ALIBABA CLOUD

Alibaba Cloud

DataWorks

Tutorials

Document Version: 20220608

Alibaba Cloud

Legal disclaimer

Alibaba Cloud reminds you to carefully read and fully understand the terms and conditions of this legal disclaimer before you read or use this document. If you have read or used this document, it shall be deemed as your total acceptance of this legal disclaimer.

1. You shall download and obtain this document from the Alibaba Cloud website or other Alibaba Cloud-authorized channels, and use this document for your own legal business activities only. The content of this document is considered confidential information of Alibaba Cloud. You shall strictly abide by the confidentiality obligations. No part of this document shall be disclosed or provided to any third party for use without the prior written consent of Alibaba Cloud.

2. No part of this document shall be excerpted, translated, reproduced, transmitted, or disseminated by any organization, company or individual in any form or by any means without the prior written consent of Alibaba Cloud.

3. The content of this document may be changed because of product version upgrade, adjustment, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and an updated version of this document will be released through Alibaba Cloud-authorized channels from time to time. You should pay attention to the version changes of this document as they occur and download and obtain the most up-to-date version of this document from Alibaba Cloud-authorized channels.

4. This document serves only as a reference guide for your use of Alibaba Cloud products and services. Alibaba Cloud provides this document based on the "status quo", "being defective", and "existing functions" of its products and services. Alibaba Cloud makes every effort to provide relevant operational guidance based on existing technologies. However, Alibaba Cloud hereby makes a clear statement that it in no way guarantees the accuracy, integrity, applicability, and reliability of the content of this document, either explicitly or implicitly. Alibaba Cloud shall not take legal responsibility for any errors or lost profits incurred by any organization, company, or individual arising from download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, take responsibility for any indirect, consequential, punitive, contingent, special, or punitive damages, including lost profits arising from the use or trust in this document (even if Alibaba Cloud has been notified of the possibility of such a loss).

5. By law, all the contents in Alibaba Cloud documents, including but not limited to pictures, architecture design, page layout, and text description, are intellectual property of Alibaba Cloud and/or its affiliates. This intellectual property includes, but is not limited to, trademark rights, patent rights, copyrights, and trade secrets. No part of this document shall be used, modified, reproduced, publicly transmitted, changed, disseminated, distributed, or published without the prior written consent of Alibaba Cloud and/or its affiliates. The names owned by Alibaba Cloud shall not be used, published, or reproduced for marketing, advertising, promotion, or other purposes without the prior written consent of Alibaba Cloud. The names owned by Alibaba Cloud include, but are not limited to, "Alibaba Cloud", "Aliyun", "HiChina", and other brands of Alibaba Cloud and/or its affiliates, which appear separately or in combination, as well as the auxiliary signs and patterns of the preceding brands, or anything similar to the company names, trade names, trademarks, product or service names, domain names, patterns, logos, marks, signs, or special descriptions that third parties identify as Alibaba Cloud and/or its affiliates.

6. Please directly contact Alibaba Cloud for any errors of this document.

Document conventions

Style	Description	Example
⚠ Danger	A danger notice indicates a situation that will cause major system changes, faults, physical injuries, and other adverse results.	⚠ **Danger:** Resetting will result in the loss of user configuration data.
🔔 Warning	A warning notice indicates a situation that may cause major system changes, faults, physical injuries, and other adverse results.	🔔 **Warning:** Restarting will cause business interruption. About 10 minutes are required to restart an instance.
🔈 Notice	A caution notice indicates warning information, supplementary instructions, and other content that the user must understand.	🔈 **Notice:** If the weight is set to 0, the server no longer receives new requests.
? Note	A note indicates supplemental instructions, best practices, tips, and other content.	? **Note:** You can use Ctrl + A to select all files.
>	Closing angle brackets are used to indicate a multi-level menu cascade.	Click **Settings**> **Network**> **Set network type**.
Bold	Bold formatting is used for buttons , menus, page names, and other UI elements.	Click **OK**.
`Courier font`	Courier font is used for commands	Run the `cd /d C:/window` command to enter the Windows system folder.
Italic	Italic formatting is used for parameters and variables.	`bae log list --instanceid` *Instance_ID*
[] or [a|b]	This format is used for an optional value, where only one item can be selected.	`ipconfig [-all|-t]`
{} or {a|b}	This format is used for a required value, where only one item can be selected.	`switch {active|stand}`

Table of Contents

1.DataWorks for MaxCompute Workshop 05
1.1. Workshop introduction 05
1.2. Prepare the environment 05
1.3. Collect data 07
1.4. Process data 26
1.5. Configure rules to monitor data quality 34
1.6. Visualize data on a dashboard 39
1.7. Use Function Studio to develop a UDF 41
2.DataWorks for EMR Workshop 44
2.1. Prepare the environment 44
2.2. Collect data 48
2.3. Process data 59
2.4. Collect and view metadata 64
2.5. Configure rules to monitor data quality 65
3.Automatically identify users who steal electricity 68
3.1. Overview 68
3.2. Prepare the environment 68
3.3. Prepare data 71
3.4. Process data 77
3.5. Build a data model 81
4.Integrate and use CDH 85

1.DataWorks for MaxCompute Workshop

1.1. Workshop introduction

This topic describes the design concepts and core features of DataWorks to help you understand DataWorks.

Workshop overview

Duration: 2 hours of online learning.

Audience: new and existing users of DataWorks, such as Java engineers and product operation staff. If you are familiar with standard SQL, you can master basic features of DataWorks without knowing much about the principles of data warehouses and MaxCompute. However, we recommend that you take DataWorks tutorials to learn more about the basic concepts and features of DataWorks. For more information, see What is DataWorks?

Objective: simulates the scenario where a large number of logs are analyzed. After you complete the workshop, you can understand the main features of DataWorks. You can follow the workshop to use the MaxCompute compute engine to perform common data operations, such as data collection, data analytics, and node O&M.

DataWorks overview

DataWorks is a one-stop big data R&D platform with features including data integration, data modeling, data analytics, O&M and monitoring, data management, data security, and data quality. In addition, it is integrated with Machine Learning Platform for AI (PAI) and optimizes the process from big data development to data mining and machine learning.

Consultation

If you have any questions during the workshop, join the DingTalk group for consultation.

1.2. Prepare the environment

To make sure that you can complete the workshop, you must activate MaxCompute and DataWorks for your Alibaba Cloud account.

Prerequisites

- An Alibaba Cloud account is created.
- Real-name verification is completed.

Context

The following Alibaba Cloud services are used in this workshop:

- MaxCompute
- DataWorks

Activate MaxCompute

ⓘ **Note** If you have already activated MaxCompute, skip this step and directly create a workspace in DataWorks.

1. Go to the Alibaba Cloud official website, click **Log In** in the upper-right corner, and then enter your account name and password.
2. Move the pointer over **Products** in the top navigation bar and choose **Analytics > Data Computing > MaxCompute** to go to the product page of MaxCompute.
3. Click **Activate Now**.
4. On the buy page of MaxCompute, select a region, read and agree to the service agreement, and then click **Confirm Order and Pay**.

   ⓘ **Note**

   - By default, DataWorks Basic Edition and the standard pay-as-you-go resource package of MaxCompute are provided on the buy page.
   - The project management, query, and editing features of MaxCompute are integrated into the features of DataWorks. Therefore, you must activate DataWorks at the same time. DataWorks Basic Edition is free of charge. You are charged only if you use Data Integration or run scheduled nodes.
   - When you activate MaxCompute, you must consider other Alibaba Cloud services that are available in each region. For example, you must consider the region where your Elastic Compute Service (ECS) instance resides and the region where the data resides.

Create a DataWorks workspace

ⓘ **Note** Data resources provided for this workshop are all stored in the China (Shanghai) region. Therefore, we recommend that you create a workspace in the China (Shanghai) region. Otherwise, the network connectivity test fails when you create a connection.

1. Log on to the DataWorks console by using your Alibaba Cloud account.
2. On the **Overview** page, click **create Workspace** in the Frequently Used Workspaces section on the right.

   You can also click **Workspaces** in the left-side navigation pane and click **Create Workspace** on the page that appears.
3. In the **Create Workspace** panel, set the parameters in the **Basic Settings** step and click **Next**.

   ⓘ **Note** In this tutorial, a workspace in standard mode is created.
4. In the **Select Engines and Services** step, select MaxCompute and click **Next**.

   DataWorks is now available as a commercial service. If you have not activated DataWorks in a region, activate it before you create a workspace in the region. By default, the following services are selected when you create a workspace: **Data Integration**, **DataStudio**, **Operation Center**, and **Data Quality**.
5. In the **Engine Details** step, set the parameters for the selected compute engines.

Engine	Parameter	Description
MaxCompute	**Instance Display Name**	The display name of the instance can be up to 27 characters in length. It must start with a letter and can contain only letters, underscores (_), and digits.
	Resource Group	The quotas of computing resources and disk space for the compute engine instance.
	MaxCompute Data Type Edition	This configuration takes effect within 5 minutes. For more information, see Data type editions.
	MaxCompute Project Name	By default, the name is the same as that of the DataWorks workspace.
	Account for Accessing MaxCompute	Valid values: **Alibaba Cloud Account** and **Node Owner**.

6. Click **Create Workspace**.

   After the workspace is created, you can view information about the workspace on the **Workspaces** page.

1.3. Collect data

This topic describes how to use DataWorks to collect logs to MaxCompute.

Context

In this workshop, you must add an Object Storage Service (OSS) bucket and an ApsaraDB RDS instance as data sources from which you want to read data. You must also create tables to which you want to write data.

Note

- You can use the data sources that are prepared for you in this workshop. You can also use your own data sources.
- The prepared data sources reside in the **China (Shanghai)** region. We recommend that you use a workspace in the China (Shanghai) region to make sure that the prepared data sources are accessible when you add these data sources.

Add an OSS data source

1. Go to the **Data Source** page.

   i. Log on to the DataWorks console.

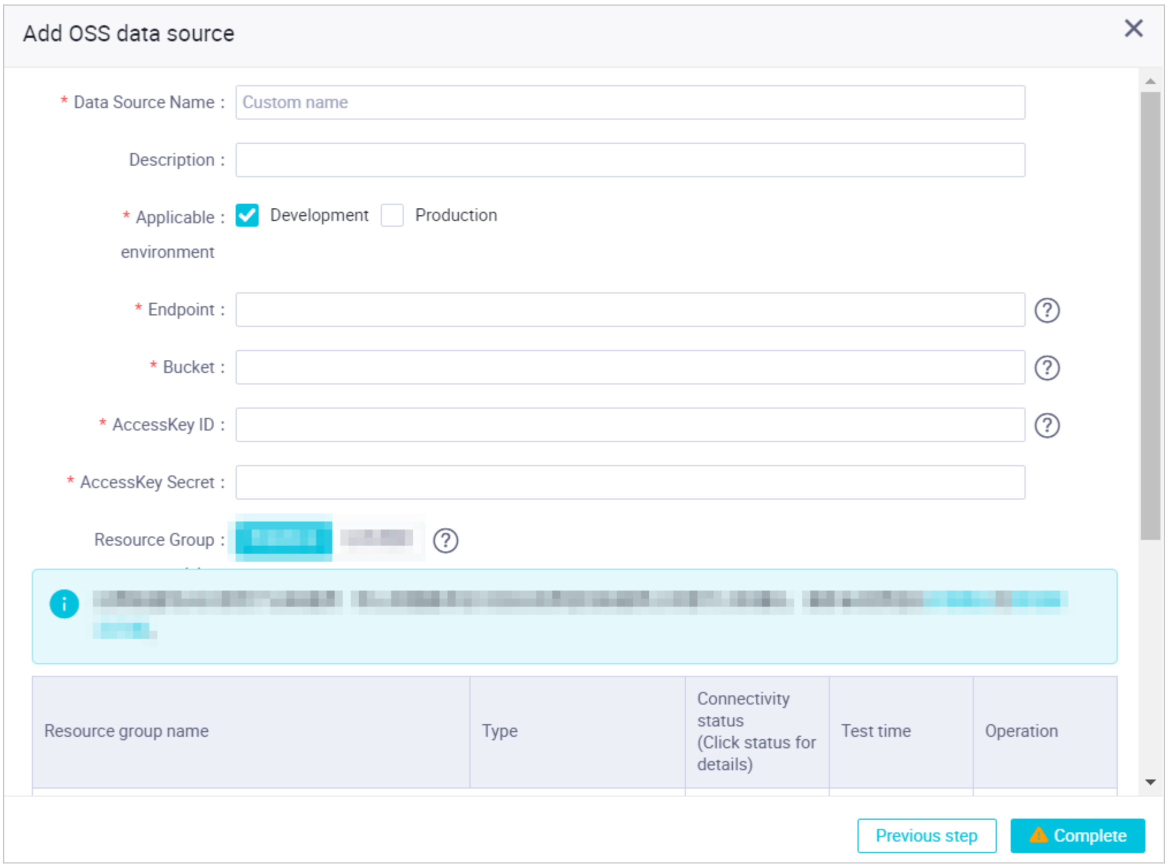
   ii. In the left-side navigation pane, click **Workspaces**.

iii. On the Workspaces page, find the workspace to which you want to add a data source and click **Data Integration** in the Actions column.

If you are using another service of DataWorks, click the ☰ icon in the upper-left corner and choose **All Products** > **Data Aggregation** > **Data Integration** to go to the **Data Integration** page.

iv. In the left-side navigation pane, choose **Data Source** > **Data Sources**. The **Data Source** page in Workspace Management appears.

2. On the **Data Source** page, click **Add data source** in the upper-right corner.
3. In the **Add data source** dialog box, click **OSS**.
4. In the **Add OSS data source** dialog box, set the parameters based on your business requirements.



Add OSS data source

* Data Source Name : Custom name

Description :

* Applicable environment : ☑ Development □ Production

* Endpoint :

* Bucket :

* AccessKey ID :

* AccessKey Secret :

Resource Group :

Resource group name	Type	Connectivity status (Click status for details)	Test time	Operation

Parameter	Description
Data Source Name	The name of the data source. Enter oss_workshop_log.
Data source description	The description of the data source.

Parameter	Description
Environment	The environment in which the data source is used. Select **Development**. Note After you add a data source in the development environment, add the same data source in the production environment by selecting **Production**. Otherwise, an error is reported when a node that uses this data source is run in the production environment.
Endpoint	The OSS endpoint. Enter `http://oss-cn-shanghai-internal.aliyuncs.com`.
Bucket	The name of the OSS bucket. Enter new-dataworks-workshop.
AccessKey ID	The AccessKey ID that is used to connect to OSS. Enter LTAI4FvGT3iU4xjKotpU****.
AccessKey Secret	The AccessKey secret that is used to connect to OSS. Enter 9RSUoRmNxpRC9EhC4m9PjuG7Jzy7px.

5. Find the desired resource group in the resource group list in the lower part of the dialog box and click **Test connectivity** in the Actions column.

   A synchronization node can use only one type of resource group. To ensure that your synchronization nodes can be normally run, you must test the connectivity of all the resource groups for Data Integration on which your synchronization nodes will be run. If you want to test the connectivity of multiple resource groups for Data Integration at a time, select the resource groups and click **Batch test connectivity**. For more information, see Select a network connectivity solution.

   Note

   - By default, the resource group list displays only exclusive resource groups for Data Integration. To ensure the stability and performance of data synchronization, we recommend that you use exclusive resource groups for Data Integration.
   - If you want to test the network connectivity between the shared resource group or a custom resource group and the data source, click **Advanced** below the resource group list. In the **Warning** message, click **Confirm**. Then, all available shared and custom resource groups appear in the resource group list.

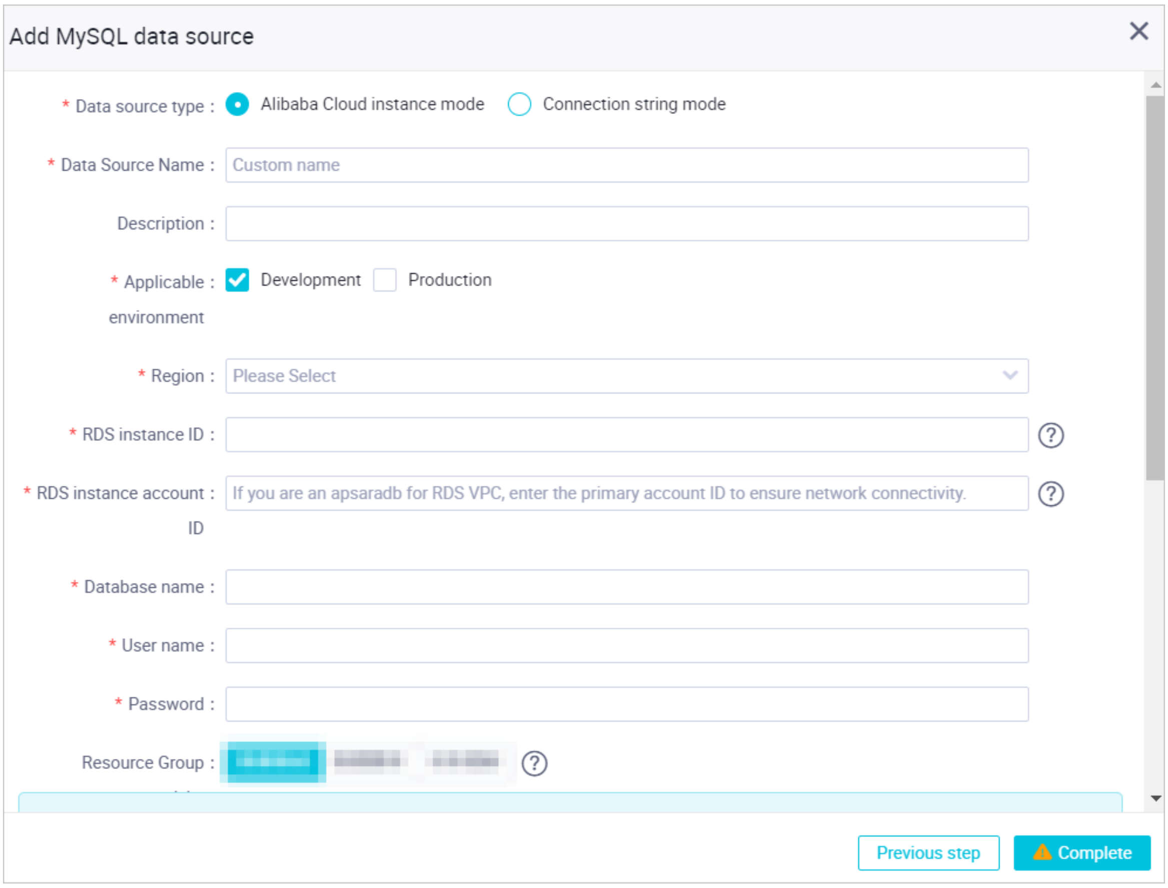
6. After the connection passes the connectivity test, click **Complete**.

   Note

   - If the connectivity test fails, check whether the AccessKey ID and AccessKey secret that you entered are correct and whether the DataWorks workspace is in the China (Shanghai) region.
   - If OSS cannot be accessed by using the internal endpoint, use the public endpoint.

Add an ApsaraDB RDS data source

1. On a service page of DataWorks, click the ☰ icon in the upper-left corner and choose **All Products > Data Aggregation > Data Integration**.
2. On the page that appears, choose **Data Source > Data Sources**. The **Data Source** page in **Workspace Management** appears.
3. On the **Data Source** page, click **Add data source** in the upper-right corner.
4. In the **Add data source** dialog box, click **MySQL**.
5. In the **Add MySQL data source** dialog box, set the parameters based on your business requirements.



Add MySQL data source

* Data source type : ● Alibaba Cloud instance mode ○ Connection string mode

* Data Source Name : Custom name

Description :

* Applicable environment : ☑ Development □ Production

* Region : Please Select

* RDS instance ID :

* RDS instance account ID : If you are an apsaradb for RDS VPC, enter the primary account ID to ensure network connectivity.

* Database name :

* User name :

* Password :

Resource Group : [redacted]

Parameter	Description
Data source type	The mode in which the data source is added. Select **Alibaba Cloud instance mode**.
Data Source Name	The name of the data source. Enter rds_workshop_log.
Data source description	The description of the data source. Enter RDS user information synchronization.

Parameter	Description
Environment	The environment in which the data source is used. Select **Development**. Note After you add a data source in the development environment, add the same data source in the production environment by selecting **Production**. Otherwise, an error is reported when a node that uses this data source is run in the production environment.
Region	The region where the ApsaraDB RDS instance resides.
RDS instance ID	The ID of the ApsaraDB RDS instance. Enter rm-bp1z69dodhh85z9qa.
RDS instance account ID	The ID of the Alibaba Cloud account that is used to purchase the ApsaraDB RDS instance. Enter 1156529087455811.
Default Database Name	The name of the ApsaraDB RDS database. Enter workshop.
User name	The username that is used to connect to the database. Enter workshop.
Password	The password that is used to connect to the database. Enter workshop#2017.

6. Find the desired resource group in the resource group list in the lower part of the dialog box and click **Test connectivity** in the Actions column.

   A synchronization node can use only one type of resource group. To ensure that your synchronization nodes can be normally run, you must test the connectivity of all the resource groups for Data Integration on which your synchronization nodes will be run. If you want to test the connectivity of multiple resource groups for Data Integration at a time, select the resource groups and click **Batch test connectivity**. For more information, see Select a network connectivity solution.

   Note

   - By default, the resource group list displays only exclusive resource groups for Data Integration. To ensure the stability and performance of data synchronization, we recommend that you use exclusive resource groups for Data Integration.
   - If you want to test the network connectivity between the shared resource group or a custom resource group and the data source, click **Advanced** below the resource group list. In the **Warning** message, click **Confirm**. Then, all available shared and custom resource groups appear in the resource group list.

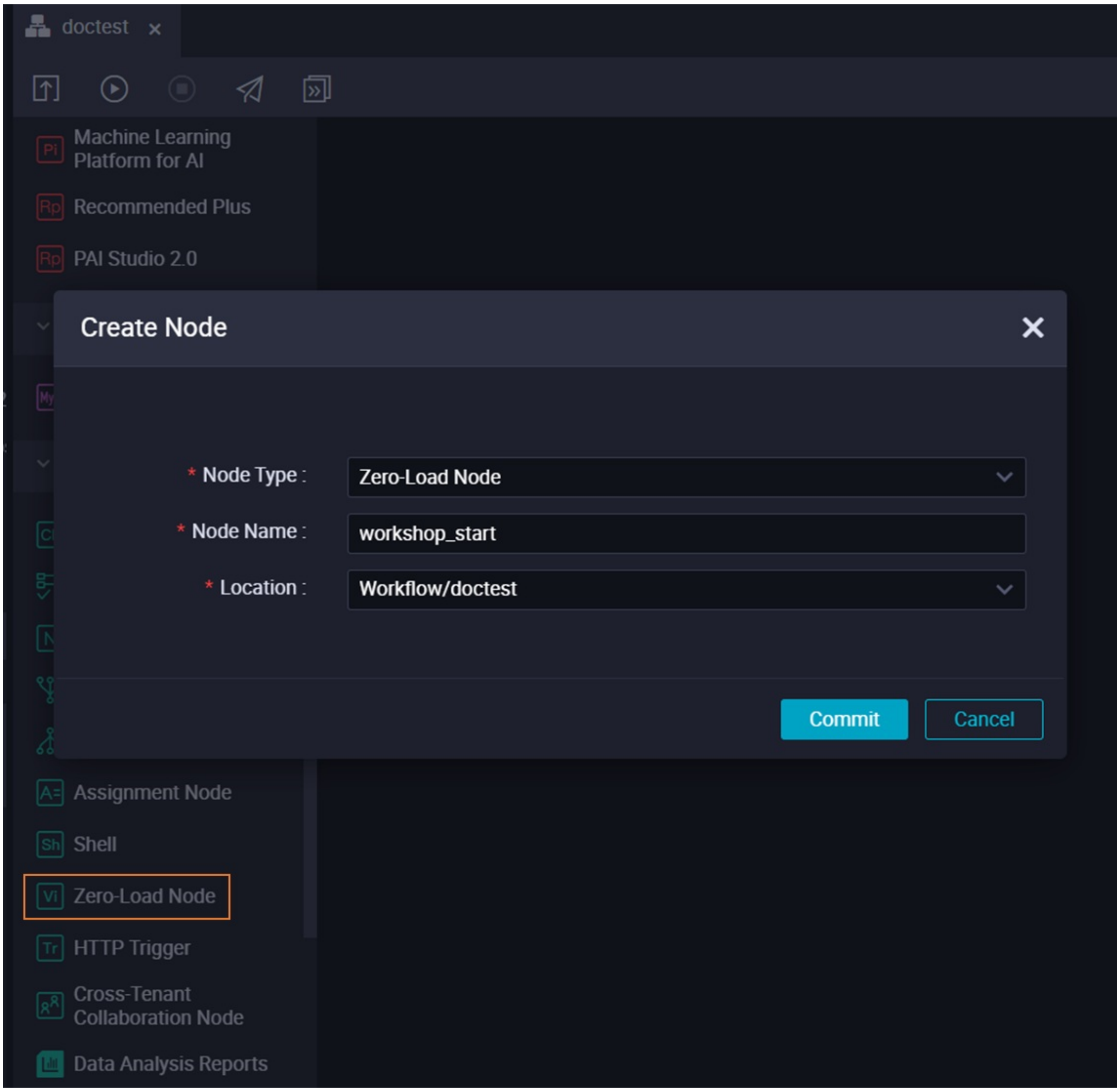
7. After the data source passes the connectivity test, click **Complete**.

Create a workflow

1. Click the ≡ icon in the upper-left corner of the Data Integration page and choose **All Products** > **Data Development** > **DataStudio**.
2. In the **Scheduled Workflow** pane, right-click **Business Flow** and select **Create Workflow**.
3. In the **Create Workflow** dialog box, set the **Workflow Name** and **Description** parameters.

> **Notice** The workflow name can be a maximum of 128 characters in length and can contain letters, digits, underscores (_), and periods (.).

4. Click **Create**.
5. Double-click the new workflow to go to the workflow configuration tab. Drag **Zero-Load Node** in the General section to the canvas on the right.
6. In the **Create Node** dialog box, set the **Node Name** parameter to workshop_start and click **Commit**.



Create Node

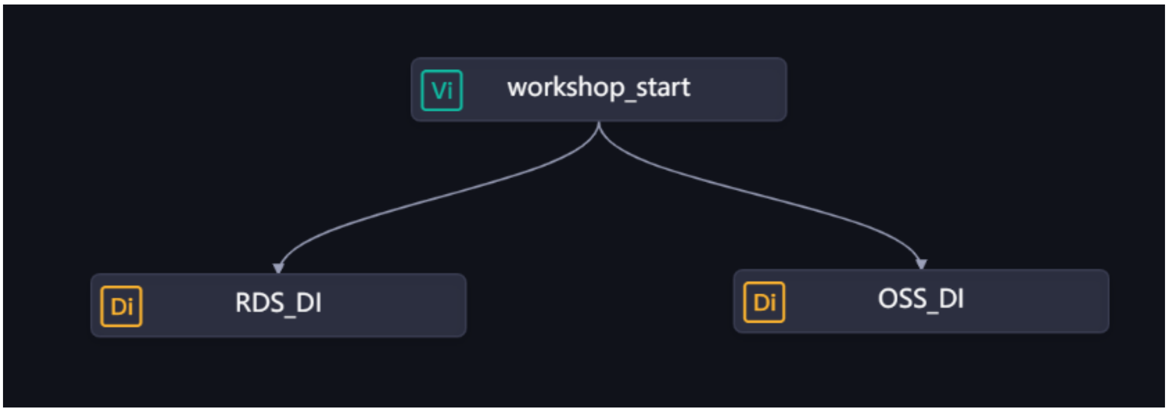
* Node Type : Zero-Load Node

* Node Name : workshop_start

* Location : Workflow/doctest

Drag **Batch Synchronization** in the Data Integration section to the canvas on the right to create two batch sync nodes named oss_synchronization and rds_synchronization.

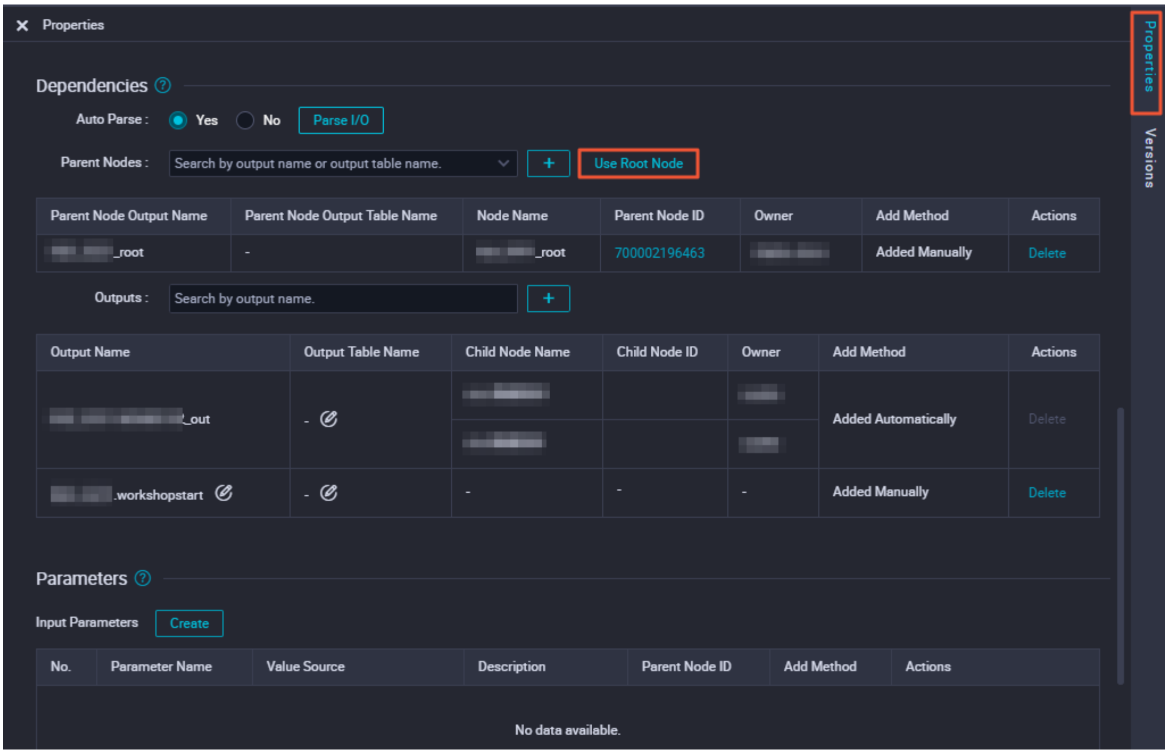
7. Drag directed lines to configure the workshop_start node as the ancestor node of the two batch sync nodes.



Configure the workshop_start node

1. In the **Scheduled Workflow** pane, double-click the workshop_start node in the workflow. On the node configuration tab, click the **Properties** tab in the right-side navigation pane.
2. In the **Dependencies** section, click **Add Root Node** to configure the root node of the workspace as the ancestor node of the workshop_start node.

   In the latest version of DataWorks, each node must have its ancestor and descendant nodes. Therefore, you must configure an ancestor node for the workshop_start node. In this example, the root node of the workspace is configured as the ancestor node of the workshop_start node. The root node of the workspace is named in the Workspace name_root format.



