>If <b>Run Mode</b> is set to <b>Agent</b>, click <b>Select</b> and select the agent created in <a href="#">Connecting to an Agent</a>.</p>	-
Use Cluster Config	<p>You can use the cluster configuration to simplify parameter settings for the Hadoop connection.</p>	No
Cluster Config Name	<p>This parameter is valid only when <b>Use Cluster Config</b> is set to <b>Yes</b>. Select a cluster configuration that has been created. For details, see <a href="#">Managing Cluster Configurations</a>.</p>	hdfs_01

### 3.3.5.15 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 3-40](#).

**Table 3-40** 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	<p>You can obtain the endpoint by either of the following means:</p> <ul style="list-style-type: none"> <li>To obtain the endpoint of an OBS bucket, go to the OBS console and click the bucket name to go to its details page.</li> <li>An endpoint is the <b>request address</b> for calling an API. Endpoints vary depending on services and regions. You can obtain endpoints from .</li> </ul> <p>You can enter a bucket-level domain name, for example, <b>test.xx.com</b>. In this case, you can query only the <b>test</b> bucket.</p>	-
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, <b>Object Storage</b> .	Object Storage
AK	AK and SK are used to log in to the OBS server.	-
SK	<p>You need to create an access key pair for the current account and obtain an AK/SK pair.</p> <p>To obtain an access key, perform the following steps:</p>	-

### 3.3.5.16 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 3-41](#).

**Table 3-41** 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
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	-

### 3.3.5.17 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 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 3-42](#).

**Table 3-42** 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: <ul style="list-style-type: none"> <li>• <b>Single</b>: installation on a single-node system</li> <li>• <b>Cluster</b>: installation on a cluster</li> <li>• <b>Proxy</b>: installation using a proxy</li> </ul>	Single

Parameter	Description	Example Value
Redis Server List	List of MongoDB server addresses. Enter each address in the format of <i>IP address or domain name of the database server.port number</i> , 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

### 3.3.5.18 Link to DDS

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

When connecting CDM to DDS, configure the parameters as described in [Table 3-43](#).

**Table 3-43** 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 <i>IP address or domain name of the database server.port number</i> , 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	-

### 3.3.5.19 Link to CloudTable

When connecting CDM to CloudTable, configure the parameters as described in [Table 3-44](#).

**Table 3-44** 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
IAM Authentication	If IAM authentication is enabled for the CloudTable cluster to be connected, set this parameter to <b>Yes</b> . Otherwise, set this to <b>No</b> . If you select <b>Yes</b> , 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 <b>Use Cluster Config</b> is set to <b>Yes</b> . Select a cluster configuration that has been created. For details, see <a href="#">Managing Cluster Configurations</a> .	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.

### 3.3.5.20 Link to CloudTable OpenTSDB

When connecting CDM to CloudTable OpenTSDB, configure the parameters as described in [Table 3-45](#).



**Table 3-45** 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	TSDB_link
OpenTSDB Link	ZK link of OpenTSDB	opentsdb-sp8afz7bgbps5ur.cloudtable.com:4242
Security Mode	Security or non-security mode If you select <b>Security</b> , enter the project ID, username, and AK/SK.	Nonsecurity
Project ID	Project ID in the region where CloudTable resides You can obtain the project ID and account ID by performing the following steps: <ol style="list-style-type: none"><li>1. Register with and log in to the management console.</li><li>2. Hover the cursor on the username in the upper right corner and select <b>My Credentials</b> from the drop-down list.</li><li>3. On the <b>My Credentials</b> page, obtain the account name and account ID, and obtain the project ID from the project list.</li></ol>	-
Username	Username for accessing CloudTable	admin
AK	AK and SK for accessing CloudTable.	-
SK	You need to create an access key for the current account and obtain an AK/SK pair.	-

### 3.3.5.21 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 3-46](#).

**Table 3-46** 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
Server List	List of MongoDB server addresses. Enter each address in the format of <i>IP address or domain name of the database server.port number</i> , 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	-

### 3.3.5.22 Link to Cassandra

**Table 3-47** 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 <b>Show Advanced Attributes</b> . Connection timeout interval, in seconds.	5
Read timeout duration	(Optional) Displayed when you click <b>Show Advanced Attributes</b> . Read timeout interval, in seconds. If the value is less than or equal to 0, no timeout occurs.	12

### 3.3.5.23 Link to Kafka

#### MRS Kafka

When connecting CDM to Kafka of MRS, configure the parameters as described in [Table 3-48](#).

**Table 3-48** 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 <b>Select</b> next to the <b>Manager IP</b> text box to select an MRS cluster. CDM automatically fills in the authentication information.	-
Username	<p>Username used for logging in to MRS Manager</p> <p>To create a data connection for an MRS security cluster, do not use user <b>admin</b>. The <b>admin</b> 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 <b>Username</b> and <b>Password</b> to the username and password of the created MRS user when creating an MRS data connection.</p> <p><b>NOTE</b></p> <ul style="list-style-type: none"> <li>• 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 <b>Manager_viewer</b> role to create links on CDM. To perform operations on databases, tables, and data of a component, you also need to add the user group permissions of the component to the user.</li> <li>• 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 <b>Manager_administrator</b> or <b>System_administrator</b> to create links on CDM.</li> <li>• A user with only the <b>Manager_tenant</b> or <b>Manager_auditor</b> permission cannot create connections.</li> </ul>	-
Password	Password used for logging in to MRS Manager	-
Authentication Method	<p>Authentication method used for accessing MRS</p> <ul style="list-style-type: none"> <li>• <b>SIMPLE</b>: for non-security mode</li> <li>• <b>KERBEROS</b>: for security mode</li> </ul>	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 3-49](#).

**Table 3-49** 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.

### 3.3.5.24 Link to DMS Kafka

When connecting CDM to DMS Kafka, configure the parameters as described in [Table 3-50](#).

**Table 3-50** 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.	-

Parameter	Description	Example Value
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 <b>Kafka SASL_SSL</b> is enabled.	-
Password	Password for connecting to DMS Kafka. This parameter is displayed when <b>Kafka SASL_SSL</b> is enabled.	-

### 3.3.5.25 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 3-51](#).

**Table 3-51** 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 <i>ip:port</i> . 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 3-52](#).

**Table 3-52** 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 <i>ip:port</i> . 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 <b>Security Mode</b> has been enabled for the CSS cluster to be connected, set this parameter to <b>Yes</b> . Otherwise, set this to <b>No</b> .	Yes
Username	This parameter is displayed when <b>Security Mode Authentication</b> is set to <b>Yes</b> . It indicates the username used for connecting to CSS.	admin
Password	This parameter is displayed when <b>Security Mode Authentication</b> is set to <b>Yes</b> . It indicates the password used for connecting to CSS.	-
HTTPS Access	This parameter is displayed when <b>Security Mode Authentication</b> is set to <b>Yes</b> . This parameter specifies whether to enable HTTPS access. HTTPS access is more secure than HTTP access.	Yes

## 3.3.6 Managing Jobs

### 3.3.6.1 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 developed a job by following 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 3-51** Creating a migration job

The screenshot shows the 'Job Configuration' page. It features a 'Job Name' input field at the top. Below it, there are two columns: 'Source Job Configuration' and 'Destination Job Configuration'. Each column has a 'Source Link Name' or 'Destination Link Name' field with a red asterisk and a dropdown menu labeled 'Select a connector'. At the bottom of the form, there are 'Cancel' and 'Next' buttons.

**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.

The parameters vary with data sources. For details about the job parameters of other types of data sources, see [Table 3-53](#) and [Table 3-54](#).

**Table 3-53** 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 <a href="#">From OBS</a> .
<ul style="list-style-type: none"> <li>• MRS HDFS</li> <li>• FusionInsight HDFS</li> <li>• Apache HDFS</li> </ul>	HDFS data can be exported in CSV, Parquet, or binary format and can be compressed in multiple formats.	For details, see <a href="#">From HDFS</a> .
<ul style="list-style-type: none"> <li>• MRS HBase</li> <li>• FusionInsight HBase</li> <li>• Apache HBase</li> <li>• CloudTable Service</li> </ul>	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 <a href="#">From HBase/CloudTable</a> .
<ul style="list-style-type: none"> <li>• MRS Hive</li> <li>• FusionInsight Hive</li> <li>• Apache Hive</li> </ul>	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 <a href="#">From Hive</a> .
DLI	Data can be exported from DLI.	For details, see <a href="#">From DLI</a> .
<ul style="list-style-type: none"> <li>• FTP</li> <li>• SFTP</li> </ul>	FTP or SFTP data can be extracted in CSV, JSON, or binary format.	For details, see <a href="#">From FTP/SFTP</a> .
<ul style="list-style-type: none"> <li>• HTTP</li> </ul>	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 <a href="#">From HTTP</a> .
<ul style="list-style-type: none"> <li>• Data Warehouse Service</li> <li>• RDS for MySQL</li> <li>• RDS for SQL Server</li> <li>• RDS for PostgreSQL</li> </ul>	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,



Migration Source	Description	Parameter Settings
<ul style="list-style-type: none"> <li>• FusionInsight LibrA</li> </ul>	Data can be exported from FusionInsight LibrA.	see <a href="#">From a Common Relational Database</a> .
<ul style="list-style-type: none"> <li>• MySQL</li> <li>• PostgreSQL</li> <li>• Oracle</li> <li>• Microsoft SQL Server</li> <li>• SAP HANA</li> <li>• MyCAT</li> <li>• Database Sharding</li> </ul>	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.	
<ul style="list-style-type: none"> <li>• MongoDB</li> <li>• Document Database Service</li> </ul>	Data can be exported from MongoDB or DDS.	For details, see <a href="#">From MongoDB/DDS</a> .
Redis	Data can be exported from open source Redis.	For details, see <a href="#">From Redis</a> .
<ul style="list-style-type: none"> <li>• Apache Kafka</li> <li>• DMS Kafka</li> <li>• MRS Kafka</li> </ul>	Data can only be exported to Cloud Search Service (CSS).	For details, see <a href="#">From Kafka/DMS Kafka</a> .
<ul style="list-style-type: none"> <li>• Cloud Search Service</li> <li>• Elasticsearch</li> </ul>	Data can be exported from CSS or Elasticsearch.	For details, see <a href="#">From Elasticsearch or CSS</a> .

**Step 5** Configure job parameters for the migration destination based on [Table 3-54](#).

**Table 3-54** 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 <a href="#">To OBS</a> .
MRS HDFS	You can select a compression format when importing data to HDFS.	For details, see <a href="#">To HDFS</a> .
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 <a href="#">To HBase/CloudTable</a> .
MRS Hive	Data can be rapidly imported to MRS Hive.	For details, see <a href="#">To Hive</a> .

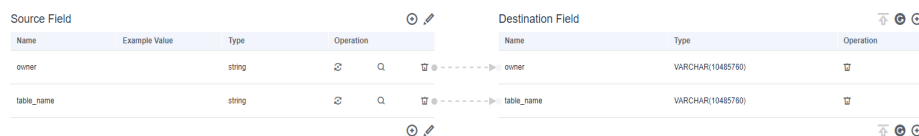
Migration Destination	Description	Parameter Settings
DLI	Data can be imported to DLI.	For details, see <a href="#">To DLI</a> .
<ul style="list-style-type: none"> <li>Data Warehouse Service</li> <li>RDS for MySQL</li> <li>RDS for SQL Server</li> <li>RDS for PostgreSQL</li> </ul>	Data can be imported to cloud database services.	For details about how to use the JDBC API to import data, see <a href="#">To a Common Relational Database</a> .
Document Database Service	Data can be imported to the DDS but cannot be imported to the local MongoDB.	For details, see <a href="#">To DDS</a> .
Distributed Cache Service	Data can be imported to DCS in the <b>String</b> or <b>HashMap</b> value type. Data cannot be imported to the local Redis.	For details, see <a href="#">To DCS</a> .
Cloud Search Service (CSS)	Data can be imported to CSS.	For details, see <a href="#">To CSS</a> .

**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 3-52** Field mapping

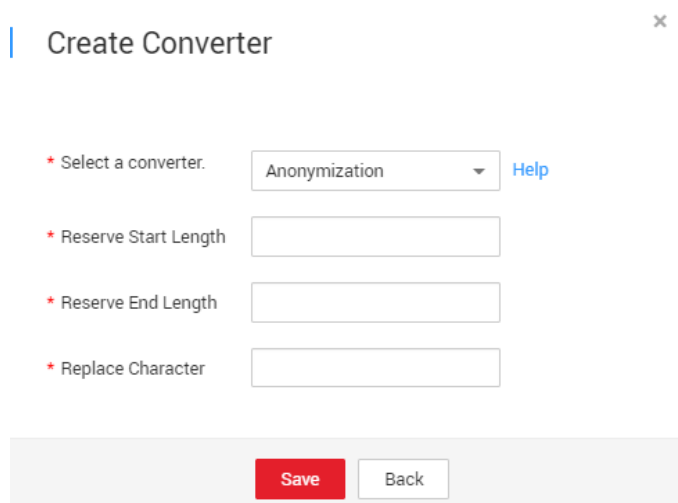


 NOTE

- 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 3-53** 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 .
- **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 3-54** Task parameters

### Configure Task

Retry if failed <sup>?</sup> Never ▼

Group <sup>?</sup> DEFAULT ▼ ⊕ Add ✎ Edit 🗑 Delete

Schedule Execution Yes No

[Hide Advanced Attributes](#)

Concurrent Extractors <sup>?</sup> 1

Write Dirty Data <sup>?</sup> Yes No

Write Dirty Data Link <sup>?</sup> OBS\_LINK1 ▼

OBS Bucket <sup>?</sup>  ⊖

Dirty Data Directory <sup>?</sup>  ⊖

Max. error records in a single shard. <sup>?</sup> 10

---

✕ Cancel
⏪ Previous
💾 Save
🏃 Save and Run

**Table 3-55** describes related parameters.

**Table 3-55** Parameter description

Parameter	Description	Example Value
Retry upon Failure	<p>You can select <b>Retry 3 times</b> or <b>Never</b>.</p> <p>You are advised to configure automatic retry for only file migration jobs or database migration jobs with <b>Import to Staging Table</b> enabled to avoid data inconsistency caused by repeated data writes.</p> <p><b>NOTE</b> 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 <b>Retry upon Failure</b> for the CDM node in DataArts Factory.</p>	Never
Job	<p>Select a group where the job resides. The default group is <b>DEFAULT</b>. On the <b>Job Management</b> page, jobs can be displayed, started, or exported by group.</p>	DEFAULT
Schedule Execution	<p>If you select <b>Yes</b>, you can set the start time, cycle, and validity period of a job. For details, see <a href="#">Scheduling Job Execution</a>.</p> <p><b>NOTE</b> 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 this parameter.</p>	No

Parameter	Description	Example Value
Concurrent Extractors	<p>Number of extraction tasks that can be concurrently executed. The value range is 1 to 300. If the value is too large, the extractors are queued.</p> <p>The number of concurrent extractors in a CDM migration job is related to the cluster specifications and table size.</p> <ul style="list-style-type: none"> <li>You are advised to set this parameter to <b>4</b> for each CU (1 CPU and 4 GB) based on the cluster specifications.</li> <li>If each row of the table contains less than or equal to 1 MB data, you can extract data concurrently. If each row contains more than 1 MB data, you are advised to extract data in a single thread.</li> </ul> <p><b>NOTE</b></p> <ul style="list-style-type: none"> <li>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.</li> <li>The number of concurrent extractors of a job is affected by <b>Maximum Concurrent Extractors</b> configured on the <b>Settings</b> page. The <b>Maximum Concurrent Extractors</b> parameter specifies the total number of concurrent extractions.</li> </ul>	1
Concurrent Loaders	<p>Number of Loaders to be concurrently executed</p> <p>This parameter is displayed only when HBase or Hive serves as the destination data source.</p>	3
Number of split retries	<p>Number of retries when a split fails to be executed. Value <b>0</b> indicates that no retry will be performed.</p>	0
Write Dirty Data	<p>Whether to record dirty data. By default, this parameter is set to <b>No</b>.</p> <p>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.</p>	Yes
Write Dirty Data Link	<p>This parameter is displayed only when <b>Write Dirty Data</b> is set to <b>Yes</b>.</p> <p>Only links to OBS support dirty data writes.</p>	obs_link
OBS Bucket	<p>This parameter is displayed only when <b>Write Dirty Data Link</b> is a link to OBS.</p> <p>Name of the OBS bucket to which the dirty data will be written.</p>	dirtydata

Parameter	Description	Example Value
Dirty Data Directory	<p>This parameter is displayed only when <b>Write Dirty Data</b> is set to <b>Yes</b>.</p> <p>Dirty data is stored in the directory for storing dirty data on OBS. Dirty data is saved only when this parameter is configured.</p> <p>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.</p>	/user/dirtydir
Max. Error Records in a Single Shard	<p>This parameter is displayed only when <b>Write Dirty Data</b> is set to <b>Yes</b>.</p> <p>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.</p>	0
Throttling	<p>Enabling throttling reduces the read pressure on the source. It controls the CDM transmission rate, not the NIC traffic.</p> <p><b>NOTE</b></p> <ul style="list-style-type: none"> <li>• It can control the rate for a job migrating data to Hive, DLI, relational databases, OBS, or HDFS.</li> <li>• To configure throttling for multiple jobs, multiply the rate by the number of concurrent jobs.</li> </ul>	Yes
Max. error records in a single shard	<p>Maximum rate for a job. To configure throttling for multiple jobs, multiply the rate by the number of concurrent jobs.</p> <p><b>NOTE</b> The rate is an integer greater than 1.</p>	20

**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

### 3.3.6.2 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 3-55](#) 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 3-55** 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 a connection by following 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 **Entire DB Migration > Create Job**. The page for configuring the job is displayed.
- Step 3** Configure the related parameters of the source database according to [Table 3-56](#).

**Table 3-56** Parameter description

Source Database	Parameter	Description	Example Value
<ul style="list-style-type: none"> <li>• DWS</li> <li>• FusionInsight LibrA</li> <li>• MySQL</li> <li>• PostgreSQL</li> <li>• SQL Server</li> <li>• Oracle</li> <li>• SAP HANA</li> <li>• MyCAT</li> </ul>	Schema/ Tablespace	Name of the schema or tablespace from which data will be extracted. This parameter is displayed when <b>Use SQL Statement</b> is set to <b>No</b> . 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
	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.	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

Source Database	Parameter	Description	Example Value
HBase CloudTable	Start Time	Start time (included). The format is <i>yyyy-MM-dd hh:mm:ss</i> . The dateformat time macro variable function is supported. Examples: <b>2017-12-31 20:00:00</b> , \$ <b>{dateformat(yyyy-MM-dd, -1, DAY)} 02:00:00</b> , and \$ <b>{dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}</b>	-
	End Time	End time (excluded) The format is <i>yyyy-MM-dd hh:mm:ss</i> . The dateformat time macro variable function is supported. Examples: <b>2018-01-01 20:00:00</b> , \$ <b>{dateformat(yyyy-MM-dd, -1, DAY)} 02:00:00</b> , and \$ <b>{dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}</b>	-
Redis	Key Filter Character	Filter character used to determine the keys to be migrated For example, if the value of this parameter is <b>a*</b> , 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.	mongod b
	Query Filter	Filter used to match documents. Example: <b>{HTTPStatusCode: {\$gt:"400"}, \$lt:"500"},HTTPMethod:"GET"}</b>	-
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 <b>cdm*</b> , CDM migrates all indexes starting with <b>cdm</b> , such as <b>cdm01</b> , <b>cdmB3</b> , <b>cdm_45</b> and so on.  If multiple indexes are migrated at the same time, <b>Index</b> cannot be configured at the migration destination.	cdm*

**Step 4** Configure the related parameters, from [Table 3-57](#), for the destination cloud service.

**Table 3-57** Destination job parameters

Source Database	Parameter	Description	Example Value
<ul style="list-style-type: none"> <li>• DWS</li> <li>• FusionInsight LibrA</li> <li>• MySQL</li> <li>• PostgreSQL</li> <li>• SQL Server</li> </ul>	-	For details about the destination job parameters required for entire DB migration to a relational database, see <a href="#">To a Common Relational Database</a> .	schema
MRS HIVE	-	For details about the destination job parameters required for entire DB migration to MRS HIVE, see <a href="#">To Hive</a> .	hivedb
MRS HBase CloudTable	-	For details about the destination job parameters required for entire DB migration to MRS HBase or CloudTable, see <a href="#">To HBase/CloudTable</a> .	Yes
MRS HDFS	-	For details about the destination job parameters required for entire DB migration to MRS HDFS, see <a href="#">To HDFS</a> .	-
OBS	-	For details about the destination job parameters required for entire database migration to OBS, see <a href="#">To OBS</a> .	-
DCS	-	For details about the destination job parameters required for entire database migration to DCS, see <a href="#">To DCS</a> .	-
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	Add If there is already a value, replace it; otherwise, add a value. Replace	-

Source Database	Parameter	Description	Example Value
CSS	Index	<p>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 <b>cdm*</b>, CDM migrates all indexes starting with <b>cdm</b>, such as <b>cdm01</b>, <b>cdmB3</b>, <b>cdm_45</b> and so on.</p> <p>If multiple indexes are migrated at the same time, <b>Index</b> cannot be configured at the migration destination.</p>	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 3-56** Task parameters

Concurrent Extractors tables ?

Concurrent Extractors ?

Write Dirty Data ? Yes No

Write Dirty Data Link ?

OBS Bucket ?  ⋮

Dirty Data Directory ?  ⋮

Max. error records in a single shard. ?

< Previous Save Save and Run

**Table 3-58** describes related parameters.

**Table 3-58** 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 <b>No</b> .	Yes
Write Dirty Data Link	This parameter is only displayed when <b>Write Dirty Data</b> is set to <b>Yes</b> . Only links to OBS support dirty data writes.	obs_link
OBS Bucket	This parameter is only displayed when <b>Write Dirty Data Link</b> is a link to OBS. Name of the OBS bucket to which the dirty data will be written.	dirtydata
Dirty Data Directory	This parameter is only displayed when <b>Write Dirty Data</b> is set to <b>Yes</b> . 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 <b>Write Dirty Data</b> is set to <b>Yes</b> . 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

### 3.3.6.3 Source Job Parameters

### 3.3.6.3.1 From OBS

If the source link of a job is the [Link to OBS](#), configure the source job parameters based on [Table 3-59](#).

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

**Table 3-59** Parameter description

Category	Parameter	Description	Example Value
Basic parameters	Bucket Name	Name of the bucket from which data will be migrated	BUCKET_2
	Source Directory/File	This parameter is available only when <b>Pull List File</b> is set to <b>No</b> . 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.  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.	FROM/ example.csv
	File Format	Format in which CDM parses data. The options are as follows: <ul style="list-style-type: none"> <li>• <b>CSV</b>: Source files will be migrated to tables after being converted to CSV format.</li> <li>• <b>Binary</b>: Files (even not in binary format) will be transferred directly. It is used for file copy.</li> <li>• <b>JSON</b>: Source files will be migrated to tables after being converted to JSON format.</li> </ul>	CSV

Category	Parameter	Description	Example Value
	Pull List File	This parameter is displayed only when <b>File Format</b> is set to <b>Binary</b> . 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 <b>Pull List File</b> is set to <b>Yes</b> . 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 <b>Pull List File</b> is set to <b>Yes</b> . 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 <b>Pull List File</b> is set to <b>Yes</b> . 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 <b>File Format</b> is set to <b>JSON</b> . Type of a JSON object stored in a JSON file. The options are <b>JSON object</b> and <b>JSON array</b> .	JSON object
	JSON Reference Node	This parameter is used only when <b>File Format</b> is set to <b>JSON</b> and <b>JSON Type</b> is set to <b>JSON Object</b> . 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	Line feed character in a file. By default, the system automatically identifies <b>\n</b> , <b>\r</b> , and <b>\r\n</b> . This parameter is displayed only when <b>File Format</b> is set to <b>CSV</b> .	\n
	Field Delimiter	Character used to separate fields in the file. To set the <b>Tab</b> key as the delimiter, set this parameter to <b>\t</b> . This parameter is displayed only when <b>File Format</b> is set to <b>CSV</b> .	,
	Use Quote Character	If you set this parameter to <b>Yes</b> , the field delimiters in the encircling symbol are regarded as a part of the string value. Currently, the default encircling symbol of CDM is <b>"</b> .	No
	Use RE to Separate Fields	Whether to use regular expressions to separate fields. If you set this parameter to <b>Yes</b> , <b>Field Delimiter</b> becomes invalid. This parameter is displayed only when <b>File Format</b> is set to <b>CSV</b> .	Yes
	Regular Expression	Regular expression used to separate fields.	^(\\d.*\\d) (\\w*) \\[(.*) \\] ([\\w\\.]* (\\w.*)*.
	Use First Row as Header	This parameter is displayed only when <b>File Format</b> is set to <b>CSV</b> . When you migrate a CSV file to a table, CDM writes all data to the table by default. If you set this parameter to <b>Yes</b> , 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, <b>UTF-8</b> or <b>GBK</b> . You can set the encoding type for text files only. This parameter is invalid when <b>File Format</b> is set to <b>Binary</b> .	GBK



Category	Parameter	Description	Example Value
	Compression Format	<p>This parameter is displayed only when <b>File Format</b> is set to <b>CSV</b> or <b>JSON</b>. The options are as follows:</p> <ul style="list-style-type: none"> <li>• <b>NONE</b>: Files in all formats can be transferred.</li> <li>• <b>GZIP</b>: Only files in gzip format can be transferred.</li> <li>• <b>ZIP</b>: Only files in Zip format can be transferred.</li> <li>• <b>TAR.GZ</b>: Files in TAR.GZ format are transferred.</li> </ul>	NONE
	Compressed File Suffix	<p>This parameter is displayed when <b>Compression Format</b> is not <b>NONE</b>. 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.</p>	*
	Source File Processing Method	<p>Operation performed on source files after the job completes.</p> <ul style="list-style-type: none"> <li>• No action</li> <li>• <b>Rename</b>: After the job completes, the source files are renamed by appending usernames and timestamps as suffixes to the file names.</li> <li>• <b>Delete</b>: After the job completes, the source files are deleted.</li> </ul>	No action
	Start Job by Marker File	<p>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 <b>Suspension Period</b>.</p>	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 <b>Start Job by Marker File</b> to <b>Yes</b> 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 <b>0</b> 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 <b>Source Directory/Files</b> , 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 <b>None</b> , <b>Wildcard</b> , and <b>Regex</b> .	Wildcard
	Directory Filter	If you set <b>Filter Type</b> to <b>Wildcard</b> , enter a wildcard character to filter paths. The paths that meet the filtering condition are migrated. You can configure multiple paths separated by commas (,).	*input
	File Filter	If you set <b>Filter Type</b> to <b>Wildcard</b> , 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 (,).	*.csv,*.txt
	Time Filter	If you select <b>Yes</b> , files are transferred based on their modification time.	Yes

Category	Parameter	Description	Example Value
	Minimum Timestamp	<p>If you set <b>Filter Type</b> to <b>Time Filter</b>, and specify a point in time for this parameter, only the files modified after the specified time are transferred. The time format must be <i>yyyy-MM-dd HH:mm:ss</i>.</p> <p>This parameter can be set to a macro variable of date and time. For example, \$<b>{timestamp(dateformat(yyyy-MM-dd HH:mm:ss,-90,DAY))}</b> indicates that only files generated within the latest 90 days are migrated.</p>	2019-06-01 00:00:00
	Maximum Timestamp	<p>If you set <b>Filter Type</b> to <b>Time Filter</b>, and specify a point in time for this parameter, only the files modified before the specified time are transferred. The time format must be <i>yyyy-MM-dd HH:mm:ss</i>.</p> <p>This parameter can be set to a macro variable of date and time. For example, \$<b>{timestamp(dateformat(yyyy-MM-dd HH:mm:ss))}</b> indicates that only the files whose modification time is earlier than the current time are migrated.</p>	2019-07-01 00:00:00
	Encryption	<p>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:</p> <ul style="list-style-type: none"> <li>• <b>NONE</b>: Export data without decrypting it.</li> <li>• <b>AES-256-GCM</b>: 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.</li> </ul>	AES-256-GCM
	Disregard Non-existent Path or File	<p>If this is set to <b>Yes</b>, the job can be successfully executed even if the source path does not exist.</p>	No

Category	Parameter	Description	Example Value
	DEK	This parameter is displayed only when <b>Encryption</b> is set to <b>AES-256-GCM</b> . The key consists of 64 hexadecimal numbers and must be the same as the <b>DEK</b> 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 FEC78BF0 51BCFDA2 5BD4E320 DB0A7AC7 5A1F3FC3D 3C56A457 DCDC1B
	IV	This parameter is displayed only when <b>Encryption</b> is set to <b>AES-256-GCM</b> . The initialization vector consists of 32 hexadecimal numbers and must be the same as the <b>IV</b> 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 <b>File Format</b> is set to <b>Binary</b> .  This parameter is used to check whether the files extracted by CDM are consistent with source files.	.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.

### 3.3.6.3.2 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 3-60](#).

**Table 3-60** 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 <b>Pull List File</b> is set to <b>No</b> . 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.	/user/cdm/
	File Format	File format used when transferring data. The options are as follows: <ul style="list-style-type: none"> <li>• <b>CSV</b>: Source files will be migrated to tables after being converted to CSV format.</li> <li>• <b>Binary</b>: Files (even not in binary format) will be transferred directly. It is used for file copy.</li> <li>• <b>Parquet</b>: Source files will be migrated to tables after being converted to Parquet format.</li> </ul>	CSV

Category	Parameter	Description	Example Value
	Pull List File	<p>This parameter is displayed only when <b>File Format</b> is set to <b>Binary</b>.</p> <p>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:</p> <pre>/mrs/job-properties/ application_1634891604621_0014/ job.properties /mrs/job-properties/ application_1634891604621_0029/ job.properties</pre>	Yes
	OBS Link of List File	This parameter is available only when <b>Pull List File</b> is set to <b>Yes</b> . 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 <b>Pull List File</b> is set to <b>Yes</b> . 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 <b>Pull List File</b> is set to <b>Yes</b> . It indicates the absolute path or directory of the list file in the OBS bucket.	/0521/ Lists.txt
Advanced attributes	Line Separator	Line feed character in a file. By default, the system automatically identifies <b>\n</b> , <b>\r</b> , and <b>\r\n</b> . This parameter is displayed only when <b>File Format</b> is set to <b>CSV</b> .	\n
	Field Delimiter	Character used to separate fields in the file. To set the <b>Tab</b> key as the delimiter, set this parameter to <b>\t</b> . This parameter is displayed only when <b>File Format</b> is set to <b>CSV</b> .	,

Category	Parameter	Description	Example Value
	Use First Row as Header	This parameter is displayed only when <b>File Format</b> is set to <b>CSV</b> . When you migrate a CSV file to a table, CDM writes all data to the table by default. If you set this parameter to <b>Yes</b> , 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. <ul style="list-style-type: none"> <li>• <b>No action</b></li> <li>• <b>Rename:</b> After the job completes, the source files are renamed by appending usernames and timestamps as suffixes to the file names.</li> <li>• <b>Delete:</b> After the job completes, the source files are deleted.</li> </ul>	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 <b>Suspension Period</b> .	ok.txt
	Filter Type	Only paths or files that meet the filtering conditions are transferred. The options are <b>None</b> , <b>Wildcard</b> , and <b>Regex</b> .	-
	Path Filter	If you set <b>Filter Type</b> to <b>Wildcard</b> , enter a wildcard character to filter paths. The paths that meet the filtering condition are migrated. You can configure multiple paths separated by commas (,).	*input

Category	Parameter	Description	Example Value
	File Filter	If you set <b>Filter Type</b> to <b>Wildcard</b> , 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 (,).	*.csv
	Time Filter	If you select <b>Yes</b> , files are transferred based on their modification time.	Yes
	Minimum Timestamp	If you set <b>Filter Type</b> to <b>Time Filter</b> , and specify a point in time for this parameter, only the files modified after the specified time are transferred. The time format must be <i>yyyy-MM-dd HH:mm:ss</i> .  This parameter can be set to a macro variable of date and time. For example, <b><code>#{timestamp(dateformat(yyyy-MM-dd HH:mm:ss,-90,DAY))}</code></b> indicates that only files generated within the latest 90 days are migrated.	2019-07-01 00:00:00
	Maximum Timestamp	If you set <b>Filter Type</b> to <b>Time Filter</b> , and specify a point in time for this parameter, only the files modified before the specified time are transferred. The time format must be <i>yyyy-MM-dd HH:mm:ss</i> .  This parameter can be set to a macro variable of date and time. For example, <b><code>#{timestamp(dateformat(yyyy-MM-dd HH:mm:ss))}</code></b> indicates that only the files whose modification time is earlier than the current time are migrated.	2019-07-30 00:00:00



Category	Parameter	Description	Example Value
	Create Snapshot	<p>If you set this parameter to <b>Yes</b>, 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.</p> <p>Only the HDFS administrator can create a snapshot. After the CDM job is completed, the snapshot is deleted.</p>	No
	Encryption	<p>This parameter is displayed only when <b>File Format</b> is set to <b>Binary</b>.</p> <p>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:</p> <ul style="list-style-type: none"> <li>• <b>NONE</b>: Export data without decrypting it.</li> <li>• <b>AES-256-GCM</b>: 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.</li> </ul>	AES-256-GCM
	DEK	<p>This parameter is displayed only when <b>Encryption</b> is set to <b>AES-256-GCM</b>. The key consists of 64 hexadecimal numbers and must be the same as the <b>DEK</b> configured during encryption. If the decryption and encryption keys are inconsistent, the system does not report an exception, but the decrypted data is incorrect.</p>	DD0AE00D FECDF8BF0 51BCFDA25 BD4E320DB 0A7AC75A1 F3FC3D3C5 6A457DCD C1B

Category	Parameter	Description	Example Value
	IV	This parameter is displayed only when <b>Encryption</b> is set to <b>AES-256-GCM</b> . The initialization vector consists of 32 hexadecimal numbers and must be the same as the <b>IV</b> configured during encryption. If the initialization vectors are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	5C91687BA 886EDCD12 ACBC3FF19 A3C3F
	MD5 File Extension	This parameter is displayed only when <b>File Format</b> is set to <b>Binary</b> .  This parameter is used to check whether the files extracted by CDM are consistent with source files.	.md5

 **NOTE**

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

### 3.3.6.3.3 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 3-61](#).

 **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 3-61** 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.	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 <b>No</b> .	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 <i>yyyy-MM-dd HH:mm:ss</i> . 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.	2019-01-01 20:00:00
End Time	(Optional) End time (excluding the value) for extracting data. The format is <i>yyyy-MM-dd HH:mm:ss</i> . Only the data generated before the time point is extracted.  This parameter can be set to a macro variable of date and time.	2019-02-01 20:00:00

### 3.3.6.3.4 From Hive

If the source link of a job is the [Link to Hive](#), configure the source job parameters based on [Table 3-62](#).

**Table 3-62** 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	<p>Hive table name. Click the icon next to the text box. The dialog box for selecting the table is displayed.</p> <p>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.</p>	TBL_E
Read Mode	<p>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.</p> <ul style="list-style-type: none"> <li>• 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.</li> <li>• The HDFS mode allows you to use the WHERE condition to filter data or add new fields on the field mapping page.</li> </ul>	HDFS

Parameter	Description	Example Value
Partition Filter Criteria	<p>This parameter is displayed when you select the HDFS read mode and click <b>Show Advanced Attributes</b>.</p> <p>You can configure multiple values (separated by spaces) or a field value range. The time macro function is supported.</p>	<ul style="list-style-type: none"> <li>• Single/ Multi-value filtering: "\$ {dateformat(yyyyMMdd, -1, DAY)} \$ {dateformat(yyyyMMdd)}"</li> <li>• Filter by range: "\${value} &gt;= \$ {dateformat(yyyyMMdd, -7, DAY)} &amp;&amp; \${value} &lt; \$ {dateformat(yyyyMMdd)}"</li> </ul>
WHERE Clause	<p>This parameter is displayed when you select the JDBC read mode and click <b>Show Advanced Attributes</b>.</p> <p>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.</p> <p>You can set a date macro variable to extract data generated on a specific date.</p>	age > 18 and age <= 60

 **NOTE**

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

### 3.3.6.3.5 From DLI

If the source link of a job is the [Link to DLI](#), configure the source job parameters based on [Table 3-63](#).

**Table 3-63** 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
Partition	Partition information. This parameter is available if <b>Clear Data Before Import</b> is set to <b>true</b> .	year=2020,location=sun

### 3.3.6.3.6 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 3-64](#).

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

**Table 3-64** 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.  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.	/ftp/a.csv ftp/b.txt

Category	Parameter	Description	Example Value
	File Format	<p>Format in which CDM parses data. The options are as follows:</p> <ul style="list-style-type: none"> <li>• <b>CSV:</b> Source files will be migrated to tables after being converted to CSV format.</li> <li>• <b>Binary:</b> Files (even not in binary format) will be transferred directly. It is used for file copy.</li> <li>• <b>JSON:</b> Source files will be migrated to tables after being converted to JSON format.</li> </ul>	CSV
	JSON Type	This parameter is displayed only when <b>File Format</b> is set to <b>JSON</b> . Type of a JSON object stored in a JSON file. The options are <b>JSON object</b> and <b>JSON array</b> .	JSON object
	JSON Reference Node	This parameter is used only when <b>File Format</b> is set to <b>JSON</b> and <b>JSON Type</b> is set to <b>JSON Object</b> . 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	Line feed character in a file. By default, the system automatically identifies <b>\n</b> , <b>\r</b> , and <b>\r\n</b> . This parameter is displayed only when <b>File Format</b> is set to <b>CSV</b> .	\n
	Field Delimiter	Character used to separate fields in the file. To set the <b>Tab</b> key as the delimiter, set this parameter to <b>\t</b> . This parameter is displayed only when <b>File Format</b> is set to <b>CSV</b> .	,
	Use Quote Character	If you set this parameter to <b>Yes</b> , the field delimiters in the encircling symbol are regarded as a part of the string value. Currently, the default encircling symbol of CDM is <b>"</b> .	No
	Use RE to Separate Fields	Whether to use regular expressions to separate fields. If you set this parameter to <b>Yes</b> , <b>Field Delimiter</b> becomes invalid. This parameter is displayed only when <b>File Format</b> is set to <b>CSV</b> .	Yes

Category	Parameter	Description	Example Value
	Regular Expression	Regular expression used to separate fields.	^(\\d.*\\d) (\\w*) \\[(.*) \\] ([\\w\\.]* (\\w.*)*.
	Use First Row as Header	This parameter is displayed only when <b>File Format</b> is set to <b>CSV</b> . When you migrate a CSV file to a table, CDM writes all data to the table by default. If you set this parameter to <b>Yes</b> , 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, <b>UTF-8</b> or <b>GBK</b> . You can set the encoding type for text files only. This parameter is invalid when <b>File Format</b> is set to <b>Binary</b> .	UTF-8
	Compression Format	This parameter is displayed only when <b>File Format</b> is set to <b>CSV</b> or <b>JSON</b> . The options are as follows: <ul style="list-style-type: none"> <li>• <b>NONE</b>: Files in all formats can be transferred.</li> <li>• <b>GZIP</b>: Only files in gzip format can be transferred.</li> <li>• <b>ZIP</b>: Only files in Zip format can be transferred.</li> <li>• <b>TAR.GZ</b>: Files in TAR.GZ format are transferred.</li> </ul>	NONE
	Compressed File Suffix	This parameter is displayed when <b>Compression Format</b> is not <b>NONE</b> . 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	<p>Operation performed on source files after the job completes.</p> <ul style="list-style-type: none"> <li>• No action</li> <li>• <b>Rename:</b> After the job completes, the source files are renamed by appending usernames and timestamps as suffixes to the file names.</li> <li>• <b>Delete:</b> After the job completes, the source files are deleted.</li> </ul>	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 <b>Suspension Period</b> .	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	<p>Waiting period for a marker file. If you set <b>Start Job by Marker File</b> to <b>Yes</b> but there is no marker file in the source path, the job fails when the suspension period times out.</p> <p>If you set this parameter to <b>0</b> and there is no marker file in the source path, the job will fail immediately.</p> <p>Unit: second</p>	10
	File Separator	File separator. If you enter multiple file paths in <b>Source Directory/Files</b> , 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 <b>None</b> , <b>Wildcard</b> , and <b>Regex</b> .	None
	Directory Filter	If you set <b>Filter Type</b> to <b>Wildcard</b> , enter a wildcard character to filter paths. The paths that meet the filtering condition are migrated. You can configure multiple paths separated by commas (,).	*input,*out

Category	Parameter	Description	Example Value
	File Filter	If you set <b>Filter Type</b> to <b>Wildcard</b> , 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 (,).	*.csv
	Time Filter	If you select <b>Yes</b> , files are transferred based on their modification time.	Yes
	Minimum Timestamp	If you set <b>Filter Type</b> to <b>Time Filter</b> , and specify a point in time for this parameter, only the files modified after the specified time are transferred. The time format must be <i>yyyy-MM-dd HH:mm:ss</i> .  This parameter can be set to a macro variable of date and time. For example, <b>timestamp(dateformat(yyyy-MM-dd HH:mm:ss,-90,DAY))</b> indicates that only files generated within the latest 90 days are migrated.	2019-07-01 00:00:00
	Maximum Timestamp	If you set <b>Filter Type</b> to <b>Time Filter</b> , and specify a point in time for this parameter, only the files modified before the specified time are transferred. The time format must be <i>yyyy-MM-dd HH:mm:ss</i> .  This parameter can be set to a macro variable of date and time. For example, <b>timestamp(dateformat(yyyy-MM-dd HH:mm:ss))</b> indicates that only the files whose modification time is earlier than the current time are migrated.	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: <ul style="list-style-type: none"> <li><b>NONE</b>: Export data without decrypting it.</li> <li><b>AES-256-GCM</b>: 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.</li> </ul>	AES-256-GCM

Category	Parameter	Description	Example Value
	Disregard Non-existent Path or File	If this is set to <b>Yes</b> , the job can be successfully executed even if the source path does not exist.	No
	DEK	This parameter is displayed only when <b>Encryption</b> is set to <b>AES-256-GCM</b> . The key consists of 64 hexadecimal numbers and must be the same as the <b>DEK</b> 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 FECDF8BF0 51BCFDA25 BD4E320DB 0A7AC75A1 F3FC3D3C5 6A457DCD C1B
	IV	This parameter is displayed only when <b>Encryption</b> is set to <b>AES-256-GCM</b> . The initialization vector consists of 32 hexadecimal numbers and must be the same as the <b>IV</b> configured during encryption. If the initialization vectors are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	5C91687BA 886EDCD12 ACBC3FF19 A3C3F
	MD5 File Extension	This parameter is displayed only when <b>File Format</b> is set to <b>Binary</b> .  This parameter is used to check whether the files extracted by CDM are consistent with source files.	.md5

### 3.3.6.3.7 From HTTP

When the source link of a job is the HTTP link, configure the source job parameters based on [Table 3-65](#). Currently, data can only be exported from the HTTP URLs.

**Table 3-65** 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.	-
Pull List File	If this parameter is set to <b>Yes</b> , 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 <b>Binary</b> 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: <ul style="list-style-type: none"> <li>• <b>NONE</b>: Files in all formats can be transferred.</li> <li>• <b>GZIP</b>: Only files in gzip format can be transferred.</li> <li>• <b>ZIP</b>: Only files in Zip format can be transferred.</li> <li>• <b>TAR.GZ</b>: Files in TAR.GZ format are transferred.</li> </ul>	NONE
Compressed File Suffix	This parameter is displayed when <b>Compression Format</b> is not <b>NONE</b> .  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.	*

Parameter	Description	Example Value
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 <b>Pull List File</b> is set to <b>Yes</b> .	
Query Parameter	<ul style="list-style-type: none"> <li>If you set this parameter to <b>Yes</b>, the name of the objects uploaded to OBS does not include the <b>query</b> parameter.</li> <li>If you set this parameter to <b>No</b>, the name of the objects uploaded to OBS includes the <b>query</b> parameter.</li> </ul>	No
Encryption	<p>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:</p> <ul style="list-style-type: none"> <li><b>NONE</b>: Export data without decrypting it.</li> <li><b>AES-256-GCM</b>: 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.</li> </ul>	AES-256-GCM
Disregard Non-existent Path or File	If this is set to <b>Yes</b> , the job can be successfully executed even if the source path does not exist.	No
DEK	This parameter is displayed only when <b>Encryption</b> is set to <b>AES-256-GCM</b> . The key consists of 64 hexadecimal numbers and must be the same as the <b>DEK</b> configured during encryption. If the encryption and decryption keys are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	DD0AE00DFEC D78BF051BCF DA25BD4E320 DB0A7AC75A1 F3FC3D3C56A 457DCDC1B
IV	This parameter is displayed only when <b>Encryption</b> is set to <b>AES-256-GCM</b> . The initialization vector consists of 32 hexadecimal numbers and must be the same as the <b>IV</b> configured during encryption. If the encryption and decryption keys are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	5C91687BA886 EDCD12ACBC3 FF19A3C3F
MD5 File Extension	This parameter is used to check whether the files extracted by CDM are consistent with source files.	.md5

### 3.3.6.3.8 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 3-66](#).

**Table 3-66** 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	<p>When <b>Use SQL Statement</b> is set to <b>Yes</b>, enter an SQL statement here. CDM exports data based on the SQL statement.</p> <p><b>NOTE</b></p> <ul style="list-style-type: none"> <li>SQL statements can only be used to query data. Join and nesting are supported, but multiple query statements are not allowed, for example, <b>select * from table a; select * from table b.</b></li> <li>With statements are not supported.</li> <li>Comments, such as -- and /*, are not supported.</li> <li>Addition, deletion, and modification operations are not supported, including but not limited to the following: <ul style="list-style-type: none"> <li>load data</li> <li>delete from</li> <li>alter table</li> <li>create table</li> <li>drop table</li> <li>into outfile</li> </ul> </li> </ul>	select id,name from sqoop.user;

Category	Parameter	Description	Example Value
	Schema/ Tablespace	<p>Name of the schema or tablespace from which data will be extracted. This parameter is displayed when <b>Use SQL Statement</b> is set to <b>No</b>. Click the icon next to the text box to go to the page for selecting a schema or directly enter a schema or tablespace.</p> <p>If the desired schema or tablespace is not displayed, check whether the login account has the permissions required to query metadata.</p> <p><b>NOTE</b> 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:</p> <ul style="list-style-type: none"> <li>● <b>SCHEMA*</b> indicates that all databases whose names starting with <b>SCHEMA</b> are exported.</li> <li>● <b>*SCHEMA</b> indicates that all databases whose names ending with <b>SCHEMA</b> are exported.</li> <li>● <b>*SCHEMA*</b> indicates that all databases whose names containing <b>SCHEMA</b> are exported.</li> </ul>	SCHEMA_E

Category	Parameter	Description	Example Value
	Table Name	<p>Name of the table from which data will be extracted. This parameter is displayed when <b>Use SQL Statement</b> is set to <b>No</b>. Click the icon next to the text box to go to the page for selecting the table or directly enter a table name.</p> <p>If the desired table is not displayed, confirm that the table exists or that the login account has the permissions required to query metadata.</p> <p>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.</p> <p><b>NOTE</b> 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:</p> <ul style="list-style-type: none"> <li>• <b>table*</b> indicates that all tables whose names starting with <b>table</b> are exported.</li> <li>• <b>*table</b> indicates that all tables whose names ending with <b>table</b> are exported.</li> <li>• <b>*table*</b> indicates that all tables whose names containing <b>table</b> are exported.</li> </ul>	table



Category	Parameter	Description	Example Value
Advanced attributes	Partition Column	<p>This parameter is displayed when <b>Use SQL Statement</b> is set to <b>No</b>, 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.</p> <p>Click the icon next to the text box to go to the page for selecting a field or directly enter a field.</p> <p><b>NOTE</b></p> <ul style="list-style-type: none"> <li>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.</li> <li>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.</li> </ul>	id
	Where Clause	<p>WHERE clause used to specify the data extraction range. This parameter is displayed when <b>Use SQL Statement</b> is set to <b>No</b>. If this parameter is not set, the entire table is extracted.</p> <p>You can set a date macro variable to extract data generated on a specific date.</p>	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.	-
	Minimum value of a split field	Specifies the minimum value of <b>Job Split Field</b> during data extraction.	-
	Maximum Split Field Value	Specifies the maximum value of <b>Job Split Field</b> during data extraction.	-

Category	Parameter	Description	Example Value
	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 <b>Job Split Field</b> .	-
	Extract by Partition	<p>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 <b>Table Partition</b> to specify specific MySQL table partitions from which data is extracted.</p> <ul style="list-style-type: none"> <li>• This function does not support non-partitioned tables.</li> <li>• This parameter is available only for RDS for PostgreSQL and RDS for MySQL.</li> <li>• The database user must have the <b>SELECT</b> permission on the system views <b>dba_tab_partitions</b> and <b>dba_tab_subpartitions</b>.</li> </ul>	No

### 3.3.6.3.9 From MySQL

If the source link of a job is the [Link to a MySQL Database](#), configure the source job parameters based on [Table 3-67](#).

**Table 3-67** 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	<p>When <b>Use SQL Statement</b> is set to <b>Yes</b>, enter an SQL statement here. CDM exports data based on the SQL statement.</p> <p><b>NOTE</b></p> <ul style="list-style-type: none"> <li>• SQL statements can only be used to query data. Join and nesting are supported, but multiple query statements are not allowed, for example, <b>select * from table a; select * from table b.</b></li> <li>• With statements are not supported.</li> <li>• Comments, such as -- and /*, are not supported.</li> <li>• Addition, deletion, and modification operations are not supported, including but not limited to the following: <ul style="list-style-type: none"> <li>• load data</li> <li>• delete from</li> <li>• alter table</li> <li>• create table</li> <li>• drop table</li> <li>• into outfile</li> </ul> </li> </ul>	select id,name from sqoop.user;
	Schema/ Tablespace	<p>Name of the schema or tablespace from which data will be extracted. This parameter is displayed when <b>Use SQL Statement</b> is set to <b>No</b>. Click the icon next to the text box to go to the page for selecting a schema or directly enter a schema or tablespace.</p> <p>If the desired schema or tablespace is not displayed, check whether the login account has the permissions required to query metadata.</p> <p><b>NOTE</b> This parameter can be set to a regular expression to export all databases that meet the rule.</p>	SCHEMA_E

Category	Parameter	Description	Example Value
	Table Name	<p>Name of the table from which data will be extracted. This parameter is displayed when <b>Use SQL Statement</b> is set to <b>No</b>. Click the icon next to the text box to go to the page for selecting the table or directly enter a table name.</p> <p>If the desired table is not displayed, confirm that the table exists or that the login account has the permissions required to query metadata.</p> <p>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.</p> <p><b>NOTE</b> This parameter can be set to a regular expression to export all databases that meet the rule.</p>	table
Advanced attributes	Partition Column	<p>This parameter is displayed when <b>Use SQL Statement</b> is set to <b>No</b>, 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.</p> <p>Click the icon next to the text box to go to the page for selecting a field or directly enter a field.</p> <p><b>NOTE</b></p> <ul style="list-style-type: none"> <li>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.</li> <li>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.</li> </ul>	id

Category	Parameter	Description	Example Value
	WHERE Clause	WHERE clause used to specify the data extraction range. This parameter is displayed when <b>Use SQL Statement</b> is set to <b>No</b> . 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.	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.	-
	Minimum value of a split field	Specifies the minimum value of <b>Job Split Field</b> during data extraction.	-
	Maximum Split Field Value	Specifies the maximum value of <b>Job Split Field</b> 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 <b>Job Split Field</b> .	-
	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 <b>Table Partition</b> to specify specific MySQL table partitions from which data is extracted.  <ul style="list-style-type: none"> <li>• This function does not support non-partitioned tables.</li> <li>• The database user must have the <b>SELECT</b> permission on the system views <b>dba_tab_partitions</b> and <b>dba_tab_subpartitions</b>.</li> </ul>	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
```

### 3.3.6.3.10 From Oracle

If the source link of a job is the [Link to an Oracle Database](#), configure the source job parameters based on [Table 3-68](#).

**Table 3-68** 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	<p>When <b>Use SQL Statement</b> is set to <b>Yes</b>, enter an SQL statement here. CDM exports data based on the SQL statement.</p> <p><b>NOTE</b></p> <ul style="list-style-type: none"> <li>• SQL statements can only be used to query data. Join and nesting are supported, but multiple query statements are not allowed, for example, <b>select * from table a; select * from table b.</b></li> <li>• With statements are not supported.</li> <li>• Comments, such as -- and /*, are not supported.</li> <li>• Addition, deletion, and modification operations are not supported, including but not limited to the following: <ul style="list-style-type: none"> <li>• load data</li> <li>• delete from</li> <li>• alter table</li> <li>• create table</li> <li>• drop table</li> <li>• into outfile</li> </ul> </li> </ul>	select id,name from sqoop.user;
	Schema/Tablespace	<p>Name of the schema or tablespace from which data will be extracted. This parameter is displayed when <b>Use SQL Statement</b> is set to <b>No</b>. Click the icon next to the text box to go to the page for selecting a schema or directly enter a schema or tablespace.</p> <p>If the desired schema or tablespace is not displayed, check whether the login account has the permissions required to query metadata.</p> <p><b>NOTE</b></p> <p>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:</p> <ul style="list-style-type: none"> <li>• <b>SCHEMA*</b> indicates that all databases whose names starting with <b>SCHEMA</b> are exported.</li> <li>• <b>*SCHEMA</b> indicates that all databases whose names ending with <b>SCHEMA</b> are exported.</li> <li>• <b>*SCHEMA*</b> indicates that all databases whose names containing <b>SCHEMA</b> are exported.</li> </ul>	SCHEMA_E

Category	Parameter	Description	Example Value
	Table Name	<p>Name of the table from which data will be extracted. This parameter is displayed when <b>Use SQL Statement</b> is set to <b>No</b>. Click the icon next to the text box to go to the page for selecting the table or directly enter a table name.</p> <p>If the desired table is not displayed, confirm that the table exists or that the login account has the permissions required to query metadata.</p> <p>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.</p> <p><b>NOTE</b> 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:</p> <ul style="list-style-type: none"> <li>• <b>table*</b> indicates that all tables whose names starting with <b>table</b> are exported.</li> <li>• <b>*table</b> indicates that all tables whose names ending with <b>table</b> are exported.</li> <li>• <b>*table*</b> indicates that all tables whose names containing <b>table</b> are exported.</li> </ul>	table



Category	Parameter	Description	Example Value
Advanced attributes	Partition Column	<p>This parameter is displayed when <b>Use SQL Statement</b> is set to <b>No</b>, 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.</p> <p>Click the icon next to the text box to go to the page for selecting a field or directly enter a field.</p> <p><b>NOTE</b></p> <ul style="list-style-type: none"> <li>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.</li> <li>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.</li> </ul>	id
	WHERE Clause	<p>WHERE clause used to specify the data extraction range. This parameter is displayed when <b>Use SQL Statement</b> is set to <b>No</b>. If this parameter is not set, the entire table is extracted.</p> <p>You can set a date macro variable to extract data generated on a specific date.</p>	DS='\$ {dateformat(yyyy-MM-dd,-1,DAY)}'
	Null in Partition Column	Whether the partition column can contain null values	Yes
	Extract by Partition	<p>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 <b>Table Partition</b> to specify specific Oracle table partitions from which data is extracted.</p> <ul style="list-style-type: none"> <li>This function does not support non-partitioned tables.</li> <li>The database user must have the <b>SELECT</b> permission on the system views <b>dba_tab_partitions</b> and <b>dba_tab_subpartitions</b>.</li> </ul>	No

Category	Parameter	Description	Example Value
	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 <i>Partition.Subpartition</i> format, for example, <b>P2.SUBP1</b> .	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 <b>Job Split Field</b> during data extraction.	-
	Maximum Split Field Value	Specifies the maximum value of <b>Job Split Field</b> 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 <b>Job Split Field</b> .	-

 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.

### 3.3.6.3.11 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 3-69](#).

**Table 3-69** Parameter description

Category	Parameter	Description	Example Value
Basic parameters	Schema/ Tablespace	<p>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.</p> <p>If the desired schema or tablespace is not displayed, check whether the login account has the permissions required to query metadata.</p> <p><b>NOTE</b> This parameter can be set to a regular expression to export all databases that meet the rule.</p>	SCHEMA_E
	Table Name	<p>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.</p> <p>If the desired table is not displayed, confirm that the table exists or that the login account has the permissions required to query metadata.</p> <p>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.</p> <p><b>NOTE</b> This parameter can be set to a regular expression to export all databases that meet the rule.</p>	table
Advanced attributes	WHERE Clause	<p>Specifies the data extraction range. If this parameter is not set, the entire table is extracted.</p> <p>You can set a date macro variable to extract data generated on a specific date.</p>	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)}``

### 3.3.6.3.12 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 3-70](#).

**Table 3-70** 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	<p>Conditions for filtering documents. CDM migrates only the data that meets the filter conditions. The examples are as follows:</p> <ol style="list-style-type: none"> <li>1. Filter by expression: <code>{'last_name': 'Smith'}</code> indicates that all files whose <code>last_name</code> value is <b>Smith</b> are queried.</li> <li>2. Filter by parameter: <code>{ x : "john" }, { z : 1 }</code> indicates that all <code>z</code> fields whose <code>x</code> is <b>john</b> are queried.</li> <li>3. Filter by condition: <code>{ "field" : { \$gt: 5 } }</code> indicates that the <b>field</b> values greater than 5 are queried.</li> <li>4. Filter by time macro: <code>{"ts":{\$gte:ISODate("\${dateformat(yyyy-MM-dd'T'HH:mm:ss.SSS'Z",-1,HOUR)}")}}</code> indicates that the values greater than those after time macro conversion in the <b>ts</b> field are queried.</li> </ol>	<code>{'last_name': 'Smith'}</code>

### 3.3.6.3.13 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 3-71](#).

**Table 3-71** 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	<p>The options are as follows:</p> <ul style="list-style-type: none"> <li>• <b>String</b>: without column name, such as <code>value1,value2</code></li> <li>• <b>Hash</b>: with column name, such as <code>column1=value1,column2=value2</code></li> </ul>	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 <b>Value Storage Type</b> is set to <b>Hash</b> . The hash key contains the same field.	Yes

### 3.3.6.3.14 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 3-72](#).

**Table 3-72** Parameter description

Parameter	Description	Example Value
Topics	One or more topics can be entered.	est1,est2
Offset	Initial offset parameter <ul style="list-style-type: none"> <li>• <b>Latest:</b> Maximum offset, indicating that the latest data will be extracted.</li> <li>• <b>Earliest:</b> Minimum offset, indicating that the earliest data will be extracted.</li> <li>• <b>Submitted:</b> data that has been submitted</li> <li>• <b>Time Range:</b> data within a specified time range</li> </ul>	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	<p>Format used for parsing data. The options are as follows:</p> <ul style="list-style-type: none"> <li>• <b>Binary</b>: Data is transferred directly. It is not converted to another format. This setting is suitable for file migration.</li> <li>• <b>CSV</b>: Source data will be migrated after being converted in CSV format.</li> <li>• <b>JSON</b>: Source data will be migrated after being converted in JSON format.</li> <li>• <b>CDC (DRS_JSON)</b>: Source data will be migrated after being converted in DRS_JSON format.</li> </ul>	Binary
Field Delimiter	The default value is space. To set the <b>Tab</b> key as the delimiter, set this parameter to <code>\t</code> .	,
Max. Poll Records	(Optional) Maximum number of records per poll	100
Max. Poll Interval	(Optional) Maximum interval between polls (seconds)	100

### 3.3.6.3.15 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 3-73](#).

**Table 3-73** 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.	type
Split Nested Field	(Optional) Whether to split the JSON content of the nested fields. For example, <code>a:{ b:{ c:1, d:{ e:2, f:3 } } }</code> can be split into <code>a.b.c</code> , <code>a.b.d.e</code> , and <code>a.b.d.f</code> .	No

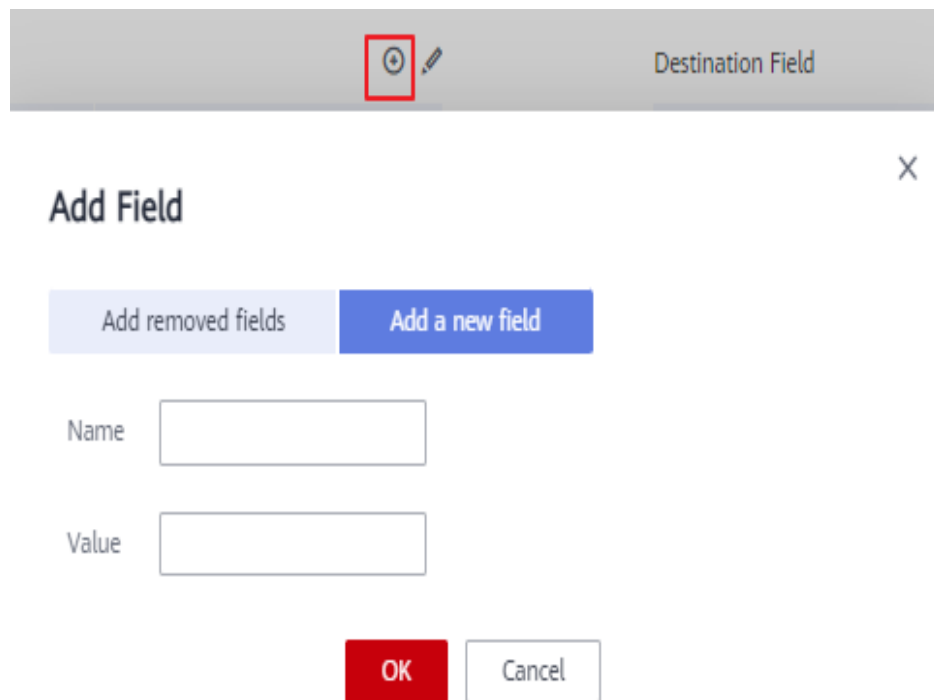
Parameter	Description	Example Value
Filter Conditions	<p>(Optional) CDM migrates only the data that meets the filter conditions.</p> <ul style="list-style-type: none"> <li>• Currently, only the query string (q syntax) of Elasticsearch can be used to filter source data. The q syntax is used in the following way: <ul style="list-style-type: none"> <li>- In exact match, the <b>column.data</b> format is used to match and filter data. <b>column</b> indicates the field name, and <b>data</b> indicates the query condition, for example, <b>last_name:Smith</b>. In addition, if <b>data</b> is a string containing spaces, it must be enclosed in double quotation marks. If <b>column</b> is not specified, all fields will be matched by <b>data</b>.</li> <li>- Multiple query conditions can be combined with connection words. The format is <b>column1.data1 AND column2.data2</b>. The connection words can be <b>AND</b>, <b>OR</b>, or <b>NOT</b>. They must be in uppercase, and there must be a space before and after each connection word. Example: <b>last_name:Smith AND last_name:John</b></li> <li>- In range matching, you can directly use a condition expression to filter data. The expression is in <b>column:&gt;data</b> format. The operator can be <b>&gt;</b>, <b>&gt;=</b>, <b>&lt;</b>, or <b>&lt;=</b>. An example is <b>time:&gt;=1636905600000 AND time:&lt;1637078400000</b>. It can also be used together with a macro variable of date and time, for example, <b>createTime:&gt;=\$ {timestamp(dateformat(yyyyMMdd,-1,D AY))} AND createTime:&lt; \$ {timestamp(dateformat(yyyyMMdd))}</b>.</li> <li>- In range matching, you can also use the range syntax to filter data. The format is <b>column:{data1 TO data2}</b>. <b>{ and }</b> indicate that a value is not included. <b>[ and ]</b> indicate that a value is included. <b>TO</b> must be capitalized, and there must be a space before and after it. <b>*</b> indicates all data. For example, <b>time:{1636992000000 TO *}</b> filters out all the data greater than 1636992000000 in the <b>time</b> field. It can also be used together with a macro variable of date and time, for example, <b>createTime:[\$</b></li> </ul> </li> </ul>	last_name:Smith



Parameter	Description	Example Value
	<pre>{timestamp(dateformat(yyyyMMdd,-1,DAY))} TO \$ {timestamp(dateformat(yyyyMMdd))}</pre> <ul style="list-style-type: none"> <li>Source data cannot be filtered using the query domain-specific language (DSL) of Elasticsearch.</li> </ul>	
Extract Meta-field	Whether to extract index meta-fields. For example, <code>_index</code> , <code>_type</code> , <code>_id</code> , and <code>_score</code> .	Yes

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

**Figure 3-57** Setting custom fields



### 3.3.6.3.16 From OpenTSDB

If the source link of a job is the [Link to CloudTable OpenTSDB](#), configure the source job parameters based on [Table 3-74](#).

**Table 3-74** Parameter description

Parameter	Description	Example Value
Start Time	Start time of the query. The value is a character string or timestamp in the format of <i>yyyyMMddHHmmdd</i> .	20180920145505
End Time	(Optional) End time of the query. The value is a string or timestamp in the format of <i>yyyyMMddHHmmdd</i> .	20180921145505
Metric	Metric of the data to be migrated. You can specify a metric or select an existing metric in OpenTSDB.	city.temp
Aggregate Function	Aggregate function	sum
Tag	(Optional) If you specify a tag, only the tagged data will be migrated.	tagk1:tagv1,tagk2:tagv2

### 3.3.6.4 Destination Job Parameters

#### 3.3.6.4.1 To OBS


If the destination link of a job is the [Link to OBS](#), configure the destination job parameters based on [Table 3-75](#).

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

**Table 3-75** 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.	directory/

Category	Parameter	Description	Example Value
	File Format	<p>Format in which data is written. The options are as follows:</p> <ul style="list-style-type: none"> <li>• <b>CSV:</b> Data is written in CSV format, which is used for migrating data tables to files.</li> <li>• <b>Binary:</b> Files will be transferred directly. CDM writes the files without changing their format. This setting is suitable for file migration.</li> </ul> <p>If data is migrated between file-related data sources, such as FTP, SFTP, HDFS, and OBS, the value of <b>File Format</b> must be the same as the source file format.</p>	CSV
	Duplicate File Processing Method	<p>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:</p> <ul style="list-style-type: none"> <li>• Replace</li> <li>• Skip</li> <li>• Stop job</li> </ul>	Skip
Advanced attributes	Encryption	<p>Whether to encrypt the uploaded data and the encryption mode. The options are as follows:</p> <ul style="list-style-type: none"> <li>• <b>None:</b> Data is written without encryption.</li> <li>• <b>KMS:</b> KMS in Data Encryption Workshop (DEW) is used for encryption. If KMS encryption is enabled, MD5 verification for data cannot be performed.</li> <li>• <b>AES-256-GCM:</b> 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.</li> </ul>	KMS

Category	Parameter	Description	Example Value
	Key ID	<p>Data encryption key. This parameter is displayed when <b>Encryption</b> is set to <b>KMS</b>. Click  next to the text box to select the KMS key that was created in DEW.</p> <ul style="list-style-type: none"> <li>• If the KMS key of the same project as that of the CDM cluster is used, you do not need to modify <b>Project ID</b>.</li> <li>• If the KMS key of another project is used, you need to modify <b>Project ID</b>.</li> </ul>	53440ccb-3e73-4700-98b5-71ff5476e621
	Project ID	<p>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.</p> <ul style="list-style-type: none"> <li>• If KMS and the CDM cluster are in the same project, retain the default value of <b>Project ID</b>.</li> <li>• If KMS of another project is used, set this parameter to the ID of the project to which KMS belongs.</li> </ul>	9bd7c4bd54e5417198f9591bef07ae67
	DEK	<p>This parameter is displayed only when <b>Encryption</b> is set to <b>AES-256-GCM</b>. The key consists of 64 hexadecimal numbers.</p> <p>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.</p>	DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B
	IV	<p>This parameter is displayed only when <b>Encryption</b> is set to <b>AES-256-GCM</b>. The initialization vector consists of 32 hexadecimal numbers.</p> <p>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.</p>	5C91687BA886EDCD12ACBC3FF19A3C3F

Category	Parameter	Description	Example Value
	Copy Content-Type	<p>This parameter is displayed only when <b>File Format</b> is <b>Binary</b>, and both the migration source and destination are object storage.</p> <p>If you set this parameter to <b>Yes</b>, the Content-Type attribute of the source file is copied during object file migration. This function is mainly used for static website migration.</p> <p>The Content-Type attribute cannot be written to Archive buckets. Therefore, if you set this parameter to <b>Yes</b>, the migration destination must be a non-Archive bucket.</p>	No
	Line Separator	<p>Line feed character in a file. By default, the system automatically identifies <b>\n</b>, <b>\r</b>, and <b>\r\n</b>. This parameter is not used when <b>File Format</b> is set to <b>Binary</b>.</p>	\n
	Field Delimiter	<p>Field delimiter in the file. This parameter is not used when <b>File Format</b> is set to <b>Binary</b>.</p>	,
	File Size	<p>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.</p>	1024
	Validate MD5 Value	<p>The MD5 value can be verified only when files are transferred in <b>Binary</b> format. KMS encryption cannot be used if the MD5 value needs to be verified.</p> <p>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.</p>	Yes
	Record MD5 Verification Result	<p>Whether to record the MD5 verification result when <b>Validate MD5 Value</b> is set to <b>Yes</b></p>	Yes
	Record MD5 Link	<p>OBS link to which the MD5 verification result will be written</p>	obslink

Category	Parameter	Description	Example Value
	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, <b>UTF-8</b> or <b>GBK</b> . This parameter is not used when <b>File Format</b> is set to <b>Binary</b> .	GBK
	Use Quote Character	This parameter is displayed only when <b>File Format</b> is <b>CSV</b> . It is used when database tables are migrated to file systems.  If you set this parameter to <b>Yes</b> 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 <b>hello,world</b> 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 <b>File Format</b> is set to <b>CSV</b> .  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 <b>Yes</b> , 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 <b>Yes</b> , 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

Category	Parameter	Description	Example Value
	Hierarchical Directory	Custom storage directory for files after migration. The time macro variable is supported.	<code>\${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}</code>
	Customize File Name	<p>This parameter is displayed only when data is exported from a relational database to OBS and <b>File Format</b> is set to <b>CSV</b>.</p> <p>This parameter specifies the name of the file generated by OBS. The options are as follows:</p> <ul style="list-style-type: none"> <li>• <b>Character string:</b> Special characters are allowed. For example, if this parameter is set to <b>cdm#</b>, the name of the generated file is <b>cdm#.csv</b>.</li> <li>• <b>Macro variable of time:</b> If this parameter is set to <b>\${timestamp()}</b>, the name of the generated file is <b>1554108737.csv</b>.</li> <li>• <b>Macro variable of table name:</b> If this parameter is set to <b>\${tableName}</b>, the name of the generated file is <b>sqltabname.csv</b>.</li> <li>• <b>Macro variable of version number:</b> If this parameter is set to <b>\${version}</b>, the name of the generated file is <b>v1.csv</b>.</li> <li>• <b>Any combination of the character string and macro variable (macro variable of time, table name, or version number).</b> For example, if this parameter is set to <b>cdm#\${timestamp()}_\${version}</b>, the name of the generated file is <b>cdm#1554108737_v1.csv</b>.</li> </ul>	cdm

### 3.3.6.4.2 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 3-76](#).

**Table 3-76** 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.	/user/output
File Format	Format in which data is written. The options are as follows: <ul style="list-style-type: none"> <li>• <b>CSV</b>: Data is written in CSV format, which is used for migrating data tables to files.</li> <li>• <b>Binary</b>: Files will be transferred directly. CDM writes the files without changing their format. This setting is suitable for file migration.</li> </ul> If data is migrated between file-related data sources, such as FTP, SFTP, HDFS, and OBS, the value of <b>File Format</b> must be 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: <ul style="list-style-type: none"> <li>• Replace</li> <li>• Skip</li> <li>• Stop job</li> </ul>	Stop job
Compression Format	File compression format after data writing. The following compression formats are supported: <ul style="list-style-type: none"> <li>• <b>None</b>: The files are not compressed.</li> <li>• <b>DEFLATE</b>: The files are compressed in DEFLATE format.</li> <li>• <b>gzip</b>: The files are compressed in gzip format.</li> <li>• <b>bzip2</b>: The files are compressed in bzip2 format.</li> <li>• <b>LZ4</b>: The files are compressed in LZ4 format.</li> <li>• <b>Snappy</b>: The files are compressed in snappy format.</li> </ul>	Snappy
Line Separator	Line feed character in a file. By default, the system automatically identifies <b>\n</b> , <b>\r</b> , and <b>\r\n</b> . This parameter is not used when <b>File Format</b> is set to <b>Binary</b> .	\n



Parameter	Description	Example Value
Field Delimiter	Field delimiter in the file. This parameter is not used when <b>File Format</b> is set to <b>Binary</b> .	,
Use Quote Character	This parameter is displayed only when <b>File Format</b> is <b>CSV</b> . It is used when database tables are migrated to file systems. If you set this parameter to <b>Yes</b> 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 <b>hello,world</b> 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 <b>Yes</b> , CDM writes the heading line of the table to the file.	No
Write to Temporary File	Whether to write the binary file to a <b>.tmp</b> file first. After the migration is successful, run the <b>rename</b> or <b>move</b> 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
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)}	-

Parameter	Description	Example Value
Encryption	<p>This parameter is displayed only when <b>File Format</b> is set to <b>Binary</b>.</p> <p>Whether to encrypt the uploaded data. The options are as follows:</p> <ul style="list-style-type: none"> <li>• <b>None</b>: Data is written without encryption.</li> <li>• <b>AES-256-GCM</b>: 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.</li> </ul>	AES-256-GCM
DEK	<p>This parameter is displayed only when <b>Encryption</b> is set to <b>AES-256-GCM</b>. The key consists of 64 hexadecimal numbers.</p> <p>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.</p>	DD0AE00DFE CD78BF051BC FDA25BD4E3 20DB0A7AC7 5A1F3FC3D3C 56A457DCDC 1B
IV	<p>This parameter is displayed only when <b>Encryption</b> is set to <b>AES-256-GCM</b>. The initialization vector consists of 32 hexadecimal numbers.</p> <p>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.</p>	5C91687BA88 6EDCD12ACB C3FF19A3C3F

 NOTE

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

### 3.3.6.4.3 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 3-77](#).

**Table 3-77** Parameter description

Parameter	Description	Example Value
Table Name	<p>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.</p> <p>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.</p>	TBL_2
Clear Data Before Import	<p>Whether the data in the destination table is cleared before data import. The options are as follows:</p> <ul style="list-style-type: none"> <li>• <b>Yes</b>: The data is cleared.</li> <li>• <b>No</b>: The data is not cleared. Instead, it will be added to the existing table.</li> </ul>	Yes
Rowkey Delimiter	<p>(Optional) Used to combine multiple columns as a rowkey. Spaces are used by default.</p>	,
Rowkey Data Redundancy	<p>(Optional) Whether to write the rowkey data into HBase columns. The default value is <b>No</b>.</p>	No
Compression Format	<p>(Optional) Compression format used in creating an HBase table. The default value is <b>None</b>.</p> <ul style="list-style-type: none"> <li>• <b>None</b>: The files are not compressed.</li> <li>• <b>Snappy</b>: The files are compressed in snappy format.</li> <li>• <b>gzip</b>: The files are compressed in gzip format.</li> </ul>	None
Write WAL	<p>Whether to enable Write Ahead Log (WAL) of HBase. The options are as follows:</p> <ul style="list-style-type: none"> <li>• <b>Yes</b>: If the HBase server breaks down after the function is enabled, you can replay the operations that have not been performed in WAL.</li> <li>• <b>No</b>: If you set this parameter to <b>No</b>, the write performance is improved. However, if the HBase server breaks down, data may be lost.</li> </ul>	No

Parameter	Description	Example Value
Match Data Type	<ul style="list-style-type: none"> <li>• <b>Yes:</b> 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 even.</li> <li>• <b>No:</b> All types of data in the source database are written into HBase as character strings.</li> </ul>	No

### 3.3.6.4.4 To Hive

If the destination link of a job is the [Link to Hive](#), configure the destination job parameters based on [Table 3-78](#).

**Table 3-78** 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
Auto Table Creation	<p>This parameter is displayed only when the source is a relational database. The options are as follows:</p> <ul style="list-style-type: none"> <li>• <b>Non-auto creation:</b> CDM will not automatically create a table.</li> <li>• <b>Auto creation:</b> If the destination database does not contain the table specified by <b>Table Name</b>, CDM will automatically create the table. If the table specified by <b>Table Name</b> already exists, no table is created and data is written to the existing table.</li> <li>• <b>Deletion before creation:</b> CDM deletes the table specified by <b>Table Name</b>, and then creates the table again.</li> </ul>	Non-auto creation

Parameter	Description	Example Value
Table Name	<p>Destination table name.</p> <p>Click the icon next to the text box. The dialog box for selecting the table is displayed.</p> <p>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.</p>	TBL_X
Clear Data Before Import	<p>Whether the data in the destination table is cleared before data import. The options are as follows:</p> <ul style="list-style-type: none"> <li>• <b>Yes:</b> The data is cleared.</li> <li>• <b>No:</b> The data is not cleared. Instead, it will be added to the existing table.</li> </ul>	Yes
Partition to Clear	<p>This parameter is available when <b>Clear Data Before Import</b> is set to <b>Yes</b>.</p> <p>When you enter the information about the partitions to be cleared, the data in the partitions will be cleared.</p>	<p>Single partition: <b>year=2020,location=sun</b></p> <p>Multiple partitions: <b>['year=2020,location=sun'</b> <b>,'year=2021,location=earth']</b></p>

 **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;
```

### 3.3.6.4.5 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 3-79](#).

**Table 3-79** Parameter description

Category	Parameter	Description	Example Value
Basic parameters	Schema/Tables space	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

Category	Parameter	Description	Example Value
	Auto Table Creation	<p>This parameter is displayed only when the source is a relational database. The options are as follows:</p> <ul style="list-style-type: none"> <li>• <b>Non-auto creation:</b> CDM will not automatically create a table.</li> <li>• <b>Auto creation:</b> If the destination database does not contain the table specified by <b>Table Name</b>, CDM will automatically create the table. If the table specified by <b>Table Name</b> already exists, no table is created and data is written to the existing table.</li> <li>• <b>Deletion before creation:</b> CDM deletes the table specified by <b>Table Name</b>, and then creates the table again.</li> </ul>	Non-auto creation
	Table Name	<p>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.</p> <p>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.</p>	table
	Clear Data Before Import	<p>Whether to clear the data in the destination table before data import. The options are as follows:</p> <ul style="list-style-type: none"> <li>• <b>Do not clear:</b> The data in the destination table is not cleared before data import. The imported data is just added to the table.</li> <li>• <b>Clear all data:</b> All data is cleared from the destination table before data import.</li> <li>• <b>Clear part of data:</b> Part of the data in the destination table is cleared before data import. If you select <b>Clear part of data</b>, you must configure <b>WHERE Clause</b> to specify which part will be deleted.</li> </ul>	Clear part of data
	WHERE Clause	<p>If <b>Clear Data Before Import</b> is set to <b>Clear part of data</b>, data in the destination table will be deleted based on the WHERE clause after the configuration is complete and before the import starts.</p>	age > 18 and age <= 60

Category	Parameter	Description	Example Value
	Constraint Conflict Handling	<p>Mode for handling conflicts in data migration</p> <ul style="list-style-type: none"> <li>• <b>insert into</b>: When a primary key or unique index conflict occurs, data cannot be written and will become dirty data.</li> <li>• <b>replace into</b>: 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.</li> <li>• <b>on duplicate key update</b>: 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.</li> </ul>	insert into
	Loader Threads	<p>Number of threads started in each loader. A larger number allows more concurrent write operations.</p> <p><b>NOTE</b> This parameter is unavailable if <b>Constraint Conflict Handling</b> is set to <b>replace into</b> or <b>on duplicate key update</b>.</p>	1
Advanced parameters	Import to Staging Tables	<p>If you set this parameter to <b>Yes</b>, 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.</p> <p>The default value is <b>No</b>, 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.</p> <p><b>NOTE</b> If you select <b>Clear part of data</b> or <b>Clear all data</b> for <b>Clear Data Before Import</b>, CDM does not roll back the deleted data in transaction mode.</p>	No



Category	Parameter	Description	Example Value
	Extend Field Length	When <b>Auto creation</b> 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.  <b>NOTE</b> 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

### 3.3.6.4.6 To DWS

If the destination link of a job is a [DWS link](#), configure the destination job parameters based on [Table 3-80](#).

**Table 3-80** Parameter description

Category	Parameter	Description	Example Value
Basic parameters	Schema/Tables space	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

Category	Parameter	Description	Example Value
	Auto Table Creation	<p>This parameter is displayed only when the source is a relational database. The options are as follows:</p> <ul style="list-style-type: none"> <li>• <b>Non-auto creation:</b> CDM will not automatically create a table.</li> <li>• <b>Auto creation:</b> If the destination database does not contain the table specified by <b>Table Name</b>, CDM will automatically create the table. If the table specified by <b>Table Name</b> already exists, no table is created and data is written to the existing table.</li> <li>• <b>Deletion before creation:</b> CDM deletes the table specified by <b>Table Name</b>, and then creates the table again.</li> </ul> <p><a href="#">Field Mapping in Automatic Table Creation on DWS</a> describes the field mapping between the DWS tables created by CDM and source tables.</p>	Non-auto creation
	Table Name	<p>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.</p> <p>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.</p>	table
	Compress Data	Whether to compress data when data is imported to DWS and <b>Auto creation</b> is selected	No
	Storage Mode	<p>When data is imported to DWS and <b>Auto Creation</b> is selected, you can specify the data storage mode:</p> <ul style="list-style-type: none"> <li>• <b>Row-based:</b> 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.</li> <li>• <b>Column-based:</b> 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).</li> </ul>	Row-based

Category	Parameter	Description	Example Value
	Import Mode	Mode for importing data to DWS <ul style="list-style-type: none"> <li>In COPY mode, the source data is copied to the DataNode of DWS after passing through the management node.</li> <li>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.</li> </ul>	COPY
	Clear Data Before Import	Whether to clear the data in the destination table before data import. The options are as follows: <ul style="list-style-type: none"> <li><b>Do not clear:</b> The data in the destination table is not cleared before data import. The imported data is just added to the table.</li> <li><b>Clear all data:</b> All data is cleared from the destination table before data import.</li> <li><b>Clear part of data:</b> Part of the data in the destination table is cleared before data import. If you select <b>Clear part of data</b>, you must configure <b>WHERE Clause</b> to specify which part will be deleted.</li> </ul>	Clear part of data
	WHERE Clause	If <b>Clear Data Before Import</b> is set to <b>Clear part of data</b> , 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 <ul style="list-style-type: none"> <li><b>insert into:</b> When a primary key or unique index conflict occurs, data cannot be written and will become dirty data.</li> <li><b>replace into:</b> 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.</li> <li><b>on duplicate key update:</b> 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.</li> </ul>	insert into

Category	Parameter	Description	Example Value
	Loader Threads	<p>Number of threads started in each loader. A larger number allows more concurrent write operations.</p> <p><b>NOTE</b> This parameter is unavailable if <b>Constraint Conflict Handling</b> is set to <b>replace into</b> or <b>on duplicate key update</b>.</p>	1
Advanced parameters	Import to Staging Tables	<p>If you set this parameter to <b>Yes</b>, 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.</p> <p>The default value is <b>No</b>, 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.</p> <p><b>NOTE</b> If you select <b>Clear part of data</b> or <b>Clear all data</b> for <b>Clear Data Before Import</b>, CDM does not roll back the deleted data in transaction mode.</p>	No
	Extending field length	<p>When <b>Auto creation</b> 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.</p> <p>When a character field is imported to DWS, the length of the character field must be automatically increased by three times.</p> <p>If a job fails to be executed and an error message similar to <b>value too long for type character varying</b> exists in the log when you import characters to DWS, you can enable this function to solve the problem.</p> <p><b>NOTE</b> When this function is enabled, some fields consume three times the storage space of the user.</p>	No

Category	Parameter	Description	Example Value
	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

## Field Mapping in Automatic Table Creation on DWS

**Figure 3-58** 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 3-58** 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

### 3.3.6.4.7 To DDS

If the destination link of a job is the [Link to DDS](#), configure the destination job parameters based on [Table 3-81](#).

**Table 3-81** 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

### 3.3.6.4.8 To DCS

If the data is imported to DCS, configure the destination job parameters based on [Table 3-82](#).

**Table 3-82** 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: <ul style="list-style-type: none"> <li><b>String</b>: without column name, such as <b>value1,value2</b></li> <li><b>Hash</b>: with column name, such as <b>column1=value1,column2=value2</b></li> </ul>	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	;

### 3.3.6.4.9 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 3-83](#).

**Table 3-83** 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.	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	<p>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:</p> <ul style="list-style-type: none"> <li>• <b>Every hour:</b> CDM creates indexes on the hour. The new indexes are named in the format of <i>Index name+Year+Month+Day+Hour</i>, for example, <b>index2018121709</b>.</li> <li>• <b>Every day:</b> CDM creates indexes at 00:00 every day. The new indexes are named in the format of <i>Index name+Year+Month+Day</i>, for example, <b>index20181217</b>.</li> <li>• <b>Every week:</b> CDM creates indexes at 00:00 every Monday. The new indexes are named in the format of <i>Index name+Year+Week</i>, for example, <b>index201842</b>.</li> <li>• <b>Every month:</b> CDM creates indexes at 00:00 on the first day of each month. The new indexes are named in the format of <i>Index name+Year+Month</i>, for example, <b>index201812</b>.</li> <li>• <b>Do not create:</b> Do not create indexes periodically.</li> </ul> <p>When extracting data from a file, you must configure a single extractor, which means setting <b>Concurrent Extractors</b> to <b>1</b>. Otherwise, this parameter is invalid.</p>	Every hour

### 3.3.6.4.10 To DLI

If the destination link of a job is the [Link to DLI](#), configure the destination job parameters based on [Table 3-84](#).

 NOTE

To use CDM to migrate data to DLI, you must obtain the read permissions of OBS.

**Table 3-84** 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 <b>Yes</b> , data in the destination table will be cleared before the task is started.	No
Data Clearing Mode	This parameter is available when <b>Clear Data Before Import</b> is set to <b>Yes</b> . <b>TRUNCATE</b> : deletes standard data. <b>INSERT_OVERWRITE</b> : overwrites existing data with inserted data.	TRUNCATE
Partition	This parameter is available when <b>Clear Data Before Import</b> is set to <b>Yes</b> .  When you enter partitions, data in these partitions will be cleared.	year=2020,location=sun

### 3.3.6.4.11 To OpenTSDB

If the destination link of a job is the [Link to CloudTable OpenTSDB](#), configure the destination job parameters based on [Table 3-85](#).

**Table 3-85** Parameter description

Parameter	Description	Example Value
Metric	(Optional) You can specify a metric or select an existing metric in OpenTSDB.	city.temp



Parameter	Description	Example Value
Time	(Optional) Data point. The value is a string or timestamp in the format of <i>yyyyMMddHHmmdd</i> .	1598870800
Tag	(Optional) Data tag	tagk:tagv, tagk2:tagv2

### 3.3.6.5 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 this parameter.

### 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.

### 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.

## 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.

## 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.

## 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.

### 3.3.6.6 Job Configuration Management

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

- [Maximum Concurrent Extractors of Jobs](#)
- [Scheduled Backup and Restoration of CDM Jobs](#)
- [Environment Variables of CDM Job Parameters](#)

#### Maximum Concurrent Extractors of Jobs

The value of this parameter ranges from 1 to 300. If the total number of extractors exceeds the value of this parameter, the excess extractors are queued. Determine the maximum number of concurrent extractors based on the number of concurrent extractors of each job.

Configure the number of concurrent extractors of a job based on the following rules:

The number of concurrent extractors in a CDM migration job is related to the cluster specifications and table size. The value range is 1 to 300. If the value is too large, the extractors are queued.

You are advised to set 4 concurrent extractors for each 1 CU (1 CU = 1 vCPU and 4 GB), as listed in [Table 3-86](#). You can also adjust the value as needed. If each row of the table contains less than or equal to 1 MB data, you can extract data concurrently. If each row contains more than 1 MB data, you are advised to extract data in a single thread.

#### NOTE

- 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.
- The number of concurrent extractors of a job is affected by **Maximum Concurrent Extractors** configured on the **Settings** page. The **Maximum Concurrent Extractors** parameter specifies the total number of concurrent extractions.

**Table 3-86** Reference configurations of concurrent extractors

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

#### Scheduled Backup and Restoration of CDM Jobs

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 3-87** 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	<ul style="list-style-type: none"> <li>• <b>All jobs:</b> 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.</li> <li>• <b>All jobs by groups:</b> You select one or more job groups to back up.</li> </ul>	All jobs
Backup Cycle	Select the backup cycle. <ul style="list-style-type: none"> <li>• <b>Day:</b> The backup is performed daily at 00:00:00.</li> <li>• <b>Week:</b> The backup is performed at 00:00:00 every Monday.</li> <li>• <b>Month:</b> The backup is performed at 00:00:00 on the first day of each month.</li> </ul>	Day
OBS Link for Writing Backups	Link used to back up jobs to OBS buckets. Select a link you have created on the <b>Links</b> 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 CDM 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 3-59** Setting the bucket names to environment variables

The screenshot shows the 'Job Configuration' window. At the top, the 'Job Name' is 'A-B'. Below are two main configuration panels: 'Source Job Configuration' and 'Destination Job Configuration'.  
**Source Job Configuration:**  
 - Source Link Name: OBS\_LINK1 (dropdown)  
 - Bucket Name: \${bucket\_1} (text field with refresh icon)  
 - Source Directory/File: FROM (text field with refresh icon)  
 - Entries Files: Yes (selected) / No (button)  
 - File Format: Binary (dropdown)  
 - Show Advanced Attributes: (link)  
**Destination Job Configuration:**  
 - Destination Link Name: OBS\_LINK1 (dropdown)  
 - Bucket Name: \${bucket\_2} (text field with refresh icon)  
 - Write Directory: TO (text field with refresh icon)  
 - File Format: Binary (dropdown)  
 - Duplicate File Processing Method: Replace (dropdown)  
 - Show Advanced Attributes: (link)  
 At the bottom, there are 'Cancel' and 'Next' buttons.

