Data Ingestion Service

Getting Started

Issue 01

Date 2023-07-14

HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD.

Trademarks and Permissions


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

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

Contents

1 General Procedure

The following is the general procedure for using DIS:

Step 1: Creating a DIS Stream

You need to create a stream before using DIS.

Step 2: Preparing a DIS Application Development Environment

Before developing a DIS application, install an application development tool, and import your SDK package and sample project into the development environment.

Step 3: Sending Data to DIS

Write a producer application and run it to send data to the cloud. The DIS stream information can be viewed on the DIS console.

Step 4: Obtaining Data from DIS

Write a consumer application and run it to retrieve data from the cloud.

2 Step 1: Creating a DIS Stream

You can create a DIS stream on the DIS management console.

Procedure

Step 1 Use the account to log in to the **DIS console**.

Step 2 Click  in the upper left corner of the page and select a region and project.

Step 3 Click **Buy Stream** and set related parameters.

Table 2-1 Stream parameters

Parameter	Description	Example
Billing Mode	Pay-per-use	Pay-per-use
Region	Physical location of the cloud service. You can select a different region from the drop-down list.	-
Basic Information		
Stream Name	Name of the DIS stream to be created. A stream name is 1 to 64 characters long. Only letters, digits, hyphens (-), and underscores (_) are allowed.	dis-Tido
Stream Type	• **Common**: Each partition supports a maximum read speed of 2 MB/s and a maximum write speed of 1 MB/s. • **Advanced**: Each partition supports a maximum read speed of 10 MB/s and a maximum write speed of 5 MB/s.	-
Partitions	Partitions are the base throughput unit of a DIS stream.	5

Parameter	Description	Example
Partition Calculator	Calculator used to calculate the estimated number of partitions based on the information you entered. 1. Click **Partition Calculator**. 2. In the **Partition Calculator** dialog box, configure the **Average Record Size (KB)**, **Max. Records Written**, and **Consumer Applications** parameters. The **Estimated Partitions** field then displays the recommended number of partitions. The value of this field cannot be modified. **NOTE** Partition calculation formulas: – Based on the traffic (the final value must be rounded up): Common stream: Average record size x (1 + 20%) x Maximum records written/ (1 x 1024 KB) (20% is the reserved partition percentage.) Advanced stream: Average record size x (1 + 20%) x Maximum records written/ (5 x 1024 KB) (20% is the reserved partition percentage.) – Based on the consumer program quantity (the final value must be rounded up): (Number of consumer programs/2) x Number of partitions calculated based on the traffic (The result of the number of consumer programs/2 must reserve two decimals.) The largest value among the values calculated based on the previous three formulas is considered as the estimated partition value. 3. Click **Use Estimated Value**. The estimated value is automatically used as the value of **Partitions**.	-
Data Retention (hours)	The maximum number of hours for which data can be preserved in DIS. Data will be deleted when the retention period expires. Value range: an integer ranging from 24 to 72.	24

Parameter	Description	Example
Source Data Type	• **BLOB**: a collection of binary data stored as a single entity in a database management system. If **Source Data Type** is set to **BLOB**, the supported **Dump Destination** can be **OBS** or **MRS**. • **JSON**: an open-standard file format that uses human-readable text to transmit data objects consisting of attribute–value pairs and array data types. If **Source Data Type** is set to **JSON**, the supported **Dump Destination** can be **OBS**, **MRS**, **DLI**, or **DWS**. • **CSV**: a simple text format for storing tabular data in a plain text file. If **Source Data Type** is set to **CSV**, the supported **Dump Destination** can be **OBS**, **MRS**, **DLI**, or **DWS**.	JSON
Auto-Scaling	You can choose to enable or disable auto-scaling when creating a stream. You can click  or  to disable or enable auto-scaling.	**NOTE** You can choose whether to enable auto-scaling when creating a stream. You can also modify the auto-scaling attributes for a created stream.
Auto-Scale Down To	Lower limit for automatic scale-down. The number of target partitions for automatic scale-down must be greater than or equal to the lower limit.	-
Auto-Scale Up To	Upper limit for automatic scale-up. The number of target partitions for automatic scale-up must be smaller than the lower limit.	-
Data Delimiter	Data delimiter when **Source Data Type** is **CSV**.	-

