Data IDE

Quick start

Quick start

Import local data

[Description] Local data importing supports .txt and .csv file types. The file size should not exceed 10MB. Partition- and table-based data imports are not supported.

Taking importing banking.txt as an example, the descriptions are as follows:

Create the MaxCompute target table ;

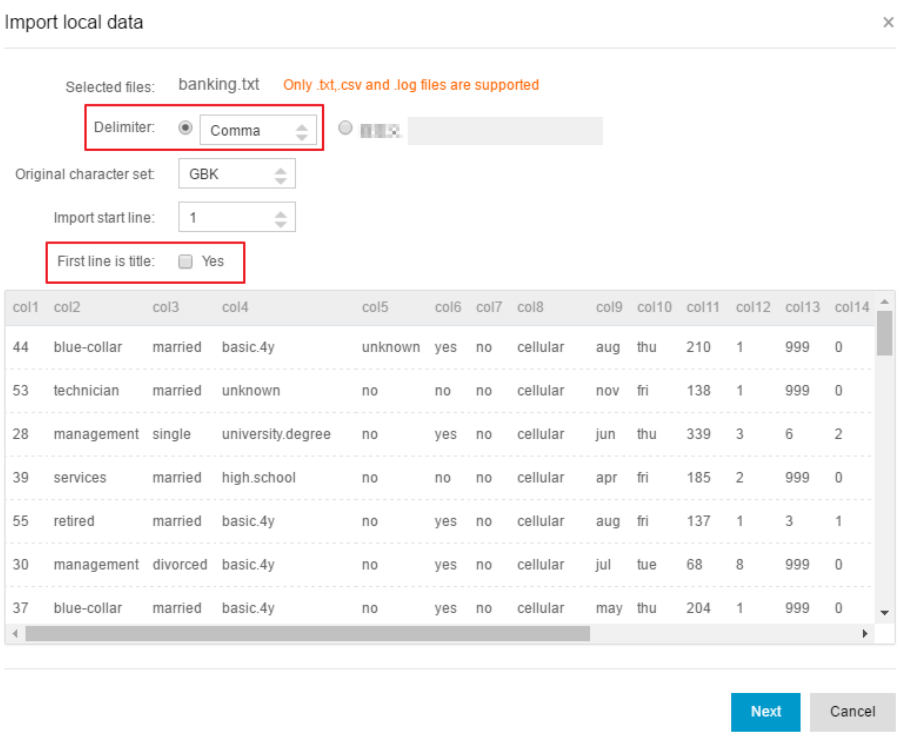
The tabulation statements are as follows:

```
create table if not exists bank_data
(age bigint comment 'Age',
job string comment 'Job type',
marital string comment 'arital status',
education string comment 'Educational level',
default string comment 'with credit card',
housing string comment 'With mortgage',
loan string comment 'With loan',
contact string comment 'Contact information',
month string comment'Month',
day_of_week string comment'Day of week',
duration string comment 'Duration',
campaign int comment 'Number of contacts during the campaign',
pdays double comment 'Time interval from last contact',
previous double comment 'Number of previous contacts with the customer',
poutcome string comment 'Outcome of previous marketing activities',
emp_var_rate double comment 'Employment change rate',
cons_price_idx double comment 'Consumer price index',
cons_conf_idx double comment 'Consumer confidence index',
euribor3m double comment 'Euro deposit interest rate',
nr_employed double comment 'Number of employees',
y bigint comment 'With fixed-term loan');
```

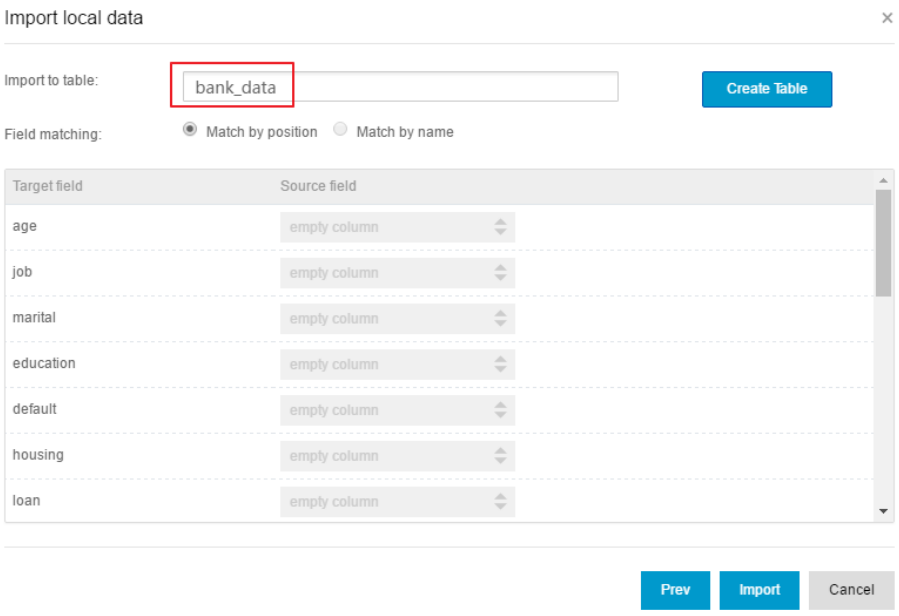
Click “Data Development” in the top menu bar to navigate to **Data IDE > New** ;

Click “Import” , and select **Import Local Data** ;

Select a local data file, configure the import information and click “Next” ;

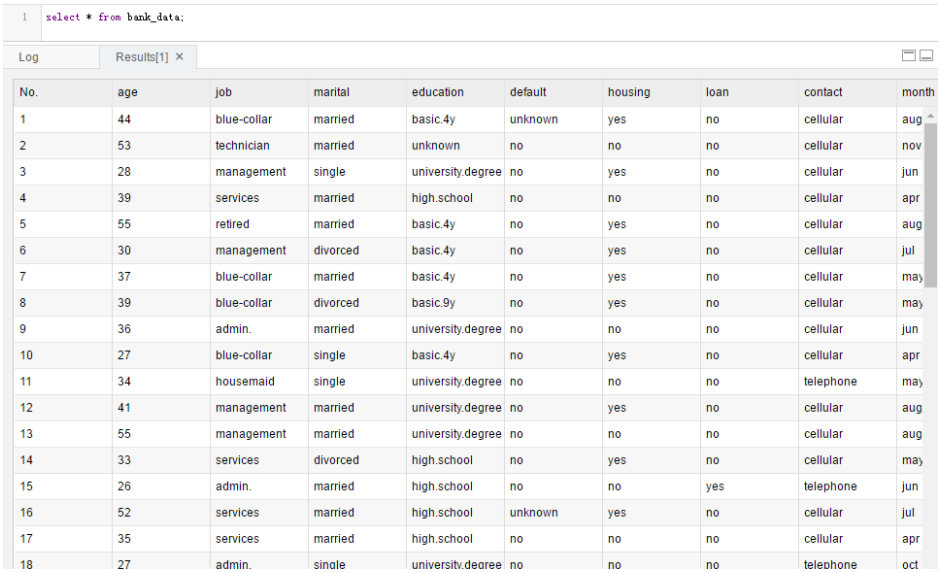


Select the target table, and the field matching method (Match by Location in this example). Then click **Import** ;



