Data Lake Governance Center

User Guide

Issue 01

Date 2020-11-10

HUAWEI TECHNOLOGIES CO., LTD.

Copyright © Huawei Technologies Co., Ltd. 2021. 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

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 DGC Introduction....1
2 Initial Preparations....3
2.1 Making Preparations Before Using DGC....3
2.2 Registering an Account and Getting Authenticated....3
2.3 Buying DGC Instances....4
2.4 (Optional) Buying DGC Incremental Packages....8
2.5 Creating Workspaces....14
2.6 Authorizing IAM Users to Use DGC....16
2.6.1 Creating Users and Granting DGC Permissions....17
2.6.2 Adding Members and Assigning Roles....17
2.7 Obtaining the Authentication Information....19
3 Workspace Management....21
3.1 Managing Workspaces....21
3.2 Changing Log Storage Paths....24
3.3 Managing Members....25
4 Preparing Data Sources....28
5 Management Center....29
5.1 Data Sources....29
5.2 Creating Data Connections....34
5.3 Migrating Resources....48
6 Data Integration....50
6.1 Cloud Data Migration....50
6.1.1 Overview....50
6.1.2 Supported Data Sources....52
6.1.3 Creating Links....75
6.1.3.1 Creating Links....75
6.1.3.2 Managing Drivers....79
6.1.3.3 Editing/Deleting Links....81
6.1.3.4 Link to Hive....82
6.1.3.5 Link to CloudTable....87
6.1.3.6 Link to an FTP or SFTP Server....89

 Copyright © Huawei Technologies Co., Ltd.

6.1.3.7 Link to NAS/SFS 90
6.1.3.8 Link to MongoDB 90
6.1.3.9 Link to DDS 91
6.1.3.10 Link to Redis/DCS 92
6.1.3.11 Link to Kafka 92
6.1.3.12 Configuring Links to DIS 93
6.1.3.13 Link to Elasticsearch/CSS 94
6.1.3.14 Link to DLI 95
6.1.3.15 Link to CloudTable OpenTSDB 96
6.1.3.16 Link to DMS Kafka 96
6.1.3.17 Link to HBase 97
6.1.3.18 Link to HDFS 102
6.1.3.19 Link to Amazon S3 109
6.1.3.20 Link to KODO/COS 109
6.1.3.21 Link to OSS on Alibaba Cloud 110
6.1.3.22 Link to Relational Databases 110
6.1.3.23 Link to OBS 113
6.1.4 Table/File Migration Jobs 114
6.1.5 Source Job Parameters 124
6.1.5.1 Source OBS Parameters 125
6.1.5.2 From HDFS 130
6.1.5.3 From HBase/CloudTable 135
6.1.5.4 From Hive 136
6.1.5.5 Source FTP/SFTP/NAS/SFS Parameters 138
6.1.5.6 Source HTTP/HTTPS Parameters 143
6.1.5.7 Source Relational Database Parameters 144
6.1.5.8 Source MongoDB/DDS Parameters 148
6.1.5.9 Source Redis Parameters 149
6.1.5.10 Source DIS Parameters 150
6.1.5.11 Source Apache Kafka/DMS Kafka Parameters 151
6.1.5.12 Source Elasticsearch or CSS Parameters 152
6.1.5.13 Source OpenTSDB Parameters 153
6.1.6 Destination Job Parameters 153
6.1.6.1 Destination OBS Parameters 153
6.1.6.2 Destination HDFS Parameters 161
6.1.6.3 Destination HBase/CloudTable Parameters 164
6.1.6.4 Destination Hive Parameters 165
6.1.6.5 Destination FTP/SFTP/NAS/SFS Parameters 167
6.1.6.6 Destination Relational Database Parameters 172
6.1.6.7 Destination DDS Parameters 176
6.1.6.8 Destination DCS Parameters 177
6.1.6.9 Destination Elasticsearch or CSS Parameters 177

 Copyright © Huawei Technologies Co., Ltd.

6.1.6.10 Destination DLI Parameters....178
6.1.6.11 Destination DIS Parameters....179
6.1.6.12 Destination OpenTSDB Parameters....179
6.1.7 Entire DB Migration....180
6.1.8 Scenario-based Migration....189
6.1.9 Incremental Migration....197
6.1.9.1 Incremental File Migration....197
6.1.9.2 Incremental Migration of Relational Databases....200
6.1.9.3 Incremental Synchronization Using the Macro Variables of Date and Time....202
6.1.9.4 HBase/CloudTable Incremental Migration....206
6.1.10 Job Management....207
6.1.10.1 Scheduling Job Execution....207
6.1.10.2 Job Configuration Management....211
6.1.10.3 Single Job Management....213
6.1.10.4 Batch Job Management....215
6.1.10.5 Agent Management....217
6.1.11 Cluster Management....220
6.1.11.1 Binding or Unbinding EIPs....220
6.1.11.2 Modifying Cluster Configurations....221
6.1.11.3 Viewing Cluster Configurations, Logs, and Monitoring Data....224
6.1.12 Typical Data Migration Scenarios....225
6.1.12.1 Migrating Data from DDS to DWS....226
6.1.12.2 Migrating Data from OBS to CSS....232
6.1.12.3 Migrating Data from OBS to DLI....236
6.1.12.4 Migrating Data from Redis to DCS....241
6.1.12.5 Migrating Data Oracle to CSS....246
6.1.12.6 Migrating Data from MySQL to MRS Hive....251
6.1.12.7 Migrating Data from OSS to OBS....261
6.1.12.8 Migrating Data from OSS to SFS....265
6.1.12.9 Migrating Data from FTP/SFTP to OBS....272
6.1.12.10 Performing Incremental Migration on CDM by Data Development....278
6.1.12.11 Migrating the Entire MySQL Database to RDS....287
6.1.12.12 Migrating the Entire Elasticsearch Database to Cloud Search Service....292
6.1.13 Advanced Operations....296
6.1.13.1 Migration in Transaction Mode....296
6.1.13.2 Encryption and Decryption During File Migration....297
6.1.13.3 MD5 Verification....301
6.1.13.4 Field Conversion....302
6.1.13.5 Migration of a List of Files....311
6.1.13.6 Regular Expressions for Separating Semi-structured Text....312
6.1.13.7 File Formats....316
6.1.14 Monitoring....325

 Copyright © Huawei Technologies Co., Ltd.

6.1.14.1 CDM Metrics 325
6.1.14.2 Configuring Alarm Rules 326
6.1.14.3 Querying Metrics 327
6.1.15 Troubleshooting 329
6.1.15.1 What Do I Do If the Log Prompts that the Date Format Fails to Be Parsed? 329
6.1.15.2 What Can I Do If the Map Field Tab Page Cannot Display All Columns? 331
6.1.15.3 How Do I Select Distribution Columns When Using CDM to Migrate Data to DWS? 335
6.1.15.4 What Do I Do If the Error Message "value too long for type character varying" Is Displayed When I Migrate Data to DWS? 336
6.1.15.5 What Can I Do If Error Message "Unable to execute the SQL statement" Is Displayed When I Import Data from OBS to SQL Server? 337
6.2 Real-time Data Ingestion 338
6.2.1 Overview 338
6.2.2 Data Sources Supported by DIS 338
6.2.3 Quick Start 339
6.2.4 Managing DIS Streams 350
6.2.4.1 Viewing Stream Monitoring Metrics 351
6.2.4.2 Changing a Source Data Type 352
6.2.4.3 Managing a Source Data Schema 353
6.2.4.4 Managing Stream Tags 356
6.2.4.5 Managing Applications 358
6.2.4.6 Managing Authorization 358
6.2.4.7 Scaling Up/Down a Stream 359
6.2.5 Uploading Data by Using Agent 361
6.2.5.1 DIS Agent Overview 361
6.2.5.2 Preparing for Installing DIS Agent 362
6.2.5.3 Installing DIS Agent 364
6.2.5.4 Configuring a DIS Agent 364
6.2.5.5 Starting DIS Agent 372
6.2.5.6 Testing DIS Agent 372
6.2.5.7 Stopping DIS Agent 375
6.2.6 Using DIS Flume Plugin to Upload and Download Data 375
6.2.6.1 Overview 375
6.2.6.2 Preparing for Installing a DIS Flume Plugin 376
6.2.6.3 Installing a DIS Flume Plugin 377
6.2.6.4 Configuring the DIS Flume Plugin 377
6.2.6.5 Testing a DIS Flume Plugin 383
6.2.6.6 (Optional) Uninstalling a DIS Flume Plugin 386
6.2.7 Using a DIS Logstash Plugin to Upload and Download Data 386
6.2.7.1 Overview 386
6.2.7.2 Preparing for Installing DIS Logstash Plugin 387
6.2.7.3 Installing the DIS Logstash Plugin 388
6.2.7.4 Installing DIS Logstash Plugin Online 388

 Copyright © Huawei Technologies Co., Ltd.

6.2.7.5 Installing DIS Logstash Plugin Offline 389
6.2.7.6 Configuring the DIS Logstash Plugin 390
6.2.7.7 Testing the DIS Logstash Plugin 393
6.2.7.8 (Optional) Uninstalling the DIS Logstash Plugin 394
6.2.8 Using Kafka Adapter to Upload and Download Data 394
6.2.8.1 Kafka Adapter Overview 394
6.2.8.2 Preparations 394
6.2.8.3 Uploading Data 396
6.2.8.4 Downloading Data 402
6.2.8.5 Consumption Modes 403
6.2.8.6 Consumption Offset 410
6.2.8.7 Adaptation to the Native KafkaConsumer API 411
6.2.9 Using SparkStreaming SDK to Download Data 416
6.2.9.1 Overview 416
6.2.9.2 Preparing a DIS Spark Streaming Environment 417
6.2.9.3 Customizing a SparkStreaming Job 418
6.2.10 Dumping Stream Data to Cloud Services 424
6.2.10.1 Creating a Dump Task 424
6.2.10.2 Dumping Data to OBS 426
6.2.10.3 Dumping Data to DLI 435
6.2.10.4 Dumping Data to DWS 436
6.2.10.5 Dumping Data to MRS 440
6.2.10.6 Dumping Data to CloudTable 444
6.2.11 Event Notification 449
6.2.11.1 Event Notification Overview 449
6.2.11.2 Subscribing to Event Notification 450
6.2.11.3 Viewing Events 453

7 Data Design 454

7.1 Data Design Introduction 454
7.2 Overview 457
7.3 Adding Reviewers 458
7.4 Information Architectures 459
7.5 Designing Processes 462
7.6 Designing Subjects 465
7.7 Managing Lookup Tables 469
7.8 Data Standards 480
7.9 ER Modeling 489
7.9.1 Designing Physical Models 489
7.9.2 Designing Logical Models 500
7.9.3 Reversing Databases 508
7.9.4 Associating with Quality Rules 510
7.9.5 Importing/Exporting Tables 515

 Copyright © Huawei Technologies Co., Ltd.

7.9.6 Viewing Tables 523
7.9.7 Modifying Subjects, Directories, and Processes 526
7.10 Dimensional Modeling 528
7.10.1 Creating Dimensions 528
7.10.2 Managing Dimension Tables 535
7.10.3 Creating Fact Tables 539
7.10.4 Creating Summary Tables 549
7.10.5 Reversing Databases 557
7.11 Business Metrics 559
7.12 Technical Metrics 564
7.12.1 Creating Atomic Metrics 564
7.12.2 Creating Time Filters 567
7.12.3 Creating Derivative Metrics 570
7.12.4 Creating Compound Metrics 574
7.12.5 Creating General Filters 577
7.13 Review Center 580
7.14 Configuration Center 584

8 Data Development 592

8.1 Overview 592
8.2 Data Management 593
8.2.1 Creating a Database 593
8.2.2 Modifying a Database 594
8.2.3 Deleting a Database 594
8.2.4 Creating a Database Schema 595
8.2.5 Modifying a Database Schema 595
8.2.6 Deleting a Database Schema 596
8.2.7 Creating a Data Table (Visualized Mode) 596
8.2.8 Creating a Data Table (DDL Mode) 603
8.2.9 Viewing Data Table Details 604
8.2.10 Deleting a Data Table 604
8.2.11 Viewing Table Details 605
8.3 Script Development 605
8.3.1 Creating a Script 605
8.3.2 Developing an SQL Script 607
8.3.3 Developing a Shell Script 611
8.3.4 Renaming a Script 613
8.3.5 Moving a Script 615
8.3.6 Exporting and Importing a Script 617
8.3.7 Deleting a Script 618
8.3.8 Copying a Script 619
8.4 Job Development 620
8.4.1 Job Development Overview 620

 Copyright © Huawei Technologies Co., Ltd.

8.4.2 Creating a Job 621
8.4.3 Developing a Job 623
8.4.4 Renaming a Job 631
8.4.5 Moving a Job 633
8.4.6 Exporting and Importing a Job 634
8.4.7 Deleting a Job 637
8.4.8 Copying a Job 638
8.5 Nodes 638
8.5.1 Node Overview 639
8.5.2 Node Lineages 639
8.5.3 CDM Job 643
8.5.4 DIS Stream 649
8.5.5 DIS Dump 651
8.5.6 DIS Client 653
8.5.7 Rest Client 655
8.5.8 Import GES 663
8.5.9 MRS Kafka 665
8.5.10 Kafka Client 667
8.5.11 DLI Flink Job 668
8.5.12 DLI SQL 672
8.5.13 DLI Spark 678
8.5.14 DWS SQL 680
8.5.15 MRS SparkSQL 686
8.5.16 MRS Hive SQL 688
8.5.17 MRS Spark 690
8.5.18 MRS Spark Python 692
8.5.19 MRS Flink Job 694
8.5.20 MRS MapReduce 696
8.5.21 CSS 698
8.5.22 Shell 700
8.5.23 RDS SQL 702
8.5.24 ETL Job 704
8.5.25 CS Job 710
8.5.26 OCR 714
8.5.27 Create OBS 716
8.5.28 Delete OBS 717
8.5.29 OBS Manager 719
8.5.30 Open/Close Resource 721
8.5.31 Subjob 722
8.5.32 SMN 724
8.5.33 Dummy 727
8.5.34 For Each 728

 Copyright © Huawei Technologies Co., Ltd.

8.6 EL 730
8.6.1 Expression Overview 730
8.6.2 Date and Time Mode 731
8.6.3 Env Embedded Objects 732
8.6.4 Job Embedded Objects 733
8.6.5 StringUtil Embedded Objects 735
8.6.6 DateUtil Embedded Objects 735
8.6.7 JSONUtil Embedded Objects 736
8.6.8 Loop Embedded Objects 737
8.6.9 OBSUtil Embedded Objects 738
8.6.10 Expression Use Example 738
8.7 Solution 741
8.8 O&M and Scheduling 742
8.8.1 Overview 742
8.8.2 Monitoring a Batch Job 743
8.8.3 Monitoring a Real-Time Job 748
8.8.4 Monitoring a Real-Time Sub-job 751
8.8.5 Monitoring an Instance 753
8.8.6 Monitoring PatchData 754
8.8.7 Managing an Notification 755
8.8.8 Cycle Overview 757
8.8.9 Backing Up and Restoring an Asset 759
8.9 Configuration and Management 761
8.9.1 Managing a Host Connection 761
8.9.2 Managing a Resource 763
8.9.3 Configuring an Environment Variable 766
8.9.4 Managing Job Labels 768
8.9.5 Configuring Agencies 769
8.10 Usage Guidance 776
8.10.1 Developing a Spark Job 776
8.10.2 Developing Hive SQL Scripts 779
8.10.3 Cross-Workspace Job Scheduling 783
9 Data Quality Control 786
9.1 Monitoring Business Metrics 786
9.1.1 Creating Metrics 786
9.1.2 Creating Rules 788
9.1.3 Creating Business Scenarios 790
9.1.4 O&M 792
9.2 Monitoring Data Quality 794
9.2.1 Purpose 794
9.2.2 Creating Rule Templates 795
9.2.3 Creating Data Quality Jobs 798

 Copyright © Huawei Technologies Co., Ltd.

9.2.4 Create Comparison Jobs 802
9.2.5 Viewing Rule Instances 808
10 Data Assets 810
10.1 Data Maps 810
10.1.1 Overview 810
10.1.2 Dashboard 810
10.1.3 Data Catalogs 811
10.1.4 Tags 813
10.2 Data Permissions 816
10.2.1 About Data Permissions 816
10.2.2 Data Catalog Permissions 816
10.2.3 Data Table Permissions 817
10.2.4 Review Center 820
10.3 Data Security 821
10.3.1 About Data Security 821
10.3.2 Data Security Levels 821
10.3.3 Data Classifications 822
10.3.4 Masking Policies 824
10.4 Metadata Collection 825
10.4.1 Overview 825
10.4.2 Task Management 825
10.4.3 Task Monitoring 833
11 Data Lake Mall 835
11.1 Overview 835
11.2 Comparison Between Shared Edition and Exclusive Edition 838
11.3 Glossary 842
11.4 Adding Reviewers 843
11.5 Generating APIs 844
11.5.1 Function Overview 844
11.5.2 Buying Exclusive Instances 844
11.5.3 Generating Wizard APIs 847
11.5.4 Generating Script APIs 852
11.6 Registering APIs 855
11.7 Debugging APIs 858
11.8 Authorizing APIs 860
11.9 Publishing APIs 861
11.10 Reviewing APIs 861
11.11 Suspending APIs 863
11.12 Deleting APIs 863
11.13 Calling APIs 864
11.14 Operating APIs 866
11.14.1 Creating Throttling Policies 866

 Copyright © Huawei Technologies Co., Ltd.

11.15 Migrating Resources 869

A Change History 873

 Copyright © Huawei Technologies Co., Ltd.

1 DGC Introduction

DGC is a one-stop data lake 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 DGC, it is easier for enterprises to construct end-to-end intelligent data systems. These systems can help eliminate data silos, unify data standards, accelerate data monetization, and promote digital transformation.

DGC Users

There are four types of DGC users: administrator, developer, O&M engineer, and visitor. The four types of users have different DGC permissions. For details, see **DGC Permission List**.

- **Administrator**

  This role is granted with the management, decision-making, and review permissions. Management personnel who are familiar with enterprise businesses can be assigned the administrator role. An administrator has the permissions of both developers and reviewers. Users who have the administrator role can perform any operations in DGC, for example, they can manage workspace members and data assets, and configure jobs. In Data Design and Data Lake Mall, operations like publish and suspend data models and APIs must be reviewed by an administrator 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 DGC 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 data service APIs provided by DGC 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 DGC permissions.

- **O&M Engineer**

  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. DGC Data Assets visualizes all data links. Data quality can be verified,

 Copyright © Huawei Technologies Co., Ltd.

controlled, and traceable. O&M personnel who have the O&M engineer role can schedule and monitor jobs from end to end. Data collection, consumption, and O&M are all one-stop services.

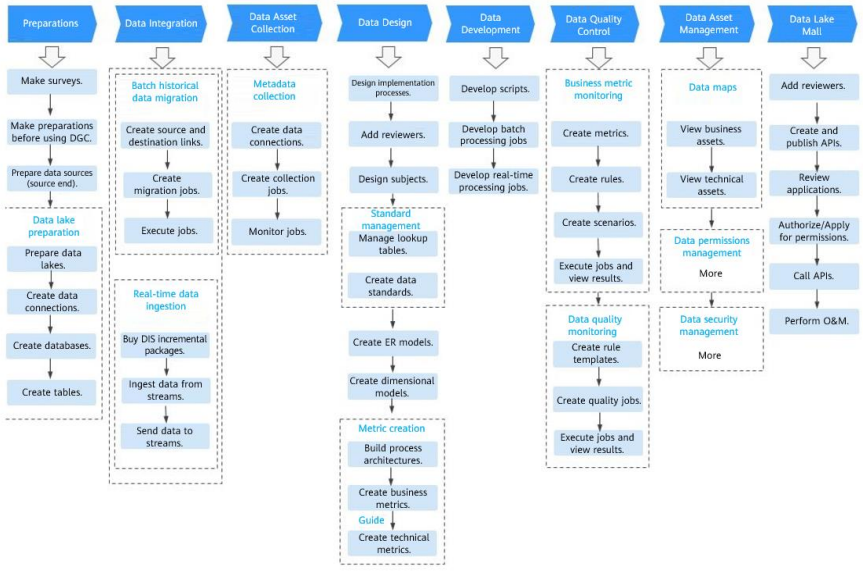
- **Visitor**

  This role is granted with the read-only permissions.

General Use Procedure

The following figure shows a typical end-to-end development process with DGC:

Figure 1-1 DGC use procedure



 Copyright © Huawei Technologies Co., Ltd.

2 Initial Preparations

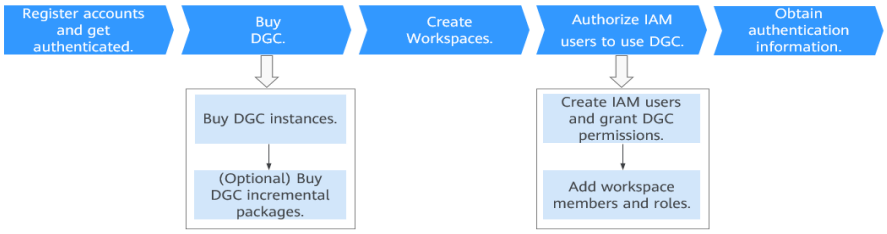
Before using DGC, you need to have an account registered with HUAWEI CLOUD or an IAM user that can access DGC. This document describes how to register a HUAWEI CLOUD account, create an IAM user, and obtain a DGC instance.

2.1 Making Preparations Before Using DGC

Before using DGC, register a HUAWEI CLOUD account and buy a DGC instance. See **Figure 2-1**.

For more details, see subsequent sections.

Figure 2-1 Preparation process



2.2 Registering an Account and Getting Authenticated

An account is an entity that accesses HUAWEI CLOUD. For details, see **My Account**. This section describes how to register a HUAWEI CLOUD account. If you already have a HUAWEI CLOUD account, skip this part.

1. Visit **HUAWEI CLOUD**.
2. In the upper right corner of the displayed page, click **Register**.
3. Complete the registration as instructed.

You are advised to complete real-name authentication for better experience in using cloud services. Real-name authentication is available for two types of accounts: individual account and enterprise account. For details on the authentication methods for different account types, click the following link: (If your account has passed the real-name authentication, skip this part.)

 Copyright © Huawei Technologies Co., Ltd.

- **Individual Real-Name Authentication Process**
- **Enterprise Real-Name Authentication Process**

NOTE

- After you send the application for real-name authentication, it will take three work days at most before the application is approved.
- The account takes effect 40 minutes later after your real-name authentication is approved.

2.3 Buying DGC Instances

DGC is billed on a package basis. The packages include **Starter**, **Basic**, **Advanced**, **Professional**, and **Enterprise**. They are billed yearly or monthly. Billing items include data integration, number of tasks scheduled per day, and number of assets supported by data governance.

- **Starter**: This version is applicable to data lake projects where no full-time data development personnel and data governance are required.
- **Basic**: This version is applicable to scenarios where 1 or 2 full-time data development personnel are required and no more than 1,000 original data tables are involved.
- **Advanced**: This version is applicable to scenarios where 5 to 10 full-time data development personnel are required, clear data standards and efficient data quality management are available, and 2,000 data tables are involved.
- **Professional**: This version is applicable to medium and large enterprises that have excellent data management teams, complete, high-quality systems, and 10 to 30 full-time data development personnel.
- **Enterprise**: This version is applicable to large enterprises with multiple branches.

Context

- Only users with the **DGC Administrator** or **Tenant Administrator** permissions can buy DGC instances or incremental packages. You need to grant the required permissions to users.

  NOTE

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

Prerequisites

You have applied for a VPC, subnet, and security group.

Buying a DGC Instance and Accessing the DGC Console

1. Log in to the **public cloud management console**.
2. Click in the upper left corner of the displayed page to select a region.

 Copyright © Huawei Technologies Co., Ltd.

3. In the upper left corner of the console, click **Service List** and choose **EI Enterprise Intelligence** > **DGC** to access the DGC console.
4. On the displayed page, click **Buy DGC Instance** to navigate to the page where you can buy an instance.
5. Configure the DGC instance parameters. See **Table 2-1** for parameter details.

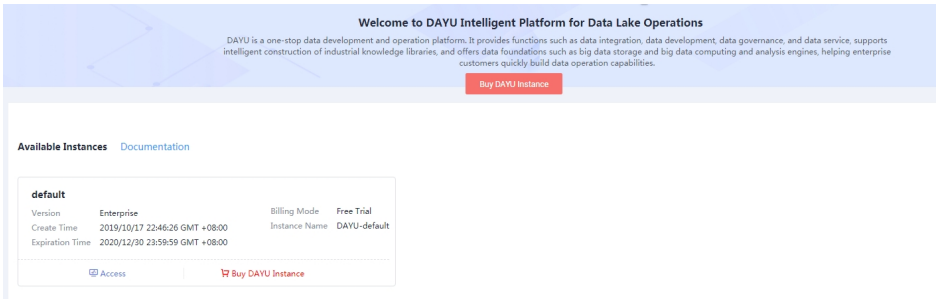
Table 2-1 DGC instance parameters

Parameter	Example	Description
Region	CN North-Beijng1	The region where the DGC instance resides. Resources in different regions cannot communicate with each other. Take care when selecting the region.
AZ	AZ2	The availability zone (AZ). Select an AZ by referring to **Region and AZ**.
Instance Name	DGC-test	The name of your DGC instance.
VPC	vpc1	The VPC, subnet, and security group where the DGC instance resides, which are used to communicate with the desired data source. For more information, see **Virtual Private Cloud User Guide**.
Subnet	subnet-1	
Security Group	sg-1	
Enterprise Project	default	The enterprise project associated with your DGC instance. An enterprise project facilitates management of cloud resources. For more information, see **Enterprise Management User Guide**. You can manage multiple projects in an enterprise and separately settle their costs. This parameter is available only when an enterprise project has been created on the **Enterprise Project Management** page.
Renew	N/A	Select **Auto-renew** to automatically renew the subscription by month or by year.

6. If there is no need to change the default settings, click **Buy Now** to display the page for confirming the order.
7. Click **Next**. After you pay for your instance, the instance you bought shows up on the DAYU homepage.

Copyright © Huawei Technologies Co., Ltd.

Figure 2-2 Viewing a DGC instance



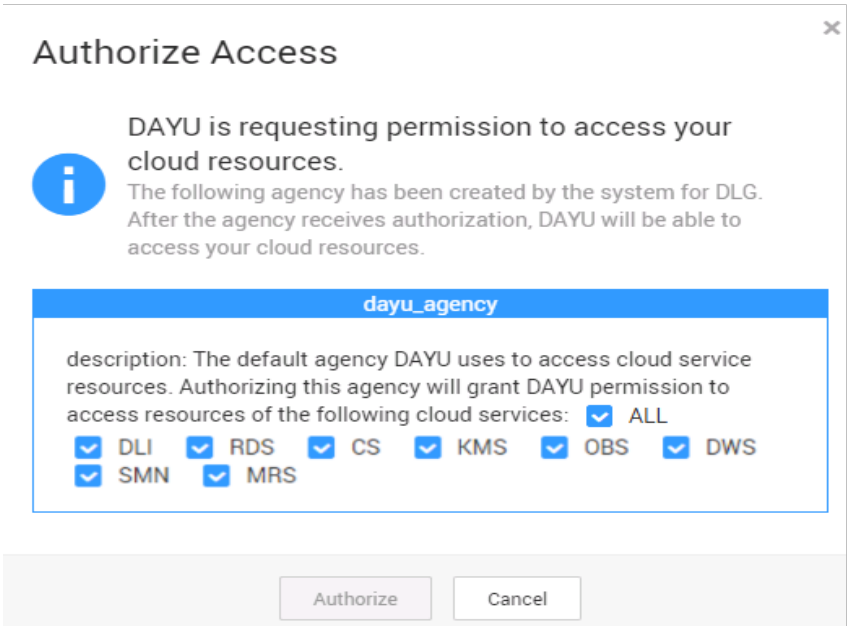
8. When you return to the DGC management console, the **Authorize Access** dialog box is displayed, prompting you to authorize the listed services.

   Select all services and click **Agree**. The platform automatically creates an agency. After the authorization is complete, the **Authorize Access** dialog box will not be displayed when you access the DGC console home page next time.

   If you select only some services for authorization, the system still displays the dialog box next time you access the DGC console homepage, prompting you to authorize access to unauthorized cloud services.

   DGC interacts with the cloud services listed in **Figure 2-3** and needs to collaborate with these cloud services. Therefore, you need to create a cloud service agency to delegate permissions to DGC so that DGC can use these cloud services and perform task scheduling and resource O&M on behalf of you.

Figure 2-3 Authorize Access dialog box



9. In the list of instances, locate the instance you just bought and click **Access** to access the DGC management console.

Trying a Basic DGC Instance and Entering the Console

Context

 Copyright © Huawei Technologies Co., Ltd.

- Each account has only one trial use opportunity.
- A basic DGC instance for free trial does not include CDM clusters. If you want to use CDM, buy a CDM incremental package. For details, see **(Optional) Buying DGC Incremental Packages**.
- You can use a basic DGC instance for free for 30 days. After the trial period expires, you need to buy an instance of one of the official versions provided.
- If you try the basic version, but you then buy the starter version, the resources created during the trial period will be released.
- During the trial period, you can buy any official version of DGC, but the trial period will expire immediately.
- Resources created during the trial period cannot be accessed once the trial period expires. After the trial period expires, data will be deleted seven days later. You will take full responsibility for any losses resulting from not buying a DGC instance in time. To access the free trial, follow the following instructions:

Procedure

1. Log in to the **public cloud management console**.
2. Click in the upper left corner of the displayed page to select a region.
3. In the upper left corner of the console, click **Service List** and choose **EI Enterprise Intelligence** > **DGC** to access the DGC console.
4. Click **Try Basic**. Enter the page for trying a basic DGC instance.
5. Configure the DGC instance parameters. See **Table 2-1** for parameter details.

Table 2-2 DGC instance parameters

Parameter	Example	Description
Region	CN North-Beijng1	The region where the DGC instance resides. Resources in different regions cannot communicate with each other. Take care when selecting the region.
Enterprise Project	default	This parameter is available only when an enterprise project has been created on the **Enterprise Project Management** page. An enterprise project facilitates management of cloud resources. For more information, see **Enterprise Management User Guide**.
Instance Name	DGC-test	The name of your DGC instance.

6. If there is no need to change the current configuration details, click **Try Now**.
7. In the **Free Trial Agreement dialog** box that is displayed, select **I have read and agreed to these terms** and click **OK**.
8. In the list of trial instances, locate your instance and click **Access** to access the DGC console.

 Copyright © Huawei Technologies Co., Ltd.

2.4 (Optional) Buying DGC Incremental Packages

DGC provides users with the basic package and the incremental package. The basic package is billed on a yearly or monthly basis. The incremental is billed by usage or yearly/monthly.

This section describes how to buy a DGC incremental package. Before buying a DGC incremental package, ensure that you have bought DGC instances.

- DIS incremental package

  DIS addresses the challenge of transmitting data from outside the cloud to the inside. It builds data intake streams for custom applications capable of processing or analyzing streaming data.

  If you want to use DGC DIS, you need to buy a DIS incremental package after buying a DGC instance. Otherwise, skip this part. For details on how to buy a DIS incremental package, see the following links:

  a. **Buying a Pay-Per-Use DIS Incremental Package**
  b. (Optional) **Buying a Yearly/Monthly DIS Incremental Package**

- CDM incremental package

  In the case that a DGC instance includes a CDM cluster (except for the trial version), and the cluster meets your requirements, you do not need to buy a CDM incremental package.

  CDM can migrate data in batches between more than 20 homogeneous or heterogeneous data sources. For details on how to buy a CDM incremental package, see the following links:

  a. **Buying a Pay-Per-Use CDM Incremental Package**
  b. (Optional) **Buying a Yearly/Monthly CDM Incremental Package**

- DLM incremental package (exclusive cluster)

  A DGC instance does not include an exclusive DLM cluster. If you want to use DLM to create APIs, you need to create a DLM exclusive cluster. See **Creating an Exclusive Cluster** for details.

Context

- When you buy a pay-per-use incremental package, the system automatically creates a CDM cluster or DIS stream based on your selected specifications.
- A yearly/monthly incremental package does not include a CDM cluster or DIS stream. If you only buy an incremental package, the package does not affect the purchased duration because there is no CDM cluster or DIS stream.
- If you want to enjoy the preferential price afforded by a yearly/monthly incremental package, you need to buy a pay-per-use incremental package and a yearly/monthly incremental package in the same region and workspace. The two packages must have the same specifications.

Buying a Pay-Per-Use DIS Incremental Package

1. Select an enabled instance and click **Buy**.

 Copyright © Huawei Technologies Co., Ltd.

2. Access the page for buying a DGC incremental package and set parameters. See **Table 2-3** for parameter details.

Figure 2-4 Configuring parameters for a DIS incremental package

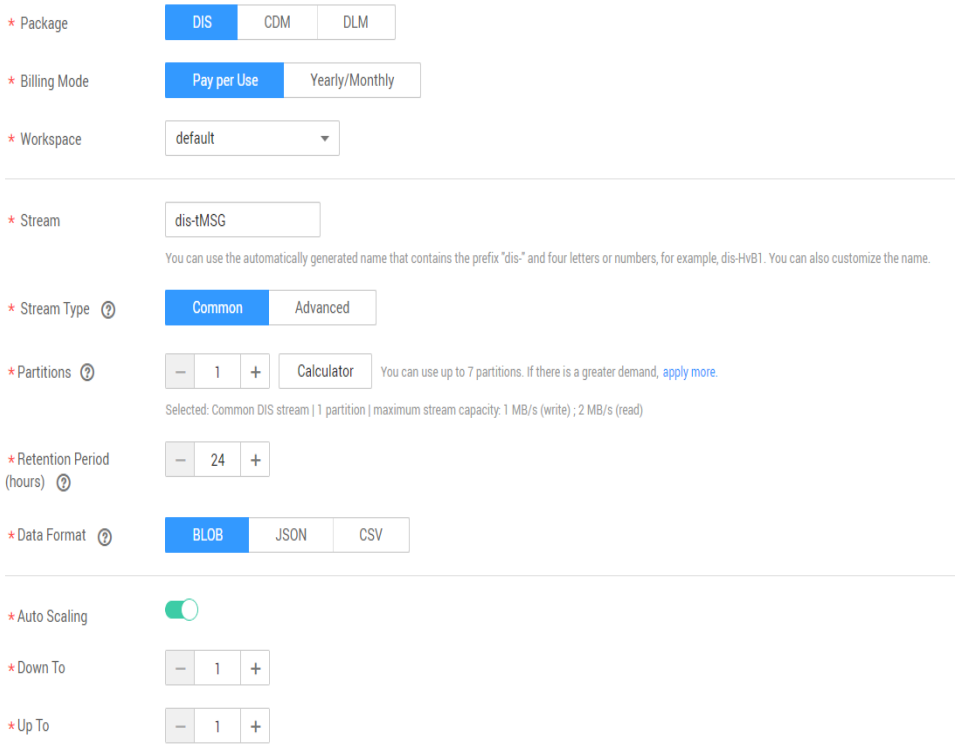


Table 2-3 Parameters for a DIS incremental package

Parameter	Description
Incremental Package Type	Select **DIS**.
Billing Mode	Select **On-demand**.
Workspace	Select the workspace where the DIS incremental package will be used. For example, if you want to buy a DIS incremental package for workspace ***A*** of instance ***test***, select ***A***. After you bought a DIS incremental package, it will show up in workspace ***A***.
Stream Name	The name of the DIS stream to be created. Stream names must be unique. They can be up to 64 characters long. Only letters, numbers, hyphens (-), and underscores (_) are allowed.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Stream Type	**Stream Type**: **Common** and **Advanced**. ● **Common**: Each partition supports a maximum read speed of 2 MB/s and a maximum write speed of 1 MB/s or 1,000 records/s. ● **Advanced**: Each partition supports a maximum read speed of 10 MB/s and a maximum write speed of 5 MB/s or 2,000 records/s.
Partitions	Partitions are the base throughput unit of a DIS stream. ● For a common DIS stream, the value of **Partitions** is an integer from 1 to 50. A user can create up 50 partitions. ● For an advanced DIS stream, the value of **Partitions** is an integer from 1 to 10. A user can create up to 10 partitions.
Partition Calculator	Used to estimate how many partitions are needed based on the information you entered. 1. Click **Partition Calculator**. 2. In the **Partition Calculator** dialog box, configure the **Average Record Size (KB)**, **Max. Records Written**, and **Consumer Applications** parameters. The **Estimated Partitions** field then displays the recommended number of partitions. The value of this field cannot be modified. **NOTE** Partition calculation formulas: ● Based on the traffic (the final value must be rounded up): Common stream: Average record size x (1 + 20%) x Max. written records/(1 x 1024 KB) (20% is the reserved partition percentage.) Advanced stream: Average record size x (1 + 20%) x Max. written records/(5 x 1024 KB) (20% is the reserved partition percentage.) ● Based on the consumer application quantity (the final value must be rounded up): (The number of consumer applications/2) x The number of partitions calculated based on the traffic (The result of the number of consumer applications/2 must reserve two decimals.) The larger value calculated based on the previous two formulas is considered as the estimated partition value. 3. Click **Use Estimated Value**. The estimated value is automatically used as the value of **Partitions**.
Data Retention (days)	The maximum number of days for which data can be preserved in DIS streams. Data will be deleted when the retention period expires.

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Source Data Type	• **BLOB**: a collection of binary data stored as a single entity in a database management system. If **Source Data Type** is set to **BLOB**, the supported **Dump Destination** can be **OBS** or **MRS**. • **JSON**: an open-standard file format that uses easy-to-read text to transmit data objects consisting of attribute values and arrays. If **Source Data Type** is set to **JSON**, the supported **Dump Destination** can be **OBS**, **MRS**, **DLI**, **CloudTable**, or **DWS**. • **CSV**: a simple text format for storing tabular data in a plain text file. Commas are used as delimiters. If **Source Data Type** is set to **CSV**, the supported **Dump Destination** can be **OBS**, **MRS**, **DLI**, or **DWS**.
Auto Scaling	You can enable or disable the auto scaling switch when creating a stream. Set the switch to or , respectively, to enable or disable auto scaling. **NOTE** You can choose whether to enable auto scaling when creating a stream. You can also modify the auto scaling attributes for a stream that has already been created.
Down To	The lower limit for automatic scale-down. The number of target partitions for automatic scale-down cannot drop below this lower limit.
Up To	The upper limit for automatic scale-up. The number of target partitions for automatic scale-up cannot exceed this upper limit.

3. Click **Buy Now**. If there are no changes made to the specifications, submit the order.
4. After you bought a DIS stream, it will show up in the corresponding workspace.

Figure 2-5 Viewing a stream

 Copyright © Huawei Technologies Co., Ltd.

Buying a Yearly/Monthly DIS Incremental Package

1. Select an enabled instance and click **Buy**.
2. Access the page for buying a yearly/monthly DGC incremental package and do the following:
   a. **Incremental Package Type**: Select **DIS**.
   b. **Billing Mode**: Select **Package**.
   c. **Resource Package Type**: Select **Common Stream Partition Hour**.
   d. **Required Duration**: Specify a validity period for the package.
   e. **Quantity**: The number of subscribed packages. For example, if you specify **1 month** for **Required Duration** and **2** for **Quantity**, you will have a 1500-hour quota and the validity period is 1 month.
3. Click **Buy Now**. If there are no changes made to the specifications, submit the order.

Buying a Pay-Per-Use CDM Incremental Package

1. Select an enabled instance and click **Buy**.
2. Access the page for buying a DGC incremental package and set parameters. See **Table 2-4** for parameter details.

Figure 2-6 Configuring parameters for a CDM incremental package

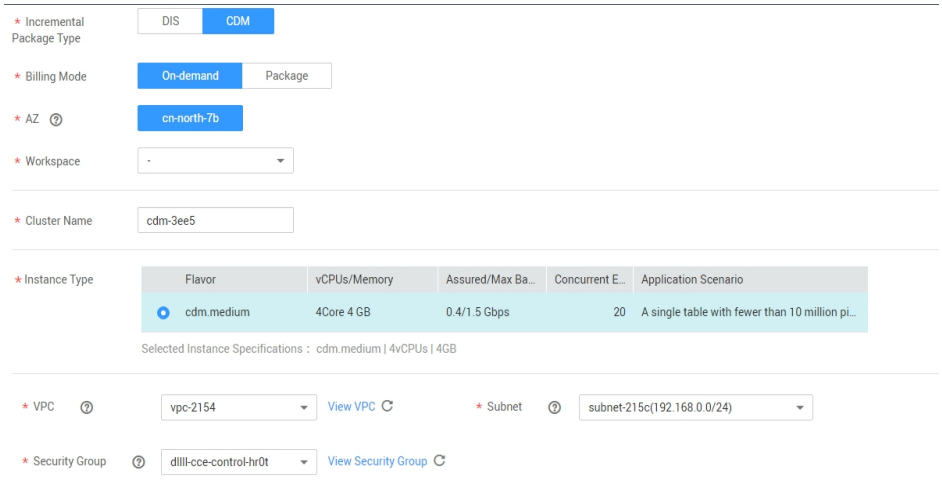


Table 2-4 Parameters for a CDM incremental package

Parameter	Description
Incremental Package Type	Select **CDM**.
Billing Mode	Select **On-demand**.
AZ	The availability zone (AZ). Select an AZ by referring to **Region and AZ**.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Workspace	Select the workspace where the CDM incremental package will be used. For example, if you want to buy a CDM incremental package for workspace ***A*** of instance ***test***, select ***A***. After you bought a cluster, it will show up in the ***A*** workspace.
Cluster Name	The name of your CDM cluster.
Instance Type	The following lists the specifications supported by the CDM cluster. • **cdm.xlarge**: 16 vCPUs and 32 GB of memory. The maximum and assured bandwidths are 10 Gbit/s and 4 Gbit/s. Up to 100 jobs can be concurrently executed. This flavor is well suited to migration of terabytes of data and 10GE high-speed bandwidth. • **cdm.4xlarge**: 64 vCPUs and 128 GB of memory. The maximum and assured bandwidths are 36 Gbit/s and 40 Gbit/s, respectively. A maximum of 200 jobs can be concurrently executed.
VPC Subnet Security Group	The VPC that the CDM cluster belongs to. The CDM cluster must be able to communicate with the data source to be connected. VPCs can be selected according to the residing networks of the migration source and destination. • If the CDM cluster and the data source to be connected belong to different VPCs or the data source is on-premises, the CDM cluster needs to be bound to an elastic IP address (EIP). • If the data source is a cloud service, and the network of the CDM cluster is the same as that of the cloud service, the CDM cluster does not need to be bound to an EIP. • If the data source is a cloud service, and the CDM cluster and the cloud service are in the same VPC but different subnets, configure security group rules to interconnect the CDM cluster with the cloud service. For more information, see ***Virtual Private Cloud User Guide***.

3. Click **Buy Now**. If there are no changes made to the specifications, submit the order.
4. After you bought a cluster, it will show up in the corresponding workspace.

Buying a Yearly/Monthly CDM Incremental Package

1. Select an enabled instance and click **Buy**.
2. Access the page for buying a yearly/monthly DGC incremental package and do the following:
   a. **Incremental Package Type**: Select **CDM**.

 Copyright © Huawei Technologies Co., Ltd.

b. **Billing Mode**: Select **Package**.

c. **Required Duration**: Specify a validity period for the package.

d. **Quantity**: The number of subscribed packages. For example, if you specify **1 month** for **Usage Duration** and **2** for **Quantity**, you will have a 1490-hour quota and the validity period is 1 month.

3. Click **Buy Now**. If there are no changes made to the specifications, submit the order.

2.5 Creating Workspaces

This section describes how to create workspaces. By default, a workspace named **default** is created for users who bought a DGC instance. They are also assigned the **Administrator** role. You can use the default workspace or create a workspace by referring to this section.

Administrators can manage user (member) permissions, resources, and compute engines in a DGC workspace.

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

You can access the Management Center, Data Assets, Data Quality Control, Data Design, Data Lake Mall, Data Development, and Data Integration modules, but only after your account is added to a workspace and assigned the permissions required to perform such operations.

In the **Workspace Management** tab, users can be added to a workspace and assigned roles such as administrator, developer, O&M engineer, and visitor that are preset in DGC.

Prerequisites

A DGC instance has been bought. See **Buying DGC Instances** for details.

Context

- A workspace is created for any user who bought a DGC instance by default. That user is also assigned the role of administrator.
- The HUAWEI CLOUD account (master account) or IAM user (sub-user) who creates the DGC instance has the permissions required to create a workspace. In a DGC instance created by a master account, if an IAM user of the account needs to create a workspace, the IAM user must be assigned the **DGC Administrator** or **Tenant Administrator** policy. By default, the master account has all the permissions for a DGC instance created by a sub-user.

  HUAWEI CLOUD accounts or users with the **DGC Administrator** or **Tenant Administrator** policy can perform all operations on DGC by default.

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

 Copyright © Huawei Technologies Co., Ltd.

- After workspaces are created, they cannot be deleted. You can disable workspaces when they are no longer needed and enable them again when you need these workspaces.
- Only users with the **Security Administrator** policy and assigned the **Administrator** role can add workspace members.
- Users assigned the **DGC User** policy can access a workspace only after being added as members of the workspace.

Creating a Workspace

1. Log in to the DGC management console using a master account.
2. Click **Workspace Management**.
3. Click **Create**. On the **Workspace Information** page, set parameters described in **Table 2-5**. After the configuration is complete, click **OK**.

Figure 2-7 Workspace information

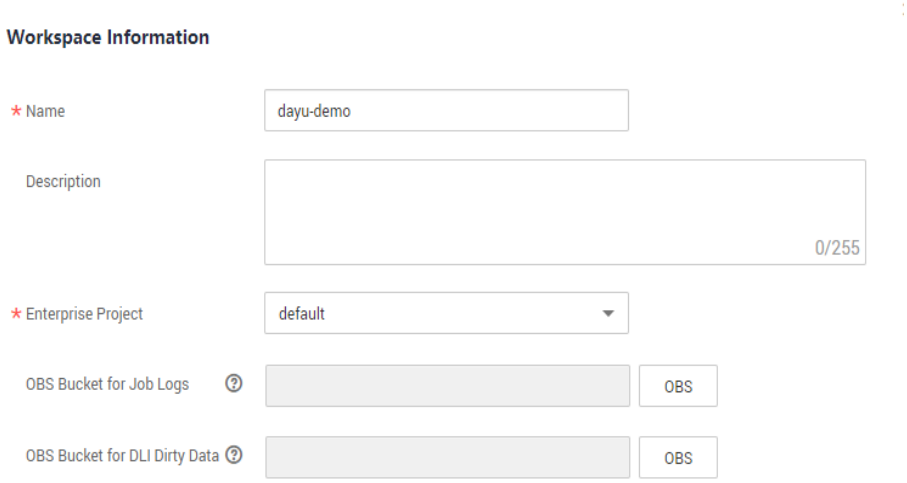


Table 2-5 Parameters for creating a workspace

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

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Enterprise Project	The enterprise project that associated with the workspace. An enterprise project facilitates management of cloud resources. For more information, see **Enterprise Management User Guide**. You can create enterprise projects based on the enterprise organization structure. Then you can manage resources across different regions by enterprise project, grant different permissions to user groups, and add them to enterprise projects. If you have not enabled an enterprise project, you can select enterprise project **default**. On the management console, click **Enterprise** in the upper right corner to access the **Enterprise Project Management** page to create an enterprise project.
OBS Bucket for Job Logs	The OBS bucket for storing logs of DGC data development jobs. To use DGC Data Development, workspace members must have the read and write permissions on the OBS bucket for storing job logs. Otherwise, the system cannot read or write job logs generated during data development. • Click **OBS**. You can select a created OBS bucket. The selected OBS bucket is globally configured in the current workspace. • If this parameter is not set, job logs generated during data development are stored in the OBS bucket named **dlf-log-{projectId}** by default. **{projectId}** indicates the project ID. You can move the pointer to the username in the upper right corner of the console and choose **My Credentials** from the drop-down list. On the **My Credentials** page, you can view the project ID you need.
OBS Bucket for DLI Dirty Data	The OBS bucket for storing dirty data generated during DLI SQL execution in DGC Data Development. To use DGC Data Development to develop and execute DLI SQL statements, workspace members must have the read and write permissions on the OBS bucket where DLI dirty data is stored. Otherwise, the system cannot read or write the dirty data generated during DLI SQL execution. • Click **OBS**. You can select a created OBS bucket. The selected OBS bucket is globally configured in the current workspace. • If this parameter is not set, dirty data generated during DLI SQL execution is stored in the OBS bucket named **dlf-log-{projectId}** by default.

2.6 Authorizing IAM Users to Use DGC

Copyright © Huawei Technologies Co., Ltd.

2.6.1 Creating Users and Granting DGC Permissions

This chapter describes **IAM** fine-grained permissions management for your DGC. With IAM, you can:

- Use your HUAWEI CLOUD account to create IAM users for employees based on the organizational structure of your enterprise. Each IAM user has their own security credentials, providing access to DGC resources.
- Grant users only the permissions required to perform a task.
- Entrust a HUAWEI CLOUD account or cloud service to perform professional and efficient O&M on your DGC resources.

If your HUAWEI CLOUD account does not need individual IAM users for permissions management, skip over this chapter.

This section describes the procedure for granting permissions. See **Creating an IAM User and Granting DGC Permissions** for details.

Context

- Before granting permissions to a user group, familiarize yourself with the workspace role permissions that can be added to the user group and select permissions based on actual requirements. For details on the supported system permissions, see **DGC Permissions Management**.
- For details on the system policies of other services, see **Permissions Policies**.

Creating an IAM User and Granting DGC Permissions

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

   Log in to the **IAM console** using a HUAWEI CLOUD account, create a user group, and assign the DGC permissions to the group. For example, assign **DGC User** to the group. In addition, assign the permissions of other cloud services that DGC depends on. For details, see **DGC Permissions Management**.

   NOTE

   - If an IAM user wants to create a workspace, you need to assign the IAM user the **DGC Administrator** policy. Users with the **DGC Administrator** policy can perform all operations on DGC.
   - Old users are not assigned the **DGC Administrator** policy. If required, log in to the IAM console and assign this role to old users.

2. **Create an IAM user.**

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

2.6.2 Adding Members and Assigning Roles

This section describes how to add workspace members and assign roles to members, as well as the differences between roles.

Workspace members include administrators, developers, O&M engineers, and visitors. The default role of a workspace creator is administrator. For details on role permissions, see **List of permissions**.

Copyright © Huawei Technologies Co., Ltd.

If you only want to use the DGC instance yourself, skip over this section.

If you plan to allow other users to use your DGC instance, create IAM users by referring to **Creating Users and Granting DGC Permissions** and create workspaces by referring to **Creating Workspaces**.

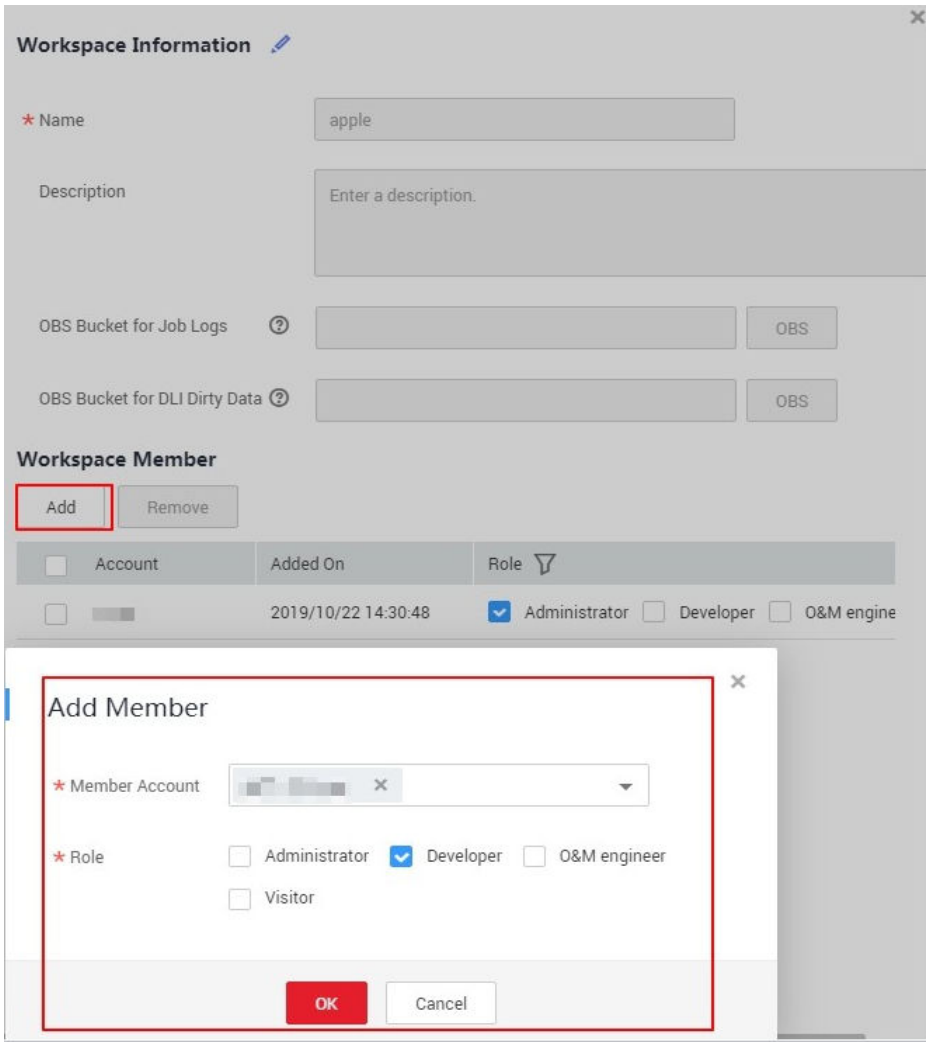
Context

Only the HUAWEI CLOUD account (that is, the master account) that creates the DGC instance or a sub-user who has the **Security Administrator** policy can add members to the workspace.

Adding a Member and Assigning a Role

1. Log in to the DGC management console. Select an instance and click **Access**. On the displayed page, choose the **Workspace Management** tab.
2. Click **Manage Member** in the **Operation** column.
3. Click **Add** under **Workspace Member**. The **Add Member** dialog box is displayed. Select a member account from the drop-down list and select a role for it. Then, click **OK**.

Figure 2-8 Adding a member



 Copyright © Huawei Technologies Co., Ltd.

4. You can view or modify the members and roles in the member list, or delete members from the workspace.

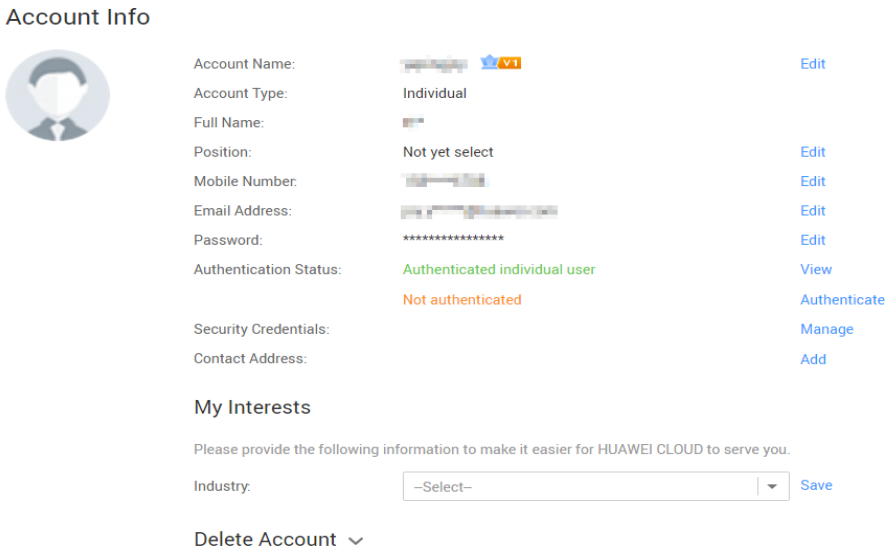
2.7 Obtaining the Authentication Information

When DGC is used for migrating batch data from on-premises servers to the cloud, Object Storage Service (OBS) is required for data storage. You must obtain an Access Key ID (AK) and Secret Access Key (SK) to access OBS.

Obtaining an AK/SK Pair

1. Log in to the **HUAWEI CLOUD**, move your pointer to the username in the upper right corner of the page, and select **Basic Information** from the drop-down list.
2. On the **My Account** page, choose **Account Info** > **Manage**. See **Figure 2-9**.

Figure 2-9 Account Info page



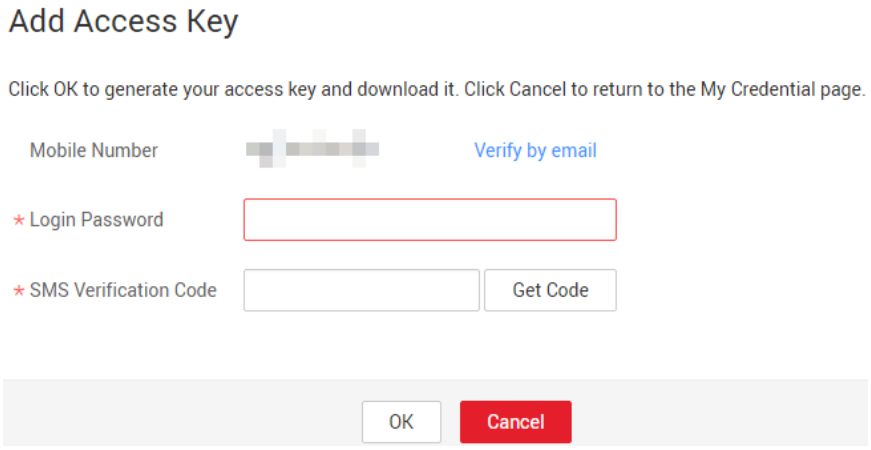
3. On the **My Credentials** page, choose **Access Keys** > **Add Access Key**. See **Figure 2-10**.

Figure 2-10 Access Keys page

4. In the **Add Access Key** dialog box, select where the verification code will be sent, and enter the verification code. See **Figure 2-11**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 2-11 Add Access Key dialog box



5. Click **OK** and save the access key file. The access key file is saved to your browser's configured download location. Open the **credentials.csv** file to view the AK and SK.

Obtaining a Project ID

A project is a group of user resources. To view the project IDs of different regions, choose **My Credentials** > **Projects**.

Obtaining a Region and Endpoint

To obtain regions and endpoints, see **Regions and Endpoints**.

 Copyright © Huawei Technologies Co., Ltd.

3 Workspace Management

3.1 Managing Workspaces

This section describes how to manage workspaces. By default, a workspace named **default** is created for users who bought a DGC instance. They are also assigned the **Administrator** role. You can use the default workspace or create a workspace by referring to this section.

Administrators can manage user (member) permissions, resources, and compute engines in a workspace.

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

You can access **Management Center**, **Data Assets**, **Data Quality Control**, **Data Design**, **Data Integration**, and other components only after you are added to a workspace and assigned the relevant permissions.

In the **Workspace Management** tab, users can be added to a workspace and assigned roles such as **Administrator**, **Developer**, **O&M engineer**, and **Visitor** that are preset in DGC.

Context

- A workspace is created for any user who bought a DGC instance by default. That user is also assigned the role of administrator.
- The HUAWEI CLOUD account (master account) or IAM user (sub-user) who creates the DGC instance has the permissions required to create a workspace. In a DGC instance created by a master account, if an IAM user of the account needs to create a workspace, the IAM user must be assigned the **DGC Administrator** or **Tenant Administrator** policy. By default, the master account has all the permissions for a DGC instance created by a sub-user.

  HUAWEI CLOUD accounts or users with the **DGC Administrator** or **Tenant Administrator** policy can perform all operations on DGC by default.

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

 Copyright © Huawei Technologies Co., Ltd.

- After workspaces are created, they cannot be deleted. You can disable workspaces when they are no longer needed and enable them again when you need these workspaces.
- Only users with the **Security Administrator** policy and assigned the **Administrator** role can add workspace members.
- Users assigned the **DGC User** policy can access a workspace only after being added as members of the workspace.

Creating a Workspace

1. Log in to the DGC management console using a master account.
2. Click **Workspace Management**.
3. Click **Create**. On the **Workspace Information** page, set parameters described in **Table 3-1**. After the configuration is complete, click **OK**.

Figure 3-1 Workspace information

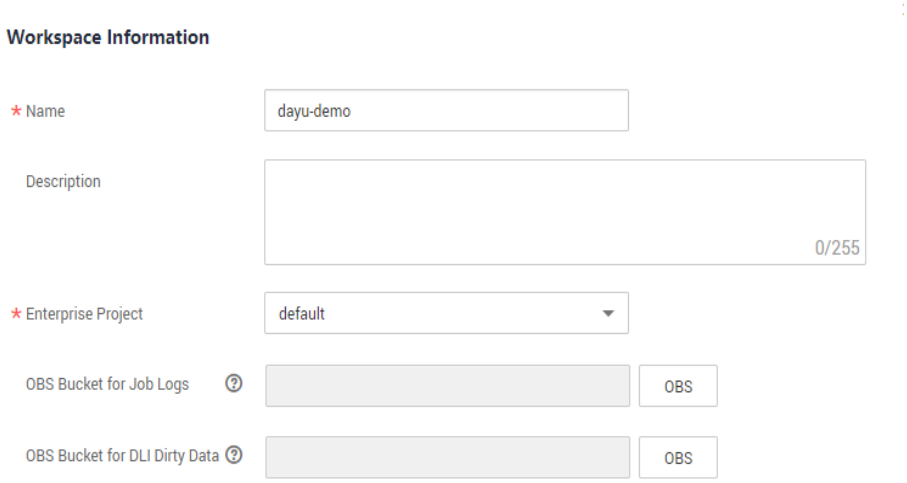


Table 3-1 Parameters for creating a workspace

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

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Enterprise Project	The enterprise project that associated with the workspace. An enterprise project facilitates management of cloud resources. For more information, see **Enterprise Management User Guide**. You can create enterprise projects based on the enterprise organization structure. Then you can manage resources across different regions by enterprise project, grant different permissions to user groups, and add them to enterprise projects. If you have not enabled an enterprise project, you can select enterprise project **default**. On the management console, click **Enterprise** in the upper right corner to access the **Enterprise Project Management** page to create an enterprise project.
OBS Bucket for Job Logs	The OBS bucket for storing logs of DGC data development jobs. To use DGC Data Development, workspace members must have the read and write permissions on the OBS bucket for storing job logs. Otherwise, the system cannot read or write job logs generated during data development. • Click **OBS**. You can select a created OBS bucket. The selected OBS bucket is globally configured in the current workspace. • If this parameter is not set, job logs generated during data development are stored in the OBS bucket named **dlf-log-{projectId}** by default. **{projectId}** indicates the project ID. You can move the pointer to the username in the upper right corner of the console and choose **My Credentials** from the drop-down list. On the **My Credentials** page, you can view the project ID you need.
OBS Bucket for DLI Dirty Data	The OBS bucket for storing dirty data generated during DLI SQL execution in DGC Data Development. To use DGC Data Development to develop and execute DLI SQL statements, workspace members must have the read and write permissions on the OBS bucket where DLI dirty data is stored. Otherwise, the system cannot read or write the dirty data generated during DLI SQL execution. • Click **OBS**. You can select a created OBS bucket. The selected OBS bucket is globally configured in the current workspace. • If this parameter is not set, dirty data generated during DLI SQL execution is stored in the OBS bucket named **dlf-log-{projectId}** by default.

Editing a Workspace

1. Log in to the DGC management console.

 Copyright © Huawei Technologies Co., Ltd.

2. Select a DGC instance and click **Access** on it. Then, click the **Workspace Management** tab.
3. On the **Workspace Management** page, locate the workspace you want to edit and click **Edit** in the row where the workspace is located. The **Workspace Information** page is displayed.
4. On the top of the **Workspace Information** page, click to edit the workspace information and manage workspace members.
5. After the configuration is complete, click at the top of the **Workspace Information** page to save the settings.

Disabling a Workspace

After a workspace is created, it is enabled by default. If you no longer need a workspace, you can disable it. If you need the workspace in the future, you can enable it again.

NOTE

After a workspace is disabled, you cannot access the workspace, edit the workspace content, or schedule jobs in the workspace.

The CDM clusters and DIS streams bought for the workspace are still billed.

1. Log in to the DGC management console.
2. Select a DGC instance and click **Access** on it. Then, click the **Workspace Management** tab.
3. On the **Workspace Management** page, locate the workspace you want to disable, and turn off in the **Status** column.
4. In the **Disable Workspace** dialog box, view the impact of disabling a workspace. If you want to continue, click **OK**.

Enabling a Workspace

1. Log in to the DGC management console.
2. Select a DGC instance and click **Access** on it. Then, click the **Workspace Management** tab.
3. On the **Workspace Management** page, locate the workspace you want to disable, and turn on in the **Status** column.
4. In the **Enable Workspace** dialog box, view the impact of enabling a workspace. If you want to continue, click **OK**.

3.2 Changing Log Storage Paths

By default, job logs and DLI dirty data are stored in an OBS bucket named **dlf-log-**{*project ID*}. You can customize a log storage path. Data Development allows you to configure an OBS bucket globally based on the workspace.

Copyright © Huawei Technologies Co., Ltd.

Constraints

Anyone who wants to change the job log storage paths must meet either of the following conditions:

- The HUAWEI CLOUD account used must be assigned with the **DGC Administrator** or **Tenant Administrator** policy.
- You must be a user assigned with the **DGC User** policy and also the administrator of the current workspace.

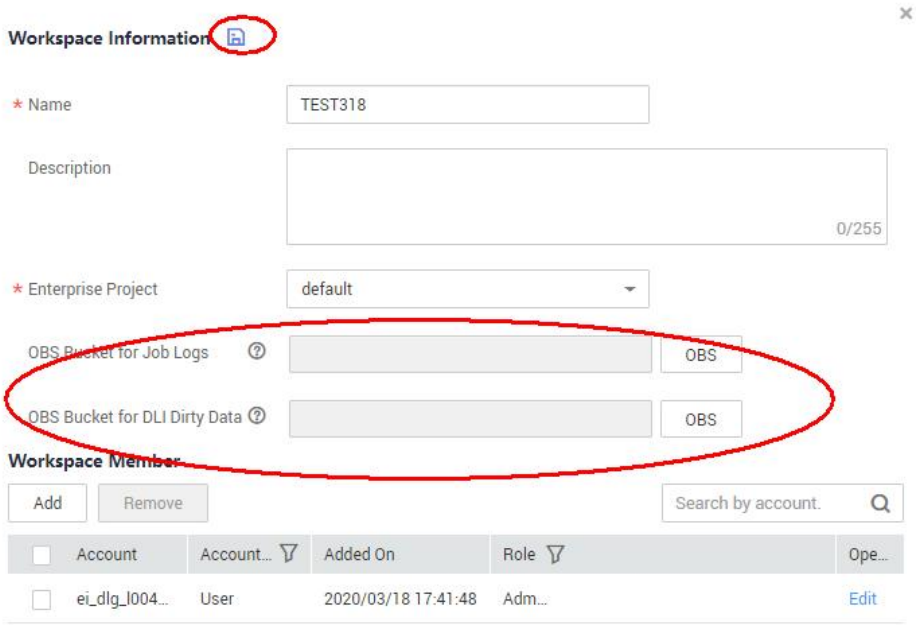
Modification Method

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Edit** corresponding to the target workspace.
3. On the **Workspace Information** page, click . Then, the workspace is in an editable state.

   Click the **OBS** button next to the OBS path for job logs to reselect a log storage path.

   Figure 3-2 Changing the log paths



4. After the modification, click to save the changes.

3.3 Managing Members

This section describes how to manage workspace members. If you plan to allow other users to use your DGC instances, grant the **DGC User** policy to these users and add them as workspace members.

 Copyright © Huawei Technologies Co., Ltd.

Workspace members include administrators, developers, O&M engineers, and visitors. The default role of a workspace creator is administrator. For details on role permissions, see **List of DGC Permissions**.

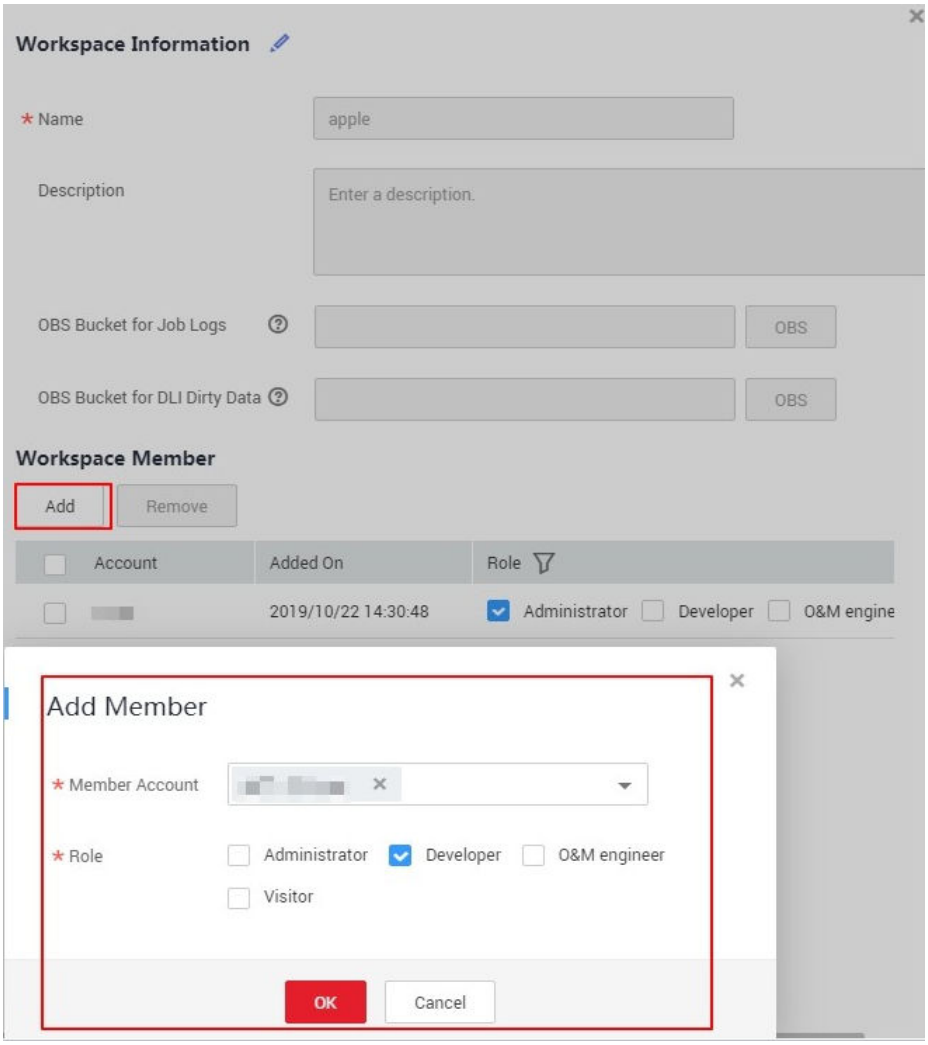
Context

Only the HUAWEI CLOUD account (that is, the master account) that creates the DGC instance or a sub-user who has the **Security Administrator** policy can add members to the workspace.

Adding a Member and Assigning a Role

1. Log in to the DGC management console. Select an instance and click **Access**. On the displayed page, choose the **Workspace Management** tab.
2. Click **Manage Member** in the **Operation** column.
3. Click **Add** under **Workspace Member**. The **Add Member** dialog box is displayed. Select a member account from the drop-down list and select a role for it. Then, click **OK**.

Figure 3-3 Adding a member



 Copyright © Huawei Technologies Co., Ltd.

4. You can view or modify the members and roles in the member list, or delete members from the workspace.

Removing a Workspace Member

1. Log in to the DGC management console.
2. Select a DGC instance and click **Access** on it. Then, click the **Workspace Management** tab.
3. On the **Workspace Management** page, locate the target workspace and click **Manage Member** in the **Operation** column.
4. On the **Workspace Information** page, select the member you want to remove from the member list and click **Remove**.

   NOTE

   The creator of a workspace cannot be deleted.

5. In the **Remove Member** dialog box, if there is no changes made to the selection, click **OK**.

 Copyright © Huawei Technologies Co., Ltd.

4 Preparing Data Sources

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.

Preparing a Data Lake

To use DGC to power data governance and operations, you need to integrate your data sources into cloud services. The cloud services are data lakes. With DGC Data Integration, you can integrate data sources to diversified data lakes. Select a cloud service as the data lake based on the site requirements.

Step 1 Select a data lake. See **Data Sources Supported By DGC**.

Step 2 Create a database.

Before using DGC Data Integration to migrate your data to the cloud, you need to create a destination database.

Step 3 (Optional) Create a table.

Before using DGC Data Integration to migrate your data to the cloud, you need to create a destination table for storing raw data. If there are a large number of table fields, you can create a table by compiling an SQL script.

NOTE

If the source database is a MySQL or PostgreSQL database, you do not need to create the destination database or table in advance because they are automatically created when DGC Data Integration is used for data migration.

----End

 Copyright © Huawei Technologies Co., Ltd.

5 Management Center

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

5.1 Data Sources

Before using DGC, select a cloud service or data warehouse as the data lake. The data lake stores raw data and data generated during data governance and serves for data development, data services, and data operations. DGC integrates various data engines and can connect to HUAWEI CLOUD data lakes and database cloud services, such as Data Lake Insight (DLI) and Data Warehouse Service (DWS), as well as traditional enterprise databases, such as MySQL and PostgreSQL.

Data Sources Supported By DGC

Supported Data Sources describes the data sources supported by DGC CDM. Note that the data connections created in DGC CDM can be used only in the CDM module.

Supported Data Sources describes the data sources supported by DGC DIS.

Table 5-1 describes the data sources supported by other DGC components. For details, see **Overview**. To connect to these data sources, create data connections using DGC Management Center.

Table 5-1 Data sources supported by other DGC components

Data Source Type	Management Center	Data Design	Data Development	Data Assets	Data Quality Control	Data Lake Mall	Data Security
Data Warehouse Service (DWS)	Supported	Supported	Supported	Supported	Supported	Supported	Supported

 Copyright © Huawei Technologies Co., Ltd.

Data Source Type	Management Center	Data Design	Data Development	Data Assets	Data Quality Control	Data Lake Mall	Data Security
Data Lake Insight (DLI)	Supported	Supported	Supported	Supported	Supported	Supported	Supported
MapReduce (MRS) HBase	Supported	Not supported	Not supported	Supported	Not supported	Not supported	Not supported
MapReduce (MRS) Hive	Supported	Supported	Supported	Supported	Supported	Not supported	Supported
MapReduce (MRS) Kafka	Supported	Not supported	Supported	Not supported	Not supported	Not supported	Not supported
MapReduce (MRS) Ranger	Supported	Not supported	Not supported	Not supported	Not supported	Not supported	Supported
MySQL	Supported	Not supported	Not supported	Not supported	Not supported	Supported	Not supported
(Optional) Oracle	Supported	Not supported	Not supported	Supported	Supported	Not supported	Not supported
SparkSQL	Supported	Not supported	Supported	Not supported	Not supported	Not supported	Not supported
Relational Database Service (RDS)	Supported	Not supported	Supported	Supported	Supported (MySQL and PostgreSQL are supported.)	Supported	Not supported
Data Ingestion Service (DIS)	Supported	Not supported	Supported	Supported	Not supported	Not supported	Not supported
Host Connection	Supported	Not supported	Supported	Not supported	Not supported	Not supported	Not supported

 Copyright © Huawei Technologies Co., Ltd.

Data Source Type	Management Center	Data Design	Data Development	Data Assets	Data Quality Control	Data Lake Mall	Data Security
PRESTO	Supported	Not supported	Supported	Not supported	Not supported	Not supported	Not supported

In addition, Data Assets supports metadata collection from the following data sources:

Overview

Table 5-2 Data source overview

Data Source Type	Description
Data Warehouse Service (DWS)	HUAWEI CLOUD 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.
Data Lake Insight (DLI)	HUAWEI CLOUD DLI is a serverless big data compute and analysis service that is fully compatible with Apache Spark and Apache Flink ecosystems and supports batch streaming. 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.
MapReduce (MRS) HBase	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. MRS HBase stores massive amount of data and supports data queries in milliseconds.

 Copyright © Huawei Technologies Co., Ltd.

Data Source Type	Description
MapReduce (MRS) Hive	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 a user familiar with SQL to query data. 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.
MapReduce (MRS) Kafka	HUAWEI CLOUD 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.
MapReduce (MRS) Ranger	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.
MySQL (pending offline)	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.

 Copyright © Huawei Technologies Co., Ltd.

Data Source Type	Description
Oracle	Oracle is a group of software that mainly applied to the distributed database. The Oracle database is one of the most popular Client/Server (C/S) and Browser/Server (B/S) databases. It is also the most widely used database management system in the world. As a general database system, the Oracle database provides complete data management functions. As a relational database, it provides complete relational models. As a distributed database, it implements distributed data processing.
SparkSQL	Spark is an open source parallel data processing framework. It helps users to easily develop unified big data applications and perform cooperative processing, stream processing, and interactive analysis on data. 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 SQLs to process structured data.
Relational Database Service (RDS)	HUAWEI CLOUD 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. Currently, DGC supports only MySQL and PostgreSQL databases in RDS.
Data Ingestion Service (DIS)	DIS streams are used to schedule jobs between workspaces. If DIS streams are used, messages can be sent to the DIS streams of another account. Otherwise, messages can be sent only to streams in all regions of the current account.
Host Connection	You can connect to a specified host during data development and execute shell 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 Connection Management** page, but do not need to edit it in scripts or jobs one by one.

 Copyright © Huawei Technologies Co., Ltd.

Data Source Type	Description
Presto	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. 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.

5.2 Creating Data Connections

You can create data connections by configuring data sources. Metadata management allows you to create, edit, and delete data connections, as well as test their connectivity. Data connections can be used for collection tasks, business metrics, and data quality jobs. If there are any changes made to the saved information, you need to update the related data connections.

Prerequisites

- Before creating a data connection, ensure that you have created a database to be connected, or have subscribed to the cloud service to be accessed.
  - To create a DWS data connection, you need to have a DWS cluster first.
  - Before creating an MRS HBase, MRS Hive, MRS Kafka, or SparkSQL data connection, ensure that you have bought an MRS cluster and selected required components when bought the cluster.
  - Before creating an RDS data connection, ensure that you have bought an RDS DB instance. Currently, DGC supports only MySQL and PostgreSQL databases in RDS.
- Before creating a data connection, ensure that the data source can communicate with the DGC instance.
  - If the data source is a cloud service (such as DWS, MRS, DIS, or RDS), ensure that the cloud service instance and the DGC instance are in the same region, AZ, VPC, and subnet, in addition, the security group rule allows the cloud service to communicate with the DGC instance. In addition, you must ensure that the cloud service instance and the DGC workspace belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace by referring to **Editing a Workspace**.
  - If the data source is an on-premises MySQL database, ensure that the host where the data source is located can access the public network and the connection port has been enabled in the firewall rule.
- You can create up to 20 data connections.

 Copyright © Huawei Technologies Co., Ltd.

Creating a Data Connection

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. On the displayed page, choose **More** > **Management Center** > **Data Connection Management** under **Quick Entry**.

3. Click **Create Data Connection** to open a dialog box. Select a data connection type and set the relevant parameters. See **Table 5-3**.

   Table 5-3 Data connections

Data Connection Type	Description
MRS Hive	**Table 5-4** describes the details of the MRS Hive data connections.
MRS HBase	**Table 5-5** describes the details of the MRS HBase data connections.
MRS Kafka	**Table 5-6** describes the details of MRS Kafka data connections.
MRS Ranger	**Table 5-13** describes the details of MRS Ranger data connections.
DWS	**Table 5-9** describes the details of the DWS data connections.
DLI	**Table 5-10** describes the details of the DLI data connections.
ORACLE	**Table 5-11** describes the details of the ORACLE data connections.
SparkSQL	**Table 5-7** describes the details of the SparkSQL data connections.
RDS	**Table 5-8** describes the details of the RDS data connections.

4. Click **Test** to test connectivity of the data connection to be created. If the test passes, the data connection is created.
5. Click **OK**.

 Copyright © Huawei Technologies Co., Ltd.

Data Connection Parameter Description

Table 5-4 MRS Hive data connection

Parameter	Mandatory	Description
Data Connection Name	Yes	The name of the data connection to be created. Connection names can contain 1 to 50 characters. They can include only letters, digits, underscores (_), and hyphens (-).
Tag	No	The attribute of the data connection to be created. 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, digits, and underscores (_) are allowed. Tag names cannot start with underscores (_). Enter up to 100 characters.
Cluster Name	Yes	The name of the MRS Hive cluster. Select an MRS cluster that Hive belongs to. If the MRS cluster is not displayed in the drop-down list, check whether the MRS cluster and the DGC instance are in the same region, AZ, and VPC. If they are not, create an MRS cluster again. In addition, you must ensure that the MRS cluster and the DGC workspace belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace by referring to **Editing a Workspace**. **NOTE** • If you select the direct connection mode, secure and non-secure clusters of MRS 2.*X* are supported. • If you select the proxy connection mode, secure and non-secure clusters of all MRS versions are supported.
Connection Method	Yes	Methods that SparkSQL uses to connect to DGC: • Proxy connection • MRS API connection When you select the MRS API connection, pay attention to the following restrictions: 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. The data governance function is not supported.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Username	Yes	The username of the MRS cluster. This parameter is mandatory when **Connection Method** is set to **agent**. 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.
Password	Yes	The password for accessing the MRS cluster. This parameter is mandatory when **Connection Method** is set to **agent**.
KMS Key	Yes	The name of the KMS key. This parameter is mandatory when **Connection Method** is set to **agent**.
Agent	Yes	MRS is not a fully managed service. It cannot be directly connected to DGC, but a CDM cluster can provide an agent for DGC to communicate with non-fully managed services. Therefore, you need to select a CDM cluster when creating an MRS Hive data connection. If there is no CDM cluster, create one. If you need to connect to an MRS 2.*x* cluster, select the CDM cluster of the 2.*x* version as the agent. The CDM cluster of version 1.8.*x* cannot connect to the MRS 2.*x* cluster.

Table 5-5 MRS HBase data connection

Parameter	Mandatory	Description
Data Connection Name	Yes	The name of the data connection to be created. Connection names can contain 1 to 50 characters. They can include only letters, digits, underscores (_), and hyphens (-).
Tag	No	The attribute of the data connection to be created. 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, digits, and underscores (_) are allowed. Tag names cannot start with underscores (_). Enter up to 100 characters.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Cluster Name	Yes	The name of the MRS HBase cluster. Select an MRS cluster that HBase belongs to. **NOTE** • If you select the direct connection mode, secure and non-secure clusters of MRS 2.*X* are supported. • If you select the proxy connection mode, secure and non-secure clusters of all MRS versions are supported.
Connection Method	Yes	Methods that SparkSQL uses to connect to DGC: • Proxy connection • Direct connection
Username	Yes	The username of the MRS cluster. This parameter is mandatory when **Connection Method** is set to **agent**. 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.
Password	Yes	The password for accessing the MRS cluster. This parameter is mandatory when **Connection Method** is set to **agent**.
KMS Key	Yes	The name of the KMS key. This parameter is mandatory when **Connection Method** is set to **agent**.
Agent	Yes	MRS is not a fully managed service. It cannot be directly connected to DGC, but a CDM cluster can provide an agent for DGC to communicate with non-fully managed services. Therefore, you need to select a CDM cluster when creating an MRS HBase data connection. If there is no CDM cluster, create one. If you need to connect to an MRS 2.*x* cluster, select the CDM cluster of the 2.*x* version as the agent.

 Copyright © Huawei Technologies Co., Ltd.

Table 5-6 MRS Kafka data connection

Parameter	Mandatory	Description
Data Connection Name	Yes	The name of the data connection to be created. Connection names can contain 1 to 50 characters. They can include only letters, digits, underscores (_), and hyphens (-).
Tag	No	The attribute of the data connection to be created. 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, digits, and underscores (_) are allowed. Tag names cannot start with underscores (_). Enter up to 100 characters.
Cluster Name	Yes	The name of the MRS Kafka cluster. Select an MRS cluster that Kafka belongs to.
Username	Yes	The username of the MRS cluster. This parameter is mandatory when **Connection Method** is set to **agent**. 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.
Password	Yes	The password for accessing the MRS cluster. This parameter is mandatory when **Connection Method** is set to **agent**.
Agent	Yes	MRS is not a fully managed service. It cannot be directly connected to DGC, but a CDM cluster can provide an agent for DGC to communicate with non-fully managed services. Therefore, you need to select a CDM cluster when creating an MRS Kafka data connection. If there is no CDM cluster, create one. If you need to connect to an MRS 2.*x* cluster, select the CDM cluster of the 2.*x* version as the agent.

 Copyright © Huawei Technologies Co., Ltd.

Table 5-7 SparkSQL data connection

Parameter	Mandatory	Description
Data Connection Name	Yes	The name of the data connection to be created. Connection names can contain 1 to 50 characters. They can include only letters, digits, underscores (_), and hyphens (-).
Tag	No	The attribute of the data connection to be created. 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, digits, and underscores (_) are allowed. Tag names cannot start with underscores (_). Enter up to 100 characters.
Cluster Name	Yes	The name of the SparkSQL cluster. Select an MRS cluster that SparkSQL belongs to. **NOTE** • If you select the direct connection mode, secure and non-secure clusters of MRS 2.*X* are supported. • If you select the proxy connection mode, secure and non-secure clusters of all MRS versions are supported.
Connection Method	Yes	Methods that SparkSQL uses to connect to DGC: • Proxy connection • MRS API connection When you select the MRS API connection, pay attention to the following restrictions: 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. The data governance function is not supported.
Username	No	The username of the MRS cluster. This parameter is mandatory when **Connection Method** is set to **agent**. 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.
Password	No	The password for accessing the MRS cluster. This parameter is mandatory when **Connection Method** is set to **agent**.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
KMS Key	No	The name of the KMS key. This parameter is mandatory when **Connection Method** is set to **agent**.
Agent	No	This parameter is mandatory when **Connection Method** is set to **agent**. MRS is not a fully managed service. It cannot be directly connected to DGC, but a CDM cluster can provide an agent for DGC to communicate with non-fully managed services. Therefore, you need to select a CDM cluster when creating a SparkSQL data connection. If there is no CDM cluster, create one. 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 allow the CDM cluster communicate with the MRS cluster. If you need to connect to an MRS 2.*x* cluster, select the CDM cluster of the 2.*x* version as the agent.

Table 5-8 RDS data connection

Parameter	Mandatory	Description
Data Connection Name	Yes	The name of the data connection to be created. Connection names can contain 1 to 50 characters. They can include only letters, digits, underscores (_), and hyphens (-).
Tag	No	The attribute of the data connection to be created. 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, digits, and underscores (_) are allowed. Tag names cannot start with underscores (_). Enter up to 100 characters.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
IP	Yes	The address for accessing RDS. To obtain the RDS access address: 1. Log in to the management console using the registered account. 2. Click **Relational Database Service** and select **Instance Management** from the list on the left. 3. Click the name of an instance. The basic information page of the instance is displayed. You can obtain the IP address on the **Connection Information** tab.
Port	Yes	The port for accessing RDS. To obtain the RDS access port: 1. Log in to the management console using the registered account. 2. Click **Relational Database Service** and select **Instance Management** from the list on the left. 3. Click the name of an instance. The basic information page of the instance is displayed. You can obtain the database port number on the **Connection Information** tab.
Driver Name	Yes	The name of the driver. • com.mysql.jdbc.Driver • org.postgresql.Driver
Driver File Path	Yes	The driver file is stored on OBS. You need to download the driver file from the official website and upload it to OBS. **NOTE** The RDS data connection driver is cached in the CDM cluster. Therefore, if the driver is updated, you need to restart the CDM cluster in Data Integration and then edit the data connection to upload the driver again.
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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
KMS Key	Yes	The name of the KMS key. To obtain the key: 1. Log in to the management console using the registered 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. It cannot be directly connected to DGC, but a CDM cluster can provide an agent for DGC to communicate with non-fully managed services. Therefore, you need to select a CDM cluster when creating an RDS data connection. If there is no CDM cluster, create one.

Table 5-9 DWS data connection

Parameter	Mandatory	Description
Data Connection Name	Yes	The name of the data connection to be created. Connection names can contain 1 to 50 characters. They can include only letters, digits, underscores (_), and hyphens (-).
Tag	No	The attribute of the data connection to be created. 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, digits, and underscores (_) are allowed. Tag names cannot start with underscores (_). Enter up to 100 characters.
Manual	Yes	You can turn off or turn on to switch off or on the **Manual** function. • When **Manual** is disabled, you do not need to enter the IP address and port number. • When **Manual** is enabled, you need to enter the IP address and port number.
IP	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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Port Number	No	The database port number specified during data warehouse 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 DGC instance can connect to the database in the DWS cluster through this port.
Cluster Name	Yes	The name of the selected DWS cluster.
Username	Yes	The database username, which is specified when the DWS cluster is created.
Password	Yes	The password for accessing the database, which is specified when the data warehouse cluster is created.
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.
Connection Method	Yes	Methods that DWS uses to connect to DGC: • **Proxy connection**: You need to bind an agent provided by the CDM cluster. • **Direct connection**
Agent	Yes	A CDM cluster can provide an agent for DGC to communicate with non-fully managed services. Therefore, you need to select a CDM cluster when creating a DWS data connection. If there is no CDM cluster, create one.

Table 5-10 DLI data connection

Parameter	Mandatory	Description
Data Connection Name	Yes	The name of the data connection to be created. Connection names can contain 1 to 50 characters. They can include only letters, digits, underscores (_), and hyphens (-).
Tag	No	The attribute of the data connection to be created. 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, digits, and underscores (_) are allowed. Tag names cannot start with underscores (_). Enter up to 100 characters.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Project Selection	Yes	The name of a project selected to connect to the DLI database in the project.

Table 5-11 Oracle data connection

Parameter	Mandatory	Description
Data Connection Name	Yes	The name of the data connection to be created. Connection names can contain 1 to 50 characters. They can include only letters, digits, underscores (_), and hyphens (-).
Tag	No	The attribute of the data connection to be created. 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, digits, and underscores (_) are allowed. Tag names cannot start with underscores (_). Enter up to 100 characters.
IP	Yes	The IP address of the database to be connected. Both public and private IP addresses are supported.
Port Number	Yes	The number of ports of the database to be connected
Username	Yes	The username of the account for accessing the database. This account must have permissions required to read and write data tables and metadata.
Password	Yes	User password.
sid	Yes	The unique identifier 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	MySQL is not a fully managed service. It cannot be directly connected to DGC, but a CDM cluster can provide an agent for DGC to communicate with non-fully managed services. Therefore, you need to select a CDM cluster when creating a MySQL data connection. If there is no CDM cluster, create one.

 Copyright © Huawei Technologies Co., Ltd.

Table 5-12 DIS connection

Parameter	Mandatory	Description
Data Connection Name	Yes	The name of the data connection to be created. Connection names can contain 1 to 50 characters. They can include only letters, digits, underscores (_), and hyphens (-).
Tag	No	The attribute of the data connection to be created. 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, digits, and underscores (_) are allowed. Tag names cannot start with underscores (_). Enter up to 100 characters.
Target Project ID	Yes	The ID of the project that the target DIS stream belongs to. The **DIS Client** node is used to send messages to the target DIS stream.
Target Region	Yes	The region that the target DIS stream belongs to. The **DIS Client** node is used to send messages to the target DIS stream.
Access Key	Yes	The AK of the tenant who creates the target DIS stream that receives messages from the **DIS Client** node.
Secret Key	Yes	The SK of the tenant who creates the target DIS stream that receives messages from the **DIS Client** node.
KMS Key	Yes	The name of the KMS key. To obtain the key: 1. Log in to the management console using the registered 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.

Table 5-13 MRS Ranger data connection

Parameter	Mandatory	Description
Data Connection Name	Yes	The name of the data connection to be created. Connection names can contain 1 to 50 characters. They can include only letters, digits, underscores (_), and hyphens (-).

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Tag	No	The attribute of the data connection to be created. 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, digits, and underscores (_) are allowed. Tag names cannot start with underscores (_). Enter up to 100 characters.
Cluster Name	Yes	The name of the MRS Ranger cluster. Select an MRS cluster that Ranger belongs to.
Username	Yes	Username of the MRS cluster. This parameter is mandatory. You cannot create a data connection for an MRS security cluster as 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 by referring to **Creating a Kerberos Authentication User for an MRS Security Cluster** and change the user type to **Admin** on the Ranger management page. For details, see **MapReduce User Guide > Ranger > Viewing Ranger Permission Information > Changing the Ranger User Type**. When creating an MRS data connection, set **Username** and **Password** to the username and password of the new MRS user.
Password	Yes	Password for accessing the MRS cluster. This parameter is mandatory.
KMS Key	Yes	KMS key name. This parameter is mandatory.
Agent	Yes	MRS is not a fully managed service. It cannot be directly connected to DGC, but a CDM cluster can provide an agent for DGC to communicate with non-fully managed services. Therefore, you need to select a CDM cluster when creating an MRS Hive data connection. If there is no CDM cluster, create one.

Creating a Host Connection

Managing a Host Connection describes the host connection details.

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:

 Copyright © Huawei Technologies Co., Ltd.

1. Log in to **MRS Manager** as user **admin**.
2. Choose **System** > **Manage User**. On the displayed page, add a dedicated user as the Kerberos authentication user, add a user group, and assign role permissions to the user. Select **superGroup** for **User Group** and select all roles, create a user as prompted.
3. Log in to **MRS Manager** as a new user and change the initial password.

Editing a Data Connection

Step 1 Log in to the DGC Management Center console 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 Then, click **Test** to test whether the data connection is valid. If the connection is normal, click **OK**.

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. Otherwise, it can be deleted.

Step 1 Log in to the DGC Management Center console and click **Data Connection Management**.

Step 2 In the data connection list, locate the data connection you want to edit and click **Delete** in the **Operation** column.

Step 3 In the displayed dialog box, confirm the data connection information. If you want to continue click **OK**.

----End

5.3 Migrating Resources

To migrate resources created in one environment to another, you can use DGC resource migration function. Resources that can be migrated include data services, metadata categories, metadata tags, metadata collection tasks, and data connections.

Prerequisites

- There must be resources that can be migrated in the system. For details on how to create data connections, see **Tags**. For details on how to classify

 Copyright © Huawei Technologies Co., Ltd.

metadata and add tags, see **Tags**. For details on how to create collection tasks, see **Task Management**.

- If OBS is supported, OBS buckets and folders have been created for storing resources. If OBS is unavailable, data can be imported and exported from the local PC.

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.

Exporting a Resource

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. In the **Lifecycle** area, choose **Management Center > Resource Migration**. The **Resource Migration** page is displayed.
3. Click **Create Export Task** to configure the save path and file name.
4. Click **Next** and select the resource to be exported.
5. Click **OK**.

Importing a Resource

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. In the **Lifecycle** area, choose **Management Center > Resource Migration**. The **Resource Migration** page is displayed.
3. Click **Create Import Task** and configure the path for saving the resources to be imported.
4. Click **Next** and select the resource to be imported.
5. Click **OK**.

 Copyright © Huawei Technologies Co., Ltd.

6 Data Integration

6.1 Cloud Data Migration

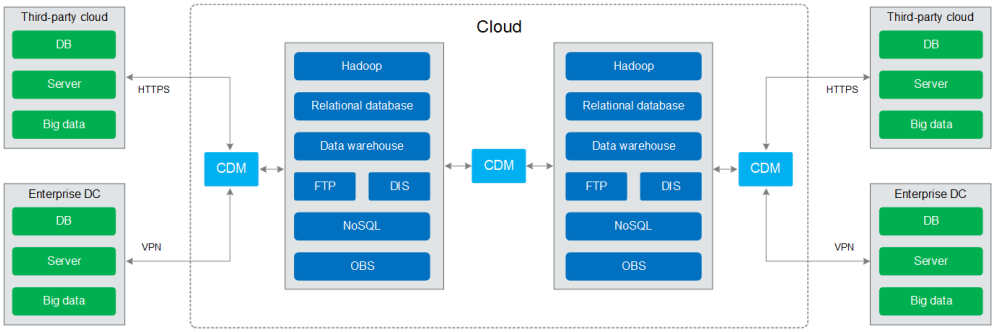
6.1.1 Overview

Cloud Data Migration (CDM) on the DGC platform enables batch data migration among homogeneous and heterogeneous data sources. You can migrate date from sources including on-premises and cloud file systems, relational databases, data warehouses, NoSQL, big data, and object storage.

Introduction to CDM

CDM is based on distributed computing and concurrent processing technologies. It helps you migrate massive sets of data stably and efficiently. You can migrate data without downtime and quickly construct the data structure you desire.

Figure 6-1 CDM application



Functions

- **Table, file, or full DB migration**

  Tables or files can be migrated in batches. An entire database can be migrated between homogeneous and heterogeneous databases. A single job can include hundreds of tables.

 Copyright © Huawei Technologies Co., Ltd.

- **Incremental data migration**

  CDM supports incremental migration of files, relational databases, and HBase/CloudTable. It also supports incremental migration implemented through WHERE clauses and macro variables of date and time.

- **Migration in transaction mode**

  If a CDM job fails, CDM rolls back the data to a state before the job started and automatically deletes data from the destination table.

- **Field conversion**

  CDM supports data conversion, including anonymization, character string operations, and date operations.

- **File encryption**

  When files are migrated to a file system, CDM can encrypt files written to cloud storage.

- **MD5 verification**

  MD5 verification is supported to check file consistency from end to end and output verification result.

- **Dirty data archiving**

  CDM can archive data that was not processed; was filtered out or converted, or not compliant with cleaning rules, to dirty data logs. A threshold for dirty data ratio can be set to determine whether a task was successful.

CDM Restrictions

1. You cannot modify the flavor of an existing cluster. If you require a flavor with different specifications, create a cluster with your desired flavor.
2. CDM does not let you control the speed of the migration. Therefore, you should not migrate data during peak hours.
3. Currently, the network bandwidth of all CDM instances is 1 Gbit/s. The theoretical maximum volume of data that can be transmitted per instance per day is 10 TB. If you have specific requirements regarding the transmission speed, use multiple CDM instances.

   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). It is recommended that you test the speed with a small amount of data before starting your migration.
4. CDM supports incremental file migration, but does not support resumable transfers of individual files.

   For example, if there are three files to be migrated and the second file fails to be migrated due to a 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. The entire file has to be moved again.
5. During file migration, a single task can include millions of files, but if there are still too many files in the directory to be migrated in a single task, you can split the files into different directories and create multiple tasks.

 Copyright © Huawei Technologies Co., Ltd.

6. A single CDM instance can execute 30 tasks at a time (cdm.large), 20 (cdm.medium), or 10 (cdm.small), or, respectively, up to 10,000, 5,000, and 2,000 jobs in the "pending" state.

   In database migration, a job is equivalent to migrating a table. In file migration, multiple files can be migrated in a single job.

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 a job configuration to CDM, you need to manually add the password to the exported JSON file.
8. Clusters cannot upgrade to a new version automatically. You need to use the job export and import functions to upgrade them.
9. CDM does not automatically back up user job configurations. You need to export and back up configuration data using the export function.
10. If a 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. The migration job does not support views.
12. During CDM migration, if the destination is a DWS or NewSQL database, constraints such as the primary key and unique index of the source database cannot be migrated together.

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.

6.1.2 Supported Data Sources

Cloud Data Migration supports the following migration modes for different data sources:

- Table/file migration: It is applicable to data migration to the cloud, data exchange on the cloud, and data migration to on-premises service systems. For details, see **Supported Data Sources in Table/File Migration**.
- Entire DB migration: It is applicable to database migration to the cloud. For details, see **Supported Data Sources in Entire DB Migration**.

 Copyright © Huawei Technologies Co., Ltd.

Supported Data Sources in Table/File Migration

Table 6-1 describes the supported data sources.

Table 6-1 Supported data sources during table/file migration

Category	Data Source	Export	Import
Data warehouse	Data Warehouse Service (DWS)	Supported	Supported
	Data Lake Insight (DLI)	Not supported	Supported
	FusionInsight LibrA	Supported	Supported
Hadoop	MRS HDFS	Supported	Supported
	MRS HBase	Supported	Supported
	MRS Hive	Supported	Supported
	FusionInsight HDFS	Supported	Supported
	Apache HDFS	Supported	Supported
	Hadoop release version	Not supported	Not supported
	FusionInsight HBase	Supported	Supported
	FusionInsight Hive	Supported	Supported
	Apache HBase	Supported	Supported
	Apache Hive	Supported	Supported
Object storage	Object Storage Service (OBS)	Supported	Supported
	Alibaba Cloud Object Storage Service (OSS)	Supported	Not supported
	Qiniu Cloud Object Storage (KODO)	Supported	Not supported
	Amazon Simple Storage Service (Amazon S3)	Supported	Not supported
	Tencent Cloud Object Storage (COS)	Supported	Not supported
File system	FTP	Supported	Supported
	SFTP	Supported	Supported
	HTTP	Supported	Not supported
	Network Attached Storage (NAS)	Supported	Supported

Copyright © Huawei Technologies Co., Ltd.

Category	Data Source	Export	Import
	Scalable File Service (SFS Turbo)	Supported	Supported
Relational database	RDS for MySQL	Supported	Supported
	Dameng database	Supported (only DWS and Hive)	Not supported
	RDS for PostgreSQL	Supported	Supported
	RDS for SQL Server	Supported	Supported
	MySQL	Supported	Supported
	PostgreSQL	Supported	Supported
	Microsoft SQL Server	Supported	Supported
	Oracle	Supported	Supported
	IBM Db2	Supported	Supported
	Derecho (GaussDB)	Supported	Supported
	NewSQL (GaussDB)	Supported	Supported
	SAP HANA	Supported	Supported
	MYCAT	Supported	Supported
NoSQL	Distributed Cache Service (DCS)	Not supported	Supported
	Document Database Service (DDS)	Supported	Supported
	Redis	Supported	Supported
	MongoDB	Supported	Supported
	Cassandra	Supported	Supported
	CloudTable Service (CloudTable)	Supported	Supported
	CloudTable OpenTSDB	Supported	Supported
Search	Cloud Search Service (CSS)	Supported	Supported
	Elasticsearch	Supported	Supported
Message system	Data Ingestion Service (DIS)	Supported	Supported
	Apache Kafka	Supported	Supported
	MRS Kafka	Supported	Supported
	DMS Kafka	Supported	Supported

 Copyright © Huawei Technologies Co., Ltd.

NOTE

- In the preceding table, the non-cloud data sources, such as MySQL, can be the on-premises MySQL, MySQL built on ECSs, or MySQL on the third-party cloud.
- When DIS, Apache Kafka, or DMS Kafka functions as the migration source, CDM can migrate data to only the following types of data sources:
  - CSS
  - DIS
  - Apache Kafka
  - DMS Kafka
- Apache Hive supports only 1.2 and later versions and 3.x.

Supported Data Sources in Entire DB Migration

Entire DB migration is applicable to the scenario where an on-premises data center or a database created on an ECS is 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 6-2 lists the data sources supporting entire DB migration using CDM.

Table 6-2 Supported data sources

Source Data Type	Destination Data Type					
	RDS for MySQL	MRS (Hive)	DWS	CSS	OBS	CloudTable
MySQL	√	√	√	×	√	×
PostgreSQL	√	√	√	×	√	×
Microsoft SQL Server	√	√	√	×	×	×
Oracle	√	√	√	×	√	×
Elasticsearch	×	×	×	√	×	×
MongoDB	×	×	×	×	×	×
HBase	×	×	×	×	×	√
IBM Db2	√	√	√	×	√	×
Derecho (GaussDB)	√	√	√	×	√	×
SAP HANA	√	√	√	×	√	×
DWS	√	√	√	×	×	×
Hive	√	×	√	×	×	×

Copyright © Huawei Technologies Co., Ltd.

Data Types Supported by the Dameng Database

When the source end is a Dameng database and the destination end is a Hive or DWS database, the following data types are supported:

Table 6-3 Supported data types

Category	Type	Description	Storage Format Example	Hive	DWS
Character	CHAR	Stores specified fixed-length character strings.	'a' or 'aaaaa'	CHAR	CHAR
	CHARACTER	Same as CHAR	'a' or 'aaaaa'	CHAR	CHAR
	VARCHAR	Stores specified variable-length strings.	'a' or 'aaaaa'	VARCHAR	VARCHAR
	VARCHAR2	Same as VARCHAR	'a' or 'aaaaa'	VARCHAR	VARCHAR
Value	NUMERIC	Stores zero, positive, and negative fixed point numbers.	52.36	DECIMAL	NUMERIC
	DECIMAL	Similar to NUMERIC	52.36	DECIMAL	NUMERIC
	DEC	Same as DECIMAL	52.36	DECIMAL	NUMERIC
	NUMBER	Same as NUMERIC	52.36	DECIMAL	NUMERIC
	INTEGER	Stores signed integers. Integer part: 10 digits; decimal part: 0 digits	5236	INT	INTEGER
	INT	Same as INTEGER	5236	INT	INTEGER
	BIGINT	Stores signed integers. Integer part: 19 digits; decimal part: 0 digits	5236	BIGINT	BIGINT

 Copyright © Huawei Technologies Co., Ltd.

Category	Type	Description	Storage Format Example	Hive	DWS
	TINYINT	Stores signed integers. Integer part: 3 digits; decimal part: 0 digits	100	TINYINT	SMALLINT
	SMALLINT	Stores signed integers. Integer part: 5 digits; decimal part: 0 digits	9999	SMALLINT	SMALLINT
	BYTE	Similar to TINYINT. Integer part: 3 digits; decimal part: 0 digits	100	TINYINT	SMALLINT
	BINARY	Stores fixed-length binary data.	0x2A3B4058	BINARY (NULL)	BYTEA (NULL)
	VARBINARY	Stores variable-length binary data.	0x2A3B4058	BINARY (NULL)	BYTEA (NULL)
	FLOAT	Stores floating-point numbers with binary precision.	52.36	FLOAT	FLOAT8
	DOUBLE	Similar to FLOAT	52.36	DOUBLE	FLOAT8
	REAL	Stores binary floating-point numbers.	52.3	FLOAT	FLOAT4
	DOUBLE PRECISION	Stores double-precision floating-point numbers.	52.3	DOUBLE	FLOAT8
Bit string	BIT	Stores 1, 0, or NULL.	1, 0, or NULL	TINYINT(1 0 NULL)	BOOLEAN(true false NULL)
Time and date	DATE	Stores information about the year, month, and day.	'1999-10-01', '1999/10/01', or '1999.10.01'	DATE	TIMESTAMP

Copyright © Huawei Technologies Co., Ltd.

Category	Type	Description	Storage Format Example	Hive	DWS
	TIME	Stores information about the hour, minute, and second.	'09:10:21' or '9:10:21'	Not supported (string)	TIME
	TIMESTAMP	Stores information about the year, month, day, hour, minute, and second.	2002-12-12 09:10:21','2002-12-12 9:10:21' '2002/12/12 09:10:21' or '2002.12.12 09:10:21'	TIMESTAMP	TIMESTAMP
	TIME WITH TIME ZONE	Stores a TIME value with a time zone. Add the time zone information to the end of the TIME type.	'09:10:21 +8:00', '09:10:21+8:00' or '9:10:21+8:00'	Not supported (string)	TIME WITH TIME ZONE
	TIMESTAMP WITH TIME ZONE	Stores a TIMESTAMP value with a time zone. Add the time zone information to the end of the TIMESTAMP type.	2002-12-12 09:10:21 +8:00','2002-12-12 9:10:21 +8:00' '2002/12/12 09:10:21 +8:00' or '2002.12.12 09:10:21 +8:00'	TIMESTAMP	TIMESTAMP WITH TIME ZONE
	TIMESTAMP WITH LOCAL TIME ZONE	Stores the TIMESTAMP value of a local time zone. The standard time zone type (TIMESTAMP WITH TIME ZONE) can be converted to the local time zone type.	2002-12-12 09:10:21 +8:00','2002-12-12 9:10:21 +8:00' '2002/12/12 09:10:21 +8:00' or '2002.12.12 09:10:21 +8:00'	Not supported (string)	Not supported (TEXT)

 Copyright © Huawei Technologies Co., Ltd.

Category	Type	Description	Storage Format Example	Hive	DWS
	DATETIME WITH TIME ZONE	Same as TIMESTAMP WITH TIME ZONE	2002-12-12 09:10:21 +8:00','2002-12-12 9:10:21 +8:00' '2002/12/12 09:10:21 +8:00' or '2002.12.12 09:10:21 +8:00'	TIMESTAMP	TIMESTAMP WITH TIME ZONE
	INTERVAL YEAR	Interval of years. The leading precision specifies the range of years.	INTERVAL '0015' YEAR	Not supported (string)	Not supported (VARCHAR)
	INTERVAL YEAR TO MONTH	Interval in the form of month and year. The leading precision specifies the range of years.	INTERVAL '0015-08' YEAR TO MONTH	Not supported (string)	Not supported (VARCHAR)
	INTERVAL MONTH	Interval of months. The leading precision specifies the range of months.	INTERVAL '0015' MONTH	Not supported (string)	Not supported (VARCHAR)
	INTERVAL DAY	Interval of days. The leading precision specifies the range of days.	INTERVAL '150' DAY	Not supported (string)	Not supported (VARCHAR)
	INTERVAL DAY TO HOUR	Interval inn the form of hour and day. The leading precision specifies the range of days.	INTERVAL '9 23' DAY TO HOUR	Not supported (string)	Not supported (VARCHAR)
	INTERVAL DAY TO MINUTE	Interval in the form of minute, hour, and day. The leading precision specifies the range of days.	INTERVAL '09 23:12' DAY TO MINUTE	Not supported (string)	Not supported (VARCHAR)

Copyright © Huawei Technologies Co., Ltd.

Category	Type	Description	Storage Format Example	Hive	DWS
	INTERVAL DAY TO SECOND	Interval in the form of second, minute, hour, and day. The leading precision specifies the range of days.	INTERVAL '09 23:12:01.1' DAY TO SECOND	Not supported (string)	Not supported (VARCHAR)
	INTERVAL HOUR	Interval of hours. The leading precision specifies the range of hours.	INTERVAL '150' HOUR	Not supported (string)	Not supported (VARCHAR)
	INTERVAL HOUR TO MINUTE	Interval in the form of minute and hour. The leading precision specifies the range of hours.	INTERVAL '23:12' HOUR TO MINUTE	Not supported (string)	Not supported (VARCHAR)
	INTERVAL HOUR TO SECOND	Interval in the form of second, minute, and hour. The leading precision specifies the range of hours.	INTERVAL '23:12:01.1' HOUR TO SECOND	Not supported (string)	Not supported (VARCHAR)
	INTERVAL MINUTE	Interval of minutes. The leading precision specifies the range of minutes.	INTERVAL '150' MINUTE	Not supported (string)	Not supported (VARCHAR)
	INTERVAL MINUTE TO SECOND	Interval in the form of second and minute. The leading precision specifies the range of minutes .	INTERVAL '12:01.1' MINUTE TO SECOND	Not supported (string)	Not supported (VARCHAR)
	INTERVAL SECOND	Interval of seconds. The leading precision specifies the range of the integer part of the second .	INTERVAL '51.1' SECOND	Not supported (string)	Not supported (VARCHAR)

Copyright © Huawei Technologies Co., Ltd.

Category	Type	Description	Storage Format Example	Hive	DWS
Multimedia	IMAGE	IMAGE specifies the image type in the multimedia information. An image consists of a pixel lattice with a maximum length of 2 GB. In addition to storing image data, other binary data can also be stored.	0x2A3B4058 (binary data)	Not supported	Not supported
	LONGVARBINARY	Same as IMAGE	0x2A3B4059 (binary data)	Not supported	Not supported
	TEXT	Stores the long string type. The maximum length of a string is 2 GB.	0x5236 (binary data)	Not supported	Not supported
	LONGVARCHAR	Similar to TEXT	0x5236 (binary data)	Not supported	Not supported
	BLOB	Stores variable-length binary large objects with a maximum length of 2 GB.	0x5236 (binary data)	Not supported	Not supported
	CLOB	Stores variable-length binary large objects with a maximum length of 2 GB.	0x5236 (binary data)	Not supported	Not supported
	BFILE	Specified the binary files stored in the operating systems. Files are stored in the operating systems instead of the databases. They can be read only.	-	Not supported	Not supported

 Copyright © Huawei Technologies Co., Ltd.

Data Types Supported by the MySQL Database

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

Table 6-4 Supported data types

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
Value	DECIMAL(M,D)	Uncompressed floating-point numbers cannot be unsigned. In unpacking decimals, each decimal corresponds to a byte. Defining the number of display lengths (M) and decimals (D) is required. NUMERIC is the synonym of DECIMAL.	52.36	DECIMAL	When D is 0, it corresponds to BIGINT. When D is not 0, it corresponds to NUMBERIC.
	NUMBERIC	Same as DECIMAL	-	DECIMAL	NUMBERIC

 Copyright © Huawei Technologies Co., Ltd.

	INTEGER	An integer of normal size, which can contain symbols. 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
	BIGINT	A large integer that can contain symbols. 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

 Copyright © Huawei Technologies Co., Ltd.

	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
	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 contain symbols. If the value is signed, it ranges from -32768 to 32767. If the value is unsigned, it ranges from 0 to 65535, and you can specify a maximum of 5-bit width.	9999	SMALLINT	SMALLINT

 Copyright © Huawei Technologies Co., Ltd.

	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
	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

Copyright © Huawei Technologies Co., Ltd.

Bit	BIT(M)	Stored bit type value. BIT(M) can store up to *M* bits of values, and *M* ranges from 1 to 64.	B'1111100' B'1100'	TINYINT	BYTEA
Time and date	DATE	The value is in the *YYYY-MM-DD* format and ranges from **1000-01-01** to **9999-12-31**. For example, **December 30, 1973** will be stored as **1973-12-30**.	1999-10-01	DATE	TIMESTAMP
	TIME	Stores information about the hour, minute, and second.	'09:10:21' or '9:10:21'	Not supported (string)	TIME
	DATETIME	The date and time are in the *YYYY-MM-DD HH:MM:SS* format and range from **1000-01-01 00:00:00** to **9999-12-31 23:59:59**. For example, **3:30 p.m. on December 30, 1973** will be stored as **1973-12-30 15:30:00**.	'1973-12-30 15:30:00'	TIMESTAMP	TIMESTAMP
	TIMESTAMP	Timestamp type. Timestamp between midnight on January 1, 1970 and a time point in 2037. Similar to the DATETIME format (YYYYMMDDHHMMSS), except that no hyphen is required. For example, **3:30 p.m. December 30, 1973** will be stored as **19731230153000**.	19731230153000	TIMESTAMP	TIMESTAMP

 Copyright © Huawei Technologies Co., Ltd.

	YEAR(M)	The year is stored in 2-digit or 4-digit number format. If the length is specified as 2 (for example, YEAR(2)), the year ranges from 1970 to 2069 (70 to 69). If the length is specified as 4, the year ranges from 1901 to 2155. The default length is 4.	2000	Not supported (string)	Not supported
Multimedia (binary)	BINARY(M)	The number of bytes is *M*. The length of a variable-length binary string ranges from 0 to *M*. *M* is the value length plus 1.	0x2A3B4058 (binary data)	Not supported	BYTEA
	VARBINARY(M)	The number of bytes is *M*. A fixed binary string with a length of 0 to *M*.	0x2A3B4059 (binary data)	Not supported	BYTEA
	TEXT	The maximum length of the field is 65535 characters. TEXT is a "binary large object" and is used to store large binary data, such as images or other types of files.	0x5236 (binary data)	Not supported	Not supported
	TINYTEXT	A binary string of 0 to 255 bytes in short text	-	-	Not supported
	MEDIUMTEXT	A binary string of 0 to 167772154 bytes in medium-length text	-	-	Not supported
	LONGTEXT	A binary string of 0 to 4294967295 bytes in large-length text	-	-	Not supported

 Copyright © Huawei Technologies Co., Ltd.

	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
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)

 Copyright © Huawei Technologies Co., Ltd.

	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 by the Oracle Database

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

Table 6-5 Supported data types

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

 Copyright © Huawei Technologies Co., Ltd.

Category	Type	Description	Hive	DWS
	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
	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

 Copyright © Huawei Technologies Co., Ltd.

Category	Type	Description	Hive	DWS
	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
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 by the SQL Server Database

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

Table 6-6 Supported data types

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

 Copyright © Huawei Technologies Co., Ltd.

Category	Type	Description	Hive	DWS	Oracle
	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
	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

 Copyright © Huawei Technologies Co., Ltd.

Category	Type	Description	Hive	DWS	Oracle
	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
	datetime2	Extended type of datetime, which has a larger data range. By default, the minimum precision is the highest, and the user-defined precision is optional.	TIMESTAMP	TIMESTAMP	Not supported
	smalldatetime	The smalldatetime type is similar to the datetime type. The difference is that the smalldatetime type stores data from January 1, 1900 to June 6, 2079. When the date and time precision is low, the smalldatetime type can be used. Data of this type occupies 4-byte storage space.	TIMESTAMP	TIMESTAMP	Not supported
	timestamp	Timestamp data type	TIMESTAMP	TIMESTAMP	TIMESTAMP
	datetimeoffset	A time that uses the 24-hour clock and combined with date and the time zone.	Not supported (string)	TIMESTAMP	Not supported
Multimedia data types (binary)	text	Stores text data.	Not supported (string)	Not supported (string)	Not supported

 Copyright © Huawei Technologies Co., Ltd.

Category	Type	Description	Hive	DWS	Oracle
	ntext	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
	varbinary	Variable-length binary data	Not supported (string)	Not supported (string)	Not supported
Currency data type	money	Stores currency values.	Not supported (string)	Not supported (string)	Not supported
	smallmoney	Similar to the money type, a currency symbol is prefixed to the input data. For example, the currency symbol of RMB 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

 Copyright © Huawei Technologies Co., Ltd.

Category	Type	Description	Hive	DWS	Oracle
	uniqueidentifier	A 16-byte globally unique identifier (GUID) is a unique number generated by the SQL Server based on the network adapter address and host CPU clock. Each GUID is a hexadecimal number ranging from 0 to 9 or a to f.	Not supported	Not supported	Not supported
	cursor	Cursor data type	Not supported	Not supported	Not supported
	sql_variant	Stores any valid SQL Server data except the text, image, and timestamp data, which facilitates the development of the SQL Server.	Not supported	Not supported	Not supported
	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

6.1.3 Creating Links

6.1.3.1 Creating Links

Scenario

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

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

Copyright © Huawei Technologies Co., Ltd.

Prerequisites

- The CDM cluster can communicate with the data source. For details on how to connect the internal network to the cloud network, see **How Do I Connect On-Premises Intranet or Third-Party Private Network to CDM?**
- You have obtained the URL and the account for accessing the data source. The account is granted with the read and write permissions on the data source.

Procedure

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

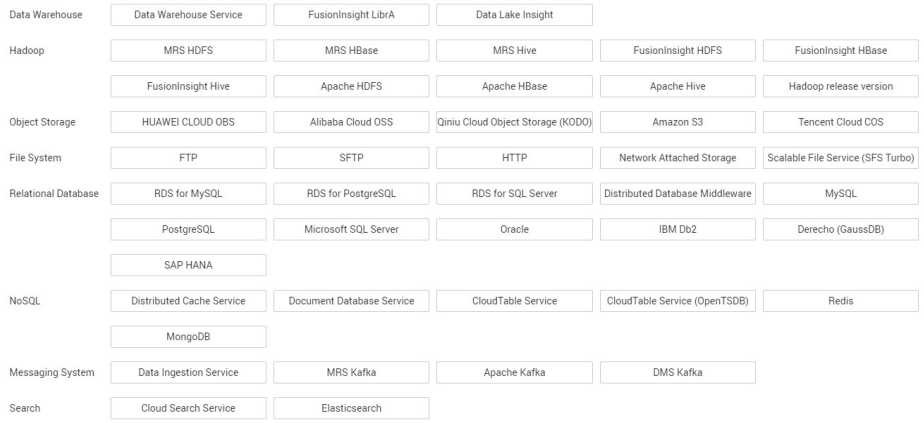
2. Click **Data Integration** of the corresponding workspace.

   The **Data Integration** page is displayed.

3. Choose **Cloud Data Migration**. On the **Cluster Management** page, choose **Job Management** > **Link Management** > **Create Link**, select one of the connector types shown in **Figure 6-2**.

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

   Figure 6-2 Selecting a connector



4. Select a data source and click **Next**.

   The link parameters of different data sources vary. **Table 6-7** describes the link parameters.

 Copyright © Huawei Technologies Co., Ltd.

Table 6-7 Link parameters

Connector	Description
• Data Warehouse Service • RDS for MySQL • RDS for PostgreSQL • RDS for SQL Server • MySQL • PostgreSQL • Microsoft SQL Server • Oracle • IBM Db2 • FusionInsight LibrA • Derecho (GaussDB)	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 Relational Databases**. • When importing data to DWS, specify the **COPY** or **GDS** import mode to improve import performance. You can specify the **Import Mode** parameter when **creating DWS links**. • When importing data to RDS for MySQL, enable the LOAD DATA function of MySQL to accelerate data import and improve the import performance. You can configure **Use Local API** to enable the function when you **create MySQL links**.
OBS	If the data source is OBS, see **Link to OBS**.
Alibaba Cloud OSS	If the data source is OSS on Alibaba Cloud, see **Link to OSS on Alibaba Cloud**. Currently, data can only be exported from OSS to OBS.
KODO COS	If the data source is KODO or COS, see **Link to KODO/COS**. Currently, data can only be exported from KODO/COS to OBS.
Amazon S3	If the data source is Amazon S3, see **Link to Amazon S3**. Currently, objects can only be exported from Amazon S3 to OBS.
• MRS HDFS • FusionInsight HDFS • Apache HDFS	If the data source is HDFS of MRS, Apache Hadoop, or FusionInsight HD, see **Link to HDFS**. **NOTE** If **Running Mode** is set to **Standalone**, CDM can migrate data between HDFSs of multiple MRS clusters.
• MRS HBase • FusionInsight HBase • Apache HBase	If the data source is HBase of MRS, Apache Hadoop, or FusionInsight HD, see **Link to HBase**.
MRS Hive	If the data source is Hive of MRS, see **Link to Hive**.

Copyright © Huawei Technologies Co., Ltd.

Connector	Description
CloudTable Service	If the data source is CloudTable, see **Link to CloudTable**.
• FTP • SFTP	If the data source is an FTP or SFTP server, see **Link to an FTP or SFTP Server**.
• HTTP • HTTPS	These connectors are used to read files with an HTTP/HTTPS URL, such as reading public files on the third-party object storage system and web disks. When creating an HTTP link, you only need to configure the link name. The URL is configured during job creation.
• NAS • SFS Turbo	If the data source is a NAS server, see **Link to NAS/SFS**. CIFS, SMB, and NFS are supported. CDM can connect to dedicated file servers, Windows file sharing servers, Linux Samba servers, and cloud services that support CIFS, SMB, or NFS file systems such as SFS.
• MongoDB • Document Database Service	If the data source is a local MongoDB or DDS, see **Link to MongoDB**.
• Redis • Distributed Cache Service	If the data source is a local Redis database or DCS, see **Link to Redis/DCS**. Currently, data can be imported to but cannot be exported from DCS. Data can be imported to and exported from the open source Redis.
Apache Kafka	If the data source is the open source Kafka, see **Link to Kafka**. Currently, Kafka data can only be exported to CSS, DIS, or DMS Kafka.
Data Ingestion Service	If the data source is DIS, see **Configuring Links to DIS**. Currently, DIS data can only be exported to CSS, Apache Kafka, or DMS Kafka.
• Cloud Search Service • 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**. Currently, data can be imported to but cannot be exported from DLI.

Copyright © Huawei Technologies Co., Ltd.

Connector	Description
OpenTSDB	If the data source is OpenTSDB, see **Link to CloudTable OpenTSDB**.
DMS Kafka	If the data source is DMS Kafka, see **Link to DMS Kafka**. Currently, data can only be exported from DMS Kafka to CSS, Apache Kafka, DIS, or DMS Kafka.

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

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

6.1.3.2 Managing Drivers

The Java Database Connectivity (JDBC) API provides programmatic access to relational databases. Using the JDBC API, applications can execute SQL statements and retrieve results. Before connecting CDM to a relational database, you need to upload a driver.

Prerequisites

- A cluster has been created.
- You have downloaded one of the drivers listed in **Table 6-8**.
- (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?

Different types of relational databases need to adapt to different drivers. Download one of the drivers listed in **Table 6-8** based on the preset driver name on the **Driver Management** page.

Table 6-8 Drivers

Relational Database Type	Driver Name	How to Obtain	Recommended Version
• RDS for MySQL • MySQL • DM	MYSQL MYCAT DM	https://downloads.mysql.com/archives/c-j/	5.1.48. The driver of the 8.x version is not supported.

Copyright © Huawei Technologies Co., Ltd.

Relational Database Type	Driver Name	How to Obtain	Recommended Version
Oracle	ORACLE_6 ORACLE_7 ORACLE_8	https://www.oracle.com/database/technologies/appdev/jdbc-downloads.html	Oracle 12c or later (recommended version: ojdbc8)
• RDS for PostgreSQL • PostgreSQL	POSTGRESQL	https://jdbc.postgresql.org/download.html	42.1.4
IBM Db2	DB2	https://www.ibm.com/support/pages/db2-jdbc-driver-versions-and-downloads	-
• RDS for SQL Server • Microsoft SQL Server	SQLServer	https://docs.microsoft.com/en-us/sql/connect/jdbc/download-microsoft-jdbc-driver-for-sql-server?view=sql-server-ver15	-

NOTE

Drivers for GaussDB and SAP HANA databases have been loaded to the system. You do not need to upload the drivers again.

Procedure

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Integration** of the corresponding workspace.

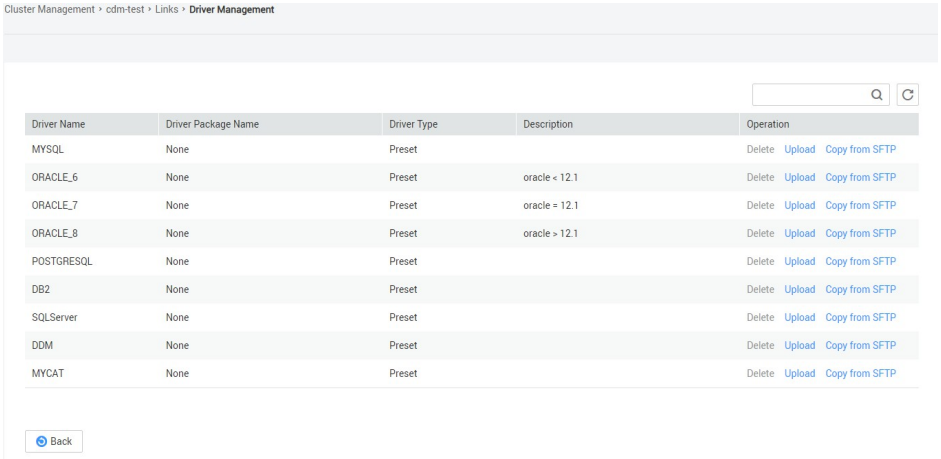
   The **Data Integration** page is displayed.

3. Choose **Cloud Data Migration**. The **Cluster Management** page is displayed.

   Choose **Job Management** > **Link Management** > **Diver Management**. The **Driver Management** page is displayed.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-3 Uploading a driver



4. Upload the 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.

6.1.3.3 Editing/Deleting Links

Scenario

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

- Edit: You can modify link parameters, but cannot re-select connectors. To modify a link, you need to re-enter the password for accessing the data source.
- Test Connectivity: You can directly test the connectivity of a created link.
- View Link JSON: Check the link parameter settings in JSON format.
- Edit Link JSON: Edit the link parameter settings in JSON format.
- Delete: You can delete links that are not used by any jobs in batches.

Prerequisites

- You have obtained the username and password for accessing the desired data source.
- Links are not used by any jobs.

Procedure

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Integration** of the corresponding workspace.

   The **Data Integration** page is displayed.

3. Choose **Cloud Data Migration**. The **Cluster Management** page is displayed. Choose **Job Management** > **Link Management**.

 Copyright © Huawei Technologies Co., Ltd.

4. On the **Link Management** page, locate the links to be deleted.
   a. Edit: 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.
   b. Test Connectivity: Click **Test Connectivity** in the **Operation** column to test the connectivity of the created link.
   c. View Link JSON: In the **Operation** column, choose **More** > **View Link JSON** to view link parameters in JSON format.
   d. Edit Link JSON: In the **Operation** column, choose **More** > **Edit Link JSON** to modify link parameters in JSON format.
   e. Delete: Select multiple links, and click **Delete Link** next to **Create Link** to batch delete unused links.

6.1.3.4 Link to Hive

CDM supports the following Hive data sources:

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

MRS Hive

The MRS Hive link is used for MapReduce Service (MRS) on HUAWEI CLOUD. **Table 6-9** describes related parameters.

NOTE

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.

Table 6-9 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

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Authentication Method	Authentication method used for accessing MRS • **SIMPLE**: Select this if MRS is in non-security mode. • **KERBEROS**: Select this if MRS is in security mode.	SIMPLE
HIVE Version	Set this to the Hive version on the server.	HIVE_3_X
Username	If **Authentication Method** is set to **KERBEROS**, you must provide the username and password used for logging in to MRS Manager.	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
Run Mode	This parameter is used only when the Hive version is **HIVE_3_X**. Possible values are: • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **STANDALONE**: The link instance runs in an independent process. The **STANDALONE** running mode is required in the following scenarios: 1. A 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.) 2. Hive1.x clusters. 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 data 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.	EMBEDDED

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Use Cluster Config	You can create cluster configurations on the **Links** page to simplify the configuration of Hadoop link parameters. When creating a link, set this parameter to **Yes**.	No
Cluster Config Name	This parameter is valid only when **Use Cluster Config** is set to **Yes**. Select a cluster configuration that has been created. For details, see Managing Cluster Configurations.	hive_01

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

FusionInsight Hive

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

Table 6-10 describes related parameters.

Table 6-10 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	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
Manager Port	FusionInsight/MRS Manager port	28443
CAS Server Port	CAS protocol port of FusionInsight/MRS Manager	20009
Authentication Method	Authentication method used for accessing MRS • **SIMPLE**: Select this if MRS is in non-security mode. • **KERBEROS**: Select this if MRS is in security mode.	SIMPLE

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
HIVE Version	Hive version	HIVE_3_X
Username	If **Authentication Method** is set to **KERBEROS**, you must provide the username and password used for logging in to MRS Manager.	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
Run Mode	This parameter is used only when the Hive version is **HIVE_3_X**. Possible values are: • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **STANDALONE**: The link instance runs in an independent process. The **STANDALONE** running mode is required in the following scenarios: 1. A 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.) 2. Hive1.x clusters. 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 data 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.	EMBEDDED
Use Cluster Config	You can create cluster configurations on the **Links** page to simplify the configuration of Hadoop link parameters. When creating a link, set this parameter to **Yes**.	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

 Copyright © Huawei Technologies Co., Ltd.

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

Apache Hive

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

Table 6-11 describes related parameters.

Table 6-11 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 address	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 MRS • **SIMPLE**: Select this if MRS is in non-security mode. • **KERBEROS**: Select this if MRS is in 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

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Run Mode	This parameter is used only when the Hive version is **HIVE_3_X**. Possible values are: • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **STANDALONE**: The link instance runs in an independent process. The **STANDALONE** running mode is required in the following scenarios: 1. A 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.) 2. Hive1.x clusters. 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 data 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.	EMBEDDED
Use Cluster Config	You can create cluster configurations on the **Links** page to simplify the configuration of Hadoop link parameters. When creating a link, set this parameter to **Yes**.	No
Cluster Config Name	This parameter is valid only when **Use Cluster Config** is set to **Yes**. Select a cluster configuration that has been created. For details, see Managing Cluster Configurations.	hive_01

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

6.1.3.5 Link to CloudTable

When connecting CDM to CloudTable, configure the parameters as described in **Table 6-12**.

Copyright © Huawei Technologies Co., Ltd.

Table 6-12 Parameter description

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	cloudtable_link
ZK Link	Obtain this parameter value from the cluster management page of CloudTable.	cloudtable-cdm-zk1.cloudtable.com:2181,cloudtable-cdm-zk2.cloudtable.com:2181
IAM Authentication	If IAM authentication is enabled for the CloudTable cluster to be connected, set this parameter to **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. If you have not created access keys for your account, create them as described in **(Optional) Obtaining the Authentication Information**.	-
SK	SK for accessing the CloudTable cluster. For details about how to obtain the SK, see the description for **AK**.	-

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Run Mode	Run mode of the HBase link. This parameter is used only when the HBase version is **HBASE_2_X**. • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **STANDALONE**: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, Hadoop, or CloudTable) with both Kerberos and Simple authentication modes, select **STANDALONE** or configure different agents. Note: The STANDALONE mode is used to solve the version conflict problem. If the connector versions of the source and destination ends of the same link are different, a JAR file conflict occurs. In this case, you need to place the source or destination end in the STANDALONE process to prevent the migration failure caused by the conflict. • **Agent**: The link instance runs on an agent.	EMBEDDED

6.1.3.6 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.

When connecting CDM to an FTP or SFTP server, configure the parameters as described in **Table 6-13**.

Table 6-13 Parameter description

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

Copyright © Huawei Technologies Co., Ltd.

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	-

6.1.3.7 Link to NAS/SFS

When connecting CDM to a NAS server or SFS Turbo, configure the parameters as described in **Table 6-14**.

CIFS, SMB, and NFS are supported. CDM can connect to dedicated file servers, Windows file sharing servers, Linux Samba servers, and cloud services that support CIFS, SMB, or NFS file systems such as SFS.

Table 6-14 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	nas_link
Protocol	NAS file protocol. Currently, only SMB, CIFS, and NFS protocols are supported.	NFS
Shared Path	Shared path of the NAS server	\\server\share
Username	Username for logging in to the NAS server, which is in the *domain name*|*username* format This parameter is not displayed if **Protocol** is set to **NFS**.	domain01\user
Password	Password for logging in to the NAS server This parameter is not displayed if **Protocol** is set to **NFS**.	-

6.1.3.8 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 6-15**.

Copyright © Huawei Technologies Co., Ltd.

Table 6-15 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
MongoDB Server List	List of server addresses. Enter each address in the format of *IP address or domain name of the database server*:*port number*, and separate the entered addresses with semicolons (;).	192.168.0.1:7300;192.168.0.2:7301
Database Name	Name of the MongoDB database to be connected	DB_mongodb
Username	Username for logging in to MongoDB	cdm
Password	Password for logging in to MongoDB	-

6.1.3.9 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 6-16**.

Table 6-16 DDS link parameters

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

Copyright © Huawei Technologies Co., Ltd.

6.1.3.10 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 6-17**.

Table 6-17 Parameter description

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	redis_link
Redis Deployment Method	Two deployment methods are available: • **Single**: installation on a single-node system • **Cluster**: installation on a cluster	Single
Redis Server List	List of server addresses. Enter each address in the format of *IP address or domain name of the database server*:*port number*, and separate the entered addresses with semicolons (;).	192.168.0.1:7300;192.168.0.2:7301
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

6.1.3.11 Link to Kafka

When connecting CDM to MRS, FusionInsight, or Kafka of local Apache Hadoop, configure the parameters as described in **Table 6-18**.

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.

Copyright © Huawei Technologies Co., Ltd.

Table 6-18 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
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.	-
Authentication Method	Authentication method used for accessing MRS • **SIMPLE**: for non-security mode • **KERBEROS**: for security mode	Yes
Username	Username used for logging in to MRS Manager	-
Password	Password used for logging in to MRS Manager	-

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

6.1.3.12 Configuring Links to DIS

When connecting links to DIS, configure the parameters as described in **Table 6-19**. Currently, data can only be exported from DIS to CSS, Apache Kafka, or DMS Kafka.

Table 6-19 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	dis_link
Region	Region where DIS is deployed.	cn-north-1
Endpoint	URL of DIS in the format of https://*Endpoint*, Obtain the endpoint from **Regions and Endpoints**.	https://dis.cn-north-1.myhuaweicloud.com

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
AK	AK used to log in to the DIS server You need to create an access key for the current account and obtain an AK/SK pair. If you have not created an AK/SK pair for your account, create one. See **(Optional) Obtaining the Authentication Information**.	N/A
SK	SK used to log in to the DIS server You need to create an access key for the current account and obtain an AK/SK pair. For details about how to obtain the SK, see the description for **AK**.	N/A
Project ID	Project ID of DIS	N/A

6.1.3.13 Link to Elasticsearch/CSS

Elasticsearch

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

When connecting CDM to Elasticsearch, configure the parameters as described in **Table 6-20**.

Table 6-20 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	List of one or more Elasticsearch servers, including the port number. The format is *ip:port*. Use semicolons (;) to separate multiple IP addresses.	192.168.0.1:9200; 192.168.0.2:9200

CSS

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

When connecting CDM to CSS, configure the parameters as described in **Table 6-21**.

Copyright © Huawei Technologies Co., Ltd.

Table 6-21 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	List of one or more Elasticsearch servers, including the port number. The format is *ip:port*. Use semicolons (;) to separate multiple IP addresses.	192.168.0.1:9200 ; 192.168.0.2:9200
Security Mode Authentication	Whether to enable security mode. If **Security Mode** has been enabled for the CSS cluster to be connected, set this parameter to **Yes**. Otherwise, set this to **No**.	Yes
Username	This parameter is displayed when **Security Mode Authentication** is set to **Yes**. It indicates the username used for connecting to CSS.	admin
Password	This parameter is displayed when **Security Mode Authentication** is set to **Yes**. It indicates the password used for connecting to CSS.	-

6.1.3.14 Link to DLI

When connecting CDM to DLI, configure the parameters as described in **Table 6-22**.

Table 6-22 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 required for authentication during access to the DLI database. If you have not created access keys for your account, create them as described in **(Optional) Obtaining the Authentication Information**.	-
SK	SK required for authentication during access to the DLI database. For details about how to obtain the SK, see the description for **AK**.	-

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Project ID	Project ID in the region where DLI resides	-

6.1.3.15 Link to CloudTable OpenTSDB

When connecting CDM to CloudTable OpenTSDB, configure the parameters as described in **Table 6-23**.

Table 6-23 Parameter description

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	TSDB_link
OpenTSDB Link	ZK link of OpenTSDB	opentsdb-sp8afz7bgbps5ur.cloudtable.com:4242
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	-
Username	Username for accessing CloudTable	admin
AK	AK for accessing CloudTable. If you have not created access keys, create them as described in **(Optional) Obtaining the Authentication Information**.	-
SK	SK for accessing CloudTable. For details about how to obtain the SK, see the description for **AK**.	-

6.1.3.16 Link to DMS Kafka

When connecting CDM to DMS Kafka, configure the parameters as described in **Table 6-24**.

Copyright © Huawei Technologies Co., Ltd.

Table 6-24 Parameter description

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	dms_link
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, providing higher security.	Yes
Username	Username used for connecting to the Kafka premium instance	-
Password	Password used for connecting to the Kafka premium instance	-

6.1.3.17 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 6-25**.

Table 6-25 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

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Username	If **Authentication Method** is set to **KERBEROS**, you must provide the username and password used for logging in to MRS Manager.	cdm
Password	Password used for logging in to MRS Manager	-
Authentication Method	Authentication method used for accessing MRS • **SIMPLE**: Select this if MRS is in non-security mode. • **KERBEROS**: Select this if MRS is in security mode.	SIMPLE
Run Mode	Run mode of the HBase link. This parameter is used only when the HBase version is **HBASE_2_X**. • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **STANDALONE**: The link instance runs in an independent process. The **STANDALONE** running mode is required in the following scenarios: 1. A 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.) 2. HBase1.x clusters. 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 data 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
Use Cluster Config	You can create cluster configurations on the **Links** page to simplify the configuration of Hadoop link parameters. When creating a link, set this parameter to **Yes**.	No

 Copyright © Huawei Technologies Co., Ltd.

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

FusionInsight HBase

When connecting CDM to HBase of FusionInsight HD, configure the parameters as described in **Table 6-26**.

Table 6-26 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 FusionInsight HD • **SIMPLE**: Select this if FusionInsight HD is in non-security mode. • **KERBEROS**: Select this if FusionInsight HD is in security mode.	Kerberos

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Run Mode	Run mode of the HBase link. This parameter is used only when the HBase version is **HBASE_2_X**. • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **STANDALONE**: The link instance runs in an independent process. The **STANDALONE** running mode is required in the following scenarios: 1. A 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.) 2. HBase1.x clusters. 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 data 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
Use Cluster Config	You can create cluster configurations on the **Links** page to simplify the configuration of Hadoop link parameters. When creating a link, set this parameter to **Yes**.	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

Apache HBase

When connecting CDM to HBase of Apache Hadoop, configure the parameters as described in **Table 6-27**.

 Copyright © Huawei Technologies Co., Ltd.

Table 6-27 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 Hadoop • **SIMPLE**: Select this if Hadoop is in non-security mode. • **KERBEROS**: Select this if Hadoop is in security mode. Obtain the **Principal** account and **Keytab File** file of the client for authentication.	Kerberos
Principal	When **Authentication Method** is set to **KERBEROS**, the **Principal** account is used for authentication. You can contact the Hadoop administrator to obtain the account.	USER@YOUR-REALM.COM
Keytab File	When **Authentication Method** is set to **KERBEROS**, this file is used for authentication. You can contact the Hadoop administrator to obtain the file.	/opt/user.keytab
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

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Run Mode	Run mode of the HBase link. This parameter is used only when the HBase version is **HBASE_2_X**. • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **STANDALONE**: The link instance runs in an independent process. The **STANDALONE** running mode is required in the following scenarios: 1. A 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.) 2. HBase1.x clusters. 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 data 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
Use Cluster Config	You can create cluster configurations on the **Links** page to simplify the configuration of Hadoop link parameters. When creating a link, set this parameter to **Yes**.	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

6.1.3.18 Link to HDFS

CDM supports the following HDFS data sources:

- **MRS HDFS**

Copyright © Huawei Technologies Co., Ltd.

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

MRS HDFS

When connecting CDM to HDFS of MRS, configure the parameters as described in **Table 6-28**.

Table 6-28 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	IP address of MRS Manager. Click **Select** next to the **Manager IP** text box to select an MRS cluster. CDM automatically fills in the authentication information.	127.0.0.1
Username	If **Authentication Method** is set to **KERBEROS**, you must provide the username and password used for logging in to MRS Manager. If you need to create a snapshot when exporting a directory from HDFS, the user configured here must have the administrator permission on HDFS.	cdm
Password	Password used for logging in to MRS Manager	-
Authentication Method	Authentication method used for accessing MRS • **SIMPLE**: Select this if MRS is in non-security mode. • **KERBEROS**: Select this if MRS is in security mode.	SIMPLE

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Run Mode	Run mode of the HDFS link. The options are as follows: • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **STANDALONE**: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, CloudTable, or Hadoop) 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 data 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 **STANDALONE** is selected, CDM can migrate data between HDFSs of multiple MRS clusters. 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 **Agent Management**. This parameter is displayed when **Run Mode** is set to **Agent**.	-
Use Cluster Config	You can create cluster configurations on the **Links** page to simplify the configuration of Hadoop link parameters. When creating a link, set this parameter to **Yes**.	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

Copyright © Huawei Technologies Co., Ltd.

FusionInsight HDFS

When connecting CDM to HDFS of FusionInsight HD, configure the parameters as described in **Table 6-29**.

Table 6-29 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 FusionInsight HD • **SIMPLE**: Select this if FusionInsight HD is in non-security mode. • **KERBEROS**: Select this if FusionInsight HD is in security mode.	KERBEROS

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Run Mode	Run mode of the HDFS link. The options are as follows: • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **STANDALONE**: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, CloudTable, or Hadoop) 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 data 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 **Agent Management**. This parameter is displayed when **Run Mode** is set to **Agent**.	-
Use Cluster Config	You can create cluster configurations on the **Links** page to simplify the configuration of Hadoop link parameters. When creating a link, set this parameter to **Yes**.	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

Copyright © Huawei Technologies Co., Ltd.

Apache HDFS

When connecting CDM to HDFS of Apache Hadoop, configure the parameters as described in **Table 6-30**.

Table 6-30 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	hdfs://nn1.example.com/
Authentication Method	Authentication method used for accessing Hadoop • **SIMPLE**: Select this if Hadoop is in non-security mode. • **KERBEROS**: Select this if Hadoop is in security mode. Obtain the **Principal** account and **Keytab File** file of the client for authentication.	KERBEROS
Principal	When **Authentication Method** is set to **KERBEROS**, the **Principal** account is used for authentication. You can contact the Hadoop administrator to obtain the account.	USER@YOUR-REALM.COM
Keytab File	When **Authentication Method** is set to **KERBEROS**, this file is used for authentication. You can contact the Hadoop administrator to obtain the file.	/opt/user.keytab

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Run Mode	Run mode of the HDFS link. The options are as follows: • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **STANDALONE**: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, CloudTable, or Hadoop) 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 data 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
IP and Host Name Mapping	This parameter is used only when **Run Mode** is set to **EMBEDDED** or **STANDALONE**. If the HDFS configuration file uses the host name, configure the mapping between the IP address and host name. Separate the IP addresses and host names by spaces and mappings by semicolons (;), carriage returns, or line feeds.	10.1.6.9 hostname01 10.2.7.9 hostname02
Agent	If **Run Mode** is set to **Agent**, click **Select** and select the agent created in **Agent Management**.	-
Use Cluster Config	You can create cluster configurations on the **Links** page to simplify the configuration of Hadoop link parameters. When creating a link, set this parameter to **Yes**.	No

Copyright © Huawei Technologies Co., Ltd.

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.	hdfs_01

6.1.3.19 Link to Amazon S3

When connecting CDM to Amazon S3, configure the parameters as described in **Table 6-31**. Currently, objects can only be exported from Amazon S3 to OBS.

Table 6-31 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	linkname
Endpoint	Endpoint of the Amazon S3 bucket	-
Region	Region to which the Amazon S3 bucket belongs	-
AK	AK for accessing the bucket of Amazon S3	-
SK	SK for accessing the bucket of Amazon S3	-

6.1.3.20 Link to KODO/COS

When connecting CDM to KODO or COS, configure the parameters as described in **Table 6-32**. Currently, data can only be exported from KODO/COS to OBS.

Table 6-32 KODO/COS 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	linkname
Region	Region where KODO/COS resides	z0
AK	AK for accessing KODO/COS	-
SK	SK for accessing KODO/COS	-

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Use Custom Domain Name to Download Objects	This parameter is available for KODO links. Whether to preferentially use the custom domain name to download objects from a bucket if the object storage bucket hosts a CDN acceleration domain name or other custom domain names.	Yes

6.1.3.21 Link to OSS on Alibaba Cloud

When connecting CDM to OSS on Alibaba Cloud, configure the parameters as described in **Table 6-33**.

Table 6-33 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	oss_link
OSS Endpoint	Endpoint of Alibaba Cloud OSS	oss-cn-hangzhou.aliyuncs.com
Authentication Method	Available identity authentication methods: • **Access Key**: Use the access key to access OSS. • **Security Token Service**: Use the temporary key and security token to access OSS.	Access Key
AK	AK used for logging in to the OSS server	-
SK	SK used for logging in to the OSS server	-
Security Token	Enter the temporary token provided by Security Token Service (STS).	-
IP and domain Name Mapping	Mapping between IP addresses and domain names	-

6.1.3.22 Link to Relational Databases

CDM supports the following relational databases:

Copyright © Huawei Technologies Co., Ltd.

- Data Warehouse Service
- RDS for MySQL
- RDS for PostgreSQL
- RDS for SQL Server
- MySQL
- PostgreSQL
- Microsoft SQL Server
- Oracle
- IBM Db2
- FusionInsight LibrA
- Derecho (GaussDB)
- SAP HANA

Prerequisites

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

Link Parameters

Table 6-34 describes the required parameters of the link to DWS, RDS for MySQL, RDS for PostgreSQL, RDS for SQL Server, MySQL, PostgreSQL, Microsoft SQL Server, Oracle, IBM Db2, or Derecho (GaussDB).

Table 6-34 Parameter description

Parameter	Description	Example Value
Name	Link name, which should be defined based on the data source type, so it is easier to remember what the link is for	mysql_link
Database Server	IP address or domain name of the database to be connected Click **Select** next to the text box to obtain the list of DWS and RDS instances.	192.168.0.1
Port	Port number of the database to be connected	3306
Connection Type	This parameter is available only for Oracle database links. The options are as follows: • **Service Name**: Use **SERVICE_NAME** to connect to the Oracle database. • **SID**: Use **SID** to connect to the Oracle database.	SID

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Instance Name	Oracle instance ID, which is used to differentiate databases by instances. This parameter is available only when an Oracle link is created and **Database Link Type** is set to **SID**.	dbname
Database Name	Name of the database to be connected	dbname
Username	Username used for accessing the database. This account must be able to read and write data tables and read metadata of the database.	cdm
Password	Password of the account	-
Use Agent	Whether to extract data from the data source through an agent	Yes
Agent	Click **Select** and select the agent created in **Agent Management**.	-
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
Fetch Size	(Optional) This parameter is displayed only after 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

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Use Local API	(Optional) Whether to use the local API of the database for acceleration. When you create a MySQL link, CDM automatically enables the **local_infile** system variable of the MySQL database to enable the LOAD DATA function, which accelerates data import to the MySQL database. If CDM fails to enable this function, contact the database administrator to enable the **local_infile** system variable. Alternatively, set **Use Local API** to **No** to disable API acceleration. If data is imported to RDS for MySQL, the LOAD DATA function is disabled by default. In such a case, you need to modify the parameter group of the MySQL instance and set **local_infile** to **ON** to enable the LOAD DATA function. **NOTE** If **local_infile** on RDS is uneditable, it is the default parameter group. You need to create a parameter group, modify its values, and apply it to the RDS for MySQL instance. For details, see the *Relational Database Service User Guide*.	Yes
SSL Encryption	(Optional) If you set this parameter to **Yes**, CDM can connect to RDS (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
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.	'
Oracle Version	This parameter is displayed only when you create an Oracle link. Select an option according to the version of the Oracle database. If the error message, "java.sql.SQLException: Protocol violation", is displayed, select another option.	Later than 12.1

6.1.3.23 Link to OBS

When connecting CDM to OBS, configure parameters according to **Table 6-35**.

Copyright © Huawei Technologies Co., Ltd.

Table 6-35 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	obs_link
OBS Endpoint	Endpoint of the OBS bucket. For details, see **Regions and Endpoints** You can enter a bucket-level domain name, for example, **test.obs.mycloud.com**. In this case, you can query only the **test** bucket.	obs.cn-north-1.myhuaweicloud.com
Port	Data transmission port. The HTTPS port number is 443 and the HTTP port number is 80.	443
AK	AK 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. If you have not created an AK/SK pair for your account, create one. See **(Optional) Obtaining the Authentication Information**.	N/A
SK	SK used to log in to the OBS server SK is used together with AK. For details about how to obtain the SK, see the description for **AK**.	N/A

6.1.4 Table/File Migration Jobs

Scenario

CDM supports table and file migration between homogeneous or heterogeneous data sources. For details about the data sources that support table/file migration, see **Supported Data Sources**.

CDM is applicable to data migration to the cloud, data exchange on the cloud, and data migration to on-premises service systems.

Prerequisites

- A connection has been created by referring to **Creating Links**.
- The CDM cluster can communicate with the data source.

Procedure

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

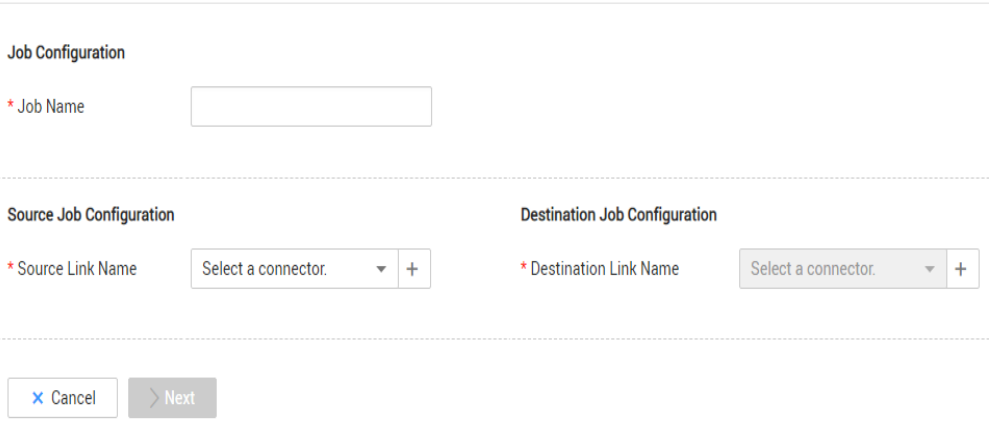
2. Click **Data Integration** of the corresponding workspace.

 Copyright © Huawei Technologies Co., Ltd.

The **Data Integration** page is displayed.

3. Choose **Cloud Data Migration**. The **Cluster Management** page is displayed. Click **Job Management** in the **Operation** column of the cluster.
4. Choose **Table/File Migration** > **Create Job**. The page for configuring the job is displayed.

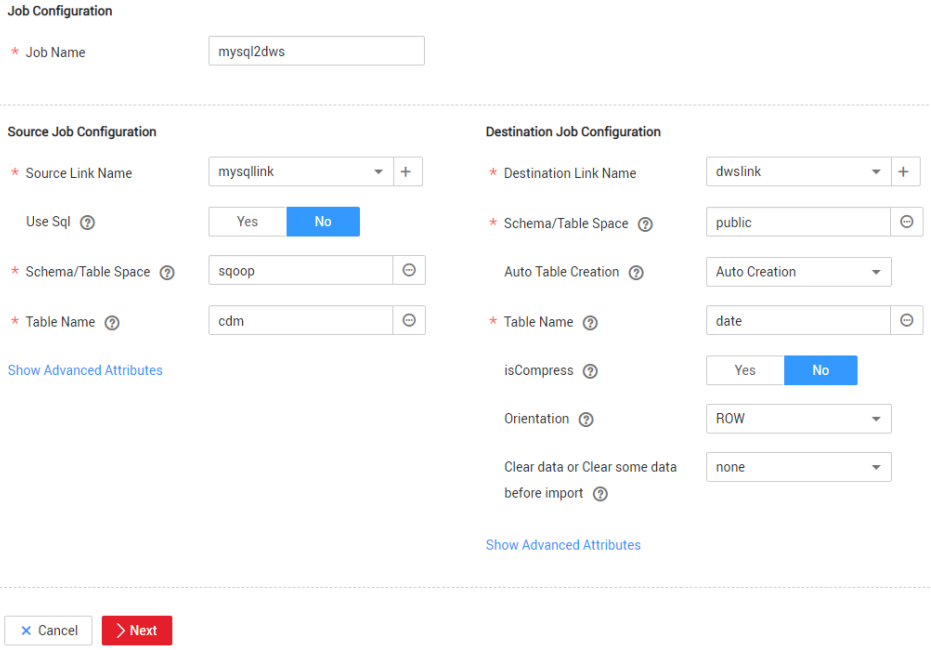
Figure 6-4 Creating a migration job



5. Select the source and destination links.
   - **Job Name**: Enter 1 to 256 characters. Only digits, letters, hyphens (-), underscores (_), and dots (.) are allowed. The first character cannot be a hyphen (-) or dot (.). Example: **oracle2obs_t**.
   - **Source Link Name**: Select the data source from which data is to be exported.
   - **Destination Link Name**: Select the data source to which data is to be imported.
6. If no link is available, click **+** or go to the **Link Management** page to create one. For details about how to create a link, go to **Creating Links**.
7. Configure the source link parameters. **Figure 6-5** shows the job configurations for migrating MySQL to DWS.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-5 Creating a job



The parameters vary with data sources. For details about the job parameters of other types of data sources, see **Table 6-36** and **Table 6-37**.

Table 6-36 Source link parameter description

Migration Source	Description	Parameter Settings
• OBS • Alibaba Cloud OSS • KODO • COS	Data can be extracted in CSV, JSON, CarbonData, or binary format. Data extracted in binary format is free from file resolution, which ensures high performance and is more applicable to file migration. Currently, data cannot be imported to Alibaba Cloud OSS, KODO, and COS.	For details, see **Source OBS Parameters**.
• 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**.

 Copyright © Huawei Technologies Co., Ltd.

Migration Source	Description	Parameter Settings
• MRS HBase • FusionInsight HBase • Apache HBase • CloudTable Service	Data can be exported from MRS, FusionInsight HD, open-source Apache Hadoop HBase, or CloudTable. You need to know all column families and field names of HBase tables.	For details, see **From HBase/CloudTable**.
MRS Hive	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**.
• FTP • SFTP • Network Attached Storage • SFS Turbo	FTP, SFTP, NAS, or SFS data can be exported in CSV, JSON, or binary format.	For details, see **Source FTP/ SFTP/NAS/SFS Parameters**.
• HTTP • HTTPS	These connectors are used to read files with an HTTP/ HTTPS URL, such as reading public files on the third-party object storage system and web disks. Currently, data can only be exported from the HTTP/ HTTPS URLs.	For details, see **Source HTTP/HTTPS Parameters**.
• 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 **Source Relational Database Parameters**.
• FusionInsight LibrA • Derecho (GaussDB)	Data can be exported from FusionInsight LibrA and Derecho.	

Copyright © Huawei Technologies Co., Ltd.

Migration Source	Description	Parameter Settings
• MySQL • PostgreSQL • Oracle • IBM Db2 • Microsoft SQL Server	The non-cloud databases can be those created in the on-premises data center or deployed on ECSs, or database services on the third-party clouds.	
• MongoDB • Document Database Service	Data can be exported from MongoDB or DDS.	For details, see **Source MongoDB/DDS Parameters**.
Redis	Data can be exported from open source Redis.	For details, see **Source Redis Parameters**.
Data Ingestion Service (DIS)	Currently, data can only be exported from DIS to CSS, Apache Kafka, or DMS Kafka.	For details, see **Source DIS Parameters**.
• Apache Kafka • DMS Kafka	Currently, Kafka data can only be exported to CSS, DIS, or DMS Kafka.	For details, see **Source Apache Kafka/DMS Kafka Parameters**.
• Cloud Search Service • Elasticsearch	Data can be exported from CSS or Elasticsearch.	For details, see **Source Elasticsearch or CSS Parameters**.

8. Configure job parameters for the migration destination based on **Table 6-37**.

Table 6-37 Parameters

Migration Destination	Description	Parameter Settings
OBS	Files (even in a large volume) can be batch migrated to OBS in CSV, CarbonData, or binary format.	For details, see **Destination OBS Parameters**.
• MRS HDFS • FusionInsight HDFS • Apache HDFS	You can select a compression format when importing data to HDFS.	For details, see **Destination HDFS Parameters**.

Copyright © Huawei Technologies Co., Ltd.

Migration Destination	Description	Parameter Settings
• MRS HBase • FusionInsight HBase • Apache HBase • CloudTable	Data can be imported to HBase. The compression algorithm can be set when a new HBase table is created.	For details, see **Destination HBase/ CloudTable Parameters**.
MRS Hive	Data can be rapidly imported to MRS Hive.	For details, see **Destination Hive Parameters**.
• FTP • SFTP • Network Attached Storage • SFS Turbo	When FTP/SFTP/NAS servers function as the migration destination, CDM usually migrates cloud data analysis results back to local file systems.	For details, see **Destination FTP/ SFTP/NAS/SFS Parameters**.
• 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 **Destination Relational Database Parameters**. • When importing data to DWS, specify the **COPY** or **GDS** import mode to improve import performance. You can specify the **Import Mode** parameter when **creating DWS links**. • When importing data to RDS for MySQL, enable the LOAD DATA function of MySQL to accelerate data import and improve the import performance. You can configure **Use**
FusionInsight LibrA	Data can be imported to FusionInsight LibrA but cannot be imported to Derecho (GaussDB).	

Copyright © Huawei Technologies Co., Ltd.

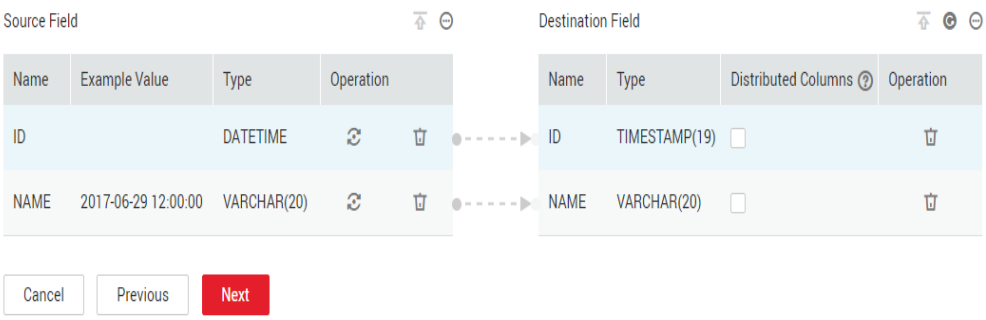
Migration Destination	Description	Parameter Settings
MySQL	The data source can be the on-premises MySQL, MySQL built on ECSs, or MySQL on the third-party cloud.	**Local API** to enable the function when you **create MySQL links**.
Document Database Service	Data can be imported to the DDS but cannot be imported to the local MongoDB.	For details, see **Destination DDS Parameters**.
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 **Destination DCS Parameters**.
• Cloud Search Service • Elasticsearch	Data can be imported to Elasticsearch or CSS.	For details, see **Destination Elasticsearch or CSS Parameters**.
Data Lake Insight	Data can be imported to DLI.	For details, see **Destination DLI Parameters**.

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

   If files are migrated between FTP, SFTP, NAS, 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 6-6 Field mapping



 Copyright © Huawei Technologies Co., Ltd.

NOTE

- If the field mapping is incorrect, you can drag the fields to adjust the mapping.
- 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 fails 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.

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

    CDM supports the following converters:

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

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

       i. Set **Reserve Start Length** to **3**.

       ii. Set **Reserve End Length** to **4**.

       iii. Set **Replace Character** to *****.

    c. **Trim**: automatically deletes the spaces before and after the character string.

    d. **Reverse string**: automatically reverses a character string. For example, reverse **ABC** into **CBA**.

    e. **Replace string**: replaces the specified character string.

    f. **Expression conversion**: uses the JSP expression language (EL) to convert the current field or a row of data. For details, see **Field Conversion**.

    g. **Remove line break**: delete the newline characters, such as \n, \r, and \r\n from the field.

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

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-7 Task parameters

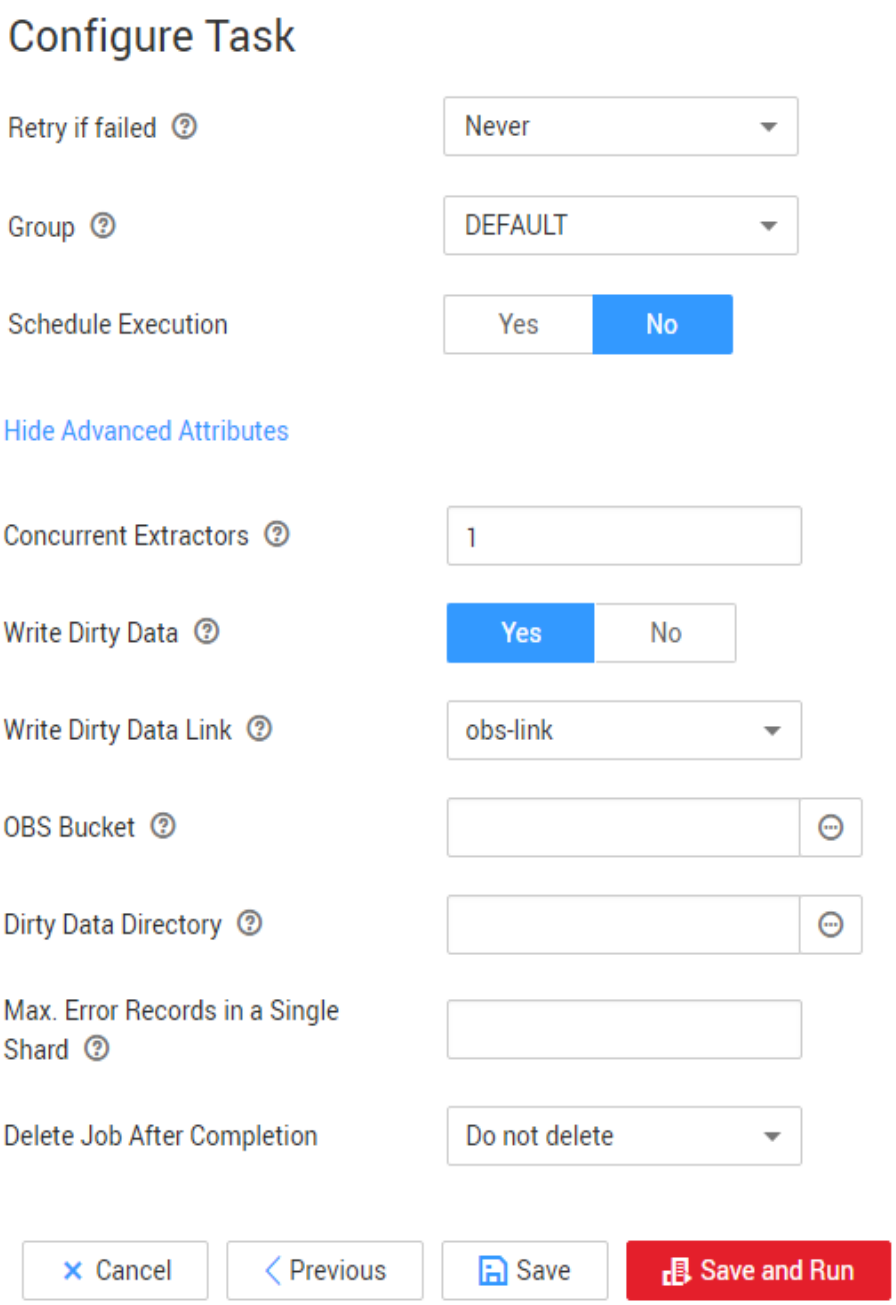


Table 6-38 describes related parameters.

