

DataArts Studio

User Guide

Issue 01
Date 2022-09-30



Copyright © Huawei Technologies Co., Ltd. 2022. All rights reserved.

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

Trademarks and Permissions



HUAWEI and other Huawei trademarks are trademarks of Huawei Technologies Co., Ltd.

All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

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

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

Contents

1 DataArts Studio Introduction.....	1
2 Preparations Before Using DataArts Studio.....	5
3 Management Center.....	7
3.1 Data Sources.....	7
3.2 Creating Data Connections.....	12
3.3 Migrating Resources.....	39
3.4 Tutorials.....	42
3.4.1 Creating an MRS Hive Connection.....	43
3.4.2 Creating a DWS Connection.....	48
3.4.3 Creating a MySQL Connection.....	53
4 DataArts Migration.....	58
4.1 Overview.....	58
4.2 Constraints.....	60
4.3 Supported Data Sources.....	65
4.4 Managing Clusters.....	90
4.4.1 Creating a CDM Cluster.....	90
4.4.2 Binding or Unbinding an EIP.....	90
4.4.3 Restarting a Cluster.....	91
4.4.4 Deleting a Cluster.....	92
4.4.5 Downloading Cluster Logs.....	94
4.4.6 Viewing Basic Cluster Information and Modifying Cluster Configurations.....	95
4.4.7 Viewing Metrics.....	98
4.4.7.1 CDM Metrics.....	98
4.4.7.2 Configuring Alarm Rules.....	101
4.4.7.3 Querying Metrics.....	101
4.5 Managing Links.....	102
4.5.1 Creating Links.....	102
4.5.2 Managing Drivers.....	107
4.5.3 Managing Agents.....	109
4.5.4 Managing Cluster Configurations.....	113
4.5.5 Link to a Common Relational Database.....	119
4.5.6 Link to a Database Shard.....	121

4.5.7 Link to MyCAT.....	123
4.5.8 Link to a Dameng Database.....	125
4.5.9 Link to an RDS for MySQL/MySQL Database.....	126
4.5.10 Link to an Oracle Database.....	129
4.5.11 Link to DLI.....	130
4.5.12 Link to Hive.....	132
4.5.13 Link to HBase.....	139
4.5.14 Link to HDFS.....	145
4.5.15 Link to OBS.....	152
4.5.16 Link to an FTP or SFTP Server.....	153
4.5.17 Link to Redis/DCS.....	154
4.5.18 Link to DDS.....	155
4.5.19 Link to CloudTable.....	155
4.5.20 Link to CloudTable OpenTSDB.....	156
4.5.21 Link to MongoDB.....	158
4.5.22 Link to Cassandra.....	159
4.5.23 Link to Kafka.....	159
4.5.24 Link to DMS Kafka.....	161
4.5.25 Link to Elasticsearch/CSS.....	162
4.6 Managing Jobs.....	163
4.6.1 Table/File Migration Jobs.....	163
4.6.2 Creating an Entire Database Migration Job.....	174
4.6.3 Source Job Parameters.....	182
4.6.3.1 From OBS.....	182
4.6.3.2 From HDFS.....	190
4.6.3.3 From HBase/CloudTable.....	198
4.6.3.4 From Hive.....	200
4.6.3.5 From DLI.....	202
4.6.3.6 From FTP/SFTP.....	203
4.6.3.7 From HTTP.....	209
4.6.3.8 From a Common Relational Database.....	212
4.6.3.9 From MySQL.....	216
4.6.3.10 From Oracle.....	221
4.6.3.11 From a Database Shard.....	225
4.6.3.12 From MongoDB/DDS.....	228
4.6.3.13 From Redis.....	229
4.6.3.14 From Kafka/DMS Kafka.....	230
4.6.3.15 From Elasticsearch or CSS.....	231
4.6.3.16 From OpenTSDB.....	233
4.6.4 Destination Job Parameters.....	234
4.6.4.1 To OBS.....	234
4.6.4.2 To HDFS.....	240

4.6.4.3 To HBase/CloudTable.....	244
4.6.4.4 To Hive.....	245
4.6.4.5 To a Common Relational Database.....	247
4.6.4.6 To DWS.....	251
4.6.4.7 To DDS.....	255
4.6.4.8 To DCS.....	256
4.6.4.9 To CSS.....	256
4.6.4.10 To DLI.....	258
4.6.4.11 To OpenTSDB.....	259
4.6.5 Scheduling Job Execution.....	260
4.6.6 Job Configuration Management.....	263
4.6.7 Managing a Single Job.....	267
4.6.8 Managing Jobs in Batches.....	269
4.7 Auditing.....	270
4.7.1 Key CDM Operations Recorded by CTS.....	271
4.7.2 Viewing Traces.....	271
4.8 Performance Reference.....	272
4.8.1 Factors Affecting Performance.....	272
4.8.2 Performance Tuning.....	275
4.8.3 Reference: Job Splitting Dimensions.....	277
4.8.4 Reference: CDM Performance Test Data.....	280
4.9 Tutorials.....	282
4.9.1 Creating an MRS Hive Link.....	282
4.9.2 Creating a MySQL Link.....	287
4.9.3 Migrating Data from MySQL to MRS Hive.....	291
4.9.4 Migrating Data from MySQL to OBS.....	302
4.9.5 Migrating Data from MySQL to DWS.....	309
4.9.6 Migrating an Entire MySQL Database to RDS.....	318
4.9.7 Migrating Data from Oracle to CSS.....	324
4.9.8 Migrating Data from Oracle to DWS.....	330
4.9.9 Migrating Data from OBS to CSS.....	337
4.9.10 Migrating Data from OBS to DLI.....	343
4.9.11 Migrating Data from MRS HDFS to OBS.....	349
4.9.12 Migrating the Entire Elasticsearch Database to CSS.....	354
4.9.13 Migrating Data from DDS to DWS.....	358
4.9.14 More Cases and Practices.....	364
4.10 Advanced Operations.....	364
4.10.1 Incremental Migration.....	364
4.10.1.1 Incremental File Migration.....	364
4.10.1.2 Incremental Migration of Relational Databases.....	367
4.10.1.3 HBase/CloudTable Incremental Migration.....	369
4.10.2 Using Macro Variables of Date and Time.....	370

4.10.3 Migration in Transaction Mode.....	374
4.10.4 Encryption and Decryption During File Migration.....	375
4.10.5 MD5 Verification.....	377
4.10.6 Field Conversion.....	378
4.10.7 Migrating Files with Specified Names.....	386
4.10.8 Regular Expressions for Separating Semi-structured Text.....	386
4.10.9 Recording the Time When Data Is Written to the Database.....	390
4.10.10 File Formats.....	393
5 DataArts Architecture.....	403
5.1 Overview.....	403
5.2 DataArts Architecture Use Process.....	407
5.3 Preparations.....	409
5.3.1 Adding Reviewers.....	411
5.3.2 Configuration Center.....	413
5.4 Data Survey.....	424
5.4.1 Designing Processes.....	424
5.4.2 Designing Subjects.....	429
5.5 Standards Design.....	435
5.5.1 Creating Lookup Tables.....	435
5.5.2 Creating Data Standards.....	446
5.6 Model Design.....	455
5.6.1 ER Modeling.....	455
5.6.1.1 Designing Logical Models.....	455
5.6.1.2 Designing Physical Models.....	466
5.6.2 Dimensional Modeling.....	479
5.6.2.1 Creating Dimensions.....	479
5.6.2.2 Managing Dimension Tables.....	488
5.6.2.3 Creating Fact Tables.....	495
5.7 Metric Design.....	506
5.7.1 Business Metrics.....	506
5.7.2 Technical Metrics.....	512
5.7.2.1 Creating Atomic Metrics.....	512
5.7.2.2 Creating Derivative Metrics.....	516
5.7.2.3 Creating Compound Metrics.....	523
5.7.2.4 Creating Time Filters.....	527
5.8 Data Mart Building.....	530
5.8.1 Creating Summary Tables.....	530
5.9 Common Operations.....	541
5.9.1 Reversing a Database (ER Modeling).....	541
5.9.2 Reversing a Database (Dimensional Modeling).....	543
5.9.3 Importing/Exporting Tables.....	545
5.9.4 Associating Quality Rules.....	554

5.9.5 Viewing Tables.....	560
5.9.6 Modifying Subjects, Directories, and Processes.....	562
5.9.7 Review Center.....	564
5.10 Tutorials.....	567
5.10.1 DataArts Architecture Example.....	567
6 DataArts Factory.....	610
6.1 Overview.....	610
6.2 Data Management.....	612
6.2.1 Data Management Process.....	612
6.2.2 Creating a Data Connection.....	613
6.2.3 Creating a Database.....	614
6.2.4 (Optional) Creating a Database Schema.....	616
6.2.5 Creating a Table.....	617
6.3 Script Development.....	624
6.3.1 Script Development Process.....	624
6.3.2 Creating a Script.....	625
6.3.3 Developing Scripts.....	627
6.3.3.1 Developing an SQL Script.....	627
6.3.3.2 Developing a Shell Script.....	632
6.3.3.3 Developing a Python Script.....	636
6.3.4 Submitting a Version and Unlocking the Script.....	638
6.3.5 (Optional) Managing Scripts.....	642
6.3.5.1 Copying a Script.....	642
6.3.5.2 Copying the Script Name and Renaming a Script.....	643
6.3.5.3 Moving a Script or Script Directory.....	645
6.3.5.4 Exporting and Importing a Script.....	648
6.3.5.5 Viewing Script References.....	649
6.3.5.6 Deleting a Script.....	650
6.3.5.7 Changing the Script Owner.....	651
6.3.5.8 Unlocking Scripts.....	652
6.4 Job Development.....	653
6.4.1 Job Development Process.....	653
6.4.2 Creating a Job.....	654
6.4.3 Developing a Job.....	658
6.4.4 Setting Up Scheduling for a Job.....	663
6.4.5 Submitting a Version and Unlocking the Script.....	669
6.4.6 (Optional) Managing Jobs.....	674
6.4.6.1 Copying a Job.....	674
6.4.6.2 Copying the Job Name and Renaming a Job.....	675
6.4.6.3 Moving a Job or Job Directory.....	676
6.4.6.4 Exporting and Importing a Job.....	676
6.4.6.5 Configuring Jobs.....	679

6.4.6.6 Deleting a Job.....	684
6.4.6.7 Changing the Job Owner.....	685
6.4.6.8 Unlocking Jobs.....	686
6.5 Solution.....	687
6.6 Execution History.....	689
6.7 O&M and Scheduling.....	690
6.7.1 Overview.....	690
6.7.2 Monitoring a Job.....	690
6.7.2.1 Monitoring a Batch Job.....	690
6.7.2.2 Monitoring a Real-Time Job.....	694
6.7.3 Monitoring an Instance.....	699
6.7.4 Monitoring PatchData.....	702
6.7.5 Managing Notifications.....	703
6.7.5.1 Managing a Notification.....	703
6.7.5.2 Cycle Overview.....	706
6.7.6 Managing Backups.....	708
6.8 Configuration and Management.....	710
6.8.1 Configuring Resources.....	710
6.8.1.1 Configuring Environment Variables.....	710
6.8.1.2 Configuring an OBS Bucket.....	713
6.8.1.3 Managing Job Labels.....	714
6.8.1.4 Configuring Agencies.....	714
6.8.1.5 Configuring a Default Item.....	722
6.8.2 Managing Resources.....	724
6.9 Node Reference.....	727
6.9.1 Node Overview.....	727
6.9.2 Node Lineages.....	727
6.9.2.1 Overview.....	728
6.9.2.2 Configuring Data Lineages.....	729
6.9.2.3 Viewing Data Lineages.....	730
6.9.3 CDM Job.....	733
6.9.4 Rest Client.....	739
6.9.5 Import GES.....	746
6.9.6 MRS Kafka.....	748
6.9.7 Kafka Client.....	750
6.9.8 ROMA FDI Job.....	752
6.9.9 DLI Flink Job.....	754
6.9.10 DLI SQL.....	758
6.9.11 DLI Spark.....	764
6.9.12 DWS SQL.....	771
6.9.13 MRS Spark SQL.....	777
6.9.14 MRS Hive SQL.....	782

6.9.15 MRS Presto SQL.....	787
6.9.16 MRS Spark.....	792
6.9.17 MRS Spark Python.....	797
6.9.18 MRS Flink Job.....	802
6.9.19 MRS MapReduce.....	804
6.9.20 CSS.....	806
6.9.21 Shell.....	808
6.9.22 RDS SQL.....	811
6.9.23 ETL Job.....	812
6.9.24 Python.....	817
6.9.25 Create OBS.....	819
6.9.26 Delete OBS.....	821
6.9.27 OBS Manager.....	823
6.9.28 Open/Close Resource.....	828
6.9.29 Data Quality Monitor.....	830
6.9.30 Subjob.....	831
6.9.31 For Each.....	833
6.9.32 SMN.....	835
6.9.33 Dummy.....	838
6.10 EL Expression Reference.....	839
6.10.1 Expression Overview.....	839
6.10.2 Basic Operators.....	843
6.10.3 Date and Time Mode.....	844
6.10.4 Env Embedded Objects.....	845
6.10.5 Job Embedded Objects.....	846
6.10.6 StringUtil Embedded Objects.....	848
6.10.7 DateUtil Embedded Objects.....	848
6.10.8 JSONUtil Embedded Objects.....	849
6.10.9 Loop Embedded Objects.....	850
6.10.10 OBSUtil Embedded Objects.....	850
6.10.11 Expression Use Example.....	851
6.11 Usage Guidance.....	854
6.11.1 Job Dependency.....	854
6.11.2 IF Statements.....	860
6.11.3 Obtaining the Return Value of a Rest Client Node.....	870
6.11.4 Using For Each Nodes.....	872
6.11.5 Developing a Python Script.....	879
6.11.6 Developing a DWS SQL Job.....	884
6.11.7 Developing a Hive SQL Job.....	887
6.11.8 Developing a DLI Spark Job.....	890
6.11.9 Developing an MRS Flink Job.....	894
6.11.10 Developing an MRS Spark Python Job.....	896

6.11.11 More Cases for Reference.....	902
7 DataArts Quality.....	903
7.1 Monitoring Business Metrics.....	903
7.1.1 Overview.....	903
7.1.2 Creating a Metric.....	904
7.1.3 Creating a Rule.....	905
7.1.4 Creating a Scenario.....	907
7.1.5 Viewing a Scenario Instance.....	909
7.2 Monitoring Data Quality.....	911
7.2.1 Overview.....	911
7.2.2 Creating Rule Templates.....	912
7.2.3 Creating Quality Jobs.....	919
7.2.4 Creating a Comparison Job.....	929
7.2.5 Viewing Job Instances.....	937
7.2.6 Viewing Quality Reports.....	939
7.3 Tutorials.....	943
7.3.1 Creating a Business Scenario.....	943
7.3.2 Creating a Quality Job.....	946
7.3.3 Creating a Comparison Job.....	949
8 DataArts Catalog.....	953
8.1 Data Maps.....	953
8.1.1 Overview.....	953
8.1.2 Overview.....	953
8.1.3 Data Catalogs.....	954
8.1.4 Tags.....	956
8.2 Data Permissions.....	959
8.2.1 Overview.....	959
8.2.2 Data Catalog Permissions.....	959
8.2.3 Data Table Permissions.....	960
8.2.4 Review Center.....	963
8.3 DataArts Security (to Be Brought Offline).....	963
8.3.1 Overview.....	964
8.3.2 Data Security Levels.....	964
8.3.3 Data Classifications.....	965
8.3.4 Masking Policies.....	966
8.4 Metadata Collection.....	968
8.4.1 Overview.....	968
8.4.2 Task Management.....	968
8.4.3 Task Monitoring.....	976
8.5 Tutorials.....	977
8.5.1 Developing an Incremental Metadata Collection Task.....	977
8.5.2 Viewing Data Lineages Through the Data Map.....	981

8.5.2.1 Overview.....	981
8.5.2.2 Configuring Data Lineages.....	982
8.5.2.3 Viewing Data Lineages.....	984
9 DataArts DataService.....	987
9.1 Overview.....	987
9.2 Specifications.....	990
9.3 API Development.....	991
9.3.1 Preparations.....	991
9.3.1.1 an Exclusive DataArts DataService instance.....	991
9.3.1.2 Adding Reviewers.....	996
9.3.2 Creating an API.....	997
9.3.2.1 Generating an API Using Configuration.....	997
9.3.2.2 Generating an API in the Script Mode.....	1005
9.3.2.3 Registering APIs.....	1011
9.3.3 Debugging an API.....	1014
9.3.4 Publishing an API.....	1016
9.3.5 Managing APIs.....	1017
9.3.5.1 Setting an API to Be Visible.....	1017
9.3.5.2 Suspending/Restoring an API.....	1018
9.3.5.3 Unpublishing/Deleting APIs.....	1019
9.3.5.4 Copying an API.....	1020
9.3.5.5 Synchronizing APIs.....	1021
9.3.5.6 Exporting All/Exporting/Importing APIs.....	1022
9.3.6 Creating Throttling Policies.....	1024
9.4 Calling APIs.....	1027
9.5 Performing Operations in Review Center.....	1029
10 Error Codes.....	1031
10.1 DataArts Migration Error Codes.....	1031

1 DataArts Studio Introduction

DataArts Studio is a one-stop data operations platform that provides intelligent data lifecycle management. It supports intelligent construction of industrial knowledge libraries and incorporates data foundations such as big data storage, computing, and analysis engines. With DataArts Studio, your enterprise can easily construct end-to-end intelligent data systems. These systems can help eliminate data silos, unify data standards, accelerate data monetization, and promote digital transformation.

DataArts Studio Users

DataArts Studio provides four preset roles. They have different DataArts Studio permissions. For details, see [DataArts Studio Permissions](#).

- **Admin**

This role is granted with the management, decision-making, and review permissions. Management personnel who are familiar with enterprise businesses can be assigned the admin role. An admin has the permissions of both developers and reviewers. Users who have the admin role can perform any operations in DataArts Studio. For example, they can manage workspaces and data assets, and configure jobs. In DataArts Architecture and DataArts DataService, operations like publishing and suspending data models and APIs must be reviewed by admins to guarantee data quality.
- **Developer**

Data modeling engineers and developers who are familiar with script development can be assigned the developer role. Developers who have the developer role can develop jobs in all DataArts Studio products. They can easily build end-to-end data systems with intelligence and at full speed. Developers can use a software development kit (SDK) to call DataArts DataService APIs provided by DataArts Studio to analyze data after intelligent data systems are built. Developers do not have the permissions required for reviewing operations and managing workspaces and workspace members. But they have most of DataArts Studio permissions.
- **Operator**

This role is granted with the permissions required to view job details, schedule O&M tasks, and monitor resources. O&M personnel can be assigned this role.

DataArts Studio DataArts Catalog visualizes all data links. Data quality can be verified, controlled, and traced. O&M personnel who have the operator role can schedule and monitor jobs from end to end. Data collection, consumption, and O&M are all one-stop services.

- **Viewer**

This role is granted with the read-only permissions.

DataArts Studio Development Process

To use DataArts Studio, perform the following steps:

Table 1-1 DataArts Studio development process

Process	Description	Task	Helpful Link
Preparations	If you access DataArts Studio for the first time, register an account, buy a DataArts Studio instance, create a workspace and a user, authorize DataArts Studio permissions to the user, and add workspace members and roles.	Prepare before you use DataArts Studio.	DataArts Studio Preparations
	Obtain the address of the data source to be connected and ensure that the host where the data source is located can communicate with the platform.	Prepare data sources.	Preparing a Data Source
	Select a cloud service as the data lake. The data lake is used to store both original and real-time data, for the purposes of data development, governance, and operations.	Prepare data lakes.	Preparing a Data Lake
	Select cloud services for data storage, query, and analysis as required. Then, create data connections required for the cloud services.	Create data connections.	Creating Data Connections

Process	Description	Task	Helpful Link
DataArts Migration	<p>Use DataArts Studio to upload data from data sources to the cloud.</p> <p>DataArts Migration migrates data between homogeneous and heterogeneous data sources such as self-built and cloud-based file systems, relational databases, data warehouses, NoSQLs, big data cloud services, and object storage.</p>	DataArts Migration	Supported Data Sources Creating a CDM Cluster Creating Links Table/File Migration Jobs
Metadata Collection	Collect metadata of raw data for data management and monitoring.	Metadata collection	Metadata Collection
DataArts Architecture	<p>Use DataArts Architecture to create entity-relationship (ER) models and dimensional models to standardize and visualize data development and output data governance methods that can guide development personnel to work with ease.</p> <p>In DataArts Architecture, you can create dimensions, fact tables, summary tables, and metrics that fit your needs.</p>	Design the data architecture implementation process.	DataArts Architecture Use Process
		Add reviewers.	Adding a Reviewer
		Design subjects.	Designing Subjects
		Manage lookup tables.	Creating Lookup Tables
		Formulate data standards.	Creating Data Standards
		Create ER models.	ER Modeling
		Create dimensional models.	Dimensional Modeling
DataArts Factory	<p>Use DataArts Factory to manage diverse big data services.</p> <p>The one-stop big data development environment enables a variety of operations such as data management,</p>	Manage data.	Data Management Process
		Develop scripts.	Script Development Process

Process	Description	Task	Helpful Link
	data integration, script development, job development, job scheduling, O&M, and monitoring, facilitating data analysis and processing.	Develop jobs.	Job Development Process
		Perform O&M and scheduling.	Overview
DataArts Quality	Use DataArts Quality to monitor business and technical metrics. Screen out unqualified data in a single column or cross columns, rows, and tables from the following perspectives: integrity, validity, timeliness, consistency, accuracy, and uniqueness. Use the automatically generated quality rules to cleanse and standardize data repeatedly.	Monitor business metrics.	Creating a Metric Creating a Rule Creating a Scenario
		Monitor data quality.	Creating Rule Templates Creating Quality Jobs Creating a Comparison Job
DataArts Catalog	Use DataArts Studio DataArts Catalog to manage data permissions. DataArts Catalog provides data maps.	N/A	Overview Overview
DataArts DataService	Use DataArts DataService to centrally manage API services, create data APIs based on tables, and register APIs with DataArts DataService itself for unified management and publication.	Develop APIs.	Preparations Creating an API Debugging an API Publishing an API Managing APIs Creating Throttling Policies
		Call APIs.	Calling APIs

2 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, buy a DataArts Studio instance and create a workspace by following the instructions provided in [Preparation](#). 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](#).

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 [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 [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 [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 [Developing an SQL Script](#).

3 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.

3.1 Data Sources

Before using DataArts Studio, you need to select cloud services or databases as the data lake foundation, which provides storage and compute capabilities. DataArts Studio provides one-stop data development, governance, and services based on the data lake foundation.

Data Sources Supported By DataArts Studio

DataArts Studio can interconnect with cloud services such as DWS, DLI, and MRS Hive as well as traditional databases such as MySQL and Oracle. For details, see [Table 3-1](#).

To connect to these data sources, go to the DataArts Studio console and choose **Management Center** to create a data connection.

NOTE

Data connections in Management Center are independent of the data links in DataArts Migration. To use the data connections in DataArts Migration, create corresponding data links in DataArts Migration first.

- The data connections in Management Center are used to connect to the data lake foundation. DataArts Studio provides one-stop data development, governance, and services based on the data lake foundation.
- Data links in DataArts Migration can be used only in DataArts Migration to integrate source datasets into the destination data lake foundation. For details about the data sources supported by DataArts Migration, see [Data Sources Supported by DataArts Migration](#).

Table 3-1 Data sources supported by DataArts Studio

Data Source Type	Management Center	DataArts Architecture	DataArts Factory	DataArts Catalog ^[1]	DataArts Quality ^[2]	DataArts DataService
DWS	Supported	Supported	Supported	Supported	Supported	Supported
DLI	Supported	Supported	Supported	Supported	Supported	Supported
MRS HBase	Supported	Not supported	Not supported	Supported	Not supported	Not supported
MapReduce (MRS) Hive	Supported	Supported	Supported	Supported	Supported	Not supported
MRS Kafka	Supported	Not supported	Supported	Not supported	Not supported	Not supported
MapReduce (MRS) Ranger	Supported	Not supported	Not supported	Not supported	Not supported	Not supported
MySQL	Supported	Not supported	Not supported	Not supported	Supported	Supported
MapReduce (MRS) Spark ^[4]	Supported	Supported	Supported	Not supported	Supported	Not supported
RDS for MySQL	Supported	Not supported	Supported	Supported	Supported	Supported
RDS for PostgreSQL	Supported	Supported	Supported	Supported	Supported	Not supported
Host Connection	Supported	Not supported	Supported	Not supported	Not supported	Not supported
MapReduce (MRS) Presto	Supported	Not supported	Supported	Not supported	Not supported	Not supported

Annotation

[1] DataArts Catalog: In addition to the data sources listed in the preceding table, DataArts Catalog can also collect metadata of the following data sources:

1. Relational databases, such as MySQL and PostgreSQL databases (You can use RDS connections to collect the metadata of these databases.)
2. Cloud Search Service (CSS)
3. Graph Engine Service (GES)
4. Object Storage Service (OBS)
5. MRS Hudi (MRS Hudi is a data format. The metadata is stored in Hive, and operations are performed using Spark.) You can enable synchronization of the Hive table configuration for Hudi tables, and then you can collect the metadata of Hudi tables by collecting the MRS Hive metadata.

[2] The quality jobs and comparison jobs of DataArts Quality are not supported by MRS clusters with decoupled storage and compute.

[3] MRS Spark: MRS Spark connections can be used to integrate data into the DataArts Architecture and DataArts Quality modules. MRS Hudi is a data format. The metadata is stored in Hive, and operations are performed using Spark. DataArts Catalog uses MRS Hive to collect Hudi metadata, and DataArts Architecture and DataArts Quality use MRS Spark to govern Hudi data sources. (Business metric monitoring of DataArts Quality does not support Hudi data sources.)

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.

Data Source Type	Description
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>
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>
MRS Ranger	<p>Ranger offers a centralized security management framework and supports unified authorization and auditing. It manages fine-grained access control over Hadoop and related components, such as HDFS, Hive, HBase, Kafka, and Storm. You can use the frontend web UI console provided by Ranger to configure policies to control users' access to these components.</p>

Data Source Type	Description
MRS Hudi	<p>Hudi is a data lake table format that provides the ability to update and delete data as well as consume new data on HDFS. It supports multiple compute engines and provides insert, update, and delete (IUD) interfaces and streaming primitives, including upsert and incremental pull, over datasets on HDFS. Hudi metadata is stored in Hive, and operations are performed using Spark.</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 Host Connections 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 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.
- For host connections, only Linux hosts are supported.
- If changes occur in the connected data lake (for example, the MRS cluster capacity is expanded), you need to edit and save the connection.

Prerequisites

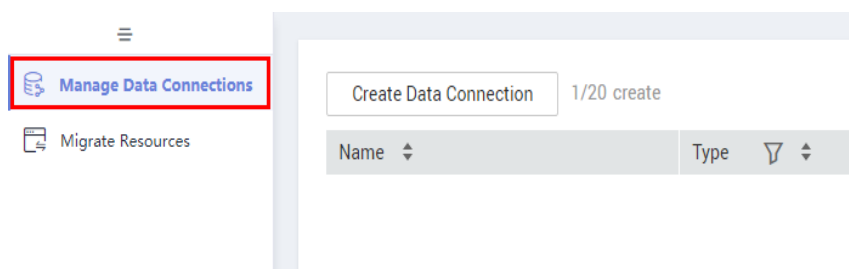
- You have created a data lake to connect, for example, a database or cloud service supported by DataArts Studio.
 - Before creating a DWS data connection, ensure that you have created a cluster in DWS and have the permissions required to view Key Management Service (KMS) keys.
 - Before creating an MRS HBase, MRS Hive, MRS Kafka, MRS Ranger, MRS Presto, or MRS Spark connection, ensure that you have bought an MRS cluster and selected required components.
 - Before creating an RDS data connection, ensure that you have bought 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 [configuring routes](#). For details about how to configure security group rules, see [configuring security group rules](#).
- The cloud service instance and the DataArts Studio workspace belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

Creating a Data Connection

1. On the DataArts Studio console, locate a workspace and click **Management Center**.
2. In the navigation pane, choose **Manage Data Connections**.

Figure 3-1 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 3-2 Create Data Connection

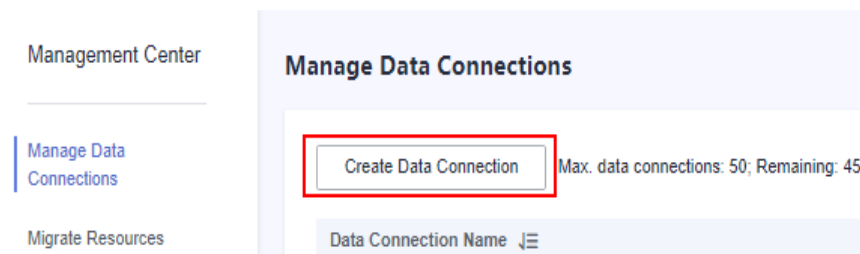


Table 3-3 Data connections

Data Connection Type	Link
MRS Hive	Table 3-4

Data Connection Type	Link
MRS HBase	Table 3-5
MRS Kafka	Table 3-6
MRS Ranger	Table 3-12
DWS	Table 3-9
DLI	Table 3-10
ORACLE	Table 3-11
MRS Spark	Table 3-7
RDS	See Table 3-8 . You can also create RDS connections to relational databases, such as MySQL, PostgreSQL, and Dameng databases.
MRS Presto	Table 3-13
MySQL (pending offline)	You are not advised to select this connection type. Instead, You are advised to select RDS . For details, see Table 3-8 .
Host Connection	See Table 3-14 .

- Click **Test** to test connectivity of the data connection. If the test passes, the data connection is created.
- 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	The attribute of the data connection to create. Tags make management easier. You can set a tag or select a tag created in Tags from the drop-down list. NOTE 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 cluster. Select an MRS cluster that Hive belongs to. Only MRS clusters are supported. A Hadoop cluster can be selected only after it is managed by MRS. All the MRS clusters with the same project ID and enterprise project are displayed.</p> <p>If the connection fails after you select a cluster, check whether the MRS cluster can communicate with the CDM instance which functions as the agent. They can communicate with each other in the following scenarios:</p> <ul style="list-style-type: none">• If the CDM cluster in the DataArts Studio instance and the MRS cluster are in different regions, a public network or a dedicated connection is required. If the Internet is used for communication, ensure that an EIP has been bound to the CDM cluster, and the MRS cluster can access the Internet and the port has been enabled in the firewall rule.• If the CDM cluster in the DataArts Studio instance and the MRS cluster are in the same region, VPC, subnet, and security group, they can communicate with each other by default. If they are in the same VPC but in different subnets or security groups, you must configure routing rules and security group rules. For details about how to configure routing rules, see configuring routes. For details about how to configure security group rules, see configuring security group rules.• The MRS cluster and the DataArts Studio workspace belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

Parameter	Mandatory	Description
Connection Type	Yes	<p>Connection type. Proxy connection is recommended.</p> <ul style="list-style-type: none"> • Proxy connection: An agent (CDM cluster) is used to access MRS clusters. This method supports all versions of MRS clusters. • MRS API connection: 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 MRS API connection, pay attention to the following restrictions: <ol style="list-style-type: none"> 1. Tables and fields cannot be viewed. 2. When the SQL editor is used to run SQL statements, the execution results can be displayed only in logs. 3. This method is not supported by data governance functions such as DataArts Architecture, DataArts Quality, and DataArts Catalog. 4. If a cluster managed by MRS is connected, disable OBS for DataArts Factory. When OBS is disabled, some functions (such as backup management and resource management) are restricted. <p>NOTE Select Proxy connection for Connection Type so that the DataArts Architecture, DataArts Quality, DataArts Catalog, and DataArts DataService components can use the MRS connection.</p>

Parameter	Mandatory	Description
Username	No	<p>The username of the MRS cluster. This parameter is mandatory when Connection Type is set to Proxy connection. 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 admin. User admin 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 Creating a Kerberos Authentication User for an MRS Security Cluster.</p> <p>When creating an MRS data connection, set Username and Password to the new MRS username and password.</p> <p>NOTE</p> <ul style="list-style-type: none"> • 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.
Password	No	<p>The password for accessing the MRS cluster. This parameter is mandatory when Connection Type is set to Proxy connection.</p>
KMS Key	No	<p>The name of the KMS key. This parameter is mandatory when Connection Type is set to Proxy connection.</p>

Parameter	Mandatory	Description
Agent	No	<p>This parameter is mandatory when Connection Type is set to Proxy connection.</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. You can set a tag or select a tag created in Tags from the drop-down list.</p> <p>NOTE 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 cluster. Select an MRS cluster that Hive belongs to. Only MRS clusters are supported. A Hadoop cluster can be selected only after it is managed by MRS. All the MRS clusters with the same project ID and enterprise project are displayed.</p> <p>If the connection fails after you select a cluster, check whether the MRS cluster can communicate with the CDM instance which functions as the agent. They can communicate with each other in the following scenarios:</p> <ul style="list-style-type: none">• If the CDM cluster in the DataArts Studio instance and the MRS cluster are in different regions, a public network or a dedicated connection is required. If the Internet is used for communication, ensure that an EIP has been bound to the CDM cluster, and the MRS cluster can access the Internet and the port has been enabled in the firewall rule.• If the CDM cluster in the DataArts Studio instance and the MRS cluster are in the same region, VPC, subnet, and security group, they can communicate with each other by default. If they are in the same VPC but in different subnets or security groups, you must configure routing rules and security group rules. For details about how to configure routing rules, see configuring routes. For details about how to configure security group rules, see configuring security group rules.• The MRS cluster and the DataArts Studio workspace belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

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 admin. User admin 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 Creating a Kerberos Authentication User for an MRS Security Cluster. When creating an MRS data connection, set Username and Password to the new MRS username and password.</p> <p>NOTE</p> <ul style="list-style-type: none"> 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.
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. You can set a tag or select a tag created in Tags from the drop-down list. NOTE 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 cluster. Select an MRS cluster that Hive belongs to. Only MRS clusters are supported. A Hadoop cluster can be selected only after it is managed by MRS. All the MRS clusters with the same project ID and enterprise project are displayed. If the connection fails after you select a cluster, check whether the MRS cluster can communicate with the CDM instance which functions as the agent. They can communicate with each other in the following scenarios: <ul style="list-style-type: none">• If the CDM cluster in the DataArts Studio instance and the MRS cluster are in different regions, a public network or a dedicated connection is required. If the Internet is used for communication, ensure that an EIP has been bound to the CDM cluster, and the MRS cluster can access the Internet and the port has been enabled in the firewall rule.• If the CDM cluster in the DataArts Studio instance and the MRS cluster are in the same region, VPC, subnet, and security group, they can communicate with each other by default. If they are in the same VPC but in different subnets or security groups, you must configure routing rules and security group rules. For details about how to configure routing rules, see configuring routes. For details about how to configure security group rules, see configuring security group rules.• The MRS cluster and the DataArts Studio workspace belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

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 admin. User admin 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 Creating a Kerberos Authentication User for an MRS Security Cluster. When creating an MRS data connection, set Username and Password to the new MRS username and password.</p> <p>NOTE</p> <ul style="list-style-type: none"> 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.
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	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. You can set a tag or select a tag created in Tags from the drop-down list. NOTE 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 cluster. Select an MRS cluster that Hive belongs to. Only MRS clusters are supported. A Hadoop cluster can be selected only after it is managed by MRS. All the MRS clusters with the same project ID and enterprise project are displayed. If the connection fails after you select a cluster, check whether the MRS cluster can communicate with the CDM instance which functions as the agent. They can communicate with each other in the following scenarios: <ul style="list-style-type: none">• If the CDM cluster in the DataArts Studio instance and the MRS cluster are in different regions, a public network or a dedicated connection is required. If the Internet is used for communication, ensure that an EIP has been bound to the CDM cluster, and the MRS cluster can access the Internet and the port has been enabled in the firewall rule.• If the CDM cluster in the DataArts Studio instance and the MRS cluster are in the same region, VPC, subnet, and security group, they can communicate with each other by default. If they are in the same VPC but in different subnets or security groups, you must configure routing rules and security group rules. For details about how to configure routing rules, see configuring routes. For details about how to configure security group rules, see configuring security group rules.• The MRS cluster and the DataArts Studio workspace belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

Parameter	Mandatory	Description
Connection Type	Yes	<p>Connection type. Proxy connection is recommended.</p> <ul style="list-style-type: none"> • Proxy connection: An agent (CDM cluster) is used to access MRS clusters. This method supports all versions of MRS clusters. • MRS API connection: 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 MRS API connection, pay attention to the following restrictions: <ol style="list-style-type: none"> 1. Tables and fields cannot be viewed. 2. When the SQL editor is used to run SQL statements, the execution results can be displayed only in logs. 3. This method is not supported by data governance functions such as DataArts Architecture, DataArts Quality, and DataArts Catalog. 4. If a cluster managed by MRS is connected, disable OBS for DataArts Factory. When OBS is disabled, some functions (such as backup management and resource management) are restricted. <p>NOTE Select Proxy connection for Connection Type so that the DataArts Architecture, DataArts Quality, DataArts Catalog, and DataArts DataService components can use the MRS connection.</p>

Parameter	Mandatory	Description
Username	No	<p>The username of the MRS cluster. This parameter is mandatory when Connection Type is set to Proxy connection. 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 admin. User admin 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 Creating a Kerberos Authentication User for an MRS Security Cluster. When creating an MRS data connection, set Username and Password to the new MRS username and password.</p> <p>NOTE</p> <ul style="list-style-type: none"> • 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.
Password	No	<p>The password for accessing the MRS cluster. This parameter is mandatory when Connection Type is set to Proxy connection.</p>
KMS Key	No	<p>The name of the KMS key. This parameter is mandatory when Connection Type is set to Proxy connection.</p>

Parameter	Mandatory	Description
Agent	No	<p>This parameter is mandatory when Connection Type is set to Proxy connection.</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 3-8 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. You can set a tag or select a tag created in Tags from the drop-down list.</p> <p>NOTE 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"> 1. Log in to the management console using the account. 2. In the Service List, choose Relational Database Service. In the left navigation pane, choose Instances. 3. Click the name of an instance. The basic information page of the instance is displayed. <p>You can obtain the IP address on the Connection Information 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"> 1. Log in to the management console using the account. 2. In the Service List, choose Relational Database Service. In the left navigation pane, choose Instances. 3. Click the name of an instance. The basic information page of the instance is displayed. <p>You can obtain the database port on the Connection Information tab.</p>
Driver Name	Yes	<p>The name of the driver. The following values are available:</p> <ul style="list-style-type: none"> • com.mysql.jdbc.Driver • org.postgresql.Driver
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"> • MySQL driver: Download it from https://downloads.mysql.com/archives/c-j/. The 5.1.48 version is recommended. • PostgreSQL driver: Download it from https://mvnrepository.com/artifact/org.postgresql/postgresql. The 42.1.4 version is recommended. <p>NOTE 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	The name of the KMS key. To obtain the key: 1. Log in to the management console using the account. 2. Click Key Management Service and select Key Management Service 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	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. You can set a tag or select a tag created in Tags from the drop-down list. NOTE 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 click  to disable or enable the Manual function. <ul style="list-style-type: none">• When Manual is disabled, you do not need to enter the IP address and port.• When Manual is enabled, you must enter the IP address and port.
IP Address	No	The IP address for accessing the cluster database through the internal network. This parameter is mandatory when Manual 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 Manual 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 SSL Connection to set the communication mode. If SSL Connection is enabled, only SSL encryption can be used. If SSL Connection is disabled, both modes can be used. SSL Connection is disabled by default.
Cluster Name	No	This parameter is mandatory when Manual is disabled. Select a DWS cluster. All the DWS clusters with the same project ID and enterprise project are displayed.
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.

Parameter	Mandatory	Description
Agent	No	<p>This parameter is mandatory when Connection Type is set to Proxy connection.</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 DLI 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. You can set a tag or select a tag created in Tags from the drop-down list.</p> <p>NOTE The name of the tag. Only letters, numbers, and underscores (_) are allowed. Tag names cannot start with underscores (_). Enter up to 100 characters.</p>

Table 3-11 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 (-).

Parameter	Mandatory	Description
Tag	No	The attribute of the data connection to create. Tags make management easier. You can set a tag or select a tag created in Tags from the drop-down list. NOTE The name of the tag. Only letters, numbers, and underscores (_) are allowed. Tag names cannot start with underscores (_). Enter up to 100 characters.
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.
Username	Yes	The username of the account for accessing the database. This account must have the permissions required to read and write data tables and metadata. NOTE 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: <ol style="list-style-type: none">1. Log in to the Oracle node as user root.2. Run the following command to switch to user oracle: su oracle3. Run the following command to log in to the database: sqlplus /nolog4. Run the following command to log in as user sys: connect sys as sysdba; Enter the password of user sys .5. Run the following SQL statement to grant permissions: GRANT SELECT ON GV_\$INSTANCE to xxx; In the preceding command, <i>xxx</i> indicates the name of the user to which the permissions will be granted.
Password	Yes	The user password.
Connection type	Yes	Select a connection type. <ul style="list-style-type: none">• SID SID indicates the ID of the Oracle database instance. One instance corresponds to only one database, but one database can correspond to multiple instances.• Service Name It was introduced since Oracle8i and indicates the external service name of the Oracle database.

Parameter	Mandatory	Description
SID	No	This parameter is mandatory when Connection type is set to SID . SID indicates the ID of the Oracle database instance. One instance corresponds to only one database, but one database can correspond to multiple instances.
Service Name	No	This parameter is mandatory when Connection type is set to Service Name . It was introduced since Oracle8i and indicates the external service name of the Oracle database.
KMS Key	Yes	The name of the KMS key. To obtain the key: 1. Log in to the management console using the created account. 2. Click Key Management Service and select Key Management Service from the list on the left. You can obtain the key name from the key list.
Agent	Yes	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. As a network proxy, the CDM cluster must be able to communicate with Oracle.

Table 3-12 MRS Ranger 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. You can set a tag or select a tag created in Tags from the drop-down list. NOTE 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 cluster. Select an MRS cluster that Hive belongs to. Only MRS clusters are supported. A Hadoop cluster can be selected only after it is managed by MRS. All the MRS clusters with the same project ID and enterprise project are displayed.</p> <p>If the connection fails after you select a cluster, check whether the MRS cluster can communicate with the CDM instance which functions as the agent. They can communicate with each other in the following scenarios:</p> <ul style="list-style-type: none"> • If the CDM cluster in the DataArts Studio instance and the MRS cluster are in different regions, a public network or a dedicated connection is required. If the Internet is used for communication, ensure that an EIP has been bound to the CDM cluster, and the MRS cluster can access the Internet and the port has been enabled in the firewall rule. • If the CDM cluster in the DataArts Studio instance and the MRS cluster are in the same region, VPC, subnet, and security group, they can communicate with each other by default. If they are in the same VPC but in different subnets or security groups, you must configure routing rules and security group rules. For details about how to configure routing rules, see configuring routes. For details about how to configure security group rules, see configuring security group rules. • The MRS cluster and the DataArts Studio workspace belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace. <p>NOTE Currently, DataArts Studio only supports connections to MRS Ranger in a cluster in security mode.</p>

Parameter	Mandatory	Description
Username	Yes	<p>The username of the MRS cluster. This parameter is mandatory when Connection Type is set to Proxy connection. 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 admin. User admin 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 Creating a Kerberos Authentication User for an MRS Security Cluster. When creating an MRS data connection, set Username and Password to the new MRS username and password.</p> <p>NOTE</p> <ul style="list-style-type: none">• 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.
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 3-13 MRS Presto 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. You can set a tag or select a tag created in Tags from the drop-down list. NOTE 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 cluster. Select an MRS cluster that Hive belongs to. Only MRS clusters are supported. A Hadoop cluster can be selected only after it is managed by MRS. All the MRS clusters with the same project ID and enterprise project are displayed. If the connection fails after you select a cluster, check whether the MRS cluster can communicate with the CDM instance which functions as the agent. They can communicate with each other in the following scenarios: <ul style="list-style-type: none">• If the CDM cluster in the DataArts Studio instance and the MRS cluster are in different regions, a public network or a dedicated connection is required. If the Internet is used for communication, ensure that an EIP has been bound to the CDM cluster, and the MRS cluster can access the Internet and the port has been enabled in the firewall rule.• If the CDM cluster in the DataArts Studio instance and the MRS cluster are in the same region, VPC, subnet, and security group, they can communicate with each other by default. If they are in the same VPC but in different subnets or security groups, you must configure routing rules and security group rules. For details about how to configure routing rules, see configuring routes. For details about how to configure security group rules, see configuring security group rules.• The MRS cluster and the DataArts Studio workspace belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.
Description	No	You can enter the description of the connection.

Table 3-14 Host Connection

Parameter	Mandatory	Description
Data Connection Name	Yes	Name of the host connection. The value can contain only letters, digits, hyphens (-), and underscores (_).
Tag	No	The attribute of the data connection to create. Tags make management easier. You can set a tag or select a tag created in Tags from the drop-down list. NOTE The tag name can contain letters, digits, and underscores (_), and cannot start with underscores (_). It can contain up to 100 characters.
Host Address	Yes	IP address of the Linux host For details, see Viewing Details About an ECS .
Agent	Yes	Agents provided by the CDM cluster, which is required if Proxy connection is selected for Connection Type .
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">• Key pair• Password
Key Pair	Yes	If you select Key pair for Login Mode , you need to obtain the private key file, upload it to OBS, and select the OBS path. This parameter is available only when Login Mode is set to Key pair . NOTE 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.
KMS Key	Yes	Key created on Key Management Service (KMS) and used for encrypting and decrypting user passwords and key pairs. You can select a created key from KMS.
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

3.3 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 data services, metadata categories, metadata tags, metadata collection tasks, and 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](#). For details on how to classify metadata and add tags, see [Tags](#). For details on how to create collection tasks, see [Task Management](#). For details on how to publish APIs, see [Publishing an API](#).

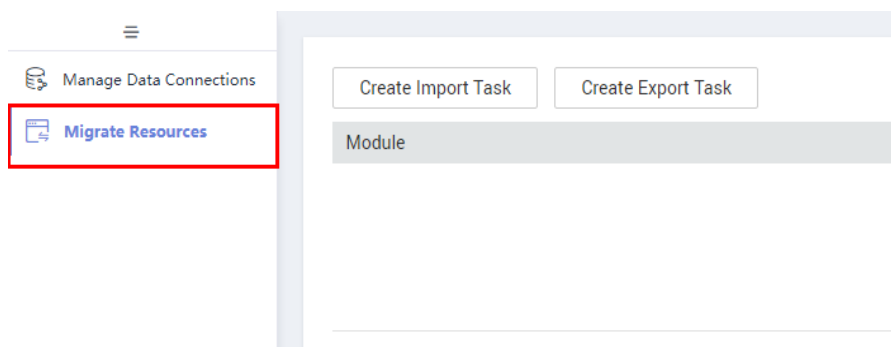
Constraints

- Collection tasks with the same name cannot be migrated repeatedly.
- Categories and tags with the same name cannot be migrated repeatedly.
- 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**.
2. In the navigation pane, choose **Migrate Resources**.

Figure 3-3 Migrating Resources



3. Click **Create Export Task** to configure the file name and the OBS path for saving the file.

Figure 3-4 Export Task

Export Task ×

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

* File Name

* OBS Bucket

* OBS Path

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

Figure 3-5 Selecting the resource to export

Export Task ×

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

DataLakeService

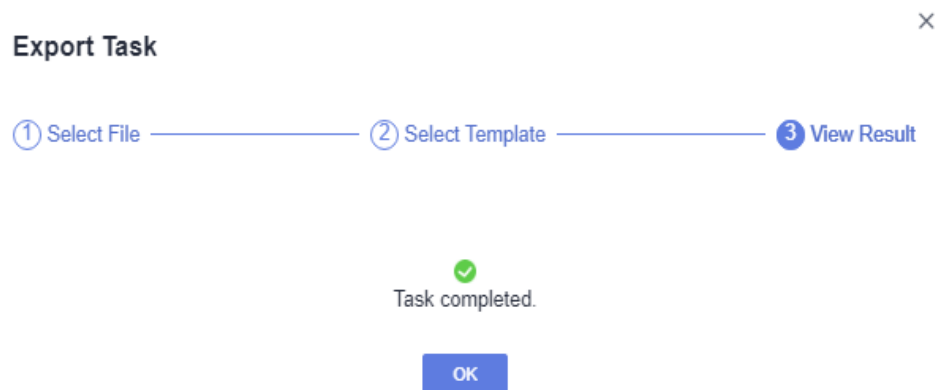
- DataService
- DataManager
- DataSource

MetaData

- Classification
- Collect
- Term

5. Click **Next** and wait until the export is complete. The resource package is exported to the OBS path set in 3.

Figure 3-6 Export completed

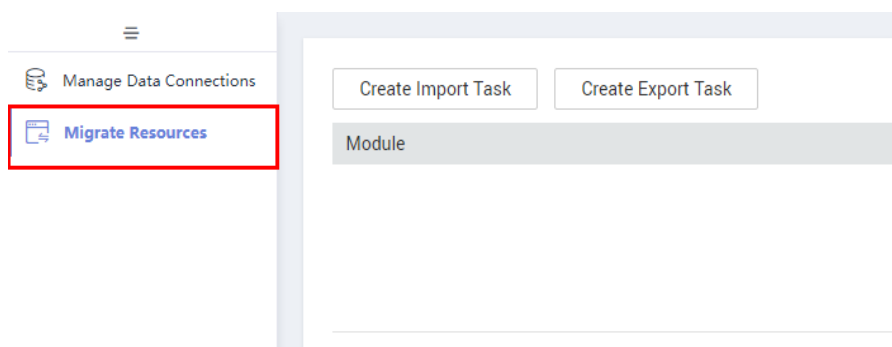


If no result is displayed in 1 minute, the export fails. Try again. If the failure persists, contact the customer service or technical support.

Importing a Resource

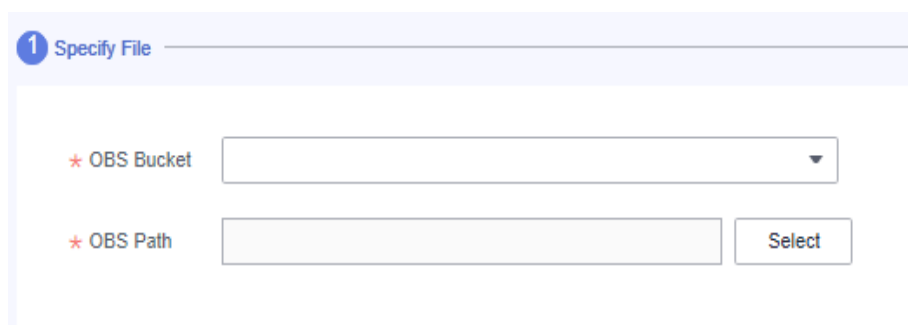
1. On the DataArts Studio console, locate a workspace and click **Management Center**.
2. In the navigation pane, choose **Migrate Resources**.

Figure 3-7 Migrating Resources



3. Click **Create Import Task** and configure the path for saving the resources to import.

Figure 3-8 Configuring the path for saving the resources to import



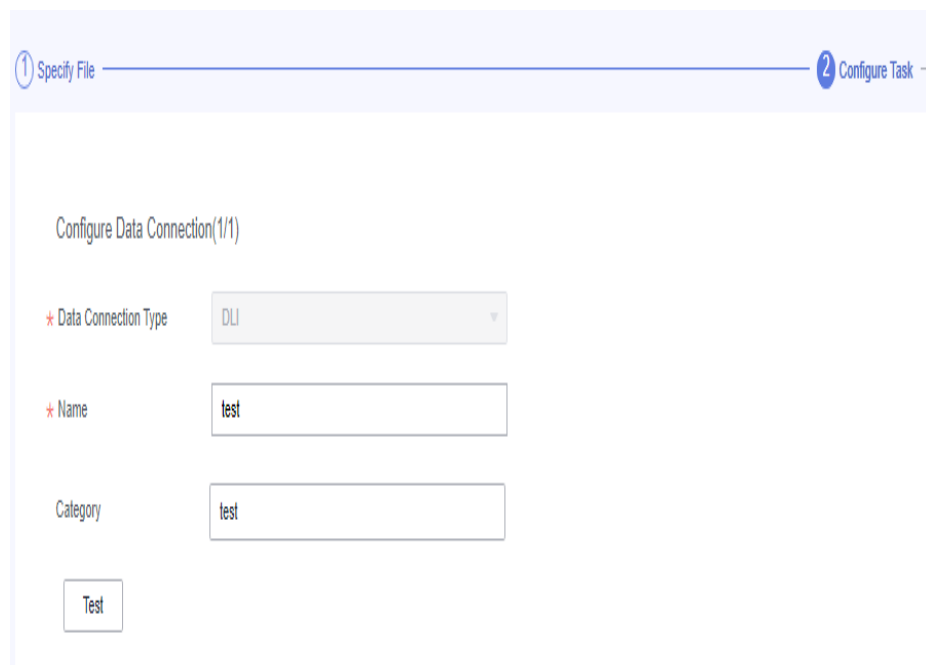
4. Click **Next** and select the resource to import.

Figure 3-9 Selecting the resource to import



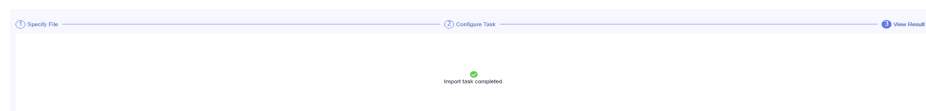
5. If you select **DataSource**, click **Next** to configure a data connection. The number of data connections required is determined by the number of data sources. Each data connection requires a password.

Figure 3-10 Configuring a data connection



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

Figure 3-11 Import completed



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

3.4 Tutorials

3.4.1 Creating an MRS Hive Connection

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

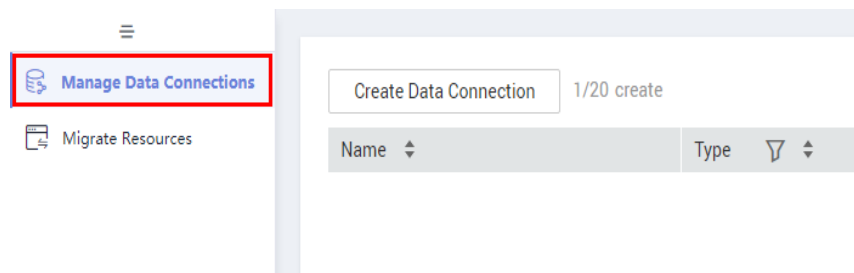
Prerequisites

- 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 Ranger, MRS Presto, or MRS Spark connection, ensure that you have bought an MRS cluster and selected required components.
 - Before creating an RDS data connection, ensure that you have bought 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 [configuring routes](#). For details about how to configure security group rules, see [configuring security group rules](#).
 - The cloud service instance and the DataArts Studio workspace belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

Creating a Data Connection

1. On the DataArts Studio console, locate a workspace and click **Management Center**.
2. In the navigation pane, choose **Manage Data Connections**.

Figure 3-12 Manage Data Connections



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 3-15](#).

Figure 3-13 Create Data Connection

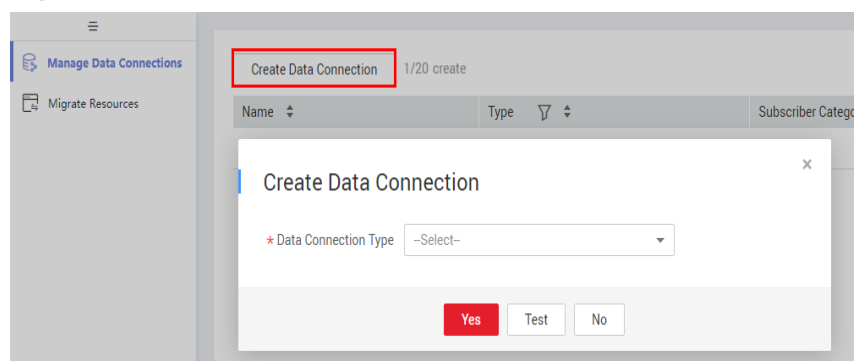


Figure 3-14 MRS Hive connection parameters

* Data Connection Type	<input type="text" value="MRS Hive"/>	
* Name	<input type="text"/>	
Category	<input type="text"/>	
* Cluster Name ?	<input type="text"/>	Manage Cluster
* Username	<input type="text"/>	
* Password	<input type="text"/>	
* KMS Key ?	<input type="text"/>	Access KMS
* Connection Type	<input checked="" type="radio"/> Proxy connection <input type="radio"/> MRS API connection	
* Agent ?	<input type="text"/>	Manage CDM Clusters
	<input type="button" value="Test"/>	

Table 3-15 MRS Hive 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. You can set a tag or select a tag created in Tags from the drop-down list. NOTE 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 cluster. Select an MRS cluster that Hive belongs to. Only MRS clusters are supported. A Hadoop cluster can be selected only after it is managed by MRS. All the MRS clusters with the same project ID and enterprise project are displayed. If the connection fails after you select a cluster, check whether the MRS cluster can communicate with the CDM instance which functions as the agent. They can communicate with each other in the following scenarios: <ul style="list-style-type: none">• If the CDM cluster in the DataArts Studio instance and the MRS cluster are in different regions, a public network or a dedicated connection is required. If the Internet is used for communication, ensure that an EIP has been bound to the CDM cluster, and the MRS cluster can access the Internet and the port has been enabled in the firewall rule.• If the CDM cluster in the DataArts Studio instance and the MRS cluster are in the same region, VPC, subnet, and security group, they can communicate with each other by default. If they are in the same VPC but in different subnets or security groups, you must configure routing rules and security group rules. For details about how to configure routing rules, see configuring routes. For details about how to configure security group rules, see configuring security group rules.• The MRS cluster and the DataArts Studio workspace belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

Parameter	Mandatory	Description
Connection Type	Yes	<p>Connection type. Proxy connection is recommended.</p> <ul style="list-style-type: none"> • Proxy connection: An agent (CDM cluster) is used to access MRS clusters. This method supports all versions of MRS clusters. • MRS API connection: 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 MRS API connection, pay attention to the following restrictions: <ol style="list-style-type: none"> 1. Tables and fields cannot be viewed. 2. When the SQL editor is used to run SQL statements, the execution results can be displayed only in logs. 3. This method is not supported by data governance functions such as DataArts Architecture, DataArts Quality, and DataArts Catalog. 4. If a cluster managed by MRS is connected, disable OBS for DataArts Factory. When OBS is disabled, some functions (such as backup management and resource management) are restricted. <p>NOTE Select Proxy connection for Connection Type so that the DataArts Architecture, DataArts Quality, DataArts Catalog, and DataArts DataService components can use the MRS connection.</p>

Parameter	Mandatory	Description
Username	No	<p>The username of the MRS cluster. This parameter is mandatory when Connection Type is set to Proxy connection. 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 admin. User admin 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 Creating a Kerberos Authentication User for an MRS Security Cluster. When creating an MRS data connection, set Username and Password to the new MRS username and password.</p> <p>NOTE</p> <ul style="list-style-type: none"> • 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.
Password	No	The password for accessing the MRS cluster. This parameter is mandatory when Connection Type is set to Proxy connection .
KMS Key	No	The name of the KMS key. This parameter is mandatory when Connection Type is set to Proxy connection .

Parameter	Mandatory	Description
Agent	No	<p>This parameter is mandatory when Connection Type is set to Proxy connection.</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>

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.

3.4.2 Creating a DWS Connection

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

Prerequisites

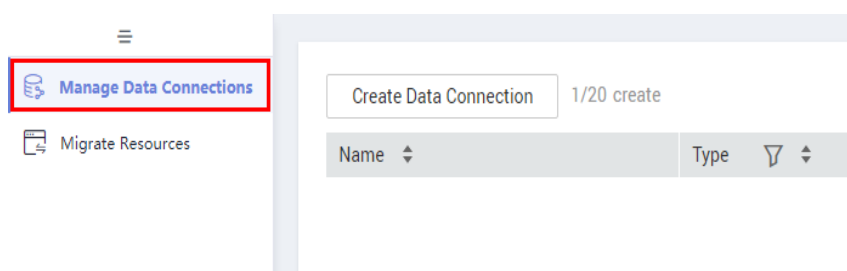
- You have created a data lake to connect, for example, a database or cloud service supported by DataArts Studio.

- 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 Ranger, MRS Presto, or MRS Spark connection, ensure that you have bought an MRS cluster and selected required components.
- Before creating an RDS data connection, ensure that you have bought 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 [configuring routes](#). For details about how to configure security group rules, see [configuring security group rules](#).
 - The cloud service instance and the DataArts Studio workspace belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

Creating a Data Connection

1. On the DataArts Studio console, locate a workspace and click **Management Center**.
2. In the navigation pane, choose **Manage Data Connections**.

Figure 3-15 Manage Data Connections



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 3-16](#).

Figure 3-16 Create Data Connection

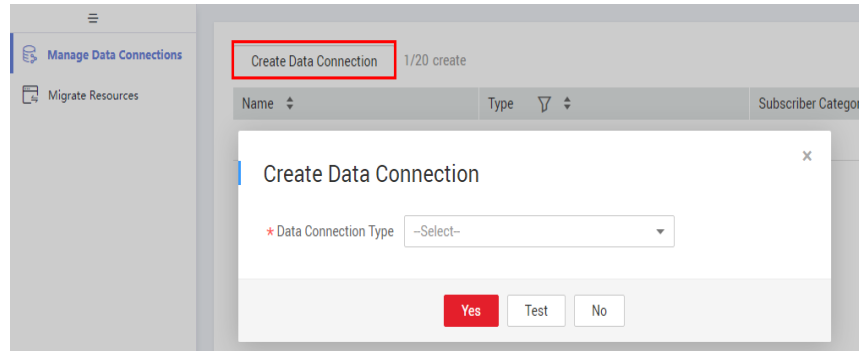


Figure 3-17 DWS connection parameters

* 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 3-16 DWS data connection

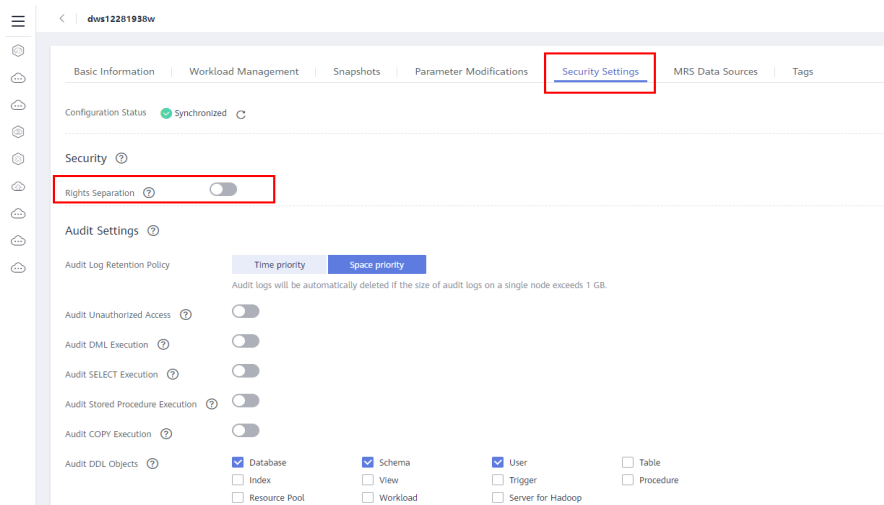
Parameter	Mandatory	Description
Data Connection Name	Yes	The name of the data connection to create. Data connection names can contain 1 to 50 characters. They can include only letters, numbers, underscores (_), and hyphens (-).
Tag	No	The attribute of the data connection to create. Tags make management easier. You can set a tag or select a tag created in Tags from the drop-down list. NOTE 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 click  to disable or enable the Manual function. <ul style="list-style-type: none"> When Manual is disabled, you do not need to enter the IP address and port. When Manual is enabled, you must enter the IP address and port.
IP Address	No	The IP address for accessing the cluster database through the internal network. This parameter is mandatory when Manual 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 Manual 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 SSL Connection to set the communication mode. If SSL Connection is enabled, only SSL encryption can be used. If SSL Connection is disabled, both modes can be used. SSL Connection is disabled by default.
Cluster Name	No	This parameter is mandatory when Manual is disabled. Select a DWS cluster. All the DWS clusters with the same project ID and enterprise project are displayed.

Parameter	Mandatory	Description
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.
Agent	No	<p>This parameter is mandatory when Connection Type is set to Proxy connection.</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>

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 3-18 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.

3.4.3 Creating a MySQL Connection

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

Prerequisites

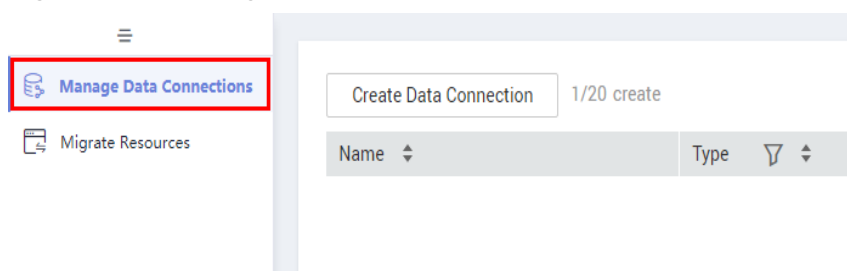
- You have created a data lake to connect, for example, a database or cloud service supported by DataArts Studio.
 - Before creating a DWS data connection, ensure that you have created a cluster in DWS and have the permissions required to view Key Management Service (KMS) keys.
 - Before creating an MRS HBase, MRS Hive, MRS Kafka, MRS Ranger, MRS Presto, or MRS Spark connection, ensure that you have bought an MRS cluster and selected required components.
 - Before creating an RDS data connection, ensure that you have bought 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 [configuring routes](#). For details about how to configure security group rules, see [configuring security group rules](#).
- The cloud service instance and the DataArts Studio workspace belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

Creating a Data Connection

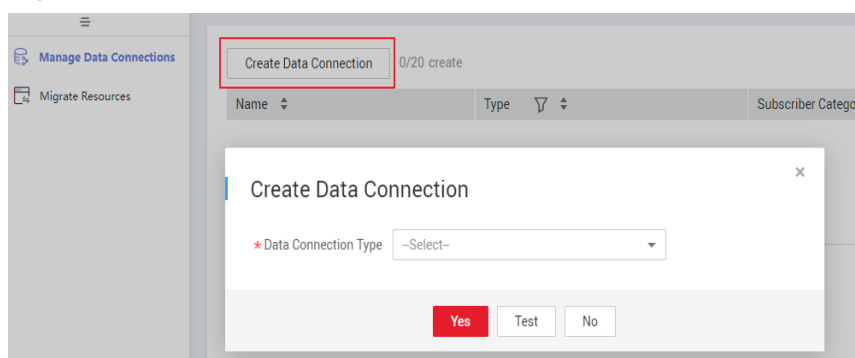
1. On the DataArts Studio console, locate a workspace and click **Management Center**.
2. In the navigation pane, choose **Manage Data Connections**.

Figure 3-19 Manage Data Connections



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 3-17](#).

Figure 3-20 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 3-21 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 'Manage CDM' link)
- Agent:** cdm-dayu (with a 'Manage CDM' link)

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

Table 3-17 RDS data connection

Parameter	Man datory	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. You can set a tag or select a tag created in Tags from the drop-down list. NOTE 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 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"> 1. Log in to the management console using the account. 2. In the Service List, choose Relational Database Service. In the left navigation pane, choose Instances. 3. Click the name of an instance. The basic information page of the instance is displayed. <p>You can obtain the IP address on the Connection Information 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"> 1. Log in to the management console using the account. 2. In the Service List, choose Relational Database Service. In the left navigation pane, choose Instances. 3. Click the name of an instance. The basic information page of the instance is displayed. <p>You can obtain the database port on the Connection Information tab.</p>
Driver Name	Yes	<p>The name of the driver. The following values are available:</p> <ul style="list-style-type: none"> • com.mysql.jdbc.Driver • org.postgresql.Driver
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"> • MySQL driver: Download it from https://downloads.mysql.com/archives/c-j/. The 5.1.48 version is recommended. • PostgreSQL driver: Download it from https://mvnrepository.com/artifact/org.postgresql/postgresql. The 42.1.4 version is recommended. <p>NOTE 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	The name of the KMS key. To obtain the key: 1. Log in to the management console using the account. 2. Click Key Management Service and select Key Management Service 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.

4 DataArts Migration

4.1 Overview

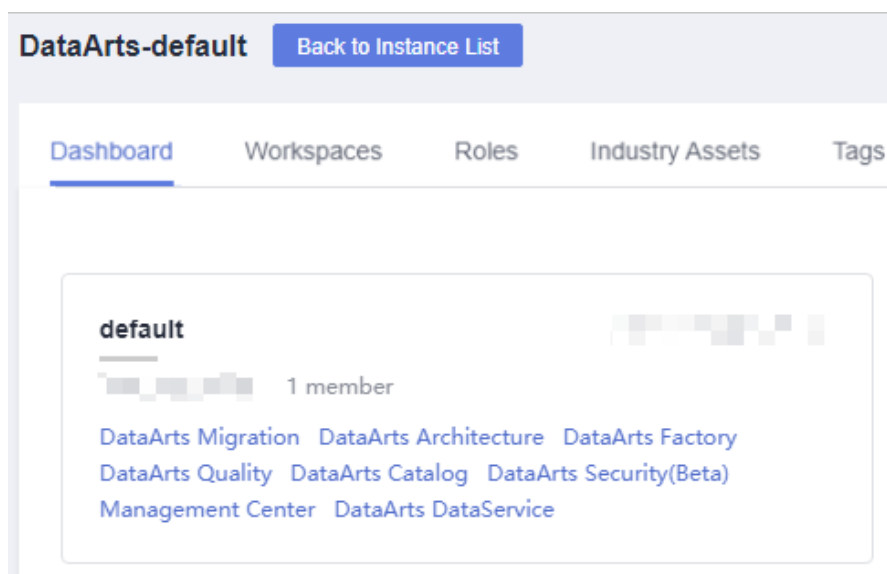
DataArts Migration is an efficient and easy-to-use data integration service. Based on the big data migration to the cloud and intelligent data lake solutions, CDM provides easy-to-use migration capabilities and can integrate various types of data sources into the data lake, which simplifies data source 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 4-1 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.

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.

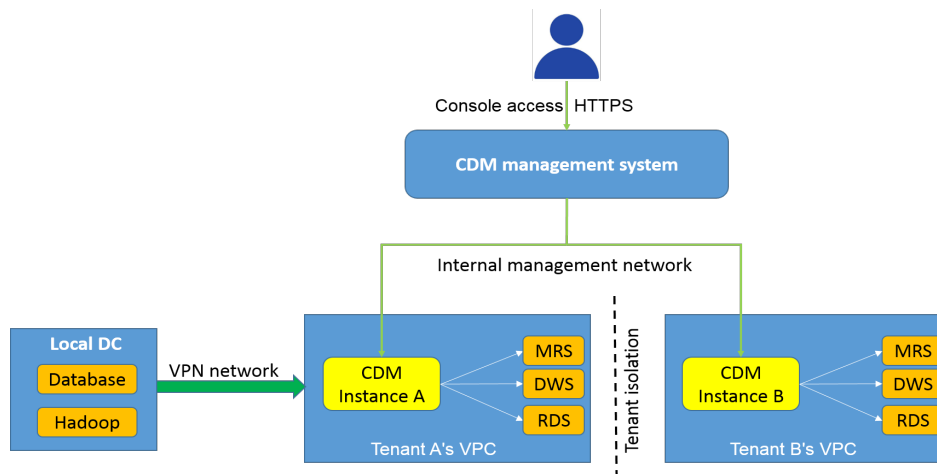
Migration Principles

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 4-2 Migration principles



4.2 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.
13. If a running job is interrupted unexpectedly, the data that has been written to the destination will not be deleted. You must manually delete the data if needed.

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. You are advised to use Data Replication Service (DRS) to migrate the incremental data of the database to RDS.
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.

You are advised to use **Direct Connect** to improve the migration speed while ensuring network access.

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 **field 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;

```

4.3 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 4-1](#) describes the supported data sources.

Table 4-1 Supported data sources during table/file migration

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

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

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

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

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

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

Category	Source	Destination	Description
	Database sharding	<ul style="list-style-type: none"> • Data warehouse: Data Lake Insight (DLI) • Hadoop: MRS HBase and MRS Hive • Search: Elasticsearch and Cloud Search Service (CSS) • Object-based storage: Object Storage Service (OBS) 	Database shards cannot serve as the destination.
NoSQL	Distributed Cache Service (DCS)	Hadoop: MRS HDFS, MRS HBase, and MRS Hive	NoSQL except CloudTable cannot serve as the destination. For how to migrate data from Redis to DCS, see Migrating Data from Self-Hosted Redis to DCS .
	Redis		
	Document Database Service (DDS)		
	MongoDB		
	CloudTable	<ul style="list-style-type: none"> • Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI) • Hadoop: MRS HDFS, MRS HBase, and MRS Hive • Object-based storage: Object Storage Service (OBS) • Relational database: RDS for MySQL, RDS for PostgreSQL, RDS for SQL Server, MySQL, PostgreSQL, Microsoft SQL Server, and Oracle • NoSQL: CloudTable • Search: Elasticsearch and Cloud Search Service (CSS) 	
Cassandra	<ul style="list-style-type: none"> • Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI) • Hadoop: MRS HDFS, MRS HBase, and MRS Hive • Object-based storage: Object Storage Service (OBS) • NoSQL: CloudTable • Search: Elasticsearch and Cloud Search Service (CSS) 		

Category	Source	Destination	Description
Message system	Apache Kafka	Search: Cloud Search Service (CSS)	The message system cannot serve as the destination.
	DMS Kafka		
	MRS Kafka	<ul style="list-style-type: none"> Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI) Hadoop: MRS HDFS, MRS HBase, and MRS Hive Object-based storage: Object Storage Service (OBS) Relational database: RDS for MySQL, RDS for PostgreSQL, and RDS for SQL Server NoSQL: CloudTable Search: Elasticsearch and Cloud Search Service (CSS) 	<ul style="list-style-type: none"> MRS Kafka cannot serve as the destination. Supported only by local storage and not in storage-compute decoupling scenarios Not supported by Ranger Not supported if SSL is enabled for ZooKeeper
Search	Elasticsearch	<ul style="list-style-type: none"> Data warehouse: GaussDB(DWS) and Data Lake Insight (DLI) Hadoop: MRS HDFS, MRS HBase, and MRS Hive 	Only the non-security mode is supported.
	Cloud Search Service (CSS)	<ul style="list-style-type: none"> Object-based storage: Object Storage Service (OBS) Relational database: RDS for MySQL, RDS for PostgreSQL, and RDS for SQL Server NoSQL: CloudTable Search: Elasticsearch and Cloud Search Service (CSS) 	N/A

 **NOTE**

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

Supported Data Sources in Entire DB Migration

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

Table 4-2 lists the data sources supporting entire DB migration using CDM.

Table 4-2 Supported data sources in entire DB migration

Category	Data Source	Read	Write	Description
Data warehouse	Data Warehouse Service (DWS)	Supported	Supported	-
	FusionInsight LibrA	Supported	Not supported	-
Hadoop (available only for local storage, and not for storage-compute decoupling, Ranger, or ZooKeeper for which SSL is enabled)	MRS HBase	Supported	Supported	Entire DB migration only to MRS HBase Recommended versions: <ul style="list-style-type: none"> • 2.1.X • 1.3.X
	MRS Hive	Supported	Supported	Entire DB migration only to a relational database 2.x versions are not supported. The following versions are recommended: <ul style="list-style-type: none"> • 1.2.X • 3.1.X
	FusionInsight HBase	Supported	Not supported	Recommended versions: <ul style="list-style-type: none"> • 2.1.X • 1.3.X
	FusionInsight Hive	Supported	Not supported	Entire DB migration only to a relational database 2.x versions are not supported. The following versions are recommended: <ul style="list-style-type: none"> • 1.2.X • 3.1.X

Category	Data Source	Read	Write	Description
	Apache HBase	Supported	Not supported	Recommended versions: <ul style="list-style-type: none"> • 2.1.X • 1.3.X
	Apache Hive	Supported	Not supported	Entire DB migration only to a relational database 2.x versions are not supported. The following versions are recommended: <ul style="list-style-type: none"> • 1.2.X • 3.1.X
Relational database	RDS for MySQL	Supported	Supported	Migration from OLTP to OLTP is not supported. In this scenario, you are advised to use the Data Replication Service (DRS).
	RDS for PostgreSQL	Supported	Supported	
	RDS for SQL Server	Supported	Supported	
	MySQL	Supported	Not supported	
	PostgreSQL	Supported	Not supported	
	Microsoft SQL Server	Supported	Not supported	
	Oracle	Supported	Not supported	

Category	Data Source	Read	Write	Description
	SAP HANA	Supported	Not supported	<ul style="list-style-type: none"> • Only the 2.00.050.00.15 92305219 version is supported. • Only the Generic Edition is supported. • BW/4 FOR HANA is not supported. • Only database names, table names, and column names consisting of English letters are supported. Special characters such as spaces and symbols are not allowed. • The following data types are supported: date, digit, Boolean, and character (except SHORTTEXT). Other data types such as binary are not supported. • During migration, tables cannot be automatically created at the destination.
	MyCAT	Supported	Not supported	-

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

Data Types Supported in Open-Source MySQL Database Migration

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

Table 4-3 Data types supported by the open-source MySQL database functioning as the source end

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

Category	Type	Description	Storage Format Example	Hive	DWS
Value	DECIMAL(M,D)	Uncompressed floating-point numbers cannot be unsigned. In unpacking decimals, each decimal corresponds to a byte. Defining the number of display lengths (M) and decimals (D) is required. NUMERIC is the synonym of DECIMAL.	52.36	DECIMAL	When D is 0, it corresponds to BIGINT. When D is not 0, it corresponds to 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 1000-01-01 to 9999-12-31 . For example, December 30, 1973 will be stored as 1973-12-30 .	1999-10-01	DATE	TIMESTAMP
	TIME	Stores information about the hour, minute, and second.	'09:10:21' or '9:10:21'	Not supported (string)	TIME

Category	Type	Description	Storage Format Example	Hive	DWS
	DATE TIME	The date and time are in the <i>YYYY-MM-DD HH:MM:SS</i> format and range from 1000-01-01 00:00:00 to 9999-12-31 23:59:59 . For example, 3:30 p.m. on December 30, 1973 will be stored as 1973-12-30 15:30:00 .	'1973-12-30 15:30:00'	TIMESTAMP	TIMESTAMP
	TIMESTAMP	Timestamp type. Timestamp between midnight on January 1, 1970 and a time point in 2037. Similar to the DATETIME format (YYYYMMDDHHMSS), except that no hyphen is required. For example, 3:30 p.m. December 30, 1973 will be stored as 19731230153000 .	19731230153000	TIMESTAMP	TIMESTAMP
	YEAR(M)	The year is stored in 2-digit or 4-digit number format. If the length is specified as 2 (for example, YEAR(2)), the year ranges from 1970 to 2069 (70 to 69). If the length is specified as 4, the year ranges from 1901 to 2155. The default length is 4.	2000	Not supported (string)	Not supported
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 4-4 Data types supported by the Oracle database

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

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

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

Data Types Supported in SQL Server Database Migration

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

Table 4-5 Data types supported by the SQL Server database functioning as the source end

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

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

Category	Type	Description	Hive	DWS	Oracle
	datetime2	Extended type of datetime, which has a larger data range. By default, the minimum precision is the highest, and the user-defined precision is optional.	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
	netxt	The function of this type is the same as that of the text type. It is non-Unicode data with variable length.	Not supported (string)	Not supported (string)	Not supported
	image	Variable-length binary data used to store pictures, catalog pictures, or paintings.	Not supported (string)	Not supported (string)	Not supported
	binary	Binary data with a fixed length of n bytes, where n ranges from 1 to 8,000.	Not supported (string)	Not supported (string)	Not supported

Category	Type	Description	Hive	DWS	Oracle
	varbinary	Variable-length binary data	Not supported (string)	Not supported (string)	Not supported
Currency data type	money	Stores currency values.	Not supported (string)	Not supported (string)	Not supported
	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 rowversion 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

4.4 Managing Clusters

4.4.1 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 a DataArts Studio instance includes a CDM cluster (except the trial version) and the cluster meets your requirements, you do not need to buy a DataArts Migration incremental package. If you need to create another CDM cluster, buy a DataArts Studio incremental package by referring to [Buying a DataArts Studio Incremental Package](#).

4.4.2 Binding or Unbinding an EIP

Scenario

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

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

NOTE

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

Prerequisites

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

Procedure

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

Figure 4-3 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

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

Figure 4-4 Cluster list



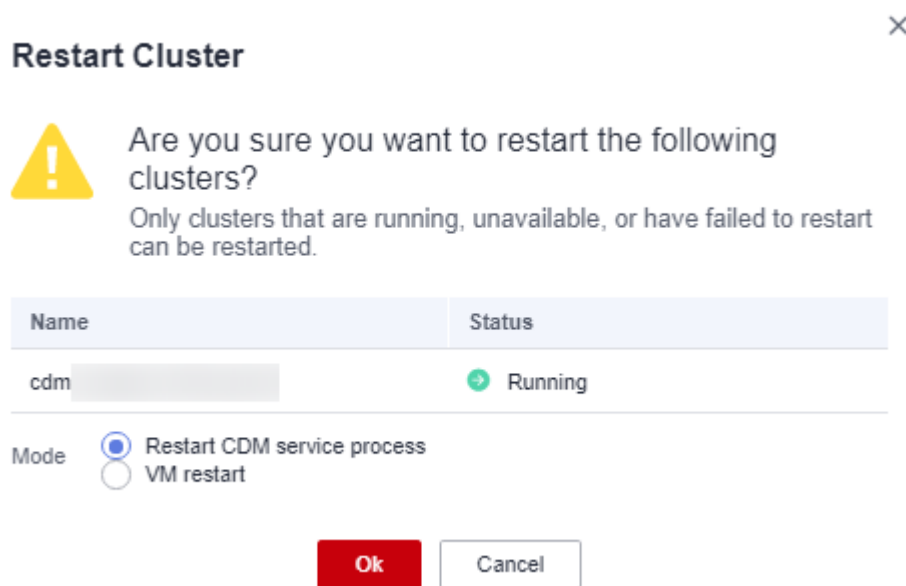
Name	Status	Internal Network Address	Public Network Address	Source	Enterprise Project	Operation
	Running			CDM	default	Job Management Bind EIP More

 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 4-5 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

4.4.4 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](#).
- You are not advised to delete the CDM cluster which is free of charge. If you delete it, you can only purchase clusters.
- After a CDM cluster is deleted, it will not be billed in pay-per-use mode and the package duration will not be deducted. If you have purchased a CDM discount package or a yearly/monthly CDM incremental package for the CDM cluster to delete, unsubscribe from the package.

Prerequisites

You have created a CDM cluster.

Deleting a Cluster

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

Figure 4-6 Cluster list

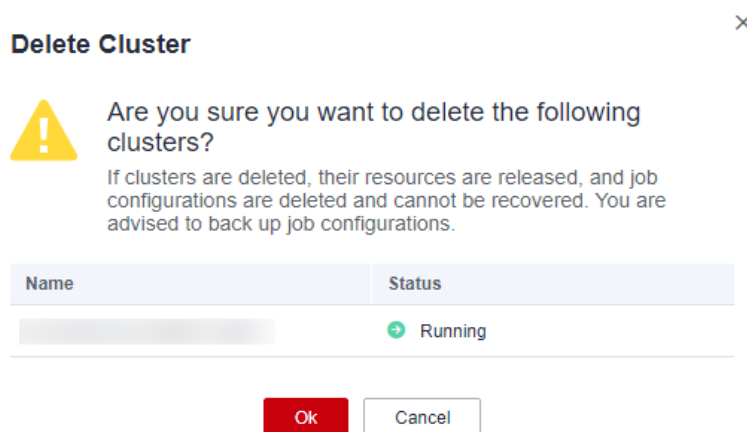


NOTE

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

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

Figure 4-7 Deleting a cluster



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

----End

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

Figure 4-8 Cluster list



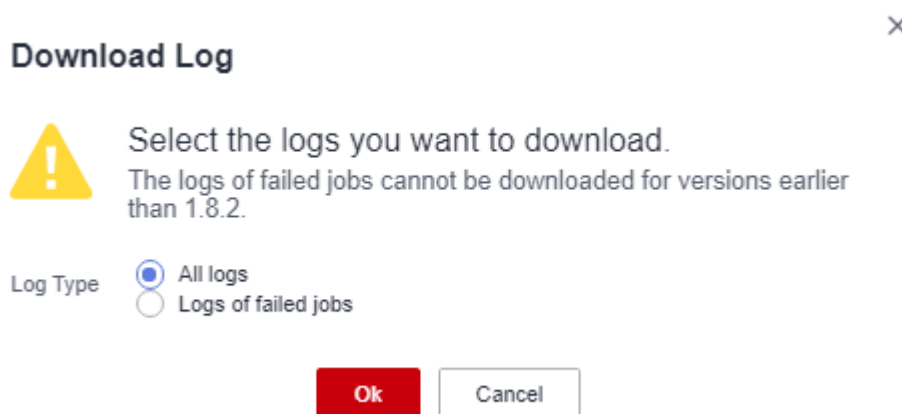
Name	Status	Internal Network Address	Public Network Address	Source	Enterprise Project	Operation
	Running			CDM	default	Job Management Bind EIP More

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 a cluster, click **More** in the **Operation** column, and select **Download Log** from the drop-down list.

Figure 4-9 Download Log



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

----End

4.4.6 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. Notifications generated by this function will not be charged.
 - User isolation: determines whether other users can operate the migration jobs or links in the cluster.
 - If this function is enabled, migration jobs and links in the cluster are isolated. Other IAM users of the HUAWEI CLOUD account cannot operate the jobs and links.
 - If this function is disabled, migration jobs and links in the cluster can be shared by users. All IAM users with the required permission in the HUAWEI CLOUD account can view and perform operations on the jobs and links in the cluster.

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

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

NOTE

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

Prerequisites

You have created a CDM cluster.

Viewing Basic Cluster Information

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

Figure 4-10 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 left navigation pane, choose **Cluster Management**.

Figure 4-11 Cluster list



 **NOTE**

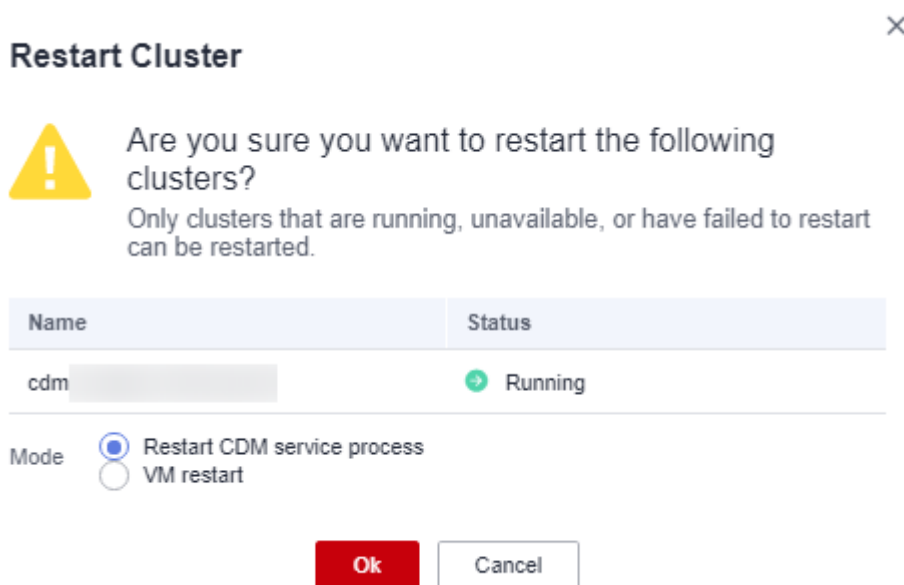
The **Source** column is displayed only when you access the **DataArts Migration** page from the DataArts Studio console.

Step 2 Click 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 4-12 Restarting a cluster



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

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

----End

Managing CDM Cluster Tags

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

Figure 4-13 Cluster list

The screenshot shows a table with columns: Name, Status, Internal Network Address, Public Network Address, Source, Enterprise Project, and Operation. A single cluster named 'cdm-ed87' is listed with a status of 'Running' and a green progress bar. The table is part of a web interface with various controls like 'Start', 'Restart', 'Authorize EP Check', and search filters.

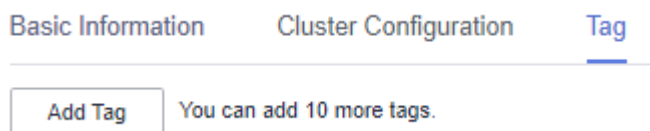
Name	Status	Internal Network Address	Public Network Address	Source	Enterprise Project	Operation
cdm-ed87	Running			CDM	default	Job Management Stop EP More

NOTE

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

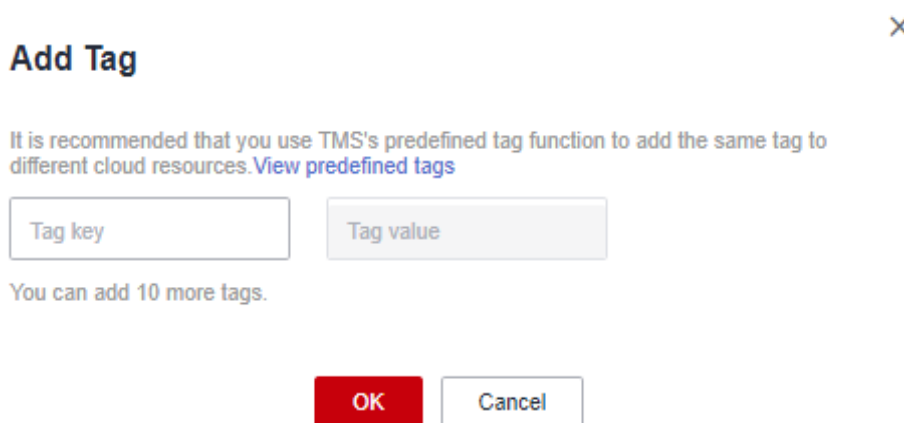
Step 2 Click a cluster name and then the **Tag** tab.

Figure 4-14 Modifying Cluster Configurations



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

Figure 4-15 Adding/Editing a tag

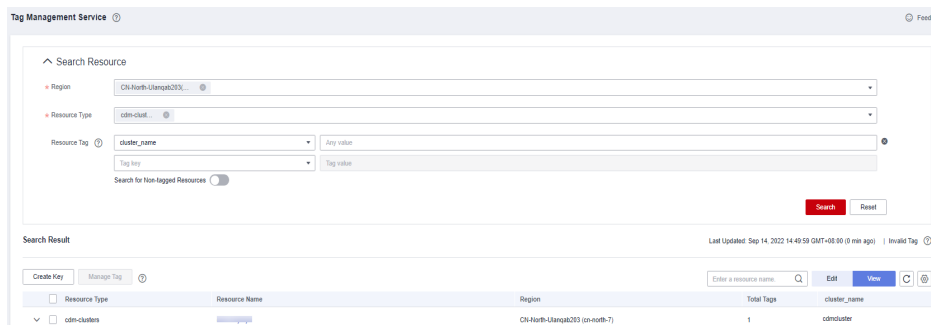


NOTE

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

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

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



----End

4.4.7 Viewing Metrics

4.4.7.1 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 4-6](#) lists the CDM metrics.

Table 4-6 CDM metrics

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

ID	Name	Description	Value Range	Monitored Object	Monitoring Period (Raw Data)
tomcat_heap_usage	Heap Memory Usage	Measures the heap memory usage of the physical server accommodating the monitored ECS, which is not accurate as that obtained on the monitored ECS. Unit: %	0.001% to 90%	Cloud Data Migration	1 minute
tomcat_connect	Tomcat Concurrent Connections	Measures the number of Tomcat concurrent connections on the physical server.	0 to 2,147,483,647	Cloud Data Migration	1 minute
tomcat_thread_count	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
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

4.4.7.2 Configuring Alarm Rules

Scenario

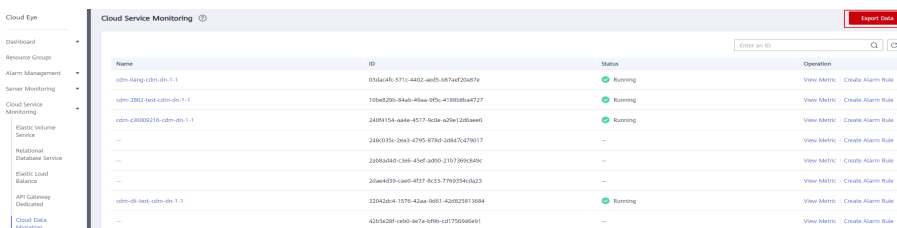
Set the alarm rules to customize the monitored objects and notification policies. Then, learn CDM running status in a timely manner.

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 4-16 Monitored CDM clusters



Name	ID	Status	Operation
cdm-12345-cdm-dm-1-1	035a4d1-371c-4402-a6b5-987af23ab7e	Running	View Metric Create Alarm Rule
cdm-23456-cdm-dm-1-1	10a8295-844b-49aa-9f5c-418880a4727	Running	View Metric Create Alarm Rule
cdm-34567-cdm-dm-1-1	2408154-a48e-4517-90de-a29a1388a4d	Running	View Metric Create Alarm Rule
---	2480035c-29a3-4795-878d-288a7a279017	---	View Metric Create Alarm Rule
---	2058a5d5-3368-43ef-a800-21073993b93c	---	View Metric Create Alarm Rule
---	2058a5d5-3368-43ef-a800-21073993b93c	---	View Metric Create Alarm Rule
---	2058a5d5-3368-43ef-a800-21073993b93c	---	View Metric Create Alarm Rule
cdm-45678-cdm-dm-1-1	320a12b-4-1576-42aa-9001-42082913684	Running	View Metric Create Alarm Rule
---	4259-28f-c480-4e7a-9f9b-cd17599d6491	---	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 .

----End

4.4.7.3 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

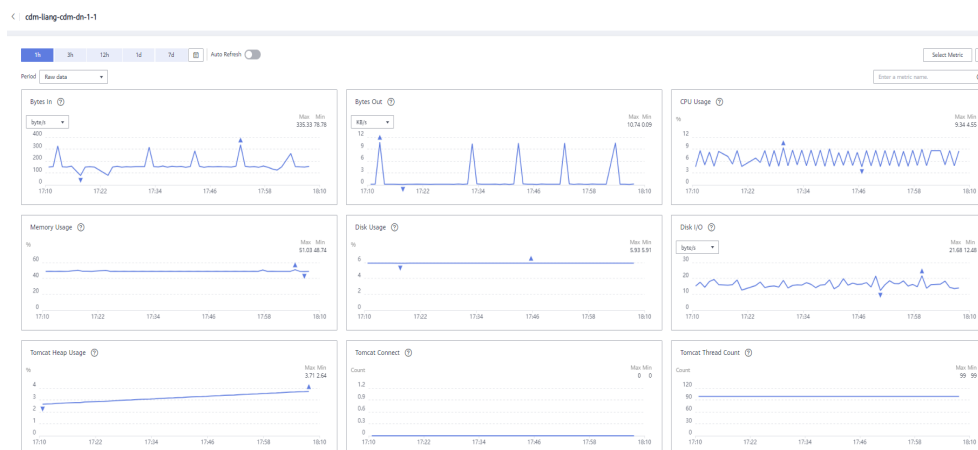
- The CDM cluster is running properly.
If a cluster fails to be restarted or is unavailable, its monitoring metrics are unavailable. You can view the monitored data only after the cluster is restarted or recovered.
- The cluster has been properly running for about 10 minutes.
The monitored data and graphs 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 4-17 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

4.5 Managing Links

4.5.1 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 [configuring routes](#). For details about how to configure security group rules, see [configuring security group rules](#).
 - The cloud service instance and the CDM cluster belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.
- You have obtained the URL and the account for accessing the data source. The account is granted with the read and write permissions for the data source.
- When using the Agent, you need to use the main account to grant the CDM operation permission to the sub-account.

Creating Links

Step 1 Log in to the management console and choose **Service List > Cloud Data Migration**. On the CDM console, choose **Cluster Management** in the left navigation pane. Locate the row that contains the target cluster and click **Job Management** in the **Operation** column. On the displayed **Links** page, click **Create Link**. On the displayed page shown in [Figure 4-18](#), 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 4-18 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 4-7](#) describes the link parameters.

Table 4-7 Link parameters

Connector	Description
<ul style="list-style-type: none"> Data Warehouse Service RDS for MySQL RDS for PostgreSQL RDS for SQL Server PostgreSQL Microsoft SQL Server SAP HANA 	Because the JDBC drivers used to connect to these relational databases are the same, the parameters to be configured are also the same and are described in Link to a Common Relational Database .
MySQL	For details about the parameters, see Link to an RDS for MySQL/MySQL Database .

Connector	Description
Oracle	For details about the parameters, see Link to an Oracle Database .
Database Sharding	For details about the parameters, see Link to a Database Shard .
HUAWEI CLOUD OBS	For details about the parameters, see Link to OBS .
<ul style="list-style-type: none"> • MRS HDFS • FusionInsight HDFS • Apache HDFS 	If the data source is HDFS of MRS, Apache Hadoop, or FusionInsight HD, see Link to HDFS .
<ul style="list-style-type: none"> • MRS HBase • FusionInsight HBase • Apache HBase 	If the data source is HBase of MRS, Apache Hadoop, or FusionInsight HD, see Link to HBase .
<ul style="list-style-type: none"> • MRS Hive • FusionInsight Hive • Apache Hive 	If the data source is Hive on MRS, Apache Hadoop, or FusionInsight HD, see Link to Hive .
CloudTable Service	If the data source is CloudTable, see Link to CloudTable .
<ul style="list-style-type: none"> • FTP • SFTP 	If the data source is an FTP or SFTP server, see Link to an FTP or SFTP Server .
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 Link to MongoDB .
Document Database Service (DDS)	If the data source is DDS, see Link to DDS .
<ul style="list-style-type: none"> • Redis • Distributed Cache Service 	If the data source is Redis or DCS, see Link to Redis/DCS .
<ul style="list-style-type: none"> • MRS Kafka • Apache Kafka 	If the data source is MRS Kafka or Apache Kafka, see Link to Kafka .
Cloud Search Service (CSS) Elasticsearch	If the data source is CSS or Elasticsearch, see Link to Elasticsearch/CSS .
Data Lake Insight	If the data source is DLI, see Link to DLI .

Connector	Description
DMS Kafka	If the data source is DMS Kafka, see Link to DMS Kafka .
Cassandra	If the data source is Cassandra, see Link to Cassandra .

 **NOTE**

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

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

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

----End

Managing Links

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

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

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

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

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

- Deleting a link: Click **Delete** in the **Operation** column to delete a link. Alternatively, select the links that are not used by any job and click **Delete Link** above the list to delete them.

- Editing the link: Click the link name or click **Edit** in the **Operation** column to access the page for modifying the link. When modifying the link, you need to enter the password for logging in to the data source again.
- Testing connectivity of the link: Click **Test Connectivity** in the **Operation** column.
- Viewing the JSON file of the link: In the **Operation** column, choose **More > View Link JSON** to view link parameters in JSON format.
- Editing the JSON file of the link: In the **Operation** column, choose **More > Edit Link JSON** to modify link parameters in JSON format.
- Viewing the backend link: Locate the row that contains a link and click **More** in the **Operation** column and select **View Backend Link** to view the backend link corresponding to the link.

----End

4.5.2 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 4-8](#).
- (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 4-8](#).

Table 4-8 Drivers

Relational Database Type	Driver Name	How to Obtain	Recommended Version
<ul style="list-style-type: none">• RDS for MySQL• MySQL	MySQL MyCAT	https://downloads.mysql.com/archives/c-j/	mysql-connector-java-5.1.48.jar

Relational Database Type	Driver Name	How to Obtain	Recommended Version
Oracle	ORACLE_6 ORACLE_7 ORACLE_8	Driver packages: https://www.oracle.com/database/technologies/appdev/jdbc-downloads.html Driver packages of historical versions: https://repo1.maven.org/maven2/com/oracle/database/jdbc/ojdbc8/12.2.0.1/	ojdbc8.jar for version 12.2.0.1 NOTE New versions (for example, Oracle Database 21c (21.3) drivers) are not supported. If they are used, the schema name cannot be obtained during job creation.
<ul style="list-style-type: none"> • RDS for PostgreSQL • PostgreSQL 	POSTGRESQL	https://mvnrepository.com/artifact/org.postgresql/postgresql	postgresql-42.1.4.jar for JDBC 4.2
<ul style="list-style-type: none"> • RDS for SQL Server • Microsoft SQL Server 	SQLServer	Driver packages: https://docs.microsoft.com/en-us/sql/connect/jdbc/download-microsoft-jdbc-driver-for-sql-server?view=sql-server-ver15 Driver packages of historical versions: https://docs.microsoft.com/en-us/sql/connect/jdbc/release-notes-for-the-jdbc-driver?view=sql-server-ver15#previous-releases	sqljdbc42.jar

Procedure

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

Figure 4-19 Uploading a driver

Driver Name	Driver Package Name	Driver Type	Description	Operation
MYSQL	None	Preset		Upload - Copy from SFTP
ORACLE_5	None	Preset	oracle + 12.1	Upload - Copy from SFTP
ORACLE_7	None	Preset	oracle + 12.1	Upload - Copy from SFTP
ORACLE_8	None	Preset	oracle + 12.1	Upload - Copy from SFTP
POSTGRESQL	None	Preset		Upload - Copy from SFTP
DB2	None	Preset		Upload - Copy from SFTP
SOLSERVER	None	Preset		Upload - Copy from SFTP
EDM	None	Preset		Upload - Copy from SFTP
MYCAT	None	Preset		Upload - Copy from SFTP
DM	None	Preset		Upload - Copy from SFTP

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

4.5.3 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 4-20 Scenario

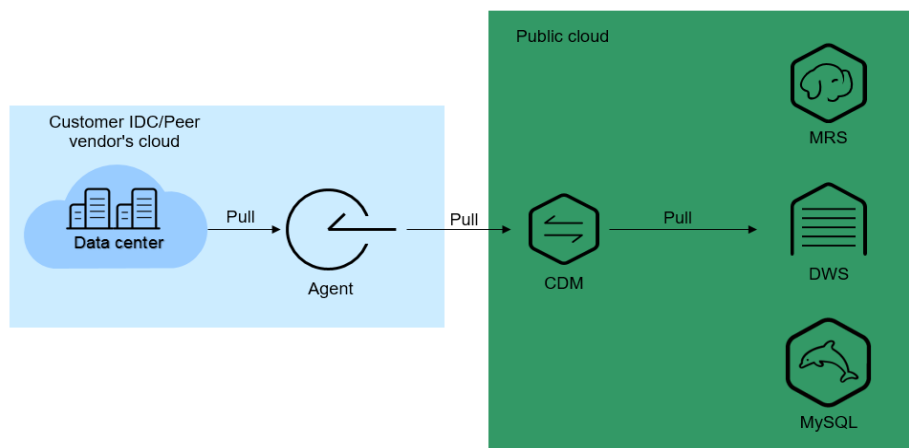
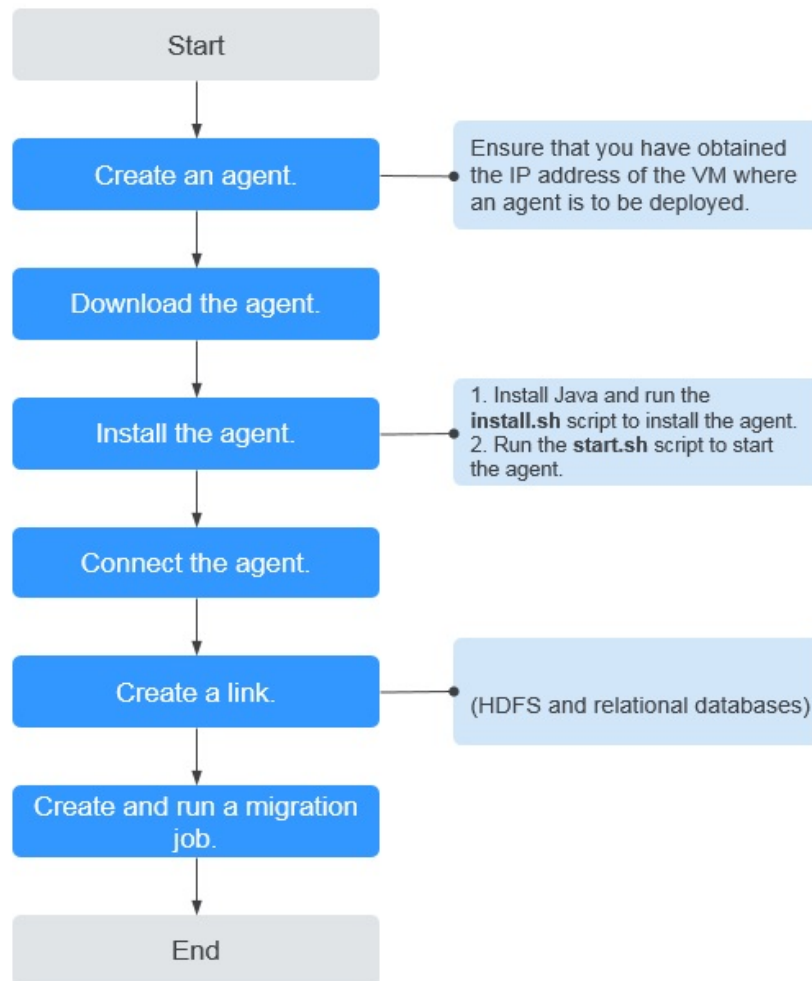


Figure 4-21 shows the process of using an agent.

Figure 4-21 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 4-22 Creating an agent

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

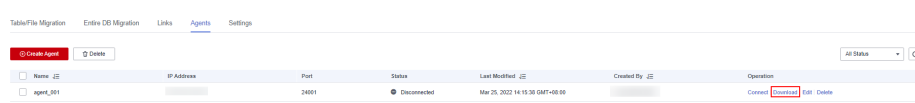
Step 2 Click **OK**. On the **Agent Management** page, view the created agent.

----End

Installing and Starting an Agent

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

Figure 4-23 Downloading an agent



Name	IP Address	Port	Status	Last Modified	Created By	Operation
agent_001		2801	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 Run the following command as user **root** to change the owner and group of the driver uploaded to the **/server/jdbc** directory to **Ruby**:

```
chown Ruby.Ruby * -R
```

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

```
su Ruby
```

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

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

