ALIBABA CLOUD

Alibaba Cloud

DataWorks

Data governance

Document Version: 20210517

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.Security Center

1.1. Overview

Security Center allows you to manage permissions with ease and submit and handle requests on a visualized interface.

> **Notice** The Security Center service is in invitational preview. The service is available in the China (Shanghai), China (Hangzhou), China (Beijing), China (Zhangjiakou), China (Shenzhen), and China (Chengdu) regions.

On the **My Permissions**, **Authorizations**, and **Approval Center** pages, Security Center provides the following features:

- Self-service permission request: Users can select the required tables to submit a permission request online. This online request mode is more efficient than the original mode in which users need to contact administrators offline.
- Permission management: Administrators can view the users who have permissions on database tables and revoke permissions as required. Users can also remove unnecessary permissions themselves.
- Permission request approval: Before permissions are granted to users, administrators approve permission requests that are submitted by users. This implements a visual and process-based permission management system, and allows you to review the approval process.

In Security Center, you can view permissions on all the tables under a tenant, request and manage table permissions, and approve or reject permission requests.

Each operation in Security Center applies to all the workspaces of a tenant in standard mode and basic mode.

1.2. Quick start

This topic describes how to use the Security Center service as users who assume different roles.

Prerequisites

Note the following items before you use the Security Center service:

- Field-level authorization and LabelSecurity

  You can request permissions on fields only in a MaxCompute project with LabelSecurity enabled. If LabelSecurity is disabled for a MaxCompute project, you can request permissions only on tables in this MaxCompute project and cannot specify the validity period of permissions. For more information about LabelSecurity, see Column-level access control.

- Validity period

  If you want to make sure that field permissions are valid in the specified validity period, specify the security level of each field higher than the security level of your account.

  After permissions on a table are granted to you, you automatically obtain permissions on the fields whose security level is not specified or not higher than the security level of your account. The permissions on these fields are permanently valid and cannot be separately revoked.

- Permissions displayed in Security Center

Security Center displays only the permissions that are granted by using access control list (ACLs) rather than the permissions that are granted by using other methods such as roles. For example, a workspace developer has the permission to access all tables in the workspace but Security Center does not display these permissions. If you can access a table but Security Center does not display the permissions on the table, contact the system administrator to check whether the permissions are granted by using other methods such as roles.

Context

If you use different accounts that assume different roles, you can perform different operations.

- RAM users that assume the common user role
  - On the **My Permissions** page, you can view permissions, request permissions, and revoke table and field permissions.
  - On the **My Requests** tab of the Approval Center page, you can view the requests that you submitted and their approval status.
- RAM users that assume the table owner role
  - On the **My Permissions** page, you can view permissions, and request and revoke permissions on a table or specific fields of the table that is not owned by you.
  - On the **My Requests** tab of the Approval Center page, you can view the requests that you submitted and their approval status.
  - On the **Pending My Approval** tab of the Approval Center page, you can view and handle the requests that are pending your approval.
  - On the **Handled by Me** tab of the Approval Center page, you can view the requests that you have handled.
- RAM users that assume the workspace administrator role
  - On the **Authorizations** page, you can view the workspace members who have permissions on tables and revoke permissions.
  - On the **My Requests** tab of the Approval Center page, you can view the requests that you submitted and their approval status.
  - On the **Pending My Approval** tab of the Approval Center page, you can view and handle the requests that are pending your approval.
  - On the **Handled by Me** tab of the Approval Center page, you can view the requests that you have handled.
- Alibaba Cloud accounts
  - On the **Authorizations** page, you can view the workspace members who have permissions on tables and revoke permissions.
  - On the **My Requests** tab of the Approval Center page, you can view the requests that you submitted and their approval status.
  - On the **Pending My Approval** tab of the Approval Center page, you can view and handle the requests that are pending your approval.
  - On the **Handled by Me** tab of the Approval Center page, you can view the requests that you have handled.

In this example, the common user, table owner, and workspace administrator roles are used.

In this example, the following operations are performed:

- Log on as RAM user A that assumes the common user role to view the permissions of RAM user A.
- Log on as RAM user A to request permissions on Table A and Table B on which RAM user A does not have permissions.
- Log on as RAM user B that is the owner of Table A to handle a request for permissions on Table A.
- Log on with an Alibaba Cloud account that assumes the workspace administrator role to handle a request for permissions on Table B.
- Log on as RAM user A to revoke permissions on specific fields in Table A.
- Log on as RAM user A to revoke permissions on Table A.
- Log on with the Alibaba Cloud account to revoke permissions on Table B that are granted to RAM user A.

Go to the Security Center page

1. Log on to the DataWorks console.
2. In the left-side navigation pane, click **Workspaces**.
3. After you select the region where the required workspace resides, find the workspace and click **Data Analytics**.
4. On the DataStudio page, click the ☰ icon in the upper-left corner and choose **All Products > Security Center**.

Manage permissions as a common user

- To view the permissions in a workspace, perform the following steps:
    i. Log on to the DataWorks console as RAM user A. Go to the **Security Center** page. By default, the **My Permissions** page appears.
    ii. On the **My Permissions** page, select a workspace and an environment to view the tables of the workspace in the environment and the tables on which you have permissions.
- To request permissions on Table A and Table B, perform the following steps:
    i. Log on to the DataWorks console as RAM user A. Go to the **Security Center** page. By default, the **My Permissions** page appears.
    ii. On the My Permissions page, select the fields in Table A and Table B on which you want to request permissions and click **Request Permission**.
    iii. On the **Table Permission Request** page, set the parameters as required.
    iv. Click **Submit**.
- To view the approval status of a request, perform the following steps:
    i. Log on to the DataWorks console as RAM user A. Go to the **Security Center** page.
    ii. In the left-side navigation pane, click **Approval Center**.
    iii. On the **My Requests** tab, view the status of a request in the **Status** column.

        If your request is in the **Approved** state, you are granted the requested table permissions.
- To revoke permissions on specific fields in Table A, perform the following steps:
    i. Log on to the DataWorks console as RAM user A. Go to the **Security Center** page. By default, the **My Permissions** page appears.
    ii. On the **Table** tab, find Table A and choose **More** > **Revoke Field Permission** in the Actions

column.

   iii. In the **Revoke Field Permission** dialog box, select the fields on which you want to revoke permissions.

   iv. Click **OK**.

- To revoke permissions on Table A, perform the following steps:

   i. Log on to the DataWorks console as RAM user A. Go to the **Security Center** page. By default, the **My Permissions** page appears.

   ii. On the **Table** tab, find Table A and choose **More** > **Revoke Permission** in the Actions column.

   iii. In the **Revoke Permission** dialog box, select the permissions that you want to revoke.

   iv. Click **OK**.

Manage permissions as a table owner

As the owner of Table A, RAM user B can handle a request for permissions on Table A.

A table owner is also a common user. In addition to the operations that can be performed by a common user, the owner of a table can also handle the requests for permissions on the table.

1. Log on to the DataWorks console as RAM user B. Go to the **Security Center** page.
2. In the left-side navigation pane, click **Approval Center**.
3. On the Approval Center page, click the **Pending My Approval** tab.
4. On the Pending My Approval tab, find the request that is submitted by RAM user A and click **Handle** in the Actions column. On the Request Details page, view the progress and objects on which permissions are requested.
5. Enter your comments and click **Approve** or **Reject** as required.

Manage permissions as a workspace administrator

- To handle a request for permissions on Table B, perform the following steps:

   i. Log on to the DataWorks console by using the Alibaba Cloud account. Go to the **Security Center** page.

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

   iii. On the Approval Center page, click the **Pending My Approval** tab.

   iv. On the Pending My Approval tab, find the request that is submitted by RAM user A and click **Handle** in the Actions column. On the Request Details page, view the progress and objects on which permissions are requested.

   v. Enter your comments and click **Approve** or **Reject** as required.

- To revoke permissions on Table B that are granted to RAM user A, perform the following steps:

   i. Log on to the DataWorks console by using the Alibaba Cloud account. Go to the **Security Center** page.

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

   iii. On the Table tab, find Table B and click the ⊞ icon before the table name to show the accounts that have permissions on the table.

   iv. Find RAM user A and click **Revoke Permission** in the Actions column.

   v. In the **Revoke Permission** dialog box, select the permissions that you want to revoke.

vi. Click **OK**.

1.3. My Permissions

On the My Permissions page, you can view, request, and revoke permissions on tables and fields in a workspace.

View permissions

1. Log on to the DataWorks console.
2. In the left-side navigation pane, click **Workspaces**.
3. On the Workspaces page, find the workspace in which you want to view permissions and click **Data Analytics** in the Actions column.
4. On the DataStudio page, click the ☰ icon in the upper-left corner and choose **All Products > Security Center**. The **My Permissions** page appears.
5. On the **Table** tab, select a workspace and an environment to view all the tables of the workspace in the environment. You can also enter a table name in the **Table** field to search for required tables in fuzzy match mode.You can view the names and owners of tables in a workspace, view your permissions on the tables, and request or revoke table and field permissions.

Request permissions

1. Go to the **My Permissions** page.
2. Select the tables or fields on which you want to request permissions.
   - Request permissions on a table or specific fields in the table

     Click the plus sign (+) before a table name, select the required fields on which you have no permissions, and then choose **More > Request Permission** in the Actions column.

     Alternatively, find a table and choose **More > Request Permission** in the Actions column without selecting fields to request permissions on all the fields in the table.

     > **Notice** You can request permissions on fields only in a MaxCompute project with LabelSecurity enabled. If LabelSecurity is disabled for a MaxCompute project, you can request permissions only on tables in this MaxCompute project.

   - Request permissions on multiple tables and fields

     Select all the required tables and fields and click **Request Permission**.

     > **Note** You can also click **Request Permission** without selecting tables or fields, and select the required tables and fields on the **Table Permission Request** page.

3. On the **Table Permission Request** page, set the parameters as required.

Parameter	Description
Workspace	The workspace, which is automatically specified based on the information you specified on the **My Permissions** page. You can change the workspace as required.

Parameter	Description
Environment	The environment of the workspace for which you request permissions. A workspace in basic mode has only the production environment.
MaxCompute Project	The name of the MaxCompute project.
Grant To	The account for which you request permissions. You can request permissions for the current account or a production account of another workspace you joined.
Valid Until	The validity period of permissions. Valid values: **1 Month**, **3 Months**, **6 Months**, **1 Year**, **Permanent**, and **Others**.
Reason for Request	The reason why you request permissions.
Objects Requested	The tables on which you request permissions.

4. Click **Submit**.

Revoke field permissions

Notice

- You can revoke permissions on fields only in a MaxCompute project with LabelSecurity enabled.
- If you want to revoke permissions on all the fields in a table, revoke the permissions on the table.