Parameter	Description	Example
Schema	Whether to create a schema when creating a stream. This parameter is available when **Source Data Type** is **JSON** or **CSV**. You can click  or  to disable or enable the schema configuration. **NOTE** If no data schema is created when a stream is created, you can also create it later after the stream is created. Create a schema on the **Stream Management** page. For details, see **Managing a Source Data Schema**.	You can create a schema only when the source data type is set to **JSON** or **CSV**.
Source Data Schema	You can enter or import source data samples in JSON or CSV format. For details, see **Managing a Source Data Schema**. 1. In the left text box, enter a JSON or CSV source data sample or click  to import a source data sample. 2. In the left text box, click  to delete your entered or imported source data sample. 3. In the left text box, click  to generate an Avro schema in the right text box according to the source data sample. 4. In the right text box, click  to delete the generated Avro schema. 5. In the right text box, click  to modify the generated Avro schema.	This parameter is mandatory only when **Schema** is set to **Enable**.
Enterprise Project	Configure the enterprise project to which streams belong. You can configure this parameter only when the Enterprise Management service is enabled. The default value is **default**. An enterprise project facilitates project-level management and grouping of cloud resources and users. You can select the default enterprise project (**default**) or other existing enterprise projects. To create an enterprise project, log in to the Enterprise Management console. For details, see the *Enterprise Management User Guide*.	-

Parameter	Description	Example
Configure	Click **Configure now**. The **Tag** parameter is displayed. For details about how to add a tag, see **Managing Stream Tags**.	-
Skip	No advanced settings need to be configured.	-
Tag	Identifier of the stream. Adding tags to streams can help you identify and manage your stream resources.	-

Step 4 Click **Next**. The **Details** page is displayed.

Step 5 Click **Submit**.

----End

3 Step 2: Preparing a DIS Application Development Environment

Before developing DIS applications, prepare an application development environment, and then obtain a software development kit (SDK) and sample project and import them to the development environment.

Prerequisites

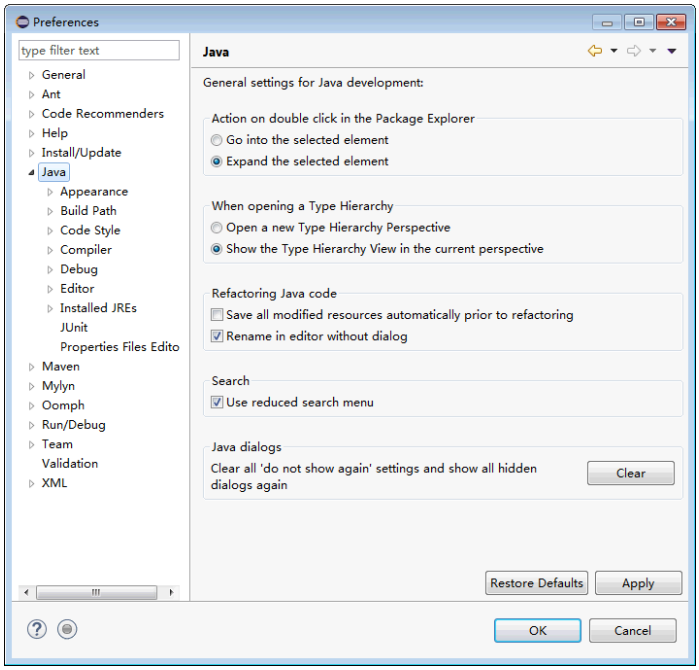
- JDK 1.8 or later has been installed.
- Eclipse has been installed.

Procedure

Step 1 Configure a JDK using Eclipse.

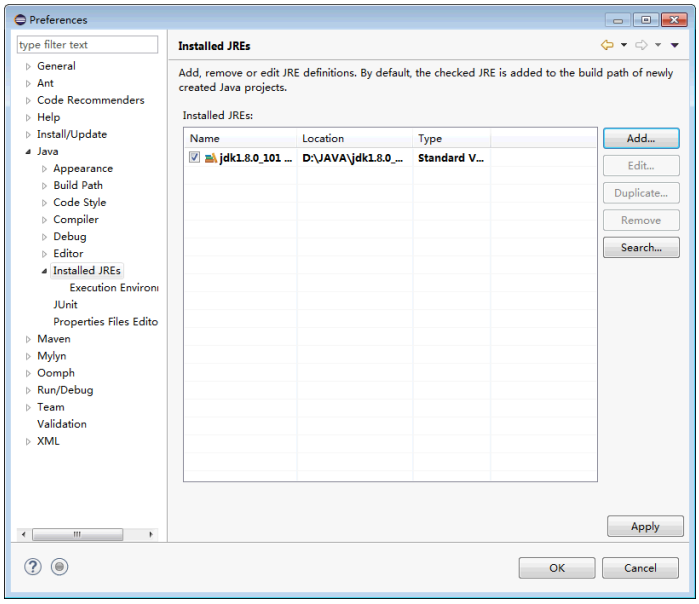
1. Start Eclipse and choose **Window** > **Preferences**. The **Preferences** dialog box is displayed.
2. In the navigation tree, choose **Java**. On the **Java** page, configure general settings for Java development and then click **OK**.