Dependencies

Auto Parse : ● Yes ○ No Parse I/O

Parent Nodes : Search by output name or output table name. + Use Root Node

Parent Node Output Name	Parent Node Output Table Name	Node Name	Parent Node ID	Owner	Add Method	Actions
[redacted]_root	-	[redacted]_root	700002196463	[redacted]	Added Manually	Delete

Outputs : Search by output name. +

Output Name	Output Table Name	Child Node Name	Child Node ID	Owner	Add Method	Actions
[redacted]_out	-	[redacted] [redacted]		[redacted] [redacted]	Added Automatically	Delete
[redacted].workshopstart	-	-	-	-	Added Manually	Delete

Parameters

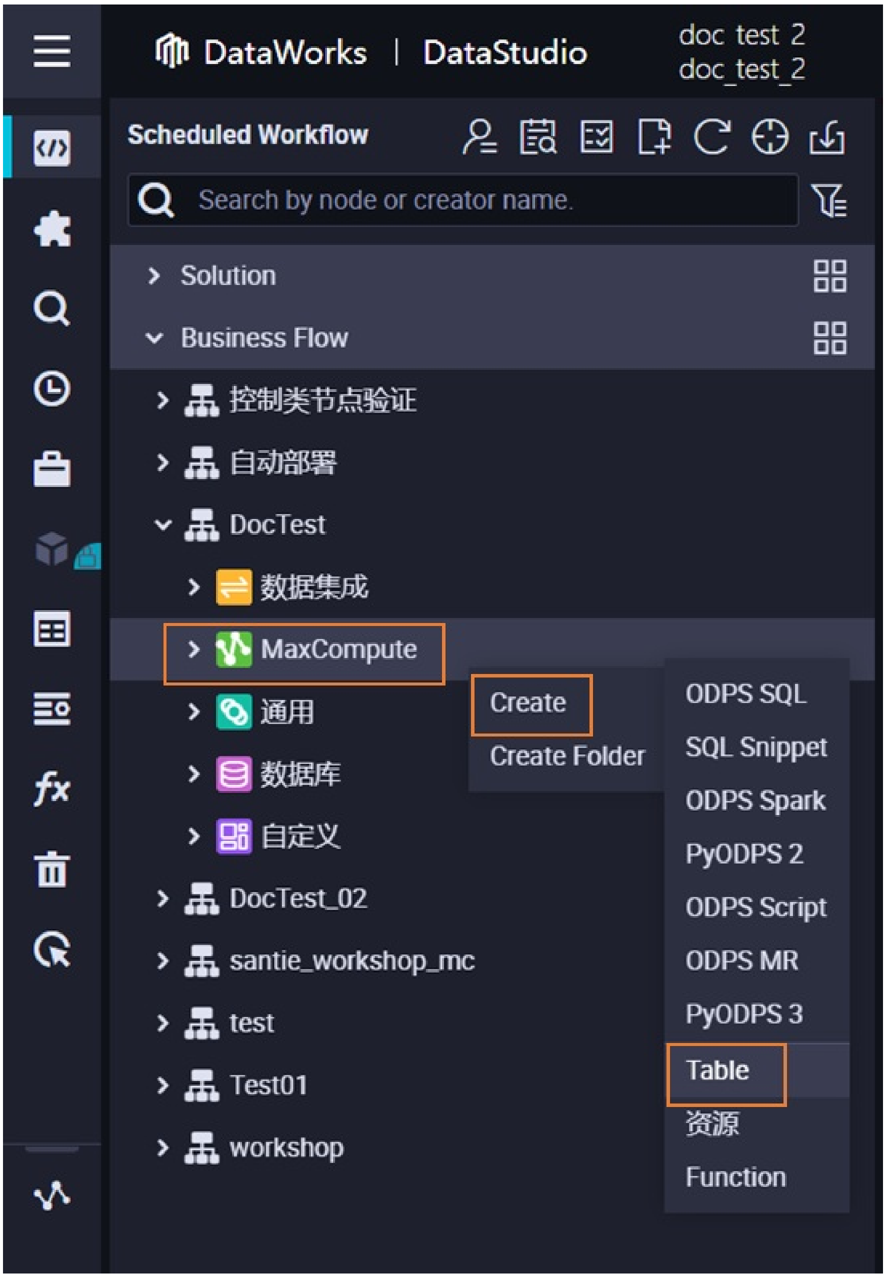
Input Parameters Create

No.	Parameter Name	Value Source	Description	Parent Node ID	Add Method	Actions
No data available.						

3. Click the 💾 icon in the top toolbar.

Create tables to which you want to write data

1. In the **Scheduled Workflow** pane, click the new workflow, right-click **MaxCompute**, and then choose **Create** > **Table**.



2. In the **Create Table** dialog box, set the **Table Name** parameter and click **Create**.

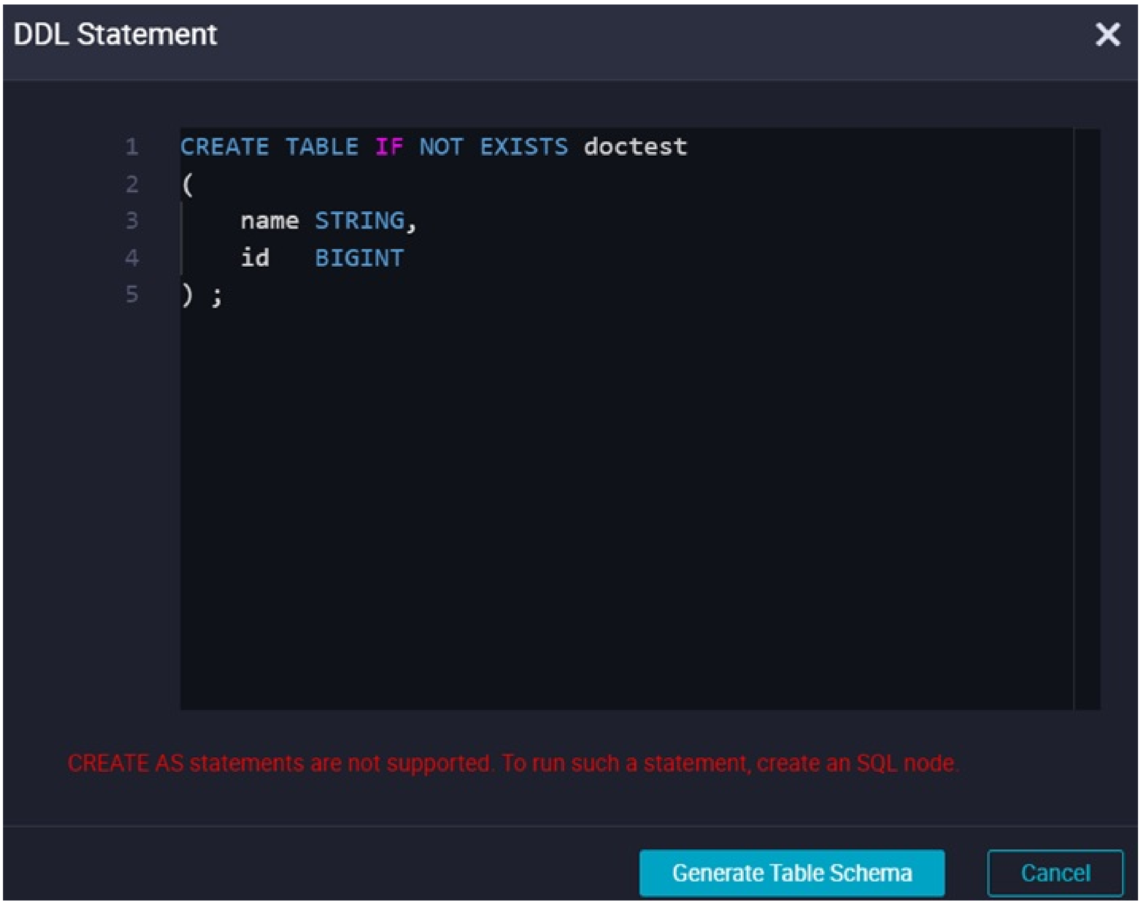
   In this workshop, you must create two tables named ods_raw_log_d and ods_user_info_d. The ods_raw_log_d table is used to store logs that are synchronized from OSS and the ods_user_info_d table is used to store user information that is synchronized from ApsaraDB RDS.

   > **Notice** The table name can be a maximum of 64 characters in length. It must start with a letter and cannot contain special characters.

3. Create the tables by executing DDL statements.

   - Create the ods_raw_log_d table.

On the creation tab of the ods_raw_log_d table, click **DDL Statement**. In the DDL Statement dialog box, enter the following statement to create the table.



```
CREATE TABLE IF NOT EXISTS doctest
(
    name STRING,
    id   BIGINT
) ;
```

CREATE AS statements are not supported. To run such a statement, create an SQL node.

```
-- Create the ods_raw_log_d table.
CREATE TABLE IF NOT EXISTS  ods_raw_log_d (
    col STRING
)
PARTITIONED BY (
    dt STRING
);
```

- Create the ods_user_info_d table.

  On the creation tab of the ods_user_info_d table, click **DDL Statement**. In the DDL Statement dialog box, enter the following statement to create the table.

```
-- Create the ods_user_info_d table.
CREATE TABLE IF NOT  EXISTS ods_user_info_d (
    uid STRING COMMENT 'User ID',
    gender STRING COMMENT 'Gender',
    age_range STRING COMMENT 'Age range',
    zodiac STRING COMMENT 'Zodiac sign'
)
PARTITIONED BY (
    dt STRING
);
```

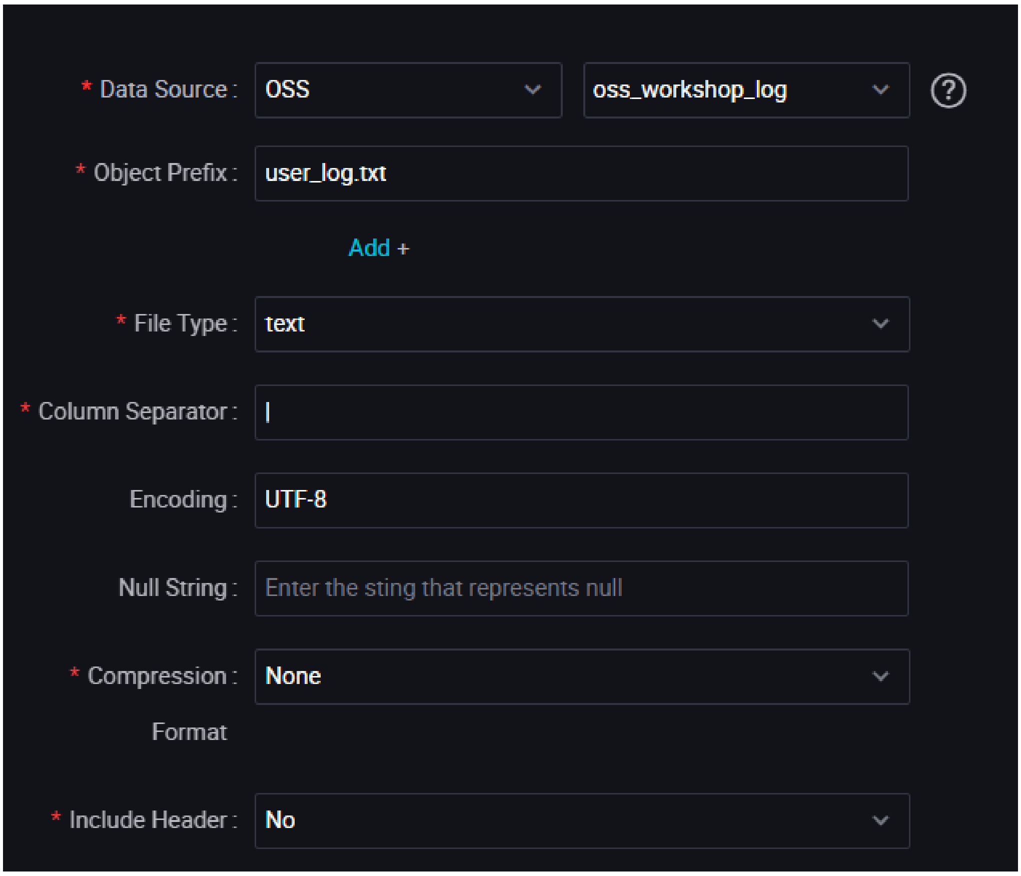
4. Click **Generate Table Schema**. In the **Confirm** message, click OK.
5. On the creation tab for each table, enter the display name in the **General** section.
6. After the creation is complete, click **Commit to Development Environment** and **Commit to Production Environment** in sequence.

Configure the batch sync nodes

> ⓘ **Note** In a workspace in standard mode, we recommend that you do not run batch sync nodes in the development environment. This means that directly running nodes on their configuration tabs is not recommended. Instead, we recommend that you deploy the nodes in the production environment and then run the nodes in test mode to obtain complete operational logs.
>
> After the nodes are deployed in the production environment, you can apply for the permissions to read data from and write data to the tables that are in the development environment.

1. Configure the oss_synchronization node.
   i. In the **Scheduled Workflow** pane, double-click the oss_synchronization node in the new workflow. The node configuration tab appears.

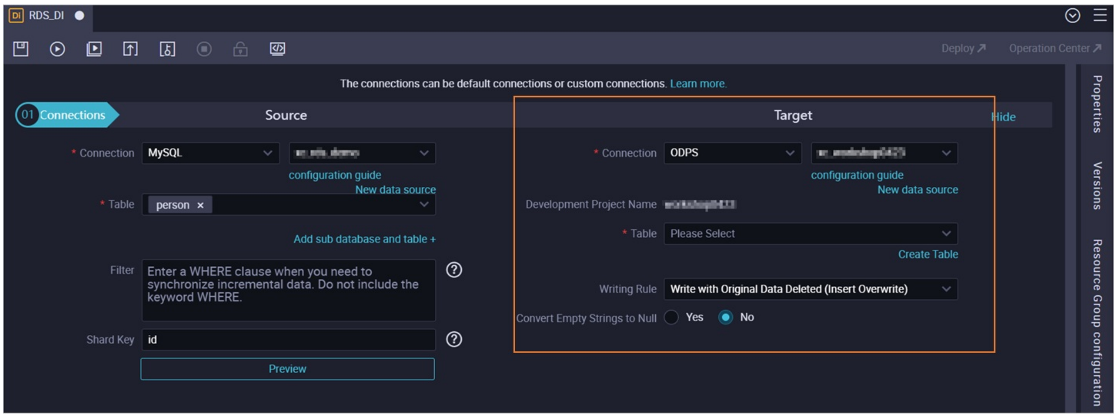
ii. Configure a source.



* Data Source :	OSS · oss_workshop_log
* Object Prefix :	user_log.txt
	Add +
* File Type :	text
* Column Separator :	|
Encoding :	UTF-8
Null String :	Enter the sting that represents null
* Compression : Format	None
* Include Header :	No

Parameter	Description
Connection	The type and name of the source. Select **OSS** and **oss_workshop_log** in sequence.
Object Name Prefix	The prefix of the OSS object for storing the data to be synchronized. Do not enter the name of the OSS bucket. In this workshop, enter user_log.txt.
File Type	The object type. Select text.
Field Delimiter	The column delimiter. Enter |.
Encoding	The encoding format. Default value: UTF-8.
Null String	The string that represents a null pointer.
Compression Format	The compression format of the OSS object. Valid values: None, Gzip, Bzip2, and Zip. Select None.
Skip Header	Specifies whether to include the table header. Default value: No.

iii. Configure a destination.



The connections can be default connections or custom connections. Learn more.

01 Connections

Source

- * Connection: MySQL, [redacted]
- configuration guide
- New data source
- * Table: person ×
- Add sub database and table +
- Filter: Enter a WHERE clause when you need to synchronize incremental data. Do not include the keyword WHERE.
- Shard Key: id
- Preview

Target

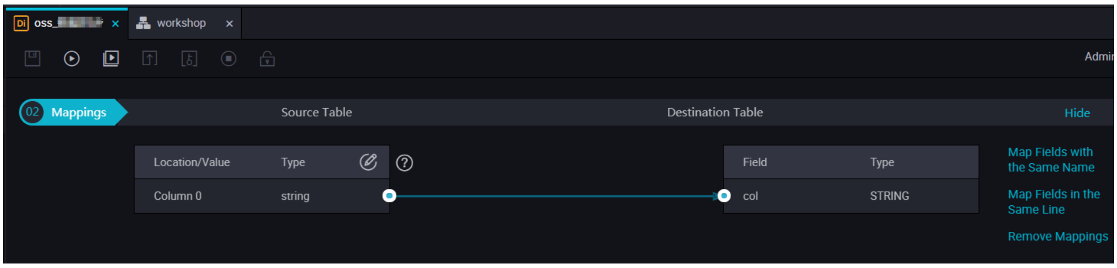
- Hide
- * Connection: ODPS, [redacted]
- configuration guide
- New data source
- Development Project Name: [redacted]
- * Table: Please Select
- Create Table
- Writing Rule: Write with Original Data Deleted (Insert Overwrite)
- Convert Empty Strings to Null: ○ Yes ● No

Parameter	Description
Connection	The type and name of the destination. Select **ODPS** and **odps_first** in sequence.
Table	The table for storing the synchronized data. Select the ods_raw_log_d table.
Partition Key Column	The partition information. Default value: ${bizdate}.
Writing Rule	The method that is used to process existing data before new data is written to MaxCompute. Default value: **Write with Original Data Deleted (Insert Overwrite)**.
Convert Empty Strings to Null	Specifies whether to convert empty strings to null. Select **No**.

Note

- The default odps_first data source is automatically generated for a workspace by DataWorks after you associate a MaxCompute compute engine instance with the workspace for the first time.
- The odps_first data source is used to write synchronized data to a MaxCompute project in the current workspace.

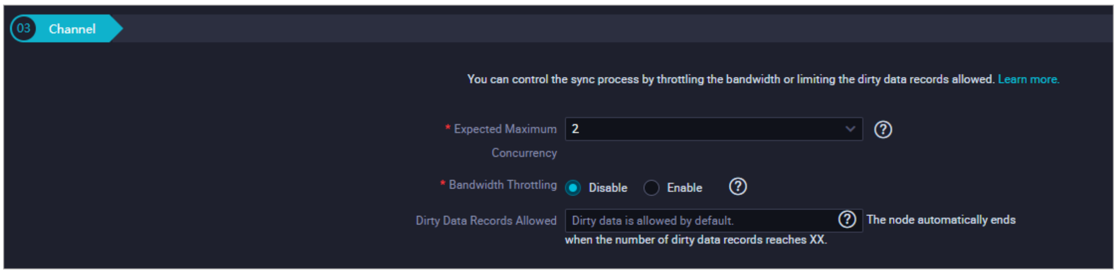
iv. Configure the mappings between fields in the source and destination.



02 Mappings

Source Table Location/Value	Type	Destination Table Field	Type
Column 0	string	col	STRING

v. Set parameters in the **Channel** section.



You can control the sync process by throttling the bandwidth or limiting the dirty data records allowed. Learn more.

* Expected Maximum Concurrency: 2

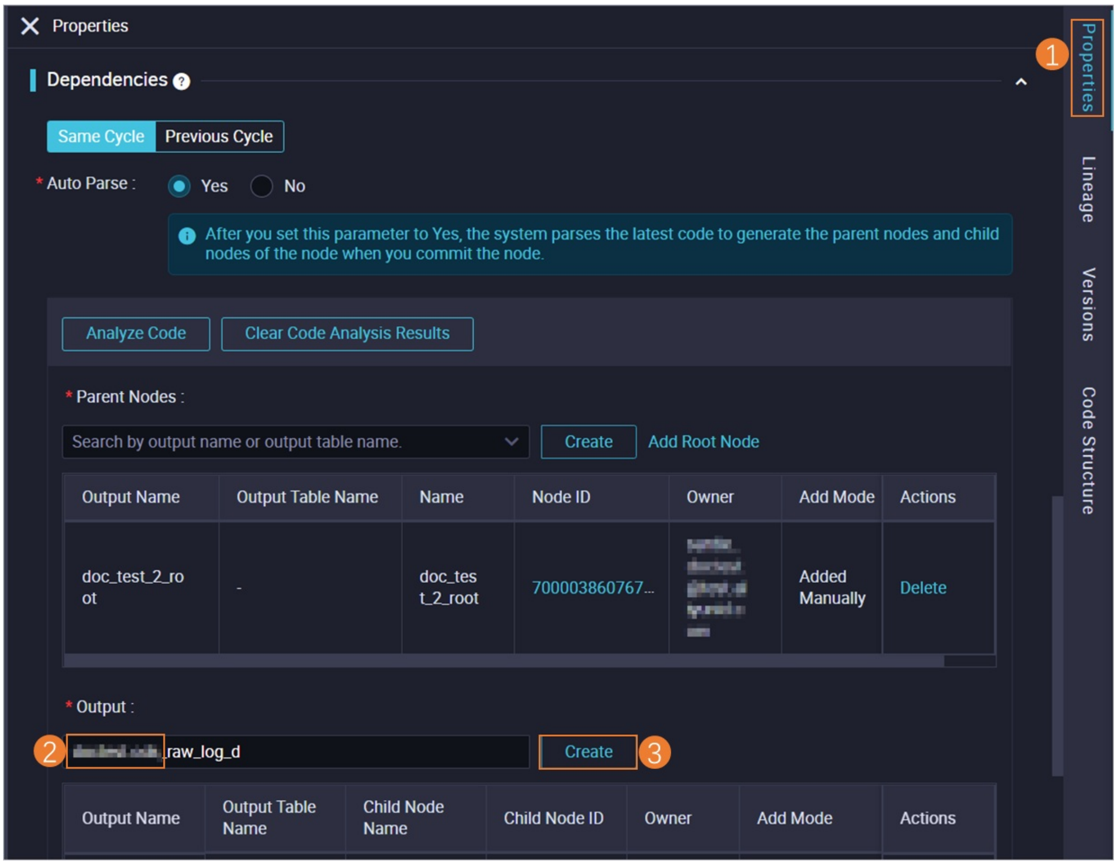
* Bandwidth Throttling: ● Disable ○ Enable

Dirty Data Records Allowed: Dirty data is allowed by default. The node automatically ends when the number of dirty data records reaches XX.

Parameter	Description
Expected Maximum Concurrency	The maximum number of parallel threads that the sync node uses to read data from the source or write data to the destination. You can configure the parallelism for the sync node on the codeless UI.
Bandwidth Throttling	Specifies whether to enable bandwidth throttling. You can enable bandwidth throttling and specify a maximum transmission rate to prevent heavy read workloads on the source. We recommend that you enable bandwidth throttling and set the maximum transmission rate to an appropriate value based on the configurations of the source.
Dirty Data Records Allowed	The maximum number of dirty data records allowed.

vi. On the node configuration tab, click the **Properties** tab in the right-side navigation pane. In the **Dependencies** section, enter an output name in the Workspace name.ods_raw_log_d format below **Outputs** and click Create.

> **Notice** We recommend that you do not include Chinese characters in the output names of the current node. Chinese characters reduce the accuracy of automatic recommendation.



Properties

Dependencies

Same Cycle | Previous Cycle

* Auto Parse : ● Yes ○ No

After you set this parameter to Yes, the system parses the latest code to generate the parent nodes and child nodes of the node when you commit the node.

Analyze Code | Clear Code Analysis Results

* Parent Nodes :

Search by output name or output table name. | Create | Add Root Node

Output Name	Output Table Name	Name	Node ID	Owner	Add Mode	Actions
doc_test_2_root	-	doc_test_2_root	700003860767...	[redacted]	Added Manually	Delete

* Output :

[redacted]_raw_log_d | Create

Output Name	Output Table Name	Child Node Name	Child Node ID	Owner	Add Mode	Actions

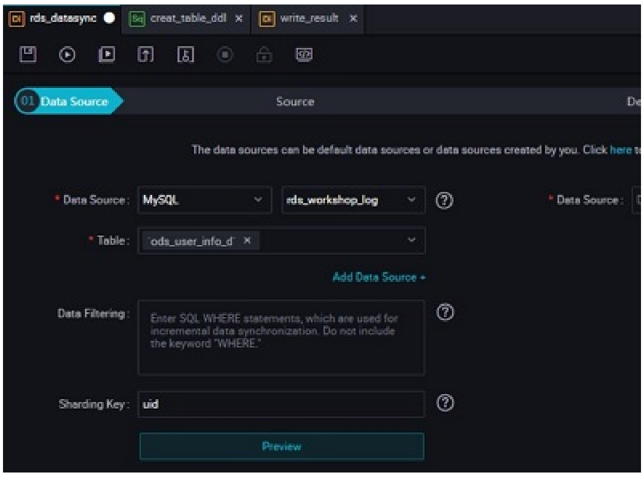
vii. Click the 💾 icon in the top toolbar.

viii. Close the node configuration tab.

2. Configure the rds_synchronization node.

   i. In the **Scheduled Workflow** pane, double-click the rds_synchronization node in the new workflow. The node configuration tab appears.

ii. Configure a source.



The data sources can be default data sources or data sources created by you. Click here t

* Data Source : MySQL rds_workshop_log

* Table : `ods_user_info_d`

Add Data Source +

Data Filtering : Enter SQL WHERE statements, which are used for incremental data synchronization. Do not include the keyword "WHERE."

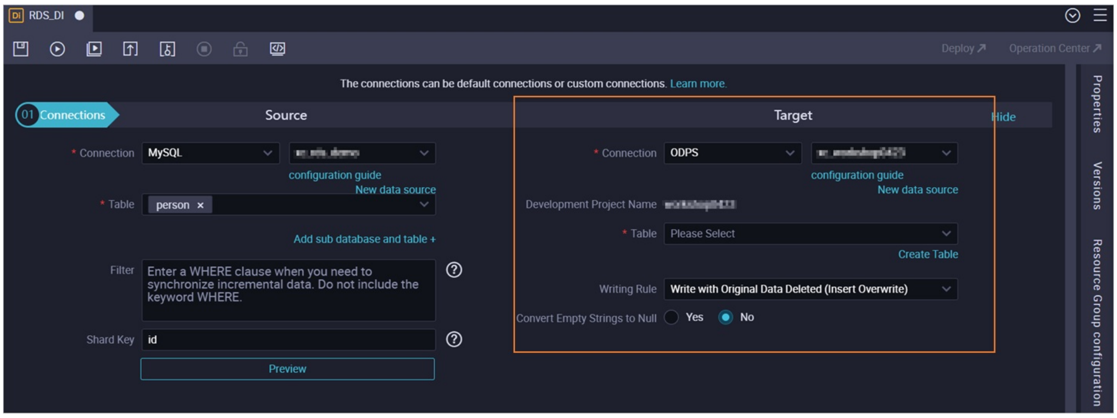
Sharding Key : uid

Preview

* Data Source :

Parameter	Description
Connection	The type and name of the source. Select **MySQL** and **rds_workshop_log** in sequence.
Table	The table for storing the synchronized data. Select the ods_user_info_d table in MaxCompute.
Filter	The condition used to filter the data that you want to synchronize. The filter condition is often used to filter incremental data. You can leave this parameter unspecified.
Shard Key	The shard key for the data to be synchronized. Default value: uid.

iii. Configure a destination.



The connections can be default connections or custom connections. Learn more.

Source

* Connection: MySQL, [redacted]

* Table: person ×

Filter: Enter a WHERE clause when you need to synchronize incremental data. Do not include the keyword WHERE.

Shard Key: id

Target

* Connection: ODPS, [redacted]

Development Project Name: [redacted]

* Table: Please Select

Writing Rule: Write with Original Data Deleted (Insert Overwrite)

Convert Empty Strings to Null: ○ Yes ● No

Parameter	Description
Connection	The type and name of the destination. Select **ODPS** and **odps_first** in sequence.
Table	The table for storing the synchronized data. Select the ods_user_info_d table in MaxCompute.
Partition Key Column	The partition information. Default value: ${bizdate}.
Writing Rule	The method that is used to process existing data before new data is written to MaxCompute. Default value: **Write with Original Data Deleted (Insert Overwrite)**.
Convert Empty Strings to Null	Specifies whether to convert empty strings to null. Select **No**.

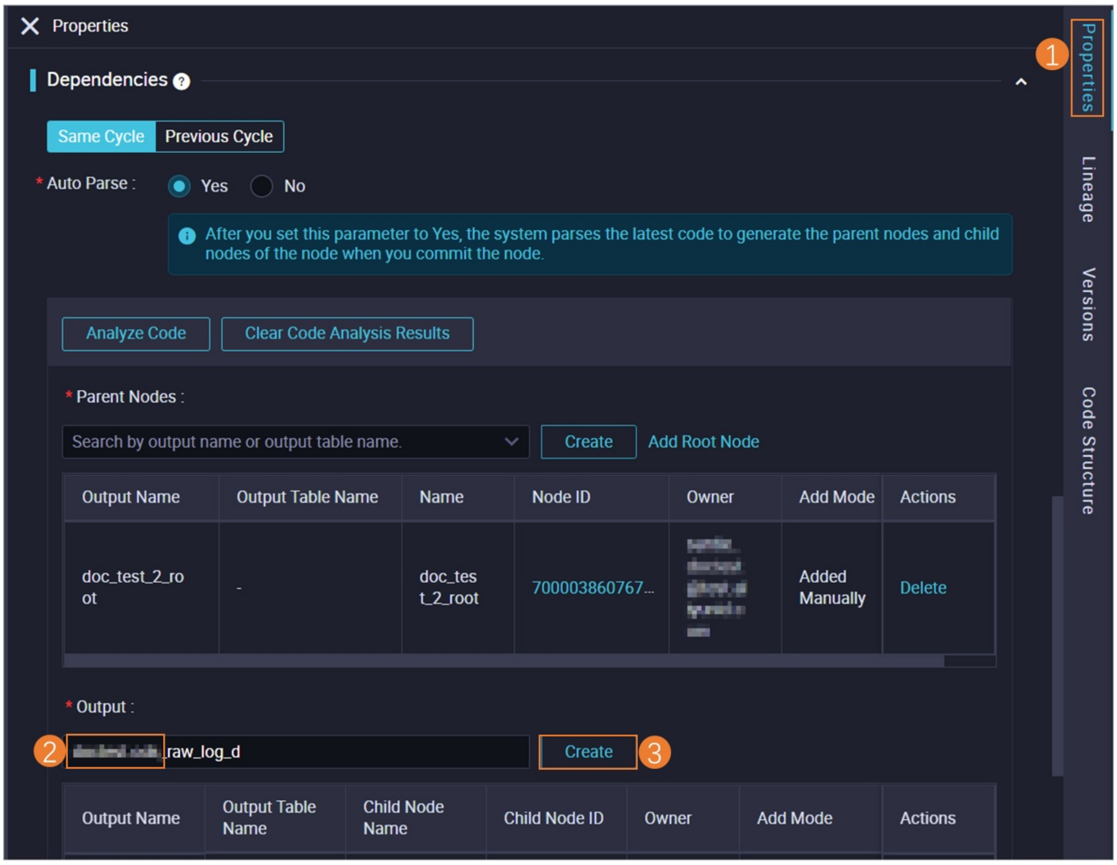
iv. Configure the mapping between fields in the source and destination.

v. Set parameters in the **Channel** section.

vi. On the node configuration tab, click the **Properties** tab in the right-side navigation pane. In the **Dependencies** section, enter an output name in the Workspace name.ods_user_info_d format below **Outputs** and click Create.

If an output name does not comply with the naming rules, click **Delete** to delete the output name.

> **Notice** We recommend that you do not include Chinese characters in the output names of the current node. Chinese characters reduce the accuracy of automatic recommendation.



Properties

Dependencies

Same Cycle | Previous Cycle

* Auto Parse : ● Yes ○ No

After you set this parameter to Yes, the system parses the latest code to generate the parent nodes and child nodes of the node when you commit the node.

Analyze Code | Clear Code Analysis Results

* Parent Nodes :

Search by output name or output table name. | Create | Add Root Node

Output Name	Output Table Name	Name	Node ID	Owner	Add Mode	Actions
doc_test_2_root	-	doc_test_2_root	700003860767...	[redacted]	Added Manually	Delete

* Output :

[redacted]raw_log_d | Create

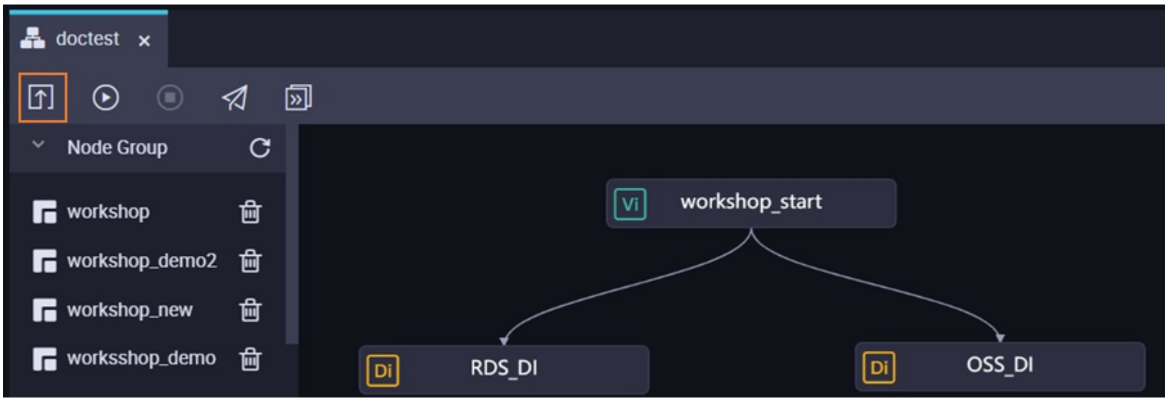
Output Name	Output Table Name	Child Node Name	Child Node ID	Owner	Add Mode	Actions

vii. Click the 💾 icon in the top toolbar.

viii. Close the node configuration tab.

