Data Lake Governance Center

FAQs

Issue	01
Date	2020-12-15



HUAWEI

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.

Huawei Technologies Co., Ltd.

Address: Huawei Industrial Base
Bantian, Longgang
Shenzhen 518129
People's Republic of China

Website: https://www.huawei.com

Email: support@huawei.com

Contents

1 Product Consulting....1
1.1 Regions and AZs....1
1.2 How Is Data Integrated into DGC?....3
1.3 What Is the Relationship Between DGC and ROMA?....3
1.4 What Is the Relationship Between DGC and Huawei Horizon Digital Platform?....3
1.5 Does DGC Support Private Cloud?....3
1.6 How Do I Create a Fine-Grained Permission Policy in IAM?....3
1.7 How Do Enterprises Prevent Data Leakage During Data Governance?....3
2 Version and Usage....5
2.1 Can I Try DGC for Free?....5
2.2 How Do I Renew My Instance When It Is About to Expire?....6
2.3 Why Are the DIS Service Fees Included in DGC?....6
2.4 What Can I Do If I Cannot Select an IAM Project When I Buy a DGC Instance?....6
3 Management Center....8
3.1 What Data Connections Are Supported?....8
3.2 What Are the Precautions for Creating Data Connections?....8
3.3 Why Do DWS/Hive/HBase Data Connections Fail to Obtain the Information About Database or Tables?....8
3.4 Why Are MRS Hive/HBase Clusters Not Displayed on the Page for Creating Data Connections?....8
3.5 What Should I Do If the Connection Test Fails When I Enable the SSL Connection During the Creation of a DWS Data Connection?....9
4 Data Integration....10
4.1 What Is the Migration Performance in the Same VPC and Different VPCs?....10
4.2 Why Is Error ORA-01555 Reported During Migration from Oracle to DWS?....10
4.3 What Should I Do If the MongoDB Connection Migration Fails?....11
4.4 Why Does the Migration Fail When the Source End Keeps Changing During the Migration in the HBase Scenario?....12
5 Data Design....13
5.1 What Is the Relationship Between Lookup Tables and Data Standards?....13
5.2 What Is the Difference Between ER Modeling and Dimensional Modeling?....13
5.3 What Data Modeling Methods Are Supported by Data Design?....13
5.4 How Can I Use Standardized Data?....14

 Copyright © Huawei Technologies Co., Ltd.

5.5 Does Data Design Support Database Reverse? 14

6 Data Development 15

6.1 How Many Jobs Can Be Created in Data Development? Is There a Limit on the Number of Nodes in a Job? 16

6.2 How Can I Quickly Rectify a Deleted CDM Cluster Associated with a Job? 16

6.3 Why Is There a Large Difference Between Job Execution Time and Start Time of a Job? 16

6.4 Will Subsequent Jobs Be Affected If a Job Fails to Be Executed During Scheduling of Dependent Jobs? What Should I Do? 17

6.5 What Do I Do If Node Error Logs Cannot Be Viewed When a Job Fails? 17

6.6 What Should I Do If the Agency List Fails to Be Obtained During Agency Configuration? 17

6.7 How Do I Locate Job Scheduling Nodes with a Large Number? 18

6.8 Why Cannot Specified Peripheral Resources Be Selected When a Data Connection Is Created in Data Development? 19

6.9 Why Cannot I Receive a Job Failure Alarm Notification After SMN Is Configured? 19

6.10 Why Is There No Job Running Scheduling Log on the Monitor Instance Page After Periodic Scheduling Is Configured for a Job? 20

6.11 Why Does the GUI Display Only the Failure Result but Not the Specific Error Cause After Hive SQL and Spark SQL Scripts Fail to Be Executed? 20

6.12 What Do I Do If the Token Is Invalid During the Running of a Data Development Node? 21

6.13 Why Cannot I View the Existing Workspaces After I Have the Required Policy? 21

6.14 How Do I View Run Logs After a Job Is Tested? 21

6.15 Why Does a Job Scheduled by Month Start Running Before the Job Scheduled by Day Is Complete? 21

6.16 How Do I Execute Presto SQL in Data Development? 22

6.17 What Should I Do If Invalid Authentication Is Reported When I Run a DLI Script? 22

6.18 Why Cannot I Select the Desired CDM Cluster in Proxy Mode When Creating a Data Connection? 22

7 Data Assets 23

7.1 What Are the Functions of the Data Assets Module? 23

7.2 What Assets Can Be Collected? 23

7.3 What Is Data Lineage? 23

7.4 How Can Data Lineage Be Displayed on a Data Map? 24

8 Data Lake Mall 25

8.1 What Languages Do Data Lake Mall SDKs Support? 25

9 Data Security 26

9.1 Why Is Data in a Data Table Not Masked Based on Rules After a Data Masking Task Is Executed? 26

9.2 Why Does the System Display a Message Indicating that Some Data Identification Rules Are in Use When They Are Deleted Although No Task Is Using Them? 26

9.3 What Should I Do If Authentication Audit Logging Is Not Enabled? 26

 Copyright © Huawei Technologies Co., Ltd.