After the successful file importation, **File imported successfully** will be prompted at the top right corner of the system. You can execute the SELECT statement to view the data at the

same time.



```
select * from bank_data;
```

Log | Results[1] ×

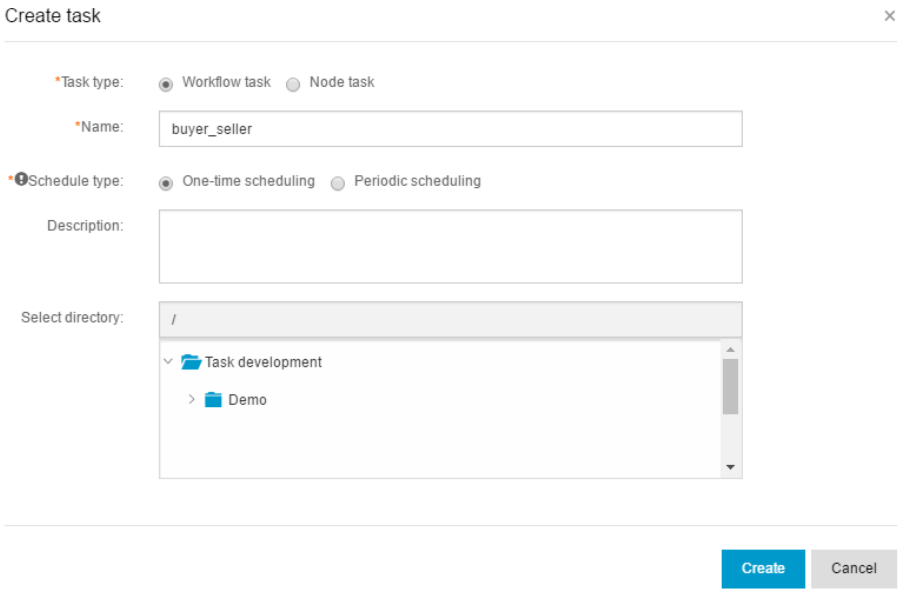
No.	age	job	marital	education	default	housing	loan	contact	month
1	44	blue-collar	married	basic.4y	unknown	yes	no	cellular	aug
2	53	technician	married	unknown	no	no	no	cellular	nov
3	28	management	single	university.degree	no	yes	no	cellular	jun
4	39	services	married	high.school	no	no	no	cellular	apr
5	55	retired	married	basic.4y	no	yes	no	cellular	aug
6	30	management	divorced	basic.4y	no	yes	no	cellular	jul
7	37	blue-collar	married	basic.4y	no	yes	no	cellular	may
8	39	blue-collar	divorced	basic.9y	no	yes	no	cellular	may
9	36	admin.	married	university.degree	no	no	no	cellular	jun
10	27	blue-collar	single	basic.4y	no	yes	no	cellular	apr
11	34	housemaid	single	university.degree	no	no	no	telephone	may
12	41	management	married	university.degree	no	yes	no	cellular	aug
13	55	management	married	university.degree	no	no	no	cellular	aug
14	33	services	divorced	high.school	no	yes	no	cellular	may
15	26	admin.	married	high.school	no	no	yes	telephone	jun
16	52	services	married	high.school	unknown	yes	no	cellular	jul
17	35	services	married	high.school	no	no	no	cellular	apr
18	27	admin.	single	university.degree	no	no	no	telephone	oct

This document takes MaxCompute for SQL as an example to illustrate the operating procedures.

Create a job

Click **New Job** in the tool bar on the “Data IDE” interface ;

Fill in the various configuration items in the New Job pop-up box. Here we take the creation of the **One-time Scheduling** workflow for example. If the workflow requires daily automatic scheduling, you can choose “Periodic Scheduling”, and then configure the scheduling cycle in the workflow attributes ;



Configuration items in the "New Job" pop-up box are described as follows:

■ **Job Type**: including the workflow jobs and node jobs. The workflow jobs can contain multiple node jobs.

■ **Job Name**: A job name is composed of numbers, letters, and underscores.

■ **Scheduling Type**: The scheduling type can be one-time scheduling or periodic scheduling. The scheduling type cannot be modified after the workflow is successfully created. The workflow attributes and node attributes of one-time scheduling do not contain the scheduling attributes. At the same time, you can directly run the current workflow on the workflow development panel.

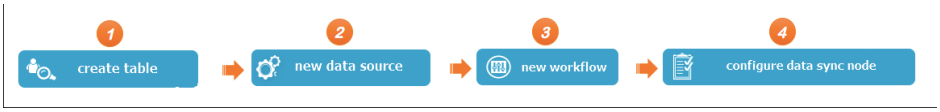
■ **Description**: A brief description of the current workflow. The description may contain Chinese characters, letters, numbers, and underscores.

■ **Select Directory**: You can select the file tree that the job belongs to.

> [Description] Currently the node job only supports periodic scheduling. You should also select the node type including: data synchronization, and MaxCompute SQL.

Click **Create**.

[Description] Currently data source types supported by the data synchronization jobs include: **MaxCompute, RDS (MySQL, SQL Server, PostgreSQL), Oracle, FTP, ADS, OSS, OCS, and DRDS**.



Taking the data synchronization from RDS to MaxCompute for example, the detailed descriptions can be found below:

Step 1: Create a data table

For details on how to create a MaxCompute table, see Create a Table.

Step 2: Create a data source

> [Description] Only users with the project administrator role are allowed to create a new data source.

