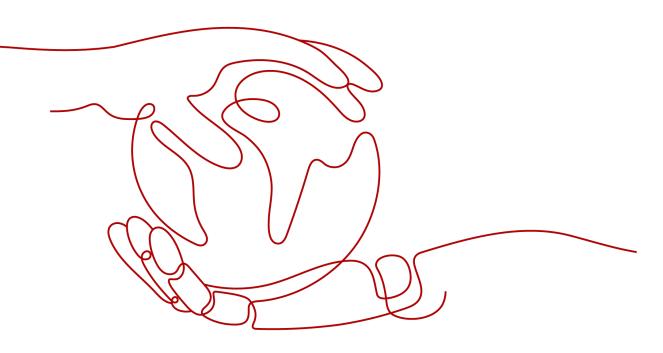
Data Lake Insight

Best Practice

 Issue
 01

 Date
 2023-10-29





HUAWEI TECHNOLOGIES CO., LTD.

Copyright © Huawei Technologies Co., Ltd. 2024. 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

NUAWEI 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.

Security Declaration

Vulnerability

Huawei's regulations on product vulnerability management are subject to the *Vul. Response Process.* For details about this process, visit the following web page:

https://www.huawei.com/en/psirt/vul-response-process

For vulnerability information, enterprise customers can visit the following web page: <u>https://securitybulletin.huawei.com/enterprise/en/security-advisory</u>

Contents

1 Overview	1
2 Data Migration	2
2.1 Overview	2
2.2 Migrating Data from Hive to DLI	4
2.3 Migrating Data from MRS Kafka to DLI	13
2.4 Migrating Data from Elasticsearch to DLI	21
2.5 Migrating Data from RDS to DLI	
2.6 Migrating Data from GaussDB(DWS) to DLI	36
3 Data Analysis	44
3.1 Analyzing Driving Behavior Data	44
3.2 Converting Data Format from CSV to Parquet	54
3.3 Analyzing E-commerce BI Reports	57
3.4 Analyzing DLI Billing Data	64
3.5 Using DLI Flink SQL to Analyze e-Commerce Business Data in Real Time	68
3.6 Interconnecting Yonghong BI with DLI to Submit Spark Jobs	83
3.6.1 Preparing for Yonghong BI Interconnection	83
3.6.2 Adding Yonghong BI Data Source	84
3.6.3 Creating Yonghong BI Data Set	
3.6.4 Creating a Chart in Yonghong BI	90
3.7 Interconnecting FineBI with DLI Trino	93
3.8 Interconnecting Power BI with DLI Trino	104
4 Connections	119
4.1 Configuring the Connection Between a DLI Queue and a Data Source in a Private Network	119
4.2 Configuring the Connection Between a DLI Queue and a Data Source in the Internet	125
A Change History	131



This document gives you best practices for data migration and analysis, helping you better use DLI for large-scale data analysis and processing.

Data Migration

You can use **Cloud Data Migration Service** (CDM) to easily migrate data from other cloud services or service platforms to DLI. You can refer to the following best practices:

- Migrating Data from Hive to DLI
- Migrating Data from MRS Kafka to DLI
- Migrating Data from Elasticsearch to DLI
- Migrating Data from RDS to DLI
- Migrating Data from GaussDB(DWS) to DLI

Data Analysis

DLI is widely used to analyze massive amounts of log data and in extract, transform, and load (ETL) processes, giving you great insight into data of a wide range of industries. You can refer to the following best practices of data analysis:

- Analyzing Driving Behavior Data
- Converting Data Format from CSV to Parquet
- Analyzing E-commerce BI Reports
- Analyzing DLI Billing Data

2 Data Migration

2.1 Overview

This section describes how you can migrate data to DLI in an efficient way. You can use **Cloud Data Migration Service** (CDM) to migrate data from other cloud services or platforms to DLI.

DLI is a serverless data processing and analysis service. It processes streaming data and batch data and supports interactive analysis. Its high-scalability framework supports the convergence of batch and streaming data analysis, and provides realtime, efficient, and diversified compute resources for TB-to EB-level data processing.

Best Practices of Data Migration

- You can migrate Hive data to DLI. For details, see Migrating Data from Hive to DLI.
- You can migrate Kafka data to DLI. For details, see Migrating Data from MRS Kafka to DLI.
- You can migrate Elasticsearch data to DLI. For details, see Migrating Data from Elasticsearch to DLI.
- You can migrate RDS data to DLI. For details, see Migrating Data from RDS to DLI.
- You can migrate GaussDB(DWS) data to DLI. For details, see Migrating Data from GaussDB(DWS) to DLI.

Data Type Mapping

If you migrate data from other cloud services or platforms to DLI, data types need to be converted and source and destination data must be mapped by type. Table 2-1 lists the mapping relationships.

Table 2-1 Data type mapping

MySQL	Hive	DWS	Oracle	Postgre SQL	Hologre s	DLI Spark
CHAR	CHAR	CHAR	CHAR	CHAR	CHAR	CHAR
VARCH AR	VARCHAR	VARCHAR	VARCHAR	VARCHA R	VARCHA R	VARCHA R/ STRING
DECIMA L	DECIMAL	NUMERIC	NUMERIC	NUMERI C	DECIMA L	DECIMAL
INT	INT	INTEGER	NUMBER	INTEGER	INTEGER	INT
BIGINT	BIGINT	BIGINT	NUMBER	BIGINT	BIGINT	BIGINT/ LONG
TINYINT	TINYINT	SMALLINT	NUMBER	SMALLI NT	SMALLI NT	TINYINT
SMALLI NT	SMALLINT	SMALLINT	NUMBER	SMALLI NT	SMALLI NT	SMALLIN T/SHORT
BINARY	BINARY	BYTEA	RAW	BYTEA	BYTEA	BINARY
VARBIN ARY	BINARY	BYTEA	RAW	BYTEA	BYTEA	BINARY
FLOAT	FLOAT	FLOAT4	FLOAT	DOUBLE	FLOAT4	FLOAT
DOUBL E	DOUBLE	FLOAT8	FLOAT	REAL/ DOUBLE	FLOAT8	DOUBLE
DATE	DATE	TIMESTAM P	DATE	DATE	DATE	DATE
TIME	Not supported (use String instead)	TIME	DATE	TIME	TIME	Not supporte d (use String instead)
DATETI ME	TIMESTA MP	TIMESTAM P	TIME	TIME	TIMESTA MP	TIMESTA MP
TINYINT	TINYINT	BOOLEAN	Not supporte d	TINYINT	BOOLEA N	BOOLEA N
Not support ed (use TEXT instead)	Not supported (use String instead)	Not supported (use TEXT instead)	Not supporte d (use VARCHAR instead)	Not supporte d (use TEXT instead)	Not supporte d (use TEXT instead)	ARRAY

MySQL	Hive	DWS	Oracle	Postgre SQL	Hologre s	DLI Spark
Not support ed (use TEXT instead)	Not supported (use String instead)	Not supported (use TEXT instead)	Not supporte d (use VARCHAR instead)	Not supporte d (use TEXT instead)	Not supporte d (use TEXT instead)	МАР
Not support ed (use TEXT instead)	Not supported (use String instead)	Not supported (use TEXT instead)	Not supporte d (use VARCHAR instead)	Not supporte d (use TEXT instead)	Not supporte d (use TEXT instead)	STRUCT

D NOTE

If a service does not support a standard data type, you can use the recommended data type.

2.2 Migrating Data from Hive to DLI

This section describes how to use the CDM data synchronization function to migrate data from MRS Hive to DLI. Data of other MRS Hadoop components can be bidirectionally synchronized between CDM and DLI.

Prerequisites

• You have created a DLI SQL queue.

When you create a queue, set its **Type** to **For SQL**.

- You have created an MRS security cluster that contains the Hive component.
 - In this example, the MRS cluster and component versions are as follows:
 - Cluster version: MRS 3.1.0
 - Hive version: 3.1.0
 - Hadoop version: 3.1.1
 - In this example, Kerberos authentication is enabled when the MRS cluster is created.
- You have created a CDM cluster. For details about how to create a cluster, see Creating a CDM Cluster.

D NOTE

- If the destination data source is an on-premises database, you need the Internet or Direct Connect. When using the Internet, ensure that an EIP has been bound to the CDM cluster, the security group of CDM allows outbound traffic from the host where the off-cloud data source is located, the host where the data source is located can access the Internet, and the connection port has been enabled in the firewall rules.
- If the data source is MRS or GaussDB(DWS) on a cloud, the network must meet the following requirements:

i. If the CDM cluster and the cloud service are in different regions, a public network or a dedicated connection is required for enabling communication between the CDM cluster and the cloud service. If the Internet is used for communication, ensure that an EIP has been bound to the CDM cluster, the host where the data source is located can access the Internet, and the port has been enabled in the firewall rules.

ii. If the CDM cluster and the cloud service are in the same region, VPC, subnet, and security group, they can communicate with each other by default. If the CDM cluster and the cloud service are in the same VPC but in different subnets or security groups, you must configure routing rules and security group rules.

For details about how to configure routes, see **Configure routes**. For details about how to configure security groups, see section **Security Group Configuration Examples**.

iii. The cloud service instance and the CDM cluster belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

In this example, the VPC, subnet, and security group of the CDM cluster are the same as those of the MRS cluster.

Step 1: Prepare Data

- Create a Hive table in the MRS cluster and insert data in the table.
 - a. Log in to MRS Manager by referring to Accessing FusionInsight Manager.
 - b. On MRS Manager, click **System** in the top navigation pane. On the page displayed, choose **Permission** > **Role** from the left navigation pane. On the displayed page, configure the following parameters:
 - **Role Name**: Enter a role name, for example, **hivetestrole**.
 - Configure Resource Permission: Select the MRS cluster name and then Hive. Select Hive Admin Privilege.

Figure 2-1 Creating a Hive role

n FusionInsight Manager	Homepage Cluster - Hosts	O&M Audit Tenant Resources System
	Role > Create Role	
System	Role Name:	hivetestrole
	Configure Resource Permission:	All resources > mrs_test_00378328 > Hive
Permission ^		View Name
• User		Hive Read Write Privileges
User Group Role	[C Ilve Admin Privilege
Security Policy	Description:	
Domain and Mutual Trust		
Interconnection ~		
Certificate		OK Cancel
OMS		
Component		

For details about how to create a role, see **Creating a Role**.

- c. On the MRS Manager console, click System in the top navigation pane.
 On the displayed page, choose Permission > User from the left navigation pane. On the displayed page, set the following parameters:
 - i. Username: Enter a username. In this example, enter hivetestusr.
 - ii. User Type: Select Human-Machine.
 - iii. **Password** and **Confirm Password**: Enter the password of the current user and enter it again.
 - iv. User Group and Primary Group: Select supergroup.
 - v. **Role**: Select the role created in **b** and the **Manager_viewer** role.

Figure 2-2 Creating a Hive User

KusionInsight Manager	Homepage Clust	ter ← Hosts O&M Audit Tenant Resources System
(Internet in the second	User > Create	
System	* Username:	hivetestusr
Permission ^	 User Type: 	Human-Machine Machine-Machine
• User	* Password:	
User Group	· Confirm Password:	
Role	User Group:	Add Clear All Create User Group
Security Policy		supergroup 🗙
Domain and Mutual Trust		
Interconnection ~		
Certificate	Primary Group:	supergroup 👻
OMS	Role:	Add Clear All Create Role
Component		hivetestrole × Manager_viewer ×
	Description:	
		OK Cancel

d. Download and install the Hive client by referring to **Installing an MRS Client**. For example, the Hive client is installed in the **/opt/hiveclient** directory on the active MRS node. e. Go to the client installation directory as user **root**.

For example, run the **cd /opt/hiveclient** command.

f. Run the following command to set environment variables:

source bigdata_env

g. Run the following command to authenticate the user created in **c** as Kerberos authentication has been enabled for the current cluster:

kinit <Username in c>

Example: kinit hivetestusr

h. Run the following command to connect to Hive:

beeline

i. Create a table and insert data into it.

Run the following statement to create a table:

create table user_info(id string,name string,gender string,age int,addr string);

Run the following statements to insert data into the table: insert into table user_info(id,name,gender,age,addr) values("12005000201", "A", "Male", 19, "City A");

insert into table user_info(id,name,gender,age,addr) values ("12005000202","B","male",20,"City B");

insert into table user_info(id,name,gender,age,addr) values ("12005000202","B","male",20,"City B");

D NOTE

In the preceding example, data is migrated by creating a table and inserting data. To migrate an existing Hive database, run the following commands to obtain Hive database and table information:

• Run the following command on the Hive client to obtain database information:

show databases

• Switch to the Hive database from which data needs to be migrated.

use Hive database name

• Run the following command to display information about all tables in the database:

show tables

• Run the following command to query the creation statement of the Hive table:

show create table table name

The queried table creation statements must be processed to comply with the DLI table creation syntax before being executed.

- Create a database and table on DLI.
 - a. Log in to the DLI management console and click **SQL Editor**. On the displayed page, set **Engine** to **spark** and **Queue** to the created SQL queue.

Enter the following statement in the editing window to create a database, for example, the migrated DLI database **testdb**: For details about the syntax for creating a DLI database, see **Creating a Database**. create database testdb:

b. Create a table in the database.

D NOTE

You need to edit the table creation statement obtained by running **show create table** *hive table name* in MRS Hive to ensure the statement complies with the table creation syntax of DLI. For details about the table creation syntax, see **Creating a DLI Table Using the DataSource Syntax**.

create table user_info(id string,name string,gender string,age int,addr string);

Step 2: Migrate Data

- 1. Create a CDM connection to MRS Hive.
 - a. Create a connection to link CDM to the data source MRS Hive.
 - i. Log in to the CDM console, choose **Cluster Management**. On the displayed page, locate the created CDM cluster, and click **Job Management** in the **Operation** column.
 - ii. On the **Job Management** page, click the **Links** tab, and click **Create Link**. On the displayed page, select **MRS Hive** and click **Next**.

Figure 2-3 Selecting the MRS Hive connector

Select Connector					Configur
Data Warehouse	Data Warehouse Service	Data Lake Insight			
Hadoop	MRS HDFS	MRS HBase	MRS Hive	Apache HDFS	
	Apache HBase	Apache Hive			
Object Storage	Object Storage Service (OBS)	Alibaba Cloud OSS			
File System	FTP	SFTP	HTTP		
Relational Database	RDS for MySQL	RDS for PostgreSQL	RDS for SQL Server	MySQL	
	PostgreSQL	Microsoft SQL Server	Oracle	IBM Db2	
NoSQL	Redis	MongoDB			
Messaging System	Data Ingestion Service	MRS Kafka	Apache Kafka		
Search	Elasticsearch				
Open Beta Test	^				
× Cancel > Ne	ter the second se				

iii. Configure the connection. The following table describes the required parameters.

Parameter	Value
Name	Name of the MRS Hive data source, for example, source_hive
Manager IP	Click Select next to the text box and select the MRS Hive cluster. The Manager IP address is automatically specified.
Authenticatio n Method	Set this parameter to KERBEROS if Kerberos authentication is enabled for the MRS cluster. Set this parameter to SIMPLE if the MRS cluster is a common cluster. In this example, set this parameter to KERBEROS .

Parameter	Value
Hive Version	Set this parameter to the Hive version you have selected during MRS cluster creation. If the current Hive version is 3.1.0, set this parameter to HIVE_3_X .
Username	Name of the MRS Hive user created on c
Password	Password of the MRS Hive user

Retain default values for other parameters.

Figure 2-	4 Configuring	the connection	to MRS Hive
-----------	---------------	----------------	-------------

aster Management / cdm-test-0037	8328 / Links / Create Link	(2) Contigue
* Name	source_hive	
* Connector	Hive	
* Hadoop Type	MRS ~	
* Manager IP 🕜	192.168.7.145	Select
Authentication Method	KERBEROS -	
* HIVE Version ⑦	HIVE_3_X *	
* Username 🕜	hivetestusr	
* Password	·····	
* OBS storage support (?)	Yes No	
* Run Mode 🕐	EMBEDDED -	
Use Cluster Config	Yes No	
Show Advanced Attributes		
X Cancel < Previo	ous 📑 Test 🕞 Save	

- iv. Click **Save** to complete the configuration.
- b. Create a connection to link CDM to DLI.
 - i. Log in to the CDM console, choose **Cluster Management**. On the displayed page, locate the created CDM cluster, and click **Job Management** in the **Operation** column.
 - ii. On the Job Management page, click the Links tab, and click Create Link. On the displayed page, select Data Lake Insight and click Next.

			-		
Data Warehouse	Data Warehouse Service	Data Lake Insight			
Hadoop	MRS HDFS	MRS HBase	MRS Hive	Apache HDFS	
	Apache HBase	Apache Hive			
Object Storage	Object Storage Service (OBS)	Alibaba Cloud OSS			
File System	FTP	SFTP	HTTP		
Relational Database	RDS for MySQL	RDS for PostgreSQL	RDS for SQL Server	MySQL	
	PostgreSQL	Microsoft SQL Server	Oracle	IBM Db2	
NoSQL	Redis	MongoDB			
Messaging System	Data Ingestion Service	MRS Kafka	Apache Kafka		
Search	Elasticsearch				

Figure 2-5 Selecting the DLI connector

iii. Configure the connection parameters.

Figure 2-6 Configuring connection parameters

Select Connector				Configur
× Name	source_hive			
* Connector	Hive			
* Hadoop Type	MRS			
* Manager IP 💮	192.168.7.145	Select		
Authentication Method	KERBEROS	~		
* HIVE Version	HIVE_3_X	×		
* Username 🛞	hivelestusr			
* Paanword		463		
* OB3: storage support	Ves No			
* Run Mode 🕐	EMBEDDED	Ψ.		
Use Cluster Config 🕐	Yes No			
Show Advanced Attributes				
× Cancel < Previ	ous of Test	C Save		

After the configuration is complete, click **Save**.

- 2. Create a CDM migration job.
 - a. Log in to the CDM console, choose **Cluster Management**. On the displayed page, locate the created CDM cluster, and click **Job Management** in the **Operation** column.
 - b. On the **Job Management** page, choose the **Table/File Migration** tab and click **Create Job**.
 - c. On the **Create Job** page, specify job information.

Figure 2-7 Configuring the CDM job

Destination Job Configuration Source Job Configuration Destination Job Configuration * Source Job Configuration * Destination University * Source Job Configuration * Source Job Configuration * Source Job Configuration * Source Job Configuration * Source Job Configuration * Source Job Configuration	(3) Configure Task
Source Job Configuration * So	
* Source Lak Surve MARE_SNA • * Controllers Norve	
* Edition Tures (b) (b) (c) (c) (c) (c) (c) (c) (c) (c) (c) (c	
nastele () weight	
Show Advanced Additions Ocean data before input 🛞 View Na	
V Genot) htt	

i. Job Name: Name of the data migration job, for example, hive_to_dli

ii. Set parameters required for **Source Job Configuration**.

Parameter	Value
Source Link Name	Select the name of the data source created in 1.a .
Database Name	Select the name of the MRS Hive database you want to migrate to DLI. For example, the default database.
Table Name	Select the name of the Hive table. In this example, a database created on DLI and the user_info table are selected.
readMode	In this example, HDFS is selected.
	Two read modes are available: HDFS and JDBC. By default, the HDFS mode is used. If you do not need to use the WHERE condition to filter data or add new fields on the field mapping page, select the HDFS mode.
	The HDFS mode shows good performance, but in this mode, you cannot use the WHERE condition to filter data or add new fields on the field mapping page.
	The JDBC mode allows you to use the WHERE condition to filter data or add new fields on the field mapping page.

 Table 2-3 Source job configuration parameters

For details about parameter settings, see **From Hive**.

iii. Set parameters required for **Destination Job Configuration**.

Parameter	Value
Destination Link Name	Select the DLI data source connection created in 1.b .
Resource Queue	Select a created DLI SQL queue.
Database	Select a created DLI database. In this example, database testdb created in Create a database and table on DLI is selected.
Table	Select the name of a table in the database. In this example, table user_info created in Create a database and table on DLI is created.

Table 2-4 Destination	job configuration	parameters
-----------------------	-------------------	------------

Parameter	Value
Clear data before import	Whether to clear data in the destination table before data import. In this example, set this parameter to No .
	If this parameter is set to Yes , data in the destination table will be cleared before the task is started.

For details about parameter settings, see **To DLI**.

- 3. 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.
 - If the type is automatically created at the migration destination, you need to configure the type and name of each field.
 - CDM allows for field conversion during migration. For details, see Field Conversion.

Figure 2-8 Field mapping

fgure Basic Information					2 1	ap Field		() c
urce Field					© /	Destination Paid		÷ e
iane	Example Value	Туре	Operation			Namo	Туре	Operation
1		sitting	2	Q	Π	-le-i M	string	Ω.
879		string	2	Q	Φ	->- ram	ating	12
enter		string	8	Q	B 0	gender	atting	τ
e*		in .	8	٩	B 0	-per apa	int.	π
40Y		string	8	Q	Ū 0	-(r-1) addr	string	0

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

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

- **Retry Upon Failure**: If the job fails to be executed, you can determine whether to automatically retry. Retain the default value **Never**.
- Group: Select the group to which the job belongs. The default group is DEFAULT. On the Job Management page, jobs can be displayed, started, or exported by group.
- Scheduled Execution: For details about how to configure scheduled execution, see Scheduling Job Execution. Retain the default value No.
- **Concurrent Extractors**: Enter the number of extractors to be concurrently executed. Retain the default value **1**.
- Write Dirty Data: Specify this parameter if data that fails to be processed or filtered out during job execution needs to be written to OBS. Before writing dirty data, create an OBS link. You can view the data on OBS later. Retain the default value No so that dirty data is not recorded.
- 5. Click **Save and Run**. On the **Job Management** page, you can view the job execution progress and result.

Figure 2-9 Job progress and execution result

Ouster Management / cdr	m-leni-00378328 / TablePi	le Migration										
Table/Tile Migration	Entire DB Migration	Links	Agents Settings									
() Create Job	🔗 Run 🔋 Delete					C Feedback	pot Ölimpot	Schedule	Al statutes	* Job name	• Jab name or link type	QC
⊙ / ▷ □	c 🛛 Norre	4	Link Details	Created By 20	Last Execution Time 32	Daration 32	Write Statistics	Status	Group Name	Operation		
Enfor a group name.	Q	lo_dl	source_hive-dest_dl	100378328	Mar 28, 2022 20:41:10 GMT+08:00	104	Written rows: 3	Succeeded	DEFAULT	Ran Historical Record Edit	More +	
Groups												
DEFAULT												

Step 3: Query Results

After the migration job is complete, log in to the DLI management console and click **SQL Editor**. In the displayed page, set **Engine** to **spark**, **Queue** to the created SQL queue, and **Database** to the database created in **a**. Execute the following query statement and check whether the Hive table data has been migrated to the **user_info** table: select * from user_info;

Figure 2-10 Querying migrated data

		Engine	spark •	Queues	test_cli_tvx	× (atabases	testdb +	O Execute	Format	Refer Syntax	Settings	More +
<pre>1create database test 2create table user_in 3 4 select * from user_info </pre>	fo(id string,name string,gender string	,age int,addr strin	01										
Line 4, Column 1									Execute: Ctrl+Ente	r, Find: CM+F, F	ormal: Shifl+Ait+F, Verif	y Syntax: Ctrl+Q,	Fulscreen: F1
ecuted Queries (Last Day)	View Result												Clear Al
Result1 O													
Direcuted successfully													
Query select * from user_in													
	d-aed6-2c1d17c1fe00												
The query takes 4.12s, and 1.34	KB scanned.A maximum of 1,000 records o	an be displayed.									Enter a keyword.	Q	<u>u</u> C1 J
id ↓≘	name 48		gender ↓≣					age J≣			addr ↓⊟		
12005000201	A							19					
12005000202	8							20					
12005000202	8							20					

2.3 Migrating Data from MRS Kafka to DLI

This section describes how to use the CDM data synchronization function to migrate data from MRS Kafka to DLI.

Prerequisites

• You have created a DLI SQL queue. For details about how to create a DLI queue, see Creating a Queue.

A CAUTION

When you create a queue, set its Type to For SQL.

- You have created an MRS security cluster that contains the Kafka component. For details about how to create an MRS cluster, see **Purchasing a Custom Cluster**.
 - In this example, the version of the MRS cluster is 3.1.0.
 - You have enabled Kerberos authentication for the MRS cluster.

• You have created a CDM cluster. For details about how to create a cluster, see **Creating a CDM Cluster**.

NOTE

- If the destination data source is an on-premises database, you need the Internet or Direct Connect. When using the Internet, ensure that an EIP has been bound to the CDM cluster, the security group of CDM allows outbound traffic from the host where the off-cloud data source is located, the host where the data source is located can access the Internet, and the connection port has been enabled in the firewall rules.
- If the data source is MRS or GaussDB(DWS), the network must meet the following requirements:

i. If the CDM cluster and the cloud service are in different regions, a public network or a dedicated connection is required for enabling communication between the CDM cluster and the cloud service. If the Internet is used for communication, ensure that an EIP has been bound to the CDM cluster, the host where the data source is located can access the Internet, and the port has been enabled in the firewall rules.

ii. If the CDM cluster and the cloud service are in the same region, VPC, subnet, and security group, they can communicate with each other by default. If the CDM cluster and the cloud service are in the same VPC but in different subnets or security groups, you must configure routing rules and security group rules.

For details about how to configure routes, see **7. Configure routes**. For details about how to configure security groups, see section **Security Group Configuration Examples**.

iii. The cloud service instance and the CDM cluster belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

In this example, the VPC, subnet, and security group of the CDM cluster are the same as those of the MRS cluster.

Step 1: Prepare Data

- Create a Kafka topic for the MRS cluster and send messages to the topic.
 - a. Log in to MRS Manager by referring to Accessing FusionInsight Manager.
 - b. On MRS Manager, click System in the top navigation pane. On the page displayed, choose Permission > User from the left navigation pane. On the displayed page, configure the following parameters:
 - i. **Username**: Enter a username. In this example, enter **testuser2**.
 - ii. User Type: Select Human-Machine.
 - iii. **Password** and **Confirm Password**: Enter the password of the current user and enter it again.
 - iv. User Group and Primary Group: Select kafkaadmin.
 - v. Role: Select Manager_viewer.

Number State	Homepage Clust	ter → Hosts O&M Audit System
and the second s	User • Create	
System	• Username:	testuser2
Parmission	• User Type:	Human-Machine Machine-Machine
Permission · User	* Password:	
• User Group	* Confirm Password:	
· Role	User Group:	Add Clear All Create User Group
Security Policy		kafkaadmin 🗙
Domain and Mutual Trust		
Interconnection ~		
Certificate	Primary Group:	kafkaadmin 👻
OMS	Role:	Add Clear All Create Role
Component		Manager_viewer ×
	Description:	
		OK Cancel

Figure 2-11 Creating a Kafka user

- c. On the MRS Manager console, choose Cluster > Name of the desired cluster > Service > ZooKeeper > Instance. On the displayed page, obtain the IP address of the ZooKeeper instance.
- d. On the MRS Manager console, choose Cluster > Name of the desired cluster > Service > Kafka > Instance. On the displayed page, obtain the IP address of the Kafka instance.
- e. Download and install the Kafka client by referring to **Installing an MRS Client**. For example, the Kafka client is installed in the **/opt/kafkaclient** directory on the active MRS node.
- f. Go to the client installation directory as user **root**.