Commit the workflow

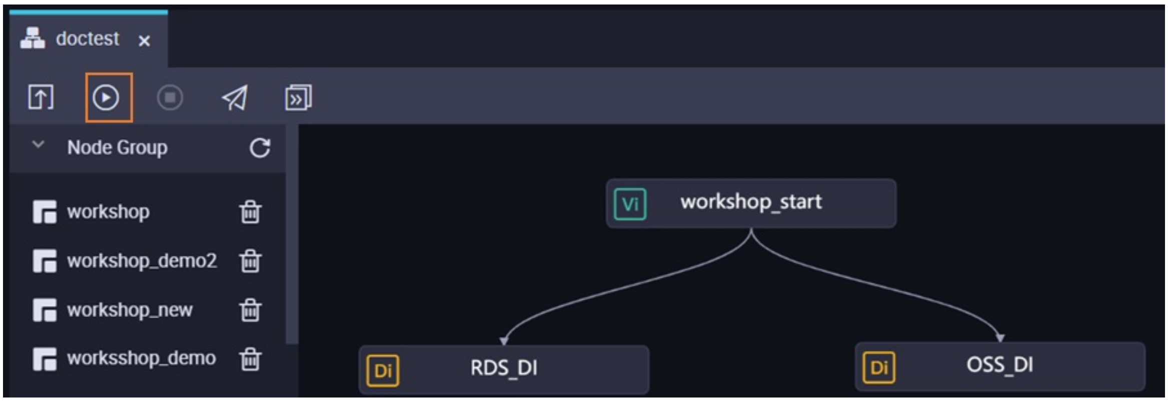
1. In the **Scheduled Workflow** pane, double-click the new workflow. On the workflow configuration tab, click the ⬆ icon in the top toolbar.



2. In the **Commit** dialog box, select the nodes to be committed, enter your comments in the **Change description** field, and then select **Ignore I/O Inconsistency Alerts**.
3. Click **Commit**. The **Committed successfully** message appears.

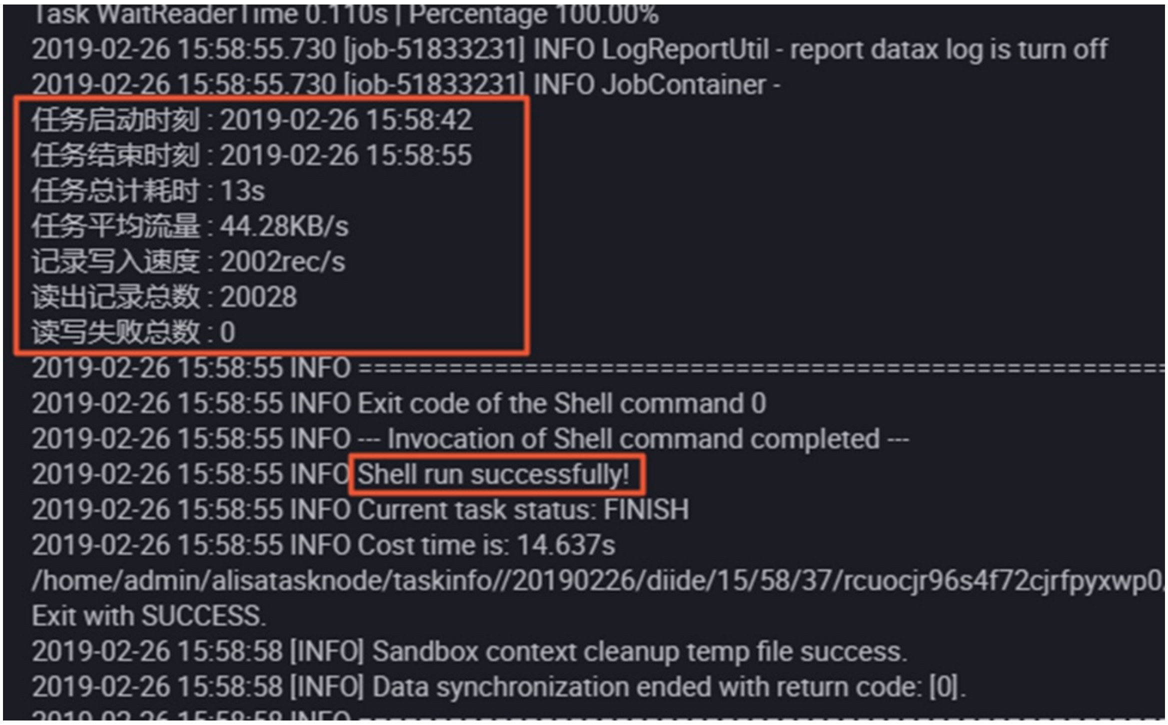
Run the workflow

1. In the **Scheduled Workflow** pane, double-click the new workflow. On the workflow configuration tab that appears, click the ▶ icon in the top toolbar.



2. Right-click the rds_synchronization node and select **View Log**.

   If the information shown in the following figure appears in the logs, the rds_synchronization node is run and data is synchronized.



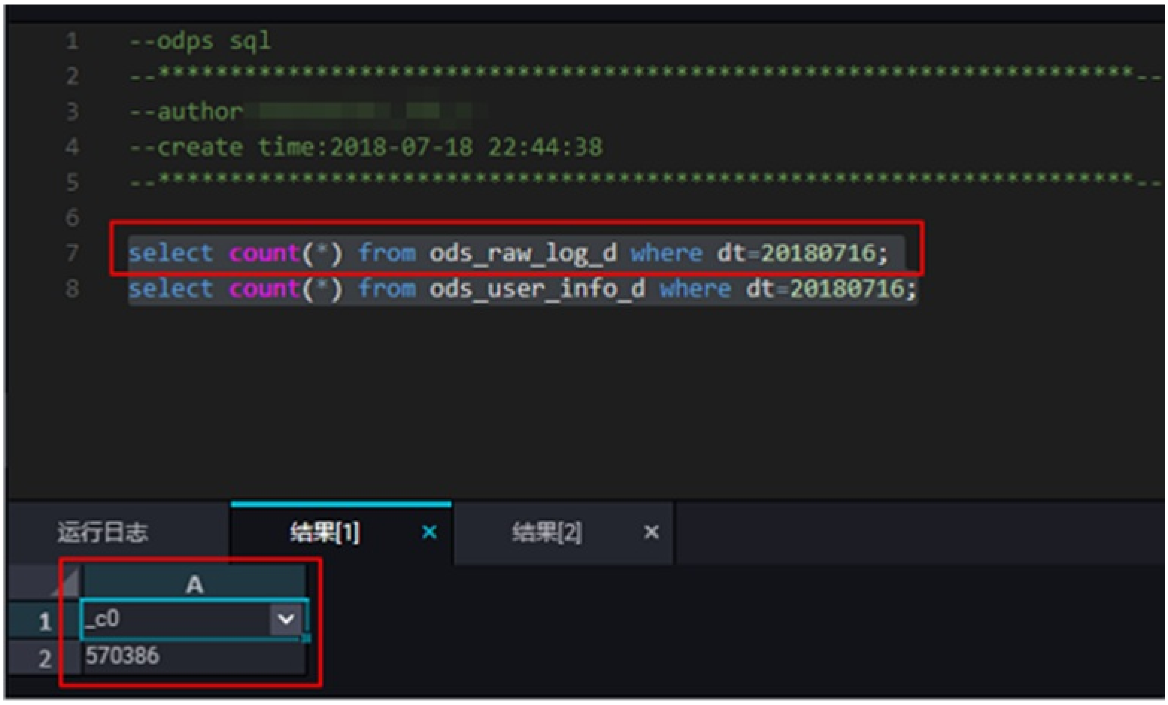
```
Task WaitReaderTime 0.110s | Percentage 100.00%
2019-02-26 15:58:55.730 [job-51833231] INFO LogReportUtil - report datax log is turn off
2019-02-26 15:58:55.730 [job-51833231] INFO JobContainer -
任务启动时刻 : 2019-02-26 15:58:42
任务结束时刻 : 2019-02-26 15:58:55
任务总计耗时 : 13s
任务平均流量 : 44.28KB/s
记录写入速度 : 2002rec/s
读出记录总数 : 20028
读写失败总数 : 0
2019-02-26 15:58:55 INFO ==================================================
2019-02-26 15:58:55 INFO Exit code of the Shell command 0
2019-02-26 15:58:55 INFO --- Invocation of Shell command completed ---
2019-02-26 15:58:55 INFO Shell run successfully!
2019-02-26 15:58:55 INFO Current task status: FINISH
2019-02-26 15:58:55 INFO Cost time is: 14.637s
/home/admin/alisatasknode/taskinfo//20190226/diide/15/58/37/rcuocjr96s4f72cjrfpyxwp0
Exit with SUCCESS.
2019-02-26 15:58:58 [INFO] Sandbox context cleanup temp file success.
2019-02-26 15:58:58 [INFO] Data synchronization ended with return code: [0].
2019-02-26 15:58:58 INFO ==================================================
```

3. Right-click the oss_synchronization node and select **View Log**. View the logs to check whether the oss_synchronization node is run and data is synchronized.

Verify data synchronization to MaxCompute

1. In the left-side navigation pane, click **Ad-Hoc Query**.
2. In the Ad-Hoc Query pane, right-click **Ad-Hoc Query** and choose **Create Node > ODPS SQL**.
3. In the Create Node dialog box, enter the node name and click Commit. On the node configuration tab that appears, write and execute SQL statements to view the number of data records that are synchronized to the ods_raw_log_d and ods_user_info_d tables.

> **Note** Execute the following SQL statements. In each statement, change the partition key value to the data timestamp of the node. For example, if the node is run on July 17, 2018, the data timestamp is 20180716, which is one day before the node is run.
>
> ```
> -- Check whether the data is written to MaxCompute.
> select count(*) from ods_raw_log_d where dt=Data timestamp of the node;
> select count(*) from ods_user_info_d where dt=Data timestamp of the node;
> ```



```
--odps sql
--**********************************************************************--
--author [redacted]
--create time:2018-07-18 22:44:38
--**********************************************************************--

select count(*) from ods_raw_log_d where dt=20180716;
select count(*) from ods_user_info_d where dt=20180716;
```

```
   A
1  _c0
2  570386
```

Subsequent steps

You understand how to collect and synchronize data. You can now proceed with the next tutorial. In the next tutorial, you will learn how to compute and analyze collected data. For more information, see Process data.

1.4. Process data

This topic describes how to compute and analyze collected data by using DataWorks.

Prerequisites

The data that is required for the workshop is collected. For more information, see Collect data.

Create tables

1. Go to the **DataStudio** page.
    i. Log on to the DataWorks console.
    ii. In the left-side navigation pane, click **Workspaces**.
    iii. In the top navigation bar, select the region in which the workspace that you want to manage resides. Find the workspace and click **DataStudio** in the Actions column.
2. On the **DataStudio** page, click the created workflow. Right-click **MaxCompute** and choose **Create > Table**.
3. In the **Create Table** dialog box, set the **Table Name** parameter and click **Create**.

    Create a table named ods_log_info_d at the operational data store (ODS) layer, a table named dw_user_info_all_d at the common data model (CDM) layer, and a table named rpt_user_info_d at the application data store (ADS) layer.
4. Run DDL statements to create tables.
    - Create the ods_log_info_d table.

Double-click the ods_log_info_d table. On the configuration tab on the right side, click **DDL Statement** and enter the following table creation statement:

```
-- Create a table at the ODS layer.
CREATE TABLE IF NOT EXISTS ods_log_info_d (
  ip STRING COMMENT 'The IP address',
  uid STRING COMMENT 'The ID of the user',
  time STRING COMMENT 'The time in the format of yyyymmddhh:mi:ss',
  status STRING COMMENT 'The status code that is returned by the server',
  bytes STRING COMMENT 'The number of bytes that are returned to the client',
  region STRING COMMENT 'The region, which is obtained based on the IP address',
  method STRING COMMENT 'The type of the HTTP request',
  url STRING COMMENT 'url',
  protocol STRING COMMENT 'The version number of HTTP',
  referer STRING COMMENT 'The source URL',
  device STRING COMMENT 'The terminal type',
  identity STRING COMMENT 'The access type, which can be crawler, feed, user, or unkn
own'
)
PARTITIONED BY (
  dt STRING
);
```

- Create the dw_user_info_all_d table.

  Double-click the dw_user_info_all_d table. On the configuration tab on the right side, click **DDL Statement** and enter the following table creation statement:

```
-- Create a table at the CDM layer.
CREATE TABLE IF NOT EXISTS dw_user_info_all_d (
  uid STRING COMMENT 'The ID of the user',
  gender STRING COMMENT 'The gender',
  age_range STRING COMMENT 'The age range',
  zodiac STRING COMMENT 'The zodiac sign',
  region STRING COMMENT 'The region, which is obtained based on the IP address',
  device STRING COMMENT 'The terminal type',
  identity STRING COMMENT 'The access type, which can be crawler, feed, user, or unkn
own',
  method STRING COMMENT 'The type of the HTTP request',
  url STRING COMMENT 'url',
  referer STRING COMMENT 'The source URL',
  time STRING COMMENT 'The time in the format of yyyymmddhh:mi:ss'
)
PARTITIONED BY (
  dt STRING
);
```

- Create the rpt_user_info_d table.

  Double-click the rpt_user_info_d table. On the configuration tab on the right side, click **DDL Statement** and enter the following table creation statement:

```
-- Create a table at the ADS layer.
CREATE TABLE IF NOT EXISTS rpt_user_info_d (
  uid STRING COMMENT 'The ID of the user',
  region STRING COMMENT 'The region, which is obtained based on the IP address',
  device STRING COMMENT 'The terminal type',
  pv BIGINT COMMENT 'pv',
  gender STRING COMMENT 'The gender',
  age_range STRING COMMENT 'The age range',
  zodiac STRING COMMENT 'The zodiac sign'
)
PARTITIONED BY (
  dt STRING
);
```

5. After you enter the table creation statement, click **Generate Table Schema**. Then, click **OK** to overwrite the current settings.
6. On the table configuration tab, enter the display name of the table in the **General** section.
7. After you complete the configuration, click **Commit in Development Environment** and **Commit to Production Environment**.

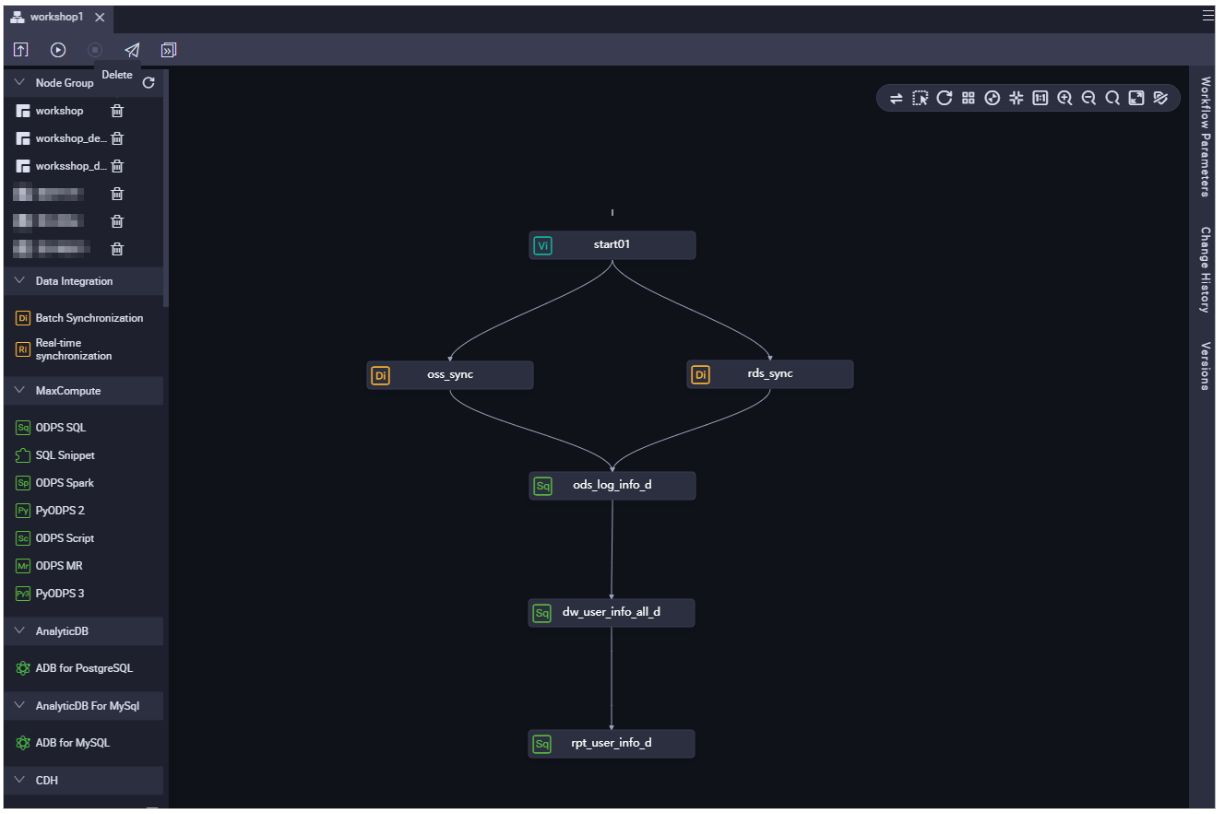
> ? **Note** If you use a workspace in basic mode, only **Commit to Production Environment** is available.

Design the workflow

For more information about how to configure the dependencies among nodes of a workflow, see Collect data.

Double-click the created workflow. On the configuration tab of the workflow, click and drag **ODPS SQL** to the configuration tab on the right. In the **Create Node** dialog box, set the **Node Name** parameter and click **Commit**.

Create three ODPS SQL nodes in total and name them ods_log_info_d, dw_user_info_all_d, and rpt_user_info_d. Then, configure the dependencies among the nodes, as shown in the following figure.



Create a UDF

1. Create a resource.

   i. Download the ip2region.jar file.

   ii. On the DataStudio page, click the created workflow, right-click **MaxCompute**, and then choose **Create** > **Resource** > **JAR**.

   iii. In the **Create Resource** dialog box, set the **Resource Name** and **Location** parameters.

   > Note
   >
   > - Select **Upload to MaxCompute**.
   > - The resource name can be different from the name of the uploaded file.
   > - A resource name can contain letters, digits, underscores (_), and periods (.), and is not case-sensitive. It must be 1 to 128 characters in length. A JAR resource name must end with .jar, and a Python resource name must end with .py.

   iv. Click **Upload**, select the *ip2region.jar* file that is downloaded to your local computer, and then click **Open**.

   v. Click **Create**.

   vi. Click the icon in the toolbar.

2. Register a function.

i. On the **DataStudio** page, click the created workflow, right-click **MaxCompute**, and then choose **Create > Function**.

ii. In the **Create Function** dialog box, set the **Function Name** parameter and click **Create**. For example, you can set the Function Name parameter to getregion.

iii. In the **Register Function** section, set the parameters.

Parameter	Description
Function Type	The type of the function.
Engine Instance MaxCompute	By default, the parameter cannot be modified.
Function Name	The name of the function that you entered when you created the function.
Owner	The owner of the function.
Class Name	Set the parameter to `org.alidata.odps.udf.Ip2Region`.
Resources	Set the parameter to `ip2region.jar`.
Description	Set the parameter to Region conversion based on the IP address.
Expression Syntax	Set the parameter to `getregion('ip')`.
Parameter Description	Set the parameter to IP address.

iv. Click the 💾 and ⇪ icons in the toolbar.

Configure the ODPS SQL nodes

1. Configure the ods_log_info_d node.

   i. Double-click the ods_log_info_d node to go to the configuration tab of the node.

   ii. On the configuration tab of the node, enter the following SQL statements:

```
INSERT OVERWRITE TABLE ods_log_info_d PARTITION (dt=${bdp.system.bizdate})
SELECT ip
  , uid
  , time
  , status
  , bytes
  , getregion(ip) AS region -- Obtain the region based on the IP address by using the user-defined function (UDF).
  , regexp_substr(request, '(^[^ ]+ )') AS method -- Use the regular expression to extract three fields from the request.
  , regexp_extract(request, '^[^ ]+ (. *) [^ ]+$') AS url
  , regexp_substr(request, '([^ ]+$)') AS protocol
  , regexp_extract(referer, '^[^/]+://([^/]+){1}') AS referer -- Use the regular expression to clarify the referrer, so as to obtain a more accurate URL.
  , CASE
    WHEN TOLOWER(agent) RLIKE 'android' THEN 'android' -- Obtain the terminal information and access types based on the agent parameter.
    WHEN TOLOWER(agent) RLIKE 'iphone' THEN 'iphone'
    WHEN TOLOWER(agent) RLIKE 'ipad' THEN 'ipad'
    WHEN TOLOWER(agent) RLIKE 'macintosh' THEN 'macintosh'
    WHEN TOLOWER(agent) RLIKE 'windows phone' THEN 'windows_phone'
    WHEN TOLOWER(agent) RLIKE 'windows' THEN 'windows_pc'
    ELSE 'unknown'
  END AS device
  , CASE
    WHEN TOLOWER(agent) RLIKE '(bot|spider|crawler|slurp)' THEN 'crawler'
    WHEN TOLOWER(agent) RLIKE 'feed'
    OR regexp_extract(request, '^[^ ]+ (. *) [^ ]+$') RLIKE 'feed' THEN 'feed'
    WHEN TOLOWER(agent) NOT RLIKE '(bot|spider|crawler|feed|slurp)'
    AND agent RLIKE '^[Mozilla|Opera]'
    AND regexp_extract(request, '^[^ ]+ (. *) [^ ]+$') NOT RLIKE 'feed' THEN 'user'
    ELSE 'unknown'
  END AS identity
  FROM (
    SELECT SPLIT(col, '##@@')[0] AS ip
    , SPLIT(col, '##@@')[1] AS uid
    , SPLIT(col, '##@@')[2] AS time
    , SPLIT(col, '##@@')[3] AS request
    , SPLIT(col, '##@@')[4] AS status
    , SPLIT(col, '##@@')[5] AS bytes
    , SPLIT(col, '##@@')[6] AS referer
    , SPLIT(col, '##@@')[7] AS agent
  FROM ods_raw_log_d
  WHERE dt = ${bdp.system.bizdate}
) a;
```

iii. Click the 💾 icon in the toolbar.

2. Configure the dw_user_info_all_d node.

   i. Double-click the dw_user_info_all_d node to go to the configuration tab of the node.

ii. On the configuration tab of the node, enter the following SQL statements:

```
INSERT OVERWRITE TABLE dw_user_info_all_d PARTITION (dt='${bdp.system.bizdate}')
SELECT COALESCE(a.uid, b.uid) AS uid
  , b.gender
  , b.age_range
  , b.zodiac
  , a.region
  , a.device
  , a.identity
  , a.method
  , a.url
  , a.referer
  , a.time
FROM (
  SELECT *
  FROM ods_log_info_d
  WHERE dt = ${bdp.system.bizdate}
) a
LEFT OUTER JOIN (
  SELECT *
  FROM ods_user_info_d
  WHERE dt = ${bdp.system.bizdate}
) b
ON a.uid = b.uid;
```

iii. Click the 💾 icon in the toolbar.

3. Configure the rpt_user_info_d node.

   i. Double-click the rpt_user_info_d node to go to the configuration tab of the node.

   ii. On the configuration tab of the node, enter the following SQL statements:

```
INSERT OVERWRITE TABLE rpt_user_info_d PARTITION (dt='${bdp.system.bizdate}')
SELECT uid
  , MAX(region)
  , MAX(device)
  , COUNT(0) AS pv
  , MAX(gender)
  , MAX(age_range)
  , MAX(zodiac)
FROM dw_user_info_all_d
WHERE dt = ${bdp.system.bizdate}
GROUP BY uid;
```

   iii. Click the 💾 icon in the toolbar.

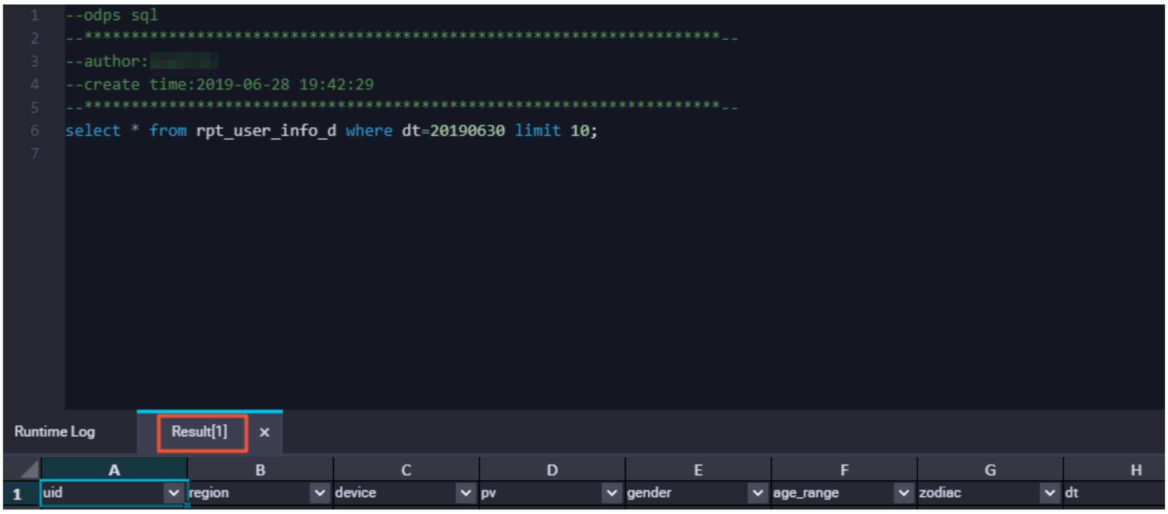
Commit the workflow

1. On the configuration tab of the workflow, click the ⬆ icon to commit the nodes that are configured in the workflow.
2. In the **Commit** dialog box, select the nodes that you want to commit and select **Ignore I/O Inconsistency Alerts**.

3. Click **Commit**.

Run the workflow

1. On the configuration tab of the workflow, click the ⊙ icon in the toolbar to verify the logic of node code.
2. After all nodes are run and a green check mark (✓) appears, click **Ad-Hoc Query** on the left-side navigation submenu.
3. On the **Ad-Hoc Query** tab, right-click **Ad-Hoc Query** and choose **Create Node** > **ODPS SQL**.
4. Write and run an SQL statement to query the node running result and check whether required data is generated.



```
--odps sql
--********************************************************************--
--author:[redacted]
--create time:2019-06-28 19:42:29
--********************************************************************--
select * from rpt_user_info_d where dt=20190630 limit 10;
```

	A	B	C	D	E	F	G	H
1	uid	region	device	pv	gender	age_range	zodiac	dt

Use the following SQL query statement. By default, the data timestamp of a node is one day before the node is run.

```
--- View data in the rpt_user_info_d table.
select * from rpt_user_info_d where dt=Data timestamp limit 10;
```

Deploy the workflow

After you commit the workflow, the nodes in the workflow are in the development environment. You must deploy the configured nodes in the production environment because nodes in the development environment cannot be automatically scheduled.

> Note
>
> - Before you deploy the nodes in the production environment, test the node code to make sure that the code is correct.
> - In a workspace in basic mode, the icon is unavailable. After you commit a node, click the icon to go to the Operation Center page.

1. On the configuration tab of the workflow, click the icon in the toolbar to go to the Deploy page.
2. Select the nodes that you want to deploy and click **Add to List**.
3. Click **To-Be-Deployed Node List** in the upper-right corner. In the Nodes to Deploy panel, click

Deploy All.

4. In the **Create Deploy Task** dialog box, click **Deploy**.
5. In the left-side navigation pane, click **Release Package** to view the deployment status.

Run the nodes in the production environment

1. After you deploy the nodes, click **Operation Center** in the upper-right corner.

   You can also click **Go to Operation Center** in the toolbar on the configuration tab of the workflow to go to the **Operation Center** page.
2. In the left-side navigation pane, choose **Cycle Task Maintenance > Cycle Task** to go to the **Cycle Task** page. Then, click the workshop workflow.
3. Double-click the zero load node in the directed acyclic graph (DAG) to show the workflow. Right-click the workshop_start node and choose **Run > Current and Descendent Nodes Retroactively**.
4. Select nodes to generate retroactive data, specify the data timestamp, and then click **OK**. The **Patch Data** page appears.
5. Click **Refresh** until all SQL nodes are run.

What to do next

You have learned how to create SQL nodes and process raw logs. You can now proceed with the next tutorial to learn how to monitor and ensure the quality of the data that is generated by the developed nodes. For more information, see Configure rules to monitor data quality.

1.5. Configure rules to monitor data quality

This topic describes how to monitor data quality by configuring a quality monitoring rule for each table and alert notifications.

Prerequisites

The data is collected and processed before this experiment. For more information, see Collect data and Process data.

Context

Data Quality is a one-stop platform that allows you to check the data quality of heterogeneous data stores, configure alert notifications, and manage connections. Data Quality monitors data in datasets and allows you to monitor MaxCompute tables and DataHub topics. When offline MaxCompute data changes, Data Quality checks the data and blocks nodes that involves the data. This prevents downstream data from being affected by dirty data. In addition, Data Quality allows you to manage the check result history so that you can analyze and evaluate the data quality.

For streaming data, Data Quality uses DataHub to monitor data streams and sends alert notifications to subscribers if it detects stream discontinuity. You can also set the alert severity such as warning and error alerts, and the alert frequency to minimize repeated alerts.

Development process in Data Quality

1. Configure a monitoring rule for an existing table and test the monitoring rule to check whether the

monitoring rule takes effect on the table.

Based on the test result, you can determine whether data that is generated in the table is as expected. We recommend that you test every monitoring rule configured for a table to verify that these monitoring rules are applicable.

2. After the test is successful, link the tested monitoring rule to scheduled nodes.

   After you configure and test the monitoring rule for the table, you must link the monitoring rule with the nodes that generate data in the table. Then, Data Quality can use the monitoring rule to check the quality of the data generated by the nodes each time the nodes are run. This ensures data accuracy.

3. After the monitoring rule is linked to the scheduled node, the monitoring rule that is used to check the data quality is triggered each time the linked node is run. This improves data accuracy.

   Data Quality allows you to subscribe to monitoring rules. You can subscribe to the monitoring rules of important tables. After the subscriptions are configured, Data Quality generates alerts based on the monitoring results. This way, you can track the monitoring results. If the monitoring results returned by Data Quality are abnormal, Data Quality send alert notifications to you based on alert rules.

Note

- Each time you configure a monitoring rule for a table, you must test the monitoring rule, link the monitoring rule to the scheduled nodes, and subscribe to the monitoring rule.
- Data Quality may charge you additional computing fees. For more information, see Overview.

Configure monitoring rules of tables

After data collection and data processing are completed, verify that you have created the following tables: ods_raw_log_d, ods_user_info_d, ods_log_info_d, dw_user_info_all_d, and rpt_user_info_d. Then, perform the following operations:

1. Go to the **Monitoring Rules** page of the ods_raw_log_d table.

   i. Log on to the DataWorks console.

   ii. In the left-side navigation pane, click **Workspaces**.

   iii. After you select the region where the required workspace resides, find the workspace and click **Data Analytics**.

   iv. Click the ☰ icon in the upper-left corner and choose **All Products > Data governance > Data Quality**.

   v. In the left-side navigation pane, click **Monitoring Rules** and select **MaxCompute** from the Engine/Data Source drop-down list.

   vi. Select an engine instance where a required table exists from the **Engine/Database Instance** drop-down list. Find the required table for which you want to configure a monitoring rule from the table list, such as the ods_raw_log_d table in this example.

   vii. Find the ods_raw_log_d table and click **View Monitoring Rules**.