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
```

### 3.3.6.7 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

### 3.3.6.8 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.

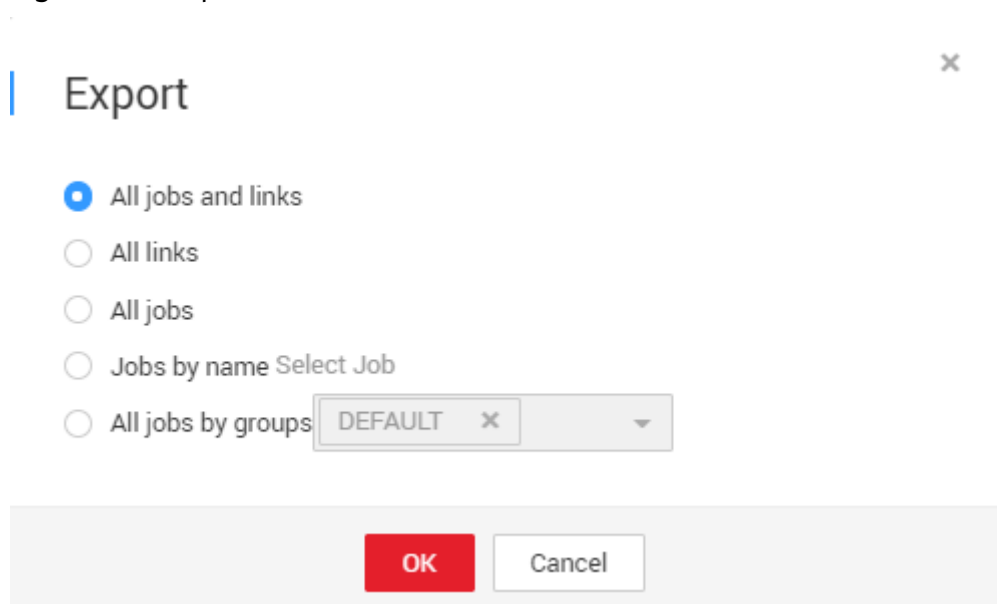
You can export and import jobs in batches in the following scenarios:

- 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 3-60 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

## 3.3.7 Auditing

### 3.3.7.1 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 3-88** 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

Operation	Resource Type	Trace Name
Starting a job	job	startJob
Stopping a job	job	stopJob

### 3.3.7.2 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.
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 the *Cloud Trace Service User Guide*.

## 3.3.8 Tutorials

### 3.3.8.1 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 **Adding Routes** in *Virtual Private Cloud (VPC) Usage Guide*. For details about how to configure security group rules, see **Security Group > Adding a Security Group Rule** in *Virtual Private Cloud (VPC) Usage Guide*.
- 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 3-61** Selecting a connector type



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

**Figure 3-62** Creating an MRS Hive link

* Name	<input type="text"/>
* Connector	Hive
* Hadoop Type	MRS
* Manager IP <span>?</span>	<input type="text"/> <a href="#">Select</a>
Authentication Method	SIMPLE
* HIVE Version <span>?</span>	HIVE_3_X
* Username <span>?</span>	<input type="text"/>
* Password	<input type="password"/>
* OBS storage support <span>?</span>	<input type="radio"/> Yes <input checked="" type="radio"/> No
* Run Mode <span>?</span>	EMBEDDED
Use Cluster Config <span>?</span>	<input type="radio"/> Yes <input checked="" type="radio"/> No
<a href="#">Show Advanced Attributes</a>	
<input type="button" value="X Cancel"/> <input type="button" value=" &lt; Previous"/> <input type="button" value=" Test"/> <input type="button" value=" Save"/>	

**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 3-89](#).

**Table 3-89** MRS Hive link parameters

Parameter	Description	Example Value
Group	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 <b>Select</b> next to the <b>Manager IP</b> 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 <ul style="list-style-type: none"> <li>• <b>SIMPLE</b>: for non-security mode</li> <li>• <b>KERBEROS</b>: for security mode</li> </ul>	SIMPLE
Hive Version	Hive version Set it to the Hive version on the server.	HIVE_3_X
Username	<p>If <b>Authentication Method</b> is set to <b>KERBEROS</b>, 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.</p> <p>To create a data connection for an MRS security cluster, do not use user <b>admin</b>. The <b>admin</b> 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 <b>Username</b> and <b>Password</b> to the username and password of the created MRS user when creating an MRS data connection.</p> <p><b>NOTE</b></p> <ul style="list-style-type: none"> <li>• 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 <b>Manager_viewer</b> role to create links on CDM. To perform operations on databases, tables, and data of a component, you also need to add the user group permissions of the component to the user.</li> <li>• 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 <b>Manager_administrator</b> or <b>System_administrator</b> to create links on CDM.</li> <li>• A user with only the <b>Manager_tenant</b> or <b>Manager_auditor</b> permission cannot create connections.</li> </ul>	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	<p>This parameter is used only when the Hive version is <b>HIVE_3_X</b>. Possible values are:</p> <ul style="list-style-type: none"> <li>• <b>EMBEDDED</b>: The link instance runs with CDM. This mode delivers better performance.</li> <li>• <b>STANDALONE</b>: 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 <b>STANDALONE</b> or configure different agents.</li> </ul> <p>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.</p>	EMBEDDED
Check Hive JDBC Connectivity	Whether to check the Hive JDBC connectivity	No
Use Cluster Config	You can create cluster configurations on the <b>Links</b> 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 **Link** page.

----End

### 3.3.8.2 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 3-63** Uploading a driver

Driver Name	Driver Package Name	Driver Type	Description	Operation
POSTGRES	None	Preset		Delete Upload Copy from SFTP
DB2	None	Preset		Delete Upload Copy from SFTP
SQLSERVER	None	Preset		Delete Upload Copy from SFTP
DDM	None	Preset		Delete Upload Copy from SFTP
MYCAT	None	Preset		Delete Upload Copy from SFTP
GAUSS_DB	None	Preset		Delete Upload Copy from SFTP
DM	None	Preset		Delete Upload Copy from SFTP
MYSQL	mysql-connector-java-5.1.44-bin.jar	Preset		Delete Upload Copy from SFTP
ORACLE_8	ojdbc6-11.2.0.4.jar	Preset	oracle > 12.1	Delete Upload Copy from SFTP
ORACLE_6	ojdbc6-11.2.0.4.jar	Preset	oracle < 12.1	Delete Upload Copy from SFTP

- 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 3-64](#).

**Figure 3-64** Selecting a connector type



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



**Figure 3-65** Creating a MySQL link

\* Name

\* Connector

Database Type

\* Database Server

\* Port

\* Database Name

\* Username

\* Password

Use Local API

Use Agent

Agent  [Select](#)

Driver Version [mysql-connector-java-5.1.48.jar](#) [Upload](#) | [Copy from SFTP](#)

[Hide Advanced Attributes](#)

Fetch Size

Commit Size

Attribute Name	Value	Operation
<input type="text" value="useCompression"/>	<input type="text" value="true"/>	<a href="#">Delete</a>

Reference Sign

**Table 3-90** MySQL link parameters

Parameter	Description	Example Value
Name	Enter a unique link name.	mysqlink
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 <b>local_infile</b> 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 <b>Select</b> and select the agent created in <a href="#">Connecting to an Agent</a> .	-
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

### 3.3.8.3 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 3-66](#).

**Figure 3-66** 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** Create a CDM cluster by following the instructions 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 3-67** Cluster list



Name	Status	Internal Network Address	Public Network Address	Source	Enterprise Project	Operation
	Running			CDM	default	Authorize EIP Check Job Management Bind EIP More

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

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 3-91](#).

**Table 3-91** MySQL link parameters

Parameter	Description	Example Value
Group	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
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 <b>local_infile</b> 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	-

Parameter	Description	Example Value
Agent	Click <b>Select</b> and select the agent created in <a href="#">Connecting to an Agent</a> .	-
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 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.

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

**Figure 3-68** Creating an MRS Hive link

* Name	<input type="text"/>
* Connector	Hive
* Hadoop Type	MRS
* Manager IP	<input type="text"/> <a href="#">Select</a>
Authentication Method	SIMPLE
* HIVE Version	HIVE_3_X
* Username	<input type="text"/>
* Password	<input type="password"/>
* OBS storage support	<input type="radio"/> Yes <input checked="" type="radio"/> No
* Run Mode	EMBEDDED
Use Cluster Config	<input type="radio"/> Yes <input checked="" type="radio"/> No
<a href="#">Show Advanced Attributes</a>	
<input type="button" value="X Cancel"/> <input type="button" value=" &lt; Previous"/> <input type="button" value=" Test"/> <input type="button" value=" Save"/>	

**Table 3-92** describes the parameters. You can configure the parameters according to the actual situation.

**Table 3-92** MRS Hive link parameters

Parameter	Description	Example Value
Group	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 <b>Select</b> next to the <b>Manager IP</b> 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 <ul style="list-style-type: none"> <li>• <b>SIMPLE</b>: for non-security mode</li> <li>• <b>KERBEROS</b>: for security mode</li> </ul>	SIMPLE
Hive Version	Hive version Set it to the Hive version on the server.	HIVE_3_X
Username	<p>If <b>Authentication Method</b> is set to <b>KERBEROS</b>, 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.</p> <p>To create a data connection for an MRS security cluster, do not use user <b>admin</b>. The <b>admin</b> 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 <b>Username</b> and <b>Password</b> to the username and password of the created MRS user when creating an MRS data connection.</p> <p><b>NOTE</b></p> <ul style="list-style-type: none"> <li>• 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 <b>Manager_viewer</b> role to create links on CDM. To perform operations on databases, tables, and data of a component, you also need to add the user group permissions of the component to the user.</li> <li>• 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 <b>Manager_administrator</b> or <b>System_administrator</b> to create links on CDM.</li> <li>• A user with only the <b>Manager_tenant</b> or <b>Manager_auditor</b> permission cannot create connections.</li> </ul>	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	<p>This parameter is used only when the Hive version is <b>HIVE_3_X</b>. Possible values are:</p> <ul style="list-style-type: none"> <li>• <b>EMBEDDED</b>: The link instance runs with CDM. This mode delivers better performance.</li> <li>• <b>STANDALONE</b>: 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 <b>STANDALONE</b> or configure different agents.</li> </ul> <p>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.</p>	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	<p>This parameter is valid only when <b>Use Cluster Config</b> is set to <b>Yes</b>. Select a cluster configuration that has been created.</p> <p>For details, see <a href="#">Managing Cluster Configurations</a>.</p>	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 3-69](#) illustrates how to create a migration job.

**Figure 3-69** Creating a job for migrating data from MySQL to Hive

**Job Configuration**

\* Job Name: mysql2hive\_partition

**Source Job Configuration**

\* Source Link Name: mysqlsink

Use SQL Statement:  Yes  No

\* Schema/Table Space: sqoop

\* Table Name: trip\_data

Show Advanced Attributes

**Destination Job Configuration**

\* Destination Link Name: hivelink

\* Database Name: default

\* Table Name: trip\_data

\* Auto Table Creation: Non-auto Creation

Clear Data Before Import:  Yes  No

**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 3-70](#).

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 3-70** Hive field mapping

Source Field						Destination Fi
Name	Example Value	Type	Operation			Name
id		BIGINT	Q		→	owner
name		VARCHAR(32)	Q		→	object_name
age		INT UNSIGNED	Q		→	object_type
sex		TINYINT	Q		→	created
date		DATETIME	Q		→	last_ddl_time
atamp		TIMESTAMP	Q		→	
Achievements		FLOAT UNSIGNED	Q		→	
timi		VARCHAR(16383)	Q		→	
yyy		CHAR(1)	Q		→	
bbb		BIGINT	Q		→	

**Step 3** Click to display the **Converter List** dialog box, and then choose **Create Converter > Expression conversion**. See [Figure 3-71](#).

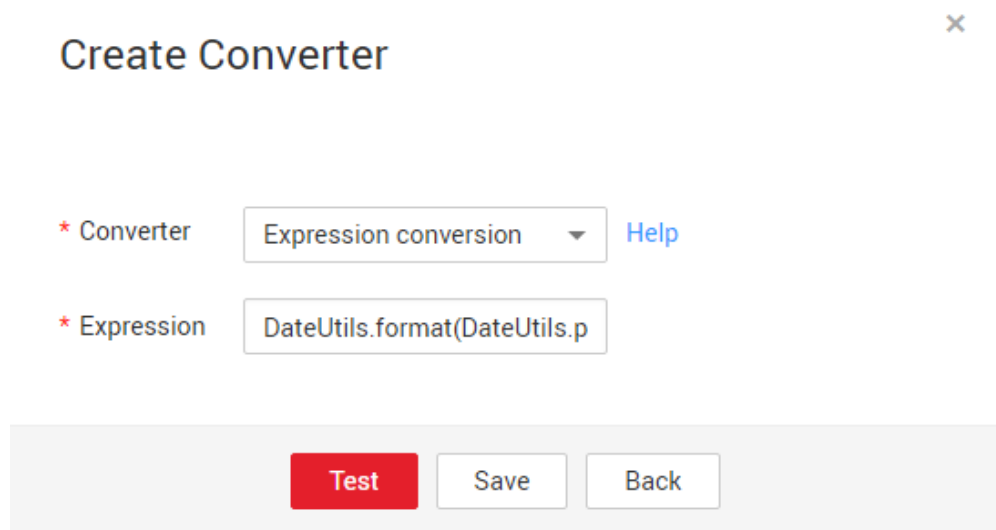
The expressions for the **y**, **ym**, and **ymd** fields are as follows:

```
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 3-71 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 log.

----End

### 3.3.8.4 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** Create a CDM cluster by following the instructions 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**.

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

Click **Show Advanced Attributes** and set optional parameters. For details, see [Link to a Common Relational Database](#). Retain the default values of the optional parameters and configure the mandatory parameters according to [Table 3-93](#).

**Table 3-93** MySQL link parameters

Parameter	Description	Example Value
Name	Unique link name	mysqllink
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 Agent	Whether to extract data from the data source through an agent	Yes
Agent	Click <b>Select</b> to select the agent created in .	-

**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 3-72** 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.

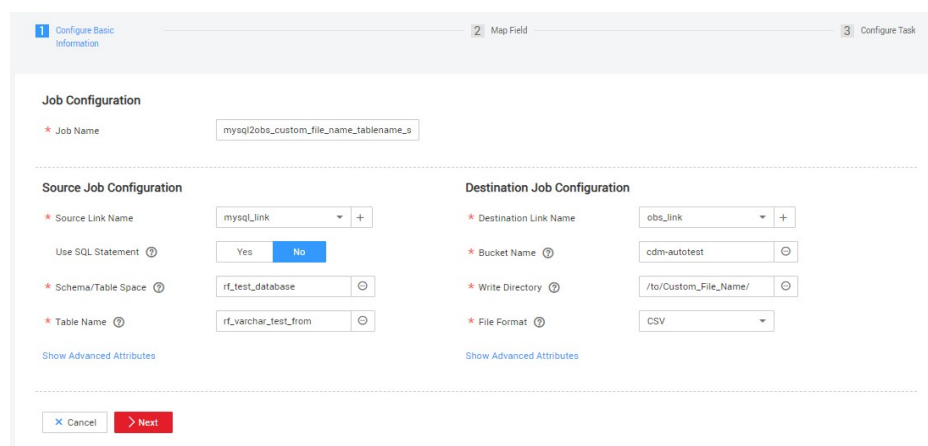
**Step 3** Click **Save**. The **Link Management** page is displayed.

----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 3-73** Creating a job for migrating data from MySQL to OBS



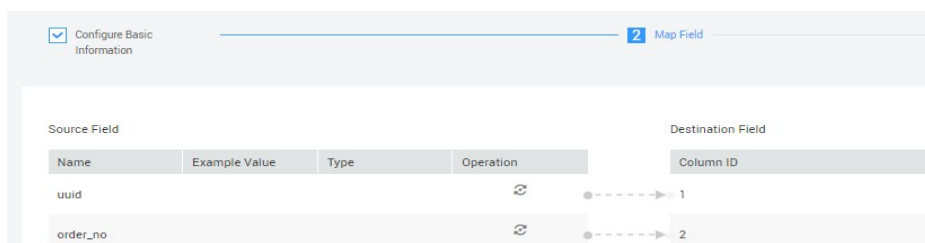
- **Job Name:** Enter a unique name.
- **Source Job Configuration**

- **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 3-74](#).

- 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.

**Figure 3-74** 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

### 3.3.8.5 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** Create a CDM cluster by following the instructions 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**.

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

Click **Show Advanced Attributes** and set optional parameters. For details, see [Link to a Common Relational Database](#). Retain the default values of the optional parameters and configure the mandatory parameters according to [Table 3-94](#).

**Table 3-94** MySQL link parameters

Parameter	Description	Example Value
Name	Unique link name	mysqllink
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 Agent	Whether to extract data from the data source through an agent	Yes
Agent	Click <b>Select</b> to select the agent created in .	-

**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**.
- Step 2** Select **Data Warehouse Service** and click **Next** to configure the DWS link parameters. Set the mandatory parameters listed in [Table 3-95](#) and retain the default values for the optional parameters.

**Table 3-95** 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 <b>Select</b> and select the agent created in <a href="#">Connecting to an Agent</a> .	-
Import Mode	<b>COPY</b> : Migrate the source data to the DWS management node and then copy the data to DataNodes. To access DWS through the Internet, select <b>COPY</b> .	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 3-75** 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).

- **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 3-76](#).

- 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.

**Figure 3-76** Table-to-table field mapping

Source Field	Example Value	Operation	Destination Field	Name	Type	Operation
1	L1		L1		string	
2	L2		L2		string	
3	L3		L3		string	
4	L4		L4		string	
5	Domain		Domain		string	
6	Type		Type		string	
7	2020YR		VR2020		string	
8	2021YR		VR2021		string	
9	2022YR		VR2022		string	
10	2023YR		VR2023		string	
11	2024YR		VR2024		string	
12	2025YR		VR2025		string	
13	2026YR		VR2026		string	

**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

### 3.3.8.6 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** Create a CDM cluster by following the instructions 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 3-77** 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**.

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

Click **Show Advanced Attributes** and set optional parameters. For details, see [Link to a Common Relational Database](#). Retain the default values of the optional parameters and configure the mandatory parameters according to [Table 3-96](#).

**Table 3-96** MySQL link parameters

Parameter	Description	Example Value
Name	Unique link name	mysqllink
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

Parameter	Description	Example Value
Password	Password of the user	-
Use Agent	Whether to extract data from the data source through an agent	Yes
Agent	Click <b>Select</b> to select the agent created in .	-

**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.

**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 3-78](#).

**Figure 3-78** Creating an entire DB migration job

Job Configuration

\* Job Name

---

Source Job Configuration

\* Source Link Name

\* Schema/Tablespace ⓘ

Destination Job Configuration

\* Destination Link Name

\* Schema/Tablespace ⓘ

Auto Table Creation ⓘ

Clear Data Before Import ⓘ

[Show Advanced Attributes](#)

---

- **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

### 3.3.8.7 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** Create a CDM cluster by following the instructions 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.

**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.

**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.

**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 3-79** 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 3-80](#).

- 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.

**Figure 3-80** Field mapping of Cloud Search Service

Source Field				Destination Field			
Name	Example Value	Type	Operation	Type	Name	Primary Key	Operation
aa	cdm-test	VARCHAR2(2000)		string	e	<input type="checkbox"/>	
bb	111	NUMBER(24-127)		string	i	<input type="checkbox"/>	

**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 . 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

### 3.3.8.8 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 created 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 cloud 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** Create a CDM cluster by following the instructions 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 3-81** Selecting a connector type



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

**Figure 3-82** Creating an Oracle link

* Name	<input type="text" value="oracle_link"/>
* Connector	<input type="text" value="Relational Database"/>
Database Type	<input type="text" value="Oracle"/>
* Database Server <span>?</span>	<input type="text" value="192.168.0.1"/>
* Port <span>?</span>	<input type="text" value="3306"/>
* Connection Type <span>?</span>	<input type="text" value="Service Name"/>
* Database Name <span>?</span>	<input type="text" value="db_user"/>
* Username <span>?</span>	<input type="text" value="sqoop"/>
* Password <span>?</span>	<input type="password"/>
Use Agent <span>?</span>	<input checked="" type="radio"/> Yes <input type="radio"/> No
Agent <span>?</span>	<input type="text"/> <a href="#">Select</a>
Oracle Version <span>?</span>	<input type="text" value="Earlier than 12.1.0.1"/>
Driver Version <span>?</span>	<a href="#">ojdbc6-11.2.0.4.jar Upload</a>   <a href="#">Copy from SFTP</a>
<a href="#">Hide Advanced Attributes</a>	
Fetch Size <span>?</span>	<input type="text" value="1000"/>
Link Attributes <span>?</span>	<input type="button" value="+ Add"/>
Reference Sign <span>?</span>	<input type="text" value=""/>
<input type="button" value="X Cancel"/> <input type="button" value="Test"/> <input type="button" value="Save"/>	

**Table 3-97** 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 <b>Select</b> and select the agent created in <a href="#">Connecting to an Agent</a> .	-
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.



**Step 2** Select **Data Warehouse Service** and click **Next** to configure the DWS link parameters. Set the mandatory parameters listed in **Table 3-98** and retain the default values for the optional parameters.

**Table 3-98** 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 <b>Select</b> and select the agent created in <b>Connecting to an Agent</b> .	-
Import Mode	<b>COPY</b> : Migrate the source data to the DWS management node and then copy the data to DataNodes. To access DWS through the Internet, select <b>COPY</b> .	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 3-83** Creating a job for migrating data from Oracle to DWS

Job Configuration

\* Job Name

**Source Job Configuration**

\* Source Link Name

Use SQL Statement

\* Schema/Table Space

\* Table Name

Show Advanced Attributes

**Destination Job Configuration**

\* Destination Link Name

\* Schema/Table Space

Auto Table Creation

\* Table Name

Clear Data Before Import

Import Mode

Show Advanced Attributes

- **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 3-84](#).

- 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.

**Figure 3-84** Table-to-table field mapping

Source Field	Example Value	Operation	Destination Field	Type
Column ID			Name	
1	L1		L1	string
2	L2		L2	string
3	L3		L3	string
4	L4		L4	string
5	Domain		Domain	string
6	type		Type	string
7	2020YR		VR2020	string
8	2021YR		VR2021	string
9	2022YR		VR2022	string
10	2023YR		VR2023	string
11	2024YR		VR2024	string
12	2025YR		VR2025	string
13	2026YR		VR2026	string

**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 log.

----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.

### 3.3.8.9 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

Create a CDM cluster by following the instructions 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.

**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.

**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 3-85** 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.

**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 3-86** Creating a job for migrating data from OBS to Cloud Search Service

Job Configuration

\* Job Name

---

Source Job Configuration	Destination Job Configuration
* Source Link Name <input type="text" value="obslink"/>	* Destination Link Name <input type="text" value="csslink"/>
* Bucket Name <input type="text" value="cdm-test"/>	* Index <input type="text" value="test-css"/>
* Source Directory/File <input type="text" value="/"/>	* Type <input type="text" value="css"/>
* File Format <input type="text" value="CSV"/>	<a href="#">Show Advanced Attributes</a>
<a href="#">Show Advanced Attributes</a>	

---

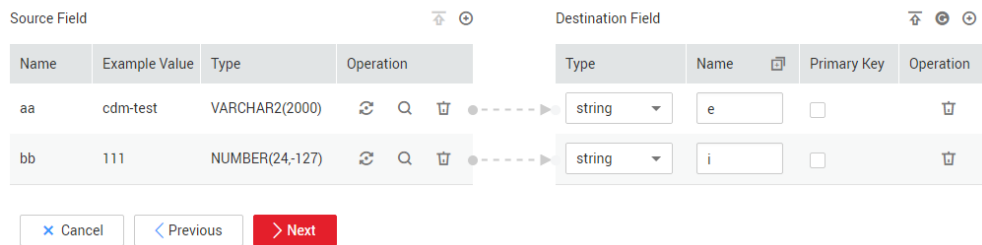
- **Job Name:** Enter a unique name.
- **Source Job Configuration**
  - **Source Link Name:** Select the **obslink** link created in [Creating an OBS Link](#).
  - **Bucket Name:** Select the bucket from which the data will be migrated.
  - **Source Directory/File:** Set this parameter to the path of 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 3-87](#).

- 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.

**Figure 3-87** 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 . 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

### 3.3.8.10 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

Create a CDM cluster by following the instructions 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.
- Step 2** Select **Data Lake Insight**, click **Next**, and configure the DLI link parameters. See [Figure 3-88](#).
  - **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 3-88** Creating a DLI link

* Name	<input type="text" value="dlilink"/>
* Connector	<input type="text" value="DLI"/>
* AK	<input type="text" value="GRC2WR0IDC6NGROYLWU2"/>
* SK	<input type="text" value="....."/>
* Project ID	<input type="text" value="c48475ce8e174a7a9f77570"/>

<input type="button" value="Cancel"/>	<input type="button" value="Previous"/>	<input type="button" value="Test"/>	<input type="button" value="Save"/>
---------------------------------------	---	-------------------------------------	-------------------------------------

**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 3-89** Selecting a connector type

Data Warehouse	<input type="button" value="Data Warehouse Service"/>	<input type="button" value="Data Lake Insight"/>		
Hadoop	<input type="button" value="MRS HDFS"/>	<input type="button" value="MRS HBase"/>	<input type="button" value="MRS Hive"/>	<input type="button" value="Apache HDFS"/>
	<input type="button" value="Apache HBase"/>	<input type="button" value="Apache Hive"/>		
Object Storage	<input type="button" value="Object Storage Service (OBS)"/>			
File System	<input type="button" value="FTP"/>	<input type="button" value="SFTP"/>	<input type="button" value="HTTP"/>	
Relational Database	<input type="button" value="RDS for MySQL"/>	<input type="button" value="RDS for PostgreSQL"/>	<input type="button" value="RDS for SQL Server"/>	<input type="button" value="MySQL"/>
	<input type="button" value="PostgreSQL"/>	<input type="button" value="Microsoft SQL Server"/>	<input type="button" value="Oracle"/>	
NoSQL	<input type="button" value="Redis"/>	<input type="button" value="MongoDB"/>		
Messaging System	<input type="button" value="Data Ingestion Service"/>	<input type="button" value="MRS Kafka"/>	<input type="button" value="Apache Kafka"/>	
Search	<input type="button" value="Elasticsearch"/>			
Open Beta Test	<input type="button" value="^"/>			
	<input type="button" value="Cancel"/>	<input type="button" value="Next"/>		

**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.

**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 3-90](#).

**Figure 3-90** Creating a job for migrating data from OBS to DLI

Job Configuration

\* Job Name

---

<p>Source Job Configuration</p> <p>* Source Link Name <input type="text" value="obslink"/> <input type="button" value="Create Link"/></p> <p>* Bucket Name <input type="text" value="obs-a0b377"/> <input type="button" value="..."/></p> <p>* Source Directory/File <input type="text" value="/obs-8909/"/> <input type="button" value="..."/></p> <p>* File Format <input type="text" value="CSV"/></p> <p><a href="#">Show advanced attributes.</a></p>	<p>Destination Job Configuration</p> <p>* Destination Link Name <input type="text" value="dlilink"/> <input type="button" value="Create Link"/></p> <p>* Resource Queue <input type="text" value="cdm"/> <input type="button" value="..."/></p> <p>* Database Name <input type="text" value="sqoop"/> <input type="button" value="..."/></p> <p>* Table Name <input type="text" value="t_test"/> <input type="button" value="..."/></p> <p>Clear Data Before Import <input type="checkbox"/> Yes <input checked="" type="checkbox"/> No</p>
--	---

- **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.

**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 log.

----End

### 3.3.8.11 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** Create a CDM cluster by following the instructions 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**.

**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 3-91** 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.

**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 3-92** Creating a job for migrating data from MRS HDFS to OBS

Job Configuration

\* Job Name: hdfs2obs\_004more

**Source Job Configuration**

- \* Source Link Name: hdfs\_link
- \* Source Directory/File: /interface/hdfsfrom/more1
- \* File Format: CSV

Show Advanced Attributes

**Destination Job Configuration**

- \* Destination Link Name: obs\_link
- \* Bucket Name: cdm-autotest
- \* Write Directory: /interface/obsto
- \* File Format: CSV
- Duplicate File Processing Method: Replace

Show Advanced Attributes

Cancel Next

- **Job Name:** Enter a unique name.
- **Source Job Configuration**
  - **Source Link Name:** Select the **hdfs\_link** 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.

**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

### 3.3.8.12 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 has been established.

## Creating a CDM Cluster and Binding an EIP to the Cluster

**Step 1** Create a CDM cluster by following the instructions 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.

**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.

**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**.

**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 3-93** Creating an entire DB migration job

The screenshot shows the 'Job Configuration' page. At the top, there is a 'Job Name' field with the value 'Elasticsearch2CSS'. Below this, the page is divided into two columns: 'Source Job Configuration' and 'Destination Job Configuration'. In the 'Source Job Configuration' column, there is a 'Source Link Name' dropdown menu set to 'es\_link', an 'Index' text box with 'test-css' and a selection icon, and a 'Clear Data Before Import' checkbox set to 'No'. In the 'Destination Job Configuration' column, there is a 'Destination Link Name' dropdown menu set to 'csslink', an 'Index' text box with 'css' and a selection icon, and a 'Clear Data Before Import' checkbox set to 'No'. At the bottom of the page, there are three buttons: 'Cancel', 'Save', and 'Save and Run'.

- **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 3-94** Historical Record

Executed By	Start Time	Last Updated	Duration	Status	Statistics	Schedule	Log
cdm	2018-07-25 11:37:20	2018-07-25 11:43:31	6m 11s	✔ Succeeded	Pending:0 / Running:0 / Succeeded:24 / Failed:0	False	No log available.

[← Back](#)

----End

## 3.3.9 Advanced Operations

### 3.3.9.1 Incremental Migration

#### 3.3.9.1.1 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

## 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](#).
  - 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 3-95 Time Filter**

**Source Job Configuration**

\* Source Link Name  [Configuration Guide](#)

\* Source Directory/File

\* File Format

**Hide Advanced Attributes**

Line Separator

Field Delimiter

Use Quote Char  Yes  No

Using RE to separate fields  Yes  No

First Row As Header  Yes  No

Encode type

Compression Format

Start Job by Marker File  Yes  No

File Separator

Filter Type

Time Filter  Yes  No

Minimum Timestamp

Maximum Timestamp

Disregard Non-existent Path/File  Yes  No

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.

### 3.3.9.1.2 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.

## WHERE Clause

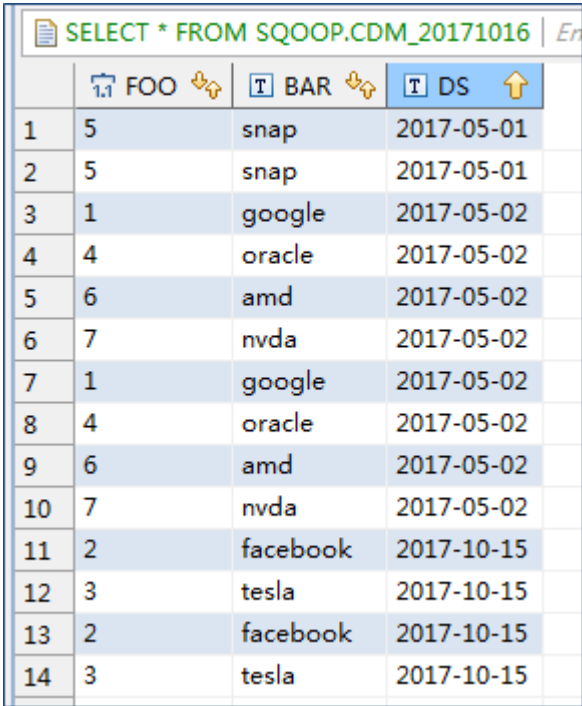
- Parameter position: When creating a table/file migration job, if the source end is a relational database, the **Where Clause** parameter is available in the advanced attributes of **Source Job Configuration**.
- Parameter principle: Set **WHERE Clause** to an SQL statement, for example, **age > 18 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 3-96**. Set the parameters as follows:

**Figure 3-96** Table data



	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

- WHERE Clause:** Set this parameter to **DS='\${dateformat(yyyy-MM-dd,-1,DAY)}'**.
- 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.

### 3.3.9.1.3 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
- Source table name
- Write directory
- Destination table name
- Where clause

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.

## 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 3-99](#) describes the macro variable definitions of time and date.

**Table 3-99** Macro variable definition of time and date

Macro Variable	Description	Display Effect
<code>\${dateformat(yyyy-MM-dd)}</code>	Returns the current date in <b>yyyy-MM-dd</b> format.	2017-10-16
<code>\${dateformat(yyyy/MM/dd)}</code>	Returns the current date in <b>yyyy/MM/dd</b> format.	2017/10/16
<code>\${dateformat(yyyy_MM_dd HH:mm:ss)}</code>	Returns the current time in <b>yyyy_MM_dd HH:mm:ss</b> format.	2017_10_16 09:00:00
<code>\${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}</code>	Returns the current time in <b>yyyy-MM-dd HH:mm:ss</b> format. The date is one day before the current day.	2017-10-15 09:00:00
<code>\${timestamp()}</code>	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
<code>\${timestamp(-10, MINUTE)}</code>	Returns the timestamp generated 10 minutes before the current time point.	1508115000000
<code>\${timestamp(dateformat(yyy yMMdd))}</code>	Returns the timestamp of 00:00:00 of the current day.	1508083200000
<code>\${timestamp(dateformat(yyy yMMdd,-1,DAY))}</code>	Returns the timestamp of 00:00:00 of the previous day.	1507996800000

Macro Variable	Description	Display Effect
\$ {timestamp(dateformat(yyy yMMddHH))}	Returns the timestamp of the current hour.	1508115600000

## Time and Date Macro Variables of Paths and Table Names

Figure 3-97 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 3-97** Setting **Table Name** and **Write Directory** to a time and date macro variable

The screenshot displays two configuration panels. The 'Source Job Configuration' panel includes fields for Source Link Name (oracle\_link), Use SQL Statement (No), Schema/Table Space (SQOOP), and Table Name (CDM\_/\${dateformat(yyyy-MM-dd)}). The 'Destination Job Configuration' panel includes fields for Destination Link Name (mshdfs\_link), Write Directory (/opt/ttxx/\${timestamp()}), and File Format (CSV).

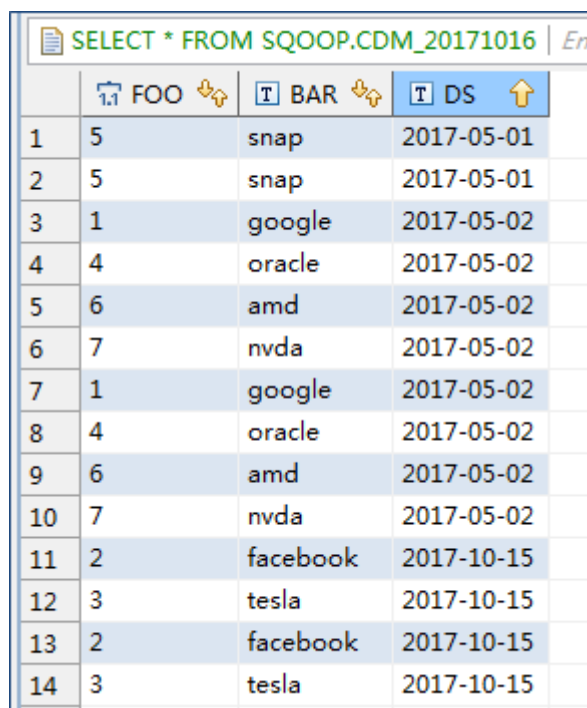
Currently, a table name or path name can contain multiple macro variables. For example, **/opt/ttxx/\${dateformat(yyyy-MM-dd)}/\${timestamp()}** is converted to **/opt/ttxx/2017-10-16/1508115701746**.

## Time and Date Macro Variables in the Where Clause

Figure 3-98 uses table **SQOOP.CDM\_20171016** as an example. The table contains column **DS**, which indicates the time.



Figure 3-98 Table data



	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.

### 3.3.9.1.4 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.

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 3-99** Time range

**Job Configuration**

\* Job Name

---

**Source Job Configuration**

\* Source Link Name  [Configuration Guide](#)

\* Table Name  ⓘ

Column Families

[Hide Advanced Attributes](#)

Split Rowkey  Yes  No

Minimum Timestamp ⓘ

Maximum Timestamp ⓘ

- 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.

### 3.3.9.2 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 3-100 Migration in transaction mode

#### Destination Job Configuration

**\* Destination Link Name**  [Configuration Guide](#)

**\* Schema/Table Space**

**\* Table Name**

**Clear Data Before Import**

[Hide Advanced Attributes](#)

**Is middle Relation table**  Yes  No

**PreSql**

**PostSql**

**Number of loader Thread**

 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.

### 3.3.9.3 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**:

- Encryption:** Select **AES-256-GCM**.
- 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.
- 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**:

- Encryption:** Select **AES-256-GCM**.
- 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.

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.

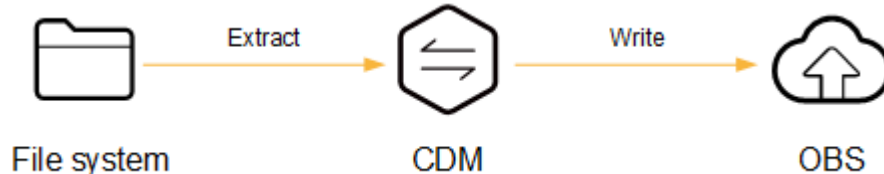
### NOTE

- 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.

### 3.3.9.4 MD5 Verification

CDM extracts data from the migration source and writes the data to the migration destination. **Figure 3-101** shows the migration mode when files are migrated to OBS.

**Figure 3-101** 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.

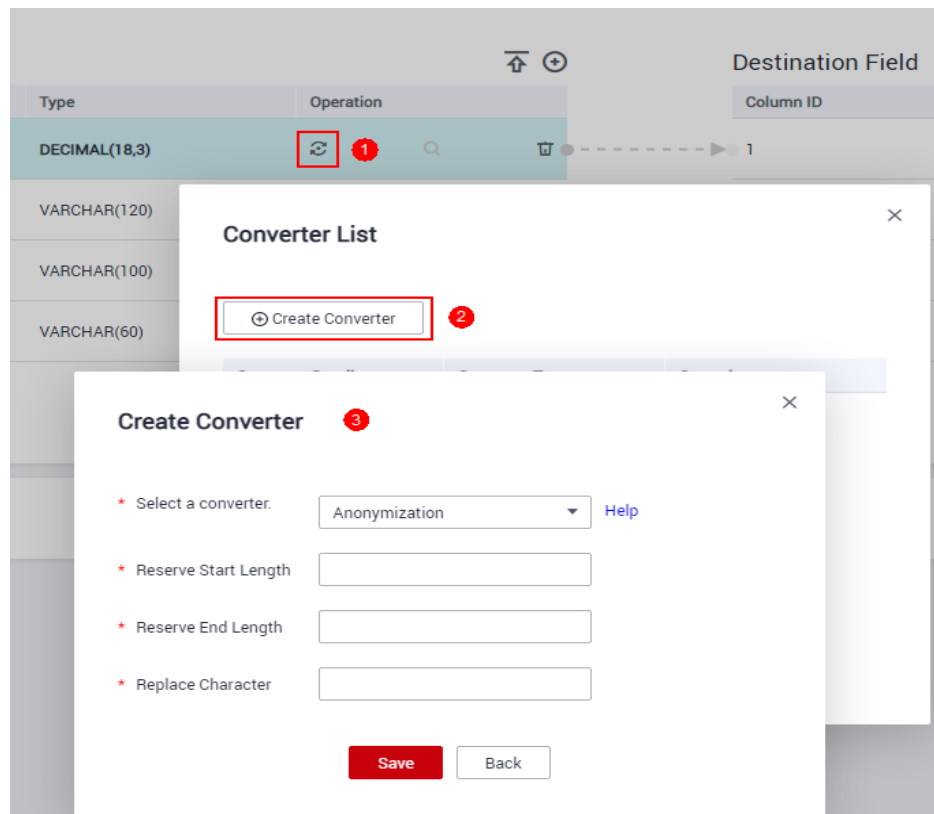
 **NOTE**

- 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.

### 3.3.9.5 Field Conversion

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

**Figure 3-102** 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 3-103 Anonymization

**Create Converter** ×

\* Converter  [Help](#)

\* Reserve Start Length

\* Reserve End Length

\* Replace Character

## Trim

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

## Reverse String

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

## Replace String

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

## 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, **zzabyxcdxx** 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 **zzabyxcdxx** is 0.

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

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

Expression: `StringUtils.isAlpha(value)`

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

Expression: `StringUtils.isAlphanumeric(value)`

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

Expression: `StringUtils.isAlphanumericSpace(value)`

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

Expression: `StringUtils.isAlphaSpace(value)`

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

Expression: `StringUtils.isAsciiPrintable(value)`

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

Expression: `StringUtils.isEmpty(value)`

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

Expression: `StringUtils.isNumeric(value)`

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

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

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

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

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

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

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

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

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

Expression: `StringUtils.length(value)`

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

- Expression: `StringUtils.remove(value,"ue")`
37. If the field is of the string type, remove the substring at the end of the field. If the specified substring is not at the end of the field, no conversion is performed. For example, remove **.com** at the end of **www.domain.com**.
- Expression: `StringUtils.removeEnd(value,".com")`
38. If the field is of the string type, delete the substring at the beginning of the field. If the specified substring is not at the beginning of the field, no conversion is performed. For example, delete **www.** at the beginning of **www.domain.com**.
- Expression: `StringUtils.removeStart(value,"www.")`
39. If the field is of the string type, replace all the specified character strings in the field. For example, replace **a** in **aba** with **z** to obtain **zbz**.
- Expression: `StringUtils.replace(value,"a","z")`
40. If the field is of the string type, replace multiple characters in the character string at a time. For example, replace **h** in **hello** with **j** and **o** with **y** to obtain **jelly**.
- Expression: `StringUtils.replaceChars(value,"ho","jy")`
41. If the string starts with the specified prefix (case sensitive), **true** is returned; otherwise, **false** is returned. For example, **abcdef** starts with **abc**, so that **true** is returned.
- Expression: `StringUtils.startsWith(value,"abc")`
42. If the field is of the string type, delete all the specified characters from the field. For example, delete all **x**, **y**, and **z** from **abcyx** to obtain **abc**.
- Expression: `StringUtils.strip(value,"xyz")`
43. If the field is of the string type, delete all the specified characters at the end of the field, for example, delete all spaces at the end of the field.
- Expression: `StringUtils.stripEnd(value,null)`
44. If the field is of the string type, delete all the specified characters at the beginning of the field, for example, delete all spaces at the beginning of the field.
- Expression: `StringUtils.stripStart(value,null)`
45. If the field is of the string type, obtain the substring after the specified position (excluding the character at the specified position) of the character string. If the specified position is a negative number, calculate the position in the descending order. For example, obtain the character string after the second character of **abcde**, that is, **cde**.
- Expression: `StringUtils.substring(value,2)`
46. If the field is of the string type, obtain the substring within the specified range of the character string. If the specified range is a negative number, calculate the range in the descending order. For example, obtain the character string between the second and fifth characters of **abcde**, that is, **cd**.
- Expression: `StringUtils.substring(value,2,5)`
47. If the field is of the string type, obtain the substring after the first specified character. For example, obtain the substring after the first **b** in **abcba**, that is, **cba**.
- Expression: `StringUtils.substringAfter(value,"b")`

48. If the field is of the string type, obtain the substring after the last specified character. For example, obtain the substring after the last **b** in **abcba**, that is, **a**.  
Expression: `StringUtils.substringAfterLast(value,"b")`
49. If the field is of the string type, obtain the substring before the first specified character. For example, obtain the substring before the first **b** in **abcba**, that is, **a**.  
Expression: `StringUtils.substringBefore(value,"b")`
50. If the field is of the string type, obtain the substring before the last specified character. For example, obtain the substring before the last **b** in **abcba**, that is, **abc**.  
Expression: `StringUtils.substringBeforeLast(value,"b")`
51. If the field is of the string type, obtain the substring nested within the specified string. If no substring is found, **null** is returned. For example, obtain the substring between **tag** in **tagabctag**, that is, **abc**.  
Expression: `StringUtils.substringBetween(value,"tag")`
52. If the field is of the string type, delete the control characters (`char≤32`) at both ends of the character string, for example, delete the spaces at both ends of the character string.  
Expression: `StringUtils.trim(value)`
53. Convert the character string to a value of the byte type. If the conversion fails, **0** is returned.  
Expression: `NumberUtils.toByte(value)`
54. Convert the character string to a value of the byte type. If the conversion fails, the specified value, for example, **1**, is returned.  
Expression: `NumberUtils.toByte(value,1)`
55. Convert the character string to a value of the double type. If the conversion fails, **0.0d** is returned.  
Expression: `NumberUtils.toDouble(value)`
56. Convert the character string to a value of the double type. If the conversion fails, the specified value, for example, **1.1d**, is returned.  
Expression: `NumberUtils.toDouble(value,1.1d)`
57. Convert the character string to a value of the float type. If the conversion fails, **0.0f** is returned.  
Expression: `NumberUtils.toFloat(value)`
58. Convert the character string to a value of the float type. If the conversion fails, the specified value, for example, **1.1f**, is returned.  
Expression: `NumberUtils.toFloat(value,1.1f)`
59. Convert the character string to a value of the int type. If the conversion fails, **0** is returned.  
Expression: `NumberUtils.toInt(value)`
60. Convert the character string to a value of the int type. If the conversion fails, the specified value, for example, **1**, is returned.  
Expression: `NumberUtils.toInt(value,1)`
61. Convert the character string to a value of the long type. If the conversion fails, **0** is returned.

- Expression: `NumberUtils.toLong(value)`
62. Convert the character string to a value of the long type. If the conversion fails, the specified value, for example, **1L**, is returned.
- Expression: `NumberUtils.toLong(value, 1L)`
63. Convert the character string to a value of the short type. If the conversion fails, **0** is returned.
- Expression: `NumberUtils.toShort(value)`
64. Convert the character string to a value of the short type. If the conversion fails, the specified value, for example, **1**, is returned.
- Expression: `NumberUtils.toShort(value, 1)`
65. Convert the IP string to a value of the long type, for example, convert **10.78.124.0** to **172915712**.
- Expression: `CommonUtils.ipToLong(value)`
66. Read an IP address and physical address mapping file from the network, and download the mapping file to the map collection. *url* indicates the address for storing the IP mapping file, for example, **http://10.114.205.45:21203/sqoop/IpList.csv**.
- Expression: `HttpsUtils.downloadMap("url")`
67. Cache the IP address and physical address mappings and specify a key for retrieval, for example, **ipList**.
- Expression: `CommonUtils.setCache("ipList", HttpsUtils.downloadMap("url"))`
68. Obtain the cached IP address and physical address mappings.
- Expression: `CommonUtils.getCache("ipList")`
69. Check whether the IP address and physical address mappings are cached.
- Expression: `CommonUtils.cacheExists("ipList")`
70. Based on the specified offset type (month/day/hour/minute/second) and offset (positive number indicates increase and negative number indicates decrease), convert the time in the specified format to a new time, for example, add 8 hours to **2019-05-21 12:00:00**.
- Expression: `DateUtils.getCurrentTimeByZone("yyyy-MM-dd HH:mm:ss", value, "hour", 8)`

### 3.3.9.6 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.

### 3.3.9.7 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 3-104** Setting regular expression parameters

### Source Job Configuration

* Source Link Name	<input type="text" value="obs-dayu-demo"/>
* Bucket Name <a href="#">?</a>	<input type="text" value="abcsze"/> <a href="#">...</a>
* Source Directory/File <a href="#">?</a>	<input type="text" value="/DAS_Imexport_Import_9e14"/> <a href="#">...</a>
* File Format <a href="#">?</a>	<input type="text" value="CSV"/>
<a href="#">Hide Advanced Attributes</a>	
Line Separator <a href="#">?</a>	<input type="text"/>
Use Quote Char <a href="#">?</a>	<input type="radio"/> Yes <input checked="" type="radio"/> No
Using RE to separate fields <a href="#">?</a>	<input checked="" type="radio"/> Yes <input type="radio"/> No
Regular Expression <a href="#">?</a>	<input type="text"/>
First Row As Header <a href="#">?</a>	<input type="radio"/> Yes <input checked="" type="radio"/> No
Encode type <a href="#">?</a>	<input type="text" value="UTF-8"/>
Compression Format <a href="#">?</a>	<input type="text" value="NONE"/>
Source File Processing Method <a href="#">?</a>	<input type="text" value="Do Nothing"/>

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 3-100** 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 3-101** 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 3-102** 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 3-103** 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 3-104** 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

### 3.3.9.8 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 3-105** Configuring the job

Job Configuration

\* Job Name

---

Source Job Configuration

\* Source Link Name  +

Use SQL Statement  Yes  No

\* Schema or Table Space  ⊖

\* Table Name  ⊖

[Show Advanced Attributes](#)

Destination Job Configuration


\* Destination Link Name  +

\* Resource Queue  ⊖

\* Database Name  ⊖

\* Table Name  ⊖

Clear Data Before Import  Yes  No

- Step 2** Click **Next** to go to the **Map Field** page and click .

**Figure 3-106** Configuring field mapping

Source #	Source Name	Operation	Destination #	Destination Name	Type	Operation
1	L3	COL	1	L3	string	COL
2	L3	COL	2	L3	string	COL
3	OrderItem	COL	3	OrderItem	string	COL
4	Type	COL	4	Type	string	COL
5	2022 VMS	COL	5	VMS001	string	COL
6	2022 VMS	COL	6	VMS002	string	COL
7	2022 VMS	COL	7	VMS003	string	COL
8	2022 VMS	COL	8	VMS004	string	COL
9	2022 VMS	COL	9	VMS005	string	COL
10	2022 VMS	COL	10	VMS006	string	COL
11	2022 VMS	COL	11	VMS007	string	COL
12	2022 VMS	COL	12	VMS008	string	COL


- 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 3-105](#).

**Figure 3-107** Add Field

Destination Field



**Add Field**

Name

Value

**Table 3-105** Macro variable definition of time and date

Macro Variable	Description	Display Effect
<code>\${dateformat(yyyy-MM-dd)}</code>	Returns the current date in <b>yyyy-MM-dd</b> format.	2017-10-16
<code>\${dateformat(yyyy/MM/dd)}</code>	Returns the current date in <b>yyyy/MM/dd</b> format.	2017/10/16
<code>\${dateformat(yyyy_MM_dd HH:mm:ss)}</code>	Returns the current time in <b>yyyy_MM_dd HH:mm:ss</b> format.	2017_10_16 09:00:00
<code>\${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}</code>	Returns the current time in <b>yyyy-MM-dd HH:mm:ss</b> format. The date is one day before the current day.	2017-10-15 09:00:00
<code>\${timestamp()}</code>	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
<code>\${timestamp(-10, MINUTE)}</code>	Returns the timestamp generated 10 minutes before the current time point.	1508115000000
<code>\${timestamp(dateformat(yyy yMMdd))}</code>	Returns the timestamp of 00:00:00 of the current day.	1508083200000
<code>\${timestamp(dateformat(yyy yMMdd,-1,DAY))}</code>	Returns the timestamp of 00:00:00 of the previous day.	1507996800000
<code>\${timestamp(dateformat(yyy yMMddHH))}</code>	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

### 3.3.9.9 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 subparameters:

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 3-106** 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 3-106](#).

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 3-108](#) shows that the value of the **name** field in the database contains a comma (,).

**Figure 3-108** Field value containing the field delimiter

The screenshot shows a database query interface. At the top, there is a search bar with the text 'city'. Below it, a SQL query is entered: 'select \* from sqoop.city | Enter a SQL expres'. Below the query, there is a table with three columns: 'id', 'name', and 'code'. The 'id' column has a value of '3', the 'name' column has a value of 'hello,world', and the 'code' column has a value of 'abc'. The table is highlighted in blue.

	T id	T name	T code
1	3	hello,world	abc

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

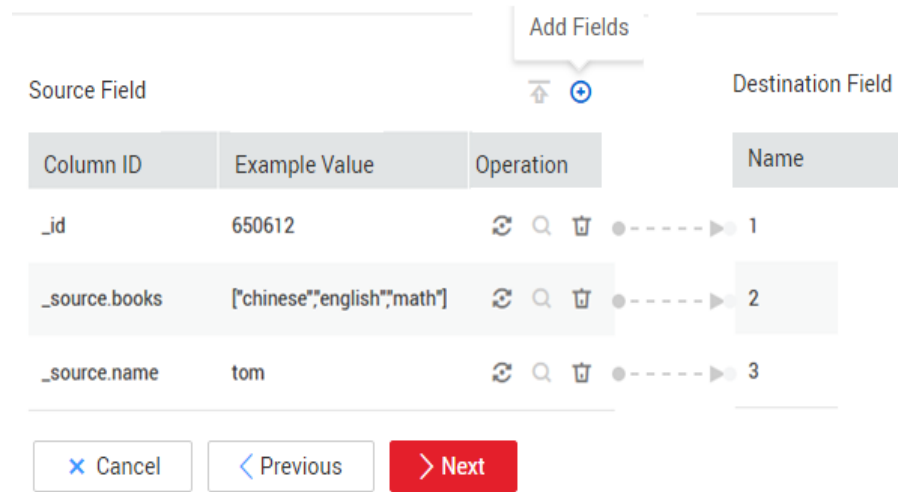
- d. Example 4: Configure a converter when parsing the JSON file. On the premise of [example 2](#), to add the **hits.max\_score** field to all records, that is, to write the data to the database in the following formats, perform the following operations:


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.

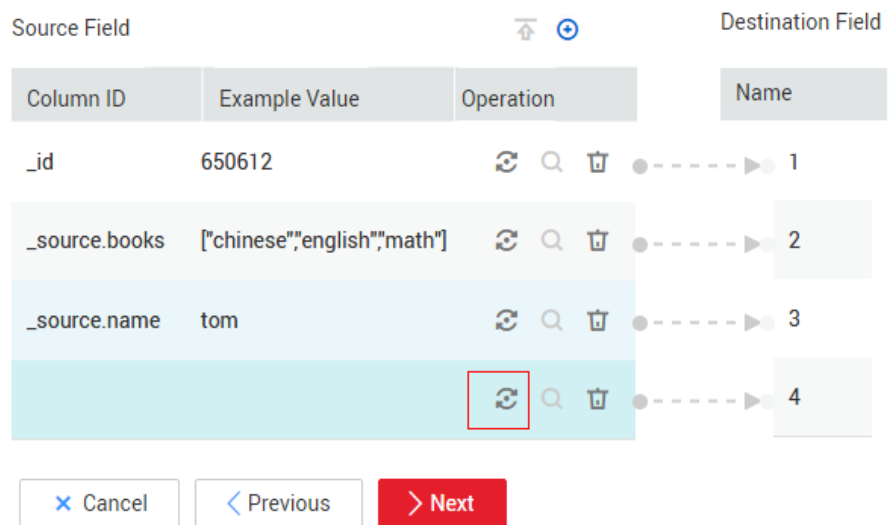
- i. Click  to add a field.

**Figure 3-109** Adding a field

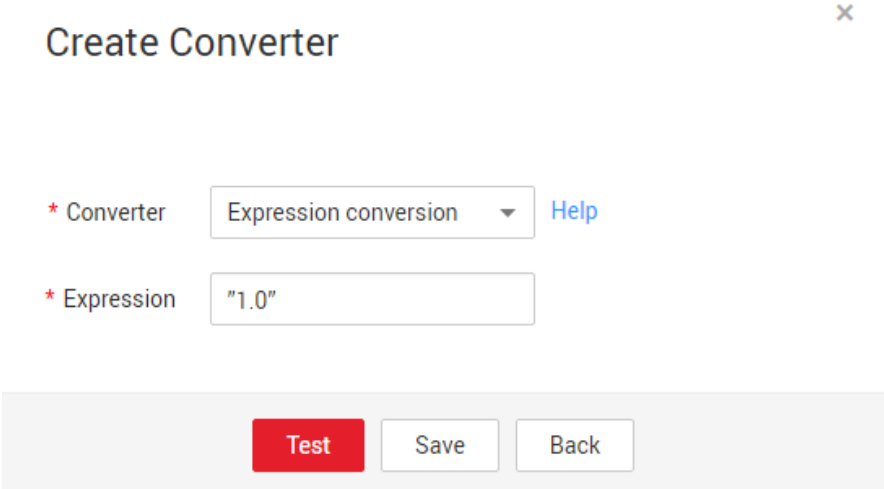


- ii. Click  to create a converter for the new field.

**Figure 3-110** Creating a field converter



- iii. Set **Converter** to **Expression conversion**, enter **"1.0"** in the **Expression** text box, and click **Save**.

**Figure 3-111** Configuring a field converter

**Create Converter** ×

\* Converter  [Help](#)

\* Expression

## 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 3-106](#).

- 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.

## 3.4 DataArts Factory

### 3.4.1 Overview

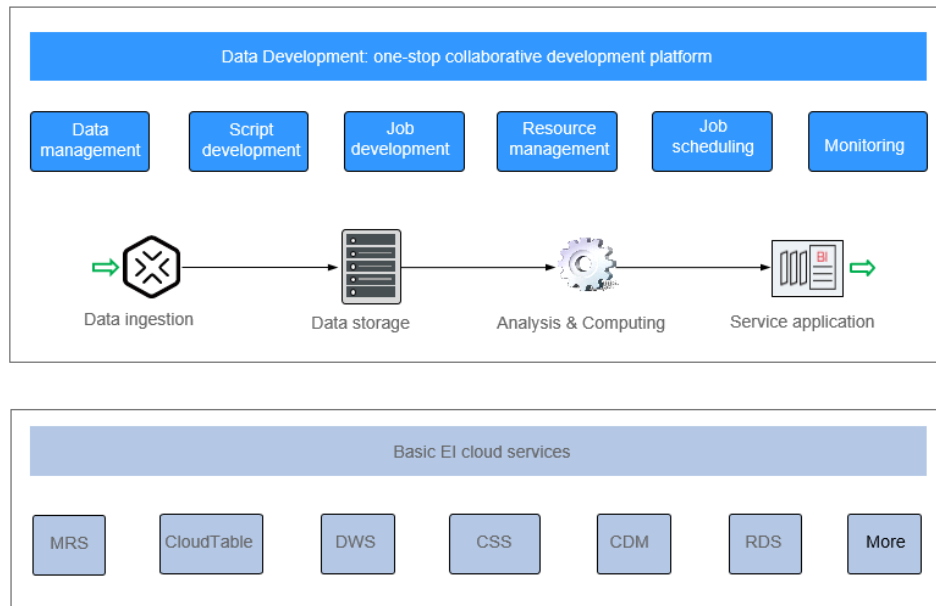
DataArts Factory is a one-stop big data collaborative development platform that provides fully managed big data scheduling capabilities. It manages various big data services, making big data more accessible than ever before and helping you effortlessly build big data processing centers.

DataArts Factory used to be Data Lake Factory (DLF). Therefore, in this document, both Data Lake Factory and DLF can be used to refer to DataArts Factory.

### Introduction to DataArts Factory

DataArts Factory enables a variety of operations such as data management, script development, job development, job scheduling, and monitoring, facilitating data analysis and processing.

**Figure 3-112** DataArts Factory architecture



## Main Functions

**Table 3-107** Main functions of DataArts Factory

Function	Description
Data management	<ul style="list-style-type: none"> <li>Manages multiple data warehouses, such as GaussDB(DWS), DLI and MRS Hive.</li> <li>Manages data tables using the GUI or data definition language (DDL).</li> </ul>
Script development	<ul style="list-style-type: none"> <li>Provides an online script editor that allows more than one operator to collaboratively develop and debug SQL, Python, and Shell scripts online.</li> <li>Allows use of variables and functions.</li> </ul>
Job development	<ul style="list-style-type: none"> <li>Provides a graphical designer that allows you to quickly build a data processing workflow by drag-and-drop.</li> <li>Presets multiple task types such as data integration, SQL, and Shell, and completes data analysis and processing by dependency between tasks.</li> <li>Supports job import and export.</li> </ul>
Resource management	Supports unified management of file, jar, and archive resources used during script and job development.
Job scheduling	Schedules jobs to run once or recursively and use events to trigger scheduling jobs.



Function	Description
Monitoring	<ul style="list-style-type: none"><li>• You can run, suspend, restore, or terminate a job.</li><li>• You can view the operation details of each job and each node in the job.</li><li>• You can use various methods to receive notifications when a job or task error occurs.</li></ul>

## Objects in DataArts Factory

- **Data connection:** A data collection is a collection of information required for accessing data storage (computing) space, including the connection type, name, and login information.
- **Solution:** A solution provides users with convenient and systematic management operations to better meet service requirements and objectives. Each solution can contain one or more business-related jobs, and one job can be used by multiple solutions.
- **Job:** A job is composed of one or more nodes that are performed collaboratively to complete data operations.
- **Script:** A script is an extension of a batch processing file. It is a program that stores text. Generally, a computer script program is a combination of a series of operations that control computers to perform operations. In the script program, certain logic branches can be implemented.
- **Node:** A node defines the operations performed on data.
- **Resource:** Resources refer to self-defined codes or text files that are uploaded by users and scheduled when node tasks are executed.
- **Expression:** Node parameter values in a node job can be dynamically generated based on the running environment by using Expression Language (EL). EL uses simple arithmetic and logic to calculate and reference embedded objects, including job objects and tool objects.
- **Environment variable:** An environment variable is an object with a specific name in the operating system. It contains information to be used by one or more applications.
- **PatchData:** PatchData refers to the instance that is generated in a period of time by a periodically scheduled job.

## 3.4.2 Data Management

### 3.4.2.1 Data Management Process

The data management function helps you quickly establish data models and provides you with data entities for script and job development. With data management, you can:

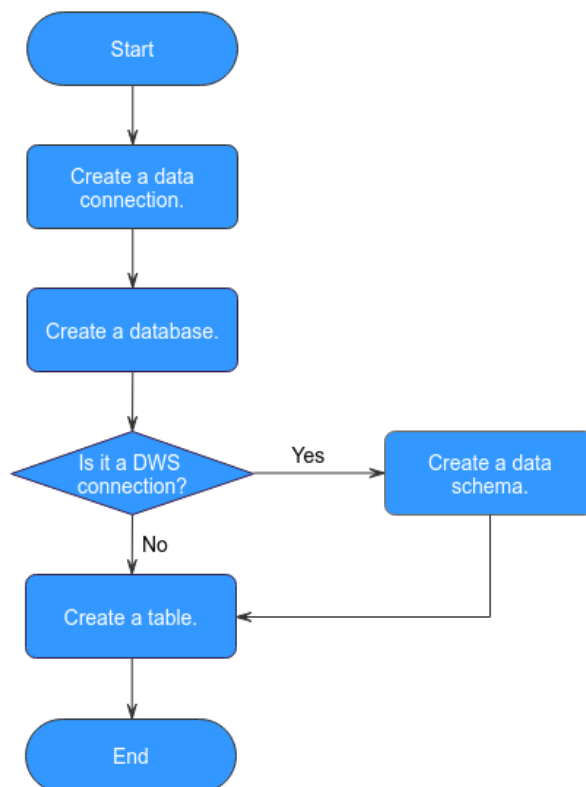
- Manage multiple types of data lakes, such as DWS and MRS Hive.
- Use the GUI and DDL to manage database tables.

**NOTE**

If you have created a data connection and a corresponding database and data table by referring to [Preparations Before Using DataArts Studio](#) before using DataArts Factory, you can skip data management operations and directly go to [Script Development](#) or [Job Development](#).

The following figure shows the process for using the data management function.

**Figure 3-113** Data management process



1. Create a data connection to connect to a data lake base service. For details, see [Creating a Data Connection](#).
2. Create a database based on the service type. For details, see [Creating a Database](#).
3. If the connection type is DWS, create a database schema and a table. If the connection type is not DWS, create a table. For details, see [\(Optional\) Creating a Database Schema](#).
4. Create a table. For details, see [Creating a Table](#).

### 3.4.2.2 Creating a Data Connection

After a data connection is created, you can perform data operations on DataArts Factory, for example, managing databases, namespaces, database schema, and tables.

With one data connection, you can run multiple jobs and develop multiple scripts. If the connection information saved in the data connection changes, you only need to modify the corresponding information in Connection Management.

## Creating a Data Connection

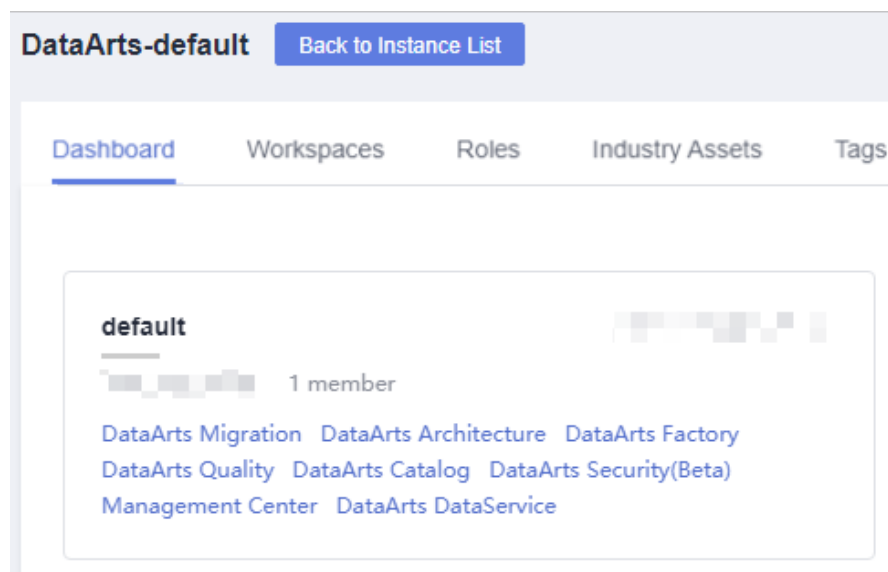
The data connection of DataArts Factory is created based on the data connection of Management Center. For details about how to create a data connection, see [Creating Data Connections](#).


## Viewing Connection References

To view the references of a connection, perform the following steps:

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-114** DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Data Development > Develop Script**.
3. Click  to display the connection list.
4. Right-click a connection in the list and select **View Reference**.
5. In the displayed **Reference List** dialog box, view the references of the connection.

### 3.4.2.3 Creating a Database

After creating a data connection, you can create a database on the console or using a SQL script.

- (Recommended) Console: You can directly create a database on the DataArts Studio DataArts Factory console with no code.
- SQL script: You can also develop and execute a SQL script for creating a database in the SQL editor of DataArts Studio's DataArts Factory module or a data lake product, and then use the script to create a database.

This section describes how to create a database on the DataArts Factory console.

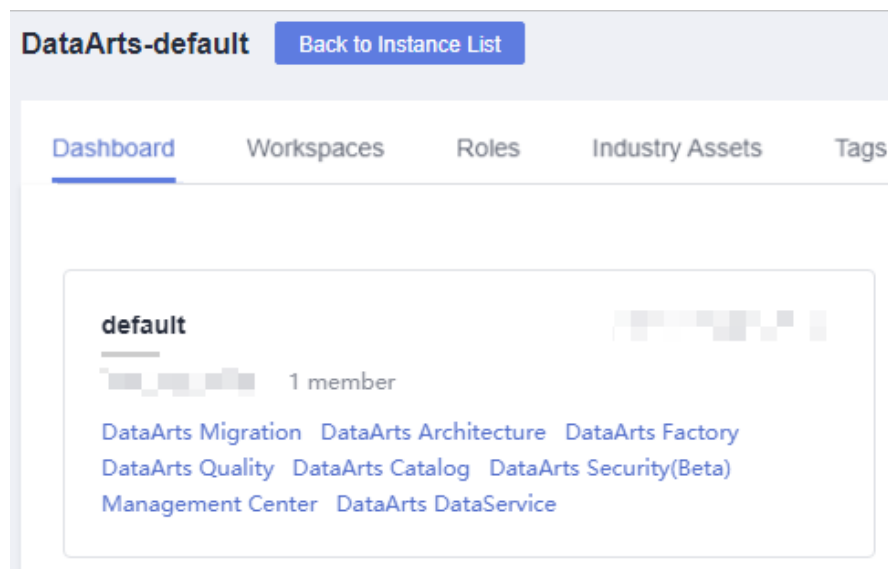
## Prerequisites


- You have already enabled the corresponding cloud services.
- A data connection has been created. For details, see [Creating a Data Connection](#).
- MRS API connections cannot be used to manage databases in a visualized mode. You are advised to create a database using SQL scripts.
- Before deleting a database, ensure that the database is not in use and is not associated with any data tables.

## Creating a Database on the DataArts Factory Console

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-115 DataArts Factory




2. In the left navigation pane of DataArts Factory, choose **Development > Develop Script** or **Development > Develop Job**.
3. In the menu on the left, click . Right-click the data connection for which you want to create a database, and choose **Create Database** from the shortcut menu. Set the parameters based on [Table 3-108](#).

**Table 3-108** Creating a database

Parameter	Mandatory	Description
Database Name	Yes	Name of a database. The naming rules are as follows: <ul style="list-style-type: none"> <li>• DLI: The value must contain only numbers, letters, and underscores (_). It cannot start with an underscore (_) or contain only numbers.</li> <li>• DWS: The value must contain only numbers, letters, and underscores (_). It cannot start with an underscore (_) or contain only numbers.</li> <li>• MRS Hive: The value must contain 1 to 128 characters, including only letters, numbers, and underscores (_). It must start with a number or letter and cannot contain only numbers.</li> </ul>
Description	No	Descriptive information about the database. The requirements are as follows: <ul style="list-style-type: none"> <li>• DLI: The value contains a maximum of 256 characters.</li> <li>• DWS: The value contains a maximum of 1,024 characters.</li> <li>• MRS Hive: The value contains a maximum of 1,024 characters.</li> </ul>

4. Click **OK**.

## Modifying a Database

1. In the left navigation pane of DataArts Factory, choose **Development > Develop Script** or **Development > Develop Job**.
2. In the menu on the left, click . Expand the data connection where the database is created, right-click the database name, and choose **Modify** from the shortcut menu.
3. In the **Modify Database** dialog box displayed, modify the database information.
4. Click **Yes**.

## Deleting a Database

1. In the left navigation pane of DataArts Factory, choose **Development > Develop Script** or **Development > Develop Job**.
2. In the menu on the left, click . Expand the data connection where the database is created, right-click the database name, and choose **Delete** from the shortcut menu.
3. In the displayed data connection list, click **Delete**.

4. Click **Yes**.

### 3.4.2.4 (Optional) Creating a Database Schema

After creating a DWS data connection, you can manage the database schemas under the DWS data connection.

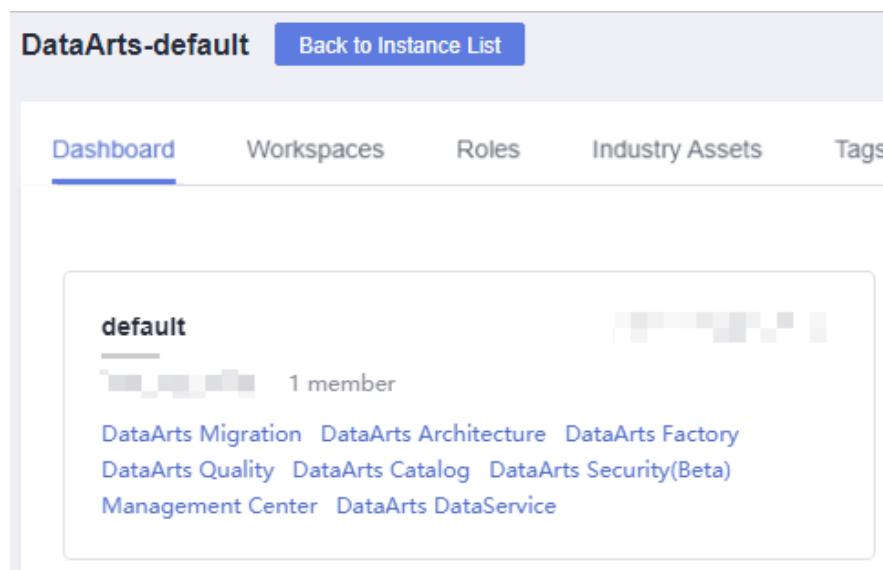
#### Prerequisites


- A DWS data connection has been created. For details, see [Creating a Data Connection](#).
- A DWS database has been created. For details, see [Creating a Database](#).

#### Creating a Database Schema

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-116** DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Development > Develop Script** or **Development > Develop Job**.
3. In the menu on the left, click . Click a DWS data connection name, select the database to be configured, and expand the directory level to **schemas**. Then right-click **schemas**, and choose **Create Schema** from the shortcut menu.
4. In the displayed dialog box, set the schema parameters based on [Table 3-109](#).


**Table 3-109** Creating a database schema

Parameter	Mandatory	Description
Mode Name	Yes	Name of a database schema.

Parameter	Mandatory	Description
Description	No	Descriptive information about the database schema.


5. Click **OK**.

## Modifying a Database Schema

1. In the left navigation pane of DataArts Factory, choose **Development > Develop Script** or **Development > Develop Job**.
2. Choose  from the menu on the left, click the data connection name, select a database, and expand the directory level to the database schema you want to modify. Right-click the database schema name and choose **Modify** from the shortcut menu.
3. In the displayed dialog box, modify the description of the database schema.
4. Click **OK**.

## Deleting a Database Schema

### NOTE

- The default database schema cannot be deleted.
  - Deleted database schemas cannot be recovered. Exercise caution when performing this operation.
1. In the left navigation pane of DataArts Factory, choose **Development > Develop Script** or **Development > Develop Job**.
  2. Choose  from the menu on the left, click the data connection name, select a database, and expand the directory level to the database schema you want to delete. Right-click the database schema name and choose **Delete** from the shortcut menu.
  3. In the displayed dialog box, click **OK**.

### 3.4.2.5 Creating a Table

You can create a table on the DataArts Factory console, in DDL mode, or using a SQL script.

- (Recommended) Console: You can directly create a table on the DataArts Studio DataArts Factory console with no code.
- (Recommended) DDL mode: You can select the DDL mode in DataArts Studio's DataArts Factory mode to create a table using a SQL script.
- SQL script: You can also develop and execute a SQL script for creating a table in the SQL editor of DataArts Studio's DataArts Factory module or a data lake product, and then use the script to create a table.

This section describes how to create a table on the DataArts Factory console and in DDL mode.

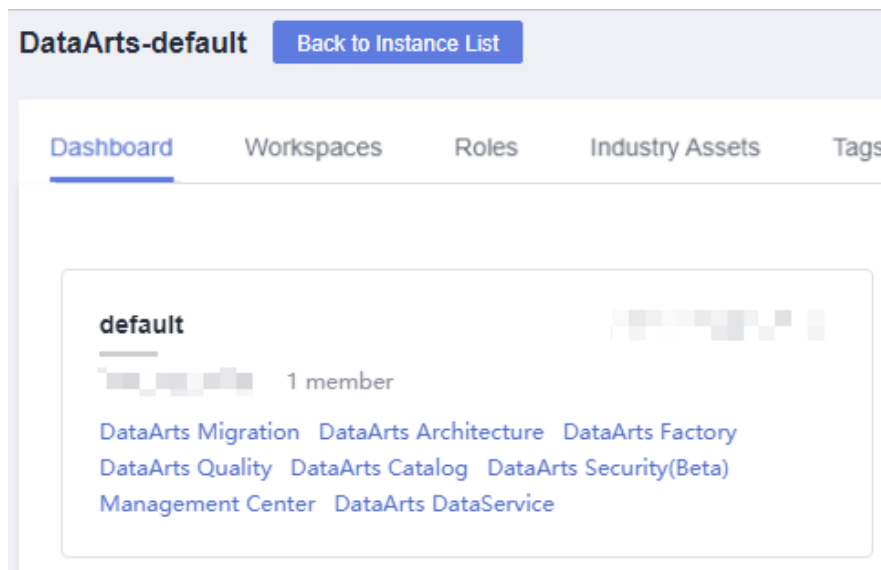
## Prerequisites


- A database has been created in the cloud service.
- A data connection that matches the table type has been created in DataArts Factory. For details, see [Creating a Data Connection](#).

## Creating a Table (GUI Mode)

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-117** DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Development** > **Develop Script** or **Development** > **Develop Job**.
3. Choose  from the menu on the left and expand the directory of a data connection to **tables** under **Data Connections**. Right-click **tables** and choose **Create Data Table** from the shortcut menu.
4. In the displayed dialog box, configure basic properties. Specific settings vary depending on the data connection type you select. [Table 3-110](#) lists the links for viewing property parameters of each type of data connection.

**Table 3-110** Basic property parameters

Data Connection Type	Description
DLI	For details, see the <b>Basic Property</b> part in <a href="#">Table 3-114</a> .
DWS	For details, see the <b>Basic Property</b> part in <a href="#">Table 3-115</a> .
MRS Hive	For details, see the <b>Basic Property</b> part in <a href="#">Table 3-116</a> .



- Click **Next**. On the **Configure Table Structure** page, configure the table structure parameters based on [Table 3-111](#).

**Table 3-111** Table structure

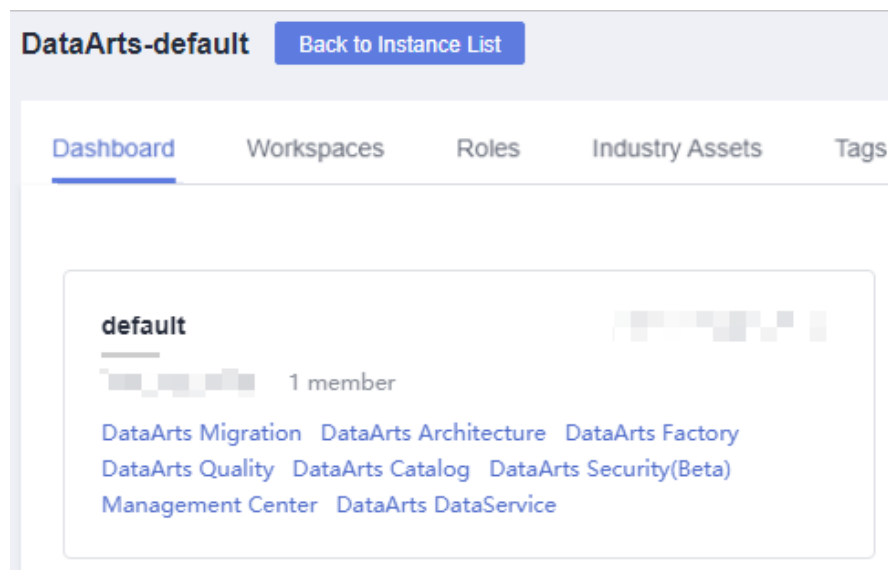
Data Connection Type	Description
DLI	For details, see the <b>Table Structure</b> part in <a href="#">Table 3-114</a> .
DWS	For details, see the <b>Table Structure</b> part in <a href="#">Table 3-115</a> .
MRS Hive	For details, see the <b>Table Structure</b> part in <a href="#">Table 3-116</a> .


- Click **OK**.

## Creating a Table (DDL Mode)

- Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-118** DataArts Factory





- In the left navigation pane of DataArts Factory, choose **Development > Develop Script** or **Development > Develop Job**.
- Choose  from the menu on the left and expand the directory of a data connection to **tables** under **Data Connections**. Right-click **tables** and choose **Create Data Table** from the shortcut menu.
- Click **DDL-based Table Creation**, configure the parameters based on [Table 3-112](#), and enter SQL statements in the editor in the lower part.

**Table 3-112** Data table parameters

Parameter	Description
Data Connection Type	Type of data connection to which the table belongs. <ul style="list-style-type: none"> <li>• DLI</li> <li>• DWS</li> <li>• HIVE</li> </ul>
Data Connection	Data connection to which the table belongs.
Database	Database to which the table belongs.

5. Click **OK**.


## Viewing Table Details

1. In the left navigation pane of DataArts Factory, choose **Development > Develop Script** or **Development > Develop Job**.
2. Choose  from the menu on the left and expand the directory of a data connection to a table name under **Data Connections**. Right-click the table name and choose **View Details** from the shortcut menu.
3. In the displayed dialog box, view the table information listed in .


**Table 3-113** Table details

Tab Name	Description
Table Information	Displays the basic information and storage information about the table.
Field Information	Displays the field information about the table.
Data Preview	Displays 10 records in the table.
DDL	Displays the DDL of the DWS, DLI, or MRS Hive data table.

## Viewing Table Column Details

1. In the left navigation pane of DataArts Factory, choose **Development > Develop Script** or **Development > Develop Job**.
2. Choose  from the menu on the left and expand the data connection directory to view column information under a desired table.


## Deleting a Table

1. In the left navigation pane of DataArts Factory, choose **Development > Develop Script** or **Development > Develop Job**.
2. Choose  from the menu on the left and expand the directory of a data connection to a table name under **Data Connections**. Right-click the table name and choose **Delete** from the shortcut menu.
3. In the **Delete Data Table** dialog box, click **OK**.

## Parameter Description

Table 3-114 DLI data table


Parameter	Mandatory	Description
<b>Basic Property</b>		
Table Name	Yes	Name of a table. The name must contain 1 to 63 characters, including only lowercase letters, numbers, and underscores (_). It cannot contain only numbers or start with an underscore.
Alias	No	Alias of a table. The alias must contain 1 to 63 characters, including only letters, numbers, and underscores (_). It cannot contain only numbers or start with an underscore.
Data Connection	Yes	Data connection to which the table belongs.
Database	Yes	Database to which the table belongs.
Data Location	Yes	Location to save data. Possible values: <ul style="list-style-type: none"> <li>• OBS</li> <li>• DLI</li> </ul>

Parameter	Mandatory	Description
Data Format	Yes	Format of data. This parameter is available only when <b>Data Location</b> is set to <b>OBS</b> . Possible values: <ul style="list-style-type: none"> <li>• <b>parquet</b>: DLF can read non-compressed parquet data and parquet data compressed using Snappy or gzip.</li> <li>• <b>csv</b>: DLF can read non-compressed CSV data and CSV data compressed using gzip.</li> <li>• <b>orc</b>: DLF can read non-compressed ORC data and ORC data compressed using Snappy.</li> <li>• <b>json</b>: DLF can read non-compressed JSON data and JSON data compressed using gzip.</li> </ul>
Path	Yes	OBS path where the data is stored. This parameter is available only when <b>Data Location</b> is set to <b>OBS</b> .
Table Description	No	Descriptive information about the table.
<b>Table Structure</b>		
Column Name	Yes	Name of the column. The name must be unique.
Type	Yes	Type of data.
Column Description	No	Descriptive information about the column.
Operation	No	To add a column, click  .


**Table 3-115** DWS data table

Parameter	Mandatory	Description
<b>Basic Property</b>		
Table Name	Yes	Name of a table. The name must contain 1 to 63 characters, including only letters, numbers, and underscores (_). It cannot contain only numbers or start with an underscore.

Parameter	Mandatory	Description
Alias	No	Alias of a table. The alias must contain 1 to 63 characters, including only letters, numbers, and underscores (_). It cannot contain only numbers or start with an underscore.
Data Connection	Yes	Data connection to which the table belongs.
Database	Yes	Database to which the table belongs.
Schema	Yes	Schema of the database.
Table Description	No	Descriptive information about the table.
Advanced Settings	No	<p>The following advanced options are available:</p> <ul style="list-style-type: none"> <li>● Storage method of a table. Possible values: <ul style="list-style-type: none"> <li>- <b>Row store</b></li> <li>- <b>Column store</b></li> </ul> </li> <li>● Compression level of a table <ul style="list-style-type: none"> <li>- Available values when the storage method is row store: <b>YES</b> or <b>NO</b>.</li> <li>- Available values when the storage method is column store: <b>YES, NO, LOW, MIDDLE, or HIGH</b>. For the same compression level in column store mode, you can configure compression grades from 0 to 3. Within any compression level, the higher the grade, the greater the compression ratio.</li> </ul> </li> </ul>
<b>Table Structure</b>		
Column Name	Yes	Name of the column. The name must be unique.

Parameter	Mandatory	Description
Data Classification	Yes	Classification of data. Possible values: <ul style="list-style-type: none"> <li>• Value</li> <li>• Currency</li> <li>• Boolean</li> <li>• Binary</li> <li>• Character</li> <li>• Time</li> <li>• Geometric</li> <li>• Network address</li> <li>• Bit string</li> <li>• Text search</li> <li>• UUID</li> <li>• JSON</li> <li>• OID</li> </ul>
Data Type	Yes	Type of data.
Column Description	No	Descriptive information about the column.
Create ES Index	No	If you click the check box, an ES index needs to be created. When creating the ES index, select the created CSS cluster from the <b>CloudSearch Cluster Name</b> drop-down list. For details about how to create a CSS cluster, see <i>Cloud Search Service User Guide</i> .
Index Data Type	No	Data type of the ES index. The options are as follows: <ul style="list-style-type: none"> <li>• text</li> <li>• keyword</li> <li>• date</li> <li>• long</li> <li>• integer</li> <li>• short</li> <li>• byte</li> <li>• double</li> <li>• boolean</li> <li>• binary</li> </ul>
Operation	No	To add a column, click  .

**Table 3-116** Basic property parameters of an MRS Hive data table

Parameter	Mandatory	Description
<b>Basic Property</b>		
Table Name	Yes	Name of a table. The name must contain 1 to 63 characters, including only lowercase letters, numbers, and underscores (_). It cannot contain only numbers or start with an underscore.
Alias	No	Alias of a table. The alias must contain 1 to 63 characters, including only letters, numbers, and underscores (_). It cannot contain only numbers or start with an underscore.
Data Connection	Yes	Data connection to which the table belongs.
Database	Yes	Database to which the table belongs.
Table Description	No	Descriptive information about the table.
<b>Table Structure</b>		
Column Name	Yes	Name of the column. The name must be unique.
Data Classification	Yes	Classification of data. Possible values: <ul style="list-style-type: none"> <li>• Original type</li> <li>• ARRAY</li> <li>• MAP</li> <li>• STRUCT</li> <li>• UNION</li> </ul>
Data Type	Yes	Type of data. See <a href="#">LanguageManual DDL</a> .
Column Description	No	Descriptive information about the column.
Operation	No	To add a column, click  .

## 3.4.3 Script Development

### 3.4.3.1 Script Development Process

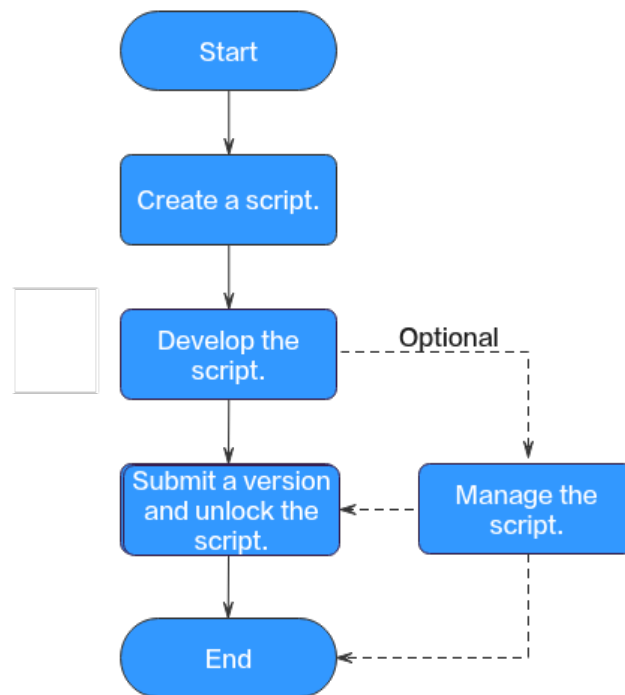
The script development function provides the following capabilities:

- Provides an online script editor for developing and debugging SQL, Python, and Shell scripts.

- Supports script import and export.
- Allows use of variables and functions.
- Provides editing locks for collaborative development.
- Supports script version management.

The following figure shows the process of script development.

**Figure 3-119** Script development process



1. Create a script of the corresponding type. For details, see [Creating a Script](#).
2. Develop the script: Develop, debug, and execute the script online. For details, see [Developing Scripts](#).
3. Submit a version and unlock the script: After performing this step, the script can be scheduled by jobs and modified by other developers. For details, see [Submitting a Version and Unlocking the Script](#).
4. (Optional) Manage the script: After the script development is complete, you can manage the script as required. For details, see [\(Optional\) Managing Scripts](#).