1 Product Consulting

1.1 Regions and AZs

1.2 How Is Data Integrated into DGC?

1.3 What Is the Relationship Between DGC and ROMA?

1.4 What Is the Relationship Between DGC and Huawei Horizon Digital Platform?

1.5 Does DGC Support Private Cloud?

1.6 How Do I Create a Fine-Grained Permission Policy in IAM?

1.7 How Do Enterprises Prevent Data Leakage During Data Governance?

1.1 Regions and AZs

Concepts

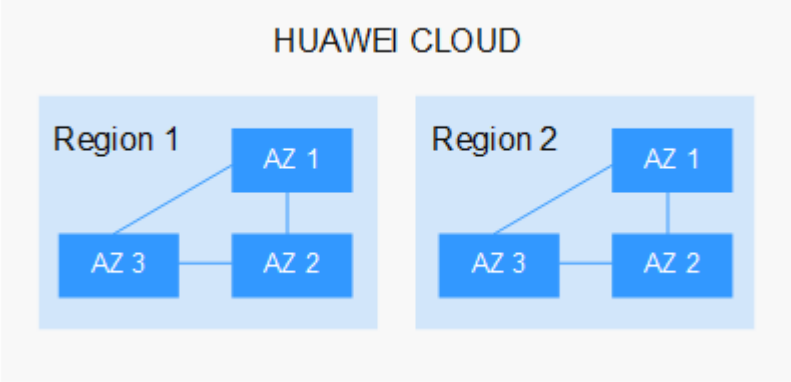
A region and availability zone (AZ) identify the location of a data center. You can create resources in a specific region and AZ.

- Regions are divided from the dimensions of geographical location and network latency. Public services, such as Elastic Cloud Server (ECS), Elastic Volume Service (EVS), Object Storage Service (OBS), Virtual Private Cloud (VPC), Elastic IP (EIP), and Image Management Service (IMS), are shared within the same region. Regions are classified as universal regions and dedicated regions. A universal region provides universal cloud services for common tenants. A dedicated region provides services of the same type only or for specific tenants.
- An AZ contains one or more physical data centers. Each AZ has independent cooling, fire extinguishing, moisture-proof, and electricity facilities. Within an AZ, computing, network, storage, and other resources are logically divided into multiple clusters. AZs within a region are interconnected using high-speed optical fibers to allow you to build cross-AZ high-availability systems.

Figure 1-1 shows the relationship between regions and AZs.

 Copyright © Huawei Technologies Co., Ltd.

Figure 1-1 Regions and AZs



HUAWEI CLOUD provides services in many regions around the world. You can select a region and AZ as needed. For more information, see **Global Products and Services**.

Region Selection

When selecting a region, consider the following factors:

- Location

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

  The countries and regions outside the Chinese mainland, such as Bangkok and Hong Kong, provide services for users outside the Chinese mainland. If you or your target users are in the Chinese mainland, these regions are not recommended due to high access latency.

- Relationship between cloud services

  Cloud services in different regions cannot communicate with each other through an internal network.

  For example, if DGC (including Management Center and CDM) needs to communicate with MRS and OBS in different regions, a public network or Direct Connect is required. If they are in the same region, instances in the same subnet and security group can communicate with each other by default.

- Resource pricing

  Resource pricing may vary in different regions. For details, see **Product Pricing Details**.

AZ Selection

When determining whether to deploy resources in the same AZ, consider your applications' requirements on disaster recovery (DR) and network latency.

- For high DR capability, deploy resources in different AZs in the same region.
- For low network latency, deploy resources in the same AZ.

 Copyright © Huawei Technologies Co., Ltd.

Regions and Endpoints

Before using an API to call resources, specify its region and endpoint. For details on HUAWEI CLOUD regions and endpoints, see **Regions and Endpoints**.

1.2 How Is Data Integrated into DGC?

DGC supports **batch data migration** and **real-time data ingestion**. It provides efficient access to more than 20 heterogeneous data sources, wizard-guided configuration and management, and integration of single tables, entire databases, data that fluctuates periodically according to a certain rule, and data added in each time segment.

1.3 What Is the Relationship Between DGC and ROMA?

Real-time Open Multi-Cloud Agile (ROMA) serves as a channel linking various systems. It does not govern or plan the ingested data. DGC analyzes the structure of ingested data and re-models the data to eliminate data silos and help enterprises build unified data models.

1.4 What Is the Relationship Between DGC and Huawei Horizon Digital Platform?

DGC is a data enablement module of Huawei Horizon Digital Platform. It helps you better manage and use data.

1.5 Does DGC Support Private Cloud?

DGC can be connected to HCS Online, providing high-quality services for customers.

For more information about HCS Online, see **HCS Online**.

1.6 How Do I Create a Fine-Grained Permission Policy in IAM?

Currently, fine-grained permission policies are not supported in DGC.

1.7 How Do Enterprises Prevent Data Leakage During Data Governance?

1. **Manage data hierarchically.** Enterprises should take the national security, public rights and interests, personal privacy, and legitimate enterprise interests into full account when formulating data classification standards and create global data asset catalogs.

 Copyright © Huawei Technologies Co., Ltd.