Example command: **cd /opt/kafkaclient**

g. Run the following command to set environment variables:

source bigdata_env

h. Run the following command to authenticate the user created in **b** since Kerberos authentication has been enabled for the cluster:

kinit <Username in b>

Example command: kinit testuser2

i. Run the following command to create a Kafka topic named **kafkatopic**: kafka-topics.sh --create --zookeeper *IP address 1 of the node where the ZooKeeper role is*.2181,*IP address 2 of the node where the ZooKeeper role is*.2181,*IP address 3 of the node where the ZooKeeper role is*.2181/kafka --replication-factor 1 --partitions 1 --topic kafkatopic

In this command, IP address of the node where the ZooKeeper role is deployed is that of the ZooKeeper instance obtained in **c**.

j. Run the following command to send a test message to **kafkatopic**: kafka-console-producer.sh --broker-list *IP address 1 of the node where the Kafka role is*:21007;*IP address 2 of the node where the Kafka role is*::21007;*IP address 3 of the node where the Kafka role is*::21007 --topic kafkatopic --producer.config /opt/kafkaclient/Kafka/kafka/config/ producer.properties In this command, IP address of the node where the Kafka role is deployed in that of the Kafka instance obtained in **d**.

The content of the test message is as follows: {"PageViews":5, "UserID":"4324182021466249494", "Duration":146,"Sign":-1}

- Create a database and table on DLI.
 - a. Log in to the DLI management console and click **SQL Editor**. On the displayed page, set **Engine** to **spark** and **Queue** to the created SQL queue.

Enter the following statement in the editing window to create a database. The following example creates the migrated DLI database **testdb**. For details about the syntax for creating a DLI database, see **Creating a Database**.

create database testdb;

b. Create a table in the database. For details about the table creation syntax, see **Creating a DLI Table Using the DataSource Syntax**. CREATE TABLE testdlitable(value STRING);

Step 2: Migrate Data

- 1. Create a CDM connection to MRS Kafka.
 - a. Create a connection to link CDM to the data source MRS Kafka.
 - i. Log in to the CDM console, choose **Cluster Management**. On the displayed page, locate the created CDM cluster, and click **Job Management** in the **Operation** column.
 - ii. On the Job Management page, click the Links tab and click Create Link. On the displayed page, select MRS Kafka and click Next.

Data Warehouse	Data Warehouse Service	Data Lake Insight			
Hadoop	MRS HDFS	MRS HBase	MRS Hive	Apache HDFS	
	Apache HBase	Apache Hive			
Object Storage	Object Storage Service (OBS)	Alibaba Cloud OSS			
File System	FTP	SFTP	HTTP		
Relational Database	RDS for MySQL	RDS for PostgreSQL	RDS for SQL Server	MySQL	
	PostgreSQL	Microsoft SQL Server	Oracle	IBM Db2	
NoSQL	Redis	MongoDB			
Messaging System	Data Ingestion Service	MRS Kafka	Apache Kafka		
Search	Elasticsearch		•		

Figure 2-12 Selecting the MRS Kafka connector

iii. Configure the connection. The following table describes the required parameters.

Parameter	Value
Name	Name of the MRS Kafka data source, for example, source_kafka .
Manager IP	Manager IP address of the cluster. The value is automatically specified after you click Select next to the text box and select the MRS Kafka cluster.
Username	Name of the MRS Kafka user created in b .
Password	Password of the MRS Kafka user.
Authenticatio n Method	KERBEROS if Kerberos authentication is enabled for the MRS cluster; SIMPLE if the MRS cluster is a common cluster
	In this example, set this parameter to KERBEROS .

Table 2-5 MRS Kafka	connection	configurations
---------------------	------------	----------------

For more details about the parameters, see Link to Kafka.

Figure 2-13 Configuring	the M	1RS Kafka	connection
-------------------------	-------	-----------	------------

Cluster Management / cdm-test	3 / Links / Create Link	
1 Select Connector		
★ Name	source_kafka	
* Connector	Kafka	
★ Kafka Type	MRS	
* Manager IP	1	Select
* Username 🕥	testuser2	
* Password	·····	
Authentication Method	KERBEROS	·
Show Advanced Attributes		
X Cancel C Previo	ous 😂 Test 🗈 Sav	re

- iv. Click **Save** to complete the configuration.
- b. Create a connection to link CDM to DLI.
 - i. Log in to the CDM console, choose **Cluster Management**. On the displayed page, locate the created CDM cluster, and click **Job Management** in the **Operation** column.
 - ii. On the Job Management page, click the Links tab, and click Create Link. On the displayed page, select Data Lake Insight and click Next.

			-		
Data Warehouse	Data Warehouse Service	Data Lake Insight			
Hadoop	MRS HDFS	MRS HBase	MRS Hive	Apache HDFS	
	Apache HBase	Apache Hive			
Object Storage	Object Storage Service (OBS)	Alibaba Cloud OSS			
File System	FTP	SFTP	HTTP		
Relational Database	RDS for MySQL	RDS for PostgreSQL	RDS for SQL Server	MySQL	
	PostgreSQL	Microsoft SQL Server	Oracle	IBM Db2	
NoSQL	Redis	MongoDB			
Messaging System	Data Ingestion Service	MRS Kafka	Apache Kafka		
Search	Elasticsearch				

Figure 2-14 Selecting the DLI connector

iii. Configure the connection parameters. For details about parameter settings, see Link to DLI.

Figure 2-15 Configuring connection parameters

Cluster Management /	/ cdm-lest-00376328 / Links / Create Link	
1 Select Connector		2 Configure
* Name	dest_dii	
* Connector	DLI ·	
* AK 🕜		
* sк 🧑	······ @	
* Project ID	05 1	
× Cancel	C Previous of Test	

- iv. After the configuration is complete, click Save.
- 2. Create a CDM migration job.
 - a. Log in to the CDM console, choose **Cluster Management**. On the displayed page, locate the created CDM cluster and click **Job Management** in the **Operation** column.
 - b. On the **Job Management** page, choose the **Table/File Migration** tab and click **Create Job**.
 - c. On the **Create Job** page, specify job information.

Figure 2-16 Configuring the CDM job

Configure Basic Information) Map Pield		(3) Configure Task
Job Configuration				
* Job Name Set				
Source Job Configuration		Destination Job Configuration	n	
* Source Link Name	alla .	* Destination Link Name	dest_di	
+ Topica 🕜 kalkato	ipie	* Resource Queue	lestdi O	
+ Data Permat	R5_ISON) *	* Database ()	Tesh/b O	
* Offset Facameter (1) EARLI	E3T *	A Table (2)	testatizatie2	
* Permanent Ranning 💮 Yes	No	Ciear data before import	Yan No.	
* Puli Data Timeout 🕥 15				
Walt Data Terreout				
* Consumer Group ID example	de-group1			
Show Advanced Attributes				
× Cancel > Next Save				

- i. Job Name: Name of the data migration job, for example, test
- ii. Set parameters required for **Source Job Configuration**.

Parameter	Value
Source Link Name	Select the name of the data source created in 1.a.
Topics	Name of the topics you want to migrate to DLI. You can select one or more topics. Example: kafkatopic .
Data Format	Select the message format as needed. In this example, CDC (DRS_JSON) is selected, indicating that the source data will be parsed in DRS_JSON format.
Offset Parameter	Initial offset when data is pulled from Kafka. In this example, select EARLIEST . Available values are as follows:
	• Latest: Maximum offset, indicating that the latest data will be extracted
	• Earliest : Minimum offset, indicating that the earliest data will be extracted
	• Submitted: Data that has been submitted
	• Time Range : Data within a specified time range
Permanent Running	Whether a job runs permanently. In this example, set this parameter to No .
Pull Data Timeout	Maximum minutes allowed for a continuous data pulling. In this example, set this parameter to 15 .
Wait Data Timeout	(Optional) Maximum seconds allowed for waiting data reading. In this example, leave this parameter empty.
Consumer Group ID	Consumer group ID. The default Kafka message group ID example-group1 is used.

Table 2-6 Source job configuration parameters

For details about parameter settings, see From Apache Kafka.

iii. Set parameters required for **Destination Job Configuration**.

Parameter	Value
Destination Link Name	Select the DLI data source connection created in 1.b .
Resource Queue	Select a created DLI SQL queue.
Database	Select a created DLI database. In this example, database testdb created in Create a database and table on DLI is selected.
Table	Select the name of a table in the database. In this example, table testdlitable created in Create a database and table on DLI is selected.
Clear data before import	Whether to clear data in the destination table before data import. In this example, set this parameter to No .
	If this parameter is set to Yes , data in the destination table will be cleared before the task is started.

Table 2-7 Destination job configuration parameters

For details about parameter settings, see **To DLI**.

- 3. 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.
 - If the type is automatically created at the migration destination, you need to configure the type and name of each field.
 - CDM allows for field conversion during migration. For details, see Field Conversion.

Figure 2-17 Field mapping

Configure Basic Information			2 Map	Field			- (3) Configure Task
Source Pield			Ð	Destrution Pield			7 G
Column ID	Dample Value ("PageViewn" 5, "User87" 4324 152821465239464", "Durate	Operation 2 Q	¥	Name	jha	Operation	
X Cancel C Previous X Ned	(Pageveen 5, Usero, 424 6222 Hoozeney, Danso.	÷ 4			ating	8	

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

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

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

- Concurrent Extractors: Enter the number of extractors to be concurrently executed. Retain the default value 1.
- Write Dirty Data: Specify this parameter if data that fails to be processed or filtered out during job execution needs to be written to OBS. Before writing dirty data, create an OBS link. You can view the data on OBS later. In this example, retain the default value **No** so that dirty data is not recorded.
- 5. Click Save and Run. On the Job Management page, you can view the job execution progress and result.

Figure 2-18 Job progress and execution result

Cluster Management 7 calm test-	1 / TablerPile Migration									
Table/Tile Migration Entire	DB Mignation Links Ager	ta Settings								
(i) Create Juli	C Debrie				C Feedback	tropost 13	Tubebule	* All statuses	• Job name • Job name or test type	Q C
0 / P 0 (Rano (II	Link Defails	Created By 28	Last Execution Taxe 28	Daraban 28	Wite Materica	354549	Group Maxie	Operation	
Deler a group name. Q	Name 28	source_kalka-dest_di	Created By 28	Last Execution Tree 28 Apr 07, 2922 15:42:25 GMT-00:00	Daration 28 15m 15e	Write Statistics Volters rave 3	Status Conceeded	Ordep Name OEFAULT	Operation Ran Hatorical Record Edit More v	

Step 3: Query Results

After the migration job is complete, log in to the DLI management console and click SQL Editor. In the displayed page, set Engine to spark, Queue to the created SQL queue, and **Database** to the database created in **a**. Execute the following query statement and check whether the Kafka table data has been migrated to the **testdlitable** table: select * from testdlitable:

2.4 Migrating Data from Elasticsearch to DLI

This section describes how to use the CDM data synchronization function to migrate data from a CSS Elasticsearch cluster to DLI. Data in a self-built Elasticsearch cluster can also be bidirectionally synchronized between CDM and DLI.

Prerequisites

You have created a DLI SQL queue. For details about how to create a DLI queue, see Creating a Queue.

When you create a queue, set its Type to For SQL.

You have created a CSS Elasticsearch cluster. For details about how to create a CSS cluster, see Creating an Elasticsearch Cluster in Non-Security Mode.

In this example, the version of the created CSS cluster is 7.6.2, and security mode is disabled for the cluster.

You have created a CDM cluster. For details about how to create a CDM cluster, see Creating a CDM Cluster.

D NOTE

- If the destination data source is an on-premises database, you need the Internet or Direct Connect. When using the Internet, ensure that an EIP has been bound to the CDM cluster, the security group of CDM allows outbound traffic from the host where the off-cloud data source is located, the host where the data source is located can access the Internet, and the connection port has been enabled in the firewall rules.
- If the data source is CSS on a cloud, the network must meet the following requirements:

i. If the CDM cluster and the cloud service are in different regions, a public network or a dedicated connection is required for enabling communication between the CDM cluster and the cloud service. If the Internet is used for communication, ensure that an EIP has been bound to the CDM cluster, the host where the data source is located can access the Internet, and the port has been enabled in the firewall rules.

ii. If the CDM cluster and the cloud service are in the same region, VPC, subnet, and security group, they can communicate with each other by default. If the CDM cluster and the cloud service are in the same VPC but in different subnets or security groups, you must configure routing rules and security group rules.

For details about how to configure routes, see **Configure routes**. For details about how to configure security groups, see section **Security Group Configuration Examples**.

iii. The cloud service instance and the CDM cluster belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

In this example, the VPC, subnet, and security group of the CDM cluster are the same as those of the CSS cluster.

Step 1: Prepare Data

- Create an index for the CSS cluster and import data.
 - a. Log in to the CSS management console and choose **Clusters** > **Elasticsearch** from the navigation pane on the left.
 - b. On the **Clusters** page, click **Access Kibana** in the **Operation** column of the created CSS cluster.
 - c. In the navigation pane of Kibana, choose **Dev Tools**. The **Console** page is displayed.
 - d. On the displayed **Console** page, run the following command to create index **my_test**:

```
PUT /my_test
{
    "settings": {
        "number_of_shards": 1
    },
    "mappings": {
            "properties": {
            "productName": {
                "type": "text",
                "analyzer": "ik_smart"
            },
            "size": {
                "type": "keyword"
            }
            }
        }
    }
}
```

e. Run the following command to import data to the **my_test** index:

POST /my_test/_doc/_bulk
{"index":{}}
{"productName":"2017 Autumn New Shirts for Women", "size":"L"}
{"index":{}}
{"productName":"2017 Autumn New Shirts for Women", "size":"M"}
{"index":{}}
{"productName":"2017 Autumn New Shirts for Women", "size":"S"}
{"index":{}}
{"productName":"2018 Spring New Jeans for Women","size":"M"}
{"index":{}}
{"productName":"2018 Spring New Jeans for Women","size":"S"}
{"index":{}}
{"productName":"2017 Spring Casual Pants for Women","size":"L"}
{"index":{}}
{"productName":"2017 Spring Casual Pants for Women","size":"S"}
If errors is false in the command output, the data is imported.

- Create a database and table on DLI.
- a. Log in to the DLI management console and click **SQL Editor**. On the displayed page, set **Engine** to **spark** and **Queue** to the created SQL queue.

Enter the following statement in the editing window to create a database, for example, the migrated DLI database **testdb**: For details about the syntax for creating a DLI database, see **Creating a Database**. create database testdb;

b. Create a table in the database. For details about the table creation syntax, see **Creating a DLI Table Using the DataSource Syntax**. create table tablecss(size string, productname string);

Step 2: Migrate Data

- 1. Create a CDM connection to MRS Hive.
 - a. Create a connection to link CDM to the data source CSS.
 - i. Log in to the CDM console, choose **Cluster Management**. On the displayed page, locate the created CDM cluster, and click **Job Management** in the **Operation** column.
 - ii. On the Job Management page, click the Links tab, and click Create Link. On the displayed page, select Cloud Search Service and click Next.

Select Connector					
Data Warehouse	Data Warehouse Service	Data Lake Insight			
Hadoop	MRS HDFS	MRS HBase	MRS Hive	Apache HDFS	
	Apache HBase	Apache Hive			
Object Storage	Object Storage Service (OBS)	Alibaba Cloud OSS			
File System	FTP	SFTP	HTTP		
Relational Database	RDS for MySQL	RDS for PostgreSQL	RDS for SQL Server	MySQL	
	PostgreSQL	Microsoft SQL Server	Oracle	IBM Db2	
NoSQL	Redis	MongoDB			
Messaging System	Data Ingestion Service	MRS Kafka	Apache Kafka		
Search	Elasticsearch				
Open Beta Test	~				
	FusionInsight LibrA	FusionInsight HDFS	FusionInsight HBase	FusionInsight Hive	
	Qiniu Cloud Object Storage (KODO)	Amazon S3	Tencent Cloud COS	Distributed Database Middleware	
	SAP HANA	MYCAT	DM	Sharding Database	
	Distributed Cache Service	Document Database Service	CloudTable Service	CloudTable Service (OpenTSDB)	
	Cassandra	DMS Kafka	Cloud Search Service		
X Cancel	ext			1	

iii. Configure the connection. The following table describes the required parameters. For details about parameter settings, see Link to Elasticsearch/CSS.

Parameter	Value.
Name	Name of the CSS data source, for example, source_css .
Elasticsearch Server List	Click Select next to the text box and select the CSS cluster. The Elasticsearch server list is automatically displayed.
Security mode Authenticatio n	If you have enabled the security mode for the CSS cluster, set this parameter to Yes . Otherwise, set this parameter to No . In this example, set this parameter to No .

Table 2-8 CSS data source configuration

Figure 2-19 Selecting the CSS connector

Cluster Management / cdm-te	/ Linis / Create Link	2 Configure
* Name	Source_cos	
* Connector	Elasticsearch *	
* Elasticsearch Server List 🕥	Select	
Security mode Authentication	Yes No	
X Cancel <br< td=""><td>¢i Test ☐ Save</td><td></td></br<>	¢i Test ☐ Save	

Figure 2-20 Configuring the CSS connection

- iv. Click **Save** to complete the configuration.
- b. Create a connection to link CDM to DLI.
 - i. Log in to the CDM console, choose **Cluster Management**. On the displayed page, locate the created CDM cluster, and click **Job Management** in the **Operation** column.
 - ii. On the **Job Management** page, click the **Links** tab, and click **Create Link**. On the displayed page, select **Data Lake Insight** and click **Next**.

elect Connector					(2 c
Data Warehouse	Data Warehouse Service	Data Lake Insight			
Hadoop	MRS HDFS	MRS HBase	MRS Hive	Apache HDFS	
	Apache HBase	Apache Hive			
Dbject Storage	Object Storage Service (OBS)	Alibaba Cloud OSS			
File System	FTP	SFTP	HTTP		
Relational Database	RDS for MySQL	RDS for PostgreSQL	RDS for SQL Server	MySQL	
	PostgreSQL	Microsoft SQL Server	Oracle	IBM Db2	
NoSQL	Redis	MongoDB			
dessaging System	Data Ingestion Service	MRS Katka	Apache Kafka		
Search	Elasticsearch				
Open Beta Test	<u>^</u>				

Figure 2-21 Selecting the DLI connector

iii. Configure the connection parameters. For details about parameter settings, see Link to DLI.

Figure 2-22 Configuring connection parameters

Cluster Management / cdm-te /	Links / Create Link	— 2 Configure
* Name	source_cos	
* Connector	Elasticsearch v	
* Elasticsearch Server List (?)	Select	
Security mode Authentication (2)	Yes No	
× Cancel < Previous	oʻ; Tost 🗋 Save	

- iv. After the configuration is complete, click **Save**.
- 2. Create a CDM migration job.
 - a. Log in to the CDM console, choose **Cluster Management**. On the displayed page, locate the created CDM cluster, and click **Job Management** in the **Operation** column.
 - b. On the **Job Management** page, choose the **Table/File Migration** tab and click **Create Job**.
 - c. On the **Create Job** page, specify job information.

Figure 2-23 Configuring the CDM job

Name css_to_dli	
ource Job Configuration	Destination Job Configuration
Source Link Name Source_css	* Destination Link Name dest_dii
Index (?) my_lett (* Resource Queue 🕥 bz
Type 🕐 🔄	* Database ⑦ testdb
how Advanced Attributes	* Table ⑦ tablecs

- i. Job Name: Name of the data migration job, for example, css_to_dli
- ii. Set parameters required for Source Job Configuration.

Parameter	Value
Source Link Name	Select the name of the data source created in 1.a .
Index	Select the Elasticsearch index created for the CSS cluster. In this example, the my_test index created in Create an index for the CSS cluster and import data is used. The index can contain only lowercase letters.
Туре	Elasticsearch type, which is similar to the table name of a relational database. The type name can contain only lowercase letters. Example: _doc .

Table 2-9 Source job configuration parameters

For details about other parameters, see From Elasticsearch or CSS.

iii. Set parameters required for **Destination Job Configuration**.

Parameter	Value
Destination Link Name	Select the DLI data source connection created in 1.b .
Resource Queue	Select a created DLI SQL queue.
Database	Select a created DLI database. In this example, database testdb created in Create a database and table on DLI is selected.
Table	Select the name of a table in the database. In this example, table tablecss created in Create a database and table on DLI is created.
Clear data before import	Whether to clear data in the destination table before data import. In this example, set this parameter to No .
	If this parameter is set to Yes , data in the destination table will be cleared before the task is started.

Table 2-10 Destination job configuration parameters

For details about parameter settings, see **To DLI**.

- 3. 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.
 - If the type is automatically created at the migration destination, you need to configure the type and name of each field.
 - CDM allows for field conversion during migration. For details, see Field Conversion.

Figure 2-24 Field mapping

Source Field					⊙ ∥	Destination Field		₫ 🕲 ⊙
Туре	Name	Example Value	Operatio	1		Name	Type	Operation
sting	productName		8	Q	¥ • · · · · ·	- In productname	shing	Ψ
keyword	size	L	8	Q	Ū 0 · · · · · ·	- (bil) size	string	σ
					•			⊕ 🕑 ⊙

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

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

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

- Group: Select the group to which the job belongs. The default group is DEFAULT. On the Job Management page, jobs can be displayed, started, or exported by group.
- Scheduled Execution: For details about how to configure scheduled execution, see Scheduling Job Execution. Retain the default value No.
- **Concurrent Extractors**: Enter the number of extractors to be concurrently executed. Retain the default value **1**.
- Write Dirty Data: Specify this parameter if data that fails to be processed or filtered out during job execution needs to be written to OBS. Before writing dirty data, create an OBS link. You can view the data on OBS later. Retain the default value No so that dirty data is not recorded.
- 5. Click **Save and Run**. On the **Job Management** page, you can view the job execution progress and result.

Figure 2-25 Job progress and execution result

Table/File Migration Er	tire DB Migration Links	igents Settings								
@ Create Jab di Ru	n C Delete				G Feedback	2 Export 2 Imp	port Schedule	• All statuses	• Job name • Job name or link type	QC
⊙ ≠ ⊳ ⊑	< □ Nette J⊞	Link Details	Created Dy 48	Last Execution Time 48	Duration JE	Write Statistics	Status	Group Name	Operation	
Enter a group nome. Q	0_00_00	source_css-dest_dl	el_dics_d003	Apr 11, 2022 19:29:39 GMT+08:00	tre 19s	Written rows: 7	Succeeded	DEFAULT	Run Historical Record Edt More +	
Groups										
DEFAULT										

Step 3: Query Results

After the migration job is complete, log in to the DLI management console and click **SQL Editor**. In the displayed page, set **Engine** to **spark**, **Queue** to the created SQL queue, and **Database** to the database created in **a**. Execute the following query statement and check whether the CSS table data has been migrated to the **tablecss** table:

select * from tablecss;

Figure 2-26 Querying migrated data

	ueues bzq_i	O Execute Format Refer Syntax Settings More
1 select * from tablecss;		
Line 1, Column 1		Execute: Ctrl+Enter, Find: Ctrl+F, Format: Shift+Alt+F, Verify Syntax: Ctrl+Q, Fullscreen: F11
Executed Queries (Last Day) View Result		Clear All
Result1 Ø		
Executed successfully		
Query select * from tablecss Job ID b5e135d0-a17f-4acc-b3d4-280853e2326a		
The query takes 35.29s, and 1.36 KB scanned.A maximum of 1,000 records can b		Enter a keyword. Q
size ↓Ξ	productname ↓≡	
L	2017 Autumn New Shirts for Women	
s	2017 Autumn New Shirts for Women	
s	2018 Spring New Jeans for Women	
S	2017 Spring Casual Pants for Women	
M	2017 Autumn New Shirts for Women	

2.5 Migrating Data from RDS to DLI

This section describes how to use the CDM data synchronization function to migrate data from an RDS DB instance to DLI. Data in other relational databases can also be bidirectionally synchronized between CDM and DLI.

Prerequisites

• You have created a DLI SQL queue. For details about how to create a DLI queue, see **Creating a Queue**.

When you create a queue, set its **Type** to **For SQL**.

- You have created an RDS for MySQL DB instance. For details about how to create an RDS cluster, see **Buy a DB Instance**.
 - In this example, the RDS DB engine is MySQL.
 - In this example, the DB engine version is 5.7.
- You have created a CDM cluster. For details about how to create a CDM cluster, see **Creating a CDM Cluster**.

NOTE

- If the destination data source is an on-premises database, you need the Internet or Direct Connect. When using the Internet, ensure that an EIP has been bound to the CDM cluster, the security group of CDM allows outbound traffic from the host where the off-cloud data source is located, the host where the data source is located can access the Internet, and the connection port has been enabled in the firewall rules.
- If the data source is RDS or MRS on a cloud, the network must meet the following requirements:

i. If the CDM cluster and the cloud service are in different regions, a public network or a dedicated connection is required for enabling communication between the CDM cluster and the cloud service. If the Internet is used for communication, ensure that an EIP has been bound to the CDM cluster, the host where the data source is located can access the Internet, and the port has been enabled in the firewall rules.

ii. If the CDM cluster and the cloud service are in the same region, VPC, subnet, and security group, they can communicate with each other by default. If the CDM cluster and the cloud service are in the same VPC but in different subnets or security groups, you must configure routing rules and security group rules.

For details about how to configure routes, see **Configure routes**. For details about how to configure security groups, see section **Security Group Configuration Examples**.

iii. The cloud service instance and the CDM cluster belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

In this example, the VPC, subnet, and security group of the CDM cluster are the same as those of the RDS for MySQL DB instance.

Step 1: Prepare Data

- Create databases and tables on the RDS for MySQL DB instance.
 - Log in to the RDS console. On the displayed page, locate the target DB a instance and choose More > Log In in the Operation column.
 - On the displayed login page, enter the correct username and password b. and click Log In.
 - On the **Databases** page, click **Create Database**. In the displayed dialog c. box, enter testrdsdb as the database name and retain default values of rest parameters. Then, click **OK**.
 - In the **Operation** column of row where the created database locates, d. click **SQL Window** and enter the following statement to create a table: CREATE TABLE tabletest (
 - `id` VARCHAR(32) NOT NULL, `name` VARCHAR(32) NOT NULL, PRIMARY KEY (`id`) ENGINE = InnoDB
 - DEFAULT CHARACTER SET = utf8mb4;
 - Run the following statements to insert data to the created table: e. insert into tabletest VALUES ('123','abc'); insert into tabletest VALUES ('456','efg'); insert into tabletest VALUES ('789','hij');
 - f. Run the following statement to query table data: select * from tabletest;

Figure 2-27 Querying table data

Homo SQL Window X				
Current Database testrdiscib 📀	Maater Switch SQL Execution Ned	in Instance Name: rds-test-00378328	192, 168 8 197, 3308 Character Set un3	Save Executed SQL Statements (1)
Cutabase: Institute V	Concols SCL (Fit) B Format in J select * from tabletest;	ida (Fil) 🔞 Brecons Ida Pan (Fil)	(BOLFANNER V)	Sids, Input Prompt 🛞 🌑 (Full Screen 22
Please search by X 0, C				
 Eddeded 	Executed SCL Statements Messaces	Read Part V		Call Controller Mark @
	The following is the execution small set of		O Cick on the cell is will be take. Also address or edites, you need is salent and save the shares	
	The following is the execution result set of			Copy Down Copy Column v Column Dellings v
			1250	
	1	112	etc.	
	2	454	*1	
	3	219	NJ	

- Create a database and table on DLI.
 - Log in to the DLI management console and click **SQL Editor**. On the а. displayed page, set **Engine** to **spark** and **Queue** to the created SQL aueue.

Enter the following statement in the editing window to create a database, for example, the migrated DLI database **testdb**: For details about the syntax for creating a DLI database, see Creating a Database. create database testdb;

b. In **SQL Editor**, select **testdb** for **Database** and run the following table creation statement to create a table in the database. For details about the table creation syntax, see Creating a DLI Table Using the DataSource Syntax.

create table tabletest(id string,name string);

Step 2: Migrate Data

- 1 Create a CDM connection to MRS Hive.
 - a. Create a connection to the RDS database.