1. Go to the **My Permissions** page.
2. Find the table on which you want to revoke permissions and choose **More > Revoke Field Permission** in the **Actions** column.
3. In the **Revoke Field Permission** dialog box, select the fields on which you want to revoke permissions.
4. Click **OK**.

Revoke table permissions

1. Go to the **My Permissions** page.
2. Find the table on which you want to revoke permissions and choose **More > Revoke Permission** in the **Actions** column.
3. In the **Revoke Permission** dialog box, select the permissions that you want to revoke.
4. Click **OK**.

1.4. Authorizations

On the Authorizations page, a workspace administrator can view the accounts that have permissions on tables and fields in each workspace, and revoke unnecessary table and field permissions.

Go to the **Security Center** page. In the left-side navigation pane, click **Authorizations**. On the **Table** tab of the Authorizations page, you can view and search for tables in workspaces of the current tenant.

On the Table tab, you can select a workspace and an environment to view all the tables of the workspace in the environment. You can also enter a table name in the Table field to search for required tables in fuzzy match mode.

View accounts that have permissions on a table

On the **Table** tab of the **Authorizations** page, click the plus sign (+) before a table name to view all the accounts that have permissions on the table.

Revoke table permissions

Find an account and click **Revoke Permission** in the Actions column to revoke the permissions of the account on the current table.

View field permissions

Find an account and click **View Field Permissions** in the Actions column. On the Field Permissions page, view the permissions of the account on the fields in the current table.

Revoke field permissions

If LabelSecurity is enabled for the corresponding MaxCompute project, select fields on the Field Permissions page and click **Revoke Field Permissions** to revoke the permissions on the fields.

1.5. Approval Center

On the Approval Center page, you can view the requests that you submitted and their approval status, view and handle the requests that are pending your approval, and view the requests that you have handled.

My Requests

1. Go to the **Security Center** page. In the left-side navigation pane, click **Approval Center**. On the Approval Center page, click the **My Requests** tab.

   On this tab, you can view the following information about each of your requests: object type, workspace, MaxCompute project, tables, request time, and status.

   > **Note** If a request contains permission requests for tables that belong to different owners, Security Center automatically splits the request into multiple requests by table owner.

2. Find a request and click **View** in the Actions column to view the details.

Pending My Approval

1. On the **Approval Center** page, click the **Pending My Approval** tab.

   On this tab, you can view the requests that are pending your approval. If a request is pending your approval, a red dot appears next to **Approval Center** and **Pending My Approval** to remind you.

   You can view the following information about each request that is pending your approval: object type, account that submits the request, workspace, MaxCompute project, tables, and request time.

2. Find a request and click **Handle** in the Actions column. On the Request Details page, view the progress and objects on which permissions are requested.
3. Enter your comments and click **Approve** or **Reject** as required.

Handled by Me

1. On the **Approval Center** page, click the **Handled by Me** tab.

   On this tab, you can view the following information about each request that you have handled: object type, account that submits the request, workspace, MaxCompute project, tables, and request time.

2. Find a request and click **View** in the Actions column. On the Request Details page, view the progress and objects on which permissions are requested.

2.Data Quality

2.1. Overview

DataWorks provides the Data Quality service for you to control the data quality of heterogeneous data stores. In Data Quality, you can check data quality, configure alert notifications, and manage connections.

In DataWorks, Data Quality provides a comprehensive data quality solution that has various features. For example, you can detect data, compare data, monitor data quality, scan SQL nodes, and use intelligent alerting.

Data Quality can monitor data processing throughout the process, detect issues based on monitoring rules, and send alert notifications to alert recipients in real time.

Data Quality monitors data quality by dataset. Data Quality allows you to monitor the data quality of E-MapReduce tables, Hologres tables, AnalyticDB for PostgreSQL tables, MaxCompute tables, and DataHub topics. When offline data changes, Data Quality checks the data and blocks nodes that use the data if it detects anomalies. This prevents the nodes from being affected. Data Quality also 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 set the alert severity, such as warning and error alerts, and the alert frequency to minimize repeated alerts.

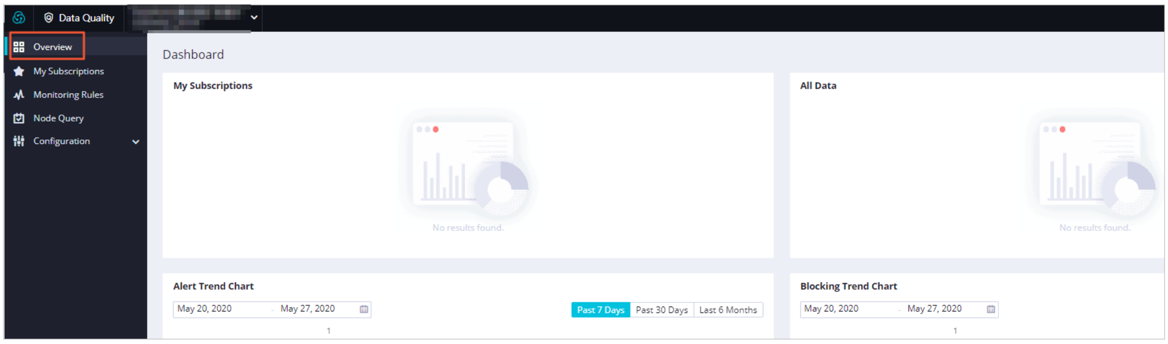
> **Note** Data Quality monitors the data quality of E-MapReduce tables, Hologres tables, AnalyticDB for PostgreSQL tables, MaxCompute tables, and DataHub topics. Before you use the Data Quality service, you must create tables or topics and write data to the tables or topics.

2.2. Go to the Overview page

The Overview page provides an overview of alerts and blocks triggered by tables and topics that you subscribed to.

Procedure

1. Log on to the DataWorks console.
2. In the left-side navigation pane, click **Workspaces**.
3. After you select the region where the required workspace resides, find the workspace and click **Data Analytics**.
4. On the DataStudio page that appears, click in the upper-left corner and choose **All Products > Data Quality**. The **Overview** page appears by default.



Section	Description
My Subscriptions	This section separately displays the number of E-MapReduce tables, AnalyticDB for PostgreSQL tables, MaxCompute tables, and Datahub topics that you subscribed to and the number of those with alerts and blocks triggered.
All Data	This section separately displays the total number of E-MapReduce tables, AnalyticDB for PostgreSQL tables, MaxCompute tables, and Datahub topics and the number of those with alerts and blocks triggered.
Alert Trend Chart	This section displays the trend of alerts triggered by E-MapReduce tables, AnalyticDB for PostgreSQL tables, MaxCompute tables, and Datahub topics in the last 7 days, 30 days, and 6 months.
Blocking Trend Chart	This section displays the trend of blocks triggered by E-MapReduce tables, AnalyticDB for PostgreSQL tables, MaxCompute tables, and Datahub topics in the last 7 days, 30 days, and 6 months.

2.3. View my subscriptions

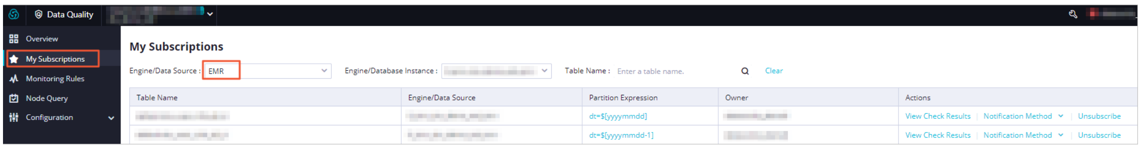
The My Subscriptions page displays E-MapReduce tables, Hologres tables, AnalyticDB for PostgreSQL tables, MaxCompute tables, and DataHub topics that you subscribed to.

Procedure

1. Log on to the DataWorks console.
2. In the left-side navigation pane, click **Workspaces**.
3. After you select the region where the required workspace resides, find the workspace and click **Data Analytics**.
4. Click the ☰ icon in the upper-left corner and choose **All Products > Data governance > Data Quality**.
5. In the left-side navigation pane, click **My Subscriptions**. Data Quality supports E-MapReduce, Hologres, AnalyticDB for PostgreSQL, MaxCompute, and DataHub data stores. You can specify a data store on the **My Subscriptions** page and view the tables or topics that you subscribed to.
   - E-MapReduce, Hologres, AnalyticDB for PostgreSQL, and MaxCompute tables

     On the **My Subscriptions** page, select **EMR** from the **Engine/Data Source** drop-down list in the upper-left corner and select an instance from the **Engine/Database Instance** drop-down list. The E-MapReduce tables that you subscribed to are displayed.

You can view Hologres, AnalyticDB for PostgreSQL, and MaxCompute tables that you subscribed to by using the same method.



My Subscriptions

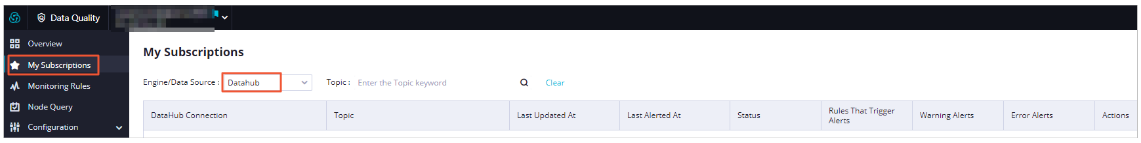
Engine/Data Source : EMR   Engine/Database Instance : [redacted]   Table Name : Enter a table name.   Clear

Table Name	Engine/Data Source	Partition Expression	Owner	Actions
[redacted]	[redacted]	dt=$[yyyymmdd]	[redacted]	View Check Results | Notification Method | Unsubscribe
[redacted]	[redacted]	dt=$[yyyymmdd-1]	[redacted]	View Check Results | Notification Method | Unsubscribe

- Find a table and click the partition filter expression in the **Partition Expression** column to go to the page where you can edit the configured monitoring rules for the table. For more information, see Configure monitoring rules.
- Find a table and click **View Check Results** to go to the page where you can view the monitoring results of the table. For more information, see View monitoring results.
- Find a table and click **Notification Method**. You can change the method for sending alert notifications that are triggered by the table. Data Quality supports the following four notification methods: **Email**, **Email and SMS**, **DingTalk Chatbot**, and **DingTalk Chatbot @ALL**. For more information about how to configure a DingTalk chatbot in the DataWorks console to receive alert notifications in a DingTalk group, see Manage custom alert rules.
- Find a table and click **Unsubscribe**. You can unsubscribe from the table.

- DataHub topics

  On the **My Subscriptions** page, select **Datahub** from the Engine/Data Source drop-down list in the upper-left corner. The DataHub topics that you subscribed to are displayed.