2. **Perform fine-grained authorization.** Enterprises can adopt differentiated control measures for data of different levels to implement refined data management. Permission authorization must comply with the principle of "minimum user authorization and full-process protection".
3. **Properly manage data sharing.** Enterprise can standardize the data sharing process to ensure that data users apply for data based on business needs on the premise of legal compliance and security assurance. The data owner reviews and determines the data usage scope and sharing mode based on rules, for the purpose of orderly data transfer and secure application through the data exchange mechanism.
4. **Audit and identify risks.** Enterprises need to record data usage to ensure that data user actions can be traced and audited throughout the process. In addition, enterprises should fully evaluate potential risks, establish a dynamic and real-time risk identification and alarm mechanism, and detect and handle risks timely.
5. **Implement security management throughout the data lifecycle.** Enterprises need to enhance data lifecycle security management to prevent user data leakage, tampering, and abuse. For example, enterprises need to honestly inform users of the purpose, method, and scope of data collection and usage and collect data only after obtaining user authorization. In the storage phase, technologies such as feature extraction and labeling are used to anonymize original information, separate the information from sensitive information, and store the information independently to implement strict access control and reduce data leakage risks. In the data usage phase, technologies such as model computing and multi-party security computing are used to provide only anonymized calculation results without collecting and sharing raw data.

 Copyright © Huawei Technologies Co., Ltd.

2 Version and Usage

2.1 Can I Try DGC for Free?

2.2 How Do I Renew My Instance When It Is About to Expire?

2.3 Why Are the DIS Service Fees Included in DGC?

2.4 What Can I Do If I Cannot Select an IAM Project When I Buy a DGC Instance?

2.1 Can I Try DGC for Free?

Yes, you can. For new users, DGC Basic provides a 30-day free trial period.

You can log in to the DGC management console and click **Try Basic** to apply for a free trial of DGC instances. Select a region with caution because resources in different regions cannot communicate with each other.

Precautions for free trial are as follows:

1. Each account can experience the free trial only once.
2. A basic DGC instance for free trial does not include CDM clusters. To use CDM, buy a CDM incremental package. For details, see **Buying DGC Incremental Packages**.
3. DGC Basic is available during the 30-day free trial period. After the trial period expires, you need to buy an instance of one of the official versions provided.
4. During the trial period, you can buy any official version of DGC. The trial period expires immediately.
5. DGC instances created during the trial period cannot be accessed when the trial period expires. After the trial period expires, the instances are retained for seven days and will be automatically deleted seven days later. You shall take full responsibility for any losses resulting from not buying a DGC instance in time. To access the free trial, follow the following instructions:

 Copyright © Huawei Technologies Co., Ltd.

2.2 How Do I Renew My Instance When It Is About to Expire?

When a free trial instance is about to expire, you can buy an instance of any official versions provided. Log in to the DGC management console, find the free trial instance that is about to expire, and click **Buy**.

For details on how to buy a DGC instance, see **Buying DGC Instances**. If you want to retain the resources and data of the original instance when buying a new one, pay attention to the following points:

- The region of the bought instance must be the same as that of the free trial instance.
- You need to buy an instance of the basic or a later version.
- By default, the resources of the trial instance are moved to the first instance you bought.

  NOTE

  Note that if you buy a basic instance, the resources and data of the Data Development and Data Integration modules in the original free trial instance will be retained, but other resources will be deleted.

When you buy an instance, the system automatically creates a CDM cluster in it. After the instance is bought, the CDM cluster will show up on the **Cluster Management** page in Data Integration.

2.3 Why Are the DIS Service Fees Included in DGC?

On February 18, 2020, HUAWEI CLOUD has notified users via emails, SMS messages, and internal messages that "HUAWEI CLOUD plans to put the advanced DIS stream into commercial use at 00:00 (Beijing time) on March 1, 2020." In addition, the advanced DIS stream has been integrated into DGC. Therefore, the fees of the advanced DIS stream displayed to customers are included in DGC.

If you no longer need to use an advanced DIS stream, delete related resources to avoid additional fees.

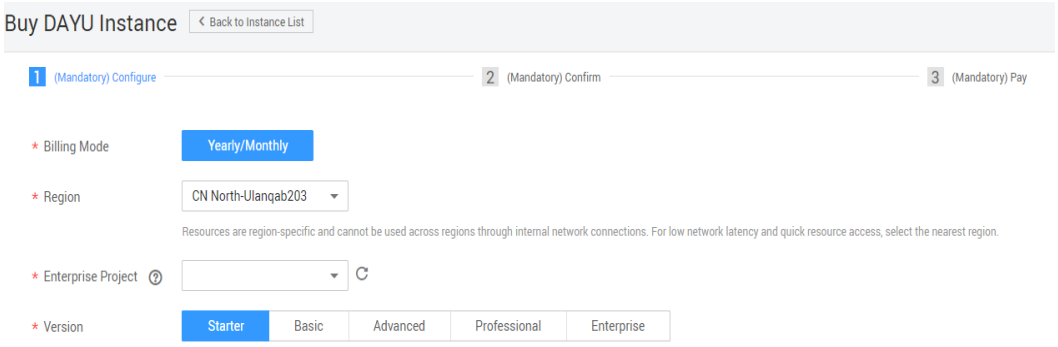
2.4 What Can I Do If I Cannot Select an IAM Project When I Buy a DGC Instance?

Check whether the current account has enabled the enterprise project function.

The enterprise project and IAM project cannot be enabled at the same time. If the enterprise project is enabled, you can buy only one instance in this enterprise project.

 Copyright © Huawei Technologies Co., Ltd.

Figure 2-1 Buying a DGC instance



 Copyright © Huawei Technologies Co., Ltd.