Preparation

Currently only the China East 1 (Hangzhou) region is supported as a RDS data source, and the Beijing region is not yet supported. In addition, when the RDS data sources in the Hangzhou region cannot be connected to during testing, you should add a whitelist of data synchronization server IP addresses onto your RDS:

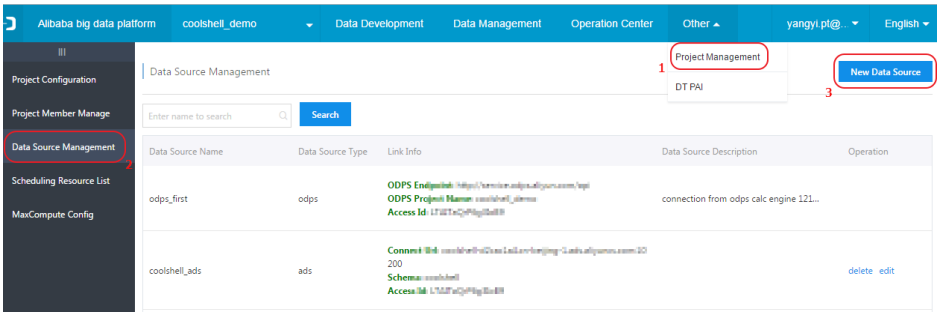
10.152.69.0/24,10.153.136.0/24,10.143.32.0/24,120.27.160.26,10.46.67.156,120.27.160.81,10.46.64.81,121.43.110.160,10.117.39.238,121.43.112.137,10.117.28.203,118.178.84.74,10.27.63.41,118.178.56.228,10.27.63.60,118.178.59.233,10.27.63.38,118.178.142.154,10.27.63.15

The specific steps are as follows:

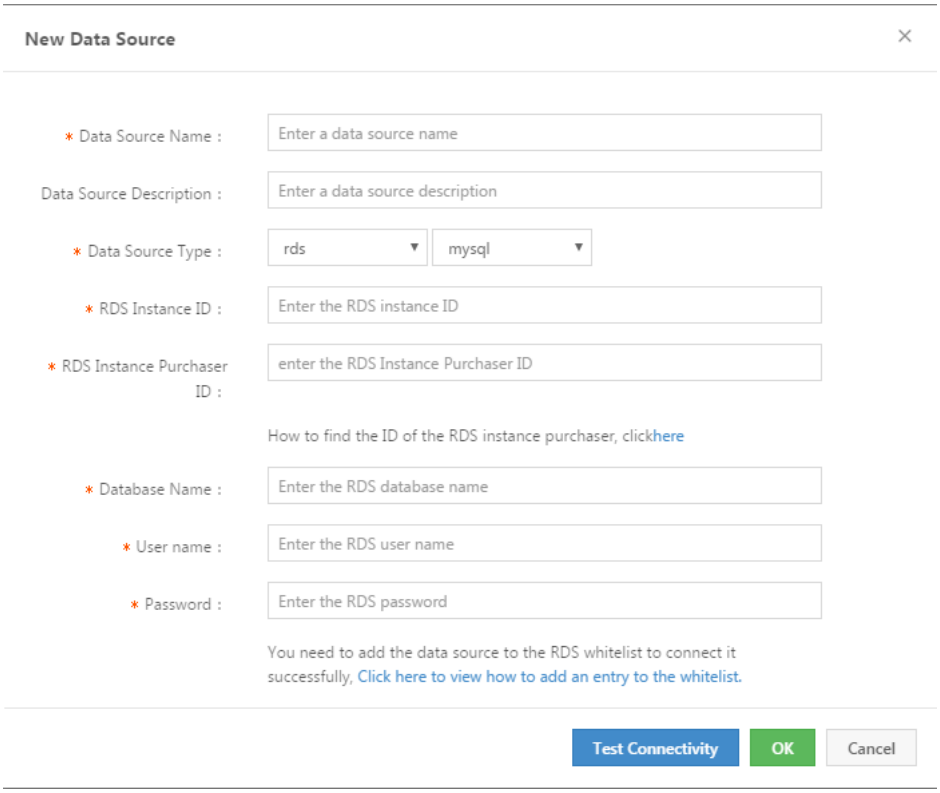
Step 2.1: Go to Alibaba Cloud Dataplus platform > Data IDE Kit > Console as a developer, click the **Enter Work Zone** in the action bar of the corresponding project.

Step 2.2: Click **Manage Projects** in the top menu bar, and then click **Manage Data Sources** in the left navigation bar.

Step 2.3: Click **New Data Source**.



Step 2.4: Fill in the configuration items in the “New Data Source” pop-up box.



Specific descriptions of the configuration items in the figure above are as follows:

- Data source name: A data source name may consist of letters, numbers, and underscores. It must begin with a letter or an underscore and cannot exceed 30 characters in length.
- Data source descriptions: A brief description of the data source. The description should not exceed 1,024 characters in length.
- Data source type: The data source type selected currently (RDS>MySQL>RDS).
- RDS instance ID: The ID of the MySQL data source RDS instance.

  RDS instance purchaser ID: The purchaser ID of the MySQL data source RDS instance.

  > [Note] If you have selected the JDBC form to configure the data source, the format of the JDBC connection information is: jdbc:mysql://IP:Port/database.

- Database name: The database name of the data source.
- User name/password: The user name and password of the database.

Step 2.5: Click **Test Connectivity**.

Step 2.6: If the test result is connected successfully, click the **Save** button to save the configuration information.

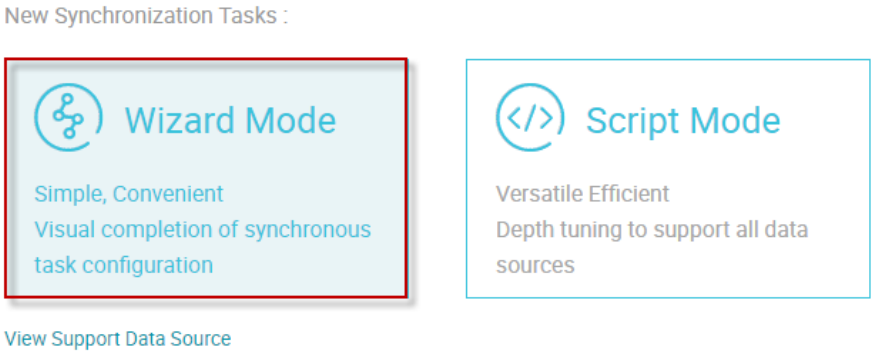
For detailed configurations of other types of data sources (MaxCompute, RDS, Oracle, FTP, ADS, OSS, OCS, and DRDS), see Data Source Configuration.

Step 3: Create a new job

Take the "wizard mode" new task as an example.

1. In the "data integration" interface, click on the left navigation bar to synchronize tasks;

   Click the "wizard mode" in the interface to get to the task configuration page.

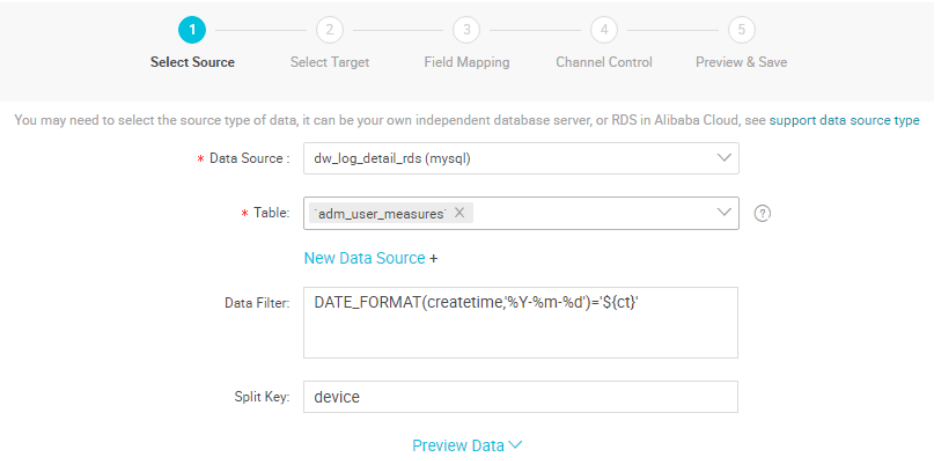


Step 4: Configure the data synchronization job

The synchronization job node includes five configuration items: **"Select Data Source and Target"**, **"Field Mapping"**, **"Channel Control"** and **"Preview & Save"**.