 Copyright © Huawei Technologies Co., Ltd.

Table 6-38 Parameter description

Parameter	Description	Example Value
Retry upon Failure	You can select **Retry 3 times** or **Never**. You are advised to configure automatic retry for only file migration jobs or database migration jobs with **Import to Staging Table** enabled to avoid data inconsistency caused by repeated data writes.	Never
Job	Select a group where the job resides. The default group is **DEFAULT**. On the **Job Management** page, jobs can be displayed, started, or exported by group.	DEFAULT
Schedule Execution	If you select **Yes**, you can set the start time, cycle, and validity period of a job. For details, see **Scheduling Job Execution**.	No
Concurrent Extractors	Number of extractors to be concurrently executed. Generally, retain the default value.	1
Concurrent Loaders	Number of Loaders to be concurrently executed This parameter is displayed only when HBase or Hive serves as the destination data source.	3
Write Dirty Data	Whether to record dirty data. By default, this parameter is set to **No**.	Yes
Write Dirty Data Link	This parameter is displayed only when **Write Dirty Data** is set to **Yes**. Only links to OBS support dirty data writes.	obs_link
OBS Bucket	This parameter is displayed only when **Write Dirty Data Link** is a link to OBS. Name of the OBS bucket to which the dirty data is to be written.	dirtydata

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Dirty Data Directory	This parameter is displayed only when **Write Dirty Data** is set to **Yes**. Dirty data is stored in the directory for storing dirty data on OBS. Dirty data is saved only when this parameter is configured. You can go to this directory to query data that fails to be processed or is filtered out during job execution, and check the source data that does not meet conversion or cleaning rules.	/user/dirtydir
Max. Error Records in a Single Shard	This parameter is displayed only when **Write Dirty Data** is set to **Yes**. When the number of error records of a single map exceeds the upper limit, the job will automatically terminate and the imported data cannot be rolled back. You are advised to use a temporary table as the destination table. After the data is imported, rename the table or combine it into the final data table.	0
Delete Job After Completion	Perform the following operations after a job is executed: • **Do not delete**: The job is not deleted after it is executed. • **Delete after success**: The job is deleted only when the job is successfully executed. It is applicable to massive one-time jobs. • **Delete**: The job is deleted regardless of whether it is successfully executed or fails to be executed.	Do not delete

12. Click **Save** or **Save and Run**. On the page that is displayed, 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.

6.1.5 Source Job Parameters

Copyright © Huawei Technologies Co., Ltd.

6.1.5.1 Source OBS Parameters

If the source link of a job is the **Link to OBS**, **Link to OSS on Alibaba Cloud**, **Link to KODO/COS**, or **Link to Amazon S3**, configure the source job parameters based on **Table 6-39**.

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

Table 6-39 Parameters

Category	Parameter	Description	Example Value
Basic parameters	Bucket Name	Name of the bucket from which data is to be migrated	BUCKET_2
	Source Directory/File	Path of the directory or file from which data is to be extracted. You can enter a maximum of 50 file paths. By default, the file paths are separated by vertical bars (|). You can also customize a file separator. For details, see **Migration of a List of Files**. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**.	FROM/example.csv
	File Format	Format in which CDM parses data. The options are as follows: • **CSV**: Source files will be migrated to tables after being parsed in CSV format. • **Binary**: Files (even not in binary format) will be directly transferred without resolution. It is applicable to file copy. • **JSON**: Source files will be migrated to tables after being parsed in JSON format. • **CarbonData**: Source files will be migrated to tables after being parsed in CarbonData format. This parameter is displayed only when the migration source is OBS.	CSV

Copyright © Huawei Technologies Co., Ltd.

Category	Parameter	Description	Example Value
	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 displayed only when **File Format** is set to **JSON**. CDM parses the data under the JSON node. If the node's corresponding data is an array of JSON, the system will extract data from the array in the same pattern. Use periods (.) to separate multi-layer nested JSON nodes.	data.list
Advanced attributes	Line Separator	Lind feed characters in a file. Supported values include **\n**, **\r**, and **\r\n**. This parameter is displayed only when **File Format** is set to **CSV**.	\n
	Field Delimiter	Character used to separate fields in the file. To set the **Tab** key as the delimiter, set this parameter to **\t**. This parameter is displayed only when **File Format** is set to **CSV**.	,
	Use Quote Character	If you set this parameter to **Yes**, the field delimiters in the encircling symbol are regarded as a part of the string value. Currently, the default encircling symbol of CDM is **"**.	No
	Use RE to Separate Fields	Whether to use regular expressions to separate fields. If you set this parameter to **Yes**, **Field Delimiter** becomes invalid. This parameter is displayed only when **File Format** is set to **CSV**.	Yes
	Regular Expression	Regular expression used to separate fields. For details about regular expressions, see **Regular Expressions for Separating Semi-structured Text**.	^(\d.*\d) (\w*) \[(.*) \] ([\w\.]*) (\w.*).*
	Use First Row as Header	This parameter is displayed only when **File Format** is set to **CSV**. When you migrate a CSV file to a table, CDM writes all data to the table by default. If you set this parameter to **Yes**, CDM uses the first line of the CSV file as the heading line and does not write the line to the destination table.	No

Copyright © Huawei Technologies Co., Ltd.

Category	Parameter	Description	Example Value
	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
	Compression Format	This parameter is displayed only when **File Format** is set to **CSV** or **JSON**. The options are as follows: • **None**: indicates that files in all formats can be transferred. • **gzip**: indicates that only files in gzip format can be transferred. • **Zip**: indicates that only files in Zip format can be transferred.	None
	Source File Processing Method	Operation on source files after the job succeeds. • **Rename**: After the job succeeds, rename the source files by adding usernames and timestamps as suffixes to file names. • **Delete**: After the job succeeds, delete the source files.	Rename
	Start Job by Marker File	Whether to start a job by a marker file. A job is started only when a marker file for starting the job exists in the source path. Otherwise, the job will be suspended for a period of time specified by **Suspension Period**.	No
	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	Period of waiting for a marker file. If you set **Start Job by Marker File** to **Yes** but no marker file exists in the source path, the job fails upon suspension timeout. If you set this parameter to 0 and no marker file exists in the source path, the job will fail immediately. Unit: second	10

 Copyright © Huawei Technologies Co., Ltd.

Category	Parameter	Description	Example Value
	File Separator	File separator. If you enter multiple file paths in **Source Directory/Files**, CDM uses the file separator to separate files. The default value is **|**.	|
	Wildcard	If you select **Yes**, enter wildcard characters to filter files. The paths or files that meet the filter conditions are transferred. For details, see **File/Path Filter**.	Yes
	Path Filter	If you set **Filter Type** to **Wildcard**, enter a wildcard character to filter paths. The paths that meet the filter condition are migrated. You can configure multiple paths separated by commas (,).	*input
	File Filter	If you set **Filter Type** to **Wildcard**, enter a wildcard character to filter files in a specified path. The files that meet the filter condition are migrated. You can configure multiple files separated by commas (,).	*.csv,*.txt
	Time Filter	If you select **Yes**, files can be transferred according to the modification time of the files.	Yes
	Minimum Timestamp	If you set **Filter Type** to **Time Filter**, specify a time point to filter files. Only the files modified after the specified time point are transferred. The time format must be *yyyy-MM-dd HH:mm:ss*. This parameter can be set to a macro variable of date and time. For example, **${timestamp(dateformat(yyyy-MM-dd HH:mm:ss,-90,DAY))}** indicates that only files generated within the latest 90 days are migrated.	2019-06-01 00:00:00

 Copyright © Huawei Technologies Co., Ltd.

Category	Parameter	Description	Example Value
	Maximum Timestamp	If you set **Filter Type** to **Time Filter**, specify a time point to filter files. Only the files modified before the specified time point are transferred. The time format must be *yyyy-MM-dd HH:mm:ss*. This parameter can be set to a macro variable of date and time. For example, **${timestamp(dateformat(yyyy-MM-dd HH:mm:ss))}** indicates that only the files whose modification time is earlier than the current time are migrated.	2019-07-01 00:00:00
	Encryption	If the source data is encrypted, CDM can decrypt the data before exporting it. Select whether to decrypt the source data and select a decryption algorithm. The options are as follows: • **None**: Directly export the data without decrypting it. • **AES-256-GCM**: The AES 256-bit encryption algorithm is used to encrypt data. Currently, only the AES-256-GCM (NoPadding) encryption algorithm is supported. This parameter is used for encryption at the migration destination and decryption at the migration source. For details, see **Encryption and Decryption During File Migration**.	AES-256-GCM
	DEK	This parameter is displayed only when **Encryption** is set to **AES-256-GCM**. The key consists of 64 hexadecimal numbers and must be the same as the **DEK** configured during encryption. If the decryption and encryption keys are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B

Copyright © Huawei Technologies Co., Ltd.

Category	Parameter	Description	Example Value
	IV	This parameter is displayed only when **Encryption** is set to **AES-256-GCM**. The initialization vector consists of 32 hexadecimal numbers and must be the same as the **IV** configured during encryption. If the initialization vectors are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	5C91687BA886EDCD12ACBC3FF19A3C3F
	MD5 File Extension	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.

6.1.5.2 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 6-40**.

Copyright © Huawei Technologies Co., Ltd.

Table 6-40 Parameter description

Category	Parameter	Description	Example Value
Basic parameters	Source Directory/File	Directory or file path from which data will be extracted. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**.	/user/cdm/
	File Format	File format used when transferring data. The options are as follows: • **CSV**: Source files will be migrated to tables after being converted to CSV format. • **Binary**: Files (even not in binary format) will be transferred directly. It is used for file copy. • **Parquet**: Source files will be migrated to tables after being converted to Parquet format.	CSV
	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-hwstaff
	Path/ Directory of entries files	Custom OBS directories that store the text file. Use slashes (/) to separate different directories.	test1
Advanced attributes	Line Separator	Lind feed character in a file. By default, the system automatically identifies **\n**, **\r**, and **\r\n**. This parameter is displayed only when **File Format** is set to **CSV**.	\n

Copyright © Huawei Technologies Co., Ltd.

Category	Parameter	Description	Example Value
	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 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
	Encode Type	The encoding type can be set only for text files. Otherwise, the setting is invalid. The value can be **UTF-8** or **GBK**.	
	Source File Processing Method	Operation performed on source files after the job completes. • **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. • Do Nothing.	Rename
	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	Select a file filter type.	-
	Wildcard	If you select **Yes**, enter wildcard characters to filter files. All paths or files that meet the search criteria are transferred. For details, see **File/Path Filter**.	Yes

Copyright © Huawei Technologies Co., Ltd.

Category	Parameter	Description	Example Value
	Directory Filter	If you set **Filter Type** to **Wildcard**, enter a wildcard character to filter paths. The paths that meet the filtering condition are migrated. You can configure multiple paths separated by commas (,).	*input
	File Filter	If you set **Filter Type** to **Wildcard**, you can enter a wildcard character to search for files in a specified path. The files that meet the search criteria are migrated. You can configure multiple files separated by commas (,).	*.csv
	Time Filter	If you select **Yes**, files are transferred based on their modification time.	Yes
	Minimum Timestamp	If you set **Filter Type** to **Time Filter**, and specify a point in time for this parameter, only the files modified after the specified time are transferred. The time format must be *yyyy-MM-dd HH:mm:ss*. This parameter can be set to a macro variable of date and time. For example, **${timestamp(dateformat(yyyy-MM-dd HH:mm:ss,-90,DAY))}** indicates that only files generated within the latest 90 days are migrated.	2019-07-01 00:00:00
	Maximum Timestamp	If you set **Filter Type** to **Time Filter**, and specify a point in time for this parameter, only the files modified before the specified time are transferred. The time format must be *yyyy-MM-dd HH:mm:ss*. This parameter can be set to a macro variable of date and time. For example, **${timestamp(dateformat(yyyy-MM-dd HH:mm:ss))}** indicates that only the files whose modification time is earlier than the current time are migrated.	2019-07-30 00:00:00

 Copyright © Huawei Technologies Co., Ltd.

Category	Parameter	Description	Example Value
	Create Snapshot	If you set this parameter to **Yes**, CDM creates a snapshot for the source directory to be migrated (the snapshot cannot be created for a single file) before it reads files from HDFS. Then CDM migrates the data in the snapshot. Only the HDFS administrator can create a snapshot. After the CDM job is completed, the snapshot is deleted.	No
	Encryption	This parameter is displayed only when **File Format** is set to **Binary**. If the source data is encrypted, CDM can decrypt the data before exporting it. Select whether to decrypt the source data and select a decryption algorithm. The options are as follows: • **NONE**: Export data without decrypting it. • **AES-256-GCM**: The AES 256-bit encryption algorithm is used to encrypt data. Currently, only the AES-256-GCM (NoPadding) encryption algorithm is supported. This parameter is used for encryption at the migration destination and decryption at the migration source. For details, see **Encryption and Decryption During File Migration**.	AES-256-GCM
	DEK	This parameter is displayed only when **Encryption** is set to **AES-256-GCM**. The key consists of 64 hexadecimal numbers and must be the same as the **DEK** configured during encryption. If the decryption and encryption keys are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B

Copyright © Huawei Technologies Co., Ltd.

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

NOTE

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

6.1.5.3 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 6-41**.

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**.

Copyright © Huawei Technologies Co., Ltd.

Table 6-41 Parameter description

Parameter	Description	Example Value
Table Name	Name of the HBase table that data will be exported from This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**.	TBL_2
Column Families	(Optional) Column families to which the exported data belongs	CF1&CF2
Split Rowkey	(Optional) Whether to split a rowkey. The default value is **No**.	Yes
Rowkey Delimiter	(Optional) Delimiter used to split a rowkey. If this parameter is left empty, the rowkey will not be split.	|
Start Time	(Optional) Start time (including the value) for extracting data. The format is *yyyy-MM-dd HH:mm:ss*. Only the data generated at the specified time and later is extracted. This parameter can be set to a macro variable of date and time. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**.	2019-01-01 20:00:00
End Time	(Optional) End time (excluding the value) for extracting data. The format is *yyyy-MM-dd HH:mm:ss*. Only the data generated before the time point is extracted. This parameter can be set to a macro variable of date and time. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**.	2019-02-01 20:00:00

6.1.5.4 From Hive

If the source link of a job is the **Link to Hive**, configure the source job parameters based on **Table 6-42**.

Copyright © Huawei Technologies Co., Ltd.

Table 6-42 Parameter description

Parameter	Description	Example Value
Database Name	Database name. Click the icon next to the text box. The dialog box for selecting the database is displayed.	default
Table Name	Hive table name. Click the icon next to the text box. The dialog box for selecting the table is displayed. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**.	TBL_E
Partition Filter Criteria	This parameter is displayed when you click **Show Advanced Attributes**. You can configure multiple values (separated by spaces) or a range. The time macro function is supported.	• Single/ Multi-value filtering: "${dateformat(yyyyMMdd, -1, DAY)} ${dateformat(yyyyMMdd)}" • Filter by range: "${value} >= ${dateformat(yyyyMMdd, -7, DAY)} && ${value} < ${dateformat(yyyyMMdd)}"

NOTE

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

 Copyright © Huawei Technologies Co., Ltd.

6.1.5.5 Source FTP/SFTP/NAS/SFS Parameters

If the source link of a job is the **Link to an FTP or SFTP Server** or **Link to NAS/SFS**, configure the source job parameters based on **Table 6-43**.

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

Table 6-43 Parameter description

Category	Parameter	Description	Example Value
Basic parameters	Source Directory/File	Directory or file path from which data will be extracted. You can enter a maximum of 50 file paths. By default, the file paths are separated by vertical bars (|). You can also customize a file separator. For details, see **Migration of a List of Files**. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**.	/ftp/a.csv|/ftp/b.txt
	File Format	Format in which CDM parses data. The options are as follows: • **CSV**: Source files will be migrated to tables after being converted to CSV format. • **Binary**: Files (even not in binary format) will be transferred directly. It is used for file copy. • **JSON**: Source files will be migrated to tables after being converted to JSON format.	CSV
	JSON Type	This parameter is displayed only when **File Format** is set to **JSON**. Type of a JSON object stored in a JSON file. The options are **JSON object** and **JSON array**.	JSON object

Copyright © Huawei Technologies Co., Ltd.

Category	Parameter	Description	Example Value
	JSON Reference Node	This parameter is used only when **File Format** is set to **JSON** and **JSON Type** is set to **JSON Object**. CDM parses the data under the JSON node. If the node's corresponding data is a JSON array, the system will extract data from the array in the same pattern. Use periods (.) to separate multi-layer nested JSON nodes.	data.list
Advanced attributes	Line Separator	Lind feed character in a file. By default, the system automatically identifies **\n**, **\r**, and **\r\n**. This parameter is displayed only when **File Format** is set to **CSV**.	\n
	Field Delimiter	Character used to separate fields in the file. To set the **Tab** key as the delimiter, set this parameter to **\t**. This parameter is displayed only when **File Format** is set to **CSV**.	,
	Use Quote Character	If you set this parameter to **Yes**, the field delimiters in the encircling symbol are regarded as a part of the string value. Currently, the default encircling symbol of CDM is **"**.	No
	Use RE to Separate Fields	Whether to use regular expressions to separate fields. If you set this parameter to **Yes**, **Field Delimiter** becomes invalid. This parameter is displayed only when **File Format** is set to **CSV**.	Yes
	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

 Copyright © Huawei Technologies Co., Ltd.

Category	Parameter	Description	Example Value
	Compression Format	This parameter is displayed only when **File Format** is set to **CSV** or **JSON**. The options are as follows: • **NONE**: Files in all formats can be transferred. • **GZIP**: Only files in gzip format can be transferred. • **ZIP**: Only files in Zip format can be transferred.	None
	Source File Processing Method	Operation performed on source files after the job completes. • **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.	Rename
	Start Job by Marker File	Whether to start a job by a marker file. A job is only started if there is a marker file for starting the job in the source path. If there is no marker file, the job will be suspended for a period of time specified by **Suspension Period**.	Yes
	Marker File	Name of the marker file for starting a job. If you specify a marker file, the migration job is executed only when the marker file exists in the source path. The marker file will not be migrated.	ok.txt
	Suspension Period	Waiting period for a marker file. If you set **Start Job by Marker File** to **Yes** but there is no marker file in the source path, the job fails when the suspension period times out. If you set this parameter to **0** and there is no marker file in the source path, the job will fail immediately. Unit: second	10
	File Separator	File separator. If you enter multiple file paths in **Source Directory/Files**, CDM uses the file separator to identify files. The default value is **|**.	|

 Copyright © Huawei Technologies Co., Ltd.

Category	Parameter	Description	Example Value
	Filter Type	Only files that meet the filtering conditions are transferred. The options are **None**, **Wildcard**, and **Regex**.	None
	Wildcard	If you select **Yes**, enter wildcard characters to filter files. All paths or files that meet the search criteria are transferred. For details, see **File/Path Filter**.	Yes
	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 (,).	*input,*out
	File Filter	If you set **Filter Type** to **Wildcard**, you can enter a wildcard character to search for files in a specified path. The files that meet the search criteria are migrated. You can configure multiple files separated by commas (,).	*.csv
	Time Filter	If you select **Yes**, files are transferred based on their modification time.	Yes
	Minimum Timestamp	If you set **Filter Type** to **Time Filter**, and specify a point in time for this parameter, only the files modified after the specified time are transferred. The time format must be *yyyy-MM-dd HH:mm:ss*. This parameter can be set to a macro variable of date and time. For example, **${timestamp(dateformat(yyyy-MM-dd HH:mm:ss,-90,DAY))}** indicates that only files generated within the latest 90 days are migrated.	2019-07-01 00:00:00
	Maximum Timestamp	If you set **Filter Type** to **Time Filter**, and specify a point in time for this parameter, only the files modified before the specified time are transferred. The time format must be *yyyy-MM-dd HH:mm:ss*. This parameter can be set to a macro variable of date and time. For example, **${timestamp(dateformat(yyyy-MM-dd HH:mm:ss))}** indicates that only the files whose modification time is earlier than the current time are migrated.	2019-07-30 00:00:00

Copyright © Huawei Technologies Co., Ltd.

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

Copyright © Huawei Technologies Co., Ltd.

6.1.5.6 Source HTTP/HTTPS Parameters

When the source link of a job is the HTTP link, configure the source job parameters based on **Table 6-44**. Currently, data can only be exported from the HTTP/HTTPS URLs.

Table 6-44 Parameter description

Parameter	Description	Example Value
File URL	Use the GET method to obtain data from the HTTP/HTTPS URL. These connectors are used to read files with an HTTP/HTTPS URL, such as reading public files on the third-party object storage system and web disks.	https://bucket.obs.myhuaweicloud.com/object-key
File Format	CDM supports **Binary** only, which indicates that files (even not in binary format) will be directly transferred.	Binary
Compression Format	Compression format of the source files. The options are as follows: • **NONE**: Files in all formats can be transferred. • **GZIP**: Only files in gzip format can be transferred. • **ZIP**: Only files in Zip format can be transferred. • **TAR.GZ**: Files in TAR.GZ format are transferred.	None
Compressed File Extension	Extension of the files to be decompressed. This parameter is only displayed when **Compression Format** is not **NONE**. The decompression operation is only performed when the filename 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**.	|
Query Parameter	• If you set this parameter to **Yes**, the name of the objects uploaded to OBS does not include the **query** parameter. • If you set this parameter to **No**, the name of the objects uploaded to OBS includes the **query** parameter.	No

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Encryption	If the source data is encrypted, CDM can decrypt the data before exporting it. Select whether to decrypt the source data and select a decryption algorithm. The options are as follows: • **NONE**: Export data without decrypting it. • **AES-256-GCM**: The AES 256-bit encryption algorithm is used to encrypt data. Currently, only the AES-256-GCM (NoPadding) encryption algorithm is supported. This parameter is used for encryption at the migration destination and decryption at the migration source. For details, see **Encryption and Decryption During File Migration**.	AES-256-GCM
DEK	This parameter is displayed only when **Encryption** is set to **AES-256-GCM**. The key consists of 64 hexadecimal numbers and must be the same as the **DEK** configured during encryption. If the decryption and encryption keys are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B
IV	This parameter is displayed only when **Encryption** is set to **AES-256-GCM**. The initialization vector consists of 32 hexadecimal numbers and must be the same as the **IV** configured during encryption. If the initialization vectors are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	5C91687BA886EDCD12ACBC3FF19A3C3F
MD5 File Extension	Check whether the files extracted by CDM are consistent with source files. For details, see **MD5 Verification**.	.md5

6.1.5.7 Source Relational Database Parameters

When the source link of a job is one of the relational databases listed in **Link to Relational Databases** (also listed here), configure the source job parameters based on **Table 6-45**.

- Data Warehouse Service
- RDS for MySQL
- RDS for SQL Server
- RDS for PostgreSQL
- FusionInsight LibrA
- Derecho (GaussDB)

Copyright © Huawei Technologies Co., Ltd.

- MySQL
- PostgreSQL
- Oracle
- IBM Db2
- Microsoft SQL Server

Table 6-45 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.	select id,name from sqoop.user;
	Schema/ Tablespace	Name of the schema or tablespace from which data will be extracted. This parameter is displayed when **Use SQL Statement** is set to **No**. Click the icon next to the text box to go to the page for selecting a schema or directly enter a schema or tablespace. If the desired schema or tablespace is not displayed, check whether the login account has the permissions required to query metadata. **NOTE** The parameter value can contain wildcard characters (*), which is used to export all databases whose names start with a certain prefix or end with a certain suffix. The examples are as follows: • **SCHEMA*** indicates that all databases whose names starting with **SCHEMA** are exported. • ***SCHEMA** indicates that all databases whose names ending with **SCHEMA** are exported. • ***SCHEMA*** indicates that all databases whose names containing **SCHEMA** are exported.	SCHEMA_E

 Copyright © Huawei Technologies Co., Ltd.

Category	Parameter	Description	Example Value
	Table Name	Name of the table from which data will be extracted. This parameter is displayed when **Use SQL Statement** is set to **No**. Click the icon next to the text box to go to the page for selecting the table or directly enter a table name. If the desired table is not displayed, confirm that the table exists or that the login account has the permissions required to query metadata. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**. **NOTE** The table name can contain wildcard characters (*), which is used to export all tables whose names start with a certain prefix or end with a certain suffix. The number and types of fields in the tables must be the same. The examples are as follows: • **table*** indicates that all tables whose names starting with **table** are exported. • ***table** indicates that all tables whose names ending with **table** are exported. • ***table*** indicates that all tables whose names containing **table** are exported.	table
Advanced attributes	Partition Column	This parameter is displayed when **Use SQL Statement** is set to **No**, indicating that a field used to split data during data extraction. CDM splits a job into multiple tasks based on this field and executes the tasks concurrently. Fields with data distributed evenly are used, such as the sequential number field. Click the icon next to the text box to go to the page for selecting a field or directly enter a field.	id

 Copyright © Huawei Technologies Co., Ltd.

Category	Parameter	Description	Example Value
	WHERE Clause	WHERE clause used to specify the data extraction range. This parameter is displayed when **Use SQL Statement** is set to **No**. If this parameter is not set, the entire table is extracted. This parameter can be configured as a macro variable of date and time to extract data generated at a specific date. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**.	DS='${dateformat(yyyy-MM-dd,-1,DAY)}'
	Null in Partition Column	Whether the partition column can contain null values	Yes
	Extract by Partition	When data is exported from an Oracle database, data can be extracted from each partition in the partitioned table. If this function is enabled, you can configure **Table Partition** to specify specific Oracle table partitions from which data is extracted. • This function does not support non-partitioned tables. • The database user must have the **SELECT** permission on the system views **dba_tab_partitions** and **dba_tab_subpartitions**.	No
	Table Partition	Oracle table partition from which data is migrated. Separate multiple partitions with ampersands (&). If you do not set this parameter, all partitions will be migrated. If there is a subpartition, enter the partition in the *Partition.Subpartition* format, for example, **P2.SUBP1**.	P0&P1&P2.SUBP1&P2.SUBP3

Copyright © Huawei Technologies Co., Ltd.

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.
- In a migration from MySQL to DWS, the constraints on the incremental data migration function in MySQL Binlog mode are as follows:
  - A single cluster supports only one incremental migration job in MySQL Binlog mode in the current version.
  - In the current version, you are not allowed to delete or update 10,000 data records at a time.
  - Entire DB migration is not supported.
  - Data Definition Language (DDL) operations are not supported.
  - Event migration is not supported.
  - If you set **Migrate Incremental Data** to **Yes**, **binlog_format** in the source MySQL database must be set to **ROW**.
  - 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.
  - 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.
  - 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
```

6.1.5.8 Source MongoDB/DDS Parameters

When you migrate data from MongoDB to a relational database, 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 a MongoDB link described in **Link to MongoDB**, that is, when data is exported from an on-premises MongoDB or DDS, configure the source job parameters based on **Table 6-46**.

Table 6-46 Parameter description

Parameter	Description	Example Value
Database Name	Name of the database from which data is to be migrated	mongodb

Copyright © Huawei Technologies Co., Ltd.

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

6.1.5.9 Source Redis Parameters

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 6-47**.

Table 6-47 Parameter description

Parameter	Description	Example Value
Redis Key Prefix	Key prefix, which is similar to the table name of a relational database	TABLE

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Value Storage Type	The options are as follows: • **String**: without column name, such as **value1,value2** • **Hash**: with column name, such as **column1=value1,column2=value2**	String
Key Delimiter	Character used to separate table names and column names of a relational database	_
Value Delimiter	Character used to separate columns when the storage type is string	;
Same Field	This parameter is displayed when **Value Storage Type** is set to **Hash**. The hash key contains the same field.	Yes

6.1.5.10 Source DIS Parameters

Currently, data can only be exported from DIS to CSS, Apache Kafka, or DMS Kafka.

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.

If the source link of a job is the **Configuring Links to DIS**, configure the source job parameters based on **Table 6-48**.

Table 6-48 Parameter description

Parameter	Description	Example Value
DIS Stream	DIS stream name	dis
Offset	Initial offset when data is pulled from DIS • **Latest**: Maximum offset, indicating that the latest data will be extracted. • **From last stop**: Data read will start from which the last read ended. • **Earliest**: Minimum offset, indicating that the earliest data will be extracted.	Latest
Permanent Running	Whether a job runs permanently. If a job is set to run for a long time, the job will fail if the DIS system is interrupted.	Yes

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
DIS Partition ID	ID of the DIS partition. You can enter multiple partition IDs separated by commas (,).	0,1,2
Data Format	Format used for parsing data. The options are as follows: • **Binary**: Data is transferred directly. It is not converted to another format. This setting is suitable for file migration. • **CSV**: Source data will be migrated after being converted in CSV format.	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
Application Name	Unique identifier of the consumer application to be used. If no application exists, CDM creates one automatically.	cdm

6.1.5.11 Source Apache Kafka/DMS Kafka Parameters

Currently, data can only be exported from Kafka to CSS, DIS, or 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 6-49**.

Table 6-49 Parameter description

Parameter	Description	Example Value
Topics	One or more topics can be entered.	est1,est2
Offset	Initial offset parameter • **Latest**: Maximum offset, indicating that the latest data will be extracted. • **Earliest**: Minimum offset, indicating that the earliest data will be extracted.	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

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Data Format	Format used for parsing data. The options are as follows: • **Binary**: Data is transferred directly. It is not converted to another format. This setting is suitable for file migration. • **CSV**: Source data will be migrated after being converted in CSV format.	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

6.1.5.12 Source Elasticsearch or CSS Parameters

If the source link of a job is the **Link to Elasticsearch/CSS**, configure the source job parameters based on **Table 6-50**.

Table 6-50 Parameter description

Parameter	Description	Example Value
Index	Elasticsearch index, which is similar to the name of a relational database. The index name can contain only lowercase letters.	index
Type	Elasticsearch type, which is similar to the table name of a relational database. The type name can contain only lowercase letters.	type
Split Nested Field	(Optional) Whether to split the JSON content of the nested fields. For example, **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
Filter Conditions	(Optional) Whether to use a query string to filter the source data. CDM only migrates the data that meets the filtering conditions.	last_name:Smith
Extract Meta-field	Whether to extract index meta-fields. For example, _index, _type, _id, and _score.	Yes

Copyright © Huawei Technologies Co., Ltd.

6.1.5.13 Source OpenTSDB Parameters

If the source link of a job is the **Link to CloudTable OpenTSDB**, configure the source job parameters based on **Table 6-51**.

Table 6-51 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 *yyyyMMddHHmmdd*.	20180920145505
End Time	(Optional) End time of the query. The value is a string or timestamp in the format of *yyyyMMddHHmmdd*.	1598870800
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

6.1.6 Destination Job Parameters

6.1.6.1 Destination OBS Parameters

If the destination link of a job is the **Link to OBS**, that is, if data is imported to OBS, configure the destination job parameters based on **Table 6-52**.

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

Table 6-52 Parameter description

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

Copyright © Huawei Technologies Co., Ltd.

Category	Parameter	Description	Example Value
	Write Directory	OBS directory to which data is to be written. Do not add **/** in front of the directory name. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**.	directory/
	File Format	Format in which data is written. The options are as follows: • **CSV**: Data is written in CSV format, which is applicable to migrating data tables to files. • **Binary**: Files will be directly transferred without resolution. CDM writes the files without changing the original file format, which is applicable to the migration of files to files. • **CarbonData**: Data is written in CarbonData format, which is applicable to migrating data tables to files. If data is migrated between file-related data sources, such as FTP, SFTP, NAS, HDFS, and OBS, the value of **File Format** must the same as the source file format.	CSV

Copyright © Huawei Technologies Co., Ltd.

Category	Parameter	Description	Example Value
	Duplicate File Processing Method	Files with the same name and size are identified as duplicate files. If there are duplicate files during data writing, the following methods are available: • Replace • Skip • Stop job For details, see **Skipping a Duplicate File**.	Skip
Advanced settings	Encryption	Whether to encrypt the uploaded data and the encryption mode. The options are as follows: • **None**: Directly write data without encryption. • **KMS**: Use KMS in Data Encryption Workshop (DEW) for encryption. If KMS encryption is enabled, MD5 verification for data cannot be performed. • **AES-256-GCM**: Use the AES 256-bit encryption algorithm to encrypt data. Currently, only the AES-256-GCM (NoPadding) encryption algorithm is supported. This parameter is used for encryption at the migration destination and decryption at the migration source. For details, see **Encryption and Decryption During File Migration**.	KMS

 Copyright © Huawei Technologies Co., Ltd.

Category	Parameter	Description	Example Value
	Key ID	Data encryption key. This parameter is displayed when **Encryption** is set to **KMS**. Click next to the text box to select the KMS key that has been created in DEW. ● If the KMS key of the same project as that of the CDM cluster is used, you do not need to modify **Project ID**. ● If the KMS key of another project is used, you need to modify **Project ID**.	53440ccb-3e73-4700-98b5-71ff5476e621
	Project ID	ID of the project to which KMS ID belongs. The default value is the ID of the project to which the current CDM cluster belongs. ● If KMS and the CDM cluster are in the same project, retain the default value of **Project ID**. ● If KMS of another project is used, set this parameter to the ID of the project to which KMS belongs.	9bd7c4bd54e5417198f9591bef07ae67
	DEK	This parameter is displayed only when **Encryption** is set to **AES-256-GCM**. The key consists of 64 hexadecimal numbers. Remember the key configured here because the decryption key must be the same as that configured here. If the encryption and decryption keys are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B

 Copyright © Huawei Technologies Co., Ltd.

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

Copyright © Huawei Technologies Co., Ltd.

Category	Parameter	Description	Example Value
	Validate MD5 Value	The MD5 value can be verified only when files are transferred in **Binary** format. KMS encryption cannot be used when 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 verify 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 is to be written	obslink
	Record MD5 Bucket	OBS bucket to which the MD5 verification result is to be written	cdm05
	Record MD5 Directory	Directory to which the MD5 verification result is to be written	/md5/
	Encoding Type	Encoding type, for example, **UTF-8** or **GBK**. This parameter is invalid when **File Format** is set to **Binary**.	GBK

Copyright © Huawei Technologies Co., Ltd.

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 **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

 Copyright © Huawei Technologies Co., Ltd.

Category	Parameter	Description	Example Value
	Customize File Name	This parameter is displayed only when data is exported from a relational database to OBS and **File Format** is set to **CSV**. This parameter specifies the name of the file generated by OBS. The options are as follows: ● **Character string**: Special characters are allowed. For example, if this parameter is set to **cdm#**, the name of the generated file is **cdm#.csv**. ● Macro variable of time: If this parameter is set to **${timestamp()}**, the name of the generated file is **1554108737.csv**. ● Macro variable of table name: If this parameter is set to **${tableName}**, the name of the generated file is **sqltabname.csv**. ● Macro variable of version number: If this parameter is set to **${version}**, the name of the generated file is **v1.csv**. ● Any combination of the character string and macro variable (macro variable of time, table name, or version number). For example, if this parameter is set to **cdm#${timestamp()}_${version}**, the name of the generated file is **cdm#1554108737_v1.csv**.	cdm
	Customize Hierarchical Directory	If this parameter is enabled, the directory hierarchy of user-defined files is used.	N/A
	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.	${dateformat(yyyy/MM/dd, -1, DAY)}

 Copyright © Huawei Technologies Co., Ltd.

6.1.6.2 Destination HDFS Parameters

When the destination link of a job is the **Link to HDFS**, that is, when data is imported to the following data sources, configure the destination job parameters based on **Table 6-53**.

Table 6-53 Parameters

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

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Compression Format	File compression format after data writing. The following compression formats are supported: • **None**: Do not compress the files. • **DEFLATE**: Compress the files in DEFLATE format. • **gzip**: Compress the files in gzip format. • **bzip2**: Compress the files in bzip2 format. • **LZ4**: Compress the files in LZ4 format. • **Snappy**: Compress the files in Snappy format.	Snappy
Line Separator	Lind feed character in a file. By default, the system automatically identifies **\n**, **\r**, and **\r\n**. This parameter is not used when **File Format** is set to **Binary**.	\n
Field Delimiter	Field delimiter in the file. This parameter is not used when **File Format** is set to **Binary**.	,
Use Quote Character	This parameter is displayed only when **File Format** is **CSV**. It is used when database tables are migrated to file systems. If you set this parameter to **Yes** and a field in the source data table contains a field delimiter or line separator, CDM uses double quotation marks (") as the quote character to quote the field content as a whole to prevent a field delimiter from dividing a field into two fields, or a line separator from dividing a field into different lines. For example, if the **hello,world** field in the database is quoted, it will be exported to the CSV file as a whole.	No
Use First Row as Header	This parameter is displayed only when **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

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
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
Encryption	This parameter is displayed only when **File Format** is set to **Binary**. Whether to encrypt the uploaded data. The options are as follows: • **None**: Data is written without encryption. • **AES-256-GCM**: The AES 256-bit encryption algorithm is used to encrypt data. Currently, only the AES-256-GCM (NoPadding) encryption algorithm is supported. This parameter is used for encryption at the migration destination and decryption at the migration source. For details, see **Encryption and Decryption During File Migration**.	AES-256-GCM
DEK	This parameter is displayed only when **Encryption** is set to **AES-256-GCM**. The key consists of 64 hexadecimal numbers. Remember the key configured here because the decryption key must be the same as that configured here. If the encryption and decryption keys are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B
IV	This parameter is displayed only when **Encryption** is set to **AES-256-GCM**. The initialization vector consists of 32 hexadecimal numbers. Remember the initialization vector configured here because the initialization vector used for decryption must be the same as that configured here. If the initialization vectors are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	5C91687BA886EDCD12ACBC3FF19A3C3F

 Copyright © Huawei Technologies Co., Ltd.

NOTE

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

6.1.6.3 Destination HBase/CloudTable Parameters

When the destination link of a job is the **Link to HBase** or **Link to CloudTable**, that is, when data is imported to the following data sources, configure the destination job parameters based on **Table 6-54**.

- MRS HBase
- FusionInsight HBase
- Apache HBase
- CloudTable

Table 6-54 Parameters

Parameter	Description	Example Value
Table Name	Name of the HBase table to which data is to be written. If you want to create an HBase table, you can copy the field names from the migration source. Click the icon next to the text box. The dialog box for selecting the table is displayed. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**.	TBL_2
Clear Data Before Import	Whether the data in the destination table is cleared before data import. The options are as follows: • **Yes**: The data is cleared. • **No**: The data is not cleared. Instead, it will be added to the existing table.	Yes
Rowkey Delimiter	(Optional) Used to combine multiple rows into 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

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Compression Format	(Optional) Compression format used in creating an HBase table. The default value is **None**. • **None**: Do not compress the files. • **Snappy**: Compress the files in Snappy format. • **gzip**: Compress the files in gzip format.	None
Write WAL	Whether to enable Write Ahead Log (WAL) of HBase. The options are as follows: • **Yes**: If the HBase server breaks down after the function is enabled, you can replay the operations that have not been performed in WAL. • **No**: If you set this parameter to **No**, the write performance is improved. However, if the HBase server breaks down, data may be lost.	No
Match Data Type	• **Yes**: Data of the Short, Int, Long, Float, Double, and Decimal columns in the source database is converted into Byte[] arrays (binary) and written into HBase. Other types of data are written as character strings. If several types of data mentioned above are combined as rowkeys, they will be written as character strings. This function saves storage space. In specific scenarios, the rowkey distribution is evener. • **No**: All types of data in the source database are written into HBase as character strings.	No

6.1.6.4 Destination Hive Parameters

If the destination link of a job is the **Link to Hive**, configure the destination job parameters based on **Table 6-55**.

Copyright © Huawei Technologies Co., Ltd.

Table 6-55 Parameter description

Parameter	Description	Example Value
Database Name	Database name. Click the icon next to the text box. The dialog box for selecting the database is displayed.	default
Auto Table Creation	This parameter is displayed only when both the migration source and destination are relational databases. The options are as follows: • **Non-auto creation**: CDM will not automatically create a table. • **Auto creation**: If the destination database does not contain the table specified by **Table Name**, CDM will automatically create the table. If the table specified by **Table Name** already exists, no table is created and data is written to the existing table. • **Deletion before creation**: CDM deletes the table specified by **Table Name**, and then creates the table again.	Non-auto creation
Table Name	Destination table name. Click the icon next to the text box. The dialog box for selecting the table is displayed. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**.	TBL_X
Clear Data Before Import	Processing mode of the data in the target table. The options are as follows: • **Yes**: Clear the data in the target table before data import. • **No**: Do not clear the data in the target table before data import. If you select **No** and the table contains data, then the data will be added to the existing table.	Yes

 Copyright © Huawei Technologies Co., Ltd.

NOTE

1. When Hive serves as the migration destination, the storage format selected during table creation will be automatically used, such as ORC and Parquet.
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;
```

6.1.6.5 Destination FTP/SFTP/NAS/SFS Parameters

If the destination link of a job is the **Link to an FTP or SFTP Server** or **Link to NAS/SFS**, configure the destination job parameters based on **Table 6-56**.

Copyright © Huawei Technologies Co., Ltd.

Table 6-56 Parameter description

Category	Parameter	Description	Example Value
Basic parameters	Write Directory	Directory to which data is to be written. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**.	/opt/ftp/
	File Format	Format in which data is written. The options are as follows: • **CSV**: Data is written in CSV format, which is used for migrating data tables to files. • **Binary**: Files will be transferred directly. CDM writes the files without changing their format. This setting is suitable for file migration. If data is migrated between file-related data sources, such as FTP, SFTP, NAS, HDFS, and OBS, the value of **File Format** must the same as the source file format.	CSV
	Duplicate File Processing Method	Files with the same name and size are identified as duplicate files. If there are duplicate files during data writing, the following methods are available: • Replace • Skip • Stop job	Skip

 Copyright © Huawei Technologies Co., Ltd.

Category	Parameter	Description	Example Value
Advanced attributes	Line Separator	Lind feed character in a file. By default, the system automatically identifies **\n**, **\r**, and **\r\n**. This parameter is 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**.	,
	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
	Encoding Type	Encoding type, for example, **UTF-8** or **GBK**. This parameter is not used when **File Format** is set to **Binary**.	GBK
	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
	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

 Copyright © Huawei Technologies Co., Ltd.

Category	Parameter	Description	Example Value
	Generate MD5 Hash Value	This parameter is displayed only when the migration source is a file system (OBS/FTP/SFTP/NAS/HDFS), the migration destination is FTP/SFP/NAS, and **File Format** is **Binary**. 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.	No
	Directory of MD5 Hash Value	Directory for storing MD5 values	/md5
	Use First Row as Header	This parameter is displayed only when **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

 Copyright © Huawei Technologies Co., Ltd.

Category	Parameter	Description	Example Value
	Encryption	Whether to encrypt the uploaded data. The options are as follows: • **None**: Data is written without encryption. • **AES-256-GCM**: The AES 256-bit encryption algorithm is used to encrypt data. Currently, only the AES-256-GCM (NoPadding) encryption algorithm is supported. This parameter is used for encryption at the migration destination and decryption at the migration source. For details, see **Encryption and Decryption During File Migration**.	AES-256-GCM
	DEK	This parameter is displayed only when **Encryption** is set to **AES-256-GCM**. The key consists of 64 hexadecimal numbers. Remember the key configured here because the decryption key must be the same as that configured here. If the encryption and decryption keys are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B

Copyright © Huawei Technologies Co., Ltd.

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. Remember the initialization vector configured here because the initialization vector used for decryption must be the same as that configured here. If the initialization vectors are inconsistent, the system does not report an exception, but the decrypted data is incorrect.	5C91687BA886EDCD12ACBC3FF19A3C3F

6.1.6.6 Destination Relational Database Parameters

When the destination link of a job is the **Link to Relational Databases**, that is, when data is imported to the following data sources, configure the destination job parameters based on **Table 6-57**.

- Data Warehouse Service
- RDS for MySQL
- RDS for SQL Server
- RDS for PostgreSQL
- FusionInsight LibrA
- MySQL

Table 6-57 Parameters

Parameter	Description	Example Value
Schema/ Tablespace	Name of the database to which data is to be written. The schema can be automatically created. Click the icon next to the text box to select a schema or tablespace.	Schema

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Auto Table Creation	This parameter is displayed only when both the migration source and destination are relational databases. The options are as follows: • **Non-auto creation**: CDM will not automatically create a table. • **Auto creation**: If the destination database does not contain the table specified by **Table Name**, CDM will automatically create the table. If the table specified by **Table Name** already exists, no table is created and data is written to the existing table. • **Deletion before creation**: CDM deletes the table specified by **Table Name**, and then creates the table again. **Field Mapping in Automatic Table Creation on DWS** describes the field mapping between the DWS tables created by CDM and source tables.	Non-auto creation
Table Name	Name of the table to which data is to be written. Click the icon next to the text box. The dialog box for selecting the table is displayed. This parameter can be configured as a macro variable of date and time and a path name can contain multiple macro variables. When the macro variable of date and time works with a scheduled job, the incremental data can be synchronized periodically. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**.	table
Compress Data	Whether to compress data when data is imported to DWS and **Auto creation** is selected	No

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Storage Mode	When data is imported to DWS and **Auto Creation** is selected, you can specify the data storage mode: • **Row-based**: Row-based storage. It is applicable to 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 applicable to statistical analysis queries (group and join scenarios) or ad hoc queries (query condition columns and row store indexes are uncertain).	Row-based
Clear Data Before Import	Whether to clear the data in the target table before data import. The options are as follows: • **Do not clear**: Do not clear the data in the target table before data import but add data to the table. • **Clear all data**: Clear all data in the target table before data import. • **Clear part of data**: Clear part of data in the target table before data import. If you select **Clear part of data**, you must configure **WHERE Clause** to specify the data to be deleted from the target table.	Clear part of data
WHERE Clause	This parameter is displayed when **Clear Data Before Import** is set to **Clear all data**. This parameter indicates that the data in the destination table is deleted based on the where condition before the import.	age > 18 and age <= 60

Copyright © Huawei Technologies Co., Ltd.

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

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Use NOT NULL Constraint	If you choose to create a target table automatically and specify the NOT NULL constraint, keep the NOT NULL constraints of the source and target tables consistent.	Yes

Field Mapping in Automatic Table Creation on DWS

Figure 6-8 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 6-8 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

6.1.6.7 Destination DDS Parameters

If the destination link of a job is the **Link to MongoDB**, configure the destination job parameters based on **Table 6-58**.

Table 6-58 Parameter description

Parameter	Description	Example Value
Database Name	Database to which data is to be imported	mongodb

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
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

6.1.6.8 Destination DCS Parameters

If the destination link of a job is the **Link to Redis/DCS**, that is, when data is imported to DCS, configure the destination job parameters based on **Table 6-59**.

Table 6-59 Parameter description

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

6.1.6.9 Destination Elasticsearch or CSS Parameters

If the destination link of a job is the **Link to Elasticsearch/CSS**, that is, when data is imported to Elasticsearch or CSS, configure the destination job parameters based on **Table 6-60**.

Copyright © Huawei Technologies Co., Ltd.

Table 6-60 Parameter description

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

6.1.6.10 Destination DLI Parameters

If the destination link of a job is the **Link to DLI**, configure the destination job parameters based on **Table 6-61**.

 Copyright © Huawei Technologies Co., Ltd.

Table 6-61 Parameter description

Parameter	Description	Example Value
Resource Queue	Resource queue to which the destination table belongs	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	No

6.1.6.11 Destination DIS Parameters

If the destination link of a job is the **Configuring Links to DIS**, configure the destination job parameters based on **Table 6-62**.

Table 6-62 Parameter description

Parameter	Description	Example Value
DIS Stream	DIS stream name	cdm
Field Delimiter	The default value is space. To set the **Tab** key as the delimiter, set this parameter to **\t**.	,
Reference Sign	Delimiter between the names of the referenced tables or columns. This parameter is left blank by default.	'

6.1.6.12 Destination OpenTSDB Parameters

If the destination link of a job is the **Link to CloudTable OpenTSDB**, configure the destination job parameters based on **Table 6-63**.

Table 6-63 Parameter description

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

Copyright © Huawei Technologies Co., Ltd.

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

6.1.7 Entire DB Migration

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 or each collection of MongoDB can be executed concurrently as a subtask.

Table 6-64 lists the data sources supporting entire DB migration using Cloud Data Migration.

Table 6-64 Supported data sources in entire DB migration

Source Data Type	Destination Data Type					
	RDS for MySQL	MRS(Hive)	DWS	CSS	OBS	CloudTable
MySQL	√	√	√	×	√	×
PostgreSQL	√	√	√	×	√	×
Microsoft SQL Server	√	√	√	×	×	×
Oracle	√	√	√	×	√	×
Elasticsearch	×	×	×	√	×	×
MongoDB	×	×	×	×	×	×
HBase	×	×	×	×	×	√
IBM Db2	√	√	√	×	√	×
Derecho(Gauss DB)	√	√	√	×	√	×
SAP HANA	√	√	√	×	√	×
DWS	√	√	√	×	×	×
Hive	√	×	√	×	×	×

Copyright © Huawei Technologies Co., Ltd.

The source databases can be deployed in on-premises data centers or built on ECSs, or third-party database services.

Field Mapping in Automatic Table Creation

CDM automatically creates tables at the destination during database migration. Figure 6-9 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 6-9 Field mapping in automatic table creation on DWS 1

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

Entire DB Migration

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Integration** of the corresponding workspace.

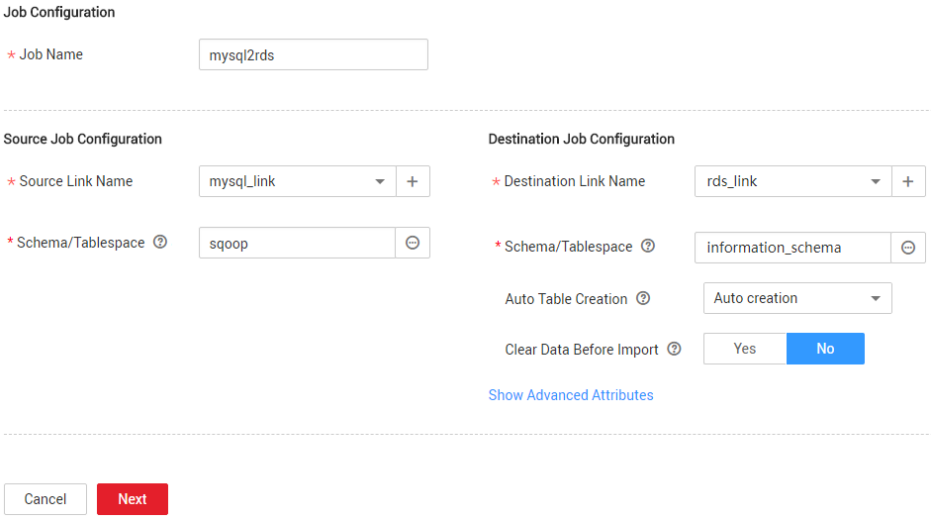
   The **Data Integration** page is displayed.

3. Choose **Cloud Data Migration**. The **Cluster Management** page is displayed. Click **Job Management** in the **Operation** column of the cluster.

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

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-10 Creating an entire DB migration job



5. Configure the related parameters of the source database according to **Table 6-65**.

Table 6-65 Parameter description

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

 Copyright © Huawei Technologies Co., Ltd.

Source Database	Parameter	Description	Example Value
	WHERE Clause	WHERE clause used to specify the tables to be extracted. This parameter applies to all subtables in the entire DB migration. If this parameter is not set, the entire table is extracted. If the table to be migrated does not contain the fields specified by the WHERE clause, the migration will fail. This parameter can be configured as a macro variable of date and time to extract data generated at a specific date. For details, see **WHERE Clause**.	age > 18 and age <= 60
Elasticsearch	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 a time, **Index** at the migration destination cannot be configured.	cdm*
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

Copyright © Huawei Technologies Co., Ltd.

Source Database	Parameter	Description	Example Value
	Filter Condition	Advanced attribute parameter. This parameter is used to create a filter for matching documents.	{HTTPStatusCode:{$gt:"400",$lt:"500"},HTTPMethod:"GET"}

6. Configure the related parameters of the destination cloud service according to **Table 6-66**.

Table 6-66 Parameter description

Cloud Service	Parameter	Description
• MRS Hive • RDS for MySQL	Schema/ Tablespace	Database name
	Auto Table Creation	The options are as follows: • **Non-auto creation**: CDM will not automatically create a table. • **Auto creation**: If no corresponding table exists in the destination database, CDM will automatically create one. • **Deletion before creation**: If a table with the same name exists in the destination database, CDM will delete the table first and create another one with the same name.

 Copyright © Huawei Technologies Co., Ltd.

Cloud Service	Parameter	Description
	Clear Data Before Import	Whether to clear the data in the target table before data import. The options are as follows: • **Do not clear**: Do not clear the data in the target table before data import but add data to the table. • **Clear all data**: Clear all data in the target table before data import. • **Clear part of data**: Clear part of data in the target table before data import. If you select **Clear part of data**, you must configure **WHERE Clause** to specify the data to be deleted from the target table.
	WHERE Clause	This parameter is displayed when **Clear Data Before Import** is set to **Clear all data**. Used to specify the data to be deleted from the target table before data import, for example, **age > 18 and age <= 60**
	Extend Field Length	Advanced attribute parameter. When a destination table is automatically created, the length of the character fields can be extended to three times the original length. 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.
	Use NOT NULL Constraint	Advanced attribute parameter. 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.

 Copyright © Huawei Technologies Co., Ltd.

Cloud Service	Parameter	Description
CSS	Index	Index to which data is written. 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	Whether to clear data of the target type before data is written
DWS	N/A	For details about the destination job parameters required for entire DB migration to DWS, see **Destination Relational Database Parameters**.
OBS	N/A	For details about the destination job parameters required for entire DB migration to OBS, see **Destination OBS Parameters**.

7. 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.
8. Click **Next**, set job parameters, and click **Show Advanced Attributes** to display and configure optional parameters.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-11 Task parameters 1

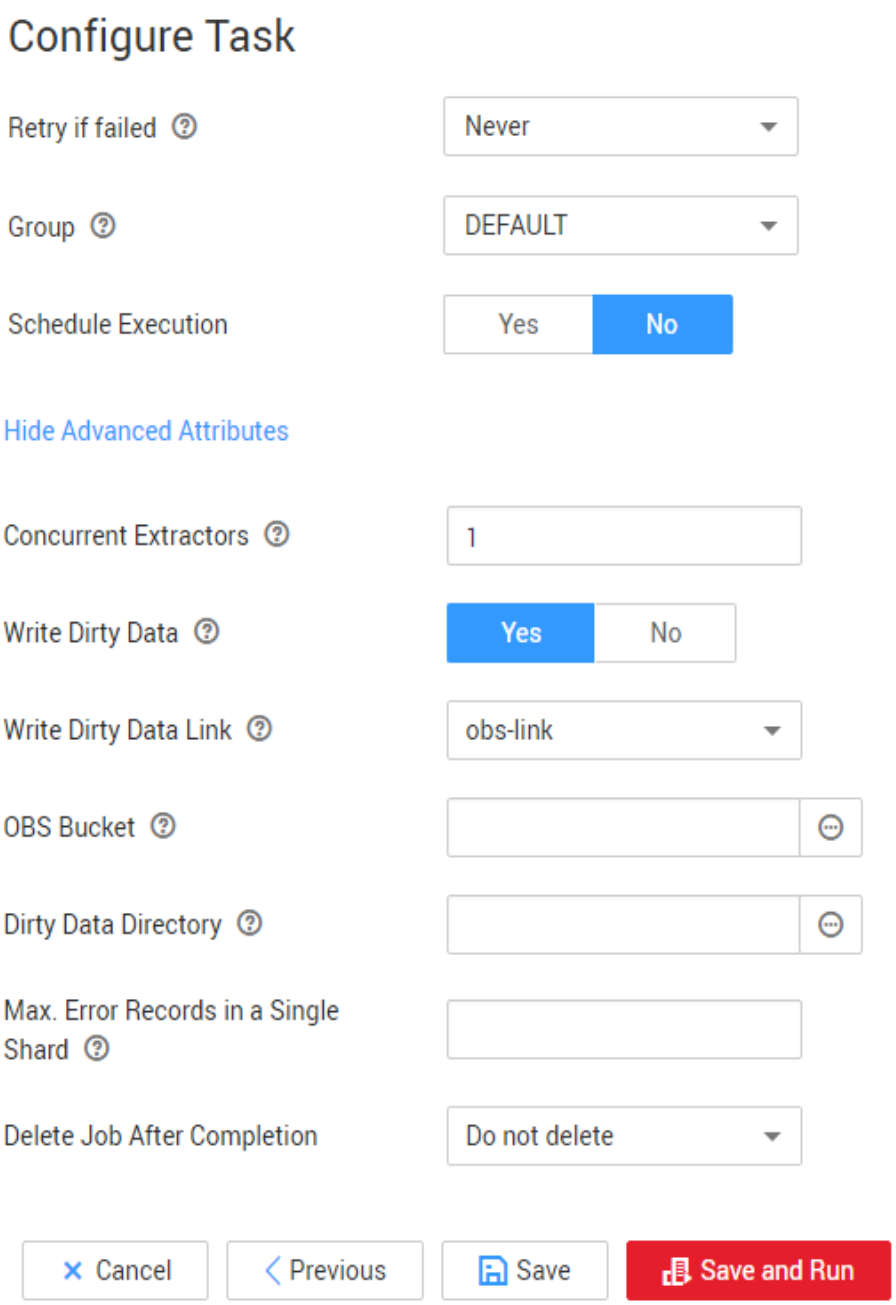


Table 6-67 describes related parameters.

 Copyright © Huawei Technologies Co., Ltd.

Table 6-67 Parameter description

Parameter	Description	Example Value
Retry upon Failure	You can select **Retry 3 times** or **Never**. You are advised to configure automatic retry for only file migration jobs or database migration jobs with **Import to Staging Table** enabled to avoid data inconsistency caused by repeated data writes.	Never
Job	Select a group where the job resides. The default group is **DEFAULT**. On the **Job Management** page, jobs can be displayed, started, or exported by group.	DEFAULT
Schedule Execution	If you select **Yes**, you can set the start time, cycle, and validity period of a job. For details, see **Scheduling Job Execution**.	No
Concurrent Extractors	Number of extractors to be concurrently executed. Generally, retain the default value.	1
Concurrent Loaders	Number of Loaders to be concurrently executed This parameter is displayed only when HBase or Hive serves as the destination data source.	3
Write Dirty Data	Whether to record dirty data. By default, this parameter is set to **No**.	Yes
Write Dirty Data Link	This parameter is displayed only when **Write Dirty Data** is set to **Yes**. Only links to OBS support dirty data writes.	obs_link
OBS Bucket	This parameter is displayed only when **Write Dirty Data Link** is a link to OBS. Name of the OBS bucket to which the dirty data is to be written.	dirtydata

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Dirty Data Directory	This parameter is displayed only when **Write Dirty Data** is set to **Yes**. Dirty data is stored in the directory for storing dirty data on OBS. Dirty data is saved only when this parameter is configured. You can go to this directory to query data that fails to be processed or is filtered out during job execution, and check the source data that does not meet conversion or cleaning rules.	/user/dirtydir
Max. Error Records in a Single Shard	This parameter is displayed only when **Write Dirty Data** is set to **Yes**. When the number of error records of a single map exceeds the upper limit, the job will automatically terminate and the imported data cannot be rolled back. You are advised to use a temporary table as the destination table. After the data is imported, rename the table or combine it into the final data table.	0
Delete Job After Completion	Perform the following operations after a job is executed: • **Do not delete**: The job is not deleted after it is executed. • **Delete after success**: The job is deleted only when the job is successfully executed. It is applicable to massive one-time jobs. • **Delete**: The job is deleted regardless of whether it is successfully executed or fails to be executed.	Do not delete

9. 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.

6.1.8 Scenario-based Migration

Scenario-based migration migrates snapshots and then restores table data to speed up migration.

Prerequisites

- The CDM cluster can communicate with the data source.

 Copyright © Huawei Technologies Co., Ltd.

- 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.

Link to Hadoop

CDM supports the following Hadoop data sources:

MRS

When connecting CDM to Hadoop of MRS, configure the parameters as described in **Table 6-68**.

Table 6-68 MRS Hadoop 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_scen_link
Manager IP	IP address of MRS Manager. Click **Select** next to the **Manager IP** text box to select an MRS cluster. CDM automatically fills in the authentication information.	127.0.0.1
Authentication Method	Authentication method used for accessing MRS • **SIMPLE**: for non-security mode • **KERBEROS**: for security mode	SIMPLE
HBase Version	Set it to the HBase version on the server.	HBASE_2_X
HIVE Version	Set it to the Hive version on the server.	HIVE_3_X
Username	If **Authentication Method** is set to **KERBEROS**, you must provide the username and password used for logging in to MRS Manager. If you need to create a snapshot when exporting a directory from HDFS, the user configured here must have the administrator permission on HDFS.	cdm
Password	Password for logging in to MRS Manager	-

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Run Mode	Run mode of the HDFS link. The options are as follows: • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **STANDALONE**: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, CloudTable, or Hadoop) with both Kerberos and Simple authentication modes, select **STANDALONE** or configure different agents. • **Agent**: The link instance runs on an agent. If **STANDALONE** is selected, CDM can migrate data between HDFSs of multiple MRS clusters.	STANDALONE
Use Cluster Config	You can create cluster configurations on the **Links** page to simplify the configuration of Hadoop link parameters. When creating a link, set this parameter to **Yes**.	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

FusionInsight Hadoop

When connecting CDM to Hadoop of FusionInsight HD, configure the parameters as described in **Table 6-69**.

Table 6-69 FusionInsight Hadoop 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

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
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 FusionInsight HD • **SIMPLE**: for non-security mode • **KERBEROS**: for security mode	KERBEROS
HBase Version	Set it to the HBase version on the server.	HBASE_2_X
HIVE Version	Set it to the Hive version on the server.	HIVE_3_X
Run Mode	Run mode of the HDFS link. The options are as follows: • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **STANDALONE**: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, CloudTable, or Hadoop) 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 data 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

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Use Cluster Config	You can create cluster configurations on the **Links** page to simplify the configuration of Hadoop link parameters. When creating a link, set this parameter to **Yes**.	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

Apache Hadoop

When connecting CDM to Apache Hadoop, configure parameters as described in **Table 6-70**.

Table 6-70 Apache Hadoop 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	hdfs://nn1.example.com/
ZooKeeper Address	ZooKeeper address, which needs to be configured for HBase scenario-based migration	hbase-node-1:2181
Hive Metastore	Hive metadata address. For details, see the **hive.metastore.uris** configuration item.	thrift://host-192-168-1-212:9083
Authentication Method	Authentication method used for accessing Hadoop • **SIMPLE**: Select this if Hadoop is in non-security mode. • **KERBEROS**: Select this if Hadoop is in security mode. Obtain the **Principal** account and **Keytab File** file of the client for authentication.	KERBEROS

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Principal	When **Authentication Method** is set to **KERBEROS**, the **Principal** account is used for authentication. You can contact the Hadoop administrator to obtain the account.	USER@YOUR-REALM.COM
Keytab File	When **Authentication Method** is set to **KERBEROS**, this file is used for authentication. You can contact the Hadoop administrator to obtain the file.	/opt/user.keytab
IP and Host Name Mapping	If the HDFS configuration file uses the host name, configure the mapping between the IP address and host name. Separate the IP addresses and host names by spaces and mappings by semicolons (;), carriage returns, or line feeds.	10.1.6.9 hostname01 10.2.7.9 hostname02
HBase Version	Set it to the HBase version on the server.	HBASE_2_X
HIVE Version	Set it to the Hive version on the server.	HIVE_3_X
Run Mode	Run mode of the HDFS link. The options are as follows: ● **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. ● **STANDALONE**: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, CloudTable, or Hadoop) 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 data 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
Use Cluster Config	You can create cluster configurations on the **Links** page to simplify the configuration of Hadoop link parameters. When creating a link, set this parameter to **Yes**.	No

 Copyright © Huawei Technologies Co., Ltd.

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.	hadoop_01

Procedure

Step 1 Log in to the **CDM management console**.

Step 2 In the left navigation pane, click **Cluster Management**. Locate the target cluster and click **Job Management**.

Step 3 Choose **Job Management** > **Link Management** > **Create Link** and set the connector type to **Hadoop release version**.

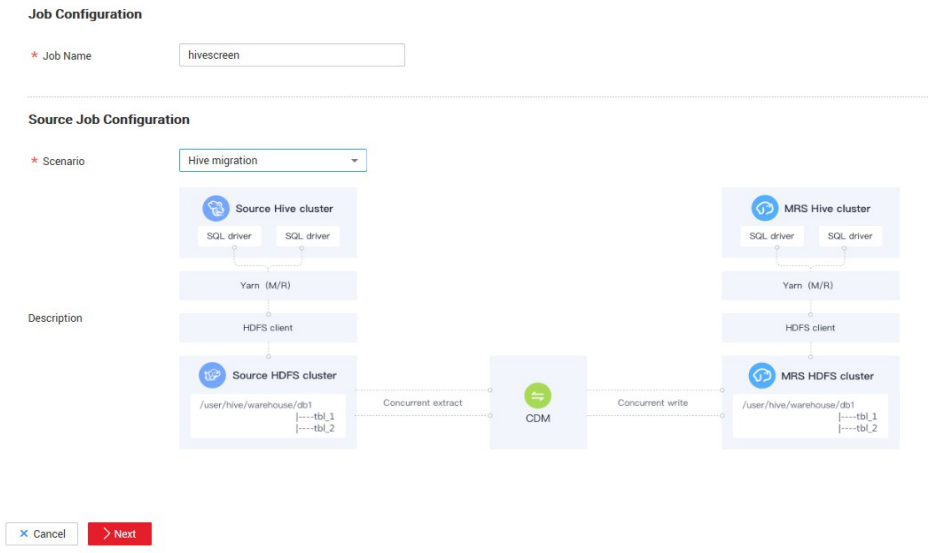
Step 4 Click **Next**. Set link parameters by referring to **Link to Hadoop**.

Step 5 Click **Test** to check whether the link is available. Alternatively, click **Save**. The system will automatically check whether the link is available.

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

Step 6 Choose **Scenario Migration** > **Create Job**. The page for configuring the job is displayed. Select a migration scenario (Hadoop migration, Hive migration, or HBase migration) and configure the job name.

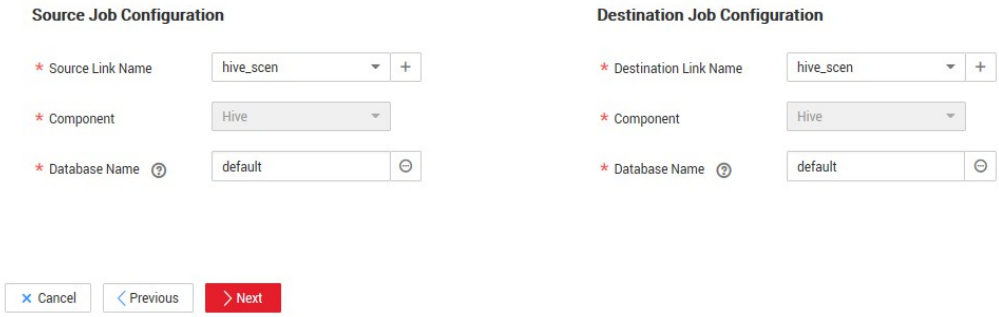
Figure 6-12 Configuring a scenario-based migration Job



Step 7 Configure the source and destination job parameters, and select the link name and name of the database to be migrated.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-13 Configuring job parameters



Step 8 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 9 Click **Next** and set job parameters.

Table 6-71 describes related parameters.

Table 6-71 Task configuration parameters

Parameter	Description	Example Value
Write Dirty Data	Whether to record dirty data. By default, this parameter is set to **No**.	Yes
Write Dirty Data Link	This parameter is displayed only when **Write Dirty Data** is set to **Yes**. Only links to OBS support dirty data writes.	obs_link
OBS Bucket	This parameter is displayed only when **Write Dirty Data Link** is a link to OBS. Name of the OBS bucket to which the dirty data will be written.	dirtydata
Dirty Data Directory	This parameter is displayed only 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

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Max. Error Records in a Single Shard	This parameter is displayed only when **Write Dirty Data** is set to **Yes**. When the number of error records of a single map exceeds the upper limit, the job will automatically terminate and the imported data cannot be rolled back. You are advised to use a temporary table as the destination table. After the data is imported, rename the table or combine it into the final data table.	0

Step 10 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

6.1.9 Incremental Migration

6.1.9.1 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 all new files**
   - Application scenarios: Both the migration source and destination are file systems (OBS/OSS/HDFS/FTP/SFTP/NAS).
   - Key configurations: **Skipping a Duplicate File** and **Schedule Execution**
   - Prerequisites: None
2. **Exporting the files in a specified directory**
   - Application scenarios: The migration source is a file system (OBS/OSS/HDFS/FTP/SFTP/NAS). The migration destination can be of any type. 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.
3. **Exporting the files modified after the specified time point**
   - Application scenarios: The migration source is a file system (OBS/OSS/HDFS/FTP/SFTP/NAS). The migration destination can be of any type. 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

Copyright © Huawei Technologies Co., Ltd.

Skipping a Duplicate File

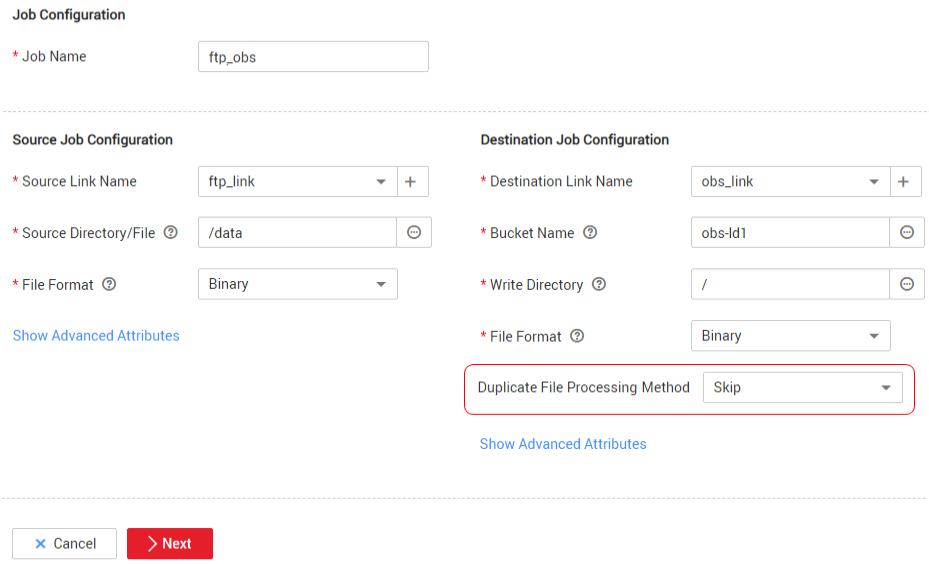
- Parameter position: When creating a table/file migration job as shown in **Table/File Migration Jobs**, if the migration source and destination are file systems, set **Duplicate File Processing Method** in **Destination Job Configuration** to **Replace**, **Skip**, or **Stop job**.
- Parameter principle: If a file with the same name and size exists on the migration source and destination, CDM determines that the file is a duplicate file.
- Example configurations:

  a. **Source Directory/File**: If you set this parameter to a directory, CDM imports all files in the directory to the migration destination.

  b. **File Format**: Select **Binary**. CDM directly copies the files without resolving the content, which is applicable to the migration of files to files.

  c. **Duplicate File Processing Method**: Select **Skip**.

  Figure 6-14 Skipping duplicate files



  d. Configure scheduled job execution.

In this way, you can import the newly added files to the destination directory periodically to implement incremental synchronization.

File/Path Filter

- Parameter position: When creating a table/file migration job as shown in **Table/File Migration Jobs**, if the migration source is a file system, set **Filter Type** in advanced attributes of **Source Job Configuration** to **Wildcard** or **Time filter**.
- Parameter principle: If you set **Filter Type** to **Wildcard**, CDM filters files or paths based on the configured wildcard character and migrates only files or paths that meet the specified condition.
- Example configurations:

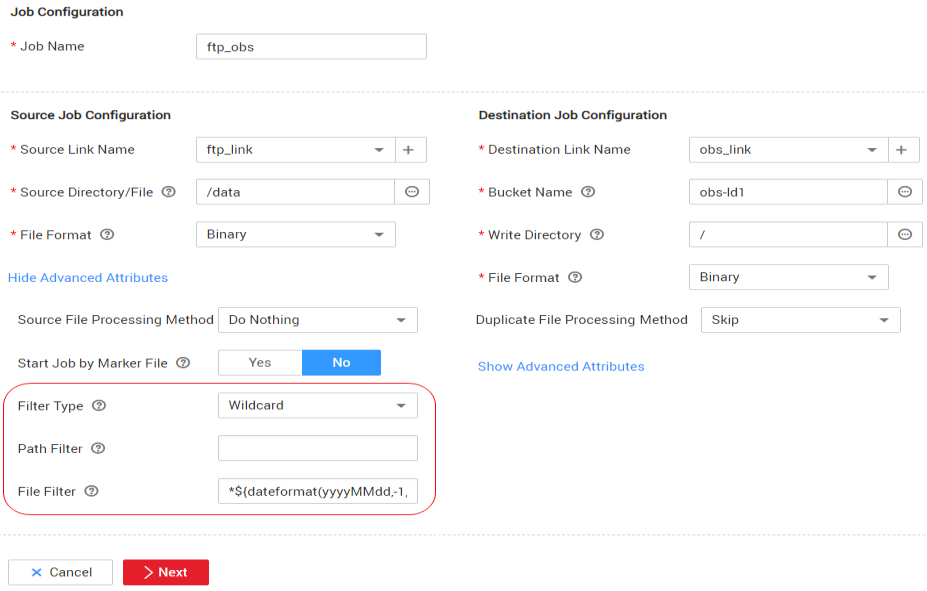
 Copyright © Huawei Technologies Co., Ltd.

Suppose that the source file name contains the date and time field, such as **2017-10-15 20:25:26**, the **/opt/data/file_20171015202526.data** file is generated. Set the parameters as follows:

a. **Filter Type**: Select **Wildcard**.

b. **File Filter**: Enter ***${dateformat(yyyyMMdd,-1,DAY)}***, which is the format of the macro variables of date and time supported by CDM. For details, see **Incremental Synchronization Using the Macro Variables of Date and Time**.

Figure 6-15 Filtering files



c. Schedule Execution: Set **Cycle (days)** to **1**.

In this way, you can import the files generated in the previous day to the destination directory every day to implement incremental synchronization.

In incremental file migration, **Path Filter** is used in the same way as **File Filter**. The path name must contain the time field. In this case, all files in the specified path can be synchronized periodically.

Time Filter

- Parameter position: When creating a table/file migration job as shown in **Table/File Migration Jobs**, if the migration source is a file system, set **Filter Type** in **Source Job Configuration** to **Time Filter**.
- Parameter principle: If you specify **Modification Time**, only the files whose modification time is later than the specified time are migrated to CDM.
- Example configurations:

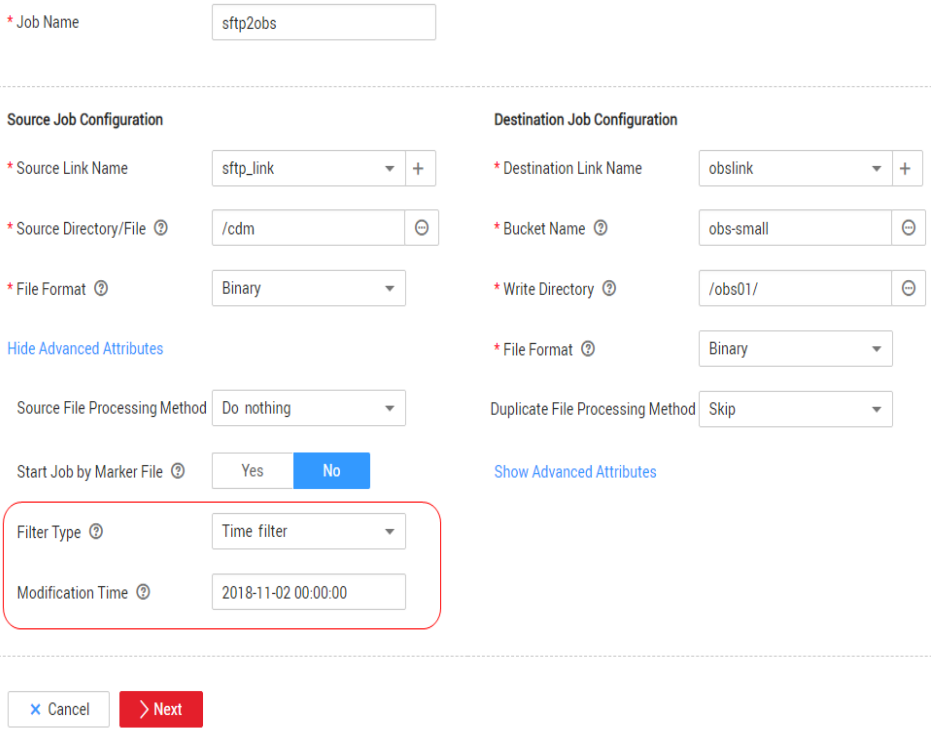
  Suppose that you want CDM to synchronize only the files generated after November 2, 2018 to the migration destination, configure the following parameters:

  a. **Filter Type**: Select **Time Filter**.

  b. **Modification Time**: Enter a value in the format of *yyyy-MM-dd HH:mm:ss*, such as **2018-11-02 00:00:00**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-16 Time Filter



c. **Duplicate File Processing Method**: Select **Skip**.

d. Configure scheduled job execution.

In this way, the CDM job migrates only files generated after November 2, 2018, and performs incremental synchronization next time it is started.

6.1.9.2 Incremental Migration of Relational Databases

CDM supports incremental migration of relational databases. After full migration is complete, you can migrate data generated in a specified time range incrementally (for example, migrate data added on the previous day at 00:00:00 every day).

- **Migrating incremental data within a specified period of time**
  - Application scenarios: The source end is a relational database. For details, see **From 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.

For incremental migration of relational databases, only the specified data is written to the data tables. The existing records are not updated or deleted.

WHERE Clause

- Parameter position: When creating a table/file migration job as shown in **Table/File Migration Jobs**, if the source end is a relational database, the **Where Clause** parameter is available in the advanced attributes of **Source Job Configuration**.

 Copyright © Huawei Technologies Co., Ltd.

- 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 6-17**. Set the parameters as follows:

  Figure 6-17 Table data 1

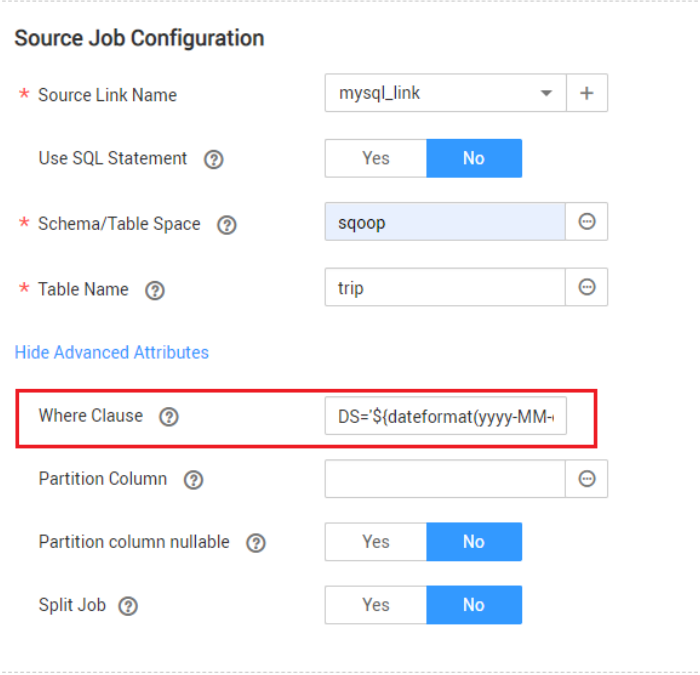
  SELECT * FROM SQOOP.CDM_20171016

	FOO	BAR	DS
1	5	snap	2017-05-01
2	5	snap	2017-05-01
3	1	google	2017-05-02
4	4	oracle	2017-05-02
5	6	amd	2017-05-02
6	7	nvda	2017-05-02
7	1	google	2017-05-02
8	4	oracle	2017-05-02
9	6	amd	2017-05-02
10	7	nvda	2017-05-02
11	2	facebook	2017-10-15
12	3	tesla	2017-10-15
13	2	facebook	2017-10-15
14	3	tesla	2017-10-15

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

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-18 Where clause



b. **Scheduling Job Execution**: Set **Cycle (days)** to **1** and **Start Time** to **00:00:00**.

In this way, all data generated on the previous day can be exported at 00:00:00 every day. **WHERE Clause** can be configured to various **macro variables of date and time**. You can use the macro variables of date and time and scheduled jobs with specified cycle of minutes, hours, days, weeks, or months together to automatically export data at a specific time.

6.1.9.3 Incremental Synchronization Using the Macro Variables of Date and Time

When creating a table/file migration job as shown in **Table/File Migration Jobs**, CDM allows you to configure time macro variables in the following parameters on the source and destination ends:

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

You can use the **${}** macro variable definition identifier to define the macros of the time type. currently, dateformat and timestamp are supported.

By using the macro variables of date and time and **scheduled job**, you can implement incremental synchronization of databases and files.

dateformat

dateformat supports two types of parameters:

 Copyright © Huawei Technologies Co., Ltd.

- **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**, **dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)** indicates the day before the current day, that is, **2017-10-15 09:00:00**.

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 6-72** describes the macro variable definitions of time and date.

Table 6-72 Macro variable definition of time and date

Macro Variable	Description	Display Effect
${dateformat(yyyy-MM-dd)}	Returns the current date in **yyyy-MM-dd** format.	2017-10-16
${dateformat(yyyy/MM/dd)}	Returns the current date in **yyyy/MM/dd** format.	2017/10/16
${dateformat(yyyy_MM_dd HH:mm:ss)}	Returns the current time in **yyyy_MM_dd HH:mm:ss** format.	2017_10_16 09:00:00

 Copyright © Huawei Technologies Co., Ltd.

Macro Variable	Description	Display Effect
${dateformat(yyyy-MM-dd HH:mm:ss, -1, DAY)}	Returns the current time in **yyyy-MM-dd HH:mm:ss** format. The date is one day before the current day.	2017-10-15 09:00:00
${timestamp()}	Returns the timestamp of the current time, that is, the number of milliseconds that have elapsed since 00:00:00 on January 1, 1970.	1508115600000
${timestamp(-10, MINUTE)}	Returns the timestamp generated 10 minutes before the current time point.	1508115000000
${timestamp(dateformat(yyyyMMdd))}	Returns the timestamp of 00:00:00 of the current day.	1508083200000
${timestamp(dateformat(yyyyMMdd,-1,DAY))}	Returns the timestamp of 00:00:00 of the previous day.	1507996800000
${timestamp(dateformat(yyyyMMddHH))}	Returns the timestamp of the current hour.	1508115600000

Time and Date Macro Variables of Paths and Table Names

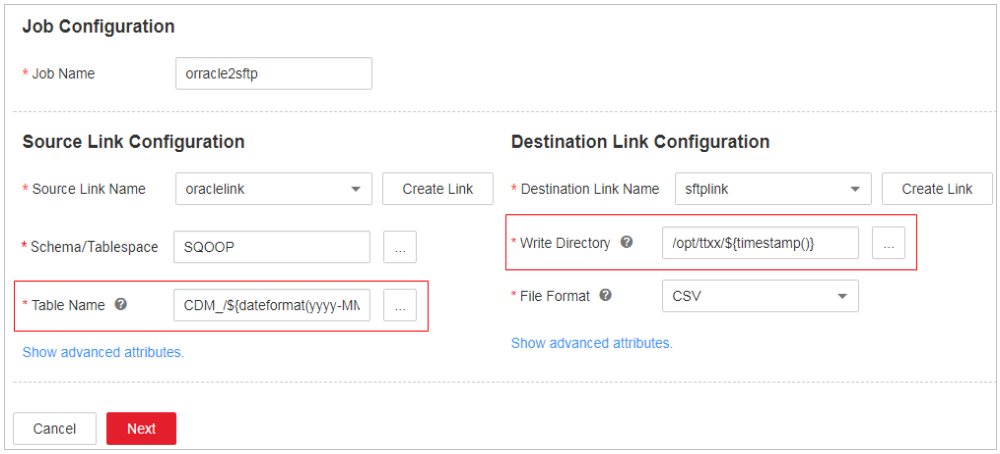
Figure 6-19 shows an example. If:

- **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 SFTP server.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-19 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 6-20 uses table **SQOOP.CDM_20171016** as an example. The table contains column **DS**, which indicates the time.

Figure 6-20 Table data

SELECT * FROM SQOOP.CDM_20171016

	FOO	BAR	DS
1	5	snap	2017-05-01
2	5	snap	2017-05-01
3	1	google	2017-05-02
4	4	oracle	2017-05-02
5	6	amd	2017-05-02
6	7	nvda	2017-05-02
7	1	google	2017-05-02
8	4	oracle	2017-05-02
9	6	amd	2017-05-02
10	7	nvda	2017-05-02
11	2	facebook	2017-10-15
12	3	tesla	2017-10-15
13	2	facebook	2017-10-15
14	3	tesla	2017-10-15

Suppose that the current date is **2017-10-16** and you want to export data generated the day before the current day (DS = 2017-10-15), then you can set the value of **Where Clause** to **DS='${dateformat(yyyy-MM-dd,-1,DAY)}'** when

 Copyright © Huawei Technologies Co., Ltd.

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 in the previous day will be exported at 00:00:00 every day.

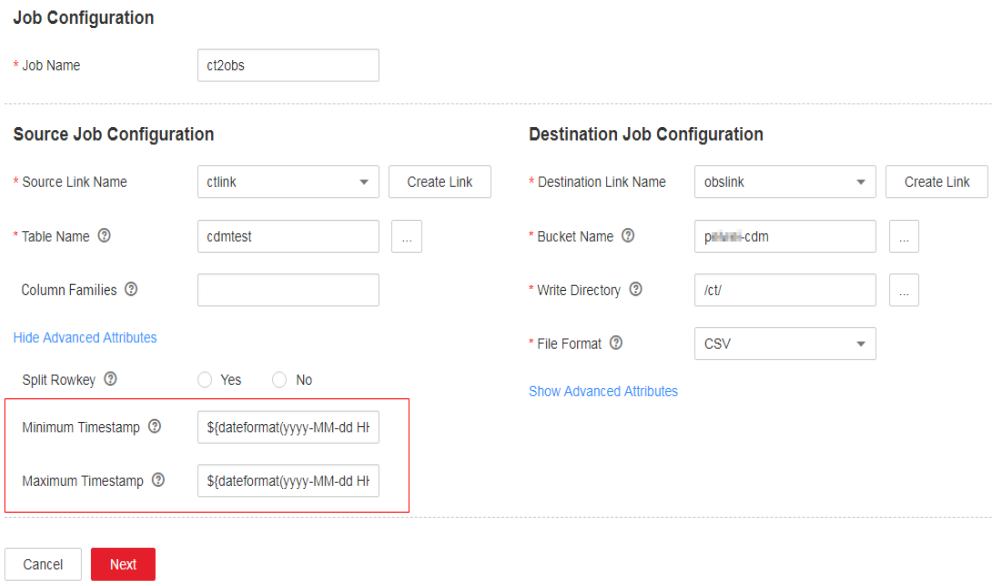
Configuration principles of other application scenarios are the same.

6.1.9.4 HBase/CloudTable Incremental Migration

You can use CDM to export data in a specified period of time from HBase (including MRS HBase, FusionInsight HBase, and Apache HBase) and CloudTable. The CDM **scheduled jobs** can be used together to implement incremental migration of HBase and CloudTable.

When creating a table/file migration job and selecting **Link to HBase** or **Link to CloudTable** as the source link, you can set the time range in advanced attributes.

Figure 6-21 Time range



 Copyright © Huawei Technologies Co., Ltd.

- 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.

Set the two parameters according to **Incremental Synchronization Using the 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.

6.1.10 Job Management

6.1.10.1 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

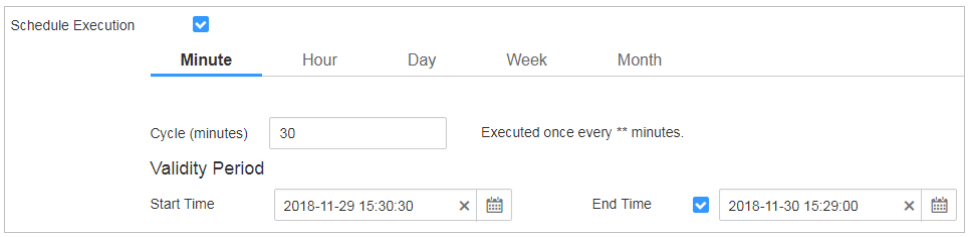
When configuring scheduled jobs, do not set the same scheduled time for different jobs. Instead, set different times to avoid exceptions.

Scheduling Job Execution by Minute

CDM allows jobs to be executed every several minutes. It is recommended that the cycle be at least 10 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 6-22 Scheduling job execution by minute



 Copyright © Huawei Technologies Co., Ltd.

Figure 6-22 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.

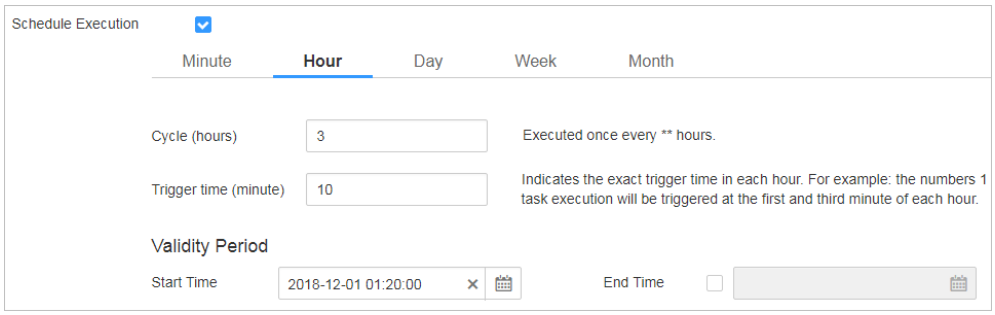
- **Cycle (hours)**: indicates the interval when a job is automatically executed.
- **Trigger Time (minute)**: indicates the exact time in each hour when a scheduled task is triggered. The value ranges from 0 to 59. You can set a maximum of 60 values and use commas (,) to separate these values. However,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

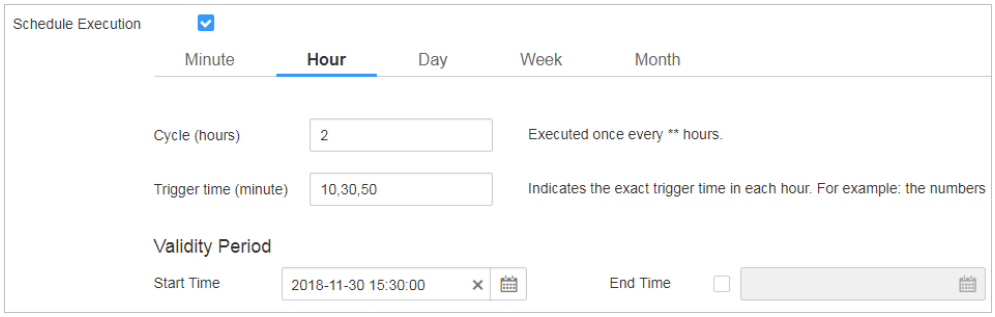
  Figure 6-23 shows that the first automatic execution time is **2:10:00**, and the second automatic execution time is **5:10:00**.

  Figure 6-23 Trigger time beyond the validity period

- **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 6-24 Scheduling job execution by hour



 Copyright © Huawei Technologies Co., Ltd.

Figure 6-24 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 6-25 Scheduling job execution by day

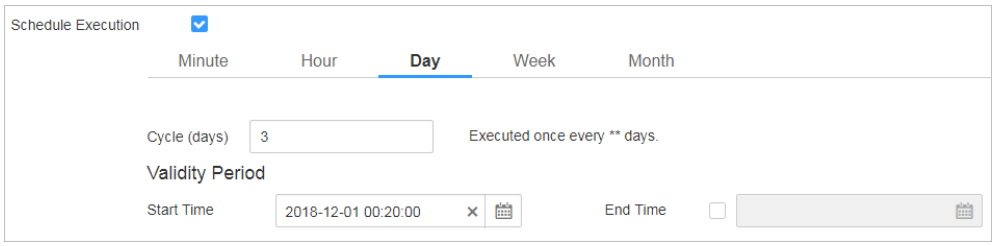


Figure 6-25 shows that the scheduled job will be automatically executed at 00:20:00 on December 1, 2018, and is executed once every three days. The configuration is always valid.

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.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-26 Scheduling job execution by week

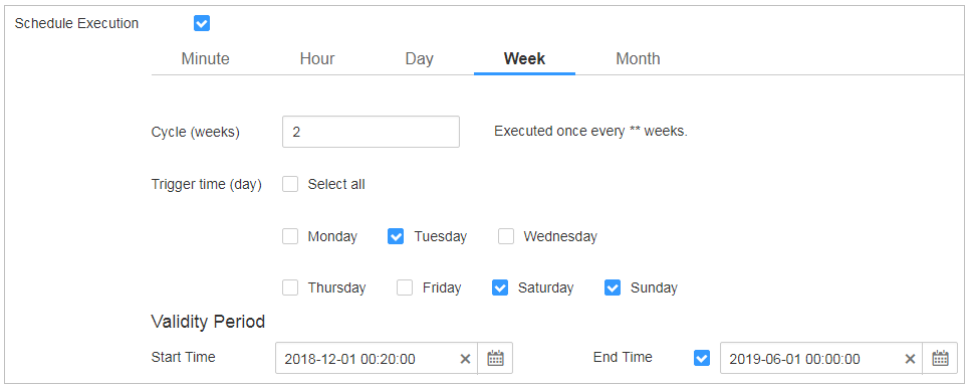


Figure 6-26 shows that the job will be automatically executed at 00:20:00 every Tuesday, Saturday, and Sunday every two weeks starting from 00:20:00 on December 1, 2018, and the job will be automatically stopped at 00:00:00 on June 1, 2019.

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 6-27 Scheduling job execution by month

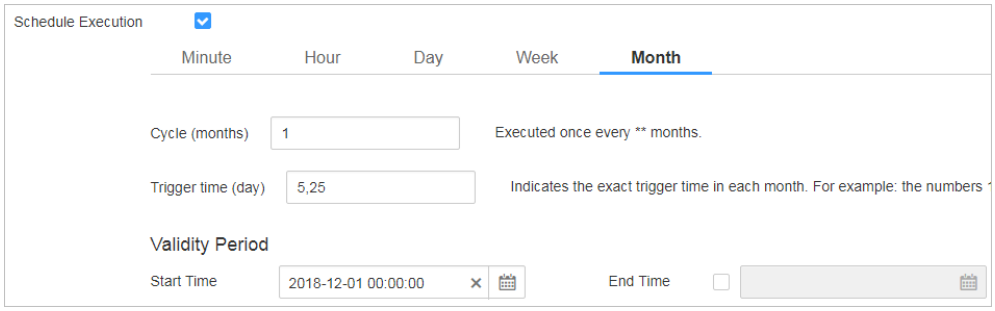


Figure 6-27 shows that the job will be automatically executed at 00:00:00 on every fifth and twenty-fifth day of each month starting from 00:00:00 on December 1, 2018. The configuration is always valid.

 Copyright © Huawei Technologies Co., Ltd.

6.1.10.2 Job Configuration Management

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

- **Automatic Backup and Restoration of CDM Jobs**
- **Global Variables of CDM Job Parameters**

Automatic Backup and Restoration of CDM Jobs

- Prerequisites

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

- Automatic backup

  On the **Job Management** page, click the **Configuration Management** tab and configure the parameters of **Scheduled Backup**.

  Table 6-73 Parameters for **Scheduled Backup**

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

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Environment Variable	Global control parameter. For example, if xxx=123 exists, you can use ${xxx} to replace 123 in jobs.	AAA=333
Maximum Concurrent Extractors	It is user-defined. Do not set the same scheduled time for different jobs. Instead, set different times to avoid exceptions.	5

- 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.

Global Variables of CDM Job Parameters

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

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

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

   ```
   buket_1=A
   buket_2=B
   ```

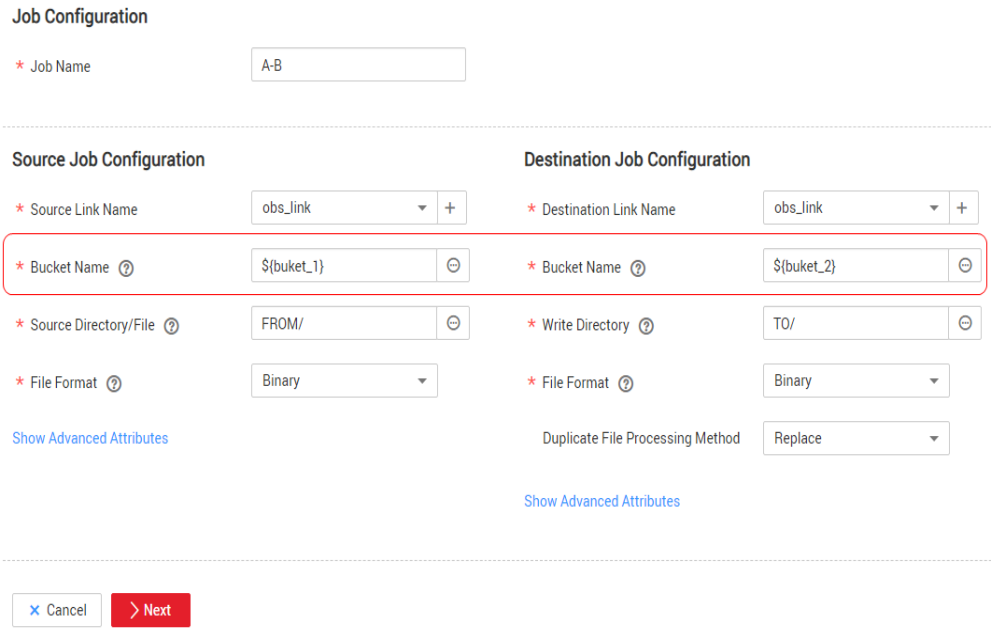
   Variable **buket_1** indicates bucket A, and variable **buket_2** indicates bucket B.

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

   Set the source bucket name to **${buket_1}** and destination bucket name to **${buket_2}**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-28 Setting the bucket names to environment variables



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

```
buket_1=C
buket_2=D
```

6.1.10.3 Single Job Management

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 historical records**

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

- **Viewing job logs**

  On the **Historical Record** page, 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**

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

- **Querying the job statistics**

 Copyright © Huawei Technologies Co., Ltd.

Open the preview window of a configured database job to preview a maximum of 1,000 pieces of data. By comparing the number of data records of the migration source and destination, you can check whether the migration is successful and whether data is lost.

- **Viewing historical jobs**

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

  For a job that is periodically executed, a historical job is generated on the **Historical Jobs** tab page each time when the job is executed, regardless of whether the job is successfully executed. Because the original job name is the same, the names of the historical jobs will carry a random character string for distinction.

Modifying a Job

- **Modifying the job parameters**

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

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

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

Procedure

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Integration** of the corresponding workspace.

   The **Data Integration** page is displayed.

3. Choose **Cloud Data Migration**. The **Cluster Management** page is displayed. Click **Job Management** in the **Operation** column of the cluster.

4. 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.

5. 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.

 Copyright © Huawei Technologies Co., Ltd.

- 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 file, which is similar to modify the job parameters.

6. After the modification, click **Save** or **Save and Run**.

6.1.10.4 Batch Job Management

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 the 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

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Integration** of the corresponding workspace.

   The **Data Integration** page is displayed.

3. Choose **Cloud Data Migration**. The **Cluster Management** page is displayed. Click **Job Management** in the **Operation** column of the cluster.

 Copyright © Huawei Technologies Co., Ltd.

4. 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.

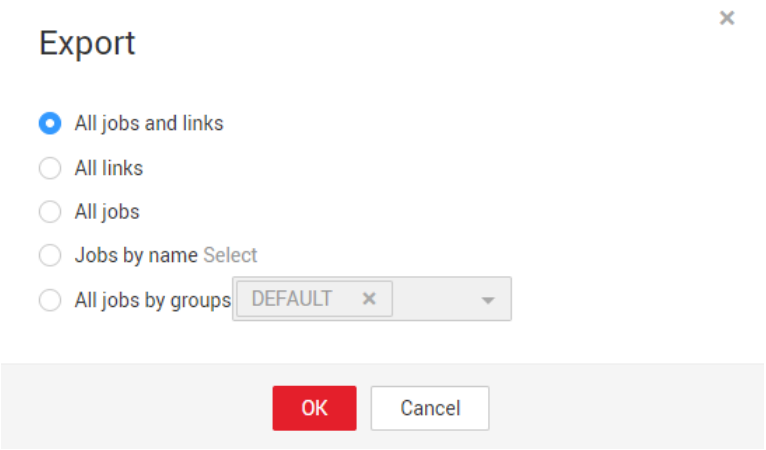
     In the third step of creating a job, if jobs have been assigned to different groups, you can display, start, or export jobs by group.
   - **Run jobs in batches.**

     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**. The page for batch export is displayed. See **Figure 6-29**.

     Figure 6-29 Batch export page

     - **All jobs and links**: If this option is selected, all jobs and links are exported at a time.
     - **All jobs**: If this option is selected, all jobs are exported at a time.
     - **All links**: If this option is selected, all links are exported at a time.
     - **Jobs by name**: If this option is selected, jobs to be exported need to be specified. Then, click **OK** to export the selected jobs.
     - **All jobs by groups**: If this option is selected, groups to be exported need to be specified. Then click **OK** to export the selected groups.

     Click **Export All** to export all jobs in JSON format. These files can be used as backups or imported to another cluster.

     You can export all jobs at a time or export jobs by job name or group.

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

 Copyright © Huawei Technologies Co., Ltd.

- **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. 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 of the jobs to the text box.

6.1.10.5 Agent Management

For HDFS and relational database data sources, if nodes cannot be exposed, you can deploy an agent on the source network. CDM uses an agent to obtain data from the internal data source of the customer. However, data cannot be written into the data source.

Figure 6-30 Scenario

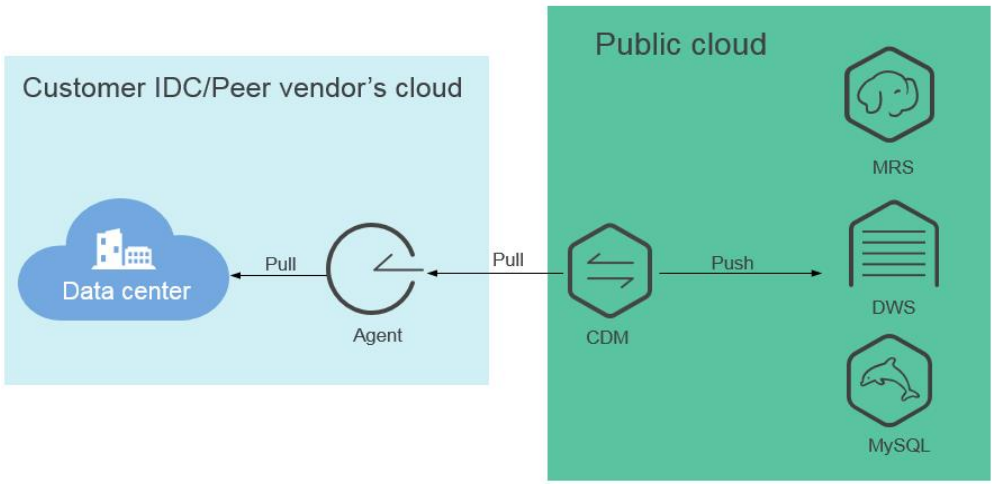
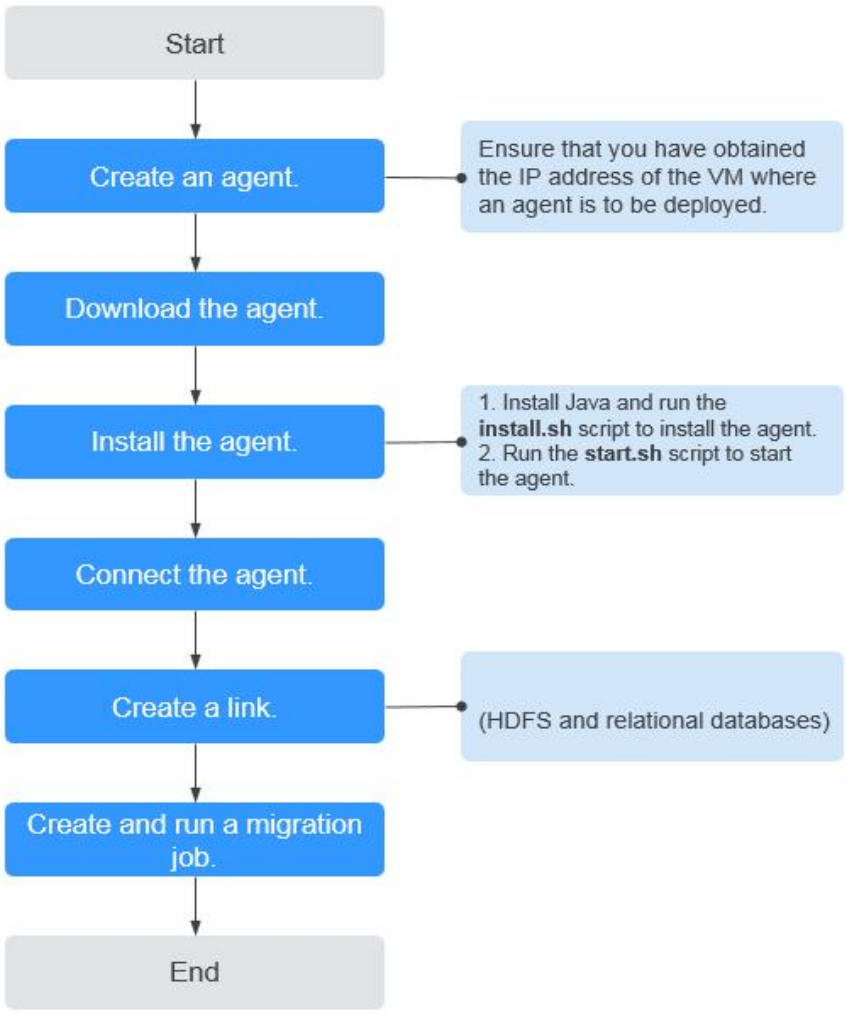


Figure 6-31 shows the process of using an agent.

Copyright © Huawei Technologies Co., Ltd.

Figure 6-31 Process of using an agent



Creating an Agent

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

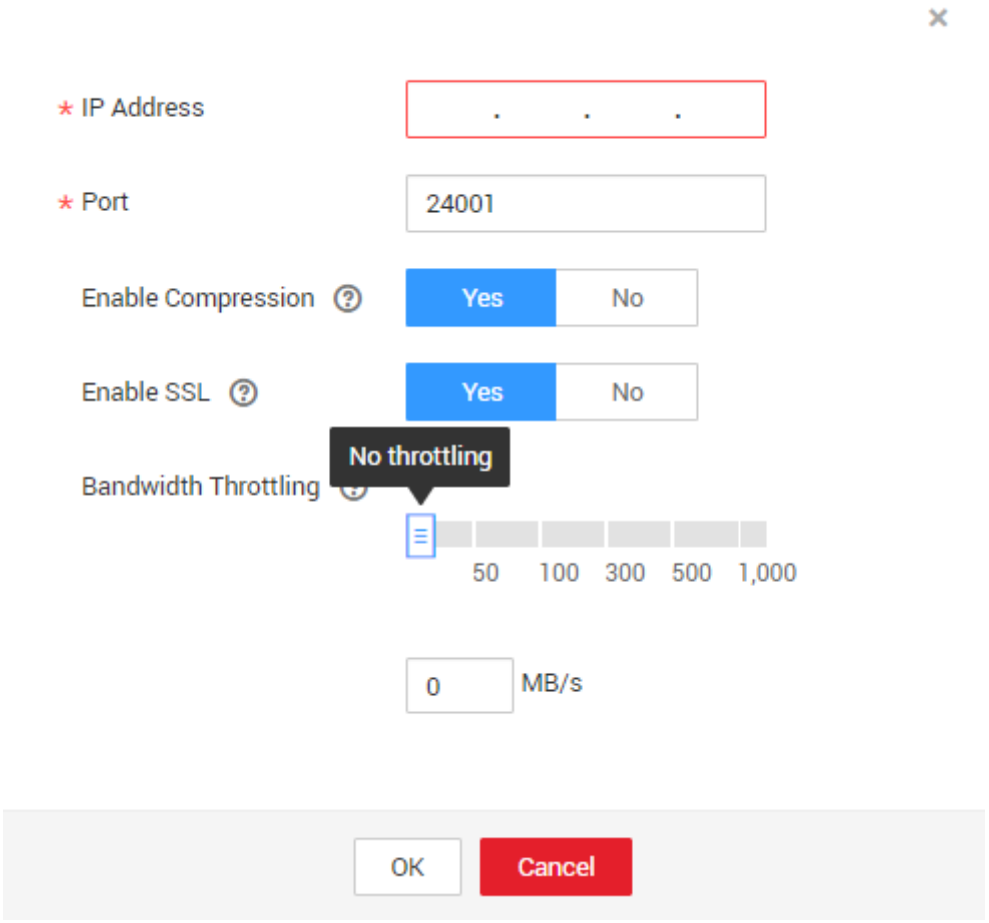
2. Click **Data Integration** of the corresponding workspace.

   The **Data Integration** page is displayed.

3. In the left navigation tree, choose **Cluster Management**. On the displayed page, choose **Job Management** > **Agent Management** > **Create Agent** and configure the agent information. For details, see **Figure 6-32**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-32 Configuring an agent



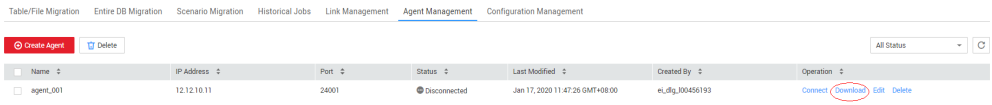
- **IP Address**: Set this parameter to the IP address of the host used to deploy the agent on the source network.
- **Port**: the port number exposed by the agent. Recommended value range: 1024–65535.
- **Enable Compression**: Whether to compress data for transmission.
- **Enable SSL**: Whether to enable SSL two-way authentication to ensure data security.
- **Bandwidth Throttling**: Set the maximum downstream rate of the agent. No throttling by default.

4. Click **OK**. On the **Agent Management** page, you can view the created agent.

Downloading and Installing an Agent

1. On the **Agent Management** page, find the created agent. Click **Download** next to the created agent to download it, as shown in **Figure 6-33**.

Figure 6-33 Downloading an agent



 Copyright © Huawei Technologies Co., Ltd.

2. Upload the downloaded agent package to the host where the agent is deployed.

   NOTE

   The agent depends on Java 8. Before deploying the agent, ensure that Java 8 (64-bit) has been installed and Java environment variables have been configured.

3. Decompress the installation package and run the following commands to install the agent:

   sh sbin/install.sh

   su Ruby

   sh sbin/start.sh

4. After the installation is complete, run the following command to check whether the agent process is started:

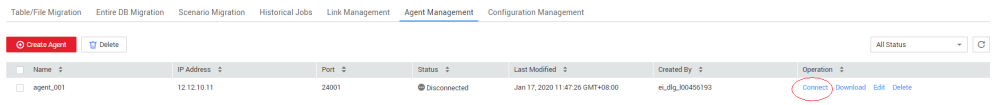
   ps -ef | grep agent

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

Connecting to an Agent

1. On the **Agent Management** page, find the created agent. Click **Connect** to connect to the created agent, as shown in **Figure 6-34**.

   Figure 6-34 Connecting to an agent

2. After the agent is successfully connected, you can select the agent in **Link to HDFS** and **Link to Relational Databases**.

6.1.11 Cluster Management

6.1.11.1 Binding or Unbinding EIPs

Scenario

If you are binding an EIP to or unbind it from a CDM cluster, the EIPs you use are billed based on the VPC service billing rules.

- 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 an SNAT rule**.
- The EIP exception notification takes effect only after the VPC policy agency of the corresponding region is created on the **IAM management console**. You can also choose **Authorize EIP Check** > **Create Agency** on the **Cluster Management** page to create an agency.

 Copyright © Huawei Technologies Co., Ltd.

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.
- An EIP can be bound only when the EIP quota is available.

Procedure

1. Log in to the DGC console, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

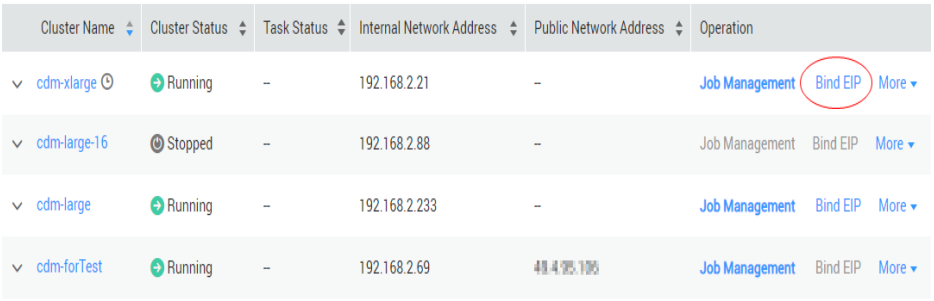
2. Click **Data Integration** of the corresponding workspace.

   The **Data Integration** page is displayed.

3. Choose **CDM**. The **Cluster Management** page is displayed.

   a. Binding an EIP: In the **Operation** column, click **Bind EIP**. The **Bind EIP** dialog box is displayed.

   Figure 6-35 Binding an EIP

   b. Unbinding an EIP: In the **Operation** column, choose **More** > **Unbind EIP**.

4. Click **OK**.

6.1.11.2 Modifying Cluster Configurations

Configuration Description

After a CDM cluster is created, you can modify the following configurations:

- **Auto Shutdown**

  If no job is running in the cluster and no scheduled job is configured, the CDM cluster automatically shuts down after 15 minutes to save costs.

- **Scheduled Startup/Shutdown**

  The scheduled startup/shutdown and auto shutdown functions cannot be enabled at the same time. You can configure scheduled startup/shutdown at a specific time every day. During scheduled shutdown, the system does not wait for unfinished jobs to complete.

 Copyright © Huawei Technologies Co., Ltd.

- **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. Up to five mobile numbers and five email addresses can be configured.

- **User Isolation**
  - If this function is enabled, migration jobs and links in the cluster are isolated, so that other IAM users under the same cloud account cannot view 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 CDM Administrator permission under the same cloud account can view and perform operations on the jobs and links. After disabling Isolate User, restart the cluster VM for the settings to take effect.

Procedure

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Integration** of the corresponding workspace.

   The **Data Integration** page is displayed.

3. Choose **Cloud Data Migration**. The **Cluster Management** page is displayed.
4. Click the name of a cluster and click the **Cluster Configuration** tab to modify auto shutdown, scheduled startup/shutdown, notification, or user isolation.

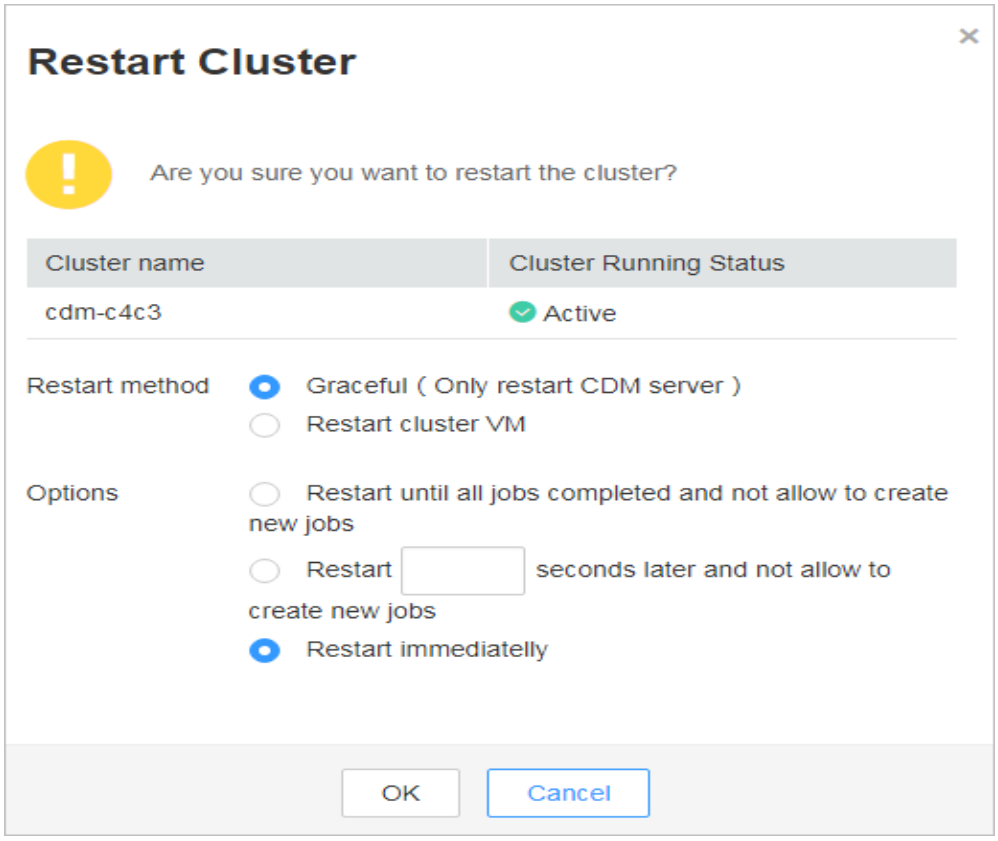
 Copyright © Huawei Technologies Co., Ltd.

Figure 6-36 Modifying cluster configurations

5. Click **Save**. The **Cluster Management** page is displayed.
6. If **User Isolation** is disabled, restart the cluster VM for the settings to take effect. In the **Operation** column, choose **More** > **Restart**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-37 Restarting a cluster



a. **Graceful restart**: Only the CDM service process is restarted. The cluster VM will not be restarted.

b. **Cluster VM restart**: The service process will be interrupted and VMs in the cluster will be restarted.

7. Select **Cluster VM restart** and click **OK**.

6.1.11.3 Viewing Cluster Configurations, Logs, and Monitoring Data

Scenario

View cluster configurations, obtain cluster logs, and view monitoring data on Cloud Eye.

Prerequisites

You have created a CDM cluster.

Procedure

1. Log in to the DGC console, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

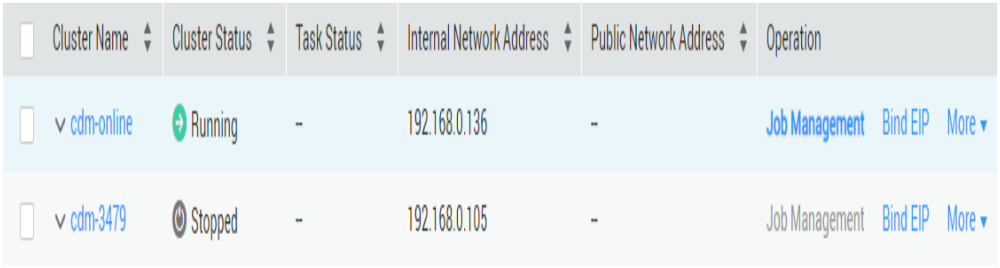
2. Click **Data Integration** of the corresponding workspace.

   The **Data Integration** page is displayed.

 Copyright © Huawei Technologies Co., Ltd.

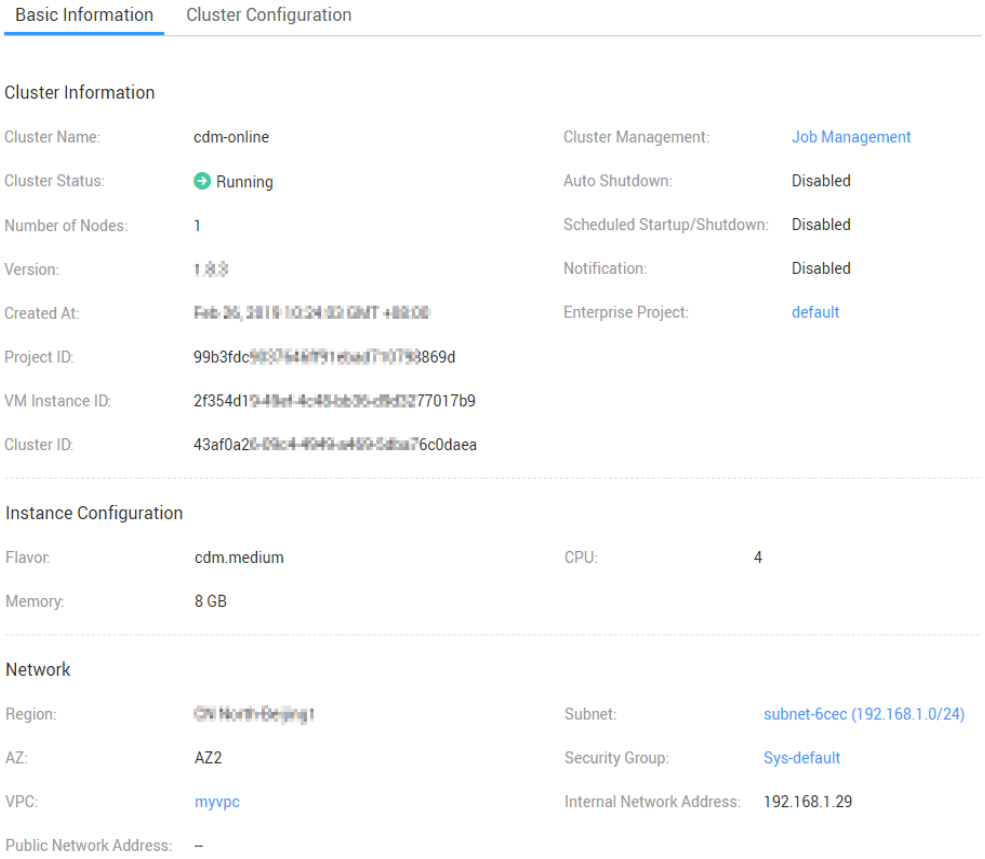
3. Choose **Cloud Data Migration**. On the **Cluster Management** page, view the cluster list.

Figure 6-38 Cluster list



4. Click the cluster name to view the basic information about the cluster, including the cluster flavor, creation time, node quantity, node configurations, network configurations, project ID, cluster ID, and instance ID.

Figure 6-39 Cluster configurations



5. In the row of the cluster, choose **More** > **Download Log** to obtain cluster logs.
6. In the **Operation** column of the cluster, choose **More** > **View Metric**. The Cloud Eye monitoring page is displayed, on which you can view the inbound and outbound rates, and CPU and memory usages.

6.1.12 Typical Data Migration Scenarios

 Copyright © Huawei Technologies Co., Ltd.

6.1.12.1 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

- You have purchased DWS/DDS clusters.
- You have obtained the IP address, port number, database name, username, and passwords for connecting to the DWS and DDS databases. In addition, you must have read, write, and delete permissions for the DDS and DWS databases.

Creating a CDM Cluster and Binding an EIP to the Cluster

Step 1 Create a CDM cluster. For details, see **Buying a DGC Incremental Package**. 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, and 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. An EIP must be bound as described in **Binding or Unbinding EIPs** 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.

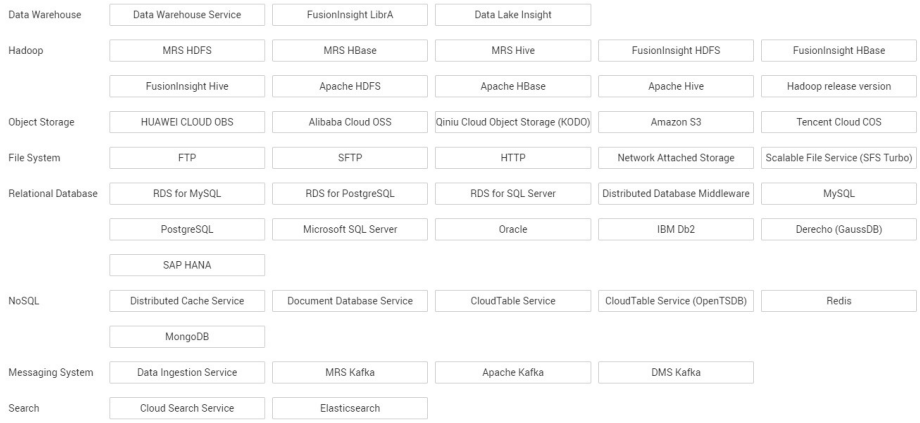
----End

Copyright © Huawei Technologies Co., Ltd.

Creating a DDS Link

Step 1 Click **Job Management** in the **Operation** column of the CDM cluster. On the page that is displayed, choose **Link Management** > **Create Link**. The page for selecting a connector is displayed. See **Figure 6-40**.

Figure 6-40 Selecting a connector



Step 2 To create a DDS link, select **Document Database Service** and click **Next**. On the page that is displayed, configure the link parameters based on **Table 6-74**.

Table 6-74 DDS link parameters

Parameter	Description	Example Value
Name	The link name, which should be chosen based on the data source type, so it is easier to remember what the link is for.	mongo_link
MongoDB Server List	And address list for the DDS cluster. The format is **IP address or domain name of the database server:port**. Separate multiple server lists with semicolons (;).	192.168.0.1:7300;192.168.0.2:7301
Database Name	The name of the DDS database to be connected	DB_mongodb
Username	The username for logging in to the DDS database	cdm
Password	The password for logging in to the DDS database	N/A

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

----End

Copyright © Huawei Technologies Co., Ltd.

Creating a DWS Link

Step 1 On the **Link Management** tab page, click **Create Link** and select **Data Warehouse Service** to create a DWS link.

Step 2 Click **Next**. The page for configuring the DWS link parameters is displayed. Configure the mandatory parameters described in **Table 6-75** and retain the default values for optional parameters.

Table 6-75 DWS link parameters

Parameter	Description	Example Value
Name	A unique link name	dwslink
Database Server	The IP address or domain name of the DWS database server	192.168.0.3
Port	The DWS database port	8000
Database Name	The name of the DWS database	db_demo
Username	A user with read, write, and delete permissions for the DWS database	dbadmin
Password	The password of the user	N/A
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**.	N/A

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Import Mode	When creating a DWS link, select the data import mode. • **COPY**: Source data is migrated to the DWS management node and then copied to DataNodes. To access DWS through the Internet, select **COPY**. • **GDS**: DataNodes of DWS concurrently request data from the GDS component of CDM and then write data to DWS. GDS mode cannot be used for data export from DWS. Theoretically, the GDS mode is more efficient than the COPY mode. If the GDS mode is used, you need to modify the security group configuration on VPC. Modify the security group rule of the CDM cluster so that the DWS data node can access port 25000 of the internal IP address of the CDM cluster.	COPY

Step 3 Click **Save**. The link is successfully created.

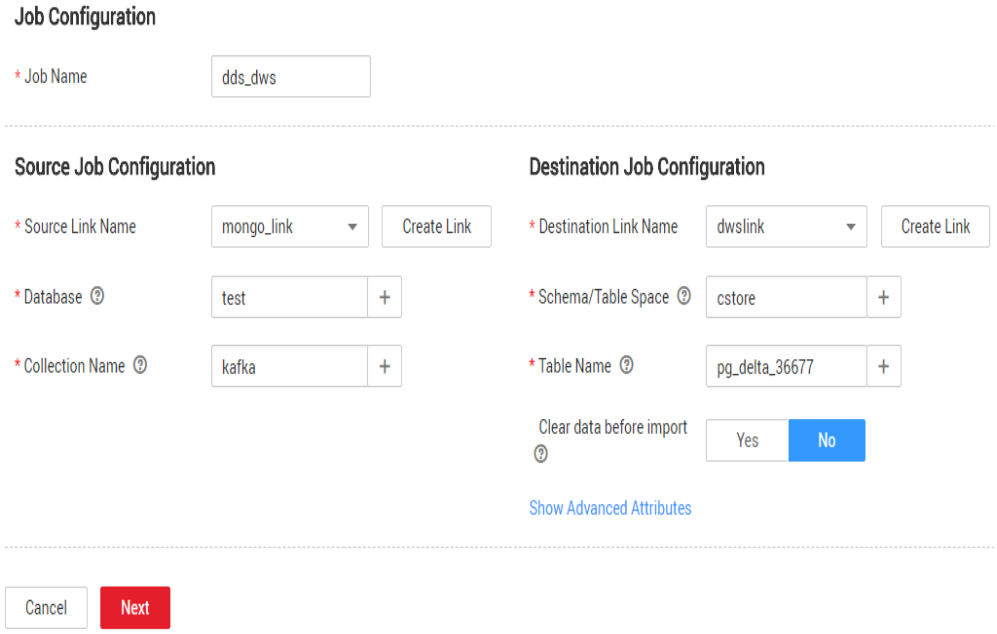
----End

Creating a Migration Job

Step 1 Choose **Table/File Migration** > **Create Job** to create a data migration job.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-41 Creating a job for migrating data from DDS to DWS



Step 2 Configure the required job details:

- **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 will be migrated.
  - **Collection Name**: Enter the name of the MongoDB collection on DDS, which should be 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 that the data will be written to.
  - **Table Name**: The name of the table the data will be written to. 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.

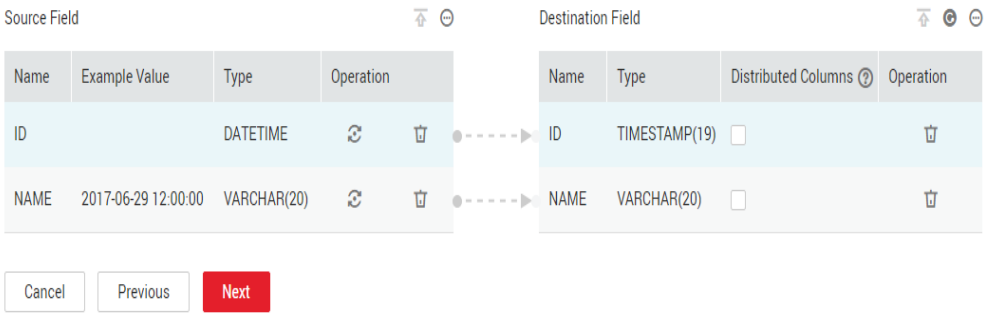
- If the field mapping is incorrect, click the row where the field is located and drag the field to adjust the mapping.
- When importing data to DWS, you need to manually select the distribution columns of DWS. You are advised to select the distribution columns according to the following principles:

 Copyright © Huawei Technologies Co., Ltd.

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. If no primary key is available, and no distribution column is selected, DWS uses the first column as the distribution column by default. As a result, the data may be skewed.

- If you need to convert the content of the source fields, refer to **Field Conversion**. In this example, the field conversion is not required.

Figure 6-42 Field mapping 1



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

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

- **Retry Upon Failure**: If the job fails, you can determine whether to automatically retry the job. Retain the default value **Never**.
- **Group**: Select which group the job belongs to. 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 executed concurrently. Retain the default value **1**. The maximum number of concurrent extractors in a single job is 500.
- **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

 Copyright © Huawei Technologies Co., Ltd.

6.1.12.2 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 Cloud Search Service. The procedure involves four steps:

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

The basic package of DGC includes a CDM cluster. If the cluster cannot meet service requirements, you can buy a DGC incremental package as required after buying a DGC instance of the basic package. The key configurations for creating CDM clusters are as follows:

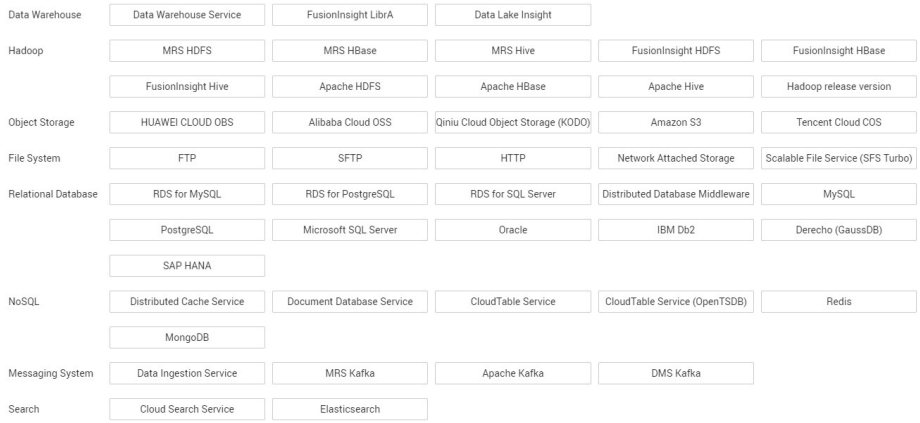
- The flavor of the CDM cluster is selected based on the amount of data to be migrated. Generally, **cdm.medium** meets requirements for most migration scenarios.
- The CDM and Cloud Search Service clusters must be in the same VPC. In addition, it is recommended that the CDM cluster be in the same subnet and security group as the Cloud Search Service cluster.
- If the same subnet and security group cannot be used, for security reasons, 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 page that is displayed, choose **Link Management** > **Create Link**. The page for selecting a connector is displayed. See **Figure 6-43**.

Copyright © Huawei Technologies Co., Ltd.

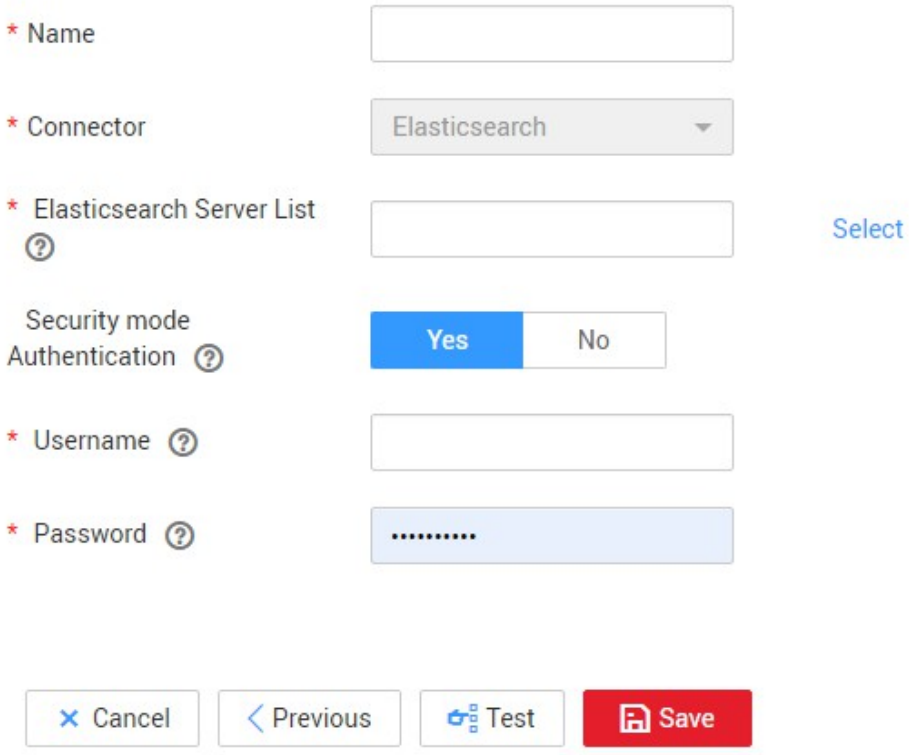
Figure 6-43 Selecting a connector



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

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

Figure 6-44 Creating a CSS link



 Copyright © Huawei Technologies Co., Ltd.

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

----End

Creating an OBS Link

Step 1 On the **Link Management** tab page, click **Create Link**. On the page that is displayed, select **HUAWEI CLOUD OBS**, click **Next**, and configure the required link parameters.

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

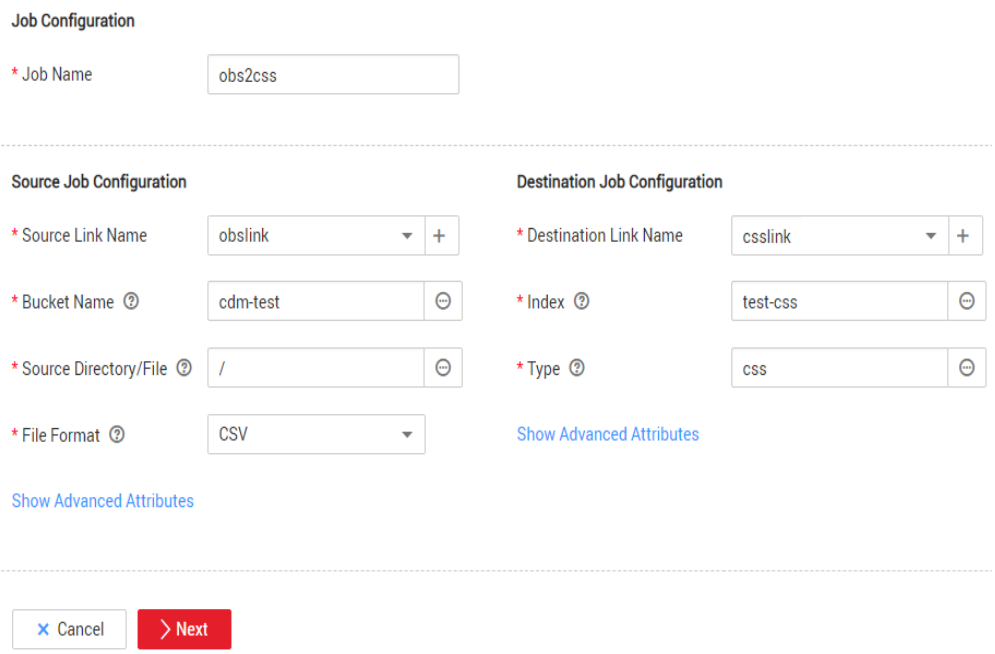
Step 2 Click **Save**. The **Link Management** page is displayed.

----End

Creating a Migration Job

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

Figure 6-45 Creating a job for migrating data from OBS to Cloud Search Service



- **Job Name**: Enter a unique name.
- **Source Job Configuration**
  - **Source Link Name**: Select the **obslink** link created in **Creating an OBS Link**.
  - **Bucket Name**: Select the bucket from which the data is to 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.

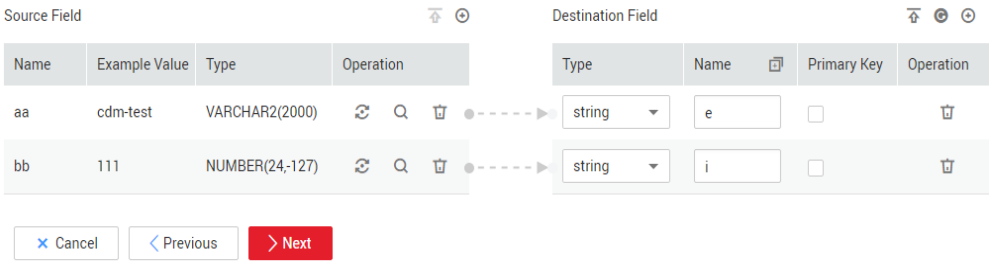
 Copyright © Huawei Technologies Co., Ltd.

  - **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 **Source OBS Parameters**.
- **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 **Destination Elasticsearch or CSS Parameters**.

Step 2 Click **Next**. The **Map Field** page is displayed. CDM automatically matches the source and destination fields, as shown in **Figure 6-46**.

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

Figure 6-46 Field mapping of CSS



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

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

- **Retry Upon Failure**: If the job fails, you can determine whether to automatically retry the job. Retain the default value **Never**.
- **Group**: Select which group the job belongs to. 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 executed concurrently. Retain the default value **1**. The maximum number of concurrent extractors in a single job is 500.

 Copyright © Huawei Technologies Co., Ltd.

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

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

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

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

----End

6.1.12.3 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 subscribed to OBS and DLI.
- You have created resource queues, databases, and tables on DLI.

Creating a CDM Cluster

Select operation parameters according to the actual conditions.

- If you already have a DGC cluster that meets the requirement, click **Job Management** in the row of the cluster and create links on the page that is displayed.
- If you do not have a cluster that meets the requirement, you can buy DGC incremental packages for batch migration to supplement the basic packages. See for details.See **Buying an Incremental Package** for details.

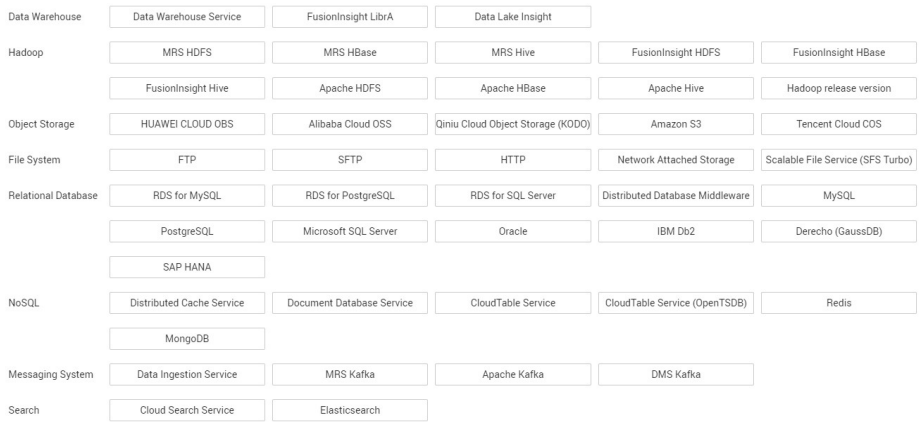
  In this scenario, if the CDM cluster is used only to migrate data from OBS to DLI and does not need to migrate data of other data sources, there is no special requirements on the VPC, subnet, and security group of the CDM cluster. You can specify them based on your needs. CDM accesses DLI and OBS through the intranet. 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 of most migration scenarios.

Copyright © Huawei Technologies Co., Ltd.

Creating a DLI Link

Step 1 Click **Job Management** in the **Operation** column of the CDM cluster. On the page that is displayed, choose **Link Management** > **Create Link**. The page for selecting a connector is displayed. See **Figure 6-47**.

Figure 6-47 Selecting a connector

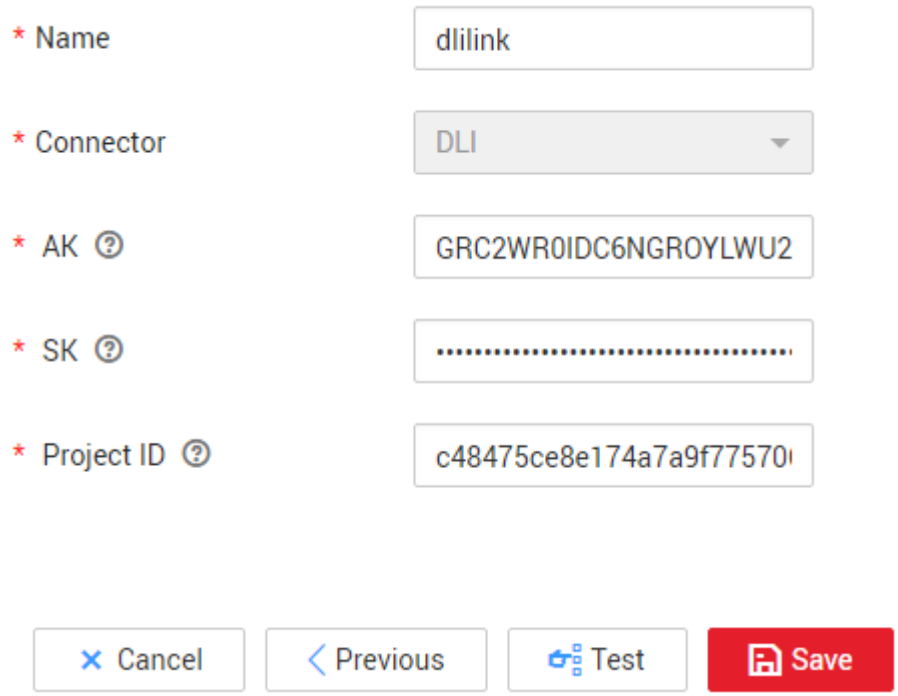


Step 2 Select **Data Lake Insight**, click **Next**, and configure the DLI link parameters. See **Figure 6-48**.

- Name: Enter a custom link name, for example, **dlilink**.
- AK and SK: Enter the AK and SK used for accessing the DLI database. For details, see the **Obtaining Authentication Information**.
- Project ID: The ID of the project to which DLI belongs. For details, see the **Obtaining Authentication Information**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-48 Creating a DLI link



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

----End

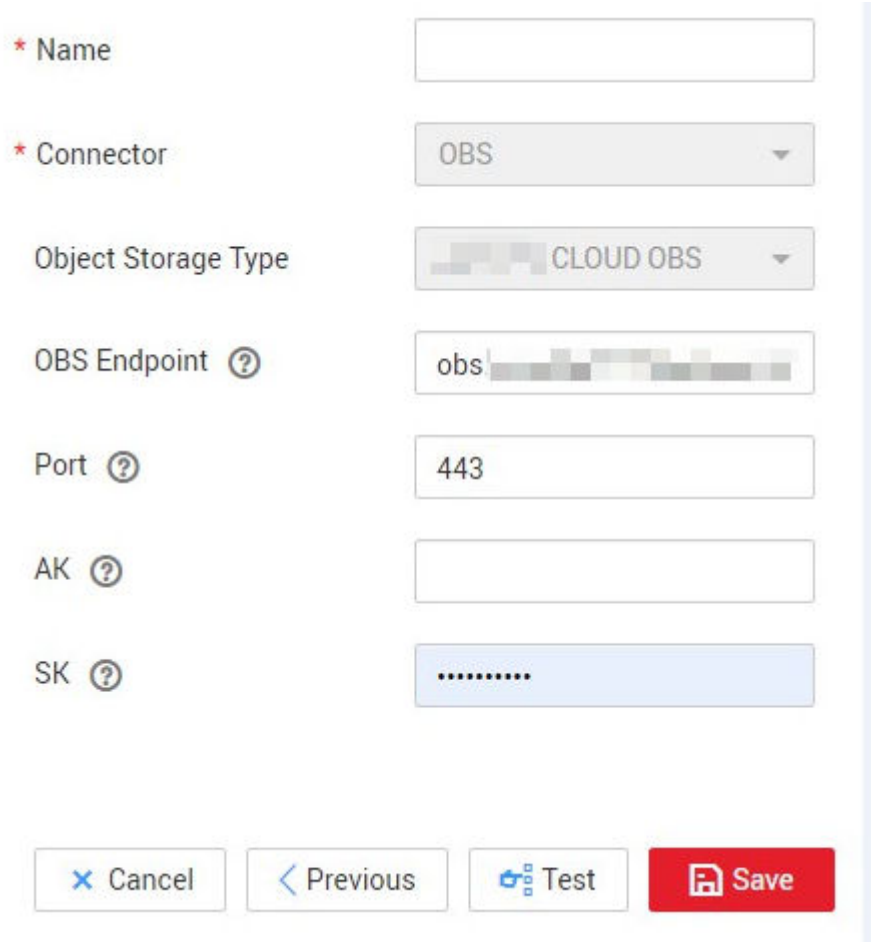
Creating an OBS Link

Step 1 On the **Link Management** tab page, click **Create Link**. On the page displayed, select **HUAWEI CLOUD OBS**, click **Next**, and configure the required link parameters. See **Figure 6-49**.

- **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.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-49 Creating an OBS link



Step 2 Click **Save**. The **Link Management** page is displayed.

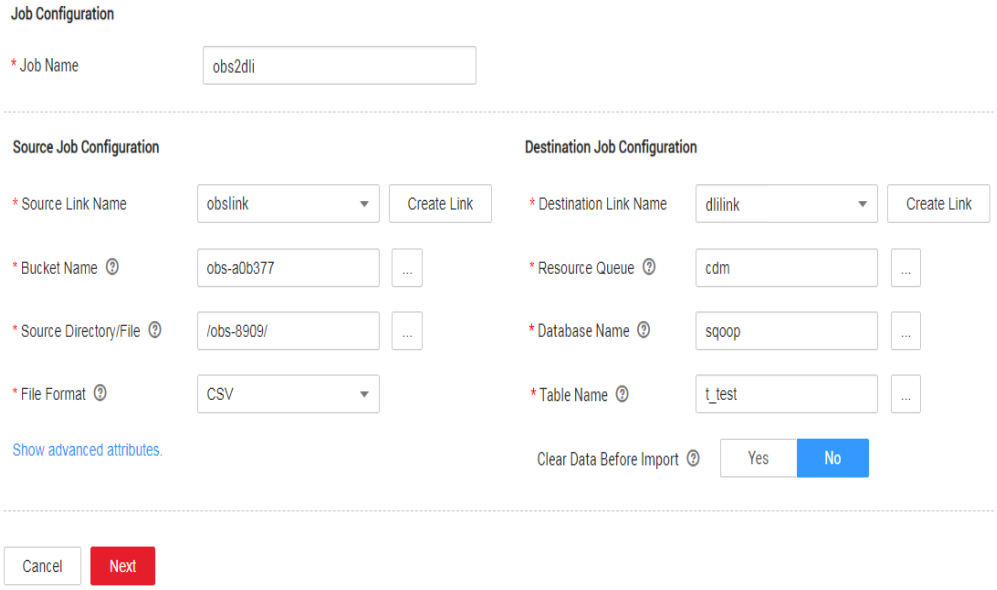
----End

Creating a Migration Job

Step 1 Choose **Table/File Migration** > **Create Job** to create a job for migrating data from OBS to DLI. See **Figure 6-50**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-50 Creating a job for migrating data from OBS to DLI



- **Job Name**: Enter a custom job name.
- **Source Link Name**: Select the **obslink** link created in **Creating an OBS Link**.
  - **Bucket Name**: Select the bucket the data will be migrated from.
  - **Source Directory/File**: Set this parameter to the path where the data to be migrated is.
  - **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 **Source OBS Parameters**.
- **Destination Link Name**: Select the **dlilink** link created in **Creating a DLI Link**.
  - **Resource Queue**: Enter the resource queue the destination table belongs to.
  - **Database Name**: Enter the name of the database the data will be written to.
  - **Table Name**: Enter the name of the table the data will be written to. 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, the default value is retained.

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 **Field Conversion**.

 Copyright © Huawei Technologies Co., Ltd.

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

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

- **Retry Upon Failure**: If the job fails, you can determine whether to automatically retry the job. Retain the default value **Never**.
- **Group**: Select which group the job belongs to. 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 executed concurrently. Retain the default value **1**. The maximum number of concurrent extractors in a single job is 500.
- **Write Dirty Data**: Specify this parameter if data that fails to be processed or filtered out during job execution needs to be written to OBS for future viewing. Before writing dirty data, create an OBS link. Retain the default value **No** so that dirty data is not recorded.
- **Delete Job After Completion**: Retain the default value **Do not delete**.

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

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

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

----End

6.1.12.4 Migrating Data from Redis to DCS

Scenario

CDM can migrate data from the on-premises Redis database or third-party Redis service to DCS without programming. The procedure includes three steps:

1. **Creating a CDM Cluster and Binding an EIP to the Cluster**
2. **Creating the Redis and DCS Links**
3. **Creating a Migration Job**

Prerequisites

- Your EIP quota is sufficient.
- You have subscribed to DCS and obtained the IP address, port, and password of the DCS database.
- The on-premises Redis database can be accessed through the public network.
- You have obtained the IP address and password of the Redis server.

Copyright © Huawei Technologies Co., Ltd.

Creating a CDM Cluster and Binding an EIP to the Cluster

Step 1 The DGC basic package includes a CDM cluster. If the cluster cannot meet your requirements, you can buy a CDM incremental package to supplement the DGC basic instance. Create a CDM cluster. For details, see **Buying a DGC Incremental Package**. The key configurations are as follows:

- The flavor of the CDM cluster is selected based on the amount of data to be migrated. Generally, select **cdm.medium** to meet the requirements of most migration scenarios.
- The CDM and DCS 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 DCS cluster.
- If the subnets and security groups are inconsistent, configure a security group rule to enable the CDM cluster to access the DCS cluster.

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

NOTE

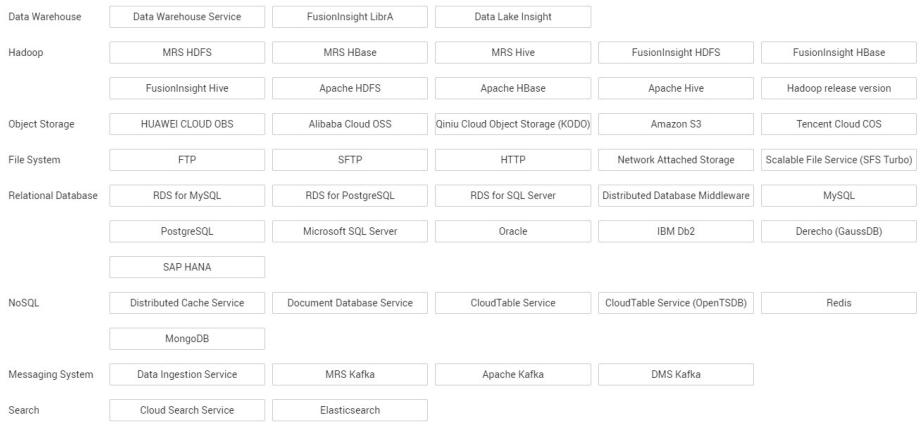
If SSL encryption is configured for the access channel of a local data source, CDM cannot connect to the data source using the EIP.

----End

Creating the Redis and DCS Links

Step 1 Click **Job Management** in the **Operation** column of the CDM cluster. On the page that is displayed, choose **Link Management** > **Create Link**. The page for selecting a connector is displayed. See **Figure 6-51**.

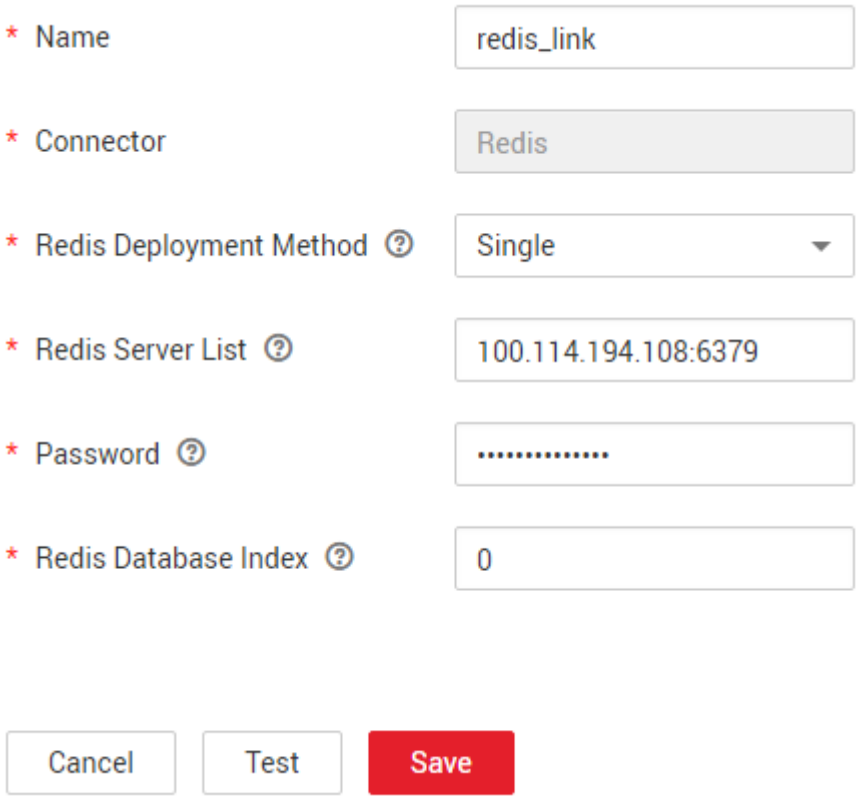
Figure 6-51 Selecting a connector



Step 2 Select **Redis** and click **Next**. On the page that is displayed, configure the Redis link parameters. See **Figure 6-52**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-52 Creating a Redis link



- **Name**: Enter a custom link name, for example, **redis_link**.
- **Redis Deployment Method**: Select a value based on the actual deployment method of the on-premises Redis database.
  - Single-node deployment: Select **Single**.
  - Cluster deployment: Select **Cluster**.
- **Redis Server List**: Set this parameter to the server address of the on-premises Redis database. Separate multiple server lists by semicolons (;).
- **Password** and **Redis Database Index**: Enter the password used for logging in to the on-premises Redis database and the index of the Redis database the data will be exported from.

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

Step 4 On the **Link Management** tab page, click **Create Link** to create a DCS link. The procedure for creating a DCS link and the link parameters is the same as those for a Redis link.

Step 5 Select **Distributed Cache Service**, click **Next**, and configure the DCS link parameters.

- **Name**: Enter a custom link name, for example, **dcs_link**.
- **Redis Deployment Method**: Select a value based on the DCS cluster deployment mode.
- **Redis Server List**: Set this parameter to the server address lists of the DCS database. Separate multiple server lists with semicolons (;).

 Copyright © Huawei Technologies Co., Ltd.

- **Password** and **Redis Database Index**: Enter the password used for logging in to the DCS database and the index of the database to which data is written.

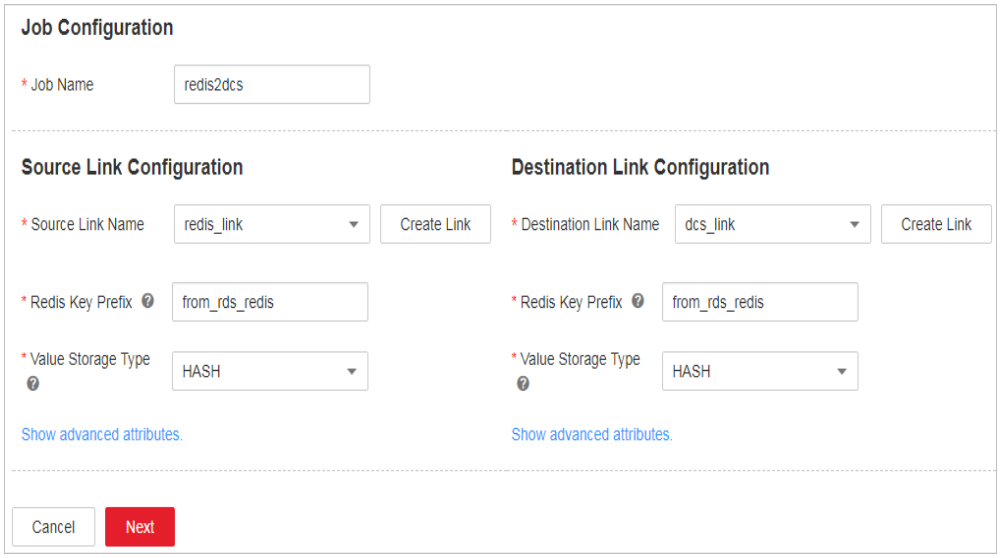
Step 6 Click **Save**. The **Link Management** page is displayed.

----End

Creating a Migration Job

Step 1 Choose **Table/File Migration** > **Create Job** to create a data migration job. See **Figure 6-53**.

Figure 6-53 Creating a job for migrating data from Redis to DCS



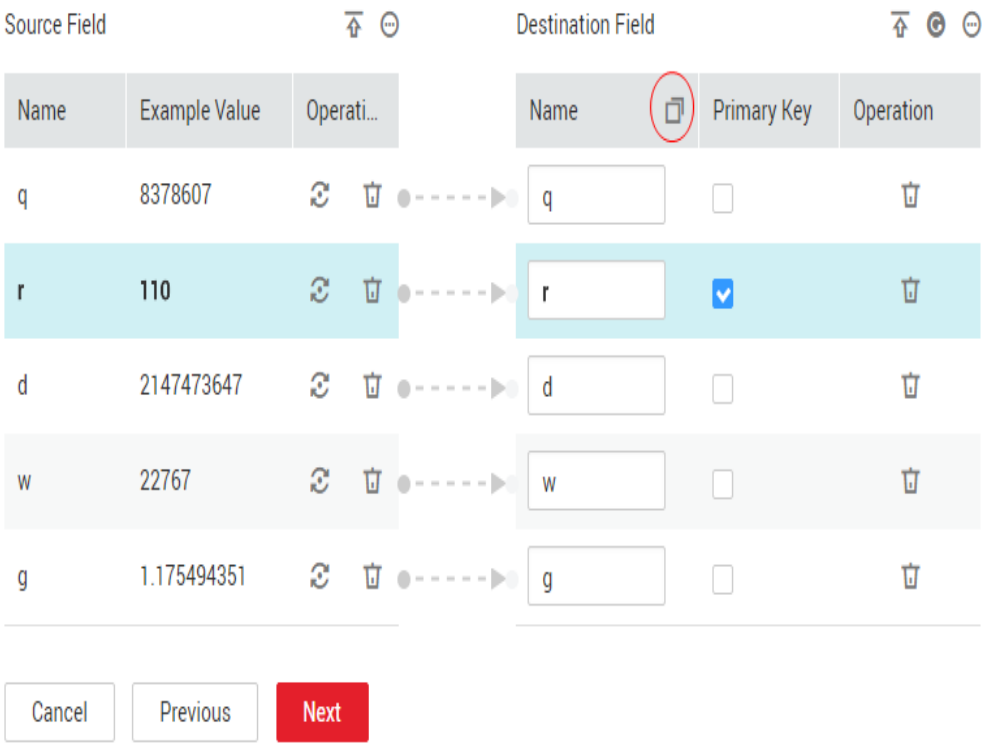
- **Job Name**: Enter a unique name, for example, **redis2dcs**.
- **Source Job Configuration**
  - **Source Link Name**: Select the Redis link created in **Creating the Redis and DCS Links**.
  - **Redis Key Prefix**: Select a key prefix that data will be exported from.
  - **Value Storage Type**: Select a value based on your needs. The example shown in the figure uses **Hash** on both the migration source and destination as an example.
  - Retain the default values for the optional parameters in **Show Advanced Attributes**. For details, see **Source Redis Parameters**.
- **Destination Job Configuration**
  - **Destination Link Name**: Select the DCS link created in **Creating the Redis and DCS Links**.
  - **Redis Key Prefix**: Select a key prefix to which data is to be imported.
  - **Value Storage Type**: Select **Hash**, which is the same as the migration source.
  - Retain the default values of the optional parameters in **Show Advanced Attributes**. For details, see **Destination DCS Parameters**.

 Copyright © Huawei Technologies Co., Ltd.

Step 2 After the basic job parameters are configured, click **Next** to enter the page for configuring field mapping.

For the hash type, you can click the button on the right of **Name** as highlighted in **Figure 6-54** to copy the fields on the migration source and then select the field that is used as the primary key.

Figure 6-54 Configuring field mapping

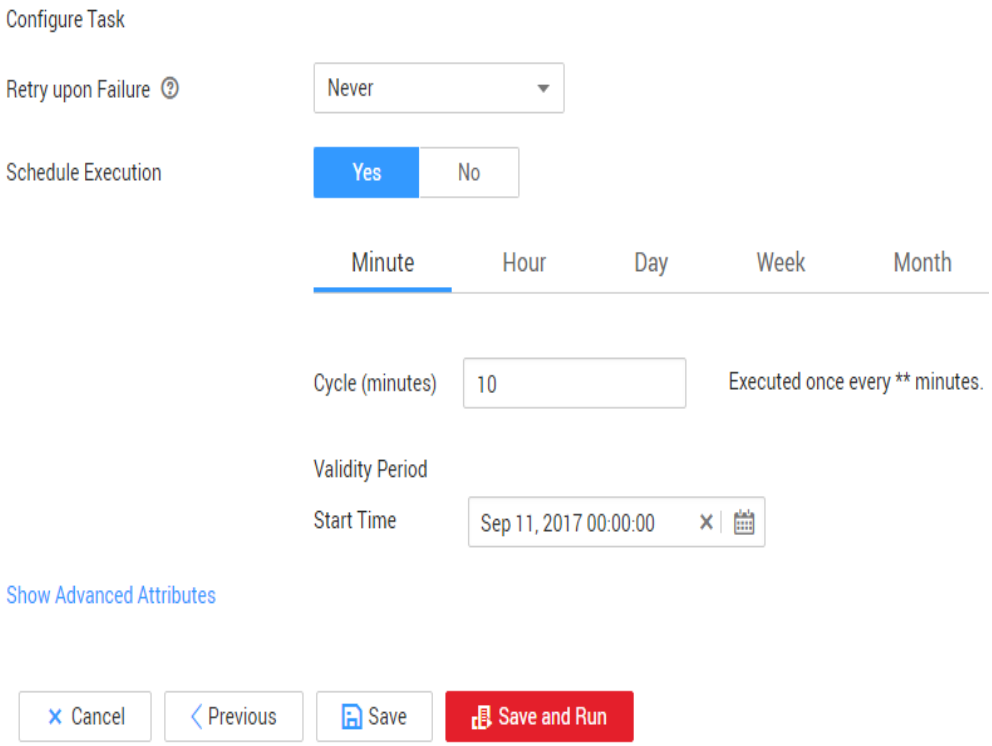


Step 3 Click **Next**. On the **Configure Task** page that is displayed, use **Schedule Execution** if required. For details, see **Figure 6-55**.