3 Management Center

3.1 What Data Connections Are Supported?

3.2 What Are the Precautions for Creating Data Connections?

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

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

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

3.1 What Data Connections Are Supported?

For details on the data connections supported by DGC, see **Data Sources**.

3.2 What Are the Precautions for Creating Data Connections?

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

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

That is probably because CDM clusters are disabled.

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

Possible causes are as follows:

 Copyright © Huawei Technologies Co., Ltd.

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

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

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

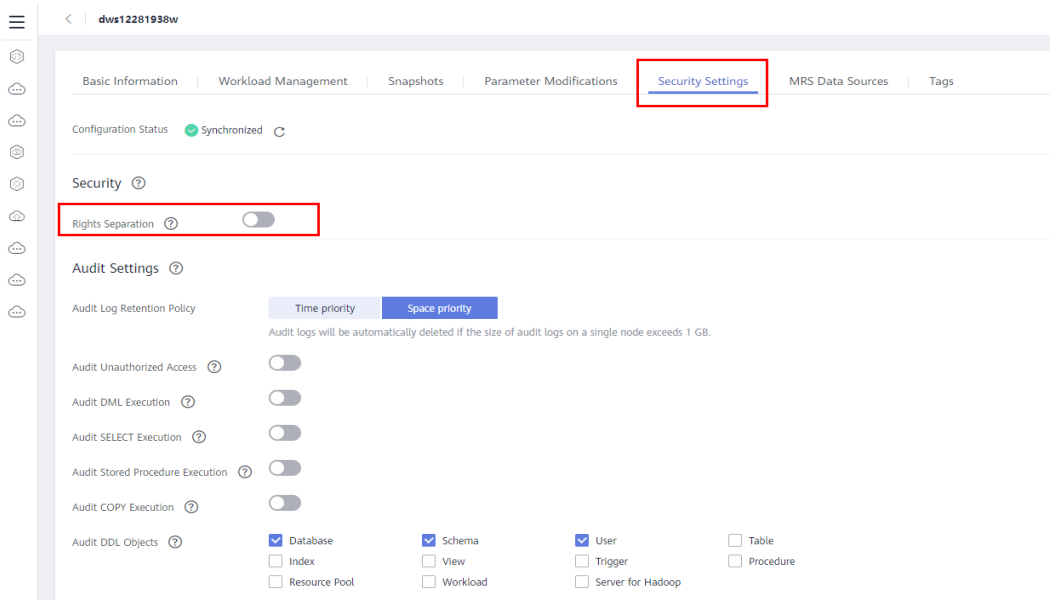
Question

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

Answer

On the DWS console, click the corresponding cluster, choose **Security Settings**, and disable **Rights Separation**.

Figure 3-1 Disabling **Rights Separation** for a DWS cluster



 Copyright © Huawei Technologies Co., Ltd.

4 Data Integration

4.1 What Is the Migration Performance in the Same VPC and Different VPCs?

4.2 Why Is Error ORA-01555 Reported During Migration from Oracle to DWS?

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

4.4 Why Does the Migration Fail When the Source End Keeps Changing During the Migration in the HBase Scenario?

4.1 What Is the Migration Performance in the Same VPC and Different VPCs?

The transmission rate depends on the bandwidth and file read/write speed.

4.2 Why Is Error ORA-01555 Reported During Migration from Oracle to DWS?

Symptom

When CDM is used to migrate Oracle data to DWS, an error is reported, as shown in **Figure 4-1**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 4-1 Symptom

```
2020-09-21 22:51:02,991 ERROR LocalJobRunner Map Task #3 [org.apache.sqoop.common.SqoopException:111] SqoopException
java.sql.SQLException: ORA-01555: snapshot too old: rollback segment number 3 with name "_SYSSMU3_20976775318" too small

    at oracle.jdbc.driver.T4CTTIoer11.processERROR(T4CTTIoer11.java:494)
    at oracle.jdbc.driver.T4CTTIoer11.processERROR(T4CTTIoer11.java:446)
    at oracle.jdbc.driver.T4C80all.processERROR(T4C80all.java:1054)
    at oracle.jdbc.driver.T4CTTIfun.receive(T4CTTIfun.java:623)
    at oracle.jdbc.driver.T4CTTIfun.doRPC(T4CTTIfun.java:252)
    at oracle.jdbc.driver.T4C80all.doOALL(T4C80all.java:612)
    at oracle.jdbc.driver.T4CPreparedStatement.doOall8(T4CPreparedStatement.java:226)
    at oracle.jdbc.driver.T4CPreparedStatement.fetch(T4CPreparedStatement.java:1023)
    at oracle.jdbc.driver.OracleStatement.fetchMoreRows(OracleStatement.java:3353)
    at oracle.jdbc.driver.InsensitiveScrollableResultSet.fetchMoreRows(InsensitiveScrollableResultSet.java:736)
    at oracle.jdbc.driver.InsensitiveScrollableResultSet.absoluteInternal(InsensitiveScrollableResultSet.java:692)
    at oracle.jdbc.driver.InsensitiveScrollableResultSet.next(InsensitiveScrollableResultSet.java:406)
    at org.apache.sqoop.connector.jdbc.sql.impl.WrapResultSet.next(WrapResultSet.java:36)
    at org.apache.sqoop.connector.jdbc.GenericJdbcExtractor.extractObjectRecord(GenericJdbcExtractor.java:151)
    at org.apache.sqoop.connector.jdbc.GenericJdbcExtractor.extract(GenericJdbcExtractor.java:129)
    at org.apache.sqoop.connector.jdbc.GenericJdbcExtractor.extract(GenericJdbcExtractor.java:59)
    at org.apache.sqoop.job.mr.SqoopMapper.runInternal(SqoopMapper.java:184)
    at org.apache.sqoop.job.mr.SqoopMapper.run(SqoopMapper.java:81)
    at org.apache.hadoop.mapred.MapTask.runNewMapper(MapTask.java:799)
    at org.apache.hadoop.mapred.MapTask.run(MapTask.java)
    at org.apache.hadoop.mapred.LocalJobRunner$Job$MapTaskRunnable.run(LocalJobRunner.java:271)
    at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
    at java.util.concurrent.FutureTask.run(FutureTask.java:266)
    at org.apache.sqoop.submission.mapreduce.MapperExecutorGroup$1.lambda$execute$0(MapperExecutorGroup.java:222)
    at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
    at java.util.concurrent.FutureTask.run(FutureTask.java:266)
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
    at java.lang.Thread.run(Thread.java:748)
Caused by: oracle.jdbc.OracleDatabaseException: ORA-01555: snapshot too old: rollback segment number 3 with name "_SYSSMU3_20976775318" too small

    at oracle.jdbc.driver.T4CTTIoer11.processERROR(T4CTTIoer11.java:498)
    ... 28 common frames omitted
```