Step 4.1: Select the data source

Select Data Source(The data source has been created in Step 2), and then select the data table.



- Extraction Filtering: You can specify the WHERE filter based on the corresponding SQL syntax (You do not need to specify the WHERE keyword). The WHERE filter will be used as a condition of incremental synchronization.

[Description] The WHERE filter is used for source data filtering. The specified column, table, and WHERE filter are concatenated to create an SQL command for data extraction. The WHERE filter

can be used for full synchronization and incremental synchronization. Specific descriptions are as follows:

1) Full synchronization:Full synchronization is usually executed when data is imported for the first time. You do not need to configure the WHERE filter. You can set the WHERE filter limit to 10 to avoid a large data size during tests.

2) Incremental synchronization:In the actual service scenario, incremental synchronization usually synchronizes the data generated on the current day. Before compiling the WHERE filter, you usually need to first determine the field that describes the increment (timestamp) in the table. For example, if in Table A, the field that describes the increment is "creat_time", you need to compile "creat_time>$yesterday" in the WHERE filter and assign a value to the parameter in parameter configuration. For more usage of nested parameters, refer to the "System Scheduling Parameters" section.

If the data synchronization job is RDS/Oracle/MaxCompute, the splitting key configuration will be displayed on the page.

- Splitting key:**only supports integer fields.** During data reading, the data will be split based on the configured fields to achieve concurrent reading, improving data synchronization efficiency. The splitting key configuration item will only be displayed when the synchronization job is for importing RDS/Oracle data into MaxCompute.

[Note] If the source is a MySQL data source, the data synchronization job also supports database- and table-based data importing (on a premise that the table structure must be consistent, no matter whether the data is stored in the same database or different databases).

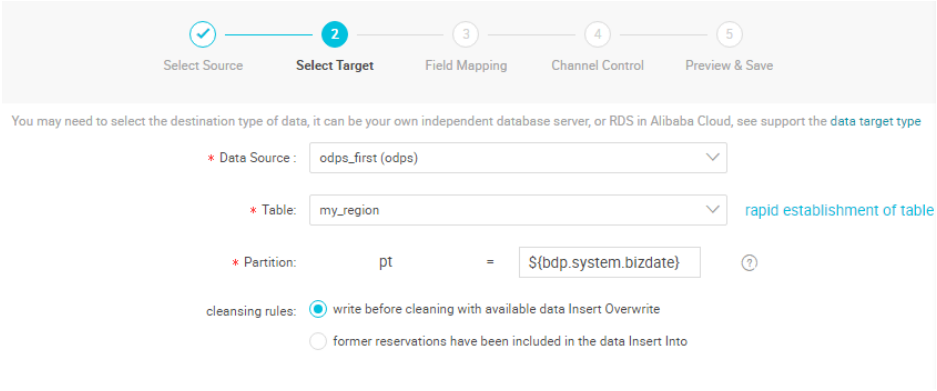
Database- and table-based data importing supports the following scenarios:

Multiple tables in the same database: Click "Search Table" to search for the tables and add the tables you want to synchronize.

Multiple tables in different databases: First click "Add" to select the source database, and then click "Search Table" to add the tables.

Step 4.2: Select the data target

Click "rapid establishment of table" and you will be able to convert the tabulation statements of the source table to DDL statements conforming to the MaxCompute SQL syntax to create a target table. After making the necessary selections, click "Next".



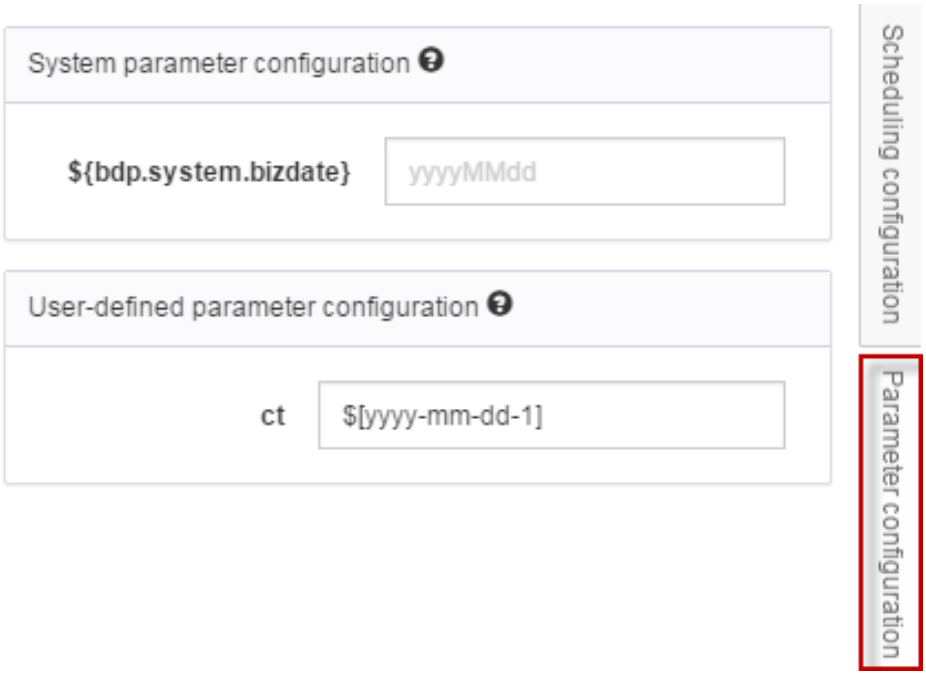
Partition information: Partitioning helps you to easily search for the special columns introduced by some data. By specifying the partition, you can quickly locate the desired data. Constant partitions and variable partitions are supported.

Clearing rules:

1) Clear existing data before writing: Before data importing, all the data in the table or partition should be cleared, which is equivalent to “Insert Overwrite”.

2) Keep existing data before writing: No data needs to be cleared before data importing. New data is always appended with each run, which is equivalent to “Insert into”.