```
ps -ef | grep cdm
```

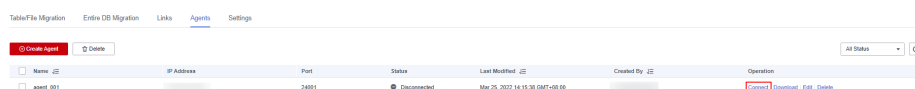
If the command output contains the running agent process, the agent process has been started.

----End

Connecting to an Agent

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

Figure 4-24 Connecting to an agent



Name	IP Address	Port	Status	Last Modified	Created By	Operation
agent_001		2801	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

4.5.4 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 4-25](#) for details.

Figure 4-25 Comparison before and after using the cluster configurations

The figure illustrates the transition from a standard Hadoop link configuration form to one that utilizes pre-defined cluster configurations. On the left, the 'Use Cluster Config' option is set to 'No', and the 'Authentication Method' and 'Run Mode' are manually selected. On the right, after selecting a cluster configuration, the 'Use Cluster Config' option is set to 'Yes', and the 'Authentication Method' and 'Run Mode' are automatically populated with the values from the selected cluster configuration. A red arrow indicates the direction of this transition.

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 4-9 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"> • MRS HDFS • MRS HBase • MRS Hive 	<p>For clusters of MRS 3.x:</p> <ol style="list-style-type: none"> 1. Log in to FusionInsight Manager. 2. Choose Cluster > <i>Name of the desired cluster</i> > Dashboard > More > Download Client. 3. In the dialog box that is displayed, select Configuration Files Only. The platform type must be the same as that on the server. Click OK to download the configuration file to the local host. 4. Obtain the downloaded TAR package, which is the FusionInsight cluster configuration file. <p>For clusters of MRS 2.x or earlier:</p> <ol style="list-style-type: none"> 1. Log in to the MRS console. 2. Choose Clusters > Active Clusters and click a cluster name to go to the cluster details page. Click the Components tab. 3. Click Download Client. Set Client Type to Only configuration files, set Download To to Server or Remote host, customize the client path, and click OK to generate the client configuration file. 4. Save the generated configuration file to a local path. <p>See MRS documentation for details.</p>	<p>For clusters of MRS 3.x:</p> <ol style="list-style-type: none"> 1. Log in to FusionInsight Manager. 2. Choose System > Permission > User, locate the row that contains the target user, and choose More > Download Authentication Credential to download the authentication credential file. 3. Obtain the downloaded TAR package, which is the keytab file of the FusionInsight cluster. <p>For clusters of MRS 2.x or earlier:</p> <ol style="list-style-type: none"> 1. Log in to MRS Manager and click System. In the Permission area, click Manage User. 2. In the row of the user for whom you want to export the keytab file, choose More > Download authentication credential to download the authentication file. After the file is automatically generated, save it to a specified path and keep it properly. <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"> • FusionInsight HDFS • FusionInsight HBase • FusionInsight Hive 	<ol style="list-style-type: none"> 1. Log in to FusionInsight Manager. 2. Choose Cluster > <i>Name of the desired cluster</i> > Dashboard > More > Download Client. 3. In the dialog box that is displayed, select Configuration Files Only. The platform type must be the same as that on the server. Click OK to download the configuration file to the local host. 4. Obtain the downloaded TAR package, which is the FusionInsight cluster configuration file. <p>See the FusionInsight documentation for details.</p>	<ol style="list-style-type: none"> 1. Log in to FusionInsight Manager. 2. Choose System > Permission > User, locate the row that contains the target user, and choose More > Download Authentication Credential to download the authentication credential file. 3. Obtain the downloaded TAR package, which is the keytab file of the FusionInsight cluster. <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"> • Apache HDFS • Apache HBase • Apache Hive 	<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"> • HDFS needs to compress the following files into a .zip package without the directory format: <ul style="list-style-type: none"> - hosts - core-site.xml - hdfs-site.xml - yarm-site.xml - mapred-site.xml - krb5.conf (optional, for clusters in security mode) • HBase needs to compress the following files into a .zip package without the directory format: <ul style="list-style-type: none"> - hosts - core-site.xml - hdfs-site.xml - yarm-site.xml - mapred-site.xml - hbase-site.xml - krb5.conf (optional, for clusters in security mode) • Hive needs to compress the following files into a .zip package without the directory format: <ul style="list-style-type: none"> - hosts - core-site.xml - hdfs-site.xml - yarm-site.xml 	<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"> 1. Rename the user's authentication credential file as user.keytab. 2. Compress the user.keytab file into a .zip package without the directory format: user.keytab.zip.

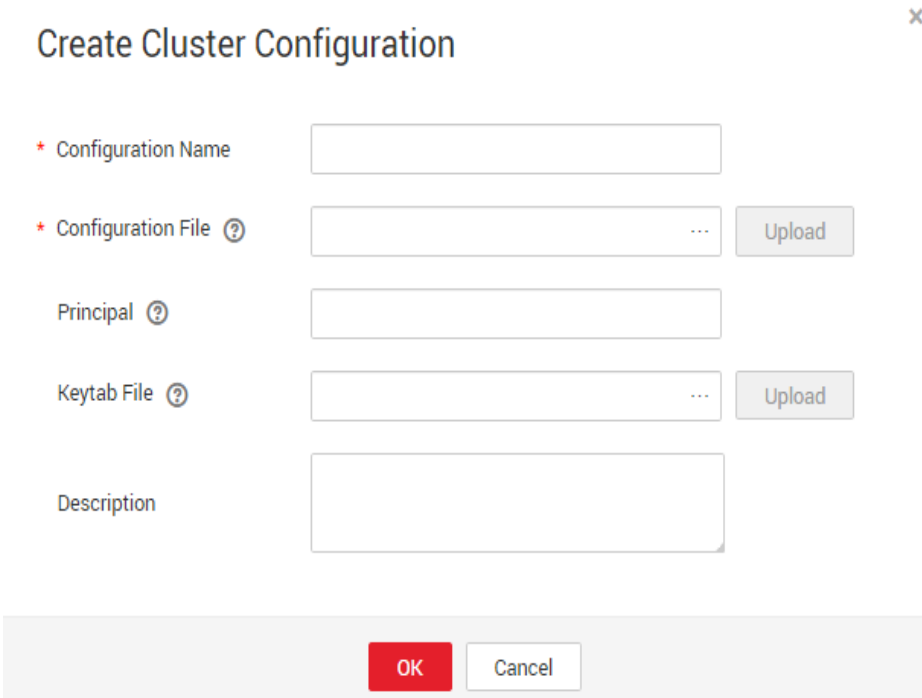
Hadoop Link	Obtaining the Cluster Configuration File	Obtaining the Keytab File
	<ul style="list-style-type: none">- mapred-site.xml- hive-site.xml- hivemetastore-site.xml- krb5.conf (optional, for clusters in security mode)	

NOTE

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

Procedure

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

Figure 4-26 Creating cluster configurations

Create Cluster Configuration ×

* Configuration Name

* Configuration File ? ...

Principal ?

Keytab File ? ...

Description

- **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 4-27 Use Cluster Config

* Name

* Connector

* Hadoop Type

* Authentication Method

* Run Mode

Use Cluster Config Yes No

Cluster Config Name

Show Advanced Attributes

4.5.5 Link to a Common Relational Database

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

Prerequisites

You have uploaded required drivers by following the instructions in [Managing Drivers](#).

Parameters for a link to a common relational database

[Table 4-10](#) lists the link parameters.

Table 4-10 Parameters for a link to a common relational database

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

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

Parameter	Description	Example Value
Link Attributes	<p>(Optional) Click Add to add the JDBC connector attributes of multiple specified data sources. For details, see the JDBC connector document of the corresponding database.</p> <p>The following are some examples:</p> <ul style="list-style-type: none">• connectTimeout=360000 and socketTimeout=360000: When a large amount of data needs to be migrated or the entire table is retrieved using query statements, the migration fails due to connection timeout. In this case, you can customize the connection timeout interval (ms) and socket timeout interval (ms) to prevent failures caused by timeout.• useCursorFetch=false: By default, useCursorFetch is enabled, indicating that the JDBC connector communicates with relational databases using a binary protocol. Some third-party systems may have compatibility issues, causing migration time conversion errors. In this case, you can disable this function. Open-source MySQL databases support the useCursorFetch parameter, and you do not need to set this parameter.	sslmode=require
Reference Sign	(Optional) Delimiter between the names of the referenced tables or columns. For details, see the product documentation of the corresponding database.	'

4.5.6 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 4-11](#) lists the link parameters.

Table 4-11 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

Parameter	Description	Example Value
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 Select and select the agent created in Managing Agents .	-
backendData source	Enter the type of the backend database. Currently, only MySQL is supported.	MySQL
Data Source List	Enter the IP address, port number, database name, account name, and password of the backend database, and separate them with colons (:). That is, ip:port:dbs:username:password. You can leave username:password empty. In this case, the username and password are used. If there are multiple backend databases, ensure that the table structures are the same and use vertical bars () to separate data sources. If the password contains a vertical bar () or colon (:), use a backslash (\) to escape the vertical bar. For example, 192.168.2.1:3306:cdm 192.168.2.2:3306:cdm:user:password indicates that the IP address of the first backend database is 192.168.2.1 , the port number is 3306 , the database name is cdm , and the account name and password are configured in <i>user</i> and <i>password</i> . The IP address of the second backend database is 192.168.2.2 , the port number is 3306 , the database name is cdm , the account name is user and the password is password .	192.168.2.1:3306:cdm 192.168.2.2:3306:cdm:user:password

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

4.5.7 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 4-12](#) lists the parameters for a MyCAT link.

Table 4-12 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 Select 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	-

Parameter	Description	Example Value
Use Local API	(Optional) Whether to use the local API of the database for acceleration. When you create a link, CDM automatically enables the local_infile system variable of the MySQL database to enable the LOAD DATA function, which accelerates data import to the MySQL database. If CDM fails to enable this function, contact the database administrator to enable the local_infile system variable. Alternatively, set Use Local API to No to disable API acceleration.	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
backendPassword	Password of the MyCAT backend database	-
Fetch Size	(Optional) Displayed when you click Show Advanced Attributes . Number of rows obtained by each request. Set this parameter based on the data source and the job's data size. If the value is either too large or too small, the job may run for a long time.	1000
Link Attributes	(Optional) Click Add to add the JDBC connector attributes of multiple specified data sources. For details, see the JDBC connector document of the corresponding database.	sslmode=require
Reference Sign	(Optional) Delimiter between the names of the referenced tables or columns. For details, see the product documentation of the corresponding database.	'

4.5.8 Link to a Dameng Database

When connecting CDM to a Dameng database, configure the parameters as described in [Table 4-13](#).

Table 4-13 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 Select 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
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 Select and select the agent created in Managing Agents .	-
Driver Version	Select a driver version that adapts to the database type.	-
Fetch Size	(Optional) Displayed when you click Show Advanced Attributes . Number of rows obtained by each request. Set this parameter based on the data source and the job's data size. If the value is either too large or too small, the job may run for a long time.	1000
SSL Encryption	(Optional) If you set this parameter to Yes , CDM can connect to the database (on-premises databases excluded) in SSL encryption mode. Security hardening has been performed on RDS for PostgreSQL. For this reason, when creating a link to RDS for PostgreSQL, set this parameter to Yes .	Yes

Parameter	Description	Example Value
Link Attributes	(Optional) Click Add to add the JDBC connector attributes of multiple specified data sources. For details, see the JDBC connector document of the corresponding database.	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.	'

4.5.9 Link to an RDS for MySQL/MySQL Database

[Table 4-14](#) lists the parameters for a link to a MySQL database.

Table 4-14 Parameters for a link to a MySQL database

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

Parameter	Description	Example Value
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 local_infile system variable of the MySQL database to enable the LOAD DATA function, which accelerates data import to the MySQL database. If this parameter is enabled, the date type that does not meet the format requirements will be stored as 0000-00-00. For details, visit the official MySQL website.</p> <p>If CDM fails to enable this function, contact the database administrator to enable the local_infile system variable. Alternatively, set Use Local API to No to disable API acceleration.</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 local_infile to ON to enable the LOAD DATA function.</p> <p>NOTE If local_infile on RDS is uneditable, it is the default parameter group. You need to create a parameter group, modify its values, and apply it to the RDS for MySQL instance. For details, see the <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 Select and select the agent created in Managing Agents .	-
local_infile Character Set	When using local_infile to import data to MySQL, you can configure the encoding format.	utf8
Driver Version	Select a driver version that adapts to the database type.	-
Fetch Size	<p>(Optional) Displayed when you click Show Advanced Attributes.</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

Parameter	Description	Example Value
Commit Size	<p>(Optional) Displayed when you click Show Advanced Attributes.</p> <p>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.</p>	-
Link Attributes	<p>(Optional) Click Add to add the JDBC connector attributes of multiple specified data sources. For details, see the JDBC connector document of the corresponding database.</p> <p>The following are some examples:</p> <ul style="list-style-type: none">• connectTimeout=360000 and socketTimeout=360000: When a large amount of data needs to be migrated or the entire table is retrieved using query statements, the migration fails due to connection timeout. In this case, you can customize the connection timeout interval (ms) and socket timeout interval (ms) to prevent failures caused by timeout.• tinyInt1isBit=false or mysql.bool.type.transform=false: By default, tinyInt1isBit is true, indicating that TINYINT(1) is processed as a bit, that is, Types.BOOLEAN, and 1 or 0 is read as true or false. As a result, the migration fails. In this case, you can set tinyInt1isBit to false to avoid migration failures.• useCursorFetch=false: By default, useCursorFetch is enabled, indicating that the JDBC connector communicates with relational databases using a binary protocol. Some third-party systems may have compatibility issues, causing migration time conversion errors. In this case, you can disable this function. Open-source MySQL databases support the useCursorFetch parameter, and you do not need to set this parameter.• allowPublicKeyRetrieval=true: By default, public key retrieval is disabled for MySQL databases. If TLS is unavailable and an RSA public key is used for encryption, connection to an MySQL database may fail. In this case, you can enable public key retrieval to avoid connection failures.	sslmode=require

Parameter	Description	Example Value
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

4.5.10 Link to an Oracle Database

[Table 4-15](#) lists the parameters for a link to an Oracle database.

Table 4-15 Parameters for a link to an Oracle database

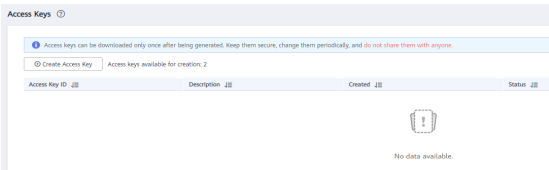
Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	oracle_link
Database Server	IP address or domain name of the database to connect	192.168.0.1
Port	Port of the database to connect	Default port: 1521
Connection Type	Oracle database connection type. The following options are available: <ul style="list-style-type: none"> • Service Name: Use SERVICE_NAME to connect to the Oracle database. • SID: Use SID to connect to the Oracle database. 	SID
Instance Name	Oracle instance ID, which is used to differentiate databases by instances. This parameter is available only when Connection Type is set to SID .	dbname
Database Name	Name of the database to connect This parameter is available only when Connection Type is set to Service Name .	dbname
Username	Username used for accessing the database This account must have the permissions required to read and write data tables and metadata.	cdm
Password	Password of the username	-
Use Agent	Whether to extract data from the data source through an agent	Yes

Parameter	Description	Example Value
Agent	Click Select and select the agent created in Managing Agents .	-
Oracle Version	Oracle database version. This parameter is available only for Oracle links. If java.sql.SQLException: Protocol violation is displayed, select another version.	Later than 12.1
Fetch Size	(Optional) Displayed when you click Show Advanced Attributes . Number of rows obtained by each request. Set this parameter based on the data source and the job's data size. If the value is either too large or too small, the job may run for a long time. A migration from the Oracle to DWS database may time out due to a long data write duration in the DWS database. In this case, reduce the value of Fetch Size for the Oracle database.	1000
Link Attributes	(Optional) Click Add to add the JDBC connector attributes of multiple specified data sources. For details, see the JDBC connector document of the corresponding database. The following are some examples: <ul style="list-style-type: none">• oracle.net.CONNECT_TIMEOUT=360000 and oracle.jdbc.ReadTimeout=360000: When a large amount of data needs to be migrated or the entire table is retrieved using query statements, the migration fails due to connection timeout. In this case, you can customize the connection timeout interval (ms) and read timeout interval (ms) to prevent failures caused by timeout.	sslmode=require
Reference Sign	(Optional) Delimiter between the names of the referenced tables or columns. For details, see the product documentation of the corresponding database.	'

4.5.11 Link to DLI

When connecting CDM to DLI, configure the parameters as described in [Table 4-16](#).

Table 4-16 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	<p>You need to create an access key for the current account and obtain an AK/SK pair.</p> <ol style="list-style-type: none"> Log in to the management console, move the cursor to the username in the upper right corner, and select My Credentials from the drop-down list. On the My Credentials page, choose Access Keys, and click Create Access Key. See Figure 4-28. <p>Figure 4-28 Clicking Create Access Key</p>  <ol style="list-style-type: none"> Click OK and save the access key file as prompted. The access key file will be saved to your browser's configured download location. Open the credentials.csv file to view Access Key Id and Secret Access Key. <p>NOTE</p> <ul style="list-style-type: none"> Only two access keys can be added for each user. To ensure access key security, the access key is automatically downloaded only when it is generated for the first time and cannot be obtained from the management console later. Keep them properly. 	-
Project ID	<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"> Register with and log in to the management console. Hover the cursor on the username in the upper right corner and select My Credentials from the drop-down list. On the My Credentials page, obtain the account name and account ID, and obtain the project ID from the project list. 	-

4.5.12 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 . [Table 4-17](#) 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 security group rules, see [configuring security group rules](#).
 - The MRS cluster and the DataArts Studio workspace belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

Table 4-17 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

Parameter	Description	Example Value
Manager IP	Floating IP address of MRS Manager. Click Select next to the Manager IP text box to select an MRS cluster. CDM automatically fills in the authentication information.	127.0.0.1
Authentication Method	Authentication method used for accessing MRS <ul style="list-style-type: none"> ● SIMPLE: Select this for non-security mode. ● KERBEROS: Select this for security mode. 	SIMPLE
HIVE Version	Set this to the Hive version on the server.	HIVE_3_X
Username	<p>If Authentication Method is set to KERBEROS, you must provide the username and password used for logging in to MRS Manager. If you need to create a snapshot when exporting a directory from HDFS, the user configured here must have the administrator permission on HDFS.</p> <p>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.</p> <p>NOTE</p> <ul style="list-style-type: none"> ● If the CDM cluster version is 2.9.0 or later and the MRS cluster version is 3.1.0 or later, the created user must have the permissions of the Manager_viewer role to create links on CDM. To perform operations on databases, tables, and columns of an MRS component, you also need to add the database, table, and column permissions of the MRS component to the user by following the instructions in the MRS documentation. ● If the CDM cluster version is earlier than 2.9.0 or the MRS cluster version is earlier than 3.1.0, the created user must have the permissions of Manager_administrator or System_administrator to create links on CDM. ● A user with only the Manager_tenant or Manager_auditor permission cannot create connections. 	cdm
Password	Password used for logging in to MRS Manager	-
OBS storage support	The server must support OBS storage. When creating a Hive table, you can store the table in OBS.	No

Parameter	Description	Example Value
Run Mode	<p>This parameter is used only when the Hive version is HIVE_3_X. Possible values are:</p> <ul style="list-style-type: none">• EMBEDDED: The link instance runs with CDM. This mode delivers better performance.• Standalone: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select STANDALONE or configure different agents. <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 Use Cluster Config is set to Yes. Select a cluster configuration that has been created.</p> <p>For details, see Managing Cluster Configurations.</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.

The following are some examples:

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

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

FusionInsight Hive

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

[Table 4-18](#) describes related parameters.

Table 4-18 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">• SIMPLE: Select this for non-security mode.• KERBEROS: Select this for security mode.	SIMPLE
HIVE Version	Hive version	HIVE_3_X
Username	Username used for logging in to FusionInsight Manager.	cdm
Password	Password used for logging in to FusionInsight Manager	-
OBS storage support	The server must support OBS storage. When creating a Hive table, you can store the table in OBS.	No

Parameter	Description	Example Value
Run Mode	<p>This parameter is used only when the Hive version is HIVE_3_X. Possible values are:</p> <ul style="list-style-type: none">• EMBEDDED: The link instance runs with CDM. This mode delivers better performance.• Standalone: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select STANDALONE or configure different agents. <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
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 Use Cluster Config is set to Yes. Select a cluster configuration that has been created.</p> <p>For details, see Managing Cluster Configurations.</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.

The following are some examples:

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

Apache Hive

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

Table 4-19 describes related parameters.

Table 4-19 Apache Hive link parameters

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	hivelink
URI	NameNode URI	hdfs:// hacluster
Hive Metastore	Hive metadata address. For details, see the hive.metastore.uris configuration item. Example: thrift://host-192-168-1-212:9083	-
Authentication Method	Authentication method used for accessing the cluster: <ul style="list-style-type: none">● SIMPLE: Select this for non-security mode.● KERBEROS: Select this for security mode.	SIMPLE
HIVE Version	Hive version	HIVE_3_X
IP and Host Name Mapping	If the Hadoop configuration file uses the host name, configure the mapping between the IP address and host name. Separate the IP addresses and host names by spaces and mappings by semicolons (;), carriage returns, or line feeds.	-
OBS storage support	The server must support OBS storage. When creating a Hive table, you can store the table in OBS.	No
Principal	When Authentication Method is set to KERBEROS , this parameter is mandatory. It is the username in the Kerberos security mode and can be obtained from the Hadoop administrator. The value of this parameter must be the same as that in the Keytab file.	-

Parameter	Description	Example Value
Keytab File	When Authentication Method is set to KERBEROS , a Keytab file must be uploaded. The Keytab file is an authentication credential and can be obtained from the Hadoop administrator. Before obtaining the keytab file, you need to change the password of this user at least once in the cluster. Otherwise, the downloaded keytab file may be unavailable. After a user password is changed, the exported keytab file becomes invalid, and you need to export a keytab file again.	-
Run Mode	This parameter is used only when the Hive version is HIVE_3_X . Possible values are: <ul style="list-style-type: none">• EMBEDDED: The link instance runs with CDM. This mode delivers better performance.• Standalone: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select STANDALONE or configure different agents. Note: The STANDALONE mode is used to solve the version conflict problem. If the connector versions of the source and destination ends of the same link are different, a JAR file conflict occurs. In this case, you need to place the source or destination end in the STANDALONE process to prevent the migration failure caused by the conflict.	EMBEDDED
Use Cluster Config	You can use the cluster configuration to simplify parameter settings for the Hadoop connection.	No
Cluster Config Name	This parameter is valid only when Use Cluster Config is set to Yes . Select a cluster configuration that has been created. For details, see Managing Cluster Configurations .	hive_01
Hive JDBC URL	URL for connecting to Hive JDBC. By default, anonymous users are used.	-

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

The following are some examples:

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

4.5.13 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 4-20](#).

NOTE

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

Table 4-20 MRS HBase link parameters

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	mrs_hbase_link
Manager IP	Floating IP address of MRS Manager. Click Select next to the Manager IP text box to select an MRS cluster. CDM automatically fills in the authentication information.	127.0.0.1
Username	<p>If Authentication Method is set to KERBEROS, you must provide the username and password used for logging in to MRS Manager. If you need to create a snapshot when exporting a directory from HDFS, the user configured here must have the administrator permission on HDFS.</p> <p>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.</p> <p>NOTE</p> <ul style="list-style-type: none">• If the CDM cluster version is 2.9.0 or later and the MRS cluster version is 3.1.0 or later, the created user must have the permissions of the Manager_viewer role to create links on CDM. To perform operations on databases, tables, and columns of an MRS component, you also need to add the database, table, and column permissions of the MRS component to the user by following the instructions in the MRS documentation.• If the CDM cluster version is earlier than 2.9.0 or the MRS cluster version is earlier than 3.1.0, the created user must have the permissions of Manager_administrator or System_administrator to create links on CDM.• A user with only the Manager_tenant or Manager_auditor permission cannot create connections.	cdm
Password	Password used for logging in to MRS Manager	-

Parameter	Description	Example Value
Authentication Method	Authentication method used for accessing the cluster: <ul style="list-style-type: none">• SIMPLE: Select this for non-security mode.• KERBEROS: Select this for security mode.	SIMPLE
HBase Version	HBase version	HBASE_2_X
Run Mode	Run mode of the HBase link. This parameter is used only when the HBase version is HBASE_2_X . <ul style="list-style-type: none">• EMBEDDED: The link instance runs with CDM. This mode delivers better performance.• Standalone: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select STANDALONE or configure different agents. Note: The STANDALONE mode is used to solve the version conflict problem. If the connector versions of the source and destination ends of the same link are different, a JAR file conflict occurs. In this case, you need to place the source or destination end in the STANDALONE process to prevent the migration failure caused by the conflict.	STANDALONE
Use Cluster Config	You can create cluster configurations on the Links page to simplify the configuration of Hadoop link parameters.	No
Cluster Config Name	This parameter is valid only when Use Cluster Config is set to Yes . Select a cluster configuration that has been created. For details, see Managing Cluster Configurations .	hbase_01

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

FusionInsight HBase

When connecting CDM to HBase of FusionInsight HD, configure the parameters as described in [Table 4-21](#).

Table 4-21 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">● SIMPLE: Select this for non-security mode.● KERBEROS: Select this for security mode.	Kerberos
HBase Version	HBase version	HBASE_2_X
Run Mode	Run mode of the HBase link. This parameter is used only when the HBase version is HBASE_2_X . <ul style="list-style-type: none">● EMBEDDED: The link instance runs with CDM. This mode delivers better performance.● Standalone: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select STANDALONE or configure different agents. Note: The STANDALONE mode is used to solve the version conflict problem. If the connector versions of the source and destination ends of the same link are different, a JAR file conflict occurs. In this case, you need to place the source or destination end in the STANDALONE process to prevent the migration failure caused by the conflict.	STANDALONE
Use Cluster Config	You can use the cluster configuration to simplify parameter settings for the Hadoop connection.	No

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

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

Apache HBase

When connecting CDM to HBase of Apache Hadoop, configure the parameters as described in [Table 4-22](#).

Table 4-22 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"> ● SIMPLE: Select this for non-security mode. ● KERBEROS: Select this for security mode. 	Kerberos
Principal	When Authentication Method is set to KERBEROS , this parameter is mandatory. It is the username in the Kerberos security mode and can be obtained from the Hadoop administrator. The value of this parameter must be the same as that in the Keytab file.	-

Parameter	Description	Example Value
Keytab File	When Authentication Method is set to KERBEROS , a Keytab file must be uploaded. The Keytab file is an authentication credential and can be obtained from the Hadoop administrator. Before obtaining the keytab file, you need to change the password of this user at least once in the cluster. Otherwise, the downloaded keytab file may be unavailable. After a user password is changed, the exported keytab file becomes invalid, and you need to export a keytab file again.	-
IP and Host Name Mapping	If the configuration file uses the host name, configure the mapping between the IP address and host name. Separate the IP addresses and host names by spaces and mappings by semicolons (;), carriage returns, or line feeds.	10.3.6.9 hostname01 10.4.7.9 hostname02
HBase Version	HBase version	HBASE_2_X
Run Mode	Run mode of the HBase link. This parameter is used only when the HBase version is HBASE_2_X . <ul style="list-style-type: none">• EMBEDDED: The link instance runs with CDM. This mode delivers better performance.• Standalone: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select STANDALONE or configure different agents. Note: The STANDALONE mode is used to solve the version conflict problem. If the connector versions of the source and destination ends of the same link are different, a JAR file conflict occurs. In this case, you need to place the source or destination end in the STANDALONE process to prevent the migration failure caused by the conflict.	STANDALONE
Use Cluster Config	You can use the cluster configuration to simplify parameter settings for the Hadoop connection.	No

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

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

4.5.14 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 4-23](#).

NOTE

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

Table 4-23 MRS HDFS link parameters

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	mrs_hdfs_link
Manager IP	Floating IP address of MRS Manager. Click Select next to the Manager IP text box to select an MRS cluster. CDM automatically fills in the authentication information.	127.0.0.1
Username	<p>If Authentication Method is set to KERBEROS, you must provide the username and password used for logging in to MRS Manager. If you need to create a snapshot when exporting a directory from HDFS, the user configured here must have the administrator permission on HDFS.</p> <p>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.</p> <p>NOTE</p> <ul style="list-style-type: none"> • If the CDM cluster version is 2.9.0 or later and the MRS cluster version is 3.1.0 or later, the created user must have the permissions of the Manager_viewer role to create links on CDM. To perform operations on databases, tables, and columns of an MRS component, you also need to add the database, table, and column permissions of the MRS component to the user by following the instructions in the MRS documentation. • If the CDM cluster version is earlier than 2.9.0 or the MRS cluster version is earlier than 3.1.0, the created user must have the permissions of Manager_administrator or System_administrator to create links on CDM. • A user with only the Manager_tenant or Manager_auditor permission cannot create connections. 	cdm
Password	Password used for logging in to MRS Manager	-

Parameter	Description	Example Value
Authentication Method	Authentication method used for accessing MRS <ul style="list-style-type: none">● SIMPLE: Select this for non-security mode.● KERBEROS: Select this for security mode.	SIMPLE
Run Mode	Run mode of the HDFS link. The options are as follows: <ul style="list-style-type: none">● EMBEDDED: The link instance runs with CDM. This mode delivers better performance.● STANDALONE: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select STANDALONE or configure different agents. Note: The STANDALONE mode is used to solve the version conflict problem. If the connector versions of the source and destination ends of the same link are different, a JAR file conflict occurs. In this case, you need to place the source or destination end in the STANDALONE process to prevent the migration failure caused by the conflict.● Agent: The link instance runs on an agent. If Agent is not used, and the CDM cluster connects to two or more clusters with Kerberos authentication enabled and the same realm, only one cluster can be connected in EMBEDDED mode, and the other clusters must be in STANDALONE mode.	STANDALONE
Agent	Click Select and select the agent created in Connecting to an Agent . This parameter is displayed when Run Mode is set to Agent .	-
Use Cluster Config	You can use the cluster configuration to simplify parameter settings for the Hadoop connection.	No
Cluster Config Name	This parameter is valid only when Use Cluster Config is set to Yes . Select a cluster configuration that has been created. For details, see Managing Cluster Configurations .	hdfs_01

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

FusionInsight HDFS

When connecting CDM to HDFS of FusionInsight HD, configure the parameters as described in [Table 4-24](#).

Table 4-24 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">● SIMPLE: Select this for non-security mode.● KERBEROS: Select this for security mode.	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">• EMBEDDED: The link instance runs with CDM. This mode delivers better performance.• STANDALONE: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select STANDALONE or configure different agents. Note: The STANDALONE mode is used to solve the version conflict problem. If the connector versions of the source and destination ends of the same link are different, a JAR file conflict occurs. In this case, you need to place the source or destination end in the STANDALONE process to prevent the migration failure caused by the conflict.• Agent: The link instance runs on an agent.	STANDALONE
Agent	Click Select and select the agent created in Connecting to an Agent . This parameter is displayed when Run Mode is set to Agent .	-
Use Cluster Config	You can use the cluster configuration to simplify parameter settings for the Hadoop connection.	No
Cluster Config Name	This parameter is valid only when Use Cluster Config is set to Yes . Select a cluster configuration that has been created. For details, see Managing Cluster Configurations .	hdfs_01

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

Apache HDFS

When connecting CDM to HDFS of Apache Hadoop, configure the parameters as described in [Table 4-25](#).

Table 4-25 Apache HDFS link parameters

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	hadoop_hdfs_link
URI	NameNode URI You can enter hdfs://IP address of the NameNode instance:8020 .	hdfs:// IP :8020
Authentication Method	Authentication method used for accessing the cluster: <ul style="list-style-type: none">● SIMPLE: Select this for non-security mode.● KERBEROS: Select this for security mode.	KERBEROS
Principal	When Authentication Method is set to KERBEROS , this parameter is mandatory. It is the username in the Kerberos security mode and can be obtained from the Hadoop administrator. The value of this parameter must be the same as that in the Keytab file.	-
Keytab File	When Authentication Method is set to KERBEROS , a Keytab file must be uploaded. The Keytab file is an authentication credential and can be obtained from the Hadoop administrator. Before obtaining the keytab file, you need to change the password of this user at least once in the cluster. Otherwise, the downloaded keytab file may be unavailable. After a user password is changed, the exported keytab file becomes invalid, and you need to export a keytab file again.	-

Parameter	Description	Example Value
Run Mode	<p>Run mode of the HDFS link. The options are as follows:</p> <ul style="list-style-type: none"> • EMBEDDED: The link instance runs with CDM. This mode delivers better performance. • STANDALONE: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select STANDALONE or configure different agents. Note: The STANDALONE mode is used to solve the version conflict problem. If the connector versions of the source and destination ends of the same link are different, a JAR file conflict occurs. In this case, you need to place the source or destination end in the STANDALONE process to prevent the migration failure caused by the conflict. • Agent: The link instance runs on an agent. For Apache HDFS, you can select Agent only if Authentication Method is set to SIMPLE. 	STANDALONE
IP and Host Name Mapping	<p>This parameter is used only when Run Mode is set to EMBEDDED or STANDALONE.</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>	10.1.6.9 hostname01 10.2.7.9 hostname02
Agent	If Run Mode is set to Agent , click Select and select the agent created in Connecting to an Agent .	-
Use Cluster Config	You can use the cluster configuration to simplify parameter settings for the Hadoop connection.	No
Cluster Config Name	<p>This parameter is valid only when Use Cluster Config is set to Yes. Select a cluster configuration that has been created.</p> <p>For details, see Managing Cluster Configurations.</p>	hdfs_01

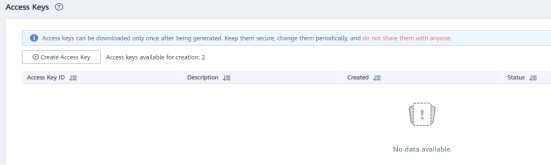
4.5.15 Link to OBS

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 4-26](#).

Table 4-26 Parameter description

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	obs_link
OBS Endpoint	You can obtain the endpoint by either of the following means: <ul style="list-style-type: none">To obtain the endpoint of an OBS bucket, go to the OBS console and click the bucket name to go to its details page.An endpoint is the request address for calling an API. Endpoints vary depending on services and regions. You can obtain endpoints from (Optional) Obtaining Authentication Information.	-
Port	Data transmission port. The HTTPS port number is 443 and the HTTP port number is 80.	443
OBS Bucket Type	Select a value from the drop-down list, generally, Object Storage .	Object Storage
AK	AK and SK are used to log in to the OBS server. You need to create an access key for the current account and obtain an AK/SK pair. To obtain an access key, perform the following steps: <ol style="list-style-type: none">Log in to the management console, move the cursor to the username in the upper right corner, and select My Credentials from the drop-down list.On the My Credentials page, choose Access Keys, and click Create Access Key. See Figure 4-29.	-

Parameter	Description	Example Value
SK	<p>Figure 4-29 Clicking Create Access Key</p>  <p>3. Click OK and save the access key file as prompted. The access key file will be saved to your browser's configured download location. Open the credentials.csv file to view Access Key Id and Secret Access Key.</p> <p>NOTE</p> <ul style="list-style-type: none"> Only two access keys can be added for each user. To ensure access key security, the access key is automatically downloaded only when it is generated for the first time and cannot be obtained from the management console later. Keep them properly. 	-

4.5.16 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 4-27](#).

Table 4-27 Parameter description

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

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

4.5.17 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 HUAWEI CLOUD. You are advised to use backup and restoration to migrate data from the third-party cloud Redis services to DCS.

When connecting CDM to an on-premises Redis database or DCS, configure the parameters as described in [Table 4-28](#).

Table 4-28 Parameter description

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	redis_link
Redis Deployment Method	Two deployment methods are available: <ul style="list-style-type: none">● Single: installation on a single-node system● Cluster: installation on a cluster● Proxy: installation using a proxy	Single
Redis Server List	List of MongoDB server addresses. Enter each address in the format of <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

4.5.18 Link to DDS

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

When connecting CDM to DDS, configure the parameters as described in [Table 4-29](#).

Table 4-29 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	-

4.5.19 Link to CloudTable

When connecting CDM to CloudTable, configure the parameters as described in [Table 4-30](#).

Table 4-30 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	cloudtable_link
ZK Link	Obtain this parameter value from the cluster management page of CloudTable.	cloudtable-cdm-zk1.cloudtable.com:2181,cloudtable-cdm-zk2.cloudtable.com:2181

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

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

4.5.20 Link to CloudTable OpenTSDB

When connecting CDM to CloudTable OpenTSDB, configure the parameters as described in [Table 4-31](#).

Table 4-31 Parameter description

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	TSDB_link
OpenTSDB Link	ZK link of OpenTSDB	opentsdb-sp8afz7bgbps5ur.cloudtable.com:4242

Parameter	Description	Example Value
Security Mode	Security or non-security mode If you select Security , 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: 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.	-
Username	Username for accessing CloudTable	admin
AK	AK and SK for accessing CloudTable. You need to create an access key for the current account and obtain an AK/SK pair. 1. Log in to the management console, move the cursor to the username in the upper right corner, and select My Credentials from the drop-down list. 2. On the My Credentials page, choose Access Keys , and click Create Access Key . See Figure 4-30 . Figure 4-30 Clicking Create Access Key 3. Click OK and save the access key file as prompted. The access key file will be saved to your browser's configured download location. Open the credentials.csv file to view Access Key Id and Secret Access Key .	-

Parameter	Description	Example Value
SK	NOTE <ul style="list-style-type: none">Only two access keys can be added for each user.To ensure access key security, the access key is automatically downloaded only when it is generated for the first time and cannot be obtained from the management console later. Keep them properly.	-

4.5.21 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 4-32](#).

Table 4-32 Parameter description

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

4.5.22 Link to Cassandra

Table 4-33 Parameter description

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

4.5.23 Link to Kafka

MRS Kafka

When connecting CDM to Kafka of MRS, configure the parameters as described in [Table 4-34](#).

Table 4-34 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

Parameter	Description	Example Value
Manager IP	Floating IP address of MRS Manager. Click Select next to the Manager IP text box to select an MRS cluster. CDM automatically fills in the authentication information.	-
Username	<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 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.</p> <p>NOTE</p> <ul style="list-style-type: none"> • If the CDM cluster version is 2.9.0 or later and the MRS cluster version is 3.1.0 or later, the created user must have the permissions of the Manager_viewer role to create links on CDM. To perform operations on databases, tables, and columns of an MRS component, you also need to add the database, table, and column permissions of the MRS component to the user by following the instructions in the MRS documentation. • If the CDM cluster version is earlier than 2.9.0 or the MRS cluster version is earlier than 3.1.0, the created user must have the permissions of Manager_administrator or System_administrator to create links on CDM. • A user with only the Manager_tenant or Manager_auditor permission cannot create connections. 	-
Password	Password used for logging in to MRS Manager	-
Authentication Method	<p>Authentication method used for accessing MRS</p> <ul style="list-style-type: none"> • SIMPLE: for non-security mode • KERBEROS: for security mode 	Yes

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

Apache Kafka

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

When connecting CDM to Kafka of Apache Hadoop, configure the parameters as described in [Table 4-35](#).

Table 4-35 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.

4.5.24 Link to DMS Kafka

When connecting CDM to DMS Kafka, configure the parameters as described in [Table 4-36](#).

Table 4-36 Parameter description

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

4.5.25 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 4-37](#).

Table 4-37 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 4-38](#).

Table 4-38 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

Parameter	Description	Example Value
Security Mode Authentication	Whether to enable security mode. If Security Mode has been enabled for the CSS cluster to be connected, set this parameter to Yes . Otherwise, set this to No .	Yes
Username	This parameter is displayed when Security Mode Authentication is set to Yes . It indicates the username used for connecting to CSS.	admin
Password	This parameter is displayed when Security Mode Authentication is set to Yes . It indicates the password used for connecting to CSS.	-
HTTPS Access	This parameter is displayed when Security Mode Authentication is set to Yes . This parameter specifies whether to enable HTTPS access. HTTPS access is more secure than HTTP access.	Yes

4.6 Managing Jobs

4.6.1 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 created links based on the instructions in [Creating Links](#).
- The CDM cluster can communicate with the data source.

Procedure

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

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

Figure 4-31 Creating a migration job

The screenshot shows the 'Job Configuration' page. At the top, there is a text input field for 'Job Name'. Below it, the page is divided into two main sections: 'Source Job Configuration' and 'Destination Job Configuration'. In the 'Source Job Configuration' section, there is a dropdown menu for 'Source Link Name' with the text 'Select a connector'. In the 'Destination Job Configuration' section, there is a dropdown menu for 'Destination Link Name' with the text 'Select a connector'. At the bottom of the page, there are two buttons: 'Cancel' and 'Next'.

Step 3 Select the source and destination links.

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

Step 4 Configure the source link parameters. **Figure 4-32** shows the job configurations for migrating MySQL to DWS.

Figure 4-32 Creating a job

The screenshot shows the 'Job Configuration' page with the following details:

- Job Name:** 'mysql2dws'
- Source Job Configuration:**
 - Source Link Name:** 'mysql_link' (with a 'Configuration Guide' link)
 - Use SQL Statement:** 'No' (with 'Yes' and 'No' buttons)
 - Schema/Table Space:** (empty text input)
 - Table Name:** (empty text input)
 - Show Advanced Attributes:** (link)
- Destination Job Configuration:**
 - Destination Link Name:** 'dws_link' (with a 'Configuration Guide' link)
 - Schema/Table Space:** (empty text input)
 - Auto Table Creation:** 'Non-auto Creation' (dropdown)
 - Table Name:** (empty text input)
 - Clear Data Before Import:** 'Do not clear' (dropdown)
 - Import Mode:** 'COPY' (dropdown)
 - Show Advanced Attributes:** (link)

 At the bottom, there are 'Cancel' and 'Next' buttons.

The parameters vary with data sources. For details about the job parameters of other types of data sources, see **Table 4-39** and **Table 4-40**.

Table 4-39 Source link parameter description

Migration Source	Description	Parameter Settings
OBS	Data can be extracted in CSV, JSON, or binary format. Data extracted in binary format is free from file resolution, which ensures high performance and is more suitable for file migration.	For details, see From OBS .
<ul style="list-style-type: none">• MRS HDFS• FusionInsight HDFS• Apache HDFS	HDFS data can be exported in CSV, Parquet, or binary format and can be compressed in multiple formats.	For details, see From HDFS .
<ul style="list-style-type: none">• MRS HBase• FusionInsight HBase• Apache HBase• CloudTable Service	Data can be exported from MRS, FusionInsight HD, open source Apache Hadoop HBase, or CloudTable. You need to know all column families and field names of HBase tables.	For details, see From HBase/CloudTable .
<ul style="list-style-type: none">• MRS Hive• FusionInsight Hive• Apache Hive	Data can be exported from Hive through the JDBC API. If the data source is Hive, CDM will automatically partition data using the Hive data partitioning file.	For details, see From Hive .
DLI	Data can be exported from DLI.	For details, see From DLI .
<ul style="list-style-type: none">• FTP• SFTP	FTP and SFTP data can be exported in CSV, JSON, or binary format.	For details, see From FTP/SFTP .

Migration Source	Description	Parameter Settings
<ul style="list-style-type: none"> • 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>Currently, data can only be exported from the HTTP URLs.</p>	For details, see From HTTP .
<ul style="list-style-type: none"> • Data Warehouse Service • RDS for MySQL • RDS for SQL Server • RDS for PostgreSQL 	Data can be exported from the cloud database services.	When data is exported from these data sources, CDM uses the JDBC API to extract data. The job parameters for the migration source are the same. For details, see From a Common Relational Database .
<ul style="list-style-type: none"> • FusionInsight LibrA 	Data can be exported from FusionInsight LibrA.	
<ul style="list-style-type: none"> • MySQL • PostgreSQL • Oracle • Microsoft SQL Server • SAP HANA • MyCAT • Database Sharding 	The non-cloud databases can be those created in the on-premises data center or deployed on ECSs, or database services on the third-party clouds.	
<ul style="list-style-type: none"> • MongoDB • Document Database Service 	Data can be exported from MongoDB or DDS.	For details, see From MongoDB/DDS .
Redis	Data can be exported from open source Redis.	For details, see From Redis .
<ul style="list-style-type: none"> • Apache Kafka • DMS Kafka • MRS Kafka 	Data can only be exported to Cloud Search Service (CSS).	For details, see From Kafka/DMS Kafka .
<ul style="list-style-type: none"> • Cloud Search Service • Elasticsearch 	Data can be exported from CSS or Elasticsearch.	For details, see From Elasticsearch or CSS .

Step 5 Configure job parameters for the migration destination based on [Table 4-40](#).

Table 4-40 Parameter description

Migration Destination	Description	Parameter Settings
OBS	Files (even in a large volume) can be batch migrated to OBS in CSV or binary format.	For details, see To OBS .
MRS HDFS	You can select a compression format when importing data to HDFS.	For details, see To HDFS .
MRS HBase CloudTable Service	Data can be imported to HBase. The compression algorithm can be set when a new HBase table is created.	For details, see To HBase/CloudTable .
MRS Hive	Data can be rapidly imported to MRS Hive.	For details, see To Hive .
DLI	Data can be imported to DLI.	For details, see To DLI .
<ul style="list-style-type: none"> • Data Warehouse Service • RDS for MySQL • RDS for SQL Server • RDS for PostgreSQL 	Data can be imported to cloud database services.	For details about how to use the JDBC API to import data, see To a Common Relational Database .
Document Database Service	Data can be imported to the DDS but cannot be imported to the local MongoDB.	For details, see To DDS .
Distributed Cache Service	Data can be imported to DCS in the String or Hashmap value type. Data cannot be imported to the local Redis.	For details, see To DCS .
Cloud Search Service (CSS)	Data can be imported to CSS.	For details, see To CSS .

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


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

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

Figure 4-33 Field mapping

Source Field				Destination Field		
Name	Example Value	Type	Operation	Name	Type	Operation
owner		string	↔ Q	owner	VARCHAR(10485760)	↔
table_name		string	↔ Q	table_name	VARCHAR(10485760)	↔

NOTE

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

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

Figure 4-34 Creating a converter

x

Create Converter

* Select a converter. Anonymization [Help](#)

* Reserve Start Length

* Reserve End Length

* Replace Character

Save
Back

CDM supports the following converters:

- **Anonymization**: hides key data in the character string.
For example, if you want to convert **12345678910** to **123****8910**, configure the parameters as follows:
 - Set **Reserve Start Length** to **3**.
 - Set **Reserve End Length** to **4**.
 - Set **Replace Character** to *****.
- **Trim** automatically deletes the spaces before and after the character string.
- **Reverse string** automatically reverses a character string. For example, reverse **ABC** into **CBA**.
- **Replace string** replaces the specified character string.
- **Expression conversion** uses the JSP expression language (EL) to convert the current field or a row of data. For details, see [Converting Fields](#).
- **Remove line break** deletes the newline characters, such as `\n`, `\r`, and `\r\n` from the field.

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

Figure 4-35 Task parameters

Configure Task

Retry if failed ?	<input type="text" value="Never"/>
Group ?	<input type="text" value="DEFAULT"/> + Add ✎ Edit 🗑 Delete
Schedule Execution	<input type="radio"/> Yes <input checked="" type="radio"/> No
Hide Advanced Attributes	
Concurrent Extractors ?	<input type="text" value="10"/>
Number of split retries ?	<input type="text" value="0"/>
Write Dirty Data ?	<input checked="" type="radio"/> Yes <input type="radio"/> No
Write Dirty Data Link ?	<input type="text" value="obs_link"/>
OBS Bucket ?	<input type="text"/> ⊖
Dirty Data Directory ?	<input type="text"/> ⊖
Max. error records in a single shard. ?	<input type="text" value="10"/>
Throttling ?	<input checked="" type="radio"/> Yes <input type="radio"/> No
byteRate(MB/s) ?	<input type="text" value="10"/>

Table 4-41 describes related parameters.

Table 4-41 Parameter description

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

Parameter	Description	Example Value
<p>Concurrent Extractors</p>	<p>Configure the number of tasks to be split from a CDM job.</p> <p>CDM migrates data through data migration jobs. It works in the following way:</p> <ol style="list-style-type: none"> 1. When data migration jobs are submitted, CDM splits each job into multiple tasks based on the Concurrent Extractors parameter in the job configuration. <p>NOTE Jobs for different data sources may be split based on different dimensions. Some jobs may not be split based on the Concurrent Extractors parameter.</p> <ol style="list-style-type: none"> 2. CDM submits the tasks to the running pool in sequence. The maximum number of tasks (defined by Maximum Concurrent Extractors) run concurrently. Excess tasks are queued. <p>By setting appropriate values for this parameter and the Maximum Concurrent Extractors parameter, you can accelerate migration.</p> <p>Configure the number of concurrent extractors based on the following rules:</p> <ol style="list-style-type: none"> 1. When data is to be migrated to files, CDM does not support multiple concurrent tasks. In this case, set a single process to extract data. 2. If each row of the table contains less than or equal to 1 MB data, data can be extracted concurrently. If each row contains more than 1 MB data, it is recommended 	<p>1</p>

Parameter	Description	Example Value
	<p>that data be extracted in a single thread.</p> <p>3. Set Concurrent Extractors for a job based on Maximum Concurrent Extractors for the cluster. It is recommended that Concurrent Extractors is less than Maximum Concurrent Extractors.</p>	
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 0 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 No.</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 Write Dirty Data is set to Yes.</p> <p>Only links to OBS support dirty data writes.</p>	obs_link
OBS Bucket	<p>This parameter is displayed only when Write Dirty Data Link 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 Write Dirty Data is set to Yes.</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 Write Dirty Data is set to Yes.</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

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

 **NOTE**

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

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

----End

4.6.2 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 4-36](#) describes the field mapping between the DWS tables created by CDM and source tables. For example, if you use CDM to migrate the Oracle database to DWS, CDM automatically creates a table on DWS and maps the **NUMBER(3,0)** field of the Oracle database to the **SMALLINT** field of DWS.

Figure 4-36 Field mapping in automatic table creation on DWS

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

Prerequisites

- You have created links according to [Creating Links](#).
- The CDM cluster can communicate with the data source.

Procedure

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

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

Figure 4-37 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](#)

Step 3 Configure the related parameters of the source database according to [Table 4-42](#).

Table 4-42 Parameter description

Source Database	Parameter	Description	Example Value
<ul style="list-style-type: none"> • DWS • FusionInsight LibrA • MySQL • PostgreSQL • SQL Server • Oracle • SAP HANA • MyCAT 	Schema/ Tablespace	<p>Name of the schema or tablespace from which data will be extracted. This parameter is displayed when Use SQL Statement is set to No. Click the icon next to the text box to go to the page for selecting a schema or directly enter a schema or tablespace.</p> <p>If the desired schema or tablespace is not displayed, check whether the login account has the permissions required to query metadata.</p>	schema

Source Database	Parameter	Description	Example Value
	WHERE Clause	<p>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.</p> <p>You can set a date macro variable to extract data generated on a specific date. For details, see Incremental Migration of Relational Databases.</p>	age > 18 and age <= 60
	Null in Partition Column	Whether a partition field can be null	Yes
Hive	Database Name	Name of the database to be migrated. The user configured in the source link must have the permission to read the database.	hivedb
HBase CloudTable	Start Time	<p>Start time (included). The format is <i>yyyy-MM-dd hh:mm:ss</i>. The dateformat time macro variable function is supported. Examples: 2017-12-31 20:00:00, \$ {dateformat(yyyy-MM-dd, -1, DAY)} 02:00:00, and \$ {dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}</p>	-

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

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

Step 4 Configure the related parameters, from [Table 4-43](#), for the destination cloud service.

Table 4-43 Destination job parameters

Source Database	Parameter	Description	Example Value
<ul style="list-style-type: none"> • DWS • FusionInsight LibrA • MySQL • PostgreSQL • SQL Server 	-	For details about the destination job parameters required for entire DB migration to a relational database, see To a Common Relational Database .	schema
MRS HIVE	-	For details about the destination job parameters required for entire DB migration to MRS HIVE, see To Hive .	hivedb
MRS HBase CloudTable	-	For details about the destination job parameters required for entire DB migration to MRS HBase or CloudTable, see To HBase/CloudTable .	Yes

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

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

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

Figure 4-38 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 4-44 describes related parameters.

Table 4-44 Task configuration parameters

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

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

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

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

----End

4.6.3 Source Job Parameters

4.6.3.1 From OBS

If the source link of a job is the [Link to OBS](#), configure the source job parameters based on [Table 4-45](#).

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

Table 4-45 Parameter description

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

Category	Parameter	Description	Example Value
	Source Directory/File	<p>This parameter is available only when Pull List File is set to No.</p> <p>Directory or file path from which data will be extracted. You can enter a maximum of 50 file paths. By default, the file paths are separated by vertical bars (). You can also customize a file separator. For details, see Migration of a List of Files.</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. For details, see Incremental Synchronization Using the Macro Variables of Date and Time.</p> <p>NOTE If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with <i>(Planned start time of the data development job - Offset)</i> rather than <i>(Actual start time of the CDM job - Offset)</i>.</p>	FROM/ example.csv
	File Format	<p>Format in which CDM parses data. The options are as follows:</p> <ul style="list-style-type: none"> ● CSV: Source files will be migrated to tables after being converted to CSV format. ● Binary: Files (even not in binary format) will be transferred directly. It is used for file copy. ● JSON: Source files will be migrated to tables after being converted to JSON format. 	CSV

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

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

Category	Parameter	Description	Example Value
	Compression Format	<p>This parameter is displayed only when File Format is set to CSV or JSON. The options are as follows:</p> <ul style="list-style-type: none"> • NONE: Files in all formats can be transferred. • GZIP: Only files in gzip format can be transferred. • ZIP: Only files in Zip format can be transferred. • TAR.GZ: Files in TAR.GZ format are transferred. 	NONE
	Compressed File Suffix	<p>This parameter is displayed when Compression Format is not NONE. This parameter specifies the extension of the files to be decompressed. The decompression operation is performed only when the file name extension is used in a batch of files. Otherwise, files are transferred in the original format. If you enter * or leave the parameter blank, all files are decompressed.</p>	*
	Source File Processing Method	<p>Operation performed on source files after the job completes.</p> <ul style="list-style-type: none"> • No action • Rename: After the job completes, the source files are renamed by appending usernames and timestamps as suffixes to the file names. • Delete: After the job completes, the source files are deleted. 	No action
	Start Job by Marker File	<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 Suspension Period.</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 Start Job by Marker File to Yes but there is no marker file in the source path, the job fails when the suspension period times out. If you set this parameter to 0 and there is no marker file in the source path, the job will fail immediately. Unit: second	10
	File Separator	File separator. If you enter multiple file paths in Source Directory/Files , CDM uses the file separator to identify files. The default value is .	
	Filter Type	Only paths or files that meet the filtering conditions are transferred. The options are None , Wildcard , and Regex . For details, see Incremental File Migration .	Wildcard
	Directory Filter	If you set Filter Type to Wildcard , enter a wildcard character to filter paths. The paths that meet the filtering condition are migrated. You can configure multiple paths separated by commas (,). NOTE If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with <i>(Planned start time of the data development job - Offset)</i> rather than <i>(Actual start time of the CDM job - Offset)</i> .	*input

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

Category	Parameter	Description	Example Value
	Maximum Timestamp	<p>If you set Filter Type to Time Filter, and specify a point in time for this parameter, only the files modified before the specified time are transferred. The time format must be <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, <code>#{timestamp(dateformat(yyyy-MM-dd HH:mm:ss))}</code> indicates that only the files whose modification time is earlier than the current time are migrated.</p> <p>NOTE If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with <i>(Planned start time of the data development job - Offset)</i> rather than <i>(Actual start time of the CDM job - Offset)</i>.</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"> • NONE: Export data without decrypting it. • AES-256-GCM: The AES 256-bit encryption algorithm is used to encrypt data. Currently, only the AES-256-GCM (NoPadding) encryption algorithm is supported. This parameter is used for encryption at the migration destination and decryption at the migration source. <p>For details, see Encryption and Decryption During File Migration.</p>	AES-256-GCM
	Disregard Non-existent Path or File	<p>If this is set to Yes, 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 Encryption is set to AES-256-GCM . The key consists of 64 hexadecimal numbers and must be the same as the DEK configured during encryption. If the decryption and encryption keys are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	DD0AE00D FEC78BF0 51BCFDA2 5BD4E320 DB0A7AC7 5A1F3FC3D 3C56A457 DCDC1B
	IV	This parameter is displayed only when Encryption is set to AES-256-GCM . The initialization vector consists of 32 hexadecimal numbers and must be the same as the IV configured during encryption. If the initialization vectors are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	5C91687BA 886EDCD1 2ACBC3FF1 9A3C3F
	MD5 File Extension	This parameter is displayed only when File Format is set to Binary . This parameter is used to check whether the files extracted by CDM are consistent with source files. For details, see MD5 Verification .	.md5

 **NOTE**

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

4.6.3.2 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 4-46](#).

Table 4-46 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	<p>This parameter is available only when Pull List File is set to No. Directory or file path from which data will be extracted.</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. For details, see Incremental Synchronization Using the Macro Variables of Date and Time.</p> <p>NOTE If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with <i>(Planned start time of the data development job - Offset)</i> rather than <i>(Actual start time of the CDM job - Offset)</i>.</p>	/user/cdm/
	File Format	<p>File format used when transferring data. The options are as follows:</p> <ul style="list-style-type: none"> • CSV: Source files will be migrated to tables after being converted to CSV format. • Binary: Files (even not in binary format) will be transferred directly. It is used for file copy. • Parquet: Source files will be migrated to tables after being converted to Parquet format. 	CSV

Category	Parameter	Description	Example Value
	Pull List File	<p>This parameter is displayed only when File Format is set to Binary.</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 Pull List File is set to Yes . You can select the OBS link where the list file is located.	OBS_test_link
	OBS Bucket of entries files	This parameter is available only when Pull List File is set to Yes . It indicates the name of the OBS bucket where the list file is located.	01
	Path/Directory of entries files	This parameter is available only when Pull List File is set to Yes . It indicates the absolute path or directory of the list file in the OBS bucket.	/0521/ Lists.txt
Advanced attributes	Line Separator	Line feed character in a file. By default, the system automatically identifies \n , \r , and \r\n . This parameter is displayed only when File Format is set to CSV .	\n
	Field Delimiter	Character used to separate fields in the file. To set the Tab key as the delimiter, set this parameter to \t . This parameter is displayed only when File Format is set to CSV .	,

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

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

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

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

Category	Parameter	Description	Example Value
	Encryption	<p>This parameter is displayed only when File Format is set to Binary.</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"> • NONE: Export data without decrypting it. • AES-256-GCM: The AES 256-bit encryption algorithm is used to encrypt data. Currently, only the AES-256-GCM (NoPadding) encryption algorithm is supported. This parameter is used for encryption at the migration destination and decryption at the migration source. <p>For details, see Encryption and Decryption During File Migration.</p>	AES-256-GCM
	DEK	<p>This parameter is displayed only when Encryption is set to AES-256-GCM. The key consists of 64 hexadecimal numbers and must be the same as the DEK configured during encryption. If the decryption and encryption keys are inconsistent, the system does not report an exception, but the decrypted data is incorrect.</p>	DD0AE00D FEC78BF0 51BCFDA25 BD4E320DB 0A7AC75A1 F3FC3D3C5 6A457DCD C1B

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

 NOTE

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

4.6.3.3 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 4-47](#).

 NOTE

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

Table 4-47 Parameter description

Parameter	Description	Example Value
Table Name	<p>Name of the HBase table that data will be exported from</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. For details, see Incremental Synchronization Using the Macro Variables of Date and Time.</p> <p>NOTE If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (<i>Planned start time of the data development job - Offset</i>) rather than (<i>Actual start time of the CDM job - Offset</i>).</p>	TBL_2
Column Families	(Optional) Column families to which the exported data belongs	CF1&CF2
Split Rowkey	(Optional) Whether to split a rowkey. The default value is No .	Yes
Rowkey Delimiter	(Optional) Delimiter used to split a rowkey. If this parameter is left empty, the rowkey will not be split.	
Start Time	<p>(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.</p> <p>This parameter can be set to a macro variable of date and time. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see Incremental Synchronization Using the Macro Variables of Date and Time.</p> <p>NOTE If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (<i>Planned start time of the data development job - Offset</i>) rather than (<i>Actual start time of the CDM job - Offset</i>).</p>	2019-01-01 20:00:00

Parameter	Description	Example Value
End Time	<p>(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.</p> <p>This parameter can be set to a macro variable of date and time. For details, see Incremental Synchronization Using the Macro Variables of Date and Time.</p> <p>NOTE If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with <i>(Planned start time of the data development job - Offset)</i> rather than <i>(Actual start time of the CDM job - Offset)</i>.</p>	2019-02-01 20:00:00

4.6.3.4 From Hive

If the source link of a job is the [Link to Hive](#), configure the source job parameters based on [Table 4-48](#).

Table 4-48 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. For details, see Incremental Synchronization Using the Macro Variables of Date and Time.</p> <p>NOTE If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with <i>(Planned start time of the data development job - Offset)</i> rather than <i>(Actual start time of the CDM job - Offset)</i>.</p>	TBL_E

Parameter	Description	Example Value
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"> • The HDFS mode shows good performance, but in this mode, you cannot use the WHERE condition to filter data or add new fields on the field mapping page. • The HDFS mode allows you to use the WHERE condition to filter data or add new fields on the field mapping page. 	HDFS
Partition Filter Criteria	<p>This parameter is displayed when you select the HDFS read mode and click Show Advanced Attributes.</p> <p>You can configure multiple values (separated by spaces) or a field value range. The time macro function is supported. For details, see Incremental Synchronization Using the Macro Variables of Date and Time.</p> <p>NOTE If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (<i>Planned start time of the data development job - Offset</i>) rather than (<i>Actual start time of the CDM job - Offset</i>).</p>	<ul style="list-style-type: none"> • Single/ Multi-value filtering: "\$ {dateformat(yyyyMMdd, -1, DAY)} \$ {dateformat(yyyyMMdd)}" • Filter by range: "\${value} >= \$ {dateformat(yyyyMMdd, -7, DAY)} && \${value} < \$ {dateformat(yyyyMMdd)}"

Parameter	Description	Example Value
WHERE Clause	<p>This parameter is displayed when you select the JDBC read mode and click Show Advanced Attributes.</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. For details, see Incremental Migration of Relational Databases.</p> <p>NOTE If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with <i>(Planned start time of the data development job - Offset)</i> rather than <i>(Actual start time of the CDM job - Offset)</i>.</p>	age > 18 and age <= 60

 **NOTE**

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

4.6.3.5 From DLI

If the source link of a job is the [Link to DLI](#), configure the source job parameters based on [Table 4-49](#).

Table 4-49 Parameter description

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

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

4.6.3.6 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 4-50](#).

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

Table 4-50 Parameter description

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

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

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

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

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

Category	Parameter	Description	Example Value
	Maximum Timestamp	<p>If you set Filter Type to Time Filter, and specify a point in time for this parameter, only the files modified before the specified time are transferred. The time format must be <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, <code>{timestamp(dateformat(yyyy-MM-dd HH:mm:ss))}</code> indicates that only the files whose modification time is earlier than the current time are migrated.</p> <p>NOTE If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with <i>(Planned start time of the data development job - Offset)</i> rather than <i>(Actual start time of the CDM job - Offset)</i>.</p>	2019-07-30 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"> • NONE: Export data without decrypting it. • AES-256-GCM: The AES 256-bit encryption algorithm is used to encrypt data. Currently, only the AES-256-GCM (NoPadding) encryption algorithm is supported. This parameter is used for encryption at the migration destination and decryption at the migration source. <p>For details, see Encryption and Decryption During File Migration.</p>	AES-256-GCM
	Disregard Non-existent Path or File	<p>If this is set to Yes, 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 Encryption is set to AES-256-GCM . The key consists of 64 hexadecimal numbers and must be the same as the DEK configured during encryption. If the decryption and encryption keys are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	DD0AE00D FEC78BF0 51BCFDA25 BD4E320DB 0A7AC75A1 F3FC3D3C5 6A457DCD C1B
	IV	This parameter is displayed only when Encryption is set to AES-256-GCM . The initialization vector consists of 32 hexadecimal numbers and must be the same as the IV configured during encryption. If the initialization vectors are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	5C91687BA 886EDCD12 ACBC3FF19 A3C3F
	MD5 File Extension	This parameter is displayed only when File Format is set to Binary . This parameter is used to check whether the files extracted by CDM are consistent with source files. For details, see MD5 Verification .	.md5

4.6.3.7 From HTTP

When the source link of a job is the HTTP link, configure the source job parameters based on [Table 4-51](#). Currently, data can only be exported from the HTTP URLs.

Table 4-51 Parameter description

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

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

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

4.6.3.8 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 4-52](#).

Table 4-52 Parameter description

Category	Parameter	Description	Example Value
Basic parameters	Use SQL Statement	Whether you can use SQL statements to export data from a relational database	No
	SQL Statement	When Use SQL Statement is set to Yes , enter an SQL statement here. CDM exports data based on the SQL statement. NOTE <ul style="list-style-type: none">• SQL statements can only be used to query data. Join and nesting are supported, but multiple query statements are not allowed, for example, select * from table a; select * from table b.• With statements are not supported.• Comments, such as -- and /*, are not supported.• Addition, deletion, and modification operations are not supported, including but not limited to the following:<ul style="list-style-type: none">• load data• delete from• alter table• create table• drop table• into outfile	select id,name from sqoop.user;

Category	Parameter	Description	Example Value
	Schema/ Tablespace	<p>Name of the schema or tablespace from which data will be extracted. This parameter is displayed when Use SQL Statement is set to No. Click the icon next to the text box to go to the page for selecting a schema or directly enter a schema or tablespace.</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>NOTE</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. The examples are as follows:</p> <ul style="list-style-type: none">• SCHEMA* indicates that all databases whose names starting with SCHEMA are exported.• *SCHEMA indicates that all databases whose names ending with SCHEMA are exported.• *SCHEMA* indicates that all databases whose names containing SCHEMA are exported.	SCHEMA_E

Category	Parameter	Description	Example Value
	Table Name	<p>Name of the table from which data will be extracted. This parameter is displayed when Use SQL Statement is set to No. Click the icon next to the text box to go to the page for selecting the table or directly enter a table name.</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. For details, see Incremental Synchronization Using the Macro Variables of Date and Time.</p> <p>NOTE If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with <i>(Planned start time of the data development job - Offset)</i> rather than <i>(Actual start time of the CDM job - Offset)</i>.</p> <p>NOTE The table name can contain wildcard characters (*), which is used to export all tables whose names start with a certain prefix or end with a certain suffix. The number and types of fields in the tables must be the same. The examples are as follows:</p> <ul style="list-style-type: none"> ● table* indicates that all tables whose names starting with table are exported. ● *table indicates that all tables whose names ending with table are exported. ● *table* indicates that all tables whose names containing table are exported. 	table

Category	Parameter	Description	Example Value
Advanced attributes	Partition Column	<p>This parameter is displayed when Use SQL Statement is set to No, indicating that a field used to split data during data extraction. CDM splits a job into multiple tasks based on this field and executes the tasks concurrently. Fields with data distributed evenly are used, such as the sequential number field.</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>NOTE</p> <ul style="list-style-type: none"> The following types of partition columns are supported: CHAR, VARCHAR, LONGVARCHAR, TINYINT, SMALLINT, INTEGER, BIGINT, REAL, FLOAT, DOUBLE, NUMERIC, DECIMAL, BIT, BOOLEAN, DATE, TIME, and TIMESTAMP. It is recommended that the partition column have an index. If the partition column type is CHAR, VARCHAR, or LONGVARCHAR, the column value cannot contain characters other than those in the ASCII character code table or Chinese characters. 	id
	Where Clause	<p>WHERE clause used to specify the data extraction range. This parameter is displayed when Use SQL Statement is set to No. If this parameter is not set, the entire table is extracted.</p> <p>You can set a date macro variable to extract data generated on a specific date. For details, see Incremental Migration of Relational Databases.</p> <p>NOTE</p> <p>If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with <i>(Planned start time of the data development job - Offset)</i> rather than <i>(Actual start time of the CDM job - Offset)</i>.</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.	-

Category	Parameter	Description	Example Value
	Minimum value of a split field	Specifies the minimum value of Job Split Field during data extraction.	-
	Maximum Split Field Value	Specifies the maximum value of Job Split Field during data extraction.	-
	Number of subjobs	Specifies the number of subjobs split from a job based on the data range specified by the minimum and maximum values of Job Split Field .	-
	Extract by Partition	<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 Table Partition to specify specific MySQL table partitions from which data is extracted.</p> <ul style="list-style-type: none"> • This function does not support non-partitioned tables. • This parameter is available only for RDS for PostgreSQL and RDS for MySQL. • The database user must have the SELECT permission on the system views dba_tab_partitions and dba_tab_subpartitions. 	No

4.6.3.9 From MySQL

If the source link of a job is the [Link to an RDS for MySQL/MySQL Database](#), configure the source job parameters based on [Table 4-53](#).

Table 4-53 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 Use SQL Statement is set to Yes, enter an SQL statement here. CDM exports data based on the SQL statement.</p> <p>NOTE</p> <ul style="list-style-type: none"> • SQL statements can only be used to query data. Join and nesting are supported, but multiple query statements are not allowed, for example, select * from table a; select * from table b. • With statements are not supported. • Comments, such as -- and /*, are not supported. • Addition, deletion, and modification operations are not supported, including but not limited to the following: <ul style="list-style-type: none"> • load data • delete from • alter table • create table • drop table • into outfile 	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 Use SQL Statement is set to No. Click the icon next to the text box to go to the page for selecting a schema or directly enter a schema or tablespace.</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>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 Use SQL Statement is set to No. Click the icon next to the text box to go to the page for selecting the table or directly enter a table name.</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. For details, see Incremental Synchronization Using the Macro Variables of Date and Time.</p> <p>This parameter can be set to a regular expression to export all databases that meet the rule. For example, if Table Name is set to <code>user_[0-9]{1,2}</code>, tables from user_0 to user_9 and from user_00 to user_99 are matched.</p> <p>NOTE</p> <p>If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with <i>(Planned start time of the data development job - Offset)</i> rather than <i>(Actual start time of the CDM job - Offset)</i>.</p>	table

Category	Parameter	Description	Example Value
Advanced attributes	Partition Column	<p>This parameter is displayed when Use SQL Statement is set to No, indicating that a field used to split data during data extraction. CDM splits a job into multiple tasks based on this field and executes the tasks concurrently. Fields with data distributed evenly are used, such as the sequential number field.</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>NOTE</p> <ul style="list-style-type: none"> The following types of partition columns are supported: CHAR, VARCHAR, LONGVARCHAR, TINYINT, SMALLINT, INTEGER, BIGINT, REAL, FLOAT, DOUBLE, NUMERIC, DECIMAL, BIT, BOOLEAN, DATE, TIME, and TIMESTAMP. It is recommended that the partition column have an index. If the partition column type is CHAR, VARCHAR, or LONGVARCHAR, the column value cannot contain characters other than those in the ASCII character code table or Chinese characters. 	id
	WHERE Clause	<p>WHERE clause used to specify the data extraction range. This parameter is displayed when Use SQL Statement is set to No. If this parameter is not set, the entire table is extracted.</p> <p>You can set a date macro variable to extract data generated on a specific date. For details, see Incremental Migration of Relational Databases.</p> <p>NOTE</p> <p>If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with <i>(Planned start time of the data development job - Offset)</i> rather than <i>(Actual start time of the CDM job - Offset)</i>.</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.	-

Category	Parameter	Description	Example Value
	Minimum value of a split field	Specifies the minimum value of Job Split Field during data extraction.	-
	Maximum Split Field Value	Specifies the maximum value of Job Split Field during data extraction.	-
	Number of subjobs	Specifies the number of subjobs split from a job based on the data range specified by the minimum and maximum values of Job Split Field .	-
	Extract by Partition	<p>When data is exported from a MySQL database, data can be extracted from each partition in the partitioned table. If this function is enabled, you can configure Table Partition to specify specific MySQL table partitions from which data is extracted.</p> <ul style="list-style-type: none"> • This function does not support non-partitioned tables. • The database user must have the SELECT permission on the system views dba_tab_partitions and dba_tab_subpartitions. 	No

 NOTE

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

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

4.6.3.10 From Oracle

If the source link of a job is the [Link to an Oracle Database](#), configure the source job parameters based on [Table 4-54](#).

Table 4-54 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 Use SQL Statement is set to Yes, enter an SQL statement here. CDM exports data based on the SQL statement.</p> <p>NOTE</p> <ul style="list-style-type: none"> • SQL statements can only be used to query data. Join and nesting are supported, but multiple query statements are not allowed, for example, select * from table a; select * from table b. • With statements are not supported. • Comments, such as -- and /*, are not supported. • Addition, deletion, and modification operations are not supported, including but not limited to the following: <ul style="list-style-type: none"> • load data • delete from • alter table • create table • drop table • into outfile 	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 Use SQL Statement is set to No. Click the icon next to the text box to go to the page for selecting a schema or directly enter a schema or tablespace.</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>NOTE</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"> • SCHEMA* indicates that all databases whose names starting with SCHEMA are exported. • *SCHEMA indicates that all databases whose names ending with SCHEMA are exported. • *SCHEMA* indicates that all databases whose names containing SCHEMA are exported. 	SCHEMA_E

Category	Parameter	Description	Example Value
	Table Name	<p>Name of the table from which data will be extracted. This parameter is displayed when Use SQL Statement is set to No. Click the icon next to the text box to go to the page for selecting the table or directly enter a table name.</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. For details, see Incremental Synchronization Using the Macro Variables of Date and Time.</p> <p>NOTE If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with <i>(Planned start time of the data development job - Offset)</i> rather than <i>(Actual start time of the CDM job - Offset)</i>.</p> <p>NOTE The table name can contain wildcard characters (*), which is used to export all tables whose names start with a certain prefix or end with a certain suffix. The number and types of fields in the tables must be the same. The examples are as follows:</p> <ul style="list-style-type: none"> ● table* indicates that all tables whose names starting with table are exported. ● *table indicates that all tables whose names ending with table are exported. ● *table* indicates that all tables whose names containing table are exported. 	table

Category	Parameter	Description	Example Value
Advanced attributes	Partition Column	<p>This parameter is displayed when Use SQL Statement is set to No, indicating that a field used to split data during data extraction. CDM splits a job into multiple tasks based on this field and executes the tasks concurrently. Fields with data distributed evenly are used, such as the sequential number field.</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>NOTE</p> <ul style="list-style-type: none"> The following types of partition columns are supported: CHAR, VARCHAR, LONGVARCHAR, TINYINT, SMALLINT, INTEGER, BIGINT, REAL, FLOAT, DOUBLE, NUMERIC, DECIMAL, BIT, BOOLEAN, DATE, TIME, and TIMESTAMP. It is recommended that the partition column have an index. If the partition column type is CHAR, VARCHAR, or LONGVARCHAR, the column value cannot contain characters other than those in the ASCII character code table or Chinese characters. 	id
	WHERE Clause	<p>WHERE clause used to specify the data extraction range. This parameter is displayed when Use SQL Statement is set to No. If this parameter is not set, the entire table is extracted.</p> <p>You can set a date macro variable to extract data generated on a specific date. For details, see Incremental Migration of Relational Databases.</p> <p>NOTE</p> <p>If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with <i>(Planned start time of the data development job - Offset)</i> rather than <i>(Actual start time of the CDM job - Offset)</i>.</p>	DS='\${dateformat(yyyy-MM-dd,-1,DAY)}'
	Null in Partition Column	Whether the partition column can contain null values	Yes

Category	Parameter	Description	Example Value
	Extract by Partition	<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 Table Partition to specify specific Oracle table partitions from which data is extracted.</p> <ul style="list-style-type: none"> This function does not support non-partitioned tables. The database user must have the SELECT permission on the system views dba_tab_partitions and dba_tab_subpartitions. 	No
	Table Partition	<p>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, P2.SUBP1.</p>	P0&P1&P2.SUBP1&P2.SUBP3
	Job Split Field	Used to split a job into multiple subjobs for concurrent execution.	-
	Minimum value of a split field	Specifies the minimum value of Job Split Field during data extraction.	-
	Maximum Split Field Value	Specifies the maximum value of Job Split Field during data extraction.	-
	Number of subjobs	Specifies the number of subjobs split from a job based on the data range specified by the minimum and maximum values of Job Split Field .	-

 NOTE

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

4.6.3.11 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 4-55](#).

Table 4-55 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>This parameter can be set to a regular expression to export all databases that meet the rule. For example, if Table Name is set to <i>user_[0-9]{1,2}</i>, tables from user_0 to user_9 and from user_00 to user_99 are matched.</p>	SCHEMA_E

Category	Parameter	Description	Example Value
	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. For details, see Incremental Synchronization Using the Macro Variables of Date and Time.</p> <p>This parameter can be set to a regular expression to export all databases that meet the rule.</p> <p>NOTE If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with <i>(Planned start time of the data development job - Offset)</i> rather than <i>(Actual start time of the CDM job - Offset)</i>.</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. For details, see Incremental Migration of Relational Databases.</p> <p>NOTE If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with <i>(Planned start time of the data development job - Offset)</i> rather than <i>(Actual start time of the CDM job - Offset)</i>.</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)}``

4.6.3.12 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 4-56](#).

Table 4-56 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"> 1. Filter by expression: <code>{'last_name': 'Smith'}</code> indicates that all files whose <code>last_name</code> value is Smith are queried. 2. Filter by parameter: <code>{ x : "john" }, { z : 1 }</code> indicates that all <code>z</code> fields whose <code>x</code> is john are queried. 3. Filter by condition: <code>{ "field" : { \$gt: 5 } }</code> indicates that the field values greater than 5 are queried. 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 ts field are queried. 	<code>{'last_name': 'Smith'}</code>

4.6.3.13 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 4-57](#).

Table 4-57 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"> • String: without column name, such as value1,value2 • Hash: with column name, such as column1=value1,column2=value2 	String
Key Delimiter	Character used to separate table names and column names of a relational database	-

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

4.6.3.14 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 4-58](#).

Table 4-58 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"> • Latest: Maximum offset, indicating that the latest data will be extracted. • Earliest: Minimum offset, indicating that the earliest data will be extracted. • Submitted: data that has been submitted • Time Range: data within a specified time range 	Latest
Permanent Running	Whether a job runs permanently.	Yes
Consumer Group ID	Consumer group ID If you export data from DMS Kafka, enter any value for Kafka Platinum but a valid consumer group ID for Kafka Basic.	sumer-group

Parameter	Description	Example Value
Data Format	Format used for parsing data. The options are as follows: <ul style="list-style-type: none"> • Binary: Data is transferred directly. It is not converted to another format. This setting is suitable for file migration. • CSV: Source data will be migrated after being converted in CSV format. • JSON: Source data will be migrated after being converted in JSON format. • CDC (DRS_JSON): Source data will be migrated after being converted in DRS_JSON format. 	Binary
Field Delimiter	The default value is space. To set the Tab key as the delimiter, set this parameter to \t .	,
Max. Poll Records	(Optional) Maximum number of records per poll	100
Max. Poll Interval	(Optional) Maximum interval between polls (seconds)	100

4.6.3.15 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 4-59](#).

Table 4-59 Job parameters when Elasticsearch or CSS is the source

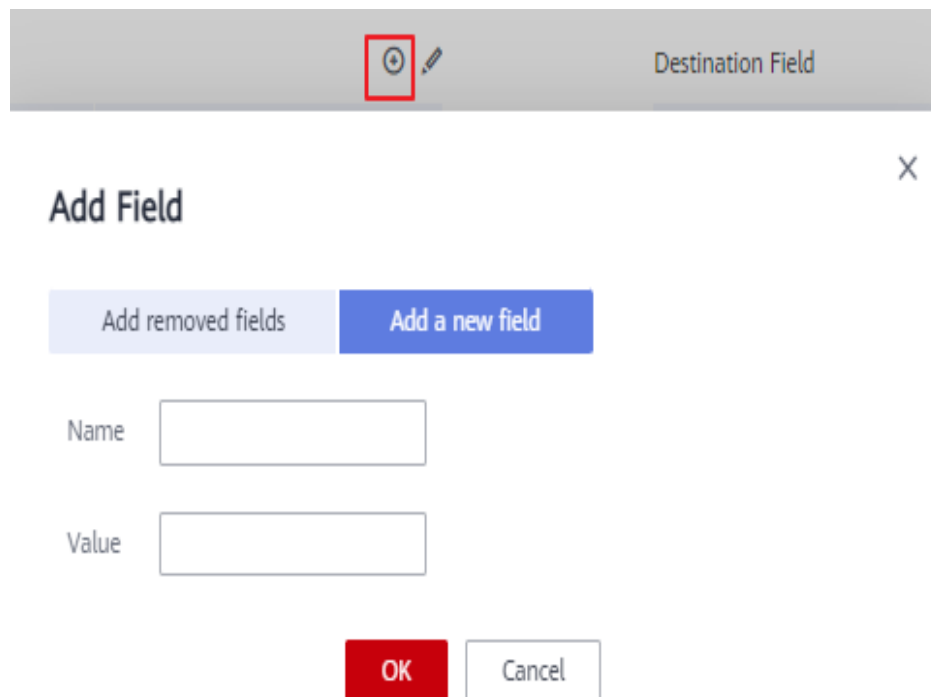
Parameter	Description	Example Value
Index	Elasticsearch index, which is similar to the name of a relational database. The index name can contain only lowercase letters.	index
Type	Elasticsearch type, which is similar to the table name of a relational database. The type name can contain only lowercase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

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

Parameter	Description	Example Value
	<p><code>{timestamp(dateformat(yyyyMMdd,-1,DAY))} TO \$ {timestamp(dateformat(yyyyMMdd))}</code>.</p> <ul style="list-style-type: none"> Source data cannot be filtered using the query domain-specific language (DSL) of Elasticsearch. 	
Extract Meta-field	Whether to extract index meta-fields. For example, <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 4-39 Setting custom fields



4.6.3.16 From OpenTSDB

If the source link of a job is the [Link to CloudTable OpenTSDB](#), configure the source job parameters based on [Table 4-60](#).

Table 4-60 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

4.6.4 Destination Job Parameters

4.6.4.1 To OBS


If the destination link of a job is the [Link to OBS](#), configure the destination job parameters based on [Table 4-61](#).

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

Table 4-61 Parameter description

Category	Parameter	Description	Example Value
Basic parameters	Bucket Name	Name of the OBS bucket that data will be written to	bucket_2

Category	Parameter	Description	Example Value
	Write Directory	<p>OBS directory to which data will be written. Do not add / in front of the directory name.</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. For details, see Incremental Synchronization Using the Macro Variables of Date and Time.</p> <p>NOTE If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with <i>(Planned start time of the data development job - Offset)</i> rather than <i>(Actual start time of the CDM job - Offset)</i>.</p>	directory/
	File Format	<p>Format in which data is written. The options are as follows:</p> <ul style="list-style-type: none"> • CSV: Data is written in CSV format, which is used for migrating data tables to files. • Binary: Files will be transferred directly. CDM writes the files without changing their format. This setting is suitable for file migration. <p>If data is migrated between file-related data sources, such as FTP, SFTP, HDFS, and OBS, the value of File Format 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"> • Replace • Skip • Stop job <p>For details, see Incremental File Migration.</p>	Skip

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

Category	Parameter	Description	Example Value
	DEK	This parameter is displayed only when Encryption is set to AES-256-GCM . The key consists of 64 hexadecimal numbers. Remember the key configured here because the decryption key must be the same as that configured here. If the encryption and decryption keys are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	DD0AE00D FECDF78BF0 51BCFDA25 BD4E320DB 0A7AC75A1 F3FC3D3C5 6A457DCD C1B
	IV	This parameter is displayed only when Encryption is set to AES-256-GCM . The initialization vector consists of 32 hexadecimal numbers. Remember the initialization vector configured here because the initialization vector used for decryption must be the same as that configured here. If the initialization vectors are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	5C91687BA 886EDCD12 ACBC3FF19 A3C3F
	Copy Content-Type	This parameter is displayed only when File Format is Binary , and both the migration source and destination are object storage. If you set this parameter to Yes , the Content-Type attribute of the source file is copied during object file migration. This function is mainly used for static website migration. The Content-Type attribute cannot be written to Archive buckets. Therefore, if you set this parameter to Yes , the migration destination must be a non-Archive bucket.	No
	Line Separator	Line feed character in a file. By default, the system automatically identifies \n , \r , and \r\n . This parameter is not used when File Format is set to Binary .	\n
	Field Delimiter	Field delimiter in the file. This parameter is not used when File Format is set to Binary .	,

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

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

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

4.6.4.2 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 4-62](#).

Table 4-62 Parameter description

Parameter	Description	Example Value
Write Directory	<p>HDFS directory to which data will be written.</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. For details, see Incremental Synchronization Using the Macro Variables of Date and Time.</p> <p>NOTE If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (<i>Planned start time of the data development job - Offset</i>) rather than (<i>Actual start time of the CDM job - Offset</i>).</p>	/user/output
File Format	<p>Format in which data is written. The options are as follows:</p> <ul style="list-style-type: none"> • CSV: Data is written in CSV format, which is used for migrating data tables to files. • Binary: Files will be transferred directly. CDM writes the files without changing their format. This setting is suitable for file migration. <p>If data is migrated between file-related data sources, such as FTP, SFTP, HDFS, and OBS, the value of File Format 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"> • Replace • Skip • Stop job 	Stop job

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

Parameter	Description	Example Value
Customize Hierarchical Directory	Users can customize the directory hierarchy of files. Example: [Table name]/[Year]/[Month]/[Day]/[Data file name]. csv	-
Hierarchical Directory	Used to specify the directory level of a file, with time macro supported (the time format is yyyy/MM/dd). If this parameter is left blank, the directory does not have a hierarchical structure. Example: \${dateformat/yyyy/MM/dd, -1, DAY)}	-
Encryption	<p>This parameter is displayed only when File Format is set to Binary.</p> <p>Whether to encrypt the uploaded data. The options are as follows:</p> <ul style="list-style-type: none"> • None: Data is written without encryption. • AES-256-GCM: The AES 256-bit encryption algorithm is used to encrypt data. Currently, only the AES-256-GCM (NoPadding) encryption algorithm is supported. This parameter is used for encryption at the migration destination and decryption at the migration source. <p>For details, see Encryption and Decryption During File Migration.</p>	AES-256-GCM
DEK	<p>This parameter is displayed only when Encryption is set to AES-256-GCM. 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 Encryption is set to AES-256-GCM. 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**.

4.6.4.3 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 4-63](#).

Table 4-63 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. For details, see Incremental Synchronization Using the Macro Variables of Date and Time.</p> <p>NOTE If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with <i>(Planned start time of the data development job - Offset)</i> rather than <i>(Actual start time of the CDM job - Offset)</i>.</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"> • Yes: The data is cleared. • No: The data is not cleared. Instead, it will be added to the existing table. 	Yes
Rowkey Delimiter	(Optional) Used to combine multiple columns as a rowkey. Spaces are used by default.	,
Rowkey Data Redundancy	(Optional) Whether to write the rowkey data into HBase columns. The default value is No .	No
Compression Format	<p>(Optional) Compression format used in creating an HBase table. The default value is None.</p> <ul style="list-style-type: none"> • None: The files are not compressed. • Snappy: The files are compressed in snappy format. • gzip: The files are compressed in gzip format. 	None

Parameter	Description	Example Value
Write WAL	Whether to enable Write Ahead Log (WAL) of HBase. The options are as follows: <ul style="list-style-type: none">• Yes: If the HBase server breaks down after the function is enabled, you can replay the operations that have not been performed in WAL.• No: If you set this parameter to No, the write performance is improved. However, if the HBase server breaks down, data may be lost.	No
Match Data Type	<ul style="list-style-type: none">• Yes: Data of the Short, Int, Long, Float, Double, and Decimal columns in the source database is converted into Byte[] arrays (binary) and written into HBase. Other types of data are written as character strings. If several types of data mentioned above are combined as rowkeys, they will be written as character strings. This function saves storage space. In specific scenarios, the rowkey distribution is evener.• No: All types of data in the source database are written into HBase as character strings.	No

4.6.4.4 To Hive

If the destination link of a job is the [Link to Hive](#), configure the destination job parameters based on [Table 4-64](#).

Table 4-64 Parameter description

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

Parameter	Description	Example Value
Auto Table Creation	<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"> • Non-auto creation: CDM will not automatically create a table. • Auto creation: If the destination database does not contain the table specified by Table Name, CDM will automatically create the table. If the table specified by Table Name already exists, no table is created and data is written to the existing table. • Deletion before creation: CDM deletes the table specified by Table Name, and then creates the table again. 	Non-auto creation
Table Name	<p>Destination 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. For details, see Incremental Synchronization Using the Macro Variables of Date and Time.</p> <p>NOTE If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with <i>(Planned start time of the data development job - Offset)</i> rather than <i>(Actual start time of the CDM job - Offset)</i>.</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"> • Yes: The data is cleared. • No: The data is not cleared. Instead, it will be added to the existing table. 	Yes

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

 **NOTE**

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

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

4.6.4.5 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 4-65](#).

Table 4-65 Parameter description

Category	Parameter	Description	Example Value
Basic parameter s	Schema/ Tables pace	Name of the database to which data will be written. The schema can be automatically created. Click the icon next to the text box to select a schema or tablespace.	schema
	Auto Table Creatio n	<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"> • Non-auto creation: CDM will not automatically create a table. • Auto creation: If the destination database does not contain the table specified by Table Name, CDM will automatically create the table. If the table specified by Table Name already exists, no table is created and data is written to the existing table. • Deletion before creation: CDM deletes the table specified by Table Name, and then creates the table again. 	Non-auto creation
	Table Name	<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. For details, see Incremental Synchronization Using the Macro Variables of Date and Time.</p> <p>NOTE If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with <i>(Planned start time of the data development job - Offset)</i> rather than <i>(Actual start time of the CDM job - Offset)</i>.</p>	table

Category	Parameter	Description	Example Value
	Clear Data Before Import	Whether to clear the data in the destination table before data import. The options are as follows: <ul style="list-style-type: none">• Do not clear: The data in the destination table is not cleared before data import. The imported data is just added to the table.• Clear all data: All data is cleared from the destination table before data import.• Clear part of data: Part of the data in the destination table is cleared before data import. If you select Clear part of data, you must configure WHERE Clause to specify which part will be deleted.	Clear part of data
	WHERE Clause	If Clear Data Before Import is set to Clear part of data , data in the destination table will be deleted based on the WHERE clause after the configuration is complete and before the import starts.	age > 18 and age <= 60
	Constraint Conflict Handling	Mode for handling conflicts in data migration <ul style="list-style-type: none">• insert into: When a primary key or unique index conflict occurs, data cannot be written and will become dirty data.• replace into: When a primary key or unique index conflict occurs, the original row is deleted and a new row is inserted to replace all the fields in the original row.• on duplicate key update: When a primary key or unique index conflict occurs in a row in the destination table, the data columns except the unique constraint column in this row are updated.	insert into
	Loader Threads	Number of threads started in each loader. A larger number allows more concurrent write operations. NOTE This parameter is unavailable if Constraint Conflict Handling is set to replace into or on duplicate key update .	1

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

4.6.4.6 To DWS

If the destination link of a job is a [DWS link](#), configure the destination job parameters based on [Table 4-66](#).

Table 4-66 Parameter description

Category	Parameter	Description	Example Value
Basic parameter s	Schema/ Tables pace	Name of the database to which data will be written. The schema can be automatically created. Click the icon next to the text box to select a schema or tablespace.	schema
	Auto Table Creatio n	<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"> • Non-auto creation: CDM will not automatically create a table. • Auto creation: If the destination database does not contain the table specified by Table Name, CDM will automatically create the table. If the table specified by Table Name already exists, no table is created and data is written to the existing table. • Deletion before creation: CDM deletes the table specified by Table Name, and then creates the table again. <p>Field Mapping in Automatic Table Creation on DWS describes the field mapping between the DWS tables created by CDM and source tables.</p>	Non-auto creation

Category	Parameter	Description	Example Value
	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. For details, see Incremental Synchronization Using the Macro Variables of Date and Time.</p> <p>NOTE If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with (<i>Planned start time of the data development job - Offset</i>) rather than (<i>Actual start time of the CDM job - Offset</i>).</p>	table
	Compress Data	Whether to compress data when data is imported to DWS and Auto creation is selected	No
	Storage Mode	<p>When data is imported to DWS and Auto Creation is selected, you can specify the data storage mode:</p> <ul style="list-style-type: none"> ● Row-based: Row-based storage. It is used for point queries (index-based simple queries with fewer return records), or the scenario that requires a large number of addition, deletion, and modification operations. ● Column-based: Column-based storage. It is used for statistical analysis queries (group and join scenarios) or ad hoc queries (query conditions are uncertain and indexes can hardly be used to scan row-based tables). 	Row-based
	Import Mode	<p>Mode for importing data to DWS</p> <ul style="list-style-type: none"> ● In COPY mode, the source data is copied to the DataNode of DWS after passing through the management node. ● In UPSERT mode, if a primary key or unique constraint conflict occurs, other data columns, except the primary key and unique constraint column, are updated. 	COPY

Category	Parameter	Description	Example Value
	Clear Data Before Import	Whether to clear the data in the destination table before data import. The options are as follows: <ul style="list-style-type: none"> • Do not clear: The data in the destination table is not cleared before data import. The imported data is just added to the table. • Clear all data: All data is cleared from the destination table before data import. • Clear part of data: Part of the data in the destination table is cleared before data import. If you select Clear part of data, you must configure WHERE Clause to specify which part will be deleted. 	Clear part of data
	WHERE Clause	If Clear Data Before Import is set to Clear part of data , data in the destination table will be deleted based on the WHERE clause after the configuration is complete and before the import starts.	age > 18 and age <= 60
	Constraint Conflict Handling	Mode for handling conflicts in data migration <ul style="list-style-type: none"> • insert into: When a primary key or unique index conflict occurs, data cannot be written and will become dirty data. • replace into: When a primary key or unique index conflict occurs, the original row is deleted and a new row is inserted to replace all the fields in the original row. • on duplicate key update: When a primary key or unique index conflict occurs in a row in the destination table, the data columns except the unique constraint column in this row are updated. 	insert into
	Loader Threads	Number of threads started in each loader. A larger number allows more concurrent write operations. NOTE This parameter is unavailable if Constraint Conflict Handling is set to replace into or on duplicate key update .	1

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

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

Field Mapping in Automatic Table Creation on DWS

Figure 4-40 describes the field mapping between DWS tables created by CDM and source tables. For example, if you use CDM to migrate the Oracle database to DWS, CDM automatically creates a table on DWS and maps the **NUMBER(3,0)** field of the Oracle database to the **SMALLINT** field of DWS.

Figure 4-40 Field mapping in automatic table creation

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

4.6.4.7 To DDS

If the destination link of a job is the [Link to DDS](#), configure the destination job parameters based on [Table 4-67](#).

Table 4-67 Parameter description

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

4.6.4.8 To DCS

If the data is imported to DCS, configure the destination job parameters based on [Table 4-68](#).

Table 4-68 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">• String: without column name, such as value1,value2• Hash: with column name, such as column1=value1,column2=value2	String
Key Delimiter	Character used to separate table names and column names of a relational database	-
Value Delimiter	Character used to separate columns when the storage type is string	;

4.6.4.9 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 4-69](#).

Table 4-69 Parameter description

Parameter	Description	Example Value
Index	Elasticsearch index, which is similar to the name of a relational database. CDM supports automatic creation of indexes and field types. The index and field type names can contain only lowercase letters.	index
Type	Elasticsearch type, which is similar to the table name of a relational database. The type name can contain only lowercase letters. NOTE Elasticsearch 7.x and later versions do not support custom types. Instead, only the <code>_doc</code> type can be used. In this case, this parameter does not take effect even if it is set.	type
Pipeline ID	Pipeline used to convert the data format after data is transferred to Elasticsearch. Pipeline IDs are ready for use after being created in Kibana.	pipeline_id
Periodically Create Index	For streaming jobs that continuously write data to Elasticsearch, CDM periodically creates indexes and writes data to the indexes, which helps you delete expired data. The indexes can be created based on the following periods: <ul style="list-style-type: none"> • Every hour: 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, index2018121709. • Every day: 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, index20181217. • Every week: 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, index201842. • Every month: CDM creates indexes at 00:00 on the first day of each month. The new indexes are named in the format of <i>Index name+Year+Month</i>, for example, index201812. • Do not create: Do not create indexes periodically. <p>When extracting data from a file, you must configure a single extractor, which means setting Concurrent Extractors to 1. Otherwise, this parameter is invalid.</p>	Every hour

4.6.4.10 To DLI

If the destination link of a job is the [Link to DLI](#), configure the destination job parameters based on [Table 4-70](#).

NOTE

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

Table 4-70 Parameter description

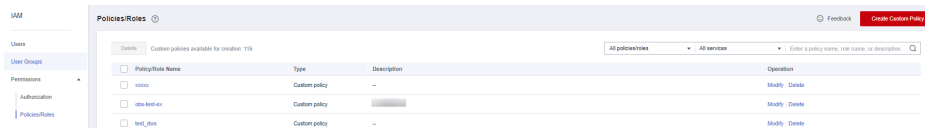
Parameter	Description	Example Value
Resource Queue	Resource queue to which the destination table belongs The default queue of DLI cannot be used for migration jobs. You need to create a SQL queue in DLI.	cdm
Database Name	Name of the database to which data will be written	dli
Table Name	Name of the table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

Adding an OBS Bucket Policy

Step 1 Log in to the IAM console.

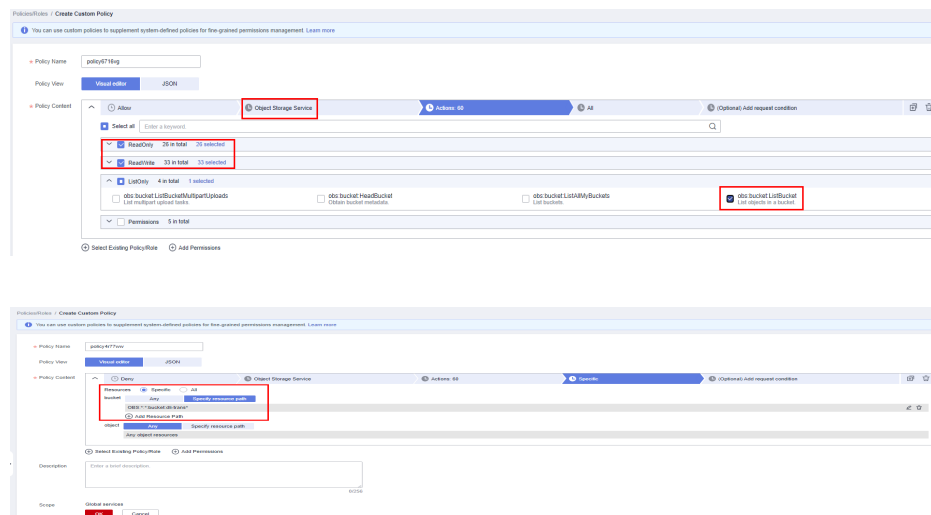
Step 2 In the navigation pane, choose **Permissions > Policies/Roles** and click **Create Custom Policy** in the upper right corner.

Figure 4-41 Creating a custom policy



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

Figure 4-42 Configuring the policy



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

----End

4.6.4.11 To OpenTSDB

If the destination link of a job is the [Link to CloudTable OpenTSDB](#), configure the destination job parameters based on [Table 4-71](#).

Table 4-71 Parameter description

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

4.6.5 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

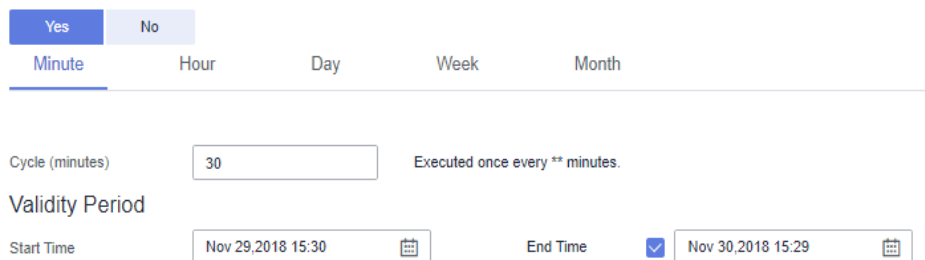
- When configuring scheduled jobs, do not set the same scheduled time for different jobs. Instead, set different times to avoid exceptions.
- If you use DataArts Studio DataArts Factory to schedule the CDM migration job and configure this parameter, both configurations take effect. To ensure unified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

Scheduling Job Execution by Minute

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

- **Start Time:** indicates the time when the scheduled configuration takes effect, or the first time when the job is automatically executed.
- **Cycle (minutes):** indicates the interval when a job is executed starting from the start time.
- **End Time:** This parameter is optional. If it is not set, the scheduled job keeps being automatically executed. If it is set, the scheduled job will be automatically stopped at the end time.

Figure 4-43 Scheduling job execution by minute



The screenshot shows a configuration interface for scheduling job execution by minute. At the top, there are two radio buttons: 'Yes' (selected) and 'No'. Below them are five tabs: 'Minute' (selected), 'Hour', 'Day', 'Week', and 'Month'. Under the 'Minute' tab, there is a 'Cycle (minutes)' field with the value '30' and a label 'Executed once every ** minutes.'. Below that is a 'Validity Period' section with 'Start Time' set to 'Nov 29, 2018 15:30' and 'End Time' checked and set to 'Nov 30, 2018 15:29'.

Figure 4-43 shows that the job will be automatically executed at 15:30:30 on November 29, 2018 for the first time at a cycle of 30 minutes, and will be automatically stopped at 15:29:00 on November 30, 2018.

Scheduling Job Execution by Hour

CDM allows jobs to be executed every several hours.

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

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

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

Figure 4-44 Scheduling job execution by hour

The screenshot shows a configuration interface for scheduling job execution. At the top, there are tabs for 'Yes' (selected) and 'No'. Below these are five tabs: 'Minute', 'Hour' (selected), 'Day', 'Week', and 'Month'. The 'Hour' tab is active, showing the following settings:

- Cycle (hours):** A text input field containing '2'. To its right, the text reads 'Executed once every ** hours.'
- Trigger Time (minute):** A text input field containing '10,30,50'. Below it, a note states: 'Exact trigger time of each hour. For example, 1,3 would indicate that task execution will be triggered at the first and third minute of each hour.'
- Validity Period:** This section includes:
 - Start Time:** A date-time picker showing 'Nov 29, 2018 15:30'.
 - End Time:** A checkbox that is currently unchecked, followed by a date-time picker showing 'Nov 30, 2018 15:29'.

Figure 4-44 shows that the scheduled configuration will take effect at 15:30:00 on November 30, 2018. The job is automatically executed for the first time upon the scheduled configuration takes effect, at 15:50:00 for the second time, and at 17:10:00 for the third time. The job is triggered for three times every 2 hours and the configuration is always valid.

Scheduling Job Execution by Day

CDM allows jobs to be executed every several days.

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

Figure 4-45 Scheduling job execution by day

The screenshot shows a scheduling configuration window with the following elements:

- Buttons: "Yes" (selected) and "No".
- Frequency tabs: "Minute", "Hour", "Day" (selected), "Week", "Month".
- Cycle (days): Input field with value "3". Text: "Executed once every ** days.".
- Validity Period: "Start Time" field with value "Dec 01,2018 00:20" and a calendar icon. "End Time" field is empty with a placeholder "Select a date and time." and a calendar icon.

Figure 4-45 shows that the scheduled job will be automatically executed at 00:20:00 on December 1, 2018, and is executed once every three days. The configuration is always valid.

Scheduling Job Execution by Week

CDM allows jobs to be executed every several weeks.

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

Figure 4-46 Scheduling job execution by week

The screenshot shows a scheduling configuration window with the following elements:

- Buttons: "Yes" (selected) and "No".
- Frequency tabs: "Minute", "Hour", "Day", "Week" (selected), "Month".
- Cycle (weeks): Input field with value "2". Text: "Executed once every ** weeks.".
- Trigger Time (day): "Select All" checkbox is unchecked. Individual day checkboxes: Monday (unchecked), Tuesday (checked), Wednesday (unchecked), Thursday (unchecked), Friday (unchecked), Saturday (checked), Sunday (checked).
- Validity Period: "Start Time" field with value "Dec 01,2018 00:20" and a calendar icon. "End Time" field has a checked checkbox and value "Jun 01,2019 00:00" and a calendar icon.

Figure 4-46 shows that the job will be automatically executed at 00:20:00 every Tuesday, Saturday, and Sunday every two weeks starting from 00:20:00 on December 1, 2018, and the job will be automatically stopped at 00:00:00 on June 1, 2019.

Scheduling Job Execution by Month

CDM allows jobs to be executed every several months.

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

Figure 4-47 Scheduling job execution by month

The screenshot shows a configuration page for scheduling job execution by month. At the top, there are two tabs: 'Yes' (selected) and 'No'. Below the tabs are five frequency options: 'Minute', 'Hour', 'Day', 'Week', and 'Month' (selected). The 'Cycle (months)' field is set to '1', with a note 'Executed once every ** months.' The 'Trigger Time (day)' field is set to '5,25', with a note 'Exact trigger time of each month. For example, 1,3 would indicate that task execution will be triggered on the first and third day of each month.' The 'Validity Period' section includes a 'Start Time' field set to 'Dec 01,2018 00:00' and an 'End Time' field set to 'Jun 01,2019 00:00'.

Figure 4-47 shows that the job will be automatically executed at 00:00:00 on every fifth and twenty-fifth day of each month starting from 00:00:00 on December 1, 2018. The configuration is always valid.

4.6.6 Job Configuration Management

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

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

Maximum Concurrent Extractors

Maximum number of concurrent extraction tasks in a cluster

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

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

NOTE

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

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

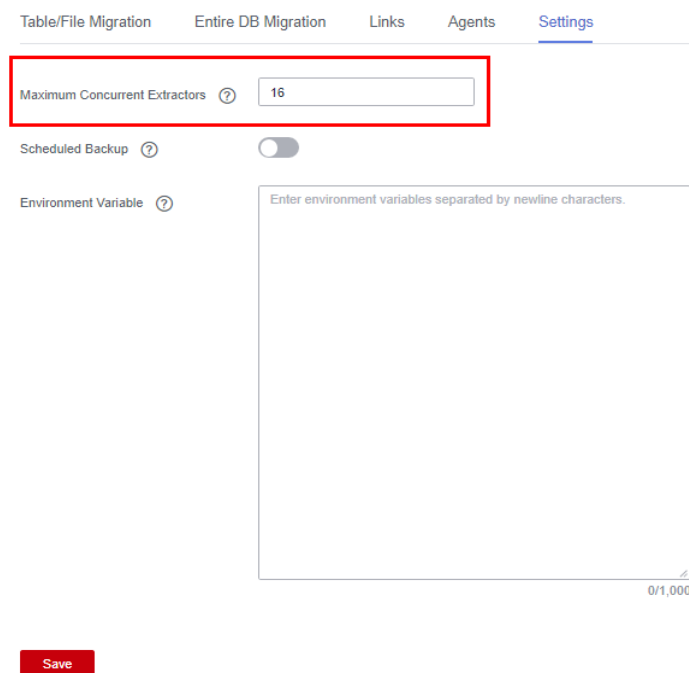
By setting an appropriate number of concurrent extractors for a job and the maximum number of concurrent extractors for the cluster, you can accelerate migration. You can configure the number of concurrent extractors as follows:

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

Table 4-72 Maximum number of concurrent extractors for a CDM cluster

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

Figure 4-48 Setting Maximum Concurrent Extractors for a CDM cluster



2. Configure the number of concurrent extractors based on the following rules:
 - a. When data is to be migrated to files, CDM does not support multiple concurrent tasks. In this case, set a single process to extract data.
 - b. If each row of the table contains less than or equal to 1 MB data, data can be extracted concurrently. If each row contains more than 1 MB data, it is recommended that data be extracted in a single thread.

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

Figure 4-49 Setting Concurrent Extractors for a job

The screenshot shows the 'Configure Task' interface. At the top, there are settings for 'Retry if failed' (set to 'Never'), 'Group' (set to 'ck2ck'), and 'Schedule Execution' (with 'No' selected). Below these is a 'Hide Advanced Attributes' link. The 'Concurrent Extractors' field is highlighted with a red box and contains the value '1'. Other fields include 'Write Dirty Data' (with 'No' selected) and 'Throttling' (with 'No' selected). At the bottom, there are buttons for 'Cancel', 'Previous', 'Save', and 'Save and Run'.

Scheduled Backup/Restoration

This function depends on the OBS service.

- Prerequisites
You have created the [Link to OBS](#).
- Scheduled backup
On the **Job Management** page, click **Settings** and configure **Scheduled Backup** and its related parameters.

Table 4-73 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"> • All jobs: CDM backs up all table/file migration jobs and entire DB migration jobs regardless of the job statuses. However, historical jobs are not backed up. • All jobs by groups: You select one or more job groups to back up. 	All jobs

Parameter	Description	Example Value
Backup Cycle	Select the backup cycle. <ul style="list-style-type: none">• Day: The backup is performed daily at 00:00:00.• Week: The backup is performed at 00:00:00 every Monday.• Month: The backup is performed at 00:00:00 on the first day of each month.	Day
OBS Link for Writing Backups	Link used to back up jobs to OBS buckets. Select a link you have created on the Links page.	obslink
OBS Bucket	OBS bucket where backup files are stored	cdm
Backup Data Directory	Directory where backup files are stored	/cdm-bk/

- Restoring jobs

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

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

Environment Variables of Job Parameters

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

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

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

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

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

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

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

Figure 4-50 Setting the bucket names to environment variables

Job Configuration

* Job Name A-B

Source Job Configuration

* Source Link Name OBS_LINK1

* Bucket Name S(bucket_1)

* Source Directory/File FROM

Entries Files Yes No

* File Format Binary

Show Advanced Attributes

Destination Job Configuration

* Destination Link Name OBS_LINK1

* Bucket Name S(bucket_2)

* Write Directory TO

* File Format Binary

Duplicate File Processing Method Replace

Show Advanced Attributes

Cancel Next

3. If you want to migrate data from bucket C to bucket D, you do not need to change the job parameters. You only need to change the environment variables on the **Configuration Management** tab page as follows:
bucket_1=C
bucket_2=D

4.6.7 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

4.6.8 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.

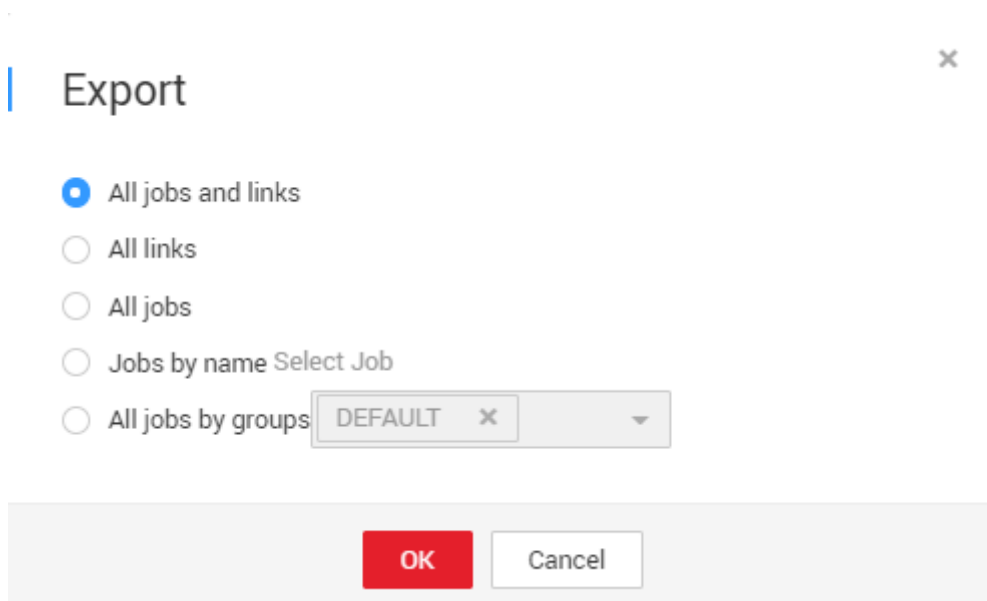
- **Delete jobs in batches.**

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

- **Export jobs in batches.**

Click **Export**.

Figure 4-51 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

4.7 Auditing

4.7.1 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 4-74 CDM operations recorded by CTS

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

4.7.2 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**.

- In the left navigation pane, click **Trace List**.
Click **Filter** and specify filter criteria as needed.

Figure 4-52 CDM traces

Trace Name	Resource Type	Trace Sour...	Resource ID	Resource Name	Trace Status	Operator	Operation Time	Operation
startJob	job	CDM	obs2obs	obs2obs	normal	billy_came	Aug 14, 2018 14:09:14 GMT+08:00	View Trace
startCluster	cluster	CDM	0fd31035-3d7e-4f...	cdm-xlarge-deng...	normal	billy_came	Aug 14, 2018 14:08:23 GMT+08:00	View Trace
startCluster	cluster	CDM	176f2fd9-62a1-4...	cdm-forTest	normal	billy_came	Aug 14, 2018 13:56:06 GMT+08:00	View Trace

- Unfold the target trace to view its details.
- Click **View Trace** in the **Operation** column to view the trace structure details.
For more information about CTS, see [Cloud Trace Service User Guide](#).

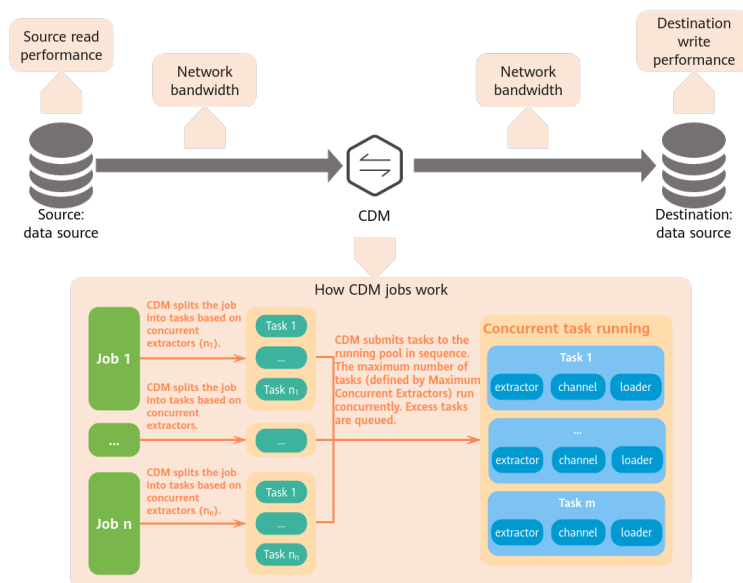
4.8 Performance Reference

4.8.1 Factors Affecting Performance

Data Migration Model

Figure 4-53 shows the simplified migration model used by CDM.

Figure 4-53 Migration model used by CDM



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

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

 **NOTE**

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

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

Factors Affecting Migration Performance

According to the migration model, the migration speed is affected by factors such as the source read speed, network bandwidth, destination write performance, and CDM cluster and job configuration.

Table 4-75 Factors affecting migration performance

Factor		Description
Service-related factors	Concurrent extractors of a job	<p>The number of concurrent extractors can be set for a CDM job during the job creation.</p> <p>Setting a proper value for this parameter can effectively improve the migration speed. If the value is too small, migration will be too slow. If the value is too large, the migration job is overloaded and may fail.</p> <ul style="list-style-type: none"> • 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. • If each row of the table contains less than or equal to 1 MB data, data can be extracted concurrently. If each row contains more than 1 MB data, it is recommended that data be extracted in a single thread.
	Maximum concurrent extractors of a cluster	<p>Setting a proper value for this parameter can effectively improve the migration speed. If the value is too small, migration will be too slow. If the value is too large, the source is overloaded and the system may be unstable.</p> <p>The maximum concurrent extractors vary depending on the CDM cluster flavor. The upper limit is twice the number of vCPUs. The following are the maximum concurrent extractors of some flavors:</p> <ul style="list-style-type: none"> • cdm.large: 16 • cdm.xlarge: 32 • cdm.4xlarge: 64
	Service model	<p>If the number of CDM jobs that run concurrently exceeds the maximum concurrent extractors for the CDM cluster, some jobs will be queued, and the migration will be prolonged.</p> <p>Avoid running too many jobs simultaneously, which may cause slow migration due to insufficient resources.</p>

Factor		Description
	Data model	<p>The migration speed is also affected by the data structure. The following are some examples:</p> <ul style="list-style-type: none">• The wider a table is and the more string types the table has, the slower the migration is.• A large file is migrated more quickly than multiple small files whose total size is the same as the large file.• The more content a message has and the higher bandwidth it uses, the less transactions per second (TPS) are.
	Source read speed	<p>It depends on the performance of the data source at the source. For details about how to increase the read speed, see the documents of data sources at the source.</p>
	Network bandwidth	<p>The CDM cluster can communicate with the data source through an intranet, public network VPN, NAT, or Direct Connect.</p> <ul style="list-style-type: none">• If they communicate through an intranet, the network bandwidth varies depending on the CDM instance flavor.<ul style="list-style-type: none">– For cdm.large instances, the baseline and maximum bandwidths of the CDM cluster NIC are 0.8 and 3 Gbit/s, respectively.– For cdm.xlarge instances, the baseline and maximum bandwidths of the CDM cluster NIC are 4 and 10 Gbit/s, respectively.– For cdm.4xlarge instances, the baseline and maximum bandwidths of the CDM cluster NIC are 36 and 40 Gbit/s, respectively.• If they communicate through the Internet, the network bandwidth is subject to the Internet bandwidth. The bandwidth for the CDM cluster depends on the EIP bound to the CDM cluster, and the bandwidth for the data source depends on the Internet bandwidth.• If they communicate through a VPN, NAT, or Direct Connect, the network bandwidth is subject to the VPN, NAT, or Direct Connect bandwidth.
	Destination write performance	<p>It depends on the performance of the data source at the destination. For details about how to improve the performance, see the documents of data sources at the destination.</p>

4.8.2 Performance Tuning

Overview

In addition to increasing the source read speed, improving the destination write performance, and increasing the bandwidth, you can accelerate migration using the following methods:

- **Use a CDM cluster of higher specifications**

The NIC bandwidth and maximum number of concurrent extractors vary depending on the CDM cluster specifications. If you want to migrate data faster, or the metrics of your CDM cluster (such as the CPU usage, disk usage, and memory usage) are often high, you may need a CDM cluster with higher specifications for data migration.

- **Use multiple CDM clusters**

In some scenarios, you are advised to use multiple CDM clusters to share workloads to improve migration efficiency and stability. The following are some examples:

- Multiple CDM clusters are required for different purposes or by multiple business departments. For example, you may need one CDM cluster for running data migration jobs and another one as an agent for DataArts Studio Management Center.
- You want to migrate a large number of tables. In this case, you can use multiple CDM clusters to run jobs simultaneously to improve migration efficiency.
- The CPU usage, disk usage, and memory usage of the in-use CDM cluster are often high. In this case, you are advised to use multiple CDM clusters to shared workloads.

- **Avoid running too many CDM jobs simultaneously**

If the number of CDM jobs that run concurrently exceeds the maximum concurrent extractors for the CDM cluster, some jobs will be queued, and the migration will be prolonged.

Avoid running too many jobs simultaneously, which may cause slow migration due to insufficient resources.

- **Change concurrent extractors**

If the number of tasks is small, adjusting the number of concurrent extractors is the best way to improve performance. You can set the number of concurrent extractors for a job and the maximum number of concurrent extractors for a cluster.

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

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

 **NOTE**

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

- b. CDM submits the tasks to the running pool in sequence. The maximum number of tasks (defined by **Maximum Concurrent Extractors**) run concurrently. Excess tasks are queued.

By setting appropriate values for parameters **Concurrent Extractors** and **Maximum Concurrent Extractors**, you can accelerate migration. For details about how to change **Concurrent Extractors**, see [Changing Concurrent Extractors](#).

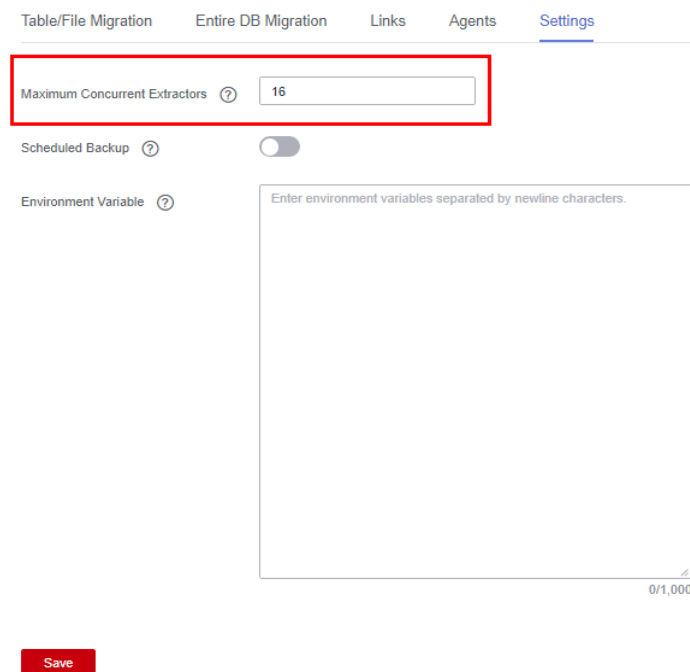
Changing Concurrent Extractors

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

Table 4-76 Maximum number of concurrent extractors for a CDM cluster

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

Figure 4-54 Setting Maximum Concurrent Extractors for a CDM cluster



2. Configure the number of concurrent extractors based on the following rules:
 - a. When data is to be migrated to files, CDM does not support multiple concurrent tasks. In this case, set a single process to extract data.

- b. If each row of the table contains less than or equal to 1 MB data, data can be extracted concurrently. If each row contains more than 1 MB data, it is recommended that data be extracted in a single thread.
- c. Set **Concurrent Extractors** for a job based on **Maximum Concurrent Extractors** for the cluster. It is recommended that **Concurrent Extractors** is less than **Maximum Concurrent Extractors**.

Figure 4-55 Setting Concurrent Extractors for a job

The screenshot shows the 'Configure Task' configuration page. The 'Concurrent Extractors' field is highlighted with a red border and contains the value '1'. Other configuration options include 'Retry if failed' set to 'Never', 'Group' set to 'ck2ck', 'Schedule Execution' set to 'No', 'Write Dirty Data' set to 'No', and 'Throttling' set to 'No'. At the bottom, there are buttons for 'Cancel', 'Previous', 'Save', and 'Save and Run'.

4.8.3 Reference: Job Splitting Dimensions

CDM splits jobs for different data sources based on different dimensions. [Table 4-77](#) lists the splitting dimensions.

Table 4-77 Job splitting dimensions for different data sources

Data Source Category	Data Source	Job Splitting Rule
Data warehouse	GaussDB(DWS)	<ul style="list-style-type: none"> • Jobs can be split based on table fields. • Jobs cannot be split based on table partitions.
	Data Lake Insight (DLI)	<ul style="list-style-type: none"> • Jobs can be split based on the partitioning information of partitioned tables. • Jobs cannot be split based on non-partitioned tables.
Hadoop	MRS HDFS	Jobs can be split based on files.
	MRS HBase	Jobs can be split based on HBase regions.

Data Source Category	Data Source	Job Splitting Rule
	MRS Hive	<ul style="list-style-type: none"> When the read mode is HDFS, jobs can be split based on Hive files. When the read mode is JDBC, jobs cannot be split.
	FusionInsight HDFS	Jobs can be split based on files.
	FusionInsight HBase	Jobs can be split based on HBase regions.
	FusionInsight Hive	<ul style="list-style-type: none"> When the read mode is HDFS, jobs can be split based on Hive files. When the read mode is JDBC, jobs cannot be split.
	Apache HDFS	Jobs can be split based on files.
	Apache HBase	Jobs can be split based on HBase regions.
	Apache Hive	<ul style="list-style-type: none"> When the read mode is HDFS, jobs can be split based on Hive files. When the read mode is JDBC, jobs cannot be split.
Object storage	Object Storage Service (OBS)	Jobs can be split based on files.
File system	FTP	Jobs can be split based on files.
	SFTP	Jobs can be split based on files.
	HTTP	Jobs can be split based on files.
Relational database	RDS for MySQL	<ul style="list-style-type: none"> Jobs can be split based on table fields. Jobs can be split based on table partitions only when Extract by Partition is configured.
	RDS for PostgreSQL	<ul style="list-style-type: none"> Jobs can be split based on table fields. Jobs can be split based on table partitions only when Extract by Partition is configured.
	RDS for SQL Server	<ul style="list-style-type: none"> Jobs can be split based on table fields. Jobs can be split based on table partitions only when Extract by Partition is configured.

Data Source Category	Data Source	Job Splitting Rule
	MySQL	<ul style="list-style-type: none"> Jobs can be split based on table fields. Jobs can be split based on table partitions only when Extract by Partition is configured.
	PostgreSQL	<ul style="list-style-type: none"> Jobs can be split based on table fields. Jobs can be split based on table partitions only when Extract by Partition is configured.
	Microsoft SQL Server	<ul style="list-style-type: none"> Jobs can be split based on table fields. Jobs cannot be split based on table partitions.
	Oracle	<ul style="list-style-type: none"> Jobs can be split based on table fields. Jobs can be split based on table partitions only when Extract by Partition is configured.
	SAP HANA	<ul style="list-style-type: none"> Jobs can be split based on table fields. Jobs cannot be split based on table partitions.
	Database shard	Each backend connects to a subjob, which can be split based on primary keys.
NoSQL	Distributed Cache Service (DCS)	Jobs cannot be split.
	Redis	Jobs cannot be split.
	Document Database Service (DDS)	Jobs cannot be split.
	MongoDB	Jobs cannot be split.
	Cassandra	Jobs can be split based on the token range of Cassandra.
Message system	Apache Kafka	Jobs can be split based on topics.
	DMS Kafka	Jobs can be split based on topics.
	MRS Kafka	Jobs can be split based on topics.
Search	Elasticsearch	Jobs cannot be split.
	Cloud Search Service (CSS)	Jobs cannot be split.

4.8.4 Reference: CDM Performance Test Data

Background

The performance metrics provided in this document are for reference only. The performance at your site may be affected by factors such as the data source performance at the source or destination, network bandwidth, latency, and the data and service model. It is recommended that you test the speed with a small amount of data before migration.

Environment

- A xlarge CDM cluster of the 2.9.1 200 version
- A table which has 50 million rows and 100 columns, and three HDFS binary files which have 35.97 million rows and 100 columns, 66.67 million rows and 100 columns, and 100 million rows and 100 columns, respectively.
- Number of concurrent extraction jobs for determining the maximum extraction/write rate: 1, 10, 20, 30, and 50

Data Source Extraction and Write Performance Test Data

[Table 4-78](#) and [Table 4-79](#) provide the data extraction and write performance, respectively.

Table 4-78 Data extraction performance

Data Source	Specifications	Version	Extraction Rate for a Single Job (Lines per Second)	Extraction Rate for Multiple Jobs (Lines per Second)
RDS for MySQL	8 vCPUs, 32 GB	MySQL 5.7	42,052	195,313 (concurrency: 40)
Oracle	8 vCPUs, 16 GB	19C	18,539	18,706 (concurrency: 10)
MRS HBase	Master: 16 vCPUs, 64 GB x 3 Node: 8 vCPUs, 32 GB x 3	MRS 3.1.0	6,296	69,156 (concurrency: 30)
MRS Hive	Master: 16 vCPUs, 64 GB x 3 Node: 8 vCPUs, 32 GB x 3	MRS 3.1.0	22,321	170,068 (concurrency: 30)

Data Source	Specifications	Version	Extraction Rate for a Single Job (Lines per Second)	Extraction Rate for Multiple Jobs (Lines per Second)
MRS HDFS (binary files)	Master: 16 vCPUs, 64 GB x 3 Node: 8 vCPUs, 32 GB x 3	MRS 3.1.0	138,727	141,468 (concurrency: 20)
			125,556	126,990 (concurrency: 10)
			120,919	120,919 (concurrency: 10)
DWS	8 vCPUs, 16 GB	8.1.1.300	13,434	/
DLI	16 vCPUs	SQL queue	71,023	19,290 (concurrency: 20)

Table 4-79 Data write performance

Data Source	Specifications	Version	Write Rate for a Single Job (Rows per Second)	Optimal Write Rate (Rows per Second)
RDS for MySQL	8 vCPUs, 32 GB	MySQL 5.7	2,658	/
Oracle	8 vCPUs, 16 GB	19C	/	/
MRS Hbase	Master: 16 vCPUs, 64 GB x 3 Node: 8 vCPUs, 32 GB x 3	MRS 3.1.0	3,959	4,120 (concurrency: 10)
MRS Hive	Master: 16 vCPUs, 64 GB x 3 Node: 8 vCPUs, 32 GB x 3	MRS 3.1.0	25,813	26,882 (concurrency: 10)
MRS HDFS (binary files)	Master: 16 vCPUs, 64 GB x 3	MRS 3.1.0	65,075	90,155 (concurrency: 10)
			86,248	86,248 (concurrency: 1)

Data Source	Specifications	Version	Write Rate for a Single Job (Rows per Second)	Optimal Write Rate (Rows per Second)
	Node: 8 vCPUs, 32 GB x 3		76,687	76,687 (concurrency: 1)
DWS	8 vCPUs, 16 GB	8.1.1.300	26,624	27,902 (concurrency: 10)
DLI	16 vCPUs	SQL queue	15,211	18,430 (concurrency: 10)

4.9 Tutorials

4.9.1 Creating an MRS Hive Link

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

Prerequisites

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

Creating an MRS Hive Link

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

Figure 4-56 Selecting a connector type



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

Figure 4-57 Creating an MRS Hive link

The screenshot shows a configuration form for creating an MRS Hive link. The form consists of several rows, each with a label, a value field, and a help icon (question mark). The labels are: Name, Connector, Hadoop Type, Manager IP, Authentication Method, HIVE Version, Username, Password, OBS storage support, Run Mode, Check Hive JDBC Connectivity, Use Cluster Config, and Hive Properties. The values are: hive_test, Hive, MRS, a greyed-out field, KERBEROS, HIVE_3_X, empty, empty, Yes, EMBEDDED, Yes, Yes, and a '+ Add' button. At the bottom, there are four buttons: 'X Cancel', '< Previous', 'Test', and 'Save'.

* Name	hive_test	Configuration Guide
* Connector	Hive	
* Hadoop Type	MRS	
* Manager IP		Select
Authentication Method	KERBEROS	
* HIVE Version	HIVE_3_X	
* Username		
* Password		
* OBS storage support	Yes No	
* Run Mode	EMBEDDED	
* Check Hive JDBC Connectivity	Yes No	
Use Cluster Config	Yes No	
Hide Advanced Attributes		
Hive Properties	+ Add	

X Cancel < Previous Test 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 4-80](#).

Table 4-80 MRS Hive link parameters

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	mrs-link
Manager IP	Floating IP address of MRS Manager. Click Select next to the Manager IP text box to select an MRS cluster. CDM automatically fills in the authentication information.	127.0.0.1
Authentication Method	Authentication method used for accessing MRS <ul style="list-style-type: none"> • SIMPLE: Select this for non-security mode. • KERBEROS: Select this for security mode. 	SIMPLE
HIVE Version	Set this to the Hive version on the server.	HIVE_3_X
Username	<p>If Authentication Method is set to KERBEROS, you must provide the username and password used for logging in to MRS Manager. If you need to create a snapshot when exporting a directory from HDFS, the user configured here must have the administrator permission on HDFS.</p> <p>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.</p> <p>NOTE</p> <ul style="list-style-type: none"> • If the CDM cluster version is 2.9.0 or later and the MRS cluster version is 3.1.0 or later, the created user must have the permissions of the Manager_viewer role to create links on CDM. To perform operations on databases, tables, and columns of an MRS component, you also need to add the database, table, and column permissions of the MRS component to the user by following the instructions in the MRS documentation. • If the CDM cluster version is earlier than 2.9.0 or the MRS cluster version is earlier than 3.1.0, the created user must have the permissions of Manager_administrator or System_administrator to create links on CDM. • A user with only the Manager_tenant or Manager_auditor permission cannot create connections. 	cdm
Password	Password used for logging in to MRS Manager	-

Parameter	Description	Example Value
OBS storage support	The server must support OBS storage. When creating a Hive table, you can store the table in OBS.	No
Run Mode	This parameter is used only when the Hive version is HIVE_3_X . Possible values are: <ul style="list-style-type: none">• EMBEDDED: The link instance runs with CDM. This mode delivers better performance.• Standalone: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select STANDALONE or configure different agents. Note: The STANDALONE mode is used to solve the version conflict problem. If the connector versions of the source and destination ends of the same link are different, a JAR file conflict occurs. In this case, you need to place the source or destination end in the STANDALONE process to prevent the migration failure caused by the conflict.	EMBEDDED
Check Hive JDBC Connectivity	Whether to check the Hive JDBC connectivity	Disabled
Use Cluster Config	You can create cluster configurations on the Links page to simplify the configuration of Hadoop link parameters.	No
Hive Properties	Other parameters for the Hive client	-

 **NOTE**

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

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

----**End**

4.9.2 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.

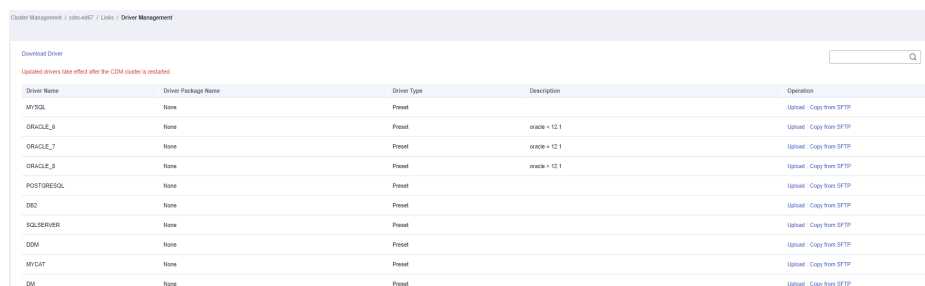
Prerequisites

- 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 4-58 Uploading a driver



Driver Name	Driver Package Name	Driver Type	Description	Operation
MYSQL	None	Preset		Upload Copy from SFTP
ORACLE_8	None	Preset	oracle = 12.1	Upload Copy from SFTP
ORACLE_7	None	Preset	oracle = 12.1	Upload Copy from SFTP
ORACLE_9	None	Preset	oracle = 12.1	Upload Copy from SFTP
POSTGRESQL	None	Preset		Upload Copy from SFTP
DB2	None	Preset		Upload Copy from SFTP
SOLSERVER	None	Preset		Upload Copy from SFTP
ODM	None	Preset		Upload Copy from SFTP
MYCAT	None	Preset		Upload Copy from SFTP
DM	None	Preset		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 4-59](#).

Figure 4-59 Selecting a connector type



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

Figure 4-60 Creating a MySQL link

* Name [Configuration Guide](#)

* Connector

Database Type

* Database Server

* Port

* Database Name

* Username

* Password

Use Local API

Use Agent

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

[Hide Advanced Attributes](#)

Fetch Size

Commit Size

Link Attributes

Reference Sign

Batch Size

Table 4-81 MySQL link parameters

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

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

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

 NOTE

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

----End

4.9.3 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 4-61](#).

Figure 4-61 MySQL table fields

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

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

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

Prerequisites

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

Creating a Hive Partition Table on MRS Hive

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

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

NOTE

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

Creating a CDM Cluster and Binding an EIP to the Cluster

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

The key configurations are as follows:

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

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

Figure 4-62 Cluster list



NOTE

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

----End

Creating a MySQL Link

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

Figure 4-63 Selecting a connector



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

Figure 4-64 Creating a MySQL link

The screenshot shows a configuration form for creating a MySQL link. The form is organized into several sections:

- Mandatory Fields:** Name (mysqllink), Connector (Relational Database), Database Type (MySQL), Database Server, Port, Database Name, Username (admin), and Password.
- Optional Fields:** Use Local API (No), Use Agent (No), and Driver Version (mysql-connector-java-5.1.48.jar).
- Advanced Attributes:** Fetch Size (1000), Commit Size (10000), Link Attributes (+ Add), Reference Sign, and Batch Size (100).

At the bottom of the form, there are four buttons: X Cancel, < Previous, Test, and Save.

Click **Show Advanced Attributes** and set optional parameters. For details, see [Link to Relational Databases](#). Retain the default values of the optional parameters and configure the mandatory parameters according to [Table 4-82](#).

Table 4-82 MySQL link parameters

Parameter	Description	Example Value
Name	Unique link name	mysqllink

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

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

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

NOTE

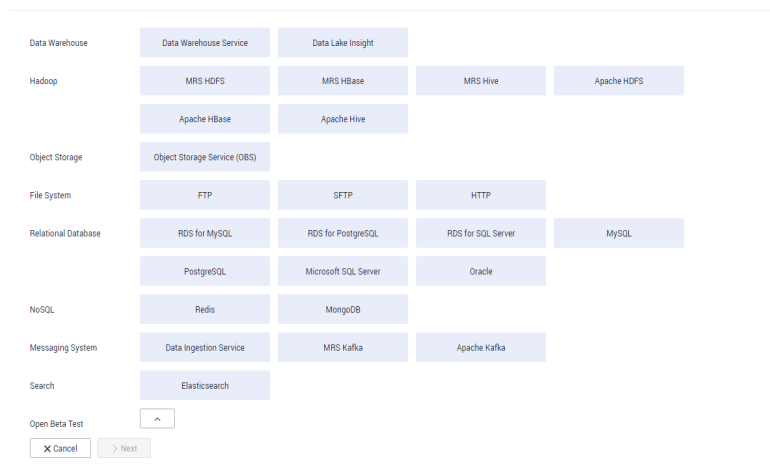
If an error occurs during the saving, the security settings of the MySQL database are incorrect. In this case, you need to enable the EIP of the CDM cluster to access the MySQL database.

----End

Creating a Hive Link

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

Figure 4-65 Selecting a connector type



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

Figure 4-66 Creating an MRS Hive link

* Name	<input type="text" value="hive_test"/>	Configuration Guide
* Connector	<input type="text" value="Hive"/>	
* Hadoop Type	<input type="text" value="MRS"/>	
* Manager IP ?	<input type="text"/>	Select
Authentication Method	<input type="text" value="KERBEROS"/>	
* HIVE Version ?	<input type="text" value="HIVE_3_X"/>	
* Username	<input type="text"/>	
* Password	<input type="password"/>	
* OBS storage support ?	<input type="button" value="Yes"/> <input checked="" type="button" value="No"/>	
* Run Mode ?	<input type="text" value="EMBEDDED"/>	
* Check Hive JDBC Connectivity ?	<input type="button" value="Yes"/> <input checked="" type="button" value="No"/>	
Use Cluster Config ?	<input type="button" value="Yes"/> <input checked="" type="button" value="No"/>	
Hide Advanced Attributes		
Hive Properties ?	<input type="button" value="+ Add"/>	
<input type="button" value="X Cancel"/> <input type="button" value=" < Previous"/> <input type="button" value="🔧 Test"/> <input checked="" type="button" value="💾 Save"/>		

Table 4-83 describes the parameters. You can configure the parameters according to the actual situation.

Table 4-83 MRS Hive link parameters

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	hivelink
Manager IP	Floating IP address of MRS Manager. Click Select next to the Manager IP text box to select an MRS cluster. CDM automatically fills in the authentication information.	127.0.0.1
Authentication Method	Authentication method used for accessing MRS <ul style="list-style-type: none"> • SIMPLE: Select this for non-security mode. • KERBEROS: Select this for security mode. 	SIMPLE
HIVE Version	Set this to the Hive version on the server.	HIVE_3_X
Username	<p>If Authentication Method is set to KERBEROS, you must provide the username and password used for logging in to MRS Manager. If you need to create a snapshot when exporting a directory from HDFS, the user configured here must have the administrator permission on HDFS.</p> <p>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.</p> <p>NOTE</p> <ul style="list-style-type: none"> • If the CDM cluster version is 2.9.0 or later and the MRS cluster version is 3.1.0 or later, the created user must have the permissions of the Manager_viewer role to create links on CDM. To perform operations on databases, tables, and columns of an MRS component, you also need to add the database, table, and column permissions of the MRS component to the user by following the instructions in the MRS documentation. • If the CDM cluster version is earlier than 2.9.0 or the MRS cluster version is earlier than 3.1.0, the created user must have the permissions of Manager_administrator or System_administrator to create links on CDM. • A user with only the Manager_tenant or Manager_auditor permission cannot create connections. 	cdm
Password	Password used for logging in to MRS Manager	-

Parameter	Description	Example Value
OBS storage support	The server must support OBS storage. When creating a Hive table, you can store the table in OBS.	No
Run Mode	This parameter is used only when the Hive version is HIVE_3_X . Possible values are: <ul style="list-style-type: none">• EMBEDDED: The link instance runs with CDM. This mode delivers better performance.• Standalone: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select STANDALONE or configure different agents. Note: The STANDALONE mode is used to solve the version conflict problem. If the connector versions of the source and destination ends of the same link are different, a JAR file conflict occurs. In this case, you need to place the source or destination end in the STANDALONE process to prevent the migration failure caused by the conflict.	EMBEDDED
Check Hive JDBC Connectivity	Whether to check the Hive JDBC connectivity	No
Use Cluster Config	You can use the cluster configuration to simplify parameter settings for the Hadoop connection.	No
Cluster Config Name	This parameter is valid only when Use Cluster Config is set to Yes . Select a cluster configuration that has been created. For details, see Managing Cluster Configurations .	hive_01

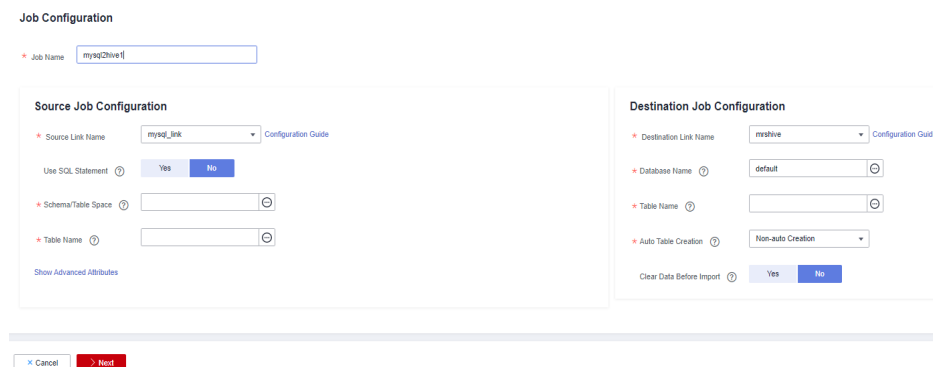
Step 3 Click **Save**. The **Link Management** page is displayed.

----End

Creating a Migration Job

Step 1 Choose **Table/File Migration > Create Job** to create a data migration job. [Figure 4-67](#) illustrates how to create a migration job.

Figure 4-67 Creating a job for migrating data from MySQL to Hive



NOTE

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

Step 2 After the parameters are configured, click **Next**. The **Map Field** tab page is displayed. See [Figure 4-68](#).

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 4-68 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 4-69](#).

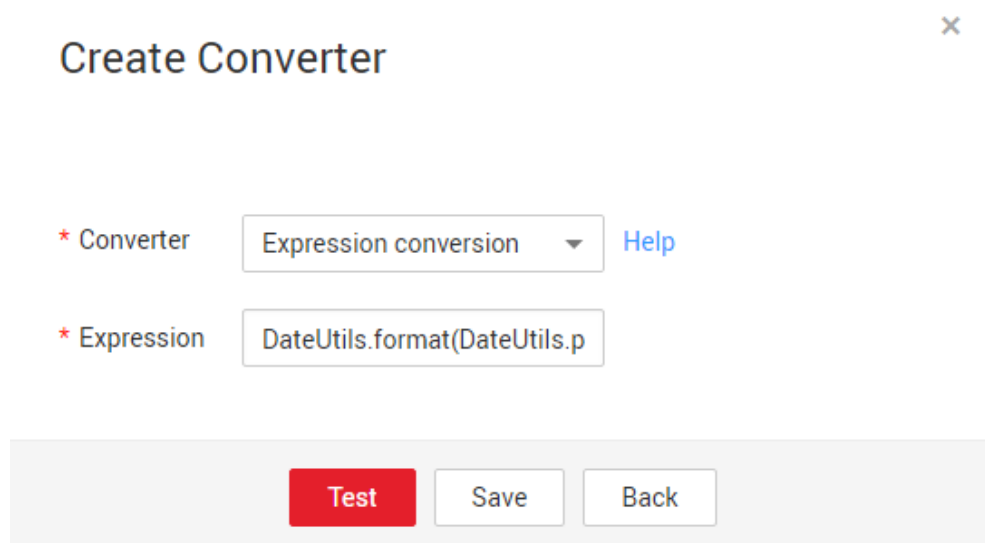
The expressions for the **y**, **ym**, and **ymd** fields are as follows:

```
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 4-69 Configuring the expression



The screenshot shows a 'Create Converter' dialog box. It has a title bar with a close button (X). The dialog contains two fields: 'Converter' with a dropdown menu set to 'Expression conversion' and a 'Help' link; and 'Expression' with a text input field containing 'DateUtils.format(DateUtils.p'. At the bottom, there are three buttons: 'Test' (red), 'Save', and 'Back'.

NOTE

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

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

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

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

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

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

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

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

----End

4.9.4 Migrating Data from MySQL to OBS

Scenario

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

1. [Creating a CDM Cluster and Binding an EIP to the Cluster](#)
2. [Creating a MySQL Link](#)
3. [Creating an OBS Link](#)
4. [Creating a Migration Job](#)

Prerequisites

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

Creating a CDM Cluster and Binding an EIP to the Cluster

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

The key configurations are as follows:

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

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

 **NOTE**

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

----End

Creating a MySQL Link

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

Figure 4-70 Selecting a connector



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

Figure 4-71 Creating a MySQL link

* Name [Configuration Guide](#)

* Connector

Database Type

* Database Server

* Port

* Database Name

* Username

* Password

Use Local API Yes No

Use Agent Yes No

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

[Hide Advanced Attributes](#)

Fetch Size

Commit Size

Link Attributes

Reference Sign

Batch Size

Click **Show Advanced Attributes** and set optional parameters. For details, see [Link to Relational Databases](#). Retain the default values of the optional parameters and configure the mandatory parameters according to [Table 4-84](#).

Table 4-84 MySQL link parameters

Parameter	Description	Example Value
Name	Unique link name	mysqllink

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

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

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

NOTE

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

----End

Creating an OBS Link

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

Figure 4-72 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.

Figure 4-73 Creating an OBS link

* Name	<input type="text" value="obslink"/>
* Connector	<input type="text" value="OBS"/>
Object Storage Type	<input type="text" value="Object Storage OBS"/>
* OBS Endpoint ?	<input type="text" value="obs.cn-north-7.ulanhqab.huav"/>
* Port ?	<input type="text" value="443"/>
* OBS Bucket Type ?	<input type="text" value="Object storage"/>
* AK ?	<input type="text"/>
* SK ?	<input type="text"/>

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.

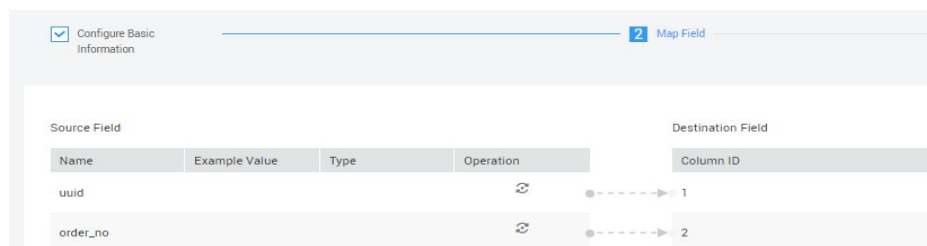
Figure 4-74 Creating a job for migrating data from MySQL to OBS

The screenshot shows the 'Job Configuration' page in DataArts Studio. It is divided into three steps: 1. Configure Basic Information, 2. Map Field, and 3. Configure Task. The 'Job Configuration' section is active. It includes a 'Job Name' field with the value 'mysql2obs_custom_file_name_tablename_s'. Below this are two columns: 'Source Job Configuration' and 'Destination Job Configuration'. The 'Source Job Configuration' includes: 'Source Link Name' (mysql_link), 'Use SQL Statement' (No), 'Schema/Table Space' (rf_test_database), and 'Table Name' (rf_varchar_test_from). The 'Destination Job Configuration' includes: 'Destination Link Name' (obs_link), 'Bucket Name' (cdm-autotest), 'Write Directory' (/to/Custom_File_Name/), and 'File Format' (CSV). At the bottom, there are 'Cancel' and 'Next' buttons.

- **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 4-75](#).

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

Figure 4-75 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

4.9.5 Migrating Data from MySQL to DWS

Scenario

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

1. [Creating a CDM Cluster and Binding an EIP to the Cluster](#)
2. [Creating a MySQL Link](#)

3. [Creating a DWS Link](#)
4. [Creating a Migration Job](#)

Prerequisites

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

Creating a CDM Cluster and Binding an EIP to the Cluster

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

The key configurations are as follows:

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

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

NOTE

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

----End

Creating a MySQL Link

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

Figure 4-76 Selecting a connector



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

Figure 4-77 Creating a MySQL link

The screenshot shows a configuration form for creating a MySQL link. The form is organized into several sections:

- Mandatory Fields:** Name (mysqllink), Connector (Relational Database), Database Type (MySQL), Database Server, Port, Database Name, Username (admin), and Password.
- Optional Fields:** Use Local API (No), Use Agent (No), and Driver Version (mysql-connector-java-5.1.48.jar).
- Advanced Attributes:** Fetch Size (1000), Commit Size (10000), Link Attributes (+ Add), Reference Sign, and Batch Size (100).

At the bottom of the form, there are four buttons: X Cancel, < Previous, Test, and Save.

Click **Show Advanced Attributes** and set optional parameters. For details, see [Link to Relational Databases](#). Retain the default values of the optional parameters and configure the mandatory parameters according to [Table 4-85](#).

Table 4-85 MySQL link parameters

Parameter	Description	Example Value
Name	Unique link name	mysqllink

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

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

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

NOTE

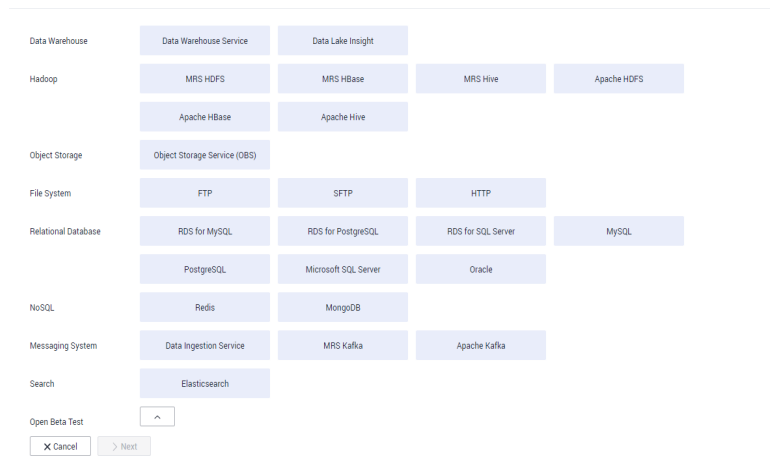
If an error occurs during the saving, the security settings of the MySQL database are incorrect. In this case, you need to enable the EIP of the CDM cluster to access the MySQL database.

----End

Creating a DWS Link

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

Figure 4-78 Selecting a connector type



Step 2 Select **Data Warehouse Service** and click **Next** to configure the DWS link parameters. Set the mandatory parameters listed in [Table 4-86](#) and retain the default values for the optional parameters.

Table 4-86 DWS link parameters

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

Step 3 Click **Save**.

----End

Creating a Migration Job

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

Figure 4-79 Creating a job for migrating data from MySQL to DWS

The screenshot shows the 'Job Configuration' window in DataArts Studio. It is divided into three main sections: 'Configure Basic Information', 'Map Field', and 'Configure Task'. The 'Configure Task' section is active and contains two sub-sections: 'Source Job Configuration' and 'Destination Job Configuration'.
In the 'Source Job Configuration' section, the 'Job Name' is 'mysql2dws_schedule'. The 'Source Link Name' is 'mysql'. The 'Use SQL Statement' is set to 'No'. The 'Schema/Table Space' is 'ppoop'. The 'Table Name' is 'test_date_char'.
In the 'Destination Job Configuration' section, the 'Destination Link Name' is 'dws'. The 'Schema/Table Space' is 'dws_job'. The 'Auto Table Creation' is set to 'Non-auto Creation'. The 'Table Name' is 'test_varchar'. The 'Clear Data Before Import' is set to 'Clear all data'. The 'Import Mode' is 'COPY'.

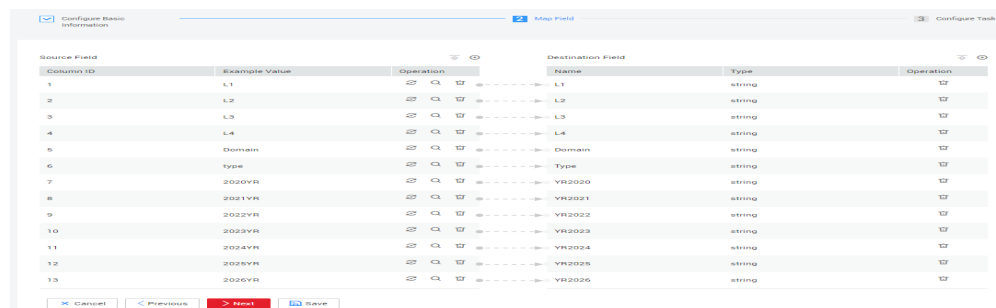
- **Job Name:** Enter a unique name.
- **Source Job Configuration**
 - **Source Link Name:** Select the **mysqllink** created in [Creating a MySQL Link](#).
 - **Use SQL Statement:** Select **No**.
 - **Schema/Tablespace:** name of the schema or tablespace from which data is to be extracted
 - **Table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

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

Step 2 Click **Next**. The **Map Field** page is displayed. CDM automatically matches the source and destination fields, as shown in [Figure 4-80](#).

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

Figure 4-80 Table-to-table field mapping



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

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

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

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

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

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

----End

4.9.6 Migrating an Entire MySQL Database to RDS

Scenario

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

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

1. [Creating a CDM Cluster and Binding an EIP to the Cluster](#)
2. [Creating a MySQL Link](#)
3. [Creating an RDS Link](#)
4. [Creating an Entire DB Migration Job](#)

Prerequisites

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

Creating a CDM Cluster and Binding an EIP to the Cluster

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

The key configurations are as follows:

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

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

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

Figure 4-81 Cluster list



NOTE

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

----End

Creating a MySQL Link

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

Figure 4-82 Selecting a connector



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

Figure 4-83 Creating a MySQL link

The screenshot shows a configuration form for creating a MySQL link. The form is organized into several sections:

- Mandatory Fields:** Name (mysqllink), Connector (Relational Database), Database Type (MySQL), Database Server, Port, Database Name, Username (admin), and Password.
- Optional Fields:** Use Local API (No), Use Agent (No), and Driver Version (mysql-connector-java-5.1.48.jar).
- Advanced Attributes:** Fetch Size (1000), Commit Size (10000), Link Attributes (+ Add), Reference Sign, and Batch Size (100).

At the bottom of the form, there are four buttons: X Cancel, < Previous, Test, and Save.

Click **Show Advanced Attributes** and set optional parameters. For details, see [Link to Relational Databases](#). Retain the default values of the optional parameters and configure the mandatory parameters according to [Table 4-87](#).

Table 4-87 MySQL link parameters

Parameter	Description	Example Value
Name	Unique link name	mysqllink

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

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

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

NOTE

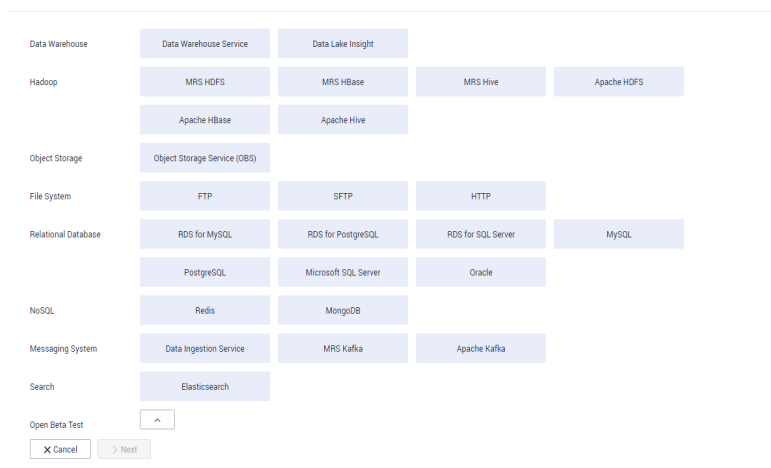
If an error occurs during the saving, the security settings of the MySQL database are incorrect. In this case, you need to enable the EIP of the CDM cluster to access the MySQL database.

----End

Creating an RDS Link

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

Figure 4-84 Selecting a connector type



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

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

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

 **NOTE**

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

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

----End

Creating an Entire DB Migration Job

Step 1 After the two links are created, choose **Entire DB Migration > Create Job** to create a migration job. See [Figure 4-85](#).

Figure 4-85 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

4.9.7 Migrating Data from Oracle to CSS

Scenario

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

1. [Creating a CDM Cluster and Binding an EIP to the Cluster](#)
2. [Creating a Cloud Search Service Link](#)
3. [Creating an Oracle Link](#)
4. [Creating a Migration Job](#)

Prerequisites

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

- You have uploaded an Oracle database driver by following the instructions provided in [Managing Drivers](#).

Creating a CDM Cluster and Binding an EIP to the Cluster

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

The key configurations are as follows:

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

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

 **NOTE**

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

----End

Creating a Cloud Search Service Link

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

Figure 4-86 Selecting a connector



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

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

Figure 4-87 Creating a CSS link

* Name

* Connector

* Elasticsearch Servers [Select](#)

Security Mode Authentication Yes No

* Username

* Password

HTTPS Access Yes No

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

----End

Creating an Oracle Link

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

Figure 4-88 Selecting a connector type



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

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

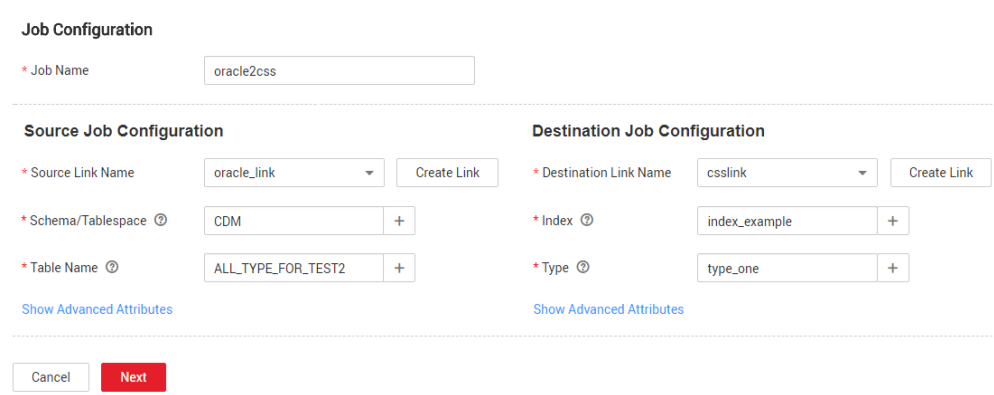
Step 3 Click **Save**. The **Link Management** page is displayed.

----End

Creating a Migration Job

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

Figure 4-89 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 4-90](#).

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

Figure 4-90 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 [Scheduling Job Execution](#). Retain the default value **No**.

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

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

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

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

----End

4.9.8 Migrating Data from Oracle to DWS

Scenario

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

1. [Creating a CDM Cluster and Binding an EIP to the Cluster](#)
2. [Creating an Oracle Link](#)
3. [Creating a DWS Link](#)
4. [Creating a Migration Job](#)

Prerequisites

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

Creating a CDM Cluster and Binding an EIP to the Cluster

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

The key configurations are as follows:

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

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

NOTE

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

----End

Creating an Oracle Link

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

Figure 4-91 Selecting a connector type



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

Figure 4-92 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 ?	<input type="text" value="192.168.0.1"/>
* Port ?	<input type="text" value="3306"/>
* Connection Type ?	<input type="text" value="Service Name"/>
* Database Name ?	<input type="text" value="db_user"/>
* Username ?	<input type="text" value="sqoop"/>
* Password ?	<input type="password"/>
Use Agent ?	<input checked="" type="radio"/> Yes <input type="radio"/> No
Agent ?	<input type="text"/> Select
Oracle Version ?	<input type="text" value="Earlier than 12.1.0.1"/>
Driver Version ?	ojdbc6-11.2.0.4.jar Upload Copy from SFTP
Hide Advanced Attributes	
Fetch Size ?	<input type="text" value="1000"/>
Link Attributes ?	<input type="button" value="+ Add"/>
Reference Sign ?	<input type="text" value=""/>
<input type="button" value="X Cancel"/> <input type="button" value="Test"/> <input type="button" value="Save"/>	

Table 4-88 Oracle link parameters

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

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

----End

Creating a DWS Link

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

Figure 4-93 Selecting a connector type



Step 2 Select **Data Warehouse Service** and click **Next** to configure the DWS link parameters. Set the mandatory parameters listed in [Table 4-89](#) and retain the default values for the optional parameters.

Table 4-89 DWS link parameters

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

Step 3 Click **Save**.

----End

Creating a Migration Job

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

Figure 4-94 Creating a job for migrating data from Oracle to DWS

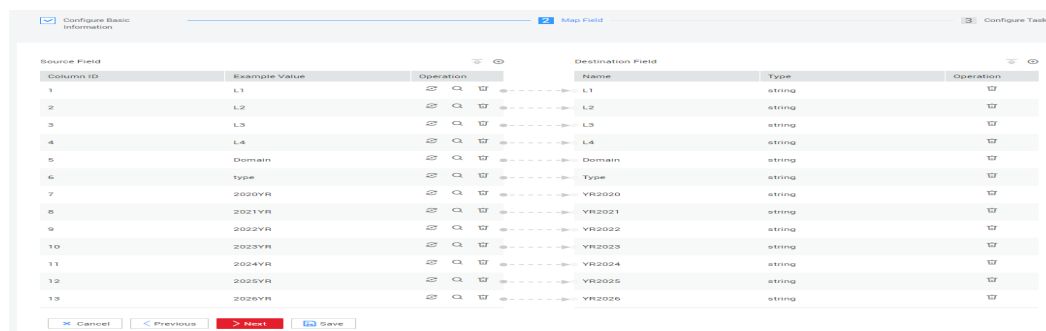
- **Job Name:** Enter a unique name.
- **Source Job Configuration**
 - **Source Link Name:** Select the **oracle_link** created in [Creating an Oracle Link](#).
 - **Schema/Tablespace:** Enter the name of the database whose data is to be migrated.
 - **Table Name:** Enter the name of the table whose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

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

Step 2 Click **Next**. The **Map Field** page is displayed. CDM automatically matches the source and destination fields, as shown in [Figure 4-95](#).

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

Figure 4-95 Table-to-table field mapping



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

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

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

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

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

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

----End

 **NOTE**

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

4.9.9 Migrating Data from OBS to CSS

Scenario

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

1. [Creating a CDM Cluster](#)
2. [Creating a Cloud Search Service Link](#)
3. [Creating an OBS Link](#)
4. [Creating a Migration Job](#)

Prerequisites

- You have obtained the domain name, port number, AK, and SK for accessing OBS.
- You have subscribed to Cloud Search Service and obtained the IP address and port number of the Cloud Search Service cluster.

Creating a CDM Cluster

If is an independent CDM service, create a CDM cluster by following the instructions provided in [Creating a Cluster](#). If is used as a CDM component of DataArts Studio, create a CDM cluster by following the instructions provided in [Creating a Cluster](#).

The key configurations are as follows:

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

Creating a Cloud Search Service Link

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

Figure 4-96 Selecting a connector



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

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

Figure 4-97 Creating a CSS link

The screenshot shows a form for creating a CSS link. The fields are as follows:

- Name:** Text input field containing "csslink".
- Connector:** Dropdown menu showing "Elasticsearch".
- Elasticsearch Servers:** Text input field, currently empty, with a "Select" link to its right.
- Security Mode Authentication:** Radio button group with "Yes" selected.
- Username:** Text input field, currently empty.
- Password:** Text input field, currently empty.
- HTTPS Access:** Radio button group with "Yes" selected.

At the bottom of the form, there are four buttons: "Cancel", "Previous", "Test", and "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 4-98 Selecting a connector type



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

- **Name:** Enter a custom link name, for example, **obslink**.
- **OBS Server** and **Port:** Enter the actual OBS address information.
- **AK** and **SK:** Enter the AK and SK used for logging in to OBS.

Figure 4-99 Creating an OBS link

* Name	<input type="text" value="obslink"/>
* Connector	<input type="text" value="OBS"/>
Object Storage Type	<input type="text" value="Object Storage OBS"/>
* OBS Endpoint ?	<input type="text" value="obs.cn-north-7.ulanhqab.huaw"/>
* Port ?	<input type="text" value="443"/>
* OBS Bucket Type ?	<input type="text" value="Object storage"/>
* AK ?	<input type="text"/>
* SK ?	<input type="text"/>

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 4-100 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"/>	Show Advanced Attributes
Show Advanced Attributes	

- **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 4-101](#).

- If the field mapping is incorrect, you can drag the fields to adjust the mapping.
- If the type is automatically created at the migration destination, you need to configure the type and name of each field.

- CDM supports field conversion during the migration. For details, see [Converting Fields](#).

Figure 4-101 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 [Scheduling Job Execution](#). Retain the default value **No**.
- **Concurrent Extractors:** Enter the number of extractors to be concurrently executed. Retain the default value **1**.
- **Write Dirty Data:** Specify this parameter if data that fails to be processed or filtered out during job execution needs to be written to OBS for future viewing. Before writing dirty data, create an OBS link. Retain the default value **No** so that dirty data is not recorded.
- **Delete Job After Completion:** Retain the default value **Do not delete**.

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

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

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

----End

4.9.10 Migrating Data from OBS to DLI

Scenario

DLI is a fully hosted big data query service. This section describes how to use CDM to migrate data from OBS to DLI. The procedure includes four steps:

1. [Creating a CDM Cluster](#)

2. [Creating a DLI Link](#)
3. [Creating an OBS Link](#)
4. [Creating a Migration Job](#)

Prerequisites

- You have enabled OBS and DLI and have the permissions to read data from OBS.
- You have created resource queues, databases, and tables on DLI.

Creating a CDM Cluster

If is an independent CDM service, create a CDM cluster by following the instructions provided in [Creating a Cluster](#). If is used as a CDM component of DataArts Studio, create a CDM cluster by following the instructions provided in [Creating a Cluster](#).

In this scenario, if the CDM cluster is used only to migrate data from OBS to DLI and does not need to migrate data of other data sources, there is no special requirements on the VPC, subnet, and security group of the CDM cluster. You can specify them based on your needs. CDM accesses DLI and OBS through the intranet. The flavor of the CDM cluster is selected based on the amount of data to be migrated. Generally, `cdm.medium` meets the requirements for most migration scenarios.

Creating a DLI Link

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




Figure 4-102 Selecting a connector



- Step 2** Select **Data Lake Insight**, click **Next**, and configure the DLI link parameters. See [Figure 4-103](#).

- **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 4-103 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="c48475ce8e174a7a9f77570i"/>

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 4-104 Selecting a connector type

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

- **Name:** Enter a custom link name, for example, **obslink**.
- **OBS Server** and **Port:** Enter the actual OBS address information.
- **AK** and **SK:** Enter the AK and SK used for logging in to OBS.

Figure 4-105 Creating an OBS link

* Name	<input type="text" value="obslink"/>
* Connector	<input type="text" value="OBS"/>
Object Storage Type	<input type="text" value="Object Storage OBS"/>
* OBS Endpoint 	<input type="text" value="obs.cn-north-7.ulanhqab.huaw"/>
* Port 	<input type="text" value="443"/>
* OBS Bucket Type 	<input type="text" value="Object storage"/>
* AK 	<input type="text"/>
* SK 	<input type="text"/>

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 4-106](#).

Figure 4-106 Creating a job for migrating data from OBS to DLI

Job Configuration

* Job Name

Source Job Configuration

* Source Link Name

* Bucket Name

* Source Directory/File

* File Format

Show advanced attributes.

Destination Job Configuration

* Destination Link Name

* Resource Queue

* Database Name

* Table Name

Clear Data Before Import

- **Job Name:** Enter a custom job name.
- **Source Link Name:** Select the **obslink** link created in [Creating an OBS Link](#).
 - **Bucket Name:** Select the bucket from which the data is to be migrated.
 - **Source Directory/File:** Set this parameter to the path of the data to be migrated.
 - **File Format:** Select **CSV** or **JSON** for transferring files to a data table.
 - Retain the default values of the optional parameters in **Show Advanced Attributes**. For details, see [From OBS](#).
- **Destination Link Name:** Select the **dlilink** link created in [Creating a DLI Link](#).
 - **Resource Queue:** Enter the resource queue to which the destination table belongs.
 - **Database Name:** Enter the name of the database to which data is to be written.
 - **Table Name:** Enter the name of the table to which data is to be written. CDM cannot automatically create tables on DLI. The table must be created on DLI in advance, and the field types and formats of the table must be consistent with those of the data to be migrated.
 - **Clear Before Importing Data:** Choose whether to clear data in the destination table before data import. In this example, retain the default value.

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

- If the field mapping is incorrect, you can drag the fields to adjust the mapping.
- CDM supports field conversion during the migration. For details, see [Converting Fields](#).

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

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

- **Retry Upon Failure:** If the job fails to be executed, you can determine whether to automatically retry. Retain the default value **Never**.

- **Group:** Select the group to which the job belongs. The default group is **DEFAULT**. On the **Job Management** page, jobs can be displayed, started, or exported by group.
- **Schedule Execution:** To configure scheduled jobs, see [Scheduling Job Execution](#). Retain the default value **No**.
- **Concurrent Extractors:** Enter the number of extractors to be concurrently executed. Retain the default value **1**.
- **Write Dirty Data:** Specify this parameter if data that fails to be processed or filtered out during job execution needs to be written to OBS for future viewing. Before writing dirty data, create an OBS link. Retain the default value **No** so that dirty data is not recorded.
- **Delete Job After Completion:** Retain the default value **Do not delete**.

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

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

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

----End

4.9.11 Migrating Data from MRS HDFS to OBS

Scenario

CDM supports file-to-file data migration. This section describes how to migrate data from MRS HDFS to OBS. The process is as follows:

1. [Creating a CDM Cluster and Binding an EIP to the Cluster](#)
2. [Creating an MRS HDFS Link](#)
3. [Creating an OBS Link](#)
4. [Creating a Migration Job](#)

Prerequisites

- You have obtained the domain name, port number, AK, and SK for accessing OBS.
- MRS is available.
- Your EIP quota is sufficient.

Creating a CDM Cluster and Binding an EIP to the Cluster

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

The key configurations are as follows:

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

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

NOTE

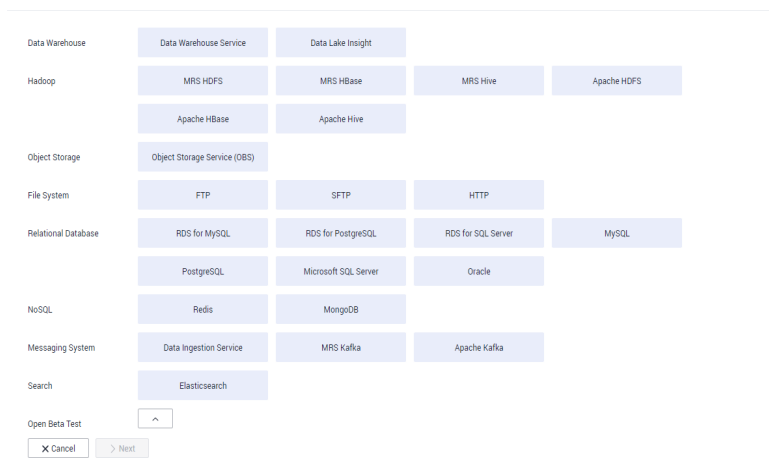
If SSL encryption is configured for the access channel of a local data source, CDM cannot connect to the data source using the EIP.

----End

Creating an MRS HDFS Link

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

Figure 4-107 Selecting a connector type



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

- **Name:** Enter a custom link name, for example, `mrs_hdfs_link`.
- **Manager IP:** IP address of MRS Manager. Click **Select** next to the **Manager IP** text box to select a created MRS cluster. CDM automatically fills in the authentication information.
- **Username:** If **Authentication Method** is set to **KERBEROS**, set the username and password for logging in to MRS Manager.
If you need to create a snapshot when exporting a directory from HDFS, the user configured here must have the administrator permission on HDFS.
- **Password:** password for logging in to MRS Manager
- **Authentication Method:** authentication method for accessing MRS

- **Run Mode:** Select the running mode of the HDFS link.
- End

Creating an OBS Link

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

Figure 4-108 Selecting a connector type



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

- **Name:** Enter a custom link name, for example, **obslink**.
- **OBS Server** and **Port:** Enter the actual OBS address information.
- **AK** and **SK:** Enter the AK and SK used for logging in to OBS.

Figure 4-109 Creating an OBS link

* Name	<input type="text" value="obslink"/>
* Connector	<input type="text" value="OBS"/>
Object Storage Type	<input type="text" value="Object Storage OBS"/>
* OBS Endpoint ?	<input type="text" value="obs.cn-north-7.ulanhqab.huav"/>
* Port ?	<input type="text" value="443"/>
* OBS Bucket Type ?	<input type="text" value="Object storage"/>
* AK ?	<input type="text"/>
* SK ?	<input type="text"/>

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 4-110 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

Destination Job Configuration

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

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. For details, see [Converting Fields](#).

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

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

- **Retry Upon Failure:** If the job fails to be executed, you can determine whether to automatically retry. Retain the default value **Never**.

- **Group:** Select the group to which the job belongs. The default group is **DEFAULT**. On the **Job Management** page, jobs can be displayed, started, or exported by group.
- **Schedule Execution:** To configure scheduled jobs, see [Scheduling Job Execution](#). Retain the default value **No**.
- **Concurrent Extractors:** Enter the number of extractors to be concurrently executed. CDM supports concurrent extraction of multiple files. Increasing the value of this parameter can improve migration efficiency.
- **Write Dirty Data:** Select **No**. The file-to-file migration is binary, and no dirty data will be generated.
- **Delete Job After Completion:** Retain the default value **Do not delete**. You can also set this parameter to **Delete** to prevent an accumulation of too many migration jobs.

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

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

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

----End

4.9.12 Migrating the Entire Elasticsearch Database to CSS

Scenario

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

1. [Creating a CDM Cluster and Binding an EIP to the Cluster](#)
2. [Creating a Cloud Search Service Link](#)
3. [Creating an Elasticsearch Link](#)
4. [Creating an Entire DB Migration Job](#)

Prerequisites

- You have sufficient EIP quota.
- You have subscribed to CSS and obtained the IP address and port number of the CSS cluster.
- You have obtained the IP address, port number, username, and password of the on-premises Elasticsearch database server.

If the Elasticsearch server is deployed on an on-premises data center or a third-party cloud, ensure that an IP address that can be accessed from the public network has been configured for the Elasticsearch server, or the VPN or Direct Connect between the on-premises data center and HUAWEI CLOUD has been established.

Creating a CDM Cluster and Binding an EIP to the Cluster

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

The key configurations are as follows:

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

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

NOTE

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

----End

Creating a Cloud Search Service Link

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

Figure 4-111 Selecting a connector



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

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

Figure 4-112 Creating a CSS link

The screenshot shows a configuration form for creating a CSS link. The fields and their values are as follows:

- Name:** csslink
- Connector:** Elasticsearch
- Elasticsearch Servers:** (empty field) with a [Select](#) button to the right.
- Security Mode Authentication:** Yes (selected)
- Username:** (empty field)
- Password:** (empty field)
- HTTPS Access:** Yes (selected)

At the bottom of the form, there are four buttons: [Cancel](#), [Previous](#), [Test](#), and [Save](#).

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

----End

Creating an Elasticsearch Link

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

Figure 4-113 Selecting a connector type



Step 2 Select **Elasticsearch** and click **Next** to configure parameters for the Elasticsearch link. The parameters are the same as those for the CSS link.

- **Name:** Enter a custom link name, for example, **es_link**.
- **Elasticsearch Server List:** Enter the IP address and port number of the on-premises Elasticsearch database. Use semicolons to separate multiple addresses.

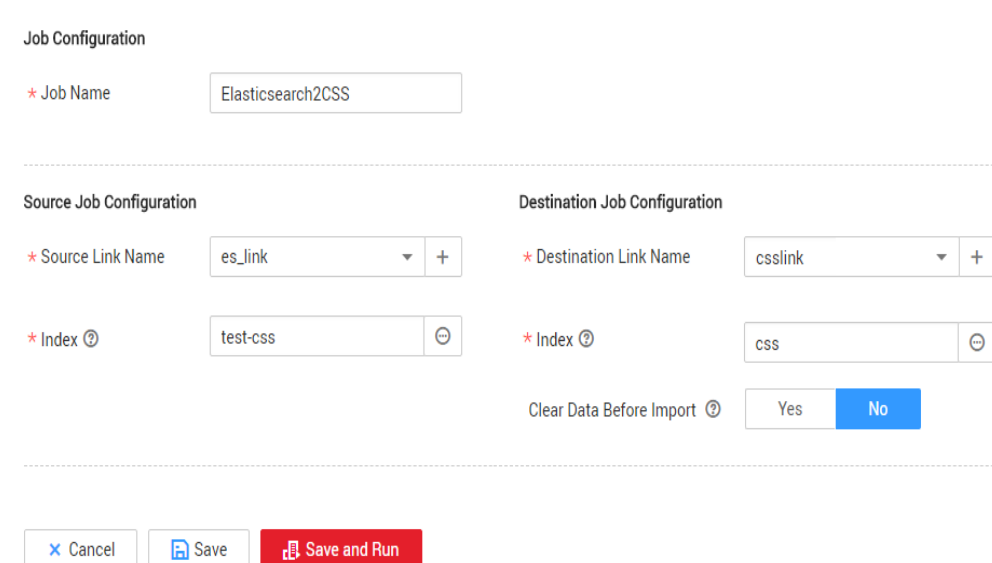
Step 3 Click **Save**. The **Link Management** page is displayed.

----End

Creating an Entire DB Migration Job

Step 1 Choose **Entire DB Migration > Create Job** to create an entire DB migration job.

Figure 4-114 Creating an entire DB migration job



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

- **Source Job Configuration**
 - **Source Link Name:** Select the **es_link** link created in [Creating an Elasticsearch Link](#).
 - **Index:** Click the icon next to the text box to select an index in the on-premises Elasticsearch database or manually enter an index name. The name can contain only lowercase letters. If multiple indexes need to be migrated at a time, set this parameter to a wildcard character. CDM migrates all indexes that meet the wildcard condition. For example, if this parameter is set to **cdm***, CDM migrates all indexes starting with **cdm**, such as **cdm01**, **cdmB3**, **cdm_45** and so on.
- **Destination Job Configuration**
 - **Destination Link Name:** Select the **csslink** link created in [Creating a Cloud Search Service Link](#).
 - **Index:** Enter the index of the data to be written. You can select an existing index in Cloud Search Service or manually enter an index name that does not exist. The name can contain only lowercase letters. CDM automatically creates the index in Cloud Search Service. If multiple indexes are migrated at a time, this parameter cannot be configured. CDM automatically creates indexes at the migration destination.
 - **Clear Data Before Import:** If the selected index already exists in Cloud Search Service, you can choose whether to clear the data in the index before importing data. If you select **No**, the data is added to the index.

Step 2 Click **Save and Run**. The **Job Management** page is displayed, on which you can view the job execution progress and result.

A sub-job will be generated for each type in the on-premises Elasticsearch index for concurrent execution. You can click the job name to view the sub-job progress.

Step 3 After the job is successfully executed, in the **Operation** column of the job, click **Historical Record** to view the job's historical execution records, read/write statistics, and job logs (only the sub-jobs have job logs).

Figure 4-115 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

4.9.13 Migrating Data from DDS to DWS

Scenario

CDM allows you to migrate data from DDS to other data sources. This section describes how to use CDM to migrate data from DDS to DWS. The procedure includes four steps:

1. [Creating a CDM Cluster and Binding an EIP to the Cluster](#)
2. [Creating a DDS Link](#)
3. [Creating a DWS Link](#)
4. [Creating a Migration Job](#)

Prerequisites

- DWS/DDS is available.
- You have obtained the IP address, port number, database name, username, and password for connecting to the DWS and DDS databases. In addition, you must have the read, write, and delete permissions for the DDS and DWS databases.

Creating a CDM Cluster and Binding an EIP to the Cluster

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

The key configurations are as follows:

- The flavor of the CDM cluster is selected based on the amount of data to be migrated. Generally, cdm.medium meets the requirements for most migration scenarios.
- If DDS and DWS are deployed in the same VPC, the newly created CDM cluster also needs to be deployed in that VPC, with no EIP bound. The CDM cluster's subnet and security group can be the same as those of the DDS or DWS cluster. You can also configure a security group rule to enable the CDM cluster to access the cluster of another service (DWS or DDS).
- If DDS and DWS are not deployed in the same VPC, the newly created CDM cluster needs to be in the same VPC as DDS and **an EIP must be bound** for the CDM cluster to access 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 DWS. If DDS and DWS are in the same VPC, do not bind an EIP to the CDM cluster.

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 DDS 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 4-116 Selecting a connector



Step 2 Select **Document Database Service** and click **Next** to configure parameters for the DDS link.

Table 4-90 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	mongo_link
Server List	Address list of the DDS cluster. The format is IP address or domain name of the database server:port number . Separate multiple server lists by semicolons (;).	192.168.0.1:7300;192.168.0.2:7301
Database Name	Name of the DDS database to be connected	DB_mongodb
Username	Username used for logging in to the DDS database	cdm
Password	Password used for logging in to the DDS database	-

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

----End

Creating a DWS Link

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

Figure 4-117 Selecting a connector



Step 2 Select **Data Warehouse Service** and click **Next** to configure the DWS link parameters. Set the mandatory parameters listed in [Table 4-91](#) and retain the default values for the optional parameters.

Table 4-91 DWS link parameters

Parameter	Description	Example Value
Name	Unique link name	dwslink
Database Server	IP address or domain name of the DWS database server	192.168.0.3
Port	DWS database port	8000
Database Name	Name of the DWS database	db_demo
Username	User who has the read, write, and delete permissions on the DWS database	dbadmin
Password	Password of the user	-
Use Agent	Whether to extract data from the data source through an agent	Yes
Agent	Click Select to select the agent created in Connecting to an Agent .	-

Step 3 Click **Save**.

----End

Creating a Migration Job

Step 1 Choose **Table/File Migration > Create Job** to create a data migration job.

Figure 4-118 Creating a job for migrating data from DDS to DWS

Job Configuration

* Job Name

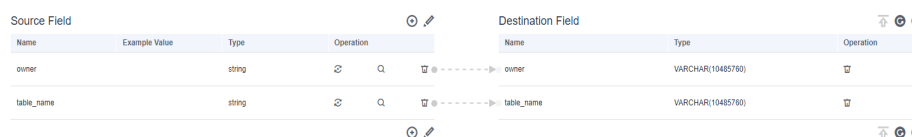
Source Job Configuration	Destination Job Configuration
* Source Link Name <input type="text" value="mongo_link"/> <input type="button" value="Create Link"/>	* Destination Link Name <input type="text" value="dwslink"/> <input type="button" value="Create Link"/>
* Database <input type="text" value="test"/> <input type="button" value="+"/> <small>?</small>	* Schema/Table Space <input type="text" value="cstore"/> <input type="button" value="+"/> <small>?</small>
* Collection Name <input type="text" value="kafka"/> <input type="button" value="+"/> <small>?</small>	* Table Name <input type="text" value="pg_delta_36677"/> <input type="button" value="+"/> <small>?</small>
	Clear data before import <input type="button" value="Yes"/> <input checked="" type="button" value="No"/> <small>?</small>
	Show Advanced Attributes

Step 2 Configure the required job information:

- **Job Name:** Enter a unique job name.
- **Source Job Configuration**
 - **Source Link Name:** Select the **mongo_link** link created in [Creating a DDS Link](#).
 - **Database Name:** Select the database whose data is to be migrated.
 - **Collection Name:** Enter the name of the MongoDB collection on DDS, which is similar to the table name in a relational database.
- **Destination Job Configuration**
 - **Destination Link Name:** Select the **dwslink** link created in [Creating a DWS Link](#).
 - **Schema/Tablespace:** Select the DWS database to which data is to be written.
 - **Table Name:** Name of the table to which data is to be written. You can manually enter a table name that does not exist. CDM automatically creates the table on DWS.
 - **Clear Data Before Import:** Choose whether to clear data in the destination table before data import.

- Step 3** Click **Next**. The **Map Field** tab page is displayed. CDM automatically maps table fields at the migration source and destination. Check whether the field mapping is correct.
- If the field mapping is incorrect, click the row where the field is located and drag the field to adjust the mapping.
 - When importing data to DWS, you need to manually select the distribution columns of DWS. You are advised to select the distribution columns according to the following principles:
 - a. Use the primary key as the distribution column.
 - b. If multiple data segments are combined as primary keys, specify all primary keys as the distribution column.
 - c. 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.
 - If you want to convert the content of the source fields, perform the operations in this step. For details, see [Converting Fields](#). In this example, field conversion is not required.

Figure 4-119 Field mapping



Source Field				Destination Field		
Name	Example Value	Type	Operation	Name	Type	Operation
owner		string	Q	owner	VARCHAR(10485760)	
table_name		string	Q	table_name	VARCHAR(10485760)	

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

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

- **Retry Upon Failure:** If the job fails to be executed, you can determine whether to automatically retry. Retain the default value **Never**.
 - **Group:** Select the group to which the job belongs. The default group is **DEFAULT**. On the **Job Management** page, jobs can be displayed, started, or exported by group.
 - **Schedule Execution:** To configure scheduled jobs, see [Scheduling Job Execution](#). Retain the default value **No**.
 - **Concurrent Extractors:** Enter the number of extractors to be concurrently executed. Retain the default value **1**.
 - **Write Dirty Data:** Specify this parameter if data that fails to be processed or filtered out during job execution needs to be written to OBS for future viewing. Before writing dirty data, create an OBS link. Retain the default value **No** so that dirty data is not recorded.
 - **Delete Job After Completion:** Retain the default value **Do not delete**.
- Step 5** Click **Save and Run**. The **Job Management** page is displayed, on which you can view the job execution progress and result.
- Step 6** After the job is successfully executed, in the **Operation** column of the job, click **Historical Record** to view the job's historical execution records and read/write statistics.

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

----End

4.9.14 More Cases and Practices

For more advanced guidance and cases of DataArts Migration, see [Best Practices](#).

4.10 Advanced Operations

4.10.1 Incremental Migration

4.10.1.1 Incremental File Migration

CDM supports incremental migration of file systems. After full migration is complete, all new files or only specified directories or files can be exported.

Currently, CDM supports the following incremental migration modes:

1. **Exporting the files in a specified directory**
 - Application scenarios: The migration source is a file system (OBS/HDFS/FTP/SFTP). In incremental migration, only the specified files are written to the migration destination. The existing records are not updated or deleted.
 - Key configurations: [File/Path Filter](#) and Schedule Execution
 - Prerequisites: The source directory or file name contains the time field.
2. **Exporting the files modified after the specified time point**
 - Application scenarios: The migration source is a file system (OBS/HDFS/FTP/SFTP). The specified time point refers to the time when the file is modified. CDM migrates the files modified after the specified time point.
 - Key configurations: [Time Filter](#) and Schedule Execution
 - Prerequisites: None

NOTE

If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with *(Planned start time of the data development job - Offset)* rather than *(Actual start time of the CDM job - Offset)*.

File/Path Filter

- Parameter position: When creating a table/file migration job, if the migration source is a file system, set **Filter Type** in advanced attributes of **Source Job Configuration** to **Wildcard** or **Regular expression**.
- Parameter principle: If you select **Wildcard** for **Filter Type**, CDM filters files or paths based on the configured wildcard character and migrates only files or paths that meet the specified condition.

- Example configurations:

Suppose that the source file name contains the date and time field, such as **2017-10-15 20:25:26**, the **/opt/data/file_20171015202526.data** file is generated. Set the parameters as follows:

 - Filter Type:** Select **Wildcard**.
 - File Filter:** Enter **"*\${dateformat(yyyyMMdd,-1,DAY)}*"**, which is the format of the macro variables of date and time supported by CDM. For details, see [Using Macro Variables of Date and Time](#).

Figure 4-120 Filtering files

Source Job Configuration

* Source Link Name [Configuration Guide](#)

* Source Directory/File

* File Format

[Hide Advanced Attributes](#)

Line Separator

Field Delimiter

Use Quote Char

Using RE to separate fields

First Row As Header

Encode Type

Compression Format

Start Job by Marker File

File Separator

Filter Type

Directory Filter

File Filter

Time Filter

Minimum Timestamp

- 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 4-121 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

Using RE to separate fields

First Row As Header

Encode type

Compression Format

Start Job by Marker File

File Separator

Filter Type

Time Filter Yes No

Minimum Timestamp

Maximum Timestamp

Disregard Non-existent Path/File

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.

4.10.1.2 Incremental Migration of Relational Databases

CDM supports incremental migration of relational databases. After a full migration is complete, data in a specified period can be incrementally migrated. For example, data added on the previous day can be exported at 00:00:00 every day.

- **Migrating incremental data within a specified period of time**
 - Application scenarios: The source end is a relational database. The destination end can be of any type.
 - Key configurations: **WHERE Clause** and Schedule Execution

- Prerequisites: The data table contains a date and time field or timestamp field.

In incremental migration, only the specified data is written to the data table. The existing records are not updated or deleted.

NOTE

If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with *(Planned start time of the data development job - Offset)* rather than *(Actual start time of the CDM job - Offset)*.

WHERE Clause

- Parameter position: When creating a table/file migration job, if the source end is a relational database, the **Where Clause** parameter is available in the advanced attributes of **Source Job Configuration**.
- Parameter principle: Set **WHERE Clause** to an SQL statement, for example, **age > 18 and age <= 60**, CDM exports only the data that meets the SQL statement requirement. If **WHERE Clause** is not specified, the entire table is exported.

Where Clause can be set to **macro variables of date and time**. When the data table contains the **date** or **timestamp** field, **Where Clause** and Schedule Execution can be used together to extract data of a specified date.

- Example configurations:

Suppose that the database table contains column **DS** indicating the time, the value type of the column is **varchar(30)**, and the inserted time format is similar to *2017-xx-xx*. See **Figure 4-122**. Set the parameters as follows:

Figure 4-122 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

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

Figure 4-123 WHERE Clause

The screenshot shows the 'Source Job Configuration' window. The 'Where Clause' field is highlighted with a red border and contains the text `DS='${dateformat(yyyy-MM-dd,-1,DAY)}'`. Other fields include: Source Link Name (mysql_link), Use SQL Statement (No), Schema/Table Space (sqoop), Table Name (trip), Partition Column, Partition column nullable (No), and Split Job (No).

- b. Scheduling job execution: Set **Cycle (days)** to **1** and **Start Time** to **00:00:00**.

In this way, all data generated on the previous day can be exported at 00:00:00 every day. **WHERE Clause** can be configured to various **macro variables of date and time**. You can use the macro variables of date and time and scheduled jobs with specified cycle of minutes, hours, days, weeks, or months together to automatically export data at a specific time.

4.10.1.3 HBase/CloudTable Incremental Migration

You can use CDM to export data in a specified period of time from HBase (including MRS HBase, FusionInsight HBase, and Apache HBase) and CloudTable. The CDM scheduled jobs can be used together to implement incremental migration of HBase and CloudTable.

NOTE

If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with *(Planned start time of the data development job – Offset)* rather than *(Actual start time of the CDM job – Offset)*.

When creating a table/file migration job and selecting the link to HBase or CloudTable as the source link, you can set the time range in advanced attributes.

Figure 4-124 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.

4.10.2 Using Macro Variables of Date and Time

During the creation of table/file migration jobs, CDM supports the macro variables of date and time in the following parameters of the source and destination links:

- Source directory or file
- Source table name
- Directory filter and file filter of the **wildcard** type
- Start time and end time of the **time filter** type
- Partition filter criteria and where clause

- Write directory
- Destination table name

You can use the `${}` macro variable definition identifier to define the macros of the time type. currently, `dateformat` and `timestamp` are supported.

By using the macro variables of date and time and scheduled job, you can implement incremental synchronization of databases and files.

NOTE

If you have configured a macro variable of date and time and schedule a CDM job through DataArts Studio DataArts Factory, the system replaces the macro variable of date and time with *(Planned start time of the data development job - Offset)* rather than *(Actual start time of the CDM job - Offset)*.

dateformat

`dateformat` supports two types of parameters:

- **dateformat(format)**
format indicates the date and time format. For details about the format definition, see the definition in `java.text.SimpleDateFormat.java`.
For example, if the current date is **2017-10-16 09:00:00**, **yyyy-MM-dd HH:mm:ss** indicates **2017-10-16 09:00:00**.

- `dateformat(format, dateOffset, dateType)`
 - **format** indicates the format of the returned date.
 - **dateOffset** indicates the date offset.
 - **dateType** indicates the type of the date offset.
Currently, **dateType** supports SECOND, MINUTE, HOUR, and DAY.

For example, if the current date is **2017-10-16 09:00:00**, then:

- `dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)` indicates the day before the current day, that is, **2017-10-15 09:00:00**.
- `dateformat(yyyy-MM-dd HH:mm:ss, -1, HOUR)` indicates one hour before the current time, that is, **2017-10-16 08:00:00**.
- `dateformat(yyyy-MM-dd HH:mm:ss, -1, MINUTE)` indicates one minute before the current time, that is, **2017-10-16 08:59:00**.
- `dateformat(yyyy-MM-dd HH:mm:ss, -1, SECOND)` indicates one second before the current time, that is, **2017-10-16 08:59:59**.

timestamp

`timestamp` supports two types of parameters:

- **timestamp()**
Indicates the returned timestamp of the current time, that is, the number of milliseconds that have elapsed since 00:00:00 on January 1, 1970 (1970-01-01 00:00:00 GMT). For example, 1508078516286.
- **timestamp(dateOffset, dateType)**
Indicates the timestamp returned after time offset. **dateOffset** and **dateType** indicate the date offset and the offset type, respectively.

For example, if the current date is **2017-10-16 09:00:00**, **timestamp(-10, MINUTE)** indicates that the timestamp generated 10 minutes before the current time point is returned, that is, **1508115000000**.

Macro Variable Definition of Time and Date

Suppose that the current time is **2017-10-16 09:00:00**, then [Table 4-92](#) describes the macro variable definitions of time and date.

Table 4-92 Macro variable definition of time and date

Macro Variable	Description	Display Effect
<code>\${dateformat(yyyy-MM-dd)}</code>	Returns the current date in yyyy-MM-dd format.	2017-10-16
<code>\${dateformat(yyyy/MM/dd)}</code>	Returns the current date in yyyy/MM/dd format.	2017/10/16
<code>\${dateformat(yyyy_MM_dd HH:mm:ss)}</code>	Returns the current time in yyyy_MM_dd HH:mm:ss format.	2017_10_16 09:00:00
<code>\${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}</code>	Returns the current time in yyyy-MM-dd HH:mm:ss format. The date is one day before the current day.	2017-10-15 09:00:00
<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

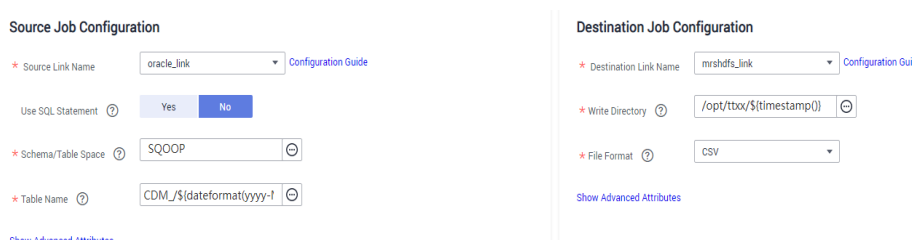
Time and Date Macro Variables of Paths and Table Names

[Figure 4-125](#) 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 4-125 Setting **Table Name** and **Write Directory** to a time and date macro variable



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

Time and Date Macro Variables in the Where Clause

Figure 4-126 uses table **SQOOP.CDM_20171016** as an example. The table contains column **DS**, which indicates the time.

Figure 4-126 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.

4.10.3 Migration in Transaction Mode

When a CDM job fails to be executed, CDM rolls back the data to the state before the job starts and automatically deletes data from the destination table.

- Parameter position: When creating a table/file migration job, if the migration source is a relational database, set **Import to Staging Table** in the advanced attributes of **Destination Job Configuration** to determine whether to enable the transaction mode.
- Parameter principle: If you set this parameter to **Yes**, CDM automatically creates a temporary table and imports the data to the temporary table. After the data is imported successfully, CDM migrates the data to the destination table in transaction mode of the database. If the import fails, the destination table is rolled back to the state before the job starts.

Figure 4-127 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.

4.10.4 Encryption and Decryption During File Migration

When you migrate files to a file system, CDM can encrypt and decrypt those files. Currently, CDM supports the following encryption modes:

- [AES-256-GCM](#)
- [KMS Encryption](#)

AES-256-GCM

Currently, only AES-256-GCM (NoPadding) is supported. This algorithm is used for encryption at the migration destination and decryption at the migration source. The supported source and destination data sources are as follows:

- Data sources supported by the migration source: OBS, FTP, SFTP, HDFS (supported in the binary format), and HTTP (applicable to scenarios where OBS shared files are downloaded)
- Data sources supported by the migration destination: OBS, FTP, SFTP, and HDFS (supported in the binary format)

The following part describes how to use AES-256-GCM to decrypt the encrypted files to be exported from OBS and encrypt the files to be imported to OBS. The methods for using the algorithm on other data sources are the same.

- **Configure decryption at the migration source.**

When you use CDM to create a job for exporting files from OBS, set the migration source to OBS and set the following parameters in the advanced settings of **Source Job Configuration**:

- a. **Encryption:** Select **AES-256-GCM**.
- b. **DEK:** The key must be the same as that configured in **Encryption**. Otherwise, the decrypted data is incorrect and the system does not display an error message.
- c. **IV:** The initialization vector must be the same as that configured in **Encryption**. Otherwise, the decrypted data is incorrect and the system does not display an error message.

In this way, after CDM exports encrypted files from OBS, the files written to the migration destination are decrypted plaintext files.

- **Configure encryption at the migration destination.**

When you use CDM to create a job for importing files to OBS, set the migration destination to OBS and set the following parameters in the advanced settings of **Destination Job Configuration**:

- a. **Encryption:** Select **AES-256-GCM**.
- b. **DEK:** custom encryption key. The key consists of 64 hexadecimal numbers. It is case-insensitive but must contain 64 characters. For example, **DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B**.
- c. **IV:** custom initialization vector. The initialization vector consists of 32 hexadecimal numbers. It is case-insensitive but must contain 32 characters. For example, **5C91687BA886EDCD12ACBC3FF19A3C3F**.

In this way, after CDM imports files to OBS, the files on the migration destination are encrypted using the AES-256-GCM algorithm.

KMS Encryption

 **NOTE**

The migration source does not support KMS encryption.

CDM supports KMS encryption if tables, files, or a whole database is migrated to OBS. In the **Advanced Attributes** area of the **Destination Job Configuration** page, set the parameters.

A key must be created in KMS of DEW in advance. For details, see the *Data Encryption Workshop User Guide*.

After KMS encryption is enabled, objects to be uploaded will be encrypted and stored on OBS. When you download the encrypted objects, the encrypted data will be decrypted on the server and displayed in plaintext to users.

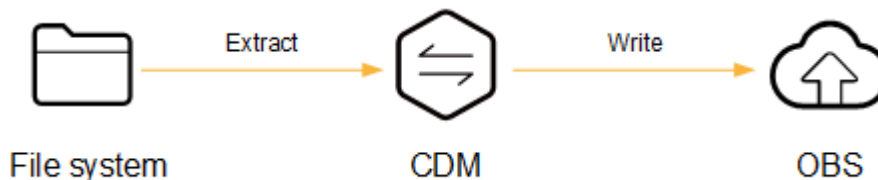
NOTE

- If KMS encryption is enabled, **MD5 verification** cannot be used.
- If the KMS ID of another project is used, change **Project ID** to the ID of the project to which KMS belongs. If KMS and CDM are in the same project, retain the default value of **Project ID**.
- After KMS encryption is performed, the encryption status of the objects on OBS cannot be changed.
- A key in use cannot be deleted. Otherwise, the object encrypted with this key cannot be downloaded.

4.10.5 MD5 Verification

CDM extracts data from the migration source and writes the data to the migration destination. **Figure 4-128** shows the migration mode when files are migrated to OBS.

Figure 4-128 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.

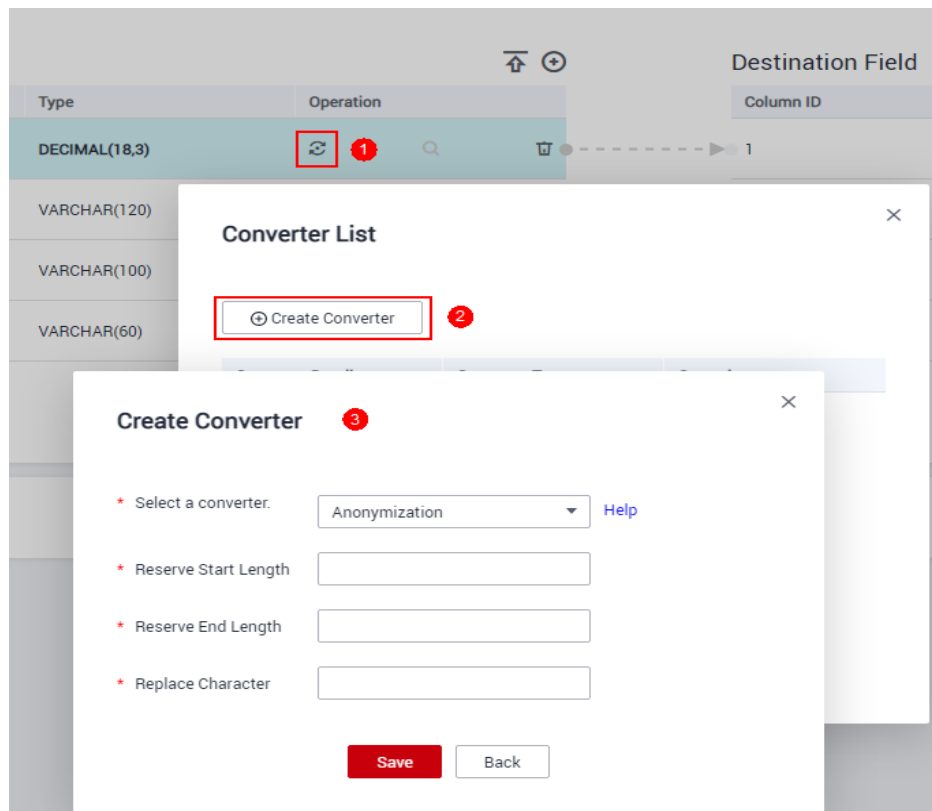
NOTE

- 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.

4.10.6 Field Conversion

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

Figure 4-129 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:

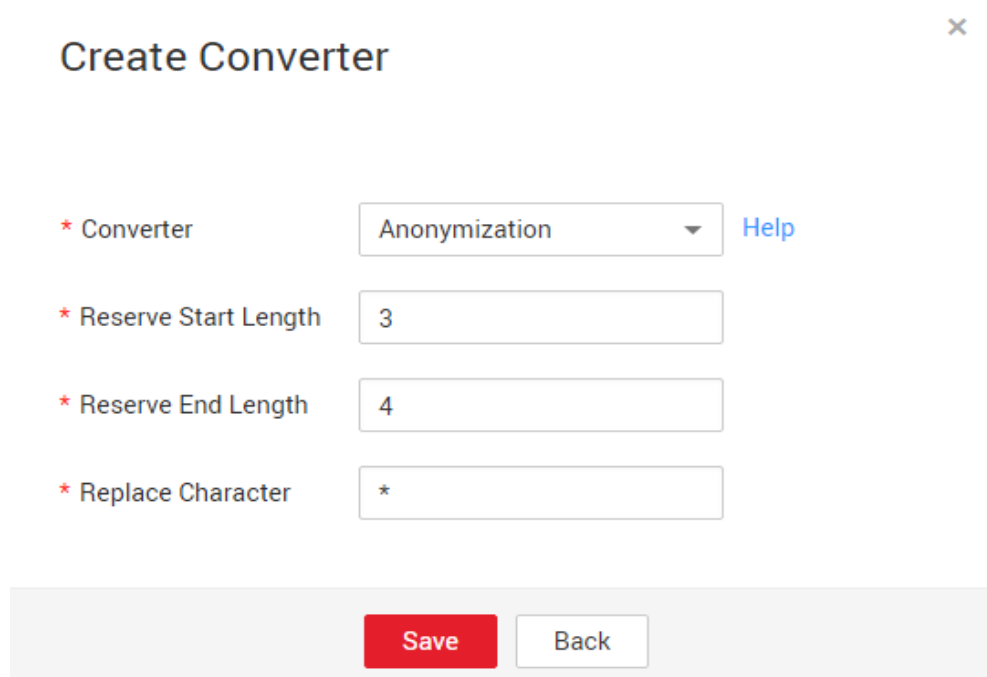
- [Anonymization](#)
- [Trim](#)
- [Reverse String](#)
- [Replace String](#)
- [Remove line break](#)
- [Expression Conversion](#)

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 4-130 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:

- * Converter**: A dropdown menu with 'Anonymization' selected and a 'Help' link to the 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 input 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.

Remove line break

This converter is used to delete the newline characters, such as `\n`, `\r`, and `\r\n` from the field.

Expression Conversion

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

The expression supports the following environment variables:

- **value**: indicates the current field value.
- **row**: indicates the current row, which is an array type.