### 3.4.3.2 Creating a Script

DataArts Factory allows you to edit, debug, and run scripts online. You must create a script before developing it.

Currently, you can create the following types of scripts in DataArts Factory:

- DLI SQL
- Hive SQL
- DWS SQL



- Spark SQL
- Flink SQL
- RDS SQL
- Presto SQL
- Shell
- Python

## Prerequisites

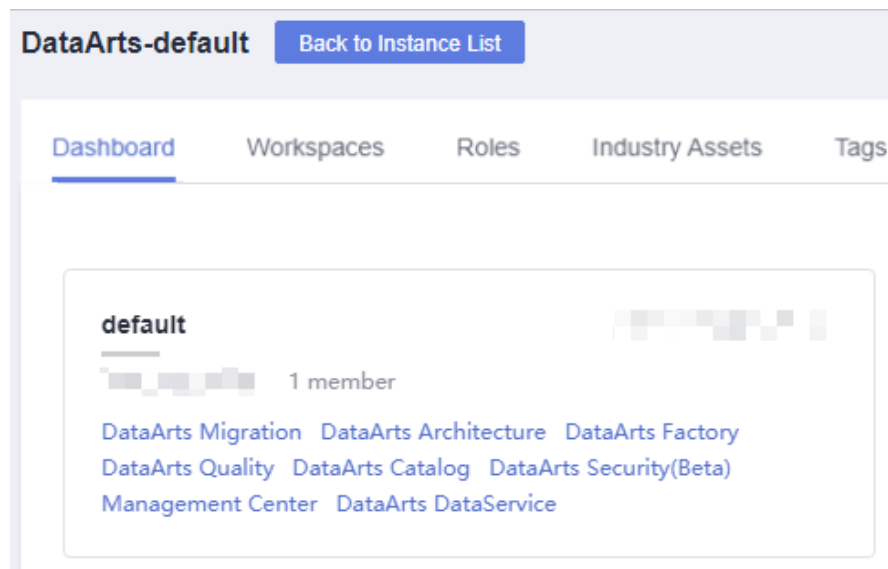
You have completed operations in [Creating a Data Connection](#) and [Creating a Database](#).

## Procedure

**Creating a Directory (If a directory already exists, you do not need to create one.)**

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-120** DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Data Development > Develop Script**.
3. In the directory list, right-click a directory and choose **Create Directory** from the shortcut menu.
4. In the displayed dialog box, configure directory parameters. [Table 3-117](#) describes the directory parameters.

**Table 3-117** Script directory parameters

Parameter	Description
Directory Name	Name of the script directory. The name must contain 1 to 64 characters, including only letters, numbers, underscores (_), and hyphens (-).
Select Directory	Parent directory of the script directory. The parent directory is the root directory by default.

5. Click **OK**.

### Creating a Script

1. In the left navigation pane of DataArts Factory, choose **Data Development > Develop Script**.
2. Create a script using either of the following methods:  
Method 1: In the right pane, click a script type to start creating a script.  
Method 2: In the directory list, right-click a directory and choose **Create Script** from the shortcut menu.
3. Go to the script development page. For details, see [Developing an SQL Script](#), [Developing a Shell Script](#), and [Developing a Python Script](#).

#### NOTE

A maximum of five temporary scripts of the same type can be created. If you close a temporary script without saving it and create a script of the same type, the closed temporary script will be opened again.

## 3.4.3.3 Developing Scripts

### 3.4.3.3.1 Developing an SQL Script

You can develop, debug, and run SQL scripts online. The developed scripts can be run in jobs. For details, see [Developing a Job](#).

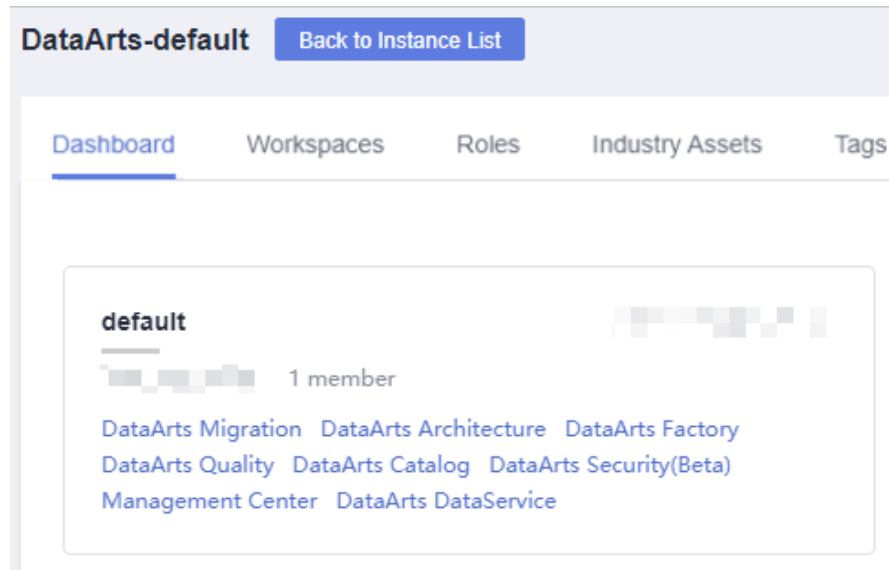
### Prerequisites

- A corresponding cloud service has been enabled and a database has been created in the cloud service. The Flink SQL script does not involve this operation.
- A data connection that matches the data connection type of the created script. For details, see [Creating Data Connections](#). The Flink SQL script does not involve this operation.
- You have locked the script. Otherwise, you must click **Lock** so that you can develop the script. A script you create or import is locked by you by default. For details, see the [lock function](#).

### Procedure

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.



**Figure 3-121** DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Data Development > Develop Script**.
3. In the script directory, double-click a script to access the script development page.
4. In the upper part of the editor, select script properties. [Table 3-118](#) describes the script properties. Skip this step when creating a Flink SQL script.

**Table 3-118** SQL script properties

Property	Description
Data Connection	Selects a data connection.
Database	Name of the database.

Property	Description
Resource Queue	<p>Selects a resource queue for executing a DLI job. Set this parameter when a DLI or SQL script is created.</p> <p>You can create a resource queue using either of the following methods:</p> <ul style="list-style-type: none"> <li>• Click . The <b>Queue Management</b> page of DLI is displayed.</li> <li>• Go to the DLI console.</li> </ul> <p><b>NOTE</b> DLI provides the default resource queue <b>default</b>, which does not support insert, load, or cat commands.</p> <p>To set properties for submitting SQL jobs in the form of <b>key/value</b>, click . A maximum of 10 properties can be set. The properties are described as follows:</p> <ul style="list-style-type: none"> <li>• <b>dli.sql.autoBroadcastJoinThreshold</b>: specifies the data volume threshold to use BroadcastJoin. If the data volume exceeds the threshold, BroadcastJoin will be automatically enabled.</li> <li>• <b>dli.sql.shuffle.partitions</b>: specifies the number of partitions during shuffling.</li> <li>• <b>dli.sql.cbo.enabled</b>: specifies whether to enable the CBO optimization policy.</li> <li>• <b>dli.sql.cbo.joinReorder.enabled</b>: specifies whether join reordering is allowed when CBO optimization is enabled.</li> <li>• <b>dli.sql.multiLevelDir.enabled</b>: specifies whether to query the content in subdirectories if there are subdirectories in the specified directory of an OBS table or in the partition directory of an OBS partition table. By default, the content in subdirectories is not queried.</li> <li>• <b>dli.sql.dynamicPartitionOverwrite.enabled</b>: specifies that only partitions used during data query are overwritten and other partitions are not deleted.</li> </ul>

5. Enter an SQL statement in the editor. You can enter multiple SQL statements.

 **NOTE**

- Note that the system date obtained by using an SQL statement is different from that obtained by using the database tool. The query result is stored in the database in the YYYY-MM-DD format, but the query result displayed on the page is in the converted format.
- SQL statements are separated by semicolons (;). If semicolons are used in other places but not used to separate SQL statements, escape them with backslashes (\).  
For example:  

```
select 1;  
select * from a where b="dsfa\";
```

To facilitate script development, DataArts Factory provides the following capabilities:

- The script editor supports the following shortcut keys, which improve the script development efficiency:
  - **Ctrl + /**: Comment out or uncomment the line or code block at the cursor.
  - **Ctrl + S**: Save
  - **Ctrl + Z**: Cancel
  - **Ctrl + Y**: Redo
  - **Ctrl + F**: Search
  - **Ctrl + Shift + R**: Replace
  - **Ctrl + X**: Cut (cut a line when the cursor selects nothing).
  - **Alt** + mouse dragging: Select columns to edit a block.
  - **Ctrl** + mouse click: Select multiple lines to edit or indent them together.
  - **Shift + Ctrl + K**: Delete the current line.
  - **Ctrl** + **→** (or **←**): Move the cursor rightwards (or leftwards) by word.
  - **Ctrl + Home** or **Ctrl + End**: Navigate to the beginning or end of the current file.
  - **Home** or **End**: Navigate to the beginning or end of the current line.
  - **Ctrl + Shift + L**: Double-click all the same character strings and add cursors to them to implement batch modification.
- System functions (Flink SQL, Spark SQL, ClickHouse SQL, and Presto SQL do not support system functions.)

To view the functions supported by this type of data connection, click **System Function** on the right of the editor. You can double-click a function to the editor to use it.
- Data tables can be read to generate SQL statements. (Flink SQL, Spark SQL, ClickHouse SQL, and Presto SQL do not support this function.)

Click **Data Tables** on the right of the editor to display all the tables in the current database or schema. You can select tables and columns and click **Generate SQL Statement** in the lower right corner to generate an SQL statement, which you need to manually format.
- Script parameters (Currently, only Flink SQL does not support script parameters.)

You can directly write script parameters in SQL statements. When debugging scripts, you can enter parameter values in the script editor. If the script is referenced by a job, you can set parameter values on the job development page. The parameter values can use EL expressions (see [Expression Overview](#)).

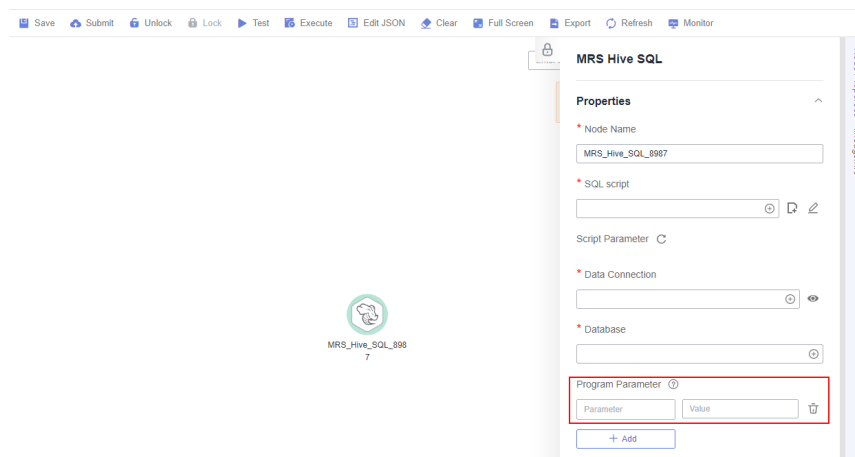
In the following script example, *str1* indicates the parameter name. It can contain only letters, numbers, hyphens (-), underscores (\_), greater-than

signs (>), and less-than signs (<), and can contain a maximum of 16 characters. The parameter name must be unique.

```
select ${str1} from data;
```

For MRS Spark SQL and MRS Hive SQL scripts, you set a program parameter by referring to **set hive.exec.parallel=true;** in the SQL statements or configure this parameter by setting **Program Parameter** on **Node Properties** of the job.

**Figure 3-122** Program Parameter




– Owner

Click **Basic Info** to set the script owner and description.

- (Optional) In the upper part of the editor, click **Format** to format the SQL statement. When developing a Flink SQL script, skip this step.
- In the upper part of the editor, click **Execute**. If you need to execute some SQL statements separately, select the SQL statements first. After executing the SQL statement, view the execution history and result of the script in the lower part of the editor. When developing a Flink SQL script, skip this step.

#### NOTE

- You can perform the following operations on execution results:
    - Double-click or right-click the name of an execution result tab to rename it. The name can contain a maximum of 16 characters.
    - Right-click the name of an execution result tab to close the current tab, all the tabs to the left or right of the current tab, all the other tabs, or all the tabs.
    - If the MRS cluster is a non-security cluster and the command whitelist is not restricted, you can easily find the corresponding task on the Yarn management page of MRS based on the script name and execution time after adding the application name information during Hive SQL execution. Note that if the default engine is **tez**, you need to set the engine to **mr** to disable the tez engine.
- Above the editor, click  to save the script.  
If the script is created but not saved, set the parameters listed in [Table 3-119](#).

**Table 3-119** Script parameters

Parameter	Mandatory	Description
Script Name	Yes	Name of the script. The name contains a maximum of 128 characters, including only letters, numbers, hyphens (-), underscores (_), and periods (.).
Owners	No	Owner of the script. By default, the creator of the script is the owner.
Description	No	Descriptive information about the script.
Select Directory	Yes	Directory to which the script belongs. The root directory is selected by default.

 **NOTE**

If you open an unsaved script, you can restore its content from the local cache.

## Downloading or Dumping a Script Execution Result

**Constraints:** This function is available only when the OBS service is available.

After the script is executed successfully, you can download or dump the execution result. Only users with the **DAYU Administrator** or **Tenant Administrator** policy can download or dump execution results..

- Download result: Download the CSV result files to the local host.
- Dump result: Dump the CSV result files to OBS. For details, see [Table 3-120](#).

 **NOTE**

The execution results of Flink SQL scripts, RDS SQL scripts, and shell scripts cannot be dumped.

**Table 3-120** Parameters for dumping results

Parameter	Mandatory	Description
Data Format	Yes	Format of the data to be exported. Only CSV result files can be exported.
Resource Queue	No	DLI queue where the export operation is to be performed. Set this parameter when a DLI or SQL script is created.

Parameter	Mandatory	Description
Compression Format	No	Format of compression. Set this parameter when a DLI or SQL script is created. <ul style="list-style-type: none"><li>• none</li><li>• bzip2</li><li>• deflate</li><li>• gzip</li></ul>
Storage Path	Yes	OBS path where the result file is stored. After selecting an OBS path, customize a folder. Then, the system will create it automatically for storing the result file.
Cover Type	No	If a folder that has the same name as your custom folder exists in the storage path, select a cover type. Set this parameter when a DLI or SQL script is created. <ul style="list-style-type: none"><li>• <b>Overwrite:</b> The existing folder will be overwritten by the customized folder.</li><li>• <b>Report:</b> The system reports an error and suspends the export operation.</li></ul>

### 3.4.3.3.2 Developing a Shell Script

You can develop, debug, and run shell scripts online. The developed scripts can be run in jobs. For details, see [Developing a Job](#).

#### Prerequisites

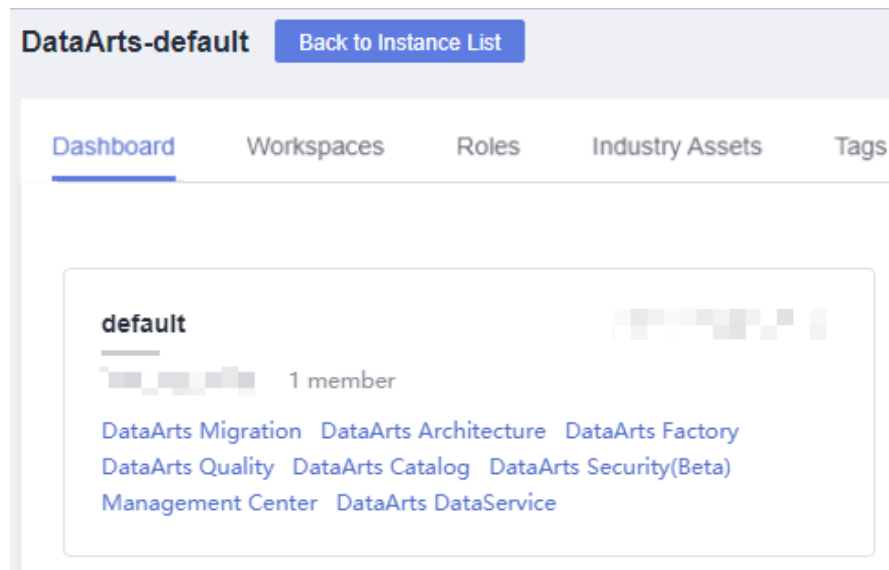
- A shell script has been added. For details, see [Creating a Script](#).
- A host connection has been created. The host is used to execute shell scripts. For details, see [Table 3-11](#).
- You have locked the script. Otherwise, you must click **Lock** so that you can develop the script. A script you create or import is locked by you by default. For details, see the [lock function](#).

#### Procedure

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.



**Figure 3-123** DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Data Development > Develop Script**.
3. In the script directory list, double-click a script that you want to develop. The script development page is displayed.
4. In the upper part of the editor, select script properties. [Table 3-121](#) describes the script properties.

**Table 3-121** Shell script properties

Parameter	Description	Example
Host Connection	Selects the host where a shell script is to be executed.	N/A

Parameter	Description	Example
Parameter	<p>Parameter transferred to the Shell script when it is executed. Parameters are separated by spaces, for example, <b>a b c</b>.</p> <p>The parameter must be referenced by a location variable (for example, \$1, \$2, or \$3) in the Shell script. Otherwise, the parameter is invalid. The location variable starts from 0. Variable 0 is reserved for storing the actual script name, variable 1 corresponds to the first parameter of the script, and so on. For example, \$1, \$2, and \$3 reference parameters <b>a</b>, <b>b</b>, and <b>c</b>, respectively.</p> <p>Note: If a variable is referenced in the shell script, use the <i>\$args</i> format instead of the <i>#{args}</i> format. Otherwise, the variable will be replaced by a parameter with the same name in the job.</p>	<p>For example, if you enter <b>a b c</b> and run the following Shell script, <b>b</b> is displayed:</p> <pre>echo \$2</pre>

Parameter	Description	Example
Interactive Input	Interactive information (for example, passwords) provided during shell script execution.	<p>For example, run the following interactive Shell script. Interaction parameters <b>1</b>, <b>2</b>, and <b>3</b> correspond to <b>begin</b>, <b>end</b>, and <b>exit</b>, respectively.</p> <ul style="list-style-type: none"> <li>• When the interaction parameter is set to <b>1</b>, the execution result is <b>start something</b>.</li> <li>• When the interaction parameter is set to <b>2</b>, the execution result is <b>stop something</b>.</li> <li>• When the interaction parameter is set to <b>3</b>, the execution result is <b>exit</b>.</li> </ul> <pre>#!/bin/bash select Actions in "begin" "end" "exit" do case \$Actions in "begin") echo "start something" break ;; "end") echo "stop something" break ;; "exit") echo "exit" break ;; *) echo "Ignorant" ;; esac done</pre>

5. Edit shell statements in the editor. To facilitate script development, DataArts Factory provides the following capabilities:
  - The script editor supports the following shortcut keys, which improve the script development efficiency:
    - **Ctrl + /**: Comment out or uncomment the line or code block at the cursor.
    - **Ctrl + S**: Save
    - **Ctrl + Z**: Cancel
    - **Ctrl + Y**: Redo

- **Ctrl + F:** Search
  - **Ctrl + Shift + R:** Replace
  - **Ctrl + X:** Cut (cut a line when the cursor selects nothing).
  - **Alt + mouse dragging:** Select columns to edit a block.
  - **Ctrl + mouse click:** Select multiple lines to edit or indent them together.
  - **Shift + Ctrl + K:** Delete the current line.
  - **Ctrl + → (or ←):** Move the cursor rightwards (or leftwards) by word.
  - **Ctrl + Home** or **Ctrl + End:** Navigate to the beginning or end of the current file.
  - **Home** or **End:** Navigate to the beginning or end of the current line.
  - **Ctrl + Shift + L:** Double-click all the same character strings and add cursors to them to implement batch modification.
  - Script parameter function. Use this function in either of the following ways:
    - i. Write the script parameter name and parameter value in the shell statement. When the shell script is referenced by a job, if the parameter name configured for the job is the same as the parameter name of the shell script, the parameter value of the shell script is replaced by the parameter value of the job.

An example is as follows:


```
a=1  
echo ${a}
```

In the preceding command, *a* indicates the parameter name. It can contain only letters, digits, hyphens (-), underscores (\_), greater-than signs (>), and less-than signs (<), and can contain a maximum of 16 characters. The parameter name must be unique.
    - ii. Configure parameters in the upper part of the editor. When you execute the shell script, the configured parameters are transferred to the script. Separate parameters by spaces, for example, **a b c**. The parameter must be referenced by the shell script. Otherwise, the parameter is invalid.

Note: If a variable is referenced in the shell script, use the *\$args* format instead of the *\${args}* format. Otherwise, the variable will be replaced by a parameter with the same name in the job.
  - Owner  
Click **Basic Info** to set the script owner and description.
6. In the lower part of the editor, click **Execute**. After executing the shell statement, view the execution history and result of the script in the lower part of the editor.

 **NOTE**

You can perform the following operations on execution results:

- Double-click or right-click the name of an execution result tab to rename it. The name can contain a maximum of 16 characters.
  - Right-click the name of an execution result tab to close the current tab, all the tabs to the left or right of the current tab, all the other tabs, or all the tabs.
7. Above the editor, click  to save the script.

If the script is created but not saved, set the parameters listed in [Table 3-122](#).

**Table 3-122** Script parameters

Parameter	Mandatory	Description
Script Name	Yes	Name of the script. It contains a maximum of 128 characters. Only letters, digits, hyphens (-), underscores (_), and periods (.) are allowed.
Description	No	Descriptive information about the script.
Select Directory	Yes	Directory to which the script belongs. The root directory is selected by default.

### 3.4.3.3.3 Developing a Python Script

You can develop, debug, and run Python scripts online. The developed scripts can be run in jobs. For details, see [Developing a Job](#).

#### Prerequisites

- A Python script has been added. For details, see [Creating a Script](#).
- A host connection has been created. The host is used to execute Python scripts. For details about how to create a host connection, see [Table 3-11](#).
- You have locked the script. Otherwise, you must click **Lock** so that you can develop the script. A script you create or import is locked by you by default. For details, see the [lock function](#).

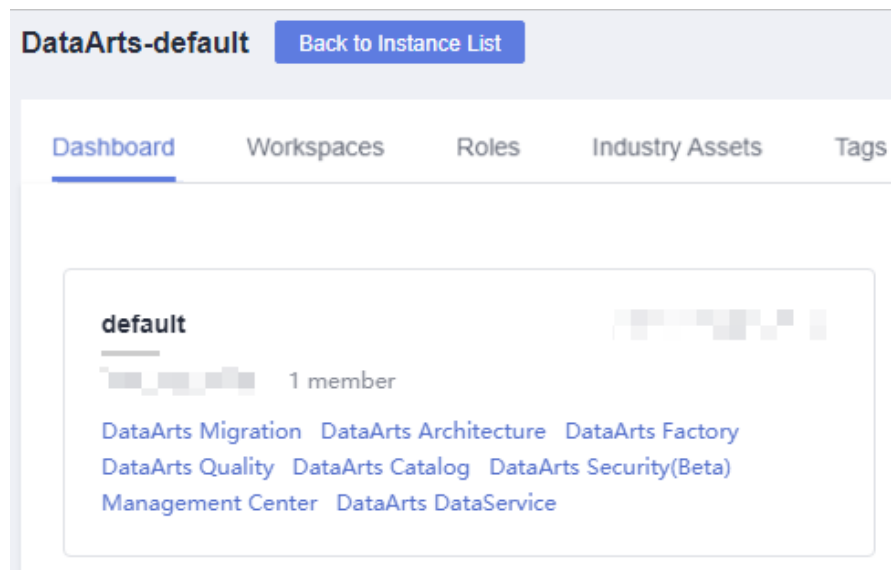
#### Constraints

Python scripts do not support script parameters or job parameters.

#### Procedure

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-124 DataArts Factory




2. In the left navigation pane of DataArts Factory, choose **Data Development > Develop Script**.
3. In the script directory list, double-click a script that you want to develop. The script development page is displayed.
4. In the upper part of the editor, configure the host connection for executing the Python script.
5. Edit Python statements in the editor. To facilitate script development, DataArts Factory provides the following capabilities:
  - The script editor supports the following shortcut keys, which improve the script development efficiency:
    - **Ctrl + /**: Comment out or uncomment the line or code block at the cursor.
    - **Ctrl + S**: Save
    - **Ctrl + Z**: Cancel
    - **Ctrl + Y**: Redo
    - **Ctrl + F**: Search
    - **Ctrl + Shift + R**: Replace
    - **Ctrl + X**: Cut (cut a line when the cursor selects nothing).
    - **Alt + mouse dragging**: Select columns to edit a block.
    - **Ctrl + mouse click**: Select multiple lines to edit or indent them together.
    - **Shift + Ctrl + K**: Delete the current line.
    - **Ctrl + → (or ←)**: Move the cursor rightwards (or leftwards) by word.

- **Ctrl + Home** or **Ctrl + End**: Navigate to the beginning or end of the current file.
  - **Home** or **End**: Navigate to the beginning or end of the current line.
  - **Ctrl + Shift + L**: Double-click all the same character strings and add cursors to them to implement batch modification.
- Owner  
Click **Basic Info** to set the script owner and description.
6. In the upper part of the editor, click **Execute**. After executing the Python statement, view the execution history and result of the script in the lower part of the editor.

 **NOTE**

You can perform the following operations on execution results:

- Double-click or right-click the name of an execution result tab to rename it. The name can contain a maximum of 16 characters.
  - Right-click the name of an execution result tab to close the current tab, all the tabs to the left or right of the current tab, all the other tabs, or all the tabs.
7. Above the editor, click  to save the script.  
If the script is created but not saved, set the parameters listed in [Table 3-123](#).

**Table 3-123** Script parameters

Parameter	Mandatory	Description
Script Name	Yes	Name of the script. It contains a maximum of 128 characters. Only letters, digits, hyphens (-), underscores (_), and periods (.) are allowed.
Description	No	Descriptive information about the script.
Select Directory	Yes	Directory to which the script belongs. The root directory is selected by default.

### 3.4.3.4 Submitting a Version and Unlocking the Script

This involves the version management and lock functions.

- Version management: traces script and job changes, and supports version comparison and rollback. The system retains 10 latest version records. In addition, version management can be used to distinguish the development state and production state.
  - Development state: Scripts or jobs have not been submitted and are used for debugging. In the development state, you can edit, save, and run scripts or jobs without affecting those being scheduled. In addition, when a job is being associated with a script or job dependency is being

- configured, the associated script or job will read the configuration in the development state.
- Production state: Script or jobs have been submitted and are used for formal scheduling. In formal scheduling, the latest submitted versions of scripts or jobs will be used in scenarios such as script invocation, instance rerunning, and job dependency and patch data configuration.
  - Lock: prevents conflict caused by collaborative script or job development. If you create or import a script or job, it is locked by you by default. You can only edit, save, or submit a script or job you have locked. To edit, save, or submit a script or job that is locked by another user or not locked by any user, you must lock the script or job first.

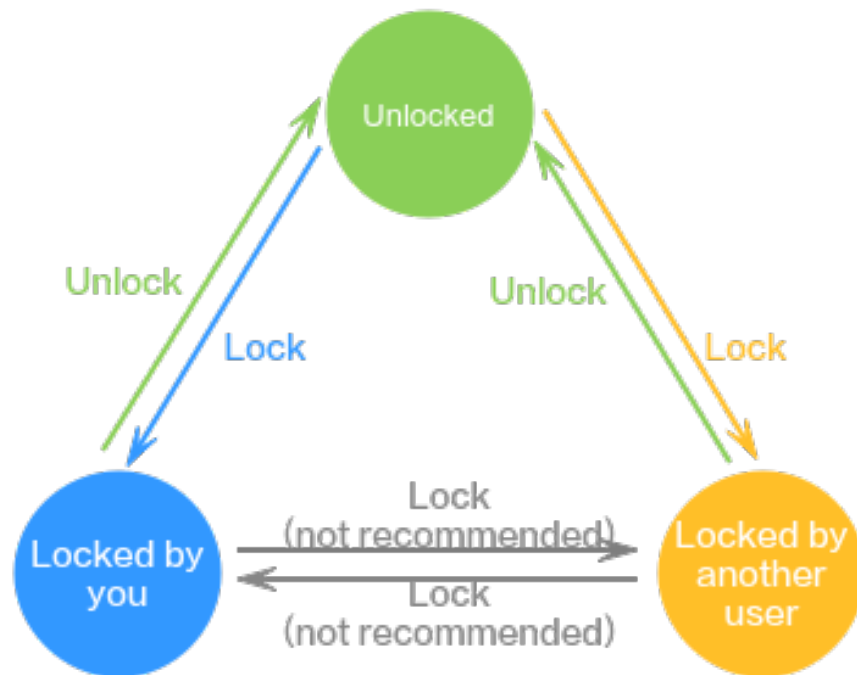
---

#### NOTICE

- You can view the lock status of a script or job in the script or job directory tree.
  - To view the latest version of an opened script or job locked by another user, you need to re-open the script or job because it is not updated in real time.
  - Scripts or jobs that were created before the lock function was available are now unlocked by default. To edit, save, or submit these scripts or jobs, you need to lock them first.
  - The locking operation depends on the soft and hard lock policies. For details about how to configure soft and hard lock policies, see [Configuring a Default Item](#).
    - Soft lock: You can lock or unlock jobs or scripts, regardless of whether they are locked by others.
    - **Hard Lock:** You can lock jobs or scripts only after they have been unlocked by other users. The space administrator and the **DAYU Administrator** user can lock and unlock jobs or scripts without any limitations.
  - Do not lock a script or job that is locked by another user because if you do so, changes to the script or job made by the user will be lost. If you want to modify the script or job, contact the user to unlock the script or job, and lock it by yourself.
-



Figure 3-125 Lock status



## Prerequisites

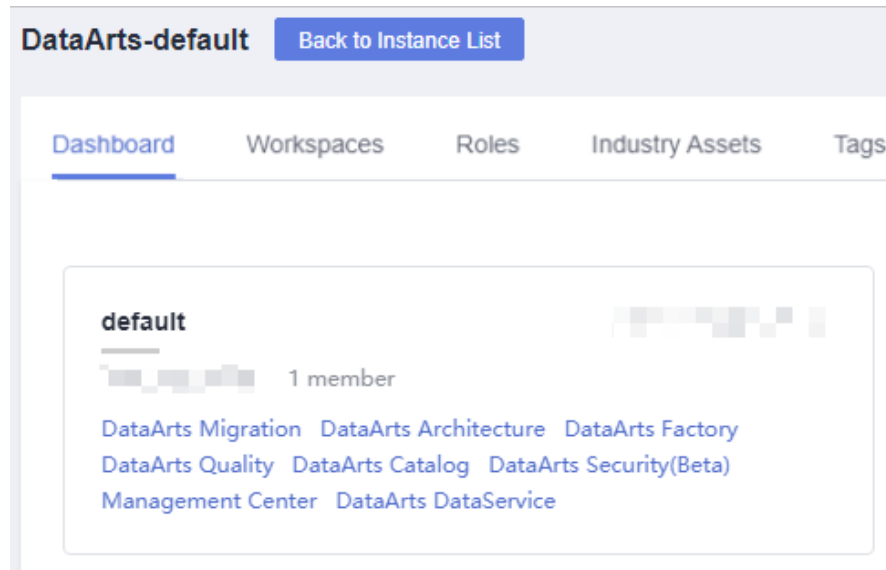
A script has been developed.

## Submitting a Version and Unlocking the Script

If you submit a version, the latest script in the development state will be saved and submitted and overwrite the previous script version. You are advised to unlock the script after submitting the version so that other developers can modify the script as needed.

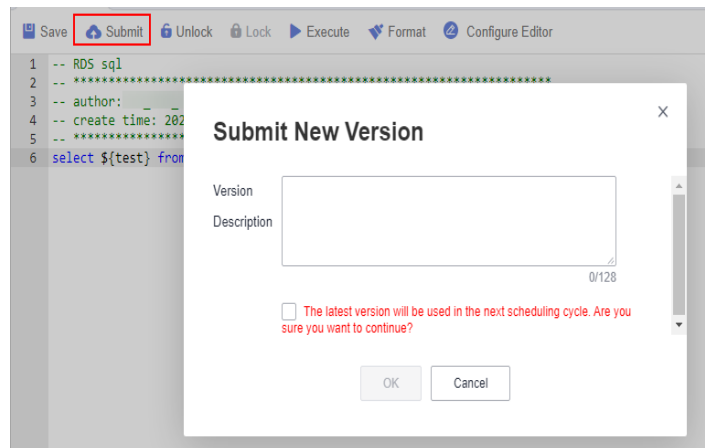
- Step 1** Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-126 DataArts Factory



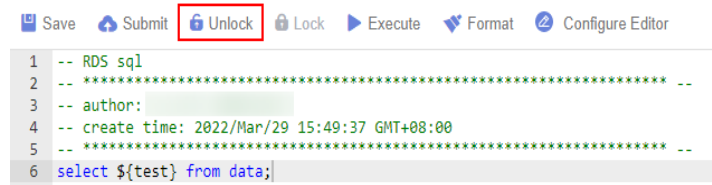
- Step 2** In the left navigation pane of DataArts Factory, choose **Data Development > Develop Script**.
- Step 3** In the script directory, double-click the developed script to access the script development page.
- Step 4** In the upper part of the script editor, click **Submit**. In the displayed dialog box, enter the change description (a maximum of 128 characters allowed) and select the check box below. If you do not select this option, you cannot click **OK**.

Figure 3-127 Submitting a version



- Step 5** In the upper part of the script editor, click **Unlock** to unlock the script.

**Figure 3-128** Unlocking a script



----End

## Version Rollback

After submitting the version, you can view it in the version list. (Currently, a maximum of 10 latest versions are saved.) Click **Roll Back** to roll back to any submitted version.

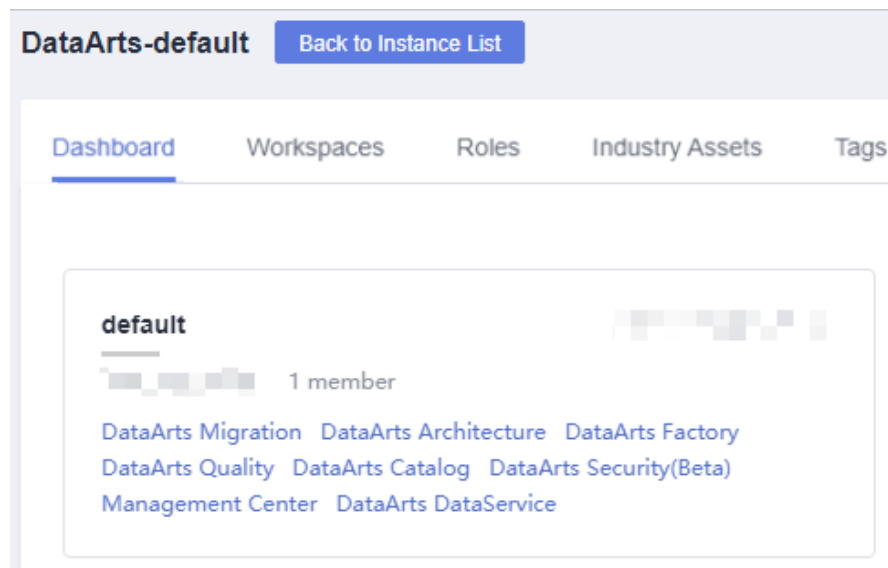
The rollback involves the following contents:

- DLI: data connections, databases, resource queues, and script contents
- DWS: data connections, databases, and script contents
- HIVE: data connections, databases, resource queues, and script contents
- SPARK: data connections, databases, and script contents
- SHELL: host connections, parameters, interactive parameters, and script contents
- RDS: data connections, databases, and script contents
- PRESTO: data connections, modes, and script contents
- PYTHON: host connections, parameters, interactive parameters, and script content
- FLINK: script content

The procedure is as follows:

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-129** DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Data Development > Develop Script**.
3. In the script directory list, double-click a script that you want to develop. The script development page is displayed.
4. On the right of the page, click the **Versions** tab and view the version submission records. Select the version to be rolled back and click **Roll Back**.  
If the content in the development state is not submitted, the content will be overwritten after the rollback. In this case, you must submit the rollback version again to make it take effect. By default, the latest submitted version is used for scheduling.

**Figure 3-130** Rolling back a version



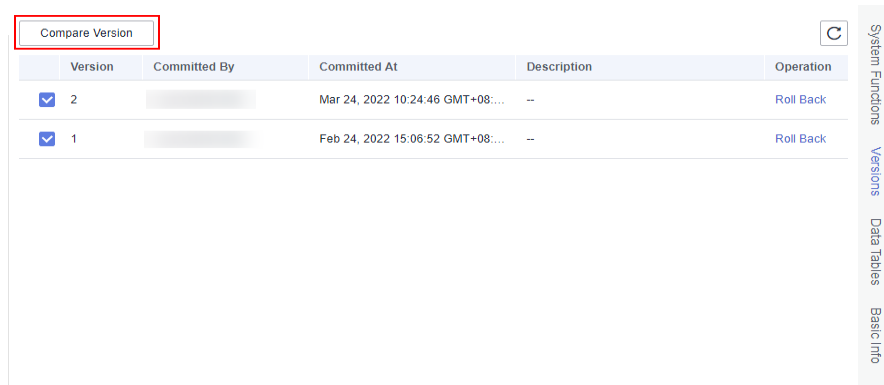
## Version Comparison

You can compare the script contents of two different versions. If you select only one version, the system compares the script content of the selected version with that in the development state. If you select two versions, the system compares the script contents of two different versions.



The procedure is as follows:

1. In the left navigation pane of DataArts Factory, choose **Data Development > Develop Script**.
2. In the script directory list, double-click a script that you want to develop. The script development page is displayed.
3. On the right of the page, click the **Versions** tab and view the version submission records. Select the versions to be compared and click **Compare Version**.

**Figure 3-131** Comparing versions



4. A new page is displayed, showing the script content of different versions on the left and right separately. The differences between the two versions have

been marked. You can use the  and  buttons in the upper right corner to go to the previous or next change.

**Figure 3-132** Version comparison details



### 3.4.3.5 (Optional) Managing Scripts

#### 3.4.3.5.1 Copying a Script

This section describes how to copy a script.

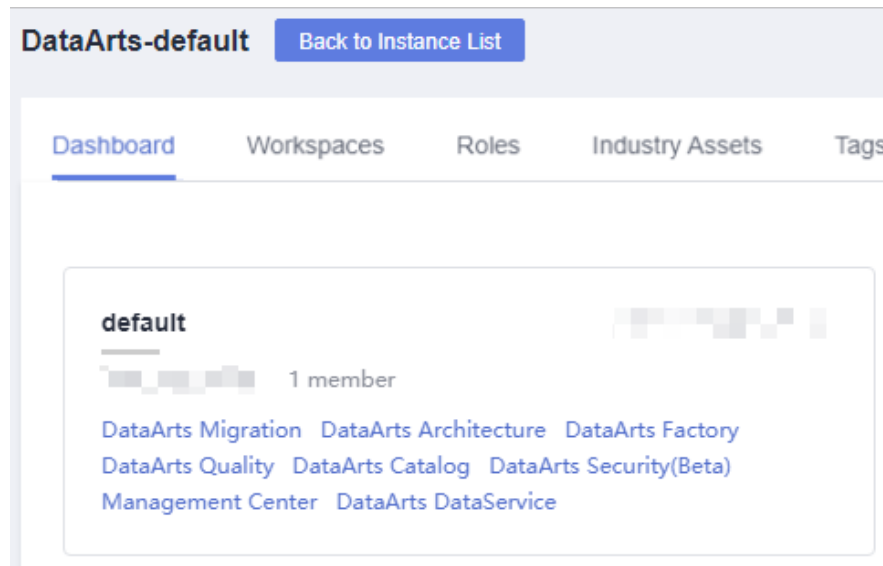
#### Prerequisites

A script has been developed. For details about how to develop scripts, see [Developing Scripts](#).

#### Procedure

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-133** DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Data Development > Develop Script**.
3. In the script directory, select the script to be copied, right-click the script name, and choose **Copy Save As**.
4. In the displayed dialog box, configure related parameters. [Table 3-124](#) describes the parameters.

**Table 3-124** Script directory parameters

Parameter	Description
Script Name	Name of the script. It contains a maximum of 128 characters. Only letters, digits, hyphens (-), underscores (_), and periods (.) are allowed. <b>NOTE</b> The name of the copied script cannot be the same as the name of the original script.
Select Directory	Parent directory of the script directory. The parent directory is the root directory by default.

5. Click **OK**.

### 3.4.3.5.2 Copying the Script Name and Renaming a Script

You can copy the name of a script and rename a script.

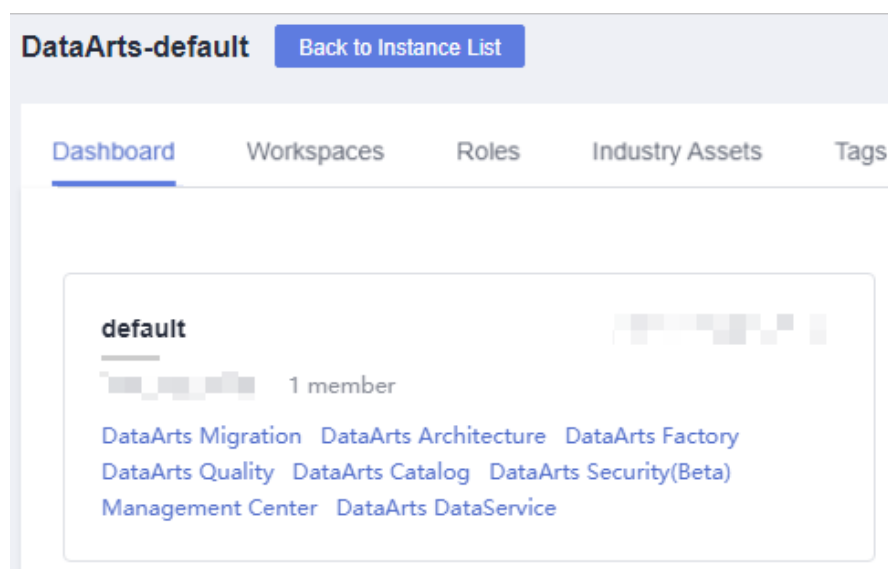
#### Prerequisites

A script has been developed. For details about how to develop scripts, see [Developing Scripts](#).

#### Copying the Script Name

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-134** DataArts Factory

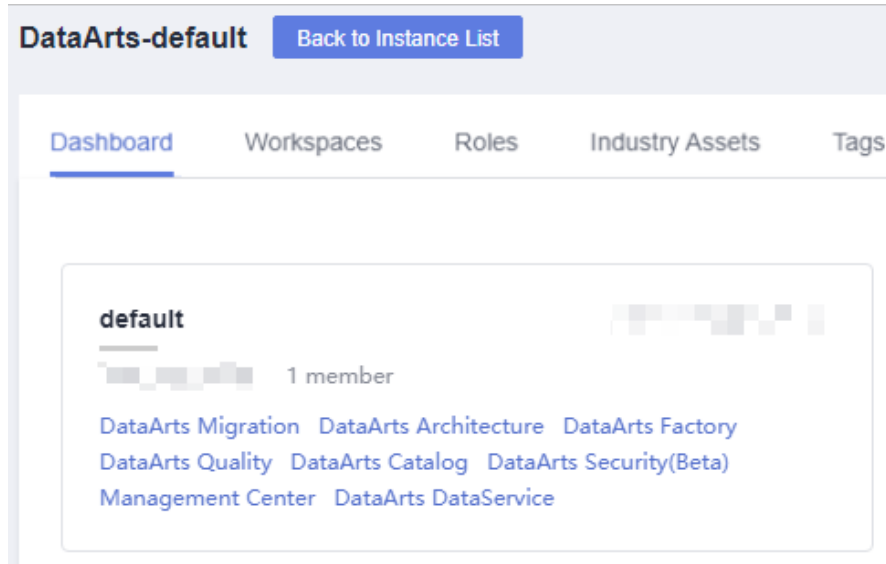


2. In the left navigation pane of DataArts Factory, choose **Data Development > Develop Script**.
3. Locate the target script in the script directory, right-click the script name, and select **Copy Name** to copy the script name to the clipboard.

## Renaming a Script

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-135** DataArts Factory



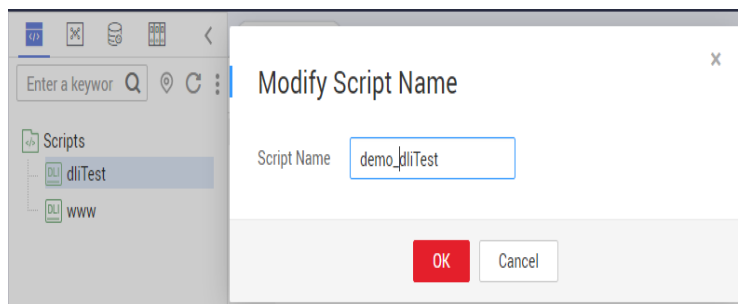
2. In the left navigation pane of DataArts Factory, choose **Data Development > Develop Script**.
3. Locate the target script In the script directory, right-click the script name, and select **Rename**.

**NOTE**

An opened script file cannot be renamed.

4. In the displayed **Modify Script Name** dialog box, change the script name.

**Figure 3-136** Renaming a script



**Table 3-125** Script renaming parameters

Parameter	Description
Script Name	Name of the script. It contains a maximum of 128 characters. Only letters, digits, hyphens (-), underscores (_), and periods (.) are allowed.

5. Click **OK**.

### 3.4.3.5.3 Moving a Script or Script Directory

You can move a script file from one directory to another or move a script directory to another directory.

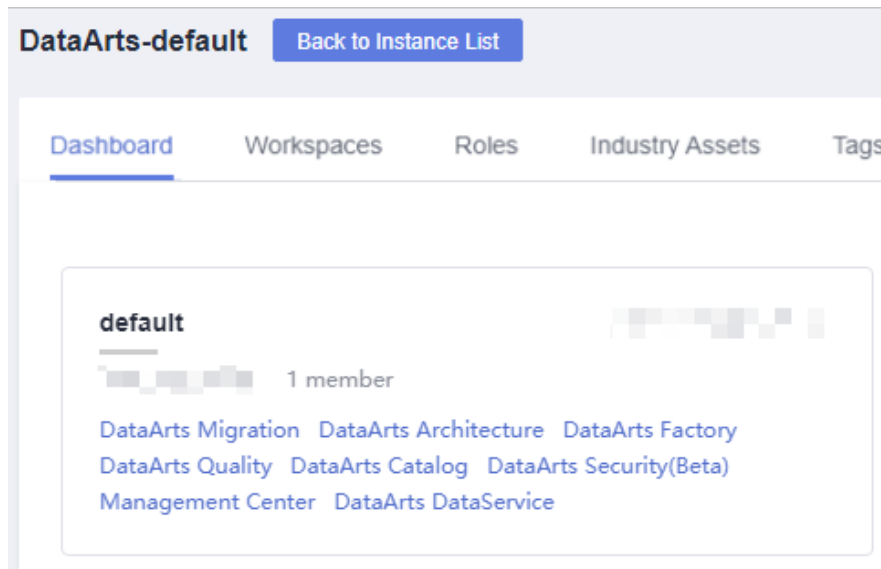
#### Prerequisites

A script has been developed. For details about how to develop scripts, see [Developing Scripts](#).

#### Procedure

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-137** DataArts Factory



2. In the navigation pane of the Data Development homepage, choose **Data Development > Develop Script**.
3. Move a script or script directory.
  - Method 1: right-click**
    - a. In the script directory, right-click a script or script folder and select **Move**.
    - b. In the displayed dialog box, configure related parameters. [Table 3-126](#) describes the parameters.

**Table 3-126** Parameters for moving a script or directory

Parameter	Description
Select Directory	Directory to which the script or script directory is to be moved. The parent directory is the <b>root</b> directory by default.



- c. Click **OK** to move the script or directory.

### Method 2: drag-and-drop

Select a script or script folder and drag and drop it to the target folder.

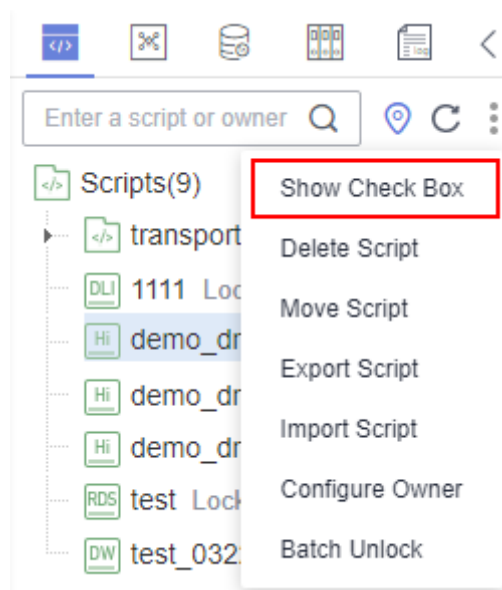
## 3.4.3.5.4 Exporting and Importing a Script


### Exporting a Script

You can export one or more script files from the script directory. The exported files store the latest content in the development state.

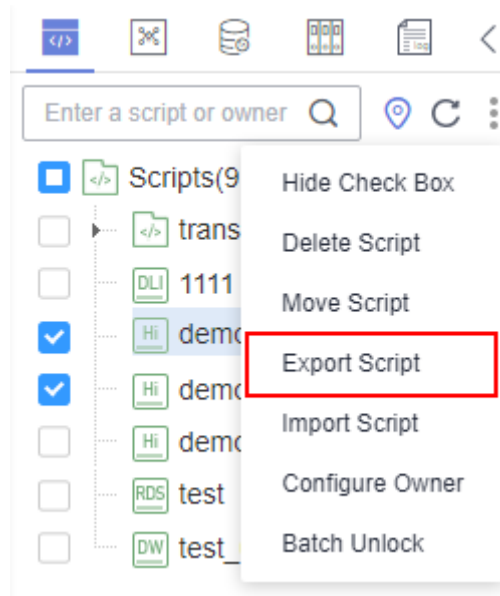
1. Click  in the script directory and select **Show Check Box**.

Figure 3-138 Clicking Show Check Box



2. Select the scripts to be exported, click , and choose **Export Script**. After the export is successful, you can obtain the exported .zip file.


**Figure 3-139** Selecting and exporting scripts



## Importing a Script

This function is available only if the OBS service is available. If OBS is unavailable, scripts can be imported from the local PC.

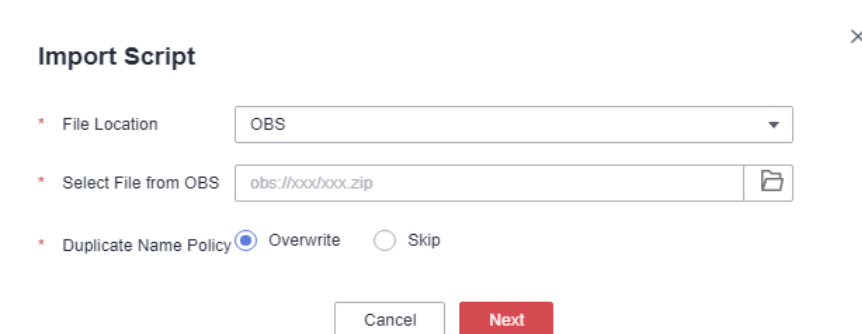
You can import one or more script files in the script directory. After a job is imported, the content in the development state is overwritten and a new version is automatically submitted.

1. Click  and choose **Import Script** in the script directory, select a script file that has been uploaded to OBS, and set **Duplicate Name Policy**.

### NOTE

If you select **Overwrite** for **Duplicate Name Policy** but the hard lock policy is used and the script is locked by another user, the overwriting will fail. For details about soft and hard lock policies, see [Configuring the Hard and Soft Lock Policy](#).

**Figure 3-140** Importing scripts



2. Click **Next**.

### 3.4.3.5.5 Viewing Script References

This section describes how to view the references of a script or all the scripts in a folder.

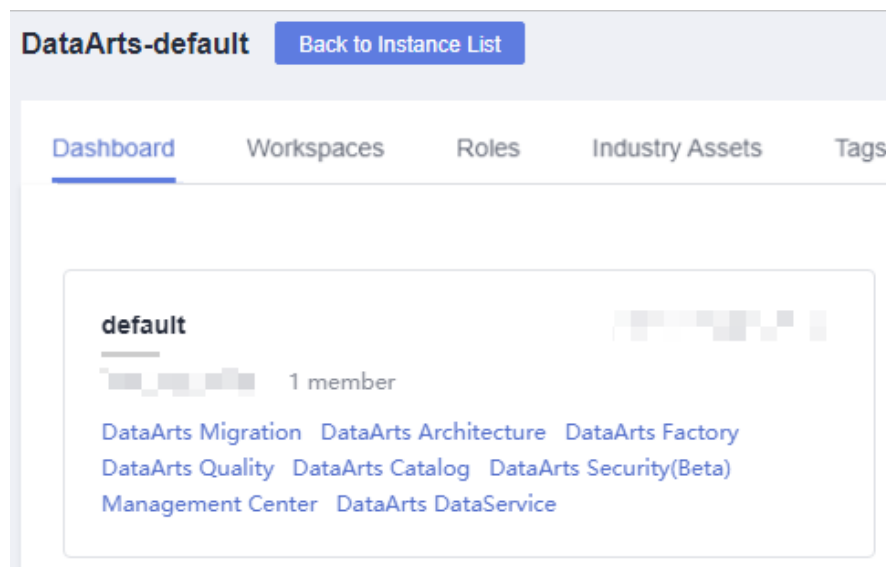
#### Prerequisites

A script has been developed. For details about how to develop scripts, see [Developing Scripts](#).

#### Procedure

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-141** DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Data Development > Develop Script**.
3. To view the references of a script, right-click the script and select **View Reference**.  
To view the references of all the scripts in a folder, right-click the folder and select **View Reference**.
4. In the displayed dialog box, you can view the references of a script or all the scripts in the folder.

### 3.4.3.5.6 Deleting a Script

If you do not need to use a script any more, perform the following operations to delete it.

When you delete a script, the system checks whether the script is being referenced by some jobs. **Version** in the reference list lists the job versions that reference the script. When you click **Delete**, the job and all its version information are deleted.

 NOTE

If a script to be deleted is being associated with a job, ensure that services are not affected after the script is forcibly deleted. If you want to continue to use the job, go to the **Develop Job** page and associate the job with an available script.

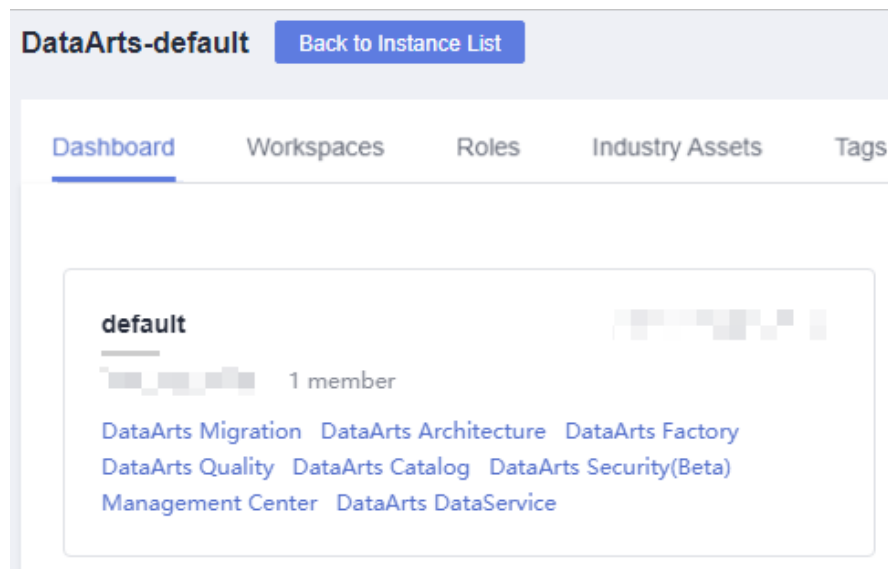
## Prerequisites

The script that you want to delete is not used by any jobs.

## Deleting a Script

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-142 DataArts Factory

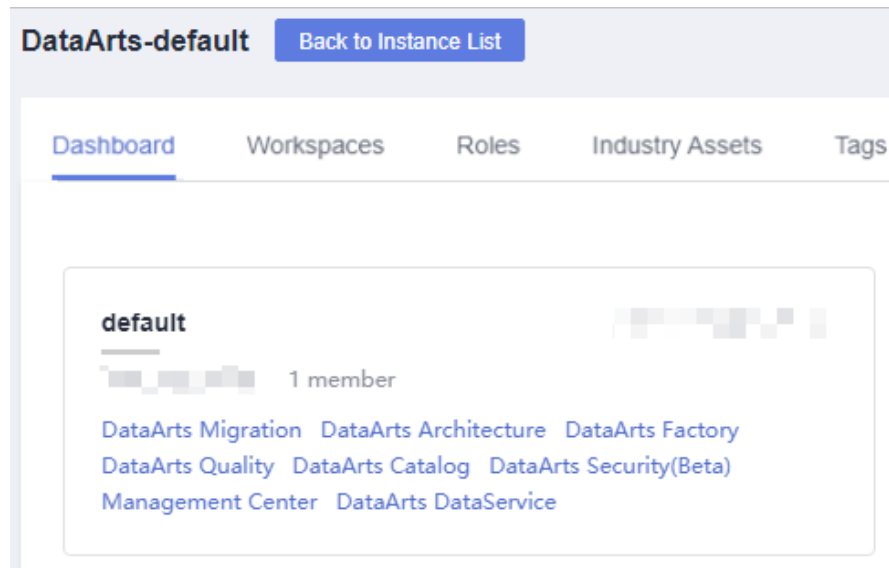




2. In the left navigation pane of DataArts Factory, choose **Data Development > Develop Script**.
3. In the script directory, right-click the script that you want to delete and choose **Delete** from the shortcut menu.
4. In the displayed dialog box, click **OK**.

## Batch Deleting Scripts

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-143 DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Data Development > Develop Script**.
3. On the top of the script directory, click  and select **Show Check Box**.
4. Select the scripts to be deleted, click , and select **Batch Delete**.
5. In the displayed dialog box, click **OK**.

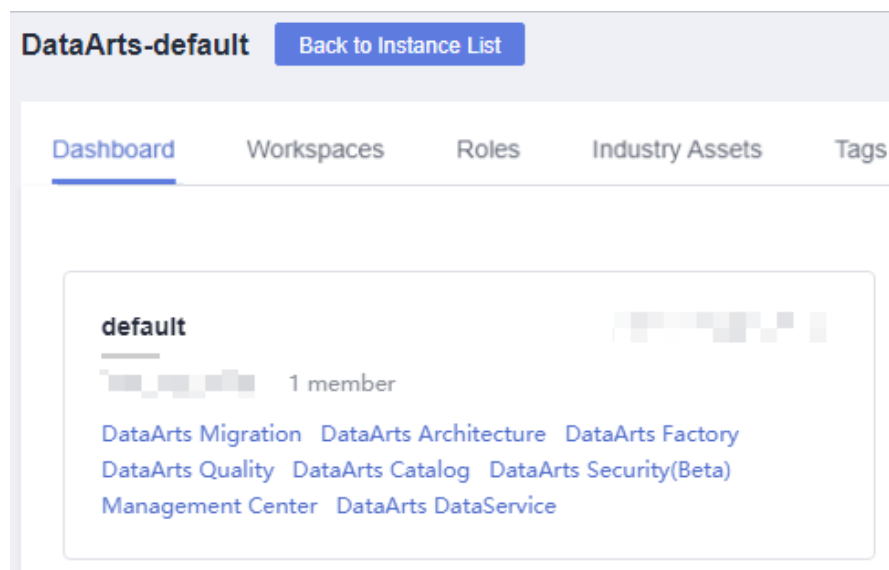
### 3.4.3.5.7 Changing the Script Owner


DataArts Factory allows you to change the owner for scripts with a few clicks.

#### Procedure

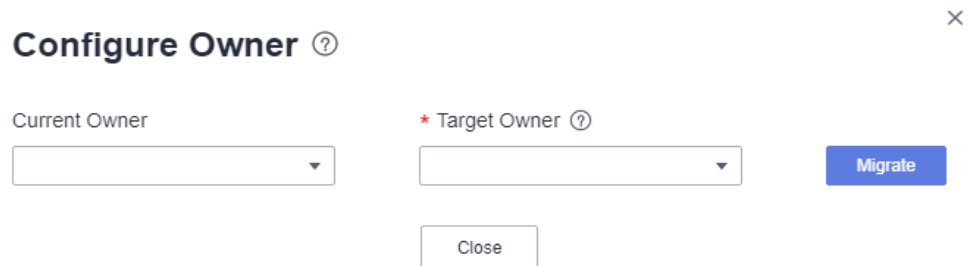
1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-144 DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Data Development > Develop Script**.
3. At the top of the script directory, click  and select **Configure Owner**.

**Figure 3-145** Configuring the owner

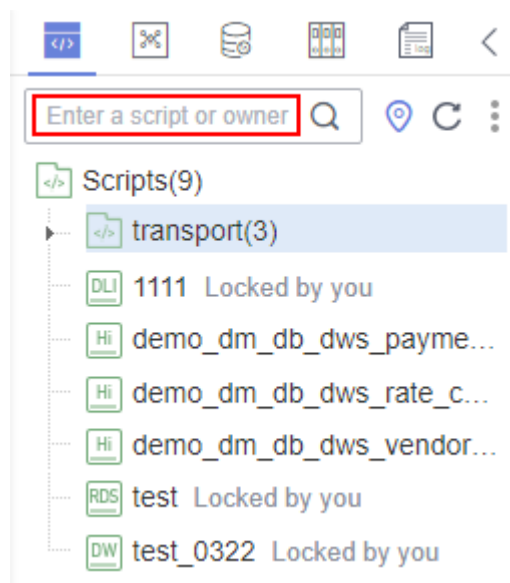


4. Set **Current Owner** and **Target Owner** and click **Migrate**.
5. When the migration succeeds, click **Close**.

## Related Operations

You can use an owner to filter scripts by entering the owner in the search box above the script directory.

**Figure 3-146** Filtering scripts by owner



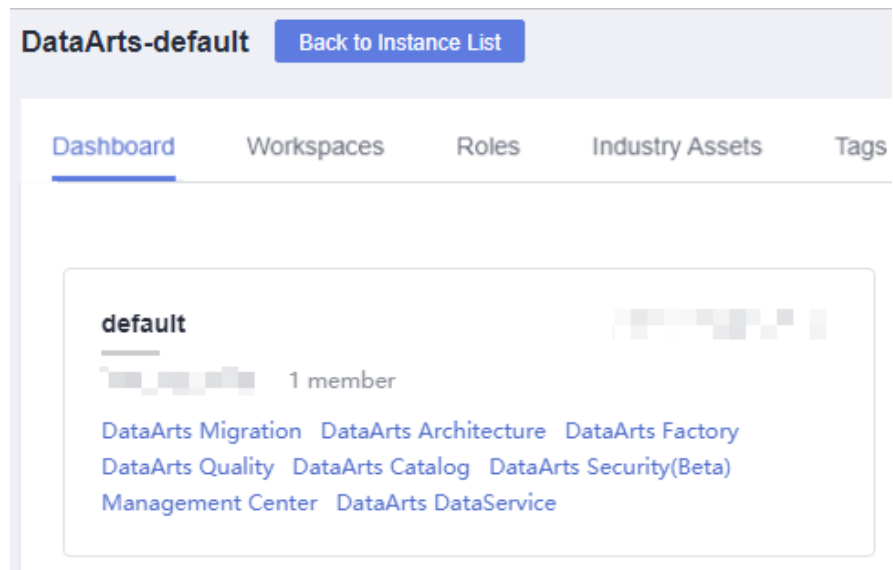
### 3.4.3.5.8 Unlocking Scripts

This section describes how to unlock scripts in batches.

#### Procedure

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-147 DataArts Factory




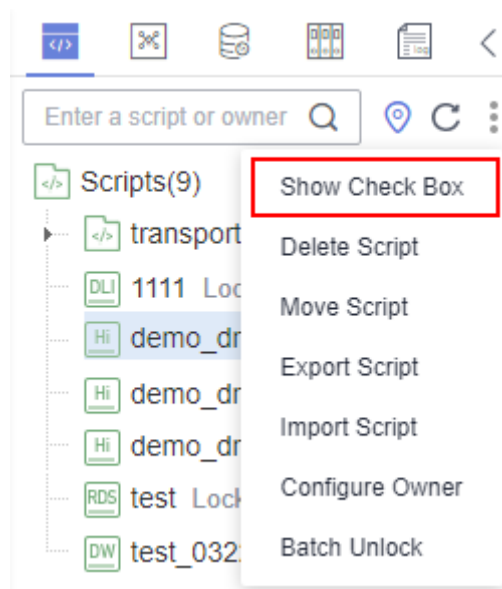
2. In the left navigation pane of DataArts Factory, choose **Data Development > Develop Script**.
3. Click  in the script directory and select **Show Check Box**.

Figure 3-148 Clicking Show Check Box




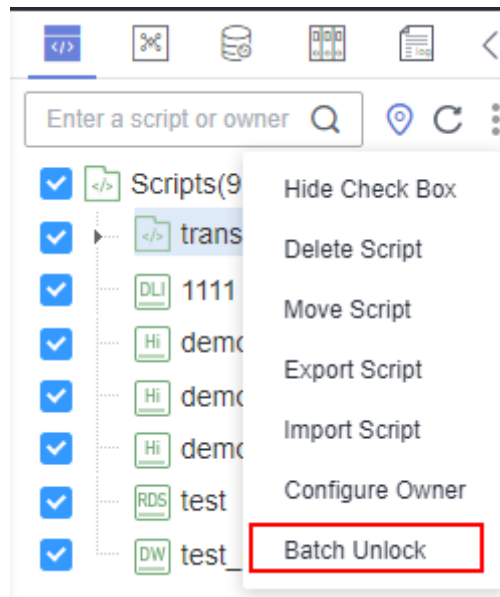
4. Select the scripts to unlock, click , and select **Batch Unlock**. The message "Unlocked." is displayed.

Figure 3-149 Batch Unlock



## 3.4.4 Job Development

### 3.4.4.1 Job Development Process

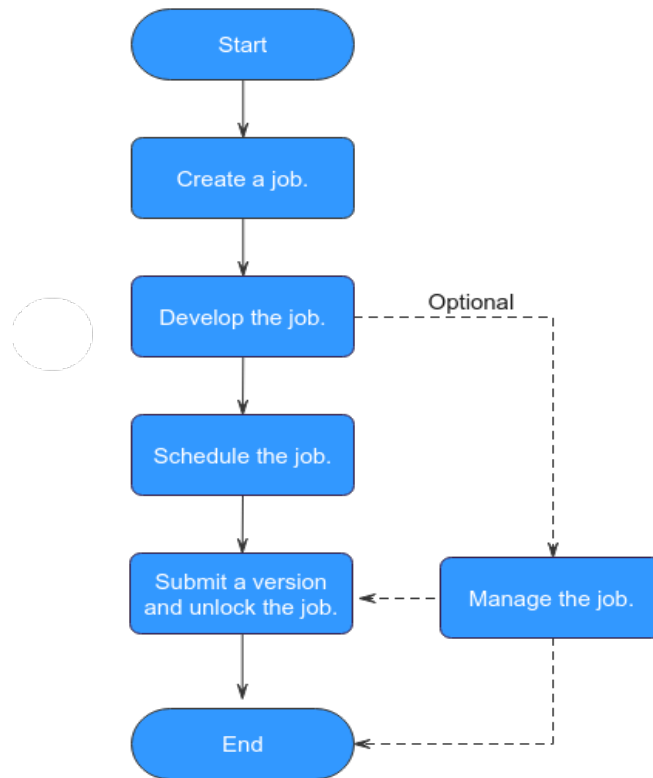
The job development function provides the following capabilities:

- Provides a graphical designer that allows you to quickly build a data processing workflow by drag-and-drop.
- Presets multiple job types, such as data integration, computing and analysis, data monitoring, and resource management, and completes complex data analysis and processing based on dependencies between jobs.
- Supports various scheduling modes.
- Supports job import and export.
- Monitors job status and sends job result notifications.
- Provides editing locks for collaborative development.
- Supports job version management.

Before developing a job, you can learn about the basic job development process from [Figure 3-150](#).



**Figure 3-150** Job development process



1. Create a job: Currently, two job types are available: batch and real-time, which are used for batch data processing and real-time connection data processing, respectively. For details, see [Creating a Job](#).
2. Develop the job: Develop the created job. You can orchestrate and configure nodes. For details, see [Developing a Job](#).
3. Schedule the job: Configure job scheduling tasks. For details, see [Setting Up Scheduling for a Job](#).
  - If the processing mode of a job is batch processing, configure scheduling types for jobs. Three scheduling types are supported: run once, run periodically, and event-based. For details, see [Setting Up Scheduling for a Job Using the Batch Processing Mode](#).
  - If the processing mode of a job is real-time processing, configure scheduling types for nodes. Three scheduling types are supported: run once, run periodically, and event-based. For details, see [Setting Up Scheduling for Nodes of a Job Using the Real-Time Processing Mode](#).
4. Submit a version and unlock the script: After performing this step, the job can be scheduled and modified by other developers. For details, see [Submitting a Version and Unlocking the Script](#).
5. (Optional) Manage the job: After the job development is complete, you can manage the job as required. For details, see [\(Optional\) Managing Jobs](#).

### 3.4.4.2 Creating a Job

A job is composed of one or more nodes that are performed collaboratively to complete data operations. Before developing a job, create a new one.

## Prerequisites

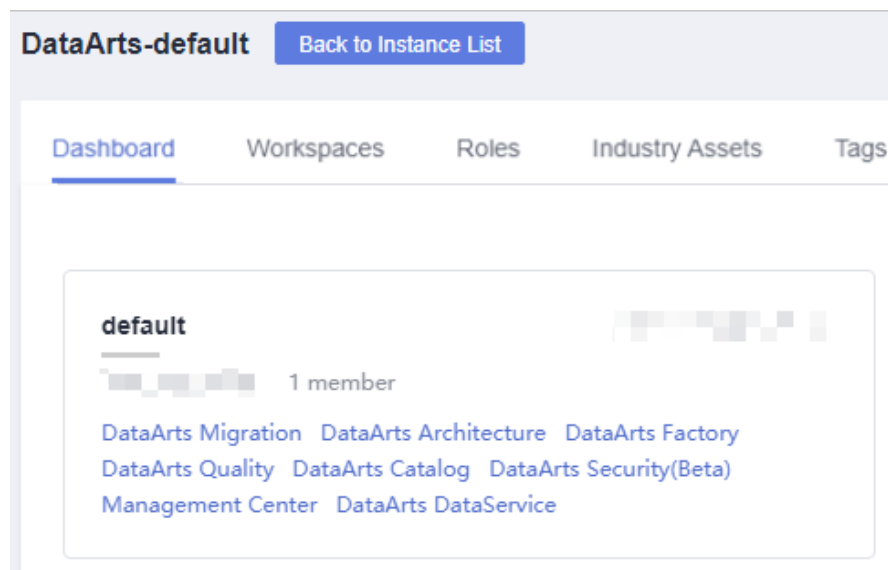
Each workspace can hold a maximum of 10,000 jobs. Ensure that the number of your jobs does not reach this upper limit.

## (Optional) Creating a Directory

If a directory exists, you do not need to create one.

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-151** DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Development > Develop Job**.
3. In the job directory list, right-click a directory and choose **Create Directory** from the shortcut menu.
4. In the **Create Directory** dialog box, configure directory parameters based on [Table 3-127](#).

**Table 3-127** Job directory parameters

Parameter	Description
Directory Name	Name of a job directory. The name must contain 1 to 64 characters, including only letters, numbers, underscores (_), and hyphens (-).
Select Directory	Parent directory of the job directory. The parent directory is the root directory by default.

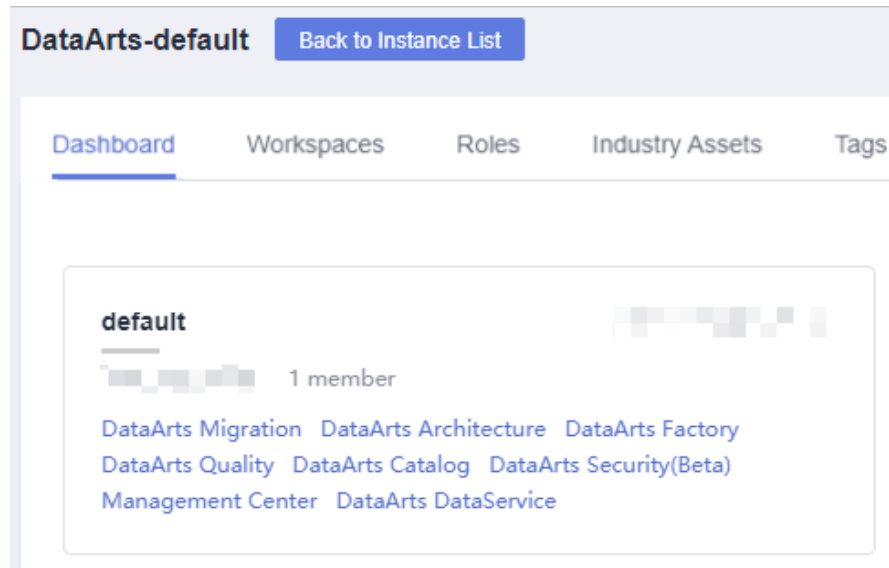
5. Click **OK**.

## Creating a Job

The quantity of jobs is less than the maximum quota (10,000).

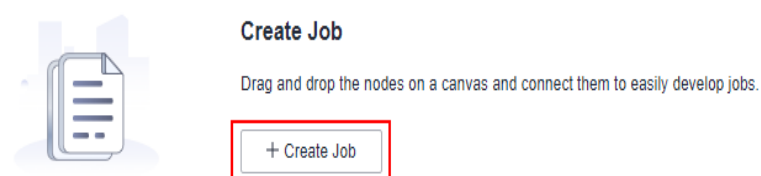
1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-152** DataArts Factory



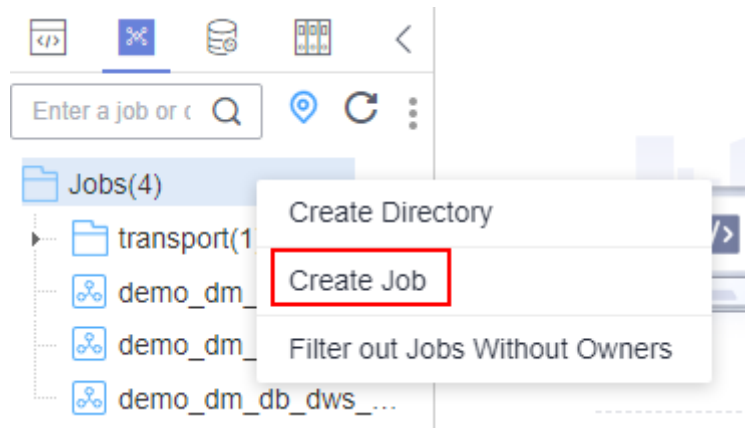
2. In the left navigation pane of DataArts Factory, choose **Development > Develop Job**.
3. Create a job using either of the following methods:  
Method 1: On the **Develop Job** page, click **Create Job**.

**Figure 3-153** Creating a job (method 1)



Method 2: In the directory list, right-click a directory and choose **Create Job** from the shortcut menu.

**Figure 3-154** Creating a job (method 2)



4. In the displayed dialog box, configure job parameters. [Table 3-128](#) describes the job parameters.

**Table 3-128** Job parameters

Parameter	Description
Job Name	Name of the job. The name must contain 1 to 128 characters, including only letters, numbers, hyphens (-), underscores (_), and periods (.).
Processing Mode	<p>Type of the job.</p> <ul style="list-style-type: none"> <li> <b>Batch processing:</b> Data is processed periodically in batches based on the scheduling plan, which is used in scenarios with low real-time requirements. This type of job is a pipeline that consists of one or more nodes and is scheduled as a whole. It cannot run for an unlimited period of time, that is, it must end after running for a certain period of time.                      You can configure job-level scheduling tasks for this type of job. That is, the job is scheduled as a whole. For details, see <a href="#">Setting Up Scheduling for a Job Using the Batch Processing Mode</a>.                 </li> <li> <b>Real-time processing:</b> Data is processed in real time, which is used in scenarios with high real-time performance. This type of job is a business relationship that consists of one or more nodes. You can configure scheduling policies for each nodes, and the tasks started by nodes can keep running for an unlimited period of time. In this type of job, lines with arrows represent only service relationships, rather than task execution processes or data flows.                      You can configure node-level scheduling tasks for this type of job, that is, each node can be independently scheduled. For details, see <a href="#">Setting Up Scheduling for Nodes of a Job Using the Real-Time Processing Mode</a>.                 </li> </ul>

Parameter	Description
Creation Method	Selects a job creation mode. <ul style="list-style-type: none"><li>• Create Empty Job: Create an empty job.</li><li>• Create Based on Template: Create a job using a template.</li></ul>
Select Directory	Directory to which the job belongs. The root directory is selected by default.
Owner	Owner of the job.
Priority	Priority of the job. The value can be <b>High</b> , <b>Medium</b> , or <b>Low</b> .
Agency	After an agency is configured, the job interacts with other services as an agency during job execution. If an agency has been configured for the workspace by referring to <a href="#">Configuring a Workspace-Level Agency</a> , the new job uses the workspace-level agency by default. You can also change the agency to a job-level agency by referring to <a href="#">Configuring a Job-level Agency</a> . <b>NOTE</b> Job-level agency takes precedence over workspace-level agency.
Log Path	Selects the OBS path to save job logs. By default, logs are stored in a bucket named <b>dlf-log-<i>{Projectid}</i></b> . <b>NOTE</b> <ul style="list-style-type: none"><li>• If you want to customize a storage path, select the bucket that you have created on OBS by following the instructions provided in <a href="#">(Optional) Changing the Job Log Storage Path</a>.</li><li>• Ensure that you have the read and write permissions on the OBS path specified by this parameter. Otherwise, the system cannot write logs or display logs.</li></ul>

5. Click **OK**.

### 3.4.4.3 Developing a Job

This section describes how to develop and configure a job.

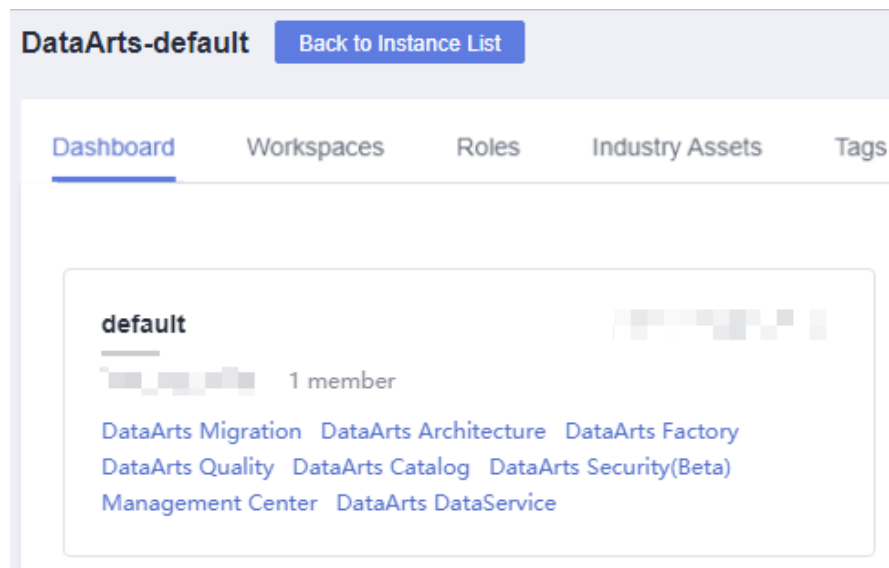
#### Prerequisites


- You have created a job. For details about how to create a job, see [Creating a Job](#).
- You have locked the job. Otherwise, you must click **Lock** so that you can develop the job. A job you create or import is locked by you by default. For details, see the [lock function](#).

#### Compiling Job Nodes

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-155** DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Development > Develop Job**.
3. In the job directory, double-click the name of a batch processing job or real-time processing job in pipeline mode to access the job development page.
4. Drag a desired node to the canvas, move the mouse over the node, and select the  icon and drag it to connect to another node.

 **NOTE**

It is recommended that each job contain a maximum of 200 nodes.

**Figure 3-156** Compiling a job



5. Configure node functions. Right-click a node icon on the canvas and select a function as needed. [Table 3-129](#) lists the available functions.

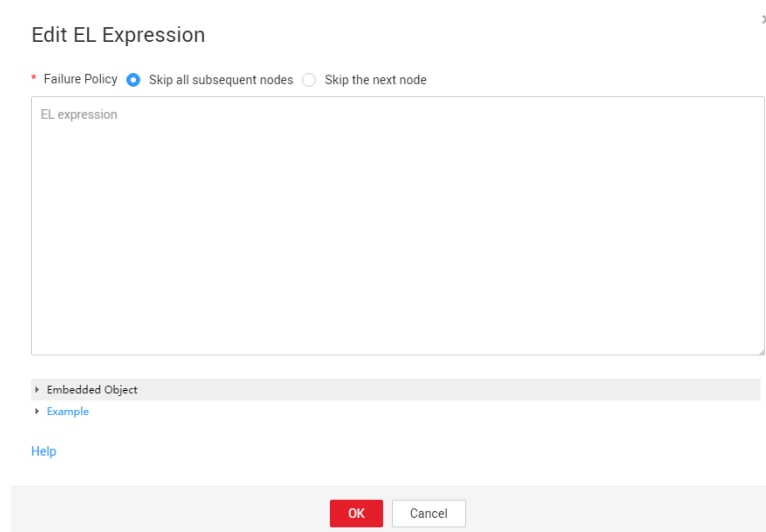
**Table 3-129** Node functions

Function	Description
Configure	Goes to the <b>Node Property</b> page of the node.

Function	Description
Delete	<p>Deletes one or more nodes at the same time.</p> <ul style="list-style-type: none"> <li>Deleting one node: Right-click the node icon in the canvas and choose <b>Delete</b> or press the <b>Delete</b> shortcut key.</li> <li>Deleting multiple nodes: Click the icons of the nodes to be deleted in the canvas while holding on <b>Ctrl</b>, right-click the blank area of the current job canvas, and choose <b>Delete</b> or press the <b>Delete</b> shortcut key.</li> </ul>
Copy	<p>Copies one or more nodes to any job.</p> <ul style="list-style-type: none"> <li>Single-node copy: You can either right-click the node icon in the canvas, choose <b>Copy</b>, and paste the node to a target location, or click the node icon in the canvas and press <b>Ctrl+C</b> and <b>Ctrl+V</b> to paste the node to a target location. The copied node carries the configuration information of the original node.</li> <li>Multi-node copy: Click the icons of the nodes to be copied in the canvas while holding on <b>Ctrl</b>. Then you can either right-click the blank area of the canvas, choose <b>Copy</b>, and paste the nodes to a target location, or press <b>Ctrl+C</b> and <b>Ctrl+V</b> to paste the nodes to a target location. The copied node carries the configuration information of the original node, but does not contain the connection relationship between nodes.</li> </ul>
Test Run	Runs the node for a test.
Test from Current Node	This option is available only for batch processing jobs. It tests the current and subsequent nodes.
Add/Delete Connection	Adds or deletes a connection between two nodes.
Edit CDM Job	This option is available only for CDM jobs. After selecting a CDM cluster and a job, you can go to the CDM job editing page to modify the job.
View Job Log	This option is available only for CDM jobs. When a CDM job is running, you can right-click the CDM job node and select <b>View Job Log</b> from the shortcut menu to go to the job monitoring page and view logs to help developers demarcate and locate job running exceptions.
Edit Script	This option is available only for the node associated with a script. Goes to the script editing page and edits the associated script.
Add Note	Adds a note to the node. Each node can have multiple notes.

6. (Optional) Configure line functions. Right-click the line connecting two nodes on the canvas. **Delete** and **Set Condition** are displayed. You can select them as needed.
  - **Delete:** Deletes the line connecting the nodes.
  - **Set Condition:** In the displayed dialog box, you can enter a ternary expression using the EL expression syntax. If the result of the ternary expression is **true**, subsequent nodes will be connected. Otherwise, subsequent nodes will be skipped.

The following figure shows a typical ternary expression. If the execution result of the DQM node is **true**, subsequent nodes will be connected. If the execution result is **false** and the **Failure Policy** is **Skip all subsequent nodes**, the next node A and all nodes following node A will be skipped.



For details about the EL expression syntax, see [Expression Overview](#).

7. Configure node properties by following the instructions in [Node Overview](#).
8. Configure node properties Click a node in the canvas. On the displayed **Node Properties** page, configure node properties. For details, see [Node Overview](#).

## Configuring Basic Job Information

After you configure the owner and priority for a job, you can search for the job by the owner and priority. The procedure is as follows:

Click the **Basic Info** tab on the right of the canvas to expand the configuration page and configure job parameters, as listed in [Table 3-130](#).

**Table 3-130** Basic job information

Parameter	Description
Owner	An owner configured during job creation is automatically matched. This parameter value can be modified.



Parameter	Description
Executor	User that executes the job. When you enter an executor, the job is executed by the executor. If the executor is left unspecified, the job is executed by the user who submitted the job for startup.
Job Agency	After an agency is configured, the job interacts with other services as an agency during job execution.
Priority	Priority configured during job creation is automatically matched. This parameter value can be modified.
Execution Timeout	Timeout of the job instance. If this parameter is set to 0 or is not set, this parameter does not take effect. If the notification function is enabled for the job and the execution time of the job instance exceeds the preset value, the system sends a specified notification.
Custom Parameter	Set the name and value of the parameter.
Job Label	Configure job labels to manage jobs by category. Click <b>Add</b> to add a tag to the job. You can also select a tag configured in <a href="#">Managing Job Labels</a> .




## Configuring Job Parameters

Job parameters can be globally used in any node in jobs. The procedure is as follows:

Click the blank area in the canvas and then the **Parameter Setup** tab on the right, and configure the parameters listed in [Table 3-131](#).

**Table 3-131** Job parameter setup

Function	Description
<b>Variable Parameter</b>	
Add	<p>Click <b>Add</b> and enter the variable parameter name and parameter value in the text boxes.</p> <ul style="list-style-type: none"> <li>● Parameter Name Only letters, numbers, hyphens, and underscores (_) are allowed.</li> <li>● Parameter Value <ul style="list-style-type: none"> <li>- The string type of parameter value is a character string, for example, <b>str1</b>.</li> <li>- The numeric type of parameter value is a number or operation expression.</li> </ul> </li> </ul> <p>After the parameter is configured, it is referenced in the format of <math>\\${parameter\ name}</math> in the job.</p>


Function	Description
Modify	Change the parameter name or value in the corresponding text boxes.
Mask	If the parameter value is a key, click  to mask the value for security purposes.
Delete	Click  next to the parameter name and value text boxes to delete the job parameter.
<b>Constant Parameter</b>	
Add	<p>Click <b>Add</b> and enter the constant parameter name and parameter value in the text boxes.</p> <ul style="list-style-type: none"> <li>Parameter name Only letters, numbers, hyphens, and underscores (_) are allowed.</li> <li>Parameter value <ul style="list-style-type: none"> <li>The string type of parameter value is a character string, for example, <b>str1</b>.</li> <li>The numeric type of parameter value is a number or operation expression.</li> </ul> </li> </ul> <p>After the parameter is configured, it is referenced in the format of <math>\\${parameter\ name}</math> in the job.</p>
Modify	Modify the parameter name and parameter value in text boxes and save the modifications.
Delete	Click  next to the parameter name and value text boxes to delete the job parameter.

## Testing and Saving the Job

After a job is configured, complete the following operations:


### Batch processing job

**Step 1** Click  to test the job.

**Step 2** After the test is completed, click  to save the job configuration information. If the test fails, modify the parameters as prompted and run the test again.

----End

### Processing jobs in real time

**Step 1** Click  to save the job configuration.

----End

### 3.4.4.4 Setting Up Scheduling for a Job

This section describes how to set up scheduling for an orchestrated job.

- If the processing mode of a job is batch processing, configure scheduling types for jobs. Three scheduling types are supported: run once, run periodically, and event-based. For details, see [Setting Up Scheduling for a Job Using the Batch Processing Mode](#).
- If the processing mode of a job is real-time processing, configure scheduling types for nodes. Three scheduling types are supported: run once, run periodically, and event-based. For details, see [Setting Up Scheduling for Nodes of a Job Using the Real-Time Processing Mode](#).

#### Prerequisites

- You have developed a job by following the instructions in [Developing a Job](#).
- You have locked the job. Otherwise, you must click **Lock** so that you can develop the job. A job you create or import is locked by you by default. For details, see the [lock function](#).

#### Constraints

- Set an appropriate value for this parameter. A maximum of five instances can be concurrently executed in a job. If the start time of a job instance is later than the configured job execution time, the job instances in the subsequent batch will be queued. As a result, the job execution costs a longer time than expected. For CDM and ETL jobs, the recurrence must be at least 5 minutes. In addition, the recurrence should be adjusted based on the data volume of the job table and the update frequency of the source table.
- If you use DataArts Studio DataArts Factory to schedule a CDM migration job and configure a scheduled task for the job in DataArts Migration, 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.