My Subscriptions

Engine/Data Source : Datahub   Topic : Enter the Topic keyword   Clear

DataHub Connection	Topic	Last Updated At	Last Alerted At	Status	Rules That Trigger Alerts	Warning Alerts	Error Alerts	Actions

- Find a topic and click **Alerts** to go to the page where you can view the details about the alerts that are triggered by the topic.
- Find a topic and click **Notification Method**. You can change the method for sending alert notifications that are triggered by the topic.
- Find a topic and click **Unsubscribe**. You can unsubscribe from the topic.

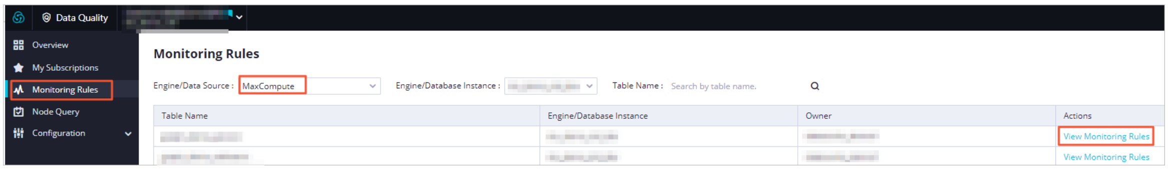
2.4. Configure monitoring rules

Data Quality allows you to configure monitoring rules for data in E-MapReduce, Hologres, AnalyticDB for PostgreSQL, MaxCompute, and DataHub data stores. This topic describes how to configure a rule for monitoring data in a MaxCompute data store.

Go to the Monitoring Rules page

1. Log on to the DataWorks console.
2. In the left-side navigation pane, click **Workspaces**.
3. After you select the region where the required workspace resides, find the workspace and click **Data Analytics**.
4. Click the ☰ icon in the upper-left corner and choose **All Products > Data governance > Data Quality**.
5. In the left-side navigation pane, click **Monitoring Rules**.

6. Select **MaxCompute** from the **Engine/Data Source** drop-down list and select a MaxCompute project from the **Engine/Database Instance** drop-down list.



Monitoring Rules

Engine/Data Source : MaxCompute Engine/Database Instance : [redacted] Table Name : Search by table name.

Table Name	Engine/Database Instance	Owner	Actions
[redacted]	[redacted]	[redacted]	View Monitoring Rules
[redacted]	[redacted]	[redacted]	View Monitoring Rules

Data Quality supports E-MapReduce, Hologres, AnalyticDB for PostgreSQL, MaxCompute, and DataHub data stores.

- If you select an E-MapReduce, a Hologres, an AnalyticDB for PostgreSQL, or a MaxCompute data store, all tables in the data store are displayed.
- If you select a DataHub data store, all topics in the data store are displayed.

7. Find a table and click **View Monitoring Rules**.Data Quality allows you to configure template rules and custom rules.

> **Notice** Before you configure a template rule, you must configure a partition filter expression. For more information, see Configure a partition filter expression.

Create a template rule

1. Find a table and click **View Monitoring Rules** to go to the **Monitoring Rules** page of the table.
2. Click **Create rules**. The **Template Rules** tab appears in a panel. To create a template rule, you can click **Add Monitoring Rule** or **Quick Create**.

- **Add Monitoring Rule**

  Click **Add Monitoring Rule**. The following table describes the parameters that are displayed if you set the Rule Source parameter to **Built-in Template**.

Parameter	Description
Rule Name	The name of the rule.
Rule Type	Valid values: Rule Type and Soft. ■ If you select Rule Type, error alerts are reported and descendant nodes are blocked, whereas warning alerts are reported but descendant nodes are not blocked. ■ If you select Soft, error alerts are reported but descendant nodes are not blocked, whereas warning alerts are not reported and descendant nodes are not blocked.
Auto-Generated Threshold	Specifies whether to use dynamic thresholds. Set this parameter as needed. **Notice** You can use the dynamic threshold feature only in DataWorks Enterprise Edition or more advanced editions.

Parameter	Description
Rule Source	Valid values: **Built-in Template** and **Rule Templates**. If you select **Rule Templates**, you must specify a rule template. For more information, see Create, manage, and use rule templates. **Notice** You can select **Rule Templates** only in DataWorks Enterprise Edition or more advanced editions.
Field	You can select All Fields in Table or a specific field of a numeric type or non-numeric type.
Template	Data Quality supports 43 rule templates. You can select only the rule templates that are displayed. For more information, see Built-in rule templates for offline data. **Note** You can set field-specific rules of the average value, accumulated value, minimum value, and maximum value only for numeric fields.
Comparison Method	Valid values: **Absolute Value**, **Raise**, and **Drop**.

Parameter	Description
Thresholds	■ Calculate the fluctuation. You can calculate the fluctuation bv using the following formula: **Fluctuation = (Sample - Baseline)/Baseline** . ■ Sample The sample value for the current day. For example, if you want to check the fluctuation of table rows on an SQL node in a day, the sample is the number of table rows on that day. ■ Baseline The comparison value from the previous N days. Examples: ■ If you want to check the fluctuation of table rows on an SQL node in a day, the baseline is the number of table rows on the previous day. ■ If you want to check the average fluctuation of table rows on an SQL node in seven days, the baseline is the average number of table rows in the last seven days. ■ Calculate the fluctuation variance. You can calculate the fluctuation variance only for numeric fields such as BIGINT and DOUBLE fields by using the following formula: Fluctuation variance = (Sample - Average value of past N days)/Standard deviation. You can specify the warning threshold and error threshold of the fluctuation to monitor data at different severities: ■ If the absolute value of the fluctuation does not exceed the warning threshold, the data is considered to be normal. ■ If the absolute value of the fluctuation does not meet the condition in Case 1 and does not exceed the error threshold, a warning alert is reported. ■ If the fluctuation does not meet the condition in Case 2, an error alert is reported.
Description	The description of the rule.

- Quick Create

  Click **Quick Create**. Set the parameters as required.

Parameter	Description
Rule Name	The name of the rule.
Field	You can select All Fields in Table or a specific field of a numeric type or non-numeric type.

Parameter	Description
Trigger	Valid values: **The number of columns is greater than 0** and **Table row number dynamic threshold**. Notice You can select **Table row number dynamic threshold** only in DataWorks Enterprise Edition or more advanced editions.

3. Click **Batch Create**.

Create a custom rule

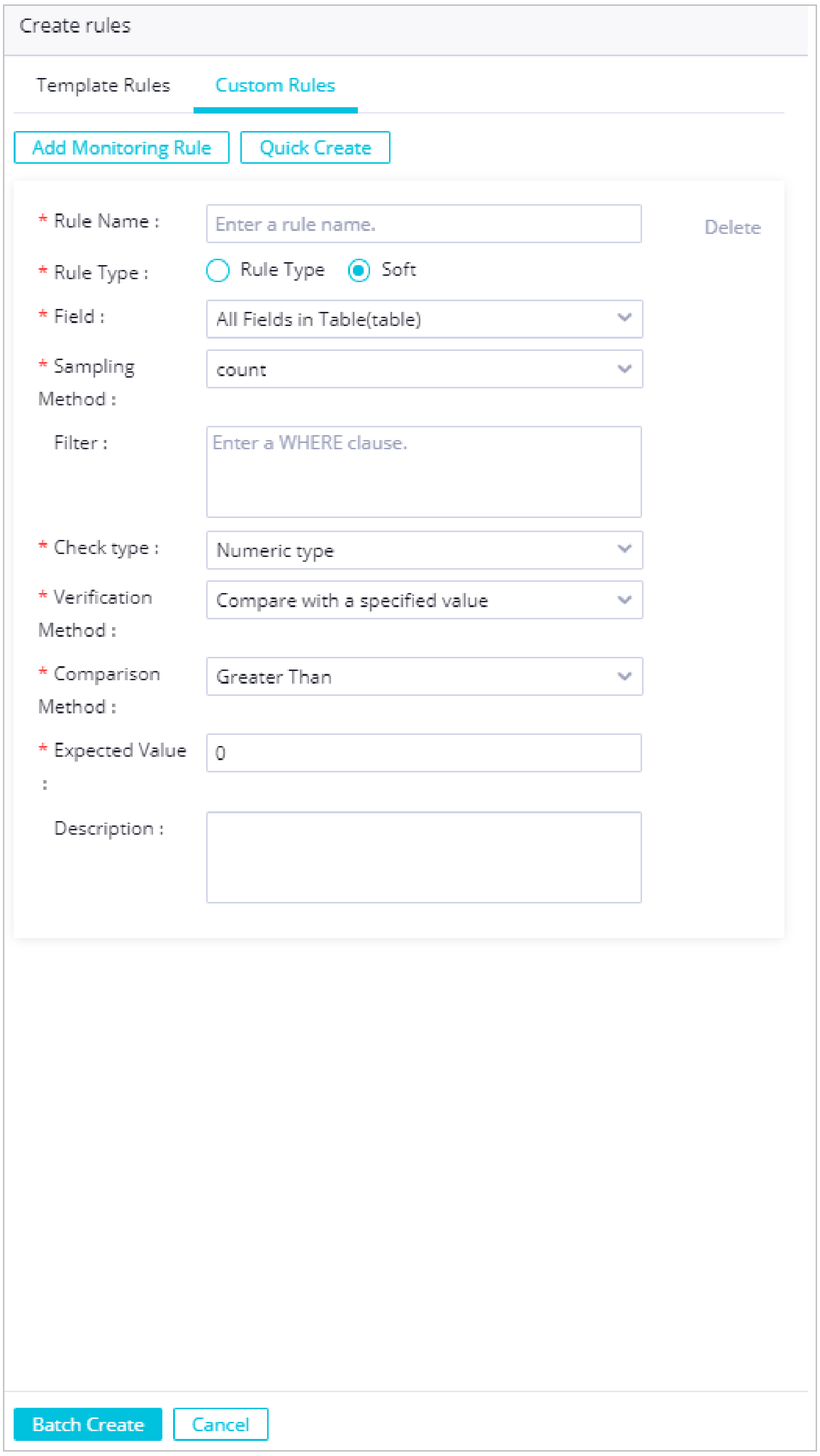
If template rules do not meet your requirements for monitoring the data quality based on a partition filter expression, you can create custom rules to meet your personalized monitoring requirements.

1. Find a table and click **View Monitoring Rules** to go to the **Monitoring Rules** page of the table.
2. Click **Create rules**. The **Template Rules** tab appears in a panel.
3. Click the **Custom Rules** tab.To create a custom rule, you can click **Add Monitoring Rule** or **Quick Create**.