The expression supports the following tool classes:

- **StringUtils**: string processing tool class. For details, see **org.apache.commons.lang.StringUtils** of the Java SDK code.
- **DateUtils**: date tool class
- **CommonUtils**: common tool class
- **NumberUtils**: string-to-value conversion class
- **HttpsUtils**: network file read class

Application examples:

1. If the field is of the string type, convert all character strings into lowercase letters, for example, convert **aBC** to **abc**.
Expression: `StringUtils.lowerCase(value)`
2. Convert all character strings of the current field to uppercase letters.
Expression: `StringUtils.upperCase(value)`
3. If the field value is a date string in *yyyy-MM-dd* format, extract the year from the field value, for example, extract **2017** from **2017-12-01**.
Expression: `StringUtils.substringBefore(value, "-")`
4. If the field value is of the numeric type, convert the value to a new value which is two times greater than the original value:
Expression: `value*2`
5. Convert the field value **true** to **Y** and other field values to **N**.
Expression: `value=="true"? "Y": "N"`
6. If the field value is of the string type and is left empty, convert it to **Default**. Otherwise, the field value will not be converted.
Expression: `empty value? "Default":value`
7. Convert date format **2018/01/05 15:15:05** to **2018-01-05 15:15:05**:

- Expression: `DateUtils.format(DateUtils.parseDate(value,"yyyy/MM/dd HH:mm:ss"),"yyyy-MM-dd HH:mm:ss")`
8. Obtain a 36-bit universally unique identifier (UUID):
Expression: `CommonUtils.randomUUID()`
 9. If the field is of the string type, capitalize the first letter, for example, convert **cat** to **Cat**.
Expression: `StringUtils.capitalize(value)`
 10. If the field is of the string type, convert the first letter to a lowercase letter, for example, convert **Cat** to **cat**.
Expression: `StringUtils.uncapitalize(value)`
 11. If the field is of the string type, use a space to fill in the character string to the specified length and center the character string. If the length of the character string is not shorter than the specified length, do not convert the character string. For example, convert **ab** to meet the specified length 4.
Expression: `StringUtils.center(value,4)`
 12. Delete a newline (including `\n`, `\r`, and `\r\n`) at the end of a character string. For example, convert **abc\r\n\r\n** to **abc\r\n**.
Expression: `StringUtils.chomp(value)`
 13. If the string contains the specified string, **true** is returned; otherwise, **false** is returned. For example, **abc** contains **a** so that **true** is returned.
Expression: `StringUtils.contains(value,"a")`
 14. If the string contains any character of the specified string, **true** is returned; otherwise, **false** is returned. For example, **zzabyycdxx** contains either **z** or **a** so that **true** is returned.
Expression: `StringUtils.containsAny("value","za")`
 15. If the string does not contain any one of the specified characters, **true** is returned. If any specified character is contained, **false** is returned. For example, **abz** contains one character of **xyz** so that **false** is returned.
Expression: `StringUtils.containsNone(value,"xyz")`
 16. If the string contains only the specified characters, **true** is returned. If any other character is contained, **false** is returned. For example, **abab** contains only characters among **abc** so that **true** is returned.
Expression: `StringUtils.containsOnly(value,"abc")`
 17. If the character string is empty or null, convert it to the specified character string. Otherwise, do not convert the character string. For example, convert the empty character string to null.
Expression: `StringUtils.defaultIfEmpty(value,null)`
 18. If the string ends with the specified suffix (case sensitive), **true** is returned; otherwise, **false** is returned. For example, if the suffix of **abcdef** is not null, **false** is returned.
Expression: `StringUtils.endsWith(value,null)`
 19. If the string is the same as the specified string (case sensitive), **true** is returned; otherwise, **false** is returned. For example, after strings **abc** and **ABC** are compared, **false** is returned.
Expression: `StringUtils.equals(value,"ABC")`

20. Obtain the first index of the specified character string in a character string. If no index is found, **-1** is returned. For example, the first index of **ab** in **aabaabaa** is 1.
Expression: `StringUtils.indexOf(value,"ab")`
21. Obtain the last index of the specified character string in a character string. If no index is found, **-1** is returned. For example, the last index of **k** in **aFkyk** is 4.
Expression: `StringUtils.lastIndexOf(value,"k")`
22. Obtain the first index of the specified character string from the position specified in the character string. If no index is found, **-1** is returned. For example, the first index of **b** obtained after the index 3 of **aabaabaa** is 5.
Expression: `StringUtils.indexOf(value,"b",3)`
23. Obtain the first index of any specified character in a character string. If no index is found, **-1** is returned. For example, the first index of **z** or **a** in **zzabyycdxx** is 0.
Expression: `StringUtils.indexOfAny(value,"za")`
24. If the string contains any Unicode character, **true** is returned; otherwise, **false** is returned. For example, **ab2c** contains only non-Unicode characters so that **false** is returned.
Expression: `StringUtils.isAlpha(value)`
25. If the string contains only Unicode characters and digits, **true** is returned; otherwise, **false** is returned. For example, **ab2c** contains only Unicode characters and digits, so that **true** is returned.
Expression: `StringUtils.isAlphanumeric(value)`
26. If the string contains only Unicode characters, digits, and spaces, **true** is returned; otherwise, **false** is returned. For example, **ab2c** contains only Unicode characters and digits, so that **true** is returned.
Expression: `StringUtils.isAlphanumericSpace(value)`
27. If the string contains only Unicode characters and spaces, **true** is returned; otherwise, **false** is returned. For example, **ab2c** contains Unicode characters and digits so that **false** is returned.
Expression: `StringUtils.isAlphaSpace(value)`
28. If the string contains only printable ASCII characters, **true** is returned; otherwise, **false** is returned. For example, for **!ab-c~**, **true** is returned.
Expression: `StringUtils.isAsciiPrintable(value)`
29. If the string is empty or null, **true** is returned; otherwise, **false** is returned.
Expression: `StringUtils.isEmpty(value)`
30. If the string contains only Unicode digits, **true** is returned; otherwise, **false** is returned.
Expression: `StringUtils.isNumeric(value)`
31. Obtain the leftmost characters of the specified length. For example, obtain the leftmost two characters **ab** from **abc**.
Expression: `StringUtils.left(value,2)`
32. Obtain the rightmost characters of the specified length. For example, obtain the rightmost two characters **bc** from **abc**.
Expression: `StringUtils.right(value,2)`

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

4.10.7 Migrating Files with Specified Names

You can migrate files (a maximum of 50) with specified names from FTP, SFTP, or OBS at a time. The exported files can only be written to the same directory on the migration destination.

When creating a table/file migration job, if the migration source is FTP, SFTP, or OBS, **Source Directory/File** can contain a maximum of 50 file names, which are separated by vertical bars (|). You can also customize a file separator.

NOTE

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

4.10.8 Regular Expressions for Separating Semi-structured Text

During table/file migration, CDM uses delimiters to separate fields in CSV files. However, delimiters cannot be used in complex semi-structured data because the field values also contain delimiters. In this case, the regular expression can be used to separate the fields.

The regular expression is configured in **Source Job Configuration**. The migration source must be an object storage or file system, and **File Format** must be **CSV**.

Figure 4-131 Setting regular expression parameters

Source Job Configuration

* Source Link Name	obs-dayu-demo
* Bucket Name ?	abcsze ...
* Source Directory/File ?	/DAS_Imexport_Import_9e14 ...
* File Format ?	CSV v
Hide Advanced Attributes	
Line Separator ?	
Use Quote Char ?	Yes No
Using RE to separate fields ?	Yes No
Regular Expression ?	
First Row As Header ?	Yes No
Encode type ?	UTF-8
Compression Format ?	NONE v
Source File Processing Method ?	Do Nothing v

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 4-93 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 4-94 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 4-95 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 4-96 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 4-97 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

4.10.9 Recording the Time When Data Is Written to the Database

When you create a job on the CDM console to migrate tables or files of a relational database, you can add a field to record the time when they were written to the database.

Prerequisites

A link has been created, and the source end of the connector is a relational database.

Creating a Table/File Migration Job

Step 1 Create a table/file migration job, and select the created source connector and destination connector.

Figure 4-132 Configuring the job

The screenshot shows the 'Job Configuration' window. At the top, the 'Job Name' is 'mz_mysqlLdli'. Below this, there are two main sections: 'Source Job Configuration' and 'Destination Job Configuration'.
Source Job Configuration:
 * Source Link Name: mz_mysql (dropdown menu)
 Use SQL Statement: Yes/No (radio buttons, 'No' is selected)
 * Schema or Table Space: mztest (dropdown menu)
 * Table Name: t_trade_order (dropdown menu)
 A link 'Show Advanced Attributes' is visible below.
Destination Job Configuration:
 * Destination Link Name: mz_dli (dropdown menu)
 * Resource Queue: dayu_demo (dropdown menu)
 * Database Name: mz_dli (dropdown menu)
 * Table Name: t_trade_order (dropdown menu)
 Clear Data Before Import: Yes/No (radio buttons, 'No' is selected)

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

Figure 4-133 Configuring field mapping

Source Field	Destination Field	Operation	Source Field	Destination Field	Operation
id	id	✓	id	id	✓
name	name	✓	name	name	✓
type	type	✓	type	type	✓
2022-11-11	2022-11-11	✓	2022-11-11	2022-11-11	✓
2022-11-11	2022-11-11	✓	2022-11-11	2022-11-11	✓
2022-11-11	2022-11-11	✓	2022-11-11	2022-11-11	✓
2022-11-11	2022-11-11	✓	2022-11-11	2022-11-11	✓
2022-11-11	2022-11-11	✓	2022-11-11	2022-11-11	✓
2022-11-11	2022-11-11	✓	2022-11-11	2022-11-11	✓
2022-11-11	2022-11-11	✓	2022-11-11	2022-11-11	✓

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 4-98](#).

Figure 4-134 Add Field

The screenshot shows the 'Add Field' dialog box. At the top, there is a 'Destination Field' header with a '+' icon in a red box. Below this, there are two tabs: 'Add removed fields' and 'Add custom fields' (which is selected).
 The 'Name' field contains 'InputTime'.
 The 'Value' field contains '\${timestamp()}'.
 At the bottom, there are 'OK' and 'Cancel' buttons.

Table 4-98 Macro variable definition of time and date

Macro Variable	Description	Display Effect
<code>\${dateformat(yyyy-MM-dd)}</code>	Returns the current date in yyyy-MM-dd format.	2017-10-16
<code>\${dateformat(yyyy/MM/dd)}</code>	Returns the current date in yyyy/MM/dd format.	2017/10/16
<code>\${dateformat(yyyy_MM_dd HH:mm:ss)}</code>	Returns the current time in yyyy_MM_dd HH:mm:ss format.	2017_10_16 09:00:00
<code>\${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}</code>	Returns the current time in yyyy-MM-dd HH:mm:ss format. The date is one day before the current day.	2017-10-15 09:00:00
<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

4.10.10 File Formats

When creating a CDM job, you need to specify **File Format** in the job parameters of the migration source and destination in some scenarios. This section describes the application scenarios, subparameters, common parameters, and usage examples of the supported file formats.

- [CSV](#)
- [JSON](#)
- [Binary](#)
- [Common parameters](#)
- [Solutions to File Format Problems](#)

CSV

To read or write a CSV file, set **File Format** to **CSV**. The CSV format can be used in the following scenarios:

- Import files to a database or NoSQL.
- Export data from a database or NoSQL to files.

After selecting the CSV format, you can also configure the following optional sub-parameters:

1. Line Separator

2. Field Delimiter

3. **Encoding Type**
4. **Use Quote Character**
5. **Use RE to Separate Fields**
6. **Use First Row as Header**
7. **File Size**

1. **Line Separator**

Character used to separate lines in a CSV file. The value can be a single character, multiple characters, or special characters. Special characters can be entered using the URL encoded characters. The following table lists the URL encoded characters of commonly used special characters.

Table 4-99 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 4-99](#).

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 4-135](#) shows that the value of the **name** field in the database contains a comma (,).

Figure 4-135 Field value containing the field delimiter

	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

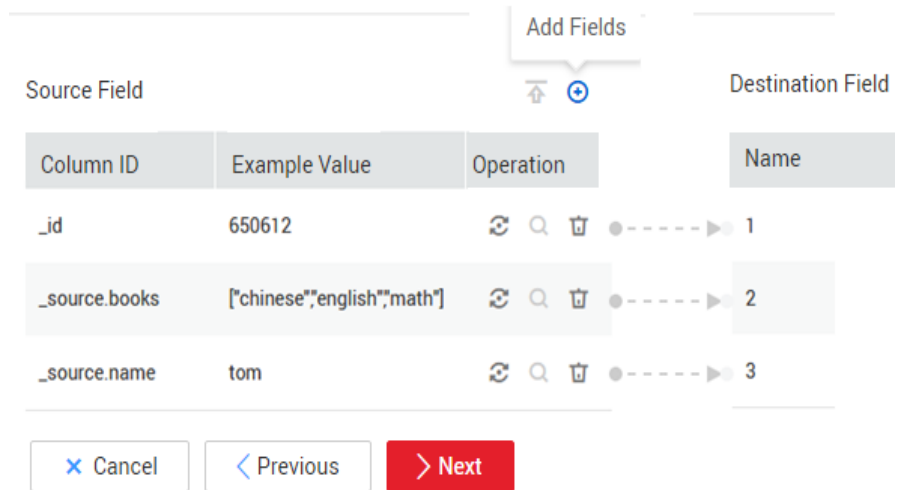
- d. Example 4: Configure a converter when parsing the JSON file. On the premise of [example 2](#), to add the **hits.max_score** field to all records, that is, to write the data to the database in the following formats, perform the following operations:

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 4-136 Adding a field




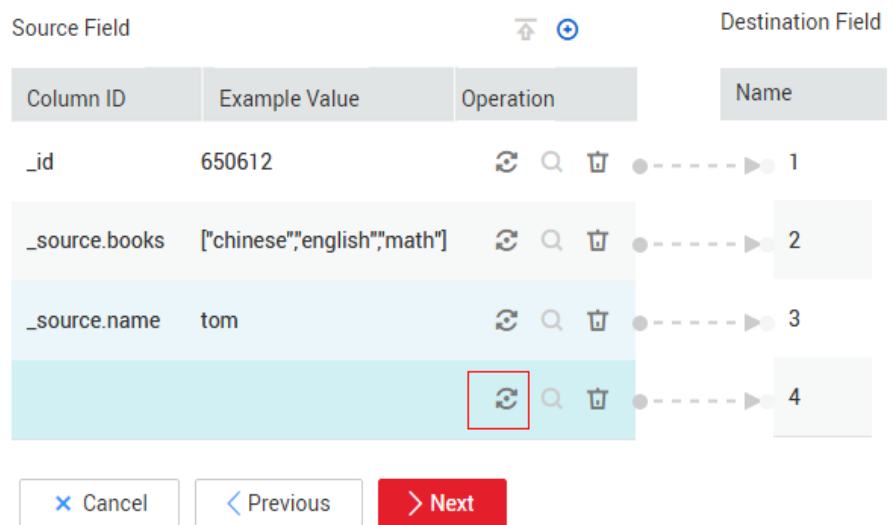
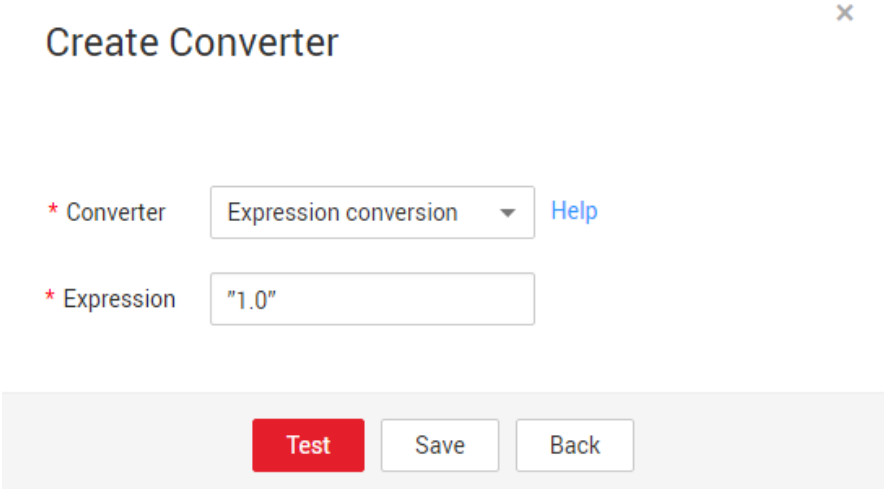
- ii. Click  to create a converter for the new field.

Figure 4-137 Creating a field converter



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

Figure 4-138 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 4-99](#).

- 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.

5 DataArts Architecture

5.1 Overview

Introduction to DataArts Architecture

DataArts Architecture can be used to create entity-relationship (ER) models and dimensional models to standardize and visualize data development and output data governance methods that can guide development personnel to work with ease.

DataArts Architecture is for processing and commercializing data, and is the core module of data governance. It consists of four parts: data survey, standards design, model design, and metric design. DataArts Architecture supports DLI, POSTGRESQL, DWS, MRS Hive, and MRS Spark connections. (It supports MRS Hudi data sources through MRS Spark.)

DataArts Architecture aims to build:

- A unified data classification system to manage all business data in directories for easier data classification, search, evaluation, and use.
- A unified data standards system that complies with national or industrial standards to standardize each row of data and each field value and improve data quality and usability.
- A unified data model system and a tiered enterprise data system from top to bottom based on standards definitions and data modeling. These systems can be used to construct enterprises' public data layers and subject libraries, facilitating data flow, sharing, creation, and innovation. They will make data usage more efficient, greatly reducing data redundancy, disorder, isolation, inconsistencies, and inaccuracies.

Model Design Method Overview

A data model can reflect the relationships between objects. It incorporates the key information features extracted based on business requirements. It visually represents how the internal information of an enterprise is organized. A data model must be capable of simulating scenarios, easy-to-understand, and easily implemented in the IT system.

ER and dimensional modeling are both used on DataArts Architecture.

- **ER modeling**

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

There are three different models involved in ER modeling: design conceptual models, logical models, and physical models.

- **Conceptual model** is used to represent business processes and business data involved in various activities. A conceptual model illustrates the relationships between business entities.
- **Logical model** is much more detailed than the conceptual model. Logical models outline business details based on entities, attributes, and relationships. They enable communication between IT and business staff. A logical model is a set of standardized logical table structures. Based on business rules, a logical model outlines business objects, data items of the business objects, and relationships between business objects.
- **Physical model:** An advanced version of the logic model and used to design the database architecture for data storage with a full consideration of various technical factors. For example, the selected data warehouse is DWS or MRS_Hive.

- **Dimensional modeling**

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

A multidimensional model is a fact table consisting of numeric metrics. The fact table is associated with a group of dimensional tables containing description attributes with primary or foreign keys.

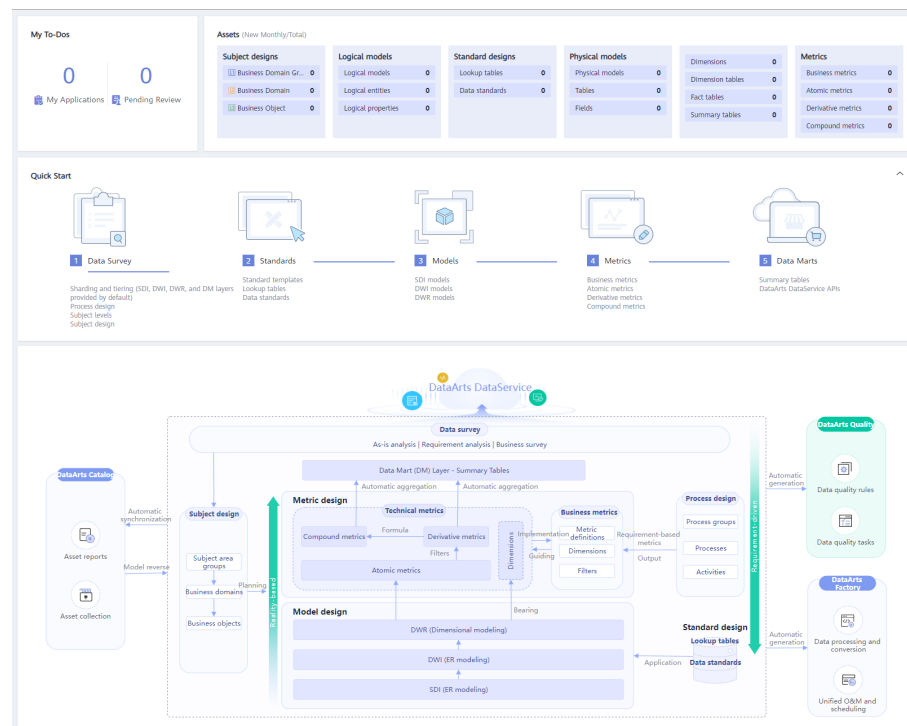
Typical dimensional models include star models and snowflake models used in some special scenarios.

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

DataArts Architecture Overview Page

On the DataArts Studio console, locate a workspace and click **DataArts Architecture**. The **Overview** page is displayed.

Figure 5-1 DataArts Architecture Overview page



- **My To-Dos**
 - The **My To-Dos** area displays the quantity of **My Applications** and **Pending Review**.
 - Click the numbers above **My Applications** and **Pending Review** to access the **My Applications** and **Pending Review** pages, respectively.
- **Assets**
 - The **Assets** area displays all the objects in DataArts Architecture.
 - Click the number next to each object name to access the object management page.
- **Quick Start**

The **Quick Start** area displays the overall process for data governance. You can click a specific operation under the process to go to the corresponding page.
- **DataArts Architecture Process**
 - This area displays the DataArts Architecture process and how the DataArts Architecture module interacts with other modules of DataArts Studio. For details about the DataArts Architecture process, see [DataArts Architecture Use Process](#).
 - You can move the cursor over the name of an object to view its description.
 - You can click the name of any object supported by DataArts Studio to access the object management page.

Information Architecture of DataArts Architecture

An information architecture is a set of component specifications that describe various types of information required for business operations and management decision-making as well as the relationships of business entities. On the **Information Architecture** page, you can view and manage all tables, including business tables, dimension tables, fact tables, and summary tables.

On the DataArts Studio console, locate a workspace and click **DataArts Architecture**. In the navigation pane, choose **Information Architecture**.

Perform the following operations on the **Information Architecture** page.

- **Search**

On the top of the **Information Architecture** page, click **Advanced Search**, set the table name, type, data source, and other filters, and click **Search** to search for a specific table. Then click the table name to access its details page.

- **Create**

Click **Create** to create a logical model, physical model, dimension table, fact table, or summary table. For details, see [Designing Logical Models](#), [Designing Physical Models](#), [Creating Dimensions](#), [Creating Fact Tables](#), or [Creating Summary Tables](#).

- **Import**

Choose **More > Import**. (Currently, only tables can be imported.) Download the table template, fill in it, and upload it. Then click **Close**. For details, see [Importing/Exporting Tables](#).

- **Export**

Choose **More > Export** to export a physical table model or DDL. For details, see [Exporting a Table or DDL](#).

- **Synchronize**

Choose **More > Synchronize** to synchronize table information to DataArts Catalog as technical assets or synchronize logical models to DataArts Catalog as logical assets.

- **Modify Subject**

Choose **More > Modify Subject** to change the selected table to another subject.

- **Delete**

Choose **More > Delete** to delete a data table. A data table in the pending publishing, published, or pending suspension state cannot be deleted. A referenced data table cannot be deleted either.

- **Suspend**

Choose **More > Suspend** to suspend a published data table. A referenced data table cannot be suspended.

 **NOTE**

Edited versions refer to the data that is re-edited after published.

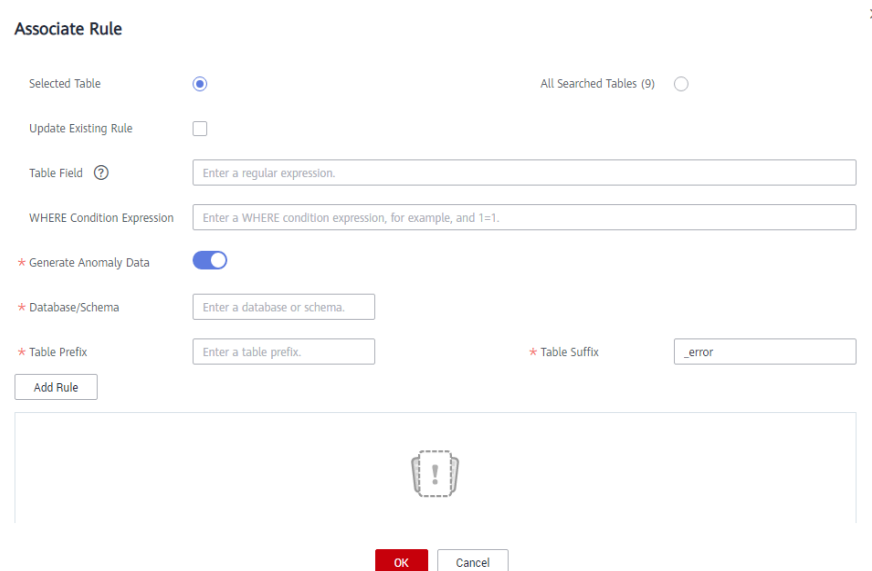
- **Publish**

Click **Publish** to publish a data table. Data tables in the pending publishing, pending suspension, or published (without edited versions) state cannot be published.

- **Associate Rule**

Click **Associate Rule** and set the parameters to associate a quality rule with the object you select. For details, see [Associating Quality Rules](#).

Figure 5-2 Associating a quality rule with an object

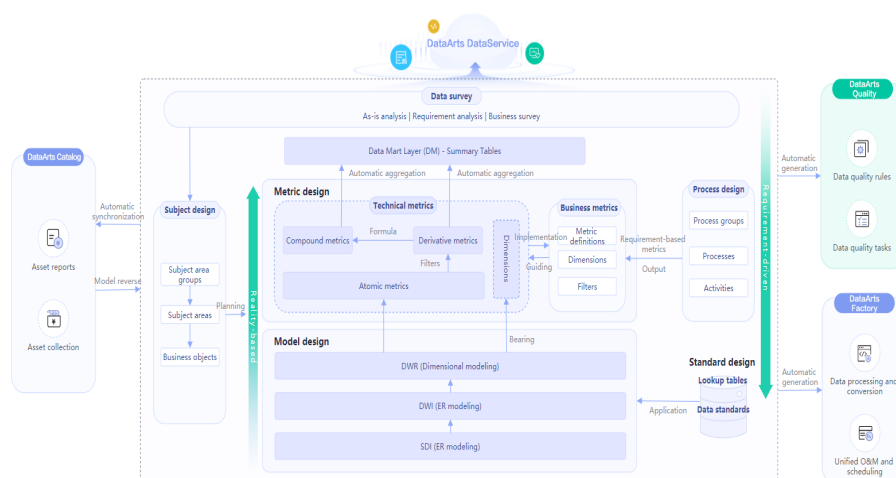


Generate Anomaly Data: If this option is selected, anomaly data is stored in the specified database based on the configured parameters.

5.2 DataArts Architecture Use Process

The process of using DataArts Architecture is as follows.

Figure 5-3 DataArts Architecture use process



1. Preparations

- **Add reviewers:** In the DataArts Architecture module, all business processes must be approved. Therefore, add reviewers first before conducting any operations. Only the workspace admin has the permissions required to add reviewers.
 - **Configuration Center** provides abundant custom options. You can customize the configuration to meet your demands.
2. **Data Survey:** A data survey involves collecting data that is generated when sorting business requirements, creating business processes, and classifying data subjects based on the existing business data and industry status.
- **Subject design** is a hierarchical architecture that classifies and defines data to help clarify data assets and specify relationships between business domains and business objects.
 - **Subject area group** is used to group business domains based on scenarios.
 - **Subject area** is the high-level data classification that does not overlap and is used to manage business objects.
 - **Business object** includes important information about people, events, and things that are indispensable to enterprise operations and management.
 - **Process design** is used to generate a structured framework of process. It describes the categories, levels, boundaries, scopes, and input/output relationships of an enterprise's processes, and reflects the business models and characteristics of the enterprise.
3. **Standards:** Create lookup tables and data standards.
- A **lookup table** includes a series of allowed values and additional text descriptions that are generally associated with data standards to generate a range of values for the verification of quality monitoring rules.
 - **Data standards** refer to the description of attribute data meanings and business rules that enterprises must comply with. It describes the common understanding of certain data at the company level.
4. **Models:** Use ER modeling and dimensional modeling methods to perform hierarchical modeling.
- **ER modeling:** Create SDI and DWI models based on ER modeling.
 - **SDI** stands for Source Data Integration and is the source data layer. SDI is a simple implementation of source system data.
 - **DWI** stands for Data Warehouse Integration, also called the data consolidation layer. DWI integrates and cleans data from multiple source systems, and implements entity relationship modeling based on the three normal forms.
 - **Dimensional modeling:** Create DWR models and release dimensions and fact tables based on ER modeling.
 - **Data Warehouse Report (DWR)** is based on the multi-dimensional model and its data granularity is the same as that of the DWI layer.

- **Dimension** is the perspective to observe and analyze business data and assist in data aggregation, drilling, slicing, and analysis, and used as a GROUP BY condition in SQL statements.
 - A **fact table** that belongs to a business process can enrich the affair information corresponding to the specific business process.
5. **Metrics:** Create business and technical metrics. Technical metrics include atomic, derivative, and compound metrics.
- A **metric** consists of its name and value. The metric name and its definition reflect the quality and quantity of the metric. The metric value reflects the quantifiable values of the specified time, location, and condition of the metric.

Business metrics are used to guide technical metrics, and technical metrics are used to implement business metrics.
 - **Atomic metrics** are generated based on dimension tables and fact tables of a multidimensional model. The objects and the finest data granularity of an atomic metric are consistent with those of the multidimensional model.

An atomic metric usually consists of measures and attributes related with measures and business objects, all of which aim to support agile self-service consumption of the metric.
 - **Derivative metrics** are aggregated from the definitions, modifiers, and dimensions of atomic metrics. Therefore, their definitions, modifiers, and dimensions are derived from the attributes of atomic metric associated tables as well.
 - **Compound metrics** are generated by adding one or more derivative metrics. The dimensions and modifiers of a compound metric are the same as those of the derivative metrics.

New dimensions and modifiers cannot be generated outside the scope of derivative metrics, dimensions, and modifiers.
6. **Data mart:** Create a DM layer and release summary tables.
- **Data Mart (DM)** is where multiple types of data are summarized. DM is designed to display the summarized data.
 - A **summary table** consists of specific analysis objects (for example, members) and related statistical metrics. The statistical metrics included in a summary table have the same statistical granularity (for example, members). The summary table provides users with all statistics-granularity-themed data (such as a member theme market).

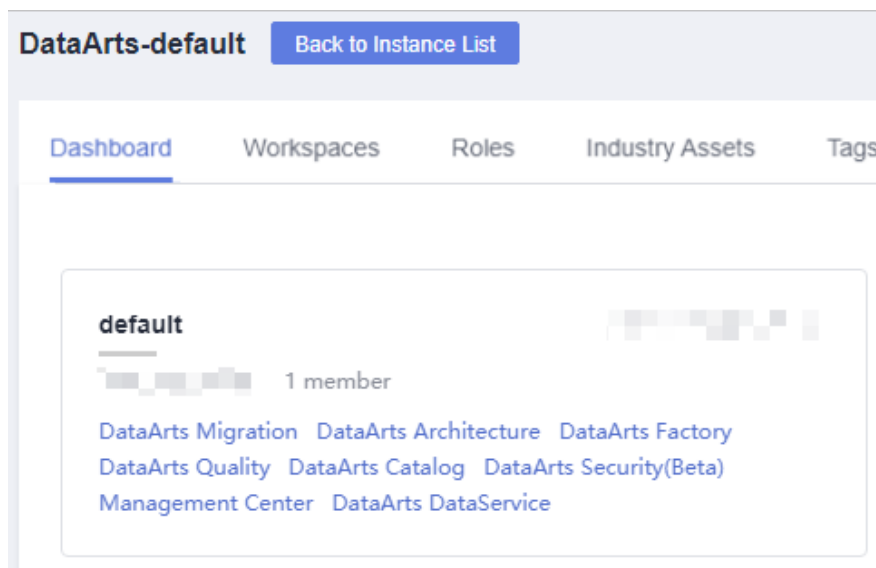
5.3 Preparations

After derivative metrics are published, you can run or schedule them in the O&M center.

Procedure

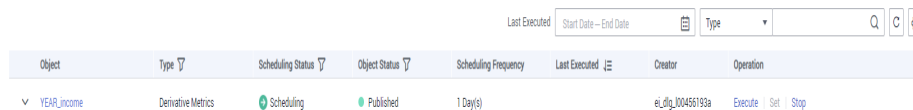
1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

Figure 5-4 DataArts Architecture



- Choose **O&M Center** from the left navigation bar. The **O&M Center** page is displayed as the figure below.


Figure 5-5 O&M Center page



- Manage your scheduling tasks as required. Refer to the following table for details.

Operation	Helpful Link
Edit	Refer to step 7 for details.
Run	Refer to step 8 for details.
Stop	Refer to step 9 for details.
View	Refer to step 10 for details.

- Edit a scheduling task.
 - To the right of the task you want to edit, click **Set**. The **Set Scheduling Task** dialog box is displayed.
 - Set parameters as prompt.
 - Click **OK**.
- Execute a scheduling task.
Click **Execute** to the right of the task you want to start.
- Stop a scheduling task.
To the right of the task you want to stop, click **Stop**.
- View a run log.

- a. In the object list, click  next to the object name to expand the object. Then, you can view the running instance of the object.
- b. Locate the target instance and click **View** in the Operation column. On the displayed page, you can view the run logs and results.

5.3.1 Adding Reviewers

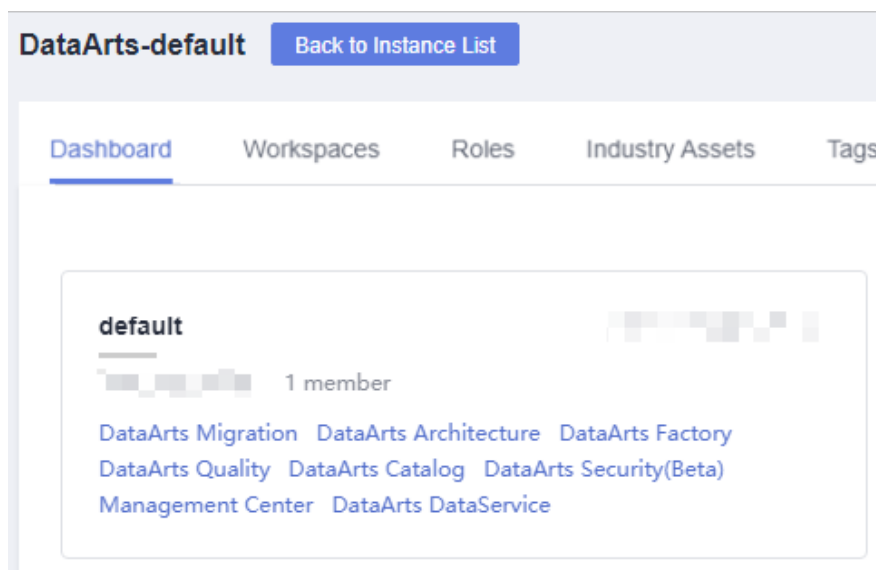
In the DataArts Architecture module, all business processes must be approved. Therefore, add reviewers first before conducting any operations. Only the workspace admin has the permissions required to add reviewers.

Adding a Reviewer

A reviewer must be a member who has the review permissions in the current workspace. You can edit and add workspace members in **Workspaces** on the DataArts Studio homepage.

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

Figure 5-6 DataArts Architecture

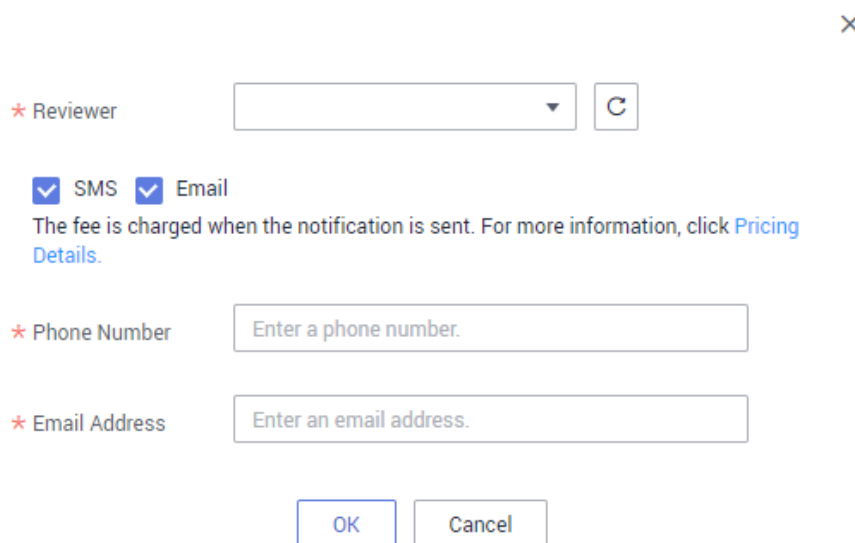



2. In the navigation pane, choose **Configuration Center**. On the displayed page, click **Reviewers**.
3. On the **Reviewer Management** tab page, click **Add**.
4. Select a reviewer, enter their mobile number and email address, and click **OK**.
The reviewer must be admins and developers of the current workspace, because only admins and developers have the review permissions of the workspace.

 **NOTE**

- You can only select reviewers from the given list. To enable a user to be available in the given list, add the user as a workspace member in **Workspaces** on the DataArts Studio homepage.
- If you select **SMS** or **Email** for **Notification Type**, DataArts Studio automatically creates a topic in SMN after the reviewer is added.
 - The topic name is in the following format: **DataArts_Subject_Reviewer_Project Name_Project ID-dlg_ds_Reviewer name**.

Figure 5-7 Adding a reviewer



* Reviewer 

SMS Email
The fee is charged when the notification is sent. For more information, click [Pricing Details](#).

* Phone Number

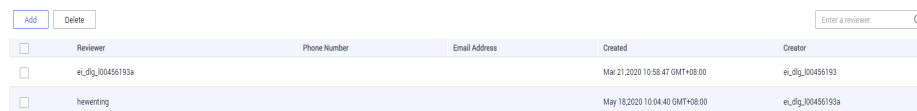
* Email Address

5. You can add multiple reviewers if needed.


Related Operations

On the DataArts Architecture page, choose **Configuration Center** in the left navigation pane. On the displayed page, click the **Reviewers** tab to manage reviewers.

Figure 5-8 Reviewer Management page



Reviewer	Phone Number	Email Address	Created	Creator
<input type="checkbox"/> ei_dlg_00456193a			Mar 21, 2020 10:58:47 GMT+08:00	ei_dlg_00456193
<input type="checkbox"/> hewenting			May 18, 2020 10:04:40 GMT+08:00	ei_dlg_00456193a

- **Searching for a reviewer**
In the upper right corner of the reviewer list, enter the name of the reviewer you are looking for and click .
- **Deleting a reviewer**
In the reviewer list, select the reviewer you want to delete, and click **Delete**.

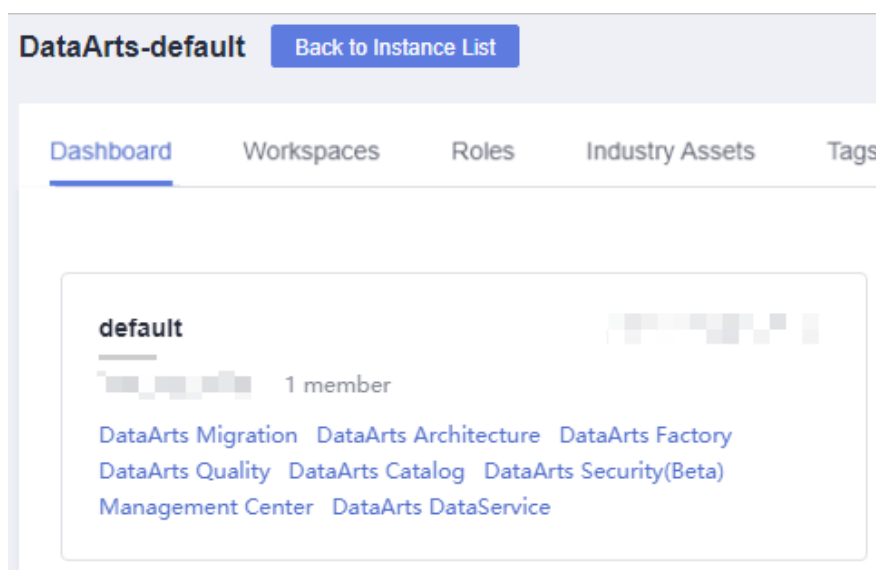
5.3.2 Configuration Center

Subjects

You can customize the subject levels and attributes in the theme design. By default, there are three levels in the system, which are named Subject Area Group (L1), Subject Area (L2), and Business Object (L3) from top to bottom. You can define a maximum of seven levels and a minimum of two levels. You can configure a maximum of 10 custom attributes.

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

Figure 5-9 DataArts Architecture



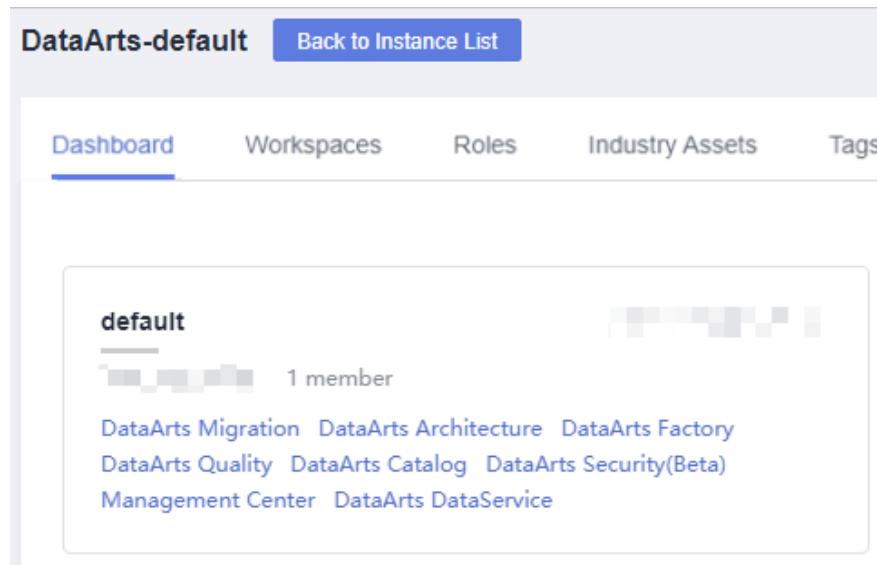
2. In the navigation pane, choose **Configuration Center**. On the displayed page, click the **Models** tab.
3. In the **Subject Level** area, you can add, delete, and edit subject levels.
 - Click **+** in the **Operation** column to add a custom subject level and click **Update**.
 - Click **🗑** in the **Operation** column to delete a subject level and click **Update**.
 - Except the business object at the last level, you can click the names of other levels to edit them.
4. In the **Customize Attribute Field** area, you can add, delete, and edit attributes.
 - Click **Create** to create a custom attribute.
 - Click **🗑** in the **Operation** column to delete a custom attribute.
 - Click the attribute name or the value in the **Mandatory** column to edit the attribute.

Standard Templates

You can customize the default options of data standards. When you access the **Standard Templates** page for the first time, the page for creating a data standard template is also displayed.

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

Figure 5-10 DataArts Architecture

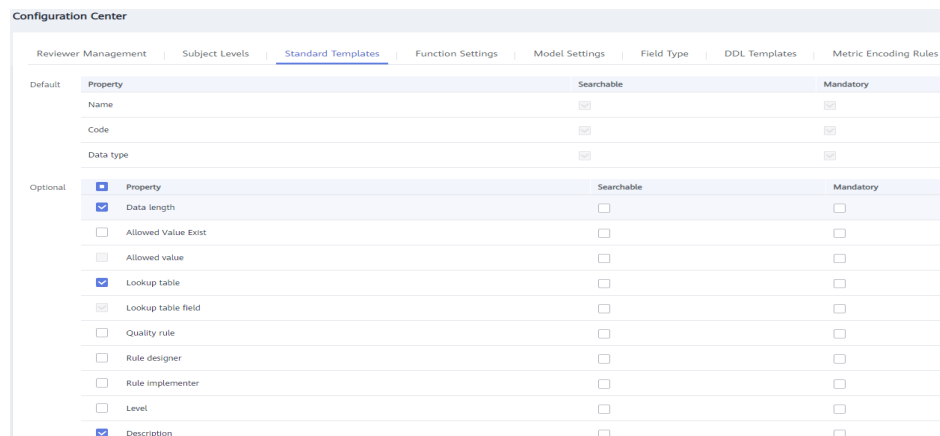


2. In the navigation pane, choose **Configuration Center**. On the displayed page, click the **Standard Templates** tab.
3. In the **Optional** area, select the parameters as required. Click **Create** next to **Custom** to add custom properties. After the configuration is complete, click **Update**.

NOTE

- **Searchable** and **Mandatory** specify whether a standard template can be searched for and whether a standard template is mandatory.
- After the template is saved, you must set values for the options selected in the template when creating a data standard.

Figure 5-11 Standard Templates tab page

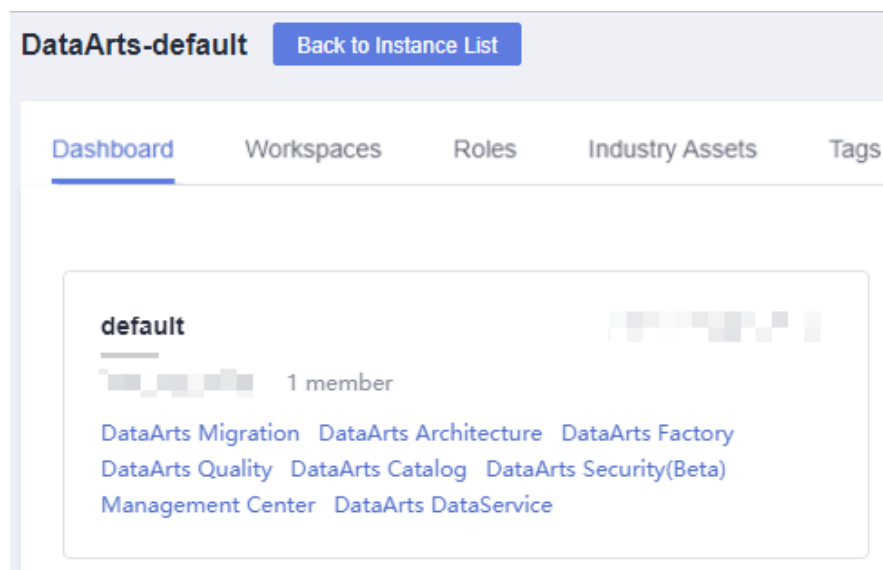


Functions

You customize functions for DataArts Architecture.

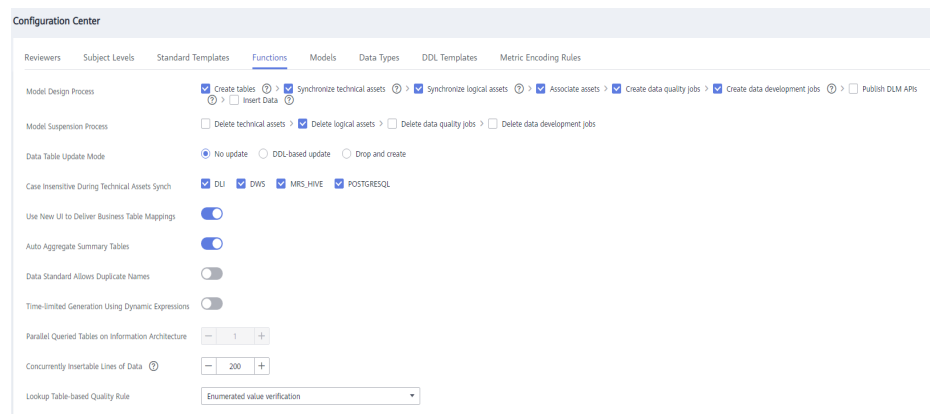
1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

Figure 5-12 DataArts Architecture



2. In the navigation pane, choose **Configuration Center**. On the displayed page, click the **Functions** tab.
3. On the page displayed, set the parameters and click **OK**. Click **Reset** to restore the default settings.

Figure 5-13 Functions



- **Model Design Process:** The selected processes are automatically executed progressively when a table created in an ER or dimension model is published and suspended. You are advised to select all the options.
 - **Create tables:** After a table publishing application is approved in DataArts Architecture, the system creates a physical table in the corresponding data source. When a table is deleted, the system deletes the corresponding physical table.
 - **Synchronize technical assets:** After a table in **ER Modeling** or **Dimensional Modeling** is published, the table is synchronized to the DataArts Catalog module as a technical asset, and the tag is synchronized to the corresponding technical asset.

NOTE

To enable **Synchronize Technical Assets**, you must create a data asset collection task for the database to which the table belongs in DataArts Catalog. Otherwise, the technical asset synchronization will fail.

- **Synchronize logical assets:** The system synchronizes logical models to DataArts Catalog as logical assets. After that, the system tags the logical assets accordingly.
- **Associate assets:** Associate logical assets with technical assets. After the logical assets and technical assets are synchronized, you can view the associated technical or logical asset when viewing the details of a logical or technical asset on the DataArts Catalog page. This function requires that the table information contains the data source information.
- **Create data quality jobs:** After a table in **ER Modeling** or **Dimensional Modeling** is published and approved, the system automatically creates a quality job in the DataArts Quality module of DataArts Studio for a table that is associated with a data standard (including the data length or allowed value) or associated with a quality rule.
- **Create data development jobs:** After a summary table is published, the system generates an E2E data development job.

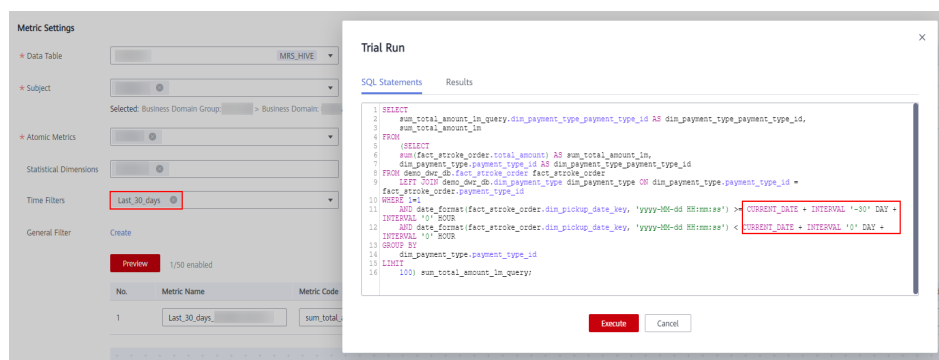
- **Publish DataArts DataService APIs:** After a summary table is published, a DataArts DataService API is automatically generated. This function takes effect only if DataArts DataService supports data connections of the summary table.
- **Insert data:** After a lookup table is published, values in the table are automatically written to the dimensional table.
- **Model Suspension Process:** Select whether to delete technical assets, logical assets, data quality jobs, and data development jobs when suspending the job.
- **Data Table Update Mode:** If a table in DataArts Architecture is modified after being published, you can choose whether to update the table in the database and how to update the table. By default, the table is not updated. However, you can set the update operation in the configuration center as required. To be more specific, configure the corresponding update statements in the DDL templates.
 - **No update:** The system does not update tables in a database.
 - **DDL-based update:** The system updates tables in the database based on the DDL update template configured in [DDL Templates](#). The underlying data warehouse engine determines whether the update is successful. Different types of data warehouses support different table update modes. If the data warehouse does not support table update operations on the DataArts Architecture page, the tables in the database may be inconsistent with those in DataArts Architecture. For example, table fields cannot be deleted when DLI tables are updated. If table fields are deleted from the tables in DataArts Architecture, the corresponding table fields cannot be deleted from the database.

If the offline database supports the syntax for updating the table architecture, you can configure the syntax in the DDL template. Then, the update operation can be performed. Otherwise, update the table by rebuilding it.
 - **Drop and create:** The system deletes an existing table in a database and then creates a table. This option ensures that the tables in the database are the same as those in DataArts Architecture. However, since the table is deleted first, you are advised to select this option only in the development and design phase or test phase. After the product is brought online, you are not advised to select this option.
- **Case Insensitive During Technical Assets Synchronizing:** When a table, whose type is the same as the data connection, is published, the data connection name is case insensitive during technology asset synchronization. If the name is the same as an existing one, the connection exists.
- **Use New UI to Deliver Business Table Mappings:** This function is enabled by default. The mapping function of the new version supports operations such as join. You are advised to use the mapping function of the new version.
- **Auto Aggregate Summary Tables:** When publishing a derivative or compound metric, the system automatically generates a summary table.

A statistical dimension corresponds to a summary table. You can click the **Automatic Aggregation** tab on the summary table page to view the automatically generated summary tables.

- **Data Standard Allows Duplicate Names:** This function is disabled by default. If it is enabled, duplicate data standard names are allowed.
- **Time-limited Generation Using Dynamic Expressions:** If you enable this function, dynamic time expressions will be used; otherwise, the default static time expressions will be used. The dynamic expression automatically updates the generated time, while the static expression does not. For example, if the current month is September and a static expression is used, data generated for the last 30 days is the data in August. Even when the current month changes to October, data generated for the last 30 days is still the data in August. However, if a dynamic expression is used, data generated for the last 30 days will automatically change to the data in September if the current month has changed to October. The following figure shows an example time function using a dynamic expression.

Figure 5-14 Dynamic expression



NOTE

If you enable this function for the first time, you need to reset the derivative metrics in the DLL template. If you have made any change to the DLL template, back up the template before resetting it. Resetting the template will overwrite any change that has been made. After the template is reset, you must make the changes again.

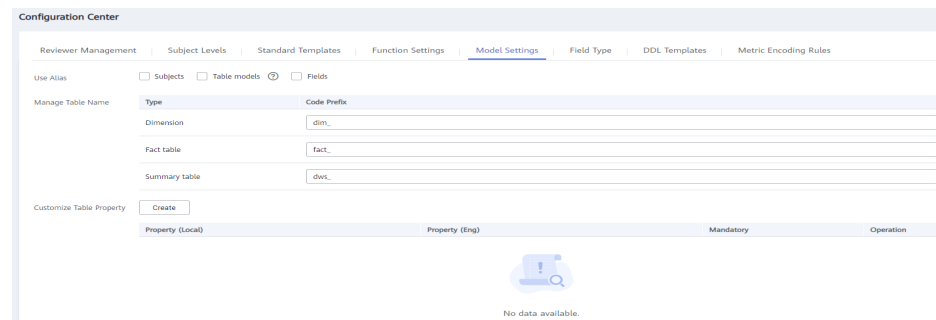
- **Parallel Queried Tables on Information Architecture:** The default value is 1. Currently, you cannot change the value.
- **Concurrently Insertable Lines of Data:** The value determines the number of lines of the dimensional table into which data of the lookup table is inserted. If the lookup table contains a large amount of data, the data may fail to be inserted into the dimensional table. In this case, you can reduce the value of this parameter.
- **Lookup Table-based Quality Rule:** Select a value from the drop-down list box. If the data volume of the lookup table is small, select **Enumerated value verification**; otherwise, select **Field value consistency verification**.

Model Settings

You can perform the following operations during subject design and model design on the **Model Settings** page.

- Add the subject alias, table model alias, and field alias.
- Set the default table code prefix for dimension tables, fact tables, and summary tables.
- Add custom fields to a table.
- Add custom fields to an attribute.

Figure 5-15 Model Settings tab page



In the navigation pane, choose **Configuration Center**. On the displayed page, click the **Models** tab.

- **Use Alias:** You can enable or disable alias.
 - The options are as follows:
 - If you select **Subjects**, you must enter an alias when creating or editing subject.
 - If you select **Table models**, you must enter an alias when creating or editing a table. Business tables, dimension tables, fact tables, and summary tables are affected when **Table models** is selected.
 - If you select **Fields**, you must enter an alias when creating or editing a table field.
- **Manage Table Names:** Set the default table code prefix for dimension tables, fact tables, and summary tables.
- **Customize Table Property:** When creating or editing a table, you can set custom fields in the basic settings of the table. Business tables, dimension tables, fact tables, and summary tables are affected.

Data Types

When you create a table, reverse a database, or convert a model, if the default data type or the data type mappings between different data sources cannot meet your requirements, you can add, delete, or modify data types. The default data type cannot be deleted.

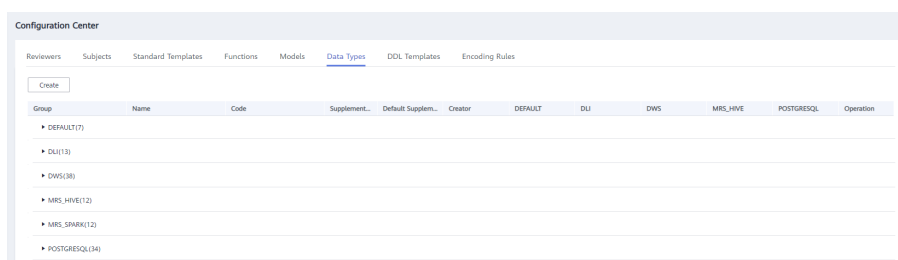
Step 1 In the navigation pane, choose **Configuration Center**. On the displayed page, click the **Data Types** tab.

Step 2 On the page displayed, you can view the data type and the data type mappings between different data sources. The type whose creator is **SYSTEM** is the default field type.

The types are described as follows:

- **DEFAULT** indicates the common data type which is used for creating a table when the data source type is not specified. For example, when you create a table of a logical model, the data type in the DEFAULT group is used.
- **DLI** indicates the data type of the table with the DLI data connection.
- **DWS** indicates the data type of the table with the DWS data connection.
- **MRS_HIVE** indicates the data type of the table with the MRS_HIVE data connection.
- **MRS_SPARK**: indicates the data type of the Hudi table with the MRS_SPARK connection.
- **POSTGRESQL**: indicates the data type of the table with the PostgreSQL connection.

Figure 5-16 Data Types tab page



Step 3 Manage field types.

- **Create**

To add a field type, click **Create**. In the dialog box displayed, set the parameters and click **OK**.

Figure 5-17 Creating a field type

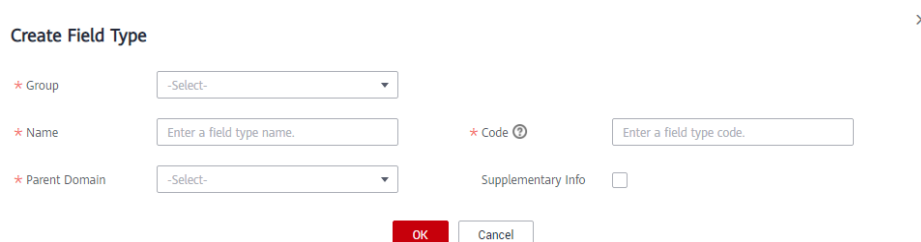



Table 5-1 Parameters for creating a field type

Parameter	Description
Group	Group that the new field type belongs to.
Name	Name of the field type to create. Field type names must start with letters. Only letters, numbers, brackets, spaces, and underscores (_) are allowed.


Parameter	Description
Code	Data type code, which must be supported by the data warehouse. Only upper-case letters, numbers, and underscore (_) are allowed.
Parent Domain	Select the domain that the new field type belongs to.
Data Types in Data Sources	Select the data type of the mapping connection of the new field type.
Supplementary Info	You can enable this function if you want to set the data length range for some data types. For example, you can enter (10,2) for the DECIMAL(p,s) data type, indicating that the total number of digits in the value is 10, and the number of digits after the decimal point is 2. You can also enter 10 for the VARCHAR data type, indicating that the maximum number of characters is 10.
DLI	Data type of the DLI data connection that the new field type is mapped to.
DWS	Data type of the DWS data connection that the new field type is mapped to.
MRS_HIVE	Data type of the MRS Hive data connection that the new field type is mapped to.
MRS_SPARK	Data type of the MRS Spark data connection that the new field type is mapped to.
POSTGRESQL	Data type of the PostgreSQL data connection that the new field type is mapped to.

- **Edit**

In the field type list, specify a field type and click  to edit the field type. For details on the parameters, see [Table 5-1](#).

- **Delete**

You can delete new field types. The field type whose creator is **SYSTEM** is the default field type and cannot be deleted.

In the field type list, specify a field type and click  to delete it. Then click **OK**.

- **Reset**

Click **Reset** at the bottom of the **Field Type** tab page to restore the default settings.

----End

DDL Templates

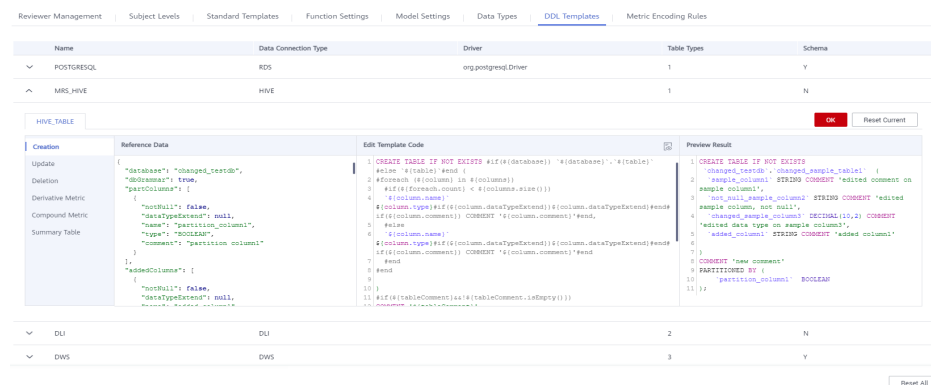
On the DataArts Architecture page, you can modify diversified types of tables (such as DWS, DLI, PostgreSQL, Hive, and Spark). If you need to generate DDL statements of other data sources for a created table of a certain type, you can modify the DDL template of the table based on the DDL syntax of the target data source.

1. In the navigation pane, choose **Configuration Center**. On the displayed page, click the **DDL Templates** tab.
2. On the page displayed, you can configure DDL templates for DLI views or diversified types of tables. You can modify the DDL templates by referring to the parameter description on this page. After the modification is complete, click **OK**. Click **Reset All** to restore the default settings.

As shown in **Figure 5-18**, the process is described as follows:

- **Creation** allows you to view or edit a new table or a DDL template of a DLI view.
- **Update** allows you to view or edit an updated table or a DDL template of a DLI view.
- **Deletion** allows you to view or edit a deleted table or a DDL template of a DLI view.
- **Derivative Metric** allows you to view or edit the SQL template of a derivative metric.
- **Compound Metric** allows you to view or edit the SQL template of a compound metric.
- **Summary Table** allows you to view or edit the SQL template of a summary table.
- The **Reference Data** area shows an example of table details. Variables in the example define table details.
- The **Edit Code Template** area allows you to edit DDL templates. If you need to generate DDL statements for other types of databases, you can modify the DDL template based on the DDL syntax of the target data source.
- The **Preview Result** area allows you can preview the DDL statements generated based on the edited template.

Figure 5-18 DDL Templates tab page



Encoding Rules

1. In the navigation pane, choose **Configuration Center**. On the displayed page, click the **Encoding Rules** tab.
2. Manage encoding rules.
 - Add an encoding rule.
Click **Add** above the encoding rule list. In the displayed dialog box, set required parameters, and click **OK**.

Figure 5-19 Adding an encoding rule

Table 5-2 Parameters for adding an encoding rule

Parameter	Description
Type	Encoding rule type. The following options are available: Business metric, Logical entity, Logical property, and Data standard.
Code Range	By default, the encoding rule takes effect globally. You can select subjects, processes, lookup tables, or data standards.
System Rule	Whether this rule is a system rule. The value is No and cannot be changed.

Parameter	Description
Encoding Rule	The value consists of a prefix and a digit code and cannot be changed.
Prefix	The value can contain characters and digits but cannot end with a digit. It cannot be changed.
Digital Code	You can select Sequential or Random .
Start Code	Start value of the digital code range
End code	End value of the digital code range
Code Example	The configured encoding rule is displayed.

- Deleting an Encoding Rule

Select an encoding rule and click **Delete** above the list. In the displayed dialog box, click **Yes**.

 **NOTE**

The four preset encoding rules cannot be deleted, including the logical property, data standard, logical entity, and business metric rules.

- Editing an Encoding Rule

Locate an encoding rule, click **Edit** in the **Operation** column, modify parameters, and click **OK**.

5.4 Data Survey

5.4.1 Designing Processes

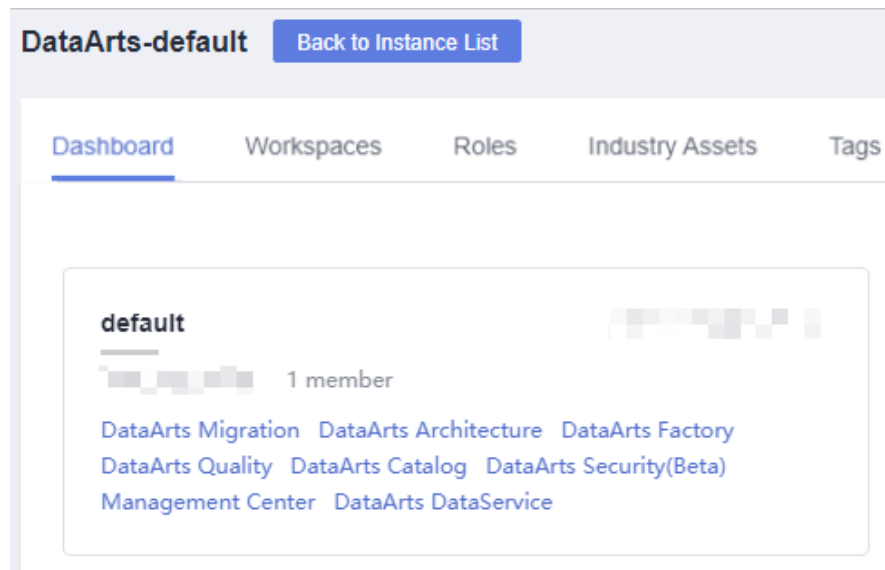
Business Process Architecture (BPA) is developed based on value streams, and is used to guide and standardize the management of BT&IT requirements and ensure the efficiency of business requirement handling, analysis, and delivery. BPA prioritizes high-value requirements, which maximizes the business value, assists in business operations, and facilitates goal achievement.

Creating a Process

Design the process from L1 to L 3 based on service requirements.

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

Figure 5-20 DataArts Architecture




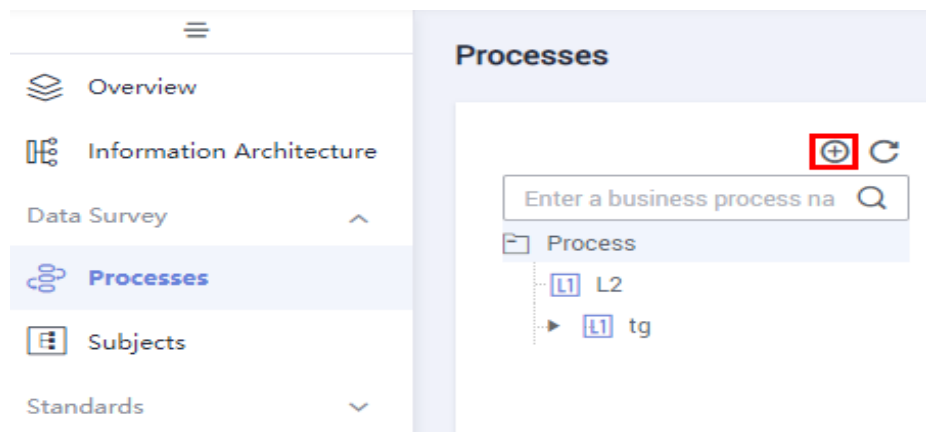
2. Choose **Data Survey** > **Processes** in the left navigation bar. Click  to create a process. When creating a process for the first time, perform the operation under the root node.

Figure 5-21 Process design



3. In the dialog box displayed, set the parameters and click **OK**.

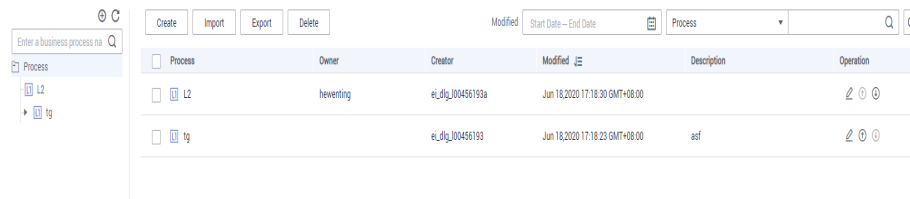
Figure 5-22 Creating a process

Table 5-3 Parameters for creating a process

Parameter	Description
Process	Process name. Only letters, numbers, and underscores (_) are allowed.
Owner	Process owner. You can enter the name of an owner or select an existing owner.
Parent Process	Parent process of the process
Description	A description of the process.

- Repeat the preceding steps in sequence to create more processes or subprocesses. Generally, you must design processes from L1 to L3. The first layer is identified as L1, the second layer as L2, and the third layer as L3. The following figure shows an example.

Figure 5-23 Process design example



Exporting a Process

You can export the processes that have been created in DataArts Architecture to files.

- Step 1** On the DataArts Architecture page, choose **Data Survey > Processes** in the left navigation pane.
- Step 2** Click **Export** above the process list. After a few seconds, a message is displayed in the upper right corner of the page, indicating that the process is exported. You can view the export process.

NOTE

A subject or process has a hierarchy. You can export only data of all levels.

----End

Importing a Process

- Step 1** On the DataArts Architecture page, choose **Data Survey > Processes** in the left navigation pane.
- Step 2** Click **Import** above the process list.
- Step 3** In the dialog box displayed, set **Update Existing Data**, click **Select File**, and click **Upload**.

Figure 5-24 Importing a process

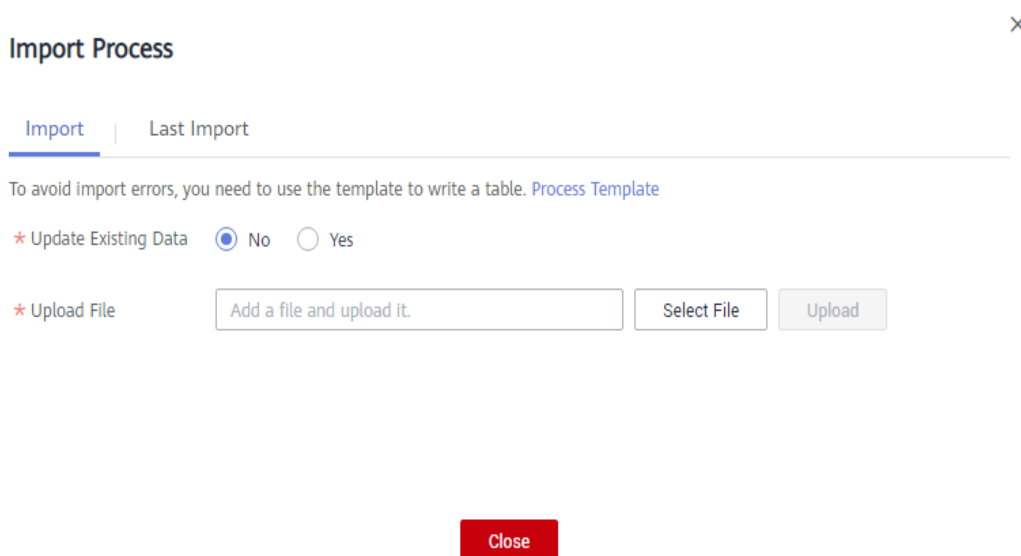


Table 5-4 Parameters for importing a process

Parameter	Description
Update Existing Data	<p>Whether to update the existing processes of DataArts Architecture. The options are as follows:</p> <ul style="list-style-type: none">• No: If you select this option, the existing process will not be updated.• Yes: If you select this option, the existing process will be updated. <p>During the import, only process creation and update are allowed.</p>
Upload File	<p>Select the file to import.</p> <p>You can use either of the following methods to obtain the file to import:</p> <ul style="list-style-type: none">• Downloading the process template and fill in it On the Import tab page, click Process Template to download the template, set related parameters in the template based on service requirements, save the settings, and upload the file. See Table 5-5 for template parameter details.• Exporting a process You can export the processes created in DataArts Architecture of a DataArts Studio instance to an Excel file. Then, import the Excel file. For details, see Exporting a Process.

[Table 5-5](#) describes the parameters in the downloaded template. Parameters whose names start with an asterisk (*) are mandatory, and parameters whose names do not start with an asterisk (*) are optional. One record is required for one process.

Table 5-5 Parameters in the process import template

Parameter	Description
Process	<p>If it is a level 1 process, this field can be left blank.</p> <p>If it is not, this field is mandatory. If there are multiple processes, separate them with slashes (/), for example, Integrated Product Development/Development Lifecycle.</p>
*Name	Process name.
*Owner	Process owner. You can enter the name of an owner or select an existing owner.
Description	A description of the process.

Step 4 The import result is displayed on the **Last Import** tab page in the **Import Process** dialog box. If the import is successful, click **Close**. If the import fails, you can view the failure cause, correct the template file, and upload it again.

----End

Deleting a Process

You can delete the processes that are no longer used. Deleted processes cannot be recovered. Exercise caution when performing this operation. If a process has subdirectories or subprocesses, you must delete the subdirectories or subprocesses first.

Step 1 On the DataArts Architecture page, choose **Data Survey > Processes** in the left navigation pane.

Step 2 In the process list, select the target process and click **Delete** above the process list.

Step 3 In the **Delete Process** dialog box displayed, confirm the process information and click **Yes**.

----End

5.4.2 Designing Subjects

A subject is a hierarchical architecture that classifies and defines data to help clarify data assets and specify relationships between subject areas and business objects.

You can design subjects in either of the following ways:

- **Creating a Subject**

Manually create a subject.

- **Importing a Subject**

If the subject information is complex, you are advised to import subjects in batches.

- You can download the provided subject design template, fill in the content, and upload the file to import the subjects in batches.
- You can export the subjects created in DataArts Architecture of a DataArts Studio instance to an Excel file. Then, import the Excel file. For details on how to export subjects, see [Exporting a Subject](#).

After creating a subject, you can search for, edit, or delete it. For details, see [Managing a Subject](#).

Subject Design Overview

By default, the system provides three subject levels: Subject Area Group (L1), Subject Area (L2), and Business Object (L3).

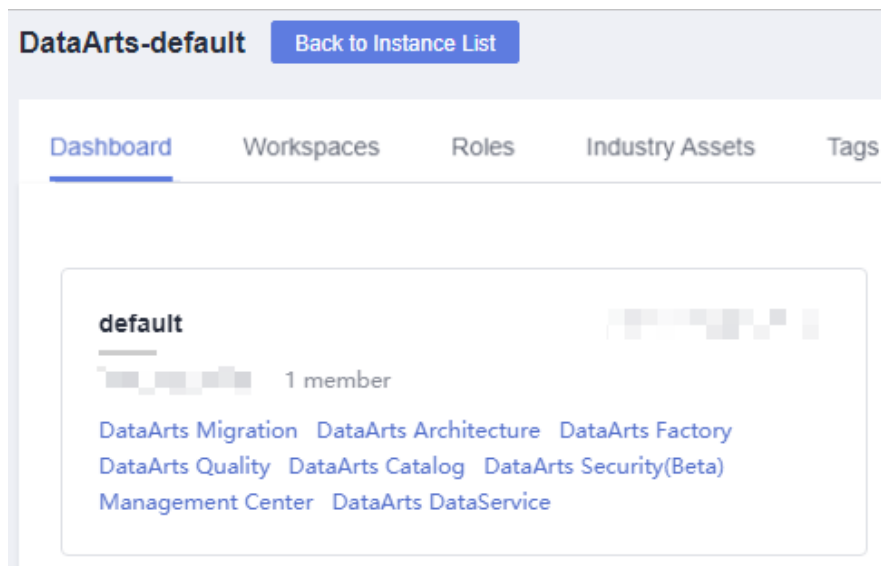
- **Subject Area Group:** used to group business domains based on scenarios
- **Subject Area:** A data domain is a dataset, in which data is of the same property.
- **Business Object** includes important information about people, events, and things that are indispensable to enterprise operations and management.

You can also customize the subject levels by referring to [Subjects](#).

Creating a Subject

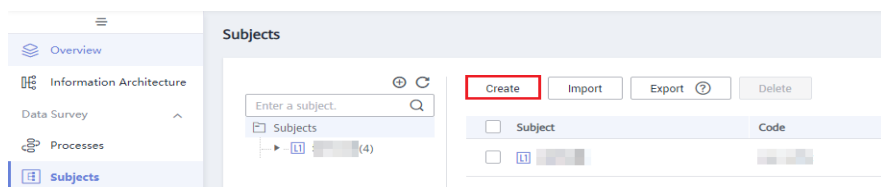
1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

Figure 5-25 DataArts Architecture



2. On the **DataArts Architecture** page, choose **Data Survey** > **Subjects** in the left navigation bar. On the page displayed, click **Create** in the upper left corner.

Figure 5-26 Designing a subject



3. In the dialog box displayed, set the parameters and click **OK**.

Table 5-6 Parameters for creating a subject area group

Parameter	Description
* Subject Name	The following characters are not allowed: / \ < >.
* Subject Code	The code of the subject area group to create. Only letters, digits, spaces, underscores (_), hyphens (-), parentheses, and ampersands (&) are allowed.

Parameter	Description
Alias	The following characters are not allowed: \ < >. NOTE Before configuring an alias, choose Metrics > Configuration Center , click the Model Settings tab, and select Subjects for Use Alias .
* Parent Subject	Parent subject of the subject area group
Data Owner's Department	The department that the data owner belongs to.
* Data Owner	Select a data owner from the drop-down list box. You can select multiple data owners or enter custom data owners.
Description	A description of the subject area group to create.

Figure 5-27 Create Subject Area Group dialog box

4. You can create multiple subjects in a subject.

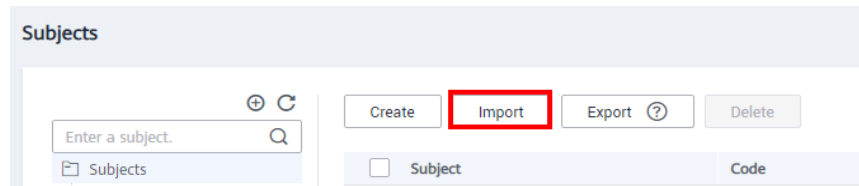
The number of subject levels is defined by users on the **Subject Levels** tab page on the **Configuration Center** page. By default, there are three levels in the system, Subject Area Group (L1), Subject Area (L2), and Business Object (L3).

Importing a Subject

- Step 1** On the DataArts Architecture page, choose **Data Survey > Subjects** in the left navigation pane.

Step 2 Click **Import** in the upper left corner.

Figure 5-28 Importing a subject



Step 3 In the dialog box displayed, set **Update Existing Data**, click **Select File**, and click **Upload**.

Figure 5-29 Importing a subject

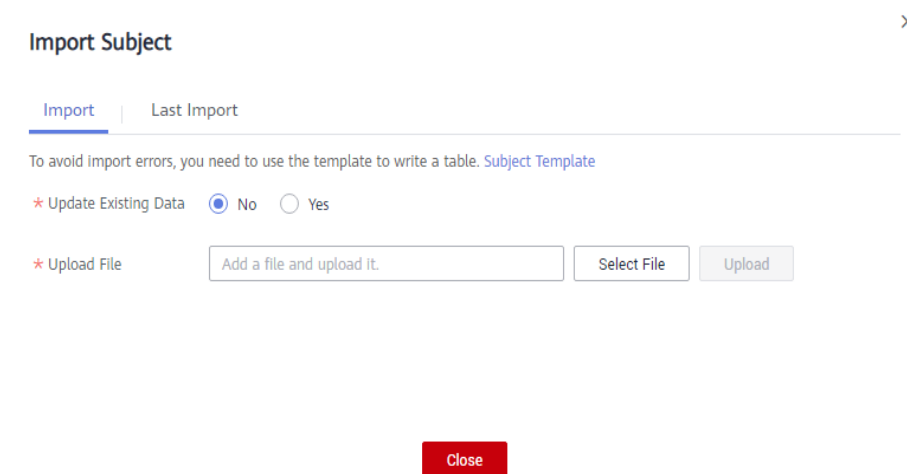


Table 5-7 Parameters for importing subjects

Parameter	Description
Update Existing Data	<p>Whether to update existing subject information (subject area group, subject area, or business object) during the import. When a subject is imported, the system checks whether the subject exists according to its code.</p> <ul style="list-style-type: none"> ● No: If you select this option, the subject information will not be updated. ● Yes: If you select this option, the subject information will be updated. <p>During the import, only subject creation and update are allowed.</p>

Parameter	Description
Upload File	<p>Select the file to import.</p> <p>You can use either of the following methods to obtain the file to import:</p> <ul style="list-style-type: none"> • Downloading the subject import template and fill in it In the Import Subject dialog box, click Subject Template to download the template, fill in the content, and save the settings. See Table 5-8 for template parameter details. • Exporting subjects to files You can export the subjects created in DataArts Architecture of a DataArts Studio instance to an Excel file. Then, import the Excel file. See Exporting a Subject for details.

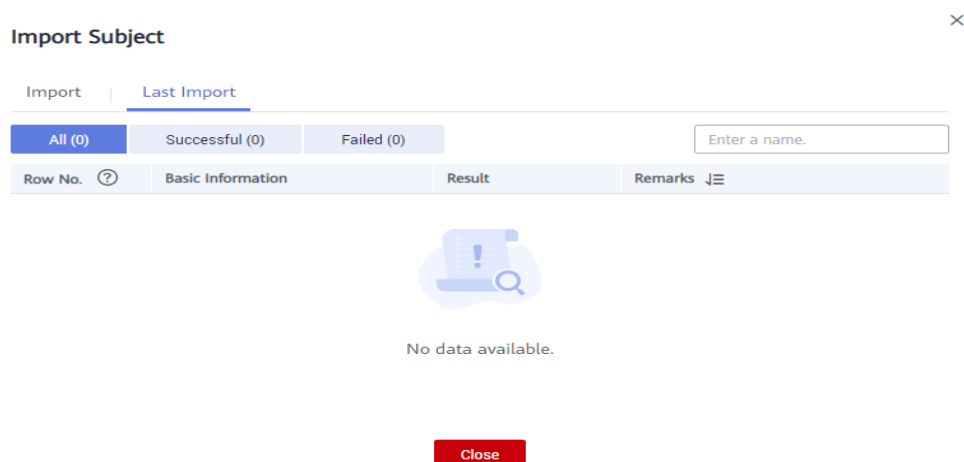
[Table 5-8](#) describes the parameters in the downloaded template. Parameters whose names start with an asterisk (*) are mandatory, and other parameters are optional. Enter the information about a subject in a line.

Table 5-8 Parameters

Parameter	Description
Parent Subject	Encoding path of the upper-level subject, which is separated by slashes (/).
*Name	The following characters are not allowed: / \ < >.
*Code	Code of the subject to create. Only letters, digits, spaces, underscores (_), hyphens (-), parentheses, and ampersands (&) are allowed.
Alias	Alias of the subject.
Description	A description of the subject. This parameter is mandatory for the lowest-level subject. You must add the description of the lowest-level subject in the file to be imported.
Data Owner's Department	The department that the data owner belongs to. This parameter is mandatory for the lowest-level subject. You must add the department of the owner of the lowest-level subject in the file to be imported.
Data Owner	The owner of the data. Multiple owners are supported. Separate owner names with commas (,)

Step 4 View the import result on the **Last Import** tab page. If the import is successful, click **Close**. If the import fails, you can view the failure cause, correct the template file, and upload it again.

Figure 5-30 Last Import tab page



----End

Exporting a Subject

- Step 1** On the DataArts Architecture page, choose **Data Survey** > **Subjects** in the left navigation pane.
- Step 2** Click **Export** in the upper left corner to export the existing subject information to an Excel file. Then, import the Excel file.

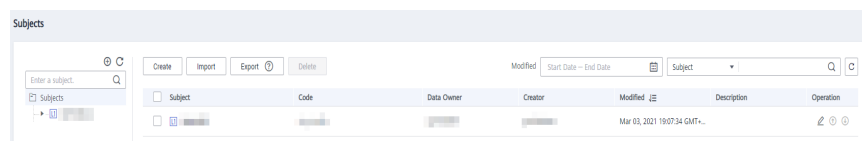
NOTE


A subject or process has a hierarchy. You can export only data of all levels.



----End

Managing a Subject

Figure 5-31 Subject design area



- Search
You can enter a keyword in the search box to search for a topic.
- Edit
Locate a subject in the list and click  in the **Operation** column to edit the subject.
- Delete
Select a subject in the list and click **Delete** above the list.
- Move Up/Down

Locate a subject in the list and click  or  in the **Operation** column to move down or up the subject.

5.5 Standards Design

5.5.1 Creating Lookup Tables

A lookup table is also called a data dictionary table. It consists of enumerable data names and codes and stores the relationships between them. A lookup table provides the following functions:

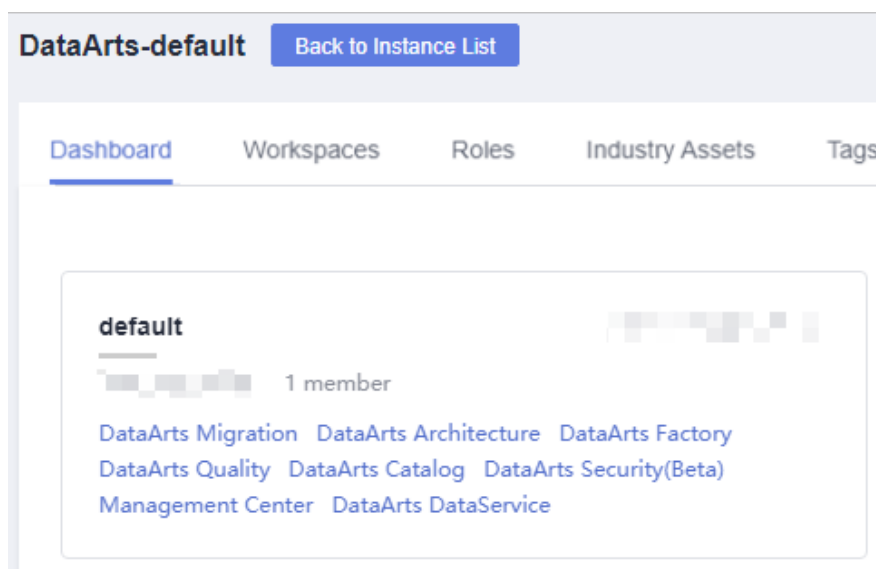
- Standardizes business data and supplements mapping fields during data cleansing.
- Monitors the value range of business data during data quality monitoring.
- Enumerates dimensions during dimensional modeling.

Creating and Publishing a Lookup Table

Manually create a lookup table. You can also add table records after creating a lookup table. For details, see [Filling in a Lookup Table](#).

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

Figure 5-32 DataArts Architecture




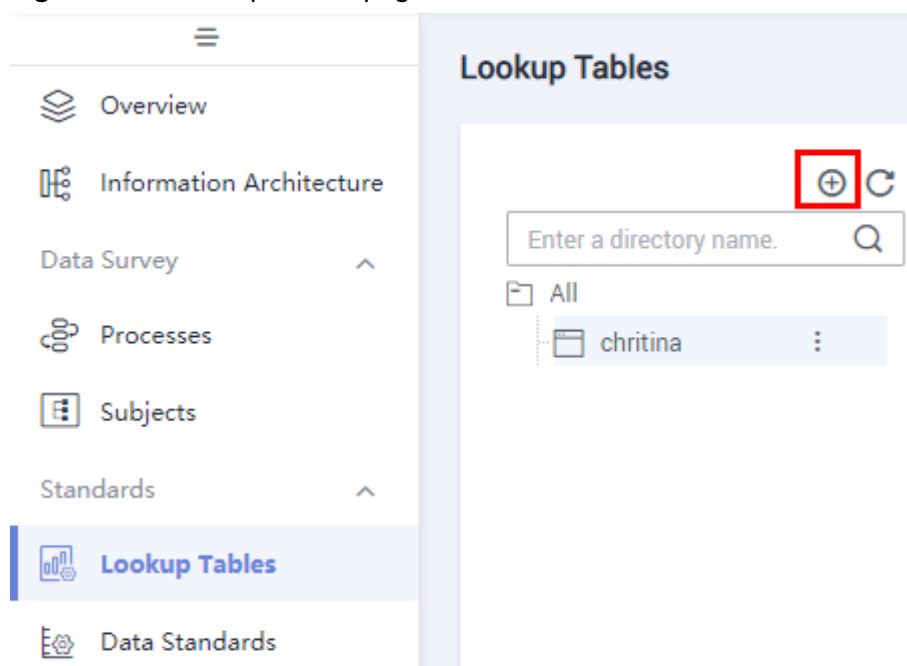
2. On the DataArts Architecture page, choose **Standards > Lookup Tables** in the left navigation pane.
3. Select a directory from the directory tree on the **Lookup Tables** page, and then click  to create a directory under the selected directory. When creating a directory for the first time, you can create a directory under the root directory.

Figure 5-33 Lookup Tables page



4. In the dialog box displayed, set the parameters and click **OK**.

Figure 5-34 Create Directory dialog box

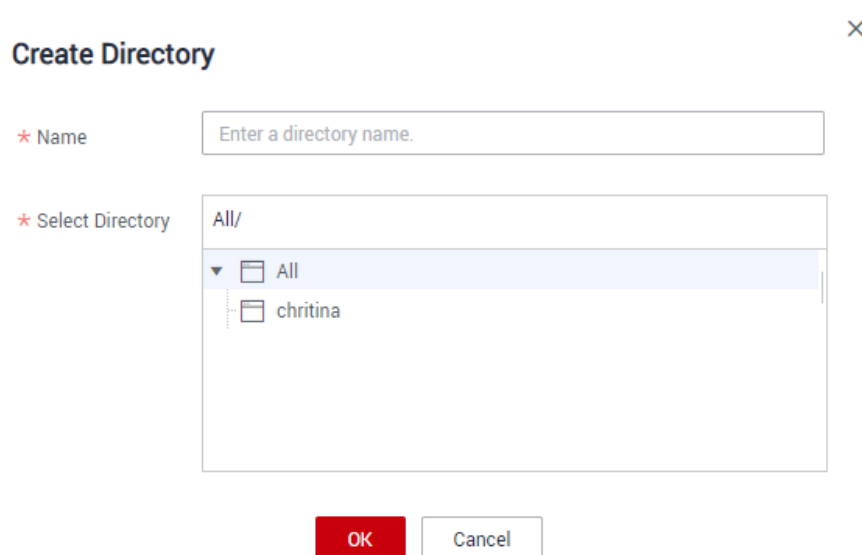


Table 5-9 Directory parameters

Parameter	Description
Name	Only letters, numbers, and underscores (_) are allowed.
Select Directory	Select an existing directory, and create a subdirectory under it.

5. Select the directory you created in the directory tree and click **Add** to create a lookup table.
6. On the **Create Lookup Table** page displayed, configure the parameters. In the **Table Details** area, set the parameters.

Figure 5-35 Table Details area

Basic Settings

Home Directory transport

* Table Name

* Table Code

Description
0/600

Table 5-10 Parameters

Parameter	Description
Table Name	The name of the lookup table to create. It must start with letters. Only letters, digits, and the following special characters are allowed: ()-_
Table Code	The code of the lookup table to create. It must start with letters. Only letters, digits, and underscores (_) are allowed.
Description	A description of the lookup table. Up to 600 characters are supported.


In the **Field Inputs** area, click **Add** or **+** to add new fields, and click  to delete unnecessary fields.

Figure 5-36 Field Inputs area

Field Inputs

2/100 configured

No.	Name	Code	Data Type	Comment	Operation
1	<input type="text" value="ID"/>	<input type="text" value="code"/>	STRING	<input type="text"/>	<input type="button" value="+"/> <input type="button" value="trash"/> <input type="button" value="refresh"/> <input type="button" value="help"/>
2	<input type="text" value="value"/>	<input type="text" value="value"/>	STRING	<input type="text"/>	<input type="button" value="+"/> <input type="button" value="trash"/> <input type="button" value="refresh"/> <input type="button" value="help"/>

7. Click **Publish**. In the **Apply for Publication** dialog box displayed, select a reviewer and click **OK**. After the application is approved, the **Lookup Tables** page is displayed. You can view the created lookup table in the list, and the status of the table is **Published**. Only published lookup tables can be used.

 **NOTE**

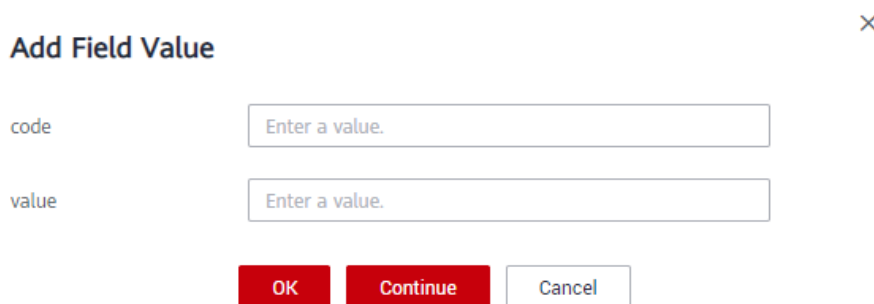
If you have been added as a reviewer, you can select **Auto-review** and click **OK**. After the application is approved, the lookup table status changes to **Published**.

Filling in a Lookup Table

Input values in the created lookup tables.

- Step 1** On the DataArts Architecture page, choose **Standards > Lookup Tables** in the left navigation pane.
- Step 2** In the list of lookup tables, find the target table and choose **More > Manage Value** in the **Operation** column.
- Step 3** On the page displayed, click **Add**. In the dialog box displayed, set the parameters.

Figure 5-37 Inputting a value



The screenshot shows a dialog box titled "Add Field Value" with a close button (X) in the top right corner. It contains two input fields: "code" and "value", both with placeholder text "Enter a value.". Below the fields are three buttons: "OK" (red), "Continue" (red), and "Cancel" (white).

- Step 4** Click **OK**. You can also click **Continue** to add more records.

----End

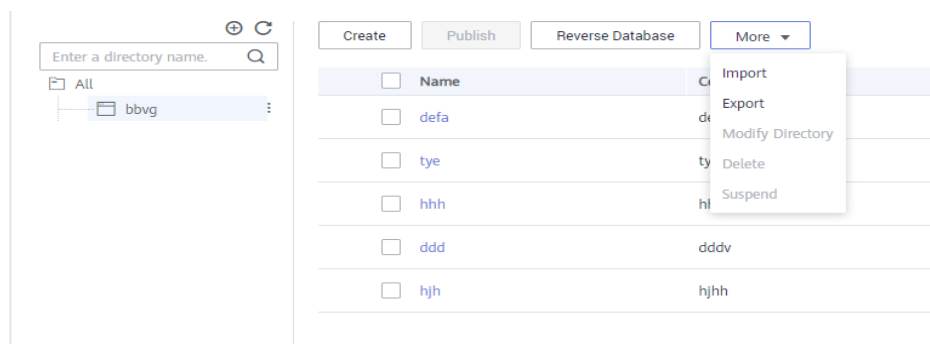
Importing a Lookup

When importing a lookup table, ensure that the table name contains a maximum of 32 characters.

You can import a new lookup table or import lookup table records in batches to an existing lookup table. If you have a large number of lookup table records, you are advised to import them in batches.

- Step 1** On the DataArts Architecture page, choose **Standards > Lookup Tables** in the left navigation pane.
- Step 2** On the page displayed, select a directory, and choose **More > Import**. You can also right-click the selected directory and choose **Import**.

Figure 5-38 Lookup Tables page



Step 3 In the **Import Lookup Table** dialog box displayed, set the parameters, and click **Upload**.

Figure 5-39 Import Lookup Table dialog box

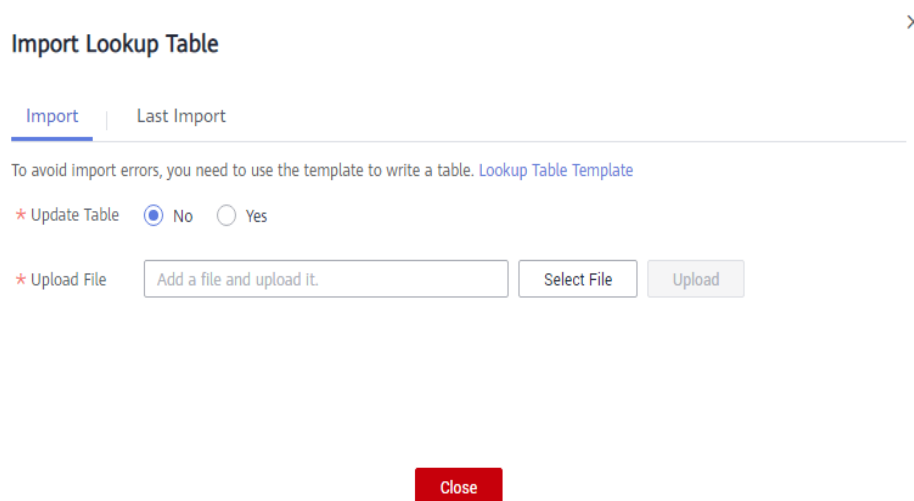


Table 5-11 Parameters for importing a lookup table

Parameter	Description
Update Table	<p>Whether to update the existing lookup table. When a lookup table is imported, the system checks whether the lookup table exists according to its code. The options are as follows:</p> <ul style="list-style-type: none"> • No: If you select this option, the existing lookup table will not be updated. • Yes: If you select this option, the existing lookup table will be updated. If a lookup table is in the Published state, you must publish the lookup table again after updating it so that the updated lookup table can take effect. <p>The import can create a lookup table or update an existing lookup table. It will not delete a lookup table.</p>

Parameter	Description
Upload File	<p>Select the file to import. You can use either of the following methods to obtain the file to import:</p> <ul style="list-style-type: none"> Downloading the lookup table template and fill in it In the Import Lookup Table dialog box, click Lookup Table Template to download the template, fill in the content, and save the settings. See Table 5-12 for template parameter details. Instructions for filling in the lookup table template: <ul style="list-style-type: none"> Parameters whose names start with an asterisk (*) are mandatory, and other parameters are optional. Multiple fields can be added to a lookup table. To import multiple lookup tables, you can add multiple sheets to the template file. The sheet name is the corresponding lookup table name. If the name of a lookup table already exists and Update Table is set to Yes, the existing lookup table will be updated during the import. If the table name does not exist, a lookup table with that name is created during the import. Exporting lookup tables to files You can export the lookup tables created in DataArts Architecture of a DataArts Studio instance to an Excel file. Then, import the Excel file. For details on how to export lookup tables, see Managing a Lookup Table.

Table 5-12 Parameters

Parameter	Description
Directory	The directory that a lookup table belongs to. Multi-level directories are separated with slashes (/), for example, dir01/dir02 .
*Table Name	The name of the lookup table to create. It must start with letters. Only letters, digits, and the following special characters are allowed: ()-_
*Table Code	The code of the lookup table to create. Only letters, numbers, and underscores (_) are allowed. A table code must start with a letter.
Table Description	A description of the lookup table. Up to 600 characters are supported.
*Field Name	The name of a field. Field names must start with letters. Only letters, numbers, spaces, and the following special characters are allowed: ()-_

Parameter	Description
*Field Code	The code of a field. Only letters, numbers, and underscores (_) are allowed. A field code must start with a letter.
*Field Data Type	The possible values are STRING , BIGINT , DOUBLE , TIMESTAMP , DATE , BOOLEAN , and DECIMAL .
Field Description	The supplementary information about a field. Up to 600 characters are supported.
Generate Standard	<ul style="list-style-type: none"> • true indicates to generate a data standard. • false indicates not to generate a data standard. The default value is false. <p>Note: To enable automatic generation of the data standard, choose Configuration Center in the navigation pane, click the Standard Templates tab, and select Lookup table.</p>

If the lookup table records need to be imported, create a sheet named after the lookup table in the template and add table fields to the sheet. Each field occupies a column. The column name includes the code and value. Enter the lookup table values to be imported. If the template contains a sheet named after the lookup table, you do not need to create the sheet. You can directly enter the table values to be imported in the sheet.

- Step 4** View the import result on the **Last Import** tab page. If the import is successful, click **Close**. If the import fails, you can view the failure cause, correct the template file, and upload it again.

----End

Importing a Lookup Table Through a Reverse Database

With reverse databases, you can import one or more created database tables from other data sources into a lookup table directory to turn them into lookup tables.

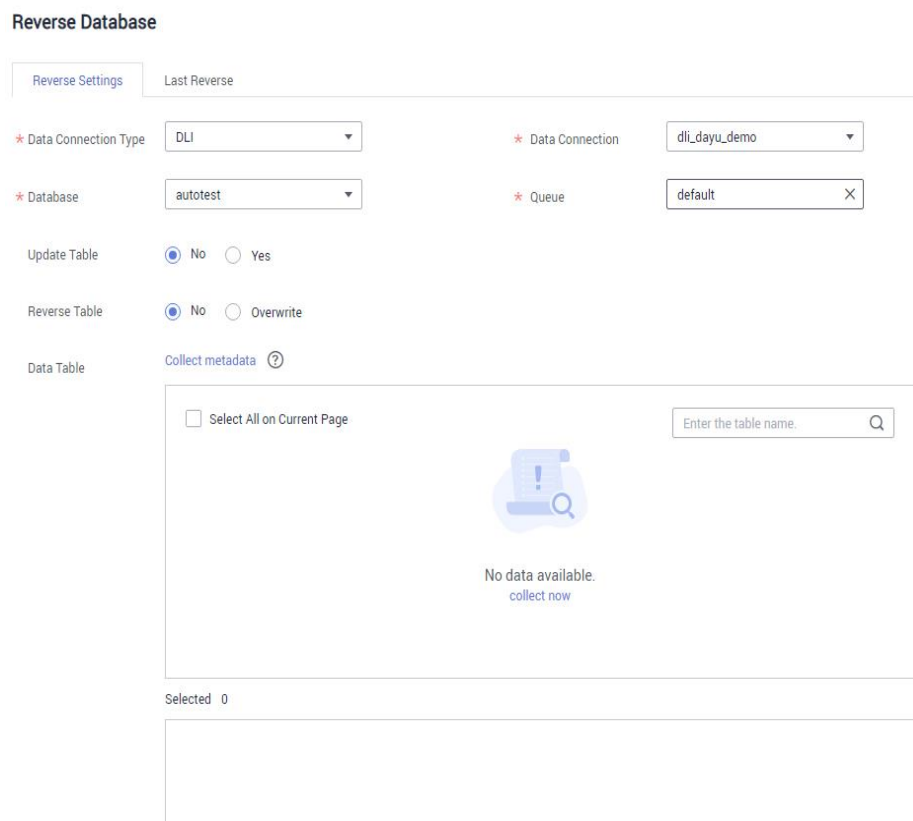
- Step 1** On the DataArts Architecture page, choose **Standards > Lookup Tables** in the left navigation pane.
- Step 2** On the page displayed, select a directory and click **Reverse Database** above the lookup table list.
- Step 3** In the dialog box displayed, set the parameters and click **OK**.

Table 5-13 Parameters for reversing a database

Parameter	Description
Data Connection Type	The data connection types supported by the reverse database are displayed in the drop-down list box. Select the required data connection type.

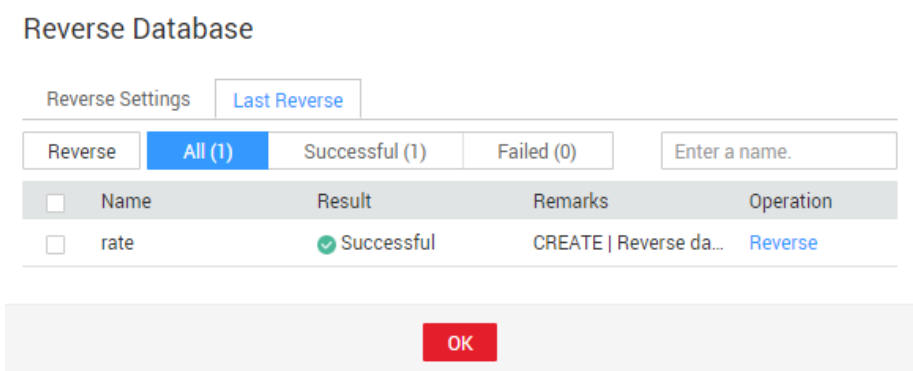
Parameter	Description
Data Connection	Select a data connection. If you want to reverse a database from other data sources to a lookup table directory, you must create a data connection in Management Center to connect to the data source. For details on how to create data connections, see Creating Data Connections .
Database	The name of the database. Select a database from the drop-down list box.
Schema	Select a value from the drop-down list box. This parameter is displayed only for DWS tables.
Queue	DLI queue. This parameter is available only when Data Connection Type is set to DLI .
Update Table	When Yes is selected, if the name of the reversed table is the same as that of an existing table in the lookup table list, the existing table is updated.
Reverse Table	<ul style="list-style-type: none">• No: If you select this option, tables are imported to the lookup table directory but table data is not imported during database reverse. After reversing a database, you can add records to the lookup table. Refer to Filling in a Lookup Table for details.• Overwrite: If you select this option, tables are imported to the lookup table directory and table data is imported as well during database reverse.
Data Table	You can select one or more data tables to import.

Figure 5-40 Reverse Database dialog box



Step 4 You can view the result on the **Last Reverse** tab page. If the reverse operation is successful, click **Close**. If the reverse operation fails, you can view the failure cause. After the fault is rectified, select the table again and click **Reverse** to retry.

Figure 5-41 Last Reverse tab page



----End

Exporting a Lookup Table

When exporting a lookup table, ensure that the table name contains a maximum of 32 characters.

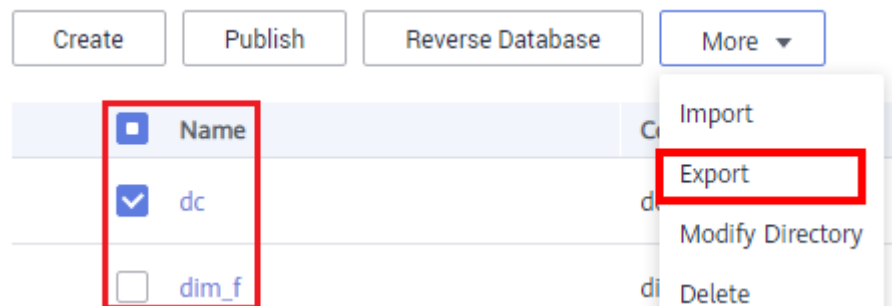
Step 1 On the DataArts Architecture page, choose **Standards > Lookup Tables** in the left navigation pane.

Step 2 Export a lookup table.

- **Export a single lookup table.**

In the lookup table list, select the target lookup table and choose **More > Export**.

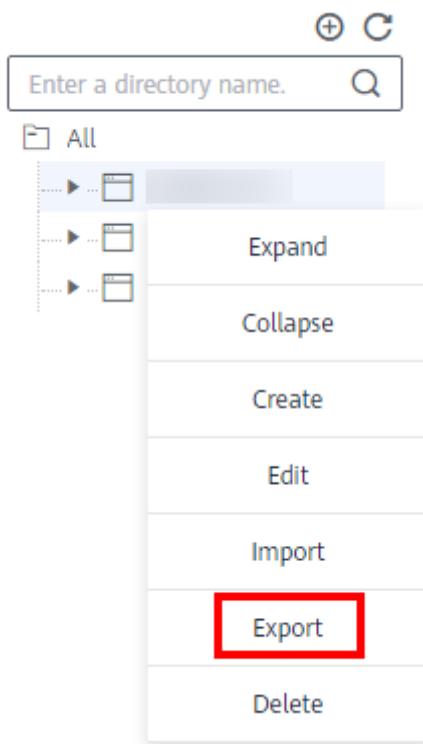
Figure 5-42 Lookup table list



- **Export all tables in the list.**

Right-click a directory in the directory tree and choose **Export**.

Figure 5-43 Directories storing exported lookup tables



----End

Deleting a Lookup Table

Deleted lookup tables cannot be recovered. Exercise caution when performing this operation. A lookup table that is to be published, has already been published, or to be suspended cannot be deleted.

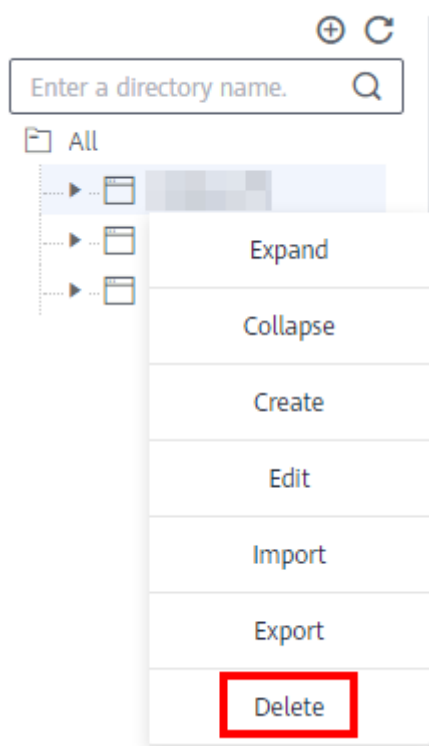
- Step 1** On the DataArts Architecture page, choose **Standards > Lookup Tables** in the left navigation pane.
 - Step 2** In the lookup table list, select the target lookup table and choose **More > Delete** above the list.
 - Step 3** In the dialog box displayed, click **Yes**.
- End

Deleting a Lookup Table Directory

A directory or its subdirectories that contain a lookup table cannot be deleted. You must delete the lookup table before deleting the directory.

- Step 1** On the DataArts Architecture page, choose **Standards > Lookup Tables** in the left navigation pane.
- Step 2** Right-click a directory in the directory tree and choose **Delete**.

Figure 5-44 Managing lookup table directories

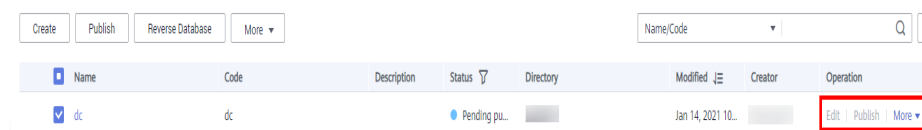


- Step 3** In the dialog box displayed, click **Yes**.
- End

Managing a Lookup Table

After a lookup table is created, you can search for, edit, or delete it.

On the DataArts Architecture page, choose **Standards > Lookup Tables** in the left navigation pane. You can manage the lookup tables as required.

Figure 5-45 Managing lookup tables

Name	Code	Description	Status	Directory	Modified	Creator	Operation
dc	dc		Pending pu...		Jan 14, 2021 10...		Edit Publish More

- **Edit**
In the lookup table list, select a table you want to edit and click **Edit** in the **Operation** column.
- **Publish**
In the lookup table list, click **Publish** in a row containing a table in the **Draft** or **Rejected** state, select a reviewer in the dialog box displayed, and click **OK**. After the application is approved, the lookup table is published.
- **Suspend**
In the lookup table list, locate a published lookup table you want to suspend, click **More** in the **Operation** column, and select **Suspend** from the drop-down list. In the displayed dialog box, select a reviewer and click **OK**. After the application is approved, the lookup table is suspended.
- **Manage Value**
In the lookup table list, locate a lookup table, click **More** in the **Operation** column, and select **Manage Value** from the drop-down list. Then you can edit the value of each field.
- **View History**
In the lookup table list, locate a lookup table, click **More** in the **Operation** column, and select **View History** from the drop-down list. Then you can view the publish history and changes of the lookup table, and compare different versions of it.

5.5.2 Creating Data Standards

Data standards describe data meanings and business rules that are stipulated and commonly recognized by enterprises and that those enterprises must comply with.

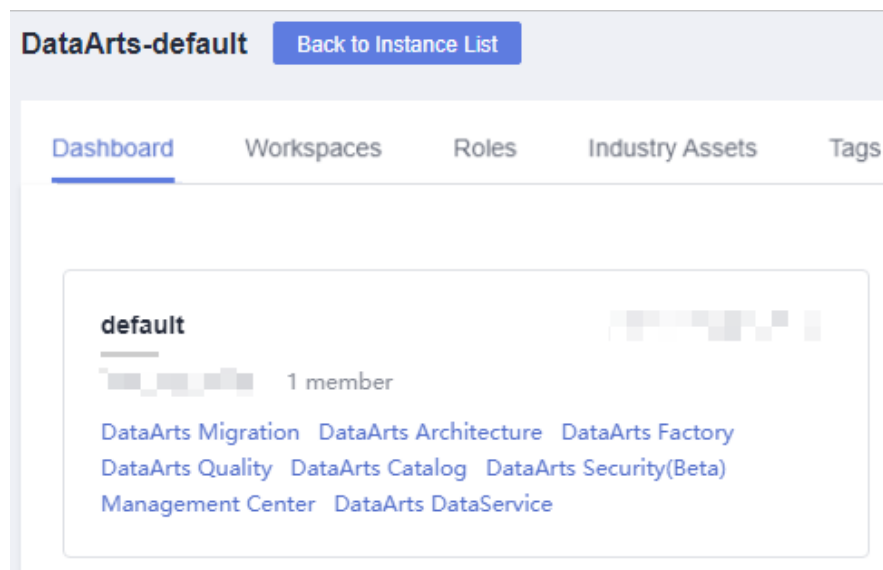
A data standard, also called a data element, is the smallest unit of data used. It cannot be further divided. A data standard is a data unit whose definition, identifiers, representations, and allowed values are specified by a group of properties. You can associate data standards with databases of a wide range of businesses. The identifier, data type, expression format, and value range are the basis of data exchange. They are used to describe field metadata of a table and standardize data information stored in a field.

This topic describes how to create a data standard. A created data standard can be associated with fields in a business table created during ER modeling, ensuring that fields in the business table comply with the specified data standards.

Creating a Data Standard Directory

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

Figure 5-46 DataArts Architecture




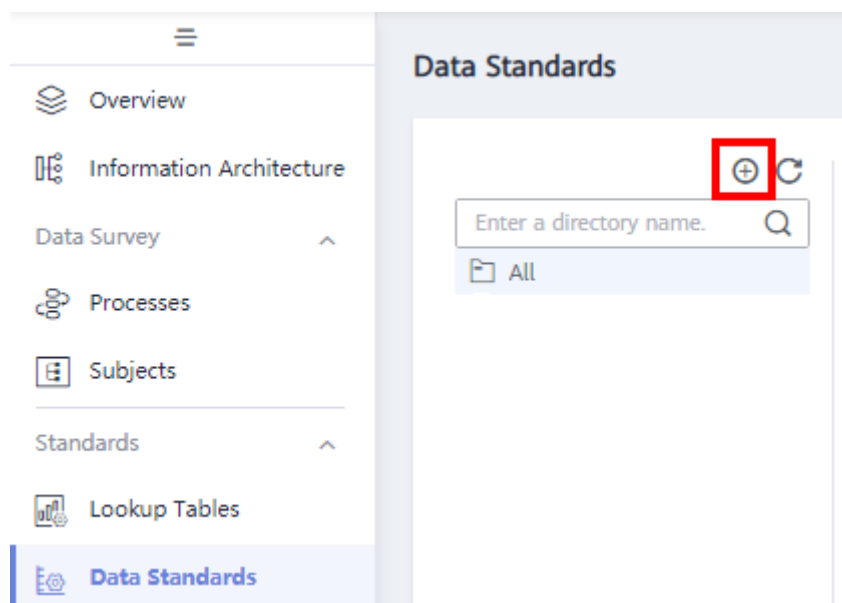
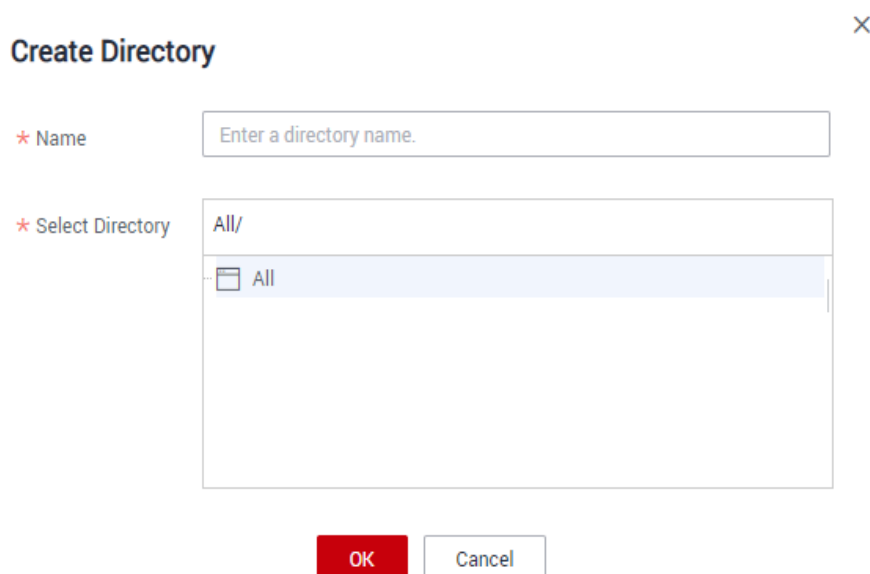
2. On the DataArts Architecture page, choose **Standards > Data Standards** in the left navigation pane.
3. When you access the **Data Standards** page for the first time, the page where you can customize a standard template is displayed. Select the required options for **Optional**, add custom items, and click **Update**.
After saving the template settings, you can modify it on the **Standard Templates** tab page of **Configuration Center**. For details, see [Standard Templates](#). When creating a data standard, you must set the selected options in the template.
4. On the **Data Standards** page, select a directory and click  to create a directory under the selected one. When creating a directory for the first time, you can create a directory under the root directory.

Figure 5-47 Data Standards page



5. In the dialog box displayed, set the parameters and click **OK**.

Figure 5-48 Create Directory dialog page

The screenshot shows a dialog box titled "Create Directory" with a close button (X) in the top right corner. The dialog contains two main sections:

- Name:** A text input field with a red asterisk icon to its left. The placeholder text inside the field is "Enter a directory name."
- Select Directory:** A tree view with a red asterisk icon to its left. The root node is "All/". Underneath it, there is a sub-item "All" which is currently selected and highlighted in light blue.

At the bottom of the dialog, there are two buttons: a red "OK" button and a white "Cancel" button.

Table 5-14 Parameters for creating directories

Parameter	Description
Name	Only letters, numbers, and underscores (_) are allowed.
Select Directory	Select an existing directory, and create a subdirectory under it.

Creating a Data Standard


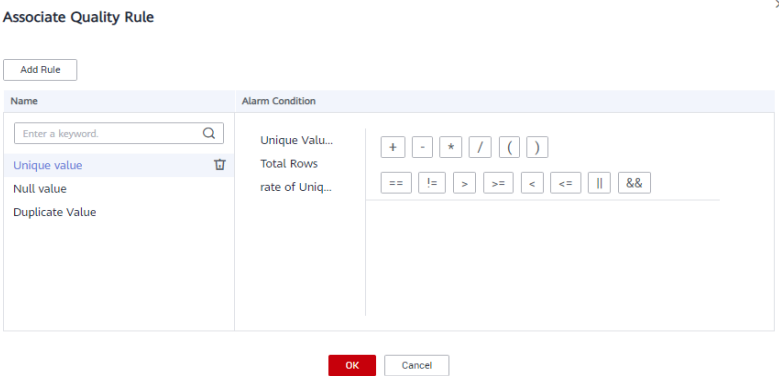
Step 1 On the **Data Standards** page, select a directory and click **Create**.

Step 2 Set the parameters based on [Table 5-15](#) and click **Publish**.

On the page for creating a data standard, only the selected parameters and custom parameters that have been added on the **Standard Templates** tab page of the **Configuration Center** are displayed. [Table 5-15](#) lists all parameters that are available in a data standard template. For details on how to configure a data standard template, see [Standard Templates](#).

Table 5-15 Parameters for creating a data standard

Parameter	Description
Standard Name	The name can contain only letters, digits, brackets, spaces, hyphens (-), and underscores (_), and must start with letters. If Data Standard Allows Duplicate Names is disabled, ensure that the standard name is unique in the current workspace. To check whether Data Standard Allows Duplicate Names is enabled, go to DataArts Architecture > Configuration Center > Functions .
Standard Code	The value can be Auto Generate or Custom . The value must be unique in the current workspace. It is used to identify a data standard record. For details, see Table 5-2 .
Data Type	The possible values are STRING, BIGINT, DOUBLE, TIMESTAMP, DATE, BOOLEAN, and DECIMAL . The data type varies according to the system. The system converts the data type internally. If the required data type does not exist, you can add one. See Data Types .
Data Length	Data length <ul style="list-style-type: none"> • You can leave this parameter blank. If it is left blank, there is no limit to the data length. • You can enter a value from 1 to 10000. • You can also set a range by entering the minimum and maximum values. If you set this parameter and select STRING for Data Type , a data quality job will be created for the attribute matching the data standard. If you select any other data type, no data quality job will be created.
Allowed Value Exist	If Allowed Value Exist is enabled, you can specify one or more allowed values.
Lookup Table	Select a created lookup table and the corresponding table fields. In this way, the lookup table fields can be associated with data standard. If no lookup table is created, create one. See Creating Lookup Tables . If Create Data Quality Jobs is selected for Model Design Process on the Function Settings tab page of Configuration Center , and the data standard of the referenced lookup tables is associated with the business tables in ER modeling, the system will automatically create quality jobs in DataArts Quality when the business tables are published, and generate quality rules based on the associated data standard and lookup tables. If the quality jobs have already been published, the system will automatically update the quality jobs and add the quality rules generated based on the data standard and lookup tables.

Parameter	Description
Quality Rule	<p>If Create Data Quality Jobs is selected for Model Design Process on the Function Settings tab page of Configuration Center, your data standard template includes the parameter Quality Rule, and the data standard is associated with a table, the system will automatically create a quality job in DataArts Quality after the table is published. The created quality job contains the quality rule added here. If the table associated with the data standard to be created has been published, the system will automatically update a quality job.</p> <p>Click . In the dialog box displayed, click Add Rule.</p> <p>For example, add a rule named Unique value, select the rule, click OK, enter an alarm condition expression in the Alarm Condition text box, add other rules in the same way, and click OK.</p> <p>An alarm condition expression consists of alarm parameters and logical operators. When a quality job is running, the system calculates the result of the alarm condition expression and determines whether to trigger the alarm based on the result of the expression. If the expression result is true, the alarm will be triggered. Otherwise, no quality alarm will be triggered.</p> <p>The alarm parameters of each data quality rule are listed as buttons.</p> <p>Figure 5-49 Associate Quality Rule dialog box</p> 
Rule Designer	<p>Select a rule designer from the drop-down list box. This owner is responsible for making quality rules. You can enter an owner name or select an existing owner.</p>
Rule Implementer	<p>Select a rule implementer from the drop-down list box. This owner is responsible for implementing quality rules. You can enter an owner name or select an existing owner.</p>
Level	<ul style="list-style-type: none"> ● global indicates the global level. ● domain indicates non-global level.

Parameter	Description
Custom Item	A custom item added on the Standard Templates tab page in Metrics > Configuration Center . You can add one or more custom items based on project requirements. For more information about adding custom items, see Standard Templates .
Description	A description of the data standard to create. Up to 600 characters are supported.

Step 3 Click **Save**.

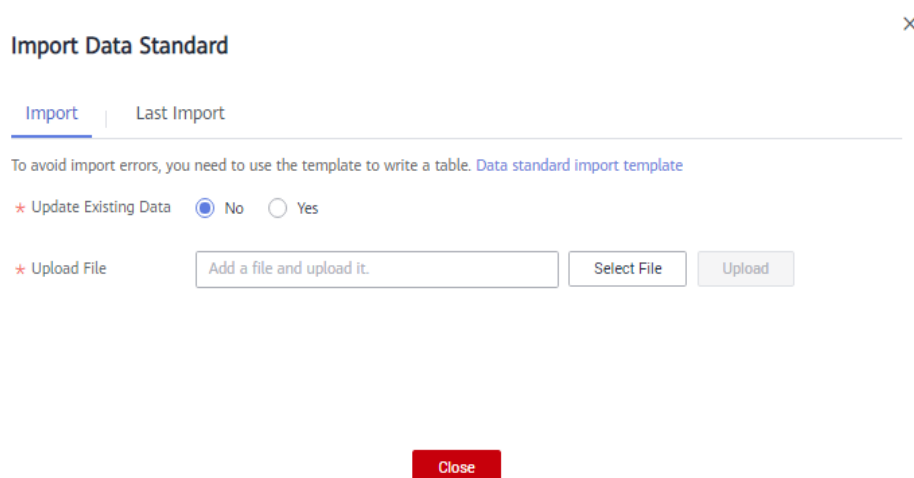
----End

Importing a Data Standard

Step 1 On the DataArts Architecture page, choose **Standards > Data Standards** in the left navigation pane.

Step 2 In the directory structure of data standards, select a directory and choose **More > Import**.

Figure 5-50 Import Data Standard dialog box



Step 3 In the **Import Data Standard** dialog box, determine whether to update the existing data. Existing data is uniquely identified by a standard code. If a standard code in the import template already exists in the current workspace, the system considers that the group of data to which the standard code in the import template belongs already exists.

Step 4 On the **Import** tab page, click **Data standard import template** to download the template. Open the template, set the parameters in the template based on service requirements, and save the settings.

[Table 5-16](#) and [Table 5-17](#) describe the parameters required for importing a data standard. Parameters whose names start with an asterisk (*) are mandatory, and other parameters are optional.

Table 5-16 Parameters in the Standards sheet

Parameter	Description
Directory	The directory that the imported data standard belongs to.
*Standard Name	The name of the data standard to import. It must start with letters. Only letters, digits, brackets, spaces, hyphens (-), and underscores (_) are allowed.
*Standard Code	The value can be Auto Generate or Custom . The value must be unique in the workspace. It is used to identify a data standard record. For details, see Table 5-2 .
*Data Type	The possible values are STRING , BIGINT , DOUBLE , TIMESTAMP , DATE , BOOLEAN , and DECIMAL . The data type varies according to the system. The system converts the data type internally. If the required data type does not exist, you can add one. See Data Types .
Data Length	You can enter a value ranging from 1 to 10,000. If it is left blank, there is no limit to the data length. If you enter a value and select STRING for Data Type , a data quality job will be created for the attribute matching the data standard. If you select any other data type, no data quality job will be created.
Allowed Value	The value true indicates that there are allowed values, and the value false indicates that there are no allowed values.
Allowed Value List	If you select true for Allowed Value , you must enter an allowed value. You can add up to 20 values. Multiple values must be separated by commas (,), for example, 1,2,3 .
Lookup Table	Set this parameter to the name of a created lookup table.
Lookup Table Field	If Lookup Table is not left blank, you must set Lookup Table Field . In this way, the code table field can be associated with the data standard.
Owner of Business Rules	Enter the business rule owner. You can enter the name of an owner or select an existing owner.
Owner of Data Monitoring	Enter the data monitoring owner. You can enter the name of an owner or select an existing owner.
Standard Level	<ul style="list-style-type: none"> ● global indicates the global level. ● domain indicates non-global level.
Description	A description of the data standard to import. Up to 600 characters are supported.

Parameter	Description
(Optional) Custom Item	If you have added one or more custom fields when customizing a data standard template, you must also fill in the corresponding fields in the import template. If no custom field is added, you do not need to fill in the fields. For details on how to customize a data standard template, see Standard Templates .

In the **Quality Rules** sheet, you can configure quality rules to be added for data standards. If **Create Data Quality Jobs** is selected for **Model Design Process** on the **Function Settings** tab page of **Configuration Center**, your data standard template includes the parameter **Quality Rule**, and the data standard is associated with a table, the system will automatically create a quality job in DataArts Quality after the table is published. The created quality job contains the quality rule added here. If the table associated with the data standard to be created has been published, the system will automatically update a quality job.

Table 5-17 Parameters in the Quality Rules sheet

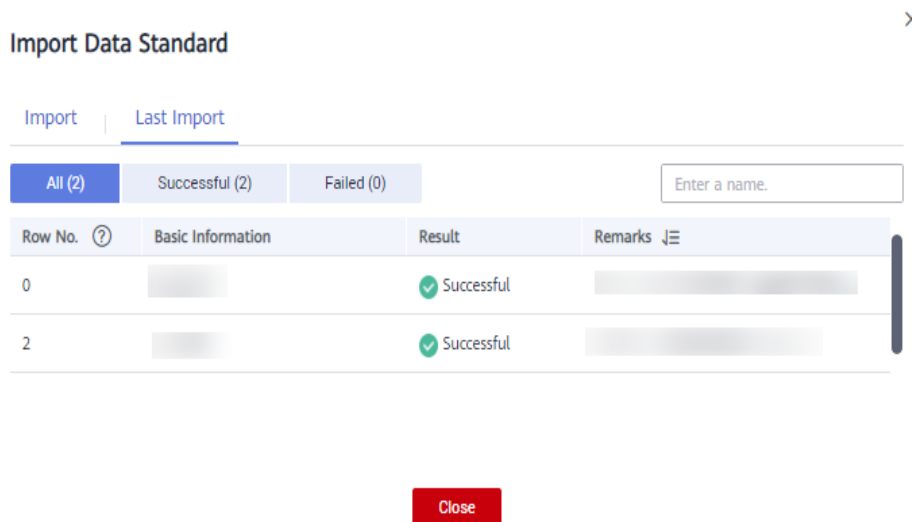
Parameter	Description
*Code	The code of the data standard that a quality rule is added to.
Rule Name	Enter an existing rule name. In the upper left corner of the DataArts Studio console, select DataArts Quality from the drop-down list box. Then, you can view the existing rule names on the Rule Templates page.
Alarm Config	An alarm condition expression consists of alarm parameters and logical operators. When a quality job is running, the system calculates the result of the alarm condition expression and determines whether to trigger the alarm based on the result of the expression. If the expression result is true , the alarm will be triggered. Otherwise, no quality alarm will be triggered. In the alarm condition expression, alarm parameters are represented by variables such as $\${1}$, $\${2}$, and $\${3}$. The variable name indicates the alarm parameter of the specified quality rule. The variable $\$1$ indicates the first alarm parameter, $\$2$ indicates the second alarm parameter, and so on. In the upper left corner of the DataArts Studio console, select DataArts Quality from the drop-down list box. Access the Rule Templates page and view the alarm parameters supported by the data quality rule in the Result Description column. Example: $\${1} > 100$
Expression	A regular expression must be configured when Rule Name is set to Regular Expression or Validity Verification .

Step 5 Return to the **Import Data Standard** dialog box, select the data standard template file configured in the previous step, and click **Upload**.

If the uploaded template file fails the verification, modify the file and upload it again.

Step 6 In the **Import Data Standard** dialog box, the import result is displayed on the **Last Import** tab page. If the import is successful, click **Close**. If the import fails, you can view the failure cause, correct the template file, and upload it again.

Figure 5-51 Last Import tab page



----End

Managing a Data Standard

On the DataArts Architecture page, choose **Standards > Data Standards** in the left navigation pane. On the page displayed, you can manage data standards as required.

Figure 5-52 List of data standards



On the **Data Standards** page, you can perform the following operations:

- **Search**

Above the data standard list, select a filter such as the standard name, data type, creator, and reviewer, and click the search icon to search for data standards.

After locating the specified data standards, you can perform the following operations:

- Edit
- Publish
- Suspend

- **Import**
Choose **More > Import** to import a data standard. Download the template, fill in it and upload it, and click **Close**.
- **Export**
 - Export data standards from a specified directory.
In the data standard directory structure, select a directory and choose **More > Export** above the data standard list to export all data standards in the directory.
 - Export specified data standards.
In the data standard list, select the data standards you want to export and choose **More > Export** above the list to export the selected data standards.
- **Delete**
Select a data standard, and choose **More > Delete**. A data standard in the pending publishing, published, or pending suspension state cannot be deleted. Referenced data standards cannot be deleted as well.
- **Publish**
Select a data standard and click **Publish**. In the displayed dialog box, perform either of the following operations:
 - Select a reviewer. If no reviewer is available in the drop-down list, click **+** to add one.
 - Select **Auto-review**.

 **NOTE**

Auto-review is available only when the current account is in the reviewer list. Click **OK**. If a reviewer is selected, the data standard is published after the application is approved. If **Auto-review** is selected, the data standard will be published immediately.

Exporting a Data Standard

- Step 1** On the DataArts Architecture page, choose **Standards > Data Standards** in the left navigation pane.
 - Step 2** In the data standard directory structure, right-click a directory name and choose **Export**.
- End

5.6 Model Design

5.6.1 ER Modeling

5.6.1.1 Designing Logical Models

A logical model is an entity relationship diagram that accurately describes business rules based on entities and their relationships. Logical models must

ensure the correctness and consistency of the data structure required by services and use a series of standard rules to reflect the features of various objects, and accurately define the relationships between entities.

In addition, logical models provide a reliable reference for constructing physical models and can be converted into physical models. Logical models are key to a successful database design.

The following parts are included in this topic:

- [Considerations in Logical Model Design](#)
- [Creating a Logical Model](#)
- [Creating and Publishing a Logical Entity](#)
- [Converting a Logical Model to a Physical Model](#)

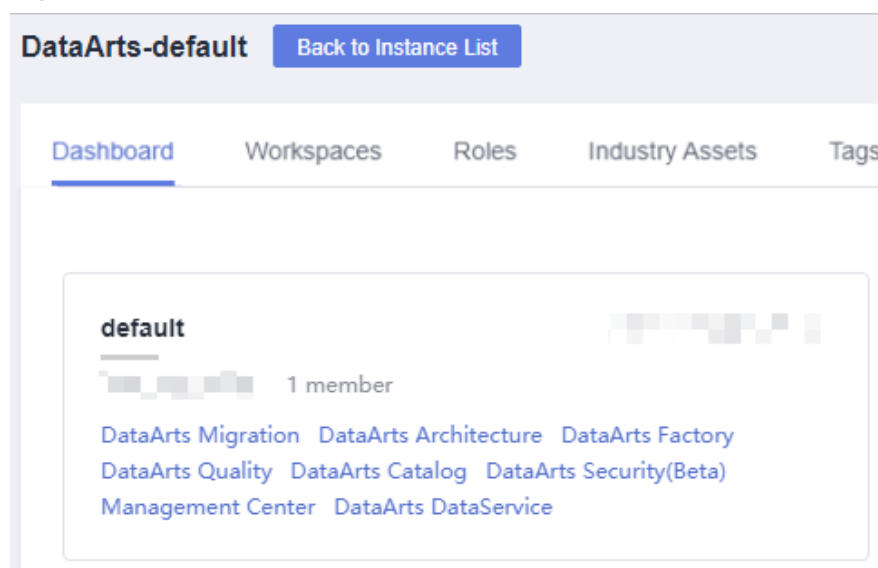
Considerations in Logical Model Design

- You must consider not only the current business status, but also the future business development.
- Personnel who are familiar with the businesses must participate in the modeling. In this way, the business requirements can be fully integrated into the models.
- Converting the logical model to the physical model must be efficient.
- You must consider physical features during physical modeling.
- Each entity, attribute, and relationship must be consistent with the information in the actual business.

Creating a Logical Model

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

Figure 5-53 DataArts Architecture



2. On the DataArts Architecture page, choose **Models > ER Modeling** in the left navigation pane.

- On the **ER Modeling** page, if no ER model has been created, the system displays a dialog box asking you to create one. If you have created ER models before, click **+** to create models.

Figure 5-54 Creating a hierarchical governance model

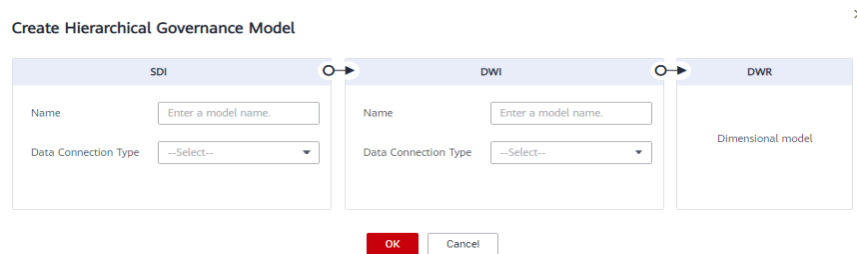
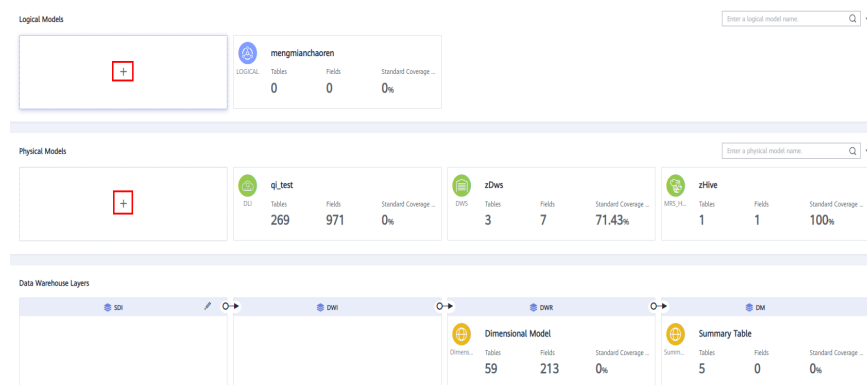


Figure 5-55 ER Modeling page



- In the dialog box displayed, set the parameters and click **OK**.

Figure 5-56 Creating a logical model

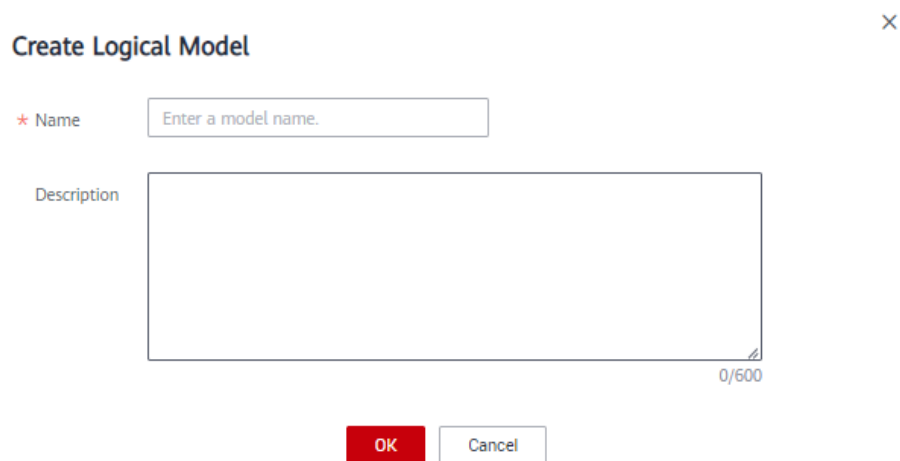


Table 5-18 Parameters for creating a logical model

Parameter	Description
Name	Only letters, numbers, and underscores (_) are allowed.
Description	A description of the logical model.

Creating and Publishing a Logical Entity

A logical entity is a logical table. After creating a logical model, you can create a logical entity in the model.


- Step 1** On the DataArts Architecture page, choose **Models > ER Modeling** in the left navigation pane.
- Step 2** On the displayed page, click a logical model to access its management page. Then, click **Create**.
- Step 3** On the displayed page, configure parameters as prompted.
1. Set the basic parameters.

Figure 5-57 Basic Settings

The screenshot shows the 'Basic Settings' tab of a configuration page. It features several input fields and a radio button group. The 'Subject' field is a dropdown menu with '--Select--'. The 'Logical Entity Name' field is a text input with the placeholder 'Enter a logical entity name.'. The 'Parent Logical Entity' field is a dropdown menu with '--Select--'. The 'Tag' field has a circular icon with a plus sign. The 'Owner' field is a text input with the placeholder 'Enter an asset owner.' and a 'C' icon. The 'Description' field is a large text area containing 'None'. To the right, there is a 'Logical Entity Code' section with two radio buttons: 'Auto Generate' (selected) and 'Custom'. Below this is a text input field with 'undefined'. The page number '4/200' is visible in the bottom right corner.

Table 5-19 Parameters on the Basic Settings tab page

Parameter	Description
* Subject	Select a subject from the drop-down list box.
Logical Entity Code	You can select Auto Generate or Custom .
* Table Name	Logic entity name. It must start with letters. Only letters, digits, and the following special characters are allowed: ()-_
* Table Code	Name of the physical table converted from the logical entity. Table codes cannot start with numbers. Only letters, numbers, and the following special characters are allowed: _\${}

Parameter	Description
Parent Logical Entity	Set a parent logical entity, which is inherited by child logical entities. Common logical entities and attributes can be logically abstracted as a parent logical entity. After specific attributes are added to the parent logical entity, a child logical entity is generated. The modifications to the attributes in a parent logical entity affect all child logical entities that inherit it.
Tag	Tags are custom identifiers that help you classify and search for data assets. After adding a tag, you can search for related data assets in the DataArts Catalog module with ease. Click  . In the dialog box displayed, select one or more existing tags, or enter a new tag name and press Enter . You can also go to the Tags page of the DataArts Catalog module to add a tag. Then, return to this page and select the newly added tag from the drop-down list box. For details, see Tags .
Owner	You can enter an owner name or select an existing owner.
* Description	A description of the table to create. It allows 1 to 200 characters.

2. On the **Logical Entity Attributes** page, add required attributes. [Table 5-20](#) lists the parameters for logical entity attributes.

Figure 5-58 Adding a logical entity attribute

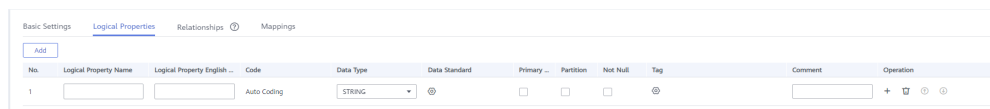




Table 5-20 Parameters for logical entity attributes

Parameter	Description
*Field Name	It must start with letters. Only letters, digits, and the following special characters are allowed: ()-_
*Field English Name	Only letters, numbers, and underscores (_) are allowed. A field code must start with a letter.
*Code	Code of the logical attribute. If the logical entity uses a custom code, the code of the logical attribute can be customized or automatically generated.
Data Type	Data type of the attribute. If you cannot find a desired data type from the drop-down list box, you can add a data type by referring to Data Types .


Parameter	Description
Data Standard	<p>If you have created data standards, click  to select one to associate with the logical entity attribute. If Create Data Quality Jobs is selected for Model Design Process on the Function Settings tab page in Configuration Center and a logical entity attribute is associated with a data standard, a quality job is automatically generated after a logical entity attribute is published. A quality rule is generated for each logical entity attribute associated with the data standard. The quality of the logical entity attribute is monitored based on the data standard. You can access the Quality Job page of DataArts Quality to view the job details.</p> <p>If no data standard is available, create one. See Creating Data Standards for details.</p>
Primary Key	If this parameter is selected, the attribute is a primary key.
Partition	If this parameter is selected, the attribute is a partition field.
Not Null	Whether the parameter value can be left empty.
Tag	<p>You can click  to add a tag for the logical entity attribute.</p> <ul style="list-style-type: none"> - In the dialog box displayed, select one or more existing tags. If no tag has been added, you can go to the Tags page of the DataArts Catalog module to add a tag. For details, see Tags. - In the dialog box displayed, enter a new tag name and press Enter. Tag names can contain letters, numbers, and underscores (_), but cannot start with underscores (_).
Description	A description of the table to create.

3. On the **Relationships** tab page, click **Add** to create a relationship.

A relationship refers to the association between a parent and a child entity (also called a primary and a secondary entity). It describes how an entity is associated with another entity, or the impact of an entity's behavior on another entity. Relationships between entities in a data model are particularly important and must be accurately defined. Otherwise, the actual business rules cannot be accurately described in the data model, and data consistency is greatly damaged.

For example, if the **student ID** attribute of a score table is the primary key for a student table, the relationship between the two tables designed according to the third normal form (3NF) is as follows:

- Child logical entity: score table
- Child logical entity attribute FK: student ID

- Child to parent:  1
- Parent logical entity: student table
- Parent logical entity attribute PK: student ID


- Parent to child:  1

Figure 5-59 Adding a relationship

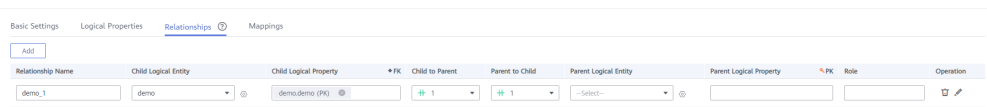













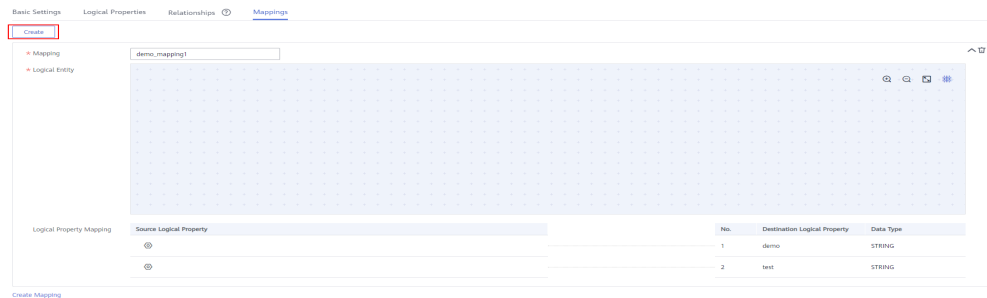
Table 5-21 Parameters on the Relationships tab page

Parameter	Description
Name	Name of the relationship
Child Logical Entity	Select a child logical entity from the drop-down list box. Click  to set the current logical entity as a child logical entity. For example, if the student ID attribute of a score table is the primary key for a student table, the child logical entity is the score table, and the corresponding parent logical entity is the student table.
Child Logical Entity Attribute FK	Foreign key of the child logical entity attribute. The attribute of the child logical entity must be the foreign key of the parent logical entity. For example, if the student ID attribute of a score table is the primary key for a student table, the foreign key of the child logical entity attribute is the student ID in the score table.
Child to Table	<p> 1 indicates that each piece of data in the child logical entity corresponds to only one piece of data in the parent logical entity.</p> <p> 0,1 indicates that each piece of data in the child logical entity corresponds to at most one piece of data in the parent logical entity.</p> <p> 0..n indicates that one piece of data in the child logical entity corresponds to multiple pieces of data in the parent logical entity.</p> <p> 1..n indicates that one piece of data in the child logical entity corresponds to one piece of data in the parent logical entity at least.</p>

Parameter	Description
Parent to Child	<p> 1 indicates that the data in the parent logical entity is in one-to-one relationship with the data in the child logical entity.</p> <p> 0,1 indicates that each piece of data in the parent logical entity corresponds to at most one piece of data in the child logical entity.</p> <p> 0..n indicates that one piece of data in the parent logical entity corresponds to multiple pieces of data in the child logical entity.</p> <p> 1..n indicates that each piece of data in the parent logical entity corresponds to at least one piece of data in the child logical entity.</p>
Parent Logical Entity	<p>Select a logical entity that has a logical relationship with the selected child logical entity.</p> <p>For example, if the student ID attribute of a score table is the primary key for a student table, the parent logical entity is the student table, and the corresponding child logical entity is the score table.</p>
Parent Logical Entity Attribute PK	<p>Primary key of the parent logical entity attribute. The attribute of the parent logical entity must be the primary key of the parent logical entity.</p> <p>For example, if the student ID attribute of a score table is the primary key for a student table, the primary key of the parent logical entity attribute is the student ID in the student table.</p>
Role	You can customize a role name to identify the relationship.
Operation	Click  to delete a relationship. Click  to edit the relationship.

4. On the **Mappings** page, click **Create** to create a mapping. Then click **Save**. Mapping means setting up a mapping relationship between the source and destination logical entity.

Figure 5-60 Creating a mapping



- **Mapping** is automatically generated when a mapping is created. You can change the value.
- **Source Logical Entity:** If data comes from multiple logical entities of a model, you can click next to a logical entity to establish a JOIN relationship between the logical entity and another logical entity.

Figure 5-61 Setting the JOIN condition for the source table

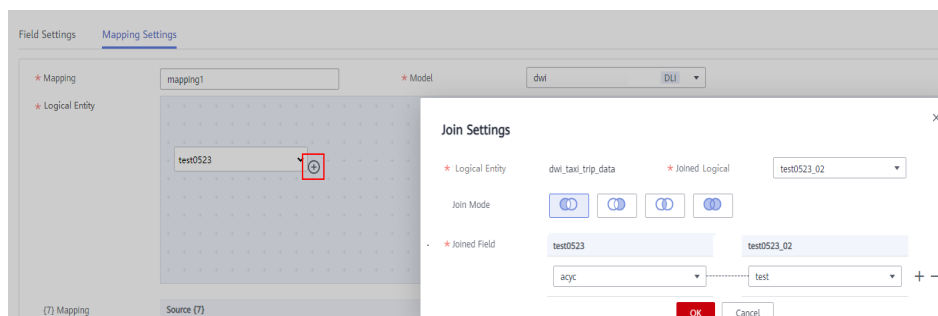


Table 5-22 JOIN conditions

Parameter	Description
Joined Logical Entity	Select a logical entity for which you want to establish a JOIN relationship with the source logical entity.
Joined Mode	Left JOIN, right JOIN, inner JOIN, and outer JOIN are represented from left to right.
Joined Attribute	Generally, the JOIN attribute in the source logical entity is the same as that in the joined logical entity. You can click or to add or delete a JOIN attribute. The relationship between JOIN attributes is AND.

- **Logical Attribute Mapping:** Select a source attribute with the same meaning as the current attribute.

Step 4 Click **Publish**, select a reviewer, and click **Submit**.

Wait for the reviewer to approve the application. After the application is approved, return to the model list and view the created logical entity in the list.

NOTE

By default, **Synchronize logical assets** is selected for **Model Design Process** on the **Functions** tab page of the **Configuration Center** page.

- For new logical models, you can click **Publish** to synchronize them to the logical assets of the DataArts Catalog module.
- For historical logical models, you can click **More** and select **Synchronize** from the drop-down list box to synchronize them to the logical assets of the DataArts Catalog module.

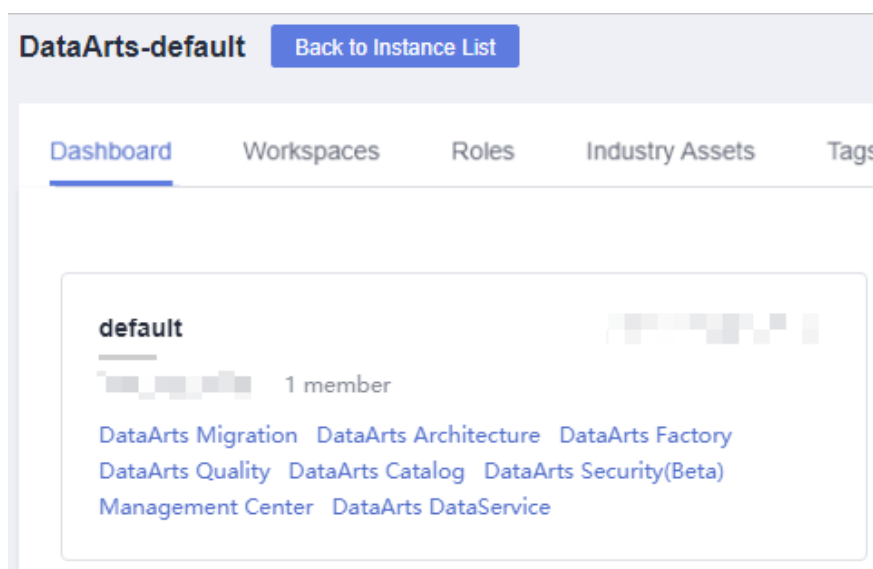
----End

Converting a Logical Model to a Physical Model

After a logical model is created, you can convert it to a new physical model or an existing physical model.

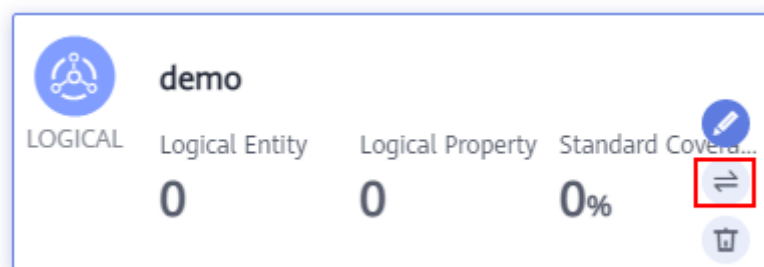
1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

Figure 5-62 DataArts Architecture



2. On the DataArts Architecture page, choose **Models > ER Modeling** in the left navigation pane.
3. Find the required logical model and click the conversion button on the model.

Figure 5-63 Logical model conversion



- In the **Convert to Physical Model** dialog box, set the parameters and click **OK**.

Figure 5-64 Convert to Physical Model dialog box

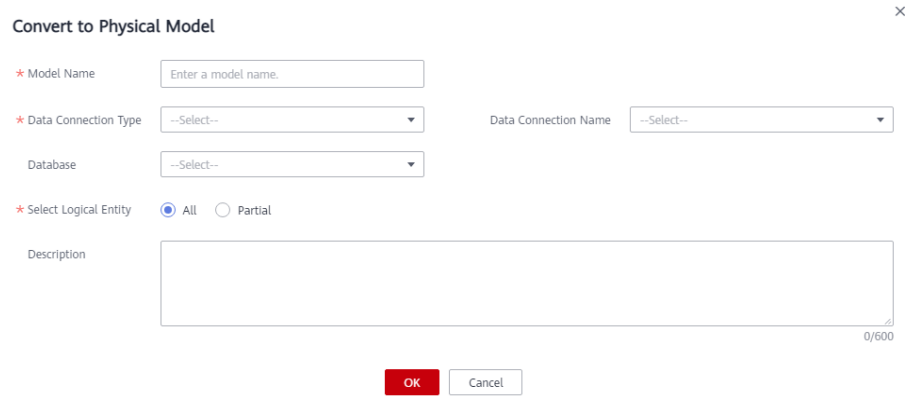


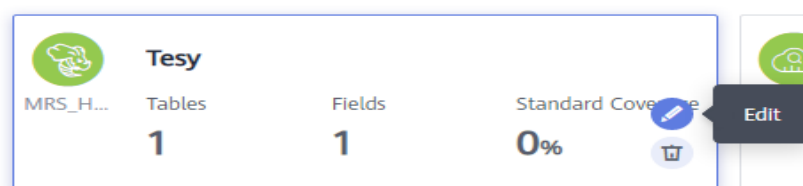
Table 5-23 Parameters

Parameter	Description
*Model Name	The name of the physical model to be converted from a logical model. You can enter a new model name, and then the system creates the model. You can also select an existing model name from the drop-down list box. Only letters, numbers, and underscores (_) are allowed.
*Data Connection Type	Select a data connection type from the drop-down list box. If the required data type does not exist, you can add one by referring to Data Types .
Data Connection	The name of the data connection. Select the required data connection. You are advised to use the same data connection for an ER model. If no data connection is available, access Management Center to create one. For details, see Creating Data Connections .
Database	The name of the database. Select a database from the drop-down list box. If no database is available, access DataArts Factory to create one. For details, see Creating a Database .
Tables	<ul style="list-style-type: none"> All: Convert all logical entities into physical tables. Partial: Convert the selected logical entities into physical tables.
Queue	DLI queue. This parameter is available only for DLI data connections.
Schema	Schema of DWS or POSTGRESQL. This parameter is available only for DWS and PostgreSQL data connections.

Parameter	Description
Description	A description of the model. Up to 600 characters are supported.

- After the model is converted to a physical model, you can set layers for the physical model. You can select the SDI or DWI layer. As shown in [Figure 5-65](#), move the cursor to the card of the physical model and click the edit button of the model.

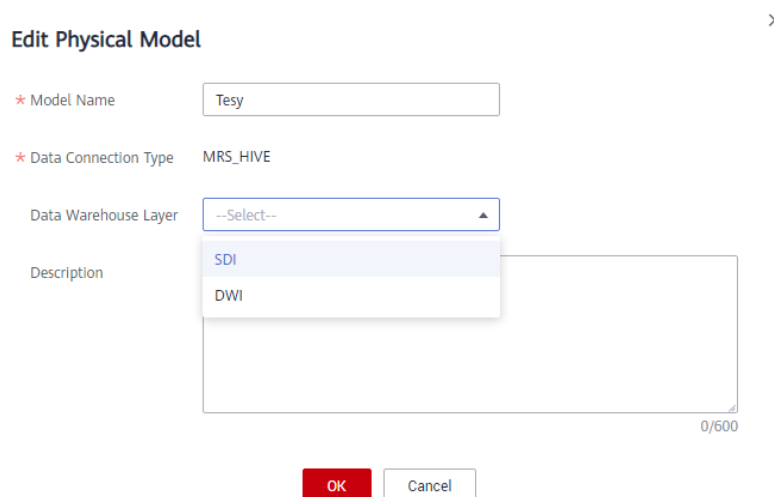
Figure 5-65 Setting layers for the physical model



In the displayed dialog box, select **SDI** or **DWI** for **Data Warehouse Layer**.

- **SDI** stands for Source Data Integration and is the source data layer. SDI is a simple implementation of source system data.
- **DWI** stands for Data Warehouse Integration, also called the data consolidation layer. DWI integrates and cleans data from multiple source systems, and implements entity relationship modeling based on the three normal forms.

Figure 5-66 Editing the physical model



5.6.1.2 Designing Physical Models

A physical model is a physical description about the conversion of elements such as entities, attributes, attribute constraints, and relationships from a logical model to a table relationship diagram that can be identified by database software using certain rules and methods.

On the **ER Modeling** page, you can create an SDI and a DWI layer. The models are implemented through physical modeling. In addition to converting a logical model to a physical model, you can directly create a physical model.

The following parts are included in this topic:

- [Considerations in Physical Model Design](#)
- [Creating a Physical Model](#)
- [Creating and Publishing a Table](#)

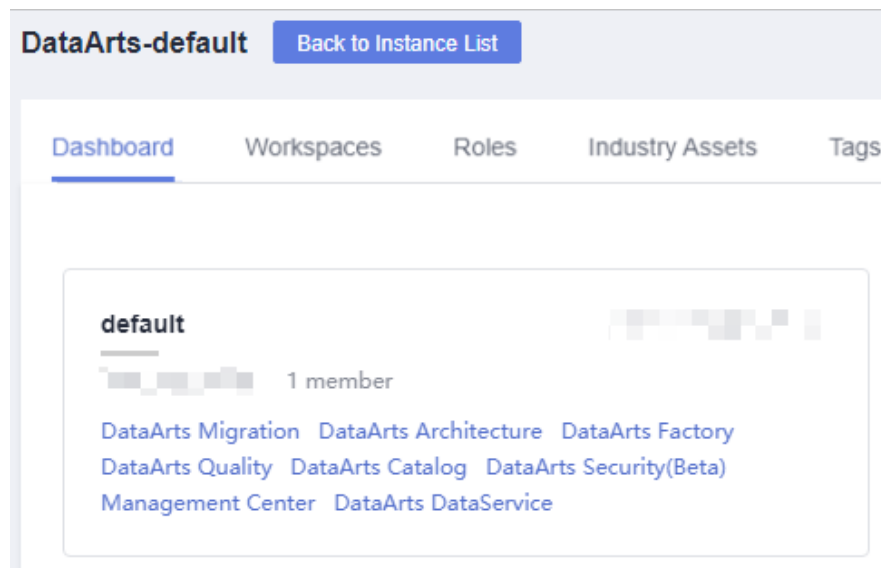
Considerations in Physical Model Design

- Physical models must ensure that the required functions are available and their performance is as good as expected.
- Physical models must ensure data consistency and quality.
- Few or no changes are made to the physical models when new services or functions are added.

Creating a Physical Model

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

Figure 5-67 DataArts Architecture



2. On the DataArts Architecture page, choose **Models > ER Modeling** in the left navigation pane.
3. On the **ER Modeling** page, if no ER model has been created, the system displays a dialog box asking you to create one. If you have created ER models before, click **+** to create models.