Cause Analysis

1. During data migration, if the entire table is queried and the table contains a large amount of data, the query takes a long time.
2. During the query, other users frequently perform the **commit** operation.
3. The RBS (the tablespace used for rollback) of Oracle is small. As a result, the migration task is not complete, the source database has been updated, and the rollback times out.

Summary and Suggestions

1. Reduce the data volume queried each time.
2. Modify the database configurations to increase the RBS of the Oracle database.

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

By default, the **userAdmin** role has only the permissions to manage roles and users and does not have the read and write permissions on a database.

If the MongoDB connection fails to be migrated, you need to view the user permission information in the MongoDB connection to ensure that the user has the read and write permissions on the specified database.

 Copyright © Huawei Technologies Co., Ltd.

4.4 Why Does the Migration Fail When the Source End Keeps Changing During the Migration in the HBase Scenario?

In the HBase scenario, the migration is performed based on the snapshot file of the HBase table. After the snapshot file is generated, if the source-end data changes, the file cannot be found. As a result, the migration fails.

When an error occurs for the first time, CDM migrates the files that do not change. For files that have changed, you need to retry the migration task.

Copyright © Huawei Technologies Co., Ltd.

5 Data Design

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

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

5.3 What Data Modeling Methods Are Supported by Data Design?

5.4 How Can I Use Standardized Data?

5.5 Does Data Design Support Database Reverse?

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

A lookup table consists of the names, codes, and data types of multiple table fields. The table fields in a code table can be associated with a data standard, and the data standard is applied to the fields in a model table.

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

ER modeling complies with 3NF modeling. Dimensional modeling mainly refers to the design of fact tables and dimension tables. Dimensional modeling is mainly used to implement multi-angle and multi-layer data query and analysis.

DGC is a data lake operations platform. Dimensional modeling is used more frequent.

5.3 What Data Modeling Methods Are Supported by Data Design?

DGC Data Design supports entity relationship (ER) modeling and dimensional modeling:

- **ER modeling**

 Copyright © Huawei Technologies Co., Ltd.

ER modeling describes the business activities within an enterprise. Compliant with the third normal form (3NF), ER modeling is designed for data integration. It is used for combining and merging data with similarities by subject. ER modeling results cannot be used directly for decision-making, but they are a useful tool.

You can divide ER modeling into three levels of abstraction: design conceptual models, logical models, and physical models.

- **Conceptual model**: A conceptual model is a representation of business processes and business data involved in different activities. It can be used to represent the relationships between business entities.
- **Logical model**: A logical model is more detailed than a conceptual model. It is used to outline the entities, attributes, and relationships of a business. It enables communication between IT and business staff. A logical model is a set of standardized logic table structures. Determined by business rules, a logical model outlines business objects, data items of the business objects, and relationships between business objects.
- **Physical model**: A physical model is based on logical models and is used to design the database architecture for data storage with a range of technical factors all considered. For example, the selected data warehouse could be defined as DWS or 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 that consists of numeric measurement metrics. The fact table is associated with a group of dimensional tables that contain description attributes through primary or foreign keys.

  Typical dimensional models include star models and snowflake models used in some special scenarios.

  In DGC Data Design, 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.

5.4 How Can I Use Standardized Data?

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

5.5 Does Data Design Support Database Reverse?

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

 Copyright © Huawei Technologies Co., Ltd.

6 Data Development

6.1 How Many Jobs Can Be Created in Data Development? Is There a Limit on the Number of Nodes in a Job?

6.2 How Can I Quickly Rectify a Deleted CDM Cluster Associated with a Job?

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

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

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

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

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

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

6.9 Why Cannot I Receive a Job Failure Alarm Notification After SMN Is Configured?

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

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

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

6.13 Why Cannot I View the Existing Workspaces After I Have the Required Policy?

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

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

6.16 How Do I Execute Presto SQL in Data Development?

 Copyright © Huawei Technologies Co., Ltd.