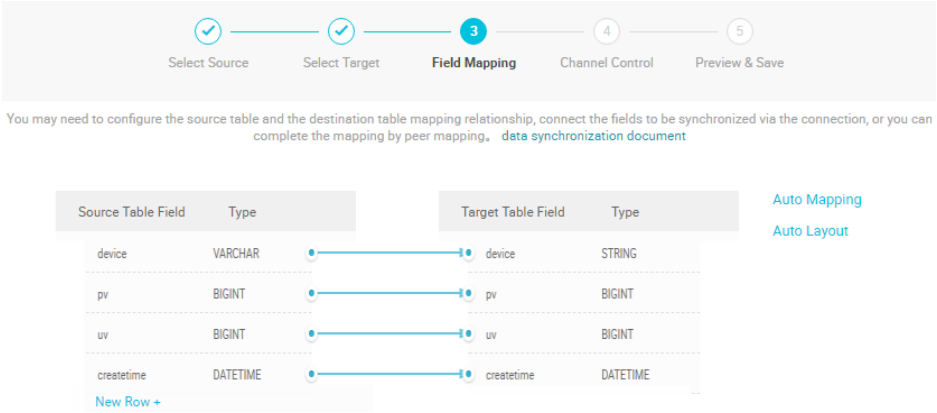
Assign values to parameters in the parameter configuration, as shown in the figure below:



Step 4.3: Field Mapping

You need to configure the field mapping relationships. The “Source Table Fields” on the left

correspond one to one with the "Target Table Fields" on the right.

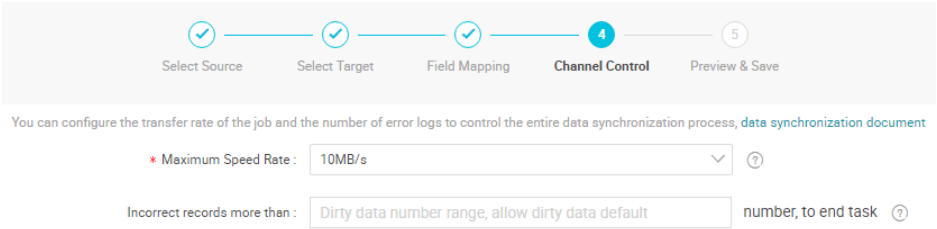


- Add/Delete: Click "Add a Line" to add a single field. Move and hover the cursor on a line above, and click the "Delete" icon, and you will delete the current field.

[Prompt] Writing method for user-defined variables and constants:To import a constant or variable to a field in the MaxCompute table, you only need to click the "Insert" button and enter the value of the constant or variable enclosed in single quotation marks. For example, for the '${yesterday}' variable, you can then assign a value to the variable using the parameter configuration component, such as yesterday=$[yyyymmdd]. For specific time parameters, see System Scheduling Parameters description.

Step 4.4: Channel Control

The **Channel Control** is used to configure the maximum speed of the job and the dirty data check rules, as shown in the figure:



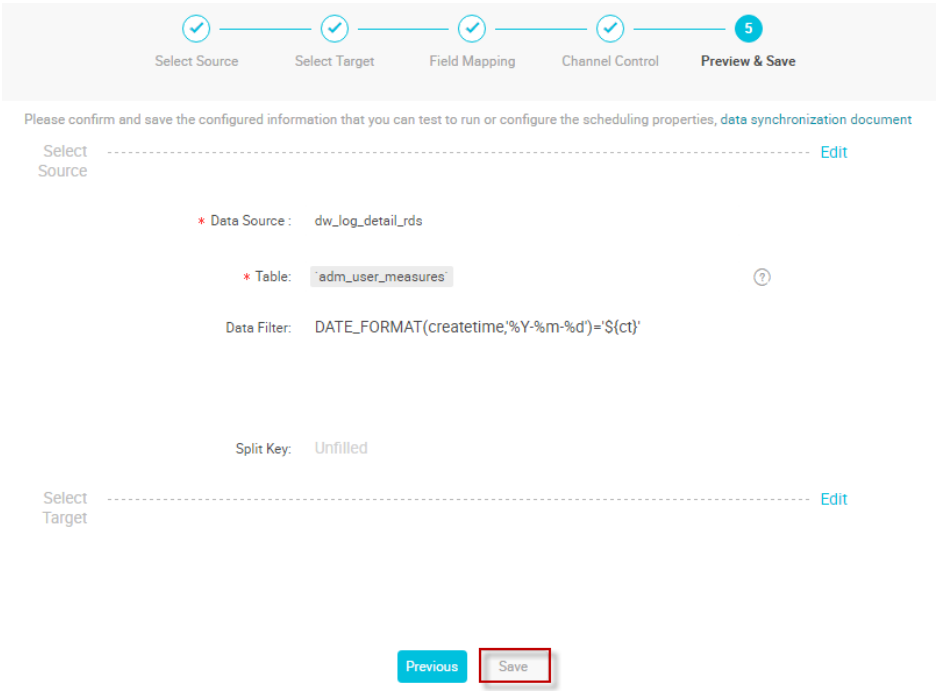
- The maximum speed of the job refers to the speed of the current data synchronization job, with a maximum value of 10 MB/s supported (The channel traffic measured value is the measured value of the data synchronization job, and does not represent the actual traffic of the network interface card).

Dirty data check rules (available for writing data to RDS and Oracle):

- When the number of error records (that is the volume of dirty data) exceeds the configured quantity, the data synchronization job ends.

Step 4.5: Preview & Save

When you complete the above configuration, click "next" to preview, if correct, click "save", as shown below:

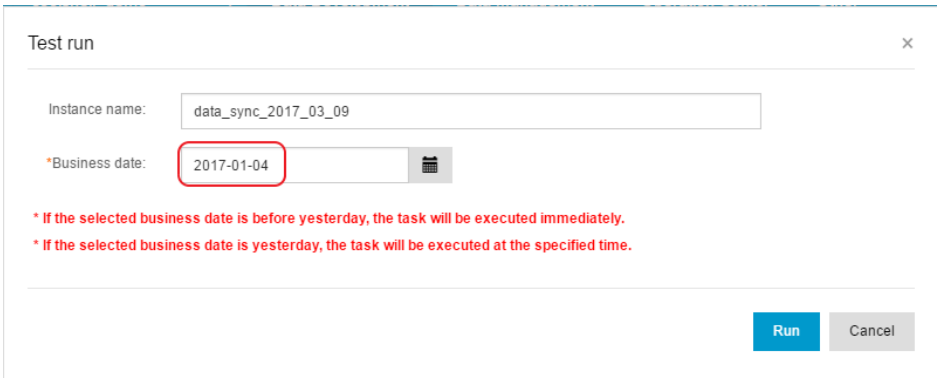


Step 5: Submit the data synchronization job and test the workflow

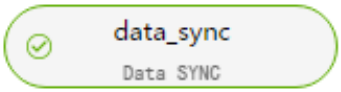
Step 5.1: Click the top menu bar to submit the job