2. Configure a monitoring rule for the ods_raw_log_d table.

i. Click the plus sign (**+**) in the **Partition Expression** section to add a partition filter expression.

The ods_raw_log_d table stores the log data that is synchronized from Object Storage Service (OSS) by using the oss_workshop_log connection. The partition key values in the table are in the format of ${bdp.system.bizdate}. The bizdate parameter specifies the date that is one day before the batch synchronization node is run.

You can configure a partition filter expression for such log data that is generated every day. In the **Add Partition** dialog box, select dt=$[yyyymmdd-1] and click **OK**. For more information about partition filter expressions, see Overview of scheduling parameters.

> ⓘ **Note** If your table does not contain any partition key columns, you can select **NOTAPARTITIONTABLE**. Select a partition filter expression based on the actual partition key values.

ii. Click **Create rules**. The **Template Rules** tab appears.

iii. Click **Add Monitoring Rule**, and set the **Template** parameter to **Number of rows, fixed value**, the Rule Type parameter to **Rule Type**, the Comparison Method parameter to **Greater Than**, and the Expected Value parameter to 0.

The data in the ods_raw_log_d table comes from the log files that are uploaded to OSS. The table is used as the source table. Therefore, you must check whether data exists in the partitions of the table as early as possible. If the partitions contain no data, prevent descendant nodes from running. If no source data can be used, no effective results are generated when the descendant nodes are run.

> ⓘ **Note** Data Quality only blocks nodes and sets the status of node instances to Failed when an error alert is generated for a hard rule.

Then, click **Batch Create**.

> ⓘ **Note** The preceding configuration is to ensure that partitions of the table contain data that can be used by descendant nodes.

iv. Click **Test**. In the **Test** dialog box, set the **Data Timestamp** parameter and click **Test**.

Data Quality tests the configured monitoring rule after you click Test. After the test is successful, click **The test is complete. Click to view the results** to go to the page of the test results.

v. Link the monitoring rule to the nodes.

Data Quality allows you to link a monitoring rule of a table to the scheduled nodes. After you link the monitoring rule to the nodes, Data Quality checks the quality of the data generated by the nodes each time the nodes are run. You can link a monitoring rule to a node in one of the following ways:

- Link the monitoring rule to the node in Operation Center

  Click the ☰ icon in the upper-left corner and choose **All Products > Operation Center**.

  In the left-side navigation pane, choose **Cycle Task Maintenance > Cycle Task**. In the directed acyclic graph (DAG), right-click the **oss_Data synchronization** node and select **Configure Data Quality Rules**.

  In the **Configure Data Quality Rules** dialog box, set the **Table Name** parameter to ods_raw_log_d and the **Partition Expression** parameter to dt=$[yyyymmdd-1] and click **Add**.

- Link the monitoring rule to the node in Data Quality

  On the **Monitoring Rules** page of the table, click **Manage Linked Nodes** to link the monitoring rule to the node.

  After you click **Manage Linked Nodes**, you can link the monitoring rule to the nodes that have been committed to the scheduling system. Data Quality lists recommended nodes based on the lineage. You can also link the monitoring rule to other nodes.

  In the **Manage Linked Nodes** dialog box, enter the node ID or name and click **Create**. Then, the monitoring rule is linked to the node.

vi. Configure subscriptions.

On the Monitoring Rules page of the table, click **Manage Subscriptions** to specify the notification method and recipient. Data Quality supports four notification methods: **Email**, **Email and SMS**, **DingTalk Chatbot**, and **DingTalk Chatbot @ALL**.

After you configure subscriptions, click **My Subscriptions** in the left-side navigation pane to view or modify the subscriptions.

> **Note** We recommend that you subscribe to all monitoring rules so that you can receive the monitoring results at the earliest opportunity.

3. Configure monitoring rules for the ods_user_info_d table.

The ods_user_info_d table stores user information. You must configure monitoring rules to verify that the table contains the specified number of rows and that the primary key values in the table are unique to avoid duplicate data.

i. Configure a monitoring rule for a partition field by adding the partition filter expression dt=$[yyyymmdd-1]. After the partition filter expression is added, you can view it in the Partition Expression section.

ii. Then, click **Create rules** to configure a monitoring rule in Data Quality.

Add a monitoring rule for all fields in a table and a monitoring rule for the values in the primary key column:

- Set the **Field** parameter to **All Fields in Table(table)**.

  Set the **Template** parameter to **Number of rows, fixed value**, the **Rule Type** parameter to **Rule Type**, the **Comparison Method** parameter to **Greater Than**, and the **Expected Value** parameter to **0**.

- Set the **Field** parameter to **uid(string)**.

  Configure a rule to monitor the values in the primary key column uid. Set the **Template** parameter to **Repeated value, fixed value**, the **Rule Type** parameter to **Soft**, the **Comparison Method** parameter to **Less Than**, and the **Expected Value** parameter to **1**.

iii. Then, click **Batch Create**.

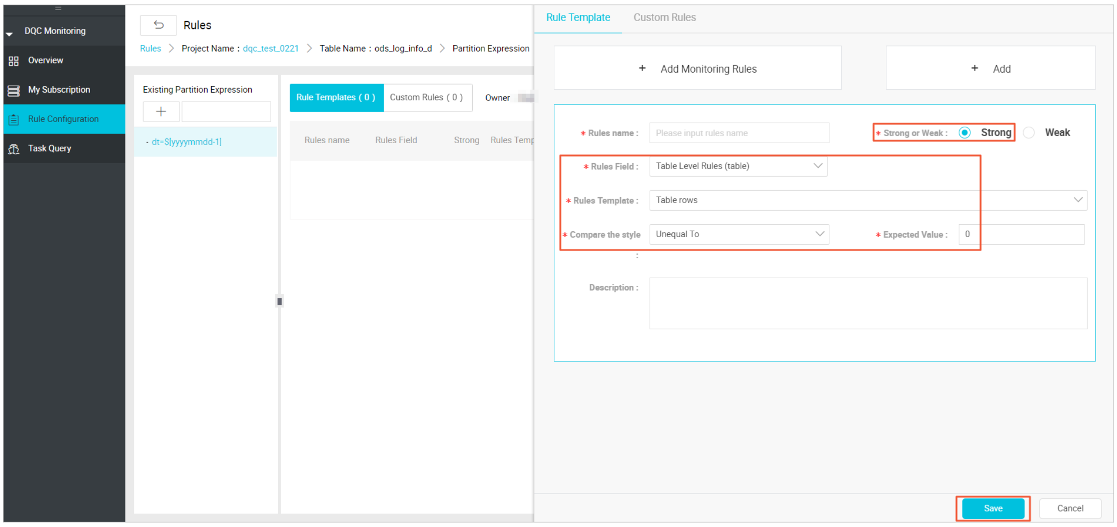
> **Note** The preceding configuration is to avoid duplicate data, which prevents downstream data from being affected by dirty data.

4. Configure a monitoring rule for the ods_log_info_d table.

The ods_log_info_d table stores the data that is parsed from the ods_raw_log_d table. The log data in the preceding table does not need to be monitored. You can configure only a monitoring rule to verify that the table contains data.

i. Add the partition filter expression dt=$[yyyymmdd-1].

ii. Click **Create rules** and click **Add Monitoring Rule** in the panel that appears.



Rules > Project Name : dqc_test_0221 > Table Name : ods_log_info_d > Partition Expression

- dt=$[yyyymmdd-1]

* Rules name : Please input rules name
* Strong or Weak : ● Strong ○ Weak
* Rules Field : Table Level Rules (table)
* Rules Template : Table rows
* Compare the style : Unequal To
* Expected Value : 0
Description :

Configure a monitoring rule to verify that the table contains data: Set the **Rule Type** parameter to **Rule Type**, the **Template** parameter to **All Fields in Table(table)**, the **Comparison Method** parameter to **Unequal To**, and the **Expected Value** parameter to **0**.

iii. Then, click **Batch Create**.

5. Configure a monitoring rule for the dw_user_info_all_d table.

The dw_user_info_all_d table aggregates data in the ods_user_info_d and ods_log_info_d tables. The workflow is simple, and a monitoring rule has been configured for the ods_user_info_d table to verify that the table contains data. Therefore, a monitoring rule for the dw_user_info_all_d table is not required. This saves computing resources.

6. Configure monitoring rules for the rpt_user_info_d table.

   The rpt_user_info_d table stores the data aggregation results. You can configure rules to monitor the number of rows in the table for any changes and verify that the primary key values are unique.

   i. Click the plus sign (**+**) in the **Partition Expression** section. Select the partition filter expression dt=$[yyyymmdd-1].

   ii. Click **Create rules**. In the panel that appears, click **Add Monitoring Rule** to configure a monitoring rule for the primary key values. Set the Field parameter to uid(string), the **Template** parameter to **Repeated value, fixed value**, the **Rule Type** parameter to **Soft**, the **Comparison Method** parameter to **Less Than**, and the **Expected Value** parameter to **1**.

   iii. Configure a rule to monitor the number of rows in the table for any changes: Set the **Template** parameter to **Number of rows, 7-day volatility**, the **Rule Type** parameter to **Soft**, the **Warning Threshold** parameter to 1%, and the **Error Threshold** parameter to 50%. Adjust the thresholds based on your business logic.

   > Note
   >
   > - The values of the **Warning Threshold** and **Error Threshold** parameters must be greater than 0%.
   > - The purpose of monitoring the number of rows is to monitor the fluctuations of daily unique visitors (UVs). Therefore, you can keep up with the traffic changes of the application at the earliest opportunity.

   iv. Then, click **Batch Create**.

A hard rule is more likely to be configured for a table at the operational data store (ODS) layer in a data warehouse. This is because data at the ODS layer is used as source data in the data warehouse and must be accurate to prevent data at other layers from being affected.

Data Quality also provides the **Node Query** module, where you can view the monitoring results of configured rules. For more information, see View monitoring results.

1.6. Visualize data on a dashboard

After you process the rpt_user_info_d table by backfilling data, you can create a dashboard in the Quick BI console to visualize the profile analysis results of website users in this table.

Prerequisites

The data that you want to visualize is processed. For more information, see Process data. You have logged on to the Quick BI console.

Context

The rpt_user_info_d table contains fields such as region, device, gender, age, and zodiac. You can view the following data on a dashboard: core metrics, periodic changes, regional distribution, age and zodiac distribution, and records of users. If you want to view changes in data over a specified period of time, we recommend that you backfill data of at least one week.

Procedure

1. In the Workspaces section of the Quick BI console, click **Default Workspace** to go to the default workspace. You can also click **Personal Workspace** to go to your personal workspace.

2. On the page that appears, click **Data Sources** in the left-side navigation pane. On the Data Sources page, click **Create Data Source** in the upper-right corner. In the Add Data Source dialog box, select **MaxCompute** on the **Cloud Data Sources** tab.

3. In the Add MaxCompute Database dialog box, enter the following information: the display name of the MaxCompute data source, the name of your MaxCompute project, your AccessKey ID, and your AccessKey secret. Use the default value for Database Address. For more information about the endpoint of MaxCompute, see Endpoints.

   Click **Test Connection**. When the message **The data source can be connected.** appears, click **Add**.

4. On the page that appears, find the rpt_user_info_d table and click **Create Dataset** in the Actions column.

   In the Create Dataset dialog box, enter the dataset name, select a location to store the dataset, and then click **OK**.

5. On the Datasets page, click the created dataset to edit the dataset.

   You can perform the following operations on a dataset: change dimensions and measures, change the dimension type, add calculated fields, create hierarchies, change the data type of a field, modify the aggregate mode of measures, and create association models.

6. Change the dimension types of fields. After you change the dimension types of fields, you can filter data based on the field values.

   i. Change the dimension type of the dt field.

      In the left-side navigation pane, right-click the dt field. Then, choose **Change Dimension Type > Date/Time (Source Format) > yyyyMMdd**.

   ii. Change the dimension type of the region field.

      In the left-side navigation pane, right-click the region field. Then, choose **Change Dimension Type > Geo > State/Province/Municipality**. After you change the dimension type of the region field, a location icon appears to the left of the field in the left-side navigation pane.

7. Create a dashboard.

   You can create a dashboard to display the most recent data. To create a dashboard, configure the display content, layout, and style. Then, create charts, and associate charts to enable filter interaction.

   i. On the Datasets page, find the rpt_user dataset and click Create Dashboard in the Actions column. In the Create Dashboard dialog box, select **Standard** for Select Dashboard Type. The dashboard edit page appears.

      > ? Note
      >
      > - If you use Quick BI Basic or Quick BI Pro, a dashboard of the standard type is created after you click Create Dashboard in the Actions column.
      > - If you use Quick BI Enterprise Standard, the Create Dashboard dialog box appears after you click Create Dashboard in the Actions column. You can create a dashboard of the standard type or full screen type in the Create Dashboard dialog box. In this topic, a dashboard of the standard type is created.

ii. In the upper part of the page, click the Kanban icon. A chart sample appears in the blank area.

On the Data tab on the right, select the rpt_user dataset from the drop-down list in the upper-right corner. Then, drag pv from the Measures list to the Metrics (Mea.) field. The rpt_user_info_d table is a partitioned table. You must select a dimension under dt, drag the dimension to the **Filters** field, and then click the Filter icon next to the dimension. In the Set Filter dialog box, specify a time range. In this example, the specified time range is 2019 to 2019. Then, click **Update** in the lower part of the Data tab.

iii. Create a trend chart. In the upper part of the page, click the **Line Chart** icon. A line chart sample appears in the blank area.

Set the parameters on the Data tab and click **Update**.

- **Value Axis (Mea.)**: Set the value to pv.
- **Category Axis (Dim.)**: Set the value to dt(day).
- **Color Legend (Dim.)**: Set the value to age_range.
- **Filters**: Drag dt(year) to this field.

iv. Create a filled map. In the upper part of the page, click the **Colored Map** icon. A map sample appears in the blank area. On the Data tab on the right side, select the rpt_user dataset from the drop-down list in the upper-right corner, drag region from the Dimensions list to the **Geo Location (Dim.)** field, and then drag pv from the Measures list to the **Colorscale (Mea.)** field. Then, click **Update**.

v. In the upper-right corner, click **Save** and then click **Preview** to view the created dashboard.

1.7. Use Function Studio to develop a UDF

This topic describes how to use Function Studio to develop a user-defined function (UDF) and commit the UDF to the development environment in DataStudio.

Limits

Function Studio is available only in the China (Beijing), China (Shanghai), China (Shenzhen), and China (Hangzhou) regions.

Create a project

If you have Git code, you can import the Git code to create a project. You can import Git code only from code.aliyun.com.

1. Go to the **DataStudio** page.

   i. Log on to the DataWorks console.

   ii. In the left-side navigation pane, click **Workspaces**.

   iii. In the top navigation bar, select the region in which the workspace that you want to manage resides. Find the workspace and click **DataStudio** in the Actions column.

2. On the DataStudio page, click the ☰ icon in the upper-left corner and choose **All Products > Data Development > Function Studio**.

3. On the **Projects** page, click **Import Git Project**.

4. On the **Create Project** page, set the **Git Repo URL**, **Name**, and **Description** parameters and select a runtime environment.

   By default, the new project is not associated with the Git service. Therefore, the **Settings** dialog box appears after you set the Git Repo URL parameter. In this dialog box, set parameters on the **SSH Key**, **Git Config**, and **Preference** tabs and click **Save**.

   - On the **SSH Key** tab, set the **Service** parameter to `code.aliyun.com`, click **Create SSH Key** to generate a Secure Shell (SSH) public key, and then click **Save**.
   - On the **Git Config** tab, set the **Username** and **Email** parameters and click **Save**.
   - On the **Preference** tab, set the **Font Size in Editor** parameter based on your needs and click **Save**.

   > ? **Note** If you want to modify Git settings after you create the project, go to the project editing page, move the pointer over **Settings** in the top navigation bar, and then select SSH Key, Git Config, or Preference as required.

5. Click **Submit**.
   After you create the project, Function Studio automatically pulls Git code from the project.

Add the SSH public key to code.aliyun.com

After you set parameters on the **SSH Key**, **Git Config**, and **Preference** tabs, you can add the SSH public key to code.aliyun.com.

1. Visit code.aliyun.com and click **Settings** in the left-side navigation pane.
2. On the page that appears, click **SSH Public Key** in the left-side navigation pane. On the SSH Public Key page, click **Add SSH Public Key** in the upper-right corner.
3. On the **Add SSH Public Key** page, enter the SSH public key that is generated on the **SSH Key** tab and click **Add**.

Test the class to run

1. On the project editing page, double-click the class that you want to test in the left-side navigation pane and click the Run Program icon in the upper-right corner.
2. In the **Run/Debug Configurations** dialog box, set parameters for the class.
3. Click **Run**. The test results appear in the Output section.

   > ? **Note**
   >
   > - The initial running process takes a longer time period. When you run the class again, the running duration is similar to that in a local integrated development environment (IDE).
   > - You can also select the class from the drop-down list in the upper-right corner and click the Run Program icon.

Commit the UDF and resource to the development environment in DataStudio

After you verify that code is correct, commit the UDF and resource to the development environment in DataStudio.

- Commit the resource to the development environment in DataStudio.
  1. On the project editing page, move the pointer over the **Submit** icon and select **Submit Resource to Development Environment**.
  2. In the **Submit Resource to DataStudio Development Environment** dialog box, set the **Target Workspace**, **Target Workflow**, and **Resource** parameters.
  3. Click **OK**.
- Commit the UDF to the development environment in DataStudio.
  1. On the project editing page, move the pointer over the **Submit** icon and select **Submit Function to Development Environment**.
  2. In the **Submit Function to DataStudio Development Environment** dialog box, set the **Target Workspace**, **Target Workflow**, **Resource**, **Class**, and **Function** parameters.
  3. Click **OK**.

After you commit the resource and UDF to the development environment in DataStudio, you can use them in SQL nodes.

2.DataWorks for EMR Workshop

2.1. Prepare the environment

To make sure that you can complete the workshop, you must activate E-MapReduce (EMR), DataWorks, and Object Storage Service (OSS) for your Alibaba Cloud account.

Prerequisites

- An Alibaba Cloud account is created.
- Real-name verification for individuals or enterprises is completed.
- An EMR compute engine instance is associated with your workspace. The EMR folder is displayed only after you associate an EMR compute engine instance with the workspace on the Workspace Management page. For more information, see Configure a workspace.
- An Alibaba Cloud EMR cluster is created, and an inbound rule that contains the following content is added to the security group to which the cluster belongs.
    - Action: Allow
    - Protocol type: Custom TCP
    - Port range: 8898/8898
    - Authorization object: 100.104.0.0/16
- If you integrate Hive with Ranger in EMR, you must modify whitelist configurations and restart Hive before you develop EMR Hive nodes in DataWorks. Otherwise, the error message **Cannot modify spark.yarn.queue at runtime** or **Cannot modify SKYNET_BIZDATE at runtime** is returned when you run EMR Hive nodes.
    i. You can modify the whitelist configurations by using custom parameters in EMR. You can append key-value pairs to the value of a custom parameter. In this example, the custom parameter for Hive components is used. The following code provides an example:

    ```
    hive.security.authorization.sqlstd.confwhitelist.append=tez.*|spark.*|mapred.*|mapreduce.*|ALISA.*|SKYNET.*
    ```

    Note In the preceding code, `ALISA.*` and `SKYNET.*` are specific to DataWorks.

    ii. After the whitelist configurations are modified, you must restart the Hive service to make the configurations take effect. For more information, see Restart a service.
- An exclusive resource group for scheduling is created, and the resource group is associated with the virtual private cloud (VPC) where the EMR cluster resides. For more information, see Create and use an exclusive resource group for scheduling.

    Note You can use only exclusive resource groups for scheduling to run EMR Hive nodes.

Context

The following Alibaba Cloud services are used in this workshop:

- E-MapReduce
- DataWorks

- OSS

Procedure

1. Create an EMR cluster.

    i. Log on to the EMR console.

    ii. In the top navigation bar, select the **China (Shanghai)** region. On the Cluster Management page, click **Cluster Wizard**.

    > Note
    >
    > - Source data used in this workshop is stored in the China (Shanghai) region. Therefore, we recommend that you create an EMR cluster in the same region as the source data.
    > - You can select **Quick Purchase** or **Cluster Wizard** to create an EMR cluster. In this topic, Cluster Wizard is selected.

    iii. On the **Cluster Wizard** page, set the **Cluster Type** parameter to **Hadoop** and use the default values for other parameters in the **Software Settings** step. Click **Next: Hardware Settings**.

    iv. In the **Hardware Settings** step, set the **Billing Method** parameter to **Pay-As-You-Go**, set the parameters in the **Network Settings** and **Instance** sections, and then click **Next: Basic Settings**.

    v. In the **Basic Settings** step, set the **Cluster Name** parameter, select a key pair from the **Key Pair** drop-down list, and then click **Next: Confirm**.

    By default, Assign Public IP Address is turned off. If you do not turn on this switch, you cannot access the cluster over the Internet after the cluster is created. In this workshop, you are not required to assign a public IP address. Therefore, click **Next** in the **Assign Public IP Address** dialog box. To access the cluster over the Internet, log on to the Elastic Compute Service (ECS) console and assign an elastic IP address (EIP) to the ECS instance that corresponds to the cluster.

    vi. In the **Confirm** step, verify your configuration, read the terms of service, select **E-MapReduce Service Terms**, and then click **Create**.

2. Initialize the cluster.

    After the purchase is complete, view the created EMR cluster on the **Cluster Management** page. It takes a few minutes to initialize the cluster.

    i. After the cluster is initialized, click the **Data Platform** tab.

    ii. On the **Data Platform** tab, click **Create Project** in the upper-right corner.

    iii. In the **Create Project** dialog box, set the **Project Name** and **Project Description** parameters.

    > Note Use your Alibaba Cloud account to create the project. The project must be associated with a DataWorks workspace in subsequent steps.

    iv. Click **Create**.

3. Create a DataWorks workspace.

> **Note** Data resources provided for this workshop are all stored in the China (Shanghai) region. Therefore, we recommend that you create a workspace in the China (Shanghai) region. Otherwise, the network connectivity test fails when you create a connection.

i. Move the pointer over the ☰ icon in the upper-left corner of the EMR console and choose **Products and Services > DTplus > DataWorks**.

ii. In the left-side navigation pane, click **Workspaces**.

iii. In the top navigation bar, select a region where you want to create a workspace.

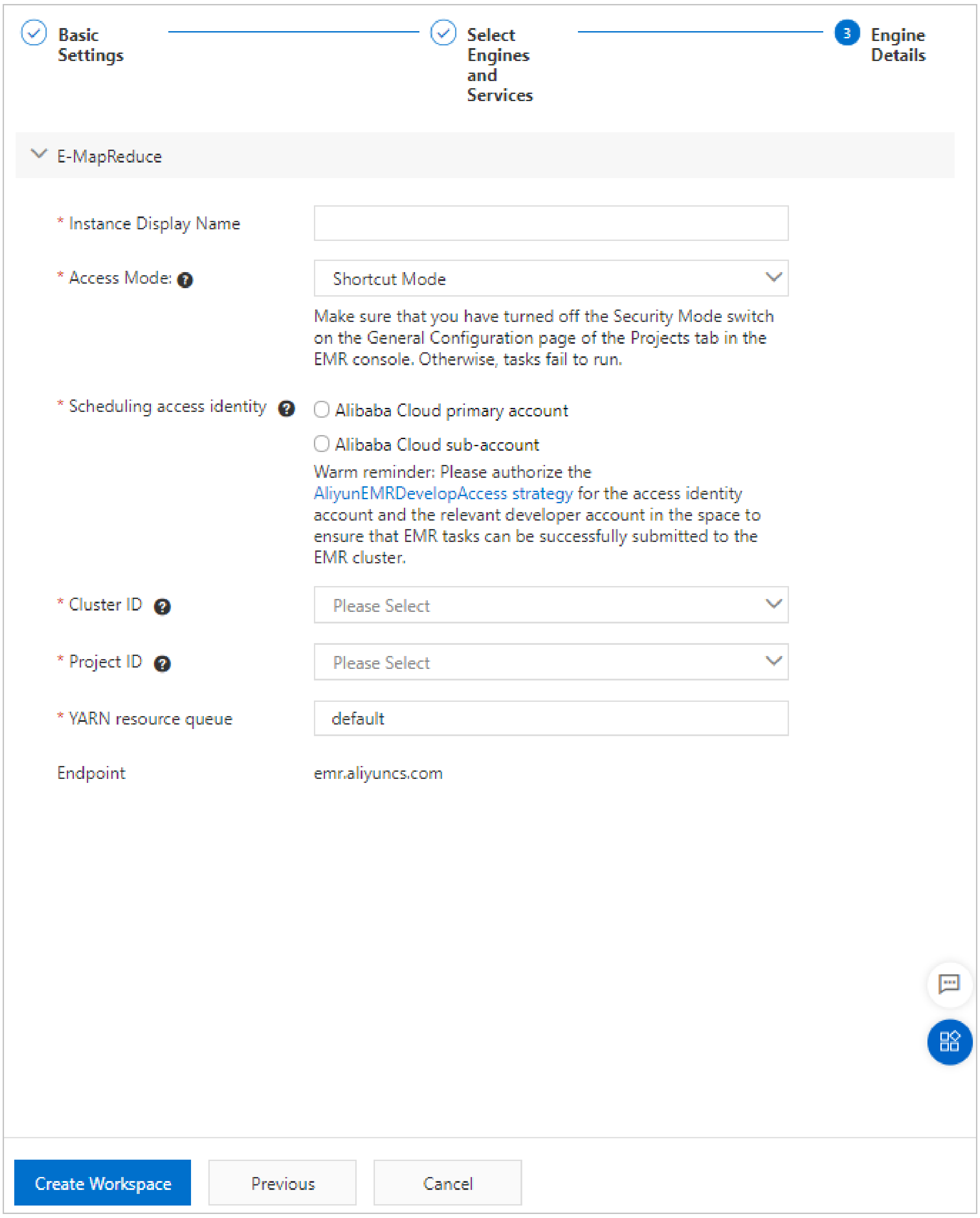
iv. On the Workspaces page, click **Create Workspace**. In the Create Workspace panel, set the parameters in the **Basic Settings** step and click **Next**.

Section	Parameter	Description
Basic Information	**Workspace Name**	The name of the workspace. The name must be 3 to 27 characters in length and start with a letter. It can contain only letters, underscores (_), and digits.
	Display Name	The display name of the workspace. The display name can be up to 27 characters in length. It must start with a letter and can contain only letters, underscores (_), and digits.
	Mode	Valid values: **Basic Mode (Production Environment Only)** and **Standard Mode (Development and Production Environments)**. In this topic, set the parameter to **Basic Mode (Production Environment Only)**.
	Description	The description of the workspace.
Advanced Settings	**Download SELECT Query Result**	Specifies whether to allow workspace members to download the results queried in DataStudio.

v. In the **Select Engines and Services** step, select **E-MapReduce** and click **Next**.

DataWorks is now available as a commercial service. If you have not activated DataWorks in a region, activate it before you create a workspace in the region.

vi. In the **Engine Details** step, set the parameters based on your business requirements.



* Instance Display Name

* Access Mode: Shortcut Mode

Make sure that you have turned off the Security Mode switch on the General Configuration page of the Projects tab in the EMR console. Otherwise, tasks fail to run.

* Scheduling access identity

○ Alibaba Cloud primary account

○ Alibaba Cloud sub-account

Warm reminder: Please authorize the AliyunEMRDevelopAccess strategy for the access identity account and the relevant developer account in the space to ensure that EMR tasks can be successfully submitted to the EMR cluster.

* Cluster ID: Please Select

* Project ID: Please Select

* YARN resource queue: default

Endpoint: emr.aliyuncs.com

Parameter	Description
Instance Display Name	The display name of the compute engine instance.
Access ID	The AccessKey ID of the account that is authorized to access the EMR cluster.
Access Key	The AccessKey secret of the account that is authorized to access the EMR cluster.
EmrClusterID	The ID of the EMR cluster. You can obtain the ID from the EMR console.

Parameter	Description
Cluster ID	The ID of the user who created the EMR cluster.
Project ID	The ID of the project in the EMR cluster.
YARN resource queue	The name of the resource queue in the EMR cluster. Unless otherwise specified, set the parameter to *default*.
Endpoint	The endpoint of the EMR cluster. You can obtain the endpoint from the EMR console.

vii. After the configuration is complete, click **Create Workspace**.

4. Activate OSS and create a bucket.

i. Activate OSS. For more information, see Activate OSS.

ii. Log on to the OSS console.

iii. In the left-side navigation pane, click **Buckets**.

iv. On the **Buckets** page, click **Create Bucket**.

v. In the **Create Bucket** panel, set the parameters and click **OK**.

> ⓘ **Note** Select **China (Shanghai)** from the **Region** drop-down list. For more information about the parameters, see Create buckets.

vi. Click the name of the created bucket in the **Bucket Name** column to go to the **Files** page.

vii. Click **Create Folder** on the Files page.

viii. In the **Create Folder** panel, set the **Folder Name** parameter and click **OK**.

> ⓘ **Note** Create three folders to store external data sources of OSS, Relational Database Service (RDS), and JAR resources.

2.2. Collect data

This topic describes how to use DataWorks to collect data to an E-MapReduce compute engine instance.

Prerequisites

The environment is prepared for performing the operations that are described in this topic. For more information, see Prepare the environment.

Context

In this workshop, you must create connections to an Object Storage Service (OSS) bucket and an ApsaraDB RDS instance from which you want to read data. You must also create a connection to an OSS bucket to which you want to write data.

Create a connection to an OSS bucket from which you want to read data

1. Go to the **Data Source** page.
   - i. Log on to the DataWorks console.
   - ii. In the left-side navigation pane, click **Workspaces**.
   - iii. On the Workspaces page, find the workspace in which you want to create a connection and click **Data Integration** in the Actions column.

     If you are using another service of DataWorks, click the ☰ icon in the upper-left corner and choose **All Products > Data Aggregation > Data Integration** to go to the **Data Integration** page.
   - iv. On the page that appears, click **Connection** in the left-side navigation pane. The **Data Source** page appears.
2. On the **Data Source** page, click **New data source** in the upper-right corner.
3. In the **Add data source** dialog box, click **OSS** in the Semi-structuredstorage section.
4. In the **Add OSS data source** dialog box, set the parameters as required. The following table describes how to set the parameters in this workshop.

Parameter	Description
Data Source Name	The name of the connection. Enter oss_workshop_log.
Data source description	The description of the connection.
Endpoint	The OSS endpoint. Enter `http://oss-cn-shanghai-internal.aliyuncs.com` .
Bucket	The name of the OSS bucket. Enter new-dataworks-workshop.
AccessKey ID	The AccessKey ID that is used to connect to OSS. Enter LTAI4FvGT3iU4xjKotpU****.
AccessKey Secret	The AccessKey secret that is used to connect to OSS. Enter 9RSUoRmNxpRC9EhC4m9PjuG7Jzy7px.

5. On the **Data Integration** tab, click **Test connectivity** in the Operation column of each resource group.
6. After the connection passes the connectivity test, click **Complete**.

Create a connection to an ApsaraDB RDS instance from which you want to read data

1. On the **Data Source** page, click **New data source** in the upper-right corner.
2. In the **Add data source** dialog box, click **MySQL** in the Relational Database section.
3. In the **Add MySQL data source** dialog box, set the parameters as required. The following table describes how to set the parameters in this workshop.

Parameter	Description
Data source type	The type of the connection. Select **Alibaba Cloud instance mode**.

Parameter	Description
Data Source Name	The name of the connection. Enter rds_workshop_log.
Data source description	The description of the connection.
Region	The region where the ApsaraDB RDS instance resides. Select China East 2 (Shanghai).
RDS instance ID	The ID of the ApsaraDB RDS instance. Enter rm-2ev0681lc7042g16u.
RDS instance account ID	The ID of the Alibaba Cloud account that is used to purchase the ApsaraDB RDS instance. Enter 5600815724958382.
Database name	The name of the ApsaraDB RDS database. Enter workshop.
User name	The username that is used to connect to the database. Enter workshop.
Password	The password that is used to connect to the database. Enter workshop#2017.