### Setting Up Scheduling for a Job Using the Batch Processing Mode

Three scheduling types are available: **Run once**, **Run periodically**, and **Event-based**. The procedure is as follows:

Click the **Scheduling Setup** tab on the right of the canvas to expand the configuration page and configure the scheduling parameters listed in [Table 3-132](#).

**Table 3-132** Job scheduling parameters

Parameter	Description
Scheduling Type	<p>Scheduling type of the job. Available options include:</p> <ul style="list-style-type: none"> <li>• <b>Run once:</b> You need to manually execute the job.</li> <li>• <b>Run periodically:</b> The job is executed periodically. For details about the parameters, see <a href="#">Table 3-133</a>.</li> <li>• <b>Event-based:</b> The job will be executed when certain external conditions are met. For details about the parameters, see <a href="#">Table 3-134</a>.</li> </ul>
Dry run	If you select this option, the job will not be executed, and a success message will be returned.

**Table 3-133** Parameters for jobs that are executed periodically

Parameter	Description
From and to	The period during which a scheduling task takes effect.
Recurrence	<p>The frequency at which the scheduling task is executed, which can be:</p> <p>Set an appropriate value for this parameter. A maximum of five instances can be concurrently executed in a job. If the start time of a job instance is later than the configured job execution time, the job instances in the subsequent batch will be queued. As a result, the job execution costs a longer time than expected. For CDM and ETL jobs, the recurrence must be at least 5 minutes. In addition, the recurrence should be adjusted based on the data volume of the job table and the update frequency of the source table.</p> <ul style="list-style-type: none"> <li>• <b>Minutes:</b> The job starts at the top of the hour. The interval is accurate to minute. After the scheduling ends at the end time of the current day, the scheduling automatically starts on the next day.</li> <li>• <b>Hours:</b> The job starts at a specified time point. The interval is accurate to hour. After the scheduling ends at the end time of the current day, the scheduling automatically starts on the next day.</li> <li>• <b>Every day:</b> The job starts at a specified time on a day. The scheduling period is one day.</li> <li>• <b>Every week:</b> You can select a specified time point of one or more days in a week.</li> <li>• <b>Every month:</b> You can select a specified time point of one or more days in a month.</li> </ul>

Parameter	Description
<p>Dependency job</p>	<p>If you select a dependency job that is executed periodically, the current job will be executed only when an instance of the dependency job is executed within a certain period of time. You can only search for jobs by name. For details about the conditions of dependency jobs and how a job runs after its dependency jobs are set, see <a href="#">Job Dependency</a>.</p> <p>If you select multiple dependency jobs, you can execute the current job only after all dependency job instances are executed within a specified time range (see <a href="#">How a Job Runs After a Dependency Job Is Set for It</a> for details.).</p> <p>The constraints are as follows:</p> <ul style="list-style-type: none"> <li>• The recurrence of job A cannot be shorter than that of job B. For example, if both job A and job B are scheduled by minute or hour and the interval of job A is shorter than that of job B, then job B cannot be set as the dependency job of job A. If job A is scheduled by minute and job B is scheduled by hour, job B cannot be set as the dependency job of job A.</li> <li>• The recurrence of neither job A nor job B can be week. For example, if the recurrence of job A or job B is week, job B cannot be set as the dependency job of job A.</li> <li>• A job whose recurrence is month can depend only on a job whose recurrence is day. For example, if the recurrence of job A is month, job B can be set as the dependency job of job A only if job B's recurrence is day.</li> </ul>
<p>Policy for Current job If Dependency job Fails</p>	<p>Policy for processing the current job when one or more instances of its dependency job fail to be executed in its period.</p> <ul style="list-style-type: none"> <li>• Suspend Suspends the current job. The suspended job will block the execution of subsequent jobs. You can force the dependency job to be executed successfully.</li> <li>• Continue Continues to execute the current job.</li> <li>• Terminate Stops executing the current job. Its status becomes <b>Canceled</b>.</li> </ul> <p>For example, the recurrence of the current job is 1 hour and that of its dependency jobs is 5 minutes.</p> <ul style="list-style-type: none"> <li>• If the value of this parameter is set to <b>Terminate</b>, the current job will be terminated as long as one of the 12 instances of its dependency job fails.</li> <li>• If the value of this parameter is set to <b>Continue</b>, the current job will be executed after the 12 instances of its dependency job are executed.</li> </ul> <p><b>NOTE</b> You can set this parameter for multiple jobs in a batch. For details, see <a href="#">Configuring a Default Item</a>.</p>

Parameter	Description
Run After Dependency Job Ends	<p>If a job depends on other jobs, the job is executed only after its dependency job instances are executed within a specified time range (see <a href="#">How a Job Runs After a Dependency Job Is Set for It</a> for details). If the dependency job instances are not successfully executed, the current job is in waiting state.</p> <p>If you select this option, the system checks whether all job instances in the previous cycle have been executed before executing the current job.</p>
Cross-Cycle Dependency	<p>Dependency between job instances</p> <ul style="list-style-type: none"> <li>• <b>Independent on the previous schedule cycle:</b> You can set <b>Concurrency</b> to set the number of job instances that are concurrently executed. If you set it to <b>1</b>, a batch is executed only after the previous batch is executed (the execution is successful, cancelled, or failed).</li> <li>• <b>Self-dependent (The current job can continue to run only after the previous schedule cycle is successfully finished.)</b></li> </ul>

**Table 3-134** Parameters for event-based jobs

Parameter	Description
Event Type	<p>Type of the event that triggers job running</p> <ul style="list-style-type: none"> <li>• <b>KAFKA</b></li> </ul>
Parameters for KAFKA event-triggered jobs	
Connection Name	<p>Before selecting a data connection, ensure that a Kafka data connection has been created in the <b>Management Center</b>.</p>
Topic	<p>Topic of the message to be sent to the Kafka.</p>
Concurrent Events	<p>Number of jobs that can be concurrently processed. The maximum number of concurrent events is 128.</p>
Event Detection Interval	<p>Interval at which the system detects the stream for new messages. The unit of the interval can be <b>Second</b> or <b>Minute</b>.</p>
Access Policy	<p>Select the location where data is to be accessed:</p> <ul style="list-style-type: none"> <li>• <b>Access from the last location:</b> For the first access, data is accessed from the most recently recorded location. For the subsequent access, data is accessed from the previously recorded location.</li> <li>• <b>Access from a new location:</b> Data is accessed from the most recently recorded location each time.</li> </ul>

Parameter	Description
Failure Policy	Select a policy to be performed after scheduling fails. <ul style="list-style-type: none"> <li>• Suspend</li> <li>• Ignore the failure and proceed with the next event</li> </ul>

## Setting Up Scheduling for Nodes of a Job Using the Real-Time Processing Mode

Three scheduling types are available: **Run once**, **Run periodically**, and **Event-based**. The procedure is as follows:

Select a node. On the node development page, click the **Scheduling Parameter Setup** tab. On the displayed page, configure the parameters listed in [Table 3-135](#).

**Table 3-135** Parameters for setting up node scheduling

Parameter	Description
Scheduling Type	Scheduling type of the job. Available options include: <ul style="list-style-type: none"> <li>• <b>Run once</b>: You need to manually execute the job.</li> <li>• <b>Run periodically</b>: The job runs automatically and periodically.</li> <li>• <b>Event-based</b>: The job runs when certain external conditions are met.</li> </ul>
<b>Parameters displayed when Scheduling Type is Run periodically</b>	
From and to	The period during which a scheduling task takes effect.
Recurrence	The frequency at which the scheduling task is executed, which can be: <ul style="list-style-type: none"> <li>• Minutes</li> <li>• Hours</li> <li>• Every day</li> <li>• Every week</li> <li>• Every month</li> </ul> For CDM and ETL jobs, the recurrence must be at least 5 minutes. In addition, the recurrence should be adjusted based on the data volume of the job table and the update frequency of the source table.
Cross-Cycle Dependency	Dependency between job instances <ul style="list-style-type: none"> <li>• Independent on the previous schedule cycle</li> <li>• Self-dependent (The current job can continue to run only after the previous schedule cycle is successfully finished.)</li> </ul>
<b>Parameters displayed when Scheduling Type is Event-based</b>	

Parameter	Description
Event Type	Type of the event that triggers job running.
Connection Name	Before selecting a data connection, ensure that a Kafka data connection has been created in the <b>Management Center</b> .
Topic	Topic of the message to be sent to the Kafka.
Consumer Group	<p>A scalable and fault-tolerant group of consumers in Kafka. Consumers in a group share the same ID. They collaborate with each other to consume all partitions of subscribed topics. A partition in a topic can be consumed by only one consumer.</p> <p><b>NOTE</b></p> <ol style="list-style-type: none"> <li>1. A consumer group can contain multiple consumers.</li> <li>2. The group ID is a string that uniquely identifies a consumer group in a Kafka cluster.</li> <li>3. Each partition of each topic subscribed to by a consumer group can be consumed by only one consumer. Consumer groups do not affect each other.</li> </ol> <p>If you select <b>DIS</b> or <b>KAFKA</b> for <b>Event Type</b>, the consumer group ID is automatically displayed. You can also manually change the consumer group ID.</p>
Concurrent Events	Number of jobs that can be concurrently processed. The maximum number of concurrent events is 10.
Event Detection Interval	Interval at which the system detects the stream for new messages. The unit of the interval can be <b>Seconds</b> or <b>Minutes</b> .
Failure Policy	<p>Select a policy to be performed after scheduling fails.</p> <ul style="list-style-type: none"> <li>● Suspend</li> <li>● Ignore failure and proceed</li> </ul>

### 3.4.4.5 Submitting a Version and Unlocking the Script

This involves the version management and lock functions.

- Version management: traces script and job changes, and supports version comparison and rollback. The system retains 10 latest version records. In addition, version management can be used to distinguish the development state and production state.
  - Development state: Scripts or jobs have not been submitted and are used for debugging. In the development state, you can edit, save, and run scripts or jobs without affecting those being scheduled. In addition, when a job is being associated with a script or job dependency is being configured, the associated script or job will read the configuration in the development state.
  - Production state: Script or jobs have been submitted and are used for formal scheduling. In formal scheduling, the latest submitted versions of



scripts or jobs will be used in scenarios such as script invocation, instance rerunning, and job dependency and patch data configuration.

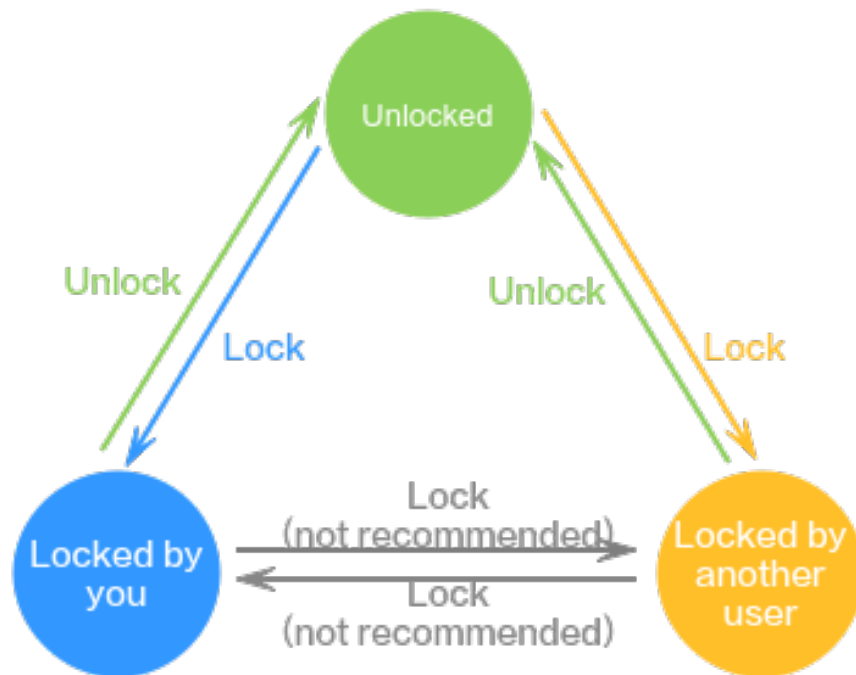
- Lock: prevents conflict caused by collaborative script or job development. If you create or import a script or job, it is locked by you by default. You can only edit, save, or submit a script or job you have locked. To edit, save, or submit a script or job that is locked by another user or not locked by any user, you must lock the script or job first.

---

#### NOTICE

- You can view the lock status of a script or job in the script or job directory tree.
  - To view the latest version of an opened script or job locked by another user, you need to re-open the script or job because it is not updated in real time.
  - Scripts or jobs that were created before the lock function was available are now unlocked by default. To edit, save, or submit these scripts or jobs, you need to lock them first.
  - The locking operation depends on the soft and hard lock policies. For details about how to configure soft and hard lock policies, see [Configuring a Default Item](#).
    - Soft lock: You can lock or unlock jobs or scripts, regardless of whether they are locked by others.
    - **Hard Lock:** You can lock jobs or scripts only after they have been unlocked by other users. The space administrator and the **DAYU Administrator** user can lock and unlock jobs or scripts without any limitations.
  - Do not lock a script or job that is locked by another user because if you do so, changes to the script or job made by the user will be lost. If you want to modify the script or job, contact the user to unlock the script or job, and lock it by yourself.
-

Figure 3-157 Lock status



## Prerequisites

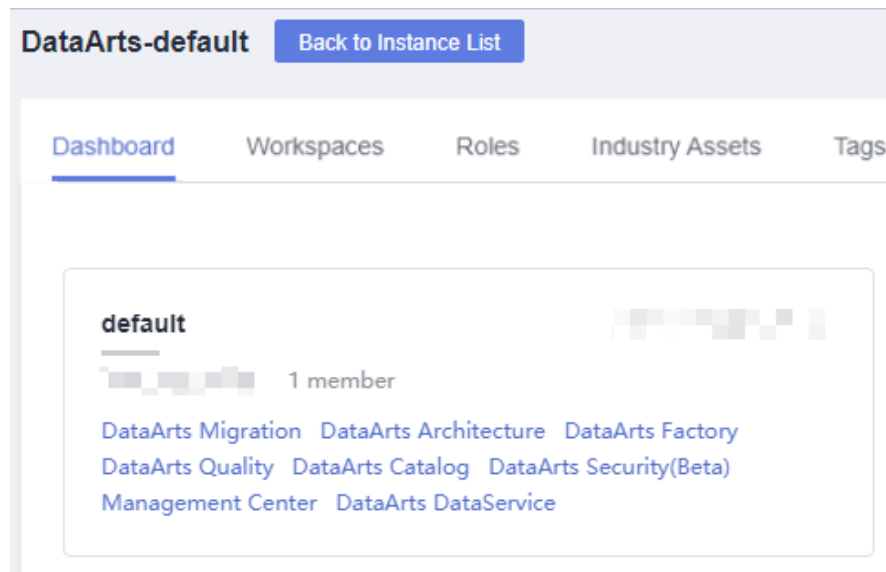
A job has been developed.

## Submitting a Version and Unlocking the Script

If you submit a version, the latest job in the development state will be saved and submitted and overwrite the previous job version. You are advised to unlock the job after submitting the version so that other developers can modify the job as needed.

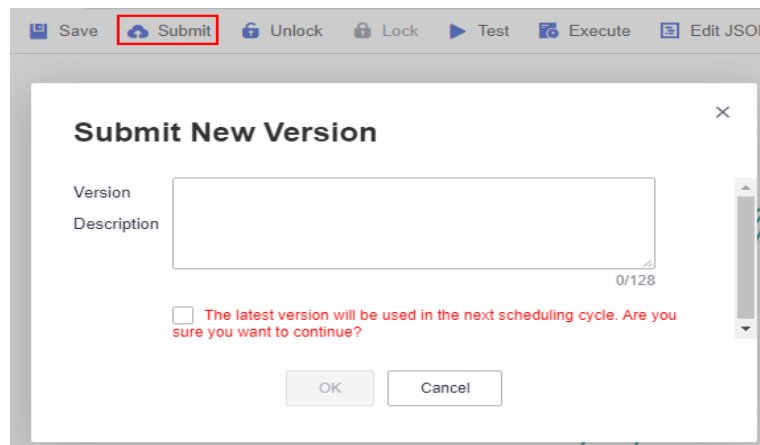
- Step 1** Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-158 DataArts Factory



- Step 2** In the left navigation pane of DataArts Factory, choose **Development > Develop Job**.
- Step 3** In the job directory, double-click the developed job to access the job development page.
- Step 4** Above the job canvas, click **Submit** to submit a version. In the displayed dialog box, enter the change description (a maximum of 128 characters allowed) and select the check box below. If you do not select this option, you cannot click **OK**.

Figure 3-159 Submitting a version



- Step 5** Above the job canvas, click **Unlock** to unlock the job.

Figure 3-160 Unlocking a job



----End

## Version Rollback

After submitting the version, you can view it in the version list. (Currently, a maximum of 10 latest versions are saved.) Click **Roll Back** to roll back to any submitted version.

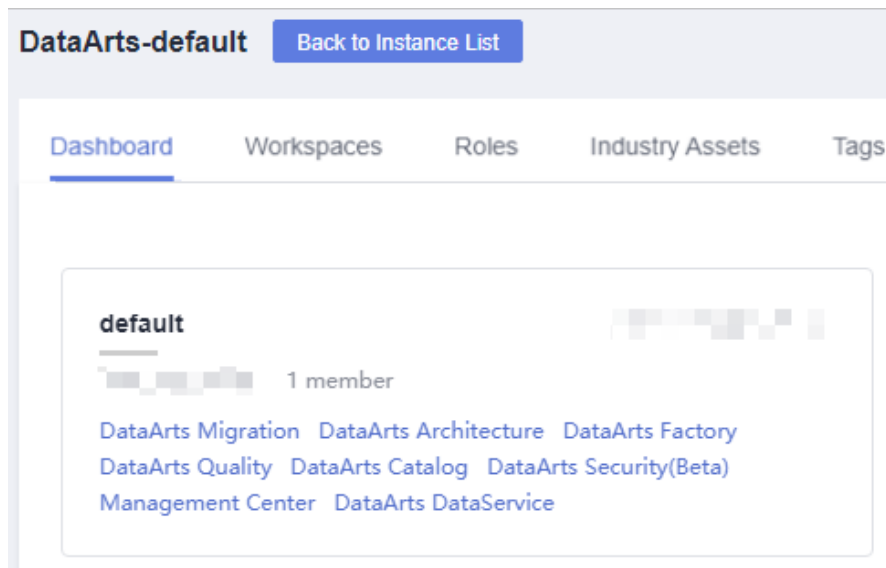
The rollback involves the following contents:

- Job definition (such as operator properties and connection lines)
- Basic job information, job scheduling configuration, job parameters, and lineage

The procedure is as follows:

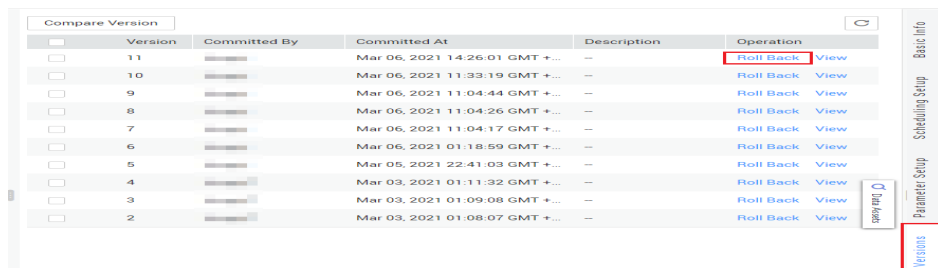
1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-161 DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Development > Develop Job**.
3. In the job directory, double-click a job to access the job development page.
4. On the right of the page, click the **Versions** tab and view the version submission records. Select the version to be rolled back and click **Roll Back**.

Figure 3-162 Rolling back the version



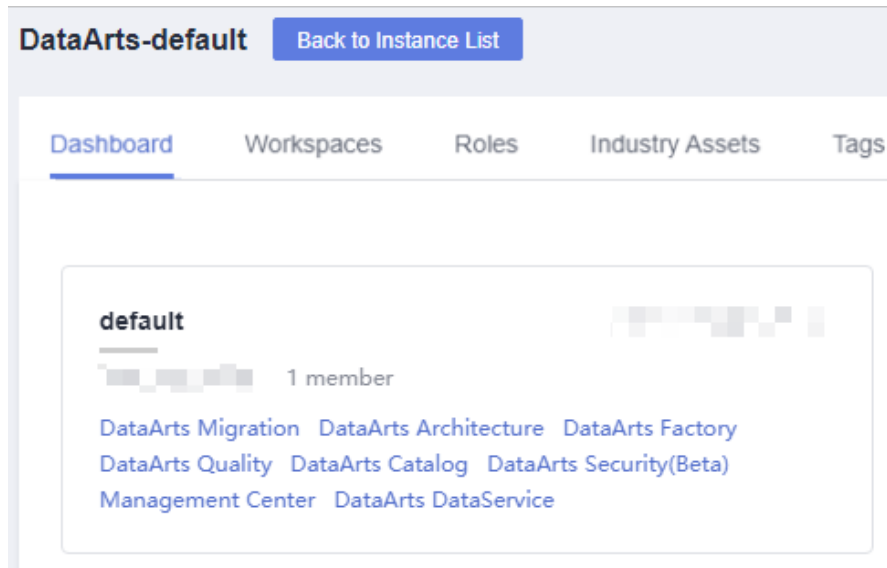
## Viewing Version Details

You can view the submitted version information in the version list.

The procedure is as follows:

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-163 DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Development** > **Develop Job**.
3. In the job directory, double-click a job to access the job development page.
4. On the right of the page, click the **Versions** tab and view the version submission records. Select the desired version and click **View** to view its details.

A new page is displayed, showing the job definition of the version. You cannot modify any job attributes in this window.

Figure 3-164 Viewing version details

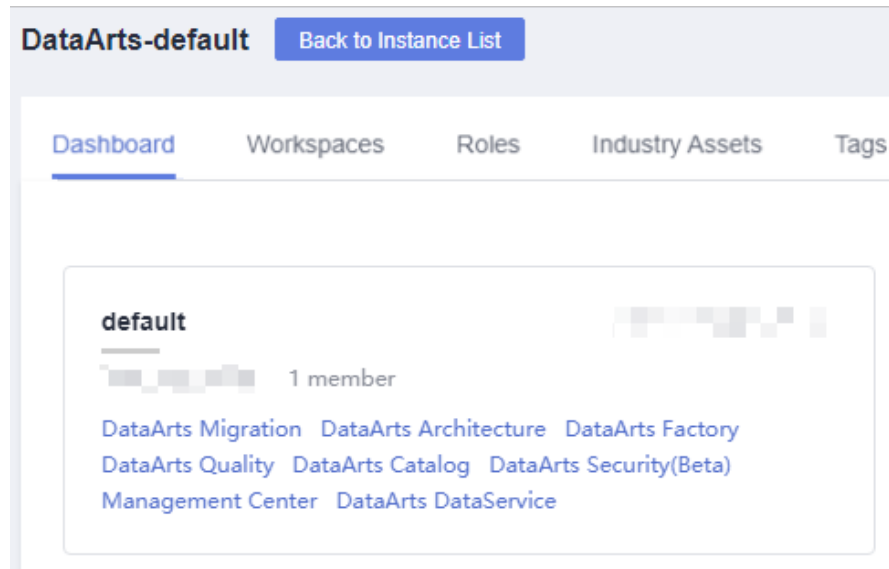
The screenshot shows the 'Viewing version details' page. It features a table with the following columns: 'Version', 'Committed By', 'Committed At', 'Description', and 'Operation'. The table contains 11 rows of data, with the 'View' button in the 'Operation' column highlighted in red. On the right side of the page, there is a vertical navigation pane with tabs for 'Versions', 'Parameter Setup', 'Scheduling Setup', and 'Basic Info'. The 'Versions' tab is selected and highlighted in red.

Version	Committed By	Committed At	Description	Operation
11		Mar 06, 2021 14:26:01 GMT +...		Roll Back View
10		Mar 06, 2021 11:33:19 GMT +...		Roll Back View
9		Mar 06, 2021 11:04:44 GMT +...		Roll Back View
8		Mar 06, 2021 11:04:26 GMT +...		Roll Back View
7		Mar 06, 2021 11:04:17 GMT +...		Roll Back View
6		Mar 06, 2021 01:18:59 GMT +...		Roll Back View
5		Mar 05, 2021 22:41:03 GMT +...		Roll Back View
4		Mar 03, 2021 01:11:32 GMT +...		Roll Back View
3		Mar 03, 2021 01:09:08 GMT +...		Roll Back View
2		Mar 03, 2021 01:08:07 GMT +...		Roll Back View

## Version Comparison

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-165 DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Development > Develop Job**.
3. In the job directory, double-click a job to access the job development page.
4. On the right of the page, click the **Versions** tab and view the version submission records. Select the versions to be compared and click **Compare Version**.

If you select only one version, the selected version is compared with the JSON of the development-state job. If you select two versions, the JSON of the two versions is compared.

Figure 3-166 Comparing versions

Version	Committed By	Committed At	Description	Operation
<input checked="" type="checkbox"/> 3	[blurred]	Mar 24, 2022 10:27:16 GMT+08:00	--	Roll Back   View
<input checked="" type="checkbox"/> 2	[blurred]	Mar 24, 2022 10:24:54 GMT+08:00	--	Roll Back   View
<input checked="" type="checkbox"/> 1	[blurred]	Feb 24, 2022 15:06:54 GMT+08:00	--	Roll Back   View

On the right side of the table, there is a vertical sidebar with the following items: 'Basic Info', 'Scheduling Setup', 'Parameter Setup', and 'Versions' (which is highlighted).

### 3.4.4.6 (Optional) Managing Jobs

#### 3.4.4.6.1 Copying a Job

This section describes how to copy a job.

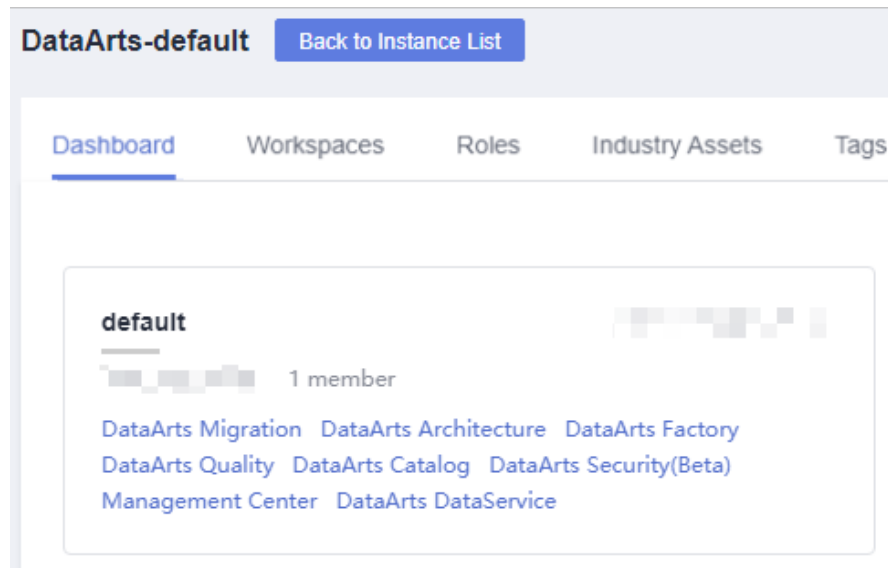
## Prerequisites

A job has been developed. For details about how to develop a job, see [Developing a Job](#).

## Procedure

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-167** DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Development > Develop Job**.
3. In the job directory, select the job to be copied, right-click the job name, and choose **Copy Save As**.
4. In the displayed dialog box, configure related parameters. [Table 3-136](#) describes the parameters.

**Table 3-136** Job and directory parameters

Parameter	Description
Job Name	Name of the job. Must consist of 1 to 128 characters and contain only letters, digits, hyphens (-), underscores (_), and periods (.).
Select Directory	Parent directory of the job directory. The parent directory is the root directory by default.

5. Click **OK**.

### 3.4.4.6.2 Copying the Job Name and Renaming a Job

You can copy the name of a job and rename a job.

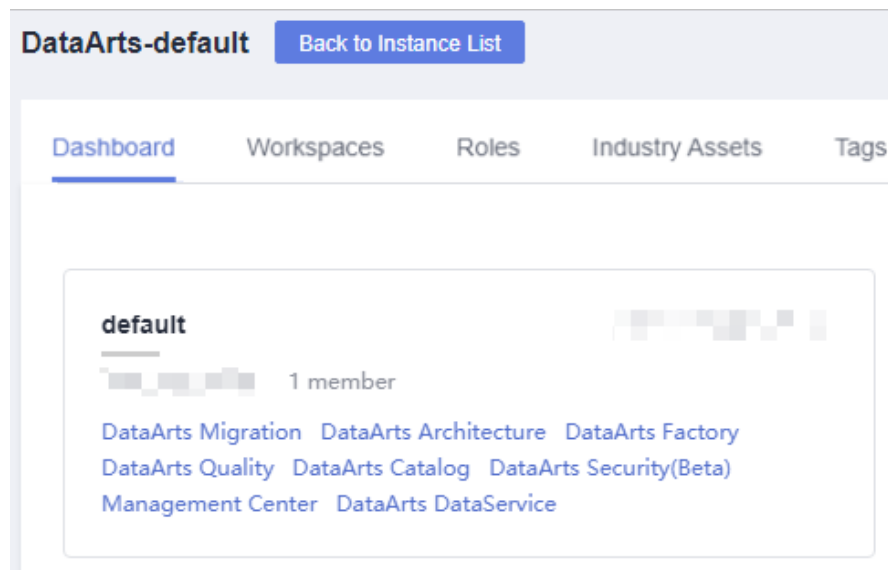
## Prerequisites

A job has been developed. For details about how to develop a job, see [Developing a Job](#).

## Copying the Job Name

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-168** DataArts Factory



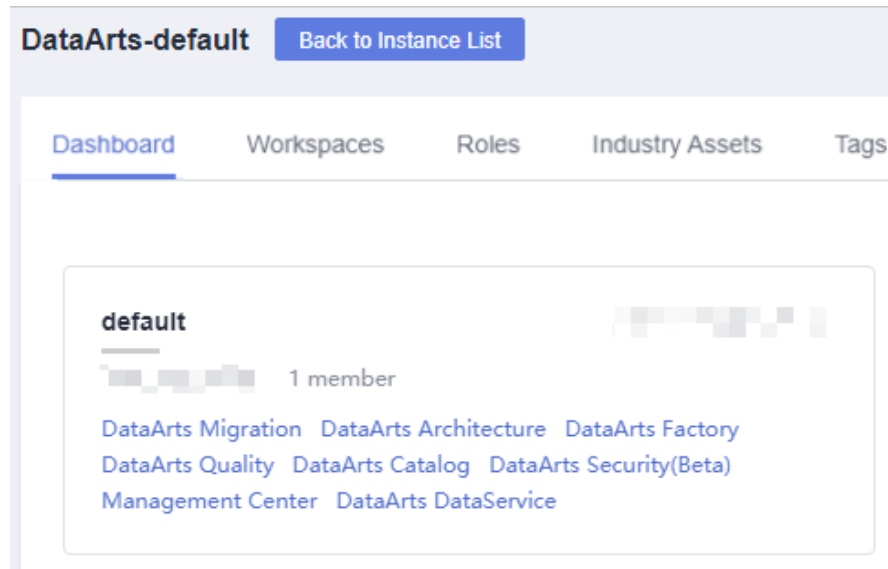
2. In the left navigation pane of DataArts Factory, choose **Development > Develop Job**.
3. Locate the target job in the job directory, right-click the job name, and select **Copy Name** to copy the job name to the clipboard.

## Renaming a job

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.



**Figure 3-169** DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Development > Develop Job**.
3. In the job directory, select the job to be renamed. Right-click the job name and choose **Rename** from the shortcut menu.
4. In the displayed **Modify Job Name** dialog box, change the job name.

**Table 3-137** Job renaming parameters

Parameter	Description
Job Name	Name of the job. Must consist of 1 to 128 characters and contain only letters, digits, hyphens (-), underscores (_), and periods (.).

5. Click **OK**.

### 3.4.4.6.3 Moving a Job or Job Directory

You can move a job file from one directory to another or move a job directory to another directory.

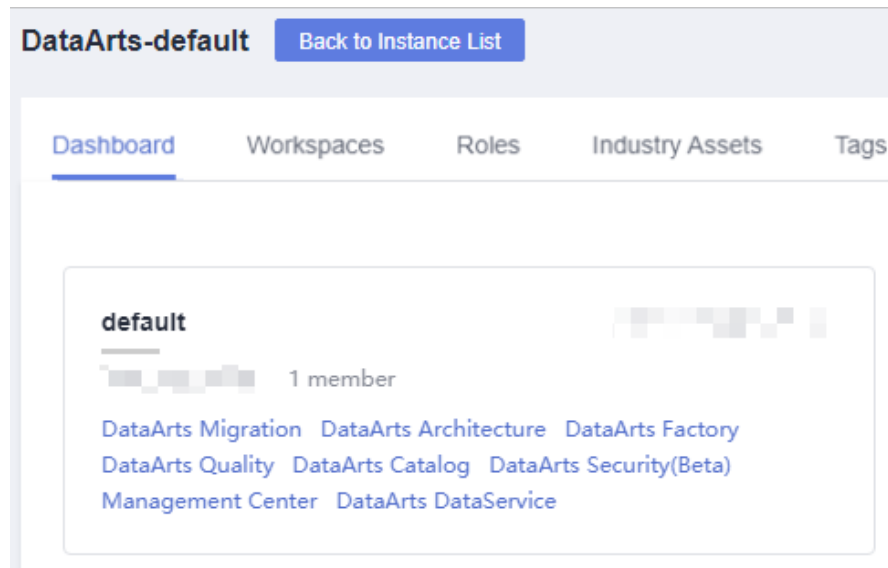
### Prerequisites

A job has been developed. For details about how to develop a job, see [Developing a Job](#).

### Procedure

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-170** DataArts Factory



2. In the navigation pane of the Data Development homepage, choose **Data Development > Develop Job**.
3. Move a job or job directory.

**Method 1: right-click**

- a. In the job directory, right-click a job or job folder and select **Move**.
- b. In the displayed dialog box, configure the target directory.

**Table 3-138** Parameters for moving a job or job directory

Parameter	Description
Select Directory	Directory to which the job or job directory is to be moved. The parent directory is the <b>root</b> directory by default.

- c. Click **OK**.

**Method 2: drag-and-drop**


Select a job or job folder and drag and drop it to the target folder.

**3.4.4.6.4 Exporting and Importing a Job**

- Exporting a job is to export the latest saved content in the development state.
- After a job is imported, the content in the development state is overwritten and a new version is automatically submitted.

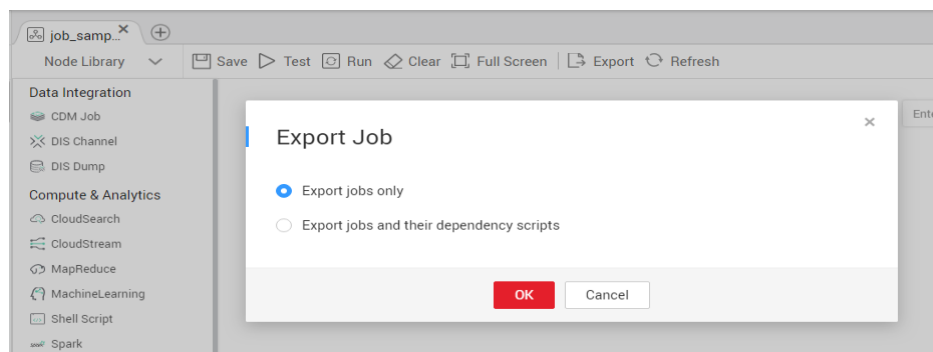
**Exporting Jobs**

**Method 1: Export a job on the job development page.**

- Step 1** Double-click a job name to access the development page of the job, click , and select the type of the job to be exported.

- **Export jobs only:** Export the connection relationships and property configurations of nodes to a local PC, excluding sensitive information such as passwords. After the export, you can use a browser to download the .zip package.
- **Export jobs and their dependency scripts:** Export the node connection relationships, node property configurations, job scheduling configurations, parameter configurations, dependency scripts, and resource definitions to a local PC, excluding sensitive information such as passwords. After the export, you can use a browser to download the .zip package.

Figure 3-171 Exporting a job (method 1)



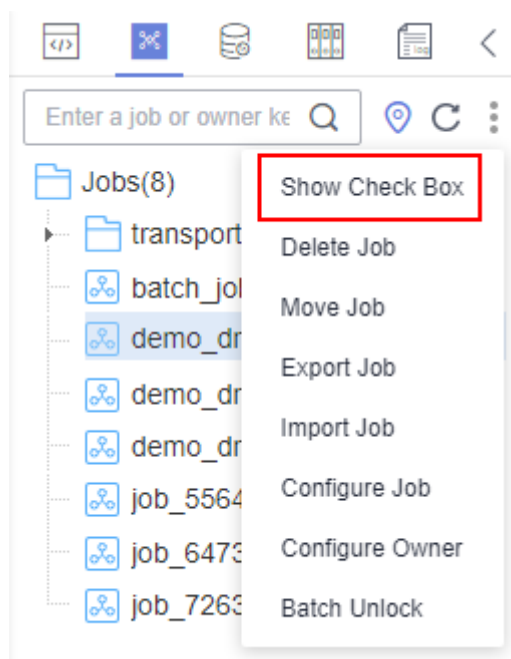
**Step 2** Click **OK** to export the required job file.


----End

**Method 2: Export one or more jobs from the job directory.**

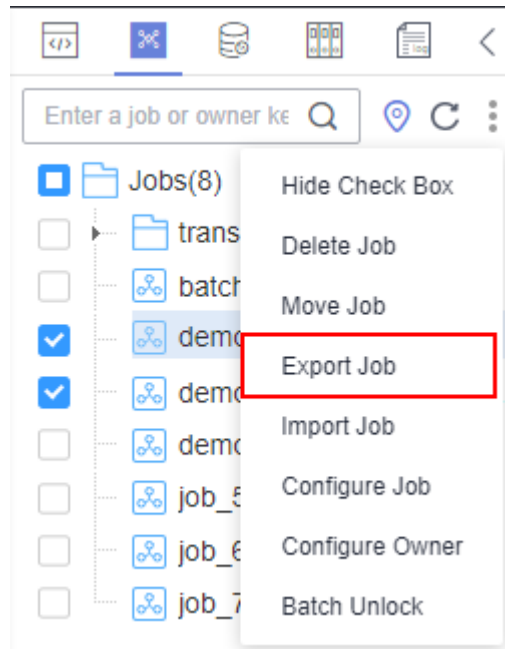
**Step 1** Click  in the job directory and select **Show Check Box**.

Figure 3-172 Clicking Show Check Box



- Step 2** Select the jobs to export, click , and select **Export Job**. In the displayed dialog box, select **Export jobs only** or **Export jobs and their dependency scripts and resource definitions**. After the export is successful, you can obtain the exported .zip file.

**Figure 3-173** Selecting and exporting a job




----End

## Importing a Job

This function is available only if the OBS service is available. If OBS is unavailable, jobs can be imported from the local PC.

**Export one or more jobs from the job directory.**

- Step 1** Click  > **Import Job** in the job directory, select the job file that has been uploaded to OBS or local directory, and rename the policy.

### NOTE

If you select **Overwrite** for **Duplicate Name Policy** but the hard lock policy is used and the script is locked by another user, the overwriting will fail. For details about soft and hard lock policies, see [Configuring the Hard and Soft Lock Policy](#).

**Figure 3-174** Importing job definitions and dependencies

**Import Job** ×

\* File Location

\* Select File from OBS

\* Duplicate Name Policy  Overwrite  Skip

**Step 2** Click **Next** to import the job as instructed.

**NOTE**

During the import, if the data connection, DLI queue, or GES graph associated with the job does not exist in DataArts Factory, the system prompts you to select one again.

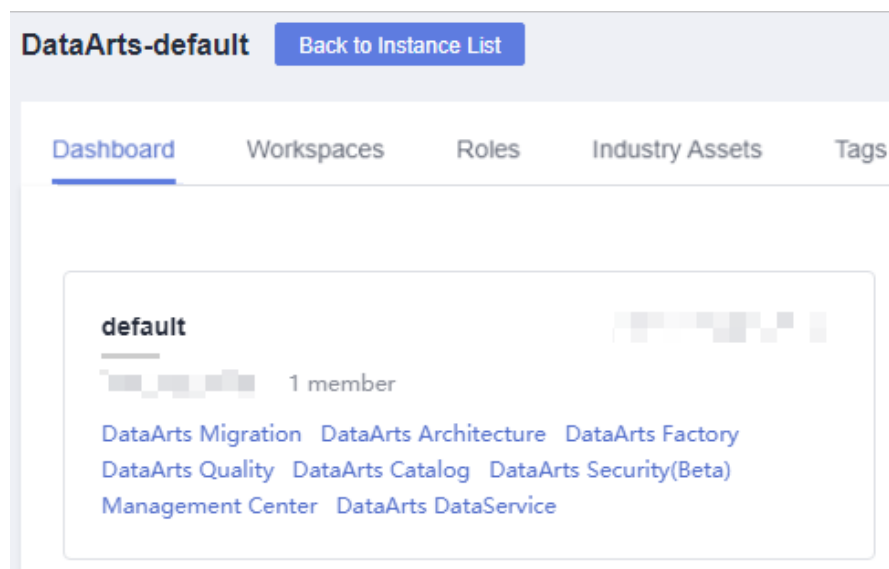
----End

## Example


Context:

- A DWS data connection **doctest** is created in DataArts Factory.
  - A real-time job **doc1** is created in the job directory. Node **DWS SQL** is added to the job. The **Data Connection** of the node is set to **doctest**. **SQL Script** and **Database** are both configured.
1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-175** DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Development > Develop Job**.

3. Search for **doc1** in the job search box, export the job to the local host, and then upload it to the OBS folder.
4. Delete the **doctest** data connection associated with the job in DataArts Factory.
5. Click  > **Import Job** in the job directory, select the job file that has been uploaded to OBS, and set the duplicate name policy.
6. Click **Next** and select another data connection as prompted.
7. Click **Next** and then **Close**.

### 3.4.4.6.5 Deleting a Job

If you do not need to use a job any more, perform the following operations to delete it to reduce the quota usage of the job.

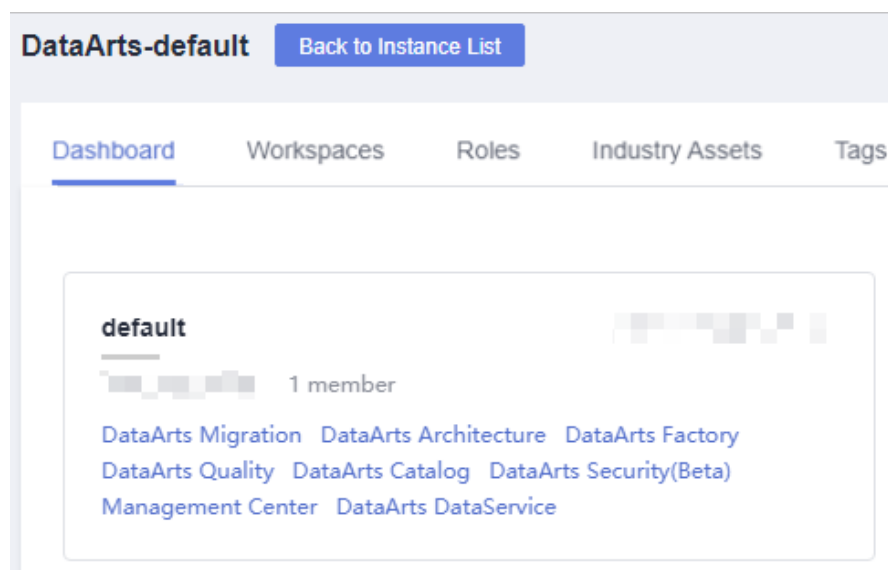
#### NOTE

Deleted jobs cannot be recovered. Exercise caution when performing this operation.

## Deleting a Script

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-176 DataArts Factory

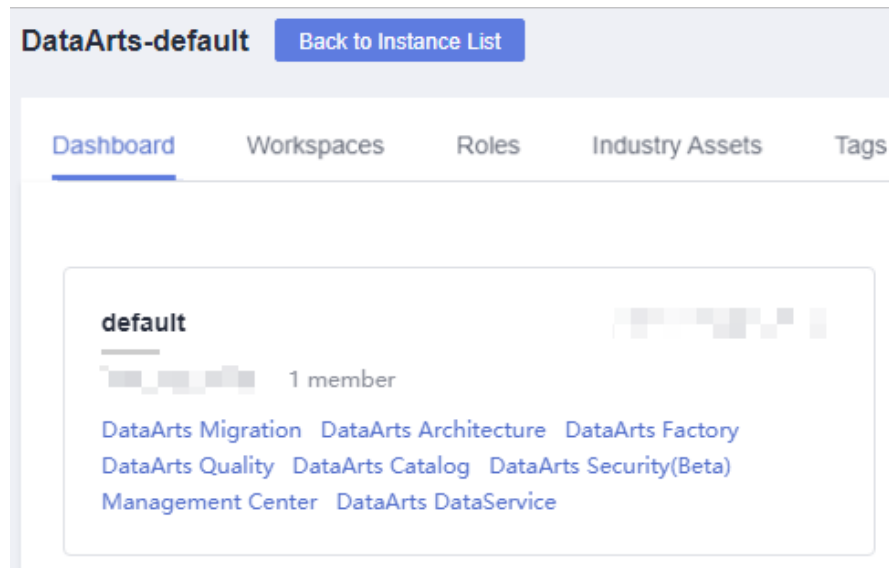




2. In the left navigation pane of DataArts Factory, choose **Development** > **Develop Job**.
3. In the job directory, right-click the job that you want to delete and choose **Delete** from the shortcut menu.
4. In the displayed dialog box, click **OK**.

## Batch Deleting Scripts

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-177 DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Development > Develop Job**.
3. On the top of the job directory, click  and select **Show Check Box**.
4. Select the jobs to be deleted, click , and select **Batch Delete**.
5. In the displayed dialog box, click **OK**.

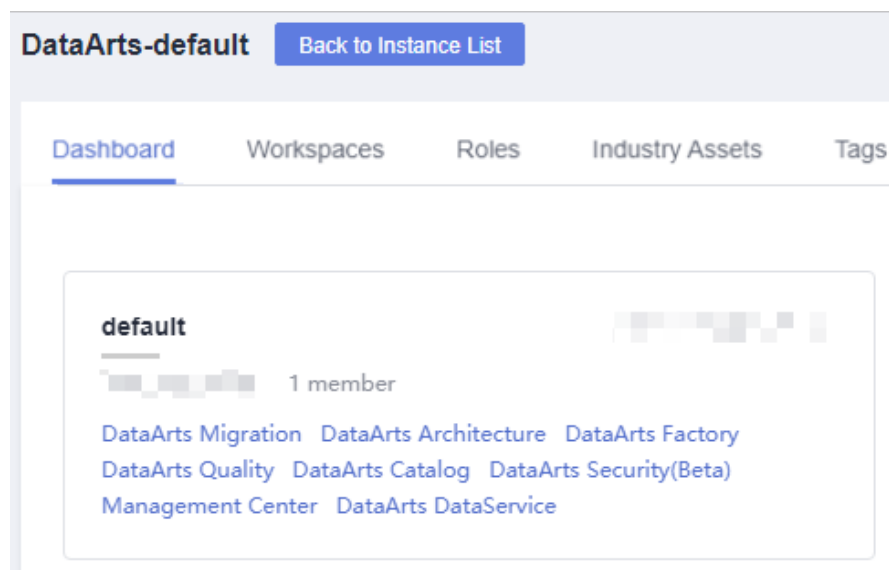
### 3.4.4.6.6 Changing the Job Owner

DataArts Factory allows you to change the owner for jobs with a few clicks.

#### Procedure

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-178 DataArts Factory




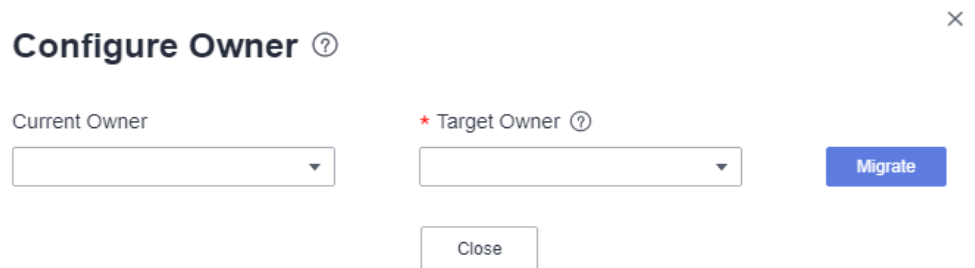
2. In the left navigation pane of DataArts Factory, choose **Development > Develop Job**.
3. At the top of the job directory, click  and select **Configure Owner**.

Figure 3-179 Configuring the owner

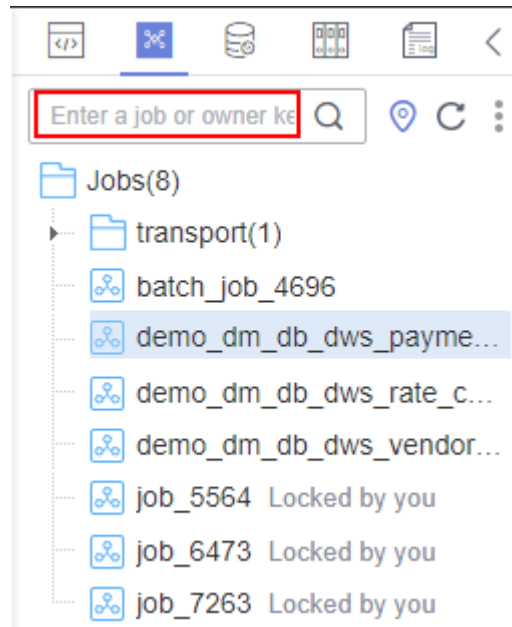


4. Set **Current Owner** and **Target Owner** and click **Migrate**.
5. When the migration succeeds, click **Close**.

## Related Operations

You can use an owner to filter jobs by entering the owner in the search box above the job directory.

Figure 3-180 Filtering jobs by owner



### 3.4.4.6.7 Unlocking Jobs

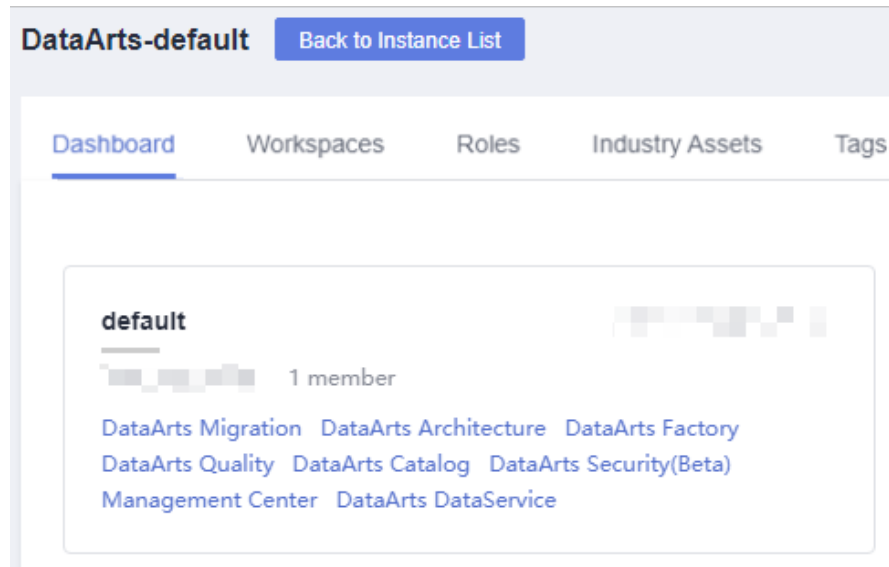
This section describes how to unlock jobs in batches.




## Procedure

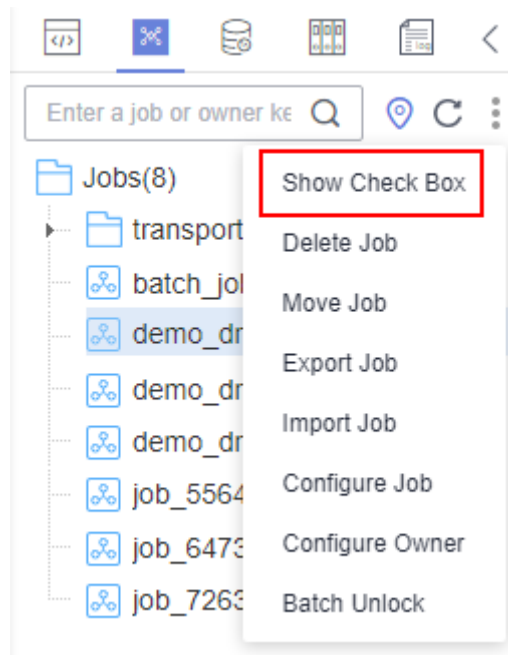
1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-181** DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Development > Develop Job**.
3. Click  in the job directory and select **Show Check Box**.

**Figure 3-182** Clicking Show Check Box




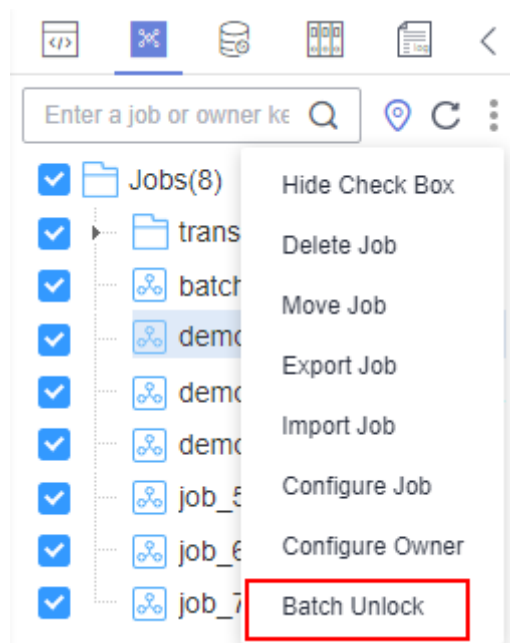
4. Select the jobs to unlock, click , and select **Batch Unlock**. The message "Unlocked." is displayed.

Figure 3-183 Batch Unlock



## 3.4.5 Solution

### Context

The solution aims to provide users with convenient and systematic management operations and better meet service requirements and objectives. Each solution can contain one or more business-related jobs, and one job can be used by multiple solutions.

You can perform the following operations on a solution:

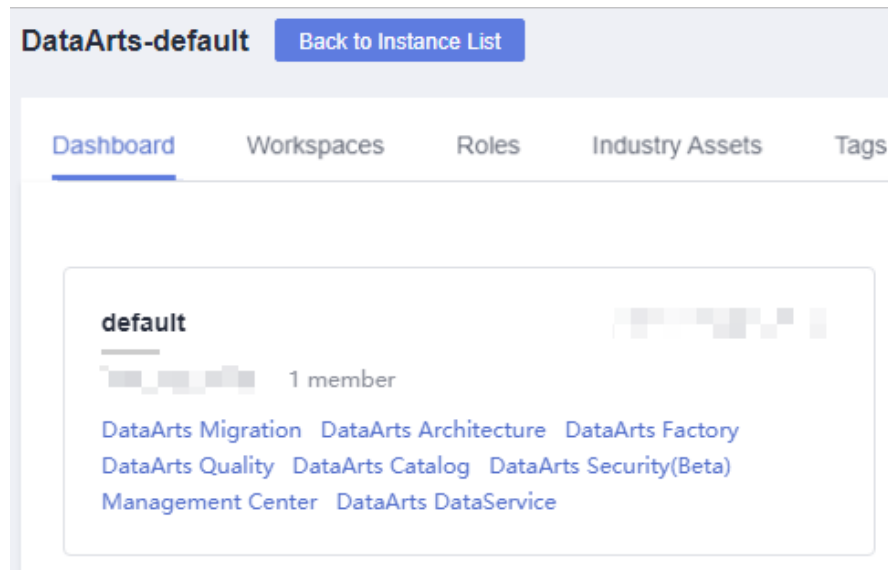
- [Creating a Solution](#)
- [Editing a Solution](#)
- [Exporting a Solution](#)
- [Importing a Solution](#)
- [Upgrading a Solution](#)
- [Deleting a Solution](#)



### Creating a Solution

On the development page of DLF, create a solution, set the solution name, and select business-related jobs.

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-184** DataArts Factory



2. In the navigation tree on the left of the data development page, choose **Development > Develop Script** or **Data Development > Develop Job**.
3. Above the directory on the left, click  to show the solution directory.
4. Click  in the upper part of the solution directory. The **Create Solution** page is displayed. [Table 3-139](#) describes the solution parameters.

**Table 3-139** Solution Parameters

Parameter	Description
Name	Name of the solution.
Select Job	Select the jobs contained in the solution.

5. Click **OK**. The new solution is displayed in the directory on the left.

## Editing a Solution

In the solution directory, right-click the solution name and select **Edit** to change the name and job.

## Exporting a Solution

In the solution directory, right-click the solution name and choose **Export** from the shortcut menu to export the solution file in ZIP format to the local host.

## Importing a Solution

This solution is available only if the OBS service is available. If OBS is unavailable, data can be imported from the local PC.

In the solution directory, right-click a solution and choose **Import Solution** from the shortcut menu to import the solution file that has been uploaded to OBS or local directory.

 NOTE

If you select **Overwrite** for **Duplicate Name Policy** but the hard lock policy is used and the script is locked by another user, the overwriting will fail. For details about soft and hard lock policies, see [Configuring the Hard and Soft Lock Policy](#).

## Upgrading a Solution

In the solution directory, right-click the solution name and choose **Upgrade** from the shortcut menu to import the solution file that has been uploaded to OBS. During the solution upgrade, the running jobs are stopped. The system determines whether to restart the jobs after the upgrade based on the configured upgrade restart policy.

## Deleting a Solution

In the solution directory, right-click the solution name and choose **Delete** from the shortcut menu. A deleted solution cannot be restored. Exercise caution when performing this operation.

## 3.4.6 Execution History

This section describes how to view the execution history of scripts, jobs, and nodes over a week.

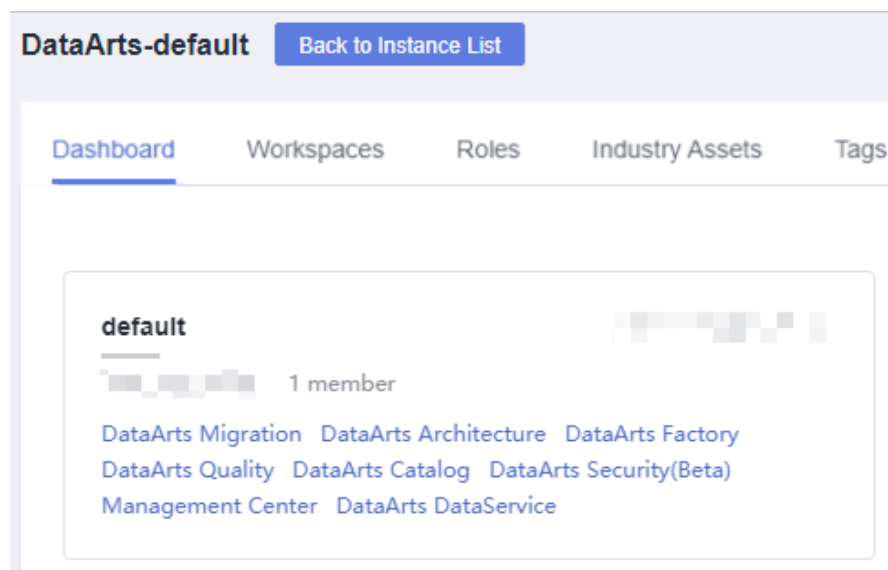
### Prerequisites


This function depends on OBS buckets. For details about how to configure OBS buckets, see [Configuring an OBS Bucket](#).

### Script Execution History

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-185 DataArts Factory

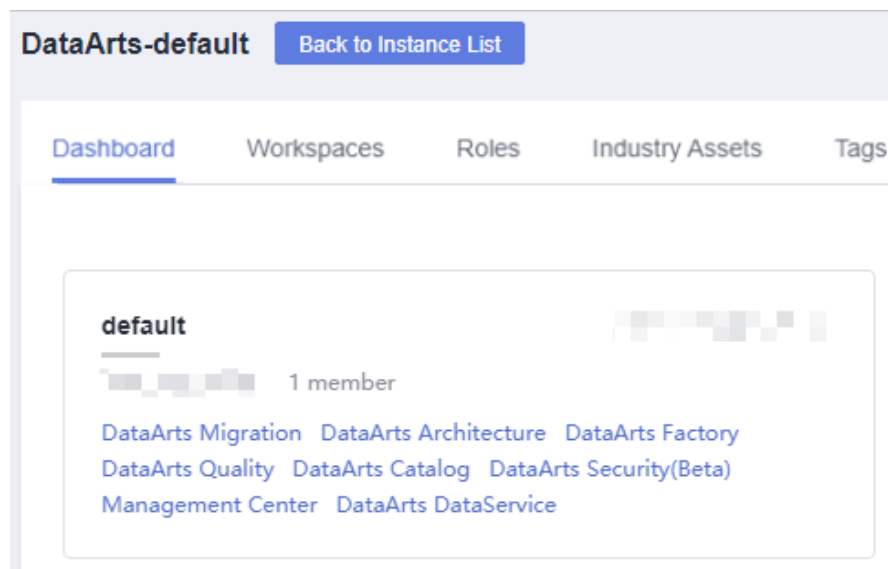



2. In the navigation pane of the DataArts Factory homepage, choose **Data Development > Develop Script**.
3. Above the directory, click  to display the script and job execution history in the past seven days.
4. Select **Scripts** from the drop-down list box to filter out the script execution history.
5. Click a record to view the script information and execution result.

## Job Execution History

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-186** DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Data Development > Develop Job**.
3. Above the directory, click  to display the script and job execution history in the past seven days.
4. Select **Jobs** from the drop-down list box to filter out the job execution history.
5. Click a record to view the job and log information.

### NOTE

If only some nodes of the job were tested, the execution history only displays information and logs for these nodes.

## 3.4.7 O&M and Scheduling

### 3.4.7.1 Overview

Choose **Monitoring > Overview**. On the **Overview** page, you can view the statistics of job instances in charts. Currently, you can view four types of statistics:

- Today's Job Instance Scheduling
- Latest 7 Days' Job Instance Scheduling
- Latest 30 Days' Top 10 Ranking in Job Instance Execution Duration

Click a job name to go to the **Monitor Instance** page and view the detailed running records of the job instance with a long execution time.

- Latest 30 Days' Top 10 Ranking in Job Instance Running Failed

Click the value in the **Failed Count** column. On the displayed **Monitor Instance** page, view the detailed running records of the job instance that is running abnormally.

### 3.4.7.2 Monitoring a Job

#### 3.4.7.2.1 Monitoring a Batch Job

In the batch processing mode, data is processed periodically in batches based on the job-level scheduling plan, which is used in scenarios with low real-time requirements. This type of job is a pipeline that consists of one or more nodes and is scheduled as a whole. It cannot run for an unlimited period of time, that is, it must end after running for a certain period of time.


You can choose **Monitor Job** and click the **Batch Job Monitoring** tab to view the scheduling status, frequency, and start time of a batch job, and perform the operations listed in [Table 3-140](#).

**Figure 3-187** Monitoring a Batch Job



**Table 3-140** Operations supported by batch job monitoring

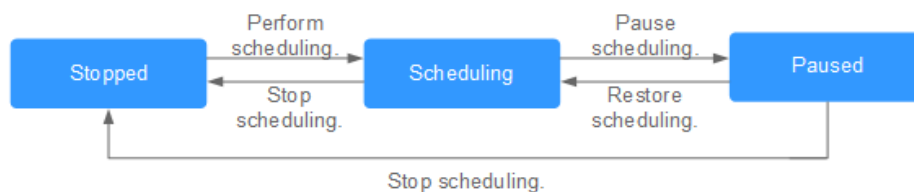
N o.	Operation	Description
1	Searching for a job based on the job name or owner	-
2	Filtering jobs by whether notifications have been configured, scheduling status, job label, or next plan time	-
3	Perform operations on jobs in a batch	Select multiple jobs and perform operations on them.

No.	Operation	Description
4	Viewing job instance status	Click  in front of the job name. The <b>Last Instance</b> page is displayed. You can view information about the last instance of the job.
5	Viewing node information of the job	Click a job name. On the displayed page, click the job node and view its associated jobs/scripts and monitoring information.
6	Job scheduling operations	In the <b>Operation</b> column of a job, you can run, pause, recover, stop, and configure scheduling. For details, see <a href="#">Batch Job Monitoring: Scheduling a Job</a> .
7	Configuring notifications	In the <b>Operation</b> column of a job, choose <b>More &gt; Set Notification</b> . In the displayed dialog box, configure notification parameters. <a href="#">Table 3-150</a> describes the notification parameters.
8	Monitoring instances	In the <b>Operation</b> column of a job, choose <b>More &gt; Monitor Instance</b> to view the running records of all instances of the job.
9	PatchData	In the <b>Operation</b> column of a job, choose <b>More &gt; PatchData</b> . For details, see <a href="#">Batch Job Monitoring: PatchData</a> .
10	Adding a job label	In the <b>Operation</b> column of a job, choose <b>More &gt; Add Job Label</b> . For details, see <a href="#">Batch Job Monitoring: Adding a Job Label</a> .

## Batch Job Monitoring: Scheduling a Job

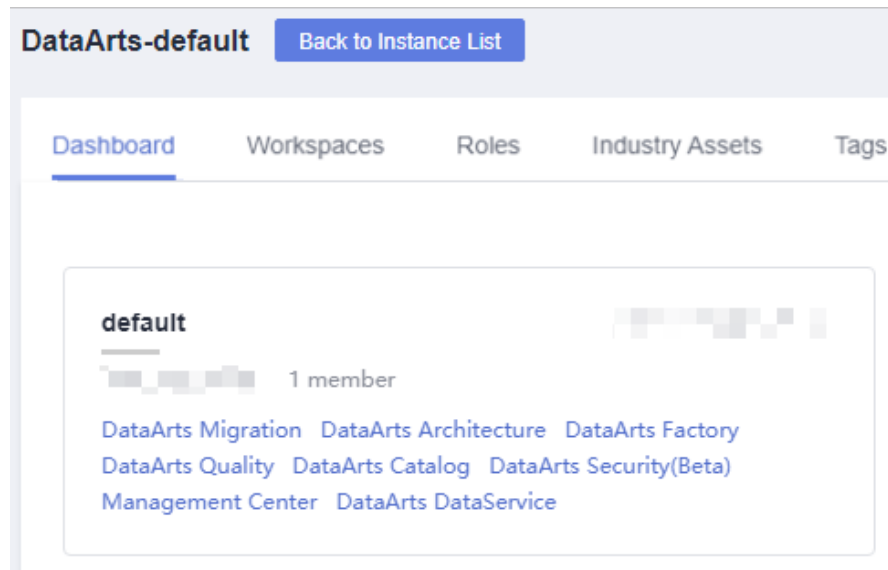
After developing a job, you can manage job scheduling tasks on the **Monitor Job** page. Specific operations include to run, pause, restore, or stop scheduling.

**Figure 3-188** Scheduling a job



1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

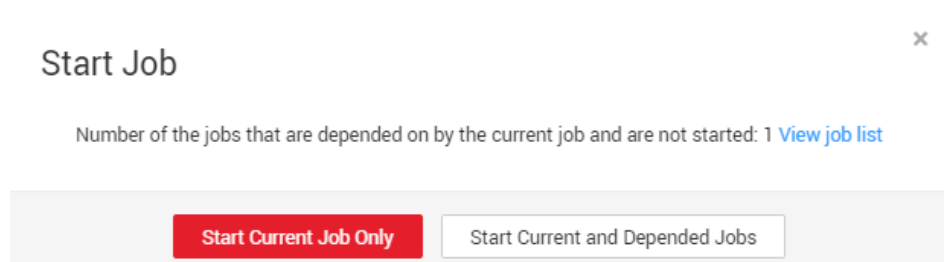
Figure 3-189 DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Monitoring > Monitor Job**.
3. Click the **Batch Job Monitoring** tab.
4. In the **Operation** column of the job, click **Submit, Pause, Restore, or Stop**.

If a dependent job has been configured for a batch job, you can select either **Start Current Job Only** or **Start Current and Depended Jobs** when submitting the batch job. For details about how to configure dependent jobs, see [Setting Up Scheduling for a Job Using the Batch Processing Mode](#).

Figure 3-190 Starting a job



## Batch Job Monitoring: PatchData

A job executes a scheduling task to generate a series of instances in a certain period of time. This series of instances are called PatchData. PatchData can be used to fix the job instances that have data errors in the historical records or to build job records for debugging programs.

Only the periodically scheduled jobs support PatchData. For details about the execution records of PatchData, see [Monitoring PatchData](#).

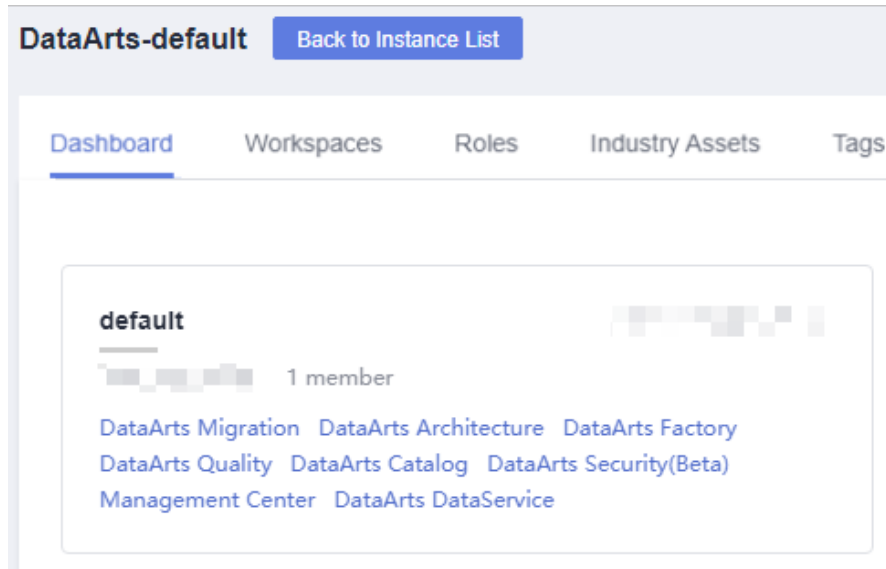
### NOTE

Do not modify the job configuration when PatchData is being performed. Otherwise, job instances generated during PatchData will be affected.



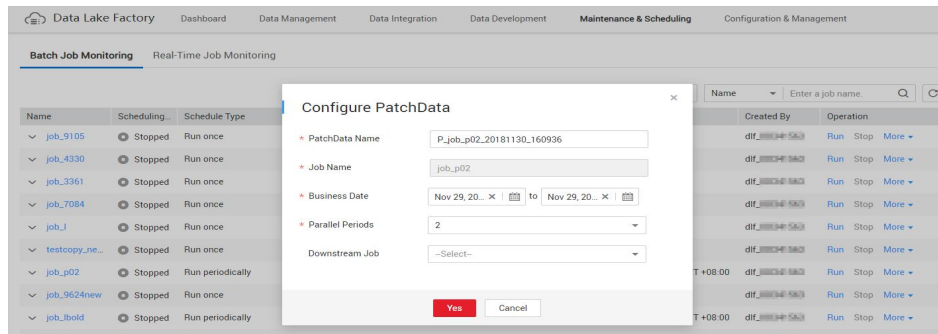
1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-191** DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Monitoring > Monitor Job**.
3. Click the **Batch Job Monitoring** tab.
4. In the **Operation** column of the job, choose **More > Configure PatchData**.
5. Configure PatchData parameters based on [Table 3-141](#).

**Figure 3-192** PatchData parameters



**Table 3-141** Parameters

Parameter	Description
PatchData Name	Name of the automatically generated PatchData task. The value can be modified.
Job Name	Name of the job that requires PatchData.

Parameter	Description
Date	<p>Period of time when PatchData is required.</p> <p><b>NOTE</b> PatchData can be configured for a job multiple times. However, avoid configuring PatchData multiple times on the same date to prevent data duplication or disorder.</p>
Parallel Instances	<p>Number of instances to be executed at the same time. A maximum of five instances can be executed at the same time.</p> <p><b>NOTE</b> Set this parameter based on the site requirements. For example, if a CDM job instance is used, data cannot be supplemented at the same time. The value of this parameter can only be set to 1.</p>
Downstream Job Requiring PatchData	<p>Select the downstream jobs (jobs that depend on the current job) that require PatchData. You can select multiple jobs.</p>

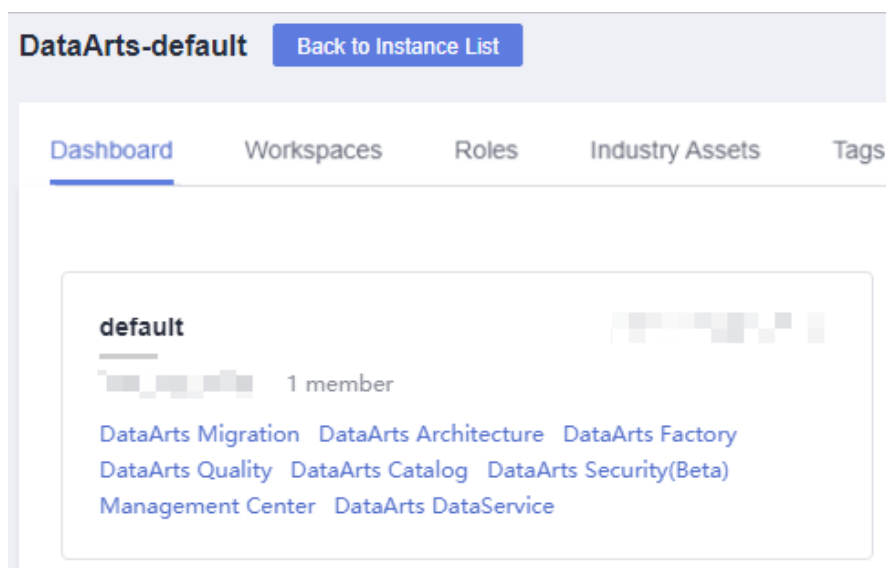
6. Click **OK**. The system starts to perform PatchData and the **PatchData Monitoring** page is displayed.

## Batch Job Monitoring: Adding a Job Label

Labels can be added to jobs to facilitate job instance filtering.

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

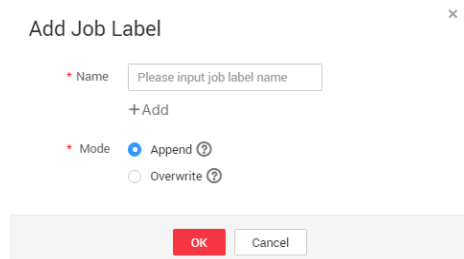
**Figure 3-193** DataArts Factory



2. In the left navigation pane of DataArts Factory, choose **Monitoring > Monitor Job**.

3. Click the **Batch Job Monitoring** tab.
4. In the **Operation** column of the job, choose **More > Add Job Label**.
5. In the **Add Job Label** dialog box displayed, set the job label parameters.

**Figure 3-194** Parameters for adding a job label



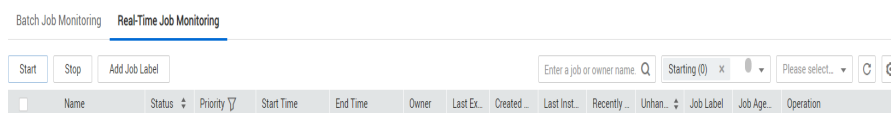
6. Click **OK**.

### 3.4.7.2.2 Monitoring a Real-Time Job

In the real-time processing mode, data is processed in real time, which is used in scenarios with high real-time performance. This type of job is a pipeline that consists of one or more nodes. You can configure scheduling policies for each node, and the tasks started by operators can keep running for an unlimited period of time. In this type of job, lines with arrows represent only service relationships, rather than task execution processes or data flows.


You can choose **Monitor Job** and click the **Real-Time Job Monitoring** tab to view the job status, start time, and end time, and perform the operations listed in [Table 3-142](#).

**Figure 3-195** Real-time job monitoring page



**Table 3-142** Operations supported by real-time job monitoring

No.	Operation	Description
1	Searching for a job based on the job name or owner	-
2	Filtering jobs based on the job status or job label	-
3	Perform operations on jobs in a batch	Select multiple jobs and perform operations on them.

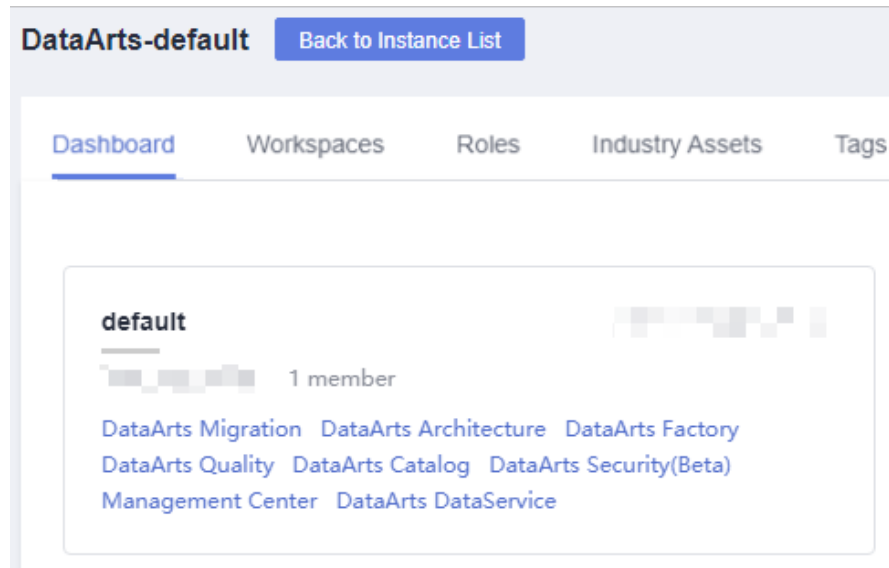
N o.	Operation	Description
4	Viewing job instance status	Click job in front of the  name. The <b>Last Instance</b> page is displayed. You can view information about the last instance of the job.
5	Job status-related operations	In the <b>Operation</b> column of the job, you can start, pause, recover, and stop job scheduling.
6	Adding a job label	In the <b>Operation</b> column of a job, choose <b>More &gt; Add Job Label</b> .
7	Viewing node information of a job	Click a job name. On the displayed page, click a node to view its associated job/scripts and monitoring information.  <b>NOTE</b> If event-driven scheduling is configured for a node in the job, the subjob monitoring page is displayed when you click the node.
8	Disabling and restoring a node	Click a job name. On the displayed page, right-click a node and select <b>Disable</b> . After the node is disabled, you can right-click it and select <b>Restore</b> to restore it on another location. For details, see <a href="#">Real-Time Job Monitoring: Disabling and Restoring a Node</a> .
9	Viewing the boot log	Click a job name. On the displayed page, right-click a node and select <b>View Run Log</b> to view logs of the node.
10	Configuring scheduling	Click a job name. On the displayed page, right-click the node where event-driven scheduling is configured and select <b>Configure Scheduling</b> to view and modify the scheduling information about the node. For details, see <a href="#">Real-Time Job Monitoring: Configuring Scheduling for a Node Where Event-driven Scheduling Is Configured</a> .
11	Monitoring subjobs	Click a job name. On the displayed page, click the node where event-driven scheduling is configured to go to the subjob monitoring page. For details, see <a href="#">Real-Time Job Monitoring: Monitoring Subjobs</a> .
12	Clearing stream messages	Click a job name. On the displayed page, right-click the node where event-driven scheduling is configured and select <b>Clear Stream Message</b> .

## Real-Time Job Monitoring: Disabling and Restoring a Node

You can disable a node in a real-time job and restore it in another location.

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-196** DataArts Factory



2. In the navigation pane on the left of the DataArts Factory page, choose **Monitoring > Monitor Job**.
3. On the **Real-Time Job Monitoring** tab page, click a job name.
4. On the displayed page, right-click the node and select **Disable**.
5. Right-click the node and choose **Resume** from the shortcut menu. The **Resume Node Running** dialog box is displayed, as shown in [Table 3-143](#).

**Table 3-143** Resumption parameters

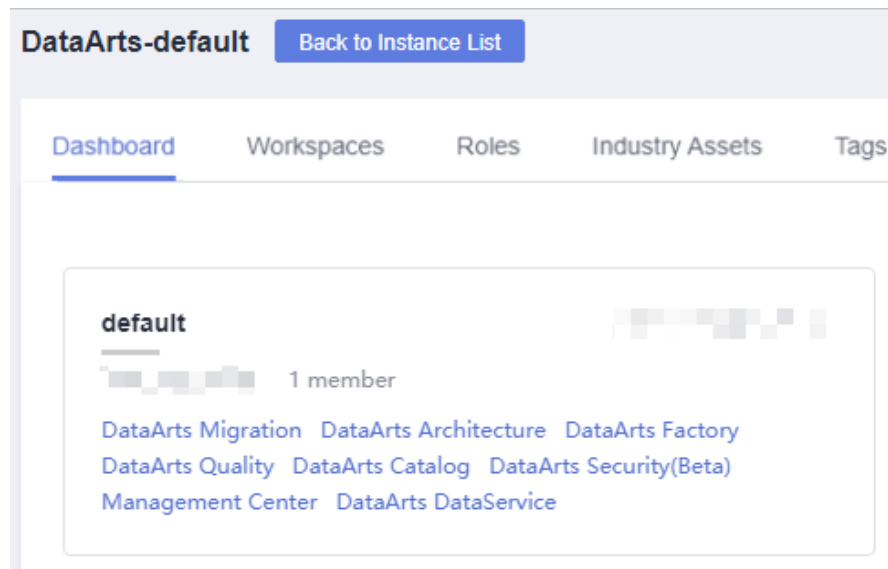
Parameter	Description
Last Paused	Start time when a node is suspended.
Tasks Not Run	Number of tasks that are not running during node suspension.
Run From	Parameters for performing the tasks generated during the pause period. Position from which running restarts. <ul style="list-style-type: none"> <li>• Paused node</li> <li>• The first node of the subjob</li> </ul>
Concurrent Tasks	Parameters for performing the tasks generated during the pause period. Number of tasks to be processed.
Task Name	Parameters for performing the tasks generated during the pause period. Task to be resumed.

## Real-Time Job Monitoring: Configuring Scheduling for a Node Where Event-driven Scheduling Is Configured

If event-driven scheduling is configured for a node in a real-time job, right-click the node on the job monitoring details page and choose **Configure Scheduling** from the shortcut menu to view and modify the scheduling information about the node.

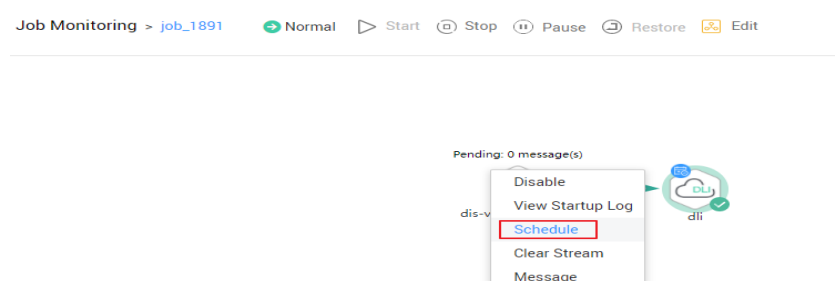
1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-197 DataArts Factory



2. In the navigation pane on the left of the DataArts Factory page, choose **Monitoring > Monitor Job**.
3. On the **Real-Time Job Monitoring** tab page, click a job name.
4. On the displayed page, right-click the node where event-driven scheduling is configured, select **Configure Scheduling**, and configure the parameters shown in [Table 3-144](#).

Figure 3-198 Configuring scheduling



**Table 3-144** Policy parameters

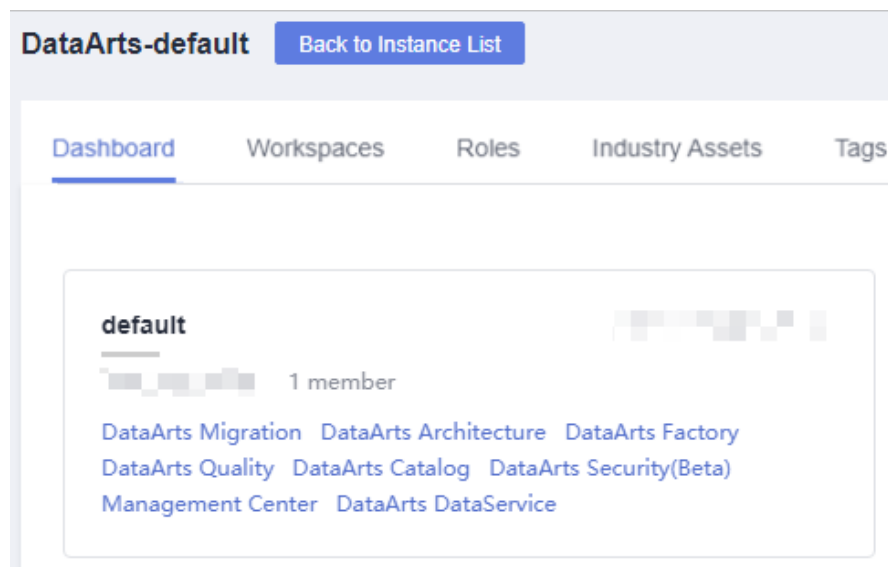
Parameter	Description
Concurrent Events	Number of jobs that can be concurrently processed. The maximum number of concurrent events is 10.
Event Detection Interval	Interval for event detection. The unit of the interval can be <b>Second</b> or <b>Minute</b> .
Failure Policy	Select a policy to be performed after scheduling fails. <ul style="list-style-type: none"> <li>• Stop scheduling</li> <li>• Ignore failure and proceed</li> </ul>

## Real-Time Job Monitoring: Monitoring Subjobs

When event-based scheduling is configured for a node in a job, you can click this node to query monitoring information of subjobs. On the **Subjob** page, you can stop, rerun, continue, and succeed subjobs as well as view subjob events.

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.


**Figure 3-199** DataArts Factory



2. In the navigation pane on the left of the DataArts Factory page, choose **Monitoring > Monitor Job**.
3. On the **Real-Time Job Monitoring** tab page, click a job name.
4. Click a node with event-based scheduling configured, [Table 3-145](#) describes the actions listed in the **Operation** column of each subjob.

**Table 3-145** Subjob monitoring operations

Operation	Description
Stop	Stops a subjob instance that is in the <b>Running</b> state.
Rerun	Reruns a subjob instance that is in the <b>Succeed</b> or <b>Failed</b> state.
Continue	If a subjob instance is in the <b>Abnormal</b> state, you can click <b>Continue</b> to begin running the subsequent nodes in the subjob instance.  <b>NOTE</b> This operation is allowed only when the <b>Failure Policy</b> of the node is set to Suspend current job execution plan.
Forcibly Succeed	Forcibly changes the status of a subjob instance from <b>Failed</b> to <b>Succeed</b> .
View Event	Displays the event content of a subjob.

5. Click  in the **Status** column. The running records of the subjob node are displayed.

[Table 3-146](#) describes the operations that can be performed on the node.

**Table 3-146** Node operations

Operation	Description
View Log	View the logs of the node.
More > Manual Retry	For a node in the <b>Failed</b> state, you can run the node again.  <b>NOTE</b> This operation is allowed only when the <b>Failure Policy</b> of the node is set to Suspend current job execution plan.
More > Succeed	Change the status of a node from <b>Failed</b> to <b>Succeed</b> .  <b>NOTE</b> This operation is allowed only when the <b>Failure Policy</b> of the node is set to Suspend current job execution plan.
More > Skip	Skip a node that is in the <b>To be run</b> or <b>Paused</b> state.
More > Pause	Pause a node that is in the <b>To be run</b> state. By doing so, the nodes following the paused node will not be run.
More > Resume	Resume a paused node.



### 3.4.7.3 Monitoring an Instance

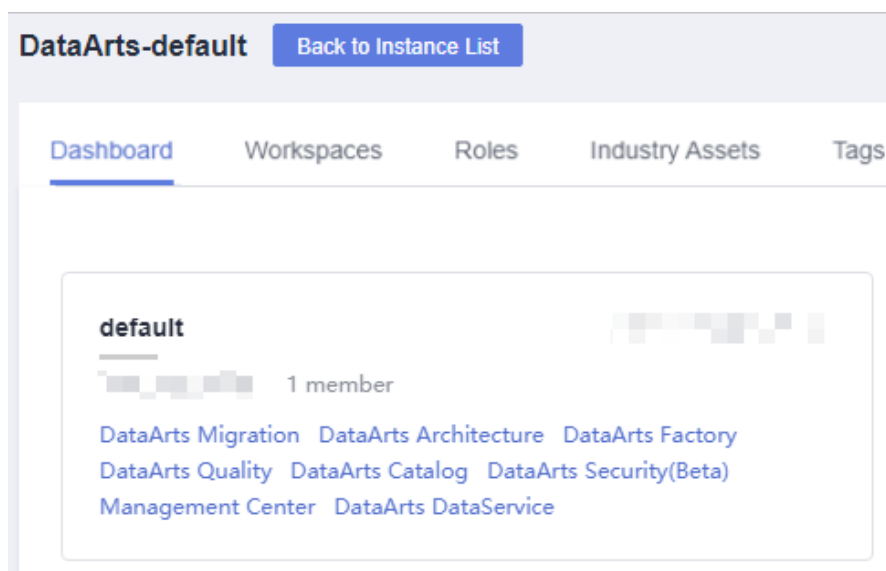
Each time a job is executed, a job instance record is generated. In the navigation pane of the DataArts Factory console, choose **Monitoring**. On the Monitor Instance page, you can view the job instance information and perform more operations on instances as required.