Figure 5-68 Creating a hierarchical governance model

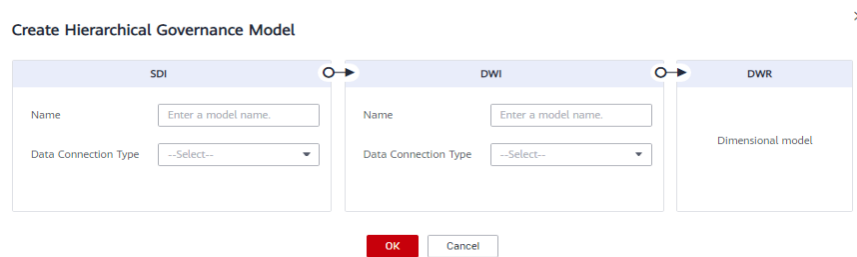
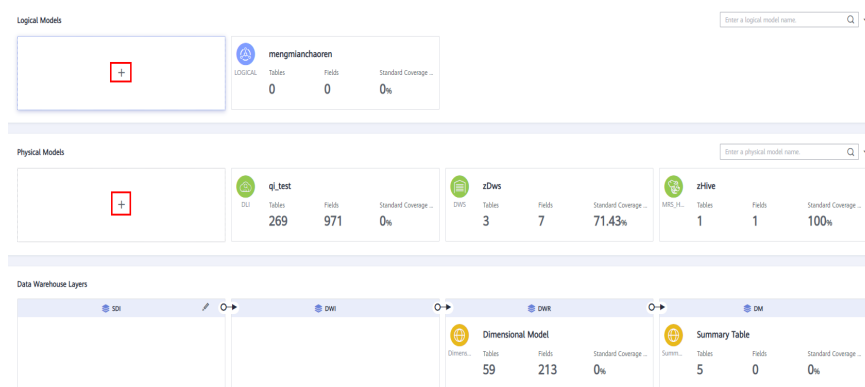


Figure 5-69 ER Modeling page



4. In the dialog box displayed, set the parameters and click **OK**.

Figure 5-70 Creating a model

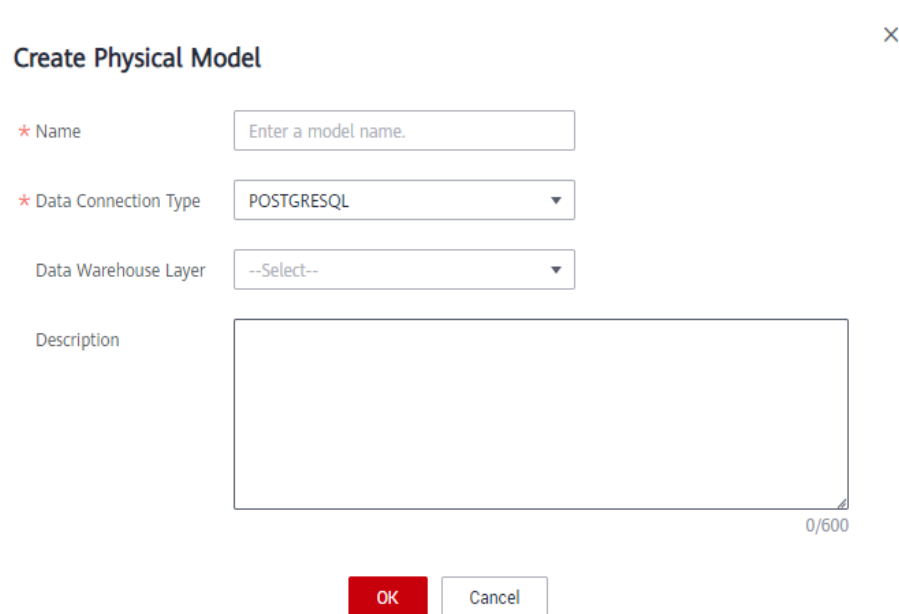


Table 5-24 Parameters for creating a physical model

Parameter	Description
Name	Only letters, numbers, and underscores (_) are allowed.

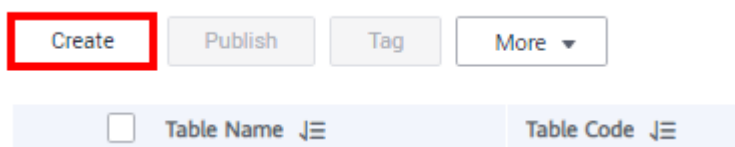
Parameter	Description
Data Connection Type	Select a data connection type from the drop-down list box.
Data Warehouse Layer	Select SDI or DWI . <ul style="list-style-type: none">• SDI stands for Source Data Integration and is the source data layer. SDI is a simple implementation of source system data.• DWI stands for Data Warehouse Integration, also called the data consolidation layer. DWI integrates and cleans data from multiple source systems, and implements entity relationship modeling based on the three normal forms.
Description	A description of the ER model. Up to 600 characters are supported.

Creating and Publishing a Table

After creating a DLI, POSTGRESQL, DWS or MRS Hive ER model, you can create a business table in the model.

- Step 1** On the DataArts Architecture page, choose **Models > ER Modeling** in the left navigation pane.
- Step 2** Select the physical model for which you want to create a table, click the physical model to access the model management page, and click **Create**.

Figure 5-71 Entry for creating a table



- Step 3** On the **Create Table** page, set the parameters as required.
1. Set the basic parameters.

Figure 5-72 Basic Settings tab page

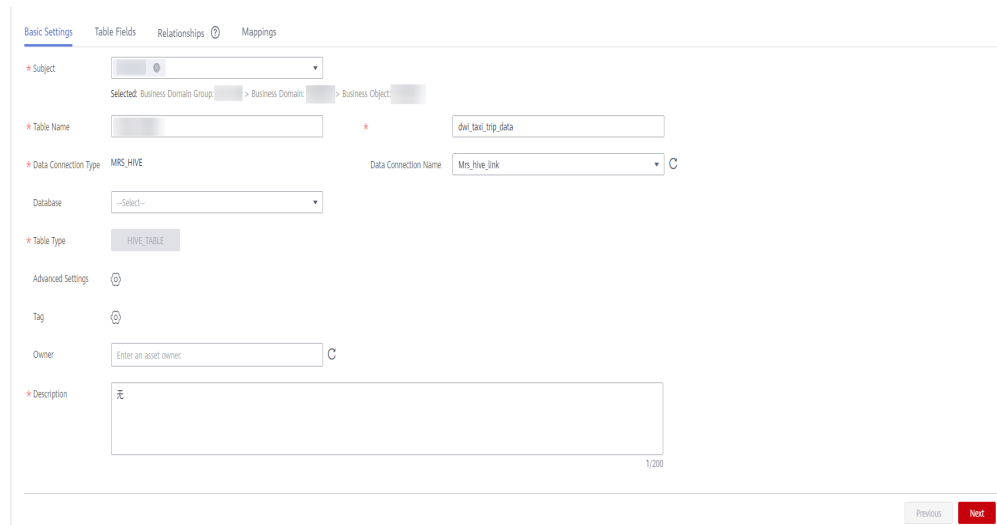



Table 5-25 Parameters on the Basic Settings tab page

Parameter	Description
Subject	Select a subject from the drop-down list box.
Name	The name of the table to create. Table names must start with letters. Only letters, numbers, and the following special characters are allowed: ()-_
Table Code	The code of the table to create. Table codes cannot start with numbers. Only letters, numbers, and the following special characters are allowed: _\${}
Data Connection Type	N/A
Data Connection	The name of the data connection. Select the required data connection. You are advised to use the same data connection for an ER model. If no data connection is available, access Management Center to create one. For details, see Creating Data Connections .
Database	The name of the database. Select a database from the drop-down list box.
Queue	DLI queue. This parameter is available only for DLI tables.
Schema	Schema of DWS or PostgreSQL. This parameter is available only for DWS and PostgreSQL tables.

Parameter	Description
Table Type	<p>DLI models support the following table types:</p> <ul style="list-style-type: none">- MANAGED: Data is stored in a DLI table.- EXTERNAL: Data is stored in an OBS table. When Table Type is set to EXTERNAL, you must set OBS Path. The OBS path format is <i>/bucket_name/filepath</i>. <p>DWS models support the following table types:</p> <ul style="list-style-type: none">- DWS_ROW: Tables are stored to disk partitions by row.- DWS_COLUMN: Tables are stored to disk partitions by column.- DWS_VIEW: Tables are stored to disk partitions by view. <p>The MRS_HIVE model supports only HIVE_TABLE.</p>
Data Format	<p>This parameter is available only for DLI tables. DLI models support the following table types:</p> <ul style="list-style-type: none">- Parquet: DLI can read non-compressed data or Parquet data that is compressed using Snappy and GZIP.- CSV: DLI can read non-compressed data or CSV data that is compressed using GZIP.- ORC: DLI can read non-compressed data or ORC data that is compressed using Snappy.- JSON: DLI can read non-compressed data or JSON data that is compressed using GZIP.- Carbon: DLI can read non-compressed Carbon data.- Avro: DLI can read non-compressed Avro data.
Advanced Settings	<p>Set custom items to describe the table. The custom items can be viewed in the table details.</p> <p>For example, if you want to identify the source of the table, you can add item source and set its value to the table source information. Then you can view the table source information in the table details.</p>
Tag	<p>Tags are custom identifiers that help you classify and search for data assets. After adding a tag, you can search for related data assets in the DataArts Catalog module with ease.</p> <p>Click . In the dialog box displayed, select one or more existing tags, or enter a new tag name and press Enter. Then press OK. You can also go to the Tags page of the DataArts Catalog module to add a tag. Then, return to this page and select the newly added tag from the drop-down list box. For details, see Tags.</p>
Owner	You can enter an owner name or select an existing owner.

Parameter	Description
Description	A description of the table. It allows 1 to 600 characters.

- Click **Add** to add required fields on the **Table Fields** page.

Figure 5-73 Adding required table fields

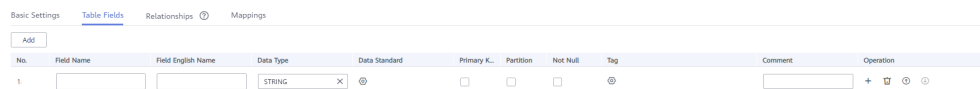




Table 5-26 Parameters on the Table Fields tab page

Parameter	Description
Name	It must start with letters. Only letters, digits, and the following special characters are allowed: ()-_
Code	Only letters, numbers, and underscores (_) are allowed. A field code must start with a letter.
Data Type	Field data type. If the required data type does not exist, you can add one. See Data Types .
Data Standard	If you have created data standards, click  to select one to associate with the field. If Create Data Quality Jobs is selected for Model Design Process on the Function Settings tab page in Configuration Center and a field is associated with a data standard, a quality job is automatically generated after a table is published. A quality rule is generated for each field associated with the data standard. The quality of the field is monitored based on the data standard. You can access the Quality Job page of DataArts Quality to view the job details. If no data standard is available, create one. See Creating Data Standards for details.
Primary Key	If this parameter is selected, the field is a primary key.
Partition	If this parameter is selected, the field is a partition field.
Not Null	Whether the parameter value can be left empty.

Parameter	Description
Tag	<p>Click  to add a tag.</p> <ul style="list-style-type: none"> - In the dialog box displayed, select one or more existing tags. If no tag has been added, you can go to the Tags page of the DataArts Catalog module to add a tag. For details, see Tags. - In the dialog box displayed, enter a new tag name and press Enter. Tag names can contain letters, numbers, and underscores (_), but cannot start with underscores (_).
Description	A description of the field to add.

3. (Optional) On the **Relationships** tab page, click **Add** to create a relationship.

A relationship refers to the association between a parent and a child table (also called a primary and a secondary table). It describes how a table is associated with another table, or the impact of a table's behavior on another table. Relationships between tables in a data model are particularly important and must be accurately defined. Otherwise, the actual business rules cannot be accurately described in the data model, and data consistency is greatly damaged.

For example, if the **student ID** attribute of a score table is the primary key for a student table, the relationship between the two tables designed according to the third normal form (3NF) is as follows:












- Child table: score table
- Child table field FK: student ID
- Child to parent:  1
- Parent table: student table
- Parent table field PK: student ID
- Parent to child:  1



Figure 5-74 (Optional) Adding a relationship



Table 5-27 Parameters on the Relations tab page

Parameter	Description
Name	Name of the relationship

Parameter	Description
Child Table	Select a table from the drop-down list box. Click  to set the current table as a child table. For example, if the student ID attribute of a score table is the primary key for a student table, the child table is the score table, and the corresponding parent table is the student table.
Child Table Field FK	Foreign key of the child table. The field of the child table must be the foreign key of the parent table. For example, if the student ID attribute of a score table is the primary key for a student table, the child table field FK is the student ID in the score table.
Child to Table	<p> 1 indicates that each piece of data in the child table corresponds to only one piece of data in the parent table.</p> <p> 0,1 indicates that each piece of data in the child table corresponds to at most one piece of data in the parent table.</p> <p> 0..n indicates that one piece of data in the child table corresponds to multiple pieces of data in the parent table.</p> <p> 1..n indicates that each piece of data in the child table corresponds to at least one piece of data in the parent table.</p>
Parent to Child	<p> 1 indicates that each piece of data in the parent table corresponds to only one piece of data in the child table.</p> <p> 0,1 indicates that each piece of data in the parent table corresponds to at most one piece of data in the child table.</p> <p> 0..n indicates that one piece of data in the parent table corresponds to multiple pieces of data in the child table.</p> <p> 1..n indicates that one piece of data in the parent table corresponds to at least one piece of data in the child table.</p>
Parent Table	Select the parent table corresponding to the selected child table. For example, if the student ID attribute of a score table is the primary key for a student table, the parent table is the student table, and the corresponding child table is the score table.

Parameter	Description
Parent Table Field PK	Primary key of the parent table. The field of the parent table must be the primary key of the parent table. For example, if the student ID attribute of a score table is the primary key for a student table, the parent table field PK is the student ID in the student table.
Role	You can customize a role name to identify the relationship.
Operation	Click  to delete a relationship. Click  to edit the relationship.

4. (Optional) On the **Mappings** tab page, click **Create** to create a mapping and design a data source based on the created mapping.

- If the table field comes from different relationship models, you must create multiple mappings.

Currently, table data can be obtained from ER models of different connection types. In each mapping, you only need to set the source field for the field that comes from the current mapping. Other fields do not need to be set.

For example, if the data of the first five fields and the last five fields in the current table comes from two different models, create the following mappings:

- **map1:** Create a table named **table01** from ER model A. In the **Field Mapping** area, set the source fields of the first to fifth fields to the corresponding fields with the same meaning in **table01**. The last five fields do not need to be set.
- **map2:** Create a table named **table02** from ER model B. In the **Field Mapping** area, set the source fields of the sixth to tenth fields to the corresponding fields with the same meaning in **table02**. The first five fields do not need to be set.
- If the field data in a table comes from multiple tables in the same ER model, you can create a mapping.

In the source table of the mapping, you can set JOIN conditions for multiple tables, and then set source fields for the fields in the table. The selected source fields must have the same meanings as the fields in the table.

For example, all fields in the current table come from ER model **d1**, the first, second, and third fields come from the **vendor**, **payment_type**, and **rate** tables respectively, and other fields come from the **dwd_taxi_trip_data** table.

You can create a mapping, as shown in [Figure 5-75](#). Join the **dwd_taxi_trip_data** table with the **vendor**, **payment_type**, and **rate** tables, and set the source fields in sequence in the field mapping.

For details on the parameters for creating a mapping, see [Table 5-28](#).

Figure 5-75 Configuring a mapping

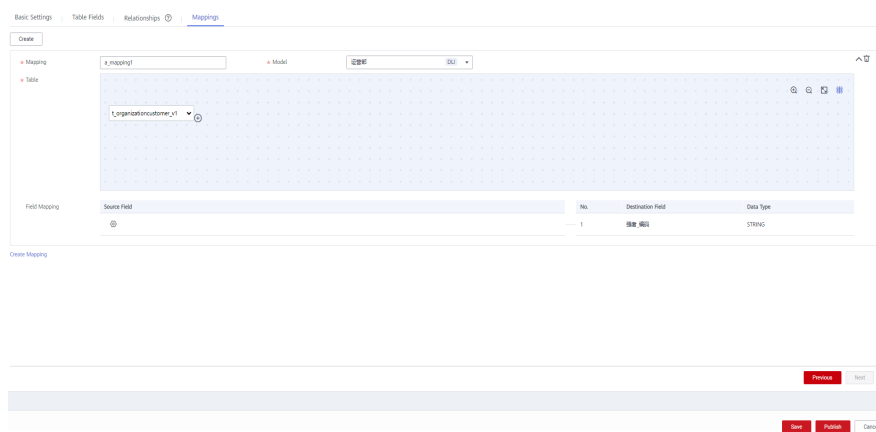




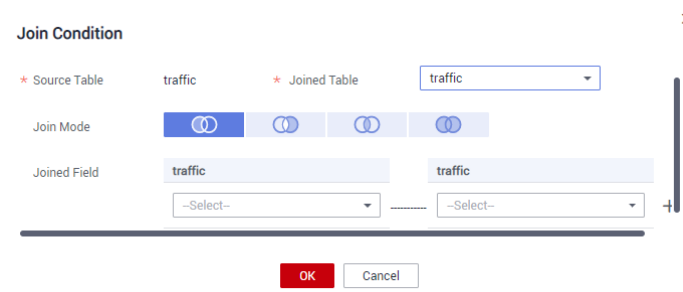




Table 5-28 Parameters of mappings

Parameter	Description
Mapping	Only letters, numbers, and underscores (_) are allowed.
Model	Select a created relationship model from the drop-down list box. If no relationship model has been created, create one. See Designing Physical Models .
Table	<p>Select a table from which data is obtained. If data is obtained from multiple tables, click  next to the table name to set the JOIN condition between the table and other tables.</p> <ol style="list-style-type: none"> 1. Select a JOIN mode. The JOIN mode includes left JOIN, right JOIN, inner JOIN, and outer JOIN from left to right. 2. Set the JOIN condition in the JOIN field. Generally, select the fields with the same meaning in the source table and joined table. Click  or  to add or delete a JOIN condition. The relationship between JOIN conditions is AND. 3. Click OK. 4. If you want to delete a joined table after setting the JOIN condition, click  next to the table name. <p>Figure 5-76 Join Condition dialog box</p> 

Parameter	Description
Field Mapping	Select a source field with the same meaning as the current mapping field. If a table field comes from multiple models, you must create multiple mappings. In each mapping, you only need to set the source field for the field that comes from the current mapping. Other fields do not need to be set.

In the upper right corner of the **Mappings** area, click  to delete a mapping or click  to collapse the mapping area.

- (Optional) If the type of the new table is **DWS_VIEW**, click **Create** to create a view.

Figure 5-77 Creating a view

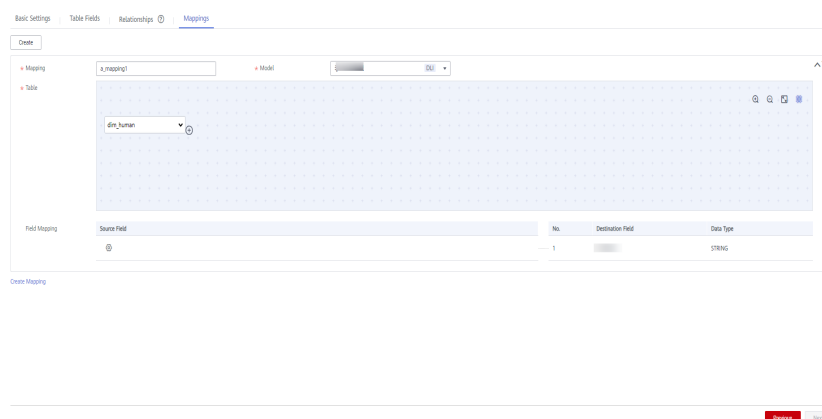




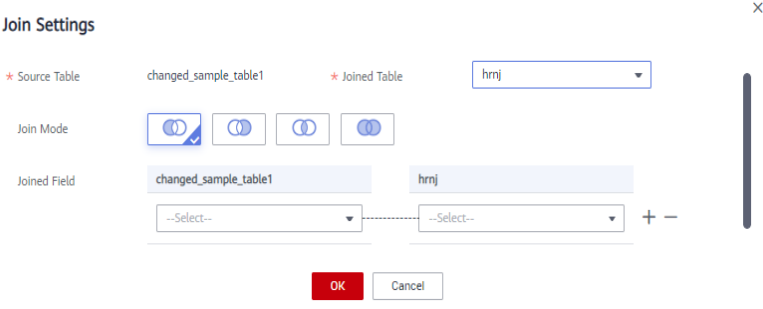




Table 5-29 Parameters


Parameter	Description
Mapping	Only letters, numbers, and underscores (_) are allowed.

Parameter	Description
Table	<p>Select a table from which data is obtained. If data is obtained from multiple tables, click  next to the table name to set the JOIN condition between the table and other tables.</p> <ol style="list-style-type: none"> 1. Select a JOIN mode. The JOIN mode includes left JOIN, right JOIN, inner JOIN, and outer JOIN from left to right. 2. Set the JOIN condition in the JOIN field. Generally, select the fields with the same meaning in the source table and joined table. Click  or  to add or delete a JOIN condition. The relationship between JOIN conditions is AND. 3. Click OK. 4. If you want to delete a joined table after setting the JOIN condition, click  next to the table name. <p>Figure 5-78 Join Settings dialog box</p> 
Field Mapping	<p>Select a source field with the same meaning as the current mapping field. If a table field comes from multiple models, you must create multiple mappings. In each mapping, you only need to set the source field for the field that comes from the current mapping. Other fields do not need to be set.</p>


In the upper right corner of the **Mappings** area, click  to delete a mapping or click  to collapse the mapping area.

Step 4 Click **Publish**, select a reviewer, and click **Submit**.

Step 5 Wait for the reviewer to approve the application. After the application is approved, return to the **ER Modeling** page to view the table status and synchronization status.

Publishing is an asynchronous operation. You can click  to refresh the status. After table publishing application is approved, the system performs operations such as creating tables and synchronizing technical assets and business assets

based on the configurations of **Model Design Process** on the **Function Settings** tab page in **Configuration Center**. The synchronization status is displayed in the **Sync Status** column of the table on the **Information Architecture** page.

- If the synchronization is successful, the table is successfully published. Move the cursor to  in the **Sync Status** column. If the message indicating "creation succeeded" is displayed, the table has been successfully created in the corresponding data source.
- If one or more items fail to be synchronized, you can refresh the status. If the fault persists, choose **More > View History** and click the **Publish Log** tab to view logs.
Troubleshoot the problem based on the logs. After the error is rectified, click **Resynchronize** on the **History** tab page to issue the synchronization command again. If the synchronization still fails, contact technical support for assistance.

----End

5.6.2 Dimensional Modeling

5.6.2.1 Creating Dimensions

A dimension is the perspective to observe and analyze business data and assist in data aggregation, drilling, slicing, and analysis, and used as a GROUP BY condition in SQL statements. Most dimensions have hierarchical structures, such as geographic dimensions (including countries, regions, provinces/states, and cities) and time dimensions (including annually, quarterly, and monthly dimensions). Creating a dimension is a way to standardize the existence and uniqueness of business entities (also called primary data) from the top down.

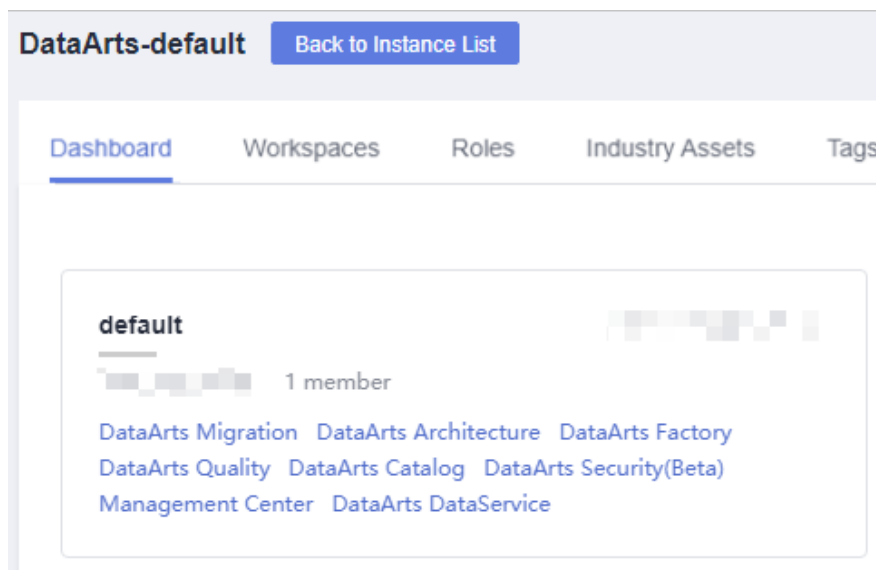
Impact on the System

After a dimension is published and approved, the system automatically creates a dimension table corresponding to the dimension. The name and code of the dimension table are the same as those of the dimension.

Creating and Publishing a Dimension

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

Figure 5-79 DataArts Architecture



2. On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane. On the displayed page, click the **Dimensions** tab.
3. Select an object from the subject directory on the left and click **Create**.
Before creating a dimension, ensure that a subject is available. For details on how to add a subject, see [Designing Subjects](#).
4. On the page displayed, set the parameters.
Set the basic settings and physicalization settings as described below.

Figure 5-80 Dimension parameters

Basic Settings

* Subject:

* Dimension Name: * Dimension English Name:

* Type: Basic Lookup table Hierarchy

* Owner: C

* Description: 4/600

Physicalization Settings

* Data Connection Type: * Data Connection Name: C

* Database:

Table 5-30 Parameters in the Basic Settings area

Parameter	Description
Subject	Select a subject from the drop-down list box.
Name	It must start with letters. Only letters, digits, and the following special characters are allowed: ()-_

Parameter	Description
Dimension Code	Only letters, numbers, and underscores (_) are allowed. Dimension codes must start with letters.
Type	<ul style="list-style-type: none">• Basic: a dimension that does not have a hierarchical structure.• Lookup Table: a dimension created based on a lookup table. The field information and data of the dimension are the same as those of the lookup table, indicating that the content is an enumerable dimension.• Hierarchy: a dimension with a hierarchical structure between attributes.
Owner	You can enter an owner name or select an existing owner.
Description	A description of the dimension to create. Up to 600 characters are supported.

Table 5-31 Parameters in the Physicalization Settings area

Parameter	Description
Data Connection Type	Select a data connection type from the drop-down list box.
Data Connection	The name of the data connection. Select the required data connection. If no data connection is available, access Management Center to create one. For details, see Creating Data Connections .
Database	The name of the database. Select a database from the drop-down list box. If no database is available, access DataArts Factory to create one. For details, see Creating a Database .
Queue	DLI queue. This parameter is displayed only for DLI data connections.
Schema	DWS or POSTGRES SQL mode. This parameter is displayed only for DWS and POSTGRES SQL data connections.
Table Type	DWS table type. Available values include: <ul style="list-style-type: none">• DWS_ROW: Tables are stored to disk partitions by row.• DWS_COLUMN: Tables are stored to disk partitions by column. The MRS_HIVE model supports only HIVE_TABLE .

Parameter	Description
Distributed By	<p>This parameter is displayed only for DWS data connections. You can select multiple fields.</p> <ul style="list-style-type: none"> REPLICATION: A full table is stored on each DN. The advantage of this option is that each DN has all the data of a table. During the join operation, data redistribution can be avoided, reducing network overhead. The disadvantage is that each DN retains the complete data of a table, resulting in data redundancy. Generally, this option is recommended for small dimension tables. HASH: If you select this option, you must specify a distribution key for the user table. When a record is inserted, the system performs hash computing based on values in the distribute keys and then stores data on the corresponding DN. In a Hash table, I/O resources on each node can be used during data read/write, which improves the read/write speed of a table. Generally, this option is recommended for large dimension tables (a large dimension table contains over 1 million records).

Add dimension fields in the **Attribute Settings** area. You can click **Add** to add multiple dimension fields.

Figure 5-81 Field configuration

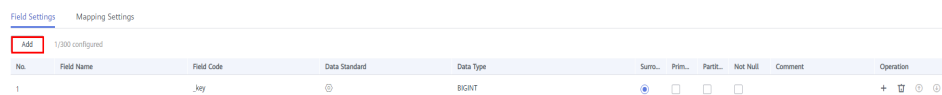



Table 5-32 Parameters in the Attribute Settings area

Parameter	Description
Name	It must start with letters. Only letters, digits, and the following special characters are allowed: ()-_ _
Code	Field codes must start with letters. Only letters, numbers, and underscores (_) are allowed.

Parameter	Description
Data Standard	<p>Click  to select a data standard to be associated with the field. If Create Data Quality Jobs is selected for Model Design Process on the Function Settings tab page in Configuration Center and a field is associated with a data standard, a quality job is automatically generated after a dimension is published. A quality rule is generated for each field associated with the data standard. The quality of the field is monitored based on the data standard. You can access the Quality Job page of DataArts Quality to view the job details.</p> <p>If no data standard is available, create one. See Creating Data Standards for details.</p>
Data Type	Type of data defined based on the original data.
Surrogate Key	Select a field as the surrogate key based on project requirements. By default, the first dimension attribute is the surrogate key.
Primary Key	Select a field as the primary key based on project requirements.
Partition	Whether to be set as a partition field.
Not Null	Whether the parameter value can be left empty.
Description	A description of the dimension field you add.

On the **Mapping Settings** tab page, click **Create** to create the mapping between dimensions and fact tables. Set the parameters.

Figure 5-82 Mapping settings

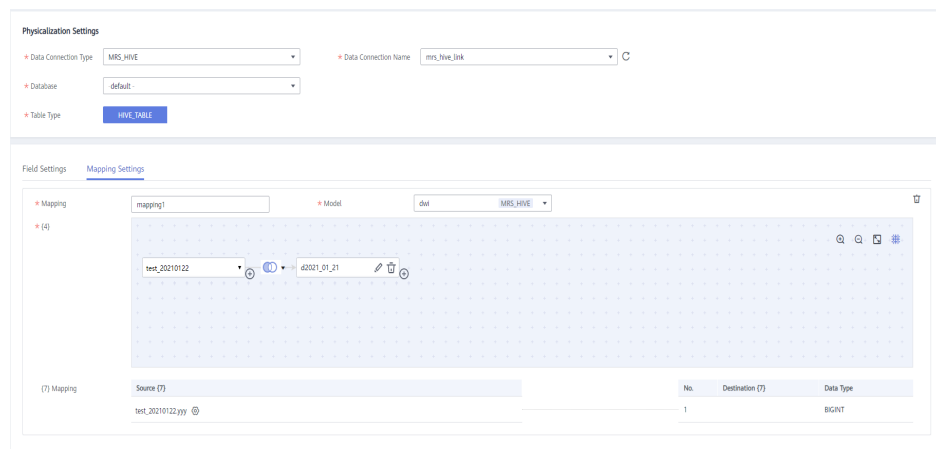




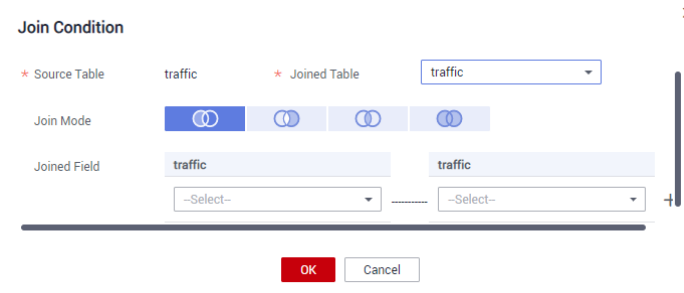




Table 5-33 Parameters of mappings

Parameter	Description
Mapping	Only letters, numbers, and underscores (_) are allowed.
Model	Select a created relationship model from the drop-down list box. If no relationship model has been created, create one. See Designing Physical Models .
Table	<p>Select a table from which data is obtained. If data is obtained from multiple tables, click  next to the table name to set the JOIN condition between the table and other tables.</p> <ol style="list-style-type: none"> 1. Select a JOIN mode. The JOIN mode includes left JOIN, right JOIN, inner JOIN, and outer JOIN from left to right. 2. Set the JOIN condition in the JOIN field. Generally, select the fields with the same meaning in the source table and joined table. Click  or  to add or delete a JOIN condition. The relationship between JOIN conditions is AND. 3. Click OK. 4. If you want to delete a joined table after setting the JOIN condition, click  next to the table name. <p>Figure 5-83 Join Condition dialog box</p> 
Field Mapping	Select a source field with the same meaning as the current mapping field. If a table field comes from multiple models, you must create multiple mappings. In each mapping, you only need to set the source field for the field that comes from the current mapping. Other fields do not need to be set.

In the upper right corner of the **Mappings** area, click  to delete a mapping or click  to collapse the mapping area.

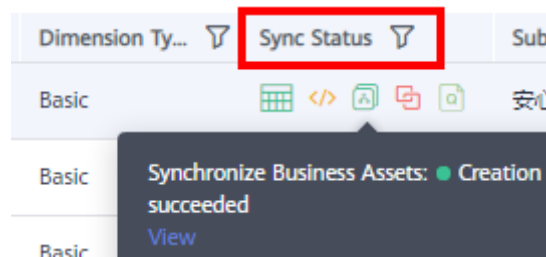
5. Click **Publish**.
6. In the dialog box displayed, select a reviewer and click **OK**.

7. Repeat 3 to 6 to create and publish other dimensions.
8. All dimensions must be approved by reviewers.

After the application is approved, the system automatically creates a dimension table corresponding to the dimension. The name and code of the dimension table are the same as those of the dimension. On the **Dimensional Modeling** page, click the **Dimension Tables** tab to view the created dimension table.

In the dimension table list, you can view the synchronization status of the dimension table in the **Sync Status** column.

Figure 5-84 Sync Status of the dimension table

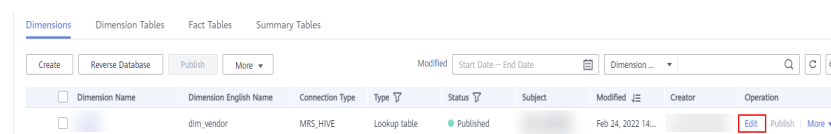


- If the synchronization is successful, the dimension is successfully published and the dimension table is successfully created in the database.
- If the synchronization failed, click **View History** in the row where the dimension table is located. On the page displayed, click the **History** tab to view logs. Troubleshoot the problem based on the logs. After the error is rectified, go back to the dimension table list and click **Synchronize** above the dimension table list to issue the synchronization command again. If the problem persists, contact technical support personnel.

Editing a Dimension

- Step 1** On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane. On the displayed page, click the **Dimensions** tab.
- Step 2** In the dimension list, select the target dimension and click **Edit** in the **Operation** column.

Figure 5-85 Editing a dimension



- Step 3** Edit the dimension information based on service requirements. For details about how to set parameters, see [Dimension parameters](#).
- Step 4** Click **Save**. Alternatively, click **Publish** to publish the edited dimension.

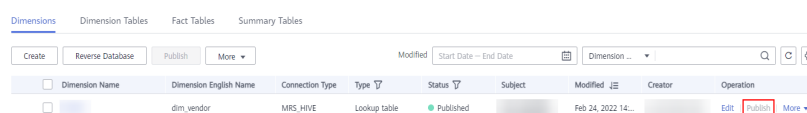
----End

Publishing a Dimension

If a dimension is created but not published, perform the following steps to publish the dimension:

- Step 1** On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane. On the displayed page, click the **Dimensions** tab.
- Step 2** In the dimension list, find the target dimension and click **Publish** in the **Operation** column.

Figure 5-86 Publishing a dimension



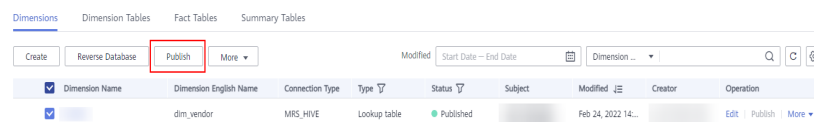
- Step 3** In the dialog box displayed, select a reviewer and click **OK**.

----End

You can also perform the following steps to publish multiple dimensions:

- Step 1** On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane. On the displayed page, click the **Dimensions** tab.
- Step 2** Select the dimensions you want to publish and click **Publish** above the dimension list.

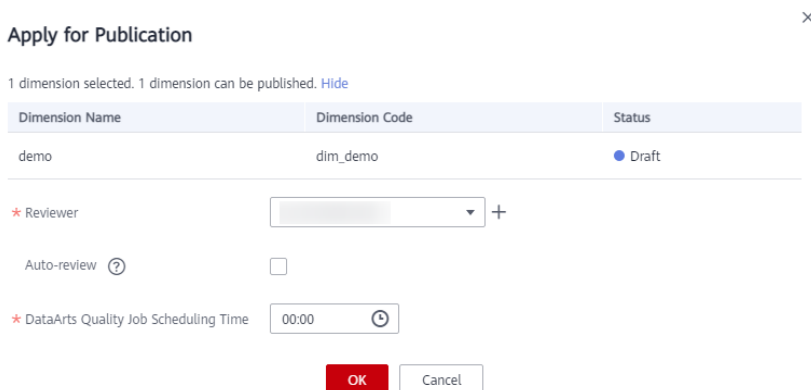
Figure 5-87 Publishing Tables multiple dimensions



- Step 3** In the displayed dialog box, select a reviewer, set **Job Scheduling Time**, and click **OK**.

Job Scheduling Time refers to the scheduling time for automatic quality job creation after the dimension is published.

Figure 5-88 Publishing multiple dimensions



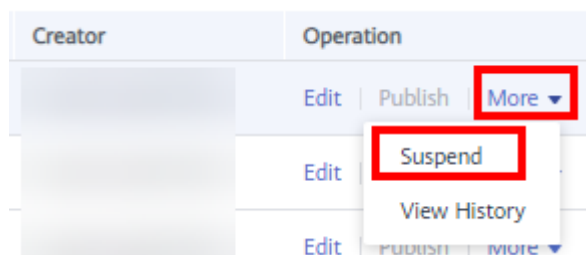
----End

Suspending a Dimension

To suspend a published dimension, perform the following steps:

- Step 1** On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane. On the displayed page, click the **Dimensions** tab.
- Step 2** In the dimension list, find the target dimension and choose **More > Suspend** in the **Operation** column.

Figure 5-89 Suspending a dimension



- Step 3** In the dialog box displayed, select a reviewer and click **OK**. The dimension is suspended after the reviewer approves it.

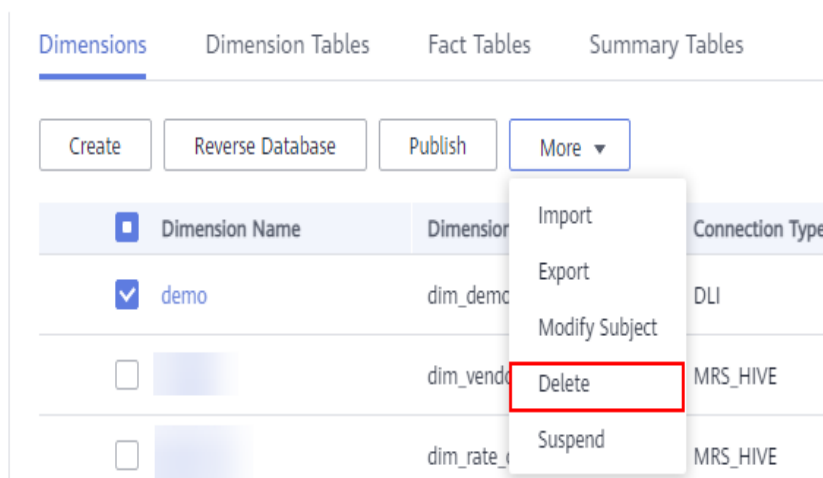
----End

Deleting a Dimension

If a dimension is no longer needed, you can delete it. However, if the dimension has been published, you must suspend the dimension before deleting it. For details, see [Suspending a Dimension](#).

- Step 1** On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane. On the displayed page, click the **Dimensions** tab.
- Step 2** In the dimension list, find the target dimension and choose **More > Delete** above the list.

Figure 5-90 Deleting a dimension



Step 3 In the **Delete Dimension** dialog box, confirm the information and click **Yes**.

If you select **Delete physical tables** in the dialog box, the physical tables in the database are also deleted when you delete the dimension.

----End

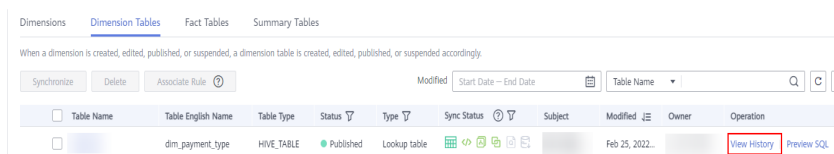
5.6.2.2 Managing Dimension Tables

A dimension table corresponds to a dimension and consists of a wide range of dimension fields. Creating, publishing, editing, and suspending a dimension table highly relate to the corresponding dimension. After a dimension is published, the system automatically creates and publishes the corresponding dimension table.

Viewing the Publish History of a Dimension Table

1. On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane.
2. Click the **Dimension Tables** tab.
3. Select a dimension table in the list and click **View History** in the **Operation** column.

Figure 5-91 Dimension Tables tab page



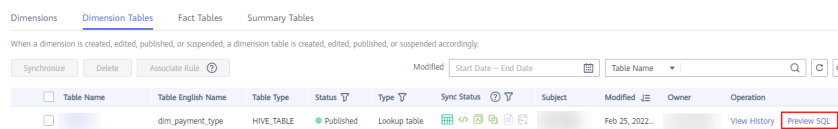
4. On the page displayed, you can view the publish history, version comparison information, and publish log of the dimension table.

If the publish log includes error logs, the publishing has failed. You can click **Resynchronize** to synchronize the table to other DataArts Studio modules.

Previewing SQL

1. On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane.
2. Click the **Dimension Tables** tab.
3. Select a dimension table in the list and click **Preview SQL** in the **Operation** column.

Figure 5-92 Previewing an SQL statement



4. On the page displayed, you can view or copy the SQL statement.

Synchronizing a Dimension Table

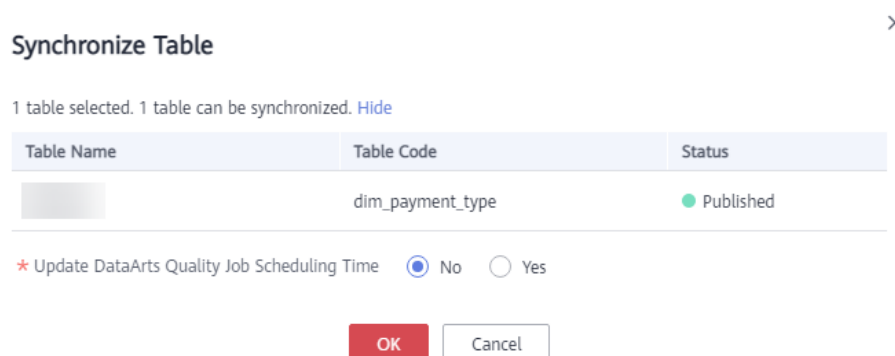
After you create or edit a dimension, you can manually synchronize the dimension table if the synchronization fails.

NOTE

- The system performs the synchronization based on the data table update mode on the **Function Settings** tab page of **Configuration Center**. For details, see [Functions](#).


1. On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane.
2. Click the **Dimension Tables** tab.
3. In the dimension table list, select the target dimension table and click **Synchronize** above the list. The dialog box for synchronizing the dimension table is displayed.

Figure 5-93 Synchronizing dimension tables



4. After confirming that the information is correct, click **OK**. The synchronization result is displayed.

After the synchronization, you can view the synchronization status of the

dimension table in the dimension table list. You can also click  above the list to refresh the status.

Associating a Dimension Table with a Quality Rule

1. On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane.
2. Click the **Dimension Tables** tab.
3. In the dimension table list, select the target dimension table, and click **Associate Rule**.

Figure 5-94 Associating a dimension table with a quality rule



4. On the page displayed, set the parameters. After the configuration is complete, click **OK**.
 - **Update Existing Rule:** If this option is selected, the newly added rule will overwrite the old rule.
 - **Table Field:** This parameter applies to all fields by default. You can enter a regular expression to filter fields as required.
 - **WHERE Clause:** This parameter can be used to filter fields.
 - **Generate Anomaly Data:** If this option is enabled, anomaly data is stored in the specified database based on the configured parameters.
 - **Database/Schema:** database or schema that stores anomaly data. This parameter is displayed when **Generate Anomaly Data** is enabled.
 - **Table Prefix:** prefix of the table that stores anomaly data. This parameter is displayed when **Generate Anomaly Data** is enabled.
 - **Table Suffix:** suffix of the table that stores anomaly data. This parameter is displayed when **Generate Anomaly Data** is enabled.
 - **Add Rule:** You can click **Add Rule** to add a rule. For example, add a rule named **Unique value**, select the rule, click **OK**, enter an alarm condition expression in the **Alarm Condition** text box, add other rules in the same way, and click **OK**.
 - An alarm condition expression consists of alarm parameters and logical operators. When a quality job is running, the system calculates the result of the alarm condition expression and determines whether to trigger the alarm based on the result of the expression. If the expression result is **true**, the alarm will be triggered. Otherwise, no quality alarm will be triggered. In the **Associate Quality Rule** dialog box, the alarm parameters of each quality rule are displayed as buttons.

Associating a Single Field with a Quality Rule


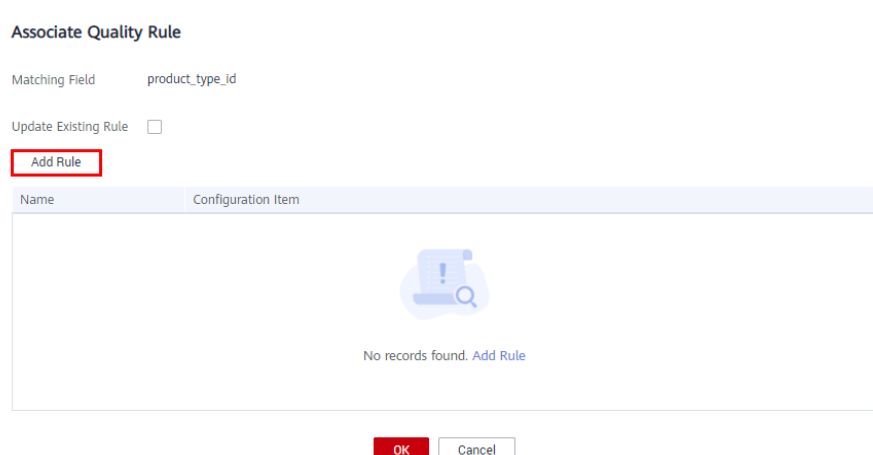
1. On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane.
2. Click the **Dimension Tables** tab.
3. In the dimension table list, click the name of the target dimension table.
4. In the field list on the dimension table details page, click  in the row of the target field to associate the field with a quality rule.

Figure 5-95 Associating a single field with a quality rule



5. After the configuration is complete, click **OK**.
 - **Update Existing Rule:** If this option is selected, the newly added rule will overwrite the old rule.
 - **Add Rule:** You can click **Add Rule** to add a rule. For example, add a rule named **Unique value**, select the rule, click **OK**, enter an alarm condition expression in the **Alarm Condition** text box, add other rules in the same way, and click **OK**.
 - An alarm condition expression consists of alarm parameters and logical operators. When a quality job is running, the system calculates the result of the alarm condition expression and determines whether to trigger the alarm based on the result of the expression. If the expression result is **true**, the alarm will be triggered. Otherwise, no quality alarm will be triggered. In the **Associate Quality Rule** dialog box, the alarm parameters of each quality rule are displayed as buttons.

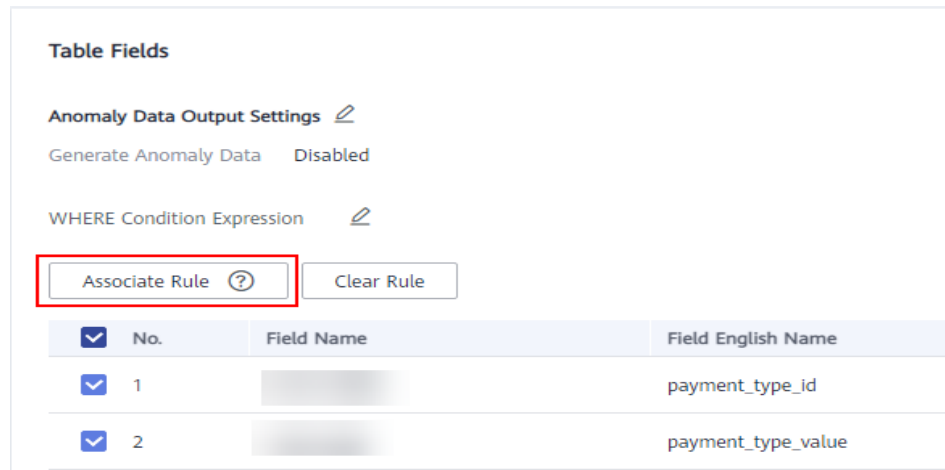
Figure 5-96 Adding a rule



Associating Table Fields with a Quality Rule in Batches

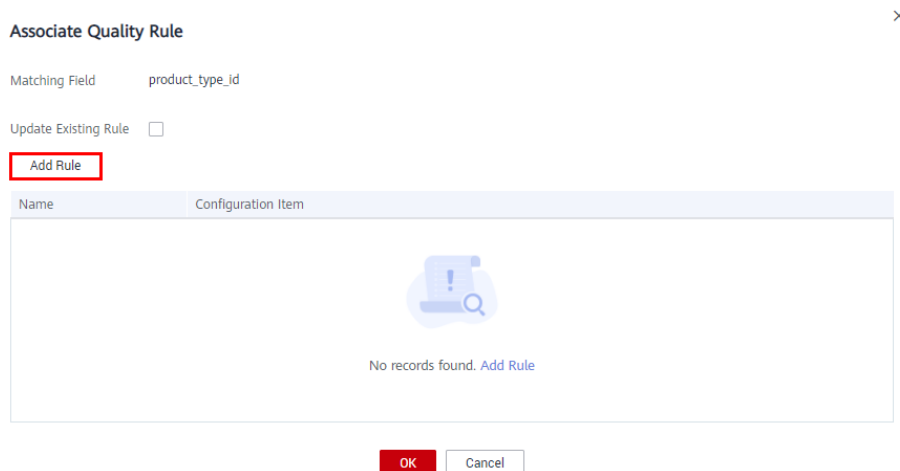
1. On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane.
2. Click the **Dimension Tables** tab.
3. In the dimension table list, click the name of the target dimension table.
4. In the table field list on the dimension table details page, select the target table fields and click **Associate Rule**.

Figure 5-97 Associating table fields with a quality rule



5. On the page displayed, add a rule and set the rule parameters.
 - **Update Existing Rule:** If this option is selected, the newly added rule will overwrite the old rule.
 - **Add Rule:** You can click **Add Rule** to add a rule. For example, add a rule named **Unique value**, select the rule, click **OK**, enter an alarm condition expression in the **Alarm Condition** text box, add other rules in the same way, and click **OK**.
 - An alarm condition expression consists of alarm parameters and logical operators. When a quality job is running, the system calculates the result of the alarm condition expression and determines whether to trigger the alarm based on the result of the expression. If the expression result is **true**, the alarm will be triggered. Otherwise, no quality alarm will be triggered. In the **Associate Quality Rule** dialog box, the alarm parameters of each quality rule are displayed as buttons.

Figure 5-98 Associating table fields with a quality rule



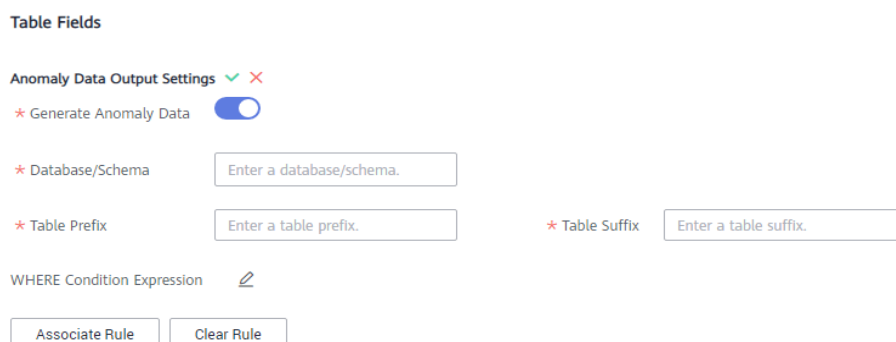
- (Optional) If you want to store anomaly data that does not comply with the preset rules in the exception table, enable **Anomaly Data Output Settings**.

Figure 5-99 Enabling Anomaly Data Output Settings



Click the pen icon next to **Anomaly Data Output Settings** and enable **Generate Anomaly Data**. The anomaly data will be stored in the specified database based on the settings.

Figure 5-100 Anomaly Data Output Settings



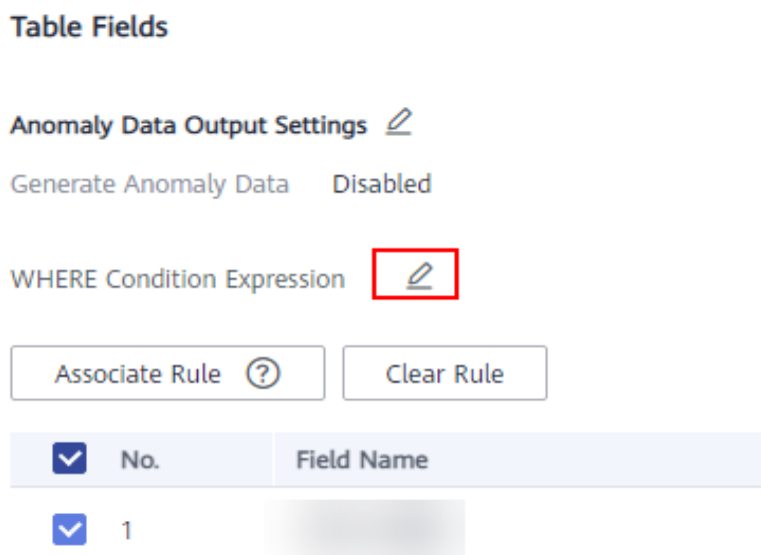
The parameters are as follows:

- **Database/Schema:** database or schema that stores anomaly data
- **Table Prefix:** prefix of the table that stores anomaly data
- **Table Suffix:** suffix of the table that stores anomaly data

Click  to save the settings.

- (Optional) By default, the quality rule applies to the entire table. If you want to query data in specified partitions, set the where condition.

Figure 5-101 Where condition



- After the configuration is complete, click **OK**.

Deleting a Dimension Table

Dimensions to be published, already published, or to be suspended cannot be deleted. You can delete a dimension table on the **Dimensions** page.

- On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane.
- Click the **Dimension Tables** tab.
- In the dimension table list, select the target dimension table and click **Delete** above the list.

Figure 5-102 Deleting a dimension table



- Confirm the dimension table to delete, and click **Yes**.

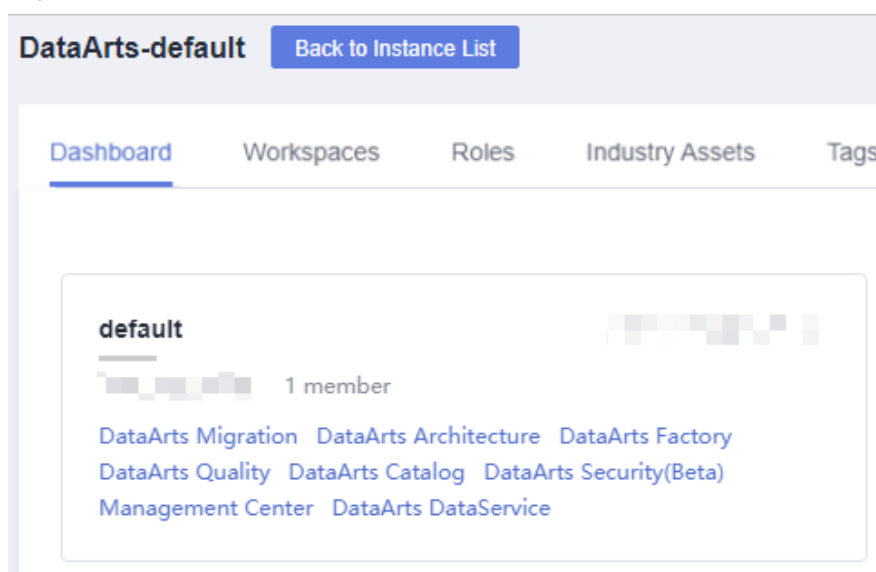
5.6.2.3 Creating Fact Tables

A fact table for a business process can provide a wealth of information about specific business processes. After a fact table is created, the public affair details are accumulated to facilitate data extraction.

Creating and Publishing a Fact Table

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

Figure 5-103 DataArts Architecture



2. On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane. On the displayed page, click the **Fact Tables** tab.
3. Select a subject from the subject tree on the left and click **Create**.
4. On the **Create Fact Table** page, perform the following operations:
 - a. Set the parameters in the **Basic Settings** area.

Figure 5-104 Basic Settings area

The screenshot shows the 'Basic Settings' form for creating a fact table. It includes the following fields:

- * Subject: A dropdown menu with '--Select--'.
- * Table Name: A text input field with the placeholder 'Enter a fact table name.'
- * Table English Name: A text input field with the value 'fact_'.
- * Data Connection Type: A dropdown menu with '--Select--'.
- * Data Connection Name: A dropdown menu with '--Select--' and a 'C' icon.
- * Database: A dropdown menu with '--Select--'.
- * Owner: A text input field with the placeholder 'Enter an asset owner.' and a 'C' icon.
- * Description: A text area with the value 'None' and a character count '4/100'.

Table 5-34 Parameters in the Basic Settings area

Parameter	Description
Subject	Select a subject (business domain group > business domain > business object) where you can place the fact table.
Name	It must start with letters. Only letters, digits, and the following special characters are allowed: ()-_
Table Code	Table codes must start with letters. Only letters, numbers, and underscores (_) are allowed.
Data Connection Type	Select a data connection type from the drop-down list box.
Data Connection	Select a data connection from the drop-down list box. It is recommended that the same data connection be used for dimension modeling.
Database	Select a database from the drop-down list box.
Queue	DLI queue. This parameter is displayed only for DLI data connections.
Schema	DWS or POSTGRESQL mode. This parameter is displayed only for DWS and POSTGRESQL data connections.
Table Type	DWS connections support the following tables: <ul style="list-style-type: none"> • DWS_ROW: Tables are stored to disk partitions by row. • DWS_COLUMN: Tables are stored to disk partitions by column. MRS_HIVE supports only HIVE_TABLE .

Parameter	Description
Distributed By	<p>This parameter is displayed only for DWS data connections. You must add a table field before selecting a table field from the drop-down list as a Distributed By field. Multiple table fields can be selected.</p> <p>Currently, only REPLICATION and HASH are supported.</p> <ul style="list-style-type: none"> • REPLICATION: A full table is stored on each DN. The advantage of this option is that each DN has all the data of a table. During the join operation, data redistribution can be avoided, reducing network overhead. The disadvantage is that each DN retains the complete data of a table, resulting in data redundancy. Generally, this option is recommended for small dimension tables. • HASH: If you select this option, you must specify a distribution key for the user table. When a record is inserted, the system performs hash computing based on values in the distribute keys and then stores data on the corresponding DN. In a Hash table, I/O resources on each node can be used during data read/write, which improves the read/write speed of a table. Generally, this option is recommended for large dimension tables (a large dimension table contains over 1 million records).
Owner	You can enter an owner name or select an existing owner.
Description	A description of the fact table. Up to 600 characters are supported.

- b. On the **Field Settings** page, click **Create** and select **Dimension** or **Measure** to add a dimension or measure field.
- If you select **Dimension**, select one or multiple dimensions in the displayed dialog box and click **OK**
 - If you select **Measure**, set required parameters to add a measure field.



For details about the field parameters, see [Table 5-35](#). After adding a field, you can click  or  to move the field up or down.

Figure 5-105 Adding a dimension or measure field

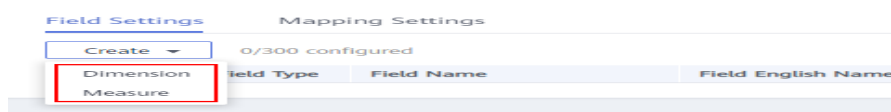




Table 5-35 Field parameters

Parameter	Description
Name	It must start with letters. Only letters, digits, and the following special characters are allowed: ()-_ The surrogate key name of the added dimension field is displayed automatically. Generally, you do not need to change the name.
Field Code	Table codes must start with letters. Only letters, numbers, and underscores (_) are allowed.
Data Type	Data type of the created dimension
Primary Key	If this parameter is selected, the field is a primary key.
Partition	If this parameter is selected, the field is a partition field.
Not Null	Whether the parameter value can be left empty.
Associate Standard	If you have created data standards, click  to select one to associate with the field. If Create Data Quality Jobs is selected for Model Design Process on the Function Settings tab page in Configuration Center and a field is associated with a data standard, a quality job is automatically generated after a table is published. A quality rule is generated for each field associated with the data standard. The quality of the field is monitored based on the data standard. You can access the Quality Job page of DataArts Quality to view the job details. If no data standard is available, create one. See Creating Data Standards for details.
Associate Dimension	Only dimension fields need to be associated with dimensions. The name of the associated dimension. Click  to replace the associated dimension.
Role	Roles need to be assigned to dimension fields which are added for multiple times. This is not required for measure fields. If a dimension is added multiple times, set different roles to distinguish the dimensions.
Description	A description of the dimension.

- c. On the **Mapping Settings** tab page, click **Create Mapping** and set mapping parameters.

Figure 5-106 Configuring mapping parameters

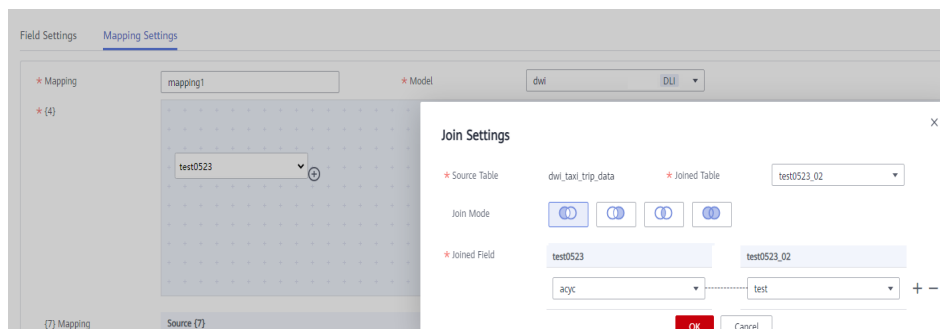




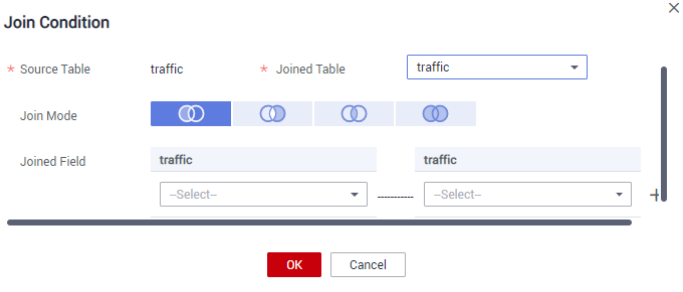


Table 5-36 Parameters of mappings

Parameter	Description
Mapping	Only letters, numbers, and underscores (_) are allowed.
Model	Select a created relationship model from the drop-down list box. If no relationship model has been created, create one. See Designing Physical Models .
Table	<p>Select a table from which data is obtained. If data is obtained from multiple tables, click  next to the table name to set the JOIN condition between the table and other tables.</p> <ol style="list-style-type: none"> Select a JOIN mode. The JOIN mode includes left JOIN, right JOIN, inner JOIN, and outer JOIN from left to right. Set the JOIN condition in the JOIN field. Generally, select the fields with the same meaning in the source table and joined table. Click  or  to add or delete a JOIN condition. The relationship between JOIN conditions is AND. Click OK. If you want to delete a joined table after setting the JOIN condition, click  next to the table name. <p>Figure 5-107 Join Condition dialog box</p> 

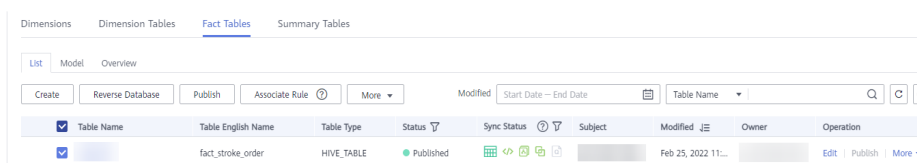
Parameter	Description
Field Mapping	Select a source field with the same meaning as the current mapping field. If a table field comes from multiple models, you must create multiple mappings. In each mapping, you only need to set the source field for the field that comes from the current mapping. Other fields do not need to be set.

5. Click **Publish**.
6. Wait for the reviewer to approve the fact table.
After the fact table is approved, it is automatically created in the database.
7. Go back to the fact table list and locate the table just published. View its synchronization status in the **Sync Status** column.
 - If the synchronization is successful, the fact table is successfully published and created in the database.
 - If the synchronization failed, choose **More > View History** in the row where the fact table is located. On the page displayed, click the **History** tab to view logs. Troubleshoot the problem based on the logs. After the error is rectified, choose **More > Synchronize** above the fact table list to issue the synchronization command again. If the problem persists, contact technical support personnel.

Managing a Fact Table

After a fact table is created, you can access the **Fact Tables** page of **Dimensional Modeling** in DataArts Architecture. On the page displayed, you can edit, publish, suspend, and delete the fact table, as well as view the publish logs.

Figure 5-108 Fact table management



- **Editing a fact table**
 - a. In the fact table list, select a fact table and click **Edit** to the right of it. The page for editing the fact table is displayed.
 - b. Edit the table as required.
 - c. Click **Save** to save the settings, or click **Publish** to publish the settings.
- **Publishing a fact table**
 - a. In the fact table list, select a fact table and click **Publish**. The dialog box for publishing the fact table is displayed.
 - b. Select a reviewer from the drop-down list box.

- c. Click **OK**.
- **Viewing the publish history**
 - a. Select a fact table in the list and choose **More > View History** on the right.
 - b. On the page displayed, you can view the publish history and version comparison information of the fact table.
If the publish log includes error logs, the publishing has failed. You can click **Resynchronize** to synchronize the table to other DataArts Studio modules.
- **Associating a fact table with a quality rule**
 - a. Select a fact table in the fact table list and click **Associate Rule** above the list.
 - b. In the **Associate Quality Rule** dialog box, you can add rules to the fields in the fact table in batches and associate the rules with the fields.
 - c. Click **OK**.
- **Previewing an SQL statement**
 - a. Select a fact table in the list and choose **More > Preview SQL** on the right.
 - b. On the page displayed, you can view or copy the SQL statement.
- **Suspending a fact table**
 - a. In the fact table list, select a fact table and click **Suspend**. The dialog box for suspending a fact table is displayed.
 - b. Select a reviewer from the drop-down list box.
 - c. Click **OK**.

 **NOTE**

- You can suspend or delete a fact table only when it is not referenced. For example, a fact table can be deleted only when it is not used by atomic metrics.
- **Deleting a fact table**

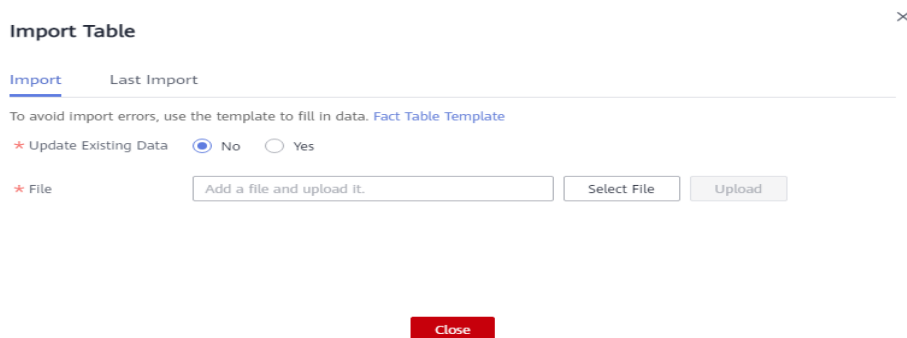
If you no longer need a fact table, you can delete it. Fact tables that are to be published, already published, or to be suspended cannot be deleted.

 - a. In the fact table list, select a fact table and choose **More > Delete** above the list.
 - b. In the dialog box displayed, click **Yes**.
- **Importing fact tables**

You can import fact tables to the system quickly.

 - a. Above the fact table list, choose **More > Import**.

Figure 5-109 Import Table dialog box



- b. Download the fact table template, and edit and save it.
- c. Choose whether to update existing data.

NOTE

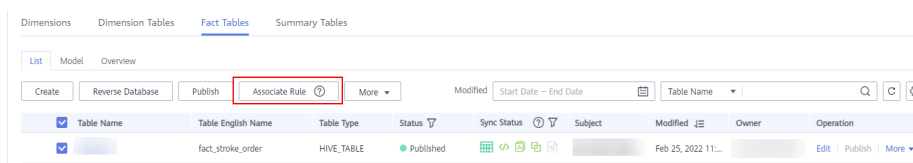
If a code in the template already exists in the system, the data is considered duplicate.

- **No:** If the data to be imported already exists in the system, the existing data in the system will not be replaced.
 - **Yes:** If the data to be imported already exists in the system:
 - If the existing data in the system is in draft state, the data will be replaced and new draft data will be generated.
 - If the existing data in the system is in published state, expanded data will be generated.
- d. Click **Select File** and select the edited template to import.
 - e. Click **Upload**. When the template is uploaded, the **Last Import** page is displayed. You can view the imported data.
 - f. Click **Close**.
- **Exporting fact tables**
Above the fact table list, choose **More > Export** to export fact tables.

Associating a Fact Table with a Quality Rule

1. On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane.
2. Click the **Fact Tables** tab.
3. In the fact table list, click the name of the target fact table. Click **Associate Rule**.

Figure 5-110 Associating a fact table with a quality rule



4. On the page displayed, set the parameters. After the configuration is complete, click **OK**.
 - **Update Existing Rule:** If this option is selected, the newly added rule will overwrite the old rule.
 - **Table Field:** This parameter applies to all fields by default. You can enter a regular expression to filter fields as required.
 - **WHERE Clause:** This parameter can be used to filter fields.
 - **Generate Anomaly Data:** If this option is enabled, anomaly data is stored in the specified database based on the configured parameters.
 - **Database/Schema:** database or schema that stores anomaly data. This parameter is displayed when **Generate Anomaly Data** is enabled.
 - **Table Prefix:** prefix of the table that stores anomaly data. This parameter is displayed when **Generate Anomaly Data** is enabled.
 - **Table Prefix:** prefix of the table that stores anomaly data. This parameter is displayed when **Generate Anomaly Data** is enabled.
 - **Add Rule:** You can click **Add Rule** to add a rule. For example, add a rule named **Unique value**, select the rule, click **OK**, enter an alarm condition expression in the **Alarm Condition** text box, add other rules in the same way, and click **OK**.
 - An alarm condition expression consists of alarm parameters and logical operators. When a quality job is running, the system calculates the result of the alarm condition expression and determines whether to trigger the alarm based on the result of the expression. If the expression result is **true**, the alarm will be triggered. Otherwise, no quality alarm will be triggered. In the **Associate Quality Rule** dialog box, the alarm parameters of each quality rule are displayed as buttons.

Figure 5-111 Associating a fact table with a quality rule

Associate Rule

Selected Table All Searched Tables (6)

Update Existing Rule

Table Field

WHERE Condition Expression

* Generate Anomaly Data

* Database/Schema

* Table Prefix * Table Suffix

Add Rule

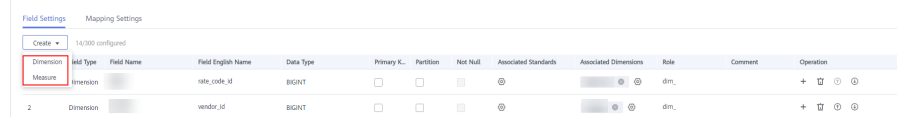
OK Cancel

Creating a Field in the Fact Table

1. On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane.

2. Click the **Fact Tables** tab.
3. In the fact table list, click the name of the target table and click **Edit** in the **Operation** column.
4. Click **Create** in the **Table Fields** area, select a new field type from the drop-down list, and set the related parameters.

Figure 5-112 Creating a field

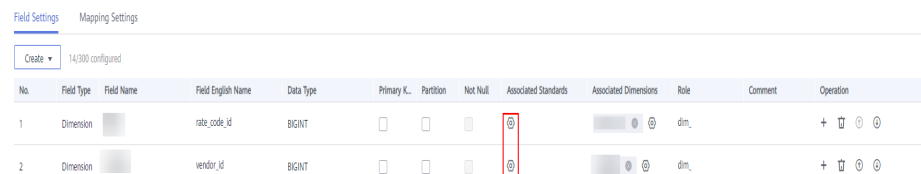


5. After the configuration is complete, click **OK**.

Associating a Fact Table Field with a Data Standard

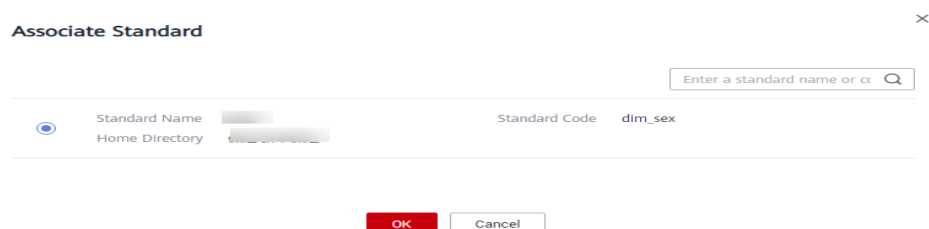
1. On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane.
2. Click the **Fact Tables** tab.
3. Click the name of the target fact table in the list.
4. In the table field list on the details page of the fact table, search for the target field, click corresponding to the field to configure the association between the field and the data standard. For details on the sources of data standards, see [Creating a Data Standard](#).

Figure 5-113 Associating a fact table field with a data standard



5. After the configuration is complete, click **OK**.

Figure 5-114 Associating a data standard



Associating a Fact Table Field with a Quality Rule

1. On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane.
2. Click the **Fact Tables** tab.
3. In the fact table list, click the name of the target fact table.


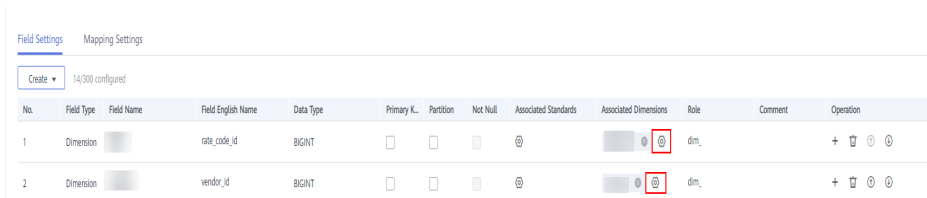
- In the table field list on the fact table details page, locate the target field and click  to associate the field with a quality rule.

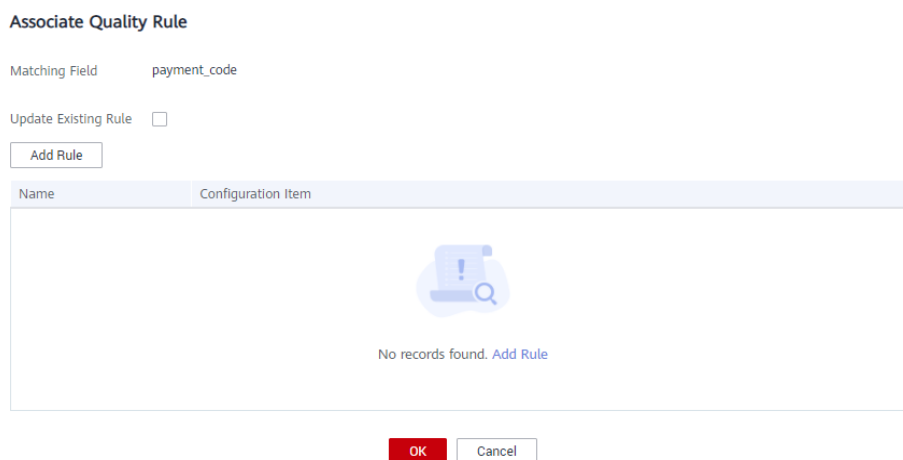
Figure 5-115 Associating a fact table field with a quality rule



No.	Field Type	Field Name	Field English Name	Data Type	Primary K.	Partition	Not Null	Associated Standards	Associated Dimensions	Role	Comment	Operation
1	Dimension		rate_code_id	BIGINT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>			dm_		+
2	Dimension		vendor_id	BIGINT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>			dm_		+

- After the configuration is complete, click **OK**.


Figure 5-116 Adding a rule



Associate Quality Rule

Matching Field:

Update Existing Rule:

Name	Configuration Item
 No records found. Add Rule	

Associating Fact Table Fields with a Quality Rule in Batches

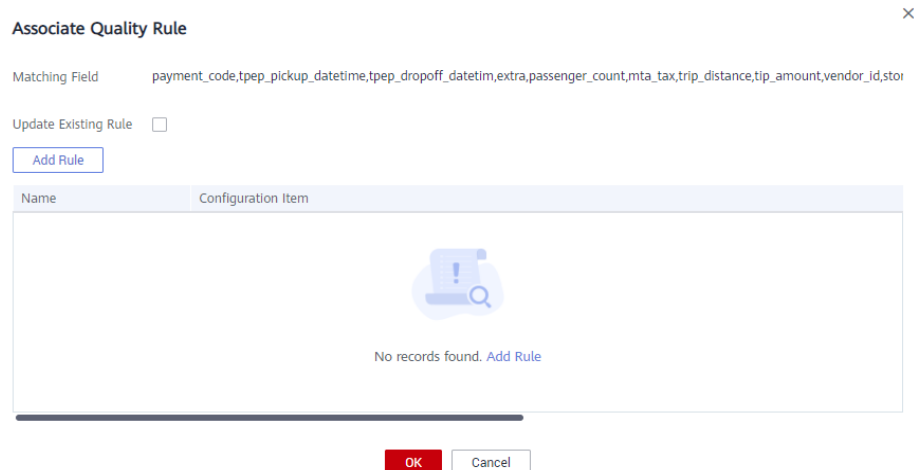
- On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane.
- Click the **Fact Tables** tab.
- In the fact table list, click the name of the target fact table.
- In the table field list on the fact table details page, select the target table fields and click **Associate Rule**.

Figure 5-117 Associating fact table fields with a quality rule



No.	Field Name	Field English Name	Data Type	Primary Key	Partition	Not Null	Associated Standards	Associated Rules	Associated Dimen...	Role	Comment
1		rate_code_id	BIGINT	N	N	N					
2		vendor_id	BIGINT	N	N	N					

- On the page displayed, add a rule and set the rule parameters.

Figure 5-118 Adding a rule

6. After the configuration is complete, click **OK**.

5.7 Metric Design

5.7.1 Business Metrics

After data survey and requirement analysis, you must implement metrics. A metric is a statistical value that measures the overall characteristic of a target and reflects the business situation in a business activity of an enterprise. A metric consists of its name and value. The metric name and its definition reflect the quality and quantity of the metric. The metric value reflects the quantifiable values of the specified time, location, and condition of the metric. Business metrics are used to guide technical metrics, and technical metrics are used to implement business metrics.

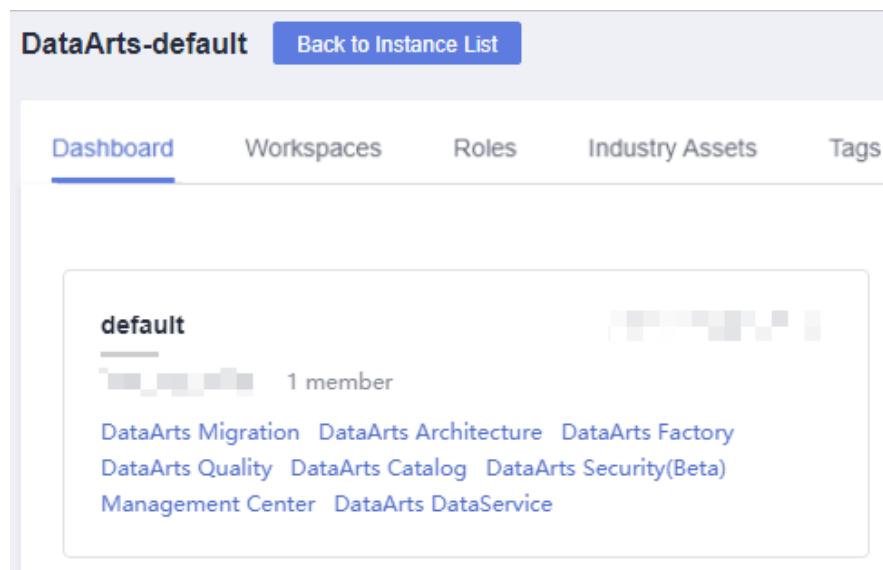
Prerequisites

You have designed a process. For details, see [Designing Processes](#).

Creating and Publishing a Business Metric

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

Figure 5-119 DataArts Architecture



2. On the DataArts Architecture page, choose **Metrics > Business Metrics** in the left navigation pane.
3. In the process tree on the left, select a process and click **Create**.
4. On the page displayed, set the parameters and click **Publish**.
 - a. Configure basic settings.

Figure 5-120 Basic Settings area

Basic Settings

* Metric Name Metric Code The code is generated automatically when you click Save, but you can modify it if needed.

Metric Alias

* Process [Manage Process](#)

* Objective 0/1,000

* Metric Definition 0/1,000

Description 0/600

Table 5-37 Parameters in the Basic Settings area

Parameter	Description
Name	The name of the business metric to create.
Code	<ul style="list-style-type: none"> The metric code is automatically generated. You can configure the generation rule on the Configuration Center page of DataArts Architecture. For details, see Encoding Rules.
Alias	This parameter is optional.
Process	Select the process that the metric belongs to. If no process is available, create one. Refer to Designing Processes for details.
Objective	Your purpose of setting the metric.
Metric Definition	The definition of the metric must be accurately described.
Remarks	Remarks for the metric to create.

b. Configure the metric information.

Figure 5-121 Metric Information area

Metric Settings

* Formula 0/1,000

* Statistical Frequency

Statistical Dimension

Standard & Modifier 0/1,000

* Refresh Frequency

Application Scenario Associated Technical Metrics Type

Associated Technical Metrics Measurement Object

Measurement Unit

Table 5-38 Parameters in the Metric Information area

Parameter	Description
Formula	The computing logic of the business metric, which guides developers to design atomic and derivative metrics. Business metrics are used to guide the implementation of technical metrics only and are not calculated.

Parameter	Description
Frequency	The statistical period of a metric, which helps developers set the time limits.
Statistical Dimension	You can select an existing dimension from the drop-down list For details on how to create a dimension, see Creating Dimensions .
Standard & Modifier	Modifiers are abstract definitions of scenarios and are used to determine the measurement scope.
Refresh Frequency	The interval for updating a metric. Developers or operators can set the scheduling frequency of derivative metrics based on the metric update frequency.
Metric Application Scenario	The application scenarios of the metric.
Associated Technical Metrics Type	Select the type of the technical metric associated with the business metric. Available options include Derivative metric and Compound metric .
Associated Technical Metrics	Select a technical metric associated with the business metric.
Measurement Object	The field for measuring a metric.
Measurement Unit	The measurement unit of a metric.

- c. Configure the management information.

Figure 5-122 Management Information area

Management Information

Data Source: * Metric Mgmt Dept:

* Metric Owner: X

Table 5-39 Parameters in Management Information area

Parameter	Description
Data Source	The generator of data.

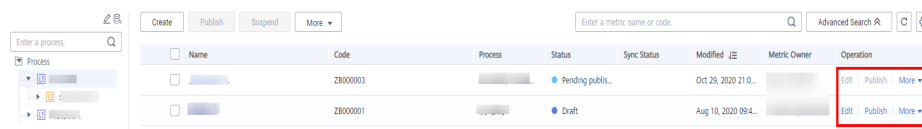
Parameter	Description
Metric Mgmt Dept	The department that manages the metric.
Metric Owner	Metric owner. You can enter the name of an owner or select an existing owner.

5. In the dialog box displayed, select a reviewer and click **OK**.
6. Repeat **3** to **5** to create and publish other business metrics.
7. All the business metrics must be approved by reviewers.
If the applications are approved, the business metrics are created.

Editing a Business Metric

1. On the DataArts Architecture page, choose **Metrics > Business Metrics** in the left navigation pane.

Figure 5-123 Managing business metrics



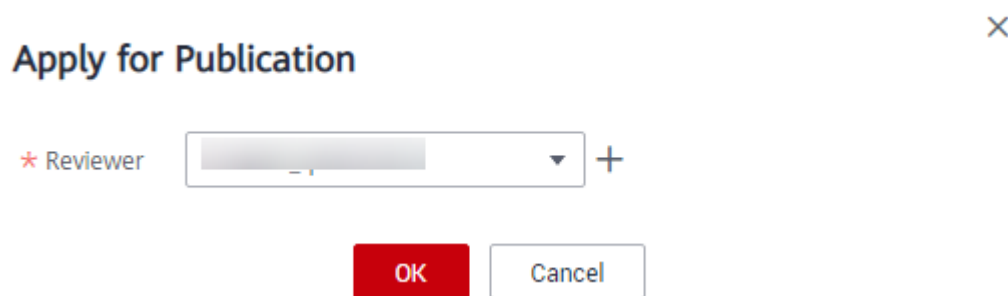
2. In the business metric list, select the target metric and click **Edit** on the right.
3. Edit the business metric information as required.
4. Click **Save** to save the settings. Alternatively, click **Publish** to publish the edited business metric.

Publishing a Business Metric

If a business metric is created but not published, perform the following steps to publish it:

- Step 1** On the DataArts Architecture page, choose **Metrics > Business Metrics** in the left navigation pane.
- Step 2** In the business metric list, select the target metric and click **Publish**.
- Step 3** In the dialog box displayed, select a reviewer and click **OK**.

Figure 5-124 Submit for Publication dialog box



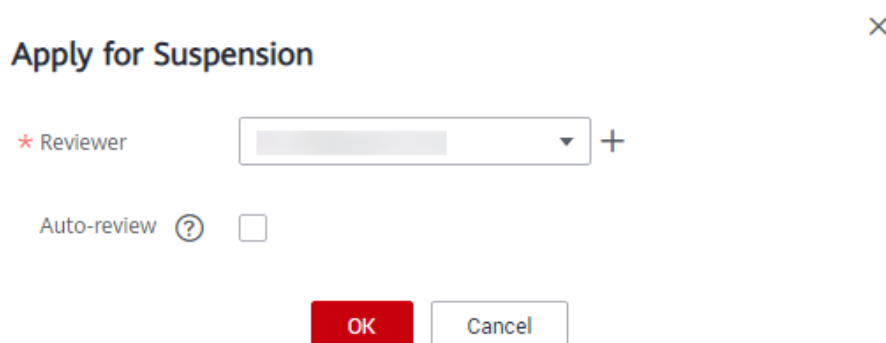
----End

Suspending a Business Metric

You can perform the following steps to suspend a published business metric:

- Step 1** On the DataArts Architecture page, choose **Metrics > Business Metrics** in the left navigation pane.
- Step 2** In the business metric list, select the target business metric and click **Suspend** in the **Operation** column.
- Step 3** In the dialog box displayed, select a reviewer and click **OK**. The business metric is suspended after the reviewer approves it.

Figure 5-125 Apply for Suspension dialog box

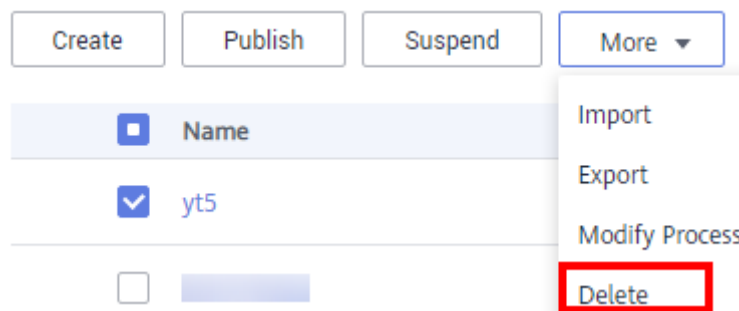


----End

Deleting a Business Metric

If a business metric is no longer needed, you can delete it. A business metric in the **Published** state can be deleted only after it is suspended.

- On the DataArts Architecture page, choose **Metrics > Business Metrics** in the left navigation pane.
- In the business metric list, select the target business metric and choose **More > Delete** above the list.

Figure 5-126 Deleting a business metric

3. In the dialog box displayed, confirm the information and click **Yes**.

5.7.2 Technical Metrics

5.7.2.1 Creating Atomic Metrics

An atomic metric is an abstract set of the statistical logic and specific algorithms. To ensure consistency between definitions and R&D, metric definitions determine the statistical logic (or the computing logic), without using ETLs to perform secondary R&D. This improves R&D efficiency and ensures consistency of statistical results.

Context

Atomic metrics come from fact tables.

- An atomic metric is a data component defined for constructing a derivative metric required by application statistical analysis. An atomic metric can only be created based on fact table details.
- A derivative metric does not have a direct source table. It belongs to the source table of the original atomic metrics that are combined into the derivative metric.

Atomic metrics and derivative metrics interact in specific ways.

- After the computing logic of an atomic metric takes effect, the related derivative metric is updated directly.
- An atomic metric referenced by any derivative metrics cannot be deleted.
- The code of an atomic metric referenced by any derivative metrics can be changed.
- The change of an atomic metric affects related derivative metrics.

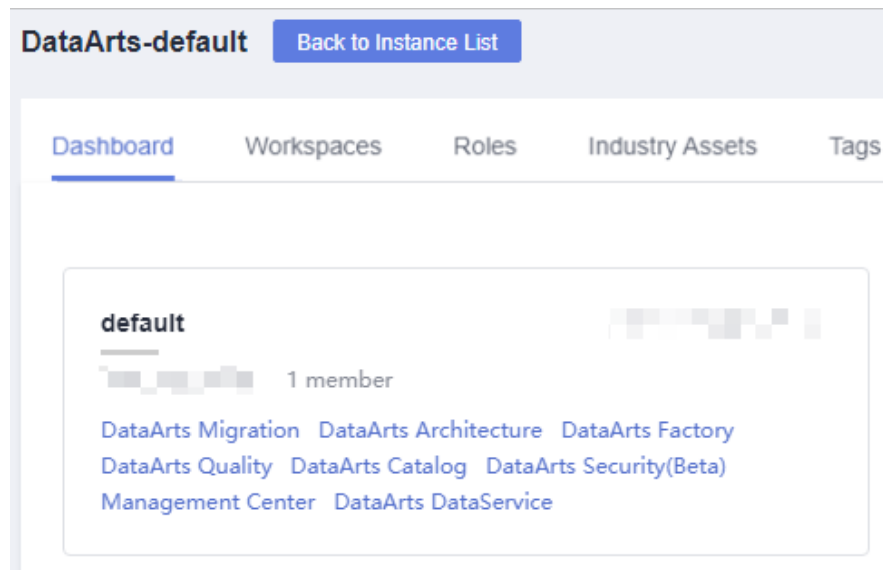
Prerequisites

You have created and published a fact table, and the fact table has been approved. For details, see [Creating Fact Tables](#).

Creating and Publishing an Atomic Metric

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

Figure 5-127 DataArts Architecture



2. On the DataArts Architecture page, choose **Metrics > Technical Metrics** in the left navigation pane. On the displayed page, click the **Atomic Metrics** tab.
3. Select a subject from the subject tree on the left and click **Create**.
4. On the **Create Atomic Metric** page, set the parameters described in [Table 5-40](#) and click **Publish**.

Figure 5-128 Creating an atomic metric

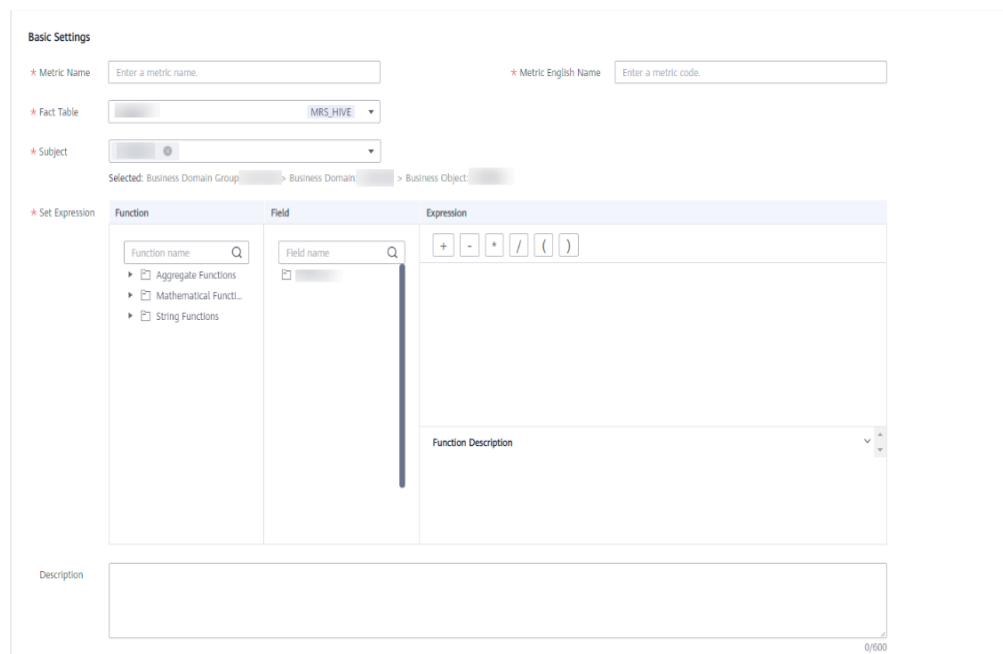


Table 5-40 Parameters for creating an atomic metric

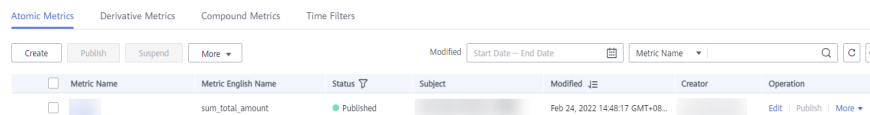
Parameter	Description
*Metric Name	Metric names must start with letters. Only letters, numbers, and underscores (_) are allowed.
*Metric Code	Metric codes must start with letters. Only letters, numbers, and underscores (_) are allowed.
*Data Table	Select a published fact table from the drop-down list box. If there are many tables, you can enter a table name in the text box to search for the desired fact table. If no fact table is available, create one. See Creating and Publishing a Fact Table .
*Subject	The subject to which the atomic metric belongs. After a fact table is selected, the information about the subject to which the fact table belongs is automatically displayed. You can also click Select to select a subject.
*Set Expression	Select the required functions and fields and set the expression.
Description	A description of the atomic metric to create. Up to 600 characters are supported.

5. In the dialog box displayed, select a reviewer and click **OK**.
6. (Optional) Create and publish other atomic metrics by repeating **3** to **5**.
7. Wait for the reviewer to approve the application.
After the application is approved, the atomic metric is created.

Managing an Atomic Metric

1. On the DataArts Architecture page, choose **Metrics > Technical Metrics** in the left navigation pane. On the displayed page, click the **Atomic Metrics** tab.

Figure 5-129 Managing an atomic metric



2. Manage your atomic metrics as required. Refer to the following table for details.

Operation	Helpful Link
Create	Creating and Publishing an Atomic Metric
Edit	3

Operation	Helpful Link
Publish	4
View Publish History	5
Suspend	6
Delete	7
Import	8
Export	9

3. Edit an atomic metric.
 - a. Click **Edit** to the right of the target atomic metric.
 - b. On the page displayed, edit the atomic metric as required.
 - c. Click **Publish**. If you do not want to immediately publish the atomic metric that you edited, click **Save** and you can publish it later.
4. Publish an atomic metric.
 - a. Click **Publish** to the right of the target atomic metric.
 - b. In the **Submit for Publication** dialog box displayed, select a reviewer from the drop-down list box.
 - c. Click **OK**.
5. View the publish history.
 - a. Select the target atomic metric in the list and choose **More > View History**.
 - b. On the **History** tab page, you can view the publish history and version comparison information of the metric.
6. Suspend an atomic metric.
 - a. Click **Suspend** to the right of the target atomic metric.
 - b. In the **Submit for Suspension** dialog box displayed, select a reviewer from the drop-down list box.
 - c. Click **OK**.

 **NOTE**

Atomic metrics cannot be suspended or deleted if they are referenced by any derivative metrics.

7. Delete an atomic metric.
 - a. Select the target atomic metric and choose **More > Delete** in the upper left corner.
 - b. In the dialog box displayed, confirm the information and click **Yes**.
8. Import

You can import atomic metrics to the system quickly.

 - a. Above the atomic metric list, choose **More > Import**.

Figure 5-130 Importing atomic metrics

Import Atomic Metric ×

Import Last Import

To avoid import errors, use the template to fill in data. [Atomic Metric Template](#)

★ Update Table No Yes

★ File

- b. Download the atomic metric template, and edit and save it.
- c. Choose whether to update existing data.

NOTE

If a code in the template already exists in the system, the data is considered duplicate.

- **No:** If the data to be imported already exists in the system, the existing data in the system will not be replaced.
 - **Yes:** If the data to be imported already exists in the system:
 - If the existing data in the system is in draft state, the data will be replaced and new draft data will be generated.
 - If the existing data in the system is in published state, expanded data will be generated.
- d. Click **Select File** and select the edited template to import.
 - e. Click **Upload**. When the template is uploaded, the **Last Import** page is displayed. You can view the imported data.
 - f. Click **Close**.
9. Export atomic metrics.

You can export atomic metrics to a local file.

- a. In the atomic metric list, select the metric to be exported.
- b. Above the atomic metric list, choose **More > Export**.

NOTE

- You can export all the atomic metrics of a subject by selecting the subject in the subject list on the left.
- You can export all the atomic metrics of a workspace, as long as there are no more than 500 atomic metrics in the workspace.

5.7.2.2 Creating Derivative Metrics

Derivative metrics are aggregated from the modifiers and dimensions of atomic metrics. Therefore, their modifiers and dimensions are derived from the attributes of atomic metrics as well. When a derivative metric is published, a summary table is automatically generated, which can be viewed in the **Automatically Aggregated** area on the **Summary Table** tab page.

Derivative metric = Atomic metric + Dimension + Time filter + General filter

- **Atomic metric** specifies the statistical standards, namely, the computing logic.
- **Dimension** is the perspective to observe and analyze business data and assist in data aggregation, drilling, slicing, and analysis, and used as a GROUP BY condition in SQL statements.
- **Time filter** is a standard definition of a time condition.
- **General filter** collects statistics on the business scope and select the records that meet the business rules (similar to the WHERE clause in SQL statements, excluding the time range).

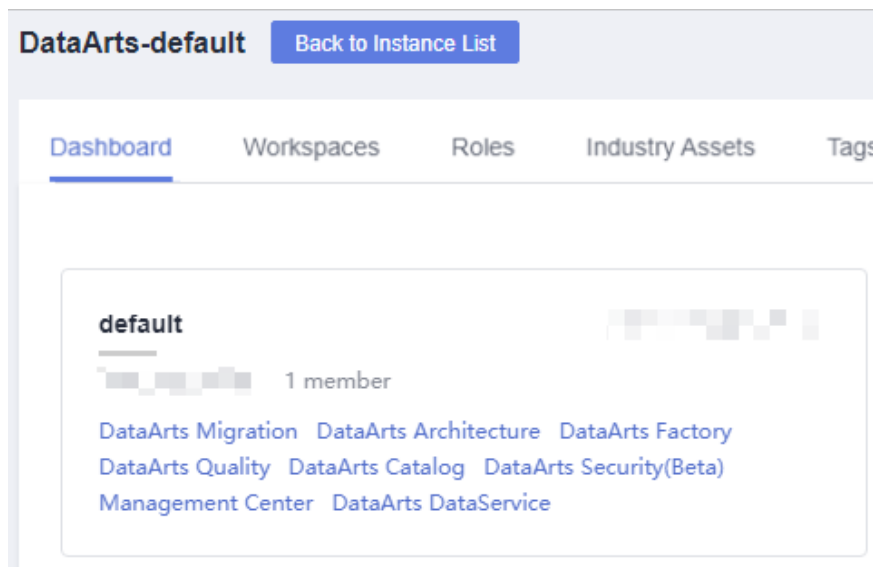
Prerequisites

- An atomic metric has been created and approved.
- A dimension and time filter have been created and approved. This prerequisite is required only if the derivative metric will use the statistical dimension or time filter.

Creating and Publishing a Derivative Metric

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

Figure 5-131 DataArts Architecture



2. On the DataArts Architecture page, choose **Metrics > Technical Metrics** in the left navigation pane. On the displayed page, click the **Derivative Metrics** tab.
3. Select a subject from the subject tree on the left and click **Create**.
4. On the page displayed, set the parameters.

Figure 5-132 Creating a derivative metric

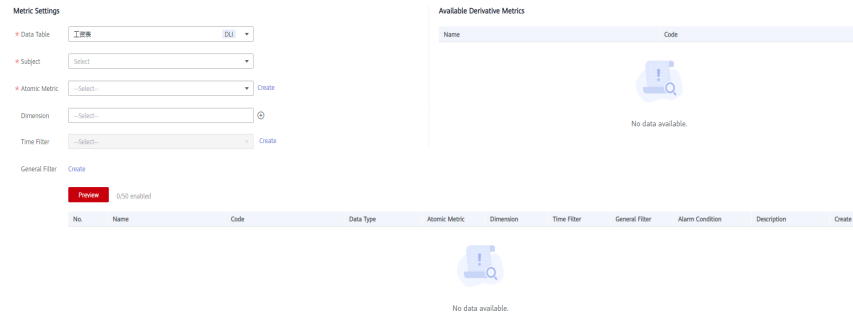


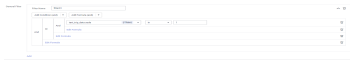


Table 5-41 Parameters for creating a derivative metric

Parameter	Description
*Data Table	Select an asset table from the drop-down list box.
*Subject	Subject information.
*Atomic Metric	Select an atomic metric.
Dimension	Select one or more dimensions from the drop-down list box. Only the attributes in the fact table associated with the atomic metrics can be selected.
Time Filter	Select the required time filter from the drop-down list and select the associated field. Some time filters are preconfigured in the system. If the available time filters cannot meet the requirements, customize one. See Creating Time Filters for details.

Parameter	Description
General Filter	<p>To set general filters, click Create.</p> <p>In the General Filter area shown in Figure 5-133, set the parameters as follows:</p> <ul style="list-style-type: none"> • Name specifies the name of a general filter. • Under Add Condition (and), you can select And condition or Or condition to add a condition. After you specify the condition, select a field from the field drop-down list and set the parameters as prompted. You can add multiple conditions. <p>You can click  to delete unwanted conditions.</p> <ul style="list-style-type: none"> • Under Add Formula (and), you can select And formula or OR formula to add a formula. Click Edit Formula if needed. In the dialog box displayed, select the required functions and fields and set the expression. <p>You can click  to delete unwanted formulas.</p> <p>Figure 5-133 Setting a general filter</p> 
Alarm Triggering Condition	<p>An alarm triggering condition consists of derivative metrics and expressions. An expression consists of alarm parameters and logical operators. When a metric is running, the system calculates the result of the alarm condition expression and determines whether to trigger the alarm based on the result of the expression. If the expression result is true, the alarm will be triggered. Otherwise, no quality alarm will be triggered.</p>
Description	<p>A description of the derivative metric to create. Up to 600 characters are supported.</p>

5. After setting the parameters, click **Preview** to view the information about the derivative metric and define the name, code, data type, alarm condition, and description for the metric.

Table 5-42 Parameters for previewing a derivative metric

Parameter	Description
Metric Name	<p>It is automatically generated by the system based on parameters such as atomic metrics, statistical dimensions, and time filters. You can also customize it.</p>

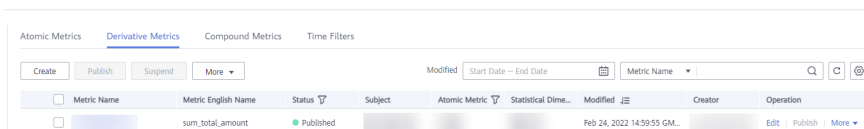
Parameter	Description
Metric Code	It is automatically generated by the system based on parameters such as atomic metrics, statistical dimensions, and time filters. You can also customize it.
Data Type	It is automatically generated by the system based on the data type of the atomic metric. You can also customize it.
Alarm Condition	An alarm condition expression consists of alarm parameters and logical operators. When a metric is running, the system calculates the result of the alarm condition expression and determines whether to trigger the alarm based on the result of the expression. If the expression result is true , the alarm will be triggered. Otherwise, no quality alarm will be triggered.
Description	A description of the derivative metric to create. Up to 600 characters are supported.

6. In the lower part of the page, click **Trial Run**. In the dialog box displayed, click **Trial Run** to check whether the derivative metric can run properly.
If the trial run fails, locate the fault based on the error message, correct the configurations, and click **Trial Run** to try again.
7. If the trial run is successful, click **Publish**.
8. In the dialog box displayed, select a reviewer and click **OK**.
9. (Optional) Create and publish other derivative metrics by repeating **2** to **8**.
10. Wait for the reviewer to approve the application.
After the application is approved, the derivative metric is created.

Managing a Derivative Metric

On the **Derivative Metrics** tab page, you can edit, publish, suspend, or delete derivative metrics.

Figure 5-134 Managing derivative metrics



1. On the DataArts Architecture page, choose **Metrics > Technical Metrics** in the left navigation pane. On the displayed page, click the **Derivative Metrics** tab.
2. Manage your derivative metrics as required. Refer to the following table for details.

Operation	Helpful Link
Create	Creating and Publishing a Derivative Metric

Operation	Helpful Link
Edit	3
Publish	4
View Publish History	5
Preview SQL	6
Suspend	7
View Summary Table	8
Delete	9
Import	10
Export	11

3. Edit a derivative metric.
 - a. Click **Edit** to the right of the target derivative metric.
 - b. On the page displayed, edit the derivative metric as required.
 - c. In the lower part of the page, click **Trial Run**. In the dialog box displayed, click **Trial Run** to check whether the derivative metric can run properly.
If the trial run fails, locate the fault based on the error message, correct the configurations, and click **Trial Run** to try again.
 - d. If the trial run is successful, click **Publish**.
4. Publish a derivative metric.
 - a. Click **Publish** to the right of the target derivative metric.
 - b. In the **Submit for Publication** dialog box displayed, select a reviewer from the drop-down list box.
 - c. Click **OK**.
5. View the publish history.
 - a. Select the target derivative metric and choose **More > View History**.
 - b. On the page displayed, you can view the publish history and version comparison information of the metric.
6. Preview an SQL statement.
 - a. Select the target derivative metric and choose **More > Preview SQL**.
 - b. On the page displayed, you can view or copy the SQL statement.
7. Suspend a derivative metric.

 **NOTE**

The prerequisite for suspending a derivative metric is that the derivative metric is not referenced to any compound metrics.

- a. Choose **More > Suspend** on the right of the target derivative metric.
- b. In the **Submit for Suspension** dialog box displayed, select a reviewer from the drop-down list box.

- c. Click **OK**.
8. View a summary table.
Currently, only details about automatically generated summary tables can be viewed. Choose **More > View Summary Table** on the right of the target derivative metric. The **Summary Tables** page is displayed.
9. Delete a derivative metric.

 **NOTE**

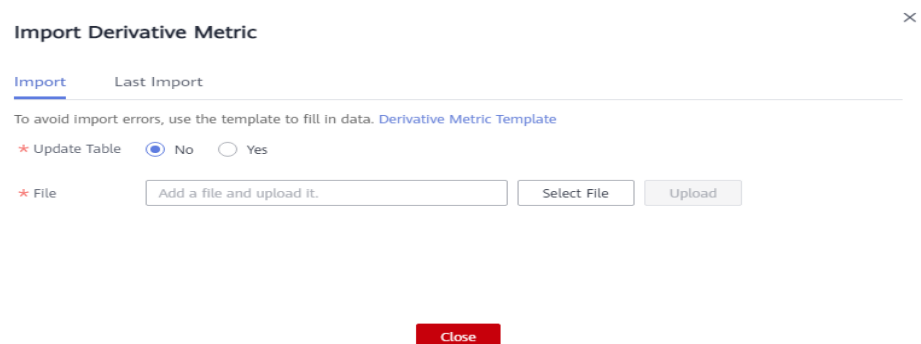
The prerequisite for deleting a derivative metric is that the derivative metric is not referenced to any compound metrics.

- a. Select the target derivative metric and choose **More > Delete** above the list.
 - b. In the dialog box displayed, confirm the information and click **Yes**.
10. Import derivative metrics.

You can import derivative metrics to the system quickly.

- a. Above the summary table list, choose **More > Import**.

Figure 5-135 Importing derivative metrics



- b. Download the derivative metric template, and edit and save it.
- c. Choose whether to update existing data.

 **NOTE**

If a code in the template already exists in the system, the data is considered duplicate.

- **No:** If the data to be imported already exists in the system, the existing data in the system will not be replaced.
 - **Yes:** If the data to be imported already exists in the system:
 - If the existing data in the system is in draft state, the data will be replaced and new draft data will be generated.
 - If the existing data in the system is in published state, expanded data will be generated.
- d. Click **Select File** and select the edited template to import.
 - e. Click **Upload**. When the template is uploaded, the **Last Import** page is displayed. You can view the imported data.
 - f. Click **Close**.

11. Exporting derivative metrics.

You can export derivative metrics to a local file.

- a. In the derivative metric list, select the metric to be exported.
- b. Above the derivative metric list, choose **More > Export**.

NOTE

- You can export all the derivative metrics of a subject by selecting the subject in the subject list on the left.
- You can export all the derivative metrics of a workspace, as long as there are no more than 500 derivative metrics in the workspace.

5.7.2.3 Creating Compound Metrics

A compound metric is generated by adding one or more derivative metrics. The dimensions and modifiers of a compound metric are the same as those of the derivative metrics. New dimensions and modifiers cannot be generated outside the scope of derivative metrics, dimensions, and modifiers.

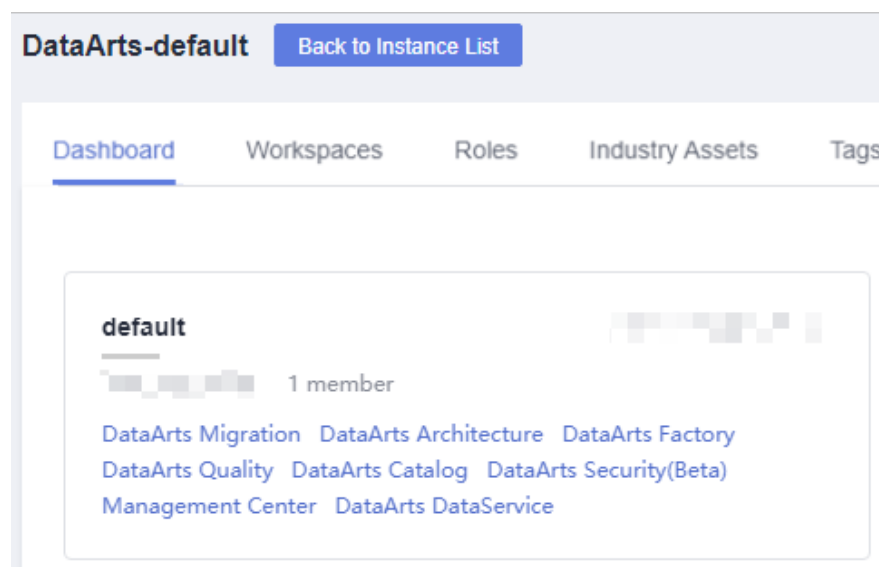
Prerequisites

A derivative metric has been created and approved. For details, see [Creating Derivative Metrics](#).

Creating a Compound Metric

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

Figure 5-136 DataArts Architecture



2. On the DataArts Architecture page, choose **Metrics > Technical Metrics** in the left navigation pane. On the displayed page, click the **Compound Metrics** tab.
3. Select a subject from the subject tree on the left and click **Create**.

- On the page displayed, set the parameters.

Figure 5-137 Creating a compound metric

The screenshot shows a web form for creating a compound metric. It contains the following fields and sections:

- * Metric Name:** A text input field with the placeholder "Enter a compound metric name."
- * Metric English Name:** A text input field with the placeholder "Enter a compound metric english name."
- * Subject:** A dropdown menu with "--Select--" as the current selection.
- * Statistical Dimension:** A dropdown menu with "--Select--" as the current selection.
- * Data Type:** A dropdown menu with "--Select--" as the current selection.
- * Expression:** A complex section divided into two parts:
 - Derivative Metric:** A search bar with the placeholder "Enter a keyword" and a magnifying glass icon.
 - Expression:** A large text area for entering mathematical or logical expressions. Above it are buttons for mathematical operators: "+", "-", "*", "/", "(", and ")", along with a search icon.
- Description:** A text area at the bottom with the placeholder "Enter a description." and a character count "0/600" at the bottom right.

Table 5-43 Parameters for creating a compound metric

Parameter	Description
*Metric Name	Compound metric names must start with letters. Only letters, numbers, and underscores (_) are allowed.
*Metric Code	Metric code names must start with letters. Only letters, numbers, and underscores (_) are allowed.
*Subject	Subject information. Select a subject.
*Statistical Dimension	The available options are those configured on the Derivative Metrics page.
*Data Type	Select a data type for the compound metric.
*Expression	Select the required derivative metrics and set the expression as required.
Alarm Triggering Condition	An alarm triggering condition consists of derivative metrics and expressions. An expression consists of alarm parameters and logical operators. When a metric is running, the system calculates the result of the alarm condition expression and determines whether to trigger the alarm based on the result of the expression. If the expression result is true , the alarm will be triggered. Otherwise, no quality alarm will be triggered. NOTE Currently, quality alarms cannot be triggered.
Description	A description of the compound metric to create. Up to 600 characters are supported.

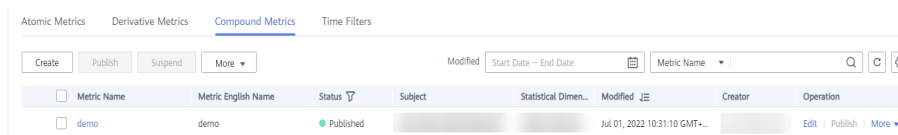
- In the lower part of the page, click **Trial Run**. In the dialog box displayed, click **Trial Run** to check whether the compound metric can run properly. If the trial run fails, locate the fault based on the error message, correct the configurations, and click **Trial Run** to try again.

6. If the trial run is successful, click **Publish**.
7. In the dialog box displayed, select a reviewer and click **OK**.
8. Wait for the reviewer to approve the application.
After the application is approved, the compound metric is created.

Editing a Compound Metric

1. On the DataArts Architecture page, choose **Metrics > Technical Metrics** in the left navigation pane. On the displayed page, click the **Compound Metrics** tab.

Figure 5-138 Compound metrics



Metric Name	Metric English Name	Status	Subject	Statistical Dimen...	Modified	Creator	Operation
demo	demo	Published			Jul 01, 2022 10:31:10 GMT+		Edit Publish More

2. In the compound metric list, select the target metric and click **Edit** on the right.
3. On the page displayed, set the parameters as prompted. For details, see [Table 5-43](#).
4. In the lower part of the page, click **Trial Run**. In the dialog box displayed, click **Trial Run** to check whether the compound metric can run properly.
If the trial run fails, locate the fault based on the error message, correct the configurations, and click **Trial Run** to try again.
5. If the trial run is successful, click **Publish**.
6. In the dialog box displayed, select a reviewer and click **OK**.

Publishing a Compound Metric

After creating or editing a compound metric, it takes effect only after it is published. Compound metrics to be published, already published, or to be suspended cannot be published.

1. On the DataArts Architecture page, choose **Metrics > Technical Metrics** in the left navigation pane. On the displayed page, click the **Compound Metrics** tab.
2. In the compound metric list, select the target compound metric and click **Publish**.
3. In the dialog box displayed, click **OK**.

Viewing the Publish History

1. On the DataArts Architecture page, choose **Metrics > Technical Metrics** in the left navigation pane. On the displayed page, click the **Compound Metrics** tab.
2. Select the target compound metric in the list and choose **More > View History**.
3. On the page displayed, you can view the publish history and version comparison information of the metric.

Previewing an SQL Statement

1. On the DataArts Architecture page, choose **Metrics > Technical Metrics** in the left navigation pane. On the displayed page, click the **Compound Metrics** tab.
2. Locate the target compound metric and choose **More > Preview SQL**.
3. In the dialog box displayed, you can view or copy the SQL statement.

Suspending a Compound Metric

You can bring a published compound metric offline if it is no longer used.

NOTE

The prerequisite for suspending a compound metric is that the metric is not referenced to any summary table.

1. On the DataArts Architecture page, choose **Metrics > Technical Metrics** in the left navigation pane. On the displayed page, click the **Compound Metrics** tab.
2. In the compound metric list, select the target compound metrics and click **Suspend** above the list.
3. In the dialog box displayed, click **OK**.

Deleting a Compound Metric

NOTE

The prerequisite for deleting a compound metric is that the metric is not referenced to any summary table.

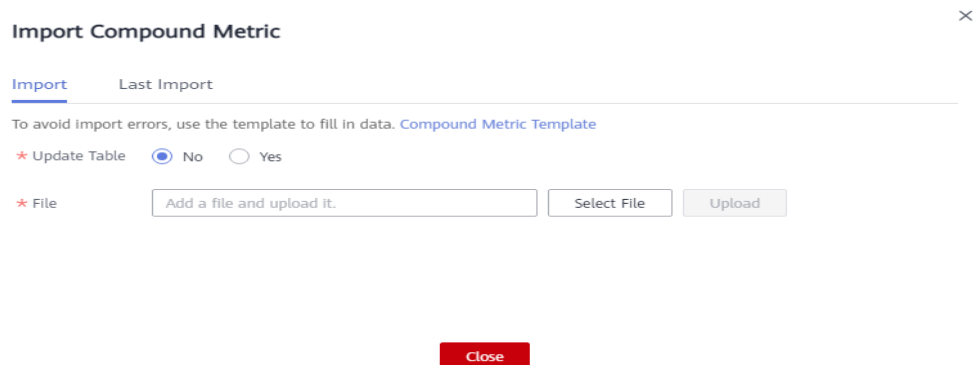
1. On the DataArts Architecture page, choose **Metrics > Technical Metrics** in the left navigation pane. On the displayed page, click the **Compound Metrics** tab.
2. Select the target compound metric and choose **More > Delete** above the list.
3. In the dialog box displayed, confirm the information and click **Yes**.

Importing Compound Metrics

You can import compound metrics to the system quickly.

1. Above the compound metric list, choose **More > Import**.

Figure 5-139 Importing Compound Metrics



2. Download the compound metric template, and edit and save it.
3. Choose whether to update existing data.

 **NOTE**

- If a code in the template already exists in the system, the data is considered duplicate.
- **No:** If the data to be imported already exists in the system, the existing data in the system will not be replaced.
 - **Yes:** If the data to be imported already exists in the system:
 - If the existing data in the system is in draft state, the data will be replaced and new draft data will be generated.
 - If the existing data in the system is in published state, expanded data will be generated.
4. Click **Select File** and select the edited template to import.
 5. Click **Upload**. When the template is uploaded, the **Last Import** page is displayed. You can view the imported data.
 6. Click **Close**.

Exporting Compound Metrics

You can export compound metrics to a local file.

1. In the compound metric list, select the metric to be exported.
2. Above the compound metric list, choose **More > Export**.

 **NOTE**

- You can export all the compound metrics of a subject by selecting the subject in the subject list on the left.
- You can export all the compound metrics of a workspace, as long as there are no more than 500 compound metrics in the workspace.

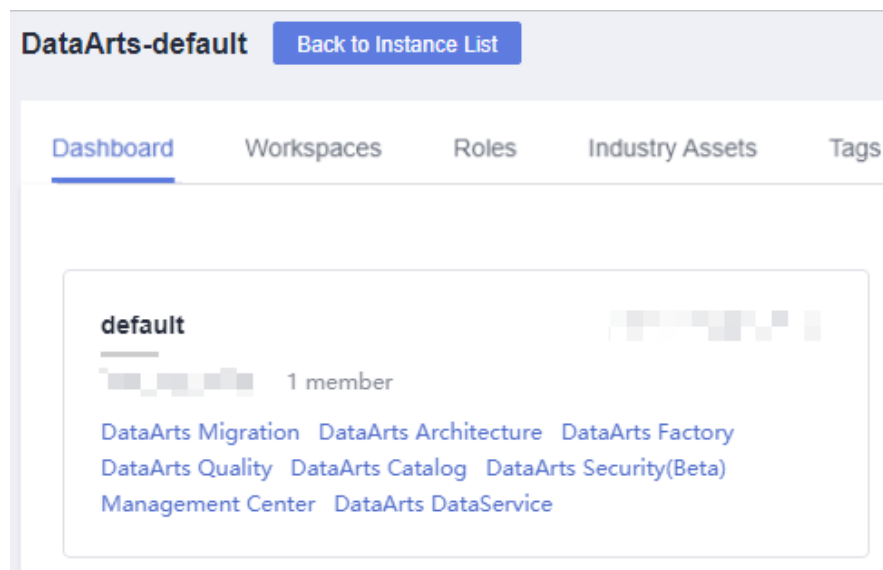
5.7.2.4 Creating Time Filters

Atomic metrics are standard definitions for computing logic. Time filters are standard definitions for conditional limits. To ensure that all statistical metrics are unified, standard, and unambiguous, time filters must be unique within a business domain and each filter can belong to a single source logic table. The computing logic is defined based on the fields of the source logic table model. A time filter may come from multiple logic tables that belong to different data domains. Therefore, a time filter may belong to multiple data domains as well.

Creating and Publishing a Time Filter

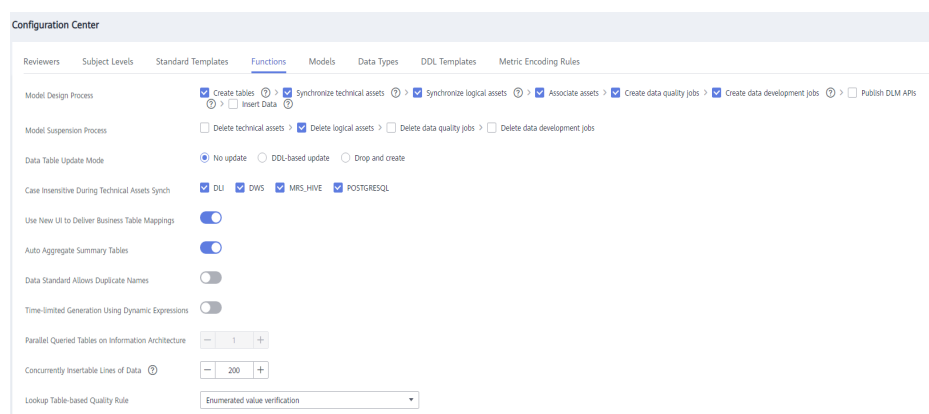
1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

Figure 5-140 DataArts Architecture



2. (Optional) On the DataArts Architecture console, choose **Configuration Center** in the left navigation pane, click the **Functions** tab, and determine whether to enable **Time-Limited Generation Using Dynamic Expressions** (disabled by default).

Figure 5-141 Functions



3. On the DataArts Architecture page, choose **Metrics > Technical Metrics** in the left navigation pane. On the displayed page, click the **Time Filters** tab.
4. On the **Time Filters** tab page, click **Create**.
5. On the **Create Time Filter** page, set the parameters described in [Table 5-44](#) and click **Publish**.

Figure 5-142 Creating a time filter

The screenshot shows a form for creating a time filter. It has two input fields at the top: '* Filter Name' and '* Filter English Name'. Below these is the '* Time Settings' section, which has tabs for 'Year', 'Month', 'Day', 'Hour', and 'Minute'. Under the 'Year' tab, there are two radio buttons: 'Quick option' (selected) and 'Custom'. The 'Quick option' has sub-options for 'Last year' and 'This year'. The 'Custom' option has two input fields separated by 'to'. At the bottom is a 'Description' text area. A small '0/490' character count is visible at the bottom right of the description field.

Table 5-44 Parameters for creating a time filter

Parameter	Description
*Filter Name	Time filter names must start with letters. Only letters, numbers, and underscores (_) are allowed.
*Filter English name	Only letters, digits, and underscores (_) are allowed.
*Time Settings	You can select Year , Month , Day , Hour , or Minute , and then select Quick option or Custom to set the time condition. If you select Custom , + and - form a time range, in which + indicates a later time and - indicates an earlier time. For example, if you want to set a time range from the past year to the next three years, set this parameter to -1 to +3 or +3 to -1 .
Description	A description of the time filter to create. Up to 490 characters are supported.

6. Then, click **Publish** to submit the application.
7. Wait for the reviewer to approve the application.
After the application is approved, the time filter is created.

Managing a Time Filter

1. On the DataArts Architecture page, choose **Metrics > Technical Metrics** in the left navigation pane. On the displayed page, click the **Time Filters** tab.

Figure 5-143 Time Filters tab page


The screenshot shows the 'Time Filters' tab page. At the top, there are tabs for 'Atomic Metrics', 'Derivative Metrics', 'Compound Metrics', and 'Time Filters'. Below the tabs are buttons for 'Create', 'Publish', 'Suspend', and 'Delete'. A search bar is present with 'Modified' and 'Filter Name' dropdowns. The main content is a table with the following data:

Filter Name	Filter English Name	Status	Modified	Creator	Operation
Next 4 weeks		Published	Feb 22, 2022 15:32:47 GMT+08:00	SYSTEM	EDIT Publish More
Next 7 days		Published	Feb 22, 2022 15:32:47 GMT+08:00	SYSTEM	EDIT Publish More

2. Manage your time filters as required. Refer to the following table for details.

Operation	Helpful Link
Create	Creating and Publishing a Time Filter

Operation	Helpful Link
Edit	3
Publish	4
View Publish History	5
Suspend	6
Delete	7

3. Edit a time filter.
 - a. Click **Edit** to the right of the target time filter.
 - b. On the page displayed, edit the time filter as required.
 - c. Click **Save** to save the time filter information, or click **Publish** to publish the edited time filter.
 4. Publish a time filter.
 - a. Click **Publish** to the right of the target time filter.
 - b. In the **Submit for Publication** dialog box displayed, select a reviewer from the drop-down list box.
 - c. Click **OK**.
 5. View the publish history.
 - a. Select the target time filter in the list and choose **More > View History**.
 - b. On the page displayed, you can view the publish history and version comparison information of the time filter.
 6. Suspend a time filter.
 - a. Select the target time filter in the list and choose **More > Suspend**.
 - b. In the **Submit for Suspension** dialog box displayed, select a reviewer from the drop-down list box.
 - c. Click **OK**.
-  **NOTE**
- Time filters cannot be suspended or deleted if they are referenced by any derivative metrics.
7. Delete a time filter.
 - a. Select the target time filter and click **Delete** above the list.
 - b. In the dialog box displayed, confirm the information and click **Yes**.

5.8 Data Mart Building

5.8.1 Creating Summary Tables

A summary table consists of specific analysis objects (such as members) and related statistical metrics. The metrics included in a summary table all have the

same level of granularity (such as members). A summary table provides users with all of the available statistics on themed data (such as a member theme market), sorted by levels of granularity.

A summary table can be manually or automatically aggregated. This topic describes how to manually create a summary table.

NOTE

On the DataArts Architecture page, choose **Metrics > Configuration Center** in the left navigation pane, and click the **Functions** tab. On the page displayed, if **Create data development jobs** is selected for **Model Design Process**, the system creates a data development job with a name starting with *Database name_Table code*. Choose **DataArts Factory > Develop Job** to view the created job. By default, this job has no scheduling configuration. You need to configure scheduling for the job in the DataArts Factory module.

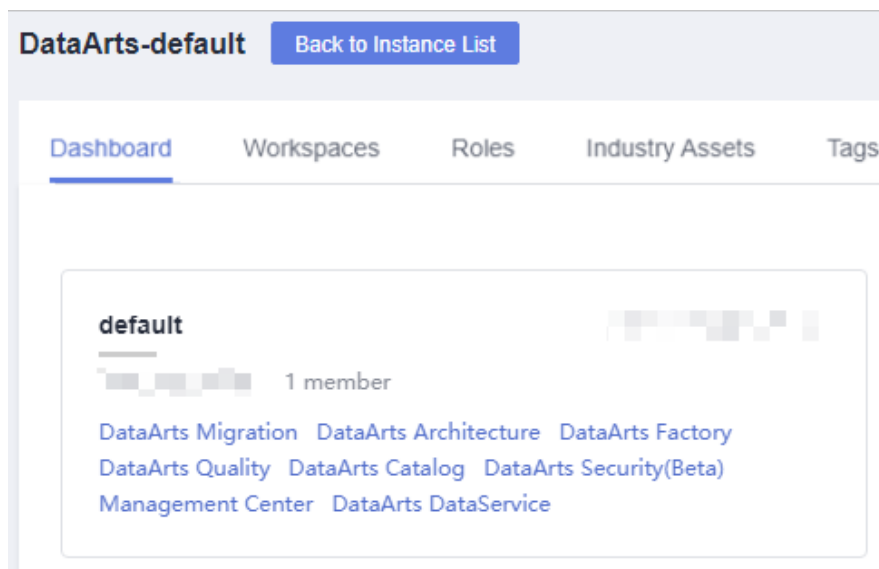
Prerequisites

A dimension, a dimension table, a fact table, and a derivative metric have been created, published, and reviewed.

Creating and Publishing a Summary Table

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

Figure 5-144 DataArts Architecture



2. On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane. On the displayed page, click the **Summary Tables** tab.
3. Select a subject from the subject tree on the left and click **Create**.
4. On the **Create Summary Table** page, perform the following operations:
 - a. Set the parameters in the **Basic Settings** area.

Figure 5-145 Basic Settings area

The screenshot shows the 'Basic Settings' configuration area. It contains the following fields:

- * Subject:** A dropdown menu with '--Select--' as the current selection.
- * Table Name:** A text input field with the placeholder text 'Enter a summary table name'.
- * Table English Name:** A text input field containing the value 'dws_'.
- * Statistical Dimension:** A dropdown menu with '--Select--' as the current selection.
- * Data Connection Type:** A dropdown menu with '--Select--' as the current selection.
- * Data Connection Name:** A dropdown menu with '--Select--' as the current selection.
- * Database:** A dropdown menu with '--Select--' as the current selection.
- * Owner:** A text input field with the placeholder text 'Enter an asset owner'.
- * Description:** A large text area containing the value 'None'.

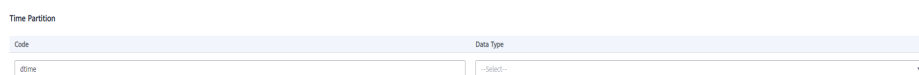
Table 5-45 Parameters in the Basic Settings area

Parameter	Description
*Subject	Select a subject catalog (business domain group > business domain > business object) where you can place the summary table.
*Table Name	The name of the table to create. It must start with letters. Only letters, digits, and the following special characters are allowed: ()-_
*Table Code	The code of the table to create. Table codes must start with letters. Only letters, numbers, and underscores (_) are allowed.
*Statistical Dimension	Select a statistical dimension. The drop-down list displays only the statistical dimensions configured on the Derivative Metrics page. If no statistical dimension is available in the drop-down list box, create one by referring to Creating Dimensions . After a summary table is created, all dimension attributes of the specified dimensions are automatically added to the summary table as fields in the summary table. After creating a summary table, go to the summary table page and click the table name to view the field details in the table.
*Data Connection Type	The parameter value must be the same as that of the dimension table and fact table.
*Data Connection Name	It is recommended that the same data connection be used for dimension modeling.
*Database	The name of the database. Select a database from the drop-down list box.

Parameter	Description
Queue	DLI queue. This parameter is available only for DLI data connections.
Schema	DWS or POSTGRESQL mode. This parameter is displayed only for DWS and POSTGRESQL data connections.
Table Type	DWS connections support the following tables: <ul style="list-style-type: none"> • DWS_ROW: Tables are stored to disk partitions by row. • DWS_COLUMN: Tables are stored to disk partitions by column. MRS_HIVE supports only HIVE_TABLE .
Distributed By	This parameter is displayed only for DWS data connections. Currently, only REPLICATION and HASH are supported. You can select multiple fields. <ul style="list-style-type: none"> • REPLICATION: A full table is stored on each DN. The advantage of this option is that each DN has all the data of a table. During the join operation, data redistribution can be avoided, reducing network overhead. The disadvantage is that each DN retains the complete data of a table, resulting in data redundancy. Generally, this option is recommended for small dimension tables. • HASH: If you select this option, you must specify a distribution key for the user table. When a record is inserted, the system performs hash computing based on values in the distribute keys and then stores data on the corresponding DN. In a Hash table, I/O resources on each node can be used during data read/write, which improves the read/write speed of a table. Generally, this option is recommended for large dimension tables (a large dimension table contains over 1 million records).
*Owner	You can enter an owner name or select an existing owner.
*Description	A description of the summary table to create. It allows 1 to 600 characters.

- b. In the **Time Partition** area, enter the field code and select the data type. After a table is published, data is written to the table based on the time partition fields.

Figure 5-146 Time Partition area



- c. In the **Metric Fields** area, click **Add** to add derivative or compound metrics that are associated with the specified statistical dimensions.

Figure 5-147 Metric Fields area

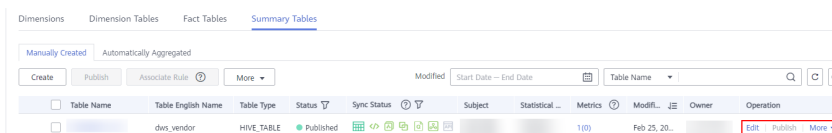


5. Click **Publish**. In the dialog box displayed, click **OK**.
6. Select a reviewer to approve the summary table.
After the summary table is approved, it is automatically created in the database.
7. Go back to the summary table list and locate the table just published. View its synchronization status in the **Sync Status** column.
 - If the synchronization is successful, the summary table is successfully published and created in the database.
 - If the synchronization failed, choose **More > View History** in the row where the summary table is located. On the page displayed, click the **History** tab to view logs. Troubleshoot the problem based on the logs. After the error is rectified, choose **More > Synchronize** above the summary table list to issue the synchronization command again. If the problem persists, contact technical support personnel.

Managing a Summary Table

1. On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane. On the displayed page, click the **Summary Tables** tab.

Figure 5-148 Summary Tables page




2. Manage your summary tables as required. Refer to the following table for details.

Operation	Helpful Link
Create	Creating and Publishing a Summary Table
Edit	3
Publish	4
View History	5
Preview SQL	6
Suspend	7
Associate Rule	8

Operation	Helpful Link
Delete	9
Import	10
Export	11

3. Edit a summary table.
 - a. Click **Edit** to the right of the target summary table.
 - b. Edit the summary table as required.
 - c. Click **Publish**.
4. Publish a summary table.
 - a. Click **Publish** to the right of the target summary table.
 - b. In the **Submit for Publication** dialog box displayed, select a reviewer from the drop-down list box.
 - c. Click **OK**.
5. View the publish history.
 - a. Select the target summary table in the list and choose **More > View History** on the right.
 - b. On the page displayed, you can view the publish history and version comparison information of the summary table.

If the publish log includes error logs, the publishing has failed. You can click **Resynchronize** to retry.
6. Previewing an SQL statement.
 - a. Select the target summary table in the list and choose **More > Preview SQL** on the right.
 - b. On the page displayed, you can view or copy the SQL statement.
7. Suspend a summary table.
 - a. Click **Suspend** to the right of the target summary table.
 - b. In the **Submit for Suspension** dialog box displayed, select a reviewer from the drop-down list box.
 - c. Click **OK**.

 **NOTE**

After a summary table is suspended, you can determine how to process APIs based on the actual situation in DataArts DataService. DataArts Architecture does not process the APIs.
8. Associate a summary table with a quality rule.
 - a. Select the target summary table in the summary table list and click **Associate Rule** above the list.
 - b. In the **Associate Quality Rule** dialog box, you can add rules to the fields in the summary table in batches and associate the rules with the fields.
 - c. Click **OK**.

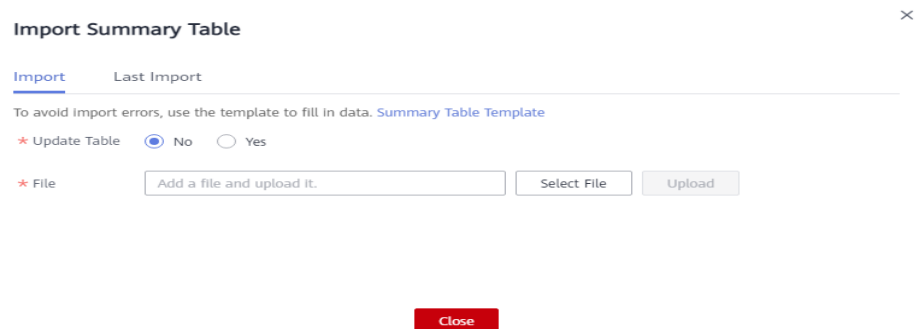
9. Delete a summary table.
 - a. Select the target summary table and choose **More > Delete** above the list.
 - b. In the dialog box displayed, click **Yes**.

10. Import

You can import summary tables to the system quickly.

- a. Above the summary table list, choose **More > Import**.

Figure 5-149 Import Summary Table



- b. Download the summary table template, and edit and save it.
- c. Choose whether to update existing data.

NOTE

If a code in the template already exists in the system, the data is considered duplicate.

- **No:** If the data to be imported already exists in the system, the existing data in the system will not be replaced.
 - **Yes:** If the data to be imported already exists in the system:
 - If the existing data in the system is in draft state, the data will be replaced and new draft data will be generated.
 - If the existing data in the system is in published state, expanded data will be generated.
- d. Click **Select File** and select the edited template to import.
 - e. Click **Upload**. When the template is uploaded, the **Last Import** page is displayed. You can view the imported data.
 - f. Click **Close**.

11. Export summary tables.

You can export summary tables to a local file.

- a. Select the summary tables to export on the **Manually Created** or **Automatically Aggregated** page.
- b. Above the summary table list, choose **More > Export**.

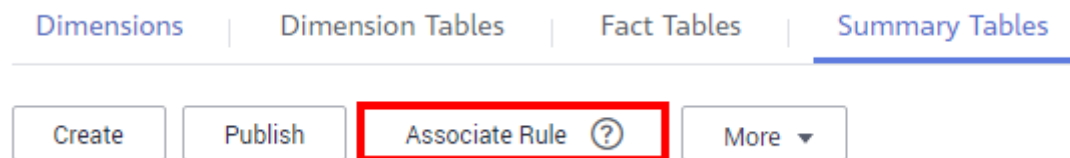
 **NOTE**

- You can export all the summary tables of a subject by selecting the subject in the subject list on the left.
- You can export all the summary tables of a workspace, as long as there are no more than 500 summary tables in the workspace.

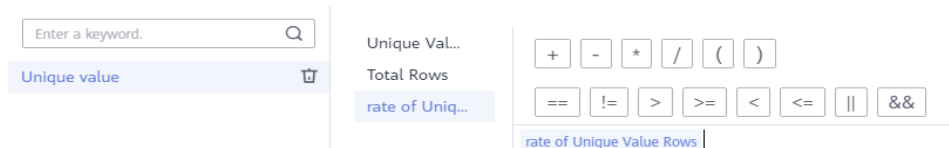
Associating a Summary Table with a Quality Rule

1. On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane.
2. Click the **Summary Tables** tab.
3. Select the target summary table in the list, and click **Associate Rule**.

Figure 5-150 Associating a summary table with a quality rule



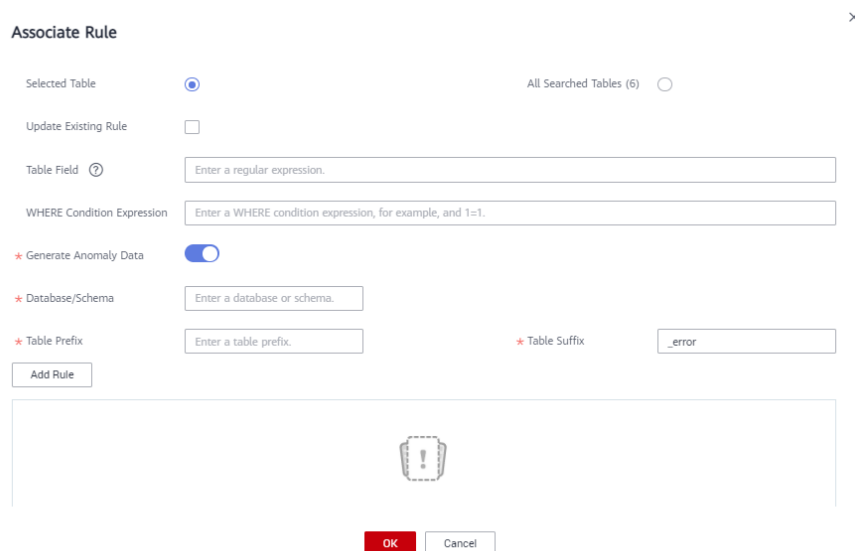
4. On the page displayed, set the parameters. After the configuration is complete, click **OK**.
 - **Update Existing Rule:** If this option is selected, the newly added rule will overwrite the old rule.
 - **Table Field:** This parameter applies to all fields by default. You can enter a regular expression to filter fields as required.
 - **WHERE Clause:** This parameter can be used to filter fields.
 - **Generate Anomaly Data:** If this option is selected, anomaly data is stored in the specified database based on the configured parameters.
 - **Database/Schema:** database or schema that stores anomaly data. This parameter is displayed when **Generate Anomaly Data** is enabled.
 - **Table Prefix:** prefix of the table that stores anomaly data. This parameter is displayed when **Generate Anomaly Data** is enabled.
 - **Table Suffix:** suffix of the table that stores anomaly data. This parameter is displayed when **Generate Anomaly Data** is enabled.
 - **Add Rule:** You can click **Add Rule** to add a rule. For example, add a rule named **Unique value**, select the rule, click **OK**, enter an alarm condition expression in the **Alarm Condition** text box, add other rules in the same way, and click **OK**. An example alarm expression is as follows:



- An alarm condition expression consists of alarm parameters and logical operators. When a quality job is running, the system calculates the result of the alarm condition expression and determines whether to trigger the alarm based on the result of the expression. If the expression result is

true, the alarm will be triggered. Otherwise, no quality alarm will be triggered. In the **Associate Quality Rule** dialog box, the alarm parameters of each quality rule are displayed as buttons.

Figure 5-151 Associating a summary table with a quality rule



Associating a Summary Table Field with a Data Standard








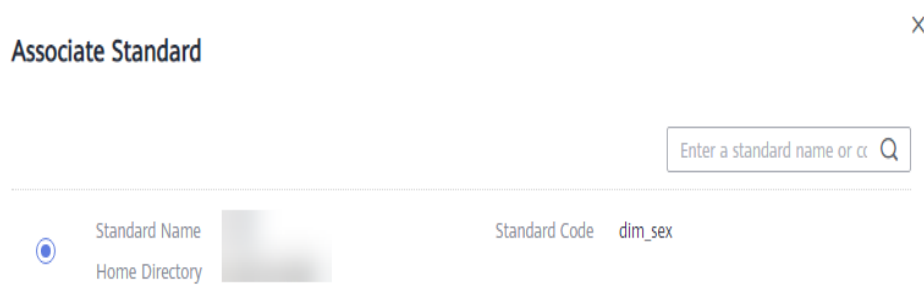
1. On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane.
2. Click the **Summary Tables** tab.
3. Click the name of the target summary table in the list.
4. In the table field list on the details page of the summary table, search for the target field, click  corresponding to the field to configure the association between the field and the data standard.

Figure 5-152 Associating a summary table field with a data standard

No.	Configuration Type	Name	English Name	Field Type	Primary Key	Partition	Not Null	Associate Data Standard	Associate Rule	Comment
1	Time period		dttime	TIMESTAMP	N	Y	N			
2	Derivative metric		sum_total_amount	STRING	N	N	N			
3	Dimension Field		dim_vendor_vendor_id	BIGINT	N	N	N			

5. After the configuration is complete, click **OK**. For details on the sources of data standards, see [Creating a Data Standard](#).

Figure 5-153 Associating a data standard



Associating a Single Field with a Quality Rule




1. On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane.
2. Click the **Summary Tables** tab.
3. In the summary table list, click the name of the target summary table.
4. In the table field list on the summary table details page, locate the target field and click  to associate the field with a quality rule.


Figure 5-154 Associating a single table field with a quality rule







Table Field

Anomaly Data Output Settings 

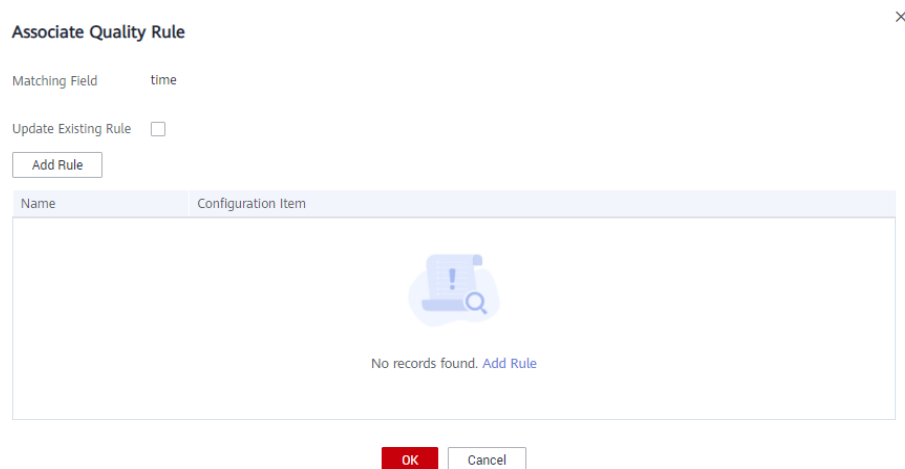
Generate Anomaly Data Disabled

WHERE Condition Expression 

Associate Rule  Clear Rule

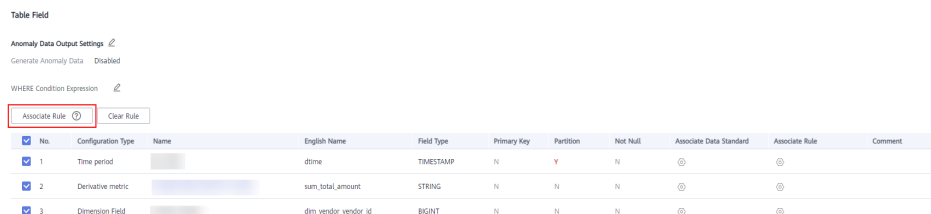
<input type="checkbox"/>	No.	Configuration Type	Name	English Name	Field Type	Primary Key	Partition	Not Null	Associate Data Standard	Associate Rule	Comment
<input type="checkbox"/>	1	Time period		dtime	TIMESTAMP	N	Y	N			
<input type="checkbox"/>	2	Derivative metric		sum_total_amount	STRING	N	N	N			
<input type="checkbox"/>	3	Dimension Field		dim_vendor_vendor_id	BIGINT	N	N	N			

5. After the configuration is complete, click **OK**.
 - **Update Existing Rule:** If this option is selected, the newly added rule will overwrite the old rule.
 - **Add Rule:** You can click **Add Rule** to add a rule. For example, add a rule named **Unique value**, select the rule, click **OK**, enter an alarm condition expression in the **Alarm Condition** text box, add other rules in the same way, and click **OK**.
 - An alarm condition expression consists of alarm parameters and logical operators. When a quality job is running, the system calculates the result of the alarm condition expression and determines whether to trigger the alarm based on the result of the expression. If the expression result is **true**, the alarm will be triggered. Otherwise, no quality alarm will be triggered. In the **Associate Quality Rule** dialog box, the alarm parameters of each quality rule are displayed as buttons.

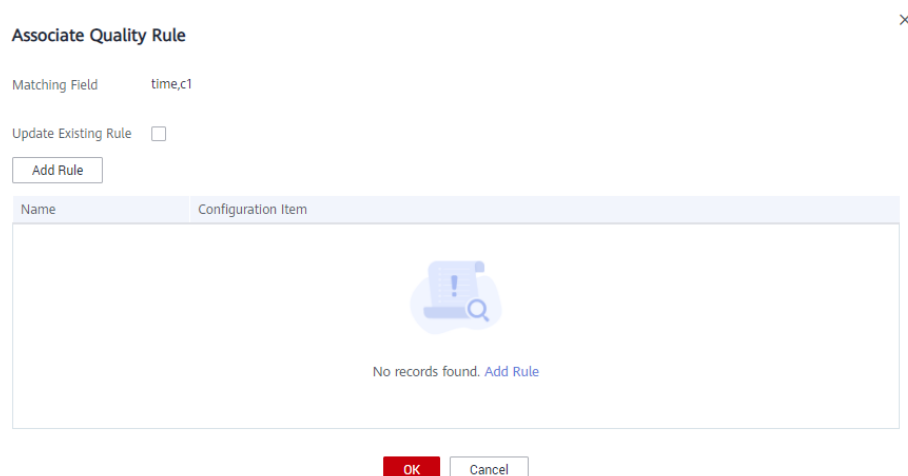
Figure 5-155 Associating a quality rule

Associating Table Fields with a Quality Rule in Batches

1. On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane.
2. Click the **Summary Tables** tab.
3. In the summary table list, click the name of the target summary table.
4. In the table field list on the summary table details page, select the target table fields and click **Associate Rule**.

Figure 5-156 Associating fields with a quality rule

5. On the page displayed, add a rule and set the rule parameters.
 - **Update Existing Rule:** If this option is selected, the newly added rule will overwrite the old rule.
 - **Add Rule:** You can click **Add Rule** to add a rule. For example, add a rule named **Unique value**, select the rule, click **OK**, enter an alarm condition expression in the **Alarm Condition** text box, add other rules in the same way, and click **OK**.
 - An alarm condition expression consists of alarm parameters and logical operators. When a quality job is running, the system calculates the result of the alarm condition expression and determines whether to trigger the alarm based on the result of the expression. If the expression result is **true**, the alarm will be triggered. Otherwise, no quality alarm will be triggered. In the **Associate Quality Rule** dialog box, the alarm parameters of each quality rule are displayed as buttons.

Figure 5-157 Adding a rule

6. After the configuration is complete, click **OK**.

5.9 Common Operations

5.9.1 Reversing a Database (ER Modeling)

You can import tables from databases of other data sources to a specific ER model.

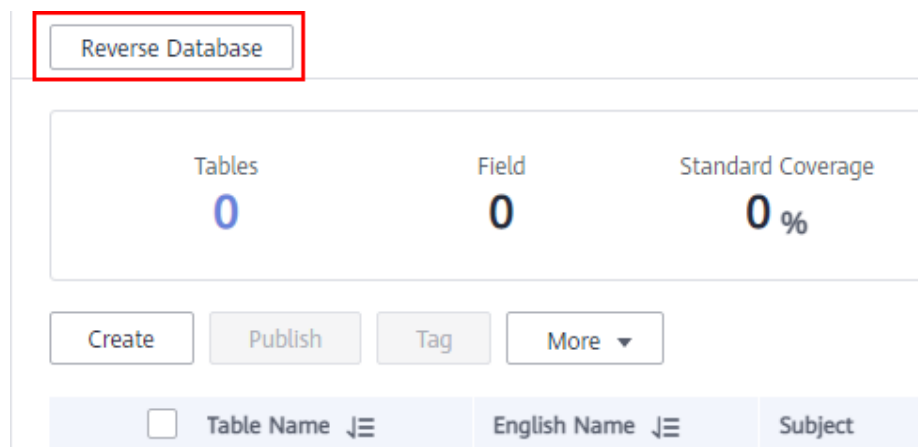
Prerequisites

You have collected metadata from databases in DataArts Catalog so that the system can synchronize tables to DataArts Catalog later. Otherwise, the synchronization tasks may fail. See [Task Management](#) for details.

Importing a Table to a Model by Reversing the Database

- Step 1** On the DataArts Architecture page, choose **Models > ER Modeling** in the left navigation pane.
- Step 2** On the model overview page, locate the target model, click the model card, and click **Reverse Database** in the upper part.

Figure 5-158 Reverse Database dialog box



Step 3 In the **Reverse Database** dialog box, set the parameters.

Figure 5-159 Setting parameters for reversing the database

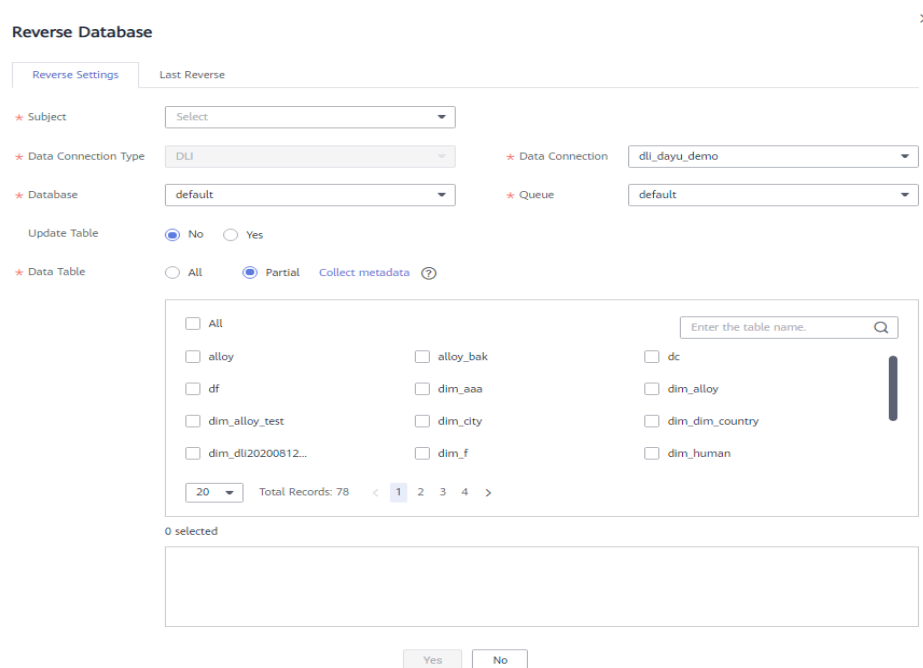


Table 5-46 Parameters for reversing a database

Parameter	Description
Subject	Select a subject from the drop-down list box.
Data Connection Type	If you reverse tables to a logical model, select a required data connection type from the drop-down list box. If you reverse tables to a physical model, the data connection type of the current model is displayed.

Parameter	Description
Data Connection	The name of the data connection. Select the required data connection. If you want to reverse a database from other data sources to an ER model, you must create a data connection in Management Center to connect to the data source. For details on how to create data connections, see Creating Data Connections .
Database	The name of the database. Select a database from the drop-down list box.
Queue	This parameter is displayed only for DLI data connections. Select a DLI queue.
Schema	This parameter is displayed only for DWS data connections.
Update Table	Whether to update the existing table if the table to be imported already exists in the ER model. When a table is imported, the system checks whether the table exists according to the table code. During the import, only table creation and update are allowed. <ul style="list-style-type: none">• No: If you select this option, the existing tables will not be updated.• Yes: If you select this option, the existing tables will be updated. If a table is in the Published state, you must publish the table again after updating it so that the updated table can take effect.
Data Table	If you select All , all tables in the database are imported to the ER model. If you select Partial , not all tables in the database are imported to the ER model.

Step 4 Click **Yes** to start reversing the database.

----End

5.9.2 Reversing a Database (Dimensional Modeling)

You can import tables from databases of other data sources to a specific ER model.

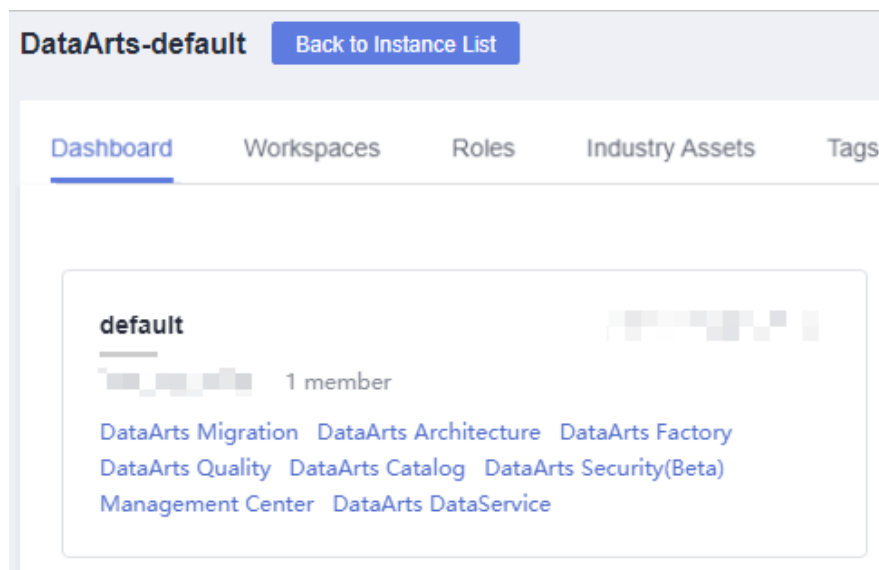
Prerequisites

You have collected metadata from databases in DataArts Catalog so that the system can synchronize tables to DataArts Catalog later. Otherwise, the synchronization tasks may fail. See [Task Management](#) for details.

Importing a Table to a Model by Reversing the Database

Step 1 On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

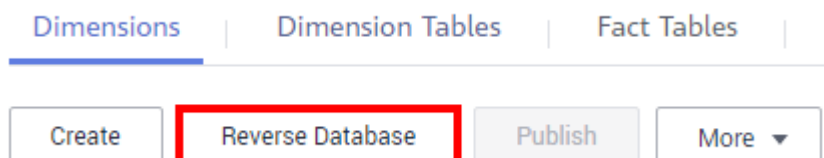
Figure 5-160 DataArts Architecture



Step 2 On the DataArts Architecture page, choose **Models > Dimensional Modeling** in the left navigation pane.

Step 3 Click the **Dimensions** or **Fact Tables** tab. Then, click **Reverse Database** above the list.

Figure 5-161 Selecting an object



Step 4 In the **Reverse Database** dialog box, set the parameters.