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

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

6.1 How Many Jobs Can Be Created in Data Development? Is There a Limit on the Number of Nodes in a Job?

By default, each user can create a maximum of 10,000 jobs, and each job can contain a maximum of 1000 nodes.

In addition, the system allows you to adjust the maximum quota as required. If you have any requirements, **submit a service ticket**.

6.2 How Can I Quickly Rectify a Deleted CDM Cluster Associated with a Job?

After the CDM cluster is deleted, the association information in the data development job remains intact. You only need to create a cluster and job with the same names on CDM. The data development job will remind you that the original CDM cluster and job will be replaced before using the newly created ones.

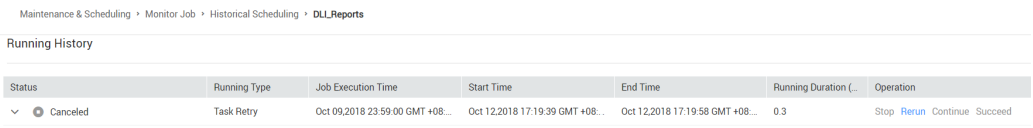
Constraints:

This function is supported in version 1.7.3 (which has been rolled out on October 24, 2018). If you want to use this function for a data development job created before this point in time, save the job again.

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

On the **Running History** page, there is a large difference between **Job Execution Time** and **Start Time**, as shown in the figure below. **Job Execution Time** is the time when the job is expected to be executed. **Start Time** is the time when the job starts to be executed.

Figure 6-1 Running History page



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

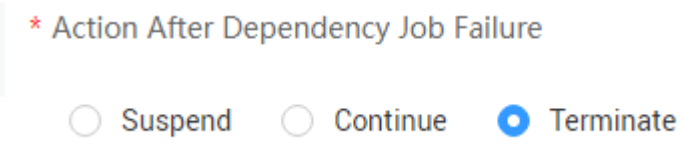
 Copyright © Huawei Technologies Co., Ltd.

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

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

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

Figure 6-2 Job dependencies



In this case, do not stop the job. You can rerun the failed job instance or stop the abnormal instance and then run it again. After the instance failure is removed, the subsequent operations will continue. If you manually process the failure not in Data Development but in other ways, you can force the job instance to succeed after the failure is removed and then subsequent jobs will continue to run properly.

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

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

NOTE

When you create a job, a bucket named **dlf-log-{projectID}** will be created by default. If the bucket exists, you do not need to create a bucket again.

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

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

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

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

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

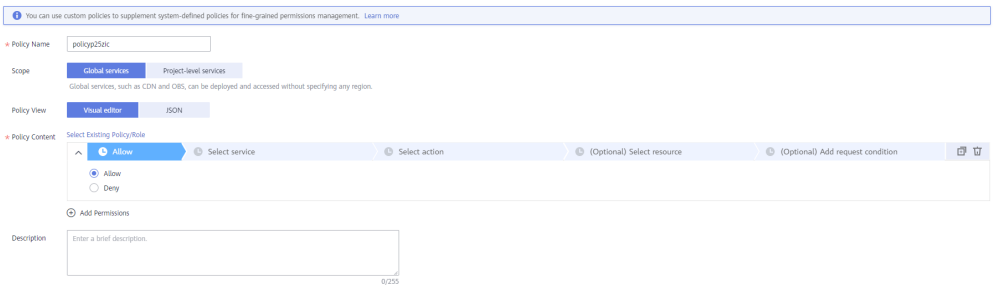
 Copyright © Huawei Technologies Co., Ltd.

Step 1 Log in to HUAWEI CLOUD and click **Console** in the upper right corner.

Step 2 On the management console, hover the mouse pointer over the username in the upper right corner, and choose **Identity and Access Management** from the drop-down list.

Step 3 In the navigation pane, choose **Permissions**. Then, click **Create Custom Policy**.

Step 4 Enter a policy name.



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

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

Step 7 Configure a policy in **Policy Content**.

1. Select **Allow**.
2. Select **Identity and Access Management (IAM)** for **Select service**.
3. Select **iam:agencies:listAgencies** for **Select action**.

Step 8 Click **OK**.

Step 9 Add the policy defined in **Step 7** to the group to which the current user belongs. For details, see **Creating a User Group and Granting Permissions**.

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

----End

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

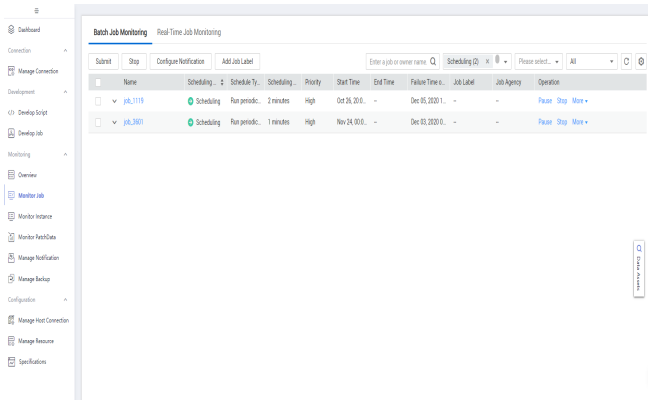
If the number of daily executed nodes exceeds the upper limit, it may be caused by frequent job scheduling. Perform the following operations:

1. In the left navigation tree of Data Development, choose **Monitoring** > **Monitor Instance**, select the current day, and view the jobs that are frequently scheduled.
2. In the left navigation tree of Data Development, choose **Monitoring** > **Monitor Job** to check whether the scheduling period of jobs that are

 Copyright © Huawei Technologies Co., Ltd.

frequently scheduled is set properly. If the scheduling period is inappropriate, adjust the scheduling period or stop the scheduling. Generally, the number of minute-level scheduling jobs executed every day exceeds the upper limit.