You can search for instances by **Job Name**, **Created By**, **CDM Job**, and **Node Type**. Search by CDM job is to search for job instances by node.

### Performing Job Instance Operations

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.


Figure 3-200 DataArts Factory



2. In the navigation tree on the left of the DataArts Factory page, choose **Monitoring > Monitor Instance**
3. You can stop, rerun, continue to run, or forcibly run jobs in batches. For details, see [Table 3-147](#).  
When multiple instances are rerun in batches, the sequence is as follows:
  - If a job does not depend on the previous schedule cycle, multiple instances run concurrently.
  - If jobs are dependent on their own, multiple instances are executed in serial mode. The instance that first finishes running in the previous schedule cycle is the first one to rerun.
4. [Table 3-147](#) describes the operations that can be performed on the instance.

**Table 3-147** Instance monitoring operations

Operation	Description
Searching for a job based on the job name or creator	If you select <b>Exact search</b> , exact search by job name is supported. If you do not select <b>Exact search</b> , fuzzy search by job name is supported.
Filtering jobs by CDM job or node type	-
Stop	Stop an instance that is in the <b>Waiting, Running, or Abnormal</b> state.
Rerun	Rerun a subjob instance that is in the <b>Succeed</b> or <b>Canceled</b> state. For details, see <a href="#">Rerunning Job Instances</a> .
View Waiting Job Instance	When the instance is in the waiting state, you can view the waiting job instance.
More > Continue	If an instance is in the <b>Abnormal</b> state, you can click <b>Continue</b> to begin running the subsequent nodes in the instance. <b>NOTE</b> This operation is allowed only when the <b>Failure Policy</b> of the node is set to Suspend current job execution plan.
More > Succeed	Forcibly change the status of an instance from <b>Abnormal, Canceled, or Failed</b> to <b>Succeed</b> .
More > View	Go to the job development page and view job information.

5. Click  in front of an instance. The running records of all nodes in the instance are displayed.
6. [Table 3-148](#) describes the actions that can be performed on the node.

**Table 3-148** Operations (node)

Operation	Description
View Log	View the log information of a node.
More > Manual Retry	To run a node again after it fails, click <b>Retry</b> . <b>NOTE</b> This operation is allowed only when the <b>Failure Policy</b> of the node is set to Suspend current job execution plan.

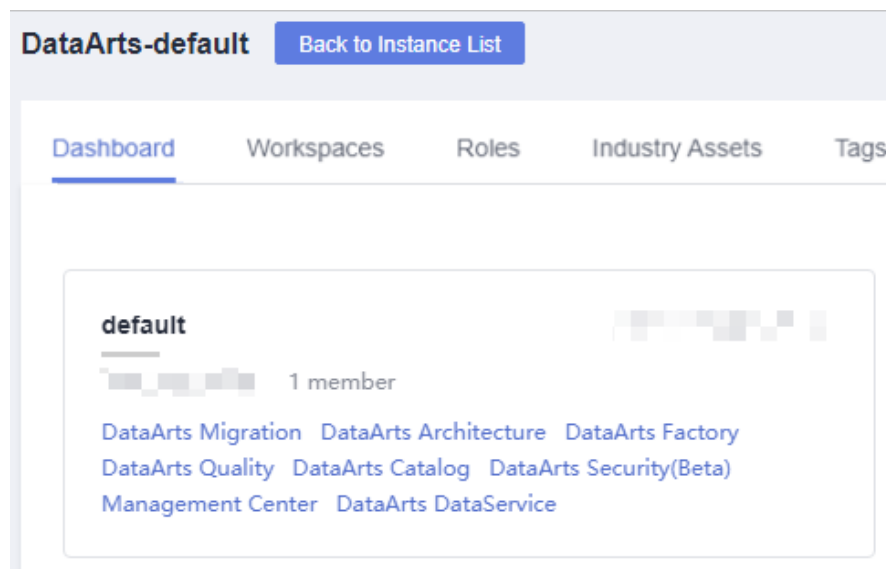
Operation	Description
More > Succeed	Change the status of a node from <b>Failed</b> to <b>Succeed</b> . <b>NOTE</b> This operation is allowed only when the <b>Failure Policy</b> of the node is set to Suspend current job execution plan.
More > Skip	To skip a node that is to be run or that has been paused, click <b>Skip</b> .
More > Pause	To pause a node that is to be run, click <b>Pause</b> . Nodes queued after the paused node will be blocked.
More > Resume	To resume a paused node, click <b>Resume</b> .

## Rerunning Job Instances

You can rerun a job instance that is successfully executed or fails to be executed by setting its rerun position.

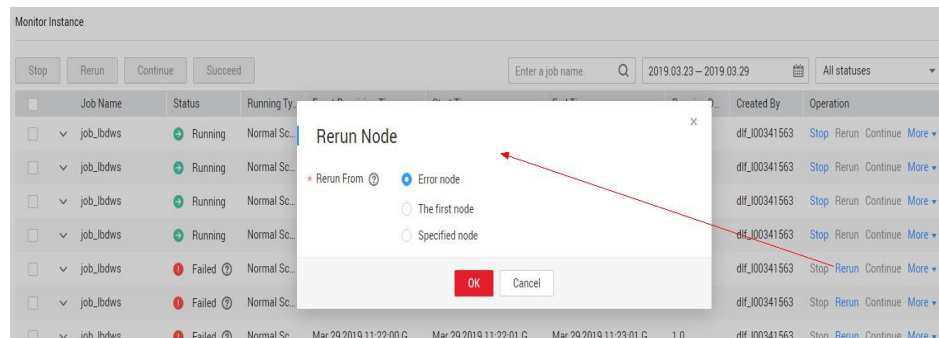
1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-201** DataArts Factory



2. In the navigation tree on the left of the DataArts Factory page, choose **Monitoring > Monitor Instance**
3. In the **Operation** column of a job, click **Rerun** to rerun the job instance. Alternatively, click the check box on the left of a job, and then click the **Rerun** button to rerun the job instance.

**Figure 3-202** Setting the rerunning position



**Table 3-149** Parameters for rerunning a job

Parameter	Description
Rerun Type	Type of the instance that you want to rerun. <ul style="list-style-type: none"> <li>• Rerun selected instance</li> <li>• Rerun instances of selected job and its upstream and downstream jobs</li> </ul>
Start Time	Time range in which instances have been run
List of Rerun Job Instances	Upstream and downstream jobs to rerun. You can select multiple jobs at a time.
Rerun From	Start position from which the job instance reruns. <ul style="list-style-type: none"> <li>• <b>Error node:</b> When a job instance fails to be run, it reruns since the error node of the job instance.</li> <li>• <b>The first node:</b> When a job instance fails to be run, it reruns since the first node of the job instance.</li> <li>• <b>Specified node:</b> When a job instance fails to run, it reruns since the node specified in the job instance. This option is available only if <b>Rerun Type</b> is set to <b>Rerun selected instance</b>.</li> </ul> <p><b>NOTE</b> A job instance reruns from its first node if either of the following cases occurs:</p> <ul style="list-style-type: none"> <li>• The quantity or name of a node in the job changes.</li> <li>• The job instance has been successfully run.</li> </ul>
Concurrent Instances	Number of job instances that can be concurrently processed.

### 3.4.7.4 Monitoring PatchData

In the navigation tree of the DataArts Factory console, choose **Monitoring > Monitor PatchData**.

On the PatchData Monitoring page, you can view the task status, service date, number of parallel periods, and PatchData job names, and stop a running task.

On the PatchData Monitoring page, click PatchData name. On the displayed page, you can view the PatchData execution status. For more information, see [Batch Job Monitoring: PatchData](#).

#### NOTE

- PatchData can be sorted by plan time, start time, and end time. Note that only one of the three sorting modes takes effect at a time.
- Click the sorting icon once to sort PatchData in ascending order, click the sorting icon twice to sort PatchData in descending order, and click the sorting icon three times to cancel sorting.

### 3.4.7.5 Managing Notifications

DataArts Studio uses Simple Message Notification (SMN) to send push notifications based on your subscription requirements, so that you can receive immediate notifications when a job encounters an exception or runs successfully.

#### 3.4.7.5.1 Managing a Notification

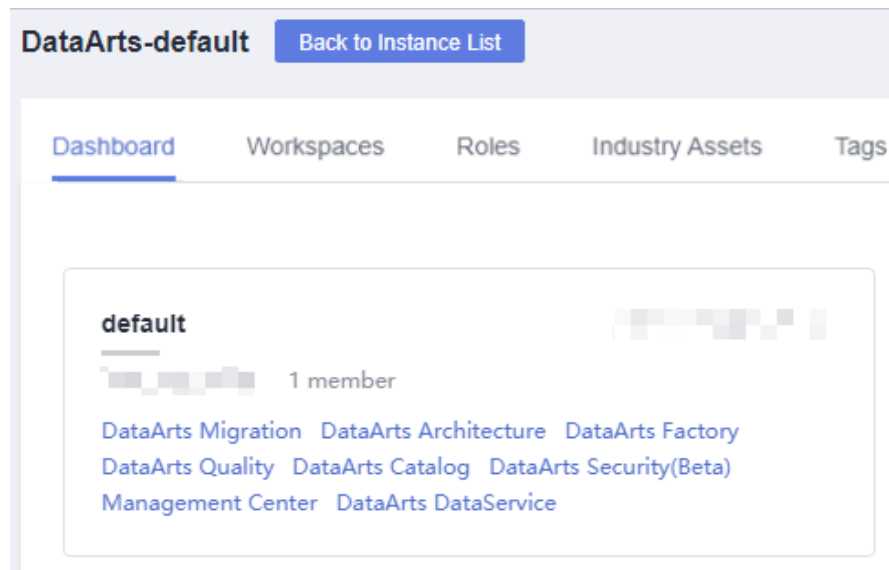
You can configure DLF to notify you of job success after it is performed.

### Configuring a Notification

Before configuring a notification for a job:

- Message notification has been enabled and a topic has been configured.
  - A job not in **Not Activated** status has been submitted.
1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-203** DataArts Factory



2. In the navigation pane on the DataArts Factory page, choose **Monitoring > Manage Notification**.
3. On the **Notification Management** tab page, click **Configure Notification**. In the displayed dialog box, configure parameters. [Table 3-150](#) describes the parameters.

**Table 3-150** Notification parameters

Parameter	Mandatory	Description
Notification Scope	Yes	Notification scope. Available options include: <ul style="list-style-type: none"> <li>• <b>One job:</b> Notifications are sent for a single job.</li> <li>• <b>All jobs:</b> Notifications are sent for all jobs.</li> </ul>
Job Name	Yes	Name of the job.

Parameter	Mandatory	Description
Notification Type	Yes	<p>Type of the notification.</p> <ul style="list-style-type: none"> <li>When <b>Notification Scope</b> is <b>One job</b>, available options for this parameter include: <ul style="list-style-type: none"> <li><b>Run abnormally/Fail:</b> When a job cannot run normally or fail to run, a notification is sent to notify the user of the abnormality.</li> <li><b>Run successfully:</b> When a job runs successfully, a notification is sent to notify the user of the success.</li> <li><b>Uncompleted:</b> This function supports only the jobs scheduled by day. If the job execution time is later than the configured time by which the job has not finished, a notification is sent.</li> <li><b>Busy resources:</b> If resources are busy during job execution, a notification is sent.</li> </ul> </li> <li>When <b>Notification Scope</b> is <b>All jobs</b>, available options for this parameter include: <ul style="list-style-type: none"> <li><b>Run abnormally/Fail:</b> When a job cannot run normally or fail to run, a notification is sent to notify the user of the abnormality.</li> <li><b>Busy resources:</b> If resources are busy during job execution, a notification is sent.</li> </ul> </li> </ul> <p><b>NOTE</b> For a real-time job, a notification is allowed to be sent only when the real-time job is in the <b>Run abnormally</b> or <b>Failed</b> state. For a batch job, a notification can be sent no matter when the batch job is in the <b>Run normally</b>, <b>Run abnormally</b>, or <b>Failed</b> state.</p>
Topic Name	Yes	<p>Select a notification topic.</p> <p><b>NOTE</b> Currently, only SMS, email, or HTTP are supported to subscribe to topics.</p>
Notification	Yes	<p>Whether to enable the notification function. The function is enabled by default.</p>

- Click **OK**.



## Editing a Notification

After a notification is created, you can modify the notification parameters as required.

1. In the navigation pane on the DataArts Factory page, choose **Monitoring > Manage Notification**.
2. Click the **Notification Management** tab.
3. In the **Operation** column of a notification, click **Edit**. In the displayed dialog box, edit notification parameters. [Table 3-150](#) describes the notification parameters.
4. Click **Yes**.

## Disabling a Notification

You can disable the notification function on the **Edit Notification** page or in the notification list.

1. In the navigation pane on the DataArts Factory page, choose **Monitoring > Manage Notification**.
2. Click the **Notification Management** tab.
3. In the **Notification Function** column, click . When it changes to , the notification function is disabled.

## Viewing a Notification

You can view all notification information on the **Notification Records** tab page.

1. In the navigation pane on the DataArts Factory page, choose **Monitoring > Manage Notification**.
2. Click the **Notification Records** tab.

## Deleting a Notification

If you do not need to use a notification any more, perform the following operations to delete it:

1. In the navigation pane on the DataArts Factory page, choose **Monitoring > Manage Notification**.
2. Click the **Notification Management** tab.
3. You can delete a notification in either of the following ways:
  - In the **Operation** column of a notification, click **Delete**.
  - Select the notifications to delete and click **Batch Delete** above the notification list.
4. In the displayed dialog box, click **OK**.



### 3.4.7.5.2 Cycle Overview

#### Scenarios

Notifications can be set to specified personnel by day, week, or month, allowing related personnel to regularly understand job scheduling information about the quantity of successfully/unsuccessfully scheduled jobs and failure details.

#### Constraints

This function depends on OBS.

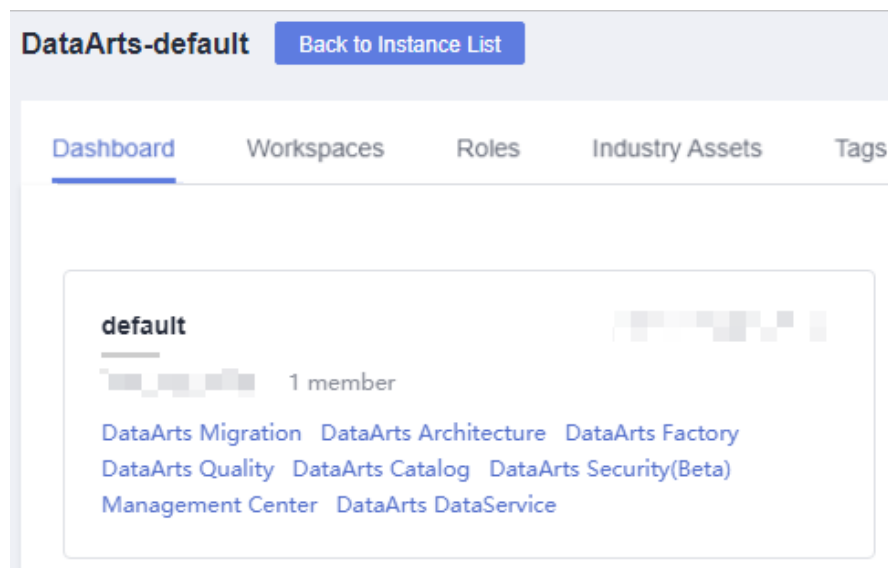
#### Prerequisites

- Simple Message Notification (SMN) has been enabled, topics have been configured, and subscriptions have been added to the topics.
- Jobs are not in **Not started** status and have been submitted.
- OBS has been enabled and a folder has been created in OBS.

#### Creating a Notification

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-204 DataArts Factory



2. In the navigation pane on the DataArts Factory page, choose **Monitoring > Manage Notification**.
3. On the **Cycles** tab page, click **Create Notification**. In the displayed dialog box, configure parameters. [Table 3-151](#) describes the notification parameters.

**Table 3-151** Notification parameters

Parameter	Mandatory	Description
Notification Name	Yes	Name of the notification to be sent.
Cycle	Yes	Interval for sending notifications, which can be set to <b>Daily</b> , <b>Weekly</b> , or <b>Monthly</b> . <b>NOTE</b> When <b>Cycle</b> is set to <b>Daily</b> , <b>Weekly</b> , or <b>Monthly</b> , a notification is sent every day, week, or month, and the notification content comes from the data generated from the last 24 hours, seven days, or 30 days.
Select Time	Yes	Time when the notification is sent. <ul style="list-style-type: none"> <li>• If <b>Cycle</b> is set to <b>Weekly</b>, the value can be any day or any several days from Monday to Sunday in a week.</li> <li>• If <b>Cycle</b> is set to <b>Monthly</b>, the value can be any day or any several days from 1st to 31st in a month.</li> </ul>
Start Time	Yes	Point in time when the notification is sent. The value can be accurate to hour or minute.
Topic	Yes	Select a notification topic from the drop-down list box.
OBS Bucket	Yes	Enter an OBS bucket in the text box or click <b>OBS</b> and select one from the displayed dialog box.
Notification	Yes	Specifies whether to enable the notification function. The function is enabled by default.

4. Click **OK**.
5. After the notification is created, you can perform the following operations on the notification:
  - Click **Edit**. In the **Create Notification** dialog box, edit the notification again.
  - Click **View Record**. In the **View Record** dialog box, view the job scheduling details.
  - Click **Delete**. In the **Delete Notification** dialog box, click **OK** to delete the notification.

### 3.4.7.6 Managing Backups

You can back up all jobs, scripts, resources, and environment variables on a daily basis.

You can also restore assets that have been backed up, including jobs, scripts, resources, and environment variables.

## Constraints

This function depends on OBS.

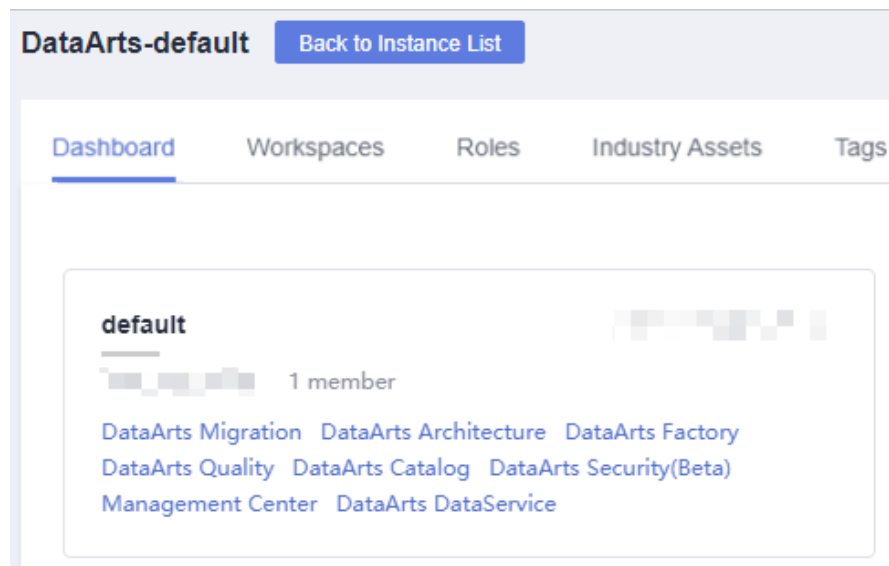
## Prerequisites

OBS has been enabled and a folder has been created in OBS.

## Backing Up Assets

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-205 DataArts Factory



2. In the navigation tree on the left, choose **Manage Backup**.
3. Click **Start Daily Backup**. In the **Browse OBS File** dialog box, select an OBS folder.

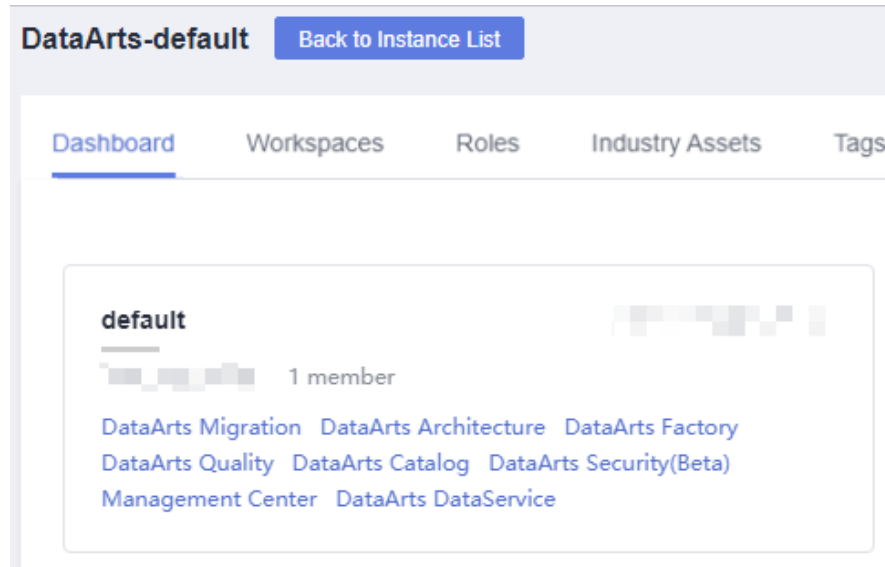
### NOTE

- Daily Backup starts at 00:00 every day to back up all jobs, scripts, resources, and environment variables of the previous day. The jobs, scripts, resources, and environment variables of the previous day are not backed up on the current day.
- If you select only the bucket name as the OBS storage path, the backup object is automatically stored in the folder named after the backup date. Environment variables, resources, scripts, and jobs are stored in the **1\_env**, **2\_resources**, **3\_scripts**, and **4\_jobs** folders, respectively.
- After the backup is successful, the **backup.json** file is automatically generated in the folder named after the backup date. The file stores job information based on the node type and can be modified before job restoration.
- To stop daily backup, click **Stop Daily Backup**.

## Restoring Assets

**Step 1** Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-206** DataArts Factory



**Step 2** In the navigation tree of the DataArts Factory console, choose **Manage Backup**.

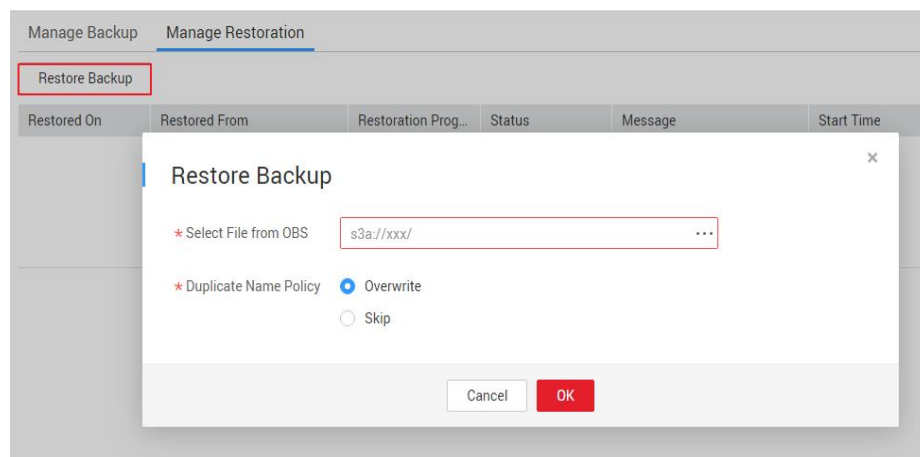
**Step 3** On the **Manage Restoration** tab, click **Restore Backup**.

In the **Restore Backup** dialog box, select the storage path of the asset to be restored from the OBS bucket and set the duplicate name policy.

### NOTE

- The storage path is the file path generated in [Backing Up Assets](#).
- Before restoring assets, you can modify the **backup.json** file in the backup path. You can change the connection name (connectionName), database name (database), and cluster name (clusterName).

**Figure 3-207** Restoring assets



**Step 4** Click **OK**.

----End

## 3.4.8 Configuration and Management

### 3.4.8.1 Configuring Resources

#### 3.4.8.1.1 Configuring Environment Variables

This topic describes how to configure and use environment variables.

### Application Scenario

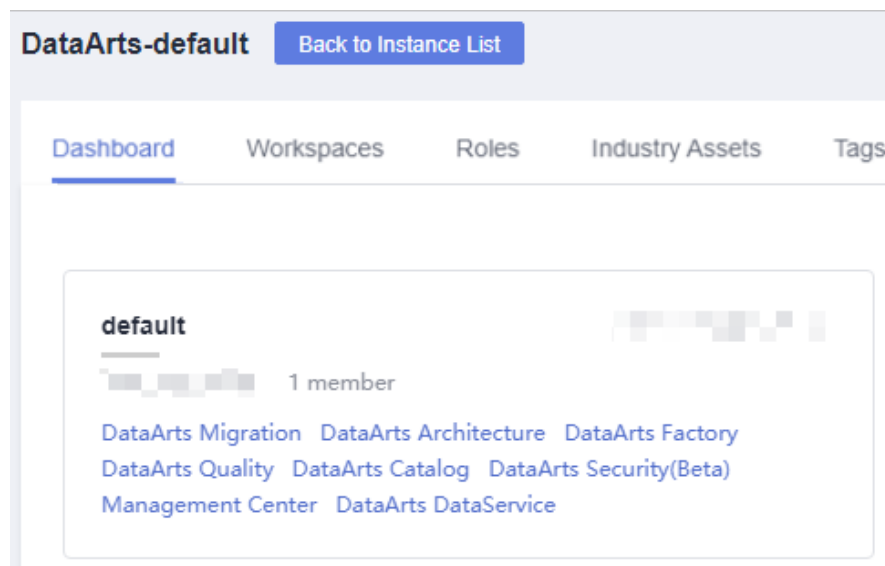
Configure job parameters. If a parameter belongs to multiple jobs, you can extract this parameter as an environment variable. Environment variables can be imported and exported.

### Importing Environment Variables

This function is available only if the OBS service is available. If OBS is unavailable, variables can be imported from the local PC.

**Step 1** Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-208** DataArts Factory



**Step 2** In the navigation tree on the left, choose **Specifications**.

**Step 3** Click **Environment Variables**. On the **Environment Variables** page, click **Import**.

**Step 4** In the **Import Environment Variable** dialog box, select the environment variable file that has been uploaded to OBS or a local directory and the duplicate name policy.

**Figure 3-209** Importing environment variables

Import Environment Variable x

\* File Location

\* Select File from OBS

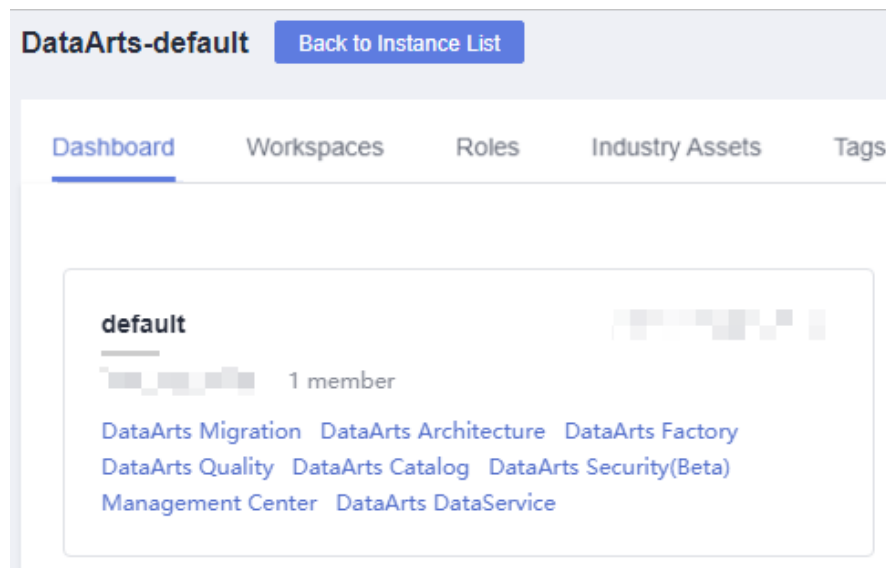
\* Duplicate Name Policy  Overwrite  
 Skip

----End

## Configuration Method

**Step 1** Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-210** DataArts Factory



**Step 2** In the navigation tree on the left, choose **Specifications**.

**Step 3** On the **Environment Variable** page, set the variables or constants listed in [Table 3-152](#) and click **Save**.

 **NOTE**



The difference between a variable and a constant lies in whether their values need to be reconfigured when they are imported to another workspace or project.

- The value of a variable (such as **workspace name**) varies depending on the workspace. When exporting a variable from a workspace and import it to another workspace, you must reconfigure its value.
- The value of a constant in different workspaces is the same. When importing a constant to another workspace, you do not need to reconfigure its value.

**Table 3-152** Configuring environment variables

Parameter	Mandatory	Description
Parameter	Yes	The parameter name must be unique, consist of 1 to 64 characters, and contain only letters, digits, underscores (_), and hyphens (-).
Value	Yes	Parameter values support constants and EL expressions but do not support system functions. For example, <b>123</b> and <b>abc</b> are supported. For details about how to use EL expressions, see <a href="#">Expression Overview</a> .

After configuring an environment variable, you can add, edit, or delete it.

- **Add:** Click **Add** to add an environment variable.
- **Edit:** If the parameter value is a constant, change the parameter value in the text box. If the parameter value is an EL expression, click  next to the text box to edit the EL expression. Click **Save**.
- **Delete:** Click  next to the parameter value text box to delete the environment variable.

----End

## How-Tos

The configured environment variables can be used in either of the following ways:

1. `${Environment variable}`
2. `#{Evn.get("environment variable")}`

## Example

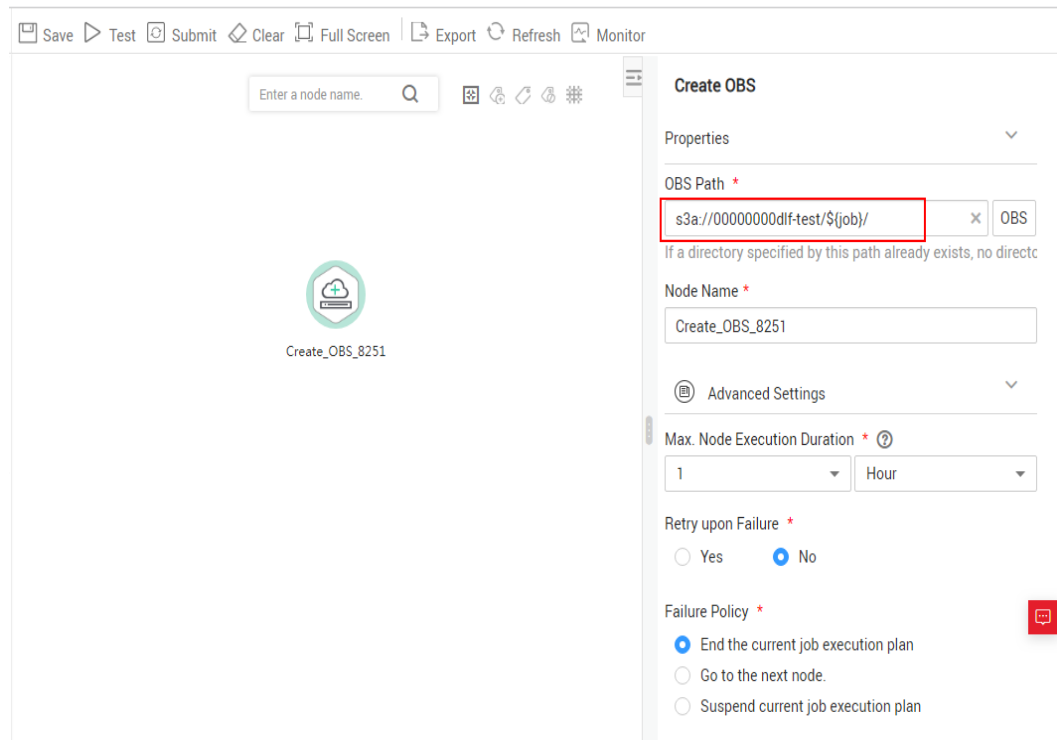
Context:

- A job named **test** has been created in the DataArts Factory module.
- An environment variable has been added. The parameter name is **job** and the parameter value is **123**.

**Step 1** Open **test** and drag a **Create OBS** node from the node library.

**Step 2** On the **Node Properties** tab page, configure the node properties.

**Figure 3-211** Create OBS



**Step 3** Click **Save** and then **Monitor** to monitor the running status of the job.

-----End

### 3.4.8.1.2 Configuring an OBS Bucket

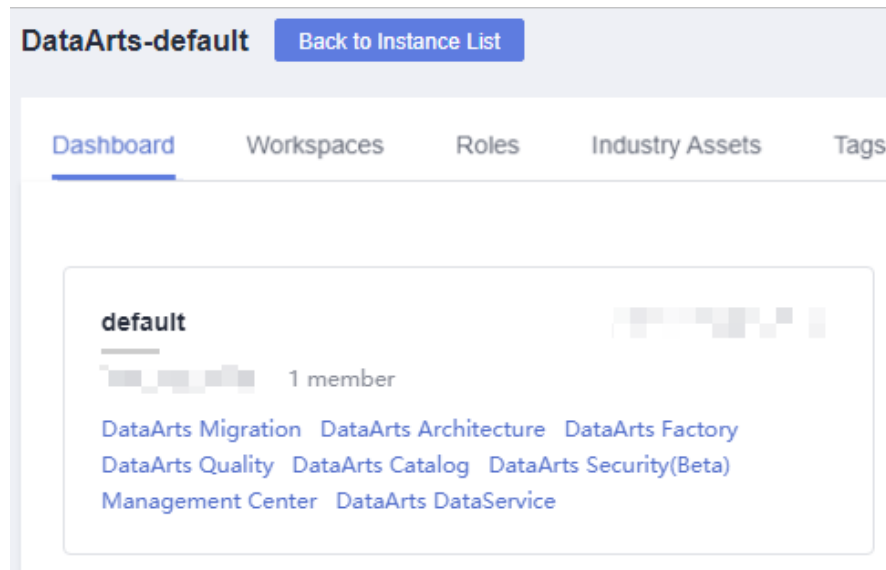
The execution history of scripts, jobs, and nodes is stored in OBS buckets. If no OBS bucket is available, you cannot view the execution history. This section describes how to configure an OBS bucket.

#### Procedure

**Step 1** Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.



Figure 3-212 DataArts Factory

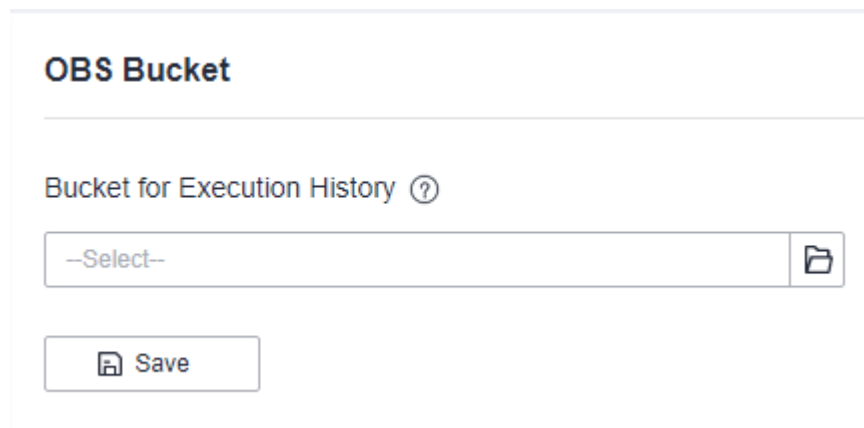


**Step 2** In the navigation pane, choose **Configuration > Configure**.

**Step 3** Choose **OBS Bucket**.

**Step 4** Select an OBS bucket.

Figure 3-213 Configuring an OBS bucket



**Step 5** Click **Save**.

----End

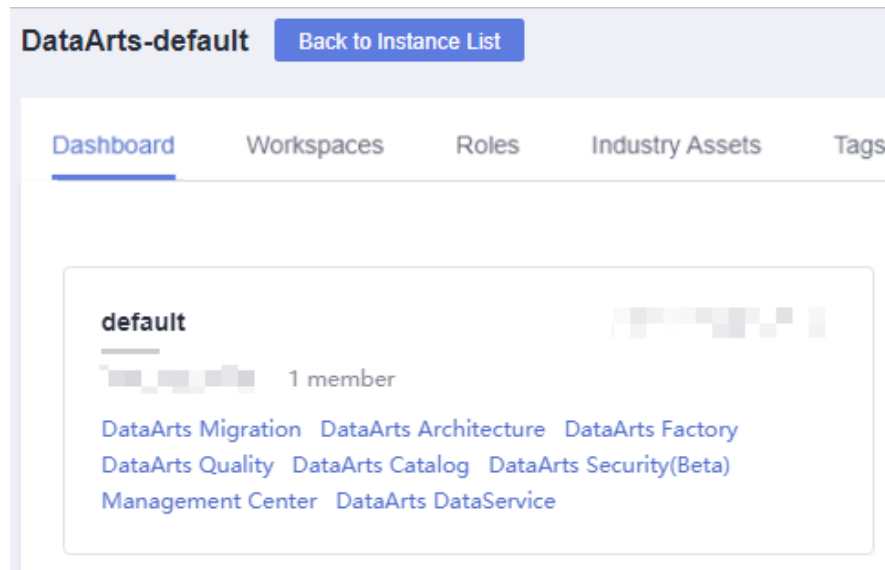
### 3.4.8.1.3 Managing Job Labels

Job labels are used to label jobs of the same or similar purposes to facilitate job management and query. This section describes how to manage job labels, including adding, modifying, and querying them.

## Configuration Method

**Step 1** Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-214 DataArts Factory



**Step 2** In the navigation tree on the left, choose **Specifications**.

**Step 3** Choose **Job Label**. On the **Manage Job Label** page, click **Add**, set the job name, and click **OK**.

**NOTE**

Up to 100 job labels can be created.

----End

### 3.4.8.1.4 Configuring Agencies

The following problems may occur during job execution in DataArts Factory:

- The job execution mechanism of the DataArts Factory module is to execute the job as the user who starts the job. For a job that is executed in periodic scheduling mode, if the IAM account used to start the job is deleted during the scheduling period, the system cannot obtain the user identity authentication information. As a result, the job fails to be executed.
- If a job is started by a low-privilege user, the job fails to be executed due to insufficient permissions.

To solve the preceding problems, configure an agency. When an agency is configured, the job interacts with other services as an agency during job execution to prevent job execution failures in the preceding scenarios.

## Role of an Agency

Cloud services interwork with each other, and some cloud services are dependent on other services. You can create an agency to delegate cloud services to access other services and perform resource O&M on your behalf.

## Agency Classification

Agencies are classified into workspace-level agencies and job-level agencies.

- Workspace-level agencies can be globally applied to all jobs in the workspace.
- Job-level agencies can only be applied to a single job.

The job-level agency has a higher priority than the workspace-level agency. If neither of them is configured, execute the job as the user who starts the job.

## Constraints

- To create or modify an agency, you must have the **Security Administrator** permissions.
- To configure a workspace-level agency, you must have the **DAYU Administrator** or **Tenant Administrator** policy.
- To configure a job-level agency, you must have the permission to view the list of agencies.

## Creating an Agency

1. Log in to the IAM console.
2. Choose **Agencies**. On the displayed page, click **Create Agency**.
3. Enter an agency name, for example, DataArts Studio\_agency.
4. Set **Agency Type** to **Cloud service** and select **DataArts Studio** for **Cloud Service** so that DataArts Studio can perform resource O&M operations on behalf of you.
5. Set **Validity Period** to **Unlimited**.

Figure 3-215 Creating an agency

The screenshot shows the 'Create Agency' form in the IAM console. The form is titled 'Agencies / Create Agency' and contains the following fields:

- Agency Name:** A text input field.
- Agency Type:** Two radio button options: 'Account' (selected) and 'Cloud service'. The 'Account' option has a sub-label 'Delegate another [redacted] account to perform operations on your resources.' The 'Cloud service' option has a sub-label 'Delegate a cloud service to access your resources in other cloud services.'
- Delegated Account:** A text input field with the placeholder text 'Specify a trusted account.'
- Validity Period:** A dropdown menu with 'Unlimited' selected.
- Description:** A text area with the placeholder text 'Enter a brief description.' and a character count '0/255' at the bottom right.

At the bottom of the form, there are two buttons: a red 'Next' button and a white 'Cancel' button.

6. Click **Assign Permissions** in the **Permissions** area.

7. On the displayed page, search for the **Tenant Administrator** policy, select it, and click **OK**. See [Figure 3-216](#).
  - Users assigned the **Tenant Administrator** policy have all permissions on all services except on IAMIAM. Therefore, delegate the **Tenant Administrator** policy to DataArts Studio so that DataArts Studio can access all related services.
  - If you want to meet the security control requirements for fewer permissions, you only need to configure the **OBS OperateAccess** permissions (During job execution, execution log information needs to be written to OBS. Therefore, you need to add the **OBS OperateAccess** permissions.) . Then, configure different agency permissions based on the node type in the job. For example, if a job contains only the **Import GES** node, you can configure the **GES Administrator** and **OBS OperateAccess** permissions. For details, see [Permissions Assignment](#).

**Figure 3-216** Assigning permissions



8. Click **OK**.

## Permissions Assignment

After the operation permissions of an account are delegated to DataArts Studio, you must configure the permissions of the agency identity so that DataArts Studio can interact with other services.

For purposes of permissions minimization, you can configure the **Admin** permissions for services based on the node types in jobs. For details, see [Table 3-153](#).

The **Admin** permissions can also be configured based on the operations, resources, and request conditions for a specific service. Based on the node types in jobs, permissions are defined by service APIs to allow for more fine-grained, secure access control of cloud resources. Configure the permissions according to [Table 3-154](#). For example, for a job containing the **Import GES** node, you only need to create a custom policy and select **ges:graph:getDetail** (viewing graph details), **ges:jobs:getDetail** (querying task status), and **ges:graph:access** (using graphs).

### NOTICE

- MRS-related nodes (MRS Presto SQL, MRS Spark, MRS Spark Python, MRS Flink Job, and MRS MapReduce) and directly connected nodes (MRS Spark SQL and MRS Hive SQL) do not support job submission in agency mode, therefore, jobs of these types cannot be configured with agencies.
- MRS clusters that support job submission in agency mode are as follows:
  - Non-security cluster
  - Security cluster whose version is later than 2.1.0 and which has MRS 2.1.0.1 or later

- Configure the service-level **Admin** permissions.  
During job execution, execution log information needs to be written to OBS. Therefore, the **OBS OperateAccess** permissions must be added for all jobs during coarse-grained authorization.

**Table 3-153** The **admin** permissions for related nodes

Node Name	System Permission	Description
CDM Job	DAYU Administrator	All DataArts Studio permissions
Import GES	GES Administrator	Permissions required to perform all operations on GES. This role depends on the <b>Tenant Guest</b> and <b>Server Administrator</b> roles in the same project.
<ul style="list-style-type: none"> <li>• MRS Presto SQL, MRS Spark, MRS Spark Python, MRS Flink Job, and MRS MapReduce</li> <li>• MRS Spark SQL and MRS Hive SQL (connecting to MRS clusters through MRS APIs)</li> </ul>	MRS Administrator KMS Administrator	<p>Users assigned the <b>MRS Administrator</b> role can perform all operations on MRS. This role depends on the <b>Tenant Guest</b> and <b>Server Administrator</b> roles in the same project.</p> <p>Users assigned the <b>KMS Administrator</b> role have the administrator permissions for encryption keys in DEW.</p>
MRS Spark SQL, MRS Hive SQL, MRS Kafka, and Kafka Client (connecting to the clusters in proxy mode)	DAYU Administrator KMS Administrator	<p><b>DAYU Administrator</b> has all permissions required for DataArts Studio.</p> <p>Users assigned the <b>KMS Administrator</b> policy have the administrator permissions for encryption keys in DEW.</p>
DLI Flink Job, DLI SQL, and DLI Spark	DLI Service Admin	All operation permissions for DLI.
DWS SQL, RDS SQL (connecting to data sources in proxy mode), and Shell	DAYU Administrator KMS Administrator	<p><b>DAYU Administrator</b> has all permissions required for DataArts Studio.</p> <p>Users assigned the <b>KMS Administrator</b> policy have the administrator permissions for encryption keys in DEW.</p>

Node Name	System Permission	Description
CSS	DAYU Administrator Elasticsearch Administrator	<b>DAYU Administrator</b> has all permissions required for DataArts Studio.  Users assigned the <b>Elasticsearch Administrator</b> policy have all permissions for CSS. This role depends on the <b>Tenant Guest</b> and <b>Server Administrator</b> roles in the same project.
Create OBS, Delete OBS, and OBS Manager	OBS OperateAccess	Basic object operation permissions, such as viewing buckets, uploading objects, obtaining objects, deleting objects, and obtaining object ACLs.
SMN	SMN Administrator	All operation permissions for SMN.

- Configure fine-grained permissions. (Create custom policies based on the actions supported by each service.)

For details on how to create a custom policy, see "Creating a Custom Policy" in the *Identity and Access Management User Guide*.

 **NOTE**

- During job execution, you must write execution logs to OBS. When the fine-grained authorization mode is used, the following OBS permissions need to be added for all types of jobs:
  - obs:bucket:GetBucketLocation
  - obs:object:GetObject
  - obs:bucket>CreateBucket
  - obs:object:PutObject
  - obs:bucket>ListAllMyBuckets
  - obs:bucket>ListBucket
- CDM Job nodes belong to the DataArts Studio module. DataArts Studio does not support fine-grained authorization. Therefore, only the **DataArts Studio Administrator** policy can be configured for jobs containing these types of nodes.
- CSS does not support fine-grained authorization and requires a proxy. Therefore, the **DataArts Studio Administrator** and **Elasticsearch Administrator** policies can be configured for jobs containing these nodes.
- SMN does not support fine-grained authorization. Therefore, jobs containing these nodes require the **SMN Administrator** permissions.

**Table 3-154** Creating a custom policy

Node Name	Action
Import GES	<ul style="list-style-type: none"> <li>● ges:graph:access</li> <li>● ges:graph:getDetail</li> <li>● ges:jobs:getDetail</li> </ul>
<ul style="list-style-type: none"> <li>● MRS Presto SQL, MRS Spark, MRS Spark Python, MRS Flink Job, and MRS MapReduce</li> <li>● MRS Spark SQL and MRS Hive SQL (connecting to MRS clusters through MRS APIs)</li> </ul>	<ul style="list-style-type: none"> <li>● mrs:job:delete</li> <li>● mrs:job:stop</li> <li>● mrs:job:submit</li> <li>● mrs:cluster:get</li> <li>● mrs:cluster:list</li> <li>● mrs:job:get</li> <li>● mrs:job:list</li> <li>● kms:dek:crypto</li> <li>● kms:cmk:get</li> </ul>
MRS Spark SQL, MRS Hive SQL, MRS Kafka, and Kafka Client (connecting to the clusters in proxy mode)	<ul style="list-style-type: none"> <li>● kms:dek:crypto</li> <li>● kms:cmk:get</li> <li>● DataArts Studio Administrator (role)</li> </ul>
DLI Flink Job, DLI SQL, and DLI Spark	<ul style="list-style-type: none"> <li>● dli:jobs:get</li> <li>● dli:jobs:update</li> <li>● dli:jobs:create</li> <li>● dli:queue:submit_job</li> <li>● dli:jobs:list</li> <li>● dli:jobs:list_all</li> </ul>
DWS SQL, RDS SQL (connecting to data sources in proxy mode), and Shell	<ul style="list-style-type: none"> <li>● kms:dek:crypto</li> <li>● kms:cmk:get</li> <li>● DataArts Studio Administrator (role)</li> </ul>
Create OBS, Delete OBS, and OBS Manager	<ul style="list-style-type: none"> <li>● obs:bucket:GetBucketLocation</li> <li>● obs:bucket:ListBucketVersions</li> <li>● obs:object:GetObject</li> <li>● obs:bucket:CreateBucket</li> <li>● obs:bucket&gt;DeleteBucket</li> <li>● obs:object&gt;DeleteObject</li> <li>● obs:object:PutObject</li> <li>● obs:bucket&gt;ListAllMyBuckets</li> <li>● obs:bucket:ListBucket</li> </ul>

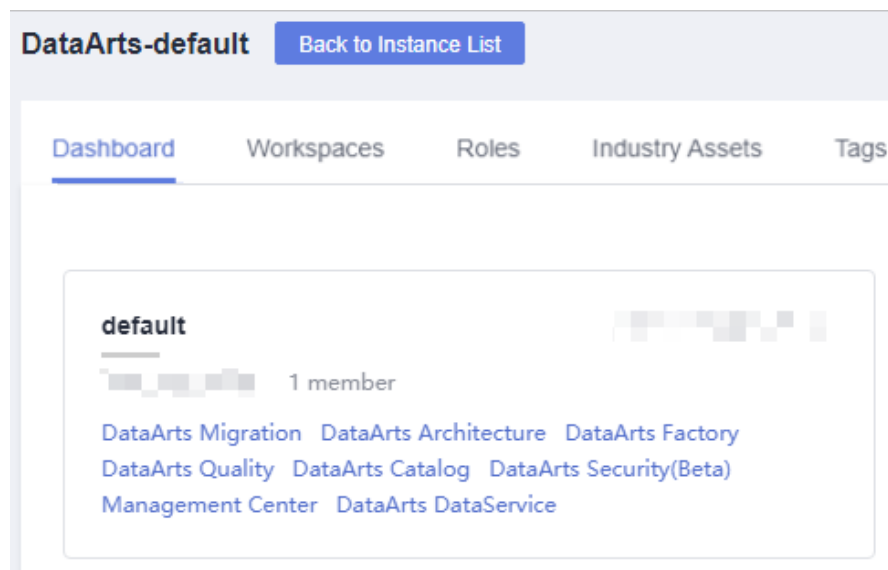
## Configuring a Workspace-Level Agency

**CAUTION**

A workspace-level agency impacts on all jobs. Some jobs contain nodes related to MRS. Exercise caution when performing this operation.

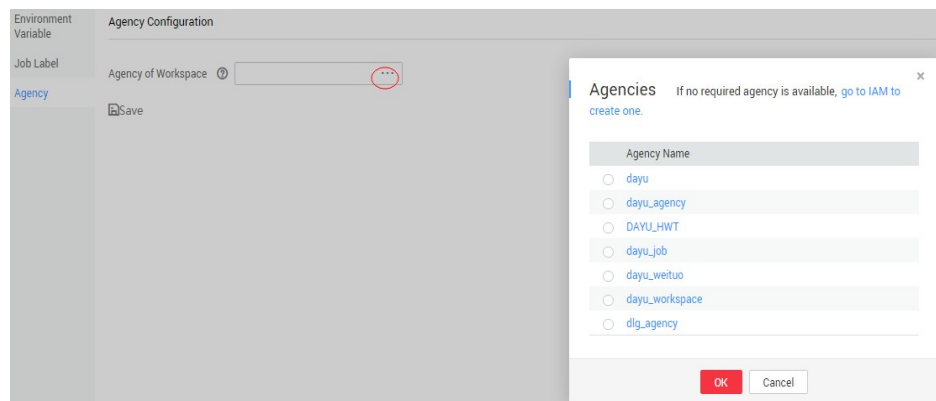
1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-217 DataArts Factory



2. In the navigation pane, choose **Configuration > Configure**.
3. Click **Agency**. On the displayed page, configure an agency.
4. You can select an agency from the agency list or create a new one. For details on how to create an agency and configure permissions, see [Creating an Agency](#).

Figure 3-218 Configuring a workspace-level agency





5. Click **OK** to return to the **Agency Configuration** page. Then, click  to save the settings.

## Configuring a Job-level Agency

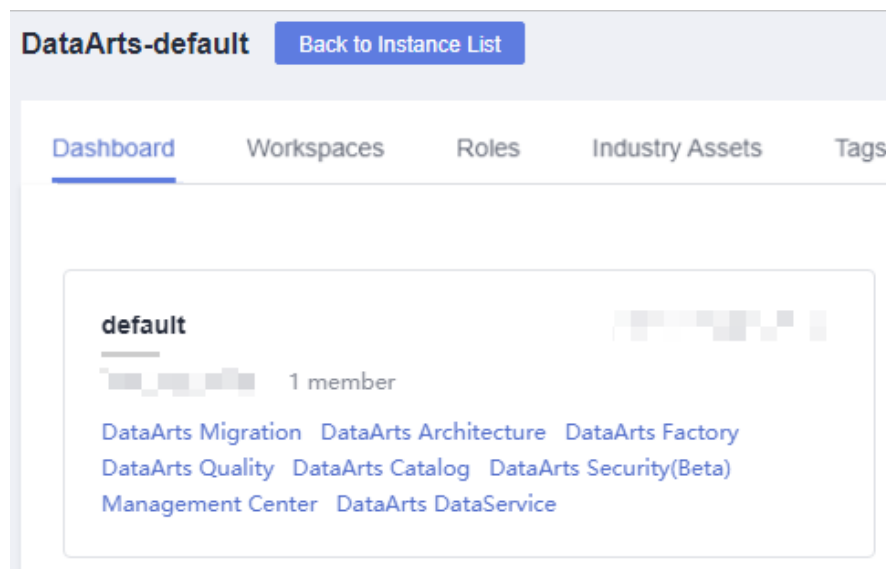
### NOTE

You can create a job-level agency when creating a job. You can also modify the agency of an existing job.

### Configuring an agency when creating a job

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-219** DataArts Factory



2. In the navigation pane of the DataArts Factory homepage, choose **Development > Develop Job**.
3. Right-click the job directory and choose **Create Job** from the shortcut menu. The **Create Job** dialog box is displayed. If a workspace-level agency has been configured, it is used for the job by default. You can also select another agency from the agency list.

**Figure 3-220** Configuring an agency for a job

### Create Job

A maximum of 10000 jobs can be created. You can create 9989 more jobs.

\* Job Name

\* Processing Mode  Batch processing  Real-time processing

\* Creation Method

\* Select Directory

Owner

Priority  High  Medium  Low

Agency

\* Log Path   
[To change the log path, go to the DAYU space management page.](#)  
[For details, see the documentation.](#)

#### Modifying the agency of an existing job

1. In the navigation pane of the DataArts Factory homepage, choose **Development > Develop Job**.
2. In the job directory, double-click an existing job. On the far right of the displayed page, click **Basic Info**. The dialog box of the job's basic settings is displayed. If a workspace-level agency has been configured, it is used by default. You can also select another agency from the agency list.

#### 3.4.8.1.5 Configuring a Default Item

This section describes how to configure a default item.

#### Scenario

If a parameter is invoked by multiple jobs, you can use this parameter as the default configuration item. In this way, you do not need to set this parameter for each job.

#### Configuring Periodic Scheduling

To configure the default action on the current job when the job it depends on fails, perform the following operations:

**Step 1** In the navigation pane, choose **Configuration > Specifications**.

**Step 2** Choose **Default Configuration**.

 **NOTE**

Three options are available. The default value is **Terminate**.

- **Suspend**: The current job is suspended.
- **Continue**: The current job continues to be executed.
- **Terminate**: The current job is terminated.

**Step 3** Click **Save** to save the settings.

----End

## Configuring the Multi-IF Policy

To configure the policy for executing nodes with multiple IF conditions, perform the following operations:

**Step 1** In the navigation pane, choose **Configuration > Specifications**.

**Step 2** Choose **Default Configuration**.

 **NOTE**

The following two options are available:

- **OR**: Nodes are executed if an IF condition is met.
- **AND**: Nodes are executed if all IF conditions are met.

For details, see [Configuring the Policy for Executing a Node with Multiple IF Statements](#).

**Step 3** Click **Save** to save the settings.

----End

## Configuring the Hard and Soft Lock Policy

The policy determines how you can grab the lock of a job or script. If you use a soft lock, you can grab the lock of a job or script regardless of whether you have the lock. If you use a hard lock, you can only unlock or grab the lock of a job or script for which you have the lock. Operations such as publish, execution, and scheduling are not restricted by locks.

You can configure the hard/soft policy based on your needs.

**Step 1** In the navigation pane, choose **Configuration > Specifications**.

**Step 2** Choose **Default Configuration**.

 **NOTE**

The default policy is **Soft Lock**.

- **Soft lock**: You can lock or unlock jobs or scripts, regardless of whether they are locked by others.
- **Hard Lock**: You can lock jobs or scripts only after they have been unlocked by other users. The space administrator and the **DAYU Administrator** user can lock and unlock jobs or scripts without any limitations.

**Step 3** Click **Save** to save the settings.

----End

### 3.4.8.2 Managing Resources

You can upload custom code or text files as resources on Manage Resource and schedule them when running nodes. Nodes that can invoke resources include DLI Spark, MRS Spark, DLI Flink Job, and MRS MapReduce.

After creating a resource, configure the file associated with the resource. Resources can be directly referenced in jobs. When the resource file is changed, you only need to change the resource reference location. You do not need to modify the job configuration. For details about resource usage examples, see [Developing a DLI Spark Job](#).

### Constraints

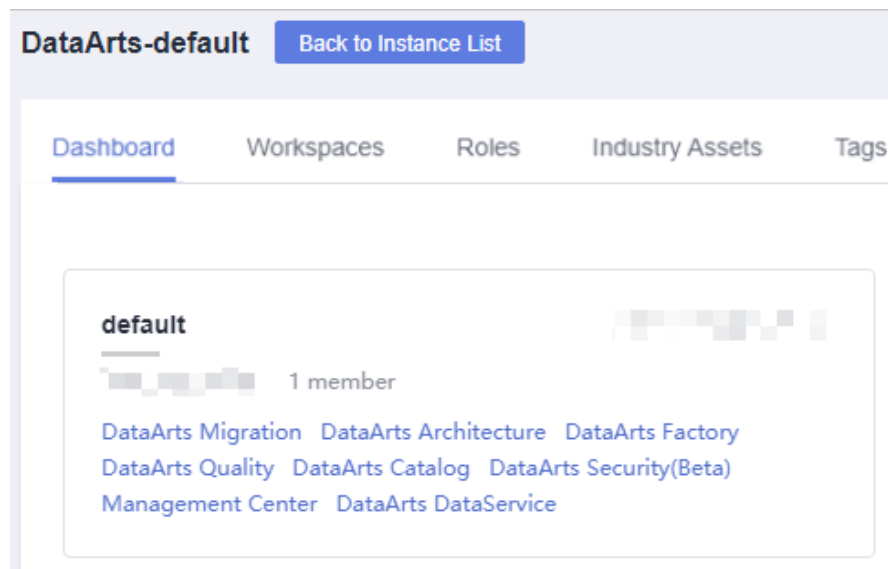
This function depends on OBS or MRS HDFS.


### (Optional) Creating a Directory

If a directory exists, you do not need to create one.

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-221** DataArts Factory



2. In the left navigation pane, choose **Configuration > Manage Resource**.
3. In the directory list, click . In the displayed dialog box, configure directory parameters. [Table 3-155](#) describes the directory parameters.

**Table 3-155** Resource directory parameters

Parameter	Description
Directory Name	Name of the resource directory. The name must contain 1 to 32 characters, including only letters, numbers, underscores (_), and hyphens (-).
Select Directory	Parent directory of the resource directory. The parent directory is the root directory by default.

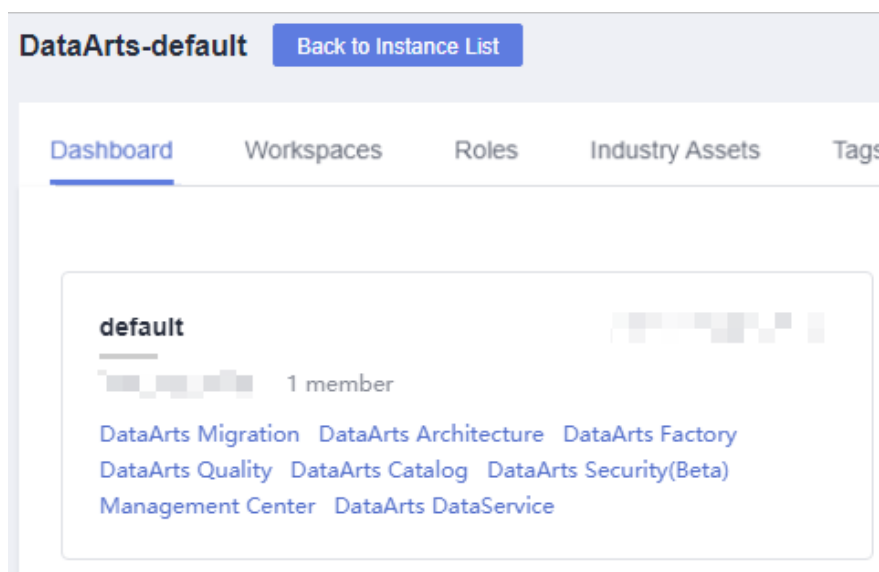
4. Click **OK**.

## Creating a Resource

You have enabled OBS before creating a resource.

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-222** DataArts Factory



2. In the left navigation pane, choose **Configuration > Manage Resource**.
3. Click **Create Resource**. In the displayed dialog box, configure resource parameters. [Table 3-156](#) describes the resource parameters. Click **OK**.

**Table 3-156** Resource management parameters

Parameter	Man datory	Description
Name	Yes	Name of the resource. The name must contain 1 to 32, including only letters, numbers, underscores (_), and hyphens (-).

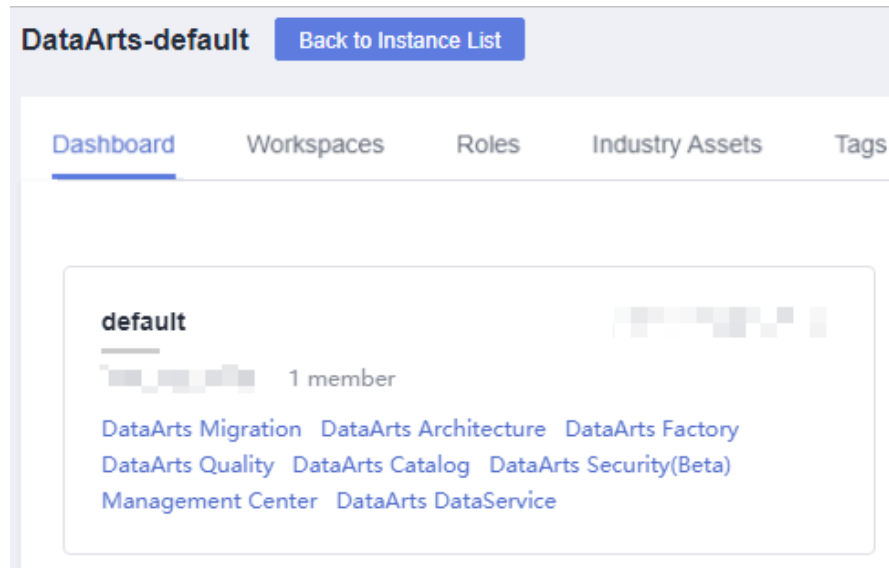
Parameter	Mandatory	Description
Type	Yes	File type of the resource. Possible values: <ul style="list-style-type: none"><li>• jar: JAR file</li><li>• pyFile: User Python file</li><li>• file: User file</li><li>• archive: User AI model file</li></ul>
Resource Location	Yes	Location of the resource. OBS and HDFS are supported. HDFS supports only MRS Spark, MRS Flink Job and MRS MapReduce nodes.
Main JAR package	Yes	<ul style="list-style-type: none"><li>• If <b>Resource Location</b> is <b>OBS</b>, select the main JAR package that has been uploaded to OBS.</li><li>• If <b>Resource Location</b> is <b>HDFS</b>, select the main JAR package that has been uploaded to HDFS.</li></ul>
Depended JAR Package	No	Depended JAR package that has been uploaded to OBS. This parameter is required when <b>Type</b> is set to <b>jar</b> and <b>Resource Location</b> is set to <b>OBS</b> or <b>HDFS</b> .
Select Resource	Yes	Specific resource file.
Storage Path	Yes	Path to a directory where the resource is stored. This parameter is required only when <b>Resource Location</b> is set to <b>Local</b> .
Description	No	Descriptive information about the resource.
Select Directory	Yes	Directory to which the resource belongs. The root directory is selected by default.

## Editing a Resource

After a resource is created, you can modify resource parameters.

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-223 DataArts Factory



2. In the left navigation pane, choose **Configuration > Manage Resource**.
3. In the **Operation** column of the resource, click **Edit**. In the displayed dialog box, modify the resource parameters. For details, see [Table 3-156](#).
4. Click **OK**.

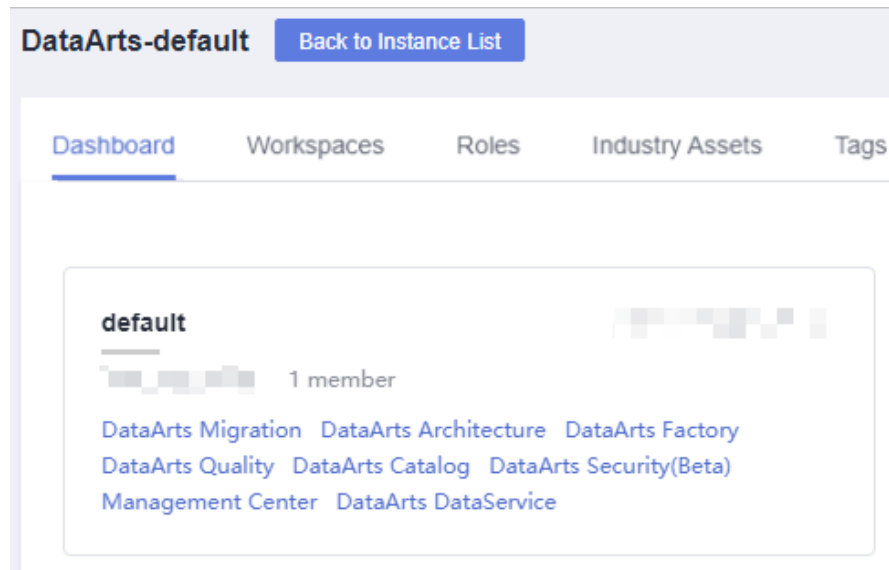
## Deleting a Resource

You can delete resources that are no longer needed.

Before deleting a resource, ensure that it is not used by any jobs. When you delete a resource, the system checks the jobs that are referencing the resource. The **Version** column in the reference list indicates the job versions that are referencing the resource. After you click **Delete**, the job will be deleted as well as all version information about the job.

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-224 DataArts Factory



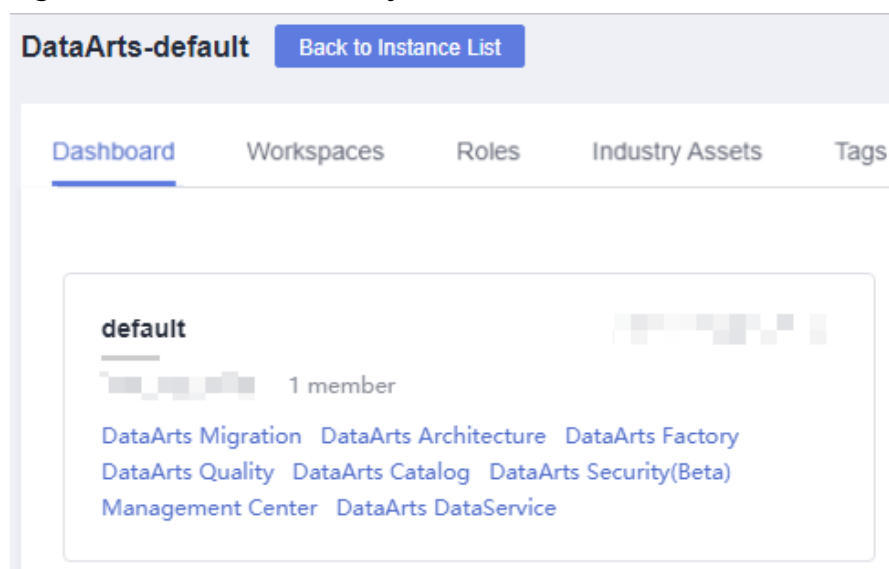
2. In the left navigation pane, choose **Configuration > Manage Resource**.
3. In the **Operation** column of the resource, click **Delete**. The **Delete Resource** dialog box is displayed.
4. Click **Yes**.


## Importing a Resource

To import a resource, perform the following operations:

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-225 DataArts Factory



2. In the left navigation pane, choose **Configuration > Manage Resource**.
3. In the resource directory, click  and select **Import Resource**. The **Import Resource** dialog box is displayed.



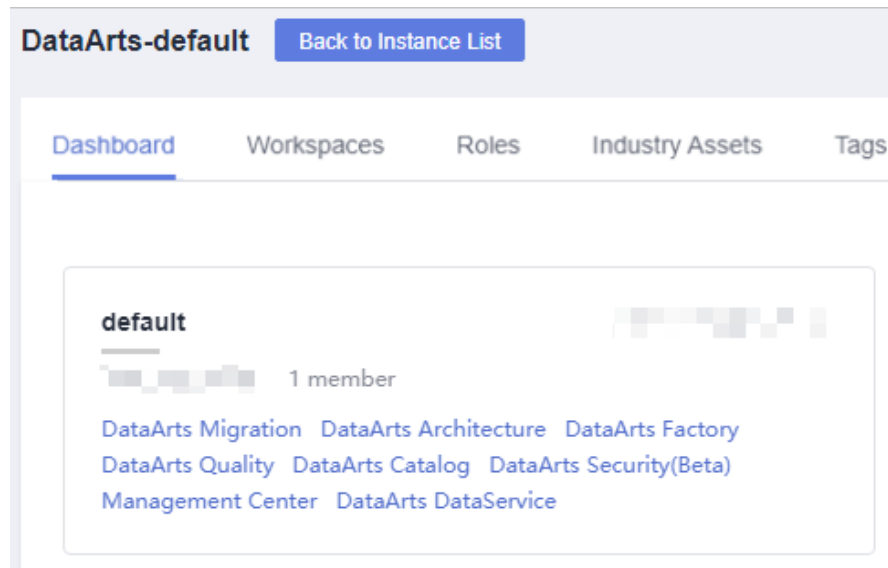
4. Select the resource file that has been uploaded to OBS and click **Next**. After the import is complete, click **Close**.

## Exporting a Resource

To export a resource, perform the following operations:

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

**Figure 3-226** DataArts Factory



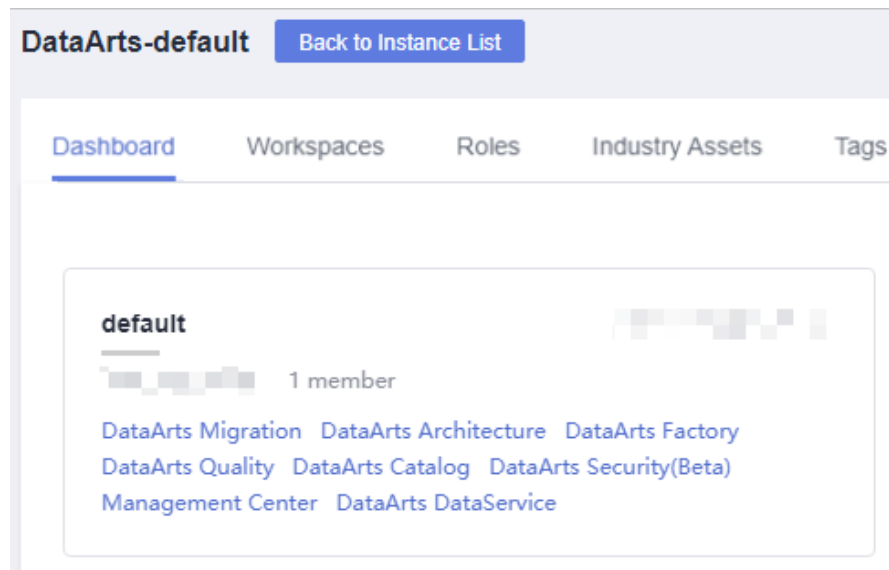
2. In the left navigation pane, choose **Configuration > Manage Resource**.
3. In the resource directory, select a resource, click **⋮**, and select **Export Resource**. The system starts downloading the resource to the local PC.

## Viewing Resource References

To view the references of a resource, perform the following operations:

1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-227 DataArts Factory



2. In the left navigation pane, choose **Configuration > Manage Resource**.
3. Right-click a resource in the list and select **View Reference**.
4. In the displayed **Reference List** dialog box, view the references of the resource.

## 3.4.9 Node Reference

### 3.4.9.1 Node Overview

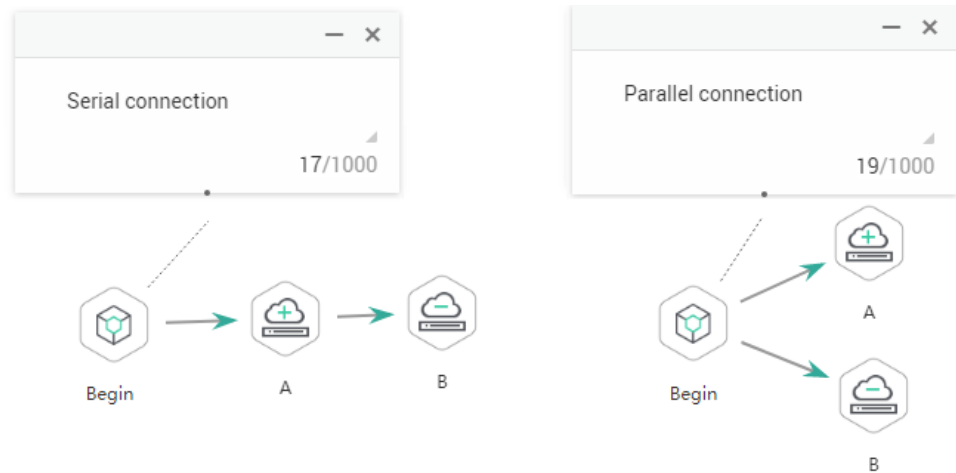
A node defines the operations performed on data. DataArts Factory provides nodes used for data integration, computing and analysis, database operations, and resource management. You can choose your desired nodes

- Node parameters can be presented using Expression Language (EL). For details about how to use EL, see [Expression Overview](#).
- Nodes cannot be connected in serial or parallel mode.

Serial connection: Nodes are run one by one. Specifically, node B runs only after node A is finished running.

Parallel connection: Nodes are run at the same time.

**Figure 3-228** Connection diagram



### 3.4.9.2 CDM Job

#### Function

The CDM Job node is used to run a predefined CDM job for data migration.

#### Parameters

[Table 3-157](#), [Table 3-158](#), and [Table 3-159](#) describe the parameters of the CDM Job node. Configure the lineage to identify the data flow direction, which can be viewed in the DataArts Catalog module.

**Table 3-157** Parameters of CDM Job nodes

Parameter	Mandatory	Description
CDM Cluster Name	Yes	<p>Name of the CDM cluster to which the CDM job to be executed belongs.</p> <p>You can select two CDM clusters to improve job reliability.</p> <ul style="list-style-type: none"> <li>If you select two clusters, the first one is the active cluster, and the second one is the standby cluster. Jobs run on the active cluster by default. If the active cluster is abnormal, jobs are migrated to the standby cluster.</li> <li>If you select two clusters, you are advised to set <b>Job Type</b> to <b>Existing jobs</b> rather than <b>New jobs</b> and ensure that the job exists in both the active and standby clusters. You can create a CDM job in the active cluster, export it, and import it to the standby cluster to implement job synchronization. For details, see <a href="#">Exporting and Importing CDM Jobs in Batches</a>.</li> </ul>
Job Type	Yes	<ul style="list-style-type: none"> <li>Existing jobs</li> <li>New jobs</li> </ul> <p><b>NOTE</b></p> <ul style="list-style-type: none"> <li>If <b>Job Type</b> is <b>Existing jobs</b>, the job node is not updated when the CDM job is modified. To update the job node, save the job where the node is located again to trigger a CDM job update.</li> <li>If <b>Job Type</b> is <b>New jobs</b>, the system checks whether a CDM job with the same name is running. <ul style="list-style-type: none"> <li>If the CDM job is not running, update the job with the same name based on the request body.</li> <li>If a CDM job with the same name is running, update the job after the job is run. During this period, the job may be started by other tasks. As a result, the extracted data may not be the same as expected (for example, the job configuration is not updated, or the macro of the running time is not correctly replaced). Therefore, do not create multiple jobs with the same name.</li> </ul> </li> </ul>
CDM Job Name	No	<p>This parameter is required only when <b>Job Type</b> is set to <b>Existing jobs</b>. Name of the CDM job to be executed.</p> <p>If the CDM job uses the <a href="#">job parameters</a> or <a href="#">environment variables</a> configured during data development, data can be indirectly migrated based on the parameters or variables during node scheduling in the DataArts Factory module.</p>

Parameter	Mandatory	Description
CDM Job Message Body	No	This parameter is required only when <b>Job Type</b> is set to <b>New jobs</b> . Enter the JSON message body of the CDM job. For convenience, you can choose <b>More &gt; View Job JSON</b> in the <b>Operation</b> column of an existing CDM job, copy the JSON content, and modify the content here.  If the CDM job uses the <b>job parameters</b> or <b>environment variables</b> configured during data development, data can be indirectly migrated based on the parameters or variables during node scheduling in the DataArts Factory module.
Node Name	Yes	Name of the node. The value must consist of 1 to 128 characters and contain only letters, digits, and the following special characters: _-/<>


**Table 3-158** Advanced parameters



Parameter	Mandatory	Description
Node Status Polling Interval (s)	Yes	Specifies how often the system check completeness of the node task. The value ranges from 1 to 60 seconds.
Max. Node Execution Duration	Yes	Execution timeout interval for the node. If retry is configured and the execution is not complete within the timeout interval, the node will not be retried and is set to the failed state.
Retry upon Failure	Yes	Whether to re-execute a node if it fails to be executed. <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> <li>- Maximum Retries</li> <li>- Retry Interval (seconds)</li> </ul> </li> <li>• <b>No:</b> The node will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b></p> <ul style="list-style-type: none"> <li>• If <b>Max. Node Execution Duration</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</li> <li>• If parameter transfer is used for scheduling the CDM job, do not configure parameter <b>Retry upon Failure</b> in the CDM job.</li> </ul>

Parameter	Mandatory	Description
Failure Policy	Yes	<p>Operation that will be performed if the node fails to be executed. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>
Dry run	No	If you select this option, the node will not be executed, and a success message will be returned.




**Table 3-159** Lineage

Parameter	Description
Input	

Parameter	Description
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● <b>DWS</b> <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>- <b>Schema:</b> Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● <b>OBS</b> <ul style="list-style-type: none"> <li>- <b>Path:</b> Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● <b>CSS</b> <ul style="list-style-type: none"> <li>- <b>Cluster Name:</b> Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>- <b>Index:</b> Enter a CSS index name.</li> </ul> </li> <li>● <b>HIVE</b> <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● <b>CUSTOM</b> <ul style="list-style-type: none"> <li>- <b>Name:</b> Enter a name of the CUSTOM type.</li> <li>- <b>Attribute:</b> Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● <b>DLI</b> <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.

Parameter	Description
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the input lineage.
<b>Output</b>	
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● DWS <ul style="list-style-type: none"> <li>- <b>Connection Name</b>: Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>- <b>Database</b>: Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>- <b>Schema</b>: Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>- <b>Table Name</b>: Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● OBS <ul style="list-style-type: none"> <li>- <b>Path</b>: Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● CSS <ul style="list-style-type: none"> <li>- <b>Cluster Name</b>: Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>- <b>Index</b>: Enter a CSS index name.</li> </ul> </li> <li>● HIVE <ul style="list-style-type: none"> <li>- <b>Connection Name</b>: Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>- <b>Database</b>: Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>- <b>Table Name</b>: Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● CUSTOM <ul style="list-style-type: none"> <li>- <b>Name</b>: Enter a name of the CUSTOM type.</li> <li>- <b>Attribute</b>: Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● DLI <ul style="list-style-type: none"> <li>- <b>Connection Name</b>: Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>- <b>Database</b>: Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>- <b>Table Name</b>: Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>



Parameter	Description
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the output lineage.

### 3.4.9.3 Rest Client

#### Functions

The Rest Client node is used to respond to RESTful requests in . Only the RESTful requests that have been authenticated by using IAM tokens are supported.

#### NOTE

If some APIs of the Rest Client node cannot be called due to network restrictions, you can use a shell script to call the APIs. To call an API using a shell script, you must have an ECS that can communicate with the API. Create a host connection and run the curl command to call the API using the shell script.

#### Parameters

[Table 3-160](#), [Table 3-161](#), and [Table 3-162](#) describe the parameters of the Rest Client node.

**Table 3-160** Parameters of Rest Client nodes

Parameter	Man dator y	Description
Node Name	Yes	Name of a node. The name must contain 1 to 128 characters, including only letters, numbers, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
Agent Name	Yes	Name of a CDM cluster. The CDM cluster provides the agent connection function.  If the selected CDM cluster is in the same VPC as the third-party service, the REST client can call APIs on the tenant plane.

Parameter	Mandatory	Description
URL Address	Yes	IP address or domain name and port number of the request host. For example: https://192.160.10.10:8080
HTTP Method	Yes	Type of the request. Possible values: <ul style="list-style-type: none"> <li>• <b>GET</b></li> <li>• <b>POST</b></li> <li>• <b>PUT</b></li> <li>• <b>DELETE</b></li> </ul>
Request Header	No	Click <b>+</b> to add a request header. The parameters are described as follows: <ul style="list-style-type: none"> <li>• Parameter Name Name of a parameter. The options are <b>Content-Type</b> and <b>Accept-Language</b>.</li> <li>• Parameter Value Value of the parameter</li> </ul>
URL Parameter	No	Enter a URL parameter. The value is a character string in <b>key=value</b> format. Character strings are separated by newlines. This parameter is available only when <b>HTTP Method</b> is set to <b>GET</b> . Set these parameters as follows: <ul style="list-style-type: none"> <li>• Parameter The parameter contains a maximum of 32 characters, including only letters, numbers, hyphens (-), and underscores (_).</li> <li>• Value The value contains a maximum of 64 characters, including only letters, numbers, hyphens (-), underscores (_), dollar signs (\$), open braces ({), and close braces (}).</li> </ul>
Request Body	Yes	The request body is in JSON format. This parameter is available only when <b>HTTP Method</b> is set to <b>POST</b> or <b>PUT</b> .
Check Return Value	No	Checks whether the value of the returned message is the same as the expected value. This parameter is available only when <b>HTTP Method</b> is set to <b>GET</b> . Possible values: <ul style="list-style-type: none"> <li>• <b>YES</b>: Check whether the return value is the same as the expected one.</li> <li>• <b>NO</b>: No need to check whether the return value is the same as the expected one. A 200 response code is returned (indicating that the node is successfully performed).</li> </ul>

Parameter	Mandatory	Description
Property Path	Yes	<p>Path of the property in the JSON response message. Each Rest Client node can have only one property path. This parameter is available only when <b>Check Returned Value</b> is set to <b>YES</b>.</p> <p>For example, the returned result is as follows:</p> <pre data-bbox="730 539 1430 871"> {   "param1": "aaaa",   "inner":   {     "inner":     {       "param4": 2014247437     },     "param3": "cccc"   },   "status": 200,   "param2": "bbbb" } </pre> <p>The <b>param4</b> path is <b>inner.inner.param4</b>.</p>
Request Success Flag	Yes	<p>Enter the request success flag. If the returned value of the response matches one of request success flags, the node is successfully performed. This parameter is available only when <b>Check Returned Value</b> is set to <b>YES</b>.</p> <p>The request success flag can contain only letters, numbers, hyphens (-), underscores (_), dollar signs (\$), open braces ({), and close braces (}). Separate values with semicolons (;).</p>
Request Failure Flag	No	<p>Enter the request failure flag. If the returned value of the response matches one of request failure flags, the node is successfully performed. This parameter is available only when <b>Check Returned Value</b> is set to <b>YES</b>.</p> <p>The request failure flag can contain only letters, numbers, hyphens (-), underscores (_), dollar signs (\$), open braces ({), and close braces (}). Separate values with semicolons (;).</p>
Retry Interval (seconds)	Yes	<p>If the return value of the response message does not match the request success flag, the node keeps querying the matching status at a specified interval until the return value of the response message is the same as the request success flag. By default, the timeout interval of the node is one hour. If the return value of the response message does not match the request success flag within this period, the node status changes to <b>Failed</b>. This parameter is available only when <b>Check Returned Value</b> is set to <b>YES</b>.</p>

Parameter	Mandatory	Description
The response message body parses the transfer parameter.	No	Specify the mapping between the job variable and JSON property path. Separate parameters by newline characters. For example: var4=inner.inner.param4 <b>var4</b> is a job variable. The job variable must contain 1 to 64 characters, including only letters and numbers. <b>inner.inner.param4</b> is the JSON property path. This parameter takes effect only when it is referenced by the subsequent node. When this parameter is referenced, the format is <b>\${var4}</b>


**Table 3-161** Advanced parameters



Parameter	Mandatory	Description
Max. Node Execution Duration	Yes	Execution timeout interval for the node. If retry is configured and the execution is not complete within the timeout interval, the node will not be retried and is set to the failed state.
Retry upon Failure	Yes	Indicates whether to re-execute a node task if its execution fails. Possible values: <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> <li>- <b>Maximum Retries</b></li> <li>- <b>Retry Interval (seconds)</b></li> </ul> </li> <li>• <b>No:</b> The node task will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Timeout Interval</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>




Parameter	Mandatory	Description
Failure Policy	Yes	<p>Operation that will be performed if the node task fails to be executed. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>

**Table 3-162** Lineage

Parameter	Description
Input	

Parameter	Description
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● <b>DWS</b> <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>- <b>Schema:</b> Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● <b>OBS</b> <ul style="list-style-type: none"> <li>- <b>Path:</b> Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● <b>CSS</b> <ul style="list-style-type: none"> <li>- <b>Cluster Name:</b> Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>- <b>Index:</b> Enter a CSS index name.</li> </ul> </li> <li>● <b>HIVE</b> <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● <b>CUSTOM</b> <ul style="list-style-type: none"> <li>- <b>Name:</b> Enter a name of the CUSTOM type.</li> <li>- <b>Attribute:</b> Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● <b>DLI</b> <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.

Parameter	Description
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the input lineage.
<b>Output</b>	
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● DWS <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>- <b>Schema:</b> Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● OBS <ul style="list-style-type: none"> <li>- <b>Path:</b> Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● CSS <ul style="list-style-type: none"> <li>- <b>Cluster Name:</b> Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>- <b>Index:</b> Enter a CSS index name.</li> </ul> </li> <li>● HIVE <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● CUSTOM <ul style="list-style-type: none"> <li>- <b>Name:</b> Enter a name of the CUSTOM type.</li> <li>- <b>Attribute:</b> Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● DLI <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>

Parameter	Description
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the output lineage.

### 3.4.9.4 Import GES

#### Function

The Import GES node is used to import files from an OBS bucket to a GES graph.

#### Parameters

[Table 3-163](#) and [Table 3-164](#) describe the parameters of the Import GES node.

**Table 3-163** Parameters of Import GES nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of a node. The name must contain 1 to 128 characters, including only letters, numbers, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
Graph Name	Yes	You can directly select the graph to import or manually enter the graph name. To create a GES graph, go to the GES console.
Metadata	Yes	You can directly select the corresponding metadata or manually enter the OBS path of the metadata.
Edge Data Set	Yes	You can directly select the corresponding edge data set or manually enter the OBS path of the edge data set.
Vertex Data Set	No	You can directly select the corresponding Vertex data set or manually enter the OBS path of the Vertex data set.  If it is not selected, the vertices in the edge dataset are used as the source of the vertex dataset.



Parameter	Mandatory	Description
Edge Processing	Yes	The edge processing supports the following modes: <ul style="list-style-type: none"> <li>• Allow repetitive edges</li> <li>• Ignore subsequent repetitive edges</li> <li>• Overwrite previous repetitive edges</li> </ul>
Offline	No	Whether offline import is used. The value is <b>Yes</b> or <b>No</b> , and the default value is <b>No</b> . <ul style="list-style-type: none"> <li>• <b>true</b>: Offline import is selected. The import speed is high, but the graph is locked and cannot be read or written during the import.</li> <li>• <b>false</b>: Online import is selected. Online import is slower than offline import. However, during online import, the graph can be read (but cannot be written).</li> </ul>
Ignore Labels on Repetitive Edges	No	Indicates whether to ignore labels on repetitive edges. The value is <b>Yes</b> or <b>No</b> , and the default value is <b>Yes</b> . <ul style="list-style-type: none"> <li>• <b>Yes</b>: Indicates that the repetitive edge definition does not contain the label. That is, the &lt;source vertex, target vertex&gt; indicates an edge, excluding the label information.</li> <li>• <b>No</b>: Indicates that the repetitive edge definition contains the label. That is, the &lt;source vertex, target vertex, label&gt; indicates an edge.</li> </ul>
Log Storage Path	No	Stores vertex and edge datasets that do not comply with the metadata definition, as well as detailed logs generated during graph import.

