Data Ingestion Service

FAQs

Issue	01
Date	2024-10-25

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 General Questions

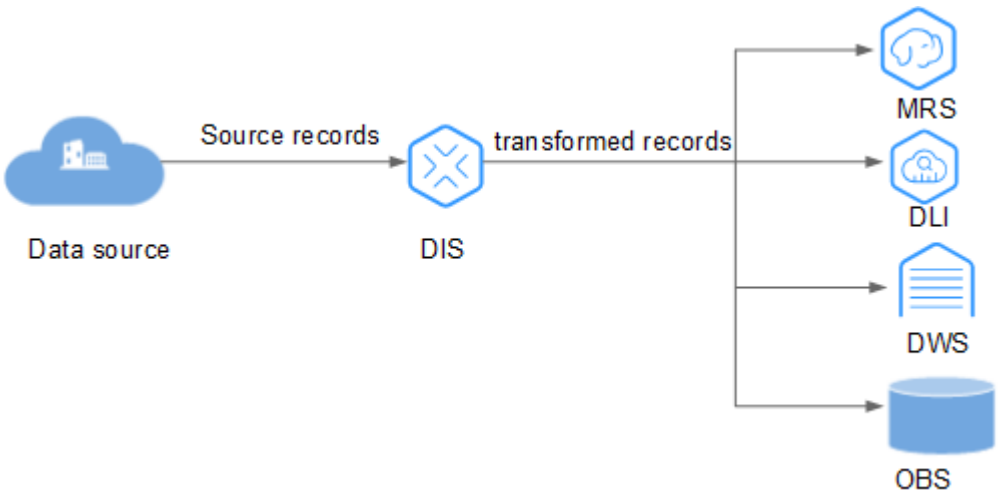
1.1 What Is DIS?

Data Ingestion Service (DIS) addresses the challenge of transmitting data from outside the cloud to inside the cloud. DIS builds data intake streams for custom applications capable of processing or analyzing streaming data. DIS continuously captures, transmits, and stores terabytes of data from hundreds of thousands of sources every hour, such as logs, Internet of Things (IoT) data, social media feeds, website clickstreams, and location-tracking events.

Data Flows

- DIS collects data from multiple data sources in real time.
- DIS transforms data continuously to MRS, DLI, DWS, and OBS for computing, analysis, and storage.

Figure 1-1 Data flows



1.2 What Is a Partition?

Partitions are the base throughput unit of a DIS stream. When creating a DIS stream, you are expected to specify the number of partitions needed within your stream.

By default, each tenant has 1 to 50 partitions. Tenants can configure the number of partitions as required.

To increase the quota, submit a **service ticket**. The upper limit is calculated based on the actual load of the cluster.

1.3 What Can I Do with DIS?

You can use DIS for rapid data intake from producers and continuous data processing. The following are typical scenarios for using DIS:

- Accelerated log or data transmission: Users do not need to wait for batch data processing. Data is put into a DIS stream immediately after it is generated by a data producer, preventing data loss in case of faults in the data producer. For example, system and application logs can be continuously put into a stream and processed within seconds.
- Real-time metrics and reporting: You can retrieve data from DIS streams for simple data analysis and reporting in real time. For example, your DIS applications can work on metrics and reporting for system and application logs as streaming data is being pushed to DIS streams using application programming interfaces (APIs), rather than wait to receive batches of data.
- Real-time data analysis: DIS combines the power of parallel processing with the value of real-time data. For example, you can transform data into valuable information and business intelligence by simply putting data into a DIS stream. This can be done in minutes instead of hours or days.
- Complex stream processing: You can create Directed Acyclic Graphs (DAGs) of DIS applications and streams. This typically involves putting data from one or multiple DIS applications into another stream for downstream processing by a different DIS application.

1.4 What Advantages Does DIS Have?

- Scalable: A DIS stream can seamlessly scale its data throughput from megabytes to terabytes per hour, and from thousands to millions of PUT records per second.
- Easy to use: You can create a DIS stream within seconds. It is easy to put data into a stream, and build a data processing application.
- Low cost: DIS has no upfront cost and you only pay for the resources you use.
- Parallel processing: DIS allows you to have multiple DIS applications processing the same stream concurrently. For example, you can have one application running real-time analytics and another sending data to OBS from the same stream.