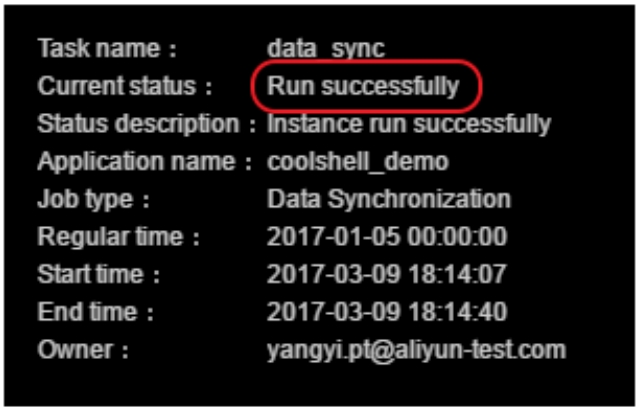
Step 5.2: After the job is submitted successfully, click Test Run

> [Description] Because some createtime values in the source table in this example are 2017-01-04, while the scheduling time parameters used in the configuration are $[yyyy-mm-dd-1] and ${bdp.system.bizdate}, we set the partition value of the target table to 20170104 to assign the value of 2017-01-04 to the createtime parameter in the test. The 2017-01-04 should be selected as the business time in the test, as shown in the figure below:



After the test task is triggered successfully, you can click "Go to O&M Center" to view the task progress.





Step 5.3: View the synchronized data

```
1 read my_region ;
```

Log | Results[1] ×

No.	device	pv	uv	createtime	pt
1	android	937	73	2017-01-04 20:51	20170104
2	iphone	428	31	2017-01-04 20:49	20170104
3	macintosh	830	107	2017-01-04 20:51	20170104
4	unknown	4124	444	2017-01-04 20:51	20170104
5	windows_pc	5650	649	2017-01-04 20:51	20170104

After a user is added to the project and granted tabulation permission, among other permissions, he or she can then perform operations on MaxCompute through the development kit. Since the operation objects in the underlying MaxCompute (input and output) are all tables, we should first create tables and partitions before processing the data. For detailed syntax on creating MaxCompute tables, see MaxCompute Introduction Documents.

Create a table

You can use the New Table function in the **New Script File** and **Data Management** modules in the Data IDE Kit to create a MaxCompute table.

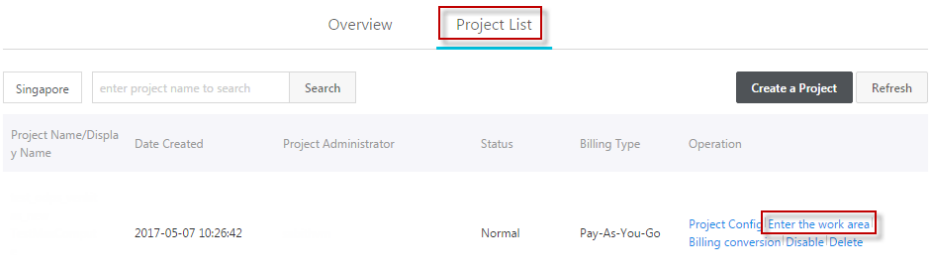
Taking, for example, the creation of a new table of tmall_user_brand (Tmall brand access log), the steps are as follows:

Tabulation statements:

```
DROP TABLE IF EXISTS tmall_user_brand;
CREATE TABLE tmall_user_brand (
user_id STRING COMMENT 'User ID',
brand_id STRING COMMENT 'Brand ID',
type STRING COMMENT 'Type of user actions of the brand: click-0, purchase-1, add to favorites-2, add to shopping cart-3',
visit_datetime STRING COMMENT 'Time of action'
)
COMMENT 'Tmall brand access log'
PARTITIONED BY (
dt STRING COMMENT 'Time range'
)
LIFECYCLE 10;
```

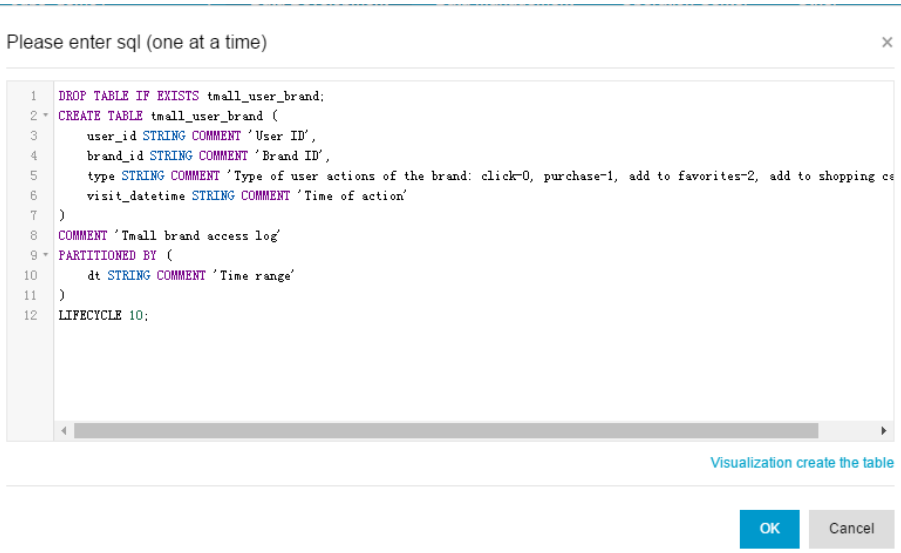
Method 1: Quick tabulation

Step 1: Go to Alibaba Cloud Dataplus platform > Data IDE Kit > Console as a developer, click the **Enter Work Zone** in the action bar of the corresponding project.



Step 2: Click New > New Table to pop up the New Table box.

Step 3: Fill in the tabulation statement, and click OK to complete the tabulation.



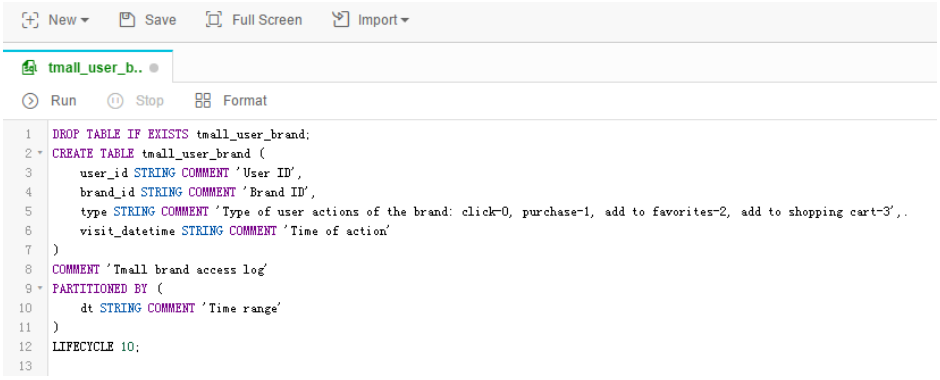
Method 2: Tabulation through script files