Table 5-47 Parameters for reversing a database

Parameter	Description
Subject	Select a subject from the drop-down list box.
Data Connection Type	Type of the database to reverse.
Data Connection	The name of the data connection. If you want to reverse a database from other data sources to an ER model, you must create a data connection in Management Center to connect to the data source. For details on how to create data connections, see Creating Data Connections .
Database	The name of the database. Select a database from the drop-down list box.

Parameter	Description
Queue	This parameter is displayed only for DLI data connections. Select a DLI queue.
Schema	DWS or POSTGRESQL mode. This parameter is displayed only for DWS and POSTGRESQL data connections.
Update Existing Table	The import operation can be used to create a table or update an existing table. It does not delete a table. <ul style="list-style-type: none">● No: If you select this option, the existing tables will not be updated.● Yes: If you select this option, the existing tables will be updated. If a table is in the Published state, you must publish the table again after updating it so that the updated table can take effect.
Data Table	If you select All , all tables in the database are imported. If you select Partial , not all tables in the database are imported.

Step 5 Click **Yes** to start reversing the database. After the operation is complete, you can view the result on the **Last Reverse** tab page or perform the reverse operation again.

----End

5.9.3 Importing/Exporting Tables

You can import tables to an ER model in batches. You can also export existing tables and import them to other models.

Importing a Table to a Logical Model

- Step 1** On the DataArts Architecture page, choose **Models > ER Modeling** in the left navigation pane.
- Step 2** On the model overview page, click the card of the target logical model, select an object in the subject directory, and choose **More > Import**.
- Step 3** In the dialog box displayed, click **ER Modeling Template**.

Figure 5-162 Import Table dialog box

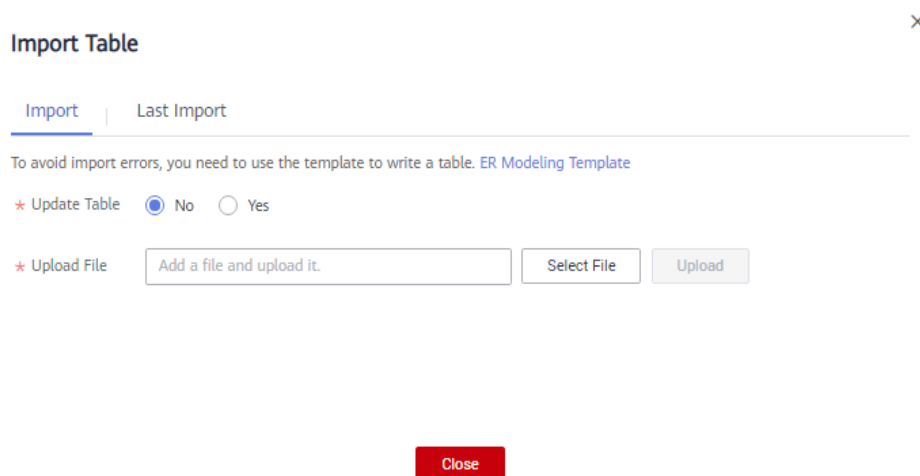


Table 5-48 Parameters for importing a table

Parameter	Description
Update Table	<p>Whether to update the existing table if the table to be imported already exists in the ER model. The system determines whether the table to import exists in the ER model based on the table code. The import operation can be used to create a table or update an existing table. It does not delete a table. The options are as follows:</p> <ul style="list-style-type: none"> • No: If you select this option, the existing tables will not be updated. • Yes: If you select this option, the existing tables will be updated. If a table is in the Published state, you must publish the table again after updating it so that the updated table can take effect.
Upload File	<p>Select the file to import. You can use either of the following methods to obtain the file to import:</p> <ul style="list-style-type: none"> • Downloading the ER modeling template and fill in the template In the Import Table dialog box, click ER Modeling Template to download the template, fill in the template, and save the settings. • Exporting tables to files You can export the lookup tables created in DataArts Architecture of a DataArts Studio instance to an Excel file. Then, import the Excel file. See Exporting a Table or DDL for details.

Step 4 Open the downloaded template, set the parameters in the template, and save the template. The **Description** sheet in the template is for reference only.

Parameters whose names start with an asterisk (*) are mandatory, and other parameters are optional.

The table below describes the parameters in the **Tables** sheet.

Table 5-49 Parameters in the Tables sheet

Parameter	Description
Subject	Enter the encoding paths of existing subjects, which are separated by slashes (/). If no subject is available, create one by referring to Designing Subjects .
*Logical Entity Name	It must start with letters. Only letters, digits, and the following special characters are allowed: ()-_
*Table Name	Name of the table. Table codes cannot start with numbers. Only letters, numbers, and the following special characters are allowed: _\${}
Table Alias	Alias of a table. This parameter is displayed when you have enabled Table Alias on the Configuration Center page.
Table Tags	Tags to be added to the table. Enter an existing or a new tag. You can add a tag on the Tags page of DataArts Catalog and go back to this page to select the tag. For details on how to add a tag, see Tags .
*Table Description	A description of the table.
Owner	You can enter an owner name or select an existing owner in the current workspace of the DataArts Studio instance.
Parent Table	You can enter only the names of other tables in this template.
DWS DISTRIBUTE BY	This field is required only for DWS data connections. The HASH (attribute name) and REPLICATION modes are supported.
*Field Name	The name of a field in the table. It must start with letters. Only letters, digits, and the following special characters are allowed: ()-_
Field Code	The code of the field in the table. It must start with letters. Only letters, digits, and underscores (_) are allowed.
Field Alias	Alias of a field. This parameter is displayed when you have enabled Field Alias on the Configuration Center page.
Field Ordinal	Sequence number of the field in the table. The value starts from 1. This parameter is optional. If this parameter is left blank, fields are sorted in the sequence in the template by default.
Field Description	A description of the field.
*Field Data Type	Data type of the logical model. For details, see the DEFAULT group in Data Types .

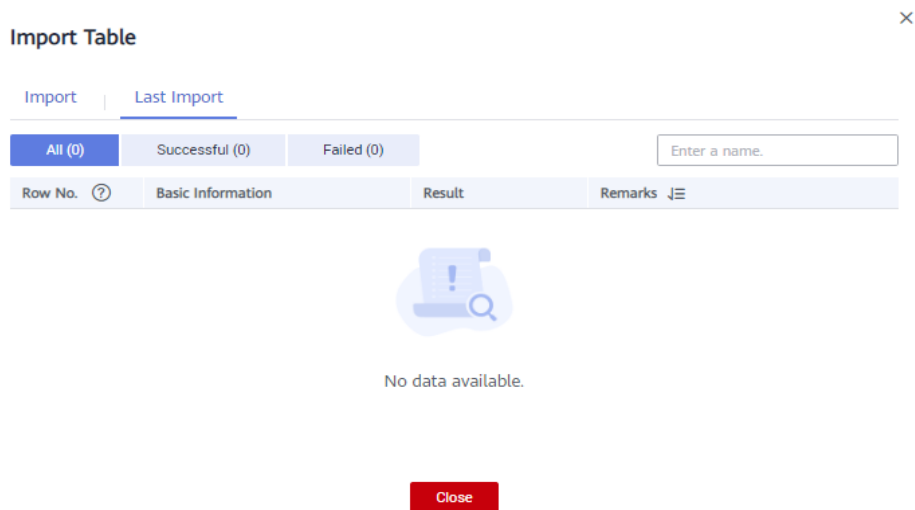
Parameter	Description
Field Data Length	Data length. For a variable-length data type, specify the data length if a data connection type supports the data length. For example, for the DWS data connection type, if the field type is CHAR(10) , set Field Data Type to CHAR and Field Data Length to 10 .
Partition	The value Y indicates that the field is a partition field, and the value N indicates that the field is not a partition field.
Primary Key	The value Y indicates that the field is a primary key, and the value N indicates that the field is not a primary key.
Not Null	The value Y indicates that the field is not empty, and the value N indicates that the field can be empty.
Associate Data Standard	The code of the data standard to be associated. This field can be left blank. If no data standard is available, create one. See Creating Data Standards for details.
Field Tags	Tags to add to the field. Enter an existing tag or a new tag.. You can add a tag on the Tags page of DataArts Catalog and go back to this page to select the tag. For details on how to add a tag, see Tags .

[Table 5-21](#) describes the parameters in the **Relations** sheet.

Currently, mappings cannot be imported. You do not need to fill in the **Mappings** sheet.

- Step 5** View the result on the **Last Import** tab page. If the import is successful, click **Close**. If the import fails, you can view the failure cause, correct the template file, and upload it again.

Figure 5-163 Last Import tab page



----End

Importing a Table to a Physical Model

- Step 1** On the DataArts Architecture page, choose **Models > ER Modeling** in the left navigation pane.
- Step 2** In the ER model tree, select a physical model, expand it, and select a target. Then, choose **More > Import**.
- Step 3** In the dialog box displayed, click **ER Modeling Template**.

Figure 5-164 Import Table dialog box

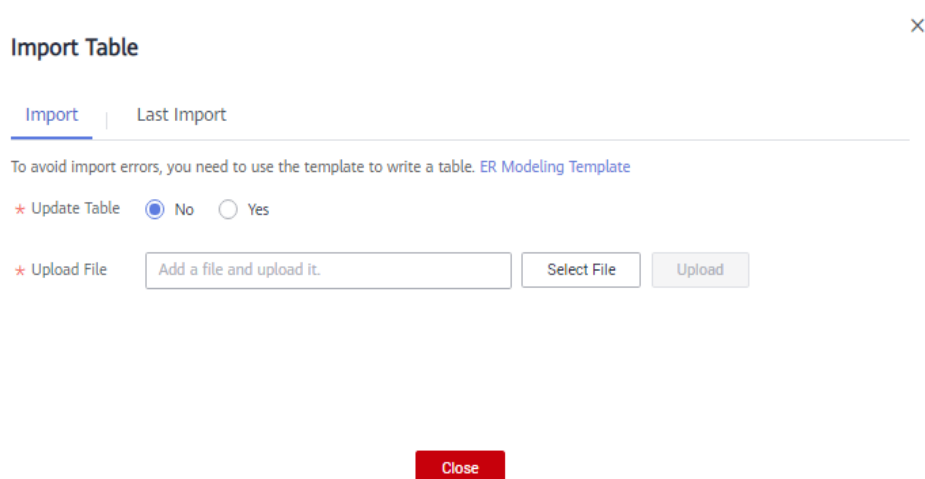


Table 5-50 Parameters for importing a table

Parameter	Description
Update Table	<p>Whether to update the existing table if the table to be imported already exists in the ER model. The system determines whether the table to import exists in the ER model based on the table code. The import operation can be used to create a table or update an existing table. It does not delete a table. The options are as follows:</p> <ul style="list-style-type: none"> • No: If you select this option, the existing tables will not be updated. • Yes: If you select this option, the existing tables will be updated. If a table is in the Published state, you must publish the table again after updating it so that the updated table can take effect.
Upload File	<p>Select the file to import. You can use either of the following methods to obtain the file to import:</p> <ul style="list-style-type: none"> • Downloading the ER modeling template and fill in the template In the Import Table dialog box, click ER Modeling Template to download the template, fill in the template, and save the settings. • Exporting tables to files You can export the lookup tables created in DataArts Architecture of a DataArts Studio instance to an Excel file. Then, import the Excel file. See Exporting a Table or DDL for details.

Step 4 Open the downloaded template, set the parameters in the template, and save the template. The **Description** sheet in the template is for reference only.

Parameters whose names start with an asterisk (*) are mandatory, and other parameters are optional.

The table below describes the parameters in the **Tables** sheet.

Table 5-51 Parameters in the Tables sheet

Parameter	Description (Importing DLI/POSTGRESQL/DWS/MRS Hive Tables)
Subject	Enter the encoding paths of existing subjects, which are separated by slashes (/). If no subject is available, create one. For details, see Designing Subjects .
*Logical Entity Name	It must start with letters. Only letters, digits, and the following special characters are allowed: ()-_
*Table Name	Name of the table. Table codes cannot start with numbers. Only letters, numbers, and the following special characters are allowed: _\${}

Parameter	Description (Importing DLI/POSTGRESQL/DWS/MRS Hive Tables)
Table Alias	Alias of a table. This parameter is displayed when you have enabled Table Alias on the Configuration Center page.
Table Tags	Tags to be added to the table. Enter an existing or a new tag. You can add a tag on the Tags page of DataArts Catalog and go back to this page to select the tag. For details on how to add a tag, see Tags .
*Table Description	A description of the table.
Asset Owner	Enter the username for entering the current workspace. Only the workspace admin, developer, or O&M personnel can be set as the designer.
Data Connection Type	The following connection types are supported: DWS, DLI, POSTGRESQL, and MRS Hive.
*Table Type	<p>DLI models support the following table types:</p> <ul style="list-style-type: none"> ● DLI_MANAGED: Data is stored in a DLI table. ● DLI_EXTERNAL: Data is stored in an OBS table. When Table Type is set to DLI_EXTERNAL, you must set OBS Path. ● DLI_VIEW is available for import only. <p>DWS models support the following table types:</p> <ul style="list-style-type: none"> ● DWS_ROW: row type ● DWS_COLUMN: column type ● DWS_VIEW: view type <p>This parameter is unavailable for the tables created in MRS Hive models.</p>
OBS Path	Enter an OBS path for storing the source data associated with the table if Table Type is set to DLI_EXTERNAL . The OBS path format is <i>bucket_name/filepath</i> .
Data Format	<p>This parameter is available only for tables created in DLI models.</p> <p>If the table type is DLI_MANAGED, the options of the data format are Parquet and Carbon.</p> <p>If the table type is DLI_EXTERNAL, the options of the data format are Parquet, Carbon, CSV, ORC, JSON, and Avro.</p>
Data Connection	Enter the name of a created data connection.
Database	Enter the name of a created database.

Parameter	Description (Importing DLI/ POSTGRESQL/DWS/MRS Hive Tables)
Connection Extra	If Data Connection Type is DLI , enter a DLI queue name. If Data Connection Type is DWS or POSTGRESQL , enter a schema name.
*Field Name	The name of a field in the table. It must start with letters. Only letters, digits, and the following special characters are allowed: ()-_
*Field Code	The code of the field in the table. It must start with letters. Only letters, digits, and underscores (_) are allowed.
Field Ordinal	Sequence number of the field in the table. The value starts from 1. This parameter is optional. If this parameter is left blank, fields are sorted in the sequence in the template by default.
Field Description	A description of the field.
*Field Data Type	The supported data types vary depending on the data connection types. For details, see Data Types .
Field Data Length	For a variable-length data type, specify the data length if a data connection type supports the data length. For example, for the DWS data connection type, if the field type is CHAR(10) , set Field Data Type to CHAR and Field Data Length to 10 .
Partition	The value Y indicates that the field is a partition field, and the value N indicates that the field is not a partition field.
Primary Key	The value Y indicates that the field is a primary key, and the value N indicates that the field is not a primary key.
Not Null	The value Y indicates that the field is not empty, and the value N indicates that the field can be empty.
Associate Data Standard	The code of the data standard to be associated. This field can be left blank. If no data standard is available, create one. For details, see Creating Data Standards .
Field Tags	Tags to add to the field. Enter an existing tag or a new tag.. You can add a tag on the Tags page of DataArts Catalog and go back to this page to select the tag. For details on how to add a tag, see Tags .

Parameter	Description (Importing DLI/POSTGRESQL/DWS/MRS Hive Tables)
configs	Additional table configuration details stored in JSON format. The format is as follows: <pre>{ "option_name1": "value", "option_name2": "value" }</pre> Example: <pre>{ "a1": "100", "a2": "30" }</pre>
Version	This parameter is optional.

Table 5-27 describes the parameters in the **Relations** sheet.

Currently, mappings cannot be imported. You do not need to fill in the **Mappings** sheet.

Step 5 View the import result on the **Last Import** tab page. If the import is successful, click **Close**. If the import fails, you can view the failure cause, correct the template file, and upload it again.

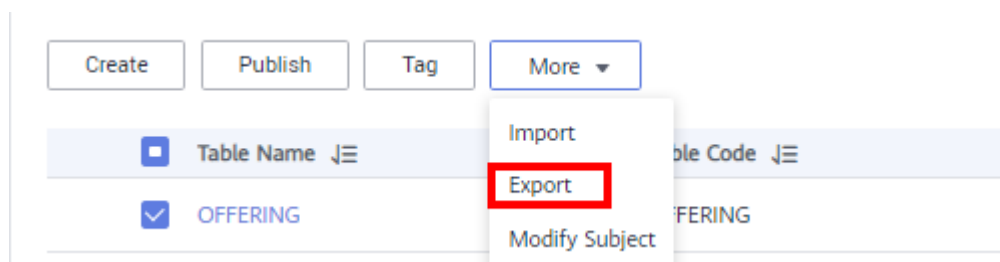
----End

Exporting a Table or DDL

Step 1 On the DataArts Architecture page, choose **Models > ER Modeling** in the left navigation pane.

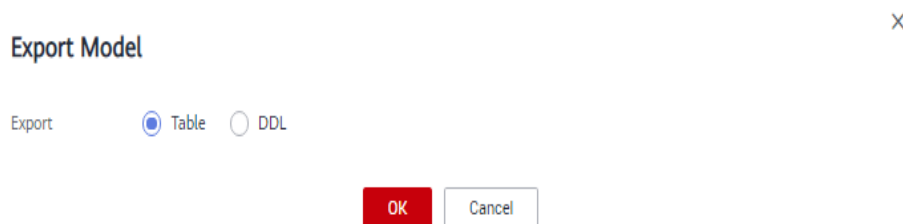
Step 2 On the model overview page, click the card of the target logical model, select an object in the subject directory, and choose **More > Export**.

Figure 5-165 Exporting a table or DDL

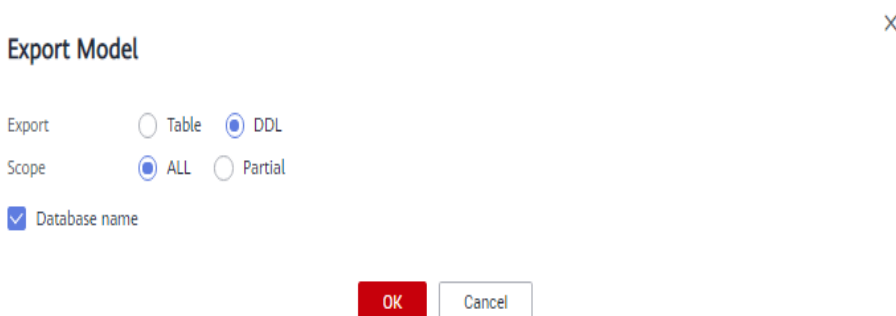


Step 3 In the dialog box displayed, select the objects to export.

The exported Excel file can be imported.

Figure 5-166 Exporting a table

When a DDL is exported, the DDL statements of the selected table are exported to TXT files.

Figure 5-167 Exporting a DDL

Step 4 Click **OK**.

----End

5.9.4 Associating Quality Rules

After creating and publishing a table, you can associate quality rules with the table. If **Create Data Quality Jobs** is selected for **Model Design Process** on the **Function Settings** tab page of **Configuration Center**, a quality job is automatically created in DataArts Quality after a quality rule is associated and the table is published. If the table has been published, the system automatically updates the corresponding quality job.

Associating a Quality Rule and Viewing a Quality Job

- Step 1** On the DataArts Architecture page, choose **Models > ER Modeling** in the left navigation pane.
- Step 2** On the page displayed, select the target model. All tables created in the model are listed on the right. You can also expand a topic structure and select an object. All tables of the object are listed on the right.
- Step 3** In the table list, select a table and click its name to access the table details page.

Figure 5-168 ER model list

<input type="checkbox"/>	Table Name	English Name	Subject	Database	Status	Sync Status	Tag	Table Type	Modified	Owner	Operation
<input type="checkbox"/>	dw_taxi_trip_data			demo_dw_ob	Publis...			HIVE_TABLE	Feb 25, 2022...		Edit Publish More

Step 4 In the **Table Field** area, select a field that you want to associate a quality rule with and click **Associate Rule**.

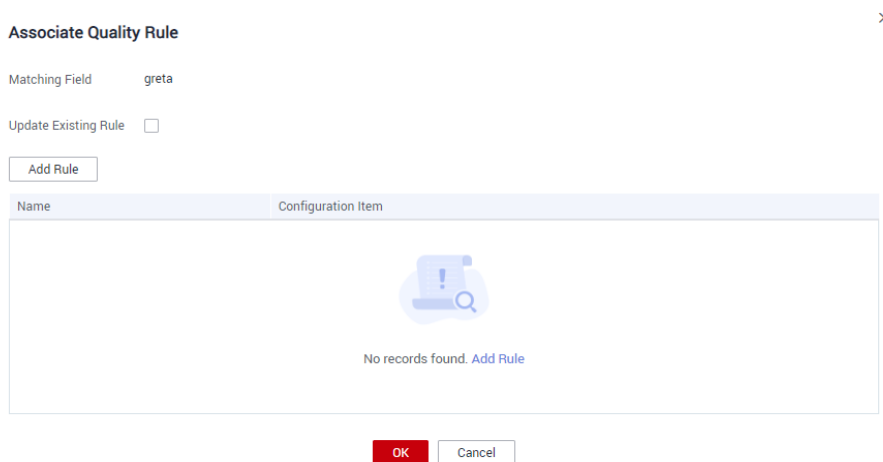
Figure 5-169 Associating Quality Rules



Anomaly Data Output Settings: If you select **Generate Anomaly Data**, the anomaly data is stored in the specified database based on the settings.

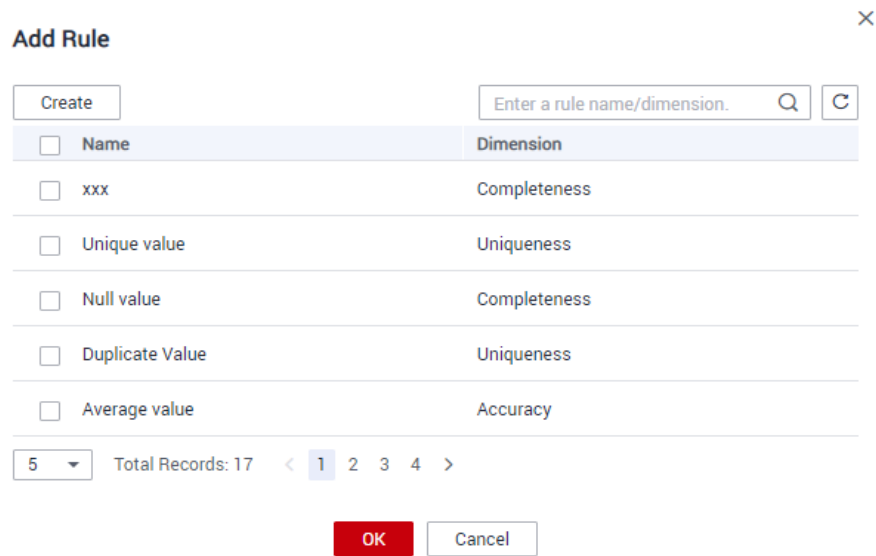
Step 5 In the dialog box displayed, click **Add Rule**.

Figure 5-170 Adding a quality rule



The **Add Rule** dialog box lists all default quality rules supported by DataArts Quality. Select a rule and click **OK**. If these quality rules cannot meet your requirements, you can customize one. In the **Add Rule** dialog box, click **Create** to navigate to DataArts Quality and create a rule on the page displayed. See [Creating Rule Templates](#).

Figure 5-171 Add Rule dialog box

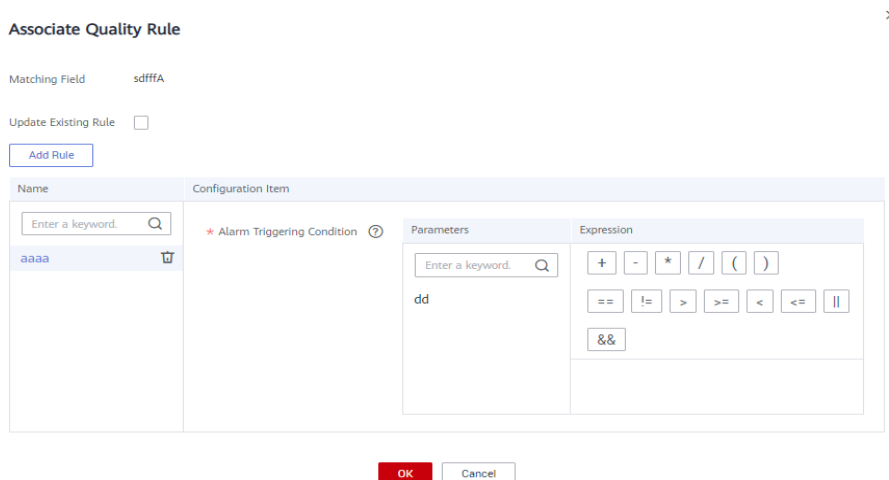


After a rule is added, the **Associate Quality Rule** dialog box is displayed. Select a rule from the rule name list, set **Alarm Condition**, and click **OK**.

- In the **Alarm Condition** text box, enter an expression. When a quality job is running, the system calculates the result of the alarm condition expression and determines whether to trigger the alarm based on the result of the expression. If the expression result is **true**, the alarm will be triggered. Otherwise, no quality alarm will be triggered.
- An alarm condition expression consists of alarm parameters and logical operators.

The alarm parameters of each rule are displayed as buttons. If you click these buttons, the alarm conditions are expressed in the sequence of alarm parameters, such as $\${1}$, $\${2}$, and $\${3}$. The variable names indicate the alarm parameters. In other words, when setting **Alarm Condition**, use the variable $\${1}$ to represent the first alarm parameter, $\${2}$ to represent the second alarm parameter, and so on.

Figure 5-172 Setting an alarm triggering condition



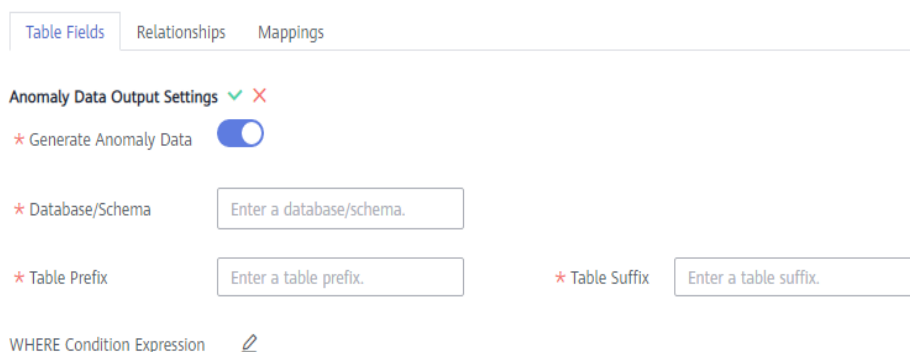
Step 6 (Optional) If you want to store anomaly data that does not comply with the preset rules in the exception table, enable **Anomaly Data Output Settings**.

Figure 5-173 Enabling Anomaly Data Output Settings



Click the pen icon next to **Anomaly Data Output Settings** and enable **Generate Anomaly Data**. The anomaly data will be stored in the specified database based on the settings.

Figure 5-174 Anomaly Data Output Settings



The parameters are as follows:

- **Database/Schema:** database or schema that stores anomaly data
- **Table Prefix:** prefix of the table that stores anomaly data
- **Table Suffix:** suffix of the table that stores anomaly data

Click  to save the settings.

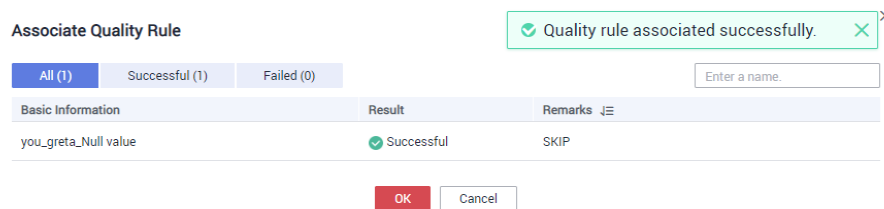
Step 7 (Optional) By default, the quality rule applies to the entire table. If you want to query data in specified partitions, set the where condition.

Figure 5-175 Where condition



Step 8 View the association result. If the association is successful, click **OK**. If the association fails, find the failure cause, correct it, and associate the quality rule again.

Figure 5-176 Association results




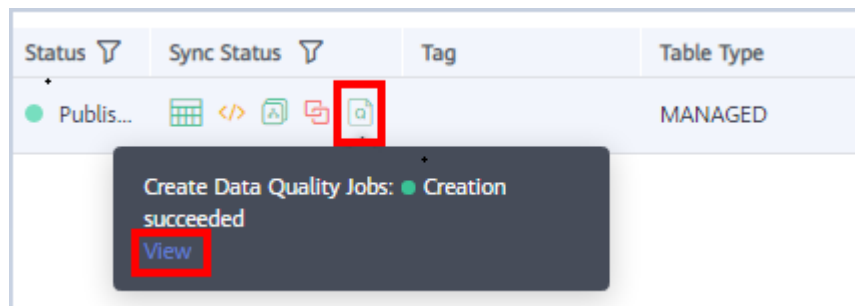
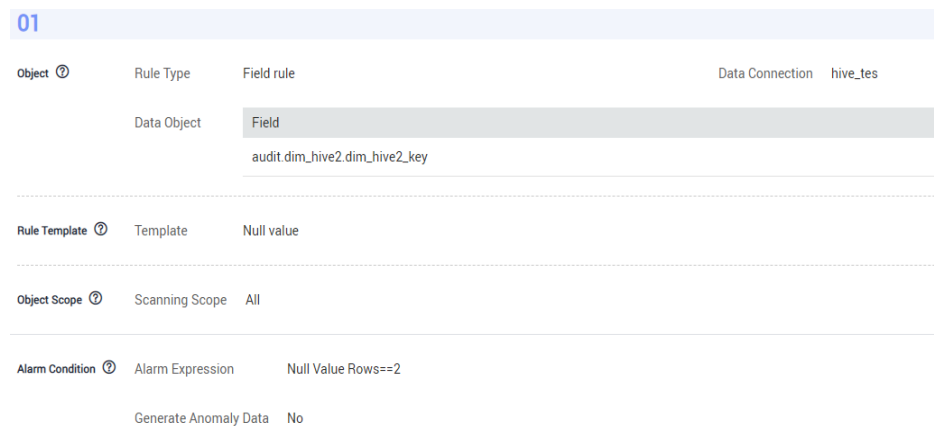
Step 9 Go back to the ER model list, locate the table that you just associated with a quality rule. In the **Sync Status** column, move your pointer to  and click **View**.

Figure 5-177 Quality job sync status



Step 10 On the page displayed, click the **Rule Configurations** tab to view the rule you just added.

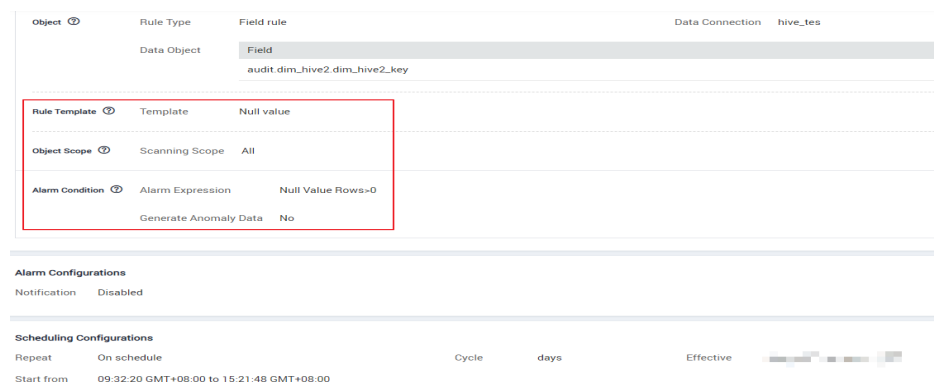
Figure 5-178 Quality rules



If a table is associated with a data standard when it is created, the corresponding quality rule is generated after the table is published. You can also view the rule on the **Quality Jobs** page.

The following provides an example of the quality rule generated based on the data standard associated with a field:

Figure 5-179 Quality rule associated with a field



The following provides an example of the quality rule generated based on the data standard associated with a lookup table:

Figure 5-180 Quality rules for data standards

Object	Rule Type	Table rule	Data Connection	hive_tes
	Data Object	Data Table dim_hive2 audit_dim_dtl20200917182915 audit_hive1		
Rule Template	Template	Table rows		
Object Scope	Scanning Scope	All		
Alarm Condition	Alarm Expression	Table Rows>0		
	Generate Anomaly Data	No		

Alarm Configurations	
Notification	Disabled

Scheduling Configurations					
Repeat	On schedule	Cycle	days	Effective	Sep 22,202 to Nov 17,202
Start from	09:32:00 GMT+08:00 to 23:59:59 GMT+08:00				

----End

5.9.5 Viewing Tables

Tables in an ER model can be displayed in the model view or list view. You can view table details, relationship diagrams, and publish history, as well as preview SQL statements.

Querying the Model View


After creating a table in an ER model, you can query the table models in the list view or model view. The created tables are displayed in the list view by default. You can switch to the model view if you like.

- Step 1** On the DataArts Architecture page, choose **Models > ER Modeling** in the left navigation pane.
- Step 2** In the ER model tree, select a model other than the OBS type, expand it, and select an object.
- Step 3** On the **ER Modeling** page, all created tables are displayed in the list view by default. You can click the **Model View** in the upper right corner on the **ER Modeling** page to change the view mode. You can click **List View** to switch back to the table list.

Figure 5-181 Model view



The following functions are supported in the model view:



- Double-click a table name to view the table details.
- Click **Export** in the upper left corner to export the model view as an image.
- Enter a table name in the search box in the upper right corner to quickly find the table you want to view.
-  represents zoom in, zoom out, full screen, switch between physical and logical models, refresh, and canvas display, respectively.

----End

Viewing Table Details and Previewing an SQL Statement

- Step 1** On the DataArts Architecture page, choose **Models > ER Modeling** in the left navigation pane.
- Step 2** On the model overview page, click the card of the required logical model, and select a subject in the subject directory. All tables under the subject are displayed in the list on the right.
- Step 3** In the table list, select a table, and choose **More > Preview SQL** in the **Operation** column to preview or copy the SQL statement. Then, click **OK** to return to the previous page.

Figure 5-182 ER model list

Table Type	Modified 	Owner	Operation
MANAGED	Aug 10, 2020 10:5...		Edit Publish More 

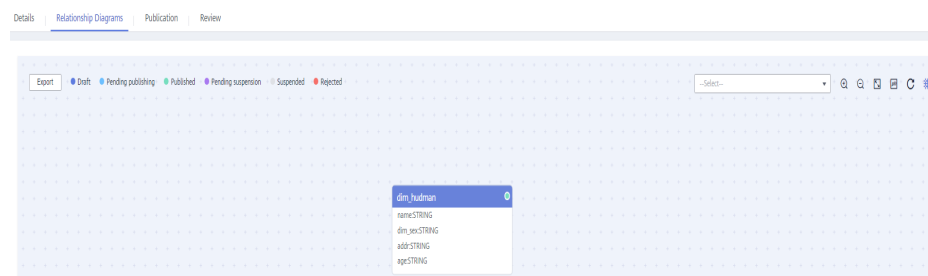
Suspend

View History

Preview SQL

- Step 4** In the table list, click a table name to access the table details page and view the table details, relationship diagrams, publish history, and review history.

Figure 5-183 Relationship diagrams



----End

Viewing Publish History

After a table is published, you can view its publish history, version comparisons, and publish logs. If a table fails to be published, or a data asset or data quality job fails to be synchronized, you can view the publish log to troubleshoot the fault and publish or synchronize it again.

- Step 1** On the DataArts Architecture page, choose **Models > ER Modeling** in the left navigation pane.
- Step 2** On the model overview page, click the card of the required logical model, and select a subject in the subject directory. All tables under the subject are displayed in the list on the right.
- Step 3** In the table list, locate the target table, and choose **More > View History**. On the page displayed, you can view the table publish history, version comparisons, and publish logs.

Figure 5-184 Viewing publish history

Table Type	Modified	Owner	Operation
MANAGED	Aug 10, 2020 10:5...		Edit Publish More
			Suspend
			View History
			Preview SQL

----End

5.9.6 Modifying Subjects, Directories, and Processes

Modifying Subjects in Batches

Currently, only subjects of information architectures, ER models, dimensions, fact tables, summary tables, and technical metrics can be modified in batches. The modification procedure is similar.

This section describes how to modify the subject of information architecture in batches.

- Step 1** On the DataArts Architecture page, choose **Information Architecture** in the left navigation pane.
- Step 2** On the page displayed, select the targets whose subjects need to be modified, and choose **More > Modify Subject**. Modify the subject, and click **OK**.

Figure 5-185 Modifying subjects



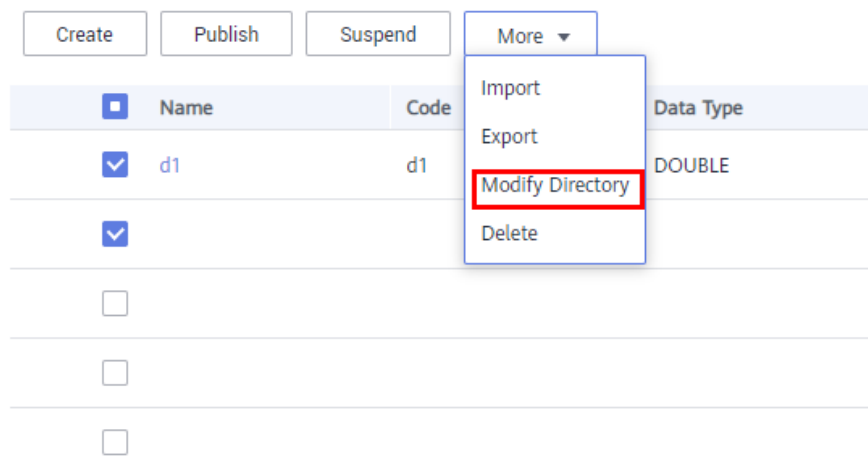
----End

Modifying Directories in Batches

Currently, only directories of lookup tables and data standards can be modified in batches.

- Step 1** On the DataArts Architecture page, choose **Standards > Lookup Tables** or **Standards > Data Standards** in the left navigation pane.
- Step 2** On the page displayed, select the targets whose directories need to be modified, and choose **More > Modify Directory**.

Figure 5-186 Modifying directories of lookup tables in batches



----End

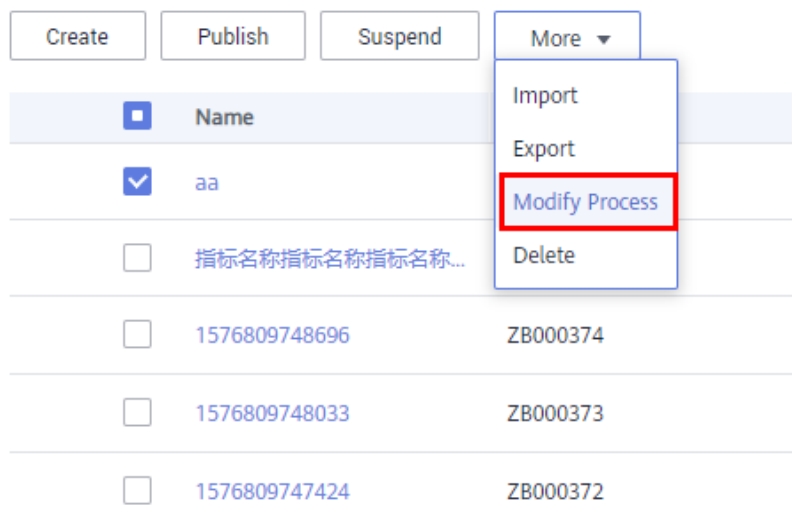
Modifying Processes in Batches

Currently, only the processes of business metrics can be modified in batches.

- Step 1** On the DataArts Architecture page, choose **Metrics > Business Metrics** in the left navigation pane.

Step 2 On the page displayed, select the metrics whose processes need to be modified, and choose **More > Modify Process**.

Figure 5-187 Modifying processes



----End

5.9.7 Review Center

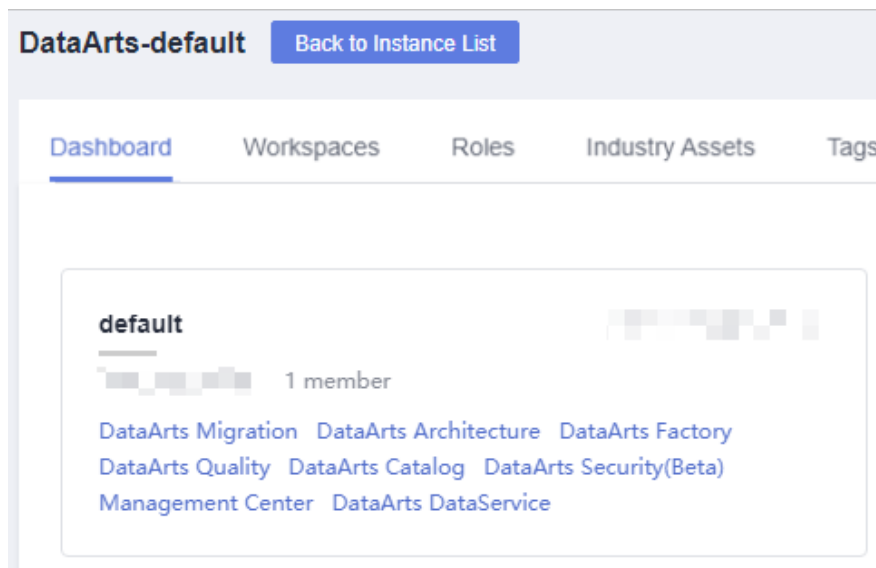
After the modeling and data processing tasks generated in the development environment are submitted, they are stored in the review center. After the tasks are approved on the **Review Center** page, these tasks are available in the production environment.

Reviewer's Audit Objects

If you are a reviewer, use the reviewer account with caution.

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

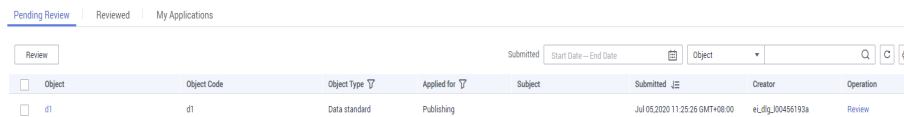
Figure 5-188 DataArts Architecture



2. Choose **Metrics > Review Center** in the left navigation bar, click the **Pending Review** tab, find the object to be reviewed in the list, and click **Review** on the right.

You can also select multiple objects to be reviewed and click **Review** in the upper left corner to review them in batches.

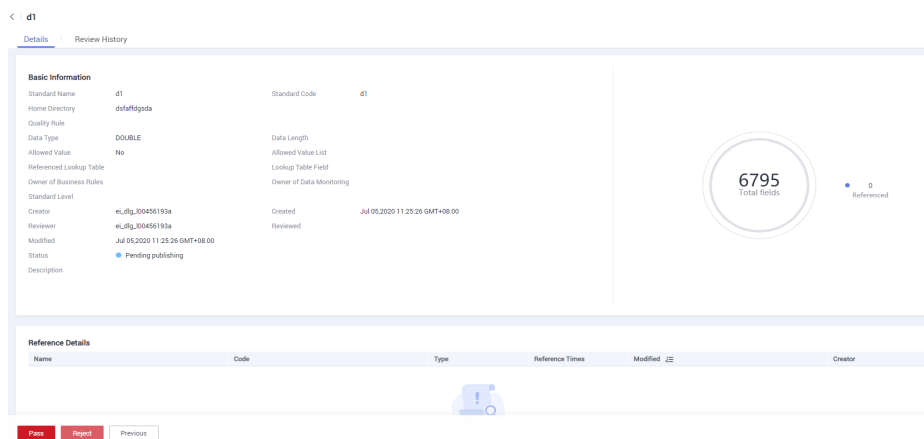
Figure 5-189 Pending Review tab page



3. On the page displayed, confirm the information and click **Accept**. In the dialog box displayed, enter the review comments and click **OK**.

If the information is incorrect, click **Reject**. In the dialog box displayed, enter the reasons for rejecting the application and click **OK**.

Figure 5-190 Review Information area



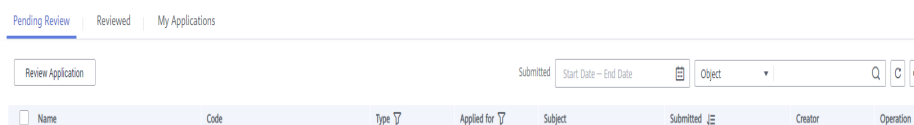
Pending Review, Reviewed, and My Applications Tab Pages

- Pending Review** tab page
 On the DataArts Architecture console, choose **Metrics > Review Center** in the left navigation pane and click the **Pending Review** tab. On the page displayed, you can view the applications to be reviewed.
- Reviewed** tab
 On the DataArts Architecture console, choose **Metrics > Review Center** in the left navigation pane and click the **Reviewed** tab. On the page displayed, you can view the applications that have been approved.
- My Applications** tab
 On the DataArts Architecture console, choose **Metrics > Review Center** in the left navigation pane and click the **My Applications** tab. On the page displayed, you can view the applications that you have submitted.

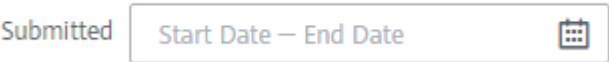



Pending Review

- Step 1** On the DataArts Architecture console, choose **Metrics > Review Center** in the left navigation pane. The **Pending Review** tab page is displayed by default.

Figure 5-191 Pending Review tab page



Function Area	Description
1	Batch Review 1. Select multiple pieces of information to be reviewed. 2. Click Review Application . 3. In the dialog box displayed, enter the valid review comments. 4. Click Accept to approve the selected targets in batches, or click Reject to reject the selected targets in batches.
2	Single Review 1. Click Review in the Operation column. The page for reviewing the information is displayed. 2. Select the review result and enter valid review comments based on service requirements. 3. Click OK .

Function Area	Description
3	<ul style="list-style-type: none">  allows you to specify a time range during which the information to be viewed is displayed.  allows you to query the to-be-reviewed information about objects and creators.  allows you to set the headers of tables to be reviewed.  allows you to refresh the current page.

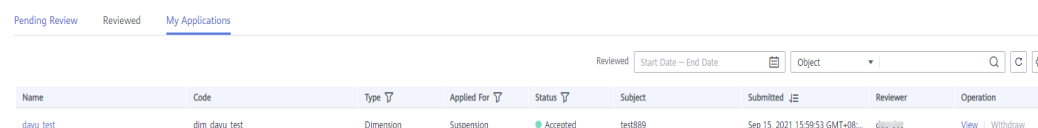
----End

My Applications

Step 1 On the DataArts Architecture console, choose **Metrics > Review Center** in the left navigation pane.

Step 2 Click **My Applications**.

Figure 5-192 My Applications tab page



You can perform the following operations:

- Click **View** in the **Operation** column to view information about a specified row.
- Click **Withdraw** in the **Operation** column to withdraw the application.

----End

5.10 Tutorials

5.10.1 DataArts Architecture Example

DataArts Architecture can be used to create entity-relationship (ER) models and dimensional models to standardize and visualize data development and output data governance methods that can guide development personnel to work with ease.

This section covers the following scenarios:

- Design a data model for the taxi travel data in an MRS Hive data lake.
- The original taxi travel data table **sdi_taxi_trip_data** is stored in the **demo_sdi_db** database.
- The following table lists the data fields in the original data table **sdi_taxi_trip_data**.

The following table lists the taxi trip data:

Table 5-52 Taxi trip data

No.	Field Name	Field Description
1	VendorID	Vendor ID. Possible values are: 1=A Company 2=B Company
2	tpep_pickup_datetime	Time when a passenger gets on a taxi.
3	tpep_dropoff_datetime	Time when a passenger gets off a taxi.
4	passenger_count	Number of passengers.
5	trip_distance	Driving distance.
6	ratecodeid	Charge rate code. Possible values are: 1=Standard rate 2=JFK 3=Newark 4=Nassau or Westchester 5=Negotiated fare 6=Group ride
7	store_fwd_flag	Store-and-forward flag.
8	PULocationID	Location at which a passenger gets on a taxi.
9	DOLocationID	Location at which a passenger gets off a taxi.

No.	Field Name	Field Description
10	payment_type	Payment type. Possible values are: 1=Credit card 2=Cash 3=No charge 4=Dispute 5=Unknown 6=Voided trip
11	fare_amount	Fare amount.
12	extra	Extra fee.
13	mta_tax	MTA tax.
14	tip_amount	Tip amount.
15	tolls_amount	Toll amount.
16	improvement_surcharge	Improvement surcharge.
17	total_amount	Total amount.

The process of using DataArts Architecture is as follows:

1. Preparations

- **Add reviewers:** In the DataArts Architecture module, all business processes must be approved. Therefore, add reviewers first before conducting any operations. Only the workspace admin has the permissions required to add reviewers.
- **Configuration Center** provides abundant custom options. You can customize the configuration to meet your demands.

2. Data Survey: A data survey involves collecting data that is generated when sorting business requirements, creating business processes, and classifying data subjects based on the existing business data and industry status.

- **Subject design** is a hierarchical architecture that classifies and defines data to help clarify data assets and specify relationships between business domains and business objects.
- **Process design:** This example does not contain this. Process design is to generate a structured framework of data processing process, including the categories, levels, boundaries, scope, and input/output relationships, and reflect the business models and characteristics of your enterprise.

3. Standards: Create lookup tables and data standards.

- **Create and publish a lookup table:** A lookup table includes a series of allowed values and additional text descriptions that are generally associated with data standards to generate a range of values for the verification of quality monitoring rules.

- **Create and publish a data standard:** A data standard refers to the description of attribute data meanings and business rules that enterprises must comply with. It describes the common understanding of certain data at the company level.
- 4. **Models:** Use ER modeling and dimensional modeling methods to perform hierarchical modeling.
 - **ER modeling: Create a model at the SDI and DWI layers, respectively.**
 - **SDI** stands for Source Data Integration and is the source data layer. SDI is a simple implementation of source system data.
 - **DWI** stands for Data Warehouse Integration, also called the data consolidation layer. DWI integrates and cleans data from multiple source systems, and implements entity relationship modeling based on the three normal forms.
 - **Dimensional modeling: Create and publish a dimension at the DWR layer. & Creating and Publishing a Fact Table for the DWR Layer.**
 - **Data Warehouse Report (DWR)** is based on the multi-dimensional model and its data granularity is the same as that of the DWI layer.
 - **Dimension** is the perspective to observe and analyze business data and assist in data aggregation, drilling, slicing, and analysis, and used as a GROUP BY condition in SQL statements.
 - A **fact table** that belongs to a business process can enrich the affair information corresponding to the specific business process.
- 5. **Metric design: Create and publish a technical metric:** Create and publish a business metric (not involved in this example) and a technical metric. Technical metrics are classified into atomic, derivative, and compound metrics.
 - A **metric** consists of its name and value. The metric name and its definition reflect the quality and quantity of the metric. The metric value reflects the quantifiable values of the specified time, location, and condition of the metric.

Business metrics are used to guide technical metrics, and technical metrics are used to implement business metrics.
 - **Atomic metrics** are generated based on dimension tables and fact tables of a multidimensional model. The objects and the finest data granularity of an atomic metric are consistent with those of the multidimensional model.

An atomic metric usually consists of measures and attributes related with measures and business objects, all of which aim to support agile self-service consumption of the metric.
 - **Derivative metrics** are aggregated from the definitions, modifiers, and dimensions of atomic metrics. Therefore, their definitions, modifiers, and dimensions are derived from the attributes of atomic metric associated tables as well.
 - **Compound metrics** are generated by adding one or more derivative metrics. The dimensions and modifiers of a compound metric are the same as those of the derivative metrics.

New dimensions and modifiers cannot be generated outside the scope of derivative metrics, dimensions, and modifiers.

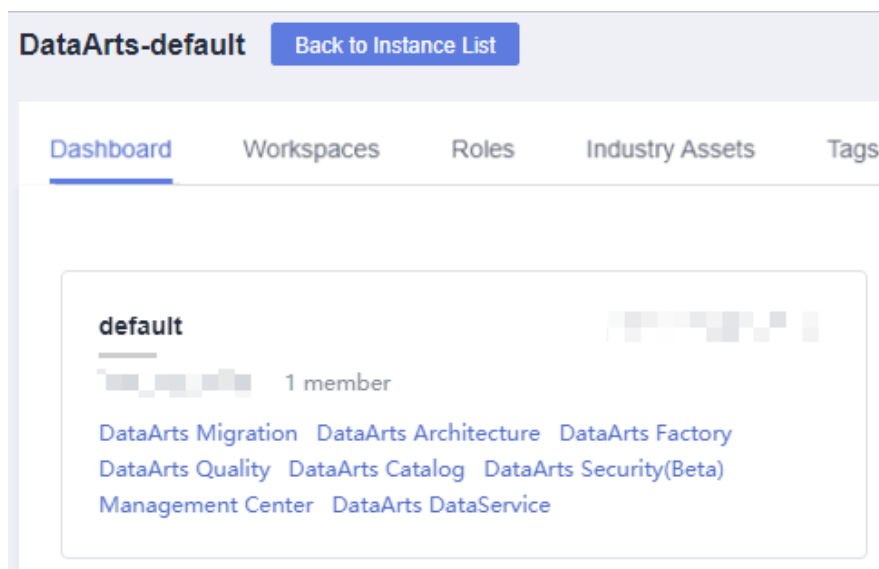
6. **Dimensional modeling: Create and publish a summary table at the DM layer.**
 - **Data Mart (DM)** is where multiple types of data are summarized. DM is designed to display the summarized data.
 - A **summary table** consists of specific analysis objects (for example, members) and related statistical metrics. The statistical metrics included in a summary table have the same statistical granularity (for example, members). The summary table provides users with all statistics-granularity-themed data (such as a member theme market).

Adding Reviewers

In the DataArts Architecture module, all modeling steps must be reviewed. Therefore, you need to add a reviewer first. DAYU Administrator or the workspace administrator has the permission to add reviewers.

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Architecture**.

Figure 5-193 DataArts Architecture



2. In the navigation pane on the left, choose **Configuration Center**. On the displayed **Reviewers** page, click **Add**.
3. Select a reviewer (administrator or developer), enter the correct email address and phone number, and click **OK**.

You can also add your current account as a reviewer. In this way, auto review is supported in subsequent operations. Add more reviewers, if required.

Figure 5-194 Adding a reviewer

Add Reviewer ×

* Reviewer ↻

A reviewer must be a member with the review permissions in the current workspace. Only admins and developers have the review permissions. You can view and edit workspace members on the Workspaces tab page of the home page.

Notification Type SMS Email
A small fee may be generated for SMS or email notifications. [Details](#)

* Phone Number

Format: country/region code-mobile number. If the country/region code is not specified, the default value 86 is used.

* Email Address

OK Cancel

Configuration Center

DataArts Architecture configuration center provides abundant custom options. You can customize the configuration to meet your demands.

1. On the DataArts Architecture page, choose **Configuration Center** in the left navigation pane.
2. Click the **Functions** tab and configure functions as needed.

Figure 5-195 Functions

Configuration Center

Reviewers Subject Levels Standard Templates **Functions** Models Data Types DDL Templates Metric Encoding Rules

Model Design Process Create tables Synchronize technical assets Synchronize logical assets Associate assets Create data quality jobs Create data development jobs Publish DLM APIs

Model Suspension Process Delete technical assets Delete logical assets Delete data quality jobs Delete data development jobs

Data Table Update Mode No update DDL-based update Drop and create

Case Inconsistent During Technical Assets Synchron DUI DWS MRS_HIVE POSTGRESQL

Use New UI to Deliver Business Table Mappings

Auto Aggregate Summary Tables

Data Standard Allows Duplicate Names

Parallel Queried Tables on Information Architecture

Concurrently Insertable Lines of Data

Lookup Table-based Quality Rule

Generate Data Lake Mail APIs

3. Click **OK**.

Designing a Subject

This section uses the subjects listed in [Table 5-53](#) as an example.

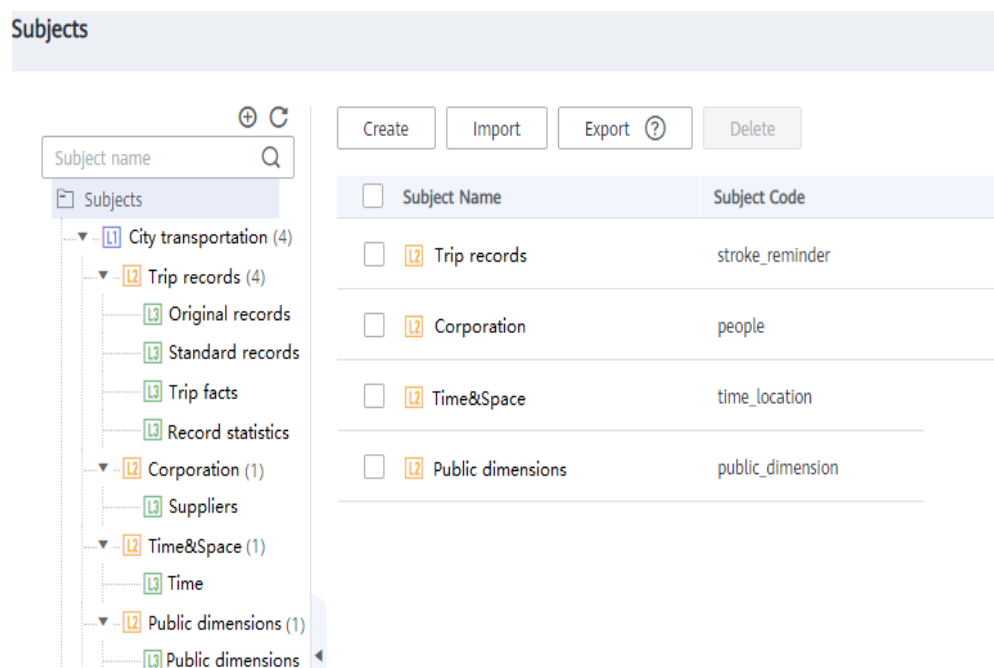
- There is a subject area group named **City transportation**.
- Under **City transportation**, there are four subject areas: **Trip records**, **Corporation**, **Time&Space**, and **Public dimensions**.
- Under **Trip records**, there are four business objects: **Original records**, **Standard records**, **Trip facts**, and **Record statistics**.

- Under **Corporation**, there is one business object: **Suppliers**.
- Under **Time&Space**, there is one business object: **Time**.
- Under **Public dimensions**, there is one business object: **Public dimensions**.

Table 5-53 Subject design

Subject Area Group Name (L1)	Subject Area Group Code (L1)	Subject Area Name (L2)	Subject Area Code (L2)	Business Object Name (L3)	Business Object Code (L3)
City transportation	city_traffic	Trip records	stroke_reminder	Original records	origin_stroke
				Standard records	stand_stroke
				Trip facts	stroke_fact
				Record statistics	stroke_statistic
		Corporation	people	Suppliers	vendor
		Time&Space	time_location	Time	date
		Public dimensions	public_dimension	Public dimensions	public_dimension

Figure 5-196 Designing a subject

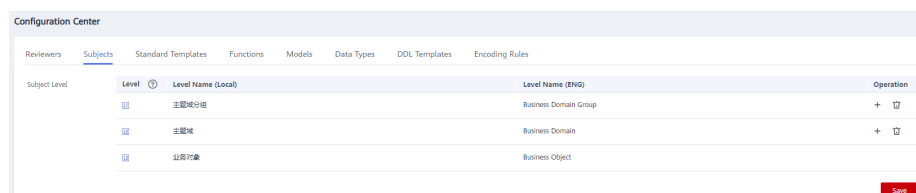


Procedure

- Step 1** Log in to the DataArts Studio console. Locate the created DataArts Studio instance and click **Access**.
- Step 2** In the workspace list, locate the target workspace and click **DataArts Architecture**.
- Step 3** Choose **Configuration Center** in the navigation pane on the left. Click the **Subject Levels** tab, and use the default three levels.

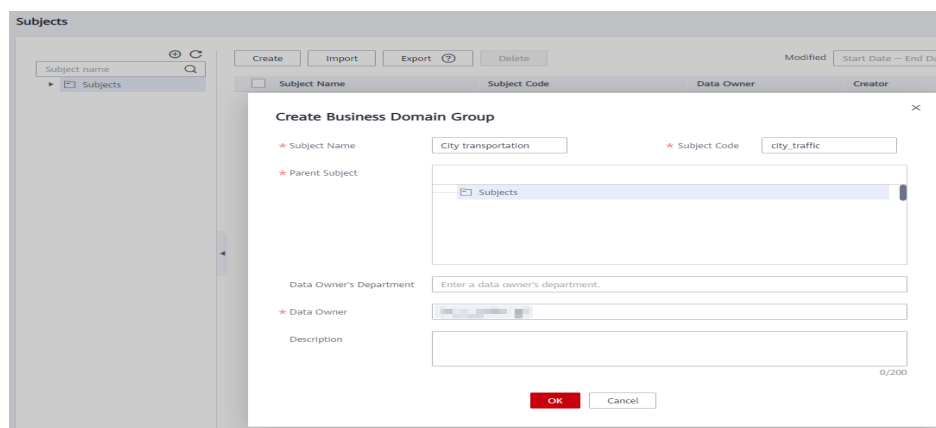
There can be a maximum of seven subject levels, a minimum of two subject levels, and three subject levels by default. L1 to L7 are used to represent the layers. The last level is **Business Object** and cannot be customized. The names of other levels can be customized. The levels configured in **Configuration Center** take effect on the **Subjects** page.

Figure 5-197 Configuring the subject levels



- Step 4** On the DataArts Architecture console, choose **Data Survey > Subjects** in the left navigation pane. On the page displayed, click **Create** to create an L1 subject, which is a subject area group.

Figure 5-198 Creating an L1 subject



In the dialog box displayed, set the parameters as shown in [Figure 5-198](#) and click **OK**.

- Step 5** Create four L2 subjects under the L1 subject **City transportation: Trip records, Corporation, Time&Space, and Public dimensions**.

Perform the following procedure to create a subject area named **Trip records**. The procedure for creating other subject areas is similar.

1. Right-click the L1 subject **City transportation** in the subject tree, and select **Create** from the shortcut menu. Alternatively, click **Create** in the right pane.

Figure 5-199 Creating an L2 subject

2. In the dialog box displayed, set **Subject Name** and **Subject Code** to the values of **Subject Area Name** and **Subject Area Code** in [Table 5-53](#), set other parameters based on project requirements, and click **OK**.

Step 6 Create business objects.

- Under **Trip records**, create four business objects: **Original records**, **Standard records**, **Trip facts**, and **Record statistics**.
- Under **Corporation**, create one business object: **Suppliers**.
- Under **Time&Space**, create one business object: **Time**.
- Under **Public dimensions**, create one business object: **Public dimensions**.

Perform the following procedure to create a business object named **Original records** in the subject area **Trip records**. The procedure for creating other business objects is similar.

1. Right-click the L2 subject **Trip records** in the subject tree, and select **Create** from the shortcut menu. Alternatively, click **Create** in the right pane.
2. In the dialog box displayed, set **Subject Name** and **Subject Code** to the values of **Business Object Name** and **Business Object Code** in [Table 5-53](#), set other parameters based on project requirements, and click **OK**.

----End

Creating and Publishing Lookup Tables

This section uses the lookup tables listed in [Table 5-54](#) as an example.

Table 5-54 Lookup tables

Directory	*Table Name	* Table Code	Table Description	* Field Name	* Field Code	* Data Type	Field Description
payment_type	payment_type	payment_type	None	payment_type_id	payment_type_id	BIGINT	None
				payment_type_value	payment_type_value	STRING	None
vendor	vendor	vendor	None	vendor_id	vendor_id	BIGINT	None
				vendor_value	vendor_value	STRING	None
rate	rate_code	rate_code	None	rate_code_id	rate_code_id	BIGINT	None
				rate_code_value	rate_code_value	STRING	None

Procedure

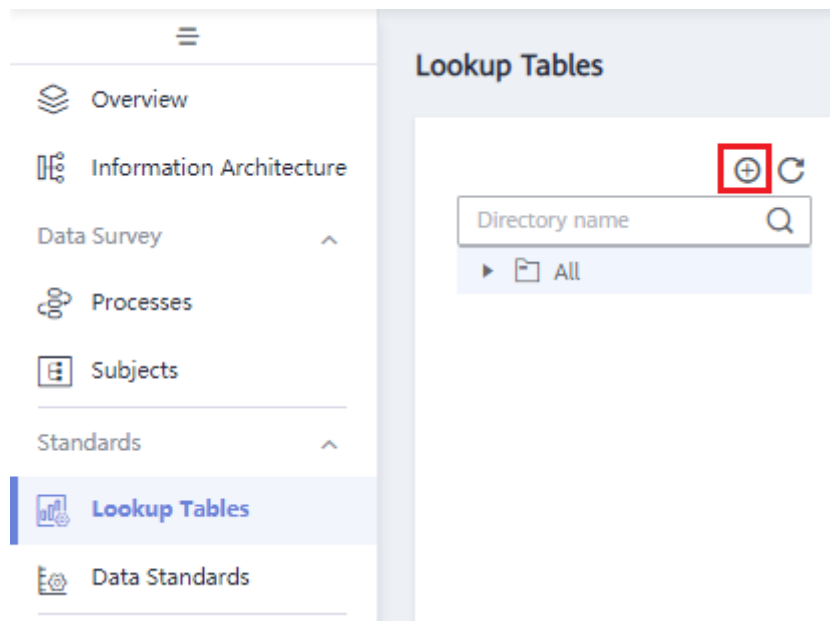
Step 1 On the DataArts Architecture console, choose **Standards > Lookup Tables** in the navigation pane on the left.

Step 2 Create three lookup table directories: **payment_type**, **vendor**, and **rate**.

Perform the following procedure to create a directory named **payment_type**. The procedure for creating other directories is similar.

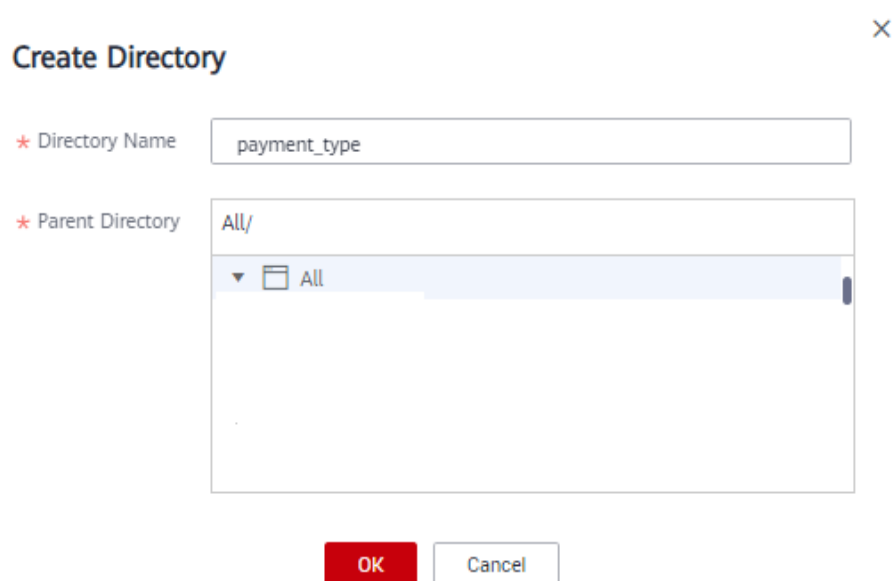
1. On the **Lookup Tables** page, click  above the directory tree to create a directory.

Figure 5-200 Lookup table directory tree



2. In the dialog box displayed, enter a directory name, select a parent directory, and click **OK**.

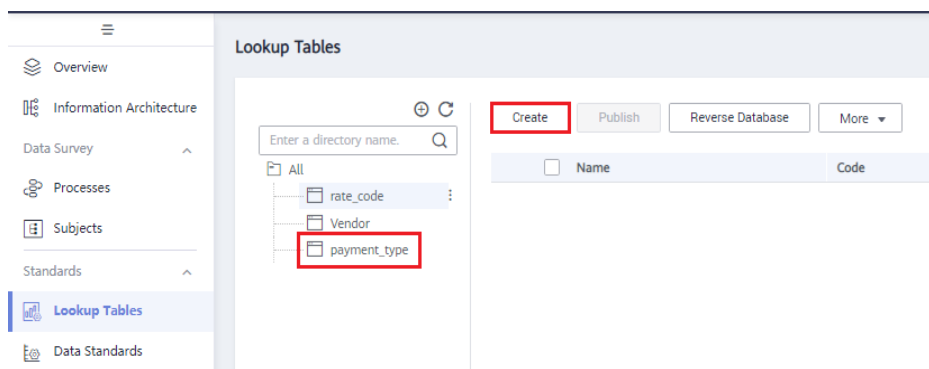
Figure 5-201 Creating a directory for lookup tables



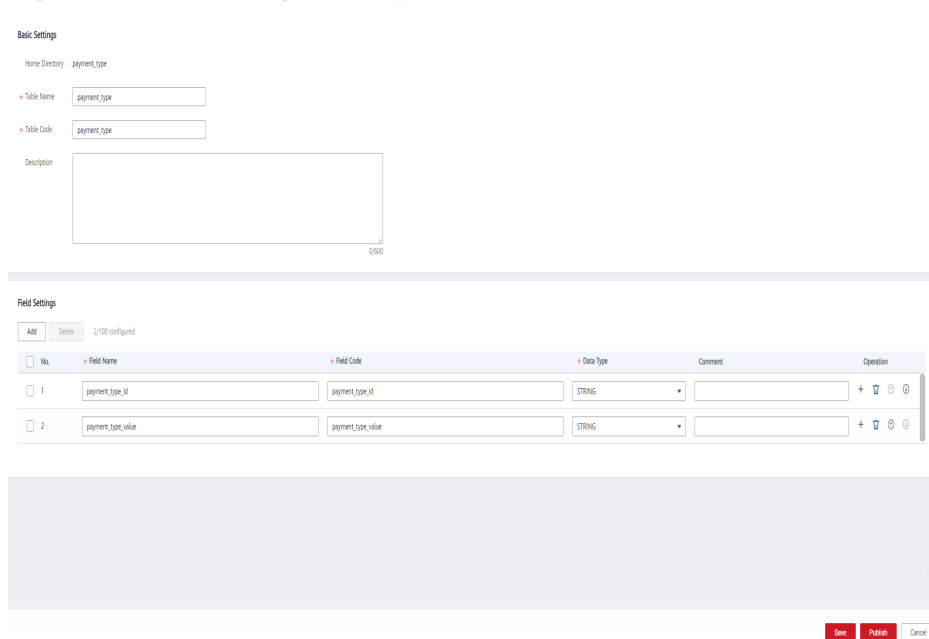
Step 3 Create three lookup tables: **payment_type**, **vendor**, and **rate_code**.

Perform the following procedure to create a lookup table named **payment_type**. The procedure for creating other lookup tables is similar.

1. On the **Lookup Tables** page, click **payment_type** in the directory tree, and click **Create** on the page displayed.

Figure 5-202 Lookup Tables page

2. Set the parameters based on [Table 5-54](#) and click **Save**.

Figure 5-203 Creating a lookup table

3. Refer to [Step 3.1](#) to [Step 3.2](#) to create the lookup table **vendor** in the **vendor** directory and the lookup table **rate_code** in the **rate** directory.

Figure 5-204 Creating a lookup table named vendor

The screenshot shows the 'Basic Settings' and 'Field Settings' sections for a new lookup table named 'vendor'.

Basic Settings:

- Home Directory: vendor
- Table Name: vendor
- Table Code: vendor
- Description: (empty text area)

Field Settings:

Buttons: Add, Delete, 2/100 configured

No.	Field Name	Field Code	Data Type	Comment	Operation
1	vendor_id	vendor_id	STRING		+ [trash] [refresh]
2	vendor_value	vendor_value	STRING		+ [trash] [refresh]

Buttons at the bottom: Save, Publish, Cancel

Figure 5-205 Creating a lookup table named rate_code

The screenshot shows the 'Table Details' and 'Field Inputs' sections for a new lookup table named 'rate_code'.

Table Details:

- Home Directory: rate_code
- Table Name: rate_code
- Table Code: rate_code
- Description: (empty text area)

Field Inputs:

Buttons: Add, Delete, 2/100 configured

No.	Name	Code	Data Type	Comment	Operation
1	rate_code_id	rate_code_id	BIGINT		+ [trash] [refresh]
2	rate_code_value	rate_code_value	STRING		+ [trash] [refresh]

Buttons at the bottom: Save, Publish, Cancel

Step 4 Enter values for the three lookup tables **payment_type**, **vendor**, and **rate_code**.

On the **Lookup Tables** page, locate the row that contains the lookup table **payment_type**, and choose **More > Manage Value** in the **Operation** column. On the page displayed, click **Add** to add the values listed in [Table 5-55](#).

Table 5-55 Values to be added for the lookup table payment_type

payment_type_id	payment_type_value
1	Credit card
2	Cash
3	No charge
4	Dispute
5	Unknown
6	Voided trip

Return to the **Lookup Tables** page, locate the row that contains the lookup table **vendor**, and choose **More > Manage Value** in the **Operation** column. On the page displayed, click **Add** to add the values listed in [Table 5-56](#).

Table 5-56 Values to be added for the lookup table vendor

vendor_id	vendor_value
1	A Company
2	B Company

Return to the **Lookup Tables** page, locate the row that contains the lookup table **rate_code**, and choose **More > Manage Value** in the **Operation** column. On the page displayed, click **Add** to add the values listed in [Table 5-57](#).

Table 5-57 Values to be added for the lookup table rate_code

rate_code_id	rate_code_value
1	Standard rate
2	JFK
3	Newark
4	Nassau or Westchester
5	Negotiated fare
6	Group ride

Step 5 Return to the **Lookup Tables** page, select the three lookup tables, and click **Publish**.

Step 6 In the **Apply for Publication** dialog box, select a reviewer and click **OK**. Wait for the reviewer to review the application. If you have the reviewer permissions, select **Auto-review** and click **Submit**.

----End

Creating and Publishing Data Standards

In this example, you need to create the three data standards listed in [Table 5-58](#).

Table 5-58 Data standards

Directory	*Standard Name	*Standard Code (Custom)	*Data Type	Data Length	Lookup Table	*Lookup Table Field	Description
payment_type	payment_type	payment_type	Long integer (BIGINT)	None	payment_type	payment_type_id	None
vendor	vendor	vendor	Long integer (BIGINT)	None	vendor	vendor_id	None
rate	rate_code	rate_code	Long integer (BIGINT)	None	rate_code	rate_code_id	None

Step 1 On the DataArts Architecture console, choose **Standards > Data Standards** in the navigation pane on the left.

Step 2 If you access the Data Standards page for the first time, you must customize a template. The custom template can be modified in Configuration Center. Additionally, select **Lookup table**, as shown in the following figure.

Figure 5-206 Customize Template

Default	Field	Searchable	Mandatory
	Standard name	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
	Standard code	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
	Data type	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Optional	Field	Searchable	Mandatory
	<input checked="" type="checkbox"/> Data length	<input type="checkbox"/>	<input type="checkbox"/>
	<input checked="" type="checkbox"/> Allowed value exist	<input type="checkbox"/>	<input type="checkbox"/>
	<input checked="" type="checkbox"/> Allowed values	<input type="checkbox"/>	<input type="checkbox"/>
	<input checked="" type="checkbox"/> Lookup table	<input type="checkbox"/>	<input type="checkbox"/>
	<input checked="" type="checkbox"/> Lookup table field	<input type="checkbox"/>	<input type="checkbox"/>
	<input checked="" type="checkbox"/> Quality rule	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/> Rule designer	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/> Rule implementer	<input type="checkbox"/>	<input type="checkbox"/>
	<input checked="" type="checkbox"/> Standard level	<input type="checkbox"/>	<input type="checkbox"/>
	<input checked="" type="checkbox"/> Description	<input type="checkbox"/>	<input type="checkbox"/>

Step 3 Create three directories for data standards: **payment_type**, **vendor**, and **rate_code**.


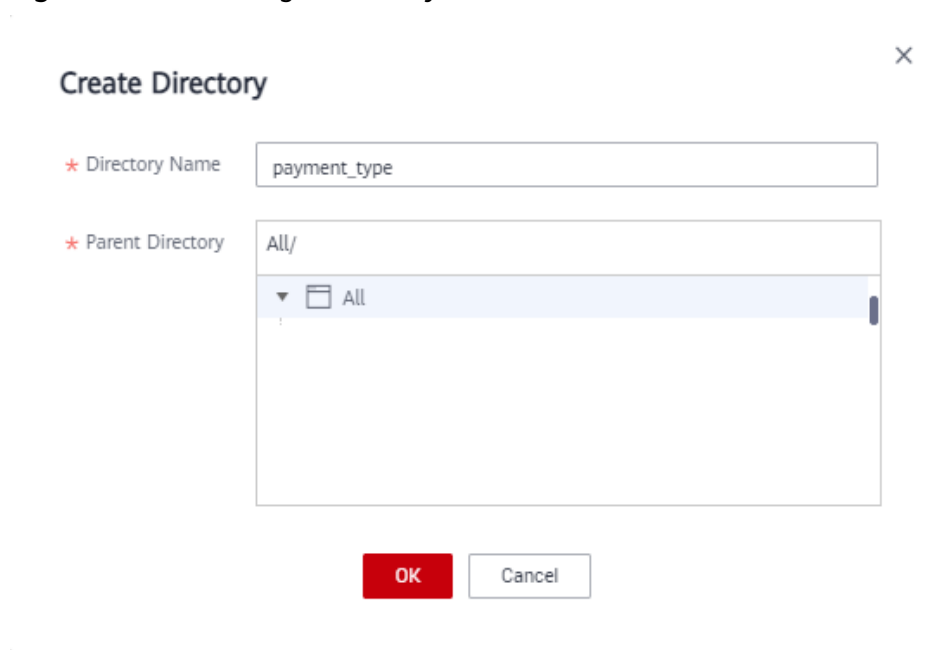
In the upper part of the directory tree on the **Data Standards** page, click . In the dialog box displayed, enter the directory name as **payment_type**, select a parent directory, and click **OK**.

Figure 5-207 Creating a directory for data standards



Step 4 Create three data standards: **payment method**, **Suppliers**, and **rate code**.

1. In the directory tree on the **Data Standards** page, select the required directory and click **Create** on the page displayed on the right.
2. On the **Create Data Standard** page, configure the three data standards by referring to the following figures, and click **Save**. In this example, only a few parameters are selected for the data standard template. You can customize a data standard template by referring to [Configuration Center](#).

Figure 5-208 Creating a data standard named payment method

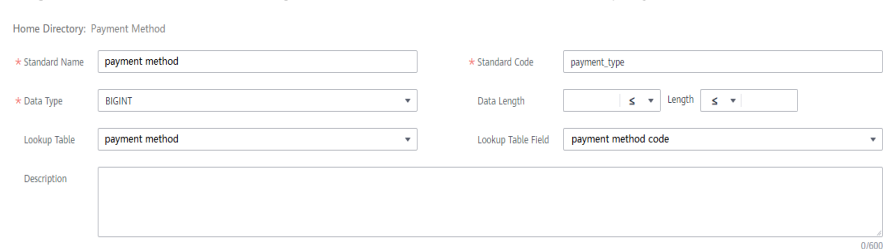


Figure 5-209 Creating a data standard named Suppliers

Home Directory: Suppliers

* Standard Name: * Standard Code:

* Data Type: Data Length: Length:

Lookup Table: Lookup Table Field:

Description:

0/600

Figure 5-210 Creating a data standard named rate code

Home Directory: Rate

* Standard Name: * Standard Code:

* Data Type: Data Length: Length:

Lookup Table: Lookup Table Field:

Description:

0/600

Step 5 Return to the **Data Standards** page, select the three data standards in the list, and click **Publish**.

Step 6 In the **Apply for Publication** dialog box, select a reviewer and click **OK**. Wait for the reviewer to review the application. If you have the reviewer permissions, select **Auto-review** and click **Submit**.

----End

ER Modeling: Creating a Model at the SDI and DWI Layers Respectively

During ER modeling, create an ER model at the SDI and DWI layer, respectively, and import the original data table to the ER model at the SDI layer through by reversing the database, and create a standard service table named **standard travel data** in the ER model at the DWI layer.

Step 1 On the DataArts Architecture page, choose **Models > ER Modeling** in the left navigation pane.

- If no ER model has been created, a dialog box is displayed, asking you to create a hierarchical governance model. You can create an SDI ER model named **sdi** and then create a DWI ER model named **dwi**. Click **OK**.

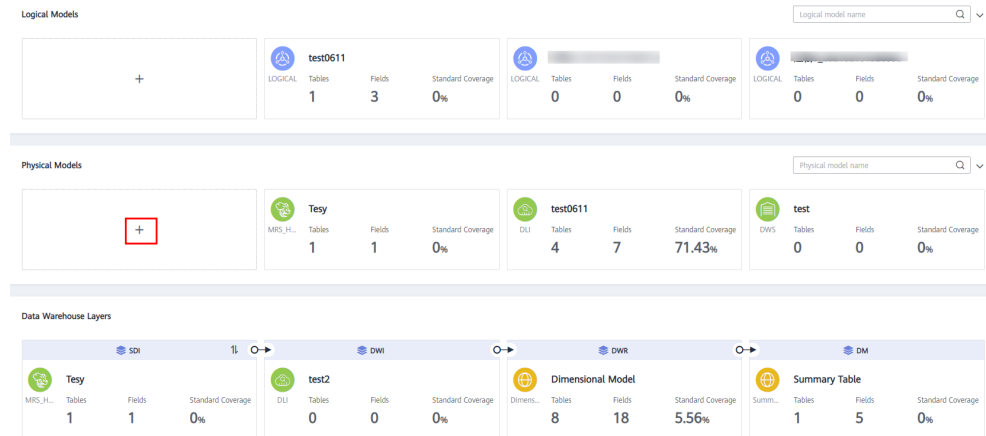
Figure 5-211 Dialog box for creating a hierarchical governance model

Create Hierarchical Governance Model ×

SDI	DWI	DWR
* Model Name: <input type="text" value="Enter a model name."/>	* Model Name: <input type="text" value="Enter a model name."/>	Dimensional model
* Data Connection Type: <input type="text" value="--Select--"/>	* Data Connection Type: <input type="text" value="--Select--"/>	

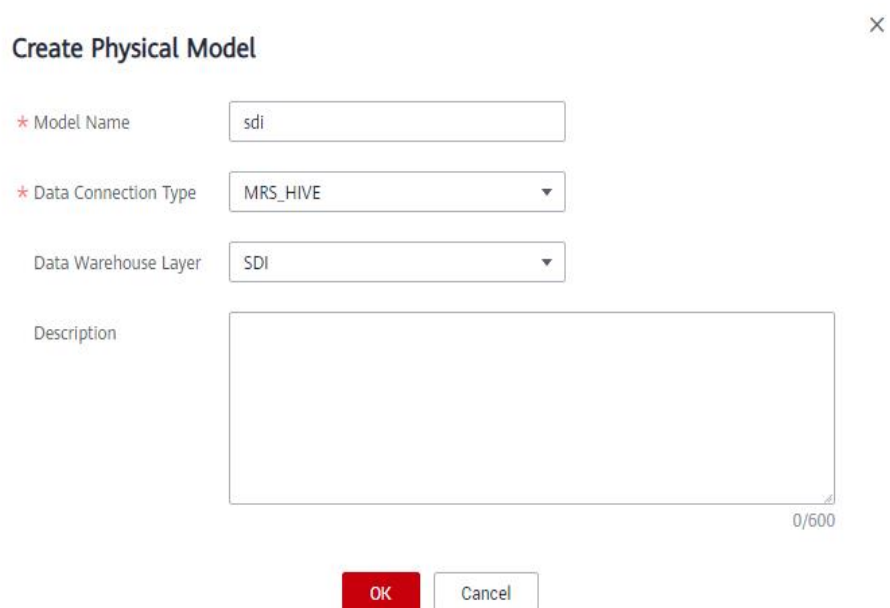
- If you have created ER models before, click **+** to create physical models, as shown in the following figure.

Figure 5-212 ER Modeling page



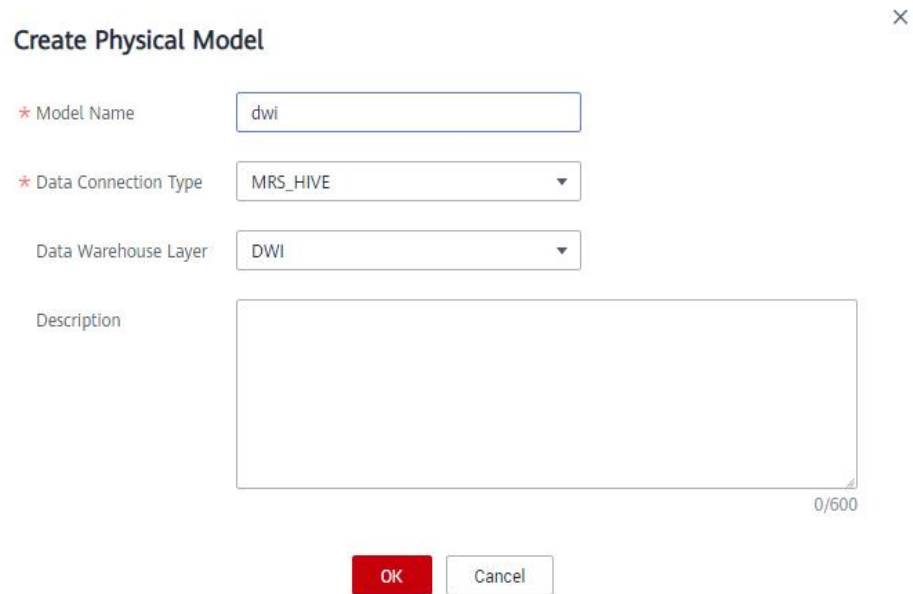
- Create an ER model at the SDI layer named **sdi**. In the **Physical Models** area, click **+**. In the displayed dialog box, configure required parameters and click **OK**.

Figure 5-213 Creating a physical model named sdi



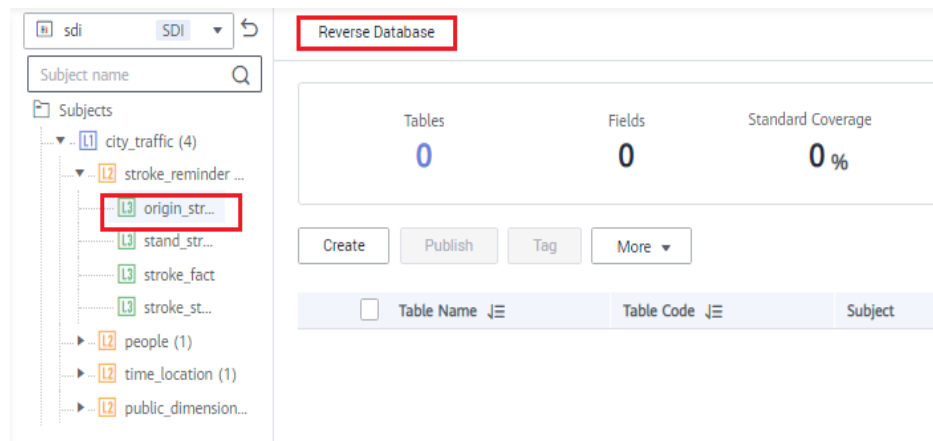
- Create an ER model at the DWI layer named **dwi**. In the **Physical Models** area, click **+**. In the displayed dialog box, configure required parameters and click **OK**.

Figure 5-214 Creating a physical model named dwi



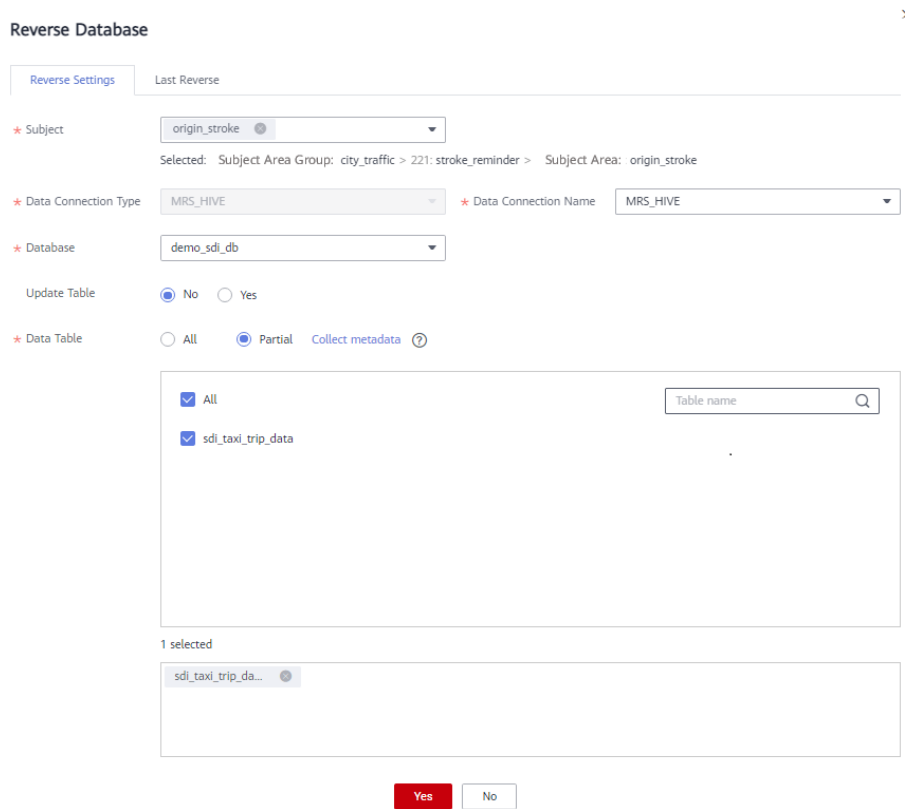
Step 2 In the **Data Warehouse Layers** part, click the newly created SDI ER model. Choose **city_traffic > stroke_reminder > origin_stroke**, and click **Reverse Database** on the page displayed on the right to import the source table.

Figure 5-215 Model directory



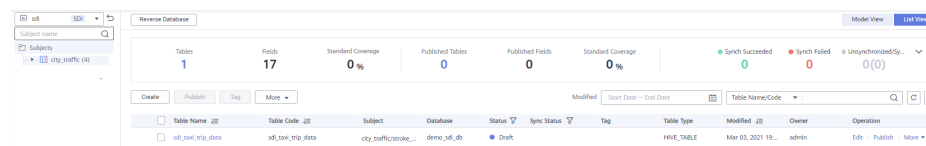
In the **Reverse Database** dialog box, set the parameters and click **OK**. In this example, select the original data table in the source layer database **demo_sdi_db**.

Figure 5-216 Reverse Database dialog box



After the database is reversed successfully, click **Close**. You can view the imported table in the table list.

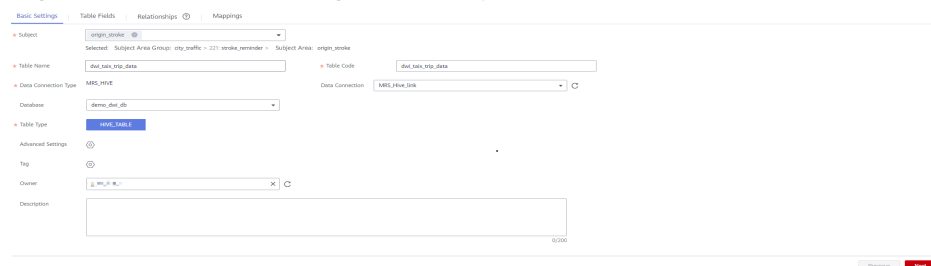
Figure 5-217 Viewing the imported table



Step 3 Perform the following steps to create a standard service table named **standard travel data**:

1. In the **Data Warehouse Layers** part, click the newly created DWI ER model. Choose **city_traffic > stroke_reminder > origin_stroke**, and click **Create** on the page displayed on the right.
2. On the **Basic Settings** page, set the parameters as follows.

Figure 5-218 Basic settings of the trip data table



- Click the **Table Fields** tab and then **Add**. Add the fields listed in [Table 5-59](#).


Then click  in the **Data Standard** column of the rows where the vendor ID, rate code ID, and payment type reside to associate with the **Vendor**, **Rate Code ID**, and **Payment Type** standards, respectively. [Figure 5-219](#) lists the fields to be added.

Table 5-59 Fields in the standard travel data table

N o.	Field Name	Field Code	Data Type	Data Standard	Primary Key	Partition	Not Null	Tag
1	Vendor ID	vendor_id	Long integer (BIGINT)	Vendor	Des elected	Des elected	Sele cted	-
2	Pick up Time	tprep_pickup_datetime	Timestamp (TIMESTAMP)	-	Des elected	Des elected	Sele cted	-
3	Drop-off Time	tprep_dropoff_datetime	Timestamp (TIMESTAMP)	-	Des elected	Des elected	Sele cted	-
4	Passenger Quantity	passenger_count	Character (STRING)	-	Des elected	Des elected	Sele cted	-
5	Trip Distance	trip_distance	High-precision (DECIMAL) (10,2)	-	Des elected	Des elected	Sele cted	-
6	Rate Code	rate_code_id	Long integer (BIGINT)	Rate code	Des elected	Des elected	Sele cted	-
7	Storage Forwarding Flag	store_fwd_flag	Character (STRING)	-	Des elected	Des elected	Sele cted	-

No.	Field Name	Field Code	Data Type	Data Standard	Primary Key	Partition	Not Null	Tag
8	Pick up Location	pu_location_id	Character (STRING)	-	Deslected	Deslected	Selected	-
9	Drop-off Location	do_location_id	Character (STRING)	-	Deslected	Deslected	Selected	-
10	Payment Type	payment_type	Long integer (BIGINT)	Payment type	Deslected	Deslected	Selected	-
11	Fare	fare_amount	High-precision (DECIMAL) (10,2)	-	Deslected	Deslected	Selected	-
12	Extra Fee	extra	High-precision (DECIMAL) (10,2)	-	Deslected	Deslected	Selected	-
13	MTA Tax	mta_tax	High-precision (DECIMAL) (10,2)	-	Deslected	Deslected	Selected	-
14	Handling Fee	tip_amount	High-precision (DECIMAL) (10,2)	-	Deslected	Deslected	Selected	-
15	Toll	tolls_amount	High-precision (DECIMAL) (10,2)	-	Deslected	Deslected	Selected	-
16	Improvement Surcharge	improvement_surcharge	High-precision (DECIMAL) (10,2)	-	Deslected	Deslected	Selected	-

No.	Field Name	Field Code	Data Type	Data Standard	Primary Key	Partition	Not Null	Tag
17	Total Fee	total_amount	High-precision (DECIMAL) (10,2)	-	Deslected	Deslected	Selected	-


Figure 5-219 Fields in the trip data table

No.	Field Name	Field Code	Data Type	Data Standard	Primary Key	Partition	Not Null	Tag	Comment	Operation
1	vendor_id	vendor_id	BIGINT	vendor	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		+ 🗑️ ⚙️
2	trip_pickup_datetime	trip_pickup_datetime	TIMESTAMP		<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		+ 🗑️ ⚙️
3	trip_dropoff_datetime	trip_dropoff_datetime	TIMESTAMP		<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		+ 🗑️ ⚙️
4	passenger_count	passenger_count	STRING		<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		+ 🗑️ ⚙️
5	trip_distance	trip_distance	DECIMAL(10,2)		<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		+ 🗑️ ⚙️
6	rate_code_id	rate_code_id	BIGINT	rate_code	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		+ 🗑️ ⚙️
7	store_fhd_flag	store_fhd_flag	STRING		<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		+ 🗑️ ⚙️
8	pu_location_id	pu_location_id	STRING		<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		+ 🗑️ ⚙️
9	do_location_id	do_location_id	STRING		<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		+ 🗑️ ⚙️
10	payment_type	payment_type	BIGINT	payment_	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		+ 🗑️ ⚙️
11	fare_amount	fare_amount	DECIMAL(10,2)		<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		+ 🗑️ ⚙️
12	extra	extra	DECIMAL(10,2)		<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		+ 🗑️ ⚙️
13	mta_tax	mta_tax	DECIMAL(10,2)		<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		+ 🗑️ ⚙️
14	tip_amount	tip_amount	DECIMAL(10,2)		<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		+ 🗑️ ⚙️
15	toll_amount	toll_amount	DECIMAL(10,2)		<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		+ 🗑️ ⚙️
16	improvement_surcharge	improvement_surcharge	DECIMAL(10,2)		<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		+ 🗑️ ⚙️
17	total_amount	total_amount	DECIMAL(10,2)		<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		+ 🗑️ ⚙️

You can perform the following operations on the fields in the standard travel data table

– **Associating with data standards**

When creating or editing a table, click the **Table Fields** tab. Locate the

row that contains a field and click  in the **Data Standard** column to associate the field with a data standard. After the field is associated with a data standard and the table is published, a quality job is automatically generated, and a quality rule is generated for each field associated with a data standard. You can monitor the fields based on the data standards and view the field statuses on the **Quality Jobs** page of the DataArts Quality console. For more information about associating data standards, see .

– **Adding tags**

Tags are user-defined identifiers. After adding a tag, you can search for related data assets in the DataArts Catalog module with ease.

When creating or editing a table, click the **Table Fields** tab, locate the

row that contains a field, and click  in the **Tag** column. In the

displayed dialog box, enter a new tag name and press **Enter** or select an existing tag from the drop-down list.

- **Associating with quality rules**

After creating a table, you can associate fields in the table with quality rules. After the association is complete and the table is published, a quality job is automatically created on the **DataArts Quality** page after the table is published. If the table has been published, the system automatically updates the quality job. For more information about associating quality rules, see [Associating with Quality Rules](#).

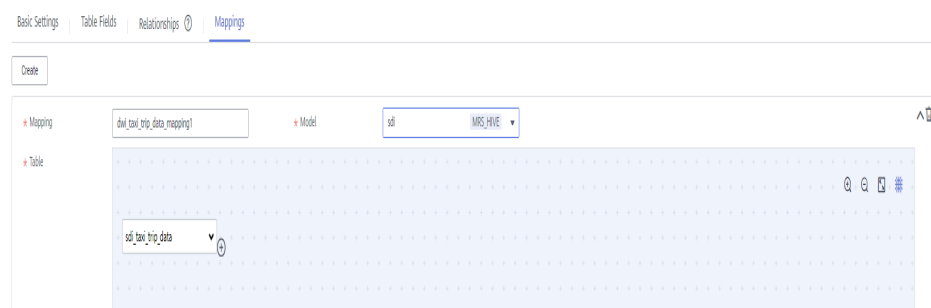
4. Click the **Mappings** tab and create mappings to design data sources of the table.

- If the table field comes from different relationship models, you must create multiple mappings. In each mapping, you only need to set the source field for the field that comes from the current mapping.
- If the table fields come from multiple tables in the same ER model, you can create a mapping. You can set **Join** for multiple tables of the mapping and set source fields for the fields in the table.

In this example, you only need to create one mapping. Click **Create** to create a mapping.

- **Mapping** is automatically generated, but is also configurable.
- **Model:** Select **sdi**.
- **Table:** Select the original data table **sdi_taxi_trip_data**, from where data of the standard travel table comes.

Figure 5-220 Creating a mapping



- **Field Mapping**

In the **Field Mapping** area, set source fields for the fields in the table in sequence. The selected source fields must have the same meaning as the fields in the table. As shown in [Figure 5-221](#), the generated SQL statement is displayed at the bottom of the **Field Mapping** area.

NOTE

- On the DataArts Architecture page, choose **Metrics > Configuration Center** in the left navigation pane, and click the **Functions** tab. On the page displayed, if **Create data development jobs** is selected for **Model Design Process**, the system creates an ETL job during data development based on the table mapping information during table release. An ETL node is generated for each mapping, and the job name starts with *Database name_Table code*. Currently, this function is in the internal test stage. Only DLI-to-DLI and DLI-to-DWS mapping jobs can be created.
You can choose **DataArts Factory > Job Development** to view the created ETL jobs. By default, ETL jobs are scheduled at 00:00 every day.
- In this example, the function of automatically creating ETL jobs is not enabled. The function provides only the data flow direction for data development. During data development, you can refer to the mapping to write SQL scripts.

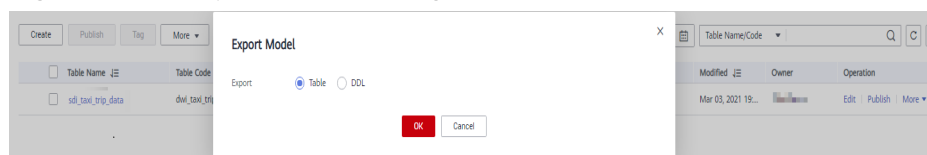
Figure 5-221 Field Mapping

No.	Field Name	Data Type
1	vendor_id	BIGINT
2	trip_pickup_datetime	TIMESTAMP
3	trip_dropoff_datetime	TIMESTAMP
4	passenger_count	STRING
5	trip_distance	DECIMAL
6	rate_code_id	BIGINT
7	store_fed_flag	STRING
8	pu_location_id	STRING
9	ds_location_id	STRING
10	payment_type	BIGINT
11	fare_amount	DECIMAL
12	extra	DECIMAL
13	mta_tax	DECIMAL
14	tip_amount	DECIMAL
15	toll_amount	DECIMAL
16	improvement_surcharge	DECIMAL
17	total_amount	DECIMAL

5. After configuring the mapping, you have finished configuring the taxi trip data table. Click **Save**.

Step 4 Select the created model and choose **More > Export**. In the displayed dialog box, select **Table** for **Export** and click **OK** to export the model. Then export the **sdi** model in the same way. The exported models can be used as backups and imported when needed in the future.

Figure 5-222 Export Model dialog box



Step 5 Publish table models.


- Publish the source table imported to the SDI ER model in **Step 2**. After the table is published, you can use DataArts Studio to manage and monitor the source table.


Return to the **ER Modeling** page, select the **sdi** model in the model directory. Select the **sdi_taxi_trip_data** table in the list on the right, and click **Publish**. In the dialog box displayed, select a reviewer and click **OK**. Wait for the reviewer to review the application. If you have the reviewer permissions, select **Auto-review** and click **Submit**.

2. Publish a table of the DWI ER model.

Return to the **ER Modeling** page, select the **dwi** model in the model directory. Select the **dwi_table_trip_data** table in the list on the right, and click **Publish**. In the dialog box displayed, select a reviewer and click **OK**. Wait for the reviewer to review the application. If you have the reviewer permissions, select **Auto-review** and click **Submit**.

Step 6 After the application is approved, you can view **Status** and **Sync Status** on the **ER Modeling** page.

Publishing is an asynchronous operation. You can click  to refresh the status. After table publishing application is approved, the system performs operations such as creating tables and synchronizing technical assets and business assets based on the configurations of **Model Design Process** on the **Function Settings** tab page in **Configuration Center**. The synchronization status is displayed in the **Sync Status** column of the table on the **Information Architecture** page.

- If the **Sync Status** is successful, the table is published successfully. Move the cursor over  in the **Sync Status** column. If the message **Table created successfully** is displayed, the table has been successfully created in the corresponding data source.
- If one or more items in the in the **Sync Status** column fail to be synchronized, you can refresh the status. If the fault persists, choose **More > View History** to view logs.

Locate the failure cause based on the error log and rectify the fault. Then return to the **ER Modeling** page, select the tables to be synchronized from the list, and choose **More > Synchronize** to synchronize the tables again. If the fault persists, contact technical support.

Figure 5-223 Checking the table status

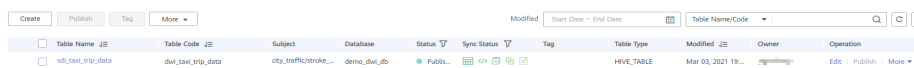
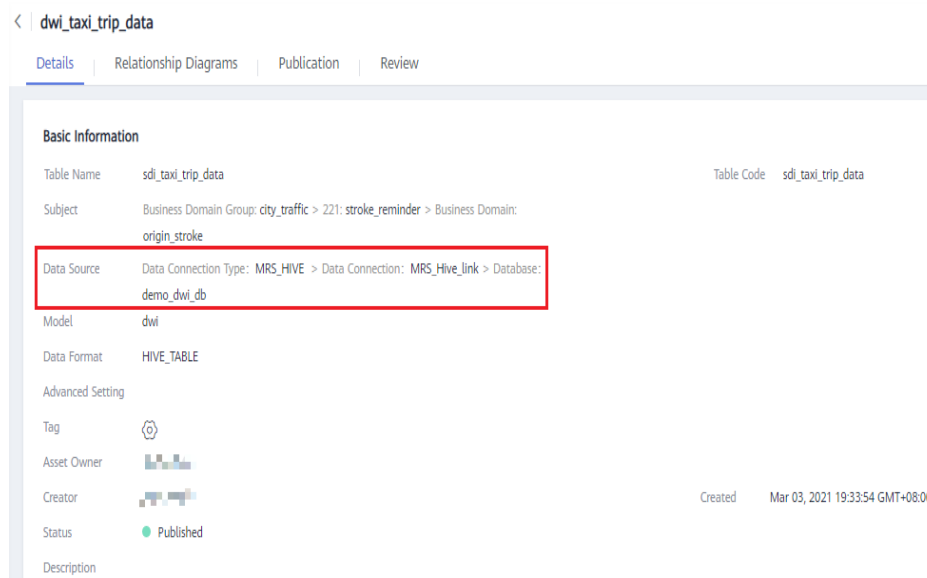


Table Name	Table Code	Subject	Database	Status	Sync Status	Tag	Table Type	Modified	Owner	Operation
sdi_taxi_trip_data	dwi_taxi_trip_data	city_traffic/traffic_...	demo_dwi_db	PUBLI...			HIVE_TABLE	Mar 03, 2021 19:...	ip@10000000	Edit Publish More

Click a table name in the list to view the table details. **Data Source** indicates the location of the table.

Figure 5-224 Table details



----End

Creating and Publishing Dimensions for the DWR Layer

During dimension modeling, create three lookup table dimensions (**vendor**, **rate_code**, and **payment_type**) and one hierarchy dimension (**date**) for the DWR layer.

- Step 1** On the DataArts Architecture console, choose **Models > Dimensional Modeling** in the navigation pane on the left.
- Step 2** Create the three lookup table dimensions listed in [Table 5-60](#).

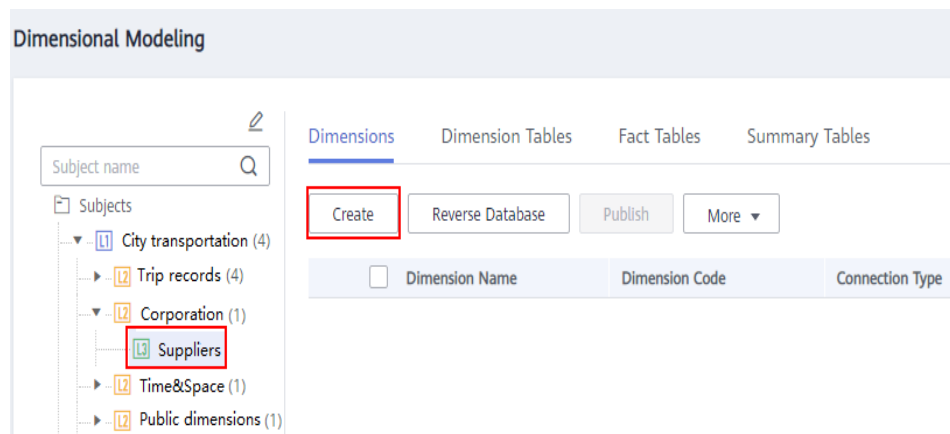
Table 5-60 Lookup table dimensions

*Subject	*Dimension Name	*Dimension Code	*Type	*Owner	Description	*Data Connection Type	*Data Connection Name	*Database	Lookup Table
vendor	vendor	dim_vendor	Lookup table	-	None	MRS_HIVE	mrs_hive_link	demo_dwr_db	vendor
public_dimension	rate_code	dim_rate_code	Lookup table	-	None	MRS_HIVE	mrs_hive_link	demo_dwr_db	rate

*Subject	*Dimension Name	*Dimension Code	*Type	*Owner	Description	*Data Connection Type	*Data Connection Name	*Database	Look up Table
public_dimension	payment_type	dim_payment_type	Look up table	-	None	MRS_HIVE	mrs_hive_link	demo_dwr_db	payment_type

1. Click the **Dimensions** tab, choose **City transportation > Corporation > Suppliers** in the subject tree, and click **Create** to create a dimension named **Suppliers**.

Figure 5-225 Dimensional modeling



2. On the **Create Dimension** page, set the parameters as shown in the figure below and click **Save**.

Figure 5-226 Creating a dimension named Suppliers

Basic Settings

Subject: **Suppliers**
Selected: Business Domain Group: **City transportation** > Business Domain: **Corporation** > Business Object: **Suppliers**

Dimension Name: **Suppliers** Dimension Code: **dim_vendor**

Type: **Basic** Lookup table Hierarchy

Owner: _____ C

Description: _____
1,600

Physicalization Settings

Data Connection Type: **MRS_HIVE** Data Connection Name: **Mrs_hive_link** C

Database: **demo_dwr_db**

Table Type: **HIVE_TABLE**

Field Settings

Lookup Table: **Suppliers**

No.	Field Name	Field Code	Data Standard	Data Type	Surrogat...	Primary ...	Partition	Not Null	Comment
1	Suppliers ID	vendor_id		BIGINT	<input checked="" type="radio"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
2	Suppliers	vendor_value		STRING	<input type="radio"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

3. Click the **Dimensions** tab, choose **City transportation > Public dimensions > Public dimensions** in the subject tree, and click **Create** to create a dimension named **rate code**. On the **Create Dimension** page, set the parameters as shown in the figure below and click **Save**.

Figure 5-227 Creating a dimension named rate code

Basic Settings

Subject: **Public dimensions**
Selected: Business Domain Group: **City transportation** > Business Domain: **Public dimensions** > Business Object: **Public dimensions**

Dimension Name: **rate code** Dimension Code: **dim_rate_code**

Type: **Basic** Lookup table Hierarchy

Owner: _____ C

Description: _____
1,600

Physicalization Settings

Data Connection Type: **MRS_HIVE** Data Connection Name: **Mrs_hive_link** X C

Database: **demo_dwr_db**

Table Type: **HIVE_TABLE**

Field Settings

Lookup Table: **rate code**

No.	Field Name	Field Code	Data Standard	Data Type	Surrogat...	Primary ...	Partition	Not Null	Comment
1	rate ID	rate_code_id		BIGINT	<input checked="" type="radio"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
2	rate description	rate_code_value		STRING	<input type="radio"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

4. Click the **Dimensions** tab, choose **City transportation > Public dimensions > Public dimensions** in the subject tree, and click **Create** to create a dimension named **payment method**. On the **Create Dimension** page, set the parameters as shown in the figure below and click **Save**.

Figure 5-228 Creating a dimension named payment method

The screenshot shows the configuration for a dimension named 'payment method'. It is divided into three sections:

- Basic Settings:**
 - Subject: Public dimensions
 - Selected: Business Domain Group: City transportation > Business Domain: Public dimensions > Business Object: Public dimensions
 - Dimension Name: payment method
 - Dimension Code: dim_payment_type
 - Type: Basic (selected), Lookup table, Hierarchy
 - Owner: [Empty field]
 - Description: [Empty text area]
- Physicalization Settings:**
 - Data Connection Type: MRS_HIVE
 - Data Connection Name: Mrs_hive_link
 - Database: demo_dwr_db
 - Table Type: HIVE_TABLE
- Field Settings:**
 - Lookup Table: payment method
 - Table with columns: No., Field Name, Field Code, Data Standard, Data Type, Summat..., Primary..., Partition, Not Null, Comment.
 - Row 1: 1, payment type ID, payment_type_id, Data Standard, BIGINT, [Checked], [Unchecked], [Unchecked], [Unchecked]
 - Row 2: 2, payment type value, payment_type_value, Data Standard, STRING, [Unchecked], [Unchecked], [Unchecked], [Unchecked]

Step 3 Create a hierarchy dimension named **date**.

1. On the **Dimensional Modeling** tab page, choose **City transportation > Time&Space > Time** in the subject tree. Then click **Create** on the **Dimensions** tab page to create a dimension named **date dimension**.
2. Configure the basic settings and physicalization settings as shown in the figure below.

Table 5-61 Date dimension

*Subject	*Dimension Name	*Dimension Code	*Type	*Owner	Description	*Data Connection Type	*Data Connection Name	*Database
date	date dimension	dim_date	Hierarchy	-	None	MRS_HIVE	mrs_hive_link	demo_dwr_db

Figure 5-229 Date dimension

The screenshot displays the configuration interface for a Date dimension. It is divided into two main sections: Basic Settings and Physicalization Settings.

- Basic Settings:**
 - Subject:** Time
 - Selected:** Business Domain Group: City transportation > Business Domain: TimeSpace > Business Object: Time
 - Dimension Name:** date dimension
 - Dimension Code:** dim_date
 - Type:** Basic (selected), Lookup table, Hierarchy
 - Owner:** [Empty field] C
 - Description:** [Empty text area]
- Physicalization Settings:**
 - Data Connection Type:** MRS_HIVE
 - Data Connection Name:** Mrs_hive_link C
 - Database:** demo_dwr_db
 - Table Type:** HIVE_TABLE

3. In the **Field Settings** area, add fields as described in the table below.

Table 5-62 Field settings

No.	Field Name	Field Code	Data Standard	Data Type	Surrogate Key	Primary Key	Partition	Not Null
1	dim_date_key	dim_date_key	-	TIMESTAMP	Selected	Selected	Not selected	Selected
2	real_time	real_time	-	TIMESTAMP	Not selected	Not selected	Not selected	Not selected
3	minute_id	minute_id	-	BIGINT	Not selected	Not selected	Not selected	Not selected
4	minute	minute	-	BIGINT	Not selected	Not selected	Not selected	Not selected
5	hour_id	hour_id	-	BIGINT	Not selected	Not selected	Not selected	Not selected

No.	Field Name	Field Code	Data Standard	Data Type	Surrogate Key	Primary Key	Partition	Not Null
6	hour	hour	-	BIGINT	Not selected	Not selected	Not selected	Not selected
7	day_id	day_id	-	BIGINT	Not selected	Not selected	Not selected	Not selected
8	day	day	-	STRING	Not selected	Not selected	Not selected	Not selected
9	month_id	month_id	-	BIGINT	Not selected	Not selected	Not selected	Not selected
10	month	month	-	STRING	Not selected	Not selected	Not selected	Not selected
11	year_id	year_id	-	BIGINT	Not selected	Not selected	Not selected	Not selected
12	year	year	-	BIGINT	Not selected	Not selected	Not selected	Not selected

Figure 5-230 Field settings

No.	Field Name	Field Code	Data Standard	Data Type	Sampled	Primary	Partition	Not Null	Comment	Operation
1	data dimension	dim_data_key		TIMESTAMP	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		+ [edit] [refresh]
2	time	real_time		TIMESTAMP	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		+ [edit] [refresh]
3	minute ID	minute_id		BIGINT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		+ [edit] [refresh]
4	minute	minute		BIGINT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		+ [edit] [refresh]
5	hour ID	hour_id		BIGINT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		+ [edit] [refresh]
6	hour	hour		BIGINT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		+ [edit] [refresh]
7	day ID	day_id		BIGINT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		+ [edit] [refresh]
8	day	day		STRING	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		+ [edit] [refresh]
9	month ID	month_id		BIGINT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		+ [edit] [refresh]
10	month	month		STRING	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		+ [edit] [refresh]
11	year ID	year_id		BIGINT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		+ [edit] [refresh]
12	year	year		BIGINT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		+ [edit] [refresh]

- In the **Hierarchy Settings** area, click **Add** to create two layers as shown in the figures below.

Figure 5-231 Layer 1

Level Name: year.month.day.time.minute

Structure: minute @ hour @ day_id @ month_id @ year_id

You can select 2 to 10 fields to create a hierarchical structure, which is arranged using the selected fields from left to right.

Field Name	Details
minute	minute_id
hour	hour_id
day_id	day_id
month_id	month_id
year_id	year_id

Figure 5-232 Layer 2

Level Name: year.month.day

Structure: day_id @ month_id @ year_id

You can select 2 to 10 fields to create a hierarchical structure, which is arranged using the selected fields from left to right.

Field Name	Details
day_id	day_id
month_id	month_id
year_id	year_id

- Click **Save**.

- Return to the **Dimensions** tab page, select the four new dimensions in the dimension list, and click **Publish**.
- In the **Apply for Publication** dialog box, select a reviewer and click **OK**. Wait for the reviewer to review the application. If you have the reviewer permissions, select **Auto-review** and click **Submit**.
- After a dimension is published and approved, the system automatically creates a dimension table for the dimension. The name and code of the dimension table are the same as those of the dimension. On the **Dimensional Modeling** page, click the **Dimension Tables** tab to view the created dimension table.

In the dimension table list, you can view **Sync Status** of the dimension tables.

- If all items in **Sync Status** are displayed as **Succeeded**, the dimension is published and the dimension table is created in the database.
- If an item in **Sync Status** is displayed as **Failed**, click **View History** in the row. On the page displayed, click the **History** tab to view logs. Troubleshoot the fault based on the logs. After the fault is rectified, select the dimension table, click **Synchronize** above the dimension table list, and click **OK** in the dialog box displayed. If the fault persists, contact technical support for assistance.

Figure 5-233 Sync Status of the dimension tables

Table Name	Table Code	Table Type	Status	Type	Sync Status	Subject	Modified	Owner	Operation
payment_type	dim_payment_type	HIVE_TABLE	Published	Lookup table		City transportation	Feb 25, 2022 11:2...		View History Preview SQL
rate_code	dim_rate_code	HIVE_TABLE	Published	Lookup table		City transportation	Feb 25, 2022 11:2...		View History Preview SQL
date dimension	dim_date	HIVE_TABLE	Published	Hierarchy		City transportation	Feb 25, 2022 11:3...		View History Preview SQL
Suppliers	dim_vendor	HIVE_TABLE	Published	Lookup table		City transportation	Feb 25, 2022 11:2...		View History Preview SQL

----End

Creating and Publishing a Fact Table for the DWR Layer

During dimensional modeling, create a fact table named **stroke_order** for the DWR layer.

Step 1 On the DataArts Architecture console, choose **Models > Dimensional Modeling** in the navigation pane on the left.

Step 2 Click the **Fact Tables** tab, choose **City transportation > Trip records > Trip facts** in the subject tree, and click **Create** to create a fact table named **stroke_order**.

In the **Basic Settings** area on the **Create Fact Table** page, set the following parameters:

- **Subject: Subject Area Group: City transportation > Subject Area: Trip records > Business Object: Trip facts**
- **Table Name: stroke_order**
- **Table Code: fact_stroke_order**
- **Data Connection Type: MRS_HIVE**
- **Data Connection Name: mrs_hive_link**
- **Database: demo_dwr_db**
- **Table Type: HIVE_TABLE**
- **Owner: an owner in the drop-down list box**
- **Description: None**

In the **Field Settings** area, choose **Create > Dimension**. In the dialog box displayed, select the dimensions **rate_code**, **vendor**, **payment_type**, and **date**, and click **OK**. Choose **Create > Dimension**. In the dialog box displayed, select the dimension **date** and click **OK**. In the dimension field list, adjust the sequence of the dimension fields and modify the information about the two **date** dimensions, as listed in [Table 5-63](#).

Table 5-63 Dimension fields

N o.	Field Name	Field Code	Data Type	Primary Key	Partition	Not Null	Associated Standard	Associated Dimension	Role	Description
1	rate_code_id	rate_code_id	BIGINT	Not selected	Not selected	Not selected	-	rate_code	dim_	-
2	vendor_id	vendor_id	BIGINT	Not selected	Not selected	Not selected	-	vendor	dim_	-
3	payment_type_id	payment_type_id	BIGINT	Not selected	Not selected	Not selected	-	payment_type	dim_	-
4	pickup_date_key	dim_pickup_date_key	TIMESTAMP	Not selected	Not selected	Not selected	-	Date	dim_pickup	Date dimension table
5	tprep_dropoff_datetime	dim_dropoff_date_key	TIMESTAMP	Not selected	Not selected	Not selected	-	Date	dim_dropoff	Date dimension table

In the **Field Settings** area, choose **Create > Measure** and create the fields listed in [Table 5-64](#) in sequence.

Table 5-64 Measure fields

No .	Field Name	Field Code	Data Type	Prim ary Key	Parti tion	Not Null	Ass ocia ted Stan dard
6	pu_loca tion_id	pu_location_i d	STRING	Not selec ted	Not selec ted	Not select ed	-
7	do_loca tion_id	do_location_i d	STRING	Not selec ted	Not selec ted	Not select ed	-
8	fare_a mount	fare_amount	DECIMAL (10,2)	Not selec ted	Not selec ted	Not select ed	-
9	extra	extra	DECIMAL (10,2)	Not selec ted	Not selec ted	Not select ed	-
10	mta_ta x	mta_tax	DECIMAL (10,2)	Not selec ted	Not selec ted	Not select ed	-
11	tip_am ount	tip_amount	DECIMAL (10,2)	Not selec ted	Not selec ted	Not select ed	-
12	tolls_a mount	tolls_amount	DECIMAL (10,2)	Not selec ted	Not selec ted	Not select ed	-
13	improv ement_ surchar ge	improvement_ surcharge	DECIMAL (10,2)	Not selec ted	Not selec ted	Not select ed	-
14	total_a mount	total_amount	DECIMAL (10,2)	Not selec ted	Not selec ted	Not select ed	-

Figure 5-234 Fact table fields

No.	Field Type	Field Name	Field Code	Data Type	Primary Key	Partition	Not Null	Associated Standards	Associated Dimensions	Role	Comment	Operation
1	Dimension	rate ID	rate_code_id	BIGINT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		rate code	dm_		+ 🗑️ 🔍
2	Dimension	Suppliers ID	vendor_id	BIGINT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		suppliers	dm_		+ 🗑️ 🔍
3	Dimension	payment type	payment_type_id	BIGINT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		payment method	dm_		+ 🗑️ 🔍
4	Dimension	pickup time	dim_pickup_date_key	TIMESTAMP	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		date dimension	dim_pickup	date dimension table	+ 🗑️ 🔍
5	Dimension	dropoff time	dim_dropoff_date_key	TIMESTAMP	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		date dimension	dim_dropoff	date dimension table	+ 🗑️ 🔍
6	Measure	pickup location	pu_location_id	STRING	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>					+ 🗑️ 🔍
7	Measure	dropoff location	do_location_id	STRING	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>					+ 🗑️ 🔍
8	Measure	fare	fare_amount	DECIMAL(10,2)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>					+ 🗑️ 🔍
9	Measure	extra	extra	DECIMAL(10,2)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>					+ 🗑️ 🔍
10	Measure	MTA tax	mta_tax	DECIMAL(10,2)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>					+ 🗑️ 🔍
11	Measure	tips	tips_amount	DECIMAL(10,2)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>					+ 🗑️ 🔍
12	Measure	tolls	tolls_amount	DECIMAL(10,2)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>					+ 🗑️ 🔍
13	Measure	improvement surcharge	improvement_surcharge	DECIMAL(10,2)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>					+ 🗑️ 🔍
14	Measure	total fare	total_amount	DECIMAL(10,2)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>					+ 🗑️ 🔍

Step 3 After the configuration, click **Publish**.

Step 4 In the dialog box displayed, select a reviewer and click **OK**. Wait for the reviewer to review the application. If you have the reviewer permissions, select **Auto-review** and click **Submit**.

Step 5 Return to the **Fact Tables** tab page, find the new fact table in the list, and view **Sync Status**.

- If all items in **Sync Status** are displayed as **Succeeded**, the fact table is published and created in the database.
- If an item in **Sync Status** is displayed as **Failed**, choose **More > View History**. On the page displayed, click the **History** tab to view logs. Troubleshoot the fault based on the logs. After the fault is rectified, choose **More > Synchronize** above the fact table list, and click **OK** in the dialog box displayed. If the fault persists, contact technical support for assistance.

----End

Creating and Publishing Technical Metrics

In this example, you need to create the technical metrics listed in [Table 5-65](#) and [Table 5-66](#).

Table 5-65 Atomic metrics

*Metric Name	* Metric Code	Data Table	*Subject	*Expression	Descrip tion
total_a mount	sum_total_ amount	Itinerary order	stroke_fa ct	sum (total amount)	None

Table 5-66 Derivative metrics

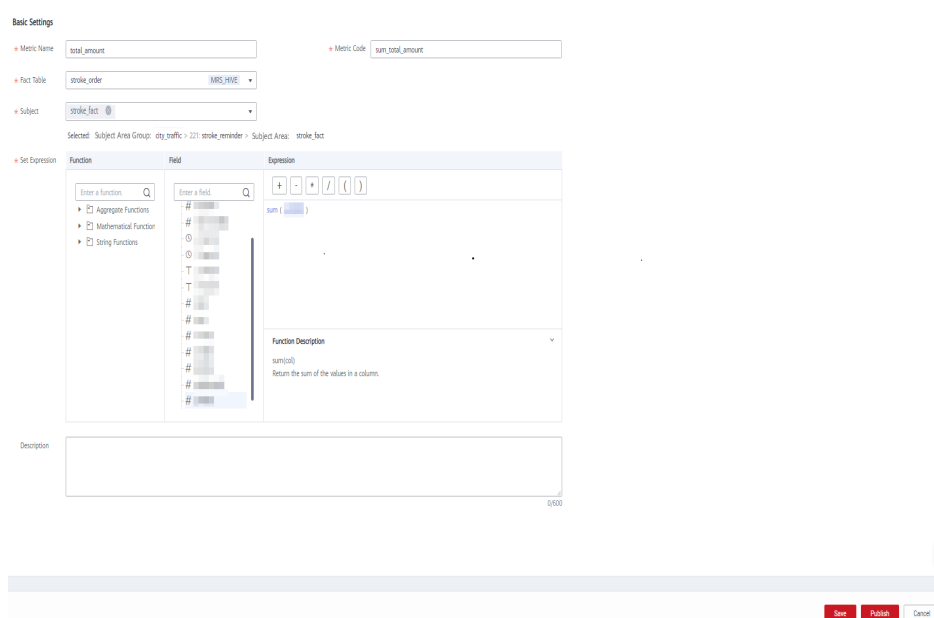
Metric	*Data Table	*Subject	*Atomic Metric	Statistical Dimension	Time Filter	General Filter
total_amount_(payment_type)	Itinerary order	stroke_statistic	total_amount	payment_type	None	None
total_amount_(rate_code)	Itinerary order	stroke_statistic	total_amount	rate_code	None	None
total_amount_(vendor,stroke_order.dim_dropoff_date_key)	Itinerary order	stroke_statistic	total_amount	vendor and stroke_order.dim_dropoff_date_key	None	None

Step 1 On the DataArts Architecture console, choose **Metrics > Technical Metrics** in the navigation pane on the left.

Step 2 Create an atomic metric named **total_amount** to collect statistics on fares.

1. Click the **Atomic Metrics** tab and click **Create**.
2. On the **Create Atomic Metric** page, set the parameters as shown in the figure below and click **Publish**.

Figure 5-235 Creating an atomic metric



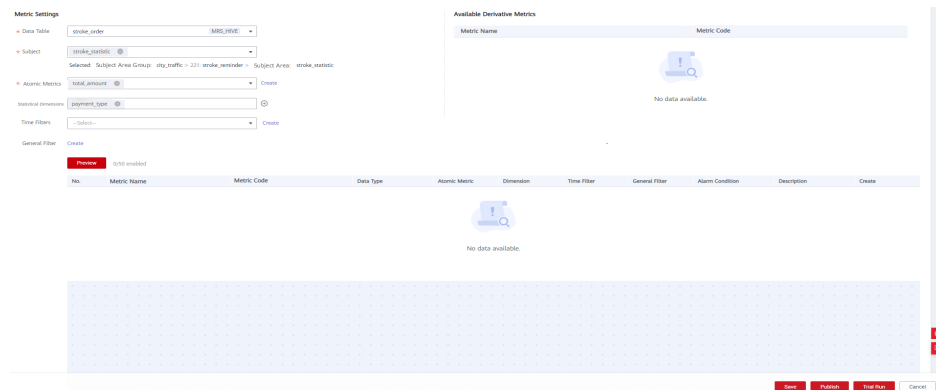
3. Wait for the reviewer to review the application. After the application is approved, the atomic metric will be created.

Step 3 Create three derivative metrics.

- Create **total_amount_(payment_type)** to collect statistics on the total fares based on **payment_type**.

On the **Technical Metrics** page, click the **Derivative Metrics** tab and click **Create**. On the **Create Derivative Metric** page, set the parameters as shown in the figure below. After the configuration is complete, click **Trial Run**. In the dialog box displayed, click **Execute**. If the trial running is successful, click **Save**.

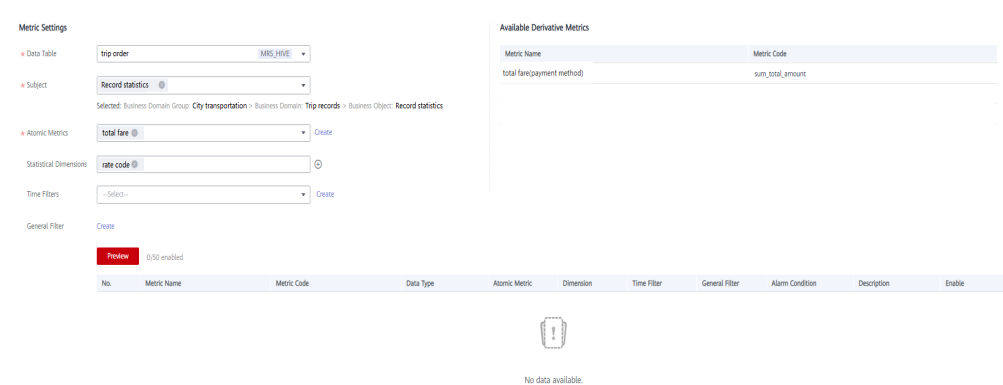
Figure 5-236 Creating a derivative metric named total_amount_(payment_type)



- Create **total_amount_(rate_code)** to collect statistics on the total fares based on **rate_code**.

On the **Technical Metrics** page, click the **Derivative Metrics** tab and click **Create**. On the **Create Derivative Metric** page, set the parameters as shown in the figure below. After the configuration is complete, click **Trial Run**. In the dialog box displayed, click **Execute**. If the trial running is successful, click **Save**.

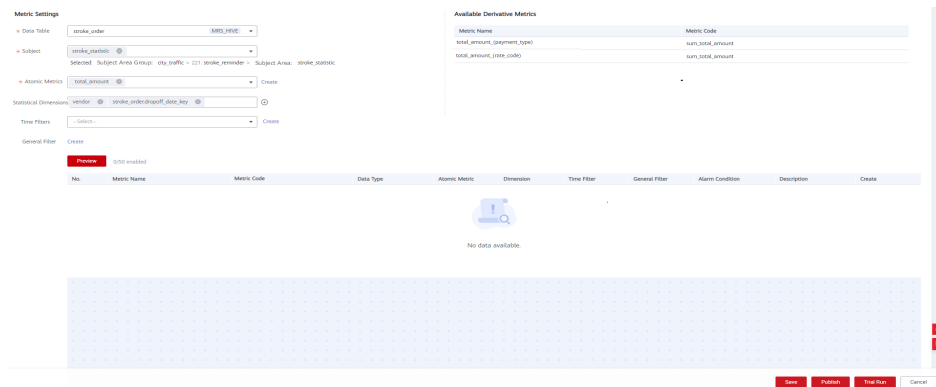
Figure 5-237 Creating a derivative metric named total_amount_(rate_code)



- Create **total_amount_(vendor,stroke_order.dim_dropoff_date_key)** to collect statistics on the total fares based on **vendor**.

On the **Technical Metrics** page, click the **Derivative Metrics** tab and click **Create**. On the **Create Derivative Metric** page, set the parameters as shown in the figure below. After the configuration is complete, click **Trial Run**. In the dialog box displayed, click **Execute**. If the trial running is successful, click **Save**.

Figure 5-238 Creating a derivative metric named total_amount_(vendor,stroke_order.dim_dropoff_date_key)



Step 4 Return to the **Derivative Metrics** tab page, select the three derivative metrics and click **Publish**. In the dialog box displayed, select a reviewer and click **OK**. Wait for the reviewer to review the application. If you have the reviewer permissions, select **Auto-review** and click **Submit**.

----End

Creating and Publish Summary Tables for the DM Layer

Create the three summary tables listed in [Table 5-67](#) for the DM layer.

Table 5-67 Summary tables

*Subject	*Table Name	*Table Code	Statistical Dimension	Data Connection Type	*Data Connection Name	*Database	Owner	Description
stroke_statistic	payment_type	dws_payment_type	payment_type	MRS_HIVE	mrs_hive_link	demo_dm_db	-	None
stroke_statistic	rate_code	dws_rate_code	rate_code	MRS_HIVE	mrs_hive_link	demo_dm_db	-	None
stroke_statistics	vendor	dws_vendor	vendor and stroke_order.dim_dropoff_date_key	MRS_HIVE	mrs_hive_link	demo_dm_db	-	None

Step 1 On the DataArts Architecture console, choose **Models > Dimensional Modeling** in the left navigation pane.

Step 2 Click the **Summary Tables** tab.

Step 3 Create three summary tables: **payment_type**, **rate_code**, and **vendor**.

1. On the **Summary Tables** page, choose **City transportation > Trip records > Record statistics** in the directory tree, and click **Create** to create a summary table named **payment method statistics**. On the **Create Summary Table** page, set the parameters and click **Save**.

Set the basic settings as shown in the figure below.

Figure 5-239 Creating a summary table named payment method statistics

The screenshot shows the 'Basic Settings' section of the 'Create Summary Table' interface. It contains the following fields and values:

- Subject:** Record statistics (Selected: Business Domain Group: City transportation > Business Domain: Trip records > Business Object: Record statistics)
- Table Name:** payment method statistics
- Table Code:** dws_payment_type
- Statistical Dimension:** payment method (MRS_HIVE)
- Data Connection Type:** MRS_HIVE (Data Connection Name: Mrs_hive_link)
- Database:** demo_dm_db
- Table Type:** HIVE_TABLE
- Owner:** (empty field)
- Description:** (empty text area)

In the **Time Partition** area, enter the field code **dttime** and select the data type **TIMESTAMP**. After a table is published, data is written to the table based on the field added here.

Figure 5-240 Time Partition settings

The screenshot shows the 'Time Partition' settings table:

Code	Data Type
dttime	TIMESTAMP

In the **Metric Settings** area, click **Add** to add the derivative metric **total_amount_(payment_mode)**. You can add only published derivative or compound metrics that are associated with the specified statistical dimension.

Figure 5-241 Metric settings

The screenshot shows the 'Metric Fields' table:

No.	Type	Name	Code	Data Type	Not Null	Comment	Operation
1	Derivative metric	total_amount_(payment_mode)	sum_posit_amount	STRING	<input type="checkbox"/>		🗑️ 🔄 🗑️

Click **Save**.

2. On the **Summary Tables** page, choose **City transportation > Trip records > Record statistics** in the directory tree, and click **Create** to create a summary table named **rate statistics**. On the **Create Summary Table** page, set the parameters and click **Save**.

Figure 5-242 Creating a summary table named rate statistics (Basic Settings)

Basic Settings

* Subject: Record statistics
Selected: Business Domain Group: City transportation > Business Domain: Trip records > Business Object: Record statistics

* Table Name: rate statistics

* Table Code: dws_rate_code

* Statistical Dimension: rate code (MRS_HIVE)

* Data Connection Type: MRS_HIVE * Data Connection Name: Mrs_hive_link

* Database: demo_dm_db

* Table Type: HIVE_TABLE

* Owner: [Empty]

* Description: [Empty]

Figure 5-243 Creating a summary table named rate statistics (Metric Settings)

Time Partition

Code: dttime Data Type: TIMESTAMP

Metric Settings

+ Metrics [Add]

No.	Metric Type	Metric Name	Metric Code	Data Type	Not Null	Comment	Operation
1	Derivative metric	total fare(rate code)	sum_total_amount	STRING	<input type="checkbox"/>		<input type="edit"/> <input type="delete"/>

3. On the **Summary Tables** page, choose **City transportation > Trip records > Record statistics** in the directory tree, and click **Create** to create a summary table named **supplier statistics**. On the **Create Summary Table** page, set the parameters and click **Save**.

Figure 5-244 Creating a summary table named supplier statistics (Basic Settings)

Basic Settings

* Subject: Record statistics
Selected: Business Domain Group: City transportation > Business Domain: Trip records > Business Object: Standard records

* Table Name: supplier statistics

* Table Code: dws_vendor

* Statistical Dimension: supplier,trip order,dropoff time (MRS_HIVE)

* Data Connection Type: MRS_HIVE * Data Connection Name: Mrs_hive_link

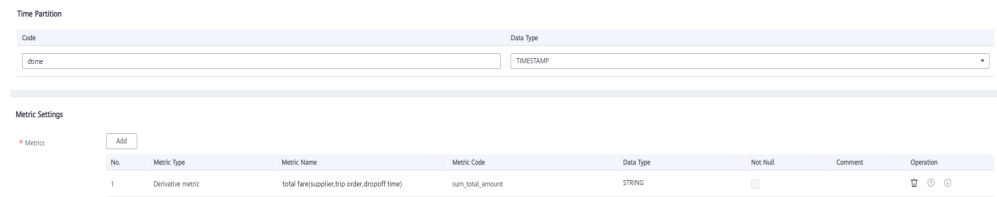
* Database: demo_dm_db

* Table Type: HIVE_TABLE

* Owner: ei_dlf_100341563

* Description: [Empty]

Figure 5-245 Creating a summary table named supplier statistics (Metric Settings)



- Step 4** Return to the **Dimensions** tab page, select the three new summary tables in the dimension list, and click **Publish**.
- Step 5** In the dialog box displayed, select a reviewer and click **OK**. After the reviewer approves the publishing application, the summary table is automatically created. If you have the reviewer permissions, select **Auto-review** and click **Submit**.
- Step 6** Return to the **Summary Tables** tab page, find the new summary tables in the list, and view **Sync Status**.
- If all items in **Sync Status** are displayed as **Succeeded**, the summary tables are published and created in the database.
 - If an item in **Sync Status** is displayed as **Failed**, choose **More > View History** in the row. On the page displayed, click the **History** tab to view logs. Troubleshoot the fault based on the logs. After the fault is rectified, choose **More > Synchronize** above the summary table list, and click **OK** in the dialog box displayed. If the fault persists, contact technical support for assistance.

----End

6 DataArts Factory

6.1 Overview

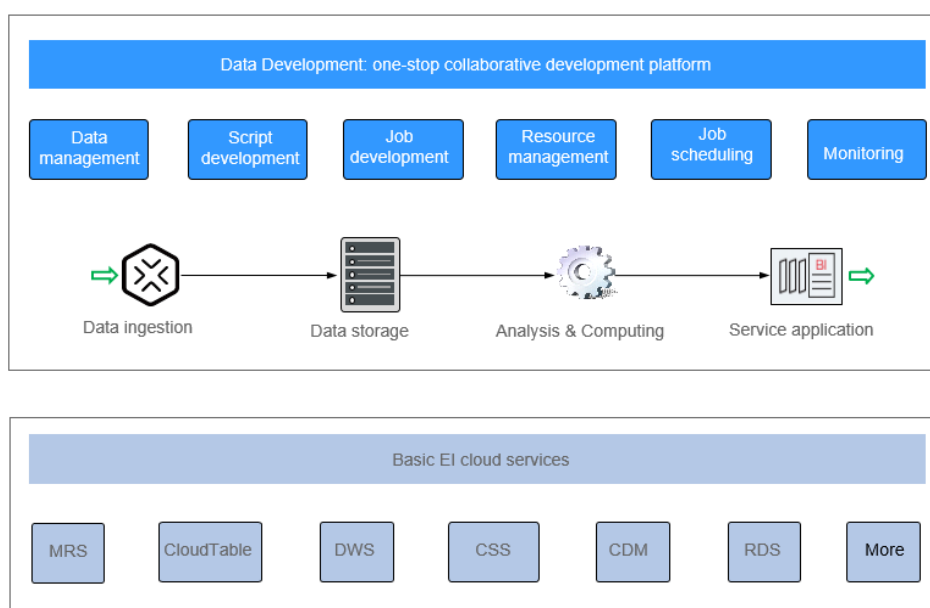
DataArts Factory is a one-stop big data collaborative development platform that provides fully managed big data scheduling capabilities. It manages various big data services, making big data more accessible than ever before and helping you effortlessly build big data processing centers.

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 6-1 DataArts Factory architecture



Main Functions

Table 6-1 Main functions of DataArts Factory

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

Objects in DataArts Factory

- **Data connection:** A data connection 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.

6.2 Data Management

6.2.1 Data Management Process

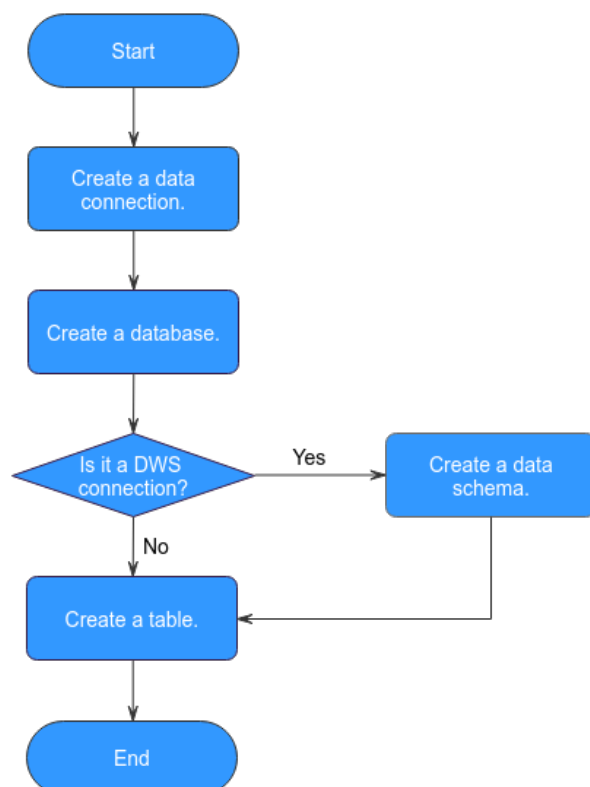
The data management function helps you quickly establish data models and provides you with data entities for script and job development. With data management, you can:

- 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 6-2 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](#).

6.2.2 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**.
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.

6.2.3 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**.
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 6-2](#).

Table 6-2 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">• DLI: The value must contain only numbers, letters, and underscores (_). It cannot start with an underscore (_) or contain only numbers.• DWS: The value must contain only numbers, letters, and underscores (_). It cannot start with an underscore (_) or contain only numbers.• 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.
Description	No	Descriptive information about the database. The requirements are as follows: <ul style="list-style-type: none">• DLI: The value contains a maximum of 256 characters.• DWS: The value contains a maximum of 1,024 characters.• MRS Hive: The value contains a maximum of 1,024 characters.

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**.

6.2.4 (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**.
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 6-3](#).

Table 6-3 Creating a database schema

Parameter	Mandatory	Description
Mode Name	Yes	Name of a database schema.
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**.

6.2.5 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**.
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 6-4](#) lists the links for viewing property parameters of each type of data connection.

Table 6-4 Basic property parameters

Data Connection Type	Description
DLI	For details, see the Basic Property part in Table 6-8 .
DWS	For details, see the Basic Property part in Table 6-9 .
MRS Hive	For details, see the Basic Property part in Table 6-10 .

5. Click **Next**. On the **Configure Table Structure** page, configure the table structure parameters based on [Table 6-5](#).

Table 6-5 Table structure

Data Connection Type	Description
DLI	For details, see the Table Structure part in Table 6-8 .
DWS	For details, see the Table Structure part in Table 6-9 .
MRS Hive	For details, see the Table Structure part in Table 6-10 .

6. Click **OK**.

Creating a Table (DDL 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**.
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. Click **DDL-based Table Creation**, configure the parameters based on [Table 6-6](#), and enter SQL statements in the editor in the lower part.

Table 6-6 Data table parameters

Parameter	Description
Data Connection Type	Type of data connection to which the table belongs. <ul style="list-style-type: none">• DLI• DWS• HIVE
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 6-7 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 6-8 DLI data table

Parameter	Mandatory	Description
Basic Property		
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">• OBS• DLI


Parameter	Mandatory	Description
Data Format	Yes	Format of data. This parameter is available only when Data Location is set to OBS . Possible values: <ul style="list-style-type: none"> • parquet: DLF can read non-compressed parquet data and parquet data compressed using Snappy or gzip. • csv: DLF can read non-compressed CSV data and CSV data compressed using gzip. • orc: DLF can read non-compressed ORC data and ORC data compressed using Snappy. • json: DLF can read non-compressed JSON data and JSON data compressed using gzip.
Path	Yes	OBS path where the data is stored. This parameter is available only when Data Location is set to OBS .
Table Description	No	Descriptive information about the table.
Table Structure		
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 6-9 DWS data table

Parameter	Mandatory	Description
Basic Property		
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"> ● Storage method of a table. Possible values: <ul style="list-style-type: none"> - Row store - Column store ● Compression level of a table <ul style="list-style-type: none"> - Available values when the storage method is row store: YES or NO. - Available values when the storage method is column store: YES, NO, LOW, MIDDLE, or HIGH. 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.
Table Structure		
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"> • Value • Currency • Boolean • Binary • Character • Time • Geometric • Network address • Bit string • Text search • UUID • JSON • OID
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 CloudSearch Cluster Name 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"> • text • keyword • date • long • integer • short • byte • double • boolean • binary
Operation	No	To add a column, click  .

Table 6-10 Basic property parameters of an MRS Hive data table

Parameter	Mandatory	Description
Basic Property		
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.
Table Structure		
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"> • Original type • ARRAY • MAP • STRUCT • UNION
Data Type	Yes	Type of data. See LanguageManual DDL .
Column Description	No	Descriptive information about the column.
Operation	No	To add a column, click  .

6.3 Script Development

6.3.1 Script Development Process

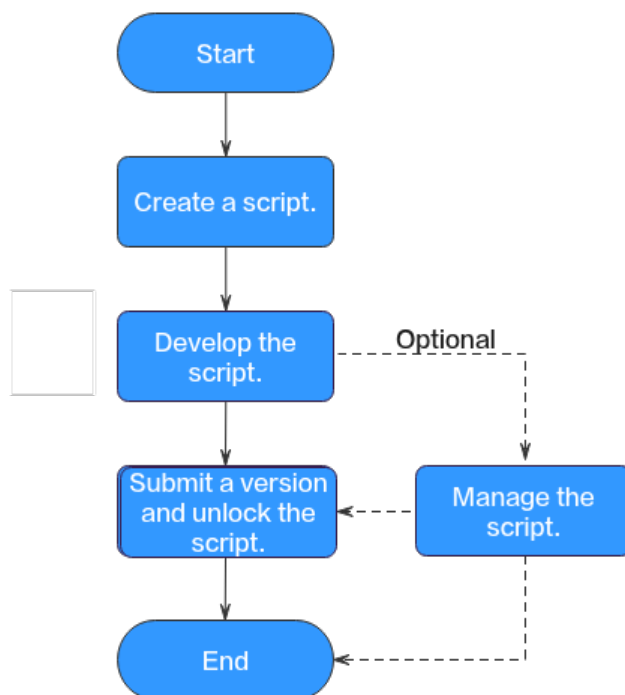
The script development function provides the following capabilities:

- 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 6-3 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](#).

6.3.2 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**.
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 6-11](#) describes the directory parameters.

Table 6-11 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.

6.3.3 Developing Scripts

6.3.3.1 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**.
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 6-12](#) describes the script properties. Skip this step when creating a Flink SQL script.

Table 6-12 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">• Click . The Buy Queue page of DLI is displayed.• Go to the DLI console. <p>NOTE</p> <p>DLI provides the default resource queue default, which does not support insert, load, or cat commands.</p> <p>To set properties for submitting SQL jobs in the form of key/value, click . A maximum of 10 properties can be set. The properties are described as follows:</p> <ul style="list-style-type: none">• dli.sql.autoBroadcastJoinThreshold: specifies the data volume threshold to use BroadcastJoin. If the data volume exceeds the threshold, BroadcastJoin will be automatically enabled.• dli.sql.shuffle.partitions: specifies the number of partitions during shuffling.• dli.sql.cbo.enabled: specifies whether to enable the CBO optimization policy.• dli.sql.cbo.joinReorder.enabled: specifies whether join reordering is allowed when CBO optimization is enabled.• dli.sql.multiLevelDir.enabled: 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.• dli.sql.dynamicPartitionOverwrite.enabled: specifies that only partitions used during data query are overwritten and other partitions are not deleted.