**Table 3-164** Advanced parameters

Parameter	Mandatory	Description
Node Status Polling Interval (s)	Yes	Specifies how often the system check completeness of the node task. The value ranges from 1 to 60 seconds.
Max. Node Execution Duration	Yes	Execution timeout interval for the node. If retry is configured and the execution is not complete within the timeout interval, the node will not be retried and is set to the failed state.

Parameter	Mandatory	Description
Retry upon Failure	Yes	<p>Indicates whether to re-execute a node task if its execution fails. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> <li>– <b>Maximum Retries</b></li> <li>– <b>Retry Interval (seconds)</b></li> </ul> </li> <li>• <b>No:</b> The node task will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Timeout Interval</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>
Failure Policy	Yes	<p>Operation that will be performed if the node task fails to be executed. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>

### 3.4.9.5 MRS Kafka

#### Functions

The MRS Kafka node is used to query the number of messages that are not consumed by a topic.

#### Parameters

[Table 3-165](#) and [Table 3-166](#) describe the parameters of the MRS Kafka node.

**Table 3-165** Parameters of MRS Kafka nodes

Parameter	Mandatory	Description
Data Connection	Yes	Select the MRS Kafka connection created in the management center.
Topic Name	Yes	Select a topic that has been created in MRS Kafka. The SDK or command line can be used to create a topic.
Node Name	Yes	Name of a node. The name must contain 1 to 128 characters, including only letters, numbers, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).

**Table 3-166** Advanced parameters

Parameter	Mandatory	Description
Max. Node Execution Duration	Yes	Execution timeout interval for the node. If retry is configured and the execution is not complete within the timeout interval, the node will not be retried and is set to the failed state.
Retry upon Failure	Yes	Indicates whether to re-execute a node task if its execution fails. Possible values: <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> <li>- <b>Maximum Retries</b></li> <li>- <b>Retry Interval (seconds)</b></li> </ul> </li> <li>• <b>No:</b> The node task will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Timeout Interval</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>

Parameter	Mandatory	Description
Failure Policy	Yes	<p>Operation that will be performed if the node task fails to be executed. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>

### 3.4.9.6 Kafka Client

#### Functions


The Kafka Client node is used to send data to Kafka topics.

#### Parameters

[Table 3-167](#) describes the parameters of the Kafka Client node.

**Table 3-167** Parameters of Kafka Client nodes

Parameter	Mandatory	Description
Data Connection	Yes	Select the MRS Kafka connection created in the management center.
Topic Name	Yes	Select the topic to which data is to be uploaded. If there are multiple partitions, data is sent to partition 0 by default.
Node Name	Yes	Name of a node. The name must contain 1 to 128 characters, including only letters, numbers, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).

Parameter	Mandatory	Description
Text	Yes	Text content sent to Kafka. You can directly enter text or click  to use the EL expression.

**Table 3-168** Advanced parameters

Parameter	Mandatory	Description
Max. Node Execution Duration	Yes	Execution timeout interval for the node. If retry is configured and the execution is not complete within the timeout interval, the node will not be retried and is set to the failed state.
Retry upon Failure	Yes	<p>Indicates whether to re-execute a node task if its execution fails. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> <li>- <b>Maximum Retries</b></li> <li>- <b>Retry Interval (seconds)</b></li> </ul> </li> <li>• <b>No:</b> The node task will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Timeout Interval</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>

Parameter	Mandatory	Description
Failure Policy	Yes	<p>Operation that will be performed if the node task fails to be executed. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>

### 3.4.9.7 ROMA FDI Job

#### Functions

The ROMA FDI Job node executes a predefined ROMA Connect data integration task to implement data integration and conversion between the source and destination.

#### Working Principles

This node enables you to start an FDI task or query whether an FDI task is running.

#### Parameters

The following table describes the parameters of a ROMA FDI Job node.

**Table 3-169** Property parameters

Parameter	Mandatory	Description
ROMA Instance	Yes	Select an existing ROMA instance.
FDI Task	Yes	Select an existing ROMA FDI task.

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. The value must consist of 1 to 128 characters and contain only letters, digits, and the following special characters: _-/<>

**Table 3-170** Advanced parameters

Parameter	Mandatory	Description
Max. Node Execution Duration	Yes	Execution timeout interval for the node. If retry is configured and the execution is not complete within the timeout interval, the node will not be retried and is set to the failed state.
Retry upon Failure	Yes	<p>Indicates whether to re-execute a node task if its execution fails. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> <li>– <b>Maximum Retries</b></li> <li>– <b>Retry Interval (seconds)</b></li> </ul> </li> <li>• <b>No:</b> The node task will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Timeout Interval</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>

Parameter	Mandatory	Description
Failure Policy	Yes	Operation that will be performed if the node task fails to be executed. Possible values: <ul style="list-style-type: none"><li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li><li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li><li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li><li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li></ul>

### 3.4.9.8 DLI Flink Job

#### Function

The DLI Flink Job node is used to execute a predefined DLI job for real-time analysis of streaming data.

#### Working Principles

This node enables you to start a DLI job or query whether a DLI job is running. If you do not select an existing Flink job, DLF creates and starts the job based on the job status configured on the node. You can customize jobs and job parameters.

#### Parameters

For details about how to configure the parameters of DLI Flink jobs, see the following:

- Property parameters:
  - **Existing Flink job:** For details, see [Table 3-171](#).
  - **Flink SQL job:** For details, see [Table 3-172](#).
  - **User-defined Flink job:** For details, see [Table 3-173](#).
- [Table 3-174](#)



**Table 3-171** Parameter parameters of an existing Flink job

Parameter	Mandatory	Description
Job Type	Yes	Select <b>Existing Flink job</b> .
Job Name	Yes	Name of an existing DLI Flink job.
Node Name	Yes	Name of a node. The name must contain 1 to 128 characters, including only letters, numbers, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).

**Table 3-172** Property parameters of a Flink SQL job

Parameter	Mandatory	Description
Job Type	Yes	Select <b>Flink SQL job</b> . You can start a job by compiling SQL statements.
Script Path	Yes	Path to a Flink SQL script to be executed. If the script is not created, create and develop the Flink SQL script by referring to <a href="#">Creating a Script</a> and <a href="#">Developing an SQL Script</a> .
DLI Queue	Yes	<b>Shared queues</b> are selected by default. You can also select a dedicated custom queue. <b>NOTE</b> During job creation, a sub-user can only select a queue that has been allocated to the user.
CUs	Yes	A CU consists of 1 vCPU compute and 4 GB memory.
Concurrency	Yes	The number of Flink SQL jobs that run at the same time. <b>NOTE</b> The value of <b>Concurrency</b> must not exceed the value obtained through the following formula: $4 \times (\text{Number of CUs} - 1)$ .
UDF Jar	No	This parameter is valid only when you select a dedicated queue for <b>Queue</b> . Before selecting a UDF JAR resource package, upload the UDF JAR package to the OBS bucket and create resources on the <b>Manage Resource</b> page. For details, see <a href="#">Creating a Resource</a> .  In SQL, you can call a user-defined function that is inserted into a JAR package.

Parameter	Mandatory	Description
Auto Restart upon Exception	No	Indicates whether to enable automatic restart. If this function is enabled, any job that has become abnormal will be automatically restarted.
Job Name	Yes	Name of the DLI Flink job. It must consist of 1 to 64 characters and contain only letters, numbers, and underscores (_). The default value is the same as the node name.
Job name must be prefixed with workspace name	No	Whether to add a workspace prefix to the created job.
Node Name	Yes	Name of a node. The name must contain 1 to 128 characters, including only letters, numbers, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).

**Table 3-173** Property parameters of a user-defined Flink job

Parameter	Mandatory	Description
Job Type	Yes	Select <b>User-defined Flink job</b> .
JAR Package Path	Yes	User-defined package. Before selecting a package, upload the JAR package to the OBS bucket and create resources on the <b>Manage Resource</b> page. For details, see <a href="#">Creating a Resource</a> .
Main Class	Yes	Name of the JAR package to be loaded, for example, <b>KafkaMessageStreaming</b> . <ul style="list-style-type: none"> <li><b>Default:</b> Specified based on the <b>Manifest</b> file in the JAR package.</li> <li><b>Manually assign:</b> Enter the class name and confirm the class arguments (separate arguments with spaces).</li> </ul> <p><b>NOTE</b> When a class belongs to a package, the package path must be carried, for example, <b>packagePath.KafkaMessageStreaming</b>.</p>
Main Class Parameter	Yes	List of parameters of a specified class. The parameters are separated by spaces.
DLI Queue	Yes	<b>Shared queues</b> are selected by default. You can also select a dedicated custom queue. <p><b>NOTE</b> During job creation, a sub-user can only select a queue that has been allocated to the user.</p>

Parameter	Mandatory	Description
Job Type	No	<p>Select a custom image and the corresponding version. This parameter is available only when the DLI queue is a containerized queue.</p> <p>A custom image is a feature of DLI. You can use the Spark or Flink basic images provided by DLI to pack the dependencies (files, JAR packages, or software) required into an image using Dockerfile, generate a custom image, and release the image to SWR. Then, select the generated image and run the job.</p> <p>Custom images can change the container runtime environments of Spark and Flink jobs. You can embed private capabilities into custom images to enhance the functions and performance of jobs.</p>
CUs	Yes	A CU consists of 1 vCPU compute and 4 GB memory.
Number of management node CUs	Yes	Set the number of CUs on a management unit. The value ranges from 1 to 4. The default value is <b>1</b> .
Concurrency	Yes	<p>The number of Flink SQL jobs that run at the same time.</p> <p><b>NOTE</b> The value of <b>Concurrency</b> must not exceed the value obtained through the following formula: <math>4 \times (\text{Number of CUs} - 1)</math>.</p>
Auto Restart upon Exception	No	Indicates whether to enable automatic restart. If this function is enabled, any job that has become abnormal will be automatically restarted.
Job Name	Yes	Name of the DLI Flink job. It must consist of 1 to 64 characters and contain only letters, numbers, and underscores (_). The default value is the same as the node name.
Job name must be prefixed with workspace name	No	Whether to add a workspace prefix to the created job.
Node Name	Yes	Name of a node. The name must contain 1 to 128 characters, including only letters, numbers, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).

**Table 3-174** Advanced parameters

Parameter	Mandatory	Description
Max. Node Execution Duration	Yes	Execution timeout interval for the node. If retry is configured and the execution is not complete within the timeout interval, the node will not be retried and is set to the failed state.
Retry upon Failure	Yes	Indicates whether to re-execute a node task if its execution fails. Possible values: <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> <li>- <b>Maximum Retries</b></li> <li>- <b>Retry Interval (seconds)</b></li> </ul> </li> <li>• <b>No:</b> The node task will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Timeout Interval</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>
Failure Policy	Yes	Operation that will be performed if the node task fails to be executed. Possible values: <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>

### 3.4.9.9 DLI SQL

#### Functions

The DLI SQL node is used to transfer SQL statements to DLI for data source analysis and exploration.

## Working Principles



This node enables you to execute DLI statements during periodical or real-time job scheduling. You can use parameter variables to perform incremental import and process partitions for your data warehouses.

## Parameters

[Table 3-175](#), [Table 3-176](#), and [Table 3-177](#) describe the parameters of the DLI SQLnode node.

**Table 3-175** Parameters of DLI SQL nodes

Parameter	Mandatory	Description
SQL Statement or Script	Yes	<p>You can select SQL statements or SQL scripts.</p> <ul style="list-style-type: none"> <li>SQL Statement In the <b>SQL statement</b> text box, enter the SQL statement to be executed.</li> <li>SQL Script Select a script to be executed. If the script is not created, create and develop the script by repeating steps <a href="#">Creating a Script</a> and <a href="#">Developing an SQL Script</a>.</li> </ul> <p><b>NOTE</b> If you select the SQL statement mode, the DataArts Factory module cannot parse the parameters contained in the SQL statement.</p>
Database Name	Yes	Database that is configured in the SQL script. The value can be changed.
DLI Environmental Variable	No	<ul style="list-style-type: none"> <li>The environment variable must start with <b>dli.sql.</b> or <b>spark.sql.</b></li> <li>If the key of the environment variable is <b>dli.sql.shuffle.partitions</b> or <b>dli.sql.autoBroadcastJoinThreshold</b>, the environment variable cannot contain the greater than (&gt;) or less than (&lt;) sign.</li> <li>If a parameter with the same name is configured in both a job and a script, the parameter value configured in the job will overwrite that configured in the script.</li> </ul>

Parameter	Mandatory	Description
Queue Name	Yes	Name of the DLI queue configured in the SQL script. The value can be changed. You can create a resource queue using either of the following methods: <ul style="list-style-type: none"> <li>Click . On the <b>Queue Management</b> page of DLI, create a resource queue.</li> <li>Go to the DLI console to create a resource queue.</li> </ul>
Script Parameter	No	If the associated SQL script uses a parameter, the parameter name is displayed. Set the parameter value in the text box next to the parameter name. The parameter value can be <a href="#">an EL expression</a> . If the parameters of the associated SQL script are changed, click  to refresh the parameters.
Node Name	Yes	Name of the SQL script. The value can be changed. The rules are as follows: Name of a node. The name must contain 1 to 128 characters, including only letters, numbers, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
Record Dirty Data	Yes	Click <input type="radio"/> to specify whether to record dirty data. <ul style="list-style-type: none"> <li>If you select <input type="radio"/>, dirty data will be recorded.</li> <li>If you do not select <input type="radio"/>, dirty data will not be recorded.</li> </ul>


**Table 3-176** Advanced parameters

Parameter	Mandatory	Description
Node Status Polling Interval (s)	Yes	Specifies how often the system check completeness of the node task. The value ranges from 1 to 60 seconds.
Max. Node Execution Duration	Yes	Execution timeout interval for the node. If retry is configured and the execution is not complete within the timeout interval, the node will not be retried and is set to the failed state.



Parameter	Mandatory	Description
Retry upon Failure	Yes	<p>Indicates whether to re-execute a node task if its execution fails. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> <li>– <b>Maximum Retries</b></li> <li>– <b>Retry Interval (seconds)</b></li> </ul> </li> <li>• <b>No:</b> The node task will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Timeout Interval</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>
Failure Policy	Yes	<p>Operation that will be performed if the node task fails to be executed. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>




Table 3-177 Lineage

Parameter	Description
Input	

Parameter	Description
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● <b>DWS</b> <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>- <b>Schema:</b> Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● <b>OBS</b> <ul style="list-style-type: none"> <li>- <b>Path:</b> Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● <b>CSS</b> <ul style="list-style-type: none"> <li>- <b>Cluster Name:</b> Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>- <b>Index:</b> Enter a CSS index name.</li> </ul> </li> <li>● <b>HIVE</b> <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● <b>CUSTOM</b> <ul style="list-style-type: none"> <li>- <b>Name:</b> Enter a name of the CUSTOM type.</li> <li>- <b>Attribute:</b> Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● <b>DLI</b> <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.



Parameter	Description
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the input lineage.
<b>Output</b>	
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● DWS <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>- <b>Schema:</b> Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● OBS <ul style="list-style-type: none"> <li>- <b>Path:</b> Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● CSS <ul style="list-style-type: none"> <li>- <b>Cluster Name:</b> Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>- <b>Index:</b> Enter a CSS index name.</li> </ul> </li> <li>● HIVE <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● CUSTOM <ul style="list-style-type: none"> <li>- <b>Name:</b> Enter a name of the CUSTOM type.</li> <li>- <b>Attribute:</b> Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● DLI <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>

Parameter	Description
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the output lineage.

### 3.4.9.10 DLI Spark

#### Function

The DLI Spark node is used to execute a predefined Spark job.

#### Parameters

[Table 3-178](#), [Table 3-179](#), and [Table 3-180](#) describe the parameters of the DLI Spark node.

**Table 3-178** Parameters of DLI Spark nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. The value must consist of 1 to 128 characters and contain only letters, digits, and the following special characters: _-/<>
DLI Queue	Yes	Select a queue from the drop-down list box.
Job Type	No	Select a custom image and the corresponding version. This parameter is available only when the DLI queue is a containerized queue.  A custom image is a feature of DLI. You can use the Spark or Flink basic images provided by DLI to pack the dependencies (files, JAR packages, or software) required into an image using Dockerfile, generate a custom image, and release the image to SWR. Then, select the generated image and run the job.  Custom images can change the container runtime environments of Spark and Flink jobs. You can embed private capabilities into custom images to enhance the functions and performance of jobs. .

Parameter	Mandatory	Description
Job Name	Yes	Name of the DLI Spark job. The name must contain 1 to 64 characters, including only letters, numbers, and underscores (_). The default value is the same as the node name.
Job Running Resources	No	Select the running resource specifications of the job. <ul style="list-style-type: none"> <li>8-core, 32 GB memory</li> <li>16-core, 64 GB memory</li> <li>32-core, 128 GB memory</li> </ul>
Major Job Class	Yes	Name of the major class of the Spark job. When the application type is <b>.jar</b> , the main class name cannot be empty.
Spark program resource package	Yes	JAR file on which the Spark job depends. You can enter the JAR package name or the corresponding OBS path. The format is as follows: <b>obs://Bucket name/Folder name/Package name</b> . Before selecting a resource package, upload the JAR package and its dependency packages to the OBS bucket and create resources on the <b>Manage Resource</b> page. For details, see <a href="#">Creating a Resource</a> .
Resource Type	Yes	Select <b>OBS path</b> or <b>DLI program package</b> . <ul style="list-style-type: none"> <li><b>OBS path:</b> The resource package file will not be uploaded to DLI resource management system before the job is executed. The OBS path where the file is located is part of the message body for starting the job. This type is recommended.</li> <li><b>DLI package:</b> The resource package file will not be uploaded to the DLI resource management system before the job is executed.</li> </ul>
Group	No	This parameter is mandatory when <b>Resource Type</b> is set to <b>DLI program package</b> . You can select <b>Use existing</b> , <b>Create new</b> , or <b>Do not use</b> .
Group Name	No	This parameter is mandatory when <b>Resource Type</b> is set to <b>DLI program package</b> . <ul style="list-style-type: none"> <li><b>Use existing:</b> Select an existing group.</li> <li><b>Create new:</b> Enter a user-defined group name.</li> <li><b>Do not use:</b> Do not select or enter a group name.</li> </ul>




Parameter	Man dator y	Description
Major-Class Entry Parameters	No	<p>User-defined parameters. Separate multiple parameters by <b>Enter</b>.</p> <p>These parameters can be replaced by global variables. For example, if you create a global variable <b>batch_num</b> on the <b>Global Configuration &gt; Global Variables</b> page, you can use <code>{{batch_num}}</code> to replace a parameter with this variable after the job is submitted.</p>
Spark Job Running Parameters	No	<p>Enter a parameter in the format of <b>key/value</b>. Press Enter to separate multiple key-value pairs. For details about the parameters, see <a href="#">Spark Configuration</a>.</p> <p>These parameters can be replaced by global variables. For example, if you create a global variable <b>custom_class</b> on the <b>Global Configuration &gt; Global Variables</b> page, you can use <code>"spark.sql.catalog"={{custom_class}}</code> to replace a parameter with this variable after the job is submitted.</p> <p><b>NOTE</b> The JVM garbage collection algorithm cannot be customized for Spark jobs.</p>
Module Name	No	<p>Dependency modules provided by DLI for executing datasource connection jobs. To access different services, you need to select different modules.</p> <ul style="list-style-type: none"> <li>● CloudTable/MRS HBase: sys.datasource.hbase</li> <li>● DDS: sys.datasource.mongo</li> <li>● CloudTable/MRS OpenTSDB: sys.datasource.opentsdb</li> <li>● DWS: sys.datasource.dws</li> <li>● RDS MySQL: sys.datasource.rds</li> <li>● RDS PostGre: sys.datasource.rds</li> <li>● DCS: sys.datasource.redis</li> <li>● CSS: sys.datasource.css</li> </ul> <p>DLI internal modules include:</p> <ul style="list-style-type: none"> <li>● sys.res.dli-v2</li> <li>● sys.res.dli</li> <li>● sys.datasource.dli-inner-table</li> </ul>
Metadata Access	Yes	<p>Whether to access metadata through Spark jobs. For details, see section "Using the Spark Job to Access DLI Metadata" in <i>Data Lake Insight Developer Guide</i>.</p>


**Table 3-179** Advanced parameters

Parameter	Mandatory	Description
Node Status Polling Interval (s)	Yes	How often the system checks completeness of the node task. The value ranges from 1 to 60 seconds.
Max. Node Execution Duration	Yes	Maximum duration of executing a node. When <b>Retry upon Failure</b> is set to <b>Yes</b> for a node, the node can be re-executed for numerous times upon an execution failure within the maximum duration.
Retry upon Failure	Yes	<p>Whether to re-execute a node after the node fails to be executed.</p> <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node will be re-executed after it fails to be executed. The following parameters must be configured: <ul style="list-style-type: none"> <li>- Maximum Retries</li> <li>- Retry Interval (seconds)</li> </ul> </li> <li>• <b>No:</b> The node will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Max. Node Execution Duration</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>
Failure Policy	Yes	<p>Policies to be performed after the node fails to be executed:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>
Dry run	No	If you select this option, the node will not be executed, and a success message will be returned.



**Table 3-180** Lineage

Parameter	Description
<b>Input</b>	
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● <b>DWS</b> <ul style="list-style-type: none"> <li>– <b>Connection Name</b>: Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>– <b>Database</b>: Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>– <b>Schema</b>: Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>– <b>Table Name</b>: Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● <b>OBS</b> <ul style="list-style-type: none"> <li>– <b>Path</b>: Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● <b>CSS</b> <ul style="list-style-type: none"> <li>– <b>Cluster Name</b>: Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>– <b>Index</b>: Enter a CSS index name.</li> </ul> </li> <li>● <b>HIVE</b> <ul style="list-style-type: none"> <li>– <b>Connection Name</b>: Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>– <b>Database</b>: Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>– <b>Table Name</b>: Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● <b>CUSTOM</b> <ul style="list-style-type: none"> <li>– <b>Name</b>: Enter a name of the CUSTOM type.</li> <li>– <b>Attribute</b>: Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● <b>DLI</b> <ul style="list-style-type: none"> <li>– <b>Connection Name</b>: Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>– <b>Database</b>: Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>– <b>Table Name</b>: Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.

Parameter	Description
Modify	Click  to modify the parameter settings. After the modification, save the settings.
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the input lineage.
<b>Output</b>	

Parameter	Description
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● <b>DWS</b> <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>- <b>Schema:</b> Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● <b>OBS</b> <ul style="list-style-type: none"> <li>- <b>Path:</b> Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● <b>CSS</b> <ul style="list-style-type: none"> <li>- <b>Cluster Name:</b> Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>- <b>Index:</b> Enter a CSS index name.</li> </ul> </li> <li>● <b>HIVE</b> <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● <b>CUSTOM</b> <ul style="list-style-type: none"> <li>- <b>Name:</b> Enter a name of the CUSTOM type.</li> <li>- <b>Attribute:</b> Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● <b>DLI</b> <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.



Parameter	Description
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the output lineage.

### 3.4.9.11 DWS SQL

#### Functions

The DWS SQL node is used to transfer SQL statements to DWS.

For details about how to use the DWS SQL operator, see [Developing a DWS SQL Job](#).

#### Context


This node enables you to execute DWS statements during batch or real-time job processing. You can use parameter variables to perform incremental import and process partitions for your data warehouses.

#### Parameters

[Table 3-181](#), [Table 3-182](#), and [Table 3-183](#) describe the parameters of the DWS SQLnode node.

**Table 3-181** Parameters of DWS SQL nodes

Parameter	Mandatory	Description
SQL or Script	Yes	<p>You can select <b>SQL statement</b> or <b>SQL script</b>.</p> <ul style="list-style-type: none"> <li>SQL Statement In the <b>SQL statement</b> text box, enter the SQL statement to be executed.</li> <li>SQL Script Select a script to be executed. If the script is not created, create and develop the script by repeating steps <a href="#">Creating a Script</a> and <a href="#">Developing an SQL Script</a>.</li> </ul> <p><b>NOTE</b> If you select the SQL statement mode, the DataArts Factory module cannot parse the parameters contained in the SQL statement.</p>
Data Connection	Yes	Data connection that is configured in the SQL script. The value can be changed.

Parameter	Mandatory	Description
Database	Yes	Database that is configured in the SQL script. The value can be changed.
Script Parameter	No	If the associated SQL script uses a parameter, the parameter name is displayed. Set the parameter value in the text box next to the parameter name. The parameter value can be <b>an EL expression</b> . If the parameters of the associated SQL script are changed, click  to refresh the parameters.
Dirty Data Table	No	Enter the name of the dirty data table defined in the SQL script.
Matching Rule	-	Enter a Java regular expression used to match the DWS SQL result. For example, if the expression is (?<= \()(-*\d+?)(?=,) and the SQL result is (1,"error message"), then the matched result is "1".
Failure Matching Value	-	If the matched content equals the set value, the node fails to be executed.
Node Name	Yes	Name of the SQL script. The value can be changed. The rules are as follows: Name of a node. The name must contain 1 to 128 characters, including only letters, numbers, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).


**Table 3-182** Advanced parameters



Parameter	Mandatory	Description
Node Status Polling Interval (s)	Yes	Specifies how often the system check completeness of the node task. The value ranges from 1 to 60 seconds.
Max. Node Execution Duration	Yes	Execution timeout interval for the node. If retry is configured and the execution is not complete within the timeout interval, the node will not be retried and is set to the failed state.




Parameter	Mandatory	Description
Retry upon Failure	Yes	<p>Indicates whether to re-execute a node task if its execution fails. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> <li>– <b>Maximum Retries</b></li> <li>– <b>Retry Interval (seconds)</b></li> </ul> </li> <li>• <b>No:</b> The node task will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Timeout Interval</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>
Failure Policy	Yes	<p>Operation that will be performed if the node task fails to be executed. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>

Table 3-183 Lineage

Parameter	Description
Input	

Parameter	Description
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● <b>DWS</b> <ul style="list-style-type: none"> <li>– <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>– <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>– <b>Schema:</b> Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>– <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● <b>OBS</b> <ul style="list-style-type: none"> <li>– <b>Path:</b> Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● <b>CSS</b> <ul style="list-style-type: none"> <li>– <b>Cluster Name:</b> Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>– <b>Index:</b> Enter a CSS index name.</li> </ul> </li> <li>● <b>HIVE</b> <ul style="list-style-type: none"> <li>– <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>– <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>– <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● <b>CUSTOM</b> <ul style="list-style-type: none"> <li>– <b>Name:</b> Enter a name of the CUSTOM type.</li> <li>– <b>Attribute:</b> Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● <b>DLI</b> <ul style="list-style-type: none"> <li>– <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>– <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>– <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.

Parameter	Description
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the input lineage.
<b>Output</b>	
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● DWS <ul style="list-style-type: none"> <li>- <b>Connection Name</b>: Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>- <b>Database</b>: Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>- <b>Schema</b>: Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>- <b>Table Name</b>: Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● OBS <ul style="list-style-type: none"> <li>- <b>Path</b>: Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● CSS <ul style="list-style-type: none"> <li>- <b>Cluster Name</b>: Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>- <b>Index</b>: Enter a CSS index name.</li> </ul> </li> <li>● HIVE <ul style="list-style-type: none"> <li>- <b>Connection Name</b>: Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>- <b>Database</b>: Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>- <b>Table Name</b>: Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● CUSTOM <ul style="list-style-type: none"> <li>- <b>Name</b>: Enter a name of the CUSTOM type.</li> <li>- <b>Attribute</b>: Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● DLI <ul style="list-style-type: none"> <li>- <b>Connection Name</b>: Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>- <b>Database</b>: Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>- <b>Table Name</b>: Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>

Parameter	Description
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the output lineage.

### 3.4.9.12 MRS Spark SQL


#### Function

The MRS Spark SQL node is used to execute a predefined SparkSQL statement on MRS.

#### Parameters

[Table 3-184](#), [Table 3-185](#), and [Table 3-186](#) describe the parameters of the MRS Spark SQL node.

**Table 3-184** Parameter of MRS Spark SQL nodes

Parameter	Mandatory	Description
SQL Script	Yes	Path of a script to be executed. If no script is available, create and develop a script by referring to <a href="#">Creating a Script</a> and <a href="#">Developing an SQL Script</a> .
Data Connection	Yes	Data connection that is configured in the SQL script. The value can be changed.
Database	Yes	Database that is configured in the SQL script. The value can be changed.
Script Parameter	No	If the associated SQL script uses a parameter, the parameter name is displayed. Set the parameter value in the text box next to the parameter name. The parameter value can be <a href="#">an EL expression</a> . If the parameters of the associated SQL script are changed, click  to refresh the parameters.

Parameter	Mandatory	Description
Program Parameter	No	Used to configure optimization parameters such as threads, memory, and vCPUs for the job to optimize resource usage and improve job execution performance. <b>NOTE</b> This parameter is mandatory if the cluster version is MRS 1.8.7 or later than MRS 2.0.1. For details about the program parameters of MRS SparkSQL jobs, see <b>Managing an Existing Cluster &gt; Job Management &gt; Running a SparkSQL Job</b> in the <i>MapReduce Service User Guide</i> .
Node Name	Yes	Name of the SQL script. The value can be changed. Node name. It contains 1 to 64 characters. Only letters, digits, hyphens (-), and underscores (_) are allowed. <b>NOTE</b> The node name cannot contain more than 64 characters. If the node name does not meet requirements, the MRS job will fail to be submitted.

**Table 3-185** Advanced parameters


Parameter	Mandatory	Description
Node Status Polling Interval (s)	Yes	How often the system checks completeness of the node task. The value ranges from 1 to 60 seconds.
Max. Node Execution Duration	Yes	Maximum duration of executing a node. When <b>Retry upon Failure</b> is set to <b>Yes</b> for a node, the node can be re-executed for numerous times upon an execution failure within the maximum duration.



Parameter	Mandatory	Description
Retry upon Failure	Yes	<p>Whether to re-execute a node after the node fails to be executed.</p> <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node will be re-executed after it fails to be executed. The following parameters must be configured: <ul style="list-style-type: none"> <li>- Maximum Retries</li> <li>- Retry Interval (seconds)</li> </ul> </li> <li>• <b>No:</b> The node will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Max. Node Execution Duration</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>
Failure Policy	Yes	<p>Policies to be performed after the node fails to be executed:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>
Dry run	No	<p>If you select this option, the node will not be executed, and a success message will be returned.</p>




**Table 3-186** Lineage

Parameter	Description
Input	



Parameter	Description
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● <b>DWS</b> <ul style="list-style-type: none"> <li>– <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>– <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>– <b>Schema:</b> Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>– <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● <b>OBS</b> <ul style="list-style-type: none"> <li>– <b>Path:</b> Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● <b>CSS</b> <ul style="list-style-type: none"> <li>– <b>Cluster Name:</b> Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>– <b>Index:</b> Enter a CSS index name.</li> </ul> </li> <li>● <b>HIVE</b> <ul style="list-style-type: none"> <li>– <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>– <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>– <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● <b>CUSTOM</b> <ul style="list-style-type: none"> <li>– <b>Name:</b> Enter a name of the CUSTOM type.</li> <li>– <b>Attribute:</b> Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● <b>DLI</b> <ul style="list-style-type: none"> <li>– <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>– <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>– <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.

Parameter	Description
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the input lineage.
<b>Output</b>	
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● DWS <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>- <b>Schema:</b> Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● OBS <ul style="list-style-type: none"> <li>- <b>Path:</b> Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● CSS <ul style="list-style-type: none"> <li>- <b>Cluster Name:</b> Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>- <b>Index:</b> Enter a CSS index name.</li> </ul> </li> <li>● HIVE <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● CUSTOM <ul style="list-style-type: none"> <li>- <b>Name:</b> Enter a name of the CUSTOM type.</li> <li>- <b>Attribute:</b> Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● DLI <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>

Parameter	Description
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the output lineage.

### 3.4.9.13 MRS Hive SQL


#### Function

The MRS Hive SQL node is used to execute a predefined Hive SQL script on DLF.

#### Parameters

[Table 3-187](#), [Table 3-188](#), and [Table 3-189](#) describe the parameters of the MRS Hive SQL node.

**Table 3-187** Parameters of MRS Hive SQL nodes

Parameter	Man dator y	Description
SQL Script	Yes	Path of a script to be executed. If no script is available, create and develop a script by referring to <a href="#">Creating a Script</a> and <a href="#">Developing an SQL Script</a> .
Data Connection	Yes	Data connection that is configured in the SQL script. The value can be changed.
Database	Yes	Database that is configured in the SQL script. The value can be changed.
Script Parameter	No	If the associated SQL script uses a parameter, the parameter name is displayed. Set the parameter value in the text box next to the parameter name. The parameter value can be <a href="#">an EL expression</a> . If the parameters of the associated SQL script are changed, click  to refresh the parameters.

Parameter	Mandatory	Description
Program Parameter	No	Used to configure optimization parameters such as threads, memory, and vCPUs for the job to optimize resource usage and improve job execution performance.  <b>NOTE</b> This parameter is mandatory if the cluster version is MRS 1.8.7 or later than MRS 2.0.1.  For details about the program parameters of MRS Hive SQL jobs, see <b>Managing an Existing Cluster &gt; Job Management &gt; Running a Hive SQL Job</b> in the <i>MapReduce Service (MRS) User Guide</i> .
Node Name	Yes	Name of the SQL script. The value can be changed. The rules are as follows:  Node name. It contains 1 to 64 characters. Only letters, digits, hyphens (-), and underscores (_) are allowed.  <b>NOTE</b> The node name cannot contain more than 64 characters. If the node name does not meet requirements, the MRS job will fail to be submitted.


**Table 3-188** Advanced parameters



Parameter	Mandatory	Description
Node Status Polling Interval (s)	Yes	How often the system checks completeness of the node task. The value ranges from 1 to 60 seconds.
Max. Node Execution Duration	Yes	Maximum duration of executing a node. When <b>Retry upon Failure</b> is set to <b>Yes</b> for a node, the node can be re-executed for numerous times upon an execution failure within the maximum duration.




Parameter	Mandatory	Description
Retry upon Failure	Yes	<p>Whether to re-execute a node after the node fails to be executed.</p> <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node will be re-executed after it fails to be executed. The following parameters must be configured: <ul style="list-style-type: none"> <li>- Maximum Retries</li> <li>- Retry Interval (seconds)</li> </ul> </li> <li>• <b>No:</b> The node will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Max. Node Execution Duration</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>
Failure Policy	Yes	<p>Policies to be performed after the node fails to be executed:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>
Dry run	No	<p>If you select this option, the node will not be executed, and a success message will be returned.</p>

**Table 3-189** Lineage

Parameter	Description
Input	

Parameter	Description
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● <b>DWS</b> <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>- <b>Schema:</b> Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● <b>OBS</b> <ul style="list-style-type: none"> <li>- <b>Path:</b> Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● <b>CSS</b> <ul style="list-style-type: none"> <li>- <b>Cluster Name:</b> Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>- <b>Index:</b> Enter a CSS index name.</li> </ul> </li> <li>● <b>HIVE</b> <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● <b>CUSTOM</b> <ul style="list-style-type: none"> <li>- <b>Name:</b> Enter a name of the CUSTOM type.</li> <li>- <b>Attribute:</b> Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● <b>DLI</b> <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.

Parameter	Description
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the input lineage.
<b>Output</b>	
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● DWS <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>- <b>Schema:</b> Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● OBS <ul style="list-style-type: none"> <li>- <b>Path:</b> Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● CSS <ul style="list-style-type: none"> <li>- <b>Cluster Name:</b> Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>- <b>Index:</b> Enter a CSS index name.</li> </ul> </li> <li>● HIVE <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● CUSTOM <ul style="list-style-type: none"> <li>- <b>Name:</b> Enter a name of the CUSTOM type.</li> <li>- <b>Attribute:</b> Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● DLI <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>

Parameter	Description
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the output lineage.

### 3.4.9.14 MRS Presto SQL

#### Function

The MRS Presto SQL node is used to execute the Presto SQL script predefined in DataArts Factory.


#### Parameters

[Table 3-190](#), [Table 3-191](#), and [Table 3-192](#) describe the parameters of the MRS Presto SQL node.

**Table 3-190** Property parameters

Parameters	Mandatory	Description
SQL or Script	Yes	<p>You can select <b>SQL statement</b> or <b>SQL script</b>.</p> <ul style="list-style-type: none"> <li>SQL statement In the <b>Statements</b> text box, enter the SQL statement to be executed.</li> <li>SQL script Select a script to be executed. If no script is available, create and develop a script by referring to <a href="#">Creating a Script</a> and <a href="#">Developing an SQL Script</a>.</li> </ul> <p><b>NOTE</b> If you select the SQL statement mode, the DataArts Factory module cannot parse the parameters contained in the SQL statement.</p>
Data Connection	Yes	Data connection that is configured in the SQL script. The value can be changed.



Parameters	Mandatory	Description
Schema	Yes	Database that is configured in the SQL script. The value can be changed.
Script Parameter	No	If the associated SQL script uses a parameter, the parameter name is displayed. Set the parameter value in the text box next to the parameter name. The parameter value can be <b>an EL expression</b> . If the parameters of the associated SQL script are changed, click  to refresh the parameters.
Node Name	Yes	Name of the SQL script. The value can be changed. Node name. It contains 1 to 64 characters. Only letters, digits, hyphens (-), and underscores (_) are allowed. <b>NOTE</b> The node name cannot contain more than 64 characters. If the node name does not meet requirements, the MRS job will fail to be submitted.


**Table 3-191** Advanced parameters



Parameters	Mandatory	Description
Node Status Polling Interval (s)	Yes	How often the system checks completeness of the node task. The value ranges from 1 to 60 seconds.
Max. Node Execution Duration	Yes	Maximum duration of executing a node. When <b>Retry upon Failure</b> is set to <b>Yes</b> for a node, the node can be re-executed for numerous times upon an execution failure within the maximum duration.
Retry upon Failure	Yes	Whether to re-execute a node after the node fails to be executed. <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node will be re-executed after it fails to be executed. The following parameters must be configured: <ul style="list-style-type: none"> <li>- Maximum Retries</li> <li>- Retry Interval (seconds)</li> </ul> </li> <li>• <b>No:</b> The node will not be re-executed. This is the default setting.</li> </ul> <b>NOTE</b> If <b>Max. Node Execution Duration</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.




Parameters	Mandatory	Description
Failure Policy	Yes	<p>Policies to be performed after the node fails to be executed:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>
Dry run	No	If you select this option, the node will not be executed, and a success message will be returned.

**Table 3-192** Lineage

Parameters	Description
Input	

Parameters	Description
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● <b>DWS</b> <ul style="list-style-type: none"> <li>– <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>– <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>– <b>Schema:</b> Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>– <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● <b>OBS</b> <ul style="list-style-type: none"> <li>– <b>Path:</b> Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● <b>CSS</b> <ul style="list-style-type: none"> <li>– <b>Cluster Name:</b> Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>– <b>Index:</b> Enter a CSS index name.</li> </ul> </li> <li>● <b>HIVE</b> <ul style="list-style-type: none"> <li>– <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>– <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>– <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● <b>CUSTOM</b> <ul style="list-style-type: none"> <li>– <b>Name:</b> Enter a name of the CUSTOM type.</li> <li>– <b>Attribute:</b> Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● <b>DLI</b> <ul style="list-style-type: none"> <li>– <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>– <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>– <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.

Parameters	Description
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the input lineage.
<b>Output</b>	
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● DWS <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>- <b>Schema:</b> Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● OBS <ul style="list-style-type: none"> <li>- <b>Path:</b> Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● CSS <ul style="list-style-type: none"> <li>- <b>Cluster Name:</b> Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>- <b>Index:</b> Enter a CSS index name.</li> </ul> </li> <li>● HIVE <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● CUSTOM <ul style="list-style-type: none"> <li>- <b>Name:</b> Enter a name of the CUSTOM type.</li> <li>- <b>Attribute:</b> Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● DLI <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>

Parameters	Description
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the output lineage.

### 3.4.9.15 MRS Spark


#### Function

The MRS Spark node is used to execute a predefined Spark job on MRS.

#### Parameters

[Table 3-193](#), [Table 3-194](#), and [Table 3-195](#) describe the parameters of the MRS Spark node.

**Table 3-193** Parameters of MRS Spark nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. The value must consist of 1 to 128 characters and contain only letters, digits, and the following special characters: _-/<>
MRS Cluster Name	Yes	Select the MRS cluster. To create an MRS cluster, use either of the following methods: <ul style="list-style-type: none"> <li>Click . On the <b>Clusters</b> page, create an MRS cluster.</li> <li>Go to the MRS console to create an MRS cluster.</li> </ul>
Spark Job Name	Yes	Name of an MRS job. The name contains 1 to 64 characters, including only letters, digits, and underscores (_). <b>NOTE</b> The job name cannot contain more than 64 characters. If the job name does not meet requirements, the MRS job will fail to be submitted.

Parameter	Mandatory	Description
JAR Package	Yes	Select <b>JAR package</b> . Before selecting a JAR package, upload the JAR package to the OBS bucket, create a resource on the <b>Manage Resource</b> page, and add the JAR package to the resource management list. For details, see <a href="#">Creating a Resource</a> .
JAR File Parameters	No	Parameters of the JAR package.
Program Parameter	No	Used to configure optimization parameters such as threads, memory, and vCPUs for the job to optimize resource usage and improve job execution performance. <b>NOTE</b> This parameter is mandatory if the cluster version is MRS 1.8.7 or later than MRS 2.0.1. For details about the program parameters of MRS Spark jobs, see <b>Managing an Existing Cluster &gt; Job Management &gt; Running a Spark Job</b> in the <i>MapReduce Service (MRS) User Guide</i> .
Input Data Path	No	Path where the input data resides.
Output Data Path	No	Path where the output data resides.


**Table 3-194** Advanced parameters

Parameter	Mandatory	Description
Node Status Polling Interval (s)	Yes	How often the system checks completeness of the node task. The value ranges from 1 to 60 seconds.
Max. Node Execution Duration	Yes	Maximum duration of executing a node. When <b>Retry upon Failure</b> is set to <b>Yes</b> for a node, the node can be re-executed for numerous times upon an execution failure within the maximum duration.



Parameter	Mandatory	Description
Retry upon Failure	Yes	<p>Whether to re-execute a node after the node fails to be executed.</p> <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node will be re-executed after it fails to be executed. The following parameters must be configured: <ul style="list-style-type: none"> <li>- Maximum Retries</li> <li>- Retry Interval (seconds)</li> </ul> </li> <li>• <b>No:</b> The node will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Max. Node Execution Duration</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>
Failure Policy	Yes	<p>Policies to be performed after the node fails to be executed:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>
Dry run	No	<p>If you select this option, the node will not be executed, and a success message will be returned.</p>




**Table 3-195** Lineage

Parameter	Description
Input	

Parameter	Description
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● <b>DWS</b> <ul style="list-style-type: none"> <li>– <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>– <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>– <b>Schema:</b> Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>– <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● <b>OBS</b> <ul style="list-style-type: none"> <li>– <b>Path:</b> Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● <b>CSS</b> <ul style="list-style-type: none"> <li>– <b>Cluster Name:</b> Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>– <b>Index:</b> Enter a CSS index name.</li> </ul> </li> <li>● <b>HIVE</b> <ul style="list-style-type: none"> <li>– <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>– <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>– <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● <b>CUSTOM</b> <ul style="list-style-type: none"> <li>– <b>Name:</b> Enter a name of the CUSTOM type.</li> <li>– <b>Attribute:</b> Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● <b>DLI</b> <ul style="list-style-type: none"> <li>– <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>– <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>– <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.



Parameter	Description
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the input lineage.
<b>Output</b>	
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● DWS <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>- <b>Schema:</b> Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● OBS <ul style="list-style-type: none"> <li>- <b>Path:</b> Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● CSS <ul style="list-style-type: none"> <li>- <b>Cluster Name:</b> Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>- <b>Index:</b> Enter a CSS index name.</li> </ul> </li> <li>● HIVE <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● CUSTOM <ul style="list-style-type: none"> <li>- <b>Name:</b> Enter a name of the CUSTOM type.</li> <li>- <b>Attribute:</b> Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● DLI <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>

Parameter	Description
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the output lineage.

### 3.4.9.16 MRS Spark Python

#### Function

The MRS Spark Python node is used to execute a predefined Spark Python job on MRS.


For details about how to use the MRS Spark Python operator, see [Developing an MRS Spark Python Job](#).

#### Parameters

[Table 3-196](#), [Table 3-197](#), and [Table 3-198](#) describe the parameters of the MRS Spark Python node.

**Table 3-196** Parameters of MRS Spark Python nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. The value must consist of 1 to 128 characters and contain only letters, digits, and the following special characters: _-/<>

Parameter	Mandatory	Description
MRS Cluster Name	Yes	<p>Select an MRS cluster that supports Spark Python. Only a specific version of MRS supports Spark Python. Test the cluster first to ensure that it supports Spark Python.</p> <p>To create an MRS cluster, use either of the following methods:</p> <ul style="list-style-type: none"> <li>Click . On the <b>Clusters</b> page, create an MRS cluster.</li> <li>Go to the MRS console to create an MRS cluster.</li> </ul> <p>For details about how to create a cluster, see section "Custom Purchase of a Cluster" in <i>MapReduce Service (MRS) Usage Guide</i>.</p>
Job Name	Yes	<p>Name of an MRS job. The name contains 1 to 64 characters, including only letters, digits, and underscores (_).</p> <p><b>NOTE</b> The job name cannot contain more than 64 characters. If the job name does not meet requirements, the MRS job will fail to be submitted.</p>
Parameter	Yes	Enter the parameters of the executable program of MRS. Use <b>Enter</b> to separate multiple parameters.
Attribute	No	Enter parameters in the key=value format. Use <b>Enter</b> to separate multiple parameters.


**Table 3-197** Advanced parameters



Parameter	Mandatory	Description
Max. Node Execution Duration	Yes	Maximum duration of executing a node. When <b>Retry upon Failure</b> is set to <b>Yes</b> for a node, the node can be re-executed for numerous times upon an execution failure within the maximum duration.




Parameter	Mandatory	Description
Retry upon Failure	Yes	<p>Whether to re-execute a node after the node fails to be executed.</p> <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node will be re-executed after it fails to be executed. The following parameters must be configured: <ul style="list-style-type: none"> <li>– Maximum Retries</li> <li>– Retry Interval (seconds)</li> </ul> </li> <li>• <b>No:</b> The node will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Max. Node Execution Duration</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>
Failure Policy	Yes	<p>Policies to be performed after the node fails to be executed:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>
Dry run	No	<p>If you select this option, the node will not be executed, and a success message will be returned.</p>

**Table 3-198** Lineage

Parameter	Description
Input	

Parameter	Description
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● <b>DWS</b> <ul style="list-style-type: none"> <li>– <b>Connection Name</b>: Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>– <b>Database</b>: Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>– <b>Schema</b>: Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>– <b>Table Name</b>: Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● <b>OBS</b> <ul style="list-style-type: none"> <li>– <b>Path</b>: Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● <b>CSS</b> <ul style="list-style-type: none"> <li>– <b>Cluster Name</b>: Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>– <b>Index</b>: Enter a CSS index name.</li> </ul> </li> <li>● <b>HIVE</b> <ul style="list-style-type: none"> <li>– <b>Connection Name</b>: Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>– <b>Database</b>: Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>– <b>Table Name</b>: Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● <b>CUSTOM</b> <ul style="list-style-type: none"> <li>– <b>Name</b>: Enter a name of the CUSTOM type.</li> <li>– <b>Attribute</b>: Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● <b>DLI</b> <ul style="list-style-type: none"> <li>– <b>Connection Name</b>: Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>– <b>Database</b>: Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>– <b>Table Name</b>: Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.

Parameter	Description
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the input lineage.
<b>Output</b>	
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● DWS <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>- <b>Schema:</b> Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● OBS <ul style="list-style-type: none"> <li>- <b>Path:</b> Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● CSS <ul style="list-style-type: none"> <li>- <b>Cluster Name:</b> Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>- <b>Index:</b> Enter a CSS index name.</li> </ul> </li> <li>● HIVE <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● CUSTOM <ul style="list-style-type: none"> <li>- <b>Name:</b> Enter a name of the CUSTOM type.</li> <li>- <b>Attribute:</b> Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● DLI <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>

Parameter	Description
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the output lineage.

### 3.4.9.17 MRS Flink Job


#### Function

The MRS Flink node is used to execute predefined Flink jobs in MRS.

#### Parameters

[Table 3-199](#) and [Table 3-200](#) describe the parameters of the MRS Flink node.

**Table 3-199** Parameters of the MRS Flink node

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. The value must consist of 1 to 128 characters and contain only letters, digits, and the following special characters: _-/<>
MRS Cluster Name	Yes	Select the MRS cluster. To create an MRS cluster, use either of the following methods: <ul style="list-style-type: none"> <li>Click . On the <b>Clusters</b> page, create an MRS cluster.</li> <li>Go to the MRS console to create an MRS cluster.</li> </ul>
Job Name	Yes	Name of an MRS job. The name contains 1 to 64 characters, including only letters, digits, and underscores (_). <b>NOTE</b> The job name cannot contain more than 64 characters. If the job name does not meet requirements, the MRS job will fail to be submitted.

Parameter	Mandatory	Description
Job Resource Package	Yes	Select a JAR package. Before selecting a JAR package, upload the JAR package to the OBS bucket, create a resource on the <b>Manage Resource</b> page, and add the JAR package to the resource management list. For details, see <a href="#">Creating a Resource</a> .
Job Execution Parameter	No	Key parameter of the program that executes the Flink job. This parameter is specified by a function in the user program. Multiple parameters are separated by space.
Program Parameter	No	Used to configure optimization parameters such as threads, memory, and vCPUs for the job to optimize resource usage and improve job execution performance. <b>NOTE</b> This parameter is mandatory if the cluster version is MRS 1.8.7 or later than MRS 2.0.1. For details about the program parameters of MRS Flink jobs, see <b>Managing an Existing Cluster &gt; Job Management &gt; Running a Flink Job</b> in the <i>MapReduce Service (MRS) User Guide</i> .
Input Data Path	No	Path where the input data resides.
Output Data Path	No	Path where the output data resides.

**Table 3-200** Advanced parameters

Parameters	Mandatory	Description
Max. Node Execution Duration	Yes	Maximum duration of executing a node. When <b>Retry upon Failure</b> is set to <b>Yes</b> for a node, the node can be re-executed for numerous times upon an execution failure within the maximum duration.



Parameters	Mandatory	Description
Retry upon Failure	Yes	<p>Whether to re-execute a node after the node fails to be executed.</p> <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node will be re-executed after it fails to be executed. The following parameters must be configured: <ul style="list-style-type: none"> <li>– Maximum Retries</li> <li>– Retry Interval (seconds)</li> </ul> </li> <li>• <b>No:</b> The node will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Max. Node Execution Duration</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>
Failure Policy	Yes	<p>Policies to be performed after the node fails to be executed:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>
Dry run	No	<p>If you select this option, the node will not be executed, and a success message will be returned.</p>

### 3.4.9.18 MRS MapReduce


#### Function

The MRS node is used to execute a predefined MapReduce program on MRS.

#### Parameters

[Table 3-201](#) and [Table 3-202](#) describe the parameters of the MRS node.

**Table 3-201** Parameters of MRS MapReduce nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. The value must consist of 1 to 128 characters and contain only letters, digits, and the following special characters: _-/<>
MRS Cluster Name	Yes	Select the MRS cluster. To create an MRS cluster, use either of the following methods: <ul style="list-style-type: none"> <li>Click . On the <b>Clusters</b> page, create an MRS cluster.</li> <li>Go to the MRS console to create an MRS cluster.</li> </ul>
MapReduce Job Name	Yes	Name of an MRS job. The name contains 1 to 64 characters, including only letters, digits, and underscores (_). <b>NOTE</b> The job name cannot contain more than 64 characters. If the job name does not meet requirements, the MRS job will fail to be submitted.
JAR Package	Yes	Select a JAR package. Before selecting a JAR package, upload the JAR package to the OBS bucket, create a resource on the <b>Manage Resource</b> page, and add the JAR package to the resource management list. For details, see <a href="#">Creating a Resource</a> .
JAR File Parameters	No	Parameters of the JAR package.
Input Data Path	No	Path where the input data resides.
Output Data Path	No	Path where the output data resides.

**Table 3-202** Advanced parameters

Parameter	Mandatory	Description
Node Status Polling Interval (s)	Yes	How often the system checks completeness of the node task. The value ranges from 1 to 60 seconds.
Max. Node Execution Duration	Yes	Maximum duration of executing a node. When <b>Retry upon Failure</b> is set to <b>Yes</b> for a node, the node can be re-executed for numerous times upon an execution failure within the maximum duration.

Parameter	Mandatory	Description
Retry upon Failure	Yes	<p>Whether to re-execute a node after the node fails to be executed.</p> <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node will be re-executed after it fails to be executed. The following parameters must be configured: <ul style="list-style-type: none"> <li>– Maximum Retries</li> <li>– Retry Interval (seconds)</li> </ul> </li> <li>• <b>No:</b> The node will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Max. Node Execution Duration</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>
Failure Policy	Yes	<p>Policies to be performed after the node fails to be executed:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>
Dry run	No	<p>If you select this option, the node will not be executed, and a success message will be returned.</p>

### 3.4.9.19 CSS

#### Functions

The CSS node is used to process CSS requests and enable online distributed searching.

#### Parameters

[Table 3-203](#) and [Table 3-204](#) describe the parameters of the CSS node.

**Table 3-203** Parameters of CSS nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of a node. The name must contain 1 to 128 characters, including only letters, numbers, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
CloudSearch Cluster	Yes	Connection to CloudSearch. A CloudSearch cluster has been created in CloudService. Currently, only clusters of version 5.5.1 is supported.
CDM Cluster Name	Yes	Name of the selected CDM cluster. The CDM cluster functions as a proxy to forward requests. If there are no CDM clusters available in the drop-down list, create one on the CDM console.
Request Type	Yes	Possible values: <ul style="list-style-type: none"> <li>• GET</li> <li>• POST</li> <li>• PUT</li> <li>• HEAD</li> <li>• DELETE</li> </ul>
Request Parameter	No	Parameter of the request. For example, to query the dlfddata mapping type in the dlf_search index, set this parameter to: <b>/dlf_search/dlfddata/_search</b>
Request Body	No	The request body is in JSON format.
CloudSearch Output Path	No	Path where output data is to be stored.

**Table 3-204** Advanced parameters

Parameter	Mandatory	Description
Node Status Polling Interval (s)	Yes	Specifies how often the system check completeness of the node task. The value ranges from 1 to 60 seconds.
Max. Node Execution Duration	Yes	Execution timeout interval for the node. If retry is configured and the execution is not complete within the timeout interval, the node will not be retried and is set to the failed state.

Parameter	Mandatory	Description
Retry upon Failure	Yes	<p>Indicates whether to re-execute a node task if its execution fails. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> <li>– <b>Maximum Retries</b></li> <li>– <b>Retry Interval (seconds)</b></li> </ul> </li> <li>• <b>No:</b> The node task will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Timeout Interval</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>
Failure Policy	Yes	<p>Operation that will be performed if the node task fails to be executed. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>

### 3.4.9.20 Shell

#### Functions

The Shell node is used to execute a shell script.

 **NOTE**

With EL expression `#{Job.getNodeOutput()}`, you can obtain the desired content (4000 characters at most and counted backwards) in the output of the shell script run by the Shell node.

Example:

To obtain `<name>jack<name1>` from a shell script (script name: shell\_job1) output, enter the following EL expression:

```
#{StringUtil.substringBetween(Job.getNodeOutput("shell_job1"),"<name>","<name1>")}
```

## Parameters

[Table 3-205](#) and [Table 3-206](#) describe the parameters of the Shell node.

**Table 3-205** Parameters of Shell nodes

Parameter	Mandatory	Description
Shell or Script	Yes	<p>You can select <b>Shell statement</b> or <b>Shell script</b>.</p> <ul style="list-style-type: none"> <li>Shell statement In the <b>Shell statement</b> text box, enter the Shell statement to be executed.</li> <li>Shell script Select a script to be executed. If no script is available, create and develop a script by referring to <a href="#">Creating a Script</a> and <a href="#">Developing an SQL Script</a>.</li> </ul> <p><b>NOTE</b> If you select <b>Shell statement</b>, the DataArts Factory module cannot parse the parameters contained in the Shell statement.</p>
Host Connection	Yes	Selects the host where a shell script is to be executed.
Script Parameter	No	Parameter transferred to the script when the shell script is executed. Parameters are separated by spaces. For example: <b>a b c</b> . The parameter must be referenced by the shell script. Otherwise, the parameter is invalid.
Interactive Input	No	Interactive information (passwords for example) provided during shell script execution. Interactive parameters are separated by carriage return characters. The shell script reads parameter values in sequence according to the interaction situation.
Node Name	Yes	Name of the node. It contains a maximum of 128 characters, including letters, digits, hyphens (-), underscores (_), slashes (/), angle brackets (<>), and periods (.).

**Table 3-206** Advanced parameters

Parameter	Mandatory	Description
Node Status Polling Interval (s)	Yes	Specifies how often the system check completeness of the node task. The value ranges from 1 to 60 seconds.
Max. Node Execution Duration	Yes	Execution timeout interval for the node. If retry is configured and the execution is not complete within the timeout interval, the node will not be retried and is set to the failed state.
Retry upon Failure	Yes	<p>Indicates whether to re-execute a node task if its execution fails. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> <li>- <b>Maximum Retries</b></li> <li>- <b>Retry Interval (seconds)</b></li> </ul> </li> <li>• <b>No:</b> The node task will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Timeout Interval</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>
Failure Policy	Yes	<p>Operation that will be performed if the node task fails to be executed. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>

### 3.4.9.21 RDS SQL

#### Functions

The RDS SQL node is used to transfer SQL statements to RDS.

#### Parameters

[Table 3-207](#) and [Table 3-208](#) describe the parameters of the RDS SQL node.

**Table 3-207** Parameters of RDS SQL nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of a node. The name must contain 1 to 128 characters, including only letters, numbers, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
Data Connection	Yes	Name of the data connection.
Database	Yes	Name of the database. The database has been created. You are advised not to use the default database.
SQL or Script	Yes	<p>You can select <b>SQL statement</b> or <b>SQL script</b>.</p> <ul style="list-style-type: none"> <li>SQL statement In the <b>Statements</b> text box, enter the SQL statement to be executed.</li> <li>SQL script Select a script to be executed. If no script is available, create and develop a script by referring to <a href="#">Creating a Script</a> and <a href="#">Developing an SQL Script</a>.</li> </ul> <p><b>NOTE</b> If you select the SQL statement mode, the DataArts Factory module cannot parse the parameters contained in the SQL statement.</p>

**Table 3-208** Advanced parameters

Parameter	Mandatory	Description
Node Status Polling Interval (s)	Yes	Specifies how often the system check completeness of the node task. The value ranges from 1 to 60 seconds.



Parameter	Mandatory	Description
Max. Node Execution Duration	Yes	Execution timeout interval for the node. If retry is configured and the execution is not complete within the timeout interval, the node will not be retried and is set to the failed state.
Retry upon Failure	Yes	<p>Indicates whether to re-execute a node task if its execution fails. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> <li>– <b>Maximum Retries</b></li> <li>– <b>Retry Interval (seconds)</b></li> </ul> </li> <li>• <b>No:</b> The node task will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Timeout Interval</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>
Failure Policy	Yes	<p>Operation that will be performed if the node task fails to be executed. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>

### 3.4.9.22 ETL Job


#### Functions

The ETL Job node is used to extract data from a specified data source, preprocess the data, and import the data to the target data source.

## Parameters

[Table 3-209](#), [Table 3-210](#), and [Table 3-211](#) describe the parameters of the ETL Job node.

**Table 3-209** Parameters of Transform Load nodes


Parameter	Man dator y	Description
Node Name	Yes	Name of a node. The name must contain 1 to 128 characters, including only letters, numbers, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
ETL Configuration	Yes	<p>Click  to edit the source and destination data to be transformed.</p> <p>The supported source data types are DLI, OBS and MySQL.</p> <ul style="list-style-type: none"> <li>• When the source data type is DLI, the supported destination data types are DWS, GES, CSS, OBS, and DLI.</li> <li>• When the source data type is MySQL, the supported destination data type is MySQL.</li> <li>• When the source data type is OBS, the supported destination data can be of the DLI type and the DWS type.</li> </ul> <p><b>NOTICE</b></p> <ul style="list-style-type: none"> <li>• Data transformation from DLI to DWS: Before importing data from DataArts Factory to DWS, ensure that a DWS data connection and a table have been created.  Before importing data from DLI to DWS, ensure that a DWS table have been created.</li> <li>• Data transformation from DLI to CSS: Before importing data from DLI to CSS, ensure that a cross-source connection associated with CSS has been created on DLI. For details about how to create a cross-source connection on DLI, see <i>Data Lake Insight User Guide</i>.</li> </ul>
Configure SQL Template	No	Click <b>Obtain Template</b> to obtain an SQL template.



**Table 3-210** Advanced parameters




Parameter	Mandatory	Description
Max. Node Execution Duration	Yes	Execution timeout interval for the node. If retry is configured and the execution is not complete within the timeout interval, the node will not be retried and is set to the failed state.
Retry upon Failure	Yes	Indicates whether to re-execute a node task if its execution fails. Possible values: <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> <li>- <b>Maximum Retries</b></li> <li>- <b>Retry Interval (seconds)</b></li> </ul> </li> <li>• <b>No:</b> The node task will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Timeout Interval</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>
Failure Policy	Yes	Operation that will be performed if the node task fails to be executed. Possible values: <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>

**Table 3-211** Lineage

Parameter	Description
Input	

Parameter	Description
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● <b>DWS</b> <ul style="list-style-type: none"> <li>– <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>– <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>– <b>Schema:</b> Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>– <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● <b>OBS</b> <ul style="list-style-type: none"> <li>– <b>Path:</b> Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● <b>CSS</b> <ul style="list-style-type: none"> <li>– <b>Cluster Name:</b> Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>– <b>Index:</b> Enter a CSS index name.</li> </ul> </li> <li>● <b>HIVE</b> <ul style="list-style-type: none"> <li>– <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>– <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>– <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● <b>CUSTOM</b> <ul style="list-style-type: none"> <li>– <b>Name:</b> Enter a name of the CUSTOM type.</li> <li>– <b>Attribute:</b> Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● <b>DLI</b> <ul style="list-style-type: none"> <li>– <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>– <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>– <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.

Parameter	Description
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the input lineage.
<b>Output</b>	
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● DWS <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>- <b>Schema:</b> Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● OBS <ul style="list-style-type: none"> <li>- <b>Path:</b> Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● CSS <ul style="list-style-type: none"> <li>- <b>Cluster Name:</b> Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>- <b>Index:</b> Enter a CSS index name.</li> </ul> </li> <li>● HIVE <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● CUSTOM <ul style="list-style-type: none"> <li>- <b>Name:</b> Enter a name of the CUSTOM type.</li> <li>- <b>Attribute:</b> Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● DLI <ul style="list-style-type: none"> <li>- <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>- <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>- <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>

Parameter	Description
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the output lineage.

### 3.4.9.23 Python

#### Functions

The Python node is used to execute Python statements.

Before using a Python node, ensure that the host connected to the node has an environment for executing Python scripts.

 **NOTE**

Python nodes do not support script parameters or job parameters.

#### Parameters

[Table 3-212](#) and [Table 3-213](#) describe the parameters of the Python node.

**Table 3-212** Parameters of the Python node

Parameter	Mandatory	Description
Python or Script	Yes	<p>You can select <b>Python statement</b> or <b>Python script</b>.</p> <ul style="list-style-type: none"> <li>Python statement In the <b>Python statement</b> text box, enter the Python statement to be executed.</li> <li>Python script Select a script to be executed for <b>Script Path</b>. If no script is available, create and develop a script by referring to <a href="#">Creating a Script</a> and <a href="#">Developing a Python Script</a>.</li> </ul> <p><b>NOTE</b> If you select <b>Python statement</b>, the DataArts Factory module cannot parse the parameters contained in the Python statement.</p>

Parameter	Mandatory	Description
Host Connection	Yes	Select the host where the Python statement is to be executed. Ensure that the host has an environment for executing Python scripts.
Node Name	Yes	Name of the node. The value must consist of 1 to 128 characters and contain only letters, digits, and the following special characters: _-/<>.

**Table 3-213** Advanced parameters

Parameter	Mandatory	Description
Node Status Polling Interval (s)	Yes	Specifies how often the system check completeness of the node task. The value ranges from 1 to 60 seconds.
Max. Node Execution Duration	Yes	Execution timeout interval for the node. If retry is configured and the execution is not complete within the timeout interval, the node will not be retried and is set to the failed state.
Retry upon Failure	Yes	<p>Indicates whether to re-execute a node task if its execution fails. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> <li>- <b>Maximum Retries</b></li> <li>- <b>Retry Interval (seconds)</b></li> </ul> </li> <li>• <b>No:</b> The node task will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Timeout Interval</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>

Parameter	Mandatory	Description
Failure Policy	Yes	<p>Operation that will be performed if the node task fails to be executed. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>

### 3.4.9.24 Create OBS

#### Constraints

This function depends on OBS.

#### Functions

The Create OBS node is used to create buckets and directories on OBS.

#### Parameters

[Table 3-214](#) and [Table 3-215](#) describe the parameters of the Create OBS node.

**Table 3-214** Parameters of Create OBS nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of a node. The name must contain 1 to 128 characters, including only letters, numbers, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).



Parameter	Mandatory	Description
OBS Path	Yes	<p>Path to the OBS bucket or directory.</p> <ul style="list-style-type: none"> <li>To create a bucket, enter <i>//OBS bucket name</i>. The OBS bucket name must be unique</li> <li>To create an OBS directory, select the path to the OBS directory to be created, and enter the <i>/ Directory name</i> following the path. The directory name must be unique.</li> </ul>

**Table 3-215** Advanced parameters

Parameter	Mandatory	Description
Max. Node Execution Duration	Yes	<p>Execution timeout interval for the node. If retry is configured and the execution is not complete within the timeout interval, the node will not be retried and is set to the failed state.</p>
Retry upon Failure	Yes	<p>Indicates whether to re-execute a node task if its execution fails. Possible values:</p> <ul style="list-style-type: none"> <li><b>Yes:</b> The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> <li><b>Maximum Retries</b></li> <li><b>Retry Interval (seconds)</b></li> </ul> </li> <li><b>No:</b> The node task will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Timeout Interval</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>

Parameter	Mandatory	Description
Failure Policy	Yes	<p>Operation that will be performed if the node task fails to be executed. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>

### 3.4.9.25 Delete OBS

#### Constraints

This function depends on OBS.

#### Functions

The Delete OBS node is used to delete a bucket or directory on OBS.

#### Parameters

[Table 3-216](#) and [Table 3-217](#) describe the parameters of the Delete OBS node.

**Table 3-216** Parameters of Delete OBS nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of a node. The name must contain 1 to 128 characters, including only letters, numbers, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).

Parameter	Mandatory	Description
OBS Path	Yes	<p>Path to the OBS bucket or directory.</p> <p><b>NOTE</b> If you delete an OBS bucket or directory, files stored in it are also deleted and cannot be restored. Before you delete a bucket or directory, back up the files stored in it if they need to be retained.</p>

**Table 3-217** Advanced parameters

Parameter	Mandatory	Description
Max. Node Execution Duration	Yes	<p>Execution timeout interval for the node. If retry is configured and the execution is not complete within the timeout interval, the node will not be retried and is set to the failed state.</p>
Retry upon Failure	Yes	<p>Indicates whether to re-execute a node task if its execution fails. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> <li>- <b>Maximum Retries</b></li> <li>- <b>Retry Interval (seconds)</b></li> </ul> </li> <li>• <b>No:</b> The node task will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Timeout Interval</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>

Parameter	Mandatory	Description
Failure Policy	Yes	<p>Operation that will be performed if the node task fails to be executed. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>

### 3.4.9.26 OBS Manager

#### Constraints

This function depends on OBS.

#### Function

The OBS Manager node is used to move or copy files from an OBS bucket to a specified directory.

#### Parameters

[Table 3-218](#), [Table 3-219](#), and [Table 3-220](#) describe the parameters of the OBS Managernode node.

**Table 3-218** Parameters of OBS Manager nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of a node. The name must contain 1 to 128 characters, including only letters, numbers, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).

Parameter	Mandatory	Description
Operation Type	Yes	<p>Operations that can be performed on the node.</p> <ul style="list-style-type: none"> <li>• <b>Move File:</b> moves a source file or directory to a new directory.</li> <li>• <b>Copy File:</b> copies the source file or directory.</li> <li>• <b>Rename File:</b> renames the last level of the directory or file. For example, you can rename the directory <b>obs://test/a/b/c/</b> as <b>obs://test/a/b/d/</b>, and rename the file <b>obs://test/a/b/hello.txt</b> as <b>obs://test/a/b/bye.txt</b>.</li> <li>• <b>Monitor File:</b> checks whether a file or directory exists. If the file or directory exists, the node is executed successfully. Otherwise, the node fails to be executed.</li> </ul>
Source File or Directory	Yes	OBS file or directory to be managed in the OBS bucket.
Target Directory	Yes	Directory for storing OBS files to be moved or copied from the OBS bucket.
File Filter	No	Wildcard for file filtering. Only the files that meet the filtering condition can be moved or copied. If this parameter is not specified, all source files are moved by default. For example, when you enter *.csv, files in this format will be moved or copied.


**Table 3-219** Advanced parameters



Parameter	Mandatory	Description
Max. Node Execution Duration	Yes	Execution timeout interval for the node. If retry is configured and the execution is not complete within the timeout interval, the node will not be retried and is set to the failed state.

Parameter	Mandatory	Description
Retry upon Failure	Yes	<p>Indicates whether to re-execute a node task if its execution fails. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> <li>– <b>Maximum Retries</b></li> <li>– <b>Retry Interval (seconds)</b></li> </ul> </li> <li>• <b>No:</b> The node task will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Timeout Interval</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>
Failure Policy	Yes	<p>Operation that will be performed if the node task fails to be executed. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>




Table 3-220 Lineage

Parameter	Description
Input	

Parameter	Description
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● <b>DWS</b> <ul style="list-style-type: none"> <li>– <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>– <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>– <b>Schema:</b> Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>– <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● <b>OBS</b> <ul style="list-style-type: none"> <li>– <b>Path:</b> Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● <b>CSS</b> <ul style="list-style-type: none"> <li>– <b>Cluster Name:</b> Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>– <b>Index:</b> Enter a CSS index name.</li> </ul> </li> <li>● <b>HIVE</b> <ul style="list-style-type: none"> <li>– <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>– <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>– <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● <b>CUSTOM</b> <ul style="list-style-type: none"> <li>– <b>Name:</b> Enter a name of the CUSTOM type.</li> <li>– <b>Attribute:</b> Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● <b>DLI</b> <ul style="list-style-type: none"> <li>– <b>Connection Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>– <b>Database:</b> Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>– <b>Table Name:</b> Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.

Parameter	Description
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the input lineage.
<b>Output</b>	
Add	<p>Click <b>Add</b>. In the <b>Type</b> drop-down list, select the type to be created. The value can be <b>DWS</b>, <b>OBS</b>, <b>CSS</b>, <b>HIVE</b>, <b>DLI</b>, or <b>CUSTOM</b>.</p> <ul style="list-style-type: none"> <li>● DWS <ul style="list-style-type: none"> <li>- <b>Connection Name</b>: Click <b>...</b>. In the displayed dialog box, select a DWS data connection.</li> <li>- <b>Database</b>: Click <b>...</b>. In the displayed dialog box, select a DWS database.</li> <li>- <b>Schema</b>: Click <b>...</b>. In the displayed dialog box, select a DWS schema.</li> <li>- <b>Table Name</b>: Click <b>...</b>. In the displayed dialog box, select a DWS table.</li> </ul> </li> <li>● OBS <ul style="list-style-type: none"> <li>- <b>Path</b>: Click <b>...</b>. In the displayed dialog box, select an OBS path.</li> </ul> </li> <li>● CSS <ul style="list-style-type: none"> <li>- <b>Cluster Name</b>: Click <b>...</b>. In the displayed dialog box, select a CSS cluster.</li> <li>- <b>Index</b>: Enter a CSS index name.</li> </ul> </li> <li>● HIVE <ul style="list-style-type: none"> <li>- <b>Connection Name</b>: Click <b>...</b>. In the displayed dialog box, select a HIVE data connection.</li> <li>- <b>Database</b>: Click <b>...</b>. In the displayed dialog box, select a HIVE database.</li> <li>- <b>Table Name</b>: Click <b>...</b>. In the displayed dialog box, select a HIVE table.</li> </ul> </li> <li>● CUSTOM <ul style="list-style-type: none"> <li>- <b>Name</b>: Enter a name of the CUSTOM type.</li> <li>- <b>Attribute</b>: Enter an attribute of the CUSTOM type. You can add more than one attribute.</li> </ul> </li> <li>● DLI <ul style="list-style-type: none"> <li>- <b>Connection Name</b>: Click <b>...</b>. In the displayed dialog box, select a DLI data connection.</li> <li>- <b>Database</b>: Click <b>...</b>. In the displayed dialog box, select a DLI database.</li> <li>- <b>Table Name</b>: Click <b>...</b>. In the displayed dialog box, select a DLI table.</li> </ul> </li> </ul>



Parameter	Description
OK	Click <b>OK</b> to save the parameter settings.
Cancel	Click <b>Cancel</b> to cancel the parameter settings.
Modify	Click  to modify the parameter settings. After the modification, save the settings.
Delete	Click  to delete the parameter settings.
View Details	Click  to view details about the table created based on the output lineage.

### 3.4.9.27 Open/Close Resource

#### Functions

You can use the Open/Close Resource node to enable or disable services as required.

#### Parameters

[Table 3-221](#) and [Table 3-222](#) describe the parameters of the Open/Close Resource node.

**Table 3-221** Parameters of Open/Close Resource nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of a node. The name must contain 1 to 128 characters, including only letters, numbers, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
Service	Yes	Service to be opened or closed. <ul style="list-style-type: none"> <li>• ECS</li> <li>• CDM</li> </ul>
Open/Close Resource	Yes	Possible values: <ul style="list-style-type: none"> <li>• On</li> <li>• Off</li> </ul>
Instance	Yes	Object to be opened or closed, for example, to open a CDM cluster.

**Table 3-222** Advanced parameters

Parameter	Mandatory	Description
Node Status Polling Interval (s)	Yes	Specifies how often the system check completeness of the node task. The value ranges from 1 to 60 seconds.
Max. Node Execution Duration	Yes	Execution timeout interval for the node. If retry is configured and the execution is not complete within the timeout interval, the node will not be retried and is set to the failed state.
Retry upon Failure	Yes	<p>Indicates whether to re-execute a node task if its execution fails. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> <li>- <b>Maximum Retries</b></li> <li>- <b>Retry Interval (seconds)</b></li> </ul> </li> <li>• <b>No:</b> The node task will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Timeout Interval</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>
Failure Policy	Yes	<p>Operation that will be performed if the node task fails to be executed. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>

### 3.4.9.28 Subjob

#### Function

The Subjob node is used to call the batch job that does not contain the subjob node.

#### Parameter

[Table 3-223](#) and [Table 3-224](#) describe the parameters of the Subjob node.

**Table 3-223** Parameters of subjob nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of a node. The name must contain 1 to 128 characters, including only letters, numbers, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
Subjob Name	Yes	Select the name of the subjob to be called. <b>NOTE</b> You can only select the name of an existing batch job that does not contain the Subjob node.
Subjob Parameter	Yes/No	<ul style="list-style-type: none"> <li>If the subjob parameters are left unspecified, the subjob is executed with its own parameter variables. The <b>Subjob Parameter Name</b> of the parent job is not displayed.</li> <li>If the subjob parameters are specified, the subjob is executed with the configured parameter values. In this case, the <b>Subjob Parameter Name</b> of the parent job is displayed, and the data or EL expression configured for the subjob is accessed and replaced according to the environment variable of the parent job.</li> </ul>

**Table 3-224** Advanced parameters

Parameter	Mandatory	Description
Node Status Polling Interval (s)	Yes	Specifies how often the system check completeness of the node task. The value ranges from 1 to 60 seconds.

Parameter	Mandatory	Description
Max. Node Execution Duration	Yes	Execution timeout interval for the node. If retry is configured and the execution is not complete within the timeout interval, the node will not be retried and is set to the failed state.
Retry upon Failure	Yes	Indicates whether to re-execute a node task if its execution fails. Possible values: <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> <li>– <b>Maximum Retries</b></li> <li>– <b>Retry Interval (seconds)</b></li> </ul> </li> <li>• <b>No:</b> The node task will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Timeout Interval</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>
Failure Policy	Yes	Operation that will be performed if the node task fails to be executed. Possible values: <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>

### 3.4.9.29 For Each

#### Functions

The For Each node specifies a subjob to be executed cyclically and assigns values to variables in a subjob with a dataset.

## Parameters

[Table 3-225](#) describes the parameters of the For Each node.

**Table 3-225** Parameters of the For Each node

Parameter	Mandatory	Description
Node Name	Yes	Name of a node. The name must contain 1 to 128 characters, including only letters, numbers, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
Subjob in a Loop	Yes	Name of the subjob to be executed cyclically.
Dataset	Yes	The For Each node needs to define a dataset. The dataset is used to cyclically replace variables in a subjob. A row of data in the dataset corresponds to a subjob instance. The dataset may come from the following sources: <ul style="list-style-type: none"> <li>Output from upstream nodes, such as the select statements of the Hive SQL, DLI SQL, or Spark SQL node, and echo of the shell node. The EL expression <code>#<b>{Job.getNodeOutput('preNodeName')}</b>}</code> is used, which means the output of the previous node.</li> <li>A specified array, for example, <code>['001'],['002'],['003']</code></li> </ul>
Concurrent Subjobs	Yes	Subjobs generated cyclically can be executed concurrently. You can set the number of concurrent subjobs.
Subjob Instance Name Suffix	No	Name of the subjob generated by For Each: For Each node name + underscore (_) + suffix. The suffix is configurable. If the suffix is not configured, the suffix increases in ascending order based on the number.
Job Running Parameter	No	This parameter is available only when you set job parameters for a subjob. <ul style="list-style-type: none"> <li>If the subjob parameters are left unspecified, the subjob is executed with its own parameter variables.</li> <li>If the subjob parameters are specified, the subjob is executed with the configured parameter values. The method or EL expression configured for the subjob parameter in the node attribute is read and replaced based on the environment variable of the parent job.</li> </ul>

**Table 3-226** Advanced parameters

Parameter	Mandatory	Description
Max. Node Execution Duration	Yes	Execution timeout interval for the node. If retry is configured and the execution is not complete within the timeout interval, the node will not be retried and is set to the failed state.
Retry upon Failure	Yes	<p>Indicates whether to re-execute a node task if its execution fails. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> <li>– <b>Maximum Retries</b></li> <li>– <b>Retry Interval (seconds)</b></li> </ul> </li> <li>• <b>No:</b> The node task will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Timeout Interval</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>
Failure Policy	Yes	<p>Operation that will be performed if the node task fails to be executed. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>

### 3.4.9.30 SMN

#### Functions

The SMN node is used to send notifications to users.

#### Parameters

[Table 3-227](#) and [Table 3-228](#) describe the parameters of the SMN node.

**Table 3-227** Parameters of SMN nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of a node. The name must contain 1 to 128 characters, including only letters, numbers, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
Topic Name	Yes	Name of the topic. The topic has been created in SMN.
Message Title	No	Title of the message. The title cannot exceed 512 characters.
Message Type	Yes	Format of the message. <ul style="list-style-type: none"> <li>• <b>Text:</b> The message is sent in text format.</li> <li>• <b>JSON:</b> The message is sent in JSON format. You can send different messages to types of subscribers. <ul style="list-style-type: none"> <li>- Manual: You can enter a message in <b>Message Content</b>.</li> <li>- Automatic: Click <b>Generate JSON Message</b>. In the displayed dialog box, enter a message and select a protocol.</li> </ul> </li> <li>• <b>Template:</b> The message is sent in template format, that is, in fixed format. The variables can be processed by tags. <ul style="list-style-type: none"> <li>- Manual: You can enter a message in <b>Message Content</b>.</li> <li>- Automatic: Click <b>Generate Template Message</b>. In the displayed dialog box, select a template name and set the value of <b>tag</b>.</li> </ul> </li> </ul>

Parameter	Mandatory	Description
Message Content	Yes	<p>Message content to be provided. The requirements for entering different types of messages are as follows:</p> <ul style="list-style-type: none"> <li>• <b>Text:</b> The size cannot exceed 10 KB.</li> <li>• <b>JSON:</b> The JSON message must contain the Default protocol and the size cannot exceed 10 KB. Example: <pre> {   "default": "Dear Sir or Madam, this is a default message.",   "email": "Dear Sir or Madam, this is an email message.",   "http": "{message:'Dear Sir or Madam, this is an HTTP message.'}",   "https": "{message:'Dear Sir or Madam, this is an HTTPS message.'}",   "sms": "This is an SMS message." } </pre> </li> <li>• <b>Template:</b> The size cannot exceed 10 KB. Example: <pre> "message_template_name":"confirm_message", "tags":{   "topic_urn":"urn:smn:regionId:xxxx:SMN_01" } </pre> </li> </ul> <p>In the preceding information, <b>message_template_name</b> indicates the template name, and <b>tags</b> indicates all tags in the template.</p> <p>For details about how to configure SMN, see section the <i>Simple Message Notification User Guide</i>.</p>

**Table 3-228** Advanced parameters

Parameter	Mandatory	Description
Max. Node Execution Duration	Yes	Execution timeout interval for the node. If retry is configured and the execution is not complete within the timeout interval, the node will not be retried and is set to the failed state.



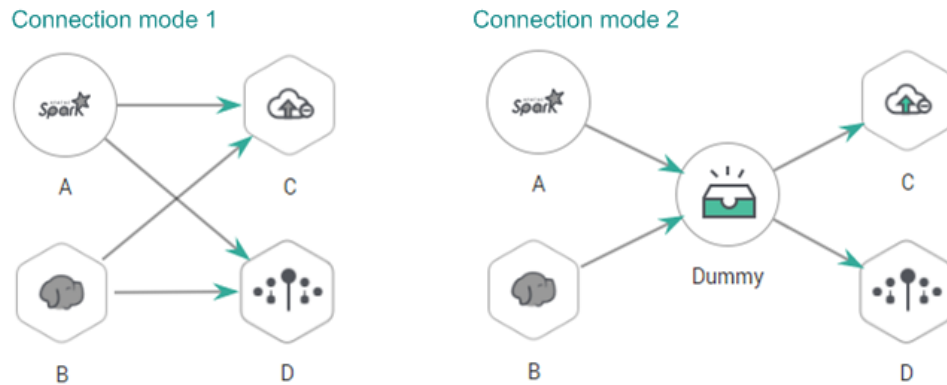
Parameter	Mandatory	Description
Retry upon Failure	Yes	<p>Indicates whether to re-execute a node task if its execution fails. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>Yes:</b> The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> <li>– <b>Maximum Retries</b></li> <li>– <b>Retry Interval (seconds)</b></li> </ul> </li> <li>• <b>No:</b> The node task will not be re-executed. This is the default setting.</li> </ul> <p><b>NOTE</b> If <b>Timeout Interval</b> is configured for the node, the node will not be executed again after the execution times out. Instead, the node is set to the failure state.</p>
Failure Policy	Yes	<p>Operation that will be performed if the node task fails to be executed. Possible values:</p> <ul style="list-style-type: none"> <li>• <b>End the current job execution plan:</b> stops running the current job. The job instance status is <b>Failed</b>.</li> <li>• <b>Go to the next node:</b> ignores the execution failure of the current node. The job instance status is <b>Failure ignored</b>.</li> <li>• <b>Suspend current job execution plan:</b> suspends running the current job. The job instance status is <b>Waiting</b>.</li> <li>• <b>Suspend execution plans of the subsequent nodes:</b> stops running subsequent nodes. The job instance status is <b>Failed</b>.</li> </ul>

### 3.4.9.31 Dummy

#### Functions

The Dummy node is empty and does not perform any operations. It is used to simplify the complex connection relationships of nodes. [Figure 3-229](#) shows an example.

**Figure 3-229** Connection modes



## Parameters

**Table 3-229** describes the parameter of Dummy nodes.

**Table 3-229** Parameter of Dummy nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of a node. The name must contain 1 to 128 characters, including only letters, numbers, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).

## 3.4.10 EL Expression Reference

### 3.4.10.1 Expression Overview

Node parameter values in a DataArts Factory job can be dynamically generated based on the running environment by using Expression Language (EL). You can determine whether to execute this node based on the input parameters of the pipeline and the output of the upstream node. EL uses simple arithmetic and logic to calculate and references embedded objects, including job objects and tool objects.

**Job object:** provides properties and methods of obtaining the output message, job scheduling plan time, and job execution time of the previous node in a job.

**Tool job:** Provides methods of operating character strings, time, and JSON. For example, truncating a substring from a string or formatting time.

## Syntax

Expression syntax:

```
#{expr}
```

In the preceding information, **expr** indicates an expression. **#** and **{ }** are common operators used in EL, allowing you to access job properties using embedded objects.

## Example

In the **URL** parameter of the Rest Client node, use expression **tableName=#{JSONUtil.path(Job.getNodeOutput("get\_cluster"),"tables[0].table\_name")}**.

Expression description:

1. **Job.getNodeOutput("get\_cluster")** is used to obtain the execution result of the **get\_cluster** node in the job. The execution result is a JSON character string.
2. **tables[0].table\_name** is used to obtain the value of a field in the JSON character string.

## Debugging Methods

You can debug EL expressions using the following methods.

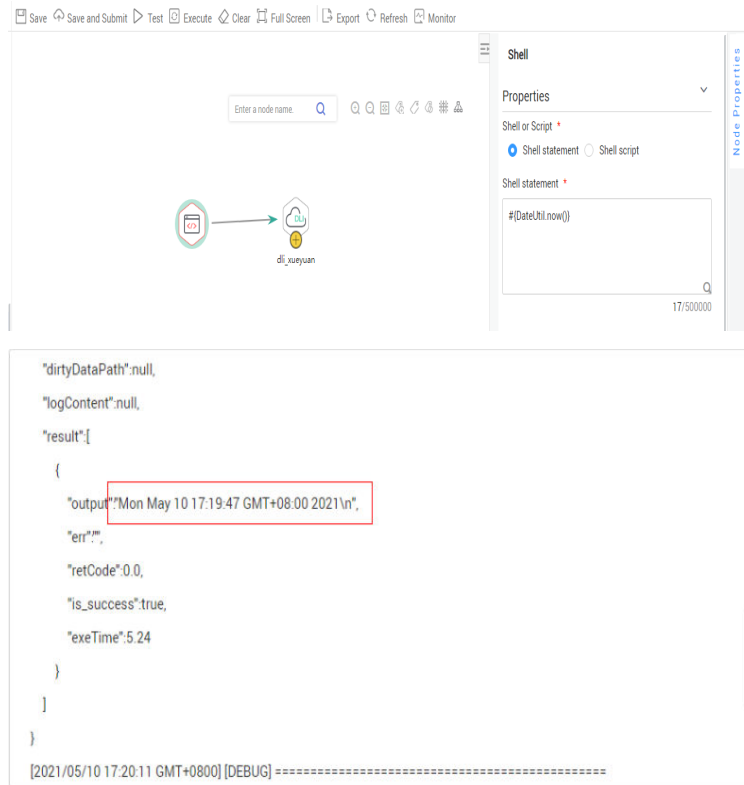
This section uses the **#{DateUtil.now()}** expression as an example.

1. Use the DIS Client node.
  - Prerequisites: A DIS stream is available.
  - Method: Select the DIS Client node, write the EL expression in the data to be sent, and click **Test**. Then right-click the node to view the log. The value of the EL expression is printed in the log.



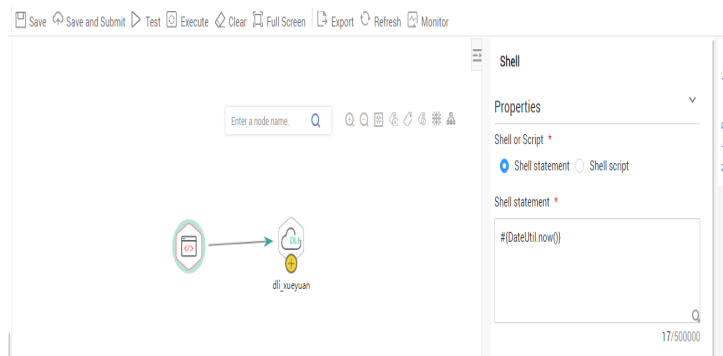
2. Use the Kafka Client node.

- Prerequisites: An MRS cluster with the Kafka component is available.
- Method: Select the Kafka Client node, write the EL expression in the data to be sent, and click **Test**. Then right-click the node to view the log. The value of the EL expression is printed in the log.



3. Use the shell node.

- Prerequisites: An ECS is available.
- Method: Create a host connection, print the EL expression using echo, and click **Test**. Then view the log. The value of the EL expression is printed in the log.



```

"dirtyDataPath":null,
"logContent":null,
"result":[
  {
    "output":"Mon May 10 17:19:47 GMT+08:00 2021\n",
    "err":",
    "retCode":0.0,
    "is_success":true,
    "exeTime":5.24
  }
]
}

```

[2021/05/10 17:20:11 GMT+0800] [DEBUG] =====

4. Use the Create OBS node.

If none of the preceding methods is available, use the Create OBS node and create an OBS path with the value of the EL expression as its name. You can click **Test** and go to the OBS console to view the name of the created path.



