Data Lake Insight

Best Practice

Issue 01

Date 2021-04-01



HUAWEI TECHNOLOGIES CO., LTD.

Copyright © Huawei Technologies Co., Ltd. 2021. All rights reserved.

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

Trademarks and Permissions


and other Huawei trademarks are trademarks of Huawei Technologies Co., Ltd.
All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

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

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

Huawei Technologies Co., Ltd.

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

Website: https://www.huawei.com

Email: support@huawei.com

Contents

1 Driving Behavior Data Analysis in IoV Scenarios.....1
2 Consulting Service Data Analysis in Financial Industry.....9
3 Using DLI to Convert Data Format from CSV to Parquet.....21
4 Using DLI to Analyze E-commerce BI Reports.....24

 Copyright © Huawei Technologies Co., Ltd.

1 Driving Behavior Data Analysis in IoV Scenarios

Application Scenarios

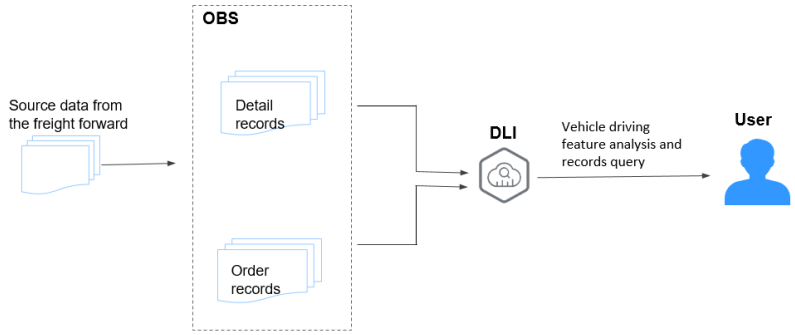
- Description

  In 2016, there were over 300 million vehicles around China, including over 14 million trucks, over 30 million commercial vehicles, and over one million fleets of different scales. The vehicles are under extensive management.

  With the support of the IoV technology, enterprises and fleet managers can use the powerful analysis and mining capabilities of cloud computing and big data to manage fleets more scientifically and conveniently.

- Solutions

  Figure 1-1 Solutions

- Tasks

  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
  - Detail records

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

 Copyright © Huawei Technologies Co., Ltd.

Table 1-1 Detail records

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

 Copyright © Huawei Technologies Co., Ltd.

- Order data

  Order data refers to the records of freight orders.

  Table 1-2 Order data

Field	Field Type	Field 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.

Process

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

Step 1: Registering an Account. Before the data analysis, register a public cloud account and perform real name authentication.

Step 2: Uploading Data. Upload the data to OBS.

Step 3: Analyzing Data. Use DLI to query the data.

Step 1: Registering an Account

Register a public cloud account and perform real name authentication.

For details about how to register an account, see **Registering a HUAWEI ID and Enabling HUAWEI CLOUD Services**.

For details about real-name authentication, see **Real-Name Authentication Entry**.

If you have registered a public cloud account and completed real-name authentication, skip this step.

 Copyright © Huawei Technologies Co., Ltd.

Step 2: Uploading Data

Upload the data to OBS for data analysis using DLI.

1. Download OBS Browser+. For details about the download address, see the **Object Storage Service Tools Guide**.
2. Install OBS Browser+. For details about the installation process, see the **Object Storage Service Tools 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 process, see the **Object Storage Service Tools Guide**.
4. Upload data using the OBS browser+.

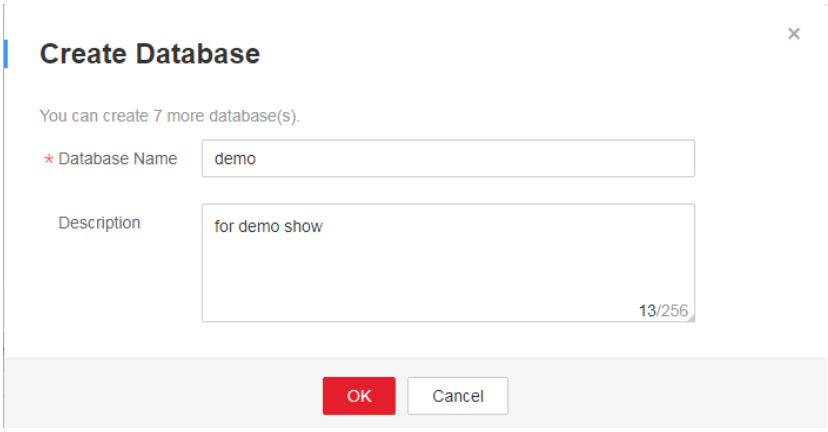
   On the OBS Browser+ page, click **Create Bucket**. Select a region and enter a bucket name (for example, **DLI_demo**). After the bucket is created, return to the bucket list and click **DLI_demo**. OBS Browser+ supports upload by dragging. You can drag one or more files or folders from a local path 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+.

Step 3: Analyzing Data

Use DLI to query the data for analysis.

1. Creating a Database and a Table
   a. On the homepage of the HUAWEI CLOUD management console, choose **Service List** > **EI Enterprise Intelligence** > **Data Lake Insight**.
   b. Create a database named **demo**. On the DLI console, click **SQL Jobs** to go to the **SQL Editor** page.
   c. In the left pane of the SQL Editor, select  and click  to create a database. For details, see **Figure 1-2**.

Figure 1-2 Creating a database



NOTE

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

 Copyright © Huawei Technologies Co., Ltd.