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\";
```

 --example 1\;example 2.

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](#)).

NOTE

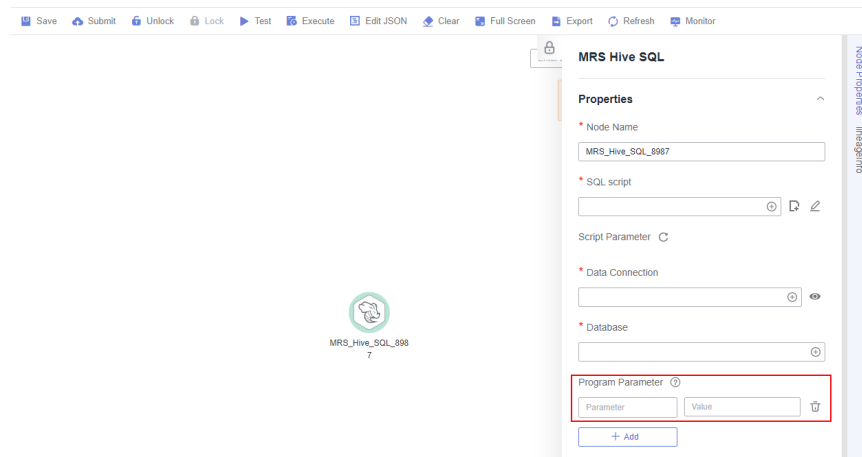
In the following script example, *str1* indicates the parameter name. It can contain only letters, numbers, hyphens (-), underscores (_), greater-than

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 6-4 Program Parameter




– Owner

Click **Basic Info** to set the script owner and description.

6. (Optional) In the upper part of the editor, click **Format** to format the SQL statement. When developing a Flink SQL script, skip this step.
7. 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.
8. Above the editor, click  to save the script.

If the script is created but not saved, set the parameters listed in [Table 6-13](#).

Table 6-13 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. A maximum of 1,000 query results and download results are supported.
- Dump result: Dump the CSV result files to OBS. For details, see [Table 6-14](#). The maximum number of dump results is not limited.

 **NOTE**

The execution results of Flink SQL scripts, RDS SQL scripts, and shell scripts cannot be dumped.

Table 6-14 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">• none• bzip2• deflate• gzip
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">• Overwrite: The existing folder will be overwritten by the customized folder.• Report: The system reports an error and suspends the export operation.

6.3.3.2 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 3-14](#).
- 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**.
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 6-15](#) describes the script properties.

Table 6-15 Shell script properties

Parameter	Description	Example
Host Connection	Selects the host where a shell script is to be executed.	N/A
Parameter	<p>Parameter transferred to the Shell script when it is executed. Parameters are separated by spaces, for example, a b c.</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 a, b, and c, 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 a b c and run the following Shell script, 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 1, 2, and 3 correspond to begin, end, and exit, respectively.</p> <ul style="list-style-type: none"> • When the interaction parameter is set to 1, the execution result is start something. • When the interaction parameter is set to 2, the execution result is stop something. • When the interaction parameter is set to 3, the execution result is exit. <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 6-16](#).

Table 6-16 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.

6.3.3.3 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 3-14](#).
- 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**.
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 6-17](#).

Table 6-17 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.

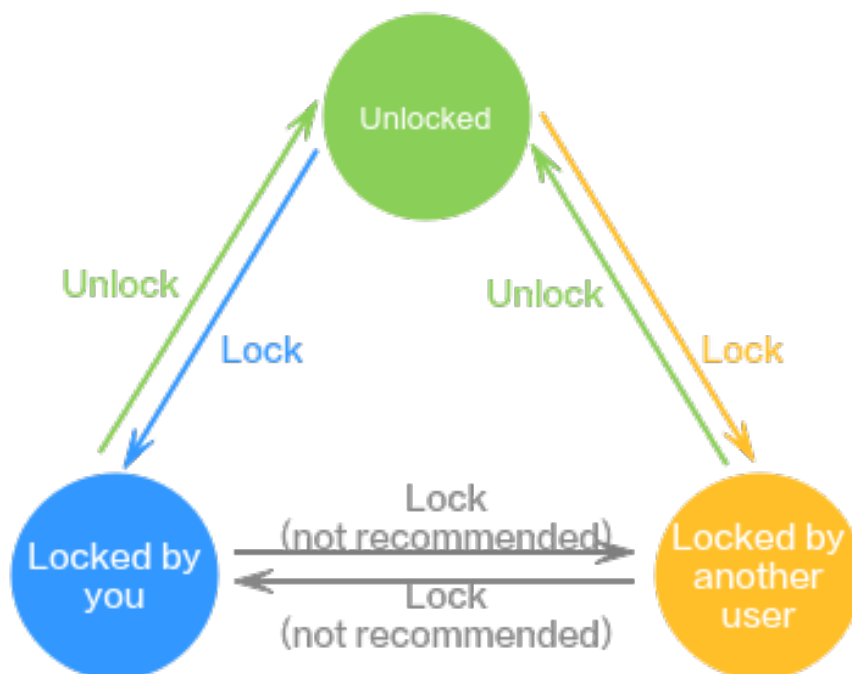
6.3.4 Submitting a Version and Unlocking the Script

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 6-5 Lock status**Prerequisites**

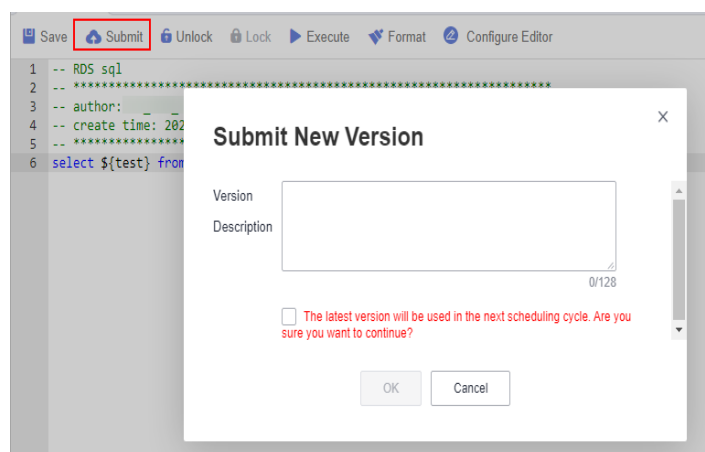
A script has been developed.

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.

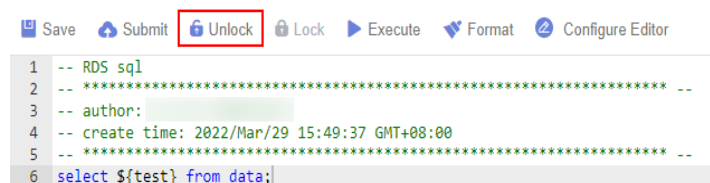
- Step 1** Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.
- Step 2** In the left navigation pane of DataArts Factory, choose **Data Development > Develop Script**.
- Step 3** In the script directory, 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 6-6 Submitting a version



- Step 5** In the upper part of the script editor, click **Unlock** to unlock the script.

Figure 6-7 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**.
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 6-8 Rolling back a version



Compare Version		Script	Operation
<input type="checkbox"/>	If you select only one version, the selected version is compared with the script in the development state. If you select two versions, the script contents of the two versions are compared.		Roll Back
<input type="checkbox"/>	5	Mar 02, 2021 16:18:22 GMT +...	Roll Back
<input type="checkbox"/>	4	Mar 02, 2021 16:16:46 GMT +...	Roll Back
<input type="checkbox"/>	3	Mar 02, 2021 16:16:22 GMT +...	Roll Back
<input type="checkbox"/>	2	Feb 23, 2021 18:39:16 GMT +...	Roll Back
<input type="checkbox"/>	1	Feb 23, 2021 18:37:35 GMT +...	Roll Back

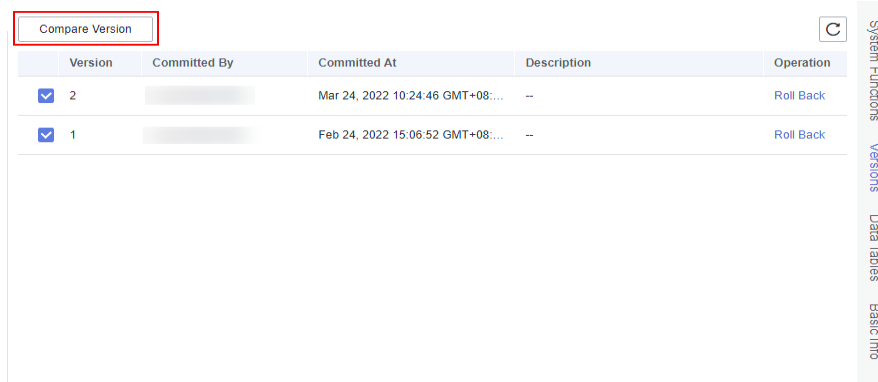
Version Comparison

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 6-9 Comparing versions



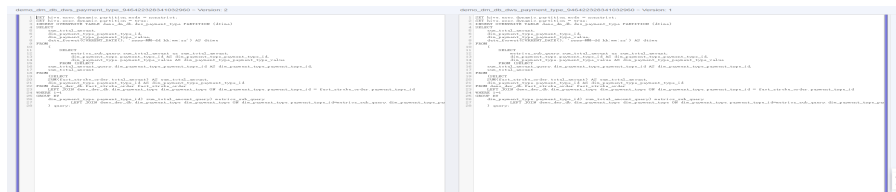
4. A new page is displayed, showing the script content of different versions on the left and right separately. The differences between the two versions have been marked. You can use the  and  buttons in the upper right corner to go to the previous or next change.

Figure 6-10 Version comparison details

6.3.5 (Optional) Managing Scripts

6.3.5.1 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**.
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 6-18](#) describes the parameters.

Table 6-18 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. NOTE 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**.

6.3.5.2 Copying the Script Name and Renaming a Script

You can copy the name of a script and rename a script.

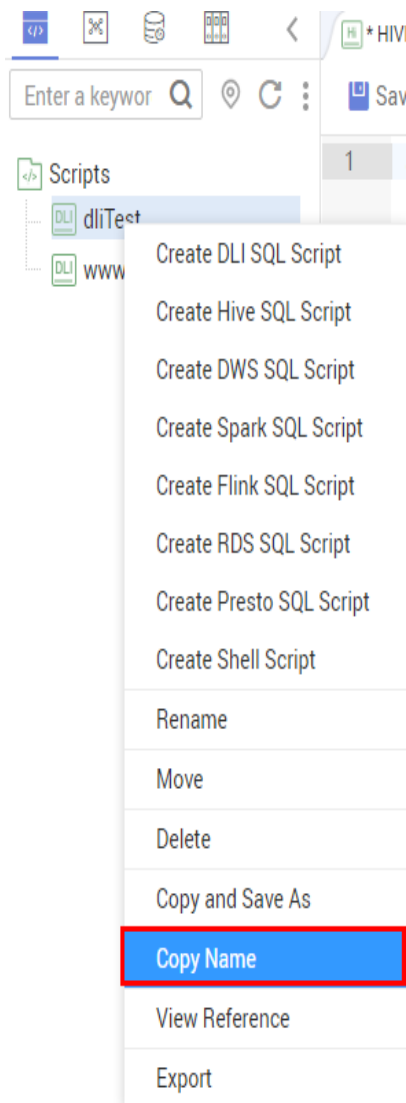
Prerequisites

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**.
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.

Figure 6-11 Copying the script name



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**.
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 6-12 Renaming a script

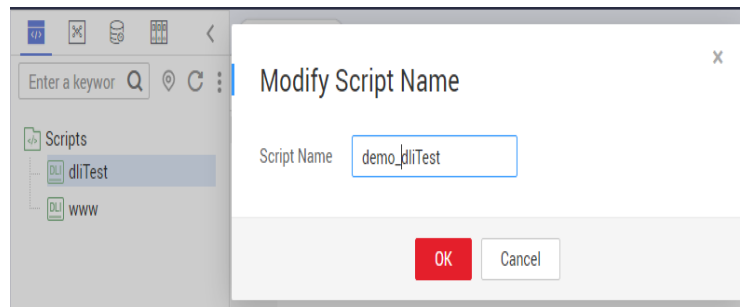


Table 6-19 Script renaming parameters

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**.

6.3.5.3 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](#).

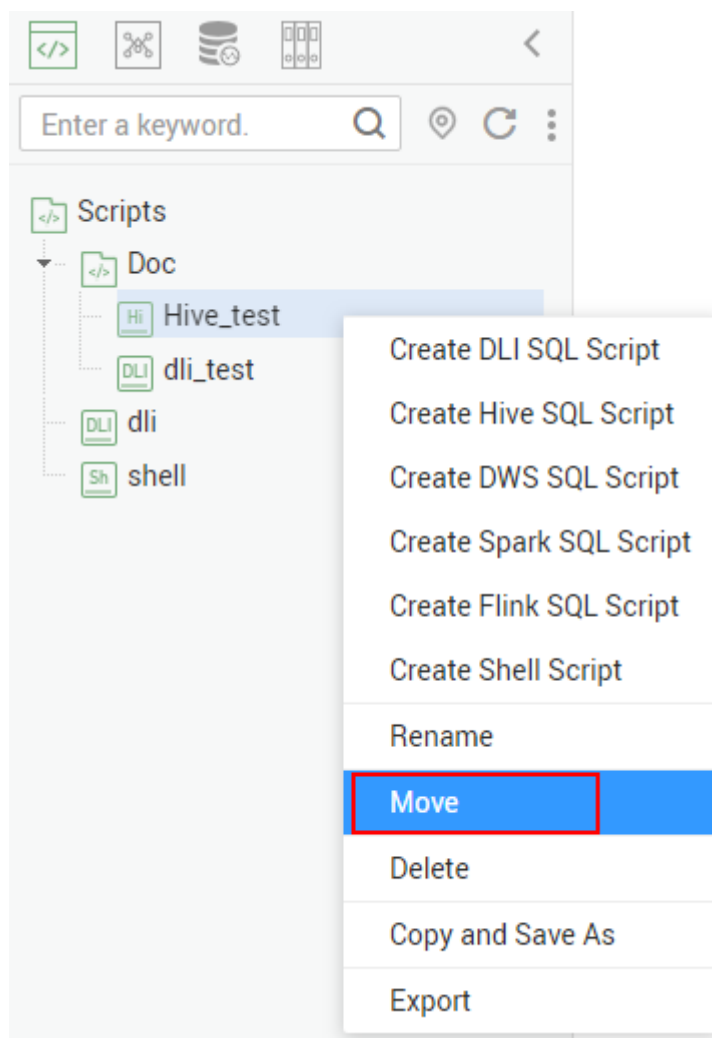
Procedure

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**.
2. In the left navigation pane of DataArts Factory, 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**.

Figure 6-13 Selecting Move



- b. In the displayed dialog box, configure related parameters. [Table 6-20](#) describes the parameters.

Figure 6-14 Moving a script

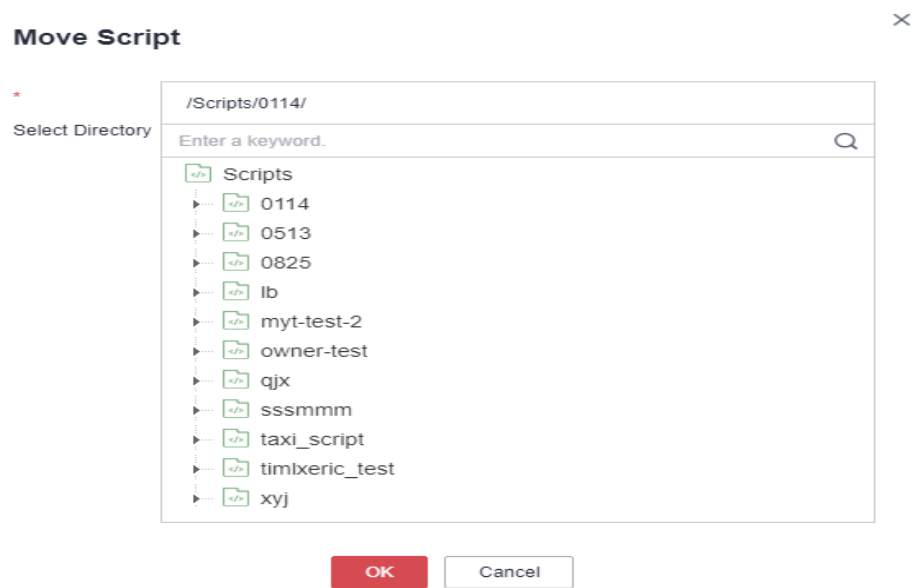


Figure 6-15 Move a directory

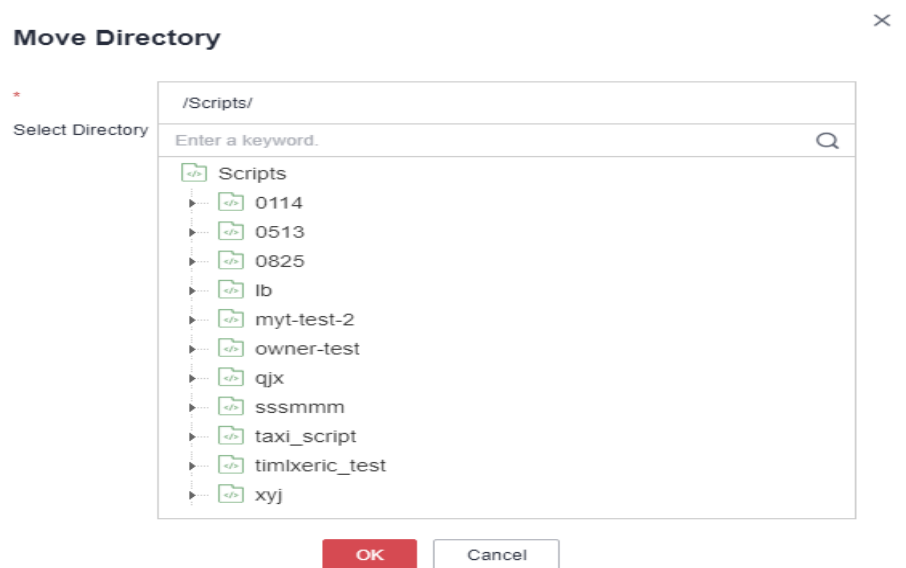


Table 6-20 Parameters for moving a script or directory

Parameter	Description
Select Directory	Directory to which the script or script directory is to be moved. The parent directory is the root 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.

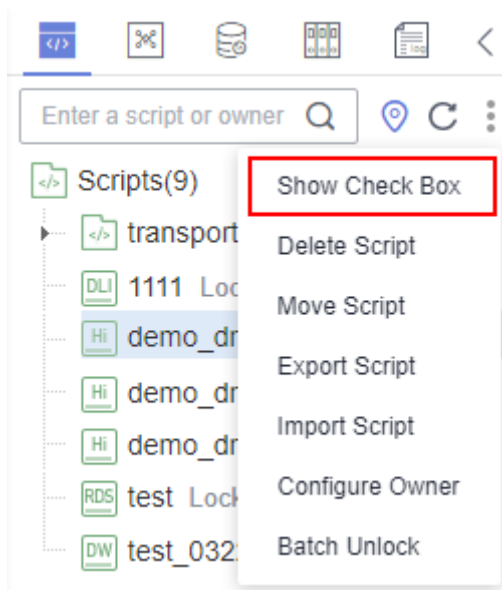
6.3.5.4 Exporting and Importing a Script

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 6-16 Clicking Show Check Box




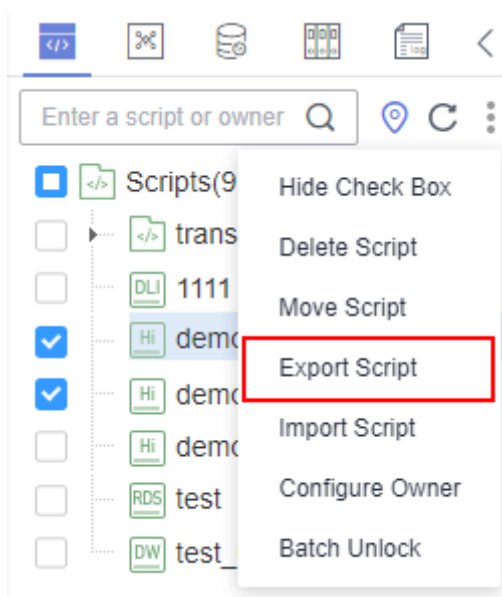
2. Select the scripts to be exported, click , and choose **Export Script**. After the export is successful, you can obtain the exported .zip file.


Figure 6-17 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.

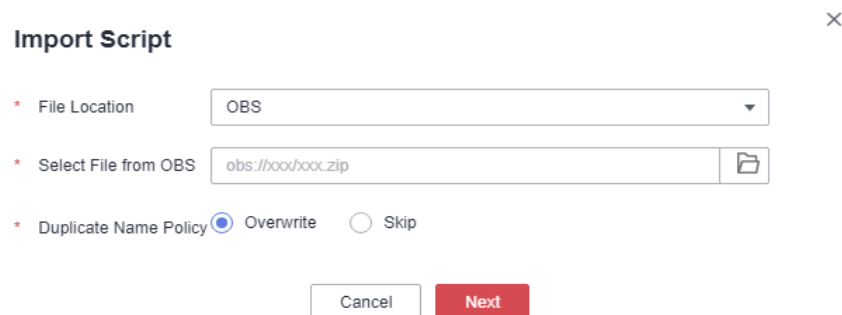
You can import one or more script files in the script directory. After a job is imported, the content in the development state is overwritten and a new version is automatically submitted.

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 6-18 Importing scripts



Import Script ×

* File Location

* Select File from OBS 

* Duplicate Name Policy Overwrite Skip

2. Click **Next**.

6.3.5.5 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**.
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**.

- In the displayed dialog box, you can view the references of a script or all the scripts in the folder.

Figure 6-19 References of a script

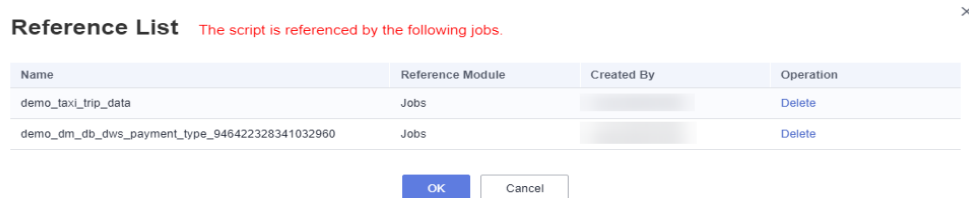
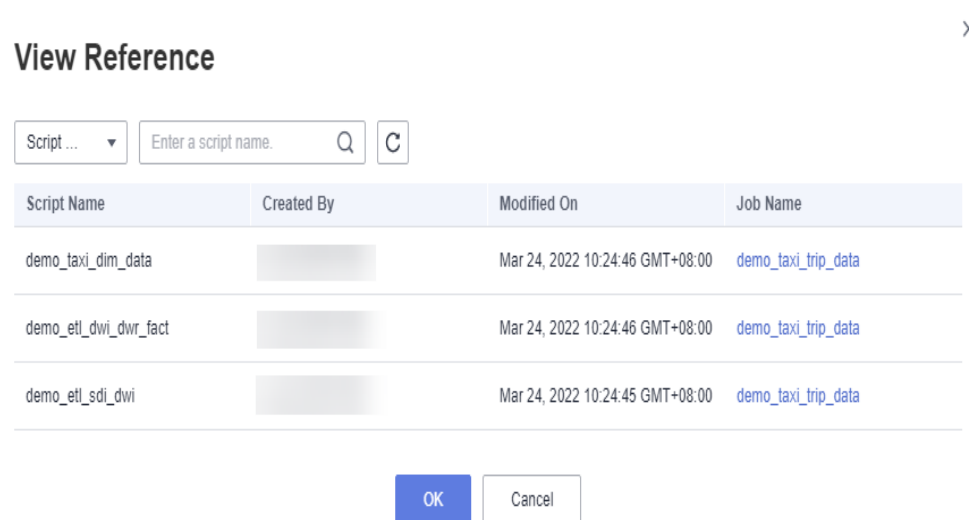


Figure 6-20 References of all the scripts in a folder



6.3.5.6 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

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

6.3.5.7 Changing the Script Owner

DataArts Factory allows you to change the owner for scripts with a few clicks.

Procedure


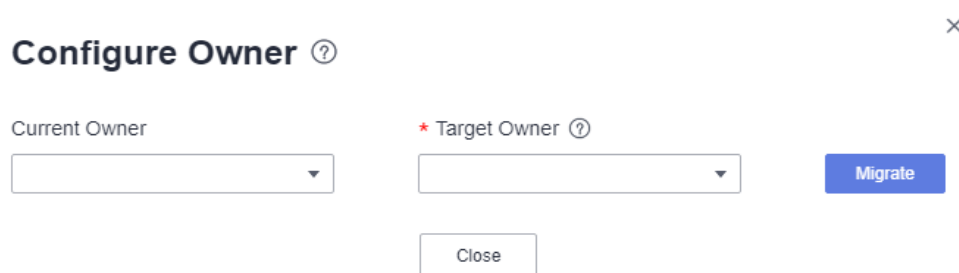
1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.
2. In the left navigation pane of DataArts Factory, choose **Data Development > Develop Script**.
3. At the top of the script directory, click  and select **Configure Owner**.

Figure 6-21 Configuring the owner



Configure Owner ⓘ

Current Owner

* Target Owner ⓘ

Migrate

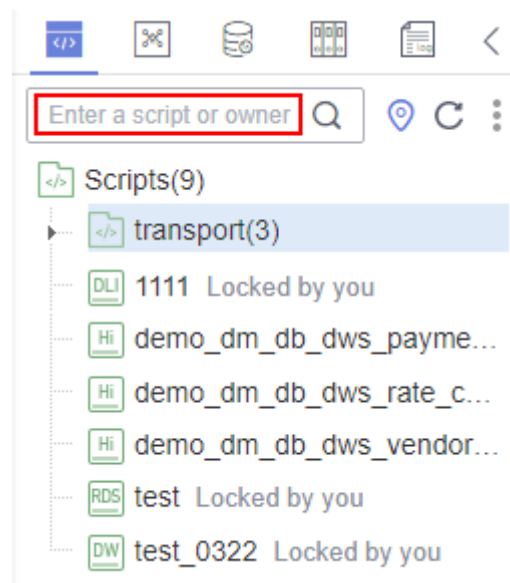
Close

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 6-22 Filtering scripts by owner



6.3.5.8 Unlocking Scripts

This section describes how to unlock scripts in batches.

Procedure


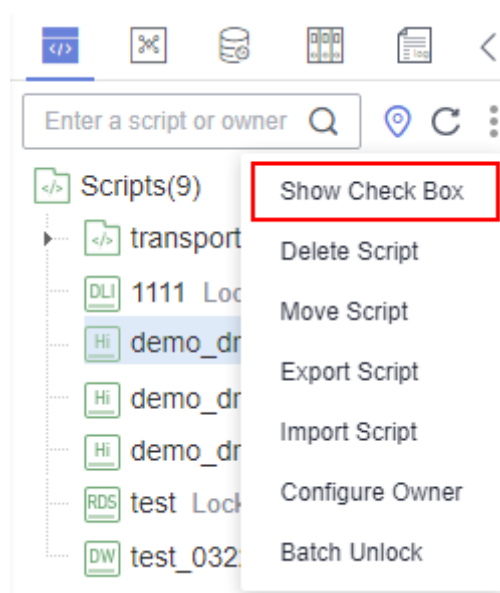
1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.
2. In the left navigation pane of DataArts Factory, choose **Data Development** > **Develop Script**.
3. Click  in the script directory and select **Show Check Box**.

Figure 6-23 Clicking Show Check Box




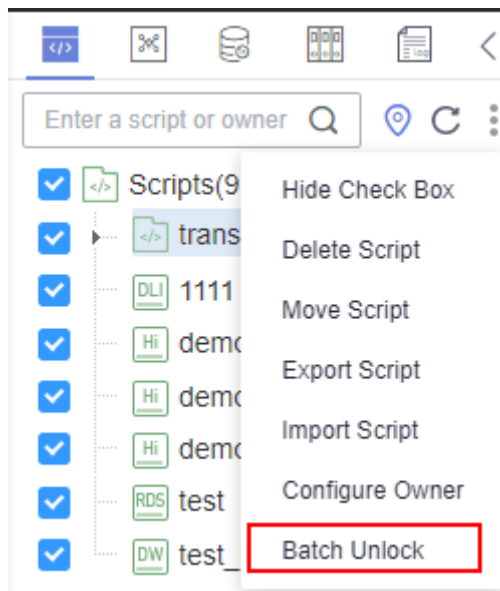
4. Select the scripts to unlock, click , and select **Batch Unlock**. The message "Unlocked." is displayed.

Figure 6-24 Batch Unlock



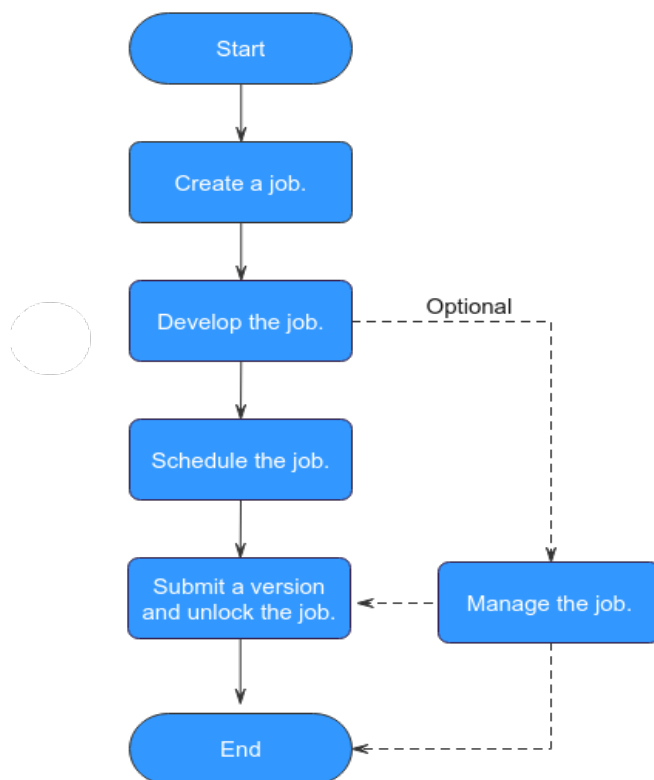
6.4 Job Development

6.4.1 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.

Figure 6-25 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](#).

6.4.2 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**.
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 6-21](#).

Table 6-21 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.

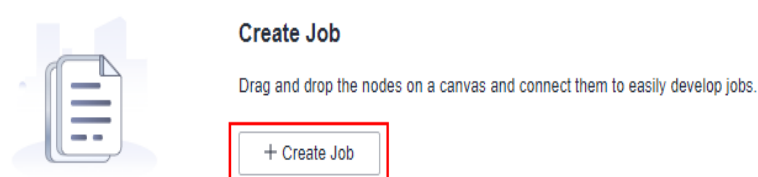
5. Click **OK**.

Creating a Job

The quantity of jobs is less than the maximum quota (10,000).

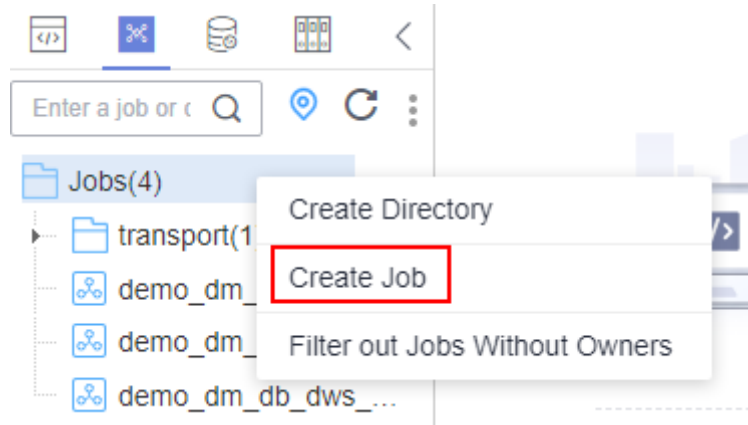
1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.
2. In the left navigation pane of DataArts Factory, choose **Development > Develop Job**.
3. Create a job using either of the following methods:
Method 1: On the **Develop Job** page, click **Create Job**.

Figure 6-26 Creating a job (method 1)



Method 2: In the directory list, right-click a directory and choose **Create Job** from the shortcut menu.

Figure 6-27 Creating a job (method 2)



4. In the displayed dialog box, configure job parameters. [Table 6-22](#) describes the job parameters.

Table 6-22 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 (.).

Parameter	Description
Processing Mode	<p>Type of the job.</p> <ul style="list-style-type: none">• Batch processing: 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 Setting Up Scheduling for a Job Using the Batch Processing Mode.• Real-time processing: 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 Setting Up Scheduling for Nodes of a Job Using the Real-Time Processing Mode.
Creation Method	<p>Selects a job creation mode.</p> <ul style="list-style-type: none">• Create Empty Job: Create an empty job.• Create Based on Template: Create a job using a template.
Select Directory	<p>Directory to which the job belongs. The root directory is selected by default.</p>
Owner	<p>Owner of the job.</p>
Priority	<p>Priority of the job. The value can be High, Medium, or Low.</p>
Agency	<p>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 (for details, see Configuring a Workspace-Level Agency), the job uses the workspace-level agency by default. You can also change the agency to a job-level agency by referring to Configuring a Job-level Agency.</p> <p>NOTE Job-level agency takes precedence over workspace-level agency.</p>

Parameter	Description
Log Path	Selects the OBS path to save job logs. By default, logs are stored in a bucket named dlf-log-<i>{Projectid}</i> . NOTE <ul style="list-style-type: none">If you want to customize a storage path, select the bucket that you have created on OBS by following the instructions provided in (Optional) Changing a Job Log Storage Path.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.

5. Click **OK**.


6.4.3 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**.
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 6-28 Compiling a job



5. Configure node functions. Right-click a node icon on the canvas and select a function as needed. [Table 6-23](#) lists the available functions.

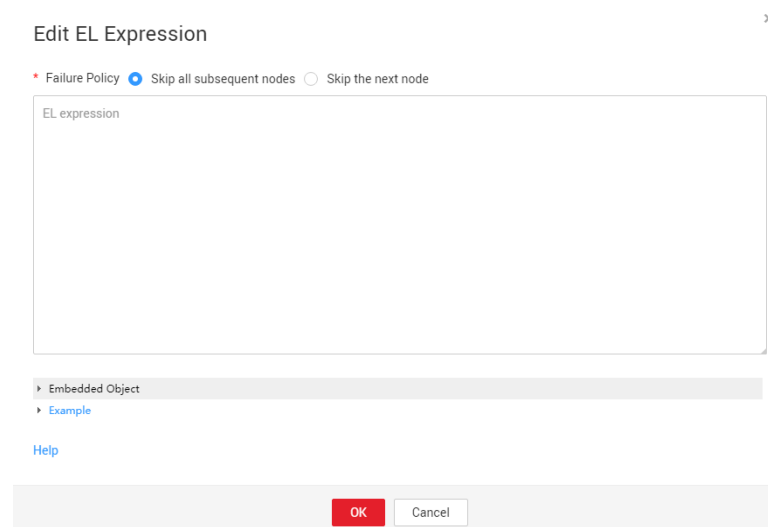
Table 6-23 Node functions

Function	Description
Configure	Goes to the Node Property page of the node.
Delete	<p>Deletes one or more nodes at the same time.</p> <ul style="list-style-type: none"> • Deleting one node: Right-click the node icon in the canvas and choose Delete or press the Delete shortcut key. • Deleting multiple nodes: Click the icons of the nodes to be deleted in the canvas while holding on Ctrl, right-click the blank area of the current job canvas, and choose Delete or press the Delete shortcut key.
Copy	<p>Copies one or more nodes to any job.</p> <ul style="list-style-type: none"> • Single-node copy: You can either right-click the node icon in the canvas, choose Copy, and paste the node to a target location, or click the node icon in the canvas and press Ctrl+C and Ctrl+V to paste the node to a target location. The copied node carries the configuration information of the original node. • Multi-node copy: Click the icons of the nodes to be copied in the canvas while holding on Ctrl. Then you can either right-click the blank area of the canvas, choose Copy, and paste the nodes to a target location, or press Ctrl+C and Ctrl+V 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.
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.

Function	Description
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 View Job Log 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](#). For details about how to use IF conditions, see [IF Condition Judgment](#).

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 6-24](#).

Table 6-24 Basic job information






Parameter	Description
Owner	An owner configured during job creation is automatically matched. This parameter value can be modified.
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 Add to add a tag to the job. You can also select a tag configured in Managing Job Labels .

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 6-25](#).

Table 6-25 Job parameter setup


Function	Description
Variable Parameter	
Add	<p>Click Add and enter the variable parameter name and parameter value in the text boxes.</p> <ul style="list-style-type: none"> Parameter Name Only letters, numbers, hyphens, and underscores (_) are allowed. Parameter Value <ul style="list-style-type: none"> The string type of parameter value is a character string, for example, str1. The numeric type of parameter value is a number or operation expression. <p>After the parameter is configured, it is referenced in the format of $\\${parameter\ name}$ in the job.</p>
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	 <p>Click  next to the parameter name and value text boxes to delete the job parameter.</p>
Constant Parameter	
Add	<p>Click Add and enter the constant parameter name and parameter value in the text boxes.</p> <ul style="list-style-type: none"> Parameter name Only letters, numbers, hyphens, and underscores (_) are allowed. Parameter value <ul style="list-style-type: none"> The string type of parameter value is a character string, for example, str1. The numeric type of parameter value is a number or operation expression. <p>After the parameter is configured, it is referenced in the format of $\\${parameter\ name}$ in the job.</p>
Modify	Modify the parameter name and parameter value in text boxes and save the modifications.
Delete	 <p>Click  next to the parameter name and value text boxes to delete the job parameter.</p>

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

6.4.4 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 6-26](#).

Table 6-26 Job scheduling parameters

Parameter	Description
Scheduling Type	Scheduling type of the job. Available options include: <ul style="list-style-type: none">• Run once: You need to manually execute the job.• Run periodically: The job is executed periodically. For details about the parameters, see Table 6-27.• Event-based: The job will be executed when certain external conditions are met. For details about the parameters, see Table 6-28.
Dry run	If you select this option, the job will not be executed, and a success message will be returned.

Table 6-27 Parameters for jobs that are executed periodically

Parameter	Description
From and to	The period during which a scheduling task takes effect.

Parameter	Description
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"> ● Minutes: 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. ● Hours: 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. ● Every day: The job starts at a specified time on a day. The scheduling period is one day. ● Every week: You can select a specified time point of one or more days in a week. ● Every month: You can select a specified time point of one or more days in a month.

Parameter	Description
Dependency job	<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 Job Dependency.</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 How a Job Runs After a Dependency Job Is Set for It for details.).</p> <p>The constraints are as follows:</p> <ul style="list-style-type: none"> • 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.
Policy for Current job If Dependency job Fails	<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"> • 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. • Continue Continues to execute the current job. • Terminate Stops executing the current job. Its status becomes Canceled. <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"> • If the value of this parameter is set to Terminate, the current job will be terminated as long as one of the 12 instances of its dependency job fails. • If the value of this parameter is set to Continue, the current job will be executed after the 12 instances of its dependency job are executed. <p>NOTE You can set this parameter for multiple jobs in a batch. For details, see Configuring a Default Item.</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 How a Job Runs After a Dependency Job Is Set for It 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"> • Independent on the previous schedule cycle: You can set Concurrency to set the number of job instances that are concurrently executed. If you set it to 1, a batch is executed only after the previous batch is executed (the execution is successful, cancelled, or failed). • Self-dependent (The current job can continue to run only after the previous schedule cycle is successfully finished.)

Table 6-28 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"> • KAFKA
Parameters for KAFKA event-triggered jobs	
Connection Name	Before selecting a data connection, ensure that a Kafka data connection has been created in the Management Center .
Topic	Topic of the message to be sent to the Kafka.
Concurrent Events	Number of jobs that can be concurrently processed. The maximum number of concurrent events is 128.
Event Detection Interval	Interval at which the system detects the stream for new messages. The unit of the interval can be Second or Minute .
Access Policy	<p>Select the location where data is to be accessed:</p> <ul style="list-style-type: none"> • Access from the last location: 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. • Access from a new location: Data is accessed from the most recently recorded location each time.

Parameter	Description
Failure Policy	Select a policy to be performed after scheduling fails. <ul style="list-style-type: none"> • Suspend • Ignore the failure and proceed with the next event

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 6-29](#).

Table 6-29 Parameters for setting up node scheduling

Parameter	Description
Scheduling Type	Scheduling type of the job. Available options include: <ul style="list-style-type: none"> • Run once: You need to manually run the job. • Run periodically: The job runs automatically and periodically. • Event-based: The job runs when certain external conditions are met.
Parameters displayed when Scheduling Type is Run periodically	
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"> • Minutes • Hours • Every day • Every week • Every month 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"> • Independent on the previous schedule cycle • Self-dependent (The current job can continue to run only after the previous schedule cycle is successfully finished.)
Parameters displayed when Scheduling Type is Event-based	

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 Management Center .
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>NOTE</p> <ol style="list-style-type: none">1. A consumer group can contain multiple consumers.2. The group ID is a string that uniquely identifies a consumer group in a Kafka cluster.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. <p>If you select KAFKA for Event Type, 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 Seconds or Minutes .
Failure Policy	Select a policy to be performed after scheduling fails. <ul style="list-style-type: none">• Suspend• Ignore failure and proceed

6.4.5 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

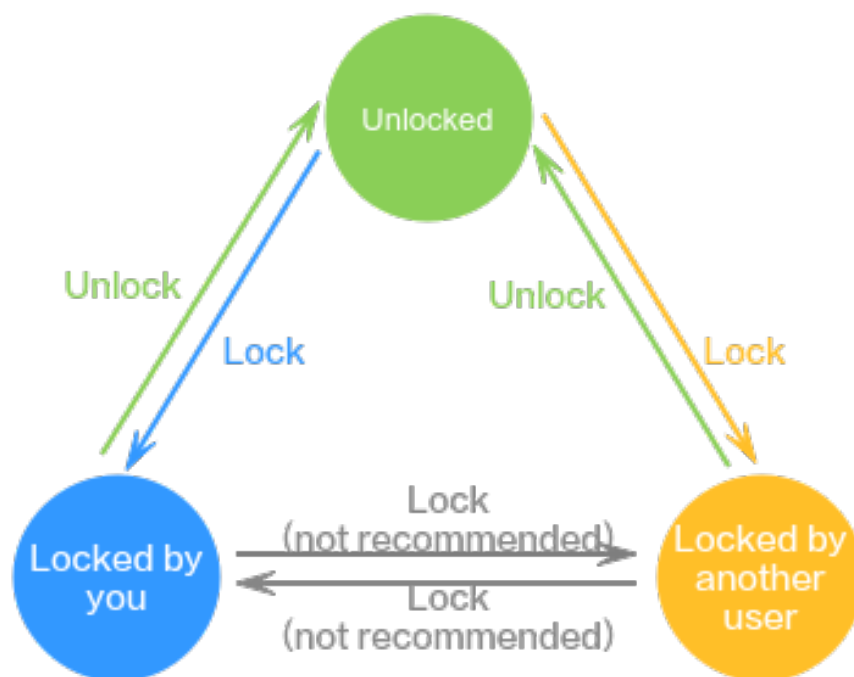
scripts or jobs will be used in scenarios such as script invocation, instance rerunning, and job dependency and patch data configuration.

- 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 6-29 Lock status



Prerequisites

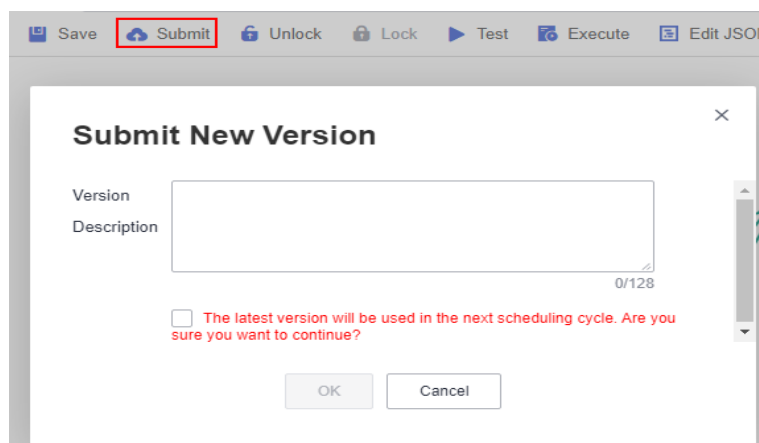
A job has been developed.

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**.
- 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 6-30 Submitting a version



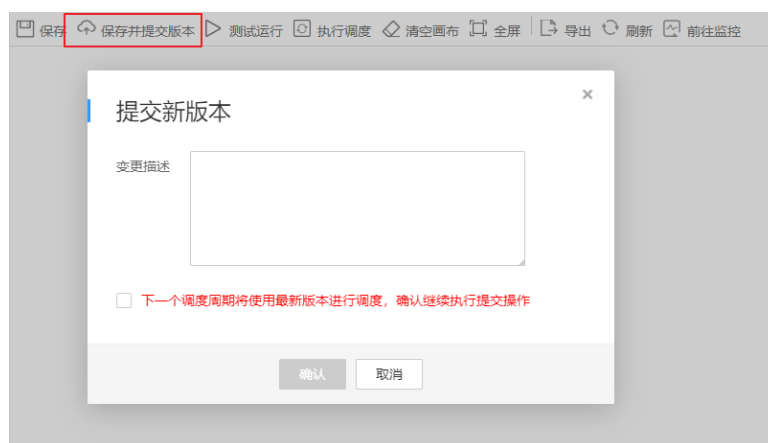
Step 5 Above the job canvas, click **Unlock** to unlock the job.

Figure 6-31 Unlocking a job



Step 6 Above the job canvas, click **Save and 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 6-32 Submitting a new version



----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**.
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 6-33 Rolling back the version

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

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**.
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 6-34 Viewing version details

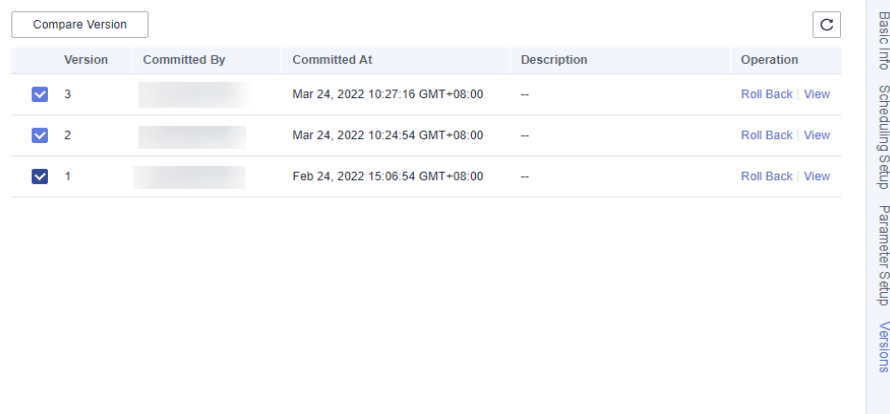
Version	Committed By	Committed At	Description	Operation
11		Mar 06, 2021 14:26:01 GMT +...		Roll Back View
10		Mar 06, 2021 11:33:19 GMT +...		Roll Back View
9		Mar 06, 2021 11:04:44 GMT +...		Roll Back View
8		Mar 06, 2021 11:04:26 GMT +...		Roll Back View
7		Mar 06, 2021 11:04:17 GMT +...		Roll Back 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**.
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 6-35 Comparing versions



Version	Committed By	Committed At	Description	Operation
<input checked="" type="checkbox"/> 3		Mar 24, 2022 10:27:16 GMT+08:00	--	Roll Back View
<input checked="" type="checkbox"/> 2		Mar 24, 2022 10:24:54 GMT+08:00	--	Roll Back View
<input checked="" type="checkbox"/> 1		Feb 24, 2022 15:06:54 GMT+08:00	--	Roll Back View

6.4.6 (Optional) Managing Jobs

6.4.6.1 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**.
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 6-30](#) describes the parameters.

Table 6-30 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**.

6.4.6.2 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**.
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**.
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 6-31 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**.

6.4.6.3 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**.
2. In the left navigation pane of DataArts Factory, choose **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 6-32 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 root 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.

6.4.6.4 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.

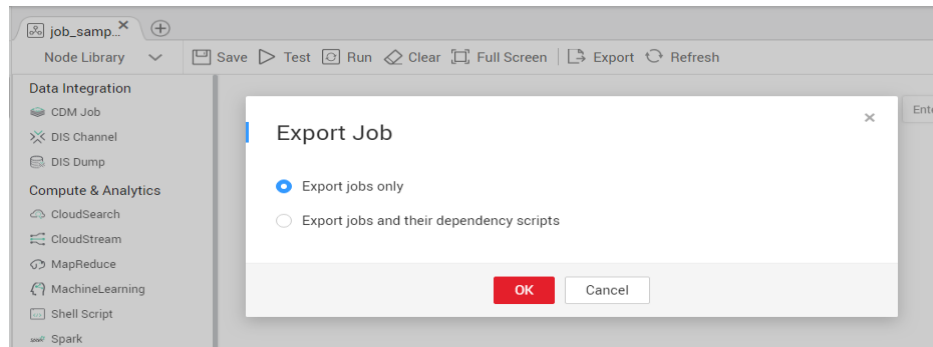
Step 1 Double-click a job name to access the development page of the job, click , and select the type of the job to be exported.

- **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 6-36 Exporting a job (method 1)



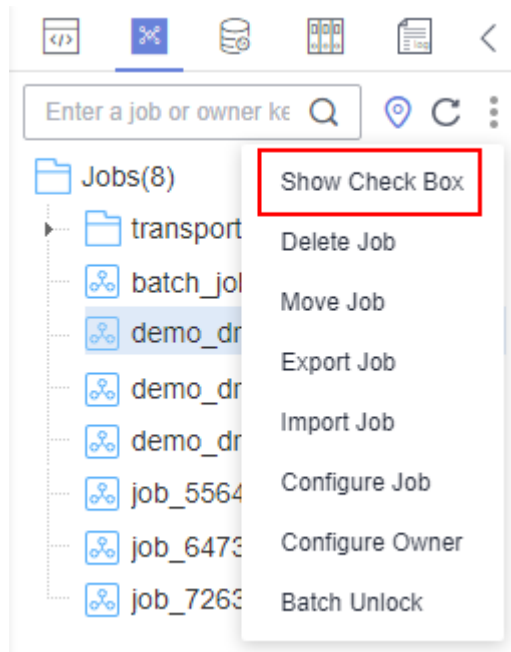
Step 2 Click **OK** to export the required job file.


----End

Method 2: Export one or more jobs from the job directory.

Step 1 Click  in the job directory and select **Show Check Box**.

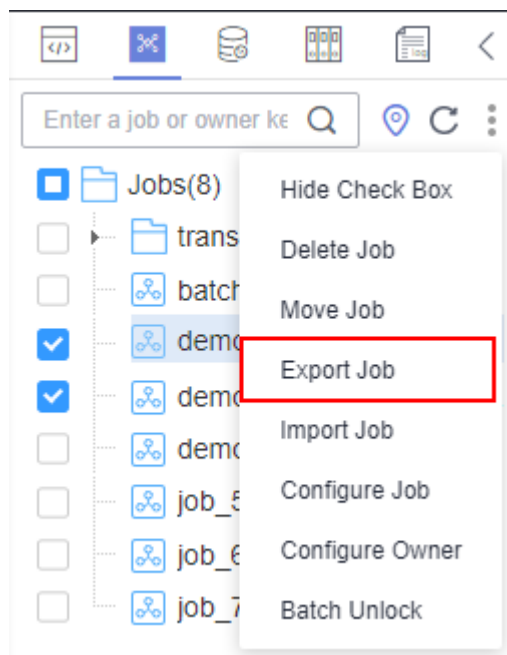
Figure 6-37 Clicking Show Check Box



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 6-38 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.

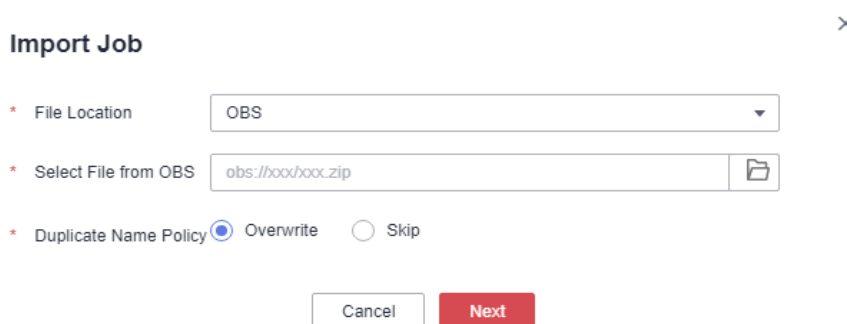
Export one or more jobs from the job directory.

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 6-39 Importing job definitions and dependencies



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**.
 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**.

6.4.6.5 Configuring Jobs


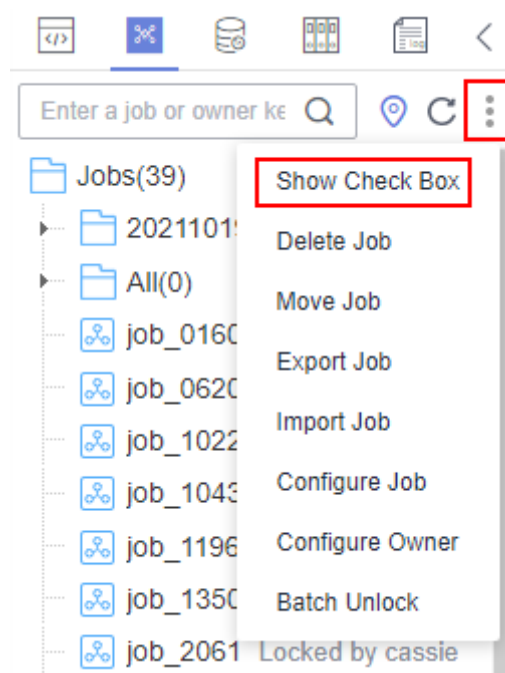
1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.
2. In the left navigation pane of DataArts Factory, choose **Development > Develop Job**.
3. Click  in the job directory and select **Show Check Box**.

Figure 6-40 Clicking Show Check Box




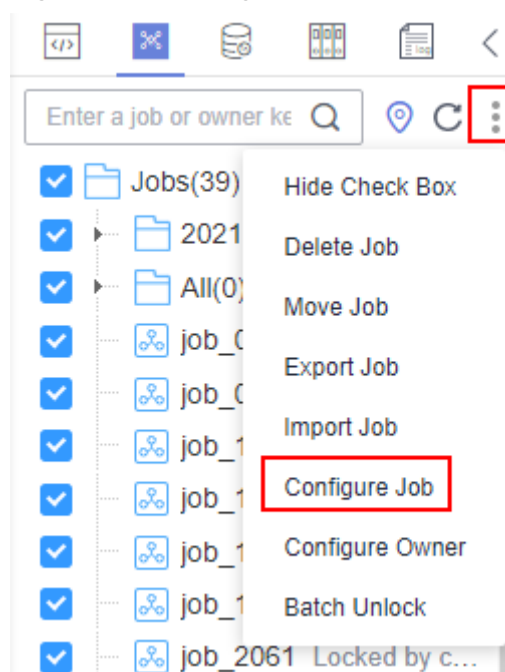
4. Select jobs, click , and select **Configure Job**.

Figure 6-41 Configure Job



5. Configure general parameters for the jobs.

Figure 6-42 General Configuration

Configure Job ×

Changes to the configuration apply to all jobs, regardless of whether the jobs have been submitted.

- General Configuration
 CDM Cluster
 DLI Queue

Node Status Polling Interval (s) Keep it unchanged



Max. Node Execution Duration



Keep it unchanged
Day

Job Agency Keep it unchanged Select an agency. +

Retry upon Failure Yes No Keep it unchanged

- Failure Policy ?
- Suspend execution plans of the subsequent nodes
 - End the current job execution plan
 - Go to the next node. ?
 - Suspend current job execution plan
 - Keep it unchanged

Owner ? Keep it unchanged

OK
Cancel

Table 6-33 General Configuration

Parameter	Description
Node Status Polling Interval	How often the system checks whether all the nodes are executed. The value ranges from 1 to 60 seconds. If you select Keep it unchanged , the poll interval remains unchanged for the nodes.
Max. Node Execution Duration	Maximum duration of executing the nodes of a job. When Retry upon Failure is set to Yes for a node, the node can be re-executed for numerous times upon an execution failure within the maximum duration. If you select Keep it unchanged , the poll interval remains unchanged for the nodes.
Job Agency	During execution of the jobs, the agency is used to communicate with other services. If you select Keep it unchanged , the agency remains unchanged for the jobs.

Parameter	Description
Retry upon Failure	Whether to re-execute the nodes of the selected jobs if the nodes fail to be executed. If you select Keep it unchanged , the retry policy remains unchanged for the nodes.
Failure Policy	Operation to be performed if all nodes of the selected jobs fail to be executed. If you select Keep it unchanged , the failure policy remains unchanged for the nodes.
Owner	Owner of the selected jobs, which can only be a member of the current workspace. If you select Keep it unchanged , the own remains unchanged for the jobs.
Concurrent Periodic Job Instances	Number of jobs that can be handled concurrently If you select Keep it unchanged , the number of concurrent periodic job instances remains unchanged.

6. Select **CDM Cluster** and configure the CDM cluster for the CDM Job node of the selected jobs.

Select the current CDM cluster from the drop-down list box on the left, and select the target CDM cluster from the drop-down list box on the right.

 **NOTE**




1. Before migrating a CDM cluster, you must create a job with the same name in the new cluster.
 2. Configure two CDM clusters for a CDM job.
 - If you select one of the source clusters, only the selected cluster will be migrated.
 - If you select both source clusters, they will be both migrated to the destination cluster.
- Search: Enter a job name and click  to filter out the jobs that contain the CDM Job node.
 - Refresh: Click  to refresh the list of jobs that contain the CDM Job node.
 - Download: Click  to download the selected jobs.

Figure 6-43 CDM Cluster

Configure Job ×

Changes to the configuration apply to all jobs, regardless of whether the jobs have been submitted.

General Configuration
 CDM Cluster
 DLI Queue

i You can only change the CDM cluster for CDM Job nodes. CDM jobs of the source cluster will not be automatically moved to the target cluster. You must export the CDM jobs from the source cluster and import them to the target cluster before changing the CDM cluster. ×

All ⇒ Keep it unchanged

Enter a keyword.

<input checked="" type="checkbox"/>	Job Name	Node Name	Sche...	Scheduled At	In-P...	CDM Cluster
<input checked="" type="checkbox"/>	guowangTest	kafka2hdfs			No	cdm-cdc
<input checked="" type="checkbox"/>	qxjForeach	11112	1 days	00:00:00	No	cdm-260-300...
<input checked="" type="checkbox"/>	guowangTest_...	kafka2hdfs			No	cdm-cdc

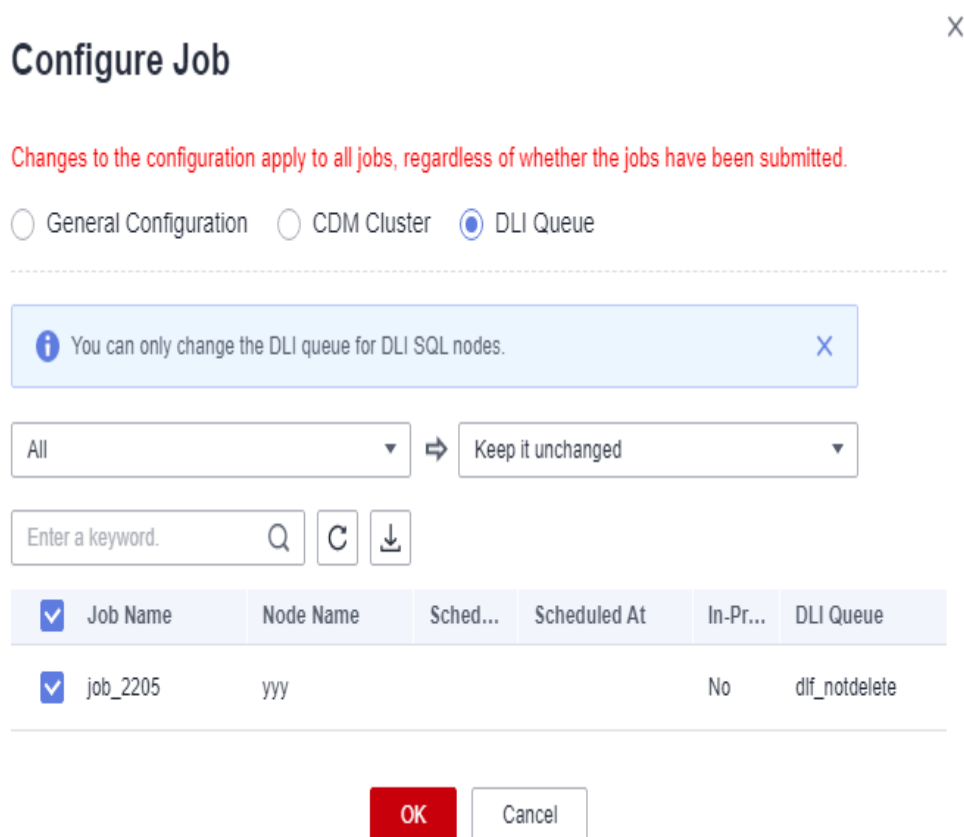
7. Select **DLI Queue** and configure the DLI queue of the DLI SQL node of the selected jobs.

Select the current DLI queue from the drop-down list box on the left, and select the target DLI queue from the drop-down list box on the right.

NOTE

- Search: Enter a job name and click to filter out the jobs that contain the DLI SQL node.
- Refresh: Click to refresh the list of jobs that contain the DLI SQL node.
- Download: Click to download the selected jobs.

Figure 6-44 DLI Queue



8. Click **OK**.

6.4.6.6 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**.
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**.

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**.

6.4.6.7 Changing the Job Owner

DataArts Factory allows you to change the owner for jobs with a few clicks.

Procedure


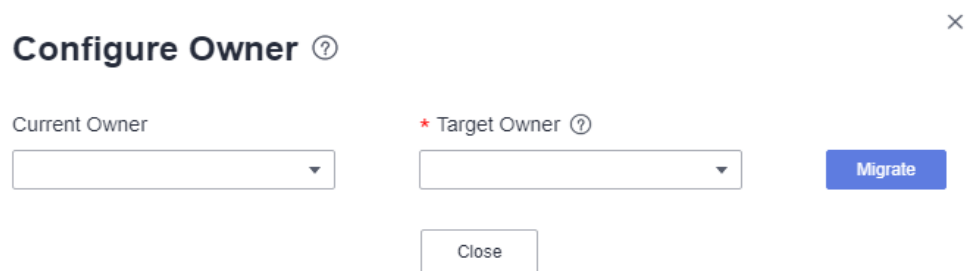
1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.
2. In the left navigation pane of DataArts Factory, choose **Development > Develop Job**.
3. At the top of the job directory, click  and select **Configure Owner**.

Figure 6-45 Configuring the owner



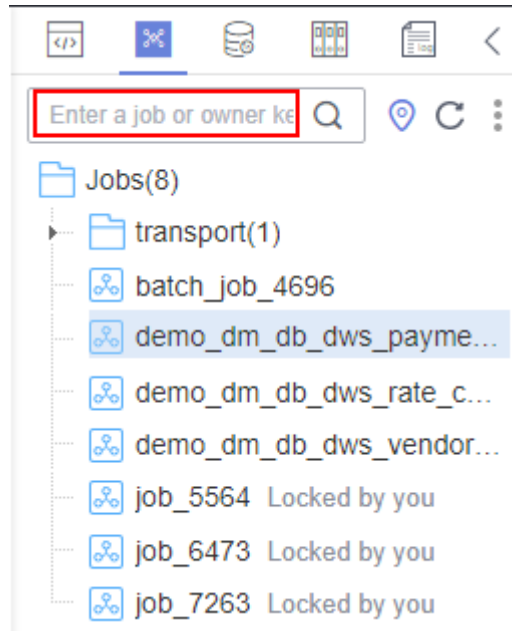
The screenshot shows a dialog box titled "Configure Owner" with a close button (X) in the top right corner. Inside the dialog, there are two dropdown menus: "Current Owner" and "* Target Owner" (with a help icon). To the right of the "Target Owner" dropdown is a blue "Migrate" button. Below the dropdowns is a "Close" button.

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 6-46 Filtering jobs by owner



6.4.6.8 Unlocking Jobs

This section describes how to unlock jobs in batches.

Procedure


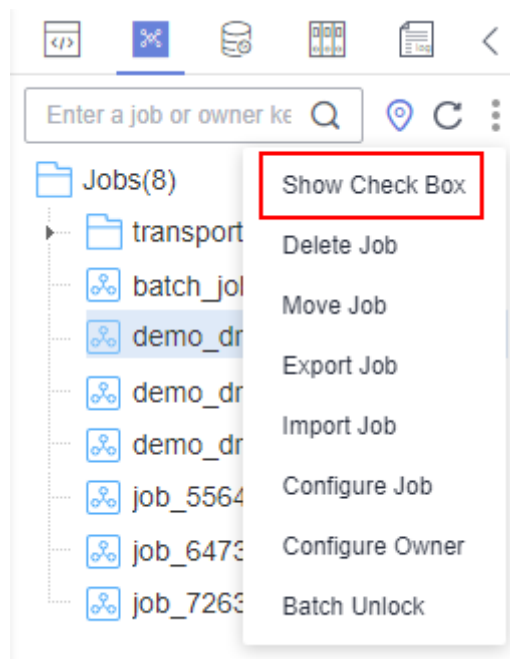
1. Log in to the DataArts Studio console. Locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Factory**.
2. In the left navigation pane of DataArts Factory, choose **Development > Develop Job**.
3. Click  in the job directory and select **Show Check Box**.

Figure 6-47 Clicking Show Check Box




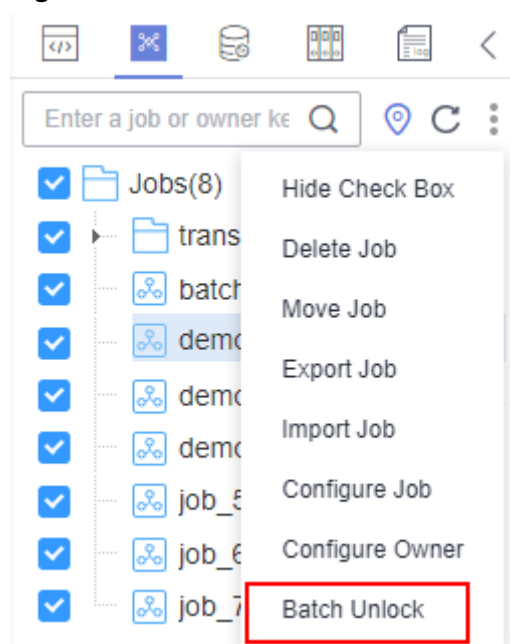
4. Select the jobs to unlock, click , and select **Batch Unlock**. The message "Unlocked." is displayed.

Figure 6-48 Batch Unlock



6.5 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**.
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 6-34](#) describes the solution parameters.

Table 6-34 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.


6.6 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**.
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**.

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.

6.7 O&M and Scheduling

6.7.1 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.

6.7.2 Monitoring a Job

6.7.2.1 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 6-35](#).

Figure 6-49 Monitoring a Batch Job



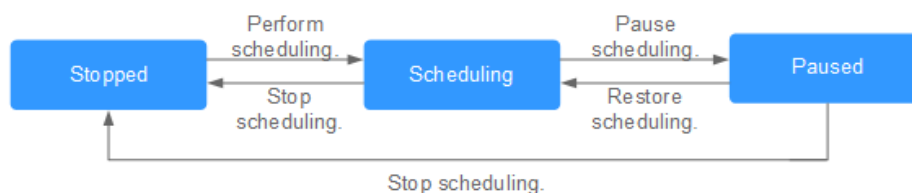
Name	Sche.	Sched.	Sched.	Priority	Next	Start	End T.	Owner	Last	Create	Last In.	Last In.	Failure	Job La.	Job Ag.	Next S.	Operation
test1111	Sch.	Run p.	2 min.	High	Sep 1.	May 1.	-	dgc_1.	dgc_test	Runnin.	Sep 15.	Sep 11.	-	-	-	1	Pause More

Table 6-35 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.
4	Viewing job instance status	Click  in front of the job name. The Last Instance 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 Operation column of a job, you can run, pause, recover, stop, and configure scheduling. For details, see Batch Job Monitoring: Scheduling a Job .
7	Configuring notifications	In the Operation column of a job, choose More > Set Notification . In the displayed dialog box, configure notification parameters. Table 6-45 describes the notification parameters.
8	Monitoring instances	In the Operation column of a job, choose More > Monitor Instance to view the running records of all instances of the job.
9	PatchData	In the Operation column of a job, choose More > PatchData . For details, see Batch Job Monitoring: PatchData .
1 0	Adding a job label	In the Operation column of a job, choose More > Add Job Label . For details, see Batch Job Monitoring: Adding a Job Label .

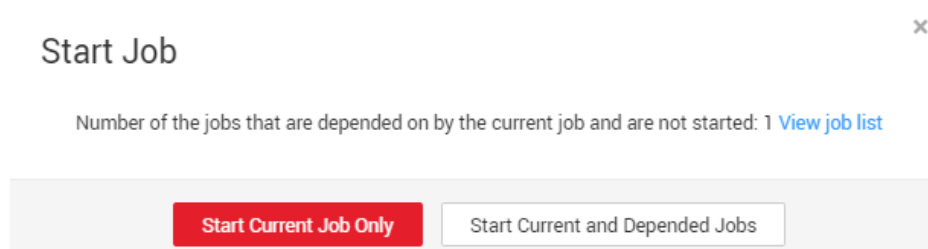
Batch Job Monitoring: Scheduling a Job

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 6-50 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**.
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 6-51 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**.
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 6-36](#).

Figure 6-52 PatchData parameters

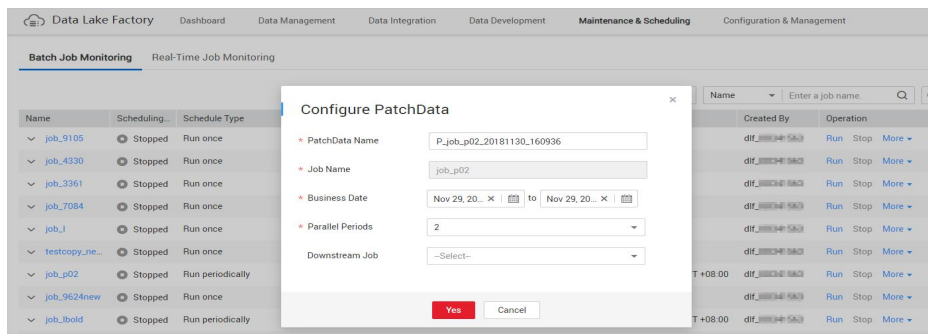


Table 6-36 Parameters

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.
Date	Period of time when PatchData is required. NOTE 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.
Parallel Instances	Number of instances to be executed at the same time. A maximum of five instances can be executed at the same time. NOTE 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.
Downstream Job Requiring PatchData	Select the downstream jobs (jobs that depend on the current job) that require PatchData. You can select multiple jobs.

6. Click **OK**. The system starts to perform PatchData and the **PatchData Monitoring** page is displayed.

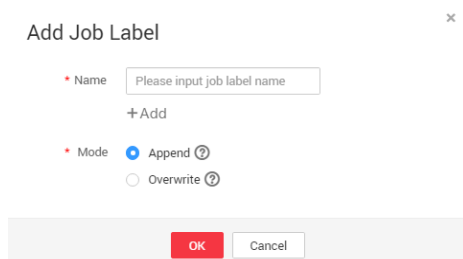
Batch Job Monitoring: Adding a Job Label

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**.
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 6-53 Parameters for adding a job label



6. Click **OK**.

6.7.2.2 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 6-37](#).

Figure 6-54 Real-time job monitoring page

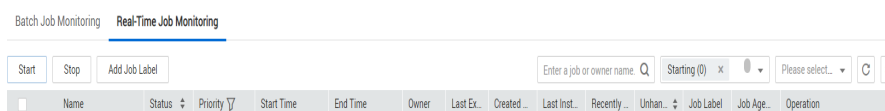



Table 6-37 Operations supported by real-time job monitoring

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.

No.	Operation	Description
4	Viewing job instance status	Click job in front of the  name. The Last Instance page is displayed. You can view information about the last instance of the job.
5	Job status-related operations	In the Operation column of the job, you can start, pause, recover, and stop job scheduling.
6	Adding a job label	In the Operation column of a job, choose More > Add Job Label .
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. NOTE 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 Disable . After the node is disabled, you can right-click it and select Restore to restore it on another location. For details, see Real-Time Job Monitoring: Disabling and Restoring a Node .
9	Viewing the boot log	Click a job name. On the displayed page, right-click a node and select View Run Log 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 Configure Scheduling to view and modify the scheduling information about the node. For details, see Real-Time Job Monitoring: Configuring Scheduling for a Node Where Event-driven Scheduling Is Configured .
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 Real-Time Job Monitoring: Monitoring Subjobs .
12	Clearing stream messages	Click a job name. On the displayed page, right-click the node where event-driven scheduling is configured and select Clear Stream Message .

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**.

2. In the left navigation pane of DataArts Factory, 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 6-38](#).

Table 6-38 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">● Paused node● The first node of the subjob
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**.
2. In the left navigation pane of DataArts Factory, 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 6-39](#).

Figure 6-55 Configuring scheduling

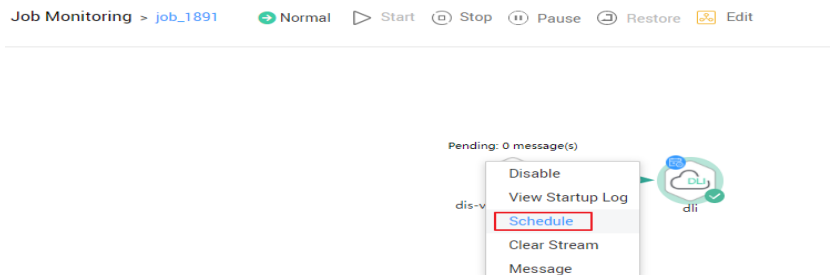


Table 6-39 Policy parameters

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 Second or Minute .
Failure Policy	Select a policy to be performed after scheduling fails. <ul style="list-style-type: none"> • Stop scheduling • Ignore failure and proceed

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**.
2. In the left navigation pane of DataArts Factory, 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, as shown in [Figure 6-56](#).

Figure 6-56 Subjob monitoring page

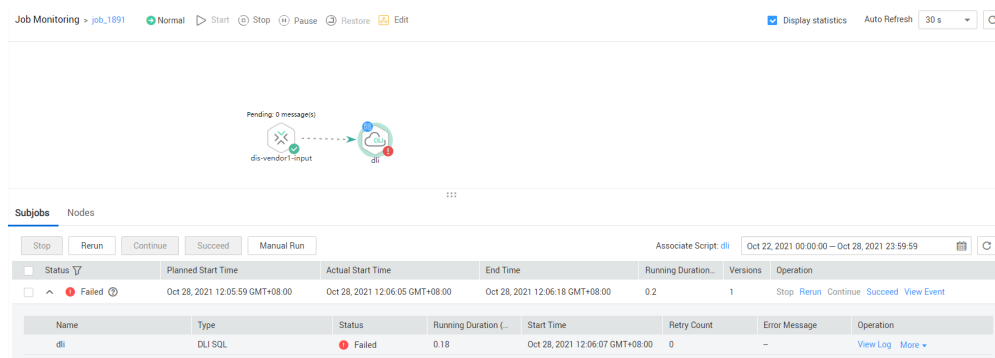


Table 6-40 describes the actions listed in the **Operation** column of each subjob.

Table 6-40 Subjob monitoring operations

Operation	Description
Stop	Stops a subjob instance that is in the Running state.
Rerun	Reruns a subjob instance that is in the Succeed or Failed state.
Continue	If a subjob instance is in the Abnormal state, you can click Continue to begin running the subsequent nodes in the subjob instance. NOTE This operation can be performed only when Failure Policy is set to Suspend the current job execution plan . To view the current failure policy, click a node and then click Advanced Settings on the Node Properties page.
Forcibly Succeed	Forcibly changes the status of a subjob instance from Failed to Succeed .
View Event	Displays the event content of a subjob.


- Click  in the **Status** column. The running records of the subjob node are displayed.

Table 6-41 describes the operations that can be performed on the node.

Table 6-41 Operations (node)

Operation	Description
View Log	View the log information of a node.

Operation	Description
More > Manual Retry	To run a node again after it fails, click Retry . NOTE This operation can be performed only when Failure Policy is set to Suspend the current job execution plan . To view the current failure policy, click a node and then click Advanced Settings on the Node Properties page.
More > Succeed	Change the status of a node from Failed to Succeed . NOTE This operation can be performed only when Failure Policy is set to Suspend the current job execution plan . To view the current failure policy, click a node and then click Advanced Settings on the Node Properties page.
More > Skip	To skip a node that is to be run or that has been paused, click Skip .
More > Pause	To pause a node that is to be run, click Pause . Nodes queued after the paused node will be blocked.
More > Resume	To resume a paused node, click Resume .

6.7.3 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**.
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 6-42](#).

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 6-42](#) describes the operations that can be performed on the instance.

Table 6-42 Instance monitoring operations

Operation	Description
Searching for a job based on the job name or creator	If you select Exact search , exact search by job name is supported. If you do not select Exact search , fuzzy search by job name is supported.
Filtering jobs by CDM job or node type	-
Stop	Stop an instance that is in the Waiting, Running, or Abnormal state.
Rerun	Rerun a subjob instance that is in the Succeed or Canceled state. For details, see Rerunning Job Instances .
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 Abnormal state, you can click Continue to begin running the subsequent nodes in the instance. NOTE This operation can be performed only when Failure Policy is set to Suspend the current job execution plan . To view the current failure policy, click a node and then click Advanced Settings on the Node Properties page.
More > Succeed	Forcibly change the status of an instance from Abnormal, Canceled, or Failed to Succeed .
More > View	Go to the job development page and view job information.


- Click  in front of an instance. The running records of all nodes in the instance are displayed.
- [Table 6-43](#) describes the actions that can be performed on the node.

Table 6-43 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 Retry . NOTE This operation can be performed only when Failure Policy is set to Suspend the current job execution plan . To view the current failure policy, click a node and then click Advanced Settings on the Node Properties page.

Operation	Description
More > Succeed	Change the status of a node from Failed to Succeed . NOTE This operation can be performed only when Failure Policy is set to Suspend the current job execution plan . To view the current failure policy, click a node and then click Advanced Settings on the Node Properties page.
More > Skip	To skip a node that is to be run or that has been paused, click Skip .
More > Pause	To pause a node that is to be run, click Pause . Nodes queued after the paused node will be blocked.
More > Resume	To resume a paused node, click Resume .

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**.
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 6-57 Setting the rerunning position

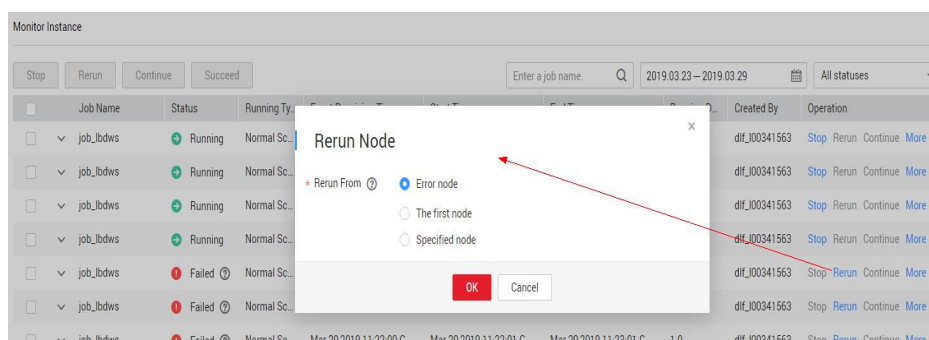


Table 6-44 Parameters for rerunning a job

Parameter	Description
Rerun Type	Type of the instance that you want to rerun. <ul style="list-style-type: none">• Rerun selected instance• Rerun instances of selected job and its upstream and downstream jobs
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">• Error node: When a job instance fails to be run, it reruns since the error node of the job instance.• The first node: When a job instance fails to be run, it reruns since the first node of the job instance.• Specified node: When a job instance fails to run, it reruns since the node specified in the job instance. This option is available only if Rerun Type is set to Rerun selected instance. <p>NOTE A job instance reruns from its first node if either of the following cases occurs:</p> <ul style="list-style-type: none">• The quantity or name of a node in the job changes.• The job instance has been successfully run.
Concurrent Instances	Number of job instances that can be concurrently processed.

6.7.4 Monitoring PatchData

In the navigation tree of the DataArts Factory console, choose **Monitoring > Monitor PatchData**.

On the , you can view the task status, service date, number of parallel periods, and PatchData job names, and stop a running task.

On the , 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.

6.7.5 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.

6.7.5.1 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**.
 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 6-45](#) describes the parameters.

Table 6-45 Notification parameters

Parameter	Mandatory	Description
Notification Scope	Yes	Notification scope. Available options include: <ul style="list-style-type: none">• One job: Notifications are sent for a single job.• All jobs: Notifications are sent for all jobs.
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">• When Notification Scope is One job, available options for this parameter include:<ul style="list-style-type: none">– Run abnormally/Fail: When a job cannot run normally or fail to run, a notification is sent to notify the user of the abnormality.– Run successfully: When a job runs successfully, a notification is sent to notify the user of the success.– Uncompleted: 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.– Busy resources: If resources are busy during job execution, a notification is sent.• When Notification Scope is All jobs, available options for this parameter include:<ul style="list-style-type: none">– Run abnormally/Fail: When a job cannot run normally or fail to run, a notification is sent to notify the user of the abnormality.– Busy resources: If resources are busy during job execution, a notification is sent. <p>NOTE For a real-time job, a notification is allowed to be sent only when the real-time job is in the Run abnormally or Failed state. For a batch job, a notification can be sent no matter when the batch job is in the Run normally, Run abnormally, or Failed state.</p>
Topic Name	Yes	<p>Select a notification topic.</p> <p>NOTE 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>

4. Click **OK**.

 **NOTE**

- The DataArts Factory module sends notifications through SMN. Using SMN may incur fees. For pricing details, contact the SMN support personnel.
- Multiple message topics can be configured for a job. When the job is successfully executed or fails to be executed, notifications can be sent to multiple subscribers.



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 6-45](#) 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**.

6.7.5.2 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**.
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 6-46](#) describes the notification parameters.

Figure 6-58 Create a notification

Table 6-46 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 Daily , Weekly , or Monthly . NOTE When Cycle is set to Daily , Weekly , or Monthly , 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"> If Cycle is set to Weekly, the value can be any day or any several days from Monday to Sunday in a week. If Cycle is set to Monthly, the value can be any day or any several days from 1st to 31st in a month.
Start Time	Yes	Point in time when the notification is sent. The value can be accurate to hour or minute.

Parameter	Mandatory	Description
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 OBS 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**.

 **NOTE**

DLF sends notifications through SMN. Using SMN may incur fees. For pricing details, contact the SMN support personnel.

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.

6.7.6 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**.
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.

Figure 6-59 Managing backup

Manage Backup

?

Date	Path	Progress	Status
2019-05-13	s3a://001zmwulanchabu3/22...	100%	finished
2019-05-12	s3a://aaaaa1111	100%	finished
2019-05-11	s3a://aaaaa1111	100%	finished
2019-05-10	s3a://aaaaa1111	100%	finished
2019-05-09	s3a://aaaaa1111	100%	finished

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**.

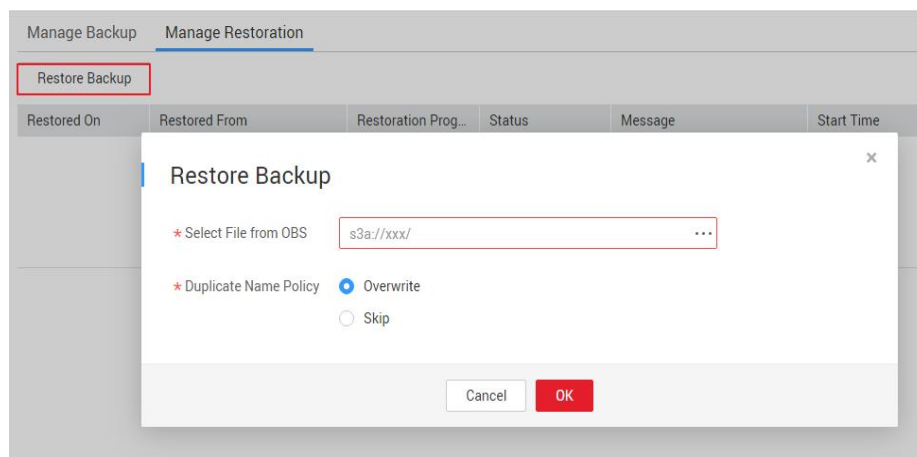
Step 2 In the navigation tree of the DataArts Factory console, choose **Manage Backup**.

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 6-60 Restoring assets

Step 4 Click **OK**.

----End

6.8 Configuration and Management

6.8.1 Configuring Resources

6.8.1.1 Configuring Environment Variables

This topic describes how to configure and use environment variables.

Application Scenario

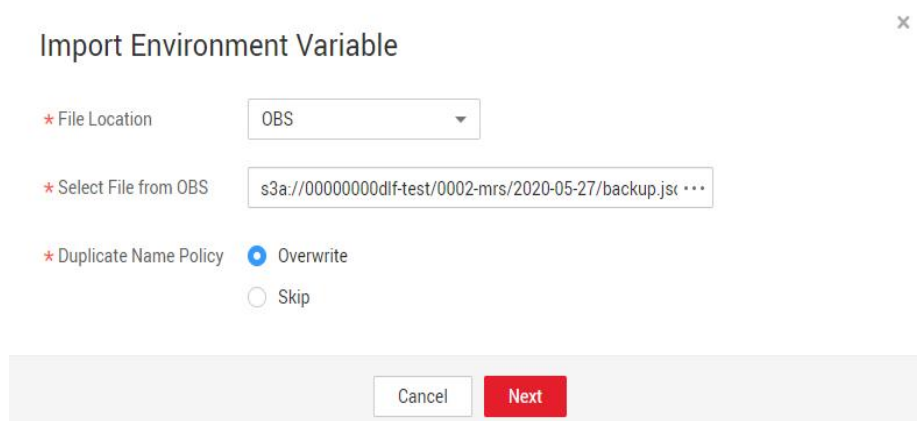
Configure job parameters. If a parameter belongs to multiple jobs, you can extract this parameter as an environment variable. Environment variables can be imported and exported.

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**.
- 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 6-61 Importing environment variables



----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**.
- 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 6-47](#) 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.

Figure 6-62 Configuring environment variables

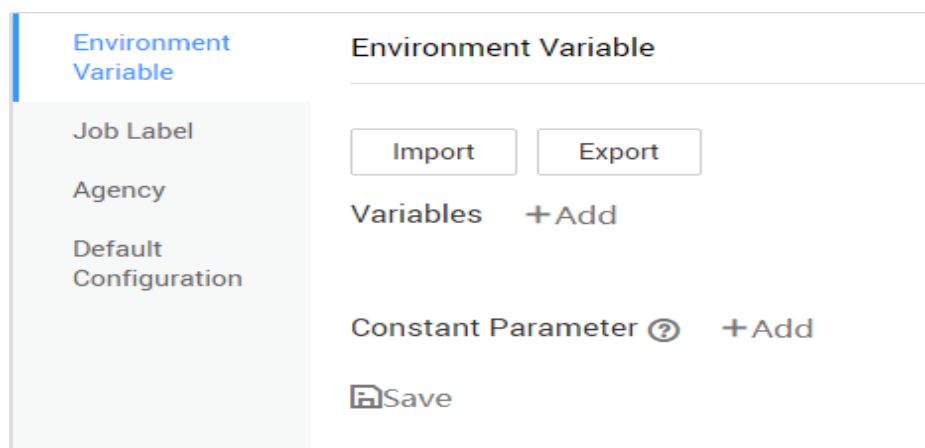




Table 6-47 Configuring environment variables

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, 123 and abc are supported. If the parameter value is a string, add double quotation marks (""), for example, " 05 ". For details about how to use EL expressions, see Expression Overview .

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")}`

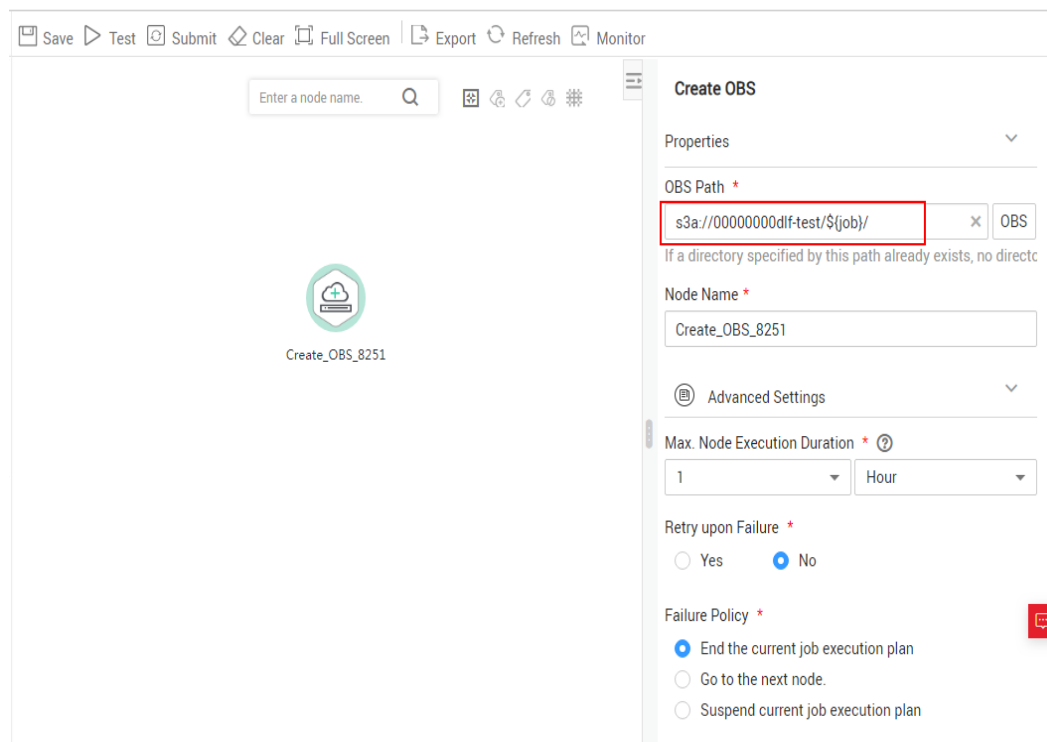
Example

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 6-63 Configuring parameters for the Create OBS node

Step 3 Click **Save** and then **Monitor** to monitor the running status of the job.

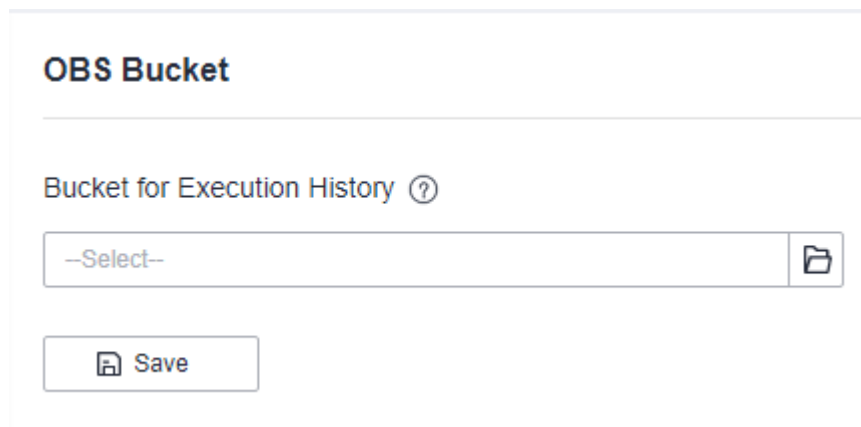
----End

6.8.1.2 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**.
- Step 2** In the navigation pane, choose **Configuration** > **Configure**.
- Step 3** Choose **OBS Bucket**.
- Step 4** Select an OBS bucket.

Figure 6-64 Configuring an OBS bucket

OBS Bucket

Bucket for Execution History ?

-Select- 

 Save

Step 5 Click **Save**.

----End

6.8.1.3 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**.

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

6.8.1.4 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 6-65 Creating an agency

Agencies / Create Agency

* Agency Name

* Agency Type Account
Delegate another HUAWEI CLOUD account to perform operations on your resources.
 Cloud service
Delegate a cloud service to access your resources in other cloud services.

* Delegated Account

* Validity Period

Description

0/255

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 6-66](#).
 - 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 6-66 Assigning permissions

Assign Permissions

Multiple policies can be selected. You can also modify or create policies.

View Selected (1) All policies/roles Tenant Administrator X Q C Policy View Project View

Policy/Role Name	Description	Project [Region]
<input checked="" type="checkbox"/> Tenant Administrator	Tenant Administrator (Exclude IAM)	All projec...

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 6-48](#).

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 6-49](#). 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 6-48 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 Tenant Guest and Server Administrator roles in the same project.

Node Name	System Permission	Description
<ul style="list-style-type: none"> MRS Presto SQL, MRS Spark, MRS Spark Python, MRS Flink Job, and MRS MapReduce MRS Spark SQL and MRS Hive SQL (connecting to MRS clusters through MRS APIs) 	MRS Administrator KMS Administrator	<p>Users assigned the MRS Administrator role can perform all operations on MRS. This role depends on the Tenant Guest and Server Administrator roles in the same project.</p> <p>Users assigned the KMS Administrator 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>DAYU Administrator has all permissions required for DataArts Studio.</p> <p>Users assigned the KMS Administrator 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>DAYU Administrator has all permissions required for DataArts Studio.</p> <p>Users assigned the KMS Administrator policy have the administrator permissions for encryption keys in DEW.</p>
CSS	DAYU Administrator Elasticsearch Administrator	<p>DAYU Administrator has all permissions required for DataArts Studio.</p> <p>Users assigned the Elasticsearch Administrator policy have all permissions for CSS. This role depends on the Tenant Guest and Server Administrator roles in the same project.</p>
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](#).

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 6-49 Creating a custom policy

Node Name	Action
Import GES	<ul style="list-style-type: none"> • ges:graph:access • ges:graph:getDetail • ges:jobs:getDetail
<ul style="list-style-type: none"> • MRS Presto SQL, MRS Spark, MRS Spark Python, MRS Flink Job, and MRS MapReduce • MRS Spark SQL and MRS Hive SQL (connecting to MRS clusters through MRS APIs) 	<ul style="list-style-type: none"> • mrs:job:delete • mrs:job:stop • mrs:job:submit • mrs:cluster:get • mrs:cluster:list • mrs:job:get • mrs:job:list • kms:dek:crypto • kms:cmk:get
MRS Spark SQL, MRS Hive SQL, MRS Kafka, and Kafka Client (connecting to the clusters in proxy mode)	<ul style="list-style-type: none"> • kms:dek:crypto • kms:cmk:get • DataArts Studio Administrator (role)

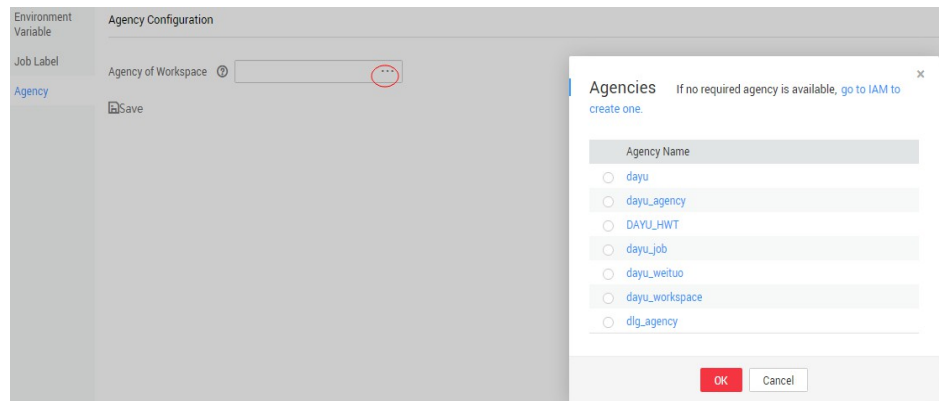
Node Name	Action
DLI Flink Job, DLI SQL, and DLI Spark	<ul style="list-style-type: none">• dli:jobs:get• dli:jobs:update• dli:jobs:create• dli:queue:submit_job• dli:jobs:list• dli:jobs:list_all
DWS SQL, RDS SQL (connecting to data sources in proxy mode), and Shell	<ul style="list-style-type: none">• kms:dek:crypto• kms:cmk:get• DataArts Studio Administrator (role)
Create OBS, Delete OBS, and OBS Manager	<ul style="list-style-type: none">• obs:bucket:GetBucketLocation• obs:bucket:ListBucketVersions• obs:object:GetObject• obs:bucket:CreateBucket• obs:bucket>DeleteBucket• obs:object>DeleteObject• obs:object:PutObject• obs:bucket:ListAllMyBuckets• obs:bucket:ListBucket

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**.
 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 6-67 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**.
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 6-68 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.

6.8.1.5 Configuring a Default Item

This section describes how to configure a default item. Currently, only DAYU Administrator and users with the DAYU Administrator role have the permission to perform operations on default configurations.

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

6.8.2 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**.
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 6-50](#) describes the directory parameters.

Table 6-50 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**.
2. In the left navigation pane, choose **Configuration > Manage Resource**.
3. Click **Create Resource**. In the displayed dialog box, configure resource parameters. [Table 6-51](#) describes the resource parameters. Click **OK**.

Table 6-51 Resource management parameters

Parameter	Man dato ry	Description
Name	Yes	Name of the resource. The name must contain 1 to 32, including only letters, numbers, underscores (_), and hyphens (-).
Type	Yes	File type of the resource. Possible values: <ul style="list-style-type: none">• jar: JAR file• pyFile: User Python file• file: User file• archive: User AI model file
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">• If Resource Location is OBS, select the main JAR package that has been uploaded to OBS.• If Resource Location is HDFS, select the main JAR package that has been uploaded to HDFS.
Depended JAR Package	No	Depended JAR package that has been uploaded to OBS. This parameter is required when Type is set to jar and Resource Location is set to OBS or HDFS .
Select Resource	Yes	Specific resource file.
Storage Path	Yes	Path to a directory where the resource is stored. This parameter is required only when Resource Location is set to Local .
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**.
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 6-51](#).
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**.
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**.
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**.
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**.
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.

6.9 Node Reference

6.9.1 Node Overview

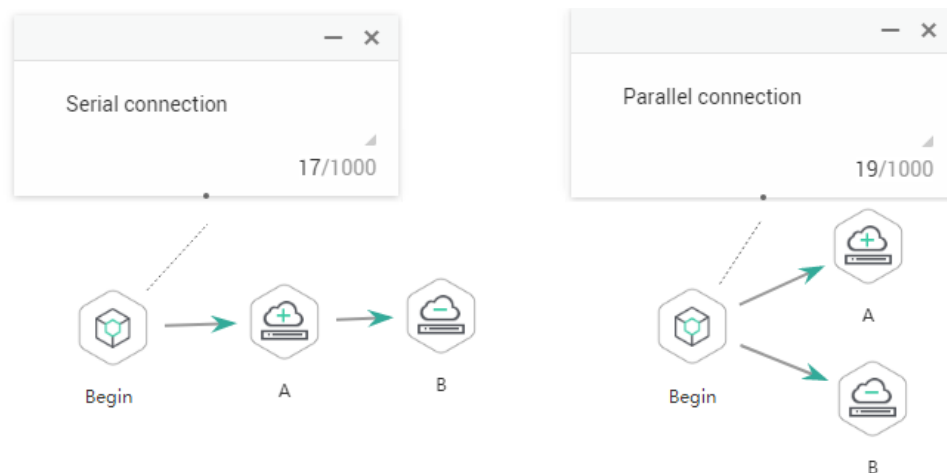
A node defines the operations performed on data. DataArts Factory provides nodes used for data integration, computing and analysis, database operations, and resource management. You can choose your desired nodes

- 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 6-69 Connection diagram



6.9.2 Node Lineages

6.9.2.1 Overview

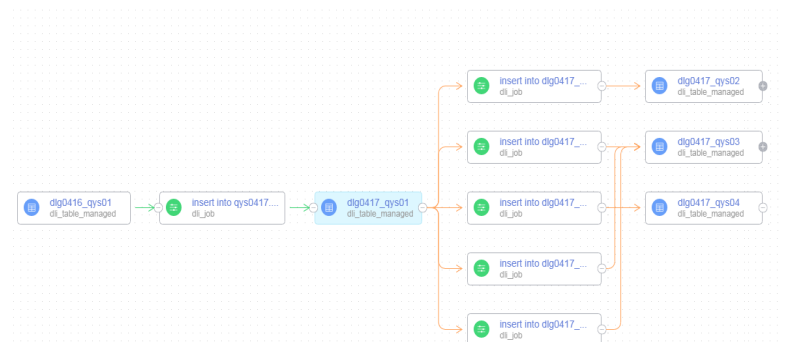
What Is Data Lineage?

In the era of big data, various types of data are rapidly generated due to explosive data growth. The massive and complex data information is converged, transformed, and transferred to generate new data and aggregate into an ocean of data.

During this process, a relationship is formed between the data, and these relationships are their lineages. They are analogous to the genetic relationships between people. However, in contrast from our human lineages, data lineages have the following distinct features:

- **Belongingness:** Specific data belongs to a specific organization or individual.
- **Multi-source:** One piece of data can have multiple sources. One piece of data may be generated by processing multiple pieces of data, and there may be multiple such processes.
- **Traceability:** The data lineage is traceable. It reflects the data lifecycle and the entire process from data generation to data disappearance.
- **Hierarchy:** The data lineage is hierarchical. Data classification and summary form new data, and different levels of description result in data layers.

Figure 6-70 Data lineage example



How DataArts Studio Data Lineage Is Implemented

- **Generation of data lineages:**
On the DataArts Studio platform, data lineages are generated by configuring data processing and migration nodes in the DataArts Factory module. Currently, the system collects the lineages generated by static node configuration and the lineages on some node instances. For details, see [Automatic Lineage Analysis](#).
In addition, DataArts Studio allows you to manually configure lineages. If you do so, automatic lineage analysis does not take effect. For details, see [Manually Configuring a Lineage](#).
- **Display of data lineages:**
If you have configured data lineages and started job scheduling in the DataArts Factory module, you can start a metadata collection task in the DataArts Catalog module to view the data lineages.

6.9.2.2 Configuring Data Lineages

On the DataArts Studio platform, data lineages are generated by configuring data processing and migration nodes in the DataArts Factory module. Currently, the system collects the lineages generated by static node configuration and the lineages on some node instances. For details, see [Automatic Lineage Analysis](#).

In addition, DataArts Studio allows you to manually configure lineages. If you do so, automatic lineage analysis does not take effect. For details, see [Manually Configuring a Lineage](#).

Automatic Lineage Analysis

Data lineages can be parsed automatically if the job contains the following nodes:

- **SQL nodes**

DataArts Studio supports lineage parsing of DLI SQL, DWS SQL and MRS Hive SQL nodes. It supports multi-SQL parsing and column-level lineage parsing.

- **DLI SQL**

- Lineages generated by data insertion between DLI tables
- Lineages between OBS files generated by table creation statements and DLI tables

- **DWS SQL**

- Lineages between DWS tables generated by DDL operations such as "Create table like/as"
- Lineages between DWS tables generated by DML operations such as "Insert into"

- **MRS Hive SQL**

- Lineages between MRS tables generated by DDL operations such as "Create table like/as"
- Lineages between MRS tables generated by DML operations such as "Insert into/overwrite"

- **Data integration nodes**

Lineages of the CDM Job, ETL Job, and OBS Manager nodes can be parsed.

- **CDM Job**

Lineages generated during table file migration between MRS Hive, DLI, RDS, CSS, DWS, and OBS

- **ETL Job**

Data lineages generated by ETL tasks between DLI, OBS, MySQL, and DWS.

- **OBS Manager**

Lineages generated by directory or file replication and migration between OBS buckets

NOTE

A single SQL statement cannot contain semicolons (;).

Manually Configuring a Lineage

In DataArts Studio DataArts Factory, you can define the input and output lineage relationships of nodes. When you manually configure a lineage, automatic lineage analysis does not take effect. Manual lineage configuration does not affect job running.

Currently, DLI, DWS, Hive, CSS, OBS, and CUSTOM are supported as the input and output data sources during manual lineage configuration. CUSTOM indicates a custom type. When manually configuring a lineage, you can add data sources that are not supported as custom types.

The following nodes support manual lineage configuration:

- [CDM Job](#)
- [Rest Client](#)
- [DLI SQL](#)
- [DLI Spark](#)
- [DWS SQL](#)
- [MRS Spark SQL](#)
- [MRS Hive SQL](#)
- [MRS Presto SQL](#)
- [MRS Spark](#)
- [MRS Spark Python](#)
- [ETL Job](#)
- [OBS Manager](#)

6.9.2.3 Viewing Data Lineages


If you have configured data lineages and started job scheduling in the DataArts Factory module, you can start a metadata collection task in the DataArts Catalog module to view the data lineages.

Prerequisites

Data lineages have been automatically or manually configured. For details, see [Configuring Data Lineages](#).

Starting Job Scheduling

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**.

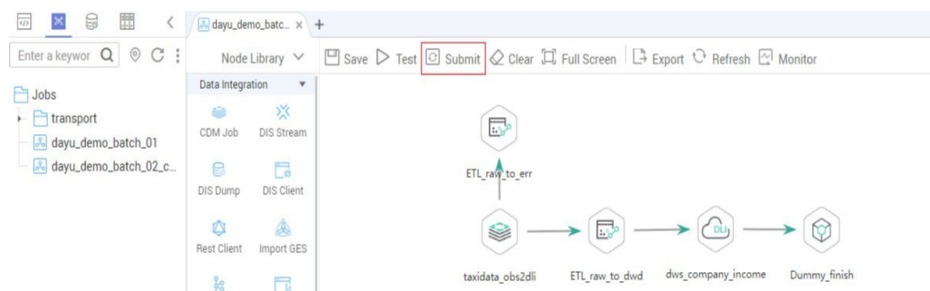
Step 2 In the navigation pane, click  and double-click the job for which lineages have been configured to open it.

Step 3 Click **Execute**. The system starts parsing lineages of the job.

NOTE

If you click **Test**, the system will not parse lineages of the job.

Figure 6-71 Starting job scheduling



----End

Creating a Metadata Collection Task

If a metadata collection task has been created, skip this part.

- Step 1** On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Catalog**.
- Step 2** Create a metadata collection task by following the instructions in [Task Management](#).

----End

Viewing Data Lineages

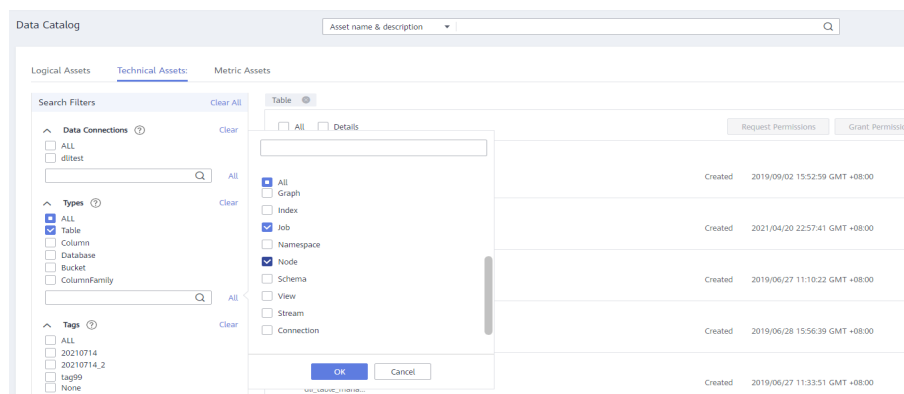
- Step 1** On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Catalog**.
- Step 2** In the navigation pane, choose **Data Catalog**. In the right pane, click the **Technical Assets** tab. On this page, you can query jobs, nodes, and tables.

In the **Types** area, click **All**, select **Job**, **Node**, and **Table**, and click **OK**.

NOTE

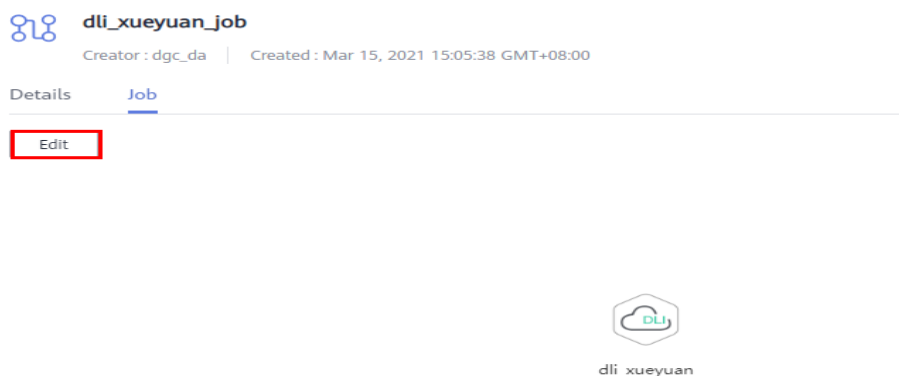
Jobs do not belong to any data connection. If you select a data connection in the search filters, no result will be returned.

Figure 6-72 Selecting types



Step 3 In the search result, click the name of an asset ending with **_job** to view its details. On the job details page, click the **Job** tab and then **Edit** to go to the job editing page.

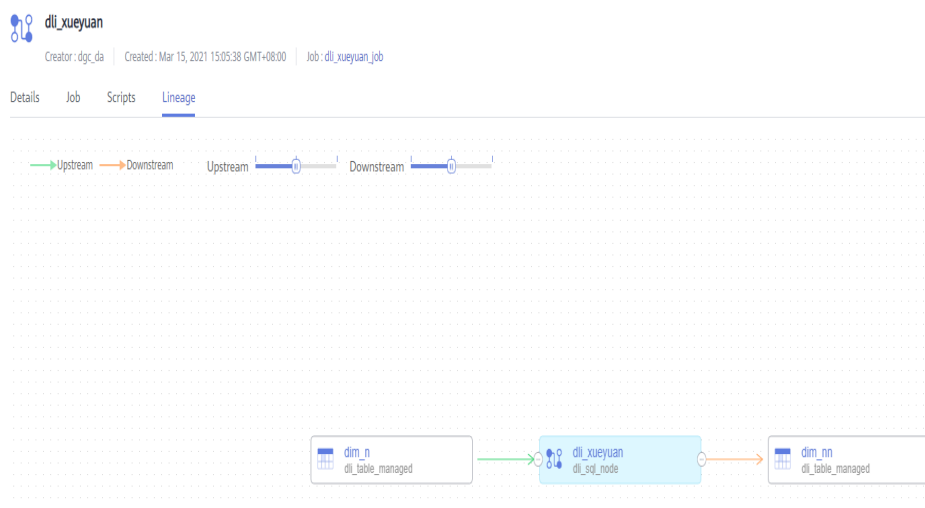
Figure 6-73 Viewing job details



Step 4 In the data asset search result, click the name of an asset ending with **_node** to view its details. On the node details page, you can view the node lineage information.

- Click the + or - icon beside the node to expand its upstream and downstream links.
- Click a node to view the its details.
- Click the **Job** tab and then **Edit** to go to the job editing page.

Figure 6-74 Viewing lineages of a node



Step 5 In the data asset search result, click the name of an asset whose icon is a table to view its details. On the table details page, you can view lineages of the table.

- Click the + or - icon beside the table to expand its upstream and downstream links.
- Click a table to view the its details.

Figure 6-75 Viewing lineages of a table

----End

6.9.3 CDM Job

Functions

The CDM Job node is used to run a predefined CDM job for data migration.

NOTE

If you have configured a macro variable of date and time in a CDM job and schedule the CDM job through DataArts Factory, the system replaces the macro variable of date and time with *(Planned start time of the data development job – Offset)* rather than *(Actual start time of the CDM job – Offset)*.

Parameters

[Table 6-52](#), [Table 6-53](#), and [Table 6-54](#) 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 6-52 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">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.If you select two clusters, you are advised to set Job Type to Existing jobs rather than New jobs 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 Exporting and Importing CDM Jobs in Batches.
Job Type	Yes	<ul style="list-style-type: none">Existing jobsNew jobs <p>NOTE</p> <ul style="list-style-type: none">If Job Type is Existing jobs, 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.If Job Type is New jobs, the system checks whether a CDM job with the same name is running.<ul style="list-style-type: none">If the CDM job is not running, update the job with the same name based on the request body.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.
CDM Job Name	No	<p>This parameter is required only when Job Type is set to Existing jobs. Name of the CDM job to be executed.</p> <p>If the CDM job uses the job parameters or environment variables 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 Job Type is set to New jobs . Enter the JSON message body of the CDM job. For convenience, you can choose More > View Job JSON in the Operation column of an existing CDM job, copy the JSON content, and modify the content here. If the CDM job uses the job parameters or environment variables 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 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 6-53 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"> • Yes: The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> - Maximum Retries - Retry Interval (seconds) • No: The node will not be re-executed. This is the default setting. <p>NOTE</p> <ul style="list-style-type: none"> • If Max. Node Execution Duration 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. • If parameter transfer is used for scheduling the CDM job, do not configure parameter Retry upon Failure in the CDM job.




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"> • End the current job execution plan: stops running the current job. The job instance status is Failed. • Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored. • Suspend current job execution plan: suspends running the current job. The job instance status is Waiting. • Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.
Dry run	No	If you select this option, the node will not be executed, and a success message will be returned.

Table 6-54 Lineage

Parameter	Description
Input	

Parameter	Description
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">• DWS<ul style="list-style-type: none">– Connection Name: Click In the displayed dialog box, select a DWS data connection.– Database: Click In the displayed dialog box, select a DWS database.– Schema: Click In the displayed dialog box, select a DWS schema.– Table Name: Click In the displayed dialog box, select a DWS table.• OBS<ul style="list-style-type: none">– Path: Click In the displayed dialog box, select an OBS path.• CSS<ul style="list-style-type: none">– Cluster Name: Click In the displayed dialog box, select a CSS cluster.– Index: Enter a CSS index name.• HIVE<ul style="list-style-type: none">– Connection Name: Click In the displayed dialog box, select a HIVE data connection.– Database: Click In the displayed dialog box, select a HIVE database.– Table Name: Click In the displayed dialog box, select a HIVE table.• CUSTOM<ul style="list-style-type: none">– Name: Enter a name of the CUSTOM type.– Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.• DLI<ul style="list-style-type: none">– Connection Name: Click In the displayed dialog box, select a DLI data connection.– Database: Click In the displayed dialog box, select a DLI database.– Table Name: Click In the displayed dialog box, select a DLI table.
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.
Output	
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">• DWS<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DWS data connection.- Database: Click In the displayed dialog box, select a DWS database.- Schema: Click In the displayed dialog box, select a DWS schema.- Table Name: Click In the displayed dialog box, select a DWS table.• OBS<ul style="list-style-type: none">- Path: Click In the displayed dialog box, select an OBS path.• CSS<ul style="list-style-type: none">- Cluster Name: Click In the displayed dialog box, select a CSS cluster.- Index: Enter a CSS index name.• HIVE<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a HIVE data connection.- Database: Click In the displayed dialog box, select a HIVE database.- Table Name: Click In the displayed dialog box, select a HIVE table.• CUSTOM<ul style="list-style-type: none">- Name: Enter a name of the CUSTOM type.- Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.• DLI<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DLI data connection.- Database: Click In the displayed dialog box, select a DLI database.- Table Name: Click In the displayed dialog box, select a DLI table.

Parameter	Description
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.

6.9.4 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 6-55](#), [Table 6-56](#), and [Table 6-57](#) describe the parameters of the Rest Client node.

Table 6-55 Parameters of Rest Client nodes

Parameter	Mandator 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"> • GET • POST • PUT • DELETE
Request Header	No	Click + to add a request header. The parameters are described as follows: <ul style="list-style-type: none"> • Parameter Name Name of a parameter. The options are Content-Type and Accept-Language. • Parameter Value Value of the parameter
URL Parameter	No	Enter a URL parameter. The value is a character string in key=value format. Character strings are separated by newlines. This parameter is available only when HTTP Method is set to GET . Set these parameters as follows: <ul style="list-style-type: none"> • Parameter The parameter contains a maximum of 32 characters, including only letters, numbers, hyphens (-), and underscores (_). • Value The value contains a maximum of 64 characters, including only letters, numbers, hyphens (-), underscores (_), dollar signs (\$), open braces ({), and close braces (}).
Request Body	Yes	The request body is in JSON format. This parameter is available only when HTTP Method is set to POST or PUT .
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 HTTP Method is set to GET . Possible values: <ul style="list-style-type: none"> • YES: Check whether the return value is the same as the expected one. • NO: 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).

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 Check Returned Value is set to YES.</p> <p>For example, the returned result is as follows:</p> <pre>{ "param1": "aaaa", "inner": { "inner": { "param4": 2014247437 }, "param3": "cccc" }, "status": 200, "param2": "bbbb" }</pre> <p>The param4 path is inner.inner.param4.</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 Check Returned Value is set to YES.</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 Check Returned Value is set to YES.</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 Failed. This parameter is available only when Check Returned Value is set to YES.</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 var4 is a job variable. The job variable must contain 1 to 64 characters, including only letters and numbers. inner.inner.param4 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 \${var4}


Table 6-56 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"> • Yes: The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> - Maximum Retries - Retry Interval (seconds) • No: The node task will not be re-executed. This is the default setting. <p>NOTE If Timeout Interval 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">• End the current job execution plan: stops running the current job. The job instance status is Failed.• Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored.• Suspend current job execution plan: suspends running the current job. The job instance status is Waiting.• Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

Table 6-57 Lineage

Parameter	Description
Input	

Parameter	Description
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">• DWS<ul style="list-style-type: none">– Connection Name: Click In the displayed dialog box, select a DWS data connection.– Database: Click In the displayed dialog box, select a DWS database.– Schema: Click In the displayed dialog box, select a DWS schema.– Table Name: Click In the displayed dialog box, select a DWS table.• OBS<ul style="list-style-type: none">– Path: Click In the displayed dialog box, select an OBS path.• CSS<ul style="list-style-type: none">– Cluster Name: Click In the displayed dialog box, select a CSS cluster.– Index: Enter a CSS index name.• HIVE<ul style="list-style-type: none">– Connection Name: Click In the displayed dialog box, select a HIVE data connection.– Database: Click In the displayed dialog box, select a HIVE database.– Table Name: Click In the displayed dialog box, select a HIVE table.• CUSTOM<ul style="list-style-type: none">– Name: Enter a name of the CUSTOM type.– Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.• DLI<ul style="list-style-type: none">– Connection Name: Click In the displayed dialog box, select a DLI data connection.– Database: Click In the displayed dialog box, select a DLI database.– Table Name: Click In the displayed dialog box, select a DLI table.
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.
Output	
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">● DWS<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DWS data connection.- Database: Click In the displayed dialog box, select a DWS database.- Schema: Click In the displayed dialog box, select a DWS schema.- Table Name: Click In the displayed dialog box, select a DWS table.● OBS<ul style="list-style-type: none">- Path: Click In the displayed dialog box, select an OBS path.● CSS<ul style="list-style-type: none">- Cluster Name: Click In the displayed dialog box, select a CSS cluster.- Index: Enter a CSS index name.● HIVE<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a HIVE data connection.- Database: Click In the displayed dialog box, select a HIVE database.- Table Name: Click In the displayed dialog box, select a HIVE table.● CUSTOM<ul style="list-style-type: none">- Name: Enter a name of the CUSTOM type.- Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.● DLI<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DLI data connection.- Database: Click In the displayed dialog box, select a DLI database.- Table Name: Click In the displayed dialog box, select a DLI table.

Parameter	Description
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.

6.9.5 Import GES

Function

The Import GES node is used to import files from an OBS bucket to a GES graph.

Parameters

[Table 6-58](#) and [Table 6-59](#) describe the parameters of the Import GES node.

Table 6-58 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"> • Allow repetitive edges • Ignore subsequent repetitive edges • Overwrite previous repetitive edges
Offline	No	Whether offline import is used. The value is Yes or No , and the default value is No . <ul style="list-style-type: none"> • true: Offline import is selected. The import speed is high, but the graph is locked and cannot be read or written during the import. • false: Online import is selected. Online import is slower than offline import. However, during online import, the graph can be read (but cannot be written).
Ignore Labels on Repetitive Edges	No	Indicates whether to ignore labels on repetitive edges. The value is Yes or No , and the default value is Yes . <ul style="list-style-type: none"> • Yes: Indicates that the repetitive edge definition does not contain the label. That is, the <source vertex, target vertex> indicates an edge, excluding the label information. • No: Indicates that the repetitive edge definition contains the label. That is, the <source vertex, target vertex, label> indicates an edge.
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 6-59 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">• Yes: The node task will be re-executed, and the following parameters must be configured:<ul style="list-style-type: none">– Maximum Retries– Retry Interval (seconds)• No: The node task will not be re-executed. This is the default setting. <p>NOTE If Timeout Interval 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">• End the current job execution plan: stops running the current job. The job instance status is Failed.• Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored.• Suspend current job execution plan: suspends running the current job. The job instance status is Waiting.• Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

6.9.6 MRS Kafka

Functions

The MRS Kafka node is used to query the number of messages that are not consumed by a topic.

Parameters

[Table 6-60](#) and [Table 6-61](#) describe the parameters of the MRS Kafka node.

Table 6-60 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 6-61 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">• Yes: The node task will be re-executed, and the following parameters must be configured:<ul style="list-style-type: none">- Maximum Retries- Retry Interval (seconds)• No: The node task will not be re-executed. This is the default setting. NOTE If Timeout Interval 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.

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">• End the current job execution plan: stops running the current job. The job instance status is Failed.• Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored.• Suspend current job execution plan: suspends running the current job. The job instance status is Waiting.• Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

6.9.7 Kafka Client

Functions

The Kafka Client node is used to send data to Kafka topics.

Parameters

[Table 6-62](#) describes the parameters of the Kafka Client node.

Table 6-62 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 6-63 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"> • Yes: The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> - Maximum Retries - Retry Interval (seconds) • No: The node task will not be re-executed. This is the default setting. <p>NOTE If Timeout Interval 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"> • End the current job execution plan: stops running the current job. The job instance status is Failed. • Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored. • Suspend current job execution plan: suspends running the current job. The job instance status is Waiting. • Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

6.9.8 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 6-64 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 6-65 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">• Yes: The node task will be re-executed, and the following parameters must be configured:<ul style="list-style-type: none">– Maximum Retries– Retry Interval (seconds)• No: The node task will not be re-executed. This is the default setting. NOTE If Timeout Interval 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.

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">• End the current job execution plan: stops running the current job. The job instance status is Failed.• Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored.• Suspend current job execution plan: suspends running the current job. The job instance status is Waiting.• Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

6.9.9 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 6-66](#).
 - **Flink SQL job:** For details, see [Table 6-67](#).
 - **User-defined Flink job:** For details, see [Table 6-68](#).
- [Table 6-69](#)

Table 6-66 Parameter parameters of an existing Flink job

Parameter	Mandatory	Description
Job Type	Yes	Select Existing Flink job .
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 6-67 Property parameters of a Flink SQL job

Parameter	Mandatory	Description
Job Type	Yes	Select Flink SQL job . 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 Creating a Script and Developing an SQL Script .
DLI Queue	Yes	Shared queues are selected by default. You can also select a dedicated custom queue. NOTE 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. NOTE The value of Concurrency 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 Queue . Before selecting a UDF JAR resource package, upload the UDF JAR package to the OBS bucket and create resources on the Manage Resource page. For details, see Creating a Resource . 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 6-68 Property parameters of a user-defined Flink job

Parameter	Mandatory	Description
Job Type	Yes	Select User-defined Flink job .
JAR Package Path	Yes	User-defined package. Before selecting a package, upload the JAR package to the OBS bucket and create resources on the Manage Resource page. For details, see Creating a Resource .
Main Class	Yes	Name of the JAR package to be loaded, for example, KafkaMessageStreaming . <ul style="list-style-type: none"> Default: Specified based on the Manifest file in the JAR package. Manually assign: Enter the class name and confirm the class arguments (separate arguments with spaces). <p>NOTE When a class belongs to a package, the package path must be carried, for example, packagePath.KafkaMessageStreaming.</p>
Main Class Parameter	Yes	List of parameters of a specified class. The parameters are separated by spaces.
DLI Queue	Yes	Shared queues are selected by default. You can also select a dedicated custom queue. <p>NOTE 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	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. For details about custom images, see Overview of Custom Images .
CUs	Yes	Compute Unit (CU) is the pricing unit for DLI. 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 1 .
Concurrency	Yes	The number of Flink SQL jobs that run at the same time. NOTE The value of Concurrency must not exceed the value obtained through the following formula: $4 \times (\text{Number of CUs} - 1)$.
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 6-69 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"> • Yes: The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> - Maximum Retries - Retry Interval (seconds) • No: The node task will not be re-executed. This is the default setting. <p>NOTE If Timeout Interval 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"> • End the current job execution plan: stops running the current job. The job instance status is Failed. • Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored. • Suspend current job execution plan: suspends running the current job. The job instance status is Waiting. • Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

6.9.10 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 6-70](#), [Table 6-71](#), and [Table 6-72](#) describe the parameters of the DLI SQLnode node.

Table 6-70 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">SQL Statement In the SQL statement text box, enter the SQL statement to be executed.SQL Script Select a script to be executed. If the script is not created, create and develop the script by repeating steps Creating a Script and Developing an SQL Script. <p>NOTE 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">The environment variable must start with dli.sql. or spark.sql.If the key of the environment variable is dli.sql.shuffle.partitions or dli.sql.autoBroadcastJoinThreshold, the environment variable cannot contain the greater than (>) or less than (<) sign.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.



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"> Click . On the Queue Management page of DLI, create a resource queue. Go to the DLI console to create a resource queue.
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 an EL expression . 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"> If you select <input type="radio"/>, dirty data will be recorded. If you do not select <input type="radio"/>, dirty data will not be recorded.


Table 6-71 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"> • Yes: The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> – Maximum Retries – Retry Interval (seconds) • No: The node task will not be re-executed. This is the default setting. <p>NOTE If Timeout Interval 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"> • End the current job execution plan: stops running the current job. The job instance status is Failed. • Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored. • Suspend current job execution plan: suspends running the current job. The job instance status is Waiting. • Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

Table 6-72 Lineage

Parameter	Description
Input	

Parameter	Description
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">• DWS<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DWS data connection.- Database: Click In the displayed dialog box, select a DWS database.- Schema: Click In the displayed dialog box, select a DWS schema.- Table Name: Click In the displayed dialog box, select a DWS table.• OBS<ul style="list-style-type: none">- Path: Click In the displayed dialog box, select an OBS path.• CSS<ul style="list-style-type: none">- Cluster Name: Click In the displayed dialog box, select a CSS cluster.- Index: Enter a CSS index name.• HIVE<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a HIVE data connection.- Database: Click In the displayed dialog box, select a HIVE database.- Table Name: Click In the displayed dialog box, select a HIVE table.• CUSTOM<ul style="list-style-type: none">- Name: Enter a name of the CUSTOM type.- Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.• DLI<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DLI data connection.- Database: Click In the displayed dialog box, select a DLI database.- Table Name: Click In the displayed dialog box, select a DLI table.
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.
Output	
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">• DWS<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DWS data connection.- Database: Click In the displayed dialog box, select a DWS database.- Schema: Click In the displayed dialog box, select a DWS schema.- Table Name: Click In the displayed dialog box, select a DWS table.• OBS<ul style="list-style-type: none">- Path: Click In the displayed dialog box, select an OBS path.• CSS<ul style="list-style-type: none">- Cluster Name: Click In the displayed dialog box, select a CSS cluster.- Index: Enter a CSS index name.• HIVE<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a HIVE data connection.- Database: Click In the displayed dialog box, select a HIVE database.- Table Name: Click In the displayed dialog box, select a HIVE table.• CUSTOM<ul style="list-style-type: none">- Name: Enter a name of the CUSTOM type.- Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.• DLI<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DLI data connection.- Database: Click In the displayed dialog box, select a DLI database.- Table Name: Click In the displayed dialog box, select a DLI table.

Parameter	Description
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.

6.9.11 DLI Spark

Functions

The DLI Spark node is used to execute a predefined Spark job.

Parameters

[Table 6-73](#), [Table 6-74](#), and [Table 6-75](#) describe the parameters of the DLI Sparknode node.

Table 6-73 Parameters of DLI Spark 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 (>).
DLI Queue	Yes	Select a queue from the drop-down list box.

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>
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">• 8-core, 32 GB memory• 16-core, 64 GB memory• 32-core, 128 GB memory
Major Job Class	Yes	Name of the major class of the Spark job. When the application type is .jar , 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: obs://Bucket name/Folder name/Package name . Before selecting a resource package, upload the JAR package and its dependency packages to the OBS bucket and create resources on the Manage Resource page. For details, see Creating a Resource .
Resource Type	Yes	Select OBS path or DLI program package . <ul style="list-style-type: none">• OBS path: 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.• DLI package: The resource package file will not be uploaded to the DLI resource management system before the job is executed.

Parameter	Mandatory	Description
Group	No	This parameter is mandatory when Resource Type is set to DLI program package . You can select Use existing , Create new , or Do not use .
Group Name	No	This parameter is mandatory when Resource Type is set to DLI program package . <ul style="list-style-type: none"> • Use existing: Select an existing group. • Create new: Enter a user-defined group name. • Do not use: Do not select or enter a group name.
Major-Class Entry Parameters	No	User-defined parameters. Separate multiple parameters by Enter . These parameters can be replaced by global variables. For example, if you create a global variable batch_num on the Global Configuration > Global Variables page, you can use {{batch_num}} to replace a parameter with this variable after the job is submitted.
Spark Job Running Parameters	No	Enter a parameter in the format of key/value . Press Enter to separate multiple key-value pairs. For details about the parameters, see Spark Configuration . These parameters can be replaced by global variables. For example, if you create a global variable custom_class on the Global Configuration > Global Variables page, you can use "spark.sql.catalog"={{custom_class}} to replace a parameter with this variable after the job is submitted. NOTE The JVM garbage collection algorithm cannot be customized for Spark jobs.

Parameter	Mandatory	Description
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"> • CloudTable/MRS HBase: sys.datasource.hbase • DDS: sys.datasource.mongo • CloudTable/MRS OpenTSDB: sys.datasource.opentsdb • DWS: sys.datasource.dws • RDS MySQL: sys.datasource.rds • RDS PostGre: sys.datasource.rds • DCS: sys.datasource.redis • CSS: sys.datasource.css <p>DLI internal modules include:</p> <ul style="list-style-type: none"> • sys.res.dli-v2 • sys.res.dli • sys.datasource.dli-inner-table
Metadata Access	Yes	Whether to access metadata through Spark jobs. For details, see Using the Spark Job to Access DLI Metadata .


Table 6-74 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"> • Yes: The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> – Maximum Retries – Retry Interval (seconds) • No: The node task will not be re-executed. This is the default setting. <p>NOTE If Timeout Interval 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"> • End the current job execution plan: stops running the current job. The job instance status is Failed. • Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored. • Suspend current job execution plan: suspends running the current job. The job instance status is Waiting. • Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

Table 6-75 Lineage

Parameter	Description
Input	

Parameter	Description
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">● DWS<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DWS data connection.- Database: Click In the displayed dialog box, select a DWS database.- Schema: Click In the displayed dialog box, select a DWS schema.- Table Name: Click In the displayed dialog box, select a DWS table.● OBS<ul style="list-style-type: none">- Path: Click In the displayed dialog box, select an OBS path.● CSS<ul style="list-style-type: none">- Cluster Name: Click In the displayed dialog box, select a CSS cluster.- Index: Enter a CSS index name.● HIVE<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a HIVE data connection.- Database: Click In the displayed dialog box, select a HIVE database.- Table Name: Click In the displayed dialog box, select a HIVE table.● CUSTOM<ul style="list-style-type: none">- Name: Enter a name of the CUSTOM type.- Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.● DLI<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DLI data connection.- Database: Click In the displayed dialog box, select a DLI database.- Table Name: Click In the displayed dialog box, select a DLI table.
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.
Output	
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">• DWS<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DWS data connection.- Database: Click In the displayed dialog box, select a DWS database.- Schema: Click In the displayed dialog box, select a DWS schema.- Table Name: Click In the displayed dialog box, select a DWS table.• OBS<ul style="list-style-type: none">- Path: Click In the displayed dialog box, select an OBS path.• CSS<ul style="list-style-type: none">- Cluster Name: Click In the displayed dialog box, select a CSS cluster.- Index: Enter a CSS index name.• HIVE<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a HIVE data connection.- Database: Click In the displayed dialog box, select a HIVE database.- Table Name: Click In the displayed dialog box, select a HIVE table.• CUSTOM<ul style="list-style-type: none">- Name: Enter a name of the CUSTOM type.- Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.• DLI<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DLI data connection.- Database: Click In the displayed dialog box, select a DLI database.- Table Name: Click In the displayed dialog box, select a DLI table.

Parameter	Description
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.

6.9.12 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 Script and 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 6-76](#), [Table 6-77](#), and [Table 6-78](#) describe the parameters of the DWS SQLnode node.

Table 6-76 Parameters of DWS SQL nodes





Parameter	Mandatory	Description
SQL or Script	Yes	<p>You can select SQL statement or SQL script.</p> <ul style="list-style-type: none"> SQL Statement In the SQL statement text box, enter the SQL statement to be executed. SQL Script Select a script to be executed. If the script is not created, create and develop the script by repeating steps Creating a Script and Developing an SQL Script. <p>NOTE 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.
Database	Yes	Database that is configured in the SQL script. The value can be changed.
Script Parameter	No	<p>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 an EL expression.</p> <p>If the parameters of the associated SQL script are changed, click  to refresh the parameters.</p>
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	<p>Name of the SQL script. The value can be changed. The rules are as follows:</p> <p>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 (>).</p>


Table 6-77 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"> • Yes: The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> - Maximum Retries - Retry Interval (seconds) • No: The node task will not be re-executed. This is the default setting. <p>NOTE If Timeout Interval 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"> • End the current job execution plan: stops running the current job. The job instance status is Failed. • Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored. • Suspend current job execution plan: suspends running the current job. The job instance status is Waiting. • Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

Table 6-78 Lineage

Parameter	Description
Input	
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">● DWS<ul style="list-style-type: none">– Connection Name: Click In the displayed dialog box, select a DWS data connection.– Database: Click In the displayed dialog box, select a DWS database.– Schema: Click In the displayed dialog box, select a DWS schema.– Table Name: Click In the displayed dialog box, select a DWS table.● OBS<ul style="list-style-type: none">– Path: Click In the displayed dialog box, select an OBS path.● CSS<ul style="list-style-type: none">– Cluster Name: Click In the displayed dialog box, select a CSS cluster.– Index: Enter a CSS index name.● HIVE<ul style="list-style-type: none">– Connection Name: Click In the displayed dialog box, select a HIVE data connection.– Database: Click In the displayed dialog box, select a HIVE database.– Table Name: Click In the displayed dialog box, select a HIVE table.● CUSTOM<ul style="list-style-type: none">– Name: Enter a name of the CUSTOM type.– Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.● DLI<ul style="list-style-type: none">– Connection Name: Click In the displayed dialog box, select a DLI data connection.– Database: Click In the displayed dialog box, select a DLI database.– Table Name: Click In the displayed dialog box, select a DLI table.
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.
Output	

Parameter	Description
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">• DWS<ul style="list-style-type: none">– Connection Name: Click In the displayed dialog box, select a DWS data connection.– Database: Click In the displayed dialog box, select a DWS database.– Schema: Click In the displayed dialog box, select a DWS schema.– Table Name: Click In the displayed dialog box, select a DWS table.• OBS<ul style="list-style-type: none">– Path: Click In the displayed dialog box, select an OBS path.• CSS<ul style="list-style-type: none">– Cluster Name: Click In the displayed dialog box, select a CSS cluster.– Index: Enter a CSS index name.• HIVE<ul style="list-style-type: none">– Connection Name: Click In the displayed dialog box, select a HIVE data connection.– Database: Click In the displayed dialog box, select a HIVE database.– Table Name: Click In the displayed dialog box, select a HIVE table.• CUSTOM<ul style="list-style-type: none">– Name: Enter a name of the CUSTOM type.– Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.• DLI<ul style="list-style-type: none">– Connection Name: Click In the displayed dialog box, select a DLI data connection.– Database: Click In the displayed dialog box, select a DLI database.– Table Name: Click In the displayed dialog box, select a DLI table.
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.

6.9.13 MRS Spark SQL


Functions

The MRS Spark SQL node is used to execute a predefined SparkSQL statement on MRS.

Parameters

[Table 6-79](#), [Table 6-80](#), and [Table 6-81](#) describe the parameters of the MRS Spark SQLnode node.

Table 6-79 Parameters of MRS Spark SQL nodes

Parameter	Mand atory	Description
SQL Script	Yes	Path of a script to be executed. If no script is available, create and develop a script by referring to Creating a Script and Developing an SQL Script .
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 an EL expression . 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. NOTE 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 Running a SparkSql Job > Table 2 Program Parameter parameters in the <i>MapReduce 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. NOTE The node name cannot contain Chinese characters or more than 64 characters. If the node name does not meet requirements, the MRS job will fail to be submitted.


Table 6-80 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"> • Yes: The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> – Maximum Retries – Retry Interval (seconds) • No: The node task will not be re-executed. This is the default setting. <p>NOTE If Timeout Interval 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"> • End the current job execution plan: stops running the current job. The job instance status is Failed. • Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored. • Suspend current job execution plan: suspends running the current job. The job instance status is Waiting. • Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

Table 6-81 Lineage

Parameter	Description
Input	

Parameter	Description
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">• DWS<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DWS data connection.- Database: Click In the displayed dialog box, select a DWS database.- Schema: Click In the displayed dialog box, select a DWS schema.- Table Name: Click In the displayed dialog box, select a DWS table.• OBS<ul style="list-style-type: none">- Path: Click In the displayed dialog box, select an OBS path.• CSS<ul style="list-style-type: none">- Cluster Name: Click In the displayed dialog box, select a CSS cluster.- Index: Enter a CSS index name.• HIVE<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a HIVE data connection.- Database: Click In the displayed dialog box, select a HIVE database.- Table Name: Click In the displayed dialog box, select a HIVE table.• CUSTOM<ul style="list-style-type: none">- Name: Enter a name of the CUSTOM type.- Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.• DLI<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DLI data connection.- Database: Click In the displayed dialog box, select a DLI database.- Table Name: Click In the displayed dialog box, select a DLI table.
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.
Output	
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">• DWS<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DWS data connection.- Database: Click In the displayed dialog box, select a DWS database.- Schema: Click In the displayed dialog box, select a DWS schema.- Table Name: Click In the displayed dialog box, select a DWS table.• OBS<ul style="list-style-type: none">- Path: Click In the displayed dialog box, select an OBS path.• CSS<ul style="list-style-type: none">- Cluster Name: Click In the displayed dialog box, select a CSS cluster.- Index: Enter a CSS index name.• HIVE<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a HIVE data connection.- Database: Click In the displayed dialog box, select a HIVE database.- Table Name: Click In the displayed dialog box, select a HIVE table.• CUSTOM<ul style="list-style-type: none">- Name: Enter a name of the CUSTOM type.- Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.• DLI<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DLI data connection.- Database: Click In the displayed dialog box, select a DLI database.- Table Name: Click In the displayed dialog box, select a DLI table.

Parameter	Description
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.

6.9.14 MRS Hive SQL


Functions

The MRS Hive SQL node is used to execute a predefined Hive SQL script on DLF.

Parameters

[Table 6-82](#), [Table 6-83](#), and [Table 6-84](#) describe the parameters of the MRS Hive SQLnode node.

Table 6-82 Parameters of MRS Hive 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 Creating a Script and Developing an SQL Script .
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 an EL expression . 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. NOTE 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 Running a HiveSql Job > Table 2 Program Parameter parameters in the <i>MapReduce Service 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. NOTE The node name cannot contain Chinese characters or more than 64 characters. If the node name does not meet requirements, the MRS job will fail to be submitted.


Table 6-83 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"> • Yes: The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> – Maximum Retries – Retry Interval (seconds) • No: The node task will not be re-executed. This is the default setting. <p>NOTE If Timeout Interval 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"> • End the current job execution plan: stops running the current job. The job instance status is Failed. • Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored. • Suspend current job execution plan: suspends running the current job. The job instance status is Waiting. • Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

Table 6-84 Lineage

Parameter	Description
Input	

Parameter	Description
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">● DWS<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DWS data connection.- Database: Click In the displayed dialog box, select a DWS database.- Schema: Click In the displayed dialog box, select a DWS schema.- Table Name: Click In the displayed dialog box, select a DWS table.● OBS<ul style="list-style-type: none">- Path: Click In the displayed dialog box, select an OBS path.● CSS<ul style="list-style-type: none">- Cluster Name: Click In the displayed dialog box, select a CSS cluster.- Index: Enter a CSS index name.● HIVE<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a HIVE data connection.- Database: Click In the displayed dialog box, select a HIVE database.- Table Name: Click In the displayed dialog box, select a HIVE table.● CUSTOM<ul style="list-style-type: none">- Name: Enter a name of the CUSTOM type.- Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.● DLI<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DLI data connection.- Database: Click In the displayed dialog box, select a DLI database.- Table Name: Click In the displayed dialog box, select a DLI table.
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.
Output	
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">• DWS<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DWS data connection.- Database: Click In the displayed dialog box, select a DWS database.- Schema: Click In the displayed dialog box, select a DWS schema.- Table Name: Click In the displayed dialog box, select a DWS table.• OBS<ul style="list-style-type: none">- Path: Click In the displayed dialog box, select an OBS path.• CSS<ul style="list-style-type: none">- Cluster Name: Click In the displayed dialog box, select a CSS cluster.- Index: Enter a CSS index name.• HIVE<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a HIVE data connection.- Database: Click In the displayed dialog box, select a HIVE database.- Table Name: Click In the displayed dialog box, select a HIVE table.• CUSTOM<ul style="list-style-type: none">- Name: Enter a name of the CUSTOM type.- Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.• DLI<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DLI data connection.- Database: Click In the displayed dialog box, select a DLI database.- Table Name: Click In the displayed dialog box, select a DLI table.

Parameter	Description
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.

6.9.15 MRS Presto SQL

Functions

The MRS Presto SQL node is used to execute the Presto SQL script predefined in DataArts Factory.

Parameters

[Table 6-85](#), [Table 6-86](#), and [Table 6-87](#) describe the parameters of the MRS Presto SQL node.

Table 6-85 Property parameters

Parameters	Mandatory	Description
SQL or Script	Yes	<p>You can select SQL statement or SQL script.</p> <ul style="list-style-type: none"> SQL Statement In the SQL statement text box, enter the SQL statement to be executed. SQL Script Select a script to be executed. If the script is not created, create and develop the script by repeating steps Creating a Script and Developing an SQL Script. <p>NOTE 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 an EL expression . 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. NOTE The node name cannot contain Chinese characters or more than 64 characters. If the node name does not meet requirements, the MRS job will fail to be submitted.


Table 6-86 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"> • Yes: The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> – Maximum Retries – Retry Interval (seconds) • No: The node task will not be re-executed. This is the default setting. <p>NOTE If Timeout Interval 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"> • End the current job execution plan: stops running the current job. The job instance status is Failed. • Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored. • Suspend current job execution plan: suspends running the current job. The job instance status is Waiting. • Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

Table 6-87 Lineage

Parameter	Description
Input	

Parameter	Description
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">• DWS<ul style="list-style-type: none">– Connection Name: Click In the displayed dialog box, select a DWS data connection.– Database: Click In the displayed dialog box, select a DWS database.– Schema: Click In the displayed dialog box, select a DWS schema.– Table Name: Click In the displayed dialog box, select a DWS table.• OBS<ul style="list-style-type: none">– Path: Click In the displayed dialog box, select an OBS path.• CSS<ul style="list-style-type: none">– Cluster Name: Click In the displayed dialog box, select a CSS cluster.– Index: Enter a CSS index name.• HIVE<ul style="list-style-type: none">– Connection Name: Click In the displayed dialog box, select a HIVE data connection.– Database: Click In the displayed dialog box, select a HIVE database.– Table Name: Click In the displayed dialog box, select a HIVE table.• CUSTOM<ul style="list-style-type: none">– Name: Enter a name of the CUSTOM type.– Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.• DLI<ul style="list-style-type: none">– Connection Name: Click In the displayed dialog box, select a DLI data connection.– Database: Click In the displayed dialog box, select a DLI database.– Table Name: Click In the displayed dialog box, select a DLI table.
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.
Output	
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">• DWS<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DWS data connection.- Database: Click In the displayed dialog box, select a DWS database.- Schema: Click In the displayed dialog box, select a DWS schema.- Table Name: Click In the displayed dialog box, select a DWS table.• OBS<ul style="list-style-type: none">- Path: Click In the displayed dialog box, select an OBS path.• CSS<ul style="list-style-type: none">- Cluster Name: Click In the displayed dialog box, select a CSS cluster.- Index: Enter a CSS index name.• HIVE<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a HIVE data connection.- Database: Click In the displayed dialog box, select a HIVE database.- Table Name: Click In the displayed dialog box, select a HIVE table.• CUSTOM<ul style="list-style-type: none">- Name: Enter a name of the CUSTOM type.- Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.• DLI<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DLI data connection.- Database: Click In the displayed dialog box, select a DLI database.- Table Name: Click In the displayed dialog box, select a DLI table.

Parameter	Description
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.

6.9.16 MRS Spark


Functions

The MRS Spark node is used to execute a predefined Spark job on MRS.

Parameters

[Table 6-88](#), [Table 6-89](#), and [Table 6-90](#) describe the parameters of the MRS Sparknode node.

Table 6-88 Parameters of MRS Spark 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 (>).
MRS Cluster Name	Yes	Name of the MRS cluster. To create an MRS cluster, use either of the following methods: <ul style="list-style-type: none"> Click . On the Clusters page, create an MRS cluster. Go to the MRS console to create an MRS cluster.
Spark Job Name	Yes	Name of an MRS job. The name contains 1 to 64 characters, including only letters, digits, and underscores (_). NOTE The job name cannot contain Chinese characters or 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 JAR package . Before selecting a JAR package, upload the JAR package to the OBS bucket, create a resource on the Manage Resource page, and add the JAR package to the resource management list. For details, see Creating a Resource .
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. NOTE 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 Running a Spark Job in the <i>MapReduce Service User Guide</i> .
Input Data Path	No	Path where the input data resides.
Output Data Path	No	Path where the output data resides.


Table 6-89 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"> • Yes: The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> – Maximum Retries – Retry Interval (seconds) • No: The node task will not be re-executed. This is the default setting. <p>NOTE If Timeout Interval 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"> • End the current job execution plan: stops running the current job. The job instance status is Failed. • Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored. • Suspend current job execution plan: suspends running the current job. The job instance status is Waiting. • Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

Table 6-90 Lineage

Parameter	Description
Input	

Parameter	Description
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">• DWS<ul style="list-style-type: none">– Connection Name: Click In the displayed dialog box, select a DWS data connection.– Database: Click In the displayed dialog box, select a DWS database.– Schema: Click In the displayed dialog box, select a DWS schema.– Table Name: Click In the displayed dialog box, select a DWS table.• OBS<ul style="list-style-type: none">– Path: Click In the displayed dialog box, select an OBS path.• CSS<ul style="list-style-type: none">– Cluster Name: Click In the displayed dialog box, select a CSS cluster.– Index: Enter a CSS index name.• HIVE<ul style="list-style-type: none">– Connection Name: Click In the displayed dialog box, select a HIVE data connection.– Database: Click In the displayed dialog box, select a HIVE database.– Table Name: Click In the displayed dialog box, select a HIVE table.• CUSTOM<ul style="list-style-type: none">– Name: Enter a name of the CUSTOM type.– Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.• DLI<ul style="list-style-type: none">– Connection Name: Click In the displayed dialog box, select a DLI data connection.– Database: Click In the displayed dialog box, select a DLI database.– Table Name: Click In the displayed dialog box, select a DLI table.
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.
Output	
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">• DWS<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DWS data connection.- Database: Click In the displayed dialog box, select a DWS database.- Schema: Click In the displayed dialog box, select a DWS schema.- Table Name: Click In the displayed dialog box, select a DWS table.• OBS<ul style="list-style-type: none">- Path: Click In the displayed dialog box, select an OBS path.• CSS<ul style="list-style-type: none">- Cluster Name: Click In the displayed dialog box, select a CSS cluster.- Index: Enter a CSS index name.• HIVE<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a HIVE data connection.- Database: Click In the displayed dialog box, select a HIVE database.- Table Name: Click In the displayed dialog box, select a HIVE table.• CUSTOM<ul style="list-style-type: none">- Name: Enter a name of the CUSTOM type.- Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.• DLI<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DLI data connection.- Database: Click In the displayed dialog box, select a DLI database.- Table Name: Click In the displayed dialog box, select a DLI table.

Parameter	Description
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.

6.9.17 MRS Spark Python

Functions


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 .

Parameters

[Table 6-91](#), [Table 6-92](#), and [Table 6-93](#) describe the parameters of the MRS Spark Pythonnode node.

Table 6-91 Parameters of MRS Spark Python 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 (>).
MRS Cluster Name	Yes	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. To create an MRS cluster, use either of the following methods: <ul style="list-style-type: none">Click . On the Clusters page, create an MRS cluster.Go to the MRS console to create an MRS cluster. For details about how to create a cluster, see Custom Purchase of a Cluster .

Parameter	Mandatory	Description
Job Name	Yes	Name of an MRS job. The name contains 1 to 64 characters, including only letters, digits, and underscores (_). NOTE The job name cannot contain Chinese characters or more than 64 characters. If the job name does not meet requirements, the MRS job will fail to be submitted.
Parameter	Yes	Enter the parameters of the executable program of MRS. Use Enter to separate multiple parameters.
Attribute	No	Enter parameters in the key=value format. Use Enter to separate multiple parameters.


Table 6-92 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"> • Yes: The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> - Maximum Retries - Retry Interval (seconds) • No: The node task will not be re-executed. This is the default setting. NOTE If Timeout Interval 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.




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"> • End the current job execution plan: stops running the current job. The job instance status is Failed. • Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored. • Suspend current job execution plan: suspends running the current job. The job instance status is Waiting. • Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

Table 6-93 Lineage

Parameter	Description
Input	

Parameter	Description
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">• DWS<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DWS data connection.- Database: Click In the displayed dialog box, select a DWS database.- Schema: Click In the displayed dialog box, select a DWS schema.- Table Name: Click In the displayed dialog box, select a DWS table.• OBS<ul style="list-style-type: none">- Path: Click In the displayed dialog box, select an OBS path.• CSS<ul style="list-style-type: none">- Cluster Name: Click In the displayed dialog box, select a CSS cluster.- Index: Enter a CSS index name.• HIVE<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a HIVE data connection.- Database: Click In the displayed dialog box, select a HIVE database.- Table Name: Click In the displayed dialog box, select a HIVE table.• CUSTOM<ul style="list-style-type: none">- Name: Enter a name of the CUSTOM type.- Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.• DLI<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DLI data connection.- Database: Click In the displayed dialog box, select a DLI database.- Table Name: Click In the displayed dialog box, select a DLI table.
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.
Output	
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">● DWS<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DWS data connection.- Database: Click In the displayed dialog box, select a DWS database.- Schema: Click In the displayed dialog box, select a DWS schema.- Table Name: Click In the displayed dialog box, select a DWS table.● OBS<ul style="list-style-type: none">- Path: Click In the displayed dialog box, select an OBS path.● CSS<ul style="list-style-type: none">- Cluster Name: Click In the displayed dialog box, select a CSS cluster.- Index: Enter a CSS index name.● HIVE<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a HIVE data connection.- Database: Click In the displayed dialog box, select a HIVE database.- Table Name: Click In the displayed dialog box, select a HIVE table.● CUSTOM<ul style="list-style-type: none">- Name: Enter a name of the CUSTOM type.- Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.● DLI<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DLI data connection.- Database: Click In the displayed dialog box, select a DLI database.- Table Name: Click In the displayed dialog box, select a DLI table.

Parameter	Description
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.

6.9.18 MRS Flink Job


Functions

The MRS Flink node is used to execute predefined Flink jobs in MRS.

Parameters

[Table 6-94](#) and [Table 6-95](#) describe the parameters of the MRS Flink node.

Table 6-94 Parameters of the MRS Flink 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 (>).
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"> Click . On the Clusters page, create an MRS cluster. Go to the MRS console to create an MRS cluster.
Job Name	Yes	Name of an MRS job. The name contains 1 to 64 characters, including only letters, digits, and underscores (_). NOTE The job name cannot contain Chinese characters or 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 Manage Resource page, and add the JAR package to the resource management list. For details, see Creating a Resource .
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. NOTE This parameter is mandatory if the cluster version is MRS 1.8.7 or later than MRS 2.0.1. For details on the program parameters of MRS Spark jobs, see Running a Flink Job in the <i>MapReduce Service User Guide</i> .
Input Data Path	No	Path where the input data resides.
Output Data Path	No	Path where the output data resides.

Table 6-95 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">• Yes: The node task will be re-executed, and the following parameters must be configured:<ul style="list-style-type: none">– Maximum Retries– Retry Interval (seconds)• No: The node task will not be re-executed. This is the default setting. <p>NOTE If Timeout Interval 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">• End the current job execution plan: stops running the current job. The job instance status is Failed.• Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored.• Suspend current job execution plan: suspends running the current job. The job instance status is Waiting.• Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

6.9.19 MRS MapReduce

Functions

The MRS MapReduce node is used to execute a predefined MapReduce program on MRS.

Parameters

[Table 6-96](#) and [Table 6-97](#) describe the parameters of the MRS MapReduce node.

Table 6-96 Parameters of MRS MapReduce 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 (>).
MRS Cluster Name	Yes	Name of the MRS cluster. To create an MRS cluster, use either of the following methods: <ul style="list-style-type: none"> Click . On the Clusters page, create an MRS cluster. Go to the MRS console to create an MRS cluster.
MapReduce Job Name	Yes	Name of an MRS job. The name contains 1 to 64 characters, including only letters, digits, and underscores (_). NOTE The job name cannot contain Chinese characters or 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 Manage Resource page, and add the JAR package to the resource management list. For details, see Creating a Resource .
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 6-97 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">• Yes: The node task will be re-executed, and the following parameters must be configured:<ul style="list-style-type: none">– Maximum Retries– Retry Interval (seconds)• No: The node task will not be re-executed. This is the default setting. NOTE If Timeout Interval 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.
Failure Policy	Yes	Operation that will be performed if the node task fails to be executed. Possible values: <ul style="list-style-type: none">• End the current job execution plan: stops running the current job. The job instance status is Failed.• Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored.• Suspend current job execution plan: suspends running the current job. The job instance status is Waiting.• Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

6.9.20 CSS

Functions

The CSS node is used to process CSS requests and enable online distributed searching.

Parameters

[Table 6-98](#) and [Table 6-99](#) describe the parameters of the CSS node.

Table 6-98 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"> • GET • POST • PUT • HEAD • DELETE
Request Parameter	No	Parameter of the request. For example, to query the dlfddata mapping type in the dlf_search index, set this parameter to: /dlf_search/dlfddata/_search
Request Body	No	The request body is in JSON format.
CloudSearch Output Path	No	Path where output data is to be stored.

Table 6-99 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"> • Yes: The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> – Maximum Retries – Retry Interval (seconds) • No: The node task will not be re-executed. This is the default setting. <p>NOTE If Timeout Interval 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"> • End the current job execution plan: stops running the current job. The job instance status is Failed. • Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored. • Suspend current job execution plan: suspends running the current job. The job instance status is Waiting. • Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

6.9.21 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 6-100](#) and [Table 6-101](#) describe the parameters of the Shell node.

Table 6-100 Parameters of Shell nodes

Parameter	Mandatory	Description
Shell or Script	Yes	<p>You can select Shell statement or Shell script.</p> <ul style="list-style-type: none"> Shell statement In the Shell statement text box, enter the Shell statement to be executed. Shell script Select a script to be executed. If no script is available, create and develop a script by referring to Creating a Script and Developing a Shell Script. <p>NOTE If you select Shell statement, 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: a b c . 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 6-101 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	Indicates whether to re-execute a node task if its execution fails. Possible values: <ul style="list-style-type: none">• Yes: The node task will be re-executed, and the following parameters must be configured:<ul style="list-style-type: none">- Maximum Retries- Retry Interval (seconds)• No: The node task will not be re-executed. This is the default setting. NOTE If Timeout Interval 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.
Failure Policy	Yes	Operation that will be performed if the node task fails to be executed. Possible values: <ul style="list-style-type: none">• End the current job execution plan: stops running the current job. The job instance status is Failed.• Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored.• Suspend current job execution plan: suspends running the current job. The job instance status is Waiting.• Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

6.9.22 RDS SQL

Functions

The RDS SQL node is used to transfer SQL statements to RDS.

Parameters

[Table 6-102](#) and [Table 6-103](#) describe the parameters of the RDS SQL node.

Table 6-102 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 SQL statement or SQL script.</p> <ul style="list-style-type: none"> SQL statement In the Statements text box, enter the SQL statement to be executed. SQL script Select a script to be executed. If no script is available, create and develop a script by referring to Creating a Script and Developing an SQL Script. <p>NOTE If you select the SQL statement mode, the DataArts Factory module cannot parse the parameters contained in the SQL statement.</p>

Table 6-103 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">• Yes: The node task will be re-executed, and the following parameters must be configured:<ul style="list-style-type: none">– Maximum Retries– Retry Interval (seconds)• No: The node task will not be re-executed. This is the default setting. NOTE If Timeout Interval 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.
Failure Policy	Yes	Operation that will be performed if the node task fails to be executed. Possible values: <ul style="list-style-type: none">• End the current job execution plan: stops running the current job. The job instance status is Failed.• Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored.• Suspend current job execution plan: suspends running the current job. The job instance status is Waiting.• Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

6.9.23 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 6-104](#), [Table 6-105](#), and [Table 6-106](#) describe the parameters of the ETL Job node.

Table 6-104 Parameters of Transform Load 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 (>).
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">• When the source data type is DLI, the supported destination data types are DWS, GES, CSS, OBS, and DLI.• When the source data type is MySQL, the supported destination data type is MySQL.• When the source data type is OBS, the supported destination data can be of the DLI type and the DWS type. <p>NOTICE</p> <ul style="list-style-type: none">• 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.• 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>.
Configure SQL Template	No	Click Obtain Template to obtain an SQL template.



Table 6-105 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"> • Yes: The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> - Maximum Retries - Retry Interval (seconds) • No: The node task will not be re-executed. This is the default setting. <p>NOTE If Timeout Interval 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"> • End the current job execution plan: stops running the current job. The job instance status is Failed. • Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored. • Suspend current job execution plan: suspends running the current job. The job instance status is Waiting. • Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

Table 6-106 Lineage

Parameter	Description
Input	

Parameter	Description
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">• DWS<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DWS data connection.- Database: Click In the displayed dialog box, select a DWS database.- Schema: Click In the displayed dialog box, select a DWS schema.- Table Name: Click In the displayed dialog box, select a DWS table.• OBS<ul style="list-style-type: none">- Path: Click In the displayed dialog box, select an OBS path.• CSS<ul style="list-style-type: none">- Cluster Name: Click In the displayed dialog box, select a CSS cluster.- Index: Enter a CSS index name.• HIVE<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a HIVE data connection.- Database: Click In the displayed dialog box, select a HIVE database.- Table Name: Click In the displayed dialog box, select a HIVE table.• CUSTOM<ul style="list-style-type: none">- Name: Enter a name of the CUSTOM type.- Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.• DLI<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DLI data connection.- Database: Click In the displayed dialog box, select a DLI database.- Table Name: Click In the displayed dialog box, select a DLI table.
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.
Output	
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">● DWS<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DWS data connection.- Database: Click In the displayed dialog box, select a DWS database.- Schema: Click In the displayed dialog box, select a DWS schema.- Table Name: Click In the displayed dialog box, select a DWS table.● OBS<ul style="list-style-type: none">- Path: Click In the displayed dialog box, select an OBS path.● CSS<ul style="list-style-type: none">- Cluster Name: Click In the displayed dialog box, select a CSS cluster.- Index: Enter a CSS index name.● HIVE<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a HIVE data connection.- Database: Click In the displayed dialog box, select a HIVE database.- Table Name: Click In the displayed dialog box, select a HIVE table.● CUSTOM<ul style="list-style-type: none">- Name: Enter a name of the CUSTOM type.- Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.● DLI<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DLI data connection.- Database: Click In the displayed dialog box, select a DLI database.- Table Name: Click In the displayed dialog box, select a DLI table.

Parameter	Description
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.

6.9.24 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 6-107](#) and [Table 6-108](#) describe the parameters of the Python node.

Table 6-107 Parameters of the Python node

Parameter	Mandatory	Description
Python or Script	Yes	<p>You can select Python statement or Python script.</p> <ul style="list-style-type: none"> Python statement In the Python statement text box, enter the Python statement to be executed. Python script Select a script to be executed for Script Path. If no script is available, create and develop a script by referring to Creating a Script and Developing a Python Script. <p>NOTE If you select Python statement, 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 6-108 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"> • Yes: The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> - Maximum Retries - Retry Interval (seconds) • No: The node task will not be re-executed. This is the default setting. <p>NOTE If Timeout Interval 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">• End the current job execution plan: stops running the current job. The job instance status is Failed.• Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored.• Suspend current job execution plan: suspends running the current job. The job instance status is Waiting.• Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

6.9.25 Create OBS

Constraints

This function depends on OBS.

Functions

The Create OBS node is used to create buckets and directories on OBS.

Parameters

[Table 6-109](#) and [Table 6-110](#) describe the parameters of the Create OBS node.

Table 6-109 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"> To create a bucket, enter <i>//OBS bucket name</i>. The OBS bucket name must be unique 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.

Table 6-110 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"> Yes: The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> Maximum Retries Retry Interval (seconds) No: The node task will not be re-executed. This is the default setting. <p>NOTE If Timeout Interval 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">• End the current job execution plan: stops running the current job. The job instance status is Failed.• Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored.• Suspend current job execution plan: suspends running the current job. The job instance status is Waiting.• Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

6.9.26 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 6-111](#) and [Table 6-112](#) describe the parameters of the Delete OBS node.

Table 6-111 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	Path to the OBS bucket or directory. NOTE 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.

Table 6-112 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"> • Yes: The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> - Maximum Retries - Retry Interval (seconds) • No: The node task will not be re-executed. This is the default setting. NOTE If Timeout Interval 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.

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">• End the current job execution plan: stops running the current job. The job instance status is Failed.• Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored.• Suspend current job execution plan: suspends running the current job. The job instance status is Waiting.• Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

6.9.27 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 6-113](#), [Table 6-114](#), and [Table 6-115](#) describe the parameters of the OBS Managernode node.

Table 6-113 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	Operations that can be performed on the node. <ul style="list-style-type: none">• Move File: moves a source file or directory to a new directory.• Copy File: copies the source file or directory.• Rename File: renames the last level of the directory or file. For example, you can rename the directory obs://test/a/b/c/ as obs://test/a/b/d/, and rename the file obs://test/a/b/hello.txt as obs://test/a/b/bye.txt.• Monitor File: 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.
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 6-114 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"> • Yes: The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> – Maximum Retries – Retry Interval (seconds) • No: The node task will not be re-executed. This is the default setting. <p>NOTE If Timeout Interval 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"> • End the current job execution plan: stops running the current job. The job instance status is Failed. • Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored. • Suspend current job execution plan: suspends running the current job. The job instance status is Waiting. • Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

Table 6-115 Lineage

Parameter	Description
Input	

Parameter	Description
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">• DWS<ul style="list-style-type: none">– Connection Name: Click In the displayed dialog box, select a DWS data connection.– Database: Click In the displayed dialog box, select a DWS database.– Schema: Click In the displayed dialog box, select a DWS schema.– Table Name: Click In the displayed dialog box, select a DWS table.• OBS<ul style="list-style-type: none">– Path: Click In the displayed dialog box, select an OBS path.• CSS<ul style="list-style-type: none">– Cluster Name: Click In the displayed dialog box, select a CSS cluster.– Index: Enter a CSS index name.• HIVE<ul style="list-style-type: none">– Connection Name: Click In the displayed dialog box, select a HIVE data connection.– Database: Click In the displayed dialog box, select a HIVE database.– Table Name: Click In the displayed dialog box, select a HIVE table.• CUSTOM<ul style="list-style-type: none">– Name: Enter a name of the CUSTOM type.– Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.• DLI<ul style="list-style-type: none">– Connection Name: Click In the displayed dialog box, select a DLI data connection.– Database: Click In the displayed dialog box, select a DLI database.– Table Name: Click In the displayed dialog box, select a DLI table.
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.
Output	
Add	<p>Click Add. In the Type drop-down list, select the type to be created. The value can be DWS, OBS, CSS, HIVE, DLI, or CUSTOM.</p> <ul style="list-style-type: none">● DWS<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DWS data connection.- Database: Click In the displayed dialog box, select a DWS database.- Schema: Click In the displayed dialog box, select a DWS schema.- Table Name: Click In the displayed dialog box, select a DWS table.● OBS<ul style="list-style-type: none">- Path: Click In the displayed dialog box, select an OBS path.● CSS<ul style="list-style-type: none">- Cluster Name: Click In the displayed dialog box, select a CSS cluster.- Index: Enter a CSS index name.● HIVE<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a HIVE data connection.- Database: Click In the displayed dialog box, select a HIVE database.- Table Name: Click In the displayed dialog box, select a HIVE table.● CUSTOM<ul style="list-style-type: none">- Name: Enter a name of the CUSTOM type.- Attribute: Enter an attribute of the CUSTOM type. You can add more than one attribute.● DLI<ul style="list-style-type: none">- Connection Name: Click In the displayed dialog box, select a DLI data connection.- Database: Click In the displayed dialog box, select a DLI database.- Table Name: Click In the displayed dialog box, select a DLI table.

Parameter	Description
OK	Click OK to save the parameter settings.
Cancel	Click Cancel 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.

6.9.28 Open/Close Resource

Functions

You can use the Open/Close Resource node to enable or disable services as required.

Parameters

[Table 6-116](#) and [Table 6-117](#) describe the parameters of the Open/Close Resource node.

Table 6-116 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"> • ECS • CDM
Open/Close Resource	Yes	Possible values: <ul style="list-style-type: none"> • On • Off
Instance	Yes	Object to be opened or closed, for example, to open a CDM cluster.

Table 6-117 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"> • Yes: The node task will be re-executed, and the following parameters must be configured: <ul style="list-style-type: none"> - Maximum Retries - Retry Interval (seconds) • No: The node task will not be re-executed. This is the default setting. <p>NOTE If Timeout Interval 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"> • End the current job execution plan: stops running the current job. The job instance status is Failed. • Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored. • Suspend current job execution plan: suspends running the current job. The job instance status is Waiting. • Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

6.9.29 Data Quality Monitor

Functions

The Data Quality Monitor node is used to monitor the quality of running data.

Parameters

[Table 6-118](#) and [Table 6-119](#) describe the parameters of the Data Quality Monitor node.

Table 6-118 Parameters of Data Quality Monitor 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 (>).
Job Type	Yes	Data quality job. The following options are available: <ul style="list-style-type: none">• Quality job• Comparison job
Quality Job Name	Yes	Name of a quality job created in DataArts Quality. This parameter is mandatory when Job Type is Quality Job . For details about how to create a quality job, see Creating Quality Jobs .
Ignore Quality Job Alarm	Yes	This parameter is mandatory when Job Type is Quality Job . <ul style="list-style-type: none">• Yes: If the quality job is in the alarm state, the status of the current node is set to successful and the subsequent nodes continue to be executed.• No: If the quality job is in the alarm state, the status of the current node is set to failed.
Comparison Job Name	Yes	Name of a comparison job created in DataArts Quality. This parameter is mandatory when Job Type is Comparison Job . For details about how to create a comparison job, see Creating a Comparison Job .
Ignore Comparison Job Alarm	Yes	This parameter is mandatory when Job Type is Comparison Job . <ul style="list-style-type: none">• Yes: If the comparison job is in the alarm state, the status of the current node is set to successful and the subsequent nodes continue to be executed.• No: If the comparison job is in the alarm state, the status of the current node is set to failed.

Table 6-119 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">• Yes: The node task will be re-executed, and the following parameters must be configured:<ul style="list-style-type: none">- Maximum Retries- Retry Interval (seconds)• No: The node task will not be re-executed. This is the default setting. NOTE If Timeout Interval 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.
Failure Policy	Yes	Operation that will be performed if the node task fails to be executed. Possible values: <ul style="list-style-type: none">• End the current job execution plan: stops running the current job. The job instance status is Failed.• Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored.• Suspend current job execution plan: suspends running the current job. The job instance status is Waiting.• Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

6.9.30 Subjob

Function

The Subjob node is used to call the batch job that does not contain the subjob node.

Parameter

[Table 6-120](#) and [Table 6-121](#) describe the parameters of the Subjob node.

Table 6-120 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. NOTE 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">• If the subjob parameters are left unspecified, the subjob is executed with its own parameter variables. The Subjob Parameter Name of the parent job is not displayed.• If the subjob parameters are specified, the subjob is executed with the configured parameter values. In this case, the Subjob Parameter Name 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.

Table 6-121 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">• Yes: The node task will be re-executed, and the following parameters must be configured:<ul style="list-style-type: none">– Maximum Retries– Retry Interval (seconds)• No: The node task will not be re-executed. This is the default setting. <p>NOTE If Timeout Interval 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">• End the current job execution plan: stops running the current job. The job instance status is Failed.• Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored.• Suspend current job execution plan: suspends running the current job. The job instance status is Waiting.• Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

6.9.31 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.

For details about how to use the For Each operator, see [Introduction to the For Each Operator](#).

Parameters

[Table 6-122](#) describes the parameters of the For Each node.

Table 6-122 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">• 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>#{Job.getNodeOutput('preNodeName')}</code> is used, which means the output of the previous node.• A specified array, for example, <code>['001'],['002'], ['003']</code>
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">• If the subjob parameters are left unspecified, the subjob is executed with its own parameter variables.• 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.

Table 6-123 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">• Yes: The node task will be re-executed, and the following parameters must be configured:<ul style="list-style-type: none">– Maximum Retries– Retry Interval (seconds)• No: The node task will not be re-executed. This is the default setting. NOTE If Timeout Interval 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.
Failure Policy	Yes	Operation that will be performed if the node task fails to be executed. Possible values: <ul style="list-style-type: none">• End the current job execution plan: stops running the current job. The job instance status is Failed.• Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored.• Suspend current job execution plan: suspends running the current job. The job instance status is Waiting.• Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

6.9.32 SMN

Functions

The SMN node is used to send notifications to users.

Parameters

[Table 6-124](#) and [Table 6-125](#) describe the parameters of the SMN node.

Table 6-124 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">• Text: The message is sent in text format.• JSON: The message is sent in JSON format. You can send different messages to types of subscribers.<ul style="list-style-type: none">– Manual: You can enter a message in Message Content.– Automatic: Click Generate JSON Message. In the displayed dialog box, enter a message and select a protocol.• Template: 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">– Manual: You can enter a message in Message Content.– Automatic: Click Generate Template Message. In the displayed dialog box, select a template name and set the value of tag.

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"> • Text: The size cannot exceed 10 KB. • JSON: 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> • Template: The size cannot exceed 10 KB. Example: <pre> "message_template_name":"confirm_message", "tags":{ "topic_urn":"urn:smn:regionId:xxxx:SMN_01" } </pre> <p>In the preceding information, message_template_name indicates the template name, and tags 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 6-125 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.

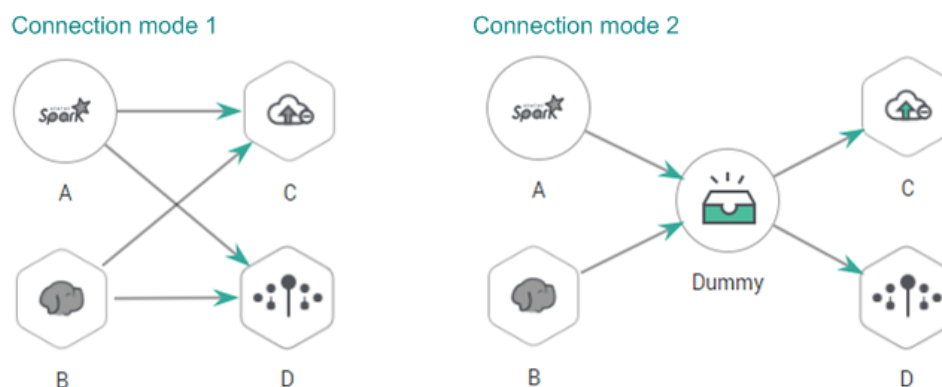
Parameter	Mandatory	Description
Retry upon Failure	Yes	<p>Indicates whether to re-execute a node task if its execution fails. Possible values:</p> <ul style="list-style-type: none">• Yes: The node task will be re-executed, and the following parameters must be configured:<ul style="list-style-type: none">– Maximum Retries– Retry Interval (seconds)• No: The node task will not be re-executed. This is the default setting. <p>NOTE If Timeout Interval 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">• End the current job execution plan: stops running the current job. The job instance status is Failed.• Go to the next node: ignores the execution failure of the current node. The job instance status is Failure ignored.• Suspend current job execution plan: suspends running the current job. The job instance status is Waiting.• Suspend execution plans of the subsequent nodes: stops running subsequent nodes. The job instance status is Failed.

6.9.33 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 6-76](#) shows an example.

Figure 6-76 Connection modes



Parameters

[Table 6-126](#) describes the parameter of Dummy nodes.

Table 6-126 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 (>).

6.10 EL Expression Reference

6.10.1 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.

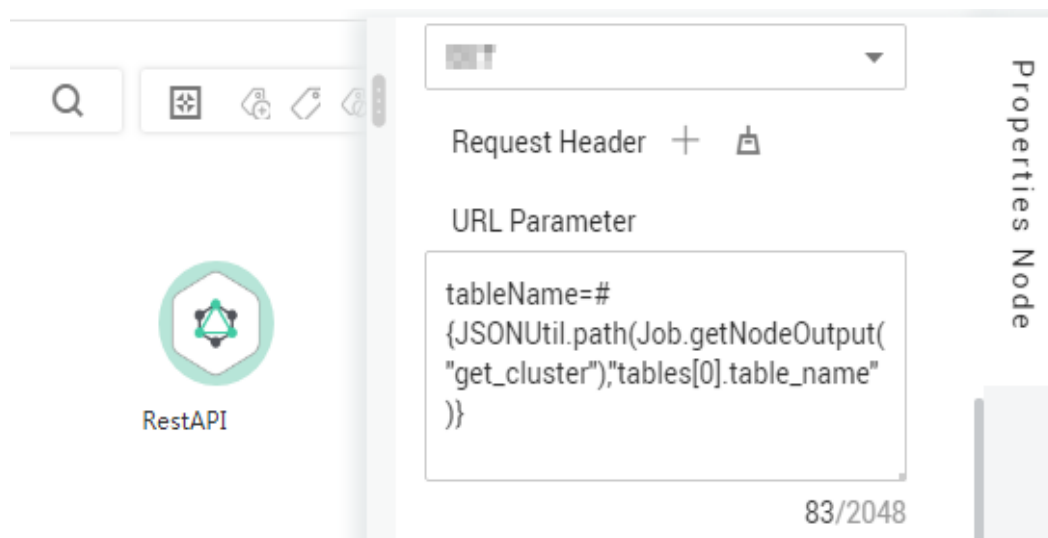
Example

In the **URL** parameter of the Rest Client node, use expression **tableName=#{JSONUtil.path(Job.getNodeOutput("get_cluster"),"tables[0].table_name")}**, as shown in [Figure 6-77](#).

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.

Figure 6-77 Expression example



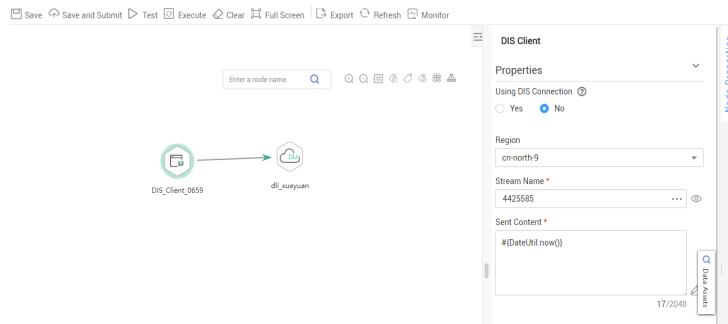
EL expressions are widely used in data development. For details, see [Best Practices](#).

Debugging Methods

You can debug EL expressions using the following methods.

This section uses the `#{DateUtil.now()}` expression as an example.

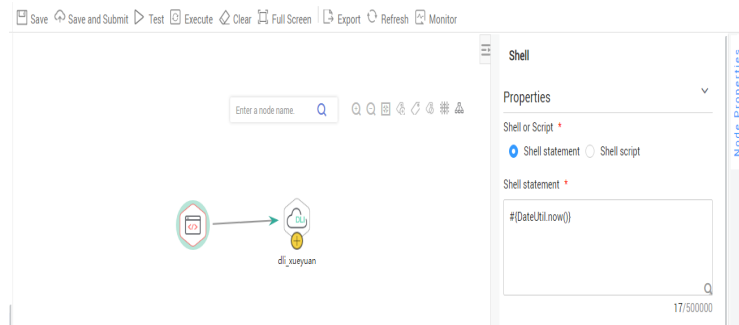
1. Use the DIS Client node.
 - Prerequisites: A DIS stream is available.
 - Method: Select the DIS Client node, write the EL expression in the data to be sent, and click **Test**. Then right-click the node to view the log. The value of the EL expression is printed in the log.