- i. Log in to the CDM console, choose **Cluster Management**. On the displayed page, locate the created CDM cluster, and click **Job Management** in the **Operation** column.
- ii. If this is your first time crating a connection to RDS for MySQL, upload the MySQL driver. Choose the Links tab and click Driver Management. The Driver Management page is displayed.
- Download the MySQL driver to your local PC by referring to Managing Drivers and decompress the driver package to obtain the JAR file.

For example, download the **mysql-connector-java-5.1.48.zip** package and decompress it to obtain the driver file **mysql-connector-java-5.1.48.jar**.

- iv. Return to the Driver Management page. Locate the MYSQL driver and click Upload in the Operation column. In the Import Driver File dialog box, click Select File to upload the driver file obtained in 1.a.iii.
- v. On the **Driver Management** page, click **Back** to return to the **Links** tab. Click **Create Link**, select **RDS for MySQL**, and click **Next**.
- vi. Configure the connection. The following table describes the required parameters.

Parameter	Value
Name	Name of the RDS data source, for example, source_rds
Database Server	Click Select next to the text box and click the name of the created RDS DB instance. The database server address is automatically entered.
Port	Port number of the RDS DB instance. The value is automatically entered after you select the database server.
Database Name	Name of the RDS DB instance you want to migrate. The testrdsdb database created in c is used in this example.
Username	Username used for accessing the database. This account must have the permissions required to read and write data tables and metadata. In this example, the default user root for creating
Password	the RDS for MySQL DB instance is used. Password of the user.

 Table 2-11
 Connection parameters

For other parameters, retain the default values. For details, see Link to Relational Databases. Click **Save** to complete the configuration.

elect Connector			2 Configu
When you create a data	atabase link for the first time, upload the r	equired driver on the Driver Management page or this page.	
* Name	source_rds		
* Connector	Relational Database v		
Database Type	MySQL v		
* Database Server		Select	
* Port (?)	3306		
* Database Name (🤉	testrdsdb		
* Username (?)	root		
* Password (?)	····· 60		
Use Local API	Yes No		
Use Agent (?)	Yes No		
Driver Version	mysql-connector-java-5.1.48.jar Uplo	ad Copy from SFTP	
Show Advanced Attributes			

Figure 2-28 Configuring the connection to the RDS for MySQL DB instance

- b. Create a connection to the DLI.
 - i. Log in to the CDM console, choose **Cluster Management**. On the displayed page, locate the created CDM cluster, and click **Job Management** in the **Operation** column.
 - ii. On the Job Management page, click the Links tab, and click Create Link. On the displayed page, select Data Lake Insight and click Next.

			-		
Data Warehouse	Data Warehouse Service	Data Lake Insight			
Hadoop	MRS HDFS	MRS HBase	MRS Hive	Apache HDFS	
	Apache HBase	Apache Hive			
Dbject Storage	Object Storage Service (OBS)	Alibaba Cloud OSS			
ile System	FTP	SFTP	HTTP		
Relational Database	RDS for MySQL	RDS for PostgreSQL	RDS for SQL Server	MySQL	
	PostgreSQL	Microsoft SQL Server	Oracle	IBM Db2	
IoSQL	Redis	MongoDB			
lessaging System	Data Ingestion Service	MRS Kafka	Apache Kafka		
Search	Elasticsearch				
Open Beta Test	^				

Figure 2-29 Selecting the DLI connector

i. Create a connection to link CDM to DLI. For details about parameter settings, see Link to DLI.

Figure 2-30 Selecting the DLI connector

Cluster Management / cdm-test-00378328 / Links / Create Link	
① Select Connector	2 Configure
* Name dest_dii	
* Connector DLI ~	
* AK ()	
* SK ()	
* Project ID 05/ f	
X Cancel Previous ei Test Save	

After the configuration is complete, click **Save**.

- 2. Create a CDM migration job.
 - a. Log in to the CDM console, choose **Cluster Management**. On the displayed page, locate the created CDM cluster, and click **Job Management** in the **Operation** column.
 - b. On the Job Management page, choose the Table/File Migration tab and click Create Job.
 - c. On the **Create Job** page, specify job information.

Figure 2-31 Configuring the migration job

Job Configuration				
* Job Name rds_to_dli				
Source Job Configuration		Destination Job Confi	guration	
* Source Link Name Source	Lrds	* Destination Link Name	dest_dl	
Use SQL Statement ⑦ Ye	No	* Resource Queue (?)	test0402	Θ
* Schema/Table Space (?) lestro	db Θ	* Database (?)	testdb	Θ
* Table Name () tablet	st Θ	* Table 🕐	tabletest	Θ
Show Advanced Altributes		Clear data before import	Yes No	

- i. Job Name: Name of the data migration job, for example, rds_to_dli
- ii. Set parameters required for Source Job Configuration.

Parameter	Value
Source Link Name	Select the name of the data source created in 1.a.
Use SQL Statement	When Use SQL Statement is set to Yes , enter an SQL statement here. CDM exports data based on the SQL statement.
	In this example, set this parameter to No .
Schema/Table Space	Select the name of the RDS for MySQL DB instance you want to migrate to DLI. For example, the testrdsdb database.

Table 2-12 Source job configuration parameters

Parameter	Value
Table Name	Name of the table you want to migrate. In this example, use tabletest created in d .

For details about parameter settings, see **From PostgreSQL/SQL Server**.

iii. Set parameters required for **Destination Job Configuration**.

Parameter	Value
Destination Link Name	Select the DLI data source connection.
Resource Queue	Select a created DLI SQL queue.
Database	Select a created DLI database. In this example, database testdb created in Create a database and table on DLI is selected.
Table	Select the name of a table in the database. In this example, table tabletest created in Create a database and table on DLI is created.
Clear data before import	Whether to clear data in the destination table before data import. In this example, set this parameter to No .
	If this parameter is set to Yes , data in the destination table will be cleared before the task is started.

Table 2-13 Destination job configuration parameters

For details about parameter settings, see **To DLI**.

- iv. 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.
 - If the type is automatically created at the migration destination, you need to configure the type and name of each field.
 - CDM allows for field conversion during migration. For details, see **Field Conversion**.

Figure 2-32 Field mapping

) Configure Basic Information						Field			(3) Configure
Source Field	Example Value	7/01	Operation		⊙∥	Destination Field	Type	Operation	• •
id	Crastia Asta	WRCHAR(32)		Q	¥ • • • • • • • • •		string	C. Cleanor	
name		WARCHAR(32)	e	Q	¥ • • • • • • • •	name .	string	12	
					⊙∥			4	6 ⊕
× Cancel (Previ	tos > Next Save								

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

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

- Retry Upon Failure: If the job fails to be executed, you can determine whether to automatically retry. Retain the default value Never.
- **Group**: Select the group to which the job belongs. The default group is **DEFAULT**. On the **Job Management** page, jobs can be displayed, started, or exported by group.
- Scheduled Execution: For details about how to configure scheduled execution, see Scheduling Job Execution. Retain the default value No.
- **Concurrent Extractors**: Enter the number of extractors to be concurrently executed. Retain the default value **1**.
- Write Dirty Data: Specify this parameter if data that fails to be processed or filtered out during job execution needs to be written to OBS. Before writing dirty data, create an OBS link. You can view the data on OBS later. Retain the default value No so that dirty data is not recorded.
- vi. Click **Save and Run**. On the **Job Management** page, you can view the job execution progress and result.

Figure 2-33 Job progress and execution result

bleTile Migration Entire DE	Migration Links	Agents Settings							
🕀 Creele Job 🌐 🖓 Run	2 Delete				G Feedback	C Expert 🛛 Im	port Schedule	* Al statuses	Job name Job name or link type Q
9≠> π <	Name J⊟	Link Delaits	Created By JE	Last Execution Time 4	Duration 48	Write Statistics	Status	Group Name	Operation
Inter a group name. Q	🗌 rds_ts_dl	source_rds-dest_dli	ei_dics_6603	Apr 02, 2022 15:45:32 GMT+08:00	1m 27s	Winkten rows: 3	Succeeded	DEFAULT	Ran Historical Record Edit Mare +
kroups									
FAULT									

Step 3: Query Results

After the migration job is complete, log in to the DLI management console and click **SQL Editor**. In the displayed page, set **Engine** to **spark**, **Queue** to the created SQL queue, and **Database** to the database created in **Create a database and table on DLI**. Execute the following query statement and check whether the table data has been migrated to the **tabletest** table: select * from tabletest;

Figure 2-34 Querying data in the table

Engine spark v Queues test	▼ Databases testdb ▼	Execute Format	Refer Syntax Settings	More +
1create table tabletest(id string,name string);				
3 select * from tabletest;				
Line 1, Column 1	6030	Execute: Ctrl+Enter, Find: Ctrl+F, Format	Shift+Alt+F, Verify Syntax: Ctrl+Q,	Fullscreen: F11
Executed Queries (Last Day) View Result				Clear All
Result1 O				
Executed successfully				
Querycreate table tabletest(id string,name string); select * from tabletest				
Job ID 9616dbe4-9fd5-49fd-8d6f-3e10b771d318				
The query takes 27.07s, and 1.05 KB scanned.A maximum of 1,000 records can be displayed.		Ent	er a keyword. Q	F C 7
id 1≣	name J≣			
456	efg			
789	hij			
123	abc			

2.6 Migrating Data from GaussDB(DWS) to DLI

This section describes how to use the CDM data synchronization function to migrate data from GaussDB(DWS) to DLI.

Prerequisites

• You have created a DLI SQL queue. For details about how to create a DLI queue, see **Creating a Queue**.

When you create a queue, set its **Type** to **For SQL**.

- You have created a GaussDB(DWS) cluster. For details about how to create a GaussDB(DWS) cluster, see Creating a Cluster.
- You have created a CDM cluster. For details about how to create a CDM cluster, see Creating a CDM Cluster.

NOTE

- If the destination data source is an on-premises database, you need the Internet or Direct Connect. When using the Internet, ensure that an EIP has been bound to the CDM cluster, the security group of CDM allows outbound traffic from the host where the off-cloud data source is located, the host where the data source is located can access the Internet, and the connection port has been enabled in the firewall rules.
- If the data source is GaussDB(DWS) or MRS on a cloud, the network must meet the following requirements:

i. If the CDM cluster and the cloud service are in different regions, a public network or a dedicated connection is required for enabling communication between the CDM cluster and the cloud service. If the Internet is used for communication, ensure that an EIP has been bound to the CDM cluster, the host where the data source is located can access the Internet, and the port has been enabled in the firewall rules.

ii. If the CDM cluster and the cloud service are in the same region, VPC, subnet, and security group, they can communicate with each other by default. If the CDM cluster and the cloud service are in the same VPC but in different subnets or security groups, you must configure routing rules and security group rules.

For details about how to configure routes, see **Configure routes**. For details about how to configure security groups, see section **Security Group Configuration Examples**.

iii. The cloud service instance and the CDM cluster belong to the same enterprise project. If they do not, you can modify the enterprise project of the workspace.

In this example, the VPC, subnet, and security group of the CDM cluster are the same as those of the GaussDB(DWS) cluster.

Step 1: Prepare Data

- Create a database and table in the GaussDB(DWS) cluster.
 - a. Connect to the existing GaussDB(DWS) cluster by referring to Using the gsql CLI Client to Connect to a Cluster.

- b. Connect to the default database **gaussdb** of a GaussDB(DWS) cluster. gsql -d gaussdb -h *Connection address of the GaussDB(DWS) cluster* -U dbadmin -p 8000 -W *password* -r
 - gaussdb: Default database of the GaussDB(DWS) cluster
 - Connection address of the DWS cluster: If a public network address is used for connection, set this parameter to Public Network Address or Public Network Access Domain Name. If a private network address is used for connection, set this parameter to Private Network Address or Private Network Access Domain Name. For details, see Obtaining the Cluster Connection Address. If an ELB is used for connection, set this parameter to ELB Address.
 - dbadmin: Default administrator username used during cluster creation
 - -W: Default password of the administrator
- c. Run the following command to create the **testdwsdb** database: CREATE DATABASE testdwsdb;
- d. Run the following command to exit the **gaussdb** database and connect to **testdwsdb**:

gsql -d testdwsdb -h *Connection address of the GaussDB(DWS) cluster* -U dbadmin -p 8000 -W *password* -r

e. Run the following commands to create a table and import data to the table.

Run the following command to create a table: CREATE TABLE table1(id int, a char(6), b varchar(6),c varchar(6))

Run the following statements to insert data into the table: INSERT INTO table1 VALUES(1,'123','456','789'); INSERT INTO table1 VALUES(2,'abc','efg','hif');

f. Query the table data to verify that the data is inserted. select * from table1;

Figure 2-35 Querying data in the table

id	a	b	I	c	table1;	
1	123 abc	-+ 456 efg	Ì	789		

- Create a database and table on DLI.
 - a. Log in to the DLI management console and click **SQL Editor**. On the displayed page, set **Engine** to **spark** and **Queue** to the created SQL queue.

Enter the following statement in the editing window to create a database, for example, the migrated DLI database **testdb**: For details about the syntax for creating a DLI database, see **Creating a Database**. create database testdb;

b. In **SQL Editor**, select **testdb** for **Database** and run the following table creation statement to create a table in the database: For details about

the table creation syntax, see **Creating a DLI Table Using the DataSource Syntax**. create table tabletest(id INT, name1 string, name2 string, name3 string);

Step 2: Migrate Data

- 1. Create a CDM connection to MRS Hive.
 - a. Create a connection to the GaussDB(DWS) database.
 - i. Log in to the CDM console, choose **Cluster Management**. On the displayed page, locate the created CDM cluster, and click **Job Management** in the **Operation** column.
 - ii. On the Job Management page, click the Links tab, and click Create Link. On the displayed page, select Data Warehouse Service and click Next.
 - iii. Configure the connection. The following table describes the required parameters.

Parameter	Value
Name	Name of the GaussDB(DWS) data source, for example, source_dws .
Database Server	Click Select next to the text box to select the name of the created GaussDB(DWS) cluster.
Port	Port number of the GaussDB(DWS) database. The default value is 8000 .
Database Name	Name of the GaussDB(DWS) database you want to migrate The testdwsdb database created in Create a database and table in the GaussDB(DWS) cluster is used in this example.
Username	Username used for accessing the database. This account must have the permissions required to read and write data tables and metadata.
	In this example, the default administrator dbadmin specified when you create the GaussDB(DWS) database is used.
Password	Password of the GaussDB(DWS) database user.

Table 2-14 GaussDB(DWS) data source configuration

5	5 5	``
Cluster Management / cdm-test-0	/ Links / Edit Link	
* Name	source_dws]
* Connector	Relational Database	
Database Type	Data warehouse 🔍	
* Database Server (?)	dws-demog.dws.myhuaweiclouds	Select
* Port (?)	8000]
* Database Name	testdwsdb]
* Username	dbadmin]
* Password (?)	····· &]
Use Agent	Yes No	
Show Advanced Attributes		
X Cancel	Save	

Figure 2-36 Configuring the GaussDB(DWS) connection

For other parameters, retain the default values. For details, see Link to Relational Databases. Click **Save** to complete the configuration.

- b. Create a connection to the DLI.
 - i. Log in to the CDM console, choose **Cluster Management**. On the displayed page, locate the created CDM cluster, and click **Job Management** in the **Operation** column.
 - ii. On the Job Management page, click the Links tab, and click Create Link. On the displayed page, select Data Lake Insight and click Next.

Data Warehouse	Data Warehouse Service	Data Lake Insight			
Hadoop	MRS HDFS	MRS HBase	MRS Hive	Apache HDFS	
	Apache HBase	Apache Hive			
Object Storage	Object Storage Service (OBS)	Alibaba Cloud OSS			
File System	FTP	SFTP	HTTP		
Relational Database	RDS for MySQL	RDS for PostgreSQL	RDS for SQL Server	MySQL	
	PostgreSQL	Microsoft SQL Server	Oracle	IBM Db2	
NoSQL	Redis	MongoDB			
Messaging System	Data Ingestion Service	MRS Kafka	Apache Kafka		
Search	Elasticsearch				

Figure 2-37 Selecting the DLI connector

i. Create a connection to link CDM to DLI. For details about parameter settings, see Link to DLI.



Cluster Management / cdm-lest-00378326 / Links / Create Link				
(1) Select Connector		2 Configure		
* Name	dest_di			
* Connector	DLI ···			
* AK 🕐				
* sk 🕜	······ · · · · · · · · · · · · · · · ·			
* Project ID	05/ 1			
X Cancel	C Previous of Test Save			

After the configuration is complete, click **Save**.

- 2. Create a CDM migration job.
 - a. Log in to the CDM console, choose **Cluster Management**. On the displayed page, locate the created CDM cluster, and click **Job Management** in the **Operation** column.
 - b. On the **Job Management** page, choose the **Table/File Migration** tab and click **Create Job**.
 - c. On the **Create Job** page, specify job information.

Figure 2-39 Configuring the migration job

Jab Cenfiguration * Januar Source Jab Cenfiguration * Januar *	part Exit: Monsular	(2) Map Teas	(3) Configure Task
Source Job Configuration Destination Job Configuration * lancola takes = ////////////////////////////////////	ob Configuration		
	Job Name Bed		
0x 500 Statement () 10x	Source Job Configuration	Destination Job Configuration	
1 Sharan Balance ()	* Source Link Name source_dvs	Destination Link Name dest_dil	
	Use 501 Statement () Yes No	* Resource Garave (1) Sealedii 💿	
a Take King (D) Matt (D)	* Schema/Table Space 🛞 public 💿	* Extenses (2) feeddo (2)	
- 100 Min	Table Name (7) Table1	* Table (2) Inteleast (3)	
Store Advanced Alababas	Show Advanced Altibulan	Clear data before import 🛞 Ven No	
× Cated 2 test	× Cancal Next Save		

- i. Job Name: Name of the data migration job, for example, test
- ii. Set parameters required for **Source Job Configuration**.

Parameter	Value
Source Link Name	Select the name of the data source created in 1.a .
Use SQL Statement	When Use SQL Statement is set to Yes , enter an SQL statement here. CDM exports data based on the SQL statement. In this example, set this parameter to No .

 Table 2-15
 Source job configuration parameters

Parameter	Value
Schema/Table Space	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.
	In this example, no schema is created in Create a database and table in the GaussDB(DWS) cluster. In this case, set this parameter to the default value public .
	If the desired schema or tablespace is not displayed, check whether the login account has the permissions required to query metadata.
	NOTE The parameter value can contain wildcard characters (*), which is used to export all databases whose names start with a certain prefix or end with a certain suffix. For example:
	SCHEMA* indicates that all databases whose names starting with SCHEMA are exported.
	*SCHEMA indicates that all databases whose names ending with SCHEMA are exported.
	SCHEMA indicates that all databases whose names containing SCHEMA are exported.
Table Name	Name of the table you want to migrate. In this example, table1 created in Create a database and table in the GaussDB(DWS) cluster is used.

For details about parameter settings, see **From a Relational Database**.

iii. Set parameters required for **Destination Job Configuration**.

Parameter	Value
Destination Link Name	Select the DLI data source connection.
Resource Queue	Select a created DLI SQL queue.
Database	Select a created DLI database. In this example, database testdb created in Create a database and table on DLI is selected.

Parameter	Value
Table	Select the name of a table in the database. In this example, table tabletest created in Create a database and table on DLI is created.
Clear data before import	Whether to clear data in the destination table before data import. In this example, set this parameter to No .
	If this parameter is set to Yes , data in the destination table will be cleared before the task is started.

For details about parameter settings, see **To DLI**.

- iv. 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.
 - If the type is automatically created at the migration destination, you need to configure the type and name of each field.
 - CDM allows for field conversion during migration. For details, see **Field Conversion**.

Figure 2-40 Field mapping

() Configure Basic Information			0	Map Field		() Contigure Task
Source Field			0 /	Destination Field		
Name Example Value	Type	Operation		Nome	Type	Operation
×	INT	8 0	Q Ø 0	- (m) H	н	α
a	CHARIE	8 0	۹ ø	- () namet	string	σ
b	WRCHAR(6)	2 (Q 12 0		ating	a
4	VARCHAR(6)	8 0	9 Ø	- () nome3	string	σ
X Caniel Previous > Next 28m						

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

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

- **Retry Upon Failure**: If the job fails to be executed, you can determine whether to automatically retry. Retain the default value **Never**.
- **Group**: Select the group to which the job belongs. The default group is **DEFAULT**. On the **Job Management** page, jobs can be displayed, started, or exported by group.
- Scheduled Execution: For details about how to configure scheduled execution, see Scheduling Job Execution. Retain the default value No.
- **Concurrent Extractors**: Enter the number of extractors to be concurrently executed. Retain the default value **1**.
- Write Dirty Data: Specify this parameter if data that fails to be processed or filtered out during job execution needs to be written to OBS. Before writing dirty data, create an OBS link. You

can view the data on OBS later. Retain the default value **No** so that dirty data is not recorded.

vi. Click **Save and Run**. On the **Job Management** page, you can view the job execution progress and result.

Figure 2-41 Job progress and execution result

Table File Migration Entire I	38 Migration Links Ager	nts Settings								
() Create Job di Pun	Colete) Feedback	ort 🖸 Import	Schedule	* All statuses	• Job name • Job name or link type	QC
⊙≠⊳≌ <	Name 48	Link Details	Created By ↓Ξ	Last Execution Time 48	Duration 4E	Write Statistics	Status	Group Name	Operation	
Enter a group name. Q	L rest	source_ovs=dest_di	(00378328	Apr 06, 2022 17:34:50 GMT+08:00	165	Willes rows: 2	Succeeded	DEFAULT	Run Historical Record Edit More +	
Groups										
DEFAULT										

Step 3: Query Results

After the migration job is complete, log in to the DLI management console and click **SQL Editor**. In the displayed page, set **Engine** to **spark**, **Queue** to the created SQL queue, and **Database** to the database created in **Create a database and table on DLI**. Execute the following query statement and check whether the table data has been migrated to the **tabletest** table: select * from tabletest;

Figure 2-42 Querying data in the table

		Engine spark • Queues testdi • Databases	testdb • O Execute Format Refer Syntax Settings More
1 create database test	tdb;		
2create table tablet	est(id DWT, namei string, name2 string, name3 string	B);	
3			
4 select * from tablete:	st		
Line 1. Column 1			Execute: Clrl+Enter, Find: Clrl+F, Format: Shift+All+F, Verify Syntax: Clrl+Q, Fullscreen:
executed Queries (Last Day)	View Result		Clear
Received General (cash pary)			
Result1 0			
Nesani U			
Executed successfully			
Executed successfully	testdb;create table tabletestjid INT, name1 string, name2	string, name3 string); sel	
Executed successfully Querycreate database t	testdb;create table tabletest(jd INT, name1 string, name2 1c-s214-3087eb40e0b5	string, name3 string); sel	
Executed successfully Querycreate database t Job ID fbda48c8-6213-40	1c-a214-3087eb40e0b5		
Executed successfully Querycreate database t Job ID fbda48c8-6213-40			Enter a keyword. Q.) (bb.) [C]
Executed successfully Querycreate database t Job ID fbda48c8-6213-40 The query takes 3.99s, and 0.9	1c-a214-3087eb40e0b5 16 KB scanned.A maximum of 1,000 records can be displayed	ed.	
Executed successfully Querycreate database t Job ID fbda48c8-6213-40 The query takes 3.99s, and 0.9	1c-a214-3087eb40e0b5		[Inter-a keyword Q]
Executed successfully Querycreate database t Job ID fbda48c8-6213-40 The query takes 3.99s, and 0.9	1c-a214-3087eb40e0b5 16 KB scanned.A maximum of 1,000 records can be displayed	ed.	
Executed successfully Querycreate database t Job ID fbds48c8-6213-40 The query takes 3.99s, and 0.9 iel JE	1c-a214-3087eb40e0b5 6 KB scanned A maximum of 1,000 records can be displayed name1 4⊟	d. namež (E	name8 ↓⊟
Executed successfully Querycreate database t Job ID fbda48c8-6213-40 The query takes 3.99s, and 0.9 id 4E	1c-a214-3087eb40e0b5 6 KB scanned A maximum of 1,000 records can be displayed name1 4⊟	d. namež (E	name8 ↓⊟

3 Data Analysis

3.1 Analyzing Driving Behavior Data

Application Scenarios

Cloud computing and big data provide companies with data analysis and mining capabilities required in the Internet of Vehicle (IoV) field, helping companies or department of motor vehicles manage and analyze vehicle and driving behavior data quickly and scientifically.

Solution Architecture

DLI can query the records of vehicle driving features based on the detail records and freight order data regularly reported by the freight forwarder.

Data Types describes the data types used by DLI to record the data.

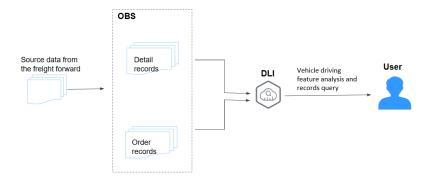


Figure 3-1 Solution Overview

Process

To use DLI to analyze driving behavior data, perform the following steps:

Step 1: Uploading Data. Upload the data to OBS.

Step 2: Analyzing Data. Use DLI to query the data.

Example Code

Download the **data package** for sample data and detailed SQL statements.

Solution Advantages

- Free of data migration: DLI can interconnect with multiple data sources. You only need to create SQL tables and map data sources.
- Easy to use: You can use standard SQL statements to compile metric analysis logic without paying attention to the complex distributed computing platform.
- Pay-per-use: Log analysis is scheduled periodically based on time-critical requirements. There is a long idle period between every two scheduling operations. DLI uses the pay-per-use billing mode, which effectively reduces your costs.

Resource Planning and Costs

Resource	Description	Cost		
OBS	You need to create an OBS bucket and upload	You will be charged for using the following OBS resources:		
	data to OBS for data analysis using DLI.	• Storage Fee for storing static website files in OBS.		
		 Request Fee for accessing static website files stored in OBS. 		
		 Traffic Fee for using a custom domain name to access OBS over the public network. 		
		The actual fee depends on the size of the stored file, the number of user access requests, and the traffic volume. Estimate the fee based on your service requirements.		
DLI	Before creating a SQL job, you need to purchase a queue. When	For example, if you purchase a pay-per- use queue, you will be billed based on the number of CUHs used by the queue.		
	using queue resources, you are billed based on the CUH of the queue.	Usage is billed by the hour. For example, 58 minutes of usage will be rounded to the hour. CUH pay-per-use billing = Unit price x Number of CUs x Number of hours.		

 Table 3-1 Resource planning and costs

Data Types

• Detail records

Detail records include the regularly reported location records and data of alarms triggered by abnormal driving behavior.

Field	Data Type	Description
driverID	string	Driver ID
carNumber	string	License plate number
latitude	double	Latitude
longitude	double	Longitude
speed	int	Speed
direction	int	Direction
siteName	string	Site name
time	timestamp	Report time of the records
isRapidlySpeedup	int	Whether the vehicle rapidly speeds up. 1 indicates that the vehicle suddenly speeds up, and 0 indicates that the vehicle does not.
isRapidlySlowdown	int	Whether the vehicle suddenly slows down.
isNeutralSlide	int	Whether the vehicle is coasting.
isNeutralSlideFinished	int	Whether vehicle coasting has stopped.
neutralSlideTime	bigint	Time length of vehicle coasting.
isOverspeed	int	Whether the vehicle is speeding.
isOverspeedFinished	int	Whether the vehicle stops speeding.
overspeedTime	bigint	Duration of the vehicle speeding
isFatigueDriving	int	Whether fatigue driving occurs.

Field	Data Type	Description
isHthrottleStop	int	Whether the driver revs the engine in neutral.
isOilLeak	int	Abnormal oil consumption

• Order data

Order data refers to the records of freight orders.