- DIS preserves data for 24 hours, reducing the probability of data loss in case of application failures, individual machine failures, or facility failures. The value of *N* is an integer from 1 to 7.

1.5 Which Modules Do DIS Have?

- Service control
  - Creates, deletes, and configures DIS streams; synchronizes user information to the data plane
  - Allocates resources on the data plane and automatically deploys DIS
- Data processing
  - Receives user requests; receives and stores authenticated data
  - Receives data read requests and returns the requested data to authorized users
  - Removes old data from DIS streams according to data aging policies
  - Stores user data into Object Storage Service (OBS) according to user options
- Service maintenance
  - Installs and upgrades DIS
  - Performs configuration, preventive maintenance, monitoring, and log collection and analysis for DIS
  - Processes service orders
- User SDK
  - Provides Java APIs for users to push and pull data
  - Encrypts data

1.6 How Do I Create a DIS Stream?

For details, see the *Data Ingestion Service User Guide*.

1.7 What Is the Difference Between Storing Data into DIS and Dumping Data Elsewhere?

You need to select **Dump Destination** when enabling the DIS stream. **Table 1-1** describes the specific differences.

- If you set **Dump Destination** to **OBS**, data from the stream will be stored in DIS and then periodically dumped to Object Storage Service (OBS).
- If you set **Dump Destination** to **MRS**, data from the stream will be stored in DIS and then periodically imported into the Hadoop Distributed File System (HDFS) of a MapReduce Service (MRS) cluster.
- If you set **Dump Destination** to **DLI**, data from the stream will be stored in DIS and then periodically dumped to DLI.
- If you set **Dump Destination** to **DWS**, data from the stream will be stored in DIS and then periodically imported to DWS.

- If you set **Dump Destination** to **CloudTable**, data from the stream will be stored in DIS and then imported into the HBase or OpenTSDB table of a CloudTable cluster in real time.

Table 1-1 Difference between storing data into DIS and dumping data elsewhere

Storing Data into DIS	Dumping Data to OBS	Dumping Data to MRS	Dumping Data to DLI	Dumping Data to DWS	Dumping Data to CloudTable
You can store data into DIS without applying for storage resources.	You must apply for OBS resources before dumping data to OBS.	You must apply for MRS resources before dumping data to MRS.	You must apply for DLI resources before dumping data to DLI.	You must apply for DWS resources before dumping data to DWS.	You must apply for CloudTable resources before dumping data to CloudTable.
Data storing is free of charge.	Additional cost for the use of OBS. For details, see the OBS price details.	Additional cost for the use of OBS and MRS. For details, see the OBS and MRS price details.	Additional cost for the use of OBS and DLI. For details, see the OBS and DLI price details.	Additional cost for the use of OBS and DWS. For details, see the OBS and DWS price details.	Additional cost for the use of CloudTable. For details, see the CloudTable price details.
Data is temporarily stored in DIS for up to 168 hours.	Data is stored in OBS until your OBS bucket expires.	Data is stored in MRS until your MRS cluster expires.	Data is stored in DLI until your DLI service expires.	Data is stored in DWS until your DWS service expires.	Data is stored in CloudTable until your CloudTable service expires.

Storing Data into DIS	Dumping Data to OBS	Dumping Data to MRS	Dumping Data to DLI	Dumping Data to DWS	Dumping Data to CloudTable
Data is stored only in DIS.	Data is stored in DIS and periodically dumped to OBS.	Data is stored in DIS and periodically imported into the HDFS of an MRS cluster. **NOTE** Before streaming data is imported into an MRS cluster, it is stored in a temporary data directory (an OBS bucket). Data in the temporary directory will be deleted after being dumped to the MRS cluster.	Data is stored in DIS and periodically dumped to DLI. **NOTE** Before streaming data is imported into a DLI table, it is stored in a temporary data directory (an OBS bucket). Data in the temporary directory will be deleted after being dumped to the DLI table.	Data is stored in DIS and periodically dumped to DWS. **NOTE** Before streaming data is imported into a DWS, it is stored in a temporary data directory (an OBS bucket). Data in the temporary directory will be deleted after being dumped to the DWS.	Data is stored in DIS and imported into the HBase or OpenTSDB table of a CloudTable cluster in real time.