```
[2021/05/10 17:13:28 GMT+0800] [INFO] Execute user name is qiujiaxin, user id is 09f65b013200d2171fbc01587ba73e6, job id is 638744FBB2F742899337D06A08A39496oHgyCFVI
[2021/05/10 17:13:28 GMT+0800] [INFO] streamName=4425585
[2021/05/10 17:13:28 GMT+0800] [INFO] data=Mon May 10 17:13:27 GMT+08:00 2021
[2021/05/10 17:13:28 GMT+0800] [INFO] response:{"records":[{"sequence_number":"120","partition_id":"shardId-0000000000"}],"failed_record_count":0}
```

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.



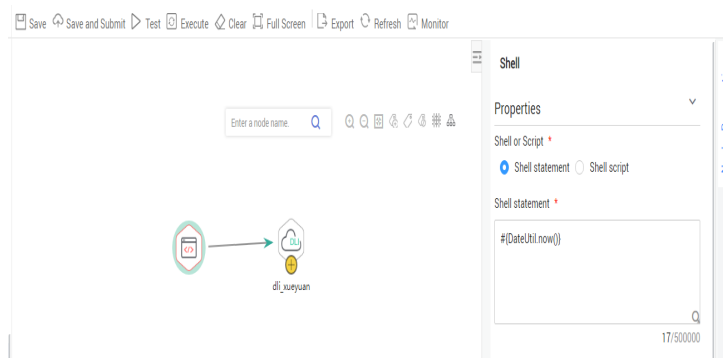
```

"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] =====

```

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.



6.10.2 Basic Operators

EL supports most of the arithmetic and logic operators provided by Java.

Operator List

Table 6-127 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.

Operator	Description
<= 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}
```

6.10.3 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 6-128](#).

Table 6-128 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

Letter	Description	Example
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")}
```

6.10.4 Env Embedded Objects

An Env embedded object provides a method of obtaining an environment variable value.

Method

Table 6-129 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")}
```


6.10.5 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 6-130 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 6-131 Method description

Method	Description
String getNodeStatus(String nodeName)	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. For example, to check whether a node is running successfully, you can use the following command, where test indicates the node name: <code>#{(Job.getNodeStatus("test")) == "success" }</code>
String getNodeOutput(String nodeName)	Obtains the output of a specified node. This method can only obtain the output of the previous dependent node.

Method	Description
String getParam(String key)	Obtains job parameters. 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. 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.
String getPlanTime(String pattern)	Obtains the plan time character string in a specified pattern. Pattern indicates the date and time mode. For details, see Date and Time Mode .
String getYesterday(String pattern)	Obtains the time character string of the day before the plan time. Pattern indicates the date and time mode. For details, see Date and Time Mode .
String getLastHour(String pattern)	Obtains the time character string of last hour before the plan time. Pattern indicates the date and time mode. For details, see Date and Time Mode .
String getRunningData(String nodeName)	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 DLI_INSERT_DATA , run the following command: <code>#{JSONUtil.path(Job.getRunningData("DLI_INSERT_DATA"),"jobIds[2]")}</code> .
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 nodeName 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 null value is returned.

Example

The expression used to obtain the output of node **test** in the job is as follows:

```
#{Job.getNodeOutput("test")}
```

6.10.6 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,".")}
```

6.10.7 DateUtil Embedded Objects

A DateUtil embedded object provides methods of formatting time and calculating time.

Methods

Table 6-132 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.
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.

Method	Description
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 Date and Time Mode .

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')}
```

6.10.8 JSONUtil Embedded Objects

A JSONUtil embedded object provides JSON object methods.

Methods

Table 6-133 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")}
```

6.10.9 Loop Embedded Objects

You can use Loop embedded objects to obtain data from the For Each dataset.

Property

Table 6-134 Property description

Property	Type	Description
dataArray	String	Dataset input by the For Each node. It is a two-dimensional array.
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]}
```

6.10.10 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 6-135 Method description

Method	Description
boolean <code>isExistOBSPath(String obsPath)</code>	Check whether the OBS file or the OBS directory that ends with a slash (/) exists. If the file or directory exists, true is returned. If not, false 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")}`

6.10.11 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.

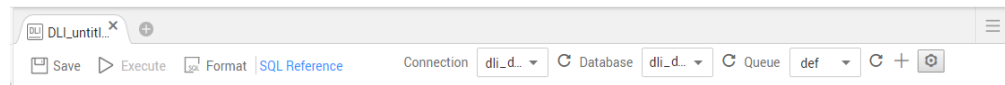
If the tables are not created, create them. For details, see [Creating a Table](#).

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.

Figure 6-78 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.

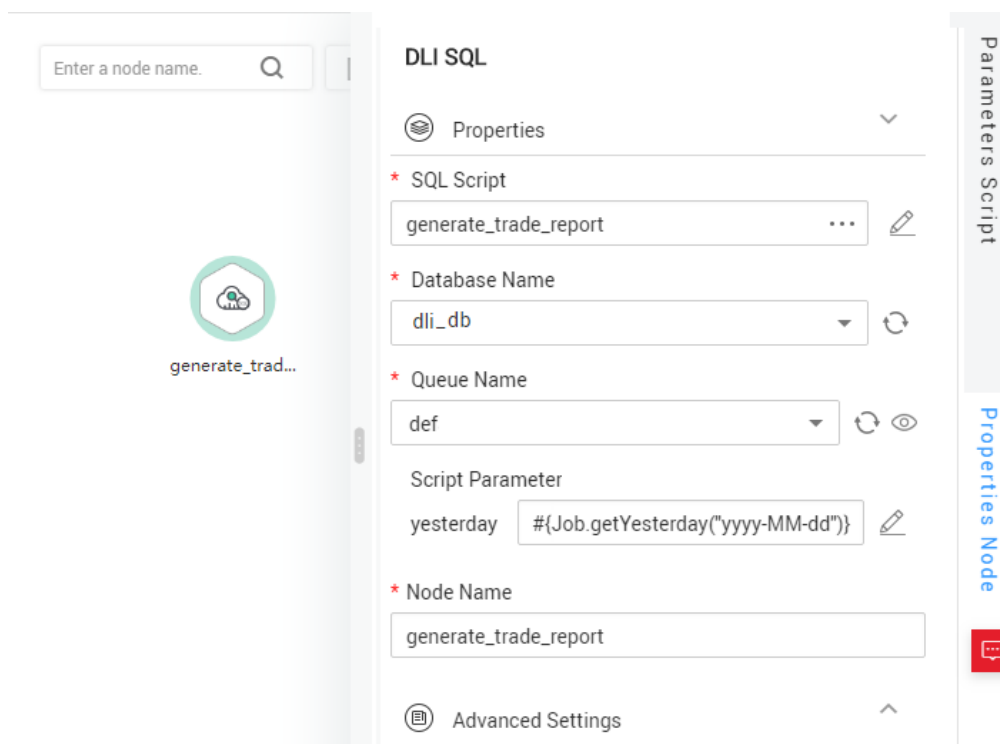
1. In the navigation tree of the DataArts Factory console, choose **Data Development > Develop Job**.
2. Access the area on the right and click **Create Job** to create an empty job named **job**.

Figure 6-79 Creating a job



3. Go to the job development page, drag the DLI SQL node to the canvas, click the icon, and configure node properties.

Figure 6-80 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

More Examples

EL expressions are widely used in data development. For details, see [Best Practices](#).

6.11 Usage Guidance

6.11.1 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 [Setting Up Scheduling for a Job Using the Batch Processing Mode](#).

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 6-81 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](#).

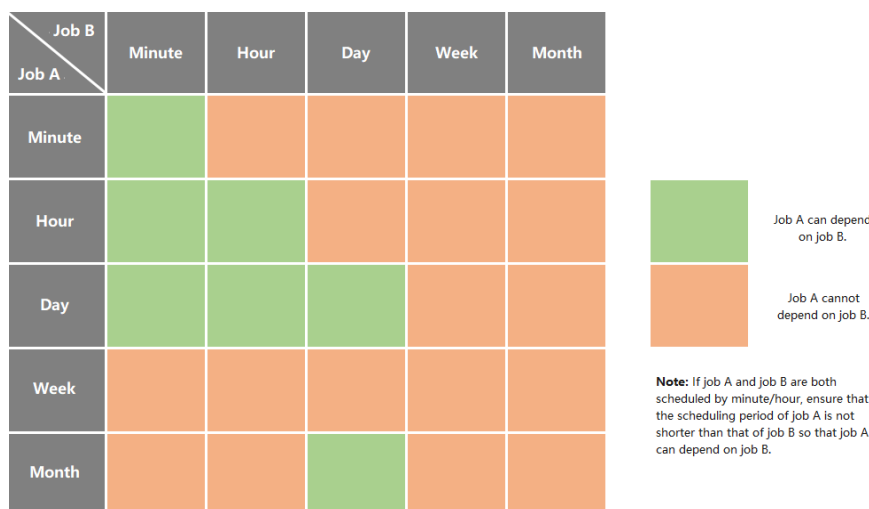
Setting Conditions of a Dependency Job

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 6-82](#) shows the requirements of the recurrences of the jobs that can function as the dependency jobs of other jobs

Figure 6-82 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 6-136 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	-
10:20	Executed	Executed after job B's instances are executed in the (10:00, 10:20] time period
10:30	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)
...

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 6-137 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 (09:00 on July 31, 09:00 on August 1] time period
10:00 on August 1	Executed	-
09:00 on August 2	-	Executed after job B's instances are executed in the (09:00 on August 1, 09:00 on August 2] 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 6-138 Example 3: dependency between jobs with different recurrences

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)
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 [00:00:00 on day 0, 00:00:00 on day 1) 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 [00:00:00 on day 1, 00:00:00 on day 2) 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 6-139 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 [00:00:00 on July 1, 00:00:00 on August 1) 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 [00:00:00 on July 1, 00:00:00 on August 1) 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
...

6.11.2 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.

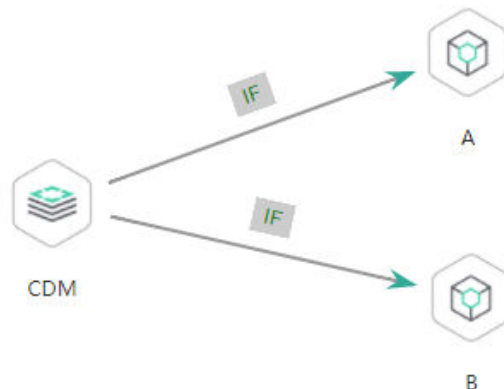
For details about how to use EL expressions, see [EL Expressions](#).

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 6-83](#).

Figure 6-83 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 6-83](#). 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.

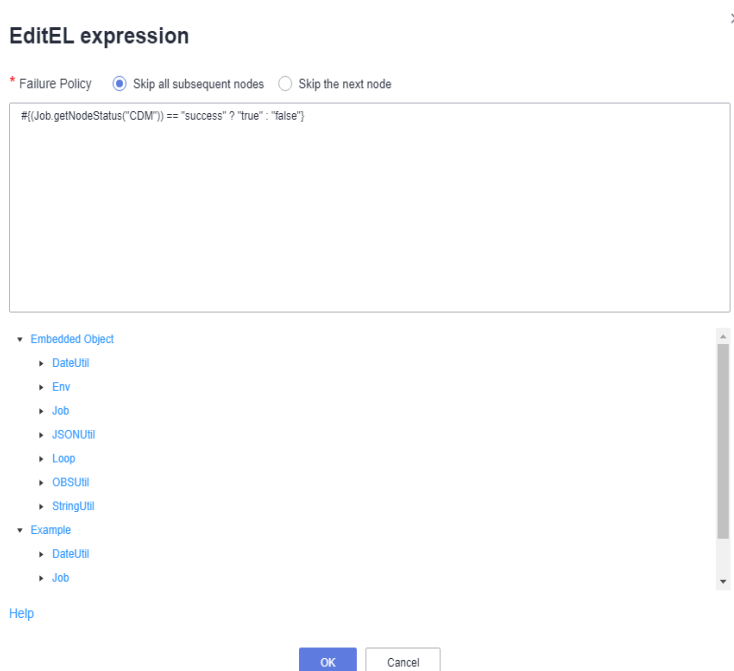
Each statement branch requires an IF statement. The IF statement is a ternary expression based on the EL expression syntax. If the result of the ternary expression is **true**, subsequent nodes will be connected. Otherwise, subsequent nodes will be skipped.

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 6-84 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 6-85 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_net	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.

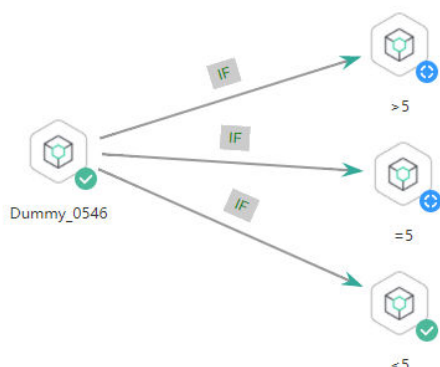
The execution result of the select statement on the HIVE SQL node is a two-dimensional array. To obtain the values in the array, use the EL expression `#{Loop.dataArray[] []}`. Currently, only the For Each node supports this expression. Therefore, you need to connect the HIVE SQL node to a For Each node. [Figure 6-86](#) shows the job orchestration.

Figure 6-86 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 6-87](#) shows the job orchestration.

Figure 6-87 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 6-87](#).
- 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 6-88 Configuring job parameters

Parameter Setup

Variable Parameter




result Enter a 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 6-86](#).

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 6-89 HIVE SQL script execution result

The screenshot shows the DataArts Studio interface. At the top, there are several icons: Save, Submit, Unlock, Lock, Execute, Format, SQL Reference, and Configure Editor. Below these is a code editor with the following content:

```
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 query on line 6 is highlighted with a red box. Below the code editor, there is a section for 'Execution History' and 'Result'. The 'Result' tab is selected, showing a table with the following data:

Row No.	count(1)
1	1

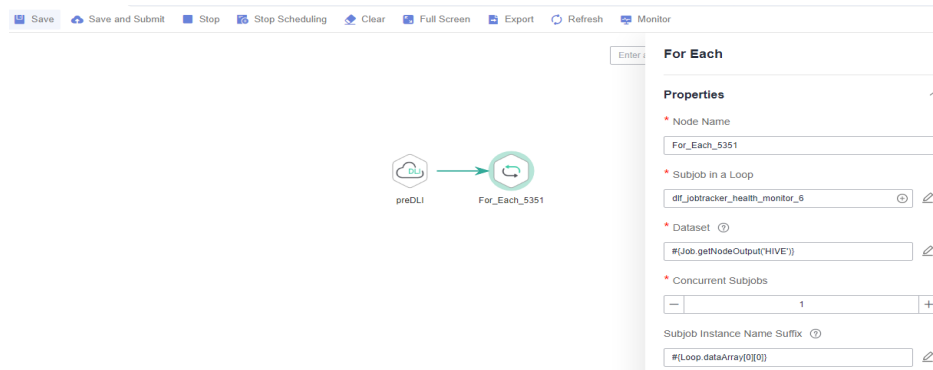
The table is also highlighted with a red box.

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 6-90 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 6-91 Execution result of the subjob

Monitor Instance

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 Reun 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 Succeed More		
Dummy_5381	Dummy	Run successfully	0.00	2022/Jan/19 14:23:59 GMT+08:00	0	--	View Log Manual Retry Succeed More		

----End

Configuring the Policy for Executing a Node with Multiple IF Statements

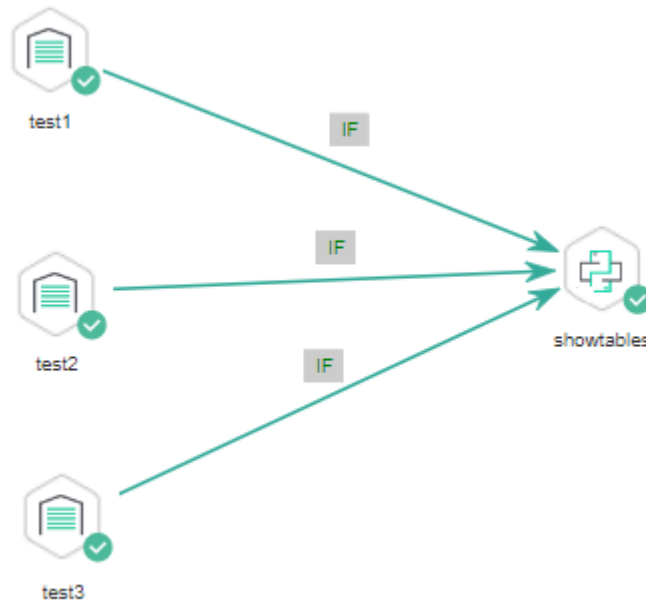
If the execution of a node depends on multiple IF statements, the policy for executing the node can be **AND** or **OR**.

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 6-92 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 6-92](#).
- 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.

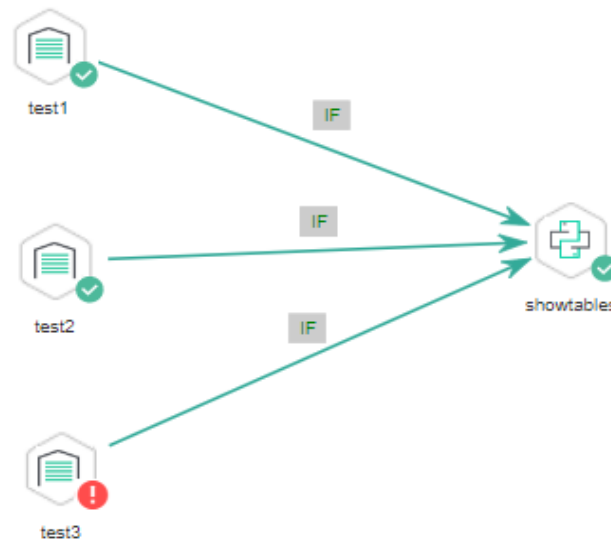
If **test1** is executed successfully, the corresponding IF statement is true.

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 6-93 How the job runs if Multi-IF Policy is OR

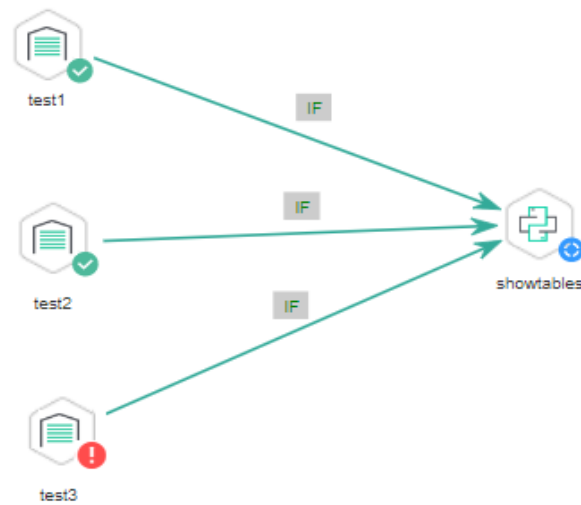


Logs

```
[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 6-94 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

6.11.3 Obtaining the Return Value of a Rest Client Node

The Rest Client node can execute RESTful requests on HUAWEI CLOUD.

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 6-95](#), the first Rest Client node invokes the API of MRS to query the cluster list. [Figure 6-96](#) 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 6-95 Rest Client job example 1

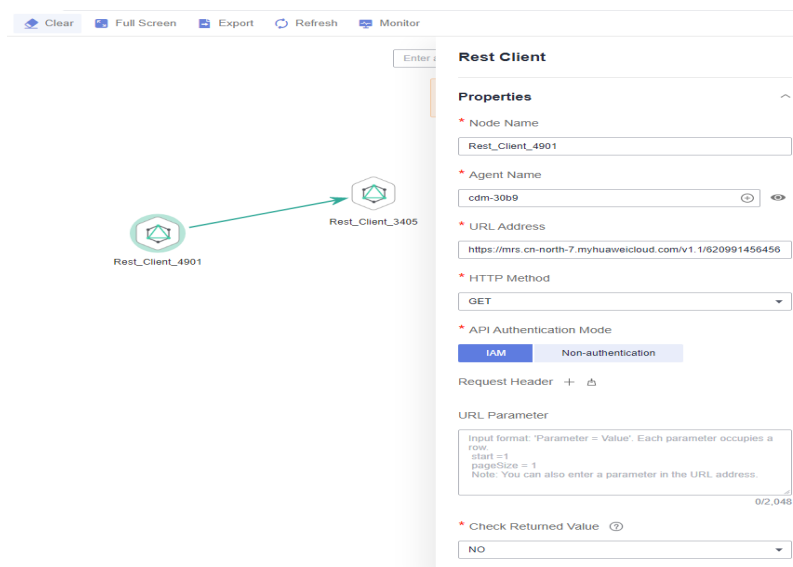


Figure 6-96 JSON message body

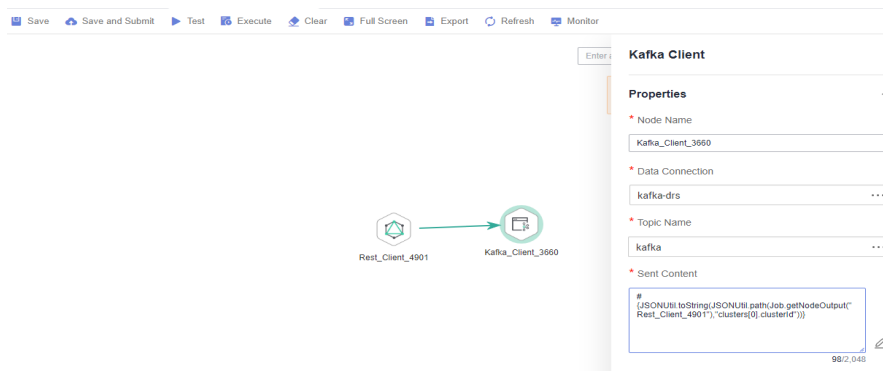


Obtaining the Return Value Using an EL Expression

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 6-97](#), 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 6-97 Rest Client job example 2

6.11.4 Using For Each Nodes

Scenario

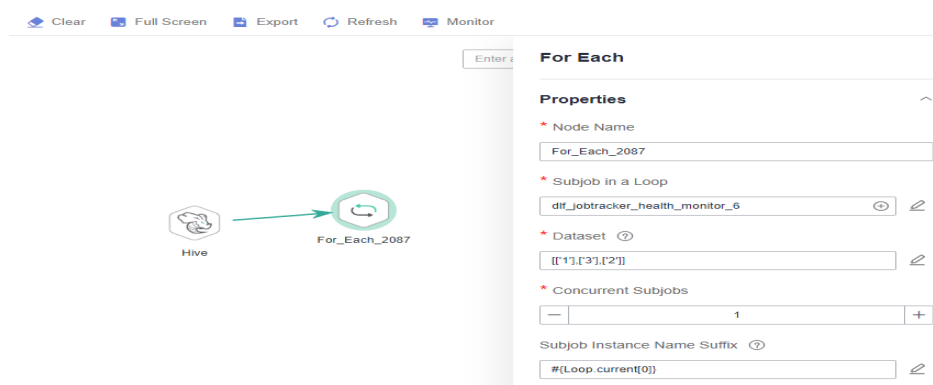
During job development, if some jobs have different parameters but the same processing logic, you can use For Each nodes to avoid repeated job development.

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 6-98 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 6-98 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 6-140 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. For details about how to create a foreign table on DLI, see **OBS Source Stream**. For details about how to create a foreign table on DWS, see **Creating a Foreign Table**.
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. For details about how to create a DLI foreign table, see **Creating a Table**.

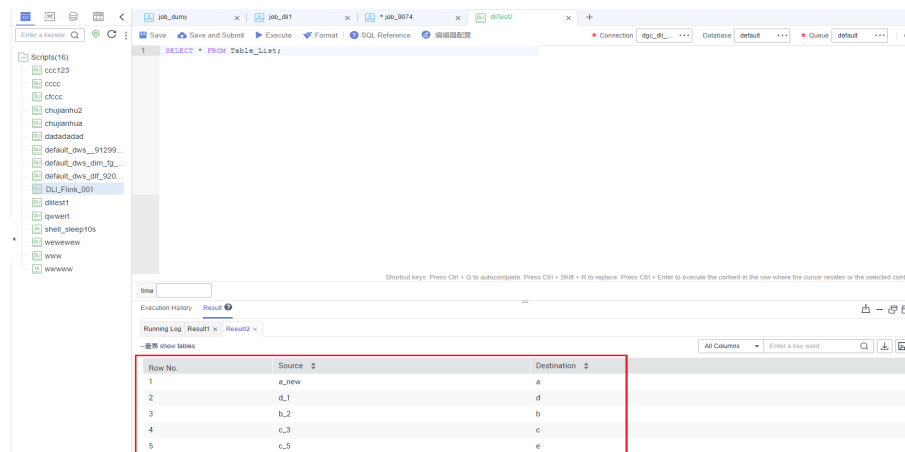
This section uses method 1 as an example to describe how to import data from **Table 1** to the DLI table (**Table_List**). You can create a DLI SQL script on the

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 6-99 Data in the Table_List table



The screenshot shows the DataArts Studio interface with a SQL editor and a results table. The SQL editor contains the query: `SELECT * FROM Table_List;`. The results table displays the following data:

Row No.	Source	Destination
1	a_new	a
2	d_1	d
3	b_2	b
4	c_3	c
5	c_5	e

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 6-100](#).

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:

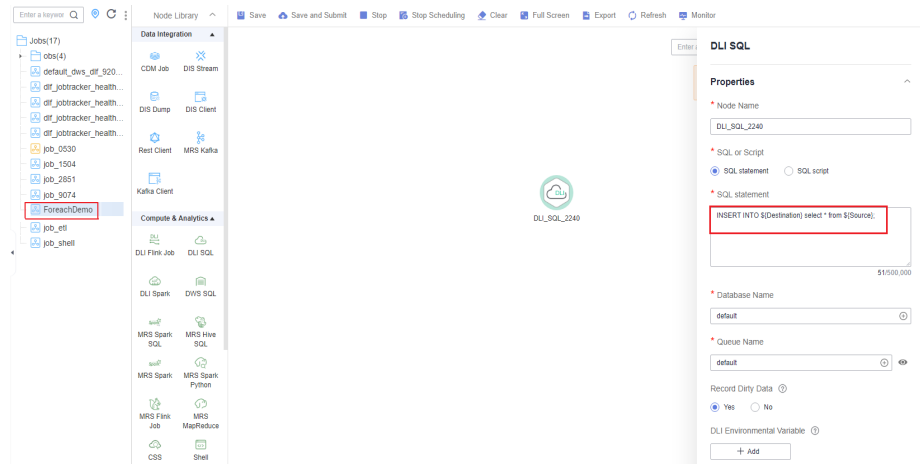
```
INSERT INTO ${Destination} select * from ${Source};
```

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.

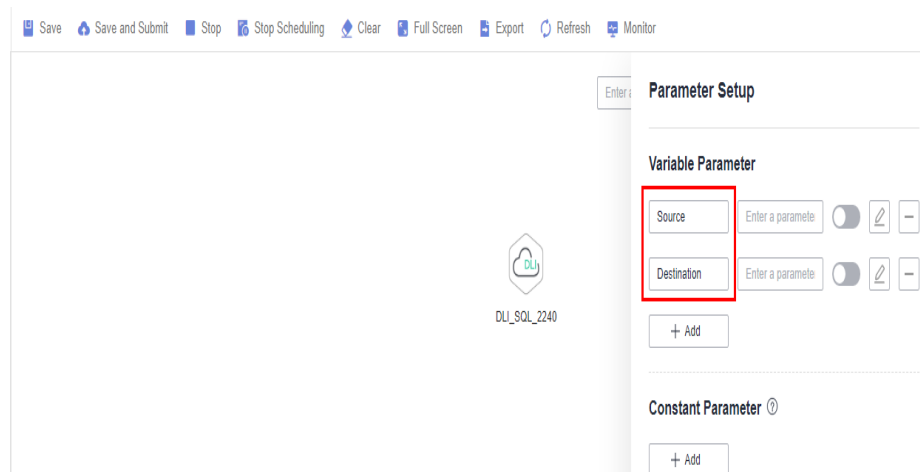
To obtain the parameter values passed from the parent job and the global variables configured for the workspace, you are advised to use the `${job_param_name}` expression.

Figure 6-100 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 6-101 Configuring subjob parameters



3. Save the job.

Step 4 Create a master job named **ForeachDemo_master** where the For Each operator is located.


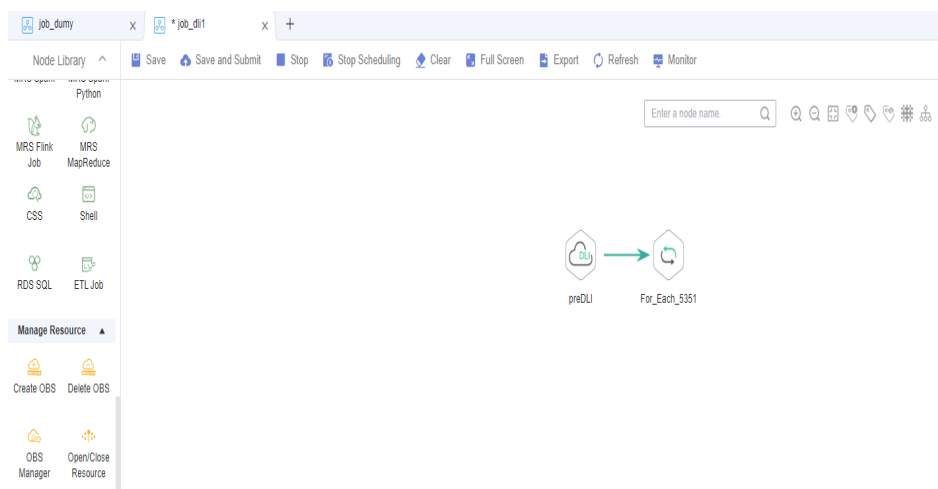
1. Access the DataArts Studio **DataArts Studio** page and choose **Develop Job**. Create a data development master job named **ForeachDemo_master**. Select the DLI SQL and For Each nodes and click and drag  to compile the job shown in [Figure 6-102](#).

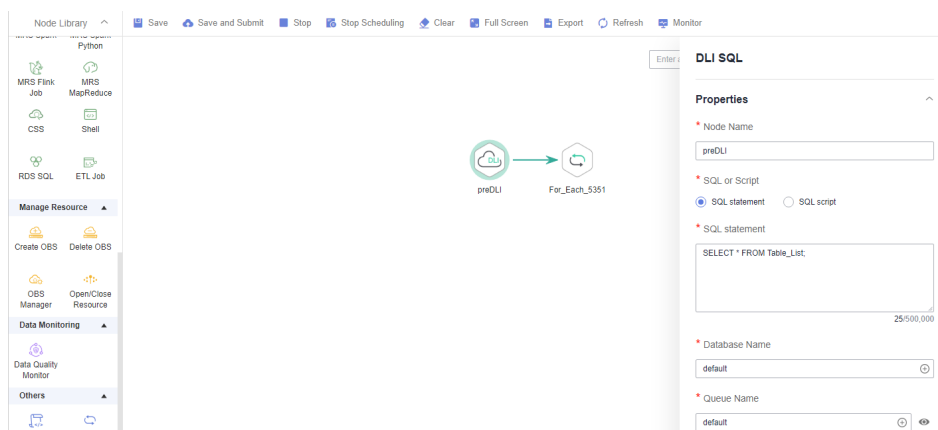
Figure 6-102 Compiling a job



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.

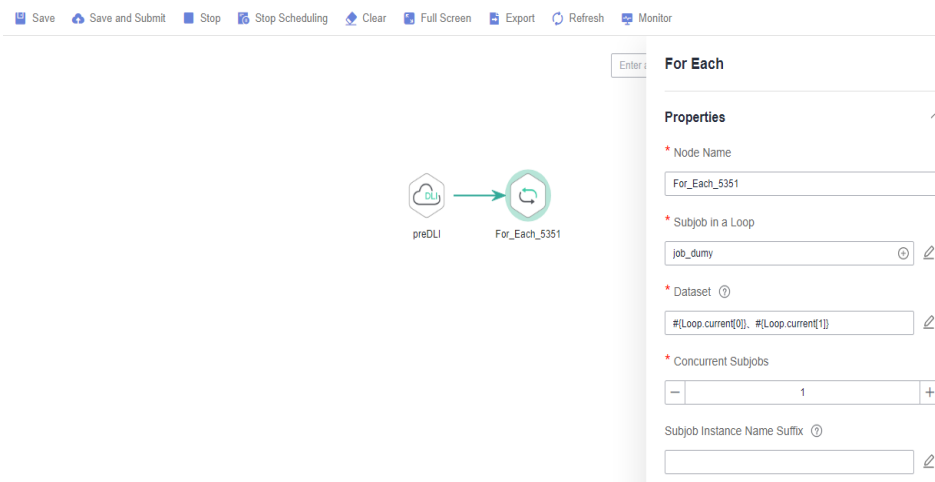
```
SELECT * FROM Table_List;
```

Figure 6-103 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 6-104 Configuring the For Each node

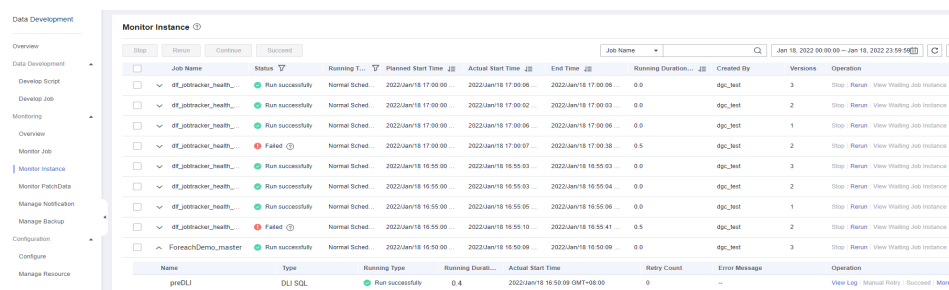


4. Save the job.

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 6-105 Viewing job instances



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 6-106 Destination table data

Row No.	name	score
1	ZHAO	90
2	QIAN	88
3	SUN	93

----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.

- [Creating Table Migration Jobs in Batches Using CDM Nodes](#)
- [Determining the IF Statement Branch to Be Executed Based on the Execution Result of the Previous Node](#)

6.11.5 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 Huawei Cloud 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-dlfpqhthn**. 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 security group rules, see [configuring security group rules](#).
 - 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.

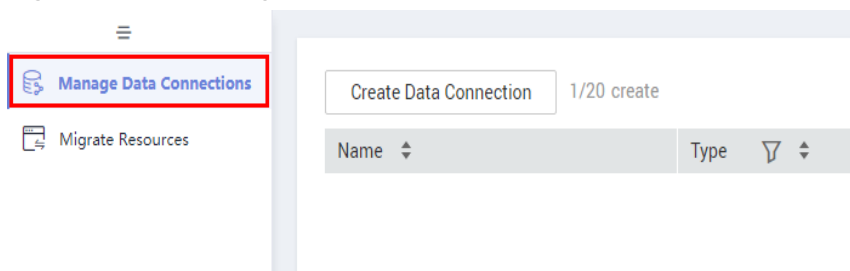
Creating an ECS Data Connection

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**.

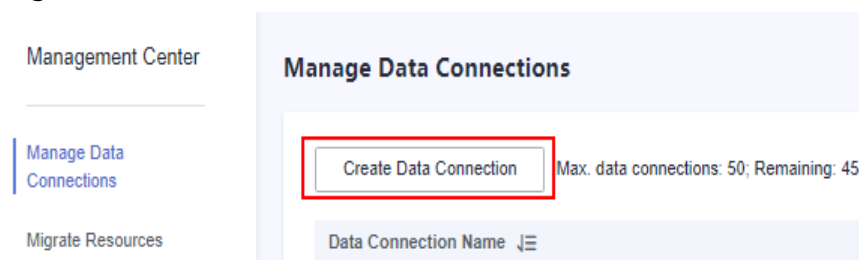
Step 2 In the navigation pane, choose **Manage Data Connections**.

Figure 6-107 Manage Data Connections



Step 3 Click **Create Data Connection**.

Figure 6-108 Create Data Connection



Step 4 Configure parameters by referring to [Table 6-141](#) and create a data connection named `python_test`.

Table 6-141 Host Connection

Parameter	Mandatory	Description
Data Connection Name	Yes	Name of the host connection. The value can contain only letters, digits, hyphens (-), and underscores (_).
Tag	No	The attribute of the data connection to create. Tags make management easier. You can set a tag or select a tag created in Tags from the drop-down list. NOTE The tag name can contain letters, digits, and underscores (_), and cannot start with underscores (_). It can contain up to 100 characters.
Host Address	Yes	IP address of the Linux host For details, see Viewing Details About an ECS .
Agent	Yes	Agents provided by the CDM cluster, which is required if Proxy connection is selected for Connection Type .
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">• Key pair• Password
Key Pair	Yes	If you select Key pair for Login Mode , you need to obtain the private key file, upload it to OBS, and select the OBS path. This parameter is available only when Login Mode is set to Key pair . NOTE 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.
KMS Key	Yes	Key created on Key Management Service (KMS) and used for encrypting and decrypting user passwords and key pairs. You can select a created key from KMS.
Host Connection Description	No	Description of the host connection

Figure 6-109 Creating a host connection

* Data Connection Type	Host Connection	
* Name	python_test	
Tag		
* Host Address		View Host
* Agent ?	cdm-suxue-test	Manage CDM Clusters
* Port	22	
* Username	root	
* Login Mode	Password	
* Password	*****	
* KMS Key ?	KMS-CDM	Access KMS
Host Connection Description	<div style="border: 1px solid #ccc; height: 50px; width: 100%;"></div> <p style="text-align: right;">0/512</p>	
<input type="button" value="Test"/>		

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 passes, the data connection is created.

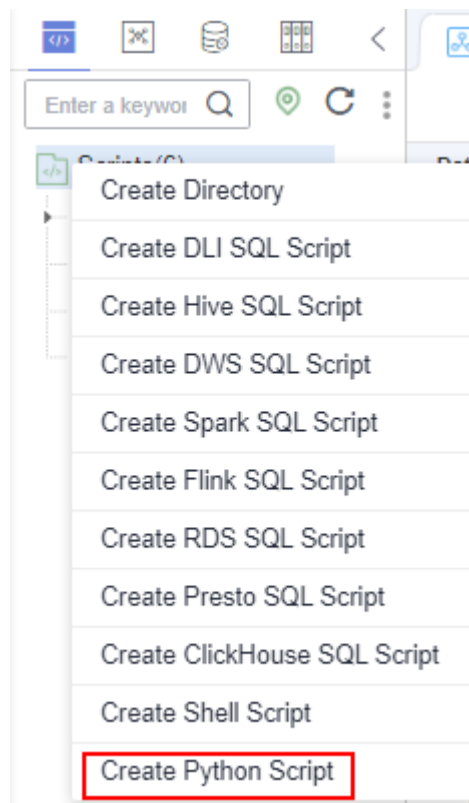
Step 6 After the test is successful, click **OK**. The system will create the data connection for you.

----End

Developing a Python Script

Step 1 Choose **DataArts Factory > Develop Script** and create a Python script named **python_test**.

Figure 6-110 Creating a Python script



Step 2 Edit the Python statement in the editor, select the host connection, and click **Submit** and **Unlock**.

NOTE

- This example defines a string template for saving company information and uses the template to output information about different companies.

```
template='No.:{:0>9s} \t CompanyName:{:s} \t Website:https://www.{:s}.com'  
context1=template.format('1','CompanyXXX','companyxxx')  
context2=template.format('2','CompanyYYY','companyyyy')  
print(context1)  
print(context2)
```
- The script development area in [Figure 6-111](#) is a temporary debugging area. After you close the script tab, the development area will be cleared.
- **Connection:** Select the data connection created in [Creating an ECS Data Connection](#).

Figure 6-111 Editing the Python statement

```
1 template='No.:{:0>9s} \t CompanyName:{:s} \t Website:https://www.{:s}.com'  
2 context1=template.format('1','CompanyXXX','companyxxx')  
3 context2=template.format('2','CompanyYYY','companyyyy')  
4 print(context1)  
5 print(context2)  
6
```

Step 3 Click **Execute** to execute the Python statement.

Step 4 View the script execution result.

----End

6.11.6 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:

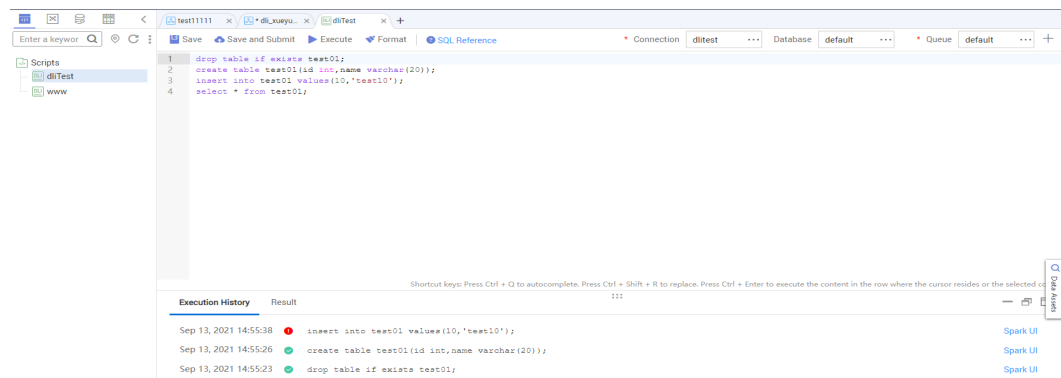
```
create schema store_sales;
set current_schema= store_sales;
drop table if exists trade_log;
CREATE TABLE trade_log
(
    sn          VARCHAR(16),
    trade_time  DATE,
    trade_count INTEGER(8)
);
set current_schema= store_sales;
drop table if exists trade_report;
```

```
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 6-112 Developing a script



Key notes:

- The script development area in **Figure 6-112** 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 6-113 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 6-114 Configuring properties for the DWS SQL node

DWS SQL

Properties ∨

SQL or Script *
 SQL statement SQL script

SQL script *
 ⋮ + ✎

Data Connection *
 ⋮ 👁

Database *
 ⋮

Script Parameter ↻

Dirty Data Table

Matching Rule ?

Failure Matching Value ?

Node Name *


⊞ **Advanced Settings** ∨

Data Assets

Key properties:

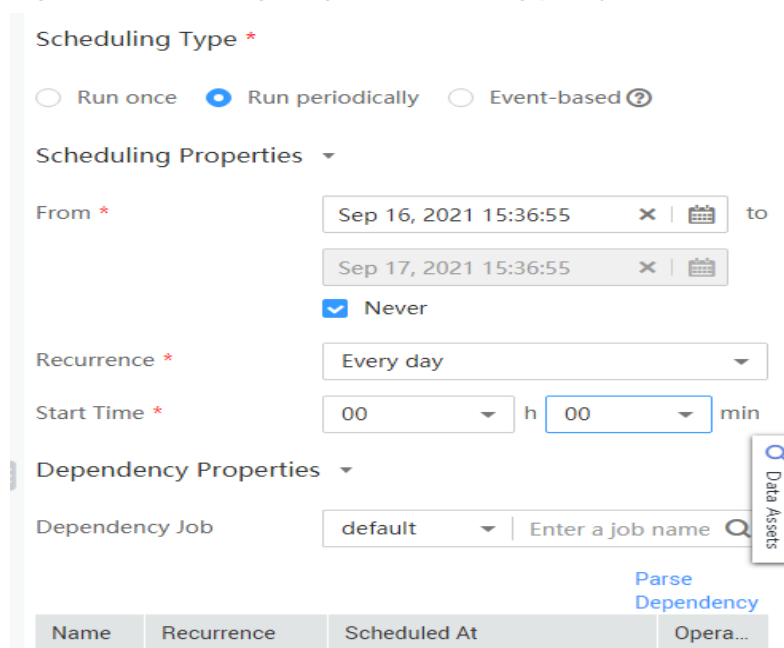
- **SQL script:** Associate with the **dws_sql** script developed in [Developing a DWS SQL Script](#).
- **Data Connection:** Select the data connection configured in the **dws_sql** script. The data connection can be changed.
- **Database:** Select the database configured in the **dws_sql** script. 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 6-115 Configuring the scheduling policy



Scheduling Type *

Run once Run periodically Event-based ?

Scheduling Properties ▾

From * × | × |

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

6.11.7 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**.

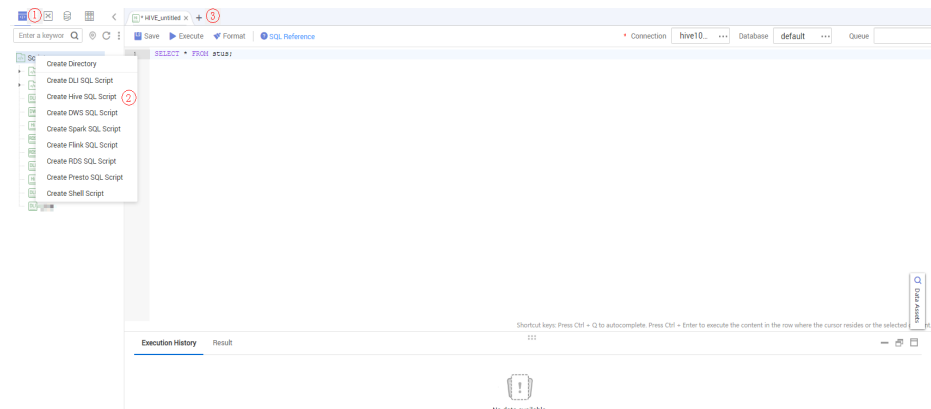
Description of key parameters:

- **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 6-116 Developing a script



Notes:

- The script development area in [Figure 6-116](#) 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 6-117 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

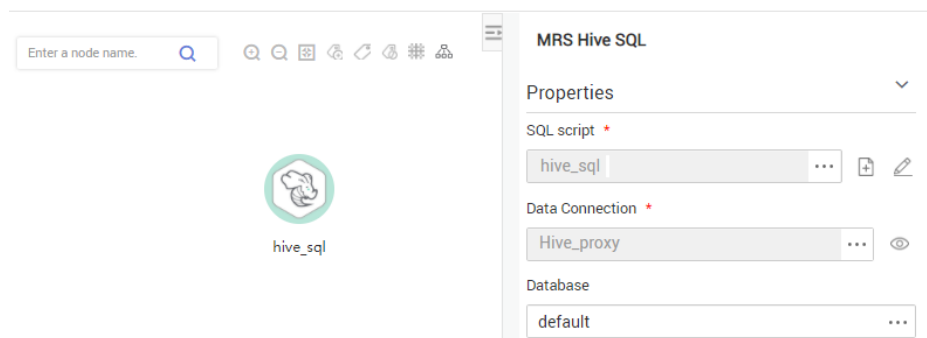
Agency ? +

* 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 6-118 Configuring properties for an MRS Hive SQL node



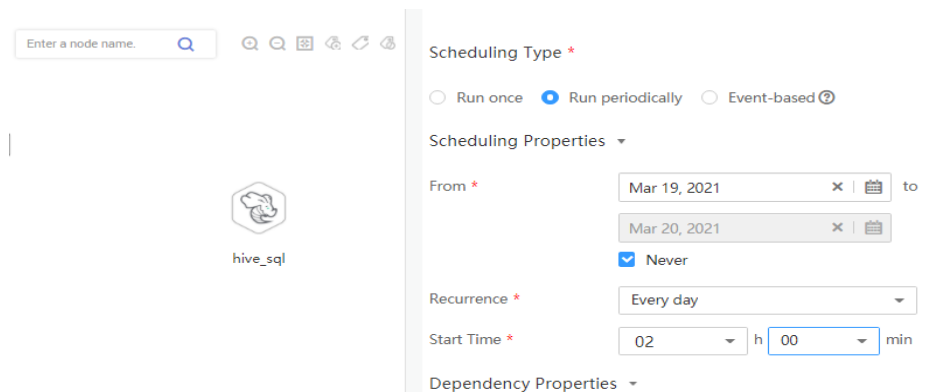
Description of key properties:

- 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 6-119 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

6.11.8 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.

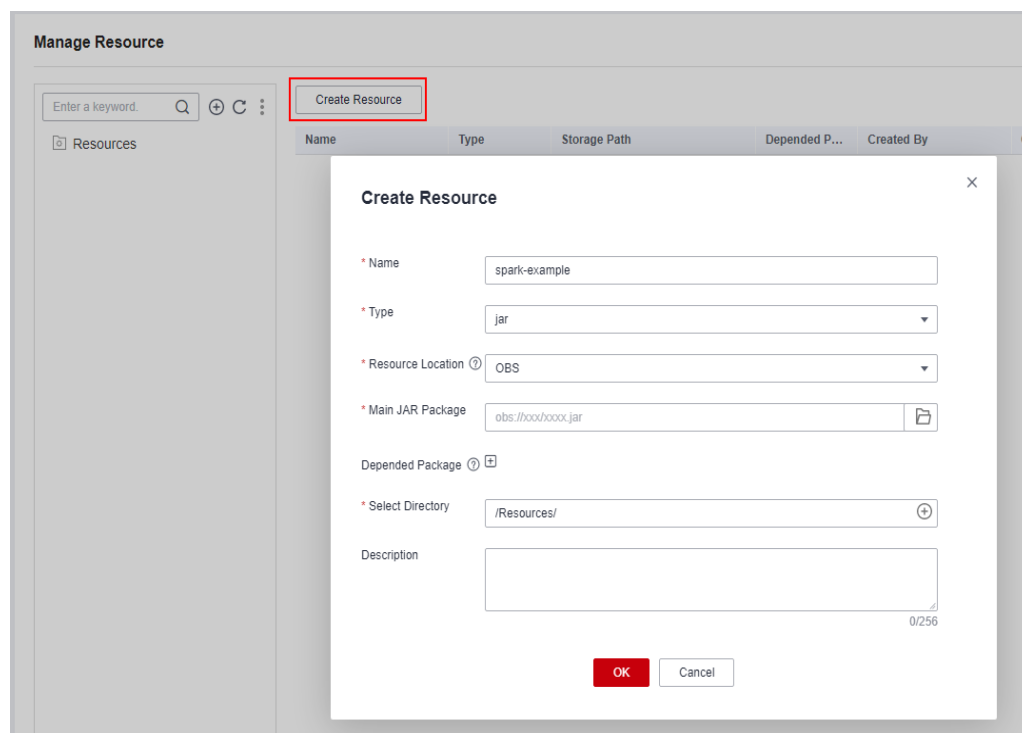
Preparations

- 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. This Spark job is to calculate the approximate value of π .

- 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**.
- 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 6-120 Creating a resource

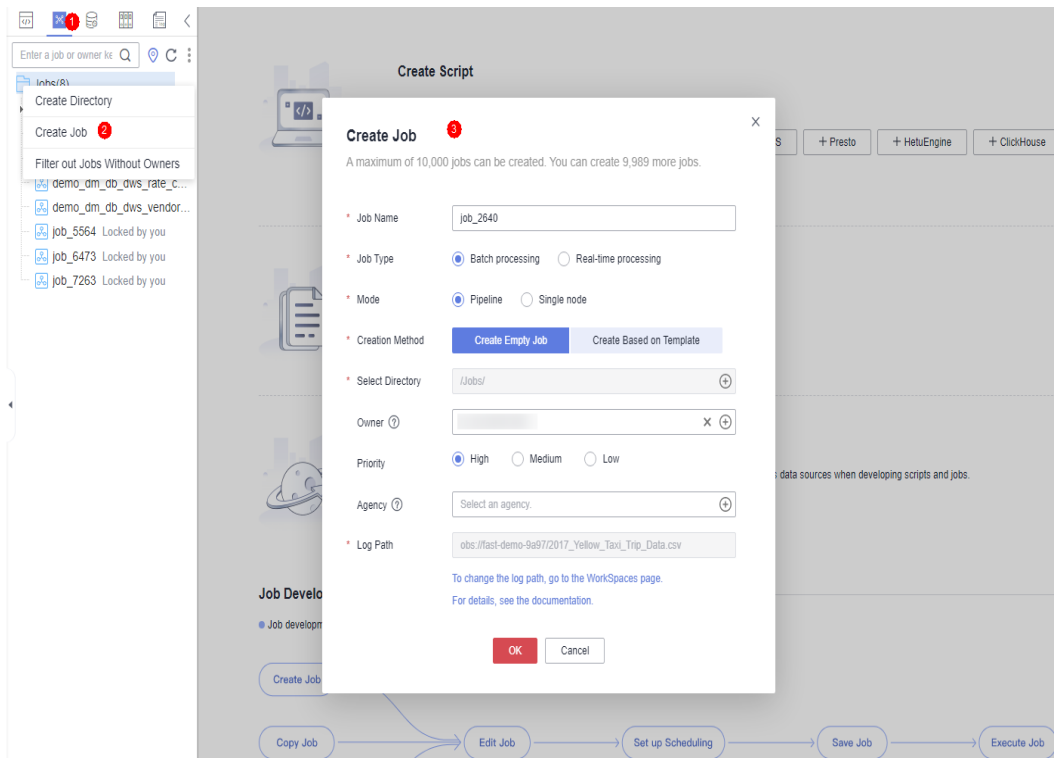
----End

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.

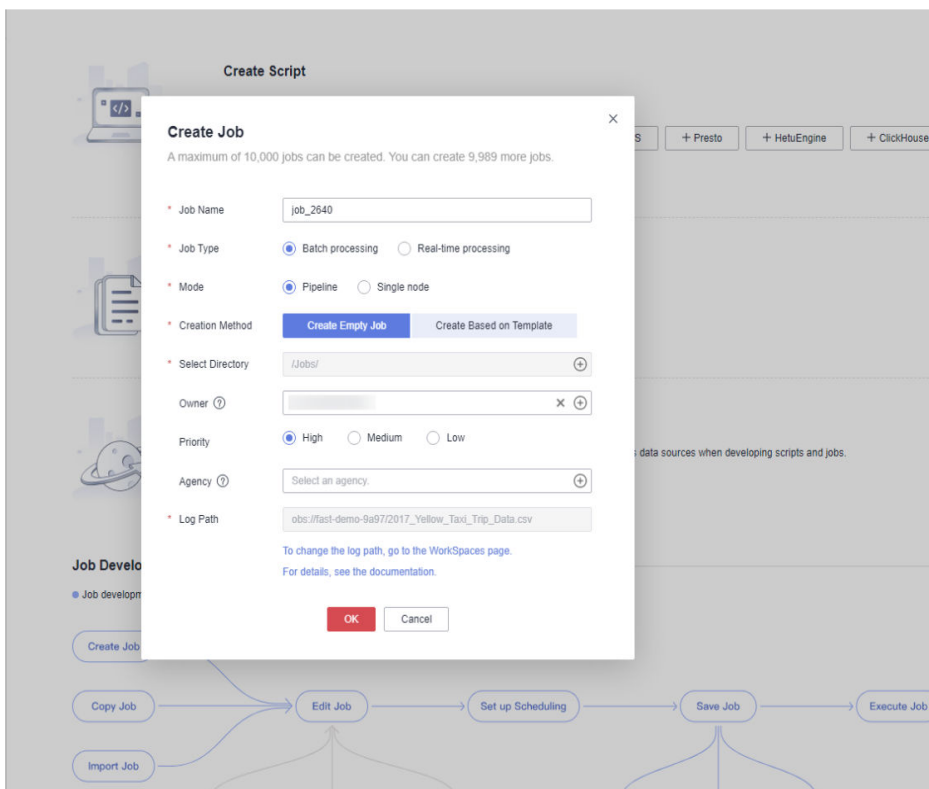
Step 1 Create a job named **job_DLI_Spark** for the DataArts Factory module.

Figure 6-121 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 6-122 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](#).


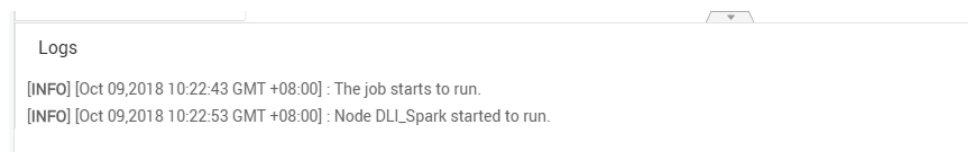
Step 3 After the job orchestration is complete, click  to test the job.

Figure 6-123 Job logs (for reference only)



Step 4 If no error is recorded in logs, save and submit the job.

----End

6.11.9 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 6-124 Creating a job

✕

Create Job

A maximum of 10,000 jobs can be created. You can create 9,999 more jobs.

* Job Name

* Job Type Batch processing Real-time processing

* Creation Method Create Empty Job Create Based on Template

* Select Directory +

Owner ? +

Priority High Medium Low

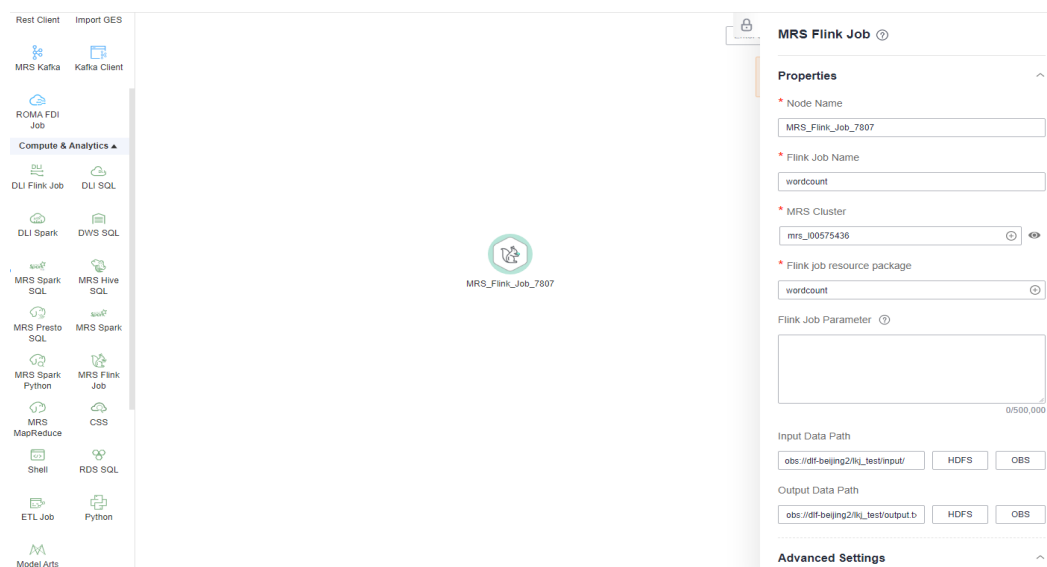
Agency ? +

* Log Path

I agree to create OBS bucket `obs://dlf-log-0621c35ef30026c92f76c005e72fd0f8/`. This bucket is used only for storing run logs of DLF jobs.
[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 6-125 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-test/lkj_test/input/word.txt
--Output data path
obs://dlf-test/lkj_test/output.txt
```

Specifically:

obs://dlf-test/lkj_test/input/word.txt is the directory where the **wordcount.jar** parameters are passed. You can pass the words to count.

obs://dlf-test/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

6.11.10 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

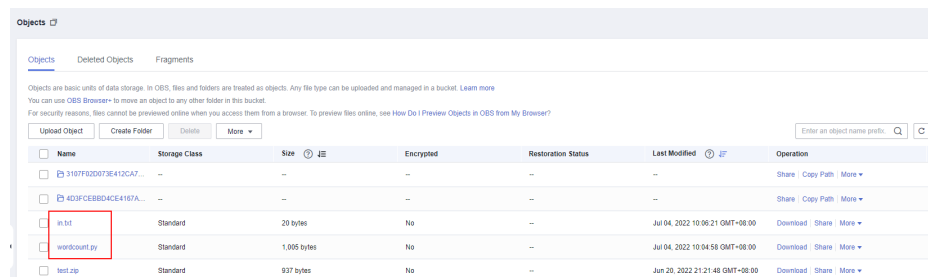
The encoding format must be set to UTF-8. Otherwise, an error will occur during script execution.

- 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 6-126 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 6-127 Creating a job

Create Job ×

A maximum of 10,000 jobs can be created. You can create 9,999 more jobs.

* Job Name

* Job Type Batch processing Real-time processing

* Creation Method

* Select Directory +

Owner +

Priority High Medium Low

Agency +

* Log Path

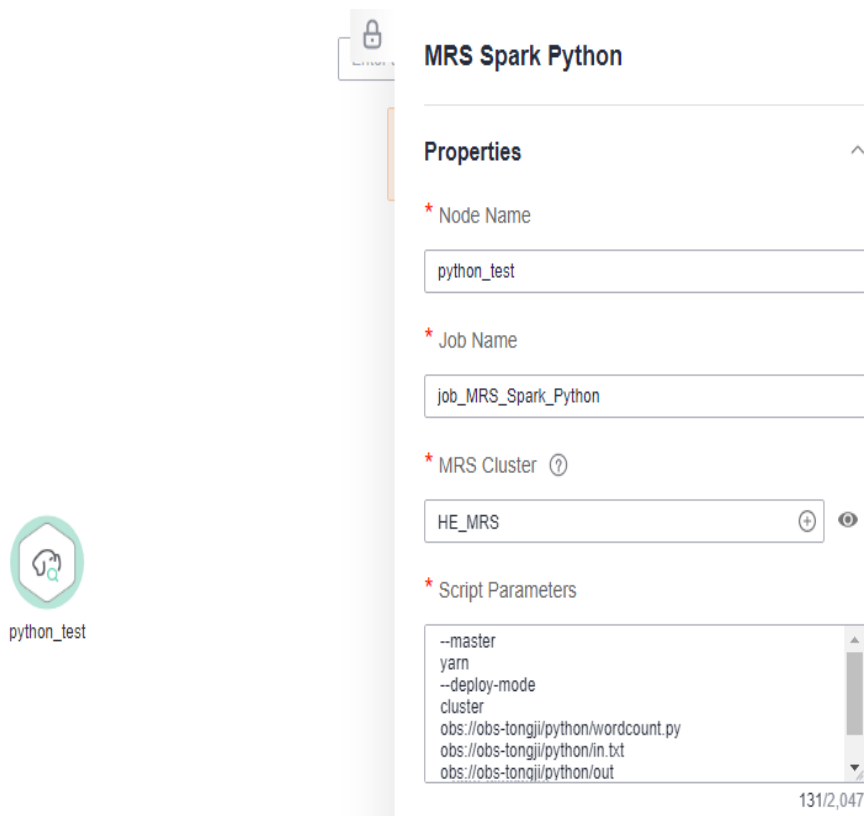
I agree to create OBS bucket `obs://dlf-log-0621c35ef30026c92f76c005e72fd0f8/`. This bucket is used only for storing run logs of DLF jobs.

[To change the log path, go to the WorkSpaces page.](#)

[For details, see the documentation.](#)

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 6-128 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.

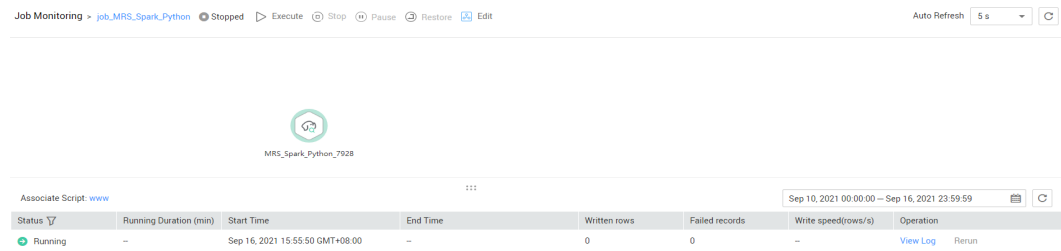
obs://obs-tongji/python/out is the directory where output parameters are stored. This directory will also be created in the OBS bucket automatically. If the **out** directory already exists in the OBS bucket, an error will occur.

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 6-129 Viewing the job execution result



The job log shows that the job was successfully executed.

Figure 6-130 Job run logs

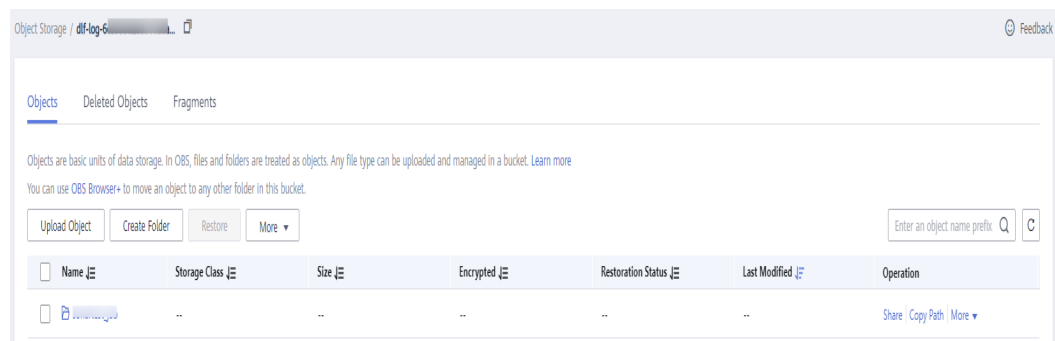


Figure 6-131 Job execution status



Step 7 View the returned records in the OBS bucket. (Skip this step if the return function is not configured.)

Figure 6-132 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.

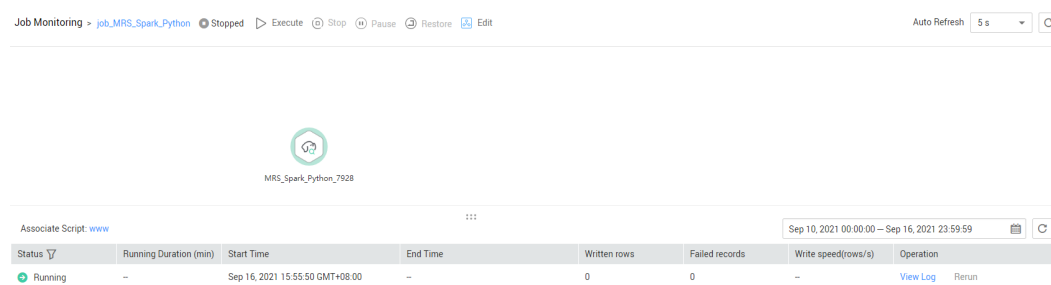
Parameter descriptions:

```
--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 6-133 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 6-134 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/codes.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

6.11.11 More Cases for Reference

For more advanced guidance and cases of DataArts Factory, see [Best Practices](#).

7 DataArts Quality

7.1 Monitoring Business Metrics

7.1.1 Overview

The Metric Monitoring module manages business metrics.

To monitor a business metric, customize a SQL metric, define a rule based on the logical expression of the metric, and create and run a business scenario. Based on the running result of the business scenario, you can determine whether the business metric meets the quality rule. The running result of the business scenario may be any of the following:

- **Normal:** The instance stops normally and the running result meets the expectation.
- **Alarming:** The instance stops normally, but the running result does not meet the expectation.
- **Abnormal:** The instance stops unexpectedly.
- **--:** The instance is running, but no running result is displayed.

The following table describes modules under **Quality Monitoring**.

Function	Description
Dashboard	Default homepage. This page contains the following parts: <ul style="list-style-type: none">• Quick Start that demonstrates how you can use metric monitoring• Running and alarm statuses for the business scenario instance over the last seven days• Alarms, scenarios, and metrics in different time periods
Metrics	You can create metrics on this page.
Rules	You can create rules based on the logical expressions of metrics on this page.

Function	Description
Scenarios	A business scenario can be considered as a business metric quality job. On this page, you can schedule and run a created rule group.
O&M	You can view the running statuses of business scenario instances and handle O&M issues. The Subscriptions page displays the running statuses of all the tasks you have subscribed to.

7.1.2 Creating a Metric

You can manage all business metrics, including the metric sources and definitions. Business metrics are stored in directories.

Metrics in DataArts Quality are independent of business metrics and technical metrics in DataArts Architecture.

Prerequisites

You have created a home directory. To create a home directory, log in to the DataArts Studio console, click **Access** under a specific instance, choose **More > DataArts Quality** on the **Workspaces** tab page, choose **Metric Monitoring > Metric Management** from the left navigation bar on the page displayed, and create a directory. Before creating a metric for a data connection, select a directory to store the metric. For details, see [Figure 7-1](#).

Figure 7-1 Directory that stores the metric to create

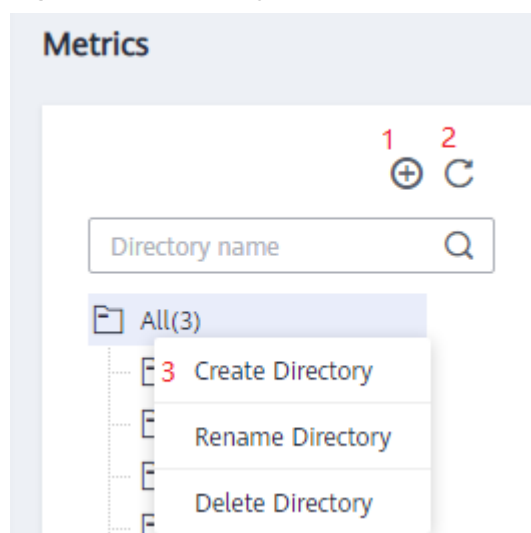


Table 7-1 Buttons in the navigation bar

No.	Description
1	Create Directory
2	Refresh Directory

No.	Description
3	Right-click All to create, rename, or delete a directory.

Creating a Metric

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Quality**.
2. Choose **Metric Monitoring > Metrics** from the left navigation bar.
3. Click **Create**. In the dialog box displayed, set the parameters based on [Table 7-2](#).

Table 7-2 Metric parameters

Parameter	Description
Metric Name	The name of a metric, which contains 1 to 64 characters and consists of only letters, numbers, and underscores (_).
Data Connection	Select a created data connection from the drop-down list box. NOTE <ul style="list-style-type: none">• Currently, only DWS, PostgreSQL, MRS Hive, DLI, and MySQL are supported.• Metrics are closely connected based on data connections. Therefore, you must establish data connections in the metadata management module before creating metrics.
Database/Queue	Select the database where the metric runs. NOTE If DLI is selected as the data connection, a running queue is required.
Description	Information to better identify a metric. It cannot exceed 4096 characters.
Directory	Directory for storing metrics. You can select a created directory. Figure 7-1 shows the directory.
Metric Type	Custom is supported. You can customize an SQL statement to define the metric source.

7.1.3 Creating a Rule

You can manage all rules that define relationships between metrics or between metrics and values. Rules are stored in directories.

Prerequisites

You have created a home directory. To create a home directory, log in to the DataArts Studio console, click **Access** under a specific instance, choose **More >**

DataArts Quality on the **Workspaces** tab page, choose **Metric Monitoring** > **Rule Management** from the left navigation bar on the page displayed, and create a directory. Before creating a rule for a metric, select a directory to store the rule. For details, see [Figure 7-2](#).

Figure 7-2 Directory that stores the rule to create

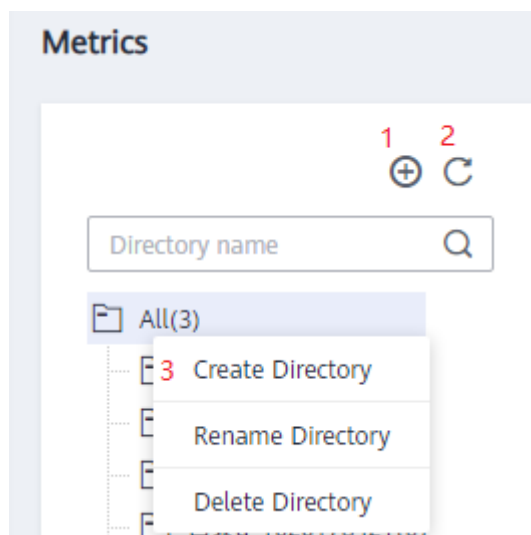


Table 7-3 Buttons in the navigation bar

No.	Description
1	Create Directory
2	Refresh Directory
3	Right-click All to create, rename, or delete a directory.

Creating a Rule

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Quality**.
2. Choose **Metric Monitoring** > **Rule Management** from the left navigation bar.
3. Click **Create**. In the dialog box displayed, set the parameters based on [Table 7-4](#).

Table 7-4 Rule parameters

Parameter	Description
Rule Name	The name of a rule, which contains 1 to 64 characters and consists of only letters, numbers, and underscores (_).
Description	Information to better identify a rule. It cannot exceed 4096 characters.

Parameter	Description
Directory	The directory that stores the rule. You can select a created directory. Figure 7-2 shows the directory.
Define Relationship	A relationship is a logical expression between a metric and a value or between metrics. The relationship can contain arithmetic operations. Metrics are abbreviated to lowercase letters a to z and are added in the alphabetic order of metric abbreviations. NOTE Only one valid logical expression and the simple four arithmetic operations are supported.

7.1.4 Creating a Scenario

You can manage all scenarios that define the logical relationships between rules. Scenarios are stored in directories.

Prerequisites

You have created a home directory. To create a home directory, log in to the DataArts Studio console, click **Access** under a specific instance, choose **More > DataArts Quality** on the **Workspaces** tab page, choose **Metric Monitoring > Business Scenario Management** from the left navigation bar on the page displayed, and create a directory. Before creating a scenario for rules, select a directory to store the scenario. For details, see **Figure 7-3**.

Figure 7-3 Directory that stores a scenario

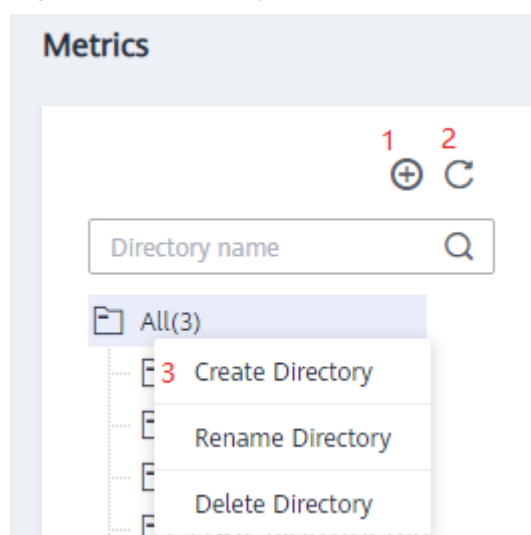


Table 7-5 Buttons in the navigation bar




No.	Description
1	Create Directory

No.	Description
2	Refresh Directory
3	Right-click All to create, rename, or delete a directory.

Creating a Scenario

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Quality**.
2. Choose **Metric Monitoring > Business Scenario Management** from the left navigation bar.
3. Click **Create**. In the dialog box displayed, set the parameters based on [Table 7-6](#).

Table 7-6 Scenario parameters

Parameter	Description
Basic Configuration	
Scenario Name	The name of a scenario, which contains 1 to 64 characters and only consists of letters, numbers, and underscores (_).
Description	Information to better identify a scenario. It cannot exceed 256 characters.
Directory	The directory that stores the scenario. You can select a created directory. Figure 7-3 shows the directory.
Business Level	The options are Warning , Minor , Major , and Critical . The business level determines the template for sending notification messages.
Rule Group Configuration	
Define Rule Group	Group of rules. Logical expressions are used between rules.
Rule A	You can select a rule from the drop-down list. You can also click  to add multiple rules.
Subscription Configuration	
Notification	Set this to  or  to enable or disable the notification function.
Notification Type	The options are as follows: <ul style="list-style-type: none">• Trigger alarms• Run successfully

Parameter	Description
Topic	Select a message notification topic.

- Click **Next** to go to the page where you can select a scheduling mode. Currently, **Schedule once** and **Schedule periodically** are supported. Set parameters for scheduling periodically by referring to [Table 7-7](#).

Table 7-7 Scheduling parameters

Parameter	Description
Effective	The period during which a scheduling task takes effect.
Scheduling Cycle	The frequency at which a scheduling task is executed. Related parameters are: <ul style="list-style-type: none"> • Minute • Hour • Day • Week
Time Interval	Interval for two consecutive scheduling tasks.
Start from	Start time and end time of the scheduling task

7.1.5 Viewing a Scenario Instance

You can manage all scenarios, view metric running statuses, query run logs, and handle issues on the **O&M Management** page.

GUI Description

The following figure shows the areas and buttons on the **O&M** page.

Figure 7-4 O&M Management page

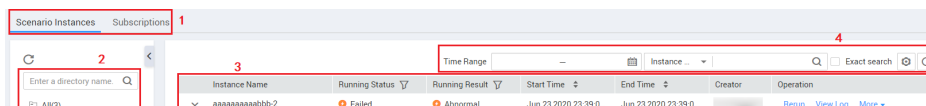


Table 7-8 Entry

No.	Area	Description
1	Menu bar	<p>The menu bar on the O&M Management page includes Scenario Instances and Subscriptions.</p> <ul style="list-style-type: none"> • The Scenario Instances tab page lists all scenario instances that you have created. • The Subscriptions tab page lists all scenarios that you have subscribed. Notification Status is available only on the Subscriptions tab page. Notification Status indicates whether the running result of a scenario instance is subscribed to, for example, sending an alarm email.
2	Navigation bar	<p>Contains the directories that store scenario instances. You can store scenarios in different directories. The number next to each directory indicates the number of scenarios stored in that directory.</p>
3	List of scenario instances	<p>Displays the instance name, running status, and running result.</p>
4	Search area	<ul style="list-style-type: none"> • Displays scenario instances selectively. For example, you can display scenario instances for a specified time range. • Displays a list of instances according to the handler, creator, or instance name. Fuzzy search is supported.

Table 7-9 Scenario instance parameters

Parameter	Description
Running Status	<p>Displays the running status of a scenario instance.</p> <ul style="list-style-type: none"> • Successful: The instance is successfully executed. • Failed: The instance fails to run. • Running: The instance is running.
Running Result	<p>Displays whether the scenario instance is running properly.</p> <ul style="list-style-type: none"> • Normal: The instance stops normally and the running result meets the expectation. • Alarming: The instance stops normally, but the running result does not meet the expectation. • Abnormal: The instance stops unexpectedly. • --: The instance is running, but no running result is displayed.
Rerun	<p>Allows you to run the scenario instance again.</p>
View Log	<p>Allows you to view the running details of the scenario instance.</p>

Parameter	Description
More > Resolve Issue	Allows you to perform further processing on the scenario instance. You can Provide handling suggestions , Close the issue , or Transfer to others . The above operations can be performed only when you are the handler of the instance.
More > View Processing Log	Allows you to view historical processing records.

7.2 Monitoring Data Quality

7.2.1 Overview

DataArts Quality is a type of quality management tool used to manage the quality of data in databases. You can filter out unqualified data in a single column or across columns, rows, and tables from the following perspectives: integrity, validity, timeliness, consistency, accuracy, and uniqueness. DataArts Quality can monitor offline data. When offline data changes, DataArts Quality verifies the data and blocks the production link to avoid the spread of the problem data. DataArts Quality also manages historical verification results so that you can analyze and grade data quality.

It can also automatically generate standardized quality rules based on the data standards in DataArts Architecture, and periodically monitor data.

The following table describes modules under **Quality Monitoring**.

Module	Description
Dashboard	The dashboard is the homepage that displays alarming and blocking information of tables. The following information is included: <ul style="list-style-type: none">• Number of jobs, instances, and anomaly tables; distributions and changes of instance running statuses in a selected period.• Statistics about alarm classifications and table alarms of the current day, as well as the alarm trend and rule quantity of the latest seven days.
Rule Template	Rule template is a major function of DataArts Quality. You can configure rules on the Rule Template page. It mainly manages functions related to rule configuration and provides built-in and custom templates.
Quality Job	Quality jobs can apply rule templates or custom rules to tables for data monitoring.

Module	Description
Comparison Job	You can create comparison jobs to apply the created rules to two existing tables to monitor their data and output the comparison results.
O&M Management	You can view the running status of rules and handle O&M problems.
Quality Report	The system automatically generates quality reports based on the job execution result.

7.2.2 Creating Rule Templates

DataArts Quality can monitor offline data, in which quality rules play a vital role. There are 26 built-in rule templates, such as database-level, table-level, field-level, and cross-field rule templates.

Table 7-10 System built-in rule templates

Rule Type	Dimension	Template	Description
Database-level	Integrity	Database null value scan	Calculates the number of rows in all the tables of a database in which all fields have a null value.
Table-level	Accuracy	Table rows	Calculates the number of rows in a data table.
	Integrity	Data table null value scan	Calculates the number of rows in a table in which all fields have a null value.
Field-level	Uniqueness	Field with a unique value	Calculates the number of rows in a data table in which a specified field has a unique value.
		Field with duplicate values	Calculates the number of rows in a data table in which a specified field has duplicate values.
		Unique combination of multiple fields	Checks whether the combination of multiple fields in a DWS table is unique. A maximum of 10 fields can be combined.
	Integrity	Field with a null value	Calculates the number of rows in a data table in which a specified field has a null value.
	Accuracy	Average field value	Calculates the average value of a specified field in a data table.

Rule Type	Dimension	Template	Description
		Total field values	Calculates the total values of a specified field in a data table.
		Maximum field value	Calculates the maximum value of a specified field in a data table.
		Minimum field value	Calculates the minimum value of a specified field in a data table.
		Field length verification	Checks whether the length of a field in the DWS table is within the allowed range.
		Field value range verification	Checks whether the value of a field in the DWS table is within the allowed range.
		Field time verification	Checks whether the time of a field in the DWS table is within the allowed range. Currently, only fields of the date and timestamp types are supported. Fields of the time type is not supported.
	Effectiveness	ID card verification	Verifies the validity of a specified field in a data table based on built-in regular expression rules.
		Mailbox verification	Verifies the validity of a specified field in a data table based on built-in regular expression rules.
		Regular expression verification	Verifies the validity of a specified field in a data table based on a custom regular expression.
		IP address verification	Verifies the validity of a specified field in a data table based on built-in regular expression rules.
		Phone number format verification	Verifies the validity of a specified field in a data table based on built-in regular expression rules.
		Postal code format verification	Verifies the validity of a specified field in a data table based on built-in regular expression rules.
		Date format verification	Verifies the validity of a specified field in a data table based on built-in regular expression rules.

Rule Type	Dimension	Template	Description
		Validity verification	Verifies the validity of a specified field in a data table based on a custom regular expression.
		Enumerated value verification	Verifies the validity of a specified field in a data table based on a custom enumerated value.
		Ignoring of null values in enumerated value verification	Verifies the validity of a specified field in a data table based on a custom enumerated value. Null values are counted in valid rows.
Cross-field level	Consistency	Field consistency verification	Checks consistency between different fields from the same data source.
	Accuracy	Cross-field time verification	Checks whether the time relationship between a specified field in the data table and the reference field meets the expectation. Currently, only fields of the date and timestamp types are supported. Fields of the time type is not supported.

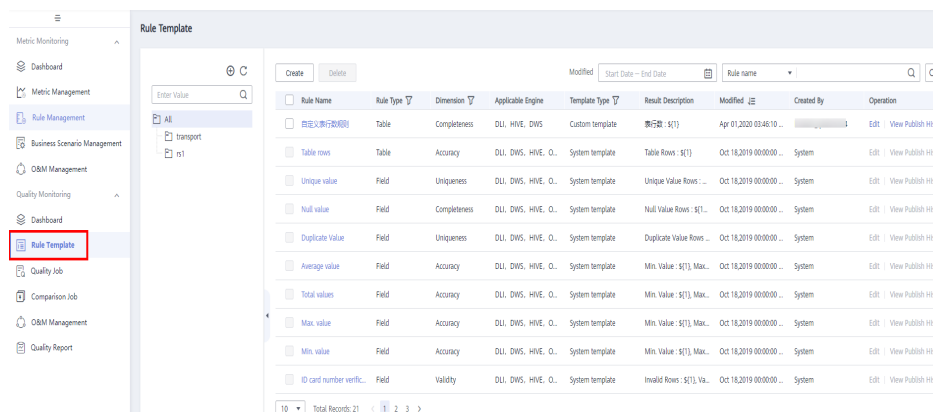
If the built-in rule templates do not meet your requirements, you can create rules in either of the following ways:

- Custom template: Choose **Quality Monitoring > Rule Templates** and click **Create**. The created rule template will be automatically classified into the corresponding rule type and displayed as a custom template. A quality job using a custom template does not support exceptional data output and quality scoring.
- Custom rule: When creating a quality job, set **Rule Type** to **Custom rule** and enter an SQL statement to define how to monitor the quality of data objects.

This section describes how to create a rule using a custom template. For details about how to create a custom rule, see [Creating Quality Jobs](#).

Step 1 Choose **Quality Monitoring > Rule Template**, and click **Create** on the page displayed.

Figure 7-5 Rule Template page



Step 2 In the dialog box displayed, enter the rule template name, select the rule matching dimension, define the SQL template, and describe the output result.

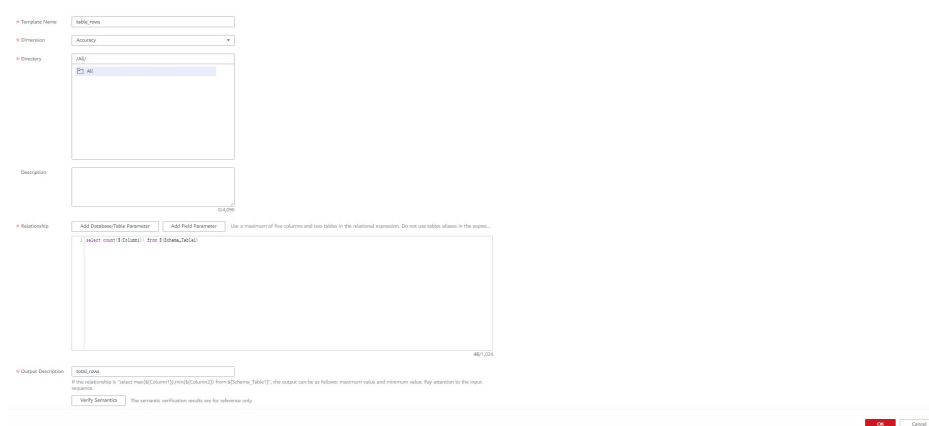
- **Dimension:** You can complete single-column, cross-column, cross-row, and cross-table analysis from six dimensions: completeness, validity, timeliness, consistency, accuracy, and uniqueness. When customizing a quality rule, select a dimension for rule matching.
- **Directory:** Select the directory where the rule is located.
- **Define Relationship:** Enter SQL statements to search for data.

For example, to count the number of rows in a table, enter **select count($\${Column1}$) from $\${Schema_Table1}$** . The value of $\${Column1}$ is generated by clicking **Add Field Parameter**, and the value of $\${Schema_Table1}$ is generated by clicking **Add Database/Table Parameter**.

- **Output Description** describes each column in the SQL result. Column descriptions are separated by commas (,).

For example, if the relationship is set to **select max ($\${Column1}$), min($\${Column2}$) from $\${Schema_Table1}$** , the output result is **Maximum value,Minimum value**. The result description must correspond to the output result sequence defined by the relationship.

Figure 7-6 Configuring a rule template



- **Abnormal Table Template:** You need to enter a complete SQL statement to specify the abnormal data to be exported.

Step 3 After you click **Yes**, the system publishes the rule template by default. The default version is V1.0.

----End

Managing a Rule Template

A published version of a custom rule template cannot be directly modified. If you want to modify a rule template, you can publish a new version. In addition, you can suspend the historical version and migrate jobs associated with the historical version to the new version.

Step 1 On the DataArts Quality homepage, select **Rule Template** from the left navigation bar. Locate the target rule template in the displayed list and click **Publish** in the **Operation** column.

Figure 7-7 Publishing a rule template

* Rule Name: 自定义表行数规则

* Dimension: Completeness

* Owner Directory: /All/

Description: 1

* Define Relationship: Add Database/Table Parameter, Add Field Parameter. Currently, only one - row and multiple - column relationship expressions are allowed.