Table 3-3 Order data

Field	Data Type	Description
orderNumber	string	Order ID
driverID	string	Driver ID
carNumber	string	License plate number
customerID	string	Customer ID
sourceCity	string	Departure
targetCity	string	Destination
expectArriveTime	timestamp	Expected delivery time
time	timestamp	Time when a record is generated.
action	string	Event type, including creating an order, dispatching goods, delivering packages, and signing orders.

Step 1: Uploading Data

Upload the data to OBS for data analysis using DLI.

- 1. Download OBS Browser+. For details about the download address, see **Object Storage Service Tool Guide**.
- 2. Install OBS Browser+. For details about the installation procedure, see **Object Storage Service Tool Guide**.
- 3. Log in to OBS Browser+. OBS Browser+ supports two login modes: AK login (using access keys) or authorization code login. For details about the login procedure, see **Object Storage Service Tool Guide**.
- 4. Upload data using the OBS browser+.

Start the OBS Browser+, click **Create Bucket** on the homepage. Select a region and enter a bucket name (for example, **DLI-demo**). After the bucket is created, return to the bucket list and click **DLI-demo**. OBS Browser+ supports

upload by dragging. You can drag one or more files or folders from a local path to the object list of a bucket or a parallel file system on OBS Browser+. You can even drag a file or folder directly to a specified folder on OBS Browser+.

Obtain the test data by downloading the **Best_Practice_01.zip** file and decompressing it. Perform the following operations:

- Detail records: Upload the **detail-records** folder in the **Data** directory to the root directory of the OBS bucket.
- Order data: Upload the order-records folder in the Data directory to the root directory of the OBS bucket.

Step 2: Analyzing Data

Use DLI to query the data for analysis.

- 1. Creating a Database and a Table
 - a. On the homepage of the management console, choose **Service List** > **Analytics** > **Data Lake Insight**.
 - b. On the DLI console, click **SQL Editor**.
 - c. In the left pane of the SQL Editor, select the **Databases** tab and click $\textcircled{\Theta}$ to create the **demo** database.

Figure	3-2	Creating	а	database
--------	-----	----------	---	----------

Create Database

'ou can create 883 mo	re databases. Increase quota.	
★ Database Name	demo	
* Enterprise Project	default C C C Create Enterprise Project	
Description		
		4
		0/256
Tags	It is recommended that you use TMS's predefined tag function to add the same tag to different cloud resources. View predefined tags $\ C$	
	Tag key Tag value	
	You can add 10 more tags.	
	OK Cancel	

D NOTE

Database Name cannot be set to **default** because **default** is the built-in database.

d. Choose the **demo** database, and enter the following SQL statement in the editing box:

create table detail_records(driverID String, carNumber String, latitude double. longitude double, speed int, direction int, siteName String, time timestamp, isRapidlySpeedup int, isRapidlySlowdown int, isNeutralSlide int, isNeutralSlideFinished int, neutralSlideTime long, isOverspeed int, isOverspeedFinished int, overspeedTime long, isFatigueDriving int, isHthrottleStop int, isOilLeak int) USING CSV OPTIONS (PATH 'obs://dli-demo/detail-records/');

D NOTE

Replace the file path in the preceding statement with the actual OBS path where the detail records are stored.

e. Click **Execute** to create the **detail_records** table. See **Figure 3-3**.

Figure 3-3 Creating the detail_records table

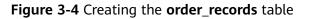


f. Run the following SQL statements to create the **event_records** table in the **demo** database. The operation is similar to **1.d** and **1.e**.

create table event_records(driverID String, carNumber String, latitude double, longitude double, speed int, direction int, siteName String, time timestamp, isRapidlySpeedup int, isRapidlySlowdown int, isNeutralSlide int, isNeutralSlideFinished int, neutralSlideTime long, isOverspeed int, isOverspeedFinished int, overspeedTime long, isFatigueDriving int, isHthrottleStop int, isOilLeak int)

- g. Run the following SQL statements to extract the alarm and event data from the detail records and insert it into the event_records table. insert into table event_records (select * from detail_records where isRapidlySpeedup > 0 OR isRapidlySlowdown > 0 OR isNeutralSlide > 0 OR isNeutralSlideFinished > 0 OR isOverspeed > 0 OR isOverspeedFinished > 0 OR isFatigueDriving > 0 OR isHthrottleStop > 0 OR isOilLeak > 0)
- h. Use another method to create the **order_records** table.

On the left of the SQL job editor, click the **Databases** tab and click the demo database. Click the plus icon (+) on the right of **Table** to create a table, and set **Data Location** to **DLI**. Set the column types according to **Order data**.



Create Table					
You can create 51 n	nore tables. Increase quota.				
* Name	order_records				
* Data Location	DLI		•	0	
Table Description	n				
			0/100	1	
* Column Name		* Type	Description	Operation	
Normal 👻	orderNumber	string 🛛 🔻		⊕ ⊕	
Normal 💌	driverID	string 🗸 🔻		ū ⊕	
Normal 💌	carNumber	string 🛛 💌		ū ⊕	
Normal 👻	customerID	string 🛛 💌		<u>ū</u> ⊕	
Normal 👻	sourceCity	string 🗸 🔻		⊕ ⊕	
Normal 👻	targetCity	string 🛛 🔻		₩ 🕀	
Normal 👻	expectArriveTime	timestamp 🔻		₩ ⊕	
Normal 👻	time	timestamp 👻		⊕ ⊕	
Normal 👻	action	string 🛛 🔻		<u>↓</u> (+)	
Number of Column column details fron	s: 9. If there are a large number of co n an Excel file. 🅜	ok	L statements to create a tal	ble or import	

i. Import the OBS data to the order_records table. Choose Data Management > Databases and Tables. Click the demo database to go to the table management page. In the Operation column of the order_records table, choose More > Import. Set File Format to CSV, the data storage path to obs://DLI-demo/order-records/, and retain default values for the rest parameters. Click OK.

D NOTE

The default timestamp format is **yyyy-MM-dd HH:mm:ss**. To use other formats, select **Advanced Settings** and enter the desired timestamp format (not modified in this example).

×

Figure 3-5 Importing table data

Import Data			
Database Name	demo		
Table Name	order_records		
★ File Format	CSV Set the options in Ad DLI supports the read of CSV data that is not compressed	_	
Queue	default	-	
★ Path	obs://DLI-demo/order-records/	Þ	
Advanced Settings			
	OK Cancel		

- 2. Querying Data
 - a. Run the following SQL statements to query the alarm events of all drivers in a certain time period.

NOTE

You can save the frequently-used query statements as a template by clicking **More** > **Save as Template** in the upper right corner of the editing window. The template is available for future use or can be modified in the SQL editor again.

Choose **Job Templates** > **SQL Templates** and click the **Custom Templates** tab. In the **Operation** column of the target template, click **Execute** to switch to the SQL editor. You can modify it as needed.

select

select
driverID,
carNumber,
sum(isRapidlySpeedup) as rapidlySpeedupTimes,
sum(isRapidlySlowdown) as rapidlySlowdownTimes,
sum(isNeutralSlide) as neutralSlideTimes,
sum(neutralSlideTime) as neutralSlideTimeTotal,
sum(isOverspeed) as overspeedTimes,
<pre>sum(overspeedTime) as overspeedTimeTotal,</pre>
sum(isFatigueDriving) as fatigueDrivingTimes,
sum(isHthrottleStop) as hthrottleStopTimes,
sum(isOilLeak) as oilLeakTimes
from
event_records
where
time >= "2017-01-01 00:00:00"
and time <= "2017-02-01 00:00:00"
group by
driverID,
carNumber
order by
rapidlySpeedupTimes desc,
rapidlySlowdownTimes desc,
neutralSlideTimes desc,
neutralSlideTimeTotal desc,
overspeedTimes desc,
overspeedTimeTotal desc,
fatigueDrivingTimes desc,
hthrottleStopTimes desc,
oilLeakTimes desc

In the query result, click \square to view graphical results.

- Set **Graph Type** to the bar chart.
- Set **X-AXIS** to **driverID**.
- Set Y-AXIS to rapidlySpeedupTimes.
- Set Results to 10.

The command output is as follows:

Figure 3-6 Rapid acceleration

Executed Que	erles (Last 7 Days)	View Res	ult							Clear All
Result1 O										
Executed suc	cessfully									
Query: select d	riverID, carNumber, sum	isRapidlySpee	dup) as rapidlySpeedup	oTimes, sum(isRi	pidlySlowdown) as rapid	lySlowdownTin	nes, sum(isNeutra	ISlide) as neutralSli	deTimes, sum(neutra	SlideTime) as neutralSlideTim
Job ID: e86f514	47-9ce9-4b78-baf3-d1cfc	107d271								
The query take	s 23.32s, and 7.95 MB sci	enned.A maxin	num of 1,000 records ci	an be displayed.				Show Table	Export Result	Submit Download Request
Graph Type:	Bar	v X-AXIS:	driverID	• Y-AXIS:	rapidlySpeedupTi	• Results:	10	¥		
					rapidlySpeedu	Times				
1,400										
1,200	1205	0	1111							
1,000				1075	1058	1047	1032	942		
800									846	731
600										731
400										
200										
200										
	ian1000005	zen	gpeng1000000	_	haowei1000008	_	zouan100000	7	duxu10000	19

b. Run the following SQL statement to query the detailed record of a driver in a certain time period.

```
select
*
from
event_records
where
driverID = "panxian1000005"
and time >= "2017-01-01 00:00:00"
and time <= "2017-02-01 00:00:00"</pre>
```

In the query result, click 🖮 to view graphical results.

- Set **Graph Type** to the bar chart.
- Set X-AXIS to driverID.
- Set Y-AXIS to speed.
- Set Results to 10.

The command output is as follows:

Figure 3-7 Speeding record

Executed Que	ries (Last 7 Days)	View Result						Clear All
Result1 🛛	Result2							
C Executed suc	cessfully							
Query: select *	from event_records where d	iriverID = "panxian1000005" a	ind time >= "2017-01-01"	00:00:00" and time <= "201	7-02-01 00:00:00*			
Job ID: 226352	58-15c7-42d8-a822-35d972	e104dc						
The query takes	s 10.53s, and 7.96 MB scann	ed.A maximum of 1,000 recor	ds can be displayed.			Show Table	Export Result	Submit Download Request
Graph Type:	Bar 💌	X-AXIS: driverID	• Y-AXIS: sp	eed 💌 P	Results: 10	¥		
				speed				
150				148				
120	130 121	120		120				
90			92			89		108
60					78		72	
30								
0								
panxia	n1000005	panxian1000005	panxi	an1000005	parxian1000005		panxian1000	005

c. Run the following SQL statement to query the order information.

from
order_records
where
orderNumber = "2017013013584419488"
order by
time desc

Figure 3-8 Order information

xecuted Queries (Last 7	Days) View	Result						Clear All
Result1 🕲 🕴 Result2 🧯	Result3 🕲							
Executed successfully								
Query: select * from order_re	cords where orderNu	mber = "2017013013	584419488" order by time des	c				
ob ID: 876d3cde-730a-49e6	-9a27-1d153aed7ee5	i i						
he query takes 9.06s, and 5	0.55 KB scanned.A ma	mimum of 1,000 reco	rds can be displayed.			Graphical Result Export Res	ult Submit (ownload Reques
orderNumber	driverID	carNumber	customerID	sourceCity	targetCity	expectArriveTime	time	action
2017013013584419488	zouan1000007	\$A58M83	zhujia151464313			Feb 1, 2017 1:58:35.000 GM	Jan 30, 20	
2017013013584419488	zouan1000007	SA58M83	zhujja151464313			Feb 1, 2017 1:58:35.000 GM	Jan 30, 20	

d. Run the following SQL statement to query a vehicle's driving feature according to the driver ID and time of departure.

```
select

driverID,

carNumber,

latitude,

longitude,

siteName,

time

from

detail_records

where

driverID = "panxian1000005"

and time > "2017-01-30 16:00:00"

and siteName IS NOT NULL

order by

time desc
```

In the query result, click to view graphical results.

- Set **Graph Type** to the bar chart.
- Set X-AXIS to time.
- Set Y-AXIS to latitude.
- Set **Results** to **10**.

The command output is as follows:

Figure 3-9 Driving information



3.2 Converting Data Format from CSV to Parquet

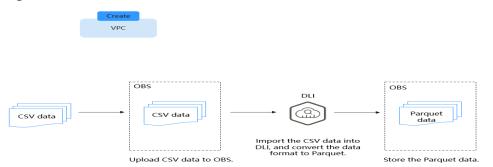
Application Scenarios

Parquet is a columnar storage substrate created for simpler data analysis. This format can speed up queries by allowing only the required columns to be read and calculated. In addition, Parquet is built to support efficient compression schemes, which maximizes the storage efficiency on disks. Using DLI, you can easily convert data format form CSV to Parquet.

Solution Overview

Upload CSV data to an OBS bucket, convert CSV data into Parquet data with DLI, and store the converted Parquet data to OBS.

Figure 3-10 Solution overview



Process

To use DLI to convert CSV data into Parquet data, perform the following steps:

Step 1: Creating and Uploading Data. Upload data to your OBS bucket.

Step 2: Using DLI to Convert CSV Data into Parquet Data. Import CSV data to DLI and convert it into Parquet data.

Solution Advantages

• The query performance is improved.

If you have text-based data files or tables in an HDFS and are using Spark SQL to query data, converting data format to Parquet can improve the query performance by about 30 times (or more in some cases), despite of the time consumed during the conversion.

• Storage is saved.

Parquet is built to support efficient compression schemes, which maximizes the storage efficiency on disks. With Parquet, the storage cost can be reduced by about 75%.

Resource Planning and Costs

Resource	Description	Cost
OBS	You need to create an OBS bucket and upload	You will be charged for using the following OBS resources:
	data to OBS for data analysis using DLI.	 Storage Fee for storing static website files in OBS.
		 Request Fee for accessing static website files stored in OBS.
		 Traffic Fee for using a custom domain name to access OBS over the public network.
		The actual fee depends on the size of the stored file, the number of user access requests, and the traffic volume. Estimate the fee based on your service requirements.
DLI	Before creating a SQL job, you need to purchase a queue. When	For example, if you purchase a pay-per- use queue, you will be billed based on the number of CUHs used by the queue.
	using queue resources, you are billed based on the CUH of the queue.	Usage is billed by the hour. For example, 58 minutes of usage will be rounded to the hour. CUH pay-per-use billing = Unit price x Number of CUs x Number of hours.

 Table 3-4 Resource planning and costs

Step 1: Creating and Uploading Data

1. Create a CSV file. See **test.csv** in **Figure 3-11**.

Figure 3-11 Creating a test.csv file

tes	t. csv🔀
1	1,2,3,4,5

2. In the OBS management console, create a bucket, name it **obs-csv-parquet**, and upload the **test.csv** file to the bucket.

Figure 3-12 Uploading CSV data to OBS

Bucket List > obs-csv-parquet							
Objects Deleted Objects	Fragments						
Objects are basic units of data stora Upload File Create Folder		ited as objects. You can u	pload any type of files and	manage them		n object name prefix.	QC
Name	Storage Class	Size	Encrypted	Restore	Last Modified 🗘	Operation	
test.csv	Standard	9 Bytes	No	-	Aug 07,2018 17:11:21 GMT+	Download Share More	

3. Create a bucket and name it **obs-parquet-data** to store the converted parquet data.

Step 2: Using DLI to Convert CSV Data into Parquet Data

- 1. Go to the DLI console, click **SQL Editor** in the navigation pane.
- 2. In the left pane of the SQL editor, click the **Databases** tab. Click ^(☉), create a database, and name it **demo**.
- 3. In the SQL editing window, set Engine to spark, Queue to default, and Database to demo. Execute the following statement to create table test_csv_hw to import the data in the test.csv file from OBS. create table test_csv_hw(id1 int, id2 int, id3 int, id4 int, id5 int) using csv options(path 'obs://obs-csv-parquet/test.csv'
- 4. In the SQL editing window, query data in the **test_csv_hw** table.

Figure 3-13 Querying data

select * from test_csv_hw;			
Line 1, Column 27		Execute: Ctrl	+Enter, Find: Ctrl+F, Format: Shift+Alt+F, Verify Syntax: Ctrl+Q, Fullscreen: F
cuted Queries (Last Day) View Result			Clear /
Result1 O			
Executed successfully			
Executed successfully			
Executed successfully uery select * from test_csv_hw	re displayed.		Enter a keyword. Q L
Executed successfully	be displayed. Iditi 4⊞	id4 i⊞	Enter a keyword. Q] [<u>H.</u>] [<u>C</u>] [Id5 42
Result1 🛛			

5. In the SQL job editing window, create a table to store the OBS data in Parquet format and name the table test_parquet_hw. create table `test_parquet_hw` (`id1` INT, `id2` INT, `id3` INT, `id4` INT, `id5` INT) using parquet options (

path 'obs://obs-parquet-data/'

You do not need to specify a file because no Parquet file exists in this OBS bucket before the data is converted.

6. In the SQL editing window, execute the following statement to convert the CSV data to Parquet format and store the data in the specified OBS folder:

insert into test_parquet_hw select * from test_csv_hw

7. Check the result. OBS automatically created a file for saving the result.

Figure 3-14 Parquet data saved in a file in OBS

Objects Deleted	bjects Fragments						
ojects are basic units o	data storage. In OBS, fil	es and folders are treat	ed as objects. You can u	upload any type of files and m	anage them in	your bucket. Learn more	
Upload File Cr	ate Folder Delet	e Restore				Ent	er an object name prefix.
Name		Storage Class	Size	Encrypted	Restore	Last Modified 🝦	Operation
A Desidence							
← Previous							

3.3 Analyzing E-commerce BI Reports

Application Scenarios

As a self-operated e-commerce company in China, the X mall has developed hundreds of millions of loyal users and accumulated massive amounts of authentic data while maintaining high-speed development. How to use the BI tool to find business opportunities from historical data is a key issue in the precision marketing of big data applications. It is also the core technology required for intelligent upgrade of all e-commerce platforms.

This case uses HUAWEI CLOUD DLI, GaussDB(DWS), and Yonghong BI to analyze data features of users and offerings based on the real user, product, and comment data (anonymized) of the mall, providing high-quality information for marketing decision-making, advertising recommendation, credit rating, brand monitoring, and user behavior prediction.

Process

To use DLI to analyze e-commerce data, perform the following steps:

Step 1: Uploading Data. Upload the data to OBS for data analysis using DLI.

Step 2: Analyzing Data. Use DLI to query the data for analysis.

Data Types

To protect user privacy and data security, all sampled data is anonymized.

User data

Table	3-5	User	data

Field	Data Type	Description	Value
user_id	int	User ID	Anonymized
age	int	Age group	-1 indicates that the user age is unknown.
gender	int	Gender	0: Male1: Female2: Confidential
rank	Int	User level	Sequenced list of user level. The higher the user level, the larger the number.
register_tim e	string	User registration date	Unit: day

• Product data

Table 3-6 Product data

Field	Data Type	Description	Value
product_id	int	Product No.	Anonymized
a1	int	Attribute 1	Enumerated value. The value -1 indicates unknown.
a2	int	Attribute 2	Enumerated value. The value -1 indicates unknown.
a3	int	Attribute 3	Enumerated value. The value -1 indicates unknown.
category	int	Category ID	Anonymized
brand	int	Brand ID	Anonymized

• Comment data

Table 3-7 Comment data

Field	Data Type	Description	Value
deadline	string	End time	Unit: day

Field	Data Type	Description	Value
product_id	int	Product No.	Anonymized
comment_num	int	Segments of accumulated comment count	 0: No comment 1: One comment 2: 2 to 10 comments 3: 11-50 comments 4: More than 50 comments
has_bad_comm ent	int	Whether there is negative feedback.	0: No; 1: Yes.
bad_comment_ rate	float	Dissatisfaction rate	Proportion of the negative feedback.

• Action data

Table 3-8 Action data

Field	Data Type	Description	Value
user_id	int	User ID	Anonymized
product_id	int	Product No.	Anonymized
time	string	Time of action	-
model_id	string	Module ID	Anonymized
type	string	 Browse (refers to the offering details page) Add to cart Remove from cart Place an order Follow Click 	-

Step 1: Uploading Data

Upload the data to OBS for data analysis using DLI.

- 1. Download OBS Browser+. For details about the download address, see **Object Storage Service Tool Guide**.
- 2. Install OBS Browser+. For details about the installation procedure, see **Object Storage Service Tool Guide**.

- 3. Log in to OBS Browser+. OBS Browser+ supports two login modes: AK login (using access keys) or authorization code login. For details about the login procedure, see **Object Storage Service Tool Guide**.
- 4. Upload data using the OBS Browser+.

On the OBS Browser+ page, click **Create Bucket**. Select a region and enter a bucket name (for example, **DLI-demo**). After the bucket is created, return to the bucket list and click **DLI-demo**. OBS Browser+ supports upload by dragging. You can drag one or more files or folders from a local path to the object list of a bucket or a parallel file system on OBS Browser+. You can even drag a file or folder directly to a specified folder on OBS Browser+.

Obtain the test data by downloading the **Best_Practice_04.zip** file, decompressing it, and uploading the **Data** folder to the root directory of the OBS bucket. The test data directory is as follows:

- data/JData_User: Data in the user table
- **data/JData_Product**:Data in the **product** table
- data/JData_Product/JData_Comment: Data in the comment table
- **data/JData_Action**: Data the **action** table

Step 2: Analyzing Data

- 1. Creating a Database and a Table
 - On the top menu bar of the portal page, choose Products > Analytics > Data Lake Insight (DLI).
 - b. Create a demo database. On the DLI console, choose Job Management
 >SQL Jobs. Click the created job on the displayed page to go to the SQL
 Editor page.
 - c. In the left pane of the SQL Editor, select the **Databases** tab and click to create the **demo** database. For details, see Figure 3-15.

Figure 3-15 Creating a database

Create Database

You can create 883 mo	pre databases. Increase quota.	
★ Database Name	demo	
★ Enterprise Project	default C ⑦ Create Enterprise Project	
Description		
		0/256
Tags	It is recommended that you use TMS's predefined tag function to add the same tag to different cloud resources. View predefined tags $\ C$	
	Tag key Tag value	
	You can add 10 more tags.	
	OK Cancel	

NOTE

The **default** database is a built-in database. You cannot create a database named **default**.

- d. Choose the **demo** database, and enter the following SQL statement in the editing box:
 - create table user(user_id int, age int, gender int, rank int, register_time string) USING csv OPTIONS (path "obs://DLI-demo/data/JData_User")

NOTE

The file path in the preceding SQL statement is the actual OBS path for storing data.

- e. Click **Execute** to create the user information table user.
- f. Create the **product**, **comment**, and **action** tables in the same way.
 - Product data
 create table product(
 product_id int,
 a1 int,
 a2 int,
 a3 int,
 category int,
 brand int
) USING csv OPTIONS (path "obs://DLI-demo/data/JData_Product")

- Comment table
 - create table comment(deadline string, product_id int, comment_num int, has_bad_comment int, bad_comment_rate float) USING csv OPTIONS (path "obs://DLI-demo/data/JData_Comment")
 - Action table create table action(user_id int, product_id int, time string, model_id string, type string) USING csv OPTIONS (path "obs://DLI-demo/data/JData_Action");
- 2. Querying Data

You can save common query statements as templates on the **Template Management** page for later use. For details, see **SQL Template Management** in *Data Lake Insight User Guide*.

- Top 10 products with the most likes
 - i. Run the following SQL statement to analyze the top 10 products with the most likes.

```
SELECT

product.brand as brand,

COUNT(product.brand) as like_count

from

action

JOIN product ON (action.product_id = product.product_id)

WHERE

action.type = 'like'

group by

brand

ORDER BY like_count desc

limit

10
```

ii. Click **Execute**. The execution results are displayed, as shown in **Figure 3-16**.

Figure 3-16 Querying results

he query takes 9.72s, and 4.15 KB scanned.A maximum of 1,000 records of	Graphical Result	Export Result	Submit Download Request	
brand	like_count			
400003	7			
400002	5			
400001	4			
400007	4			
400009	2			
400006	2			
400004	2			
400008	1			
400008	1			

iii. Click to view the result in a chart.

Figure 3-17 Graphical results

raph Type:	Bar	*	X-AXIS: brand	▼ Y-AXIS: I	ike_count	 Results: 10 	•	
				li	ke_count			
7	7							
6								
5		5						
4			4	4				
3								
2							2	
1								
0								

- Top 10 worst-rated products
 - i. Run the following SQL statement to analyze the top 10 worst-rated products:

SELECT
DISTINCT product_id,
comment_num,
bad_comment_rate
from
comment
where
comment_num > 3
order by
bad_comment_rate desc
limit
10

ii. Click **Execute**. The execution results are displayed, as shown in **Figure 3-18**.

Figure 3-18 Querying results

C Executed successfully											
Query: SELECT DISTINCT product_id, comment_num, bad_comment_rate from comment where comment_num > 3 order by bad_comment_rate desc limit 10											
Job ID: 5b7457f5-42f2-4b54-b757-6ald(70b0fb2a											
The query takes 6.53s, and 0.96 KB scanned.A max	imum of 1,000 records can be displayed.	Graphical Result Export Result Submit Download Request									
product_id	comment_num	bad_comment_rate									
200040	4	0.009									
200024	4	0.006									
200032	4	0.003									
200016	4	0.003									
200008	4	0.001									
200009	4	0									
200033	4	0									
200001	4	0									
200017	4	0									
200025	4	0									

iii. Click

Шı

to view the result in a chart.

Figure 3-19 Graphical result

h Type: Bar	*	X-AXIS: prod	uct_id	• Y-AXIS: ba	id_comment_rate	▼ Results:	10	*	
				bad_cor	mment_rate				
0.01									
0.009									
0.008									
0.006	0.006								
0.004									
		0.003	0.003						
0.002				0.001		0		0	0
200040	200024	200032	200016	200008	200009	200033	200001	200017	200025

You can also analyze data for age distribution, gender ratio, offering evaluation, purchase number, and browsing statistics of users.

3.4 Analyzing DLI Billing Data

Application Scenarios

You can analyze DLI billing data (account information has been masked) on the big data analysis platform of DLI, find possible optimization, and figure out some measures to reduce costs for using DLI.

Analysis Process

Perform the following steps to analyze billing data and reduce costs:

Step 1: Obtaining Consumption Data. Obtain billing data of an account.

Step 2: Analyzing Billing Data and Reducing Costs. Analyze the consumption data, find the resources or users with high expenditure, and provide optimization measures to reduce cost.

Resources and Costs

DLI DLI is a big data analytics platform on Spark jobs on DLI.	
 HUAWE LCLOUD You are billed for using storage and compute resources. DLI supports three billing modes: yearly/monthly, package, and pay-per-use. For SQL jobs, you are billed for both storage and compute resources can be billed bar on a yearly/monthly basis or pay-peruse. If you choose the yearly/monthly basis or pay-peruse. If you choose the yearly/monthly basis or pay-peruse. If you choose the yearly/monthly billing mode, fees are deducted be on the subscription period. In pay-per-use mode, fees are deducted by hour. You can choose either billing by CUH or by the amount of data scanned. Billing 1 CUH is recommended, for you can have exclusive resources and clear costing. In addition, you can purce and use packages. Billing for CUH used = Number CUs x Usage duration x Unit price. If a computing task times out or f no fee is charged for the task. For Flink and Spark jobs, you will billed for compute resources only billing rules are same to those of jobs. 	sed r- based illing e by n r hase r of rice. n is an hour. t ails, t be t The

Step 1: Obtaining Consumption Data

- 1. Obtain billing details.
 - a. Log in to the DLI console.
 - b. Click **Billing & Costs** on the upper right corner of the page. Choose **Bills**.

Figure 3-20 Bills

Search	Q	Billing & Costs				
	Unpaid Orders					
	Renewal					
npatible with Apach ata ETL. DLI suppo	My Packages					
	Bills					
	Invoices					
link Jobs	Cost Center					

c. On the **Dashboard** page of the **Billing Center**, click **Expenditure Details**. On the displayed page, set **Data Type** to **Usage Type** and **Data Period** to **Details**. Set time to the billing cycle you want.