4. On the **Data Integration** tab, click **Test connectivity** in the Operation column of each resource group.
5. After the connection passes the connectivity test, click **Complete**.

Create a connection to an OSS bucket to which you want to write data

In this workshop, data of the E-MapReduce compute engine instance is stored in the OSS bucket that you created in the Prepare the environment topic.

1. On the **Data Source** page, click **New data source** in the upper-right corner.
2. In the **Add data source** dialog box, click **OSS** in the Semi-structuredstorage section.
3. In the **Add OSS data source** dialog box, set the parameters as required.

Parameter	Description
Data Source Name	The name of the connection.
Data source description	The description of the connection.
Endpoint	The OSS endpoint. Enter `http://oss-cn-shanghai-internal.aliyuncs.com`.
Bucket	The name of the OSS bucket that you created in the Prepare the environment topic. Enter dw-emr-demo.

Parameter	Description
AccessKey ID	The AccessKey ID of the account that is used to log on to DataWorks. You can go to the **Security Management** page to copy the AccessKey ID.
AccessKey Secret	The AccessKey secret of the account that is used to log on to DataWorks.

4. On the **Data Integration** tab, click **Test connectivity** in the Operation column of each resource group.
5. After the connection passes the connectivity test, click **Complete**.

Create a workflow

1. On the Data Source page, click the ☰ icon in the upper-left corner and choose **All Products** > **Data Development** > **DataStudio**. The DataStudio page appears.
2. On the **Data Analytics** tab, right-click **Business Flow** and select **Create Workflow**.
3. In the **Create Workflow** dialog box, set the **Workflow Name** and **Description** parameters.

   > **Notice** The workflow name can be up to 128 characters in length and can contain letters, digits, underscores (_), and periods (.).

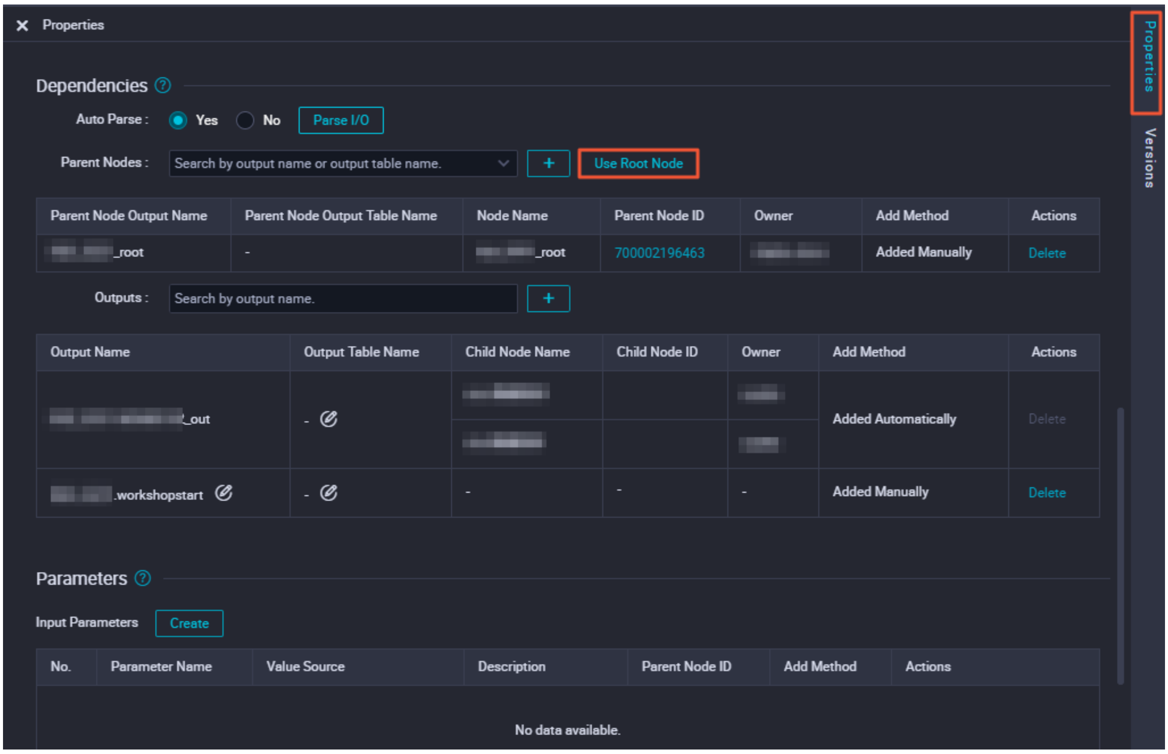
4. Click **Create**.
5. Double-click the new workflow to go to the workflow configuration tab. Drag **Zero-Load Node** under General to the canvas on the right. In the **Create Node** dialog box, set the **Node Name** parameter to workstart and click **Commit**.

   Drag **Batch Synchronization** under Data Integration to the canvas on the right to create two batch sync nodes named Log2oss and User2oss.
6. Drag directed lines to configure the workshopstart node as the parent node of the two batch sync nodes.

Configure the workstart node

1. On the **Data Analytics** tab, double-click the workstart node in the new workflow. On the node configuration tab that appears, click **Properties** in the right-side navigation pane.
2. In the **Dependencies** section, click **Use Root Node** to set the root node of the workspace as the parent node of the workstart node.

   In the latest version of DataWorks, each node must have its parent and child nodes. Therefore, you must set a parent node for the workstart node. In this workshop, the root node of the workspace is set as the parent node of the workstart node. The root node of the workspace is named in the Workspace name_root format.



× Properties

Dependencies

Auto Parse : ● Yes ○ No Parse I/O

Parent Nodes : Search by output name or output table name. + Use Root Node

Parent Node Output Name	Parent Node Output Table Name	Node Name	Parent Node ID	Owner	Add Method	Actions
[redacted]_root	-	[redacted]_root	700002196463	[redacted]	Added Manually	Delete

Outputs : Search by output name. +

Output Name	Output Table Name	Child Node Name	Child Node ID	Owner	Add Method	Actions
[redacted]_out	-	[redacted]		[redacted]	Added Automatically	Delete
		[redacted]		[redacted]		
[redacted].workshopstart	-	-	-	-	Added Manually	Delete

Parameters

Input Parameters Create

No.	Parameter Name	Value Source	Description	Parent Node ID	Add Method	Actions

No data available.

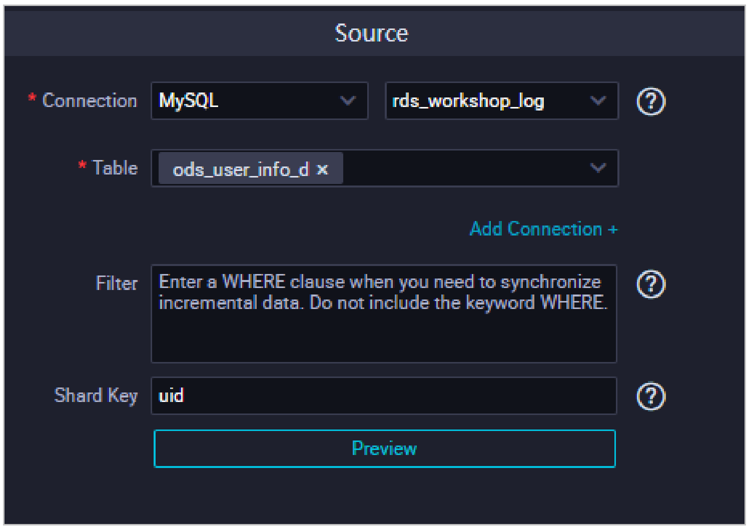
3. Click the 💾 icon in the upper-left corner.

Configure batch sync nodes

1. Configure the User2oss node.

   i. On the **Data Analytics** tab, double-click the User2oss node in the new workflow. The node configuration tab appears.

ii. Configure a connection to the source data store.



Source

* Connection: MySQL ▾ rds_workshop_log ▾ ?

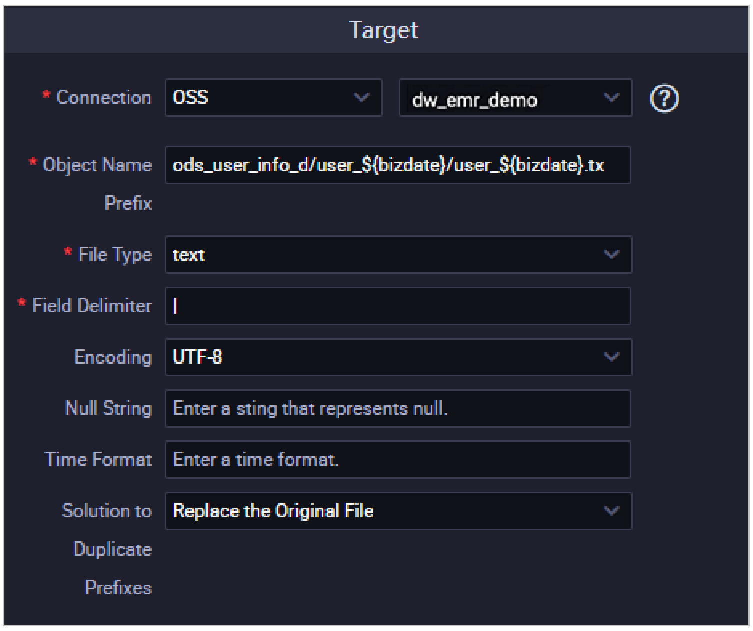
* Table: ods_user_info_d × ▾

Filter: Enter a WHERE clause when you need to synchronize incremental data. Do not include the keyword WHERE. ?

Shard Key: uid ?

Parameter	Description
Connection	The type and name of the connection. Select **MySQL** and **rds_workshop_log** in sequence.
Table	The table from which data is synchronized. Select the ods_user_info_d table.
Filter	The filter condition for the data to be synchronized. Filtering based on the limit keyword is not supported. You can leave this parameter unspecified.
Shard Key	The shard key for the data to be synchronized. We recommend that you use the primary key or an indexed column as the shard key. Only fields of the INTEGER type are supported. In this workshop, enter uid.

iii. Configure a connection to the destination data store.



Target

* Connection: OSS | dw_emr_demo

* Object Name Prefix: ods_user_info_d/user_${bizdate}/user_${bizdate}.tx

* File Type: text

* Field Delimiter: |

Encoding: UTF-8

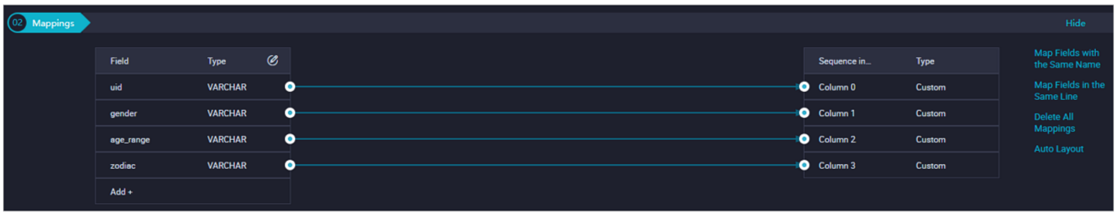
Null String: Enter a sting that represents null.

Time Format: Enter a time format.

Solution to Duplicate Prefixes: Replace the Original File

Parameter	Description
Connection	The type and name of the connection. Select **OSS** and **dw_emr_demo** in sequence.
Object Name Prefix	The prefix of the OSS object for storing synchronized data. Set this parameter based on the folder that you created. In this workshop, enter ods_user_info_d/user_${bizdate}/user_${bizdate}.txt.
File Type	The object type. Select text.
Field Delimiter	The column delimiter. Enter |.
Encoding	The encoding format. Default value: UTF-8.
Null String	The string that represents null. You can leave this parameter unspecified.
Time Format	The time format. You can leave this parameter unspecified.
Solution to Duplicate Prefixes	The method that is used to process duplicate prefixes. Select **Replace the Original File**.

iv. Configure the mappings between fields in the source and destination.

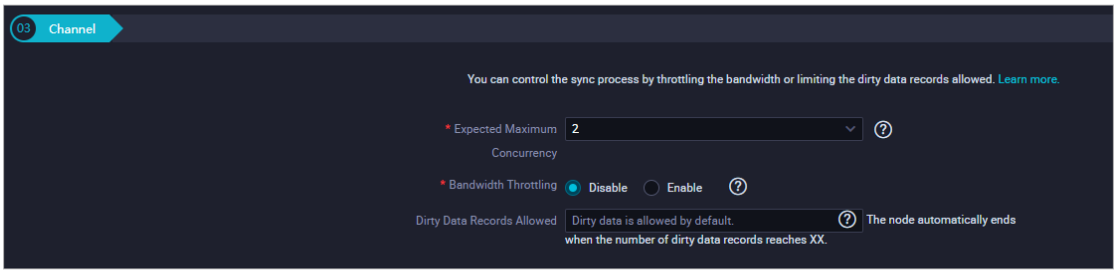


02 Mappings

Field	Type
uid	VARCHAR
gender	VARCHAR
age_range	VARCHAR
zodiac	VARCHAR
Add +	

Sequence in...	Type
Column 0	Custom
Column 1	Custom
Column 2	Custom
Column 3	Custom

v. Configure channel control policies and click the icon in the top toolbar.



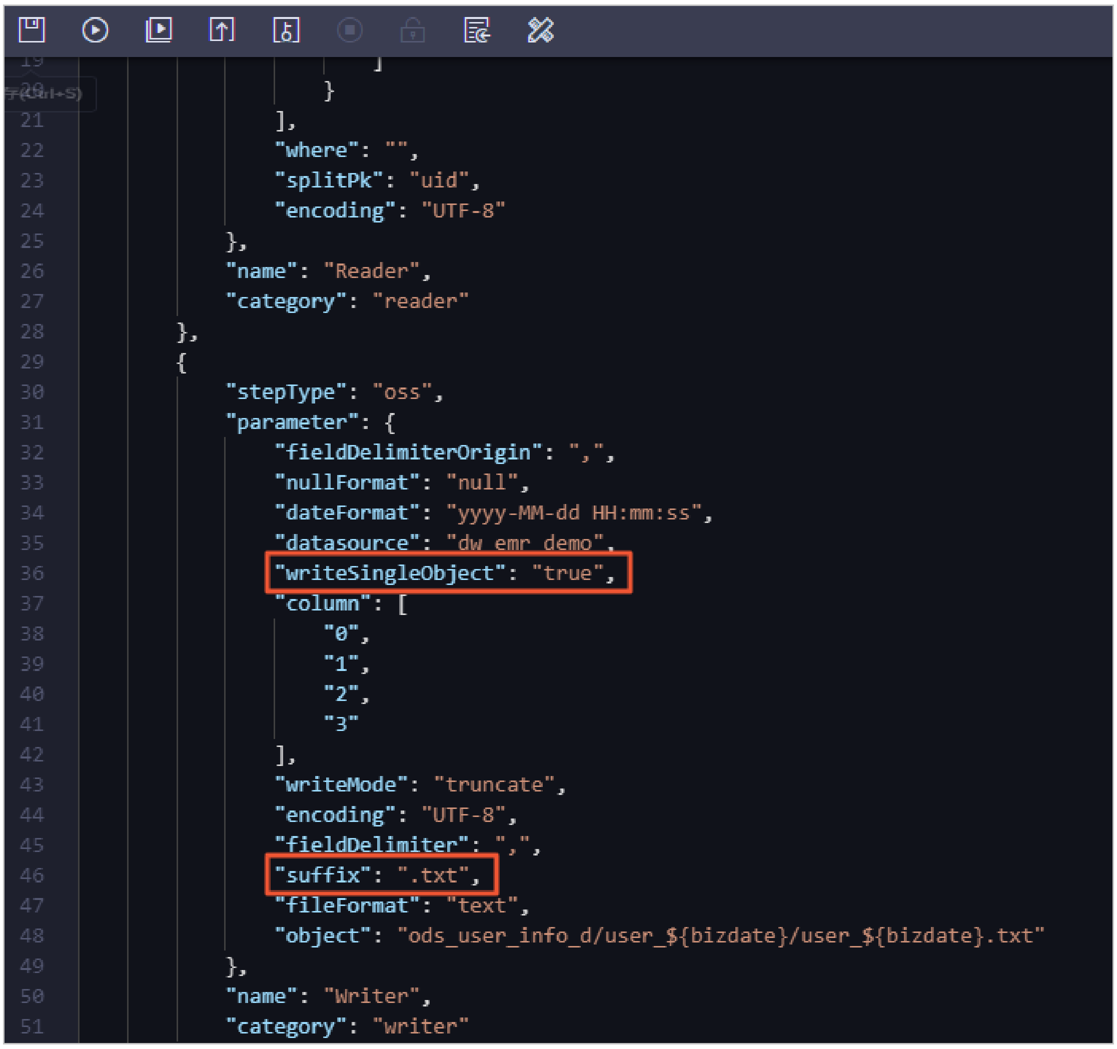
You can control the sync process by throttling the bandwidth or limiting the dirty data records allowed. Learn more.

* Expected Maximum Concurrency: 2

* Bandwidth Throttling: ● Disable ○ Enable

Dirty Data Records Allowed: Dirty data is allowed by default. The node automatically ends when the number of dirty data records reaches XX.

vi. Click the icon in the top toolbar. Add the following content to the existing code: "writeSingleObject": "true" and "suffix": ".txt".



```
                        ]
                    }
                ],
                "where": "",
                "splitPk": "uid",
                "encoding": "UTF-8"
            },
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "oss",
            "parameter": {
                "fieldDelimiterOrigin": ",",
                "nullFormat": "null",
                "dateFormat": "yyyy-MM-dd HH:mm:ss",
                "datasource": "dw_emr_demo",
                "writeSingleObject": "true",
                "column": [
                    "0",
                    "1",
                    "2",
                    "3"
                ],
                "writeMode": "truncate",
                "encoding": "UTF-8",
                "fieldDelimiter": ",",
                "suffix": ".txt",
                "fileFormat": "text",
                "object": "ods_user_info_d/user_${bizdate}/user_${bizdate}.txt"
            },
            "name": "Writer",
            "category": "writer"
```

Note

- You can add the writeSingleObject and suffix parameters only in the code editor.
- The value of the object parameter must be the same as the folder that is created in the OSS bucket.

vii. Click the icon in the top toolbar.

2. Configure the Log2oss node.

   i. On the **Data Analytics** tab, double-click the Log2oss node in the new workflow. The node configuration tab appears.

   ii. Configure a connection to the source data store.

Parameter	Description
Connection	The type and name of the connection. Select **OSS** and **oss_workshop_log** in sequence.
Object Name Prefix	The prefix of the OSS object for storing the data to be synchronized. Enter user_log.txt.
File Type	The object type. Select text.
Field Delimiter	The column delimiter. Enter |.
Encoding	The encoding format. Default value: UTF-8.
Null String	The string that represents null. You can leave this parameter unspecified.
Compression Format	The compression format of the OSS object. Valid values: None, Gzip, Bzip2, and Zip. Select None.
Include Header	Specifies whether to include the table header. Default value: No.

iii. Configure a connection to the destination data store.

Parameter	Description
Connection	The type and name of the connection. Select **OSS** and **dw_emr_demo** in sequence.
Object Name Prefix	The prefix of the OSS object for storing synchronized data. Set this parameter based on the folder that you created. In this workshop, enter ods_raw_log_d/user_log_${bizdate}/user_log_${bizdate}.txt.
File Type	The object type. Select text.
Field Delimiter	The column delimiter. Enter |.
Encoding	The encoding format. Default value: UTF-8.
Null String	The string that represents null. You can leave this parameter unspecified.
Time Format	The time format. You can leave this parameter unspecified.
Solution to Duplicate Prefixes	The method that is used to process duplicate prefixes. Select **Replace the Original File**.

iv. Configure the mappings between fields in the source and destination.

> Notice Only one column in the source table contains data. Remove the empty columns from the Source section.

v. Configure channel control policies and click the 💾 icon in the top toolbar.

vi. Click the ⟨/⟩ icon in the top toolbar. Add the following content to the existing code: "writeSingleObject": "true" and "suffix": ".txt".

> Note
>
> - You can add the writeSingleObject and suffix parameters only in the code editor.
> - The value of the object parameter must be the same as the folder that is created in the OSS bucket.

vii. Click the 💾 icon in the top toolbar.

Create tables to which you want to write data

1. On the **Data Analytics** tab, click the new workflow, right-click **EMR**, and then choose **Create > EMR Hive**.

2. In the **Create Node** dialog box, set the **Node Name** parameter and click **Commit**.

   In this workshop, you must create two EMR Hive nodes named ods_user_info_d and ods_raw_log_d. The former is used to create a table to store user information that is synchronized from ApsaraDB RDS and the latter is used to create a table to store logs that are synchronized from OSS.

3. On the configuration tab of each EMR Hive node, select an E-MapReduce compute engine instance, enter the table creation statements, and then click the **Save** and **Run** icons in sequence to execute the statements.

   ○ Create the ods_user_info_d table.

     Double-click the ods_user_info_d node. On the node configuration tab that appears, enter the table creation statements.

     ```
     CREATE EXTERNAL TABLE IF NOT EXISTS ods_user_info_d
     (
         `uid` STRING COMMENT 'User ID',
         `gender` STRING COMMENT 'Gender',
         `age_range` STRING COMMENT 'Age range',
         `zodiac` STRING COMMENT 'Zodiac sign'
     ) PARTITIONED BY (
       dt STRING
     )
     ROW FORMAT  delimited fields terminated by '|'
     LOCATION 'oss://dw-emr-demo/ods_user_info_d/';
     ALTER TABLE ods_user_info_d ADD IF NOT EXISTS PARTITION (dt=${bizdate})
     LOCATION 'oss://dw-emr-demo/ods_user_info_d/user_${bizdate}/';
     ```

     > **Note** In the preceding code, the path specified in the location parameter is used as an example. Set the parameter to the path of the created folder.

   ○ Create the ods_raw_log_d table.

     Double-click the ods_raw_log_d node. On the node configuration tab that appears, enter the table creation statements.

     ```
     -- Create a table to store logs that are synchronized from OSS.
     CREATE EXTERNAL TABLE IF NOT EXISTS ods_raw_log_d
     (
       `col` STRING
     ) PARTITIONED BY (
       dt STRING
     );
     ALTER TABLE ods_raw_log_d ADD IF NOT EXISTS PARTITION (dt=${bizdate})
     LOCATION 'oss://dw-emr-demo/ods_raw_log_d/user_log_${bizdate}/';
     ```

     > **Note** In the preceding code, the path specified in the location parameter is used as an example. Set the parameter to the path of the created folder.

4. View the data synchronization results.

   After the table creation statements are executed, enter a query statement on the configuration tab of each EMR Hive node.

> ⓘ **Note** In the query statement, change the partition key value to the data timestamp of the node. For example, if the node is run on November 7, 2019, the data timestamp is 20191106, which is one day before the node is run.

- Query data in the ods_user_info_d table.

  ```
  SELECT * from ods_user_info_d where dt=Data timestamp of the node; -- The data timest
  amp is one day before the node is run.
  ```

- Query data in the ods_raw_log_d table.

  ```
  SELECT * from ods_raw_log_d where dt=Data timestamp of the node; -- The data timestam
  p is one day before the node is run.
  ```

What to do next

You have learned how to collect and synchronize data. You can now proceed with the next tutorial. The next tutorial describes how to compute and analyze collected data. For more information, see Process data.

2.3. Process data

This topic describes how to create E-MapReduce Hive nodes to process collected logs in DataWorks.

Prerequisites

The data is collected. For more information, see Collect data.

Upload resources in the OSS console

1. Download the ip2region-emr.jar package and store it in a local directory.
2. Log on to the OSS console.
3. In the left-side navigation pane, click **Buckets**. On the Buckets page that appears, click the target bucket, for example, dw-emr-demo.
4. On the bucket page, click **Files** in the left-side navigation pane. On the Files page, click the folder that is created in the *Prepare the environment* topic for storing JAR resources, for example, ip2region.
5. Click **Upload** in the upper-left corner. In the **Upload** dialog box that appears, set parameters for uploading the ip2region-emr.jar package.

Parameter	Description
Upload To	The folder to which the package is uploaded. Set the value to **Current**. In this example, the folder is *oss://dw-emr-demo/ip2region/*.
File ACL	The access control list (ACL) of the package. The default value is **Inherited from Bucket**, which indicates that the ACL of each object is the same as that of the bucket.

Parameter	Description
Upload	Click **Upload** and select the downloaded ip2region-emr.jar package.

Design the workflow

For more information about how to configure the dependencies among nodes of a workflow, see Collect data.

In the DataStudio console, double-click the created workflow in the left-side navigation pane. On the workflow editing tab that appears, click and hold **EMR Hive** on the left and drag it to the editing section on the right. In the **Create Node** dialog box that appears, set **Node Name** and click **Commit**.

Create three E-MapReduce Hive nodes in total and name them ods_log_info_d, dw_user_info_all_d, and rpt_user_info_d respectively. Then, configure the dependencies among the nodes.

Configure the E-MapReduce Hive nodes

1. Configure the ods_log_info_d node.

   i. Double-click the ods_log_info_d node.

   ii. On the node editing tab that appears, enter the following statements:

   > **Note** If the current workspace is bound to multiple E-MapReduce compute engine instances, you must select an E-MapReduce compute engine instance. If the current workspace is only bound to one E-MapReduce compute engine instance, you do not need to do so.

```
-- Create a table at the ODS layer.
CREATE TABLE IF NOT EXISTS ods_log_info_d (
  ip STRING COMMENT 'The IP address of the client that sends the request',
  uid STRING COMMENT 'The ID of the client user',
  `time` STRING COMMENT 'The time when the user accessed the webpage, in the format of yyyymmddhh:mi:ss',
  status STRING COMMENT 'The status code returned by the server',
  bytes STRING COMMENT 'The number of bytes returned to the client',
  region STRING COMMENT 'The region where the user resides, which is obtained based on the IP address',
  method STRING COMMENT 'The type of the HTTP request',
  url STRING COMMENT 'The URL of the webpage accessed by the user',
  protocol STRING COMMENT 'The version number of HTTP',
  referer STRING COMMENT 'The URL of the webpage linked to the resource being requested',
  device STRING COMMENT 'The terminal type',
  identity STRING COMMENT 'The access type, which can be crawler, feed, user, or unknown'
)
PARTITIONED BY (
  dt STRING
);
create function  getregion as 'org.alidata.emr.udf.Ip2Region'
using jar 'oss://dw-emr-demo/ip2region/ip2region-emr.jar';
ALTER TABLE ods_log_info_d ADD IF NOT EXISTS PARTITION (dt=${bizdate});
```

```
set hive.vectorized.execution.enabled = false;
INSERT OVERWRITE TABLE ods_log_info_d PARTITION (dt=${bizdate})
SELECT ip
  , uid
  , tm
  , status
  , bytes
  , getregion(ip) AS region -- Obtain the region by using the user defined function
(UDF) based on the IP address.
  , regexp_extract(request, '(^[^ ]+) . *') AS method -- Use the regular expression
to extract three fields from the request.
  , regexp_extract(request, '^[^ ]+ (. *) [^ ]+$') AS url
  , regexp_extract(request, '. * ([^ ]+$)') AS protocol
  , regexp_extract(referer, '^[^/]+://([^/]+){1}') AS referer  -- Use the regular e
xpression to clean the HTTP referrer so as to obtain a more accurate URL.
  , CASE
    WHEN lower (agent) RLIKE 'android' THEN 'android' -- Obtain the terminal and ac
cess types from the value of the agent parameter.
    WHEN lower(agent) RLIKE 'iphone' THEN 'iphone'
    WHEN lower(agent) RLIKE 'ipad' THEN 'ipad'
    WHEN lower(agent) RLIKE 'macintosh' THEN 'macintosh'
    WHEN lower(agent) RLIKE 'windows phone' THEN 'windows_phone'
    WHEN lower(agent) RLIKE 'windows' THEN 'windows_pc'
    ELSE 'unknown'
  END AS device
  , CASE
    WHEN lower(agent) RLIKE '(bot|spider|crawler|slurp)' THEN 'crawler'
    WHEN lower(agent) RLIKE 'feed'
    OR regexp_extract(request, '^[^ ]+ (. *) [^ ]+$') RLIKE 'feed' THEN 'feed'
    WHEN lower(agent) NOT RLIKE '(bot|spider|crawler|feed|slurp)'
    AND agent RLIKE '^[Mozilla|Opera]'
    AND regexp_extract(request, '^[^ ]+ (. *) [^ ]+$') NOT RLIKE 'feed' THEN 'user'
    ELSE 'unknown'
  END AS identity
  FROM (
    SELECT SPLIT(col, '##@@')[0] AS ip
    , SPLIT(col, '##@@')[1] AS uid
    , SPLIT(col, '##@@')[2] AS tm
    , SPLIT(col, '##@@')[3] AS request
    , SPLIT(col, '##@@')[4] AS status
    , SPLIT(col, '##@@')[5] AS bytes
    , SPLIT(col, '##@@')[6] AS referer
    , SPLIT(col, '##@@')[7] AS agent
    FROM ods_raw_log_d
  WHERE dt = ${bizdate}
) a;
```

iii. Click the 💾 icon in the toolbar.

2. Configure the dw_user_info_all_d node.

   i. Double-click the dw_user_info_all_d node.

ii. On the node editing tab that appears, enter the following statements:

> **Note** If the current workspace is bound to multiple E-MapReduce compute engine instances, you must select an E-MapReduce compute engine instance. If the current workspace is only bound to one E-MapReduce compute engine instance, you do not need to do so.

```
-- Create a table at the DW layer.
CREATE TABLE IF NOT EXISTS dw_user_info_all_d (
  uid STRING COMMENT 'The ID of the client user',
  gender STRING COMMENT 'The gender of the user',
  age_range STRING COMMENT 'The age range of the user',
  zodiac STRING COMMENT 'The zodiac sign of the user',
  region STRING COMMENT 'The region where the user resides, which is obtained based on the IP address',
  device STRING COMMENT 'The terminal type',
  identity STRING COMMENT 'The access type, which can be crawler, feed, user, or unknown',
  method STRING COMMENT 'The type of the HTTP request',
  url STRING COMMENT 'The URL of the webpage accessed by the user',
  referer STRING COMMENT 'The URL of the webpage linked to the resource being requested',
  `time` STRING COMMENT 'The time when the user accessed the webpage, in the format of yyyymmddhh:mi:ss'
)
PARTITIONED BY (
  dt STRING
);
ALTER TABLE dw_user_info_all_d ADD IF NOT EXISTS PARTITION (dt = ${bizdate});
INSERT OVERWRITE TABLE dw_user_info_all_d PARTITION (dt=${bizdate})
SELECT COALESCE(a.uid, b.uid) AS uid
  , b.gender
  , b.age_range
  , b.zodiac
  , a.region
  , a.device
  , a.identity
  , a.method
  , a.url
  , a.referer
  , a.`time`
FROM (
  SELECT *
  FROM ods_log_info_d
  WHERE dt = ${bizdate}
) a
LEFT OUTER JOIN (
  SELECT *
  FROM ods_user_info_d
  WHERE dt = ${bizdate}
) b
ON a.uid = b.uid;
```

iii. Click the 💾 icon in the toolbar.

3. Configure the rpt_user_info_d node.
   i. Double-click the rpt_user_info_d node.
   ii. On the node editing tab that appears, enter the following statements:

> ⓘ **Note** If the current workspace is bound to multiple E-MapReduce compute engine instances, you must select an E-MapReduce compute engine instance. If the current workspace is only bound to one E-MapReduce compute engine instance, you do not need to do so.

```
-- Create a table at the RPT layer.
CREATE TABLE IF NOT EXISTS rpt_user_info_d (
  uid STRING COMMENT 'The ID of the client user',
  region STRING COMMENT 'The region where the user resides, which is obtained based on the IP address',
  device STRING COMMENT 'The terminal type',
  pv BIGINT COMMENT 'The number of times that the user viewed the webpage',
  gender STRING COMMENT 'The gender of the user',
  age_range STRING COMMENT 'The age range of the user',
  zodiac STRING COMMENT 'The zodiac sign of the user'
)
PARTITIONED BY (
  dt STRING
);
ALTER TABLE rpt_user_info_d ADD IF NOT EXISTS PARTITION (dt=${bizdate});
INSERT OVERWRITE TABLE rpt_user_info_d PARTITION (dt=${bizdate})
SELECT uid
  , MAX(region)
  , MAX(device)
  , COUNT(0) AS pv
  , MAX(gender)
  , MAX(age_range)
  , MAX(zodiac)
FROM dw_user_info_all_d
WHERE dt = ${bizdate}
GROUP BY uid;
```

   iii. Click the 💾 icon in the toolbar.