If **Schedule Execution** is enabled, CDM periodically synchronizes data. If data with duplicate primary keys exists, CDM automatically overwrites the existing primary key.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-55 Configuring scheduled tasks



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

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

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

----End

6.1.12.5 Migrating Data Oracle to CSS

Scenario

Cloud Search Service provides users with structured and unstructured data search, statistics, and reports. This section describes how to use CDM to migrate data from an Oracle database to Cloud Search Service. The procedure includes four steps:

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

- Your EIP quota is sufficient.

 Copyright © Huawei Technologies Co., Ltd.

- 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 HUAWEI CLOUD has been established.

Creating a CDM Cluster and Binding an EIP to the Cluster

Step 1 The DGC basic package includes a CDM cluster. If the cluster cannot meet your requirements, you can buy an incremental package to supplement the DGC basic instance. Create a CDM cluster. For details, see **Buying a DGC Incremental Package**. 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 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 reasons, 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.

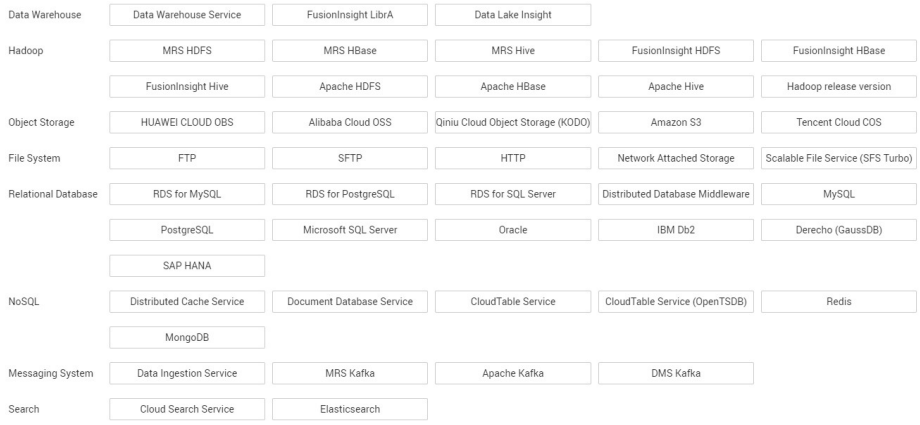
----End

Creating a Cloud Search Service Link

Step 1 Click **Job Management** in the **Operation** column of the CDM cluster. On the page that is displayed, choose **Link Management** > **Create Link**. The page for selecting a connector is displayed. See **Figure 6-56**.

 Copyright © Huawei Technologies Co., Ltd.

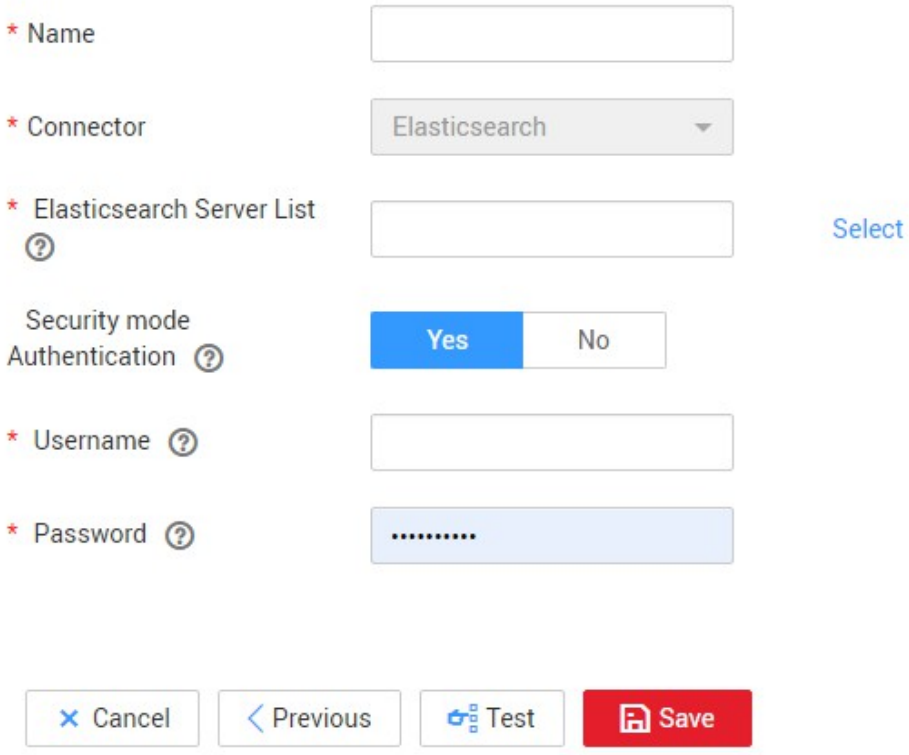
Figure 6-56 Selecting a connector



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

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

Figure 6-57 Creating a Cloud Search Service link 1



 Copyright © Huawei Technologies Co., Ltd.

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

----End

Creating an Oracle Link

Step 1 On the **Link Management** tab page, click **Create Link**. On the page that is displayed, select **Oracle**, click **Next**, and configure the Oracle link parameters.

- **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 will 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.

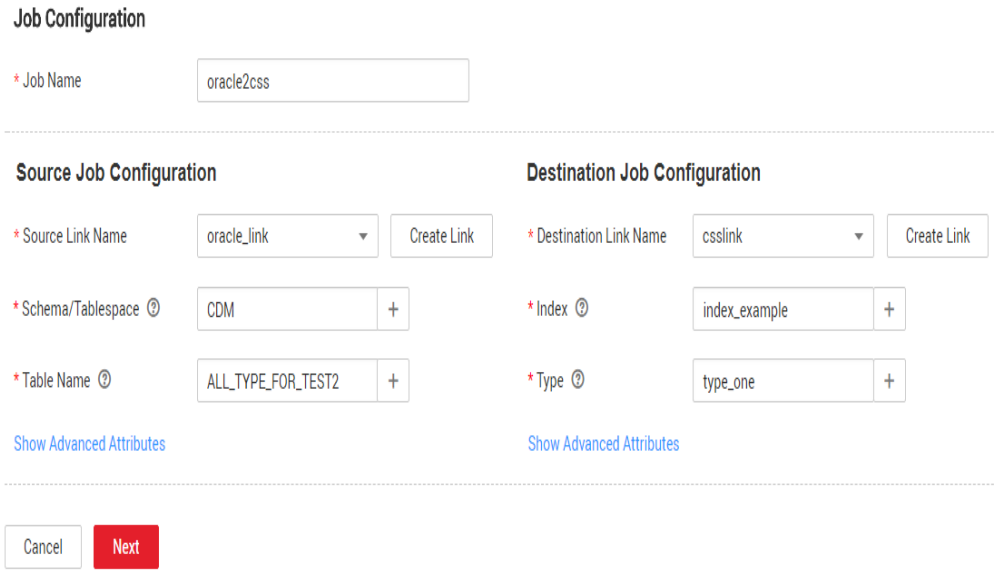
Step 2 Click **Save**. The **Link Management** page is displayed.

----End

Creating a Migration Job

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

Figure 6-58 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**.

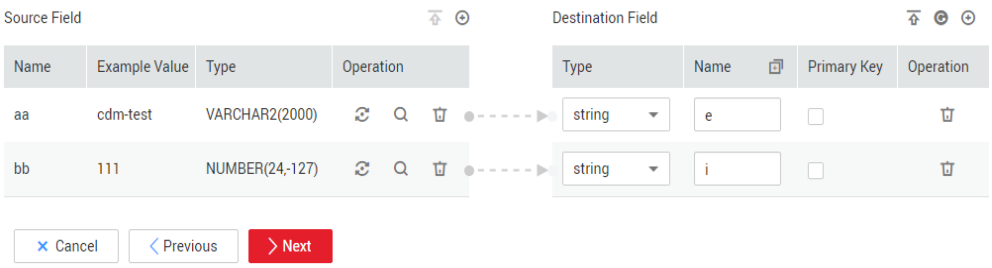
 Copyright © Huawei Technologies Co., Ltd.

  - **Schema/Tablespace**: Enter the name of the database whose data will 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 **Source Relational Database Parameters**.
- **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, or 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, or you can enter a new type. CDM automatically creates a type at the migration destination.
  - Retain the default values for the optional parameters in **Show Advanced Attributes**. For details, see **Destination Elasticsearch or CSS Parameters**.

Step 2 Click **Next**. The **Map Field** page is displayed. CDM automatically matches the source and destination fields. See **Figure 6-59**.

- 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 **Field Conversion**.

Figure 6-59 Field mapping of Cloud Search Service 1



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

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

- **Retry Upon Failure**: If the job fails, you can determine whether to automatically retry the job. Retain the default value **Never**.
- **Group**: Select which group the job belongs to. 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**.

 Copyright © Huawei Technologies Co., Ltd.

- **Concurrent Extractors**: Enter the number of extractors to be executed concurrently. Retain the default value **1**. The maximum number of concurrent extractors in a single job is 500.
- **Write Dirty Data**: Specify this parameter if data that fails to be processed or filtered out during job execution needs to be written to OBS for future viewing. Before writing dirty data, create an OBS link. Retain the default value **No** so that dirty data is not recorded.
- **Delete Job After Completion**: Retain the default value **Do not delete**.

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

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

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

----End

6.1.12.6 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 used for massive data analysis of enterprises.

Hive supports SQL to help users perform extraction, transformation, and loading (ETL) operations on large-scale data sets. Querying large-scale data sets can take 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 HDFS subdirectories. Each subdirectory contains the column names and values of each partition. If there are multiple partitions, there are many HDFS subdirectories. 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 from external data sources (relational databases, object storage services, and file system services) to Hive partition tables.

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

Scenario

Suppose that there is a **trip_data** table in the MySQL database. The table stores cycling records such as the start time, end time, start sites, end sites, and rider IDs. For details about the fields in the **trip_data** table, see **Figure 6-60**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-60 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 includes 5 steps:

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

- You have bought the MRS clusters.
- You have sufficient EIP quota.
- You have obtained the IP address, port number, database name, username, and password for connecting to the MySQL database. In addition, the user must have the read and write permissions for the MySQL database.

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.

 Copyright © Huawei Technologies Co., Ltd.

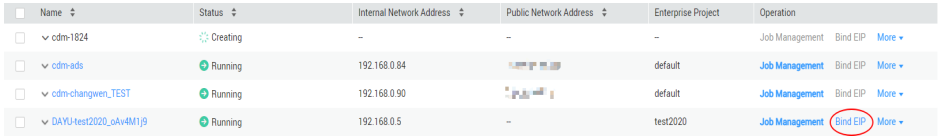
Creating a CDM Cluster and Binding an EIP to the Cluster

Step 1 The DGC basic package includes a CDM cluster. If the cluster cannot meet your requirements, you can buy an incremental package to supplement the DGC basic instance. Create a CDM cluster. For details, see **Buying a DGC Incremental Package**. 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 requirements for most migration scenarios.
- The CDM and MRS clusters must be in the same VPC, subnet, and security group.

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

Figure 6-61 Binding an EIP



NOTE

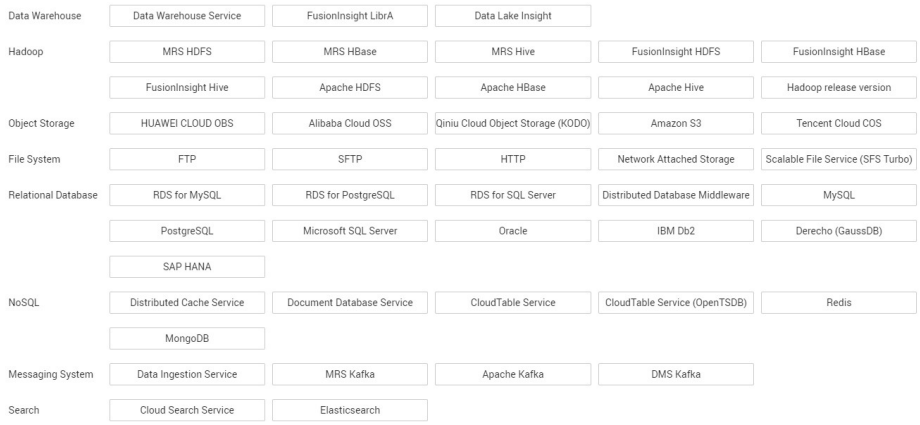
If SSL encryption is configured for the access channel of a local data source, CDM cannot connect to the data source using the EIP.

----End

Creating a MySQL Link

Step 1 On the **Cluster Management** page, click **Job Management** of the cluster and choose **Link Management** > **Create Link** to enter the page for selecting the connector. See **Figure 6-62**.

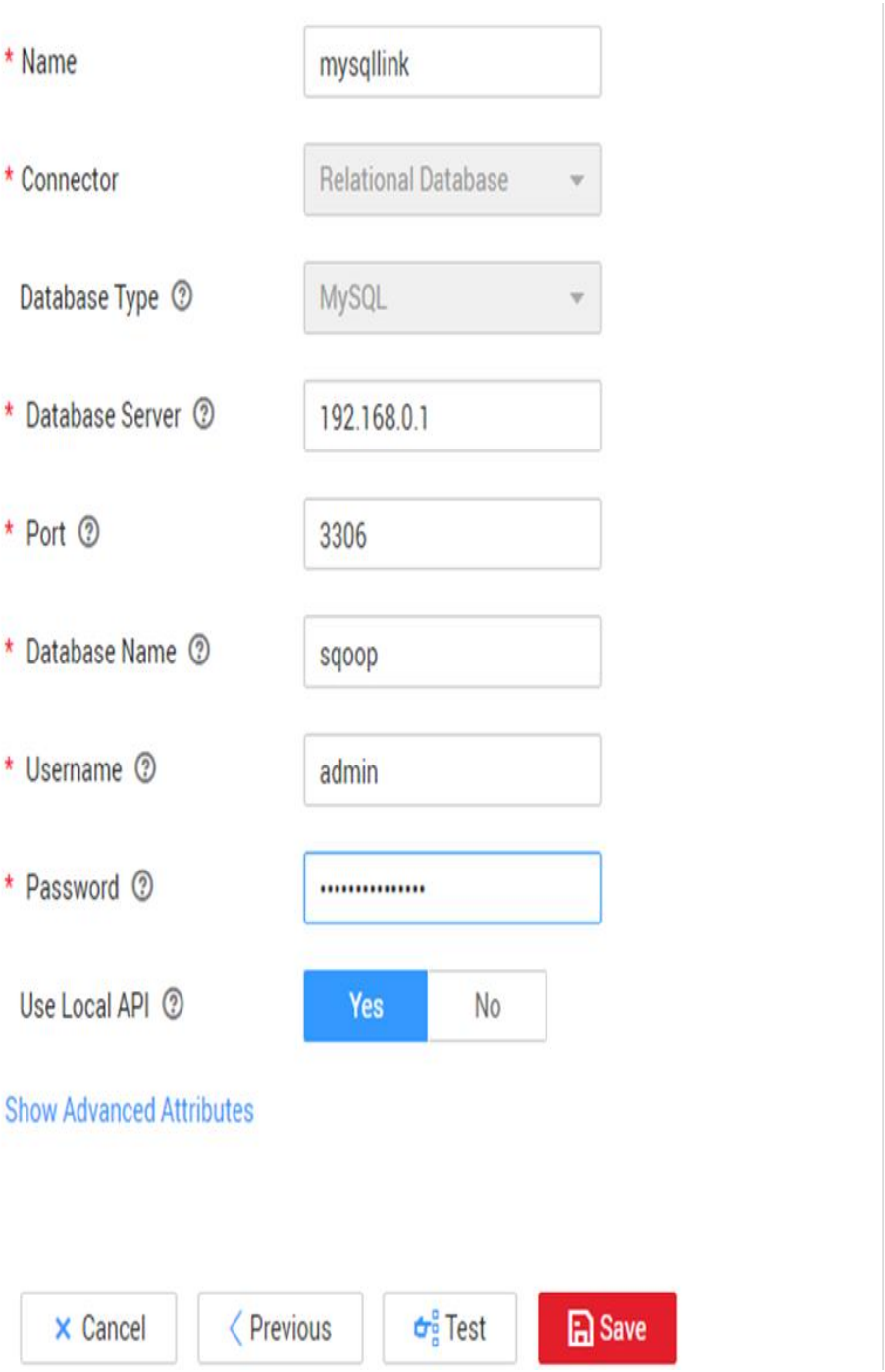
Figure 6-62 Selecting a connector



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

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-63 Creating a MySQL link



Click **Show Advanced Attributes** to display optional parameters. For details, see **Link to Relational Databases**. Retain the default values for the optional parameters and configure the mandatory parameters described in **Table 6-76**.

Table 6-76 MySQL link parameters

Parameter	Description	Example Value
Name	A unique link name	mysqllink

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Database Server	IP address or domain name of the MySQL database	192.168.0.1
Port	MySQL database port	3306
Database Name	The name of the MySQL database	sqoop
Username	A user who has the read, write, and delete permissions on the MySQL database	admin
Password	The password of the user	N/A
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**.	N/A

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

NOTE

If an error occurs when saving the settings, it is likely that the security settings of the MySQL database are incorrect. If this is the case, you may need to enable the EIP of the CDM cluster to access the MySQL database.

----End

Creating a Hive Link

Step 1 Click **Create Link** and select **MRS Hive** to create an MRS Hive link.

Step 2 Click **Next** and configure the MRS Hive link parameters. See **Figure 6-64**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-64 Creating a Hive link

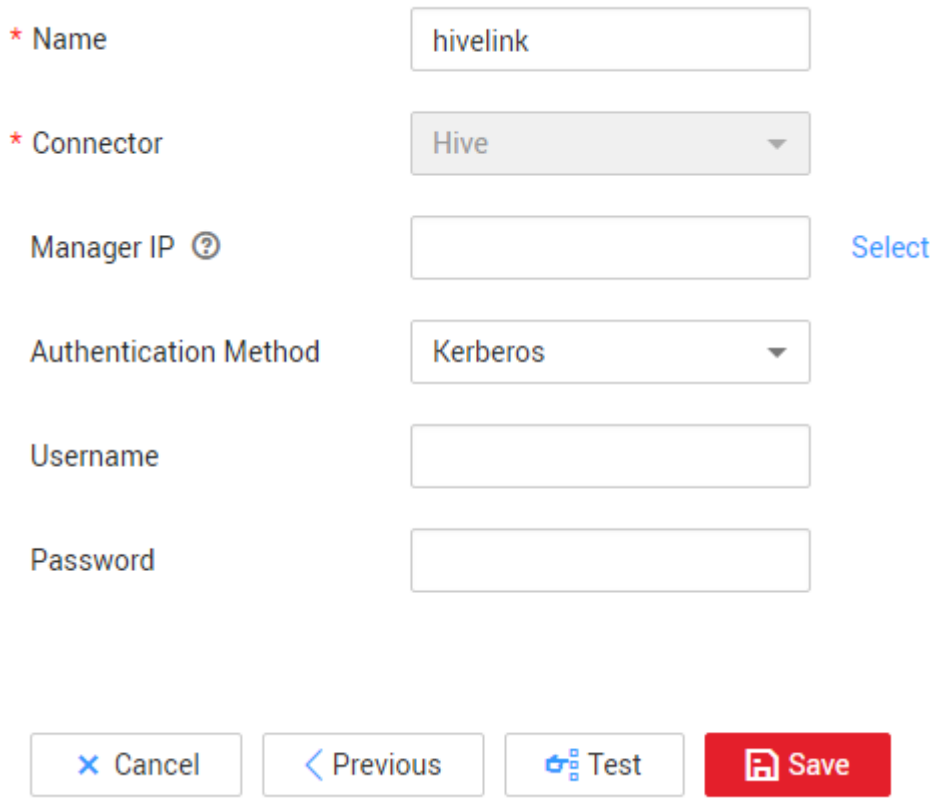


Table 6-77 describes the parameters. Configure these parameters based on your actual situation.

Table 6-77 FusionInsight Hive link parameters

Parameter	Description	Example Value
Name	Link name, which can be defined based on the data source type for easy memorization.	hivelink
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.	127.0.0.1
Manager Port	FusionInsight/MRS Manager port	28443
CAS Server Port	CAS protocol port of FusionInsight/MRS Manager	20009

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Verification Type	Authentication method used for accessing MRS • **Simple**: Select this if FusionInsight HD is in non-security mode. • **Kerberos**: Select this if FusionInsight HD is in security mode.	SIMPLE
Hive Version	Hive version	HIVE_3_X
Username	When **Authentication Method** is set to **Kerberos**, set the username and password for logging in to MRS Manager.	cdm
Password	Password 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
Running Mode	This parameter is used only when the Hive version is **HIVE_3_X**. Possible values are: • **EMBEDDED**: The link instance runs with CDM. This mode delivers better performance. • **STANDALONE**: The link instance runs in an independent process. If CDM needs to connect to multiple Hadoop data sources (MRS, CloudTable, or Hadoop) 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 data 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.	EMBEDDED

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

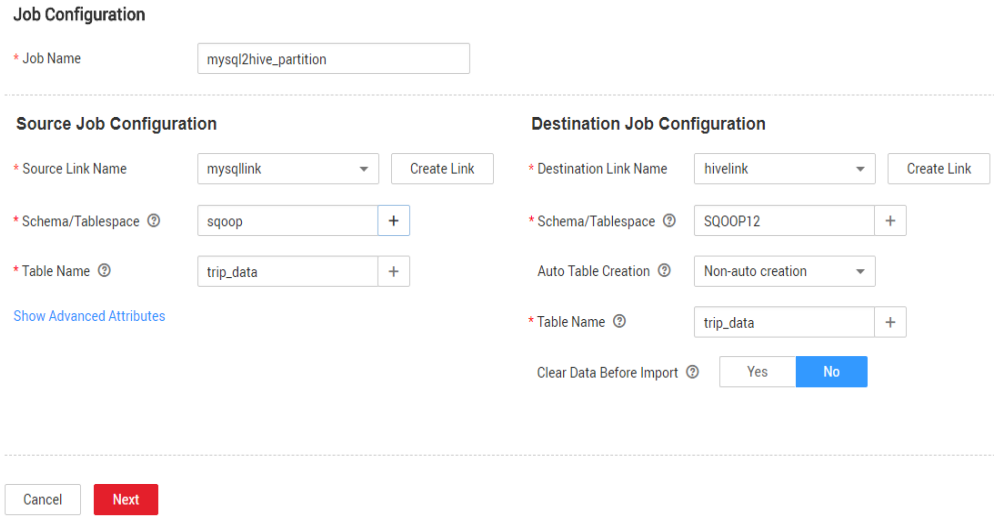
----End

Creating a Migration Job

Step 1 Choose **Table/File Migration** > **Create Job** to create a data migration job. **Figure 6-65** illustrates how to create a migration job.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-65 Creating a job for migrating data from MySQL to Hive



NOTE

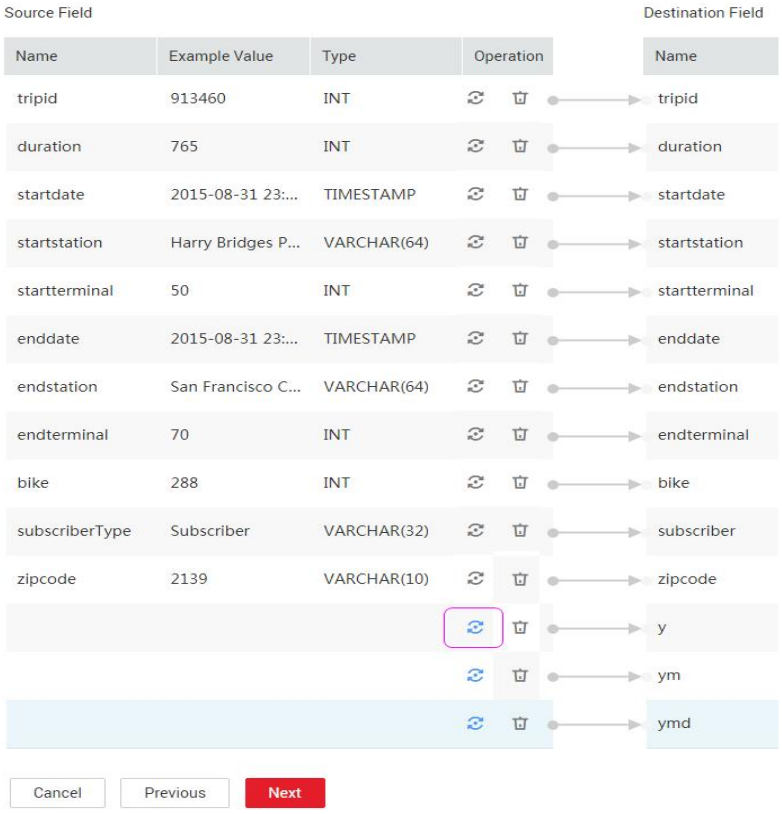
Set **Clear Data Before Import** to **Yes**, so that the data in the Hive table will be cleared before data import.

Step 2 After the parameters are configured, click **Next**. The **Map Field** tab page is displayed. See **Figure 6-66**.

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. These 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.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-66 Hive field mapping



Step 3 Click to display the **Converter List** dialog box, and then choose **Create Converter** > **Expression conversion**. See **Figure 6-67**.

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

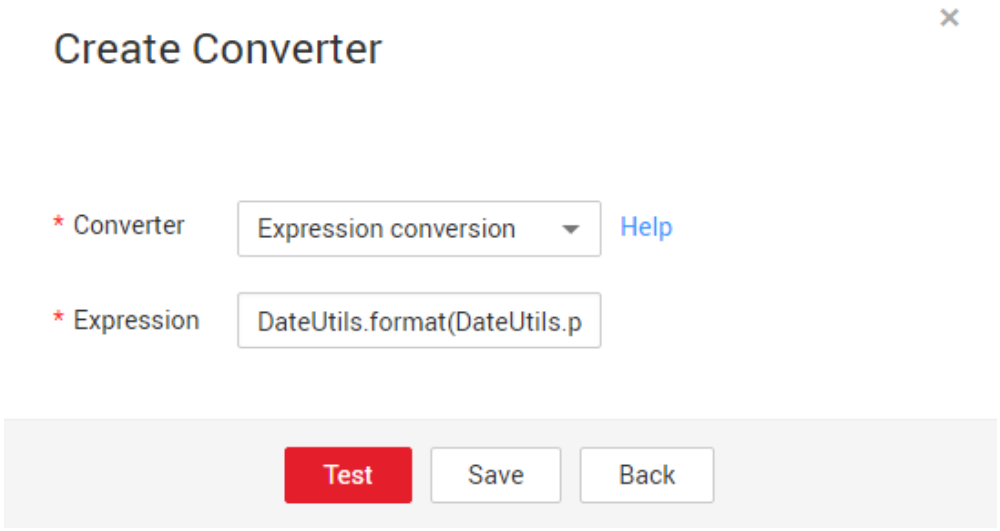
DateUtils.format(DateUtils.parseDate(row[2],"yyyy-MM-dd HH:mm:ss.SSS"),"yyyy")

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")

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-67 Configuring the expression



NOTE

The expressions in CDM support field conversion of common character strings, dates, and values. For details, see **Field Conversion**.

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

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

- **Retry Upon Failure**: If the job fails, you can determine whether to automatically retry the job. Retain the default value **Never**.
- **Group**: Select which group the job belongs to. 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 executed concurrently. Retain the default value **1**. The maximum number of concurrent extractors in a single job is 500.
- **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

 Copyright © Huawei Technologies Co., Ltd.

6.1.12.7 Migrating Data from OSS to OBS

Scenario

CDM allows you to directly migrate object storage data from a third-party cloud to OBS without forwarding or writing code.

This section describes how to use CDM to migrate files from OSS on Alibaba Cloud to OBS. The procedure is as follows:

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

Preparing Data

- Endpoint for accessing OSS, for example, oss-cn-hangzhou.aliyuncs.com
- AK, temporary credential, or security token for accessing OSS
- Domain name, port number, AK, and SK for accessing OBS

Creating a CDM Cluster and Binding an EIP to the Cluster

Step 1 The DGC basic package includes a CDM cluster. If the cluster cannot meet your requirements, you can buy an incremental package to supplement the DGC basic instance. Create a CDM cluster. For details, see **Buying a DGC Incremental Package**. The key configurations are as follows:

- Select the **cdm.medium** instance, which is applicable to most migration scenarios.
- If the cluster is used only to migrate data from third-party data sources to OBS on HUAWEI CLOUD, there is no special requirements on the VPC, subnet, and security group of the CDM cluster. You can specify them based on your needs.

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 accesses Alibaba Cloud OSS through the public network.

Because data is imported to HUAWEI CLOUD, 5 Mbit/s bandwidth for the EIP is enough.

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.

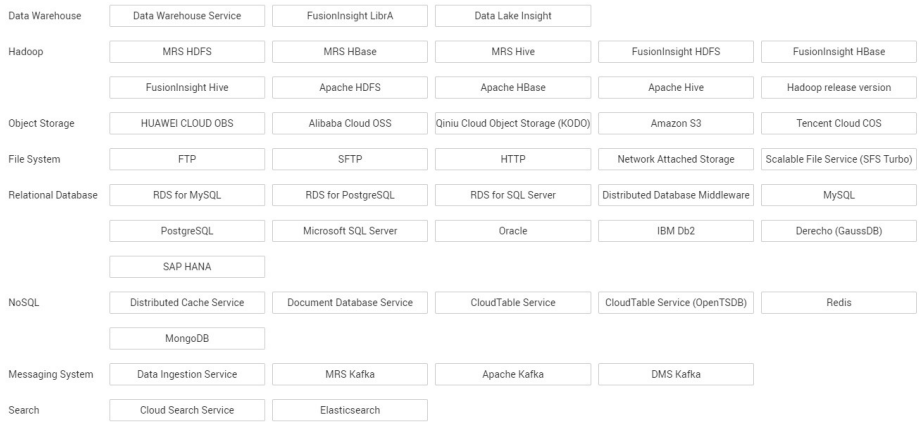
----End

Creating an OBS Link

Step 1 Click **Job Management** in the **Operation** column of the CDM cluster. On the page that is displayed, choose **Link Management** > **Create Link**. The page for selecting a connector is displayed. See **Figure 6-68**.

 Copyright © Huawei Technologies Co., Ltd.

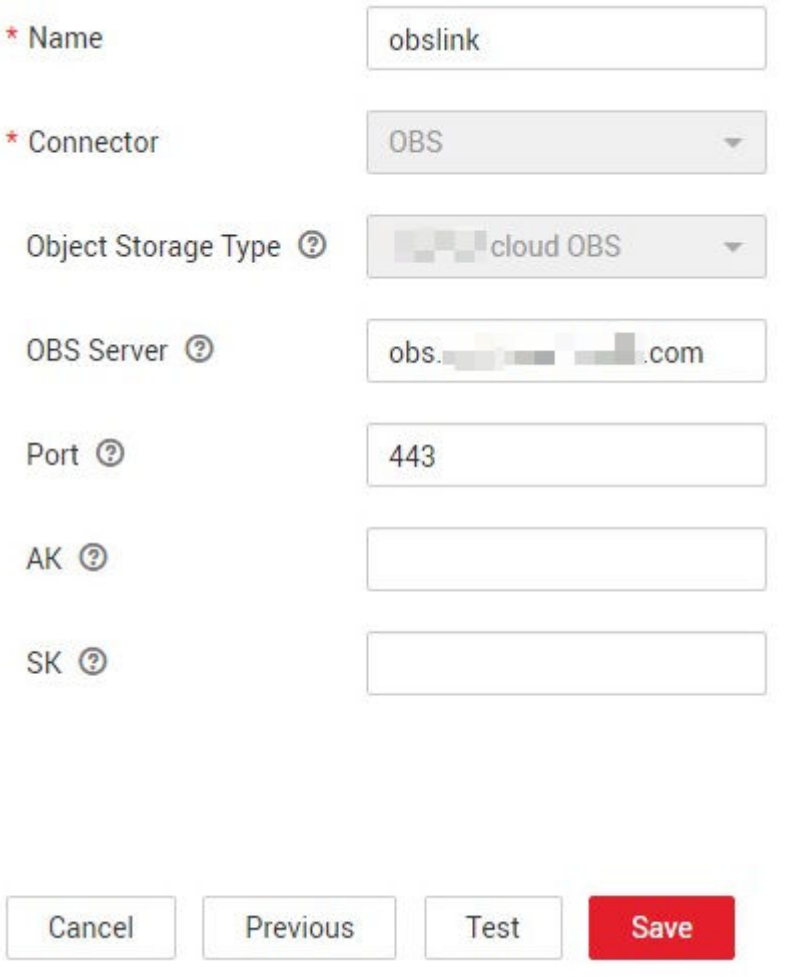
Figure 6-68 Selecting a connector



Step 2 Select **Object Storage Service** and click **Next**. On the page that is displayed, configure the OBS link parameters.

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

Figure 6-69 Creating an OBS link



 Copyright © Huawei Technologies Co., Ltd.

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

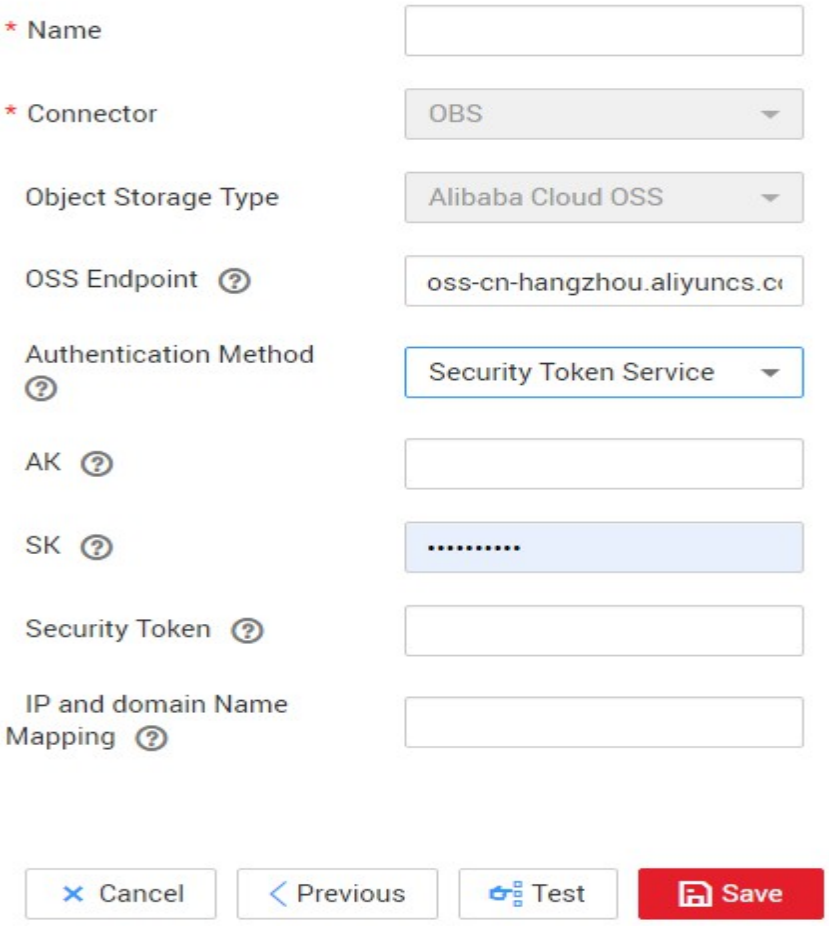
----End

Creating an OSS Link

Step 1 On the **Link Management** tab page, click **Create Link**. On the page that is displayed, select **Alibaba Cloud OSS**, click **Next**, and configure the required link parameters. See **Figure 6-70**.

- **Name**: Enter a custom link name.
- **OSS Endpoint**: Configure the endpoint of the data to be migrated. Select the endpoint based on the region where the bucket resides.
- **Authentication Method**: Select an authentication method based on your needs, for example, **Access key**.
- **AK** and **SK**: Enter the AK and SK used for logging in to OSS.
- **IP and domain Name Mapping**: Configure the mapping between IP addresses and domain names.

Figure 6-70 Creating an OSS link



Step 2 Click **Save**. The **Link Management** page is displayed.

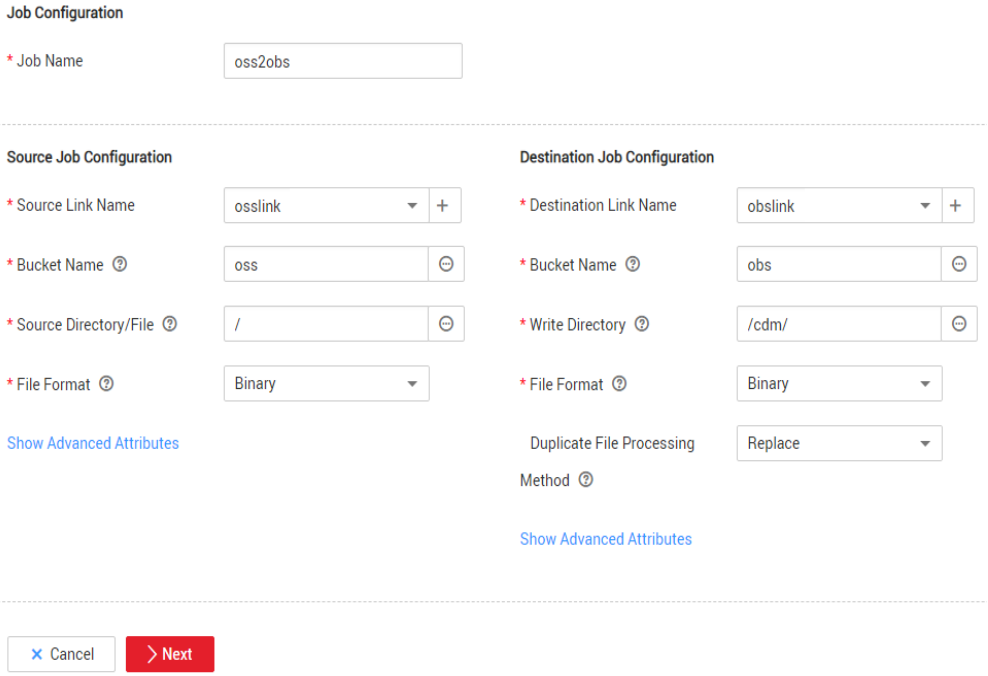
----End

 Copyright © Huawei Technologies Co., Ltd.

Creating a Migration Job

Step 1 Choose **Table/File Migration** > **Create Job** to create a job for migrating data from OSS to OBS. See **Figure 6-71**.

Figure 6-71 Creating a job for migrating data from OSS to OBS



- **Job Name**: Enter a custom job name.
- **Source Job Configuration**
  - **Source Link Name**: Select the **osslink** link created in **Creating an OSS 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. You can migrate all files in the bucket.
  - **File Format**: Select **Binary**. It is applicable to file copy. To write files to databases, select **CSV** or **JSON**.
  - Retain the default values of the optional parameters in **Show Advanced Attributes**. For details, see **Source OBS Parameters**.
- **Destination Job Configuration**
  - **Destination Link Name**: Select the **obslink** link created in **Creating an OBS Link**.
  - **Bucket Name**: Select the bucket to which data is to be written.
  - **Write Directory**: Select the path for storing data.
  - **File Format**: Select **Binary**. The value must be the same as that on the migration source.
  - Retain the default values of other optional parameters. For details, see **Destination OBS Parameters**.

 Copyright © Huawei Technologies Co., Ltd.

Step 2 Click **Next** and set task parameters. Generally, you can retain the default values for all parameters.

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

- **Retry Upon Failure**: If the job fails, you can determine whether to automatically retry the job. Retain the default value **Never**.
- **Group**: Select which group the job belongs to. 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 executed concurrently. Retain the default value **1**. The maximum number of concurrent extractors in a single job is 500.
- **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 3 Click **Save and Run**. The **Job Management** page is displayed, on which you can view the job execution progress and result.

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

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

----End

6.1.12.8 Migrating Data from OSS to SFS

Scenario

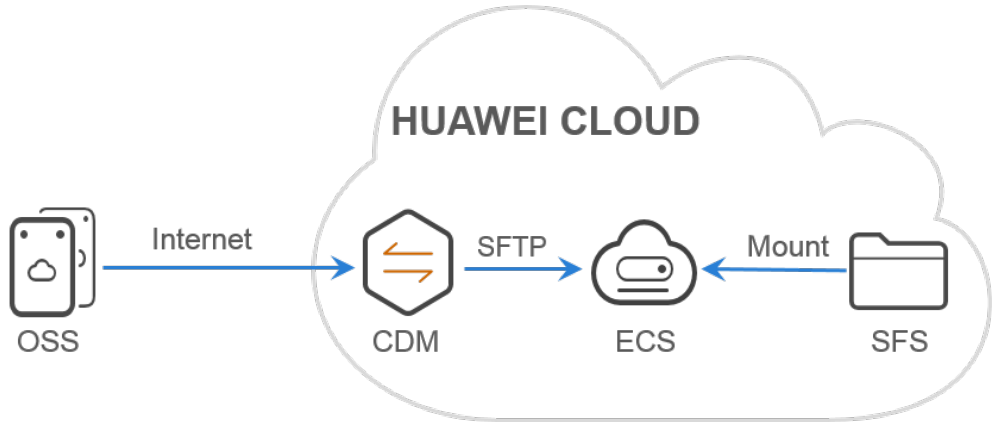
SFS provides fully hosted file storage with complete share capabilities for ECSs. Compatible with the Network File System protocol, SFS is expandable to the petabyte range while ensuring top performance in data-intensive and bandwidth-intensive applications.

CDM allows you to directly migrate object storage data from a third-party cloud to SFS on HUAWEI CLOUD without forwarding or writing code.

This section describes how to use CDM to migrate data from OSS on Alibaba Cloud to SFS on HUAWEI CLOUD.

Copyright © Huawei Technologies Co., Ltd.

Figure 6-72 Migration principle



Mount the SFS file system to a directory of an ECS or your HPC cluster. CDM accesses the directory through the SFTP port. The procedure is as follows:

1. **Mounting the SFS File System to the ECS**
2. **Creating a CDM Cluster and Binding an EIP to the Cluster**
3. **Creating an SFTP Link**
4. **Creating an OSS Link**
5. **Creating a Migration Job**

NOTE

- The CDM cluster, SFS file system, and ECS to which the SFS file system is mounted must be in the same VPC.
- Currently, the SFS file system can be mounted to only the Linux ECS through the NFSv3 protocol. Therefore, the ECS must run on a Linux system.

Preparing Data

- Endpoint for accessing OSS, for example, oss-cn-hangzhou.aliyuncs.com
- AK, temporary credential, or security token for accessing OSS
- Intranet access IP address, SFTP port number, username, and password for accessing the ECS

Mounting the SFS File System to the ECS

Run the following command on the ECS to mount the SFS file system to the ECS. For details, see **Mounting a File System to ECSs (Linux)**.

```
cd /
mkdir sfs
mount -t nfs -o vers=3,timeo=600,nolock Shared path Local path
```

If the error message "**wrong fs type**" is displayed during the mount, the **nfs-utils** package is missing. Run the **yum install –y nfs-utils** command to install the package.

 Copyright © Huawei Technologies Co., Ltd.

Creating a CDM Cluster and Binding an EIP to the Cluster

Step 1 The DGC basic package includes a CDM cluster. If the cluster cannot meet your requirements, you can buy a CDM incremental package to supplement the DGC basic instance. Create a CDM cluster. For details, see **Buying a DGC Incremental Package**. The key configurations are as follows:

- Select the instance type based on the amount of data to be migrated. If the data volume is large, you are advised to select a cluster with a large size, for example, **cdm.large**.
- The CDM cluster must be in the same VPC as the ECS and SFS file system. The subnet and security group rules ensure that they can communicate with each other.

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 accesses Alibaba Cloud OSS through the public network.

Data is uploaded to the cloud. You will not be billed for the uplink traffic. The EIP bandwidth is 10 Mbit/s.

NOTE

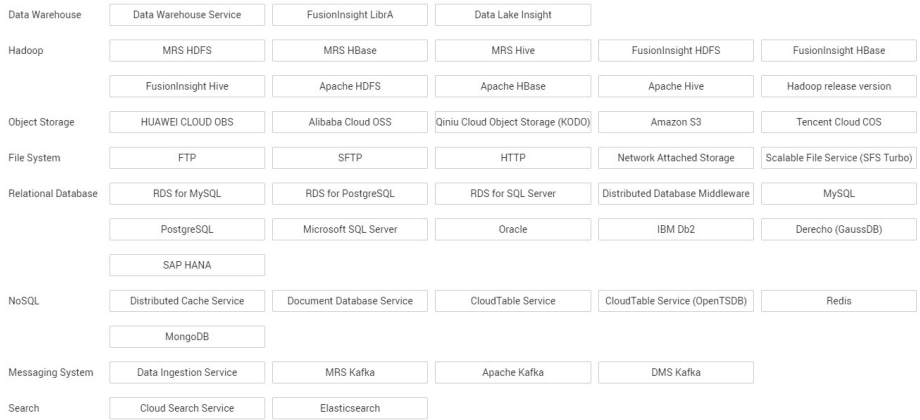
If SSL encryption is configured for the access channel of a local data source, CDM cannot connect to the data source using the EIP.

----End

Creating an SFTP Link

Step 1 Click **Job Management** in the **Operation** column of the CDM cluster. On the page that is displayed, choose **Link Management** > **Create Link**. The page for selecting a connector is displayed. See **Figure 6-73**.

Figure 6-73 Selecting a connector



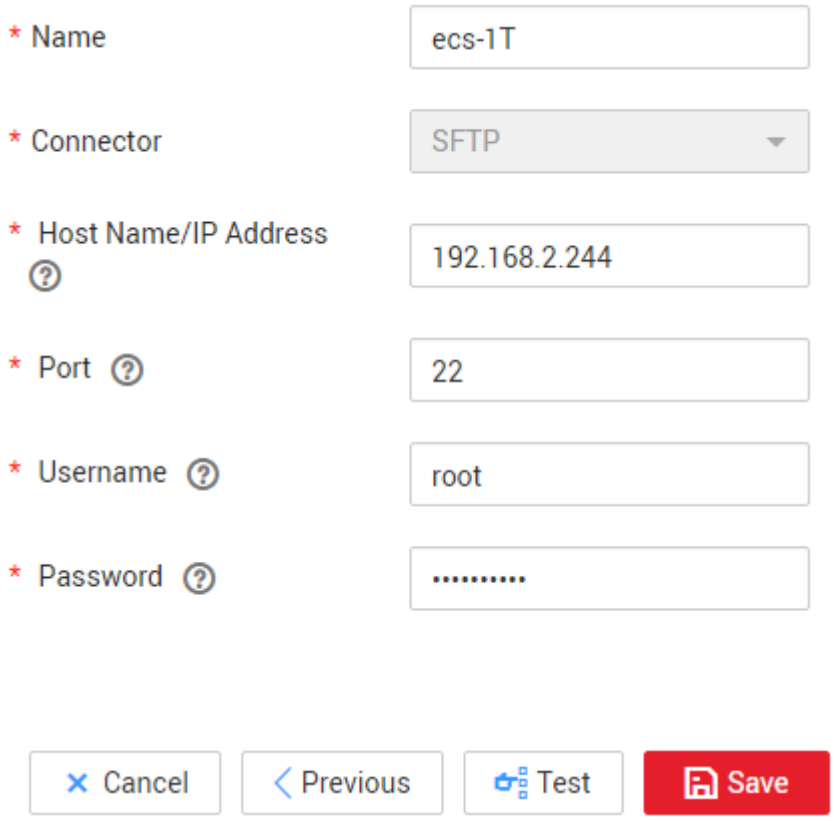
Step 2 Select **SFTP** and click **Next**. On the page that is displayed, configure the SFTP link parameters. See **Figure 6-74**.

- **Name**: Enter a link name.
- **Host Name/IP Address** and **Port**: Set these parameters to the intranet access IP address and port number of the ECS.

 Copyright © Huawei Technologies Co., Ltd.

- **Username** and **Password**: Enter the username and password for accessing the ECS.

Figure 6-74 Creating an SFTP link



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

----End

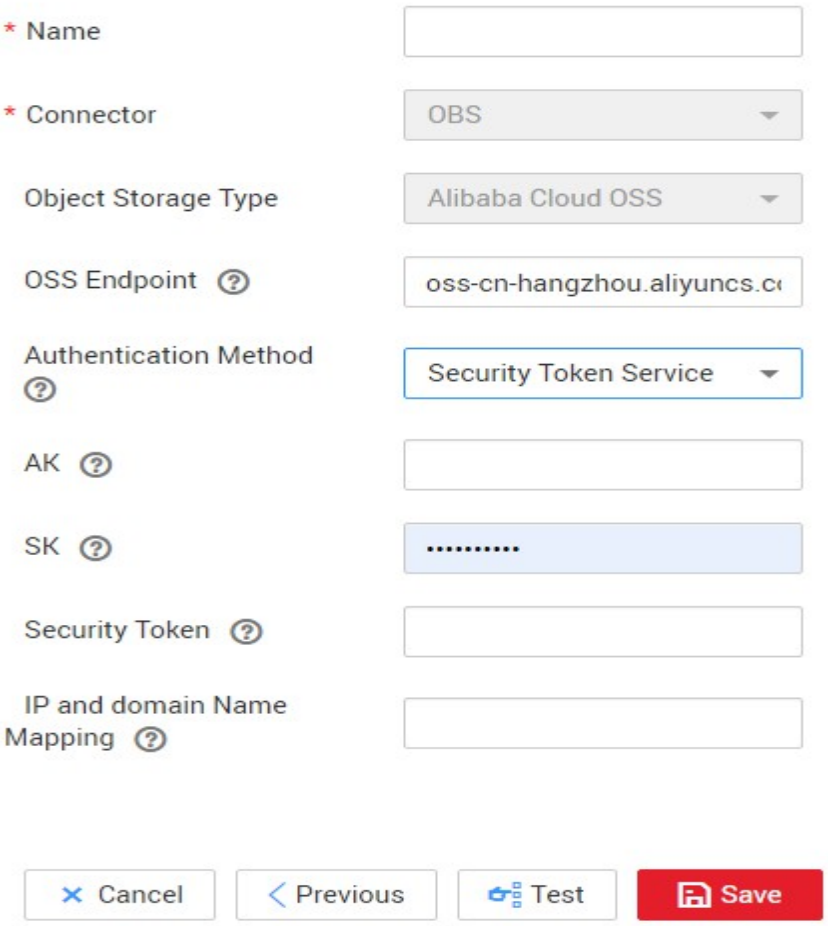
Creating an OSS Link

Step 1 On the **Link Management** tab page, click **Create Link**. On the page that is displayed, select **Alibaba Cloud OSS**, click **Next**, and configure the required link parameters. See **Figure 6-75**.

- **Name**: Enter a custom link name.
- **OSS Endpoint**: Configure the endpoint of the data to be migrated. Select the endpoint based on the region where the bucket resides.
- **Authentication Method**: Select an authentication method based on your needs, for example, **Access key**.
- **AK** and **SK**: Enter the AK and SK used for logging in to OSS.
- **IP and domain Name Mapping**: Configure the mapping between IP addresses and domain names.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-75 Creating an OSS link



Step 2 Click **Save**. The **Link Management** page is displayed.

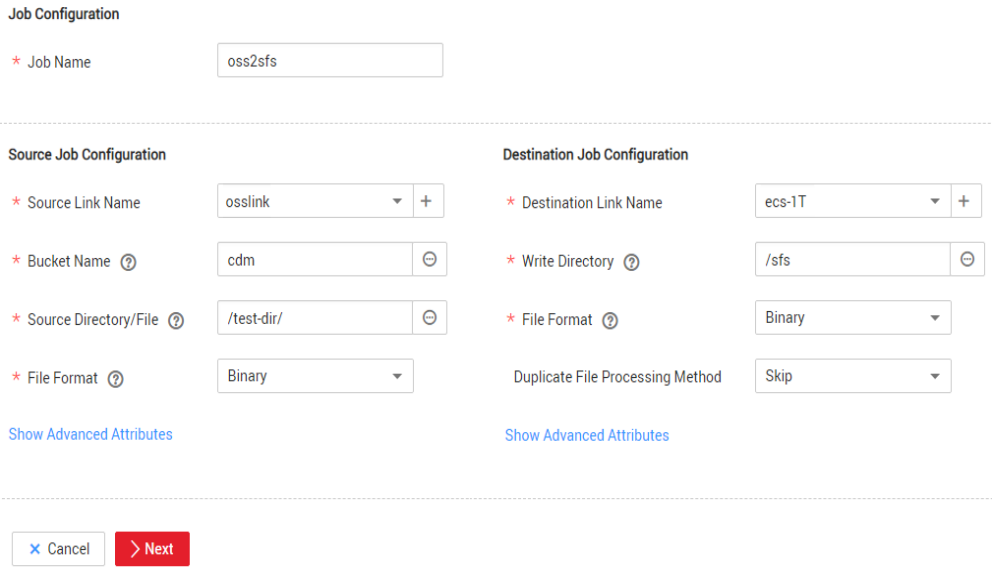
----End

Creating a Migration Job

Step 1 Choose **Table/File Migration** > **Create Job** to create a job for migrating data from OSS to SFS. See **Figure 6-76**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-76 Creating a job for migrating data from OSS to SFS



- **Job Name**: Enter a job name.
- **Source Job Configuration**
  - **Source Link Name**: Select the OSS link created in **Creating an OSS Link**.
  - **Bucket Name**: Enter the name of an OSS bucket.
  - **Source Directory/File**: Enter an object name or prefix. If you enter an object prefix, all objects with the prefix are migrated.
  - **File Format**: Select **Binary**. This setting is used to copy files. To write files to databases, select **CSV** or **JSON**.
  - Retain the default values of the optional parameters in **Show Advanced Attributes**. For details, see **Source OBS Parameters**.
- **Destination Job Configuration**
  - **Destination Link Name**: Select the SFTP link created in **Creating an SFTP Link**.
  - **Write Directory**: Enter the directory of the ECS the SFS file system is mounted to.
  - **File Format**: Select **Binary**. The value must be the same as that on the migration source.
  - **Duplicate File Processing Method**: Select **Skip**. CDM compares files before migrating them. If a file with the same name and size exists on the migration source and destination, CDM determines that the file is a duplicate file and skips it during the migration.
  - Retain the default values of other optional parameters. For details, see **Destination FTP/SFTP/NAS/SFS Parameters**.

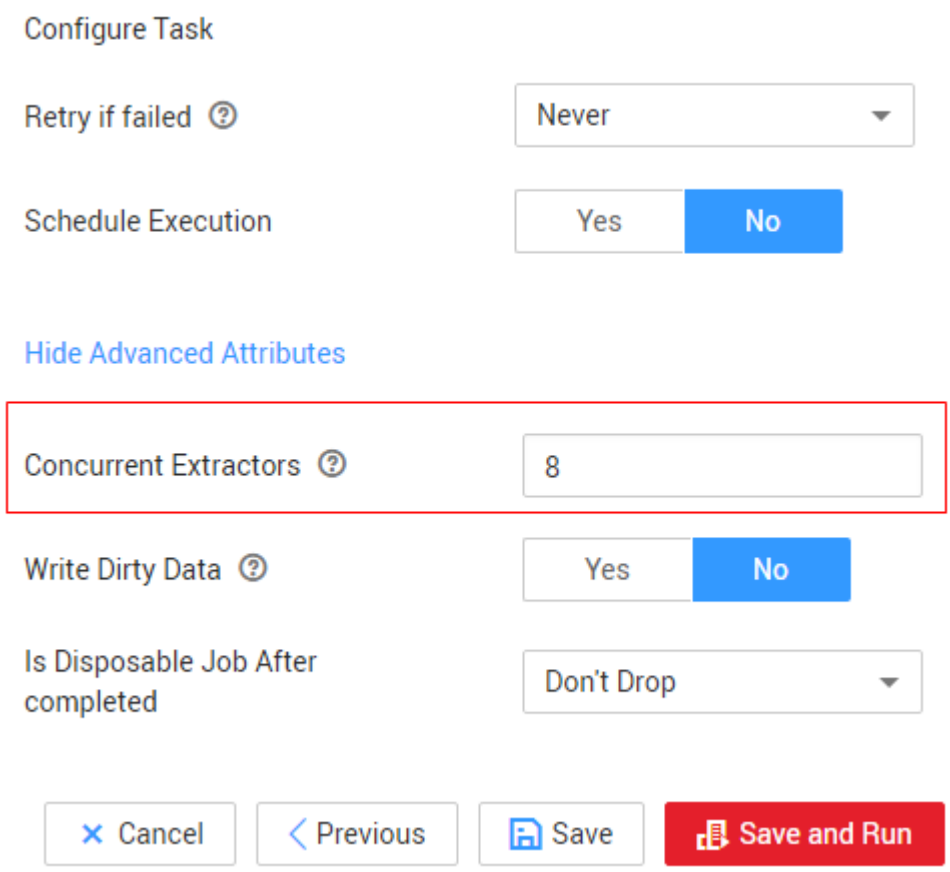
Step 2 Click **Next** and set task parameters. When migrating data from OSS to SFS, you are advised to set **Concurrent Extractors** to **8**, which indicates the number of concurrent jobs. If multiple files are being migrated, modify **Concurrent Extractors** to improve the migration performance. Retain the default values of other parameters. Retain the default values of other parameters.

 Copyright © Huawei Technologies Co., Ltd.

If you increase the value of **Concurrent Extractors**, you must also increase the maximum number of SSH connections on the ECS.

1. Change the value of **#MaxStartups 10:30:100** in **/etc/ssh/sshd_config** to **MaxStartups 1000**.
2. Change **#MaxSessions 10** in **/etc/ssh/sshd_config** to **MaxSessions 1000**.
3. Run the **service sshd restart** command to restart the **sshd** process.

Figure 6-77 Concurrent Extractors



Step 3 Click **Save and Run**. The **Job Management** page is displayed, on which you can view the job execution progress and result.

Step 4 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

 Copyright © Huawei Technologies Co., Ltd.

6.1.12.9 Migrating Data from FTP/SFTP to OBS

Scenario

CDM can periodically upload new files to OBS. You do not need to compile code or manually upload the files. You can use the massive storage capabilities of OBS directly to back up files.

This section describes how to periodically back up FTP files to OBS with CDM.

For example, the **to_obs_test** directory on the FTP server contains one subdirectory **another_dir** and two files **file1** and **file2**. **file2** is in the **another_dir** directory. **Figure 6-78** shows the files. Configure a scheduled job of CDM to transfer these files to OBS and add **file3** and **file4** to the directory to verify that CDM can periodically transfer new files to OBS.

Figure 6-78 Files on the FTP server

```
to_obs_test/:
drwx------ 2 ftptest users 4096 Nov  9 11:54 another_dir
-rw------- 1 ftptest users 5933 Nov  9 11:50 file1.rar
to_obs_test/another_dir:
-rw------- 1 ftptest users 2199050 Nov  9 11:54 file2.zip
```

Prerequisites

- Your EIP quota is sufficient.
- You have created an OBS bucket and obtained the access key (AK and SK).
- You have obtained the IP address, username, and password of the FTP server.
- If the FTP server is in the on-premises environment, ensure the FTP server must be accessible through the public network, 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 The DGC basic package includes a CDM cluster. If the cluster cannot meet your requirements, you can buy an incremental package to supplement the DGC basic instance. Click **Buy CDM** to create a CDM cluster. The key configurations are as follows:

- Select the **cdm.medium** instance, which is user for most migration scenarios.
- If the cluster is only used to migrate data from third-party data sources to OBS on HUAWEI CLOUD, there are no special requirements for the VPC, subnet, and security group of the CDM cluster. You can specify them based on your needs.

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 FTP server.

 Copyright © Huawei Technologies Co., Ltd.

NOTE

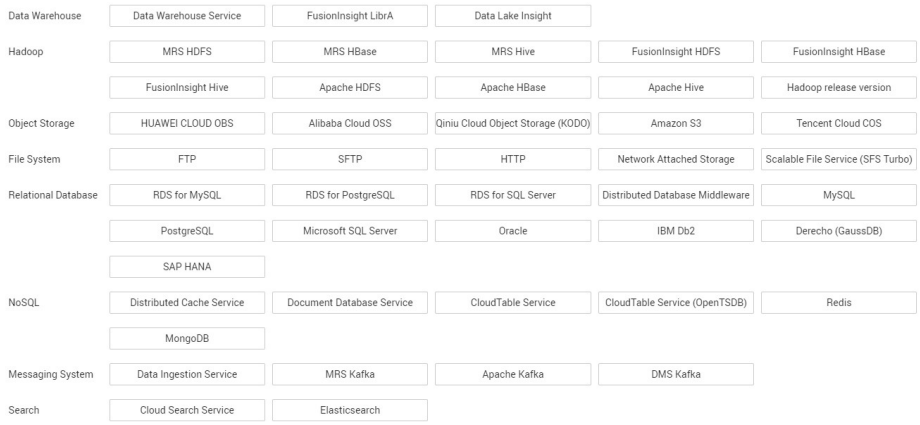
If SSL encryption is configured for the access channel of a local data source, CDM cannot connect to the data source using the EIP.

----End

Creating an OBS Link

Step 1 Click **Job Management** in the **Operation** column of the CDM cluster. On the page that is displayed, choose **Link Management** > **Create Link**. The page for selecting a connector is displayed. See **Figure 6-79**.

Figure 6-79 Selecting a connector

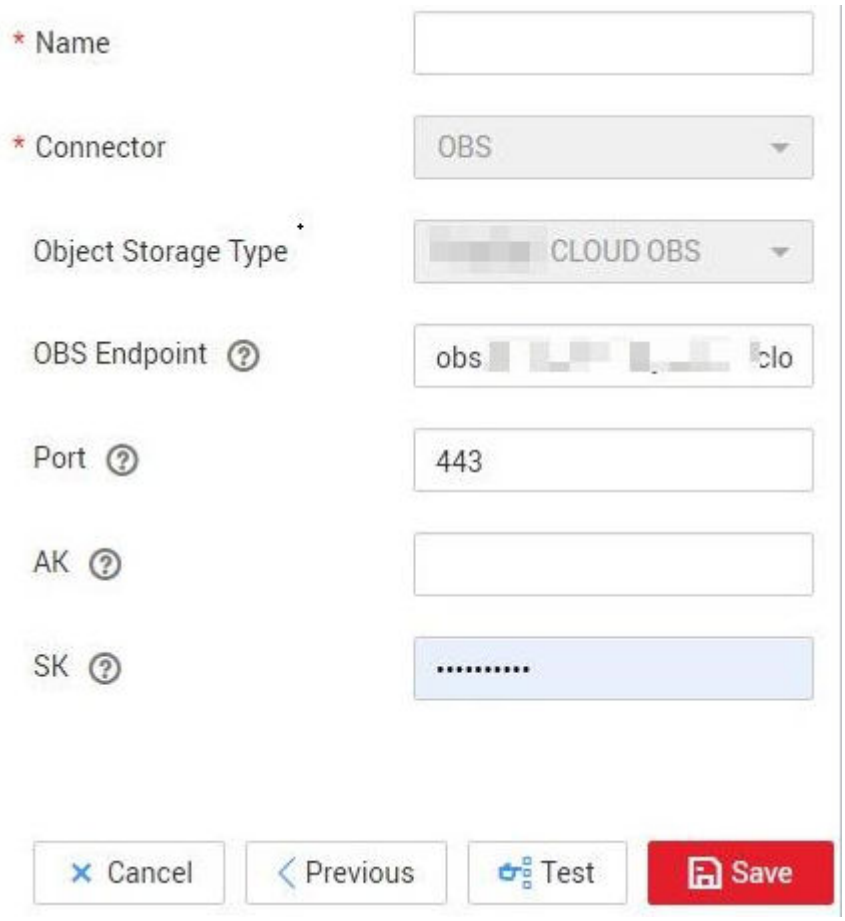


Step 2 Select **Object Storage Service** and click **Next**. On the page that is displayed, configure the OBS link parameters.

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

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-80 Creating an OBS link



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

----End

Creating an FTP Link

Step 1 On the **Link Management** tab page, click **Create Link**. On the page that is displayed, select **FTP**, click **Next**, and configure the FTP link parameters.

- **Name**: Enter a custom link name, for example, **ftplink**.
- **Host Name/IP Address** and **Port**: Enter the username and password used for logging in to the FTP server.
- **Username** and **Password**: Enter the username and password used for logging in to the FTP server.

Step 2 Click **Save**. The **Link Management** page is displayed.

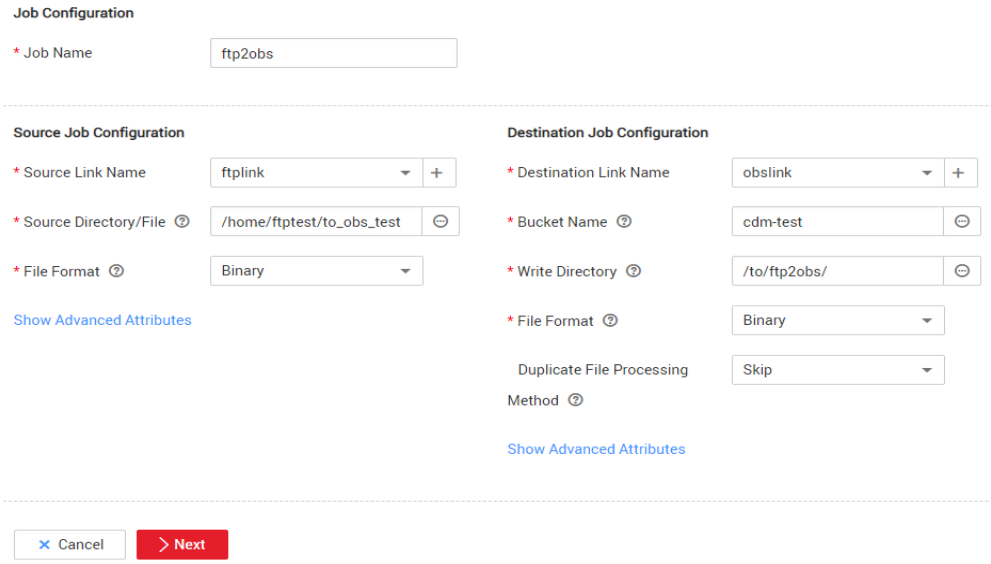
----End

Creating a Scheduled Migration Job

Step 1 Choose **Table/File Migration** > **Create Job** to create a data migration job.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-81 Creating a job for migrating data from FTP to OBS

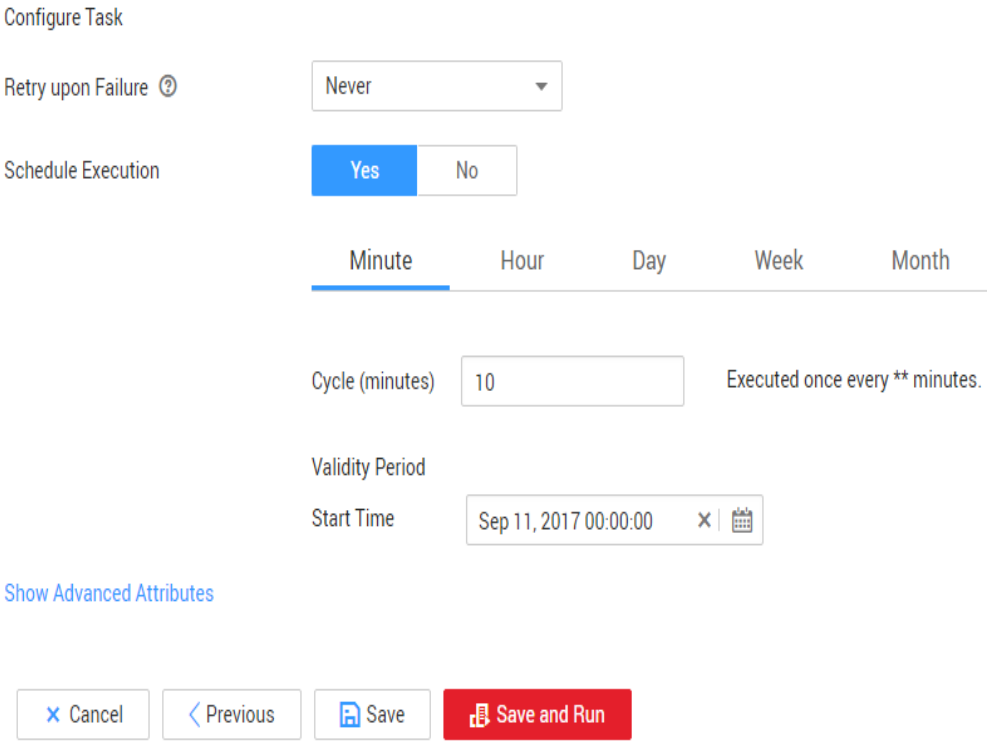


- **Job Name**: Enter a job name.
- **Source Link Name**: Select the **ftplink** link created in **Creating an FTP Link**.
  - **Source Directory/File**: Select the path where **to_obs_test** is located.
  - **File Format**: Select **Binary**. This setting is used to copy files. To write files to a database, select **CSV** or **JSON**.
- **Destination Link Name**: Select the **obslink** link created in **Creating an OBS Link**.
  - **Bucket Name**: Select the OBS bucket for storing FTP files.
  - **Write Directory**: Select an existing directory or enter one manually. If the entered directory does not exist, CDM automatically creates one, for example, **/to/ftp2obs/**.
  - **File Format**: Select **Binary**. The value must be the same as that on the migration source.
  - **Duplicate File Processing Method**: Select **Skip** to avoid transferring duplicate files.

Step 2 Click **Next** and configure the scheduled task. In this example, with the settings shown in the above figure, the scheduled task is executed every 10 minutes. Retain the default values of other parameters.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-82 Scheduling job execution



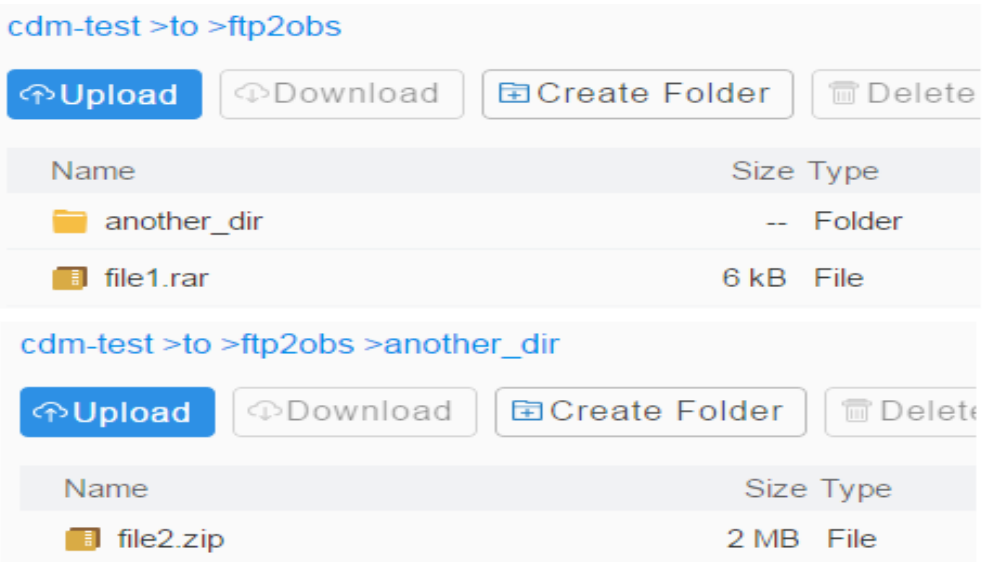
Step 3 Click **Save and Run**.

----End

Verifying Backup

Step 1 After the job is executed successfully, log in to the OBS client, and you can see the files on OBS. **Figure 6-83** shows an example.

Figure 6-83 Files on the OBS client



 Copyright © Huawei Technologies Co., Ltd.

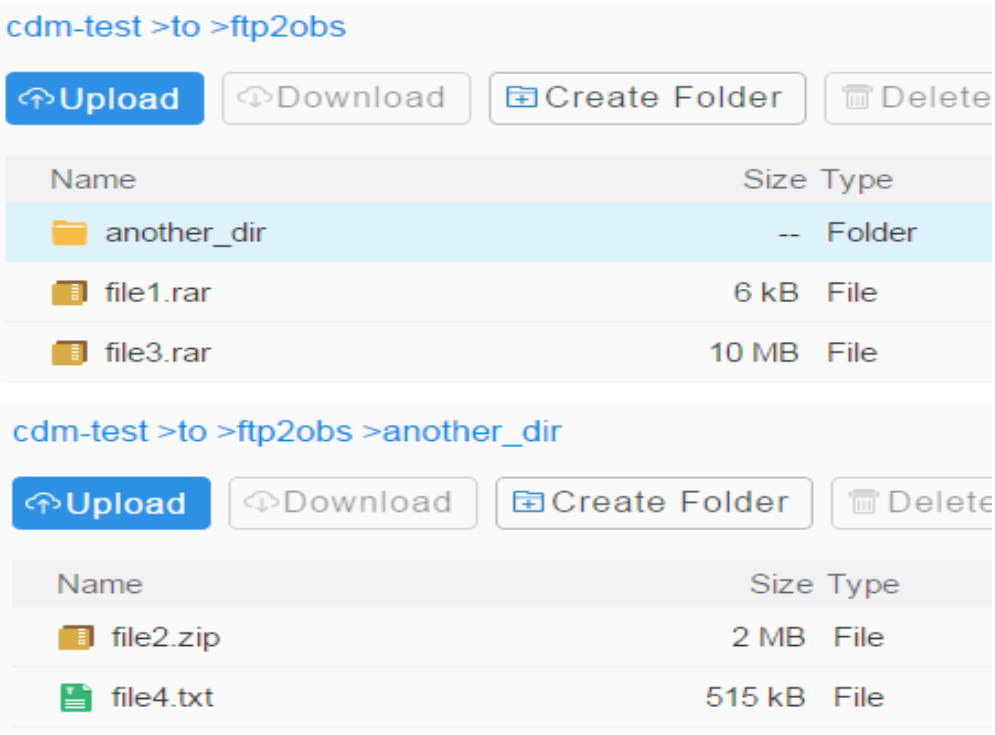
Step 2 In the FTP server directories, add files **file3** and **file4**. **file3** and **file1** in the figure are in the same directory, and **file2** and **file4** are in another directory. See **Figure 6-84**.

Figure 6-84 New files on the FTP server

```
to_obs_test/:
total 10356
drwx------ 2 ftptest users     4096 Nov  9 12:18 another_dir
-rw------- 1 ftptest users     5933 Nov  9 11:50 file1.rar
-rw------- 1 ftptest users 10575867 Nov  9 12:18 file3.rar

to_obs_test/another_dir:
total 2672
-rw------- 1 ftptest users 2199050 Nov  9 11:54 file2.zip
-rw------- 1 ftptest users  526921 Nov  9 12:18 file4.txt
```

Step 3 Wait 10 minutes and CDM automatically triggers the scheduled job. Then you can view the new files **file3** and **file4** after logging in to OBS. **Figure 6-85** shows the new files on OBS.

Figure 6-85 New files on OBS



Step 4 On the **Job Management** page, click **Historical Record** in the **Operation** column to view the job's historical execution records and read/write statistics.

Step 5 Click **Log** to view the job logs.

----End

 Copyright © Huawei Technologies Co., Ltd.

6.1.12.10 Performing Incremental Migration on CDM by Data Development

CDM supports incremental migration. For details, see **Incremental File Migration**, **Incremental Migration of Relational Databases**, **HBase/CloudTable Incremental Migration**, and **Incremental Synchronization Using the Macro Variables of Date and Time**.

Data Development is a one-stop big data collaboration development platform. With DLF's online script editing, CDM migration jobs can be scheduled to implement incremental migration.

This section describes how Data Development helps CDM migrate data from DWS to OBS. The procedure is as follows:

1. **Viewing a Job JSON File**
2. **Modifying a Job JSON File**
3. **Creating and Running a CDM Job on the Data Development Module**
4. **Waiting Until the Job Execution Is Completed**
5. **(Optional) Deleting a Job**
6. **Job Parameter Settings of Data Development**

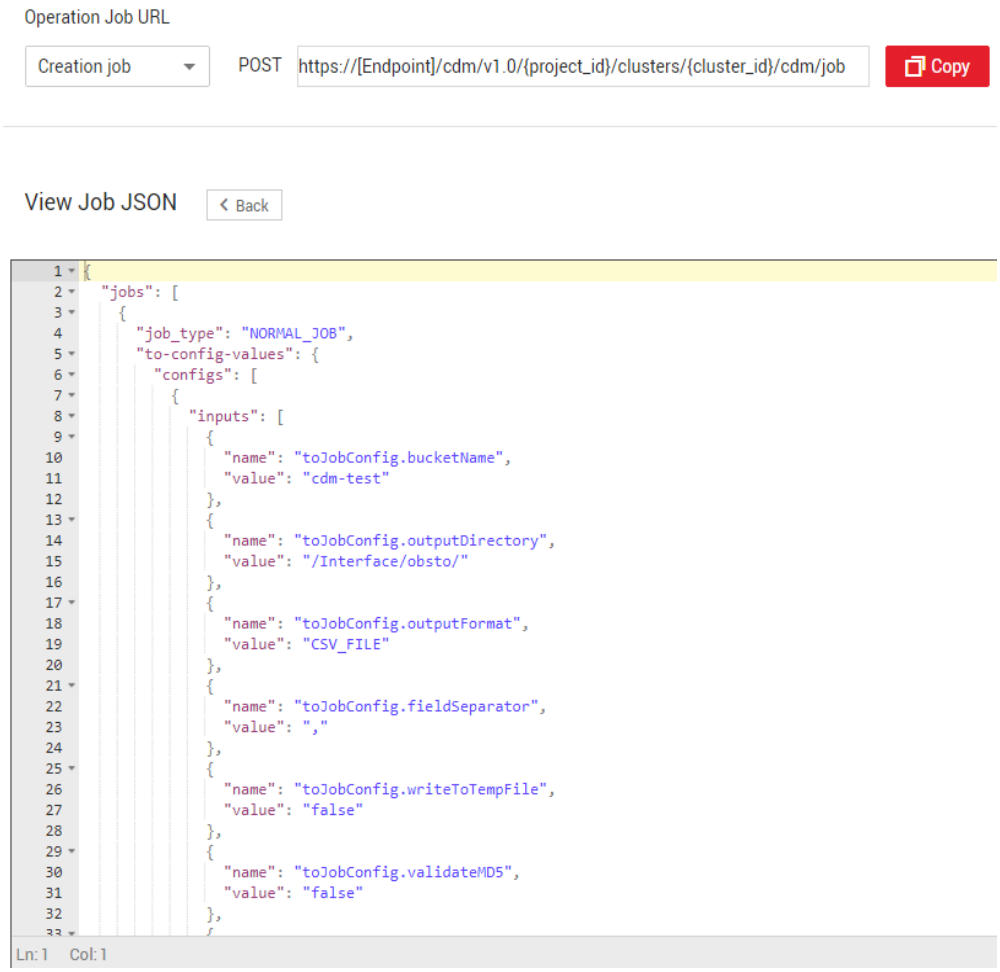
Viewing a Job JSON File

1. Create a table/file migration job from DWS to OBS.
2. On the **Table/File Migration** tab page of the **Job Management** page, find the job you have created, and choose **More** > **View Job JSON** in the **Operation** column to view the job JSON file. See **Figure 6-86**.

   You can also use other job JSON files.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-86 Viewing a job JSON file



3. A job JSON file is the request message body template for creating a CDM job. In the URL, **[Endpoint]**, **{project_id}**, and **{cluster_id}** must be replaced with the actual information.
   - **[Endpoint]**: Obtain the value from **Regions and Endpoints**, for example, **cdm.cn-north-1.myhuaweicloud.com**.
   - **{project_id}**: Project ID. Obtain the value from the project list on the **My Credentials** page
   - **{cluster_id}**: Cluster ID. Click the cluster name on the **Cluster Management** page to view the cluster ID.

Modifying a Job JSON File

You can modify the JSON body as required. The following uses 1 day as the cycle. The WHERE clause is used as the judgment condition for data extraction (generally, the time field is used as the judgment condition for incremental migration), and the data generated on the previous day is migrated every day.

1. Modify the WHERE clause to migrate incremental data in a certain period.

```
{
  "name": "fromJobConfig.whereClause",
  "value": "_timestamp >= '${startTime}' and _timestamp < '${currentTime}'"
}
```

 Copyright © Huawei Technologies Co., Ltd.

NOTE

- If the migration source is DWS or a MySQL database, the time can be determined as follows:
  ```
  _timestamp >= '2018-10-10 00:00:00' and _timestamp < '2018-10-11 00:00:00'
  Or
  _timestamp between '2018-10-10 00:00:00' and '2018-10-11 00:00:00'
  ```
- If the migration source is an Oracle database, the WHERE clause is as follows:
  ```
  _timestamp >= to_date (2018-10-10 00:00:00' , 'yyyy-mm-dd hh24:mi:ss' ) and _timestamp < to_date (2018-10-10 00:00:00' , 'yyyy-mm-dd hh24:mi:ss' )
  ```

2. Incremental data in each period is imported to different directories.

   ```
   {
     "name": "toJobConfig.outputDirectory",
     "value": "dws2obs/${currentTime}"
   }
   ```

3. Dynamically generate the job name. Otherwise, the job cannot be created because the job name is duplicate.

   ```
   "to-connector-name": "obs-connector",
   "from-link-name": "dws_link",
   "name": "dws2obs-${currentTime}"
   ```

For details on how to modify more parameters, see the **Cloud Data Migration API Reference**.The modified JSON example is as follows:

```
{
  "jobs": [
    {
      "job_type": "NORMAL_JOB",
      "to-config-values": {
        "configs": [
          {
            "inputs": [
              {
                "name": "toJobConfig.bucketName",
                "value": "cdm-test"
              },
              {
                "name": "toJobConfig.outputDirectory",
                "value": "dws2obs/${currentTime}"
              },
              {
                "name": "toJobConfig.outputFormat",
                "value": "CSV_FILE"
              },
              {
                "name": "toJobConfig.fieldSeparator",
                "value": ","
              },
              {
                "name": "toJobConfig.writeToTempFile",
                "value": "false"
              },
              {
                "name": "toJobConfig.validateMD5",
                "value": "false"
              },
              {
                "name": "toJobConfig.encodeType",
                "value": "UTF-8"
              },
              {
                "name": "toJobConfig.duplicateFileOpType",
                "value": "REPLACE"
              },
              {
                "name": "toJobConfig.kmsEncryption",
                "value": "false"
```

 Copyright © Huawei Technologies Co., Ltd.

```
          }
         ],
         "name": "toJobConfig"
        }
       ]
      },
      "from-config-values": {
       "configs": [
        {
         "inputs": [
          {
           "name": "fromJobConfig.schemaName",
           "value": "dws_database"
          },
          {
           "name": "fromJobConfig.tableName",
           "value": "dws_from"
          },
          {
           "name": "fromJobConfig.whereClause",
           "value": "_timestamp >= '${startTime}' and _timestamp < '${currentTime}'"
          },
          {
           "name": "fromJobConfig.columnList",
           "value":
"_tiny&_small&_int&_integer&_bigint&_float&_double&_date&_timestamp&_char&_varchar&_text"
          }
         ],
         "name": "fromJobConfig"
        }
       ]
      },
      "from-connector-name": "generic-jdbc-connector",
      "to-link-name": "obs_link",
      "driver-config-values": {
       "configs": [
        {
         "inputs": [
          {
           "name": "throttlingConfig.numExtractors",
           "value": "1"
          },
          {
           "name": "throttlingConfig.submitToCluster",
           "value": "false"
          },
          {
           "name": "throttlingConfig.numLoaders",
           "value": "1"
          },
          {
           "name": "throttlingConfig.recordDirtyData",
           "value": "false"
          },
          {
           "name": "throttlingConfig.writeToLink",
           "value": "obs_link"
          }
         ],
         "name": "throttlingConfig"
        },
        {
         "inputs": [],
         "name": "jarConfig"
        },
        {
         "inputs": [],
         "name": "schedulerConfig"
        },
```

 Copyright © Huawei Technologies Co., Ltd.

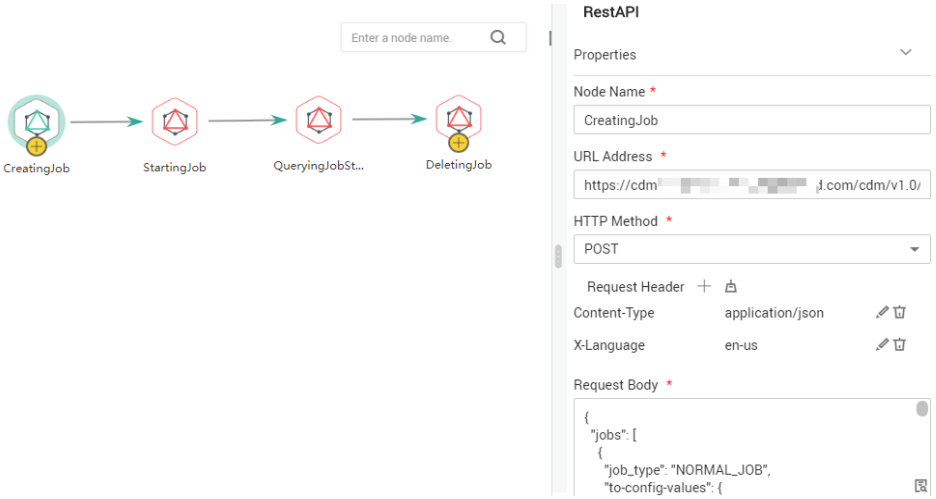
```
        {
          "inputs": [],
          "name": "transformConfig"
        },
        {
          "inputs": [],
          "name": "smnConfig"
        },
        {
          "inputs": [],
          "name": "retryJobConfig"
        }
      ]
    },
    "to-connector-name": "obs-connector",
    "from-link-name": "dws_link",
    "name": "dws2obs-${currentTime}"
  }
 ]
}
```

Creating and Running a CDM Job on the Data Development Module

1. For details on how to create a data development job, see section "Creating a Job" in the *Data Lake Factory User Guide*.
2. After the data development job is created, double-click the job name to go to the job development page. Data Development uses the RestAPI node to call a RESTful API to create a CDM migration job.
3. Configure the properties of the RestAPI node.
   a. **Node Name**: Customize a name, for example, **CreatingJob**. Note that the CDM migration job is only used as a node in the data development job.
   b. **URL Address**: Enter the URL obtained in **Viewing a Job JSON File**, for example, **https://cdm.cn-north-1.myhuaweicloud.com/cdm/v1.0/1551c7f6c808414d8e9f3c514a170f2e/clusters/6ec9a0a4-76be-4262-8697-e7af1fac7920/cdm/job**.
   c. **HTTP Method**: Enter **POST**.
   d. Add the following request headers:
      - Content-Type = application/json
      - X-Language = en-us
   e. **Request Body**: Enter the modified JSON of the CDM job in **Modifying a Job JSON File**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-87 Properties of the node for creating the CDM job

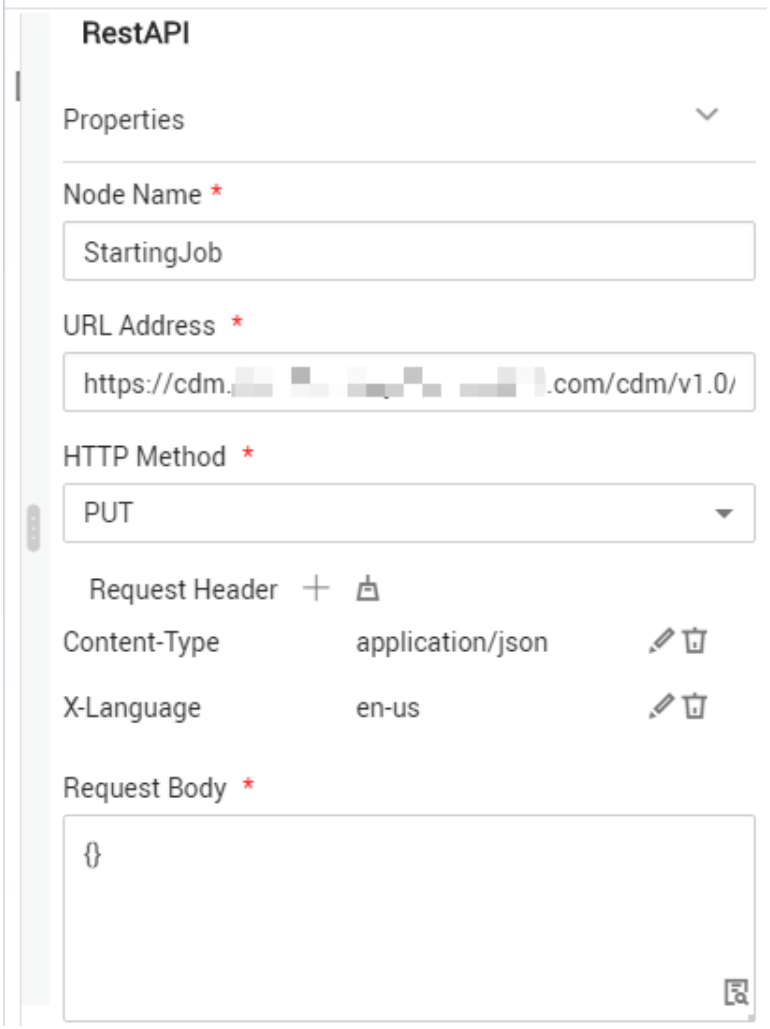


4. After configuring the RestAPI node for creating a CDM job, you need to add the RestAPI node for running the CDM job. For details, see **Starting a Job** in the *Cloud Data Migration API Reference*.
   - **Node Name**: Customize a name, for example, **StartingJob**.
   - **URL Address**: Keep the values of **project_id** and **cluster_id** consistent with those in **the RestAPI node for creating the CDM job**. Set the job name to **dws2obs-${currentTime}**.

     For example, **https://cdm.cn-north-1.myhuaweicloud.com/cdm/v1.0/1551c7f6c808414d8e9f3c514a170f2e/clusters/6ec9a0a4-76be-4262-8697-e7af1fac7920/cdm/job/dws2obs-${currentTime}/start**.
   - **HTTP Method**: Enter **PUT**.
   - Add the following request headers:
     - Content-Type = application/json
     - X-Language = en-us

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-88 Properties of the node for running the CDM job



Waiting Until the Job Execution Is Completed

Because the CDM job is running asynchronously, the REST request for running the job returns **200**, which does not indicate that the data has been migrated successfully. If a computing job depends on the CDM migration job, a RestAPI node is required to periodically check whether the migration is successful. Computing is performed only when the migration is successful.

1. For details about the API for querying the migration job status, see **Querying Job Status** in the *Cloud Data Migration API Reference*.
2. After configuring the RestAPI node for running the CDM job, add the node for waiting for the CDM job completion. The node properties are as follows:
   - **Node Name**: Customize a name, for example, **WaitingJobCompletion**.
   - **URL Address**: For example, **https://cdm.cn-north-1.myhuaweicloud.com/cdm/v1.0/1551c7f6c808414d8e9f3c514a170f2e/clusters/6ec9a0a4-76be-4262-8697-e7af1fac7920/cdm/job/dws2obs-${currentTime}/status**.
   - **HTTP Method**: Enter **GET**.

 Copyright © Huawei Technologies Co., Ltd.

  - Add the following request headers:
    - Content-Type = application/json
    - X-Language = en-us
  - Check Return Value: Select **YES**.
  - Property Path: Enter **submissions[0].status**.
  - **Request Success Flag**: Select **SUCCEEDED**.
  - Retain the default values of other parameters.

(Optional) Deleting a Job

If computing operations need to be performed after the migration, add various computing nodes to complete data computing.

You can delete the CDM jobs as required. Data Development periodically creates CDM jobs to implement incremental migration. Therefore, a large number of CDM jobs are accumulated on the CDM cluster. Therefore, after the migration is successful, you can delete the jobs that have been successfully executed.

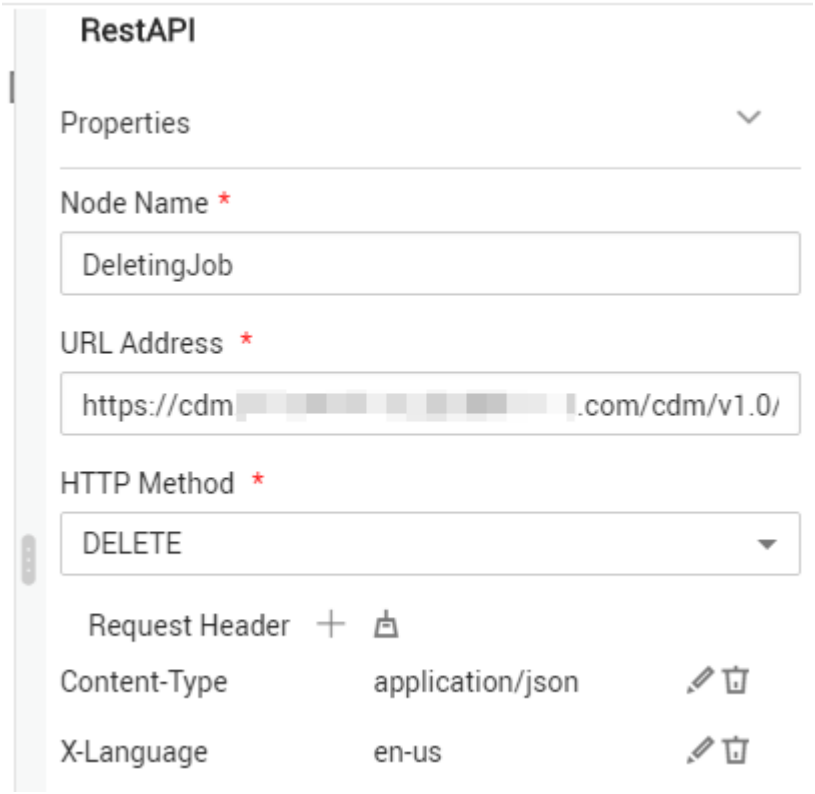
If you need to delete a CDM job, add the RestAPI node for deleting the CDM job. Then, Data Integration calls the API in **Deleting a Job** in the *Cloud Data Migration API Reference*.

The node properties are as follows:

- **Node Name**: Customize a name, for example, **DeletingJob**.
- **URL Address**: For example, **https://cdm.cn-north-1.myhuaweicloud.com/cdm/v1.0/1551c7f6c808414d8e9f3c514a170f2e/clusters/6ec9a0a4-76be-4262-8697-e7af1fac7920/cdm/job/dws2obs-${currentTime}**.
- **HTTP Method**: Enter **DELETE**.
- Add the following request headers:
  - Content-Type = application/json
  - X-Language = en-us
- Retain the default values of other parameters.

 Copyright © Huawei Technologies Co., Ltd.

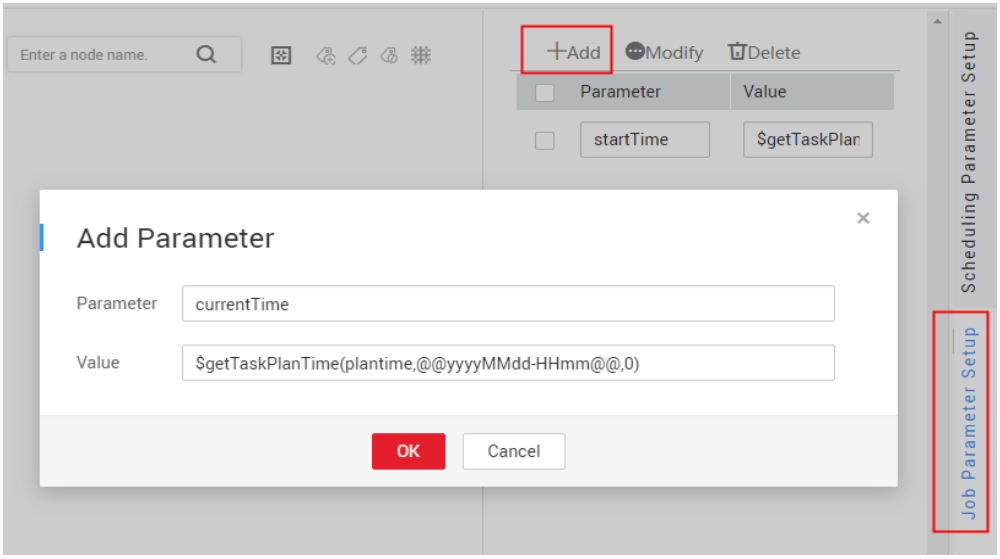
Figure 6-89 Properties of the node for deleting the CDM job



Job Parameter Settings of Data Development

1. Set job parameters shown in **Figure 6-90**.
   - startTime = $getTaskPlanTime(plantime,@@yyyyMMddHHmmss@@,-24*60*60)
   - currentTime = $getTaskPlanTime(plantime,@@yyyyMMdd-HHmm@@,0)

Figure 6-90 Job parameter settings of Data Development



 Copyright © Huawei Technologies Co., Ltd.

2. After saving the data development job, choose **Scheduling Configuration** > **Schedule periodically** and set the value to one day.

   In this way, Data Development works with CDM to migrate data generated on the previous day.

6.1.12.11 Migrating the 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 for MySQL. 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 bought 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 HUAWEI CLOUD has been established.
- You have obtained the IP addresses, names, usernames, and passwords of the on-premises MySQL database and RDS for MySQL.

Creating a CDM Cluster and Binding an EIP to the Cluster

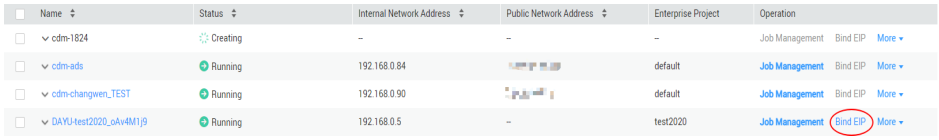
Step 1 The DGC basic package includes a CDM cluster. If the cluster cannot meet your requirements, you can buy an incremental package to supplement the DGC basic instance. Create a CDM cluster. For details, see **Buying a DGC Incremental Package**. The key configurations are as follows:

- The flavor of the CDM cluster is selected based on the amount of data to be migrated. Generally, select **cdm.medium** to meet the requirements of 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.

 Copyright © Huawei Technologies Co., Ltd.

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

Figure 6-91 Binding an EIP



NOTE

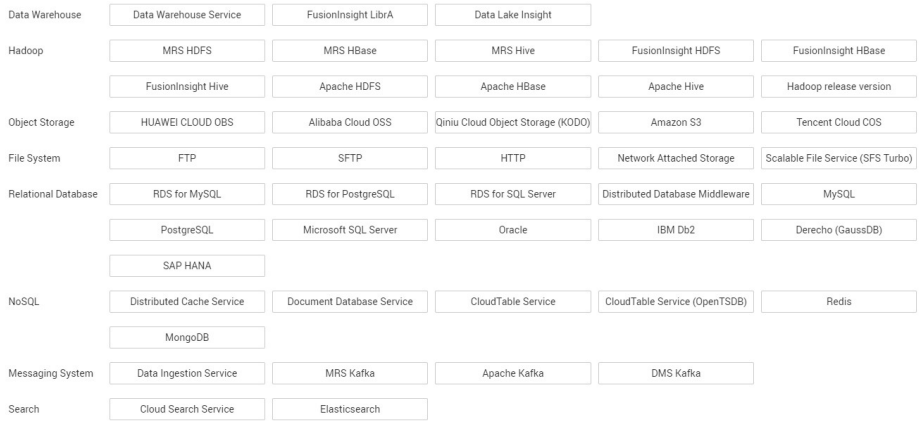
If SSL encryption is configured for the access channel of a local data source, CDM cannot connect to the data source using the EIP.

----End

Creating a MySQL Link

Step 1 On the **Cluster Management** page, click **Job Management** of the cluster and choose **Link Management** > **Create Link** to enter the page for selecting the connector. See **Figure 6-92**.

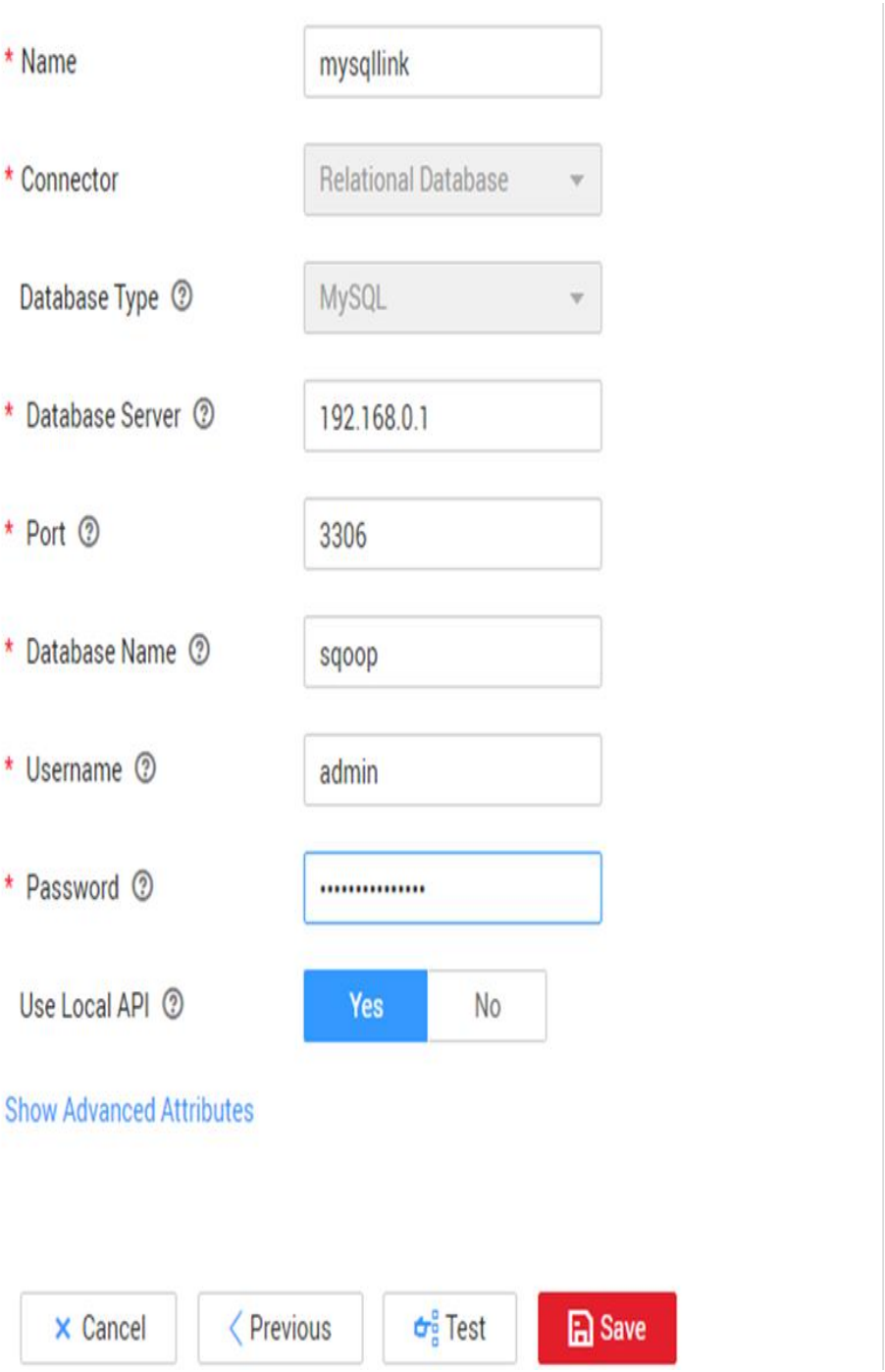
Figure 6-92 Selecting a connector



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

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-93 Creating a MySQL link



Click **Show Advanced Attributes** to display optional parameters. For details, see **Link to Relational Databases**. Retain the default values for the optional parameters and configure the mandatory parameters described in **Table 6-78**.

Table 6-78 MySQL link parameters

Parameter	Description	Example Value
Name	A unique link name	mysqllink

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Database Server	IP address or domain name of the MySQL database	192.168.0.1
Port	MySQL database port	3306
Database Name	The name of the MySQL database	sqoop
Username	A user who has the read, write, and delete permissions on the MySQL database	admin
Password	The password of the user	N/A
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**.	N/A

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

NOTE

If an error occurs when saving the settings, it is likely that the security settings of the MySQL database are incorrect. If this is the case, you may need to enable the EIP of the CDM cluster to access the MySQL database.

----End

Creating an RDS Link

Step 1 Select **RDS (MySQL)**, click **Next**, and configure the RDS link parameters.

- **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.

 Copyright © Huawei Technologies Co., Ltd.

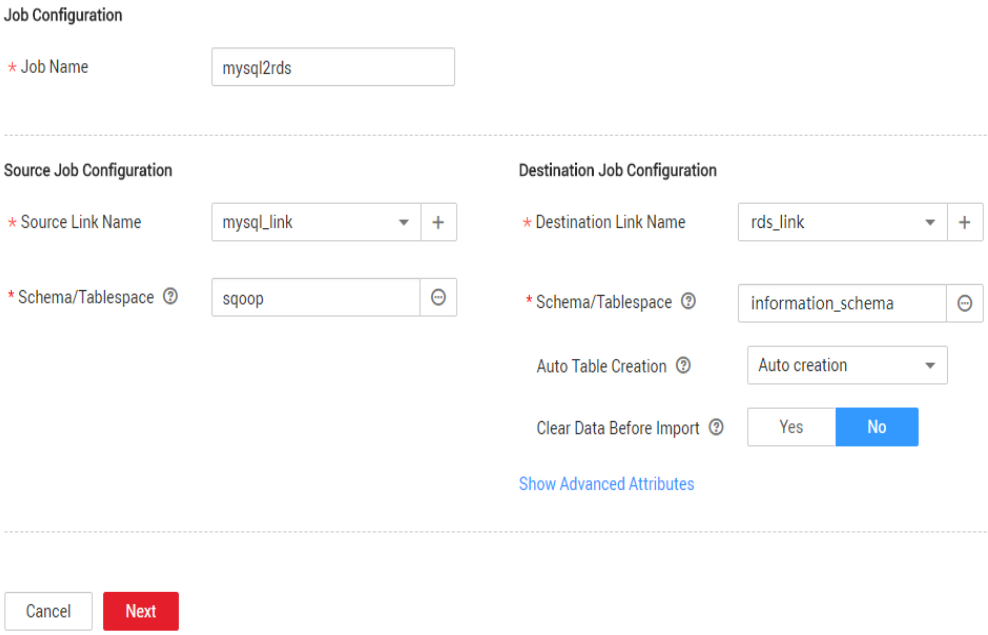
Step 2 Click **Save**. The **Link Management** page is displayed.

----End

Creating an Entire DB Migration Job

Step 1 After the two links are created, choose **Entire DB Migration** > **Create Job** to create a migration job. See **Figure 6-94**.

Figure 6-94 Creating an entire DB migration job 3



- **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**.

 Copyright © Huawei Technologies Co., Ltd.

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

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

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

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

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

----End

6.1.12.12 Migrating the Entire Elasticsearch Database to Cloud Search Service

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 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 Cloud Search Service and obtained the IP address and port number of the Cloud Search Service 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 The DGC basic package includes a CDM cluster. If the cluster cannot meet your requirements, you can buy an incremental package to supplement the DGC basic instance. Create a CDM cluster. For details, see **Buying a DGC Incremental Package**. The key configurations are as follows:

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

 Copyright © Huawei Technologies Co., Ltd.

- 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

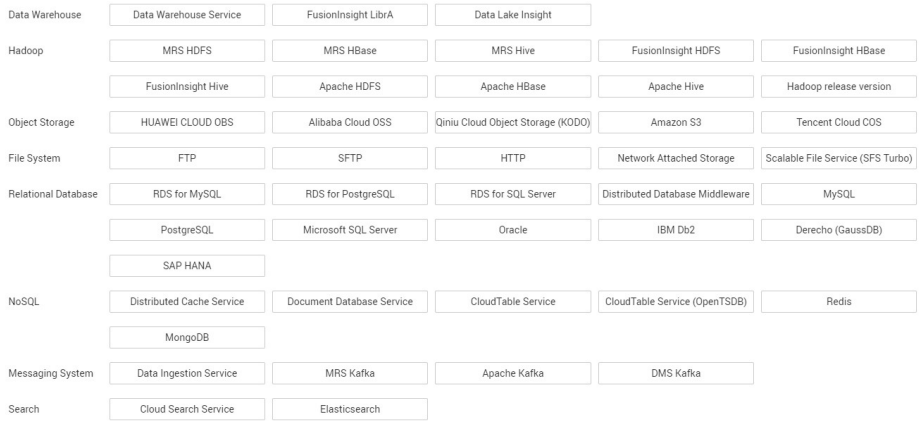
If SSL encryption is configured for the access channel of a local data source, CDM cannot connect to the data source using the EIP.

----End

Creating a Cloud Search Service Link

Step 1 Click **Job Management** in the **Operation** column of the CDM cluster. On the page that is displayed, choose **Link Management** > **Create Link**. The page for selecting a connector is displayed. See **Figure 6-95**.

Figure 6-95 Selecting a connector

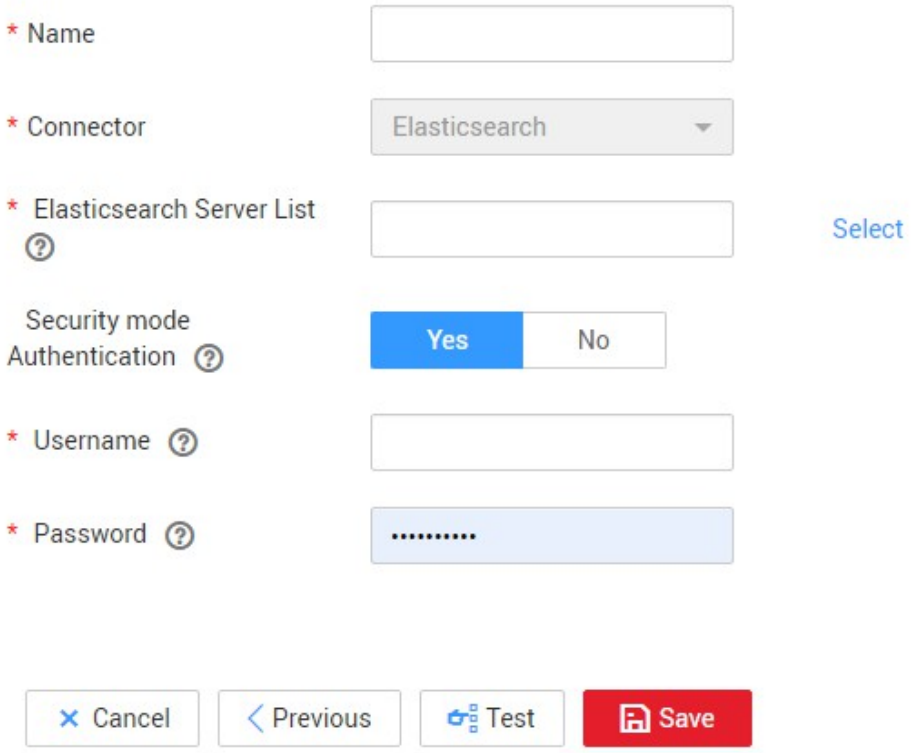


Step 2 Select **Cloud Search Service** and click **Next**. On the page that is displayed, configure the CSS link parameters.

- **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.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-96 Creating a CSS link



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

----End

Creating an Elasticsearch Link

Step 1 On the **Link Management** tab page, click **Create Link**. On the page that is displayed, select **Elasticsearch**, click **Next**, and configure the Elasticsearch link parameters. The Elasticsearch link parameters are the same as those of the Cloud Search Service 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.
- **Username** and **Password**: If the Elasticsearch database has user restrictions, select the user who has the read and write permissions on the Elasticsearch database. If there is no user restriction, you do not need to configure the two parameters.

Step 2 Click **Save**. The **Link Management** page is displayed.

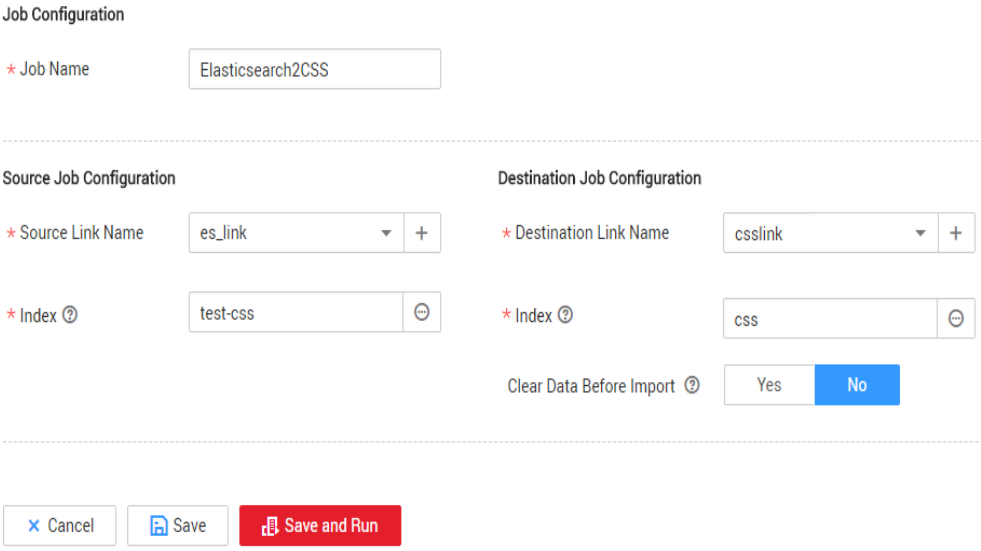
----End

Creating an Entire DB Migration Job

Step 1 Choose **Entire DB Migration** > **Create Job** to create an entire DB migration job.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-97 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.

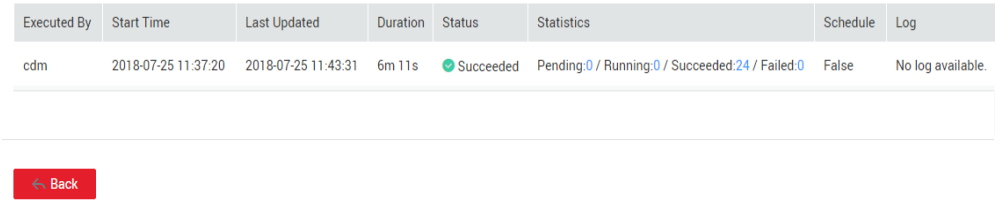
Step 2 Click **Save and Run**. The **Job Management** page is displayed, on which you can view the job execution progress and result.

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).

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-98 Historical Record



----End

6.1.13 Advanced Operations

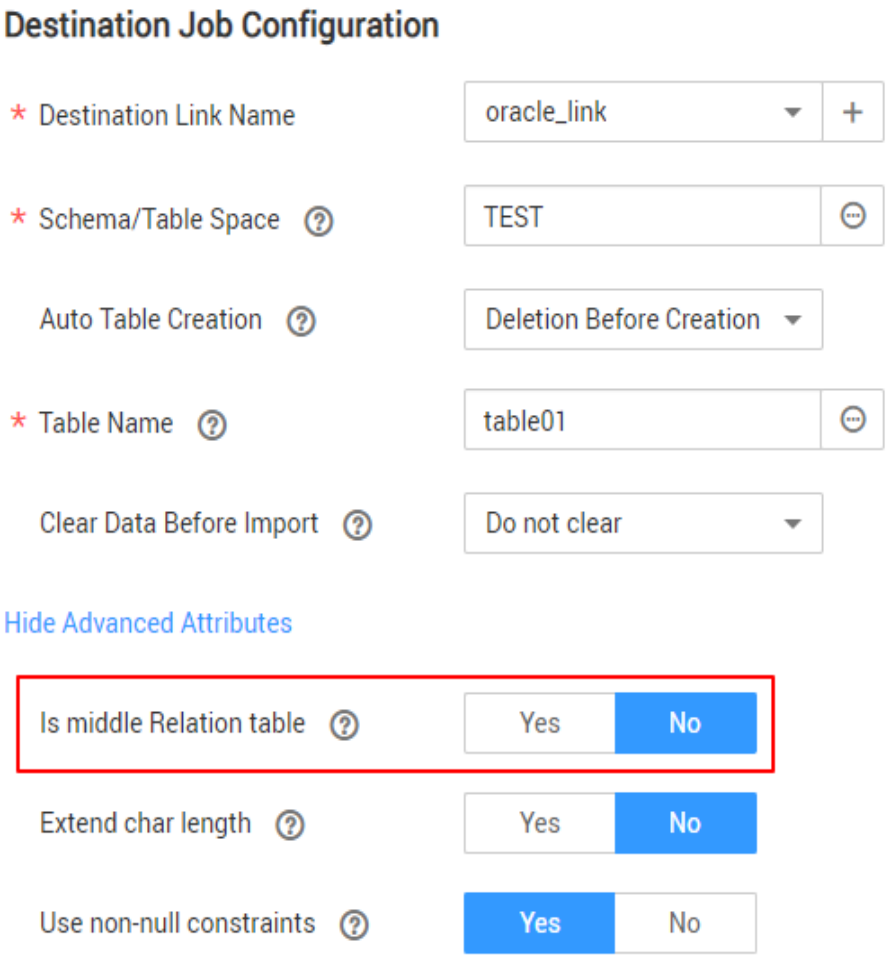
6.1.13.1 Migration in Transaction Mode

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.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-99 Migration in transaction mode



NOTE

If you set **Clear Data Before Import** to **Yes**, CDM does not roll back the deleted data even in transaction mode.

6.1.13.2 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, NAS, SFS, HDFS (supported in the binary format), and HTTP (applicable to scenarios where OBS shared files are downloaded)

 Copyright © Huawei Technologies Co., Ltd.

- Data sources supported by the migration destination: OBS, FTP, SFTP, NAS, SFS, and HDFS (supported in the binary format)

The following describes how to use the AES-256-GCM algorithm for encryption and decryption in OBS file migration. 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**.

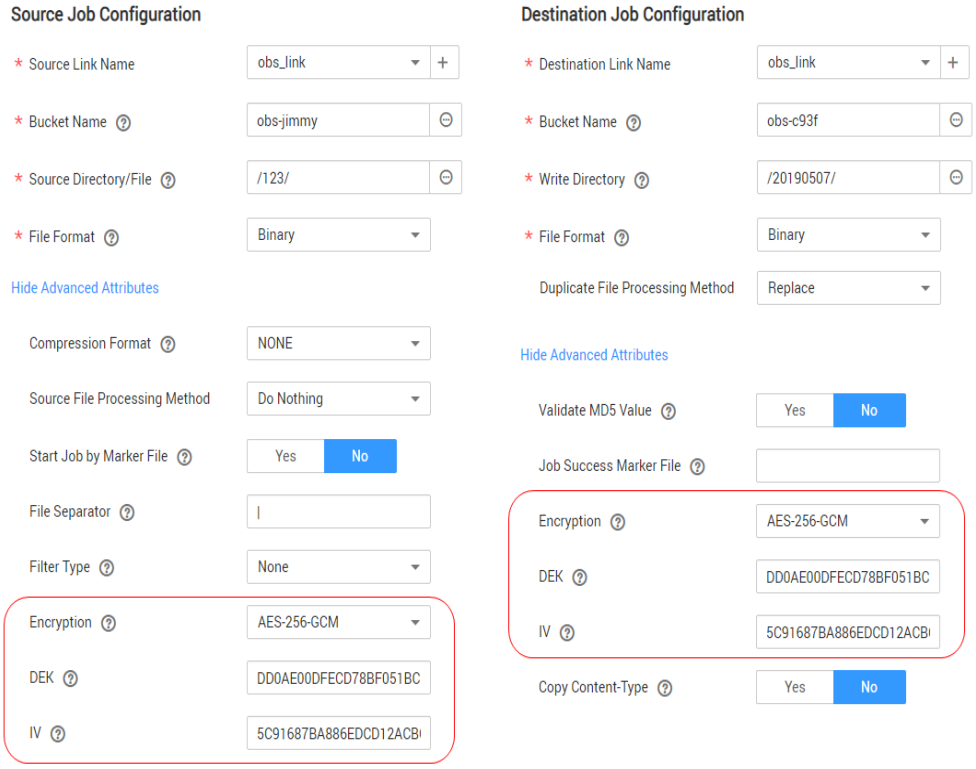
  b. **DEK**: custom encryption key. The key consists of 64 hexadecimal numbers. It is case-insensitive but must contain 64 characters. For example, **DD0AE00DFECD78BF051BCFDA25BD4E320DB0A7AC75A1F3FC3D3C56A457DCDC1B**.

  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.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-100 AES-256-GCM encryption and decryption



KMS Encryption

NOTE

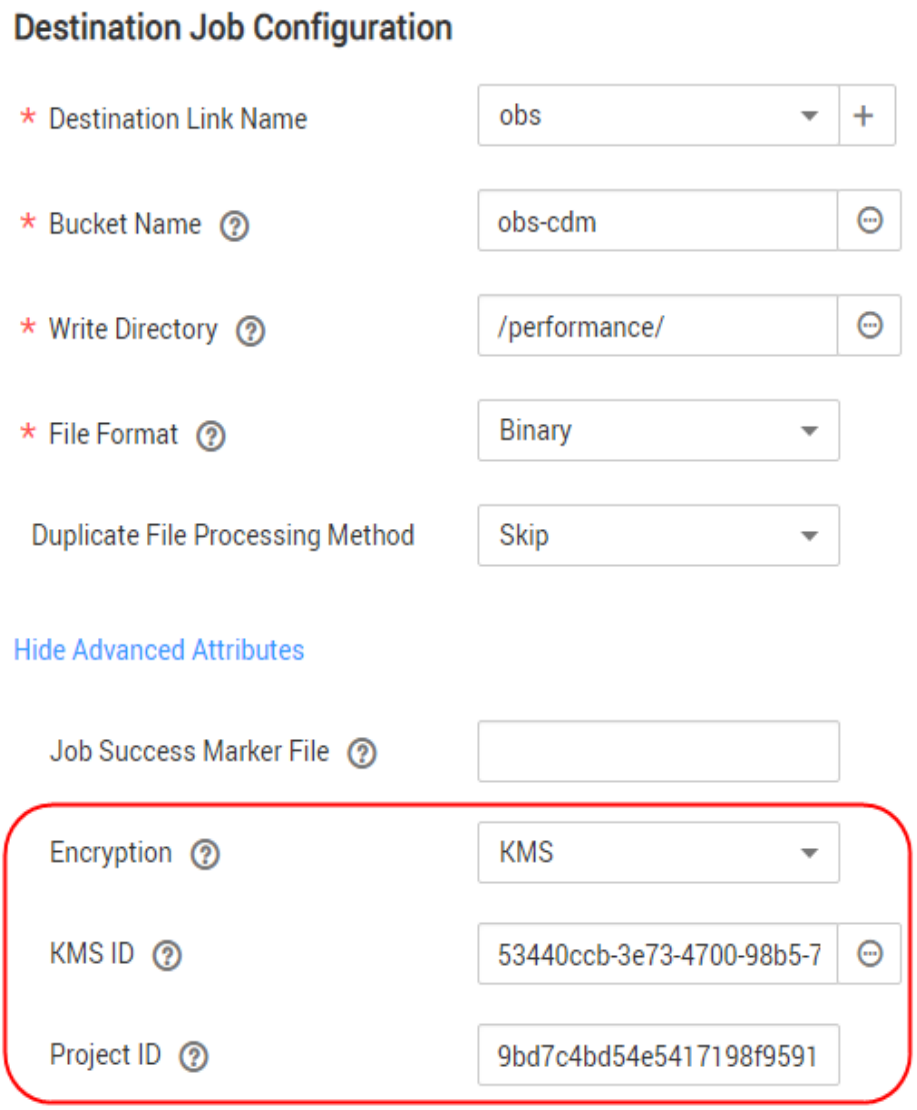
The migration source does not support KMS encryption.

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. See **Figure 6-101**.

A key must be created in KMS of DEW in advance. For details, see the *Data Encryption Workshop User Guide*.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-101 Enabling KMS encryption



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

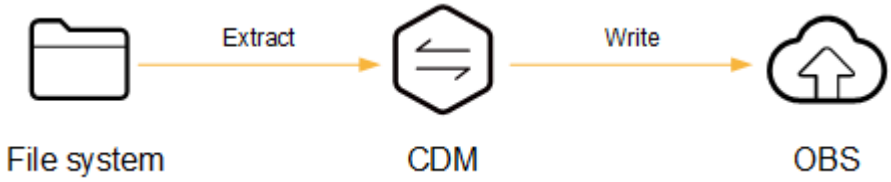
- If KMS encryption is enabled, **MD5 verification** cannot be used.
- If the KMS ID of another project is used, change **Project ID** to the ID of the project to which KMS belongs. If KMS and CDM are in the same project, retain the default value of **Project ID**.
- After KMS encryption is performed, 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.

 Copyright © Huawei Technologies Co., Ltd.

6.1.13.3 MD5 Verification

CDM extracts data from the migration source and writes the data to the migration destination. **Figure 6-102** shows the migration mode when files are migrated to OBS.

Figure 6-102 Migrating files to OBS

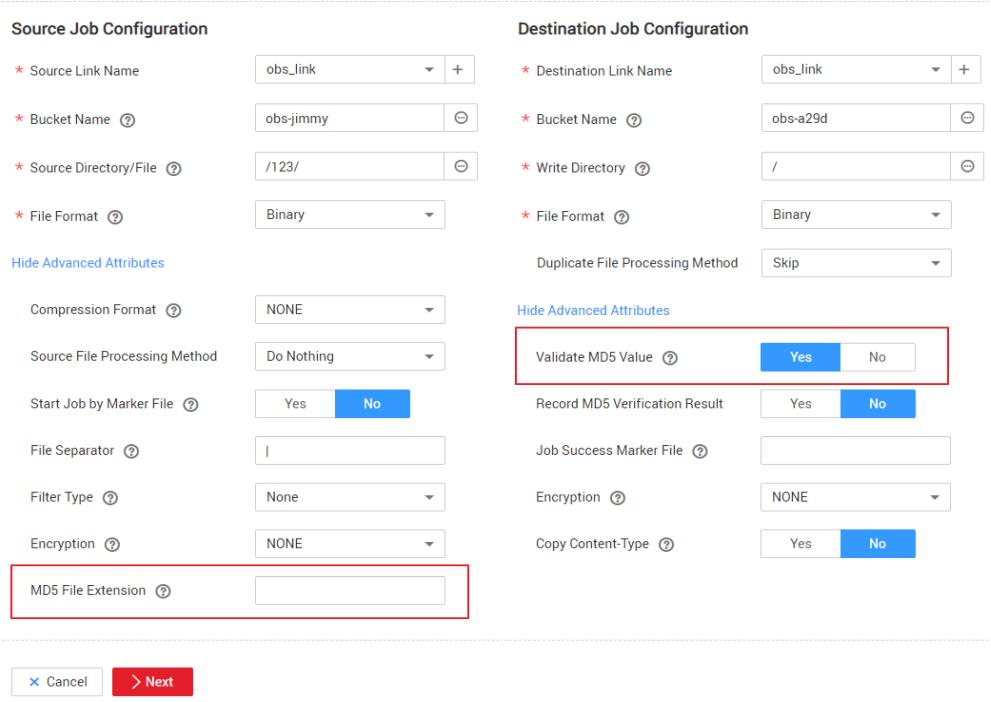


During the process, CDM uses MD5 to verify file consistency. For details about the parameters, see **Figure 6-103**.

- **Extract**
  - Check whether the files extracted by CDM are consistent with source files.
  - This function is controlled by the **MD5 File Extension** parameter 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.
  - The migration source can be OBS, FTP, SFTP, NAS, SFS, or HTTP.
- **Write**
  - 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.
  - Currently, this function can be used only when OBS serves as the migration destination.

Copyright © Huawei Technologies Co., Ltd.

Figure 6-103 Enabling MD5 verification to verify file consistency



NOTE

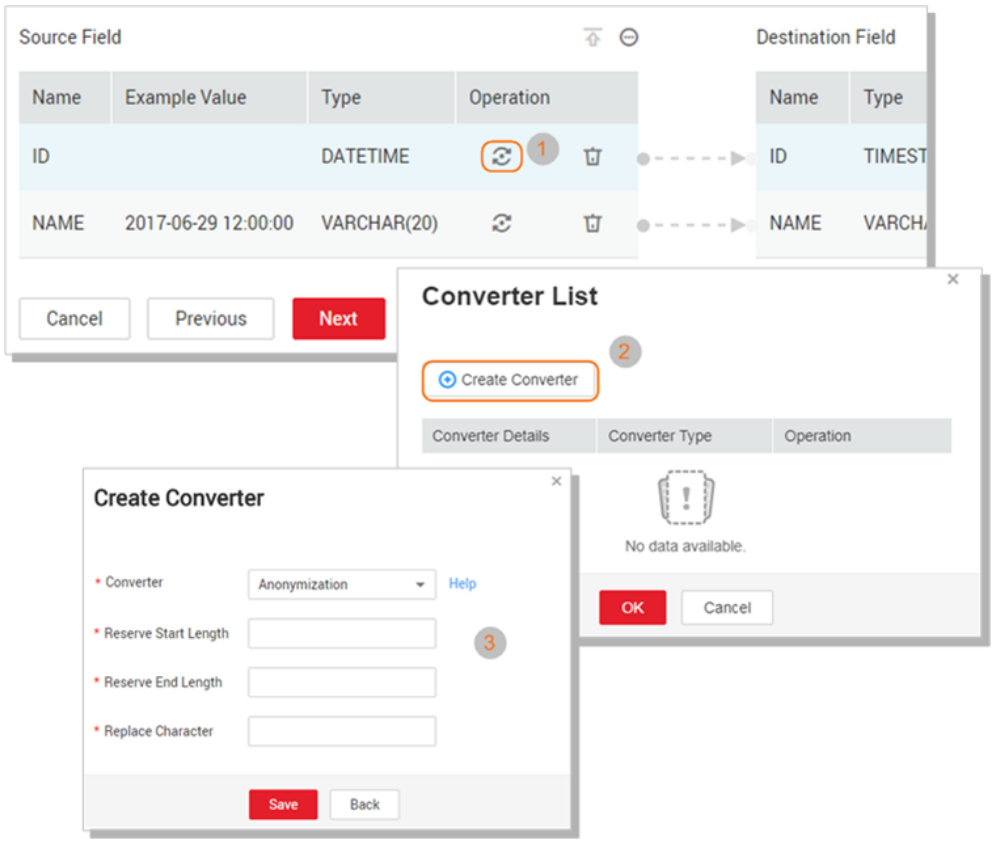
- When files are migrated to a file system, only the extracted files are verified.
- When files are migrated to OBS, both the extracted files and files written to OBS are verified.
- If MD5 verification is used, **KMS encryption** cannot be used.

6.1.13.4 Field Conversion

You can create a field converter on the **Map Field** tab page when creating a table/file migration job. See **Figure 6-104**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-104 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**

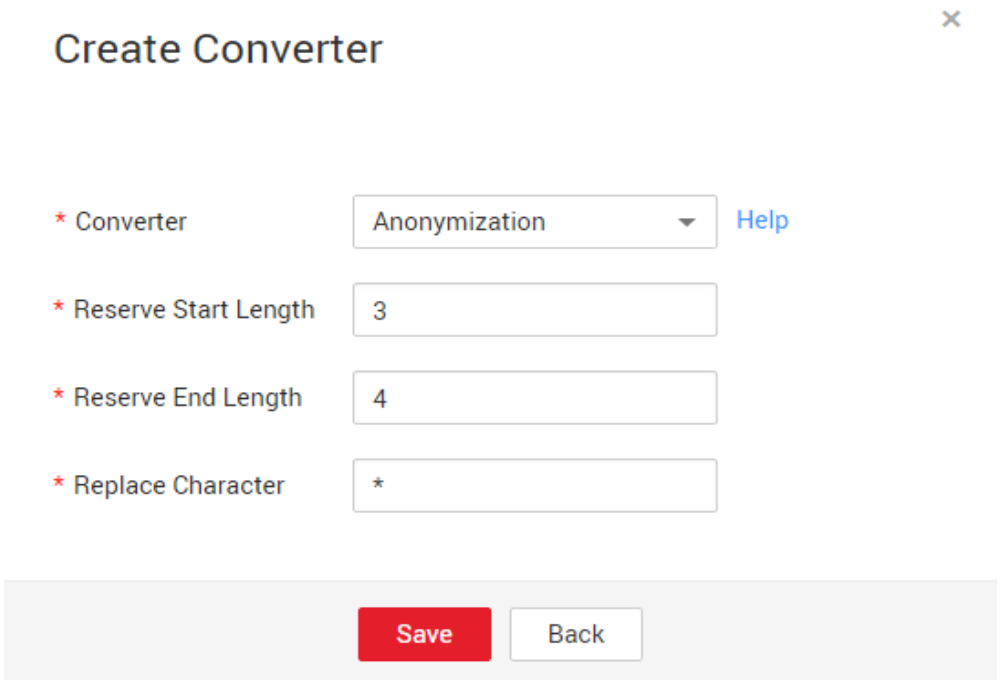
Anonymization

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 *****.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-105 Anonymization



Trim

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

Reverse String

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

Replace String

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

Remove line break

This converter is used to delete the newline characters, such as \n, \r, and \r\n from the field.

Expression Conversion

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

The expression supports the following environment variables:

- **value**: indicates the current field value.

 Copyright © Huawei Technologies Co., Ltd.

- **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. Set a string constant for the current field, for example, VIP.
   Expression: **"***VIP***"**
2. 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)
3. Convert all character strings of the current field to uppercase letters.
   Expression: StringUtils.upperCase(value)
4. 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,"-")
5. 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
6. Convert the field value **true** to **Y** and other field values to **N**.
   Expression: value=="true"?"Y":"N"
7. 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
8. If the first and second fields are of the numeric type, convert the field to the combination of the first and second field values.
   Expression: row[0]+row[1]
9. If the field is of the date or timestamp type, return the current year after conversion. The data type is int.
   Expression: DateUtils.getYear(value)
10. If the field is a date and time string in *yyyy-MM-dd* format, convert it to the date type:
    Expression: DateUtils.format(value,"yyyy-MM-dd")
11. 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")
12. Obtain a 36-bit universally unique identifier (UUID):
    Expression: CommonUtils.randomUUID()
13. If the field is of the string type, capitalize the first letter, for example, convert **cat** to **Cat**.

 Copyright © Huawei Technologies Co., Ltd.

Expression: StringUtils.capitalize(value)

14. 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)

15. 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*)

16. 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)

17. 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*")

18. 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*")

19. 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*")

20. 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*")

21. 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*)

22. 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*)

23. 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*")

24. 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*")

25. 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.

 Copyright © Huawei Technologies Co., Ltd.

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

26. 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*)

27. 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*")

28. 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)

29. 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)

30. 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)

31. 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)

32. 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)

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

    Expression: StringUtils.isEmpty(value)

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

    Expression: StringUtils.isNumeric(value)

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

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

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

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

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

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

 Copyright © Huawei Technologies Co., Ltd.

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

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

39. 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)

40. 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*")

41. 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*")

42. 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.*")

43. 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*")

44. 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*")

45. If the field is of the string type, use the specified delimiter to split the text into arrays. For example, use **:** to split **ab:cd:ef** into **["ab","cd","ef"]**.

    Expression: StringUtils.split(value,":")

46. 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*")

47. 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*")

48. 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*)

49. 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*)

 Copyright © Huawei Technologies Co., Ltd.

50. 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*)

51. 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*)

52. 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*")

53. 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*")

54. 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*")

55. 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*")

56. 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*")

57. 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)

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

    Expression: NumberUtils.toByte(value)

59. 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*)

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

    Expression: NumberUtils.toDouble(value)

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

 Copyright © Huawei Technologies Co., Ltd.

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

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

    Expression: NumberUtils.toFloat(value)

63. 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*)

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

    Expression: NumberUtils.toInt(value)

65. 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*)

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

    Expression: NumberUtils.toLong(value)

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

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

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

    Expression: NumberUtils.toShort(value)

69. 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*)

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

    Expression: CommonUtils.ipToLong(value)

71. 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*")

72. Cache the IP address and physical address mappings and specify a key for retrieval, for example, **ipList**.

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

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

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

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

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

75. Obtain the physical addresses corresponding to the IP address in *Country_Province_City_Carrier* format. For example, the physical address corresponding to **1xx.78.124.0** is **China_Guangdong_Shenzhen_China Telecom**. If the corresponding physical address cannot be obtained, the

 Copyright © Huawei Technologies Co., Ltd.

default value **_**_**_** is returned. If necessary, you can use the StringUtil class expression to further split the addresses.

Expression: CommonUtils.getMapValue(CommonUtils.ipToLong(value),CommonUtils.cacheExists("*ipLis*")? CommonUtils.getCache("*ipLis*"):CommonUtils.setCache("*ipLis*",HttpsUtils.downloadMap("*url*")))

76. 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**.

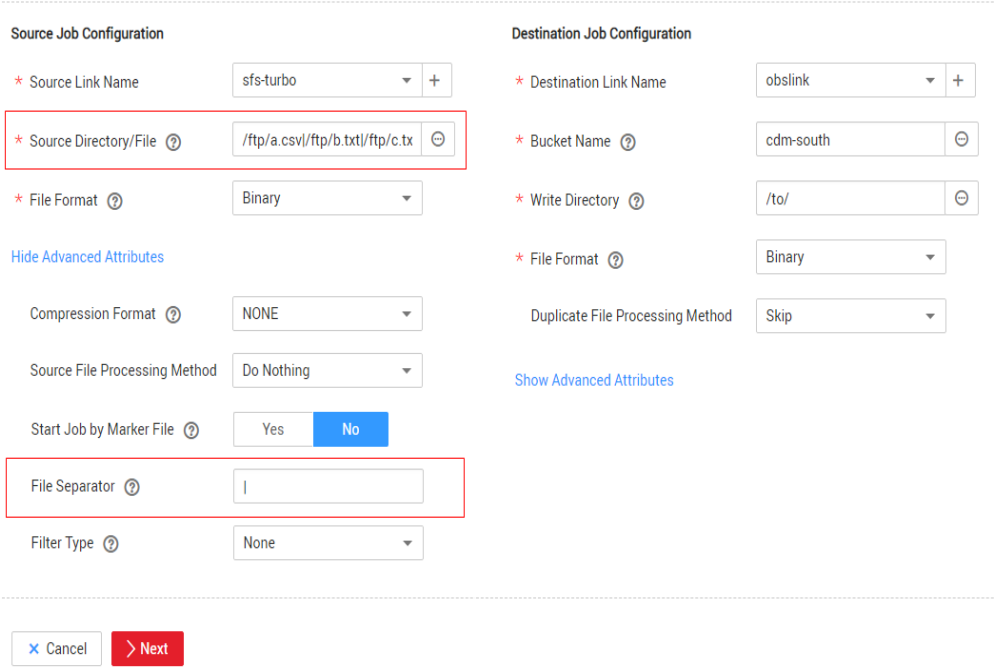
    Expression: DateUtils.getCurrentTimeByZone("*yyyy-MM-dd HH:mm:ss*",value,"*hour*", *8*)

6.1.13.5 Migration of a List of Files

You can migrate a list of files (a maximum of 50 files) from FTP, SFTP, NAS, OBS, OSS, KODO, or SFS 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, NAS, OBS, OSS, KODO, or SFS, **Source Directory/File** can contain a maximum of 50 file names, which are separated by vertical bars (|). You can also customize a file separator.

Figure 6-106 Migrating a list of files



 Copyright © Huawei Technologies Co., Ltd.

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.

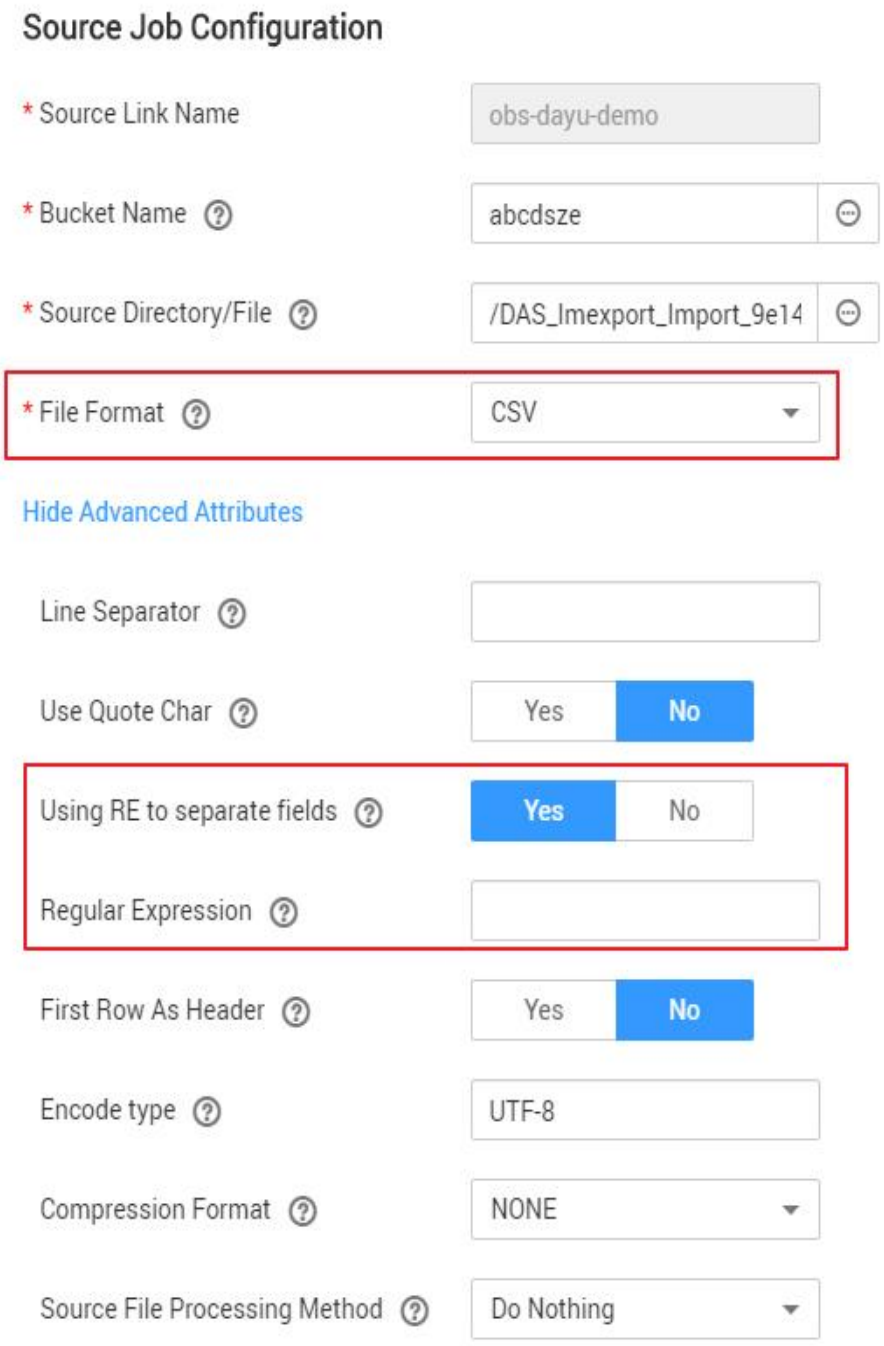
6.1.13.6 Regular Expressions for Separating Semi-structured Text

When jobs are migrated according to **Table/File Migration Jobs**, 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**.

Copyright © Huawei Technologies Co., Ltd.

Figure 6-107 Setting regular expression parameters



During the migration of CSV files, CDM can use regular expressions to separate fields and write parsed results to the migration destination. For details about the syntax of the regular expression, refer to the related documents. This section describes the regular expressions of the following log files:

- **Log4J Log**
- **Log4J Audit Log**
- **Tomcat Log**
- **Django Log**

 Copyright © Huawei Technologies Co., Ltd.

- **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 6-79 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 6-80 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

 Copyright © Huawei Technologies Co., Ltd.

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 6-81 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:

 Copyright © Huawei Technologies Co., Ltd.

Table 6-82 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 6-83 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

6.1.13.7 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**

Copyright © Huawei Technologies Co., Ltd.

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 6-84 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 6-84**.

3. **Encoding Type**

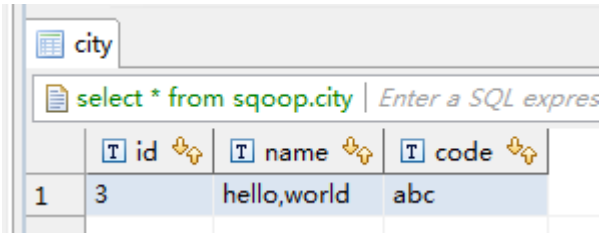
   Encoding type of a CSV file. The default value is **UTF-8**. Some Chinese characters are encoded by GBK.

   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.

Copyright © Huawei Technologies Co., Ltd.

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 6-108** shows that the value of the **name** field in the database contains a comma (,).

     Figure 6-108 Field value containing the field delimiter

     If you do not use the quote character, the exported CSV file is displayed as follows:

     ```
     3.hello,world,abc
     ```

     If you use the quote character, the exported CSV file is displayed as follows:

     ```
     3,"hello,world",abc
     ```

     If the data in the database contains double quotation marks ("") and you set **Use Quote Character** to **Yes**, the quote character in the exported CSV file is displayed as three double quotation marks (""""). For example, if the value of a field is **a"hello,world"c**, the exported data is as follows:

     ```
     """a"hello,world"c"""
     ```

   - Exporting CSV files to a database or NoSQL (configuring **Use Quote Character** at the migration source): If you want to import the CSV files with quoted values to a database correctly, set **Use Quote Character** to **Yes** at the migration source to write the quoted values as a whole.

5. **Use RE to Separate Fields**

   This function is used to parse complex semi-structured text, such as log files. For details, see **Using Regular Expressions to Separate Semi-structured Text**.

6. **Use First Row as Header**

   This parameter is used when CSV files are exported to other locations. If this parameter is specified at the migration source, CDM uses the first row as the header when extracting data. When the CSV files are transferred, the headers are skipped. The number of rows extracted from the migration source is more than the number of rows written to the migration destination. The log files will output the information that the header is skipped during the migration.

7. **File Size**

   This parameter is used when data is exported from the database to a CSV file. If a table contains a large amount of data, a large CSV file is generated after migration, which is inconvenient to download or view. In this case, you can

 Copyright © Huawei Technologies Co., Ltd.

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**

 Copyright © Huawei Technologies Co., Ltd.

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": ["chinese","english","math"]
        }
      },
      {
          "_id": "650616",
          "_source": {
              "name": "tom",
              "books": ["chinese","english","math"]
          }
      },
     {
          "_id": "650618",
          "_source": {
              "name": "tom",
              "books": ["chinese","english","math"]
          }
      }]
```

 Copyright © Huawei Technologies Co., Ltd.

```
   }
}
```

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	["chinese","english","math"]
650616	tom	["chinese","english","math"]
650618	tom	["chinese","english","math"]

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:

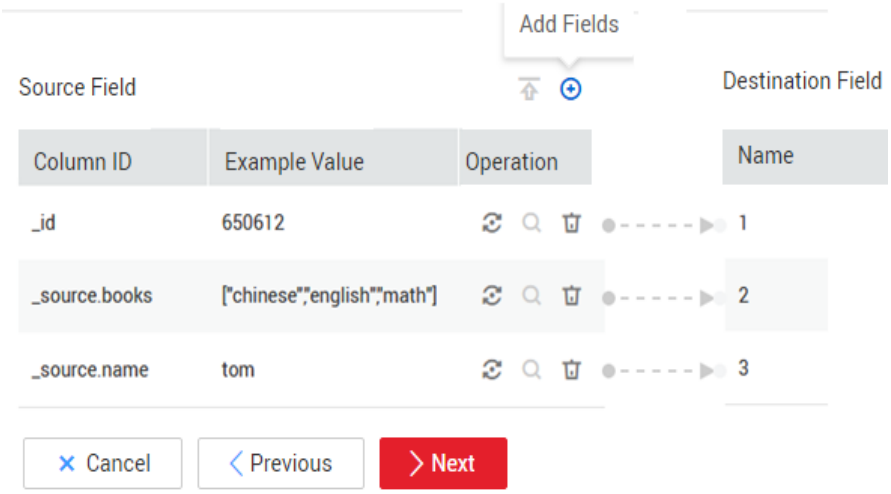
ID	SourceName	SourceBooks	MaxScore
650612	tom	["chinese","english","math"]	1.0
650616	tom	["chinese","english","math"]	1.0
650618	tom	["chinese","english","math"]	1.0

Set **File Format** to **JSON**, **JSON Type** to **JSON object**, and **JSON Reference Node** to **hits.hits**, and then create a converter.

Copyright © Huawei Technologies Co., Ltd.

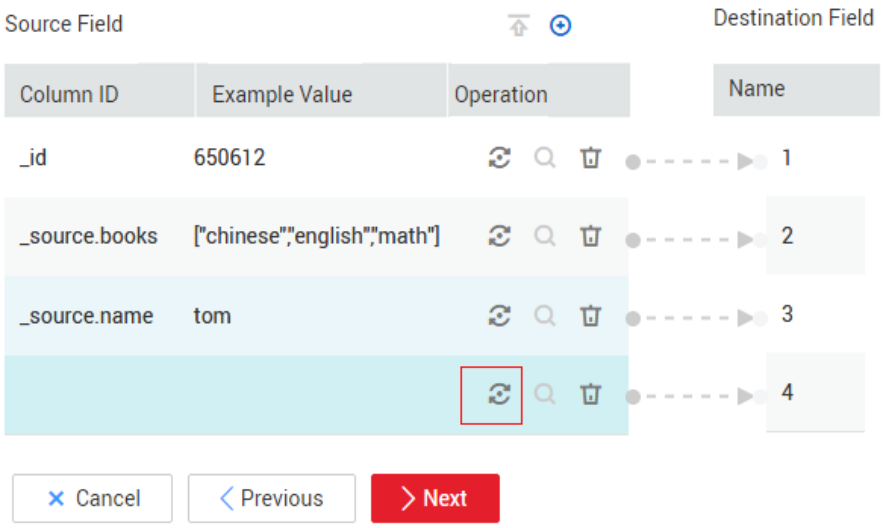
i. Click ⊕ to add a field.

Figure 6-109 Adding a field



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

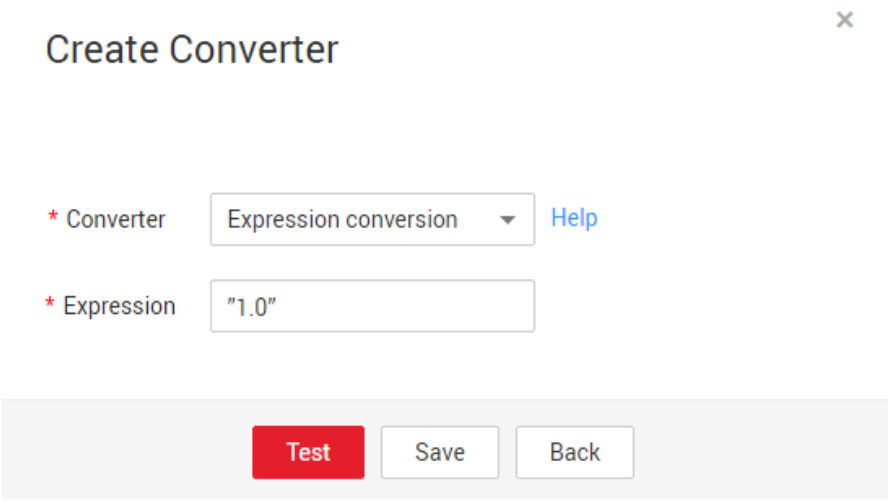
Figure 6-110 Creating a field converter



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

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-111 Configuring a field converter



Binary

If you want to copy files between file systems, you can select the binary format. The binary format delivers the optimal rate and performance in file transfer, and does not require field mapping.

- **Directory structure for file transfer**

  CDM can transfer a single file or all files in a directory at a time. After the files are transferred to the migration destination, the directory structure remains unchanged.

- **Migrating incremental files**

  When you use CDM to transfer files in binary format, configure **Duplicate File Processing Method** at the migration destination for incremental file migration. For details, see **Incremental File Migration**.

  During incremental file migration, set **Duplicate File Processing Method** to **Skip**. If new files exist at the migration source or a failure occurs during the migration, run the job again, so that the migrated files will not be migrated repeatedly.

- **Write to Temporary File**

  When migrating files in binary format, you can specify whether to write the files to a temporary file at the migration destination. If this parameter is specified, the file is written to a temporary file during file replication. After the file is successfully migrated, run the **rename** or **move** command to restore the file at the migration destination.

- **Generate MD5 Hash Value**

  An MD5 hash value is generated for each transferred file, and the value is recorded in a new **.md5** file. You can specify the directory where the MD5 value is generated.

Common parameters

- **Source File Processing Method**

  After a file is copied successfully, CDM can perform operations on the source file. The options are **Rename** and **Delete**.

 Copyright © Huawei Technologies Co., Ltd.

- **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 6-84**.

   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.

 Copyright © Huawei Technologies Co., Ltd.

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.

6.1.14 Monitoring

6.1.14.1 CDM Metrics

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 6-85 lists the CDM metrics.

Table 6-85 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

Copyright © Huawei Technologies Co., Ltd.

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

Dimension

Key	Value
instance_id	CDM instance

6.1.14.2 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 Choose **Cluster Management**. Choose **More** > **View Metric**. The Cloud Eye management console is displayed.

Step 2 In the left navigation pane, choose **Alarm Management** > **Alarm Rules** > **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 CDM alarm rules, see **Cloud Eye User Guide**.

----End

Copyright © Huawei Technologies Co., Ltd.

6.1.14.3 Querying Metrics

Scenario

Cloud Eye monitors the running status of the CDM cluster. You can obtain the monitoring metrics of CDM on the Cloud Eye management console.

Monitored data requires a period of time for transmission and display. The status of CDM displayed on the Cloud Eye page 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 shut down or restart, or is unavailable, its monitoring metrics cannot be viewed on Cloud Eye. 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.

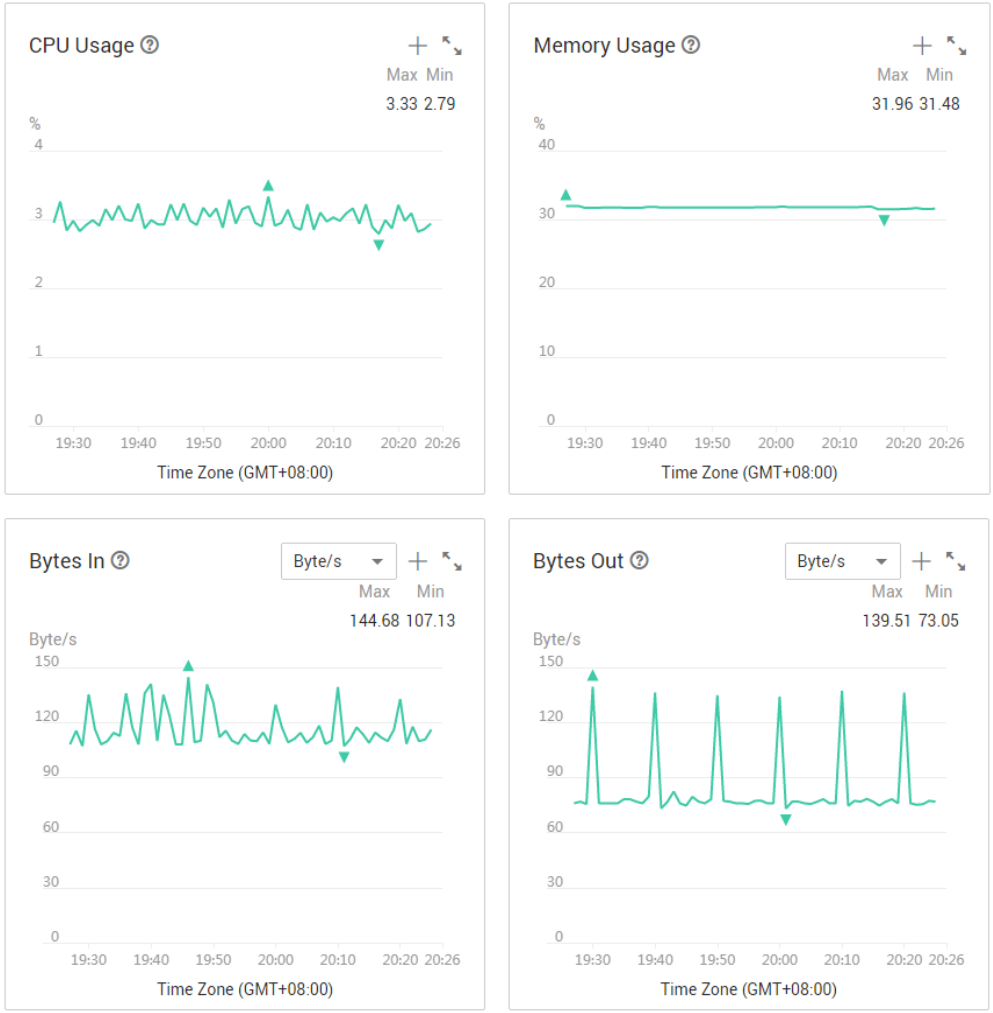
Procedure

Step 1 Choose **Cluster Management**. Choose **More** > **View Metric**. The Cloud Eye console is displayed.

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

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-112 Viewing metrics



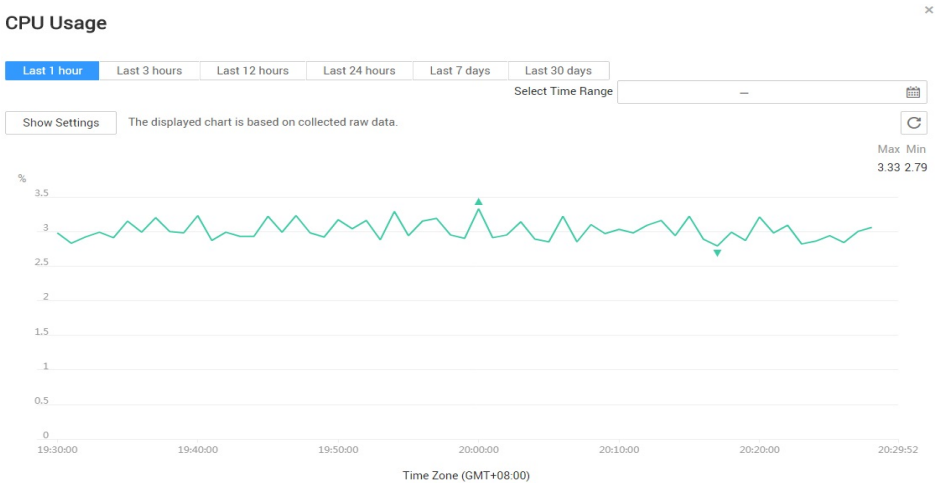
Step 3 Click in the upper right corner of the graphs to zoom in the graphs.

The system allows you to select a fixed time range or customize the time range.

1. Fixed time ranges include **Last 1 hour**, **Last 3 hours**, **Last 12 hours**, **Last 24 hours**, **Last 7 days**, and **Last 30 days**.
2. A customized time range can be specified within the latest seven days.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-113 Zoomed in monitoring graph



----End

6.1.15 Troubleshooting

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

Symptom

When CDM is used to migrate other data sources to Cloud Search Service, the job fails to be executed and the error message "Unparseable date" is displayed in the log. See **Figure 6-114**.

Figure 6-114 Log output

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

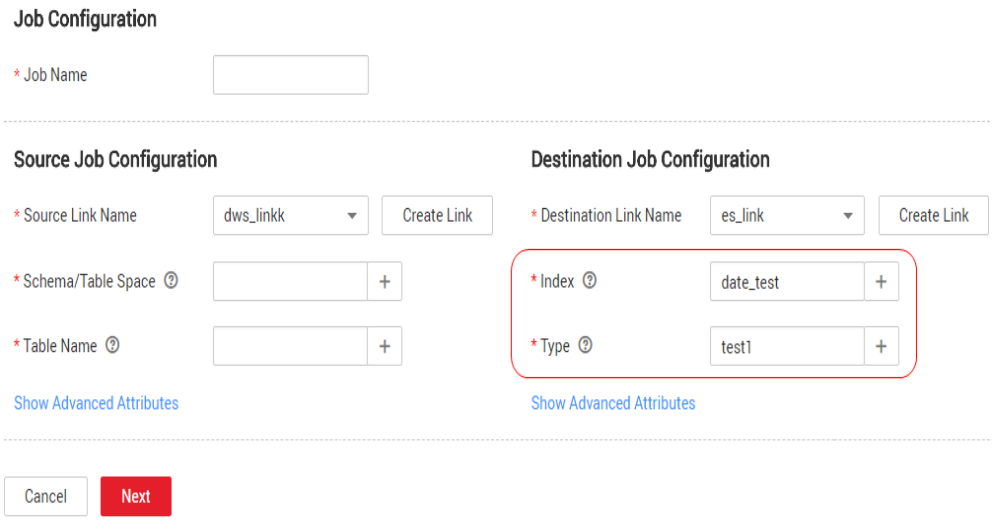
Possible Cause

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

In China, the displayed time is eight hours earlier than the actual time. Therefore, when you use CDM to migrate data to the cloud search service, if the index and type are automatically created by CDM (for example, in **Figure 6-115**, if **date_test** and **test1** of the destination end do not exist, CDM automatically creates the index and type in the cloud search service), CDM sets the format of the time type field to **yyyy-MM-dd HH:mm:ss.SSS Z** by default, for example, **2018-01-08 08:08:08.666 +0800**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-115 Job configuration

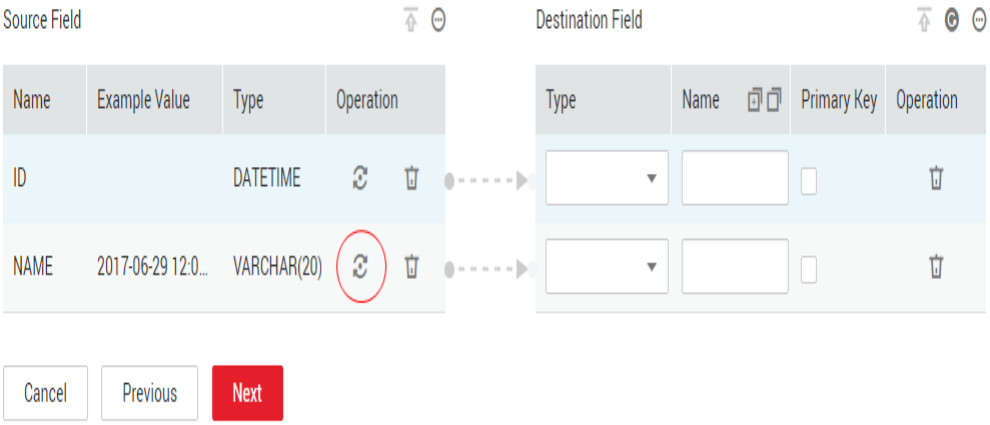


When data is imported from another data source to Cloud Search Service, if the date format in the source data is not the standard format, for example, **2018/01/05 15:15:46**, the CDM job fails to be executed, and the log shows that the date format cannot be parsed. You need to configure a field converter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

Solution

1. Edit the job and go to the **Map Field** tab page. Click the icon for creating a converter in the row of the source field to create a converter. See **Figure 6-116**.

   Figure 6-116 Creating a converter 03

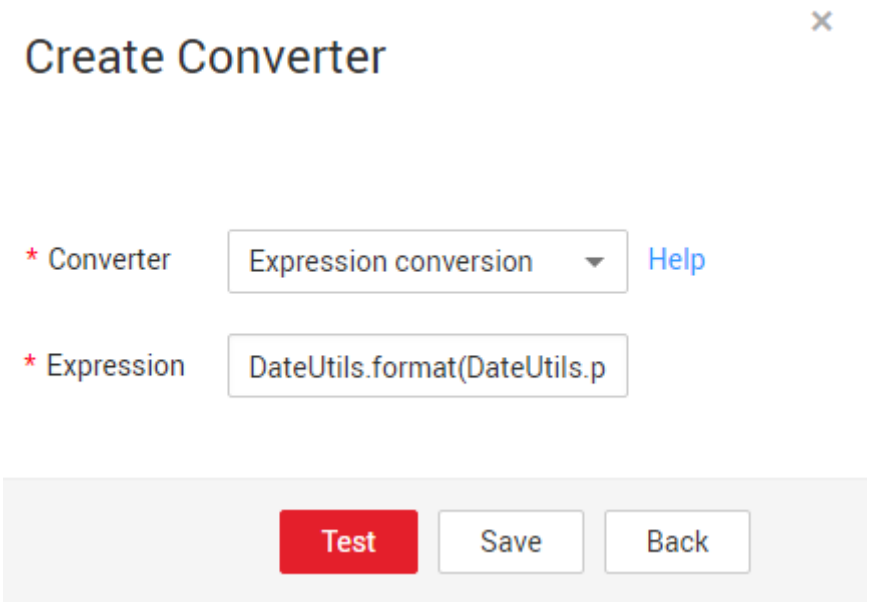
2. Select **Expression conversion** as the converter. Currently, expression conversion supports functions of the character string and date types. The syntax is similar to the Java character string and time functions. For details about how to compile the expression, see **Expression Conversion**.

 Copyright © Huawei Technologies Co., Ltd.

3. In this example, the source time format is *yyyy/MM/dd HH:mm:ss*. To convert the source time format to *yyyy-MM-dd HH:mm:ss.SSS Z*, perform the following operations:
   a. Add the time zone information **+0800** to the end of the original date character string. The corresponding expression is **value+" +0800"**.
   b. Use the original date format to parse the string to a date object. You can use the DateUtils.parseDate function for parsing. The syntax is **DateUtils.parseDate(String value, String format)**.
   c. Format the date object into a character string in target format by using the DateUtils.format function. The syntax is **DateUtils.format(Date date, String format)**.

   In this example, the complete expression is **DateUtils.format(DateUtils.parseDate(value+" +0800","yyyy/MM/dd HH:mm:ss Z"),"yyyy-MM-dd HH:mm:ss.SSS Z")**. See **Figure 6-117**.

   Figure 6-117 Configuring the expression 01



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

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

Symptom

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

Possible Cause

HBase/CloudTable is schema-less, and the number of columns in each data is not fixed. On the **Map Field** page, there is a high probability that all columns cannot

 Copyright © Huawei Technologies Co., Ltd.

be obtained by obtaining example values. In this case, the data on the migration destination is incomplete after the job is executed.

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

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

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

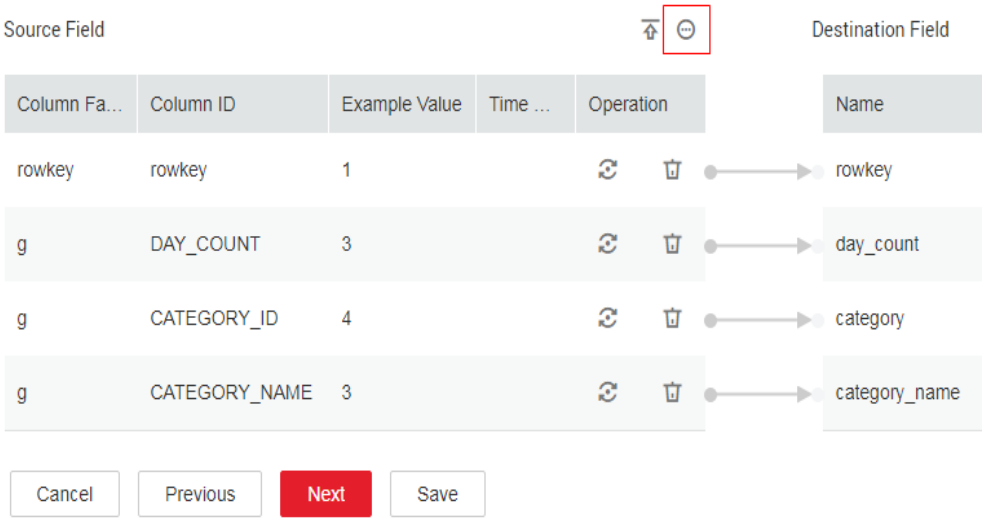
Solution 1: Adding Fields on the Map Field Tab Page

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

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

2. On the **Job Management** page, locate the job for exporting data from HBase to DWS, click **Edit** in the row where the job resides, and go to the **Map Field** tab page. See **Figure 6-118**.

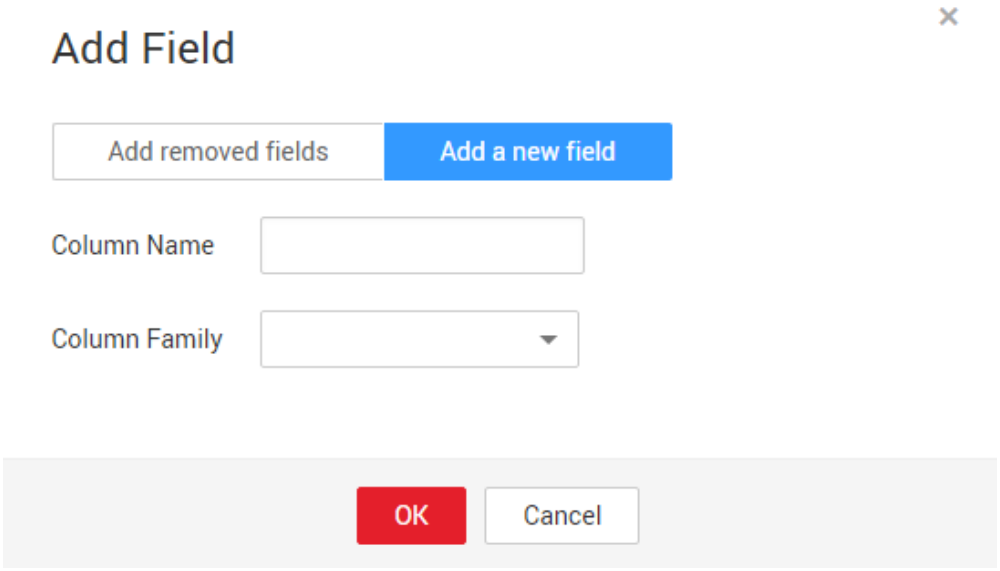
Figure 6-118 Field mapping 03



 Copyright © Huawei Technologies Co., Ltd.

3. Click ⊕. In the dialog box that is displayed, select **Add a new field**. See **Figure 6-119**.

   Figure 6-119 Adding a field 04

   NOTE

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

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

Solution 2: Modifying a JSON File

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

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

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

 Copyright © Huawei Technologies Co., Ltd.

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

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

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

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

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

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

```
"to-config-values": {
            "configs": [
               {
                  "inputs": [
                     {
                        "name": "toJobConfig.schemaName",
                        "value": "dbadmin"
```

 Copyright © Huawei Technologies Co., Ltd.