In the title row of the displayed table, set **Service Type** to **Data Lake Insight (DLI)** and **Resource Type** to **DLI cuh**. Click **Export**. On the **Export** page, configure **Export Content** and **Period** as you need, and click **Export**. The **Export History** page is displayed.

Figure 3-21 DLI Bills

enditure l	Details													1 Hel	ip Ce
Expenditure		t updated in	real time. For rea	al-time data, see Expenditu at is equal to the used nurr		d by the unit price. O	ther pricing modes, such as	tiered pricing, do no	t provide unit prices.					More	~
ata Type	Usage Ty	pe -	Resource	Resource Type	Service Type	Account									
ata Period	By billing	cycle	By day	Details											
Export			-				Apr 2022 • R	isource ID 💌		Q Enter an	order No. or trans	action No.	Q © Custo	omize Colun	nn
Billing	Enterp 🕅	Account	Service T 7	Resource Type 🍸	Billing 🍞	Expenditure Time	Order No./Transaction	Bill Type 🍞	Transaction Time	Resource Na	Resource Tag	Specifications	Region 🍞	AZ	
Apr 2022	default			DLI cuh	Pay-per-Use	Apr 01, 2022 14: Apr 01, 2022 15:		Expenditure	Apr 01, 2022 15:	testqueue_ei	-	DLI x86 Ond	-	AZ1	
Apr 2022	default			DLI cuh	Pay-per-Use	Apr 01, 2022 14: Apr 01, 2022 15:		Expenditure	Apr 01, 2022 15:			DLI x86 Ond	(AZ1	
Apr 2022	default			DLI cuh	Pay-per-Use	Apr 01, 2022 14: Apr 01, 2022 15:		4 Expenditure	Apr 01, 2022 15:			DLI x86 Ond	i.	AZ1	
Apr 2022	default			DLI cuh	Pay-per-Use	Apr 01, 2022 13: Apr 01, 2022 14:		Expenditure	Apr 01, 2022 14:			DLI x85 Ond		AZ1	

d. On the **Export History** page, wait until the file status changes to **Successful**. Click **Download**.

Step 2: Analyzing Billing Data and Reducing Costs

- 1. Analyze billing details.
 - a. Upload the billing details downloaded in **Step 1: Obtaining Consumption Data** to the created OBS bucket.

- b. Create a table on DLI.
 - i. Log in to the DLI console. In the navigation pane, choose **SQL Editor**. Select **spark** for **Engine**, and select the queue and database. In this example, the default queue and database are used.
 - ii. The downloaded file contains information such as time and usage. Create a table on DLI based on these table headers. For details, see the following example

the following example. CREATE TABLE `spending` (account_period string, EnterpriseProject string, EnterpriseProjectID string, accountID string, product_type_code string, product_type string, product_code string, product_name string, product_id string, mode string, time1 string, use_start string, use_end string, orderid string, ordertime string, resource_type string, resource_id string, resouce_name string, tag string, skuid string, `c22name` STRING, `c23name` STRING, `c24name` STRING, `c25name` STRING, `c26name` STRING, `c27name` STRING. `c28name` STRING, `c29name` STRING, size STRING, `c31name` STRING, `c32name` STRING, `c33name` STRING, `c34name` STRING. `c35name` STRING, `amount` STRING, `c37name` STRING, `c38name` STRING, `c39name` STRING, `c40name` STRING, `c41name` STRING, `c42name` STRING, `c43name` STRING, `c44name` STRING, `c45name` STRING, `c46name` STRING, `c47name` STRING, `c48name` STRING, `c49name` STRING, `c50name` STRING, `c51name` STRING, `c52name` STRING, `c53name` STRING, `c54name` STRING) USING csv options (path 'obs://xxx/Spendings(ByTransaction)_20200501_20200531.csv', header true)

c. Query **resource_id** and **resource_name** with the highest amount within the period.

The following statement shows the amount charged for using the SQL and Flink queues.

select resource_id, resouce_name, sum(size) as usage, sum(amount) as sum_amount from spending group by resource_id, resouce_name order by sum_amount desc

Figure 3-22 Query results

resource_id	resouce_name	usage	sum_amount
d91d4616-b10c-471a-820d-e676e6c514b4	sql	5264	1842.3999999999895
8163cc27-89ce-4bac-aa85-38cb753ee425	flink	5264	1842.3999999999896
9bd0736b-f8ca-4bfb-b3e7-0e391ef7dd8b	nul	48	14.3999999999999999
dd3a12ff-c0af-4ad1-bbc1-858bf4d3661c	ditest	32	11.2
f8265ef5-eb5f-4eff-b8d6-9ca91ed20009	test	16	5.6

d. Run the following statements to analyze the usage periods of SQL and Flink resources:

select * from spending where resource_id = 'd91d4616-b10c-471a-820d-e676e6c5f4b4' order by ordertime

The SQL queue was billed each hour from May 14 2020 17:00:00 GMT +08:00 to May 28, 2020 10:00:00 GMT+08:00.

Similarly, the Flink queue was continuously used from May 14, 2020 17:00:00 GMT+08:00 to May 28 2020 10:00:00 GMT+08:00.

2. Suggestion for reducing the cost

You can change the SQL and Flink queues to yearly/monthly queues for lower costs. If you are sure about the number of CUHs required for a job, you can purchase a package to reduce the cost.

DLI helps you to analyze billing details of your enterprise to quickly find the unreasonable expenses and control costs. You can also use DLI to reduce your cost on HUAWEI CLOUD.

3.5 Using DLI Flink SQL to Analyze e-Commerce Business Data in Real Time

Application Scenarios

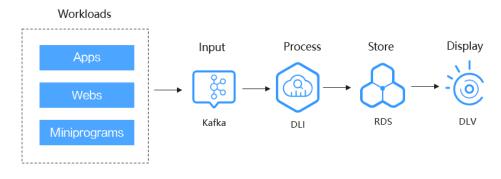
Online shopping is very popular for its convenience and flexibility. e-Commerce platform can be accessed via an array of methods, such as visiting the websites, using shopping apps, and accessing through mini-programs. A large volume of statistics data such as the real-time access volume, number of orders, and number of visitors needs to be collected and analyzed on each e-commerce platform every day. These data needs to be displayed in an intuitive way and updated in real time to help managers learn about data changes in a timely manner and adjust marketing policies accordingly. How can we efficiently and quickly collect statistics based on these metrics?

Assume the order information of each offering is written into Kafka in real time. The information includes the order ID, channel (websites or apps), order creation time, amount, actual payment amount after discount, payment time, user ID, username, and region ID. We need to collect statistics on such information based on metrics of each sales channel in real time, store the statistics in a database, and display the statistics on screens.

Solution Overview

The following figure gives you an overview to user DLI Flink to analyze and process real-time e-commerce business data and sales data of all channels.

Figure 3-23 Solution overview



Process

To analyze real-time e-commerce data with DLI Flink, perform the following steps:

Step 1: Creating Resources. Create resources required for creating jobs belong to your account, including VPC, DMS, DLI, and RDS.

Step 2: Obtaining the DMS Connection Address and Creating a Topic. Obtain the connection address of the DMS Kafka instance and create a DMS topic.

Step 3: Creating an RDS Database Table. Obtain the private IP address of the RDS DB instance and log in to the instance to create an RDS database and MySQL table.

Step 4: Creating an Enhanced Datasource Connection. Create an enhanced datasource connection for the queue and test the connectivity between the queue and the RDS instance and the queue and the DMS instance, respectively.

Step 5: Creating and Submitting a Flink Job. Create a DLI Flink OpenSource SQL job and run it.

Step 6: Querying the Result. Query the Flink job results and display the results on a screen in DLV.

Solution Advantages

- Cross-source analysis: You can perform association analysis on sales summary data of each channel stored in OBS. There is no need for data migration.
- SQL only: DLI has interconnected with multiple data sources. You can create tables using SQL statements to complete data source mapping.

Resource Planning and Costs

Resource	Description	Cost
OBS	You need to create an OBS bucket and upload data to OBS for data analysis using DLI.	 You will be charged for using the following OBS resources: Storage Fee for storing static website files in OBS. Request Fee for accessing static website files stored in OBS. Traffic Fee for using a custom domain name to access OBS over the public network. The actual fee depends on the size of the stored file, the number of user access requests, and the traffic volume. Estimate the fee based on your service requirements.
DLI	Before creating a SQL job, you need to purchase a queue. When using queue resources, you are billed based on the CUH of the queue.	For example, if you purchase a pay-per-use queue, you will be billed based on the number of CUHs used by the queue. Usage is billed by the hour. For example, 58 minutes of usage will be rounded to the hour. CUH pay-per-use billing = Unit price x Number of CUs x Number of hours.
VPC	You can customize subnets, security groups, network ACLs, and assign EIPs and bandwidths.	The VPC service is free of charge. EIPs are required if your resources need to access the Internet. EIP supports two billing modes: pay-per-use and yearly/monthly. For more, see VPC Billing Description .
DMS Kafka	Kafka provides premium instances with computing, storage, and exclusive bandwidth resources.	Kafka supports two billing modes: pay-per-use and yearly/monthly. Billing items include Kafka instances and Kafka disk storage space. For details, see DMS for Kafka Billing Description.
RDS MySQL	RDS for MySQL provides online cloud database services.	You are billed for RDS DB instances, database storage, and backup storage (optional). For details, see RDS Billing Description .

 Table 3-10 Resource planning and costs

Resource	Description	Cost
DLV	DLV adapts to a wide range of on- premise and cloud data sources, and provides diverse visualized components for you to quickly customize your data screens.	If you use the DLV service, you will be charged for the purchased yearly/monthly DLV package.

Example Data

• Order details wide table

Field	Data Type	Description
order_id	string	Order ID.
order_channel	string	Order channel (websites or apps)
order_time	string	Time
pay_amount	double	Order amount
real_pay	double	Actual amount paid
pay_time	string	Payment time
user_id	string	User ID
user_name	string	Username
area_id	string	Region ID

• Result table: real-time statistics of the total sales amount in each channel

Field	Data Type	Description
begin_time	varchar(32)	Start time for collecting statistics on metrics
channel_code	varchar(32)	Channel code
channel_name	varchar(32)	Channel
cur_gmv	double	Gross merchandises value (GMV) of the day

Field	Data Type	Description
cur_order_user_count	bigint	Number of users who settled the payment in the day
cur_order_count	bigint	Number of orders paid on the day
last_pay_time	varchar(32)	Latest settlement time
flink_current_time	varchar(32)	Flink data processing time

Step 1: Creating Resources

Create VPC, DMS, RDS, DLI, and DLV resources listed in Table 3-11.

Resource	Description	Instructions
VPC	A VPC manages network resources on the cloud.	Creating a VPC and Subnet
	The network planning is described as follows:	
	• The VPCs specified for the Kafka and MySQL instances must be the same.	
	• The VPC network segment where the Kafka and MySQL instances belong cannot conflict with that of the DLI queue.	
DMS Kafka	In this example, the DMS for Kafka instance is the data source.	Getting Started with DMS for Kafka
RDS MySQL	In this example, an RDS for MySQL instance provides the cloud database service.	Getting Started with RDS for MySQL
DLI	DLI provides real-time data analysis.	Creating a
	Create a general-purpose queue that uses dedicated resources in yearly/monthly or pay-per-use billing mode. Otherwise, an enhanced network connection cannot be created.	Queue
DLV	DLV displays the result data processed by the DLI queue in real time.	Creating Screens

Table 3-11 Cloud resources required

Step 2: Obtaining the DMS Connection Address and Creating a Topic

1. Hover the mouth on the **Service List** icon and choose **Distributed Message Service** in **Application**. The DMS console is displayed. On the **DMS for Kafka** page, locate the Kafka instance you have created.

Figure 3-24 Kafka instances

IS for Kafka ③							Buy Instance	Create Kal	ka Instance for Free
Quick start: Buy RabbitMQ Premium	Buy RocketMC	Instance							
Restart Change to Yearly	Monthly Billing	Renew	More +	Instance Creation	Failures				C @ []
									() Q
Name	Monit	Status	Version	Flavor	Used/Available Storage	Maximu	Billing Mode	Operation	
 kafkc 9e612875-c950-48c1-8b 	89	Running		kafka.2u4g.clu	0/11	750	Pay-per-Use Created on Nov 17,	Restart More 👻	

2. The instance details page is displayed. Obtain the **Instance Address (Private Network)** in the **Connection** pane.

Figure 3-25 Connection address

Connection		
Username		
Kafka SASL_SSL	Disabled	d Fixed for this instance
Instance Address (Private Network)		192.168.168.99:9092,192.168.168.249:9092,192.168.16 8.113:9092 ロ

3. Create a topic and name it **trade_order_detail_info**.

Figure 3-26 Creating a topic

Create Topic	
Topic Name	topic-
Partitions ⑦	- 1 + Value range: 1 to 100
Replicas	- 1 + Value range: 1 to 3
	Number of message copies.
Aging Time (h)	72 + Value range: 1 to 168 Time after which data in the topic expires.
Synchronous Replication ⑦	
Synchronous Flushing	

Configure the required topic parameters as follows:

- **Partitions**: Set it to **1**.
- **Replicas**: Set it to **1**.
- Aging Time: Set it to 72 hours.
- Synchronous Flushing: Disable this function.

Step 3: Creating an RDS Database Table

1. Log in to the console, hover your mouse over the service list icon and choose **Relational Database Service** in **Databases**. The RDS console is displayed. On the **Instances** page, locate the created DB instance and view its floating IP address.

×

Figure 3-27 Viewing the floating IP address



 Click More > Log In in the Operation column. On the displayed page, enter the username and password for logging in to the instance and click Test Connection. After Connection is successful is displayed, click Log In.

Figure 3-28 Logging in to an Instance

Instance Login Ir	formation
DB Instance Name	no_delete D8 Engine Version MySQL 5.7
* Login Username	root
* Password	Test Connection
	Remember Password Your password will be encrypted and stored securely.
Description	created by sync rds instance
Collect Metadata Periodically ⑦	(In or enabled, DAS can query the real-time structure information only from databases, which may affect the real-time performance of databases.
Show Executed SQL Statements ⑦	If not enabled, the executed SQL statements cannot be viewed, and you need to input each SQL statement manually.
	Log In Cancel

3. Click **Create Database**. In the displayed dialog box, enter database name **dlidemo**. Then, click **OK**.

Figure 3-29 Creating a database

Create Da	tabase	Х
* Name	dli-demo Only user databases can be created	
Character Set	utf8mb4	V
	OK Cancel	

 Choose SQL Operation > SQL Query and run the following SQL statement to create a MySQL table for test (Example Data describes the fields): DROP TABLE `dli-demo`.`trade channel collect`.

υ	KOP TABLE dil-defilo : trade_channet_collect ;
C	REATE TABLE `dli-demo`.`trade_channel_collect` (
	`begin_time` VARCHAR(32) NOT NULL,
	`channel_code` VARCHAR(32) NOT NULL,
	`channel_name` VARCHAR(32) NULL,
	`cur_gmv` DOUBLE UNSIGNED NULL,
	`cur_order_user_count` BIGINT UNSIGNED NULL,
	`cur_order_count` BIGINT UNSIGNED NULL,
	`last_pay_time` VARCHAR(32) NULL,
	`flink_current_time` VARCHAR(32) NULL.

)

```
PRIMARY KEY (`begin_time`, `channel_code`)
ENGINE = InnoDB
DEFAULT CHARACTER SET = utf8mb4
COLLATE = utf8mb4_general_ci
COMMENT = 'Real-time statistics on the total sales amount of each channel';
```

Figure 3-30 Creating a table

Data Admin Service MySQL	SQL Operations	Database Management	Import and Export	Structure Management	Data Scheme	Background Tasks
Home SQL Window X						
Current Database: dli-demo) 丨 🚦 Master S	vitch SQL Execution Node	Instance Name: no_del	ete 192.168	Character Set	utf8 V
Database: dii-demo V	Execute SO	QL (F8)	(F9) (ii) Execute SQ	L Plan (F6) SQL Favorites	v	
Tables Views		.E `dli-demo`.`trade_cha NBLE `dli-demo`.`trade cl	- /			
Please search by key ۹, C	4 °chai 5 °chai	in_time` VARCHAR(32) NOT nnel_code` VARCHAR(32) NU nnel_name` VARCHAR(32) NU _gmv` DOUBLE UNSIGNED NU	ΣT NULL, JLL,			
	8 `cur 9 `las	order_user_count' BIGIN order_count' BIGINT UNS: :_pay_time' VARCHAR(32) ik current time' VARCHAR	EGNED NULL, NULL,			
		Statements Messages				

Step 4: Creating an Enhanced Datasource Connection

 On the management console, hover the mouse on the service list icon and choose Analytics > Data Lake Insight. The DLI management console is displayed. Choose Resources > Queue Management to query the created DLI queue.

Figure 3-31 Queue list

ueue Management						0	Feedback		Buy Queue Buy DLI Package
Create SMN Topic						Search b	y name by	defaul	a c
Name	Type 🔽	Specificatio	Actual CUs JΞ	Elastic Scaling	Billing Mode	User	Ente	Det	Operation
∽ default	For SQL	-		Max: CUs Min: CUs	By SQL computations Created on Mar 21, 2019 19:36:38 G	DLI		Sys	
~	For general pur	16 CUs	16 CUs	Max: 48 CUs Min: 16 CUs	Pay-per-use Created on Sep 07, 2022 10:19:33 G	ei_dli	default		Delete Permissions More 💌

 In the navigation pane of the DLI management console, choose Global Configuration > Service Authorization. On the displayed page, select VPC Administrator, and click Update to grant the DLI user the permission to access VPC resources. The permission is used to create a VPC peering connection.

Figure 3-32 Updating agency permissions

 VICE	Autorization
Assi	ign Agency Permissions
A	DLI agency permissions are stictly restricted. Permissions can be customized based on different service scenarios. You are advised to assign agency permissions.
	Select all
	Tomant Administrator(Global service) Trend Administrator permission a required to access date from 666 to exercise Trink jobs on 501, for example, obtaining 666/0V/d date sources, log dump (including budiet authorization), checkpairl enabling, and job import and export. Due to could invite could memory could enable access a doubt for models to take effect.
	DIS Administrator DIS Administrator permissions are required to use DIS data as the data source of DU Flink jobs. Due to cloud service cache differences, operations require about 50 minutes to take effect.
	CloudTable Administrator To use CloudTable data as the data source of GU Plank jobs, CloudTable Administrator permissions are required.
	VPC Administrator VPC Administrator permissions are required to use the VPC, subnet, route, VPC peering connection, and port for DLI datasource connections.
	SMN Administrator To receive natioations when a CL (x0 Table, DMN Administrator permissions are required.
	Tenant Administrator(Project-level) Tenant Administrator(Project-level) permission is needed if you use services that can run only with this permission. Due to service cache differences, operations require about 2 minutes to take effect.
	IAM ReadOnlyAccess Mult ReadOnlyAccess permissions are required to obtain information about the Mul users who have never logged in to DLI.
Once r	service authorization has succeeded, an agency named dt_admin_agency on IAAI will be created. Go to the agency isi to view the details.
Note	36
	y the lensert account or sub-accounts under Liter Group admin can perform authorization.
2. Do r	not debte the created agency dLadmm_agency.

- 3. Choose **Datasource Connections**. On the displayed **Enhanced** tab, click **Create**. Configure the following parameters, and click **OK**.
 - **Connection Name**: Enter a name.
 - **Resource Pool**: Select the general-purpose queue you have created.
 - **VPC**: Select the VPC where the Kafka and MySQL instances are.
 - **Subnet**: Select the subnet where the Kafka and MySQL instances are.

Figure 3-33 Creating an enhanced datasource

After you create the enha connection and required	anced datasource connection, the system will automatically create a VPC peering routes.
* Connection Name	peer_
Resource Pool	•
* VPC	vpc-default(
* Subnet	subnet-default •
Host Information	Enter host information in the format "host IP address host name". Specify the information for each host on a separate line.
Tags	It is recommended that you use TMS's predefined tag function to add the same to different cloud resources. View predefined tags C
	To add a tag, enter a tag key and a tag value below.

The status of the created datasource connection is **Active** in the **Enhanced** tab.

Click the name of the datasource connection. On the details page, the connection status is **ACTIVE**.

Figure 3-34 Connection status

Datasou	irce C	Conr	ections		
Enha	nced	10	Basic	Datasource A	Authentication
Cr	eate	ть	e enhanced dat	tasource connectio	on supports queues created in
	Con	necti	on Name		Connection Status
~					 Active

Figure 3-35 Details

Create				Enter a name. Q
VPC Peering ID	Resource Pool	Connection Status	Updated	Operation
✓ 5b6a1c43-ab62-42b7-a3ce-161139a6feee	auto_erp	Active	Oct 13, 2022 16:48:11 GMT+08:00	Unbind Resource Pool

- 4. Test whether the queue can connect to RDS for MySQL and DMS for Kafka instances, respectively.
 - a. On the **Queue Management** page, locate the target queue. In the **Operation** column, click **More** > **Test Address Connectivity**.

Figure 3-36 Testing address connectivity

e N	anagen	nent						Feed	back	Buy Queue Buy DLI Pack
Cre	ate SMN To	ppic					Se	arch by nan	ne by defaul	L Q
	Name	Type 🕜	Specificatio	Actual CUs JΞ	Elastic Scaling	Billing Mode	User	Ente	Desc	Operation
~	default	For SQL		-	Max: CUs Min: CUs	By SQL computations Created on Mar 21, 2019 19:36:38 G	DLI		Syst	
~		For general pur	16 CUs	16 CUs	Max: 48 CUs Min: 16 CUs	Pay-per-use Created on Sep 07, 2022 10:19:33 G	ei_dli	default	-	Delete Permissions More 🔺
~		For general pur	16 CUs	16 CUs	Max: CUs Min: CUs	Pay-per-use Created on Jul 03, 2021 11:25:58 G	ei_dli	default		Modify Enterprise Project Elastic Scaling Schedule CU Changes
~		For SQL	16 CUs	32 CUs	Max: 32 CUs Min: 16 CUs		ei_dli	(Test Address Connectivity Change to Yearly/Monthly
~		For general pur	16 CU8	16 CUs	Max: CUs Min: CUs		ei_dli		-	Tags

b. Enter the connection address of the DMS for Kafka instance and the private IP address of the RDS for MySQL instance to test the connectivity.

If the test is successful, the DLI queue can connect to the Kafka and MySQL instances.

Figure 3-37 Testing address connectivity

Test Address Connectivity					
	an address is reachable from a specified cluster. The address can be a an IP address, or a specified port.				
* Address	192.168.0.233:3306				
	Test Cancel				

If the test fails, modify the security group rules of the VPC where the Kafka and MySQL instances are to allow DLI queue access on ports 9092 and 3306. You can obtain the network segment of the queue on its details page.

Figure 3-38 VPC security group rules



Step 5: Creating and Submitting a Flink Job

- 1. In the navigation pane on the left, choose **Job Management** > **Flink Jobs**. Click **Create Job**.
 - Type: Select Flink OpenSource SQL.
 - Name: Enter a name.

Figure 3-39 Creating a Flink Job

Create Job	
Туре	Flink OpenSource SQL 🔹
* Name	kafka-2-msql
Description	Description
Template Name	Select
Tags	It is recommended that you use TMS's predefined tag function to add the same tag to different cloud resources. View predefined tags $\ {f C}$
	To add a tag, enter a tag key and a tag value below.
	Enter a tag key Add
	20 tags available for addition.
	OK Cancel

Click **OK**. The job editing page is displayed. The following is a simple SQL statement. You need to modify some parameter values based on the RDS and DMS instance information.

```
-- Data source: trade_order_detail_info (order details wide table)
create table trade_order_detail (
 order_id string, -- Order ID
 order_channel string, -- Channel
order_time string, -- Order creation time
pay_amount double, -- Order amount
real_pay double, -- Actual payment amount
pay_time string, -- Payment time
user_id string, -- User ID
user_name string, -- Username
area_id string -- Region ID
) with (
 "connector.type" = "kafka",
 "connector.version" = "0.10",
 "connector.properties.bootstrap.servers" = "xxxx:9092,xxxx:9092,xxxx:9092", -- Kafka connection
address
 "connector.properties.group.id" = "trade_order", -- Kafka groupID
 "connector.topic" = "trade_order_detail_info", -- Kafka topic
 "format.type" = "json",
 "connector.startup-mode" = "latest-offset"
);
*************************
-- Result table: trade_channel_collect (real-time statistics on the total sales amount of each channel)
```

```
create table trade_channel_collect(
 begin_time string, --Start time of statistics collection
                      -- Channel ID
 channel_code string,
 channel_name string, -- Channel name
                      -- GMV
cur amv double.
 cur_order_user_count bigint, -- Number of payers
cur_order_count bigint, -- Number of orders paid on the day last_pay_time string, -- Latest settlement time
 flink_current_time string,
 primary key (begin_time, channel_code) not enforced
with (
 "connector.type" = "jdbc",
 "connector.url" = "jdbc:mysql://xxxx:3306/xxxx", -- MySQL connection address, in JDBC format
 "connector.table" = "xxxx",
                                 -- MySQL table name
 "connector.driver" = "com.mysql.jdbc.Driver",
 'pwd_auth_name'= 'xxxxx', -- Name of the datasource authentication of the password type created
on DLI. If datasource authentication is used, you do not need to set the username and password for
the job.
 "connector.write.flush.max-rows" = "1000",
 "connector.write.flush.interval" = "1s"
);
-- Temporary intermediate table
                             create view tmp order detail
as
select *
  , case when t.order_channel not in ("webShop", "appShop", "miniAppShop") then "other"
       else t.order_channel end as channel_code -- Redefine channels. Only four enumeration values
are available: webShop, appShop, miniAppShop, and other.
  , case when t.order_channel = "webShop" then _UTF16"Website"
       when t.order_channel = "appShop" then _UTF16"Shopping App"
       when t.order_channel = "miniAppShop" then _UTF16" Miniprogram"
       else _UTF16"Other" end as channel_name -- Channel name
from (
  select *
     , row_number() over(partition by order_id order by order_time desc ) as rn -- Remove duplicate
order data
     , concat(substr("2021-03-25 12:03:00", 1, 10), " 00:00:00") as begin_time
     , concat(substr("2021-03-25 12:03:00", 1, 10), " 23:59:59") as end_time
  from trade_order_detail
  where pay_time >= concat(substr("2021-03-25 12:03:00", 1, 10), " 00:00:00") --Obtain the data of
the current day. To accelerate running, 2021-03-25 12:03:00 is used to replace
cast(LOCALTIMESTAMP as string).
  and real_pay is not null
) t
where t.rn = 1;
-- Collect data statistics by channel.
insert into trade_channel_collect
select
   begin_time --Start time of statistics collection
  , channel_code
  , channel_name
  , cast(COALESCE(sum(real_pay), 0) as double) as cur_gmv -- GMV
  , count(distinct user_id) as cur_order_user_count -- Number of payers
  , count(1) as cur_order_count -- Number of orders paid on the day
  , max(pay_time) as last_pay_time -- Settlement time
  , cast(LOCALTIMESTAMP as string) as flink_current_time -- Current time of the flink task
from tmp_order_detail
where pay_time >= concat(substr("2021-03-25 12:03:00", 1, 10), " 00:00:00")
```

where pay_time >= concat(substr("2021-03-25 12:03:00", 1, 10), " 00:0 group by begin_time, channel_code, channel_name;

D NOTE

Job logic

- 1. Create a Kafka source table to read consumption data from a specified Kafka topic.
- 2. Create a result table to write result data into MySQL through JDBC.
- 3. Implement the processing logic to collect statistics on each metric.

To simplify the final processing logic, create a view to preprocess the data.