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

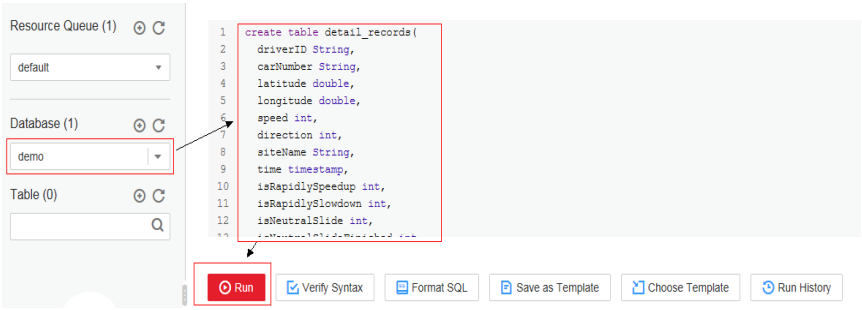
```
create table detail_records(
  driverID String,
  carNumber String,
  latitude double,
  longitude double,
  speed int,
  direction int,
  siteName String,
  time timestamp,
  isRapidlySpeedup int,
  isRapidlySlowdown int,
  isNeutralSlide int,
  isNeutralSlideFinished int,
  neutralSlideTime long,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

NOTE

Set the file path in the preceding statement to the actual OBS path where the detail records are stored.

e. Click **Execute** to create the **detail_records** table. See **Figure 1-3**.

Figure 1-3 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**.

```
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)
```

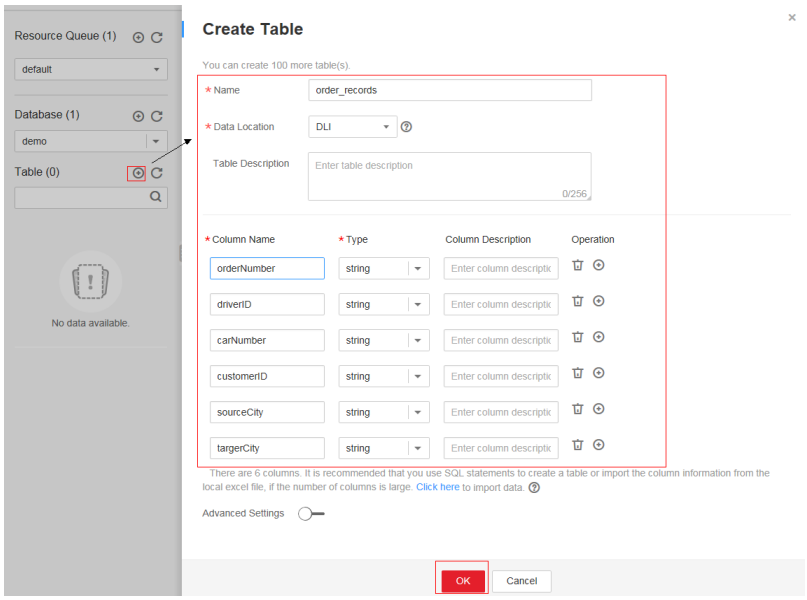
 Copyright © Huawei Technologies Co., Ltd.

g. Run the following SQL statement 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)
```

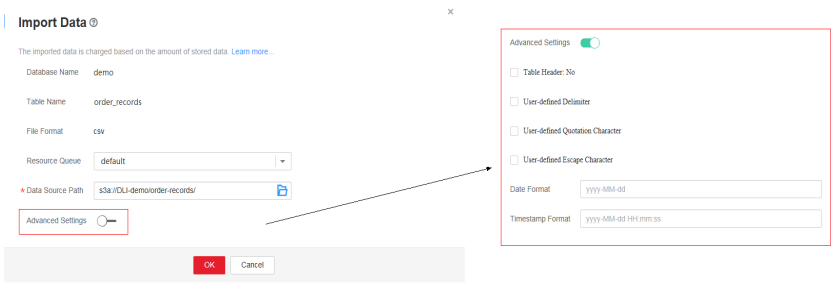
h. Click **+** next to **Table** to create the **order_records** table, for which **Data Location** is set to **DLI**. See **Figure 1-4**. Set the column types according to **Order data**.

Figure 1-4 Creating the **order_records** table



i. Import OBS data to the **order_records** table. Click **Data Management**, and unfold the **demo** database. Click **Table Management**, find the **order_records** table, and click **Operation** < **More** < **Import**. Set **Data Source Path** to **obs://DLI-demo/order-records/**. The default timestamp format is **yyyy-MM-dd HH:mm:ss**. To use other formats, select **Advanced Settings** and manually enter the desired format. Click **OK**. See **Figure 1-5**.

Figure 1-5 Importing table data



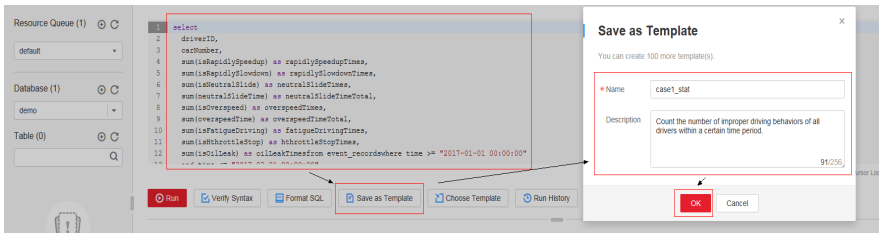
 Copyright © Huawei Technologies Co., Ltd.

2. Querying Data

   a. Run the following SQL statement to query the alarm events of all drivers in a certain time period. You can save the frequently-used query statements as templates that are available in the **Template Management** page for later use. For details, see **Figure 1-6**. Click **Execute** to switch to the **SQL Editor** page. You can query data easily by modifying the query statements.

```
select
  driverID,
  carNumber,
  sum(isRapidlySpeedup) as rapidlySpeedupTimes,
  sum(isRapidlySlowdown) as rapidlySlowdownTimes,
  sum(isNeutralSlide) as neutralSlideTimes,
  sum(neutralSlideTime) as neutralSlideTimeTotal,
  sum(isOverspeed) as overspeedTimes,
  sum(overspeedTime) as overspeedTimeTotal,
  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
```

Figure 1-6 Saving the query as a template



   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"
```

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

```
select
  *
from
  order_records
where
```

 Copyright © Huawei Technologies Co., Ltd.