```
                    },
                    {
                        "name": "toJobConfig.tablePreparation",
                        "value": "DO_NOTHING"
                    },
                    {
                        "name": "toJobConfig.tableName",
                        "value": "DWS "
                    },
                    {
                        "name": "toJobConfig.columnList",
                        "value":
"rowkey&day_count&category&category_name&find_time&upload_people&id&infomation_
id&title&coordinate_x&coordinate_y&coordinate_z&content&images&state"
                    },
                    {
                        "name": "toJobConfig.shouldClearTable",
                        "value": "true"
                    }
                ],
                "name": "toJobConfig"
            }
        ]
    }
```

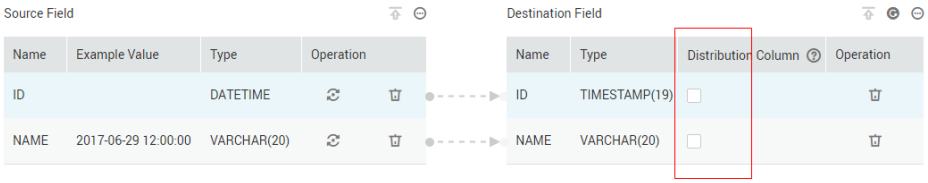
c. Retain the settings of other parameters, and then click **Save and Run**.

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

6.1.15.3 How Do I Select Distribution Columns When Using CDM to Migrate Data to DWS?

When using CDM to migrate data to DWS or FusionInsight LibrA and create a table on DWS, select the distribution columns on the **Map Field** tab page. See **Figure 6-120**.

Figure 6-120 Selecting distribution columns



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

1. Use the primary key as the distribution column.
2. If multiple data segments are combined as primary keys, specify all primary keys as the distribution column.
3. In the scenario where no primary key is available, if no distribution column is selected, DWS uses the first column as the distribution column by default. As a result, data skew risks exist.

Therefore, when a single table or entire database is imported to DWS/FusionInsight LibrA, you are advised to manually select a distribution column;

 Copyright © Huawei Technologies Co., Ltd.

otherwise, CDM automatically selects one. For more information about the distribution column, see **Selecting a Distribution Column**.

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

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

Symptom

When you use CDM to migrate data to DWS/FusionInsight LibrA, the migration fails and the error message "**value too long for type character varying**" is displayed in the log. See **Figure 6-121**.

Figure 6-121 Log output

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

Possible Cause

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

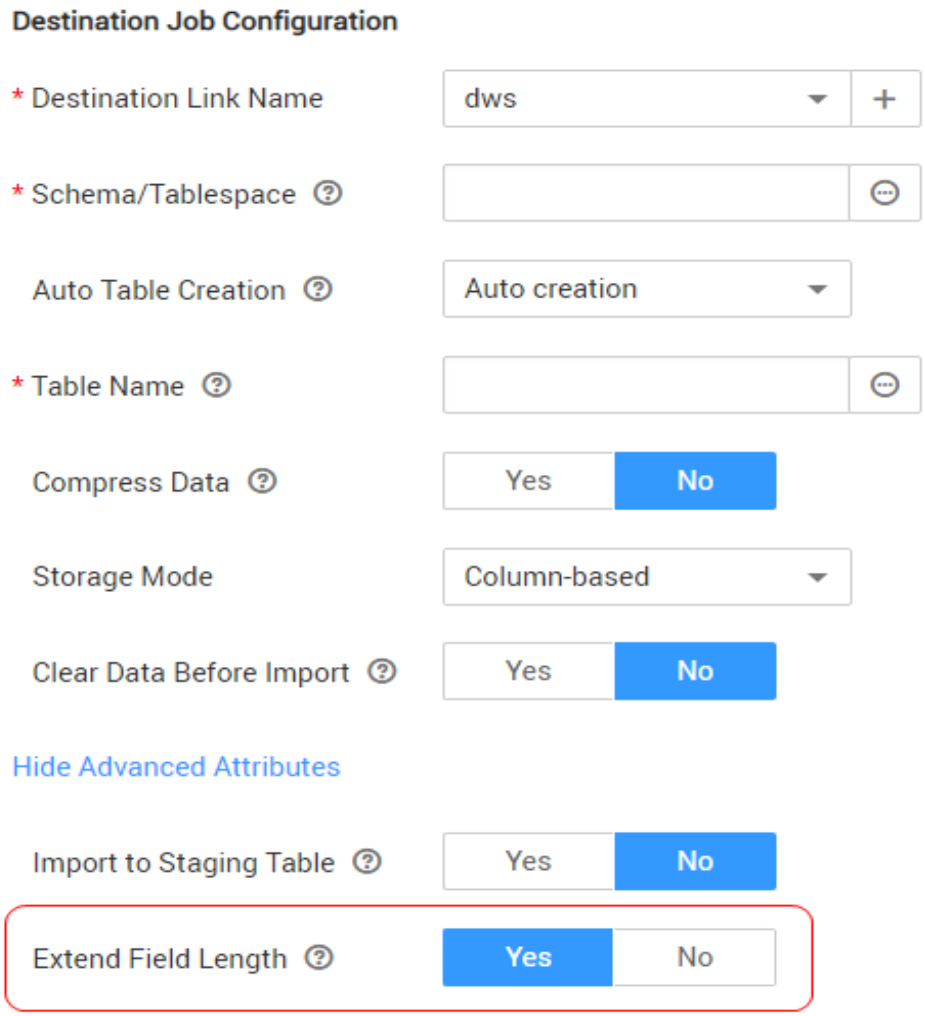
Solution

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

Edit the table/file migration job on CDM. In **Destination Job Configuration**, set **Auto Table Creation** to **Auto creation**, **Extend Field Length** is displayed in **Show Advanced Attributes**. Set **Extend Field Length** to **Yes**. See **Figure 6-122**.

Copyright © Huawei Technologies Co., Ltd.

Figure 6-122 Extending field length



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

Symptom

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

Possible Cause

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

Solution

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

 Copyright © Huawei Technologies Co., Ltd.

6.2 Real-time Data Ingestion

6.2.1 Overview

Data Ingestion Service (DIS) addresses the challenge of transmitting data from outside the cloud to inside the cloud. DIS builds data intake streams for custom applications capable of processing or analyzing streaming data. DIS continuously captures, transmits, and stores terabytes of data from hundreds of thousands of sources every hour, such as logs, social media feeds, website clickstreams, and location-tracking events.

6.2.2 Data Sources Supported by DIS

This section describes the data sources supported by DIS.

Data Source

Table 6-86 Data sources supported by DIS

Data Flow	Data Source Type	Operation Guide
Writing data to a stream	Source data types are BLOB, JSON, and CSV.	**Uploading Data by Using Agent**
	Flume Channel	**Using DIS Flume Plugin to Upload and Download Data**
	Logstash	**Using a DIS Logstash Plugin to Upload and Download Data**
	Kafka	**Using Kafka Adapter to Upload and Download Data**
Reading data from a stream	Flume Channel	**Using DIS Flume Plugin to Upload and Download Data**
	Logstash	**Using a DIS Logstash Plugin to Upload and Download Data**
	Kafka	**Using Kafka Adapter to Upload and Download Data**
	Spark Streaming	**Using SparkStreaming SDK to Download Data**
Dumping data to cloud services	OBS	**Dumping Data to OBS**
	DLI	**Dumping Data to DLI**
	DWS	**Dumping Data to DWS**
	MRS	**Dumping Data to MRS**

Copyright © Huawei Technologies Co., Ltd.

Data Flow	Data Source Type	Operation Guide
	CloudTable	**Dumping Data to CloudTable**

6.2.3 Quick Start

General Procedure

The following is the general procedure for using DIS:

Buying a Pay-Per-Use DIS Incremental Package

You need to create a stream before using DIS.

Preparing a DIS Application Development Environment

Before developing DIS applications, you need to install application development tools. Then, obtain the SDK and sample project and import them to the development environment.

Sending Data to a DIS Stream

Install an application based on data uploading services and run it to send data to the cloud. The DIS stream information can be viewed on the DIS console.

Obtaining Data from a DIS Stream

Write a consumer application and run it to retrieve data from the cloud.

Buying a Pay-Per-Use DIS Incremental Package

1. Select an enabled instance and click **Buy**.
2. Access the page for buying a DGC incremental package and set parameters. See **Table 6-87** for parameter details.

Copyright © Huawei Technologies Co., Ltd.

Figure 6-123 Configuring parameters for a DIS incremental package

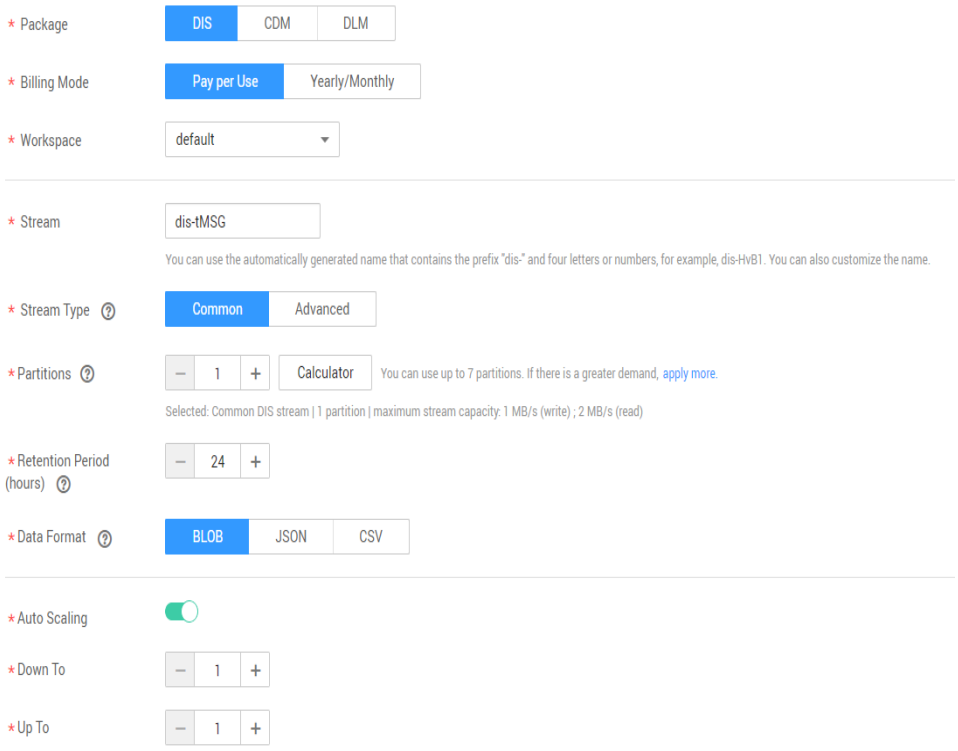


Table 6-87 Parameters for a DIS incremental package

Parameter	Description
Incremental Package Type	Select **DIS**.
Billing Mode	Select **On-demand**.
Workspace	Select the workspace where the DIS incremental package will be used. For example, if you want to buy a DIS incremental package for workspace ***A*** of instance ***test***, select ***A***. After you bought a DIS incremental package, it will show up in workspace ***A***.
Stream Name	The name of the DIS stream to be created. Stream names must be unique. They can be up to 64 characters long. Only letters, numbers, hyphens (-), and underscores (_) are allowed.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Stream Type	**Stream Type**: **Common** and **Advanced**. ● **Common**: Each partition supports a maximum read speed of 2 MB/s and a maximum write speed of 1 MB/s or 1,000 records/s. ● **Advanced**: Each partition supports a maximum read speed of 10 MB/s and a maximum write speed of 5 MB/s or 2,000 records/s.
Partitions	Partitions are the base throughput unit of a DIS stream. ● For a common DIS stream, the value of **Partitions** is an integer from 1 to 50. A user can create up 50 partitions. ● For an advanced DIS stream, the value of **Partitions** is an integer from 1 to 10. A user can create up to 10 partitions.
Partition Calculator	Used to estimate how many partitions are needed based on the information you entered. 1. Click **Partition Calculator**. 2. In the **Partition Calculator** dialog box, configure the **Average Record Size (KB)**, **Max. Records Written**, and **Consumer Applications** parameters. The **Estimated Partitions** field then displays the recommended number of partitions. The value of this field cannot be modified. **NOTE** Partition calculation formulas: ● Based on the traffic (the final value must be rounded up): Common stream: Average record size x (1 + 20%) x Max. written records/(1 x 1024 KB) (20% is the reserved partition percentage.) Advanced stream: Average record size x (1 + 20%) x Max. written records/(5 x 1024 KB) (20% is the reserved partition percentage.) ● Based on the consumer application quantity (the final value must be rounded up): (The number of consumer applications/2) x The number of partitions calculated based on the traffic (The result of the number of consumer applications/2 must reserve two decimals.) The larger value calculated based on the previous two formulas is considered as the estimated partition value. 3. Click **Use Estimated Value**. The estimated value is automatically used as the value of **Partitions**.
Data Retention (days)	The maximum number of days for which data can be preserved in DIS streams. Data will be deleted when the retention period expires.

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Source Data Type	• **BLOB**: a collection of binary data stored as a single entity in a database management system. If **Source Data Type** is set to **BLOB**, the supported **Dump Destination** can be **OBS** or **MRS**. • **JSON**: an open-standard file format that uses easy-to-read text to transmit data objects consisting of attribute values and arrays. If **Source Data Type** is set to **JSON**, the supported **Dump Destination** can be **OBS**, **MRS**, **DLI**, **CloudTable**, or **DWS**. • **CSV**: a simple text format for storing tabular data in a plain text file. Commas are used as delimiters. If **Source Data Type** is set to **CSV**, the supported **Dump Destination** can be **OBS**, **MRS**, **DLI**, or **DWS**.
Auto Scaling	You can enable or disable the auto scaling switch when creating a stream. Set the switch to or , respectively, to enable or disable auto scaling. **NOTE** You can choose whether to enable auto scaling when creating a stream. You can also modify the auto scaling attributes for a stream that has already been created.
Down To	The lower limit for automatic scale-down. The number of target partitions for automatic scale-down cannot drop below this lower limit.
Up To	The upper limit for automatic scale-up. The number of target partitions for automatic scale-up cannot exceed this upper limit.

3. Click **Buy Now**. If there are no changes made to the specifications, submit the order.
4. After you bought a DIS stream, it will show up in the corresponding workspace.

Figure 6-124 Viewing a stream

Preparing a DIS Application Development Environment

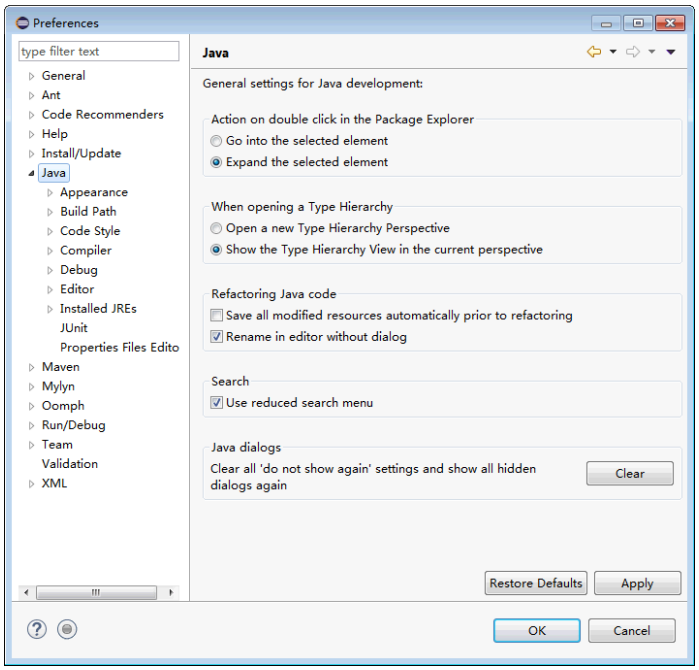
- JDK 1.8 or later has been installed.

 Copyright © Huawei Technologies Co., Ltd.

- Eclipse has been installed.

1. Configure a JDK using Eclipse.
   a. Start Eclipse and choose **Window** > **Preferences**. The **Preferences** dialog box is displayed.
   b. In the navigation tree, choose **Java**. On the **Java** page, configure general settings for Java development and then click **OK**. See **Figure 6-125**.

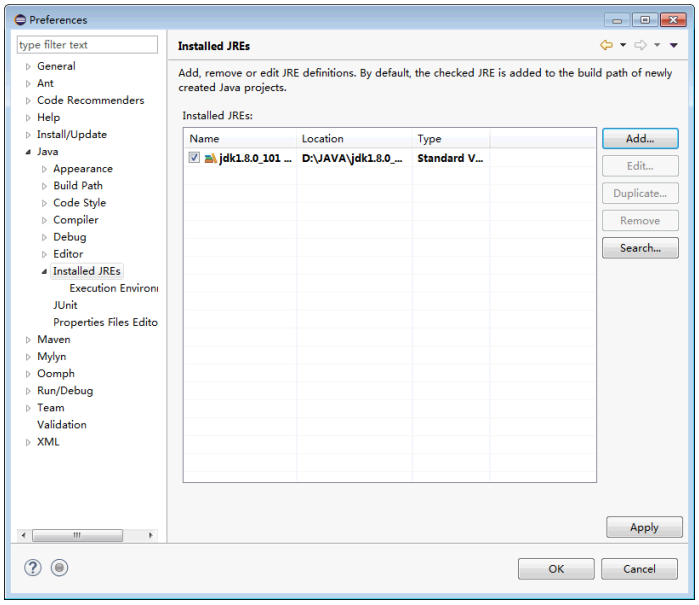
Figure 6-125 Preferences



   c. In the navigation tree, choose **Java** > **Installed JREs**. See **Figure 6-126**.
      - Ensure that configured JDK environmental variables are displayed on the **Installed JREs** page. Then go to **1.c.i**.
      - To configure different variables for different versions of JDK, perform **1.c.ii** to **1.c.iv**.

 Copyright © Huawei Technologies Co., Ltd.

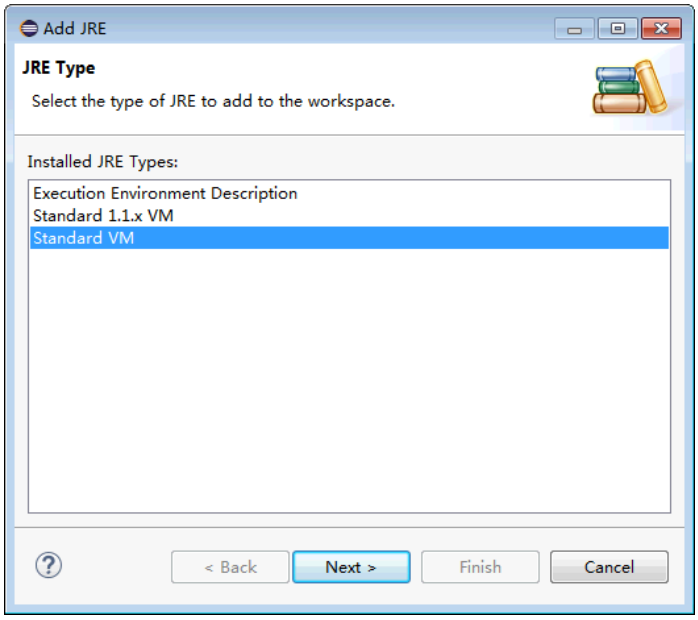
Figure 6-126 Installed JREs



i. Select the installed JDK and click **OK**.

ii. Click **Add**. The **Add JRE** dialog box is displayed. See **Figure 6-127**.

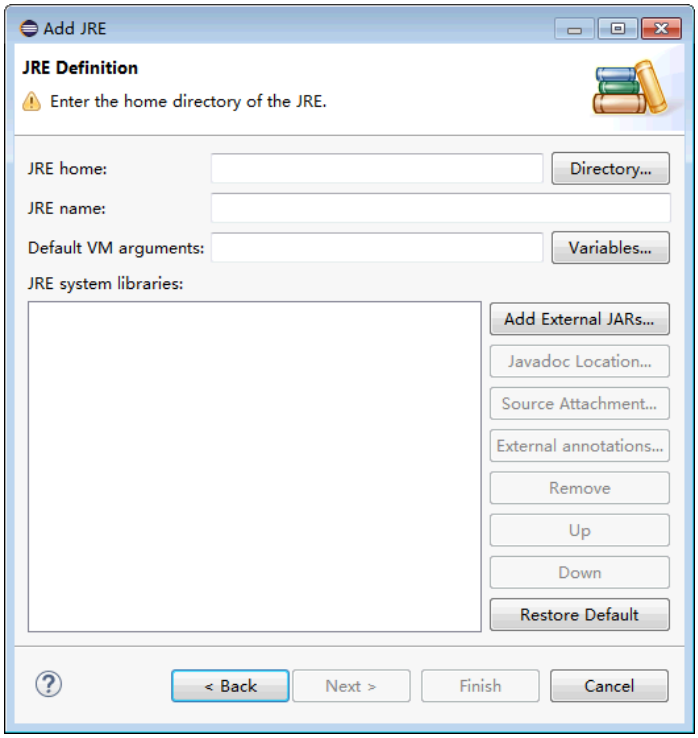
Figure 6-127 JRE Type



iii. Select a JRE type and click **Next**. See **Figure 6-128**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-128 JRE Definition

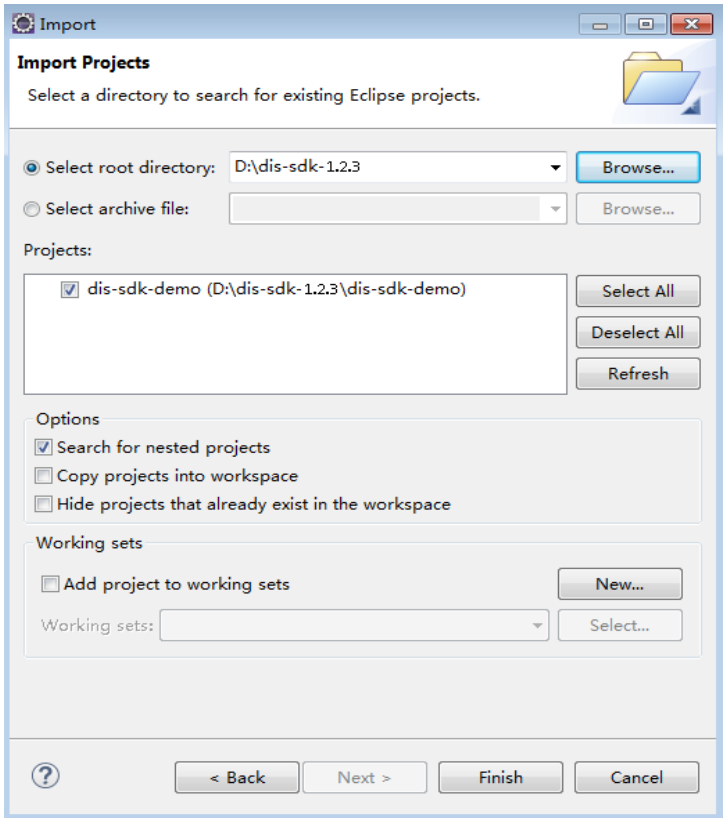


iv. Configure the basic information about JDK and click **Finish**.
   - JRE home: JDK installation path.
   - Default VM arguments: JDK running parameters.

2. Download the **huaweicloud-sdk-dis-java-*X.X.X*.zip** package from **DIS SDK** and decompress it.
3. Import the Eclipse project.
   a. Start Eclipse. Choose **File** > **Import**. The **Import** dialog box is displayed.
   b. Choose **Maven** > **Existing Maven Projects**, and click **Next**. The **Import** dialog box is displayed.
   c. Click **Browse** and select a save location for the **dis-sdk-demo** sample project. In the **Projects** area, select a sample project. See **Figure 6-129**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-129 Import Maven Projects1



d. Click **Finish**.

4. Import the Eclipse project.

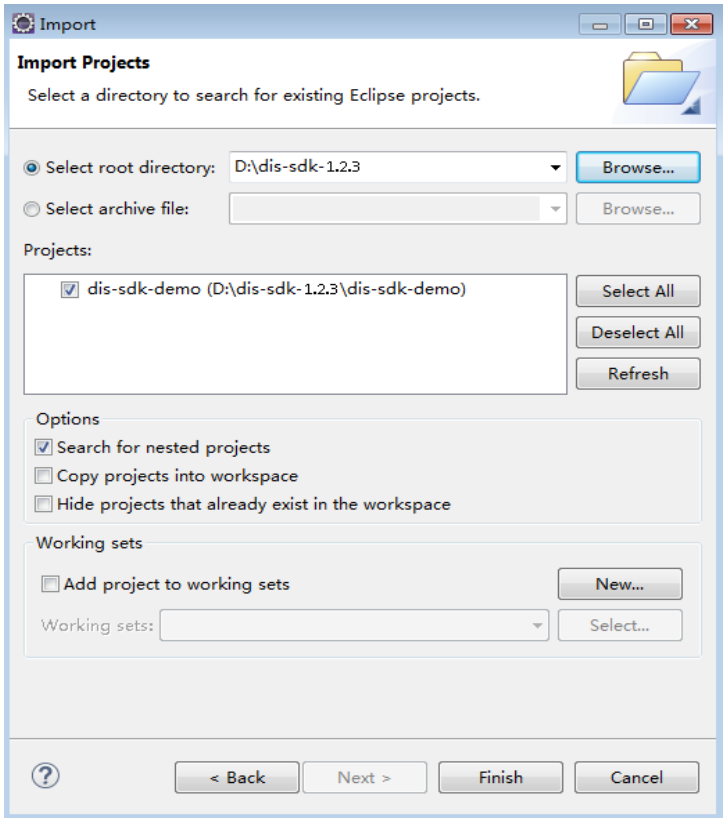
   a. Start Eclipse. Choose **File** > **Import**. The **Import** dialog box is displayed.

   b. Choose **General** > **Existing Projects into Workspace** and click **Next**. The **Import** dialog box is displayed.

   c. Click **Browse** and select a save location for the **dis-sdk-demo** sample project. In the **Projects** area, select a sample project. See **Figure 6-130**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-130 Import Projects



d. Click **Finish**.

5. Import the maven project.

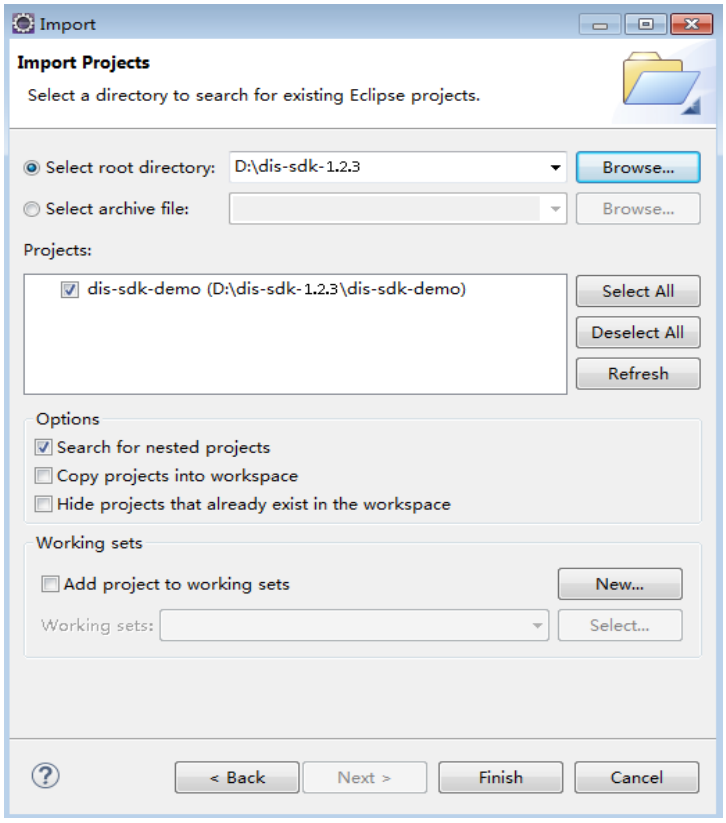
   a. Start Eclipse. Choose **File** > **Import**. The **Import** dialog box is displayed.

   b. Choose **Maven** > **Existing Maven Projects**, and click **Next**. The **Import Maven Projects** dialog box is displayed.

   c. Click **Browse** and select a save location for the **dis-sdk-demo** sample project. In the **Projects** area, select a sample project. See **Figure 6-131**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-131 Import Maven Projects2



d. Click **Finish**.

6. Configure the demo project.

   a. Set the project code to UTF-8.

      i. In the navigation tree on the left, right-click the required project under **Project Explorer** and choose **Properties** from the shortcut menu. The **Properties for dis-sdk-demo** page is displayed.

      ii. In the left list, select **Resource**. The **Resource** pane is displayed.

      iii. Set **Text file encoding** to **Other**. In the **Others** drop-down list, select **UTF-8**.

      iv. Click **Apply and Close** to complete the encoding configuration.

   b. Import the JAR in dependency package.

      i. In the navigation pane, choose **Project Explorer**. Right-click the chosen project and choose **Properties** from the shortcut menu.

      ii. In the left list, select **Java Build Path**. The **Java Build Path** pane is displayed.

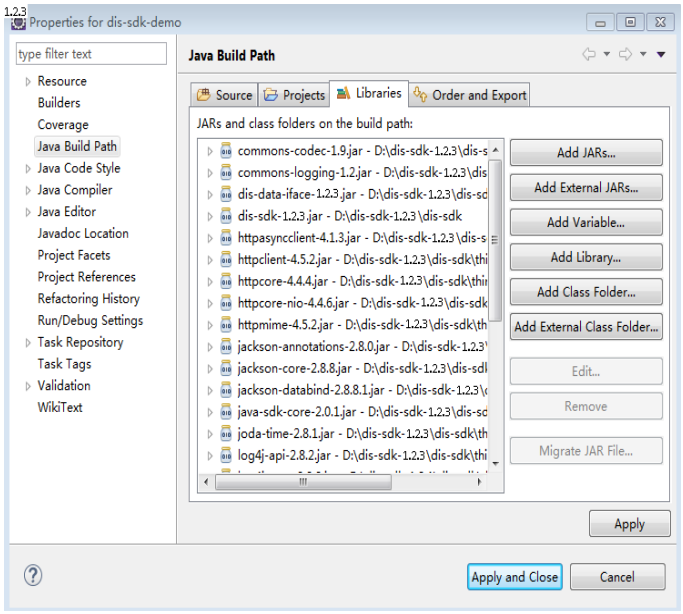
      iii. Click the **Libraries** tab, and then click **Add External JARs**. The **JAR Selection** dialog box is displayed.

      iv. Select the directory where the **dis-sdk** folder is located and click **Open**.

      v. On the **Properties for dis-sdk-demo** page, click **Apply and Close** to import all JAR files from the current path and the **third_lib** folder to the project.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-132 Java Build Path



c. Add a JDK.

i. In the navigation pane, choose **Project Explorer**. Right-click the chosen project and choose **Properties** from the shortcut menu.

ii. In the left list, select **Java Build Path**. The **Java Build Path** pane is displayed.

iii. Click the **Libraries** tab, and then click **Add Library**. The **Add Library** dialog box is displayed.

iv. Select **JRE System Library** and click **Next**. Ensure that the version of **Workspace default JRE** is jdk1.8 or later.

v. Click **Finish** to exit the **Add Library** dialog box.

vi. Click **Apply and Close** to add the JDK.

7. Initialize a DIS client sample. For details about endpoint, ak, sk, region and project ID, see **Obtaining Authentication Information**.

Sending Data to a DIS Stream

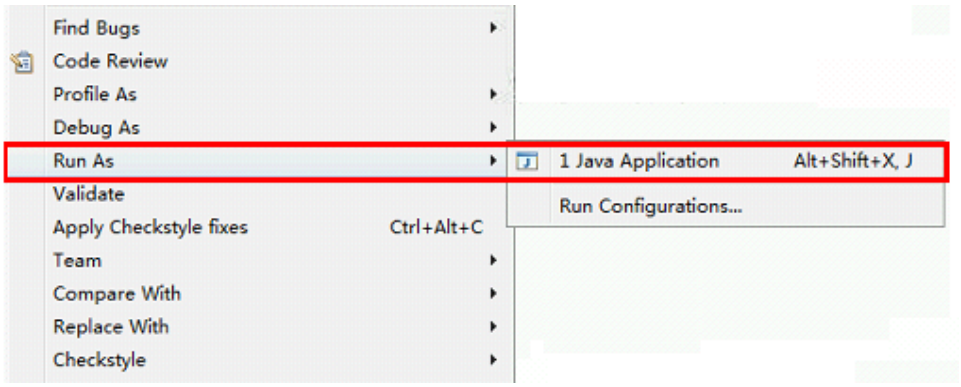
Local data is continuously uploaded to DIS.

The example code file is the **ProducerDemo.java** file in the **\dis-sdk-demo\src\main\java\com\bigdata\dis\sdk\demo** directory decompressed from the **huaweicloud-sdk-dis-java-*X.X.X*.zip** package. The compression package is downloaded in **Preparing a DIS Application Development Environment**.

Right-click the producer application and choose **Run As** > **1 Java Application** from the shortcut menu. See **Figure 6-133**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-133 Running the application



While data is being sent to DIS, the DIS console displays DIS stream information. If information similar to the following is displayed, the data has been successfully sent to DIS:

```
14:40:20.090 [main] INFOcom.bigdata.dis.sdk.DISConfig - get from classLoader
14:40:20.093 [main] INFODEMOT - ========== BEGIN PUT ============
14:40:21.186 [main] INFOcom.bigdata.dis.sdk.util.config.ConfigurationUtils - get from classLoader
14:40:21.187 [main] INFOcom.bigdata.dis.sdk.util.config.ConfigurationUtils - propertyMapFromFile size : 2
14:40:22.092 [main] INFOcom.bigdata.dis.sdk.demo.ProducerDemo - Put 3 records[3 successful / 0 failed].
14:40:22.092 [main] INFOcom.bigdata.dis.sdk.demo.ProducerDemo - [hello world.] put success, partitionId
[shardId-0000000000], partitionKey [964885], sequenceNumber [0]
14:40:22.092 [main] INFOcom.bigdata.dis.sdk.demo.ProducerDemo - [hello world.] put success, partitionId
[shardId-0000000000], partitionKey [910960], sequenceNumber [1]
14:40:22.092 [main] INFOcom.bigdata.dis.sdk.demo.ProducerDemo - [hello world.] put success, partitionId
[shardId-0000000000], partitionKey [528377], sequenceNumber [2]
14:40:22.092 [main] INFOcom.bigdata.dis.sdk.demo.ProducerDemo - ========== PUT OVER ============
```

Obtaining Data from a DIS Stream

The example code file is the **ConsumerDemo.java** file in the **\dis-sdk-demo\src\main\java\com\bigdata\dis\sdk\demo** directory decompressed from the **huaweicloud-sdk-dis-java-*X.X.X*zipdis-sdk-XXX.zip** package. The compression package is downloaded in **Preparing a DIS Application Development Environment**.

If information similar to the following appears, data has been successfully retrieved from DIS:

```
14:55:42.954 [main] INFOcom.bigdata.dis.sdk.DISConfig - get from classLoader
14:55:44.103 [main] INFOcom.bigdata.dis.sdk.util.config.ConfigurationUtils - get from classLoader
14:55:44.105 [main] INFOcom.bigdata.dis.sdk.util.config.ConfigurationUtils - propertyMapFromFile size : 2
14:55:45.235 [main] INFOcom.bigdata.dis.sdk.demo.ConsumerDemo - Get stream
streamName[partitionId=0] cursor success :
eyJnZXRJdGVyYXRvclBhcmFtIjp7InN0cmVhbS1uYW1lIjoiZGlzLTEzbW9uZXkiLCJwYXJ0aXRpb24taWQiOiIwIiwiY
3Vyc29yLXR5cGUiOiJBVF9TRVFVRU5DRV9OVU1CRVIiLCJzdGFydGluZy1zZXF1ZW5jZS1udW1iZXIiOiIxMDY4O
TcyIn0sImdlbmVyYXRlVGltZXN0YW1wIjoxNTEzNjY2NjMxMTYxfQ
14:55:45.305 [main] INFOcom.bigdata.dis.sdk.demo.ConsumerDemo - Get Record [hello world.],
partitionKey [964885], sequenceNumber [0].
14:55:45.305 [main] INFOcom.bigdata.dis.sdk.demo.ConsumerDemo - Get Record [hello world.],
partitionKey [910960], sequenceNumber [1].
14:55:46.359 [main] INFOcom.bigdata.dis.sdk.demo.ConsumerDemo - Get Record [hello world.],
partitionKey [528377], sequenceNumber [2].
```

6.2.4 Managing DIS Streams

Copyright © Huawei Technologies Co., Ltd.

6.2.4.1 Viewing Stream Monitoring Metrics

You can view stream monitoring metrics on the DIS console.

1. In the stream list, click the name of the DIS stream whose monitoring metrics will be viewed. The monitoring page is displayed.
2. On the **Monitoring** page, click the **Streams** or **Partitions** tab to view stream or partition monitoring metrics. **Table 6-88** describes the monitoring parameters.

Table 6-88 DIS monitoring information

Parameter	Description
Time Range	• Monitoring time range. Values: – 1h – 2h – 3h • You can customize the time range for viewing monitoring information. – Click next to **Custom**, and then set the start time and end time. – The end time cannot be later than the current system time. – The difference between the start time and end time cannot exceed 168 hours.
Partitions	
Partition ID	ID of the partition. It starts from 0 by default. Select any of the following values from the **Partition ID** drop-down.
Data Rate (KB/s)	Rates at which data is sent to and retrieved from the chosen partition within the specified time range. Unit: KB/s
Records Per Second	The number of records sent to and retrieved from the chosen partition within the specified time range.
Streams	
Data Rate (KB/s)	Rates at which data is sent to and retrieved from the chosen DIS stream within the specified time range. Unit: KB/s
Records Per Second	The number of records sent to and retrieved from the chosen DIS stream within the specified time range.

Copyright © Huawei Technologies Co., Ltd.

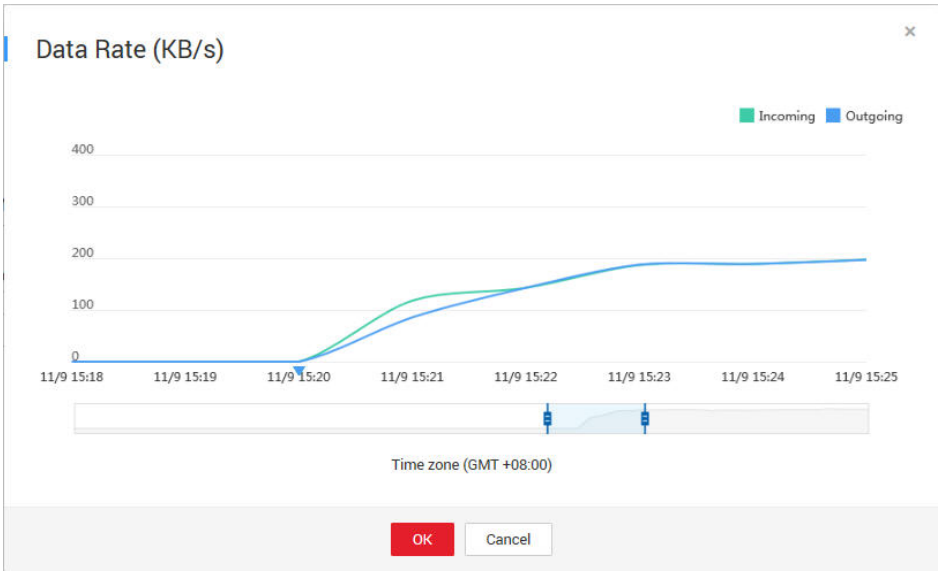
Parameter	Description
Successful Requests Per Second	The number of PutRecords and GetRecords requests successfully fulfilled within the specified time range.
Throttled Requests Per Second	The number of PutRecords and GetRecords requests rejected within the specified time range due to flow control.
Average Request Processing Time (ms)	The average amount of time spent in processing a PutRecords or GetRecords request.

3. In the upper right corner of the graph, click to enlarge the graph for viewing details.
4. You can adjust the positions of the sliders to view the monitoring metrics within the time range specified by the sliders.

   NOTE

   The time range specified by the sliders cannot exceed the total time range defined on the monitoring panel.

Figure 6-134 Adjusting sliders



6.2.4.2 Changing a Source Data Type

The source data schema is required for data conversion of a specific dump task in a stream. If it is incorrectly configured, data conversion fails and the dump task is abnormal. You can configure the source data schema when creating a stream or when creating a dump task. You can modify the configured source data schema on the stream details page.

1. Click the name of a stream that you want to view. The **Stream Management** page is displayed.

 Copyright © Huawei Technologies Co., Ltd.

2. Click next to **Source Data Type** and select a desired source data type from the drop-down list. Alternatively, in the **Operation** column of the stream for which you want to change its source data type, choose **More** > **Change Source Data Type**. In the displayed dialog box, change the source data type.

NOTE

- You can change the source data type for a stream whose **Source Data Type** is **BLOB**, **JSON**, or **CSV** only when the stream has no dump tasks.
- After you change the source data type of a stream for which you have configured source data schema, the existing source data schema will become invalid and cannot be recovered. You need to reconfigure the source data schema for the stream.

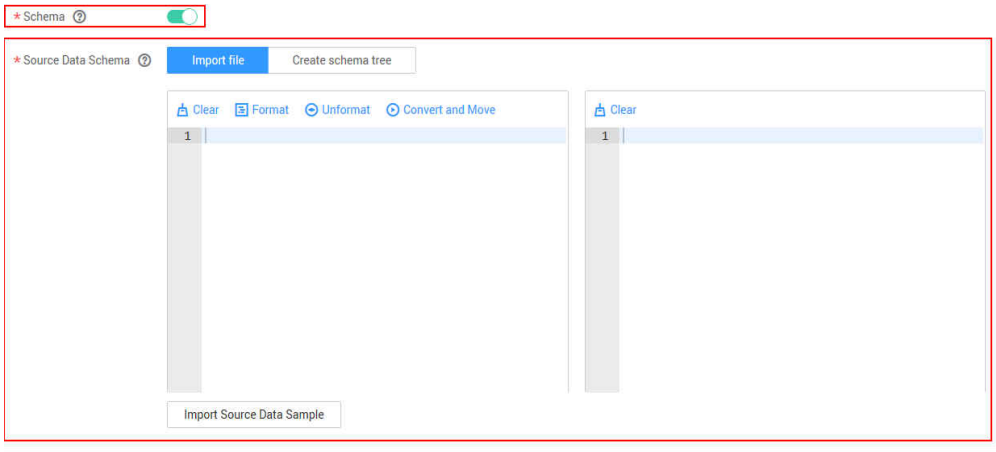
6.2.4.3 Managing a Source Data Schema

A source data schema is a user's JSON or CSV data sample used to describe the JSON or CSV data format. For example, DIS can generate an Avro schema based on the JSON data sample and convert the JSON data uploaded to a stream to the Parquet or CarbonData format.

Three entrances are available for creating a source data schema:

- Enable **Schema** when creating a stream. For details, see **Figure 6-135**.
- Keep **Schema** disabled when creating a stream. After the stream is created, choose **Stream Management** in the navigation tree and click the created stream. Click **Create Source Data Schema** next to **Source Data Type**. For details, see **Figure 6-136**.
- Keep **Schema** disabled when creating a stream. After the stream is created, choose **Stream Management** in the navigation tree and click the created stream. On the **Dump Management** tab page, click **Create Dump Task**. On the displayed page, create a source data schema. For details, see **Figure 6-137**.

Figure 6-135 Method 1 for creating a schema



 Copyright © Huawei Technologies Co., Ltd.

Figure 6-136 Entrance 2 for creating a schema

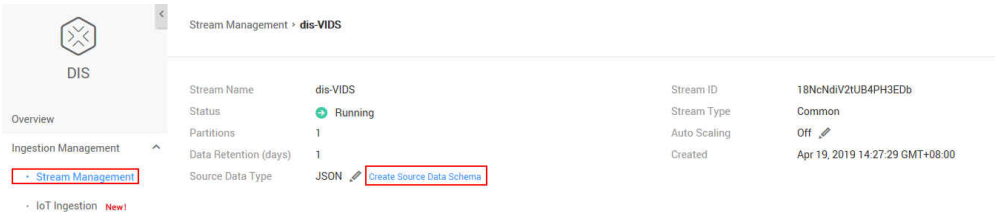
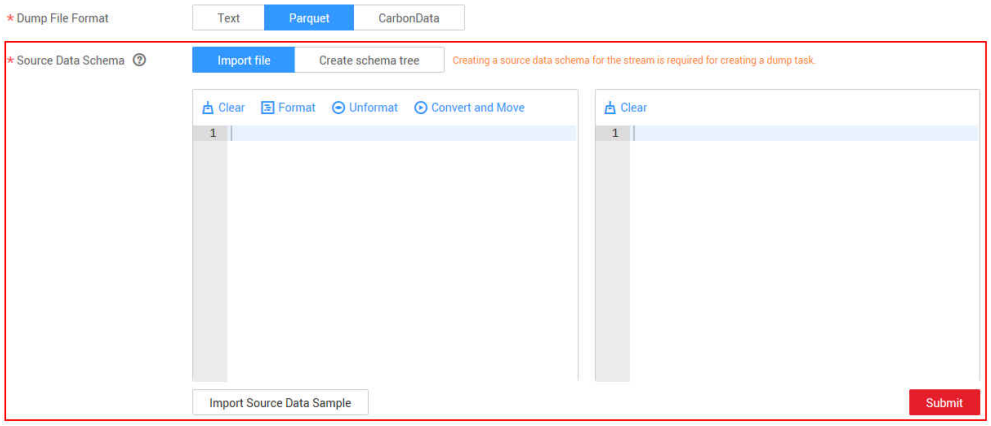


Figure 6-137 Entrance 3 for creating a schema



Creating a Schema for Source Data by Importing Files

Use the following method to create a source data schema:

1. When configuring **Source Data Schema**, click **Import File**.
2. In the left text box, enter a JSON or CSV source data sample or click Import Source Data Sample to import a source data sample.

   NOTE

   When importing source data samples, you can import only .txt, .json, .csv, and .java files.

3. To delete the source data sample, click .
4. In the left text box, click to generate an Avro schema in the right text box according to the source data sample.
5. In the right text box, click to modify the Avro schema.
6. To format the parsed data, click **Format** and **Unformat**.

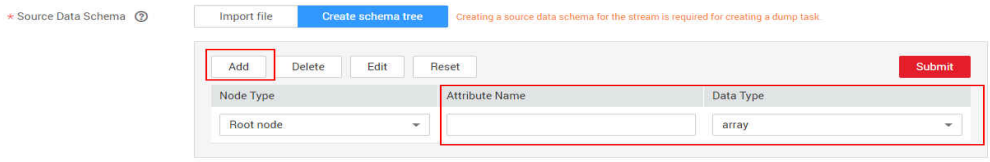
Creating a Schema for Source Data by Creating a Schema Tree

Use the following method to create a source data schema:

1. When configuring **Source Data Schema**, click **Create Schema Tree**.
2. After configuring an attribute name and data type, click **Add** to add a root node. For details, see **Figure 6-138**.

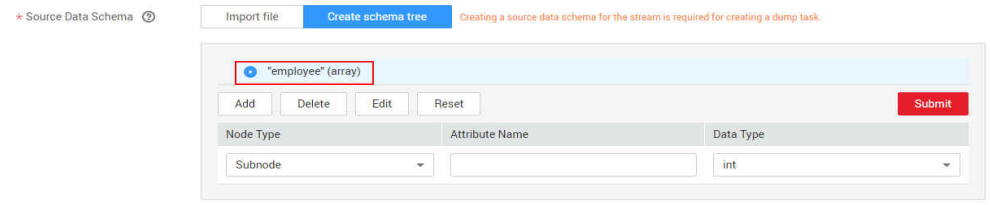
 Copyright © Huawei Technologies Co., Ltd.

Figure 6-138 Adding the root node



3. Select the created root node and configure an attribute name and data type in the same way to add subnodes.

Figure 6-139 Creating a subnode



NOTE

- To delete a node, select the check box of the node and click **Delete**.
- To edit the attributes of a node, select the check box of the node and click **Edit**.
- To delete all nodes, click **Reset**.

4. Click **Submit**.

Modifying a Source Data Schema

NOTE

Do not modify the source data schema of a stream if the stream has dump tasks.

1. Click a stream name. The **Stream Management** page is displayed.
2. Click **View Existing Source Data Schema** next to **Source Data Type**.
3. In the **Source Data Schema** text box, click to modify the source data schema.

Figure 6-140 Modifying the source data schema

 Copyright © Huawei Technologies Co., Ltd.

NOTE

If a stream has dump tasks, modifying the source data schema of the stream will cause some data unable to be successfully dumped.

4. After the modification is complete, click **Submit**. Click **Cancel** to give up modifying the source data schema.

6.2.4.4 Managing Stream Tags

A tag is an identifier of a stream. Adding tags to streams can help you identify and manage your stream resources.

You can add a maximum of 10 tags to a stream when creating the stream or add them on the details page of the created stream.

A tag consists of a tag key and a tag value. **Table 6-89** describes the rules for naming the tag key and value.

Table 6-89 Naming rules for a tag key and value

Parameter	Rule	Example
Key	A tag key cannot be left blank. A tag key must be unique for a stream. A tag key contains a maximum of 36 characters. A tag value cannot contain special characters =*<> \,|/ or start or end with a space.	Organization
Tag value	A tag value contains a maximum of 43 characters. A tag value cannot contain special characters =*<> \,|/ or start or end with a space. This parameter can be left blank.	Apache

Adding a Tag to a Stream

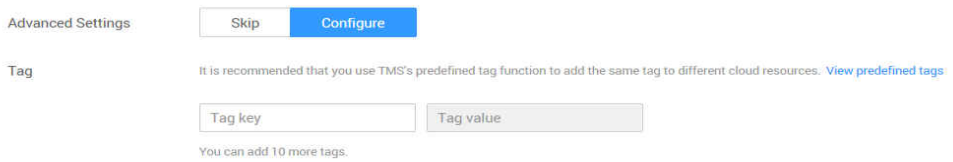
You can add a tag to a stream on the Buy Stream page.

1. Click Buy Stream. The Buy Stream page is displayed.
2. On the **Advanced Settings** tab page, select **Configure**.

   Enter the key and value of a tag to be added.

   You can add a maximum of 10 tags to the stream and use intersections of tags to search for the target stream.

Figure 6-141 Adding a tag



 Copyright © Huawei Technologies Co., Ltd.

NOTE

You can also add tags to existing streams. For details, see **Managing Tags**.

Searching for a Target Stream

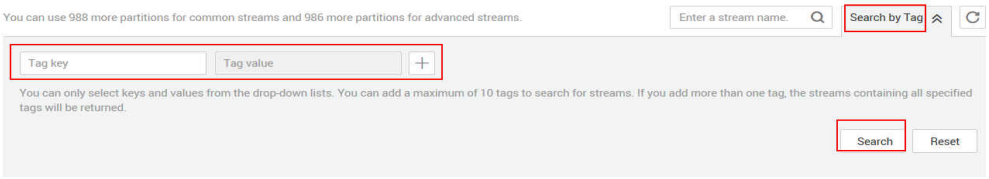
You can search for a target stream by tag on the **Stream Management** page.

1. In the navigation tree, choose **Ingestion Management** > **Stream Management**. In the upper right corner of the page, click **Search by Tag**.
2. Enter the tag key and value of the stream you are searching for.

   You can select a tag key or tag value from its drop-down list. When the tag key or tag value is exactly matched, the system can automatically locate the target stream. If you enter multiple tags, their intersections are used to search for the stream.
3. Click **Search**.

   The system searches for the target stream by tag key or value.

Figure 6-142 Searching for a target stream



Managing Tags

You can add, delete, modify, and view tags on the **Tags** tab page of a stream.

1. In the navigation tree, choose **Ingestion Management** > **Stream Management**. Click a stream to which the tags to be managed belong to.

   The stream details page is displayed.
2. Click the **Tags** tab and add, deleted, modify, and view tags.
   - View

     On the **Tags** tab page, you can view details about tags of the stream, including the number of tags and the key and value of each tag.
   - Add

     Click **Add Tag** in the upper left corner. In the displayed **Add Tag** dialog box, enter the key and value of the tag to be added, and click **OK**.
   - Modify

     In the **Operation** column of a tag, click **Edit**. In the displayed **Edit Tag** page, enter a new tag key and value and click **OK**.
   - Delete

     In the **Operation** column of the tag, click **Delete**. After confirmation, click **OK** on the displayed **Delete Tag** page.

 Copyright © Huawei Technologies Co., Ltd.

6.2.4.5 Managing Applications

An application is a program identifier. Multiple applications can access data in the same stream. Checkpoints generated for each application are used to record the consumed data in the stream by each application.

You can create applications and view details about existing application on the DIS console.

Creating an Application

Step 1 Log in to DGC, click **Access** on an instance, and choose **Data Integration**. Then, go to the **Stream Management** page.

Step 2 In the navigation tree, select **App Management**.

On the **Applications** page, click **Create App** and enter an application name.

----End

Viewing an Application

Step 1 Log in to DGC, click **Access** on an instance, and choose **Data Integration**. Then, go to the **DIS** module.

Step 2 In the navigation tree on the left, choose **Stream Management**.

Step 3 Click the name of a stream that you want to view.

Step 4 Click **App Management** to view all applications that access the stream.

You can view the names, IDs, and creation time of the applications that access the stream.

After you click **Remove All Checkpoints**, all checkpoints of the application will be deleted.

NOTE

When an application consumes data, the latest SN of the consumed data is recorded as a checkpoint. When the data is reconsumed, the consumption can be continued based on this checkpoint.

Figure 6-143 Viewing applications

----End

6.2.4.6 Managing Authorization

Authorized users can upload data to or download data from DIS after an upload or download permission is granted.

Step 1 Click a stream that you want to view.

 Copyright © Huawei Technologies Co., Ltd.

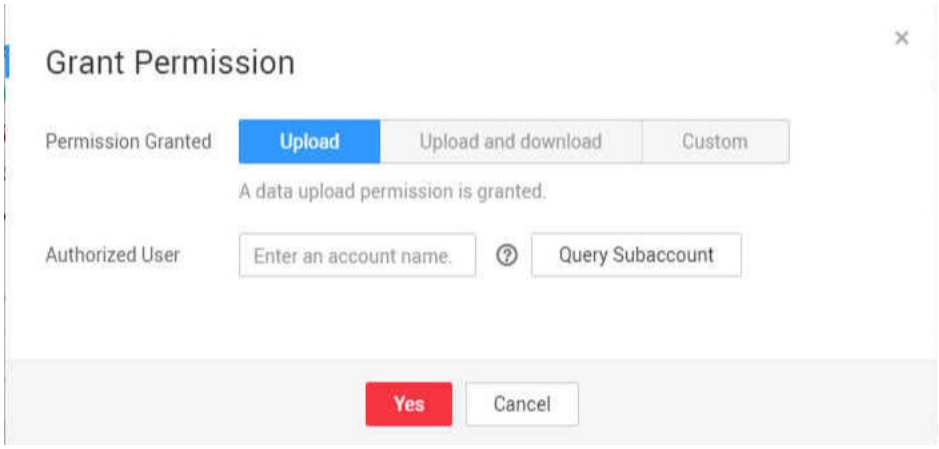
Step 2 On the **Permissions** tab page, click **Grant Permission**.

Select a permission, and then set the user information in the **Authorized User** text box.

NOTE

- To authorize all accounts, enter a wildcard (*).
- To authorize multiple accounts, enter these account names. These account names must be separated by commas (,).
- To authorize a subaccount of an account, enter the account name and then click **Query Subaccount**.

Figure 6-144 Granting permissions



----End

6.2.4.7 Scaling Up/Down a Stream

After a stream is created successfully, you can scale up or down a stream by adding or reducing partitions to meet capacity change requirements. Streams can be scaled automatically or manually.

Function

Scaling rules

- When scaling up a stream, ensure that the number of target partitions is greater than the number of the current partitions but is not greater than the total number of the remaining quotas and current partitions.
- When scaling up a stream, restore the **DELETED** and **EXPIRED** partitions to the **ACTIVE** partitions to make the partitions accessible. If the stream capacity is still insufficient, the system will create new partitions.
- When scaling down a stream, ensure that the number of target partitions is smaller than the number of current partitions but is not smaller than **1**.
- The partitions that have been scaled down will not be charged or occupy the quota. During the lifecycle configured in the incremental package that is billed in **pay-per-use** mode, data can be read from but cannot be written to the partitions that are successfully scaled in. After the lifecycle elapses, data cannot be read from or written to the partitions.

 Copyright © Huawei Technologies Co., Ltd.

- A maximum of 10 automatic scale-up, 10 manual scale-up, and 1 scale-down operations can be performed for each stream within one hour.
- The interval between performing automatic scale-up and scale-down must be greater than 1 minute. Automatic scale-down cannot be triggered within 2 minutes after automatic or manual scale-up is performed, and meanwhile automatic scale-up cannot be triggered within 2 minutes after automatic or manual scale-down is performed.

Precautions

- When uploading data, you need to set PartitionKey based on your specific demands so that DIS can automatically upload data to all partitions according to your requirements.
- When downloading data, you need to periodically use the descriptStream API to check the partition quantity change so that DIS can download data from all partitions.

Manually Scaling Up/Down a Stream

Use either of the following methods to manually scale up or down a stream:

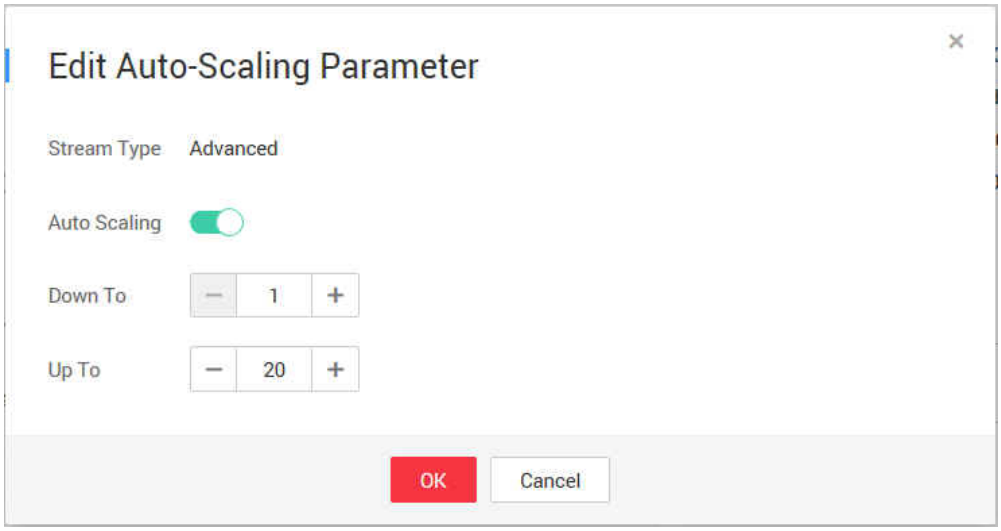
- Method 1
  a. On the **Stream Management** page, click the name of the stream to be scaled.
  b. In the upper right corner of the displayed page, click **Scale Up/Down**. The **Scale Up/Down Stream** dialog box is displayed.
  c. Change the number of target partitions and click **Yes**.
- Method 2
  a. In the **Operation** column of the stream to be scaled, click **More** and choose **Scale Up/Down** from the drop-down list.
  b. The **Scale Up/Down Partition** dialog box is displayed.
  c. Change the number of target partitions and click **Yes**.

Automatically Scaling Up/Down a Stream

1. On the **Stream Management** page, click the name of the stream to be scaled.
2. Click **Edit** next to **Auto-Scaling**.
3. In the displayed **Edit Auto-Scaling Parameter** dialog box, enable **Auto-Scaling**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-145 Edit Auto-Scaling Parameter



4. Set the upper and lower thresholds for automatic scaling and click **OK**.

6.2.5 Uploading Data by Using Agent

6.2.5.1 DIS Agent Overview

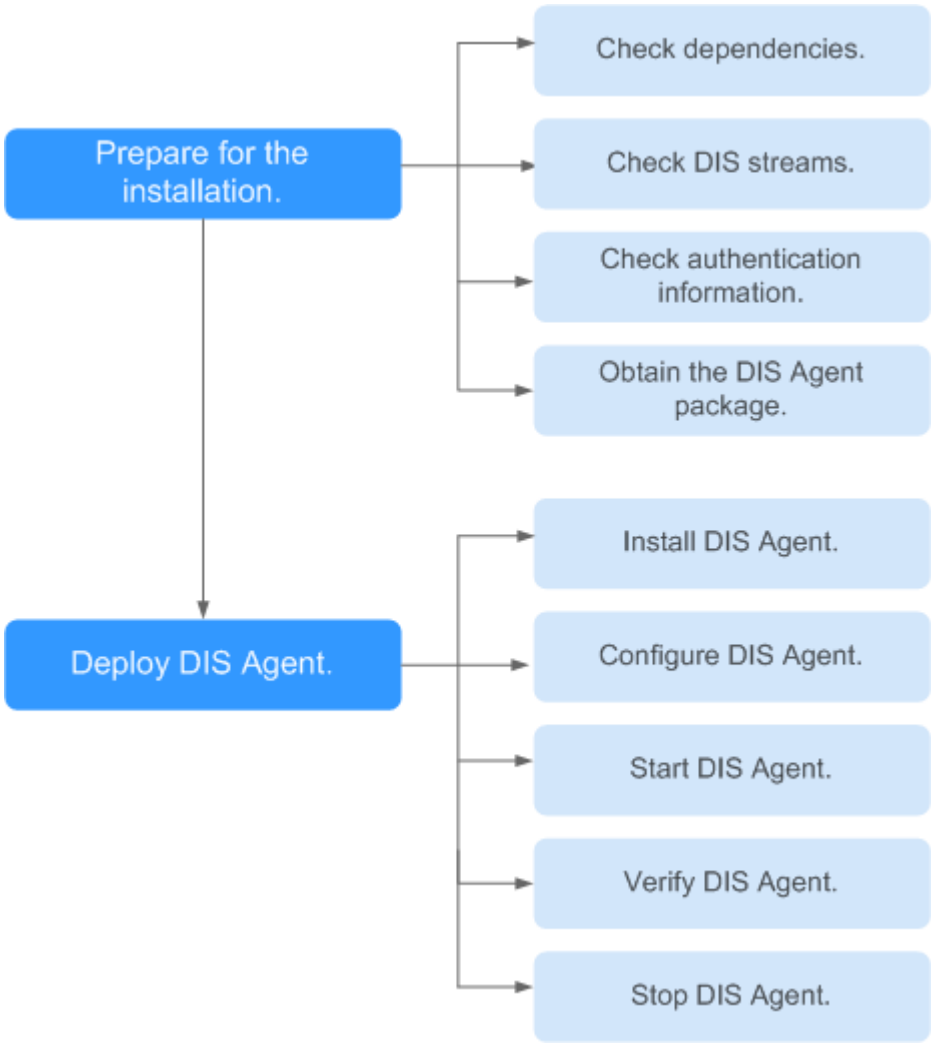
DIS Agent is a client-side program provided by the Data Ingestion Service (DIS) to fulfill the following tasks:

Monitor text files continuously, collect incremental data in real time, parse the data by delimiter and upload it to DIS streams. The source data type of a stream can be BLOB, JSON, and CSV.

Figure 6-146 depicts the process for installing the DIS agent.

Copyright © Huawei Technologies Co., Ltd.

Figure 6-146 Installation flowchart



6.2.5.2 Preparing for Installing DIS Agent

Checking Dependencies

Step 1 Check the server type.

- Linux x86-64 (64-bit) server, for example, EulerOS, Ubuntu, Debian, CentOS, or OpenSUSE
- Windows 7 or a later version

Step 2 Ensure that Java 1.8.0 or later has been installed.

To download JRE, go to **https://www.java.com/en/download/manual.jsp**.

If Java 1.8.0 or later is not installed, perform the following steps to install it on a Linux server:

1. As the **root** user, run the following command to navigate to the **/opt** directory:

   cd /opt

 Copyright © Huawei Technologies Co., Ltd.

2. Run the following command to create the JDK installation directory **jre**:

   mkdir -p jre

3. Run the following command to assign permission 755 to the JDK installation directory:

   chmod -R 777 jre/

4. Download the Java runtime environment (JRE) installation package from the following website:

   tar -zxvf 'JRE installation package name'.tar.gz

5. Modify the **/etc/profile** file.

   a. Run the following command to open the **/etc/profile** file: **vim /etc/profile**

   b. Add the following JDK installation directory information to the configuration option **JAVA_HOME** in the **/etc/profile** file:

      export JAVA_HOME=path to the **jre** folder

      export PATH=$PATH:$JAVA_HOME/bin

      export CLASSPATH=.:$JAVA_HOME/lib/rt.jar:$JAVA_HOME/lib/ext

   c. Run the following command to save the modification and exit:

      :wq

6. Run the following command to validate the JDK configuration:

   source /etc/profile

----End

Checking DIS Streams

Step 1 **Log in to the DGC console**, select an instance, and click **Access** on it.

Click **Workspace Management** to create a workspace.

The workspace you created will show up in the list of workspaces.

Step 2 Click **Data Integration** of the corresponding workspace.

The **Data Integration** page is displayed.

Step 3 In the navigation pane, choose **Stream Management**.

Step 4 In the navigation tree, choose **Stream Management**.

Step 5 Ensure that at least one DIS stream is in the Running state and ready to receive incoming data.

----End

Checking Authentication Information

- AK/SK file

  Access Key ID/Secret Access Key (AK/SK) files are created by the Identity and Access Management (IAM) service to authenticate calls to application programming interfaces (APIs). For details, see the **Obtaining Authentication Information.**

 Copyright © Huawei Technologies Co., Ltd.

- Project ID

  A project is a group of tenant resources. Each region has a unique project ID. For details, see the **Obtaining Authentication Information.**

Obtaining the DIS Agent Package

Obtain the **dis-agent-***X.X.X***.zip** package from **https://dis-publish.obs-website.cn-north-1.myhuaweicloud.com/**.

6.2.5.3 Installing DIS Agent

Prerequisites

PuTTY has been installed.

Installing DIS Agent on a Linux Server

Step 1 Start PuTTY and log in to the server on which logs reside, that is, the server that experiences a dependency check in section **Checking Dependencies**.

Step 2 Upload the DIS Agent package **dis-agent-***X.X.X***.zip** obtained in section **Obtaining the DIS Agent Package** to the **/opt** directory.

Step 3 Run the following command to decompress the **dis-agent-*X.X.X*** directory from the DIS Agent package **dis-agent-*X.X.X*.zip**:

unzip dis-agent-X.X.X.zip

Step 4 Run the following command to navigate to the **dis-agent-*X.X.X*** directory:

cd dis-agent-X.X.X

----End

Installing DIS Agent on a Windows Server

Step 1 Save the **dis-agent-***X.X.X***.zip** package obtained in **Obtaining the DIS Agent Package** to the local host.

Step 2 Decompress the **dis-agent-***X.X.X***.zip** package to the current directory.

----End

6.2.5.4 Configuring a DIS Agent

The DIS Agent configuration file is in the YAML format. Configuration parameters and values must be separated by colon (:) and space. **Table 6-90** describes the configuration parameters in the file.

NOTE

After the configuration is complete, delete unnecessary examples from flows in **agent.yml** or use # to comment out them. For example, if only one DISStream is configured, delete or comment out the following CustomFileStream and other DISStream modules.

 Copyright © Huawei Technologies Co., Ltd.

Configuring DIS Agent on a Linux Server

Step 1 Start PuTTY and log in to the Linux server on which the DIS Agent is installed.

Step 2 Run the **cd /opt/dis-agent-X.X.X/** command to open the **dis-agent-X.X.X** directory.

Step 3 Run the **vim conf/agent.yml** command to open the DIS Agent configuration file **agent.yml**. Modify parameter values in the file to meet specific requirements. **Table 6-90** describes the configuration parameters in the file.

Table 6-90 Parameters in the **agent.yml** file

Parameter	Mandatory	Description	Default Value
region	Yes	Region where DIS resides. **NOTE** For details about how to obtain the region where DIS is deployed, see **Regions and Endpoints**.	cn-north-1
ak	Yes	User's AK. For details about how to obtain the user's AK, see **Checking Authentication Information**.	Set it based on the actual environment requirements.
sk	Yes	User's SK. For details about how to obtain the user's SK, see **Checking Authentication Information**.	Set it based on the actual environment requirements.
projectId	Yes	Project ID specific to your region. For details about how to obtain the project ID, see **Checking Authentication Information**.	Set it based on the actual environment requirements.
endpoint	Yes	DIS gateway address. Format: https://DIS endpoint **NOTE** For details about how to obtain the DIS endpoint, see **Regions and Endpoints**.	https://dis.cn-north-1.myhuaweicloud.com

Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description	Default Value
body.serialize.type	No	Format of the DIS data package to be uploaded (non-original data format). • **json**: The DIS data packet is encapsulated in the format of JSON. • **protobuf**: The DIS data packet is encapsulated in the binary format. After being encapsulated, the volume of the data packet is reduced by 1/3. This format is recommended when a massive amount of data is generated.	json
body.compress.enabled	No	Specifies whether to enable data compression.	false
body.compress.type	No	Data compression format selected when compression is enabled. Currently, the following compression formats are supported: lz4: a compression algorithm with a fast compression speed and high compression efficiency zstd: a new lossless compression algorithm with a fast compression speed and high compression ratio	lz4
PROXY_HOST	No	Proxy IP address. This parameter is mandatory when requests are sent through the proxy server.	-
PROXY_PORT	No	Proxy port.	80
PROXY_PROTOCOL	No	Proxy protocol. **http** and **https** are supported.	http
PROXY_USERNAME	No	Proxy username.	-

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description	Default Value
PROXY_PASSWORD	No	Proxy password.	-
[flows] The [flows] section presents information about the files that will be uploaded to DIS. The following upload mode is supported: **DISStream**: monitors text files continuously, collects incremental data in real time, parses the data by delimiter, and uploads it to DIS streams (source data type: BLOB, JSON, and CSV). **Table 6-91** describes configuration parameters. The **agent.yml** file provides example parameter settings.			

Table 6-91 DISStream configuration parameters

Parameter	Mandatory	Description	Default Value
DISStream	Yes	Name of the DIS stream. Parses the file content matching **filePattern** by delimiter and uploads the file to the stream.	-
filePattern	Yes	File monitoring path. Files in only one directory can be monitored. Directories cannot be monitored recursively. To monitor multiple directories, configure multiple DIS streams in **flows**. The file names can be matched by asterisk (*) • **/tmp/*.log**: Matches all files whose names end with **.log** in the **/tmp** directory. • **/tmp/access-*.log**: Matches all files whose names start with **access-** and end with **.log** in the **/tmp** directory. • In Windows, the example path is **D:\logs*.log**.	-

Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description	Default Value
directoryRecursionEnabled	No	Specify whether to search for a subdirectory. Possible values: • **false**: Not to search for subdirectories recursively and match only files in the **root** directory. • **true**: Search for all subdirectories recursively. For example, if **filePattern** is set to **/tmp/*.log**, **/tmp/one.log**, **/tmp/child/two.log**, and **/tmp/child/child/three.log** can be matched.	false
initialPosition	No	Initial position from which the file started to be monitored. Possible values: • **END_OF_FILE**: After monitoring starts, the system does not parse the files that match **filePattern**. Instead, the newly added file or file content will be parsed by delimiter and uploaded to DIS. • **START_OF_FILE**: All the files that match **filePattern** will be parsed by delimiter and uploaded to DIS based on the file modification time (from the earliest modified to the latest modified).	START_OF_FILE
maxBufferAgeMillis	No	The maximum number of milliseconds that must elapse before data can be uploaded to the DIS. Unit: ms • If the buffer is full with data waiting to be uploaded, data will be immediately uploaded to the DIS. • If the record queue is not full, files will be uploaded to DIS only after the specified period of time is reached.	5000
maxBufferSizeRecords	No	The maximum number of records for which the agent buffers data before sending it to DIS. If the number of records in a queue reaches the value, the data will be uploaded to DIS immediately.	500

Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description	Default Value
partitionKeyOption	No	Method for generating the partition key. Each record carries a partition key. Records with the same partition key are allocated to the same partition. Possible values: • **RANDOM_INT**: The partition key is a random numeric string. Records with such a key are evenly distributed to each partition. • **FILE_NAME**: The partition key is a file name string. Records with such a key is distributed to a specific partition. • **FILE_NAME,RANDOM_INT**: The partition key is a combination of a file name string and a random numeric string, which are separated by comma (,). Records with such a key carries file names and are evenly distributed to all partitions.	RANDOM_INT
recordDelimiter	No	Delimiter used to separate records. Value range: any character that is enclosed in double quotation marks. The value cannot be empty. That is, this parameter cannot be set to "". **NOTE** If the value is a special character, use a backslash (\) to escape. For example, if the value is a quotation mark ("), set this parameter to **\"**. If the value is a backslash (\), set this parameter to ****. If the value is a control character, for example, STX, set this parameter to **\u0002**.	"\n"
isRemainRecordDelimiter	No	Specify whether a delimiter is contained in records to be uploaded. Possible values: • **true**: The delimiter is contained in records to be uploaded. • **false**: The delimiter is not contained in records to be uploaded.	false

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description	Default Value
isFileAppendable	No	Specify whether the file contains additional content. Possible values: • **true**: The file may contain additional content. Agent continuously monitors files. If content is added to a file, Agent parses the file by **recordDelimiter** and uploads records. In this case, ensure that the file ends with **recordDelimiter**. Otherwise, Agent considers that the content has not been added to the file and waits for **recordDelimiter** to be written. • **false**: The file will not contain additional content. If the last row of the file does not end with **recordDelimiter**, Agent still uploads the file as the last record. After the upload is complete, Agent will delete or rename the file based on the configuration of **deletePolicy** and **fileSuffix**.	true
maxFileCheckingMillis	No	Maximum time for checking file changes. If the file size, modification time, and file ID do not change within this period of time, a complete file is generated and starts to be uploaded. Set this parameter based on the actual file change frequency to prevent an incomplete file from being uploaded. If the file is changed after being uploaded, it will be fully uploaded again. Unit: ms **NOTE** This parameter is available only when **isFileAppendable** is set to **false**.	5000
deletePolicy	No	Policy for deleting a file after the file content is uploaded. Possible values: • **never**: The file will not be deleted after the file content is uploaded. • **immediate**: The file will be deleted after the file content is uploaded. **NOTE** This parameter is available only when **isFileAppendable** is set to **false**.	never

Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description	Default Value
fileSuffix	No	Suffix of the file name that is added after the file content is uploaded. If the original file name is **x.txt** and **fileSuffix** is set to **.COMPLETED**, the name of the uploaded file is **x.txt.COMPLETED**. **NOTE** This parameter is available only when **isFileAppendable** is set to **false** and **deletePolicy** is set to **never**.	.COMPLETED
sendingThreadSize	No	The number of sender threads. By default, there is only one sender thread. **NOTICE** If multiple threads are used, the following problems may occur: • Data may not be sent in order. • Some data is lost after the program stops abnormally and restarts.	1
fileEncoding	No	File encoding format. Possible values: **UTF8**, **GBK**, **GB2312**, and **ISO-8859-1**.	UTF8
resultLogLevel	No	Level of the calling result log generated each time when the DIS data sending API is called. • OFF: Each API calling result is not logged. • INFO: Each API calling result is logged at the INFO level. • WARN: Each API calling result is logged at the WARN level. • ERROR: Each API calling result is logged at the ERROR level.	INFO

----End

Configuring DIS Agent on a Windows Server

Step 1 Use a file manager to open the directory (for example, **C:\dis-agent-X.X.X**) where the installation package is decompressed.

Step 2 Open the **agent.yml** file using an editor and modify parameter values in the file to meet specific requirements. **Table 6-90** describes the configuration parameters in the file.

NOTE

The **agent.yml** file is in the Linux format. You are advised to use the general-purpose text editor to edit the file.

----End

Copyright © Huawei Technologies Co., Ltd.

6.2.5.5 Starting DIS Agent

Starting DIS Agent on a Linux Server

Step 1 Start PuTTY and log in to the server on which logs reside.

Step 2 Run the following command to navigate to the DIS Agent installation directory:

cd /opt/dis-agent-x.x.x/*x.x.x* indicates the version number.

Step 3 Run the following script to start the DIS Agent:

bash bin/start-dis-agent.sh

If multiple DIS Agent processes need to be started, you need to use **-c** to specify the configuration file and **-n** to specify the name of the new Agent process.

bash bin/start-dis-agent.sh -c config/anotherAgent.yml -n anotherAgent

NOTICE

Run the bash script to start the DIS Agent. Running another script, such as sh or ./ may fail to start the DIS Agent.

If information similar to the following appears, the DIS Agent has started successfully:

```
Success to start DIS Agent [xxxxx].
```

If no Java variables can be found after the DIS Agent starts, run the following command to restart the DIS Agent:

source /etc/profile

----End

Starting DIS Agent on a Windows Server

Step 1 Navigate to the bin directory of DIS Agent, for example, **C:\dis-agent-X.X.X\bin**.

Step 2 Double-click **start-dis-agent.bat** to start the DIS Agent.

If information similar to the following appears in log files, the DIS Agent has started successfully:

```
[INFO ] (main) com.bigdata.dis.agent.Agent Agent: Startup completed in XXX ms.
```

----End

6.2.5.6 Testing DIS Agent

Testing DIS Agent on a Linux Server

Step 1 Start PuTTY and log in to the server on which logs reside.

Step 2 Run the following command to navigate to the directory in which the DIS Agent stores logs:

 Copyright © Huawei Technologies Co., Ltd.

cd /opt/dis-agent-X.X.X/logs

Step 3 Run the following command to view logs:

tail -100f dis-agent.log

- If information similar to the following appears, the DIS Agent is running properly.
  Agent: Startup completed in xx ms
- If any of the following information appears, the DIS Agent is not running properly and you need to troubleshoot the DIS Agent:
  - **HttpClientErrorException: 400 Bad Request**

    Possible cause: The value of **DISStream** or **projectId** in **Configuring a DIS Agent** is incorrect.

    Solution: Stop the DIS Agent process and correct the parameter values.
  - **HttpClientErrorException: 403 Forbidden**

    Possible cause: The DIS gateway blacklists the IP address of the Linux server on which the DIS Agent is installed and consequently the request to view logs is blocked. Usually, the Linux server IP address is blacklisted because a DIS API is repeatedly called using incorrect configuration.

    Solution: Stop the DIS Agent process, modify the **agent.yml** file (see **Configuring a DIS Agent**). Restart the DIS Agent process 30 minutes after it stops.
  - **UnknownHttpStatusCodeException: Unknown status code [441]**

    Possible cause: AK/SK configuration is incorrect.

    Solution: Stop the DIS Agent process and correct the AK/SK configuration.
  - **ConnectTimeoutException: Connect to DOMAIN[DOMAIN/IP] failed: connect timed out**

    Possible cause: The connection between the Linux server on which the DIS Agent is installed and the DIS gateway times out.

    Solution: Ensure that the Linux server on which the DIS Agent is installed can connect to the Internet.

Step 4 Check whether the DIS Agent can upload logs.

- If the monitoring directory configured in **agent.yml** contains matched files, logs will record the information similar to the following, indicating that [N1 records (B1 bytes)/N2 files (B2 bytes)] are parsed and [N3 records/N4 files] are successfully uploaded.
  Agent: Progress: [**N1** records (**B1** bytes) / **N2** files (**B2** bytes)] parsed, and [**N3** records / **N4** files] sent successfully to destinations. Uptime: 30146ms
- If no files match the **filePattern**, run the following command to generate log files required for testing log uploading:

  echo "`date` Hello world." >> /tmp/test.log

Step 5 Log in to the DIS console and view monitoring data of the DIS stream specified by the **DISStream** or **CustomFileStream** parameter in **Configuring a DIS Agent**. If the stream monitoring data shows that data has been uploaded to the DIS stream, the DIS Agent has been successfully installed.

----End

 Copyright © Huawei Technologies Co., Ltd.

Testing DIS Agent on a Windows Server

Step 1 Use a file manager to navigate to the **logs** directory.

Step 2 Open the **dis-agent.log** file using an editor. View logs in the file.

- If information similar to the following appears, the DIS Agent is running properly.
  ```
  Agent: Startup completed in xx ms
  ```
- If any of the following information appears, the DIS Agent is not running properly and you need to troubleshoot the DIS Agent:
  - **HttpClientErrorException: 400 Bad Request**

    Possible cause: The value of **DISStream** or **projectId** in **Configuring a DIS Agent** is incorrect.

    Solution: Stop the DIS Agent process and correct the parameter values.
  - **HttpClientErrorException: 403 Forbidden**

    Possible cause: The DIS gateway blacklists the IP address of the Linux server on which the DIS Agent is installed and consequently the request to view logs is blocked. Usually, the Linux server IP address is blacklisted because a DIS API is repeatedly called using incorrect configuration.

    Solution: Stop the DIS Agent process, modify the **agent.yml** file (see **Configuring a DIS Agent**). Restart the DIS Agent process 30 minutes after it stops.
  - **UnknownHttpStatusCodeException: Unknown status code [441]**

    Possible cause: AK/SK configuration is incorrect.

    Solution: Stop the DIS Agent process and correct the AK/SK configuration.
  - **ConnectTimeoutException: Connect to DOMAIN[DOMAIN/IP] failed: connect timed out**

    Possible cause: The connection between the Linux server on which the DIS Agent is installed and the DIS gateway times out.

    Solution: Ensure that the Linux server on which the DIS Agent is installed can connect to the Internet.

Step 3 Check whether the DIS Agent can upload logs.

- If the monitoring directory configured in **agent.yml** contains matched files, logs will record the information similar to the following, indicating that [N1 records (B1 bytes)/N2 files (B2 bytes)] are parsed and [N3 records/N4 files] are successfully uploaded.
  ```
  Agent: Progress: [N1 records (B1 bytes) / N2 files (B2 bytes)] parsed, and [N3 records / N4 files] sent successfully to destinations. Uptime: 30146ms
  ```
- If no files match the **filePattern**, run the following command to generate log files required for testing log uploading:

  echo %date%time%Hello world. >> C:\test.log

Step 4 Log in to the DIS console and view monitoring data of the DIS stream specified by the **DISStream** or **CustomFileStream** parameter in **Configuring a DIS Agent**. If the stream monitoring data shows that data has been uploaded to the DIS stream, the DIS Agent has been successfully installed.

----End

 Copyright © Huawei Technologies Co., Ltd.

6.2.5.7 Stopping DIS Agent

Stopping DIS Agent on a Linux Server

Step 1 Start PuTTY and log in to the server on which logs reside.

Step 2 Run the following command to navigate to the DIS Agent installation directory:

cd /opt/dis-agent-X.X.X/

Step 3 Run the following script to stop the DIS Agent:

bash bin/stop-dis-agent.sh

NOTICE

Run the bash script to stop the DIS Agent. Running another script, such as sh or ./ may fail to stop the DIS Agent.

If information similar to the following appears, the DIS Agent is stopping. *xxxxx* indicates the process ID.

```
Stopping Agent [xxxxx].....
```

If information similar to the following appears, the DIS Agent has stopped:

```
Stopping Agent [xxxxx]..... Successfully.
```

To forcibly stop the DIS Agent process, perform the following steps:

1. Run the command to obtain the DIS Agent process ID (PID):
   ps -ef | grep dis-agent | grep -v grep
   The second field in the command output indicates the PID.
2. Run the following command to forcibly stop the DIS Agent process:
   kill -9 PID

----End

Stopping DIS Agent on a Windows Server

Step 1 Press **Ctrl+C** on the CMD terminal. The following information appears:

```
[INFO ] (Agent STOPPING) com.bigdata.dis.agent.Agent Agent: Shutting down...
```

Step 2 When the following information appears, it indicates that the DIS Agent has stopped. Enter **Y** and then press **Enter** to exit.

```
Terminate batch job (Y/N)?
```

----End

6.2.6 Using DIS Flume Plugin to Upload and Download Data

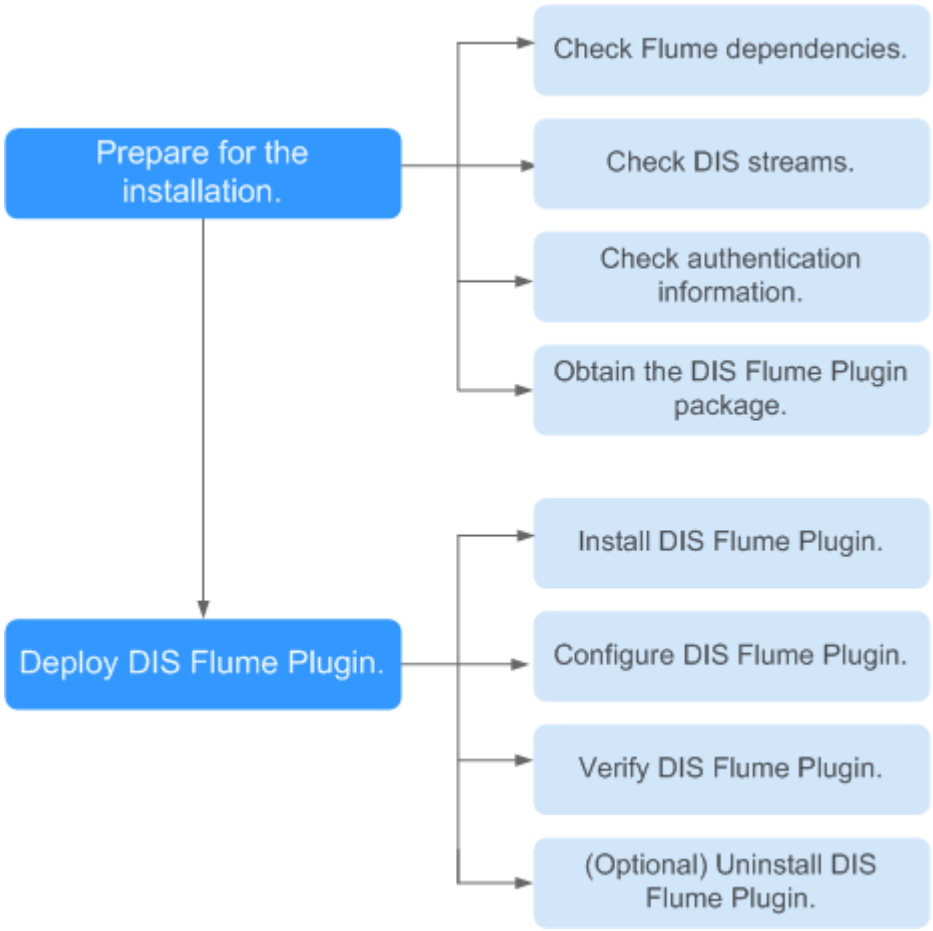
6.2.6.1 Overview

A DIS Flume Plugin is a Flume plugin provided by DIS and consists of DIS Source and DIS Sink. DIS Source is used to download data from DIS to Flume Channel;

 Copyright © Huawei Technologies Co., Ltd.

DIS Sink is used to upload data from Flume Channel to DIS. **Figure 6-147** depicts the process for installing a DIS Flume Plugin.

Figure 6-147 Process for installing a DIS Flume Plugin



6.2.6.2 Preparing for Installing a DIS Flume Plugin

Checking Dependencies

Step 1 Check that Flume is running properly.

Step 2 In the Flume installation directory, run the following command to check that the Flume version is 1.4.0 or later:

```
$ bin/flume-ng version | grep Flume
```

Step 3 Run the following command to check that the Java runtime environment (JRE) version is 1.8.0 or later:

```
java -version
```

----End

Checking DIS Streams

Ensure that at least one DIS stream is in the Running state and ready to receive incoming data.

 Copyright © Huawei Technologies Co., Ltd.

Checking Authentication Information

- AK/SK file

  Access Key ID/Secret Access Key (AK/SK) files are created by the Identity and Access Management (IAM) service to authenticate calls to application programming interfaces (APIs). To obtain AK/SK, choose .

- Project ID

  A project is a group of tenant resources. Each region has a unique project ID. To view the project IDs of different regions, choose **My Credentials** > **Projects**.

Obtaining a DIS Flume Plugin Package

Obtain the **dis-flume-Plugin-***X.X.X***.zip** package from **https://dis-publish.obs-website.cn-north-1.myhuaweicloud.com/**.

6.2.6.3 Installing a DIS Flume Plugin

Prerequisites

PuTTY has been installed.

Procedure

Step 1 Log in to the remote Flume server using PuTTY or another remote login tool.

Step 2 Run the following command to navigate to the Flume installation directory:

cd ${FLUME_HOME}

Step 3 Upload the **dis-flume-plugin-***X.X.X***.zip** installation package to this directory.

Step 4 Decompress the installation package.

unzip dis-flume-plugin-X.X.X.zip

Step 5 Run the following command to navigate to the directory where the package is decompressed.

cd dis-flume-plugin

Step 6 Run the following command to execute the installation program.

bash install.sh

If the following information appears, the ***DIS Flume Plugin*** has been successfully installed in the **${FLUME_HOME}/plugin.d/dis-flume-plugin** directory:

```
Install dis-flume-plugin successfully.
```

----End

6.2.6.4 Configuring the DIS Flume Plugin

A DIS Flume Plugin consists of DIS Source and DIS Sink. The **dis-flume-plugin.conf.template** file in the installation package lists the configuration

 Copyright © Huawei Technologies Co., Ltd.

methods. This section describes the configuration items of DIS Source and DIS Sink.

NOTE

Dis-flume-plugin.conf.template is a configuration sample for DIS plug-in and is not a configuration file that will be accessed when Flume is run. Flume provides a configuration sample file in *FLUME_HOME*/**conf/flume-conf.properties.template**, where *FLUME_HOME* is the installation path of Flume. You can modify the configuration file based on site requirements.

Configuring DIS Source

Table 6-92 DIS Source configuration parameters

Parameter	Mandatory	Description	Default Value
channels	Yes	Name of the Flume channel.	-
type	Yes	DIS Source type.	com.huaweicloud.dis.adapter.flume.source.DISSource
streams	Yes	DIS stream name.	The entered DIS stream name must be the same as the stream name specified when you are creating a DIS stream on the DIS console.
ak	Yes	User's AK. For details about how to obtain an AK, see **Checking Authentication Information**.	-
sk	Yes	User's SK. For details about how to obtain an SK, see **Checking Authentication Information**.	-
region	Yes	Region in which the DIS is located.	-

Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description	Default Value
projectId	Yes	Project ID specific to your region. For details about how to obtain a project ID, see **Checking Authentication Information**.	-
endpoint	Yes	Data API address of the region where DIS resides.	-
group.id	Yes	Application name, which is used to identify a consumer group and consists of letters, digits, hyphens (-), and underscores (_).	-

Configuring DIS Sink

Table 6-93 DIS Sink configuration parameters

Parameter	Mandatory	Description	Default Value
channel	Yes	Name of the Flume channel.	-
type	Yes	Sink type.	com.huaweicloud.dis.adapter.flume.sink.DISSink
streamName	Yes	Name of the DIS stream.	The entered DIS stream name must be the same as the stream name specified when you are creating a DIS stream on the DIS console.

Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description	Default Value
ak	Yes	User's AK. For details about how to obtain an AK, see **Checking Authentication Information**.	-
sk	Yes	User's SK. For details about how to obtain an SK, see **Checking Authentication Information**.	-
region	Yes	Region in which the DIS is located.	cn-north-1
projectId	Yes	Project ID specific to your region. For details about how to obtain a project ID, see **Checking Authentication Information**.	-
endpoint	Yes	Data API address of the region where DIS resides.	https://dis.cn-north-1.myhuaweicloud.com
partitionNumber	No	The number of partitions that the chosen DIS stream has. The value is used to calculate batchSize.	1
batchSize	No	Number of data records that can be batch processed in a single Flume transaction.	batchSize = partitionNumber * 250
sendingThreadSize	No	The number of sender threads. By default, there is only one sender thread. **NOTE** If multiple sender threads are used, the following situations will occur: • There is no guarantee on the order in which data will be sent. • Certain data will be resent if the Flume application restarts after it stops abruptly.	1

Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description	Default Value
sendingRecordSize	No	Number of data records that can be sent in a single call to the DIS API that is used to put data into DIS streams. **NOTE** **batchSize** indicates the number of data records that can be batch processed in a single Flume transaction, whereas **sendingRecordSize** indicates the number of data records that can be batch processed in a single API call. For example, if **batchSize** is 1000 and **sendingRecordSize** is 250, it indicates that four API calls will be made to complete the Flume transaction. A Flume transaction is completed and submitted only after the **batchSize** amount of data is successfully sent. If the application restarts before a Flume transaction is submitted, data will be resent. If **sendingThreadSize** is set to **1**, it indicates that **sendingRecordSize** and **batchSize** will have the same value. This prevents unnecessary data resending.	250
retrySize	No	The maximum number of times that the DIS Flume Sink can retry to call a DIS API when the initial call to the DIS API fails. The default value 2147483647 is recommended, indicating that Sink can retry the API call for an unlimited number of times. Exponential backoff is used to incrementally increase the wait between retry attempts in order to reduce server load and increase the likelihood that repeated requests will succeed.	2147483647

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description	Default Value
resultLogLevel	No	The level of logs generated to print out the latest sequenceNumber at the end of each DIS API call. Log levels are listed in the order of from low to high: OFF < DEBUG < INFO < WARN < ERROR. The value **OFF** indicates that no logs will be generated. If the log level of Flume log4j is higher than **resultLogLevel**, no logs will be generated.	OFF
maxBufferAgeMillis	No	The maximum number of milliseconds that must elapse before data can be uploaded to the DIS. ● If the buffer is full with data waiting to be uploaded, data will be immediately uploaded to the DIS. ● If the buffer is not full, data will be uploaded to the DIS after the specified number of milliseconds elapses.	5000
connectionTimeOutSeconds	No	The amount of time that must elapse before a DIS API call times out. Unit: second	30
socketTimeOutSeconds	No	The amount of time that must elapse before a response to a DIS API call times out. Unit: second	60

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description	Default Value
dataEncryptEnabled	No	An indicator of whether data is encrypted using the Advanced Encryption Standard (AES) algorithm. • **true** • **false**	false
dataPassword	No	Password used to encrypt or decrypt data. This parameter is mandatory if **dataEncryptEnabled** is set to **true**.	-
bodySerializeType	No	Upload format of the DIS data packet (not original data format). Possible values: • **json**: The DIS data packet is encapsulated in the format of JSON. • **protobuf**: The DIS data packet is encapsulated in the binary format. After being encapsulated, the volume of the data packet is reduced by 1/3. This format is recommended when a massive amount of data is generated.	json

6.2.6.5 Testing a DIS Flume Plugin

Testing DIS Source

Step 1 Start PuTTY and log in to the server on which Flume is installed.

Step 2 Ensure that the configuration file containing information about the DIS Source is ready.

You can modify the configuration file based on the **flume-conf.properties.template** provided by Flume. The following provides a file sample:

```
agent.sources = dissource
agent.channels = memoryChannel
```

 Copyright © Huawei Technologies Co., Ltd.

```
agent.sinks = loggerSink
#Define DIS Source (which is used to obtain data from DIS).
agent.sources.dissource.channels = memoryChannel
agent.sources.dissource.type = com.huaweicloud.dis.adapter.flume.source.DISSource
agent.sources.dissource.streams = YOU_DIS_STREAM_NAME
agent.sources.dissource.ak = YOU_ACCESS_KEY_ID
agent.sources.dissource.sk = YOU_SECRET_KEY_ID
agent.sources.dissource.region = YOU_Region
agent.sources.dissource.projectId = YOU_PROJECT_ID
agent.sources.dissource.endpoint = https://dis.${region}.myhuaweicloud.com
agent.sources.dissource.group.id = YOU_APP_NAME
#Define a stream.
agent.channels.memoryChannel.type = memory
agent.channels.memoryChannel.capacity = 10000
#Define Logger Sink (which is used to output data to the console).
agent.sinks.loggerSink.type = logger
agent.sinks.loggerSink.channel = memoryChannel
```

Step 3 Start Flume. For details about the startup command, see the instructions at the Apache Flume official website.

To start Flume from the Flume installation directory, run the following sample command:

```
bin/flume-ng agent --conf-file conf/flume-conf.properties.template --name agent --conf conf/ -
Dflume.root.logger=INFO,console
```

In the preceding information, **bin/flume-ng agent** indicates that Flume Agent will be started; **conf-file** indicates the path of the configuration file written by the user; **name** indicates the agent name in the configuration file; **conf** indicates the **conf/** path of Flume.

View the log file. If the log file contains information similar to "source disSource started", DIS Source starts normally. In the preceding information, **disSource** indicates the DIS Source name configured by the user.

Step 4 Check that DIS Source can successfully download data from DIS.

Upload data to a stream to which DIS Source points. If Flume does not report an error and DIS Sink can obtain data, the download is successful.

NOTE

If the sample configuration in **Step 2** is used, the data obtained from DIS is output to the console in byte array format.

Step 5 Log in to the DIS console. Two minutes later, check monitoring data of the DIS stream specified in **Table 6-92**. If data download (blue lines) is displayed, DIS Source is running successfully.

----End

Testing DIS Sink

Step 1 Start PuTTY and log in to the server on which Flume is installed.

Step 2 Ensure that the configuration file containing information about DIS Sink is ready.

You can modify the configuration file based on the **flume-conf.properties.template** provided by Flume. The following provides a file sample:

```
agent.sources = exec
agent.channels = memoryChannel
```

 Copyright © Huawei Technologies Co., Ltd.

```
agent.sinks = dissink
#Define EXEC Source (which is used to monitor the dis.txt file in the tmp directory).
agent.sources.exec.type = exec
agent.sources.exec.command = tail -F /tmp/dis.txt
agent.sources.exec.shell = /bin/bash -c
agent.sources.exec.channels = memoryChannel
#Define a stream.
agent.channels.memoryChannel.type = memory
agent.channels.memoryChannel.capacity = 10000
# Define DIS Sink (which is used to output data to the DIS stream).
agent.sinks.dissink.channel = memoryChannel
agent.sinks.dissink.type = com.huaweicloud.dis.adapter.flume.sink.DISSink
agent.sinks.dissink.streamName = YOU_DIS_STREAM_NAME
agent.sinks.dissink.ak = YOU_ACCESS_KEY_ID
agent.sinks.dissink.sk = YOU_SECRET_KEY_ID
agent.sinks.dissink.region = YOU_Region
agent.sinks.dissink.projectId = YOU_PROJECT_ID
agent.sinks.dissink.endpoint = https://dis.${region}.myhuaweicloud.com
agent.sinks.dissink.resultLogLevel = INFO
```

Step 3 Start Flume. For details about the startup command, see the instructions at the Apache Flume official website.

To start Flume from the Flume installation directory, run the following sample command:

```
bin/flume-ng agent --conf-file conf/flume-conf.properties.template --name agent --conf conf/ -
Dflume.root.logger=INFO,console
```

In the preceding information, **bin/flume-ng agent** indicates that Flume Agent will be started; **conf-file** indicates the path of the configuration file written by the user; **name** indicates the agent name in the configuration file; **conf** indicates the **conf/** path of Flume.

View the log file. If the log file contains information similar to "Dis flume sink [dissink] start", it indicates that DIS Sink starts normally. The value of **dissink** is the DIS Sink name configured by the user.

Step 4 Check that the DIS Sink can successfully upload data to DIS.

Ingest data into the Flume source. Set the **resultLogLevel** of the DIS Sink to a value that is not **OFF** and higher than the log level of log4j. If logs similar to the following are displayed, the DIS Flume Sink has successfully uploaded data to the DIS.

```
CurrentPut 5 events[success 5 / failed 0] spend 131 ms.
```

NOTE

If the sample configuration described in **Step 2** is used, you can create a **dis.txt** file in the **tmp** directory and append content to the file. After Flume is started, each row of the appended content is read by Flume and sent to the DIS stream through DIS Sink.

Step 5 Log in to the DIS console. Two minutes later, check monitoring data of the DIS stream specified in **Table 6-93**. If data upload (indicated in green lines) is displayed, DIS Sink is running successfully.

----End

 Copyright © Huawei Technologies Co., Ltd.

6.2.6.6 (Optional) Uninstalling a DIS Flume Plugin

Procedure

Step 1 Start PuTTY and log in to the server on which Flume is installed.

Step 2 Stop the Flume program.

Step 3 Go to the directory where the DIS Flume Plugin is located.

cd ${FLUME_HOME}

cd dis-flume-plugin

Step 4 Run the following commands to uninstall the DIS Flume Plugin:

dos2unix install.sh

bash install.sh uninstall

If information similar to the following appears, the DIS Flume Sink has been successfully uninstalled:

```
Uninstall dis-flume-plugin successfully.
```

----End

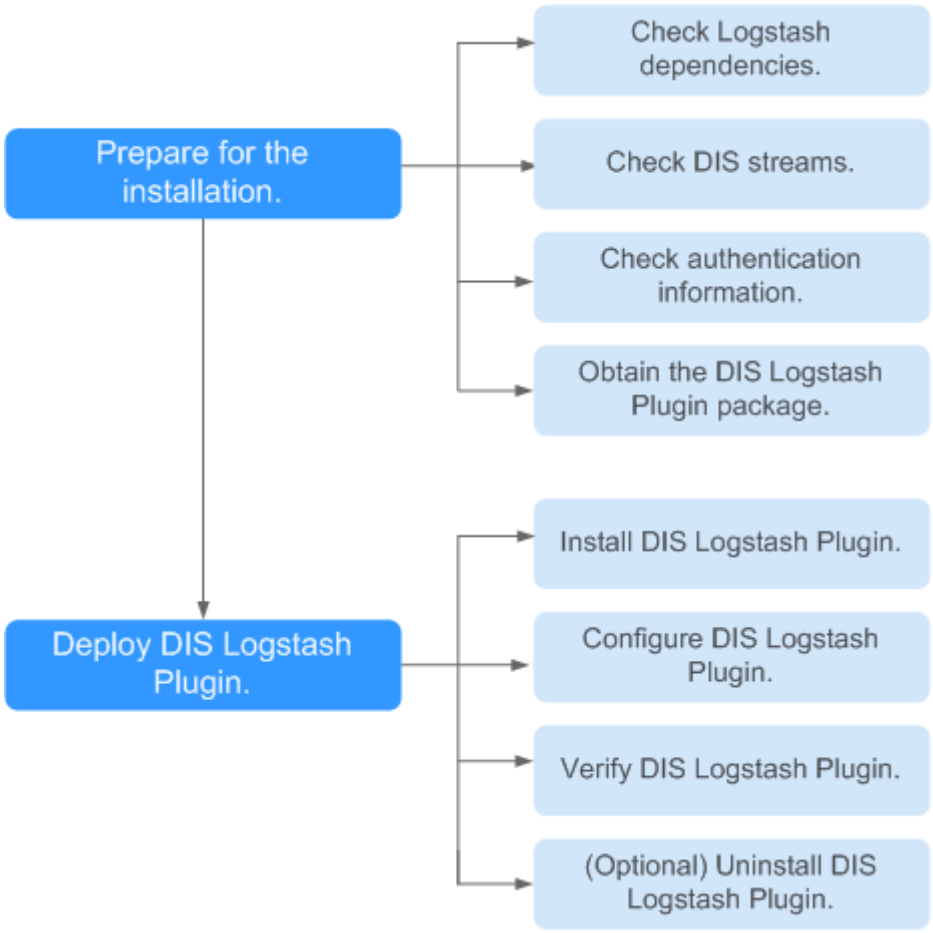
6.2.7 Using a DIS Logstash Plugin to Upload and Download Data

6.2.7.1 Overview

A DIS Logstash Plugin is a Logstash plugin provided by DIS and consists of DIS Input and DIS Output. DIS Input is used to download data from DIS to Logstash; DIS Output is used to upload data from Logstash to DIS. **Figure 6-148** depicts the process for installing a DIS Logstash Plugin.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-148 Installation flowchart



6.2.7.2 Preparing for Installing DIS Logstash Plugin

Checking Dependencies

Step 1 Check that Logstash is running properly.

Step 2 Ensure that the Logstash version ranges from 6.0.0 to 6.1.0. Go to the Logstash installation directory and run the following command to view the Logstash version:

```
$ bin/logstash --version
```

Step 3 Ensure that the Java version is 1.8.0 or later. Run the following command to view the Java version:

```
java -version
```

Step 4 Ensure that the JRuby version is 9.0.0.0 or later. Run the following command to view the JRuby version:

```
$ bin/jruby -v
```

----End

 Copyright © Huawei Technologies Co., Ltd.

Checking DIS Streams

Step 1 Log in to the DGC console, select an instance, and click **Access** on it.

Click **Workspace Management** to create a workspace.

The workspace you created will show up in the list of workspaces.

Step 2 Click **Data Integration** of the corresponding workspace.

The **Data Integration** page is displayed.

Step 3 In the navigation pane, choose **Stream Management**.

Step 4 Ensure that at least one DIS stream is in the Running state and ready to receive incoming data.

----End

Checking Authentication Information

- AK/SK file

  Access Key ID/Secret Access Key (AK/SK) files are created by the Identity and Access Management (IAM) service to authenticate calls to application programming interfaces (APIs). To obtain AK/SK, choose **My Credentials** > **Access Keys**.

- Project ID

  A project is a group of tenant resources. Each region has a unique project ID. To view the project IDs of different regions, choose .

Obtaining a DIS Logstash Plugin Package

Obtain the **dis-logstash-Plugin-***X.X.X***.zip** package from **https://dis-publish.obs-website.cn-north-1.myhuaweicloud.com/**.

6.2.7.3 Installing the DIS Logstash Plugin

The DIS Logstash Plugin can be installed online or offline.

For offline installation, you need to obtain the DIS Logstash Plugin package and execute the installation script.

You do not need to download the DIS Logstash Plugin package for online installation. You can directly connect to the public network to install the plugin.

Prerequisites

PuTTY has been installed.

6.2.7.4 Installing DIS Logstash Plugin Online

The DIS Logstash Plugin can be installed online or offline.

You do not need to download the DIS Logstash Plugin package for online installation. You can directly connect to the public network to install the plugin.

 Copyright © Huawei Technologies Co., Ltd.

Prerequisites

PuTTY has been installed.

Installing logstash-input-dis

Step 1 Start PuTTY or another remote login tool to log in to the server where Logstash is installed.

Step 2 Run the following command to navigate to the Logstash installation directory:

cd ${LOGSTASH_HOME}

Step 3 Run the following command to install **logstash-input-dis**:

bin/logstash-plugin install logstash-input-dis

If information similar to the following is displayed, the installation is successful:

```
Validating logstash-input-dis
Installing logstash-input-dis
Installation successful
```

----End

Installing logstash-output-dis

Step 1 Start PuTTY or another remote login tool to log in to the server where Logstash is installed.

Step 2 Run the following command to navigate to the Logstash installation directory:

cd ${LOGSTASH_HOME}

Step 3 Run the following command to install **logstash-output-dis**:

bin/logstash-plugin install logstash-output-dis

If information similar to the following is displayed, the installation is successful:

```
Validating logstash-output-dis
Installing logstash-output-dis
Installation successful
```

----End

6.2.7.5 Installing DIS Logstash Plugin Offline

The DIS Logstash Plugin can be installed online or offline.

For offline installation, you need to obtain the DIS Logstash Plugin package and execute the installation script.

Prerequisites

PuTTY has been installed.

 Copyright © Huawei Technologies Co., Ltd.

Procedure

Step 1 Start PuTTY or another remote login tool to log in to the server where Logstash is installed.

Step 2 Run the following command to navigate to the Logstash installation directory:

cd ${LOGSTASH_HOME}

Step 3 Upload the **dis-logstash-plugins-*X.X.X*.zip** installation package to this directory.

Step 4 Decompress the installation package.

unzip dis-logstash-plugins-X.X.X.zip

Step 5 Run the following command to navigate to the directory where the package is decompressed:

cd logstash-plugins

Step 6 Run the following command to run the installation program:

bash install.sh –p ${LOGSTASH_HOME}

If information similar to the following is displayed, the installation is successful:

```
Install dis-logstash-plugins successfully.
```

----End

6.2.7.6 Configuring the DIS Logstash Plugin

DIS Logstash Plugins consist of the Input and Output plugins. This section describes the configuration items of the plugins.

Configuring DIS Logstash Input

The configuration template (used to download data from a DIS stream to a local file) is as follows:

```
input
{
  dis {
      streams => ["YOU_DIS_STREAM_NAME"]
      endpoint => "https://${endpoint}"
      ak => "YOU_ACCESS_KEY_ID"
      sk => "YOU_SECRET_KEY_ID"
      region => "${region}"
      project_id => "YOU_PROJECT_ID"
      group_id => "YOU_APP_ID"
      auto_offset_reset => "earliest"
   }
}
output
{
    file {
      path => ["/tmp/test.log"]
  }
}
```

 Copyright © Huawei Technologies Co., Ltd.

Table 6-94 DIS Logstash Input configuration parameters

Parameter	Mandatory	Description	Default Value
stream	Yes	DIS stream name.	The value of this parameter must be the same as that of **Stream Name** configured on the DIS console.
ak	Yes	User's AK. For details about how to obtain an AK, see **Checking Authentication Information**.	-
sk	Yes	User's SK. For details about how to obtain an SK, see **Checking Authentication Information**.	-
region	Yes	Region in which DIS is located.	cn-north-1
project_id	Yes	Project ID specific to your region. For details about how to obtain a project ID, see **Checking Authentication Information**.	-
endpoint	Yes	Data API address of the region where DIS resides.	https://dis.cn-north-1.myhuaweicloud.com
group_id	Yes	DIS App name, used to identify a consumer group. The value can be any character string.	-

Configuring DIS Logstash Output

The configuration template (used to read data from a local file and upload it to a DIS stream) is as follows:

 Copyright © Huawei Technologies Co., Ltd.

```
input
{
  file {
     path => ["/tmp/test.log"]
     type => "log4j"
     start_position => "beginning"
  }
}
output
{
   dis {
      stream => ["YOU_DIS_STREAM_NAME"]
      endpoint => "https://dis.${endpoint}"
      ak => "YOU_ACCESS_KEY_ID"
      sk => "YOU_SECRET_KEY_ID"
      region => "${region}"
      project_id => "YOU_PROJECT_ID"
   }
}
```

Table 6-95 DIS Logstash Output configuration parameters

Parameter	Mandatory	Description	Default Value
stream	Yes	DIS stream name.	The value of this parameter must be the same as that of **Stream Name** configured on the DIS console.
ak	Yes	User's AK. For details about how to obtain an AK, see **Checking Authentication Information**.	-
sk	Yes	User's SK. For details about how to obtain an SK, see **Checking Authentication Information**.	-
region	Yes	Region in which DIS is located.	cn-north-1
project_id	Yes	Project ID specific to your region. For details about how to obtain a project ID, see **Checking Authentication Information**.	-

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description	Default Value
body_compress_enabled	No	Specifies whether to enable data compression.	No
body_compress_type	No	Data compression type. Currently, the following compression algorithms are supported: lz4: a compression algorithm with a fast compression speed and high compression efficiency zstd: a new lossless compression algorithm with a fast compression speed and high compression ratio	lz4

6.2.7.7 Testing the DIS Logstash Plugin

Testing DIS Logstash Input

Step 1 Start PuTTY and log in to the server on which Logstash is installed.

Step 2 Start the Logstash program.

```
bin/logstash -f dis_to_local.conf
```

In the preceding command, **-f** indicates the path of the configuration file.

Step 3 Check that DIS Input can successfully download data from DIS.

Upload data to a stream to which DIS Input points. If Logstash does not report an error and DIS Output can obtain data, the download is successful.

Step 4 Log in to the DIS console. Two minutes later, check monitoring data of the DIS stream specified in **Table 6-94**. If data download (blue lines) is displayed, DIS Logstash Input is running successfully.

----End

Testing DIS Logstash Output

Step 1 Start PuTTY and log in to the server on which Logstash is installed.

 Copyright © Huawei Technologies Co., Ltd.

Step 2 Start the Logstash program.

```
bin/logstash -f local_to_dis.conf
```

In the preceding command, **-f** indicates the path of the configuration file.

Step 3 Check that DIS Logstash Output can successfully upload data to DIS.

Ingest data to the input end of Logstash. If Logstash does not report an error and can upload data to the specified stream, the upload is successful.

Step 4 Log in to the DIS console. Two minutes later, check monitoring data of the DIS stream specified in **Table 6-95**. If data upload (green lines) is displayed, DIS Logstash Output is running successfully.

----End

6.2.7.8 (Optional) Uninstalling the DIS Logstash Plugin

Procedure

Step 1 Start PuTTY and log in to the server on which Logstash is installed.

Step 2 Stop the Logstash program.

Step 3 Go to the directory where the DIS Logstash Plugin is located.

cd ${LOGSTASH_HOME}

cd logstash-plugins

Step 4 Run the following command to uninstall the DIS Logstash Plugin:

bash uninstall.sh –p ${LOGSTASH_HOME}

If information similar to the following appears, the DIS Logstash Plugin has been successfully uninstalled:

```
Uninstall dis-logstash-plugins successfully.
```

----End

6.2.8 Using Kafka Adapter to Upload and Download Data

6.2.8.1 Kafka Adapter Overview

dis-kafka-adapter is a software development kit (SDK) provided by DIS. Users who originally use Kafka Client can use dis-kafka-adapter instead to upload data to DIS in the similar way.

6.2.8.2 Preparations

Configuring the pom.xml File

If a maven project exists, use the following dependency in **pom.xml**:

```
<dependency>
   <groupId>com.huaweicloud.dis</groupId>
```

 Copyright © Huawei Technologies Co., Ltd.

```
    <artifactId>huaweicloud-dis-kafka-adapter</artifactId>
    <version>1.2.9</version>
</dependency>
```

Using the DIS Sample Project

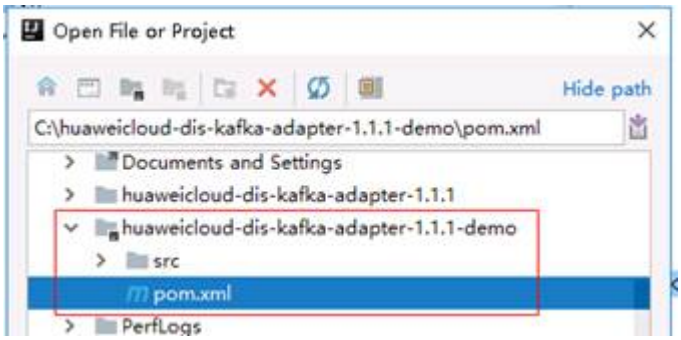
Obtain the **huaweicloud-dis-kafka-adapter-*X.X.X*** package of the latest version from **https://dis-publish.obs-website.cn-north-1.myhuaweicloud.com/**. The package contains two directories.

- The **huaweicloud-dis-kafka-adapter-*X.X.X*** directory contains all JAR packages. If a non-maven project is used, import all JAR packages in the lib directory to the environment.
- **huaweicloud-dis-kafka-adapter--*X.X.X*-demo** is a sample project and is compiled using maven.

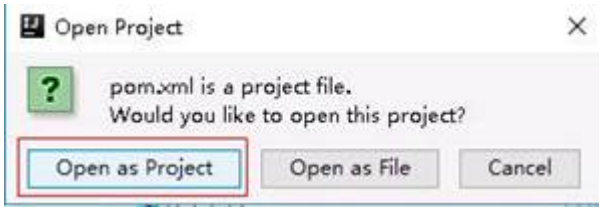
You can use IntelliJ IDEA to import the sample project as follows:

Step 1 Run IntelliJ IDEA and choose **File** > **Open**.

In the displayed dialog box, expand **huaweicloud-dis-kafka-adapter-*X.X.X*-demo** and double-click **pom.xml**.



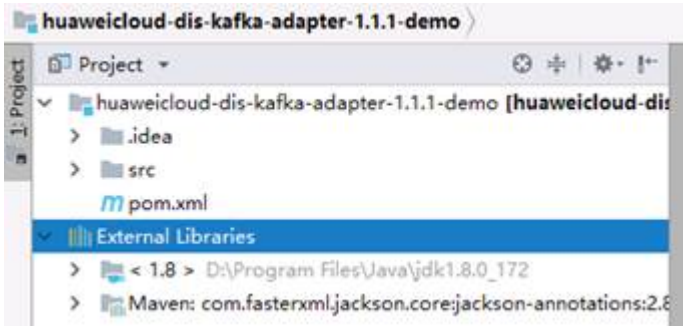
Step 2 When the following dialog box is displayed, select **Open as Project**.



Step 3 Click **New Window** to open the project in a new window.

Step 4 Wait when IntelliJ IDEA is building the project. After the project is built, the directory and files are displayed.

 Copyright © Huawei Technologies Co., Ltd.



----End

Checking Authentication Information

The Access Key ID/Secret Access Key (AK/SK) or AuthToken can be used for authentication. By default, the AK/SK is used for authentication.

- If AK/SK is used for authentication, check the AK/SK.

  AK/SK files are created by the Identity and Access Management (IAM) service to authenticate calls to application programming interfaces (APIs) on the public cloud. To obtain AK/SK, choose **My Credentials > Access Keys**.

- If AuthToken is used for authentication, check **AuthType** and **AuthToken**.

  AuthType: authentication type. The value is set to **authtoken**.

  AuthToken: created by the user in IAM and is valid for 24 hours.

- Project ID

  A project is a group of tenant resources. A tenant can have multiple projects, one for each region. Each region has a unique Project ID. To view the project IDs of different regions, choose .

6.2.8.3 Uploading Data

Sample Code

For details about the data upload code, see **DISKafkaProducerDemo.java** in the sample project. For details about how to obtain the AK, SK, and Product ID, see **Checking Authentication Information**.

```
package com.huaweicloud.dis.demo.adapter;
import com.huaweicloud.dis.DISConfig;
import com.huaweicloud.dis.adapter.kafka.clients.producer.*;
import com.huaweicloud.dis.adapter.kafka.common.serialization.StringSerializer;
import org.slf4j.Logger;
import org.slf4j.LoggerFactory;
import java.util.Properties;
import java.util.concurrent.CountDownLatch;
import java.util.concurrent.Future;
import java.util.concurrent.ThreadLocalRandom;

public class DISKafkaProducerDemo
{
    private static final Logger LOGGER = LoggerFactory.getLogger(DISKafkaProducerDemo.class);

    public static void main(String[] args)
    {
        // YOU AK
        String ak = "YOU_AK";
```

 Copyright © Huawei Technologies Co., Ltd.

```
        // YOU SK
        String sk = "YOU_SK";
        // YOU ProjectId
        String projectId = "YOU_PROJECT_ID";
        // YOU DIS Stream
        String streamName = "YOU_STREAM_NAME";
        // DIS region
        String region = "YOU_Region";

        Properties props = new Properties();
        props.setProperty(DISConfig.PROPERTY_AK, ak);
        props.setProperty(DISConfig.PROPERTY_SK, sk);
        props.setProperty(DISConfig.PROPERTY_PROJECT_ID, projectId);
        props.setProperty(DISConfig.PROPERTY_REGION_ID, region);
        props.setProperty(ProducerConfig.KEY_SERIALIZER_CLASS_CONFIG, StringSerializer.class.getName());
        props.setProperty(ProducerConfig.VALUE_SERIALIZER_CLASS_CONFIG,
StringSerializer.class.getName());

//By default, the domain name is automatically used for access instead of configuring an endpoint. If an
endpoint is required, remove the following comments and set the endpoint:
        // props.setProperty(DISConfig.PROPERTY_ENDPOINT, "https://dis-${region}.myhuaweicloud.com");

        // Create dis producer
        Producer<String, String> producer = new DISKafkaProducer<>(props);

                //Send data synchronously.
        synchronousSendDemo(producer, streamName);

              //Send data asynchronously.
        asynchronousSendDemo(producer, streamName);

// Disable the producer to prevent resource leakage.
        producer.close();
    }

    public static void synchronousSendDemo(Producer<String, String> producer, String streamName)
    {
        LOGGER.info("===== synchronous send =====");
        for (int i = 0; i < 5; i++)
        {
 //If the key is set to Random or Null), data is evenly distributed to all partitions.
            String key = String.valueOf(ThreadLocalRandom.current().nextInt(1000000));
            String value = "Hello world[sync]. " + i;

            Future<RecordMetadata> future = producer.send(new ProducerRecord<>(streamName, key, value));

            try
            {
 //Calling future.get will block waiting until sending is complete.
                RecordMetadata recordMetadata = future.get();
//Data is successfully sent.
                LOGGER.info("Success to send [{}], Partition [{}], Offset [{}].",
                        value, recordMetadata.partition(), recordMetadata.offset());
            }
            catch (Exception e)
            {
//Data failed to be sent.
                LOGGER.error("Failed to send [{}], Error [{}]", value, e.getMessage(), e);
            }
        }
    }

    public static void asynchronousSendDemo(Producer<String, String> producer, String streamName)
    {
        LOGGER.info("===== asynchronous send =====");
        int totalSendCount = 5;
        CountDownLatch countDownLatch = new CountDownLatch(totalSendCount);
        for (int i = 0; i < totalSendCount; i++)
        {
```

 Copyright © Huawei Technologies Co., Ltd.

```
 //If the key is set to Random or Null), data is evenly distributed to all partitions.
        String key = String.valueOf(ThreadLocalRandom.current().nextInt(1000000));
        String value = "Hello world[async]. " + i;

        try
        {
//Data is sent in callback mode and is not blocked.
            producer.send(new ProducerRecord<>(streamName, key, value), new Callback()
            {
                @Override
                public void onCompletion(RecordMetadata recordMetadata, Exception e)
                {
                    countDownLatch.countDown();
                    if (e == null)
                    {
//Data is successfully sent.
                        LOGGER.info("Success to send [{}], Partition [{}], Offset [{}].",
                                value, recordMetadata.partition(), recordMetadata.offset());
                    }
                    else
                    {
//Data failed to be sent.
                        LOGGER.error("Failed to send [{}], Error [{}]", value, e.getMessage(), e);
                    }
                }
            });
        }
        catch (Exception e)
        {
            countDownLatch.countDown();
            LOGGER.error(e.getMessage(), e);
        }
    }

    try
    {
// Wait until all data is sent.
        countDownLatch.await();
    }
    catch (InterruptedException e)
    {
        LOGGER.error(e.getMessage(), e);
    }
  }
}
```

After running the preceding program, if the data is successfully sent, the following log is generated:

```
09:32:52.001 INFO  c.h.d.d.a.DISKafkaProducerDemo - ===== synchronous send =====
09:32:53.523 INFO  c.h.d.d.a.DISKafkaProducerDemo - Success to send [Hello world[sync]. 0], Partition [0],
Offset [114].
09:32:53.706 INFO  c.h.d.d.a.DISKafkaProducerDemo - Success to send [Hello world[sync]. 1], Partition [0],
Offset [115].
09:32:53.956 INFO  c.h.d.d.a.DISKafkaProducerDemo - Success to send [Hello world[sync]. 2], Partition [0],
Offset [116].
09:32:54.160 INFO  c.h.d.d.a.DISKafkaProducerDemo - Success to send [Hello world[sync]. 3], Partition [0],
Offset [117].
09:32:54.450 INFO  c.h.d.d.a.DISKafkaProducerDemo - Success to send [Hello world[sync]. 4], Partition [0],
Offset [118].
09:32:54.450 INFO  c.h.d.d.a.DISKafkaProducerDemo - ===== asynchronous send =====
09:32:54.673 INFO  c.h.d.d.a.DISKafkaProducerDemo - Success to send [Hello world[async]. 0], Partition [0],
Offset [119].
09:32:54.674 INFO  c.h.d.d.a.DISKafkaProducerDemo - Success to send [Hello world[async]. 1], Partition [0],
Offset [120].
09:32:54.674 INFO  c.h.d.d.a.DISKafkaProducerDemo - Success to send [Hello world[async]. 2], Partition [0],
Offset [121].
09:32:54.674 INFO  c.h.d.d.a.DISKafkaProducerDemo - Success to send [Hello world[async]. 3], Partition [0],
Offset [122].
```

 Copyright © Huawei Technologies Co., Ltd.

```
09:32:54.674 INFO  c.h.d.d.a.DISKafkaProducerDemo - Success to send [Hello world[async]. 4], Partition [0], Offset [123].
```

Adaptation to the Native KafkaProducer API

DISKafkaProducer is implemented in a different way from KafkaProducer. The DISKafkaProducer client and server are implemented through the REST API, whereas KafkaProducer is implemented based on TCP. Their API compatibility differences are as follows:

Table 6-96 Adaptation description

Native KafkaProducer	Type	DISKafkaProducer	Description
Future<RecordMetadata> send(ProducerRecord<K, V> record)	API	Supported	Send a single data record.
Future<RecordMetadata> send(ProducerRecord<K, V> record, Callback callback)	API	Supported	Send a single data record and set the callback processing function.
void close()	API	Supported	Disable Producer.
void close(long timeout, TimeUnit timeUnit)	API	Supported	Disable Producer and set the timeout period.
List<PartitionInfo> partitionsFor(String topic)	API	Supported	Obtain the partition information of the stream.
void flush(long timeout, TimeUnit unit)	API	Not supported	Forcibly send the current cached data.
Map<MetricName, ? extends Metric> metrics()	API	Not supported	Obtain statistics.

 Copyright © Huawei Technologies Co., Ltd.

Native KafkaProducer	Type	DISKafkaProducer	Description
key.serializer	Parameter	Supported	The meaning of this parameter is the same as that in Kafka. The default value is **StringDeserializer**. In Kafka, this parameter has no default value, and you must configure a value for it.
value.serializer	Parameter	Supported	The meaning of this parameter is the same as that in Kafka. The default value is **StringDeserializer**. In Kafka, this parameter has no default value, and you must configure a value for it.
linger.ms	Parameter	Supported	The meaning of this parameter is the same as that in Kafka. The default value is **50**. In Kafka, the default value is **0**. This parameter is configured for improving the upload efficiency of the REST API.

 Copyright © Huawei Technologies Co., Ltd.

Native KafkaProducer	Type	DISKafkaProducer	Description
batch.size	Parameter	Supported	The meaning of this parameter is the same as that in Kafka. The default value is 1 MB. In Kafka, the default value is 16 KB). This parameter is configured for matching the flow control.
buffer.memory	Parameter	Supported	Same as the default setting in Kafka, which is 32 MB.
max.in.flight.requests.per.connection	Parameter	Supported	Limit the maximum number of not-responded requests that can be sent by a client on a single connection. The default value is **100**. In Kafka, the default value is 5. This parameter is configured for improving the sending performance. However, the data sequence may be inconsistent. You are advised to set this parameter to 1 to ensure data sequence.

 Copyright © Huawei Technologies Co., Ltd.

Native KafkaProducer	Type	DISKafkaProducer	Description
block.on.buffer.full	Parameter	Supported	Same as the default setting in Kafka, which is **false**. ● true: When the sending buffer is full, sending is blocked and does not time out. ● false: After the sending buffer is full, data is blocked based on max.block.ms. If the time is exceeded, an exception is issued.
max.block.ms	Parameter	Supported	Same as the default setting in Kafka, which is **60000**. When the sending buffer is full and **block.on.buffer.full** is **false**, control the block time (ms) of send ().
retries	Parameter	Supported, but the parameter name is changed to **exception.retries**.	The default value in Kafka is **0**, and the default value in DIS is **8**. Number of retry attempts that are made when the network or server is abnormal.
Others	Parameter	Not supported	-

6.2.8.4 Downloading Data

 Copyright © Huawei Technologies Co., Ltd.

6.2.8.5 Consumption Modes

Similar to Kafka, dis kafka adapter supports three consumption modes.

assign Mode

Users specify the partitions to be consumed by the consumer instance are manually. In this case, the group management mechanism is not used. That is, when the number of consumers in the group changes or the stream scaling is performed, the partitions will not be reallocated.

For details about the sample code, see DISKafkaConsumerAssignDemo.java in the sample project, as shown in the following:

```
package com.huaweicloud.dis.demo.adapter;

import com.huaweicloud.dis.DISConfig;
import com.huaweicloud.dis.adapter.kafka.clients.consumer.*;
import com.huaweicloud.dis.adapter.kafka.common.PartitionInfo;
import com.huaweicloud.dis.adapter.kafka.common.TopicPartition;
import com.huaweicloud.dis.adapter.kafka.common.serialization.StringDeserializer;
import org.slf4j.Logger;
import org.slf4j.LoggerFactory;

import java.util.ArrayList;
import java.util.List;
import java.util.Map;
import java.util.Properties;

public class DISKafkaConsumerAssignDemo
{
    private static final Logger LOGGER = LoggerFactory.getLogger(DISKafkaConsumerAssignDemo.class);

    public static void main(String[] args)
    {
        // YOU AK
        String ak = "YOU_AK";
        // YOU SK
        String sk = "YOU_SK";
        // YOU ProjectId
        String projectId = "YOU_PROJECT_ID";
        // YOU DIS Stream
        String streamName = "YOU_STREAM_NAME";
//Consumption group ID, which is used to record the offset.
        String groupId = "YOU_GROUP_ID";
        // DIS region
        String region = "YOU_Region";

        Properties props = new Properties();
        props.setProperty(DISConfig.PROPERTY_AK, ak);
        props.setProperty(DISConfig.PROPERTY_SK, sk);
        props.setProperty(DISConfig.PROPERTY_PROJECT_ID, projectId);
        props.setProperty(DISConfig.PROPERTY_REGION_ID, region);
        props.setProperty(ConsumerConfig.KEY_DESERIALIZER_CLASS_CONFIG,
StringDeserializer.class.getName());
        props.setProperty(ConsumerConfig.VALUE_DESERIALIZER_CLASS_CONFIG,
StringDeserializer.class.getName());
        props.setProperty(ConsumerConfig.GROUP_ID_CONFIG, groupId);
        props.setProperty(ConsumerConfig.ENABLE_AUTO_COMMIT_CONFIG, "false");
        props.setProperty(ConsumerConfig.AUTO_OFFSET_RESET_CONFIG,
OffsetResetStrategy.LATEST.name());

//By default, the domain name is automatically used for access instead of configuring an endpoint. If an
endpoint is required, remove the following comments and set the endpoint:
        // props.setProperty(DISConfig.PROPERTY_ENDPOINT, "https://dis-${region}.myhuaweicloud.com");

        Consumer<String, String> consumer = new DISKafkaConsumer<>(props);
```

 Copyright © Huawei Technologies Co., Ltd.

```
        List<TopicPartition> topicPartitions = new ArrayList<>();
        for (PartitionInfo partitionInfo : consumer.partitionsFor(streamName))
        {
            topicPartitions.add(new TopicPartition(partitionInfo.topic(), partitionInfo.partition()));
        }

//Use the assign mode to specify the partition to be consumed.
        consumer.assign(topicPartitions);

        while (true)
        {
            try
            {
                ConsumerRecords<String, String> records = consumer.poll(Long.MAX_VALUE);

                if (!records.isEmpty())
                {
                    for (TopicPartition partition : records.partitions())
                    {
                        List<ConsumerRecord<String, String>> partitionRecords = records.records(partition);
                        for (ConsumerRecord<String, String> record : partitionRecords)
                        {
                            LOGGER.info("Value [{}], Partition [{}], Offset [{}], Key [{}]",
                                    record.value(), record.partition(), record.offset(), record.key());
                        }
                    }
//Submit the current offset asynchronously after data processing is complete or submit commitSync
synchronously.
                    consumer.commitAsync(new OffsetCommitCallback()
                    {
                        @Override
                        public void onComplete(Map<TopicPartition, OffsetAndMetadata> map, Exception e)
                        {
                            if (e == null)
                            {
                                LOGGER.debug("Success to commit offset [{}]", map);
                            }
                            else
                            {
                                LOGGER.error("Failed to commit offset [{}]", e.getMessage(), e);
                            }
                        }
                    });
                }
            }
            catch (Exception e)
            {
                LOGGER.info(e.getMessage(), e);
            }
        }
    }
}
```

After running the preceding program, if the data is sent to the stream, the following log is generated:

```
09:36:45.071 INFO  c.h.d.a.k.c.DISKafkaConsumer - create DISKafkaConsumer successfully
09:36:49.842 INFO  c.h.d.d.a.DISKafkaConsumerAssignDemo - Value [Hello world[sync]. 0], Partition [0],
Offset [134], Key [769066]
09:36:49.963 INFO  c.h.d.d.a.DISKafkaConsumerAssignDemo - Value [Hello world[sync]. 1], Partition [0],
Offset [135], Key [700005]
09:36:50.145 INFO  c.h.d.d.a.DISKafkaConsumerAssignDemo - Value [Hello world[sync]. 2], Partition [0],
Offset [136], Key [338044]
09:36:51.093 INFO  c.h.d.d.a.DISKafkaConsumerAssignDemo - Value [Hello world[sync]. 3], Partition [0],
Offset [137], Key [537495]
09:36:51.093 INFO  c.h.d.d.a.DISKafkaConsumerAssignDemo - Value [Hello world[sync]. 4], Partition [0],
Offset [138], Key [980016]
09:36:51.093 INFO  c.h.d.d.a.DISKafkaConsumerAssignDemo - Value [Hello world[async]. 0], Partition [0],
Offset [139], Key [182772]
09:36:51.093 INFO  c.h.d.d.a.DISKafkaConsumerAssignDemo - Value [Hello world[async]. 1], Partition [0],
```

 Copyright © Huawei Technologies Co., Ltd.

```
Offset [140], Key [830271]
09:36:51.093 INFO  c.h.d.d.a.DISKafkaConsumerAssignDemo - Value [Hello world[async]. 2], Partition [0],
Offset [141], Key [927054]
09:36:51.093 INFO  c.h.d.d.a.DISKafkaConsumerAssignDemo - Value [Hello world[async]. 3], Partition [0],
Offset [142], Key [268122]
09:36:51.093 INFO  c.h.d.d.a.DISKafkaConsumerAssignDemo - Value [Hello world[async]. 4], Partition [0],
Offset [143], Key [992787]
```

subscribe Mode

Users only need to specify stream names without specifying specific partitions. The server will trigger the group management mechanism based on the number of consumers or stream scaling changes to automatically allocate partitions to each consumer. This ensures that a partition is consumed by only one consumer.

For details about the sample code, see DISKafkaConsumerSubscribeDemo.java in the sample project, as shown in the following:

```
package com.huaweicloud.dis.demo.adapter;

import com.huaweicloud.dis.DISConfig;
import com.huaweicloud.dis.adapter.kafka.clients.consumer.*;
import com.huaweicloud.dis.adapter.kafka.common.TopicPartition;
import com.huaweicloud.dis.adapter.kafka.common.serialization.StringDeserializer;
import org.slf4j.Logger;
import org.slf4j.LoggerFactory;

import java.util.Collections;
import java.util.List;
import java.util.Map;
import java.util.Properties;

public class DISKafkaConsumerSubscribeDemo
{
    private static final Logger LOGGER = LoggerFactory.getLogger(DISKafkaConsumerSubscribeDemo.class);

    public static void main(String[] args)
    {
        // YOU AK
        String ak = "YOU_AK";
        // YOU SK
        String sk = "YOU_SK";
        // YOU ProjectId
        String projectId = "YOU_PROJECT_ID";
        // YOU DIS Stream
        String streamName = "YOU_STREAM_NAME";
//Consumption group ID, which is used to record the offset and perform group rebalance.
        String groupId = "YOU_GROUP_ID";
        // DIS region
        String region = "YOU_Region";

        Properties props = new Properties();
        props.setProperty(DISConfig.PROPERTY_AK, ak);
        props.setProperty(DISConfig.PROPERTY_SK, sk);
        props.setProperty(DISConfig.PROPERTY_PROJECT_ID, projectId);
        props.setProperty(DISConfig.PROPERTY_REGION_ID, region);
        props.setProperty(ConsumerConfig.KEY_DESERIALIZER_CLASS_CONFIG,
StringDeserializer.class.getName());
        props.setProperty(ConsumerConfig.VALUE_DESERIALIZER_CLASS_CONFIG,
StringDeserializer.class.getName());
        props.setProperty(ConsumerConfig.GROUP_ID_CONFIG, groupId);
        props.setProperty(ConsumerConfig.ENABLE_AUTO_COMMIT_CONFIG, "false");
        props.setProperty(ConsumerConfig.AUTO_OFFSET_RESET_CONFIG,
OffsetResetStrategy.LATEST.name());

//By default, the domain name is automatically used for access instead of configuring an endpoint. If an
```

 Copyright © Huawei Technologies Co., Ltd.

```
endpoint is required, remove the following comments and set the endpoint:
       // props.setProperty(DISConfig.PROPERTY_ENDPOINT, "https://dis-${region}.myhuaweicloud.com");

       Consumer<String, String> consumer = new DISKafkaConsumer<>(props);
      //When using the subscribe mode, specify the name of the stream to be consumed.
       consumer.subscribe(Collections.singleton(streamName));

       while (true)
       {
          try
          {
             ConsumerRecords<String, String> records = consumer.poll(Long.MAX_VALUE);

             if (!records.isEmpty())
             {
                for (TopicPartition partition : records.partitions())
                {
                   List<ConsumerRecord<String, String>> partitionRecords = records.records(partition);
                   for (ConsumerRecord<String, String> record : partitionRecords)
                   {
                      LOGGER.info("Value [{}], Partition [{}], Offset [{}], Key [{}]",
                            record.value(), record.partition(), record.offset(), record.key());
                   }
                }
//Submit the current offset aynchronously after data processing is complete or submit commitSync
synchronously.
                consumer.commitAsync(new OffsetCommitCallback()
                {
                   @Override
                   public void onComplete(Map<TopicPartition, OffsetAndMetadata> map, Exception e)
                   {
                      if (e == null)
                      {
                         LOGGER.debug("Success to commit offset [{}]", map);
                      }
                      else
                      {
                         LOGGER.error("Failed to commit offset [{}]", e.getMessage(), e);
                      }
                   }
                });
             }
          }
          catch (Exception e)
          {
             LOGGER.info(e.getMessage(), e);
          }
       }
    }
}
```

When the program runs, it sends a heartbeat request (Heartbeat) every 10 seconds and then sends a (JoinGroup) request to join the consumer group. The server starts to allocate partitions to consumers in the consumer group. This process takes about 20s. After the operation is complete, the consumers send synchronization requests (SyncGroup) to obtain allocation results. If the information about heartbeat {"state": "STABLE"} is recorded in the log, it indicates that allocation to the entire consumer group has been completed, data is ready to be consumed.

The key logs in this process are described as follows:

- Heartbeat {"state":"JOINING"}

Heartbeat: indicates a heartbeat request. The heartbeat request is initiated every 10 seconds and is used to keep a connection with the server. If the server does not receive the heartbeat message within 1 minute, it considers that the consumer is

 Copyright © Huawei Technologies Co., Ltd.

offline and the partitions will be reallocated to the consumers in the consumer group. If the heartbeat result is JOINING, the consumer needs to join the consumer group again. If the heartbeat result is STABLE, the consumer group is stable.

- JoinGroup

If the heartbeat result is not STABLE, the consumer sends a joinGroup request to notify the server that it needs to join the consumer group. After receiving the join request from the client, the server will reallocate partitions for the consumer group. In this case, syncDelayedTimeMs is returned, indicating the allocation duration. After allocation is completed, the client sends a synchronization request (SyncGroup) to obtain the allocation result.

- SyncGroup

A SyncGroup request is used to obtain the allocation result. The returned assignment contains the stream name and partition to be consumed.

Run the sample program. After allocation is completed, send data to the stream. The complete log is as follows:

```
09:42:37.296 INFO  c.h.d.a.k.c.DISKafkaConsumer - create DISKafkaConsumer successfully
09:42:37.354 INFO  c.h.d.a.k.c.Coordinator - Heartbeat {"state":"JOINING"}
09:42:37.363 INFO  c.h.d.a.k.c.Coordinator - joinGroupRequest {"groupId":"ding","clientId":"consumer-c2d43144-0823-4eea-aaa8-7af95c536144","interestedStream":["liuhao12"]}
09:42:37.406 INFO  c.h.d.a.k.c.Coordinator - joinGroupResponse {"state":"OK","syncDelayedTimeMs":21000}
09:42:58.408 INFO  c.h.d.a.k.c.Coordinator - syncGroup {"groupId":"ding","clientId":"consumer-c2d43144-0823-4eea-aaa8-7af95c536144","generation":-1}
09:42:58.451 INFO  c.h.d.a.k.c.Coordinator - syncGroup {"state":"OK","generation":33,"assignment":{"dis-test":[0]}}
09:42:58.488 INFO  c.h.d.a.k.c.Coordinator - Heartbeat {"state":"STABLE"}
09:43:08.960 INFO  c.h.d.a.k.c.Coordinator - Heartbeat {"state":"STABLE"}
09:46:56.227 INFO  c.h.d.d.a.DISKafkaConsumerSubscribeDemo - Value [Hello world[sync]. 0], Partition [0], Offset [144], Key [799704]
09:46:56.327 INFO  c.h.d.d.a.DISKafkaConsumerSubscribeDemo - Value [Hello world[sync]. 1], Partition [0], Offset [145], Key [683757]
09:46:56.449 INFO  c.h.d.d.a.DISKafkaConsumerSubscribeDemo - Value [Hello world[sync]. 2], Partition [0], Offset [146], Key [439062]
09:46:56.535 INFO  c.h.d.d.a.DISKafkaConsumerSubscribeDemo - Value [Hello world[sync]. 3], Partition [0], Offset [147], Key [374939]
09:46:56.654 INFO  c.h.d.d.a.DISKafkaConsumerSubscribeDemo - Value [Hello world[sync]. 4], Partition [0], Offset [148], Key [321528]
09:46:56.749 INFO  c.h.d.d.a.DISKafkaConsumerSubscribeDemo - Value [Hello world[async]. 0], Partition [0], Offset [149], Key [964841]
09:46:56.749 INFO  c.h.d.d.a.DISKafkaConsumerSubscribeDemo - Value [Hello world[async]. 1], Partition [0], Offset [150], Key [520262]
09:46:56.749 INFO  c.h.d.d.a.DISKafkaConsumerSubscribeDemo - Value [Hello world[async]. 2], Partition [0], Offset [151], Key [619119]
09:46:56.749 INFO  c.h.d.d.a.DISKafkaConsumerSubscribeDemo - Value [Hello world[async]. 3], Partition [0], Offset [152], Key [257094]
09:46:56.749 INFO  c.h.d.d.a.DISKafkaConsumerSubscribeDemo - Value [Hello world[async]. 4], Partition [0], Offset [153], Key [310331]
```

subscribePattern Mode

Users specify a wildcard instead of specifying a stream name. For example, stream.* indicates that stream1, stream2, stream_123, and so on are consumed. An existing, added, or deleted stream can be consumed by a consumer group as long as it matches the wildcard.

 Copyright © Huawei Technologies Co., Ltd.

For details about the sample code, see DISKafkaConsumerSubscribePatternDemo.java in the sample project, as shown in the following:

```
package com.huaweicloud.dis.demo.adapter;

import com.huaweicloud.dis.DISConfig;
import com.huaweicloud.dis.adapter.kafka.clients.consumer.*;
import com.huaweicloud.dis.adapter.kafka.common.TopicPartition;
import com.huaweicloud.dis.adapter.kafka.common.serialization.StringDeserializer;
import org.slf4j.Logger;
import org.slf4j.LoggerFactory;

import java.util.Collection;
import java.util.List;
import java.util.Map;
import java.util.Properties;
import java.util.regex.Pattern;

public class DISKafkaConsumerSubscribePatternDemo
{
    private static final Logger LOGGER =
LoggerFactory.getLogger(DISKafkaConsumerSubscribePatternDemo.class);

    public static void main(String[] args)
    {
        // YOU AK
        String ak = "YOU_AK";
        // YOU SK
        String sk = "YOU_SK";
        // YOU ProjectId
        String projectId = "YOU_PROJECT_ID";
// Wildcard of the stream to be consumed (stream.* indicates that stream1, stream2, stream_123, and so on
will be consumed.)
        String streamNamePattern = "stream.*";
//Consumption group ID, which is used to record the offset and perform group rebalance.
        String groupId = "YOU_GROUP_ID";
        // DIS region
        String region = "YOU_Region";

        Properties props = new Properties();
        props.setProperty(DISConfig.PROPERTY_AK, ak);
        props.setProperty(DISConfig.PROPERTY_SK, sk);
        props.setProperty(DISConfig.PROPERTY_PROJECT_ID, projectId);
        props.setProperty(DISConfig.PROPERTY_REGION_ID, region);
        props.setProperty(ConsumerConfig.KEY_DESERIALIZER_CLASS_CONFIG,
StringDeserializer.class.getName());
        props.setProperty(ConsumerConfig.VALUE_DESERIALIZER_CLASS_CONFIG,
StringDeserializer.class.getName());
        props.setProperty(ConsumerConfig.GROUP_ID_CONFIG, groupId);
        props.setProperty(ConsumerConfig.ENABLE_AUTO_COMMIT_CONFIG, "false");
        props.setProperty(ConsumerConfig.AUTO_OFFSET_RESET_CONFIG,
OffsetResetStrategy.LATEST.name());

//By default, the domain name is automatically used for access instead of configuring an endpoint. If an
endpoint is required, remove the following comments and set the endpoint:
        // props.setProperty(DISConfig.PROPERTY_ENDPOINT, "https://dis-${region}.myhuaweicloud.com");

        Consumer<String, String> consumer = new DISKafkaConsumer<>(props);
//With the subscribePattern mode, you only need to specify a wildcard.
        consumer.subscribe(Pattern.compile(streamNamePattern), new ConsumerRebalanceListener()
        {
            @Override
            public void onPartitionsRevoked(Collection<TopicPartition> collection)
            {
                LOGGER.info("onPartitionsRevoked [{}]", collection);
            }

            @Override
```

 Copyright © Huawei Technologies Co., Ltd.

```
            public void onPartitionsAssigned(Collection<TopicPartition> collection)
            {
                LOGGER.info("onPartitionsAssigned [{}]", collection);
            }
        });

        while (true)
        {
            try
            {
                ConsumerRecords<String, String> records = consumer.poll(Long.MAX_VALUE);

                if (!records.isEmpty())
                {
                    for (TopicPartition partition : records.partitions())
                    {
                        List<ConsumerRecord<String, String>> partitionRecords = records.records(partition);
                        for (ConsumerRecord<String, String> record : partitionRecords)
                        {
                            LOGGER.info("Value [{}], Partition [{}], Offset [{}], Key [{}]",
                                    record.value(), record.partition(), record.offset(), record.key());
                        }
                    }
//Submit the current offset asynchronously after data processing is complete or submit commitSync synchronously.
                    consumer.commitAsync(new OffsetCommitCallback()
                    {
                        @Override
                        public void onComplete(Map<TopicPartition, OffsetAndMetadata> map, Exception e)
                        {
                            if (e == null)
                            {
                                LOGGER.debug("Success to commit offset [{}]", map);
                            }
                            else
                            {
                                LOGGER.error("Failed to commit offset [{}]", e.getMessage(), e);
                            }
                        }
                    });
                }
            }
            catch (Exception e)
            {
                LOGGER.info(e.getMessage(), e);
            }
        }
    }
}
```

After the program runs, wait about 20s until the allocation is completed. The data can be consumed. The following is an example code execution log:

```
10:10:36.420 INFO  c.h.d.a.k.c.DISKafkaConsumer - create DISKafkaConsumer successfully
10:10:36.481 INFO  c.h.d.a.k.c.Coordinator - Heartbeat {"state":"JOINING"}
10:10:36.486 INFO  c.h.d.a.k.c.Coordinator - joinGroupRequest {"groupId":"ding","clientId":"consumer-cad967ba-70ab-4e02-b184-f60b95fe3256","streamPattern":"stream.*"}
10:10:36.697 INFO  c.h.d.a.k.c.Coordinator - joinGroupResponse {"state":"OK","subscription":["stream_hello","stream_world"],"syncDelayedTimeMs":21000}
10:10:57.699 INFO  c.h.d.a.k.c.Coordinator - syncGroup {"groupId":"ding","clientId":"consumer-cad967ba-70ab-4e02-b184-f60b95fe3256","generation":-1}
10:10:57.746 INFO  c.h.d.a.k.c.Coordinator - syncGroup {"state":"OK","generation":34,"assignment":{"stream_hello":[0],"stream_world":[0]}}
10:10:57.770 INFO  c.h.d.d.a.DISKafkaConsumerSubscribePatternDemo - onPartitionsAssigned [[stream_hello-0, stream_world-0]]
10:10:57.770 INFO  c.h.d.a.k.c.Coordinator - Heartbeat {"state":"STABLE"}
10:11:08.466 INFO  c.h.d.a.k.c.Coordinator - Heartbeat {"state":"STABLE"}
10:11:09.992 INFO  c.h.d.d.a.DISKafkaConsumerSubscribePatternDemo - Value [Hello world[sync]. 0], Partition [0], Offset [154], Key [181881]
10:11:09.993 INFO  c.h.d.d.a.DISKafkaConsumerSubscribePatternDemo - Value [Hello world[sync]. 1],
```

 Copyright © Huawei Technologies Co., Ltd.

```
Partition [0], Offset [155], Key [483023]
10:11:09.993 INFO  c.h.d.d.a.DISKafkaConsumerSubscribePatternDemo - Value [Hello world[sync]. 2],
Partition [0], Offset [156], Key [32453]
10:11:10.093 INFO  c.h.d.d.a.DISKafkaConsumerSubscribePatternDemo - Value [Hello world[sync]. 3],
Partition [0], Offset [157], Key [111948]
10:11:10.180 INFO  c.h.d.d.a.DISKafkaConsumerSubscribePatternDemo - Value [Hello world[sync]. 4],
Partition [0], Offset [158], Key [822860]
```

6.2.8.6 Consumption Offset

Two offset committing policies are available: automatic and manual. When creating a DISKafkaConsumer object, set the **enable.auto.commit** parameter to specify a desired offset committing policy. If the value is set to **true**, automatic offset committing is used. If the value is set to **false**, manual offset committing is used.

With automatic offset committing, the consumer enables the coordinator to commit offsets every *auto.commit.interval.ms*. With manual offset committing, instead of relying on the consumer to periodically commit consumed offsets, users can control when records should be considered as consumed and hence commit their offsets.

- Automatic

  When a consumer is created, automatic offset committing is set by default. The default committing interval is 5,000 ms. Parameters for automatic offset committing are as follows:

```
Props.setProperty("enable.auto.commit", "true");//Automatic offset committing is used.
Props.setProperty("auto.commit.interval.ms", "5000");//Offsets are committed at an interval of 5,000 ms.
```

- Automatic

In some scenarios, offsets need to be more strictly managed to ensure that messages are not repeatedly consumed or are not lost. For example, the pulled messages need to be written to the database for processing, or are used for complex service processing such as processing other network access requests. In such scenarios, the messages are regarded as successfully consumed only after all services are processed. In this case, you must manually control offset committing. Parameters for manual offset committing are as follows:

```
props.put("enable.auto.commit", "false");//Manual offset committing is used.
```

After the services are successfully processed, call the commitAsync() or commitSync() method to commit offsets. commitAsync() is used to commit offsets asynchronously. With this method, the consumer threads will not be blocked, and the next pull operation may be started before the offset committing result is returned. To obtain the committing result, add the OffsetCommitCallback method. After the offsets are committed, the onComplete() method is automatically called for processing of different logics based on the callback results.

CommitSync() is used to commit offsets synchronously. With this method, the consumer thread will be blocked until the offset committing result is returned.

In addition, the specific offset data of the specific partition may be further controlled. The confirmed offset is the maximum offset of the accepted data plus 1. For example, when a batch of data is consumed and the offset of the last record is 100, commit offset 101. In this case, consumption starts from the record whose offset is 101 and data will not be consumed repeatedly. The code sample is as follows:

 Copyright © Huawei Technologies Co., Ltd.