Figure 6-3 Viewing the scheduling period



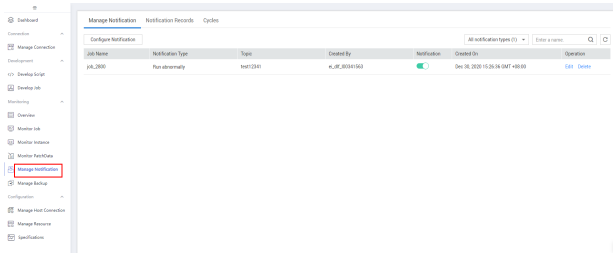
6.8 Why Cannot Specified Peripheral Resources Be Selected When a Data Connection Is Created in Data Development?

Ensure that the current instance and peripheral resources are in the same region and IAM project. If the enterprise project function is enabled for your account, the current instance and peripheral resources must be in the same enterprise project.

6.9 Why Cannot I Receive a Job Failure Alarm Notification After SMN Is Configured?

What can I do if the SMN notification for job exception/failure is configured but the alarm notification for job exception/failure is not received?

Figure 6-4 Notification management



To solve the problem, perform the following steps:

Step 1 Check whether the failed job is being scheduled. No notification is sent for jobs in the test running state. SMN notifications are sent only for jobs in the scheduling state.

 Copyright © Huawei Technologies Co., Ltd.

Step 2 On the Data Development page, choose **Monitoring** > **Manage Notification** to check whether the notification function is enabled.

Step 3 Log in to the SMN console and check whether the SMN topic has been subscribed to.

Step 4 Check whether the subscription endpoint of the SMN topic has its own name and whether the subscription is confirmed.

Step 5 Check whether the SMN channel is normal. You can send messages to your topic on the SMN console to check whether you can receive notifications from SMN.

----End

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

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

   Figure 6-5 Viewing the job scheduling status

2. If a job depends on other jobs, choose **Monitoring** > **Monitor Instance** to view the running status of the dependent jobs. If the job is self-dependent, expand the search time to check whether the job is waiting for running due to the failure of a historical job instance.

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

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

In direct connection mode, DGC users submit the scripts to MRS through APIs and then check whether the scripts are executed successfully. MRS does not send the specific error cause to DGC. Therefore, the GUI displays only the execution result (success or failure) but does not display the error cause.

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

 Copyright © Huawei Technologies Co., Ltd.

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

Check whether the permissions of the current user in IAM are changed, whether the user is removed from the user group, or whether the permission policy of the user group to which the user belongs is changed.

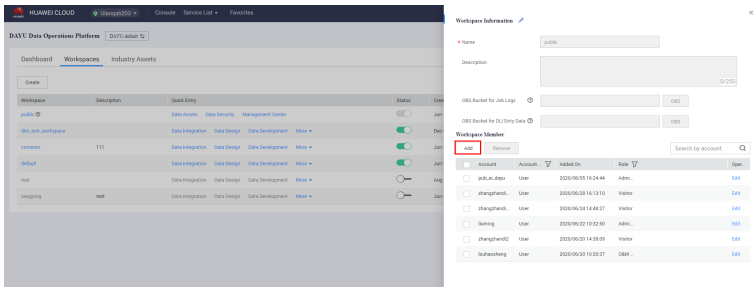
If they are indeed changed, log in to the system again.

6.13 Why Cannot I View the Existing Workspaces After I Have the Required Policy?

Log in to the system as a user who has the permissions required by the current workspace and check whether the required permissions are assigned to all users in the workspace.

If they are not, assign them to users.

Figure 6-6 Viewing workspace members



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

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

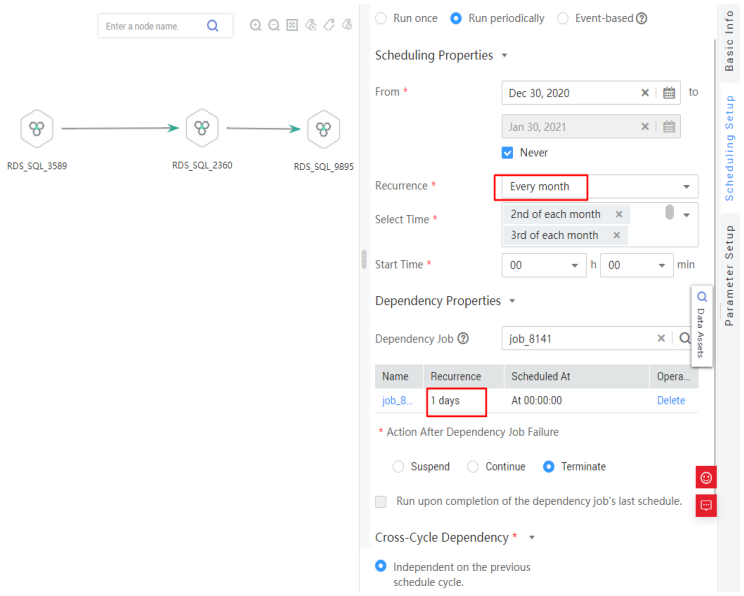
Method 2: Click **Monitor** in the upper part of the canvas, expand the job instance on the **Monitor Instance** page, and view node logs.