Commit the workflow

1. On the workflow editing tab, click the ▶ icon to run the workflow.
2. After the ✓ icon appears next to all nodes on the workflow editing tab, click the ⇧ icon to commit the workflow.
3. In the **Commit** dialog box that appears, select the nodes to be committed, and then select **Ignore I/O Inconsistency Alerts**.
4. Click **Commit**.

Run the nodes in the production environment

1. After you commit the workflow, click **Operation Center** in the upper-right corner.

   You can also click **Go to Operation Center** in the toolbar on the workflow editing tab to go to the **Operation Center** page.
2. On the Operation Center page, choose **Cycle Task Maintenance > Cycle Task** in the left-side navigation pane. On the **Cycle Task** page, click the workstart zero load node.
3. In the directed acyclic graph (DAG) on the right, right-click the workstart node and choose **Run > Current and Descendent Nodes Retroactively**.
4. In the Patch Data dialog box that appears, select a node to generate retroactive data, specify the data timestamp, and then click **OK**. The **Patch Data** page appears.
5. Click **Refresh** until the instance status is Successful.

What to do next

Now, you have learned how to create E-MapReduce Hive nodes and process raw logs. You can proceed with the next tutorial to learn how to collect metadata and view table information in Data Map. For more information, see Collect and view metadata.

2.4. Collect and view metadata

This topic describes how to collect metadata and view table information in Data Map.

Prerequisites

The data is processed. For more information, see Process data.

Collect metadata

1.
2.
3.
4.
5.
6.
7.
8. On the **Obtain Metadata from E-MapReduce** page, find the created crawler and click **Obtain All**.

   Click **Refresh** in the upper-right corner of the page and verify that the running status of the E-MapReduce data collection instance changes to **Collected successfully**.

   > **Note** After full metadata is collected from E-MapReduce, the system automatically collects incremental metadata and synchronizes new metadata from E-MapReduce tables.

View table information

1. In the top navigation bar of the current page, click **All Data**.
2. On the **All Data** page, click the **E-MapReduce** tab.

3. On the **E-MapReduce** tab, click the table that is named rpt_user_info_d to view the details of the table.

   You can also enter a keyword in the search box at the top to search for an E-MapReduce table and view the details of the table.
4. Click the **Lineage** tab to view the lineages of the table.

What to do next

You have learned how to collect metadata and view table information in Data Map. You can now proceed with the next tutorial to learn how to monitor and ensure the quality of the data that is generated by the developed nodes. For more information, see Configure rules to monitor data quality.

2.5. Configure rules to monitor data quality

This topic describes how to configure rules to monitor the data quality of the ods_log_info_d table.

Prerequisites

The metadata is collected. For more information, see Collect and view metadata.

Procedure

1. Go to the **DataStudio** page.
   i. Log on to the DataWorks console.
   ii. In the left-side navigation pane, click **Workspaces**.
   iii. In the top navigation bar, select the region in which the workspace that you want to manage resides. Find the workspace and click **DataStudio** in the Actions column.
2. Go to the **Monitoring Rules** page of the ods_log_info_d table.
   i. Click the ≡ icon in the upper-left corner and choose **All Products > Data Quality**.
   ii. In the left-side navigation pane, click **Monitoring Rules**. Select **EMR** from the Engine/Data Source drop-down list.
   iii. Find the ods_log_info_d table and click **View Monitoring Rules**.
3. Add a partition filter expression.
   i. Click **+** in the **Partition Expression** section.
   ii. In the **Add Partition** dialog box, set the **Partition Expression** parameter to dt=$[yyyymmdd-1] and select the corresponding data quality wrapper.
   iii. Click **Verify** to view the scheduling result.
   iv. Verify that the scheduling result is correct and click **OK**.
4. Create a monitoring rule.
   i. Select a partition and click **Create rules** in the upper-right corner.
   ii. On the **Template Rules** tab, click **Add Monitoring Rule**.

iii. Configure the monitoring rule.

Parameter	Description
Rule Name	The name of the monitoring rule.
Rule Type	The type of the monitoring rule. Set this parameter to **Rule Type**.
Auto-Generated Threshold	Specifies whether to use dynamic thresholds. Set this parameter as needed. ? **Note** You can use the dynamic threshold feature only in DataWorks Enterprise Edition or more advanced editions.
Rule Source	Valid values: **Built-in Template** and **Rule Templates**. ? **Note** You can select **Rule Templates** only in DataWorks Enterprise Edition or more advanced editions.
Field	Set this parameter to **All Fields in Table(table)**.
Template	Set this parameter to **Number of rows, fixed value**.
Comparison Method	Set this parameter to **Greater Than**.
Expected Value	Set this parameter to 0. In this case, you expect the actual value to be greater than 0.

iv. After the configuration is completed, click **Batch Create**.

5. Test the monitoring rule.

   i. Click **Test** in the upper-right corner of the page.

   ii. In the **Test** dialog box, set the **Data Timestamp** and **Resource Group** parameters and click **Test**.

   iii. After the test is completed, click **The test is complete. Click to view the results** to go to the page of the test results.

6. Link the monitoring rule to nodes.

   i. On the **Monitoring Rules** page of the ods_log_info_d table, click **Manage Linked Nodes**.

   ii. In the **Manage Linked Nodes** dialog box, enter the IDs or names of the nodes and click **Create**.

   iii. After the nodes are added, the monitoring rule is linked to the nodes. Verify that Data Quality checks the data quality of a node instance after the instance is run.

7. Configure subscriptions.

   i. On the **Monitoring Rules** page of the ods_log_info_d table, click **Manage Subscriptions**.

   ii. In the **Manage Subscriptions** dialog box, set the **Notification Method** and **Recipient** parameters.

       Data Quality supports the following notification methods: **Email**, **Email and SMS**, **DingTalk Chatbot**, and **DingTalk Chatbot @ALL**.

   iii. After the configuration is completed, click **Save**. You can go to the **My Subscriptions** page to view your subscriptions and modify the subscription configuration.

3.Automatically identify users who steal electricity

3.1. Overview

This tutorial describes how to use DataWorks together with Machine Learning Platform for AI (PAI) to automatically identify users who steal electricity. This ensures that users use electricity in a safe manner.

The traditional methods of identifying electricity theft and metering device failures include regular inspection, regular check of electricity meters, and reporting of electricity theft from users. These methods require manual operations. In addition, these methods are inefficient if you want to identify users who steal electricity or are involved in electricity leakage.

The staff of power supply bureaus, such as those who inspect electricity marketing and who check and meter electricity usage, use the existing automated system for monitor electricity usage online. The system triggers alerts for abnormal electricity usage and provides electricity usage data. The system collects data about abnormal electricity usage, abnormal load, abnormal line loss, and alerts that are reported by terminals and primary sites, and builds models for analyzing the data. This way, relevant staff can identify electricity theft, electricity leakage, and metering device failures in real time. After alerts are triggered, the system builds models for analyzing abnormal electricity usage based on the current, voltage, and load before and after the alert time. This helps identify electricity theft, electricity leakage, and metering device failures.

Information about abnormal electricity usage can be collected by using the traditional methods of identifying electricity theft and electricity leakage. However, due to frequent false positives and false negatives, these methods cannot precisely identify users who steal electricity or are involved in electricity leakage. In addition, experts need to determine the weight of each metric for the model to be built based on their knowledge and experience. This process is subjective.

The existing automated system for metering electricity usage can collect electricity load data, such as the current, voltage, and power data, and alert data that terminals report. Electrical inspection staff can also collect electricity theft and leakage data from the online inspection system or by conducting on-site inspection and enter the data into the system.

Based on the preceding data, DataWorks together with PAI can abstract key features of users who steal electricity or are involved in electricity leakage. In addition, DataWorks together with PAI can also build a model for identifying such users. This way, electricity theft or leakage can be automatically detected. This reduces the inspection workload of electrical inspection staff and ensures normal and secure electricity usage.

3.2. Prepare the environment

To make sure that you can complete the workshop, you must activate MaxCompute, DataWorks, and Machine Learning Platform for AI (PAI) for your Alibaba Cloud account.

Prerequisites

- An Alibaba Cloud account is created.
- Real-name verification for individuals or enterprises is completed.

Context

The following Alibaba Cloud services are used in this workshop:

- MaxCompute
- DataWorks
- PAI

Activate MaxCompute

> **Note** If you have already activated MaxCompute, skip this step and directly create a workspace in DataWorks.

1. Go to the Alibaba Cloud official website, click **Log In** in the upper-right corner, and then enter your account name and password.
2. Move the pointer over **Products** in the top navigation bar and choose **Analytics** > **Data Computing** > **MaxCompute** to go to the product page of MaxCompute.
3. Click **Activate Now**.
4. On the buy page of MaxCompute, select a region, read and agree to the service agreement, and then click **Confirm Order and Pay**.

   > **Note**
   >
   > - By default, DataWorks Basic Edition and the standard pay-as-you-go resource package of MaxCompute are provided on the buy page.
   > - The project management, query, and editing features of MaxCompute are integrated into the features of DataWorks. Therefore, you must activate DataWorks at the same time. DataWorks Basic Edition is free of charge. You are charged only if you use Data Integration or run scheduled nodes.
   > - When you activate MaxCompute, you must consider other Alibaba Cloud services that are available in each region. For example, you must consider the region where your Elastic Compute Service (ECS) instance resides and the region where the data resides.

Create a DataWorks workspace

1. Log on to the DataWorks console by using your Alibaba Cloud account.
2. On the **Overview** page, click **create Workspace** in the Frequently Used Workspaces section on the right.

   You can also click **Workspaces** in the left-side navigation pane and click **Create Workspace** on the page that appears.
3. In the **Create Workspace** panel, set the parameters in the Basic Settings step and click **Next**.

Section	Parameter	Description
	Workspace Name	The name of the workspace. The name must be 3 to 23 characters in length and can contain letters, underscores (_), and digits. The name must start with a letter.

Section	Parameter	Description
Basic Information	**Display Name**	The display name of the workspace. The display name can be a maximum of 23 characters in length. It can contain letters, underscores (_), and digits and must start with a letter.
	Mode	The mode of the workspace. Valid values: **Basic Mode (Production Environment Only)** and **Standard Mode (Development and Production Environments)**. ○ **Basic Mode (Production Environment Only)**: A workspace in basic mode is associated with only one MaxCompute project. Workspaces in basic mode do not isolate the development environment from the production environment. In these workspaces, you can perform only basic data development and cannot strictly control the data development process and the permissions on tables. ○ **Standard Mode (Development and Production Environments)**: A workspace in standard mode is associated with two MaxCompute projects. One serves as the development environment, and the other serves as the production environment. Workspaces in standard mode allow you to develop code in a standard way and strictly control the permissions on tables. These workspaces impose limits on table operations in the production environment for data security. For more information, see Basic mode and standard mode.
	Description	The description of the workspace.
Advanced Settings	**Download SELECT Query Result**	Specifies whether the query results that are returned by SELECT statements in DataStudio can be downloaded. If you turn off this switch, the query results cannot be downloaded. You can change the setting of this parameter for the workspace in the Workspace Settings panel after the workspace is created. For more information, see Configure security settings.

4. In the **Select Engines and Services** step, select required compute engines and services and click **Next**.

   DataWorks is now available as a commercial service. If you have not activated DataWorks in a region, activate it before you create a workspace in the region. By default, the following services are selected when you create a workspace: **Data Integration**, **Data Analytics**, **Operation Center**, and **Data Quality**.

> **Note** In this workshop, you must select PAI Studio and MaxCompute.

5. In the **Engine Details** step, set the parameters for the selected compute engines.

Engine	Parameter	Description
MaxCompute	**Instance Display Name**	The display name of the compute engine instance. The display name must start with a letter and can contain only letters, underscores (_), and digits.
	Resource Group	The quotas of computing resources and disk space for the compute engine instance.
	MaxCompute Data Type Edition	The edition of the MaxCompute data type. This configuration takes effect within 5 minutes. For more information, see Data type editions. If you do not know which edition to select, we recommend that you contact the workspace administrator.
	Whether to encrypt	Specifies whether to encrypt data. Valid values: **No encryption** and **Encryption**.
	MaxCompute Project Name	The name of the MaxCompute project. By default, the MaxCompute project that serves as the production environment is named after the DataWorks workspace. The MaxCompute project that serves as the development environment is named in the format of DataWorks workspace name_dev.
	Account for Accessing MaxCompute	The identity that you can use to access the MaxCompute project. For the development environment, the value is fixed to **Node Owner**. For the production environment, the valid values are **Alibaba Cloud Account** and **RAM User**.

6. Click **Create Workspace**.

   After the workspace is created, you can view information about the workspace on the **Workspaces** page.

3.3. Prepare data

You must synchronize raw data to MaxCompute during data preparation.

Prepare the data source

1. Create an ApsaraDB RDS for MySQL instance in the ApsaraDB RDS console and record the instance ID. For more information, see Create an ApsaraDB RDS for MySQL instance.
2. Configure a whitelist for the ApsaraDB RDS for MySQL instance in the ApsaraDB RDS console. For more information, see Configure a whitelist.

   > ? **Note** If you use a custom resource group to run the synchronization node for the ApsaraDB RDS for MySQL instance, you must add the IP addresses of the servers in the custom resource group to the whitelist of the ApsaraDB RDS for MySQL instance.

3. Download the raw data required in this tutorial: indicators_data, steal_flag_data, and trend_data.
4. Upload the raw data to the ApsaraDB RDS for MySQL instance. For more information, see Import data from Excel to ApsaraDB RDS for MySQL.

Create a data source

> ? **Note** In this example, you must create an ApsaraDB RDS for MySQL data source.

1. Go to the **Data Source** page.
   i. Log on to the DataWorks console.
   ii. In the left-side navigation pane, click **Workspaces**.
   iii. After you select the region where the required workspace resides, find the workspace and click **Data Integration**.
   iv. In the left-side navigation pane, choose **Data Source > Data Sources**.
2. On the **Data Source** page, click **Add data source** in the upper-right corner.
3. In the **Add data source** dialog box, click **MySQL**.
4. In the **Add MySQL data source** dialog box, set the parameters based on your business requirements.

Parameter	Description
Data source type	The type of the data source. Set the parameter to **Alibaba Cloud instance mode**.
Data Source Name	The name of the data source. The name can contain letters, digits, and underscores (_) and must start with a letter.
Data source description	The description of the data source. The description can be a maximum of 80 characters in length.
Environment	The environment in which the data source is used. Valid values: **Development** and **Production**. ? **Note** This parameter is displayed only when the workspace is in standard mode.

Parameter	Description
Region	The region of the ApsaraDB RDS for MySQL instance.
RDS instance ID	The ID of your ApsaraDB RDS for MySQL instance. You can view the ID in the ApsaraDB RDS console.
RDS instance account ID	The ID of the Alibaba Cloud account that is used to purchase the ApsaraDB RDS for MySQL instance. You can view the ID on the security settings page in the ApsaraDB RDS console.
Database name	The name of the ApsaraDB RDS for MySQL database.
User name	The username that is used to connect to the ApsaraDB RDS for MySQL database.
Password	The password that is used to connect to the ApsaraDB RDS for MySQL database.

5. Click **Test connectivity**.
6.
7. After the data source passes the connectivity test, click **Complete**.

Create a workflow

1. Click the ≡ icon in the upper-left corner and choose **All Products** > **Data Development** > **DataStudio**.
2. Right-click **Business Flow** and select **Create Workflow**.
3. In the **Create Workflow** dialog box, set the **Workflow Name** and **Description** parameters.

   DataStudio

   > ? **Note** The workflow name can be a maximum of 128 characters in length and can contain letters, digits, underscores (_), and periods (.).

4. Click **Create**.
5. On the workflow configuration tab that appears, drag Zero-Load Node to the canvas, name the zero-load node start, and then click **Commit**. Create three batch synchronization nodes in the same way for synchronizing power consumption trend data, electricity-stealing flag data, and metric data.
6. Draw lines between nodes and set the start node as the ancestor node of the three batch synchronization nodes.

Configure the start node

1. Double-click the start node. In the right-side navigation pane, click the **Properties** tab.
2. Set the root node of the workspace as the ancestor node of the start node.

   In the latest version of DataWorks, each node must have its ancestor and descendant nodes. Therefore, you must set an ancestor node for the start node. In this example, the root node of the workspace is set as the ancestor node of the start node. The root node of the workspace is named in the Workspace name_root format.

3. After the configuration is complete, click the 💾 icon in the upper-left corner.

Create tables

1. Click the created workflow. Then, click **MaxCompute**.
2. Right-click **Table** in the **MaxCompute** folder and select **Create Table**.
3. In the **Create Table** dialog box, set the **Table Name** parameter and click **Create**.

   Create three tables named trend_data, indicators_data, and steal_flag_data. The trend_data table is used to store power consumption trend data, the indicators_data table is used to store metric data, and the steal_flag_data table is used to store electricity-stealing flag data.

   > **Note** The table name can be a maximum of 64 characters in length. It cannot contain special characters and must start with a letter.

4. On the configuration tab of each table, click **DDL Statement** and enter the following CREATE TABLE statements:

```
-- Create a table to store power consumption trend data.
CREATE TABLE trend_data (
    uid bigint,
    trend bigint
)
PARTITIONED BY (dt string);
```

```
-- Create a table to store metric data.
CREATE TABLE indicators_data (
    uid bigint,
    xiansun bigint,
    warnindicator bigint
)
COMMENT '*'
PARTITIONED BY (ds string)
LIFECYCLE 36000;
```

```
-- Create a table to store electricity-stealing flag data.
CREATE TABLE steal_flag_data (
    uid bigint,
    flag bigint
)
COMMENT '*'
PARTITIONED BY (ds string)
LIFECYCLE 36000;
```

5. After you enter the CREATE TABLE statements, click **Generate Table Schema**. Then, click **OK**.
6. On the configuration tab of each table, enter the display name in the **General** section.
7. After the configuration is complete, click **Commit in Development Environment** and **Commit to Production Environment** in sequence.

Configure the batch synchronization nodes

1. Configure the node to synchronize power consumption trend data.
   i. Double-click the node to go to the node configuration tab.
   ii. Configure a source.

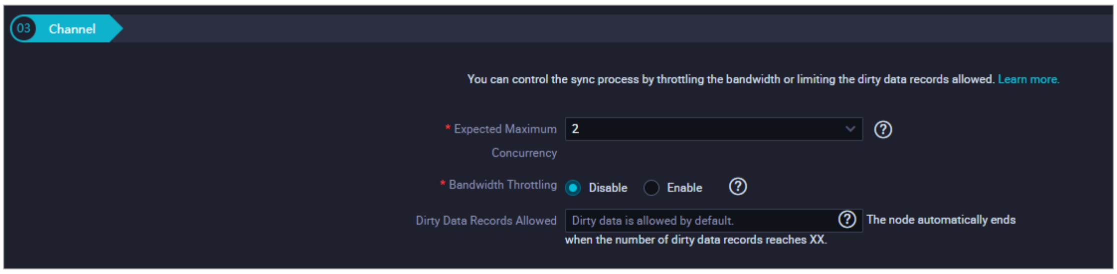
Parameter	Description
Connection	Select **MySQL** and **workshop** in sequence.
Table	Select the trending table from which data is to be synchronized.
Filter	The condition used to filter the data that you want to synchronize. Filtering based on the LIMIT keyword is not supported. The SQL syntax is determined based on the selected data source. This parameter is optional.
Shard Key	If you specify this parameter, data sharding is performed based on the value of this parameter, and parallel threads can be used to read data. This improves data synchronization efficiency. This parameter is optional.

   iii. Configure a destination.

Parameter	Description
Connection	Select **ODPS** and **odps_first** in sequence.
Table	Select the trend_data table to store the source data.
Partition Key Column	Enter the partition key column to be synchronized. Default value: `dt=${bdp.system.bizdate}` .
Writing Rule	Select **Write with Original Data Deleted (Insert Overwrite)**.
Convert Empty Strings to Null	Select **No**.

   iv. Configure the mappings between fields in the source and destination.

v. Set parameters in the **Channel** section.



You can control the sync process by throttling the bandwidth or limiting the dirty data records allowed. Learn more.

* Expected Maximum Concurrency: 2

* Bandwidth Throttling: ● Disable ○ Enable

Dirty Data Records Allowed: Dirty data is allowed by default. The node automatically ends when the number of dirty data records reaches XX.

Parameter	Description
Expected Maximum Concurrency	The maximum number of parallel threads that the synchronization node uses to read data from the source and write data to the destination. You can configure the parallelism for the synchronization node on the codeless UI.
Bandwidth Throttling	Specifies whether to enable bandwidth throttling. You can enable bandwidth throttling and set a maximum transmission rate to avoid heavy read workloads on the source. We recommend that you enable bandwidth throttling and set the maximum transmission rate to an appropriate value based on the configurations of the source.
Dirty Data Records Allowed	The maximum number of dirty data records allowed.

vi. Verify that the preceding configurations are correct and click the 💾 icon in the upper-left corner.

Commit a workflow

1. Go to the workflow configuration tab and click the icon in the upper-left corner.
2. In the **Commit** dialog box, select the nodes to be committed, enter your comments in the **Change description** field, and then select **Ignore I/O Inconsistency Alerts**.
3. Click **Commit**. The **Committed successfully** message appears.

Verify data synchronization to MaxCompute

1. In the left-side navigation pane, click **Ad-Hoc Query**. The **Ad-Hoc Query** tab appears.
2. On the Ad-Hoc Query tab, right-click **Ad-Hoc Query** and choose **Create Node > ODPS SQL**.
3. Write and execute SQL statements to query the number of data records synchronized to the trend_data, indicators_data, and steal_flag_data tables.

   Use the following SQL statements. In each statement, change the partition key value to the data timestamp. For example, if the node is run on August 9, 2019, the data timestamp is 20190808, which is one day before the node is run.

```
-- Check whether the data is written to MaxCompute.
SELECT count(*) from trend_data where dt=Data timestamp of the ad-hoc query node;
SELECT count(*) from indicators_data where ds=Data timestamp of the ad-hoc query node;
SELECT count(*) from steal_flag_data where ds=Data timestamp of the ad-hoc query node;
```

What's next

You understand how to collect and synchronize data. You can now proceed with the next tutorial. The next tutorial describes how to compute and analyze collected data.

3.4. Process data

This topic describes how to process data that is collected to MaxCompute and obtain cleansed data in DataWorks.

Prerequisites

Data is prepared. For more information, see Prepare data.

Create tables

1. Go to the **DataStudio** page.
   1. Log on to the DataWorks console.
   2. In the left-side navigation pane, click **Workspaces**.
   3. In the top navigation bar, select the region in which the workspace that you want to manage resides. Find the workspace and click **DataStudio** in the Actions column.
2. On the **Data Analytics** tab, click the ⌄ icon to the left of the required workflow to show the content of the workflow.
3. Right-click **MaxCompute** and choose **Create** > **Table**.
4. In the **Create Table** dialog box, set the **Table Name** parameter and click **Create**.

   > **Notice** The table name must be 1 to 64 characters in length. It must start with a letter and cannot contain special characters.

   In this example, you must create the following tables:

   - Create three tables to store the power consumption trend data, metrics data, and electricity-stealing flag data that are synchronized to MaxCompute and cleansed in DataWorks. Name the tables clean_trend_data, clean_indicators_data, and clean_steal_flag_data.
   - Create a table named data4ml to store the aggregated data.
5. On the configuration tab of each table, click **DDL Statement**. Enter the following CREATE TABLE statements:

```
-- Create a table for storing the cleansed power consumption trend data.
CREATE TABLE clean_trend_data (
    uid bigint,
    trend bigint
)
PARTITIONED BY (dt string)
LIFECYCLE 7;
```

```
-- Create a table for storing the cleansed metrics data.
CREATE TABLE clean_indicators_data (
    uid bigint,
    xiansun bigint,
    warnindicator bigint
)
COMMENT '*'
PARTITIONED BY (ds string)
LIFECYCLE 36000;
```

```
-- Create a table for storing the cleansed electricity-stealing flag data.
CREATE TABLE clean_steal_flag_data (
    uid bigint,
    flag bigint
)
COMMENT '*'
PARTITIONED BY (ds string)
LIFECYCLE 36000;
```

```
-- Create a table for storing the aggregated data.
CREATE TABLE data4ml (
    uid bigint,
    trend bigint,
    xiansun bigint,
    warnindicator bigint,
    flag bigint
)
COMMENT '*'
PARTITIONED BY (ds string)
LIFECYCLE 36000;
```

6. After you enter the CREATE TABLE statements, click **Generate Table Schema**. Then, click **OK**.
7. On the configuration tab of each table, enter the display name in the **General** section.
8. After the configuration is completed, click **Commit in Development Environment** and **Commit to Production Environment** in sequence.

Design the workflow

For information about how to create a workflow and configure the dependencies among nodes in the workflow, see Create a workflow.

On the workflow configuration tab, create two ODPS SQL nodes for data cleansing and data aggregation and configure the dependencies between nodes.

Configure ODPS SQL nodes

- Configure the data cleansing node.
    i. Double-click the data cleansing node to go to the node configuration tab.
    ii. Write the processing logic.

        Write the following SQL statements:

```
INSERT OVERWRITE TABLE clean_trend_data PARTITION(dt=${bdp.system.bizdate})
SELECT  uid
        ,trend
FROM    trend_data
WHERE   trend IS NOT NULL
AND     uid != 0
AND     dt = ${bdp.system.bizdate}
;
INSERT OVERWRITE TABLE clean_steal_flag_data PARTITION(ds=${bdp.system.bizdate})
SELECT  uid
        ,flag
FROM    steal_flag_data
WHERE   uid != 0
AND     ds = ${bdp.system.bizdate}
;
INSERT OVERWRITE TABLE clean_indicators_data PARTITION(ds=${bdp.system.bizdate})
SELECT  uid
        ,xiansun,warnindicator
FROM    indicators_data
WHERE   uid != 0
AND     ds = ${bdp.system.bizdate}
;
```

iii. Click the 💾 icon in the top toolbar.

- Configure the data aggregation node.

  i. Double-click the data aggregation node to go to the node configuration tab.

  ii. Write the processing logic.

  Write the following SQL statements:

```
INSERT OVERWRITE TABLE data4ml PARTITION (ds=${bdp.system.bizdate})
SELECT  a.uid
        ,trend
        ,xiansun
        ,warnindicator
        ,flag
FROM
(
    SELECT uid,trend FROM clean_trend_data where dt=${bdp.system.bizdate}
)a
FULL OUTER JOIN
(
    SELECT uid,xiansun,warnindicator FROM  clean_indicators_data where ds=${bdp.syste
m.bizdate}
)b
ON      a.uid = b.uid
FULL OUTER JOIN
(
    SELECT uid,flag FROM  clean_steal_flag_data where ds=${bdp.system.bizdate}
)c
ON      b.uid = c.uid
;
```

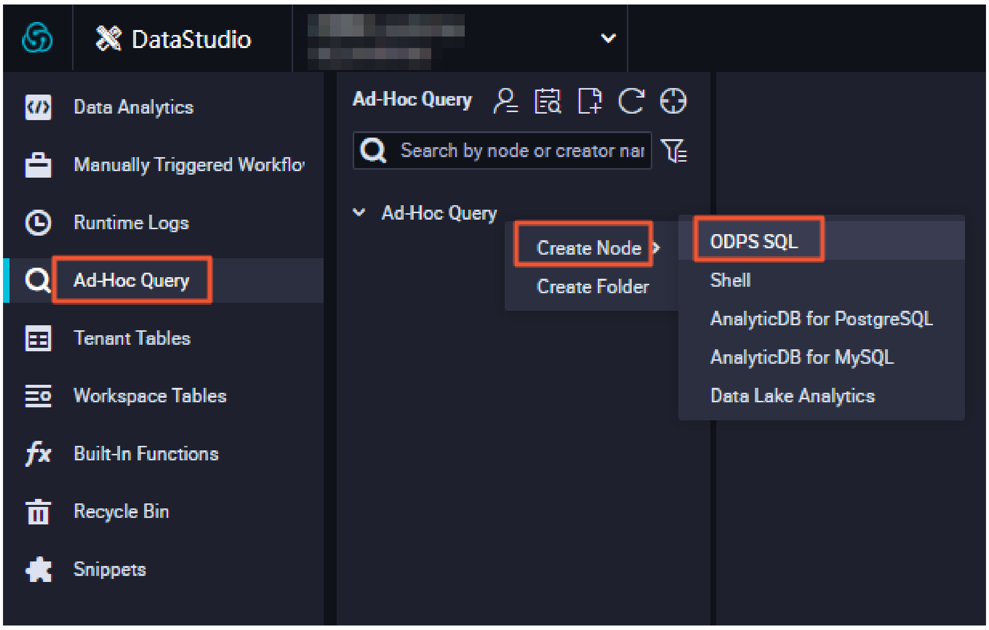
iii. Click the 💾 icon in the top toolbar.

Commit the workflow

1. On the workflow configuration tab, click the ⇧ icon in the top toolbar.
2. In the **Commit** dialog box, select the nodes to be committed, set the **Change description** parameter, and then select **Ignore I/O Inconsistency Alerts**.
3. Click **Commit**. The **Committed successfully** message appears.

Run the workflow

1. On the workflow configuration tab, click the ▶ icon in the top toolbar.
2. On the left-side navigation submenu, click the **Ad-Hoc Query** icon.
3. On the **Ad-Hoc Query** tab, right-click **Ad-Hoc Query** and choose **Create Node > ODPS SQL**.



4. Write and execute SQL statements to query the number of data records that are written to the clean_trend_data, clean_indicators_data, clean_steal_flag_data, and data4ml tables.

   Use the following SQL statements. In each statement, change the partition key value to the data timestamp. For example, if the node is run on August 9, 2019, the data timestamp is 20190808.

```
-- Check whether the data is written to MaxCompute.
SELECT count(*) from clean_trend_data where dt=Data timestamp;
SELECT count(*) from clean_indicators_data where ds=Data timestamp;
SELECT count(*) from clean_steal_flag_data where ds=Data timestamp;
SELECT count(*) from data4ml where ds=Data timestamp;
```

Deploy the workflow

After you commit the workflow, the nodes in the workflow are in the development environment. You must deploy the configured nodes to the production environment because nodes in the development environment cannot be automatically scheduled.

> **Note** Before you deploy the nodes to the production environment, test the node code to ensure that the code is correct.

1. On the workflow configuration tab, click the icon in the top toolbar.
2. On the **Create Deploy Task** page, select the nodes to be deployed and click **Add to List**.
3. Click **To-Be-Deployed Node List** in the upper-right corner and click **Deploy All**.
4. Go to the **Deploy Tasks** page and view the deployed nodes.

Run the nodes in the production environment

1. After the nodes are deployed, click **Operation Center** in the upper-right corner.
2. Choose **Cycle Task Maintenance > Cycle Task**. Select the required nodes.
3. In the directed acyclic graph (DAG), right-click the start node and choose **Run > Current and Descendent Nodes Retroactively**.
4. Select nodes to generate retroactive data and set the **Data Timestamp** parameter.
5. Click **OK**.
6. On the **Patch Data** page, click **Refresh** until all the SQL nodes are run.

What's next

You have learned how to create SQL nodes and process raw data. You can now proceed with the next step to learn how to load processed data and build a model for identifying users who steal electricity or are involved in electricity leakage by using Machine Learning Platform for AI (PAI).

3.5. Build a data model

This topic describes how to load processed data from DataWorks to Machine Learning Platform for AI (PAI) and build a model for identifying users who steal electricity or are involved in electricity leakage.

Prerequisites

Data is processed. For more information, see Process data.