```
ConsumerRecords<String, String> records = consumer.poll(Long.MAX_VALUE);

if (!records.isEmpty())
{
    for (TopicPartition partition : records.partitions())
    {
        List<ConsumerRecord<String, String>> partitionRecords = records.records(partition);
        for (ConsumerRecord<String, String> record : partitionRecords)
        {
            LOGGER.info("Value [{}], Partition [{}], Offset [{}], Key [{}]",
                    record.value(), record.partition(), record.offset(), record.key());
        }
        if (!partitionRecords.isEmpty())
        {
// Confirm the specific offset of a partition synchronously.
            long lastOffset = partitionRecords.get(partitionRecords.size() - 1).offset();
            consumer.commitSync(Collections.singletonMap(partition, new OffsetAndMetadata(lastOffset
+ 1)));
        }
    }
}
```

6.2.8.7 Adaptation to the Native KafkaConsumer API

Table 6-97 API adaptation description

Native KafkaConsumer	Type	DISKafkaConsumer	Description
Set<TopicPartition> assignment()	API	Supported	Obtain information about the consumed stream and partition.
Set<String> subscription()	API	Supported	Obtain the name of the stream that has been subscribed to by the consumer.
void assign(Collection<TopicPartition> var1)	API	Supported	Allocate a specified partition.
void subscribe(Collection<String> var1)	API	Supported	Subscribe to a specified stream.
void subscribe(Collection<String> var1, ConsumerRebalanceListener var2)	API	Supported	Subscribe to a specified stream and supports callback of ConsumerRebalanceListener.

 Copyright © Huawei Technologies Co., Ltd.

Native KafkaConsumer	Type	DISKafkaConsumer	Description
void subscribe(Pattern var1, ConsumerRebalanceListener var2)	API	Supported	Subscribe to all streams that match wildcards and supports callback of ConsumerRebalanceListener.
void unsubscribe()	API	Supported	Cancels all subscription.
ConsumerRecords <K, V> poll(long var1)	API	Supported	The message is obtained. Checksum (CRC32 check value of the message), serializedKeySize (byte length after key serialization), and serializedValue-Size (byte length after key serialization).
void commitSync()	API	Supported	Commit the consumed offsets synchronously.
void commitSync(final Map<TopicPartition, OffsetAndMetadata> offsets)	API	Supported	Commit the specified offset synchronously.
void commitAsync()	API	Supported	Commit the consumed offsets asynchronously.
public void commitAsync(OffsetCommitCallback callback)	API	Supported	Commit the current consumed offset asynchronously and supports callback of OffsetCommitCall-back.

 Copyright © Huawei Technologies Co., Ltd.

Native KafkaConsumer	Type	DISKafkaConsumer	Description
void commitAsync(Map<TopicPartition, OffsetAndMetadata> offsets, OffsetCommitCallback callback)	API	Supported	Commit the specified offset asynchronously and supports callback of OffsetCommitCallback.
void seek(TopicPartition partition, long offset)	API	Supported	Set the specified offset for a partition.
void seekToBeginning(Collection<TopicPartition> partitions)	API	Supported	Set the earliest value for the partition offset.
void seekToEnd(Collection<TopicPartition> partitions)	API	Supported	Set the latest value for the partition offset.
long position(TopicPartition partition)	API	Supported	Obtain the offset of the current consumed data in the partition.
OffsetAndMetadata committed(TopicPartition partition)	API	Supported	Obtain the committed offset of the partition.
List<PartitionInfo> partitionsFor(String topic)	API	Supported	Obtain the partition information of the stream, but leader, replicas, and inSyncReplicas in PartitionInfo are not implemented.
Map<String, List<PartitionInfo>> listTopics()	API	Supported	Obtain the stream information, but leader, replicas, and inSyncReplicas in PartitionInfo are not implemented.

 Copyright © Huawei Technologies Co., Ltd.

Native KafkaConsumer	Type	DISKafkaConsumer	Description
void pause(Collection<TopicPartition> partitions)	API	Supported	Suspends the partition consumption.
void resume(Collection<TopicPartition> partitions)	API	Supported	Resumes the partition consumption.
Set<TopicPartition> paused()	API	Supported	Obtain all partitions that stop being consumed.
close()	API	Supported	Disable the consumer.
Map<MetricName, ? extends Metric> metrics()	API	Not supported	Obtain statistics.
wakeup()	API	Not supported	The internal implementation principles are different.
group.id	Parameter	Supported	Consumer group ID.
client.id	Parameter	Supported	**client.id** of each consumer must be unique. If **client.id** is not specified, the dis kafka consumer will generate a UUID as a **client.id**.
key.deserializer	Parameter	Supported	The meaning of this parameter is the same as that in Kafka. The default value is **StringDeserializer**. In Kafka, **StringDeserializer** has no default value, and you must configure a value for it.

 Copyright © Huawei Technologies Co., Ltd.

Native KafkaConsumer	Type	DISKafkaConsumer	Description
value.deserializer	Parameter	Supported	The meaning of this parameter is the same as that in Kafka. The default value is **StringDeserializer**. In Kafka, this parameter has no default value, and you must configure a value for it.
enable.auto.commit	Parameter	Supported	The default value is the same as that in Kafka, which is **true**. • **true**: The automatic offset committing is enabled and offsets are automatically committed at an interval of *auto.commit.interval.ms*. • **false**: Offsets are not automatically committed.
auto.commit.interval.ms	Parameter	Supported	Interval for automatically committing offsets, in milliseconds. The default value is **5000**.

 Copyright © Huawei Technologies Co., Ltd.

Native KafkaConsumer	Type	DISKafkaConsumer	Description
auto.offset.reset	Parameter	Supported	The default value is the same as that in Kafka, which is **latest**. This parameter is used to automatically set the offset position when there is no initial offset or the offset is incorrect. • **earliest**: The offset is automatically reset to the earliest value. • **latest**: The offset is automatically reset to the latest value. • **none**: If the previous offset is not found in the consumer group, an exception is issued to the consumer.
Others	Parameter	Not supported	-

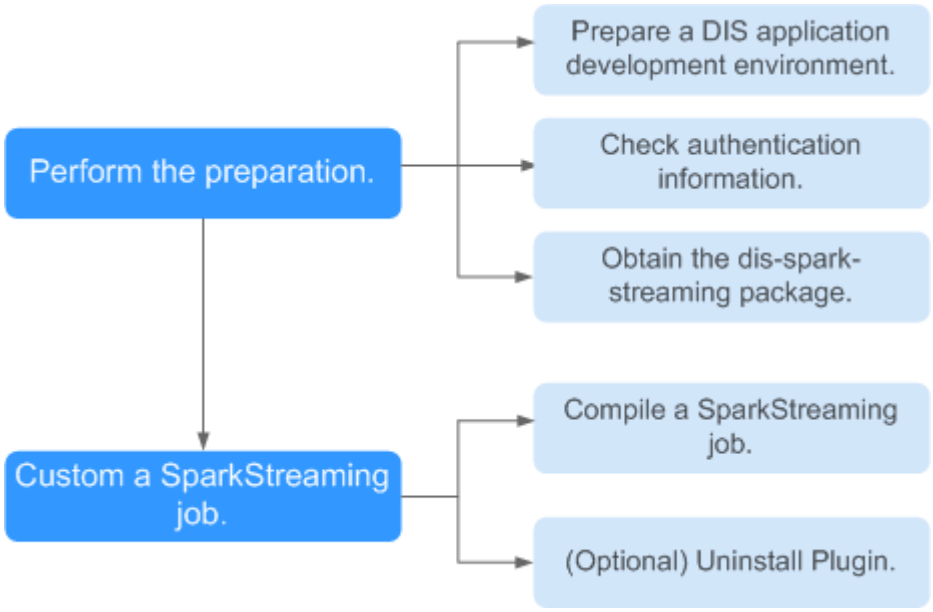
6.2.9 Using SparkStreaming SDK to Download Data

6.2.9.1 Overview

DIS Spark Streaming is a software development kit (SDK) provided by DIS to create a DStream that uses the DIS as a data source and runs with SparkStreaming. **Figure 6-149** depicts the process for using DIS Spark Streaming.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-149 DIS Spark Streaming usage process



6.2.9.2 Preparing a DIS Spark Streaming Environment

Preparing a DIS Application Development Environment

Step 1 Set up DIS application development environments by following the procedure in **Buying a Pay-Per-Use DIS Incremental Package**.

Step 2 Install Maven. Configure a local repository address.

Step 3 Install scala-sdk.

----End

Configuring the DIS Spark Streaming Dependency

The DIS Spark Streaming dependency can be introduced to the project through the following configuration:

```
<dependency>
   <groupId>com.huaweicloud.dis</groupId>
   <artifactId>huaweicloud-dis-spark-streaming_2.11</artifactId>
   <version>1.2.1</version>
   <scope>compile</scope>
</dependency>
```

Checking Authentication Information

- AK/SK file

  Access Key ID/Secret Access Key (AK/SK) files are created by the Identity and Access Management (IAM) service to authenticate calls to application programming interfaces (APIs) on the cloud. To obtain AK/SK, choose **My Credentials > Access Keys**.

- Project ID

 Copyright © Huawei Technologies Co., Ltd.

A project is a group of tenant resources. A tenant can have multiple projects, one for each region. Each region has a unique Project ID. To view the project IDs of different regions, choose .

6.2.9.3 Customizing a SparkStreaming Job

Obtaining the DIS Spark Streaming Demo

Step 1 Obtain the **dis-spark-streaming-*X.X.X*.zip** package from **https://dis-publish.obs-website.cn-north-1.myhuaweicloud.com/**. Decompress the **dis-spark-streaming-*X.X.X*.zip** package to obtain the following directory that contains a maven project sample:

- dis-spark-streaming-demo: The **dis-spark-streaming-demo** directory contains a maven project example.

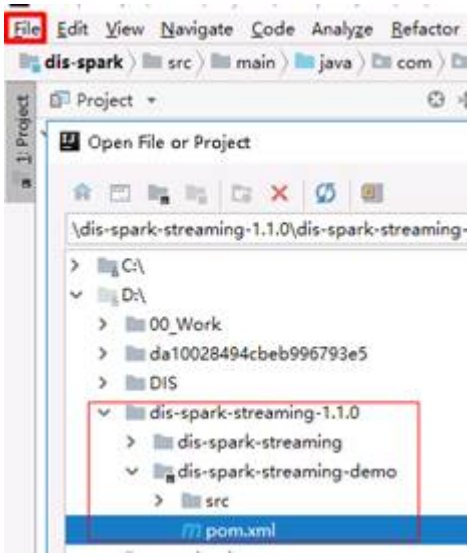
----End

Building a SparkStreaming Job

The following uses the IntelliJ IDEA community version as an example to describe how to build a SparkStreaming job. Make sure that the SparkStreaming job has been configured on the IDEA.

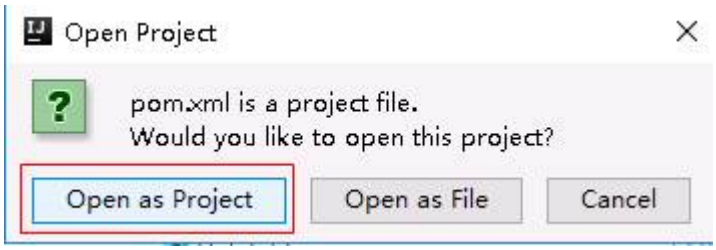
- JDK 1.8+
- Scala-sdk-2.11
- Maven 3.3.*

Step 1 Start IntelliJIDEA and choose **File** > **Open**. Select the **dis-spark-streaming-demo** directory and double-click **pom.xml**.



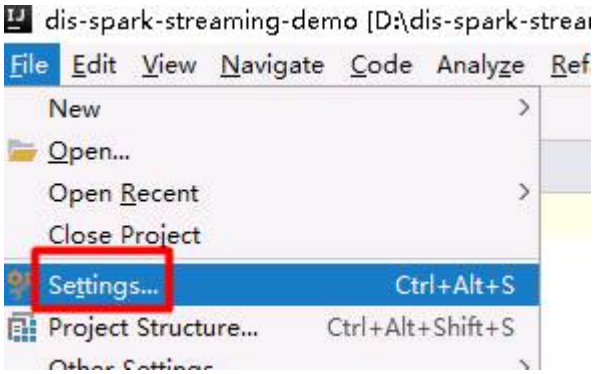
Step 2 When the following dialog box is displayed, select **Open as Project**.

 Copyright © Huawei Technologies Co., Ltd.



Step 3 Click **New Window** to open the project in a new window.

Step 4 In the displayed IDEA window, choose **File** > **Settings**.

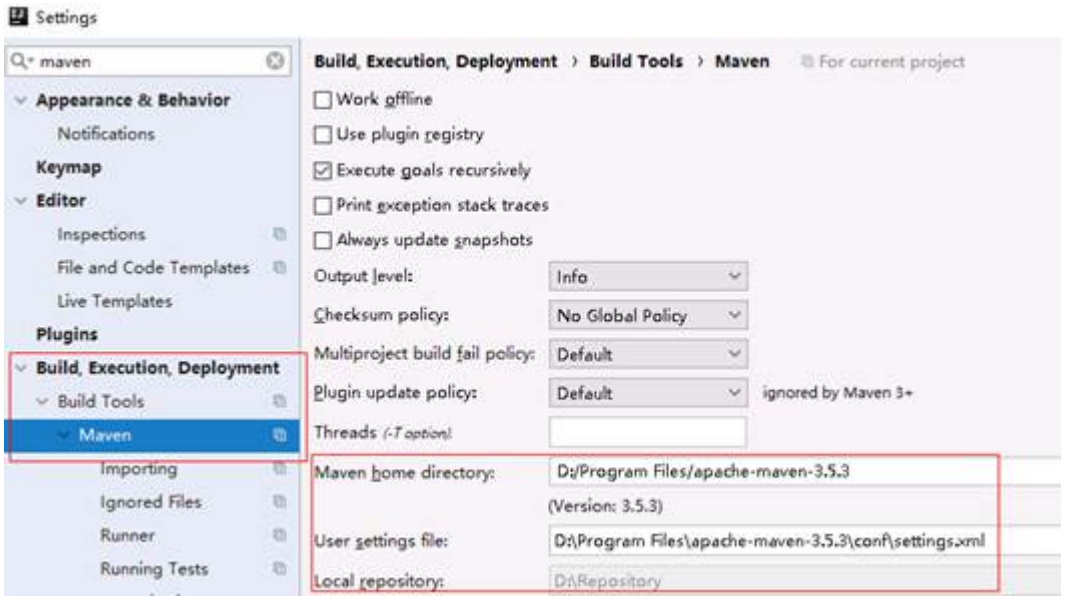


Step 5 Enter **maven** in the search box and find the maven configuration. Ensure that **Maven home directory** (maven installation path), **User settings file** (**settings.xml** file location), and **Local repository** (local repository address) are correctly configured.

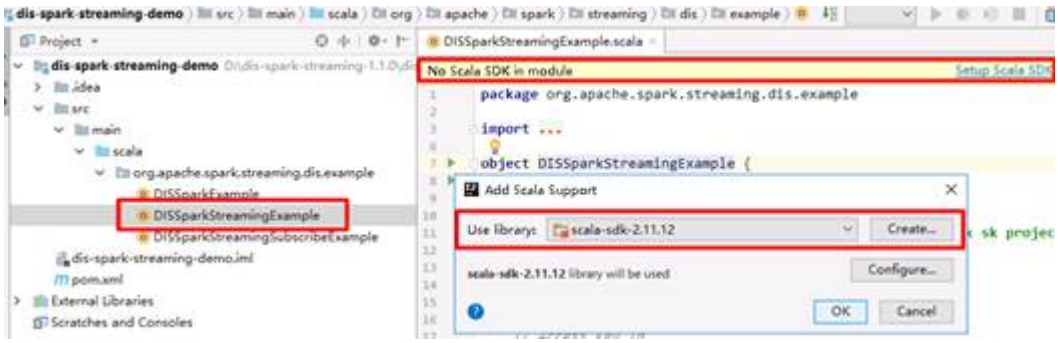
NOTE

If configurations are incorrect, modify them. Otherwise, the SDK installed in **Step 2** cannot be found.

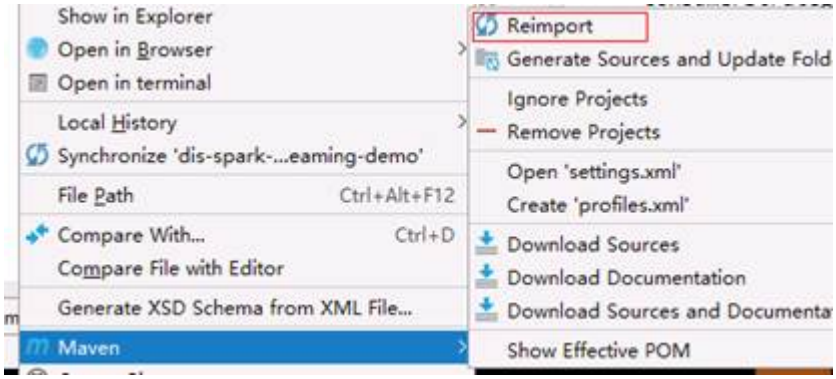
 Copyright © Huawei Technologies Co., Ltd.



Step 6 Open the **DISSparkStreamingExample** file. If the IDEA displays the message "No Scala SDK in module", click **Setup Scala SDK** next to it and select one from the Scala SDK list. If no Scala SDK is available, create one and associate it with the Scala path. Version 2.11 can be selected.



Step 7 Right-click **pom.xml** and choose **Maven** > **Reimport** from the shortcut menu to import the maven dependency library again.



Step 8 If there is no error in the **DISSparkStreamingExample** file opened by IDEA, the development environment is configured successfully. The logic of this file is to read the data in DIS streams and count the occurrence times of each word.

 Copyright © Huawei Technologies Co., Ltd.

1. DISSparkStreamingExample is an example of using the Assign mode. It does not have the capability to start from the last stop. The SDK construction method is as follows:

```
ConsumerStrategies.Assign[String, String](streamName, params, startingOffsets)
```

- StreamName: a DIS stream name.
- params: a set of parameter maps. It includes at least **endpoint** (DIS gateway address), **region** (region where DIS resides), **ak** (user AK), **sk** (user SK), and **projectId** (user project ID).
- StartingOffsets: start position for reading DIS data. LATEST indicates read from the latest data. EARLIEST indicates reading from the oldest data. If the exact start position of each partition needs to be specified, the value can be a JSON character string. For example, {"0" :23, "1" :-1, "2" :-2}, which indicates that the start position of partition 0 is 23, the first partition starts from the location of the latest data, and the second partition starts from the position of the earliest data. If there is no specified location for a partition, the system starts from the latest data position by default.

2. DISSparkStreamingExample: an example of using the Subscribe mode. It has the capability to start from the last stop. The SDK construction method is as follows:

```
ConsumerStrategies.Subscribe[String, String](Array(streamName), params)
```

- StreamName: a DIS stream name.
- params: a set of parameter maps. It includes at least **endpoint** (DIS gateway address), **region** (region where DIS resides), **ak** (user AK), **sk** (user SK), **projectId** (user project ID), and **group.id** (app name, indicating a consumer group). It can also include **auto.offset.reset**. The meaning of this parameter is the same as that of **startingOffsets** in Assign mode. If **enable.auto.commit** is set to **true**, the system automatically submits the offset every 5000 ms (which can be modified by setting **auto.commit.interval.ms**). If the parameter is set to **false**, the system does not automatically submit the offset. You can manually invoke **commitAsync** to submit the offset. For details, see the following part in the sample code:

```
stream.foreachRDD { rdd =>
    val offsetRanges = rdd.asInstanceOf[HasOffsetRanges].offsetRanges
    // commit offset to DIS async.
    stream.asInstanceOf[CanCommitOffsets].commitAsync(offsetRanges)
}
```

----End

Testing a SparkStreaming Job

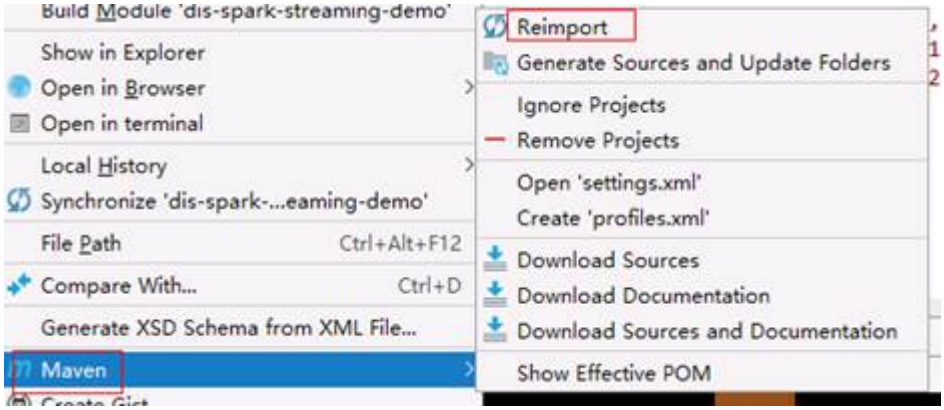
In a real-world scenario, the SparkStreming job needs to be submitted to the Spark cluster. However, this demonstration only describes how to test the job in the local IDE to understand the basic usage of the SDK. After the test is complete, you can create clusters (such as MRS clusters orCS clusters) and submit a job for verification.

Step 1 Create a DIS stream by referring to **Buying a Pay-Per-Use DIS Incremental Package** and continuously upload data to the newly created DIS stream. In this example, the content to be uploaded is **hello world**.

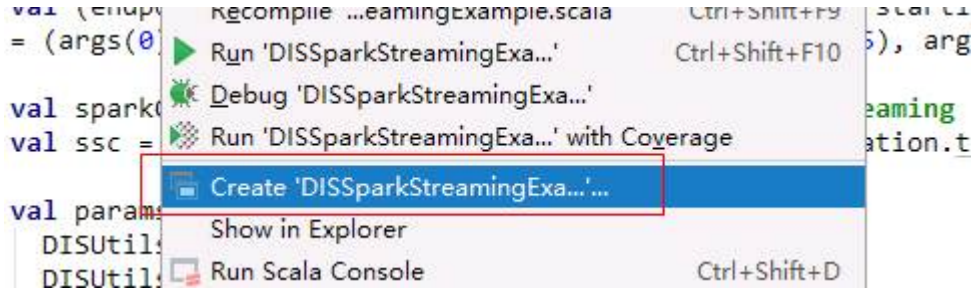
 Copyright © Huawei Technologies Co., Ltd.

Step 2 Open the **pom.xml** file, press **Ctrl+/** to comment out the **<scope>provided</scope>** row, and save the setting.

Step 3 Right-click **pom.xml** and choose **Maven** > **Reimport** from the shortcut menu to import the dependency package again.



Step 4 Right-click in the **DISSparkStreamingExample** file and choose **Create 'DISSparkStreamingExample'** from the shortcut menu.



Step 5 On the displayed configuration page, enter **-Dspark.master=local[*]** in **VM options**, indicating that a Spark job runs in local mode. Enter running parameters in **Program arguments** in the following format:

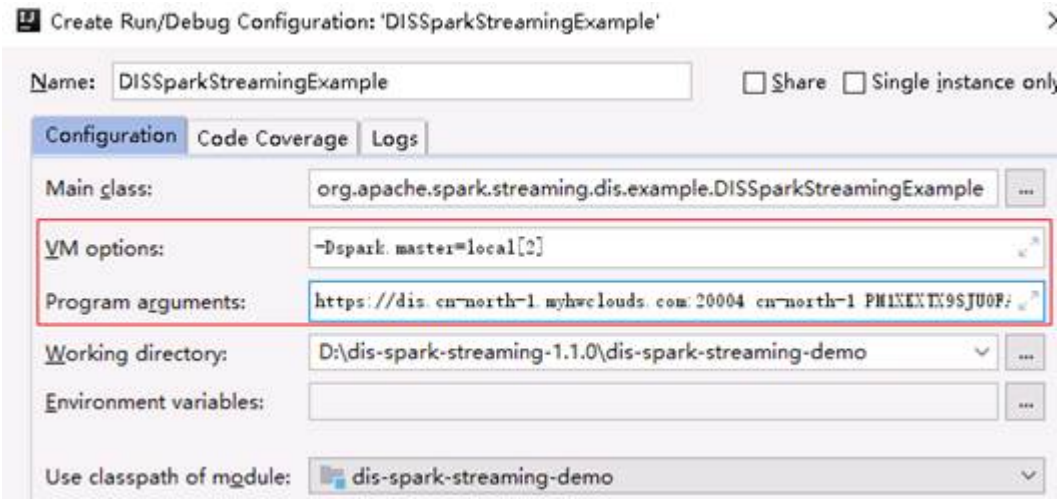
```
DIS gateway address Region name AK SK ProjectID Stream name Start position Streaming batch time

For example, if a test is performed in North China -Beijing1, the parameter example is as follows:
https://dis.${region}.myhuaweicloud.com ${region} YOU_AK YOU_SK YOU_PROJECTID YOU_STREAM_NAME latest 10
```

The parameter sequence and meaning are available in the sample code. For details, see the following information:

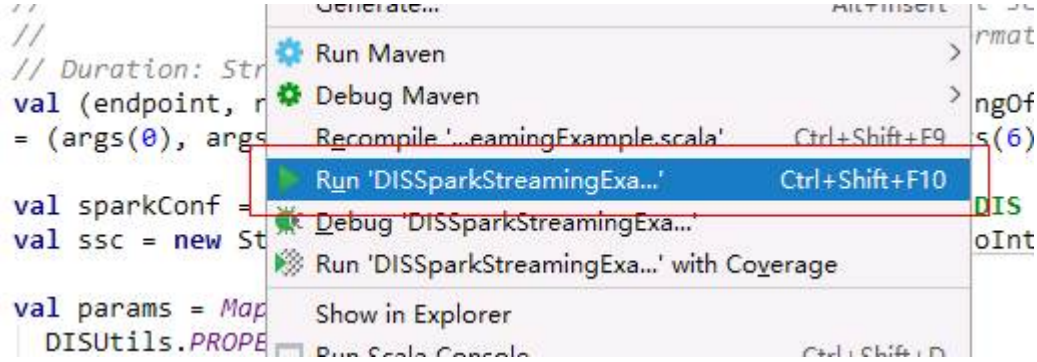
 Copyright © Huawei Technologies Co., Ltd.

```
// DIS GW url
// Region ID
// Access Key Id
// Secret Access Key
// User ProjectId
// DIS stream name
// Starting offsets:  'LATEST'    (Starting with the latest sequenceNumber)
//                    'EARLIEST'  (Starting with the earliest sequenceNumber)
//                    '{"0":23,"1":-1,"2":-2}'  (Use json format to specify the st
// Duration: StreamingContext duration(second)
val (endpoint, region, ak, sk, projectId, streamName, startingOffsets, duration)
= (args(0), args(1), args(2), args(3), args(4), args(5), args(6), args(7))
```

The final configurations of the IDEA are as the follow figure. After confirming that all configurations are correct, click **OK** to close the window.



Step 6 Right-click in the **DISSparkStreamingExample** file and choose **Run 'DISSparkStreamingExample'** from the shortcut menu to start the job.



Step 7 During the startup, an error message "hadoop binary path" is displayed, which can be ignored.

```
18/08/28 10:26:10 ERROR Shell: Failed to locate the winutils binary in the hadoop binary path
java.io.IOException: Could not locate executable null\bin\winutils.exe in the Hadoop binaries.
```

Step 8 If there are no other errors, the job runs a batch at a fixed interval, reads data from the batch, and exports the read result. The following gives an example:

```
-------------------------------------------
Time: 1535423650000 ms
-------------------------------------------
```

 Copyright © Huawei Technologies Co., Ltd.