```
  orderNumber = "2017013013584419488"
order by
  time desc
```

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 = "zouan1000007"
  and time > "2017-01-30 16:00:00"
  and siteName IS NOT NULL
order by
  time desc
```

 Copyright © Huawei Technologies Co., Ltd.

2 Consulting Service Data Analysis in Financial Industry

Application Scenarios

- Description

  Company *H* is the first commercial organization in China engaged in gathering trade statistics and buyer information of major trading participants worldwide. It has the world's largest trade statistics database for industry research, international trade promotion, and other uses.

  Before using DLI, company *H* used its own big data queue. It had to hire O&M personnel, purchase the dedicated bandwidth from China Telecom and China Unicom, and invest heavily in equipment room, electric power, private networks, and servers. Even with the huge input, the company cannot satisfy customers' ever-changing service requirements due to insufficient manpower and capability restrictions of its big data queue. As a result, only 4% of 100 TB data is actually used.

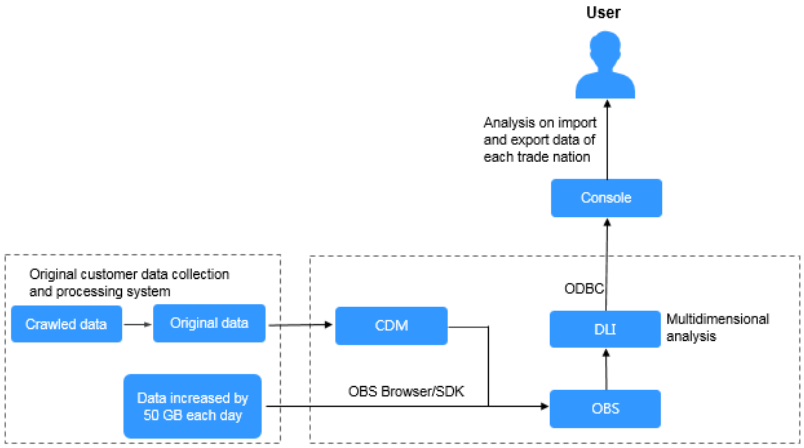
  After migrating to the cloud and adopting DLI as its data analysis tool, company *H* made full use of the 100 TB database to maximize asset monetization, without the need to construct and maintain infrastructures. Using the pay-per-use billing mode, DLI requires no O&M personnel and lower dedicated bandwidth cost, which in total saves the maintenance cost by 70% compared with the offline data center. In addition, DLI requires no expertise and enables smooth migration of existing services, shortening the service rollout period by 50%.

  DLI helps company *H* build its big data analysis platform with extremely low cost, so that it can focus on constant innovation on business needs.

- Solutions

 Copyright © Huawei Technologies Co., Ltd.

Figure 2-1 Solutions



Process the existing data (for example: trade details and basic information) of company *H* in the system based on the original customer data. Then, complete trade statistics analysis using CDM, OBS, and DLI.

- Data Types
  - Trade detail records

    Trade detail records include trade statistics of major trading nations around the world.

    Table 2-1 Trade detail records

Field	Field Type	Field Description
m_month	smallint	Month
hs_code	string	Product code
country	smallint	Country code
dollar_value	double	Transaction amount
quantity	double	Transaction volume
unit	smallint	Measurement unit
b_country	smallint	Destination country code
imex	smallint	Import or export
y_year	smallint	Year

  - Basic information

    The basic information includes dictionary data corresponding to the fields in the trade detail records.

 Copyright © Huawei Technologies Co., Ltd.

Table 2-2 Basic country information

Field	Field Type	Field Description
countryid	smallint	Country code
country_en	string	English name of a country
country_cn	string	Chinese name of a country

Table 2-3 Update time information

Field	Field Type	Field Description
countryid	smallint	Country code
imex	smallint	Import or export
hs_len	smallint	Product code length
minstartdate	string	Minimum start time
startdate	string	Start time
newdate	string	Update time
minnewdate	string	Last update time

Table 2-4 Import and export product code information table (hs246)

Field	Field Type	Field Description
id	bigint	ID
hs	string	Product code
hs_cn	string	Chinese name of a product
hs_en	string	English name of a product

Table 2-5 Unit information (unit_general)

Field	Field Type	Field Description
id	smallint	Measurement unit code

 Copyright © Huawei Technologies Co., Ltd.

Field	Field Type	Field Description
unit_en	string	English name of the measurement unit
unit_cn	string	Chinese name of the measurement unit

Analysis Process

To use DLI to analyze trade statistics data, you need to perform the following three steps:

Step 1: Registering an Account. Before the data analysis, register a public cloud account and perform real name authentication.

Step 2: Uploading Data. Upload incremental data every day using OBS Browser+ and use CDM to migrate inventory data from OBS to DLI to prepare for data analysis using DLI.

Step 3: Analyzing Data. Interconnect the customer service system with DLI to enable trade statistics analysis for company *H*.

Step 1: Registering an Account

Register a public cloud account and perform real name authentication.

For details about how to register an account, see **Registering a HUAWEI ID and Enabling HUAWEI CLOUD Services**.

For details about real-name authentication, see **Real-Name Authentication Entry**.

If you have registered a public cloud account and completed real-name authentication, skip this step.

Step 2: Uploading Data

- Upload Incremental Data

  Upload the incremental data to OBS.

  a. Download OBS Browser+. For details about the download address, see the **Object Storage Service Tools Guide**.

  b. Install OBS Browser+. For details about the installation process, see the **Object Storage Service Tools Guide**.

  c. 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 process, see the **Object Storage Service Tools Guide**.

  d. 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

 Copyright © Huawei Technologies Co., Ltd.