1.8 How Do I Check Software Package Integrity?

This section describes how to verify integrity of the DIS SDK software package on a Linux system by using a verification file.

Prerequisites

- The PuTTY tool is available.
- The WinSCP tool is available.

Procedure

Step 1 Upload the DIS SDK software package **huaweicloud-sdk-dis-x.x.x1.2.3.zip** to any directory on the Linux system by using WinSCP.

NOTE

In **huaweicloud-sdk-dis-x.x.x.zip**, **x.x.x** indicates the version number of the DIS SDK software package.

Step 2 Log in to the Linux system by using PuTTY. In the directory in which **huaweicloud-sdk-dis-x.x.x1.2.3.zip** is stored, run the following command to obtain the verification code of the DIS SDK software package

sha256sum huaweicloud-sdk-dis-x.x.x.zip

Example verification code:

```
# sha256sum dis-sdk-x.x.x.zip
8be2c937e8d78b1a9b99777cee4e7131f8bf231de3f839cf214e7c5b5ba3c088 huaweicloud-sdk-dis-x.x.x.zip
```

Step 3 Open the DIS SDK verification file **huaweicloud-sdk-dis-x.x.x1.2.3.zip.sha256sum** and compare it with the verification code obtained in **Step 2**.

- If they are consistent, the DIS-SDK compression package has not been tampered with.
- If they are inconsistent, the DIS SDK software package is tampered with and you need to obtain it again.

----End

1.9 How Do I Send and Retrieve Data Using DIS?

Step 1 Create a DIS stream. Obtain your Access Key ID/Secret Access Key (AK/SK) pair from the Identity and Access Management (IAM) service.

Step 2 Download the **dis-sdk-*X.X.X*.zip** package at **https://dis-publish.obs-website.cn-north-1.myhuaweicloud.com/** and decompress it.

Step 3 Create a producer project and configure the user AK/SK, endpoint, project ID, region, stream name, and the number of partitions.

Step 4 Run the producer application to send data.

Step 5 Create a consumer project and configure the user AK/SK, endpoint, project, region, stream name, partition ID, and startingSequenceNumber.

Step 6 Run the consumer application to retrieve data.

----End

1.10 What Is Data Control?

Data control is performed when the maximum throughput of a stream is exceeded. It does not affect the charge and data.

2 Dump Questions

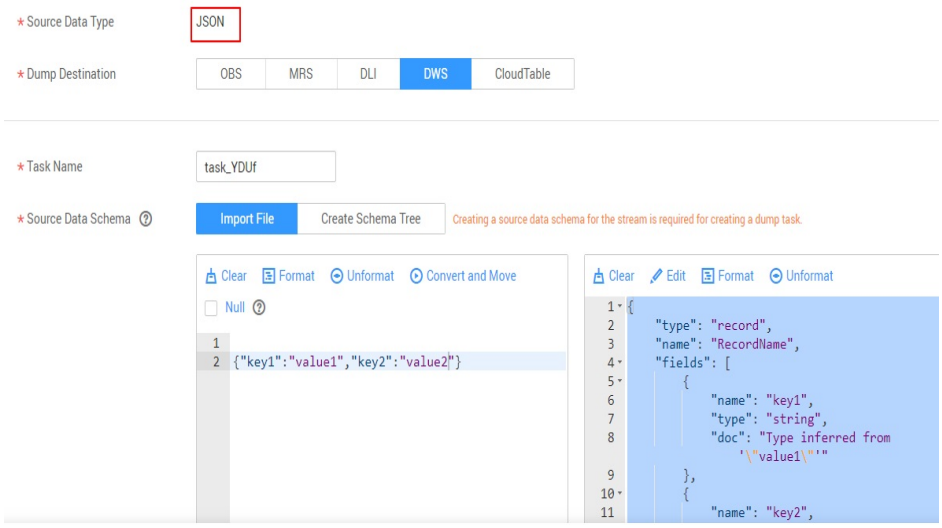
2.1 How Does DIS Dump Data to a Specific Column of DWS?

DIS can dump source data in JSON format to DWS. Before dumping data, you need to configure the source data schema.

A source data schema is a user's JSON data sample used to describe the JSON data format. DIS can generate an Avro schema based on the JSON data sample and convert the JSON data uploaded to a stream to the Parquet or CarbonData format.