```
(hello,30)
(world.,30)
```

Step 9 After verifying that the job can run locally without error, remove the comment tag from the <scope>provided</scope> row in **pom.xml** to prevent the Spark dependency from being packaged. Then stop the data upload program.

----End

6.2.10 Dumping Stream Data to Cloud Services

6.2.10.1 Creating a Dump Task

If a dump task is created for a DIS stream, data sent to the DIS stream can be automatically dumped to the selected target specified in the dump task.

1. Click the name of a stream that you want to view. On the displayed page, click the **Dump Management** tab. Alternatively, in the **Operation** column of a stream that you want to view, click **More** and choose **View Dump Task** from the drop-down list.
2. Click **Add Dump Task**. On the **Add Dump Task** page, configure dump parameters. Dump task parameters are described in **Dumping Data to OBS** to **Dumping Data to CloudTable**.

   NOTE

   A maximum of five dump tasks can be created for each stream.
3. Click **Create Now**.
4. In the **Operation** column of the corresponding **Task Name**, click **More** > **View Dump Log** to view the dump task details of the stream. **Table 6-98** describes the dump log parameters.

Table 6-98 Dump log parameters

Parameter	Description
Start Time	Time when the dump log is created. Format: YYYY/MM/dd HH:mm:ss GTM • YYYY: year. • MM: month. • dd: date. • HH: hour. • mm: minute. • ss: second. • GMT: time zone. For example: 2018/03/19 14:04:43 GMT +08:00

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
End Time	Time when you finish creating the dump log. Format: YYYY/MM/dd HH:mm:ss GTM ● YYYY: year. ● MM: month. ● dd: date. ● HH: hour. ● mm: minute. ● ss: second. ● GMT: time zone. For example: 2018/03/19 14:09:52 GMT +08:00
Status	Dump status. ● Succeeded ● Failed ● Abnormal
Read Records	Number of user records read from the stream.
Write Records	Number of records written from the stream to the target service, such as CloudTable. For example, a user record in a stream may be converted into multiple records and written to OpenTSDB of CloudTable.
Dump File Name	Name of the file that is dumped to the target service. The user records read from the stream are written into the file and then dumped to the target service (such as OBS, and MRS) in the file format.
Records	Number of the records uploaded between the time when you start to create a dump log to the time when you finish creating it.
Data Amount (bytes)	Amount of the data uploaded between the time when you start to create the dump log to the time when you finish creating it. Unit: byte

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Operation	Dump failure details. • If **Status** is **Succeeded**, the column is not operable. • If **Status** is **Failed**, click **View Details** to view dump details. • If **Status** is **Abnormal**, click **View Details** to view dump details.

Modifying a Dump Task

After creating a stream and adding a dump task successfully, you can modify the attributes of the created stream.

Step 1 **Log in to the DGC console**, select an instance, and click **Access** on it.

Click **Workspace Management** to create a workspace.

The workspace you created will show up in the list of workspaces.

Step 2 Click **Data Integration** of the corresponding workspace.

The **Data Integration** page is displayed.

Step 3 In the navigation pane, choose **Stream Management**.

Step 4 Click the name of a stream that you want to view. On the displayed page, click the **Dump Management** tab. Alternatively, in the **Operation** column of a stream that you want to view, click **More** and choose **View Dump Task** from the drop-down list.

Step 5 In the **Operation** column of the dump task that has been added to the stream, choose **More** > **Edit**. In the displayed dialog box, edit the dump task.

Step 6 Click **Submit**.

----End

6.2.10.2 Dumping Data to OBS

Source Data Type: JSON, BLOB, and CSV; Dump File Format: Text

Table 6-99 Parameters for configuring a Text dump file

Parameter	Description	Value
Task Name	Name of the dump task. The names of dump tasks created for the same stream must be unique. A dump task name is 1 to 64 characters long. Only letters, digits, hyphens (-), and underscores (_) are allowed.	N/A

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Value
Dump Bucket	Name of the OBS bucket used to store data from the DIS stream. The bucket name is created when you create a bucket in OBS.	N/A
File Directory	Directory created in OBS to store files from the DIS stream. This directory name is 0 to 50 characters long. By default, this parameter is left unspecified.	N/A
Time Directory Format	Data will be saved according to the time format in the file directory of the OBS bucket. For example, if the time directory is accurate to day, the save path will be in the format of bucket name/file directory/year/month/day. Possible values are as follows: • N/A: If this field is left unspecified, the time directory format will not be used. • yyyy: year. • yyyy/MM: year and month. • yyyy/MM/dd: year, month, and day. • yyyy/MM/dd/HH: year, month, day, and hour. • yyyy/MM/dd/HH/mm: year, month, day, hour, and minute. You can only select but not enter a value in this field.	N/A
Record Delimiter	Delimiter used to separate different dump records. Possible values are as follows: • Comma (,) • Semicolon (;) • Vertical bar (|) • Newline (\n) • NULL You can only select but not enter a value in this field.	N/A
Offset	• **Latest**: Maximum offset, indicating that the latest data will be read. • **Earliest**: Minimum offset, indicating that the earliest data will be read.	Latest

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Value
Dump Interval (s)	Interval at which data from the DIS stream will be imported into dump destination, such as OBS, MRS, DLI, DWS, and CloudTable. If no data was pushed to the DIS stream during the time specified here, the dump file will not be generated. Value range: 30s to 900s Unit: second Default value: 300s	N/A

Dumping the JSON and CSV Data to the CarbonData Data

Table 6-100 Parameters for configuring a CSV dump file

Parameter	Description	Value
Task Name	Name of the dump task. The names of dump tasks created for the same stream must be unique. A dump task name is 1 to 64 characters long. Only letters, digits, hyphens (-), and underscores (_) are allowed.	N/A
Dump Bucket	Name of the OBS bucket used to store data from the DIS stream. The bucket name is created when you create a bucket in OBS.	N/A
File Directory	Directory created in OBS to store files from the DIS stream. This directory name is 0 to 50 characters long. By default, this parameter is left unspecified.	N/A

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Value
Time Directory Format	Data will be saved according to the time format in the file directory of the OBS bucket. For example, if the time directory is accurate to day, the save path will be in the format of bucket name/file directory/year/month/day. Possible values are as follows: • N/A: If this field is left unspecified, the time directory format will not be used. • yyyy: year. • yyyy/MM: year and month. • yyyy/MM/dd: year, month, and day. • yyyy/MM/dd/HH: year, month, day, and hour. • yyyy/MM/dd/HH/mm: year, month, day, hour, and minute. You can only select but not enter a value in this field.	N/A
Offset	• **Latest**: Maximum offset, indicating that the latest data will be read. • **Earliest**: Minimum offset, indicating that the earliest data will be read.	Latest
Dump Interval (s)	User-defined interval at which data is imported from the current DIS stream into the target place. If no data is pushed to the DIS stream during the current interval, no dump file package will be generated. Value range: 30s to 900s Unit: second Default value: 300s	N/A

Source Data Type: JSON and CSV; Dump File Format: Parquet

NOTE

Table 6-101 lists the differentiated parameters that need to be set when the source data type is JSON or CSV, the dump destination is OBS, and the dump file format is Parquet. For details about how to configure other common parameters, see Table 6-99.

Copyright © Huawei Technologies Co., Ltd.

Table 6-101 Parameters for configuring a Parquet dump file

Parameter	Description	Value
Source Data Schema	JSON or CSV data sample, used to describe the JSON or CSV data format. DIS can generate an Avro schema based on the JSON or CSV data sample and convert the uploaded JSON or CSV data to the Parquet format.	N/A
File Directory	Directory created in OBS to store files from the DIS stream. This directory name is 0 to 100 characters long. By default, this parameter is left unspecified. **NOTE** When the source data type is JSON, EL and built-in functions are supported.	• **EL example** Source data: {"name":"Andy","city":"", "time":1556323141582} The dump file directory is configured as follows: dis/basePath/app_key_p=${name} The final directory structure is as follows: {$Bucket name}/dis/basePath/app_key_p=Andy • **Example of built-in functions** Source data: {"name":"Andy","city":"", "time":1556323141582} The dump file directory is configured as follows: dis/basePath/date_p=toDate(${time}, "yyMMdd") The final directory structure is as follows: {$Bucket name}/dis-basePath/date_p=20190427 **Supported built-in functions** toDate(timestamp, format): Convert the time stamp to a specified time format. For example: toDate (1556323141582, 'yymmdd')

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Value
Custom Time Directory	You can click or to disable or enable the custom time directory. • If the custom timestamp is disabled, the directory where the object file written to OBS resides is named after the creation time of the dump file. For example, if a dump file was created on October 16, 2018 and the time directory format is accurate to day, the file will be saved in *OBS bucket name*/*dump file directory*/**2018**/**10**/**16**. • If the custom timestamp is disabled, the directory where the object file written to OBS resides is named in the time format specified in the source data. For example, if a dump file was created on October 16, 2018, the time directory format is accurate to day, and the time format specified in the data source is 2017/09/08 11:01:01, the file will be saved in *OBS bucket name*/*dump file directory*/**2017**/**09**/**08**. The storage directory is defined based on the time format defined in the source data instead of the time when the dump file is created.	• Example 1: Dump simple JSON data. Source data: { "id":"1", "date":"2018/10/16 11:00:05"} The configuration is as follows: Set the timestamp attribute name to **date**, data type to **String**, and timestamp format to **yyyy/MM/dd HH:mm:ss** based on the source data type of the data to be dumped. After the data is dumped successfully, the storage directory structure depends on the source data timestamp and the time directory format. In this example, the time directory format is accurate to day. Therefore, the final data storage directory is *OBS bucket name*/*dump file directory*/**2018**/**10**/**16**. • Example 2: Dump multiply layers of nested JSON data. Source data: { "id":"1", "detail": { "detID":"057901100 00000000103#567fd3 cb13a4493eaa430769 53253eed", "endTime":"2018/10/0 7 13:26:35" }} The configuration is as follows: Set the timestamp attribute name to **detail.endTime**, data type to **String**, and timestamp format to **yyyy/MM/dd HH:mm:ss**

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Value
Source Data Timestamp	• Attribute name of the timestamp. **NOTE** Enter the field name corresponding to the timestamp defined in the source data to be uploaded. • Timestamp format. Possible values can be: yyyy/MM/dd HH:mm:ss MM/dd/yyyy HH:mm:ss dd/MM/yyyy HH:mm:ss yyyy-MM-dd HH:mm:ss MM-dd-yyyy HH:mm:ss dd-MM-yyyy HH:mm:ss • Data type. Possible values can be: – String – Timestamp **NOTE** If the type of the source data to be uploaded is **Timestamp**, the value must be accurate to milliseconds.	based on the source data type of the data to be dumped. After the data is dumped successfully, the storage directory structure depends on the source data timestamp and the time directory format. In this example, the time directory format is accurate to day. Therefore, the final data storage directory is *OBS bucket name*/*dump file directory*/**2018**/**10**/**07**. • Example 3: Dump CSV data. Source data: a,2010-10-12 11:00:00,b,2011-10-12 11:00:10 The configuration is as follows: Select timestamp **2010-10-12 11:00:00** based on the source data to be dumped. After data is converted to the Parquet format, the corresponding attribute field name is **field_1**. When creating a dump task, set the timestamp attribute to **field_1**, data type to **String**, and timestamp format to **yyyy/MM/dd HH:mm:ss**. After the data is dumped successfully, the storage directory structure depends on the source data timestamp and the time directory format. In this example, the time directory format is accurate to day. Therefore, the final data storage directory is *OBS*

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Value
		bucket name/*dump file directory*/**2010**/**10**/**12**.

Source Data Type: JSON and CSV; Dump File Format: CarbonData

NOTE

Table 6-102 lists the differentiated parameters that need to be set when the source data type is JSON or CSV, the dump destination is OBS, and the dump file format is CarbonData. For details about how to configure other common parameters, see Table 6-99.

Table 6-102 Parameters for configuring a CarbonData dump file

Parameter	Description	Value
Source Data Schema	JSON or CSV data example, used to describe the JSON or CSV data format. DIS can generate an Avro schema based on the JSON or CSV data sample and convert the uploaded JSON or CSV data to the CarbonData format.	N/A

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Value
CarbonData Retrieval Attribute	Attribute of the carbon table, used to create a carbon writer. The following keys are supported: • **table_blocksize**: Size of a table block. The value ranges from 1 MB to 2048 MB. The default value is 1024 MB. • **table_blocklet_size**: Size of the blocklet in a file. The default value is 64 MB. • local_dictionary_threshold • **local_dictionary_enable**: Possible values can be **true** or **false**. The default value is **false**. • **sort_columns**: Specifies the index column. Multi-level index columns are separated by commas (,). • **sort_scope**: Specifies the scope where data is sorted during loading. Currently, the following types are supported: – **local_sort**: Default value, indicating that data is sorted in a node. – **no_sort**: Data is not sorted. It is used when data needs to be saved to a database quickly. After the data is saved to the database, you can use the **Compaction** command to create an index when the system is idle. – **batch_sort**: A CarbonData file is generated after the memory is sorted in a node and no full sorting is performed on the node. This configuration improves the loading speed, but the query performance is inferior to that of **LOCAL_SORT**. • long_string_columns	N/A

 Copyright © Huawei Technologies Co., Ltd.

6.2.10.3 Dumping Data to DLI

Source Data Type: JSON and CSV

Table 6-103 Dump parameters

Parameter	Description	Value
Task Name	Name of the dump task. The names of dump tasks created for the same stream must be unique. A dump task name is 1 to 64 characters long. Only letters, digits, hyphens (-), and underscores (_) are allowed.	-
DLI Database	Click **Select**. In the **Select DLI Database** dialog box, select a database. You can only select but not enter a value in this field.	-
DLI Table	Click **Select**. In the **Select DLI Table** dialog box, select a data table. Only tables in DLI are supported. Before selecting a DLI table, ensure that you have the permission to insert data into the table. You can only select but not enter a value in this field.	This parameter is available only after you select a DLI database.
Offset	• **Latest**: Maximum offset, indicating that the latest data will be read. • **Earliest**: Minimum offset, indicating that the earliest data will be read.	Latest
Dump Interval (s)	Interval at which data from the DIS stream will be imported into dump destination, such as OBS, MRS, DLI, DWS, and CloudTable. If no data was pushed to the DIS stream during the time specified here, the dump file will not be generated. Value range: 30s to 900s Unit: second Default value: 300s	-
Temporary Bucket	OBS bucket in which a directory is created for temporarily storing user data. The data in the directory is deleted after being dumped to a specific destination.	-

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Value
Temporary Directory	Directory in the chosen **Temporary Bucket** for temporarily storing data. The data in the directory is deleted after being dumped to a specific destination. If this field is left blank, the data is stored directly to the **Temporary Bucket**.	-

6.2.10.4 Dumping Data to DWS

Source Data Type: JSON and CSV

Table 6-104 Dump parameters

Parameter	Description	Value
Task Name	Name of the dump task. The names of dump tasks created for the same stream must be unique. A dump task name is 1 to 64 characters long. Only letters, digits, hyphens (-), and underscores (_) are allowed.	N/A
DWS Cluster	Name of the DWS cluster that stores the data in the stream. Click **Select**. In the **Select DWS Cluster** dialog box, select a DWS cluster. You can only select but not enter a value in this field.	N/A
DWS Database	Name of the DWS database that stores the data in the stream. The value must be manually entered and cannot be left blank.	N/A
Database Schema	A database contains one or more named schemas, and schemas contain tables. Schemas also contain other named objects, including data types, functions, and operators. The same object name can be used in different modes without causing conflicts.	N/A
DWS Table	Name of the DWS table that stores the data in the stream.	N/A

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Value
Delimiter	Delimiter used to separate the columns in the DWS tables. This parameter cannot be left blank.	N/A
Offset	• **Latest**: Maximum offset, indicating that the latest data will be read. • **Earliest**: Minimum offset, indicating that the earliest data will be read.	Latest
Dump Interval (s)	Interval at which data from the DIS stream will be imported into dump destination, such as OBS, MRS, DLI, DWS, and CloudTable. If no data was pushed to the DIS stream during the time specified here, the dump file will not be generated. Value range: 30s to 900s Unit: second Default value: 300s	N/A
Username	Username for logging in to the DWS cluster.	N/A
Password	Password for logging in to the DWS cluster.	N/A
KMS Key	Database encryption key of the cluster	N/A
Temporary Bucket	OBS bucket in which a directory is created for temporarily storing user data. The data in the directory is deleted after being dumped to a specific destination.	N/A
Temporary Directory	Directory in the chosen **Temporary Bucket** for temporarily storing data. The data in the directory is deleted after being dumped to a specific destination. If this field is left blank, the data is stored directly to the **Temporary Bucket**.	N/A

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Value

Fault Processing	You can click or to disable or enable Fault Processing.
	• fill_missing_fields
	Specify whether to generate an error message when the last field in one row of data in the source file is missing during data import.
	Value: **true/on** or **false/off** Default value: **false/off**
	– **true/on**: The value of the last field is set to **NULL** and no error message is reported.
	– **false/off**: The following error message is reported:
	`missing data for column "tt"`
	• ignore_extra_data
	Specify whether to ignore the extra columns when the number of fields in the data source file is greater than the number of columns defined in the external table. This parameter is used only during data import.
	Value: **true/on** or **false/off** Default value: **false/off**
	– **true/on**: The extra columns at the end of a row are ignored.
	– **false/off**: The following error message is reported.
	`extra data after last expected column`
	NOTE
	If a linefeed at the end of a row is missing so that the row and another row are integrated into one, data in another row is ignored after the parameter is set to **true**.
	• compatible_illegal_chars
	Specify whether to tolerate invalid characters during data import. This parameter is valid only for READ ONLY foreign tables.
	Value: **true/on** or **false/off** Default value: **false/off**
	– **true/on**: Invalid characters are converted into valid ones before being imported to the database. No error message is reported and data import is not interrupted.
	– **false/off**: Data import is interrupted.
		N/A

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Value
	NOTICE In the Windows operating system, if OBS reads data files in text format, 0x1A is used as the EOF symbol to end data reading. As a result, a parsing error occurs. This is a restriction on the Windows platform. OBS can read BINARY data files on Linux only. **NOTE** • The rules for converting invalid characters are as follows: 1. \0 is converted to a space. 2. Others invalid characters are converted into question marks. 3. When **compatible_illegal_chars** is set to **true/on**, after invalid characters such as **NULL**, **DELIMITER**, **QUOTE**, and **ESCAPE** are converted to spaces or question marks, an error message stating "illegal chars conversion may confuse COPY escape 0x20" will be displayed to remind you of possible parameter confusion caused by the conversion. • **PER NODE REJECT LIMIT 'value'** Specify the number of data format errors allowed in each DN instance during data import. If the number of errors in a DN instance is greater than the preset value, the import fails and an error message is displayed. Value range: integer, unlimited. The default value is 0, indicating that error information is returned immediately. **NOTE** This parameter specifies the error tolerance of a single node. Data format errors refer to missing or redundant field values, data type errors, or coding errors. If a non-data format error occurs, the entire data import fails.	

 Copyright © Huawei Technologies Co., Ltd.

6.2.10.5 Dumping Data to MRS

Source Data Type: JSON, BLOB, and CSV; Dump File Format: Text

Table 6-105 Parameters for configuring a Text dump file

Parameter	Description	Value
Task Name	Name of the dump task. The names of dump tasks created for the same stream must be unique. A dump task name is 1 to 64 characters long. Only letters, digits, hyphens (-), and underscores (_) are allowed.	-
MRS Cluster	Click **Select**. In the **Select MRS Cluster** dialog box, select an MRS cluster. Data is dumped only to an MRS cluster that is not authenticated by Kerberos. You can only select but not enter a value in this field.	-
HDFS Path	Click **Select**. In the **Select HDFS Path** dialog box, select an HDFS path. You can only select but not enter a value in this field.	This parameter is available only after you select an MRS cluster. -
File Directory	Directory created in OBS to store files from the DIS stream. This directory name is 0 to 50 characters long. By default, this parameter is left unspecified.	-
Offset	• **Latest**: Maximum offset, indicating that the latest data will be read. • **Earliest**: Minimum offset, indicating that the earliest data will be read.	Latest
Dump Interval (s)	Interval at which data from the DIS stream will be imported into dump destination, such as OBS, MRS, DLI, DWS, and CloudTable. If no data was pushed to the DIS stream during the time specified here, the dump file will not be generated. Value range: 30s to 900s Unit: second Default value: 300s	-

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Value
Temporary Bucket	OBS bucket in which a directory is created for temporarily storing user data. The data in the directory is deleted after being dumped to a specific destination.	-
Temporary Directory	Directory in the chosen **Temporary Bucket** for temporarily storing data. The data in the directory is deleted after being dumped to a specific destination. If this field is left blank, the data is stored directly to the **Temporary Bucket**.	-

Source Data Type: JSON and CSV; Dump File Format: Parquet

NOTE

Table 6-106 lists the differentiated parameters that need to be set when the source data type is JSON or CSV, the dump destination is MRS, and the dump file format is Parquet. For details about how to configure other common parameters, see **Table 6-105**.

Table 6-106 Parameters for configuring a Parquet dump file

Parameter	Description	Value
Source Data Schema	JSON or CSV data example, used to describe the JSON or CSV data format. DIS can generate an Avro schema based on the JSON or CSV data sample and convert the uploaded JSON or CSV data to the Parquet format.	-

Source Data Type: JSON and CSV; Dump File Format: CarbonData

NOTE

Table 6-107 lists the differentiated parameters that need to be set when the source data type is JSON or CSV, the dump destination is MRS, and the dump file format is CarbonData. For details about how to configure other common parameters, see **Table 6-105**.

Copyright © Huawei Technologies Co., Ltd.

Table 6-107 Parameters for configuring a CarbonData dump file

Parameter	Description	Value
Source Data Schema	JSON or CSV data example, used to describe the JSON or CSV data format. DIS can generate an Avro schema based on the JSON or CSV data sample and convert the uploaded JSON or CSV data to the CarbonData format.	-

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Value
CarbonData Retrieval Attribute	Attribute of the carbon table, used to create a carbon writer. The following keys are supported: • **table_blocksize**: Size of a table block. The value ranges from 1 MB to 2048 MB. The default value is 1024 MB. • **table_blocklet_size**: Size of the blocklet in a file. The default value is 64 MB. • local_dictionary_threshold • **local_dictionary_enable**: Possible values can be **true** or **false**. The default value is **false**. • **sort_columns**: Specifies the index column. Multi-level index columns are separated by commas (,). • **sort_scope**: Specifies the scope where data is sorted during loading. Currently, the following types are supported: – **local_sort**: Default value, indicating that data is sorted in a node. – **no_sort**: Data is not sorted. It is used when data needs to be saved to a database quickly. After the data is saved to the database, you can use the **Compaction** command to create an index when the system is idle. – **batch_sort**: A CarbonData file is generated after the memory is sorted in a node and no full sorting is performed on the node. This configuration improves the loading speed, but the query performance is inferior to that of **LOCAL_SORT**. • long_string_columns	-

 Copyright © Huawei Technologies Co., Ltd.

6.2.10.6 Dumping Data to CloudTable

Source Data Type: JSON and CSV; Dump Destination: CloudTable HBase

Table 6-108 Dump parameters

Parameter	Description	Value
Task Name	Name of the dump task. The names of dump tasks created for the same stream must be unique. A dump task name is 1 to 64 characters long. Only letters, digits, hyphens (-), and underscores (_) are allowed.	N/A
Source Data Schema	JSON or CSV data sample, used to describe the JSON or CSV data format. DIS can generate an Avro schema based on the JSON or CSV data sample.	N/A
CloudTable Cluster	Click **Select**. In the **Select CloudTable Cluster** dialog box, select a CloudTable cluster. This parameter cannot be left blank. You can only select but not enter a value in this field.	N/A
Table Type	Two CloudTable types are available: **HBase** and **OpenTSDB**.	N/A
Table	Click **Select**. In the **Select CloudTable Table** dialog box, select a CloudTable table. You can only select but not enter a value in this field.	This parameter is available only after you select a CloudTable cluster and create a HBase table.
Offset	• **Latest**: Maximum offset, indicating that the latest data will be read. • **Earliest**: Minimum offset, indicating that the earliest data will be read.	Latest
Backup	Specifies whether to back up the data that fails to be dumped to CloudTable to OBS. • Enable: The data will be backed up to OBS. • Disable: The data will not be backed up to OBS. Backup is disabled by default.	If this function is disabled, data that fails the dump will be stored in DIS and be cleared when **Data Retention** is expired.

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Value
Backup Bucket	OSB bucket that stores the dump failure records. You can only select but not enter a value in this field.	N/A
Backup Directory	Directory used to store dump failure records. If this parameter is left empty, the dump failure records are stored in the backup bucket. The value consists only of letters, digits, underscores (_), and slashes (/) and cannot start with a slash. The value contains a maximum of 50 characters.	N/A
Row Key	• JSON attribute name. The value contains a maximum of 32 characters and consists only of letters, digits, underscores (_), and periods (.). The value cannot start or end with a period, contain consecutive periods, or be left unspecified. A maximum of 64 attributes can be added. • Data type. Possible values can be: – Bigint – Double – Boolean – Timestamp – String – Decimal	N/A
Row Key Delimiter	Delimiter used to separate row keys. The value can only be a period (**.**), comma (**,**), vertical bar (**|**), semicolon (**;**), hyphen (**-**), underscore (_), or tilde (**~**). It can also be set to **NULL**. The maximum length is one character.	N/A

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Value
Schema Column	• Column name. The value contains a maximum of 32 characters and consists only of letters, digits, and underscores (_). The value cannot be left unspecified. A maximum of 4096 columns can be added. • Data type. Possible values can be: – Bigint – Double – Boolean – Timestamp – String – Decimal • JSON attribute name. The value contains a maximum of 32 characters and consists only of letters, digits, underscores (_), and periods (.). The value cannot start or end with a period, contain consecutive periods, or be left unspecified. • Column family. The value can only be selected from the drop-down list box and cannot be left unspecified. This parameter is available only after you select a CloudTable cluster and set **Type** of a CloudTable table to **HBase**.	N/A

Source Data Type: JSON and CSV; Dump Destination: CloudTable openTSDB

Table 6-109 Dump parameters

Parameter	Description	Value
Task Name	Name of the dump task. The names of dump tasks created for the same stream must be unique. A dump task name is 1 to 64 characters long. Only letters, digits, hyphens (-), and underscores (_) are allowed.	N/A
Source Data Schema	JSON or CSV data example, used to describe the JSON or CSV data format. DIS can generate an Avro schema based on the JSON or CSV data sample.	N/A

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Value
CloudTable Cluster	Click **Select**. In the **Select CloudTable Cluster** dialog box, select a CloudTable cluster. This parameter cannot be left blank. You can only select but not enter a value in this field.	N/A
Table Type	Two CloudTable types are available: **HBase** and **OpenTSDB**.	N/A
Table	Click **Select**. In the **Select CloudTable Table** dialog box, select a CloudTable table. You can only select but not enter a value in this field.	This parameter is available only after you select a CloudTable cluster and create an openTSDB table.
Offset	• **Latest**: Maximum offset, indicating that the latest data will be read. • **Earliest**: Minimum offset, indicating that the earliest data will be read.	Latest
Backup	Specifies whether to back up the data that fails to be dumped to CloudTable to OBS. • Enable: The data will be backed up to OBS. • Disable: The data will not be backed up to OBS. Backup is disabled by default.	If this function is disabled, data that fails the dump will be stored in DIS and be cleared when **Data Retention** is expired.
Backup Bucket	OSB bucket that stores the dump failure records. You can only select but not enter a value in this field.	N/A
Backup Directory	Directory used to store dump failure records. If this parameter is left empty, the dump failure records are stored in the backup bucket. The value consists only of letters, digits, underscores (_), and slashes (/) and cannot start with a slash. The value contains a maximum of 50 characters.	N/A

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Value
Metric: Schema configuration of the OpenTSDB data metric in the CloudTable cluster. After this parameter is set, the JSON data in the stream can be converted to the metric of the OpenTSDB data.	• JSON attribute name. The value contains a maximum of 32 characters and consists only of letters, digits, underscores (_), and periods (.). The value cannot start or end with a period, contain consecutive periods, or be left unspecified. • Data type. Possible values can be: – String – Constant	N/A
Timestamp: Schema configuration of the OpenTSDB data timestamp in the CloudTable cluster. After this parameter is set, the JSON data in the stream can be converted to the timestamp of the OpenTSDB data.	• JSON attribute name. The value contains a maximum of 32 characters and consists only of letters, digits, underscores (_), and periods (.). The value cannot start or end with a period, contain consecutive periods, or be left unspecified. • Timestamp format. Possible values can be: yyyy/MM/dd HH:mm:ss MM/dd/yyyy HH:mm:ss dd/MM/yyyy HH:mm:ss yyyy-MM-dd HH:mm:ss MM-dd-yyyy HH:mm:ss dd-MM-yyyy HH:mm:ss • Data type. Possible values can be: – String – Timestamp	N/A

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Value
Value: Schema configuration of the OpenTSDB data value in the CloudTable cluster. After this parameter is set, the JSON data in the stream can be converted to the value of the OpenTSDB data.	• JSON attribute name. The value contains a maximum of 32 characters and consists only of letters, digits, underscores (_), and periods (.). The value cannot start or end with a period, contain consecutive periods, or be left unspecified. • Data type. Possible values can be: – Bigint – Double – Boolean – Timestamp – String – Decimal	N/A
Tags: Schema configuration of the OpenTSDB data tags in the CloudTable cluster. After this parameter is set, the JSON data in the stream can be converted to the tags of the OpenTSDB data.	• Tag name. The value contains a maximum of 32 characters and consists only of letters, digits, and underscores (_). The value cannot be left unspecified. • Data type. Possible values can be: – Bigint – Double – Boolean – Timestamp – String – Decimal – Constant • JSON attribute name. The value contains a maximum of 32 characters and consists only of letters, digits, underscores (_), and periods (.). The value cannot start or end with a period, contain consecutive periods, or be left unspecified.	N/A

6.2.11 Event Notification

6.2.11.1 Event Notification Overview

Overview

DIS uses Simple Message Notification (SMN) to send notifications of DIS events. In a subscription, you need to specify one or more event filtering conditions. When an event that matches all filtering conditions occurs, DIS sends a notification based on the subscription. The filter conditions include the **Event Type** (for

 Copyright © Huawei Technologies Co., Ltd.

example, **Management**, **Monitoring**, or **Security**), **Event Level** (for example, **Normal** or **Warning**), and **Event Source** (for example, **Stream**, **Dump task**, or **User**).

Supported Event Types and Events

Events are records of changes in the tenant's stream status. It can be triggered by a user operation (for example, an audit event), or may be caused by a stream status change (for example, a dump task is abnormal or a dump task is recovered). The following tables list the events and event types supported by DIS:

- The following table lists the events whose **Event Source** is **Stream**.

  Table 6-110 Events whose **Event Source** is **Stream**

Event Source	Event Level	Event
Stream	Warning	Traffic limited
Stream	Warning	Automatic stream scaling succeeded
Stream	Warning	Automatic stream scaling failed
Stream	Warning	Stream traffic abnormal
Stream	Warning	Stream traffic restored

- The following table lists the events whose **Event Source** is **Event**.

  Table 6-111 Events whose **Event Source** is **User**

Event Source	Event Level	Event
User	Warning	Quota insufficient

- The following table lists the events whose **Event Source** is **Dump task**.

  Table 6-112 Events whose **Event Source** is **Dump task**

Event Source	Event Level	Event
Dump task	Normal	Dump task restored
Dump task	Warning	Dump task abnormal

6.2.11.2 Subscribing to Event Notification

After enabling **Notification** for subscribed events, you will receive notifications by email or SMS when management, monitoring, or security events occur in a specific cluster or dump task.

 Copyright © Huawei Technologies Co., Ltd.

Creating a Subscription

Step 1 Log in to the DGC console, select an instance, and click **Access** on it.

Click **Workspace Management** to create a workspace.

The workspace you created will show up in the list of workspaces.

Step 2 Click **Data Integration** of the corresponding workspace.

The **Data Integration** page is displayed.

Step 3 In the navigation pane, choose **Stream Management**.

Step 4 Choose **Event Management -Subscriptions** > **Create Subscription**.

Step 5 In the **Subscription Settings** area, set basic subscription information and event filtering.

The **Subscribed Event List** area displays the events filtered by the system based on the subscription settings.

Table 6-113 Subscription parameters

Parameter	Description
Notification	Enable or disable event subscription. indicates that notification is enabled and indicates that notification is disabled. By default, notification is disabled. After notification is disabled, the system stops sending notifications of subscribed events but does not delete the subscription.
Subscription Name	Enter the name of a subscription. • The name can contain letters (upper or lower case), digits, hyphens (-), and underscores (_) and must start with a letter or digit. • The name must be between 1 and 256 characters in length.
Subscription Stream	Specifies whether to enable subscription to alarms of a specified stream. indicates that subscription to alarms of a specified stream is enabled. indicates that subscription to alarms of a specified channel is disabled. By default, the function is disabled. After this function is enabled, the alarms sent from your specified stream are received, and sent from other streams will not be received.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Subscription Type	Two subscription types are supported: SMN and DIS. **NOTE** • If **Subscription Type** is set to **SMN**, go to **Step 7** to select an SMN topic. • If **Subscription Type** is set to **DIS**, go to **Step 6** to select a stream.

Step 6 Select a DIS stream.

Step 7 Select a message notification topic from the **SMN Topic** drop-down list.

You can perform the following operations to create a message notification topic as required.

1. Click **Create SMN Topic**. The **Topics** page of SMN is displayed. You can click **Create Topic** in the upper right corner to create a topic. For details, see section **Creating a Topic** in the *Simple Message Notification User Guide*.
2. In the **Operation** column of a topic, choose **More** > **Configure Topic Policy** and select **DIS** under **Services that can publish messages to this topic** to enable SMN to publish DIS topics.
3. In the row containing the created topic, click **Add Subscription** to add a subscription to the topic. For details, see **Adding a Subscription** in the *Simple Message Notification User Guide*.

Step 8 Click **Submit**.

----End

Modifying the Subscription

Step 1 On the **Event Management** page, click **Subscription**.

Step 2 In the **Operation** column of a specified subscription name, click **More** > **Modify**.

Step 3 On the **Subscription Settings** page, set the parameters that you want to modify. For details, see **Step 5** to **Step 8** in section **Creating a Subscription**.

----End

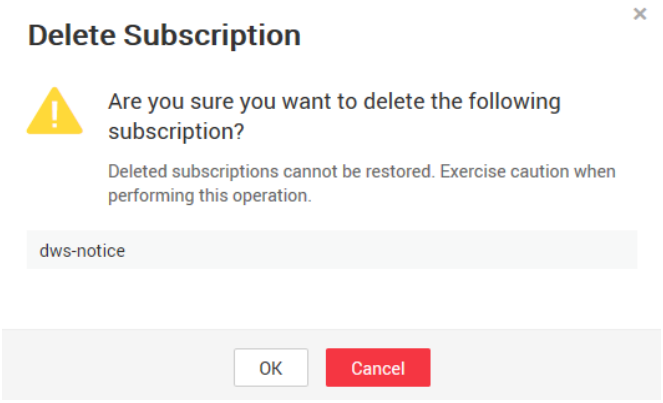
Deleting the Subscription

Step 1 On the **Event Management** page, click **Subscription**.

Step 2 In the **Operation** column of a specified subscription name, click **More** > **Delete**. The **Delete Subscription** dialog box shown in **Figure 6-150** is displayed.

 Copyright © Huawei Technologies Co., Ltd.

Figure 6-150 Confirming deletion



Step 3 Click **OK**.

----End

6.2.11.3 Viewing Events

This section describes how to search for events that occur in a stream or dump task.

Step 1 In the navigation tree, choose **Event Management** > **Events**. All events are displayed by default.

Step 2 Select different filter criteria from the drop-down list in the upper right corner of the event list to search for events. Events can be filtered based on the event level or event source.

- In the **Event Levels** drop-down list, select **All**, **Normal**, or **Warning**
- In the **Event source** drop-down list, select **Event source**. In the search box, enter a stream name or a dump task name, for example, **demo**.

Step 3 Click . The filtered event query results are displayed.

Step 4 Click on the right of **Event** and select an event name, for example, **Dump task restored** to filter the corresponding events.

----End

 Copyright © Huawei Technologies Co., Ltd.

7 Data Design

7.1 Data Design Introduction

Overview

Use DGC Data Design to create 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.

Data Design is designed for processing and commercializing data. Data Design consists of four parts: data survey, standard design, model design, and metric design.

Data Design is used for building:

- a unified data classification system to manage all business data in directories for easier data classification, search, evaluation, and use.
- a unified data standard system that conforms to country-level 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 standard 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.

Data Design Process

The figure below describes the process of data design.

- **Data survey** is used to sort business requirements, create business processes, and classify data subjects based on the existing business data and industry status.
- **Business subject** is a hierarchical architecture that classifies and defines data to help clarify data assets and specify relationships between business domains and business objects.

 Copyright © Huawei Technologies Co., Ltd.

  - **Subject area group** is used to group subject areas based on business scenarios.
  - **Subject area** is a high-level data classification that does not overlap and used to manage business objects.
  - **Business object** refers to the important information about people, events, and things that are indispensable to enterprise operations and management.
- **Business process design** is used to generate a structured framework of process. It describes the categories, levels, boundaries, scope, and input/output relationships of an enterprise's processes, and reflects the business models and characteristics of the enterprise.
- **Standard design**
  - **Data standard** refers to the description of attribute data meanings and business rules that must be complied with by enterprises. It describes the common understanding of certain data at the company level. Once the understanding is determined, it should be complied with inside an enterprise.
  - **Lookup table** contains a series of allowed values and additional text descriptions. The allowed values and additional text descriptions are generally associated with data standards to generate a range of values to verify the quality monitoring rules.
- **Model design** refers to designing data analysis models.
  - **Data analysis model** is defined for a specific business scenario to explore and analyze data to reveal the causalities, correlations, internal relationships, and service rules. A data analysis model often uses technologies such as statistics or data mining, machine learning, and AI technologies to solve specific business problems.
  - **Source Data Integration (SDI)** is called the source data layer. SDI is a simple implementation of source system data.
  - **Data Warehouse Integration (DWI)** is called the data integration layer. This layer integrates and cleans data from multiple source systems, and implements entity relationship modeling based on the three normal forms.
  - **Data Warehouse Report (DWR)** is called the data report layer. DWR is based on the multi-dimensional model and its data granularity is the same as that of the DWI layer.
- **Metric design**
  - **Business 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 meaning 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.
  - **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.

 Copyright © Huawei Technologies Co., Ltd.

  - **Modifier** is the abstract definition of a business scenario, which is used to determine the measurement scope and used as the WHERE condition in SQL statements.
  - **Atomic metric** refers to the measures and attributes of atomic metrics that come from the dimension tables and fact tables of the multidimensional models. The business objects and the finest data granularity of atomic metrics are consistent with those of the multidimensional models.

    An atomic metric consists of measures and attributes related with measures and business objects, all of which support agile self-service consumption of the metric.
  - **Derivative metric** is aggregated from the definitions, modifiers, and dimensions of atomic metrics. Therefore, their definitions, modifiers, and dimensions are derived from the attributes of atomic metrics.
  - **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.

    No new dimensions and modifiers for the compound metrics generated if the derivative metrics have no dimensions and modifiers.
- Data Mart (DM) is also called data mall. Multiple types of data are summarized at the DM layer. DM is designed to display the summarized data.

Model Design Method Overview

A data model can reflect the relationships between business 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 business scenarios, be easily understood, and be easily implemented in the IT system.

Entity relationship (ER) and dimensional modeling are both used on DGC Data Design:

- **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 a conceptual model can be 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 compared with 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.

 Copyright © Huawei Technologies Co., Ltd.

  - **Physical model** is an advanced version of the logical model. It is used to design the database architecture for data storage with a range of technical factors considered if the selected data warehouse is DWSor DLI.
- **Dimensional modeling**

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

  A multidimensional model is a fact table 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 Data Design, you can construct a bus matrix, a dimension model and a fact model by extracting facts and dimensions; and a summary model incorporating metrics abstracted from BI analysis.

Highlights

- Use Huawei's mature business classification methodology to effectively manage all business assets.
- Manage asset classifications and models in Data Assets.

7.2 Overview

Logging In to the Data Design Console

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Design** of the corresponding workspace to access the Data Design console.

Data Design Overview Page

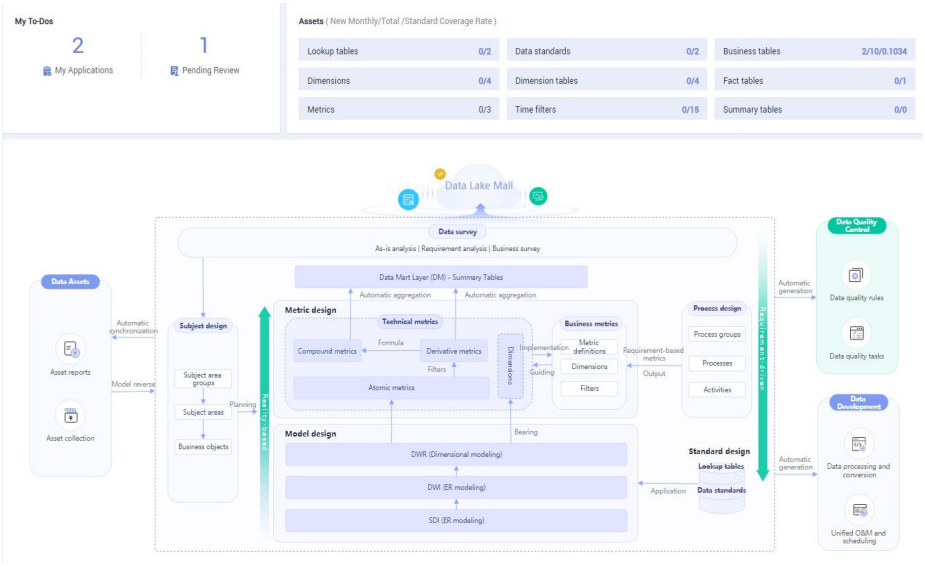
Go to the **Overview** page of the DGC Data Design. See **Figure 7-1** for details.

- **My To-Dos**
  - The **My To-Dos** area displays the quantity of **My Requests** and **Pending Approval**.
  - Click the statistics above **My Requests** and **Pending Approval** and switch to the **My Applications** and **Pending Review** pages, respectively.
- **Assets**
  - The **Assets** area displays the number of monthly new objects, total number of objects, and the coverage rate of standards.
  - Click the number next to each object name to go to the object management page.
- **Data Design Process**

 Copyright © Huawei Technologies Co., Ltd.

- **Data Design Process** displays the data design process and relationships with other DGC modules.
- Move your pointer to any object name in the flowchart. The object description is displayed on the page.
- You can click the name of any object supported by DGC go to the object management page.

Figure 7-1 Data design overview



7.3 Adding Reviewers

In the Data Design module, all business processes need to be approved. Therefore, add reviewers first before conducting any operations. Only the workspace administrator has the permissions required to add reviewers.

Adding a Reviewer

The reviewer must be a member who has the review permission in the current workspace. You need to edit and add workspace members in **Workspaces** on the DGC homepage.

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Design** of the corresponding workspace to access the Data Design console.
3. Choose **Configuration Center** from the left navigation bar. On the displayed page, choose **Reviewer Management**.
4. On the **Reviewer Management** page, click **Add**.
5. Select a reviewer, enter their mobile number and email address, and click **OK**.

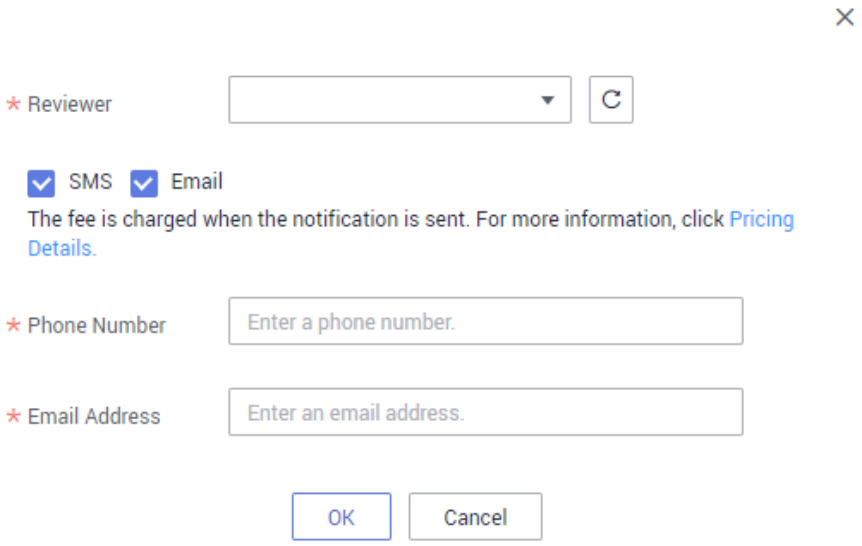
   The reviewer must be a member with the review permissions for the current workspace. Only administrators and developers have the review permissions.

 Copyright © Huawei Technologies Co., Ltd.

NOTE

The reviewer cannot be manually added. You need to edit and add workspace members in **Workspaces** on the DGC homepage.

Figure 7-2 Adding a reviewer

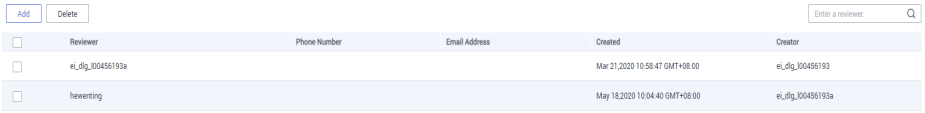


6. You can add multiple reviewers if needed.

Managing a Reviewer

On the Data Design console, choose **Configuration Center** > **Reviewer Management** from the left navigation bar. On the displayed page, you can manage reviewers as required.

Figure 7-3 Reviewer Management page



- **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, locate the reviewer you want to delete, select the reviewer, and click **Delete**.

7.4 Information Architectures

An information architecture is a set of component standardizations that describes the various types of information required for business operations and management decision making as well as the relationships of business entities. On

 Copyright © Huawei Technologies Co., Ltd.

the **Information Architecture** page, you can view and manage business tables, dimension tables, fact tables, and summary tables.

Overview

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Design** of the corresponding workspace to access the Data Design console.
3. Choose **Information Architecture** from the left navigation bar.
4. On the **Information Architecture** page, you can perform the following operations:
   - **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 specified table. Then click the table name to go to its **Basic Information** page.

     After finding a specified table, you can perform any of the following operations:

     - Edit
     - Publish
     - More > Suspend
     - More > Preview SQL
     - More > View Publish History
     - More > Modify Subject
     - More > Import
     - More > Export
     - More > Synchronize
     - More > Delete
   - **Create**

     Click **Create** to create a logical model, physical model, fact table, or summary table. For details, see **Designing Physical Models**, **Creating Fact Tables**, or **Creating Summary Tables**.
   - **Import**

     Choose **More** > **Import** to import a business table. Download the template, fill it in and upload it, and click **OK**. For more information on importing business tables, see **Importing/Exporting Tables**.
   - **Export**

 Copyright © Huawei Technologies Co., Ltd.

Choose **More** > **Export** to export a business table model or DDL. For more information on exporting business tables or DDLs, see **Exporting a Table/DDL**.

- **Synchronize**

  Choose **More** > **Synchronize** to synchronize the table information to Data Assets as the technical assets and logical models as business 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 with the pending publishing, published, or pending suspension status cannot be deleted. The referenced data table cannot be deleted either.

- **Suspend**

  Choose **More** > **Suspend** to take a published data table offline. The referenced data table cannot be brought offline.

  NOTE

  Data with the drop-down arrow on the left has been re-edited.

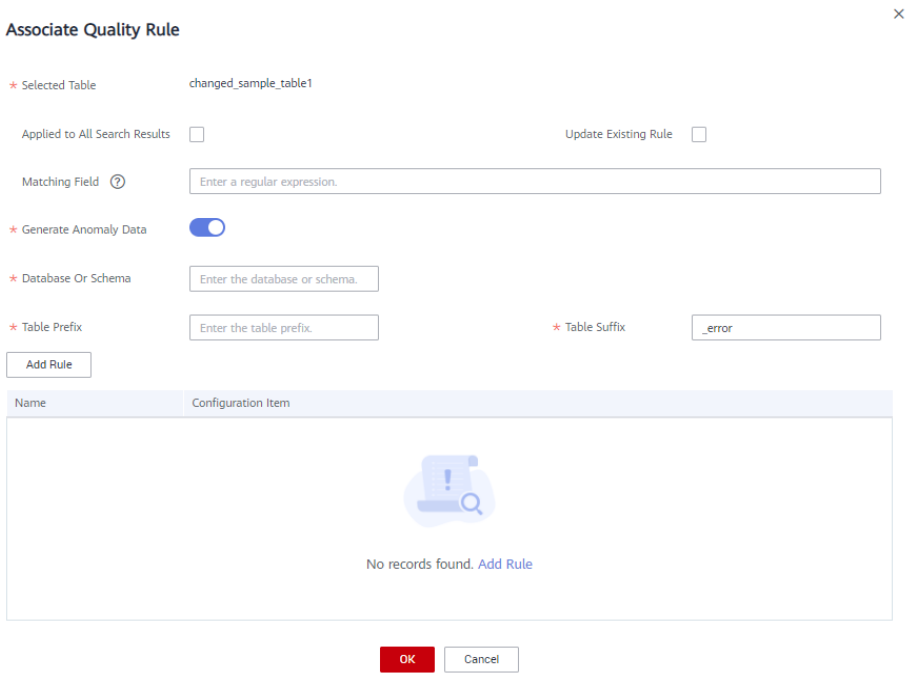
- **Publish**

  Click **Publish** to publish the data table. Data tables with the pending publishing, pending suspension, or published (without edited versions) status cannot be published.

- **Associate Quality Rule**

  Click **Associate Quality Rule** and set the parameters as prompted to associate a quality rule with the object you select. For more information on associated quality rules, see **Associating with Quality Rules**.

Figure 7-4 Associating quality rules with the object you select



 Copyright © Huawei Technologies Co., Ltd.

Generate Anomaly Data: If this option is selected, anomaly data is stored in the specified database based on the configured parameters.

7.5 Designing Processes

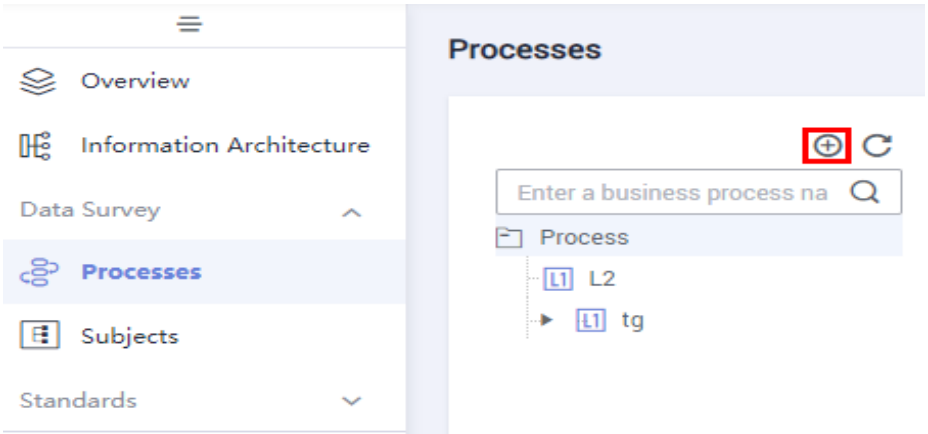
Business Process Architecture (BPA) is developed based on the value streams, and is used to guide and standardize the management of BT&IT requirements and ensure the efficiency of business requirement handling, analysis, and delivery. The BPA puts high-value requirements first, which maximizes the business value, assists in business operations, and facilitates goal achievement.

Creating a Business Process

Design the process from L1 to L 3 based on service requirements.

1. **Log in to the DGC console**, select an instance, and click **Access** on it.
   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Design** of the corresponding workspace to access the Data Design console.
3. On the Data Design console, choose **Processes** from the left navigation bar.
   Click ⊕ to create a process. When creating a process for the first time, you can create a process under the root node.

   Figure 7-5 Process design



4. In the displayed dialog box, set parameters as prompted and click **OK**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 7-6 Creating a process

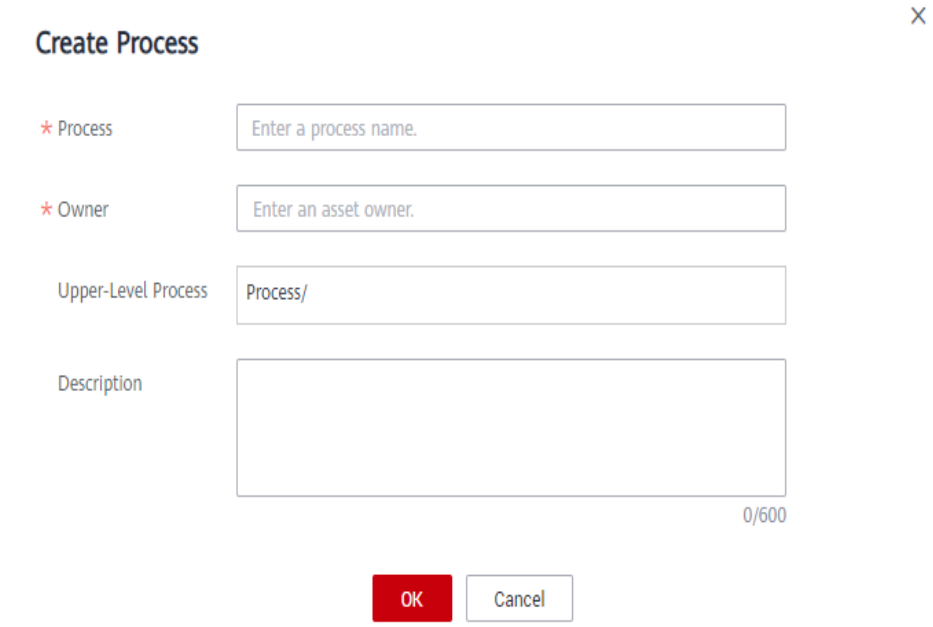


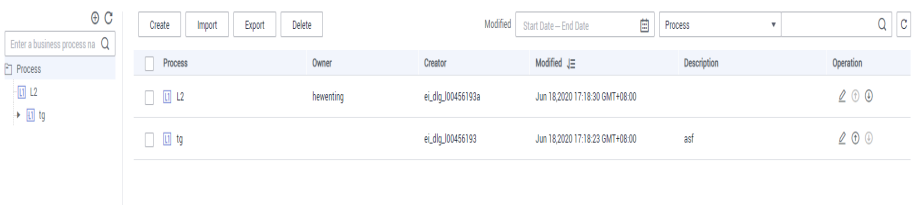
Table 7-1 Parameters for creating processes

Parameter	Description
Process	Process name. Only letters, digits, and underscores (_) are allowed.
Owner	Process owner. Only the workspace administrator or developer can be set as the owner.
Select Process	Home process
Description	A description of the process

5. Repeat the above steps in sequence to create more processes or subprocesses. Generally, processes from L1 to L3 need to be designed. The first layer is identified as L1, the second layer as L2, and the third layer as L3.

   The following is an example:

Figure 7-7 Process design example



 Copyright © Huawei Technologies Co., Ltd.

Exporting a Process

You can export the processes that have been created in Data Design to files.

Step 1 On the Data Design console, choose **Processes** from the left navigation bar.

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 successfully.

NOTE

A subjector process has a hierarchy. You can export only data of all levels.

----End

Importing a Process

Step 1 On the Data Design console, choose **Processes** from the left navigation bar.

Step 2 Click **Import** above the process list.

Step 3 In the **Import Process** dialog box, set the parameters as prompt, and click **Upload and Verify**.

Table 7-2 Parameters for importing processes

Parameter	Description
Update Existing Data	Whether to update the existing processes of DGC Data Design. The options are as follows: • **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. During the import, only process creation and update are allowed.
Upload File	Select the file you want to import. You can use either of the following methods to obtain the file to be imported: • **Downloading the process template and fill in it** In the **Import Table** dialog box, click **Process Template** to download the template, set parameters in the template as required, and save the settings. See **Table 7-3** for template parameter details. • **Exporting a pocess** You can export the processes created in Data Design of a DGC instance to Excel files. Then, import the Excel files. For details, see **Exporting a Process**.

Table 7-3 describes the parameters in the downloaded template. Parameters whose names start with an asterisk (*) are mandatory, and parameters whose

 Copyright © Huawei Technologies Co., Ltd.

names do not start with an asterisk (*) are optional. One record is required for one process.

Table 7-3 Parameters in the process import template

Parameter	Description
Process	If it is a level 1 process, this field can be left blank. If it is not, this field is mandatory. If there are multiple processes, separate them with slashes (/), for example, "Integrated Product Development/Development Lifecycle".
*Name	Process name
*Owner	Process owner Only the workspace administrator or developer can be set as the owner.
Description	A description of the process

Step 4 The import result is displayed in **Last Import** 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 business process has subdirectories or subprocesses, you need to delete the subdirectories or subprocesses progressively.

Step 1 On the Data Design console, choose **Processes** from the left navigation bar.

Step 2 In the process list, select the process to be deleted and click **Delete** above the process list.

Step 3 In the dialog box that is displayed, click **Yes**.

----End

7.6 Designing Subjects

A business subject is a hierarchical architecture that classifies and defines data to help clarify data assets and specify relationships between business domains 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.

 Copyright © Huawei Technologies Co., Ltd.

- You can download the provided subject design template, set the parameters, and use the template to import subjects in batches.
- You can export the subjects created in Data Design of a DGC instance to Excel files. Then, import the Excel files. 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**.

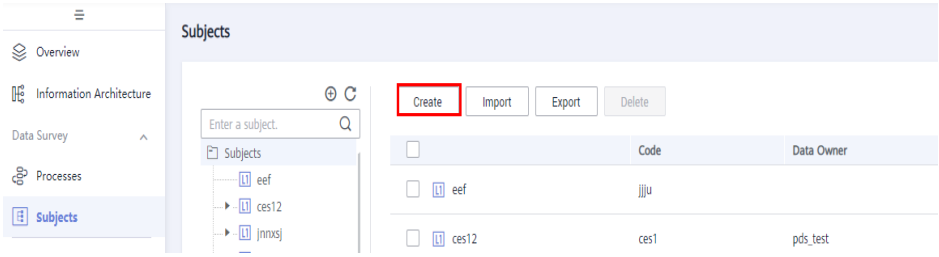
Creating a Subject

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Design** of the corresponding workspace to access the Data Design console.
3. Click **Subjects** from the left navigation bar. On the displayed page, click **Create** in the upper left corner.

   Figure 7-8 Designing a subject



4. In the displayed dialog box, set the parameters described in the table below and click **OK**.

   You can create multiple sub-subjects under a subject. The number of subject levels is defined by users in the configuration center. For details, see **Suject Levels**. 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.

   Table 7-4 Parameters

Parameter	Description
Name	The following characters are not allowed: / \ < >.
Code	The code of the subject area group to be created. The code can include letters, numbers, spaces, colons (:), underscores (_), hyphens (-), parentheses, slashes (/), and ampersands (&) and cannot start with a space, <, >, or /.
Alias	The following characters are not allowed: \ < >. **NOTE** Before configuring an alias, go to the **Configuration Center** > **Model Settings**, and set **Use Alias** to **Subjects**.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Subject	Select a created subject.
Data Owner's Department	The department the data owner belongs to.
Data Owner	Select a data owner from the drop-down list. You can select multiple data owners or enter custom data owners.
Description	A description of the subject area group to be created.

Importing a Subject

Step 1 On the **Data Design** console, choose **Subjects** from the left navigation bar.

Step 2 Click **Import** in the upper left corner. The **Import Subject** dialog box is displayed.

Step 3 In the **Import Subject** dialog box, set the parameters as prompt, and click **Upload and Verify**.

Table 7-5 Parameters for importing subjects

Parameter	Description
Update Existing Data	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. • **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. During the import, only subject creation and update are allowed.
Upload File	Select the file you want to import. You can use either of the following methods to obtain the file to be imported: • **Downloading the subject import template and fill in it** In the **Import Subject** dialog box, click **Subject Template** to download the template, set parameters in the template as required, and save the settings. Refer to **Table 7-6** for template parameter details. • **Exporting subjects to files** You can export the subjects created in Data Design of a DGC instance to Excel files in advance. Then, import the Excel files. See **Exporting a Subject** for details.

 Copyright © Huawei Technologies Co., Ltd.

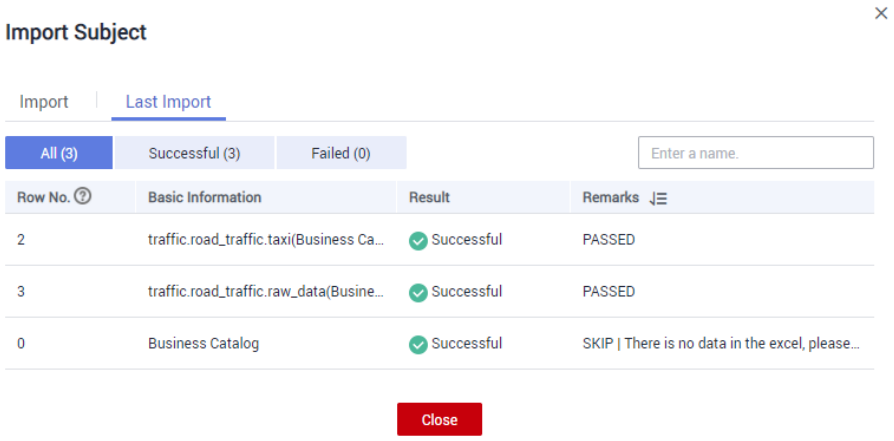
Table 7-6 describes the parameters required for importing a subject. Parameters whose names start with * are mandatory, and those that do not start with * are optional. You need to enter a line of information for a subject.

Table 7-6 Parameters

Parameter	Description
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 area group to be created. The code can include letters, numbers, spaces, colons (:), underscores (_), hyphens (-), parentheses, slashes (/), and ampersands (&) and cannot start with a space, <, >, or /.
Alias	Alias of a subject object.
Description	A description of the subject to be created.
Data Owner's Department	The department the data owner belongs to.
Data Owner	The owner of the data. Multiple owners are supported. Separate owner names with commas (,)

Step 4 You can 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 7-9 Last Import tab page



----End

 Copyright © Huawei Technologies Co., Ltd.

Exporting a Subject

Step 1 On the DGC Data Design console, choose **Subjects** from the left navigation bar.

Step 2 Click **Export** in the upper left corner to export the existing subject information to Excel files. Then, import the Excel files.

NOTE

A subjector process has a hierarchy. You can export only data of all levels.

----End

Managing a Subject

- Search

  You can enter a keyword in the search box of **Subject**, and search for your target.

- Edit

  You can select a target from the subject list, and then click to the right of the target to edit it.

- Delete

  You can select one or more targets from the subject list and click **Delete** above the list to delete the selected subjects.

- Move Up/Down

  You can select a target from the subject list and then click to the right of the target to adjust its ranking sequence.

7.7 Managing Lookup Tables

A lookup table, also called data dictionary table, consists of enumerable data names and codes. It stores the relationships between enumeration data names and codes. A lookup table:

- Standardizes business data and supplement mapping fields during data cleaning.
- Monitors the value range of the 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. **Log in to the DGC console**, select an instance, and click **Access** on it.

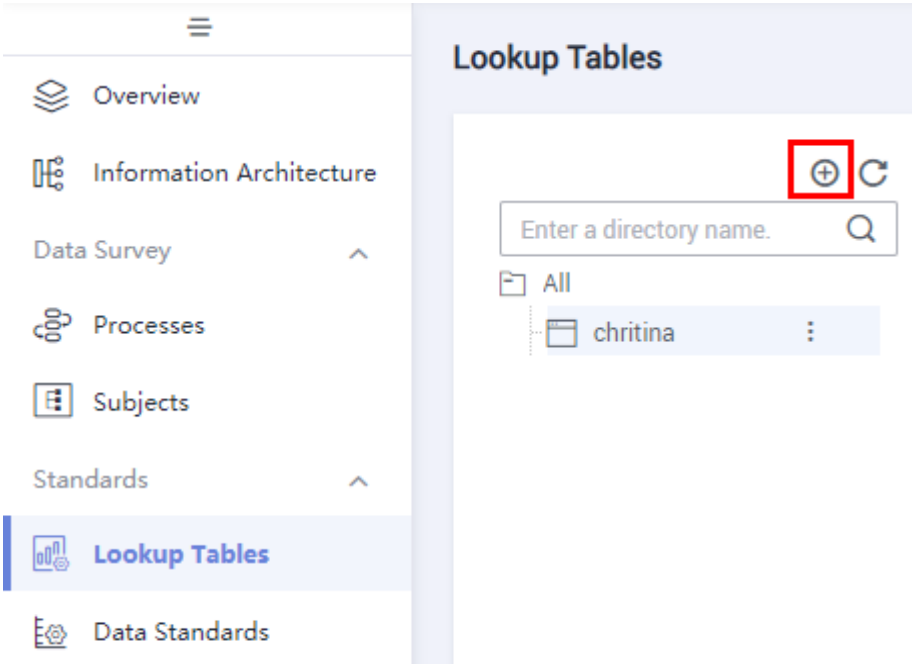
   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Design** of the corresponding workspace to access the Data Design console.

 Copyright © Huawei Technologies Co., Ltd.

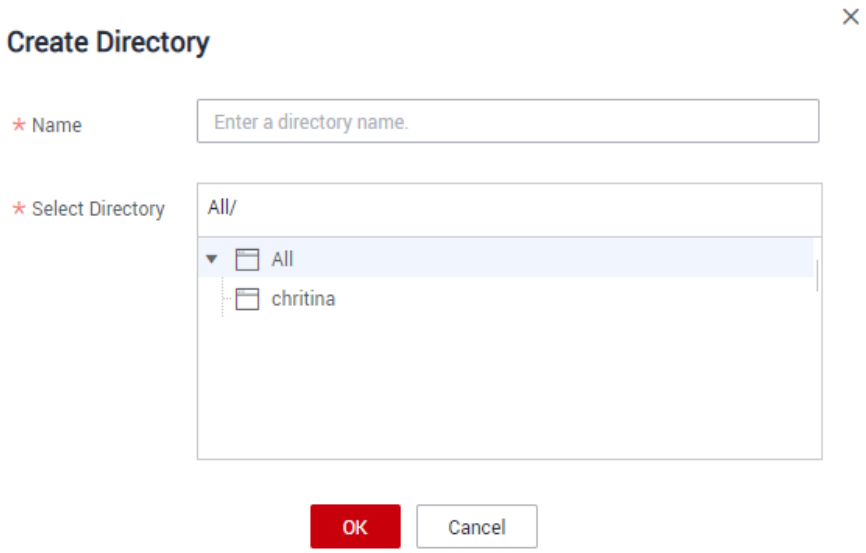
3. Choose **Lookup Tables** from the left navigation bar.
4. Select a directory from the directory tree on the **Lookup Tables** page, and then click ⊕ to create a directory under the selected directory. When creating a directory for the first time, you can create a directory under the root directory.

Figure 7-10 Lookup Tables page



5. In the displayed dialog box, set the parameters as prompted and click **OK**.

Figure 7-11 Create Directory dialog box



 Copyright © Huawei Technologies Co., Ltd.

Table 7-7 Parameters

Parameter	Description
Directory Name	Only letters, digits, and underscores (_) are allowed.
Select Directory	Select an existing directory, and create a sub-directory under it.

6. Select the directory you created in the directory tree and click **Create** to create a lookup table.
7. On the **Create Lookup Table** page, configure the parameters.

   In the **Basic Settings** area, set the parameters described in the figure below:

Figure 7-12 Basic Settings tab

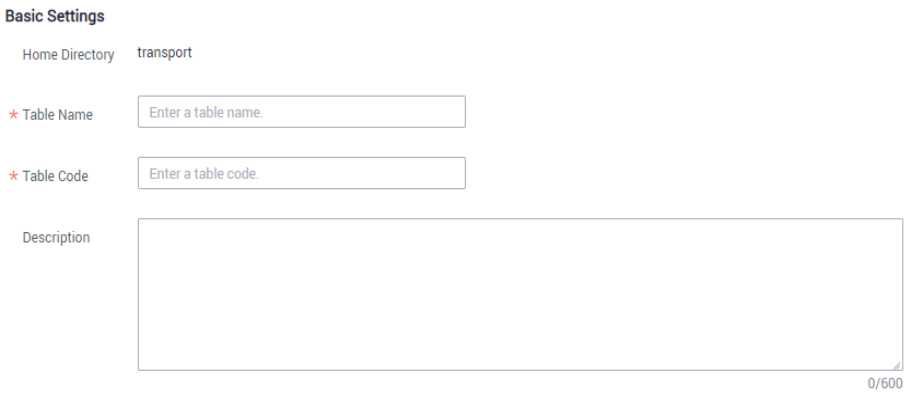


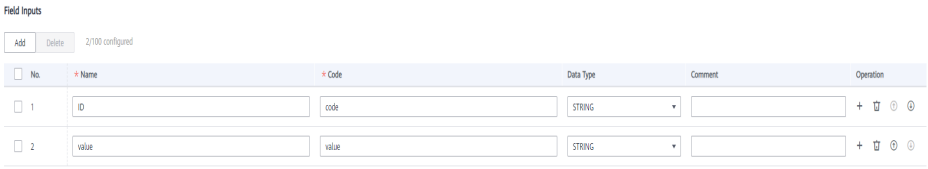
Table 7-8 Parameters

Parameter	Description
Table Name	The name of the lookup table to be created. Only letters, digits, brackets, spaces, hyphens (-) and underscores (_) are allowed. Table names must start with letters.
Table Code	The code of the lookup table to be created. The code can include letters, numbers, spaces, colons (:), underscores (_), hyphens (-), parentheses, slashes (/), and ampersands (&) and cannot start with a space, <, >, or /.
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 to delete unnecessary fields.

 Copyright © Huawei Technologies Co., Ltd.

Figure 7-13 Field Inputs tab



8. Click **Publish**. In the displayed **Apply for Publication** dialog box, 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

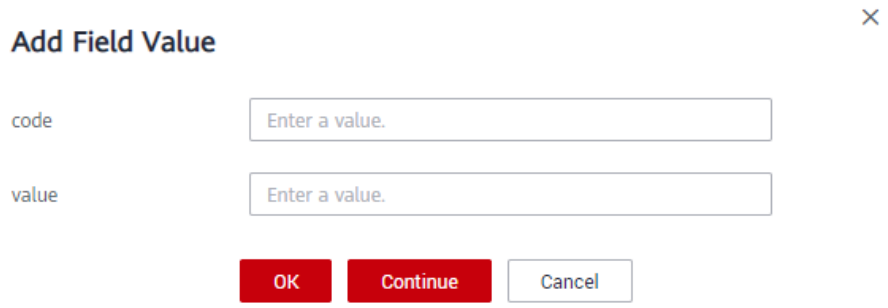
   If the current user has been added as a reviewer, you can select **Auto-review** and click **OK**. After the application is approved, the lookup table status changes to **Published**.

Filling in a Lookup Table

Input values in the created lookup tables.

1. On the Data Design console, choose **Lookup Tables** from the left navigation bar.
2. In the list of lookup tables, find the target table and choose **More** > **Value Input** in the **Operation** column.
3. On the page that is displayed, click **Add**. In the dialog box that is displayed, set parameters.

Figure 7-14 Inputting a value



4. Click **OK**. You can also click **Continue** to add additional records.

Importing a Lookup

When importing a lookup table, ensure that the table name contains a maximum of 32 characters.

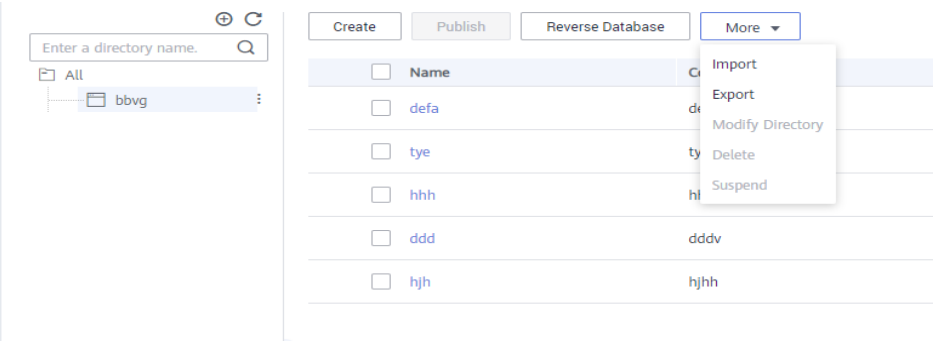
You can import a new lookup table or import lookup table records in batches from an existing lookup table. If you have a large number of lookup table records, you are advised to import them in batches.

Step 1 On the Data Design console, choose **Lookup Tables** from the left navigation bar.

 Copyright © Huawei Technologies Co., Ltd.

Step 2 On the displayed page, select a directory, and choose **More** > **Import**. You can also right-click the selected directory and choose **Import** from the shortcut menu.

Figure 7-15 Lookup Tables page



Step 3 In the **Import Lookup Table** dialog box, set the parameters as prompt, and click **Upload and Verify**.

Table 7-9 Parameters for importing lookup tables

Parameter	Description
Update Existing Table	Whether to update the existing lookup tables. When a lookup table is imported, the system checks whether the lookup table exists according to its code. The options are as follows: • **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 need to publish the lookup table again after updating it so that the updated lookup table can take effect. During the import, only lookup table creation and update are allowed.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Upload File	Select the file to be imported. You can use either of the following methods to obtain the file to be imported: • **Downloading the lookup table template and fill in it** In the **Import Lookup Table** dialog box, click **Lookup Table Template** to download the template, set parameters in the template as required, and save the settings. Refer to **Table 7-10** for template parameter details. Instructions for filling in the lookup table template: – Parameters whose names start with * are mandatory, and those that do not start with * 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 Data Design of a DGC instance to Excel files in advance. Then, import the Excel files. For details on how to export lookup tables, see **Managing a Lookup Table**.

Table 7-10 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 be created. Only letters, digits, brackets, spaces, hyphens (-) and underscores (_) are allowed. Table names must start with letters.
*Table Code	The code of the lookup table to be created. 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. Only letters, digits, brackets, spaces, hyphens (-) and underscores (_) are allowed. Table names must start with letters.

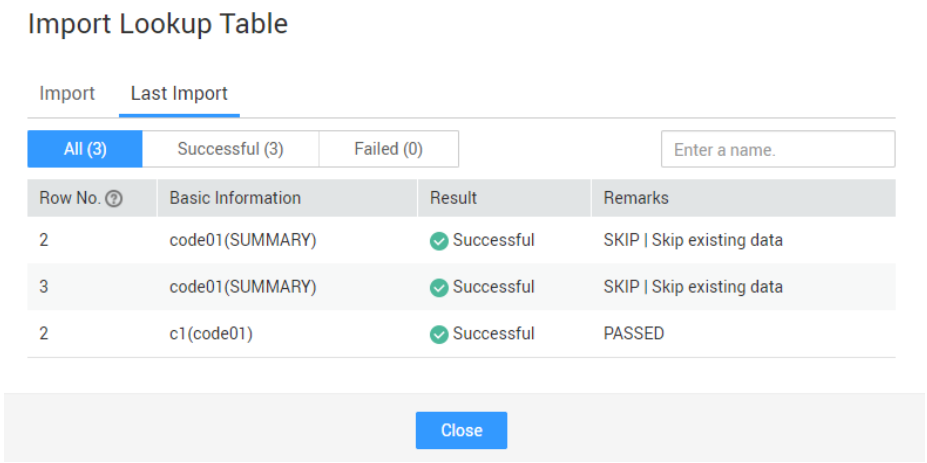
 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
*Field Code	The code of a field. Only letters, numbers, and underscores (_) are allowed. A table 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	• **true** indicates to generate a data standard. • **false** indicates not to generate a data standard. The default value is **false**.

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 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 You can 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 7-16 Last Import tab page



----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 Data Design console, choose **Lookup Tables** from the left navigation bar.

 Copyright © Huawei Technologies Co., Ltd.

Step 2 On the displayed page, select a directory and click **Reverse Database** above the lookup table list.

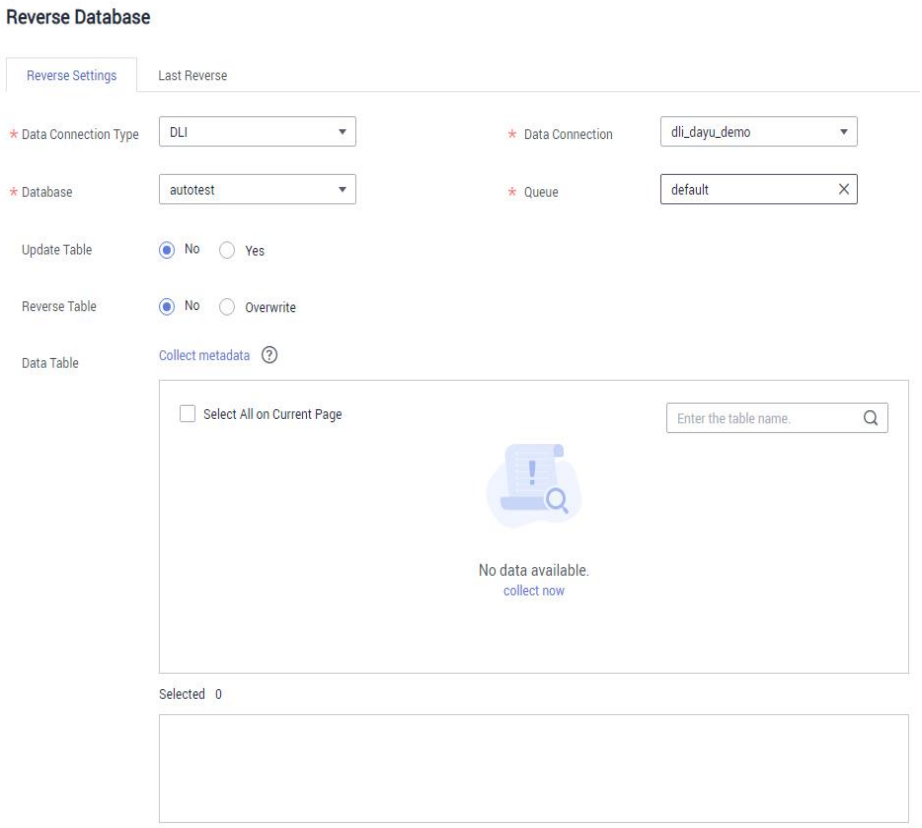
Step 3 In the **Reverse Database** dialog box, set the parameters described in the table below and click **Yes**.

Table 7-11 Parameters

Parameter	Description
Data Connection Type	The data connection types supported by the reverse database are displayed in the drop-down list. Select the required data connection type.
Data Connection	Select a data connection. If you want to reverse a database from other data sources to a lookup table directory, you need to create a data connection in DGC 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.
Queue	DLI queue. This parameter is available only when **Data Connection Type** is set to **DLI**.
Update Table	When you reverse a table from another data source, if the name of the reversed table is the same as that of an existing table in lookup table list, you can select this option.
Reverse Table	• **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 be imported.

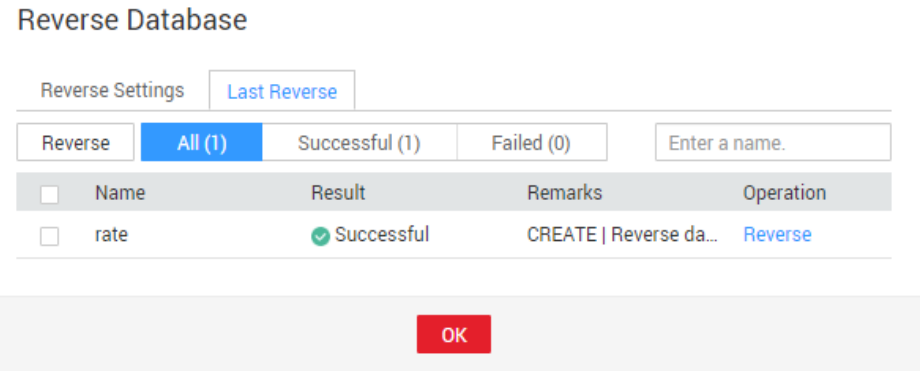
 Copyright © Huawei Technologies Co., Ltd.

Figure 7-17 Reverse Database dialog box



Step 4 You can view the result in 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 7-18 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 Data Design console, choose **Lookup Tables** from the left navigation bar.

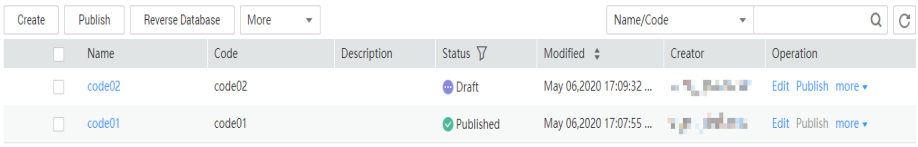
 Copyright © Huawei Technologies Co., Ltd.

Step 2 Export a lookup table

- Export a lookup table.

  In the lookup table list, select the lookup table you want to export and choose **More** > **Export**.

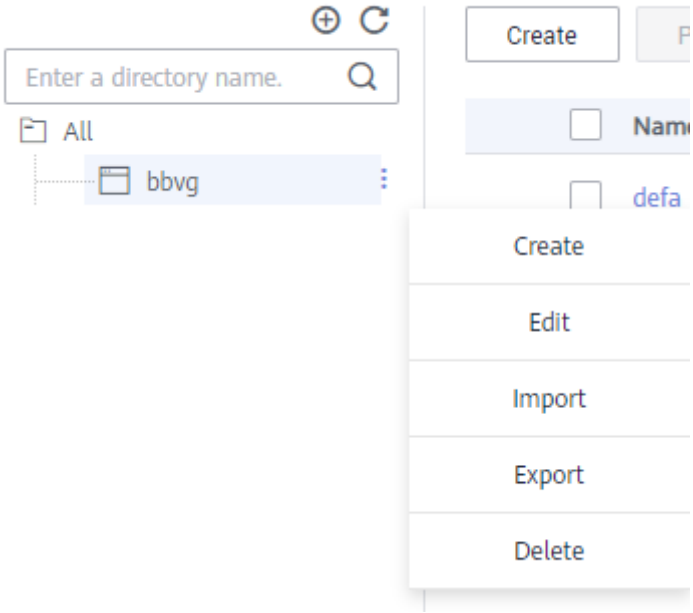
  Figure 7-19 Lookup table list



- **Export all tables in the list.**

  Right-click a directory in the directory tree and choose **Export** from the shortcut menu.

  Figure 7-20 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. You can delete a lookup table only after the status of the lookup table changed.

Step 1 On the Data Design console, choose **Lookup Tables** from the left navigation bar.

Step 2 In the lookup table list, select the lookup table you want to delete and choose **More** > **Delete** above the list.

 Copyright © Huawei Technologies Co., Ltd.

Step 3 In the displayed dialog box, click **Yes** to delete it.

----End

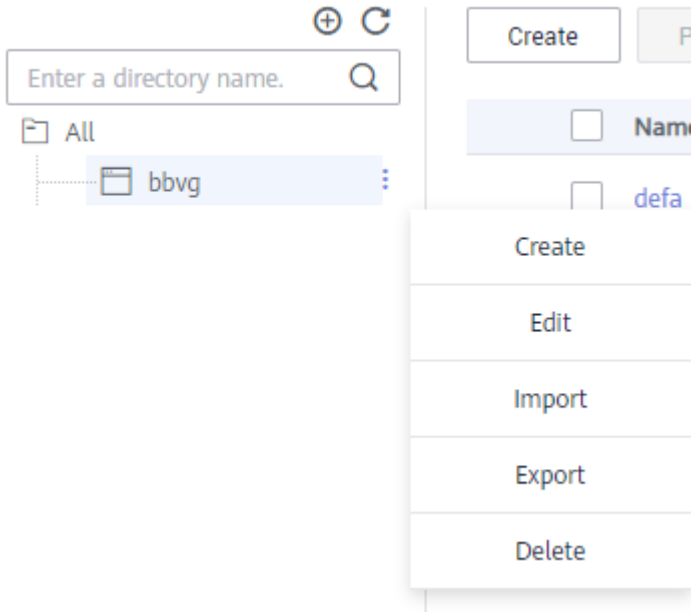
Deleting a Lookup Table Directory

If a directory or its subdirectories contain a lookup table, the lookup table directory or the subdirectories cannot be deleted. You need to delete the lookup table before deleting the directory.

Step 1 On the Data Design console, choose **Lookup Tables** from the left navigation bar.

Step 2 On the displayed page, right-click the directory you want to delete and choose **Delete** from the shortcut menu.

Figure 7-21 Managing lookup table directories



Step 3 In the displayed dialog box, click **Yes** to delete it.

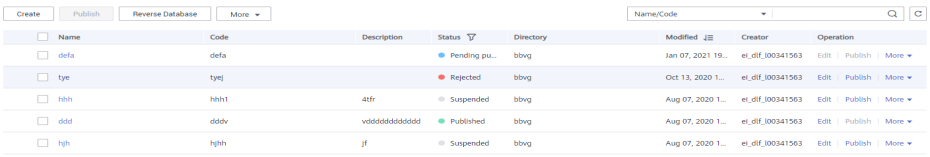
----End

Managing a Lookup Table

After a lookup table is created, you can search for, edit, or delete it.

On the Data Design console, choose **Lookup Tables** from the left navigation bar. You can manage the lookup tables as required.

Figure 7-22 Manage lookup tables



 Copyright © Huawei Technologies Co., Ltd.

- **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 where there is a table with the **Draft** or **Rejected** status, select a reviewer in the displayed dialog box, and click **Confirm**. After the application is approved, the lookup table is published.

- **Suspend**

  In the lookup table list, select a published lookup table you want to suspend, click **Suspend** in the **Operation** column, select a reviewer in the displayed dialog box, and submit the application. If the application is approved, the lookup table is suspended.

7.8 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 subdivided. A data standard is a data unit whose definition, identifiers, representations, and allowed values are specified by a group of attributes. 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 section 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. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

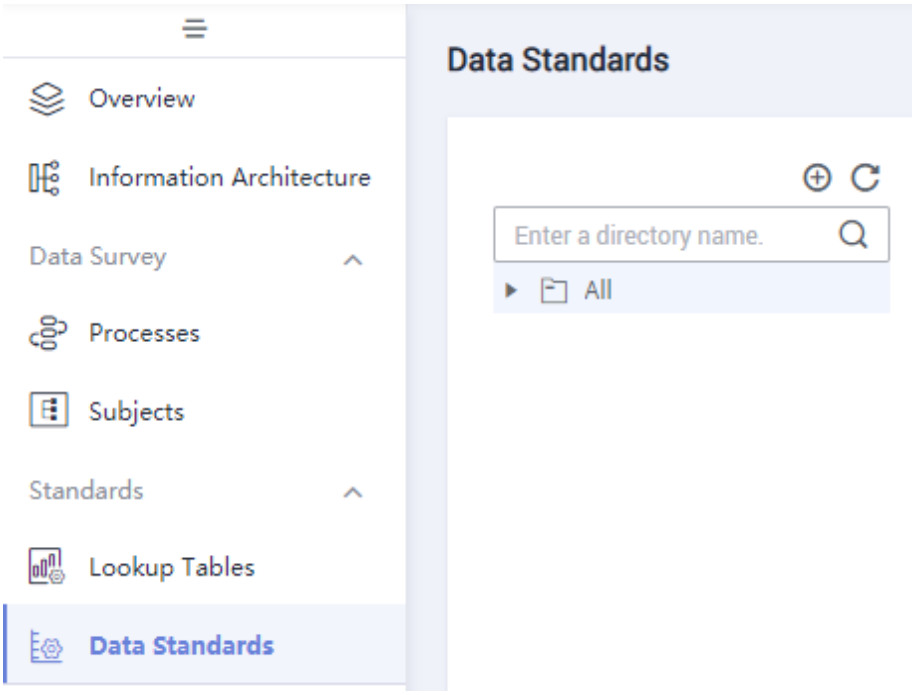
2. Click **Data Design** of the corresponding workspace.
3. Choose **Data Standards** from the left navigation bar.
4. When you access the **Data Design** page for the first time, the page where you can customize a standard template, shown in the following figure, is displayed. Select the required options for **Optional**, click + to add custom items, and then click **OK**.

   After saving the template settings, you can modify it on the **Standard Template Management** page of **Configuration Center**. For details, see **Managing a Standard Template**. When creating a data standard, you need to set the selected options in the template.

5. 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.

 Copyright © Huawei Technologies Co., Ltd.

Figure 7-23 Data Standards page



6. In the displayed dialog box, set parameters as prompted and click **OK**.

Figure 7-24 Create Directory dialog page

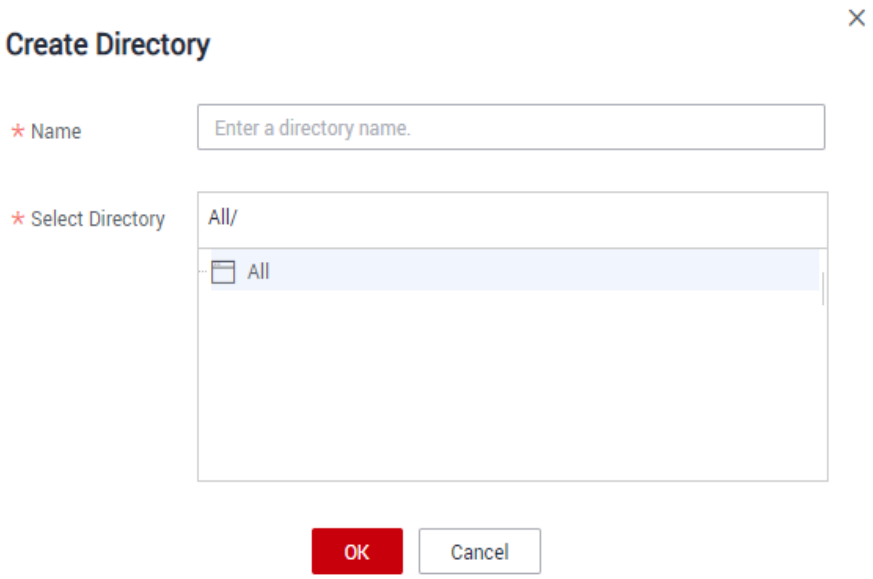


Table 7-12 Parameters for creating data standards

Parameter	Description
Directory Name	Only letters, digits, and underscores (_) are allowed.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Select Directory	Select an existing directory, and create a sub-directory under it.

Creating a Data Standard

Step 1 In the **Data Standards** page, select a directory and click **Create**.

Step 2 Set parameters by referring to **Table 7-13** 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 Template Management** tab page of the **Configuration Center** are displayed. **Table 7-13** lists all parameters that are available in a data standard template. For details on how to configure a data standard template, see **Managing a Standard Template**.

Figure 7-25 Creating a data standard

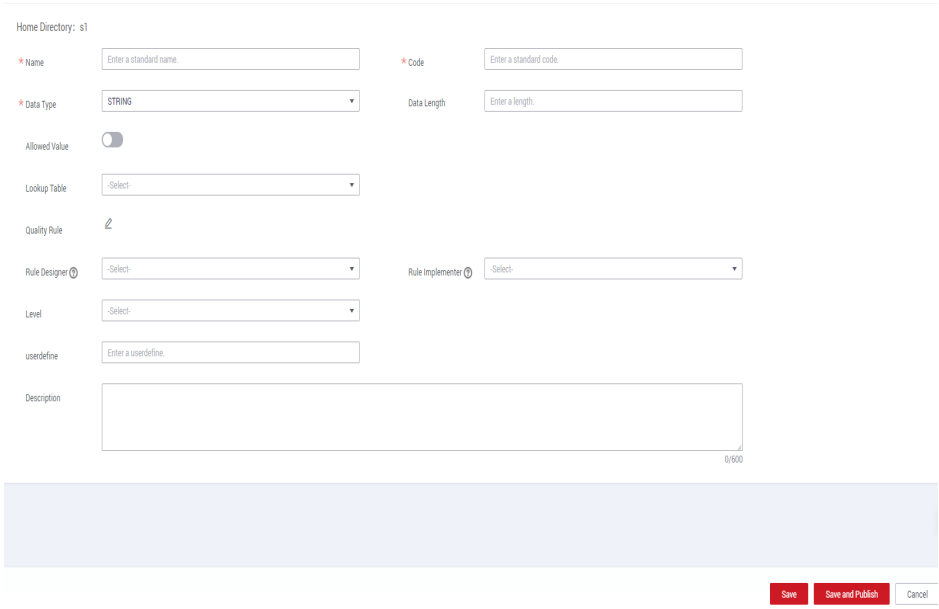


Table 7-13 Parameters

Parameter	Description
*Standard Name	Letters, digits, brackets, spaces, and underscores (_) are allowed, but field type names must start with letters.
*Standard Code	The value must be unique in the workspace. It is used to identify a data standard record. Letters, digits, and brackets, spaces, and underscores (_) are allowed, but standard codes must start with letters.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
*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 **Managing a Field Type**.
Data Length	This field can be left blank. You can also enter a value that ranges from 0 to 10,000. If it is left blank, it means there is no limit to the data length.
Allowed Value	If this field is selected, enter a value and press **Enter** to add an allowed value. You can add up to 20 allowed values.
Referenced Lookup Table	Select a created lookup table and corresponding table fields. In this way, the lookup table fields can be associated with data standards. If no lookup table is created, create one. See **Managing Lookup Tables**. If **Create Quality Jobs** is selected under **Model Design Process** on the **Function Configuration** tab of **Configuration Center**, and the data standards of the reference lookup tables are associated with the business tables in ER modeling, the system will automatically create quality jobs in the Data Quality Control module of DGC when the business tables are published, and generates quality rules based on the associated data standards 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 standards and lookup tables.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Quality Rule	If **Create Quality Jobs** is selected under **Model Design Process** on the **Function Configuration** tab of **Configuration Center**, your data standard template includes parameter **Quality Rule**, and the data standard is associated with a table, the system will automatically create a quality job in Data Quality Control of DGC 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. Click . In the displayed dialog box, click **Add Rule**. For example, add a rule named **Field unique value**, select the rule, 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. The alarm parameters are listed as buttons. If you click these buttons, the alarm condition value is displayed in the sequence of variables such as *$1*, *$2*, and *$3*. The variable names indicate the alarm parameters. In other words, when setting **Alarm Triggering Condition**, use the variable $1 to represent the first alarm parameter, $2 to represent the second alarm parameter, and so on.
Rule Designer	Select a business rule owner from the drop-down list. This owner is responsible for developing quality rules. Only the workspace administrator or developer can be set as the owner.
Rule Implementer	Select a data monitoring owner from the drop-down list. This owner is responsible for implementing quality rules. Only the workspace administrator or developer can be set as the owner.
Standard Level	• **global** indicates the global level. • **domain** indicates non-global level.
userdefine	Choose **Configuration Center** > **Standard Template Management** to add custom items. You can add one or more custom items based on the site requirements. For more information about adding custom items, see **Managing a Standard Template**.
Description	A description of the data standard to be created. Up to 600 characters are supported.

Step 3 Click **Save and Publish**. Wait for the reviewer to approve the application.

 Copyright © Huawei Technologies Co., Ltd.

After the data standard is approved, the status of the data standard changes to **published**, indicating that the data standard is released successfully.

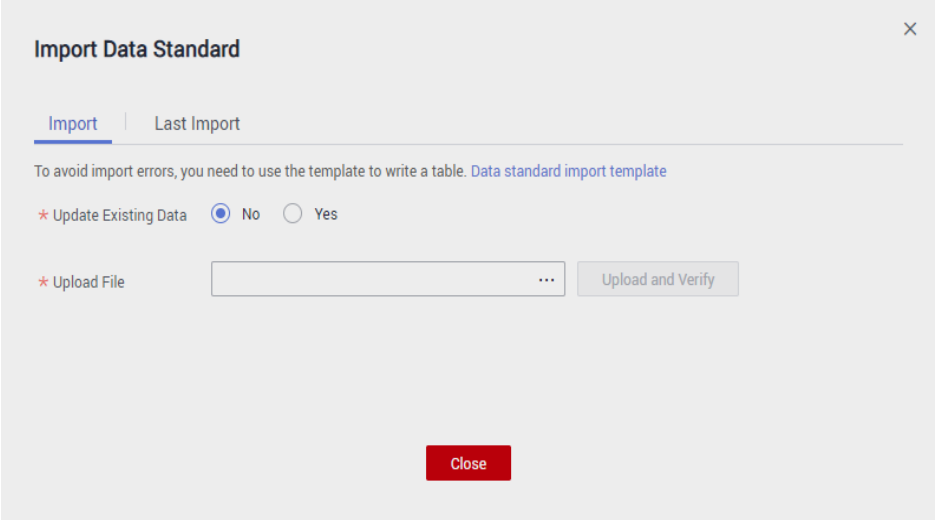
----End

Importing a Data Standard

Step 1 On the Data Design console, choose **Data Standards** from the left navigation bar. The **Data Standards** page is displayed.

Step 2 On the displayed page, specify a directory and choose **More** > **Import**. A dialog box is displayed.

Figure 7-26 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. That is, 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 parameters in the template based on business requirements, and save the settings.

Table 7-14 and Table 7-15 describes the parameters required for importing a data standard. Parameters whose names start with * are mandatory, and those that do not start with * are optional.

Table 7-14 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 be created. Letters, digits, brackets, spaces, and underscores (_) are allowed, but field type names must start with letters.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
*Standard Code	The name of the data standard code to be created, which uniquely identifies a data standard record. Letters, digits, and brackets, spaces, and underscores (_) are allowed, but standard codes must start with letters.
*Data Type	The possible values are STRING, BIGINT, DOUBLE, TIMESTAMP, DATE, BOOLEAN, and DECIMAL. If the required data type does not exist, you can add one. See **Managing a Field Type**. The data type varies according to the system. The system converts the data type internally.
Data Length	This field can be left blank. You can also enter a value that ranges from 0 to 10,000. If it is left blank, it means there is no limit to the data length.
Allowed Value	The value **true** indicates that the value is allowed, and the value **false** indicates that the value is not allowed.
Yes	If you select **Yes** 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.
Referenced Lookup Table	Set this parameter to the name of a created lookup table.
Lookup Table Field	If **Referenced Lookup Table** is not left blank, you need to set **Lookup Table Field**. In this way, the code table field can be associated with the data standard.
Owner of Business Rules	Enter a rule designer. Only the workspace administrator or developer can be set as the owner.
Owner of Data Monitoring	Enter a rule implementer. Only the workspace administrator or developer can be set as the owner.
Standard Level	• **global** indicates the global level. • **domain** indicates non-global level.
Description	A description of the data standard to be imported. Up to 600 characters are supported.
(Optional) Custom Item	If you have added one or more custom fields when customizing a data standard template, you also need to 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 **Managing a Standard Template**.

Copyright © Huawei Technologies Co., Ltd.

In the **Quality Rule** sheet, you can configure quality rules to be added for data standards. If **Create Quality Jobs** is selected under **Model Design Process** on the **Function Configuration** tab of **Configuration Center**, your data standard template includes parameter **Quality Rule**, and the data standard is associated with a table, the system will automatically create a quality job in Data Quality Control of DGC 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 7-15 Parameters in the "Quality Rule" sheet

Parameter	Description
*Standard Name	The name 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 DGC console, select **Data Quality Control** from the drop-down list. Then, you can view the existing rule names on the **Rule Template** page.
Alarm Triggering Condition	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 DGC console, select **Data Quality Control** from the drop-down list. Then, you can go to the **Rule Template** page and view the alarm parameters supported by the quality rule in the **Result Description** column. Example: ${1} > 100.
Regular Expression	A regular expression needs to be configured only 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 and Verify**.

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.

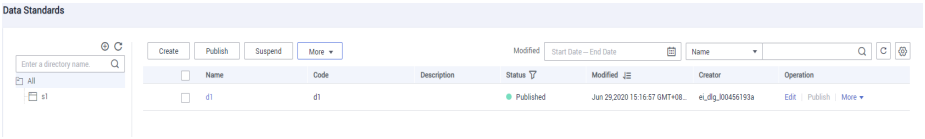
----End

 Copyright © Huawei Technologies Co., Ltd.

Managing a Data Standard

On the DGC Data Design console, choose **Data Standards** from the left navigation bar. On the displayed page, you can manage data standards as required.

Figure 7-27 List of data standards



On the **Data Standards** page, you can perform the following operations:

- **Search**

  In the upper part of the **Data Standards** area, set criteria such as the standard name, creator, reviewer, and modification time, and click the search icon to search for the specified 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 tree, 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**

  Choose **More** > **Delete** to delete a data standard. A data standard with the pending publishing, published, or pending suspension status cannot be deleted. Referenced data standards cannot be deleted as well.

- **Publish**

  Click **Publish** to publish a data standard.

Exporting a Data Standard

Step 1 On the Data Design console, choose **Data Standards** from the left navigation bar. The **Data Standards** page is displayed.

Step 2 In the directory tree of data standards, specify a directory and click **Export**.

----End

 Copyright © Huawei Technologies Co., Ltd.

7.9 ER Modeling

7.9.1 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.

To create an ER model tailored to your business scenario, you need to create a physical model first, and then create tables in the model.

The following parts are included in this section:

- **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.

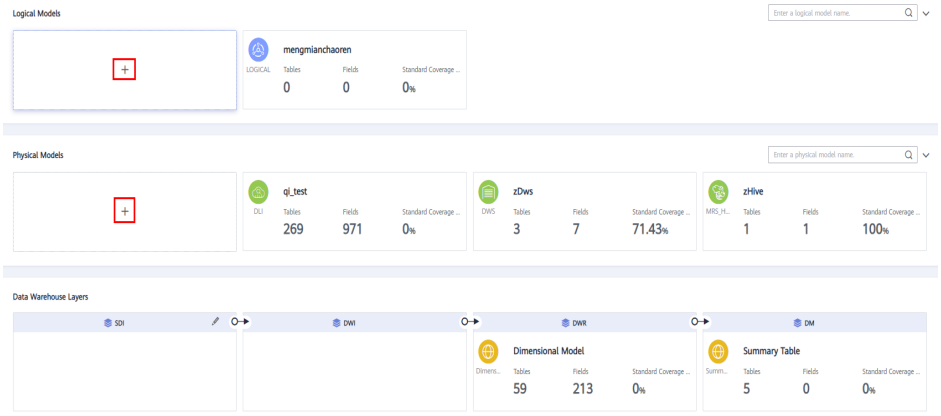
Creating a Physical Model

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Design** of the corresponding workspace to access the Data Design console.
3. Choose **ER Modeling** from the left navigation bar.
4. 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 new models. See the figure below.

Copyright © Huawei Technologies Co., Ltd.

Figure 7-28 ER Modeling page



5. In the displayed dialog box, set parameters as prompted and click **OK**.

Figure 7-29 Creating a model

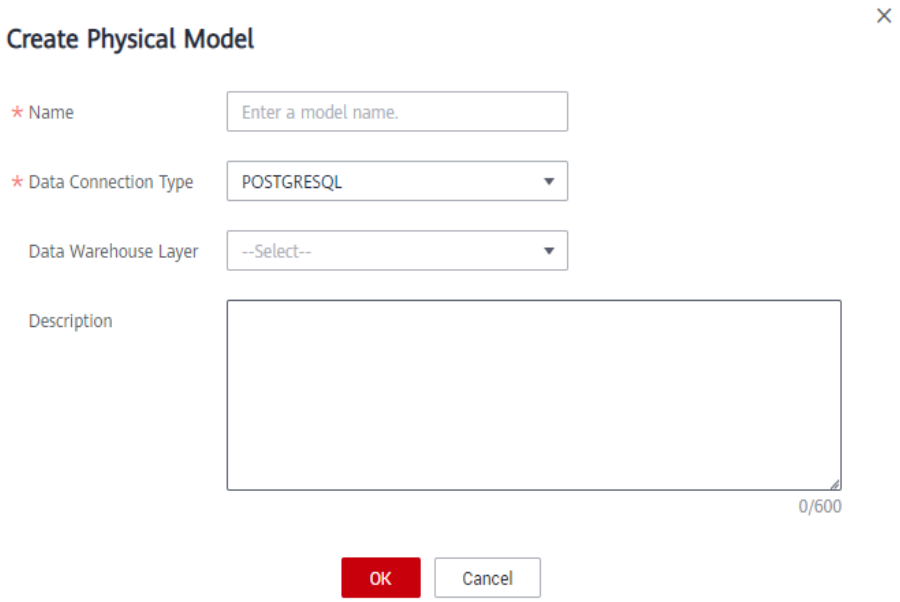


Table 7-16 Parameters

Parameter	Description
Model Name	Only letters, digits, and underscores (_) are allowed.
Model Type	Select **Physical model**. For details on the physical model, see the **physical model description**.
Data Connection Type	Currently, DLI, DWS, and MRS_HIVE are supported.
Description	A description of the ER model. Up to 600 characters are supported.

 Copyright © Huawei Technologies Co., Ltd.

Creating and Publishing a Table

After creating a DLI, DWS, or an MRS_HIVE ER model, you can create a business table in the model.

Step 1 On the DGC Data Design console, choose **ER Modeling** from the left navigation bar.

Step 2 On the displayed page, select a model, expand it, and select a subdirectory. Then, click **Create**.

Step 3 On the **Create Table** page, set parameters as required.

1. Set the basic parameters described in the figure below.

Figure 7-30 Basic Settings tab page

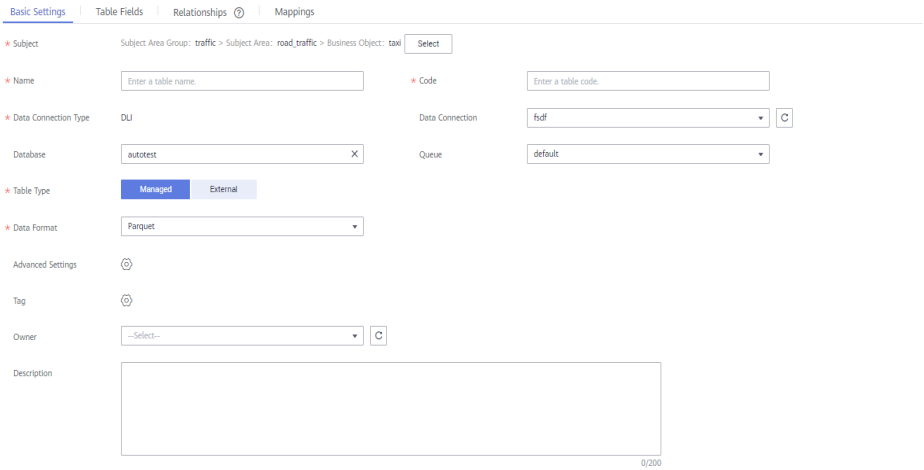


Table 7-17 Parameters

Parameter	Description
Subject	Click **Select** to select a subject.
Table Name	The name of the table to be created. Only letters, digits, and underscores (_) are allowed. Table names cannot start with digits.
Table Code	The code of the table to be created. Only letters, digits, underscores (_), $, {, and } are allowed. Table codes cannot start with digits.
Data Connection Type	N/A

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
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 you have not created a data connection yet, go to the DGC 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.
Queue	DLI queue. This parameter is available only for tables created in DLI models.
Schema	DWS mode. This parameter is valid only for DWS tables.
Table Type	DLI models support the following table types: – **Managed** indicates that data is stored in a DLI table. – **External** indicates that data is stored in an OBS table. When **Table Type** is set to **External**, you need to set **OBS Path**. The OBS path format is ***/bucket_name/filepath***. DWS models support the following table types: – **DWS_ROW** indicates that tables are stored to disk partitions by row. – **DWS_COLUMN** indicates that tables are stored to disk partitions by column. This parameter is unavailable for the tables created in MRS_HIVE models.
Data Format	This parameter is available only for tables created in DLI models. The supported data formats are as follows: – **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 uncompressed avro data.
Advanced Settings	You can set custom items.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Tag	Tags are custom identifiers that help users classify and search for data assets. After adding a tag, you can search for related data assets in the DGC Data Assets module with ease. Click to add a tag to the table. In the dialog box that is displayed, select one or more existing tags, or enter a new tag name and press **Enter**. You can also go to the **Tag Management** page of the DGC Data Assets module to add a tag. Then, return to this page and select the newly added tag from the tag drop-down list. See **Tags**.
Owner	Select an asset owner from the drop-down list. Only the workspace administrator or developer can be set as the owner.
Description	A description of the table. Up to 200 characters are supported.

2. Add required fields on the **Table Fields** page.

Figure 7-31 Table Fields tab page

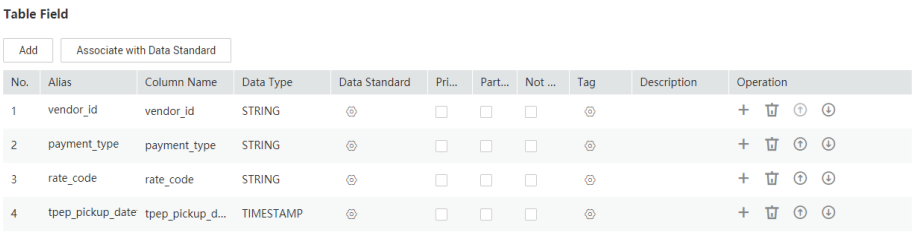


Table 7-18 Parameters of table fields

Parameter	Description
Name	Table names must start with letters. Only letters, digits, and underscores (_) are allowed.
Code	Table codes must start with letters. Only letters, digits, and underscores (_) are allowed.
Data Type	Field data type. If the required data type does not exist, you can add one. See **Managing a Field Type**.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Data Standard	If you have created data standards, click to select one to associate with the field. Choose **Configuration Center** > **Function Configuration** and select **Create Quality Jobs**. Associate fields with a data standard. After a table is published, a quality job is automatically generated. 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 go to the **Quality Job** page of the DGC Data Quality Control to view the job details. If you click **Associate with Data Standard** above the field list and select a data standard that you want to associate the field with, a new table field associated with the selected data standard is automatically created. The default field name and code are the same as those of the associated data standard. You can change the field name and code if necessary. If you have not created a data standard, create one. See **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.
Tag	Click to add a tag. – In the displayed dialog box, select one or more existing tags. If no tag has been added, you can go to the **Tag Management** page of the DGC Data Assets to add a tag. See **Tags**. – In the dialog box that is displayed, enter a new tag name and press Enter. Tag names can contain letters, digits, and underscores (_), but cannot start with underscores (_).
Description	A description of the field to be added.

3. (Optional) On the **Relationships** tab page, click **Create** to create a relationship.

   A relationship refers to the logical relationship between two entities. 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.

 Copyright © Huawei Technologies Co., Ltd.

Figure 7-32 Relationships tab page

Table 7-19 Parameters

Parameter	Description
Name	Each relationship has a name to describe the function of the relationship.
Child Table	Click this field to select a table from the drop-down list. Click to set the current table as a child table.
Child Table Field	Select fields in the child table.
Child to Table	indicates that each piece of data in the child table corresponds to only one piece of data in the parent table. indicates that each piece of data in the child table corresponds to one piece of data at most in the parent table. indicates that one piece of data in the child table corresponds to multiple pieces of data in the parent table. indicates that each piece of data in the child table corresponds to at least one piece of data in the parent table.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Parent to Child	indicates that each piece of data in the parent table corresponds to only one piece of data in the child table. indicates that each piece of data in the parent table corresponds to one piece of data at most in the child table. indicates that one piece of data in the parent table corresponds to multiple pieces of data in the child table. indicates that one piece of data in the parent table corresponds to one piece of data in the child table at least.
Parent Table	Select a table that has a logical relationship with the selected child table.
Table Field	Select fields in the parent table.
Role Name	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 need to 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, you need to 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.
     - **map1**: Create a table named **table02** from ER model A. 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.

 Copyright © Huawei Technologies Co., Ltd.

- 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, assume that all fields in the current table come from ER model **d1**, the first field comes from the **vendor** table, the second field comes from the **payment_type** table, the third field comes from the **rate** table, and other fields come from the **dwd_taxi_trip_data** table.

  You can create a mapping, as shown in **Figure 7-33**. 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 7-20**.

Figure 7-33 Configuring a mapping

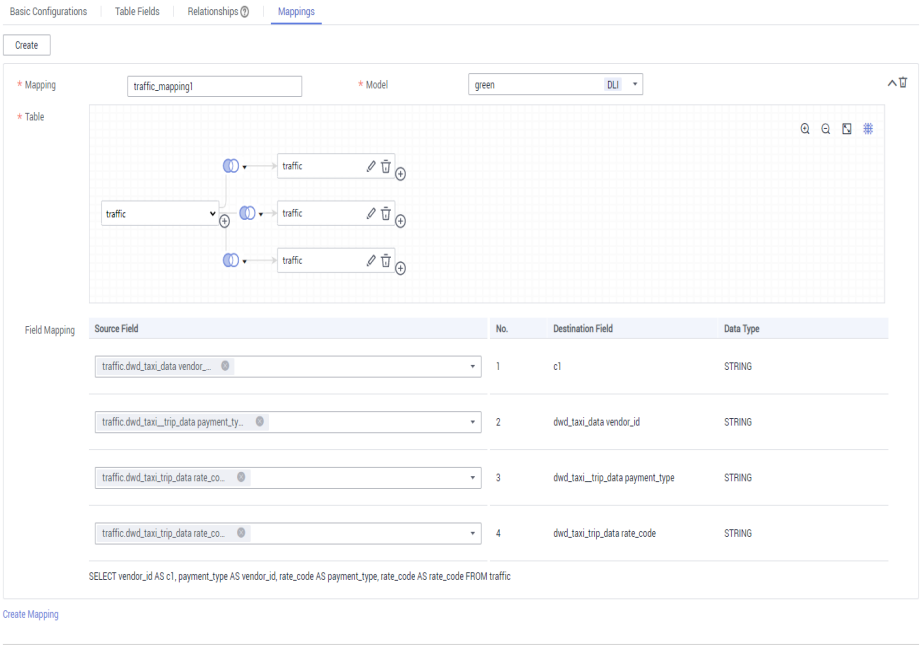
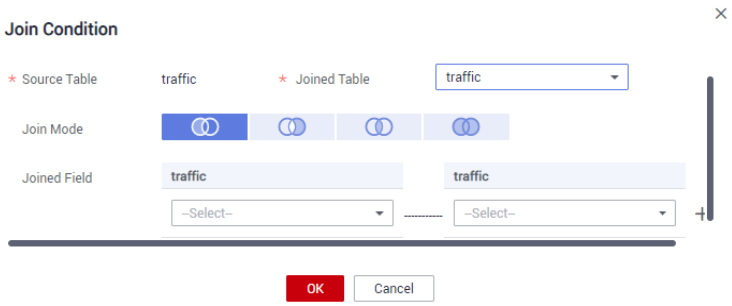


Table 7-20 Parameters of mappings

Parameter	Description
Mapping	Only letters, digits, 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**.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Table	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 another table. 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. **Figure 7-34 Join Condition** dialog box 
Field Mapping	Select a source field with the same meaning as the current mapping field. If a table field comes from multiple models, you need to 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 specified mapping or click to collapse the mapping area.

5. (Optional) If the type of the new table is **DWS_VIEW**, click **Create** to create a view.

 Copyright © Huawei Technologies Co., Ltd.

Figure 7-35 Creating a view

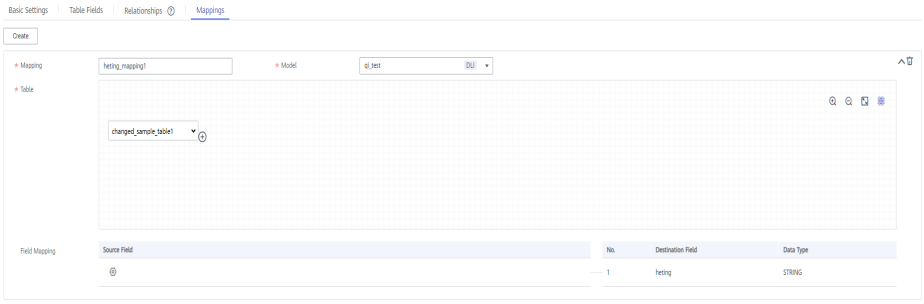
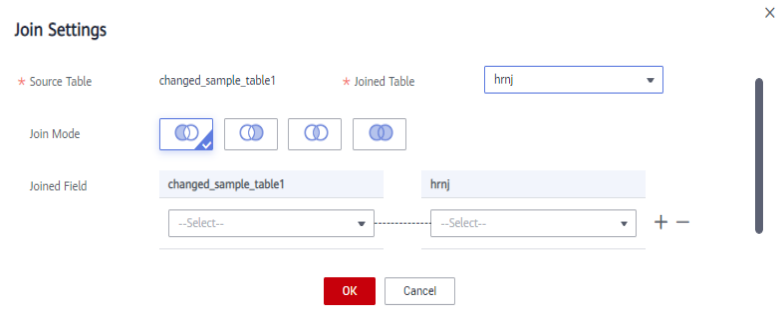


Table 7-21 Parameters

Parameter	Description
Mapping	Only letters, digits, 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	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 another table. 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. **Figure 7-36 Join Condition** dialog box 

 Copyright © Huawei Technologies Co., Ltd.

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 need to 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 specified 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, operations such as creating tables and synchronizing technical assets and business assets based on the configurations of **Model Design Process** on the **Function Configuration** page. 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** > **Publish History** to view logs.

  Troubleshoot the problem based on the logs. After the error is rectified, click **Resynchronize** in the **Publish Log** tab to issue the synchronization command again. If the synchronization still fails, contact technical support for assistance.

----End

7.9.2 Designing Logical Models

A logical model is an entity relationship diagram that accurately describes business rules based on entities and their relationships. The logical model must ensure the correctness and consistency of the data structure required by business, reflect the features of diverse objects based on a series of standard rules, and accurately define the relationships between entities. In addition, the logical model provides a powerful reference for constructing the physical model, which is an indispensable phase in the process of designing a database.

The following parts are included in this section:

- **Considerations in Logical Model Design**

 Copyright © Huawei Technologies Co., Ltd.

- Creating a Logical Model
- Creating and Publishing a Logical Table
- Converting a Logical Model to a Physical Model

Considerations in Logical Model Design

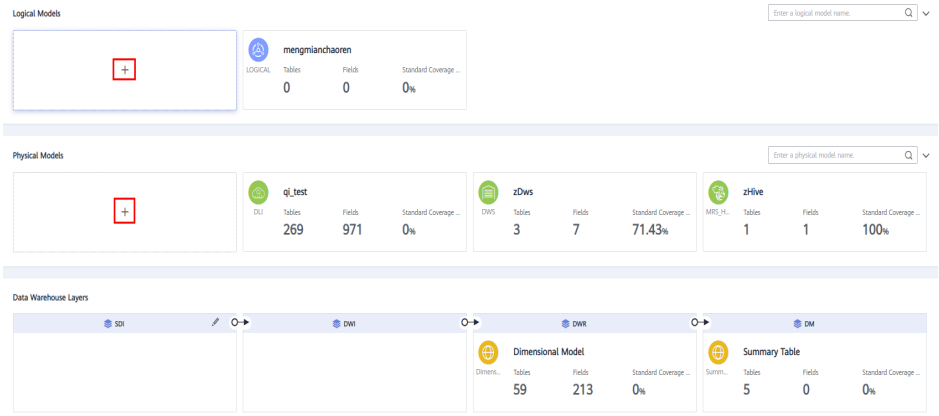
- You must consider not only the current business status, but also the future business development.
- Personnel who are familiar with the businesses must participate in the modeling. In this way, the business requirements can be fully integrated into the models.
- High efficiency of transforming the logical model to the physical model must be ensured.
- Take physical features into consideration in the physical modeling phase.
- Each entity, attribute, and relationship must be consistent with the information in the actual business.

Creating a Logical Model

1. Log in to the DGC console, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Design** of the corresponding workspace to access the Data Design console.
3. Choose **ER Modeling** from the left navigation bar.
4. 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 new models. See the figure below.

   Figure 7-37 ER Modeling page



5. In the displayed dialog box, set parameters as prompted and click **OK**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 7-38 Creating a logical model

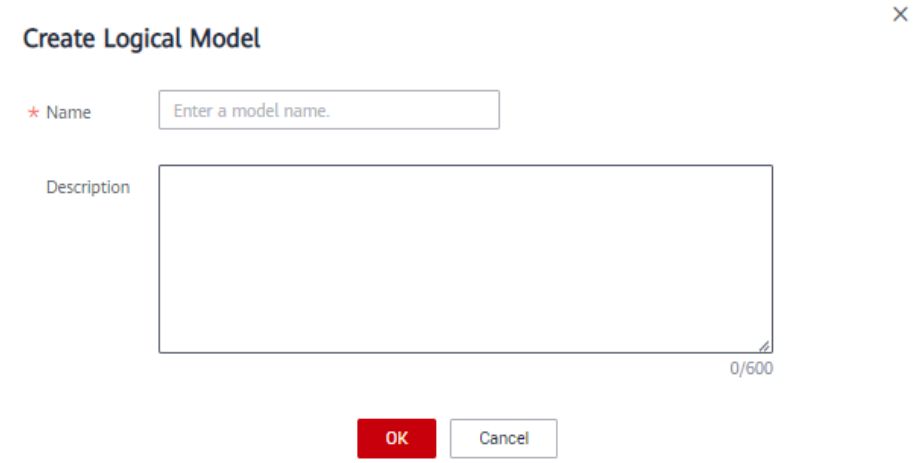


Table 7-22 Parameters

Parameter	Description
Name	Only letters, digits, and underscores (_) are allowed.
Description	A description of the logical model.

Creating and Publishing a Logical Table

After creating a logical model, you can create a logical table.

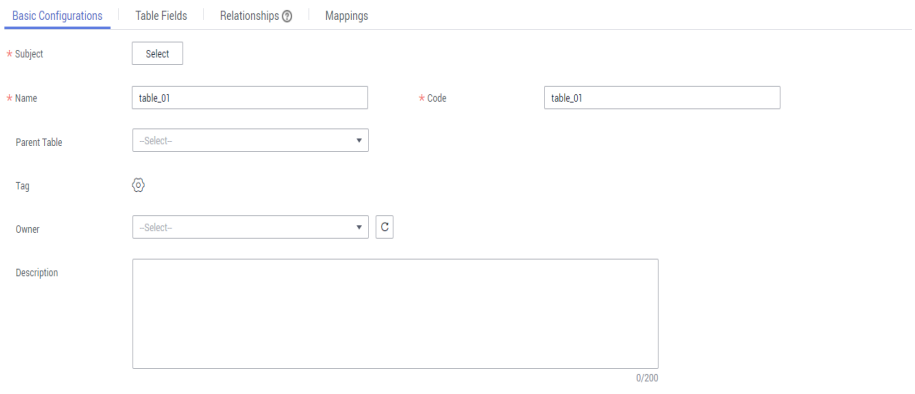
Step 1 On the DGC Data Design console, choose **ER Modeling** from the left navigation bar.

Step 2 On the displayed page, select a logical model, expand it, and select a subdirectory. Then, click **Create**.

Step 3 On the **Create Table** page, set parameters as required.

1. Set the basic parameters described in the figure below:

Figure 7-39 Basic configurations of logical tables



 Copyright © Huawei Technologies Co., Ltd.

Table 7-23 Parameters

Parameter	Description
Subject	Click **Select** to select a subject.
Name	The name of the table to be created. Only letters, digits, and underscores (_) are allowed. Table names cannot start with digits.
Code	The code of the table to be created. Only letters, digits, underscores (_), $, {, and } are allowed. Table codes cannot start with digits.
Parent Table	Set a parent table.
Tag	Tags are custom identifiers that help users classify and search for data assets. After adding a tag, you can search for related data assets in the DGC Data Assets module with ease. Click to add a tag to the table. In the dialog box that is displayed, select one or more existing tags, or enter a new tag name and press **Enter**. You can also go to the **Tag Management** page of the DGC Data Assets module to add a tag. Then, return to this page and select the newly added tag from the tag drop-down list. See **Tags**.
Owner	Select an asset owner from the drop-down list. Only the workspace administrator or developer can be set as the owner.
Description	A description of the table to be created. Up to 200 characters are supported.

2. Add required fields on the **Table Fields** tab page. See **Table 7-24** for details.

Figure 7-40 Adding a table field

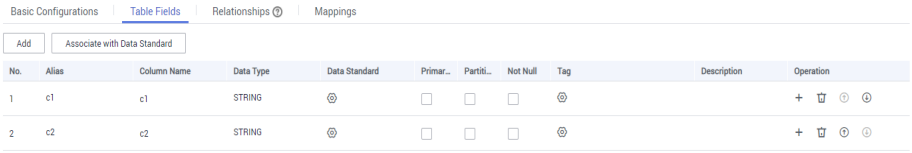


Table 7-24 Parameters of table fields

Parameter	Description
Name	Table names must start with letters. Only letters, digits, and underscores (_) are allowed.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Code	Table codes must start with letters. Only letters, digits, and underscores (_) are allowed.
Data Type	Field data type. If the required data type does not exist, you can add one. See **Managing a Field Type**.
Data Standard	If you have created data standards, click to select one to associate with the field. Choose **Configuration Center** > **Function Configuration** and select **Create Quality Jobs**. Associate fields with a data standard. After a table is published, a quality job is automatically generated. 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 go to the **Quality Job** page of the DGC Data Quality Control to view the job details. If you click **Associate with Data Standard** above the field list and select a data standard that you want to associate the field with, a new table field associated with the selected data standard is automatically created. The default field name and code are the same as those of the associated data standard. You can change the field name and code if necessary. If you have not created a data standard, create one. See **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.
Tag	Click to add a tag. – In the displayed dialog box, select one or more existing tags. If no tag has been added, you can go to the **Tag Management** page of the DGC Data Assets to add a tag. See **Tags**. – In the dialog box that is displayed, enter a new tag name and press Enter. Tag names can contain letters, digits, and underscores (_), but cannot start with underscores (_).
Description	A description of the field to be added.

3. Click **Create** on the **Relationships** tab page to create a relationship.

   A relationship refers to the logical relationship between two entities. 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,

 Copyright © Huawei Technologies Co., Ltd.

the actual business rules cannot be accurately described in the data model, and data consistency is greatly damaged.

Figure 7-41 Creating a relationship

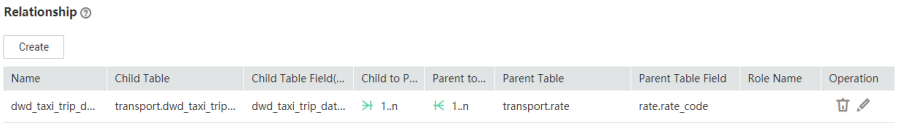
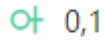


Table 7-25 Parameters

Parameter	Description
Name	Each relationship has a name to describe the function of the relationship.
Child Table	Click this field to select a table from the drop-down list. Click to set the current table as a child table.
Child Table Field	Select fields in the child table.
Child to Table	indicates that each piece of data in the child table corresponds to only one piece of data in the parent table.  indicates that each piece of data in the child table corresponds to one piece of data at most in the parent table. indicates that one piece of data in the child table corresponds to multiple pieces of data in the parent table. indicates that each piece of data in the child table corresponds to at least one piece of data in the parent table.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Parent to Child	1 indicates that each piece of data in the parent table corresponds to only one piece of data in the child table. 0,1 indicates that each piece of data in the parent table corresponds to one piece of data at most in the child table. 0..n indicates that one piece of data in the parent table corresponds to multiple pieces of data in the child table. 1..n indicates that one piece of data in the parent table corresponds to one piece of data in the child table at least.
Parent Table	Select a table that has a logical relationship with the selected child table.
Table Field	Select fields in the parent table.
Role Name	You can customize a role name to identify the relationship.
Operation	Click to delete a relationship. Click to edit the relationship.

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 table in the list.

----End

Converting a Logical Model to a Physical Model

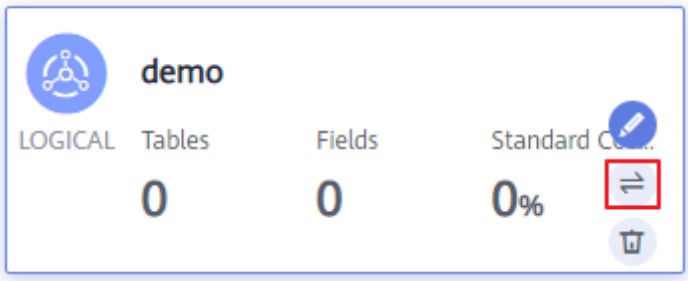
After a logical model is created, you can convert it to a new physical model or an existing physical model.

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Design** of the corresponding workspace to access the Data Design console.
3. Choose **ER Modeling** from the left navigation bar.
4. Find the required logical model and click the conversion button on the model.

 Copyright © Huawei Technologies Co., Ltd.

Figure 7-42 Logical model conversion



5. In the **Convert to Physical Model** dialog box, set the parameters in the following figure and click **OK**.

Figure 7-43 Convert to Physical Model dialog box

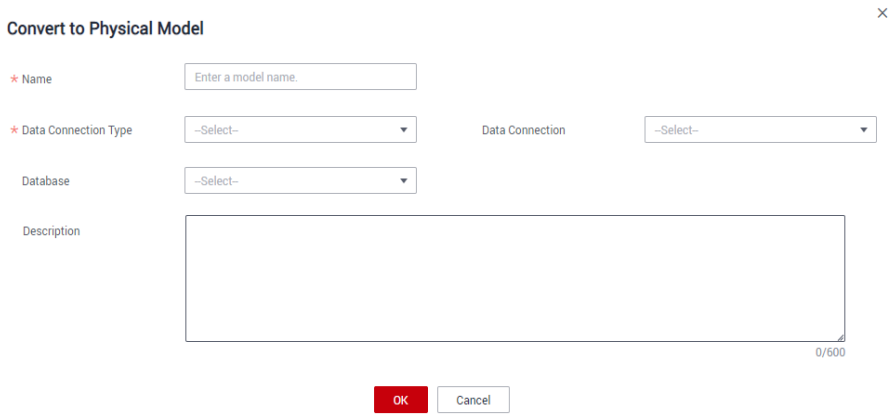


Table 7-26 Parameters

Parameter	Description
Name	The name of the physical model to be converted from a logical model. You can enter a new model name. The system creates the model. You can also select an existing model name from the drop-down list. Only letters, digits, and underscores (_) are allowed.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Update Existing Table	This parameter is displayed only when **Name** is set to an existing model name. It indicates whether to update the tables in the physical model if the tables in the logical model already exist in the target physical model during the model conversion. The system determines whether the table to be converted exists in the physical model based on the table code. When converting a logical model, you can only create or update the existing tables, but cannot delete them. • **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 need to publish the table again after updating it so that the updated table can take effect.
Data Connection Type	Select a data connection type from the drop-down list. If the required data type does not exist, you can add one by referring to **Managing a Field Type**.
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 you have not created a data connection yet, access to the DGC 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. If you have not created a database yet, access to the DGC Data Development console to create one. For details, see **Creating a Database**.
Queue	DLI queue. This parameter is displayed only for DLI data connections.
Schema	This parameter is displayed for DWS data connections.
Description	A description of the model. Up to 600 characters are supported.

7.9.3 Reversing Databases

You can import tables from databases of other data sources to a specified ER model (except the OBS type).

Prerequisites

Before reversing databases, collect data assets from databases in DGC Data Assets so that the system can synchronize tables to DGC Data Assets later. Otherwise, the synchronization tasks may fail. See **Task Management** for details.

Copyright © Huawei Technologies Co., Ltd.

Importing a Table to a Specified Model by Reversing the Database

Step 1 On the DGC Data Design console, choose **ER Modeling** from the left navigation bar.

Step 2 In the ER model tree, select a model, expand it, and select a sub-item. On the displayed page, click **Reverse Database** in the upper left corner.

Step 3 In the **Reverse Database** dialog box, set the parameters in the figure below:

Figure 7-44 Reverse Database dialog box

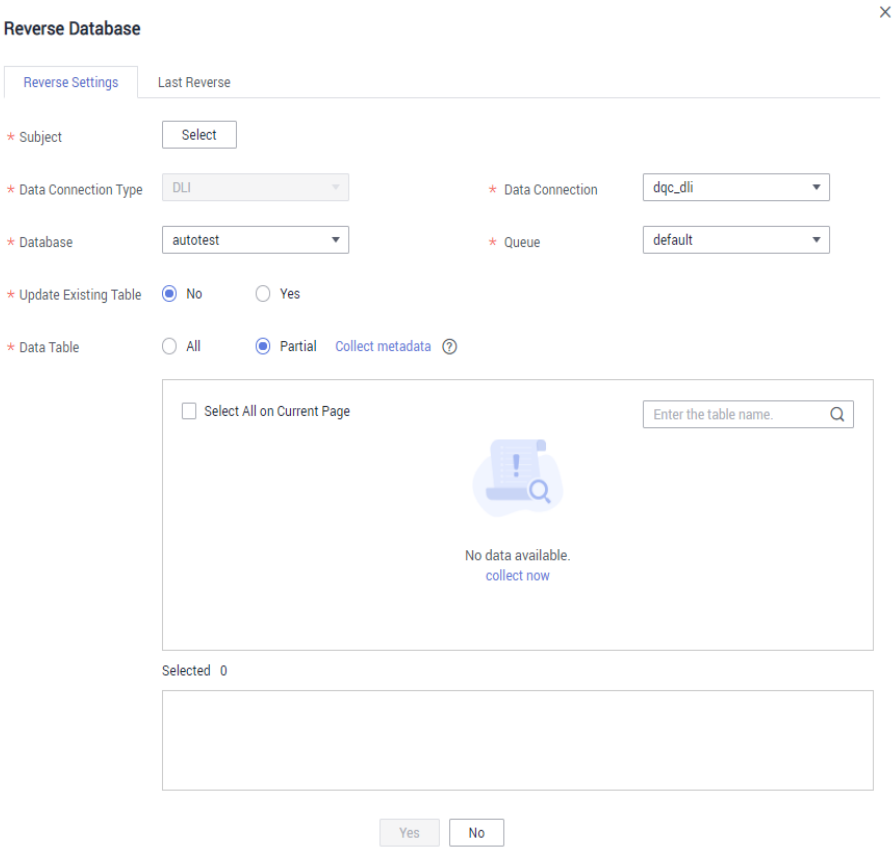


Table 7-27 Parameters

Parameter	Description
Subject	Click **Select** to select a subject.
Data Connection Type	If you reverse tables to a logical model, select a required data connection type from the drop-down list. If you reverse tables to a physical model, the data connection type of the current model is displayed.

 Copyright © Huawei Technologies Co., Ltd.

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 need to create a data connection in the DGC 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.
Queue	This parameter is displayed only for the DLI data connections. Select a DLI queue.
Schema	This parameter is displayed only for DWS data connections. You need to set **Data Connection Type** to **DWS**.
Update Existing 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 its code. During the import, only table creation and update are allowed. • **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 need to 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

7.9.4 Associating with Quality Rules

After creating and publishing a table in an ER model (except the OBS type), you can associate quality rules with the table. If **Create Quality Jobs** is selected in **Model Design Process** on the **Function Configuration** tab of **Configuration Center**, after a quality rule is associated, a quality job is automatically created in DGC Data Quality Control after the table is published. If the table has been published, the system automatically updates the corresponding quality job.

Associating with a Quality Rule and Viewing a Quality Job

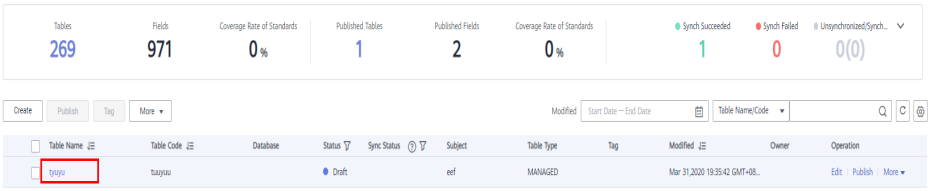
Step 1 On the DGC Data Design console, choose **ER Modeling** from the left navigation bar.

 Copyright © Huawei Technologies Co., Ltd.

Step 2 On the displayed page, select a model other than the OBS type. All tables created in the model are listed on the right. You can also expand a model 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 go to the table details page.

Figure 7-45 ER model list



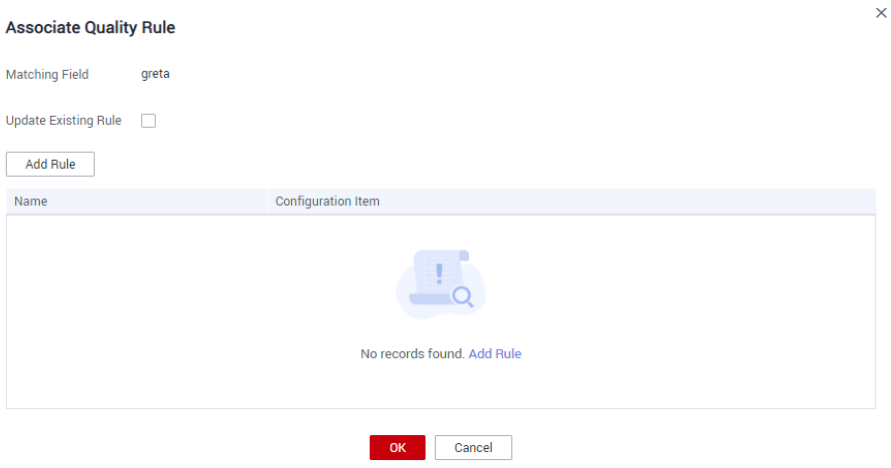
Step 4 On the displayed page, click the **Standards and Rules** tab. In the field list, select a field that you want to associate a quality rule with and then click **Associate Rule**.

Figure 7-46 Standards and Rules tab

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 displayed dialog box, click **Add Rule**.

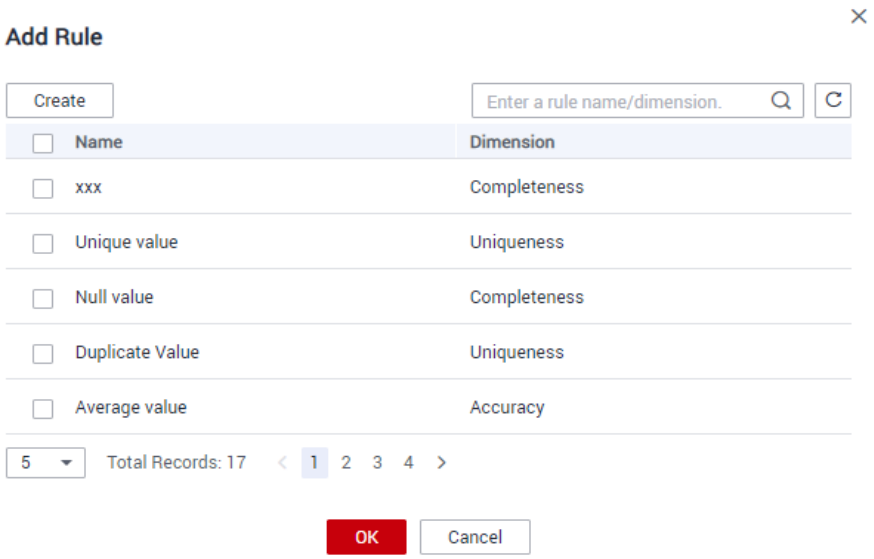
Figure 7-47 Associating with quality rules



In the **Add Rule** dialog box lists all default quality rules supported by DGC Data Quality Control. 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 DGC Data Quality Control. On the displayed page, create a rule. See **Creating Rule Templates**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 7-48 Add Rule dialog box



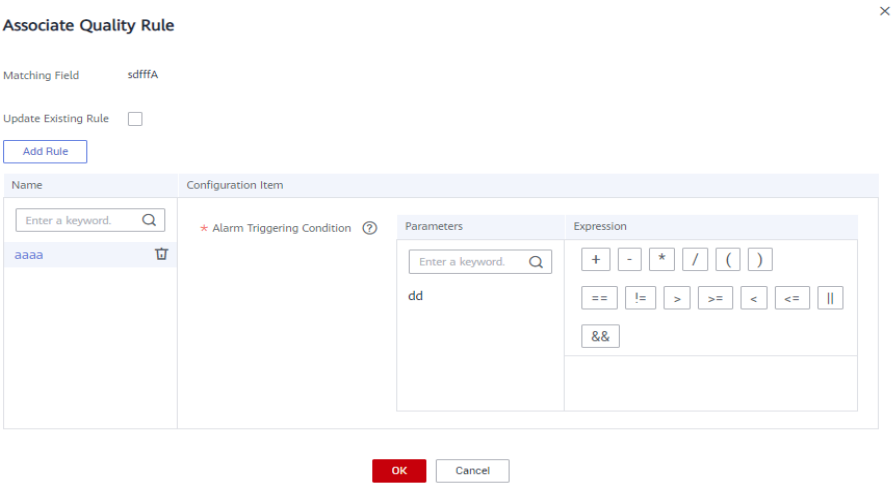
After a rule is added, go back to the **Associate Quality Rule** dialog box. Select a rule from the rule name list, set **Alarm Triggering Condition**, and click **OK**.

- In the **Alarm Triggering Condition** text box, enter an expression. When a quality job is running, the system will calculate the result of the alarm condition expression and check whether to trigger an 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 name indicates the alarm parameter. In other words, when setting **Alarm Triggering Condition**, use the variable $1 to represent the first alarm parameter, $2 to represent the second alarm parameter, and so on.

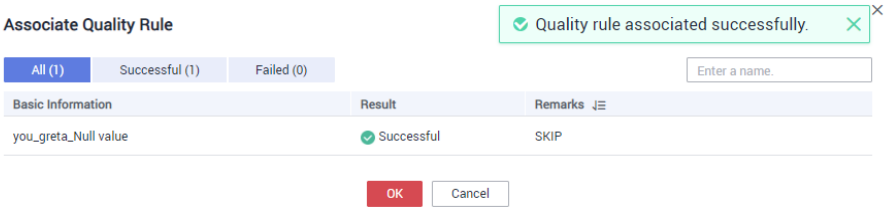
 Copyright © Huawei Technologies Co., Ltd.

Figure 7-49 Setting an alarm triggering condition



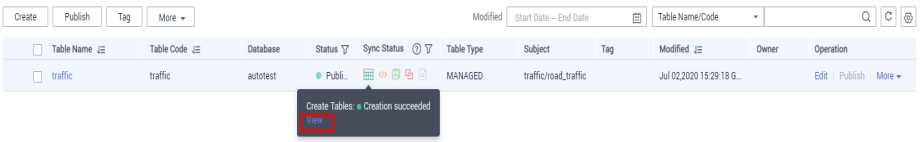
Step 6 View the association result. If the association is successful, click **OK**. If the association fails, find the failure cause, correct it, and associate the quality rule again.

Figure 7-50 Association results



Step 7 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**.

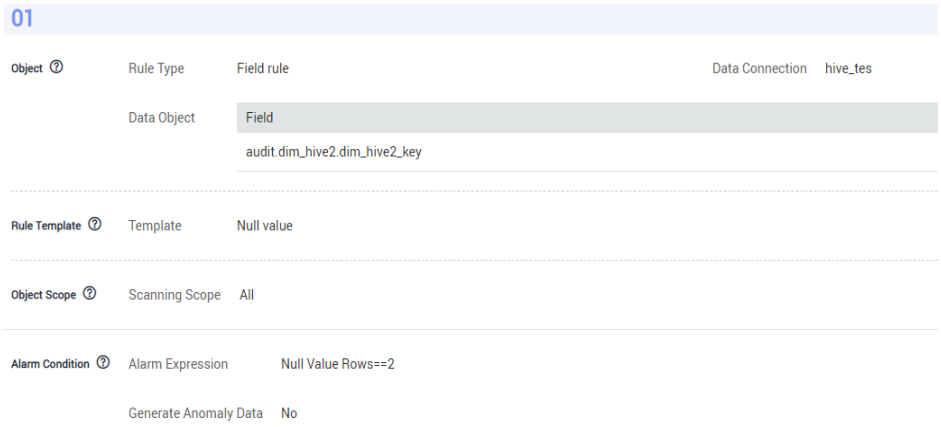
Figure 7-51 Quality job sync status



Step 8 On the displayed page, choose the **Rule Configurations** tab to the rule you just added.

 Copyright © Huawei Technologies Co., Ltd.

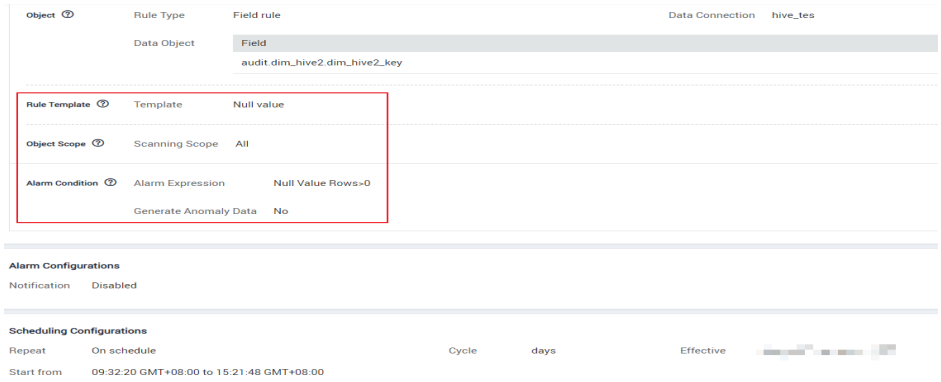
Figure 7-52 Quality rules



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:

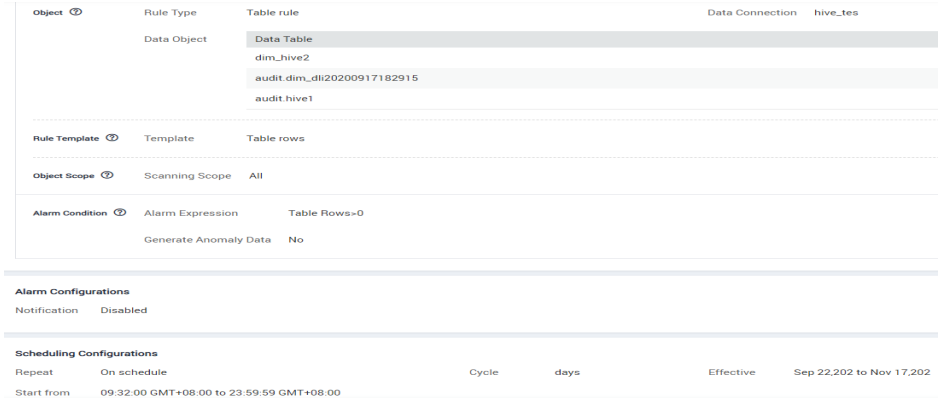
Figure 7-53 Quality rules for data standards



The following provides an example of the quality rule generated based on the data standard associated with a lookup table:

 Copyright © Huawei Technologies Co., Ltd.

Figure 7-54 Quality rules for data standards



----End

7.9.5 Importing/Exporting Tables

You can import tables to an ER model in batches. You can also export existing ER models and import the exported model files to other ER models.

The following parts are included in this section:

- **Importing a Table to a Logical Model**
- **Importing a Table to a Physical Model**
- **Exporting a Table/DDL**

Importing a Table to a Logical Model

Step 1 On the DGC Data Design console, choose **ER Modeling** from the left navigation bar.

Step 2 In the ER model tree, select a logical model, expand it, and select a target. Then, choose **More** > **Import**.

Step 3 In the displayed dialog box, click **ER Modeling Template**.

Table 7-28 Parameters for importing tables

Parameter	Description
Update Existing Table	Whether to update the existing table information. The system determines whether the table to be imported exists in the ER model based on the table code. During the import, only table creation and update are allowed. The options are as follows: • **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 need to publish the table again after updating it so that the updated table can take effect.

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Upload File	Select the file to be imported. You can use either of the following methods to obtain the file to be imported: • **Downloading the ER modeling template and fill in the template** In the **Import Table** dialog box, click **ER Modeling Template** to download the template, set parameters in the template as required, and save the settings. • **Exporting tables to files** You can export the tables created in Data Design of a DGC instance to Excel files in advance. Then, import the Excel files. See **Exporting a Table/DDL** for details.

Step 4 Open the downloaded template, set parameters in the template based on actual requirements, and save the template. The "Description" sheet in the template is for reference only.

Parameters whose names start with * are mandatory, and those that do not start with * are optional.

The table below describes the parameters in the **Table Model** sheet.

Table 7-29 Parameters in the "Table Model" sheet

Parameter	Parameter
Subject Area	Enter the encoding paths of existing subjects, which are separated by slashes (/). If you have not created a subject, create one by referring to **Designing Subjects**.
*Logical Entity	The name of the table. Only letters, digits, and underscores (_) are allowed. Table names cannot start with digits.
*Table Name	The code of the table. Only letters, digits, underscores (_), $, {, and } are allowed. Table codes cannot start with digits.
Alias	Alias of a table. This parameter is displayed when a user selects the table alias option in the configuration center. The value can include only letters, numbers, underscores (_), dollar signs ($), left braces ({), and right braces (}) and cannot start with a number.
Attribute Alias	Alias of an attribute. This parameter is displayed when a user selects the attribute alias option in the configuration center. The value can include only letters, numbers, underscores (_), dollar signs ($), left braces ({), and right braces (}) and cannot start with a number.
Table Tag	Tag to be added to a table. Enter an existing tag or a new tag.. You can add a tag on the **Tag** page of DGC Data Assets module and then go back to add this tag. For details on how to add a tag, see **Tags**.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Parameter
Description	A description of the table.
Asset Owner	Enter the username for entering the current workspace. Only the workspace administrator or developer can be set as the owner.
Parent Table	You can enter only the names of other tables in this template.
*Attribute Name	The name of the attribute field in a table. Only letters, digits, brackets, spaces, hyphens (-) and underscores (_) are allowed. Table names must start with letters.
Attribute Code	The name of the OBS folder. Only letters, digits, brackets, spaces, hyphens (-) and underscores (_) are allowed. Attribute names must start with letters.
No.	Sequence number of an attribute field in a table. The value starts from 1. This parameter is optional. If this parameter is left blank, attribute fields are sorted in the sequence in the template by default.
Description	A description of an attribute field.
*Data Type	Data type of the logical model. For details, see the DEFAULT group in **Managing a Field Type**.
Data Length	Data length. For a variable-length data type, specify the data length if the specified data connection type supports the specified data length. For example, for the DWS connection type, if the field type is **CHAR(10)**, set **Data Type** to **CHAR** and **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.
Referenced Data Standard Code	The code of the data standard to be referenced. This field can be left blank. If you have not created a data standard, create one. See **Data Standards** for details.
Attribute Tag	Tag to be added to an attribute field. Enter an existing tag or a new tag.. You can add a tag on the **Tag** page of DGC Data Assets module and then go back to add this tag. For details on how to add a tag, see **Tags**.

Table 7-25 describes the parameters in the **Relationships** sheet.

The table below describes the parameters in the **Mappings** sheet.

 Copyright © Huawei Technologies Co., Ltd.

Table 7-30 Parameters in the "Mappings" sheet

Parameter	Parameter
Mapping	Only letters, digits, and underscores (_) are allowed.
Mapping Description	A description of the mapping.
*Table	Enter the name of an existing table.
*Table Field	The code of the mapping field.
Source Table Name	The name of the table from which data is obtained. You need to enter the name of an existing table.
Source Table Field Code	Enter the code of the source table field.
Field Mapping Description	A description of the field.

Step 5 View the result in 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 Table to a Physical Model

Step 1 On the DGC Data Design console, choose **ER Modeling** from the left navigation bar.

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 displayed dialog box, click **ER Modeling Template**.

Table 7-31 Parameters for importing tables

Parameter	Description
Update Existing Table	Whether to update the existing table information. The system determines whether the table to be imported exists in the ER model based on the table code. During the import, only table creation and update are allowed. The options are as follows: • **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 need to publish the table again after updating it so that the updated table can take effect.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Upload File	Select the file to be imported. You can use either of the following methods to obtain the file to be imported: • **Downloading the ER modeling template and fill in the template** In the **Import Table** dialog box, click **ER Modeling Template** to download the template, set parameters in the template as required, and save the settings. • **Exporting tables to files** You can export the tables created in Data Design of a DGC instance to Excel files in advance. Then, import the Excel files. See **Exporting a Table/DDL** for details.

Step 4 Open the downloaded template, set parameters in the template based on actual requirements, and save the template. The "Description" sheet in the template is for reference only.

Parameters whose names start with * are mandatory, and those that do not start with * are optional.

The table below describes the parameters in the **Table Model** sheet.

Table 7-32 Parameters in the "Table Model" sheet

Parameter	Description (Importing DLI/DWS/MRS_HIVE Tables)
Subject Area	Enter the encoding paths of existing subjects, which are separated by slashes (/). If you have not created a subject, create one by referring to **Designing Subjects**.
*Logical Entity	The name of the table. Only letters, digits, and underscores (_) are allowed. Table names cannot start with digits.
*Table Name	The code of the table. Only letters, digits, underscores (_), $, {, and } are allowed. Table codes cannot start with digits.
Table Tag	Tag to be added to a table. Enter an existing tag or a new tag.. You can add a tag on the **Tag** page of DGC Data Assets module and then go back to add this tag. For details on how to add a tag, see **Tags**.
Description	A description of the table.
Asset Owner	Enter the username for entering the current workspace. Only the workspace administrator or developer can be set as the owner.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description (Importing DLI/DWS/MRS_HIVE Tables)
Data Connection Type	The following connection types are supported: DLI, DWS and MRS_HIVE.
*Table Type	DLI models support the following table types: • **Managed** indicates that data is stored in a DLI table. • **External** indicates that data is stored in an OBS table. When **Table Type** is set to **External**, you need to set **OBS Path**. • **DLI_VIEW** is available for import only. DWS models support the following table types: • DWS_ROW: row type • DWS_COLUMN: column type • DWS_VIEW: view type This parameter is unavailable for the tables created in MRS_HIVE models.
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 ***bucket_name/filepath***.
Data Format	This parameter is available only for tables created in DLI models. If the table type is "DLI_MANAGED", the options of the data format are **Parquet** and **Carbon**. If the table type is "DLI_EXTERNAL", the options of the data format are **Parquet**, **Carbon**, **CSV**, **ORC**, **JSON**, and **Avro**.
Data Connection	Enter the name of a created data connection.
Database	Enter the name of a created database.
Connection Extra	If the connection type is DLI, enter a DLI queue name. If the connection type is DWS, enter a DWS schema name.
*Attribute Name	The name of the attribute field in a table. Only letters, digits, brackets, spaces, hyphens (-) and underscores (_) are allowed. Table names must start with letters.
Attribute Code	The name of the OBS folder. Only letters, digits, brackets, spaces, hyphens (-) and underscores (_) are allowed. Attribute names must start with letters.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description (Importing DLI/DWS/MRS_HIVE Tables)
No.	Sequence number of an attribute field in a table. The value starts from 1. This parameter is optional. If this parameter is left blank, attribute fields are sorted in the sequence in the template by default.
Description	A description of an attribute field.
*Data Type	The supported data types vary depending on the data connection types. For details, see **Managing a Field Type**.
Data Length	For a variable-length data type, specify the data length if the specified data connection type supports the specified data length. For example, for the DWS connection type, if the field type is **CHAR(10)**, set **Data Type** to **CHAR** and **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.
Referenced Data Standard Code	The code of the data standard to be referenced. This field can be left blank. If you have not created a data standard, create one. See **Data Standards** for details.
Attribute Tag	Tag to be added to an attribute field. Enter an existing tag or a new tag.. You can add a tag on the **Tag** page of DGC Data Assets module and then go back to add this tag. For details on how to add a tag, see **Tags**.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description (Importing DLI/DWS/MRS_HIVE Tables)

Other Configurations	Additional table configuration details stored in JSON format. The format is as follows:
	{
	"option_name1": "value",
	"option_name2": "value"
	.....
	}
	Examples:
	{
	"a1": "100",
	"a2": "30"
	}

Version	This parameter is optional.

Table 7-19 describes the parameters in the **Relationships** sheet.

The table below describes the parameters in the **Mappings** sheet.

Table 7-33 Parameters in the "Mappings" sheet

Parameter	Parameter
Mapping	Only letters, digits, and underscores (_) are allowed.
Mapping Description	A description of the mapping.
*Table	Enter the name of an existing table.
*Table Database	Enter the name of the database that the mapping table belongs to.
*Table Field	The code of the mapping field.
Source Table Name	The table from which data is obtained. You need to enter the name of an existing table. All source tables in a mapping must come from the same model.
Source Table Database	Enter the database that the source table belongs to.
Source Table Model	Enter the name of the ER model that the source table belongs to. If this parameter is not set, the current model is used by default.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Parameter
Source Table Field Code	Enter the code of the source table field.
Field Mapping Description	A description of the field.

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/DDL

Step 1 On the DGC Data Design console, choose **ER Modeling** from the left navigation bar.

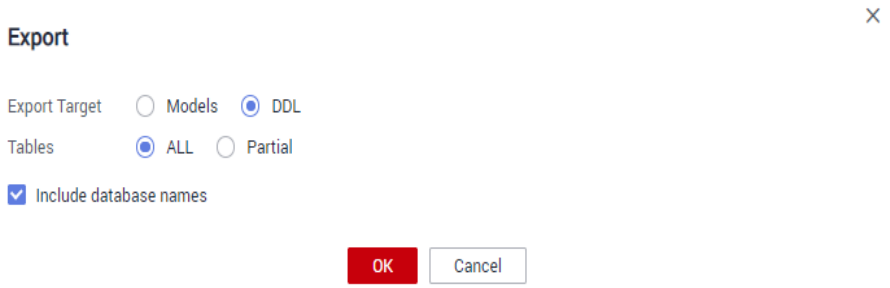
Step 2 In the ER model tree, select a model, right-click the model, and choose **More > Export** from the short-cut menu.

Step 3 In the displayed dialog box, select the objects to be exported.

The exported Excel file can also be imported.

When a DDL is exported, the DDL statements of the selected table are exported to TXT files.

Figure 7-55 Exporting a DDL



Step 4 Click **OK**.

----End

7.9.6 Viewing Tables

Tables in an ER model (except the OBS type) can be displayed in the model view or list view. You can view table details, relationship diagrams, and publish history, as well as preview SQLs.

Copyright © Huawei Technologies Co., Ltd.

Querying the Model View

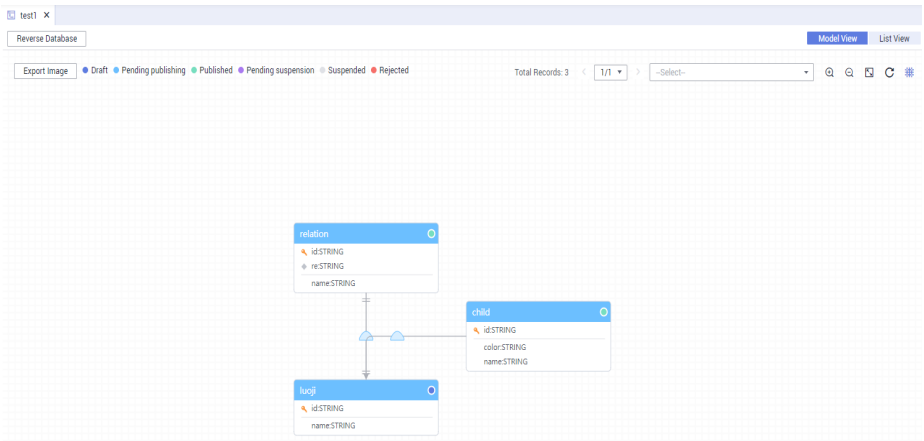
After creating a table in an ER model (except the OBS type), you can query the table models in the list view or model view. The created tables are displayed in the list view by default. You can switch to the model view if you like.

Step 1 On the DGC Data Design console, choose **ER Modeling** from the left navigation bar.

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. Or, click **List View** to switch back to the table list.

Figure 7-56 Model view



The following functions are supported in the model view:

- Double-click a table name to view the table details.
- Click **Export Image** 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 a SQL Statement

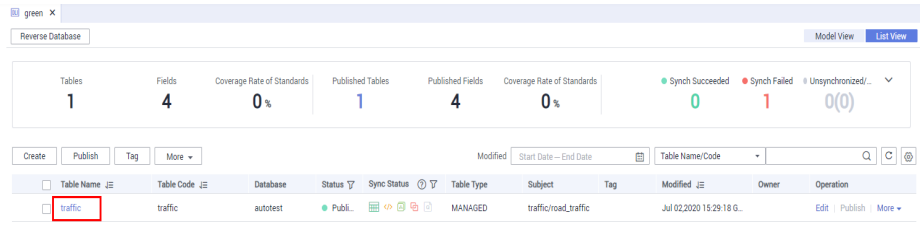
Step 1 On the DGC Data Design console, choose **ER Modeling** from the left navigation bar.

Step 2 On the displayed page, select a model other than the OBS type. All tables created in the model are listed on the right. You can also expand a model and select an object. All tables of the object are listed on the right.

 Copyright © Huawei Technologies Co., Ltd.

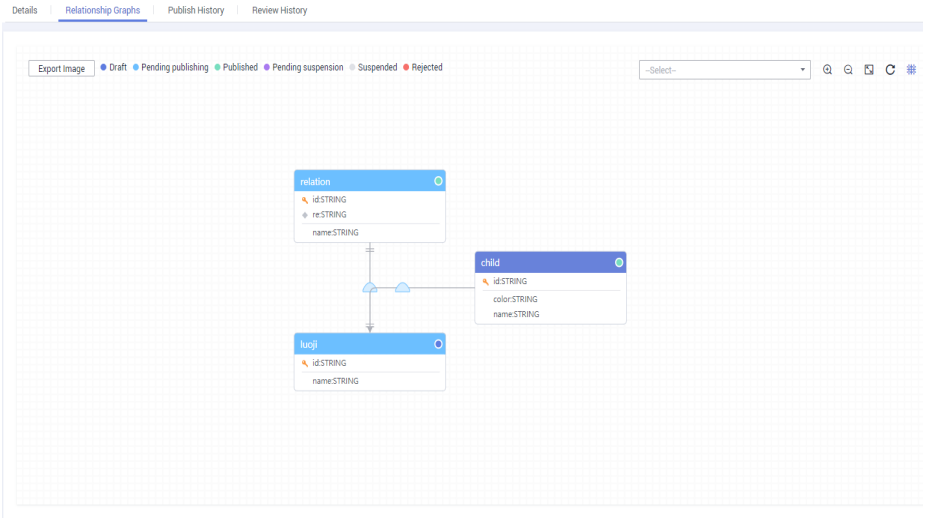
Step 3 In the table list, select a table, and choose **More** > **Preview SQL** in the **Operation** column to preview or copy the SQL statement. Then, click **OK** to return to the previous page.

Figure 7-57 ER model list



Step 4 In the table list, click a table name to go to the table details page and view the table details, relationship diagram, publish history, and review history.

Figure 7-58 Relationship diagram



----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.

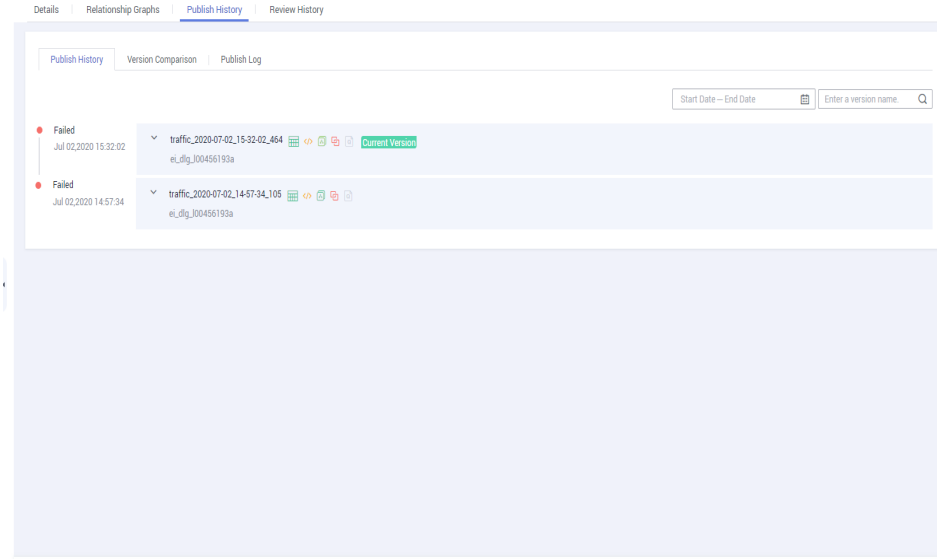
Step 1 On the DGC Data Design console, choose **ER Modeling** from the left navigation bar.

Step 2 On the displayed page, select a model. All tables created in the model are listed on the right. You can also expand a model and select an object. All tables of the object are listed on the right.

Step 3 In the table list, locate the target table, and choose **More** > **View Publish History**. On the displayed page, you can view the table publish history, version comparisons, and publish logs.

 Copyright © Huawei Technologies Co., Ltd.

Figure 7-59 Viewing publish history



----End

7.9.7 Modifying Subjects, Directories, and Processes

Modifying Subjects in Batches

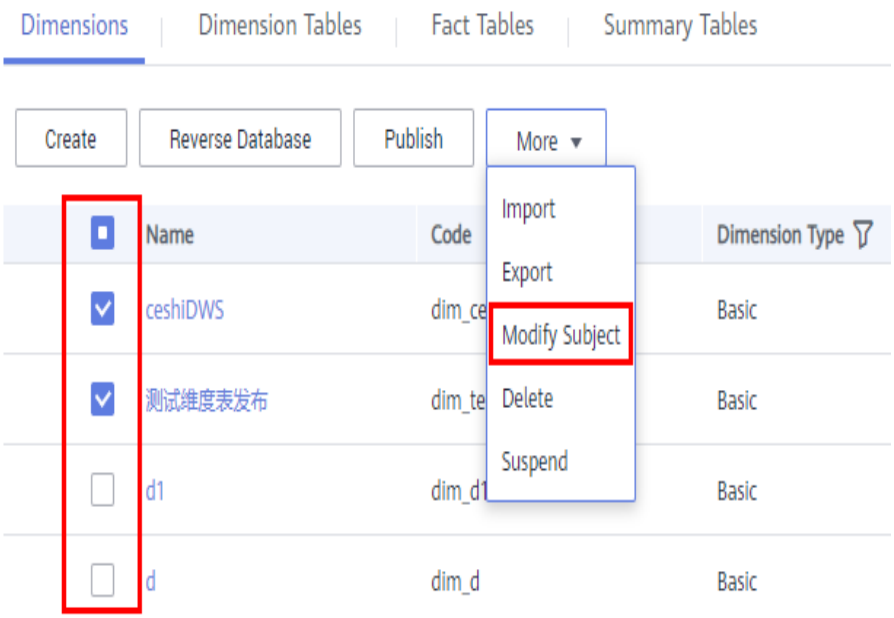
Currently, only subjects of information architectures, ER models, dimensions, fact tables, summary tables, and technical metrics can be modified in batches.

Step 1 On the DGC Data Design console, choose the module whose subjects you want to modify from the left navigation bar.

Step 2 On the displayed page, select the targets whose subjects need to be modified, and choose **More** > **Modify Subject**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 7-60 Modifying subjects



----End

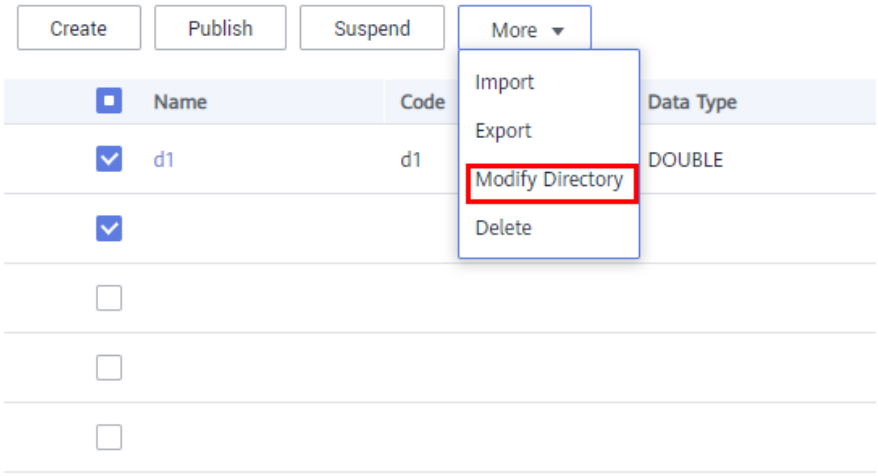
Modifying Directories in Batches

Currently, only directories of lookup tables and data standards can be modified in batches.

Step 1 On the DGC Data Design console, choose **Lookup Tables** or **Data Standards** from the left navigation bar.

Step 2 On the displayed page, select the targets whose directories need to be modified, and choose **More** > **Modify Directory**.

Figure 7-61 Modifying directories



----End

 Copyright © Huawei Technologies Co., Ltd.

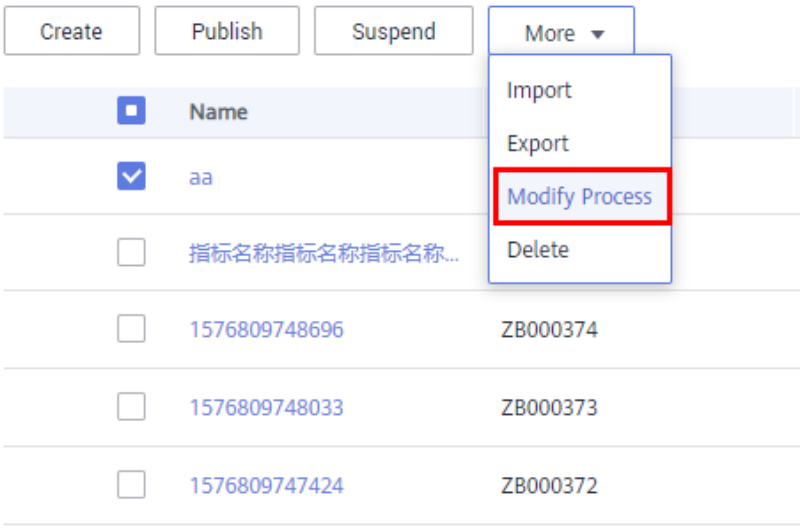
Modifying Processes in Batches

Currently, only the processes of business metrics can be modified in batches.

Step 1 On the DGC Data Design console, choose **Business Metrics** from the left navigation bar.

Step 2 On the displayed page, select the metrics whose processes to be modified and choose **More** > **Modify Process**.

Figure 7-62 Modifying processes



----End

7.10 Dimensional Modeling

7.10.1 Creating Dimensions

A dimension is a perspective to observe and analyze business data. It supports data aggregation, drilling, and slicing analysis and is used as the GROUP BY condition in SQL statements. Most dimensions have hierarchical structures, such as geographic dimensions (including countries, regions, provinces, 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.

Creating and Publishing a Dimension

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Design** of the corresponding workspace to access the Data Design console.

 Copyright © Huawei Technologies Co., Ltd.

3. Choose **Dimensional Modeling** from the left navigation bar and click the **Dimensions** tab.
4. Select an object from the subject tree on the left and click **Create** to create a dimension.

   Before creating a dimension, if you have not added any subject information, add it. See **Designing Subjects**.
5. On the displayed page, set parameters as prompted.

   Set the basic configurations and physicalization configurations shown in the following figure:

Figure 7-63 Creating a dimension

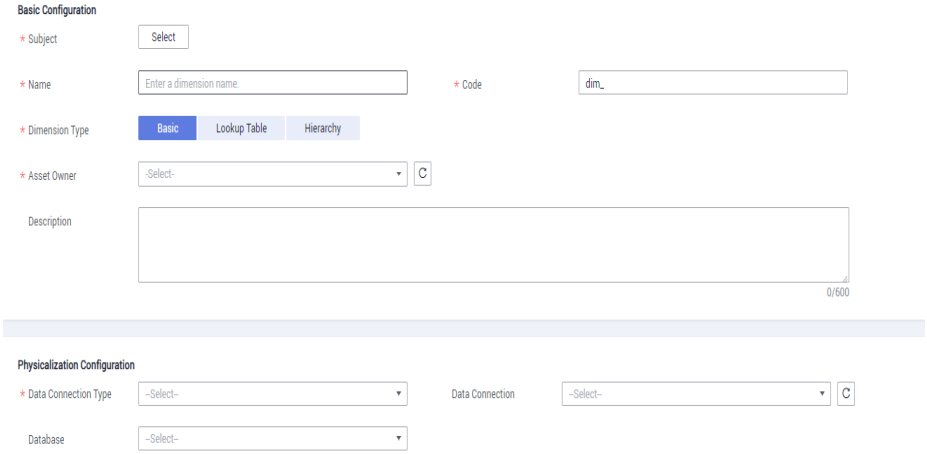


Table 7-34 Parameters of **Basic Configuration**

Parameter	Description
Subject	Click **Select**. In the dialog box that is displayed, select the subject you need, and click **OK**.
Name	Only letters, digits, and underscores (_) are allowed. Dimension names must start with letters.
Code	The code of the dimension to be created. The system automatically synchronizes the dimension name to the **Code** text box. Only letters, digits, and underscores (_) are allowed. Dimension codes must start with letters.
Dimension Type	• **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**: dimension with a hierarchical structure between attributes.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Asset Owner	Select the asset owner that the dimension belongs to from the drop-down list. Only the workspace administrator or developer can be set as the owner.
Description	A description of the dimension to be created. Up to 600 characters are supported.

Table 7-35 Parameters of **Physicalization Configuration**

Parameter	Description
Data Connection Type	Select a data connection type from the drop-down list.
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 you have not created a data connection yet, access to the DGC 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. If you have not created a database yet, access to the DGC Data Development console to create one. For details, see **Creating a Database**.
Queue	DLI queue. This parameter is displayed only for DLI data connections.
Schema	This parameter is displayed for DWS data connections.
Table Type	This parameter is displayed only for DWS data connections. The options are: • **DWS_ROW** indicates that tables are stored to disk partitions by row. • **DWS_COLUMN** indicates that tables are stored to disk partitions by column.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
DISTRIBUTE BY	This parameter is displayed only for DWS data connections. • **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 need to 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 attributes in the **Attribute Configuration** area. You can click **Add** to add multiple dimension attributes.

Table 7-36 Parameters of **Attribute Configuration**

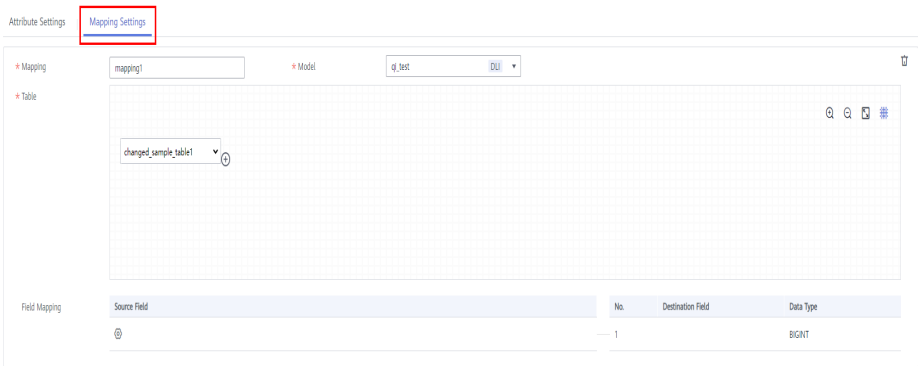
Parameter	Description
Name	Only letters, digits, and underscores (_) are allowed. Attribute names must start with letters.
Code	Only letters, digits, and underscores (_) are allowed. Attribute codes must start with letters.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Data Standard	Click to select a data standard to be associated with the field. Choose **Configuration Center** > **Function Configuration** and select **Create Quality Jobs**. Associate fields with a data standard. After a dimension table is published, a quality job is automatically generated. 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 go to the **Quality Job** page of the DGC Data Quality Control to view the job details. If you click **Associate with Data Standard** above the field list and select a data standard that you want to associate the field with, a new table field associated with the selected data standard is automatically created. The default field name and code are the same as those of the associated data standard. You can change the field name and code if necessary. If you have not created a data standard, create one. See **Data Standards** for details.
Data Type	Type of data defined based on the original data.
Primary Key	Select a proper field as the primary key based on service requirements. By default, the first dimension attribute is the primary key.
Surrogate Key	Select a proper field as the surrogate key based on service 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 attribute you add.

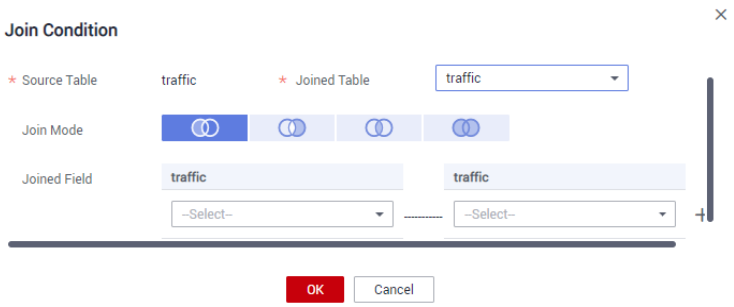
On the **Mapping Settings** tab page, click **Create** to create the mapping between dimensions and fact tables. Set the parameters shown the figure below.

Figure 7-64 Mapping settings



 Copyright © Huawei Technologies Co., Ltd.

Table 7-37 Parameters of mappings

Parameter	Description
Mapping	Only letters, digits, 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	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 another table. 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. **Figure 7-65 Join Condition** dialog box 
Field Mapping	Select a source field with the same meaning as the current mapping field. If a table field comes from multiple models, you need to 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 specified mapping or click to collapse the mapping area.

6. Click **Save and Publish**.
7. In the displayed dialog box, select a reviewer and click **Confirm**.

 Copyright © Huawei Technologies Co., Ltd.

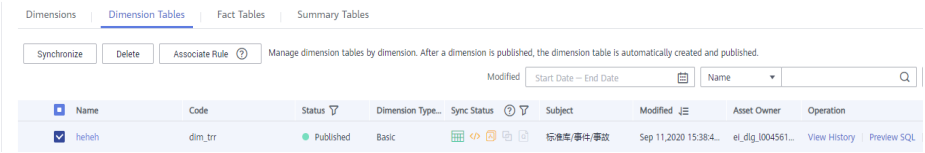
8. Repeat steps 4 to 7 to create and publish other dimensions.
9. All dimensions need to be approved by reviewers.

   After a dimension publishing 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.

   - 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 Publish History** in the row where the dimension table is located. On the displayed page, click the **Publish Log** 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.

Figure 7-66 Synchronization status



Editing a Dimension

1. On the Data Design console, choose **Dimensional Modeling** from the left navigation bar. On the displayed page, click the **Dimensions** tab.
2. In the dimension list, select the dimension you want to edit and click **Edit** in the **Operation** column. The **Edit Dimension** page is displayed.
3. Edit the dimension information as required.
4. Click **Save** to save the settings. Alternatively, click **Save and Publish** to publish the edited dimension.

Publishing a Dimension

If a dimension is created but not published, perform the following steps to publish the dimension:

Step 1 On the Data Design console, choose **Dimensional Modeling** from the left navigation bar. On the displayed page, click the **Dimensions** tab.

Step 2 In the dimension list, find the dimension you want to publish and click **Publish** in the **Operation** column.

Step 3 In the displayed dialog box, select a reviewer and click **Confirm**.

----End

 Copyright © Huawei Technologies Co., Ltd.

Suspending a Dimension

To bring a published dimension offline, perform the following steps:

Step 1 On the Data Design console, choose **Dimensional Modeling** from the left navigation bar. On the displayed page, click the **Dimensions** tab.

Step 2 In the dimension list, find the dimension you want to suspend and click **Suspend** in the **Operation** column.

Step 3 In the displayed dialog box, select a reviewer and click **Confirm**.

----End

Deleting a Dimension

If a dimension is no longer needed, you can delete it. However, if the dimension has been published, you must bring the dimension offline before deleting it. For details, see **Suspending a Dimension**.

1. On the Data Design console, choose **Dimensional Modeling** from the left navigation bar. On the displayed page, click the **Dimensions** tab.
2. Select the dimension you want to delete from the dimension list and click **Delete** in the upper left.
3. In the **Delete Dimension** dialog box, confirm whether the information is correct. If there is no need of modification, click **OK**.

   After you select Delete **physical tables** in the dialog box, the physical tables in the database are also deleted when you delete the table.

7.10.2 Managing Dimension Tables

A dimension table corresponds to a dimension and consists of a wide range of dimension attributes. Creating, publishing, editing, and suspending a dimension table highly relates 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 Data Design console, choose **Dimensional Modeling** from the left navigation bar.
2. Click the **Dimension Tables** tab.

   Figure 7-67 Dimension Tables page

3. Select a dimension table in the list and choose **More** > **View Publish History**.
4. On the displayed page, you can view the publish history, version comparison information, and publish log of the dimension table.

 Copyright © Huawei Technologies Co., Ltd.

If the publish log includes error logs, the publishing has failed. You can click **Resynchronize** to synchronize the table to other modules of DGC.

Previewing SQL

1. On the Data Design console, choose **Dimensional Modeling** from the left navigation bar.
2. Click the **Dimension Tables** tab.

   Figure 7-68 Dimension Tables page

3. Select a dimension table in the list and choose **More** > **Preview SQL**.
4. On the displayed page, 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

- During the synchronization, the system performs the synchronization based on the data table update mode on the **Configuration Center** > **Function Configuration** page. For details, see **Function Configuration**.

1. On the Data Design console, choose **Dimensional Modeling** from the left navigation bar.
2. Click the **Dimension Tables** tab.
3. In the dimension table list, select the dimension tables to be synchronized and click **Synchronize** above the list. The dialog box for synchronizing dimension tables in batches is displayed.
4. If there are no changes made to the information, click **Confirm**.

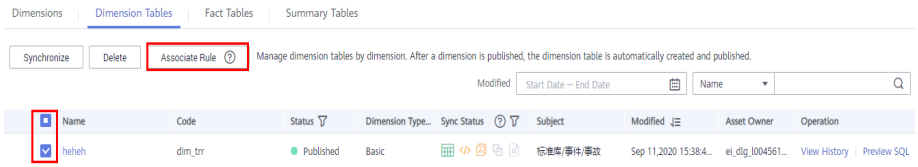
   After the synchronization, you can view the synchronization status of the dimension table in the dimension table list. Click the refresh button above the list to refresh the status.

Associating a Dimension Table with a Quality Rule

1. On the Data Design console, choose **Dimensional Modeling** from the left navigation bar.
2. Click the **Dimension Tables** tab.
3. In the dimension table list, click the name of the dimension table you want to associate with a quality rule. Click **Associate Rule**.

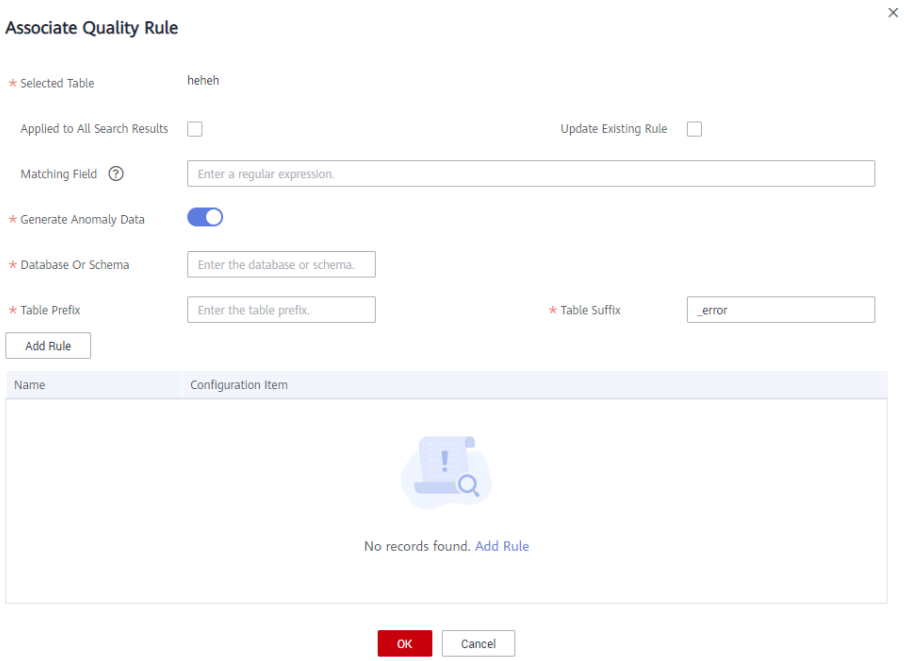
 Copyright © Huawei Technologies Co., Ltd.

Figure 7-69 Associating with a quality rule



4. On the displayed page, set parameters as prompt. After the configuration is completed, click **OK**.

Figure 7-70 Associating a dimension table with a quality rule



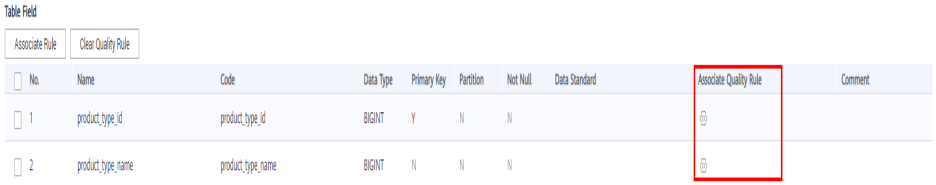
Generate Anomaly Data: If this option is selected, anomaly data is stored in the specified database based on the configured parameters.

Associating a Single Field with a Quality Rule

1. On the Data Design console, choose **Dimensional Modeling** from the left navigation bar.
2. Click the **Dimension Tables** tab.
3. In the dimension table list, click the name of the dimension table you want to associate with a quality rule.
4. In the table field list on the details page of the dimension table, find the field and click to associate the field with a quality rule.

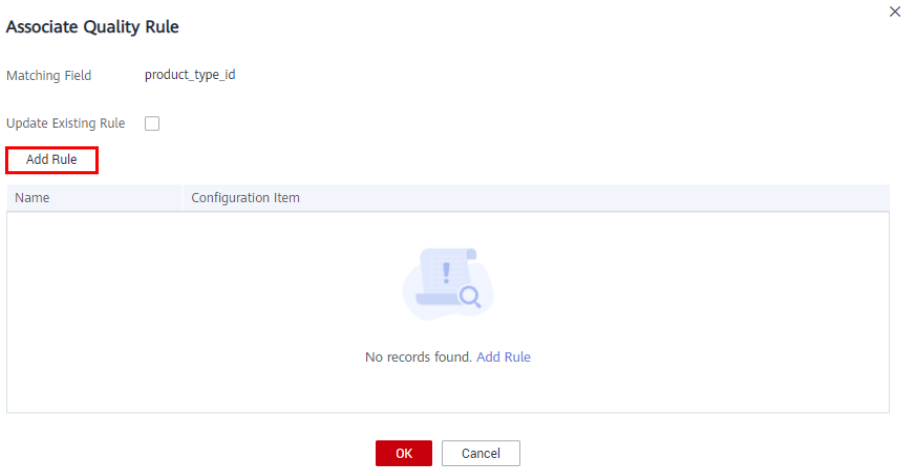
 Copyright © Huawei Technologies Co., Ltd.

Figure 7-71 Associating a single field with a quality rule



5. After the configuration is complete, click **OK**.

Figure 7-72 Adding a rule



Associating Table Fields with Quality Rules in Batches

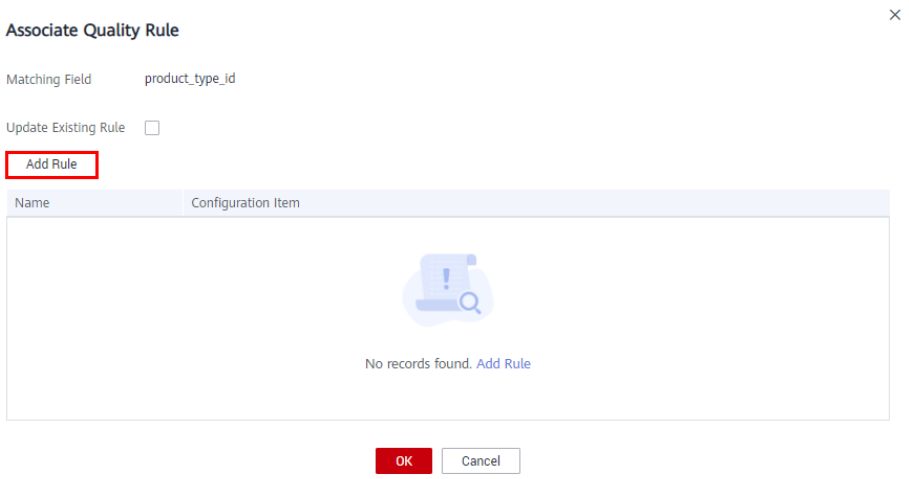
1. On the Data Design console, choose **Dimensional Modeling** from the left navigation bar.
2. Click the **Dimension Tables** tab.
3. In the dimension table list, click the name of the dimension table you want to associate with quality rules.
4. In the table field list on the details page of a dimension table, select the table fields you want to associate with quality rules and click **Associate Rule**.

Figure 7-73 Associating table fields with quality rules

5. On the page that is displayed, add rules and set rule parameters.

 Copyright © Huawei Technologies Co., Ltd.

Figure 7-74 Adding rules



6. After the configuration is complete, click **OK**.

Deleting a Dimension Table

Dimensions to be published, already published, or to be suspended cannot be deleted.

1. On the Data Design console, choose **Dimensional Modeling** from the left navigation bar.
2. Click the **Dimension Tables** tab.
3. In the dimension table list, select the dimension tables to be deleted and click **Delete** above the list. The dialog box for deleting dimension tables in batches is displayed.
4. If there are no changes made, click **OK**.

7.10.3 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. Log in to the DGC console, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Design** of the corresponding workspace to access the Data Design console.
3. Choose **Dimensional Modeling** from the left navigation bar. On the displayed page, click the **Fact Tables** tab.
4. Select a subject from the subject tree on the left and click **Create**.
5. On the **Create Fact Table** page, perform the following operations:

   a. Set the parameters in the **Basic Configuration** area.

 Copyright © Huawei Technologies Co., Ltd.

Figure 7-75 Basic Configuration area

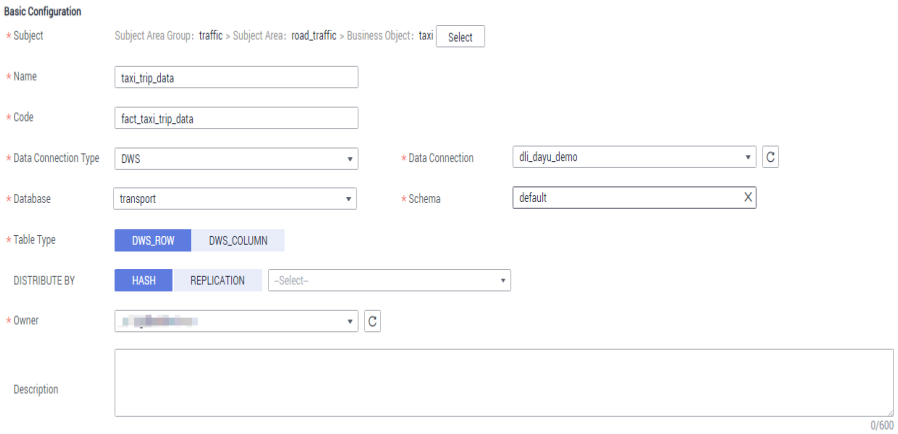


Table 7-38 Parameters of the **Basic Configuration** area

Parameter	Description
Subject	Click **Select** to select a subject catalog (subject area group > subject area > business object) where you can place the fact table.
Name	Table names must start with letters. Only letters, digits, and underscores (_) are allowed.
Code	Table codes must start with letters. Only letters, digits, and underscores (_) are allowed.
Data Connection Type	Select a data connection type from the drop-down list.
Data Connection	Select a data connection from the drop-down list. It is recommended that a same data connection be used for dimension modeling.
Database	Select a database from the drop-down list.
Queue	DLI queue. This parameter is displayed only for DLI data connections.
Schema	This parameter is displayed only for DWS data connections.
Table Type	This parameter is displayed only for DWS data connections. The options are: • **DWS_ROW** indicates that tables are stored to disk partitions by row. • **DWS_COLUMN** indicates that tables are stored to disk partitions by column.

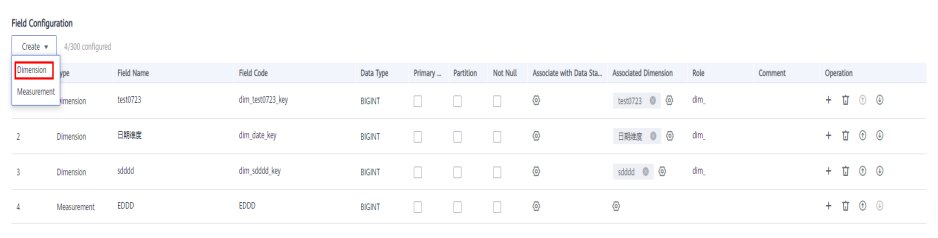
 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
DISTRIBUTE BY	This optional parameter is valid only for the DWS connection type. You need to add a table field before selecting a table field from the drop-down list as the **DISTRIBUTE BY** field. Currently, only **REPLICATION** and **HASH** are supported. • **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 need to 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	Select an asset owner from the drop-down list. Only the workspace administrator or developer can be set as the owner.
Description	A description of the fact table. Up to 600 characters are supported.

b. In the **Field Configuration** area, click **Create** to select one or more created dimensions. The primary key fields of the dimensions are then added to the list.

Click the **Up** or **Down** icon to adjust the rankings of dimensions.

Figure 7-76 Dimension Configuration area



 Copyright © Huawei Technologies Co., Ltd.

Table 7-39 Parameters of the **Dimension Configuration** area

Parameter	Description
Field Name	The primary key name of the added dimension is displayed automatically. Generally, you do not need to change the name. Field names must start with letters. Only letters, digits, and underscores (_) are allowed.
Field Code	The primary key code of the added dimension is displayed automatically. Generally, you do not need to change the code. Field codes must start with letters. Only letters, digits, and underscores (_) are allowed.
Data Type	The 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.
Associating with Data Standards	If you have created data standards, click to select one to associate with the field. Choose **Configuration Center** > **Function Configuration** and select **Create Quality Jobs**. Associate fields with a data standard. After a table is published, a quality job is automatically generated. 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 go to the **Quality Job** page of the DGC Data Quality Control to view the job details. If you have not created a data standard, create one. See **Data Standards** for details.
Associated Dimension	The name of the associated dimension. Click to replace the associated dimension.
Role	If a dimension is added multiple times, you need to set different roles to distinguish the dimensions. See **Figure 7-76**.
Description	A description of the dimension.

c. In the **Field Configuration** area, click **Create** to add a measurement.

 Copyright © Huawei Technologies Co., Ltd.

Figure 7-77 Configuring a measurement

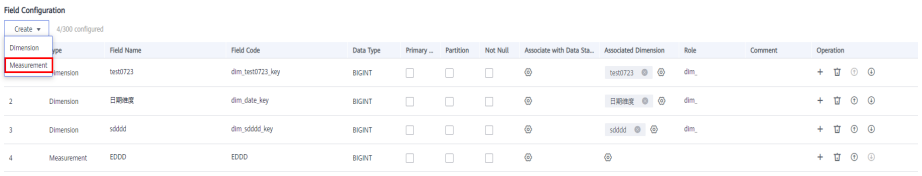


Table 7-40 Parameters of configuring a measurement

Parameter	Description
Field Name	Field names must start with letters. Only letters, digits, and underscores (_) are allowed.
Field Code	Field codes must start with letters. Only letters, digits, and underscores (_) are allowed.
Data Type	Field data type. If the required data type does not exist, you can add one. See **Managing a Field Type**.
Associating with Data Standards	If you have created data standards, click to select one to associate with the field. Choose **Configuration Center** > **Function Configuration** and select **Create Quality Jobs**. Associate fields with a data standard. After a table is published, a quality job is automatically generated. 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 go to the **Quality Job** page of the DGC Data Quality Control to view the job details. If you have not created a data standard, create one. See **Data Standards** for details.
Associated Dimension	The name of the associated dimension. Click to replace the associated dimension.
Description	A description of the measurement.

6. Click **Save and Publish** to apply for approval.
7. Wait for the reviewer to approve the fact table.

   After the fact table is approved, it is automatically created in the database.
8. 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 Publish History** in the row where the fact table is located. On the displayed page, click the

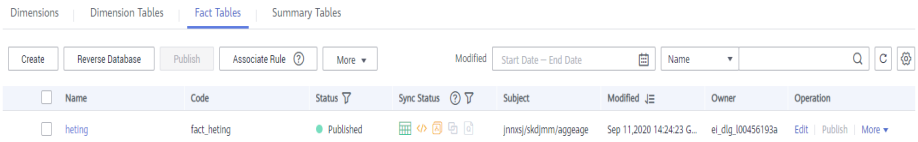
 Copyright © Huawei Technologies Co., Ltd.

Publish Log 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 Data Design. On the displayed page, you can edit, publish, suspend, and delete the fact table, as well as view the publish logs.

Figure 7-78 Fact table management



- **Editing a fact table**
  a. In the fact table list, select a fact table you want to edit 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 **Save and Publish** to publish the settings.
- **Publishing a fact table**
  a. In the fact table list, select the fact tables you want to publish and click **Publish**. The dialog box for publishing fact tables is displayed.
  b. Select a reviewer from the drop-down list.
  c. Click **Confirm**.
- **Viewing the publish history**
  a. Select a fact table in the list and choose **More** > **View Publish History** on the right.
  b. On the displayed page, 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 modules of DGC.
- **Associate a fact table with a quality rule.**
  a. In the fact table list, select the fact table you want to associate with a quality rule and click **Associate Rule** above the fact table list. The **Associate Quality Rule** dialog box is displayed.
  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 a SQL statement**
  a. Select a fact table in the list and choose **More** > **Preview SQL** on the right.

 Copyright © Huawei Technologies Co., Ltd.

b. On the displayed page, you can view or copy the SQL statement.

- **Suspending a fact table**

  a. In the fact table list, select the fact tables you want to suspend and click **Suspend**. The dialog box for suspending fact tables is displayed.

  b. Select a reviewer from the drop-down list.

  c. Click **Confirm**.

  NOTE

  - The prerequisite for suspending or deleting a fact table is that the fact table is not referenced to define 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.

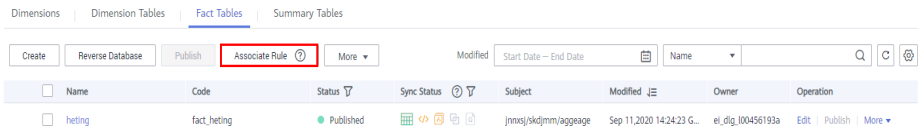
  a. In the fact table list, select the fact tables you want to delete and choose **More** > **Delete** above the list. The dialog box for deleting fact tables is displayed.

  b. Click **OK**.

Associating a Fact Table with a Quality Rule

1. On the Data Design console, choose **Dimensional Modeling** from the left navigation bar.
2. Click the **Fact Tables** tab. The **Fact Tables** tab is displayed.
3. In the fact table list, click the name of the fact table you want to associate with a quality rule. Click **Associate Rule**.

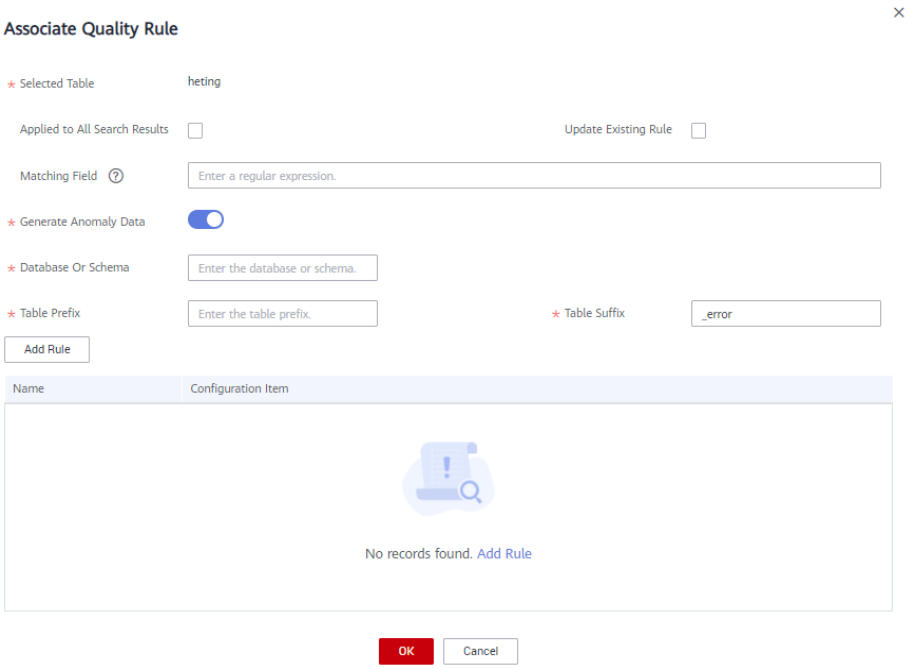
   Figure 7-79 Associating with a quality rule



4. On the displayed page, set parameters as prompt. After the configuration is completed, click **OK**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 7-80 Associating a fact table with a quality rule

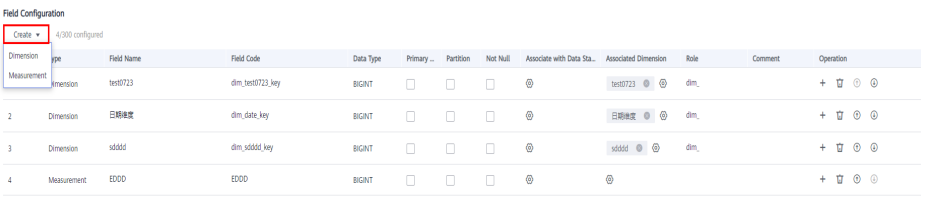


Generate Anomaly Data: If this option is selected, anomaly data is stored in the specified database based on the configured parameters.

Creating a Field in the Fact Table

1. On the Data Design console, choose **Dimensional Modeling** from the left navigation bar.
2. Click the **Fact Tables** tab. The **Fact Tables** tab is displayed.
3. In the fact table list, click the name of the table for which you want to create a field and click **Edit** in the **Operation** column. The editing page is displayed.
4. Click **Create** in the **Field Configuration** area, select a new field type from the drop-down list, and set related parameters.

Figure 7-81 Creating a field



5. After the configuration is complete, click **OK**.

Associating a Fact Table Field with a Data Standard

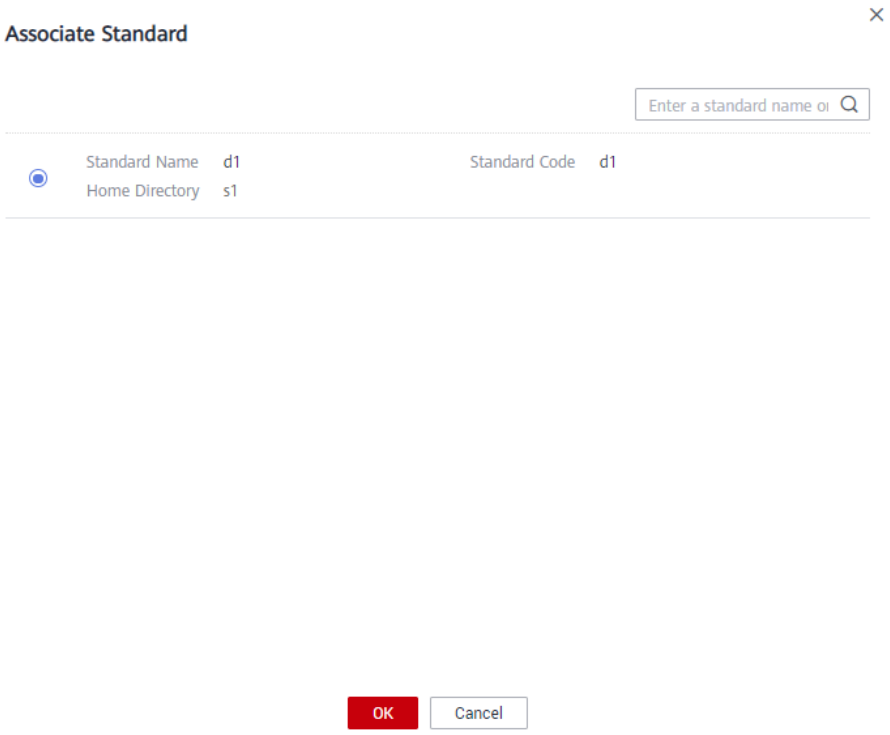
1. On the Data Design console, choose **Dimensional Modeling** from the left navigation bar.
2. Click the **Fact Tables** tab. The **Fact Tables** tab is displayed.

 Copyright © Huawei Technologies Co., Ltd.

3. In the fact table list, click the name of the fact table you want to associate with a data standard.
4. In the table field list on the details page of the fact table, search for the field you want to associate with a data standard, 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 7-82 Associating a fact table field with a data standard

5. After the configuration is complete, click **OK**.

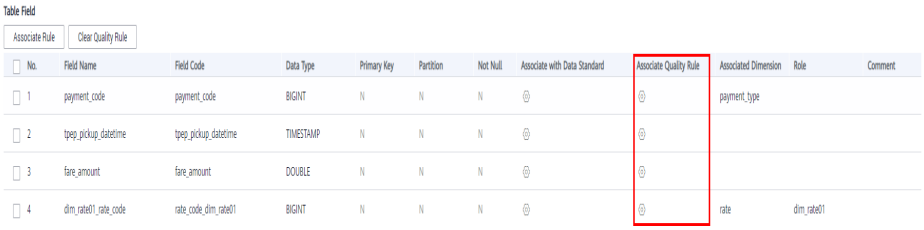


AssociatIng a Fact Table Field with a Quality Rule

1. On the Data Design console, choose **Dimensional Modeling** from the left navigation bar.
2. Click the **Fact Tables** tab. The **Fact Tables** tab is displayed.
3. In the fact table list, click the name of the fact table you want to associate with a quality rule.
4. In the table field list on the details page of the fact table, locate the target field and click to associate the field with a quality rule.

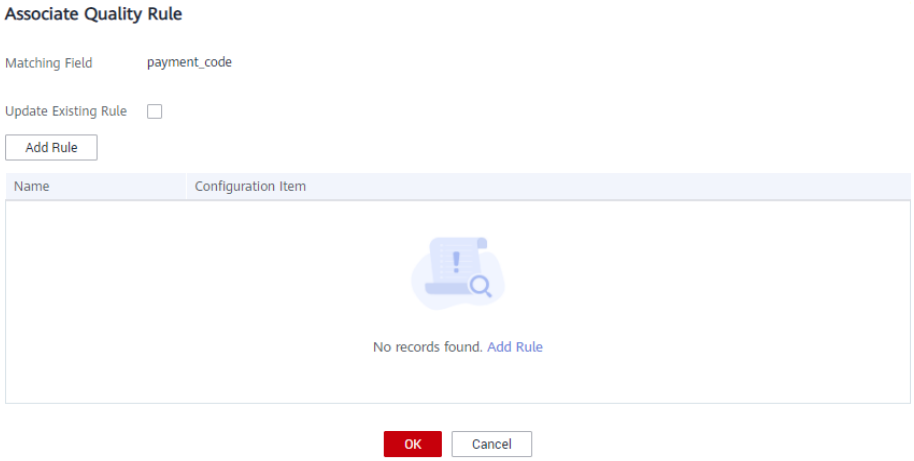
 Copyright © Huawei Technologies Co., Ltd.

Figure 7-83 Associating a single table field with a quality rule



5. After the configuration is complete, click **OK**.

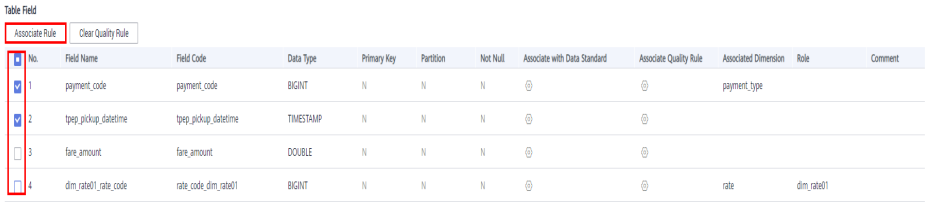
Figure 7-84 Adding rules



AssociatIng Fact Table Fields with Quality Rules in Batches

1. On the Data Design console, choose **Dimensional Modeling** from the left navigation bar.
2. Click the **Fact Tables** tab. The **Fact Tables** tab is displayed.
3. In the fact table list, click the name of the fact table you want to associate with a quality rule.
4. In the table field list on the details page of the fact table, select the table fields you want to associate with quality rules and click **Associate Rule**.

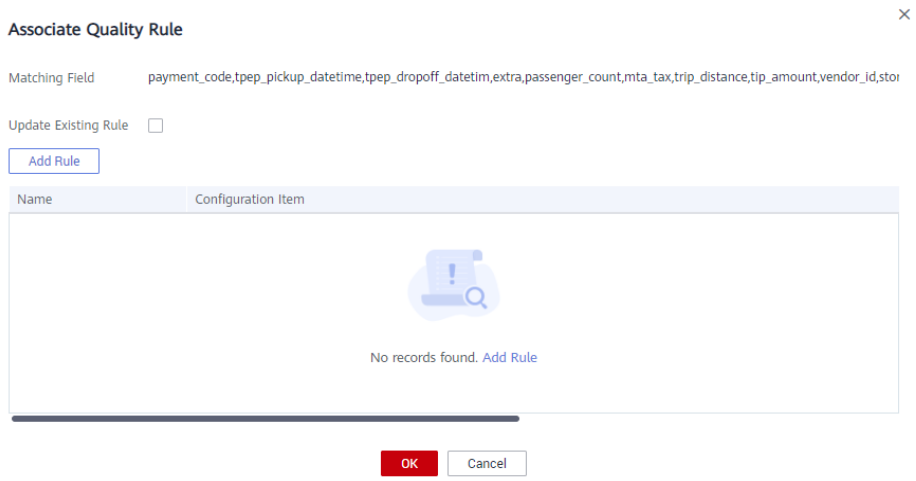
Figure 7-85 Associating fact table fields with quality rules



5. On the page that is displayed, add rules and set rule parameters.

 Copyright © Huawei Technologies Co., Ltd.

Figure 7-86 Adding rules



6. After the configuration is complete, click **OK**.

7.10.4 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.

Prerequisites

Before creating a summary table, ensure that you have created, published, and reviewed dimensions, dimension tables, fact tables, and derivative metrics.

Creating and Publishing a Summary Table

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Design** of the corresponding workspace to access the Data Design console.
3. Choose **Dimensional Modeling** from the left navigation bar. On the displayed page, click the **Summary Tables** tab.
4. Select a subject from the subject tree on the left and click **Create**.
5. On the **Create Summary Table** page, perform the following operations:
   a. Set the parameters in the **Basic Configuration** area.

 Copyright © Huawei Technologies Co., Ltd.

Figure 7-87 Basic Configuration area

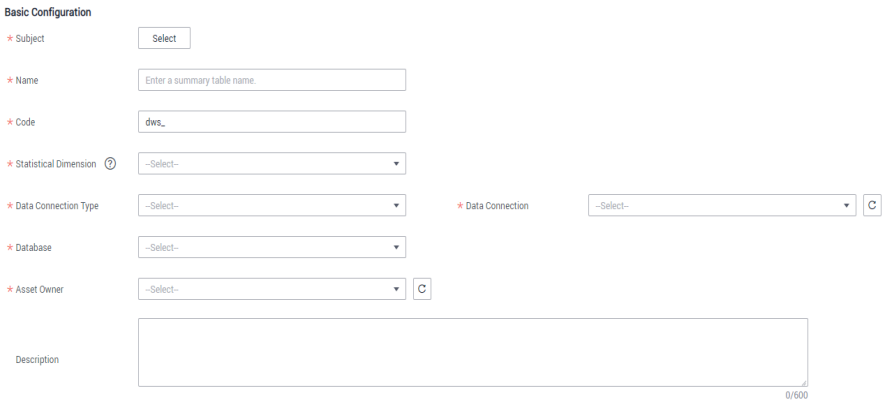


Table 7-41 Parameters of the **Basic Configuration** area

Parameter	Description
Subject	Click **Select** to select a subject catalog (subject area group > subject area > business object) where you can place the summary table.
Name	The name of the table you created. Only letters, digits, and underscores (_) are allowed. Table names must start with letters.
Code	The code of the table you created. Only letters, digits, and underscores (_) are allowed. Table names must start with letters.
Statistical Dimension	Select a statistical dimension. The drop-down list displays only the statistical dimensions configured on the Derivative Metrics page. 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	It is recommended that a same data connection be used for dimension modeling.
Database	The name of the database. Select a database from the drop-down list.
Queue	DLI queue. This parameter is displayed only for DLI data connections.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Schema	This parameter is displayed only for DWS data connections.
Table Type	This parameter is displayed only for DWS data connections. The options are: • **DWS_ROW** indicates that tables are stored to disk partitions by row. • **DWS_COLUMN** indicates that tables are stored to disk partitions by column.
DISTRIBUTE BY	This parameter is displayed only for DWS data connections. Currently, only **REPLICATION** and **HASH** are supported. • **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 need to 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	Select an asset owner from the drop-down list. Only the workspace administrator or developer can be set as the owner.
Description	A description of the summary table to be created. Up to 600 characters are supported.

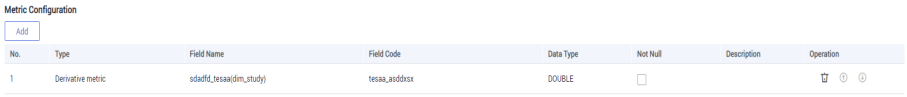
b. In the **Time Partition** area, enter the field code and select the data type. After a table is published, data is written to the table based on the time partition fields.

Figure 7-88 Time Period Configuration area

 Copyright © Huawei Technologies Co., Ltd.

c. In the **Metric Configuration** area, click **Add** to add derivative or compound metrics that are associated with the specified statistical dimensions.

Figure 7-89 Metric Configuration area



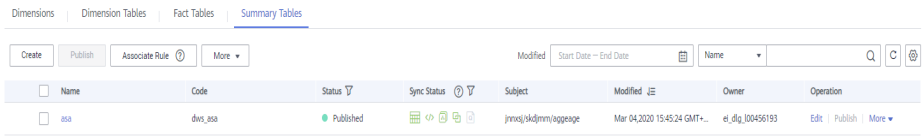
6. Click **Save and Publish**. In the displayed dialogue box, click **Confirm**.
7. Select a reviewer to approve the summary table.

   After the summary table is approved, it is automatically created in the database.
8. 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 Publish History** in the row where the summary table is located. On the displayed page, click the **Publish Log** 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. Choose **Dimensional Modeling** from the left navigation bar and click the **Summary Tables** tab.

Figure 7-90 Summary Tables page



2. Manage your summary tables as required. Refer to the following table for details.

Operation	Helpful Link
Create	Refer to step **Creating and Publishing a Summary Table** for details.
Edit	Refer to step **3** for details.
Publish	Refer to step **4** for details.
View Publish History	Refer to step **5** for details.
Preview SQL	Refer to step **6** for details.
Suspend	Refer to step **7** for details.

 Copyright © Huawei Technologies Co., Ltd.

Operation	Helpful Link
Associate Rule	Refer to step **8** for details.
Delete	Refer to step **9** for details.

3. Edit a summary table.
   a. Click **Edit** to the right of the summary table you want to edit. The page for editing the summary table is displayed.
   b. Edit the summary table as required.
   c. Click **Publish**.
4. Publish a summary table.
   a. Click **Publish** to the right of the summary table you want to publish. The **Submit for Publishing** dialog box is displayed.
   b. Select a reviewer from the drop-down list.
   c. Click **Confirm**.
5. View the publish history.
   a. Select a summary table in the list and choose **More** > **View Publish History** on the right.
   b. On the displayed page, you can view the publish history and version comparison information of the summary table.

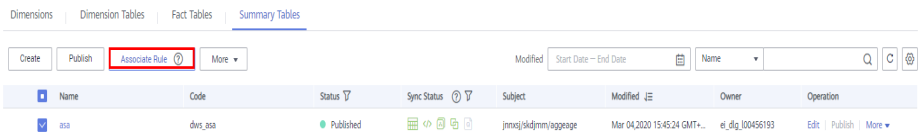
      If the publish log includes error logs, the publishing has failed. You can click **Resynchronize** to retry.
6. Previewing a SQL statement.
   a. Select a summary table in the list and choose **More** > **Preview SQL** on the right.
   b. On the displayed page, you can view or copy the SQL statement.
7. Suspend a summary table.
   a. Click **Suspend** to the right of the summary table you want to suspend. The **Submit for Suspension** dialog box is displayed.
   b. Select a reviewer from the drop-down list.
   c. Click **Confirm**.
8. Associate a summary table with a quality rule.
   a. In the summary table list, select the summary table you want to associate with a quality rule and click **Associate Rule** above the summary table list. The **Associate Quality Rule** dialog box is displayed.
   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 summary table to be deleted and click **Delete** in the upper left corner. The **Delete Summary Table** dialog box is displayed.
   b. Click **OK**.

 Copyright © Huawei Technologies Co., Ltd.

Associating a Summary Table with a Quality Rule

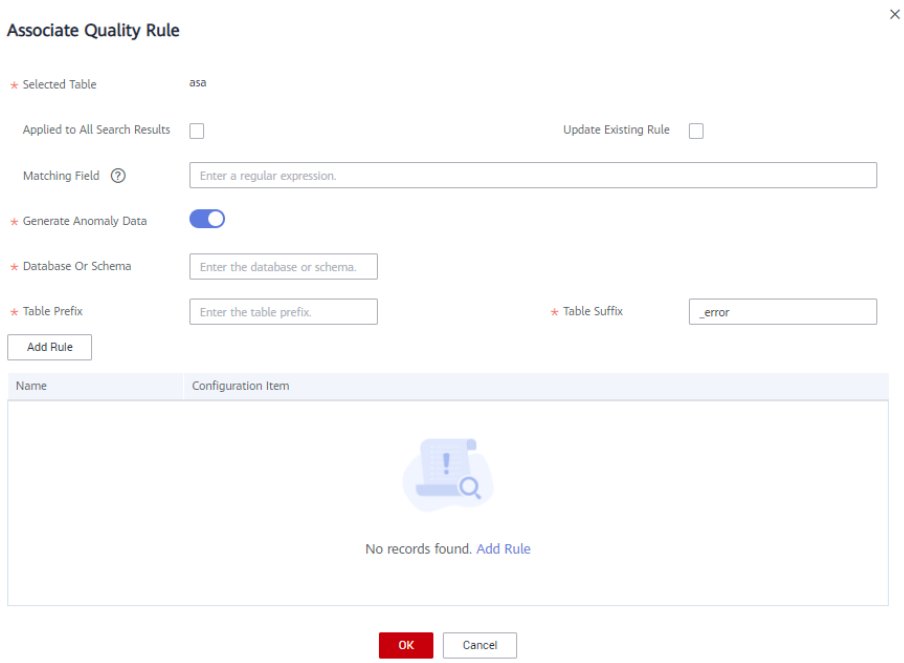
1. On the Data Design console, choose **Dimensional Modeling** from the left navigation bar.
2. Click the **Summary Tables** tab.
3. In the summary table list, click the name of the summary table you want to associate with a quality rule. Click **Associate Rule**.

   Figure 7-91 Associating with a quality rule



4. On the displayed page, set parameters as prompt. After the configuration is completed, click **OK**.

   Figure 7-92 Associating a summary table with a quality rule



   Generate Anomaly Data: If this option is selected, anomaly data is stored in the specified database based on the configured parameters.

Associating a Summary Table Field with a Data Standard

1. On the Data Design console, choose **Dimensional Modeling** from the left navigation bar.
2. Click the **Summary Tables** tab.
3. In the summary table list, click the name of the summary table you want to associate with a data standard.

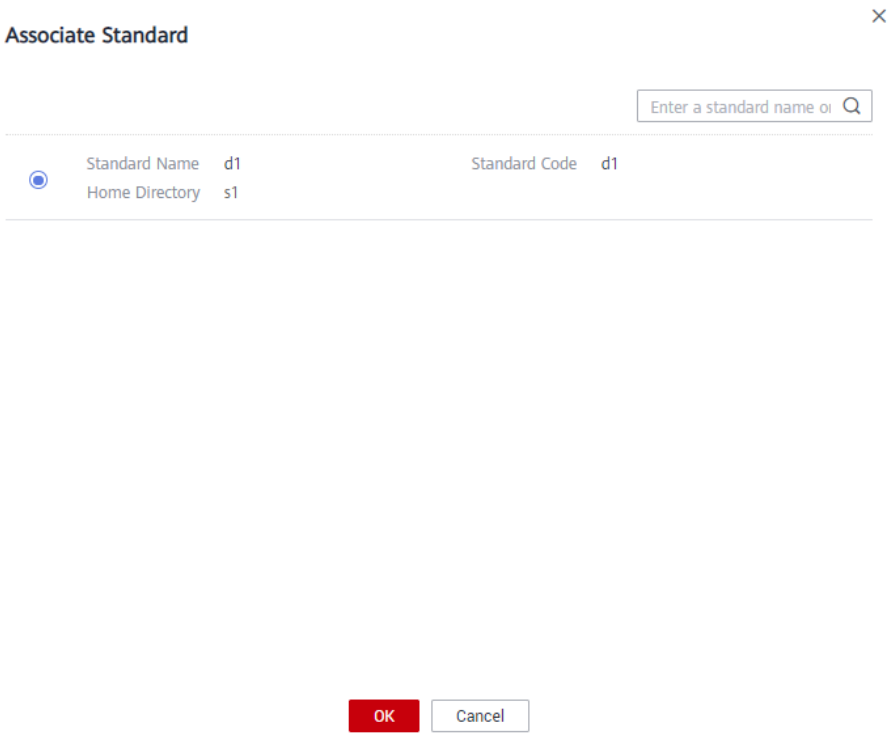
 Copyright © Huawei Technologies Co., Ltd.

4. In the table field list on the details page of the summary table, search for the field you want to associate with a data standard, click corresponding to the field to configure the association between the field and the data standard.

Figure 7-93 Associating a summary table field with a data standard

5. After the configuration is complete, click **OK**. For details on the sources of data standards, see **Creating a Data Standard**.

Figure 7-94 Associating a data standard



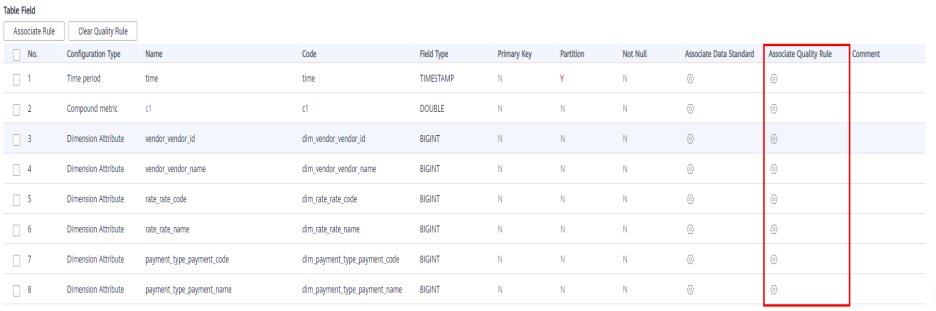
Associating a Single Table Field with a Quality Rule

1. On the Data Design console, choose **Dimensional Modeling** from the left navigation bar.
2. Click the **Summary Tables** tab.
3. In the summary table list, click the name of the summary table you want to associate with a data standard.

 Copyright © Huawei Technologies Co., Ltd.

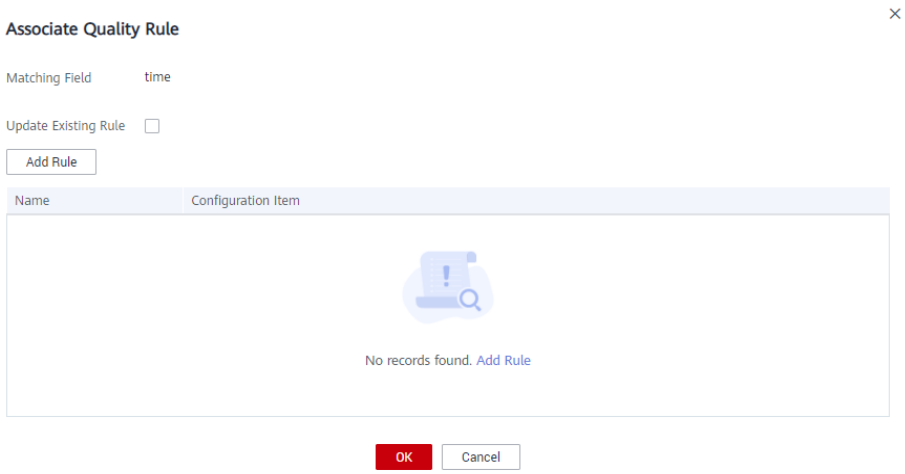
4. In the table field list on the details page of the summary table, locate the target field and click to associate the field with a quality rule.

   Figure 7-95 Associating a single table field with a quality rule



5. After the configuration is complete, click **OK**.

   Figure 7-96 Associating a quality rule



Associating Table Fields with Quality Rules in Batches

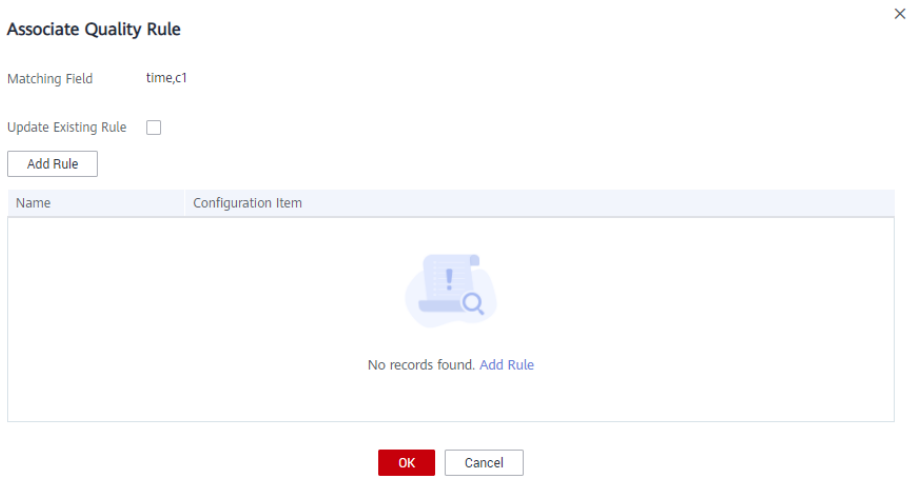
1. On the Data Design console, choose **Dimensional Modeling** from the left navigation bar.
2. Click the **Summary Tables** tab.
3. In the summary table list, click the name of the summary table you want to associate with a data standard.
4. In the table field list on the details page of the summary table, select the table fields you want to associate with quality rules and click **Associate Rule**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 7-97 Associating table fields with quality rules

5. On the page that is displayed, add rules and set rule parameters.

Figure 7-98 Adding rules



6. After the configuration is complete, click **OK**.

7.10.5 Reversing Databases

You can import tables from databases of other data sources to a specified ER model (except the OBS type).

Prerequisites

Before reversing databases, collect data assets from databases in DGC Data Assets so that the system can synchronize tables to DGC Data Assets later. Otherwise, the synchronization tasks may fail. See **Task Management** for details.

Importing a Table to a Specified Model by Reversing the Database

Step 1 On the DGC Data Design console, choose **Dimensional Modeling** from the left navigation bar.

Step 2 Click the **Dimensions** tab or the **Fact Tables** tab. Then, click **Reverse Database** above the list.

 Copyright © Huawei Technologies Co., Ltd.

Figure 7-99 Selecting an object

Step 3 In the **Reverse Database** dialog box, set the parameters as prompt.

Figure 7-100 Reversing a database

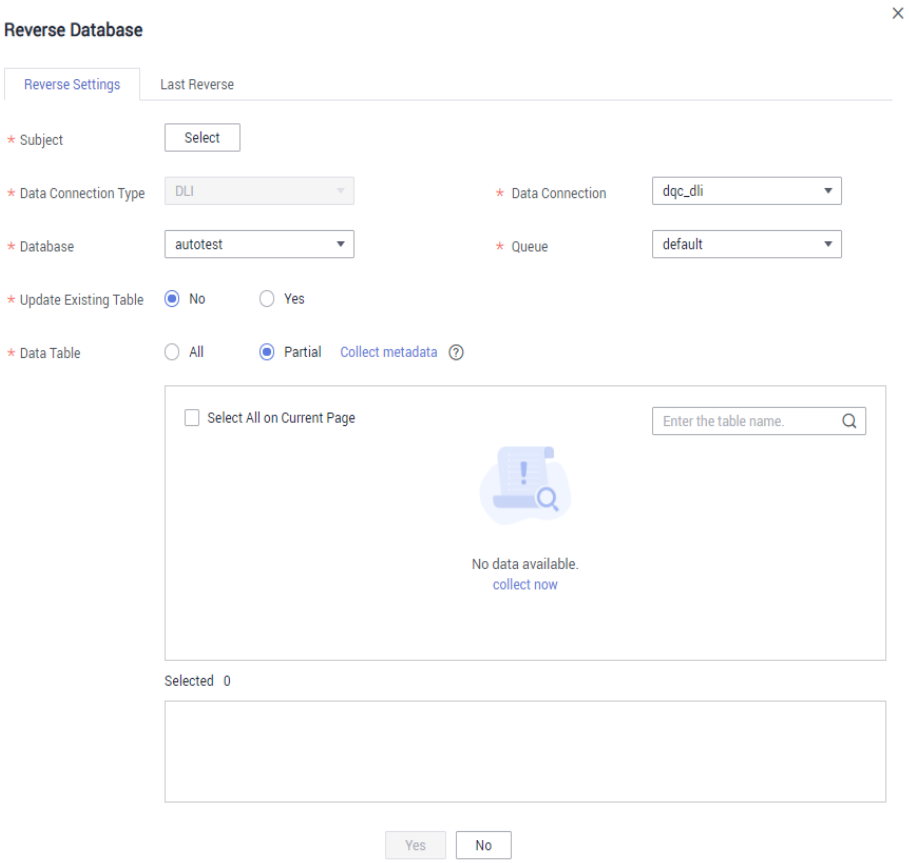


Table 7-42 Parameters

Parameter	Description
Subject	Click **Select** to select a subject.
Data Connection Type	If you reverse tables to a logical model, select a required data connection type from the drop-down list. If you reverse tables to a physical model, the data connection type of the current model is displayed.

 Copyright © Huawei Technologies Co., Ltd.

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 need to create a data connection in the DGC 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.
Queue	This parameter is displayed only for the DLI data connections. Select a DLI queue.
Schema	This parameter is displayed only for DWS data connections. You need to set **Data Connection Type** to **DWS**.
Update Existing 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 its code. During the import, only table creation and update are allowed. • **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 need to 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

7.11 Business Metrics

After data survey and requirement analysis, you need to implement metrics. A metric is a statistical value that measures the overall characteristic of a target and indicates the business situation in a business activity of an enterprise. A metric consists of its name and value. The metric name and its meaning 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.

 Copyright © Huawei Technologies Co., Ltd.

Prerequisites

Before creating a business metric, you need to design a process. For details, see **Designing Processes**.

Creating and Publishing a Business Metric

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Design** of the corresponding workspace to access the Data Design console.
3. On the Data Design console, choose **Business Metrics** from the left navigation bar.
4. In the process tree on the left, select a process and click **Create**.
5. On the displayed page, set parameters as prompted and click **Publish**.

   a. Set the basic parameters described in the figure below:

Figure 7-101 Create Business Metric page

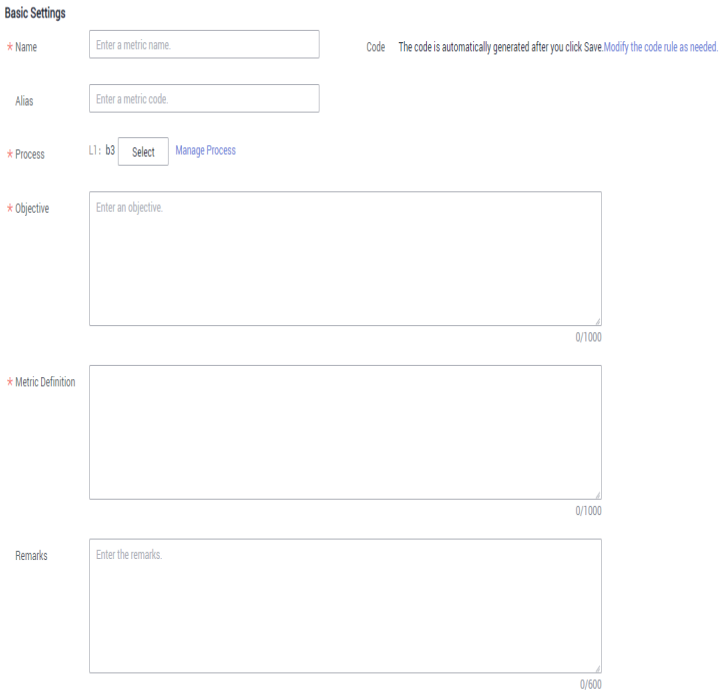


Table 7-43 Parameters of the **Basic Settings** area

Parameter	Description
Name	The name of the business metric to be created.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Code	The metric code is automatically generated. You can configure the generated rule on the **Configuration Center** page. For details, see **Metric Encoding Rule**.
Alias	This parameter is optional.
Process	Select the business process that the metric belongs to. If you have not created a business process, 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 be created.

b. Configure the metric information.

Figure 7-102 Metric Information area

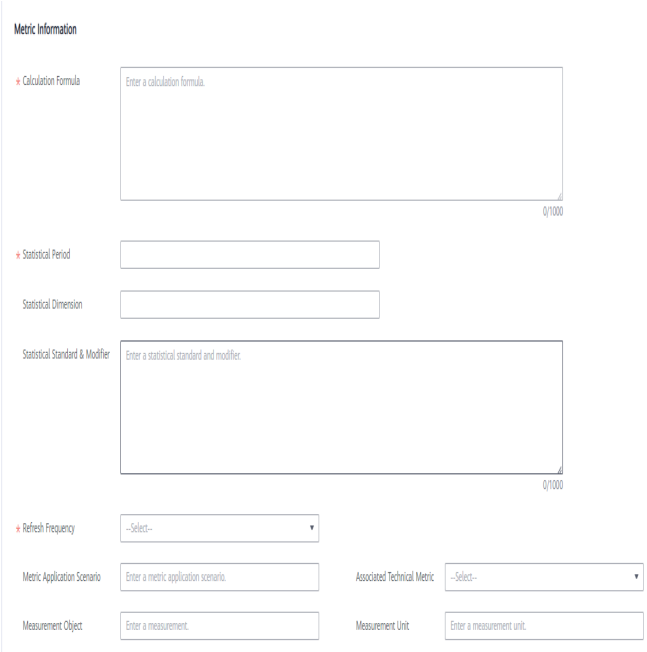


Table 7-44 Parameters of the **Metric Information** area

Parameter	Description
Calculation 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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Statistical Period	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**.
Statistical Standard & Modifier	Standards and modifiers are abstract definitions of business scenarios and are used to determine the measurement scope.
Refresh Frequency	The interval for updating a metric. Developers or O&M personnel can set the scheduling frequency of derivative metrics based on the metric update frequency.
Metric Application Scenario	The application scenarios of a metric.
Associated Technical Metric	Select a technical metric (a derivative metric or compound metric) associated with business metric from the drop-down list.
Measurement Object	The field for measuring a metric.
Measurement Unit	The measurement unit of a metric.

c. Configure the management information.

Figure 7-103 Management Information area

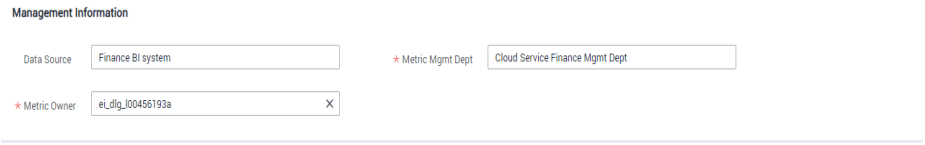


Table 7-45 Parameters of the **Management Information** area

Parameter	Description
Data Source	The generator of data.
Metric Mgmt Dept	The department that manages the metric.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Metric Owner	The owner of the metric. Only the workspace administrator or developer can be set as the owner.

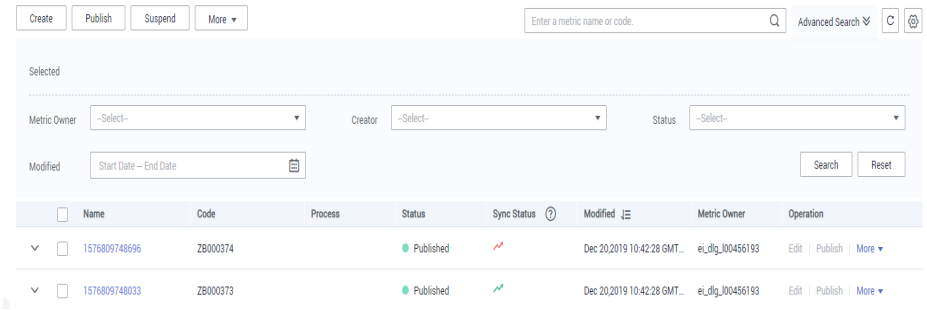
6. In the displayed dialog box, select a reviewer and click **Confirm**.
7. Repeat steps **4** to **6** to create and publish other business metrics.
8. All metrics need to be approved by reviewers.

   If the applications are approved, the business metrics are created.

Editing a Business Metric

1. On the Data Design console, choose **Business Metrics** from the left navigation bar.

   Figure 7-104 Managing business metrics



2. In the business metric list, select a metric that you want to edit and click **Edit** on the right. The page for editing business metrics is displayed.
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 Data Design console, choose **Business Metrics** from the left navigation bar.

Step 2 In the business metric list, select a metric that you want to publish and click **Publish**.

Step 3 In the displayed dialog box, select a reviewer and click **Confirm**.

----End

Suspending a Business Metric

You can perform the following steps to bring a published business metric offline:

 Copyright © Huawei Technologies Co., Ltd.

Step 1 On the Data Design console, choose **Business Metrics** from the left navigation bar.

Step 2 In the business metric list, select a business metric that you want to suspend and click **Suspend** in the **Operation** column.

Step 3 In the displayed dialog box, select a reviewer and click **Confirm**.

----End

Deleting a Business Metric

If a business metric is no longer needed, you can delete it. If the business metric to be deleted has been published, the business metric cannot be deleted. You must suspend the business metric before deleting it.

1. On the Data Design console, choose **Business Metrics** from the left navigation bar.
2. In the business metric list, select one or more business metrics that you want to delete and click **Delete** above the business metric list.
3. In the **Delete Business Metric** dialog box, confirm whether the information is correct. If there is no need of modification, click **OK**.

7.12 Technical Metrics

7.12.1 Creating Atomic Metrics

An atomic metric is an abstract set of the statistical logic and specific algorithms. To solve the inconsistency problem between definitions and R&D, metric definitions determine the statistical logic (or the computing logic), without using ETLs to perform secondary R&D, which improves the R&D efficiency and ensures the consistency of statistical results.

Context

Atomic metrics come from the fact table.

- 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 has no 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.
- If an atomic metric is deleted, check whether a derivative metric is used in later steps. If it is, the atomic metric cannot be deleted.
- Currently, the code of an atomic metric can be changed when the atomic metric is used by the later steps.

 Copyright © Huawei Technologies Co., Ltd.

- If an atomic metric is changed, the related derivative metric will be affected.

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. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Design** of the corresponding workspace to access the Data Design console.
3. Choose **Technical Metrics** from the left navigation bar and click the **Atomic Metrics** tab.
4. Select a subject from the subject tree on the left and click **Create**.
5. On the **Create Atomic Metric** page, set parameters described in **Table 7-46** and click **Save and Publish**.

Figure 7-105 Creating an atomic metric

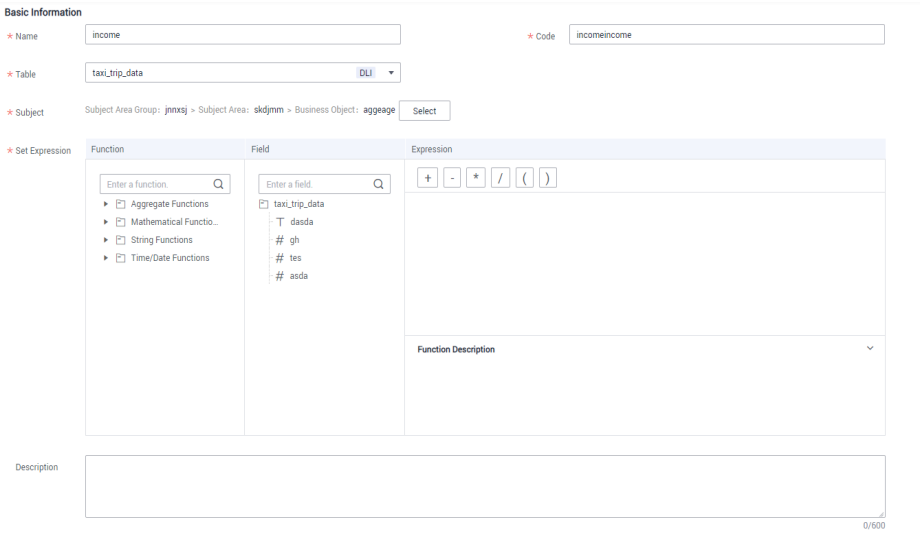


Table 7-46 Parameters for creating atomic metrics

Parameter	Description
Name	Only letters, digits, and underscores (_) are allowed. Metric names must start with letters.
Code	Only letters, digits, and underscores (_) are allowed. Metric codes must start with letters.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Data Table	Select a published fact table from the drop-down list. If there are many tables, you can enter a table name in the text box to search for the fact table you want. If you have not created a fact table, create one. See **Creating and Publishing a Fact Table**.
Subject	The subject to which an 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 be created. Up to 600 characters are supported.

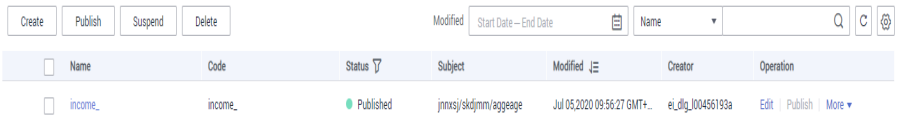
6. In the dialog box that is displayed, select a reviewer and click **Confirm**.
7. (Optional) Publish other atomic metrics by repeating steps **4** to **6**.
8. Wait for the reviewer to approve the application.

   After the application is approved, the atomic metric is created.

Managing an Atomic Metric

1. Click **Technical Metrics** in the navigation tree on the left and click the **Atomic Metrics** tab.

   Figure 7-106 Managing an atomic metric



2. Manage your atomic metrics as required. Refer to the following table for details.

Operation	Helpful Link
Create	Refer to step **Creating and Publishing an Atomic Metric** for details.
Edit	Refer to step **3** for details.
Publish	Refer to step **4** for details.
View Publish History	Refer to step **5** for details.
Suspend	Refer to step **6** for details.

 Copyright © Huawei Technologies Co., Ltd.

Operation	Helpful Link
Delete	Refer to step 7 for details.

3. Edit an atomic metric.
   a. Click **Edit** to the right of the atomic metric you want to edit. The **Edit Atomic Metric** page is displayed.
   b. Edit the atomic metric as required.
   c. Click **Save and Publish**. If you do not want to immediately publish the atomic metric you edited, click **Save** and you can publish it later.
4. Publish an atomic metric.
   a. Click **Publish** to the right of the atomic metric you want to publish. The **Submit for Publishing** dialog box is displayed.
   b. Select a reviewer from the drop-down list.
   c. Click **Confirm**.
5. View the publish history.
   a. Select an atomic metric in the list and choose **More** > **View Publish History**.
   b. On the displayed 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 atomic metric you want to suspend. The **Submit for Suspension** dialog box is displayed.
   b. Select a reviewer from the drop-down list.
   c. Click **Confirm**.

   NOTE

   An atomic metric cannot be suspended or deleted if it is referenced to any derivative metrics.
7. Delete an atomic metric.
   a. Select an atomic metric you want to delete and click **Delete** in the upper left corner. The **Delete Atomic Metric** dialog box is displayed.
   b. Click **OK**.

7.12.2 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 are unique in the business domain and one filter belongs to only one 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.

 Copyright © Huawei Technologies Co., Ltd.

Creating and Publishing a Time Filter

1. Log in to the DGC console, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Design** of the corresponding workspace to access the Data Design console.
3. Click **Technical Metrics** in the navigation tree on the left. On the displayed page, click **Time Filters**.
4. On the **Time Filters** tab, click **Create**.
5. On the **Create Time Filter** page, set parameters described in **Table 7-47** and click **Save and Publish**.

   Figure 7-107 Creating a time filter

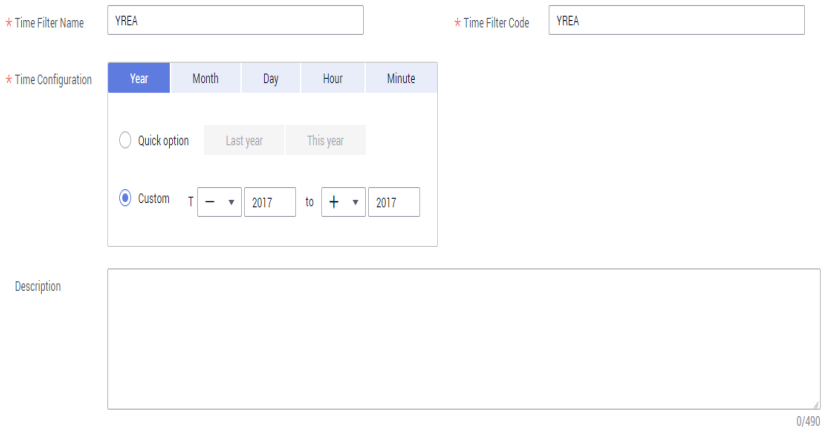


   Table 7-47 Parameters for creating time filters

Parameter	Description
Time Filter Name	Only letters, digits, and underscores (_) are allowed. Time filter names must start with letters.
Time Filter Code	Only letters, digits, and underscores (_) are allowed. Time filter codes must start with letters.
Time Configuration	You can select **Year**, **Month**, **Day**, or **Minute**, and then select **Quick option** or **Custom** to set the time condition.
Description	A description of the time filter to be created. Up to 490 characters are supported.

6. Then, click **Save and Publish** to submit the application.

Copyright © Huawei Technologies Co., Ltd.

7. Wait for the reviewer to approve the application.

   If the application is approved, the time filter is created.

Managing a Time Filter

1. Click **Technical Metrics** in the navigation tree on the left and click the **Time Filters** tab.

   Figure 7-108 Time Filters tab

2. Manage your time filters as required. Refer to the following table for details.

Operation	Helpful Link
Create	Refer to step **Creating and Publishing a Time Filter** for details.
Edit	Refer to step **3** for details.
Publish	Refer to step **4** for details.
View Publish History	Refer to step **5** for details.
Suspend	Refer to step **6** for details.
Delete	Refer to step **7** for details.

3. Edit a time filter.
   a. Click **Edit** to the right of the time filter you want to edit. The page for editing the time filter is displayed.
   b. Edit the time filter as required.
   c. Click **Save** to save the time filter information, or click **Save and Publish** to publish the edited time filter.
4. Publish a time filter.
   a. Click **Publish** to the right of the time filter you want to publish. The **Submit for Publishing** dialog box is displayed.
   b. Select a reviewer from the drop-down list.
   c. Click **Confirm**.
5. View publish history.
   a. Select a time filter in the list and choose **More** > **View Publish History**.
   b. On the displayed page, you can view the publish history and version comparison information of the time filter.
6. Suspend a time filter.

 Copyright © Huawei Technologies Co., Ltd.

a. Click **Suspend** to the right of the time filter you want to suspend. The **Submit for Suspension** dialog box is displayed.
b. Select a reviewer from the drop-down list.
c. Click **Confirm**.

NOTE

Time filters with references cannot be suspended or deleted. They cannot have any derivative metrics referenced.

7. Delete a time filter.
   a. Select the time filter you want to delete and click **Delete** in the upper left corner. The **Delete Time Filter** dialog box is displayed.
   b. Click **OK**.

7.12.3 Creating Derivative Metrics

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 metrics as well.

Derivative metric = Atomic metric + Statistical Dimension + Time filter + General filter.

- **Atomic metric** specifies the statistical standards, namely, the computing logic.
- **Statistical 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 SQLs.
- **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 the SQL statement, excluding the time range).

Prerequisites

- Before creating a derivative metric, ensure that an atomic metric has been created and approved.
- If the derivative metric will use a statistical dimension or time filter, ensure that the dimension and time filter have been created and approved.

Creating and Publishing a Derivative Metric

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Design** of the corresponding workspace to access the Data Design console.
3. Choose **Technical Metrics** from the left navigation bar and click the **Derivative Metrics** tab.
4. Select a subject from the subject tree on the left and click **Create**.
5. On the displayed page, set parameters as prompted.

 Copyright © Huawei Technologies Co., Ltd.

Figure 7-109 Creating a derivative metric

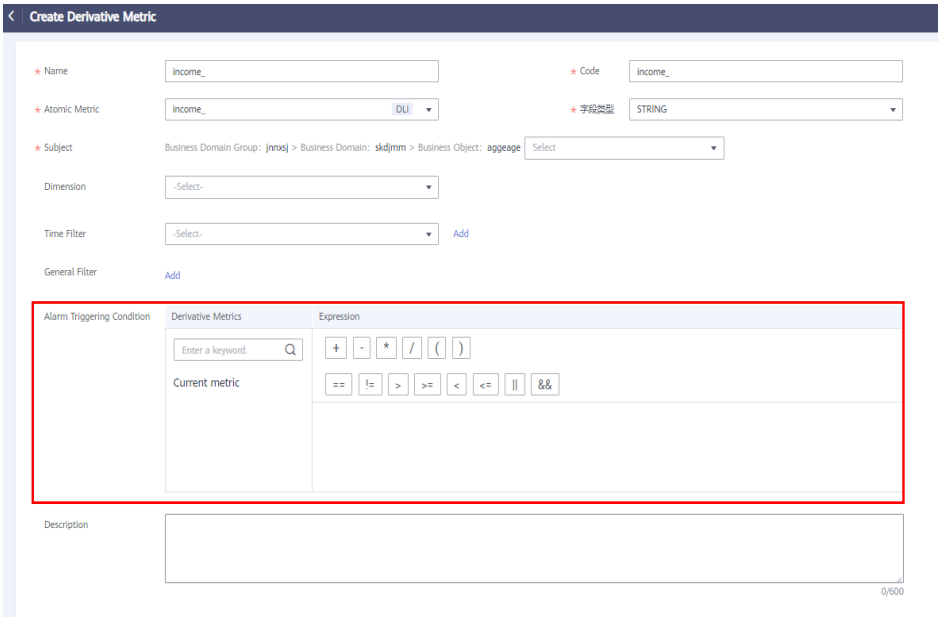
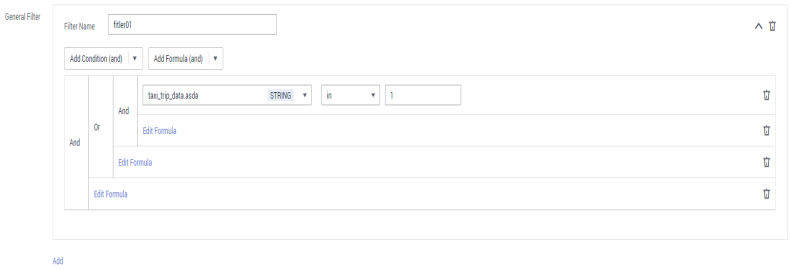


Table 7-48 Parameters for creating derivative metrics

Parameter	Description
Atomic Metric	Select an atomic metric. Only approved atomic metrics are displayed in the drop-down list.
Subject	Subject information.
Statistical Dimension	Select one or more dimensions from the drop-down list. Only the dimensions 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 reconfigured time filters cannot meet the requirements, customize one. See **Creating Time Filters** for details,

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
General Filter	To set general filters, click **Create** to create one or more general filters. In the **General Filter** area shown in **Figure 7-110**, set the following parameters: • **Filter Name** specifies the name of a general filter. • **Adding Condition (and)** allows you to 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. You can click to delete the unwanted conditions. • **Add Formula (and)** allows you to select **And Formula** or **OR formula** to add a formula. Click **Edit Formula** if needed. In the displayed dialog box, select the required functions and fields and set the expression. You can click to delete the unwanted formulas. **Figure 7-110** Setting a general filter 
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.
Description	A description of the derivative metric to be created. Up to 600 characters are supported.

6. In the lower part of the page, click **Trial Run**. In the dialog box that is 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 **Save and Publish**.

 Copyright © Huawei Technologies Co., Ltd.

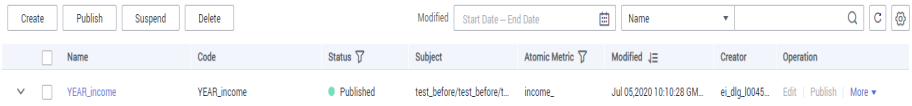
8. In the dialog box that is displayed, select a reviewer and click **Confirm**.
9. (Optional) Create and publish other derivative metrics by repeating steps **3** to **8**.
10. Wait for the reviewer to approve the application.

    After the application is approved, the derivative metric is created.

Managing a Derivative Metric

On the Data Design console, choose **Technical Metrics** > **Derivative Metrics**. On the displayed page, you can edit, publish, suspend, or delete derivative metrics.

Figure 7-111 Managing derivative metrics



1. Choose **Technical Metrics** from the left navigation bar and click the **Derivative Metrics** tab.
2. Manage your derivative metrics as required. Refer to the following table for details.

Operation	Helpful Link
Create	Refer to step **Creating and Publishing a Derivative Metric** for details.
Edit	Refer to step **3** for details.
Publish	Refer to step **4** for details.
View Publish History	Refer to step **5** for details.
Preview SQL	Refer to step **6** for details.
Suspend	Refer to step **7** for details.
Delete	Refer to step **8** for details.

3. Edit a derivative metric.
   a. Click **Edit** to the right of the derivative metric you want to edit. The **Edit Derivative Metric** page is displayed.
   b. Edit the derivative metric as required.
   c. In the lower part of the page, click **Trial Run**. In the dialog box that is 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 **Save and Publish**.
4. Publish a derivative metric.

 Copyright © Huawei Technologies Co., Ltd.

a. Click **Publish** to the right of the derivative metric you selected. The **Submit for Publishing** dialog box is displayed.
b. Select a reviewer from the drop-down list.
c. Click **Confirm**.

5. View the publish history.
   a. Select a derivative metric in the list and choose **More** > **View Publish History**.
   b. On the displayed page, you can view the publish history and version comparison information of the metric.
6. Preview a SQL statement.
   a. Select a derivative metric in the list and choose **More** > **Preview SQL**.
   b. On the displayed page, you can view or copy the SQL statement.
7. Suspend a derivative metric.
   a. On the right of the derivative metric you want to suspend, choose **More** > **Suspend**. The **Submit for Suspension** dialog box is displayed.
   b. Select a reviewer from the drop-down list.
   c. Click **Confirm**.
8. Delete a derivative metric.
   a. Select one or more derivative metrics in the list and click **Delete** in the upper left corner. The **Delete Derivative Metric** dialog box is displayed.
   b. Click **OK**.

7.12.4 Creating Compound Metrics

A compound metric is generated by adding one or more derivative metrics. The dimensions, definitions, standards, and modifiers of a compound metric are the same as those of the derivative metrics. No new dimensions and modifiers for the compound metrics generated if the derivative metrics have no dimensions and modifiers.

Prerequisites

You have created a derivative metric and the metric has been approved. For details, see **Creating Derivative Metrics**.

Creating a Compound Metric

To create a compound metric, do the following:

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Design** of the corresponding workspace to access the Data Design console.
3. Choose **Technical Metrics** from the left navigation bar. On the displayed page, click the **Compound Metric** tab.

 Copyright © Huawei Technologies Co., Ltd.

4. On the **Create Compound Metric** page, select a subject from the subject tree on the left and click **Add**.
5. On the displayed page, set parameters as prompted.

Figure 7-112 Creating a compound metric

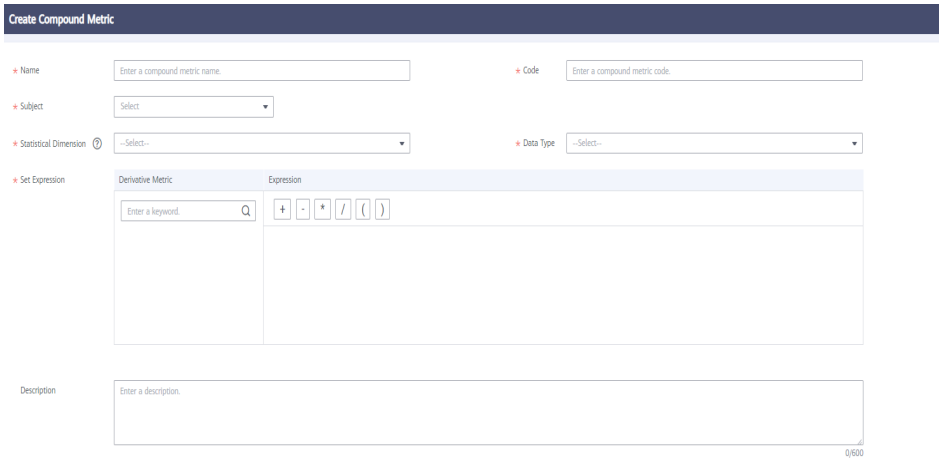


Table 7-49 Parameters for creating compound metrics

Parameter	Description
Name	Only letters, digits, and underscores (_) are allowed. Compound metric names must start with letters.
Code	Only letters, digits, and underscores (_) are allowed. Metric code names must start with letters.
Subject	Subject information. Click **Select** to select a subject.
Statistical Dimension	Configured on the **Derivative Metrics** page.
Data Type	Select a data type for the compound metric to be created.
Set 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.
Description	A description of the compound metric to be created. Up to 600 characters are supported.

 Copyright © Huawei Technologies Co., Ltd.

6. In the lower part of the page, click **Trial Run**. In the dialog box that is displayed, click **Trial Run** to check whether the compound metric can run properly.

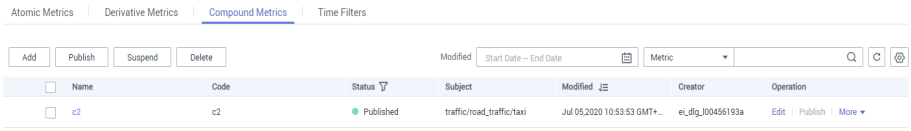
   If the trial run fails, locate the fault based on the error message, correct the configurations, and click **Trial Run** to try again.
7. If the trial run is successful, click **Publish**.
8. In the dialog box that is displayed, select a reviewer and click **Confirm**.
9. Wait for the reviewer to approve the application.

   After the application is approved, the compound metric is created.

Editing a Compound Metric

1. Choose **Technical Metrics** from the left navigation bar and click the **Compound Metrics** tab.

   Figure 7-113 Compound metrics



2. In the compound metric list, select a metric that you want to edit and click **Edit** on the right. The page for editing the compound metric is displayed.
3. Set the parameters as prompt. For details, see **Table 7-49**.
4. In the lower part of the page, click **Trial Run**. In the dialog box that is 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 that is displayed, select a reviewer and click **Confirm**.

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. Choose **Technical Metrics** from the left navigation bar and click the **Compound Metrics** tab.
2. In the compound metric list, select the compound metrics you want to publish and click **Publish**. The dialog box for publishing compound metrics is displayed.
3. If there are no changes made to the information, click **Confirm**.

Viewing the Publish History

1. Choose **Technical Metrics** from the left navigation bar and click the **Compound Metrics** tab.

 Copyright © Huawei Technologies Co., Ltd.

2. Select a compound metric in the list and choose **More** > **View Publish History**.
3. On the displayed page, you can view the publish history and version comparison information of the metric.

Previewing a SQL Statement

1. Choose **Technical Metrics** from the left navigation bar and click the **Compound Metrics** tab.
2. In the compound metric list, locate the target compound metric and choose **More** > **Preview SQL**. The **Preview SQL** dialog box is displayed.
3. On the displayed page, 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.

1. Choose **Technical Metrics** from the left navigation bar and click the **Compound Metrics** tab.
2. In the compound metric list, select the compound metrics you want to suspend and click **Suspend** above the list. The dialog box for suspending compound metrics is displayed.
3. If there are no changes made to the information, click **Confirm**.

Deleting a Compound Metric

1. Choose **Technical Metrics** from the left navigation bar and click the **Compound Metrics** tab.
2. Select compound metrics in the list and click **Delete** above the list. The dialog box for deleting the compound metrics is displayed.
3. If there are changes made to your selection, click **OK**.

7.12.5 Creating General Filters

Atomic metrics are standard definitions for computing logic. General filters are standard definitions for conditional limits. To ensure that all statistical metrics are unified, standard, and unambiguous, these general filters must be unique within a business domain and each general filter can only belong to only a single source logic table. The computing logic is defined based on the fields of the source logic table model. A general filter may come from multiple logic tables that belong to different data domains. Therefore, a general filter may belong to multiple data domains as well.

Creating and Publishing a General Filter

Before creating a general filter, ensure that you have created, published, and reviewed dimension tables, fact tables, and general filter groups.

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

 Copyright © Huawei Technologies Co., Ltd.

2. Click **Data Design** of the corresponding workspace to access the Data Design console.
3. Choose **Technical Metrics** from the left navigation bar. On the displayed page, click the **General Filters** tab.
4. On the **General Filters** tab, click **Create**.
5. Set the parameters shown in the following figure and click **Save and Publish**.

Figure 7-114 Creating a general filter

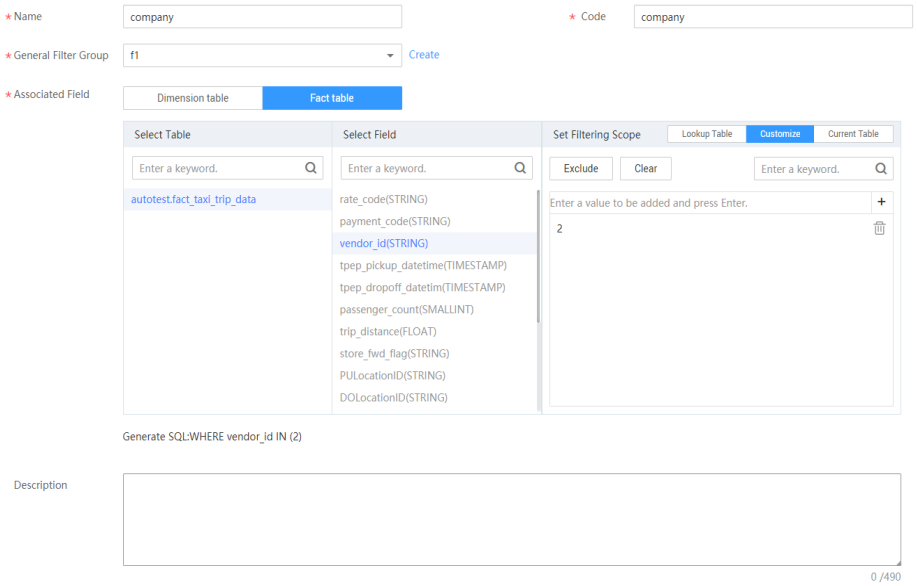


Table 7-50 Parameters for creating general filters

Parameter	Description
Name	Only letters, digits, and underscores (_) are allowed. General filter names must start with letters.
Code	Only letters, digits, and underscores (_) are allowed. General filter codes must start with letters.
General Filter Group	Select a general filter from the drop-down list. If you have not created a general filter, click **Create**.

 Copyright © Huawei Technologies Co., Ltd.

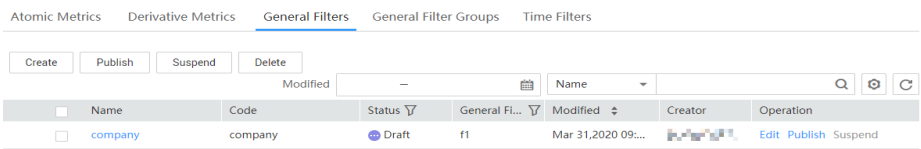
Parameter	Description
Associated Field	Select the required fields to be associated and set the filtering scope. You can set the filtering scope using one of the following methods: • Selecting from the lookup table: You can select a value from the created lookup table. • Selecting from the list of custom values: You can add one or more custom values. • Selecting from the existing values: You can select one or more values from the existing values. Click Exclude to filter out a scope that is displayed as NOT IN in the generated SQL statement. Click Exclude again to not filter out this scope. That is, the filter scope is restored as IN in the generated SQL statement.
Description	A description of the general filter to be created. Up to 490 characters are supported.

6. Then, click **Save and Publish** to submit the application.
7. (Optional) Create and publish other general filters by repeating steps 5 to 8.
8. Wait for the reviewer to approve the application.
   If the application is approved, the general filter is created.

Managing a General Filter

1. On the Data Design console, choose **Technical Metrics** from the left navigation bar and click the **General Filters** tab.

Figure 7-115 General Filters tab



2. Manage your general filters as required. Refer to the following table for details.

Operation	Helpful Link
Create	Refer to step **Creating and Publishing a General Filter** for details.
Edit	Refer to step 3 for details.

 Copyright © Huawei Technologies Co., Ltd.

Operation	Helpful Link
Publish	Refer to step 4 for details.
Suspend	Refer to step 5 for details.
Delete	Refer to step 6 for details.

3. Edit a general filter.
   a. Click **Edit** to the right of the general filter you want to edit. The page for editing a general filter is displayed.
   b. Edit the target general filter. Refer to step 2 for details.
   c. Click **Save** to save the general filter information, or click **Publish** to publish the general filter.
4. Publish a general filter.
   a. Click **Publish** to the right of the general filter you want to publish. The **Submit for Publishing** dialog box is displayed.
   b. Select a reviewer from the drop-down list.
   c. Click **Confirm**.
5. Suspend a general filter.
   a. Click **Suspend** to the right of the general filter you want to suspend. The **Submit for Suspension** dialog box is displayed.
   b. Select a reviewer from the drop-down list.
   c. Click **Confirm**.

   NOTE

   1. A general filter cannot be suspended or deleted if it is referenced to any derivative metrics.
6. Delete a general filter.
   a. Select the general filter to be deleted and click **Delete** in the upper left corner. The **Delete General Filter** dialog box is displayed.
   b. Click **OK**.

7.13 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 released on the **Review Center** page, these tasks can be brought online in the production environment.

Reviewer's Audit Object

If you are a reviewer, use the reviewer account with caution.

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

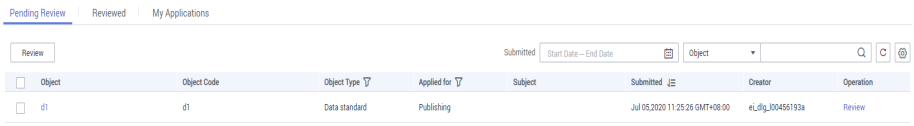
   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

 Copyright © Huawei Technologies Co., Ltd.

2. Click **Data Design** of the corresponding workspace to access the Data Design console.
3. Choose **Review Center** from 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.

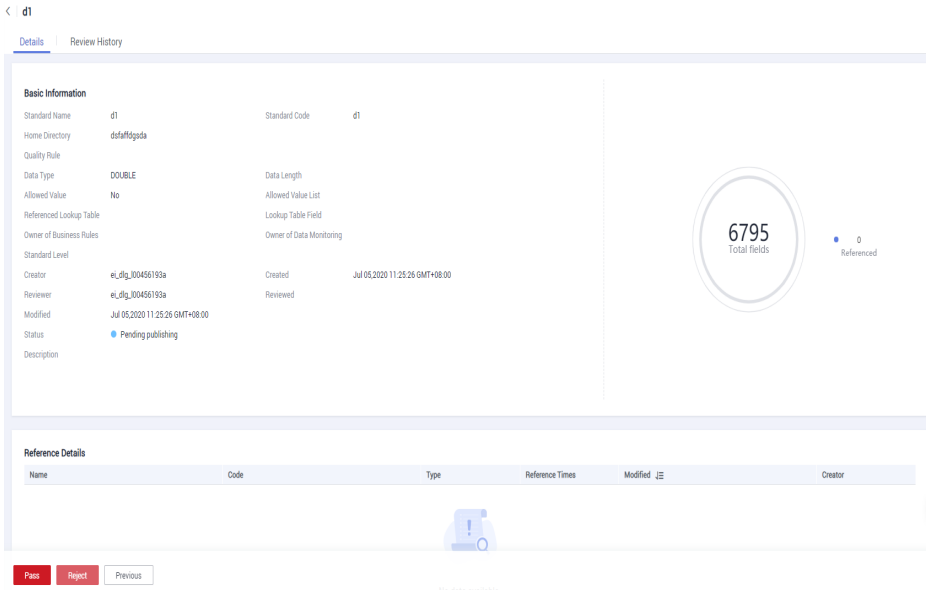
   Figure 7-116 Pending Review tab

4. On the displayed page, confirm the information and click **Pass**. In the dialog box that is displayed, enter the review comments and click **OK**.

   If the information is incorrect, click **Reject**. In the dialog box that is displayed, enter the reasons for rejecting the application and click **OK**.

   Figure 7-117 Review Information area

Viewing the Pending Review, Reviewed, and My Applications Tabs

- **Pending Review** tab

  On the DGC Data Design console, choose **Review Center** from the left navigation bar and click the **Pending Review** tab. On the displayed page, you can view the objects to be reviewed.

- **Reviewed** tab

  On the DGC Data Design console, choose **Review Center** from the left navigation bar and click the **Reviewed** tab. On the displayed page, you can view the objects that have been approved.

 Copyright © Huawei Technologies Co., Ltd.

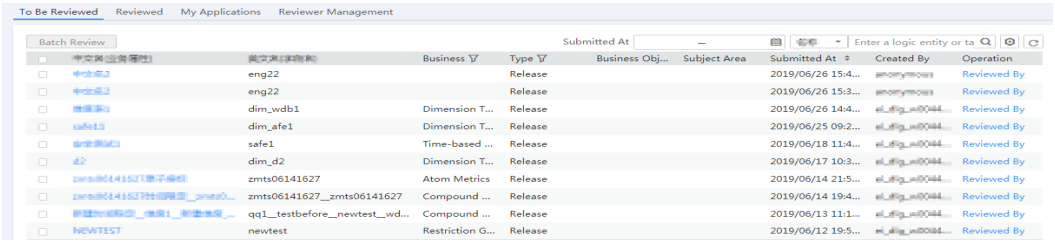
- **My Applications** tab

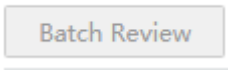
  On the DGC Data Design console, choose **Review Center** from the left navigation bar and click the **My Applications** tab. On the displayed page, you can view the applications you have submitted.

Pending Review

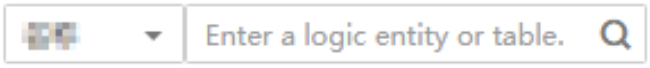
Step 1 On the DGC Data Design console, choose **Review Center** form the left navigation bar. The **Review Center** page is displayed.

Figure 7-118 Pending Review tab



Function Area	Description
1	Batch Review 1. Allows users to select multiple pieces of information to be reviewed. 2. Click  . The **Batch Review** dialog box is displayed. 3. Enter the valid review comments. 4. Click **Batch Review**. The selected targets are approved in batches. Click **Batch Reject**. The selected targets are rejected 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 the site requirements. 3. Click **Confirm**.

 Copyright © Huawei Technologies Co., Ltd.

Function Area	Description
3	• 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 such as tables, creators, reviewers, or owners. • allows you to set the headers of tables to be reviewed. • allows you to refresh the current page.

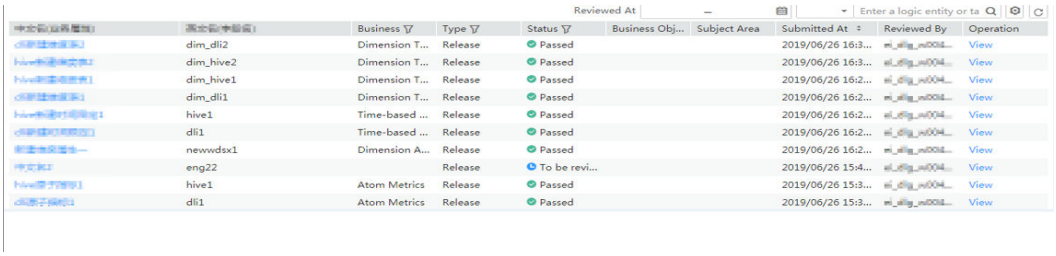
----End

My Applications

Step 1 On the Data Design console, choose **Review Center** from the left navigation bar. The **Review Center** page is displayed.

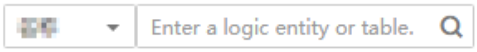
Step 2 Click **My Applications**. The **My Applications** page is displayed.

Figure 7-119 My Applications page



Function Area	Description
1	Displays my application information. You can click **View** to view the information about a specified row in the **Operation** column.

 Copyright © Huawei Technologies Co., Ltd.

Function Area	Description
2	• allows you to specify a time range during which the submitted applications are displayed. •  allows you to select the required names and reviewers. • allows you to set the headers of the reviewed tables. • allows you to refresh the current page.

----End

7.14 Configuration Center

Subject Levels

1. Log in to the DGC console, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Design** of the corresponding workspace to access the Data Design console.
3. On the displayed page, click **Configuration Center** in the navigation tree on the left. Then, click **Subject Levels**.
4. Click + in the **Operation** column to add a custom subject level and click **OK**.
5. Click in the **Operation** column to delete a subject level and click **OK**.
6. Except the business object at the last level, you can click the names of other levels to edit them.

   NOTE

   - 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.
   - Currently, a maximum of seven layers and a minimum of two layers are supported.

Managing a Standard Template

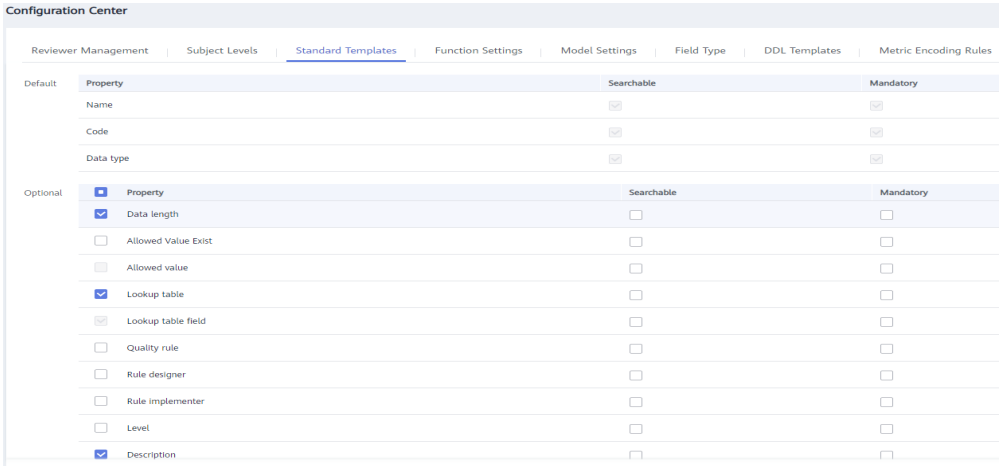
1. Log in to the DGC console, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Design** of the corresponding workspace to access the Data Design console.

 Copyright © Huawei Technologies Co., Ltd.

3. Choose **Configuration Center** from the left navigation bar. On the displayed page, click **Standard Template Management**.
4. In the **Optional** area, select parameters as required. Click + to add custom properties. After the configuration is complete, click **OK**.

Figure 7-120 Standard Template Management tab



A standard template includes fields **Searchable** and **Mandatory**.

After the template is saved, you need to set the options in the template when creating a data standard.

Function Configuration

1. On the Data Design console, choose **Configuration Center** from the left navigation bar and click **Function Configuration**.
2. On the **Function Configuration** tab, set the parameters as prompt and click **OK**. Click **Reset** to restore the default settings.
   - **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. If a process is not selected, you need to manually execute it in the DGC Data Development, Data Assets, and Data Quality Control modules.

     NOTE

     - To enable **Synchronize Technical Assets**, you must create a data asset collection task for the database to which the table belongs in the DGC Data Assets module. Otherwise, the technical asset synchronization will fail.
     - If **Create Tables** is selected, the physical table in the database is automatically deleted when the table is deleted.
   - **Data Table Update Mode**: If a table in Data Design 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.

 Copyright © Huawei Technologies Co., Ltd.

- **Not update**: Do not update tables in a database.
- **DDL-based update**: Update tables in the database based on the DDL update template configured in **Managing a DDL Template**. Whether the update is successful depends on whether the underlying data warehouse engine. Different types of data warehouses support different table update modes. If the data warehouse does not support table update operations on the Data Design console, the tables in the database may be inconsistent with those in Data Design. For example, table fields cannot be deleted when DLI tables are updated. If table fields are deleted from the tables in Data Design, 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**: Delete an existing table in a database and then create a table. This option ensures that the tables in the database are the same as those in Data Design. 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 Synch**: 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.
– **Number of Parallel Query Subjects On the Information Architecture Page**: The default value is **1**. Currently, you cannot change the value.

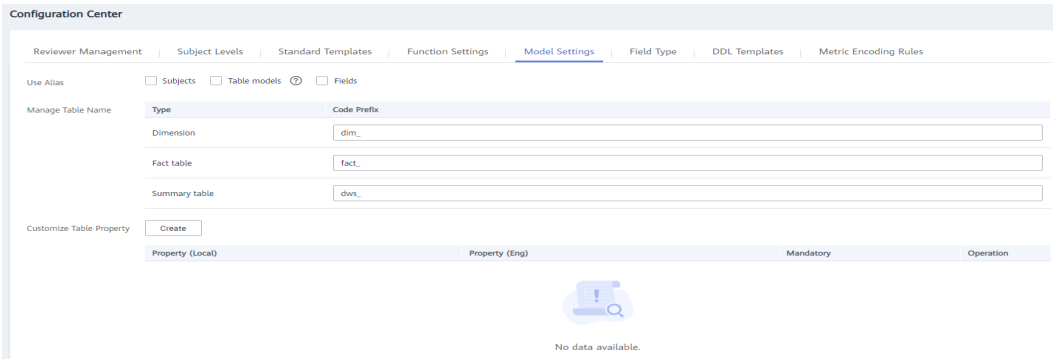
Configuring a Model

You can perform the following operations during subject design and model design on the **Model Settings** page.

- Added the subject alias, table model alias, and field alias.
- Set the default table code prefix for the dimension table, fact table, and summary table.
- Add custom fields to a table.

 Copyright © Huawei Technologies Co., Ltd.

Figure 7-121 Configuring a model



Step 1 On the displayed page, click **Configuration Center** in the navigation tree on the left. Then, click **Model Settings**.

Step 2 Enable the alias. On the **Model Settings** page, you can add an 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 can enter an alias when creating or editing a table field.

Step 3 Manage table names. You can set the default table code prefix for a dimension table, a fact table, or a summary table.

Step 4 Customize table properties. 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 when **Fields** is selected.

----End

Managing a Field Type

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 On the Data Design console, choose **Configuration Center** from the left navigation bar and click **Field Type**.

Step 2 On the displayed page, 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.

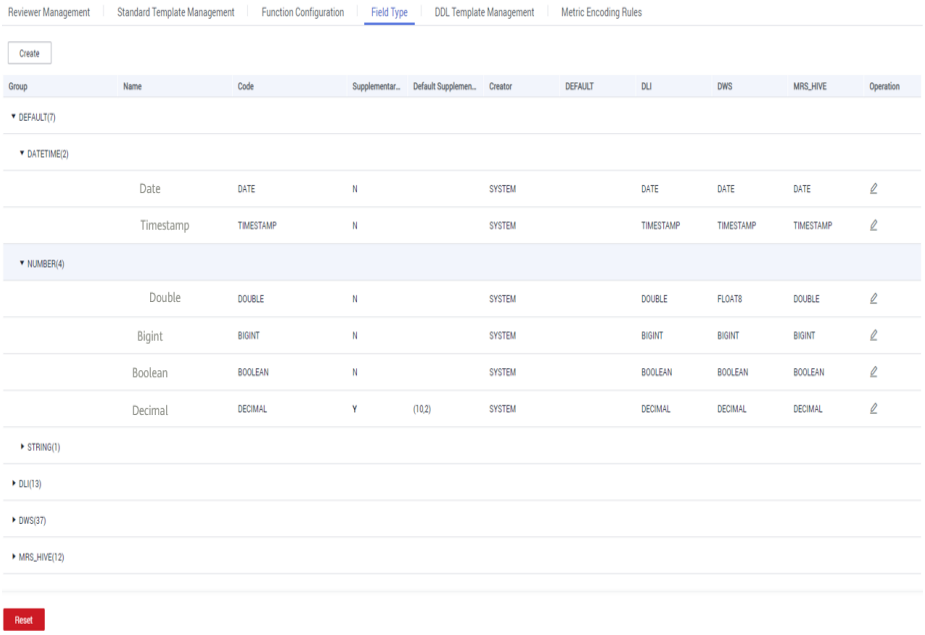
The types are described as follows:

- **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.

 Copyright © Huawei Technologies Co., Ltd.

- **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.

Figure 7-122 Field Type tab

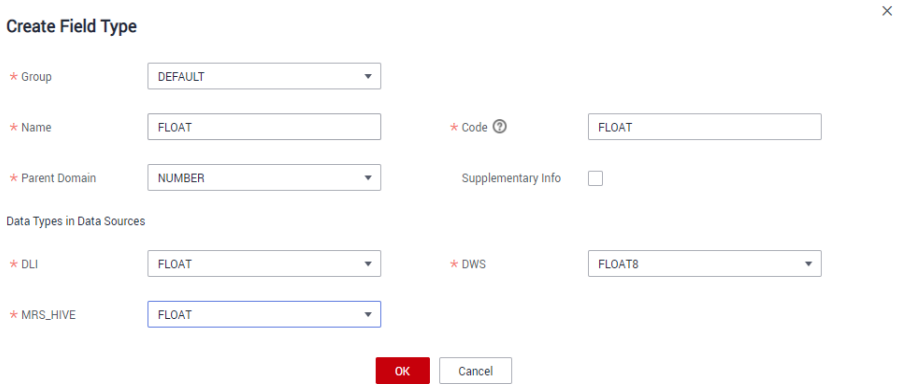


Step 3 Manage field types.

- **Create**

  To add a field type, click **Create**. In the displayed dialog box, set the parameters as prompted and click **OK**.

  Figure 7-123 Creating a field type



 Copyright © Huawei Technologies Co., Ltd.

Table 7-51 Parameters

Parameter	Description
Application Scope	Type group that the new type belongs to.
Name	Name of the field type to be created. Letters, digits, brackets, spaces, and underscores (_) are allowed, but field type names must start with letters.
Code	Data type code, which must be supported by the data warehouse. Only upper-case letters, digits, and underscore (_) are allowed.
Parent Domain	Select the domain that the new field type belongs to.
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.

- **Edit**

  In the field type list, specify a field type and click to edit the field type. For details on the parameters, see **Table 7-51**.

- **Delete**

  You can delete 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

Managing a DDL Template

In DGC Data Design, you can modify DDL templates of DLI views or diversified types of tables (such as DLI, DWS and Hive). 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. On the Data Design console, choose **Configuration Center** from the left navigation bar. Then, click **DDL Template Management**.
2. On the displayed page, you can configure DDL templates for DLI views or diversified types of tables. You can modify the DDL templates by referring to

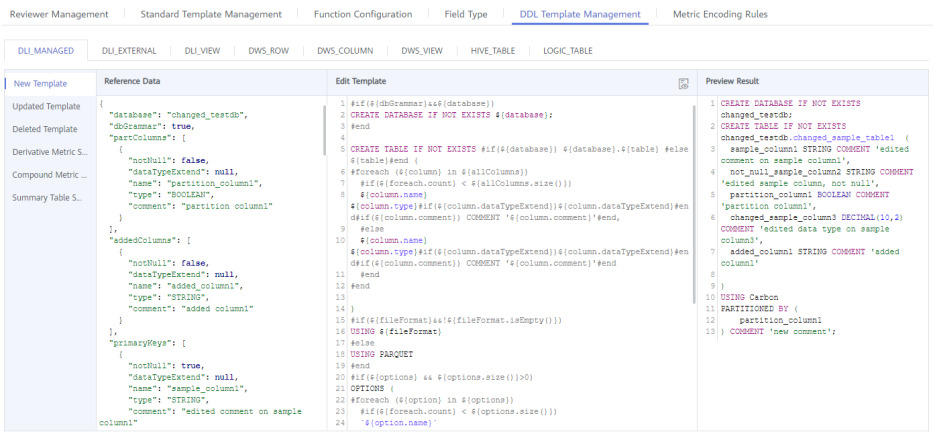
 Copyright © Huawei Technologies Co., Ltd.

the parameter description on this page. After the modification is complete, click **OK**. Click **Reset** to restore the default settings.

Parameters in **Figure 7-124** are described as follows:

- **New Template** allows you to view or edit a new table or a DDL template of a DLI view.
- **Updated Template** allows you to view or edit an updated table or a DDL template of a DLI view.
- **Deleted Template** allows you to view or edit a deleted table or a DDL template of a DLI view.
- **Derivative Metric SQL Template** allows you to view or edit the SQL template of a derivative metric.
- **Compound Metric SQL Template** allows you to view or edit the SQL template of a compound metric.
- **Summary Table SQL Template** 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 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 7-124 DDL Template Management tab

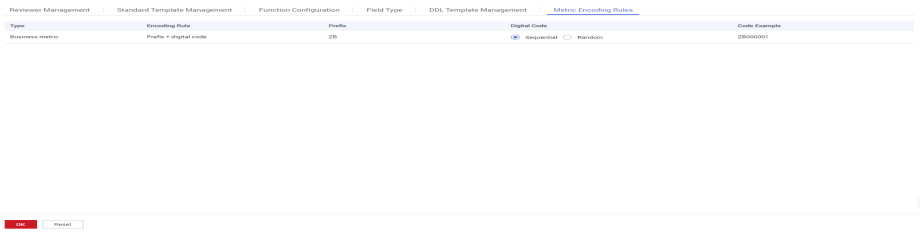


Metric Encoding Rule

1. On the Data Design console, choose **Configuration Center** from the left navigation bar. On the displayed page, click **Metric Encoding Rules**.
2. In the **Digital Code** column, select **Sequential** or **Random** and click **OK**. Click **Reset** to restore the default settings.

 Copyright © Huawei Technologies Co., Ltd.

Figure 7-125 Metric Encoding Rule tab



 Copyright © Huawei Technologies Co., Ltd.

8 Data Development

8.1 Overview

Purpose

Data Development is a one-stop big data platform for collaborative development. It provides multiple functions, such as data modeling, data integration, script development, job scheduling, and monitoring. Data modeling supports multiple EI services. Cloud Data Migration is integrated for data migration between more than 30 types of data sources. Script development provides the SQL and Shell script development capabilities to achieve easy and convenient online development. Job scheduling supports hybrid orchestration of types of cloud services. The high-performance scheduling engine has been verified by hundreds of applications. DLF provides activity monitoring and supports on-demand resource scheduling to reduce the costs of using the big data platform.

Objects in Data Development

- Workspace: A workspace belongs to an enterprise project. To facilitate unified management of scripts and jobs, DLF supports workspace configuration.
- Data connection: A data collection is a collection of information required for accessing data storage (computing) space, including the connection type, name, and login information.
- Solution: A solution provides users with convenient and systematic management operations and better meet service requirements and objectives. Each solution can contain one or more business-related jobs, and one job can be reused 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.

 Copyright © Huawei Technologies Co., Ltd.

- 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 references 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 refer to the instance that is generated in a past period of time by a periodically scheduled job.

8.2 Data Management

8.2.1 Creating a Database

After creating a data connection, you can manage the databases under the data connection.

The following types of databases can be created:

- DLI
- DWS
- MRS Hive

Prerequisites

A data connection has been created. For details, see **Creating Data Connections**.

Procedure

1. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Script**or **Data Development** > **Develop Job**.
2. 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 database parameters. **Table 8-1** describes the database parameters.

   NOTE

   You can create a maximum of 10 databases for a DLI data connection. No quantity limit is set on other types of data connections.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-1 Creating a database

Parameter	Mandatory	Description
Database Name	Yes	Name of the database. The naming rules are as follows: • : The value must consist of 1 to 128 characters and contain only letters, digits, and underscores (_). It must start with a digit or letter and cannot contain only digits. • DWS: The value must consist of 1 to 63 characters and contain only letters, digits, underscores (_), and dollar signs (**$**). It must start with a letter or an underscore and cannot contain only digits. • MRS Hive: The value must consist of 1 to 128 characters and contain only letters, digits, and underscores (_). It must start with a digit or letter and cannot contain only digits.
Description	No	Descriptive information about the database. The requirements are as follows: • 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.

3. Click **OK**.

8.2.2 Modifying a Database

After creating a database, you can modify the description of the DWS or MRS Hive database as required.

Procedure

1. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Script**or **Data Development** > **Develop Job**.
2. In the menu on the left, click . Right-click the database that you want to edit, and choose **Edit** from the shortcut menu.
3. In the displayed dialog box, modify the description of the database.
4. Click **Yes**.

8.2.3 Deleting a Database

If you do not need to use a database any more, perform the following operations to delete it. Deleted databases cannot be recovered. Exercise caution when performing this operation.

 Copyright © Huawei Technologies Co., Ltd.

Prerequisites

The database that you want to delete is not used and is not associated with any data tables.

Procedure

1. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Script**or **Data Development** > **Develop Job**.
2. In the menu on the left, click , right-click the database that you want to delete, and choose **Delete** from the shortcut menu.
3. In the displayed dialog box, click **OK**.

8.2.4 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 Data Connections**.
- A DWS database has been created. For details, see **Creating a Database**.

Procedure

1. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Script**or **Data Development** > **Develop Job**.
2. In the menu on the left, click , right-click **schemas**, and choose **Create Schema** from the shortcut menu.
3. In the displayed dialog box, configure schema parameters. **Table 8-2** describes the database schema parameters.

   Table 8-2 Creating a database schema

Parameter	Mandatory	Description
Schema Name	Yes	Name of the database schema.
Description	No	Descriptive information about the database schema.

4. Click **OK**.

8.2.5 Modifying a Database Schema

After creating a database schema, you can modify the description of the database schema as required.

Copyright © Huawei Technologies Co., Ltd.

Procedure

1. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Script**or **Data Development** > **Develop Job**.
2. In the menu on the left, click , right-click the database schema that you want to modify, and choose **Modify** from the shortcut menu.
3. In the displayed dialog box, modify the description of the database.
4. Click **Yes**.

8.2.6 Deleting a Database Schema

If you do not need to use a database schema any more, perform the following operations to delete it. Deleted database schemas cannot be recovered. Exercise caution when performing this operation.

Constraints

The default database schema cannot be deleted.

Procedure

1. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Script**or **Data Development** > **Develop Job**.
2. In the menu on the left, click , right-click the database schema that you want to delete, and choose **Delete** from the shortcut menu.
3. In the displayed dialog box, click **OK**.

8.2.7 Creating a Data Table (Visualized Mode)

You can create permanent data tables in visualized mode. After creating a data table, you can use it for job and script development.

The following types of data tables can be created:

- DLI
- DWS
- MRS Hive
- CloudTable

Prerequisites

- A corresponding cloud service has been enabled and a database has been created in the cloud service. For example, before creating a table, DLI has been enabled and a database has been created in DLI.
- A data connection that matches the data table type has been created in Data Development. For details, see **Creating Data Connections**.

Copyright © Huawei Technologies Co., Ltd.

Procedure

1. To create a data table in the right pane, perform the following steps:
   a. In the navigation tree of the data development page, choose **Development** > **Develop Script** or **Development** > **Develop Job**.
   b. In the menu on the left, click , right-click **tables**, and choose **Create Data Table** from the shortcut menu.
2. On the displayed page, configure basic properties. Specific settings vary depending on the data connection type you select. **Table 8-3** lists the links for viewing property parameters of each type of data connection.

   Table 8-3 Basic property parameters

Data Connection Type	Parameter Description
DLI	For details, see the **Basic Property** part in **Table 8-5**.
DWS	For details, see the **Basic Property** part in **Table 8-6**.
MRS Hive	For details, see the **Basic Property** part in **Table 8-7**.
CloudTable	For details, see the **Basic Property** part in **Table 8-8**.

3. Click **Next**. On the **Configure Table Structure** page, configure table structure parameters. **Table 8-4** describes the table structure parameters.

   Table 8-4 Table structure

Data Connection Type	Description
DLI	For details, see the **Table Structure** part in **Table 8-5**.
DWS	For details, see the **Table Structure** part in **Table 8-6**.
MRS Hive	For details, see the **Table Structure** part in **Table 8-7**.
CloudTable	For details, see the **Table Structure** part in **Table 8-8**.

4. Click **OK**.

 Copyright © Huawei Technologies Co., Ltd.

Parameter Description

Table 8-5 DLI data table

Parameter	Mandatory	Description
Basic Property		
Table Name	Yes	Name of the data table. Must consist of 1 to 63 characters and contain only lowercase letters, digits, and underscores (_). Cannot contain only digits or start with an underscore.
Alias	No	Alias of the data table. Must consist of 1 to 63 characters and contain only letters, digits, and underscores (_). Cannot contain only digits or start with an underscore.
Data Connection	Yes	Data connection to which the data table belongs.
Database	Yes	Database to which the data table belongs.
Data Location	Yes	Location to save data. Possible values: • OBS • DLI
Data Format	Yes	Format of data. This parameter is available only when **Data Location** is set to **OBS**. Possible values: • **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 data table.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Table Structure		
Column Name	Yes	Name of the column. Must be unique.
Type	Yes	Type of data. For details about the data types, see **Data Types** of *Data Lake Insight SQL Syntax Reference*.
Column Description	No	Descriptive information about the column.
Operation	No	To add a column, click .

Table 8-6 DWS data table

Parameter	Mandatory	Description
Basic Property		
Table Name	Yes	Name of the data table. Must consist of 1 to 63 characters and contain only letters, digits, and underscores (_). Cannot contain only digits or start with an underscore.
Alias	No	Alias of the data table. Must consist of 1 to 63 characters and contain only letters, digits, and underscores (_). Cannot contain only digits or start with an underscore.
Data Connection	Yes	Data connection to which the data table belongs.
Database	Yes	Database to which the data table belongs.
Schema	Yes	Schema of the database.
Table Description	No	Descriptive information about the data table.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Advanced Settings	No	The following advanced options are available: • Storage method of a data table. Possible values: – **Row store** – **Column store** • Compression level of a data table – Possible values if the storage method is row store: **YES** or **NO**. – Possible values if 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. Must be unique.
Data Classification	Yes	Classification of data. Possible values: • **Value** • **Currency** • **Boolean** • **Binary** • **Character** • **Time** • **Geometric** • **Network address** • **Bit string** • **Text search** • **UUID** • **JSON** • **OID**
Data Type	Yes	Type of data. For details about the data types, see **Data Types** of *Data Warehouse Service Developer Guide*.
Column Description	No	Descriptive information about the column.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
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 **Creating a Cluster** of *Cloud Search Service User Guide*.
Index Data Type	No	Data type of the ES index. The options are as follows: • text • keyword • date • long • integer • short • byte • double • boolean • binary
Operation	No	To add a column, click .

Table 8-7 Basic property parameters of an MRS Hive data table

Parameter	Mandatory	Description
Basic Property		
Table Name	Yes	Name of the data table. Must consist of 1 to 63 characters and contain only lowercase letters, digits, and underscores (_). Cannot contain only digits or start with an underscore.
Alias	No	Alias of the data table. Must consist of 1 to 63 characters and contain only letters, digits, and underscores (_). Cannot contain only digits or start with an underscore.
Data Connection	Yes	Data connection to which the data table belongs.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Database	Yes	Database to which the data table belongs.
Table Description	No	Descriptive information about the data table.
Table Structure		
Column Name	Yes	Name of the column. Must be unique.
Data Classification	Yes	Classification of data. Possible values: • **Original** • **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 .

Table 8-8 Basic property parameters of a CloudTable data table

Parameter	Mandatory	Description
Basic Property		
Table Name	Yes	Name of the data table. Must consist of 1 to 63 characters and contain only letters, digits, and underscores (_). Cannot contain only digits or start with an underscore.
Alias	No	Alias of the data table. Must consist of 1 to 63 characters and contain only letters, digits, and underscores (_). Cannot contain only digits or start with an underscore.
Data Connection	Yes	Data connection to which the data table belongs.
Namespace	Yes	Namespace to which the data table belongs.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Table Description	No	Descriptive information about the data table.
Table Structure		
Column Family Name	Yes	Name of the column family. Must be unique.
Column Family Description	No	Descriptive information about the column family.
Operation	No	To add a column, click .

8.2.8 Creating a Data Table (DDL Mode)

You can create permanent and temporary data tables in DDL mode. After creating a data table, you can use it for job and script development.

The following types of data tables can be created:

- DLI
- DWS
- MRS Hive

Prerequisites

- A corresponding cloud service has been enabled and a database has been created in the cloud service. For example, before creating a table, DLI has been enabled and a database has been created in DLI.
- A data connection that matches the data table type has been created in Data Development. For details, see **Creating Data Connections**.

Procedure

1. Perform the following steps:
   a. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Script/Data Development** > **Develop Job**.
   b. In the menu on the left, click , right-click **tables**, and choose **Create Data Table** from the shortcut menu.
2. Click **DDL-based Table Creation**, configure parameters described in **Table 8-9**, and enter SQL statements in the editor in the lower part.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-9 Data table parameters

Parameter	Description
Data Connection Type	Type of data connection to which the data table belongs. • DLI • DWS • HIVE
Data Connection	Data connection to which the data table belongs.
Database	Database to which the data table belongs.

3. Click **OK**.

8.2.9 Viewing Data Table Details

After creating a data table, you can view the basic information, storage information, field information, and preview data of the data table.

Procedure

1. Perform the following steps:
   a. In the navigation tree of the data development page, choose **Development** > **Develop Script** or **Development** > **Develop Job**.
   b. In the menu on the left, click , right-click the data table that you want to view, and choose **View Details** from the shortcut menu.
2. In the displayed dialog box, view the data table information. See **Table 8-10**.

Table 8-10 Table details page

Tab Name	Description
Table Information	Displays the basic information and storage information about the data table.
Field Information	Displays the field information about the data table.
Data Preview	Displays 10 records about the data table.
DDL	Displays the DDL of the DLI, MRS Hive or DWS data table.

8.2.10 Deleting a Data Table

If you do not need to use a data table any more, perform the following operations to delete it.

 Copyright © Huawei Technologies Co., Ltd.

Procedure

1. Perform the following steps:
   a. In the navigation tree of the data development page, choose **Development** > **Develop Script** or **Development** > **Develop Job**.
   b. In the menu on the left, click , right-click the data table that you want to delete, and choose **Delete** from the shortcut menu.
2. In the displayed dialog box, click **OK**.

8.2.11 Viewing Table Details

You can view the column information of a data table on the right area of the page.

Procedure

1. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Script**or **Data Development** > **Develop Job**.
2. In the menu on the left, click , and expand the data connection directory to view column information under a desired data table.

8.3 Script Development

8.3.1 Creating a Script

DLF allows you to edit, debug, and run scripts online. You must add a script before developing it.

(Optional) Creating a Directory

If a directory exists, you do not need to create one.

1. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Script**.
2. In the directory list, right-click a directory and choose **Create Directory** from the shortcut menu.
3. In the displayed dialog box, configure directory parameters. **Table 8-11** describes the directory parameters.

Table 8-11 Script directory parameters

Parameter	Description
Directory Name	Name of the script directory. Must consist of 1 to 32 characters and contain only letters, digits, underscores (_), and hyphens (-).

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Select Directory	Parent directory of the script directory. The parent directory is the root directory by default.

4. Click **OK**.

Creating a Script

Currently, you can create the following types of scripts in DLF:

- DLI SQL
- Hive SQL
- DWS SQL
- Spark SQL
- Flink SQL
- RDS SQL
- Shell

Prerequisites

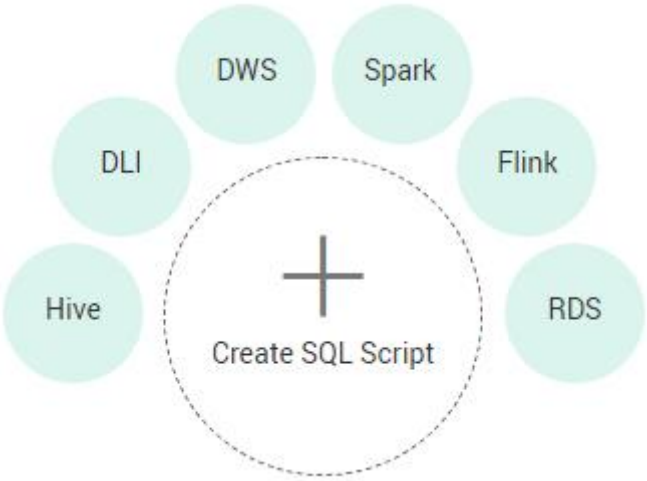
The quantity of scripts is less than the maximum quota (1,000).

Procedure

1. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Script**.
2. Create a script using either of the following methods:

   Method 1: On the right area of the page, click **Create SQL Script**/**Create Shell Script**.

   Figure 8-1 Creating an SQL script (method 1)

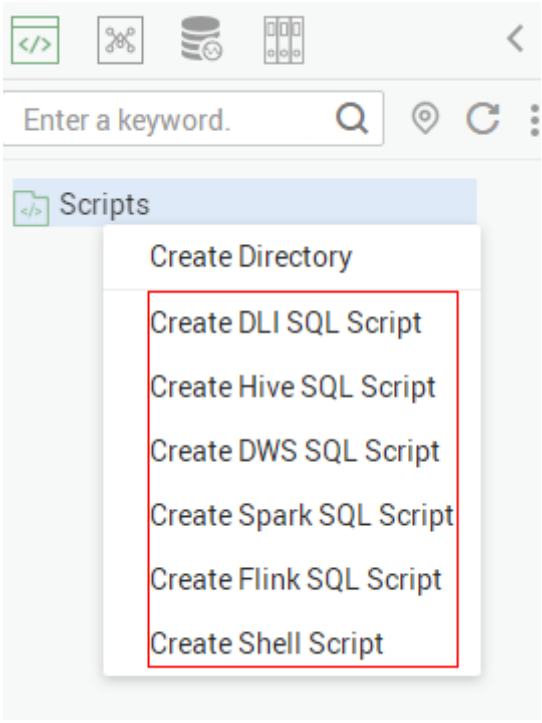


 Copyright © Huawei Technologies Co., Ltd.

Figure 8-2 Creating a shell script (method 1)

Method 2: In the directory list, right-click a directory and choose **Create Script** from the shortcut menu.

Figure 8-3 Creating a script (method 2)



3. Go to the script development page. For details, see **Developing an SQL Script** and **Developing a Shell Script**.

8.3.2 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. For example, before developing a DLI script, DLI has been enabled and a database has been created in DLI. The Flink SQL script is not involved.

 Copyright © Huawei Technologies Co., Ltd.

- A data connection that matches the data connection type of the script has been created in Data connection. For details, see **Creating Data Connections**. Flink SQL scripts are not involved.
- An SQL script has been added. For details, see **Creating a Script**.

Procedure

1. In the navigation tree on the left of the data development page, 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. In the upper part of the editor, select script properties. **Table 8-12** describes the script properties. When developing a Flink SQL script, skip this step.

Table 8-12 SQL script properties

Property	Description
Data Connection	Selects a data connection.
Database	Name of the database.

 Copyright © Huawei Technologies Co., Ltd.

Property	Description
Resource Queue	Selects a resource queue for executing a DLI job. Set this parameter when a DLI or SQL script is created. You can create a resource queue using either of the following methods: • Click + to go to the **Queue Management** page. • Go to the DLI console. **NOTE** DLI provides the default resource queue **default**, which does not support insert, load, or cat commands. 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: • **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.

4. Enter an SQL statement in the editor. You can enter multiple SQL statements. To facilitate script development, Data Development provides system functions and script parameters (Flink SQL and RDS SQL scripts are excluded).

   NOTE

   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.
   ```

   – System Functions

     To view the functions supported by this type of data connection, click **System Functions** on the right of the editor. You can drag a function to the editor to use it.

 Copyright © Huawei Technologies Co., Ltd.

- 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**).

  An example is as follows:

  ```
  select ${str1} from data;
  ```

  In the preceding command, *str1* 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.

5. (Optional) In the upper part of the editor, click **Format** to format the SQL statement. Skip this step because Flink SQL scripts are not involved.
6. In the upper part of the editor, click **Execute**. If you need to run some SQL statements separately, select the SQL statements first. After the SQL statement is executed, you can view the execution history and execution result of the script in the lower part of the editor. When developing a Flink SQL script, skip this step.
7. Above the editor, click to save the script.

   If the script is created but not saved, set the parameters listed in **Table 8-13**.

Table 8-13 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.

Downloading or Dumping a Script Execution Result

You can download or dump execution results as required. Only users with the **DGC Administrator** or **Tenant Administrator** policy can download or dump execution results. Download address:

- Download result: Download the CSV result files to the local host.
- Dump result: Dump the CSV result files to OBS. For details, see **Table 8-14**.

  NOTE

  The execution results of Flink SQL scripts, RDS SQL scripts, and Shell scripts cannot be dumped.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-14 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.
Compression Format	No	Format of compression. Set this parameter when a DLI or SQL script is created. • **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. • **Overwrite**: The existing folder will be overwritten by the customized folder. • **Report**: The system reports an error and suspends the export operation.

8.3.3 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 **Managing a Host Connection**.