   - **Add Monitoring Rule**

     You can select **All Fields in Table**, **SQL Statement**, or a specific field for the **Field** parameter.

     - Select **All Fields in Table** or a specific field.



Create rules

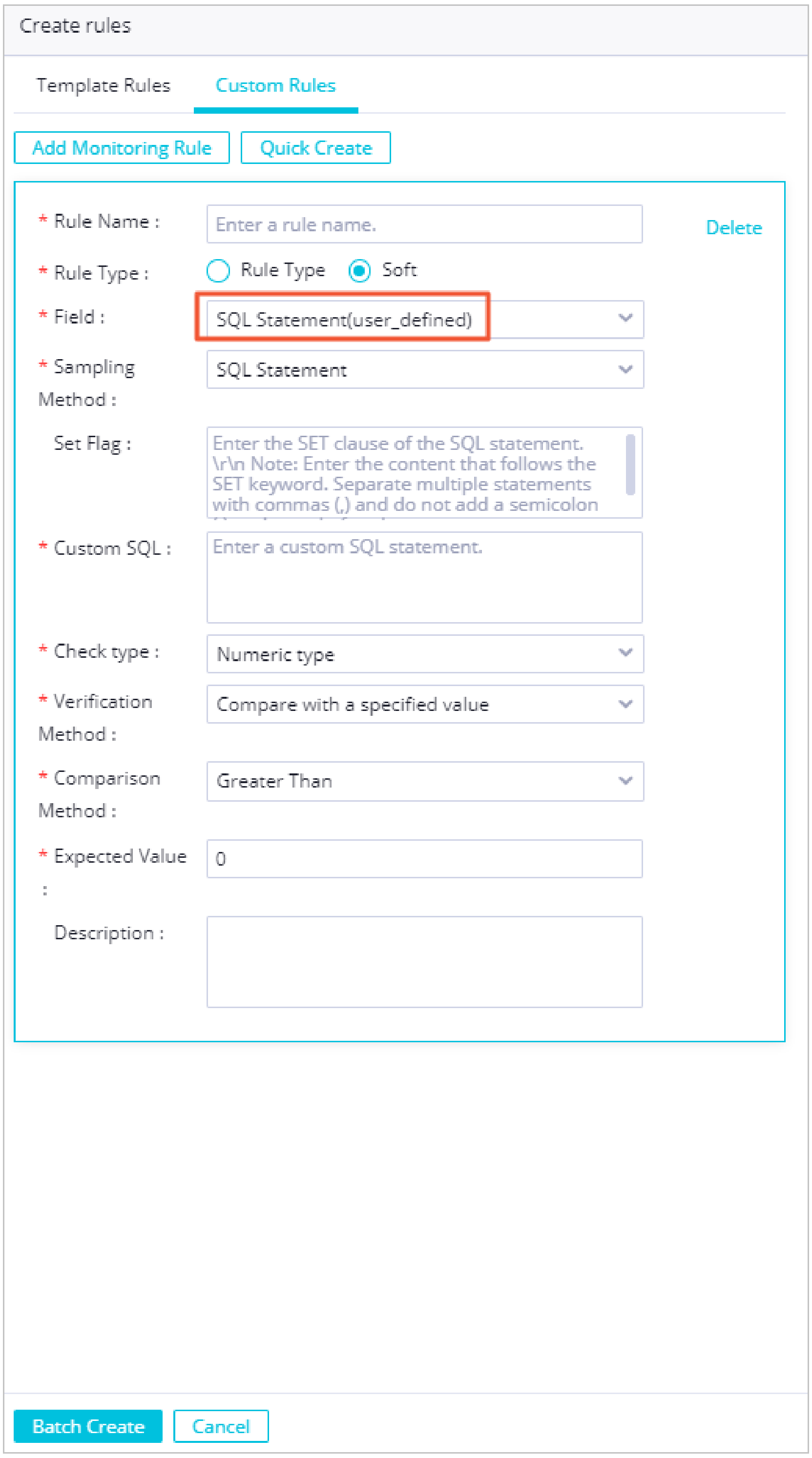
* Rule Name :	Enter a rule name.
* Rule Type :	○ Rule Type ● Soft
* Field :	All Fields in Table(table)
* Sampling Method :	count
Filter :	Enter a WHERE clause.
* Check type :	Numeric type
* Verification Method :	Compare with a specified value
* Comparison Method :	Greater Than
* Expected Value :	0
Description :	

Parameter	Description
Rule Name	The name of the rule.

Parameter	Description
Rule Type	Valid values: Rule Type and Soft. ■ If you select Rule Type, error alerts are reported and descendant nodes are blocked, whereas warning alerts are reported but descendant nodes are not blocked. ■ If you select Soft, error alerts are reported but descendant nodes are not blocked, whereas warning alerts are not reported and descendant nodes are not blocked.
Field	In this example, select **All Fields in Table**. If you select All Fields in Table, you can use the WHERE clause to customize filter conditions based on business requirements.
Sampling Method	Valid values: **count** and **count/table_count**.
Filter	The filter condition. For example, if you want to query the partitions of the table based on a specific data timestamp, you can specify `pt=$[yyyymmdd-1]` as the filter condition.
Check type	Valid values: **Numeric type**, **Fluctuation**, and **Auto-Generated Threshold**. **Note** You can select **Auto-Generated Threshold** only in DataWorks Enterprise Edition or more advanced editions.
Comparison Method	The comparison methods that can be selected vary based on the threshold type. ■ If you set the **Check type** parameter to **Numeric type**, the valid values of the Comparison Method parameter are **Greater Than**, **Greater Than or Equal To**, **Equal To**, **Unequal To**, **Less Than**, and **Less Than or Equal To**. ■ If you set the **Check type** parameter to **Fluctuation**, the valid values of the Comparison Method parameter are **Absolute Value**, **Raise**, and **Drop**.

Parameter	Description
Verification Method	The verification methods that can be selected vary based on the threshold type. - If you set the **Check type** parameter to **Numeric type**, you can set the Verification Method parameter only to **Compare with a specified value**. - If you set the **Check type** parameter to **Fluctuation**, the valid values of the Verification Method parameter are **Compare the current value with the average value of the last 7 days**, **Compare the current value with the average value of the last 30 days**, **Compare the current value with the value 1 day before**, **Compare the current value with the value 7 days before**, **Compare the current value with the value 30 days before**, **The variance between the current value and the value 7 days before**, **The variance between the current value and the value 30 days before**, **Compare with the value 1, 7, and 30 days before**, and **Compare with the value of the previous cycle**.
Expected Value	If you set the **Check type** parameter to **Numeric type**, you must specify an expected value.
Thresholds	If you set the **Check type** parameter to **Fluctuation**, you must specify a warning threshold and an error threshold for the fluctuation. You can enter thresholds or adjust the slider to specify thresholds.
Description	The description of the rule.

- Select **SQL Statement**.



* Rule Name : Enter a rule name.

* Rule Type : ○ Rule Type ● Soft

* Field : SQL Statement(user_defined)

* Sampling Method : SQL Statement

Set Flag : Enter the SET clause of the SQL statement. \r\n Note: Enter the content that follows the SET keyword. Separate multiple statements with commas (,) and do not add a semicolon

* Custom SQL : Enter a custom SQL statement.

* Check type : Numeric type

* Verification Method : Compare with a specified value

* Comparison Method : Greater Than

* Expected Value : 0

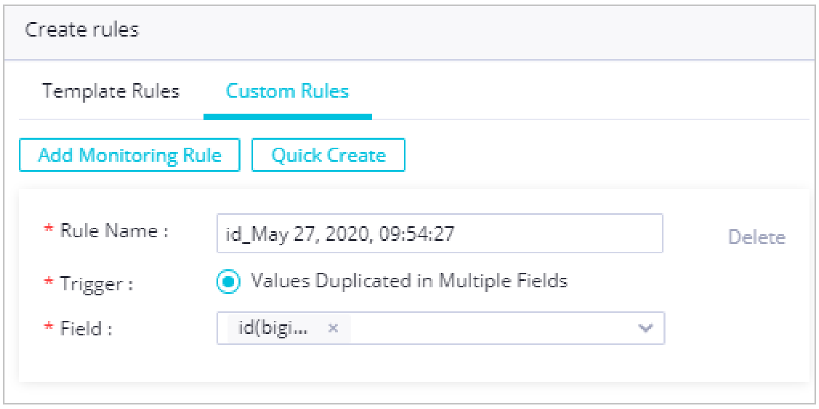
Description :

Parameter	Description
Rule Name	The name of the rule.

Parameter	Description
Rule Type	Valid values: Rule Type and Soft. ■ If you select Rule Type, error alerts are reported and descendant nodes are blocked, whereas warning alerts are reported but descendant nodes are not blocked. ■ If you select Soft, error alerts are reported but descendant nodes are not blocked, whereas warning alerts are not reported and descendant nodes are not blocked.
Field	If you select **SQL Statement**, you can customize the SQL logic. The return value is the value in a row of a column.
Sampling Method	You can set this parameter only to **SQL Statement**.
Set Flag	The SET clause of the SQL statement to be used.
Custom SQL	The SQL statement to be used. You can specify only an SQL statement that returns the value in a row of a column. In the SQL statement, enclose the partition filter expression in brackets [].
Check type	Valid values: **Numeric type** and **Fluctuation**.
Comparison Method	The comparison methods that can be selected vary based on the threshold type. ■ If you set the **Check type** parameter to **Numeric type**, the valid values of the Comparison Method parameter are **Greater Than**, **Greater Than or Equal To**, **Equal To**, **Unequal To**, **Less Than**, and **Less Than or Equal To**. ■ If you set the **Check type** parameter to **Fluctuation**, the valid values of the Comparison Method parameter are **Absolute Value**, **Raise**, and **Drop**.
Verification Method	The verification methods that can be selected vary based on the threshold type. ■ If you set the **Check type** parameter to **Numeric type**, you can set the Verification Method parameter only to **Compare with a specified value**. ■ If you set the **Check type** parameter to **Fluctuation**, the valid values of the Verification Method parameter are **Compare the current value with the average value of the last 7 days**, **Compare the current value with the average value of the last 30 days**, **Compare the current value with the value 1 day before**, **Compare the current value with the value 7 days before**, **Compare the current value with the value 30 days before**, **The variance between the current value and the value 7 days before**, **The variance between the current value and the value 30 days before**, **Compare with the value 1, 7, and 30 days before**, and **Compare with the value of the previous cycle**.

Parameter	Description
Expected Value	If you set the **Check type** parameter to **Numeric type**, you must specify an expected value.
Thresholds	If you set the **Check type** parameter to **Fluctuation**, you must specify a warning threshold and an error threshold for the fluctuation. You can enter thresholds or adjust the slider to specify thresholds.
Description	The description of the rule.