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+.

- Upload Inventory Data

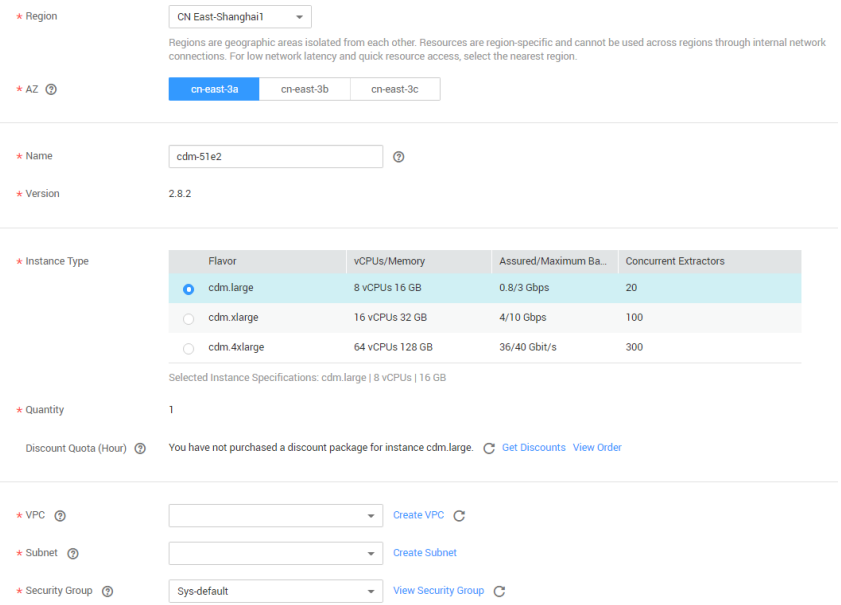
  Cloud Data Migration (CDM) enables data migration among various data sources, including services of public cloud or between public cloud and on-premises service systems. You can use CDM to upload the inventory data of company *H* stored on OBS to DLI through table or file migration for data analysis.

  a. By default, data is stored on OBS. For details about how to upload data to OBS, see **Incremental Data Upload**.

  b. Create a resource queue, a database, and a table in DLI. For details, see the **Data Lake Insight User Guide**.

  c. Log in to the CDM management console and click **Buy CDM** in the upper right corner to create a CDM cluster, as shown in **Figure 2-2**.

Figure 2-2 Creating a CDM cluster



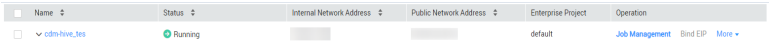
NOTE

The VPC chosen must be the same on both CDM and DLI. You are advised to retain the same **Subnet** and **Security Group** settings on both CDM and DLI.

If the **Subnet** and **Security Group** settings cannot be the same due to security control, ensure that the security group rule is configured to allow CDM to access DLI databases.

  d. After a CDM cluster is created, click **Job Management** under **Operation**. See **Figure 2-3**.

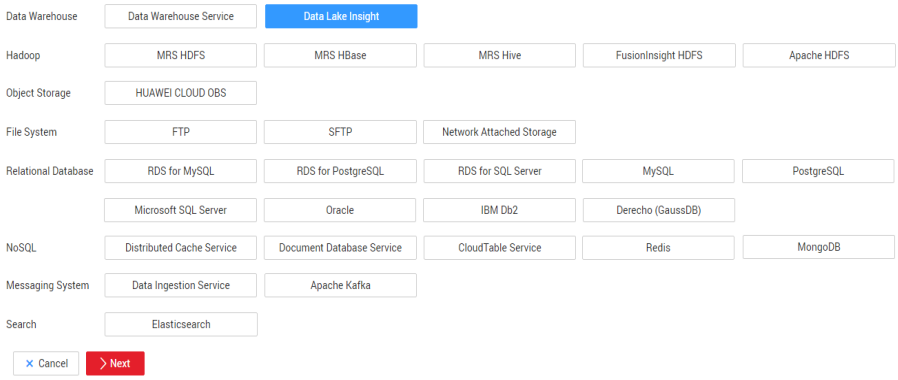
Figure 2-3 Job management



Copyright © Huawei Technologies Co., Ltd.

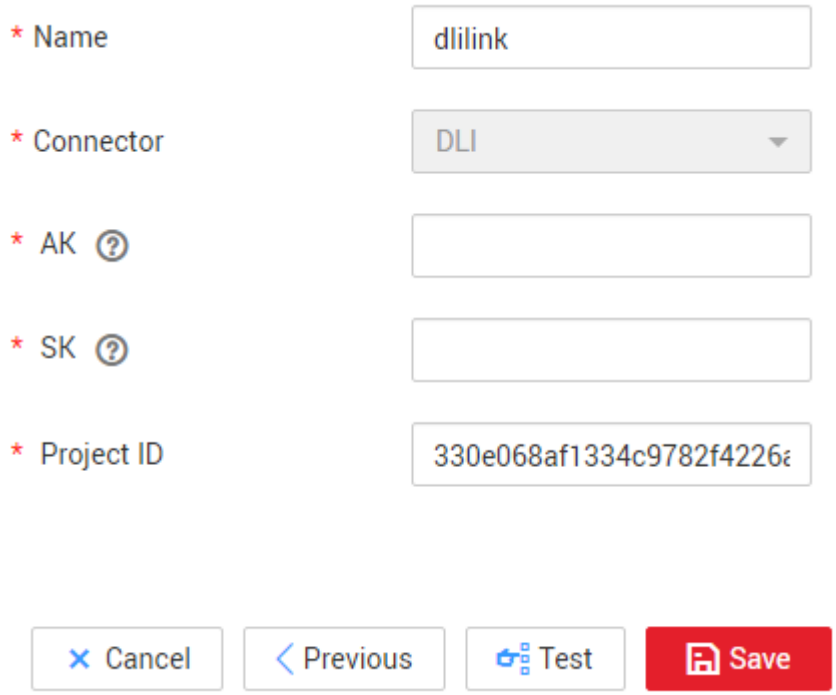
e. On the **Job Management** page, choose **Link Management > Create Link**. See **Figure 2-4**.