### 3.4.10.2 Basic Operators

EL supports most of the arithmetic and logic operators provided by Java.

## Operator List

**Table 3-230** Basic operators

Operator	Description
.	Accesses a Bean property or a mapping entry.
[]	Accesses an array or linked list.
()	Organizes a subexpression to change priority.
+	Plus sign
-	Minus or negative sign
*	Multiplication sign
/ or div	Division sign
% or mod	Modulo
== or eq	Test whether equal to.
!= or ne	Test whether unequal to.
< or lt	Test whether less than.
> or gt	Test whether greater than.
<= or le	Check whether less than or equal to.
>= or ge	Test whether greater than or equal to.
&& or and	Test logic and.
or or	Test logic or.
! or not	Test negation.
empty	Test whether empty.
?:	The expression is similar to if else. If the statement in front of ? is true, the value of the expression between ? and : is returned. Otherwise, the value following : is returned.

### Example

If variable a is empty, default is returned. If variable a is not empty, a itself is returned. The EL expression is as follows:

```
{empty a?"default":a}
```

### 3.4.10.3 Date and Time Mode

The date and time in the EL expression can be displayed in a user-specified format. The date and time format is specified by the date and time mode

character string. The date and time mode character string consists of letters from A to Z and from a to z, as shown in [Table 3-231](#).

**Table 3-231** Letter description

Letter	Description	Example
G	Epoch	AD
y	Year	2001
M	Month in a year	July or 07
d	Day in a month	10
h	Hour in the 12-hour clock	12
H	Hour in the 24-hour clock	22
m	Minute	30
s	Second	55
S	Millisecond	234
E	Day of a week	Mon, Tue, Wed, Thu, Fri, Sat, or Sun
D	Date in the year	360
F	Day in a week of a month	2(second Wed. in July)
w	Week in a year	40
W	Week in a month	1
a	A.M. /P.M.	PM
k	Hour in the 24-hour clock	24
K	Hour in the 12-hour clock	10
z	Time zone	Eastern Standard Time
'	Text delimiter	None
"	Single quotation mark	No example

## Example

To obtain the date of the day before the planned scheduling time of a job, use the following EL expression:

```
#{DateUtil.format(DateUtil.addDays(Job.planTime,-1),"yyyy-MM-dd")}
```

### 3.4.10.4 Env Embedded Objects

An Env embedded object provides a method of obtaining an environment variable value.

#### Method

**Table 3-232** Method description

Method	Description
String get(String name)	Obtains the value of a specified environment variable.

#### Example

The EL expression used to obtain the value of environment variable **test** is as follows:

```
#{Env.get("test")}
```

### 3.4.10.5 Job Embedded Objects

A job object provides properties and methods of obtaining the output message, job scheduling plan time, and job execution time of the previous node in a job.

#### Properties and Methods

**Table 3-233** Property description

Property	Type	Description
name	String	Job name.
planTime	java.util.Date	Job scheduling plane time, that is, the time configured for periodic scheduling, for example, to schedule a job at 1:01 a.m. every day.
startTime	java.util.Date	Job execution time. It may be the same as or later than the planTime (because the job engine is busy).
eventData	String	Message obtained from the stream when the event-driven scheduling is used.
projectId	String	ID of the project where the DataArts Factory module is located.



**Table 3-234** Method description

Method	Description
String getNodeStatus(String nodeName)	<p>Obtains the running status of a specified node. If the node runs properly, success is returned. If the node fails to run, fail is returned.</p> <p>For example, to check whether a node is running successfully, you can use the following command, where <b>test</b> indicates the node name:</p> <pre><b>#{(Job.getNodeStatus("test")) == "success" }</b></pre>
String getNodeOutput(String nodeName)	<p>Obtains the output of a specified node. This method can only obtain the output of the previous dependent node.</p>
String getParam(String key)	<p>Obtains job parameters.</p> <p>This method only obtains the parameter values configured for the current job, but not parameter values passed from the parent job or the global variables configured for the workspace.</p> <p>To obtain the parameter values passed from the parent job and the global variables configured for the workspace, you are advised to use the <code>#{job_param_name}</code> expression.</p>
String getPlanTime(String pattern)	<p>Obtains the plan time character string in a specified pattern. Pattern indicates the date and time mode. For details, see <a href="#">Date and Time Mode</a>.</p>
String getYesterday(String pattern)	<p>Obtains the time character string of the day before the plan time. Pattern indicates the date and time mode. For details, see <a href="#">Date and Time Mode</a>.</p>
String getLastHour(String pattern)	<p>Obtains the time character string of last hour before the plan time. Pattern indicates the date and time mode. For details, see <a href="#">Date and Time Mode</a>.</p>
String getRunningData(String nodeName)	<p>Obtains the data recorded during the running of a specified node. This method can only obtain the output of the previous dependent node. Currently, only the IDs of the DLI jobs recorded during the running of the DLI SQL node can be obtained. For example, to obtain the job ID of the third statement on DLI node <b>DLI_INSERT_DATA</b>, run the following command:</p> <pre><b>#{JSONUtil.path(Job.getRunningData("DLI_INSERT_DATA"),"jobIds[2])}</b></pre>

Method	Description
String getInsertJobId(String nodeName)	Returns the job ID in the first DLI Insert SQL statement of the specified DLI SQL or Transform Load node. If the <b>nodeName</b> parameter is not specified, the job ID in the first DLI Insert SQL statement of the DLI SQL node is obtained. If the job ID cannot be obtained, the <b>null</b> value is returned.

## Example

The expression used to obtain the output of node **test** in the job is as follows:

```
#{Job.getNodeOutput("test")}
```

### 3.4.10.6 StringUtil Embedded Objects

A StringUtil embedded object provides methods of operating character strings, for example, truncating a substring from a character string.

StringUtil is implemented through org.apache.commons.lang3.StringUtils. For details about how to use the object, see the apache commons document.

## Example

If variable a is character string No.0010, the substring after . is returned. The EL expression is as follows:

```
#{StringUtil.substringAfter(a,".")}
```

### 3.4.10.7 DateUtil Embedded Objects

A DateUtil embedded object provides methods of formatting time and calculating time.

## Methods

**Table 3-235** Method description

Method	Description
String format(Date date, String pattern)	Formats Date to character strings according to the specified pattern.
Date addMonths(Date date, int amount)	After the specified number of months is added to Date, the new Date object is returned. The amount can be a negative number.

Method	Description
Date addDays(Date date, int amount)	After the specified number of days is added to Date, the new Date object is returned. The amount can be a negative number.
Date addHours(Date date, int amount)	After the specified number of hours is added to Date, the new Date object is returned. The amount can be a negative number.
Date addMinutes(Date date, int amount)	After the specified number of minutes is added to Date, the new Date object is returned. The amount can be a negative number.
int getDay(Date date)	Obtains the day from the date. For example, if the date is 2018-09-14, 14 is returned.
int getMonth(Date date)	Obtains the month from the date. For example, if the date is 2018-09-14, 9 is returned.
int getYear(Date date)	Obtains the year from the date. For example, if the date is 2018-09-14, 2018 is returned.
Date now()	Returns the current time.
long getTime(Date date)	Converts the date type to the long type.
Date parseDate(String str, String pattern)	Converts the character string to the date by pattern. The pattern is the date and time mode. For details, see <a href="#">Date and Time Mode</a> .

## Example

The previous day of the job scheduling plan time is used as the subdirectory name to generate an OBS path. The EL expression is as follows:

```
#{"obs://test/"+DateUtil.format(DateUtil.addDays(Job.planTime,-1),"yyyy-MM-dd")}
```

### 3.4.10.8 JSONUtil Embedded Objects

A JSONUtil embedded object provides JSON object methods.

## Methods

**Table 3-236** Method description

Method	Description
Object parse(String jsonStr)	Converts a JSON character string into an object.
String toString(Object jsonObject)	Converts an object to a JSON character string.
Object path(String jsonStr,String jsonPath)	Returns the field value in a path specified by the JSON character string. This method is similar to XPath and can be used to retrieve or set JSON by path. You can use . or [] in the path to access members and values. For example, tables[0].table_name.

## Example

The content of variable str is as follows:

```
{
  "cities": [{
    "name": "city1",
    "areaCode": "1000"
  },
  {
    "name": "city2",
    "areaCode": "2000"
  },
  {
    "name": "city3",
    "areaCode": "3000"
  }
]}
```

The expression for obtaining the area code of city1 is as follows:

```
#{JSONUtil.path(str,"cities[0].areaCode")}
```

### 3.4.10.9 Loop Embedded Objects

You can use Loop embedded objects to obtain data from the For Each dataset.

## Property

**Table 3-237** Property description

Property	Type	Description
dataArray	String	Dataset input by the For Each node. It is a two-dimensional array.

Property	Type	Description
current	String	Data row traversed by the For Each node. It is a one-dimensional array.
offset	Int	Current offset of the For Each node, starting from 0. Loop.dataArray[Loop.offset] = Loop.current.

## Example

The EL expression for the Foreach operator to cyclically obtain the first column of the output (a two-dimensional array) of the previous node is as follows:

```
#{Loop.current[0]}
```

### 3.4.10.10 OBSUtil Embedded Objects

The OBSUtil embedded objects provide a series of OBS operation methods, for example, checking whether an OBS file or directory exists.

## Methods

**Table 3-238** Method description

Method	Description
boolean isExistOBSPath(String obsPath)	Check whether the OBS file or the OBS directory that ends with a slash (/) exists. If the file or directory exists, <b>true</b> is returned. If not, <b>false</b> is returned.

## Examples

- The following is the EL expression for checking whether the OBS directory that ends with a slash (/) exists:  
#{OBSUtil.isExistOBSPath("obs://test/jobs/")}
- The following is the EL expression for checking whether the OBS file exists:  
#{OBSUtil.isExistOBSPath("obs://test/jobs/job.log")}

### 3.4.10.11 Expression Use Example

With this example, you can understand how to use EL expressions in the following applications:

- Using variables in the SQL script of DataArts Factory

- Transferring parameters to SQL script variables?
- Using EL expressions in parameters?

## Context

Use the job orchestration and job scheduling functions to generate daily transaction statistics reports according to transaction details tables.

The tables involved in this example are as follows:

- `trade_log`: This table records data generated in each transaction.
- `trade_report`: This table is generated based on `trade_log` and records the daily transaction summary.

## Prerequisites


- A DLI data connection named **dli\_demo** has been created.  
If this data connection is not created, create one. For details, see [Creating Data Connections](#).
- A database named **dli\_db** has been created in DLI.  
If this database is not created, create one. For details, see [Creating a Database](#).
- Tables **trade\_log** and **trade\_report** have been created in the **dli\_db** database.  
If the tables are not created, create them. For details, see [Creating a Table](#).

## Procedure

### Step 1 Create and develop a SQL script.

1. In the navigation tree of the DataArts Factory console, choose **Data Development > Develop Script**.
2. Access the area on the right and choose **Create SQL Script > DLI**.
3. Go to the SQL script development page and set the data connection, database, and resource queue on the script property bar.
4. Enter the following SQL statements in the script editor:

```
INSERT OVERWRITE TABLE trade_report
SELECT
  sum(trade_count),
  '${yesterday}'
FROM
  trade_log
where
  date_format(trade_time, 'yyyy-MM-dd') = '${yesterday}'
```

5. Click  and set the script name to **generate\_trade\_report**.

### Step 2 Create and develop a job.

1. In the navigation tree of the DataArts Factory console, choose **Data Development > Develop Job**.
2. Access the area on the right and click **Create Job** to create an empty job named **job**.
3. Go to the job development page, drag the DLI SQL node to the canvas, click the icon, and configure node properties.

Description of key properties:



- SQL Script: SQL script **generate\_trade\_report** that is developed in [Step 1](#).
- Database Name: Database configured in SQL script **generate\_trade\_report**.
- Queue Name: Resource queue configured in SQL script **generate\_trade\_report**.
- Script Parameter: Parameter **yesterday** configured in SQL script **generate\_trade\_report**. Enter the following EL expression as the parameter values:

```
#{Job.getYesterday("yyyy-MM-dd")}
```

Expression Description: The job object uses the `getYesterday` method to obtain the time of the day before the job plan execution time. The time format is `yyyy-MM-dd`.

If the job plan time is 2018/9/26 01:00:00, the calculation result of this expression is 2018-09-25. The calculation result will replace the value of parameter `#{yesterday}` in the SQL script. The SQL statements after the replacement are as follows:

```
INSERT OVERWRITE TABLE trade_report
SELECT
  sum(trade_count),
  '2018-09-25'
FROM
  trade_log
where
  date_format(trade_time, 'yyyy-MM-dd') = '2018-09-25'
```

4. Click  to test the running job.
5. After the job test is complete, click  to save the job configuration.

----End

## 3.4.11 Usage Guidance

### 3.4.11.1 Job Dependency

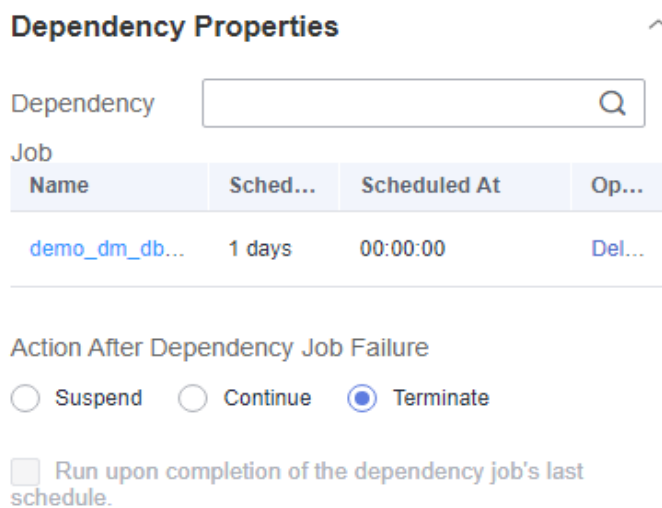
You can set a job that meets the scheduling period conditions as the dependency jobs for a job that is scheduled periodically. For details about how to set a dependency job, see **DataArts Factory > Job Development > Setting Up Scheduling for a Job** in *DataArts Studio User Guide*.

For example, you can set a dependency job (job B) for job A which is scheduled periodically. In this case, job A will be executed only when all the instances of job B are executed successfully within a specified period.

 NOTE

- The specified period is calculated as follows (see [How a Job Runs After a Dependency Job Is Set for It](#) for details):
  - Same-cycle dependency: If the scheduling periods of the two jobs are accurate to the same level (for example, minute, hour, or day), the specified period is **(Execution time of job A – Recurrence of job A, Execution time of job A)**.
  - Cross-cycle dependency: If the scheduling periods of the two jobs are accurate to different levels, the specified period is **[Natural start time of the previous recurrence of job A, Natural start time of the current recurrence of job A)**.
- Parameter **Policy for Current job If Dependency job Fails** determines whether job A will check the status of job B's instances.
  - If this parameter is set to **Suspend** or **Terminate**, job A will be suspended or terminated if instances of job B fail during a specified time period.
  - If this parameter is set to **Continue**, job A will be executed only if all the instances of job B are executed (regardless of whether the execution is successful or not).

Figure 3-230 Job dependency attributes



**Dependency Properties** ^

Dependency

Job

Name	Sched...	Scheduled At	Op...
demo_dm_db...	1 days	00:00:00	Del...

Action After Dependency Job Failure

Suspend  Continue  Terminate

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

This section describes [how to set the conditions of a dependency job](#) and [how a job runs after a dependency job is set for it](#).

## Setting Conditions of a Dependency Job

The recurrence of a periodically scheduled job can be minute, hour, day, week, or month. If job A and job B are both periodically scheduled jobs, and you want to set job B as the dependency job of job A, their recurrences must meet the following requirements:

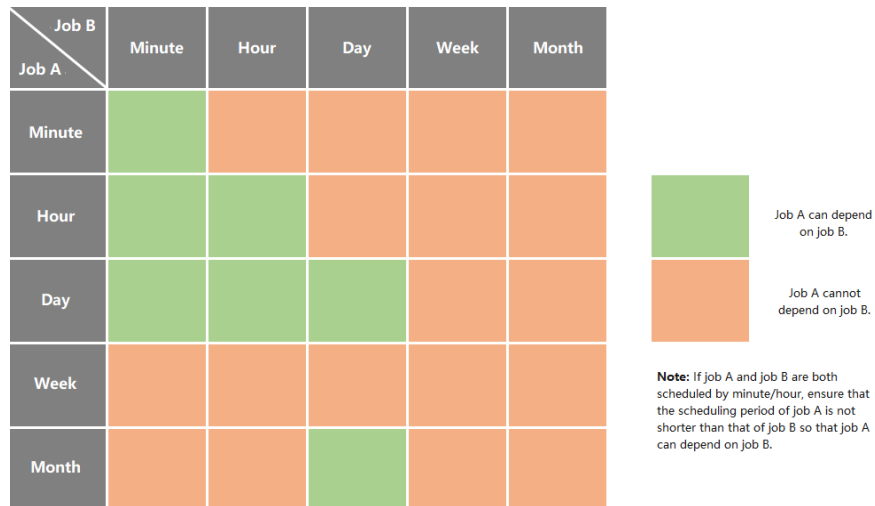
- The recurrence of job A cannot be shorter than that of job B. For example, if both job A and job B are scheduled by minute or hour and the interval of job A is shorter than that of job B, then job B cannot be set as the dependency job of job A. If job A is scheduled by minute and job B is scheduled by hour, job B cannot be set as the dependency job of job A.
- The recurrence of neither job A nor job B can be week. For example, if the recurrence of job A or job B is week, job B cannot be set as the dependency job of job A.



- A job whose recurrence is month can depend only on a job whose recurrence is day. For example, if the recurrence of job A is month, job B can be set as the dependency job of job A only if job B's recurrence is day.

**Figure 3-231** shows the requirements of the recurrences of the jobs that can function as the dependency jobs of other jobs

**Figure 3-231** Job dependency



## How a Job Runs After a Dependency Job Is Set for It

It varies depending on whether a job and its dependency job has the same recurrence. In this example, assume that the **Policy for Current job If Dependency job Fails** parameter is set to **Continue**, and job A does not check the running statuses of job B's instances. If this parameter is set to **Suspend** or **Terminate**, job A will also check whether there are failed instances in job B.

- **Same-cycle dependency:** Job A and its dependency job B have the same recurrence, for example, minute, hour, or day.

After job B is set as the dependency job of job A, job A checks whether instances of job B are running within a specified time range (**Execution time of job A – Recurrence of job A, Execution time of job A**). Job A will be executed only if all the instances of job B are executed.

Example 1: Job A depends on job B and they are both scheduled by minute. Job A starts at 10:00 and the interval is 20 minutes. Job B starts at 10:00 and the interval is 10 minutes. The following table lists how the two jobs run.

**Table 3-239** Example 1: dependency between jobs with the same recurrence

Time Point	Job B (Starting at 10:00 and Scheduled Every 10 Minutes)	Job A (Starting at 10:00 and Scheduled Every 20 Minutes)
10:00	Executed	Executed after job B's instances are executed in the (09:40, 10:00] time period
10:10	Executed	-

Time Point	Job B (Starting at 10:00 and Scheduled Every 10 Minutes)	Job A (Starting at 10:00 and Scheduled Every 20 Minutes)
10:20	Executed	Executed after job B's instances are executed in the <b>(10:00, 10:20]</b> time period
10:30	Executed	-
...	...	...

Example 2: Job A depends on job B and they are both scheduled by day. Job A starts at 09:00 on August 1, and job B starts at 10:00 on August 1. The following table lists how the two jobs run.

**Table 3-240** Example 2: dependency between jobs with the same recurrence

Time Point	Job B (Starting at 10:00 on August 1 and Scheduled by Day)	Job A (Starting at 09:00 on August 1 and Scheduled by Day)
09:00 on August 1	-	Not executed if no instance of job B is running in the <b>(09:00 on July 31, 09:00 on August 1]</b> time period
10:00 on August 1	Executed	-
09:00 on August 2	-	Executed after job B's instances are executed in the <b>(09:00 on August 1, 09:00 on August 2]</b> time period
10:00 on August 2	Executed	-
...	...	...

- **Cross-cycle dependency:** Job A and its dependency job B have different recurrences.

After job B is set as the dependent job of job A, job A checks whether any instance of job B is running in the time range **(Natural start time of the previous recurrence of job A, Natural start time of the current recurrence of job A)**. Job A will be executed only after all the instances of job B are executed.

 **NOTE**

The natural start time of a recurrence is defined as follows:

- If the recurrence is hour, the **natural start time of the previous recurrence** is 00:00 of the previous hour, and the **natural start time of the current recurrence** is 00:00 of the current hour.
- If the recurrence is day, the **natural start time of the previous recurrence** is 00:00:00 of the previous day, and the **natural start time of the current recurrence** is 00:00:00 of the current day.
- If the recurrence is month, the **natural start time of the previous recurrence** is 00:00:00 on 1st of the previous month, and the **natural start time of the current recurrence** is 00:00:00 on 1st of the current month.

Example 3: Job A depends on job B. Job A is scheduled by day, and job B is scheduled by hour. Job A is executed at 02:00 every day. Job B starts at 00:00 and is executed at an interval of 10 hours. The following table lists how the two jobs run.

**Table 3-241** Example 3: dependency between jobs with different recurrences

<b>Time Point</b>	<b>Job B (Starting at 00:00 at an Interval of 10 hours and Scheduled by Hour)</b>	<b>Job A (Scheduled at 02:00 Every Day)</b>
00:00 on the first day	Executed	-
02:00 on the first day	-	Not executed if no instance of job B is running in the <b>[00:00:00 on day 0, 00:00:00 on day 1)</b> time period
10:00 on the first day	Executed	-
20:00 on the first day	Executed	-
00:00 on the second day	Executed	-
02:00 on the second day	-	Executed if instances of job B are executed in the <b>[00:00:00 on day 1, 00:00:00 on day 2)</b> time period

Time Point	Job B (Starting at 00:00 at an Interval of 10 hours and Scheduled by Hour)	Job A (Scheduled at 02:00 Every Day)
10:00 on the second day	Executed	-
20:00 on the second day	Executed	-
...	...	...

Example 4: Job A depends on job B. Job A is scheduled by month, and job B is scheduled by day. Job A is executed at 02:00 on the first and second days of each month. Job B is executed at 00:00 on August 1. The following table lists how the two jobs run.

**Table 3-242** Example 4: dependency between jobs with different recurrences

Time Point	Job B (Scheduled by Day and Executed at 00:00 on August 1)	Job A (Scheduled by Month and Executed at 02:00 on the First and Second Days of Each Month)
00:00 on August 1	Executed	-
02:00 on August 1	-	Not executed if no instance of job B is running in the <b>[00:00:00 on July 1, 00:00:00 on August 1)</b> time period
00:00 on August 2	Executed	-
02:00 on August 2	-	Not executed if no instance of job B is running in the <b>[00:00:00 on July 1, 00:00:00 on August 1)</b> time period
...	-	...

Time Point	Job B (Scheduled by Day and Executed at 00:00 on August 1)	Job A (Scheduled by Month and Executed at 02:00 on the First and Second Days of Each Month)
00:00 on September 1	Executed	-
02:00 on September 1	-	Executed if instances of job B are executed in the [00:00:00 on August 1, 00:00:00 on September 1) time period
00:00 on September 2	Executed	-
02:00 on September 2	-	Executed if instances of job B are executed in the [00:00:00 on August 1, 00:00:00 on September 1) time period
...	...	...

### 3.4.11.2 IF Statements

When developing and orchestrating jobs in DataArts Factory, you can use IF statements to determine the branch to execute.

This section describes how to use IF statements in the following scenarios:

- [Determining the IF Statement Branch to Be Executed Based on the Execution Status of the Previous Node](#)
- [Determining the IF Statement Branch to Be Executed Based on the Execution Result of the Previous Node](#)
- [Configuring the Policy for Executing a Node with Multiple IF Statements](#)

IF statements use EL expressions. You can select EL expressions and follow the instruction in this section to develop jobs.

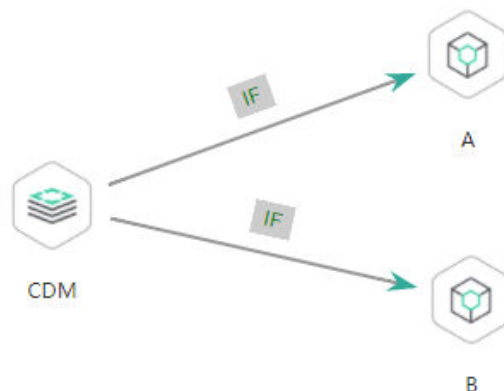
For details about how to use EL expressions, see [EL Expressions](#).

#### Determining the IF Statement Branch to Be Executed Based on the Execution Status of the Previous Node


Scenario

Generally, you can determine the IF statement branch to be executed based on whether the previous CDM node is successfully executed. For details on how to set IF statements, see [Figure 3-232](#).

**Figure 3-232** Example job



### Configuration Method

- Step 1** Log in to the DataArts Studio console, locate the target DataArts Studio instance, and click **Access** on the instance card.
- Step 2** Click the **Workspaces** tab. In the workspace list, locate the target workspace and click **DataArts Factory**. The DataArts Factory console is displayed.
- Step 3** On the **Develop Job** page, create a job, drag a CDM node and two Dummy nodes and drop them on the canvas in the right pane. Click and hold  to connect the CDM node to the Dummy nodes, as shown in [Figure 3-232](#). Set the **Failure Policy** for the CDM node to **Go to the next node**.
- Step 4** Right-click the connection line and select **Set Condition**. In the **Edit EL Expression** dialog box, enter the IF statement in the text box.

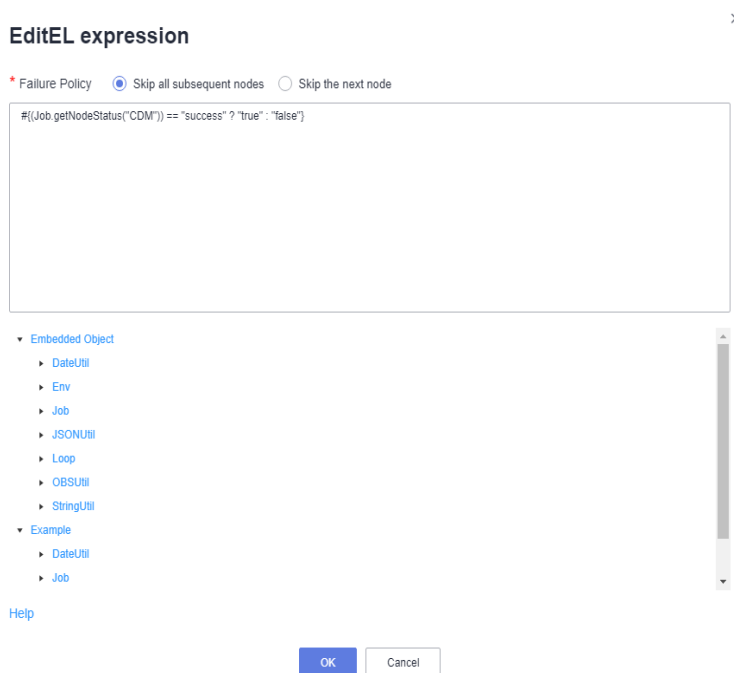
Each statement branch requires an IF statement. The IF statement is a ternary expression based on the EL expression syntax. If the result of the ternary expression is **true**, subsequent nodes will be connected. Otherwise, subsequent nodes will be skipped.

In this demo, the `#{Job.getNodeStatus("node_name")}` EL expression is used to obtain the execution status of a specified node. If the execution is successful, **success** is returned; otherwise, **fail** is returned. In this example, the IF statement expressions are as follows:

- The IF statement expression for branch A is `#{(Job.getNodeStatus("CDM")) == "success" ? "true" : "false"}`
- The IF statement expression for branch B is `#{(Job.getNodeStatus("CDM")) == "fail" ? "true" : "false"}`

After entering the IF statement expression, you can select either **Skip all subsequent nodes** or **Skip the next node** for **Failure Policy**. After the configuration is complete, click **OK** to save the job.

**Figure 3-233** Configuring a failure policy



**Step 5** Click **Test** to test the job and view the execution result on the **Monitor Instance** page.

**Step 6** After the job is executed, view the job instance running result on the **Monitor Instance** page. The execution result meets the expectation. If the execution result is **fail**, branch A is skipped and branch B is executed.

**Figure 3-234** Job execution result

Job Name	Status	Running Type	Planned Start Time	Actual Start Time	End Time	Running Duration	Created By	Versions	Operation
job_2051	Run successfully	Manual Sched.	2022/Jan/19 14:23:52	2022/Jan/19 14:23:58	2022/Jan/19 14:23:59	0:0	opc_test	0	Stop, Renew, View Waiting Job Instance

----End

## Determining the IF Statement Branch to Be Executed Based on the Execution Result of the Previous Node

### Scenario Description

Scenario: Use the execution result of the select statement on the HIVE SQL node as a parameter to determine the IF statement branch to be executed.

The execution result of the select statement on the HIVE SQL node is a two-dimensional array. To obtain the values in the array, use the EL expression **#{Loop.dataArray[] []}**. Currently, only the For Each node supports this expression. Therefore, you need to connect the HIVE SQL node to a For Each node. **Figure 3-235** shows the job orchestration.

Figure 3-235 Example job

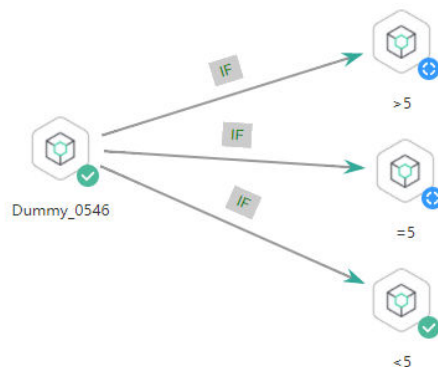


Key configurations of the For Each node are as follows:

- **Dataset:** Enter the execution result of the select statement on the HIVE SQL node. Use the `#{Job.getNodeOutput('HIVE')}` expression, where **HIVE** is the name of the previous node.
- **Job Running Parameter:** Enter the parameter defined in the sub-job. Transfer the output of the previous node of the main job to the sub-job for use. The variable name is **result**, and its value is a column in the dataset. The EL expression `#{Loop.dataArray[0][0]}` is used.

The sub-job selected on the For Each node determines the IF statement branch to be executed based on the job running parameter transferred from the For Each node. [Figure 3-236](#) shows the job orchestration.

Figure 3-236 Example sub-job



The IF statement is the key configuration of the subjob. This example uses the expression `#{result}` to obtain the value of the job parameter.

**NOTE**


Do not use the `#{Job.getParam("job_param_name")}` EL expression because this expression can only obtain the values of the parameters configured in the current job, but cannot obtain the parameter values transferred from the parent job or the global variables configured in the workspace. The expression only works for the current job.

To obtain the parameter values passed from the parent job and the global variables configured for the workspace, you are advised to use the `#{job_param_name}` expression.

**Configuration Method**

Developing a Subjob



- Step 1** Log in to the DataArts Studio console, locate the target DataArts Studio instance, and click **Access** on the instance card.
- Step 2** Click the **Workspaces** tab. In the workspace list, locate the target workspace and click **DataArts Factory**. The DataArts Factory console is displayed.
- Step 3** On the **Develop Job** page, create a data development subjob named **foreach**.  
Drag four Dummy nodes and drop them on the canvas, click and hold  to connect them, as shown in [Figure 3-236](#).
- Step 4** Right-click the connection line and select **Set Condition**. In the **Edit EL Expression** dialog box, enter the IF statement in the text box.  
Each statement branch requires an IF statement. The IF statement is a ternary expression based on the EL expression syntax. If the result of the ternary expression is **true**, subsequent nodes will be connected. Otherwise, subsequent nodes will be skipped.
- For the >5 branch, the IF statement expression is `#{${result} > 5 ? "true" : "false"}`.
  - For the =5 branch, the IF statement expression is `#{${result} == 5 ? "true" : "false"}`.
  - For the <5 branch, the IF statement expression is `#{${result} < 5 ? "true" : "false"}`.
- After entering the IF statement expression, you can select either **Skip all subsequent nodes** or **Skip the next node** for **Failure Policy**.
- Step 5** Configure job parameters. Set the parameter name to **result**. This parameter is only used by the For Each node in the main job **testif** to identify subjob parameters. You do not need to set the parameter value.

**Figure 3-237** Configuring job parameters

### Parameter Setup


#### Variable Parameter



- Step 6** Save the job.

----End

## Developing a Job

- Step 1** On the **Develop Job** page, create a data development job named **testif**. Drag a HIVE SQL node and a For Each node and drop them on the canvas. Click and hold  to connect the nodes, as shown in [Figure 3-235](#).

**Step 2** Configure properties for the HIVE SQL node. Reference the following SQL script (there is no special requirement for other properties):

```
SELECT count(*) FROM student // Count from the student table. The script execution result is a two-dimensional array.
```

**Figure 3-238** HIVE SQL script execution result

The screenshot shows the SQL editor interface with a toolbar at the top containing icons for Save, Submit, Unlock, Lock, Execute, Format, SQL Reference, and Configure Editor. The script content is as follows:

```
1 -- DLI sql
2 -- *****
3 -- author:
4 -- create time: 2022/03/22 16:21:19 GMT+08:00
5 -- *****
6 SELECT count(*) FROM student
```

The SQL statement on line 6 is highlighted with a red box. Below the editor, the 'Execution History' tab is active, showing the 'Result' of the query:

Row No.	count(1)
1	1

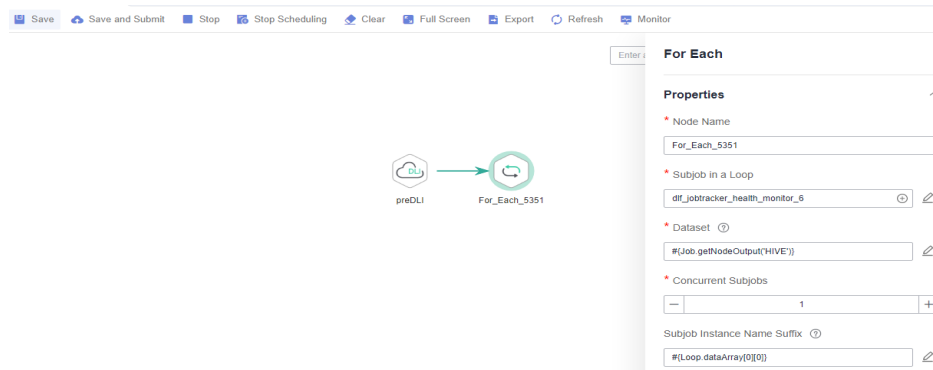
The result table is also highlighted with a red box.

**Step 3** Configure properties for the For Each node.

- **Subjob in a Loop:** Select **foreach**, the subjob that has been developed.
- **Dataset:** Enter the execution result of the select statement on the HIVE SQL node. Use the `#{Job.getNodeOutput('HIVE')}` expression, where **HIVE** is the name of the previous node.

- Job Running Parameter:** Enter the parameter defined in the sub-job. Transfer the output of the previous node of the main job to the sub-job for use. The variable name is **result** (parameter name of the subjob), and its value is a column in the dataset. The EL expression **`#{Loop.dataArray[0][0]}`** is used.

Figure 3-239 Properties of the For Each node



**Step 4** Save the job.

----End

Testing the Main Job

- Step 1** Click **Test** above the main job canvas to test the job. After the main job is executed, the subjob is automatically invoked through the For Each node and executed.
- Step 2** In the navigation pane on the left, choose **Monitor Instance** to view the job execution result.
- Step 3** After the job is executed, view the execution result of the subjob **foreach** on the **Monitor Instance** page. The execution result meets the expectation. Currently, the execution result of the Hive SQL statement is **1**. Therefore, the **>5** and **=5** branches are skipped, and the **<5** branch is successfully executed.

Figure 3-240 Execution result of the subjob

Monitor Instance ⓘ

Stop | **Run** | Continue | Succeeded

Job Name: [ ] | Search: [ ] | Jan 19, 2022 00:00:00 - Jan 19, 2022 23:59:59 | [ ] [ ]

Job Name	Status	Running T...	Planned Start Time	Actual Start Time	End Time	Running Duration...	Created By	Versions	Operation
foreach_1	Run successfully	Manual Sched...	2022/Jan/19 14:23:52	2022/Jan/19 14:23:58	2022/Jan/19 14:23:59	0.0	dgc_test	0	Stop   <b>Run</b>   View Waiting Job Instance
Name	Type	Running Type	Running Durati...	Actual Start Time	Retry Count	Error Message	Operation		
Dummy_4141	Dummy	Run successfully	0.00	2022/Jan/19 14:23:58 GMT+08:00	0	--	View Log   Manual Retry   Succeeded   More		
Dummy_5381	Dummy	Run successfully	0.00	2022/Jan/19 14:23:59 GMT+08:00	0	--	View Log   Manual Retry   Succeeded   More		

----End

## Configuring the Policy for Executing a Node with Multiple IF Statements

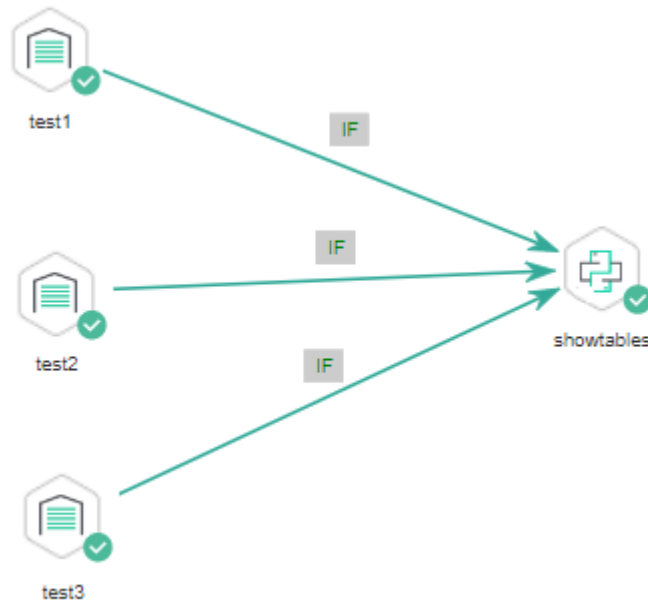
If the execution of a node depends on multiple IF statements, the policy for executing the node can be **AND** or **OR**.

If you choose the **OR** policy, the node will be executed if any one of the IF statements is met.

If you choose the **AND** policy, the node will be executed only if all of the IF statements are met.

If you choose neither, the **OR** policy will be used.

**Figure 3-241** A job with multiple IF statements




### Configuration Method

Configure the execution policy.

- Step 1** Log in to the DataArts Studio console, locate the target DataArts Studio instance, and click **Access** on the instance card.
- Step 2** Click the **Workspaces** tab. In the workspace list, locate the target workspace and click **DataArts Factory**. The DataArts Factory console is displayed.
- Step 3** On the DataArts Factory console, choose **Configuration > Configure > Default Configuration**.
- Step 4** Select **AND** or **OR** for **Multi-IF Policy**.
- Step 5** Click **Save**.

----End

Develop a job.

- Step 1** On the **Develop Job** page, create a data development job.
- Step 2** Drag three DWS SQL operators as parent nodes and one Python operator as a child node to the canvas. Click and hold  to connect the nodes to orchestrate the job shown in [Figure 3-241](#).
- Step 3** Right-click the connection line and select **Set Condition**. In the **Edit EL Expression** dialog box, enter the IF statement in the text box.

Each statement branch requires an IF statement. The IF statement is a ternary expression based on the EL expression syntax.

- The IF statement expression for the test1 node is  
`#{(Job.getNodeStatus("test1")) == "success" ? "true" : "false"},`
- The IF statement expression for the test2 node is  
`#{(Job.getNodeStatus("test2")) == "success" ? "true" : "false"},`
- The IF statement expression for the test3 node is  
`#{(Job.getNodeStatus("test3")) == "success" ? "true" : "false"},`

The expression of each node is determined using the IF statement based on the execution status of the previous node.

After entering the IF statement expression, you can select either **Skip all subsequent nodes** or **Skip the next node** for **Failure Policy**.

----End

Test the job.

**Step 1** Click **Save** above the canvas to save the job.

**Step 2** Click **Test** above the canvas to test the job.

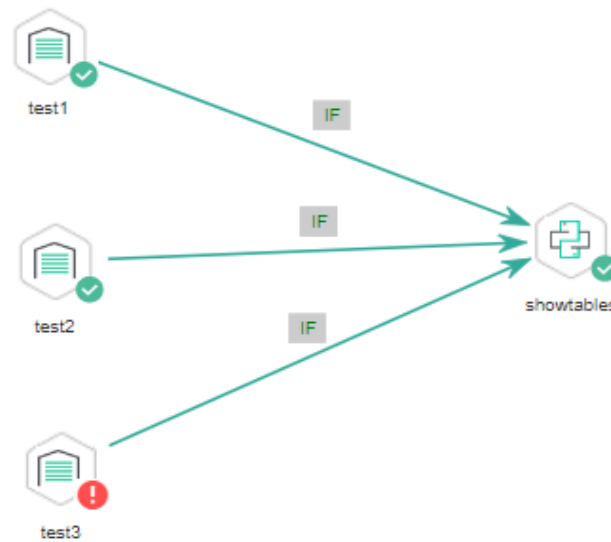
If **test1** is executed successfully, the corresponding IF statement is true.

If **test2** is executed successfully, the corresponding IF statement is true.

If **test3** fails to be executed, the corresponding IF statement is false.

If **Multi-IF Policy** is set to **OR**, the **showtables** node is executed and the job execution is complete.

Figure 3-242 How the job runs if Multi-IF Policy is OR

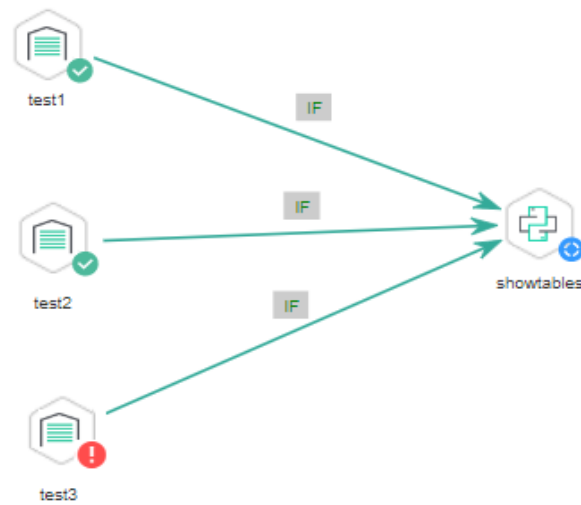


### Logs

[INFO][Jul 04, 2022 17:28:23 GMT+08:00] : The job starts to run.  
[INFO][Jul 04, 2022 17:30:31 GMT+08:00] : Node test1 started to run.  
[INFO][Jul 04, 2022 17:30:31 GMT+08:00] : Node test2 started to run.  
[INFO][Jul 04, 2022 17:30:31 GMT+08:00] : Node test3 started to run.  
[ERROR][Jul 04, 2022 17:30:51 GMT+08:00] : Node test3 failed to run.  
[INFO][Jul 04, 2022 17:30:51 GMT+08:00] : Node test1 finished to run.  
[INFO][Jul 04, 2022 17:30:51 GMT+08:00] : Node test2 finished to run.  
[INFO][Jul 04, 2022 17:30:51 GMT+08:00] : Node showtables started to run.  
[INFO][Jul 04, 2022 17:30:51 GMT+08:00] : Node showtables finished to run.  
[INFO][Jul 04, 2022 17:30:51 GMT+08:00] : Job running is completed.]

If **Multi-IF Policy** is set to **AND**, the **showtables** node is skipped and the job execution is complete.

**Figure 3-243** How the job runs if Multi-IF Policy is AND



### Logs

```
[INFO][Jul 05, 2022 09:05:33 GMT+08:00] : The job starts to run.
[INFO][Jul 05, 2022 09:07:42 GMT+08:00] : Node test1 started to run.
[INFO][Jul 05, 2022 09:07:42 GMT+08:00] : Node test2 started to run.
[INFO][Jul 05, 2022 09:07:42 GMT+08:00] : Node test3 started to run.
[ERROR][Jul 05, 2022 09:08:03 GMT+08:00] : Node test3 failed to run.
[INFO][Jul 05, 2022 09:08:03 GMT+08:00] : Node test1 finished to run.
[INFO][Jul 05, 2022 09:08:03 GMT+08:00] : Node test2 finished to run.
[INFO][Jul 05, 2022 09:08:03 GMT+08:00] : Node showtables finished to run.
```

----End

### 3.4.11.3 Obtaining the Return Value of a Rest Client Node

The Rest Client node can execute RESTful requests.

This tutorial describes how to obtain the return value of the Rest Client node, covering the following two application scenarios:

- [Obtaining the Return Value Through Parameter "The response message body parses the transfer parameter"](#)
- [Obtaining the Return Value Using an EL Expression](#)

#### Obtaining the Return Value Through Parameter "The response message body parses the transfer parameter"

As shown in [Figure 3-244](#), the first Rest Client node invokes the API of MRS to query the cluster list. [Figure 3-245](#) shows the JSON message body returned by the API.

- Scenario: The ID of the first cluster in the cluster list needs to be obtained and transferred to other nodes as a parameter.
- Key configurations: Set **The response message body parses the transfer parameter** of the first Rest Client to **clusterId=clusters[0].clusterId**. Other Rest Client nodes can reference the ID of the first cluster in `${clusterId}` mode.

Figure 3-244 Rest Client job example 1

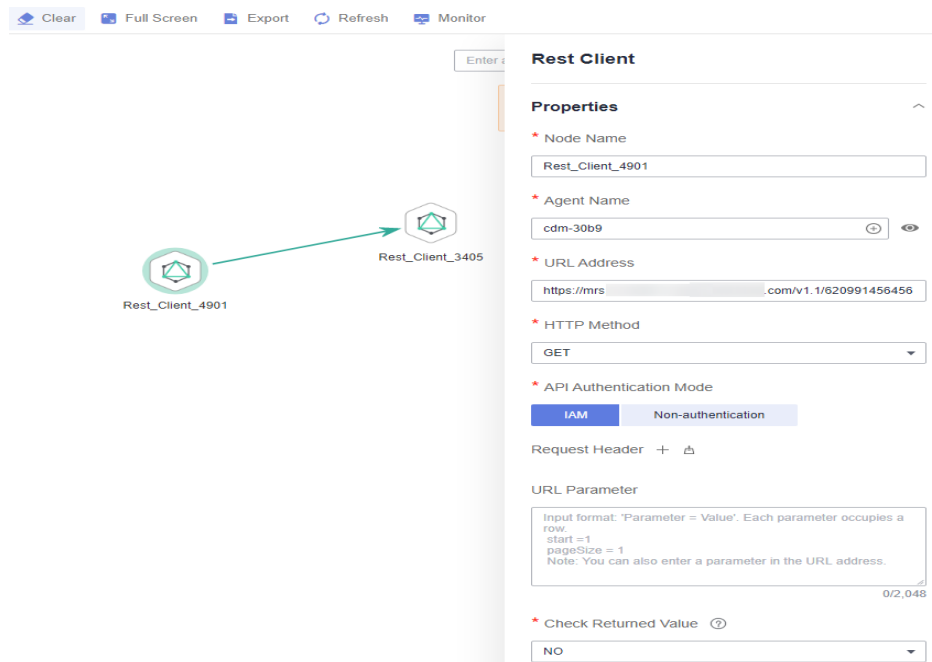


Figure 3-245 JSON message body

```

{
  "clusterTotal": 31,
  "clusters": [
    {
      "clusterId": "6e1b5c2-6526-4ef8-9c8f-4105b63fa893",
      "clusterName": "azs_hbase22",
      "totalNodeNum": "2",
      "clusterState": "running",
      "stageDesc": null,
      "createAt": "1620378935",
      "updateAt": "1620611307",
      "chargingStartTime": "1620380067",
      "billingType": "Metered",
      "dataCenter": "cn-north-7",
      "vpc": "vpc-dlf",
      "vpcId": "f35aee01-c4a3-47c1-8d92-9df430537de4",
      "duration": "0",
      "fee": "0.0",
      "hadoopVersion": "",
      "componentList": [
        {
          "id": "218051",
          "componentId": "HRS 2.1.0_001",
          "componentName": "Hadoop",
          "componentVersion": "3.1.1",
          "external_datasources": null,
          "componentDesc": "A distributed data storage and processing framework for large da
ta sets, including core components such as HDFS, YARN, and MapReduce.",
          "componentDescEn": null,
          "multi_service_name": null
        }
      ]
    }
  ]
}

```

## Obtaining the Return Value Using an EL Expression

The Rest Client node can be used together with EL expressions. You can select different EL expressions based on scenarios. This section describes how to develop

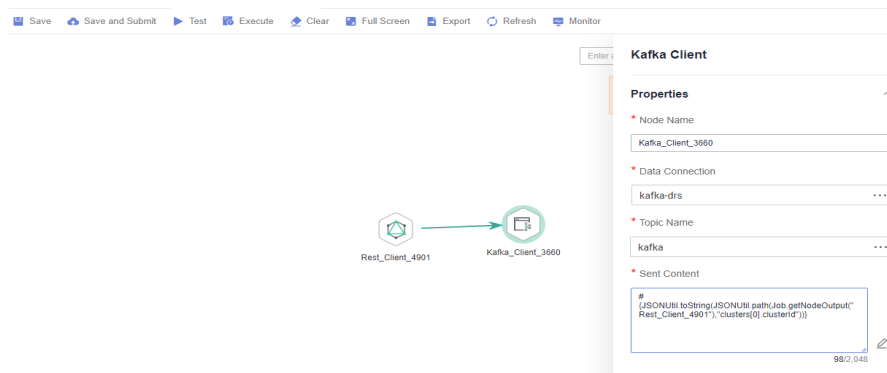


your own jobs based on your service requirements. For details about how to use EL expressions, see [EL Expressions](#).

As shown in [Figure 3-246](#), the Rest Client invokes the API of MRS to query the cluster list and then invokes the Kafka Client to send a message.

- Scenario: The Kafka Client sends a character string message. The message content is the ID of the first cluster in the cluster list.
- Key configurations: When you configure the Kafka Client, use the following EL expression to obtain a specific field in the message body returned by the REST API:  
`#{JSONUtil.toString(JSONUtil.path(Job.getNodeOutput("Rest_Client_4901"),"clusters[0].clusterId"))}`

**Figure 3-246** Rest Client job example 2



### 3.4.11.4 Using For Each Nodes

#### Scenario

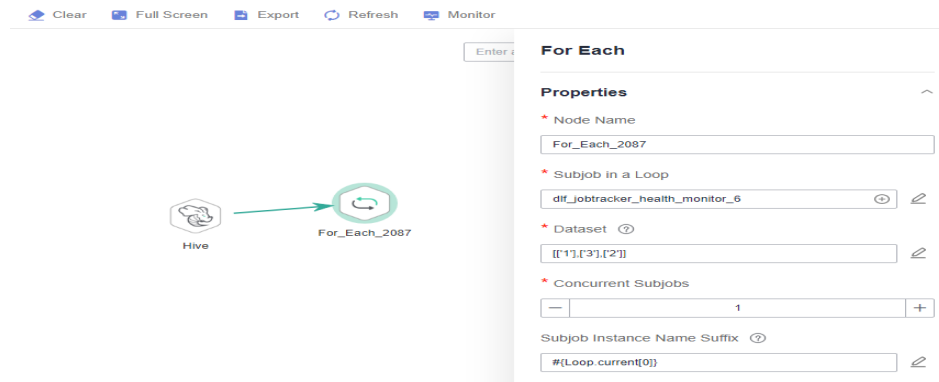
During job development, if some jobs have different parameters but the same processing logic, you can use For Each nodes to avoid repeated job development.

You can use a For Each node to execute a subjob in a loop and use a dataset to replace the parameters in the subjob. The key parameters are as follows:

- **Subjob in a Loop:** Select the subjob to be executed in a loop.
- **Dataset:** Enter a set of parameter values of the subjobs. The value can be a specified dataset such as `[[ '1' ], [ '3' ], [ '2' ]]` or an EL expression such as `#{Job.getNodeOutput('preNodeName')}`, which is the output value of the previous node.
- **Job Running Parameter:** The parameter name is the variable defined in the subjob. The parameter value is usually set to a group of data in the dataset. Each time the job is run, the parameter value is transferred to the subjob for use. For example, parameter value `#{Loop.current[0]}` indicates that the first value of each group of data in the dataset is traversed and transferred to the subjob.

[Figure 3-247](#) shows an example For Each node. As shown in the figure, the parameter name of the **foreach** subjob is **result**, and the parameter value is the traversal of the one-dimensional array dataset `[[ '1' ], [ '3' ], [ '2' ]]` (that is, the value is **1**, **3**, and **2** in the first, second, and third loop, respectively).

**Figure 3-247** For Each node



## For Each Nodes and EL Expressions

To use For Each nodes properly, you must be familiar with EL expressions. For details about how to use EL expressions, see [EL Expressions](#).

For Each nodes use the following EL expressions most:

- `#{Loop.dataArray}`: dataset input by the For Each node. It is a two-dimensional array.
- `#{Loop.current}`: The For Loop node processes a dataset line by line. *Loop.current* indicates a line of data that is being processed. *Loop.current* is a one-dimensional array, and its format is `#{Loop.current[0]}`, `#{Loop.current[1]}`, or others. The value 0 indicates that the first value in the current line is traversed.
- `#{Loop.offset}`: current offset when the For Each node processes the dataset. The value starts from 0.
- `#{Job.getNodeOutput('preNodeName')}`: obtains the output of the previous node.

## Examples

### Scenario

To meet data normalization requirements, you need to periodically import data from multiple source DLI tables to the corresponding destination DLI tables, as listed in [Table 1](#).

**Table 3-243** Tables to be imported

Source Table	Destination Table
a_new	a
b_2	b
c_3	c
d_1	d
c_5	e

Source Table	Destination Table
b_1	f

If you use SQL nodes to execute import scripts, a large number of scripts and nodes need to be developed, resulting in repeated work. In this case, you can use the For Each operator to perform cyclic jobs to reduce the development workload.

### Configuration Method

**Step 1** Prepare the source and destination tables. To facilitate subsequent job execution and verification, you need to create a source DLI table and a destination DLI table and insert data into the tables.

1. Create a DLI table. You can create a DLI SQL script on the **DataArts Factory** page and run the following commands to create a DLI table. You can also run the following SQL commands in the SQL editor on the DLI console.

```
/* Create a data table. */  
CREATE TABLE a_new (name STRING, score INT) STORED AS PARQUET;  
CREATE TABLE b_2 (name STRING, score INT) STORED AS PARQUET;  
CREATE TABLE c_3 (name STRING, score INT) STORED AS PARQUET;  
CREATE TABLE d_1 (name STRING, score INT) STORED AS PARQUET;  
CREATE TABLE c_5 (name STRING, score INT) STORED AS PARQUET;  
CREATE TABLE b_1 (name STRING, score INT) STORED AS PARQUET;  
CREATE TABLE a (name STRING, score INT) STORED AS PARQUET;  
CREATE TABLE b (name STRING, score INT) STORED AS PARQUET;  
CREATE TABLE c (name STRING, score INT) STORED AS PARQUET;  
CREATE TABLE d (name STRING, score INT) STORED AS PARQUET;  
CREATE TABLE e (name STRING, score INT) STORED AS PARQUET;  
CREATE TABLE f (name STRING, score INT) STORED AS PARQUET;
```

2. Insert data into the source data table. You can create a DLI SQL script on the **DataArts Factory** page and run the following commands to create a DLI table. You can also run the following SQL commands in the SQL editor on the DLI console.

```
/* Insert data into the source data table. */  
INSERT INTO a_new VALUES ('ZHAO','90'),('QIAN','88'),('SUN','93');  
INSERT INTO b_2 VALUES ('LI','94'),('ZHOU','85');  
INSERT INTO c_3 VALUES ('WU','79');  
INSERT INTO d_1 VALUES ('ZHENG','87'),('WANG','97');  
INSERT INTO c_5 VALUES ('FENG','83');  
INSERT INTO b_1 VALUES ('CEHN','99');
```

**Step 2** Prepare dataset data. You can obtain a dataset in any of the following ways:

1. Import the data in **Table 1** into the DLI table and use the result read by the SQL script as the dataset.
2. You can save the data in **Table 1** to a CSV file in the OBS bucket. Then use a DLI SQL or DWS SQL statement to create an OBS foreign table, associate it with the CSV file, and use the query result of the OBS foreign table as the dataset.
3. You can save the data in **Table 1** to a CSV file in the HDFS. Then use a Hive SQL statement to create a Hive foreign table, associate it with the CSV file, and use the query result of the Hive foreign table as the dataset.

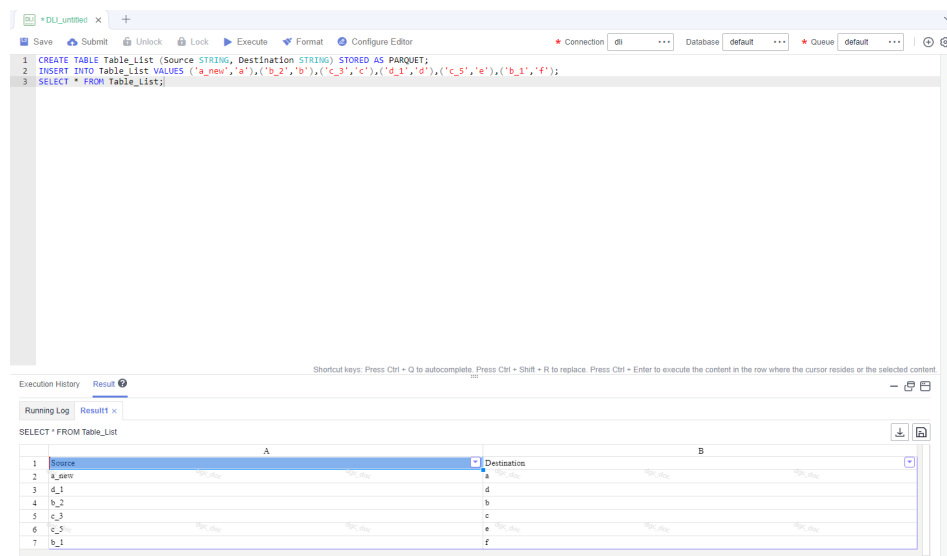
This section uses method 1 as an example to describe how to import data from **Table 1** to the DLI table (**Table\_List**). You can create a DLI SQL script on the **DataArts Factory** page and run the following commands to import data into the

table. You can also run the following SQL commands in the SQL editor on the DLI console.

```
/* Create the Table_List data table, insert data in Table 1 into the table, and check the generated data. */  
CREATE TABLE Table_List (Source STRING, Destination STRING) STORED AS PARQUET;  
INSERT INTO Table_List VALUES ('a_new','a'),('b_2','b'),('c_3','c'),('d_1','d'),('c_5','e'),('b_1','f');  
SELECT * FROM Table_List;
```

The generated data in the **Table\_List** table is as follows:

**Figure 3-248** Data in the Table\_List table



**Step 3** Create a subjob named **ForeachDemo** to be executed cyclically. In this operation, a task containing the DLI SQL node is defined to be executed cyclically.

1. Access the DataArts Studio **DataArts Factory** page, choose **Develop Job**. Create a job named **ForeachDemo**, select the DLI SQL node, and configure the job as shown in [Figure 3-249](#).

In the DLI SQL statement, set the variable to be replaced to **\${}**. The following SQL statement is used to import all data in the **\${Source}** table to the **\${Destination}** table. **\${fromTable}** and **\${toTable}** are the variables. The SQL statement is as follows:

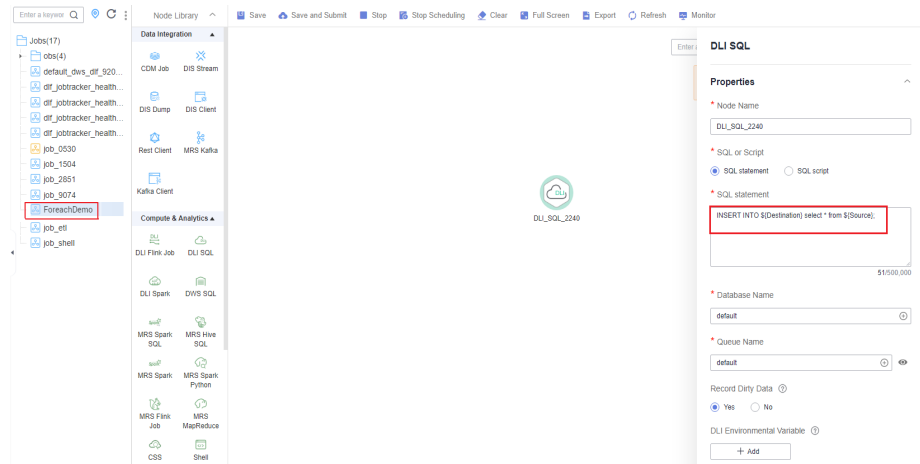
```
INSERT INTO ${Destination} select * from ${Source};
```

#### NOTE

Do not use the `#{Job.getParam("job_param_name")}` EL expression because this expression can only obtain the values of the parameters configured in the current job, but cannot obtain the parameter values transferred from the parent job or the global variables configured in the workspace. The expression only works for the current job.

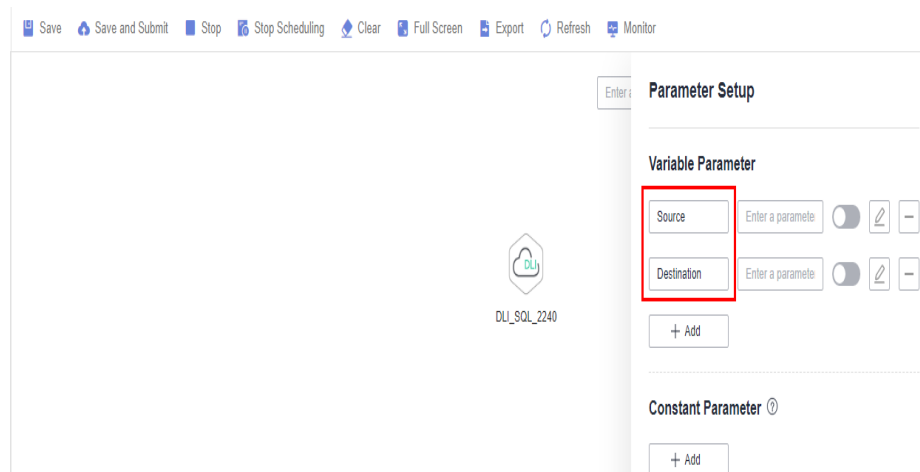
To obtain the parameter values passed from the parent job and the global variables configured for the workspace, you are advised to use the `#{job_param_name}` expression.

**Figure 3-249** Cyclically executing a subjob




2. After configuring the SQL statement, configure parameters for the subjob. You only need to set the parameter names, which are used by the For Each operator of the **ForeachDemo\_master** job to identify subjob parameters.

**Figure 3-250** Configuring subjob parameters

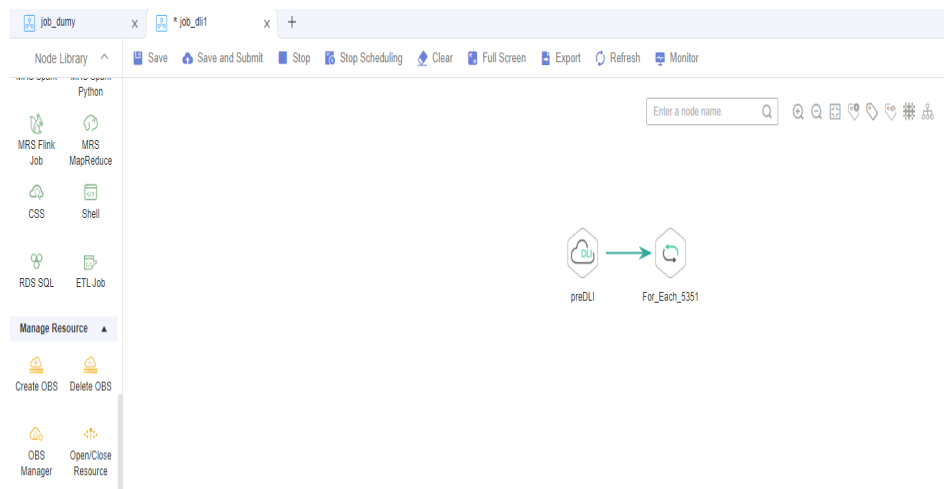


3. Save the job.

**Step 4** Create a master job named **ForeachDemo\_master** where the For Each operator is located.

1. Access the DataArts Studio **DataArts Studio** page and choose **Develop Job**. Create a data development master job named **ForeachDemo\_master**. Select the DLI SQL and For Each nodes and click and drag  to compile the job shown in **Figure 3-251**.

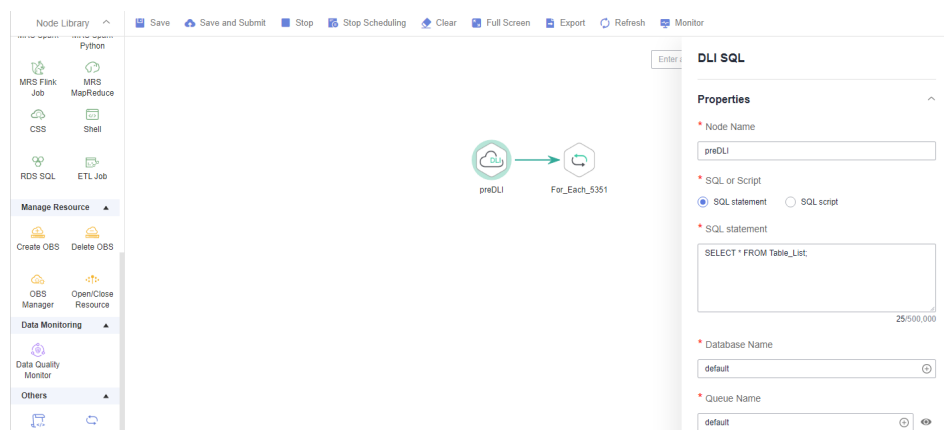
**Figure 3-251** Compiling a job



2. Configure the properties of the DLI SQL node. Select **SQL statement** and enter the following statement. The DLI SQL node reads data from the DLI table **Table\_List** and uses it as the dataset.  

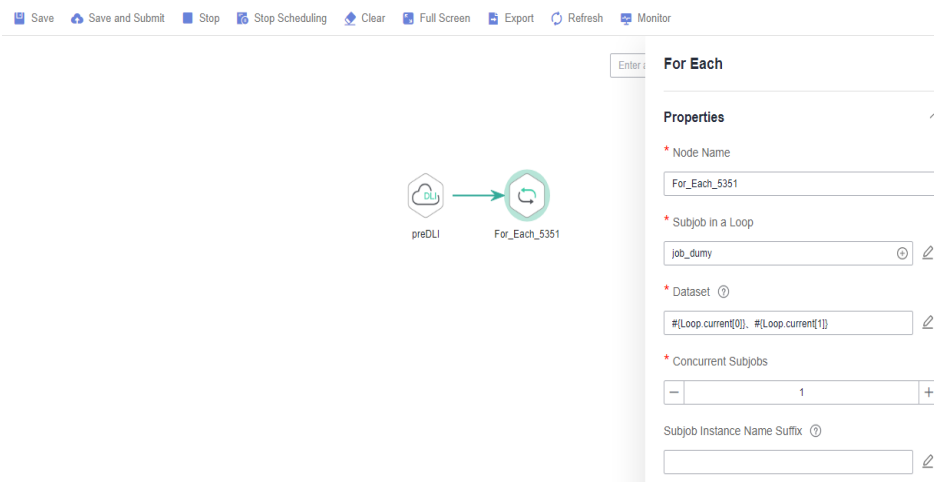
```
SELECT * FROM Table_List;
```

**Figure 3-252** DLI SQL node configuration



3. Configure properties for the For Each node.
  - **Subjob in a Loop:** Select **ForeachDemo**, which is the subjob that has been developed in [step 2](#).
  - **Dataset:** Enter the execution result of the select statement on the DLI SQL node. Use the `#{Job.getNodeOutput('preDLI')}` expression, where **preDLI** is the name of the previous node.
  - **Job Running Parameters:** used to transfer data in the dataset to the subjob **Source** corresponds to the first column in the **Table\_List** table of the dataset, and **Destination** corresponds to the second column. Therefore, enter EL expression `#{Loop.current[0]}` for **Source** and `#{Loop.current[1]}` for **Destination**.

Figure 3-253 Configuring the For Each node



4. Save the job.

**Step 5** Test the main job.

1. Click **Test** above the main job canvas to test the job. After the main job is executed, the subjob is automatically invoked through the For Each node and executed.
2. In the navigation pane on the left, choose **Monitor Instance** to view the job execution status. After the job is successfully executed, you can view the subjob instances generated on the For Each node. Because the dataset contains six rows of data, six subjob instances are generated.

Figure 3-254 Viewing job instances

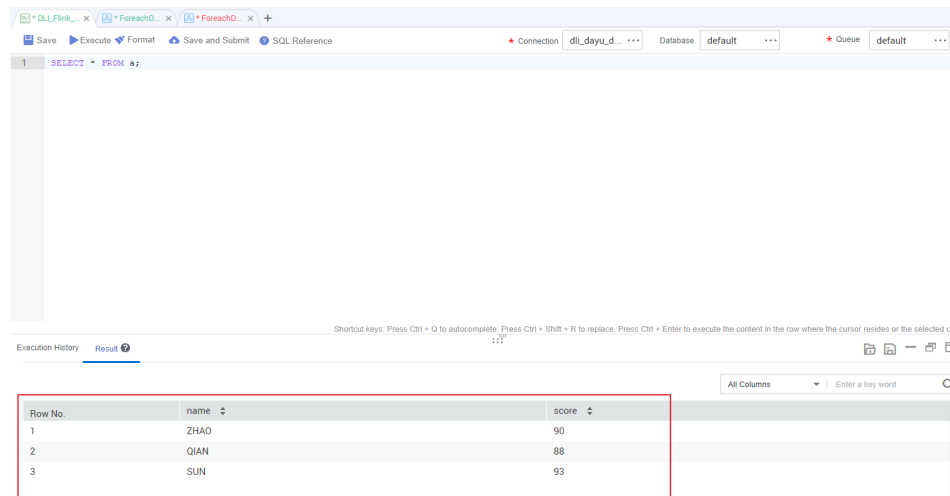
Job Name	Status	Running Time	Planned Start Time	Actual Start Time	End Time	Running Duration	Created By	Versions	Operation
#_jobtracker_health...	Run successfully	Normal Sched.	2022/Jan/18 17:00:00	2022/Jan/18 17:00:06	2022/Jan/18 17:00:06	0.0	qpc_hst	3	Stop   Run   View   Waiting Job Instance
#_jobtracker_health...	Run successfully	Normal Sched.	2022/Jan/18 17:00:00	2022/Jan/18 17:00:02	2022/Jan/18 17:00:03	0.0	qpc_hst	2	Stop   Run   View   Waiting Job Instance
#_jobtracker_health...	Run successfully	Normal Sched.	2022/Jan/18 17:00:00	2022/Jan/18 17:00:06	2022/Jan/18 17:00:06	0.0	qpc_hst	1	Stop   Run   View   Waiting Job Instance
#_jobtracker_health...	Failed	Normal Sched.	2022/Jan/18 17:00:00	2022/Jan/18 17:00:07	2022/Jan/18 17:00:38	0.5	qpc_hst	2	Stop   Run   View   Waiting Job Instance
#_jobtracker_health...	Run successfully	Normal Sched.	2022/Jan/18 16:55:00	2022/Jan/18 16:55:03	2022/Jan/18 16:55:03	0.0	qpc_hst	3	Stop   Run   View   Waiting Job Instance
#_jobtracker_health...	Run successfully	Normal Sched.	2022/Jan/18 16:55:00	2022/Jan/18 16:55:03	2022/Jan/18 16:55:04	0.0	qpc_hst	2	Stop   Run   View   Waiting Job Instance
#_jobtracker_health...	Run successfully	Normal Sched.	2022/Jan/18 16:55:00	2022/Jan/18 16:55:05	2022/Jan/18 16:55:06	0.0	qpc_hst	1	Stop   Run   View   Waiting Job Instance
#_jobtracker_health...	Failed	Normal Sched.	2022/Jan/18 16:55:00	2022/Jan/18 16:55:10	2022/Jan/18 16:55:41	0.5	qpc_hst	2	Stop   Run   View   Waiting Job Instance
ForEachDemo_master	Run successfully	Normal Sched.	2022/Jan/18 16:50:00	2022/Jan/18 16:50:09	2022/Jan/18 16:50:09	0.0	qpc_hst	3	Stop   Run   View   Waiting Job Instance
preDLI	DLI SQL	Run successfully	0.4	2022/Jan/18 16:50:09 GMT+08:00	0	0	--	--	View Log   Manual Retry   Succeeded   More
For_Each_5351	ForEachJob	Run successfully	5.7	2022/Jan/18 16:50:09 GMT+08:00	0	0	--	--	View Log   Manual Retry   Succeeded   More

3. Check whether the data has been inserted into the six DLI destination tables. You can create a DLI SQL script on the **DataArts Factory** page and run the following commands to import data into the table. You can also run the following SQL commands in the SQL editor on the DLI console.

```
/* Run the following command to query the data in a table (table a is used as an example): */
SELECT * FROM a;
```

Compare the obtained data with the data in **Insert data into the source data table**. The inserted data meets the expectation.

Figure 3-255 Destination table data



----End

## More Cases for Reference

For Each nodes can work with other nodes to implement more functions. You can refer to the following cases to learn more about how to use For Each nodes.

- [Determining the IF Statement Branch to Be Executed Based on the Execution Result of the Previous Node](#)

### 3.4.11.5 Developing a Python Script

This section describes how to develop and execute a Python script using DataArts Factory.

#### Preparing the Environment

- An ECS named **ecs-dgc** has been created.

#### NOTE

In this example, the ECS uses the **CentOS 8.0 64bit with ARM (40 GB)** public image and the Python environment. You can log in to the ECS and run the **python** command to check the Python environment.

```
CentOS Linux 7 (AltArch)
Kernel 4.14.0-115.el7a.0.1.aarch64 on an aarch64

ecs-dgc login: root
Password:

Welcome to [REDACTED] Service

[root@ecs-dgc ~]# python
Python 2.7.5 (default, Aug 7 2019, 00:57:09)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-39)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
```



- You have enabled the DataArts Migration incremental package and created a CDM cluster named **cdm-dlfpython**. The cluster provides an agent for the DataArts Factory module to communicate with the ECS.
- Ensure that the ECS can communicate with the CDM cluster, which depends on the following conditions:
  - If the CDM cluster and the ECS 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 **Adding Routes** in *Virtual Private Cloud (VPC) Usage Guide*. For details about how to configure security group rules, see **Security Group > Adding a Security Group Rule** in *Virtual Private Cloud (VPC) Usage Guide*.
  - If the CDM cluster and the ECS 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.
  - The ECS and the CDM cluster belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

## Constraints

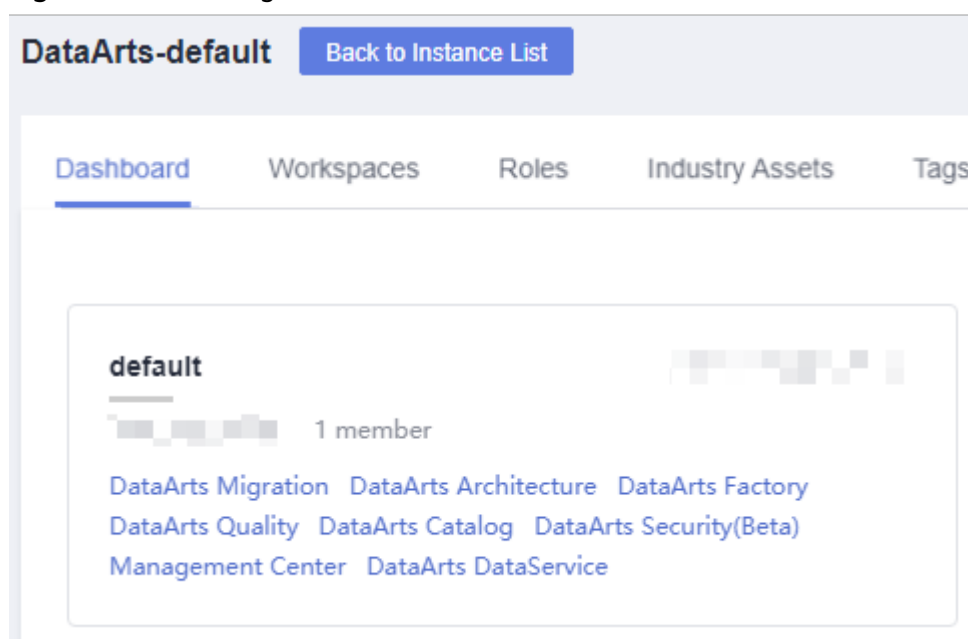
- Python scripts do not support script parameters or job parameters.

## Creating an ECS Data Connection

Before developing a Python script, you need to create a connection to the ECS.

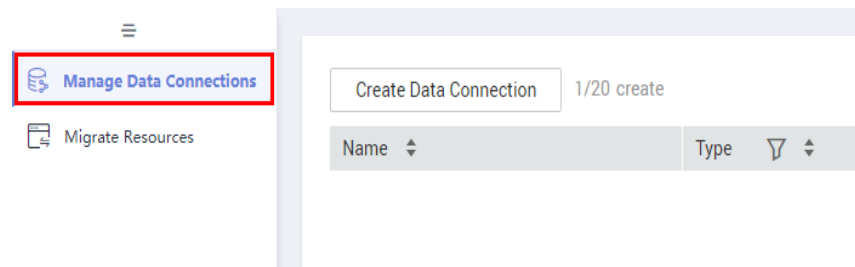
- Step 1** On the DataArts Studio console, locate a workspace and click **Management Center**.

Figure 3-256 Management Center



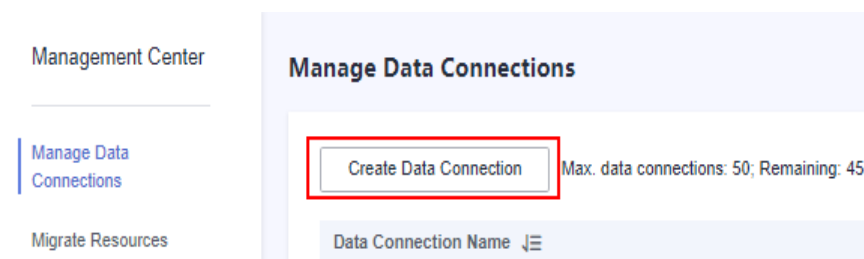
**Step 2** In the navigation pane, choose **Manage Data Connections**.

**Figure 3-257** Creating a data connection



**Step 3** Click **Create Data Connection**.

**Figure 3-258** Creating a data connection



**Step 4** Configure parameters by referring to [Table 3-244](#) and create a data connection named **python\_test**.

**Table 3-244** Host connection parameters

Parameter	Mandatory	Description
Data Connection Name	Yes	Name of the host connection. The value can contain only letters, digits, hyphens (-), and underscores (_).
Host Address	Yes	IP address of the host. For details, see section "Viewing Details About an ECS" in <i>Elastic Cloud Server User Guide</i> .
Agent	Yes	Agents provided by the CDM cluster.
Port	Yes	SSH port number of the host.
Username	Yes	Username of the host
Login Mode	Yes	Selects the login mode of the host. <ul style="list-style-type: none"> <li>Key pair</li> <li>Password</li> </ul>

Parameter	Man dator y	Description
Key pair	Yes	If <b>Key Pair</b> is the login mode of the host, the user needs to obtain the private key file, upload it to OBS, and select an OBS path. This parameter is available only when <b>Login Mode</b> is set to <b>Key Pair</b> . <b>NOTE</b> The uploaded private key file must be in PEM format, and the uploaded private key file and the public key configured on the host must be in the same key pair.
Key Pair Password	No	If no password is set for the key pair, you do not need to set this parameter.
Password	Yes	If the login mode of the host is to use a password, enter a login password.
Host Connection Description	No	Descriptive information about the host connection.

 **NOTE**

The key parameters are as follows:

- **Host Address:** Enter the IP address of the **ECS**.
- **Agent:** Select the **CDM cluster**.

**Step 5** Click **Test** to test connectivity of the data connection. If the test fails, the data connection fails to be created.

**Step 6** After the test is successful, click **OK**. The system will create the data connection for you.

----End

## Developing a Python Script

**Step 1** Choose **DataArts Factory > Develop Script** and create a Python script named **python\_test**.

**Step 2** Edit the Python statement in the editor, select the host connection, and click **Submit** and **Unlock**.

**Step 3** Click **Execute** to execute the Python statement.

**Step 4** View the script execution result.

----End

### 3.4.11.6 Developing a DWS SQL Job

This section describes how to use the DWS SQL operator to develop a job on DataArts Factory.

## Scenario

This tutorial describes how to develop a DWS job to collect the sales volume of a store on the previous day.

## Preparing the Environment

- Enable DWS and create a DWS cluster for running DWS SQL jobs.
- Enable CDM incremental packages and create a CDM cluster.  
Ensure that the VPC, subnet, and security group of the CDM cluster are the same as those of the DWS cluster so that the two clusters can communicate with each other.

## Creating a DWS Data Connection

Before developing a DWS SQL job, you must create a data connection to DWS on the **Manage Data Connections** page of **Management Center**. The data connection name is **dws\_link**.

The key parameters are as follows:

- **Cluster Name:** Select the DWS cluster you have created when preparing the environment.
- **Agent:** Select the CDM cluster you have created when preparing the environment.

## Creating a Database

Create a **gaussdb** database by following the instructions in [Creating a Database](#).

## Creating Data Tables

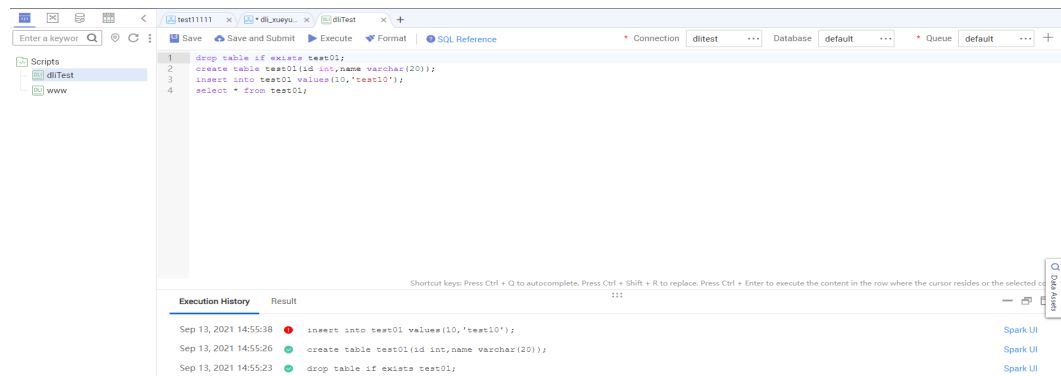
Create tables **trade\_log** and **trade\_report** in the **gaussdb** database. The following is an example script for creating the tables:

```
create schema store_sales;
set current_schema= store_sales;
drop table if exists trade_log;
CREATE TABLE trade_log
(
    sn          VARCHAR(16),
    trade_time  DATE,
    trade_count INTEGER(8)
);
set current_schema= store_sales;
drop table if exists trade_report;
CREATE TABLE trade_report
(
    rq         DATE,
    trade_total INTEGER(8)
);
```

## Developing a DWS SQL Script

Choose **Development > Develop Script** and create a DWS SQL script named **dws\_sql**. Enter an SQL statement in the editor to collect the sales amount of the previous day.

**Figure 3-259** Developing a script



**Key notes:**

- The script development area in **Figure 3-259** is a temporary debugging area. After you close the script tab, the development area will be cleared. You can click **Submit** to save and submit a script version.
- **Connection:** Select the data connection created in **Creating a DWS Data Connection**.

**Developing a DWS SQL Job**

After developing the DWS SQL script, create a job for periodically executing the DWS SQL script.

**Step 1** Create an empty job named **job\_dws\_sql**.

**Figure 3-260** Creating the job\_dws\_sql job

**Create Job** ×

A maximum of 10,000 jobs can be created. You can create 9,989 more jobs.

\* Job Name:

\* Job Type:  Batch processing  Real-time processing

\* Mode:  Pipeline  Single node

\* Creation Method:

\* Select Directory:  +

Owner ?:  × +

Priority:  High  Medium  Low

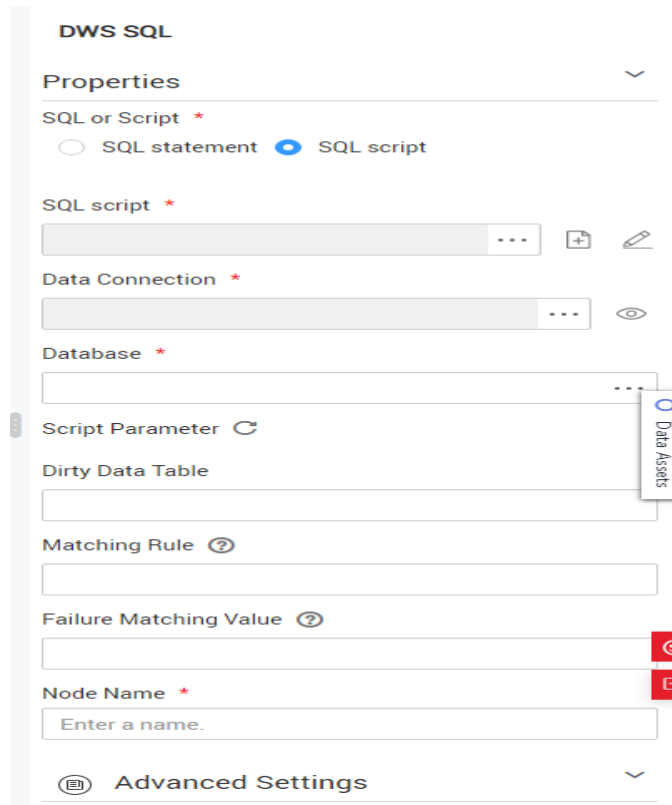
Agency ?:  +

\* Log Path:

To change the log path, go to the WorkSpaces page.  
For details, see the documentation.

**Step 2** Go to the job development page, drag the DWS SQL node to the canvas, and click the node to configure its properties.


**Figure 3-261** Configuring properties for the DWS SQL node



Key properties:

- **SQL script:** Associate with the **dws\_sql** script developed in [Developing a DWS SQL Script](#).
- **Data Connection:** Select the data connection configured in the **dws\_sql** script. The data connection can be changed.
- **Database:** Select the database configured in the **dws\_sql** script. The database can be changed.
- **Script Parameter:** Obtain the value of **yesterday** using the following EL expression:  

```
#{Job.getYesterday("yyyy-MM-dd")}
```
- **Node Name:** The name of the **dws\_sql** script is displayed by default. The name can be changed.

**Step 3** After configuring the job, click  to test it.

**Step 4** If the test is successful, click the blank area on the canvas and then the **Scheduling Setup** tab on the right. On the displayed page, configure the scheduling policy.

**Figure 3-262** Configuring the scheduling policy

Scheduling Type \*

Run once  Run periodically  Event-based ?

Scheduling Properties ▾

From \*  × |  to  × |

Never

Recurrence \*  ▾

Start Time \*  ▾ h  ▾ min

Dependency Properties ▾

Dependency Job  ▾ |  🔍

[Parse Dependency](#)

Name	Recurrence	Scheduled At	Opera...
------	------------	--------------	----------

Parameter descriptions:

From Aug 6 to Aug 31 in 2021, the job was executed once at 02:00 every day.

**Step 5** Click **Submit** and then **Execute**. The job will be executed automatically every day.

----End

### 3.4.11.7 Developing a Hive SQL Job

This section introduces how to develop Hive SQL scripts on DataArts Factory.

#### Scenario Description

As a one-stop big data development platform, DataArts Factory supports development of multiple big data tools. Hive is a data warehouse tool running on Hadoop. It can map structured data files to a database table and provides a simple SQL search function that converts SQL statements into MapReduce tasks.

#### Preparations

- MRS has been enabled and an MRS cluster has been created for running Hive SQL jobs.  
The MRS cluster must contain the Hive component.
- Cloud Data Migration (CDM) has been enabled and a CDM cluster has been created for providing an agent for communication between DataArts Factory and MRS.  
Ensure that the VPC, subnet, and security group of the CDM cluster are the same as those of the MRS cluster so that the two clusters can communicate with each other.

## Creating a Hive Data Connection

Before developing a Hive SQL script, you must create a data connection to MRS Hive on the **Manage Data Connections** page of **Management Center**. The data connection name is **hive1009**.