- Quick Create



* Rule Name : id_May 27, 2020, 09:54:27

* Trigger : ● Values Duplicated in Multiple Fields

* Field : id(bigi... ×

Parameter	Description
Rule Name	The name of the rule.
Trigger	You can select only **Values Duplicated in Multiple Fields**.
Field	The fields to be monitored.

4. Click **Batch Create**.

2.5. View monitoring results

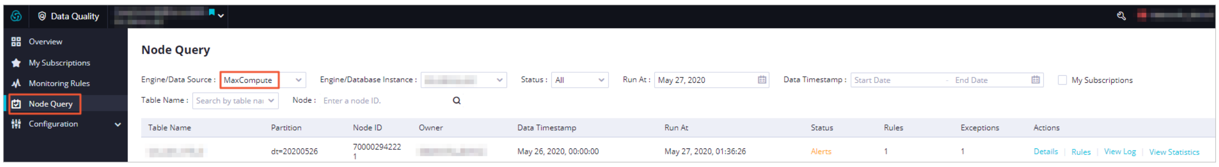
The Node Query page displays the monitoring results of tables and topics based on monitoring rules. After monitoring rules are triggered, you can go to the Node Query page to view monitoring results.

Go to the Node Query page

1. Log on to the DataWorks console.
2. In the left-side navigation pane, click **Workspaces**.
3. After you select the region where the required workspace resides, find the workspace and click **Data Analytics**.
4. Click the ☰ icon in the upper-left corner and choose **All Products** > **Data governance** > **Data Quality**.
5. On the Data Quality page, click **Node Query** in the left-side navigation pane.On the **Node Query**

page, you can set parameters such as **Engine/Data Source**, **Status**, and **My Subscriptions**. You can set the parameters to filter nodes and view the monitoring results of E-MapReduce tables, Hologres tables, AnalyticDB for PostgreSQL tables, MaxCompute tables, and DataHub topics.

View the monitoring results of E-MapReduce, Hologres, AnalyticDB for PostgreSQL, and MaxCompute tables



Node Query

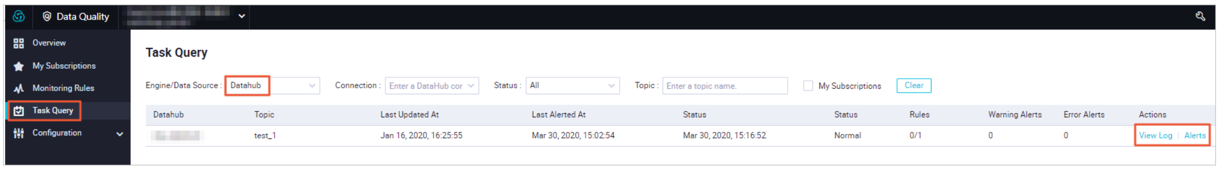
Engine/Data Source : MaxCompute
Engine/Database Instance : [redacted]
Status : All
Run At : May 27, 2020
Data Timestamp : Start Date - End Date
□ My Subscriptions
Table Name : Search by table nai
Node : Enter a node ID.

Table Name	Partition	Node ID	Owner	Data Timestamp	Run At	Status	Rules	Exceptions	Actions
[redacted]	dt=20200526	700002942221	[redacted]	May 26, 2020, 00:00:00	May 27, 2020, 01:36:26	Alerts	1	1	Details | Rules | View Log | View Statistics

GUI element	Description
Engine/Data Source	The name of the compute engine. In this example, select **EMR**, **Hologres**, **AnalyticDB for PostgreSQL**, or **MaxCompute**.
Engine/Database Instance	The name of the E-MapReduce instance, Hologres instance, AnalyticDB for PostgreSQL instance, or MaxCompute project where the desired tables reside.
Status	The monitoring result of tables. Pay attention to partitions that trigger alerts or blocks.
Data Timestamp	The data timestamp.
My Subscriptions	Specifies whether to display only the monitoring results of tables that you subscribed to.
Run At	The time when monitoring rules were triggered.
Table Name	The name of the table that you want to view.
Node	The node that triggered monitoring rules.
Details	Find a table and click **Details** in the **Actions** column. On the page that appears, you can perform the following operations on each monitoring rule that is configured for the table: • Find a rule and click **View History Check Results** in the Actions column to view the monitoring result history of the rule. • Enter comments on a rule based on the execution status of the rule. Perform the following steps to enter comments on a rule: i. Find the rule and click **Problem Handling** in the Actions column. ii. In the **Problem Handling** dialog box, set the **Handling Method** and **Comments** parameters. iii. Click **OK**. Notice You can use the problem handling feature only in DataWorks Enterprise Edition or more advanced editions. • Find a rule and click **Handling Logs** in the Actions column to view the processing history of the rule.

GUI element	Description
Rules	Find a table and click **Rules** in the Actions column to go to the rule configuration page of the table. On this page, you can view partition filter expressions and monitoring rules that are configured for the table and modify the rules as required. For more information, see Configure monitoring rules for MaxCompute.
View Log	Find a table and click **View Log** in the Actions column to view the operational logs of monitoring rules that are configured for the table.
View Statistics	Find a table and click **View Statistics** in the Actions column to view rule execution information about the table, including the number of rows and the table size.

View the monitoring results of DataHub topics



Task Query

Engine/Data Source : Datahub Connection : Enter a DataHub cor Status : All Topic : Enter a topic name. □ My Subscriptions Clear

Datahub	Topic	Last Updated At	Last Alerted At	Status	Status	Rules	Warning Alerts	Error Alerts	Actions
[redacted]	test_1	Jan 16, 2020, 16:25:55	Mar 30, 2020, 15:02:54	Mar 30, 2020, 15:16:52	Normal	0/1	0	0	View Log | Alerts

GUI element	Description
Engine/Data Source	The name of the compute engine. In this example, select **Datahub**.
Connection	The name of the DataHub connection.
Status	The monitoring result of topics. Pay attention to topics that trigger alerts or blocks.
Topic	The name of the topic that you want to view.
My Subscriptions	Specifies whether to display only the monitoring results of topics that you subscribed to.
Clear	Click **Clear** to clear the filter conditions you specified.
View Log	Find a topic and click **View Log** in the Actions column to view the operational logs of monitoring rules that are configured for the topic.
Alerts	Find a topic and click **Alerts** in the Actions column. On the Alerts page, you can view the details about alerts that are triggered by the topic. On the **Alerts** page, you can find an alert and click **Close** in the **Actions** column. In the message that appears, click **OK** to disable the alert.

2.6. Configuration

2.6.1. Create and manage report templates

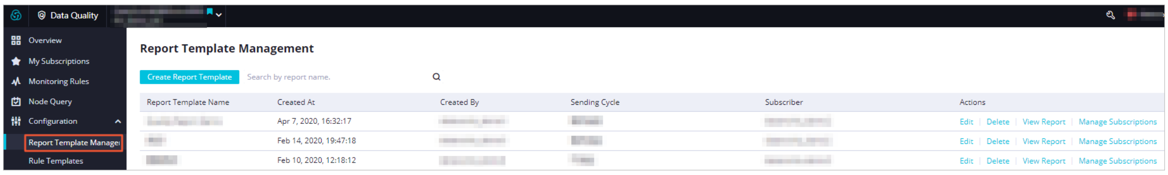
You can dynamically configure report templates on the Report Template Management page. Data Quality can generate and send reports based on a report template as scheduled.

Prerequisites

DataWorks Enterprise Edition or a more advanced edition is activated so that you can use the report template management feature.

Create a report template

1. Log on to the DataWorks console.
2. In the left-side navigation pane, click **Workspaces**.
3. After you select the region where the required workspace resides, find the workspace and click **Data Analytics**.
4. Click the ☰ icon in the upper-left corner and choose **All Products** > **Data governance** > **Data Quality**.
5. In the left-side navigation pane, choose **Configuration** > **Report Template Management**.
6. Click **Create Report Template**.
7. On the **Create Report Template** page, set the parameters.



Report Template Management

Report Template Name	Created At	Created By	Sending Cycle	Subscriber	Actions
[redacted]	Apr 7, 2020, 16:32:17	[redacted]	[redacted]	[redacted]	Edit | Delete | View Report | Manage Subscriptions
[redacted]	Feb 14, 2020, 19:47:18	[redacted]	[redacted]	[redacted]	Edit | Delete | View Report | Manage Subscriptions
[redacted]	Feb 10, 2020, 12:18:12	[redacted]	[redacted]	[redacted]	Edit | Delete | View Report | Manage Subscriptions

Section	Parameter	Description
Basic Settings	**Name**	Enter a name for the report template.
	Sending Cycle	Valid values: **Every Day**, **Every Week**, **Every Month**, and **Do Not Send**. If you select **Every Week** or **Every Month**, you must specify the specific time.
	Timespan	Enter the number of days that have elapsed before the current day. The maximum value of this parameter is 30.
Statistics of Rule Configuration The metrics in the Statistics of Rule Configuration section include offline metrics and real-time metrics. You can select metrics as needed.	**Offline data**	Offline monitoring coverage indicates the ratio of tables that are configured with quality rules to all tables in the workspace. The metrics include **Table count**, **Partition expression count**, **Count of rule on offline data**, and **Rule coverage on tables**.
	Realtime data	Real-time monitoring coverage indicates the ratio of topics that are configured with quality rules to all topics in the workspace. The metrics include **Topic count**, **Count of rule on realtime data**, **Count of rule on cut off data**, **Rule coverage on topic**, **Count of rule on delayed data**, and **Count of customized rule**.

Section	Parameter	Description
Statistics of Rule Execution The metrics in the Statistics of Rule Execution section include offline metrics and real-time metrics. You can select metrics as needed. Quality reports display the selected metrics in charts.	**Offline data**	The metrics are classified into the following types: **About rules**, **About partitions**, and **About tables**.
	Realtime data	The metrics are classified into the following types: **About messages**, **About alarms**, and **About cut-offs**.
Subscriptions	**Notification Method**	DataWorks sends report notifications to subscribers by using emails.
	Recipient	Select the recipient of report notifications. You can add multiple recipients.
	Actions	You can modify or delete subscriptions that you have added.
	Add Subscription	Click **Add Subscription** and configure a subscription.

8. Click **Save** in the upper-right corner. A template of data quality reports is generated.

   You can also perform the following operations:

   - Click **Preview** in the upper-right corner to view the display format of the report template.

     > **Note** If report subscribers view reports in emails, they can view the reports only in tables. If report subscribers view reports on the Data Quality page, they can view the reports in tables or charts.

   - Click **Cancel** in the upper-right corner. In the **Confirm** message, click **OK** to cancel the creation of the report template.