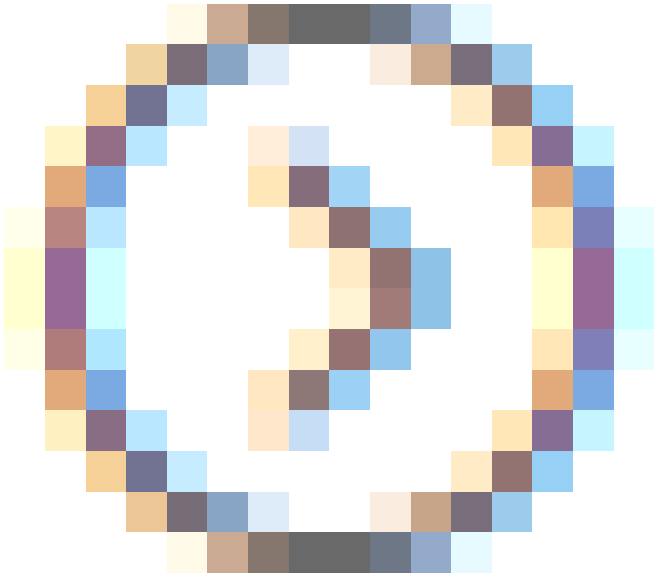
Step 1: Go to Alibaba Cloud Dataplus platform > Data IDE Kit > Console as a developer, click the

Enter Work Zone in the action bar of the corresponding project.

Step 2: Create a script file. Click Data Development in the top menu bar, click "New" to create a new script, or you can click the "New Script" task box directly.

Step 3: Edit the tabulation statement.





Step 4: Click the button to run the tabulation DDL statement.

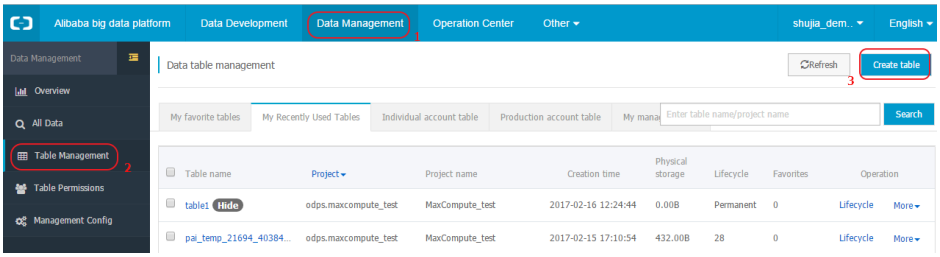
Step 5: If the statement is run successfully, it will indicate that the table has been created successfully.

Method 3: Tabulation through Data Management module

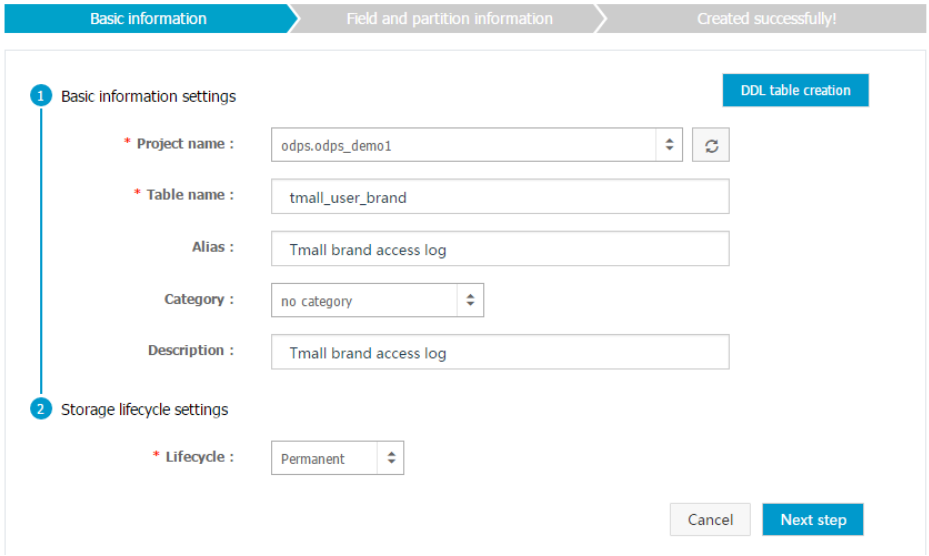
Step 1: Enter the Data IDE Kit.

Step 2: Click **Data Management** in the top menu bar and navigate to **Manage Data Tables**.

Step 3: Click **New Table**.



Step 4: Fill in the basic information configuration items on the New Table page.



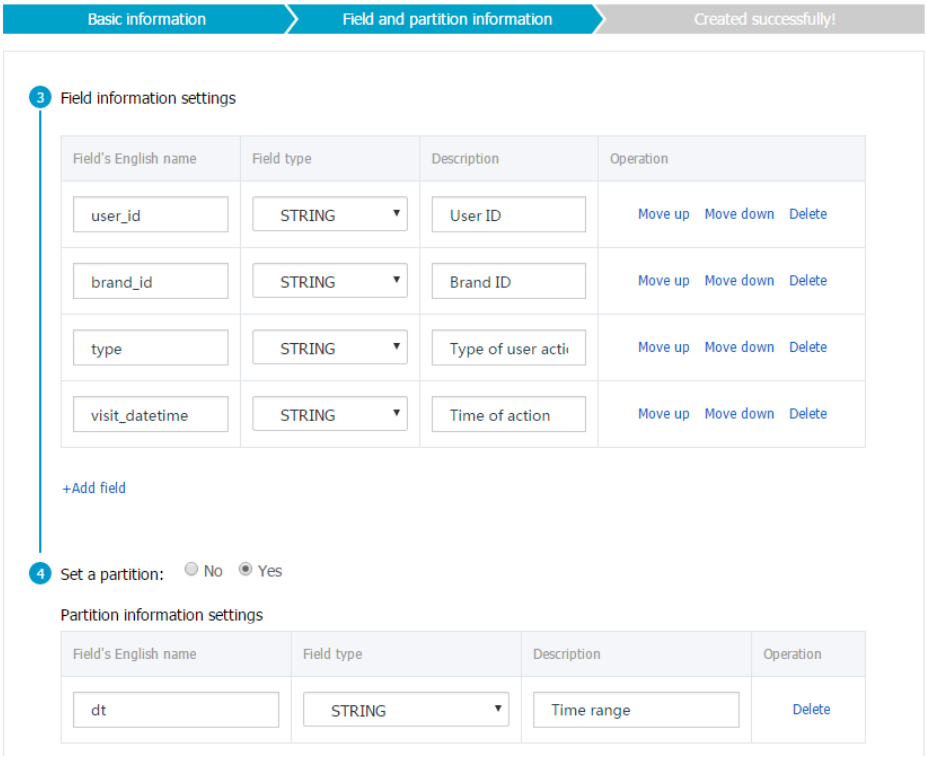
Specific descriptions on the configuration items on the Basic Info page are as follows:

- Project Name: The list shows the MaxCompute projects that the user is currently in.
- Table Name: A table name may contain letters, numbers, and underscores.
- Alias: The Chinese name of the table.
- Category: The category of the current table. Up to four category levels are supported. For details about the configuration of existing category navigation, see Category Navigation Configuration.