Procedure

1. In the navigation tree on the left of the data development page, 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. In the upper part of the editor, select script properties. **Table 8-15** describes the script properties.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-15 Shell script properties

Parameter	Description	Example
Host Connection	Selects the host where a shell script is to be executed.	N/A
Parameter	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.	For example, if you enter the following shell interactive script and the interactive parameters **1**, **2**, and **3** correspond to **begin**, **end**, and **exit**, you need to enter parameters **1**, **2**, and **3**. #!/bin/bash select ch in "begin" "end" "exit"; do case $ch in "begin") echo "start something" ;; "end") echo "stop something" ;; "exit") echo "exit" break; ;; *) echo "Ignorant" ;; esac done
Interactive Input	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.	N/A

4. Edit shell statements in the editor.

   To facilitate script development, the Data Development module provides the script parameter function. The usage method is as follows:

   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}
   ```

 Copyright © Huawei Technologies Co., Ltd.

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.

5. 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.
6. Above the editor, click to save the script.

   If the script is created but not saved, set the parameters listed in **Table 8-16**.

Table 8-16 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.

8.3.4 Renaming a Script

You can change the name of a script.

This section describes how to rename a script.

Prerequisites

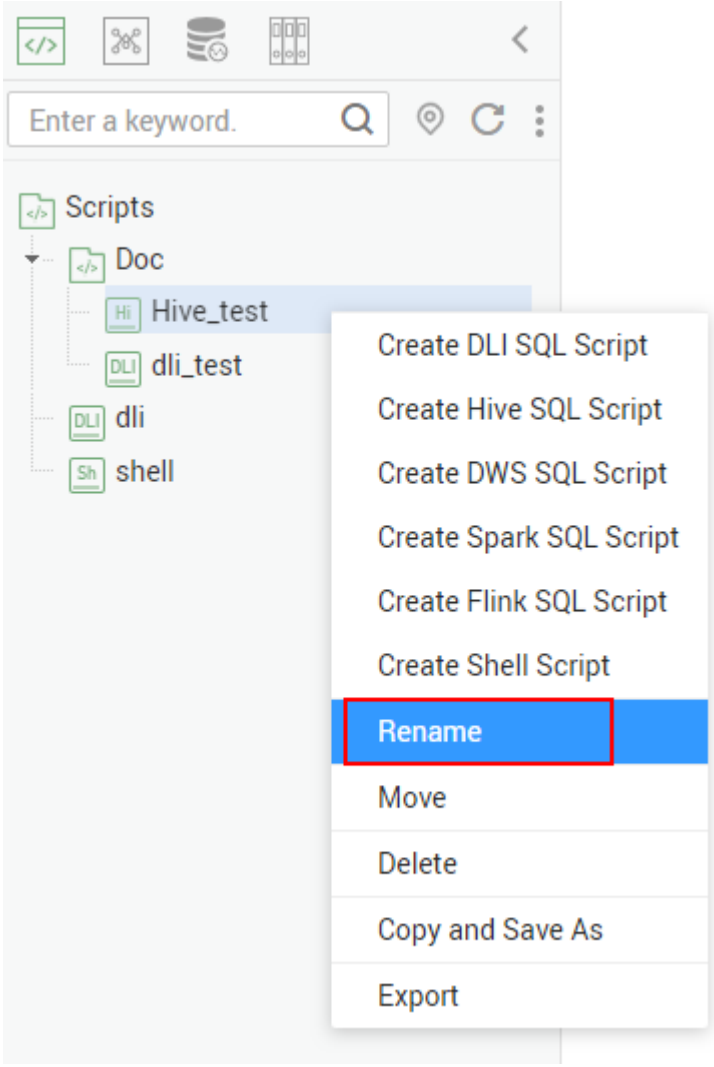
- You have developed a script. The script to be renamed exists in the script directory.
- For details about how to develop scripts, see **Developing an SQL Script** and **Developing a Shell Script**.

Procedure

1. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Script**.
2. In the script directory, select the script to be renamed. Right-click the script name and choose **Rename** from the shortcut menu.

 Copyright © Huawei Technologies Co., Ltd.

Figure 8-4 Choosing Rename

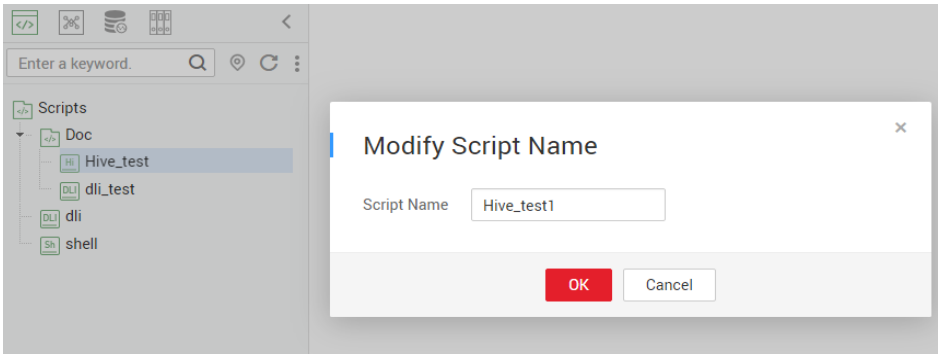


NOTE

An opened script file cannot be renamed.

3. On the page that is displayed, configure related parameters. **Table 8-17** describes the parameters.

Figure 8-5 Renaming a script



 Copyright © Huawei Technologies Co., Ltd.

Table 8-17 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.

4. Click **OK**.

8.3.5 Moving a Script

You can move a script from the current directory to another directory.

This section describes how to move a script.

Prerequisites

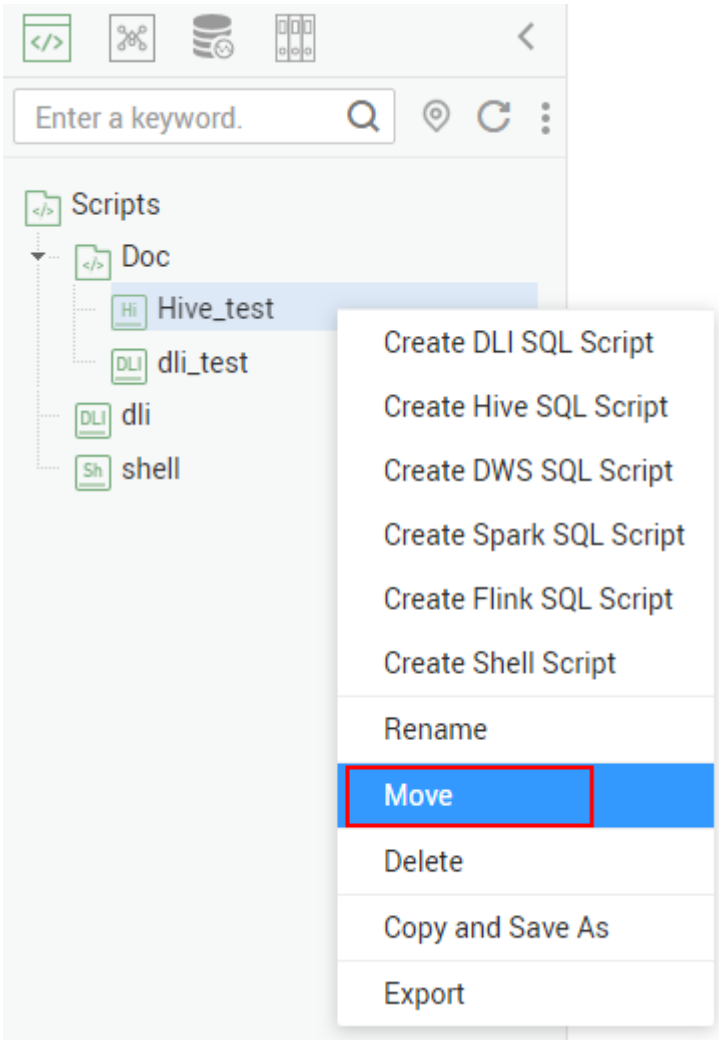
- You have developed a script. The script to be moved exists in the script directory.
- For details about how to develop scripts, see **Developing an SQL Script** and **Developing a Shell Script**.

Procedure

1. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Script**.
2. In the script directory, select the script to be moved. Right-click the script name and choose **Move** from the shortcut menu.

 Copyright © Huawei Technologies Co., Ltd.

Figure 8-6 Choosing Move



3. In the displayed dialog box, configure related parameters. **Table 8-18** describes the parameters.

 Copyright © Huawei Technologies Co., Ltd.

Figure 8-7 Moving a script

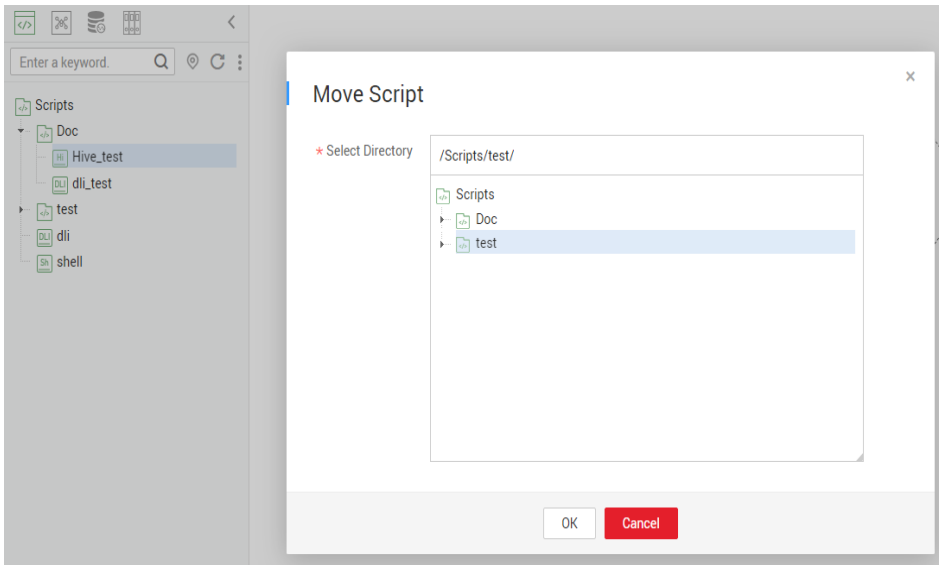


Table 8-18 Script moving parameters

Parameter	Description
Select Directory	Directory to which the script is to be moved. The parent directory is the **root** directory by default.

4. Click **OK**.

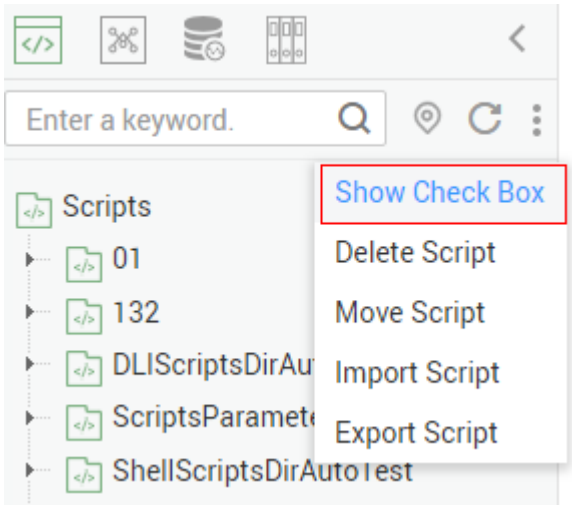
8.3.6 Exporting and Importing a Script

Exporting a Script

You can export one or more script files from the script directory.

1. Click in the script directory and select **Show Check Box**.

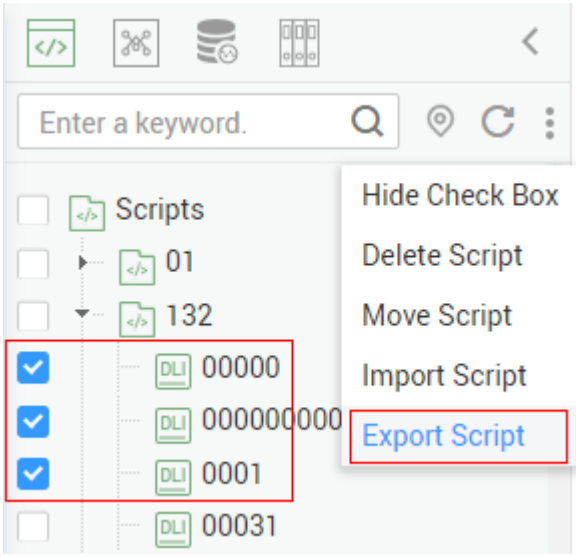
Figure 8-8 Clicking Show Check Box



 Copyright © Huawei Technologies Co., Ltd.

2. Select the scripts to be exported, click , and choose **Export Script**.

Figure 8-9 Selecting and exporting scripts

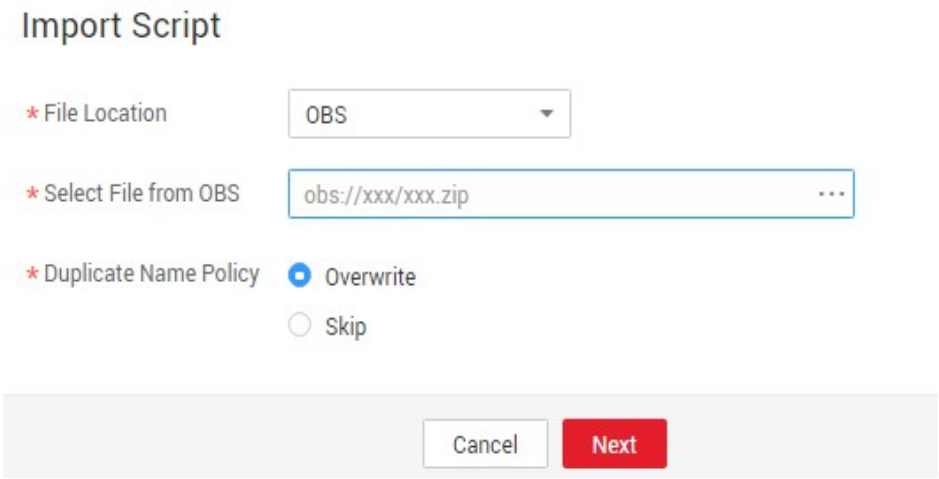


Importing a Script

You can import one or more script files in the script directory.

1. Click and choose **Import Script** in the script directory, select the script file that has been uploaded to OBS, and set **Duplicate Name Policy**.

Figure 8-10 Importing scripts (with OBS)



2. Click **Next**.

8.3.7 Deleting a Script

If you do not need to use a script any more, perform the following operations to delete it.

 Copyright © Huawei Technologies Co., Ltd.

NOTICE

If you forcibly delete a script that is being associated with a job, ensure that services are not affected by going to the job development page and reassociating an available script with the job.

Deleting a Script

1. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Script**.
2. In the script directory, right-click the script that you want to delete and choose **Delete** from the shortcut menu.
3. In the displayed dialog box, click **OK**.

Batch Deleting Scripts

1. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Script**.
2. On the top of the script directory, click and select **Show Check Box**.
3. Select the scripts to be deleted, click , and select **Batch Delete**.
4. In the displayed dialog box, click **OK** to delete scripts in batches.

8.3.8 Copying a Script

This section describes how to copy a script.

Prerequisites

The script file to be copied exists in the script directory.

Procedure

1. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Script**.
2. In the script directory, select the script to be copied, right-click the script name, and choose **Copy Save As**.
3. In the displayed dialog box, configure related parameters. **Table 8-19** describes the parameters.

Table 8-19 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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Select Directory	Parent directory of the script directory. The parent directory is the root directory by default.

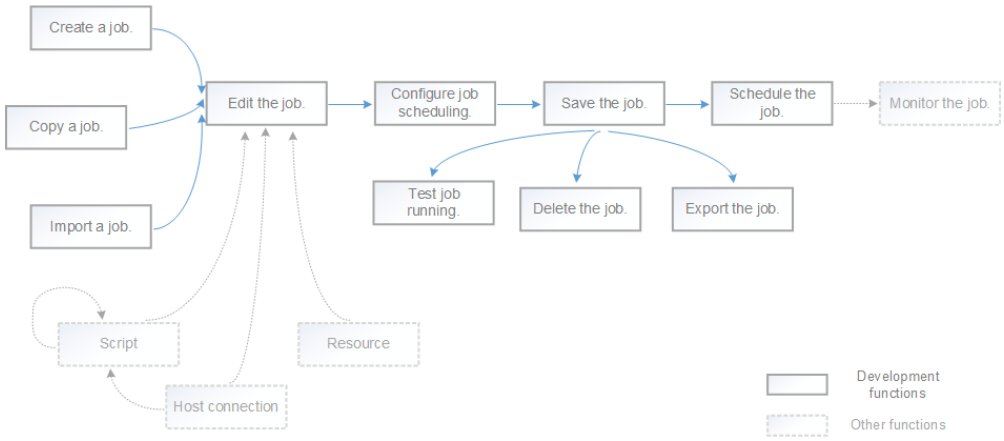
4. Click **OK**.

8.4 Job Development

8.4.1 Job Development Overview

Figure 8-11 illustrates the process for developing a job in the Data Development module.

Figure 8-11 Job development process



- **Creating a Job**

  Two job processing modes are provided: batch processing and real-time processing. For details, see **Creating a Job**.

- **Copying a Job**

  An existing job can be copied for development. For details, see **Copying a Job**.

- **Importing a Job**

  A job file in ZIP format can be imported from the local host to Data Development and be modified using the editor. For details, see **Exporting and Importing a Job**.

- Editing a Job

  Nodes can be orchestrated and configured according to service requirements. For details, see **Developing a Job**.

- **Configuring Task Scheduling**

  – 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 **Configuring Job Scheduling Tasks**.

 Copyright © Huawei Technologies Co., Ltd.

  - If the processing mode of a job is real-time processing, configure scheduling types for nodes in jobs. Three scheduling types are supported: run once, run periodically, and event-based. For details, see **Configuring Node Scheduling Tasks**.

- **Saving a Job**

  Jobs can be saved in a timely manner to avoid losing job configuration information. For details, see **Saving and Starting Jobs**.

- **Testing Job Run**

  Jobs can be manually triggered to run for a run test. You can view the running process and results at the bottom of the canvas. This function is mainly used for debugging jobs.

- **Deleting a Job**

  Jobs that are no longer used can be deleted to reduce resource and quota usage. Multiple jobs can be deleted at the same time. For details, see **Deleting a Job**.

- **Exporting a Job**

  To back up or share jobs, you can export job files to the local host. For details, see **Exporting and Importing a Job**.

- **Scheduling a Job**

  After a job is developed, you can execute the job.

8.4.2 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.

(Optional) Creating a Directory

If a directory exists, you do not need to create one.

1. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Job**.
2. In the directory list, right-click a directory and choose **Create Directory** from the shortcut menu.
3. In the displayed dialog box, configure directory parameters. **Table 8-20** describes the directory parameters.

Table 8-20 Job directory parameters

Parameter	Description
Directory Name	Name of the job directory. Must consist of 1 to 32 characters and contain only letters, digits, underscores (_), and hyphens (-).
Select Directory	Parent directory of the job directory. The parent directory is the root directory by default.

Copyright © Huawei Technologies Co., Ltd.

4. Click **OK**.

Creating a Job

The quantity of jobs is less than the maximum quota (10,000).

1. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Job**.
2. Create a job using either of the following methods:

   Method 1: On the right area of the page, click **Create Job**.

   Method 2: In the directory list, right-click a directory and choose **Create Job** from the shortcut menu.
3. In the displayed dialog box, configure job parameters. **Table 8-21** describes the job parameters.

Table 8-21 Job 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 (.).
Processing Mode	Type of the job. • **Batch**: Data is processed periodically in batches based on the scheduling plan, which is used in scenarios with low real-time requirements. • **Real-Time**: Data is processed in real time, which is used in scenarios with high real-time performance.
Creation Method	Selects a job creation mode. • Create Empty Job: Create an empty job. • Create Based on Template: Create a job using a template.
Select Directory	Directory to which the job belongs. The root directory is selected by default.
Job Owner	Owner of the job.
Job Priority	Priority of the job. The value can be **High**, **Medium**, or **Low**.
Agency	After an agency is configured, the job interacts with other services as an agency during job execution. If an agency has been configured for the workspace, the new job uses the agency by default. For details, see **Configuring a Workspace-Level Agency**. You can also change the agency to a job-level agency by referring to **Configuring a Job-level Agency**. **NOTE** Job-level agency takes precedence over workspace-level agency.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Log Path	Selects the OBS path to save job logs. By default, logs are stored in a bucket named **dlf-log-**{*Projectid*}. **NOTE** • If you want to customize a storage path, select the bucket that you have created on OBS by referring to the instructions in **Changing Log Storage Paths**. • 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.

4. Click **OK**.

8.4.3 Developing a Job

DLF allows you to develop existing jobs.

Prerequisites

You have created a job. For details about how to create a job, go to **Creating a Job**.

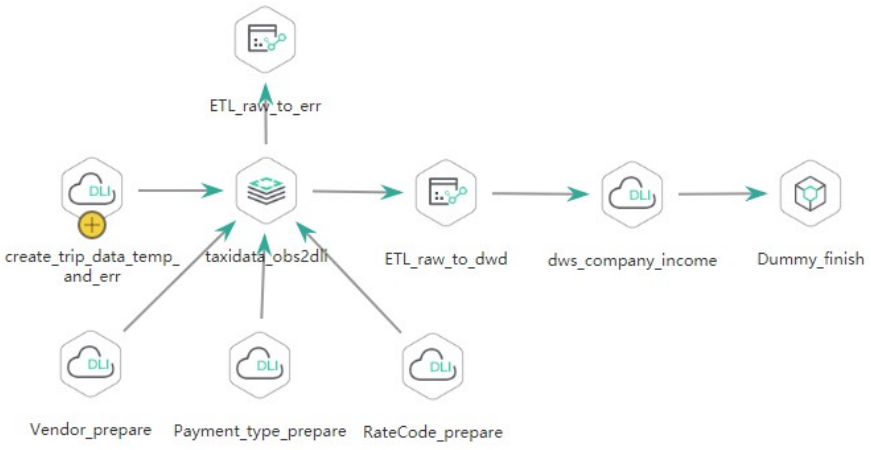
Compiling Job Nodes

1. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Job**.
2. In the job directory, double-click a job that you want to develop. The job development page is displayed.
3. Drag the desired node to the canvas, move the mouse over the node, and select the icon and drag it to connect to another node.

   NOTE

   Each job can contain a maximum of 200 nodes.

Figure 8-12 Compiling a Job



 Copyright © Huawei Technologies Co., Ltd.

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:

Select a job. On the job development page, click the **Basic Job Information** tab. On the displayed page, configure parameters. **Table 8-22** describes the parameters.

Table 8-22 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.
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:

Select a job. On the job development page, click the **Job Parameter Setup** tab. On the displayed page, configure parameters. **Table 8-23** describes the parameters.

Table 8-23 Job parameter setup

Function	Description
Variable Parameter	

 Copyright © Huawei Technologies Co., Ltd.

Function	Description
Add	Click **Add** and enter the variable parameter name and parameter value in the text boxes. ● Parameter Name The parameter name must be unique, consist of 1 to 64 characters, and contain only letters, digits, underscores (_), hyphens (-), less-than signs (<), and greater-than signs (>). ● Parameter Value – The string type of parameter value is a character string. For example: **str1** When a character string and function are used together, use **@@** to enclose the character string and use **+** to connect the character string and function. For example: **@@str1@@+ $getCurrentTime(@@yyyyMMdd@@,0)** – The numeric type of parameter value is a number or operation expression. After the parameter is configured, it is referenced in the format of **$**{*parameter name*} in the job.
Modify	Modify the parameter name and parameter value in text boxes and save the modifications.
Save	Click **Save** to save the settings.
Delete	Click next to the parameter name and value text boxes to delete the job parameter.
Constant Parameter	

 Copyright © Huawei Technologies Co., Ltd.

Function	Description
Add	Click **Add** and enter the constant parameter name and parameter value in the text boxes. ● Parameter name The parameter name must be unique, consist of 1 to 64 characters, and contain only letters, digits, underscores (_), hyphens (-), less-than signs (<), and greater-than signs (>). ● Parameter Value – The function type of parameter value starts with a dollar sign (**$**). For example: **$getCurrentTime(@@yyyyMMdd@@,0)** – The string type of parameter value is a character string. For example: **str1** When a character string and function are used together, use **@@** to enclose the character string and use **+** to connect the character string and function. For example: **@@str1@@+ $getCurrentTime(@@yyyyMMdd@@,0)** – The numeric type of parameter value is a number or operation expression. After the parameter is configured, it is referenced in the format of **$**{*parameter name*} in the job.
Modify	Modify the parameter name and parameter value in text boxes and save the modifications.
Save	Click **Save** to save the settings.
Delete	Click next to the parameter name and value text boxes to delete the job parameter.

Configuring Job Scheduling Tasks

You can configure job scheduling tasks for batch jobs. There are three scheduling types available: **Run once**, **Run periodically**, and **Event-driven**. The procedure is as follows:

Select a job. On the job development page, click the **Scheduling Parameter Setup** tab. On the displayed page, configure parameters. **Table 8-24** describes the parameters.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-24 Scheduling parameter setup

Parameter	Description
Schedule Type	Job schedule type. Possible values: • **Run once**: The job will be run only once. • **Run periodically**: The job will be run periodically. • Event-driven: The job will be run when certain external conditions are met.
Parameters for **Run periodically**	
Effective Time	Period during which a job runs.
Schedule Cycle	Frequency at which a job is run. The job can be run once every: • **Minute** • **Hour** • **Day** • **Week** • **Month**
Dependency Job	Job that is depended on. The constraints are as follows: • A short-cycle job cannot depend on a long-cycle job. • A job whose schedule cycle is **Week** cannot depend on a job whose schedule cycle is **Minute**. • A job whose schedule cycle is **Week** cannot depend on or be depended on by another job. • A job whose schedule cycle is **Month** can depend only on the job whose schedule cycle is **Day**.
Action After Dependency Job Failure	Action that will be performed on the current job if the dependency job fails to be executed. Possible values: • **Suspend** The current job will be suspended. A suspended job will block the execution of the subsequent jobs. You can manually force the dependency job to succeed to solve the blocking problem. For details, see **Monitoring a Batch Job**. • **Continue** The current job will continue to be performed. • **Terminate** The current job will be terminated. Then the status of the current job will become **Canceled**.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Run upon completion of the dependency job's last schedule cycle.	Specify whether, before executing the current job, the dependency job's last schedule cycle must end.
Cross-Cycle Dependency	Dependency between instances in the job. Possible values: • **Independent on the previous schedule cycle** • **Self-dependent** (The current job can continue to run only after the previous schedule cycle ends.)
Parameters for **Event-driven**	
Triggering Event Type	Type of the event that triggers job running. • DIS • OBS
DIS Stream Name	Name of the DIS stream. When a new message is sent to the specified DIS stream, Data Development transfers the new message to the job to trigger the job running.
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 DIS stream or OBS path for new messages. The unit of the interval can be set to second or minute.
Access Policy	Select the location where data is to be read: • Access from the previous location: For the first access, data is accessed from the most recently recorded location. For the subsequent access, data is accessed from the previously recoded location. • Access from the latest location: Data is accessed from the most recently recorded location each time.
Failure Policy	Select a policy to be performed after scheduling fails. • Stop scheduling • Ignore failure and proceed

Configuring Node Properties

Click a node in the canvas. On the displayed **Node Properties** page, configure node properties. For details, see **Node Overview**.

 Copyright © Huawei Technologies Co., Ltd.

Configuring Node Scheduling Tasks

You can configure node scheduling tasks for real-time jobs. There are three scheduling types available: **Run once**, **Run periodically**, and **Event-driven**. The procedure is as follows:

Select a node. On the node development page, click the **Scheduling Parameter Setup** tab. On the displayed page, configure parameters. **Table 8-25** describes the parameters.

Table 8-25 Node scheduling setup

Parameter	Description
Schedule Type	Job schedule type. Possible values: • **Run once**: The job will be run only once. • **Run periodically**: The job will be run periodically. • Event-driven: The job will be run when certain external conditions are met.
Parameters for **Run periodically**	
Effective Time	Period during which a job runs.
Schedule Cycle	Frequency at which a job is run. The job can be run once every: • **Minute** • **Hour** • **Day** • **Week** • **Month**
Cross-Cycle Dependency	Dependency between instances in the job. Possible values: • **Independent on the previous schedule cycle** • **Self-dependent** (The current job can continue to run only after the previous schedule cycle ends.)
Parameters for **Event-driven**	
Triggering Event Type	Type of the event that triggers job running.
DIS Stream Name	Name of the DIS stream. When a new message is sent to the specified DIS stream, Data Development transfers the new message to the job to trigger the job running.
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 DIS stream for new messages. The unit of the interval can be set to second or minute.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Failure Policy	Select a policy to be performed after scheduling fails. • Suspend • Ignore failure and proceed

More Node Functions

To obtain more node functions, right-click the node icon in the canvas and choose a function as required. **Table 8-26** describes the node functions.

Table 8-26 More mode functions

Function	Description
Configure	Goes to the **Node Property** page of the node.
Delete	Deletes one or more nodes at the same time. • 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	Copies one or more nodes to any job. • 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.
Add Note	Adds a note to the node. Each node can have multiple notes.
Edit Script	Goes to the script editing page and edits the associated script. This option is available only for the node associated with a script.

Saving and Starting Jobs

After a job is configured, complete the following operations:

 Copyright © Huawei Technologies Co., Ltd.

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.

Step 3 Click to start the job.

----End

Processing jobs in real time

Step 1 Click to save the job configuration.

Step 2 Click to submit and start the job.

NOTE

Data Development allows you to modify a running real-time processing job. After the job is modified, click **Save**. On the page that is displayed, click **OK** to save the job.

----End

8.4.4 Renaming a Job

You can change the name of a job.

This section describes how to rename a job.

Prerequisites

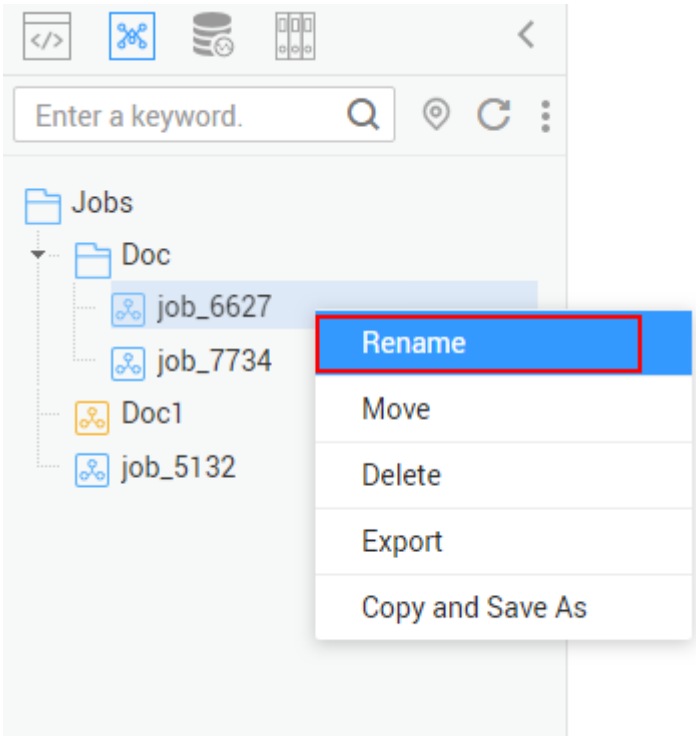
You have created a job. The job to be renamed exists in the job directory. For details about how to create a job, see **Creating a Job**.

Procedure

1. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Job**.
2. In the job directory, select the job to be renamed. Right-click the job name and choose **Rename** from the shortcut menu.

 Copyright © Huawei Technologies Co., Ltd.

Figure 8-13 Renaming a job



3. On the page that is displayed, configure related parameters. **Table 8-27** describes the parameters.

Figure 8-14 Renaming a job

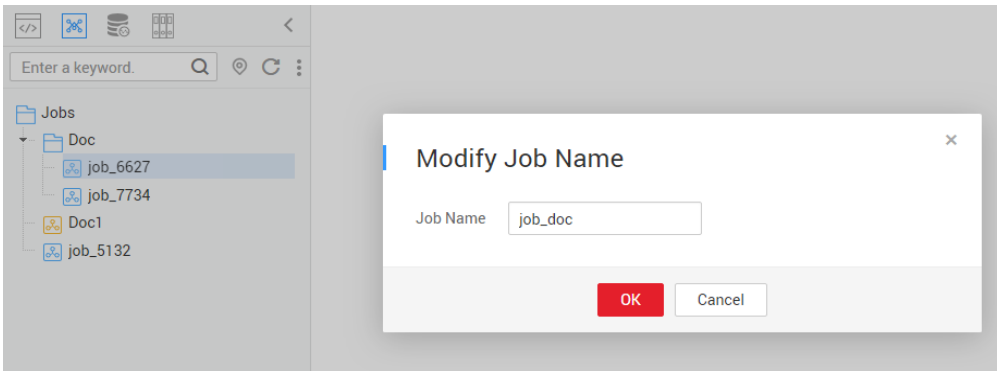


Table 8-27 Job renaming parameters

Parameter	Description
Job Name	Name of the job. Must consist of 1 to 128 characters and contain only letters, digits, hyphens (-), underscores (_), and periods (.).

4. Click **OK**.

 Copyright © Huawei Technologies Co., Ltd.

8.4.5 Moving a Job

You can move a job from the current directory to another directory.

This section describes how to move a job.

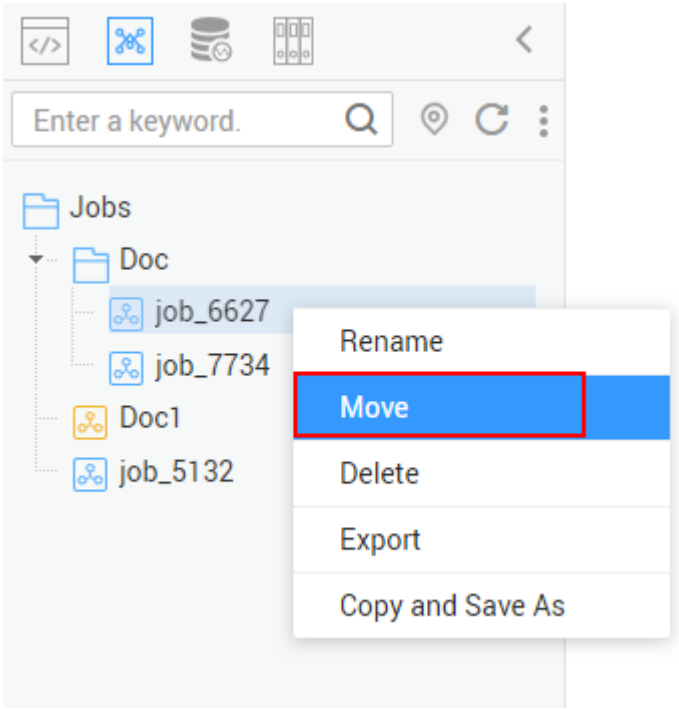
Prerequisites

You have created a job. The job to be moved exists in the job directory. For details about how to create a job, see **Creating a Job**.

Procedure

1. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Job**.
2. In the job directory, select the job to be moved. Right-click the job name and choose **Move** from the shortcut menu.

   Figure 8-15 Selecting a job to be moved



3. In the displayed dialog box, configure related parameters. **Table 8-28** describes the parameters.

 Copyright © Huawei Technologies Co., Ltd.

Figure 8-16 Moving a job

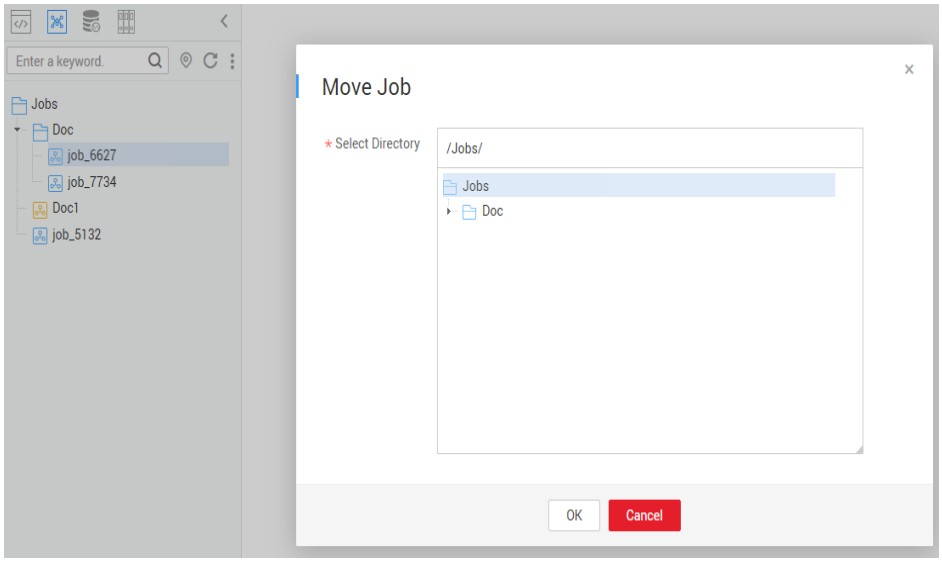


Table 8-28 Job moving parameters

Parameter	Description
Select Directory	Directory to which the job is to be moved. The parent directory is the **root** directory by default.

4. Click **OK**.

8.4.6 Exporting and Importing a Job

Exporting Jobs

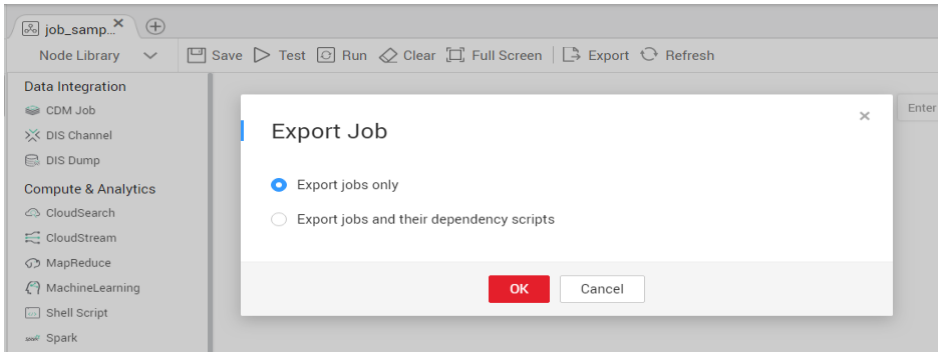
Method 1: Export a job on the job development page.

Step 1 Go to the development page of a job, click , and select the type of the job to be exported.

- **Export jobs only**: Export the connection relationships and property configurations of nodes, excluding sensitive information such as passwords. After the export is completed, the user obtains a .zip file.
- **Export jobs and their dependency scripts and resource definitions**: Export the node connection relationships, node property configurations, job scheduling configurations, parameter configurations, dependency scripts, and resource definitions, excluding sensitive information such as passwords. After the export is completed, the user obtains a .zip file.

 Copyright © Huawei Technologies Co., Ltd.

Figure 8-17 Exporting a job (method 1)



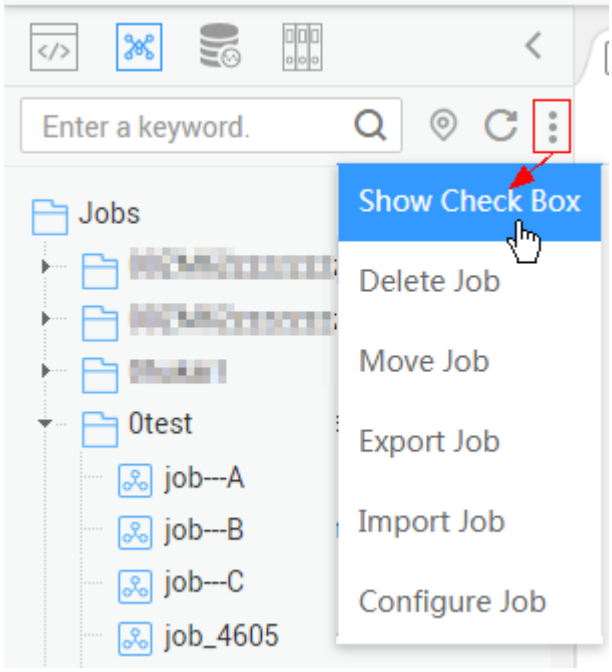
Step 2 Click **OK** to export the required job file.

----End

Method 2: Export one or more jobs from the job directory.

Step 1 Click in the job directory and select **Show Check Box**.

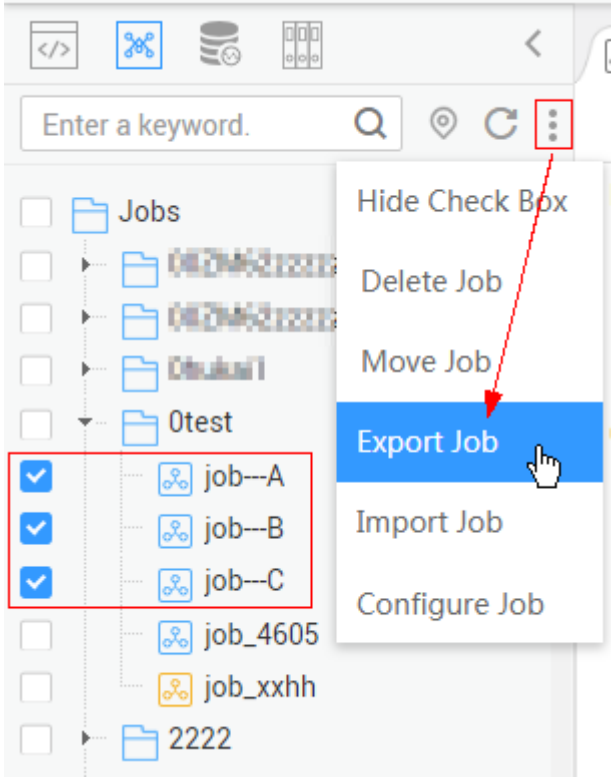
Figure 8-18 Show Check Box 1



Step 2 Select the job to be exported and choose > **Export Job** to export the job and its dependencies.

 Copyright © Huawei Technologies Co., Ltd.

Figure 8-19 Selecting and exporting a job



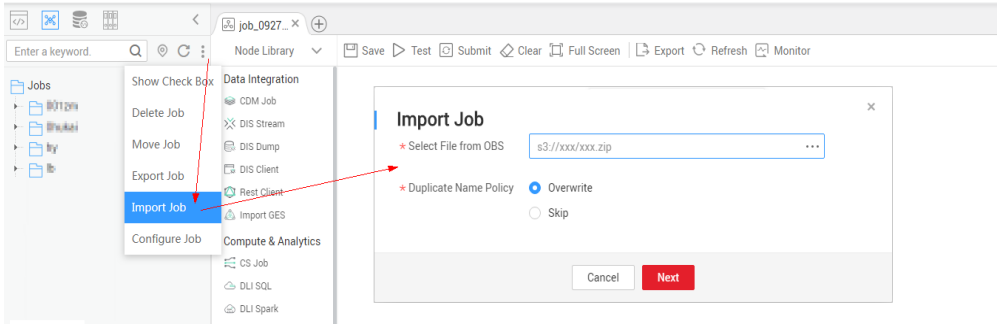
----End

Importing a Job

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.

Figure 8-20 Importing job definitions and dependencies (with OBS)



Step 2 Click **Next** to import the job as instructed.

 Copyright © Huawei Technologies Co., Ltd.

NOTE

During the import, if the data connection, DIS stream, DLI queue, or GES graph associated with the job does not exist in Data Development, the system prompts you to select one again.

----End

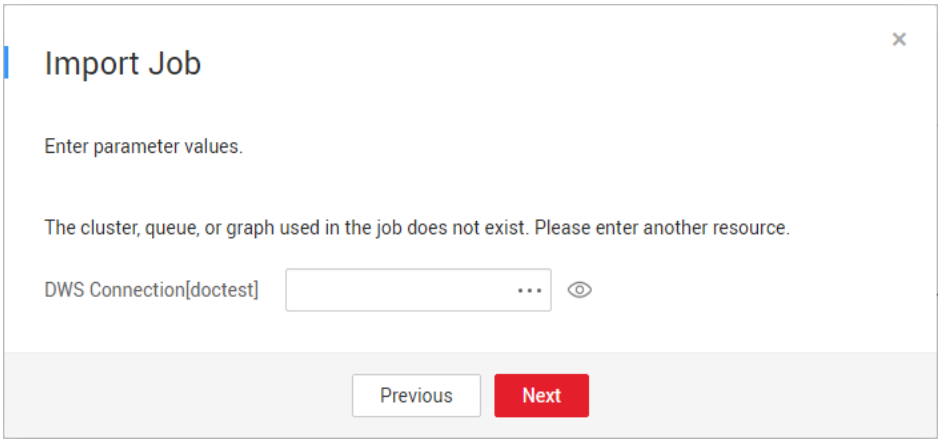
Example

Context:

- A DWS data connection **doctest** is created in Data Development.
- 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. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Job**.
2. Search for **doc1** in the job search box, export the job to the local host, and then upload it to the OBS folder.
3. Delete the **doctest** data connection associated with the job in Data Development.
4. Click > **Import Job** in the job directory, select the job file that has been uploaded to OBS, and set the duplicate name policy.
5. Click **Next** and select another data connection as prompted.

Figure 8-21 Importing a job



6. Click **Next** and then **Close**.

8.4.7 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.

 Copyright © Huawei Technologies Co., Ltd.

Deleting a Script

1. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Job**.
2. In the job directory, right-click the job that you want to delete and choose **Delete** from the shortcut menu.
3. In the displayed dialog box, click **OK**.

Batch Deleting Scripts

1. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Job**.
2. On the top of the job directory, click and select **Show Check Box**.
3. Select the jobs to be deleted, click , and select **Batch Delete**.
4. In the displayed dialog box, click **OK**.

8.4.8 Copying a Job

This section describes how to copy a job.

Prerequisites

The job file to be copied exists in the job directory.

Procedure

1. In the navigation tree on the left of the data development page, choose **Data Development** > **Develop Job**.
2. In the job directory, select the job to be copied, right-click the job name, and choose **Copy Save As**.
3. In the displayed dialog box, configure related parameters. **Table 8-29** describes the parameters.

   Table 8-29 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.

4. Click **OK**.

8.5 Nodes

Copyright © Huawei Technologies Co., Ltd.

8.5.1 Node Overview

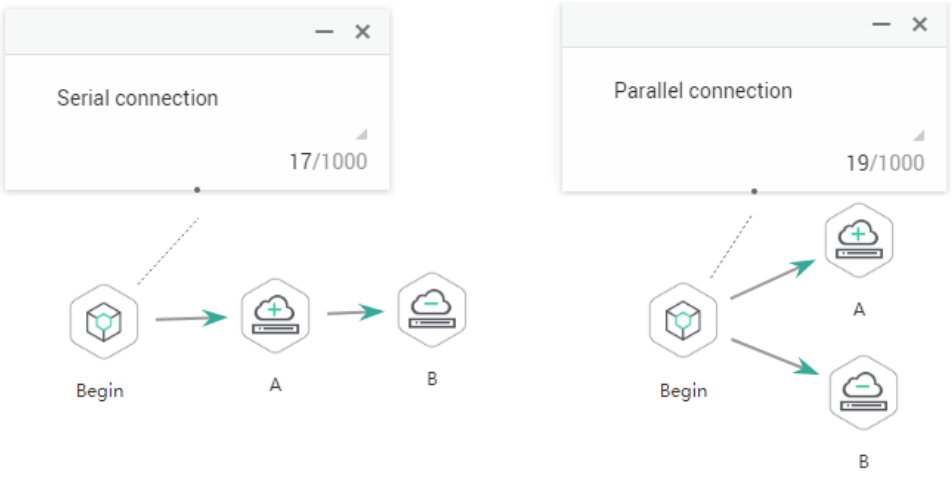
A node defines operations performed on data. Data Development provides types of node used for data integration, computing and analysis, database operations, and resource management. Users can select required nodes based on service models.

- 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, run node B only after node A is finished running.

  Parallel connection: Nodes are run at the same time.

  Figure 8-22 Connection diagram



8.5.2 Node Lineages

Implementation Solution

On the DGC platform, data lineages are generated by configuring data processing and migration nodes in the Data Development module. Currently, the system collects the lineages generated by static node configuration and the lineages on some node instances.

Node Types Supported by Data Lineage

- **SQL nodes**

  DGC supports lineage parsing of DLI SQL, DWS SQL and MRS Hive SQL nodes. It supports multi-SQL parsing and column-level lineage parsing. If a statement contains a temporary table, the temporary table entity is automatically created in the Data Assets module.

  - **DLI SQL**
    - Support for parsing the lineages generated during data insertion between DLI tables.

 Copyright © Huawei Technologies Co., Ltd.

        - Support for lineages between OBS files generated by table creation statements and DLI tables.
      - **DWS SQL**
        - Support for lineages between DWS tables generated by DDL operations such as "Create table like/as".
        - Support for lineages between DWS tables generated by DML operations such as "Insert into".
      - **MRS Hive SQL**
        - Support for lineages between Hive tables generated by DDL operations such as "Create table like/as".
        - Support for lineages between DWS tables generated by DML operations such as "Insert into/overwrite".
  - **Data integration nodes**

    Support for lineage parsing of the **CDM Job**, **ETL Job**, and **OBS Manager** nodes.
    - **CDM Job**

      Support for data lineages generated during table file migration between MRS Hive, DLI, DWS, RDS, OBS, and CSS.
    - **ETL Job**

      Support for data lineages generated by ETL tasks between DLI, OBS, MySQL, and DWS.
    - **OBS Manager**

      Support for data lineages generated by file replication and migration between OBS directories.

Automatic Lineage Analysis

In DGC, when a job in Data Development is scheduled, the system automatically analyzes the lineage relationships and synchronizes the job information to the Data Assets module for display.

NOTE

The prerequisite for automatic lineage analysis is that you have collected metadata from the data source in the job. For details on metadata collection, see **Task Management**.

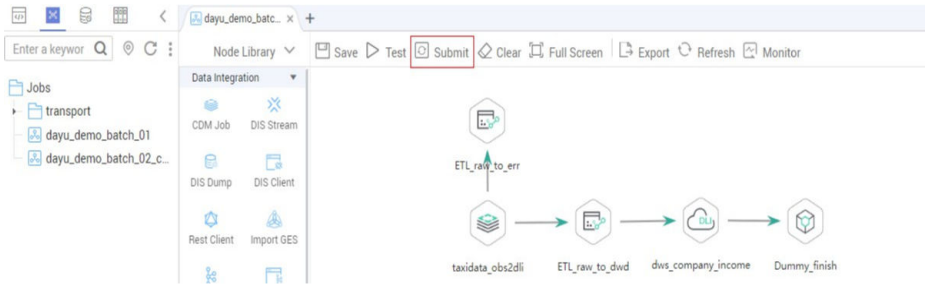
Step 1 Log in to the DGC console, select a DGC instance and a workspace, and click **Data Development** on the workspace to access the **Data Development** console.

Step 2 Choose the **Develop Job** icon from the left navigation bar. On the displayed page, click the required job.

Step 3 During data development, when a job is scheduled, the job information is synchronized to Data Assets.

 Copyright © Huawei Technologies Co., Ltd.

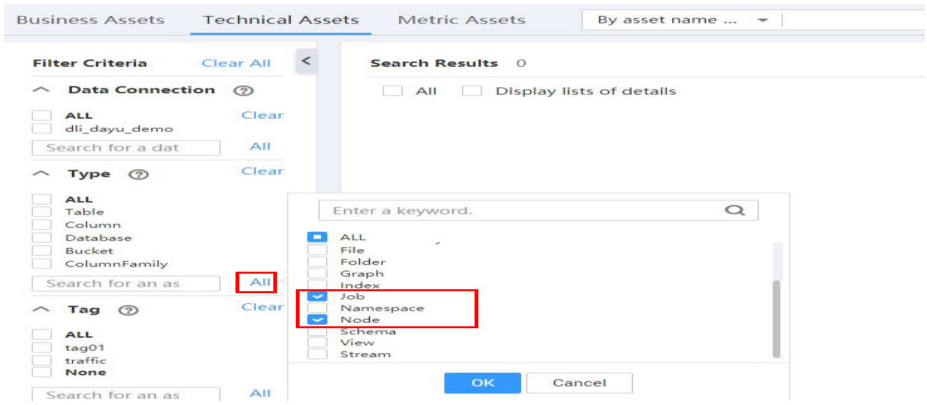
Figure 8-23 Scheduling



Step 4 Select **Data Assets** from the drop-down list in the upper left corner of the console to access the Data Assets console. Choose **Data Catalog** > **Technical Assets** to query data development jobs and nodes. In the **Types** filter area, click **ALL**, select **Job** and **Node**, and click **OK**.

Jobs and nodes in Data Development correspond to the **Job** and **Node** types, respectively. Note that job information in Data Development does not belong to any data connection. Therefore, if you select a data connection in the search criteria, no result is displayed.

Figure 8-24 Job and **Node** types



In the data asset search result, the data asset whose type name ends with **_job** is a job. You can click the job name to view the job details. On the job details page that is displayed, click **Edit** to go to the job editing page of Data Development.

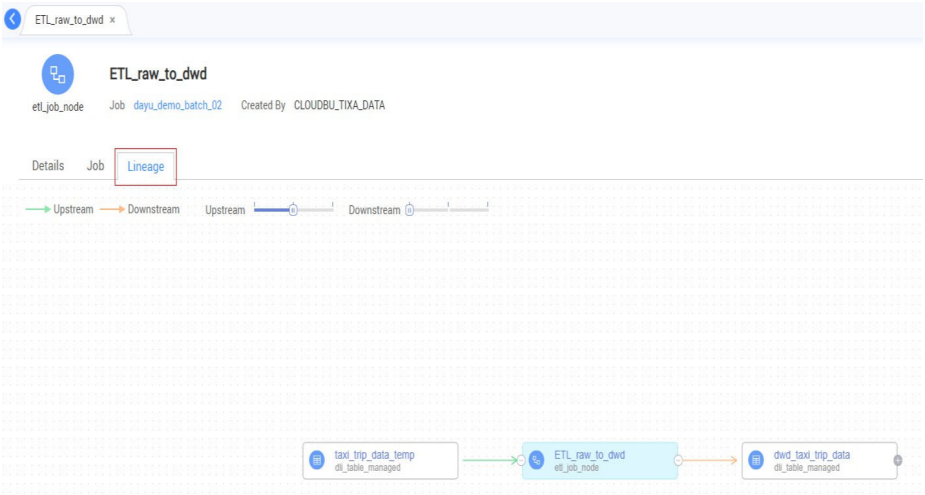
 Copyright © Huawei Technologies Co., Ltd.

Figure 8-25 Viewing a job

In the data asset search result, the data asset whose type name ends with **_node** is a node. You can click the node name to view the node details. On the node details page, you can view the script content and lineage of the node.

- Click the **+** and **-** icons of a node in the lineage diagram to view the upstream and downstream links of the lineage.
- Click a node in the lineage diagram to view the details about the node.
- On the job details page that is displayed, click **Edit** to go to the job editing page of Data Development.

Figure 8-26 Viewing a node lineage



----End

Manually Configuring a Lineage

In DGC Data Development, 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

 Copyright © Huawei Technologies Co., Ltd.

custom type. When manually configuring a lineage, you can add data sources that are not supported as custom types.

The following describes the nodes that support manual lineage configuration. See the description of related nodes for more details.

- **CDM Job**
- **DWS SQL**
- **MRS Hive SQL**
- **DLI SQL**
- **ETL Job**
- **MRS Spark**

8.5.3 CDM Job

Functions

The CDM Job node is used to execute a predefined CDM job for data migration.

Parameters

Table 8-30, **Table 8-31**, and **Table 8-32** describe the parameters of the CDM Job node. Configure the lineage to identify the data flow direction, which can be viewed in the data asset module.

Table 8-30 Parameters of the CDM Job node

Parameter	Mandatory	Description
CDM Cluster Name	Yes	Name of the CDM cluster to which the CDM job to be executed belongs. You can select two CDM clusters to improve job reliability.
Job Type	Yes	• Existing jobs • New jobs **NOTE** 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.
CDM Job Name	No	This parameter is required only when **Job Type** is set to **Existing jobs**. Name of the CDM job to be executed. 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 Data Development module.

 Copyright © Huawei Technologies Co., Ltd.

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 Data Development module.
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).

Table 8-31 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: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Failure Policy	Yes	Operation that will be performed if the node task fails to be executed. Possible values: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • Suspend execution plans of the current and subsequent nodes

Table 8-32 Lineage

Parameter	Description
Input	

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Add	Click **Add**. In the **Type** drop-down list, select the type to be created. The value can be **DWS**, **OBS**, **HIVE**, **CUSTOM**, **CloudTable, DLI**, or **CSS**. • DWS – **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 – **Path**: Click ···. In the displayed dialog box, select an OBS path. • CSS – **Cluster Name**: Click ···. In the displayed dialog box, select a CSS cluster. – **Index**: name of the CSS index. • HIVE – **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 – **Name** (mandatory): Enter a name of the CUSTOM type. – **Attribute** (mandatory): Enter an attribute of the CUSTOM type. You can add more than one attribute. • CloudTable (supported only in North China–Beijing1) – **Connection Name**: Click ···. In the displayed dialog box, select a CloudTable data connection. – **Namespace**: Click ···. In the displayed dialog box, select a CloudTable namespace. – **Table Name**: Click ···. In the displayed dialog box, select a CloudTable table. • DLI – **Connection Name**: Click ···. In the displayed dialog box, select a DLI data connection. – **Database**: Click ···. In the displayed dialog box, select a DLI database.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
	– **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.
Edit	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	

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Add	Click **Add**. In the **Type** drop-down list, select the type to be created. The value can be **DWS**, **OBS**, **HIVE**, **CUSTOM**, **CloudTable, DLI**, or **CSS**. • DWS – **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 – **Path**: Click ···. In the displayed dialog box, select an OBS path. • CSS – **Cluster Name**: Click ···. In the displayed dialog box, select a CSS cluster. – **Index**: Enter a CSS index name. • HIVE – **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 – **Name** (mandatory): Enter a name of the CUSTOM type. – **Attribute** (mandatory): Enter an attribute of the CUSTOM type. You can add more than one attribute. • CloudTable (supported only in North China–Beijing1) – **Connection Name**: Click ···. In the displayed dialog box, select a CloudTable data connection. – **Namespace**: Click ···. In the displayed dialog box, select a CloudTable namespace. – **Table Name**: Click ···. In the displayed dialog box, select a CloudTable table. • DLI – **Connection Name**: Click ···. In the displayed dialog box, select a DLI data connection. – **Database**: Click ···. In the displayed dialog box, select a DLI database.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
	– **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.
Edit	Click to modify the parameter settings. After the modification, save the settings.
Delete	Click to delete the parameter settings.
Table Details	Click to view details about the table created based on the output lineage.

8.5.4 DIS Stream

Functions

The DIS Stream node is used to query the status of a DIS stream. If the DIS stream is normal, you can perform other nodes. If the DIS stream is abnormal, the DIS Stream node will send an error message and exit. If you want to perform other nodes, you must set **Failure policy** to **Proceed to the next node**. For details about how to set **Failure policy**, see **Table 8-34**.

Parameters

Table 8-33 and **Table 8-34** describe the parameters of the DIS Stream node.

Table 8-33 Parameters of DIS Stream nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).

Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Stream Name	Yes	Select or enter the DIS stream to be queried. When entering the stream name, you can reference job parameters and use the EL expression (for details, see **Expression Overview**). To create a DIS stream, you can use either of the following methods: • Click . On the **Stream Management** page of DLF, create a DIS stream. • Go to the DIS console to create a DIS stream.

Table 8-34 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: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Failure Policy	Yes	Operation that will be performed if the node task fails to be executed. Possible values: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • **Suspend execution plans of the current and subsequent nodes**

8.5.5 DIS Dump

Functions

The DIS Dump node is used to configure data dump tasks in DIS.

Parameters

Table 8-35 and **Table 8-36** describe the parameters of the DIS Dump node.

Table 8-35 Parameters of DIS Dump nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
Stream Name	Yes	Select or enter the DIS stream to be queried. When entering the stream name, you can reference job parameters and use the EL expression (for details, see **Expression Overview**). To create a DIS stream, you can use either of the following methods: • Click . On the **Stream Management** page of DLF, create a DIS stream. • Go to the DIS console to create a DIS stream.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Duplicate Name Policy	Yes	Select a duplicate name policy. If the name of a dump task already exists, you can adopt either of the following policies based on site requirements: • Ignore: Give up adding the dump task and exit DIS Dump. The status of DIS Dump is **Succeeded**. • Overwrite: Continue to add the dump task by overwriting the one with the same name.
Dump Destination	Yes	1. Destination to which data is dumped. Possible values: • **CloudTable**: The streaming data is stored to DIS while being imported to the HBase or OpenTSDB table of CloudTable in real time. • **OBS**: After the streaming data is stored to DIS, it is then periodically imported to OBS. After the real-time file data is stored to DIS, it is imported to OBS immediately. **NOTE** CloudTable (supported only in North China–Beijing1) 2. Click . In the dialog box that is displayed, set dump parameters. For details, see **Managing a Dump Task** of *Data Ingestion Service User Guide*.

Table 8-36 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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Retry upon Failure	Yes	Indicates whether to re-execute a node task if its execution fails. Possible values: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • **Suspend execution plans of the current and subsequent nodes**

8.5.6 DIS Client

Functions

The DIS Client node is used to send messages to a DIS stream.

Parameters

Table 8-37 describes the parameters of the DIS Client node.

Table 8-37 Parameters of DIS Client nodes

Parameter	Mandatory	Description
Using DIS Connection	No	If this function is enabled, messages can be sent to the DIS streams of another account. If this function is not enabled, messages can be sent only to the DIS streams in all regions of the current account.

Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
DIS Connection	No	Select a created DIS connection.
Region	No	Region that the target DIS stream belongs to. The DIS Client node is used to send messages to the target DIS stream.
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
Stream Name	Yes	DIS stream to which messages will be sent. You can select a DIS stream by entering the DIS stream address or clicking ***.
Sent Data	Yes	Text sent to the DIS stream. You can directly enter text or click to use the EL expression.
Related Job	No	Click *** to select a related job. You can select a batch job or a real-time job. **NOTE** A maximum of 10 jobs can be selected. After selecting a job, click **Monitor**. On the **Monitory Job** page, select the DIS Client node and click **View Related Job** on the lower part of the page. In the **View Related Job** dialog box, click **View** in the **Operation** column of a job to view the details about the job.

Table 8-38 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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Retry upon Failure	Yes	Indicates whether to re-execute a node task if its execution fails. Possible values: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • **Suspend execution plans of the current and subsequent nodes**

8.5.7 Rest Client

Functions

The Rest Client node is used to respond to RESTful requests in HUAWEI CLOUD. Only the RESTful requests that have been authenticated by IAM Token are supported.

Parameters

Table 8-39, **Table 8-40**, and **Table 8-41** describe the parameters of the Rest Client node.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-39 Parameters of RestAPI nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
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: ● **GET** ● **POST** ● **PUT** ● **DELETE**
Request Header	No	Click + to add a request header. The parameters are described as follows: ● Parameter Name Name of a parameter. The options are **Content-Type**, **Accept-Language**, and **X-Auth-Token**. ● Parameter Value The value can contain a maximum of 64 characters and consists only of letters, digits, hyphens (-), underscores (_), slashes (/), and semicolons (;).
URL Parameter	No	Enters 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: ● Parameter The value can contain a maximum of 32 characters and consists only of letters, digits, hyphens (-), and underscores (_). ● Value The value can contain a maximum of 64 characters and consists only of letters, digits, hyphens (-), underscores (_), dollar signs ($), open braces ({), and close braces (}).

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
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: • **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).
Property Path	Yes	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**. For example, the returned result is as follows: `{ "param1": "aaaa", "inner": { "inner": { "param4": 2014247437 }, "param3": "cccc" }, "status": 200, "param2": "bbbb" }` The **param4** path is **inner.inner.param4**.
Request Success Flag	Yes	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**. The request success flag can contain only letters, digits, hyphens (**-**), underscores (_), dollar signs (**$**), open braces (**{**), and close braces (**}**). Separate values with semicolons (;).

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Request Failure Flag	No	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**. The request success flag can contain only letters, digits, hyphens (**-**), underscores (_), dollar signs (**$**), open braces (**{**), and close braces (**}**). Separate values with semicolons (;).
Retry Interval (seconds)	Yes	If the return value of the response message does not match the request success flag, the node keeps querying the matching status at a specified interval until the return value of the response message is the same as the request success flag. By default, the timeout interval of the node is one hour. If the return value of the response message does not match the request success flag within this period, the node status changes to **Failed**. This parameter is available only when **Check Returned Value** is set to **YES**.
The response message body parses the transfer parameter.	No	Specify the mapping between the job variable and JSON property path. Separate parameters by newline characters. For example: var4=inner.inner.param4 **var4** is a job variable. The job variable must consist of 1 to 16 characters and contain only letters and digits. **inner.inner.param4** is the JSON property path. This parameter takes effect only when it is referenced by the subsequent node. When this parameter is referenced, the format is **${var4}**

 Copyright © Huawei Technologies Co., Ltd.

Table 8-40 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: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • **Suspend execution plans of the current and subsequent nodes**

Table 8-41 Lineage

Parameter	Description
Input	

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Create	Click **Add**. In the **Type** drop-down list, select the type to be created. The value can be **DWS**, **OBS**, **HIVE**, **CUSTOM**, **CloudTable, DLI**, or **CSS**. • DWS – **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 – **Path**: Click ···. In the displayed dialog box, select an OBS path. • CSS – **Cluster Name**: Click ···. In the displayed dialog box, select a CSS cluster. – **Index**: Enter a CSS index name. • HIVE – **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 – **Name**: Enter a name of the CUSTOM type. – **Attribute**: Enter an attribute of the CUSTOM type. You can add more than one attribute. • CloudTable (supported only in North China–Beijing1) – **Connection Name**: Click ···. In the displayed dialog box, select a CloudTable data connection. – **Namespace**: Click ···. In the displayed dialog box, select a CloudTable namespace. – **Table Name**: Click ···. In the displayed dialog box, select a CloudTable table. • DLI – **Connection Name**: Click ···. In the displayed dialog box, select a DLI data connection. – **Database**: Click ···. In the displayed dialog box, select a DLI database.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
	– **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.
Delete	Click to delete the parameter settings.
View Details	Click to view details about the table created based on the input lineage.
Output	

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Create	Click **Add**. In the **Type** drop-down list, select the type to be created. The value can be **DWS**, **OBS**, **HIVE**, **CUSTOM**, **CloudTable, DLI**, or **CSS**. • DWS – **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 – **Path**: Click ···. In the displayed dialog box, select an OBS path. • CSS – **Cluster Name**: Click ···. In the displayed dialog box, select a CSS cluster. – **Index**: Enter a CSS index name. • HIVE – **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 – **Name**: Enter a name of the CUSTOM type. – **Attribute**: Enter an attribute of the CUSTOM type. You can add more than one attribute. • CloudTable (supported only in North China–Beijing1) – **Connection Name**: Click ···. In the displayed dialog box, select a CloudTable data connection. – **Namespace**: Click ···. In the displayed dialog box, select a CloudTable namespace. – **Table Name**: Click ···. In the displayed dialog box, select a CloudTable table. • DLI – **Connection Name**: Click ···. In the displayed dialog box, select a DLI data connection. – **Database**: Click ···. In the displayed dialog box, select a DLI database.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
	– **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.
Delete	Click to delete the parameter settings.
View Details	Click to view details about the table created based on the output lineage.

8.5.8 Import GES

Function

The Import GES node is used to import files from an OBS bucket to a GES graph.

Parameters

Table 8-42 and Table 8-43 describe the parameters of the Import GES node.

Table 8-42 Parameters of Import GES nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
Graph Name	Yes	Select the graph to be imported. To create a GES graph, go to the GES console.
Metadata	Yes	Select the corresponding metadata.
Edge Dataset	Yes	Select the corresponding edge dataset.
Vertex Dataset	No	Select the corresponding vertex dataset. If it is not selected, the vertices in the edge dataset are used as the source of the vertex dataset.

Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Edge Processing	Yes	The edge processing supports the following modes: • Allow repetitive edges • Ignore subsequent repetitive edges • Overwrite previous repetitive edges
Offline	No	Indicates whether offline import is selected. The value is **true** or **false**, and the default value is **false**. • **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. Compared with offline import, online import is slower. However, the graph can be read (cannot be written) during the import.
Ignore Labels on Repetitive Edges	No	Indicates whether to ignore labels on repetitive edges. The value is **true** or **false**, and the default value is **true**. • **true**: 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. • **false**: 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 8-43 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.

 Copyright © Huawei Technologies Co., Ltd.

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: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • Suspend execution plans of the current and subsequent nodes

8.5.9 MRS Kafka

Functions

The MRS Kafka node is used to query the number of messages that are not consumed by a topic.

Parameters

Table 8-44 and **Table 8-45** describe the parameters of the MRS Kafka node.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-44 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 the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).

Table 8-45 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: ● **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Failure Policy	Yes	Operation that will be performed if the node task fails to be executed. Possible values: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • **Suspend execution plans of the current and subsequent nodes**

8.5.10 Kafka Client

Functions

The Kafka Client node is used to send data to Kafka topics.

Parameters

Table 8-46 describes the parameters of the Kafka Client node.

Table 8-46 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 the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
Text	Yes	Text content sent to Kafka. You can directly enter text or click to use the EL expression.

Copyright © Huawei Technologies Co., Ltd.

Table 8-47 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: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • **Suspend execution plans of the current and subsequent nodes**

8.5.11 DLI Flink Job

Functions

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.

 Copyright © Huawei Technologies Co., Ltd.

Property 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 8-48**.
  - **Flink SQL job**: For details, see **Table 8-49**.
  - **User-defined Flink job**: For details, see **Table 8-50**.
- **Table 8-51**

Table 8-48 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 the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).

Table 8-49 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 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	Compute Unit (CU) is the pricing unit for DLI. 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 x (Number of CUs – 1).

Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
UDF Jar	No	This parameter is valid only when you select a dedicated queue for **Queue**. Before selecting a UDF JAR resource package, you need to upload the JAR package and its dependency packages to the OBS bucket and create resources on the resource management page. For details, see **Creating a Resource**. In SQL, you can call a user-defined function that is inserted into a JAR package.
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, digits, and underscores (_). The default value is the same as the node name.
Job name must be prefixed with workspace name	No	Indicates whether to add a workspace prefix to the created job.
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).

Table 8-50 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, you need to upload the JAR package and its dependency packages to the OBS bucket and create resources on the resource management page. For details, see **Creating a Resource**.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Main Class	Yes	Name of the JAR package to be loaded, for example, **KafkaMessageStreaming**. • **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). **NOTE** When a class belongs to a package, the package path must be carried, for example, **packagePath.KafkaMessageStreaming**.
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. **NOTE** During job creation, a sub-user can only select a queue that has been allocated to the user.
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 x (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, digits, 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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).

Table 8-51 Advanced parameters

Parameter	Mandatory	Description
Max. Node Execution Duration	Yes	Execution timeout interval for 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: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: • **End the current job execution plan** • **Go to the next node** • **Suspend the current job execution plan**

8.5.12 DLI SQL

Functions

The DLI SQL node is used to transfer SQL statements to DLI for data source analysis and exploration.

 Copyright © Huawei Technologies Co., Ltd.

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 8-52, **Table 8-53**, and **Table 8-54** describe the parameters of the DLI SQL node.

Table 8-52 Parameters of DLI SQL nodes

Parameter	Mandatory	Description
SQL Statement or Script	Yes	You can select SQL statements or SQL scripts. • 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**. **NOTE** If you select the SQL statement mode, the Data Development module cannot parse the parameters contained in the SQL statement.
Database Name	Yes	Database that is configured in the SQL script. The value can be changed.
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: • Click . On the **Queue Management** page of DLF, 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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Node Name	Yes	Name of the SQL script. The value can be changed. The rules are as follows: Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
Record Dirty Data	Yes	Click to specify whether to record dirty data. • If you select , dirty data will be recorded. • If you do not select , dirty data will not be recorded.

Table 8-53 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	Indicates whether to re-execute a node task if its execution fails. Possible values: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Failure Policy	Yes	Operation that will be performed if the node task fails to be executed. Possible values: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • Suspend execution plans of the current and subsequent nodes

Table 8-54 Lineage

Parameter	Description
Input	

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Add	Click **Add**. In the **Type** drop-down list, select the type to be created. The value can be CloudTableDLI, DWS, OBS, or CSS. • DWS – **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 – **Path**: Click ···. In the displayed dialog box, select an OBS path. • CSS – **Cluster Name**: Click ···. In the displayed dialog box, select a CSS cluster. – **Index**: name of the CSS index. • CloudTable (supported only in North China–Beijing1) – **Connection Name**: Click ···. In the displayed dialog box, select a CloudTable data connection. – **Namespace**: Click ···. In the displayed dialog box, select a CloudTable namespace. – **Table Name**: Click ···. In the displayed dialog box, select a CloudTable table. • DLI – **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.
Edit	Click to modify the parameter settings. After the modification, save the settings.
Delete	Click to delete the parameter settings.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
View Details	Click to view details about the table created based on the input lineage.
Output	
Add	Click **Add**. In the **Type** drop-down list, select the type to be created. The value can be CloudTableDLI, DWS, OBS, or CSS. • DWS – **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 – **Path**: Click ···. In the displayed dialog box, select an OBS path. • CSS – **Cluster Name**: Click ···. In the displayed dialog box, select a CSS cluster. – **Index**: Enter a CSS index name. • CloudTable (supported only in North China–Beijing1) – **Connection Name**: Click ···. In the displayed dialog box, select a CloudTable data connection. – **Namespace**: Click ···. In the displayed dialog box, select a CloudTable namespace. – **Table Name**: Click ···. In the displayed dialog box, select a CloudTable table. • DLI – **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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Edit	Click to modify the parameter settings. After the modification, save the settings.
Delete	Click to delete the parameter settings.
Table Details	Click to view details about the table created based on the output lineage.

8.5.13 DLI Spark

Functions

The DLI Spark node is used to execute a predefined Spark job.

Parameters

Table 8-55 and Table 8-56 describe the parameters of the DLI Spark node.

Table 8-55 Parameters of DLI Spark nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
DLI Queue	Yes	Select a queue from the drop-down list box.
Job Name	Yes	Name of the DLI Spark job. It must consist of 1 to 64 characters and contain only letters, digits, and underscores (_). The default value is the same as the node name.
Job Running Resources	No	Select the running resource specifications of the job. • 8-core, 32 GB memory • 16-core, 64 GB memory • 32-core, 128 GB memory
Major Job Class	Yes	Main class of the DLI Spark job, that is, the main class of the JAR package.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Spark program resource package	Yes	JAR package of the user-defined Spark application. Before selecting a resource package, you need to upload the JAR package and its dependency packages to the OBS bucket and create resources on the resource management page. For details, see **Creating a Resource**.
Major-Class Entry Parameters	No	Enter the entry parameters of the program. Press **Enter** to separate the parameters.
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**.
Module Name	No	Dependency modules provided by DLI for executing datasource connection jobs. To access different services, you need to select different modules. • CloudTable/MRS HBase: sys.datasource.hbase • CloudTable/MRS OpenTSDB: sys.datasource.opentsdb • RDS MySQL: sys.datasource.rds • RDS PostGre: sys.datasource.rds • DWS: sys.datasource.dws • CSS: sys.datasource.css

Table 8-56 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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Retry upon Failure	Yes	Indicates whether to re-execute a node task if its execution fails. Possible values: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • Suspend execution plans of the current and subsequent nodes

8.5.14 DWS SQL

Functions

The DWS SQL node is used to transfer an SQL statement to DWS.

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 8-57, **Table 8-58**, and **Table 8-59** describe the parameters of the DWS SQL node.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-57 Parameters of DWS SQL nodes

Parameter	Mandatory	Description
SQL Statement or Script	Yes	You can select **SQL statement** or **SQL script**. • 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**. **NOTE** If you select the SQL statement mode, the Data Development module cannot parse the parameters contained in the SQL statement.
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.
Dirty Data Table	No	Enter the name of the dirty data table defined in the SQL script.
Node Name	Yes	Name of the SQL script. The value can be changed. The rules are as follows: Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).

Table 8-58 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.

 Copyright © Huawei Technologies Co., Ltd.

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: ● **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: ● **End the current job execution plan** ● **Go to the next job** ● **Suspend the current job execution plan** ● Suspend execution plans of the current and subsequent nodes

Table 8-59 Lineage

Parameter	Description
Input	

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Create	Click **Add**. In the **Type** drop-down list, select the type to be created. The value can be **DWS**, **OBS**, **HIVE**, **CUSTOM**, **CloudTable, DLI**, or **CSS**. • DWS – **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 – **Path**: Click ···. In the displayed dialog box, select an OBS path. • CSS – **Cluster Name**: Click ···. In the displayed dialog box, select a CSS cluster. – **Index**: Enter a CSS index name. • HIVE – **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 – **Name**: Enter a name of the CUSTOM type. – **Attribute**: Enter an attribute of the CUSTOM type. You can add more than one attribute. • CloudTable (supported only in North China–Beijing1) – **Connection Name**: Click ···. In the displayed dialog box, select a CloudTable data connection. – **Namespace**: Click ···. In the displayed dialog box, select a CloudTable namespace. – **Table Name**: Click ···. In the displayed dialog box, select a CloudTable table. • DLI – **Connection Name**: Click ···. In the displayed dialog box, select a DLI data connection. – **Database**: Click ···. In the displayed dialog box, select a DLI database.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
	– **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.
Delete	Click to delete the parameter settings.
View Details	Click to view details about the table created based on the input lineage.
Output	

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Create	Click **Add**. In the **Type** drop-down list, select the type to be created. The value can be **DWS**, **OBS**, **HIVE**, **CUSTOM**, **CloudTable, DLI**, or **CSS**. • DWS – **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 – **Path**: Click ···. In the displayed dialog box, select an OBS path. • CSS – **Cluster Name**: Click ···. In the displayed dialog box, select a CSS cluster. – **Index**: Enter a CSS index name. • HIVE – **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 – **Name**: Enter a name of the CUSTOM type. – **Attribute**: Enter an attribute of the CUSTOM type. You can add more than one attribute. • CloudTable (supported only in North China–Beijing1) – **Connection Name**: Click ···. In the displayed dialog box, select a CloudTable data connection. – **Namespace**: Click ···. In the displayed dialog box, select a CloudTable namespace. – **Table Name**: Click ···. In the displayed dialog box, select a CloudTable table. • DLI – **Connection Name**: Click ···. In the displayed dialog box, select a DLI data connection. – **Database**: Click ···. In the displayed dialog box, select a DLI database.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
	– **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.
Delete	Click to delete the parameter settings.
View Details	Click to view details about the table created based on the output lineage.

8.5.15 MRS SparkSQL

Functions

The MRS Spark SQL node is used to execute a predefined SparkSQL statement on MRS.

Parameters

Table 8-60 and **Table 8-61** describe the parameters of the MRS Spark SQL node.

Table 8-60 Parameters of MRS Spark SQL nodes

Parameter	Mandatory	Description
SQL Script	Yes	Path to a script to be executed. If the script is not created, create and develop the 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.

Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Script Parameter	No	If the associated SQL script uses a parameter, the parameter name is displayed. Set the parameter value in the text box next to the parameter name. The parameter value can be a built-in function and EL expression. For details about built-in functions and EL expressions, see **Expression Overview**. If the parameters of the associated SQL script are changed, click to refresh the parameters.
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 *MapReduce User Guide*.
Node Name	Yes	Name of the SQL script. The value can be changed. The rules are as follows: Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).

Table 8-61 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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Retry upon Failure	Yes	Indicates whether to re-execute a node task if its execution fails. Possible values: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • Suspend execution plans of the current and subsequent nodes

8.5.16 MRS Hive SQL

Functions

The MRS Hive SQL node is used to execute a predefined Hive SQL script on DLF.

Parameters

Table 8-62 and **Table 8-63** describe the parameters of the MRS Hive SQL node.

Table 8-62 Parameters of MRS Hive SQL nodes

Parameter	Mandatory	Description
SQL Script	Yes	Path to a script to be executed. If the script is not created, create and develop the script by referring to **Creating a Script** and **Developing an SQL Script**.

Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Data Connection	Yes	Data connection that is configured in the SQL script. The value can be changed.
Database	Yes	Database that is configured in the SQL script. The value can be changed.
Script Parameter	No	If the associated SQL script uses a parameter, the parameter name is displayed. Set the parameter value in the text box next to the parameter name. The parameter value can be a built-in function and EL expression. For details about built-in functions and EL expressions, see **Expression Overview**. If the parameters of the associated SQL script are changed, click to refresh the parameters.
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 MRSHiveSQL jobs, see **Running a HiveSql Job** > **Table 2 Program Parameter parameters** in the *MapReduce User Guide*.
Node Name	Yes	Name of the SQL script. The value can be changed. The rules are as follows: Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).

Table 8-63 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.

 Copyright © Huawei Technologies Co., Ltd.

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: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • Suspend execution plans of the current and subsequent nodes

8.5.17 MRS Spark

Functions

The MRS Spark node is used to execute a predefined Spark job on MRS.

Parameters

Table 8-64 and **Table 8-65** describe the parameters of the MRS Spark node.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-64 Parameters of MRS Spark nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, 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: • 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 the MRS job. Must consist of 1 to 64 characters and contain only letters, digits, and underscores (_).
JAR Package	Yes	Select a JAR package. Before selecting a JAR package, you need to upload the JAR package to the OBS bucket, create a resource on the **Resource Management** 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 *MapReduce User Guide*.
Input Data Path	No	Path where the input data resides.
Output Data Path	No	Path where the output data resides.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-65 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: ● **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: ● **End the current job execution plan** ● **Go to the next job** ● **Suspend the current job execution plan** ● Suspend execution plans of the current and subsequent nodes

8.5.18 MRS Spark Python

Functions

The MRS Spark Python node is used to execute a predefined Spark Python job on MRS.

 Copyright © Huawei Technologies Co., Ltd.

Parameters

Table 8-66 and Table 8-67 describe the parameters of the MRS Spark Python node.

Table 8-66 Parameters of MRS Spark Python nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, 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: • 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 an MRS cluster, see **Custom Purchase of a Cluster** in *MapReduce User Guide*.
Job Name	Yes	Name of the MRS Spark Python job. It must consist of 1 to 64 characters and contain only letters, digits, and underscores (_).
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 8-67 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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Retry upon Failure	Yes	Indicates whether to re-execute a node task if its execution fails. Possible values: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • **Suspend execution plans of the current and subsequent nodes**

8.5.19 MRS Flink Job

Functions

The MRS Flink node is used to execute predefined Flink jobs in MRS.

Parameters

Configure the parameters of the MRS Flink node by referring to **Table 8-68** and **Table 8-69**.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-68 Parameters of the MRS Flink node

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, 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: • Click . On the **Clusters** page, create an MRS cluster. • Go to the MRS console to create an MRS cluster.
Job Name	Yes	Name of the MRS Flink job. Must consist of 1 to 64 characters and contain only letters, digits, and underscores (_).
Job Resource Package	Yes	Select a JAR package. Before selecting a JAR package, you need to upload the JAR package to the OBS bucket, create a resource on the **Resource Management** 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 Spark Job** in the *MapReduce User Guide*.
Input Data Path	No	Path where the input data resides.
Output Data Path	No	Path where the output data resides.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-69 Advanced parameters

Parameter	Mandatory	Description
Max. Node Execution Duration	Yes	Maximum duration of executing a node. 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.
Retry upon Failure	Yes	Whether to re-execute a node after the node fails to be executed. • **Yes**: The node will be re-executed after it fails to be executed. The following parameters must be configured: – Maximum Retries – Retry Interval (seconds) • **No**: The node 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	Policies to be performed after the node fails to be executed: • **End the current job execution plan** • **Go to the next node** • **Suspend the current job execution plan** • **End execution plans of the current and subsequent nodes**

8.5.20 MRS MapReduce

Functions

The MRS MapReduce node is used to execute a predefined MapReduce program on MRS.

Parameters

Table 8-70 and **Table 8-71** describe the parameters of the MRS MapReduce node.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-70 Parameters of MRS MapReduce nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, 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: • 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 the MRS job. Must consist of 1 to 64 characters and contain only letters, digits, and underscores (_).
JAR Package	Yes	Select a JAR package. Before selecting a JAR package, you need to upload the JAR package to the OBS bucket, create a resource on the **Resource Management** 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 8-71 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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Retry upon Failure	Yes	Indicates whether to re-execute a node task if its execution fails. Possible values: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • Suspend execution plans of the current and subsequent nodes

8.5.21 CSS

Functions

The CSS node is used to process CSS requests and enable online distributed searching.

Parameters

Table 8-72 and **Table 8-73** describe the parameters of the CSS node.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-72 Parameters of CSS nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, 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	A CDM cluster. The Agent provided by the CDM cluster is the operating environment of CSS. If there are no CDM clusters available in the drop-down list, create one on the CDM console.
Request Type	Yes	Possible values: • **GET** • **POST** • **PUT** • **HEAD** • **DELETE**
Request Parameter	No	Parameter of the request. For example, to query the dlfdata mapping type in the dlf_search index, set this parameter to: **/dlf_search/dlfdata/_search**
Request Body	No	The request body is in JSON format.
CloudSearch Output Path	No	Path where output data is to be stored.

Table 8-73 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.

 Copyright © Huawei Technologies Co., Ltd.

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: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • Suspend execution plans of the current and subsequent nodes

8.5.22 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>")}
```

 Copyright © Huawei Technologies Co., Ltd.

Parameters

Table 8-74 and Table 8-75 describe the parameters of the Shell node.

Table 8-74 Parameters of Shell nodes

Parameter	Mandatory	Description
SQL Statement or Script	Yes	You can select SQL statements or SQL scripts. • 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**. **NOTE** If you select the SQL statement mode, the Data Development module cannot parse the parameters contained in the SQL statement.
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. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).

 Copyright © Huawei Technologies Co., Ltd.

Table 8-75 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: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • Suspend execution plans of the current and subsequent nodes

8.5.23 RDS SQL

Functions

The RDS SQL node is used to transfer SQL statements to RDS.

Parameters

Table 8-76 and **Table 8-77** describe the parameters of the RDS SQL node.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-76 Parameters of RDS SQL nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, 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	You can select **SQL statement** or **SQL script**. • SQL statement In the **Statements** 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**. **NOTE** If you select the SQL statement mode, the Data Development module cannot parse the parameters contained in the SQL statement.

Table 8-77 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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Retry upon Failure	Yes	Indicates whether to re-execute a node task if its execution fails. Possible values: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • Suspend execution plans of the current and subsequent nodes

8.5.24 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 8-78 and **Table 8-79** describe the parameters of the ETL Job node.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-78 Parameters of Transform Load nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
ETL Configuration	Yes	Click to edit the source and destination data to be transformed. The supported source data types are DLI, OBS and MySQL. • When the source data type is DLI, the supported destination data types are CloudTable (supported only in CN North-Beijing1), DWS, GES, CSS, DLI and OBS. • 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. **NOTICE** • Data transformation from DLI to DWS: Before importing data from Data Development 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 and CloudTable: Before importing data from DLI to CSS or CloudTable, ensure that a cross-source connection associated with CSS or CloudTable has been created on DLI. For details about how to create a cross-source connection on DLI, see **SQL Datasource Connection** of *Data Lake Insight User Guide*. Before importing data from DLI to CloudTable, ensure that a CloudTable table and a cross-source connection associated with CloudTable have been created on DLI.
Configure SQL Template	No	Click **Obtain Template** to obtain an SQL template.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-79 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: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • **Suspend execution plans of the current and subsequent nodes**

Table 8-80 Lineage

Parameter	Description
Input	

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Create	Click **Add**. In the **Type** drop-down list, select the type to be created. The value can be **DWS**, **OBS**, **HIVE**, **CUSTOM**, **CloudTable, DLI**, or **CSS**. • DWS – **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 – **Path**: Click ···. In the displayed dialog box, select an OBS path. • CSS – **Cluster Name**: Click ···. In the displayed dialog box, select a CSS cluster. – **Index**: Enter a CSS index name. • HIVE – **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 – **Name**: Enter a name of the CUSTOM type. – **Attribute**: Enter an attribute of the CUSTOM type. You can add more than one attribute. • CloudTable (supported only in North China–Beijing1) – **Connection Name**: Click ···. In the displayed dialog box, select a CloudTable data connection. – **Namespace**: Click ···. In the displayed dialog box, select a CloudTable namespace. – **Table Name**: Click ···. In the displayed dialog box, select a CloudTable table. • DLI – **Connection Name**: Click ···. In the displayed dialog box, select a DLI data connection. – **Database**: Click ···. In the displayed dialog box, select a DLI database.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
	– **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.
Delete	Click to delete the parameter settings.
View Details	Click to view details about the table created based on the input lineage.
Output	

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Create	Click **Add**. In the **Type** drop-down list, select the type to be created. The value can be **DWS**, **OBS**, **HIVE**, **CUSTOM**, **CloudTable, DLI**, or **CSS**. • DWS – **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 – **Path**: Click ···. In the displayed dialog box, select an OBS path. • CSS – **Cluster Name**: Click ···. In the displayed dialog box, select a CSS cluster. – **Index**: Enter a CSS index name. • HIVE – **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 – **Name**: Enter a name of the CUSTOM type. – **Attribute**: Enter an attribute of the CUSTOM type. You can add more than one attribute. • CloudTable (supported only in North China–Beijing1) – **Connection Name**: Click ···. In the displayed dialog box, select a CloudTable data connection. – **Namespace**: Click ···. In the displayed dialog box, select a CloudTable namespace. – **Table Name**: Click ···. In the displayed dialog box, select a CloudTable table. • DLI – **Connection Name**: Click ···. In the displayed dialog box, select a DLI data connection. – **Database**: Click ···. In the displayed dialog box, select a DLI database.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
	– **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.
Delete	Click to delete the parameter settings.
View Details	Click to view details about the table created based on the output lineage.

8.5.25 CS Job

Functions

The CS Job node is used to execute a predefined Cloud Stream Service (CS) job for real-time analysis of streaming data.

Context

This node enables you to start a CS job or query whether a CS job is running. If you do not select an existing real-time job, Data Development creates and starts the job based on the job status configured on the node. You can customize jobs and use data development job parameters.

Parameters

Table 8-81 and Table 8-82 describe the parameters of the CS Job node.

Table 8-81 Parameters of CS Job nodes

Parameter	Mandatory	Description
Job Type	Yes	Select a job type for CS. • Existing CS job • Flink SQL job • User-defined Flink job • User-defined Spark job
Existing CS job		

Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Streaming Job Name	Yes	Name of the CS job to be executed. To create a CS job, you can use either of the following methods: • Click . On the **Data Integration** page of DLF, create a CS job. • Go to the CS console to create a CS job.
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
Flink SQL job		
SQL Script	Yes	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**.
Script Parameter	Yes	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**.
CloudStream Cluster	Yes	Name of the CS cluster. To create a CS cluster, go to the CS console.
SPUs	Yes	1 SPU = 1 core and 4 GB memory
Parallelism	Yes	Number of tasks that run CS jobs at the same time. You are advised to set this parameter to 1 or 2 times of the SPU.
UDF JAR	No	After the JAR package is imported, you can use SQL statements to call the custom functions in the package. You need to upload the JAR package to the OBS bucket.
Auto Restart upon Exception	No	If you enable this function, the system automatically restarts and restores abnormal CS jobs upon job exceptions.
Streaming Job Name	Yes	Name of the Flink SQL job. The name is 1 to 57 characters long and consists only of letters, digits, hyphens (-), and underlines (_).

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
User-defined Flink job		
JAR Package Path	Yes	The JAR package path can be selected only after you have uploaded the custom JAR package to the OBS bucket.
Main Class	No	Name of the main class in the JAR file to be uploaded, for example, **KafkaMessageStreaming**. If this parameter is not specified, the main class name is determined based on the **Manifest** file in the JAR file.
Main Class Parameter	No	List of parameters for the main class. Parameters are separated by spaces, for example, test tmp/result.txt.
CloudStream Cluster	Yes	Name of the CS cluster. To create a CS cluster, go to the CS console.
SPUs	Yes	1 SPU = 1 core and 4 GB memory
Driver SPU	Yes	Number of SPUs used for each driver node.
Parallelism	Yes	Number of tasks that run CS jobs at the same time. You are advised to set this parameter to 1 or 2 times of the SPU.
Auto Restart upon Exception	No	If you enable this function, the system automatically restarts and restores abnormal CS jobs upon job exceptions.
Streaming Job Name	Yes	Name of the user-defined Flink job. The name is 1 to 57 characters long and consists only of letters, digits, hyphens (-), and underlines (_).
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
User-defined Spark job		

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
JAR Package Path	Yes	The JAR package path can be selected only after you have uploaded the custom JAR package to the OBS bucket.
Main Class	No	Name of the main class in the JAR file to be uploaded, for example, **KafkaMessageStreaming**. If this parameter is not specified, the main class name is determined based on the **Manifest** file in the JAR file.
Main Class Parameter	No	List of parameters for the main class. Parameters are separated by spaces, for example, test tmp/result.txt.
CloudStream Cluster	Yes	Name of the CS cluster. To create a CS cluster, go to the CS console.
SPUs	Yes	1 SPU = 1 core and 4 GB memory
Driver SPU	Yes	Number of SPUs used for each driver node.
Executors	Yes	Number of the executor nodes.
Executor SPUs	Yes	Number of SPUs used for each executor node.
Auto Restart upon Exception	No	If you enable this function, the system automatically restarts and restores abnormal CS jobs upon job exceptions.
Streaming Job Name	Yes	Name of the user-defined Spark job. The name is 1 to 57 characters long and consists only of letters, digits, hyphens (-), and underlines (_).
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).

Table 8-82 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.

 Copyright © Huawei Technologies Co., Ltd.

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: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • Suspend execution plans of the current and subsequent nodes

8.5.26 OCR

Functions

The OCR node is used to recognize text on images and store the recognition result in a specified OBS directory.

NOTE

The OCR node can be used only in CN North-Beijing4.

Parameters

Table 8-83 and **Table 8-84** describe the parameters of the OCR node.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-83 Parameters of OCR nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
Image Path	Yes	OBS path of the image to be recognized.
Category	Yes	Image type that can be recognized. Currently, only General Text OCR is supported. It can be used in various scenarios, such as scanned files, electronic documents, books, documents, and forms.
Storage Directory	Yes	OBS directory for storing the recognition result.

Table 8-84 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: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Failure Policy	Yes	Operation that will be performed if the node task fails to be executed. Possible values: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • **Suspend execution plans of the current and subsequent nodes**

8.5.27 Create OBS

Functions

The Create OBS node is used to create buckets and directories in OBS.

Parameters

Table 8-85 and **Table 8-86** describe the parameters of the Create OBS node.

Table 8-85 Parameters of Create OBS nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
OBS Path	Yes	Path to the OBS bucket or directory. • To create a bucket, enter **//***OBS bucket name*. 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 /*Directory name* following the path. The directory name must be unique.

Copyright © Huawei Technologies Co., Ltd.

Table 8-86 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: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • **Suspend execution plans of the current and subsequent nodes**

8.5.28 Delete OBS

Functions

The Delete OBS node is used to delete a bucket or directory on OBS.

Parameters

Table 8-87 and **Table 8-88** describe the parameters of the Delete OBS node.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-87 Parameters of Delete OBS nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
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 8-88 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: ● **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Failure Policy	Yes	Operation that will be performed if the node task fails to be executed. Possible values: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • **Suspend execution plans of the current and subsequent nodes**

8.5.29 OBS Manager

Function

The OBS Manager node is used to move or copy files from an OBS bucket to a specified directory.

Parameters

Table 8-89 and **Table 8-90** describe the parameters of the OBS Manager node.

Table 8-89 Parameters of OBS Manager nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
Operation Type	Yes	Operations that can be performed on the node. • Move File • Copy File
Source File	Yes	OBS file to be moved or copied from the OBS bucket.
Target Directory	Yes	Directory for storing OBS files to be moved or copied from the OBS bucket.

Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
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 or copied by default. For example, when you enter *.csv, files in this format will be moved or copied.

Table 8-90 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: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • **Suspend execution plans of the current and subsequent nodes**

 Copyright © Huawei Technologies Co., Ltd.

8.5.30 Open/Close Resource

Functions

The Open/Close Resource node is used to enable or disable the HUAWEI CLOUD service.

Parameters

Table 8-91 and Table 8-92 describe the parameters of the Open/Close Resource node.

Table 8-91 Parameters of Open/Close Resource nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).
Service	Yes	Service to be opened or closed. • **ECS** • **CDM** • CS
Open/Close Resource	Yes	Possible values: • On • Off
Instance	Yes	Object to be opened or closed, for example, to open a CDM cluster.

Table 8-92 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.

Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Retry upon Failure	Yes	Indicates whether to re-execute a node task if its execution fails. Possible values: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • Suspend execution plans of the current and subsequent nodes

8.5.31 Subjob

Function

The Subjob node is used to call the batch job that does not contain the subjob node.

Parameter

Table 8-93 and **Table 8-94** describe the parameters of the Subjob node.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-93 Parameters of subjob nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, 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	• 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 8-94 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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Retry upon Failure	Yes	Indicates whether to re-execute a node task if its execution fails. Possible values: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • Suspend execution plans of the current and subsequent nodes

8.5.32 SMN

Functions

The SMN node is used to send notifications to users.

Parameters

Table 8-95 and **Table 8-96** describe the parameters of the SMN node.

Table 8-95 Parameters of SMN nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).

Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
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. ● **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. – 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. – 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**.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description

Message Content	Yes	Message content to be provided. The requirements for entering different types of messages are as follows:
		• **Text**: The size cannot exceed 10 KB.
		• **JSON**: The JSON message must contain the Default protocol and the size cannot exceed 10 KB.
		Example:
		{
		"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."
		}
		• **Template**: The size cannot exceed 10 KB.
		Example:
		"message_template_name":"confirm_message",
		"tags":{
		"topic_urn":"urn:smn:regionId:xxxx:SMN_01"
		}
		In the preceding information, **message_template_name** indicates the template name, and **tags** indicates all tags in the template.
		For details about how to configure SMN, see section **Introduction** of *Simple Message Notification User Guide*.

Table 8-96 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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Retry upon Failure	Yes	Indicates whether to re-execute a node task if its execution fails. Possible values: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • **Suspend execution plans of the current and subsequent nodes**

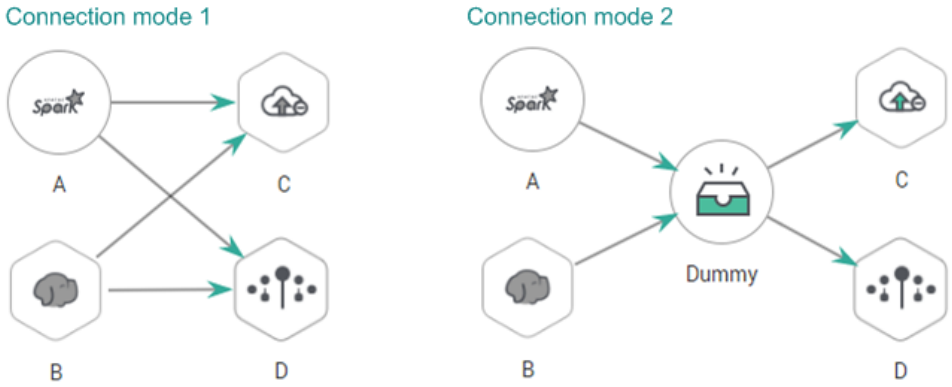
8.5.33 Dummy

Functions

The Dummy node is empty and does not perform any operations. It is used to simplify the complex connection relationships of nodes. **Figure 8-27** shows an example.

Copyright © Huawei Technologies Co., Ltd.

Figure 8-27 Connection modes



Parameters

Table 8-97 describes the parameter of the Dummy node.

Table 8-97 Parameters of Dummy nodes

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, underscores (_), hyphens (-), slashes (/), less-than signs (<), and greater-than signs (>).

8.5.34 For Each

Functions

Specifies a subjob to be executed cyclically and assigns values to variables in a subjob with a dataset.

Parameters

Table 8-98 describes the parameters of the For Each node.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-98 Parameters of the For Each node

Parameter	Mandatory	Description
Node Name	Yes	Name of the node. Must consist of 1 to 128 characters and contain only letters, digits, 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 Loop operator needs to define a dataset, which is used to cyclically replace the variables in subjobs. The dataset comes from: • Upstream nodes For example, the SELECT statement of DLI SQL or the echo of the shell node. Currently, only the SELECT statements of Hive SQL, Spark SQL, and DLI SQL are supported. • Files in CSV format
Concurrent Subjobs	Yes	Subjobs generated cyclically can be executed concurrently. You can set the number of concurrent subjobs.
Subjob 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	• 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.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-99 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: • **Yes**: The node task will be re-executed, and the following parameters must be configured: – **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: • **End the current job execution plan** • **Go to the next job** • **Suspend the current job execution plan** • **Suspend execution plans of the current and subsequent nodes**

8.6 EL

8.6.1 Expression Overview

Node parameter values in a Data Development 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.

 Copyright © Huawei Technologies Co., Ltd.

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.

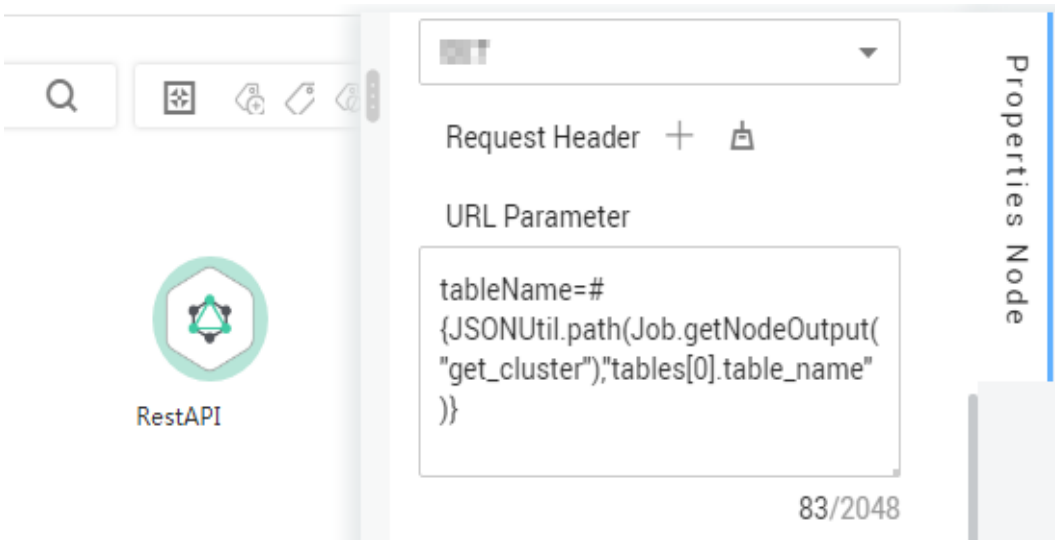
Example

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 8-28**.

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 8-28 Expression example



8.6.2 Date and Time Mode

The date and time in the EL expression can be displayed in the 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 8-100**.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-100 Letter description

Letter	Description	Example
G	Epoch	AD
y	Year	2001
M	Month in a year	July or 07
d	Day in a month	10
h	Hour in the 12-hour clock	12
H	Hour in the 24-hour clock	22
m	Minute	30
s	Second	55
S	Millisecond	234
E	Day of a week	Tuesday
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	No example
''	Single quotation mark	No example

8.6.3 Env Embedded Objects

The Env embedded object provides a method to obtain an environment variable value.

 Copyright © Huawei Technologies Co., Ltd.

Method

Table 8-101 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")}
```

8.6.4 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 8-102 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 DIS stream when the event-driven scheduling is used.
projectId	String	ID of the project where the Data Development module is located.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-103 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: **#{(Job.getNodeStatus("test")) == "success" }**
String getNodeOutput(String nodeName)	Obtains the output of a specified node. This method can only obtain the output of the previous dependent node. Currently, only the output of the Rest Client node, that is, the returned message of the REST request, can be obtained.
String getParam(String key)	Obtains a job parameter.
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 DLI job ID recorded during the running of the DLI SQL node can be obtained. For example, to obtain the job ID in the third statement of DLI node **DLI_INSERT_DATA**, run the following command: #{JSONUtil.path(Job.getRunningData("DLI_INSERT_DATA"),"jobIds[2]")}
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.

 Copyright © Huawei Technologies Co., Ltd.

Example

The expression used to obtain the output of node **test** in the job is as follows:

```
#{Job.getNodeOutput("test")}
```

8.6.5 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 appache 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,".")}
```

8.6.6 DateUtil Embedded Objects

A DateUtil embedded object provides methods of formatting time and calculating time.

Methods

Table 8-104 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.

 Copyright © Huawei Technologies Co., Ltd.

Method	Description
Date addMinutes(Date date, int amount)	After the specified number of minutes is added to Date, the new Date object is returned. The amount can be a negative number.
int getDay(Date date)	Obtains the day from the date. For example, if the date is 2018-09-14, 14 is returned.
int getMonth(Date date)	Obtains the month from the date. For example, if the date is 2018-09-14, 9 is returned.
int getYear(Date date)	Obtains the year from the date. For example, if the date is 2018-09-14, 2018 is returned.
Date now()	Returns the current time.
long getTime(Date date)	Converts the date type to the long type.
Date parseDate(String str, String pattern)	Converts the character string to the date by pattern. The pattern is the date and time mode. For details, see **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")}
```

8.6.7 JSONUtil Embedded Objects

A JSONUtil embedded object provides JSON object methods.

Methods

Table 8-105 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.

 Copyright © Huawei Technologies Co., Ltd.

Method	Description
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": "Shenzhen",
		"areaCode": "0755"
	},
	{
		"name": "city2",
		"areaCode": "010"
	},
	{
		"name": "Shanghai",
		"areaCode": "021"
	}]
}
```

The expression for obtaining the area code of Shenzhen is as follows:

```
#{JSONUtil.path(str,"cities[0].areaCode")}
```

8.6.8 Loop Embedded Objects

You can use Loop embedded objects to obtain data from the For Each dataset.

Property

Table 8-106 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.

 Copyright © Huawei Technologies Co., Ltd.

8.6.9 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 8-107 Method description

Method	Description
boolean isExistOBSPath(String obsPath)	Check whether the OBS file or the OBS directory that ends with a slash (/) exists. If the file or directory exists, **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")}

8.6.10 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 Data Development
- 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 data tables involved in this example are as follows:

- trade_log: Records data generated in each transaction.
- trade_report: 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**.

 Copyright © Huawei Technologies Co., Ltd.

- A database named **dli_db** has been created in DLI.

  If this database is not created, create one. For details, see **Creating a Database**.

- Data tables **trade_log** and **trade_report** have been created in the **dli_db** database.

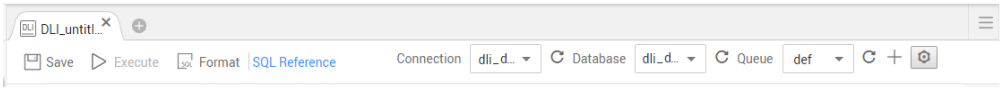
  If the data tables are not created, create them. For details, see **Creating a Data Table (Visualized Mode)** or **Creating a Data Table (DDL Mode)**.

Procedure

Step 1 Create and develop a SQL script.

1. In the navigation tree of the Data Development 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 8-29 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.

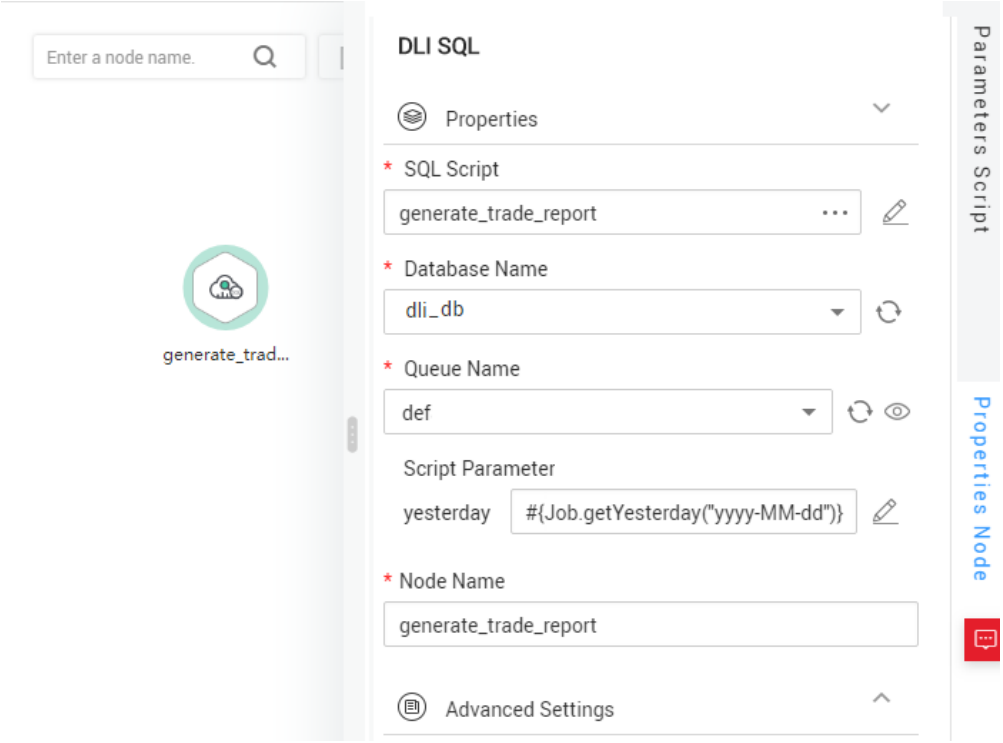
1. In the navigation tree of the Data Development console, choose **Data Development** > **Develop Job**.
2. Access the area on the right and click **Create Job** to create an empty job named **job**.

   Figure 8-30 Creating a job

 Copyright © Huawei Technologies Co., Ltd.

3. Go to the job development page, drag the DLI SQL node to the canvas, click the icon, and configure node properties.

Figure 8-31 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'
  ```

 Copyright © Huawei Technologies Co., Ltd.

4. Click to test the running job.
5. After the job test is complete, click to save the job configuration.

----End

8.7 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 reused by multiple solutions.

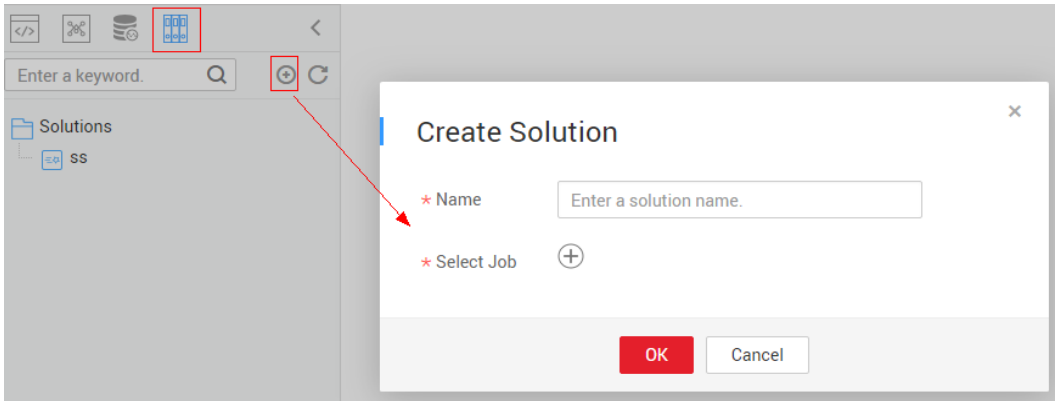
You can perform the following operations on a solution:

- **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.

Figure 8-32 Creating a solution



1. In the navigation tree on the left of the data development page, choose **Development** > **Develop Script** or **Data Development** > **Develop Job**.
2. In the upper part of the directory on the left, click . The solution directory is displayed.
3. Click in the upper part of the solution directory. The **Create Solution** page is displayed. **Table 8-108** describes the solution parameters.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-108 Solution Parameters

Parameter	Description
Name	Name of the solution.
Select Job	Select the jobs contained in the solution.

4. 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 choose **Edit** from the shortcut menu. See **Table 8-108**.

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

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.

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.

8.8 O&M and Scheduling

8.8.1 Overview

Log in to the DLF console and 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

 Copyright © Huawei Technologies Co., Ltd.

- Top 10 Job Instances with the Longest Execution Time in the Last 30 Days

  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.

- Top 10 Job Instances with the Largest Number of Running Exceptions in the Last 30 Days

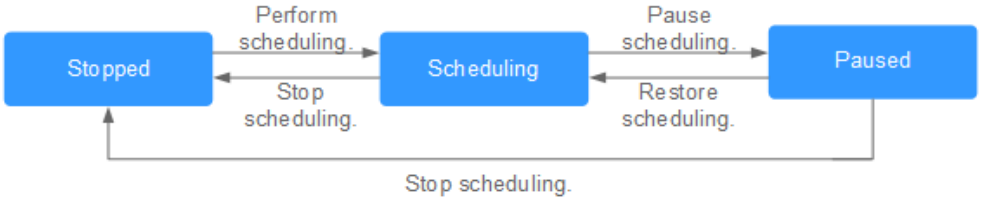
  Click the value in the **Exception Count** column. On the displayed **Monitor Instance** page, view the detailed running records of the job instance that is running abnormally.

8.8.2 Monitoring a Batch Job

Batch Processing: Scheduling Jobs

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 8-33 Scheduling a job



1. In the navigation tree on the left of the data development page, choose **Monitoring** > **Monitor Job**.
2. Click the **Batch Job Monitor** tab.
3. In the **Operation** column of the job, click **Run**/**Pause**/**Restore**/**Stop**.

Batch Processing: Scheduling the Depended Jobs

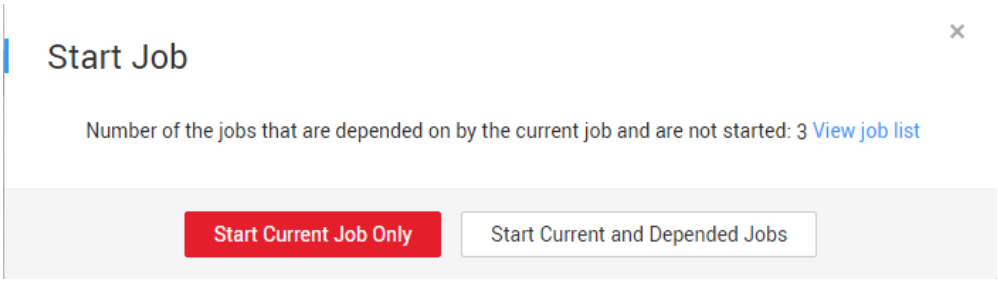
You can configure whether to start the depended jobs when scheduling a batch job on the job monitoring page. For details about how to configure depended jobs, see **Configuring Job Scheduling Tasks**.

1. In the navigation tree on the left of the data development page, choose **Monitoring** > **Monitor Job**.
2. Click the **Batch Job Monitor** tab and select a job that has depended jobs.
3. In the **Operation** column of the job, click **Schedule**.

   You can start only the current job or start the depended jobs at the same time when scheduling the job.

 Copyright © Huawei Technologies Co., Ltd.

Figure 8-34 Starting a job



Batch Processing: Notification Settings

You can configure DLF to notify you of job success or failure. The following provides the method for configuring a notification task. You can also configure a notification task on the **Monitoring** page. For details, see **Managing an Notification**.

1. In the navigation tree on the left of the data development page, choose **Monitoring** > **Monitor Job**.
2. Click the **Batch Job Monitor** tab.
3. In the **Operation** column of the job, choose **More** > **Set Notification**. In the displayed dialog box, configure notification parameters. **Table 8-117** describes the notification parameters.
4. Click **OK**.

Batch Processing: Instance Monitoring

You can view the running records of all instances of a job on the **Monitor Instance** page.

1. In the navigation tree on the left of the data development page, choose **Monitoring** > **Monitor Job**.
2. Click the **Batch Job Monitor** tab.
3. In the **Operation** column of a job, choose **More** > **Monitor Instance** to view the running records of all instances of the job.
   - For details about the **Operation** column of the instance, see **Table 8-110**.
   - For details about the **Operation** column of the node, see **Table 8-111**.

Batch Processing: Scheduling Configuration

You can perform the following steps to go to the development page of a specific job.

1. In the navigation tree on the left of the data development page, choose **Monitoring** > **Monitor Job**.
2. Click the **Batch Job Monitor** tab.
3. In the **Operation** column of a job, choose **More** > **Configure Scheduling**.

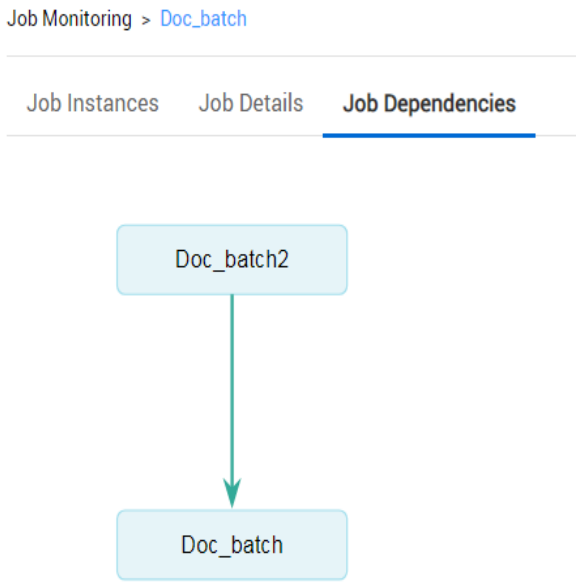
 Copyright © Huawei Technologies Co., Ltd.

Batch Processing: Job Dependency View

You can view the dependencies between jobs.

1. In the navigation tree on the left of the data development page, choose **Monitoring** > **Monitor Job**.
2. Click the **Batch Job Monitor** tab.
3. Click the job name and click the **Job Dependencies** tab. View the dependencies between jobs.

   Figure 8-35 Job dependencies view

   Click a job in the view. The development page of the job will be displayed.

Batch Processing: 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. In the navigation tree on the left of the data development page, choose **Monitoring** > **Monitor Job**.
2. Click the **Batch Job Monitor** tab.
3. In the **Operation** column of the job, choose **More** > **PatchData**.
4. Configure PatchData parameters. **Table 8-109** shows the PatchData parameters.

 Copyright © Huawei Technologies Co., Ltd.

Figure 8-36 PatchData parameters

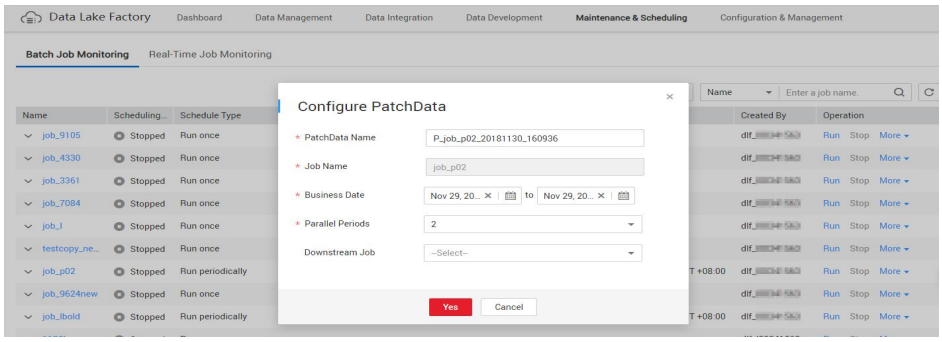


Table 8-109 Parameters

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.
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	Downstream job (job that depends on the current job) that requires PatchData. You can select more than one downstream jobs.

5. Click **OK**. The system starts to perform PatchData and the **PatchData Monitoring** page is displayed.

Batch Processing: Batch Processing Jobs

You can schedule and stop jobs and configure notification tasks in batches.

1. In the navigation tree on the left of the data development page, choose **Monitoring** > **Monitor Job**.
2. Click the **Batch Job Monitor** tab.
3. Select the jobs and click **Schedule**/**Stop**/**Configure Notification** to process the jobs in batches.

Batch Processing: Viewing Running Instances

This function enables you to view the information about five instances that are running in a job.

 Copyright © Huawei Technologies Co., Ltd.

1. In the navigation tree on the left of the data development page, choose **Monitoring** > **Monitor Job**.
2. Click the **Batch Job Monitor** tab.
3. Click in front of the job name. The **Latest Instances** page is displayed. You can view the information about the five instances that are running in the job.

Batch Processing: Viewing All Instances

You can view all running records of a job on the **Running History** page and perform more operations on instances or nodes based on site requirements.

1. In the navigation tree on the left of the data development page, choose **Monitoring** > **Monitor Job**.
2. Click the **Batch Job Monitor** tab.
3. Click the job name to go to the job instance page.

   You can stop, rerun, continue to run, or forcibly run jobs in batches. For details, see **Table 8-110**.

   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. View actions in the **Operation** column of an instance. **Table 8-110** describes the actions that can be performed on the instance.

Table 8-110 Actions for an instance

Action	Description
Stop	Stops an instance that is in the **Waiting**, **Running**, or **Abnormal** state.
Rerun	Reruns an instance that is in the **Succeeded** or **Canceled**.
View Waiting Job Instance	When the instance is in the waiting state, you can view the waiting job instance.
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.
Succeed	Forcibly changes the status of an instance from **Abnormal**, **Canceled**, **Failed** to **Succeed**.

 Copyright © Huawei Technologies Co., Ltd.

Action	Description
View	Goes to the job development page and view job information.

5. Click in front of an instance. The running records of all nodes in the instance are displayed.
6. View actions in the **Operation** column of a node. **Table 8-111** describes the actions that can be performed on the node.

Table 8-111 Actions for a node

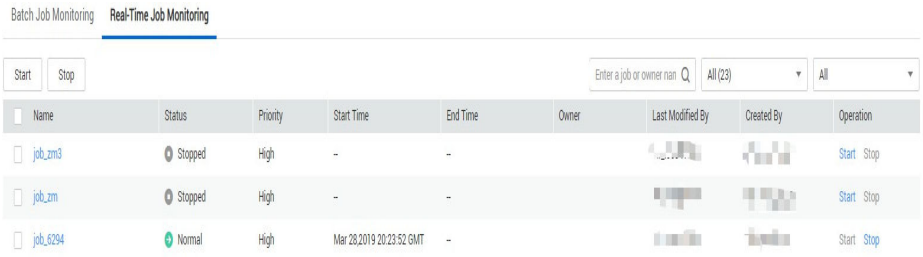
Action	Description
View Log	View the log information of the node.
Manual Retry	To run a node again after it fails, click **Retry**. **NOTE** This operation can be performed only when **Failure Policy** is set to **Suspend the current job execution plan**. To view the current failure policy, click a node and then click **Advanced Settings** on the **Node Properties** page.
Succeed	To change the status of a failed node to **Succeed**, click **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.
Skip	To skip a node that is to be run or that has been paused, click **Skip**.
Pause	To pause a node that is to be run, click **Pause**. Nodes queued after the paused node will be blocked.
Resume	To resume a paused node, click **Resume**.

8.8.3 Monitoring a Real-Time Job

Each real-time processing job generates only one instance. You can view information such as the running status, start time, and end time of the real-time processing job on the **Real-Time Job Monitoring** page, and start and stop the real-time processing job.

Copyright © Huawei Technologies Co., Ltd.

Figure 8-37 Real-time job monitoring page



- Batch job processing: Select jobs and click **Start**/**Stop**.
- Viewing node information of the job: Click a job name. On the page that is displayed, view the node connection relationship and monitoring information.
- Viewing status of the latest instances: Click the drop-down arrow next to the name of a job to view the latest instance status.

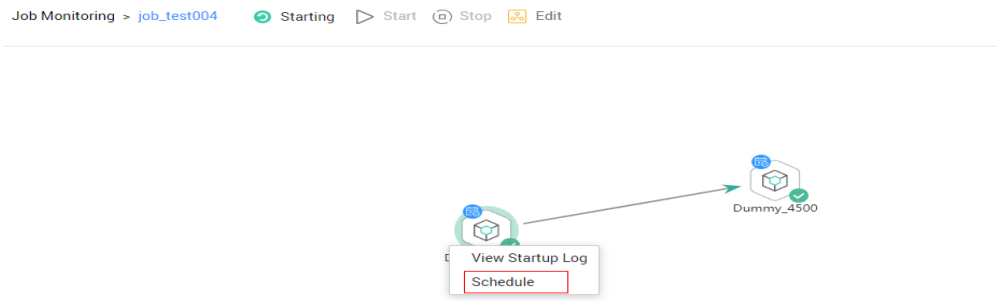
Real-Time Jobs: Configuring Event-based Scheduling on a Node

If event-driven scheduling is configured for a node in a real-time job, right-click the node and choose **Configure Scheduling** from the shortcut menu to view and modify the scheduling information about the node.

1. In the navigation tree on the left of the data development page, choose **Monitoring** > **Monitor Job**.
2. On the **Real-Time Job Monitoring** tab page, click a job name.
3. Right-click the node where event-driven scheduling is configured and choose **Configure Scheduling**. Set the parameters described in **Table 8-112**.

   When you click **View Startup Log**, you can view the job startup log information.

   Figure 8-38 Configuring scheduling

   When **Triggering Event Type** of **Event-Driven Scheduling** is set to **DIS**, configure a DIS scheduling policy as follows:

 Copyright © Huawei Technologies Co., Ltd.

Figure 8-39 Configuring a DIS scheduling policy

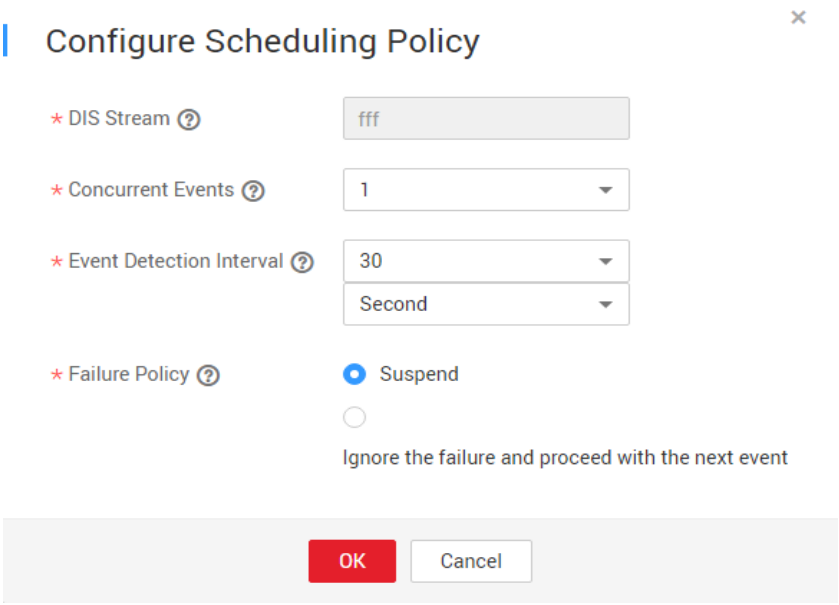


Table 8-112 Policy parameters

Parameter	Description
DIS Stream	Name of the DIS stream. When a new message is sent to the specified DIS stream, Data Development transfers the new message to the job to trigger the job running.
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 DIS stream or OBS path for new messages. The unit of the interval can be set to second or minute.
Failure Policy	Select a policy to be performed after scheduling fails. • Stop scheduling • Ignore failure and proceed

Rerunning a Job After the Job Is Paused

If event-based scheduling is configured for a node in a real-time job, you can disable the node and then restore it. When the node is restored, you can select where you can resume running.

1. In the navigation tree on the left of the data development page, choose **Monitoring** > **Monitor Job**.
2. On the **Real-Time Job Monitoring** tab page, click a job name.

 Copyright © Huawei Technologies Co., Ltd.

3. On the job monitoring page, right-click a node with event-driven scheduling configured and choose **Disable** from the shortcut menu.
4. Right-click the node and choose **Resume** from the shortcut menu. The **Resume Node Running** dialog box is displayed, as shown in **Table 8-113**.

Figure 8-40 Resuming node running

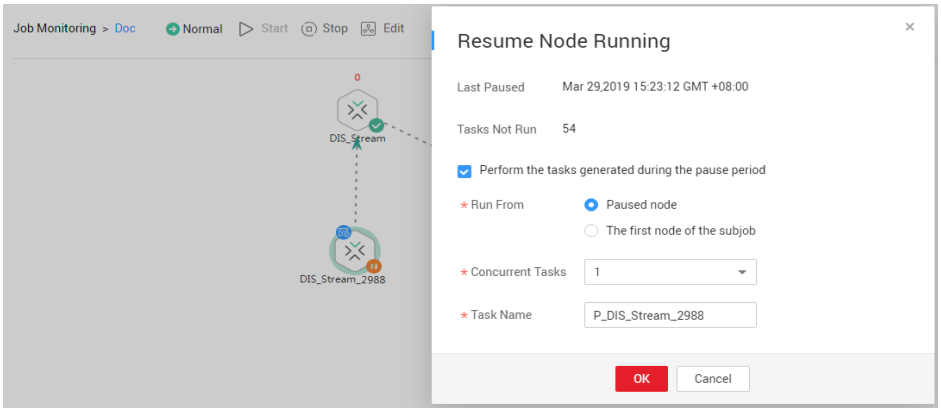


Table 8-113 Resumption parameters

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. • 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.

8.8.4 Monitoring a Real-Time Sub-job

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. In the navigation tree on the left of the data development page, choose **Monitoring** > **Monitor Job**.

 Copyright © Huawei Technologies Co., Ltd.

2. On the **Real-Time Job Monitoring** tab page, click a job name.
3. Click a node with event-based scheduling configured, as shown in **Figure 8-41**.

Figure 8-41 Subjob monitoring page

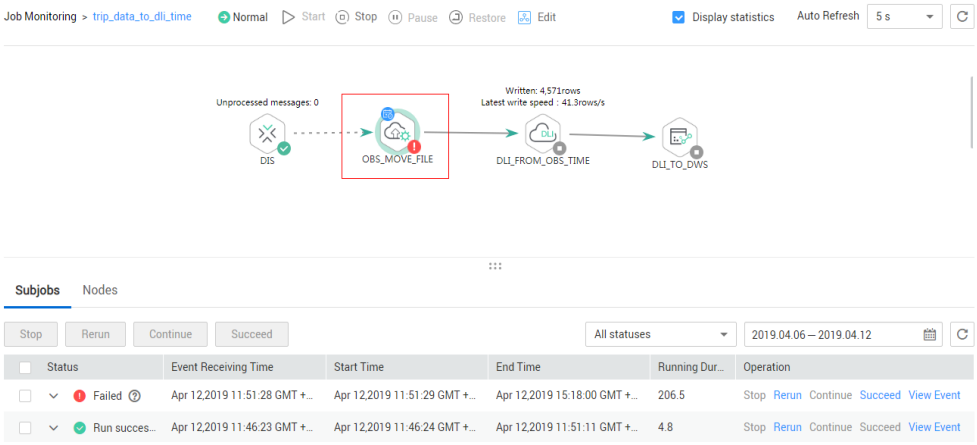


Table 8-114 describes the actions listed in the **Operation** column of each subjob.

Table 8-114 Actions in the **Operation** column

Action	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.
Succeed	Forcibly changes the status of a subjob instance from **Failed** to **Succeed**.
View Event	Displays the event content of a subjob.

4. Click in the **Status** column. The running records of the subjob node are displayed.

   Table 8-115 describes the actions that can be performed on the node.

 Copyright © Huawei Technologies Co., Ltd.

Table 8-115 Actions for a node

Action	Description
View Log	View the log information of the node.
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.
Succeed	To change the status of a failed node to **Succeed**, click **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.
Skip	To skip a node that is to be run or that has been paused, click **Skip**.
Pause	To pause a node that is to be run, click **Pause**. Nodes queued after the paused node will be blocked.
Resume	To resume a paused node, click **Resume**.

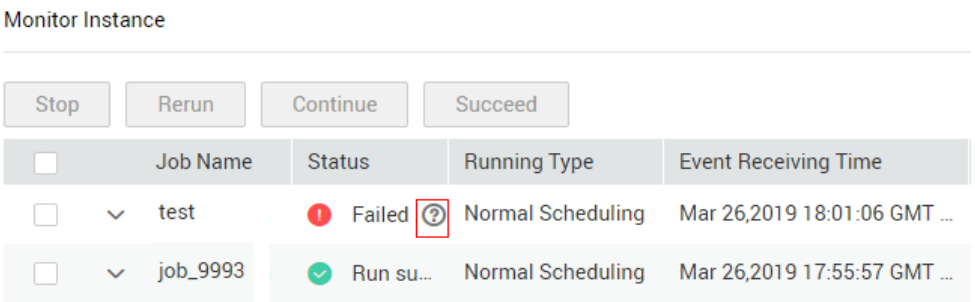
8.8.5 Monitoring an Instance

In the navigation tree of the Data Development module, choose **Monitoring**. On the Monitor Instance page, you can view the job instance information and perform more operations on instances and nodes as required. For details, see **Batch Processing: Viewing All Instances**.

You can search for instances by job name, creator, or CDM job. Search by CDM job is to search for job instances by node.

If the job instance is in the **Failed** state, you can mouse over the question mark and view the pop-up message to understand the failure cause, as shown in **Figure 8-42**.

Figure 8-42 Job instance failure prompt



 Copyright © Huawei Technologies Co., Ltd.

Rerunning Job Instances

You can rerun a job instance that is successfully executed or fails to be executed by setting its rerun position.

1. In the navigation tree on the left of the data development page, choose **Monitoring** > **Monitor Instance**
2. 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.

Figure 8-43 Setting the rerunning position

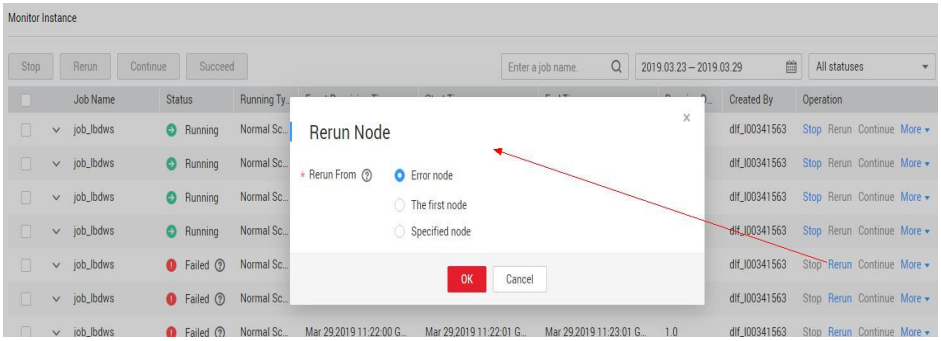


Table 8-116 Parameter description

Parameter	Description
Rerun From	Select the start position from which the job instance reruns. • **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 be run, it reruns since the node specified in the job instance. **NOTE** A job instance reruns from its first node if either of the following cases occurs: • The quantity or name of a node in the job changes. • The job instance has been successfully run.

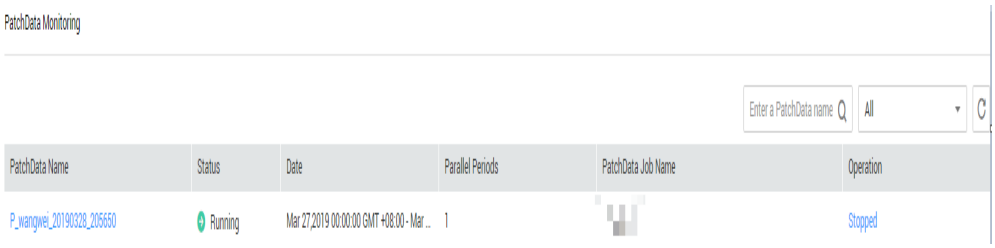
8.8.6 Monitoring PatchData

In the navigation tree of the Data Development console, choose **Monitoring** > **Monitor PatchData**.

 Copyright © Huawei Technologies Co., Ltd.

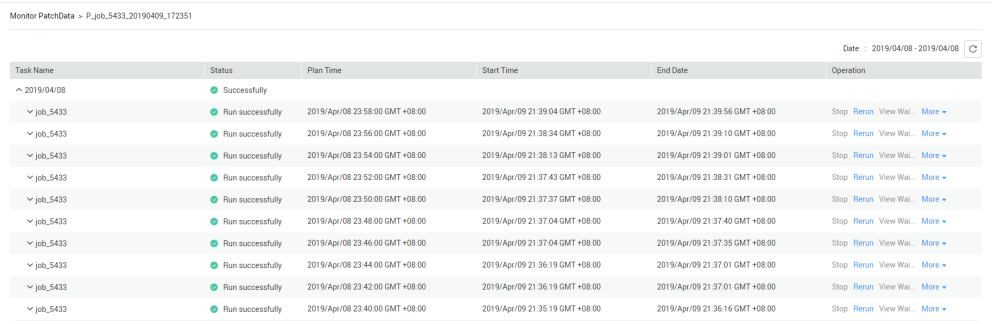
On the page shown in **Figure 8-44**, you can view the task status, service date, number of parallel periods, and PatchData job names, and stop a running task.

Figure 8-44 PatchData Monitoring page



On the page shown in **Figure 8-44**, click PatchData name. On the displayed page, you can view the PatchData execution status. For more information, see **Batch Processing: Viewing All Instances**.

Figure 8-45 PatchData monitoring details



8.8.7 Managing an 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. In the navigation tree on the left of the data development page, choose **Monitoring** > **Manage Notification**.
2. On the **Notification Management** tab page, click **Configure Notification**. In the displayed dialog box, configure parameters. **Table 8-117** describes the parameters.

Table 8-117 Notification parameters

Parameter	Mandatory	Description
Job Name	Yes	Name of the job.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Notification Type	Yes	Type of the notification. • **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. **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.
Topic for Abnormal/Failed Running	Yes	Select a topic for abnormal job running or job failure. **NOTE** Currently, only SMS, email, or HTTP are supported to subscribe to topics.
Topic for Successful Running	Yes	Select a topic for successful job running.
Notification	Yes	Whether to enable the notification function. The function is enabled by default.

3. Click **OK**.

   NOTE

   - The Data Development 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 tree on the left of the data development page, choose **Monitoring** > **Manage Notification**.
2. Click the **Notification Management** tab.

 Copyright © Huawei Technologies Co., Ltd.

3. In the **Operation** column of a notification, click **Edit**. In the displayed dialog box, edit notification parameters. **Table 8-117** 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 tree on the left of the data development 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 tree on the left of the data development 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 tree on the left of the data development page, choose **Monitoring** > **Manage Notification**.
2. Click the **Notification Management** tab.
3. In the **Operation** column of the notification, click **Delete**. The **Delete Notification** dialog box is displayed.
4. Click **OK**.

8.8.8 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.

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.

Copyright © Huawei Technologies Co., Ltd.

- OBS has been enabled and a folder has been created in OBS.

Creating a Notification

1. In the navigation tree on the left of the data development page, choose **Monitoring** > **Manage Notification**.
2. On the **Cycles** tab page, click **Create Notification**. In the displayed dialog box, configure parameters. **Table 8-118** describes the notification parameters.

Figure 8-46 Create a notification

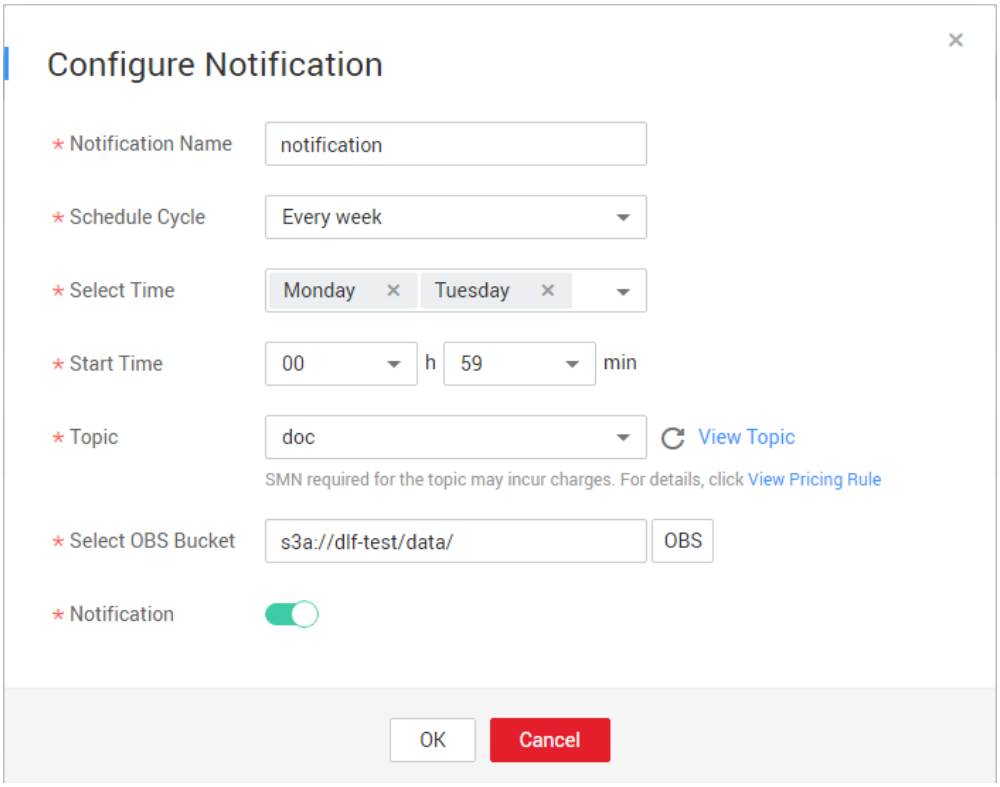


Table 8-118 Notification parameters

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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Select Time	Yes	Time when the notification is sent. • 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.
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.

3. Click **OK**.

   NOTE

   DLF sends notifications through SMN. Using SMN may incur fees. For pricing details, contact the SMN support personnel.

4. 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.

8.8.9 Backing Up and Restoring an Asset

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.

Prerequisites

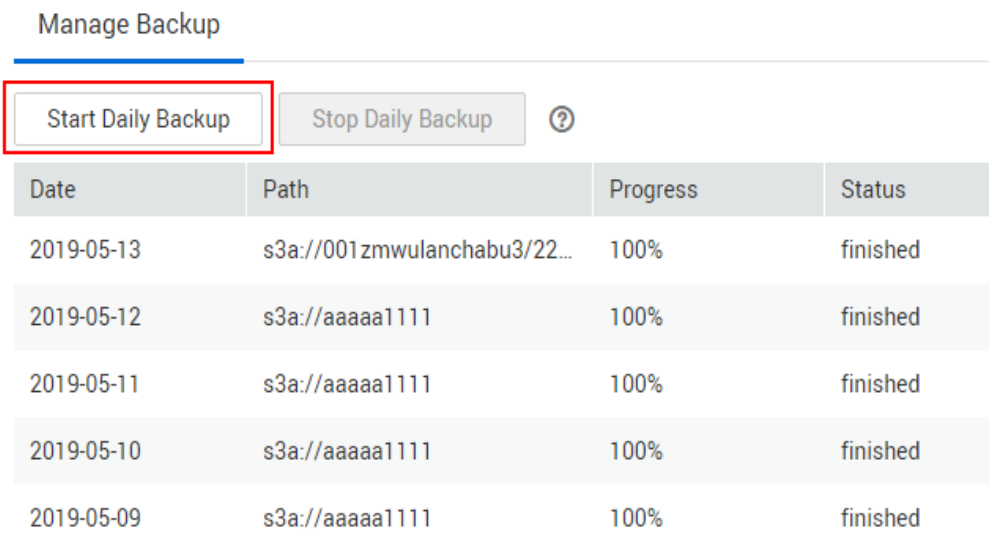
OBS has been enabled and a folder has been created in OBS.

 Copyright © Huawei Technologies Co., Ltd.

Backing Up an Asset

1. In the navigation tree on the left, choose **Manage Backup**.
2. Click **Start Daily Backup**. In the **Browse OBS File** dialog box, select an OBS folder.

Figure 8-47 Managing backup



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 an Asset

Step 1 In the navigation tree of the Data Development console, choose **Manage Backup**.

Step 2 On the **Manage Restoration** tab, click **Restore Backup**.

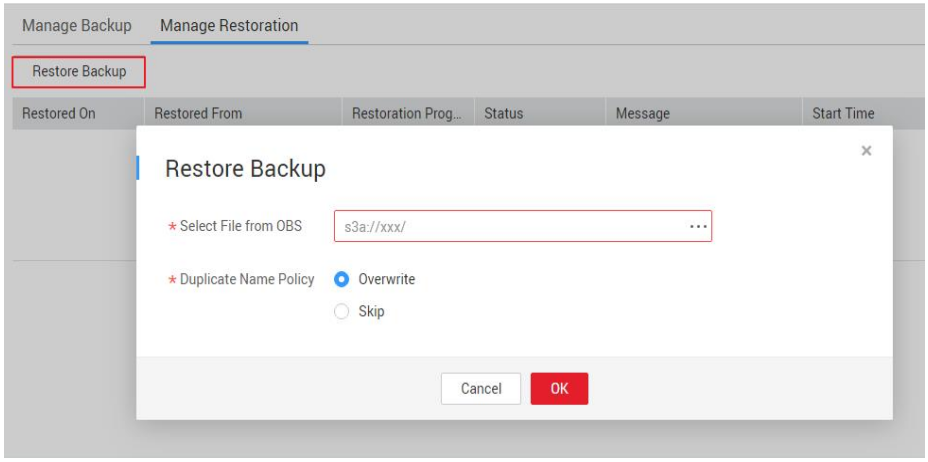
In the **Restore Backup** dialog box, select the storage path of the asset to be restored from the OBS bucket and set the duplicate name policy.

NOTE

- The storage path is the file path generated in **Backing Up an Asset**.
- 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).

 Copyright © Huawei Technologies Co., Ltd.

Figure 8-48 Restoring assets



Step 3 Click **OK**.

----End

8.9 Configuration and Management

8.9.1 Managing a Host Connection

DLF can save the host connection information required for scripts and jobs. If the host connection information changes, you only need to edit it on the **Host Connection Management** page, but do not need to edit it in scripts or jobs one by one.

Creating a Host Connection

Ensure that the quantity of host connections is less than the maximum connection quota (20).

1. In the navigation tree on the left of the data development page, choose **Configuration** > **Manage Host Connection**.
2. Click **Create Host Connection**. The **Data Connection Management** page of DGC Management Center is displayed. Click **Create Data Connection** and set **Data Connection Type** to **Host Connection**. In the displayed **Create Data Connection** dialog box, set the parameters listed in **Table 8-119**.

Table 8-119 Host connection parameters

Parameter	Mandatory	Description
Host Connection Name	Yes	Name of the host connection. Must consist of 1 to 100 characters, contain only letters, digits, and underscores (_), and be unique.

Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
Host IP Address	Yes	IP address of the host. For details, see **Viewing ECS Details**.
Connection Type	Yes	• Proxy connection • Direct connection
Host ID	Yes	The ID of the host where the ECS is deployed.
Port	Yes	Port number of the host.
Username	Yes	Username of the host.
Login Mode	Yes	Selects the login mode of the host. • Key pair • Password
Key Pair	Yes	If Key Pair is the login mode of the host, the user needs to obtain the private key file, upload it to OBS, and select an OBS path. This parameter is available only when **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	If the login mode of the host is to use a password, enter a login password.
CDM Cluster	Yes	Name of the Cloud Data Migration (CDM) cluster.
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	Descriptive information about the host connection.

3. Click **Test** to test connectivity to the host. If the test fails, the host connection cannot be created. In this case, contact technical support.
4. Click **OK**.

 Copyright © Huawei Technologies Co., Ltd.

Editing a Host Connection

After a host connection is created, you can modify host connection parameters as required.

1. In the navigation tree on the left of the data development page, choose **Configuration** > **Manage Host Connection**.
2. In the **Operation** column of the host connection, click **Edit**. In the displayed dialog box, modify the host connection parameters. Refer to **Table 8-119** for details.

   NOTE

   For security purposes, you need to enter the password of the host again when modifying the host connection parameters.
3. Click **Test** to test connectivity to the host. If the host can be connected, the host connection has been successfully created.
4. Click **OK**.

Deleting a Host Connection

If you do not need to use a host connection any more, perform the following operations to delete the host connection:

1. In the navigation tree on the left of the data development page, choose **Configuration** > **Manage Host Connection**.
2. In the **Operation** column of the host connection, click **Delete**. The **Delete Host Connection** dialog box is displayed.
3. Click **OK**.

NOTICE

If you forcibly delete a host connection that is being associated with a script or job, ensure that services are not affected by going to the script or job development page and reassociating an available host connection with the script or job.

8.9.2 Managing a Resource

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 Spark Job**.

(Optional) Creating a Directory

If a directory exists, you do not need to create one.

Copyright © Huawei Technologies Co., Ltd.

1. In the navigation tree on the left, choose **Configuration** > **Manage Resource**.
2. In the directory list, click . In the displayed dialog box, configure directory parameters. **Table 8-120** describes the directory parameters.

Table 8-120 Resource directory parameters

Parameter	Description
Directory Name	Name of the resource directory. Must consist of 1 to 32 characters and contain only letters, digits, underscores (_), and hyphens (-).
Select Directory	Parent directory of the resource directory. The parent directory is the root directory by default.

3. Click **OK**.

Creating a Resource

You have enabled OBS before creating a resource.

1. In the navigation tree on the left, choose **Configuration** > **Manage Resource**.
2. Click **Create Resource**. In the displayed dialog box, configure resource parameters. **Table 8-121** describes the resource parameters. Click **Yes**.

Table 8-121 Resource management parameters

Parameter	Mandatory	Description
Name	Yes	Name of the resource. Must consist of 1 to 32 characters and contain only letters, digits, underscores (_), and hyphens (-).
Type	Yes	File type of the resource. Possible values: • jar: JAR file • pyFile: User Python file • file: User file • ModelFile: User AI model file
Resource Location	Yes	Location of the resource. Currently, only OBS is supported.
Main JAR package	Yes	Main 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**.
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**.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Mandatory	Description
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. In the navigation tree on the left, choose **Configuration** > **Manage Resource**.
2. In the **Operation** column of the resource, click **Edit**. In the displayed dialog box, modify the resource parameters. For details, see **Table 8-121**.
3. Click **OK**.

Deleting a Resource

If you do not need to use a resource any more, perform the following operations to delete it. Before deleting a resource, ensure that it is not used by any jobs.

1. In the navigation tree on the left, choose **Configuration** > **Manage Resource**.
2. In the **Operation** column of the resource, click **Delete**. The **Delete Resource** dialog box is displayed.
3. Click **Yes**.

Importing a Resource

To import a resource, perform the following operations:

1. In the navigation tree on the left, choose **Configuration** > **Manage Resource**.
2. In the resource directory, right-click and select **Import Resource**. The **Import Resource** dialog box is displayed.
3. 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. In the navigation tree on the left, choose **Configuration** > **Manage Resource**.
2. In the **Operation** column of the resource, click **Export**. The **Export Resource** dialog box is displayed.

 Copyright © Huawei Technologies Co., Ltd.

3. Select **Export resource definitions only** or **Export resource definitions and files**, and click **OK**.

8.9.3 Configuring an Environment Variable

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.

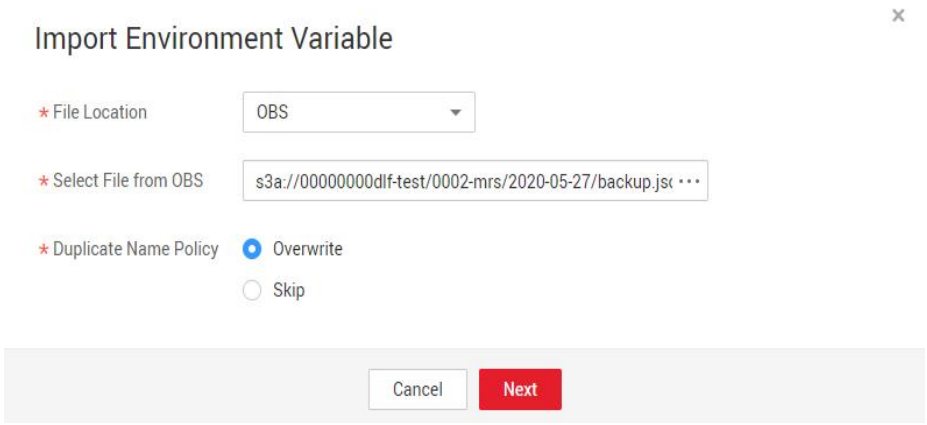
Importing Environment Variables

Step 1 In the navigation tree on the left, choose **Specifications**.

Step 2 Click **Environment Variables**. On the **Environment Variables** page, click **Import**.

Step 3 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 8-49 Importing environment variables (with OBS)



----End

Configuration Method

Step 1 In the navigation tree on the left, choose **Specifications**.

Step 2 On the **Environment Variable** page, set the parameters described in **Table 8-122** and click **Save**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 8-50 Configuring environment variables

Table 8-122 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. 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 Data Development module.

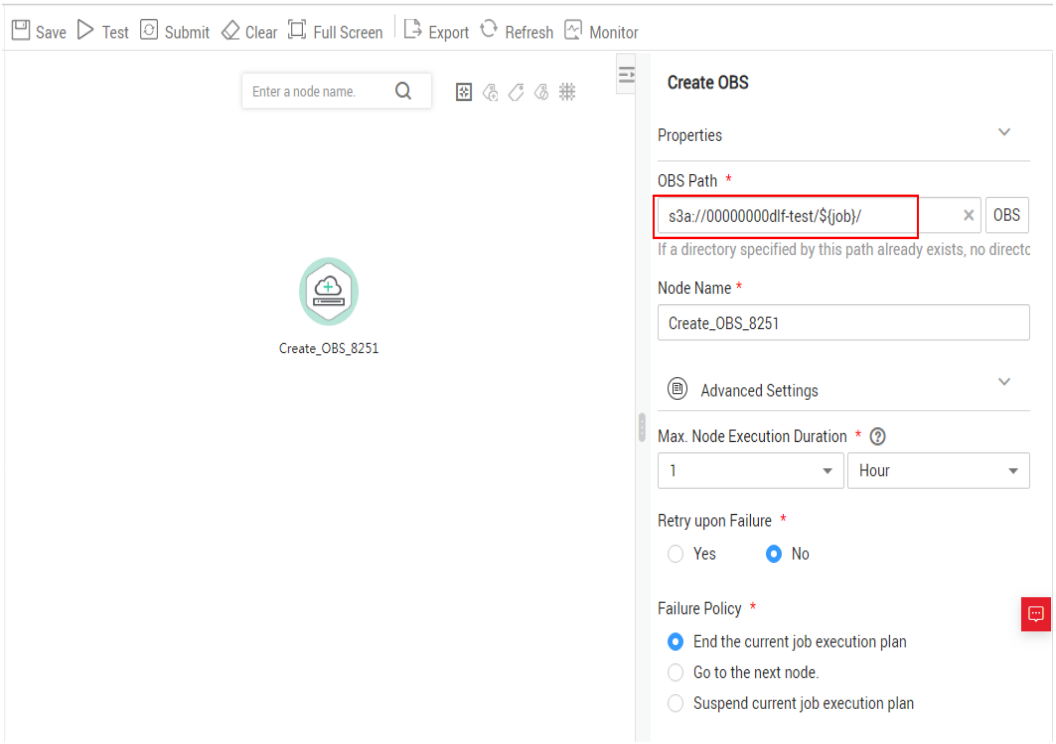
 Copyright © Huawei Technologies Co., Ltd.

- 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 8-51 Create OBS



Step 3 Click **Save** and then **Monitor** to monitor the running status of the job.

----End

8.9.4 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 In the navigation tree on the left, choose **Specifications**.

Step 2 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

 Copyright © Huawei Technologies Co., Ltd.

8.9.5 Configuring Agencies

The following problems may occur during job execution in Data Development:

- The job execution mechanism of the Data Development 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 a cloud service 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

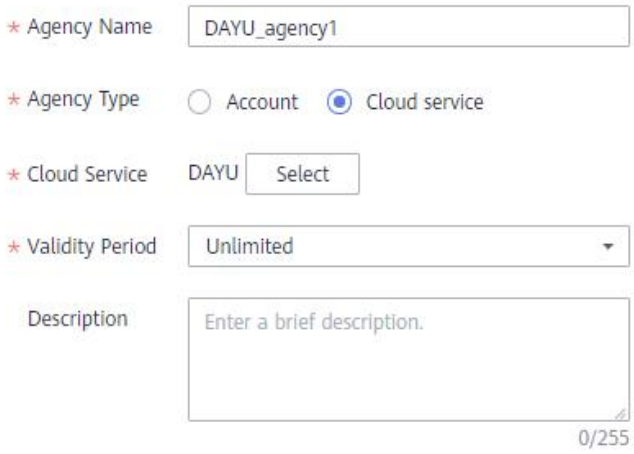
- To create or modify an agency, you must have the **Security Administrator** permissions.
- To configure a workspace-level agency, you must have the **DGC Administrator** or **Tenant Administrator** policy.
- All users who have the permissions to view the list of agencies can configure a job-level agency.

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, Example: DGC _agency.
4. Set **Agency Type** to **Cloud service**. Select **DGC** for **Cloud Service**. DGC then can perform resource O&M operations on behalf of you. See **Figure 8-52**.
5. Set **Validity Period** to **Unlimited**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 8-52 Creating an agency



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 8-53**.
   - Users assigned the **Tenant Administrator** policy have all permissions on all services except on IAM. Therefore, delegate the **Tenant Administrator** policy to DGC so thatDGC 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 **Permission Assignment**.

   Figure 8-53 Assigning permissions

8. Click **OK**.

Permission Assignment

After the operation permissions of an account are delegated to DGC, you need to configure the permissions of the agency identity so that DGC can interact with other services.

To meet the security control requirements for fewer permissions, you can configure the **Admin** permissions for services based on the node types in jobs. For details, see **Table 8-123**.

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

 Copyright © Huawei Technologies Co., Ltd.

access control of cloud resources. Configure the permissions according to **Table 8-124**. 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 that has MRS 2.1.0.1 or later installed

- 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 8-123 The **admin** permissions for related nodes

Node Name	System Permission	Description
CDM Job, DIS Stream, DIS Dump, and DIS Client	DGC Administrator	All Data Lake Governance Center 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.
• 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	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. Users assigned the **KMS Administrator** role have the administrator permissions for encryption keys in DEW.

 Copyright © Huawei Technologies Co., Ltd.

Node Name	System Permission	Description
MRS Spark SQL, MRS Hive SQL, MRS Kafka, and Kafka Client (connecting to the clusters in proxy mode)	DGC Administrator KMS Administrator	DGC Administrator has all permissions required for DGC. Users assigned the **KMS Administrator** role have the administrator permissions for encryption keys in DEW.
DLI Flink Job, DLI SQL, and DLI Spark	DLI Service Admin	All operation permissions for DLI.
DWS SQL, RDS SQL, and Shell (connecting to data sources in proxy mode)	DGC Administrator KMS Administrator	DGC Administrator has all permissions required for DGC. Users assigned the **KMS Administrator** role have the administrator permissions for encryption keys in DEW.
CSS	DGC Administrator Elasticsearch Administrator	DGC Administrator has all permissions required for DGC. Users assigned the **Elasticsearch Administrator** role have all permissions for CSS. This role depends on the **Tenant Guest** and **Server Administrator** roles in the same project.
Create OBS, Delete OBS, and OBS Manager	OBS OperateAccess	Basic object operation permissions, such as viewing buckets, uploading objects, obtaining objects, deleting objects, and obtaining object ACLs.
SMN	SMN Administrator	All operation permissions for SMN.

- Configure fine-grained permissions. (Create custom policies based on the actions supported by each service.)

  For details on how to create a custom policy, see **Creating Custom Policies**.

 Copyright © Huawei Technologies Co., Ltd.

NOTE

- During job execution, you need to 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, DIS Stream, DIS Dump and DIS Client nodes belong to the DGC module. DGC does not support fine-grained authorization. Therefore, only the **DGC 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 **DGC 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 8-124 Creating a custom policy

Node Name	Action
Import GES	• ges:graph:access • ges:graph:getDetail • ges:jobs:getDetail
• 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: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)	• kms:dek:crypto • kms:cmk:get • DGC Administrator (role)

 Copyright © Huawei Technologies Co., Ltd.

Node Name	Action
DLI Flink Job, DLI SQL, and DLI Spark	• dli:jobs:get • dli:jobs:update • dli:jobs:create • dli:queue:submit_job • dli:jobs:list • dli:jobs:list_all
DWS SQL, RDS SQL, and Shell (connecting to data sources in proxy mode)	• kms:dek:crypto • kms:cmk:get • DGC Administrator (role)
Create OBS, Delete OBS, and OBS Manager	• 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

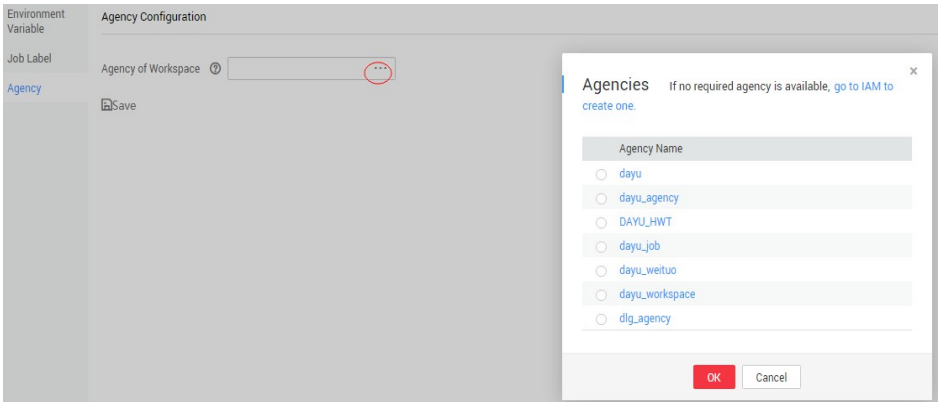
CAUTION

A workspace-level agency impacts on all jobs. Some jobs contain nodes related to MRS. Exercise caution when performing this operation.

1. In the navigation tree on the left, choose **Specifications**.
2. Click **Agency**. On the displayed page, configure an agency.
3. 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**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 8-54 Configuring a workspace-level agency



4. Click **OK** to return to the **Agency Configuration** page. Then, click to save the settings.

Configuring a Job-level Agency

NOTE

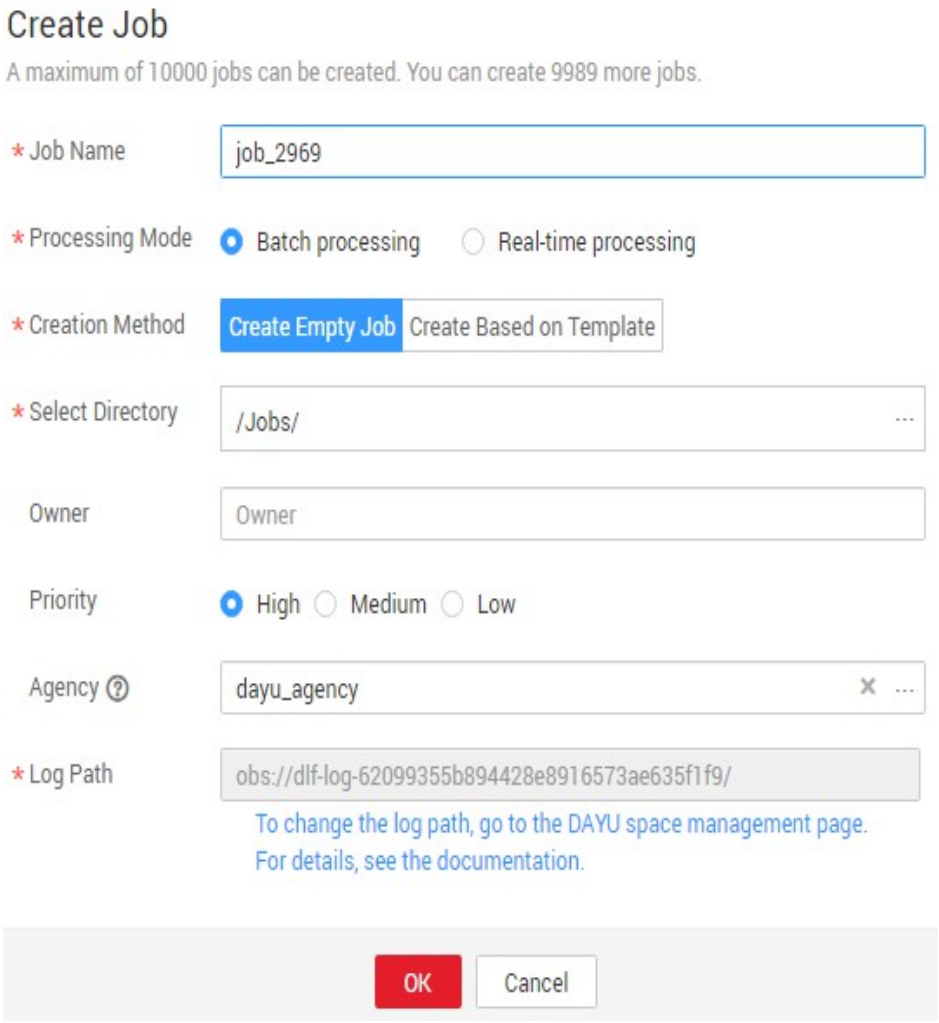
You can create a job-level agency when creating a job. You can also modify the agency of an existing job.

Configuring an agency when creating a job

1. In the navigation pane of the Data Development homepage, choose **Development** > **Develop Job**.
2. 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.

 Copyright © Huawei Technologies Co., Ltd.

Figure 8-55 Configuring a Job Agency



Modifying the agency of an existing job

1. In the navigation pane of the Data Development 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.

8.10 Usage Guidance

8.10.1 Developing a Spark Job

This section introduces how to develop Spark jobs on Data Development.

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

 Copyright © Huawei Technologies Co., Ltd.

logic. In this case, Spark jobs can help. This section uses an example to demonstrate how to submit a Spark job on Data Development.

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 Data Development.
3. Create a Data Development job and submit it using the DLI Spark node.

Preparations

- Object Storage Service (OBS) has been enabled and a bucket, for example, **s3a://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

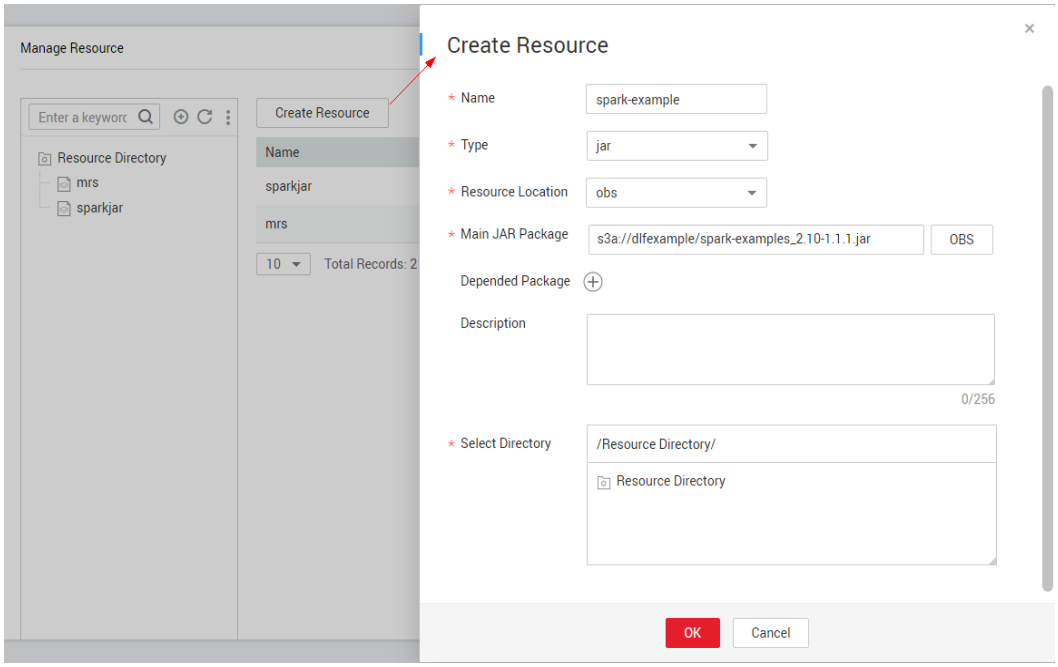
The Spark job code used in this example comes from the maven repository that can be download from **https://repo.maven.apache.org/maven2/org/apache/spark/spark-examples_2.10/1.1.1/spark-examples_2.10-1.1.1.jar**. This Spark job is to calculate the approximate value of π.

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 In the navigation tree on the left, choose **Configuration** > **Manage Resource**. Create resource **spark-example** on Data Development and associate it with the JAR package obtained in **Step 1**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 8-56 Creating a resource



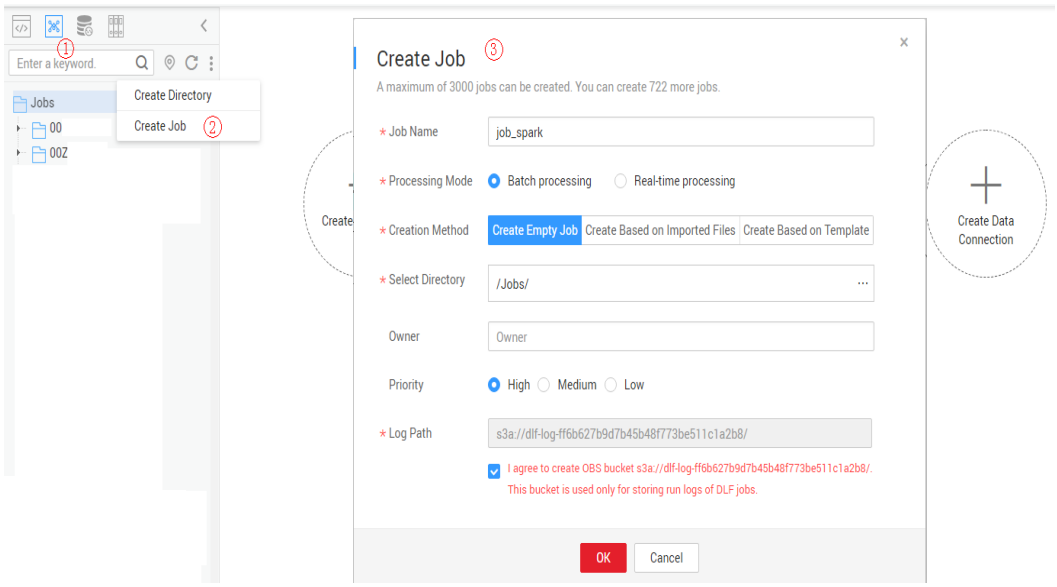
----End

Submitting a Spark Job

You need to create a job on Data Development and submit the Spark job using the DLI Spark node of the job.

Step 1 Create a job named **job_spark** for the Data Development module.

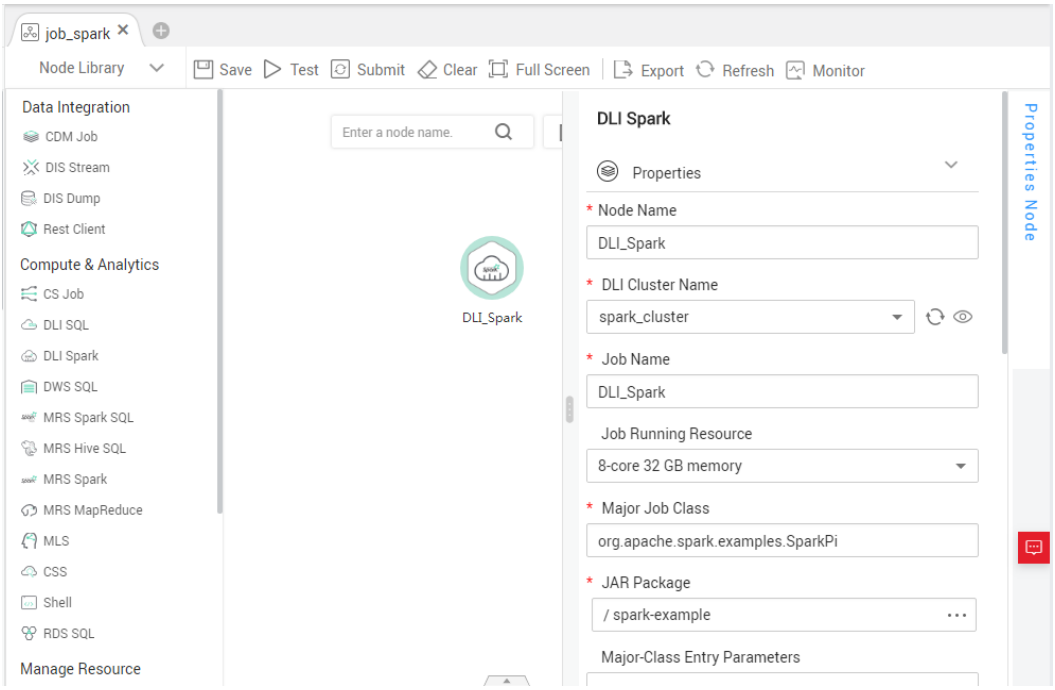
Figure 8-57 Creating a job



Step 2 Go to the job development page, drag the DLI Spark node to the canvas, and click the node to configure node properties.

 Copyright © Huawei Technologies Co., Ltd.

Figure 8-58 Configuring node properties

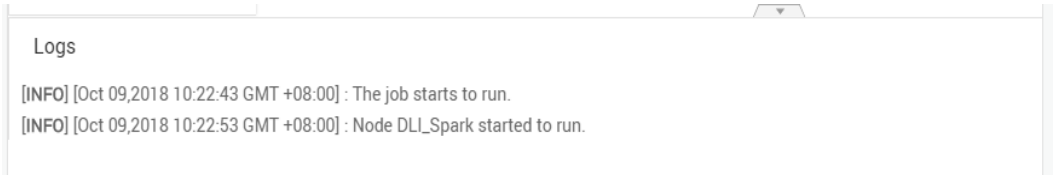


Description of key properties:

- DLI Cluster Name: Name of the Spark cluster created in **Preparations**.
- Job Running Resource: Maximum CPU and memory resources that can be used when a DLI Spark node is running.
- Major Job Class: Main class of a DLI Spark node. In this example, the main class is **org.apache.spark.examples.SparkPi**.
- JAR Package: Resource created in **Step 2**.

Step 3 After the job orchestration is complete, click to test the job.

Figure 8-59 Job logs (for reference only)



Step 4 If no problems are recorded in logs, click to save the job.

----End

8.10.2 Developing Hive SQL Scripts

This section introduces how to develop Hive SQL scripts on Data Development.

Scenario Description

As a one-stop big data development platform, Data Development 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

- MapReduce Service (MRS) has been enabled and the MRS cluster **MRS_1009** has been created for providing an operating environment for Hive SQL.

  When creating an MRS cluster, note the following:

  – Kerberos authentication is disabled for the cluster.

  – Hive is available.

- Cloud Data Migration (CDM) has been enabled and the CDM cluster **cdm-7357** has been created for providing a proxy for communication between Data Development and MRS.

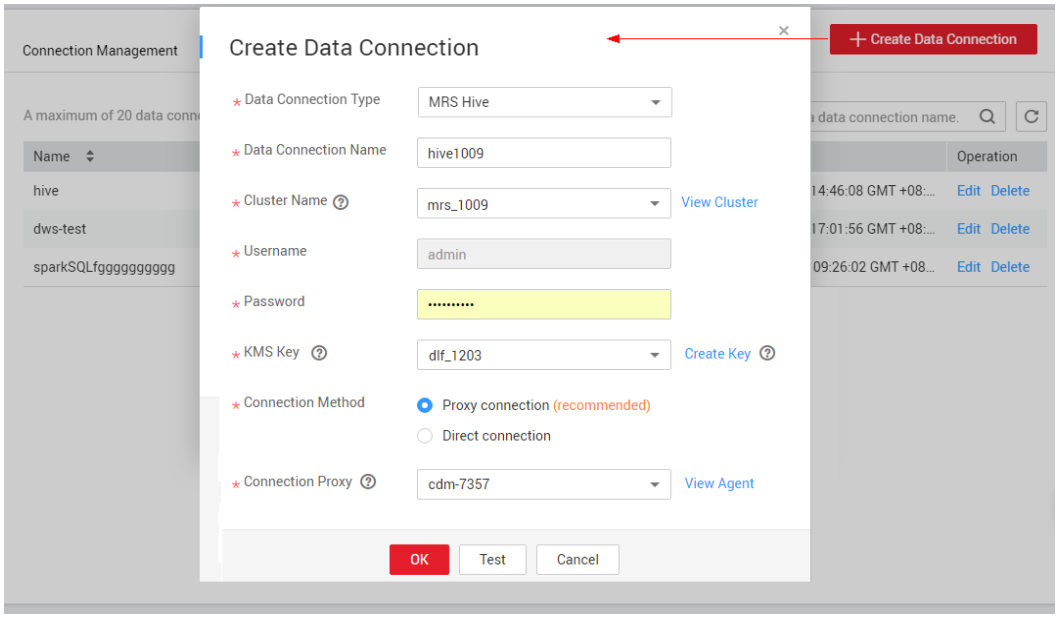
  When creating a CDM cluster, note the following:

  – The virtual private cloud (VPC), subnet, and security group can communicate with the MRS cluster **MRS_1009**.

Creating a Hive Data Connection

Before developing a Hive SQL script, create a connection to MRS Hive on Data Development. In this example, the data connection is named **hive1009**.

Figure 8-60 Creating a data connection



Description of key parameters:

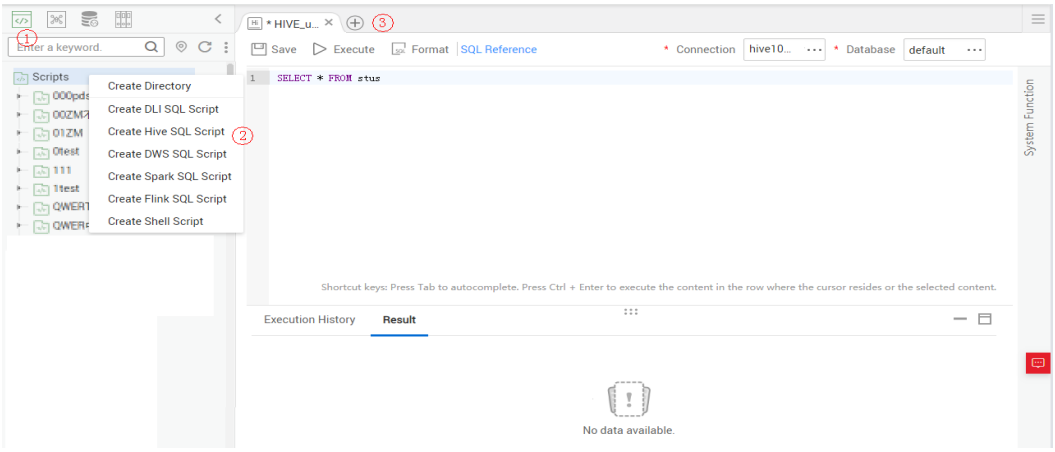
- Cluster Name: Name of the MRS cluster created in **MapReduce Preparations**.
- CDM Cluster: Name of the CDM cluster created in **CDM Preparations**.

Developing a Hive SQL Script

Create a Hive SQL script named **hive_sql** on Data Development. Then enter SQL statements in the editor to fulfill business requirements.

 Copyright © Huawei Technologies Co., Ltd.

Figure 8-61 Developing a script



Notes:

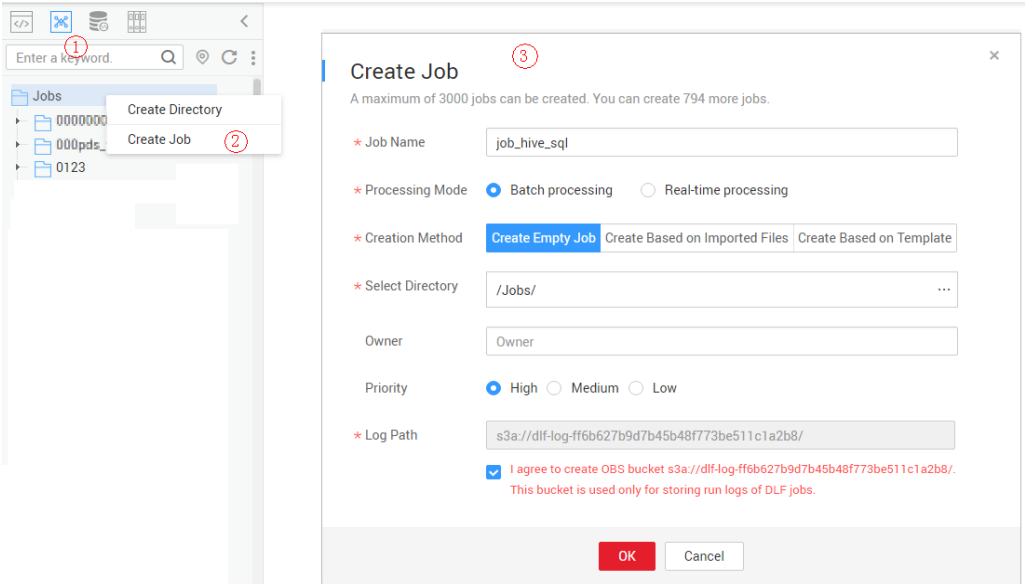
- The script development area in **Figure 8-61** is a temporary debugging area.
  After you close the tab page, the development area will be cleared. Click to save the script to a specified directory.
- Data Connection: Connection created in **Creating a Hive Data Connection**.

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 Data Development job named **job_hive_sql**.

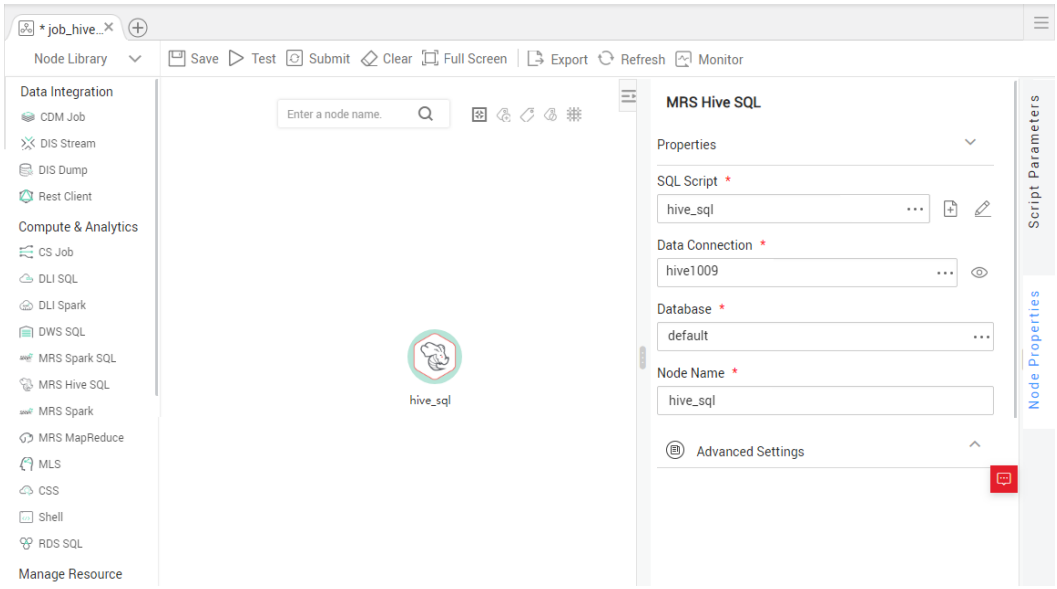
Figure 8-62 Creating a job named **job_hive_sql**



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.

 Copyright © Huawei Technologies Co., Ltd.

Figure 8-63 Configuring properties for an MRS Hive SQL node



Description of key properties:

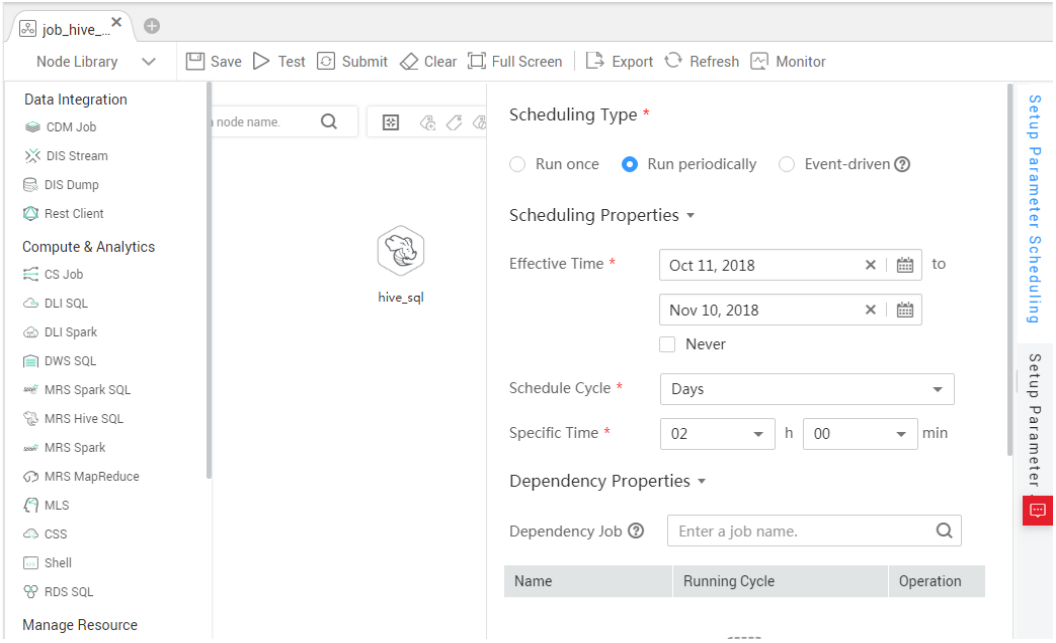
- SQL Script: Hive SQL script **hive_sql** that is developed in **Developing a Hive SQL Script**.
- Data Connection: Data connection that is configured in the SQL script **hive_sql** is selected by default. The value can be changed.
- Database: Database that is configured in the SQL script **hive_sql** 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 the job orchestration is complete, click to test the job.

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.

 Copyright © Huawei Technologies Co., Ltd.

Figure 8-64 Configuring the scheduling mode



Note:

- 2018/10/11 to 2018/11/10: The job is executed at 02:00 a.m. every day.

Step 5 Click to save the job and click to schedule the job to enable the job to run automatically every day.

----End

8.10.3 Cross-Workspace Job Scheduling

Applicable Scenario

If you have assigned permissions based on workspaces, users in different workspaces can only perform operations on jobs in their own workspaces. However, jobs in different workspaces depend on each other. Therefore, they need to be scheduled cross workspaces. Refer to this section for cross-workspace job scheduling details.

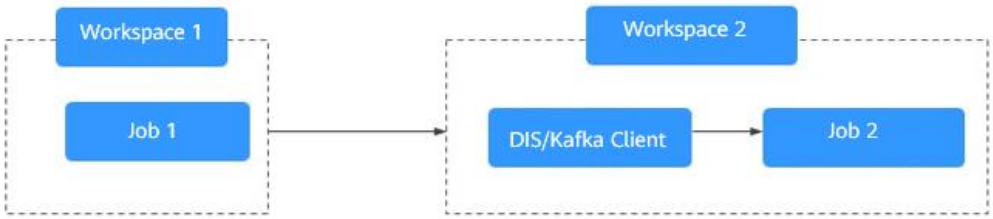
Solution Overview

The DGC Data Development module supports job running triggered by events. Therefore, DIS or MRS Kafka can be used as the job dependency to implement cross-workspace job scheduling.

As shown in the following figure, after Job 1 in workspace 1 is executed, you can use the DIS Client or Kafka Client to send a message to trigger Job 2; Configure event-triggered scheduling for Job 2, that is, trigger job 2 based on the messages sent by the DIS Client or Kafka Client.

 Copyright © Huawei Technologies Co., Ltd.

Figure 8-65 Scheduling solution



Configuration Method

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

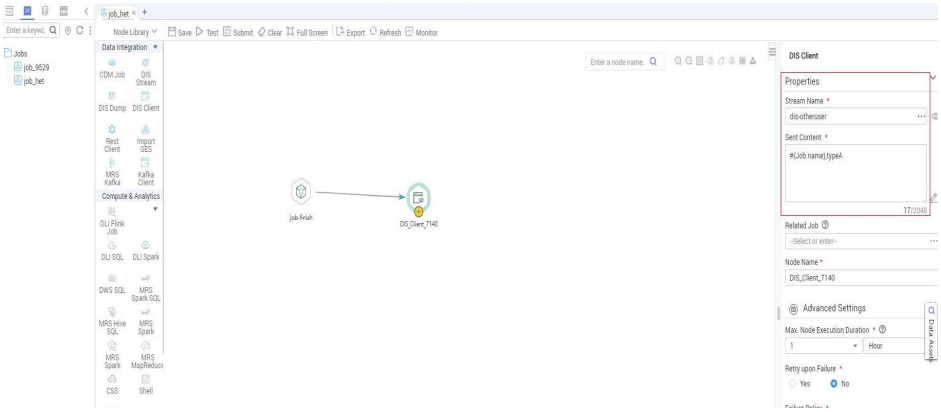
2. Click **Data Development** of the corresponding workspace.

   The **Data Development** page is displayed.

3. Create jobs in two different workspaces. For example, create a job named **job_het** in one workspace and create a job named **job_het1** in another workspace.

   After the job in one workspace is executed, use the DIS Client to send a trigger message. Send a single character string message to trigger the job in another workspace.

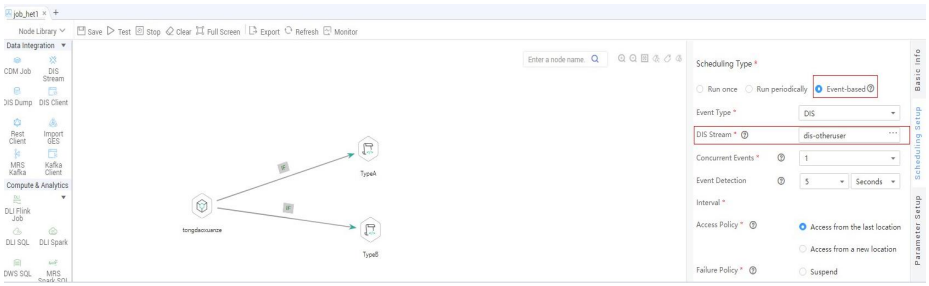
   Figure 8-66 Creating a job

4. Configure the **job_het1** job in another workspace. **Figure 8-67** shows the configuration node. An empty node is selected for the stream. **TYPEA** and **TYPEB** are Subjob nodes.

 Copyright © Huawei Technologies Co., Ltd.

Figure 8-67 Creating a job



a. Move the pointer to the empty node icon, and drag the icon to connect the node task to another node task.

b. Right-click a line and choose **Set Condition** from the shortcut menu. In the **Edit EL Expression** dialog box, enter the IF condition.

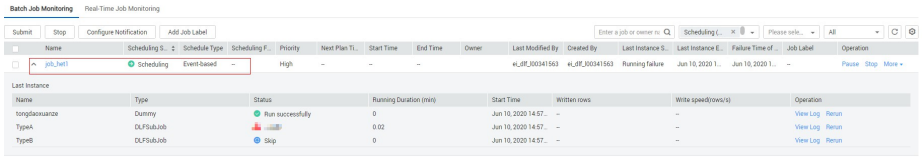
c. The IF condition is configured as follows:

```
#{StringUtil.equals(StringUtil.split(Job.eventData,',')[1],'typeA')}
#{StringUtil.equals(StringUtil.split(Job.eventData,',')[1],'typeB')}
```

> **NOTE**
>
> When the trigger message sent in **3** is "workspace A, type A", the **TYPEA** node is executed. If the message is "workspace A, type B", the **TYPEB** node is executed.

d. After the configuration is complete, start scheduling job in workspace B. After the job in workspace A is tested, the job in space B is triggered.

Figure 8-68 Viewing results



9 Data Quality Control

This module provides functions such as intelligent data planning, data quality monitoring, and data standardization, manages data models and standards in a unified manner, and implements data models and standards to improve data quality and effectively support decision-making.

9.1 Monitoring Business Metrics

9.1.1 Creating Metrics

You can manage all business metrics, including the metric sources and definitions. Business metrics are stored in directories.

Note that metrics in DQC are independent of the business and technical metrics in Data Design.

Entry

To create a metric, you must select a directory first. For details, see **Figure 9-1**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 9-1 Directory that stores the metric to be created

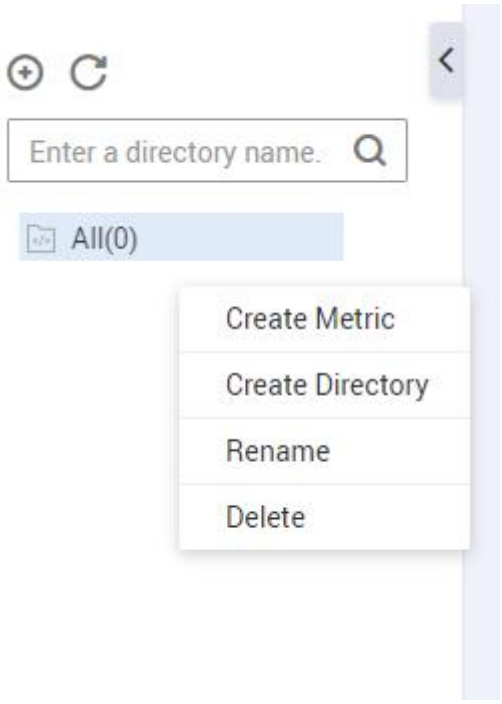


Table 9-1 Buttons in navigation bar

No.	Description
1	Create Directory
2	Refresh Directory
3	Select **All**. Right-click to create metrics, as well as to create, delete, and rename directories.

Creating a Metric

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Quality Control** of the corresponding workspace.

   The system navigates to the Data Quality Control console.

3. Choose **Metric Management** from the left navigation bar.
4. Click **Create**. In the dialog box that is displayed, set parameters described in **Table 9-2**.

 Copyright © Huawei Technologies Co., Ltd.

Table 9-2 Parameters

Parameter	Description
Metric Name	The name of a metric, which contains 1 to 64 characters and consists of only letters, digits, and underscores (_).
Data Connection	Select a created data connection from the drop-down list. **NOTE** • Currently, only DWS, PostgreSQL, MRS Hive, DLI, and MySQL are supported. • Metrics are closely connected based on data connections. Therefore, you need to establish data connections in the metadata management module before creating metrics.
Database/ Queue	Select the database where the metric runs. **NOTE** When DLI is selected as the data connection, a running queue is required.
Metric Description	Information to better identify a metric. It cannot exceed 256 characters.
Metric Directory	Directory for storing metrics. You can select a created directory. **Figure 9-1** describes how to create a directory.
Metric Type	**Custom** is supported. You can customize a SQL statement, to define the metric source.

9.1.2 Creating Rules

You can manage all business rules that define relationships between metrics or between metrics and values. Business rules are stored in directories.

Entry

To create a business rule using metrics, you must add a directory that stores the rule to be created. For details, see **Figure 9-2**.

Copyright © Huawei Technologies Co., Ltd.

Figure 9-2 Directory that stores the rule to be created

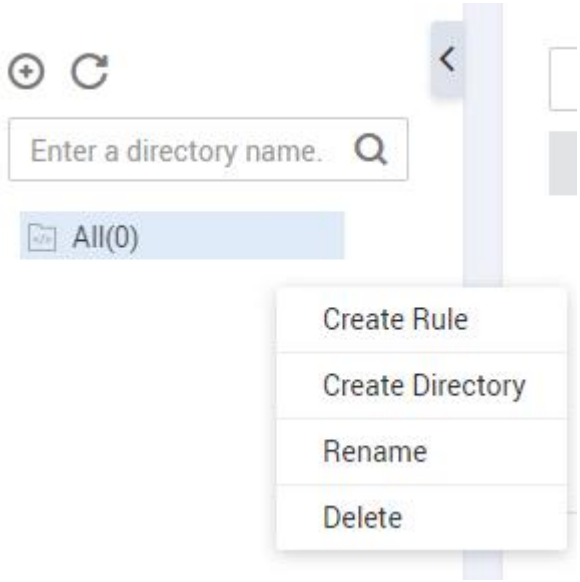


Table 9-3 Buttons in navigation bar

No.	Description
1	Create Directory
2	Refresh Directory
3	Select **All**. Right-click to create rules, as well as create, delete, and rename directories.

Creating a Rule

1. **Log in to the DGC console**, select an instance, and click **Access** on it.
   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Quality Control** of the corresponding workspace.
   The system navigates to the Data Quality Control console.
3. Choose **Rule Management** from the left navigation bar.
4. Click **Create**. In the dialog box that is displayed, set parameters described in **Table 9-4**.

Table 9-4 Parameters

Parameter	Description
Rule Name	The name of a rule, which contains 1 to 64 characters and consists of only letters, digits, and underscores (_).
Description	Information to better identify a rule. It cannot exceed 256 characters.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Directory	The directory that stores rules. You can select a created directory. Figure 9-2 describes how to create a directory.
Define Relationship	A logical expression between metrics. The relationship can contain arithmetic operations. **NOTE** Simple four arithmetic operations are supported.

9.1.3 Creating Business Scenarios

You can manage all business scenarios that define the logical relationships between rules. Business scenarios are stored in directories.

Entry

To create a business scenario using rules, you must add a directory that stores the business scenario to be created. For details, see Figure 9-3.

Figure 9-3 Directory that stores a business scenario

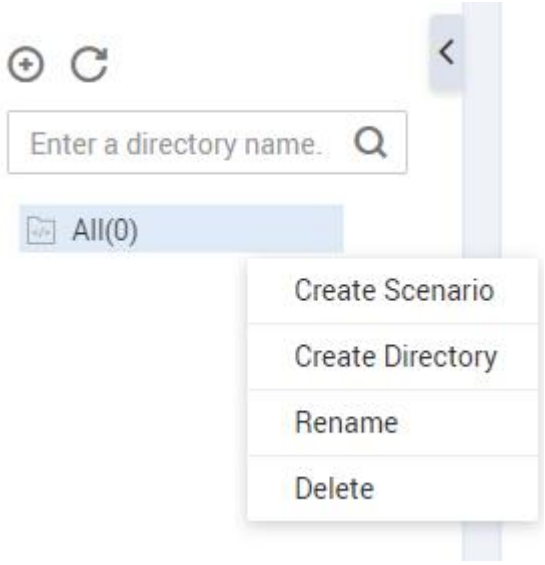


Table 9-5 Buttons in navigation bar

No.	Description
1	Create Directory
2	Refresh Directory
3	Select **All**. Right-click to create business scenarios, as well as to create, delete, and rename directories.

Copyright © Huawei Technologies Co., Ltd.

Creating a Business Scenario

1. Log in to the DGC console, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Quality Control** of the corresponding workspace.

   The system navigates to the Data Quality Control console.

3. Choose **Business Scenario Management** from the left navigation bar.
4. Click **Create**. In the dialog box that is displayed, set parameters described in Table 9-6.

Table 9-6 Parameters

Parameter	Description
Basic Configuration	
Scenario Name	The name of a business scenario, which contains 1 to 64 characters and consists of letters, numbers, and underscores (_).
Description	Information to better identify a business scenario. It cannot exceed 256 characters.
Select Directory	The directory that stores business scenarios. You can select a created directory. Figure 9-3 describes how to create a directory.
Business Level	Currently, 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 **Add** to add multiple rules.
Subscription Configuration	
Notification	Set this to or , respectively, to disable or enable the notification function.
Notification Type	The options are as follows: • **Trigger alarms** • **Run successfully**
Topic	Select a message notification topic.

 Copyright © Huawei Technologies Co., Ltd.

5. Click **Next** to go to a 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 9-7**.

Table 9-7 Parameters

Parameter	Description
Scheduling Date	The period during which a scheduling task takes effect
Scheduling Cycle	The frequency at which a scheduling task is executed. Related parameters are: • Minute • Hour • Day • Week
Start scheduling	If this check box is selected, the task is scheduled immediately.

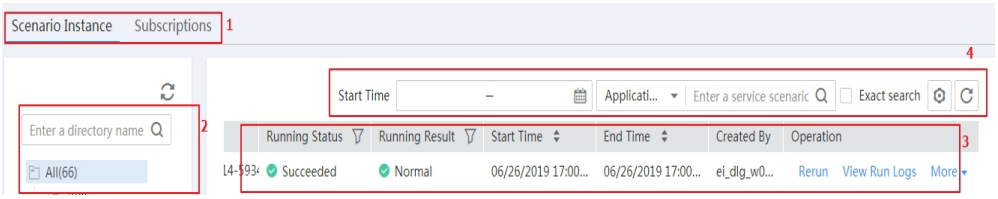
9.1.4 O&M

You can manage all business scenarios, view metric running statuses, query run logs, and handle issues on the **O&M Management** page.

GUI Description

This part describes the areas and buttons on the **O&M Management** page.

Figure 9-4 O&M Management page



 Copyright © Huawei Technologies Co., Ltd.

Table 9-8 Entry

No.	Area	Description
1	Menu bar	The menu bar on the **O&M Management** page includes **Scenario Instances** and **Subscriptions**. • The **Scenario Instances** tab lists all scenario instances created by the current account. • The **Subscriptions** tab lists all business scenarios that are subscribed to by the current user. In the **Subscriptions** tab, **Notification Status** is added. **Notification Status** indicates whether the running result of the business instance is successfully subscribed to, for example, sending an alarm email.
2	Navigation bar	The navigation bar on the left contains the directories that store business scenario instances. You can store business scenarios in different directories. The number next to each directory indicates the number of business scenarios stored in that directory.
3	List of business instances	Displays the instance name, running status, and running result.
4	Search area	• Displays business instances selectively. For example, you can only display business scenario instances for a specified time range. • Displays a list of business scenario instances according to the handler, creator, or instance name. Fuzzy search is supported.

Table 9-9 Business scenario parameters

Parameter	Description
Running Result	Displays whether the instance is running properly. • **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. • **End unexpectedly**: The instance stops unexpectedly. • --: The instance is running, but no running result is displayed.
Rerun	Allows you to run a business scenario instance again.
View Log	Allows you to view the running details of a scenario instance.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
More > Resolve Issue	Allows you to perform further processing on the current business scenario instance. For example, fill in the **Provide handling suggestions**, **Close the issue**, or **Transfer to others**. The above operations can be performed only on the condition that the handler of the instance is the current login user.
More > View Processing Log	Allows you to view historical processing records.
	Indicates that the displayed content can be filtered by the column properties.
	Indicates that the list content can be displayed in the order of the column properties.
	Allows you to select properties of scenarios to be displayed.

9.2 Monitoring Data Quality

9.2.1 Purpose

Data Quality Control (DQC) is a kind of data quality management tool used to manage quality of the data in databases and analyze the data from the perspectives of completeness, validity, timeliness, consistency, accuracy, and uniqueness in single-column, cross-column, cross-row, and cross-table mode. It is also used for data cleansing and standardization, automatic generation of cleansing and standardization rules based on data standards, and periodic monitoring and cleansing.

NOTE

Dataset is the target of DQC. Currently, offline data can be monitored. When the offline data changes, DQC verifies the data and blocks the production link to avoid the spread of the problem data. In addition, DQC manages historical verification results so that you can analyze and grade data quality.

The following table describes modules under **Data Monitoring**:

 Copyright © Huawei Technologies Co., Ltd.

Module	Description
Dashboard	The dashboard is the homepage that displays the alarm and blocking information of data tables. To be more specific, the following information is included: • Number of jobs, instances, and anomaly tables; distributions and changes of instance running status in a selected period. • Statistics about alarm classifications and data table alarms of the current day, as well as the alarm trend and rule quantity of the latest seven days are displayed.
Rule Template	Rule template is vital for data quality control. The **Rule Template** page is the entry for configuring rules. Users can manage the configured rules (template rules and self-service rules).
Quality Job	Users can create quality jobs to apply the created rules to existing tables.
Comparison Job	Users can create comparison jobs to apply the created rules to two existing tables to monitor and control their data and output the comparison results.
O&M Management	O&M management is used to view the running status of rules and handle O&M problems. The **Subscriptions** page displays the running status of all subscribed tasks.
Quality Report	The system automatically generates a quality report based on the job execution result.

9.2.2 Creating Rule Templates

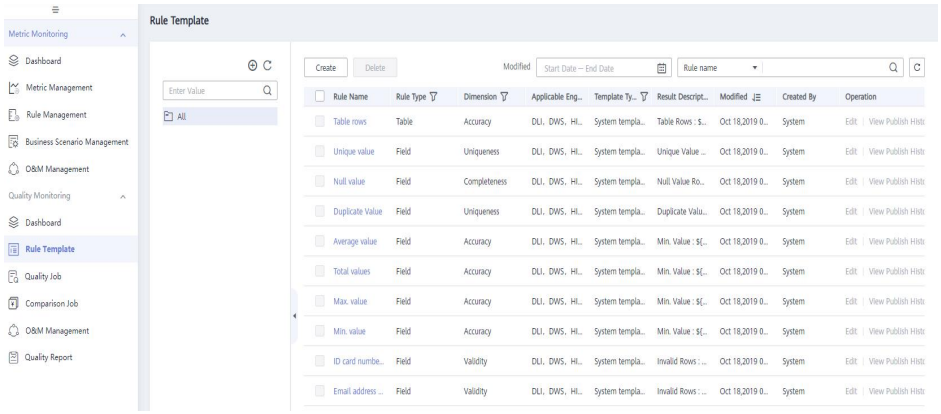
Creating rules is the core functional module of Data Quality Control (DQC). Offline data monitoring is supported. Therefore, you can create rules based on the actual requirements. You can create a rule template or customize a rule based on the requirements.

There are 20 built-in rule templates, including database-level rule templates, table-level rule templates, field-level rule templates, and cross-field rule templates. If these built-in rule templates cannot meet your requirements, you can customize a rule template. See the description below for details.

Step 1 Choose **Quality Monitoring** > **Rule Template** and click **Create**. On the **Create Rule Template** page, configure the required parameters.

 Copyright © Huawei Technologies Co., Ltd.

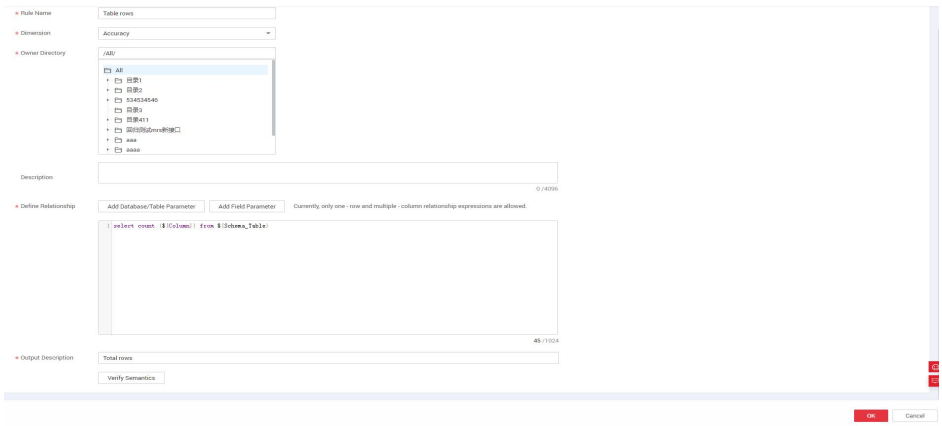
Figure 9-5 Creating rule templates



Step 2 In the displayed dialog box, enter the rule name, select the rule matching dimension, define the SQL template, describe the output result, and click **OK**.

- **Dimension**: You can complete single-column, cross-column, cross-row, and cross-table analysis from six dimensions: integrity, validity, timeliness, consistency, accuracy, and uniqueness. When customizing a quality rule, select a dimension for rule matching.
- **Define Relationship**: Enter SQL statements to search for data.
  - 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 **${Schema_Table1}** is generated by clicking **Add Database Table Parameter**.
- **Output Description**: describes each column in the SQL result. Column descriptions are separated with commas (,).
  - Example: If the relationship is set to **select max (${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 9-6 Configuring a rule template



----End

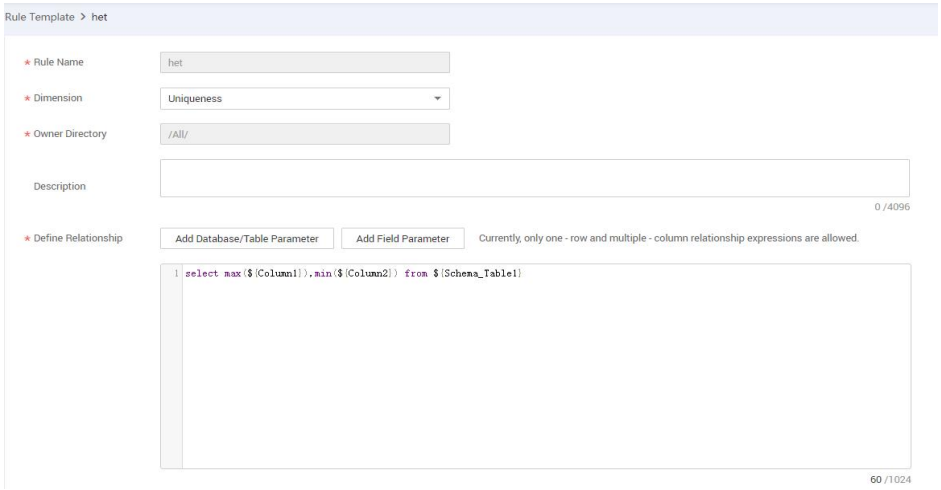
 Copyright © Huawei Technologies Co., Ltd.

Managing a Rule Template

You can modify and view the publication history of a custom rule template. To modify a custom rule template, do as follows:

Step 1 On the DQC homepage, select **Rule Template** from the left navigation bar. Locate the target rule template in the displayed list and click **Edit** in the **Operation** column.

Figure 9-7 Editing a rule template

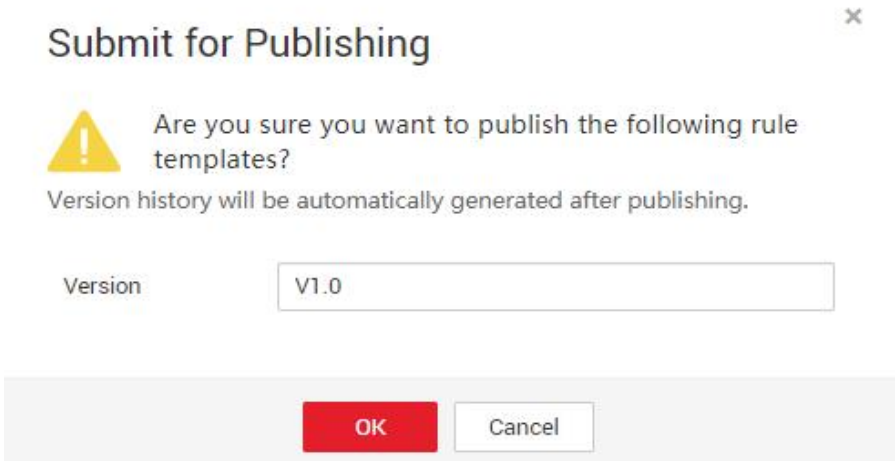


Step 2 Dimensions can be modified, output description can be modified, and relationships can be redefined.

Step 3 Click **Save** to save the settings.

Step 4 Click **Publish**. In the displayed dialog box, set the version. Save the new rule template and upgrade it.

Figure 9-8 Publishing the new rule template



Step 5 Click **OK**. Go back to the rule template list and click **View Publish History** in the **Operation** column. View the publication history and compare the versions to display the modifications.

 Copyright © Huawei Technologies Co., Ltd.

Figure 9-9 Version comparison

----End

9.2.3 Creating Data Quality Jobs

Entry

To create a quality job by a data connection, you must select a directory to store the data quality job. For details, see **Figure 9-10**.

Figure 9-10 Directory that stores the quality job to be created

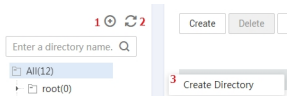


Table 9-10 Buttons in navigation bar

No.	Description
1	Create Directory
2	Refresh Directory
3	Select **All**. You can right-click to create, delete, and rename directories.

Process

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Quality Control** of the corresponding workspace.

   The system navigates to the Data Quality Control console.

3. Choose **Quality Monitoring** > **Quality Job**.
4. Click **Create**. In the dialog box that is displayed, set related parameters. See **Table 9-11** for details.

 Copyright © Huawei Technologies Co., Ltd.

Table 9-11 Quality job parameters

Parameter	Description
Name	The name of the data quality job, which contains 1 to 64 characters and consists of letters, numbers, and underscores (_).
Description	Information to better identify a data quality job. It cannot exceed 256 characters.
Select Directory	Directory for storing data quality jobs. You can select a created directory. **Figure 9-10** describes how to create a directory.
Job Level	Currently, the options are **Warning**, **Minor**, **Major**, and **Critical**. The job level determines the template for sending notification messages.

5. Click **Next**. Configure a data quality rule. See **Table 9-12** for details. You can also add a monitoring rule to create a sub-rule. Then, click **Batch Save**. All created rules will be displayed in the created tables.

Table 9-12 Parameters for configuring a rule template

Parameter	Sub-parameter	Description
Basic Info	Rule Type	Database rules. table rules, field rules. cross-field rules, and custom filed rules are rules configured for specific fields in a table. Set this parameter to a table-level rule. Other configuration items on the page are set to table-level rule configuration items.
Object	Data Connection Type	Currently, only DWS, MRS Hive, DLI, RDS (MySQL and PostgreSQL), and ORACLE are supported.
	Data Connection	Select a created data connection from the drop-down list. **NOTE** Rules are based on data connections. Therefore, you need to create data connections in **Management Center** before creating data quality rules.
	Database	Select a database to which the configured data quality rules are applied. **NOTE** • 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.

Copyright © Huawei Technologies Co., Ltd.

Parameter	Sub-parameter	Description
	Data Table	Select the table to which the configured data quality rules apply. **NOTE** • The data table is closely related to the database. • When **Rule Type** is set to **Table rule**, set the data object to the corresponding data table.
	Select Field	Field in the data table to which the data quality rule applies. **NOTE** When **Rule Type** is set to **Field rule**, set the data object to the field in the corresponding data table.
	Reference Data Object	This parameter is mandatory when **Rule Type** is set to **Cross-field rule**. Set this parameter to the data field to be referenced.
Rule Template	Template Type	Select a built-in or custom rule template. **NOTE** The template type is closely related to the rule type. • If **Rule Type** is set to **Database rule**, the built-in template type is **Null value scanning in databases**. • If **Rule Type** is set to **Table rule**, the built-in template types are **Table rows** and **Null value scan in tables**. • If **Rule Type** is set to **Field rule**, the built-in template types are **Phone number format verification**, **IP address verification**, and **Email address verification**. • If **Rule Type** is set to **Cross-field rule**, the built-in template type is **Field value consistency verification**. • In addition to built-in rule templates, you can also associate custom rule templates. For details on how to customize a rule template, see **Creating Rule Templates**.
Reference Data Object	Reference Data Object	Reference data needs to be configured for some rule templates, for example, specifying standard values. This parameter is valid only when **Rule Type** is set to **Cross-field rule.**
Custom SQL	SQL	Enter a complete SQL statement to define how to monitor the data quality of data objects.
Alarm Condition	Alarm Expression	Replace the **Output Description** variable defined in the rule template with the actual value and use the calculated result as the alarm condition.
	Parameter	Set the alarm parameter of the alarm condition. The parameter value comes from the output of the rule template.

Copyright © Huawei Technologies Co., Ltd.

Parameter	Sub-parameter	Description
	Logical Operator	Logical calculation is performed on multiple alarm parameters to form more complex alarm conditions.
	Generate Anomaly Data	Enable the **Generate Anomaly Data** function and click **Select** next to **Anomaly Table** to store the anomaly data that does not comply with the preset rules. **NOTE** • The custom rule template does not support the generation of anomaly data. • 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.
	Output Settings	• **Output Rule Settings**: If you select this option, the quality job settings will show up in the anomaly tables so that users 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	Users can select to export all anomaly data or the specified amount of anomaly data.
	View Rule	Click this button and you can: • Determine the rule repetition based on tables and fields. • View the related sub-rules and quality jobs that already exist.
Compute Engine	Cluster type	Select the engine for running the quality job.
Object Scope	Scanning Scope	Determine the scope that the created rule is applied to. • If **All** is selected, all tables are scanned. • If **Partial** is selected, you need to enter a where condition expression to precisely locate the partitions to query data.

6. Click **Next** to set the **Alarm Configuration**. If the SMN notification needs to be received, enable **Notification**, and set **Notification Type** and **Topic**.

 Copyright © Huawei Technologies Co., Ltd.

7. Click **Next** to go to a 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 9-13**.

   NOTE

   1. If **Schedule once** is selected, a manual task instance is generated. A manual task has no dependency on scheduling and needs to be manually triggered.
   2. If **Schedule periodically** 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 management on scheduled instance tasks, such as viewing the running status, stopping and rerunning the scheduled tasks.

Table 9-13 Parameters

Parameter	Description
Effective Date	Effective date of a scheduling task
Scheduling Cycle	The frequency at which a scheduling task is executed. Related parameters are: • Minute • Hour • Day • Week **NOTE** • If **Scheduling Cycle** is set to **Minute** or **Hour**, you need to set **Start Time**, **Time Interval**, and **End Time** for the scheduling task. Currently, the start time is in minute for stagger scheduling. • If **Scheduling Cycle** is set to **Day**, you need to set **Specific Time** for the scheduling task, namely, a specified time when the scheduling task is enabled every day. • If **Scheduling Cycle** is set to Week, you need to set **Scheduling Time** and **Specific Time** for the scheduling task, that is, *XX* o'clock *XX* minutes on *XXX* every week.

9.2.4 Create Comparison Jobs

Prerequisites

To create a comparison job by a data connection, you must select a directory to store the comparison job. For details, see **Figure 9-11**.

For details on directory-related operations, see **Table 9-14**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 9-11 Directory that stores the comparison job to be created

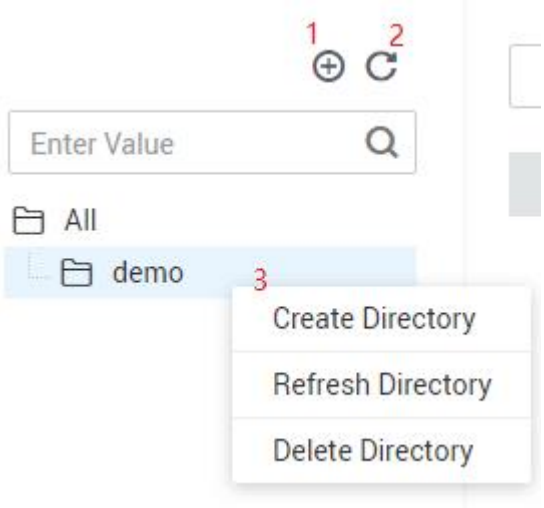


Table 9-14 Buttons in navigation bar

No.	Description
1	Create Directory
2	Refresh Directory
3	Select **All**. You can right-click to create, delete, and rename directories.

Create a Job

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Quality Control** of the corresponding workspace.

   The system navigates to the Data Quality Control console.
3. Choose **Quality Monitoring** > **Comparison Job** from the left navigation bar.
4. Click **Create**. In the dialog box that is displayed, set related parameters. See **Table 9-15** for details.

   Table 9-15 Parameters for creating a comparison job

Parameter	Description
Job Name	The name of a comparison job, which can contain 1 to 64 characters and consist of letters, numbers, and underscores (_).
Description	Information to better identify a comparison job. It cannot exceed 256 characters.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Directory	The directory for storing the comparison job to be created. You can select a created directory. **Figure 9-11** describes how to create a directory.
Job Level	Currently, the options are **Warning**, **Minor**, **Major**, and **Critical**. The job level determines the template for sending notification messages.

5. Click **Next**. Configure a comparison rule. See **Table 9-16** for details. You can also add a monitoring rule to create a sub-rule.

Table 9-16 Parameters for configuring a rule template

Module	Parameter	Description
Configure Source	Rule Type	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**. Then, other configuration items on the page are set to table-level rule configuration items correspondingly. The rule type of the destination object is automatically generated based on that of the source object.
	Data Connection	Source and destination objects support the following data connection types: DWS, MRS Hive, DLI, ORACLE, and RDS (MySQL and PostgreSQL). Select a created data connection from the drop-down list. **NOTE** Rules are based on data connections. Therefore, you need to create data connections in Management Center before creating data quality rules.
	Mapping	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. **NOTE** The data table is closely related to the database. The database is tailored to the created data connection.

 Copyright © Huawei Technologies Co., Ltd.

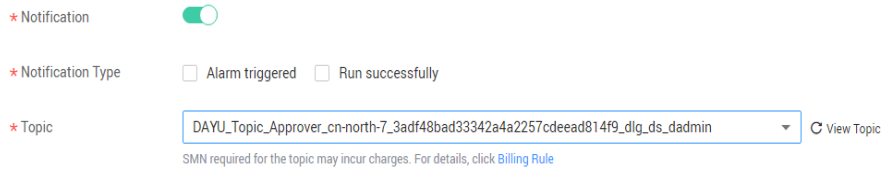
Module	Parameter	Description
Rule Template	Template	This parameter defines how to monitor the data quality of data objects. The template name of the source object contains the built-in rule template and custom rule template. The template name of the destination object is automatically generated based on the rule type of the source object. **NOTE** The template type is closely related to the rule type. This parameter is displayed only when **Rule Type** is set to **Table rule** or **Field rule**. • If **Rule Type** is set to **Table rule**, the built-in template types are **Table rows** and **Null value scan in tables**. • When **Rule Type** is set to **Field rule**, the system has multiple built-in template types, such as **Unique field value**, **Duplicate field value**, **Integrity**, **Accuracy**, and **Validity**. • In addition to built-in rule templates, you can also associate custom rule templates. For details on how to customize a rule template, see **Creating Rule Templates**.
Compute Engine	Cluster Name	Select the engine for running the comparison job.
Object Scope	Scanning Scope	Specify the scope that the created rule is applied to. • If **All** is selected, all tables are scanned. • If **Partial** is selected, you need to enter a where condition expression to precisely locate the partitions to query data.
Alarm Condition	Alarm Expression	Replace the **Output Description** variable defined in the rule template with the actual value and use the calculated result as the alarm condition. • For example, the number of table rows is greater than 100. • The supported logical operators including && (and) and || (or). For example, **table rows > 0 && unique values = 1** indicate that the number of table rows is greater than 0 and the number of unique values is 1; **table rows > 0 || unique values = 1** indicates that the number of table rows is greater than 0 or the number of unique values is 1.

Copyright © Huawei Technologies Co., Ltd.

Module	Parameter	Description
	Parameter	Set the alarm parameter of the alarm condition. The parameter value comes from the output of the rule template.
	Logical Operator	Logical calculation is performed on multiple alarm parameters to form more complex alarm conditions.

6. Click **Next** to set the subscription configuration. If the SMN notification needs to be received, enable **Notification**, and set **Notification Type** and **Topic**. See **Figure 9-12**.

Figure 9-12 Subscription configuration



7. Click **Next** to go to a 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 9-17**.

NOTE

1. If **Schedule once** is selected, a manual task instance is generated. A manual task has no dependency on scheduling and needs to be manually triggered.
2. If **Schedule periodically** 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 management on scheduled instance tasks, such as viewing the running status, stopping and rerunning the scheduled tasks.

Table 9-17 Parameters for setting the scheduling mode

Parameter	Description
Effective Date	Effective date of a scheduling task

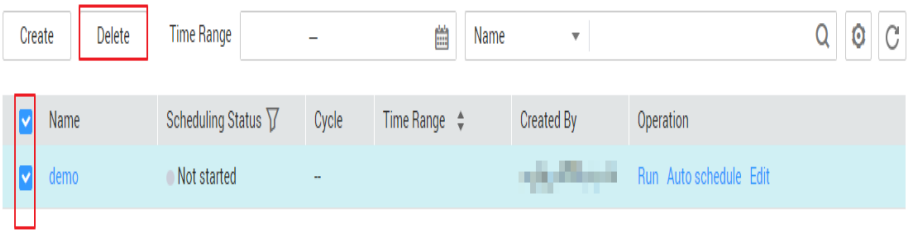
 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Scheduling Cycle	The frequency at which a scheduling task is executed. Related parameters are: • Minute • Hour • Day • Week **NOTE** • If **Scheduling Cycle** is set to **Minute** or **Hour**, you need to set **Start Time**, **Time Interval**, and **End Time** for the scheduling task. • If **Scheduling Cycle** is set to **Day**, you need to set **Specific Time** for the scheduling task, namely, a specified time when the scheduling task is enabled every day. • If **Scheduling Cycle** is set to Week, you need to set **Scheduling Time** and **Specific Time** for the scheduling task, that is, *XX* o'clock *XX* minutes on *XXX* every week.

Deleting a Job

1. **Log in to the DGC console**, select an instance, and click **Access** on it.
   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Quality Control** of the corresponding workspace.
   The system navigates to the Data Quality Control console.
3. Choose **Quality Monitoring** > **Comparison Job** from the left navigation bar.
4. Select the jobs you want to delete and click **Delete** in the upper part of the page. See **Figure 9-13**.

Figure 9-13 Deleting jobs



NOTE

Deleted comparison jobs cannot be recovered. Exercise caution when performing this operation.

Running/Starting a Scheduling Job

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

 Copyright © Huawei Technologies Co., Ltd.

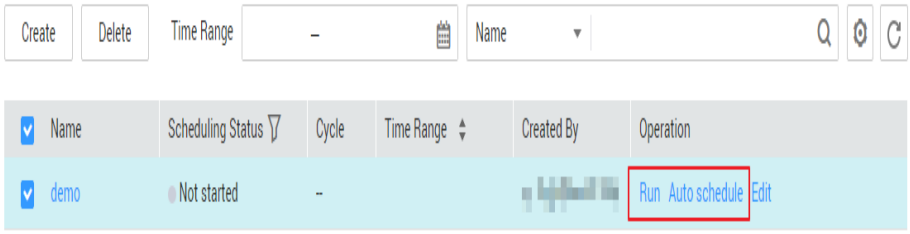
On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Quality Control** of the corresponding workspace.

   The system navigates to the Data Quality Control console.

3. Choose **Quality Monitoring** > **Comparison Job** from the left navigation bar.
4. Run or start the corresponding scheduling tasks. For details, see **Figure 9-14**.

Figure 9-14 Running/Starting scheduling tasks



9.2.5 Viewing Rule Instances

GUI Description

This topic describes the areas and buttons on the **O&M Management** page.

Figure 9-15 Description of areas on the **SQL Editor** page

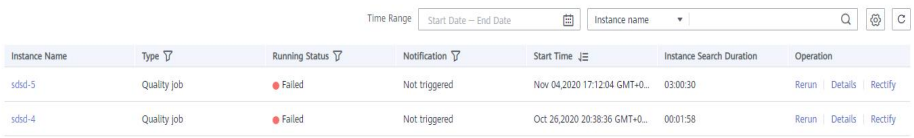


Table 9-18 Areas

No.	Area	Description
1	Navigation bar	Contains the storage directory of data quality rules. You can store rules in different directories tailored to the site 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, running status, and running result.
3	Search area	• Rule instances can be displayed selectively. For example, you can display rule instances for a specified time range. • The system displays a list of instances according to the handler, creator, or instance name. Fuzzy search is supported.

 Copyright © Huawei Technologies Co., Ltd.

Table 9-19 List of rule instances

Parameter	Description
Instance Name	Consists of a rule name and a digit. The larger the digit is, the later the instance is created.
Running Status	Displays the running status of an instance, such as **Successful**, **Failed**, **Running**, and **Alarming**. In the right pane, you can view the detailed run logs of the rule instance. • **Successful**: The instance stops normally and the running result meets the expectation. • **Failed**: The instance stops normally, but the running result does not meet the expectation. • **Alarming**: The instance stops unexpectedly. • **Running**: The instance is running, but no running result is displayed.
Rerun	Allows users to run a rule instance again.
Processing Record	Allows users to perform further processing on the current rule instance. For example, fill in the **Provide handling suggestions**, **Close the issue**, or **Transfer to others**. The operations above can be performed on condition that the handler of the instance is the current login user.
Details	Displays the running results and logs of job instances.
	Indicates that the displayed content can be filtered by the column properties.
	Indicates that the list content can be displayed in the order of the column properties.
	Allows you to select properties of scenarios to be displayed.
	Allows users to view the details (calculation results included) about all sub-rules of a rule instance.

 Copyright © Huawei Technologies Co., Ltd.

10 Data Assets

This module provides enterprise-class metadata management to clarify information assets. A data map is available to display a data lineage and view of the data assets for intelligent data search and monitoring.

10.1 Data Maps

10.1.1 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.

10.1.2 Dashboard

The **Dashboard** page displays data assets and asset reports. You can query information such as the assets quantity and size, and 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**.

 Copyright © Huawei Technologies Co., Ltd.

Procedure

1. Log in to the DGC console, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Assets** of the corresponding workspace to navigate to the Data Assets console.
3. Choose **Dashboard** from the left navigation bar.
4. Click **Business Assets** to display the number and details of business objects, logical entities, and business attributes.
5. Click **Technical Assets** to display the number and details of databases and tables, and data volumes.
6. Click **Metric Assets** to display the metrics and their details.

10.1.3 Data Catalogs

Data Assets provides data catalogs for users 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. Log in to the DGC console, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Assets** of the corresponding workspace to navigate to the Data Assets console.
3. Choose **Data Catalog** from the left navigation bar.
4. Enter a keyword in the search text box, and click the search icon. The search results are listed below the search text box. You can search assets in either of the following ways:

   Search for assets by their names and description.

   Search for assets by their attributes (attributes 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 criteria as follows:

- Data connection: the connection that your target asset uses.
- Type: the type of your target asset.

Copyright © Huawei Technologies Co., Ltd.

- 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.

Use type as an example to demonstrate how to filter an asset.

Step 1 Select **Table** under **Types**. Then, table assets will show up.

Step 2 Five asset types are displayed by default. Click **All** if you want to view data assets of all types.

----End

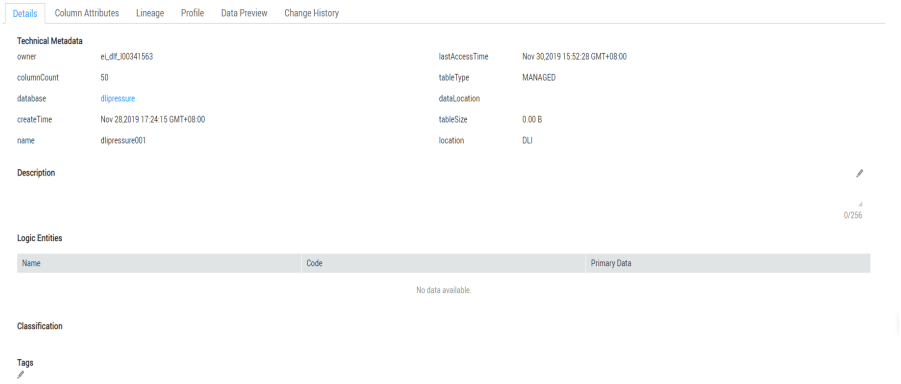
Viewing the Details of an Asset

Use 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 navigate to its details page.

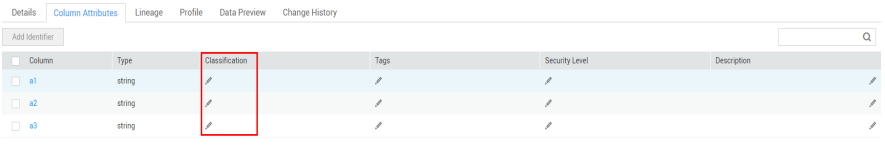
Step 2 On the **Details** tab page, you can view the basic attributes of the technical metadata; edit the description; and add or delete classifications, tags, and security levels for the table, table columns, or OBS objects.

Figure 10-1 Viewing asset details



Step 3 On the **Column Attributes** tab page, you can view the column attributes of the table; add or delete classifications, tags, and security levels for the data columns; and edit the description.

Figure 10-2 Managing column attributes



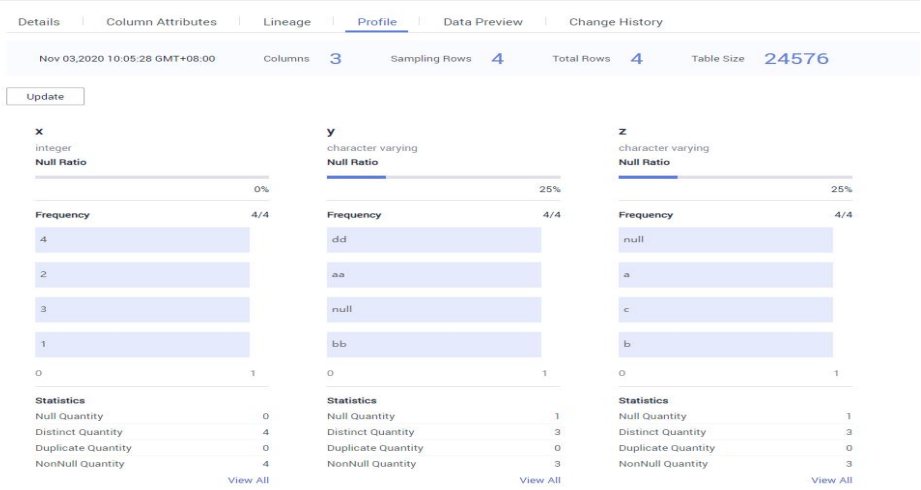
Step 4 On the **Lineage** tab page, you can view table lineages and impacts. For details on how to set a data lineage, see **Node Lineages**.

Step 5 On the **Profile** tab page, you can view the table profile.

Click **Update** to update the table profile.

 Copyright © Huawei Technologies Co., Ltd.

Figure 10-3 Viewing a table profile



Step 6 On the **Data Preview** tab page, you can view the effect after data is masked.

Step 7 On the **Change History** tab page, you can view the update history of the table.

----End

10.1.4 Tags

Data Assets provides tags for users to identify their business data. For example, a tag can be used to identify an ODS table or a DWD table. Tags often depict business activities and they are associated with technical assets. Therefore, technical assets can be managed with ease.

Managing a Tag

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Assets** of the corresponding workspace to navigate to the Data Assets console.
3. Choose **Tag Management** from the left navigation bar.
4. Click **Create** to create a tag.
   - **Tag Name**: Tag names can include only letters, numbers, and underscores (_). They cannot start with underscores (_) and exceed 100 characters.
   - **Description**: Up to 255 characters are allowed.
5. Click **Delete** to delete a tag.
6. Click **Edit** to modify the description of a tag.

Adding a Tag to Identify Data

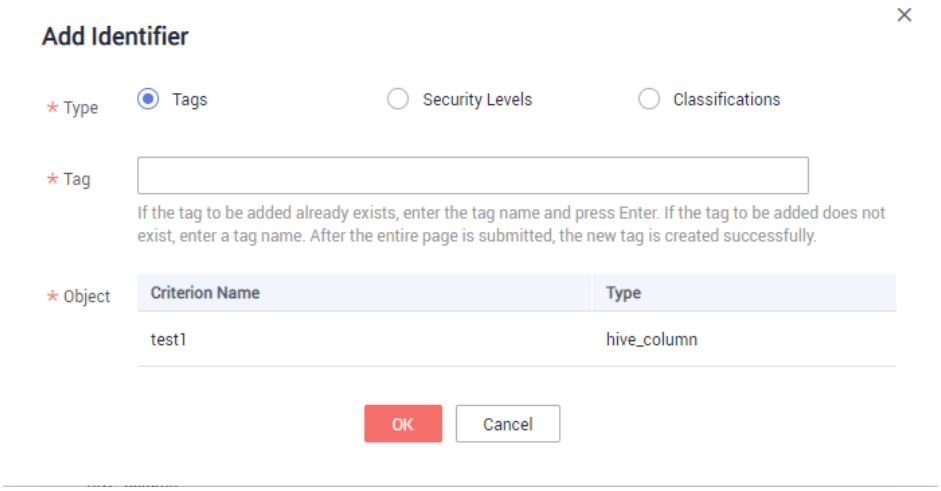
1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

 Copyright © Huawei Technologies Co., Ltd.

2. Click **Data Assets** of the corresponding workspace to navigate to the Data Assets console.
3. Choose **Data Catalog** from the left navigation bar.
4. Enter a keyword in the search text box, and click the search icon. The searched results are listed below the search box.
5. 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 10-4 Adding an identifier



6. Set parameters as prompt 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**.

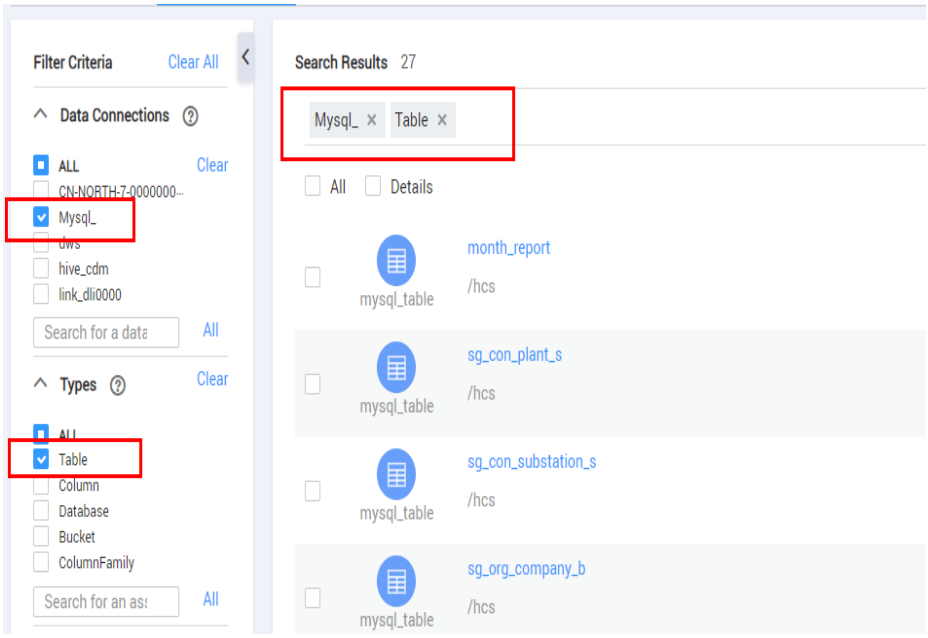
Viewing the Details of a Table

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Assets** of the corresponding workspace to navigate to the Data Assets console.
3. Choose **Data Catalog** from the left navigation bar. On the **Technical Assets** tab page, select a data connection and set **Type** to **Table**. Filter all tables in the data connection. For details, see **Figure 10-5**.

 Copyright © Huawei Technologies Co., Ltd.

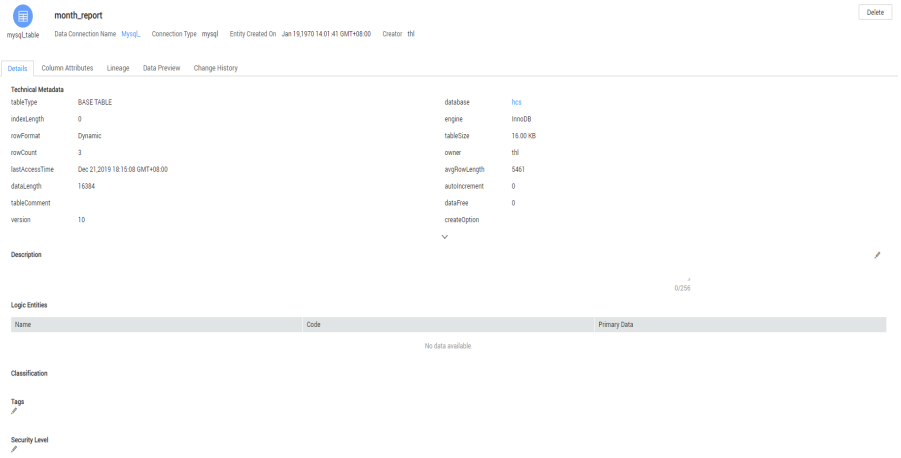
Figure 10-5 Filtering tables



4. Click a table name to go to its details page. See **Figure 10-6**.

   The displayed page displays permissions information, column attributes, lineage information, table profile, data preview, and change history.

Figure 10-6 Table details page



Previewing Data

You can preview the 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, Hive, and MySQL data connections can be previewed.
- The column classification information can be added manually or by creating data classification tasks. For data assets that use the MySQL data connection, you cannot add the column classification information by creating data classification tasks.

 Copyright © Huawei Technologies Co., Ltd.

10.2 Data Permissions

10.2.1 About Data Permissions

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.

10.2.2 Data Catalog Permissions

This section describes how to manage the data catalog permissions.

Constraints

- Only administrators can create, delete, and modify data catalog permission rules and set the permission effective status.
- Developers, O&M personnel, and visitors can only view data permissions.

Managing a Data Catalog Rule

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Assets** of the corresponding workspace to navigate to the Data Assets console.
3. Choose **Permissions** > **Data Catalog Permission**, click **Create**, and configure a data catalog permission rule.

   a. **Rule**: Name of a data permission rule.

   b. **Type**: Currently, only tags, confidentiality, and categories can be used for filtering.

   c. **Scope**: Select the actual tags, confidentiality, and categories.

 Copyright © Huawei Technologies Co., Ltd.

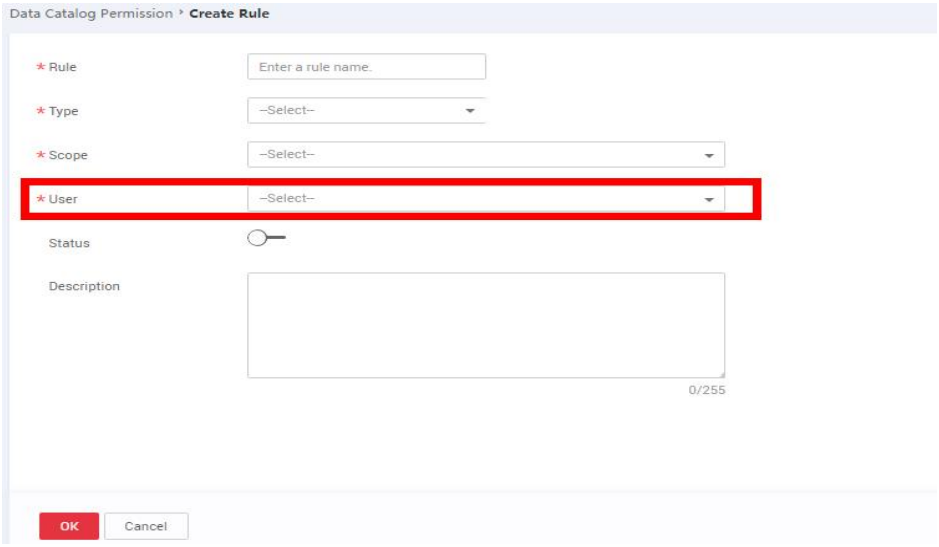
d. **User**: User who has the permissions to access the configured data catalog.

e. **Status**: If the status is enabled, the data catalog permission rule takes effect. Otherwise, the data catalog permission rule does not take effect.

NOTE

After a data catalog permission rule takes effect, only users assigned the data catalog permission 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 only the assets with tag **test** after the permission rule is enabled.

Figure 10-7 Creating a rule



4. In the data permission rule list, click **Edit** or **Delete** in the **Operation** column to modify or delete the rule.

10.2.3 Data Table Permissions

On the **My Permissions** tab page, you can view your table and column permissions in the workspace, and apply for or return the table and column permissions.

Administrators have the permissions to manage user permissions. That is, an administrator can view the resource permissions of all users in the workspace.

Applying for Table/Field Permissions

NOTE

- The current version supports permission 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. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

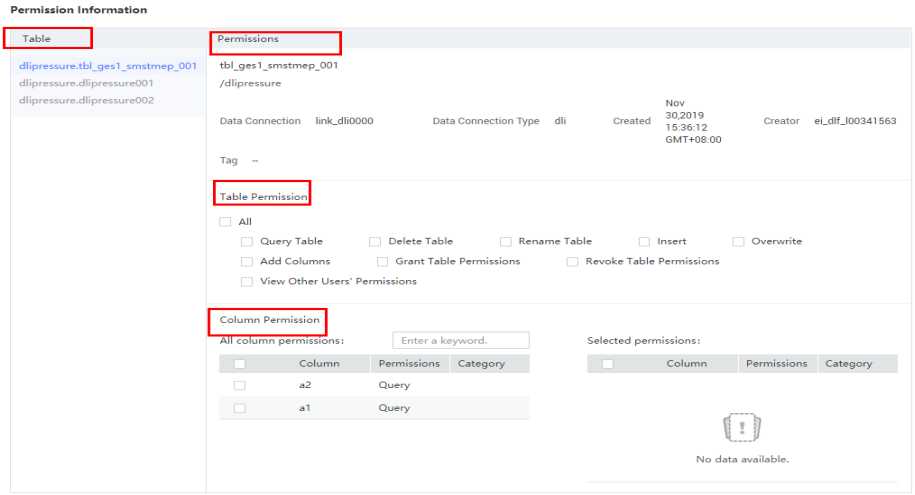
 Copyright © Huawei Technologies Co., Ltd.

2. Click **Data Assets** of the corresponding workspace to navigate to the Data Assets console.
3. Choose **Permissions** > **Data Table Permissions**. On the **My Permissions** tab page, click **Apply**.
4. On the displayed page, describe the scenario where the permissions are required, and select the data connection, database, and data table.
5. 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 on in the **Permission Information** area.

Figure 10-8 Table/column permissions



6. Click **OK**. Configure a reviewer and click **OK**.
7. The **My Applications** page is displayed. Wait for the reviewer to approve your application. After the application is approved, the permission takes effect.

Managing Existing Table Permissions

This section describes how to manage the table or field permissions you already have, including viewing, editing, and returning permissions.

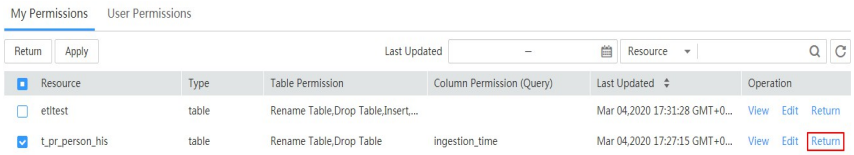
1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Assets** of the corresponding workspace to navigate to the Data Assets console.
3. Choose **Permissions** > **Data Table Permissions**. On the **My Permissions** tab page, select the check box before a permission name.
   - Click **View** under **Operation** to view the permission details.
   - Click **Edit** under **Operation** to modify data table permissions as needed.

 Copyright © Huawei Technologies Co., Ltd.

- Click **Return** under **Operation** to return data table permissions as needed.

Figure 10-9 Managing table permissions



Auditing User Permissions

On the **User Permissions** tab page, administrators can view the accounts that have permissions on tables and fields in the same workspace, reclaim the table and field permissions as needed, or grant permissions to users in batches.

NOTE

Only workspace administrators 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

  Choose **Data Table Permissions** > **User Permissions** to view the accounts with applied permissions in the same workspace.

Figure 10-10 Viewing accounts with table permissions

- Reclaiming users' asset permissions
  - Choose **Data Table Permissions** > **User Permissions** and click **Reclaim** under **Operation** to the right of the account to reclaim all its permissions.
  - Choose **Data Table Permissions** > **User Permissions**, select the check boxes to the left of one or more usernames, and click **Reclaim** in the upper left corner to revoke user asset permissions in batches.

Figure 10-11 Reclaiming users' asset permissions

- Granting permissions to users
  - Choose **Data Table Permissions** > **User Permissions**, select the check box to the right of an account, and click **Authorize** in the upper left corner to grant permissions.

 Copyright © Huawei Technologies Co., Ltd.

– Choose **Data Table Permissions** > **User Permissions**, select the check boxes to the left of one or more usernames, and click **Authorize** in the upper left corner to grant permissions in batches.

Figure 10-12 Authorization

- Managing user permissions

  Choose **Data Table Permissions** > **User Permissions**, and 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 10-13 Managing user permissions

10.2.4 Review Center

Constraints

Only administrators can manage reviewers, including creating and deleting reviewers.

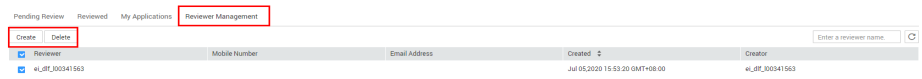
Approval Management

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**. On the **Reviewer Management** tab page, create and delete reviewers as required. See **Figure 10-14**. The reviewer data refers to the person added in the workspace.

Figure 10-14 Managing reviewers



- Pending review

  a. Choose **Permissions** > **Review Center** and click the **Pending Review** tab.

     On this page, you can view the applications that need to be approved.

  b. Click **Review** under **Operation** 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.

 Copyright © Huawei Technologies Co., Ltd.

- Reviewed applications
  a. Choose **Permissions** > **Review Center** 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** 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.

10.3 Data Security

10.3.1 About Data Security

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.

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 the database.
- Masking policies
  You can create data masking policies to protect data privacy.

10.3.2 Data Security Levels

This section describes how to manage data security levels, including creating and deleting security levels and adjusting their ranking sequences.

Accessing the Data Security Levels Page

1. **Log in to the DGC console**, select an instance, and click **Access** on it.
   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Assets** of the corresponding workspace to navigate to the Data Assets console.

 Copyright © Huawei Technologies Co., Ltd.

3. Choose **Data Security** > **Data Security Levels**. On the page that is displayed, you can create, delete, edit, move up or 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 **Data 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.

10.3.3 Data Classifications

This section describes how to create a data classification rule.

Creating a Classification Rule

1. Log in to the DGC console, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Assets** of the corresponding workspace to navigate to the Data Assets console.
3. Choose **Data Security** > **Data Classifications**. On the **Classification Rule** tab page, click **Create**.

   In the **Create Classification** page, set related parameters to create a classification rule. You can either create a rule based on the built-in template or customize one.

Figure 10-15 Creating a classification rule

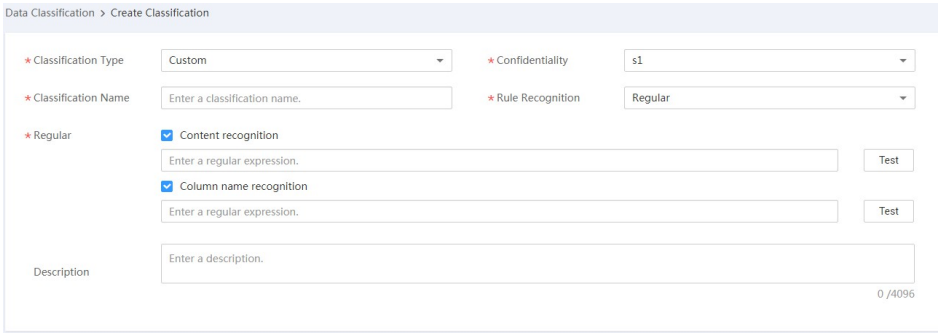


Table 10-1 Parameters for creating a classification rule

Parameter	Description
Classification Type	The category to which a rule belongs. You can either create a rule based on built-in templates or customize one.
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**.

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Classification Template	This parameter is available when **Classification Type** is set to **Built-in**. You can select a built-in sensitive data identification template based on the site requirements, for example, **Time**, **Mobile number**, and **License plate number**.
Classification Name	• If **Classification Type** is set to **Built-in**, the classification name is automatically generated based on the classification template. • If you select **Custom** for **Classification Type**, 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	• **Content recognition**: one of the data recognition methods. 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 classification rule to be created.

Creating a Group

1. Log in to the DGC console, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Assets** of the corresponding workspace to navigate to the Data Assets console.
3. Choose **Data Security** > **Data Classification**. On the **Groups** tab page, click **Create**.

   In the **Create Group** page, set related parameters as prompt.

   Set parameters by referring to **Table 10-2** and select classification rules in the list.

   The selected rules are displayed in the list on the right.

Table 10-2 Parameters for creating a group

Parameter	Description
Name	The name of the group. Only letters, digits, and underscores (_) are allowed.
Description	Information to better identify a group. The description contains up to 4,096 characters and consists of only letters, digits, and underscores (_).

Copyright © Huawei Technologies Co., Ltd.

10.3.4 Masking Policies

This section describes how to create a data masking policy and perform masking query in Data Assets.

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.

Creating a Masking Policy

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

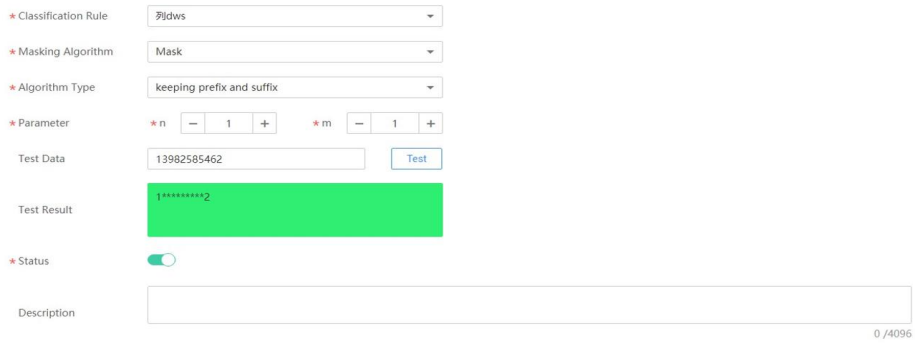
   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Assets** of the corresponding workspace to navigate to the Data Assets console.
3. Choose **Masking Policies** from the left navigation bar. Click **Create**.
4. Set **Classification Rule**, **Masking Algorithm**, and **Algorithm Type**. The options for **Masking Algorithm** include **Mask**, **Truncate**, and **Hash**. Each masking algorithm corresponds to multiple algorithm types. Select an algorithm type as required.

   NOTE

   A classification rule that has been bound to a masking algorithm cannot be bound again.

   Figure 10-16 Creating a masking policy

5. After you configured the making algorithm, you can perform the online test. Enter the test data, and then click **Test**. You can verify the result in the **Test Result** text box.
6. Enable or disable **Status**. The masking policy takes effect only when **Status** is enabled.

Viewing the Data Masking Effect

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

 Copyright © Huawei Technologies Co., Ltd.

On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Assets** of the corresponding workspace to navigate to the Data Assets console.
3. Choose **Data Catalog** from the left navigation bar.
4. In a list of searched asset results, click a table name. A page for data table details is displayed.
5. On the details page, click **Data Preview** to view the data masking effect.

10.4 Metadata Collection

10.4.1 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.

10.4.2 Task Management

This section describes how to create collection tasks by configuring metadata collection policies. Different types of data sources require different collection policies. Metadata management allows users to collect the technical metadata using the configured collection policies.

Prerequisites

Metadata collection supports various data source types. For DWS, DLI, MRS HBase, MRS Hive, RDS (MySQL and PostgresSQL), and ORACLE data sources, you need to create a data connection in Management Center.

 Copyright © Huawei Technologies Co., Ltd.

Entry

To create a collection task, you must select a directory that stores the collection task to be created. For details, see **Figure 10-17**.

Figure 10-17 Directory that stores the collection task to be created

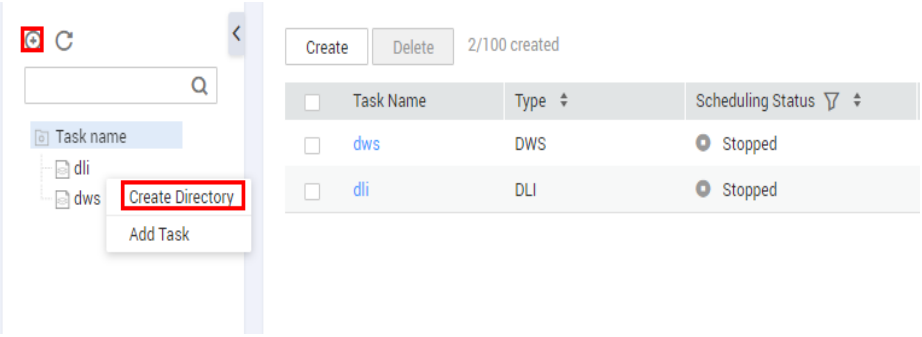
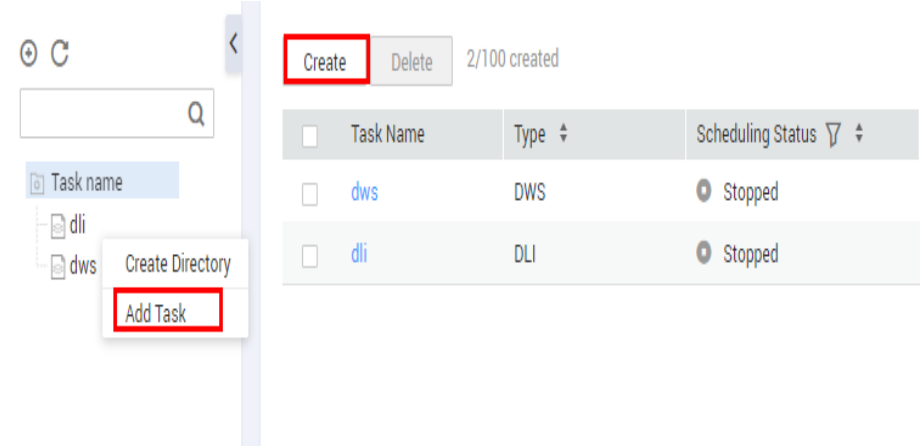


Figure 10-18 shows the entries for creating tasks.

Figure 10-18 Entries for creating a collection task



Creating a Collection Task

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Assets** of the corresponding workspace to navigate to the Data Assets console.
3. Choose **Task Management** from the left navigation bar.
4. Click **Create** in the upper part of the displayed page or right-click the **Task name** and choose **Add Task** from the shortcut menu. In the displayed dialog box, set parameters as prompt.

   a. Set basic parameters by referring to **Table 10-3**.

 Copyright © Huawei Technologies Co., Ltd.

Table 10-3 Parameters

Parameter	Description
Task Name	The name of the collection task, which contains 1 to 63 characters and consists of only letters, digits, and underscores (_).
Description	Information to better identify a collection task. The description contains up to 255 characters and consists of only letters, digits, and underscores (_).
Select Directory	The directory that stores collection tasks. You can select an existing one. **Figure 10-17** describes how to create a directory.

b. Configure data source information. For details, see **Table 10-4**.

Table 10-4 Parameters

Parameter	Description
Data Connection Type	Select a data source type from the drop-down list. **NOTE** Collection tasks support various data source types. For DWS, DLI, MRS HBase, MRS Hive, RDS (MySQL), and ORACLE data sources, you need to create data connections in Management Center. See **Creating Data Connections**.
OBS Bucket	An OBS bucket that data to be collected belongs to. This parameter is mandatory only when the data source type is **OBS**.
Data Connection Name	• Data connection in the created data connection types. You can select a value from the drop-down list. • Data connection in the data connection types that are not created. Click **Create** to add one.
OBS Path	The OBS path where data to be collected resides. This parameter is mandatory only when the data source type is **OBS**.
Collection Scope	The scope of data to be collected. This parameter is displayed only when the data source type is OBS. The reason is that the OBS bucket is hierarchical. • If you select **This folder**, the collection task collects only the objects in the folder set in the OBS path. • If you select **This folders and subfolders**, the collection task collects all objects in the folder set in the OBS path, including the objects in the sub-folders.

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Collected Content	The content of data to be collected. This parameter is displayed only when the data source type is OBS. The reason is that the OBS bucket is hierarchical. • If you select **Folders and objects**, the collection task collects folders and objects. • If you select **Folders**, the collection task collects only folders.
Database and Schema	This parameter is displayed only when the data source type is set to **DWS**. Click **Select** to set the range to be scanned by the collection task. If this parameter is not set, all databases under the data connection are scanned by default. Click **Clear** to delete the selected databases and data tables.
Namespace	This parameter is displayed only when the data source type is set to **MRS HBase**. Click **Select** next to the **Namespace** text box to set the namespace range to be scanned by the collection task. If this parameter is not set, all namespaces under the data connection are scanned by default. You can click **Clear** to modify the selected namespace.
Database Data Table	The databases and data tables to be collected. • Click **Set** to set the range to be scanned by the collection task. If this parameter is not set, all databases under the data connection are scanned by default. • Click **Set** to set the range to be scanned by the collection task. If this parameter is not set, all data tables under the data connection are scanned by default. For data tables whose data connection type is MySQL or DLI, the tables to be collected can be filtered by regular expressions. • If neither the database nor the data table is set, the collection task scans all data tables under the data connection. • Click **Delete** to delete the selected databases and data tables.
Collect Dump Task	This parameter is displayed only when the data source type is set to **DIS**. If **Yes** is selected, the dump task is collected.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Collection Channel	This parameter is displayed only when the data source type is set to **DIS**. A DIS instance is a channel. Therefore, this parameter indicates that a channel is selected for data collection.
Select Graph	This parameter is displayed only when the data source type is set to **GES**. Select graphs that store structured data based on "relationships".
Select Cluster	This parameter is displayed only when the data source type is set to **CSS**. Select the CSS cluster for storing the data to be collected. You can also click **Create** to create a CSS cluster. After the CSS cluster is created, click **Refresh** and select the new CSS cluster.
CDM Cluster	To manage CloudTable/GES/CSS data connections, select an 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	This parameter is displayed only when the data source type is set to **CSS**. Index, similar to "database" in the relational database (RDB), stores Elasticsearch data. It refers to a logical space that consists of one or more shards.

c. Set metadata collection parameters. See **Table 10-5**.

 Copyright © Huawei Technologies Co., Ltd.

Table 10-5 Parameters for metadata collection

Parameter	Description
The data source metadata has been updated.	When metadata in a data connection changes, you can configure an update policy to set the metadata update mode in the data directory. Note that the configured update and deletion policies apply only to the databases and data tables configured by yourself. • If you select **Update metadata in the data directory only**, the collection task updates only the metadata that has been collected in the data directory. • 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 directory. • 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 updated.	When metadata in a data connection changes, you can configure a deletion policy to set the metadata update mode in the data directory. • 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 directory. • If you select **Ignore the deletion**, when some metadata in the data source has been deleted, the corresponding metadata is not deleted from the data directory.
Timeout Interval for Collecting a Data Record	This parameter is displayed only when the data source type is DWS. If the metadata collection duration of a single data table exceeds the specified timeout interval, the metadata collection of the data table fails. If this parameter is not set, the default timeout interval is 30 minutes. If the size of a table is large, the metadata collection takes a long time. You can set the timeout interval and configure a collection task for each large table.

d. Set parameters when **Data Summary** is selected. See **Table 10-6** for details.

 Copyright © Huawei Technologies Co., Ltd.

Table 10-6 Parameters

Parameter	Description
Full data	**Full data**: If this option is selected, a data profile is generated in the data directory based on all data collected. This mode applies to scenarios where the data volume is less than 1 million.
Sample data	**Full data**: If this option is selected, a data profile is generated in the data directory based on all data collected. This mode is applicable to scenarios with a large amount of data.
Randomly collect {} % records of data from all data	**Full data**: If this option is selected, a data profile is generated in the data directory based on all data collected. This mode is applicable to scenarios with a large amount of data.
Data Lake Insight Queue	The queue that is used to obtain profile data and execute DLI SQLs. 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 need to 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 data classification parameters as prompt.

- If you select **Data Classification** and create a classification rule group or select an existing classification rule group, 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 highest security level of the matched classification rules.
- If you select **Manually**, no classification rule and security level are not automatically added to **Column Attributes** of **Data Catalog** under **Data Map**. Go to the **Task Monitoring** page. Locate the target

 Copyright © Huawei Technologies Co., Ltd.

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 DWS, DLI, and MRS Hive data sources, 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.

   Schedule once: If the execution duration of a task exceeds the configured timeout duration, the task is considered failed.

   Schedule periodically: See **Table 10-7** for details.

   NOTE

   1. If **Schedule once** is selected, a manual task instance is generated. A manual task has no dependency on scheduling and needs to be manually triggered.
   2. If **Schedule periodically** 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 management on scheduled instance tasks, such as viewing the running status, stopping and rerunning the scheduled tasks.

Table 10-7 Parameters

Parameter	Description
Scheduling Date	The period during which a scheduling task takes effect
Scheduling Cycle	The frequency at which a scheduling task is executed. Related parameters are: • Minute • Hour • Day • Week
Start Time	The time at which the scheduling task starts.
Interval	The interval between the last scheduling end time and the next scheduling start time.
End Time	The time at which the scheduling task ends.
Timeout Interval	If the running duration of a task exceeds the timeout interval set here, the task is regarded as failed.
Start	If this check box is selected, the task is scheduled immediately.

 Copyright © Huawei Technologies Co., Ltd.

6. Click **Submit**. The collection task is created successfully.

Managing a Collection Task

Enter the workspace where to the collection task resides and click **Data Assets**. The Data Assets console is displayed.

On the Data Assets console, choose **Task Management** from the left navigation bar.

Then, you can view all created collection tasks.

Table 10-8 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.
Data Source	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
Created By	The creator of a collection task
Last Run	Displays the latest running time of a collection task.
Operation	You can perform the following operations on a created collection task: • **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**: If the task status is **Stopped**, you can reschedule the task. • **Stop Scheduling**: This parameter is valid only for collection tasks that are scheduled periodically. If the task status is **Scheduling**, you can stop it.

10.4.3 Task Monitoring

You can monitor the running statuses of the metadata collection tasks, view collection logs, and perform operations such as rerunning collection tasks.

 Copyright © Huawei Technologies Co., Ltd.

On the Data Assets console, choose **Task Monitoring** from the left navigation bar. On the **Task Monitoring** page, monitor the created collection tasks. See **Table 10-9** for details.

Table 10-9 Parameters for monitoring a collection task

Parameter	Description
Task Name	The name of a collection task
Instance Status	The status of an instance (collection task), which are: • Successful • Executing • Failed
Schedule	The scheduling mode of a collection task. The options are **Schedule once** and **Schedule periodically**.
Scheduling Cycle	The scheduling period of a collection task
Start Time	The time when a collection task restarts running
End Time	The time when a collection task stops running
Running Duration (min)	The running time of a collection task
Operation	The operations that can be performed on a collection task under monitoring: • **Rerun**: Instances whose statuses are **Canceled** or **Succeeded** can be rerun. • **View Log**: Only logs of instances that are successfully executed can be viewed. **NOTE** Click **View Log** to view the run logs of the metadata collection, data summary, and data classification tasks in real time. • **Cancel**: Instances whose statuses are **Executing** can be stopped. • **View Scanning Result**: 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.

 Copyright © Huawei Technologies Co., Ltd.

11 Data Lake Mall

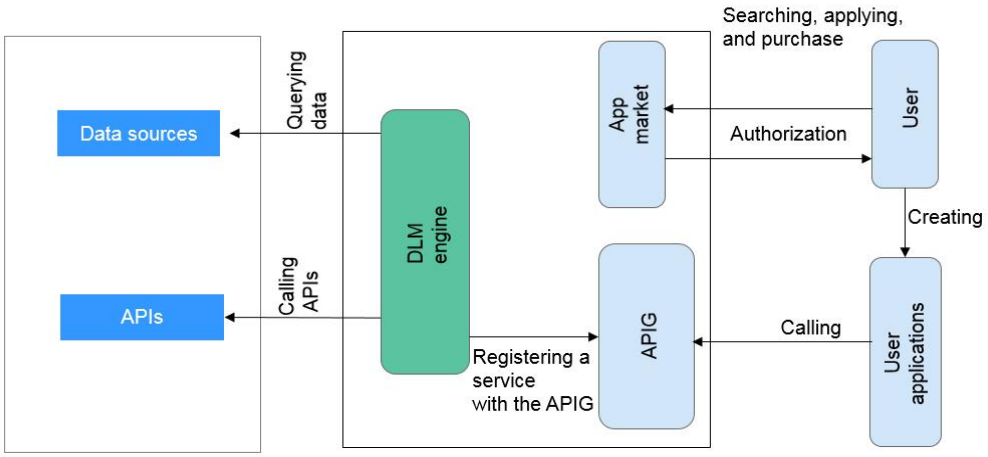
11.1 Overview

This section describes the functions of Data Lake Mall (DLM), including data API generation, registration, calling, and operations.

DGC DLM aims to build a unified data service bus for enterprises to centrally manage internal and external API services. DLM helps to quickly generate data APIs based on data tables and allows you to register existing APIs with the DLM platform for integrated management and release.

DLM 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 running environment. DLM prepares compute resources and supports for elastic scaling, bringing you low costs and no O&M expenditure.

Figure 11-1 DLM architecture



Overview Page

On the **Overview** page, you can view various monitoring data views. The **Overview** page displays **API Development** and **API Calls**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-2 API development

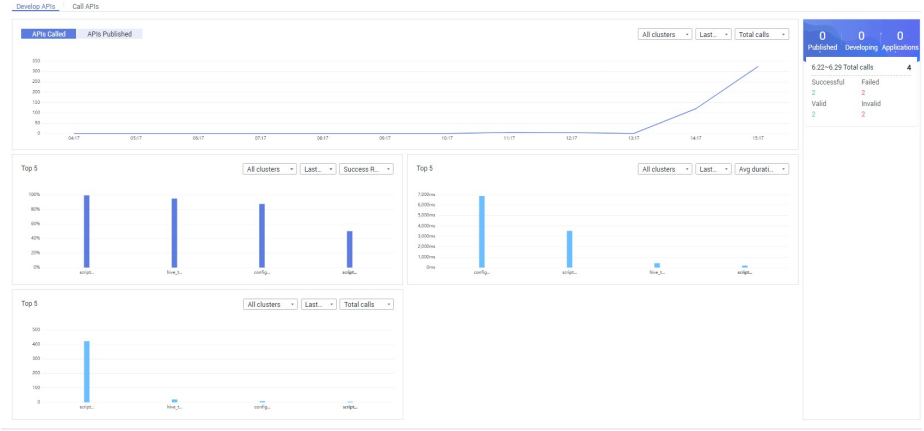


Table 11-1 Parameters of API publishing

Parameter	Description
API Published	The number of APIs published every day, week, month, and year.
API 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 successfully published on the service market.
Developing	The number of APIs that are being developed.
Applying	The number of applications that are being applied.
Successful API Calls	The number of successful API calls.
Failed	The total number of failed API calls.
Total API Calls	The total number of API calls.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-3 API calls

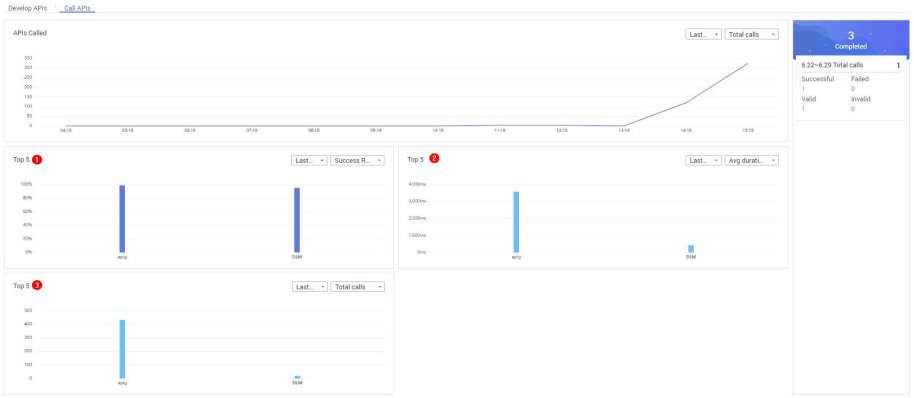


Table 11-2 Parameters about API calling

Parameter	Description
API Called	The number of API calls made every day, week, month, and year.
API Call Ratio	The ratio of successful and failed API calls in the last seven days.
Applied APIs	The number of APIs applied on the DLM platform.
Successful API Calls	The number of successful API calls on the DLM platform.
Total API Calls	The number of total API calls on the DLM platform.

Procedure for Publishing an API

To publish an API or group of APIs, do as follows:

1. **Create an API group**.

   An API group facilitates management of APIs used for the same service. Create an API group before creating APIs.

2. Create an API.

   You can **generate** and **register** APIs. An API can be generated in the **wizard mode** or **script mode**.

3. **Publish an API**.

   An API can be called on the DLM platform only after it is published.

Calling an API

To call an API, do as follows:

1. Obtain an API.

   You can obtain an API from the service market of DLM.

 Copyright © Huawei Technologies Co., Ltd.

2. **Create an application and obtain the application authorization**.

   To call an API that is accessed through application authentication, you must create an application to generate an application ID and key pair (AppKey and AppSecret). You can call the API only after you bind the application to it. Provide your application information to the API developer to get authorized. When an API is called, the key pair replaces the one in the SDK, and DLM authenticates the identity of the API caller by verifying the key pair. For details on how to use the application authentication method, see the **DLM SDK Reference**.

3. Call an API.

   After completing the preceding operations, you can call an API by following the procedures in **Calling an API**.

11.2 Comparison Between Shared Edition and Exclusive Edition

Currently, DLM provides two editions: Shared and Exclusive. Shared DLM is available immediately after being enabled. To use DLM Exclusive Edition, you need to buy an incremental package of the exclusive cluster and manage APIs in the instance.

Exclusive DLM is usually used to decouple internal systems of an enterprise. Services are deployed in a VPC on the cloud and communicate with each other through RESTful APIs. Communication links are established within the VPC, further ensuring network security. In addition, you can deploy the frontend or backend services of an exclusive instance on the public network and bind an EIP to the instance for network interaction.

Table 11-3 Differences between Shared DLM and Exclusive DLM

Difference	Shared Edition	Exclusive Edition
Billing	Billed by API calling times	Billed by instance specifications and usage duration
Network access	Public network access	Instances run in a VPC. In the VPC, the instance subnet address is used to call APIs. You can enable public network ingress and egress access to allow external services to call APIs and deploy backend services in the external network environment.
Sales in the cloud market	APIs can be sold on the market place.	APIs cannot be sold on the market place.

 Copyright © Huawei Technologies Co., Ltd.

Difference	Shared Edition	Exclusive Edition
Intended customers	Small-sized enterprises. Low requirements on physical isolation, out-of-the-box, and quick API capability openness.	Large- and medium-sized enterprises. Exclusive DLM has physically isolated independent clusters and richer features. Internal APIs can be called and opened, and independent inbound and outbound bandwidths can be selected.

The following table lists the similarities and differences between the shared edition and exclusive edition.

Table 11-4 Function differences between Shared DLM and Exclusive DLM

Category	Description	Shared Edition	Exclusive Edition
Basic functions	Refined request throttling policies	30	500
	IP and user access control	√	√
	Security authentication	√	√
	API lifecycle management	√	√
	Custom domain name	√	√
	VPC channels	√	√
	API parameter orchestration	√	√
	API group variable management	√	√
Advanced functions	Custom authentication	√	√
	Interconnection with the API marketplace	√	To be supported
	Policy-based routing	√	√
	O&M analysis display	√	√

 Copyright © Huawei Technologies Co., Ltd.

Category	Description	Shared Edition	Exclusive Edition
	Backend load balancing	×	√
	Internal API management	×	√
	Access to backend services in private clouds	×	√
	Service access through Direct Connect	×	√
Performance metrics	Physically isolated clusters	×	√
	Different bandwidths for inbound and outbound access	×	√
	Number of recommended requests per second	200TPS	8000-30000TPS

Specifications of Shared Edition

There are no specific specifications for Shared Edition. For details on the quotas of APIs created and used by users, see **Table 11-5**.

If you need to change the default quota, see **How Do I Apply for a Higher Quota?**

Table 11-5 Usage restrictions

Item	Default Restriction	Modifiable
API groups	Up to 50 API groups can be created by a user.	√
APIs	Up to 200 APIs can be created by a user.	√
Backend policies	Up to 5 backend policies can be created by a user.	√
Applications	Up to 50 apps can be created by a user. The app quota includes the apps that you have created and apps generated when you purchase APIs from the service market.	√

 Copyright © Huawei Technologies Co., Ltd.

Item	Default Restriction	Modifiable
Request throttling policies	• Up to 30 throttling policies can be created by a user. • The maximum number of requests cannot exceed the maximum number of API requests. • Maximum application requests cannot exceed the maximum user requests. • The maximum number of source IP requests cannot exceed the maximum number of API requests.	√
Access control policies	Up to 100 access control policies can be created by a user.	√
Number of VPC channels	Up to 30 VPC channels can be created by a user.	√
Number of variables	Up to 50 variables for an API group in each environment can be created by a user.	√
Number of ECSs	Up to 200 ECSs can be added to a VPC channel.	√
Number of parameters	Up to 50 parameters can be created for each API.	√
Number of release records	Up to 10 publication records of an API can be retained for each environment.	√
Access frequency of each API	No more than 200 times per second.	√
Special apps	The threshold cannot exceed the maximum API requests.	x
Number of times that the subdomain name is accessed	Each subdomain name can be accessed up to 1000 times a day.	x
Size of an invoking request package	The maximum size of a request package that can be invoked by an API at a time is 12 MB.	x
TLS protocol	TLS1.1 and TLS1.2 are supported. TLS1.2 is recommended.	x
Real-name authentication	Users who have not passed real-name authentication cannot perform any creation operations.	x

 Copyright © Huawei Technologies Co., Ltd.

Specifications of Exclusive Edition

Table 11-6 lists the specifications of API Gateway exclusive instances.

Table 11-6 Specifications of exclusive instances

Instance	Max. Requests / Second
Basic	8,000
Advanced	15,000
Professional	30,000

11.3 Glossary

- A data connection refers to the database connection information. DGC DLM accesses data in data sources that are configured in DGC Management Center.
- API generation is to generate data APIs through a series of configurations.
- API registration is to register existing APIs with DLM for unified management.
- Wizard mode is to generate APIs based on step-by-step visual configurations on GUI. No code needs to be written, which is suitable for simple APIs and beginners.
- Script mode is to generate APIs based on custom SQL scripts. Multi-table join, complex query conditions, and aggregation functions are supported, which are suitable for complex APIs and developers with programming experience.
- A data service instance is a container that bears API service functions. Each instance is an independent carrier of a group of API service functions. API services are provided as instances.
- 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.
- An API gateway is a fully managed service provided by HUAWEI CLOUD. API gateway handles all the tasks involved in accepting and processing concurrent API calls, including traffic management, authorization and access control, and API life cycle management.
- DLM service market is an API transaction platform built in DGC. It allows you to find the services you need and apply for permissions.
- An application is a carrier for authenticating the permissions to call APIs. A user can create multiple applications, and an application can be used to apply for multiple API permissions.
- An SDK is a tool for simplifying operations when users call APIs. The SDKs in nine languages are available in DLM.
- A throttling policy controls the number of API calls made by a user or an application within a specified time to protect the backend services. The throttling can be accurate to the minute or second.

 Copyright © Huawei Technologies Co., Ltd.

11.4 Adding Reviewers

Add a reviewer first because the application of creating an API needs to be reviewed. Only the workspace administrator has the permissions required to add reviewers.

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

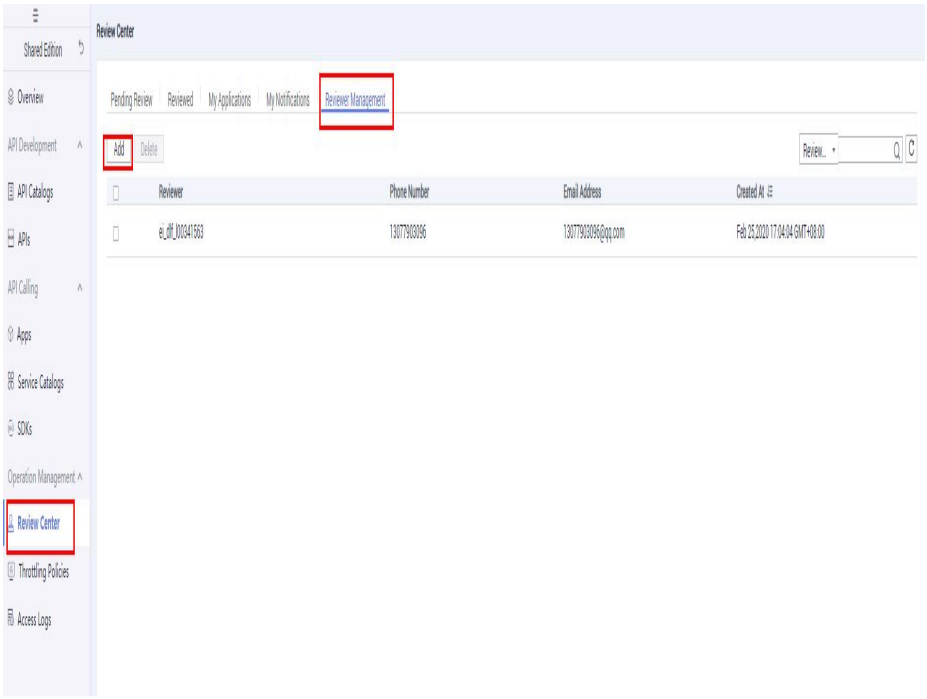
   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Lake Mall** of the corresponding workspace.

   Navigate to the Data Lake Mall console.

3. Choose **Review Center** from the left navigation bar. On the page that is displayed, choose **Reviewer Management** and click **Add**.

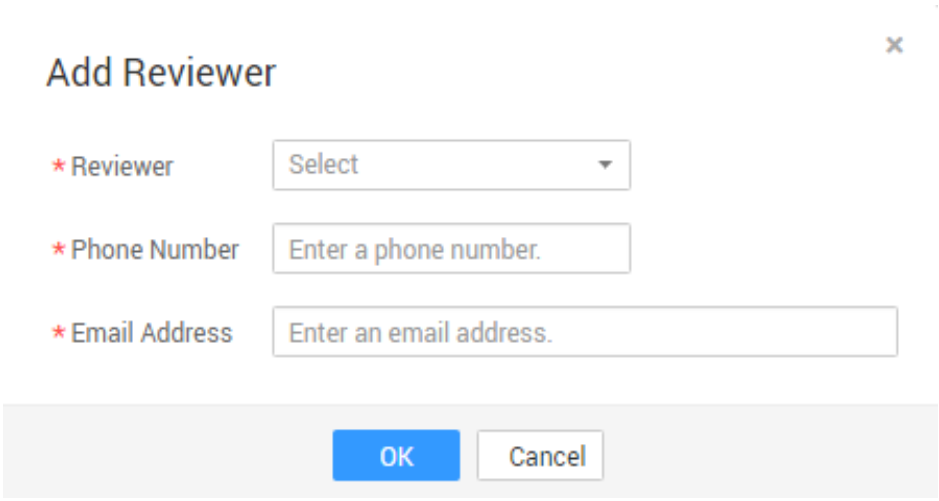
   Figure 11-4 Reviewer Management tab page



4. Select a reviewer (account or IAM user), enter the correct phone number and email address, and click **OK**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-5 Adding a reviewer



5. Add multiple reviewers if needed.

11.5 Generating APIs

11.5.1 Function Overview

This section describes the differences between API generation in the wizard mode and script mode.

DLM supports quick generation of data APIs based on the DWS, DLI, MRS Hive, RDS, and MRS HBase tables in the Wizard mode. You can configure a data API within several minutes without coding.

To meet personalized query requirements of users, DLM also supports API generation in the custom SQL Script mode. You can compile API query SQL statements. Multi-table join, complex query conditions, and aggregation functions are supported in this mode.

11.5.2 Buying Exclusive Instances

This section describes how to buy an exclusive instance. You can create an API and use it to provide services only after the instance is created. If you do not have high performance requirements, you can use Shared Edition and create APIs in the wizard mode by following instructions in **Generating Wizard APIs** without buying exclusive instances.

Network Environment Preparation

- Virtual Private Cloud (VPC)

  A VPC must be configured for an exclusive instance. Resources (such as ECSs) in the same VPC can use the private address of the exclusive instance to call APIs.

  When buying an exclusive instance, you are advised to configure the same VPC as other associated services to ensure network security and facilitate network configuration.

 Copyright © Huawei Technologies Co., Ltd.

- Security Group (SG)

  An SG is similar to a firewall. It controls who can access the specified port of an instance and makes the communication data flow of the instance to the specified destination address. You are advised to enable the IP address and port in the inbound direction of the SG as required to protect the network security of the instance to the maximum extent.

  The SG bound to an exclusive instance must meet the following requirements:

  - Inbound rule: To call APIs from the public network or from resources in other SGs, enable ports 80 (HTTP) and 443 (HTTPS) in the inbound direction of the SG bound to the exclusive instance.
  - Outbound direction: If the backend service is deployed on the public network or in another SG, you need to enable the backend service address and API calling listening port in the outbound direction of the SG bound to the exclusive instance.
  - If the frontend and backend services of the API are bound to the same SG and VPC as the exclusive instance, you do not need to enable the preceding ports for the exclusive instance.

Procedure

Step 1 Locate an enabled instance and click **Buy**.

Step 2 Access the page for buying a DGC incremental package and set parameters. See **Table 11-7** for parameter details.

Table 11-7 Parameters for buying an exclusive DLM instance

Parameter	Description
Package	Select **DLM**.
Billing Mode	Specify the billing mode of the instance. Currently, pay-per-use is supported.
Workspace	Select the workspace for which you want to use the incremental package. For example, if you want to buy a DLM incremental package in the ***A*** workspace of instance ***test***, select ***A***. If the purchase is successful, you can view the exclusive DLM cluster in the ***A*** workspace.
AZ	A physical region where resources use independent power supplies and networks. AZs are physically isolated but interconnected through an internal network. You can select multiple AZs to deploy a DLM instance across AZs, improving instance availability.
Cluster	N/A
Cluster Specifications	Different instance specifications support different concurrent API requests..

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Enterprise Project	When you log in as an enterprise user, you can select the enterprise project to which the instance belongs. For details on resource usage, migration, and user permissions of enterprise projects, see the **Enterprise Management User Guide**.
VPC/Subnet	Bind an instance to a VPC and allocate a subnet to the instance. Cloud service resources (such as ECSs) in the same VPC can use the private IP address of an exclusive DLM instance to call APIs. You are advised to configure the same VPC for the exclusive instance and other associated services to ensure network security and facilitate network configuration.
Security Group	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. For example, if the backend service is deployed on an external network, you need to configure security group rules to allow access to the IP address and listening port of the backend service. **NOTE** If the public network entry is enabled, the security group must allow access from ports 80 (HTTP) and 443 (HTTPS) in the inbound direction.
Purchase Quantity	N/A

Step 3 Click Buy Now. If there are no changes made to the specifications, submit the order.

----End

Follow-up Procedures

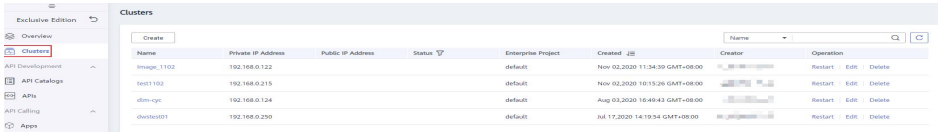
After an exclusive cluster is created, you can view the cluster status on the **Exclusive Edition** module. If a cluster is abnormal, you can locate the cause based on the pop-up message. The cluster status includes the following:

- CREATING
- CREATED
- STARTING
- RUNNING
- LOST
- EXCEPTION
- RESTARTING
- MODIFYING

 Copyright © Huawei Technologies Co., Ltd.

- GROWING
- UNKOWN
- SHUTDOWN
- SHUTTING_DOWN
- WAITTING_IP_ALLOCATION

Figure 11-6 Viewing exclusive instances



11.5.3 Generating Wizard APIs

This section describes how to generate an API in the wizard mode.

Generating data APIs in the wizard mode is simple and easy. An API can be generated without writing 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**.

Managing an API Catalog

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. **Log in to the DGC console**, select an instance, and click **Access** on it.
   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Lake Mall** of the corresponding workspace.
   Navigate to the Data Lake Mall console.
3. On the displayed page, choose **API Catalogs** and click .
   In the displayed dialog box, enter an API catalog name.
4. In the **Operation** column of API catalog list, edit API catalogs.
   Click **Edit** to the right of the API catalog that you want to edit.

Configuring Basic API Information

1. **Log in to the DGC console**, select an instance, and click **Access** on it.
   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

 Copyright © Huawei Technologies Co., Ltd.

2. Click **Data Lake Mall** of the corresponding workspace.
   Navigate to the Data Lake Mall console.
3. Choose **APIs** > **Create**. On the displayed page, enter the basic information.

Table 11-8 API basic info

Parameter	Description
API	An API name consists of 3 to 64 characters and starts with a letter. Only letters, digits, and underscores (_) are allowed.
API Catalog	An API catalog is a set of APIs with a specific function or scenario. It is the minimum organization unit of APIs in DLM and the minimum management unit of the API gateway. Click **Select Catalog** to create an API catalog or select an existing one created in **Managing an API Catalog**.
Request Path	The path for accessing an API. Example: **/v2/{project_id}/streams**.
Parameter Protocol	A parameter protocol is used to transmit requests. HTTP and HTTPS are supported.
Request Mode	HTTP defines the following request modes that can be used to send a request to the server: **GET** requests the server to return specified resources. **POST** requests the server to add a resource or perform special operations.
Description	A brief description of the API to be created.
Tag	The tag set for an API. A maximum of 20 tags can be set for an API.
Reviewer	An owner who has permissions to review APIs. Click **Add new** to enter the **Review Center** page. On the displayed page, choose **Reviewer Management** > **Add**.
Security Authentication	Security authentication mode, which can be: • **App Authentication**: API Gateway authenticates API requests. • **IAM Authentication**: IAM authenticates API requests. • **None-authentication**: No authentication is required.
Display Scope	After the API is published, all users in the selected scope can view the API in the service catalog. • Current workspace APIs • Current project APIs • Current tenant APIs

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Access Log	If this parameter is selected, you can view the generated API query results that will be kept for seven days on the **Access Logs** page.
Min. Retention Period	Minimum duration reserved before API unbinding. Before an API developer suspends, brings offline, or cancels the authorization of an API, the system notifies the authorized API callers and reserves at least *X* 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	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. 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. 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 the Data Extraction Logic

Set **Data Acquisition Method** to **Configuration Mode**.

1. Select a data source, data connection, database, and data sheet to obtain the tables to be configured.

   NOTE

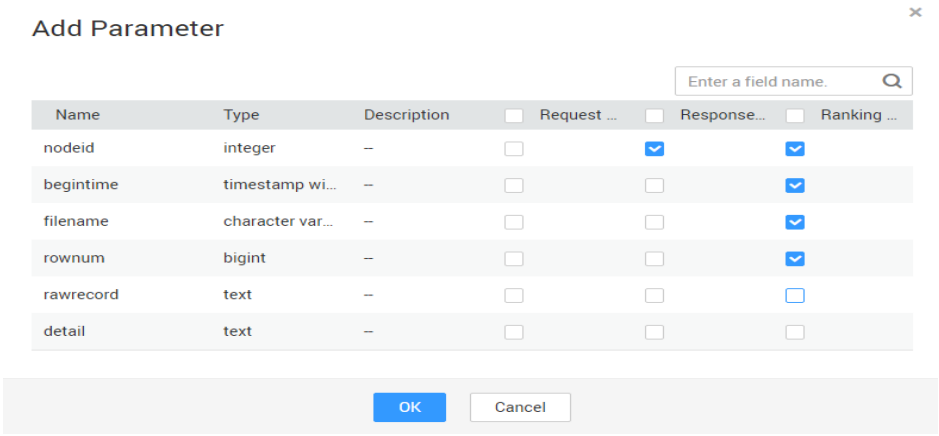
   Currently, only DWS, DLI, HBASE, HIVE, MySQL and RDS data sources are supported. Configure data sources in DGC Management Center in advance. Data tables can be searched by their names.

2. Configure parameter fields.

   Click **Add** next to **Parameter Setting**. 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.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-7 Add Parameter dialog box



3. Edit request parameters.

   In the request parameter list, select a bound parameter and an operator.

4. Edit response parameters.

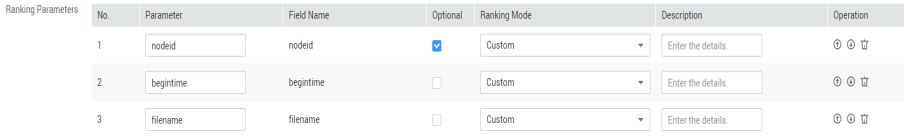
   In the response parameter list, enter the values for **Parameter**, **Parameter Type**, **Example Value**, and **Description**.

5. Edit ranking parameters.

   In the list of ranking parameters, set whether the ranking parameters are optional, set the ranking mode, and enter the description.

   Click the buttons in the **Operation** column to move up, down, or delete parameters.

Figure 11-8 Editing ranking parameters



6. Click **Next** and set the value of **pre_order_by** as the description of all ranking parameters and separate them with semicolons (;). See **Figure 11-9**.

   Figure 11-8 is an example.

Table 11-9 Ranking parameters

Parameter	Description
id	**a:asc**, where **a** is the parameter name. **asc** indicates that parameters are listed in the ascending order.

 Copyright © Huawei Technologies Co., Ltd.

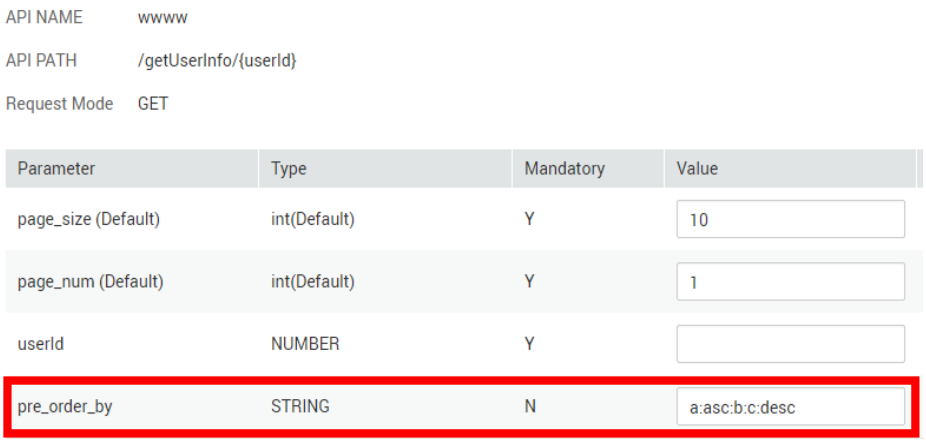
Parameter	Description
name	• **b:asc** • **b** • **b:desc**, where **b** is the parameter name. **asc** and **desc** indicate that parameters are listed in ascending and descending orders respectively. The ranking order can be customized or left blank.
age	**c:desc**, where **c** is the parameter name. **desc** indicates that parameters are listed in the descending order.

Table 11-9 lists the ranking parameters. The following table describes how to configure **pre_order_by**.

Table 11-10 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 11-9 Setting ranking parameters



 Copyright © Huawei Technologies Co., Ltd.

NOTICE

- **pre_order_by** is optional. If **pre_order_by** is not set, use the mandatory fields as the ranking criterion.
- If **pre_order_by** is set, configure an API strictly following the ranking sequence specified in **Figure 11-8**. 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 an API

After all parameters are set, click **Debug**. Enter the values for parameter **Value** and click **Debug** in the lower left corner. 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**.

11.5.4 Generating Script APIs

This section describes how to generate an API in the script mode.

To meet personalized query requirements of users, DLM also supports the API generation in the custom 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. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Lake Mall** of the corresponding workspace.

   Navigate to the Data Lake Mall console.

3. Choose **APIs** > **Create**. On the displayed page, enter the basic information.

Table 11-11 API basic info

Parameter	Description
API	An API name consists of 3 to 64 characters and starts with a letter. Only letters, digits, and underscores (_) are allowed.
API Catalog	An API catalog is a set of APIs with a specific function or scenario. It is the minimum organization unit of APIs in DLM and the minimum management unit of the API gateway. Click **Select Catalog** to create an API catalog or select an existing one created in **Managing an API Catalog**.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Request Path	The path for accessing an API. Example: **/v2/{project_id}/streams**.
Parameter Protocol	A parameter protocol is used to transmit requests. HTTP and HTTPS are supported.
Request Mode	HTTP defines the following request modes that can be used to send a request to the server: **GET** requests the server to return specified resources. **POST** requests the server to add a resource or perform special operations.
Description	A brief description of the API to be created.
Tag	The tag set for an API. A maximum of 20 tags can be set for an API.
Reviewer	An owner who has permissions to review APIs. Click **Add new** to enter the **Review Center** page. On the displayed page, choose **Reviewer Management** > **Add**.
Security Authentication	Security authentication mode, which can be: ● **App Authentication**: API Gateway authenticates API requests. ● **IAM Authentication**: IAM authenticates API requests. ● **None-authentication**: No authentication is required.
Display Scope	After the API is published, all users in the selected scope can view the API in the service catalog. ● Current workspace APIs ● Current project APIs ● Current tenant APIs
Access Log	If this parameter is selected, you can view the generated API query results that will be kept for seven days on the **Access Logs** page.
Min. Retention Period	Minimum duration reserved before API unbinding. Before an API developer suspends, brings offline, or cancels the authorization of an API, the system notifies the authorized API callers and reserves at least *X* 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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Input parameter	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. 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. 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 the Data Extraction Logic

Set **Data Acquisition Method** to **Script mode**.

1. Select a data source, data connection, database, and queue to obtain the tables to be configured.

   NOTE

   Currently, only DWS, DLI, HBASE, HIVE, MySQL and RDS data sources are supported. Configure data sources in DGC Management Center in advance and enter SQL statements as prompted.

2. Compile a SQL statement to query APIs.

   On the script editing page, enter the SQL statement as prompted.

   NOTE

   - The fields queried by the SELECT statement are parameters returned by APIs.
   - The parameters in the WHERE statement are parameters requested by APIs. The format is **${*parameter name*}**.

3. Select a pagination mode.
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**.

 Copyright © Huawei Technologies Co., Ltd.

NOTICE

pre_order_by is optional. If **pre_order_by** is not set, use the mandatory fields as the ranking criterion.

Testing an API

After all parameters are set, click **Debug**. Enter the values for parameter **Value** and click **Debug** in the lower left corner. 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**.

11.6 Registering APIs

This section describes how to register APIs, and manage and publish APIs generated based on data tables to the API gateways.

DLM supports the registration of RESTful APIs, including GET and POST APIs.

Configuring Basic API Information

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Lake Mall** of the corresponding workspace.

   Navigate to the Data Lake Mall console.

3. Choose **APIs** > **Register**. Configure the basic information.

Table 11-12 API basic info

Parameter	Description
API	An API name consists of 3 to 64 characters and starts with a letter. Only letters, digits, and underscores (_) are allowed.
API Catalog	An API catalog is a set of APIs with a specific function or scenario. It is the minimum organization unit of APIs in DLM and the minimum management unit of the API gateway. Click **Select Catalog** to create an API catalog or select an existing one created in **Managing an API Catalog**.
Request Path	The path for accessing an API. Example: **/v2/{project_id}/streams**.
Protocol	A parameter protocol is used to transmit requests. HTTP and HTTPS are supported.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Request Mode	HTTP defines the following request modes that can be used to send a request to the server: **GET** requests the server to return specified resources. **POST** requests the server to add a resource or perform special operations.
Description	A brief description of the API to be registered.
Tag	Tag names can include letters, digits, and underscores (_), and cannot start with underscores (_).
Reviewer	An owner who has permissions to review APIs. Click **Add new** to enter the **Review Center** page. On the displayed page, choose **Reviewer Management** > **Add**.
Security Authentication	Security authentication mode, which can be: ● App Authentication: API Gateway authenticates API requests. ● **IAM Authentication**: IAM authenticates API requests. ● **None-authentication**: No authentication is required.
Display Scope	After the API is published, all users in the selected scope can view the API in the service catalog. ● Current workspace APIs ● Current project APIs ● Current tenant APIs
Access Log	If this parameter is selected, you can view the generated API query results that will be kept for seven days on the **Access Logs** page.
Min. Retention Period	Minimum duration reserved before API unbinding. Before an API developer suspends, brings offline, or cancels the authorization of an API, the system notifies the authorized API callers and reserves at least *X* 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.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Input parameter	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. 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. 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 the 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 11-13 API parameters

Parameter	Description
Protocol	A parameter protocol is used to transmit requests. HTTP and HTTPS are supported. This parameter is used by DLM to transmit requests to the APIs to be registered.
Request Mode	HTTP defines the following request modes that can be used to send a request to the server: This parameter is used by DLM to transmit requests to the APIs to be registered. **GET** requests the server to return specified resources. **POST** requests the server to add a resource or perform special operations.
Backend Service Host	Backend service host is the host of the API to be registered. The value cannot start with **http://** or **https://** and cannot contain **Path**.
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)	This parameter refers to the backend timeout interval.

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
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**. Enter the values for parameter **Value** and click **Debug** in the lower left corner. 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**.

11.7 Debugging APIs

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.
- When an API is bound to a signature key, the API 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. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Lake Mall** of the corresponding workspace.

   The **Data Lake Mall** page is displayed.

 Copyright © Huawei Technologies Co., Ltd.

3. In the navigation pane on the left, click **Exclusive Edition** or **Shared Edition**. The required page is displayed.
4. Choose **APIs**.
5. Use either of the following methods to debug an API:
   - Locate the row that contains the target API, and choose **More** > **Debug**.
   - Click an API name to access the details page. Click **Debug** in the lower left corner.

     You can configure API request parameters in the left pane. See **Table 11-14** 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 11-14 Debugging APIs

Parameter	Description
Query Parameter	Query parameters and their values.
Cluster Settings	Supported only by the 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.
6. After adding the request parameters, click **Send Request**.

   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 is returned.
   - If the debugging fails, the HTTP status code 4*xx* or 5*xx* is returned.
7. 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 Procedure

After an API is debugged, you can publish the API. See **Publishing APIs** for details. Alternatively, you can create a request throttling policy for the API and bind it to the API for security purposes. See **Creating Throttling Policies** for details.

 Copyright © Huawei Technologies Co., Ltd.

11.8 Authorizing APIs

Scenarios

After an API is created, you can authorize specified applications to invoke it.

NOTE

- An API can be invoked only after it has been published in an environment.
- An application can be associated with an API only when the API uses App authentication.

Prerequisites

- An API has been created.
- An application has been created.

Procedure

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Lake Mall** of the corresponding workspace.

   The **Data Lake Mall** page is displayed.

3. Authorize the API to the application in either of the following ways:

   Giving API authorization:

   a. In the navigation pane, choose **API Development** > **API Management**.

   b. Locate the row that contains the API to be bound, and click **View Authorization**.

      On the **Complete Information** tab page, click **Assign Authorization**.

   c. 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 to be bound to an application. The API information page is displayed.

   c. On the displayed page, click **Permission Application**.

   d. Set an expiry, select an application, and click **OK**.

   e. After the application is submitted, the authorization takes effect only after it is approved in the review center.

4. After the authorization is complete, you can view the bound APIs on the application details page.

 Copyright © Huawei Technologies Co., Ltd.

NOTE

- In the API list, if you no longer access an API through the application, click **Unbind** in the **Operation** column.
- To commission an API to which the application is bound, choose **More** > **Commission** in the Operation column
- To extend the authorization period for the bound API, click **Renew**.

11.9 Publishing APIs

This section describes how to publish APIs on DLM.

DLM provides API hosting services through the API gateway, including API publishing, management, O&M, and sales. It helps you easily implement microservice aggregation, frontend and backend separation, and system integration. With DLM, you can access data services in a quick, cost-effective, and low risk manner.

In addition, you can publish and call APIs on DLM. The functions provided by DLM include permissions management, traffic control, access control, and others. For security reasons, the generated and registered APIs on DLM must be published to the service market before being provided outside.

Publishing an API to Service Market

When being published, the API is synchronized to **Review Center** under **Operation Management** of **Data Lake Mall**. After this API is approved, you can publish it to **Service Market** in a few clicks.

When the API is published, go to the **Service Catalog** page to view its information. You can also control the access traffic on the **Throttling Policies** page. For details, see **Creating Throttling Policies**.

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Lake Mall** of the corresponding workspace.

   Navigate to the Data Lake Mall console.

3. On the displayed page, choose **APIs** from the left navigation bar. In the **Operation** column, choose **More** > **Publish**.

11.10 Reviewing APIs

The review center of DLM 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 service market, remove an API from the service market, and reclaim the authorization of an app, these operations take effect only after the reviewers approve the operations.
- If an API caller wants to authorize an API or apply for renewal, these operations take effect only after the reviewers approve the operations.

 Copyright © Huawei Technologies Co., Ltd.

- An API developer or caller can call an API application to be reviewed in the review center.

Reviewing an Application

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Lake Mall** of the corresponding workspace.

   Navigate to the Data Lake Mall console.

3. Choose **Review Center** from 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 DLM, you can add and delete reviewers.

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Lake Mall** of the corresponding workspace.

   Navigate to the Data Lake Mall console.

3. Choose **Review Center** from the left navigation bar and click the **Reviewer Management** tab.
4. Click **Add**. In the displayed **Add Reviewer** dialog box, set the basic information.

Canceling an API Application

DLM provides the function of canceling applications to be reviewed. Choose **Review Center** > **My Applications** to cancel applications to be reviewed.

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Lake Mall** of the corresponding workspace.

   Navigate to the Data Lake Mall console.

3. Choose **Review Center** from the left navigation bar and click the **My Applications** tab.
4. Locate the row that contains the API to be canceled, and click **Cancel** in the **Operation** column.

 Copyright © Huawei Technologies Co., Ltd.

11.11 Suspending APIs

Scenarios

If you want to suspend an API that has been published, you can temporarily remove the API.

NOTICE

After this operation, the API cannot be accessed in the specified environment. Ensure that you have notified users of the consequence.

Prerequisites

- An API has been created.
- An API has been published in the environment.

Procedure

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Lake Mall** of the corresponding workspace.

   The **Data Lake Mall** page is displayed.

3. In the navigation pane on the left, click **Exclusive Edition** or **Shared Edition**. The required page is displayed.
4. Choose **APIs**.
5. Locate the row that contains the target API, click **More**, and select **Remove**.

   NOTE

   To suspend APIs in batches, select the APIs and click **Suspend**. A maximum of 1000 APIs can be suspended at a time.

6. In the displayed dialog box, select the environment where the API needs to be suspended and click **OK**.

11.12 Deleting APIs

Scenarios

You can delete a published API that no longer needs to provide services.

Copyright © Huawei Technologies Co., Ltd.

NOTICE

- If this operation is performed, the API cannot be accessed, which may affect the applications or users that are using the API. Ensure that users have been notified of this consequence.
- For a published API, you need to suspend it before deleting it.

Prerequisites

- An API group and APIs have been created.
- The API does not need to provide services.

Procedure

1. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Lake Mall** of the corresponding workspace.

   The **Data Lake Mall** page is displayed.

3. Choose **API Catalogs**. On the displayed page, select the API you want to delete and click **Delete**.
4. Click **OK** to delete the API.

   NOTE

   To delete APIs in batches, select the APIs and click **Delete**. A maximum of 1000 APIs can be deleted at a time.

11.13 Calling APIs

Preparation

To call an API, do as follows:

1. Obtain an API.

   You can obtain an API from the service market of DLM.

2. Create and authorize an application. See **Creating an Application and Getting Authorized** for details.

   To call an API that is accessed through application authentication, you must create an application to generate an application ID and key pair (**AppKey** and **AppSecret**). You can call the API only after you bind the application to it. Provide your application information to the API developer to get authorized. When an API is called, the key pair replaces the one in the SDK, and DLM authenticates the identity of the API caller by verifying the key pair. For details on how to use the application authentication method, see the **DLM SDK Reference**.

3. Call an API.

 Copyright © Huawei Technologies Co., Ltd.

After completing the preceding operations, you can call an API by following the procedures in **Calling an API**.

Creating an Application and Getting Authorized

Creating an application

1. **Log in to the DGC console**, select an instance, and click **Access** on it.
   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Lake Mall** of the corresponding workspace.
   Navigate to the Data Lake Mall console.
3. Choose **API Calling** > **Apps**. On the displayed page, click **Create**. The **Create App** dialog box is displayed. Set the parameters listed in **Table 11-15**.

   Table 11-15 App information

Parameter	Description
Name	The name of the application to be created.
Applied To	Exclusive DLM, which does not connect to a gateway. Exclusive ROMA_APIC, which connects to the ROMA APIC gateway. Exclusive APIGW, which connects to the APIGW gateway. IAM authentication, which means access using a token.
Description	A description of the application to be created.

4. Click **OK**.
   After the app is created, its name and ID are displayed in the app list.
5. Click an app name, and view the **AppKey** and **AppSecret** on the displayed app details page.

   Figure 11-10 App details page

Authorizing an API to an App

1. **Log in to the DGC console**, select an instance, and click **Access** on it.
   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Lake Mall** of the corresponding workspace.
   Navigate to the Data Lake Mall console.
3. Authorize an API to an app in either of the following ways:
   Giving API authorization:

 Copyright © Huawei Technologies Co., Ltd.

a. Choose **API Development** > **APIs**.
b. Locate the row that contains the API to be bound, and click **View**.
   On the displayed page, click **Authorize**.
c. 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 app. The API information page is displayed.
c. On the displayed page, click **Permission Application**.
d. Set an expiry, select an application, and click **OK**.
e. 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 app, 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 a bound API, click **Renew**.

Calling an API

DLM provides the application authentication mode for calling an API.

When the application authentication is used, the SDK is required for access.

Java, Go, Python, JavaScript, C#, PHP, C++, C, and Android SDKs are available.

For details on how to use the application authentication method, see the **DLM SDK Reference**.

11.14 Operating APIs

11.14.1 Creating Throttling Policies

Scenario

A throttling policy is used to limit the number of API calls within a specified period to protect backend services.

To ensure service stability, create throttling policies to control the access traffic of APIs.

Note that a throttling policy takes effect only after it is bound to an API.

 Copyright © Huawei Technologies Co., Ltd.

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. Log in to the DGC console, select an instance, and click **Access** on it.
   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Lake Mall** of the corresponding workspace.
   Navigate to the Data Lake Mall console.
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 11-16.

Figure 11-11 Creating a throttling policy

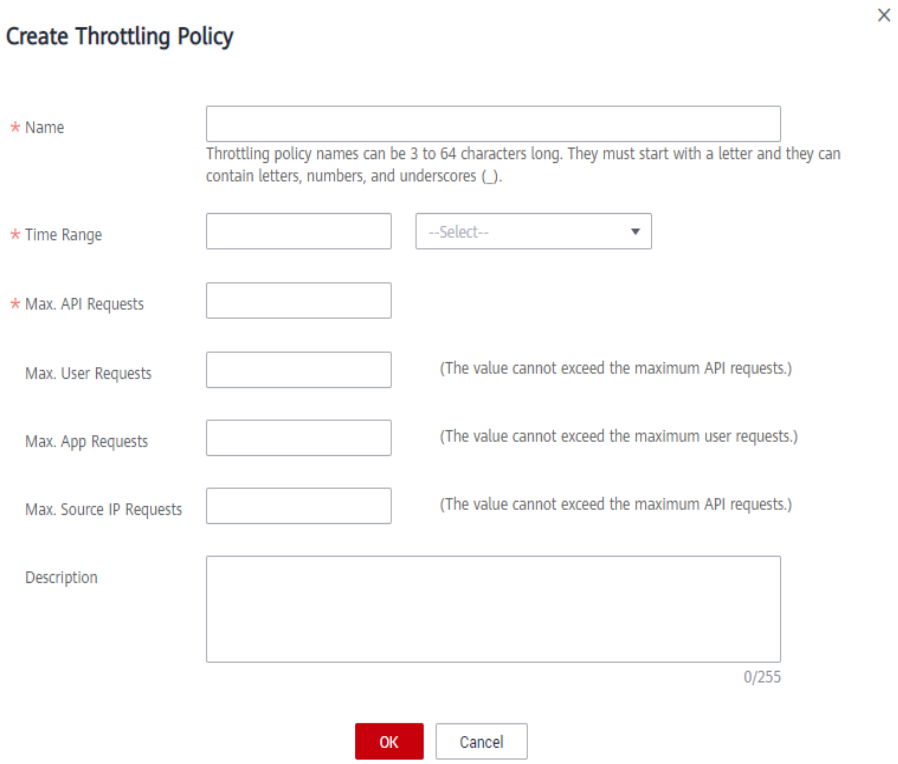


Table 11-16 Parameters

Parameter	Description
Name	The throttling policy name

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description
Time Range	The time duration for limiting the number of API calls • 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. • 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. • 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. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. Click **Data Lake Mall** of the corresponding workspace.

   Navigate to the Data Lake Mall console.

3. On the displayed page, choose **Throttling Policies** from the left navigation bar.

 Copyright © Huawei Technologies Co., Ltd.

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. **Log in to the DGC console**, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.
2. Click **Data Lake Mall** of the corresponding workspace.

   Navigate to the Data Lake Mall console.
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**.

11.15 Migrating Resources

To migrate data services created in one environment to another, you can use DGC resource migration function.

Prerequisites

- There are resources that can be migrated in the system. Create data connections by referring to **Creating Data Connections**.
- You have created OBS buckets and folders for data storage.

NOTE

Imported and exported resources are stored in JSON format.

 Copyright © Huawei Technologies Co., Ltd.

Exporting a Resource

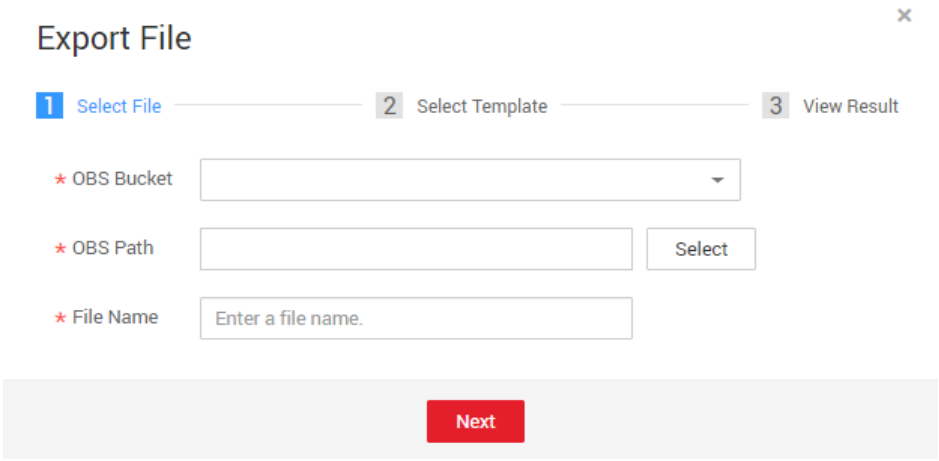
1. Log in to the DGC console, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. In the **Lifecycle** area, choose **Management Center** > **Resource Migration**. The **Resource Migration** page is displayed.

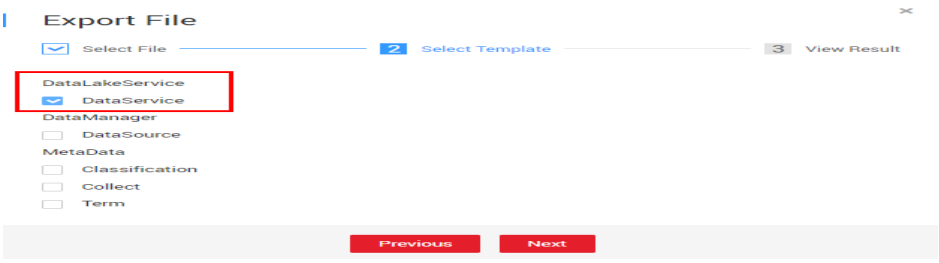
3. Click **Create Export Task** to configure the save path and file name.

   Figure 11-12 Configuring the save path and file name

4. Click **Next** and select the resource to be exported.

   Figure 11-13 Selecting the resource to be exported

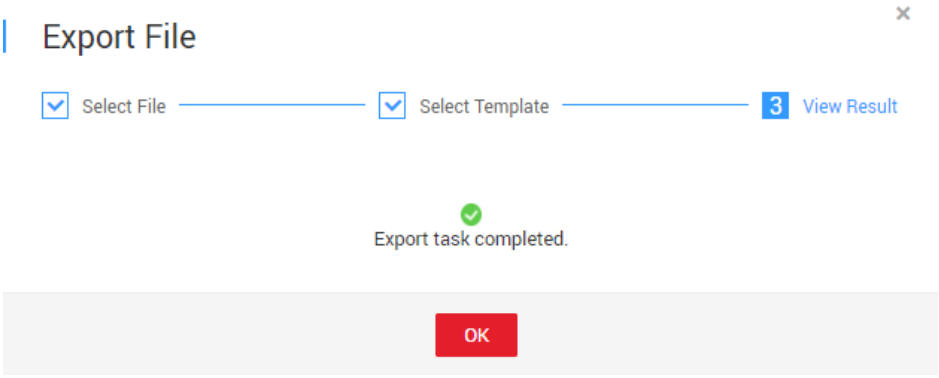
5. Click **Next** and wait until the export is complete.

   The exported resource is in JSON format and saved in the path specified in step 3.

   If no result is displayed within 1 minute, the export fails. Try again or contact after-sales personnel for technical support.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-14 Export complete



Importing a Resource

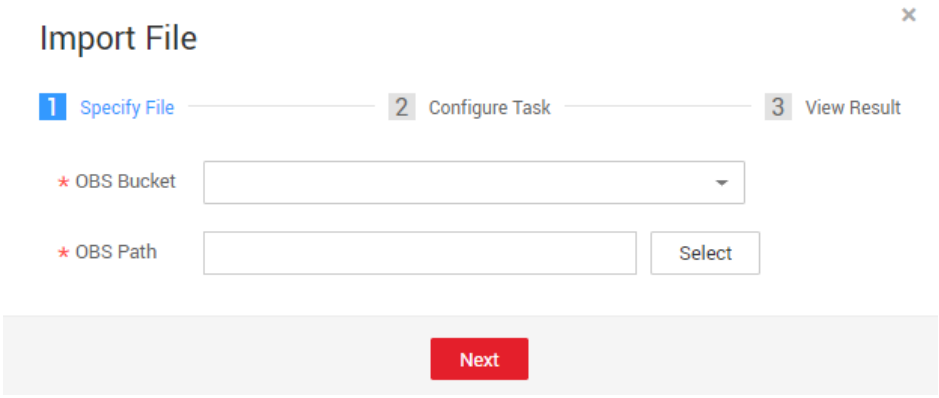
1. Log in to the DGC console, select an instance, and click **Access** on it.

   On the displayed page, click the **Workspace Management** tab. In the workspace list, locate the required workspace.

2. In the **Lifecycle** area, choose **Management Center** > **Resource Migration**. The **Resource Migration** page is displayed.

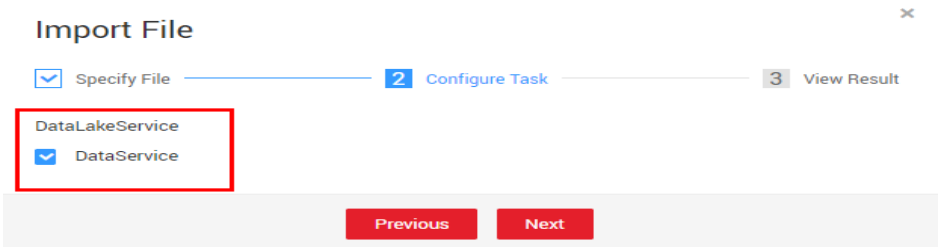
3. Click **Create Import Task** and configure the path for saving the resources to be imported.

   Figure 11-15 Configuring the storage path of the resource to be imported

4. Click **Next** and select the resource to be imported.

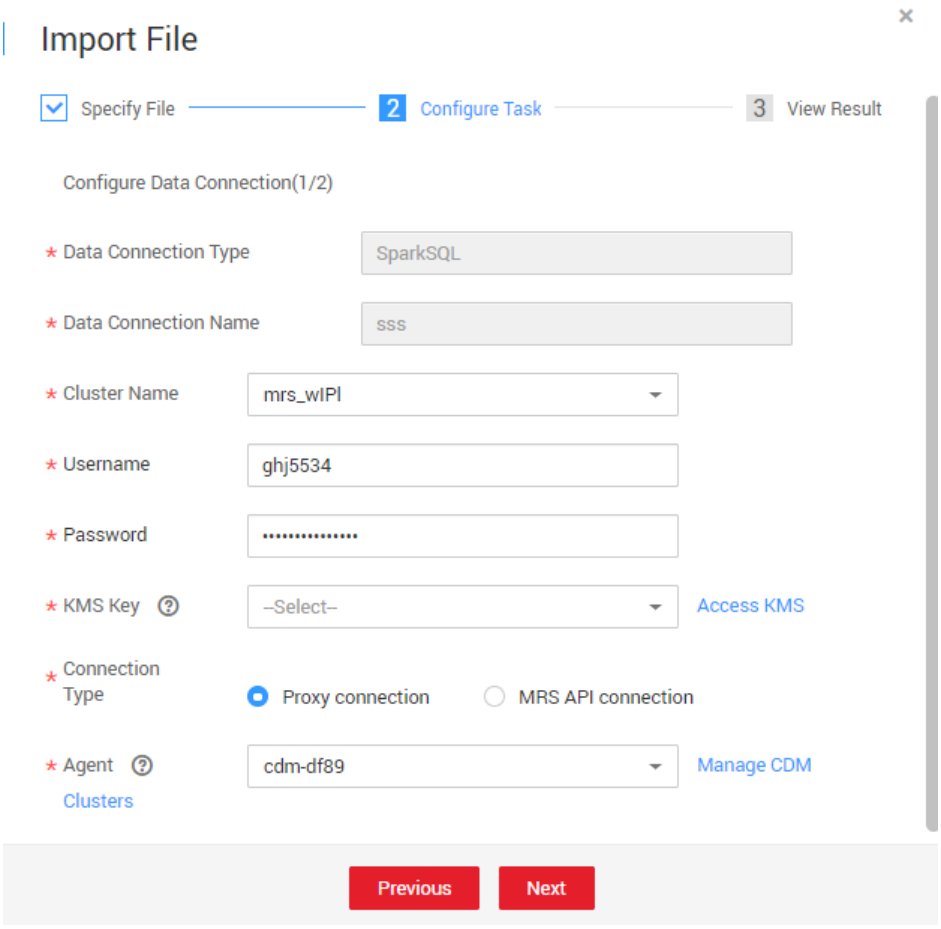
   Figure 11-16 Selecting the resource to be imported

 Copyright © Huawei Technologies Co., Ltd.

5. Click **Next** and configure a data connection. The number of data connections to be configured is the same as the number of data sources. You need to enter a new password each time.

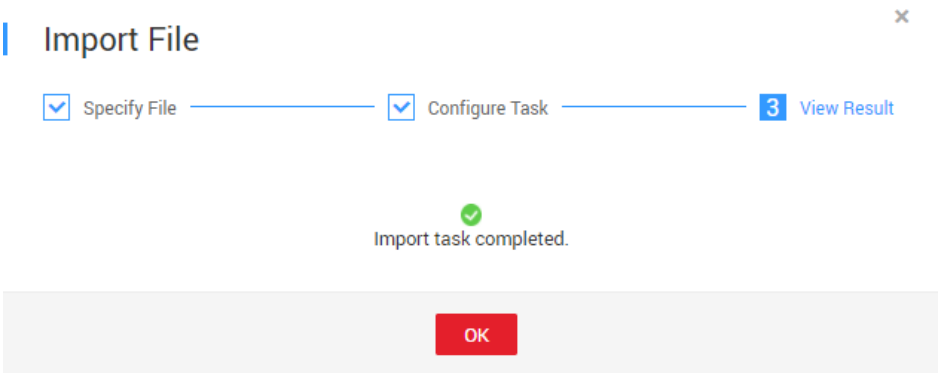
Figure 11-17 Configuring a data connection



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

   If no result is displayed within 1 minute, the import fails. Try again or contact after-sales personnel for technical support.

Figure 11-18 Import complete



 Copyright © Huawei Technologies Co., Ltd.

A Change History

Released On	Description
2020-11-5	This is the thirteenth official release. • Added the following sections: – **Debugging APIs** – **Authorizing APIs** – **Suspending APIs** – **Deleting APIs** • Modified the following sections: – Added how ro debug an API in **Debugging APIs**. – Added cluster status descriptions in **Buying Exclusive Instances**. – Added the timeout interval for collecting a single data table in **Task Management**. – Updated the screenshots in **Configuration Center**. – Added the description that the standard name and code must be unique. See **Data Standards**. – Added the **Generate Null** parameter in **Creating Data Quality Jobs**. – Optimized source-end decryption and destination-end encryption in **Encryption and Decryption During File Migration**. – Updated the screenshots in **Creating Rule Templates**. – Modified periodic scheduling parameters in **Creating Data Quality Jobs** and **Create Comparison Jobs**. – Updated the screenshots and descriptions in **Viewing Rule Instances**. – Optimized the script parameters in **Developing an SQL Script** and **Developing a Shell Script**.

 Copyright © Huawei Technologies Co., Ltd.

Released On	Description
2020-10-27	This is the twelfth official release. ● Added the following sections: – Added **Managing Drivers**. ● Modified the following sections: – Added the subject levels in **Configuration Center**.

 Copyright © Huawei Technologies Co., Ltd.

Released On	Description
2020-08-07	This is the eleventh official release. • Added the following sections: – Added how to export and import data service resources from the workspace. The resources include API data and traffic control data. For details, see **Migrating Resources**. – Added how to change the path for storing job logs. For details, see **Changing Log Storage Paths**. – Added how to configure agencies. After an agency is configured, a job can interact with other services as an agency during job execution. For details, see **Configuring Agencies**. – Added the description of MRS Flink node. For details, see **MRS Flink Job**. • Modified the following sections: – Added how to configure workspace-level and job-level agencies in section **Creating a Job**. – Added data service migration objects in section **Migrating Resources**. – Added the operation guide for associating dimension tables with quality rules in section **Managing Dimension Tables**. – Added the operation guide for associating fact tables with quality rules and data standards, and operation guide for adding fields (dimension and measurement type) to the fact tables in section **Creating Fact Tables**. – Added the operation guide for associating summary tables with quality rules and data standards in section **Creating Summary Tables**. – Supported rule template upgrade in section **Creating Rule Templates**. – Added support for determining rule repetition and generating dirty data based on tables and fields in section **Creating Data Quality Jobs**. – Added the Oracle data source for the metadata collection tasks in section **Task Management**. – Added the minimum retention period of basic API configuration parameters and the RDS data source for DLM in section **Generating APIs** and **Registering APIs**, respectively. – Added the description about how to cancel an application for reviewing an API in section **Reviewing APIs**.

Copyright © Huawei Technologies Co., Ltd.

Released On	Description
2020-06-30	This is the tenth official release. • Added the following sections: – **Managing Workspaces** – **Managing Members** – **Data Sources** – **Data Sources Supported by DIS** – **Comparison Between Shared Edition and Exclusive Edition** – **Buying Exclusive Instances** – Added the cross-space job scheduling guide. For details, see **Cross-Workspace Job Scheduling**. – Added OBSUtil embedded objects. For details, see **OBSUtil Embedded Objects**. – **Create Comparison Jobs** • Modified the following sections: – Added dimension interaction to the fact table. For details, see **Creating Fact Tables**. – Modified the summary table interaction. For details, see **Creating Summary Tables**. – Added the subject directory to the dimension modeling list page. For details, see **Creating Dimensions**. – Enhanced atomic metrics. For details, see **Creating Atomic Metrics**. – Enhanced derivative metrics. For details, see **Creating Derivative Metrics**. – Added compound metrics. For details, see **Creating Compound Metrics**. – Provided the derivative metric SQL template. Asset synchronization is case insensitive during function configuration. For details, see **Configuration Center**. – Deleted general filter groups and general filers for Data Design. Old data is still compatible. – Supported notifications of multiple SMN topics. For details, see **Managing an Notification**. – Allowed CDM job content saving. – Supported YARN queue scheduling in sections **MRS SparkSQL**, **MRS Hive SQL**, and **MRS Spark**. – The basic operators supported by EL expressions are deleted. – Updated the GUI style and added statistical parameters. For details, see **Overview**. – Masked the domain name binding function and instance creation function off. Added the service catalog

Copyright © Huawei Technologies Co., Ltd.

Released On	Description
	visibility setting. For details, see **Generating Wizard APIs**. – Added service catalog visibility parameters and the pagination function. For details, see **Generating Script APIs**. – Added app categories. For details, see **Creating an Application and Getting Authorized**. – Added the local import function in the scenario without OBS. For details, see **Importing a Script**. – Added the local import function in the scenario with OBS. For details, see **Importing a Job**. – Added the local import function in the scenario without OBS. For details, see **Importing Environment Variables**.

Copyright © Huawei Technologies Co., Ltd.

Released On	Description
2020-05-08	This is the ninth official release. • Added the following sections: – Added the overview page description. For details, see **Overview**. – Supported scenario migration. For details, see **Scenario-based Migration**. – Added the OCR operator. For details, see **OCR**. – Added the reverse database supported by the logical model. For details, see **Reversing Databases**. – Added publish history, version comparison, and publish logs. For details, see **Viewing Tables**. • Modified the following sections: – Supported the creation of DIS links. For details, see **Creating Data Connections**. – Supported the migration of an entire Hive database to DWS. For details, see **Supported Data Sources**. – Supported DWS and MRS_Hive connections by reverse databases. For details, see **Managing Lookup Tables**. – Supported the conversion of logical models to physical models. For details, see **Designing Logical Models**. – Supported the modification of the OBS path format. For details, see **Designing Physical Models**. – Supported the code table dimensions and hierarchy dimensions. When a dimension is published, the corresponding dimension table is published as well. For details, see **Creating Dimensions**. – Changed the compound metric to derivative metric. For details, see **Creating Derivative Metrics**. – Enhanced the DDL template management function. Field type management and mapping configuration management of table field types in different data warehouses are supported. For details, see **Configuration Center**. – Supported the script execution result download and permission control. For details, see **Developing an SQL Script**. – Supported the cross-account job scheduling. For details, see **DIS Client**. – Supported for user confirmation on automatic data classification. Foe details, see **Task Management** and **Task Monitoring**. – Supported metric management in data catalogs. For details, see **Dashboard**.

Copyright © Huawei Technologies Co., Ltd.

Released On	Description
2020-03-31	This is the eighth official release. • Modified the following sections: – Added how to create, import, and export business processes. For details, see **Designing Processes**. – Added how to automatically generate data quality rules based on data standards. For details, see **Data Standards**. – Added how to import and export all standards and lookup tables with directories. For details, see **Managing Lookup Tables**. – Supported DWS and MRS_HIVE ER models. You can export specified ER models. ER models can be visualized as well. For details, see **Designing Physical Models**. Users also can manage the publish history of a single table (version management). – Added the publish history management. For details, see **Managing Dimension Tables**. – Added the publish history management. For details, see **Creating Fact Tables**. – Added the publish history management. For details, see **Creating Summary Tables**. – Added the **Offline** parameter. For details, see **Import GES**. – Added the **Subscription Stream** function. For details, see **Subscribing to Event Notification**. – Added how to grant the permissions required to download data to other users. For details, see **Managing Authorization**. – Added how to migrate the **Host Connection** function to DGC Management Center. For details, see **Managing a Host Connection**. – Added how to select cross-source dependency packages for the DLI Spark operators. For details, see **DLI Spark**. – Added the Shell and MRS Kafka data connections. For details, see **Creating Data Connections**. – Added support for creating data quality jobs using custom SQL statements and generating dirty data in **Creating Data Quality Jobs**. – Added how to manage instances. For details, see **Glossary**. – Added non-authentication, blacklist and whitelist, and IAM token authentication. For details, see **Generating Wizard APIs**. Data service catalogs of any levels are supported. • Added the following sections:

Copyright © Huawei Technologies Co., Ltd.

Released On	Description
	– Added logical models. For details, see **Designing Logical Models**. – Added the parameters of the MRS Kafka node. For details, see **MRS Kafka**. – Added the parameters of the MRS Kafka node. For details, see **Kafka Client**.
2020-03-06	This is the seventh official release. • Added DLI Flink job parameters. For details, see **DLI Flink Job**. • Added how to stop subsequent nodes and modify all nodes if a node fails to run. • Added MRS security clusters. For details, see **Creating Data Connections**. • Added job labels. For details, see **Managing Job Labels**. • Added the FOR operator. For details, see **For Each**. • Added DLI Spark Python job development. For details, see **Developing a Job**. • Added parameters of Hive/Spark SQL nodes. For details, see **MRS SparkSQL** and **MRS Hive SQL**. • Added data catalog search based on levels and categories. For details, see **Data Catalogs**. • Added the DLI data table permission control. For details, see **Data Permissions**. • Optimized the display of metadata details. For details, see **Task Management**.
2020-1-16	This is the sixth official release. • Supported the update of DLI, OBS, and DWS profiles. For details, see **Data Catalogs**. • Supported that DLI MySQL can filter tables to be collected based on the regular expressions. For details, see **Task Management**. • Added the process design. For details, see **Designing Processes**. • Added business metric management. For details, see **Business Metrics**. • Added technical metric management. For details, see **Technical Metrics**.

Copyright © Huawei Technologies Co., Ltd.

Released On	Description
2019-12-03	This is the fifth official release. • Supported the agent management function. For details, see **Agent Management**. • Modified the data retention period. For details, see **Quick Start**. • Added custom rules for data quality. For details, see **Creating Rule Templates**.
2019-11-04	This is the fourth official release. Added "Generating API to Rank Parameters"; Modified **Generating Wizard APIs** and **Generating Script APIs**
2019-10-18	This is the third official release. Added "Permissions Control" and "Workspaces"; Modified **Data Design**, **Data Integration**, **Data Development** and all entries
2019-09-26	This is the second official release. Deleted OBS from the **Triggering Event Type** options. For details, see **Developing a Job**, **Job Embedded Objects**, and **Monitoring a Real-Time Job**.
2019-09-11	This is the first official release.

Copyright © Huawei Technologies Co., Ltd.