Figure 2-4 Selecting a link type



f. Create a DLI link. Specifically, select **Data Lake Insight** for **Data Warehouse** and click **Next**. Specify DLI link parameters as required. See **Figure 2-5**.

Figure 2-5 Configuring the DLI link



 Copyright © Huawei Technologies Co., Ltd.

Table 2-6 Parameters used for linking CDM to DLI

Parameter	Description	Example Value
Name	Link name, which can be defined based on the data source type for easy memorization.	dli_link
Connector	Data source to be connected, which is the data type selected in the previous step.	DLI
AK	AK required for authentication during accessing the DLI database. To obtain the AK and SK, log in to the management console, click the username in the upper right corner, select **My Credentials** from the drop-down list, and click **Access Keys** on the displayed page.	-
SK	SK required for authentication during accessing the DLI database. To obtain the SK, log in to the management console, click the username in the upper right corner, select **My Credentials** from the drop-down list, and click **Access Keys** on the displayed page.	-
Project ID	Project ID of the region where DLI resides. Obtain the value from **My Credentials**.	a46ed0f02bde42e7afe36777eb9d0f42

g. Click **Save**. CDM automatically checks whether the link is available. If the link is available, a message is displayed, indicating that the link is successfully saved, and the link management page is displayed. If the link

 Copyright © Huawei Technologies Co., Ltd.

is unavailable, check whether the link parameters are correctly configured.

h. Click **New Link** to create an OBS link. Specifically, select **Object Storage Service** for **Data Warehouse** and click **Next**. Set the OBS link parameters as required. See **Figure 2-6**.

Figure 2-6 Creating an OBS link

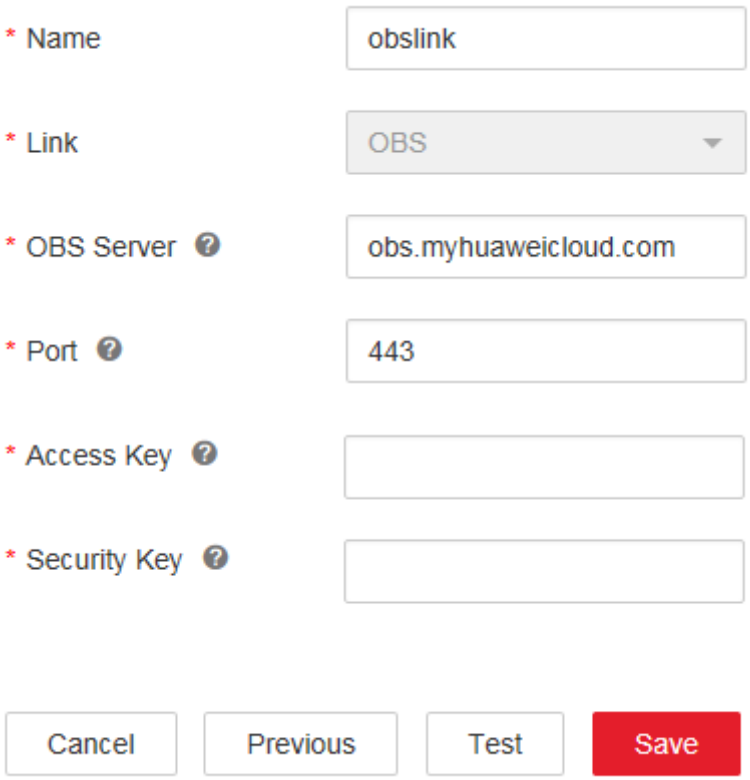


Table 2-7 Parameters used for linking CDM to OBS

Parameter	Description	Example Value
Name	Link name, which can be defined based on the data source type for easy memorization.	obslink
OBS Server	Set this parameter to the domain name or IP address of OBS.	obs.cn-north-1.myhuaweicloud.com
Port	Port number of the OBS server	443

Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
AK	Access Key ID required for authentication during accessing the DLI database. To obtain the AK, log in to the management console, click the username in the upper right corner, select **My Credentials** from the drop-down list, and click **Access Keys** on the displayed page.	-
SK	Secret Key required for authentication during accessing the DLI database. To obtain the SK, log in to the management console, click the username in the upper right corner, select **My Credentials** from the drop-down list, and click **Access Keys** on the displayed page.	-

i. Click **Save**. CDM automatically checks whether the link is available. If the link is available, a message is displayed, indicating that the link is successfully saved, and the link management page is displayed. If the link is unavailable, check whether the link parameters are correctly configured.

j. Choose **Table/File Migration** > **Create Job** to create a task for migrating OBS data to DLI. See **Figure 2-7**.

Copyright © Huawei Technologies Co., Ltd.