Manage a report template

After a report template is created, the **Report Template Management** page appears. On this page, you can view the details of the report template. You can also perform the following operations:

- Find the required report template and click **Edit**. On the **Edit Report Template** page, modify the report template.
- Find the required report template and click **Delete**. In the **Confirm** message, click **OK** to delete the report template.
- Find the required report template and click **View Report**. Set the **Query Range** parameter and view related reports.
- Add, modify, and delete subscriptions.

- Add a subscription.
  1. Find the required report template and click **Manage Subscriptions**.
  2. In the **Subscriptions** dialog box, click **Add Subscription**.
  3. Select the required recipient and click **Save**.
  4. Click **OK**.
- Modify a subscription.
  1. Find the required report template and click **Manage Subscriptions**.
  2. In the **Subscriptions** dialog box, find the required subscription and click **Modify**.
  3. Select the required recipient and click **Save**.
  4. Click **OK**.
- Delete a subscription.
  1. Find the required report template and click **Manage Subscriptions**.
  2. In the **Subscriptions** dialog box, find the required subscription and click **Delete**.

2.6.2. Create, manage, and use rule templates

In Data Quality, you can manage a set of custom rule templates and use the rule templates to improve the efficiency of rule configuration.

Prerequisites

DataWorks Enterprise Edition or a more advanced edition is activated.

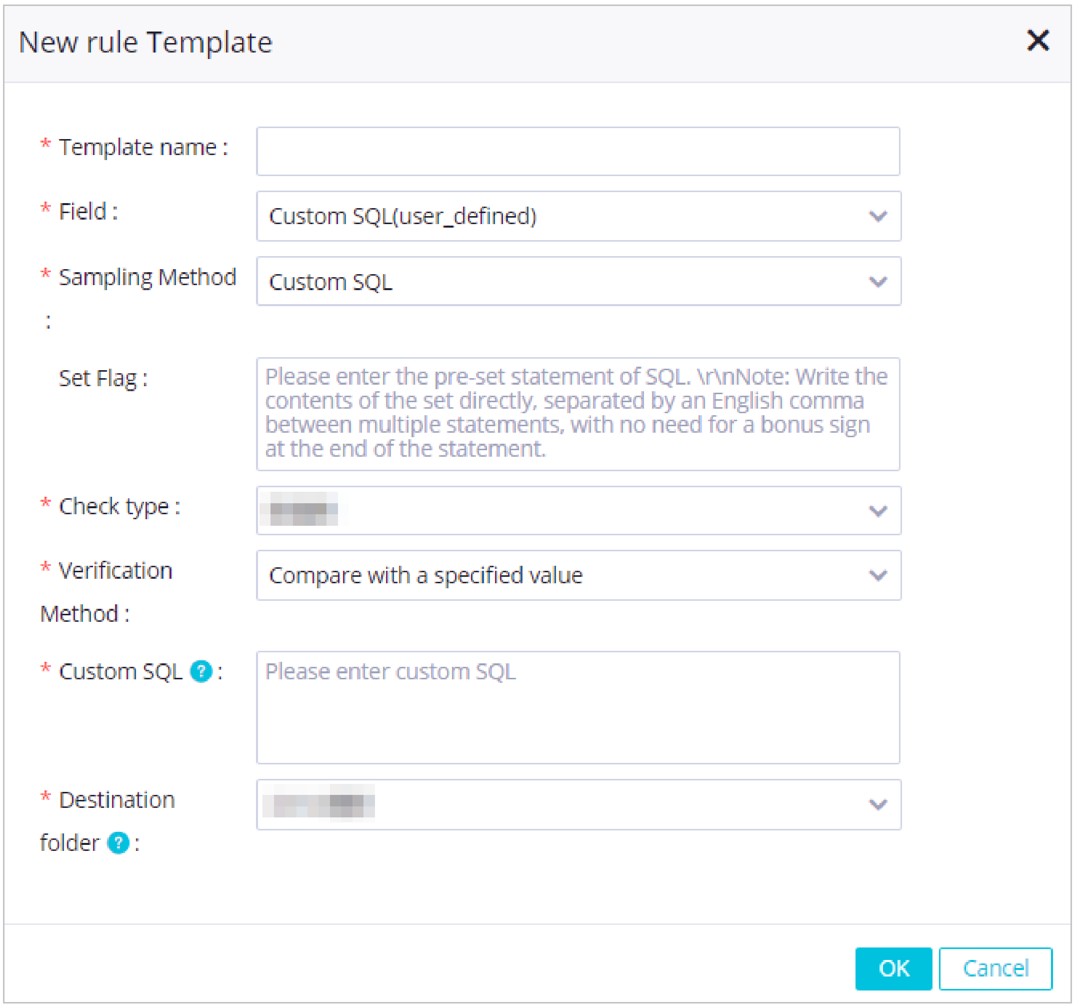
Context

You can create a rule template on the **Rule Templates** and **Monitoring Rules** pages. After the rule template is created, you can manage and use it.

Create a rule template on the Rule Templates page

1. Go to the **Data Quality** 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 rule template and click **Data Analytics** in the Actions column.
   iv. On the DataStudio page, click the ☰ icon in the upper-left corner and choose **All Products > Data governance > Data Quality**.
2. On the Data Quality page, choose **Configuration > Rule Templates** in the left-side navigation pane.
3. On the Rule Templates page, click the 🗋 icon and select **Create Folder**.
4. In the **Create Folder** dialog box, set the **Name** and **Location** parameters and click **OK**.
5. Right-click the created folder and select **Create Rule Template**. You can also rename or delete a folder.

6. In the **Create Rule Template** dialog box, set the parameters as required.



New rule Template

* Template name :

* Field : Custom SQL(user_defined)

* Sampling Method : Custom SQL

Set Flag : Please enter the pre-set statement of SQL. \r\nNote: Write the contents of the set directly, separated by an English comma between multiple statements, with no need for a bonus sign at the end of the statement.

* Check type : [redacted]

* Verification Method : Compare with a specified value

* Custom SQL ? : Please enter custom SQL

* Destination folder ? : [redacted]

Parameter	Description
Template Name	The name of the rule template.
Field	The fields to be monitored. You can set this parameter only to **Custom SQL**.
Sampling Method	The statistical function of the rule. You can set this parameter only to **Custom SQL**.
Set Flag	The `SET` clause of the SQL statement that is used to query the fields to be monitored. **Note** Separate multiple statements with commas (,). You do not need to add a semicolon (;) at the end of each statement.

Parameter	Description
Check type	The threshold type of the rule. Valid values: **Numeric type**, **Fluctuation**, and **Auto-Generated Threshold**. 🔈 **Notice** You can select **Auto-Generated Threshold** only in DataWorks Enterprise Edition or more advanced editions.
Verification Method	The verification method of the rule. The verification methods that can be selected vary based on the threshold type. ○ If you set the **Check type** parameter to **Numeric type**, you can set this parameter only to **Compare with a specified value**. Only the value that is calculated by the COUNT or SUM function can be returned. The return value is compared with a specified value. ○ If you set the **Check type** parameter to **Fluctuation**, the valid values of this parameter are **Compare the current value with the average value of the last 7 days**, **Compare the current value with the average value of the last 30 days**, **Compare the current value with the value 1 day before**, **Compare the current value with the value 7 days before**, **Compare the current value with the value 30 days before**, **The variance between the current value and the value 7 days before**, **The variance between the current value and the value 30 days before**, **Compare with the value 1, 7, and 30 days before**, and **Compare with the value of the previous cycle**. ○ If you set the **Check type** parameter to **Auto-Generated Threshold**, you can set this parameter only to **Dynamic threshold**.
Custom SQL	The SQL statement that is used to query the fields to be monitored. You can use ${tableName} to specify a table name. ⓘ **Note** Make sure that the return value is the value in a row of a column and can be compared with the specified threshold.
Location	The name of the folder in which you want to store the rule template.

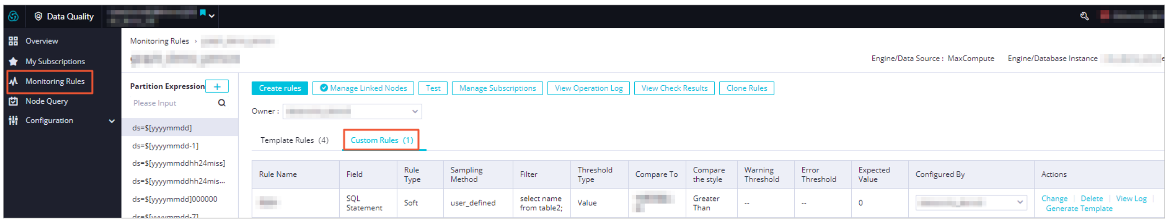
7. Click **OK**.

Create a rule template on the Monitoring Rules page

1. Go to the **Data Quality** page.
2. On the Data Quality page, click **Monitoring Rules** in the left-side navigation pane.
3. On the **Monitoring Rules** page, select a compute engine or data store, find a table or topic, and then click **View Monitoring Rules** in the **Actions** column.

> ⓘ **Note** In this example, a MaxCompute table is used.