Create a PAI experiment

1. Log on to the PAI console. In the left-side navigation pane, choose Model Training > **Studio-Modeling Visualization**.
2. On the page that appears, find the target workspace and click **Machine Learning** in the Operation column.
3. On the left-side navigation submenu, click **Experiments**. In the left-side navigation pane, right-click **My Experiments** and select **New Experiment**.
4. In the **New Experiment** dialog box that appears, set **Name** and **Description**.
5. Click **OK**.

Load datasets

1. On the left-side navigation submenu, click **Data Source**.
2. Enter data4ml in the search box and click the search icon to search for the final output table of the

target workflow. For more information, see Process data.

3. Drag the data4ml table in the **Table Search Result** section to the canvas on the right.

   On the canvas, right-click the data4ml node and select **View Data**. In the dialog box that appears, view the loaded data. The data includes electricity theft and leakage metrics, such as the power consumption trend, the line loss, and the number of alerts. The data also includes the electricity-stealing flag data that indicates whether users steal electricity or are involved in electricity leakage.

Explore the data

1. Analyze the correlation between data.
   i. On the left-side navigation submenu, click **Components**. In the left-side navigation pane, drag the **Correlation Coefficient Matrix** component under **Statistical Analysis** to the canvas on the right.
   ii. On the canvas, draw a line from the source MaxCompute table output port of the **data4ml** node to the input port of the **Correlation Coefficient Matrix** node.
   iii. Right-click the **Correlation Coefficient Matrix** node and select **Run from Here**.
   iv. After the **Correlation Coefficient Matrix** node is run, right-click the node and select **View Analytics Report**.

      As shown in the correlation coefficient matrix, the three electricity theft and leakage metrics are not enough to identify users who steal electricity or are involved in electricity leakage. To identify such users, you must analyze sufficient features.
2. Analyze features.
   i. On the left-side navigation submenu, click **Components**. In the left-side navigation pane, drag the **Data View** component under **Statistical Analysis** to the canvas on the right.
   ii. On the canvas, draw a line from the source MaxCompute table output port of the **data4ml** node to the input port of the **Data View** node.
   iii. Double-click the **Data View** node. In the right-side pane, click the **Fields Setting** tab. Click **Select Column** for **Feature Columns**. Select the flag field for **Target Column**.
   iv. In the **Select Column** dialog box that appears, select the trend, xiansun, and warnindicator fields as the feature columns and click **OK**.
   v. Right-click the **Data View** node and select **Run from Here**.
   vi. After the Data View node is run, right-click the node and select **View Analytics Report** to view the relationship between each feature column and the flag column in terms of data distribution.

Perform data modeling

After you explore and analyze the data, you can select appropriate algorithm models for data modeling.

1. Use the **Split** component to divide data into training datasets and test datasets.
   i. On the left-side navigation submenu, click **Components**. In the left-side navigation pane, drag the **Split** component under **Data Preprocessing** to the canvas on the right.
   ii. On the canvas, draw a line from the source MaxCompute table output port of the **data4ml** node to the input port of the **Split** node.
   iii. Right-click the **Split** node and select **Run from Here**.

    iv. After the **Split** node is run, right-click the node and choose **View Data > View Output Port**.

2. Use the **Logistic Regression for Binary Classification** component to perform regression modeling on data.

    i. On the left-side navigation submenu, click **Components**. In the left-side navigation pane, choose **Machine Learning > Binary Classification** and drag the **Logistic Regression for Binary Classification** component to the canvas on the right.

    ii. On the canvas, draw a line from the output table 1 of the **Split** node to the training table of the **Logistic Regression for Binary Classification** node.

    iii. Double-click the **Logistic Regression for Binary Classification** node. In the right-side pane, click the **Fields Setting** tab. Click **Select Column** for **Training Feature Columns**. Select the flag field for **Target Columns**.

    iv. In the **Select Column** dialog box that appears, select the trend, xiansun, and warnindicator fields as the training feature columns and click **OK**.

    v. Right-click the **Logistic Regression for Binary Classification** node and select **Run from Here**.

    vi. After the Logistic Regression for Binary Classification node is run, right-click the node and choose **Model Option > Show Model** to view the data model.

Predict and evaluate the regression model

1. Use the **Prediction** component to predict the result of applying the model to test datasets.

    i. On the left-side navigation submenu, click **Components**. In the left-side navigation pane, drag the **Prediction** component under **Machine Learning** to the canvas on the right.

    ii. On the canvas, draw a line from the logistic regression model of the **Logistic Regression for Binary Classification** node to the model result input port of the **Prediction** node. Draw a line from the output table 2 of the **Split** node to the prediction data input port of the **Prediction** node.

    iii. Double-click the **Prediction** node. In the right-side pane, set fields on the **Fields Setting** tab.

    Click **Select Column** separately for **Feature Columns** and **Reserved Output Column**.

    iv. In the **Select Column** dialog box that appears, select all the five fields and click **OK**.

    v. Right-click the **Prediction** node and select **Run from Here**.

    vi. After the Prediction node is run, right-click the node and select **View Data**.

2. Use the **Binary Classification Evaluation** component to obtain the modeling result.

    i. On the left-side navigation submenu, click **Components**. In the left-side navigation pane, choose **Machine Learning > Evaluation** and drag the **Binary Classification Evaluation** component to the canvas on the right.

    ii. On the canvas, draw a line from the prediction result output port of the **Prediction** node to the input port of the **Binary Classification Evaluation** node.

    iii. Double-click the **Binary Classification Evaluation** node. In the right-side **Fields Setting** pane, select the flag field for **Original Label Column**.

    iv. Right-click the **Binary Classification Evaluation** node and select **Run from Here**.

    v. After the Binary Classification Evaluation node is run, right-click the node and select **View Evaluation Report** to view the modeling effect.

What's next

Now, you have learned how to use PAI to identify users who steal electricity or are involved in electricity leakage. You can also use Elastic Algorithm Service to deploy an online service for identifying electricity theft and leakage.

4.Integrate and use CDH

Cloudera's Distribution including Apache Hadoop (CDH) can be integrated into DataWorks. This allows you to configure your CDH clusters as storage and compute engines in DataWorks. This way, you can use DataWorks features, such as node development, node scheduling, Data Map (metadata management), and Data Quality, to develop and manage data and nodes. This topic describes how to integrate CDH into DataWorks and use CDH in DataWorks.

Prerequisites

- A CDH cluster is deployed on an Elastic Compute Service (ECS) instance.

  The CDH cluster can also be deployed in an environment other than Alibaba Cloud ECS. You must make sure that the environment can be connected to Alibaba Cloud. You can use Express Connect and VPN Gateway to ensure the network connectivity between the environment and Alibaba Cloud.

- DataWorks is activated, and a workspace is created to connect the CDH cluster.

  > ? **Note** The workspaces that are used to connect CDH clusters do not need to be associated with compute engines. Therefore, when you create a workspace, you do not need to select an engine. For more information about how to create a workspace, see Create a workspace.

- An account that has administrative permissions on the workspace is created. Only workspace administrators can associate CDH clusters with a DataWorks workspace. For more information about how to grant administrative permissions on a workspace to an account, see Manage workspace-level roles and members.
- A DataWorks exclusive resource group for scheduling is created. For more information, see Exclusive resource group mode.

Before you use CDH in DataWorks, you must perform the following operations to integrate CDH into DataWorks:

1. Step 1: Obtain the configuration information of the CDH cluster
2. Step 2: Configure network connectivity
3. Step 3: Add the configurations of the CDH cluster to DataWorks

After you complete the preceding operations, you can develop and run CDH nodes in DataWorks and view the status of the nodes in DataWorks Operation Center. For more information, see Use DataWorks to develop nodes and Configure O&M and monitoring settings.

You can also use the Data Quality and Data Map services of DataWorks to manage CDH data and nodes. For more information, see Configure data quality rules and Use Data Map to collect data.

Limits

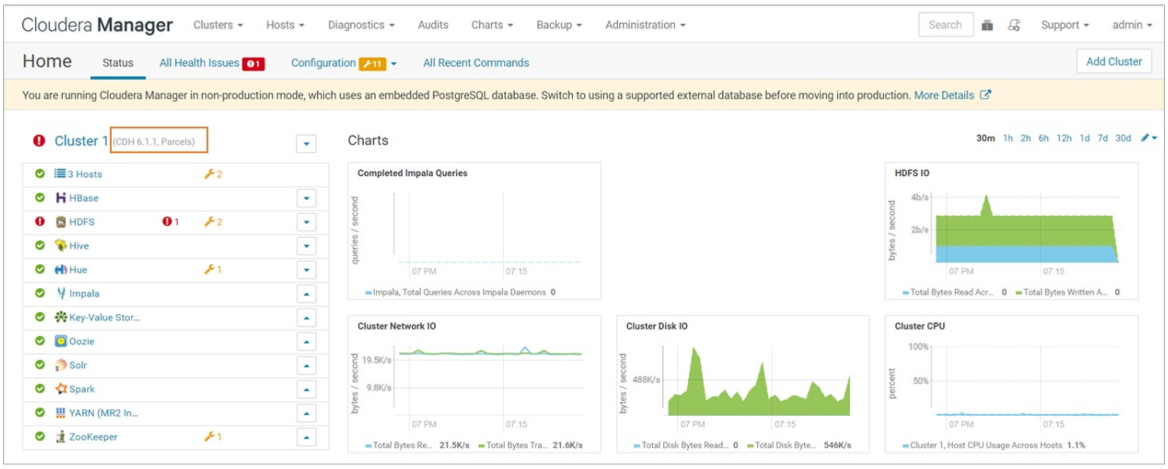
- To use CDH features in DataWorks, you must purchase and use a DataWorks exclusive resource group for scheduling.
- The CDH cluster must be connected to the exclusive resource group for scheduling.
- DataWorks supports CDH 6.1.1, CDH 5.16.2, CDH 6.2.1, and CDH 6.3.2.

Step 1: Obtain the configuration information of the CDH cluster

1. Obtain the version information of the CDH cluster. The version information is required when you add

the configurations of the CDH cluster to DataWorks.

Log on to the Cloudera Manager Admin Console. On the page that appears, you can view the version information on the right side of the cluster name, as shown in the following figure.



You are running Cloudera Manager in non-production mode, which uses an embedded PostgreSQL database. Switch to using a supported external database before moving into production. More Details

2. Obtain the host and component addresses of the CDH cluster. The addresses are required when you add the configurations of the CDH cluster to DataWorks.
   - Method 1: Use the DataWorks JAR package to obtain the addresses.
     a. Log on to the Cloudera Manager Admin Console and download the DataWorks JAR package.

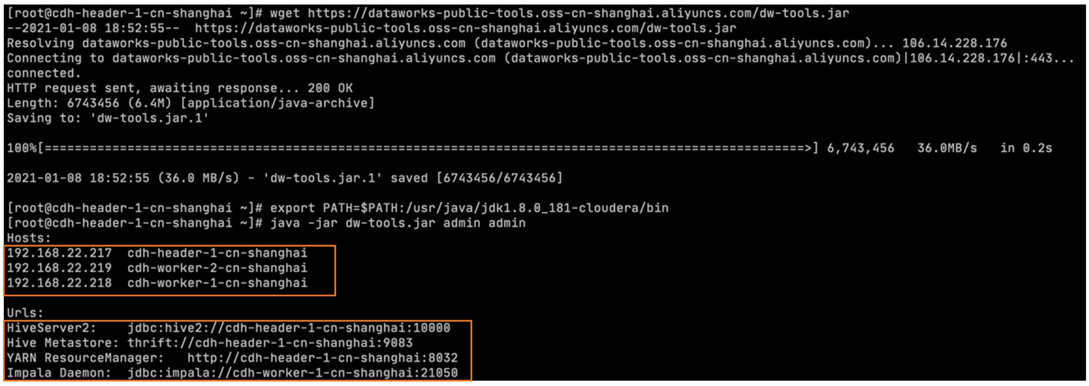
     ```
     wget https://dataworks-public-tools.oss-cn-shanghai.aliyuncs.com/dw-tools.jar
     ```

     b. Run the JAR package.

     ```
     export PATH=$PATH:/usr/java/jdk1.8.0_181-cloudera/bin
     java -jar dw-tools.jar <user> <password>
     ```

     Set `<user>` to the username that you use to log on to the Cloudera Manager Admin Console and `<password>` to the password that you use to log on to the Cloudera Manager Admin Console.

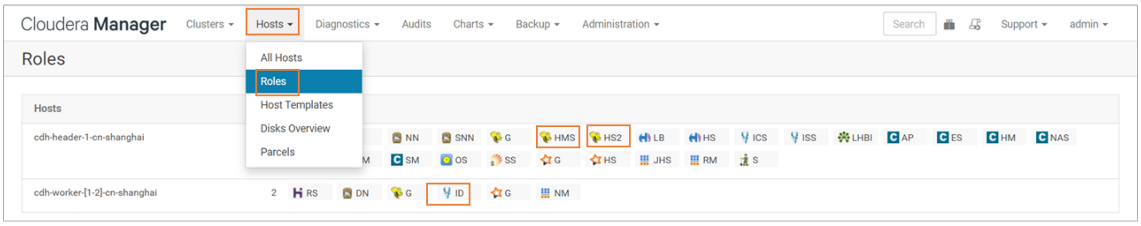
     c. View the host and component addresses of the CDH cluster in the returned results. Then, record the addresses.



```
[root@cdh-header-1-cn-shanghai ~]# wget https://dataworks-public-tools.oss-cn-shanghai.aliyuncs.com/dw-tools.jar
--2021-01-08 18:52:55--  https://dataworks-public-tools.oss-cn-shanghai.aliyuncs.com/dw-tools.jar
Resolving dataworks-public-tools.oss-cn-shanghai.aliyuncs.com (dataworks-public-tools.oss-cn-shanghai.aliyuncs.com)... 106.14.228.176
Connecting to dataworks-public-tools.oss-cn-shanghai.aliyuncs.com (dataworks-public-tools.oss-cn-shanghai.aliyuncs.com)|106.14.228.176|:443...
connected.
HTTP request sent, awaiting response... 200 OK
Length: 6743456 (6.4M) [application/java-archive]
Saving to: 'dw-tools.jar.1'

100%[===================================================================================================>] 6,743,456   36.0MB/s   in 0.2s

2021-01-08 18:52:55 (36.0 MB/s) - 'dw-tools.jar.1' saved [6743456/6743456]

[root@cdh-header-1-cn-shanghai ~]# export PATH=$PATH:/usr/java/jdk1.8.0_181-cloudera/bin
[root@cdh-header-1-cn-shanghai ~]# java -jar dw-tools.jar admin admin
Hosts:
192.168.22.217  cdh-header-1-cn-shanghai
192.168.22.219  cdh-worker-2-cn-shanghai
192.168.22.218  cdh-worker-1-cn-shanghai

Urls:
HiveServer2:    jdbc:hive2://cdh-header-1-cn-shanghai:10000
Hive Metastore: thrift://cdh-header-1-cn-shanghai:9083
YARN ResourceManager:   http://cdh-header-1-cn-shanghai:8032
Impala Daemon:  jdbc:impala://cdh-worker-1-cn-shanghai:21050
```

   - Method 2: Obtain the addresses from the Cloudera Manager Admin Console.

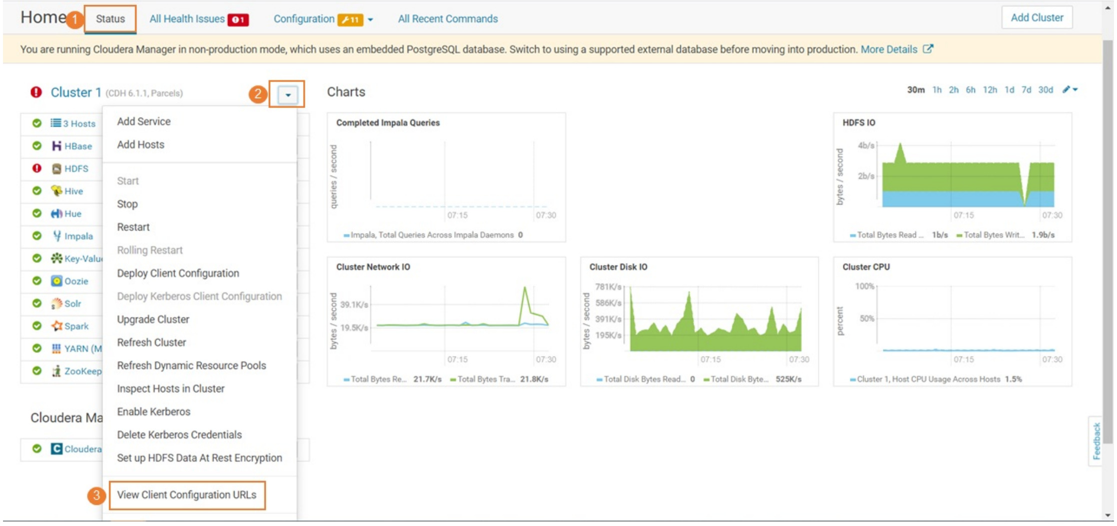
Log on to the Cloudera Manager Admin Console and select Roles from the Hosts drop-down list. Find the components that you want to configure based on keywords and icons. Then, view and record the hostnames displayed on the left, and complete component addresses based on the hostnames and the address format. For more information about the default port numbers in the addresses, see the returned results in Method 1.



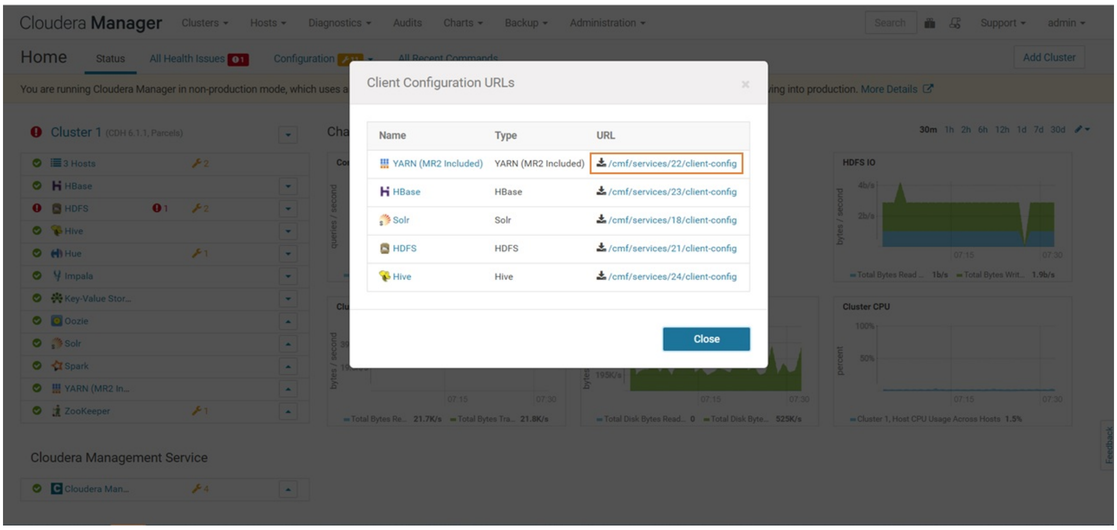
Components:

- HS2: HiveServer2
- HMS: Hive Metastore
- ID: Impala Daemon
- RM: YARN ResourceManager

3. Obtain the configuration files of the CDH cluster. The configuration files must be uploaded when you add the configurations of the CDH cluster to DataWorks.
   i. Log on to the Cloudera Manager Admin Console.
   ii. On the **Status** tab, click the drop-down arrow on the right of the cluster name and select **View Client Configuration URLs**.


Home
1
Status
All Health Issues
Configuration
All Recent Commands
Add Cluster
You are running Cloudera Manager in non-production mode, which uses an embedded PostgreSQL database. Switch to using a supported external database before moving into production. More Details
Cluster 1 (CDH 6.1.1, Parcels)
2
3 Hosts
HBase
HDFS
Hive
Hue
Impala
Key-Valu
Oozie
Solr
Spark
YARN (M
ZooKeep
Cloudera Ma
Cloudera
Add Service
Add Hosts
Start
Stop
Restart
Rolling Restart
Deploy Client Configuration
Deploy Kerberos Client Configuration
Upgrade Cluster
Refresh Cluster
Refresh Dynamic Resource Pools
Inspect Hosts in Cluster
Enable Kerberos
Delete Kerberos Credentials
Set up HDFS Data At Rest Encryption
3
View Client Configuration URLs
Charts
30m 1h 2h 6h 12h 1d 7d 30d
Completed Impala Queries
queries / second
07:15
07:30
Impala, Total Queries Across Impala Daemons 0
HDFS IO
bytes / second
4b/s
2b/s
Total Bytes Read... 1b/s
Total Bytes Writ... 1.9b/s
Cluster Network IO
bytes / second
39.1K/s
19.5K/s
Total Bytes Re... 21.7K/s
Total Bytes Tra... 21.8K/s
Cluster Disk IO
781K/s
586K/s
391K/s
195K/s
Total Disk Bytes Read... 0
Total Disk Byte... 525K/s
Cluster CPU
percent
100%
50%
Cluster 1, Host CPU Usage Across Hosts 1.5%
Feedback

iii. In the Client Configuration URLs dialog box, download the YARN configuration package.



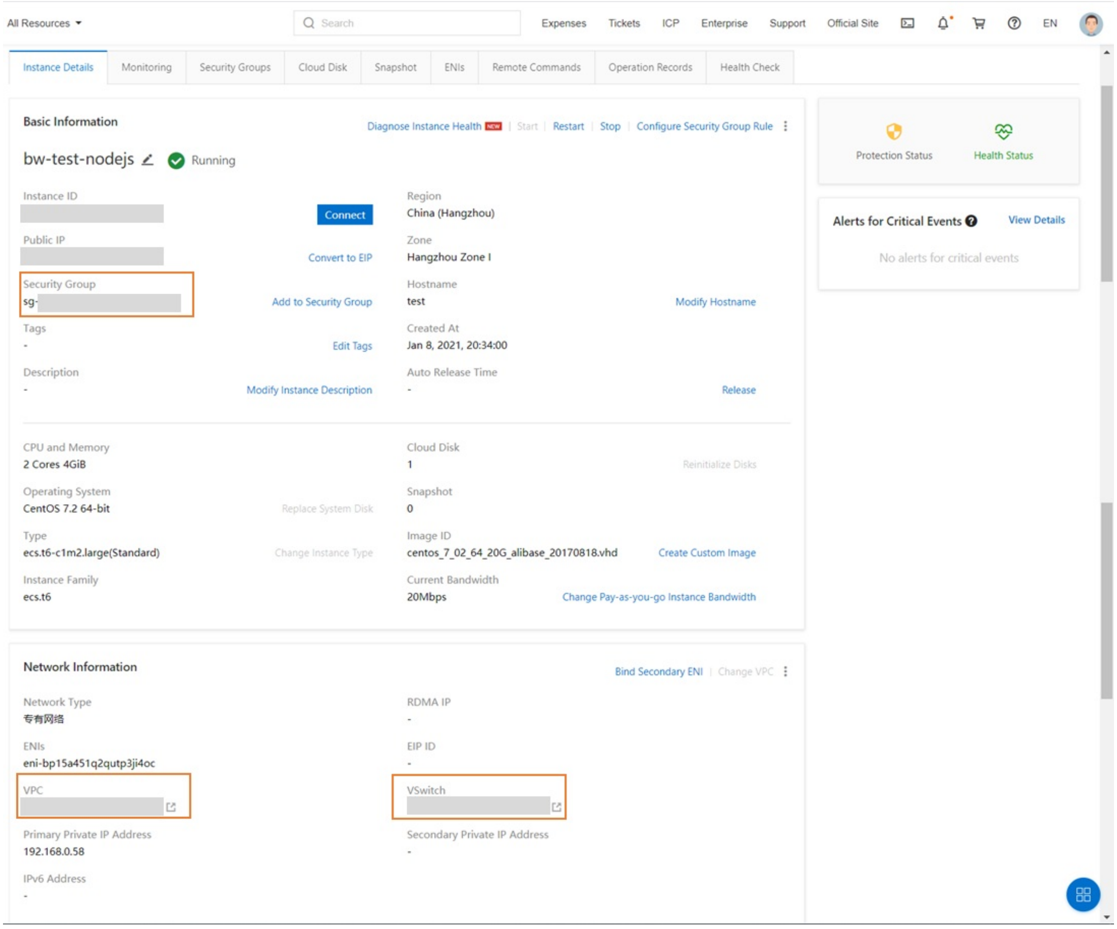
Client Configuration URLs

Name	Type	URL
YARN (MR2 Included)	YARN (MR2 Included)	/cmf/services/22/client-config
HBase	HBase	/cmf/services/23/client-config
Solr	Solr	/cmf/services/18/client-config
HDFS	HDFS	/cmf/services/21/client-config
Hive	Hive	/cmf/services/24/client-config

4. Obtain the network information of the CDH cluster. The network information is used to configure network connectivity between the CDH cluster and DataWorks exclusive resource group for scheduling.

   i. Log on to the ECS console.

   ii. In the left-side navigation pane, choose Instances & Images > Instances. In the top navigation bar, select the region where the ECS instance that hosts the CDH cluster resides. On the Instances page, find the ECS instance and click its ID. On the **Instance Details** tab of the page that appears, view the information about the instance, such as security group, VPC, and vSwitch. Then, record the information.



Basic Information

bw-test-nodejs Running

Instance ID [redacted]

Public IP [redacted]

Security Group sg-[redacted]

Tags -

Description -

Region China (Hangzhou)

Zone Hangzhou Zone I

Hostname test

Created At Jan 8, 2021, 20:34:00

Auto Release Time -

CPU and Memory 2 Cores 4GiB

Operating System CentOS 7.2 64-bit

Type ecs.t6-c1m2.large(Standard)

Instance Family ecs.t6

Cloud Disk 1

Snapshot 0

Image ID centos_7_02_64_20G_alibase_20170818.vhd

Current Bandwidth 20Mbps

Alerts for Critical Events

No alerts for critical events

Network Information

Network Type 专有网络

ENIs eni-bp15a451q2qutp3ji4oc

VPC [redacted]

Primary Private IP Address 192.168.0.58

IPv6 Address -

RDMA IP -

EIP ID -

VSwitch [redacted]

Secondary Private IP Address -

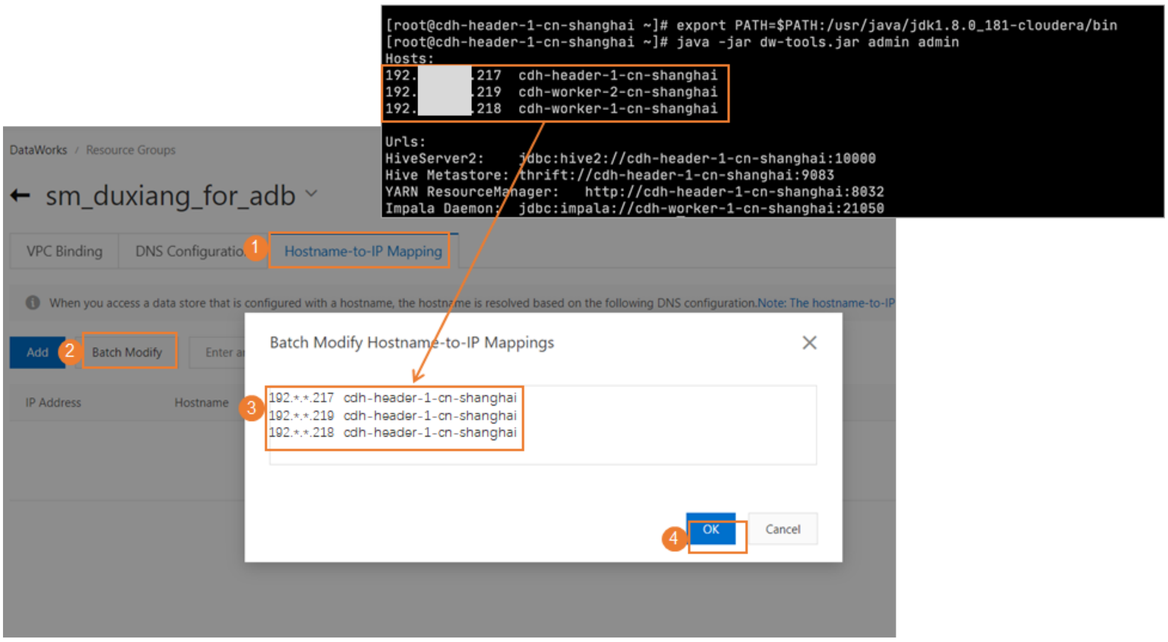
Step 2: Configure network connectivity

By default, DataWorks exclusive resource groups for scheduling are not connected to the networks of resources for other Alibaba Cloud services after the resource groups are created. Therefore, before you use CDH, you must obtain the network information of your CDH cluster. Then, associate your DataWorks exclusive resource group for scheduling with the VPC to which the CDH cluster belongs. This ensures network connectivity between the CDH cluster and DataWorks exclusive resource group for scheduling.

1. Go to the network configuration page of the exclusive resource group for scheduling.
   i. Log on to the DataWorks console.
   ii. In the left-side navigation pane, click **Resource Groups**. The **Exclusive Resource Groups** tab appears.
   iii. Find the desired exclusive resource group for scheduling and click **Network Settings** in the Actions column.
2. Associate the exclusive resource group for scheduling with the VPC to which the CDH cluster belongs.

   On the **VPC Binding** tab, click **Add Binding**. In the Add VPC Binding panel, select the VPC, vSwitch, and security group that are recorded in . Then, click OK.
3. Configure hosts.

   Click the **Hostname-to-IP Mapping** tab. On this tab, click **Batch Modify**. In the Batch Modify Hostname-to-IP Mappings dialog box, enter the host addresses that are recorded in .



```
[root@cdh-header-1-cn-shanghai ~]# export PATH=$PATH:/usr/java/jdk1.8.0_181-cloudera/bin
[root@cdh-header-1-cn-shanghai ~]# java -jar dw-tools.jar admin admin
Hosts:
192.[redacted].217  cdh-header-1-cn-shanghai
192.      .219  cdh-worker-2-cn-shanghai
192.      .218  cdh-worker-1-cn-shanghai

Urls:
HiveServer2:    jdbc:hive2://cdh-header-1-cn-shanghai:10000
Hive Metastore: thrift://cdh-header-1-cn-shanghai:9083
YARN ResourceManager:   http://cdh-header-1-cn-shanghai:8032
Impala Daemon:  jdbc:impala://cdh-worker-1-cn-shanghai:21050
```

Batch Modify Hostname-to-IP Mappings

```
192.*.*.217  cdh-header-1-cn-shanghai
192.*.*.219  cdh-header-1-cn-shanghai
192.*.*.218  cdh-header-1-cn-shanghai
```

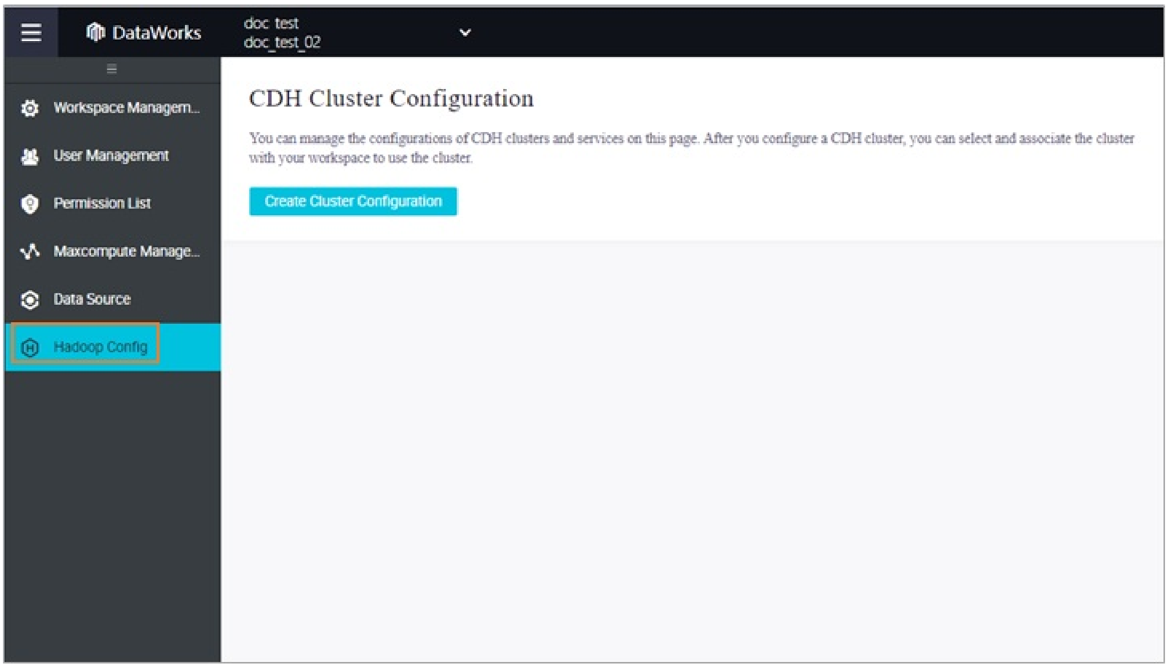
DataWorks / Resource Groups
sm_duxiang_for_adb
VPC Binding
DNS Configuratio
1
Hostname-to-IP Mapping
When you access a data store that is configured with a hostname, the hostname is resolved based on the following DNS configuration.Note: The hostname-to-IP
Add
2
Batch Modify
IP Address
Hostname
3
4
OK
Cancel

Step 3: Add the configurations of the CDH cluster to DataWorks

Only workspace administrators can add the configurations of CDH clusters to DataWorks. Therefore, you must use an account that has administrative permissions on your workspace to perform this operation.