```
select count(1) from ${Schema.Table1}
```

Step 2 Dimensions and output description can be modified, and relationships can be redefined.

Step 3 Click **Publish**. In the displayed dialog box, set the version.

Figure 7-8 Publishing a new version

Submit for Publishing

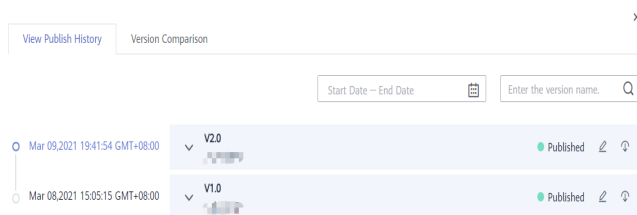
* Version: V1.0

Version history will be automatically generated after publishing.

Yes No

Step 4 After the rule template is submitted for publishing, you can click **View Publish History** in the **Operation** column. You can view the publish history, change the version, and suspend the version.

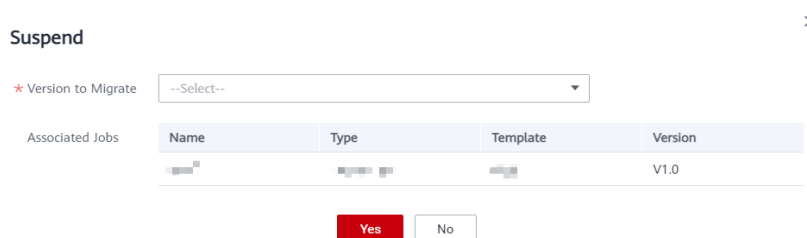
Figure 7-9 Publish History page



Step 5 To suspend a historical version, click **Suspend** on the right of the historical version.

- If the version is not associated with any job, click **OK** to suspend it.
- If the version has associated jobs, select a new version, associate the jobs with the new version, and click **OK**.

Figure 7-10 Migrating and suspending a version



Step 6 On the **Version Comparison** tab page, you can compare the versions to see their differences.

Figure 7-11 Version comparison



----End

Exporting Rule Templates

To export custom rule templates, perform the following steps (you can export a maximum of 200 rule templates at a time):

- Step 1** In the left navigation pane, choose **Quality Monitoring > Rule Templates**, and select the templates to export in the right pane.
- Step 2** Click **Export**. The **Export Rule Template** dialog box is displayed.
- Step 3** Click **Export** to switch to the **Export Records** tab.

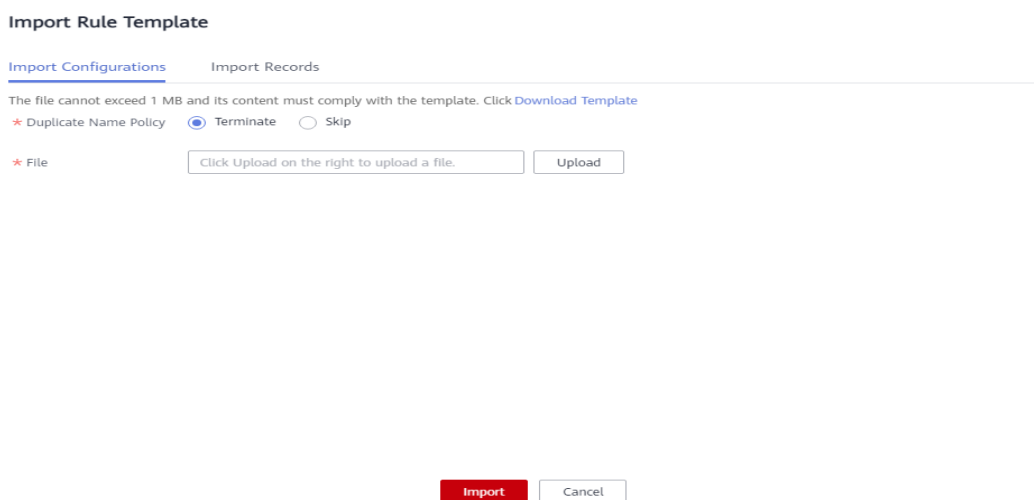
Step 4 In the list of exported files, locate an exported template and click **Download** in the **Operation** column to download the Excel file of the rule template to the local PC.

----End

Importing Rule Templates

You can import a file containing a maximum of 4 MB data.

Step 1 In the left navigation pane, choose **Quality Monitoring > Rule Templates**. In the right pane, click **Import**.



Step 2 On the **Import Configurations** tab page, set **Duplicate Name Policy**.

- **Terminate**: If template names repeat, all templates will fail to be imported.
- **Skip**: If template names repeat, the templates will still be imported.

Step 3 Click **Upload** and select the prepared data file.

NOTE

Edit the data file using either of the following methods:

- (Recommended) Click **Export** to export data and import the data to the system directly or import it after modification.
- Click **Download Template**, fill in the template with the data to import, and import the data to the system.

Step 4 Configure the mapped resource for the catalog and select the directory where rule template has been imported.



Step 5 Click **Import** to import the Excel template to the system.

Step 6 Click the **Import Records** tab to view the import records.

----End

7.2.3 Creating Quality Jobs

You can create quality jobs to apply the created rules to existing tables.

Prerequisites

You have created a directory for storing the quality job. To create a directory, choose **Quality Monitoring > Quality Jobs** in the navigation pane. Before creating a quality job for a data connection, select a directory to store the quality job. For details, see [Figure 7-12](#).

Figure 7-12 Directory that stores the quality job to create



Table 7-11 Buttons in the navigation bar

No.	Description
1	Create Directory
2	Refresh Directory
3	Select All . Right-click to create, delete, and rename directories.

Procedure

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Quality**.
2. Choose **Quality Monitoring > Quality Job** in the left navigation bar.
3. Click **Create**. In the dialog box displayed, set the parameters based on [Table 7-12](#).

Table 7-12 Quality job parameters

Parameter	Description
Name	Quality job name
Description	Information to better identify the quality job. It cannot exceed 256 characters.
Directory	Directory for storing the quality job. You can select a created directory. Figure 7-12 shows the directory.
Job Level	The options are Warning , Minor , Major , and Critical . The job level determines the template for sending notification messages.


4. Click **Next** to go to the **Define Rule** page, on which each rule card corresponds to a subjob. Click  on the rule card and configure it based on

Table 7-13. You can also add more quality rules and click **Next** to apply them to a created database or table.

Figure 7-13 Configuring rules for a quality job

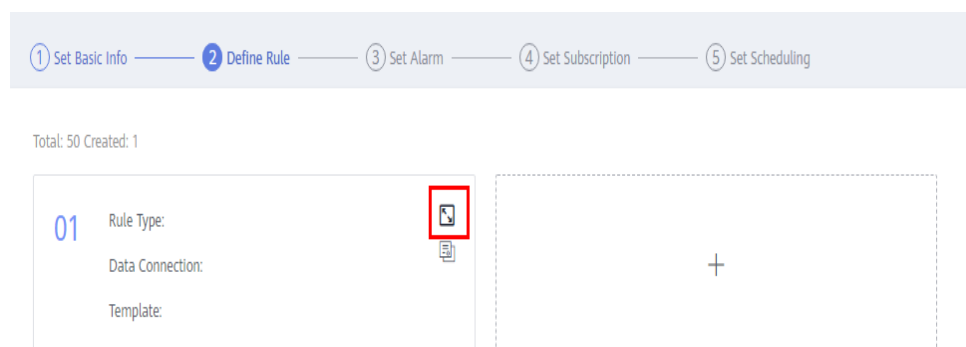


Table 7-13 Parameters for configuring a rule

Parameter	Sub-parameter	Description
Basic Information	Subjob Name	In the job execution result, each rule corresponds to a subjob. You are advised to set the subjob information so that you can view the job execution result and locate faults through logs more easily.
	Description	Information to better identify the subjob
Object	Rule Type	Database rules, table rules, field rules, cross-field rules, or custom rules configured for specific fields in a table.

Parameter	Sub-parameter	Description
	Data Connection	<p>Source and destination objects support the following data connection types: DWS, MRS Hive, DLI, RDS (MySQL and PostgreSQL), Oracle, and MRS Spark (Hudi).</p> <p>Select a created data connection from the drop-down list box.</p> <p>NOTE</p> <ul style="list-style-type: none"> Rules are based on data connections. Therefore, you must create data connections in Management Center before creating data quality rules. For MRS Hive connected through a proxy, select the MRS API mode or proxy mode. <ul style="list-style-type: none"> MRS API mode: An MRS API is used for submission. By default, historical jobs are submitted through MRS APIs. You are advised keep the default settings when editing the job. Proxy mode: A username and a password are used for submission. You are advised to select this mode for new jobs to prevent job submission failures caused by permission issues.
	Database	<p>Select the database to which the configured data quality rules are applied.</p> <p>NOTE</p> <ul style="list-style-type: none"> The database is tailored to the created data connection. When Rule Type is set to Database rule, set the data object to the corresponding database.
	Data Table	<p>Select the table to which the configured data quality rules apply.</p> <p>NOTE</p> <ul style="list-style-type: none"> The table is closely related to the database. When Rule Type is set to Table rule, set the data object to the corresponding table.
	SQL	<p>This parameter is mandatory if you select Custom rule for Rule Type. Enter a complete SQL statement to define how to monitor the quality of data objects.</p>
	Failure Policy	<p>Select Ignore rule errors as required.</p>
	Select Fields	<p>This parameter is mandatory if you select Field rule for Rule Type. Select a field in the corresponding data table.</p> <p>NOTE</p> <p>Fields names containing only one letter (such as a, b, c, and d) cannot be verified.</p>

Parameter	Sub-parameter	Description
	Reference Data Object	This parameter is mandatory if you select Cross-field rule for Rule Type . Select a reference data field.
	Dimension	This parameter is mandatory if you select Custom rule for Rule Type . It associates the custom rule with one of the six quality attributes, including completeness, validity, timeliness, consistency, accuracy, and uniqueness.
Compute Engine	Cluster Name	Select the engine for running the quality job. This parameter is valid only for DLI data connections.
Rule Template	Template	Select a system or custom rule template. NOTE The template type is closely related to the rule type. For details, see Table 7-10 . In addition to system rule templates, you can select the custom rule template created in Creating Rule Templates .
	Version	This parameter is required only when you select a custom rule template. Select the version of the published custom rule template.
	Scoring Weight	Set the weight for the rule based on the field level. The value is an integer from 1 to 9. The default value is 5.
Object Scope	Scanning Scope	You can select All or Partial . The default value is All . If you want only part of data to be computed or quality jobs to be executed periodically based on a timestamp, you can set a WHERE condition for scanning.
	WHERE Clause	Enter a WHERE clause. The system will scan the data that matches the clause. For example, if you want to filter out the data for which the value range of the age field is (18, 60], enter the following WHERE clause: <code>age > 18 and age <= 60</code> You can also enter a dynamic SQL expression. For example, if you want to filter out the data generated 24 hours ago based on the time field, enter the following WHERE clause: <code>time >= (date_trunc('hour', now()) - interval '24 h') and time <= (date_trunc('hour', now()))</code>

Parameter	Sub-parameter	Description
Alarm Condition	Alarm Expression	<p>Set this parameter if you want to set an alarm condition for the current rule. If you want to use the logical operations of multiple rules to set a unified alarm condition expression, you do not need to set this parameter. Instead, you can set it on the next Set Alarm page.</p> <p>After the alarm conditions are configured, the system determines whether to generate an alarm based on the value of Parameter and the alarm condition. Apart from a single alarm expression, you can also use more complex alarm conditions consisting of logical operators. The alarm expression supports the following logical operators, which can be enclosed by "(" and ")".</p> <ul style="list-style-type: none"> • +: addition • -: subtraction • *: multiplying • /: division • ==: equal to • !=: not equal to • >: greater than • <: less than • >=: greater than or equal to • <=: less than or equal to • !: non • : or • &&: and <p>For example, if Rule Template is set to Null value, you can set this parameter as follows:</p> <ul style="list-style-type: none"> • If you want an alarm to be generated when the number of rows with a null value is greater than 10, enter $\\${1}>10$ ($\\${1}$ is the number of rows with a null value). • If you want an alarm to be generated when the ratio of fields with a null value is greater than 80%, enter $\\${3}>0.8$ ($\\${3}$ is the ratio of fields with a null value). • If you want an alarm to be generated when the number of rows with a null value is greater than 10 or the ratio of fields with a null value is greater than 80%, enter $(\\${1}>10) (\\${3}>0.8)$ ($\\${1}$ is the number of rows with a null value, $\\${3}$ is the ratio of fields with a null value, and

Parameter	Sub-parameter	Description
		indicates that an alarm will be generated if either of the conditions is met).
	Parameter	<p>The value of this parameter is obtained from the output of the rule template. If you can click a parameter, the expression of the parameter is displayed in Alarm Expression.</p> <p>For example, if Template is set to Null value, \${1} is displayed in Alarm Expression when you click alarm parameter Null Value Rows.</p>
	Logical Operator	<p>This parameter is optional. You can perform logical operations on the result of an alarm expression to generate more complex alarm conditions.</p> <p>You can move the cursor between two alarm expressions in Alarm Expression and click one of the following operators to insert them. You can also manually enter an operator. The current expression supports the following logical operators which can be enclosed by brackets ().</p> <ul style="list-style-type: none"> • +: addition • -: subtraction • *: multiplying • /: division • ==: equal to • !=: not equal to • >: greater than • <: less than • >=: greater than or equal to • <=: less than or equal to • !: non • : or • &&: and <p>For example, if Template is set to Null value and you want an alarm to be generated when the number of rows with a null value is greater than 10 or the ratio of fields with a null value is greater than 80%, enter (\${1}>10) (\${3}>0.8) for Alarm Expression (\${1} is the number of rows with a null value, \${3} is the ratio of fields with a null value, and indicates that an alarm will be generated if either of the conditions is met).</p>

Parameter	Sub-parameter	Description
	Score Quality	This parameter is mandatory if you select Custom rule for Rule Type .
	Generate Anomaly Data	Enable Generate Anomaly Data and click Select next to Anomaly Table to store the anomaly data that does not comply with the preset rules. NOTE <ul style="list-style-type: none"> For a field rule, the average value, total value, maximum value, and minimum value of a field in the field-level rule template cannot be used to generate anomaly data. If periodic scheduling or re-execution is configured for a quality job, abnormal data detected in each instance scan is inserted into the anomaly table. You are advised to periodically delete the data in the anomaly table to reduce cost and ensure good performance.
	Anomaly Table	Select a database table. You can configure the prefix and suffix of the output table name.
	Output Settings	<ul style="list-style-type: none"> Output Rule Settings: If you select this option, the quality job settings will show up in the anomaly tables so that you can view the anomaly data sources with ease. Output null: If you select this option, and the preset rules are not complied, the null value will show up in anomaly tables.
	Anomaly Data Amount	You can choose to export all anomaly data or the specified amount of anomaly data.
	Anomaly Table SQL	This parameter is mandatory if you select Custom rule for Rule Type . You need to enter a complete SQL statement to specify the abnormal data to be exported.
	View Duplicate Rules	Click it to view the following duplicate rules: <ul style="list-style-type: none"> Determine the rule repetition based on tables and fields. View the related sub-rules and quality jobs that already exist.
Object Scope	Scanning Scope <ul style="list-style-type: none"> If All is selected, all tables are scanned. If Partial is selected, you must enter a WHERE condition expression to precisely locate the partitions to query data. 	

5. Click **Next** and set alarm information. If you have configured an alarm expression in the previous step, the configured expression is automatically displayed. If there are two or more sub-rules, you can use either of the following methods to configure alarms:
 - a. Use the alarm conditions of sub-rules to report alarms.
 - b. Perform mathematical and logical operations on the alarm parameter values to generate a universal alarm expression to specify whether to report alarms for jobs.

The alarm expression supports the following logical operators, which can be enclosed by "(" and ")".

- +: addition
- -: subtraction
- *: multiplying
- /: division
- ==: equal to
- !=: not equal to
- >: greater than
- <: less than
- >=: greater than or equal to
- <=: less than or equal to
- !: non
- ||: or
- &&: and

6. Click **Next** and set the subscription information. If the SMN notification is required, enable **Notification**, and set **Notification Type** and **Topic**.

 **NOTE**

After notification is enabled, a notification is sent for all the subjobs of the configured notification type.

7. Click **Next** to go to the page where you can select a scheduling mode. Currently, **Once** and **On schedule** are supported. Set parameters for scheduling periodically by referring to [Table 7-14](#). Click **Submit**.

 **NOTE**

1. If **Once** is selected, a manual task instance is generated. A manual task has no dependency on scheduling and must be manually triggered.
2. If **On schedule** is selected, a periodic instance is generated. A periodic instance is an instance snapshot that is automatically scheduled when a periodic task reaches the scheduled execution time.
3. When a periodic task is scheduled once, an instance workflow is generated. You can perform routine O&M on scheduled instance tasks, such as viewing the running status, stopping and rerunning the scheduled tasks.
4. Only MRS clusters that support job submission through an agency support periodic scheduling of quality jobs. MRS clusters that support job submission through an agency are as follows:
 - Non-security MRS cluster
 - MRS security cluster whose version is later than 2.1.0, and that has MRS 2.1.0.1 or later installed

Table 7-14 Parameters

Parameter	Description
Effective	Effective date of a scheduling task.
Cycle	The frequency at which a scheduling task is executed. Related parameters are: <ul style="list-style-type: none">• Minutes• Hours• Days• Weeks NOTE <ul style="list-style-type: none">• If Cycle is set to Minutes or Hours, set the start time, end time, and interval for the scheduling task. Currently, the start time is in minute for stagger scheduling.• If Cycle is set to Days, set a specified time when the scheduling task is enabled every day.• If Cycle is set to Weeks, set Scheduling Time and Start from for the scheduling task, that is, <i>XX</i> o'clock <i>XX</i> minutes on <i>XXX</i> every week.

Exporting Quality Jobs

You can export a maximum of 200 quality jobs.

- Step 1** In the left navigation pane on the **DataArts Quality** page, choose **Quality Monitoring > Quality Jobs**. In the right pane, select the quality jobs to export.
- Step 2** Click **Export**. The **Export Quality Job** dialog box is displayed.
- Step 3** Click **Export** to switch to the **Export Records** tab.

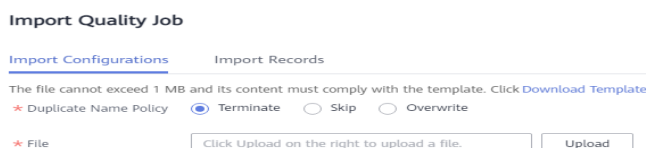
Step 4 In the list of exported files, locate an exported quality job and click **Download** in the **Operation** column to download the Excel file of the job to the local PC.

----End

Importing Quality Jobs

You can import a file containing a maximum of 4 MB data.

Step 1 In the left navigation pane on the **DataArts Quality** page, choose **Quality Monitoring > Quality Jobs**. In the right pane, click **Import**. The **Import Quality Job** dialog box is displayed.



Step 2 On the **Import Configurations** tab page, set **Duplicate Name Policy**.

- **Terminate:** If quality job names repeat, all quality jobs will fail to be imported.
- **Skip:** If quality job names repeat, the quality jobs will still be imported.
- **Overwrite:** If quality job names repeat, new jobs will replace existing ones with the same names.

Step 3 Click **Upload** and select the prepared data file.

NOTE

Edit the data file using either of the following methods:

- (Recommended) Click **Export** to export data and import the data to the system directly or import it after modification.
- Click **Download Template**, fill in the template with the data to import, and import the data to the system.

Step 4 Configure resource mapping for the data connection, cluster, directory, and topic.



- **Data Connection:** Select the type of the imported data connection.

- **Cluster:** If the data connection type is DLI, select the corresponding queue.
- **Directory:** Select the directory where the imported quality job is stored.
- **Topic:** If SMN is configured, you need to select a topic.

Step 5 Click **Import** to import the Excel template to the system.

Step 6 Click the **Import Records** tab to view the import records.

----End

7.2.4 Creating a Comparison Job

Data comparison is critical to ensure data consistency in data development and migration. The cross-source data comparison capability is the key to checking consistency of the data before and after migration or processing.

Comparison jobs in Quality Monitoring support cross-source data comparison. You can apply created rules to two tables for quality monitoring and output the comparison result.

Prerequisites

You have created a directory for storing the comparison job. To create a directory, choose **Quality Monitoring > Comparison Jobs** in the navigation pane. Before creating a comparison job for a data connection, select a directory to store the comparison job. For details, see [Figure 7-14](#).

[Table 7-15](#) describes the directory-related operations.

Figure 7-14 Directory that stores the comparison job to create

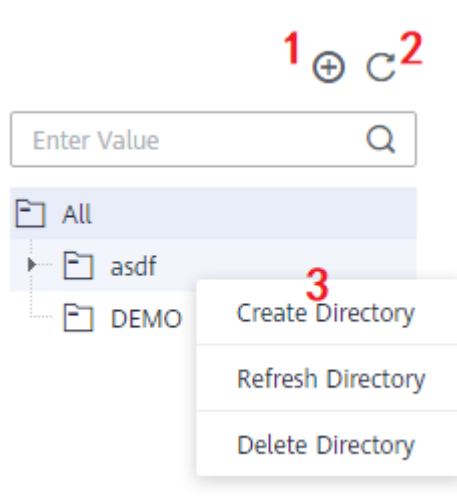


Table 7-15 Buttons in the navigation bar

No.	Description
1	Create Directory
2	Refresh Directory

No.	Description
3	Select All . Right-click to create, delete, and rename directories.

Creating a Job

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Quality**.
2. Choose **Quality Monitoring > Comparison Job** from the left navigation bar.
3. Click **Create**. In the dialog box displayed, set the parameters based on [Table 7-16](#).

Table 7-16 Comparison job parameters

Parameter	Description
Name	Comparison job name
Description	Information to better identify a comparison job. It cannot exceed 256 characters.
Directory	The directory for storing the comparison job to create. You can select a created directory. Figure 7-14 shows the directory.
Job Level	The options are Warning , Minor , Major , and Critical . The job level determines the template for sending notification messages.

4. Click **Next** to go to the **Define Rule** page. Click  on the rule card and configure it based on [Table 7-17](#). You can also add comparison rules.

Figure 7-15 Configuring rules for a comparison job

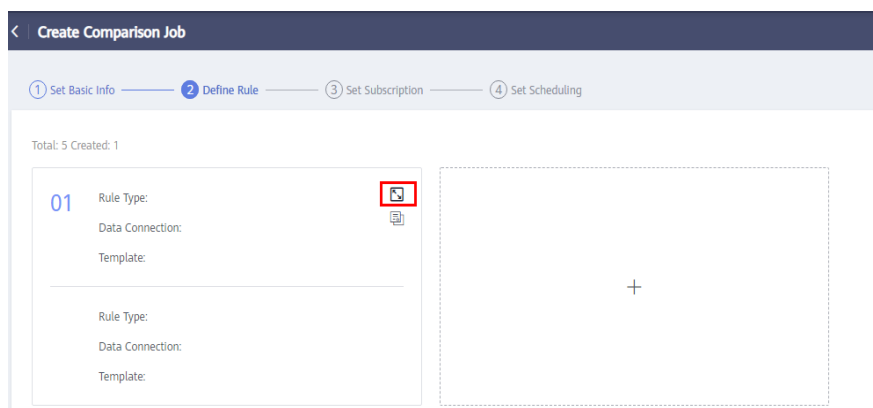


Table 7-17 Parameters for configuring a rule template

Module	Parameter	Description
Basic Information	Subjob Name	In the job execution result, each rule corresponds to a subjob. You are advised to set the subjob information so that you can view the job execution result and locate faults through logs more easily.
	Description	Information to better identify the subjob
Object	Rule Type	<p>The options are Table rule, Field rule, and Custom rule. Field-level rules can be used to configure monitoring rules for specific fields in tables. For example, set this parameter to Table rule, and set other configuration items on the page to table-level rule configuration items correspondingly.</p> <p>The rule type of the destination object is automatically generated based on that of the source object.</p>
	Data Connection	<p>Source and destination objects support the following data connection types: DWS, MRS Hive, DLI, RDS (MySQL and PostgreSQL), Oracle, and MRS Spark (Hudi).</p> <p>Select a created data connection from the drop-down list box.</p> <p>NOTE</p> <ul style="list-style-type: none"> Rules are based on data connections. Therefore, you must create data connections in Management Center before creating data quality rules. For MRS Hive connected through a proxy, select the MRS API mode or proxy mode. <ul style="list-style-type: none"> MRS API mode: An MRS API is used for submission. By default, historical jobs are submitted through MRS APIs. You are advised keep the default settings when editing the job. Proxy mode: A username and a password are used for submission. You are advised to select this mode for new jobs to prevent job submission failures caused by permission issues.
	Data Object	<p>The data table selected for the source object is compared with the data table of the destination object on the right. Select the table to which the configured comparison rule applies.</p> <p>NOTE</p> <p>The table is closely related to the database. The database is tailored to the created data connection.</p>

Module	Parameter	Description
	SQL	This parameter is mandatory if you select Custom rule for Rule Type . Enter a complete SQL statement to define how to monitor the quality of data objects.
Compute Engine	Cluster Name	Select the engine for running the comparison job. This parameter is valid only for DLI data connections.
Rule Template	Template	<p>This parameter defines how to monitor the quality of data objects.</p> <p>The template name of the source object contains the system rule template and custom rule template.</p> <p>The template name of the destination object is automatically generated based on the rule type of the source object.</p> <p>NOTE The template type is closely related to the rule type. For details, see Table 7-10. In addition to system rule templates, you can select the custom rule template created in Creating Rule Templates.</p>
	Version	This parameter is required only when you select a custom rule template. Select the version of the published custom rule template.
Object Scope	Scanning Scope	<p>You can select All or Partial. The default value is All.</p> <p>If you want only part of data to be computed or quality jobs to be executed periodically based on a timestamp, you can set a WHERE condition for scanning.</p>
	WHERE Clause	<p>Enter a WHERE clause. The system will scan the data that matches the clause.</p> <p>For example, if you want to filter out the data for which the value range of the age field is (18, 60], enter the following WHERE clause:</p> <pre>age > 18 and age <= 60</pre> <p>You can also enter a dynamic SQL expression. For example, if you want to filter out the data generated 24 hours ago based on the time field, enter the following WHERE clause:</p> <pre>time >= (date_trunc('hour', now()) - interval '24 h') and time <= (date_trunc('hour', now()))</pre>

Module	Parameter	Description
Alarm Condition	Alarm Expression	<p>Set this parameter if you want to set an alarm condition for the current rule.</p> <p>After the alarm conditions are configured, the system determines whether to generate an alarm based on the value of Parameter and the alarm condition. Apart from a single alarm expression, you can also use more complex alarm conditions consisting of logical operators. The alarm expression supports the following logical operators, which can be enclosed by "(" and ")".</p> <ul style="list-style-type: none">• +: addition• -: subtraction• *: multiplying• /: division• ==: equal to• !=: not equal to• >: greater than• <: less than• >=: greater than or equal to• <=: less than or equal to• !: non• : or• &&: and <p>For example, if Rule Template of the source and destination of the comparison job is set to Table Rows, you can configure the alarm expression as follows:</p> <ul style="list-style-type: none">• To generate an alarm when the number of rows in the source table is less than 100, enter `\${1_1}<100, where `\${1_1} indicates the total number of rows in the source table.• To generate an alarm when the number of rows in the source table is not equal to that in the destination table, enter `\${1_1}!=`\${2_1}, where `\${1_1} indicates the total number of rows in the source table and `\${2_1} indicates the total number of rows in the destination table.• To generate an alarm when the number of rows in the source table is less than 100 or when the number of rows in the source table is not equal to that in the destination table, enter `\${1_1}<100 `\${1_1}!=`\${2_1}, where `\${1_1} and `\${2_1} indicate the total number of rows in the source and destination tables, respectively,

Module	Parameter	Description
		<p>and indicates that an alarm is generated if either condition is met.</p>
	Parameter	<p>The value of this parameter is obtained from the output of the rule template. If you can click a parameter, the expression of the parameter is displayed in Alarm Expression.</p> <p>For example, if Template is set to Table Rows, #{1_1} is displayed in Alarm Expression when you click alarm parameter Table Rows.</p>
	Logical Operator	<p>This parameter is optional. You can perform logical operations on the result of an alarm expression to generate more complex alarm conditions.</p> <p>You can move the cursor between two alarm expressions in Alarm Expression and click one of the following operators to insert them. You can also manually enter an operator. The current expression supports the following logical operators which can be enclosed by brackets ().</p> <ul style="list-style-type: none"> ● +: addition ● -: subtraction ● *: multiplying ● /: division ● ==: equal to ● !=: not equal to ● >: greater than ● <: less than ● >=: greater than or equal to ● <=: less than or equal to ● !: non ● : or ● &&: and <p>For example, if Template is Table Rows and if you want to generate an alarm when the number of rows in the source table is less than 100 or when the number of rows in the source table is not equal to that in the destination table, enter #{1_1}<100 #{1_1}!={2_1}, where #{1_1} and #{2_1} indicate the total number of rows in the source and destination tables, respectively, and indicates that an alarm is generated if either condition is met.</p>

- Click **Next** and set the subscription configuration. If the SMN notification is required, enable **Notification**, and set **Notification Type** and **Topic**. See [Figure 7-16](#).

Figure 7-16 Subscription configuration

* Notification

* Notification Type Alarm triggered Run successfully

* Topic [View Topic](#)
SMN required for the topic may incur charges. For details, click [Billing Rule](#)

- Click **Next** to go to the page where you can select a scheduling mode. Currently, **Once** and **On schedule** are supported. Set parameters for scheduling periodically by referring to [Table 7-18](#). Click **Submit**.

NOTE

- If **Once** is selected, a manual task instance is generated. A manual task has no dependency on scheduling and must be manually triggered.
- If **On schedule** is selected, a periodic instance is generated. A periodic instance is an instance snapshot that is automatically scheduled when a periodic task reaches the scheduled execution time.
- When a periodic task is scheduled once, an instance workflow is generated. You can perform routine O&M management on scheduled instance tasks, such as viewing the running status, stopping and rerunning the scheduled tasks.
- Only MRS clusters that support job submission through an agency support periodic scheduling of comparison jobs. MRS clusters that support job submission through an agency are as follows:
 - Non-security MRS cluster
 - MRS security cluster whose version is later than 2.1.0, and that has MRS 2.1.0.1 or later installed

Table 7-18 Parameters for setting the scheduling mode

Parameter	Description
Effective	Effective date of a scheduling task.
Cycle	<p>The frequency at which a scheduling task is executed. Related parameters are:</p> <ul style="list-style-type: none"> Minutes Hours Days Weeks <p>NOTE</p> <ul style="list-style-type: none"> If Cycle is set to Minutes or Hours, set the start time, end time, and interval for the scheduling task. If Cycle is set to Days, set the start time of the scheduling task. If Cycle is set to Weeks, set Scheduling Time and Start from for the scheduling task, that is, <i>XX</i> o'clock <i>XX</i> minutes on <i>XXX</i> every week.

Exporting Comparison Jobs

You can export a maximum of 200 comparison jobs.

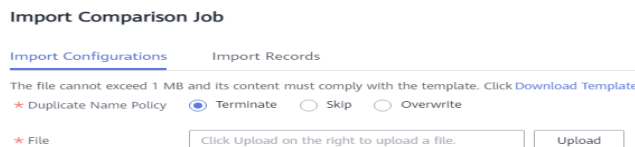
- Step 1** In the left navigation pane on the **DataArts Quality** page, choose **Quality Monitoring > Comparison Jobs**. In the right pane, select the comparison jobs to export.
- Step 2** Click **Export**. The **Export Comparison Job** dialog box is displayed.
- Step 3** Click **Export** to switch to the **Export Records** tab.
- Step 4** In the list of exported files, locate an exported quality job and click **Download** in the **Operation** column to download the Excel file of the job to the local PC.

----End

Importing Comparison Jobs

You can import a file containing a maximum of 4 MB data.

- Step 1** In the left navigation pane on the **DataArts Quality** page, choose **Quality Monitoring > Comparison Jobs**. In the right pane, click **Import**. The **Import Comparison Job** dialog box is displayed.



- Step 2** On the **Import Configurations** tab page, set **Duplicate Name Policy**.
 - **Terminate**: If comparison job names repeat, all comparison jobs will fail to be imported.
 - **Skip**: If comparison job names repeat, the comparison jobs will still be imported.
 - **Overwrite**: If comparison job names repeat, new jobs will replace existing ones with the same names.
- Step 3** Click **Upload** and select the prepared data file.

NOTE

Edit the data file using either of the following methods:

- (Recommended) Click **Export** to export data and import the data to the system directly or import it after modification.
- Click **Download Template**, fill in the template with the data to import, and import the data to the system.

Step 4 Configure resource mapping for the data connection, cluster, directory, and topic. Click **Import** to import the Excel template to the system.



- **Data Connection:** Select the type of the imported data connection.
- **Cluster:** If the data connection type is DLI, select the corresponding queue.
- **Directory:** Select the directory where the imported comparison job is stored.
- **Topic:** If SMN is configured, you need to select a topic.

Step 5 Click the **Import Records** tab to view the import records.

----End

7.2.5 Viewing Job Instances

GUI Description

The following figure shows the areas and buttons on the **O&M** page.

Figure 7-17 O&M page

Instance Name	Type	Running Status	Notification	Start Time	Instance Search Duration	Operation
#123-280	Quality job	Failed	Not triggered	Jun 18, 2022 00:00:25 GMT+08:00	00:01:47	Run Details Rectify
#123-279	Quality job	Successful	Not triggered	Jun 17, 2022 00:00:48 GMT+08:00	00:01:15	Run Details Rectify
#123-278	Quality job	Successful	Not triggered	Jun 16, 2022 00:05:14 GMT+08:00	00:01:46	Run Details Rectify

Table 7-19 O&M page

No.	Area	Description
1	Navigation bar	Contains the storage directory of data quality rules. You can store rules in different directories tailored to service requirements. The number next to each directory indicates the number of rule instances stored in the directory.
2	List of rule instances	Displays the instance name, type, running status, and running result.

No.	Area	Description
3	Management area	Provides buttons for exporting and deleting selected instances.
4	Search area	<ul style="list-style-type: none">• Displays rule instances based on specified conditions. For example, you can display rule instances for a specified time range.• Displays a list of instances according to the handler, creator, or instance name. Fuzzy search is supported.

Table 7-20 List of rule instances

Parameter	Description
Instance Name	Consists of a rule name and a number. The larger the number is, the later the instance is created.
Type	Displays the job type. The value can be Quality Job or Comparison Job .
Running Status	Displays the running status of an instance, such as Successfully , Failed , Running , and Alarming . In the right pane, you can view the detailed run logs of the rule instances. <ul style="list-style-type: none">• Successfully: The instance stops normally and the running result meets the expectation.• Failed: The instance stops unexpectedly.• Alarming: The instance stops normally, but the running result does not meet the expectation.• Running: The instance is running, but no running result is displayed.
Notification	Displays the notification status of an instance, such as Successfully , Failed , and Not triggered .
Start Time	Displays the time when the instance starts to run.
Running Duration	Displays the running duration of the instance.
Rerun	Allows you to run a rule instance again.

Parameter	Description
Details	<p>Displays the running results and logs of job instances.</p> <ul style="list-style-type: none">• Comparison Job Result In the running result, the left pane displays the execution result of the rule for source table rows, and the right pane displays the execution result of the rule for destination table rows. The error rate indicates the difference between the number of rows of the source and destination tables. If the error rate is 0, the source and destination tables have the same number of rows.
Rectify	<p>Allows you to perform further processing on a rule instance. For example, you can Provide defects directly, Close defects, or Specify a user to rectify fault.</p> <p>The above operations can be performed only when you are the handler of the instance.</p>

7.2.6 Viewing Quality Reports

You can query the quality reports of business metrics and data objects to determine whether their quality meets the requirements.

Querying Business Quality Reports

The full quality score can be set to 5, 10, or 100 points. By default, a five-point scale is used for quality scoring based on table-associated rules. The scores in different dimensions, such as tables, business objects, and subject areas, are calculated based on the weighted average values of rule scores in different dimensions.

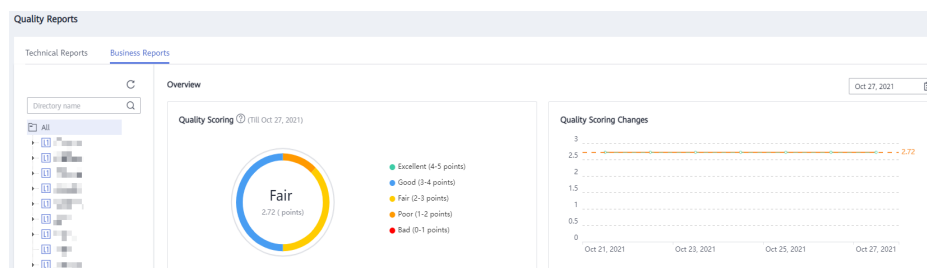
You can query the quality scores of subject area groups, subject areas, business objects, tables, and table-associated rules. For details on the calculation formulas, see [Table 7-21](#).

Table 7-21 Formulas for calculating scores

Object	Formula
Rule	<p>When a quality job that contains a percentage-related rule (either built-in or custom) is created, a quality report can be generated.</p> <ul style="list-style-type: none"> Percentage-related rules can be classified into positive rules and negative rules. For a positive rule, the higher the percentage is, the better the data quality is. For a negative rule, the higher the percentage is, the poorer the data quality is. Rules that contain the unique value percentage, duplicate value percentage, and valid percentage are positive rules, and rules that contain the null value percentage are negative rules. Positive rule score = Number of data rows that meet the rule/ Total number of data rows x Full score (5, 10, or 100 points). Negative rule score = (1 - Number of data rows that meet the rule/Total number of data rows) x Full score (5, 10, or 100 points). If the table is empty (the total number of rows is 0), the positive rule score is fixed at the full score and the negative rule score is fixed at 0 points.
Table	The table score is calculated as follows: $\sum(\text{Scores of all rules associated with the table} \times \text{Rule weight}) / \sum \text{Rule weight}$.
Business object	Weighted average value of the scores of all tables under the business object, that is, $\sum \text{Scores of all tables under the business object} / \text{Number of tables}$.
Subject area	Weighted average value of scores of all business objects in the subject area, that is, $\sum \text{Scores of all business objects in the subject area} / \text{Number of business objects}$.
Subject area group	Average weighted value of the scores of all subject areas in the group, that is, $\sum \text{Scores of all subject areas in the group} / \text{Number of subject areas}$.

- Step 1** On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Quality**.
- Step 2** Choose **Quality Monitoring > Quality Job** in the left navigation bar.
- Step 3** Click the **Business Reports** tab, and select a subject and an end date to query the quality scores of the end date and the previous seven days, as shown in [Figure 7-18](#).

Figure 7-18 Business object



NOTE

- Take the full score 5 points as an example. Points 4 to 5: excellent; 3 to 4: good; 2 to 3: fair; 1 to 2: qualified; 0 to 1: unqualified.
- The quality score data of a day is generated in the early morning of the next day.
- In the **Quality Scoring Changes** area, the solid line consists of the quality scores of the end date and the previous seven days, and the dashed line indicates the average quality score of these days.
- If the job is executed multiple times on a day, the last score is used as the quality score of the day.

Step 4 Click the score link in the **Table Score** column to expand the scores of the rules associated with the table.

Step 5 Click the score link in the **Rule Score** column to expand the scores of the fields associated with the rule.

Figure 7-19 Table-associated rule scores

X

Sub-rule Field Score

Name	Rule Desc	Score	Column ...	空值行数	总行数	空值率	Alarm St...
ycr.dwr.d...	COLUM...	5.0	5	0	2	0.0	false

----End

Viewing Data Quality Reports

The full quality score can be set to 5, 10, or 100 points. By default, a five-point scale is used for quality scoring based on table-associated rules. Scores in different dimensions, such as tables and databases, are calculated based on the weighted average values of rule scores in different dimensions.

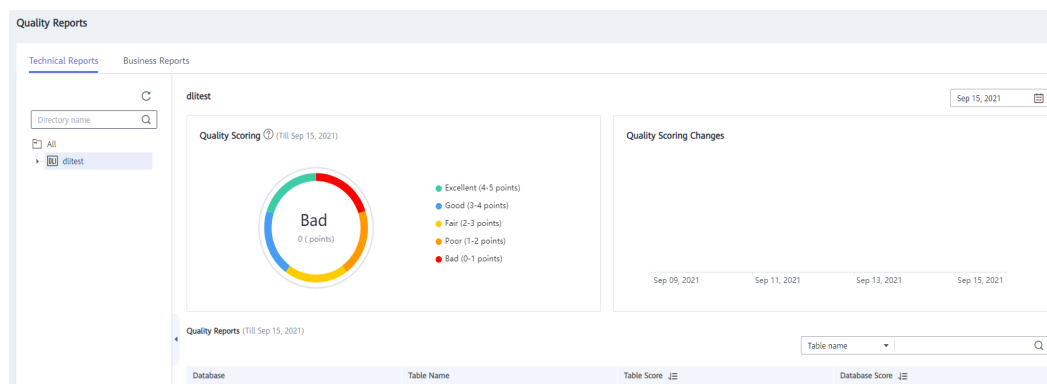
You can query the scores of databases, tables, and table-associated rules. For details on the calculation formulas, see [Table 7-22](#).

Table 7-22 Formulas for calculating scores

Object	Formula
Rule	<p>When a quality job that contains a percentage-related rule (either built-in or custom) is created, a quality report can be generated.</p> <ul style="list-style-type: none"> Percentage-related rules can be classified into positive rules and negative rules. For a positive rule, the higher the percentage is, the better the data quality is. For a negative rule, the higher the percentage is, the poorer the data quality is. Rules that contain the unique value percentage, duplicate value percentage, and valid percentage are positive rules, and rules that contain the null value percentage are negative rules. Positive rule score = Number of data rows that meet the rule/ Total number of data rows x 5. Negative rule score = (1 - Number of data rows that meet the rule/Total number of data rows) x 5.
Table	The table score is calculated as follows: $\sum(\text{Scores of all rules associated with the table} \times \text{Rule weight}) / \sum \text{Rule weight}$.
Database	Weighted average value of the scores of all data tables in the database, that is, $\sum \text{Scores of all data tables in the database} / \text{Number of tables}$.

- Step 1** On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Quality**.
- Step 2** Choose **Quality Monitoring > Quality Job** in the left navigation bar.
- Step 3** Click the **Technical Report** tab, and select a data connection and an end date to query the quality scores of the end date and the previous seven days, as shown in [Figure 7-20](#).

Figure 7-20 Selecting a data connection



 NOTE

- Take the full score 5 points as an example. Points 4 to 5: excellent; 3 to 4: good; 2 to 3: unqualified; 1 to 2: poor; 0 to 1: very poor.
- The quality score data of a day is generated in the early morning of the next day.
- In the **Quality Scoring Changes** area, the solid line consists of the quality scores of the end date and the previous seven days, and the dashed line indicates the average quality score of these days.
- If the job is executed multiple times on a day, the last score is used as the quality score of the day.

Step 4 Click the score link in the **Table Score** column to expand the scores of the rules associated with the table.

Step 5 Click the score link in the **Rule Score** column to expand the scores of the fields associated with the rule.

Figure 7-21 Table-associated rule scores



Name	Rule Desc	Score	Column ...	空值行数	总行数	空值率	Alarm St...
ycr.dwr.d...	COLUM...	5.0	5	0	2	0.0	false

----End

7.3 Tutorials

7.3.1 Creating a Business Scenario

Scenario

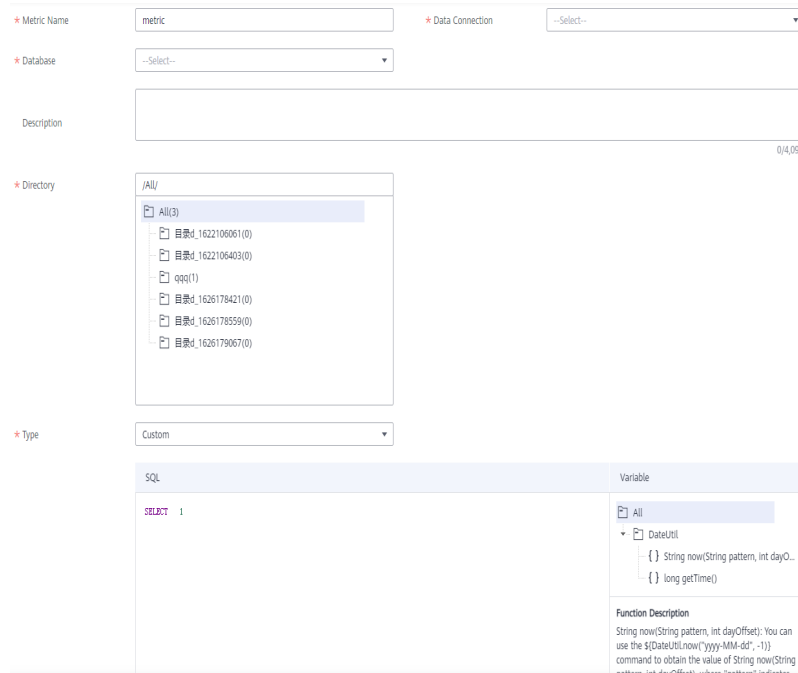
Business scenarios are used to monitor business metrics. This section describes how to create a business scenario.

Procedure

Step 1 On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Quality**.

Step 2 Create a metric.

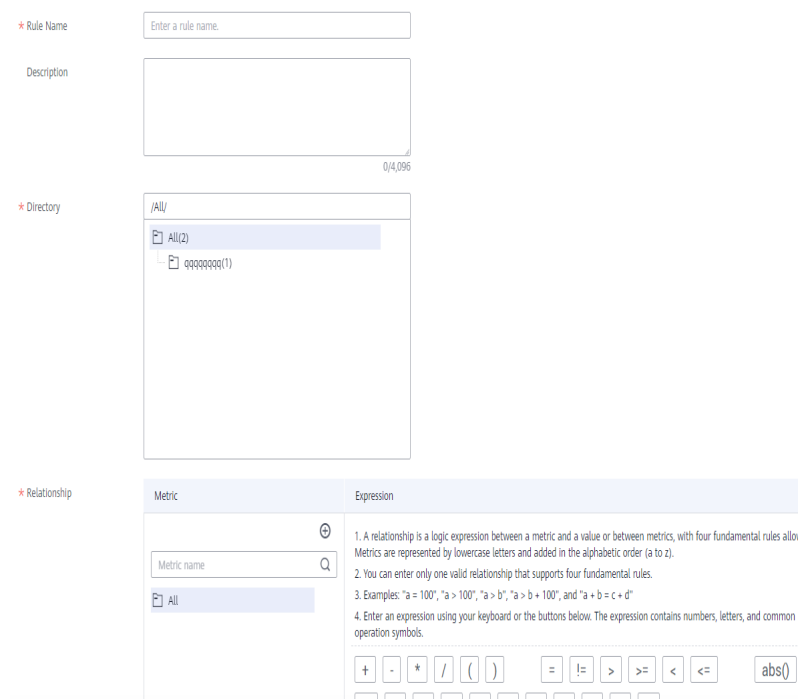
1. In the navigation pane on the left, choose **Metrics**.
2. On the **Metrics** page, click **Create**.



3. Click **Trial Run** to check whether the metric runs properly.
4. Click **OK**.

Step 3 Create a rule.

1. In the navigation pane on the left, choose **Rules**.
2. On the **Rules** page, click **Create**.
3. Set the parameters shown in the following figure.



4. Click **OK**.
5. On the **Rules** page, click **Create** to create another rule.

6. Set the parameters shown in the following figure.

The screenshot shows a configuration window for a rule. It contains the following sections:

- * Rule Name:** A text input field with the placeholder "Enter a rule name."
- Description:** A large text area with a character count of 0/4,096.
- * Directory:** A tree view showing a folder structure: "/All/" containing "All(2)" and "qqqqqqqq(1)".
- * Relationship:** A section with a "Metric" dropdown menu (currently showing "All") and an "Expression" field. The Expression field includes a rich text editor with the following instructions:
 1. A relationship is a logic expression between a metric and a value or between metrics, with four fundamental rules allow Metrics are represented by lowercase letters and added in the alphabetic order (a to z).
 2. You can enter only one valid relationship that supports four fundamental rules.
 3. Examples: "a = 100", "a > 100", "a > b", "a > b + 100", and "a + b = c + d"
 4. Enter an expression using your keyboard or the buttons below. The expression contains numbers, letters, and common operation symbols.
 Below the text is a toolbar with mathematical symbols: +, -, *, /, (,), =, !=, >, >=, <, <=, and abs().

7. Click **OK**.

Step 4 Create a scenario.

1. In the navigation pane on the left, choose **Scenarios**.
2. On the **Scenarios** page, click **Create**. On the displayed **Create Scenario** page shown in the following figure, set the required parameters.

The screenshot shows the "Create Scenario" configuration window. It features a progress bar at the top with four steps: 1. Set Basic Info (active), 2. Define Rule Group, 3. Set Subscription, and 4. Set Scheduling. The main configuration area includes:

- * Scenario Name:** A text input field with the placeholder "Enter a scenario name."
- Description:** A large text area with a character count of 0/256.
- * Directory:** A tree view showing a folder structure: "/All/" containing "All(2)" and "qqqqqq(1)".
- * Level:** A dropdown menu currently set to "Warning".

3. Click **Next** and set the parameters for the rule group.

4. Click **Next** and set subscription parameters.

5. Click **Next** and set scheduling parameters.

6. Click **Submit**.

Step 5 In the scenario list, locate the created scenario and click **Run** in the **Operation** column.

1. Click the refresh button in the upper right corner. The **Running Status** of the scenario is **Succeeded**.
2. Click the running result to view details.

----End

7.3.2 Creating a Quality Job

Scenario

You can use a quality job to monitor data quality. This section describes how to create a quality job.

Procedure

Step 1 On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Quality**.

Step 2 Create a rule template.

1. In the navigation pane on the left, choose **Rule Templates**. System templates are displayed. Rule templates have six dimensions: completeness, uniqueness, timeliness, validity, accuracy, and consistency.

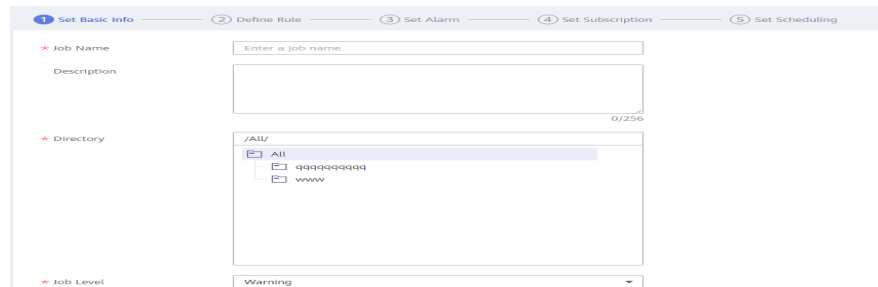
2. **Optional:** Click **Create** to create a rule template.

 **NOTE**

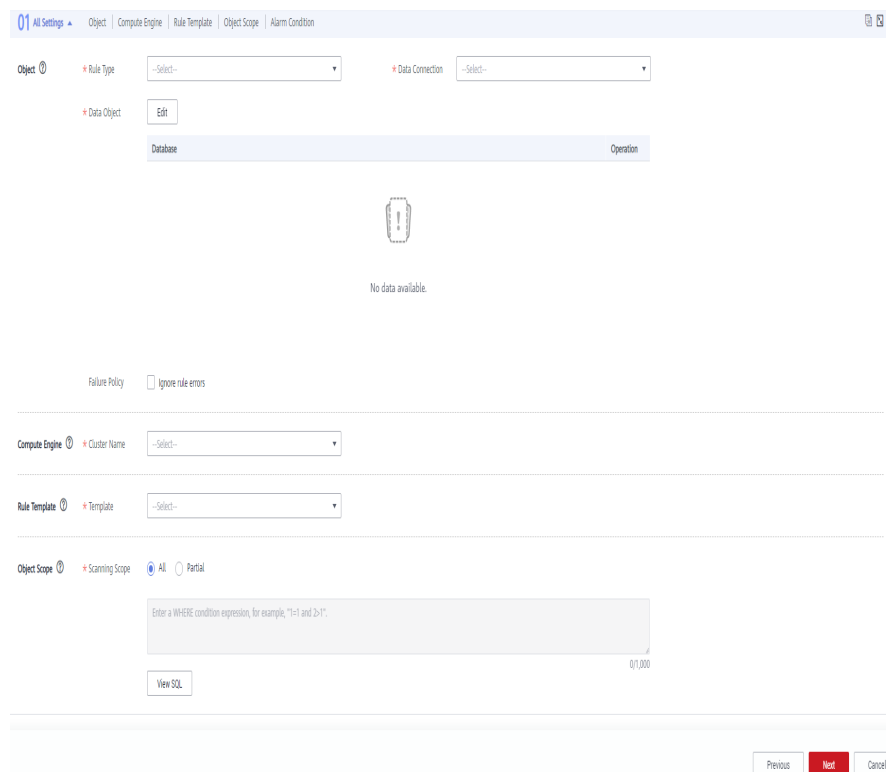
In this example, use a system rule.

Step 3 Create a quality job.

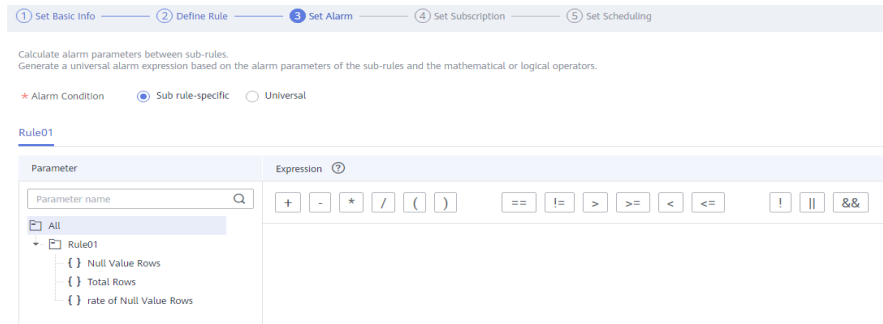
1. In the navigation pane on the left, choose **Quality Jobs**.
2. Click **Create**. On the **Create Quality Job** page, set basic information about the quality job.



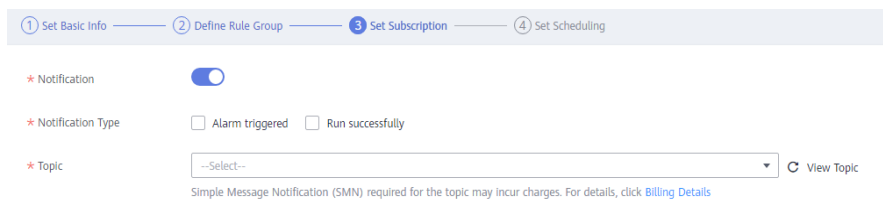
3. Click **Next** to go to the **Define Rule** page. Click  on the rule card to configure the rule.



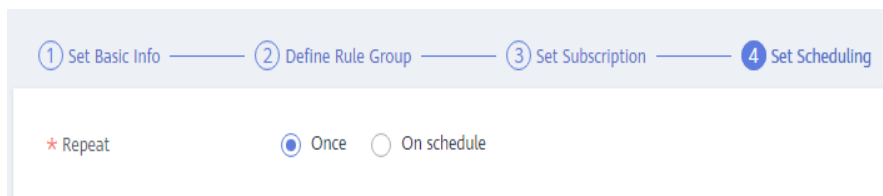
4. Click **Next** and set alarm parameters.



5. Click **Next** and set subscription parameters.



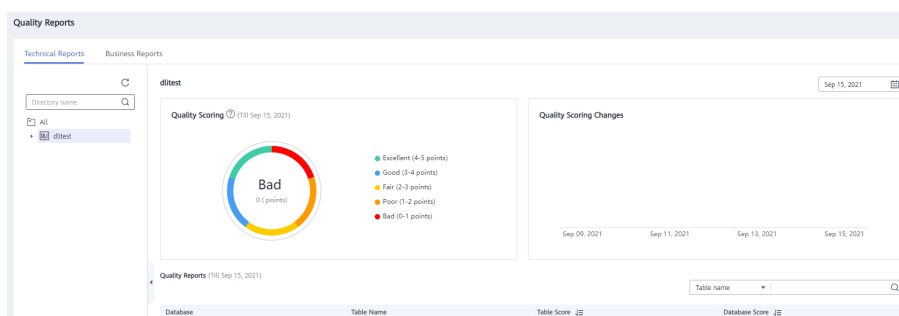
6. Click **Next** and set scheduling parameters.



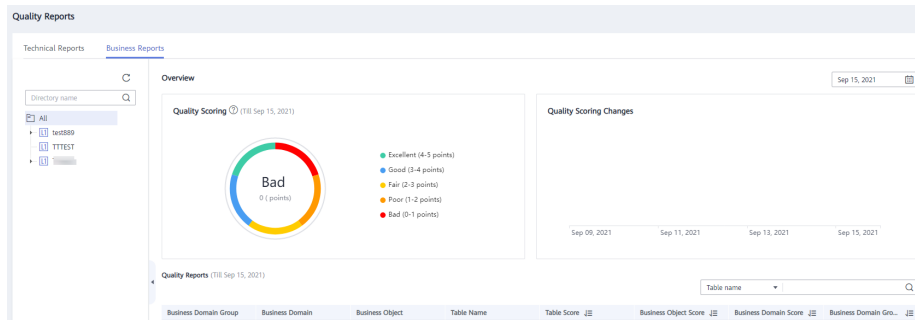
7. Click **Submit**.

Step 4 In the quality job list, locate the created job and click **Run** in the **Operation** column.

1. After the quality job is successfully run, choose **Quality Reports** in the navigation pane on the left.
2. The **Technical Reports** page is displayed by default.



3. Click the **Business Reports** tab and view the business reports.



----End

7.3.3 Creating a Comparison Job

Scenario

Data comparison is critical to ensure data consistency in data development and migration. The cross-source data comparison capability is the key to checking consistency of the data before and after migration or processing. This section describes how to create a comparison job in the DataArts Quality module of DataArts Studio to verify consistency between a DLI and DWS connection.

Environment Preparations

Create the data sources to compare, that is, create different types of data connections in the Management Center.

Procedure

Step 1 Create different types of data connections.

1. Create a DLI data connection. On the **Management Center** page, click **Create Data Connection**. In the displayed dialog box, select **DLI** for **Data Connection Type**, enter a connection name, and click **Test**. If the message "Connected." is displayed, click **OK**.

✕

Edit Data Connection

* Data Connection Type

* Name

Tag

2. Create a DWS data connection. On the **Management Center** page, click **Create Data Connection**. In the displayed dialog box, select **DWS** for **Data Connection Type**, enter a connection name, set other required parameters, and click **Test**. If the message "Connected." is displayed, click **OK**.

✕

Edit Data Connection

* Data Connection Type: DWS

* Name: test1027

Tag: --

* Manual:

* SSL Connection:

* Cluster Name ?: ttd1027 [Manage Cluster](#)

* Username: dbadmin

* Password:

* KMS Key ?: dlf/default [Access KMS](#)

* Connection Type: Proxy connection Direct connection

* ?: ... [Manage CDM](#)

OK
Test
Cancel

Step 2 Create a comparison job.

1. On the **DataArts Quality** page, choose **Comparison Jobs** in the navigation pane.
2. Click **Create**. On the **Create Comparison Job** page, set basic information about the comparison job.

Basic Settings
Rule Settings
Subscription Settings
Scheduling Settings

* Job Name: compare_dws_dli

Description: 0/256

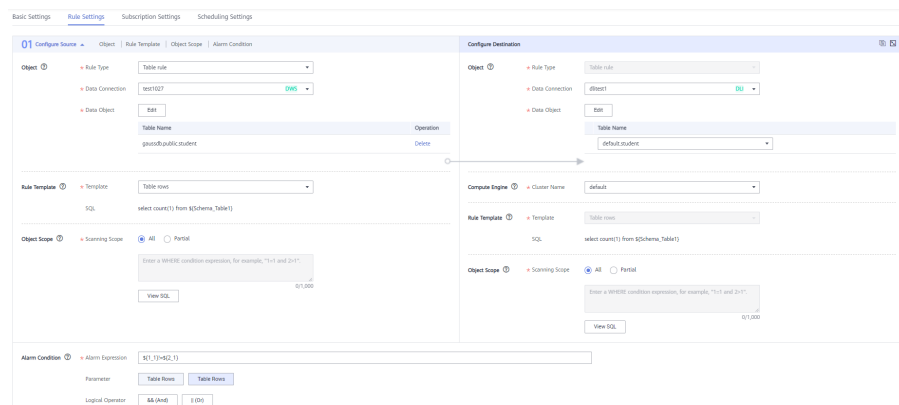
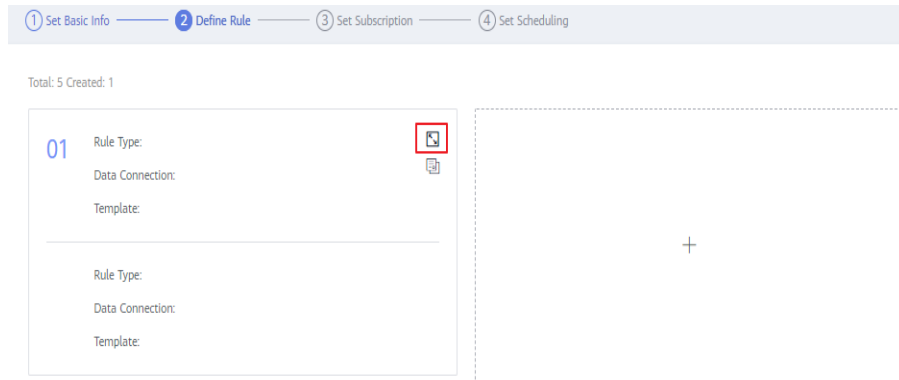
* Directory:

/All/

All

* Job Level: Warning

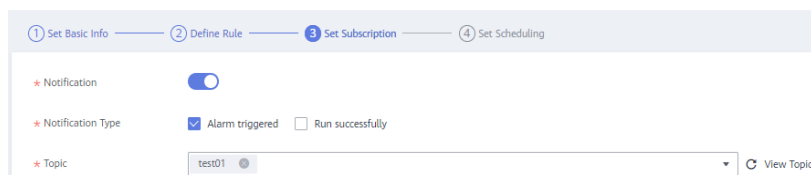
3. Click **Next** to go to the **Define Rule** page. Click on the rule card to configure the rule.



NOTE

- You need to configure information about both the source and destination.
- When configuring **Alarm Condition**, **#{1_1}** indicates the number of rows in the source table, and **#{2_1}** indicates the number of rows in the destination table. In the preceding figure, the alarm condition **#{1_1}!=\$2_1** indicates that an alarm is generated when the number of rows in the source table is inconsistent with that in the destination table.

4. Click **Next** and set subscription parameters.



NOTE

If you enable notification, **Alarm triggered** indicates that a notification is sent to the SMN topic when an alarm is generated for the job, and **Run successfully** indicates that a notification is sent to the SMN topic when no alarm is generated for the job.

5. Click **Next** and set scheduling parameters.

NOTE

Once indicates that the job needs to be manually executed, and **On schedule** indicates that the job is executed automatically based on your configuration. The configuration in the preceding figure indicates that the job is automatically executed every 15 minutes on Oct 27, 2020.

6. Click **Submit**.

Step 3 View the comparison job.

1. In the comparison job list, locate the created job and click **Run** in the **Operation** column.
2. On the displayed **O&M** page, locate the row that contains the comparison job and click **Details** in the **Operation** column to view the running results and logs.

----End

Analyzing the Comparison Result

In the running result, the left pane displays the execution result of the rule for source table rows, and the right pane displays the execution result of the rule for destination table rows.

The error rate indicates the difference between the number of rows of the source and destination tables. If the error rate is 0, the source and destination tables have the same number of rows.

Source Settings		Destination Settings		Comparison Result	
Rule Type	Data Connection	Rule Type	Data Connection	Result Data	
Table rule	spe1027	Table rule	ddest1	总行数	
Data Object	Export (up to 10,000 records can be exported)	Data Object	Export (up to 10,000 records can be exported)	Error Value	Error Rate
Name	源行数	Name	目标行数	0	0%
Value	3	Value	3		
Template	Table rows	Template	Table rows		
Alarm Condition	{Source}Table Rows<{Destination}Table Rows	Alarm Condition	{Source}Table Rows<{Destination}Table Rows		

8 DataArts Catalog

This module provides enterprise-class metadata management to clarify information assets. It uses a data map to display a data lineage and panorama of data assets for intelligent data search, operations, and monitoring.

8.1 Data Maps

8.1.1 Overview

Data map facilitates data search and powers data analysis, development, mining, and operations. With data map, you can search for data quickly and make lineage and impact analysis with ease.

- Before data analysis, a data map can be used to search for keywords to narrow down the scope of data to be analyzed.
- A data map can be used to query table details by table names, letting you know how to use a table.
- Through lineage analysis, a data map displays you how a table is generated and where it is applied, and the logic used for processing table fields.

8.1.2 Overview

The **Dashboard** page displays data assets and asset reports. You can query information such as the asset quantity and size, as well as databases and tables classified by data connections.

Prerequisites

- A data connection has been created. For details on how to create a data connection, see [Creating Data Connections](#).
- A collection task has been created. For details on how to create a collection task, see [Creating a Collection Task](#).
- The collection task has been successfully executed. For details on how to view the status of a collection task, see [Monitoring Collection Tasks](#).

Procedure

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Catalog**.
2. Choose **Data Map > Dashboard** from the left navigation bar.
3. On the **Assets** tab page, click **Logical Assets** to display the number and details of business objects, logical entities, and business attributes.
4. Click **Technical Assets** to display the number and details of databases and tables, and data volumes.
5. Click **Metrics Assets** to display the metrics and their details.

Tags and Classifications

Tags are highly related keywords that help you classify and describe assets for easy retrieval.

Classification is the process of categorizing assets by category, level, or nature. Classification is top-down. Assets are classified according to certain standards.

The table below lists the differences between tags and classifications.

Table 8-1 Differences between tags and classifications

Item	Classification	Tag
Exclusiveness	Yes	No
Relationship	Dependent	Relevant (associated)
Creation	Pre-event planning	Any time
Cost	High	Low

8.1.3 Data Catalogs

You can use data catalogs to search for and filter data assets; view asset details, lineages, and relationships; and add asset classifications and tags.

Searching for a Data Asset

An asset can be searched by its name, description, or attributes. Fuzzy search is supported.

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Catalog**.
2. Choose **Data Map > Catalog** from the left navigation bar.
3. Enter a keyword in the search box, and search for assets. The search results are listed below the search box. You can search for assets in either of the following ways:
 - By their names and description

- By their attributes, which are displayed on the asset details page

NOTE

- You can save the search criteria you set.
- You can import the search criteria you need.

Filtering an Asset

Assets can be filtered by the following criteria:

- Data connection: the data connection that your target asset uses.
- Type: the type of your target asset.
- Classification: the category that your asset is classified into.
- Tag: the tag that your asset includes.
- Security level: the security level of your target asset.

The following uses **type** as an example to demonstrate how to filter an asset.

Step 1 Select **Table** under **Types**. Table assets are displayed.

Step 2 In the **Types** area, **Table**, **Column**, **Database**, **Bucket**, and **ColumnFamily** are supported by default. If you select **All**, the system displays assets of all types.

----End

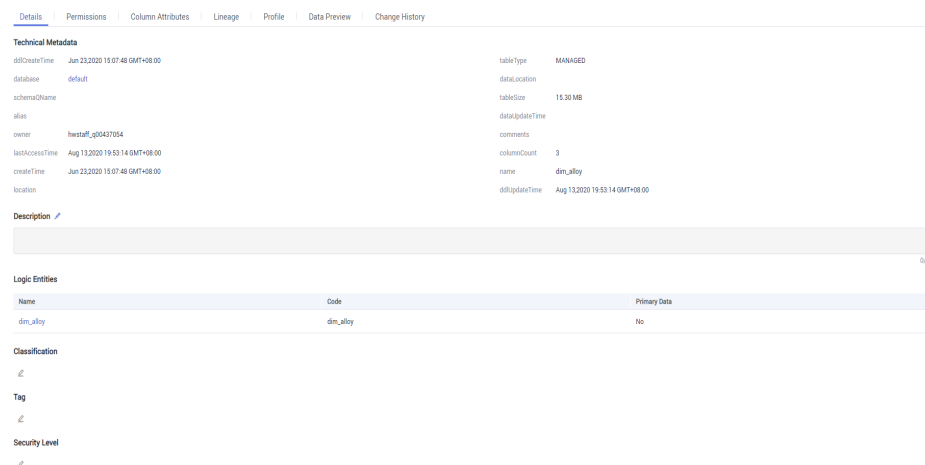
Viewing the Details of an Asset

The following uses a table as an example to demonstrate how to view the details of an asset.

Step 1 In the list of searched assets, select a table and click its name to access its details page.

Step 2 On the **Details** tab page, view the basic attributes of the technical metadata; edit the description; add or delete classifications, tags, and security levels for the table, table columns, or OBS objects.

Figure 8-1 Details tab page



- Step 3** On the **Column Attributes** tab page, view the column attributes of the table; add or delete classifications, tags, and security levels for the data columns; edit the description.

Figure 8-2 Managing column attributes

Identifier	Column	Type	Associated Business Attribute	Classification	Tag	Security Level	Metadata Description	Description
	name	string	name	ℓ	ℓ	ℓ		

- Step 4** On the **Lineage** tab page, view table lineages and impacts. For details on how to set a data lineage, see **Node Lineages**. After the lineage is set for a node, the node can be automatically parsed during job execution. Then, the metadata is collected during data asset collection and is displayed on the **DataArts Catalog** page.

- Step 5** On the **Summary** tab page, view the summary of the data table. Currently, only DWS and DLI data tables can be viewed.

Click **Update** to update the table profile.

- Step 6** On the **Data Preview** tab page, view the effect after data is masked.

- Step 7** On the **Change History** tab page, view the change history of the table.

----End

8.1.4 Tags

Tags are keywords used to identify the business meaning of data. They help you classify and describe assets for easy search.

Tags can be defined and associated with technical assets for better asset management. For example, you can tag a table as the SDI source data layer or DWI data integration layer.

Managing a Tag

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Catalog**.
2. Choose **Data Map > Tag Management** from the left navigation bar.
3. Click **Create** to create a tag.
 - **Tag Name:** Tag names can include only letters, numbers, and underscores (_). They cannot start with underscores (_) or exceed 100 characters.
 - **Description:** Up to 255 characters are allowed.
4. Select a tag and click **Delete** to delete the tag.
5. Click **Edit** to modify the description of a tag.

Adding a Tag to Identify Data

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Catalog**.
2. Choose **Data Map > Catalog** from the left navigation bar.

3. Enter a keyword in the search box, and click the search icon. The search results are listed below the search box.
4. Select the asset that you want to add a tag for and click **Add Identifier** in the upper right corner. In the **Add Identifier** dialog box, select **Tags** for **Type**.

Figure 8-3 Adding an identifier

Add Identifier ×

* Type Tags Security Levels Classifications

* Tag
If the tag to be added already exists, enter the tag name and press Enter. If the tag to be added does not exist, enter a tag name. After the entire page is submitted, the new tag is created successfully.

Name	Type
a	dli_column

OK Cancel

5. Set the parameters and click **OK**.

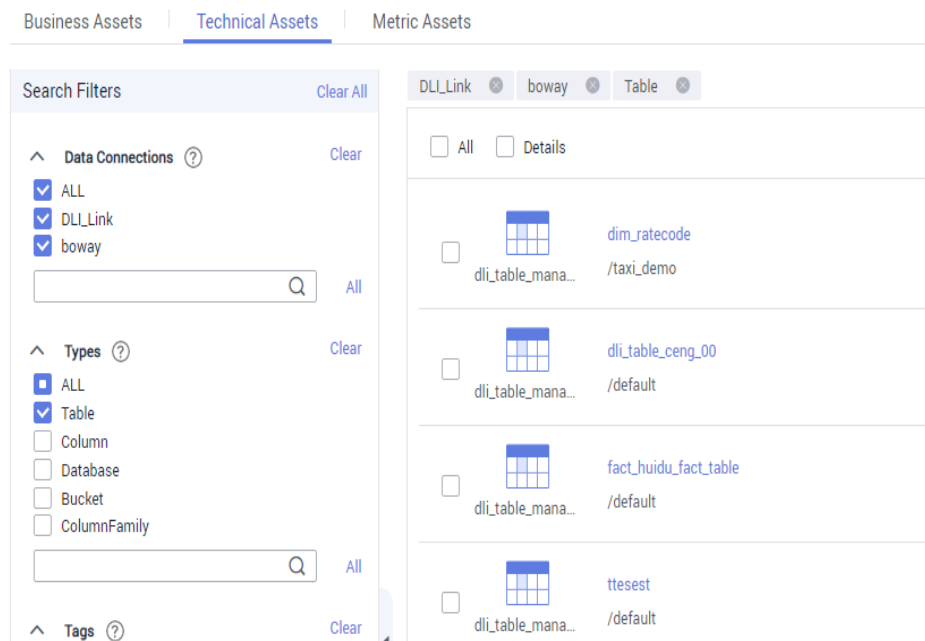
NOTE

You can add a new tag or select an existing tag. Existing tags are created by following instructions in [Managing a Tag](#).

Viewing the Details of a Table

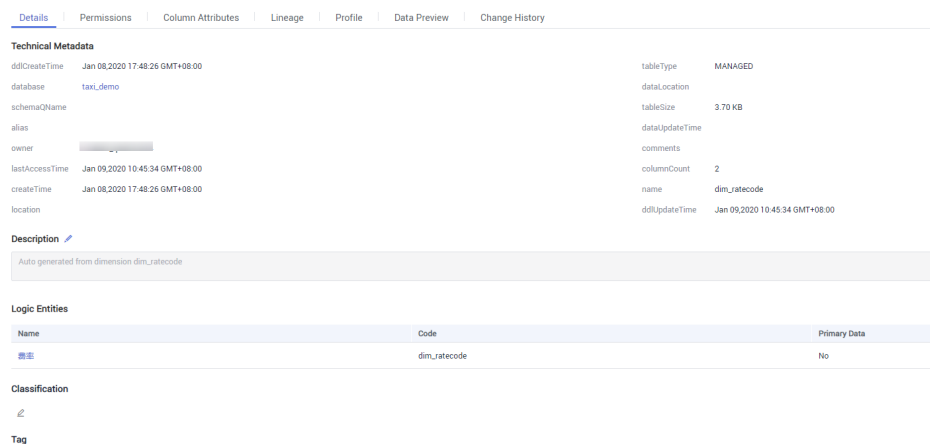
1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Catalog**.
1. Choose **More > DataArts Catalog** in the **Quick Entry** column of the corresponding workspace.
2. Choose **Data Map > Catalog** from the left navigation bar. On the **Technical Assets** tab page, select a data connection and set **Types** to **Table**. Filter all tables in the data connection. For details, see [Figure 8-4](#).

Figure 8-4 Filtering tables



3. Click a table name to access its details page. See [Figure 8-5](#).
The displayed page displays permissions information, column attributes, lineage information, table profile, data preview, and change history.

Figure 8-5 Table details page



Previewing Data

You can preview business data of a table on the **Data Preview** tab page. The data can be masked in real time based on the column classification information.

- Data assets that use DWS, DLI, MRS Hive, and MySQL data connections can be previewed.
- Column classification information can be automatically set when a collection task is created or manually added in the data classification menu. Automatic classification setting is available only for DWS and DLI data collections.

8.2 Data Permissions

8.2.1 Overview

To ensure data security and controllability, you need to apply for permissions before using data tables. The **Permissions** module facilitates permission control, provides visualized application and approval processes, and supports for permission audit and management. Data is secure and data permission control is convenient.

The **Permissions** module consists of **Data Catalog Permissions**, **Data Table Permissions**, and **Review Center**. The provided functions are:

- Self-service permission application: You can select a data table and quickly apply for the needed permissions online.
- Permission audit: Administrators can quickly and easily view the personnel with the corresponding database table permissions and perform audit management.
- Permission revoking and returning: Administrators can revoke user permissions in a timely manner. Users can also proactively return unnecessary permissions.
- Permission approval and management: Visualized and process-based management and authorization mechanism facilitates post-event tracing.

8.2.2 Data Catalog Permissions

You can manage data catalog permissions.

Constraints

- Only workspace admins can create, delete, and modify data catalog permissions rules and set the permissions effective status.
- Workspace developers, operators, and viewers can only view data permissions.

Managing a Data Catalog Permissions Rule

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Catalog**.
2. Choose **Permissions > Data Catalog Permissions** from the left navigation bar, and click **Create** on the page displayed to configure a data catalog permissions rule.
 - a. **Rule**: Name of a data catalog permissions rule.
 - b. **Type**: Currently, only **Tag**, **Security level**, and **Classification** can be used for filtering.
 - c. **Scope**: Select available tags, security levels, and classifications.
 - d. **User**: User to whom the configured data catalog permissions rule applies.
 - e. **Validate**: If this function is enabled, the data catalog permissions rule takes effect. Otherwise, the rule does not take effect.

NOTE

After a data catalog permissions rule takes effect, only users to whom the configured data directory permissions rule applies can manage data assets with specified tags or classifications. For example, if **Type** is set to **Tag**, **Scope** is set to **test**, and **User** is set to **A**, user A can manage assets with tag **test** after the permissions rule is enabled.

Figure 8-6 Creating a rule

★ Rule

★ Type

★ Scope

★ User

Validate

Description

0/255

3. In the data catalog permissions rule list, click **Edit** or **Delete** in the **Operation** column to modify or delete the rule.

8.2.3 Data Table Permissions

On the **My Permissions** page, you can view your table and column permissions in the workspace, and apply for or return the permissions.

Workspace admins have the permissions to manage user permissions. An admin can view the resource permissions of all users in the workspace.

Applying for Table or Column Permissions

NOTE

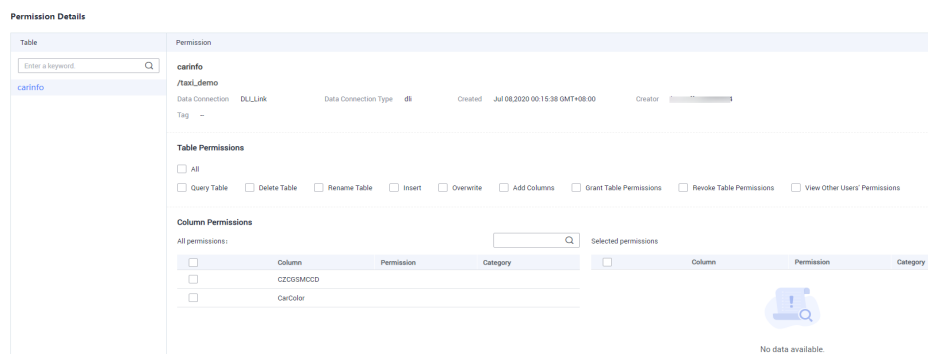
- The current version supports permissions control only on DLI data tables.
 - The table or column permissions you applied for take effect only after being approved by reviewers. Therefore, before applying for the permissions, create a reviewer by referring to [Managing Reviewers](#).
1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Catalog**.
 2. Choose **Permissions > Data Table Permissions** from the left navigation bar. On the **My Permissions** tab page, click **Apply**.
 3. On the page displayed, describe the scenario where the permissions are required, and select the data connection, database, and data table.
 4. Select the table or column permissions you want to apply for.
 - Applying for the permissions of a single table or column

Select the table or column permissions that you do not have but need to use.

- Applying for the permissions of multiple tables or columns

After selecting multiple tables, select the table or column permissions to be used in the **Permission Details** area.

Figure 8-7 Applying for permissions on tables and columns



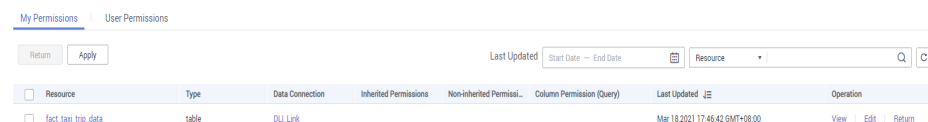
5. Click **OK**. Configure a reviewer and click **OK**.
6. Wait for the reviewer to approve the application. After the application is approved, the permissions take effect.

Managing Existing Table Permissions

You can manage the table or field permissions you already have, including viewing, editing, and returning permissions.

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Catalog**.
2. Choose **Permissions > Table Permissions**. On the **My Permissions** tab page, you can perform the following operations:
 - Click **View** in the **Operation** column to view the permissions details.
 - Click **Edit** in the **Operation** column to modify table permissions as needed.
 - Click **Return** in the **Operation** column to return table permissions as needed.

Figure 8-8 Managing table permissions



Auditing User Permissions

On the **User Permissions** tab page, admins can view the accounts that have permissions on tables and fields in the same workspace, reclaim the table and field permissions, or grant permissions to users in batches.

 **NOTE**

Only workspace admins can audit user permissions, including viewing the user list, reclaiming user permissions, or granting permissions to users.

- Viewing accounts with table permissions and the corresponding asset list
On the **User Permissions** tab page, view the accounts with applied permissions.

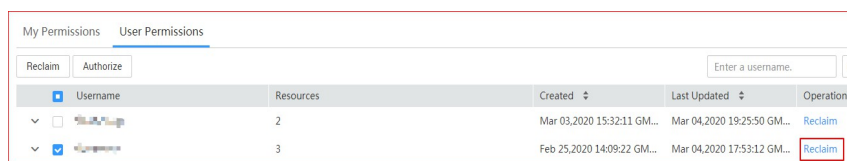
Figure 8-9 Viewing accounts with table permissions



Username	Resources	Created	Last Updated	Operation
<input type="checkbox"/>	1	Mar 03,2020 15:32:11 GM...	Mar 04,2020 17:31:28 GM...	Reclaim
<input type="checkbox"/>	3	Feb 25,2020 14:09:22 GM...	Mar 04,2020 17:53:12 GM...	Reclaim

- Reclaiming user permissions
 - On the **User Permissions** tab page, click **Reclaim** in the **Operation** column to the right of the account to reclaim all its permissions.
 - On the **User Permissions** tab page, select the check boxes to the left of one or more usernames, and click **Reclaim** in the upper left corner to revoke their permissions in batches.

Figure 8-10 Reclaiming user permissions



Username	Resources	Created	Last Updated	Operation
<input type="checkbox"/>	2	Mar 03,2020 15:32:11 GM...	Mar 04,2020 19:25:50 GM...	Reclaim
<input checked="" type="checkbox"/>	3	Feb 25,2020 14:09:22 GM...	Mar 04,2020 17:53:12 GM...	Reclaim

- Granting permissions to users

Figure 8-11 Authorization



Username	Resources	Created	Last Updated	Operation
<input checked="" type="checkbox"/>	2	Mar 03,2020 15:32:11 GM...	Mar 05,2020 10:21:55 GM...	Reclaim
<input checked="" type="checkbox"/>	4	Feb 25,2020 14:09:22 GM...	Mar 05,2020 10:21:55 GM...	Reclaim

- Managing user permissions
On the **User Permissions** tab page, click the drop-down arrow to the utmost left of an account to display the assets of the user. Click **View**, **Edit**, and **Return** in the **Operation** column to the right of a specific resource as required.

Figure 8-12 Managing user permissions

Resource	Type	Data Connection	Inherited Permissions	Non-inherited Permissions	Column Permission (Query)	Last Updated	Operation
fact_taxi_trip_data	table	DLLink		ALL		Mar 18,2021 17:46:42 GMT+08:00	View Edit Return
shop	table	DLLink		ALL		Jan 15,2021 16:22:57 GMT+08:00	View Edit Return

8.2.4 Review Center

Constraints

Only workspace admins can manage reviewers, including creating and deleting reviewers.

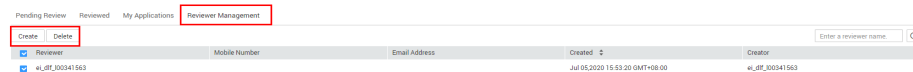
Approval Management

On the **Review Center** page, you can view the application status, applications to be approved, and approved applications, and manage reviewers.

- Reviewer management

Choose **Permissions** > **Review Center** from the left navigation bar. On the **Reviewer Management** tab page, create and delete reviewers as required. See [Figure 8-13](#). The reviewer data refers to the person added in the workspace.

Figure 8-13 Managing reviewers



- Pending review
 - a. Choose **Permissions** > **Review Center** from the left navigation bar, and click the **Pending Review** tab.
On this page, you can view the applications that need to be approved.
 - b. Click **Review** in the **Operation** column to view the application details and approve the application.
 - c. After entering the approval comments, approve or reject the application based on the actual situation.
- Reviewed applications
 - a. Choose **Permissions** > **Review Center** from the left navigation bar, and click the **Reviewed** tab.
 - b. Click **View Details** in the **Operation** column to view the approval records and application content.
- My Applications
 - a. Choose **Permissions** > **Review Center** from the left navigation bar, and click the **My Applications** tab.
 - b. Click **View Details** in the **Operation** column to view details about an application.
 - c. Click **Retry** in the **Operation** column to re-authorize an application.

8.3 DataArts Security (to Be Brought Offline)

8.3.1 Overview

Background

Data security provides data lakes with unified data usage protection capabilities throughout the data lifecycle. Sensitive data identification, classification, privacy protection, resource permission control, encrypted data transmission, encrypted storage, data risk identification, and compliance audit help users establish a security warning mechanism and enhance the overall security protection capability, to ensure data security.

NOTE

In regions where the DataArts Security module is available, DataArts Security, rather than DataArts Catalog, provides functions to ensure data security.

Functional Module

Data security includes:

- Data security levels
You can classify your data into different levels to facilitate data management.
- Data classification rules
You can classify data to effectively identify sensitive data in databases.
- Masking policies
Based on the data classification, you can create masking policies to mask data assets and protect privacy.

8.3.2 Data Security Levels

You can manage data security levels, including creating and deleting security levels and adjusting their ranking sequences.

You can create a data classification rule and data masking policy only after you have created a data security level.

Prerequisites

None

Accessing the Data Security Levels Page

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Catalog**.
2. Choose **DataArts Security > Security Levels** from the left navigation bar. On the page displayed, you can create, delete, edit, move up, and move down data security levels as required.
 - Creating a security level: Click **Create** in the upper left corner of the **Data Security Levels** page and enter the name and description.
 - Deleting a security level: Select unnecessary security levels and click **Delete** in the upper left corner of the **Security Levels** page.

- Adjusting the ranking sequence of a security level: Click **Up** or **Down** to the right of a security level to adjust its sequence.

8.3.3 Data Classifications

You can create data classification rules.

You can create a data masking policy to mask data only after you have created a data classification rule.

Prerequisites

A data security level has been created. For details, see [Data Security Levels](#).

Creating a Data Classification Rule

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Catalog**.
2. Choose **DataArts Security > Classifications** from the left navigation bar. On the **Classification Rule** tab page, click **Create**.

On the page displayed, set the parameters to create a data classification rule. You can either create a rule by using a system template or custom template.

Figure 8-14 Creating a data classification rule

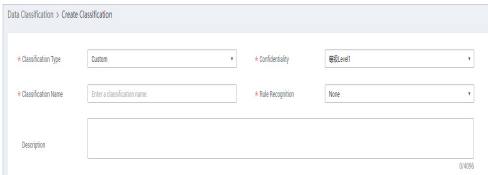


Table 8-2 Parameters for creating a data classification rule

Parameter	Description
Classification Type	The category to which a rule belongs. You can either create a rule by using a system template or custom template.
Confidentiality	Classify the configured data into different levels. If the existing confidentiality does not meet the requirements, go to the confidentiality management page to set security levels. For details, see Data Security Levels .
Classification Template	This parameter is available when Classification Type is set to Built-in . You can select a system sensitive data identification template based on service requirements, for example, Time , Mobile number , and License plate number .

Parameter	Description
Classification Name	<ul style="list-style-type: none">If Classification Type is set to Built-in, a classification name is automatically generated based on the classification template selected.If Classification Type is set to Custom, you can customize a classification name. NOTE The name of a data classification rule must be unique.
Rule Recognition	This parameter is available when Classification Type is set to Custom . Regular expressions are supported.
Regular Expression	<ul style="list-style-type: none">Content recognition: You can customize a regular expression.Column name recognition: Both exact match and fuzzy match are supported. Multiple fields can be matched.
Description	A description of the data classification rule to create.

Creating a Group

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Catalog**.
2. Choose **DataArts Security > Classifications** from the left navigation bar. On the **Groups** tab page, click **Create**.

In the **Create Group** dialog box, set the parameters and click **OK**.

Set the parameters by referring to [Table 8-3](#) and select classification rules in the list.

The selected rules are displayed in the list on the right.

Table 8-3 Parameters for creating a group

Parameter	Description
Name	The name of a group. Only letters, numbers, and underscores (_) are allowed.
Description	Information to better identify the group. It cannot exceed 4,096 characters.

8.3.4 Masking Policies

You can create a data masking policy and perform masking query in DataArts Catalog.

Prerequisites

- A data classification rule has been created. For details on how to create a classification rule, see [Data Classifications](#).
- A data connection and a data table have been created, and sensitive data has been collected by DataArts Catalog.

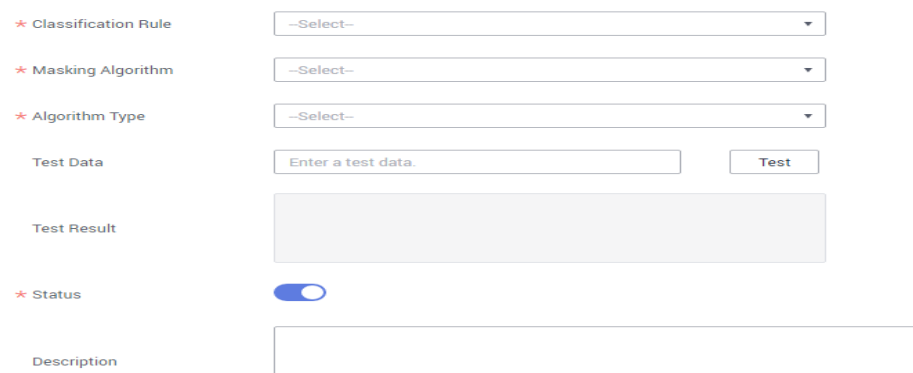
Creating a Masking Policy

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Catalog**.
2. Choose **DataArts Security > Masking Policies** from the left navigation bar, and click **Create** on the page displayed.
3. Set **Classification Rule**, **Masking Algorithm**, and **Algorithm Type**. The options for **Masking Algorithm** include **Mask**, **Truncate**, and **Hash**. Each masking algorithm has multiple algorithm types. Select an algorithm type as required. After the configuration, click **OK**.

NOTE

A data classification rule can be bound to only one masking algorithm.

Figure 8-15 Creating a masking policy



* Classification Rule

* Masking Algorithm

* Algorithm Type

Test Data

Test Result

* Status

Description

4. After you configured the making algorithm, you can perform an online test. Enter the test data, and click **Test**. You can verify the result in the **Test Result** text box.
5. Enable or disable **Status**. The masking policy takes effect only when **Status** is enabled.

Viewing the Data Masking Effect

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Catalog**.
2. Choose **Data Map > Catalog** from the left navigation bar.
3. In a list of asset results, click a table name to access its details page.
4. Click **Data Preview** to view the data masking effect.

8.4 Metadata Collection

8.4.1 Overview

Metadata is data about data. Metadata streamlines source data, data warehouses, and data applications, and records the entire process from data generation to data consumption. Metadata mainly refers to model definitions in the data warehouse and mappings between layers. It also describes the monitoring data status of the data warehouse and running status of ETL tasks. In the data warehouse system, metadata helps data warehouse administrators and developers easily locate the data they are looking for, improving the efficiency of data management and development.

Metadata is classified into technical metadata and business metadata by function.

- Technical metadata is data that stores technical details of a data warehouse system and is used to develop and manage data warehouses.
- Business metadata describes data in a data warehouse from the business perspective. It provides a semantic layer between users and actual systems, enabling business personnel who do not understand computer technologies to understand data in the data warehouse.

The metadata management module is the cornerstone of data lake governance. It allows you to create collection tasks by custom collection policies to collect technical metadata from data sources, customize business metamodels to batch import business metadata, associate business metadata with technical metadata, and manage and apply linkages throughout the entire link.

8.4.2 Task Management

You can create collection tasks by configuring metadata collection policies. Different types of data sources require different collection policies. Metadata management allows you to collect technical metadata using the configured collection policies.

Prerequisites

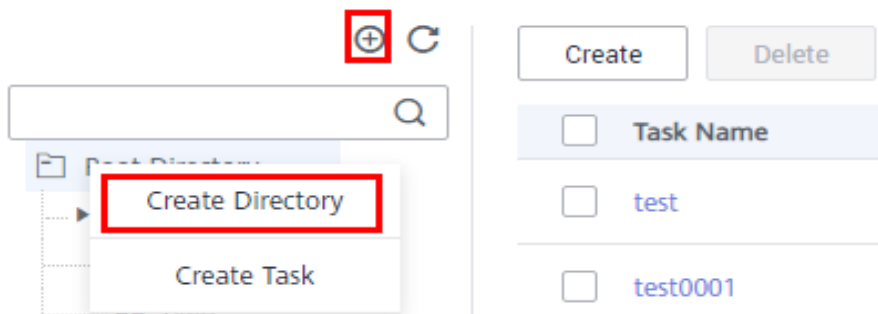
- Metadata of the following types of data sources can be collected: DWS, DLI, MRS HBase, MRS Hive, RDS (MySQL), RDS (PostgreSQL), and Oracle. To obtain metadata, you must first create data connections in Management Center.
- Before you can collect the metadata of Hudi tables by collecting the MRS Hive metadata, you must enable synchronization of the Hive table configuration for Hudi tables.

Creating a Collection Task

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Catalog**.
2. Choose **Metadata Collection > Task Management** from the left navigation bar.

3. Select the directory for the collection task. If no directory is available, create one as **Figure 8-16** shows.

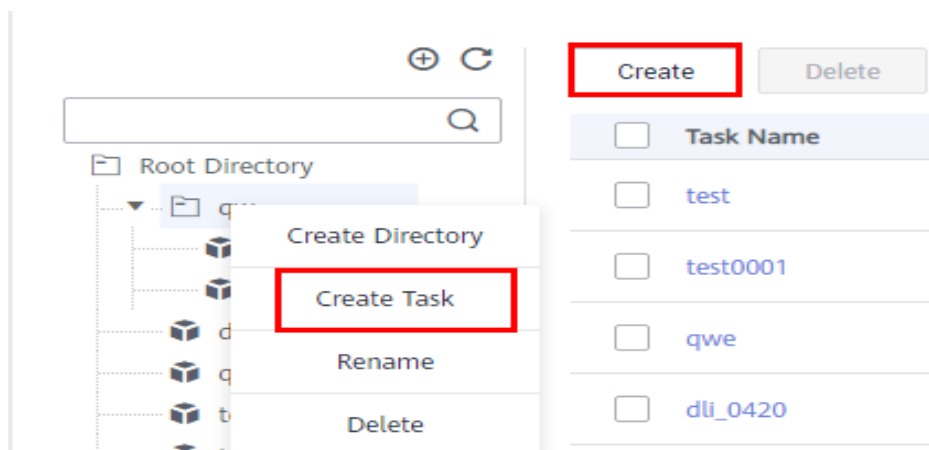
Figure 8-16 Directory that stores the collection task to create



4. Click **Create** in the upper part of the displayed page or right-click **Task name** and choose **Add Task** from the shortcut menu. On the page displayed, set the parameters.

Figure 8-17 shows the entries for creating a task.

Figure 8-17 Entries for creating a collection task



- a. Set the basic configuration based on **Table 8-4**.

Table 8-4 Basic configuration parameters

Parameter	Description
Task Name	Name of a collection task. The value can contain only letters, numbers, and underscores (_), and cannot exceed 62 characters.
Description	Information to better identify the collection task. Length of the description cannot exceed 255 characters.
Select Directory	The directory that stores the collection task. You can select an existing one. Figure 8-16 shows the directory.

- b. Configure data source information based on [Table 8-5](#).

Table 8-5 Data source parameters

Parameter		Description
Data Connection Type		<p>Select a data connection type from the drop-down list box.</p> <p>NOTE Metadata of the following types of data sources can be collected: DWS, DLI, MRS HBase, MRS Hive, RDS (MySQL), RDS (PostgreSQL), and Oracle. To obtain metadata, you must first create data connections in Management Center.</p>
<ul style="list-style-type: none"> • DWS • DLI • MRS HBase • MRS Hive • ORACLE • RDS 	Data Connection Name	<ul style="list-style-type: none"> • To use an existing data connection, select a value from the drop-down list. • To use a data connection that does not exist, click Create to add one.
	Database (or Database and Schema and Namespace)	<p>Database, schema, or namespace and data table from which data will be collected</p> <ul style="list-style-type: none"> • Click Set next to Database (or Database and Schema or Namespace) to set the range of databases (or databases and schemas or namespaces) to be scanned by the collection task. If this parameter is not set, all databases (or databases and schemas or namespaces) under the data connection are scanned by default.
	Table	<ul style="list-style-type: none"> • Click Set next to Table to set the range of tables to be scanned by the collection task. If this parameter is not set, all tables in the database (or database and schema or namespace) are scanned by default. For data tables whose data connection type is Oracle, MySQL, DLI, the tables to be collected can be filtered by regular expressions. • If neither the database (or database and schema or namespace) nor the data table is set, the task scans all data tables of the selected data connection. • Click Clear to delete the selected database (or database and schema or namespace) and data table.
CSS	Cluster	<p>Select the CSS cluster for storing the data to be collected.</p> <p>You can also click Create to create a CSS cluster. After the CSS cluster is created, click Refresh and select the new CSS cluster.</p>

Parameter		Description
	CDM Cluster	Select the agent provided by the CDM cluster. You can also click Create to create an agent. After the agent is created, click Refresh and select the new agent.
	Index	Index, similar to "database" in the relational database (RDB), stores Elasticsearch data. It is a logical space that consists of one or more shards.
GES	Graph	Select graphs that store structured data based on "relationships".
	CDM Cluster	Select the agent provided by the CDM cluster. You can also click Create to create an agent. After the agent is created, click Refresh and select the new agent.
OBS	OBS Bucket	Select the OBS bucket from which data will be collected.
	OBS Path	Select the path of the OBS bucket from which data will be collected.
	Collection Scope	Select the range of data to be collected. <ul style="list-style-type: none"> If you select This folder, the collection task collects only the objects in the folder set in the OBS path. If you select This folder and subfolders, the collection task collects all objects in the folder set in the OBS path, including the objects in the sub-folders.
	Collected Content	Select the content of data to be collected. <ul style="list-style-type: none"> If you select Folders and objects, the collection task collects folders and objects. If you select Folders, the collection task collects only folders.

- c. Set parameters under **Metadata Collection**. See [Table 8-6](#).

 **NOTE**

Metadata collection parameters are available only for DWS, DLI, MRS HBase, MRS Hive, RDS, or Oracle connections.

Table 8-6 Parameters for metadata collection

Parameter	Description
The data source metadata has been updated.	<p>When metadata in a data connection changes, you can configure an update policy to set the metadata update mode in the data catalog.</p> <p>Note that the configured update and deletion policies apply only to the databases and data tables configured by yourself.</p> <ul style="list-style-type: none">• If you select Update metadata in the data directory only, the collection task updates only the metadata that has been collected in the data catalog.• If you select Add new metadata to the data directory only, the collection task collects only metadata that exists in the data source but does not exist in the data catalog.• If you select Update metadata in the data directory and add metadata, the collection task fully synchronizes metadata from the data source.• If you select Ignore the update and addition operations, the metadata in the data source is not collected.
The data source metadata has been deleted.	<p>When metadata in a data connection changes, you can configure a deletion policy to set the metadata update mode in the data catalog.</p> <ul style="list-style-type: none">• If you select Delete metadata from data directory, when some metadata in the data source is deleted, the corresponding metadata is also deleted from the data catalog.• If you select Ignore the deletion, when some metadata in the data source is deleted, the corresponding metadata is not deleted from the data catalog.

- d. Set parameters when **Data Summary** is selected. See [Table 8-7](#) for details.

 **NOTE**

- **Data Summary** parameters are available only for DWS, DLI, and OBS connections.
- You are advised not to select **Data Summary** unless necessary. Selecting this option will increase the SQL execution workload. As a result, the metadata collection task may take a longer time than expected.

Table 8-7 Parameters

Parameter	Description
Full data	If this option is selected, a data profile is generated in the data catalog based on all data collected. This mode applies to scenarios where the data volume is less than 1 million.
Sampled data, first <i>x</i> rows	If this option is selected, a data profile is generated in the data catalog based on all data collected. This mode is applicable to scenarios with a large amount of data.
Randomly collect <i>x</i> % records of data from all data	If this option is selected, a data profile is generated in the data catalog based on all data collected. This mode is applicable to scenarios with a large amount of data.
Data Lake Insight Queue	The queue used to obtain profile data and execute DLI SQL statements. If you select Collect unique value , the number of unique values in the collected table is calculated and displayed on the Profile tab page in the data catalog.
Data Format	If the data stored in the OBS bucket is in CSV format, you must determine whether to have a table header, whether to customize a delimiter, whether to customize a reference character, and whether to customize an escape character based on the actual data attributes.
Date Format	If the data stored in the OBS bucket is in CSV format, configure the date format based on the actual attributes to prevent data from being incorrectly parsed.
Timestamp Format	If the data stored in the OBS bucket is in CSV format, configure the timestamp format based on the actual attributes to prevent data from being incorrectly parsed.

- e. Set parameters when **Data Classification** is selected. (This option is available only when DataArts Catalog provides data security functions. The data classification cannot be associated with a sensitive data identification rule created in the independent DataArts Security module.)
- If you select **Data Classification** and create a classification rule group or select an existing classification rule group by referring to [Data Classifications](#), data will be automatically identified and a classification will be added.
 - If you select **Update the data table security level based on the data classification result**, the table security level must be the same as the highest security level of the matched classification rules.

- If you select **Manually** for **Synchronize Data**, classification rules and security levels are not automatically added to **Column Attributes** of **Data Catalog** under **Data Map**. Go to the **Task Monitoring** page. Locate the target instance and choose **More > View Scanning Result** to view the execution result of the collection task and check whether the classification result matches. Select the check box of the classification matching field and click **Synchronize** to manually synchronize the classification rule and security level.

 **NOTE**

Only when you choose the DWS or DLI data source, you can add data classifications for automatic data identification. In addition, you can add classification rules only for columns in the data tables and OBS objects.

5. Click **Next** and select a scheduling mode.

Once: If the execution duration of a task exceeds the configured timeout duration, the task is considered failed.

Repeating: See [Table 8-8](#) for details.

 **NOTE**

1. If **Once** is selected, a manual task instance is generated. A manual task has no dependency on scheduling and must be manually triggered.
2. If **Repeating** is selected, a periodic instance is generated. A periodic instance is an instance snapshot that is automatically scheduled when the scheduled execution time is arrived.
3. When a periodic task is scheduled once, an instance workflow is generated. You can perform routine O&M on scheduled instance tasks, such as viewing the running status, stopping and rerunning the scheduled tasks.

Table 8-8 Parameters

Parameter	Description
Scheduling Date	The period during which a scheduling task takes effect.
Scheduling Cycle	The frequency at which the scheduling task is executed, which can be: <ul style="list-style-type: none"> • Minutes • Hours • Days • Weeks
Start Time	Start time of periodic scheduling, which is used together with the start time in Scheduling Date .
Time Interval	Interval between two periodic scheduling operations A scheduling task instance starts even if the previous scheduling task instance has not ended. A collection task supports concurrent running of multiple instances.

Parameter	Description
End Time	End time of periodic scheduling, which is used together with the end time in Scheduling Date .
Timeout	Timeout duration for a task instance. If a task runs longer than the value of this parameter, the task fails to be executed.
Start	If this check box is selected, the task is scheduled immediately.



6. Click **Submit**. The collection task is created.

Managing a Collection Task

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Catalog**.
2. Choose **Metadata Collection > Task Management** from the left navigation bar.

Then, you can view all created collection tasks.

Table 8-9 Parameters for managing collection tasks

Parameter	Description
Task Name	The name of a collection task. Click a collection task name to view the collection policies and scheduling properties.
Type	The name of a data connection.
Scheduling Status	The scheduling status of a collection task. You can click  to view only tasks of the specified statuses.
Scheduling Cycle	The scheduling frequency of a collection task. You can click  to view only tasks of the specified frequencies.
Description	The description of a collection task.
Creator	The creator of a collection task.
Last Executed On	The last time when the collection task ran.

Parameter	Description
Operation	<p>You can perform the following operations on a created collection task:</p> <ul style="list-style-type: none"> • Edit: Modify the parameters that are closely related to the policies of collection tasks whose status is Started, Not started, or Failed. The data source type cannot be modified. • Run: Click Run to run a collection task and view its status and related logs on the Task Monitoring page. • Start Scheduling: When the scheduling status is Stopped, you can reschedule the task. • Stop Scheduling: When the scheduling status is Scheduling, you can stop the scheduling.

8.4.3 Task Monitoring

You can monitor the running status of metadata collection tasks, view collection logs, and perform operations such as rerunning collection tasks.

On the **DataArts Catalog** page, choose **Metadata Collection > Task Monitoring** in the left navigation pane. On the page displayed, monitor the created collection tasks. See [Table 8-10](#) for details.

Table 8-10 Parameters for monitoring a collection task

Parameter	Description
Task Name	The name of a collection task.
Instance Status	<p>The status of an instance (collection task), which can be:</p> <ul style="list-style-type: none"> • Successful • Partially successful • Executing • Failed • Running exception • Paused: Task monitoring is paused due to management plane upgrade. After the upgrade is complete, the monitoring will recover.
Schedule	The scheduling mode of the collection task. The options are Schedule once and Schedule periodically .
Time Interval	The scheduling period of the collection task.
Start Time	The time when the collection task restarts running.
End Time	The time when the collection task stops running.

Parameter	Description
Running Duration (min)	The duration that the collection task has run.
Operation	<p>The operations that can be performed on the collection task under monitoring:</p> <ul style="list-style-type: none">● Rerun: Instances whose statuses are Failed or Succeeded can be rerun.● View Log: You can view instance logs. <p>NOTE Click View Log to view the run logs of metadata collection, data summary, and data classification tasks in real time.</p> <ul style="list-style-type: none">● More > Cancel: You can perform this operation only when Manually is selected for Synchronize Data under Data Classification during the creation of the collection task. Instances whose statuses are Executing can be stopped.● More > View Scanning Result: You can perform this operation only when Manually is selected for Synchronize Data under Data Classification during the creation of the collection task. You can view the execution result of the collection task instance to check whether the classification result is matched. Select the check box of the classification matching field and click Synchronize to manually synchronize the classification rule and security level.

8.5 Tutorials

8.5.1 Developing an Incremental Metadata Collection Task

Configuring and running a collection task is the prerequisite for building data assets. This section describes how to create different types of metadata collection tasks.

Scenario 1: Adding Metadata Only

Create a collection task to collect new tables only.

For example, if table4 is newly added:

- Data table metadata before collection: table1, table2, table3
- Data table metadata after collection: table1, table2, table3, **table4**

If you only want to collect table 4 in the preceding figure, perform the following steps (on condition that table 1, 2, and 3 are already in DataArts Catalog):

Step 1 Access the **DataArts Catalog** module on the DataArts Studio console.

Step 2 In the navigation pane on the left, choose **Collection Tasks**.

Step 3 Click **Create**.

Step 4 Configure parameters for the task.

The screenshot shows the configuration interface for a data collection task. It is divided into two main sections: "Data Source Information" and "Metadata Collection".

Data Source Information:

- Data Connection Type:** A dropdown menu set to "MRS Hive". Below it is a note: "Select a data source. You can manage data from a wide range of sources, such as DWS, DLI, MRS HBase, MRS Hive, MySQL, and RDS. However, you need to create data connections in Management Center before creating a collection task."
- Data Connection Name:** A dropdown menu set to "test_hive_agent" with a "Create" link.
- Database:** A text input field set to "default" with "Set" and "Clear" buttons.
- Table:** A text input field set to "All" with "Set" and "Clear" buttons.

Metadata Collection:

- Update & Addition Policy:** Four radio button options:
 - Update metadata only
 - Add metadata only (highlighted with a red box)
 - Update and add metadata
 - Do not update or add metadata
- Deletion Policy:** Two radio button options:
 - Delete metadata
 - Do not delete metadata

Step 5 Click **Next** and set scheduling parameters.

The screenshot shows the "Scheduling Settings" configuration interface. It features a progress bar at the top with two steps: "1 Configure" and "2 Scheduling Settings", with the second step being active.

Schedule: Two radio button options: "Once" (selected) and "Repeating".

Timeout: A dropdown menu set to "1" and another dropdown menu set to "Hour".

Step 6 Click **Submit** to create a collection task.

Step 7 In the task list, locate the created task and click **Run** or **Start Schedule** in the **Operation** column to go to the **Task Monitoring** page and view the task status.

----End

Scenario 2: Updating Existing Metadata and Adding New Metadata

Create a collection task to collect all tables, including existing and new ones.

For example, if table4 is newly added:

- Data table metadata before collection: table1, table2, table3
- Data table metadata after collection: **table1, table2, table3, table4**

If you want to collect all tables in the preceding figure, perform the following steps:

Step 1 Access the **DataArts Catalog** module on the DataArts Studio console.

Step 2 In the navigation pane on the left, choose **Collection Tasks**.

Step 3 Click **Create**.

Step 4 Configure parameters for the task.

The screenshot shows two sections of a configuration form:

- Data Source Information:**
 - Data Connection Type: MRS Hive
 - Data Connection Name: test_hive_agent (with a 'Create' link)
 - Database: default (with 'Set' and 'Clear' buttons)
 - Table: All (with 'Set' and 'Clear' buttons)
- Metadata Collection:**
 - Update & Addition Policy:
 - Update metadata only
 - Add metadata only
 - Update and add metadata
 - Do not update or add metadata
 - Deletion Policy:
 - Delete metadata
 - Do not delete metadata

Step 5 Click **Next** and set scheduling parameters.

The screenshot shows the 'Scheduling Settings' section of the configuration form:

- Progress indicator: 1 Configure — 2 Scheduling Settings
- Schedule: Once Repeating
- Timeout: 1 (dropdown) Hour (dropdown)

Step 6 Click **Submit** to create a collection task.

Step 7 In the task list, locate the created task and click **Run** or **Start Schedule** in the **Operation** column to go to the **Task Monitoring** page and view the task status.

----End

Scenario 3: Updating Existing Metadata Only

Create a collection task to collect existing tables.

For example, if table4 is newly added:

- Data table metadata before collection: table1, table2, table3
- Data table metadata after collection: **table1, table2, table3**

If you want to collect table 1, 2, and 3 in the preceding figure, perform the following steps:

Step 1 Access the **DataArts Catalog** module on the DataArts Studio console.

Step 2 In the navigation pane on the left, choose **Collection Tasks**.

Step 3 Click **Create**.

Step 4 Configure parameters for the task.

The screenshot shows two sections of a configuration form. The top section, 'Data Source Information', includes a dropdown for 'Data Connection Type' (MRS Hive), a text input for 'Data Connection Name' (test_hive_agent), and two input fields for 'Database' (default) and 'Table' (All), each with 'Set' and 'Clear' buttons. The bottom section, 'Metadata Collection', has an 'Update & Addition Policy' with radio buttons for 'Update metadata only' (selected), 'Add metadata only', 'Update and add metadata', and 'Do not update or add metadata'. It also has a 'Deletion Policy' with radio buttons for 'Delete metadata' and 'Do not delete metadata' (selected).

Step 5 Click **Next** and set scheduling parameters.

The screenshot shows the 'Scheduling Settings' section of a configuration form. It features a progress bar with '1 Configure' and '2 Scheduling Settings'. Below, there is a 'Schedule' section with radio buttons for 'Once' (selected) and 'Repeating'. A 'Timeout' section has a dropdown menu set to '1' and another dropdown menu set to 'Hour'.

Step 6 Click **Submit** to create a collection task.

Step 7 In the task list, locate the created task and click **Run** or **Start Schedule** in the **Operation** column to go to the **Task Monitoring** page and view the task status.

----End

Scenario 4: Updating and Deleting Existing Metadata and Adding New Metadata

Create a collection task to delete existing tables.

For example, if table1 is deleted from the database:

- Data table metadata before collection: table1, table2, table3
- Data table metadata after collection: **table2, table3**

If you want to delete table1, perform the following steps:

Step 1 Access the **DataArts Catalog** module on the DataArts Studio console.

Step 2 In the navigation pane on the left, choose **Collection Tasks**.

Step 3 Click **Create**.

Step 4 Configure parameters for the task.

The screenshot shows two sections of a configuration interface. The top section, 'Data Source Information', includes a 'Data Connection Type' dropdown set to 'MRS Hive', a 'Data Connection Name' dropdown set to 'testhive_agent', and input fields for 'Database' (set to 'default') and 'Table' (set to 'All'). The bottom section, 'Metadata Collection', has two sub-sections: 'Update & Addition Policy' with radio buttons for 'Update metadata only', 'Add metadata only', 'Update and add metadata' (selected), and 'Do not update or add metadata'; and 'Deletion Policy' with radio buttons for 'Delete metadata' (selected) and 'Do not delete metadata'.

Step 5 Click **Next** and set scheduling parameters.

The screenshot shows the 'Scheduling Settings' section of the configuration interface. It features a progress bar with '1 Configure' and '2 Scheduling Settings'. Below the progress bar, there are two main settings: 'Schedule' with radio buttons for 'Once' (selected) and 'Repeating', and 'Timeout' with a dropdown menu set to '1' and a unit dropdown set to 'Hour'.

Step 6 Click **Submit** to create a collection task.

Step 7 In the task list, locate the created task and click **Run** or **Start Schedule** in the **Operation** column to go to the **Task Monitoring** page and view the task status.

----End

8.5.2 Viewing Data Lineages Through the Data Map

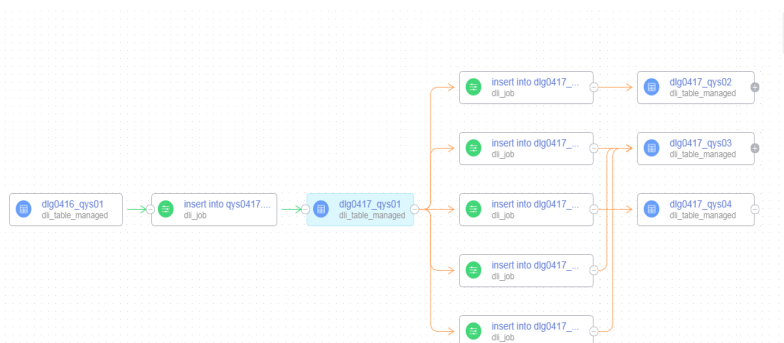
8.5.2.1 Overview

What Is Data Lineage?

In the era of big data, various types of data are rapidly generated due to explosive data growth. The massive and complex data information is converged, transformed, and transferred to generate new data and aggregate into an ocean of data.

During this process, a relationship is formed between the data, and these relationships are their lineages. They are analogous to the genetic relationships between people. However, in contrast from our human lineages, data lineages have the following distinct features:

- **Belongingness:** Specific data belongs to a specific organization or individual.
- **Multi-source:** One piece of data can have multiple sources. One piece of data may be generated by processing multiple pieces of data, and there may be multiple such processes.
- **Traceability:** The data lineage is traceable. It reflects the data lifecycle and the entire process from data generation to data disappearance.
- **Hierarchy:** The data lineage is hierarchical. Data classification and summary form new data, and different levels of description result in data layers.

Figure 8-18 Data lineage example

How DataArts Studio Data Lineage Is Implemented

- Generation of data lineages:

On the DataArts Studio platform, data lineages are generated by configuring data processing and migration nodes in the DataArts Factory module. Currently, the system collects the lineages generated by static node configuration and the lineages on some node instances. For details, see [Automatic Lineage Analysis](#).

In addition, DataArts Studio allows you to manually configure lineages. If you do so, automatic lineage analysis does not take effect. For details, see [Manually Configuring a Lineage](#).

- Display of data lineages:

If you have configured data lineages and started job scheduling in the DataArts Factory module, you can start a metadata collection task in the DataArts Catalog module to view the data lineages.

8.5.2.2 Configuring Data Lineages

On the DataArts Studio platform, data lineages are generated by configuring data processing and migration nodes in the DataArts Factory module. Currently, the system collects the lineages generated by static node configuration and the lineages on some node instances. For details, see [Automatic Lineage Analysis](#).

In addition, DataArts Studio allows you to manually configure lineages. If you do so, automatic lineage analysis does not take effect. For details, see [Manually Configuring a Lineage](#).

Automatic Lineage Analysis

Data lineages can be parsed automatically if the job contains the following nodes:

- **SQL nodes**

DataArts Studio supports lineage parsing of DLI SQL, DWS SQL and MRS Hive SQL nodes. It supports multi-SQL parsing and column-level lineage parsing.

- **DLI SQL**

- Lineages generated by data insertion between DLI tables
- Lineages between OBS files generated by table creation statements and DLI tables

- **DWS SQL**
 - Lineages between DWS tables generated by DDL operations such as "Create table like/as"
 - Lineages between DWS tables generated by DML operations such as "Insert into"
- **MRS Hive SQL**
 - Lineages between MRS tables generated by DDL operations such as "Create table like/as"
 - Lineages between MRS tables generated by DML operations such as "Insert into/overwrite"
- **Data integration nodes**

Lineages of the CDM Job, ETL Job, and OBS Manager nodes can be parsed.

 - **CDM Job**

Lineages generated during table file migration between MRS Hive, DLI, RDS, CSS, DWS, and OBS
 - **ETL Job**

Data lineages generated by ETL tasks between DLI, OBS, MySQL, and DWS.
 - **OBS Manager**

Lineages generated by directory or file replication and migration between OBS buckets

 **NOTE**

A single SQL statement cannot contain semicolons (;).

Manually Configuring a Lineage

In DataArts Studio DataArts Factory, you can define the input and output lineage relationships of nodes. When you manually configure a lineage, automatic lineage analysis does not take effect. Manual lineage configuration does not affect job running.

Currently, DLI, DWS, Hive, CSS, OBS, and CUSTOM are supported as the input and output data sources during manual lineage configuration. CUSTOM indicates a custom type. When manually configuring a lineage, you can add data sources that are not supported as custom types.

The following nodes support manual lineage configuration:

- **CDM Job**
- **Rest Client**
- **DLI SQL**
- **DLI Spark**
- **DWS SQL**
- **MRS Spark SQL**
- **MRS Hive SQL**

- [MRS Presto SQL](#)
- [MRS Spark](#)
- [MRS Spark Python](#)
- [ETL Job](#)
- [OBS Manager](#)

8.5.2.3 Viewing Data Lineages


If you have configured data lineages and started job scheduling in the DataArts Factory module, you can start a metadata collection task in the DataArts Catalog module to view the data lineages.

Prerequisites

Data lineages have been automatically or manually configured. For details, see [Configuring Data Lineages](#).

Starting Job Scheduling

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**.

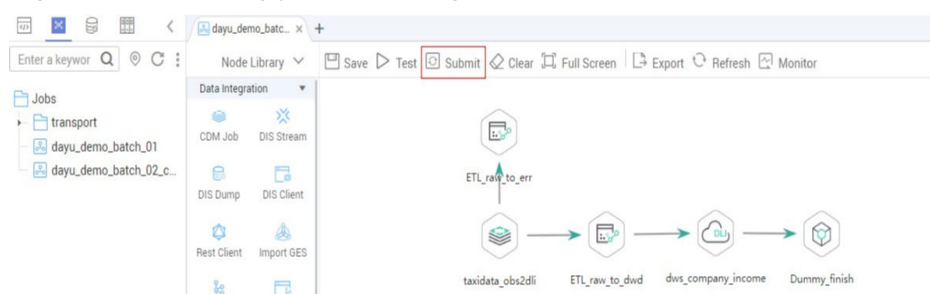
Step 2 In the navigation pane, click  and double-click the job for which lineages have been configured to open it.

Step 3 Click **Execute**. The system starts parsing lineages of the job.

NOTE

If you click **Test**, the system will not parse lineages of the job.

Figure 8-19 Starting job scheduling



----End

Creating a Metadata Collection Task

If a metadata collection task has been created, skip this part.

Step 1 On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Catalog**.

Step 2 Create a metadata collection task by following the instructions in [Task Management](#).

----End

Viewing Data Lineages

Step 1 On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts Catalog**.

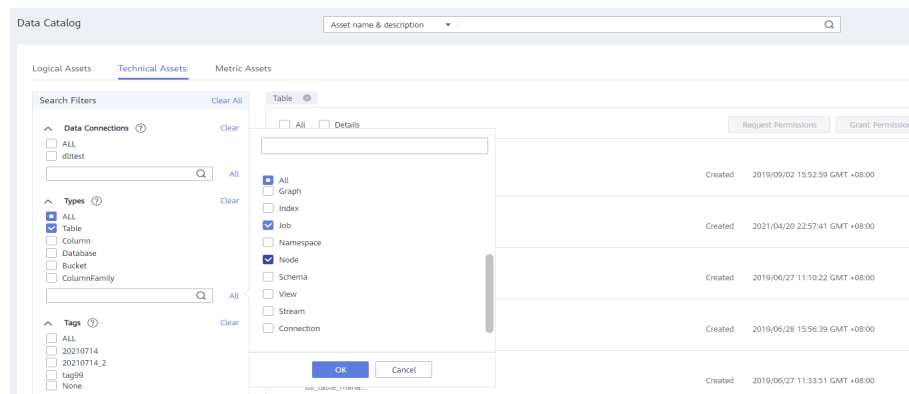
Step 2 In the navigation pane, choose **Data Catalog**. In the right pane, click the **Technical Assets** tab. On this page, you can query jobs, nodes, and tables.

In the **Types** area, click **All**, select **Job**, **Node**, and **Table**, and click **OK**.

NOTE

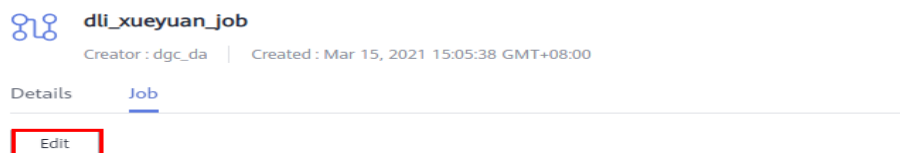
Jobs do not belong to any data connection. If you select a data connection in the search filters, no result will be returned.

Figure 8-20 Selecting types



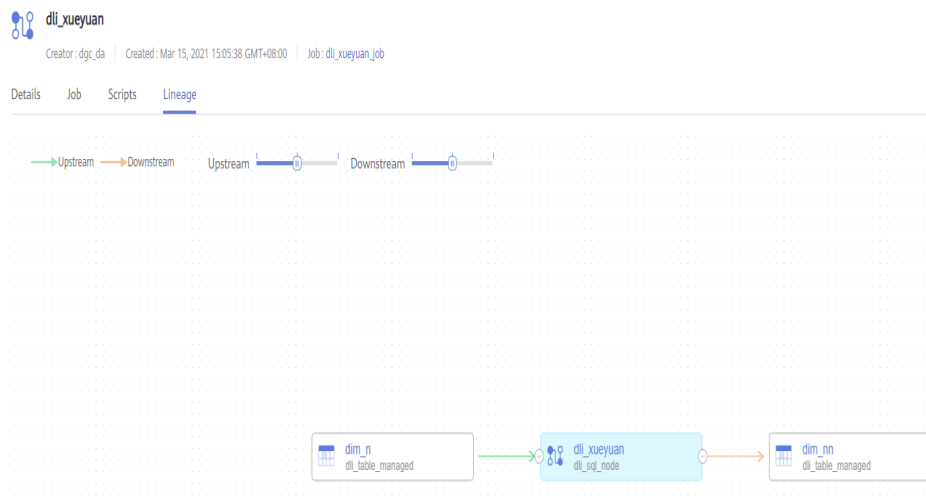
Step 3 In the search result, click the name of an asset ending with **_job** to view its details. On the job details page, click the **Job** tab and then **Edit** to go to the job editing page.

Figure 8-21 Viewing job details



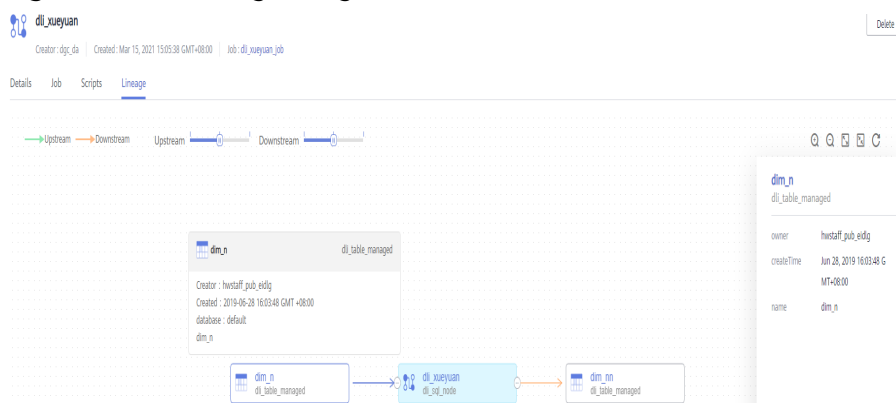
- Step 4** In the data asset search result, click the name of an asset ending with **_node** to view its details. On the node details page, you can view the node lineage information.
- Click the **+** or **-** icon beside the node to expand its upstream and downstream links.
 - Click a node to view the its details.
 - Click the **Job** tab and then **Edit** to go to the job editing page.

Figure 8-22 Viewing lineages of a node



- Step 5** In the data asset search result, click the name of an asset whose icon is a table to view its details. On the table details page, you can view lineages of the table.
- Click the **+** or **-** icon beside the table to expand its upstream and downstream links.
 - Click a table to view the its details.

Figure 8-23 Viewing lineages of a table



----End

9 DataArts DataService

9.1 Overview

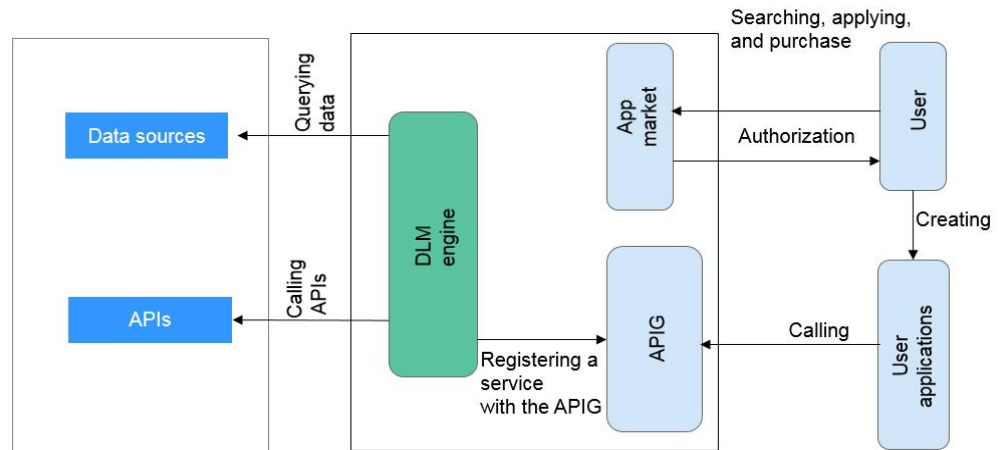
DataArts Studio DataArts DataService aims to build a unified data service bus for enterprises to centrally manage internal and external API services. DataArts DataService helps you quickly generate data APIs based on data tables and allows you manage the full lifecycle of APIs, covering API publishing, management, and O&M. With DataArts DataService, you can implement microservice aggregation, frontend-backend separation, and system integration, and provide functions and data for partners and developers easily and quickly at a low cost and risk.

DataArts DataService has the following advantages over other data sharing and exchange methods:

- Unified interface standards reduce the workload for interconnection with upper-layer applications.
- Data logic is deployed on the data platform and is therefore decoupled from the application logic. This reduces repeated development of data models and avoids frequent changes caused by data logic adjustment.
- Data logic-related storage and compute resources are deployed on the data platform, reducing resource consumption on applications.
- A large amount of detailed and sensitive data is inaccessible to applications. In addition, DataArts DataService improves data security by means of API review and publishing, authentication and throttling, and dynamic anonymization.

DataArts DataService encapsulates data logic into RESTful APIs of a unified standard that can be used to access data. DataArts DataService applies to quick response to the requests for accessing a small amount of data. To open a large amount of data, you are advised to adopt data sharing and exchange or other solutions. .

DataArts DataService uses the serverless architecture. You only need to focus on the API query logic and do not need to worry about the infrastructure such as the operating environment. DataArts DataService prepares compute resources, supports elastic scaling, and spares O&M expenditure.

Figure 9-1 DataArts DataService architecture

Publishing an API

To publish an API or a group of APIs, do as follows:

1. **Make preparations.**

If you want to use DataArts DataService, you must perform the operations in [an Exclusive DataArts DataService instance](#).

In addition, before creating an API, you must add a reviewer by following the instructions in [Adding Reviewers](#).

2. **Create an API.**

You can **generate** and **register** APIs. An API can be generated in the **wizard mode** or **script mode**.

3. **Debug the API.**

Debug the created API on the management console to check whether it runs properly.

4. **Publish the API.**

The API can be called only after it is published.

5. **(Optional) Manage the API.**

You can manage the published API as needed.

6. **(Optional) Perform throttling.**

To ensure the stability of backend services, you can perform throttling on the API.

Calling an API

To call an API, perform the following operations:

1. Obtain an API.

Obtain the API from the service catalog. An API can be called only after it is published.

2. (Optional) Create an application and get authorized.

For an API that is accessed using application or IAM authentication, you need to **create an application** and **authorize the application to use the API**. When you call an API, DataArts DataService verifies your identity based on the key pair (AppKey and AppSecret) of the created application.

3. **Call the API.**

After completing the preceding steps, you can call the API.

Overview Page

On the **Overview** page, you can view various monitoring data views. The **Overview** page displays **Develop APIs** and **Call APIs**.

Figure 9-2 Develop APIs tab page

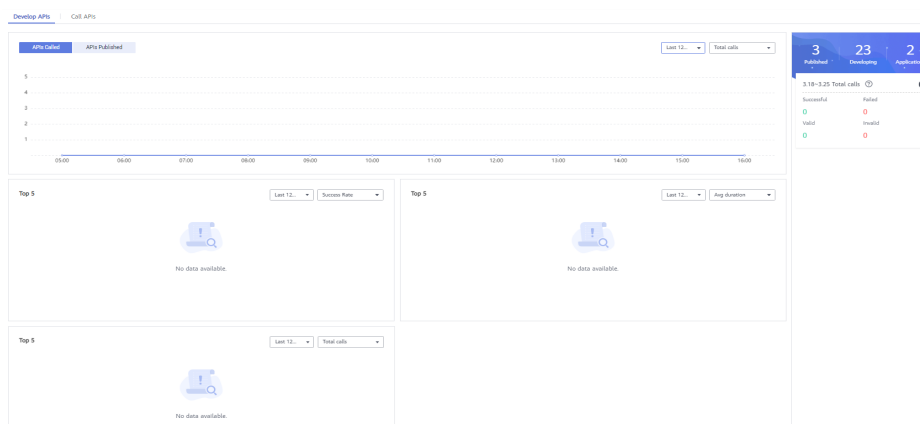


Table 9-1 Parameters on the Develop APIs tab page

Parameter	Description
APIs Published	The number of APIs published every day, week, month, and year.
APIs Called	The number of times that APIs are called in half a day, every day, every week, and every month.
Top 5 (1)	The call rate of APIs, including the success rate, failure rate, validity rate, and invalidity rate.
Top 5 (2)	The calling duration of APIs, average duration, success duration, and failure duration.
Top 5 (3)	The top 5 APIs that are called, successful API calls, failed API calls, valid API calls, and invalid API calls.
Published	The number of APIs that are published on the API marketplace.
Developing	The number of APIs that are being developed.
Applications	The number of APIs that are requested by applications.
Successful	The number of successful API calls.
Failed	The number of failed API calls.

Parameter	Description
Total	The total number of API calls.

Figure 9-3 Call APIs tab page

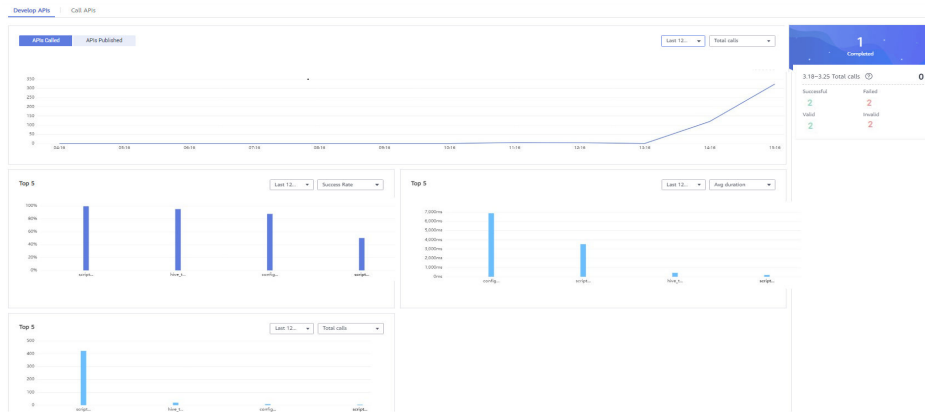


Table 9-2 Parameters on the Call APIs tab page

Parameter	Description
APIs Called	The number of API calls made every day, week, month, and year.
Top 5	The ratio of successful and failed API calls in the last seven days.
Completed	The number of APIs applied on the DataArts DataService platform.
Successful	The number of successful API calls on the DataArts DataService platform.
Total	The number of total API calls on the DataArts DataService platform.

9.2 Specifications

Specifications of Exclusive DataArts DataService

Table 9-3 lists the specifications of DataArts DataService Exclusive.

Table 9-3 Specifications of Exclusive DataArts DataService

Instance	Max. APIs That Can Be Published	Delay (Unit: ms)
Small	500	<20
Medium	1,000	<15

Instance	Max. APIs That Can Be Published	Delay (Unit: ms)
Large	2,000	<10

Specifications of API Return Data

DataArts DataService is applicable to interactions involving a small amount of data, and is not applicable to returning a large amount of data through APIs. The following table lists the specifications of the data returned by DataArts DataService APIs.

Table 9-4 Restrictions on the number of data records returned by an API

API Category	Scenario	Data Source	Default Specifications
Configuration	Debugging	DLI/ MySQL/RDS/DWS	10
	Call	DLI/ MySQL/RDS/DWS	100
Script	Test SQL	-	10
	Debugging	DLI	<ul style="list-style-type: none">• Default pages: 100• Custom pages: 1,000
		MySQL/RDS/DWS	<ul style="list-style-type: none">• Default pages: 10• Custom pages: 2,000
	Call	DLI	<ul style="list-style-type: none">• Default pages: 100• Custom pages: 1,000
		MySQL/RDS/DWS	<ul style="list-style-type: none">• Default pages: 10• Custom pages: 2,000

9.3 API Development

9.3.1 Preparations

9.3.1.1 an Exclusive DataArts DataService instance

This topic describes how to an exclusive DataArts DataService instance. You can create an API in Exclusive DataArts DataService and use it to provide services only after the instance is available.

NOTICE

To create or delete an exclusive cluster or change API quotas, you must have either of the following accounts:

- DAYU Administrator with the VPC Endpoint Administrator permission
- Tenant Administrator with the VPC Endpoint Administrator permission

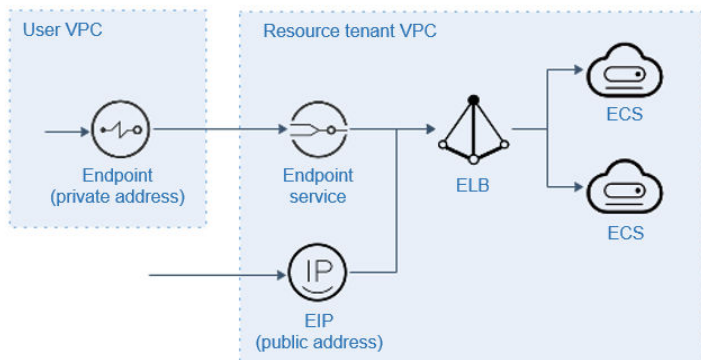
Network Environment Preparation

After a DataArts DataService exclusive cluster is created, resources are located in the resource tenant zone. ELB performs load balancing for the nodes in the cluster.

You can access the cluster in either of the following ways:

- Private address: IP address of the VPC endpoint
- Public address (optional): EIP bound to ELB The EIP is available only when you enable the Internet access when creating the DataArts DataService cluster.

Figure 9-4 Networking of the DataArts DataService exclusive cluster



To ensure that the created exclusive cluster is accessible, pay attention to the following network configurations:

- **Virtual Private Cloud (VPC)**
A VPC must be configured for an exclusive DataArts DataService instance. Resources (such as ECSs) in the same VPC can use the private address of the exclusive instance to call APIs.
When an exclusive instance, you are advised to configure the same VPC as other associated services to ensure network security and facilitate network configuration.
- **Elastic IP (EIP)**
If you want to call an API of an exclusive instance, buy an EIP and bind it to the instance. The EIP will be used as the Internet entry of the instance.
- **Security Group**
A security group is similar to a firewall. It controls who can access the specified port of an instance and enables the communication data flow of the instance to move to the specified destination address. You are advised to

enable the IP address and port in the inbound direction of the security group to protect the network security of the instance to the maximum extent.

The security group bound to an exclusive instance must meet the following requirements:

- Inbound rule: To call APIs from the Internet or from resources in other security groups, enable ports 80 (HTTP) and 443 (HTTPS) in the inbound direction of the security group bound to the exclusive instance.
- Outbound direction: If the backend service is deployed on the Internet or in another security group, enable the backend service address and API calling listening port in the outbound direction of the security group bound to the exclusive instance.
- If the frontend and backend services of the API are bound to the same security group and VPC as the exclusive instance, you do not need to enable the preceding ports for the exclusive instance.

- Route

In the physical machine management scenario, if the physical machine and the cluster have different network segments, you need to configure a route.

On the **Basic Details** page of the cluster, you can add or delete routes.

Basic Details		Nodes	Published APIs
Billing is not enabled for this cluster, and some functions are unavailable. Enable billing to use more functions			
Name	dlm-...	Cluster ID	...
AZ	cn-north-4a	Specifications	基础版 8CPUs 16GB
VPC	vpc-...	Subnet	s-...
Security Group	Sys-default	Nodes	2
Status	Abnormal	Version	2.3.1
Public IP		Private IP	
Creator	...	Created	Jun 23, 2021 14:48:47 GMT+08:00
Description		Expires	
Order Type	Pay-per-use	Order Period	
Dump Log	<input type="checkbox"/>		

NOTE

If the DataArts DataService cluster does not support routes, you can contact related support personnel to modify the configuration item **dlm.instance.route.action.support** to enable this function.

Procedure

a DataArts DataService incremental package. The system automatically creates a cluster based on your selected specifications.

Step 1 Locate an enabled instance and click .

Step 2 On the displayed page, set parameters based on [Table 9-5](#).

Table 9-5 Parameters for an exclusive DataArts DataService instance

Parameter	Description
Package	Select DataArts DataService .
Billing Mode	Currently, Yearly/Monthly is supported.
Workspace	The workspace for which you want to use the incremental package. For example, if you want to use DataArts DataService Exclusive in workspace A of the DataArts Studio instance, select workspace A. After an exclusive DataArts DataService cluster, you can view it in workspace A.
AZ	<p>When you buy a DataArts Studio instance or incremental package for the first time, you can select any available AZ.</p> <p>When you buy another DataArts Studio instance or incremental package, determine whether to deploy your resources in the same AZ based on your DR and network latency demands.</p> <ul style="list-style-type: none">• If your application requires good DR capability, deploy resources in different AZs in the same region.• If your application requires a low network latency between instances, deploy resources in the same AZ. <p>For details, see AZs.</p>
Name	N/A
Description	A description of the exclusive DataArts DataService cluster.
Version	Cluster version of the exclusive DataArts DataService cluster.
Cluster Details	The number of concurrent API requests supported varies depending on the instance specifications.
Enabling public IP address	If you select Enabling public IP address , external services can call the APIs created in exclusive instances through the Internet address.
Bandwidth	Bandwidth range on the Internet.
VPC	<p>A VPC is a secure, isolated, and logical network environment. Cloud resources (such as ECSs) within the same VPC can call APIs using the private IP address of DataArts DataService Exclusive.</p> <p>Deploy the DataArts DataService Exclusive instance in the same VPC as your other services to facilitate network configuration and secure network access.</p> <p>NOTE After the DataArts DataService instance is created, the VPC cannot be changed.</p>

Parameter	Description
Subnet	<p>A subnet provides dedicated network resources that are logically isolated from other networks for network security.</p> <p>Deploy the DataArts DataService Exclusive instance in the same subnet of the same VPC as your other services to facilitate network configuration and secure network access.</p> <p>NOTE After the DataArts DataService instance is created, the subnet cannot be changed.</p>
Security Group	<p>A security group is used to set port access rules, define ports that can be accessed by external services, and determine the IP addresses and ports that can be accessed externally.</p> <p>For example, if the backend service is deployed on an external network, configure security group rules to allow access to the IP address and listening port of the backend service.</p> <p>NOTE</p> <ol style="list-style-type: none">1. If Enabling the public IP address is selected, the security group must allow access from ports 80 (HTTP) and 443 (HTTPS) in the inbound direction.2. After the DataArts DataService instance is created, the security group cannot be changed.
Managing Cluster Resources Using an Enterprise Project	<p>Enterprise project associated with the exclusive DataArts DataService cluster. An enterprise project facilitates management of cloud resources. For details, see Enterprise Management User Guide.</p>
Nodes	N/A
Required Duration	N/A

Step 3 Click **Now**, confirm the settings, and click **Next**.

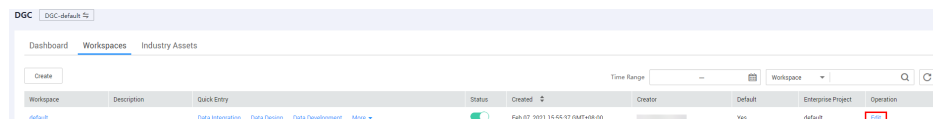
----End

Setting the Allocated API Quota

After creating an exclusive cluster, you need to set the allocated API quota so that you can create APIs. To set the quota, perform the following steps:

Step 1 On the **Workspaces** page, locate a workspace and click **Edit** in the **Operation** column.

Figure 9-5 Editing a workspace



Step 2 In the displayed **Workspace Information** dialog box, click **Edit** to set the allocated quota.

Figure 9-6 Setting the allocated quota

Workspace Information

* Name

Description

* Enterprise Project

Job Log Path

Dirty Data Path

* API Quota of DLM Exclusive Used: 0
 Allocated: 0
 Total used: 0
 Total allocated: 0
 Total: 0

NOTE

You will be charged for the APIs you create. If you increase the API quota, more APIs can be created in the workspace and the fees may increase.

Step 3 Set the allocated API quota for DataArts DataService Exclusive.

Figure 9-7 Setting the quota

* API Quota of DLM Exclusive Used: 1
 Allocated: 5
 Total used: 345
 Total allocated: 2,377
 Total: 5,000

NOTE

The allocated quota cannot be less than the used quota and not greater than the total quota minus the total allocated quota plus the previously allocated quota.

----End

9.3.1.2 Adding Reviewers

APIs must be reviewed and approved before they can be published. APIs are reviewed in the following way:

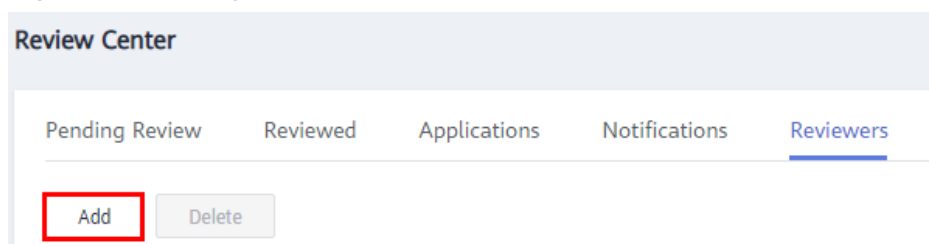
- An API publisher who does not have the reviewer permission must submit the API to the reviewer for review.
- An API publisher who has the reviewer permission can publish an API without review or approval. By default, a workspace administrator has the reviewer permission.

Therefore, if you do not have the reviewer permission and want to publish an API, you must add a reviewer first. Only the workspace admin has the permissions required to add reviewers.

Procedure

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. Choose **Operation Management > Review Center** from the left navigation pane. On the page displayed, choose **Reviewer Management** and click **Add**.

Figure 9-8 Adding reviewers



4. Select a reviewer (workspace member), enter a correct phone number and email address, and click **OK**.
5. Add more reviewers, if required.

9.3.2 Creating an API

9.3.2.1 Generating an API Using Configuration

This topic describes how to generate an API using configuration.


Generating data APIs using configuration is simple. You do not need to write any code. Wizard mode is designed for users who do not have high requirements on API functions or have no experience in code development.

Prerequisites

You have configured data sources on the **Data Connection Management** page of **Management Center**.

Creating an API Directory

An API catalog is an API index that is orchestrated and recorded in a certain sequence. It is a tool for reflecting categories, guiding API usage, and searching for APIs, helping API developers effectively classify and manage API services.


1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. Choose **API Development > API Catalogs** and click .
In the dialog box displayed, enter an API catalog name, and click **OK**.
4. In the **Operation** column of an API catalog, edit or manage the API catalog.
Click **Edit** to the right of the API catalog that you want to edit. An API can be edited only when it is in the **Created**, **Rejected**, **Offline**, or **Disabled** state.

Configuring Basic API Information

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. Choose **API Development > APIs** from the left navigation bar, and click **Create**. On the displayed page, enter the basic information.

Table 9-6 API basic configuration

Parameter	Description
API	An API name consists of 3 to 64 characters and starts with a letter. Only letters, numbers, and underscores (_) are allowed.
API Catalog	A collection of APIs for a specific function or scenario. It is similar to a folder and specifies the location of APIs. You can search for APIs in a specified API catalog. The API catalog is the minimum organization unit of APIs in DataArts DataService and also the minimum management unit in the API gateway. Click Select Catalog to create an API catalog or select an existing one created in Creating an API Directory .

Parameter	Description
Request Path	<p>API access path, for example, <code>/v2/{project_id}/streams</code>. It is the part between the domain name and query parameters in the URL of a request path, for example, <code>/blogs/xxxx</code> shown in the following figure.</p> <p>Figure 9-9 API access path in the URL</p>  <p>Braces ({}) can be used to identify parameters in a request path as wildcard characters. For example, <code>/blogs/{blog_id}</code> indicates that any parameter can follow <code>/blogs</code>. <code>/blogs/188138</code> and <code>/blogs/0</code> can both match <code>/blogs/{blog_id}</code>, and are processed by this API.</p> <p>In addition, duplicate request paths are not allowed for the same domain name. When a path parameter is used as a wildcard, the name is not unique. For example, <code>/blogs/{blog_id}</code> and <code>/blogs/{xxxx}</code> are considered as the same path.</p>
Parameter Protocol	<p>A protocol used to transmit requests. HTTP and HTTPS are supported.</p> <ul style="list-style-type: none"> • HTTP is a basic network transmission protocol. It is stateless, connectionless, simple, fast, and flexible, and uses plaintext for transmission. It is easy to use but has poor security. • HTTPS is an HTTP-based protocol with SSL or TLS encryption verification. It can effectively verify identities and protect data integrity. To access HTTPS APIs, you need to configure related SSL certificates or skip SSL verification.
Request Method	<p>HTTP request method, indicating the type of the requested operation, such as GET and POST. The method complies with the resultful style.</p> <ul style="list-style-type: none"> • GET requests the server to return specified resources. This method is recommended. • POST requests the server to add resources or perform special operations. This method is used only for API registration. The POST request does not have a body. Instead, it involves transparent transmission.
Description	A brief description of the API to create.
Tag	API tag. The tag is used to mark the API attributes. After the API is created, you can quickly search for the API by tag. A maximum of 20 tags can be set for an API.

Parameter	Description
Reviewer	An owner who has permissions to review APIs. Click Add to enter the Review Center page. On the displayed page, click Add on the Reviewer Management tab page to add a reviewer.
Security Authentication	Security authentication mode, which can be: <ul style="list-style-type: none"> ● App Authentication: API Gateway authenticates API requests. This mode has the highest security level. ● IAM Authentication: IAM authenticates API requests. This mode has a medium security level. ● No authentication: No authentication is required for accessing the API. This mode has a low security level, and is not recommended.
Display Scope	After the API is published, all users in the selected scope can view the API in the service catalog. <ul style="list-style-type: none"> ● Current workspace APIs ● Current project APIs ● Current tenant's APIs
Access Log	If you select this option, the API query result will be recorded and retained for seven days. You can choose Operations Management > Access Logs and select the request date to view the logs.
Min. Retention Period	Minimum retention period of the API publishing status, in hours. Value 0 indicates that the retention period is not limited. You can suspend, unpublish, or cancel authorization for an API only after the minimum retention period ends. The system notifies the authorized users. If all authorized users have processed the notifications or unbound the API from their apps, the API will be suspended or unpublished, or the API authorization will be canceled. Otherwise, the system will forcibly suspend, unpublish, or cancel authorization for the API when the minimum retention periods ends. For example, if the minimum retention period is set to 24 hours, the API can be suspended 24 hours after it is published. If the authorized user handles the notifications in the review center or unbind the API from the app, the API will be directly suspended. Otherwise, the API will be forcibly suspended when the minimum retention period ends.

Parameter	Description
Input Parameter	<p>Configure parameters in the API request. An input parameter consists of the parameter location, parameter type, whether the parameter is mandatory, and the default value.</p> <ul style="list-style-type: none"> • The parameter location can be Query, Header, Path, or Body. In addition, static parameters are supported. <ul style="list-style-type: none"> – Query is the query parameter following the URL. It starts with a question mark (?) and connects multiple parameters with &. – Header is located in the request header and is used to transfer current information, for example, host and token. – Path is a request parameter in the request path. If you configure a path parameter, you must also add this parameter to the request path. – Body is a parameter in the request body and is generally in JSON format. – Static is a static parameter that does not change with the value passed by API callers. The parameter value is determined upon API authorization. If the parameter value is not set during authorization, the default value of the API input parameter is used. • The parameter type can be Number or String. Number corresponds to numeric data types such as int, double, and long. String corresponds to text data types such as char, varchar, and text. • Mandatory and Default Value: If you select Yes for Mandatory, parameters must be passed for accessing the API. Otherwise, the default value of the parameter will be used if the parameter is not passed for accessing the API. <p>Constraints for the parameters are as follows:</p> <ul style="list-style-type: none"> • Query and Path: 32 KB. • HEADER: The maximum size is 128 KB. • BODY: The maximum size is 128 KB. <p>For instance, to configure the dynamic parameter project_id in the /v2/{project_id}/streams request path, do as follows:</p> <ol style="list-style-type: none"> 1. Click Add and enter project_id for Name. 2. Set Parameter Location to PATH. 3. Set Type to STRING. 4. Select Yes for Mandatory. 5. Leave Default Value blank.

4. After the basic API information is complete, click **Next** to go to the **Data Extract Logic** page.

Configuring the Data Extraction Logic

Set **Data Acquisition Method** to **GUI based**.

1. Select a data source, data connection, database, and data table to obtain the tables to be configured.

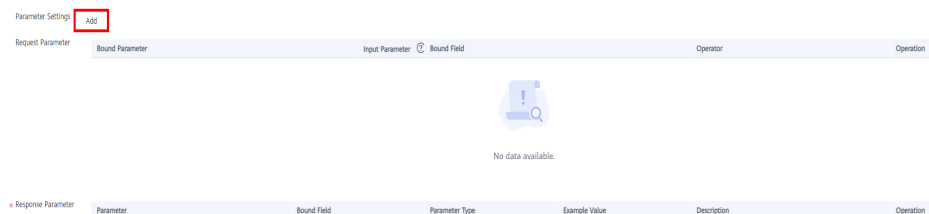
 **NOTE**

For details on the data sources supported by DataArts DataService, see [Data Sources](#). Configure data sources in Management Center in advance. You can search for a data table by name.

2. Configure parameter fields.

Click **Add** next to **Parameter Settings**. All fields in the table are displayed on the page for adding parameters. Select the request parameters, response parameters, and ranking parameters that you want to add to the corresponding lists.

Figure 9-10 Add Parameter dialog box



3. Edit request parameters.

A request parameter consists of a bound parameter, bound field, and operator. In the request parameter list, select a bound parameter and an operator.

- The bound parameter is directly used to access APIs.
- The bound field is the content that is accessed through an API call.
- The operator determines how to process parameters in the access request. The following table lists the available operators.

Table 9-7 Available operators

Operator	Description
=	Checks whether the values of two operands are the same. If yes, the condition is true.
<>	Checks whether the values of two operands are the same. If no, the condition is true.
>	Checks whether the value of the left operand is greater than that of the right operand. If yes, the condition is true.
>=	Checks whether the value of the left operand is greater than or equal to that of the right operand. If yes, the condition is true.

Operator	Description
<	Checks whether the value of the left operand is less than that of the right operand. If yes, the condition is true.
<=	Checks whether the value of the left operand is less than or equal to that of the right operand. If yes, the condition is true.
%like%	Ignores the prefix and suffix in character matching.
%like	Ignores the prefix in character matching.
like%	Ignores the suffix in character matching.
in	Compares a value with a specified list of values.
not in	Compares a value with values not in a specified list. It is the opposite of the in operator.

4. Edit response parameters.

A response parameter consists of the parameter name, bound field, and parameter type.

- The parameter names are returned by the API.
- The bound fields are the actual content returned by the API.
- The parameter type is the data display format when the API is called, and can be a numeric or character.

5. Edit ranking parameters.

A ranking parameter consists of the parameter name, field name, whether the parameter is optional, and ranking mode.

- The parameter names are returned by the API.
- The field names are the content that is accessed when the API is called.
- Whether a ranking parameter is optional determines whether the ranking condition can be removed. If it is selected, the parameter is optional.
- The ranking mode can be ascending, descending, or custom.

Click the buttons in the **Operation** column to move up, down, or delete parameters.

6. Click **Next** and set the value of **pre_order_by** as the description of all ranking parameters and separate them with semicolons (;).

Data in [Table 9-8](#) is used as an example.

Table 9-8 Ranking parameters

Parameter	Description
id	a:asc a is the parameter name. asc indicates that parameters are listed in the ascending order.

Parameter	Description
name	<ul style="list-style-type: none"> • b:asc • b • b:desc <p>b is the parameter name. asc and desc indicate that parameters are listed in ascending and descending orders, respectively. The ranking order can be customized or left blank.</p>
age	<p>c:desc</p> <p>c is the parameter name. desc indicates that parameters are listed in the descending order.</p>

Table 9-8 lists the ranking parameters. The following table describes how to configure **pre_order_by**.

Table 9-9 Configuring parameter pre_order_by

Parameter	Backend Statement	Remarks
a:asc;b:c:desc	order by id ASC, name, age DESC	N/A
b;c:desc	order by name, age DESC	a is optional and can be left blank.
b:asc;c:desc	order by name ASC; age DESC	b can be customized. b can be listed in the ascending order.
b:desc;c:desc	order by name DESC; age DESC	b can be customized. b can be listed in the descending order.

Figure 9-11 Setting ranking parameters

API NAME www
API PATH /getUserInfo/{userId}
Request Mode GET

Parameter	Type	Mandatory	Value
page_size (Default)	int(Default)	Y	<input type="text" value="10"/>
page_num (Default)	int(Default)	Y	<input type="text" value="1"/>
userId	NUMBER	Y	<input type="text"/>
pre_order_by	STRING	N	<input type="text" value="a.asc:b:c:desc"/>

 NOTE

- **pre_order_by** is optional. If **pre_order_by** is not set, the mandatory fields are used as the ranking criterion.
- If **pre_order_by** is set, configure an API following the parameter sequence.
Example:
If **pre_order_by** is set to **a:asc;b:c:desc**, no error will be generated. If **pre_order_by** is set to **b;a:asc;c:desc**, errors will be reported.

Testing the API

After setting and saving all parameters, click **Next**.

Specify **Value** and click **Debug**. You can view the **Request** and **Response** details on the right part of the displayed page. If the test fails, follow the instructions as prompted and restart the test. During the configuration, pay attention to the settings of the normal response example.

After the test is complete, click **OK**.

Modifying the API

To modify an API, choose **API Development > API Catalogs** or **API Development > APIs**, locate the API, and click **Edit** to modify the API.

 NOTE

An API can be edited only when it is in the **Created**, **Rejected**, **Offline**, or **Disabled** state.

9.3.2.2 Generating an API in the Script Mode

This topic describes how to generate an API in the script mode.


To meet personalized query requirements of users, DataArts DataService also supports API generation in the SQL script mode. It allows you to compile API query SQL statements and provides multi-table join, complex query conditions, and aggregation functions.

Configuring Basic API Information

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. Choose **API Development > APIs** from the left navigation bar, and click **Create**. On the displayed page, enter the basic information.

Table 9-10 API basic configuration

Parameter	Description
API	An API name consists of 3 to 64 characters and starts with a letter. Only letters, numbers, and underscores (_) are allowed.

Parameter	Description
API Catalog	<p>A collection of APIs for a specific function or scenario. It is similar to a folder and specifies the location of APIs. You can search for APIs in a specified API catalog.</p> <p>The API catalog is the minimum organization unit of APIs in DataArts DataService and also the minimum management unit in the API gateway. Click Select Catalog to create an API catalog or select an existing one created in Creating an API Directory.</p>
Request Path	<p>API access path, for example, <code>/v2/{project_id}/streams</code>.</p> <p>It is the part between the domain name and query parameters in the URL of a request path, for example, <code>/blogs/xxxx</code> shown in the following figure.</p> <p>Figure 9-12 API access path in the URL</p>  <p>Braces ({}) can be used to identify parameters in a request path as wildcard characters. For example, <code>/blogs/{blog_id}</code> indicates that any parameter can follow <code>/blogs</code>. <code>/blogs/188138</code> and <code>/blogs/0</code> can both match <code>/blogs/{blog_id}</code>, and are processed by this API.</p> <p>In addition, duplicate request paths are not allowed for the same domain name. When a path parameter is used as a wildcard, the name is not unique. For example, <code>/blogs/{blog_id}</code> and <code>/blogs/{xxxx}</code> are considered as the same path.</p>
Parameter Protocol	<p>A protocol used to transmit requests. HTTP and HTTPS are supported.</p> <ul style="list-style-type: none">• HTTP is a basic network transmission protocol. It is stateless, connectionless, simple, fast, and flexible, and uses plaintext for transmission. It is easy to use but has poor security.• HTTPS is an HTTP-based protocol with SSL or TLS encryption verification. It can effectively verify identities and protect data integrity. To access HTTPS APIs, you need to configure related SSL certificates or skip SSL verification.

Parameter	Description
Request Method	<p>HTTP request method, indicating the type of the requested operation, such as GET and POST. The method complies with the resultful style.</p> <ul style="list-style-type: none">● GET requests the server to return specified resources. This method is recommended.● POST requests the server to add resources or perform special operations. This method is used only for API registration. The POST request does not have a body. Instead, it involves transparent transmission.
Description	A brief description of the API to create.
Tag	API tag. The tag is used to mark the API attributes. After the API is created, you can quickly search for the API by tag. A maximum of 20 tags can be set for an API.
Reviewer	<p>An owner who has permissions to review APIs.</p> <p>Click Add to enter the Review Center page. On the displayed page, click Add on the Reviewer Management tab page to add a reviewer.</p>
Security Authentication	<p>Security authentication mode, which can be:</p> <ul style="list-style-type: none">● App Authentication: API Gateway authenticates API requests. This mode has the highest security level.● IAM Authentication: IAM authenticates API requests. This mode has a medium security level.● No authentication: No authentication is required for accessing the API. This mode has a low security level, and is not recommended.
Display Scope	<p>After the API is published, all users in the selected scope can view the API in the service catalog.</p> <ul style="list-style-type: none">● Current workspace APIs● Current project APIs● Current tenant's APIs
Access Log	<p>If you select this option, the API query result will be recorded and retained for seven days. You can choose Operations Management > Access Logs and select the request date to view the logs.</p>

Parameter	Description
Min. Retention Period	<p>Minimum retention period of the API publishing status, in hours. Value 0 indicates that the retention period is not limited.</p> <p>You can suspend, unpublish, or cancel authorization for an API only after the minimum retention period ends. The system notifies the authorized users. If all authorized users have processed the notifications or unbound the API from their apps, the API will be suspended or unpublished, or the API authorization will be canceled. Otherwise, the system will forcibly suspend, unpublish, or cancel authorization for the API when the minimum retention periods ends.</p> <p>For example, if the minimum retention period is set to 24 hours, the API can be suspended 24 hours after it is published. If the authorized user handles the notifications in the review center or unbind the API from the app, the API will be directly suspended. Otherwise, the API will be forcibly suspended when the minimum retention period ends.</p>

Parameter	Description
Input Parameter	<p>Configure parameters in the API request. An input parameter consists of the parameter location, parameter type, whether the parameter is mandatory, and the default value.</p> <ul style="list-style-type: none">• The parameter location can be Query, Header, Path, or Body. In addition, static parameters are supported.<ul style="list-style-type: none">– Query is the query parameter following the URL. It starts with a question mark (?) and connects multiple parameters with &.– Header is located in the request header and is used to transfer current information, for example, host and token.– Path is a request parameter in the request path. If you configure a path parameter, you must also add this parameter to the request path.– Body is a parameter in the request body and is generally in JSON format.– Static is a static parameter that does not change with the value passed by API callers. The parameter value is determined upon API authorization. If the parameter value is not set during authorization, the default value of the API input parameter is used.• The parameter type can be Number or String. Number corresponds to numeric data types such as int, double, and long. String corresponds to text data types such as char, varchar, and text.• Mandatory and Default Value: If you select Yes for Mandatory, parameters must be passed for accessing the API. Otherwise, the default value of the parameter will be used if the parameter is not passed for accessing the API. <p>Constraints for the parameters are as follows:</p> <ul style="list-style-type: none">• Query and Path: 32 KB.• HEADER: The maximum size is 128 KB.• BODY: The maximum size is 128 KB. <p>For instance, to configure the dynamic parameter project_id in the /v2/{project_id}/streams request path, do as follows:</p> <ol style="list-style-type: none">1. Click Add and enter project_id for Name.2. Set Parameter Location to PATH.3. Set Type to STRING.4. Select Yes for Mandatory.5. Leave Default Value blank.

4. After the basic API information is complete, click **Next** to go to the **Data Extract Logic** page.

Configuring the Data Extraction Logic

Set **Data Acquisition Method** to **Script**.

1. Select a data source, data connection, database, and queue to obtain the tables to be configured.

NOTE

For details on the data sources supported by DataArts DataService, see [Data Sources](#). Configure data sources in Management Center in advance and enter SQL statements as prompted.

2. Compile an SQL statement to query APIs.

On the script editing page, enter the SQL statement as prompted.

NOTE

- The fields obtained by SELECT are the returned parameters of the API. (The alias is supported.)
 - The parameters in the WHERE statement are parameters requested by APIs. The parameter format is **$\${parameter\ name}$** .
3. Select a pagination mode.

- Default pagination: If you enter a SQL script when creating an API, DataArts DataService automatically adds the pagination logic to the SQL script. For example, if you enter the following SQL script:

```
SELECT name as Student_Name FROM tableofresults
```

When processing API debugging or calling, DataArts DataService automatically adds the pagination logic to the preceding SQL script and generates the following script:

```
SELECT * FROM (SELECT name as Student_Name FROM tableofresults) LIMIT {pageSize} OFFSET {offsetValue}
```


pageNum and **offsetValue** are the input parameters for API debugging or calling. If the **pageNum** parameter is not specified, DataArts DataService sets it for the API by default. **offsetValue** is calculated based on the value of the input parameter **pageSize**. If **pageSize** is not specified, DataArts DataService sets it for the API by default.

- Custom pagination: DataArts DataService does not process the SQL script for creating an API. The pagination logic is defined by you. If you want to create an API that supports pagination, you can add the pagination logic when writing the SQL statement. Example:

```
SELECT name as Student_Name FROM tableofresults LIMIT {pageSize} OFFSET {offsetValue}
```

4. Add ranking parameters.

In the list of ranking parameters, set whether the ranking parameters are optional, set the ranking mode, and enter the description.

Click  to add the input and ranking parameters to the API requests of the SQL statement.

NOTE

Before adding ranking parameters, ensure that the SQL statement is correct.

5. Edit request parameters.

After the SQL statement is compiled, click **Test SQL** and bind the HTTP input parameters on the database field tab page. Set **pre_order_by** by referring to **6** in [Configuring the Data Extraction Logic](#).

 **NOTE**

pre_order_by is optional. If **pre_order_by** is not set, the mandatory fields are used as the ranking criterion.

Testing the API

After setting and saving all parameters, click **Next**.

Specify **Value** and click **Debug**. You can view the **Request** and **Response** details on the right part of the displayed page. If the test fails, follow the instructions as prompted and restart the test. During the configuration, pay attention to the settings of the normal response example.

After the test is complete, click **OK**.

Modifying the API

To modify an API, choose **API Development > API Catalogs** or **API Development > APIs**, locate the API, and click **Edit** to modify the API.

 **NOTE**

An API can be edited only when it is in the **Created**, **Rejected**, **Offline**, or **Disabled** state.

9.3.2.3 Registering APIs

This topic describes how to register APIs, manage APIs generated based on data tables, and publish APIs to API Gateway.

DataArts DataService Shared supports the registration of RESTful APIs using GET and POST methods.

Configuring Basic API Information

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose **DataArts DataService Shared**. The **Overview** page is displayed.
3. Choose **API Development > APIs** in the left navigation bar, and click **Register**. Configure the basic information.

Table 9-11 API basic configuration

Parameter	Description
API	An API name consists of 3 to 64 characters and starts with a letter. Only letters, numbers, and underscores (_) are allowed.

Parameter	Description
API Catalog	<p>An API catalog is a set of APIs with a specific function or scenario. It is the minimum organization unit of APIs in DataArts DataService and the minimum management unit of API Gateway.</p> <p>Click Select Catalog to create an API catalog or select an existing one created in Creating an API Directory.</p>
Request Path	<p>The path for accessing an API.</p> <p>Example: <code>/v2/{project_id}/streams</code>.</p>
Protocol	<p>A protocol used to transmit requests. HTTP and HTTPS are supported.</p>
Request Mode	<p>HTTP defines the following request modes that can be used to send a request to the server.</p> <p>GET requests the server to return specified resources.</p> <p>POST requests the server to add resources or perform special operations. This method is recommended for API registration. The POST request does not have a body. Instead, it involves transparent transmission.</p>
Description	<p>A brief description of the API to be registered.</p>
Tag	<p>The name of the tag. Only letters, numbers, and underscores (_) are allowed. Tag names cannot start with underscores (_).</p>
Reviewer	<p>An owner who has permissions to review APIs.</p> <p>Click Add to enter the Review Center page. On the page displayed, click Add on the Reviewer Management tab page to add a reviewer.</p>
Security Authentication	<p>Security authentication mode, which can be:</p> <ul style="list-style-type: none">● App Authentication: API Gateway authenticates API requests.● IAM Authentication: IAM authenticates API requests.● Non-authentication: No authentication is required.
Display Scope	<p>After the API is published, all users in the selected scope can view the API in the service catalog.</p> <ul style="list-style-type: none">● Current workspace APIs● Current project APIs● Current tenant's APIs
Access Log	<p>If you select this option, the API query result will be recorded and retained for seven days. You can choose Operations Management > Access Logs and select the request date to view the logs.</p>

Parameter	Description
Min. Retention Period	Minimum duration reserved before API unbinding. Before an API developer suspends, unpublishes, or cancels the authorization of an API, the system notifies the authorized API callers and reserves at least <i>X</i> hours for them to unbind the API. During the retention period, the API can be used if it is not unbound. The value 0 indicates that there is no minimum retention period.
Input Parameter	<p>Input parameter is a set of parameters in the API request, including dynamic parameters in the resource path, query parameters in the request URI, and header parameters.</p> <p>The following is an example that describes the dynamic parameters in the resource path (request path): /v2/{project_id}/streams, where {project_id} is a dynamic parameter that needs to be configured.</p> <ol style="list-style-type: none"> 1. Click Add and enter project_id for Name. 2. Set Parameter Location to PATH. 3. Set Type to STRING. 4. Set Example Value and Description as required.

4. After the basic API information is complete, click **Next** to go to the **Data Extract Logic** page.

Configuring API Parameters

After configuring basic API information, you can set API parameters. The following describes how to configure the API backend services and request parameters.

Table 9-12 API parameters

Parameter	Description
Protocol	<p>A protocol used to transmit requests. HTTP and HTTPS are supported.</p> <p>This parameter is used by DataArts DataService to transmit requests to the APIs to be registered.</p>
Request Mode	<p>HTTP defines the following request modes that can be used to send a request to the server. This parameter is used by DataArts DataService to transmit requests to the APIs to be registered.</p> <p>GET requests the server to return specified resources.</p> <p>POST requests the server to add resources or perform special operations.</p>
Backend Service Host	<p>Backend service host is the host of the API to be registered. The value cannot start with http:// or https:// and cannot contain Path.</p>

Parameter	Description
Backend Service Path	Backend service path is the path of the API to be registered. The path can contain parameters placed in {}, for example, /user/{userid} .
Backend Timeout (ms)	Backend timeout interval.
Backend Service Parameter	The optional parameters can be placed in PATH , Header , and Query . The positions of optional parameters vary depending on the request mode. Select a parameter position as required.
Constant Parameter	Constant parameter is the fixed parameter invisible to the caller. Constant parameter does not need to be transferred during API calling. However, the background service always receives the constant parameter and parameter value defined here. This parameter applies to scenarios in which you want to set a parameter of an API to a fixed value and hide the parameter from the caller.

Testing an API

After all parameters are set, click **Debug**. Specify **Value** and click **Debug**. You can view the **Request** and **Response** details on the right part of the page displayed. If the test fails, follow the instructions as prompted and restart the test. During the configuration, pay attention to the settings of the normal response example.

After the test is complete, click **OK**.

9.3.3 Debugging an API

Scenarios

You can debug an API on the management console by adding HTTP header parameters and body parameters.

NOTE

- APIs whose backend paths contain environment variables cannot be debugged.
- APIs bound to a signature key cannot be debugged.
- If a request throttling policy has been bound to an API, the policy does not take effect when you debug the API.

Prerequisites

- An API has been created.
- The backend service has been set up.

Procedure

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. Choose **API Development > APIs**.
4. Use either of the following methods to debug an API:
 - Locate the row that contains the target API, and choose **More > Debug**.
 - Click the name of the target API, and click **Test** on the displayed API details page.

You can configure API request parameters in the left pane. See [Table 9-13](#) for parameter details. The request information sent by the API and the returned result after the API request is invoked are displayed on the right.

Table 9-13 Debugging APIs

Parameter	Description
Parameters	Query parameters and their values.
Cluster Settings	Supported only by Exclusive Edition. Select the instance where the API to be debugged resides.

NOTE

The information displayed on the debugging page varies according to the request type.

5. After request parameters are added, click **Debug**.

The API calling response information is displayed in the command output area in the right pane.

 - If the API is successfully called, HTTP status code 200 and response information are returned.
 - If the debugging fails, the HTTP status code 4xx or 5xx is returned.
6. You can send different requests using varied parameters and values to verify the API.

NOTE

To modify the API parameters, click **Edit** in the upper right corner. The API editing page is displayed.

Follow-up Procedures

After an API is debugged, you can publish the API. See [Publishing an API](#) for details.

9.3.4 Publishing an API

This section describes how to publish an API to service catalogs.

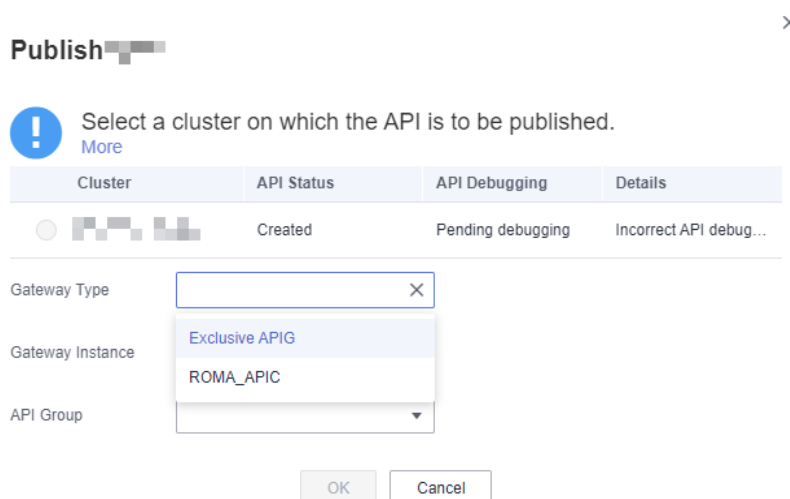
Scenario

For security purposes, APIs generated and registered in DataArts DataService must be published to service catalogs so that they can provide services.

Procedure

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. On the displayed page, choose **APIs** from the left navigation bar. In the **Operation** column, choose **More > Publish**.
4. On the confirmation page, you can click **More** to view publish details.

Figure 9-13 Publish



- In DataArts DataService Exclusive, the API is published to a DataArts DataService Exclusive cluster by default. After the API is published, it can be called through the intranet. You can also publish the API to an APIG Exclusive or ROMA Connect instance.
 - APIG Exclusive: To publish an API to APIG Exclusive, you must buy an APIG instance on the APIG console in advance. After the instance is created, a default API group is available. The system automatically assigns a debugging domain name for internal tests to the API group. This debugging domain name is unique and cannot be changed, and it can be accessed for a maximum of 1,000 times each day. If you want to create an API group exclusively for DataArts DataService APIs, see [Creating an API Group](#). In addition, you can bind one or more independent domain names to an API group. For

details, see [Binding a Domain Name](#). The domain names can be used to call APIs for more than 1,000 times each day.

5. APIs must be reviewed and approved before they can be published. APIs are reviewed in the following way:
 - An API publisher who does not have the reviewer permission must submit the API to the reviewer for review.
 - An API publisher who has the reviewer permission can publish an API without review or approval. By default, a workspace administrator has the reviewer permission.

An API submitted by a non-reviewer is published after it is approved by the reviewer.

Follow-up Operations

After the API is published, you can go to the **Service Catalogs** page to view the API information.

You can also manage APIs. For details, see [Managing APIs](#). Alternatively, you can choose **Operations Management > Throttling Policies** and configure throttling for the API. For details, see [Creating Throttling Policies](#).

9.3.5 Managing APIs

9.3.5.1 Setting an API to Be Visible

Scenario

If you want to change the visibility scope of an API in the service catalog, you can use the **Display** function or set the **Display Scope** parameter for the API.

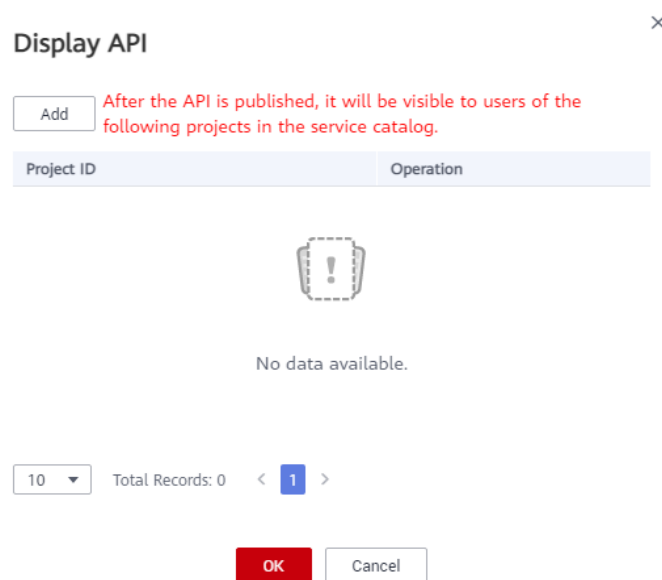
Prerequisites

An API has been created.

Changing the API Visibility Scope Using the Display Function

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
1. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.

2. Choose **API Development > API Catalogs** or **API Development > APIs**. Locate an API, click **More** in the **Operation** column, and select **Display**.
3. In the displayed dialog box, click **Add**, enter a project ID, and click **OK** to make the API visible to users in the project.
For how to obtain the project ID, see [\(Optional\) Obtaining Authentication Information](#).

Figure 9-14 Display API

Changing the API Visibility Scope by Setting the Display Scope Parameter

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. Choose **API Development > API Catalogs** or **API Development > APIs**. Locate an API and click **Edit** in the **Operation** column. An API can be edited only when it is in the **Created**, **Rejected**, **Offline**, or **Disabled** state.
4. On the **Configure Basic Details** page, select a value for the **Display Scope** parameter. The value can be **Current workspace's APIs**, **Current project's APIs**, or **Current tenant's APIs**. Then save the modification.
5. Restore or publish the API again to change the visibility scope of the API in the service catalog.

9.3.5.2 Suspending/Restoring an API

Scenarios

To edit or debug a published API, you must suspend the API first. After the API is suspended, its original authorization information is retained. You can edit and debug the API.

You can restore the API so that it can continue to provide services.

 **NOTE**

The suspended API cannot be accessed in the specified time, which may affect the applications or users who are using the API. Ensure that users have been notified of this consequence.

Prerequisites

- An API has been created.
- An API has been published in the environment.

Suspending an API

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. Choose **API Development > APIs**.
4. Locate the row that contains the API to be suspended, click **More** in the **Operation** column, and select **Suspend**.
5. In the displayed dialog box, select the time period when the API needs to be suspended and click **OK**.

 **NOTE**

The API suspension time must be later than its minimum retention period. Authorized users will be notified of the suspension. If all authorized users process the notifications in the review center or unbind the API from their apps, the API will be directly suspended. Otherwise, the API will be forcibly suspended when the minimum retention period ends.

Restoring an API

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. Locate the row that contains the API to be restored, click **More** in the **Operation** column, and select **Restore**.

9.3.5.3 Unpublishing/Deleting APIs

Scenario

If you want to stop an API that has been published from providing services, you can unpublish the API. For details, see [Unpublishing an API](#).

- If you want to continue to use an API that has been unpublished, you need to publish it again. Note that the original authorization information of the API will not be retained once the API is unpublished.
- If you no longer need the API, you can delete it. For details, see [Deleting APIs](#).

 **NOTE**

The unpublished API cannot be accessed in the specified time, which may affect the applications or users who are using the API. Ensure that users have been notified of this consequence.

Prerequisites

- An API has been created.
- The API has been published.

Unpublishing an API

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. Choose **API Development > APIs**.
4. Locate the row that contains the target API, choose **More > Unpublish**.
5. In the displayed dialog box, select the time period where the API needs to be unpublished and click **OK**.

 **NOTE**

The API unpublishing time must be later than its minimum retention period. Authorized users will be notified of the unpublishing. If all authorized users process the notifications in the review center or unbind the API from their apps, the API will be directly unpublished. Otherwise, the API will be forcibly unpublished when the minimum retention period ends.

Deleting APIs

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. Choose **API Development > API Catalogs**. On the page displayed, select the API you want to delete and click **Delete**.

 **NOTE**

- Only APIs in an unpublished state can be deleted. APIs in suspended or published state cannot be deleted.
 - A maximum of 1,000 APIs can be deleted at a time.
4. Click **OK** to delete the API.

9.3.5.4 Copying an API

Scenario

You can copy an API to obtain another API with the same configuration.

Prerequisites

An API has been created.

Procedure

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. Choose **API Development > APIs**.
4. Select the target API, click **More** above the API list, and select **Copy**.
5. In the displayed dialog box, enter the new API name and request path, and click **OK**.

Figure 9-15 Copying an API

Copy ×

* API Name
API names can be 4 to 50 characters long. They must start with a letter, and they can contain letters, numbers, and underscores ().

* Request Path
API paths can be 200 characters long. They must start with a slash (/), and they can contain request parameters included in {}, for example, /getUserInfo/{userId}. They can contain letters, numbers, and special characters _-*%_-.

OK Cancel

9.3.5.5 Synchronizing APIs

Operation Scenario

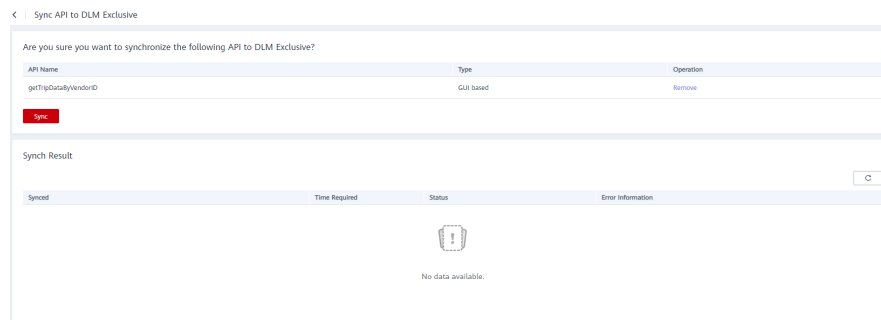
You can use the API synchronization function to synchronize APIs between DataArts DataService Exclusive.

Prerequisites

An API has been created.

Procedure

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. Choose **API Development > APIs**.
4. Select the target APIs, click **More** above the API list, and select **Sync**.
5. On the displayed page, click **Sync**. The synchronization status is displayed in the **Synch Result** area.

Figure 9-16 Synchronizing APIs

9.3.5.6 Exporting All/Exporting/Importing APIs

Operation Scenario

DataArts DataService allows you to import and export (including exporting all) APIs to quickly copy or migrate existing APIs.

Prerequisites

- An API has been created.
- To export all APIs, you must have the DAYU Administrator or Tenant Administrator permission.
- Only one task for exporting all APIs can be executed at a time.
- All the APIs of a workspace can be exported only once every minute.

Exporting All APIs

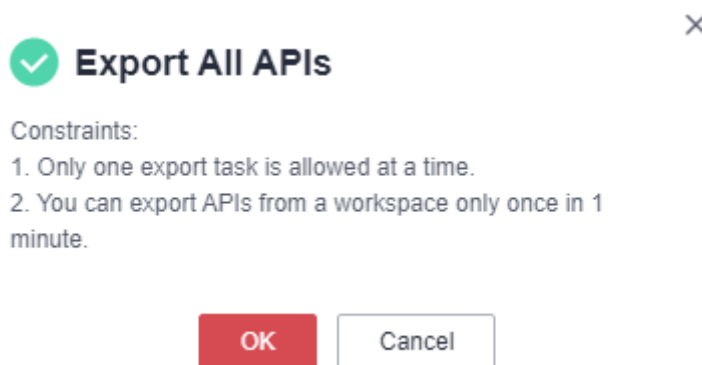
1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. Choose **API Development > APIs**.
4. Above the API list, choose **More > Export All**.

NOTE

- To export all APIs, you must have the DAYU Administrator or Tenant Administrator permission.
- Only one task for exporting all APIs can be executed at a time.
- All the APIs of a workspace can be exported only once every minute.

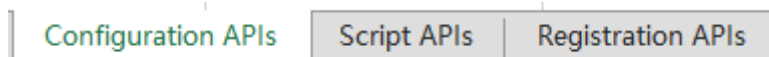
In the displayed dialog box, click **Yes** to export all the APIs to an Excel file.

Figure 9-17 Exporting all APIs



5. Open the downloaded Excel file to view the exported APIs. APIs of different types are exported to different sheets.

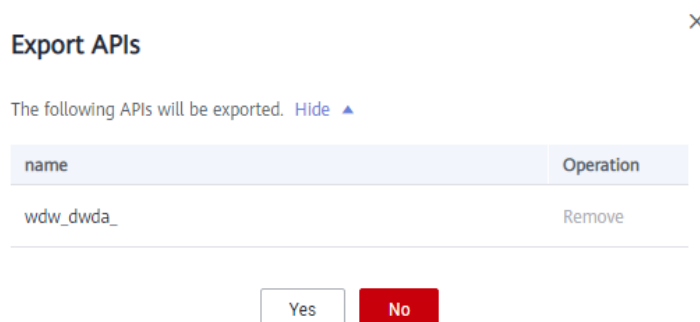
Figure 9-18 Exported Excel file



Exporting APIs

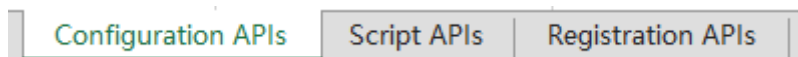
1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. Choose **API Development > APIs**.
4. Select the target APIs, click **More** above the API list, and select **Export**.
5. In the displayed dialog box, confirm the APIs to export and click **Yes** to export the APIs to an Excel file.

Figure 9-19 Exporting APIs



6. Open the downloaded Excel file to view the exported APIs. APIs of different types are exported to different sheets.

Figure 9-20 Exported Excel file



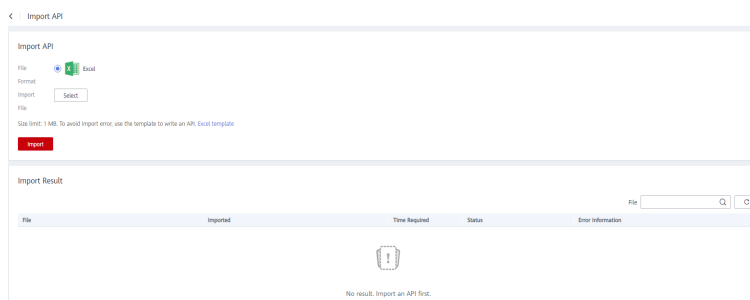
Importing APIs

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. Choose **API Development > APIs**.
4. Click **More** above the API list and select **Import**.
5. On the displayed page, click **Select**, select an API file, and click **Import**. The import status is displayed in the **Import Result** area.

NOTE

The API file can be one exported from another project or an Excel file edited based on the template specifications.

Figure 9-21 Importing APIs



6. After the APIs are imported successfully, you can view them in the API list.

9.3.6 Creating Throttling Policies

Scenario

A throttling policy limits the maximum number of times that an API can be called within a specific period. Throttling policies can protect the backend service from getting overloaded. Currently, API throttling can limit the number of API calls by user, application, and time period.

To ensure the stability of services, you can create throttling policies to control the calls made to specified APIs. Throttling policies take effect for an API only if they are bound to the API.

NOTE

An API can be bound to only one throttling policy in an environment, but each throttling policy can be bound to multiple APIs.

Prerequisites

The API to be bound has been published.

Creating a Throttling Policy

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. On the displayed page, choose **Throttling Policies** from the left navigation bar.
4. On the displayed page, click **Create**. Set the parameters listed in [Table 9-14](#).

Figure 9-22 Creating a throttling policy

Create Throttling Policy ×

* Name
Throttling policy names can be 3 to 64 characters long. They must start with a letter and they can contain letters, numbers, and underscores (_).

* Time Range --Select-- ▼

* Max. API Requests

Max. User Requests (The value cannot exceed the maximum API requests.)

Max. App Requests (The value cannot exceed the maximum user requests.)

Max. Source IP Requests (The value cannot exceed the maximum API requests.)

Description
0/255

OK Cancel

Table 9-14 Parameters

Parameter	Description
Name	The throttling policy name.

Parameter	Description
Time Range	The time duration for limiting the number of API calls <ul style="list-style-type: none">Used together with Max. API Requests to specify the total number of times an API can be called within a time period.Used together with Max. User Requests to specify the number of times an API can be called by a user within a time period.Used together with Max. App Requests to specify the total number of times an API can be called by an app within a time period.
Max. API Requests	The maximum number of times an API can be called within the specified time period. Used together with Time Range to specify the maximum number of times an API can be called within the period.
Max. User Requests	The maximum number of times an API can be called by a user within the specified period. <ul style="list-style-type: none">The value of this parameter must be less than that of Max. API Requests.Used together with Time Range to specify the maximum number of times an API can be called by a user within the specified period.
Max. App Requests	The maximum number of times an application can be called by a user within the specified period. <ul style="list-style-type: none">The value of this parameter must be less than that of Max. User Requests.Used together with Time Range to specify the maximum number of requests an app can make within the specified period.
Description	A description of the throttling policy to be created

5. Click **OK**.

After the throttling policy is created, it is listed in the throttling policy list. Bind the throttling policy to an API to limit the access traffic.

Binding a Throttling Policy to an API

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. Choose **Throttling Policies** from the left navigation bar.
4. Bind a throttling policy to an API in either of the following ways:
 - Locate the throttling policy to be bound and click **Associate with APIs**.

- Click the target policy name to go to its details page and click **Associate with APIs** on the **List of Associated APIs** tab page.
5. Enter an API group and API name to search for the target API.
 6. Select the API and click **OK**.

 **NOTE**

If a throttling policy is no longer needed, click **Unbind** on the **List of Associated APIs** tab page. To unbind multiple APIs at a time, select the APIs to be unbound and click **Unbind**. Up to 1000 APIs can be unbound at a time.

Deleting a Throttling Policy

You can delete a throttling policy if it is no longer needed.

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. On the displayed page, choose **Throttling Policies** from the left navigation bar.
4. On the displayed page, locate the policy you want to unbind and click **Delete** in the **Operation** column.

 **NOTE**

- Throttling policies bound to APIs cannot be deleted. Therefore, you need to unbind them from APIs before deleting them.
 - To delete multiple throttling policies at a time, select the policies, and click **Delete**. Up to 1000 throttling policies can be deleted at a time.
5. Click **Yes**.

9.4 Calling APIs

Overview

To call an API, perform the following operations:

1. Obtain an API.

Obtain the API from the service catalog. An API can be called only after it is published.

2. (Optional) Create an application and get authorized.

For an API that is accessed using application or IAM authentication, you need to **create an application** and **authorize the application to use the API**. When you call an API, DataArts DataService verifies your identity based on the key pair (AppKey and AppSecret) of the created application.

3. **Call the API**.

After completing the preceding steps, you can call the API.

Creating an App

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. Choose **API Calling > Apps**. On the page displayed, click **Create**. The **Create App** dialog box is displayed. Set the parameters listed in [Table 9-15](#).

Table 9-15 App information

Parameter	Description
Name	The name of the application to create.
Type	IAM : IAM authentication is used, which means access using a token. APP : access through app authentication
Description	A description of the application to create.

4. Click **OK**.
After the application is created, its name and ID are displayed in the application list.
5. Click an application name, and view the **AppKey** and **AppSecret** on the displayed application details page.

Figure 9-23 Application details page



Authorizing an Application to Use an API

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. Authorize an application to use an API in either of the following ways:
Giving API authorization:
 - a. Choose **API Development > APIs**.
 - b. Locate the row that contains the API to be bound, and click **View**.
On the page displayed, click **Authorize**.
 - c. (Optional) If **Parameter Location** was set to **Static** for an input parameter during API creation, you must set a static parameter value. If you do not set a value, the default value of the API input parameter is used.

- d. Set an expiry time, select an application, and click **OK**.

Applying for authorization:

- a. Choose **API Calls > Service Catalogs**.
 - b. Click the name of the API you want to bind to an application.
 - c. On the page displayed, click **Permission Application**.
 - d. (Optional) If **Parameter Location** was set to **Static** for an input parameter during API creation, you must set a static parameter value. If you do not set a value, the default value of the API input parameter is used.
 - e. Set an expiry time, select an application, and click **OK**.
 - f. After the application is submitted, the authorization takes effect only after it is approved in the review center.
4. After the authorization is complete, view the bound APIs on the application details page.

 **NOTE**

- In the API list, if you no longer access an API through the application, click **Unbind** in the **Operation** column.
- To test an API to which the application is bound, choose **More > Debug** in the **Operation** column
- To extend the authorization period for the bound API, click **Renew**.

Calling an API

The only difference between the three authentication methods is the authentication content. The methods for calling APIs are the same.

- **IAM Authentication:** IAM authenticates API requests.
- **Non-authentication:** No authentication is required. You can directly call an API.
- **App Authentication:** Application authentication is used for calling an API.
 - When **App Authentication** is used, an SDK is required for access.
 - Currently, Java, Go, Python, JavaScript, C#, PHP, C++, C, and Android SDKs are available.
 - For details on API calling examples in different programming languages, see [Java](#), [Go](#), [Python](#), [C#](#), [JavaScript](#), [PHP](#), [C++](#), [C](#), and [Android](#).

9.5 Performing Operations in Review Center

The review center of DataArts DataService is designed to approve the applications of publishing APIs, suspending APIs, permission authorization, renewal, and other operations.

- If an API developer wants to publish an API to the API marketplace, remove an API from the API marketplace, and reclaim the authorization of an application, these operations take effect only after being approved by the reviewers.
- If an API caller wants to authorize an API or apply for renewal, these operations take effect only after the reviewers approve the operations.

- An API developer or caller can cancel an API application to be reviewed in the review center.
- The reviewer can only view applications for publishing APIs from non-reviewers. APIs can only be published to the release environment of API Gateway.

Reviewing an Application

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. Choose **Operation Management > Review Center** in the left navigation bar and click the **Pending Review** tab.
4. Locate the task you want to review based on the search criteria such as the review type and submission time. Then, select **Review** in the **Operation** column.

NOTE

Multiple APIs can be selected for batch review.

Managing a Reviewer

In the review center of DataArts DataService, you can add and delete reviewers.

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. Choose **Operation Management > Review Center** in the left navigation bar and click the **Reviewer Management** tab.
4. Click **Add**. In the **Add Reviewer** dialog box displayed, set the basic information.

Canceling an API Application

DataArts DataService provides the function of canceling applications to be reviewed. You can cancel applications to be reviewed on the **Applications** tab page on the **Review Center** page.

1. On the DataArts Studio console, locate an instance and click **Access**. On the displayed page, locate a workspace and click **DataArts DataService**.
2. In the left navigation pane, choose an edition, for example, **Exclusive Edition**. The **Overview** page is displayed.
3. Choose **Operations Management > Review Center** in the left navigation pane and click the **Applications** tab.
4. Locate the row that contains the API to be canceled, and click **Cancel** in the **Operation** column.

10 Error Codes

10.1 DataArts Migration Error Codes

If an exception occurs during the execution of an operation request and the request is not processed, an error message is returned. The error information contains the error code and error description. [Table 10-1](#) lists some common error code in CDM error messages. You can handle the exceptions by referring to the solutions in [Table 10-1](#).

Error Code Description

Table 10-1 Description

Error Code	Error Message	Solution
Cdm. 0000	System error.	Contact customer service or technical support.
Cdm. 0003	Kerberos login failed.	Check whether the keytab and principal configuration files are correct.
Cdm. 0009	<i>%s</i> is not an integer or is beyond the value range [0, 2147483647].	Modify the parameter settings based on the error message and try again.
Cdm. 0010	The integer must be within the range of [<i>%s</i>].	Check whether the parameter value is valid based on the error message. If it is not, correct it and try again.
Cdm. 0011	The parameter value exceeds the value range.	Check whether the parameter value is valid based on the error message. If it is not, correct it and try again.
Cdm. 0012	JDBC driver class is not found.	Contact customer service or technical support.

Error Code	Error Message	Solution
Cdm. 0013	Failed to connect to the agent.	It is possible that the network is disconnected, or no security group or firewall rule is configured to allow access. If the fault persists, contact customer service or technical support.
Cdm. 0014	The parameter is invalid.	Change the parameter value and try again.
Cdm. 0015	An error occurred during file parse.	Check whether the content or format of the uploaded file is correct. If it is not, correct it and try again.
Cdm. 0016	The file to be uploaded cannot be empty.	Ensure that the file you uploaded is not empty and try again.
Cdm. 0017	MRS Kerberos authentication failed.	Check whether the password used for Kerberos authentication is strong. If it is not, change to a strong password and try again.
Cdm. 0018	The content of jobs or links is invalid.	Contact customer service or technical support.
Cdm. 0019	Invalid IP address and port number.	Try again later or contact customer service or technical support.
Cdm. 0020	The string must contain the following substring: %s.	Modify the parameter settings based on the error message and try again.
Cdm. 0021	Failed to connect to the server: %s.	Contact customer service or technical support.
Cdm. 0023	Failed to write data. Cause: %s.	Contact customer service or technical support.
Cdm. 0024	[%s] must be within the range of [%s].	Modify the parameter settings based on the error message and try again.
Cdm. 0025	The length of the written data exceeds the length defined by the table field. Error message: %s.	Modify the length of the data to be written based on the error message and try again.
Cdm. 0026	The primary key already exists. Error message: %s.	Check the data based on the error message and resolve the primary key conflict.

Error Code	Error Message	Solution
Cdm. 0027	The code of the written character string may be different from the code defined in the table. Error message: %s.	Modify the character string code based on the error message.
Cdm. 0028	Incorrect username or password. Error message: %s.	Change the username or password and try again.
Cdm. 0029	The database name does not exist. Error message: %s.	Select a correct database and try again.
Cdm. 0030	Incorrect username, password, or database name. Error message: %s.	Correct the username, password, and database name as prompted and try again.
Cdm. 0031	The connection timed out.	Connection timed out. Check whether the IP address, host name, and port number are correct, and whether the security group and firewall are correctly configured.
Cdm. 0032	Incorrect username or password. See the error message returned by the server: %s.	Change the username and password based on the error message and try again.
Cdm. 0033	SIMPLE authentication is not supported.	Select the Kerberos authentication type and try again.
Cdm. 0034	Restart the CDM cluster to reload MRS or FusionInsight configurations.	Restart the CDM cluster to reload MRS or FusionInsight configurations.
Cdm. 0035	You do not have the write permission on the file. Error message: %s.	Configure the permission based on the error message and try again.
Cdm. 0036	Invalid datestamp or date format. Error message: %s.	Configure the datestamp or date format based on the error message and try again.
Cdm. 0037	The parameter is invalid. Error message: %s.	Correct the parameter settings based on the error message and try again.
Cdm. 0038	The connection timed out.	Check the VPC and security group rules.
Cdm. 0039	The connection name cannot be modified.	The connection name cannot be changed.

Error Code	Error Message	Solution
Cdm. 0040	Logs are deleted because they are periodically cleared.	Contact customer service or technical support.
Cdm. 0041	The group in use cannot be updated or deleted.	Do not modify the group.
Cdm. 0042	Failed to operate the group. Error message: %s.	Select a correct group based on the error message and try again.
Cdm. 0043	Failed to trigger data extraction or loading failed. Cause: %s.	Contact customer service or technical support.
Cdm. 0051	Invalid submission engine: %s.	Specify a correct job engine and try again.
Cdm. 0052	Job %s is running.	The operation cannot be performed because the job is running. Try again after the job completes.
Cdm. 0053	Job %s is not running.	Run the job and try again.
Cdm. 0054	Job %s does not exist.	Check whether the job exists.
Cdm. 0055	Unsupported job type.	Specify a correct job type and try again.
Cdm. 0056	Failed to submit the job. Cause: %s.	Locate the cause based on the error message, rectify the fault, and try again.
Cdm. 0057	Invalid job execution engine: %s.	Specify a correct job engine and try again.
Cdm. 0058	Invalid combination of submission and execution engines.	Specify a correct job engine and try again.
Cdm. 0059	Job %s has been disabled. Failed to submit the job.	Create a job and try again. Alternatively, contact customer service or technical support.
Cdm. 0060	Link %s for this job has been disabled. Failed to submit the job.	Change the link and submit the job again.
Cdm. 0061	Connector %s does not support the specified direction. Failed to submit the job.	The connector cannot be used as the source or destination of a job. Change the link and submit the job again.

Error Code	Error Message	Solution
Cdm. 0062	The binary file is applicable only to the SFTP, FTP, HDFS, or OBS connector.	Specify a correct connector and try again.
Cdm. 0063	An error occurred during table creation. Cause: %s.	Locate the cause based on the error message, rectify the fault, and try again.
Cdm. 0064	The data format is incorrect.	Check whether the data format is correct based on the error message. If it is not, correct it and try again.
Cdm. 0065	Failed to start the scheduler. Cause: %s.	Contact customer service or technical support.
Cdm. 0066	Failed to obtain the sample value. Cause: %s.	Contact customer service or technical support.
Cdm. 0067	Failed to obtain the schema. Cause: %s.	Contact customer service or technical support.
Cdm. 0068	Failed to clear table data. Cause: %s.	<ul style="list-style-type: none"> • Check whether the current account has the operation permissions on the table. • Check whether the table is locked. • If neither of the preceding methods is feasible, contact customer service or technical support.
Cdm. 0070	Failed to run task %s because the maximum number of running jobs has been reached.	Contact customer service or technical support.
Cdm. 0071	Failed to obtain table data. Cause: %s.	Contact customer service or technical support.
Cdm. 0074	Failed to repair the table. Cause: %s.	Contact customer service or technical support.
Cdm. 0075	Failed to delete the table. Cause: %s.	<ul style="list-style-type: none"> • Check whether the current account has the operation permissions on the table. • Check whether the table is locked. • If neither of the preceding methods is feasible, contact customer service or technical support.

Error Code	Error Message	Solution
Cdm. 0080	Invalid username.	Correct the username based on the error message and try again.
Cdm. 0081	Invalid certificate.	Contact customer service or technical support.
Cdm. 0082	The certificate is not readable.	Contact customer service or technical support.
Cdm. 0083	A process cannot be configured with multiple certificates. Restart to use the new certificate.	Modify the certificate based on the error message and restart the system.
Cdm. 0085	The value exceeds the upper limit.	Contact customer service or technical support.
Cdm. 0088	Incorrect <i>XX</i> configuration item.	Modify the configuration item based on the error message and try again.
Cdm. 0089	The configuration item <i>XX</i> does not exist.	<ul style="list-style-type: none"> Modify the configuration item based on the error message and try again. During the switchover from a CDM cluster of an earlier version to a CDM cluster of a later version, configuration items may be unavailable occasionally when you create a data connection or save a job. In this case, manually clear the cache and try again.
Cdm. 0091	The patches cannot be installed.	Contact customer service or technical support.
Cdm. 0092	The backup file does not exist.	Contact customer service or technical support.
Cdm. 0093	Failed to load the krb5.conf file.	Contact customer service or technical support.
Cdm. 0094	The link named <i>XX</i> does not exist.	Check whether the <i>XX</i> link exists based on the error message and try again.
Cdm. 0095	The job named <i>XX</i> does not exist.	Check whether the <i>XX</i> job exists based on the error message and try again.
Cdm. 0100	Job [%s] does not exist.	Specify a correct job and try again.

Error Code	Error Message	Solution
Cdm. 0101	Link [%s] does not exist.	Specify a correct link and try again.
Cdm. 0102	Connector [%s] does not exist.	Specify a correct connector and try again.
Cdm. 0104	The job name exists.	Rename the job and try again.
Cdm. 0105	The expression is empty.	<ul style="list-style-type: none"> Check whether the expression is valid by referring to the help document. If the fault persists, contact customer service or technical support.
Cdm. 0106	Failed to calculate the <i>XX</i> expression.	<ul style="list-style-type: none"> Check whether the expression is valid by referring to the help document. If the fault persists, contact customer service or technical support.
Cdm. 0107	The task is being executed. Modify job configurations later.	After the task is complete, modify the job configurations.
Cdm. 0108	Failed to query table records.	<ul style="list-style-type: none"> Ensure that the custom SQL statement is correct. Ensure that the query does not time out (less than 60s). If the preceding errors cannot be avoided, contact customer service or technical support.
Cdm. 0109	The length of a job or link name cannot exceed %s.	Modify the job or link name based on the error message.
Cdm. 0110	Invalid name. The name must start with a character or digit and consist of only letters, digits, underscores (_), hyphens (-), and dots (.).	Change the name based on the error message.
Cdm. 0201	Failed to obtain the instance.	Contact customer service or technical support.
Cdm. 0202	Unknown job status.	Try again later or contact customer service or technical support.
Cdm. 0204	No MRS link is created.	Go to the Link Management page to create an MRS link and try again.

Error Code	Error Message	Solution
Cdm. 0230	Failed to load the specified class: %s.	Contact customer service or technical support.
Cdm. 0231	Failed to initialize the specified class: %s.	Contact customer service or technical support.
Cdm. 0232	Failed to write data. Cause: %s.	Contact customer service or technical support.
Cdm. 0233	An exception occurred during data extraction. Cause: %s.	Contact customer service or technical support.
Cdm. 0234	An exception occurred during data loading. Cause: %s.	Contact customer service or technical support.
Cdm. 0235	All data has been consumed. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact customer service or technical support.
Cdm. 0236	Invalid partitions have been retrieved from Partitioner.	Locate the cause based on the error message. If the fault persists, contact customer service or technical support.
Cdm. 0237	Failed to find the JAR file of the connector.	Contact customer service or technical support.
Cdm. 0238	%s cannot be empty.	Modify the parameter settings based on the error message and try again.
Cdm. 0239	Failed to obtain HDFS. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact customer service or technical support.
Cdm. 0240	Failed to obtain the status of the %s file.	Contact customer service or technical support.
Cdm. 0241	Failed to obtain the type of the %s file.	Contact customer service or technical support.
Cdm. 0242	An exception occurred during file check: %s.	Contact customer service or technical support.
Cdm. 0243	Failed to rename %s to %s.	Rename the job and try again.
Cdm. 0244	Failed to create the %s file.	Check whether you have the permissions or try again later. If the fault persists, contact customer service or technical support.

Error Code	Error Message	Solution
Cdm. 0245	Failed to delete the %s file.	Check whether you have the permissions or try again later. If the fault persists, contact customer service or technical support.
Cdm. 0246	Failed to create the %s directory.	Check whether you have the permissions or try again later. If the fault persists, contact customer service or technical support.
Cdm. 0247	HBase operation failure. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact customer service or technical support.
Cdm. 0248	Failed to clear data in %s. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 0249	The file name %s is invalid.	Modify the file name and try again.
Cdm. 0250	Failed to perform operations in the path: %s.	Check whether you have the permissions or try again later. If the fault persists, contact .
Cdm. 0251	Failed to load data to HBase. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 0307	Failed to release the link. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 0315	The link name %s already exists.	Specify another link name and try again.
Cdm. 0316	The link that does not exist cannot be updated.	Specify a correct link and try again.
Cdm. 0317	Link %s is invalid.	Specify a correct link and try again.
Cdm. 0318	The job exists and cannot be created repeatedly.	Specify another job name and try again.
Cdm. 0319	The job that does not exist cannot be updated.	Check whether the job to be updated exists. If it does, modify the job name and try again.
Cdm. 0320	Job %s is invalid.	Contact .

Error Code	Error Message	Solution
Cdm. 0321	Link %s has been used.	Release the link and try again.
Cdm. 0322	Job %s has been used.	Contact .
Cdm. 0323	The submission already exists and cannot be created repeatedly.	Try again later.
Cdm. 0327	Invalid link or job: %s.	Specify a correct link or job and try again.
Cdm. 0411	An error occurred when connecting to the file server.	Contact .
Cdm. 0412	An error occurred when disconnecting from the file server.	Contact .
Cdm. 0413	An error occurred in data transfer to the file server.	Contact .
Cdm. 0415	An error occurred when downloading files from the file server.	Contact .
Cdm. 0416	An error occurred during data extraction.	Contact .
Cdm. 0420	The source file or source directory does not exist.	Check whether the source file or source directory exists. If it does not, specify a correct source file or directory and try again.
Cdm. 0423	Duplicate files exist in the destination path.	Delete duplicate files from the destination path and try again.
Cdm. 0500	The source directory or the [%s] file does not exist.	Specify a correct source file or directory and try again.
Cdm. 0501	Invalid URI [%s].	Specify a correct URI and try again.
Cdm. 0518	Failed to connect to HDFS. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 0523	Connection timed out due to insufficient user permissions.	Create another service user, grant required permissions to the user, and try again.

Error Code	Error Message	Solution
Cdm. 0600	Failed to connect to the FTP server. Cause: %s.	It is possible that the network is disconnected, no security group or firewall rule is configured to allow access, the FTP host name cannot be parsed, or the FTP username or password is incorrect. If the fault persists, contact .
Cdm. 0700	Failed to connect to the SFTP server. Cause: %s.	It is possible that the network is disconnected, no security group or firewall rule is configured to allow access, the SFTP host name cannot be parsed, or the SFTP username or password is incorrect. If the fault persists, contact .
Cdm. 0800	Failed to connect to the OBS server. Cause: %s.	It is possible that the OBS endpoint is inconsistent with the current region, the AK/SK pair is incorrect, the AK/SK pair is not the current user's, or no security group or firewall rule is configured to allow access. If the fault persists, contact .
Cdm. 0801	OBS bucket [%s] does not exist.	The OBS bucket may not exist or is not in the current region. Specify a correct OBS bucket and try again.
Cdm. 0900	Table [%s] does not exist.	Specify a correct table name and try again.
Cdm. 0901	Failed to connect to the database server. Cause: %s.	Contact .
Cdm. 0902	Failed to execute the SQL statement. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 0903	Failed to obtain metadata. Cause: %s.	Check whether the quote character is correct or whether the database table exists when you create the link. If the fault persists, contact .
Cdm. 0904	An error occurred while retrieving data from the result. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 0905	No partition column is found.	Specify a partition column and try again.
Cdm. 0906	No boundary is found in the partition column.	Contact .

Error Code	Error Message	Solution
Cdm. 0911	The table name or SQL must be specified.	Specify a table name or SQL statement and try again.
Cdm. 0912	The table name and SQL cannot be specified at the same time.	Specify one of them and try again.
Cdm. 0913	Schema and SQL cannot be specified at the same time.	Specify one of them and try again.
Cdm. 0914	Partition column is mandatory for query-based import.	Specify a partition column and try again.
Cdm. 0915	The SQL-based import mode and columnList cannot be used at the same time.	Use either of them and try again.
Cdm. 0916	Last value is mandatory for incremental read.	Specify the last value and try again.
Cdm. 0917	Last value cannot be obtained without field check.	Contact .
Cdm. 0918	If no transfer table is specified, shouldClearStageTable cannot be specified.	Specify a transfer table and try again.
Cdm. 0921	Type <i>%s</i> is not supported.	Specify a correct type and try again.
Cdm. 0925	The partition column contains unsupported values.	Correct the values and try again.
Cdm. 0926	Failed to obtain the schema. Cause: <i>%s</i> .	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 0927	The transfer table is not empty.	Specify an empty transfer table and try again.
Cdm. 0928	An error occurred when data is migrated from the transfer table to the destination table.	Contact .
Cdm. 0931	Schema column size [<i>%s</i>] does not match the result set column size [<i>%s</i>].	Change the schema column size to be the same as the result set column size and try again.
Cdm. 0932	Failed to obtain the maximum value of the field.	Contact .

Error Code	Error Message	Solution
Cdm. 0934	Multiple tables of the same name exist in different schemas or catalogs.	Contact .
Cdm. 0935	No primary key. Specify the partition column.	Specify a partition column and try again.
Cdm. 0936	The maximum number of error dirty data records has been reached.	Edit the job and increase the number of error dirty data records.
Cdm. 0940	Failed to match the exact table name.	Specify a correct table name and try again.
Cdm. 0941	Failed to connect to the server. Cause: %s.	Check whether the IP address, host name, and port number are correct, and whether the network security group and firewall are correctly configured. Locate the fault based on the error message. If the fault persists, contact .
Cdm. 0950	Failed to connect to the database with the existing authentication information.	Incorrect authentication information. Correct it and try again.
Cdm. 0960	Server address must be specified.	Specify the server address and try again.
Cdm. 0961	Invalid server address format.	Change to the correct format and try again.
Cdm. 0962	The host IP address must be specified.	Specify the host IP address and try again.
Cdm. 0963	The host port must be specified.	Specify the host port and try again.
Cdm. 0964	The database must be specified.	Specify a database and try again.
Cdm. 1000	Hive table [%s] does not exist.	Specify a correct Hive table name and try again.
Cdm. 1010	Invalid URI [%s]. The URI must be either null or a valid URI.	Specify a correct URI and try again. Correct URI examples: <ul style="list-style-type: none"> • hdfs://example.com:8020/ • hdfs://example.com/ • file:/// • file:///tmp • file://localhost/tmp

Error Code	Error Message	Solution
Cdm. 1011	Failed to connect to Hive. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 1012	Failed to initialize the Hive client. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 1100	Table [%s] does not exist.	Enter a correct table name and try again.
Cdm. 1101	Failed to obtain the link. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 1102	Failed to create the table. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 1103	No rowkey is set.	Set the rowkey and try again.
Cdm. 1104	Failed to open the table. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 1105	Failed to initialize the job. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 1111	The table name is mandatory.	Specify a correct table name and try again.
Cdm. 1112	The import mode is mandatory.	Set the import mode and try again.
Cdm. 1113	Whether to clear data before import has not been specified.	Set Clear Data Before Import and try again.
Cdm. 1114	The rowkey is empty. Set it in field mapping.	Fix the error based on the error message.
Cdm. 1115	Columns is empty. Set it in field mapping.	Fix the error based on the error message.
Cdm. 1116	Duplicate column names. Set it in field mapping.	Fix the error based on the error message.
Cdm. 1117	An error occurred when checking whether the table exists. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact .

Error Code	Error Message	Solution
Cdm. 1118	Table %s does not contain the %s column family.	Specify a column family and try again.
Cdm. 1119	The number of column families is %s and the number of columns is %s.	Change the number of column families to the same as the number of columns and try again.
Cdm. 1120	The table contains data. Clear the table data or set the configuration item to specify whether to clear the table data before the import.	Fix the error based on the error message.
Cdm. 1121	Failed to close the link. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 1201	Failed to connect to the Redis server. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 1202	Failed to connect to the Redis cluster in single-node mode.	Connect to the Redis cluster in another mode.
Cdm. 1203	Failed to extract data from the Redis server. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 1205	Redis Key Prefix cannot be blank.	Delete the whitespace before the Redis prefix and try again.
Cdm. 1206	The storage type of the Redis value must be STRING or HASH .	Fix the error based on the error message.
Cdm. 1207	When the value of the storage type is STRING , Value Delimiter must be specified.	Specify a value delimiter and try again.
Cdm. 1208	columnList of Redis must be specified.	Specify columnList and try again.
Cdm. 1209	Redis Key Delimiter cannot be empty.	Enter a correct delimiter and try again.
Cdm. 1210	primaryKeyList of Redis must be specified.	Specify primaryKeyList and try again.
Cdm. 1211	primaryKeyList of Redis must exist in columnList .	Specify primaryKeyList and try again.

Error Code	Error Message	Solution
Cdm. 1212	databaseType of Redis must be Original or DCS .	Fix the error based on the error message.
Cdm. 1213	Redis Server Address must be specified.	Specify Redis Server Address and try again.
Cdm. 1301	Failed to connect to the MongoDB server. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 1302	Failed to extract data from the MongoDB server. Cause: %s	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 1304	The collection of MongoDB servers must be specified.	Specify the collection of MongoDB servers and try again.
Cdm. 1305	Server Address of MongoDB must be specified.	Specify Server Address and try again.
Cdm. 1306	The database name of the MongoDB service must be specified.	Specify a database and try again.
Cdm. 1307	serverlist of MongoDB must be specified.	Specify serverlist and try again.
Cdm. 1400	Failed to connect to the NAS server.	Contact .
Cdm. 1401	No permissions to access the NAS server.	Apply for the permissions and try again.
Cdm. 1501	Failed to connect to the Elasticsearch server. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 1502	Failed to write data to the Elasticsearch server. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 1503	Failed to close the Elasticsearch link. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 1504	An error occurred when obtaining the Elasticsearch index. Cause: %s	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 1505	An error occurred when obtaining the Elasticsearch type. Cause: %s	Locate the cause based on the error message. If the fault persists, contact .

Error Code	Error Message	Solution
Cdm. 1506	An error occurred when obtaining the Elasticsearch field. Cause: %s	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 1507	An error occurred when obtaining the Elasticsearch sample data. Cause: %s	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 1508	The host name or IP address of the Elasticsearch server must be specified.	Specify the host name or IP address and try again.
Cdm. 1509	The port of the Elasticsearch server must be specified.	Specify a port and try again.
Cdm. 1510	The Elasticsearch index must be specified.	Specify an index and try again.
Cdm. 1511	The Elasticsearch type must be specified.	Specify a type and try again.
Cdm. 1512	columnList of Elasticsearch must be specified.	Specify columnList and try again.
Cdm. 1513	columnList must contain the field type definition.	Include the field type definition and try again.
Cdm. 1514	columnList must contain primaryKey .	Set the primary key field and try again.
Cdm. 1515	An error occurred when resolving the JSON character string. Cause: %s.	Locate the cause based on the error message, rectify the fault, and try again. If the fault persists, contact .
Cdm. 1516	The column name %s is invalid.	Enter a correct column name and try again.
Cdm. 1517	An error occurred when obtaining the number of documents.	Contact .
Cdm. 1518	The partition fails to be created.	Contact .
Cdm. 1519	An error occurred during data extraction.	Contact .
Cdm. 1520	Failed to obtain the type. Cause: %s.	Locate the cause based on the error message. If the fault persists, contact .
Cdm. 1601	Failed to connect to the server.	Contact .

Error Code	Error Message	Solution
Cdm. 1603	Failed to obtain the sample value of the %s topic.	Contact .
Cdm. 1604	No data exists in topic %s.	Locate the cause. Alternatively, change the topic and try again.
Cdm. 1605	Invalid brokerList .	Specify a correct brokerList and try again.