- Use over window condition and filters to remove duplicate data (the top N method is used). In addition, the built-in functions concat and substr are used to set 00:00:00 as the start time and 23:59:59 of the same day as the end time, and to collect statistics on orders paid later than 00:00:00 on the day. (To facilitate data simulation, replace cast(LOCALTIMESTAMP as string) with 2021-03-25 12:03:00.)
- Based on the channels of the order data, the built-in condition function is used to set channel_code and channel_name to obtain the field information in the source table and the values of begin_time, end_time, channel_code, and channel_name.
- 4. Collect statistics on the required metrics, filter the data as required, and write the results to the result table.
- 3. Select the created general-purpose queue and submit the job.

· · ·	
k Queue	•
★ Flink Version	1.12
UDF Jar	Select
★ CUs	- 2 + ?
★ Job Manager CUs	- 1 +
★ Parallelism	- 1 + ?
Task Manager Configu	
★ OBS Bucket	
Save Job Log	
Alarm Generation upo	
Enable Checkpointing	
Auto Restart upon Exc	
Idle State Retention Time	e — 1 + h ▼

Figure 3-40 Flink OpenSource SQL Job

4. Wait until the job status changes to **Running**. Click the job name to view the details.

Figure 3-41 Job status

[Running] ID: 104542 Job Type						
Job Detail Task List	Execution Plan	Commit Logs	Run Log	Tags		
Name	Duration	Max C Task		Status	Back P	Delay
	15d 14	1 000	100000	Runni	ОК	
	15d 14	1 000	100000	Runni	OK	
	. 15d 14	1 000	100000	Runni	ОК	53

5. Use the Kafka client to send data to a specified topic to simulate real-time data streaming.

For details, see Connecting to an Instance Without SASL.

Figure 3-42 Real-time data streaming

[dli@kafka-client bin]\$./kafka-console-producer.shbroker-list 192.168.0.3:9092,192.168.0.147:9092,192.168.0.192:9092topi
c trade_order_detail_info
>("order_id":"202103241000000001", "order_channel":"webShop", "order_time":"2021-03-24 10:00:00", "pay_amount":"100.00", "real_p
ay":"100.00", "pay_time":"2021-03-24 10:02:03", "user_id":"0001", "user_name":"Alice", "area_id":"330106"}
>{"order_id":"202103241606060001", "order_channel":"appShop", "order_time":"2021-03-24 16:06:06", "pay_amount":"200.00", "real_p
au":"188.00", "pau time":"2021-03-24 16:10:06", "user id":"9001", "user name":"Alice", "area id":"330106")
>("order id":"202103251202020001", "order channel":"miniAppShop", "order time":"2021-03-25 12:02:02", "pay amount":"60.00", "rea
l_pay":"60.00", "pay_time":"2021-03-25 12:03:00", "user_id":"0002", "user_name":"Bob", "area_id":"330110"}
>("order_id":"20210325150505050001", "order_channel":"qgShop", "order_time":"2021-03-25 15:05:05", "pay_amount":"500.00", "real_pa
u":"400.00", "pau time":"2021-03-25 15:10:00", "user id":"0003", "user name":"Cindu", "area id":"300100")
>("order_id:"282103250200200001", "order_chame!:"webShop", "order_time":"2021-03-24 20:201, "pay_amount":"600.00", "real_p
au":"488.66", "pau time":"2821-63-25 96:06":60", "user id":"9694", "user name":"Daisu", "area id":"39192")
[3] 'do'do', 'pag-init' of the object of a start of the object of the
au":"240.80", "bay time":"2821-83-25 08:10:80", "user id":"0804", "user name":"Daisu", "area id":"330102")
mg : 210:07, pag_int : 220105 12 05105 02, 0510 00 , 0501 00 7, 0501 pag. bats of the start of
au": "180.88". "pau time": "2821-63-25 16:16:6", "user id": "8884". "user name": "Daisu", "area id": "388182")
ay . 100.00 , pay_ime . 2021 mo . 2011 no 10 , user_ia . 0007 , user_iame . valsy , area_ia . 300102 , X("order id": "202103270606609001", "order channel": "auxNou", "order time: "2021-03-25 06:06:06", "pau amount": "59.50", "real pa
<pre>vurder_id : 202103270000000001 ; Urder_challel : appshop , Urder_chme : 202103220000000000 ; pag_amount : 50.50 ; real_pa u":"50 :50", "pag time:":"202103270000000001 ; Urder_challer : appshop , Urder_challer : "3100"; "330106")</pre>
""""""""""""""""""""""""""""""""""""""
y":"66.60", "pay_time":"2821-03-25 06:07:00", "user_id":"0002", "user_name":"Bob", "area_id":"330110")
""""""""""""""""""""""""""""""""""""""
l_pay":"88.88", "pay_time":"2021-03-25 06:07:00", "user_id":"0003", "user_name":"Cindy", "area_id":"330108")
>{"order_id":"28218327868686868804", "order_channel":"webShop", "order_time":"2821-83-25 86:86:86", "pay_amount":"99.98", "real_pa
y":"99.98", "pay_time":"2821-83-25 06:87:08", "user_id":"0004", "user_name":"Daisy", "area_id":"330102")

6. Run the following command:

sh kafka_2.11-2.3.0/bin/kafka-console-producer.sh --broker-list *KafKa connection address --*topic *Topic name*

Example data is as follows:

{"order_id":"20210324100000001", "order_channel":"webShop", "order_time":"2021-03-24 10:00:00", "pay_amount":"100.00", "real_pay":"100.00", "pay_time":"2021-03-24 10:02:03", "user_id":"0001", "user_name":"Alice", "area_id":"330106"} {"order_id":"202103241606060001", "order_channel":"appShop", "order_time":"2021-03-24 16:06:06",

[order_id : 202103241000000001 , order_channet : appshop , order_unite : 2021-03-24 10:00:00 , "pay_amount":"200.00", "real_pay":"180.00", "pay_time":"2021-03-24 16:10:06", "user_id":"0001", "user_name":"Alice", "area_id":"330106"}

{"order_id":"202103251202020001", "order_channel":"miniAppShop", "order_time":"2021-03-25 12:02:02", "pay_amount":"60.00", "real_pay":"60.00", "pay_time":"2021-03-25 12:02:02", "base and "base and

"user_id":"0002", "user_name":"Bob", "area_id":"330110"} {"order_id":"202103251505050001", "order_channel":"qGhop", "order_time":"2021-03-25 15:05:05", "pay_amount":"500.00", "real_pay":"400.00", "pay_time":"2021-03-25 15:10:00", "user_id":"0003",

"user_name":"Cindy", "area_id":"330108"}

{"order_id":"202103252020200001", "order_channel":"webShop", "order_time":"2021-03-24 20:20:20", "pay_amount":"600.00", "real_pay":"480.00", "pay_time":"2021-03-25 00:00:00", "user_id":"0004", "user_name":"Daisy", "area_id":"330102"}

{"order_id":"202103260808080001", "order_channel":"webShop", "order_time":"2021-03-25 08:08:08", "pay_amount":"300.00", "real_pay":"240.00", "pay_time":"2021-03-25 08:10:00", "user_id":"0004", "user_name":"Daisy", "area_id":"330102"}

{"order_id":"202103261313130001", "order_channel":"webShop", "order_time":"2021-03-25 13:13:13", "pay_amount":"100.00", "real_pay":"100.00", "pay_time":"2021-03-25 16:16:16", "user_id":"0004", "user_name":"Daisy", "area_id":"330102"}

{"order_id":"202103270606060001", "order_channel":"appShop", "order_time":"2021-03-25 06:06:06", "pay_amount":"50.50", "real_pay":"50.50", "pay_time":"2021-03-25 06:07:00", "user_id":"0001", "user_name":"Alice", "area_id":"330106"}

{"order_id":"202103270606060002", "order_channel":"webShop", "order_time":"2021-03-25 06:06:06", "pay_amount":"66.60", "real_pay":"66.60", "pay_time":"2021-03-25 06:07:00", "user_id":"0002", "user_name":"Bob", "area_id":"330110"}

{"order_id":"202103270606060003", "order_channel":"miniAppShop", "order_time":"2021-03-25 06:06:06", "pay_amount":"88.80", "real_pay":"88.80", "pay_time":"2021-03-25 06:07:00", "user_id":"0003", "user_name":"Cindy", "area_id":"330108"}

{"order_id":"2021032706060600004", "order_channel":"webShop", "order_time":"2021-03-25 06:06:06", "pay_amount":"99.90", "real_pay":"99.90", "pay_time":"2021-03-25 06:07:00", "user_id":"0004", "user_name":"Daisy", "area_id":"330102"}

 In the navigation pane on the left, choose Job Management > Flink Jobs, and click the job submitted in 3. On the job details page, view the number of processed data records.

Figure 3-43 Job details

ID: 10454	[Running] 2 Job Typ	be:							C Job Monit	oring Edit	Start	More 🔻
Job Det	ail Tas	ik List	Execution Plan C	Commit Logs	Run	Log	Tags					
												С
	Duration	Max C	Task	Status	Back P	Delay	Sent Records	Sent Bytes	Received Records	Received Bytes	Started	Ended
P	15d 14	1	000100000	Runni	OK)						Nov 15, 202	
p	15d 14	1	000100000	Runni	OK)						Nov 15, 202	
dows	15d 14	1	000100000	Runni	OK	53					Nov 15, 202	
4												

Step 6: Querying the Result

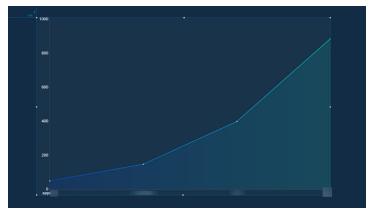
 Log in to the MySQL instance by referring to 2 and run the following SQL statement to query the result data processed by the Flink job: SELECT * FROM `dli-demo`.`trade_channel_collect`;

Figure 3-44 Querying results



 Log in to the DLV console, configure a DLV screen, and run SQL statements to query data in the RDS for MySQL instance to display the data on the screen.
 For details, see Editing Screens.

Figure 3-45 DLV screen



3.6 Interconnecting Yonghong BI with DLI to Submit Spark Jobs

3.6.1 Preparing for Yonghong BI Interconnection

Scenario

Prepare for the interconnection between Yonghong BI system and DLI.

Procedure

- Step 1 (Optional) In the upper left corner of the Huawei Cloud management console, click Service List and choose Analytics > Data Lake Insight. On the Overview page displayed, find the Common Links area on the right, and click SDK Download. On the DLI SDK DOWNLOAD page displayed, download a DLI JDBC driver, for example, dli-jdbc-1.1.0-jar-with-dependencies-jdk1.7.jar. For details, see Downloading the JDBC Driver Package.
- **Step 2** The AK/SK and token authentication modes can be used for JDBC authentication. The AK/SK authentication mode is recommended.
- **Step 3** Contact Yonghong customer service personnel to obtain the username and password of the Yonghong SaaS production environment.
- **Step 4** Log in to the Yonghong SaaS production environment and enter the username and password.

----End

3.6.2 Adding Yonghong BI Data Source

Scenario

Add the DLI data source to the Yonghong SaaS production environment.

Procedure

Step 1 On the homepage of the Yonghong SaaS production environment, click **Create Connection** from the left navigation tree. See **Figure 3-46**.

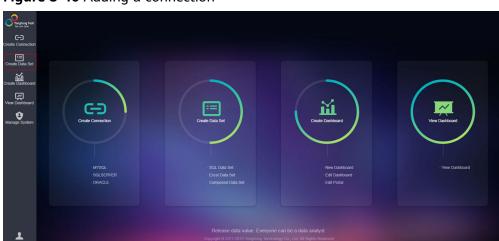


Figure 3-46 Adding a connection

Step 2 On the **New Connection Type** page, choose **GENERIC** for the type of the new connection. See **Figure 3-47**.

	5	5	51
0			
œ	11011503 GO public_db	New Connection Type	
=		DATA MART	POSTGRESQL
		DB2	PRESTO
影		DERBY	SPARK
(A)		GENERIC	SQLSERVER
9		HANA	SYBASE
		HBASE	VERTICA
		HIVE	
		IMPALA	
		INFORMIX	
		KINGBASE	
		MONGO	
		MYSQL	
		ORACLE	
2			

Figure 3-47 Choosing a new connection type

Step 3 Configure the new connection. See Figure 3-48.

In Driver, enter com.huawei.dli.jdbc.DliDriver.

In **URL**, select **Self-defined Protocol**. Enter the URL of the DLI JDBC driver. For details about the URL format and the attributes, see **Table 3-12** and **Table 3-13**, respectively.

NOTE

- In **Schema**, you can optionally enter the name of the database to be accessed. If you enter the name, only tables in the database are displayed during data set creation. If you do not enter the name, tables in all databases are displayed during data set creation. For details about how to create a data set, see **Creating Yonghong BI Data Set**.
- Retain default values of other parameters. You do not need to select **Request Login**.

0	Q Enter a keyw ord to search.	New Connection Save Save As Test Connection Edit Parameters Refresh Parameter Cancel Close
G	74022130	Detabase: GBNERIC v
	▶ GD public_db	☐ <u>Visible</u> to users have write permission
::		Driver: comhuswei.dlijdb: DiDriver
叉		URL jibb: dii./dii.cn-north-1.myhuase eic.loud.com/8fc20d9784444.afba3c3a86393800
		Default DB:
[[]		Request Login: 🕑 Max Connection: 10
		User. Passw ord:
9		DB <u>Q</u> haracter Sets:
		Please refer to database administrator for URL and user / passw ord.

Figure 3-48 Configuring the new connection

Parameter	Description
URL	The URL format is as follows:
	jdbc:dli:// <endpoint>/<projectid>? <key1>=<val1>;<key2>=<val2></val2></key2></val1></key1></projectid></endpoint>
	NOTE
	 endpoint indicates the domain name of DLI. For details, see Regions and Endpoints.
	 projectId indicates the project ID, which can be obtained from the My Credentials page of the public cloud platform.
	• The question mark (?) is followed by other configuration items. Each configuration item is listed in the "key=value" format. Semicolons (;) are used to separate configuration items. For details, see Table 3-13.

 Table 3-12 Database connection parameters

Attribute (key)	Mandatory	Defaul t Value (value)	Description
queuename	Yes	-	Queue name of DLI.
databasename	No	-	Default database to be accessed. If this parameter is not specified in the URL, you need to use db.table (for example, select * from dbother.tabletest) to access tables in the database.
authentication mode	Yes	token	Authentication method, which can be token or aksk . Value aksk is recommended during the interconnection with Yonghong BI system.
accesskey	This parameter must be configured if authentication mode is set to aksk .	-	For details, see Preparing for Yonghong BI Interconnection .

Attribute (key)	Mandatory	Defaul t Value (value)	Description
secretkey	This parameter must be configured if authentication mode is set to aksk .	-	For details, see Preparing for Yonghong BI Interconnection .
regionname	This parameter must be configured if authentication mode is set to aksk .	-	For details, see Regions and Endpoints.
servicename	This parameter must be configured if authentication mode is set to aksk .	-	servicename=dli
dli.sql.checkNo ResultQuery	No	false	 Whether to allow invoking the executeQuery API to execute statements (for example, DDL) that do not return results. Value false indicates that
			invoking of the executeQuery API is allowed.
			 Value true indicates that invoking of the executeQuery API is not allowed.
			NOTE If dli.sql.checkNoResultQuery is set to false, non-query statements will be executed twice.

Step 4 On the tool bar of the displayed page, click **Test Connection**. After the test is complete, click **Save**. Enter the data source name, and save the data source.

NOTE

Currently, you are not allowed to save the data source to the root directory. Therefore, you can only save the data source to an existing folder.

----End

3.6.3 Creating Yonghong BI Data Set

Scenario

Create a DLI data set in the Yonghong SaaS production environment.

Procedure

Step 1 On the home page of the Yonghong SaaS production environment, click Create Data Set in the left navigation tree. See Figure 3-49.



Figure 3-49 Creating a data set





Step 3 On the displayed page, select the added DLI data source from the **Connection** drop-down list box. See **Figure 3-51**.

Figure 3-50 Creating a SQL data set

	5								
Q Enter a keyw ord to search.	New Data Set 🔻	Save Save As Test Connection Edit Parameters Refresh Parameter Cano	cel Close						
Connection	Performance Te	st	Dat	ta	Row Filter				
74022130	Connection:	Connection/74022130/dll_test 🔍	===	=				🗌 Aļi Data	Sample Row
Ξ	Database:	GBNERC v	Name		Alias	Data Type	Format	Visibility	Column Filter
	Drjver:	comhuaw ei.dli.jdbc.DliDriver	tin Din						
题	URL:	y jdbc:dlt//dll.cn-north-1.myhuaw.elcloud.com/8fc20d97a4444cafba3c3a8639380003	tin Me	asure					
風	Default DB:	Scheme:							
	Request Login:	Max Connection: 10							
•	Us <u>e</u> r:	Passw ord:							
	DB Character Sets:								
		lease refer to database administrator for URL and user / passw ord.							
	Tables:]						
		Right click to refresh							
	SQL Statement:		-						
•									
		Merge SQL Refresh Metadata(B)							

Figure 3-51 Selecting a data source

Step 4 In the Table area on the left pane, right-click and choose Update to update tables. All databases and their tables are listed in the area. Figure 3-52 shows the page displayed when Table Structure is not configured during connection creation.

Figure 3-52 Updating tables

Connection	Performance	Test		Data	Row Filter			
74022130	Connection:	Connection/74022130/dli_test 🔹						All Data Sample Rows:
	Database:	GENER I C		Name	Alias	Data Type Format	Visibi	. Column Filter
	Drjver:	com. huavei. dli. jdbc. DliDriver		Dimension				
	URL:	y jdbc:dli://dli.ontmorth=1.myhuaweicloud.com/8fc20d97a4444cafba3c3a863	7380003?qu	🛅 Measure				
	De <u>f</u> ault D0:	Sohena:						
	Request Login:	Max Connection: 10						
	Us <u>e</u> r:	Password:						
	S Character Sets:							
		Please refer to database administrator for URL and user ${\sc /}$ password.						
		▶ ▲ db_xd ▶ ▲ default						
	<u>S</u> OL Statement:	 > ▲ discriming_states > ▲ color_db > ▲ testdb 						

Step 5 In the **SQL Statement** area on the left pane, enter the **select * from table_name** command to query tables. On the **Preview Data** page on the right pane, click

. Metadata of the table, including fields and field types, is displayed. See **Figure 3-53**.

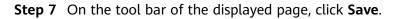
늘 Connection	Performance	Test				Data	Row Filter				
74022130	Connection:	Connection/74022130/dli_test		Ŧ	*						A <u>l</u> i Data <u>S</u> emple Rows:
	Qatabase:	GENERIC				Namo	Alias	Data Type	Format	Visibi	Column Filter
	Dr <u>i</u> ver:	com. huamei. dli. jdbo. DliDriver				🔻 🛅 Dimensio	•				
	URL:	dbc:dli://dli.onmort	th=1. myhuaweicloud	com/8fc20d97a4444caf	a8c3a8639380003?qu	Abc name		String		0	
	De <u>f</u> ault DB:		Schena :			# age		Integer		•	
	Reques <u>t</u> Login:		Max Connection:	10		# number		Integer		0	
	Usgr:		Password:								
	3 <u>C</u> haracter Sets:										
		Please refer to database administrator (for URL and user /	password.							
	Tables:	***									
		Connection/74022130/dli_test									
		v III Tables									
		III table_child									
		tableallot									
		III testbi									
		III aasthia	_								
	<u>S</u> QL Statement:	SELECT * from table_child									
		Merge SOL		Re	resh Metadata(B)						

Figure 3-53 Querying the table

Step 6 Click and the right pane to query data details. See **Figure 3-54**.

Figure 3-54 Querying data of the table

Connection	Performance	Test	Data	Row Filter		
74022130	Connection:	Connection/74022130/dli_test		Show Row Cou	nt (0)	Max Rows for Preview 100
	Database:	GENERIC			# number	Abc name
	Dr <u>i</u> ver:	com. husvel. dll. jdbo. DliDriver	1		0	zhangsan
	URL:	jdbc:dli://dli.cn=north=1.myhuaweicloud.com/8fc20d97a4444cafba3c3a8639380003	2		2	lisi
	Default DB:		3		4	zheosu
			4		6	xisoning
	Reques <u>t</u> Login:		5		8	xiaoqi xiaoba
	Usgr:	Bassmord:			10	A reveal
	<u>C</u> haracter Sets:					
		Please refer to database administrator for URL and user / password.				
			5			
	Tables:	Connection/74022130/dli_test V 👗 testdo				
		v III Tables				
		table_child				
		tableallet				
		III testbi				
		III testoi				
	SOL Statement:	SELECT * from table_child				
		-				
		Megge SOL Refresh Metadata(B)				



----End

3.6.4 Creating a Chart in Yonghong BI

Scenario

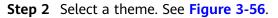
Create a chart in the Yonghong SaaS production environment.

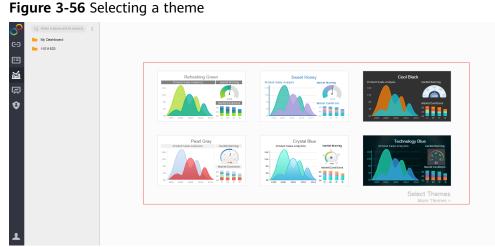
Procedure

Step 1 On the home page of the Yonghong SaaS production environment, click **Create** Dashboard in the left navigation tree. See Figure 3-55.



Figure 3-55 Creating a dashboard





- **Step 3** In this example, the Refreshing Green theme is selected. On the left pane, select
 - the created data set from the drop-down list box and choose a table as the data source (for example, table_child). Metadata (including fields and field types) of the table is displayed in the lower part of the **Data** column. See Figure 3-57.

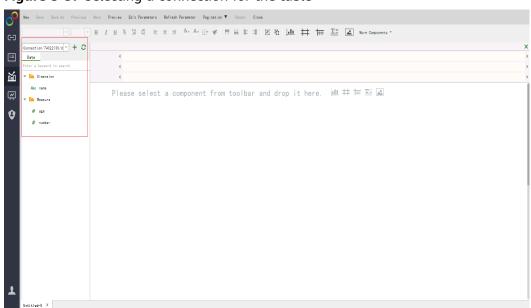
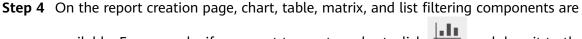
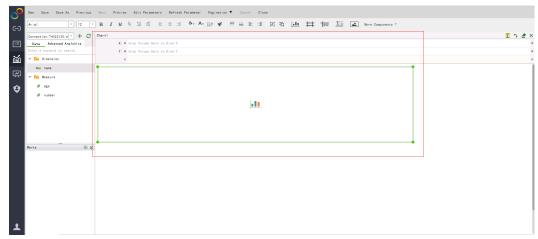


Figure 3-57 Selecting a connection for the table



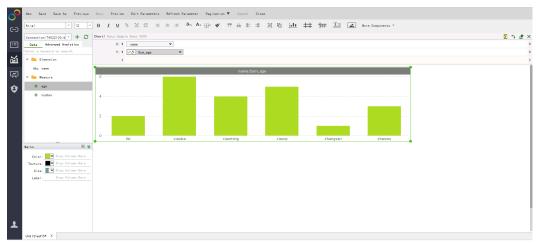
available. For example, if you want to create a chart, click and drag it to the editing area. See **Figure 3-58**.

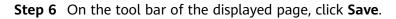
Figure 3-58 Creating a chart



Step 5 In **X**, choose **name**. In **Y**, choose **age**. Drag them to the corresponding area, and the system automatically generates a bar chart. See **Figure 3-59**.

Figure 3-59 Generating a chart





----End

3.7 Interconnecting FineBI with DLI Trino

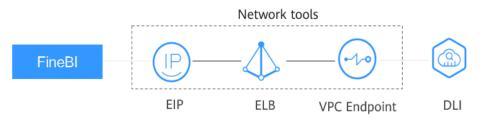
FineBI is a BI tool for big data analytics developed by FanRuan Software. It provides business personnel and data analysts with data exploration capabilities such as data management, editing, and visualization. Huawei Cloud DLI integrates data analysis and processing. SQL jobs using the Trino interactive engine are more suitable for interactive analysis and query. It provides FineBI with efficient engine compute capabilities and effective high-quality data for subsequent data statistics and analysis, helping enterprises make data decisions.

This section describes how to interconnect FineBI with DLI.

Solution Overview

This solution uses VPCEP to connect FinBI to DLI.

Figure 3-60 Architecture



Constraints

- Trino engine queues support only HTTPS connections.
- When the Trino engine is used, the created SQL queues cannot be scaled in or out.

To adjust the CU size of a queue, you need to first **delete** the queue in the elastic resource pool and then **create** a queue with Trino as the engine and with an appropriate CU size in the elastic resource pool.

• The DLI Trino engine is in the open beta test (OBT) phase. If you need it, contact customer service to apply for it.

The DLI Trino engine is available in the following regions: **CN North-Beijing4**, **CN East-Shanghai1**, **CN-Hong Kong**, **AP-Bangkok**, **AP-Singapore**, and **AF-Johannesburg**.

• Currently, only foreign tables created using the Hive syntax can be used for FineBI interconnection.

Process

Figure 3-61 Process of interconnecting DLI with FlineBI



To interconnect FineBI with Huawei Cloud DLI, perform the following steps:

- 1. Creating an Elastic Resource Pool and Queue
- 2. Configuring Network Connectivity for the DLI Cluster
- 3. Installing the Trino Driver for FineBI
- 4. Interconnecting FineBI with DLI Trino
- 5. Testing the Connection

Solution Advantages

- As a next-gen BI tool for self-service big data analytics, FineBI provides enterprises with one-stop solutions for enterprise business intelligence, such as multi-source data collection, self-service exploratory analysis, multi-screen support, and enterprise-level management and control.
- Huawei Cloud DLI provides convergent data analysis and processing capabilities. DLI can interconnect with multiple data sources and map data sources by creating tables using SQL statements. You can use standard SQL statements to compile metric analysis logic without paying attention to the complex distributed computing platform.
- FineBI interconnects with Huawei Cloud DLI for real-time data ingestion, efficient data processing, and good data visualization. DLI can connect to FineBI from multiple data sources. Fine BI can display DLI data in charts and reports, making data more intuitive and improving decision-making accuracy and efficiency.

Resource Planning and Costs

_		
Resource	Description	Cost
OBS	DLI needs to use OBS buckets to store logs.	You will be charged for using the following OBS resources:
		• Storage Fee for storing static website files in OBS.
		 Request Fee for accessing static website files stored in OBS.
		 Traffic Fee for using a custom domain name to access OBS over the public network.
		The actual fee depends on the size of the stored file, the number of user access requests, and the traffic volume. Estimate the fee based on your service requirements.
DLI	In this example, an elastic resource pool is used to create SQL jobs.	When using a DLI elastic resource pool, you are billed based on the CUH of the elastic resource pool.
VPCEP	Used to enable the network connection between FineBI and DLI.	For details about the billing for VPCEP, see Billing .
ELB	Elastic Load Balance (ELB) distributs access traffic to multiple backend servers based on distribution policies.	For details about the billing for ELB, see Billing .
EIP	Provides independent public IP addresses and bandwidth for Internet access.	For details about the billing for EIP, see Billing .