1. Create a source data schema. For details, see **Managing a Source Data Schema**. The following describes how to create a source data schema when adding a dump task.
   a. Select a stream whose source data type is JSON.
   b. On the **Dump Tasks** tab page, click **Create Dump Task**.
   c. Set **Dump Destination** to **DWS** and configure **Source Data Schema** by importing a file.
   d. Enter the source data sample, click **Convert Source Data Sample**, and click **Submit**.

Figure 2-1 Creating a source data schema



2. Configure schema filtering.

   NOTE

   Schema filtering is valid only for the root node or level-1 subnode of the source data schema that is not of the array type. For details about how to create a source data schema, see **Managing a Source Data Schema**.

   a. Enable schema filtering.

   b. In the **Source Data Attribute Name** list, select the corresponding attribute names to map the specific columns in the DWS table.

      NOTE

      Attributes in the source data attribute list are generated by the **name** field in the source data schema and match the column name in the DWS table.

Figure 2-2 Configuring schema attributes

   c. As shown in **Figure 2-2**, only **id** is selected as the source data attribute name, which is less than the total number of fields in the corresponding table.

      Create a cluster on DWS and run the following command to create a table:

      CREATE TABLE dis_test3(id TEXT,dev TEXT,online BIGINT,module TEXT default 'a',logTime TEXT,appId TEXT,event TEXT);

d. After data is dumped from DIS to DWS, log in to the cluster database and query data in the **dis_test3** table. You can find that data is inserted only into the **id** and **module** columns. The data in the **module** column is the default data, as shown in **Figure 2-3**.

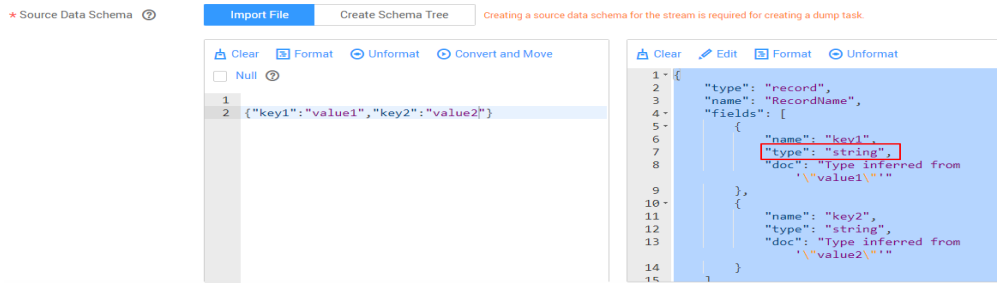
Figure 2-3 Schema filtering result

```
postgres=> select * from dis_test3;
           id           | dev | online | module | logtime | appid | event
------------------------+-----+--------+--------+---------+-------+-------
 xsadas1212331232l3sadasd |     |        | a      |         |       |
 xsadas1212331232l3sadasd |     |        | a      |         |       |
(2 rows)
```

2.2 How Does a Schema Support Default Fields or NULL Fields?

A source data schema is a user's JSON data sample used to describe the JSON data format. DIS can generate the Avro schema based on the JSON data sample. By default, the default field or NULL is not supported, as shown in **Figure 2-4**.

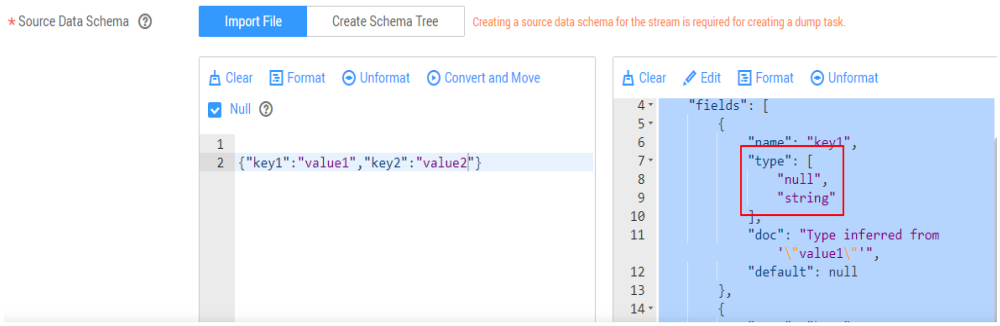
Figure 2-4 Example of not supporting the default field



The type of the **key1** field is **String** (**type**: **string** in Avro Schema). If the **key1** field in the source data is not transferred or the transferred value is null, the dump task reports an error.