1.

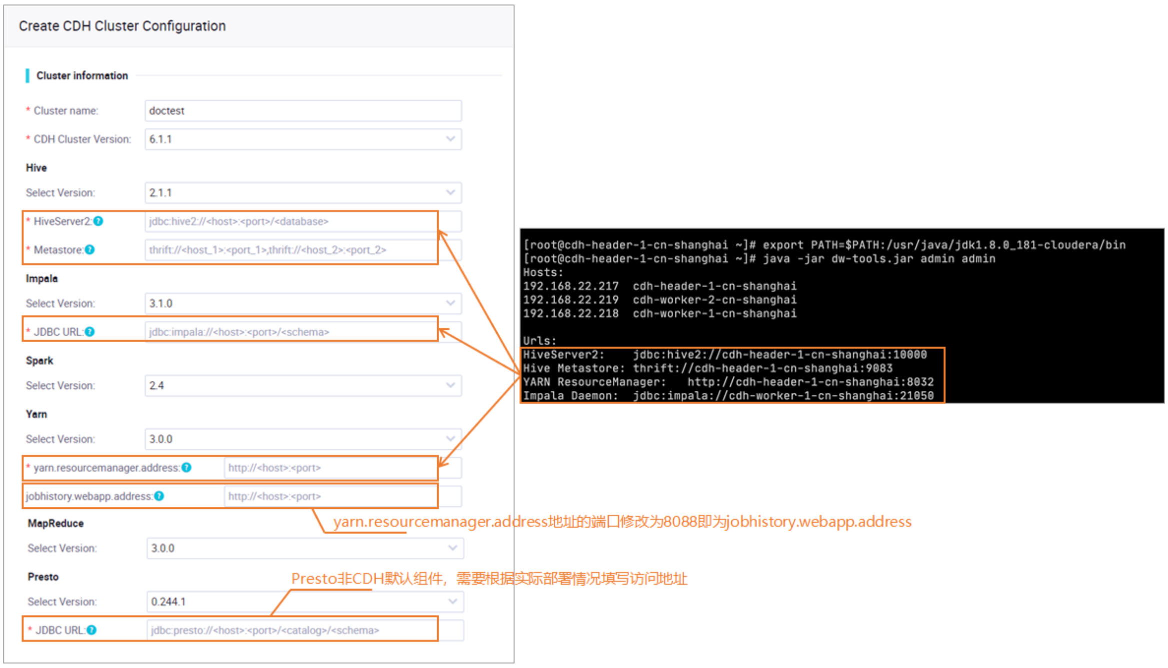
2. In the left-side navigation pane of the page that appears, choose Opensource Cluster Management > **Hadoop Config**.



CDH Cluster Configuration

You can manage the configurations of CDH clusters and services on this page. After you configure a CDH cluster, you can select and associate the cluster with your workspace to use the cluster.

3. On the **CDH Cluster Configuration** page, click **Create Now**. In the **Create CDH Cluster Configuration** dialog box, enter the component addresses that are recorded in Step 2: Configure network connectivity in the related fields.



Create CDH Cluster Configuration

Cluster information

* Cluster name:	doctest
* CDH Cluster Version:	6.1.1
Hive	
Select Version:	2.1.1
* HiveServer2:	jdbc:hive2://<host>:<port>/<database>
* Metastore:	thrift://<host_1>:<port_1>,thrift://<host_2>:<port_2>
Impala	
Select Version:	3.1.0
* JDBC URL:	jdbc:impala://<host>:<port>/<schema>
Spark	
Select Version:	2.4
Yarn	
Select Version:	3.0.0
* yarn.resourcemanager.address:	http://<host>:<port>
jobhistory.webapp.address:	http://<host>:<port>
MapReduce	
Select Version:	3.0.0
Presto	
Select Version:	0.244.1
* JDBC URL:	jdbc:presto://<host>:<port>/<catalog>/<schema>

```
[root@cdh-header-1-cn-shanghai ~]# export PATH=$PATH:/usr/java/jdk1.8.0_181-cloudera/bin
[root@cdh-header-1-cn-shanghai ~]# java -jar dw-tools.jar admin admin
Hosts:
192.168.22.217  cdh-header-1-cn-shanghai
192.168.22.219  cdh-worker-2-cn-shanghai
192.168.22.218  cdh-worker-1-cn-shanghai

Urls:
HiveServer2:    jdbc:hive2://cdh-header-1-cn-shanghai:10000
Hive Metastore: thrift://cdh-header-1-cn-shanghai:9083
YARN ResourceManager:   http://cdh-header-1-cn-shanghai:8032
Impala Daemon:  jdbc:impala://cdh-worker-1-cn-shanghai:21050
```

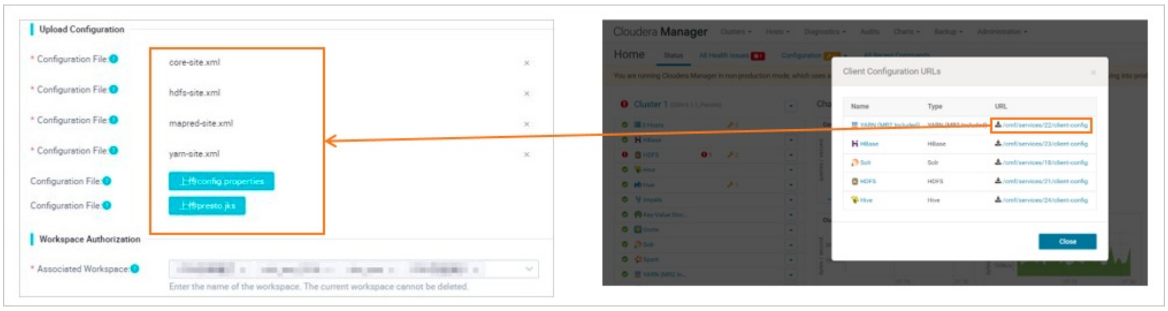
yarn.resourcemanager.address地址的端口修改为8088即为jobhistory.webapp.address

Presto非CDH默认组件，需要根据实际部署情况填写访问地址

Configuration information:

- Cluster name: the name of your CDH cluster. You can customize the name.
- Versions: Select the CDH cluster version and component versions based on actual conditions.
- Addresses: Enter the recorded component addresses. Configuration information:
  - **jobhistory.webapp.address** for YARN: Change the port number in the value of **yarn.resourcemanager.address** to 8088.
  - **JDBC URL** for Presto: Presto is not a default component for CDH. You must configure this parameter based on actual conditions.

4. Upload configuration files and associate the CDH cluster with the workspace.



Upload Configuration

* Configuration File:	core-site.xml
* Configuration File:	hdfs-site.xml
* Configuration File:	mapred-site.xml
* Configuration File:	yarn-site.xml
Configuration File:	上传config.properties
Configuration File:	上传presto.jks

Workspace Authorization

* Associated Workspace:	[redacted]

Enter the name of the workspace. The current workspace cannot be deleted.

Client Configuration URLs

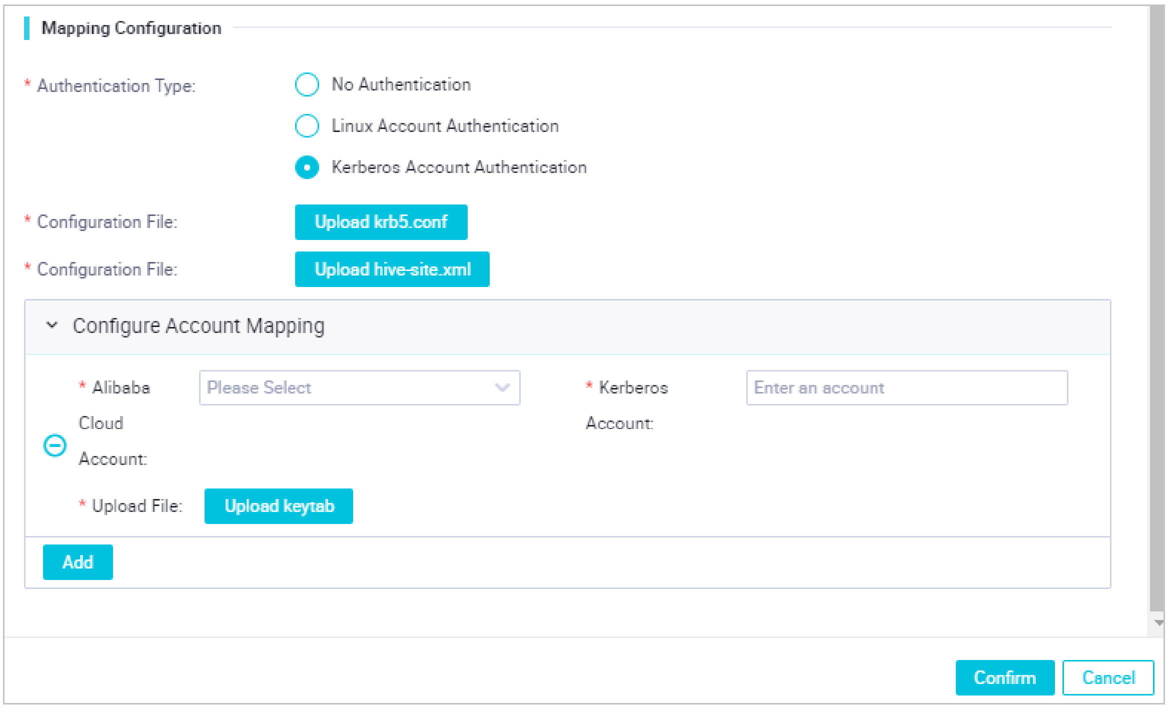
Name	Type	URL
YARN (MR2 Included)	YARN (MR2 Included)	/cmf/services/22/client-config
HBase	HBase	/cmf/services/23/client-config
Solr	Solr	/cmf/services/18/client-config
HDFS	HDFS	/cmf/services/21/client-config
Hive	Hive	/cmf/services/24/client-config

Close

5. Configure mappings between Alibaba Cloud accounts or RAM users and Kerberos accounts.

   If you want to isolate permissions on the data that can be accessed by different Alibaba Cloud accounts or RAM users in a CDH cluster, enable **Kerberos Account Authentication** and configure the mappings between **Alibaba Cloud accounts or RAM users** and **Kerberos accounts**.

> Note **Kerberos Account** specifies an account that you use to access the CDH cluster. You can use the Sentry or Ranger component to configure different permissions for different **Kerberos accounts** in the CDH cluster to isolate data permissions. The Alibaba Cloud accounts or RAM users that are mapped to the same **Kerberos account** have the same permissions on the data in the CDH cluster. Specify a Kerberos account (also referred to as a Kerberos principal) in the format of `Instance name@Domain name`, such as cdn_test@HADOOP.COM.


Mapping Configuration

* Authentication Type:
○ No Authentication
○ Linux Account Authentication
● Kerberos Account Authentication

* Configuration File: Upload krb5.conf

* Configuration File: Upload hive-site.xml

Configure Account Mapping

* Alibaba Cloud Account: Please Select

* Kerberos Account: Enter an account

* Upload File: Upload keytab

6. Click **Confirm**.

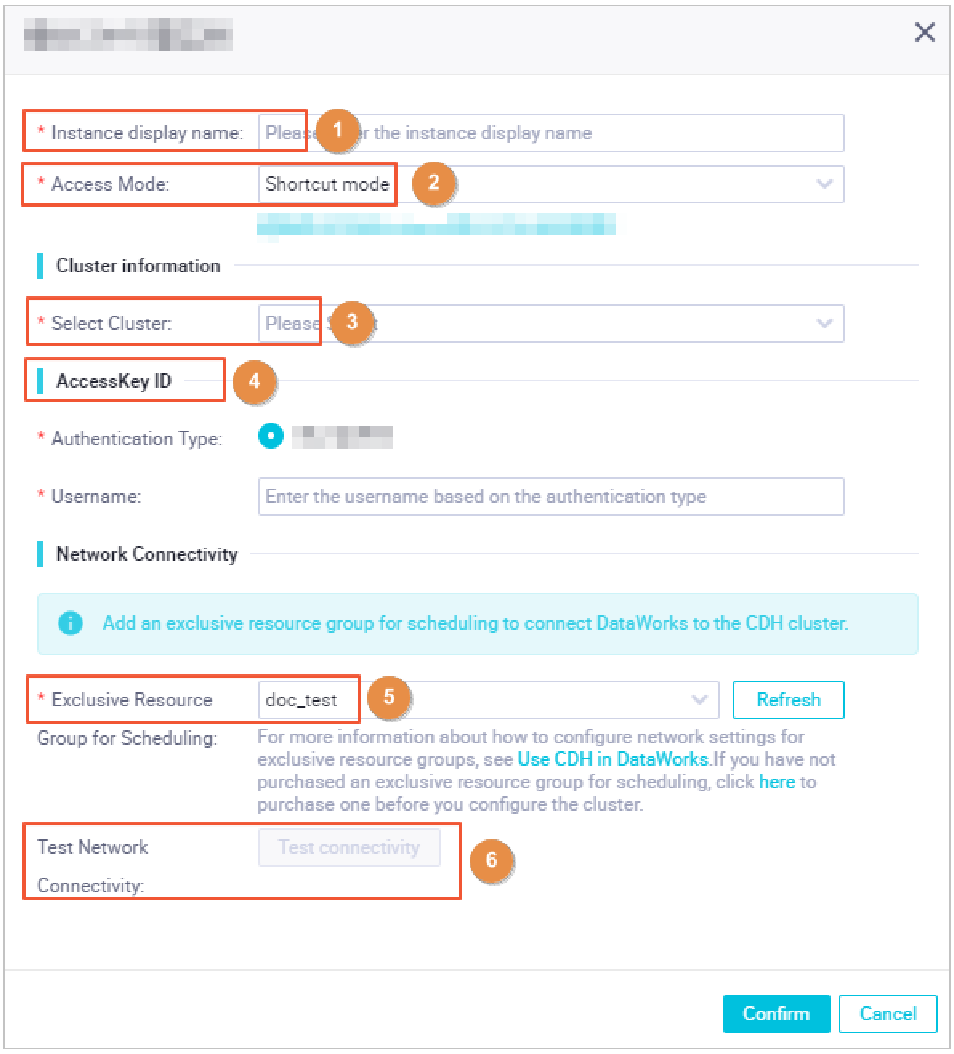
   After the configurations of the CDH cluster are added to DataWorks, you can add the CDH cluster to the associated workspace as a compute engine instance. Then, you can develop and run CDH nodes in the workspace.

Step 4: Add the CDH cluster to the associated workspace as a compute engine instance

1. On the Workspaces page, click **Workspace Settings** in the Actions column that corresponds to the associated workspace.
2. In the lower part of the Workspace Settings panel, click More. In the **Compute Engine Information** section of the Configuration page, click the **CDH** tab. On the CDH tab, click **Add Instance**. In the Add CDH Compute Engine dialog box, configure the parameters.

   You can set Access Mode to **Shortcut mode** or **Security mode**. If **Security mode** is selected, the permissions on the data of the node that is run by different Alibaba Cloud accounts or RAM users can be isolated. The parameters that need to be configured vary based on the value of the Access Mode parameter.

   - The following figure shows the parameters you must configure if you set Access Mode to **Shortcut mode**.



[redacted]

* Instance display name: Please enter the instance display name

* Access Mode: Shortcut mode

[redacted]

Cluster information

* Select Cluster: Please select

AccessKey ID

* Authentication Type: ● [redacted]

* Username: Enter the username based on the authentication type

Network Connectivity

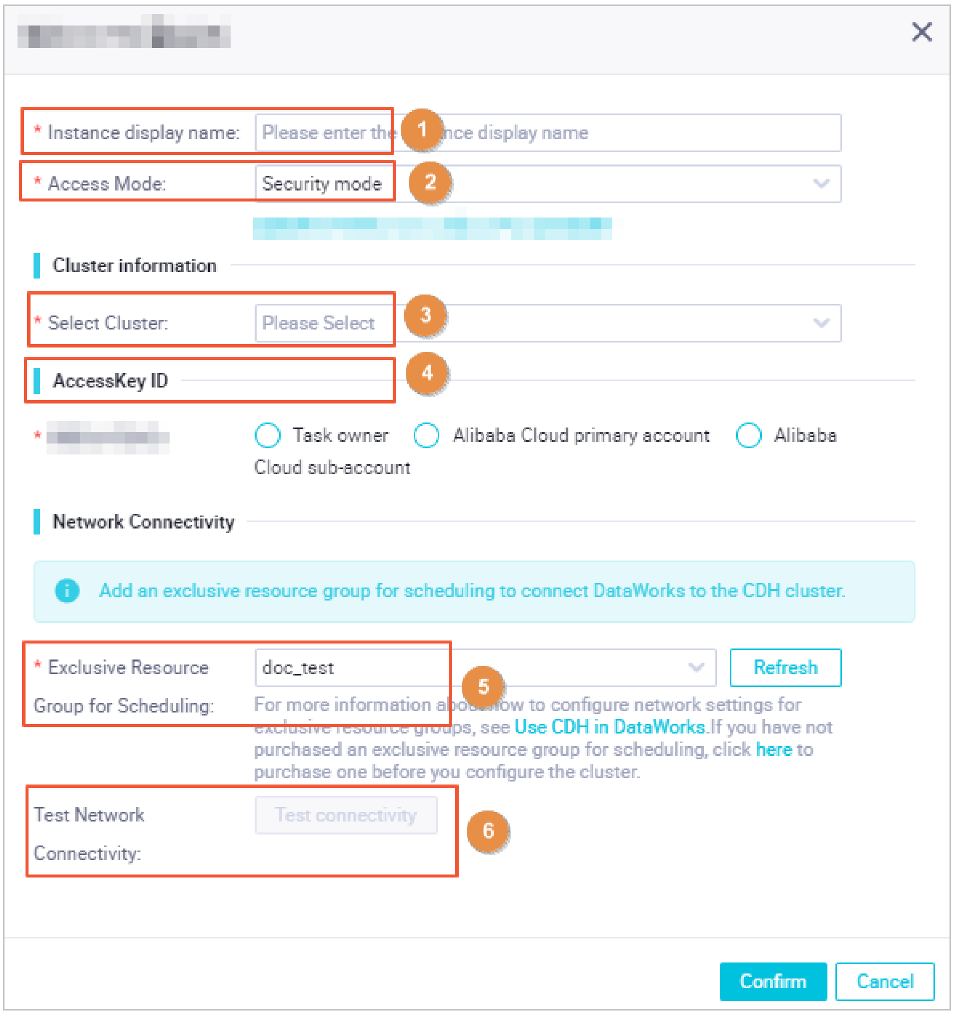
Add an exclusive resource group for scheduling to connect DataWorks to the CDH cluster.

* Exclusive Resource Group for Scheduling: doc_test

For more information about how to configure network settings for exclusive resource groups, see Use CDH in DataWorks.If you have not purchased an exclusive resource group for scheduling, click here to purchase one before you configure the cluster.

Test Network Connectivity:

- The following figure shows the parameters you must configure if you set Access Mode to **Security mode**.



[redacted]

1 * Instance display name: Please enter the instance display name

2 * Access Mode: Security mode

[redacted]

Cluster information

3 * Select Cluster: Please Select

4 AccessKey ID

* [redacted] ○ Task owner ○ Alibaba Cloud primary account ○ Alibaba Cloud sub-account

Network Connectivity

Add an exclusive resource group for scheduling to connect DataWorks to the CDH cluster.

5 * Exclusive Resource Group for Scheduling: doc_test Refresh

For more information about how to configure network settings for exclusive resource groups, see Use CDH in DataWorks.If you have not purchased an exclusive resource group for scheduling, click here to purchase one before you configure the cluster.

6 Test Network Connectivity: Test connectivity

Confirm Cancel

i. Specify **Instance Display Name**.

ii. Specify **Access Mode**.

- **Shortcut mode**

  If this access mode is used, multiple Alibaba Cloud accounts or RAM users map to the same CDH cluster account. These Alibaba Cloud accounts or RAM users can access data in the same CDH cluster account. In this case, data permissions are not isolated.

- **Security mode**

  If this access mode is used, you can configure the mappings between the Alibaba Cloud accounts or RAM users and CDH cluster accounts to isolate the permissions on the data of the node that is run by the Alibaba Cloud accounts or RAM users.

iii. Select the CDH cluster whose configurations you added.

If **Shortcut mode** is selected for **Access Mode**, you must select a CDH cluster whose Authentication Type is not set to Kerberos Account Authentication. If **Security mode** is selected for **Access Mode**, you must select a CDH cluster whose Authentication Type is set to Kerberos Account Authentication. For more information about how to check whether Kerberos Account Authentication is selected for the CDH cluster, see Go to the Workspace Management page.

iv. Configure access authentication information for the CDH cluster.

- **Shortcut mode**

  You can use only the specified accounts, such as admin and hadoop. These accounts are used only to commit nodes.

- **Security mode**

  You can set **Account for Scheduling Nodes** based on your business requirements. This identity is used to automatically schedule and run a node after the node is committed. You must configure mappings between the Alibaba Cloud accounts or RAM users and CDH cluster accounts. For more information about how to configure the mappings, see Configure mappings between Alibaba Cloud accounts or RAM users and Kerberos accounts.

  > ⓘ **Note** On the DataStudio page, the identity used to run nodes is the CDH cluster account that is mapped to the logon Alibaba Cloud account or RAM user. Therefore, you must configure the identity mappings not only for scheduling access identities but also for the workspace developers to prevent nodes from failing to run.

v. Select the created exclusive resource group for scheduling.

vi. Click **Test Connectivity**.

If the connectivity test fails, the exclusive resource group for scheduling is not associated with the VPC to which the CDH cluster belongs or is not configured with hosts. For more information about how to configure the network settings of the exclusive resource group for scheduling, see Step 2: Configure network connectivity.

3. Click **Confirm**.

   Then, the system starts to initialize the exclusive resource group for scheduling. During the initialization, the system installs the client that is used to access the CDH cluster and uploads the configuration files of the CDH cluster. After the value of **Initialization Status of Resource Group** on the CDH tab changes from **Preparing** to **Complete**, the CDH cluster is added to the workspace as a compute engine instance.

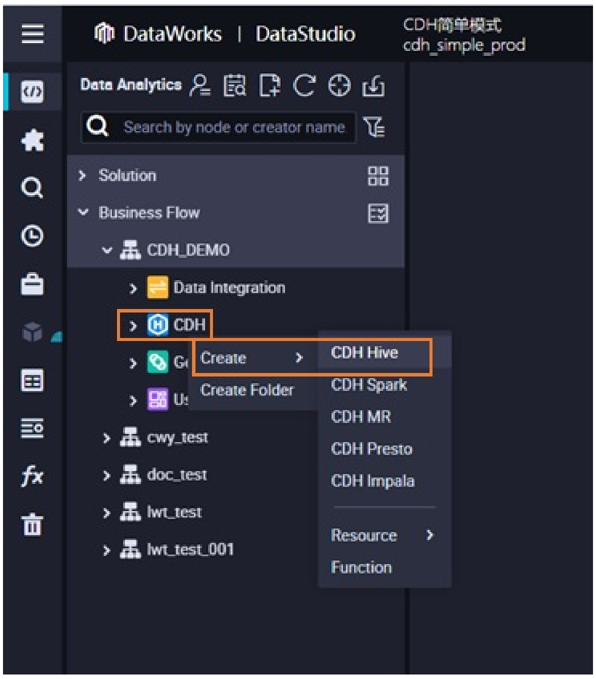
4. Click **Test Connectivity** next to Test Service Connectivity on the CDH tab. Then, DataWorks runs a test task to check whether the client is installed and the configuration files are uploaded.

   If the test fails, you can view the logs and submit a ticket to consult technical support of DataWorks.

Use DataWorks to develop nodes

After you add the CDH compute engine instance, you can create and run CDH Hive, CDH Spark, CDH MR, CDH Impala, or CDH Presto nodes in DataStudio. You can also configure properties for the nodes. In this section, a CDH Hive node is created and run to demonstrate how to use a CDH node to develop data.

1.
2. On the DataStudio page, move the pointer over the Create icon and click Workflow. In the Create Workflow dialog box, configure the parameters and click Create.
3. In the left-side navigation pane, click Business Flow, find the created workflow, and then click the workflow name. Right-click **CDH** and choose **Create > CDH Hive**.



4. In the code editor, write SQL code for the CDH Hive node and click the ▶ icon in the top toolbar.

   In the Parameters dialog box, select the exclusive resource group for scheduling you want to use and click OK. After the code is run, you can view the results.

5. If you want to configure properties for the node, click the **Properties** tab in the right-side navigation pane. On the Properties tab, configure time properties, resource properties, and scheduling dependencies for the node. Then, commit the node. After the node is committed, the system runs the node based on the configured properties. For more information about how to configure properties for a node, see Configure basic properties.

6. Go to the Operation Center page and view the status of the node on the Cycle Task page. For more information, see View auto triggered nodes.

Configure O&M and monitoring settings

CDH nodes support the intelligent monitoring feature provided by DataWorks Operation Center. This feature allows you to customize alert rules and configure alerting for CDH nodes. The system automatically generates alerts if errors occur on the CDH nodes based on the configured alert rules. For more information about how to create custom alert rules, see 自定义规则. For more information about how to configure alerting for nodes, see Manage baselines.

Configure data quality rules

When you use CDH in DataWorks, you can use the Data Quality service of DataWorks to query and compare data, monitor data quality, scan SQL code, and perform intelligent alerting. For more information about the Data Quality service, see Overview.

Use Data Map to collect data

When you use CDH in DataWorks, you can use the Data Map service of DataWorks to collect the metadata of Hive databases, tables, fields, and partitions in the CDH cluster. This facilitates global data searches, viewing of metadata details, data preview, data lineage management, and data category management.

> **Note** You can use Data Map to collect the metadata only of Hive databases in CDH clusters.

For more information about the Data Map service and related configurations, see Overview.

If you want to monitor the metadata changes of Hive databases in a CDH cluster in real time or view lineage and metadata change records in Data Map, associate DataWorks Hive hooks with the CDH cluster. Then, use Log Service to collect the logs generated by the hooks.

After the Hive hooks are configured, metadata changes are recorded in the log file */tmp/hive/hook.event.*.log* on the HiveServer2 and Hive Metastore hosts. In this case, you can use Log Service to collect the change records for DataWorks to read. Download the DataWorks tool `dw-tools.jar`, create a `config.json` file in the same directory, and then specify the configuration items in the file. Then, run the tool to enable log collection.

To configure Hive hooks and collect logs from the hooks, perform the following steps:

1. Configure Hive hooks.

   i. Log on to the HiveServer2 and Hive Metastore hosts and go to the */var/lib/hive* directory to download DataWorks Hive hooks.

   ```
   # Download dataworks-hive-hook-2.1.1.jar for CDH 6.X clusters.
   wget https://dataworks-public-tools.oss-cn-shanghai.aliyuncs.com/dataworks-hive-hook-2.1.1.jar
   # Download dataworks-hive-hook-1.1.0-cdh5.16.2.jar for CDH 5.X clusters.
   wget https://dataworks-public-tools.oss-cn-shanghai.aliyuncs.com/dataworks-hive-hook-1.1.0-cdh5.16.2.jar
   ```

   ii. Log on to the Cloudera Manager Admin Console and click **Hive** below the cluster name. On the page that appears, click the **Configuration** tab. Then, set Hive Auxiliary JARs Directory to `/var/lib/hive`.

   iii. For **Hive Service Advanced Configuration Snippet (Safety Valve) for hive-site.xml**, specify the Name and Value fields based on the following information:

   ```
   <property>
     <name>hive.exec.post.hooks</name>
     <value>com.cloudera.navigator.audit.hive.HiveExecHookContext,org.apache.hadoop.hive.ql.hooks.LineageLogger,com.aliyun.dataworks.meta.hive.hook.LineageLoggerHook</value>
   </property>
   ```

   iv. For **Hive Metastore Server Advanced Configuration Snippet (Safety Valve) for hive-site.xml**, specify the Name and Value fields based on the following information:

   ```
   <property>
     <name>hive.metastore.event.listeners</name>
     <value>com.aliyun.dataworks.meta.hive.listener.MetaStoreListener</value>
   </property>
   <property>
     <name>hive.metastore.pre.event.listeners</name>
     <value>com.aliyun.dataworks.meta.hive.listener.MetaStorePreAuditListener</value>
   </property>
   ```

v. After the Hive hooks are configured, you must perform configurations on clients as prompted in the Cloudera Manager Admin Console. Then, restart the Hive service.

> **Note** If the restart fails, retain the logs for troubleshooting. To prevent normal operations from being affected, you can remove the added information and restart the Hive service again. If the restart succeeds after the information is added, check whether the log files whose names start with `hook.event`, such as `hook.event.1608728145871.log`, are generated in the */tmp/hive/* directory on the hosts.

2. Collect logs from the Hive hooks.

   i. Log on to the Cloudera Manager Admin Console and download the DataWorks JAR package.

   ```
   wget https://dataworks-public-tools.oss-cn-shanghai.aliyuncs.com/dw-tools.jar
   ```

   ii. Create a `config.json` file in the directory in which the DataWorks tool is stored. Then, modify the file based on the following code and save the file:

   ```
   // config.json
   {
       "accessId": "<accessId>",
       "accessKey": "<accessKey>",
       "endpoint": "cn-shanghai-intranet.log.aliyuncs.com",
       "project": "onefall-test-pre",
       "clusterId": "1234",
       "ipList": "192.168.0.1,192.168.0.2,192.168.0.3"
   }
   ```

   Configuration information:

   - **accessId**: the AccessKey ID of your Alibaba Cloud account.
   - **accessKey**: the AccessKey secret of your Alibaba Cloud account.
   - **endpoint**: the internal endpoint that is used to access your Log Service project. For more information, see Endpoints.
   - **project**: the name of your Log Service project. For more information about how to obtain the name, see Manage a project.
   - **clusterId**: the ID of the CDH cluster generated for DataWorks. You can submit a ticket to obtain the ID.
   - **ipList**: the IP addresses of all HiveServer2 and Hive Metastore hosts. Separate the IP addresses with commas (,). The hosts are those on which the DataWorks Hive hooks are deployed.

   iii. Run the config.json file.

   ```
   java -cp dw-tools.jar com.aliyun.dataworks.tools.CreateLogConfig config.json
   ```

   iv. Install the client.

   ```
   wget http://logtail-release-cn-shanghai.oss-cn-shanghai.aliyuncs.com/linux64/logtail.sh -O logtail.sh; chmod 755 logtail.sh; ./logtail.sh install cn-shanghai
   ```

   Replace **cn-shanghai** with the region where your Log Service project resides.

3. After you complete the preceding steps, a Logstore named hive-event, a Logtail configuration

named hive-event-config, and a log group named hive-servers are generated in your Log Service project. You can view and record the ID of your Alibaba Cloud account, the endpoint of your Log Service project, and other information about the project. Then, submit a ticket to send the recorded information to the technical support personnel of DataWorks. This way, the technical personnel can perform subsequent configurations.