Figure 3-1 Preferences



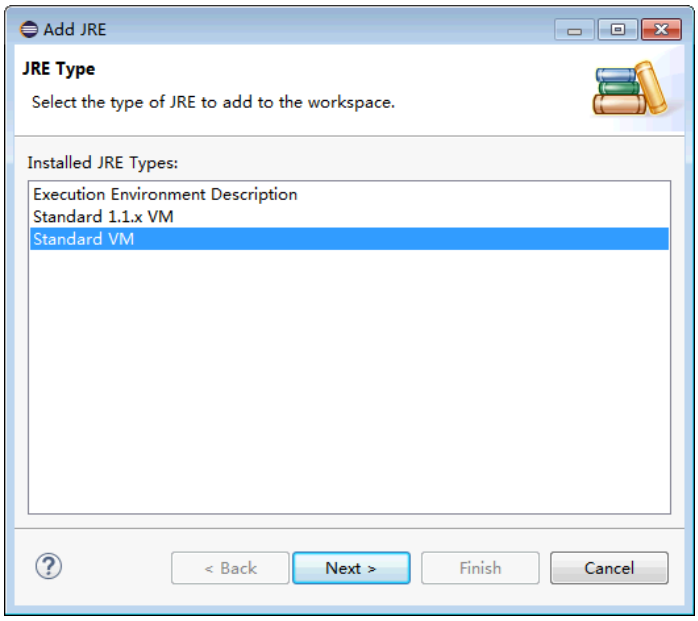
3. In the navigation tree, choose **Java** > **Installed JREs**.
   - Ensure that configured JDK environmental variables are displayed on the **Installed JREs** page. Then go to **Step 1.3.a**.
   - To configure different variables for different versions of JDK, perform **Step 1.3.b** to **Step 1.3.d**.

Figure 3-2 Installed JREs



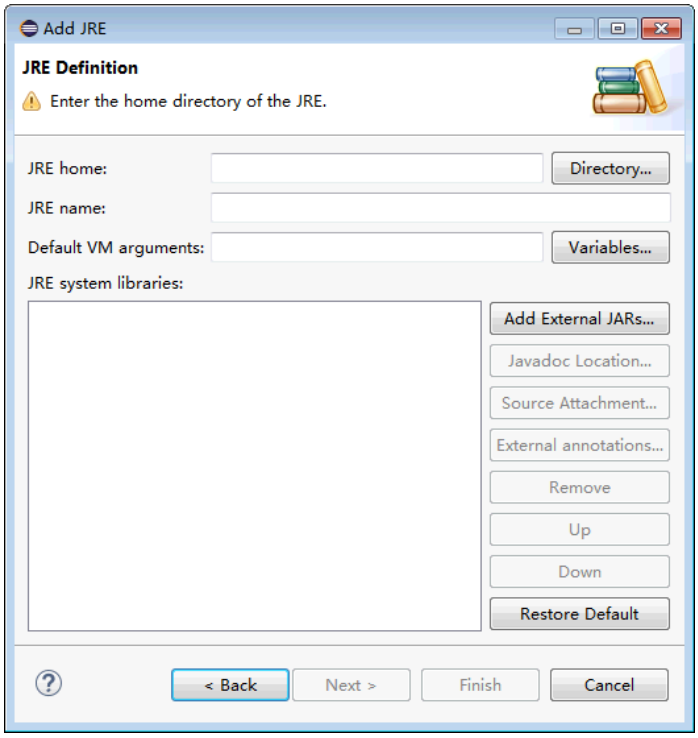
   a. Select the installed JDK and click **OK**.
   b. Click **Add**. The **Add JRE** dialog box is displayed.

Figure 3-3 JRE Type



c. Select a JRE type and click **Next**.