- Description: A brief description of the current table.
- Lifecycle: The lifecycle function of MaxCompute. Data in the table (or partition) that has not been updated within the period of time specified by "Lifecycle" (in the unit of days) will be cleared. Five options are available, including "1 day", "7 days", "32 days", "Permanent", and "User-defined".

Step 5: Click **Next**.

Step 6: Fill in the configuration items of the field and partition information on the New Table page.



The configuration items on the field and partition pages are described as follows:

- English Name of the Field: The English name of a field, which may contain letters, numbers, and underscores.
- Field Type: The MaxCompute data type (string, bigint, double, datetime, or boolean).
- Description: Detailed description of a field.
- Action: The options include "Move Up", "Move Down", and "Delete".
- Whether to Set Partitions: If you select to set partitions, you need to configure the partitioning key information. String and bigint data types are supported.

[Description] The sensitivity level tag of the field, with a value range of 0-9, indicating the sensitivity level from low to high. After a new user enters the project, the default security permission tag is 0, so the user can only view fields with a sensitivity level of 0 in the table. The user needs to get the authorization for viewing fields with a higher sensitivity level (that is, project members can only view data with a sensitivity level not higher than the member's security permission tag level).

Step 7: Click **Submit**.

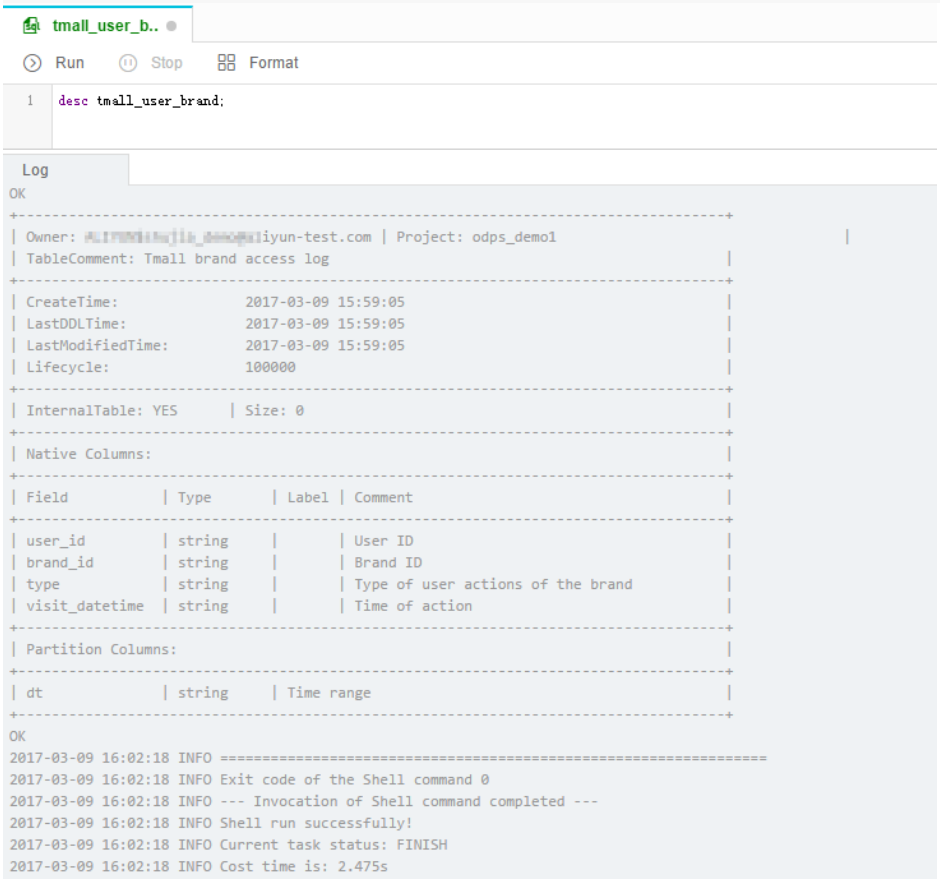
After the newly created table has been submitted successfully, the system will automatically jump back to the Manage Data Tables page. Click **My Managed Tables** and you will be able to see the newly created table.

Get table information

After the table is created successfully, we can get the table information by writing the following command in the script file and then clicking the "Run" button:

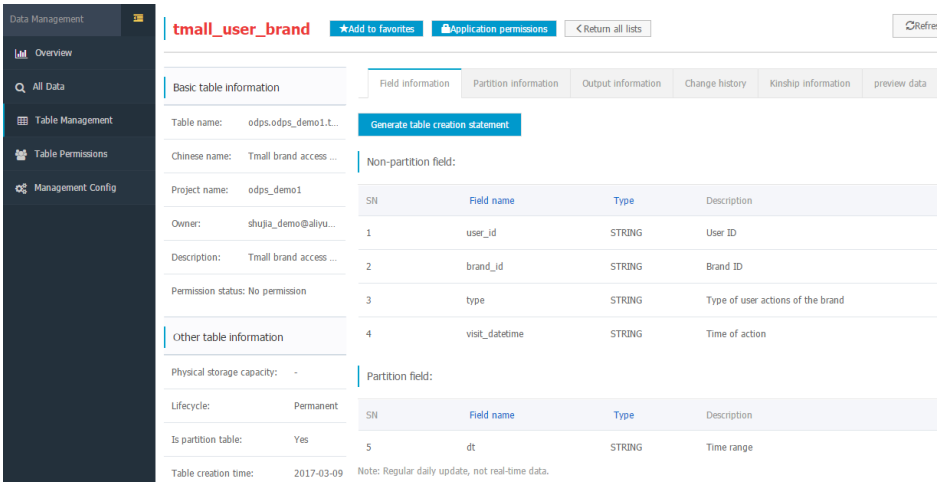
Method 1: Query through the script file

```
desc <tablename>
```



Method 2: Query through the table details

Click the table name to enter the table details page:



Delete a table

The operations for deleting a table are the same as for creating a table. You can delete a table by compiling DDL statements in the script file, or you can delete the table in the **Data Management** module.

Method 1: Query through the script file

Execute the table drop command through the SQL statement.

```
DROP TABLE [IF EXISTS] table_name;
```

Method 2: Query through Manage Data Tables

> [Description] You can delete a table on the **Data Management** > **Manage Data Tables** > **My Managed Tables** page.