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

Jobs scheduled by month depend on jobs scheduled by day. Why does a job scheduled by month start running before the job scheduled by day is complete?

Copyright © Huawei Technologies Co., Ltd.

Figure 6-7 Viewing the job scheduling period and dependency attributes



Although jobs scheduled by month depend on jobs scheduled by day, whether jobs scheduled by month in the current month are executed depends on whether all jobs scheduled by day in the previous month are complete, not the jobs scheduled by day in the current month.

For example, whether the monthly scheduled jobs run in November depends on whether the daily scheduled jobs were complete in October.

6.16 How Do I Execute Presto SQL in Data Development?

For details, see **Using Presto to Dump Data in DLF**.

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

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

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

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

 Copyright © Huawei Technologies Co., Ltd.

7 Data Assets

7.1 What Are the Functions of the Data Assets Module?
7.2 What Assets Can Be Collected?
7.3 What Is Data Lineage?
7.4 How Can Data Lineage Be Displayed on a Data Map?

7.1 What Are the Functions of the Data Assets Module?

The Data Assets module displays enterprise data assets in the form of data maps, including all metadata information and data lineage.

7.2 What Assets Can Be Collected?

Currently, the following assets can be collected: DWS, DLI, MRS HBase, MRS Hive, MySQL, RDS MySQL, and RDS PosetSQL

7.3 What Is Data Lineage?

In the era of big data, various types of data are rapidly generated due to explosive data growth. The massive and complex data information is converged, transformed, and transferred to generate new data and aggregate into an ocean of data.

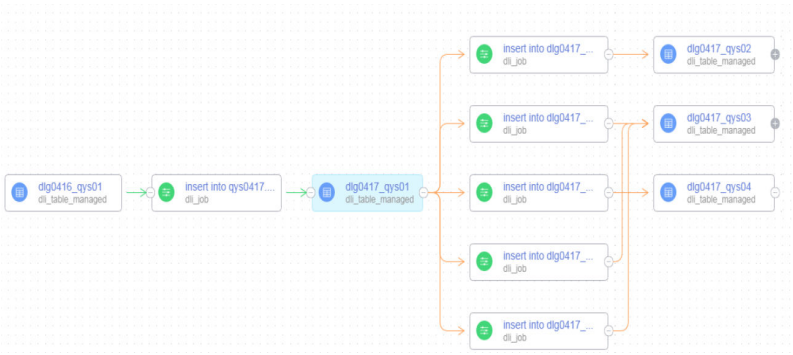
During this process, a relationship is formed between the data, and these relationships are their lineages. They are analogous to the genetic relationships between people. However, in contrast from our human lineages, data lineages have the following distinct features:

- **Belongingness**: Specific data belongs to a specific organization or individual.
- **Multi-source**: One piece of data can have multiple sources. One piece of data may be generated by processing multiple pieces of data, and there may be multiple such processes.
- **Traceability**: The data lineage is traceable. It reflects the data lifecycle and the entire process from data generation to data disappearance.

 Copyright © Huawei Technologies Co., Ltd.

- **Hierarchy**: The data lineage is hierarchical. Data classification and summary form new data, and different levels of description result in data layers.

Figure 7-1 Data lineage example



7.4 How Can Data Lineage Be Displayed on a Data Map?

To display data lineage, you need to collect metadata first and then schedule related jobs in Data Development.

 Copyright © Huawei Technologies Co., Ltd.

8 Data Lake Mall

8.1 What Languages Do Data Lake Mall SDKs Support?

8.1 What Languages Do Data Lake Mall SDKs Support?

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

 Copyright © Huawei Technologies Co., Ltd.

9 Data Security

9.1 Why Is Data in a Data Table Not Masked Based on Rules After a Data Masking Task Is Executed?

9.2 Why Does the System Display a Message Indicating that Some Data Identification Rules Are in Use When They Are Deleted Although No Task Is Using Them?

9.3 What Should I Do If Authentication Audit Logging Is Not Enabled?

9.1 Why Is Data in a Data Table Not Masked Based on Rules After a Data Masking Task Is Executed?

This is because the masking task depends on the sensitive data discovery task. You must create a sensitive data discovery task first. After sensitive fields are discovered, the sensitive fields are masked based on rules.

9.2 Why Does the System Display a Message Indicating that Some Data Identification Rules Are in Use When They Are Deleted Although No Task Is Using Them?

This is because Data Security and Data Assets use the same set of data identification rules. Although a data identification rule is not used in any data security tasks, it may be used in Data Assets. Therefore, when you delete the identification rule, the system displays a message indicating that the identification rule is in use.

9.3 What Should I Do If Authentication Audit Logging Is Not Enabled?

Authentication audit logging is unavailable in this version.

 Copyright © Huawei Technologies Co., Ltd.