Figure 2-7 Creating a job

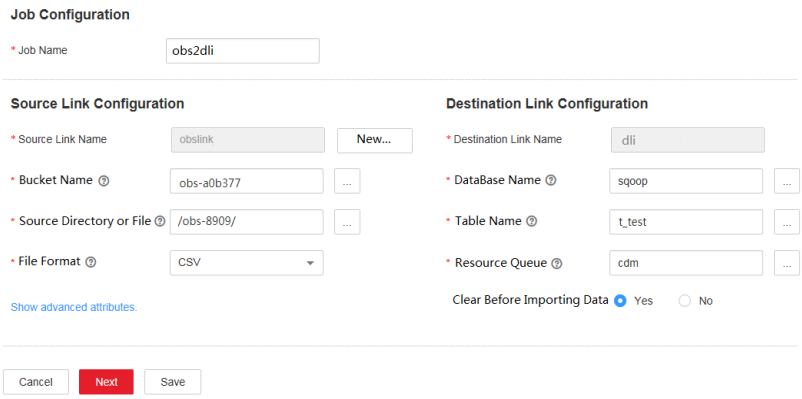


Table 2-8 Job configuration parameters

Parameter	Description	Example Value
Job Name	User-defined job name.	obs2dli

Table 2-9 Parameters for configuring the OBS as the source connection

Parameter	Description	Example Value
Source link name	Select the created OBS connection.	obslink
Bucket Name	Name of the bucket from which data is to be migrated	-
Source Directory/File	Path of the data to be migrated.	-
File format	File format used for data transmission. The value can be **CSV** or **JSON**.	CSV

Table 2-10 Parameters for configuring DLI as the destination connection

Parameter	Description	Example Value
Destination connection name	Select the created DLI connection.	dlilink
Database Name	DLI service database to which data is to be migrated.	sqoop

 Copyright © Huawei Technologies Co., Ltd.

Parameter	Description	Example Value
Table Name	Name of the DLI table to which data is to be migrated. It is recommended that the field type and format of the table be consistent with those of the data to be migrated.	car_detail
Resource Queue	Queue to which data is to be migrated.	cdm
Clear Data Before Import	Before importing data, determine whether to clear the original data in the destination queue.	No

k. After basic job information is configured, click **Next** to verify whether the field mapping is correct.

Specifically, check the following items:

- Whether the source field type can be converted to the destination field type
- Whether the mapping between the source and destination fields is correct. If the mapping is incorrect, you can click the arrows in the **Operation** column to adjust the sequence of the destination fields.
- Whether the date and time format is correct. If the format is incorrect, modify the format or select an existing time format.

NOTE

Users can customize field content conversion.

l. Click **Next** to set task parameters. See **Figure 2-8**.

Copyright © Huawei Technologies Co., Ltd.

Figure 2-8 Task configuration

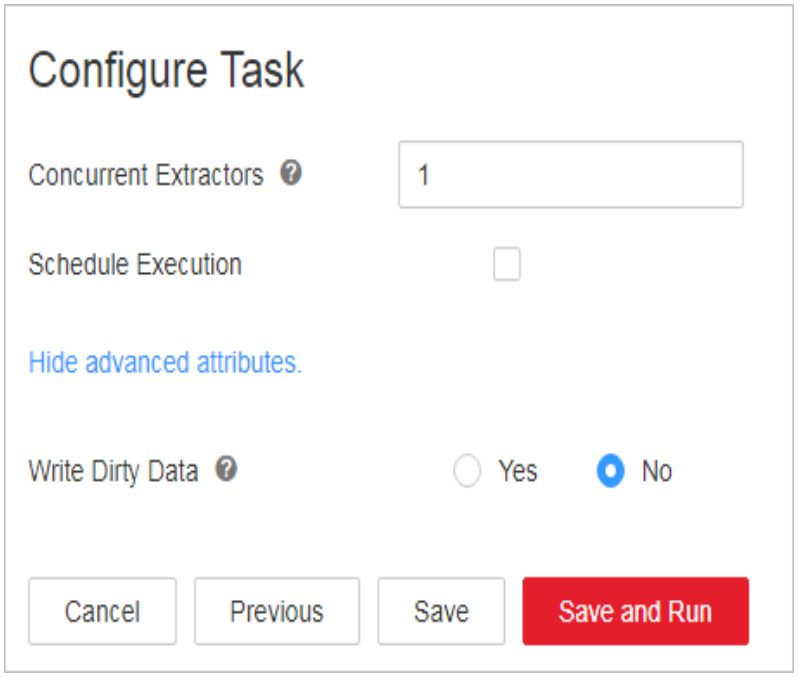


Table 2-11 Task configuration parameters

Parameter	Description	Example Value
Concurrent Extractors	Indicates the number of concurrent extraction jobs that are performed concurrently.	1
Schedule Execution	If you select **Yes**, you can customize a scheduled task.	No
Write Dirty Data	Indicates whether to record dirty data.	No

m. After the job is created, click **Save and Run**. The **Job Management** page is displayed. You can view the job execution progress and result.

Step 3: Analyzing Data

Query data in the corresponding table. For details, see **Step 3: Analyzing Data**.

 Copyright © Huawei Technologies Co., Ltd.

3 Using DLI to Convert Data Format from CSV to Parquet

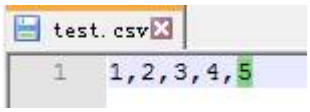
Application Scenarios

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 very efficient compression schemes, which maximizes the storage efficiency on disks. Using DLI, you can easily convert data format form CSV to Parquet.

Operation Process

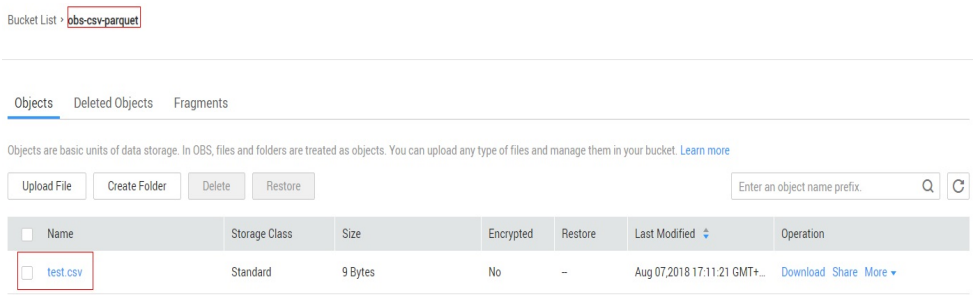
Step 1 Create a CSV file. See **test.csv** in **Figure 3-1**.