4. On the page that appears, select a partition filter expression and click the **Custom Rules** tab.

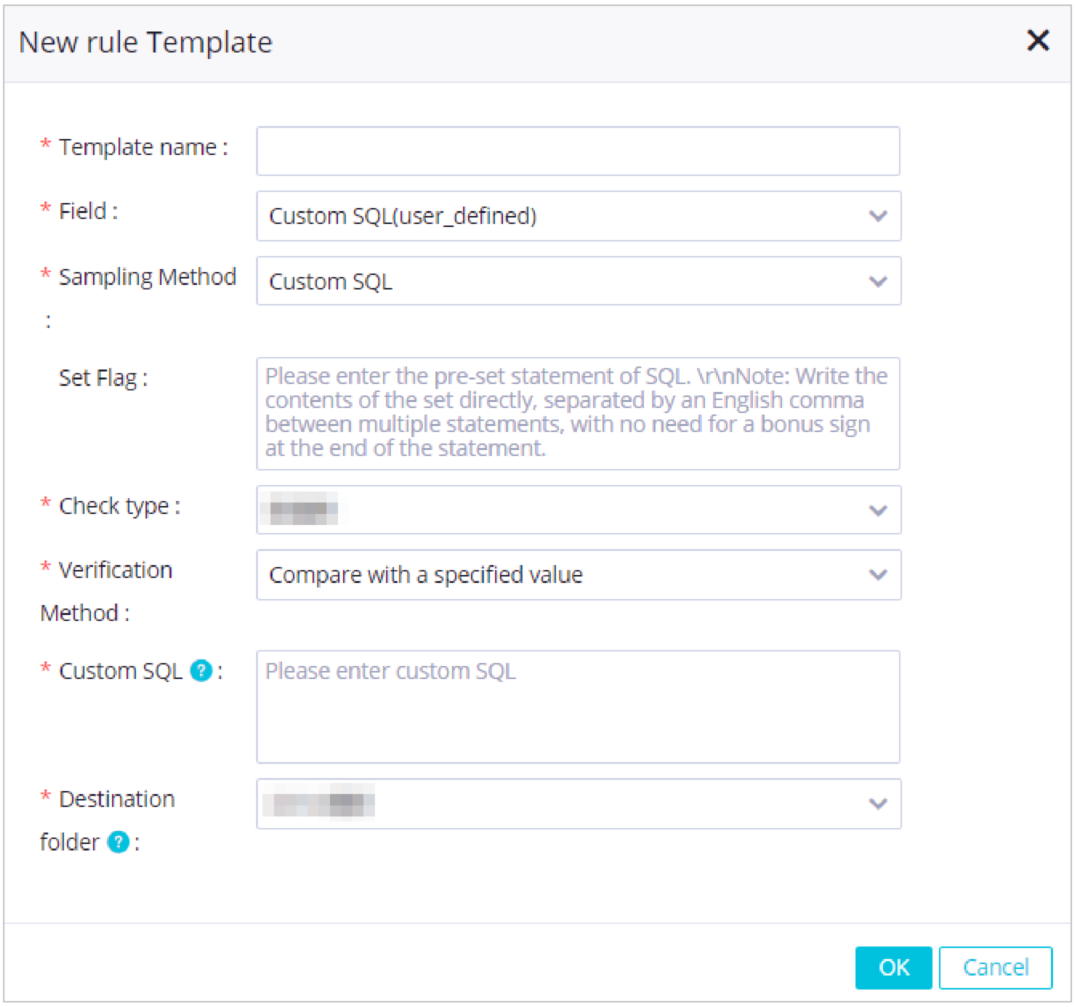


Engine/Data Source : MaxCompute Engine/Database Instance [redacted]

Owner : [redacted]

Rule Name	Field	Rule Type	Sampling Method	Filter	Threshold Type	Compare To	Compare the style	Warning Threshold	Error Threshold	Expected Value	Configured By	Actions
[redacted]	SQL Statement	Soft	user_defined	select name from table2;	Value	[redacted]	Greater Than	--	--	0	[redacted]	Change | Delete | View Log | Generate Template

5. On the Custom Rules tab, find the custom rule based on which you want to create a rule template and click **Generate Template** in the Actions column.
6. In the **Create Rule Template** dialog box, set the parameters as required.


New rule Template

* Template name :	
* Field :	Custom SQL(user_defined)
* Sampling Method :	Custom SQL
Set Flag :	Please enter the pre-set statement of SQL. \r\nNote: Write the contents of the set directly, separated by an English comma between multiple statements, with no need for a bonus sign at the end of the statement.
* Check type :	[redacted]
* Verification Method :	Compare with a specified value
* Custom SQL ? :	Please enter custom SQL
* Destination folder ? :	[redacted]

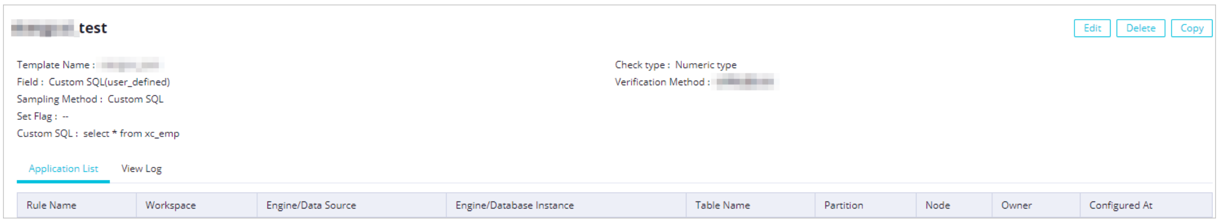
Parameter	Description
Template Name	The name of the rule template.
Field	The fields to be monitored. You can set this parameter only to **Custom SQL**.
Sampling Method	The statistical function of the rule. You can set this parameter only to **Custom SQL**.

Parameter	Description
Set Flag	The `SET` clause of the SQL statement that is used to query the fields to be monitored. **Note** Separate multiple statements with commas (,). You do not need to add a semicolon (;) at the end of each statement.
Check type	The threshold type of the rule. Valid values: **Numeric type**, **Fluctuation**, and **Auto-Generated Threshold**. **Notice** You can select **Auto-Generated Threshold** only in DataWorks Enterprise Edition or more advanced editions.
Verification Method	The verification method of the rule. The verification methods that can be selected vary based on the threshold type. ○ If you set the **Check type** parameter to **Numeric type**, you can set this parameter only to **Compare with a specified value**. Only the value that is calculated by the COUNT or SUM function can be returned. The return value is compared with a specified value. ○ If you set the **Check type** parameter to **Fluctuation**, the valid values of this parameter are **Compare the current value with the average value of the last 7 days**, **Compare the current value with the average value of the last 30 days**, **Compare the current value with the value 1 day before**, **Compare the current value with the value 7 days before**, **Compare the current value with the value 30 days before**, **The variance between the current value and the value 7 days before**, **The variance between the current value and the value 30 days before**, **Compare with the value 1, 7, and 30 days before**, and **Compare with the value of the previous cycle**. ○ If you set the **Check type** parameter to **Auto-Generated Threshold**, you can set this parameter only to **Dynamic threshold**.
Custom SQL	The SQL statement that is used to query the fields to be monitored. You can use ${tableName} to specify a table name. **Note** Make sure that the return value is the value in a row of a column and can be compared with the specified threshold.
Location	The name of the folder in which you want to store the rule template.

7. Click **OK**.
8. In the left-side navigation pane, choose **Configuration** > **Rule Templates** to view the created rule template.

Manage a rule template

On the Rule Templates page, you can click the name of a rule template to go to the template details page. On this page, you can view, edit, delete, or copy the rule template.



[redacted] test

Template Name : [redacted]
Field : Custom SQL(user_defined)
Sampling Method : Custom SQL
Set Flag : --
Custom SQL : select * from xc_emp

Check type : Numeric type
Verification Method : [redacted]

Rule Name	Workspace	Engine/Data Source	Engine/Database Instance	Table Name	Partition	Node	Owner	Configured At

Action	Description
View	You can view the parameter settings and logs of the rule template and the rules that use the rule template. ● The **Application List** tab displays the rules that use the rule template. ● The **View Log** tab displays the operations logs of the rule template, including the user who performed each operation, the time when each operation was performed, and the operation details.
Edit	Click **Edit** in the upper-right corner. In the **Edit Rule Template** dialog box, modify the parameters as required and click **OK**.
Delete	Click **Delete** in the upper-right corner. In the **Delete Template** message, click **OK**.
Copy	Click **Copy** in the upper-right corner. In the **Clone Rule Template** dialog box, set the **Template Name** and **Location** parameters and click **OK**.

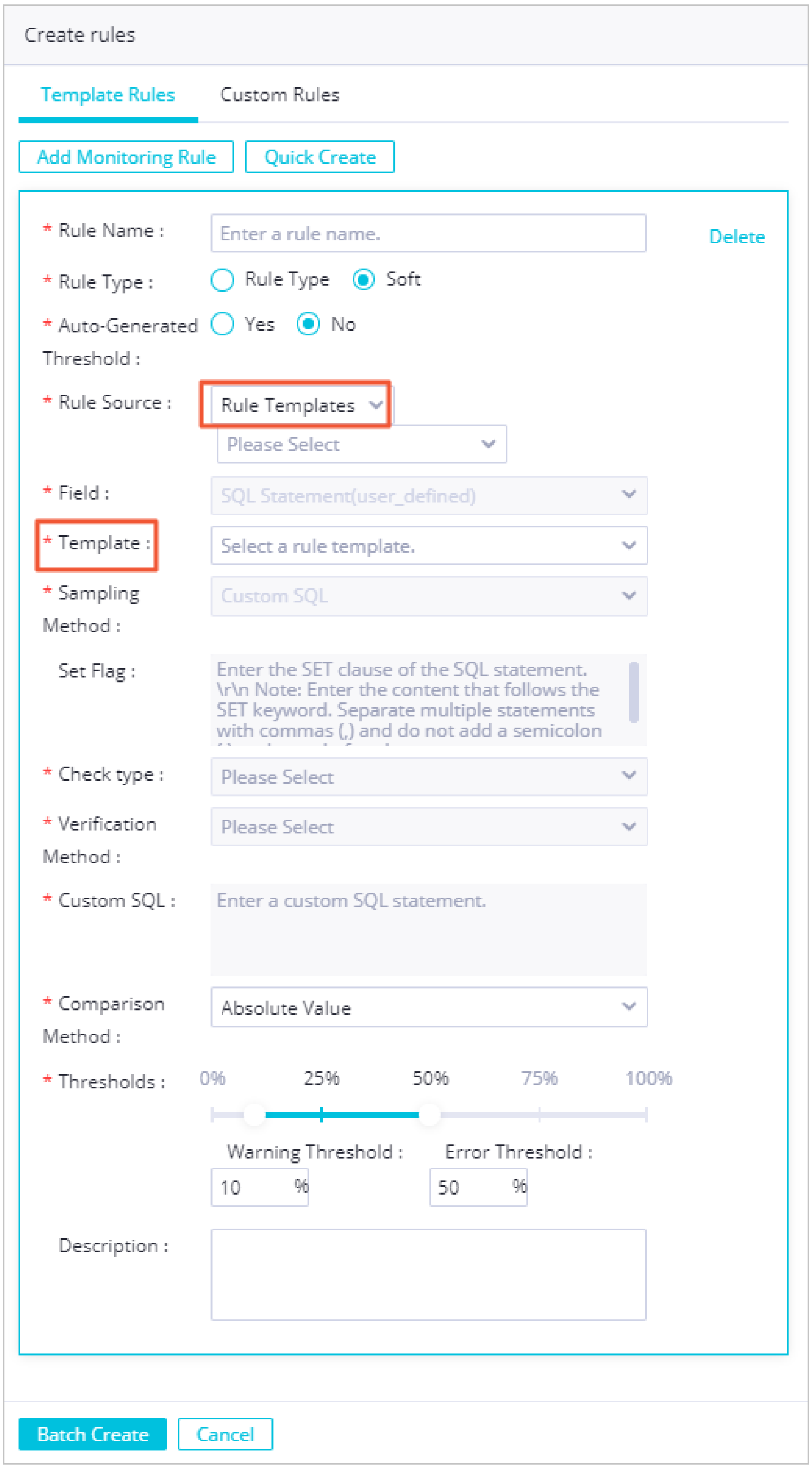
Use a rule template

When you create a monitoring rule, you can select a custom rule template to create the rule based on the rule template.