To enable the schema generated based on the JSON data sample supports the default value or null, select the **Convert and Move** and click **Convert Source Data Sample**, as shown in **Figure 2-5**.

Figure 2-5 Example of supporting the default field



In this case, the type of the **key1** field is **Union** (**type**: **[null, string]** in Avro Schema). If the **key1** field in the source data is not transferred or the transferred value is null, null is automatically filled as the default value. The dump task can properly convert the format.

2.3 How Do I Access DIS Through a Direct Connect Connection?

Step 1 Enable Direct Connect by referring to **Using Direct Connect in self-service mode**. Then, use the cloud account of Direct Connect to access VPC.

Step 2 Choose **VPC Endpoint** > **VPC Endpoints** and click **Buy VPC Endpoint**.

Step 3 Set **Service Category** to **Cloud services** and select a DIS endpoint. Select a VPC and subnet where the connection resides. After the VPC endpoint is created, a node IP address is automatically assigned to the endpoint.

Step 4 Use this IP address to access DIS.

----End

2.4 How Do I Distinguish Different Types of Data When Accessing Data in Streams?

- By stream: Different types of data use different streams and can be distinguished by stream.
- By partition: Different types of data use different partitions on one stream. Each partition is assigned one partition key. Different types of data can be distinguished by partition key of each partition.

3 DIS Agent Questions

3.1 How Can I Configure Agent to Listen To Multiple Directories or Files?

DIS Agent can listen to multiple directories or files. For example, to collect logs of the **/home/folder1/file1** and **/home/folder2/file2** files, configure multiple DIS streams.

```
---
region: REGION
ak: YOUR_AK
sk: YOUR_SK
projectId: YOUR_PROJECTID
endpoint: ENDPOINT
flows:
  - DISStream: YOUR_STREAM
    filePattern: /home/folder1/file1
    initialPosition: START_OF_FILE
    maxBufferAgeMillis: 5000

  - DISStream: YOUR_STREAM
    filePattern: /home/folder2/file2
    initialPosition: START_OF_FILE
    maxBufferAgeMillis: 5000
```

3.2 How Can I Configure Recursive Listening for Directories on DIS Agent?

On DIS Agent, set **directoryRecursionEnabled** to **true**. For example, the following configuration can match **/home/one.log**, **/home/child/two.log**, and **/home/child/child/three.log**:

```
---
region: REGION
ak: YOUR_AK
sk: YOUR_SK
projectId: YOUR_PROJECTID
endpoint: ENDPOINT
flows:
  - DISStream: YOUR_STREAM
    filePattern: /home/*.log
```



```
  directoryRecursionEnabled: true
  initialPosition: START_OF_FILE
  maxBufferAgeMillis: 5000
```

3.3 How Can I Configure DIS Agent?

DIS Agent can upload data to DIS by configuring **PROXY_HOST**, **PROXY_PORT**, **PROXY_USERNAME**, and **PROXY_PASSWORD**. For details about these configuration items, see **Configuring DIS Agent**.

```
---
region: REGION
ak: YOUR_AK
sk: YOUR_SK
projectId: YOUR_PROJECTID
endpoint: ENDPOINT
PROXY_HOST: YOUR_PROXY_HOST
PROXY_PORT: YOUR_PROXY_PORT
PROXY_USERNAME: YOUR_PROXY_USERNAME
PROXY_PASSWORD: YOUR_PROXY_PASSWORD
flows:
 - DISStream: YOUR_STREAM
   filePattern: /home/*.log
   initialPosition: START_OF_FILE
   maxBufferAgeMillis: 5000
```

3.4 How Do I Use Agent to Encrypt AK/SK?

Secret Access Key (SK) is sensitive information. To encrypt the SK, perform the following steps:

Step 1 Go to the **bin/** directory.

cd /opt/dis-agent-X.X.X/bin

Step 2 Run the encryption script, enter the password, and press **Enter**.

bash dis-encrypt.sh

Step 3 View the encryption result. The character string following "Encrypt result:" displayed on the console is the encryption result. Use this method to encrypt the MySQL password and SK, respectively and record the ciphertext in the configuration file.

----End