Figure 3-1 Creating a **test.csv** file



Step 2 In the OBS management console, create a bucket named **obs-csv-parquet** and upload the **test.csv** file to OBS. See **Figure 3-2**.

Figure 3-2 Uploading CSV data to OBS



Step 3 In the SQL editing box of the DLI management console, create an OBS table named **test_csv_hw** and import data in the **test.csv** file to **test_csv_hw**.

```
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')
```

 Copyright © Huawei Technologies Co., Ltd.

Step 4 In the SQL editing box of the DLI management console, query data in the **test_csv_hw** table.

Figure 3-3 Querying data in the **test_csv_hw** table



Step 5 Create a bucket named **obs-parquet-data** on OBS to store Parquet data.

Step 6 In the SQL job editing window of the DLI management console, create an OBS table named **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/'
)
```

NOTE

You do not need to specify a file because no Parquet file exists before the data is converted from the CSV format to the Parquet format.

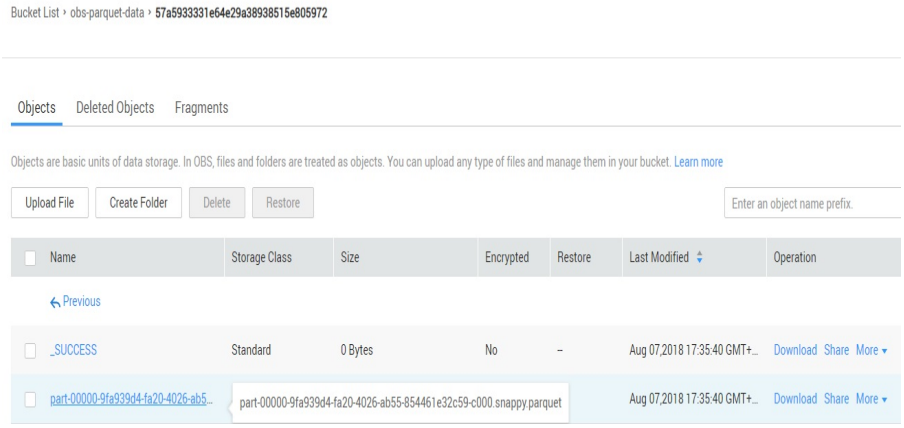
Step 7 In the SQL job editing window of the DLI management console, convert CSV data into Parquet data and store the data in OBS.

insert into test_parquet_hw select * from test_csv_hw

Step 8 Check the result. As shown in **Figure 3-4**, the system automatically creates a file for saving the result.

 Copyright © Huawei Technologies Co., Ltd.

Figure 3-4 Saving Parquet data



----End

 Copyright © Huawei Technologies Co., Ltd.

4 Using DLI to Analyze E-commerce BI Reports

Application Scenarios

- Description

  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 the public cloud DLI, DWS, and Yonghong BI to analyze data features of users and offerings based on the real user, offering, 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.

- Data Types

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

  – User data

  Table 4-1 User data

Field Name	Field Type	Description	Value Range
user_id	int	User ID	Anonymized
age	int	Age group	**-1** indicates that the user age is unknown.
sexuality	int	Gender	• 0: Male • 1: Female • 2: Confidential

 Copyright © Huawei Technologies Co., Ltd.

Field Name	Field Type	Description	Value Range
rank	Int	User level	Sequenced list of user level. The higher the user level, the larger the number.
register_time	string	User registration date	Unit: day

- Product data

Table 4-2 Product data

Field Name	Field Type	Description	Value Range
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

- Evaluation data

Table 4-3 Evaluation data

Field Name	Field Type	Description	Value Range
deadline	string	End time	Unit: day
product_id	int	Product No.	Anonymized

 Copyright © Huawei Technologies Co., Ltd.

Field Name	Field Type	Description	Value Range
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_comment	int	Whether there is negative feedback.	0: No; 1: Yes.
bad_comment_rate	float	Dissatisfaction rate	Proportion of the negative feedback.

- Behavior data

Table 4-4 Behavior data

Field Name	Field Type	Description	Value Range
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	-

Analysis Process

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

Step 1: Registering an Account. Before the data analysis, register a public cloud account and perform real name authentication.

Step 2: Uploading Data. Upload the data to OBS for data analysis using DLI.

 Copyright © Huawei Technologies Co., Ltd.

Step 3: Analyzing Data. Use DLI to query the data for analysis.

Step 1: Registering an Account

Register a public cloud account and perform real name authentication.

For details about how to register an account, see **Registering a HUAWEI ID and Enabling HUAWEI CLOUD Services**.

For details about real-name authentication, see **Real-Name Authentication Entry**.

If you have registered a public cloud account and completed real-name authentication, skip this step.

Step 2: Uploading Data

Upload the data to OBS for data analysis using DLI.

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

NOTE

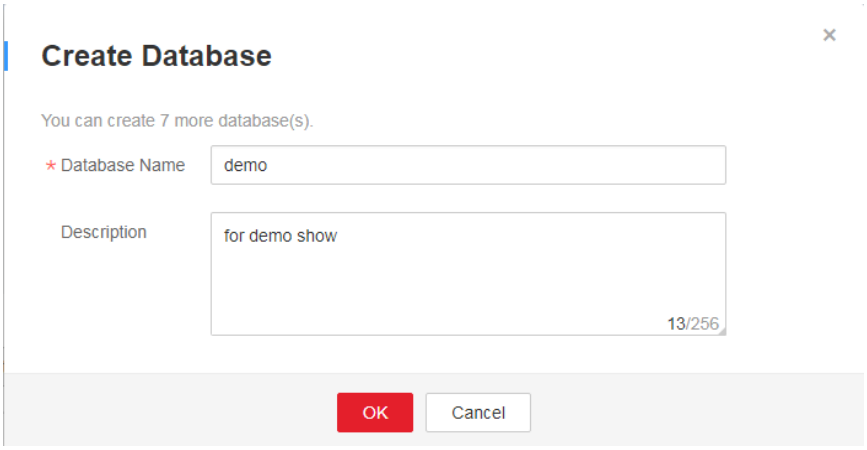
Data has been uploaded to OBS. The procedure for uploading data can be omitted.

Step 3: Analyzing Data

1. Creating a Database and a Table
   a. On the homepage of the HUAWEI CLOUD management console, choose **Service List** > **EI Enterprise Intelligence** > **Data Lake Insight**.
   b. Create a demo database. On the DLI console, click **SQL Jobs** to go to the SQL Editor page.
   c. In the left pane of the SQL Editor, select **Database** and click to create a database. For details, see **Figure 4-1**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 4-1 Creating a database



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,
  sexuality int,
  rank int,
  register_time string
) USING csv OPTIONS (path "obs://yonghong/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 offering table, evaluation table, and behavior table in the same way.

- Product table