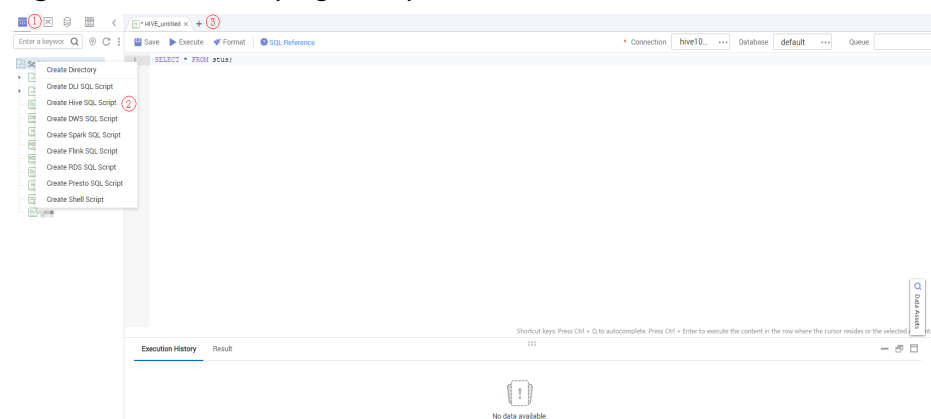
Description of key parameters:

- **Cluster Name:** Enter the name of the created MRS cluster.
- **Agent:** Select the created CDM cluster.

## Developing a Hive SQL Script

Choose **Development > Develop Script** and create a Hive SQL script named **hive\_sql**. Then enter SQL statements in the editor to fulfill business requirements.

Figure 3-263 Developing a script



Notes:

- The script development area in [Figure 3-263](#) is a temporary debugging area. After you close the tab page, the development area will be cleared. You can click **Submit** to save and submit a script version.
- Data Connection: Connection created in [Creating a Hive Data Connection](#).

## Developing a Hive SQL Job

After the Hive SQL script is developed, build a periodically deducted job for the Hive SQL script so that the script can be executed periodically.

**Step 1** Create an empty DataArts Factory job named **job\_hive\_sql**.



**Figure 3-264** Creating a job named job\_hive\_sql

**Create Job** ×

A maximum of 10,000 jobs can be created. You can create 9,989 more jobs.

\* Job Name:

\* Job Type:  Batch processing  Real-time processing

\* Mode:  Pipeline  Single node

\* Creation Method:

\* Select Directory:  +

Owner ?:  × +

Priority:  High  Medium  Low

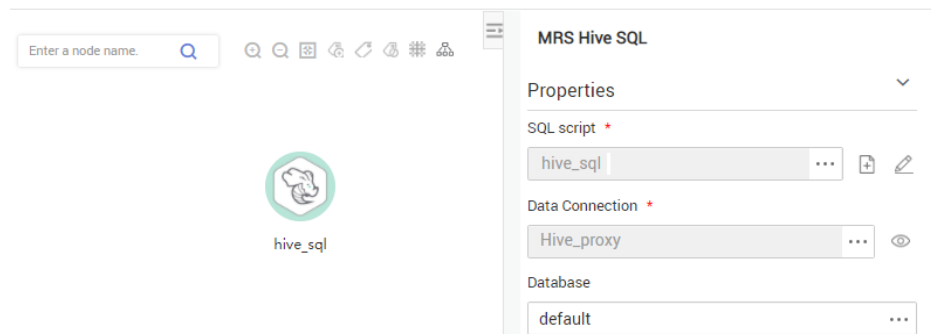
Agency ?:  +

\* Log Path:

[To change the log path, go to the WorkSpaces page.](#)  
[For details, see the documentation.](#)

**Step 2** Go to the job development page, drag the MRS Hive SQL node to the canvas, and click the node to configure node properties.


**Figure 3-265** Configuring properties for an MRS Hive SQL node



Description of key properties:

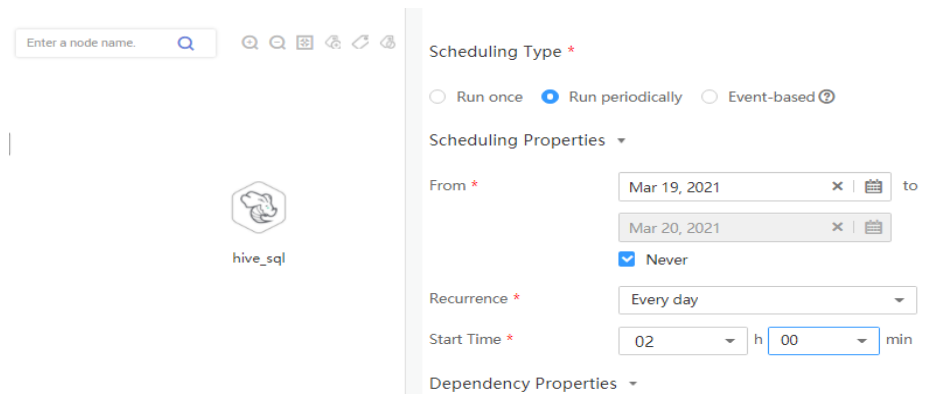
- SQL Script: Hive SQL script **hive\_sql** that is developed in [Developing a Hive SQL Script](#).
- Data Connection: Data connection that is configured in the SQL script **hive\_sql** is selected by default. The value can be changed.
- Database: Database that is configured in the SQL script **hive\_sql** and is selected by default. The value can be changed.

- Node Name: Name of the SQL script **hive\_sql** by default. The value can be changed.

**Step 3** After configuring the job, click  to test it.

**Step 4** If the job runs successfully, click the blank area on the canvas and configure the job scheduling policy on the scheduling configuration page on the right.

**Figure 3-266** Configuring the scheduling mode



Note:

From Jan 1 to Jan 25 in 2021, the job was executed at 02:00 every day.

**Step 5** Click **Submit** and **Execute**. The job will be automatically executed every day.

----End

### 3.4.11.8 Developing a DLI Spark Job

This section introduces how to develop a DLI Spark job on DataArts Factory.

#### Scenario Description

In most cases, SQL is used to analyze and process data when using Data Lake Insight (DLI). However, SQL is usually unable to deal with complex processing logic. In this case, Spark jobs can help. This section uses an example to demonstrate how to submit a Spark job on DataArts Factory.

The general submission procedure is as follows:

1. Create a DLI cluster and run a Spark job using physical resources of the DLI cluster.
2. Obtain a demo JAR package of the Spark job and associate with the JAR package on DataArts Factory.
3. Create a DataArts Factory job and submit it using the DLI Spark node.

#### Preparations

- Object Storage Service (OBS) has been enabled and a bucket, for example, **obs://dlfexample**, has been created for storing the JAR package of the Spark job.

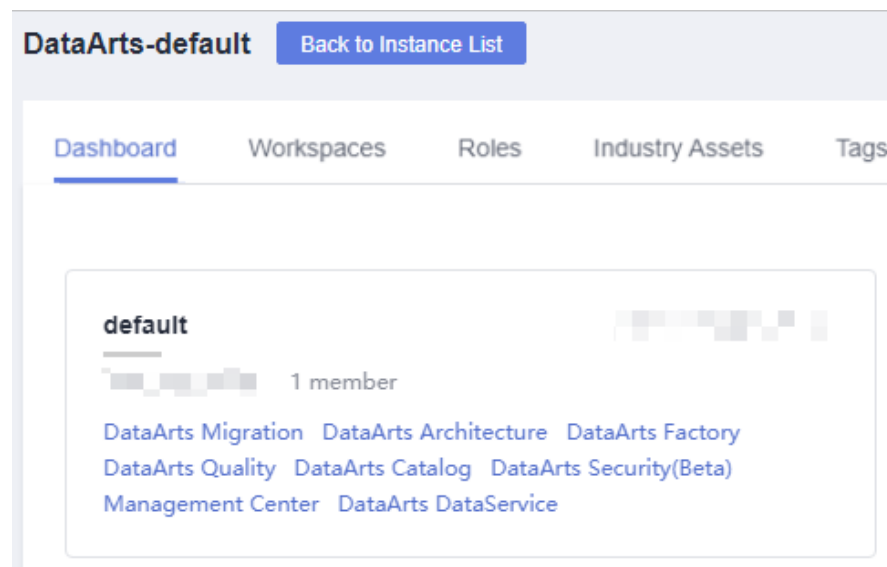
- DLI has been enabled, and the Spark cluster **spark\_cluster** has been created for providing physical resources required for the Spark job.

## Obtaining Spark Job Code

The Spark job code used in this example comes from the maven repository that can be download from [https://repo.maven.apache.org/maven2/org/apache/spark/spark-examples\\_2.10/1.1.1/spark-examples\\_2.10-1.1.1.jar](https://repo.maven.apache.org/maven2/org/apache/spark/spark-examples_2.10/1.1.1/spark-examples_2.10-1.1.1.jar). This Spark job is to calculate the approximate value of  $\pi$ .

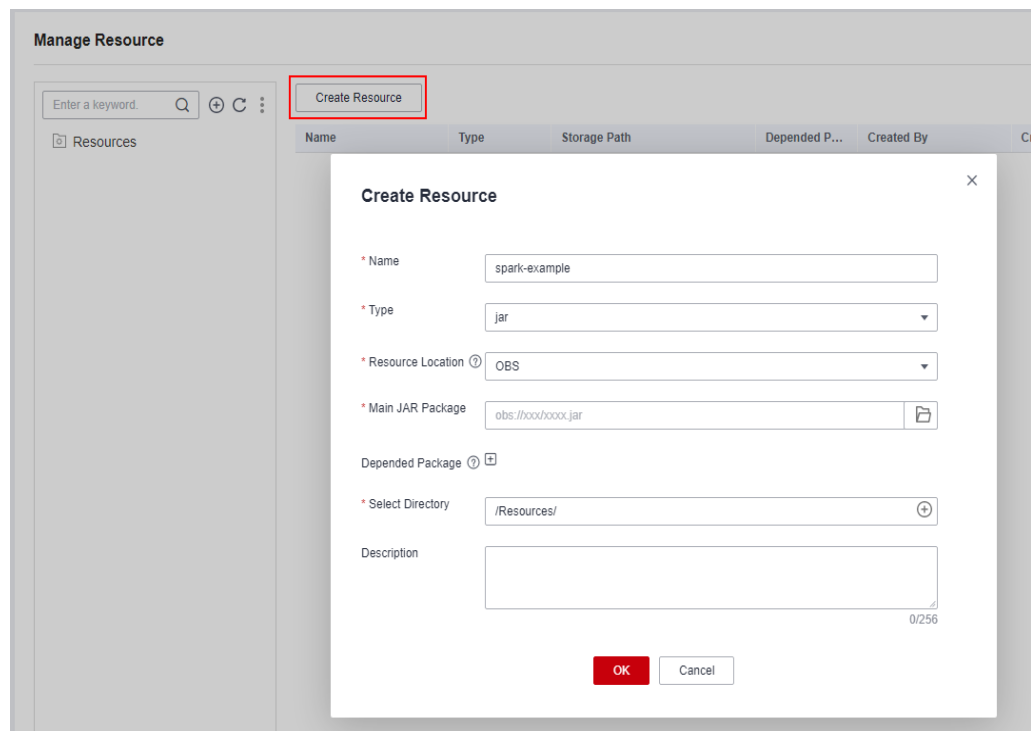
- Step 1** After obtaining the JAR package of the Spark job codes, upload it to the OBS bucket. The save path is **obs://dlfexample/spark-examples\_2.10-1.1.1.jar**.
- Step 2** Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.

Figure 3-267 DataArts Factory



- Step 3** In the navigation tree on the left, choose **Configuration > Manage Resource**. Click **Create Resource** and create resource **spark-example** on DataArts Factory and associate it with the JAR package obtained in [Step 1](#).

**Figure 3-268** Creating a resource



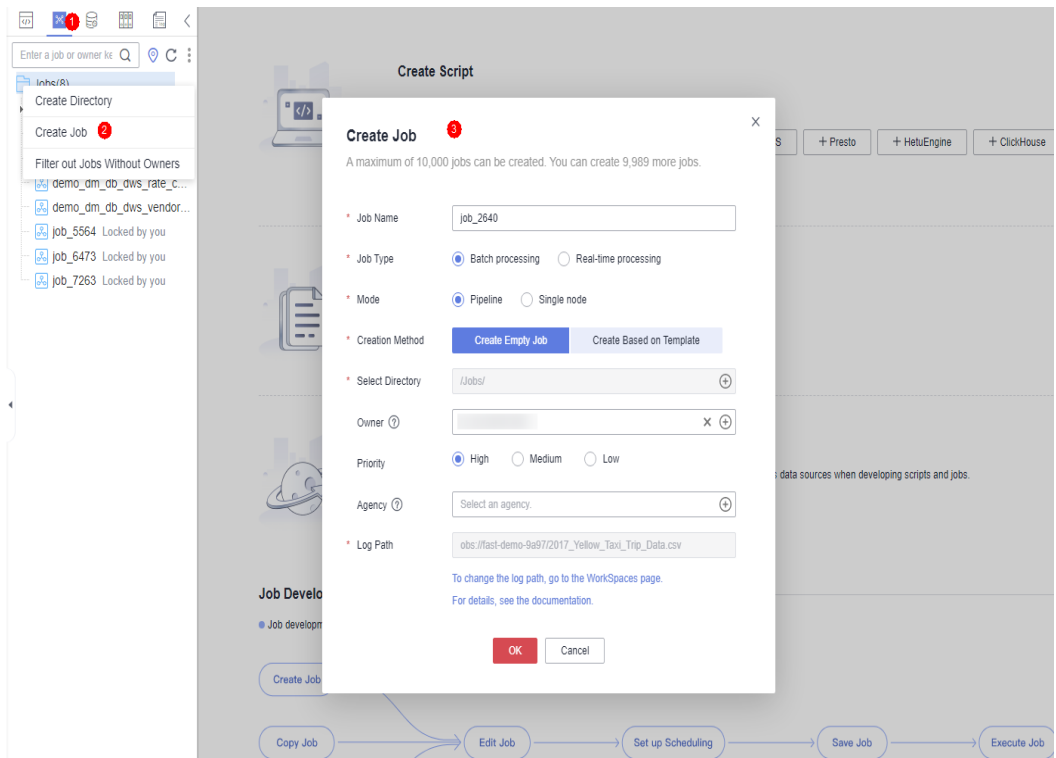
----End

## Submitting a Spark Job

You need to create a job on DataArts Factory and submit the Spark job using the DLI Spark node of the job.

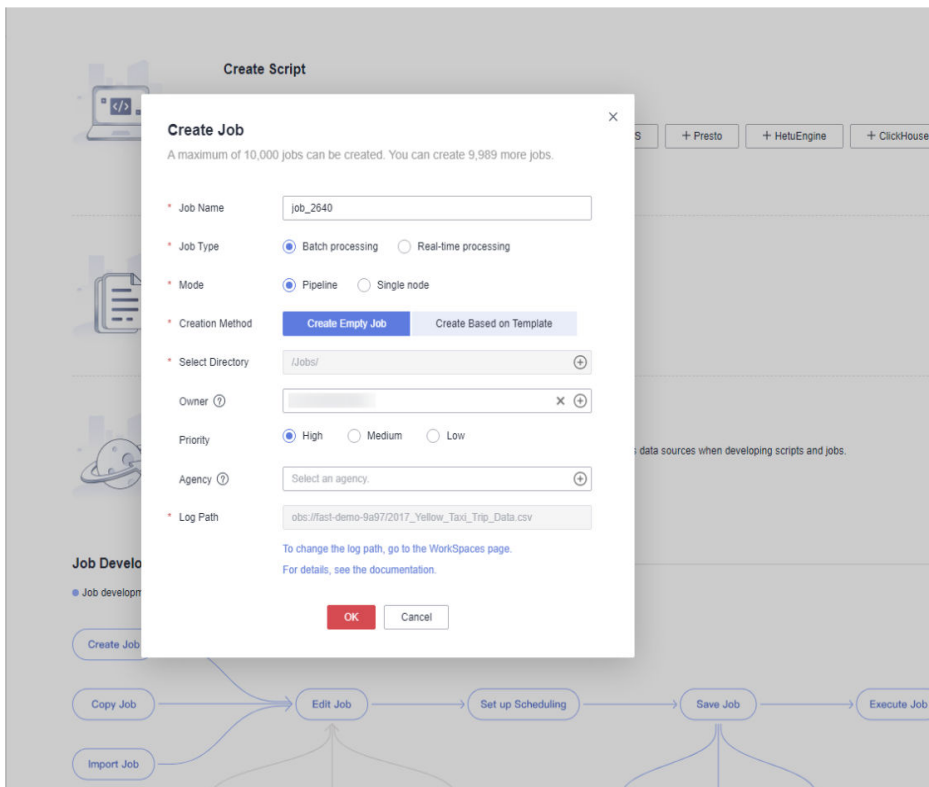
**Step 1** Create a job named **job\_DLI\_Spark** for the DataArts Factory module.

Figure 3-269 Creating a job




**Step 2** Go to the job development page, drag the DLI Spark node to the canvas, and click the node to configure node properties.

Figure 3-270 Configuring node properties

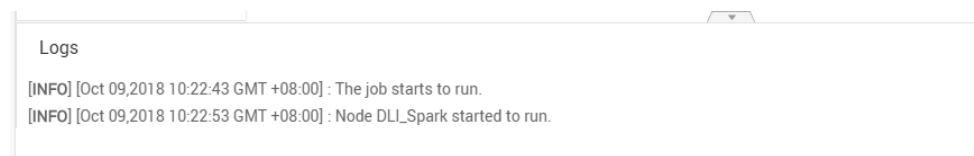


Description of key properties:

- **DLI Cluster Name:** name of the Spark cluster created in DLI
- **Job Running Resource:** Maximum CPU and memory resources that can be used when a DLI Spark node is running.
- **Major Job Class:** major class of a DLI Spark node. In this example, the major class is **org.apache.spark.examples.SparkPi**.
- **JAR Package:** Resource created in [Step 3](#).

**Step 3** After the job orchestration is complete, click  to test the job.

**Figure 3-271** Job logs (for reference only)



**Step 4** If no error is recorded in logs, save and submit the job.

----End

### 3.4.11.9 Developing an MRS Flink Job

This section describes how to develop an MRS Flink job on DataArts Factory. Use an MRS Flink job to count the number of words.

#### Prerequisites

- You have the permission to access OBS paths.
- MRS has been enabled and an MRS cluster has been created.

#### Data Preparation

- Download the Flink job resource package **wordcount.jar** from <https://github.com/apache/flink/tree/master/flink-examples/flink-examples-streaming/src/main/java/org/apache/flink/streaming/examples/wordcount>.
- Prepare the data file **in.txt**, which contains some English words.

#### Procedure

**Step 1** Upload the job resource package and data file to the OBS bucket.

 **NOTE**

In this example, upload **WordCount.jar** to **lkj\_test/WordCount.jar** and **word.txt** to **lkj\_test/input/word.txt**.

**Step 2** Create an empty job named **job\_MRS\_Flink**.

**Figure 3-272** Creating a job

### Create Job

✕

A maximum of 10,000 jobs can be created. You can create 9,989 more jobs.

- \* Job Name
- \* Job Type  Batch processing  Real-time processing
- \* Mode  Pipeline  Single node
- \* Creation Method Create Empty Job Create Based on Template
- \* Select Directory  +
- Owner ?  ✕ +
- Priority  High  Medium  Low
- Agency ?  +
- \* Log Path

[To change the log path, go to the WorkSpaces page.](#)  
[For details, see the documentation.](#)

OK
Cancel

**Step 3** Go to the job development page, drag the **MRS Flink** node to the canvas, and click the node to configure its properties.

**Figure 3-273** Configuring properties for an MRS Flink node

Parameter descriptions:

```
--Flink job name
wordcount
--MRS cluster name
Select an MRS cluster.
--Program parameter
-c org.apache.flink.streaming.examples.wordcount.WordCount
--Flink job resource package
wordcount
--Input data path
obs://dlf/lkj_test/input/word.txt
--Output data path
obs://dlf/lkj_test/output.txt
```

Specifically:

**obs://dlf/lkj\_test/input/word.txt** is the directory where the **wordcount.jar** parameters are passed. You can pass the words to count.

**obs://dlf/lkj\_test/output.txt** is the directory where the output parameter file is stored. (If the **output.txt** file already exists, an error is reported.)

**Step 4** Click **Test** to execute the MRS Flink job.

**Step 5** After the test is complete, click **Submit**.

**Step 6** Choose **Monitor Job** in the navigation pane and view the job execution result.

**Step 7** View the returned records in the OBS bucket. (Skip this step if the return function is not configured.)

----End

### 3.4.11.10 Developing an MRS Spark Python Job

This section describes how to develop an MRS Spark Python on DataArts Factory.

## Case 1: Using an MRS Spark Python Job to Count the Number of Words

### Prerequisites

You have the permission to access OBS paths.

### Data preparation

- Prepare the script file **wordcount.py** with the following content:

```
# -*- coding: utf-8 -*-
import sys
from pyspark import SparkConf, SparkContext
def show(x):
    print(x)
if __name__ == "__main__":
    if len(sys.argv) < 2:
        print ("Usage: wordcount <inputPath> <outputPath>")
        exit(-1)
    # Create SparkConf.
    conf = SparkConf().setAppName("wordcount")
    # Create SparkContext. Pass the conf=conf parameter.
    sc = SparkContext(conf=conf)
    inputPath = sys.argv[1]
    outputPath = sys.argv[2]
    lines = sc.textFile(name = inputPath)
    # Split each line of data by space to obtain words.
```



```

words = lines.flatMap(lambda line:line.split(" "),True)
# Pair each word into a tuple count 1.
pairWords = words.map(lambda word:(word,1),True)
# Use three partitions (reduceByKey) for summarization.
result = pairWords.reduceByKey(lambda v1,v2:v1+v2)
# Print the result.
result.foreach(lambda t :show(t))
# Save the result to a file.
result.saveAsTextFile(outputPath)
# Stop SparkContext.
sc.stop()
    
```

**NOTE**

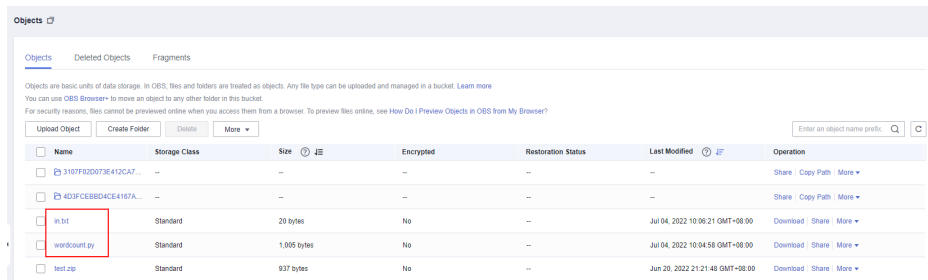
The encoding format must be set to UTF-8. Otherwise, an error will occur during script execution.

- Prepare the data file **in.txt**, which contains some English words.

**Procedure**

**Step 1** Upload the script and data file to the OBS bucket.

**Figure 3-274** Uploading files to an OBS bucket



**NOTE**

In this example, upload **wordcount.py** and **in.txt** to **obs://obs-tongji/python/**.

**Step 2** Create an empty job named **job\_MRS\_Spark\_Python**.

Figure 3-275 Creating a job

### Create Job

×

A maximum of 10,000 jobs can be created. You can create 9,989 more jobs.

\* Job Name

\* Job Type  Batch processing  Real-time processing

\* Mode  Pipeline  Single node

\* Creation Method Create Empty Job Create Based on Template

\* Select Directory  +

Owner  ? × +

Priority  High  Medium  Low

Agency  ? +

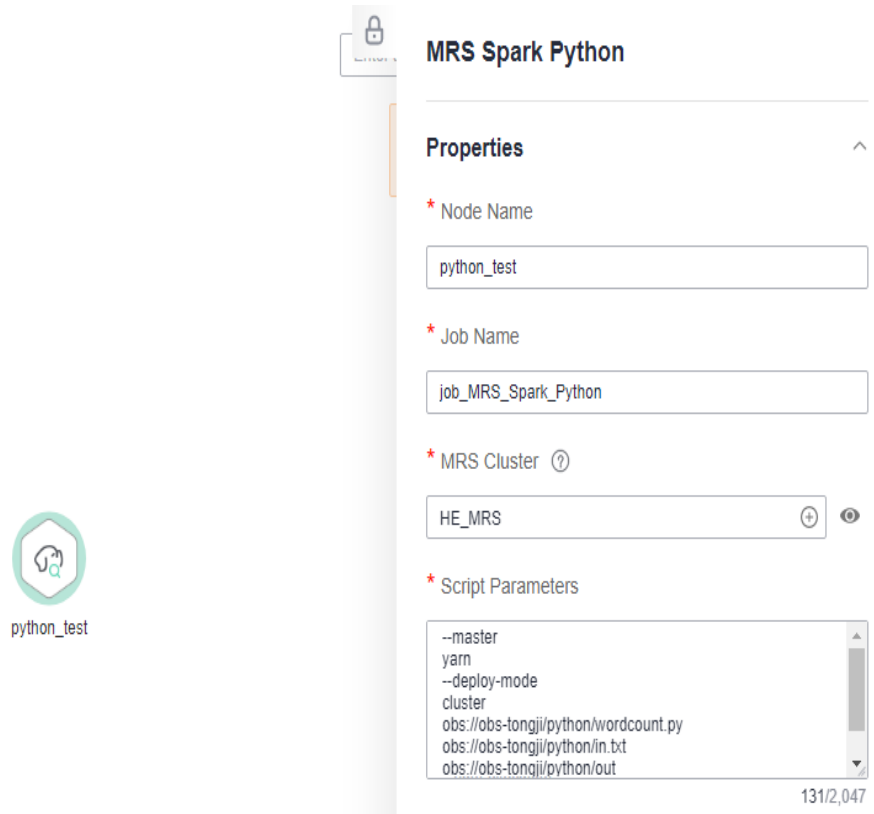
\* Log Path

[To change the log path, go to the WorkSpaces page.](#)  
[For details, see the documentation.](#)

OK Cancel

**Step 3** Go to the job development page, drag the **MRS Spark Python** node to the canvas, and click the node to configure its properties.

Figure 3-276 Configuring properties for an MRS Spark Python node



Parameter descriptions:

```
--master  
yarn  
--deploy-mode  
cluster  
obs://obs-tongji/python/wordcount.py  
obs://obs-tongji/python/in.txt  
obs://obs-tongji/python/out
```

Specifically:

**obs://obs-tongji/python/wordcount.py** is the directory where the script is stored.

**obs://obs-tongji/python/in.txt** is the directory where the **wordcount.py** parameters are passed. You can pass the words to count.

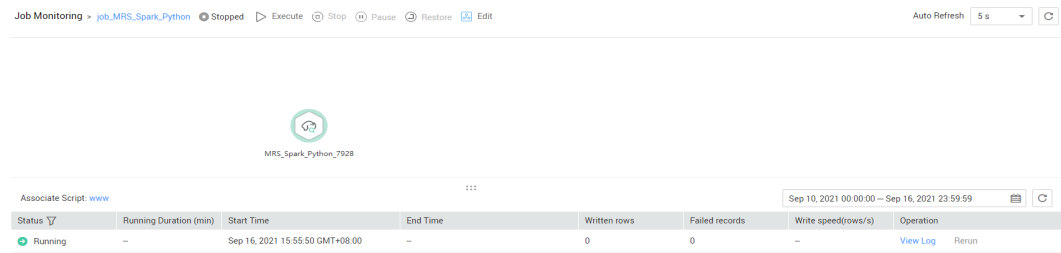
**obs://obs-tongji/python/out** is the directory where output parameters are stored. This directory will also be created in the OBS bucket automatically. If the **out** directory already exists in the OBS bucket, an error will occur.

**Step 4** Click **Test** to execute the script job.

**Step 5** After the test is complete, click **Submit**.

**Step 6** Choose **Monitor Job** in the navigation pane and view the job execution result.

**Figure 3-277** Viewing the job execution result



The job log shows that the job was successfully executed.

**Figure 3-278** Job run logs

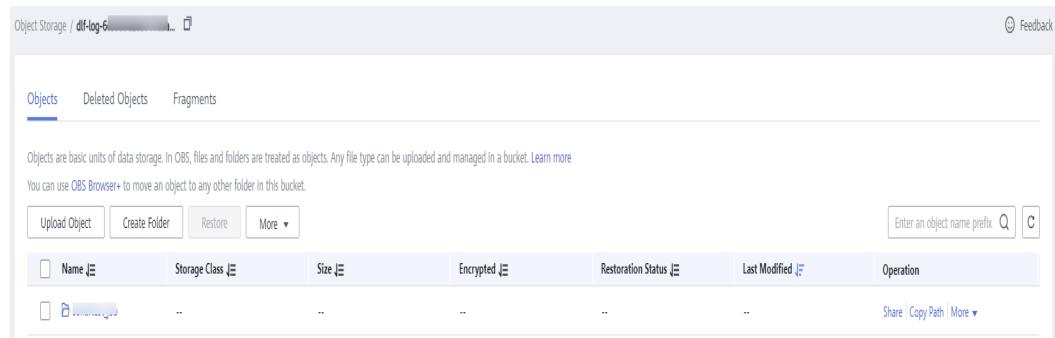


**Figure 3-279** Job execution status



**Step 7** View the returned records in the OBS bucket. (Skip this step if the return function is not configured.)

**Figure 3-280** Viewing the returned records in the OBS bucket



----End

## Case 2: Using an MRS Spark Python Job to Print hello python

### Prerequisites

You have the permission to access OBS paths.

### Data preparation

Prepare the script file **zt\_test\_sparkPython1.py** with the following content:

```
from pyspark import SparkContext, SparkConf
conf = SparkConf().setAppName("master"). setMaster("yarn")
sc = SparkContext(conf=conf)
print("hello python")
sc.stop()
```

### Procedure

- Step 1** Upload the script file to an OBS bucket.
- Step 2** Create an empty job.
- Step 3** Go to the job development page, drag the **MRS Spark Python** node to the canvas, and click the node to configure its properties.

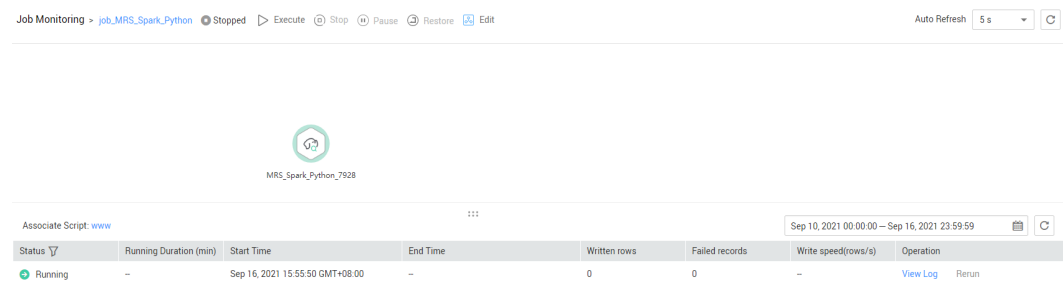
Parameter descriptions:

```
--master
yarn
--deploy-mode
cluster
obs://obs-tongji/python/zt_test_sparkPython1.py
```

**zt\_test\_sparkPython1.py** indicates the directory where the script is stored.

- Step 4** Click **Test** to execute the script job.
- Step 5** After the test is complete, click **Submit**.
- Step 6** Choose **Monitor Job** in the navigation pane and view the job execution result.

**Figure 3-281** Viewing the job execution result



- Step 7** Verify the log.

Login to MRS Manager and check that the log on YARN contains **hello python**.

Figure 3-282 Viewing logs on YARN

```
Log Type: prelaunch.err
Log Upload Time: Mon May 31 10:35:05 +0800 2021
Log Length: 0

Log Type: prelaunch.out
Log Upload Time: Mon May 31 10:35:05 +0800 2021
Log Length: 100
Setting up env variables
Setting up job resources
Copying debugging information
Launching container

Log Type: stderr
Log Upload Time: Mon May 31 10:35:05 +0800 2021
Log Length: 510
SLF4J: Class path contains multiple SLF4J bindings
SLF4J: Found binding in [jar:file:/usr/lib/gdata/hadoop/data24/am/localdir/filescache/S27/spark-wdhuve-2x.rzp/sl4j-log4j12-1.7.16.jar/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/share/sl4j-log4j12-1.7.25/sl4j-log4j12-1.7.25.jar/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/notes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]

Log Type: stdout
Log Upload Time: Mon May 31 10:35:05 +0800 2021
Log Length: 11
hello python

Log Type: stdout.log
Log Upload Time: Mon May 31 10:35:05 +0800 2021
Log Length: 42817
Showing 4096 bytes of 42817 total. Click here for the full log.
```

----End

# 4 FAQs

---

## 4.1 Consultation

### 4.1.1 Regions

#### Concepts

We use a region to identify the location of a data center. You can create resources in a specific region.

- A region is a physical data center. Each region is completely independent, improving fault tolerance and stability. After a resource is created, its region cannot be changed.

#### Region Selection

You are advised to select a region close to you or your target users. This reduces network latency and improves access rate.

#### Changing the Region of an Instance

- You cannot change the region of an instance.

#### Regions and Endpoints

An endpoint is the **request address** for calling an API. Endpoints vary depending on services and regions. You can obtain endpoints from [Regions and Endpoints](#).

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

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

## Adding a Member and Assigning a Role

1. Log in to the DataArts Studio console and access the **Workspaces** page.
2. On the **Workspaces** page, locate the target workspace and click **Edit** in the **Operation** column.
3. Click **Add** under **Workspace Members**. In the displayed **Add Member** dialog box, select **Add User** or **Add Group**, select a member account from the drop-down list, and select a role for it.
4. Click **OK**. You can view or modify the members and roles in the member list, or remove members from the workspace.

### 4.1.3 Can I Delete DataArts Studio Workspaces?

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

### 4.1.4 Can I Transfer a Trial Instance to Another Account?

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

### 4.1.5 Does DataArts Studio Support Version Downgrade?

No. You cannot downgrade a created DataArts Studio instance.

## 4.2 Management Center

### 4.2.1 What Are the Precautions for Creating Data Connections?

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

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

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

To resolve this issue, perform the following steps:

**Step 1** Check whether the CDM cluster is stopped.

- If yes, start the CDM cluster and check whether the data connection in Management Center recovers.
- If no, go to [step 2](#).

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



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

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

----End

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

Possible causes are as follows:

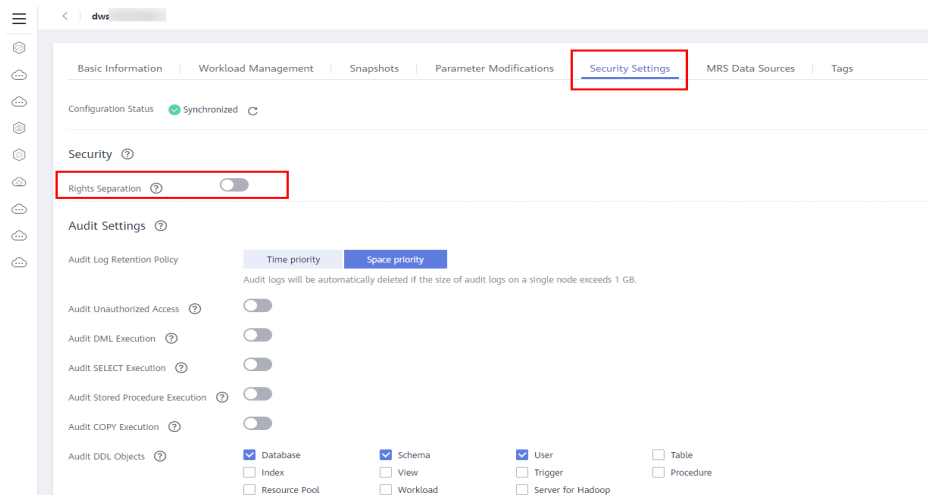
- Hive/HBase components were not selected during MRS cluster creation.
- The network between the CDM cluster and MRS cluster was disconnected when an MRS data connection is created.

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

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

The failure may be caused by the rights separation function of the DWS cluster. On the DWS console, click the corresponding cluster, choose **Security Settings**, and disable **Rights Separation**.

**Figure 4-1** Disabling Rights Separation for the DWS cluster



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

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

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

You are advised to choose a proxy connection.

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

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

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

## 4.2.8 Can I Delete Workspaces?

No, but you can change the names of workspaces.

# 4.3 DataArts Migration

## 4.3.1 General

### 4.3.1.1 What Are the Advantages of CDM?

CDM is developed based on a distributed computing framework and leverages the parallel data processing technology. [Table 4-1](#) details the advantages of CDM.

**Table 4-1** CDM advantages

Item	User-Developed Script	CDM
Ease of use	<p>You need to prepare server resources, and install and configure software, which is time-consuming.</p> <p>Because the data source types are different, the program uses different access interfaces, such as JDBC and native APIs, to read and write data. In this case, various libraries and SDKs are required when you write data migration scripts, resulting in high development and management costs.</p>	<p>CDM provides a web-based management console for enabling services on web pages in real time.</p> <p>You can migrate data by configuring data sources and migration jobs on the GUI and CDM will manage and maintain the data sources and migration jobs for you. In other words, you only need to focus on the data migration logic without worrying about the environment, which greatly reduces development and maintenance costs.</p> <p>CDM also provides RESTful APIs to support third-party system calling and integration.</p>
Real-time monitoring	<p>You need to select specific versions to develop as required.</p>	<p>You can use Cloud Eye to automatically monitor CDM clusters in real time and manage alarms and notifications, so that you can keep track of CDM cluster performance metrics.</p>
O&M free	<p>You need to develop and optimize O&amp;M functions, especially alarm and notification functions, to ensure system availability. Otherwise, manual attendance is required.</p>	<p>With CDM, you do not need to maintain resources such as servers and VMs. CDM has the log, monitoring, and alarm functions, which send notifications to related personnel in a timely manner to avoid 24/7 hours of manual O&amp;M.</p>
High efficiency	<p>During data migration, the read and write process is completed in one job. Limited by available resources, the performance is poor and cannot meet the requirements of scenarios where massive sets of data need to be migrated.</p>	<p>Based on the distributed computing framework, CDM jobs are split into independent sub-jobs and executed concurrently, which drastically improves data migration efficiency. In addition, efficient data import interfaces are provided to import data from Hive, HBase, MySQL databases, and Data Warehouse Service (DWS).</p>

Item	User-Developed Script	CDM
Various data sources	Different tasks must be developed for different data sources, generating a number of scripts.	Data sources such as databases, Hadoop services, NoSQL databases, data warehouses, and files are supported.
Different network environments	As the cloud computing technology develops, user data may be stored in different environments, such as public clouds, on-premises or hosted Internet data centers (IDCs), and hybrid scenarios. In heterogeneous environments, data migration is subject to various factors, for example, network connectivity, which causes inconvenience for development and maintenance.	CDM helps you easily cope with various data migration scenarios, including data migration to the cloud, data exchange on the cloud, and data migration to on-premises service systems, regardless of whether the data is stored on on-premises IDCs, cloud services, third-party clouds, or self-built databases or file systems on ECSs.

#### 4.3.1.2 What Are the Security Protection Mechanisms of CDM?

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

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

#### 4.3.1.3 How Do I Reduce the Cost of Using CDM?

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

The following details the operations:

1. Suppose that you have created a CDM cluster (no dedicated EIP needs to be bound to the CDM cluster). Record the VPC and subnet where the CDM cluster is located.
2. Create a NAT gateway. Select the same VPC and subnet as the CDM cluster.

3. After the NAT gateway is created, return to the NAT gateway console list, click the created gateway name, and then click **Add SNAT Rule**.
4. Select a subnet and an EIP. If no EIP is available, apply for one.

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

 **NOTE**

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

#### 4.3.1.4 Can I Upgrade a CDM Cluster?

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

#### 4.3.1.5 How Is the Migration Performance of CDM?

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

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

[Table 4-2](#) lists the number of concurrent jobs for different CDM cluster versions.

**Table 4-2** Concurrent tasks

Flavor	<code>cdm.large</code>	<code>cdm.xlarge</code>	<code>cdm.4xlarge</code>
Specifications	Node quantity: 1 vCPUs   memory: 8 vCPUs   16 GB Baseline/Max. bandwidth: 0.8/3 Gbit/s	Node quantity: 1 vCPUs   memory: 16 vCPUs   32 GB Baseline/Max. bandwidth: 4/10 Gbit/s	Node quantity: 1 vCPUs   memory: 64 vCPUs   128 GB Baseline/Max. bandwidth: 36/40 Gbit/s
Number of concurrent jobs	30	100	300

You are advised to use multiple CDM clusters in the following and other scenarios as needed:

- Use different CDM clusters for different purposes, for example, for data migration jobs or as connection agents in the DataArts Studio Management Center.
- Use different CDM clusters for different business departments, such as the finance department and online store.

## 4.3.2 Functions

### 4.3.2.1 Does CDM Support Incremental Data Migration?

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

- Incremental file migration
- Incremental migration of relational databases
- Incremental synchronization using the macro variables of date and time
- Incremental migration of HBase/CloudTable

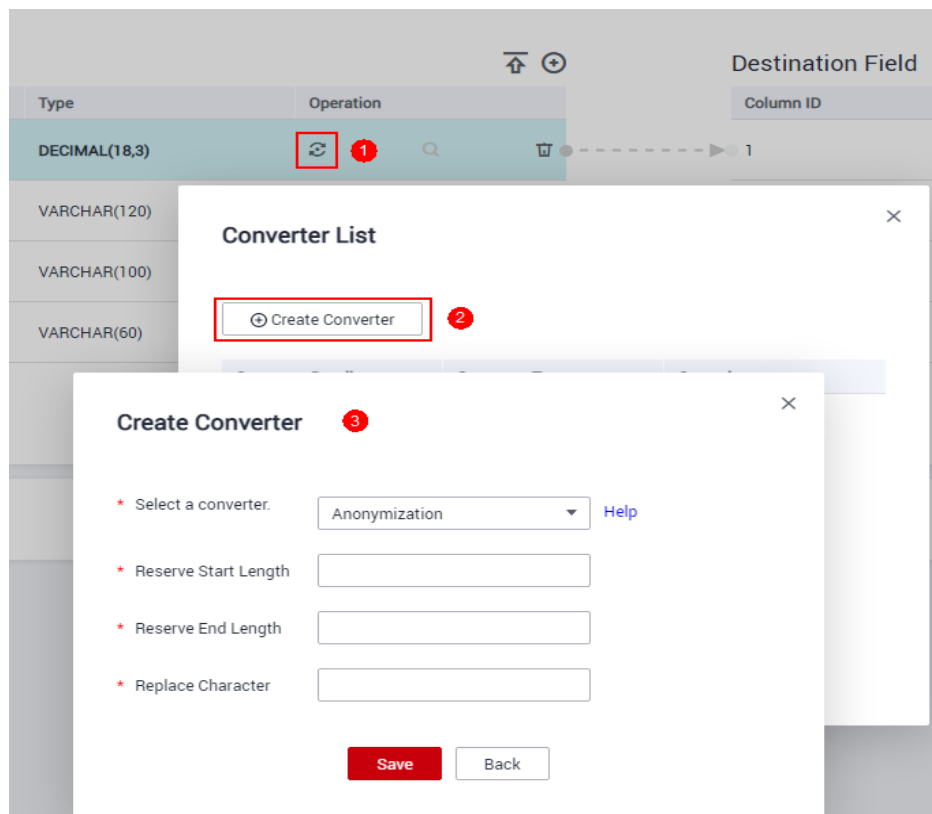
### 4.3.2.2 Does CDM Support Field Conversion?

Yes. CDM supports the following field converters:

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

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

**Figure 4-2** Creating a field converter

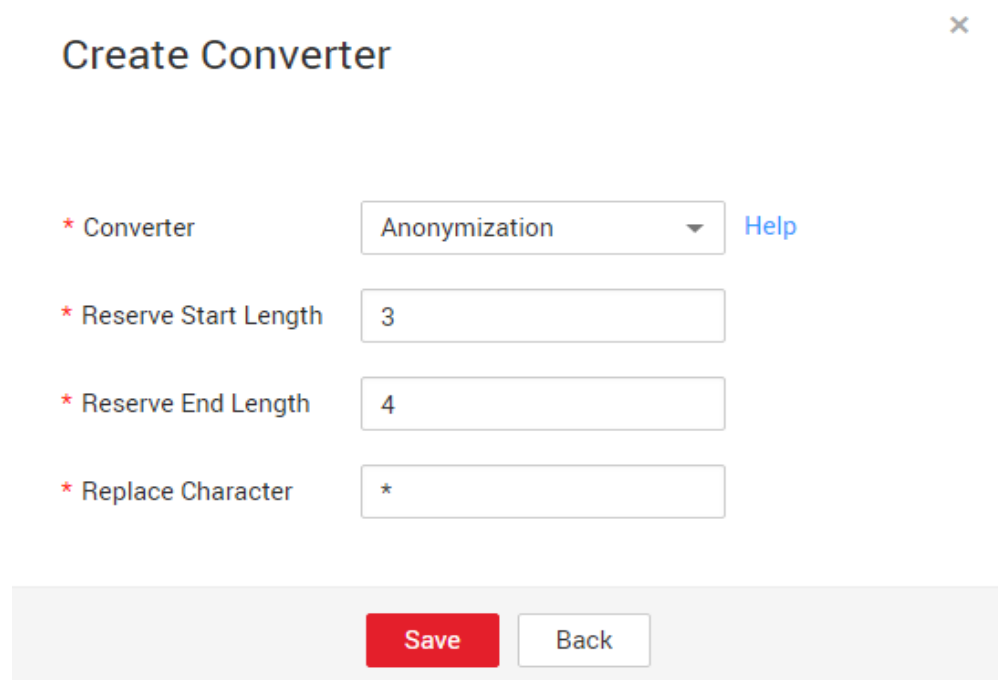


## Anonymization

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

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

Figure 4-3 Anonymization



The screenshot shows a 'Create Converter' dialog box with a close button (x) in the top right corner. The dialog contains the following fields and controls:

- Converter:** A dropdown menu set to 'Anonymization' with a 'Help' link to its right.
- Reserve Start Length:** A text input field containing the number '3'.
- Reserve End Length:** A text input field containing the number '4'.
- Replace Character:** A text input field containing the asterisk character '\*'. A red asterisk is visible to the left of the field.

At the bottom of the dialog, there are two buttons: a red 'Save' button and a white 'Back' button.

## Trim

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

## Reverse String

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

## Replace String

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

## Expression Conversion

This converter uses the JSP expression language (EL) to convert the current field or a row of data. The JSP EL is used to create arithmetic and logical expressions.

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

The expression supports the following environment variables:

- **value**: indicates the current field value.
- **row**: indicates the current row, which is an array type.

The expression supports the following tool classes:

- **StringUtils**: string processing tool class. For details, see **org.apache.commons.lang.StringUtils** of the Java SDK code.
- **DateUtils**: date tool class
- **CommonUtils**: common tool class
- **NumberUtils**: string-to-value conversion class
- **HttpsUtils**: network file read class

Application examples:

1. If the field is of the string type, convert all character strings into lowercase letters, for example, convert **aBC** to **abc**.  
Expression: `StringUtils.lowerCase(value)`
2. Convert all character strings of the current field to uppercase letters.  
Expression: `StringUtils.upperCase(value)`
3. If the field value is a date string in *yyyy-MM-dd* format, extract the year from the field value, for example, extract **2017** from **2017-12-01**.  
Expression: `StringUtils.substringBefore(value,"-")`
4. If the field value is of the numeric type, convert the value to a new value which is two times greater than the original value:  
Expression: `value*2`
5. Convert the field value **true** to **Y** and other field values to **N**.  
Expression: `value=="true"? "Y": "N"`
6. If the field value is of the string type and is left empty, convert it to **Default**. Otherwise, the field value will not be converted.  
Expression: `empty value? "Default":value`
7. Convert date format **2018/01/05 15:15:05** to **2018-01-05 15:15:05**:  
Expression: `DateUtils.format(DateUtils.parseDate(value,"yyyy/MM/dd HH:mm:ss"),"yyyy-MM-dd HH:mm:ss")`
8. Obtain a 36-bit universally unique identifier (UUID):  
Expression: `CommonUtils.randomUUID()`
9. If the field is of the string type, capitalize the first letter, for example, convert **cat** to **Cat**.  
Expression: `StringUtils.capitalize(value)`
10. If the field is of the string type, convert the first letter to a lowercase letter, for example, convert **Cat** to **cat**.  
Expression: `StringUtils.uncapitalize(value)`
11. If the field is of the string type, use a space to fill in the character string to the specified length and center the character string. If the length of the



character string is not shorter than the specified length, do not convert the character string. For example, convert **ab** to meet the specified length 4.

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

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

Expression: `StringUtils.chomp(value)`

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

23. Obtain the first index of any specified character in a character string. If no index is found, **-1** is returned. For example, the first index of **z** or **a** in **zzabyxcdxx**. 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 **yzzybat** after conversion.  
Expression: `StringUtils.leftPad(value,8,"yz")`
34. Concatenate the specified character string to the right of the current character string and specify the length of the concatenated character string. If the length of the current character string is not shorter than the specified length, the character string will not be converted. For example, if **yz** is concatenated to the right of **bat** and the length must be 8 after concatenation, the character string is **batzyzy** after conversion.

- Expression: `StringUtils.rightPad(value,8,"yz")`
35. If the field is of the string type, obtain the length of the current character string. If the character string is null, **0** is returned.  
Expression: `StringUtils.length(value)`
36. If the field is of the string type, delete all the specified character strings from it. For example, delete **ue** from **queued** to obtain **qd**.  
Expression: `StringUtils.remove(value,"ue")`
37. If the field is of the string type, remove the substring at the end of the field. If the specified substring is not at the end of the field, no conversion is performed. For example, remove **.com** at the end of **www.domain.com**.  
Expression: `StringUtils.removeEnd(value,".com")`
38. If the field is of the string type, delete the substring at the beginning of the field. If the specified substring is not at the beginning of the field, no conversion is performed. For example, delete **www.** at the beginning of **www.domain.com**.  
Expression: `StringUtils.removeStart(value,"www.")`
39. If the field is of the string type, replace all the specified character strings in the field. For example, replace **a** in **aba** with **z** to obtain **zbz**.  
Expression: `StringUtils.replace(value,"a","z")`
40. If the field is of the string type, replace multiple characters in the character string at a time. For example, replace **h** in **hello** with **j** and **o** with **y** to obtain **jelly**.  
Expression: `StringUtils.replaceChars(value,"ho","jy")`
41. If the string starts with the specified prefix (case sensitive), **true** is returned; otherwise, **false** is returned. For example, **abcdef** starts with **abc**, so that **true** is returned.  
Expression: `StringUtils.startsWith(value,"abc")`
42. If the field is of the string type, delete all the specified characters from the field. For example, delete all **x**, **y**, and **z** from **abcyx** to obtain **abc**.  
Expression: `StringUtils.strip(value,"xyz")`
43. If the field is of the string type, delete all the specified characters at the end of the field, for example, delete all spaces at the end of the field.  
Expression: `StringUtils.stripEnd(value,null)`
44. If the field is of the string type, delete all the specified characters at the beginning of the field, for example, delete all spaces at the beginning of the field.  
Expression: `StringUtils.stripStart(value,null)`
45. If the field is of the string type, obtain the substring after the specified position (excluding the character at the specified position) of the character string. If the specified position is a negative number, calculate the position in the descending order. For example, obtain the character string after the second character of **abcde**, that is, **cde**.  
Expression: `StringUtils.substring(value,2)`
46. If the field is of the string type, obtain the substring within the specified range of the character string. If the specified range is a negative number, calculate the range in the descending order. For example, obtain the character string between the second and fifth characters of **abcde**, that is, **cd**.

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

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

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

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

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

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

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

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

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

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

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

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

Expression: `StringUtils.trim(value)`

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

Expression: `NumberUtils.toByte(value)`

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

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

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

Expression: `NumberUtils.toDouble(value)`

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

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

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

Expression: `NumberUtils.toFloat(value)`

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

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

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

- Expression: `NumberUtils.toInt(value)`
60. Convert the character string to a value of the int type. If the conversion fails, the specified value, for example, **1**, is returned.
- Expression: `NumberUtils.toInt(value, 1)`
61. Convert the character string to a value of the long type. If the conversion fails, **0** is returned.
- Expression: `NumberUtils.toLong(value)`
62. Convert the character string to a value of the long type. If the conversion fails, the specified value, for example, **1L**, is returned.
- Expression: `NumberUtils.toLong(value, 1L)`
63. Convert the character string to a value of the short type. If the conversion fails, **0** is returned.
- Expression: `NumberUtils.toShort(value)`
64. Convert the character string to a value of the short type. If the conversion fails, the specified value, for example, **1**, is returned.
- Expression: `NumberUtils.toShort(value, 1)`
65. Convert the IP string to a value of the long type, for example, convert **10.78.124.0** to **172915712**.
- Expression: `CommonUtils.ipToLong(value)`
66. Read an IP address and physical address mapping file from the network, and download the mapping file to the map collection. *url* indicates the address for storing the IP mapping file, for example, **http://10.114.205.45:21203/sqoop/IpList.csv**.
- Expression: `HttpsUtils.downloadMap("url")`
67. Cache the IP address and physical address mappings and specify a key for retrieval, for example, **ipList**.
- Expression: `CommonUtils.setCache("ipList", HttpsUtils.downloadMap("url"))`
68. Obtain the cached IP address and physical address mappings.
- Expression: `CommonUtils.getCache("ipList")`
69. Check whether the IP address and physical address mappings are cached.
- Expression: `CommonUtils.cacheExists("ipList")`
70. Based on the specified offset type (month/day/hour/minute/second) and offset (positive number indicates increase and negative number indicates decrease), convert the time in the specified format to a new time, for example, add 8 hours to **2019-05-21 12:00:00**.
- Expression: `DateUtils.getCurrentTimeByZone("yyyy-MM-dd HH:mm:ss", value, "hour", 8)`

### 4.3.2.3 What Component Versions Are Recommended for Migrating Hadoop Data Sources?

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

**Table 4-3** Recommended component versions

Hadoop Type	Component	Description
MRS/Apache/ FusionInsight HD	Hive	2.x versions are not supported. The following versions are recommended: <ul style="list-style-type: none"> <li>• 1.2.X</li> <li>• 3.1.X</li> </ul>
	HDFS	Recommended versions: <ul style="list-style-type: none"> <li>• 2.8.X</li> <li>• 3.1.X</li> </ul>
	HBase	Recommended versions: <ul style="list-style-type: none"> <li>• 2.1.X</li> <li>• 1.3.X</li> </ul>

#### 4.3.2.4 What Data Formats Are Supported When the Data Source Is Hive?

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

#### 4.3.2.5 Can I Synchronize Jobs to Other Clusters?

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

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

#### 4.3.2.6 Can I Create Jobs in Batches?

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

1. Create a job manually.
2. Export the job and save the job's JSON file to a local PC.
3. Edit the JSON file and replicate more jobs in the JSON file according to the job configuration.
4. Import the JSON file to the CDM cluster to implement batch job creation.

### 4.3.2.7 Can I Schedule Jobs in Batches?

Yes.

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

### 4.3.2.8 How Do I Back Up CDM Jobs?

Yes. If you do not need to use the CDM cluster for a long time, you can stop or delete it to reduce costs.

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

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

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

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

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

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

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

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

Prepare the following data in advance:

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

On the **Cluster Management** page, click the CDM cluster name to view the cluster ID, for example, **c110beff-0f11-4e75-8b10-da7cd882b0ef**.

3. Create a MySQL database and a DWS database, and create tables **city1** and **city2**. The statements for creating tables are as follows:

MySQL:

```
create table city1(code varchar(10),name varchar(32));  
insert into city1 values('NY','New York');
```

DWS:

```
create table city2(code varchar(10),name varchar(32));
```

4. In the CDM cluster, create a link to MySQL, such as a link named **mysqltestlink**. Create a link to DWS, such as a link named **dwstestlink**.
5. Run the following code. You are advised to use the HttpClient package of version 4.5. Maven configuration is as follows:

```
<project>
<modelVersion>4.0.0</modelVersion>
<groupId>cdm</groupId>
<artifactId>cdm-client</artifactId>
<version>1</version>
<dependencies>
<dependency>
<groupId>org.apache.httpcomponents</groupId>
<artifactId>httpclient</artifactId>
<version>4.5</version>
</dependency>
</dependencies>
</project>
```

## Sample Code

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

```
package cdmclient;
import java.io.IOException;
import org.apache.http.Header;
import org.apache.http.HttpEntity;
import org.apache.http.HttpHost;
import org.apache.http.auth.AuthScope;
import org.apache.http.auth.UsernamePasswordCredentials;
import org.apache.http.client.CredentialsProvider;
import org.apache.http.client.config.RequestConfig;
import org.apache.http.client.methods.CloseableHttpResponse;
import org.apache.http.client.methods.HttpDelete;
import org.apache.http.client.methods.HttpGet;
import org.apache.http.client.methods.HttpPost;
import org.apache.http.client.methods.HttpPut;
import org.apache.http.entity.StringEntity;
import org.apache.http.impl.client.BasicCredentialsProvider;
import org.apache.http.impl.client.CloseableHttpClient;
import org.apache.http.impl.client.HttpClients;
import org.apache.http.util.EntityUtils;
public class CdmClient {
private final static String DOMAIN_NAME=" account name";
private final static String USER_NAME=" username";
Private final static String USER_PASSWORD= "Password of the cloud user";
private final static String PROJECT_ID=" Project ID";
private final static String CLUSTER_ID=" CDM cluster ID";
private final static String JOB_NAME=" Job name";
private final static String FROM_LINKNAME=" Source link name";
private final static String TO_LINKNAME=" Destination link name";
Private final static String IAM_ENDPOINT= "IAM endpoint";
Private final static String CDM_ENDPOINT= "CDM endpoint";
private CloseableHttpClient httpClient;
private String token;

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



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

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

/*Create a job.*/

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

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

    String json =
        "{\r\n"+
        "\  \"jobs\": [\r\n"+
        "\    {\r\n"+
        "\      \"from-connector-name\": \"generic-jdbc-connector\",\r\n"+
        "\      \"name\": \""+JOB_NAME+"\",\r\n"+
        "\      \"to-connector-name\": \"generic-jdbc-connector\",

```

```

"\driver-config-values\": {\r\n"+
"\configs\": [\r\n"+
"{\r\n"+
"\"inputs\": [\r\n"+
"{\r\n"+
"\"name\": \"throttlingConfig.numExtractors\", \r\n"+
"\"value\": \"1\" \r\n"+
"}\r\n"+
"], \r\n"+
"\"validators\": [], \r\n"+
"\"type\": \"JOB\", \r\n"+
"\"id\": 30, \r\n"+
"\"name\": \"throttlingConfig\" \r\n"+
"}\r\n"+
"]\r\n"+
"}, \r\n"+
"\"from-link-name\": \"\"+FROM_LINKNAME+\"\", \r\n"+
"\"from-config-values\": {\r\n"+
"\configs\": [\r\n"+
"{\r\n"+
"\"inputs\": [\r\n"+
"{\r\n"+
"\"name\": \"fromJobConfig.schemaName\", \r\n"+
"\"value\": \"sqoop\" \r\n"+
"}, \r\n"+
"{\r\n"+
"\"name\": \"fromJobConfig.tableName\", \r\n"+
"\"value\": \"city1\" \r\n"+
"}, \r\n"+
"{\r\n"+
"\"name\": \"fromJobConfig.columnList\", \r\n"+
"\"value\": \"code&name\" \r\n"+
"}, \r\n"+
"{\r\n"+
"\"name\": \"fromJobConfig.partitionColumn\", \r\n"+
"\"value\": \"code\" \r\n"+
"}\r\n"+
"], \r\n"+
"\"validators\": [], \r\n"+
"\"type\": \"JOB\", \r\n"+
"\"id\": 7, \r\n"+
"\"name\": \"fromJobConfig\" \r\n"+
"}\r\n"+
"]\r\n"+
"}, \r\n"+
"\"to-link-name\": \"\"+TO_LINKNAME+\"\", \r\n"+
"\"to-config-values\": {\r\n"+
"\configs\": [\r\n"+
"{\r\n"+
"\"inputs\": [\r\n"+
"{\r\n"+
"\"name\": \"toJobConfig.schemaName\", \r\n"+
"\"value\": \"sqoop\" \r\n"+
"}, \r\n"+
"{\r\n"+
"\"name\": \"toJobConfig.tableName\", \r\n"+
"\"value\": \"city2\" \r\n"+
"}, \r\n"+
"{\r\n"+
"\"name\": \"toJobConfig.columnList\", \r\n"+
"\"value\": \"code&name\" \r\n"+
"}, \r\n"+

```

```
{\r\n"+
  "\name\": \toJobConfig.shouldClearTable\", \r\n"+
  "\value\": \"true\" \r\n"+
  }\r\n"+
  ], \r\n"+
  "\validators\": [], \r\n"+
  "\type\": \"JOB\", \r\n"+
  "\id\": 9, \r\n"+
  "\name\": \toJobConfig\" \r\n"+
  }\r\n"+
  ] \r\n"+
  } \r\n"+
  } \r\n"+
  } \r\n"+
  } \r\n";
try {
StringEntity s = new StringEntity(json);
s.setContentEncoding("UTF-8");
s.setContentType("application/json");
httpPost.setEntity(s);
httpPost.addHeader("X-Auth-Token", this.token);
httpPost.addHeader("X-Language", "en-us");
CloseableHttpResponse response = httpClient.execute(httpPost);
int status = response.getStatusLine().getStatusCode();
if(status == 200){
System.out.println("Create job successful.");
}else{
System.out.println("Create job failed.");
HttpEntity entity = response.getEntity();
System.out.println(EntityUtils.toString(entity));
}
} catch (Exception e) {
e.printStackTrace();
throw new RuntimeException("Create job failed.", e);
}
}
/*Start the job.*/

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

```
}
}
/*Query the job running status cyclically until the job is complete.*/

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

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

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

```
/*Close the process.*/  
  
public void close(){  
    try {  
        httpClient.close();  
    } catch (IOException e) {  
        throw new RuntimeException("Close failed.", e);  
    }  
}  
  
public static void main(String[] args){  
    CdmClient cdmClient = new CdmClient();  
    cdmClient.createJob();  
    cdmClient.startJob();  
    cdmClient.getJobStatus();  
    cdmClient.deleteJob();  
    cdmClient.close();  
}  
}
```

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

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

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

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

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

---

#### NOTICE

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

---

### Scenario

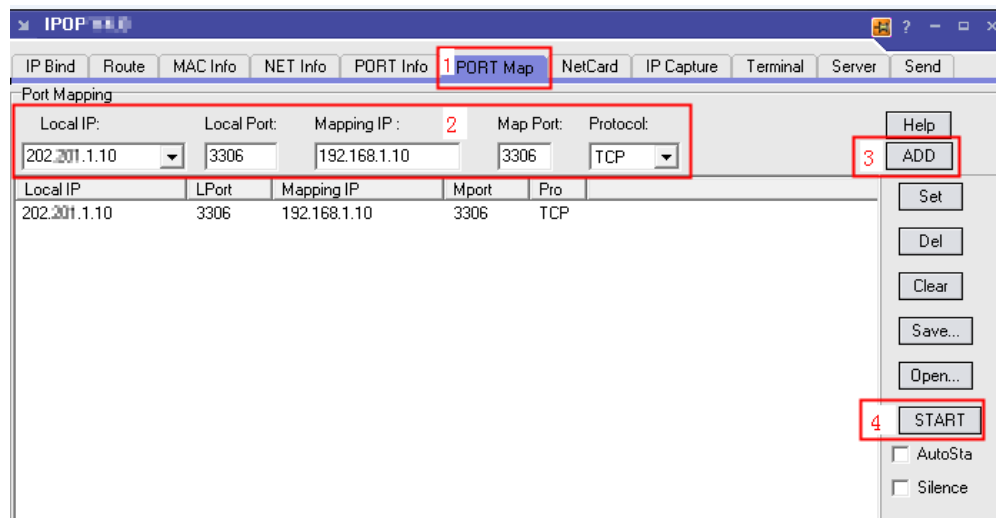
Suppose that the MySQL database on the intranet is migrated to DWS.

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

## Procedure

- Step 1** Use a Windows computer as the gateway. Configure both the intranet and Internet IP addresses on the computer. Conduct the following test to check whether the gateway computer can fulfill service needs.
1. Run the **ping** command on the computer to check whether the intranet address of the MySQL database is pingable. For example, run **ping 192.168.1.8**.
  2. Run the **ping** command on another computer that can access the Internet to check whether the public network address of the gateway computer is pingable. For example, run **ping 202.xx.xx.10**.
- Step 2** Download the port mapping tool IPOP and install it on the gateway computer.
- Step 3** Run the port mapping tool and select **PORT Map**. See [Figure 4-4](#).
- **Local IP** and **Local Port**: Configure these two parameters to the public network address and port number of the gateway computer respectively, which must be entered when creating MySQL links on CDM.
  - **Mapping IP** and **Map Port**: Configure these two parameters to the IP address and port number of the MySQL database on the intranet.

**Figure 4-4** Configuring port mapping



- Step 4** Click **ADD** to add a port mapping relationship.
- Step 5** Click **START** to start mapping and receive data packets.

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

 NOTE

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

----End

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

The number of concurrent extractors in a CDM migration job is related to the cluster specifications and table size. The value range is 1 to 300. If the value is too large, the extractors are queued.

You are advised to set 4 concurrent extractors for each 1 CU (1 CU = 1 vCPU and 4 GB), as listed in [Table 4-4](#). You can also adjust the value as needed. If each row of the table contains less than or equal to 1 MB data, you can extract data concurrently. If each row contains more than 1 MB data, you are advised to extract data in a single thread.

 NOTE

- 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.
- The number of concurrent extractors of a job is affected by **Maximum Concurrent Extractors** configured on the **Settings** page. The **Maximum Concurrent Extractors** parameter specifies the total number of concurrent extractions.

**Table 4-4** Reference configurations of concurrent extractors

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

#### 4.3.2.13 Does CDM Support Real-Time Migration of Dynamic Data?

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

## 4.3.3 Troubleshooting

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

#### Symptom

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

#### Possible Cause

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

#### Solution

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

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

#### Symptom

When CDM is used to migrate Oracle data to DWS, an error is reported, as shown in [Figure 4-5](#).

Figure 4-5 Symptom

```

665 2020-09-21 22:51:02,991 ERROR LocalJobRunner Map Task #3 [org.apache.sqoop.common.SqoopException:111] SqoopException
666 java.sql.SQLException: ORA-01555: snapshot too old: rollback segment number 3 with name "_SYSMU3_2097677531$" too small
667
668     at oracle.jdbc.driver.T4CTTIoer11.processERROR(T4CTTIoer11.java:494)
669     at oracle.jdbc.driver.T4CTTIoer11.processERROR(T4CTTIoer11.java:446)
670     at oracle.jdbc.driver.T4C8Oall.processERROR(T4C8Oall.java:1054)
671     at oracle.jdbc.driver.T4CTTIifun.receive(T4CTTIifun.java:623)
672     at oracle.jdbc.driver.T4CTTIifun.doRPC(T4CTTIifun.java:252)
673     at oracle.jdbc.driver.T4C8Oall.doOALL(T4C8Oall.java:612)
674     at oracle.jdbc.driver.T4CPreparedStatement.doOall8(T4CPreparedStatement.java:226)
675     at oracle.jdbc.driver.T4CPreparedStatement.fetch(T4CPreparedStatement.java:1023)
676     at oracle.jdbc.driver.OracleStatement.fetchMoreRows(OracleStatement.java:1353)
677     at oracle.jdbc.driver.InsensitiveScrollableResultSet.fetchMoreRows(InsensitiveScrollableResultSet.java:736)
678     at oracle.jdbc.driver.InsensitiveScrollableResultSet.absoluteInternal(InsensitiveScrollableResultSet.java:692)
679     at oracle.jdbc.driver.InsensitiveScrollableResultSet.next(InsensitiveScrollableResultSet.java:406)
680     at org.apache.sqoop.connector.jdbc.sql.impl.WrapResultSet.next(WrapResultSet.java:36)
681     at org.apache.sqoop.connector.jdbc.GenericJdbcExtractor.extractObjectRecord(GenericJdbcExtractor.java:151)
682     at org.apache.sqoop.connector.jdbc.GenericJdbcExtractor.extract(GenericJdbcExtractor.java:129)
683     at org.apache.sqoop.connector.jdbc.GenericJdbcExtractor.extract(GenericJdbcExtractor.java:59)
684     at org.apache.sqoop.job.map.SqoopMapper.runInternal(SqoopMapper.java:184)
685     at org.apache.sqoop.job.map.SqoopMapper.run(SqoopMapper.java:81)
686     at org.apache.hadoop.mapred.MapTask.runNewMapper(MapTask.java:799)
687     at org.apache.hadoop.mapred.MapTask.run(MapTask.java)
688     at org.apache.hadoop.mapred.LocalJobRunner$JobMapTaskRunnable.run(LocalJobRunner.java:271)
689     at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
690     at java.util.concurrent.FutureTask.run(FutureTask.java:266)
691     at org.apache.sqoop.submission.mapreduce.MapperExecutorGroup$1.lambda$execute$0(MapperExecutorGroup.java:222)
692     at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
693     at java.util.concurrent.FutureTask.run(FutureTask.java:266)
694     at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
695     at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
696     at java.lang.Thread.run(Thread.java:748)
697 Caused by: oracle.jdbc.OracleDatabaseException: ORA-01555: snapshot too old: rollback segment number 3 with name "_SYSMU3_2097677531$" too small
698
699     at oracle.jdbc.driver.T4CTTIoer11.processERROR(T4CTTIoer11.java:498)
700     ... 28 common frames omitted

```



## Cause Analysis

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

## Summary and Suggestions

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

### 4.3.3.3 What Should I Do If the MongoDB Connection Migration Fails?

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

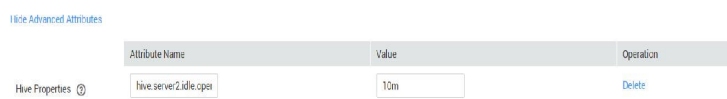
If the MongoDB connection fails to be migrated, you need to view the user permission information in the MongoDB connection to ensure that the user has the read and write permissions on the specified database.

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

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

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

In the figure on the left:



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

#### Symptom

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

#### Possible Cause

The possible cause is that the type of the source table does not match that of the target table. For example, the **dli** field of the source is of the string type, and the

**dws** field of the destination is of the varchar(50) type. As a result, the precision is default and the error message "value too long for type character varying" is reported. This issue also occurs for conversion from string to bigint and from bigint to int.

## Solution

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

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

## Symptom

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

## Possible Cause

The table has a large data volume, and the source end uses the where statement to filter data. However, the column is not an index column or the column values are not discrete. As a result, the entire table is scanned during the query, causing a JDBC connection timeout. As shown in [Figure 4-6](#), the **c\_date** field is not an index column.

**Figure 4-6** Non-index column

The image shows two side-by-side configuration panels for a data migration job. The left panel is titled "Source Job Configuration" and includes fields for "Source Link Name" (mysql), "Use SQL Statement" (Yes/No), "Schema/Table Space" (SQOOP), "Table Name" (rf\_BaoWeiFu\_test\_sql\_To), "Where Clause" (c\_date > '2021-02-27 10:43:04.123'), "Partition Column", and "Partition column nullable" (Yes/No). The right panel is titled "Destination Job Configuration" and includes fields for "Destination Link Name" (dlj), "Resource Queue" (dlf\_notdelete), "Database" (abcd), "Table" (dddd), and "Clear data before import" (Yes/No).

## Solution

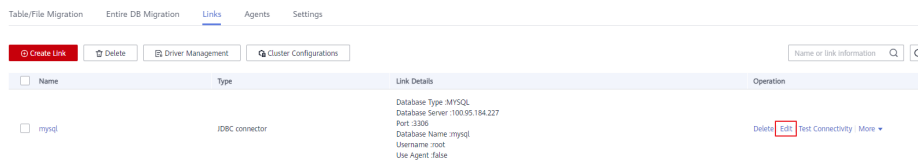
1. Contact the DBA to modify the table structure, set the columns to be filtered as index columns, and try again.  
If the failure persists because the data is not discrete, perform [2](#) to [4](#) and increase the JDBC timeout duration.
2. Locate the MySQL link name based on the job and obtain the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

**Figure 4-7** Link information



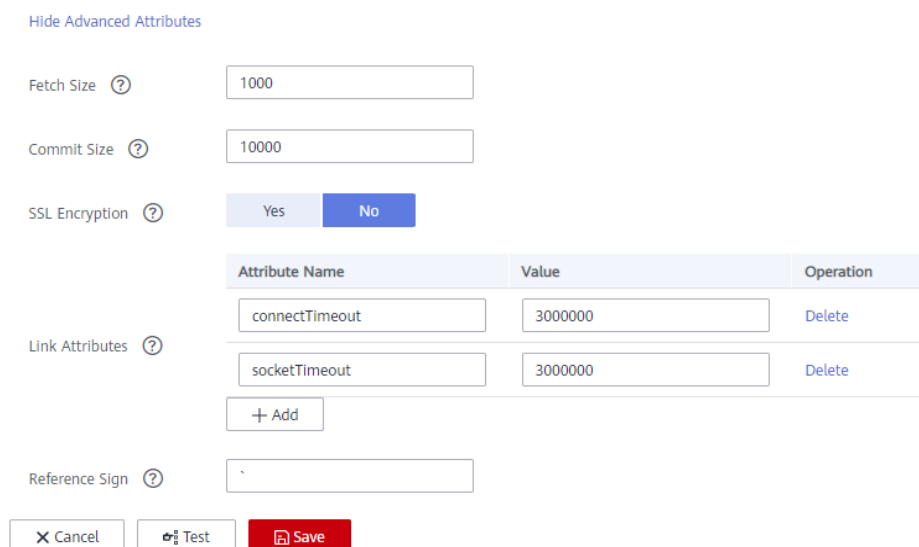
3. Click the **Links** tab and click **Edit** to edit the link.

**Figure 4-8** Editing the link



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

**Figure 4-9** Editing advanced attributes



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

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

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

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

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

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

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

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

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

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

Manually clear the browser cache to avoid this error.

#### 4.3.3.12 What Should I Do If Message "CORE\_0031:Connect time out. (Cdm. 0523)" Is Displayed During the Creation of an MRS Hive Link?

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

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

##### NOTE

- If the CDM cluster version is 2.9.0 or later and the MRS cluster version is 3.1.0 or later, the created user must have the permissions of the **Manager\_viewer** role to create links on CDM. To perform operations on databases, tables, and data of a component, you also need to add the user group permissions of the component to the user.
- 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.

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

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

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

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

This may be because you have uploaded the latest ORACLE\_8 driver (for example, Oracle Database 21c (21.3) driver), which is not supported yet. You are advised to use the ojdbc8.jar driver in Oracle Database 12c. You can download it from <https://www.oracle.com/database/technologies/jdbc-ucp-122-downloads.html>.

## 4.4 DataArts Factory

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

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

In addition, the system allows you to adjust the maximum quota as required. If you have any requirements, .

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

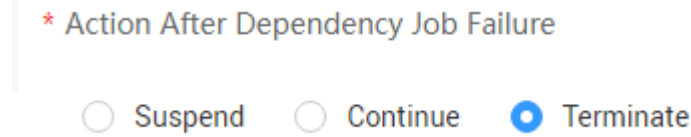
On the **Running History** page, there is a large difference between **Job Execution Time** and **Start Time**, as shown in the figure below. **Job Execution Time** is the time when the job is expected to be executed. **Start Time** is the time when the job starts to be executed.

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

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

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

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

**Figure 4-10** Job dependencies

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

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

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

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

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

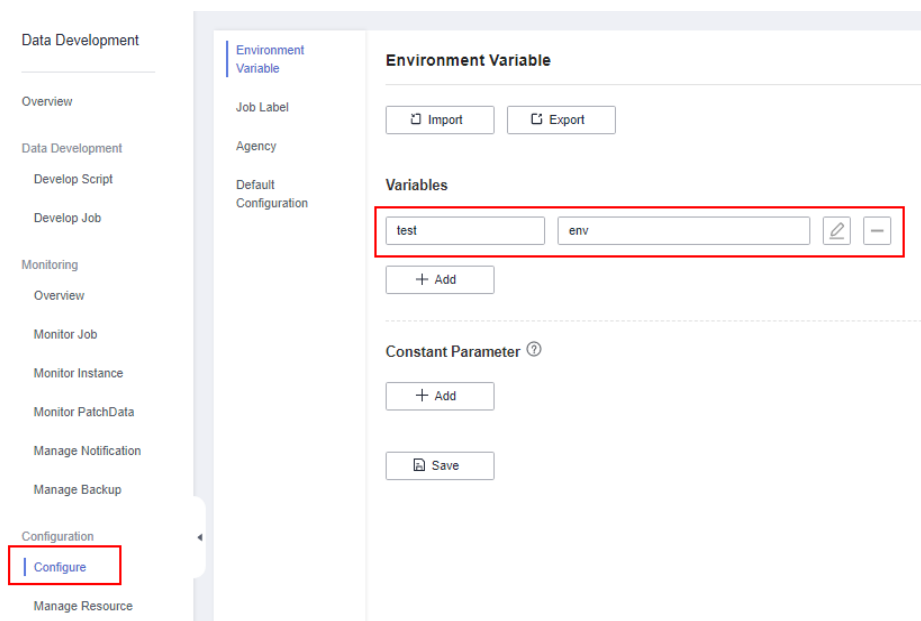
#### 4.4.5 What Are the Differences and Connections Among Environment Variables, Job Parameters, and Script Parameters?

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

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

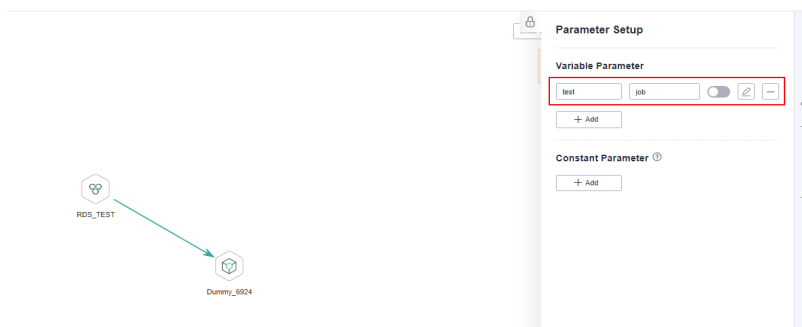
- Variables and constants can be defined in environment variables. Environment variables take effect in current workspace.
  - The value of a variable (such as **workspace name**) varies depending on the workspace. When exporting a variable from a workspace and import it to another workspace, you must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

**Figure 4-11** Environment variable



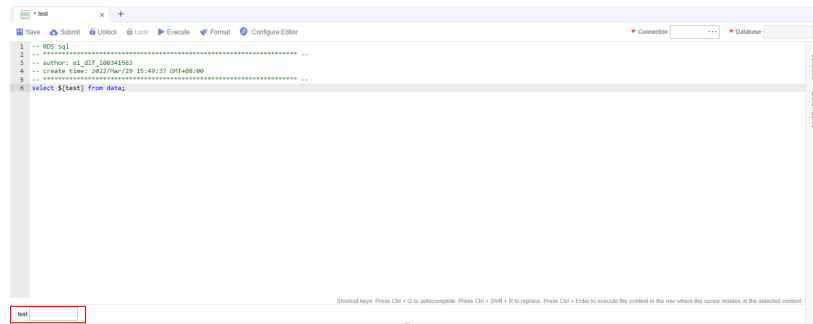
- Parameters and constants can be defined in job parameters. Job parameters take effect in current job.
  - The value of a parameter varies depending on jobs. When exporting a parameter from a workspace and import it to another workspace, you must reconfigure its value.
  - The value of a constant in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

**Figure 4-12** Job parameter.

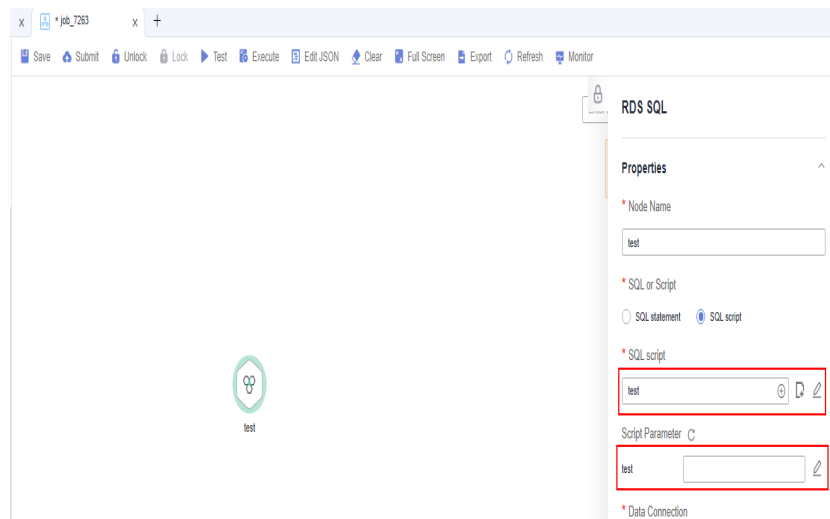


- Script parameters take effect in current script and it can be used in the following ways.
  - Enter SQL script parameters in the script editor (Flink SQL is not supported). If the script is executed independently, you can configure the parameters in the lower part of the editor, as shown in [Figure 4-13](#). If the script is executed by job scheduling, you can assign values to the parameters based on node attributes, as shown in [Figure 4-14](#).
  - For Shell scripts, you can enter a parameter and an interactive parameter in the upper part of the editor to transfer the parameters.
  - Python scripts do not support parameter transfer.

**Figure 4-13** Configuring script parameters when the script is executed independently



**Figure 4-14** Configuring script parameters when the script is executed by job scheduling



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

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

**NOTE**

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

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

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

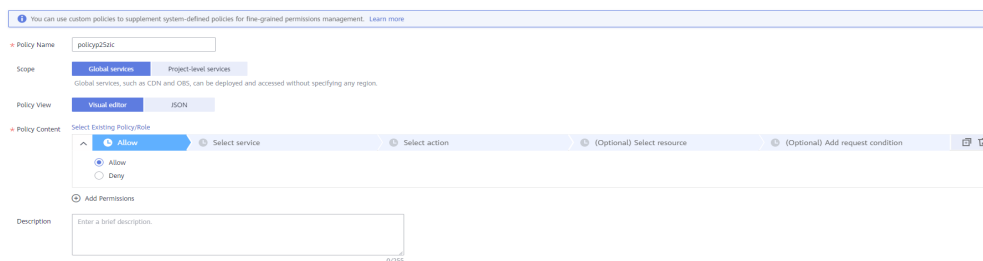
Policy doesn't allow iam:agencies:listAgencies to be performed.

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



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

- Step 1** Log in to the management console.
- Step 2** On the management console, hover the mouse pointer over the username in the upper right corner, and choose **Identity and Access Management** from the drop-down list.
- Step 3** In the navigation pane, choose **Permissions**. Then, click **Create Custom Policy**.
- Step 4** Enter a policy name.



- Step 5** Set **Scope** to **Global services**. The scope you set is where the custom policy takes effect. In this example, the custom policy has the permissions required to view the agency lists based on specified conditions.
- Step 6** Set **Policy View** to **Visual editor**.
- Step 7** Configure a policy in **Policy Content**.
  1. Select **Allow**.
  2. Select **Identity and Access Management (IAM)** for **Select service**.
  3. Select **iam:agencies:listAgencies** for **Select action**.
- Step 8** Click **OK**.
- Step 9** Add the policy defined in **Step 7** to the group to which the current user belongs. For details, see .

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

----End

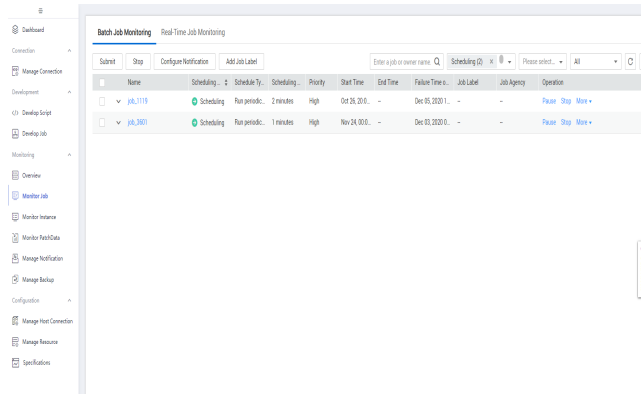
## 4.4.8 How Do I Locate Job Scheduling Nodes with a Large Number?

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

1. In the left navigation tree of Data Development, choose **Monitoring** > **Monitor Instance**, select the current day, and view the jobs that are frequently scheduled.
2. In the left navigation tree of Data Development, choose **Monitoring** > **Monitor Job** to check whether the scheduling period of jobs that are

frequently scheduled is set properly. If the scheduling period is inappropriate, adjust the scheduling period or stop the scheduling. Generally, the number of minute-level scheduling jobs executed every day exceeds the upper limit.

**Figure 4-15** Viewing the scheduling period



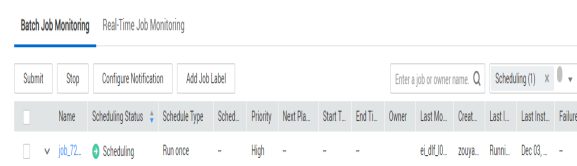
#### 4.4.9 Why Cannot Specified Peripheral Resources Be Selected When a Data Connection Is Created in Data Development?

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

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

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

**Figure 4-16** Viewing the job scheduling status



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

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

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

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

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

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

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

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

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

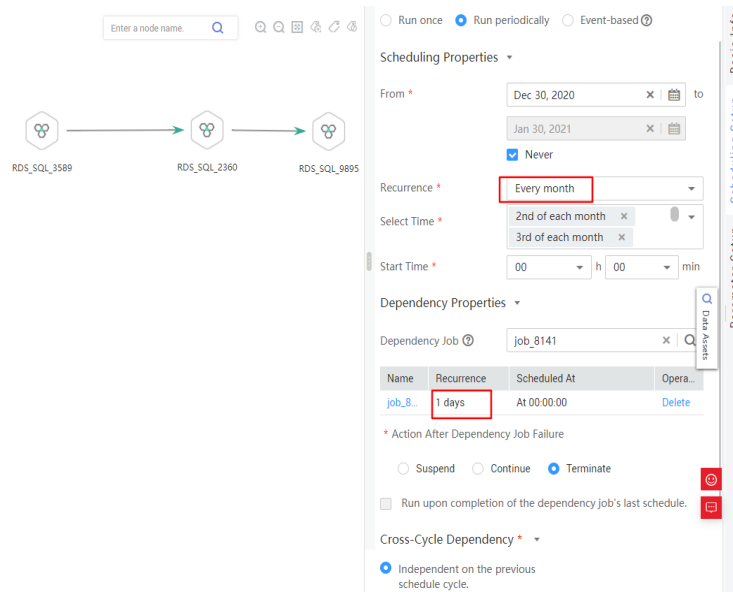
Method 1: After the node test is complete, right-click the current node and choose **View Log** from the shortcut menu.

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

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

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

**Figure 4-17** Viewing the job scheduling period and dependency attributes



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

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

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

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

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

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

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

##### Symptom

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

##### Cause Analysis

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

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

## Solutions

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

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

#### Symptom

There is no content contained in the job log.

#### Cause Analysis

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

## Solutions

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

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

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

#### Symptom

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

#### Cause Analysis

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

## Solutions

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

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

### Symptom

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

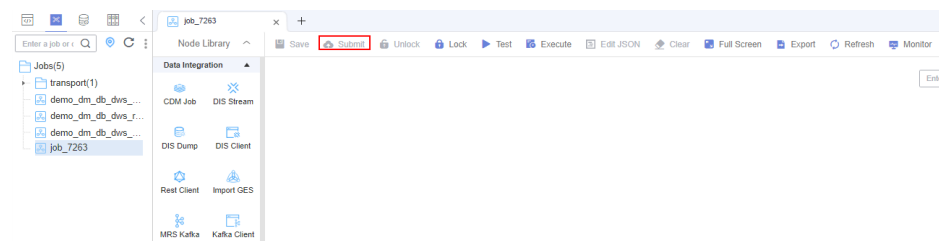
### Cause Analysis

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

### Solutions

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

Figure 4-18 Submitting a version



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

### Symptom

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

### Cause Analysis

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

### Solutions

1. Step 1: Switch to the script development page and find the corresponding script.

2. Step 2: Submit the script version.
3. Step 3: Schedule the job.

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

##### Symptom

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

##### Cause Analysis

The upstream dependency job is not in the running state.

##### Solutions

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

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

Databases and data tables can be created in DLI.

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

#### 4.4.24 Why Is No Result Displayed After an HIVE Task Is Executed?

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

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

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

#### 4.4.26 How Do I Create a Notification for All Jobs?

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

**Figure 4-19** Creating a notification

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

### 4.4.27 How Many Nodes Can Be Executed Concurrently in Each DataArts Studio Version?

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

**Table 4-5** Number of nodes that can be executed concurrently in each DataArts Studio version

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

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

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

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

- After an agency is configured for a job, the job interacts with other services through the agency, preventing job execution failures caused by permission issues. There are two types of agencies, workspace agencies and job agencies. Workspace agencies have a higher priority than job agencies.



- Workspace agency: applies to all the jobs in a workspace. You can choose **Configuration > Configure > Agency** to configure a workspace agency.
- Job agency: applies to a single job. You can configure a job agency in the basic information of a job.
- After an execution user is configured, the user will be used to start the job. You can configure an execution user in the basic information of a job.