Figure 3-4 JRE Definition



d. Configure the basic information about JDK and click **Finish**.

- JRE home: JDK installation path.
- Default VM arguments: JDK running parameters.

Step 2 Download resource packages.

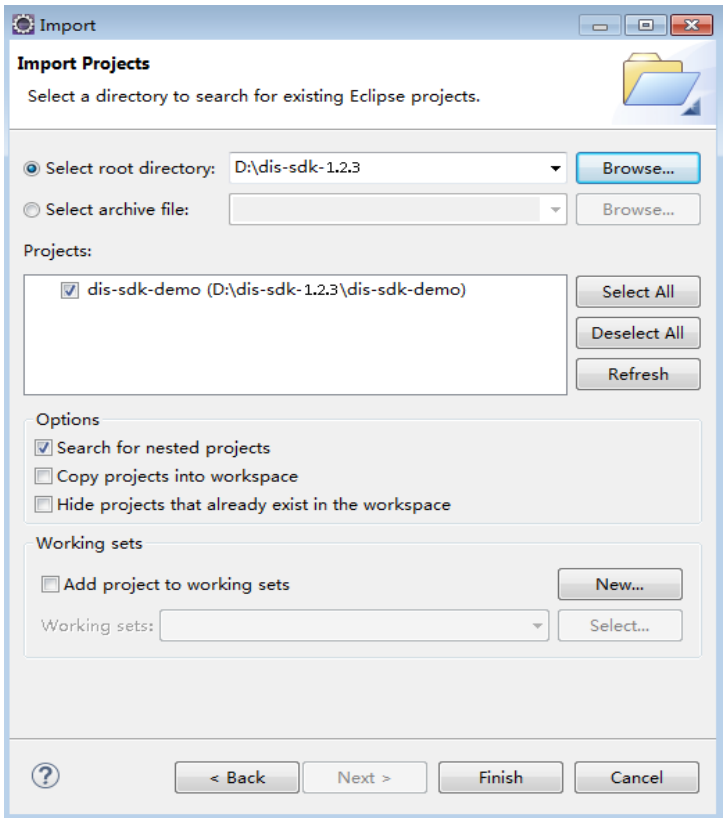
Download the DIS Java SDK from **https://github.com/huaweicloud/huaweicloud-sdk-java-dis**.

Obtain **huaweicloud-sdk-dis-java-*X.X.X*.zip** from the **DIS SDK**. The package contains the demo package of the sample project.

Step 3 Import the Eclipse project.

1. Start Eclipse. Choose **File** > **Import**. The **Import** dialog box is displayed.
2. Choose **Maven** > **Existing Maven Projects**, and click **Next**. The **Import** dialog box is displayed.
3. Click **Browse** and select a save location for the **dis-sdk-demo** sample project. In the **Projects** area, select a sample project.

Figure 3-5 Importing a project



4. Click **Finish** to import the project.

Step 4 Configure the demo project.

1. Set the project code to **UTF-8**.
   a. In the navigation tree, right-click the required project under **Project Explorer** and choose **Properties** from the shortcut menu. The **Properties for dis-sdk-demo** dialog box is displayed.
   b. In the navigation tree, choose **Resource**. The **Resource** page is displayed in the right pane.
   c. In the **Other** drop-down list, select **UTF-8**.
   d. Click **Apply and Close**.

2. Add the JDK.
   a. In the navigation pane, choose **Project Explorer**. Right-click the chosen project and choose **Properties** from the shortcut menu.
   b. In the navigation tree, choose **Java Build Path**. The **Java Build Path** page is displayed in the right pane.
   c. Click the **Libraries** tab, and then click **Add Library**. The **Add Library** dialog box is displayed.
   d. Select **JRE System Library** and click **Next**. Verify that the version of **Workspace default JRE** is **jdk1.8** or later.
   e. Click **Finish** to exit the **Add Library** dialog box.
   f. Click **Apply and Close**.

Step 5 Initialize a DIS client sample. For details about **endpoint**, **ak**, **sk**, **region**, and **projectId**, see **Obtaining Authentication Information**.

----End

4 Step 3: Sending Data to DIS

Function

Local data is continuously uploaded to DIS.

NOTE

Data can be stored in MRS, DIS, OBS, and DLI. For details about how to configure a storage location, see **Creating a Dump Task**.

The maximum number of days for DIS to preserve data cannot exceed **Data Retention (days)**.