-

Step 1: Creating an Elastic Resource Pool and Queue

- **Step 1** Log in to the DLI management console.
- **Step 2** In the navigation pane on the left, choose **Resources** > **Resource Pool**.
- **Step 3** On the **Resource Pool** page, click **Buy Resource Pool** in the upper right corner.
- **Step 4** On the displayed page, set the following parameters:

Parameter	Description
Billing Mode	Pay-per-use/Yearly/Monthly
Region	Select a region near you to ensure the lowest latency possible.
Project	Each region corresponds to a project.
Name	Name of the elastic resource pool.
CU Range	The maximum and minimum CUs allowed for the elastic resource pool.
Description	Description of the elastic resource pool.
CIDR Block	CIDR block the elastic resource pool belongs. If DLI enhanced datasource connections are required, the CIDR block of the elastic resource pool cannot overlap with that of the data source. The CIDR block of the elastic resource pool cannot be changed after being set. Recommended CIDR blocks: 10.0.0~10.255.0.0/16~19 172.16.0.0~172.31.0.0/16~19 192.168.0.0~192.168.0.0/16~19
Enterprise Project	If the created elastic resource pool belongs to an enterprise project, select the enterprise project.
Required Duration	You must specify the Required Duration if Billing Mode is set to Yearly/Monthly . The longer the subscription duration is, the more discounts you can get. If Auto renew is selected, monthly subscriptions are renewed each month. And yearly subscriptions are renewed each year.
Tag	Tags used to identify cloud resources.

Table 3-15 Parameters

- **Step 5** Click **Buy** and confirm the configurations.
- **Step 6** Wait until the status of the elastic resource pool changes to **Available**. The elastic resource pool is successfully created.
- **Step 7** Add a SQL queue to the elastic resource pool and select Trino as the execution engine.

NOTE

When buying a SQL queue, select Trino as the execution engine.

----End

Step 2: Configuring Network Connectivity for the DLI Cluster

Step 1 On the **Queue Management** page of the created DLI queue, view the VPC endpoint information of the queue.

- 1. On the DLI console, choose **Resources** > **Queue Management**, and view the VPC endpoint information about one minute after the queue is created.
- 2. Locate the created queue and click in front of the queue name to obtain the VPC endpoint information of the queue.

Figure 3-62 VPC endpoint information

Data Lake Insight	Queue Ma	anagement									G Fee	dback Buy Quave	Buy DLI Package
Overview	Crea	te SMN Topic								Name:	Add fiber		x Q C
SQL Editor		Name	Type 🍞	Specificatio	Actual CUs ↓⊟	Elastic Scaling	Billing Mode		Username	Enterprise Project	Description	Operation	
Job Management +	^	gaze	For SQL			Max - CUs Min: - CUs	Resource pool Created on Jun 20, 20	19:23:43 G				Delete Perm	ssions More 🕶
Resource Pool	Nar		gaze					Engine	openLooKeng				
Queue Management			neihezhuanyong					CPU Architecture					
		licated Resource						AZ Mode	Single AZ				
Data Management 🔹	CID	R Block	1					Username	el_dlcs_d00352221				
Job Templates 🔹	Cre	bete	Jun 20, 2023 19:23:43 G	MT+08:00				VPC					

Step 2 Create a VPC endpoint.

- 1. Log in to the VPC Endpoint management console.
- 2. On the VPC Endpoints page displayed, click Buy VPC Endpoint.
- 3. Set Service Category to Find a service by name.
- 4. In the **VPC Endpoint Service Name** field box, enter the **obtained** VPC endpoint information, excluding the port.

Example:

The VPC endpoint information of the queue is **xxx.3a715f69-b1b0-45d0-bc4a-d917137bcd08:18090**.

Enter xxx.3a715f69-b1b0-45d0-bc4a-d917137bcd08 in the field box.

Figure 3-63 Buy VPC Endpoint page

* Region	Q
	Regions are geographic areas isolated from each other. Resources are region-specific and cannot be used across regions through internal network connections. For low network latency and quick resou
* Billing Mode	Pay-per-use (?)
* Service Category	Cloud services Find a service by name
* VPC Endpoint Service Name	Enter a private service name and verify. Verify
* VPC	di_vpc_rds(10.0.0.016) ▼ C View VPCs
Тад	It is recommended that you use TMS's predefined tag function to add the same tag to different cloud resources. View predefined tags C
	Tag key Tag value
	You can add 10 more tags.
Description	
	0/512

Step 3 Obtain the IP address of the VPC endpoint.

- In the navigation pane of the VPCEP console, choose VPC Endpoint > VPC Endpoints.
- 2. Click the ID of the VPC endpoint and view the node IP address on the **Summary** tab page.

Figure 3-64 IP address of the VPC endpoint

_						
*	HUAWEI CLOUD					'я 🧬
Ξ	<					C
	Summary Tags					
ఉ						
0	ID.		Status	Rejected		
~	VPC		Тура	Interface		
٢	VPC Endpoint Service Name		Assigned			
	Private IP Address	15	Private Domain Name			
	Description					

Step 4 Create an ELB.

- 1. Log in to the ELB console.
- 2. Click **Buy Elastic Load Balancer** and then configure the parameters.

Figure	3-65	Buy	Elastic	Load	Balancer	page
--------	------	-----	---------	------	----------	------

Basic Information	
* Type	Dedicated Shared Learn more
* Billing Mode	Yearly/Monthly Payper-use
* Region	CN North-Ulangat200
	Regions are geographic areas isolated from each other. Resources are region specific and cannot be used across regions through internal network connections. For low network latency and quick resource access, select the nearest region.
* AZ	A22 (4) True can choose to deploy the load balancer in multiple AZs for higher availability.
	too unit unione au ongoy inte owa waarken in toologen waarke ingene analoning.
Network Configurat	tion
* IP as a Backend	0
Network Type	🥑 Public IPv4 metwork/Public network traffic) 😨 Private IPv4 metwork/Private metwork traffic) 🗌 IPv6 metwork/Public and private metwork traffic) 🔞
* VPC	vpcdi • C View VPCs
* Subnet	-Select- • C Vew Subnet
* Specifications	Elastic Billed by how many LCUs you will use Fixed Silled by the specifications you will select
	For fluctuating traffic
	Application load balancing(HTTP/HTTPS) 👿 Network load balancing(TCPIUDP) 🔞
	Network load balancingTCP/UDP)
	e appropris na mais na na markaza na performance vili multiply by the number of AZs.
* Namo	
	eb foic
* Enterprise Project	-Salect- C 🕲 Cheate Enterprise Project
* EIP	⊕ New DP ∪ Use existing ⑦
* EIP Type	6.gvm
* Billed By	Est-duidh 🔐 Taffic 🚈 Taffic 🖾 For lightharph fuctuating traffic

Step 5 Obtain the service IP address of the ELB.

1. On the ELB console, choose **Elastic Load Balance** > **Load Balancers**.

Figure 3-66 Load balancer list

HUAWEI CLOUD					Biling & Costs [®] Re	sources Enterprise Develop	er Tools ICP License	Support Se		4 G
Network Console	Elastic Load Balance ③								BP Quick Links	Buy Elastic Load Balance
Dashboard	Dedicated load balancers now prov	ide elastic scaling to handle traffic peaks	and troughs. The	resources needed for so	aling up are charged per use. Try	now				
Virtual Private Cloud 🔹	Renew Charge Billing Mc	de Utsubscribe								C [] 0
Access Control + Routing Control +	Name-1D	Monit Status	Туре	Specifications	IP Address and Network	Listener (Frontend Proto	Bandwidth Infor	Billing Mode	Enterprise Project	Operation
VPC Row Logs Elastic IP and		۲			antipetar (seco)					Add Listener More +
Bandwidth NAT Galeway • Elastic Load Balance •	-					Istener-Dell (TCP/11096)				Add Listener More +
Backend Server Groups										Add Listener More +
Certificates IP Address										Add Listener More +

2. Click the ID of the created load balancer. On the **Summary** tab page, view the load balancer information, and record the IPv4 EIP address.

Figure 3-67 Dedicated load balancer

We would	HU/	WEICLOUD 습 Con	sole V ···································	Billing & Costs Resources	Enterprise Developer Tools ICP License Support Service Tick	its Erglish	ă 🛛 🖉
Ξ		Load balancer (lwz-quick)	i-Hest) 🧿 Running			Add Listener	Backend Server Groups v
	S	ummary Listeners	Monitoring Access Logs Tags				
යි							
0		Name	Iwz-quickbi-Hest 🖉	VPC	vpc-NoDelete_AutoTest		
۲		ID	4885c165-7a03-43ea-8a82-45696ee5074e 🗇	IPv4 Subnet	vpc-NoDelete_AutoTest		
		Туре	Dedicated	IPv6 Subnet	-		
۵		AZ	A22	Backend Subnet	vpc-NoDelete_AutoTest 🖉		
Θ		Specification	Network load balancing(TCP/UDP) Small	IP as a Backend	Enabled (?)		
6		Billing Mode	Pay-per-use	IP Address	Private IPv4 address 15		
					IPv4 EIP 10		
۲					IPv6 address Bind		
6		Enterprise Project	default	Bandwidth Information	IPv4 5 Mbit/s Pay-per-use - By traffic		
		Description	- 🖉	Created	Nov 28, 2022 18:54:07 GMT+08:00		
	•	Modification Protection (?)	Disabled Configure				

Step 6 Create a datasource connection.

- 1. Log in to the DLI management console.
- 2. In the left navigation pane, choose **Datasource Connections**.

On the **Enhanced** tab page displayed, click **Create**. In the **Create Enhanced Connection** dialog box, enter a connection name in **Connection Name**, set **Resource Pool** to the elastic resource pool that contains the Trino engine queue created in **step 1**, and configure **VPC** and **Subnet**. For details about the parameters, see **Table 3-16**.

Figure 3-68 Creating an enhanced datasource connection

	anced datasource connection, the system will automatically create a VPC peering routes. Learn more about how to connect DLI queues.
Connection Name	Enter a name.
Resource Pool	· .
▶ VPC	vpc-nodel(10.128.0.0/10)
★ Subnet	subnet-nodel(10.128.0.0/24)
Route Table	rtb-vpc-nodel(Default)
Host Information	Enter host information in the format "host IP address host name". Specify the information for each host on a separate line.
Tags	It is recommended that you use TMS's predefined tag function to add the same tag to different cloud resources. View predefined tags ${f C}$
	To add a tag, enter a tag key and a tag value below.

Table 3-16 Parameters

Paramete r	Description
Connectio n Name	 Name of the datasource connection to be created The name can contain only digits, letters, and underscores (_) but cannot be left blank. Enter a maximum of 64 characters.
Resource Pool	It binds an elastic resource pool or queue that uses a datasource connection. This parameter is optional. Only dedicated queues charged in yearly/monthly or pay-per- use billing mode can be bound to elastic resource pools. In regions where this function is available, an elastic resource pool with the same name is created by default for the yearly/ monthly or pay-per-use dedicated queue created in "Creating a Queue." NOTE Before using an enhanced datasource connection, you must bind a queue to the connection and ensure that the VPC peering connection is in the Active state.
VPC	VPC used by the destination data source
Subnet	Subnet used by the destination data source
Route Table	 Route table of the subnet NOTE The route table is associated with the subnet used by the destination data source, which is not the table containing the route you add by Manage Route in the Operation column. The route you add on the Manage Route page is contained in the route table associated with the subnet used by the queue to be bound. The subnet used by the destination data source must be different from that used by the queue to be bound. Otherwise, a segment conflict occurs.
Tags	Tags used to identify cloud resources.

- 3. Click **OK**.
- 4. Check whether the datasource connection is successfully created.

Click the name of the created datasource connection to view its connection status. If the status is **Active**, the datasource connection is successfully created.

- **Step 7** Add a backend server group as the VPC backend. The cross-VPC backend IP address is the IP address of the purchased VPC endpoint.
 - On the ELB console, choose Elastic Load Balance > Backend Server Groups. On the page displayed, click Create Backend Server Group.
 - 2. Select the created load balancer for **Load Balancer** and click **Next**. On the **Backend Servers** tab page, click **Next**. On the **Confirm** page, click **Create Now**.

F	Figure 3-69 Creating a backend server group											
	< Create Backend Serve	Group										
	Configure Routing Policy ———	— (2) Add Backend Server ———	3 Confirm									
	★ Load Balancing Type	Dedicated Shared										
	* Load Balancer	-Select- C	View load balancers									
	* Forwarding Mode	Load balancing										
	★ Backend server group name	server_group-3315										
	* Backend Protocol	HTTP •										
	★ Load Balancing Algorithm	Weighted round robin We	eighted least connections	Source IP hash								
•	Sticky Session	0										
	Slow Start	0										
	Description		h									
			0/255									

3. On the **Backend Servers** tab page, click **Add Backend Server** in the **Operation** column of the created backend server group to add backend servers.

Figure 3-70 Cross-VPC backend IP address and service port

Backed Server P. Address Bagementary Nations Image: Control of the c	Summary Bac	kend Servers			
Adi tudor Wright Remove C V Specify Rise rubes Q Backend Server & Publics Q Backend Server & Publics Backend Fort	Backend Server	IP as Backend Servers Supplementary Network	Interfaces		
Bacterd Server IP Address Health Check Result () Weight Bacterd Port	Add	odity Weight Remove			C
Backend Server IP Address Health Check Result O Weight Backend Pert	V Specify filte	r criteria.			Q
		Server IP Address	Health Check Result ③	Weight	Backend Port

Step 8 Verify that the network connection between VPCEP and DLI is normal.

On the **IP as Backend Servers** tab page, check whether the health check result is normal. If yes, the network connection is normal.

Figure 3-71 Successful network connection

Bactand Servers IP as Bactand Servers Supplementary Network Interfaces									
Add Modty Weight Remove									
▼ Specify filter criteria.									
Backend Server IP Address	Health Check Result (?)	Weight	Backend Port						
Dackend Server IP Address	-	weight	Backend Port						
10.0.0.72		1	18090						

End

Step 3: Installing the Trino Driver for FineBI

- 1. Install FineBI.
- 2. Install the Trino driver for FineBI.

Visit Trino to download the Trino driver JAR file.

On the FineBI console, choose **Management System** > **Data Connection Management**, click **New Driver**, click **Upload File**, and upload the downloaded driver package.

Figure 3-72 Configuring the Presto driver

Save
3
· · · ·
uploaded Upload File

Step 4: Interconnecting FineBI with DLI Trino

Configure the interconnection between FineBI and DLI.

- On the FineBI management console, choose Data Connection Management
 > Create Data Connection > Other > Other JDBC.
- 2. Enter information about the data connection.
 - a. Enter the data connection name.
 - b. Set **Driver** to **Custom** and select **io.prestosql.jdbc.PrestoDriver** as the driver.
 - c. Enter the cross-VPC backend IP address for **Host** and enter the service port in **Port**. For details, see **Creating a backend server group**.
 - d. Enter the username and password. The username is in the format of Account name/Username/Project ID. For details about how to obtain a project ID, see Obtaining a Project ID. If a primary account is used for connection, Account name and Username are both the account name.
 - e. Example data connection URL: jdbc:presto://{ip}/dli/default? SSL=true

D NOTE

In the URL, **SSL=true** indicates that backend requests use HTTPS. Currently, Trino engine queues support only HTTPS connections.

igure 57.	- connigui		
Directory	¢		
/≞ User	S Other JDBC		
Permission	Data Connection Name	Data Connections	
₽ Publication Mana	Driver	Default v org.h2.Driver v	
🗄 Public data manage	Database Name	Database Name	
🗣 Appearance			
System	Host	Host	
Task Schedule	Port	Port	
Mobile Platform	Username	Username	
Registration	Password	Password	
Intelligent Op ∨	Encoding	Default v	
🖉 Data Connect 🔨	Pattern	Click to Connect Database Read Mode List	
🖀 Data Connectio	Pattern	~	0
器 Server Dataset			
Plugin	Data Connection URL	jdbc:h2://\${ENV_HOME}//database	
③ Security	Max active connection	50	
🗟 Template Authenti	Test before connect	Yes 🗸	
Map Configuration	SQL Verification Query	Use the default statement if is empty	
🖥 Data developmen	oqe ronnoadon quorj	oo no solan ola	
B Resource Handover			
Shuzhiniao			
Extracted Data C	Max Wait Time	10000	Millisecond
	 SSH Setting 		
	 More Settings 		

Figure 3-73 Configuration information

Step 5: Testing the Connection

Click **Test Connection** in the upper right corner of the FineBI data connection management page. If the connection is successful, you can use the connection to query DLI tables for BI report analytics.

Figure 3-74 Testing the connection

illebi				4
Directory	0		Car	ncel Test Con
R⊨ User	5 Other JDBC			
⑦ Permission	Data Connection Name	Data Connections		
Publication Mana	Driver	Default v org.h2.Driver v		
9 Public data manage	Database Name	Database Name		
Appearance	Host	Host		
9 System	Port	Port		
5 Task Schedule	Usemame	Usemame		
Registration		Password		
🖇 Intelligent Op 🗸	Password			
Ø Data Connect ^	Encoding	Default ~		
2 Data Connectio	Pattern	Click to Connect Database Read Mode List	0	
🕱 Server Dataset				
> Plugin	Data Connection URL	(dbc/h2://\$(ENV_HOME)/./database		
Security	Max active connection	50		
P Template Authenti	Test before connect	Yes ~		
Map Configuration	SQL Verification Query	Use the default statement if is empty		
§ Data developmen				
B Resource Handover				
Extracted Data C	Max Wait Time	10000	Milliscond	
	 SSH Setting 			

Related Operations

• Trino supports the SQL syntax. For details about the Trino SQL syntax, see Trino SQL Syntax. Currently, the Trino engine supports only the SELECT query operation.

3.8 Interconnecting Power BI with DLI Trino

Application Scenarios

Power BI is a unified, scalable self-service and enterprise business intelligence (BI) platform. You can use it to connect to and visualize any data, and seamlessly integrate visual objects into your daily applications.

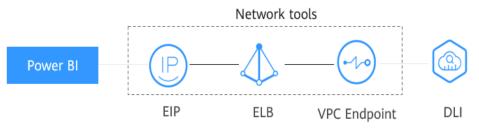
Huawei Cloud DLI provides Power BI with standard, effective, and high-quality data through converged data analysis and processing for subsequent data statistics and analysis, helping enterprises make data decisions.

For more information about Power BI, see Power BI.

Solution Overview

This solution uses VPCEP to connect Power BI to DLI.

Figure 3-75 Architecture



Constraints

- Trino engine queues support only HTTPS connections.
- When the Trino engine is used, the created SQL queues cannot be scaled in or out.

To adjust the CU size of a queue, you need to first **delete** the queue in the elastic resource pool and then **create** a queue with Trino as the engine and with an appropriate CU size in the elastic resource pool.

• The DLI Trino engine is in the open beta test (OBT) phase. If you need it, contact customer service to apply for it.

The DLI Trino engine is available in the following regions: **CN North-Beijing4**, **CN East-Shanghai1**, **CN-Hong Kong**, **AP-Bangkok**, **AP-Singapore**, and **AF-Johannesburg**.

• Currently, only foreign tables created using the Hive syntax can be used for Power BI interconnection.

Process

Figure 3-76 Process of interconnecting Power BI with DLI Trino



To interconnect DLI Trino with Power BI, perform the following steps:

- 1. Creating a DLI Elastic Resource Pool and Queue
- 2. Configuring Network Connectivity for the DLI Cluster
- 3. Installing Power BI and Its Driver
- 4. Interconnecting Power BI with DLI Trino
- 5. Testing the Connection

Solution Advantages

- This BI tool for big data analytics provides data exploration capabilities by converting data into various forms, such as charts, tables, maps, and dashboards, making data more intuitive and easy to understand.
- Huawei Cloud DLI provides convergent data analysis and processing capabilities. DLI can interconnect with multiple data sources and map data sources by creating tables using SQL statements. DLI provides powerful data exploration capabilities to deeply explore data potentials through data filtering, sorting, and grouping.
- Power BI interconnects with Huawei Cloud DLI for real-time data ingestion, high data accuracy, efficient data processing, and good data visualization. After DLI interconnects with Power BI, data from different data sources can be integrated. DLI supports big data processing. It can process a large amount of data and mine the potential of the data. Power BI can quickly visualize the data to improve the efficiency and precision of data analytics.

Resource Planning and Costs

Table 3-17 Resource planning and costs	
--	--

Resource	Description	Cost
OBS	DLI needs to use OBS buckets to store logs.	You will be charged for using the following OBS resources:
		 Storage Fee for storing static website files in OBS.
		 Request Fee for accessing static website files stored in OBS.
		 Traffic Fee for using a custom domain name to access OBS over the public network.
		The actual fee depends on the size of the stored file, the number of user access requests, and the traffic volume. Estimate the fee based on your service requirements.
DLI	In this example, an elastic resource pool is used to create SQL jobs.	When using a DLI elastic resource pool, you are billed based on the CUH of the elastic resource pool.
VPCEP	Used to enable the network connection between FineBI and DLI.	For details about the billing for VPCEP, see Billing .
ELB	Elastic Load Balance (ELB) distributs access traffic to multiple backend servers based on distribution policies.	For details about the billing for ELB, see Billing .
EIP	Provides independent public IP addresses and bandwidth for Internet access.	For details about the billing for EIP, see Billing .

Step 1: Creating an Elastic Resource Pool and Queue

- **Step 1** Log in to the DLI management console.
- **Step 2** In the navigation pane on the left, choose **Resources** > **Resource Pool**.
- **Step 3** On the **Resource Pool** page, click **Buy Resource Pool** in the upper right corner.
- **Step 4** On the displayed page, set the following parameters:

Parameter	Description
Billing Mode	Pay-per-use/Yearly/Monthly
Region	Select a region. Select a region near you to ensure the lowest latency possible.
Project	Each region corresponds to a project.
Name	Name of the elastic resource pool.
CU Range	The maximum and minimum CUs allowed for the elastic resource pool.
Description	Description of the elastic resource pool.
CIDR Block	CIDR block the elastic resource pool belongs. If DLI enhanced datasource connections are required, the CIDR block of the elastic resource pool cannot overlap with that of the data source. The CIDR block of the elastic resource pool cannot be changed after being set. Recommended CIDR blocks: 10.0.0~10.255.0.0/16~19 172.16.0.0~172.31.0.0/16~19 192.168.0.0~192.168.0.0/16~19
Enterprise Project	If the created elastic resource pool belongs to an enterprise project, select the enterprise project.
Required Duration	You must specify the Required Duration if Billing Mode is set to Yearly/Monthly . The longer the subscription duration is, the more discounts you can get. If Auto renew is selected, monthly subscriptions are renewed each month. And yearly subscriptions are renewed each year.
Tags	Tags used to identify cloud resources.

- **Step 5** Click **Buy** and confirm the configurations.
- **Step 6** Wait until the status of the elastic resource pool changes to **Available**. The elastic resource pool is successfully created.
- **Step 7** Add a SQL queue to the elastic resource pool. The selected engine is Trino.

NOTE

When buying a SQL queue, select Trino as the execution engine.

----End

Step 2: Configuring Network Connectivity for the DLI Cluster

Step 1 On the **Queue Management** page of the created DLI queue, view the VPC endpoint information of the queue.

- 1. On the DLI console, choose **Resources** > **Queue Management**.
- 2. Locate the created queue and click in front of the queue name to obtain the VPC endpoint information of the queue.

Figure 3-77 VPC endpoint information

Data Lake Insight	Que	eue Management									Feedt	Buy Queue	Buy DLI Package
Overview		Create SMN Topic								Name:	Add filter		x Q C
SQL Editor		Name	Type 🍞	Specificatio	Actual CUs ↓Ξ	Elastic Scaling	Billing Mode		Username	Enterprise Project	Description	Operation	
Job Management +		∧ gaze	For SQL			Max - CUs Min: - CUs	Resource pool Created on Jun 20, 20	23 19:23:43 G				Delete Permis	sions More 🕶
Resource Pool		Name	gaze					Engine	openLooKeng				
Queue Management		Resource Pool Dedicated Resource	neihezhuanyong Yes					CPU Architecture AZ Mode	X86 Single AZ				
Data Management 🔹		CIDR Block	1					Usemame	el_dlics_d00352221				
Job Templates 💌		Created	Jun 20, 2023 19:23:43 GP	/T+08:00				VPC					

Step 2 Create a VPC endpoint.

- 1. Log in to the VPC Endpoint management console.
- 2. Click **Buy VPC Endpoint**. The **Buy VPC Endpoint** page is displayed.
- 3. Set Service Category to Find a service by name.
- In the VPC Endpoint Service Name field box, enter the obtained VPC endpoint information, excluding the port.
 Example:

The VPC endpoint information of the queue is **xxx.3a715f69-b1b0-45d0-bc4a-d917137bcd08:18090**.

Enter xxx.3a715f69-b1b0-45d0-bc4a-d917137bcd08 in the field box.

Figure 3-78 Buy VPC Endpoint page

* Region	Regions are geographic areas isolated from each other. Resources are region-specific and cannot be used across regions through internal network connections. For low network latency and quick resources
* Billing Mode	Pay-per-use 0
* Service Category	Cloud services Find a service by name
* VPC Endpoint Service Name	Enter a private service name and verify. Verify 🕖
* VPC	dL_vpc_rds(10.0.0.016) • C View VPCs
Tag	Il is recommended that you use TMS's predefined tag function to add the same tag to different cloud resources. View predefined tags C
	Tag key Tag value
	You can add 10 more tags.
Description	

Step 3 Obtain the IP address of the VPC endpoint.

- In the navigation pane of the VPCEP console, choose VPC Endpoint > VPC Endpoints.
- 2. Click the ID of the VPC endpoint and view the node IP address on the **Summary** tab page.

Figure 3-79 IP address of the VPC endpoint

<u>*</u>	HUAWEI CLOUD									'e 🛃
≣	<[С
٢	Summary Tags									
ස										
0	ID		Status		Rej	ected				
0	VPC		Type		Inte	rface				
S	VPC Endpoint Service Name		Assigned							
	Private IP Address	15	Private D	omain Name						
	Description									

Step 4 Create an ELB.

- 1. Log in to the ELB console.
- 2. Click **Buy Elastic Load Balancer** and then configure the parameters.

Figure 3-80 Buy Elastic Load Balancer page

Basic Information	
* Type	Dedicated Shared Learn more
* Billing Mode	Vearly/Monthly Payper-use
* Region	CNHarth Ulampib203 *
	Regions are geographic areas isolated from each other. Resources are region-specific and cannot be used across regions through internal network connections. For low network latency and quick resource access, select the network region.
* AZ	AZ2 (4)
Network Configuration	
* IP as a Backend	
Network Type	👽 Public IPV4 network(Public network traffic) 💟 Private IPV4 network(Private network traffic) 🗌 IPV6 network(Public and private network traffic) 🔞
* VPC	vpodi • C VeerVPOs
* Subnet	-Select- Vew Subnet
* Specifications	Eastic Brief by how many LCUs you will use Find Brief by the specifications you will see
	For fuctuating traffic
	Application load belanching(HTTPHTTPS) 🛃 Network load belanching(TCPUDP) 💿
	Network new usaitung (U-Hour)
	deploy the load balancer in more AZs, its performance will multiply by the number of AZs.
* Namo	ab-fde
* Enterprise Project	-Select- • C O Create Enterprise Project
* EIP	New EIP Use existing
* EIP Type	5_gvm
* Billed By	For heavy/stable traffic

Step 5 Obtain the service IP address of the ELB.

1. On the ELB console, choose **Elastic Load Balance** > **Load Balancers**.

Figure 3-81 Load balancer list

Network Console	Elastic Load E	Balance 🛞									BP Quick Links	Buy Elastic Load Ba
Dashboard	Dedicated is	pad balancers now provide elec	stic scaling to handle	traffic peaks an	d troughs. The n	esources needed for sc	aling up are charged per use. Try r	now				
Virtual Private Cloud	Renew	Charge Billing Mode	Unsubscribe									C
Access Control +	T Specify I	filter criteria.										
Routing Control +	Name	e1D	Monit	Status	Type	Specifications	IP Address and Network	Listener (Frontend Proto	Bandwidth Infor	Billing Mode	Enterprise Project	Operation
VPC Flow Logs Elastic IP and			۵				actification (series)					Add Listener More +
Bandwidth NAT Galleway Elastic Load Balance Load Balancers								Istener-Daß (TCP/11096)				Add Listener More +
Backend Server Groups												Add Listener More +
Certificates												Add Listener More +

2. Click the ID of the created load balancer. On the **Summary** tab page, view the load balancer information, and record the IPv4 EIP address.