```
create table product(
  product_id int,
  a1 int,
  a2 int,
  a3 int,
  category int,
  brand int
) USING csv OPTIONS (path "obs://yonghong/data/JData_Product")
```

- Evaluation 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://yonghong/data/JData_Comment")
```

- Behavior table

```
create table action(
  user_id int,
  product_id int,
```

 Copyright © Huawei Technologies Co., Ltd.

```
  time string,
  model_id string,
  type string
) USING csv OPTIONS (path "obs://yonghong/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 the *Data Lake Insight User Guide*.

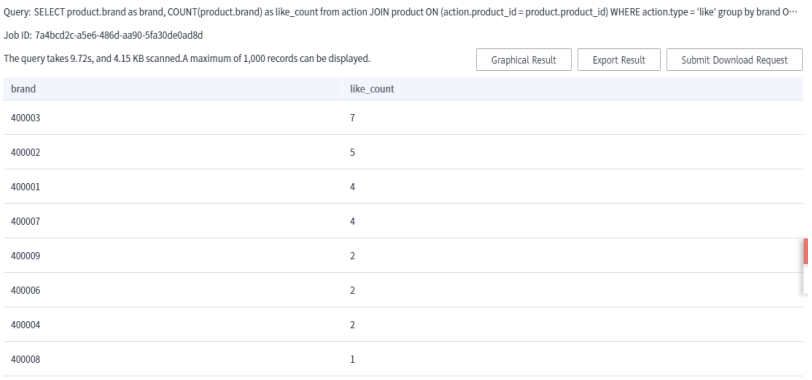
   – Analyze the top 10 products with the most likes.

     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 4-2**.

Figure 4-2 Querying results



     iii. Click **Graphical Result** to display the result in a chart.

Figure 4-3 Graphical results



   – Analyze the top 10 worst-rated products.

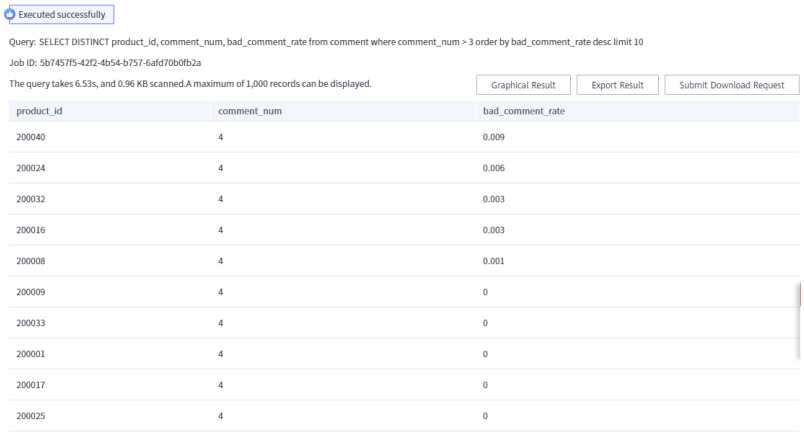
 Copyright © Huawei Technologies Co., Ltd.

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
```

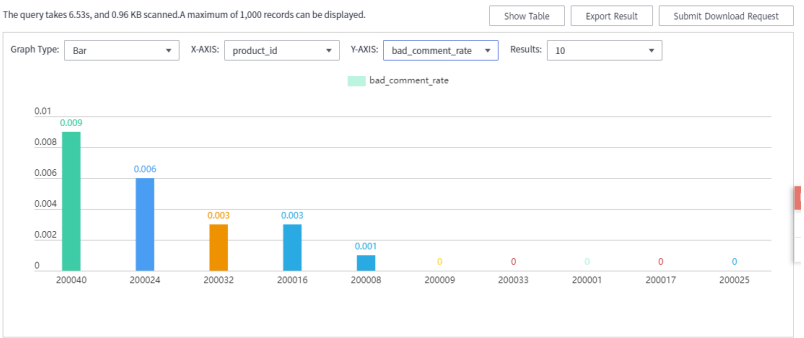
ii. Click **Execute**. The execution results are displayed, as shown in **Figure 4-4**.

Figure 4-4 Querying results



iii. Click **Graphical Result** to display the result in a chart.

Figure 4-5 Graphical result



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

Click the following link to download the sample data and detailed SQL statements:

https://dli-demo-ce3.obs.cn-east-3.myhuaweicloud.com:443/Best_Practice_4/Best_Practice_4.zip

 Copyright © Huawei Technologies Co., Ltd.