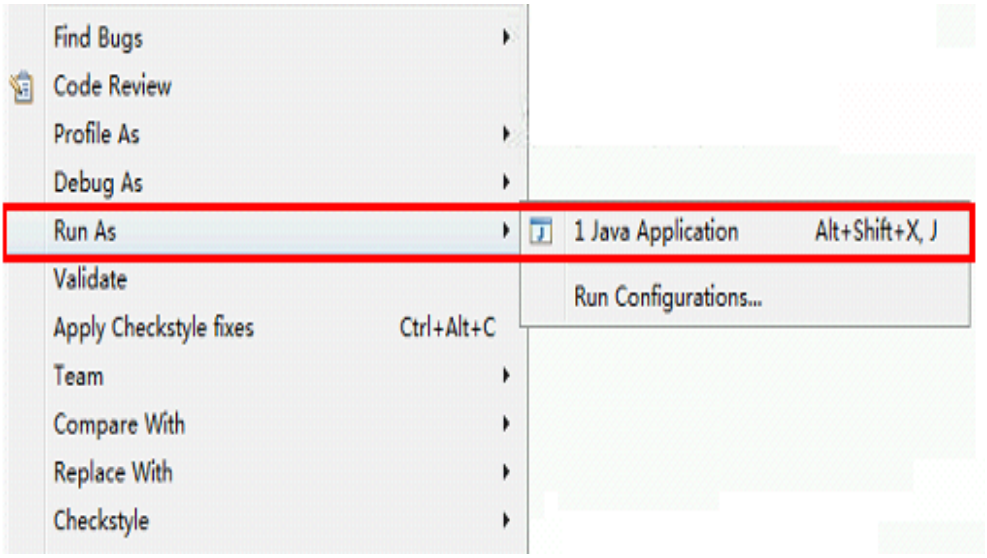
Sample Code

The example code file is the **ProducerDemo.java** file in the **\dis-sdk-demo\src\main\java\com\bigdata\dis\sdk\demo** directory decompressed from the **huaweicloud-sdk-dis-java-*X.X.X*.zip** package. The compression package is downloaded from the **DIS SDK**.

Running the Producer Program

Right-click the producer application and choose **Run As** > **1 Java Application** from the shortcut menu.

Figure 4-1 Running a producer application



While data is being sent to DIS, the DIS console displays DIS stream information. If information similar to the following is displayed, the data has been successfully sent to DIS:

```
14:40:20.090 [main] INFOcom.bigdata.dis.sdk.DISConfig - get from classLoader
14:40:20.093 [main] INFODEMOT - ========== BEGIN PUT ============
14:40:21.186 [main] INFOcom.bigdata.dis.sdk.util.config.ConfigurationUtils - get from classLoader
14:40:21.187 [main] INFOcom.bigdata.dis.sdk.util.config.ConfigurationUtils - propertyMapFromFile size : 2
14:40:22.092 [main] INFOcom.bigdata.dis.sdk.demo.ProducerDemo - Put 3 records[3 successful / 0 failed].
14:40:22.092 [main] INFOcom.bigdata.dis.sdk.demo.ProducerDemo - [hello world.] put success, partitionId
[shardId-0000000000], partitionKey [964885], sequenceNumber [0]
14:40:22.092 [main] INFOcom.bigdata.dis.sdk.demo.ProducerDemo - [hello world.] put success, partitionId
[shardId-0000000000], partitionKey [910960], sequenceNumber [1]
14:40:22.092 [main] INFOcom.bigdata.dis.sdk.demo.ProducerDemo - [hello world.] put success, partitionId
[shardId-0000000000], partitionKey [528377], sequenceNumber [2]
14:40:22.092 [main] INFOcom.bigdata.dis.sdk.demo.ProducerDemo - ========== PUT OVER ============
```


5 Step 4: Obtaining Data from DIS

Function

You can retrieve data from DIS when needed.

Sample Code

The example code file is the **ConsumerDemo.java** file in the **\dis-sdk-demo\src\main\java\com\bigdata\dis\sdk\demo** directory decompressed from the **huaweicloud-sdk-dis-java-*X.X.X*.zip** package. The compression package is downloaded from the **DIS SDK**.