Figure 3-82 Dedicated load balancer

н	UAWEI CLOUD 🏠 Con	ssole Q V	Billing & Costs® Resources	Enterprise Developer Tools ICP License	Support Service Tick	xts English ei_	dics_d00352221 🕞 'ç
Ξ	C Load balancer (N/2-quick)	bl-Best) 🔕 Running				Add Listener	Backend Server Groups 🔹
	Summary Listeners	Monitoring Access Logs Tags					
ది							
0	Name	tvz-quictéritest 🖉	VPC	vpc-NoDelete_AutoTest			
0	ID	4885c165-7a03-43ee-8a82-45696ee5074e 🗇	IPv4 Subnet	vpc-NoDelete_AutoTest			
~	Туре	Dedicated	IPv6 Subnet				
	AZ	A22	Backend Subnet	vpc-NoDelete_AutoTest 🖉			
<u>ن</u>	Specification	Network load balancing(TCP/UDP) Small I	IP as a Backend	Enabled (?)			
۵	Billing Mode	Pay-per-use	IP Address	Private IPv4 address 15			
				IPv4 EIP 10 🗗 Unbind			
۲				IPv6 address Bind			
6	Enterprise Project	default	Bandwidth Information	IPv4 5 Mbitis Pay-per-use - By traffic			
	Description	- 🖉	Created	Nov 28, 2022 18:54:07 GMT+08:00			
•	Modification Protection (?)	Disabled Configure					

Step 6 Create a datasource connection.

- 1. Log in to the DLI management console.
- 2. In the left navigation pane, choose **Datasource Connections**.

On the **Enhanced** tab page displayed, click **Create**. In the **Create Enhanced Connection** dialog box, enter a connection name in **Connection Name**, set **Resource Pool** to the elastic resource pool that contains the Trino engine queue created in **step 1**, and configure **VPC** and **Subnet**. For details about the parameters, see **Table 3-19**.

Figure 3-83 Creating an enhanced datasource connection

Enter a name.
¥ J
vpc-nodel(10.128.0.0/10)
subnet-nodel(10.128.0.0/24)
rtb-vpc-nodel(Default)
Enter host information in the format "host IP address host name". Specify the information for each host on a separate line.
It is recommended that you use TMS's predefined tag function to add the same tag to different cloud resources. View predefined tags $\ C$
To add a tag, enter a tag key and a tag value below.

Table 3-19 Parameters

Paramete r	Description
Connectio n Name	 Name of the datasource connection to be created The name can contain only digits, letters, and underscores (_) but cannot be left blank. Enter a maximum of 64 characters.
Resource Pool	It binds an elastic resource pool or queue that uses a datasource connection. This parameter is optional. Only dedicated queues charged in yearly/monthly or pay-per- use billing mode can be bound to elastic resource pools. In regions where this function is available, an elastic resource pool with the same name is created by default for the yearly/ monthly or pay-per-use dedicated queue created in "Creating a Queue." NOTE Before using an enhanced datasource connection, you must bind a queue to the connection and ensure that the VPC peering connection is in the Active state.
VPC	VPC used by the destination data source
Subnet	Subnet used by the destination data source
Route Table	 Route table of the subnet NOTE The route table is associated with the subnet used by the destination data source, which is not the table containing the route you add by Manage Route in the Operation column. The route you add on the Manage Route page is contained in the route table associated with the subnet used by the queue to be bound. The subnet used by the destination data source must be different from that used by the queue to be bound. Otherwise, a segment conflict occurs.
Tags	Tags used to identify cloud resources.

- 3. Click **OK**.
- 4. Check whether the datasource connection is successfully created.

Click the name of the created datasource connection to view its connection status. If the status is **Active**, the datasource connection is successfully created.

- **Step 7** Add a backend server group as the VPC backend. The cross-VPC backend IP address is the IP address of the purchased VPC endpoint.
 - On the ELB console, choose Elastic Load Balance > Backend Server Groups. On the page displayed, click Create Backend Server Group.
 - 2. Select the created load balancer for **Load Balancer** and click **Next**. On the **Backend Servers** tab page, click **Next**. On the **Confirm** page, click **Create Now**.

	ig a backena server group
< Create Backend Server	r Group
Configure Routing Policy	— (2) Add Backend Server — (3) Confirm
* Load Balancing Type	Dedicated Shared
* Load Balancer	-Select- View load balancers
* Forwarding Mode	Load balancing
★ Backend server group name	server_group-3315
* Backend Protocol	HTTP •
★ Load Balancing Algorithm	Weighted round robin Weighted least connections Source IP hash
Sticky Session	0
Slow Start	0
Description	
	0/255

Figure 3-84 Creating a backend server group

- 3. On the **Backend Servers** tab page, click **Add Backend Server** in the **Operation** column of the created backend server group to add backend servers.
- Step 8 Verify that the network connection between VPCEP and DLI is normal.

On the **IP as Backend Servers** tab page, check whether the health check result is normal. If yes, the network connection is normal.

----End

Step 3: Installing Power BI and Its Driver

Step 1 Install Power BI of the desktop version.

Download the Power BI desktop version.

Step 2 Install the openLooKeng ODBC driver.

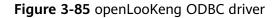
Before installing the **driver**, ensure that you have the administrator rights.

- 1. Double-click the .msi installation package. The welcome page is displayed. Click **Next**.
- 2. The second page is the user agreement. Accept the terms and click **Next**.
- 3. On the third page, select an installation mode. You are advised to select **Complete**.
- 4. On the fourth page, select an installation path and click **Next**.
- 5. After the preceding installation settings are complete, click **Install** on the last page to start the installation.

NOTE

During the installation, the CLI is displayed to show the process of installing the driver components. After the installation is complete, the CLI is automatically closed. The openLooKeng ODBC driver is installed.

In the dialog box displayed, use DSN for new installation if you have configured the user DSN for an earlier version and click **Finish**.



	Community Newsroom Code	Q English ~
anani asi/ang Sanjar	hada araan 4.40 A ka ar	
openLooKeng Server	hetu-server-1.10.0.tar.gz	hetu-server-rpm-1.10.0.x86_64.rpm
openLooKeng deployment packages. See Deploying openLooKeng for deployment instructions.	业 Download SHA256 d	业 Download SHA256 d
Command Line Interface	JDBC Driver	ODBC Driver
openLooKeng CLI provides a simple way to execute SQL statements in interactive mode. See Command Line Interface for installation instructions.	JOBC Driver enables users to connect with live openLookeng Server, directly from any Java applications that support JDBC connectivity. See JDBC Driver for installation instructions.	ODBC Driver lets you communicate with openLooKeng Server via standard ODBC protocol. See openLooKeng ODBC User Manual for installation instructions.
hetu-cli-1.10.0-executable.jar	hetu-jdbc-1.10.0.jar	hetu-odbc-win64-1.10.0.msi
业 Download SHA256 ₫	→ Download SHA256	➡ Download SHA256 🗐

----End

Step 4: Interconnecting Power BI with DLI Trino

- 1. Stop the ODBC service.
 - a. Run the following command to go to the C:\Program Files \openLooKeng\openLooKeng ODBC Driver 64-bit\odbc_gateway \mycat\bin directory:

cd C:\Program Files\openLooKeng\openLooKeng ODBC Driver 64-bit\odbc_gateway\mycat\bin

- b. Run the following command to stop the ODBC service: mycat.bat stop
- 2. Replace the JDBC driver.
 - a. Copy the JDBC JAR file obtained from driver to the C:\ProgramFiles \openLooKeng\openLooKeng ODBC Driver 64-bit\odbc_gateway \mycat\lib directory.
 - b. Delete the existing hetu-jdbc-1.0.1.jar file from the directory.
- 3. Edit the protocol prefix of the ODBC server.xml file.

Change the property value of the **server.xml** file in the **C:\Program Files \openLooKeng\openLooKeng ODBC Driver 64- bit\odbc_gateway\mycat \conf** directory from <property name="jdbcUrlPrefix">jdbcUrlPrefix">jdbc:lk://</property> to <property name="jdbcUrlPrefix">jdbc:presto://</property>.

4. Configure the connection mode of using the username and password.

Create the **jdbc_param.properties** file in a user-defined path, for example, **D:**\, and add the following content to the file:

```
SSL=true
user={Account name}/{Username}/{Project ID}
password={Password}
```

- 5. Restart the ODBC service.
 - a. Run the following command to go to the C:\Program Files \openLooKeng\openLooKeng ODBC Driver 64-bit\odbc_gateway \mycat\bin directory:

cd C:\Program Files\openLooKeng\openLooKeng ODBC Driver 64-bit\odbc_gateway\mycat\bin

- b. Run the following command to restart the ODBC service: mycat.bat restart
- 6. Set up ODBC data sources (64-bit).

Enter **odbc** in the control panel of the Windows OS to search for the ODBC management program. Click **Set up ODBC data sources (64-bit)**.

Figure 3-86 Clicking Set up ODBC data sources (64-bit)



7. Add a driver.

In the dialog box displayed, click **Add**. In the **Create New Data Source** dialog box, click **openLookeng ODBC 1.9 Driver** and click **Finish**.

Figure 3-87 Adding a driver

odb		BC Data Source A	Administra	itor (64-bi	t)					\times	
ab		SN System DSN	File DSN	Drivers	Tracing	Connection Po	ooling	About			
	User	Data Sources:								_	
	Nan	-		Driver					Add		
		el Files Access Database	64-bit 64-bit 64-bit	Microsoft I	Excel Driv	C 1.9 Driver er (* xls, * xlsx, *. iver (* .mdb, * .ac		xlsb)	Remove		
									Configure		
	Create N	ew Data Source	Name Micro Micro Micro	e osoft Acces osoft Acces	s Driver (* s Text Dri Driver (* x	u want to set up 'mdb, *.accdb) ver (* bt., *.csv) ls, *.viex, *.vien, Driver		version 15.00.4(15.00.4(ted data provider. A		
				<	Back	Finish		Cancel			

8. Enter data source information.

Enter the name and description by referring to Figure 3-88 and click Next.

Figure 3-88 Entering data source information

×

- 9. Configure related information.
 - Connect URL: indicates the address for accessing openLooKeng. Enter the value in the format of *IP address.Port*. For details about how to obtain the IP address and port, see Obtaining the IP address and port.

 Connect Config: Select the jdbc_param.properties file in 4.
 Example content of the properties file: SSL=true user=xxx_d21/xx_352221/xxxxacc00a2e2 password=xxxx12*

NOTE

In the URL, **SSL=true** indicates that backend requests use HTTPS. Currently, Trino engine queues support only HTTPS connections.

- Set Catalog to dli.

Figure 3-89 Configuration information

t up ODBC data sources (64-bit)

	Please specify par	ams to connect to o	openLooKeng		π	
	Connect URL:					Add
	Connect Config:			Browse		Remove
	User name:					
R	Password:			Test DSN		Configure
openLooKeng	Please specify a c	atalog and schema	to use			
	Catalog: c	đli	✓ Schema:	~		
(1) -	< Pr	evious Next	> Car	ncel Help		
and the					e indica	ated data provider

Step 5: Testing the Connection

1. Click **Test DSN**. If the connection is successful, click **Next** > **Finish** to complete the test.

Figure 3-90	Testing the	connection
-------------	-------------	------------

Create a new Data S	ource to openLo	ooKeng	×
	Please specify par	rams to connect to openLooKeng	
	Connect URL:	100.95.149.134:18090]
	Connect Config:	D:\jdbc_param.properties	Browse
	User name:	[
B	Password:		Test DSN
penLooKeng	Please specify a c	atalog and schema to use	
()) () () () () () () () () (Catalog:	dli 🗸 Schema:	~
	< P1	revious Next > Cance	l Help

Figure 3-91 Successful connection

Connecti	on test	×
1	Connection successfully established Server Information: openLooKeng 05.06.000029 Connection String: DRIVER=	
	ОК]

2. Use Power BI to interconnect with DLI. Choose **Get data** > **All** > **ODBC** > **Connect**.

You need to enter the password for logging in to the DLI console for the first connection.

Get Data		<
odbc X	AII	
All	◆ ODBC	
Other		
Certified Connectors Templa	ate Apps Connect Cancel	

Figure 3-92 Connecting to Power BI

Related Operations

• Trino supports the SQL syntax. For details about the Trino SQL syntax, see Trino SQL Syntax. Currently, the Trino engine supports only the SELECT query operation.

4 Connections

4.1 Configuring the Connection Between a DLI Queue and a Data Source in a Private Network

Background

If your DLI jobs need to connect to external data sources, for example, MRS, RDS, CSS, Kafka, or GaussDB(DWS), you need to enable the network between DLI and the external data sources. DLI enhanced datasource connection uses VPC peering to directly connect the VPC networks of the destination data sources for point-to-point data exchanges.

This section provides a guide to help you connect to data sources. You can also refer to this section to rectify connection faults.

Development Process

Figure 4-1 Configuration process of an enhanced datasource connection



Prerequisites

• You have created a queue. For details about how to create a queue, see **Creating a Queue**.

The queue billing mode must be **Pay-per-use**, and **Dedicated Resource Mode** must be selected after you select a queue type.

Enhanced datasource connections can be created only for pay-per-use resources in dedicated resource mode.

• A cluster of the external data source has been created. You can select a data source as needed.

Service Name	Reference Documents			
RDS	Getting Started with RDS for MySQL			
GaussDB(DWS)	Creating a GaussDB(DWS) Cluster			
DMS Kafka	Creating a Kafka Instance CAUTION When you create the instance, do not enable Kafka SASL_SSL.			
CSS	Creating a CSS Cluster			
MRS	Creating an MRS Cluster			

- The CIDR block of the DLI queue bound with a datasource connection cannot overlap with the CIDR block of other data sources.
- Datasource connections cannot be bound with the **default** queue.

Step 1: Obtain the Floating IP Address, Port Number, and Security Group of an External Data Source

Table 4-2 Data source in	formation to	be obtained
--------------------------	--------------	-------------

Dat a Sour ce	Obtain Method
DMS Kafk a	 On the Kafka management console, click an instance name on the DMS for Kafka page. Basic information of the Kafka instance is displayed. In the Connection page, obtain the Instance Address (Private)
	 In the Connection pane, obtain the Instance Address (Private Network) value. In the Network pane, obtain the VPC and subnet of the instance.
	3. In the Network pane, obtain the security group of the instance.
RDS	On the Instances page of the RDS console, click the target DB instance name. In the displayed page, locate the Connection Information pane and obtain the Floating IP Address , VPC , Subnet , Database Port , and Security Group .

Dat a Sour ce	Obtain Method
CSS	 On the CSS management console, choose Clusters > Elasticsearch. On the displayed page, click the name of the created CSS cluster to view basic information.
	 On the Cluster Information page, obtain the Private Network Address, VPC, Subnet, and Security Group.
Gaus sDB(DWS	1. On the GaussDB(DWS) management console, choose Clusters . On the displayed page, click the name of the created GaussDB(DWS) cluster to view basic information.
)	2. On the Basic Information tab, locate the Database Attributes pane and obtain the private IP address and port number of the DB instance. In the Network pane, obtain the VPC, subnet, and security group information.

Dat Obtain Method a Sour ce	
 An MRS 3.x cluster is used as an example. HBa se Log in to the MRS management console, click a cluster name on Clusters > Active Clusters page to view basic information. On the dashboard, obtain VPC, subnet, and security group from the Basic Information pane. The ZooKeeper instance and its port of the MRS cluster are required for creating a job that connects DLI to MRS HBase. You need to obtain the host information of the MRS cluster. Log in to MRS Manager by referring to Accessing FusionInsig Manager. On MRS Manager, choose Cluster > Name of the desired cluster > Services > ZooKeeper. Click the Instance tal obtain the ZooKeeper host information such as the host name service IP address. On MRS Manager, choose Cluster and click the name of the desired cluster. Choose Services > ZooKeeper. Click the Configurations tab and select All Configurations, search for clientPort parameter, and obtain its value, that is, the ZooKeep port number. Log in to any MRS node as user root in SSH mode. For details Logging In to an ECS. Run the following command to obtain MRS hosts information Copy and save the information. cat /etc/hosts An example query result is as follows: 	the red Jht o and e and the eper 5, see

Step 2: Obtain the CIDR Block of the DLI Queue

On the DLI management console, choose **Resources** > **Queue Management** from

the navigation pane. Locate the queue you have created, and click $\stackrel{\textstyle{\scriptstyle \bigvee}}{=}$ next to the queue name to view the CIDR block of the queue.

Step 3: Add a Rule to the Security Group of the External Data Source to Allow Access from the DLI Queue

- 1. Log in to the VPC console.
- 2. In the navigation pane on the left, choose **Access Control** > **Security Groups**.
- 3. Click the name of the security group to which the external data source belongs.

To obtain the security group information, go to the management console of the data source service and follow the steps provided in **Step 1: Obtain the Floating IP Address, Port Number, and Security Group of an External Data Source**.

4. In the **Inbound Rules** tab, add a rule to allow access from the queue network segment.

For details about how to set the inbound rule parameters, see **Table 4-3**.

Figure 4-2 Adding an inbound rule

Add Inbound	I Rule Learn	more about security group co	nfiguration.			>
Some securi	ity group rules will no	t take effect for ECSs with certain spe	cifications. Learn more	e		
Security Group d You can import multi	efault iple rules in a batch.					
Priority (?)	Action (?)	Protocol & Port (?)	Туре	Source (?)	Description	Operation
1-100	Allow 🔻	Protocols/TCP (Custo Example: 22 or 22,24 or 22-31	IPv4 v	IP address 0.0.0.0/0		Replicate Delete
Add Rule Cancel						

Parameter	Description	Example
Priority	The security group rule priority.	1
	The priority value ranges from 1 to 100. The default value is 1 , indicating the highest priority. A smaller value indicates a higher priority of a security group rule.	
Action	Action of the security group rule.	Select Allow .

Parameter	Description	Example
Protocol &Port	 Network protocol: The value can be All, TCP, UDP, ICMP, or GRE. Port: Port or port range over which the traffic can reach your instance. The port ranges from 1 to 65535. 	In this example, select TCP. Leave the port blank or set it to the data source port obtained in Step 1: Obtain the Floating IP Address, Port Number, and Security Group of an External Data Source.
Туре	Type of IP addresses.	IPv4
Source	Allow access from IP addresses or instances in another security group.	In this example, enter the queue network segment obtained in Step 2: Obtain the CIDR Block of the DLI Queue .
Description	Supplementary information about the security group rule. This parameter is optional.	_

Step 4: Create an Enhanced Datasource Connection

- 1. Log in to the DLI management console. In the navigation pane on the left, choose **Datasource Connections**. On the displayed page, click **Create** in the **Enhanced** tab.
- 2. In the displayed dialog box, set the following parameters:
 - **Connection Name**: Name of the enhanced datasource connection
 - Resource Pool: Select the target DLI queue. (Queues that are not added to a resource pool are displayed in this list.)
 - VPC: VPC of the data source obtained in Step 1: Obtain the Floating IP Address, Port Number, and Security Group of an External Data Source
 - Subnet: Subnet of the data source obtained in Step 1: Obtain the Floating IP Address, Port Number, and Security Group of an External Data Source
 - Set other parameters as you need.
- 3. Click **OK**. Click the name of the created datasource connection to view its status. You can perform subsequent steps only after the connection status changes to **Active**.
- 4. To connect to MRS HBase, you need to add MRS host information. The procedure is as follows:
 - a. On the **Datasource Connections** page, click the **Enhanced** tab and locate the row that contains the created enhanced datasource connection. Click **More** > **Modify Host** in the **Operation** column.

Х

 In displayed dialog box, enter the MRS HBase host information obtained in Step 1: Obtain the Floating IP Address, Port Number, and Security Group of an External Data Source to the Host Information box.

Figure 4-3 Modifying host information

Modify Host	
Connection Name) ⁴⁴ 4 - 85
Host Information	Enter host information in the format "host IP address host name". Specify the information for each host on a separate line.
	OK Cancel

c. Click OK.

Step 5: Test Network Connectivity

- Choose Resources > Queue Management from the left navigation pane, locate the target queue. In the Operation column, click More > Test Address Connectivity.
- 2. In the displayed dialog box, enter the obtained IP address and port number of the data source in the address box, and click **Test**. If the queue passes the test, it can access the data source.

NOTE

For MRS HBase, use **ZooKeeper IP address:ZooKeeper port** or **ZooKeeper host** information:ZooKeeper port for the test.

4.2 Configuring the Connection Between a DLI Queue and a Data Source in the Internet

Scenario

This section provides instructions to enable network connectivity for DLI queues to be accessed from the Internet. You can configure SNAT rules and add routes to the public network to enable communications between a queue and the Internet.

Procedure

Figure 4-4 Configuration process



Step 1: Create a VPC

Log in to the VPC console and create a VPC. The created VPC is used for NAT to access the public network.

For details about how to create a VPC, see Creating a VPC.

Figure 4-5 Creating a VPC

Basic Information	
Region	·
	Regions are geographic areas isolated from each other. Resources are region-specific and cannot latency and quick resource access, select the nearest region.
Name	vpc-9334
IPv4 CIDR Block	10 · 0 · 0 · 0 / 8 ·
	Recommended:10.0.0.0/8-24 (Select) 172.16.0.0/12-24 (Select) 192.168.0.0/16-24 (Select)

Step 2: Create a Dedicated Queue

In this example, you will create a pay-per-use queue that uses dedicated resources.

The billing mode of the queue must be **Yearly/Monthly** or **Pay-per-use**. (If you select **Pay-per-use**, select **Dedicated Resource Mode** after you select a queue type.)

Enhanced datasource connections can be created only for yearly/monthly resources or pay-per-use resources in dedicated resource mode.

- 1. Log in to the DLI management console.
- Click **Buy Queue** in the upper left corner on the homepage page. On the displayed page, specify specifications and other required parameters.
 For details about the parameters for purchasing a queue, see Creating a Queue.

Step 3: Create an Enhanced Datasource Connection Between the Queue and a VPC

- 1. In the navigation pane of the DLI management console, choose **Datasource Connections**.
- 2. In the **Enhanced** tab, click **Create**.

Enter the connection name, select the created queue, VPC, and subnet, and enter the host information (optional).

Figure 4-6 Creating an enhanced datasource connection

Create Enhanced Connection

After you create the enhanced datasource connection, the system will automatically create a connection and required routes.

* Connection Name	dli_peer_0927	
Resource Pool		•
* VPC	vpc-9334(10.0.0/8)	•
* Subnet	subnet-9344(10.0.0/24)	▼

Step 4: Buy an EIP

- 1. Log in to the **EIPs** page of the network console, click **Buy EIP**.
- In the displayed page, configure the parameters as required.
 For details about how to set the parameters, see Buy EIP.

Step 5: Configure a NAT Gateway

Step 1 Create a NAT gateway.

- 1. Log in to the console and search for **NAT Gateway** in the Service List. The **Public NAT Gateways** page of the network console is displayed.
- 2. Click **Buy Public NAT Gateway** and configure the required parameters. For details, see **Buying a Public NAT Gateway**.

Figure 4-7 Buying a NAT gateway

★ Billing Mode	Yearly/Monthly	Pay-per-use		
★ Region		V	#	
	Regions are geographic ar cannot be used across reg and quick resource access	ions through internal net	work connections. I	
* Name	nat-32c8			
* VPC	vpc-9334	• C	View VPCs	
* Subnet	subnet-9344(10.0.0.0/24) • C	View Subnets Av	vailable private IP addresses: 25
	The selected subnet is for after the NAT gateway is c			ations over the Internet,
* Specifications	Small Med	lium Large	Extra-large	
	Supports up to 10,000 con	nections.Learn more		
Advanced Settings <	Description Tag			

3. Click Next, confirm the configurations, and click Submit.

NOTE

During the configuration, you need to set **VPC** to the one created in **Step 1: Create a VPC**.

Step 2 Add a route.

In the navigation pane on the left of the network console, choose Virtual **Private Cloud** > **Route Tables**. After a NAT gateway instance is created, a route to that gateway is automatically created. Click the route table name to view the automatically created route.

The destination address is the public IP address you want to access, and the next hop is the NAT gateway.

Figure 4-8 Viewing the route

Routes					
Delete	Add Route	Replicate Route	Q Learn how to cont	figure routes.	
Destin	nation (?)	Next	Нор Туре 🕐	Next Hop ⑦	Туре 🕐
∧ Local		Local		Local	System
Dest	tination		Next Hop Type		Next Hop
14.	38/32		NAT gateway		nat-32c8

Step 3 Add an SNAT rule.

You need to add SNAT rules for the new NAT gateway to allow the hosts in the subnet to communicate with the Internet.

- 1. Click the name of the created NAT gateway on the **Public NAT Gateways** page of the network console.
- 2. On the **SNAT Rules** tab, click **Add SNAT Rule**. For details, see **Adding an SNAT Rule**.
- 3. Scenario: Select Direct Connect/Cloud Connect.

- 4. **Subnet**: Select the subnet where the queue you want to connect locates.
- 5. **EIP**: Select the target EIP.

Figure 4-9 Adding an SNAT rule

Add SNAT Rule

It is not recommended t	F gateway are configured for nat an SNAT rule and a DNA are an EIP with a DNAT rule	rule share the same	EIP because there r			
Public NAT Gateway Name	nat-lishenrui					
* Scenario	VPC	Direct Conne	ct/Cloud Connect			
	172 · 16 · 0	. 0 / 16	?			
* EIP	You can select more	EIPs. 🥐 View Ell	Specify filter c	riteria.		Q
	EIP	EIP Type	Bandwidth Na	Bandwidth(M	Billing Mode	Enterprise Pr

6. Click OK.

----End

Step 6: Adding a Custom Route

Add a custom route for the enhanced datasource connection you have created. Specify the route information of the IP address you want to access.

For details, see **Custom Route Information**.

Figure 4-10 Adding route information for test

Add Route		×
* Route Name		
★ IP Address	14 . 0 24	
	OK Cancel	

Step 7: Testing the Connectivity to the Public Network

Test the connectivity between the queue and the public network. Click **More** > **Test Address Connectivity** in the **Operation** column of the target queue and enter the public IP address you want to access.

Figure 4-11	Testing	address	connectivity
-------------	---------	---------	--------------

Test Address Connectivity

 \times

Tests whether an address is reachable from a specified cluster. The address can be a domain name, an IP address, or a specified port.

* Address	114.	:80		
		Test	Cancel	

A Change History

Released On	What's New
2023-10-09	 Modified the following sections: Added the regions that support the Trino engine to Interconnecting FineBI with DLI Trino and Interconnecting Power BI with DLI Trino.
2023-08-19	Added the following section: Using DLI Flink SQL to Analyze e-Commerce Business Data in Real Time
2023-07-21	Added the following sections: • Interconnecting FineBI with DLI Trino • Interconnecting Power BI with DLI Trino
2023-03-09	Adjusted the document structure and moved the content related to DLI data development to <i>Data Lake Insight Development Guide</i> .
2023-03-02	Modified the prerequisites in Migrating Data from MRS Kafka to DLI.
2023-01-18	Added the description of configuring migration job scenarios to Migrating Data from MRS Kafka to DLI.
2023-01-06	 Optimized the procedure for adding an inbound rule to the security group of the external data source to allow access from the DLI queue in Configuring the Connection Between a DLI Queue and a Data Source in a Private Network.
2022-10-31	Optimized the following sections and added information about solution advantages, process guidance, and resource planning and costs. • Analyzing Driving Behavior Data • Converting Data Format from CSV to Parquet

Released On	What's New
2022-09-27	Added Configuring the Connection Between a DLI Queue and a Data Source in the Internet.