Running the Consumer Application

If information similar to the following appears, data has been successfully retrieved from DIS:

```
14:55:42.954 [main] INFOcom.bigdata.dis.sdk.DISConfig - get from classLoader
14:55:44.103 [main] INFOcom.bigdata.dis.sdk.util.config.ConfigurationUtils - get from classLoader
14:55:44.105 [main] INFOcom.bigdata.dis.sdk.util.config.ConfigurationUtils - propertyMapFromFile size : 2
14:55:45.235 [main] INFOcom.bigdata.dis.sdk.demo.ConsumerDemo - Get stream
streamName[partitionId=0] cursor success :
eyJnZXRJdGVyYXRvclBhcmFtIjp7InN0cmVhbS1uYW1lIjoiZGlzLTEzbW9uZXkiLCJwYXJ0aXRpb24taWQiOiIwIiwiY
3Vyc29yLXR5cGUiOiJBVF9TRVFVRU5DRV9OVU1CRVIiLCJzdGFydGluZy1zZXF1ZW5jZS1udW1iZXIiOiIxMDY4O
TcyIn0sImdlbmVyYXRlVGltZXN0YW1wIjoxNTEzNjY2NjMxMTYxfQ
14:55:45.305 [main] INFOcom.bigdata.dis.sdk.demo.ConsumerDemo - Get Record [hello world.],
partitionKey [964885], sequenceNumber [0].
14:55:45.305 [main] INFOcom.bigdata.dis.sdk.demo.ConsumerDemo - Get Record [hello world.],
partitionKey [910960], sequenceNumber [1].
14:55:46.359 [main] INFOcom.bigdata.dis.sdk.demo.ConsumerDemo - Get Record [hello world.],
partitionKey [528377], sequenceNumber [2].
```


6 Obtaining Authentication Information

Obtaining AK/SK

Access Key ID/Secret Access Key (AK/SK) is created on Identity and Access Management (IAM) to authenticate calls to application programming interfaces (APIs) on the public cloud. To obtain an AK/SK pair, choose **My Credentials** > **Access Keys**.

Obtaining Project ID

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

Obtaining Region Information and Endpoint Information

For details about regions and endpoints, see **Regions and Endpoints**.

7 Interconnecting with OBS

Introduction

DIS can upload data to Object Storage Service (OBS).

Prerequisites

An IAM agency has been created by following the procedure in **Creating an IAM Agency**. This IAM agency entrusts DIS to access your OBS resources.

Data Dumping

You can set **Dump Bucket** when **creating a dump task**. If Dump Destination is set to **OBS**, DIS periodically imports data from DIS streams to OBS.

8 Creating an IAM Agency

Introduction

If you choose to dump data from DIS to OBS, MRS, or DLI, create an IAM agency that grants DIS permissions to access OBS, MRS, or DLI.

Creating an IAM Agency

Step 1 Log in to the management console.

Step 2 Click **Service List**. Under **Management & Deployment**, select **Identify and Access Management**.

Step 3 Select **Agencies** in the navigation tree pane, and click **Create Agency**.

Step 4 Configure agency parameters and click **OK**.

Table 8-1 Agency parameters

Parameter	Description
Agency Name	Name of the agency to be created. The value of this parameter is 1 to 64 characters long and cannot be left unspecified.
Agency Type	Type of the agency to be created. This parameter must be set to **Cloud service**.
Cloud Service	Click **Select** next to **Cloud Service**. In the **Select Cloud Service** dialog box, select **DIS** and click **OK**.
Validity Period	Select **Unlimited**. **NOTE** Currently, this parameter must be set to **Unlimited**. Using another value may result in authorization failures.
Description	Agency description. The entered description cannot exceed 255 characters.

Parameter	Description
Permissions	To modify agency policies, click **Modify** in the **Operation** column. In the **Available Policies** area, select your required policy and click **OK**. **NOTE** After an agency is created, its policies cannot be modified.

----End

9 Practices

After performing the operations in **Creating a DIS Stream** and **Obtaining Data from DIS**, you can use a series of common practices provided by DIS as you need.

Table 9-1 Common best practices

Practice		Description
Case study	**Using DIS to Analyze Vehicle Locations in Real Time**	Data Ingestion Service (DIS) collects vehicle location data in real time and uploads the data to CloudTable Service (CloudTable). You can use CloudTable to query locations of a vehicle in a specified period.
	Collecting Incremental Log Data of Driving Behavior	DIS collects incremental driving behavior log data and uploads the data to Huawei Cloud Object Storage Service (OBS). Data Lake Insight (DLI) analyzes the uploaded log data to obtain the driving behavior and helps automobile manufacturers provide value-added services such as improvement of driving habits.