1. Go to the **Data Quality** page.
2. On the Data Quality page, click **Monitoring Rules** in the left-side navigation pane.
3. On the **Monitoring Rules** page, select a compute engine or data store, find a table or topic, and then click **View Monitoring Rules** in the **Actions** column.

   > **Note** In this example, a MaxCompute table is used.

4. On the page that appears, select a partition filter expression and click **Create rules**.
5. On the **Template Rules** tab of the **Create rules** panel, click **Add Monitoring Rule**.
6. Set the parameters for the rule. In this example, set the **Rule Source** parameter to **Rule Templates** and select a rule template. For more information about the parameter description, see Configure monitoring rules.



* Rule Name :	Enter a rule name.
* Rule Type :	○ Rule Type ● Soft
* Auto-Generated Threshold :	○ Yes ● No
* Rule Source :	Rule Templates Please Select
* Field :	SQL Statement(user_defined)
* Template :	Select a rule template.
* Sampling Method :	Custom SQL
Set Flag :	Enter the SET clause of the SQL statement. \r\n Note: Enter the content that follows the SET keyword. Separate multiple statements with commas (,) and do not add a semicolon
* Check type :	Please Select
* Verification Method :	Please Select
* Custom SQL :	Enter a custom SQL statement.
* Comparison Method :	Absolute Value
* Thresholds :	0% 25% 50% 75% 100% Warning Threshold : 10 % Error Threshold : 50 %
Description :	

7. Click **Batch Create**.

2.7. Instructions

2.7.1. Configure monitoring rules for DataHub

The Monitoring Rules page is the most important part of Data Quality. On this page, you can configure rules to monitor data in E-MapReduce, Hologres, AnalyticDB for PostgreSQL, MaxCompute, and DataHub. This topic describes how to configure monitoring rules for DataHub.

Context

DataHub monitoring supports the following features:

- Templates for monitoring stream discontinuity and data latency
- Stream processing features, such as custom Flink SQL, dimension table JOIN, multi-stream JOIN, and window functions

Procedure

1. Create a DataHub connection.
    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 where your workspace resides. Find the workspace and click **Data Integration** in the Actions column.
    iv. On the Data Integration page, click **Connection** in the left-side navigation pane. The **Data Source** page appeas.
    v. Click **New data source** in the upper-right corner. In the Add data source dialog box, set the parameters as required to create a DataHub connection. For more information, see Configure a DataHub connection.
2. Select the DataHub connection.
    i. On the current page, click the ≡ icon in the upper-left corner and choose **All Products > Data governance > Data Quality**.
    ii. On the Data Quality page, click **Monitoring Rules** in the left-side navigation pane.

iii. On the Monitoring Rules page, select **Datahub** from the **Engine/Data Source** drop-down list and select the DataHub connection. All the topics in the selected DataHub data store are displayed.

GUI element	Description
Configure Flink/SLS Resources	After you create a connection, click Configure Flink/SLS Resources to configure Realtime Compute and Log Service resources related to the connection.
Topics	The Topics tab lists all topics in the DataHub data store. You can click the following buttons in the Actions column for a topic: ■ **View Monitoring Rules**: Click it to create rules for the topic. You can create template rules and custom rules as needed. ■ **Manage Subscriptions**: Click it to view and modify subscribers to the topic, and change the notification method. You can use a DingTalk chatbot to receive notifications. The changed notification method takes effect for all subscribers to the topic.
Dimension Tables	When you create custom rules for a topic, you can create dimension tables and use the JOIN clause to join dimension tables. If the collected data streams lack some fields for a dimension table, you must supplement fields to data streams before data analysis and declare the dimension table in Data Quality. DataHub supports the dimension tables of ApsaraDB for HBase, Lindorm, ApsaraDB RDS, Tablestore, Taobao Distributed Data Layer (TDDL), and MaxCompute. Flink SQL does not design the data definition language (DDL) syntax for dimension tables. You can use the standard CREATE TABLE statement. However, you must add `period for system_time` to specify the period of a dimension table and declare that the dimension table stores time-varying data. ⓘ **Note** When you declare a dimension table, you must specify the primary key. When you join a dimension table with another table, the ON condition must contain an equivalence condition that includes the primary key of either table.

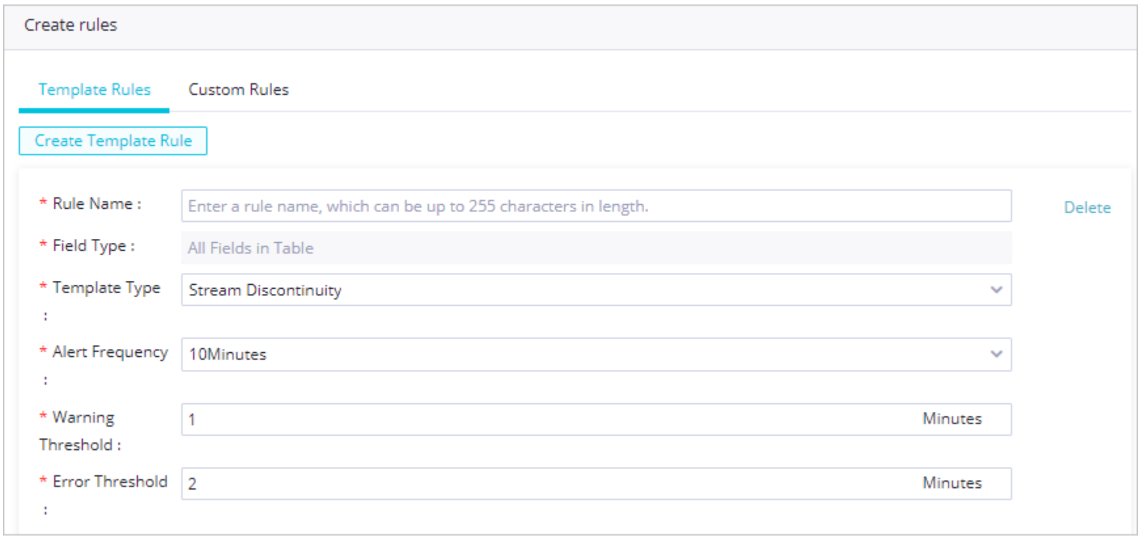
iv. Click the **Topics** tab. Find the topic for which you want to configure monitoring rules and click **View Monitoring Rules** in the Actions column.

3. On the rule configuration page of the topic, click **Create Rule**.

4. Create a monitoring rule.In Data Quality, you can create template rules and custom rules as needed.

   ○ On the Template Rules tab of the Create rules panel, click **Create Template Rule**. Two templates are available: **Data Delay** and **Stream Discontinuity**.

For example, you can select **Data Delay** for the Template Type parameter.



Create rules

Template Rules | Custom Rules

Create Template Rule

* Rule Name : Enter a rule name, which can be up to 255 characters in length.

* Field Type : All Fields in Table

* Template Type : Stream Discontinuity

* Alert Frequency : 10Minutes

* Warning Threshold : 1 Minutes

* Error Threshold : 2 Minutes

Parameter	Description
Rule Name	The name of the rule. The name can be up to 255 characters in length.
Field Type	The fields to be monitored. By default, this parameter is set to All Fields in Table.
Template Type	■ **Data Delay**: monitors the interval between the time when data is generated and the time when data is written to DataHub based on the data timestamp field. If the interval exceeds a specified threshold, an alert is generated. **Note** The data timestamp field supports two data types: TIMESTAMP and STRING (yyyy-MM-dd HH:mm:ss). ■ **Stream Discontinuity**: monitors the period during which no data is written to DataHub. If the period exceeds a specified threshold, an alert is generated. Before you configure a stream discontinuity rule, you must activate Realtime Compute and create a project. On the Monitoring Rules page, click **Configure Flink/SLS Resources** in the upper-right corner. In the dialog box that appears, specify the Realtime Compute project and click **OK**.
Alerts Threshold	The maximum number of alerts generated for data latency. Data Quality reports an alert when the number of alerts generated for data latency exceeds this threshold. This parameter is displayed only when you select Data Delay for the Template Type parameter.

Parameter	Description
Data Timestamp Field	The data timestamp field of the topic for which the rule is created. This field supports two data types: TIMESTAMP and STRING (yyyy-MM-dd HH:mm:ss). This parameter is displayed only when you select Data Delay for the Template Type parameter.
Alert Frequency	The interval at which alerts are reported. You can set the alert interval to 10 minutes, 30 minutes, 1 hour, or 2 hours.
Warning Threshold	The warning threshold, in seconds. The value must be an integer and less than the error threshold.
Error Threshold	The error threshold, in seconds. The value must be an integer and greater than the warning threshold.

- If template rules do not meet your requirements for monitoring the data quality of DataHub topics, you can create a custom rule. On the Custom Rules tab of the Create rules panel, click **Create Custom Rule**.

> Note
>
> - The field in the SELECT clause must be a column. Make sure that you can compare the field values with the warning threshold and error threshold.
> - The FROM clause must include the current topic and all its columns.

Parameter	Description
Rule Name	The name of the rule. The name must be unique in the topic and can be up to 20 characters in length.

Script	The custom SQL script that can be used to set a rule. The return value of the SELECT clause must be unique. Examples:
	- Use a simple SQL statement.
	`select id as a from zmr_tst02;`
	- Join the topic and a dimension table named test_dim.
	`select e.id as eid`
	`from zmr_test02 as e`
	`join test_dim for system_time as of proctime() as w`
	`on e.id=w.id`
	- Join the topic and another topic named dp1test_zmr01.
	`select count(newtab.biz_date) as aa`
	`from (select o.*`
	`from zmr_test02 as o`
	`join dp1test_zmr01 as p`
	`on o.id=p.id)newtab`
	`group by id.biz_date,biz_date_str,total_price,'timestamp'`

Parameter	Description
Warning Threshold	The warning threshold, in minutes. The value must be an integer and less than the error threshold.
Error Threshold	The error threshold, in minutes. The value must be an integer and greater than the warning threshold.
Minimum Alert Interval	The minimum interval at which alert are reported, in minutes.
Description	The description of the rule.

5. Click **Batch Create**. After rules are created for the topic, you can perform the following operations:
    - **View Log**: Click it to view the operational logs of the rules.
    - **Manage Subscriptions**: Click it to view and modify subscribers to the rules, and change the notification method. The changed notification method takes effect for all subscribers to the rules.

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

      > **Note** Add a DingTalk chatbot and obtain a webhook URL. Then, copy the webhook URL to the Manage Subscriptions dialog box. For more information, see Add a DingTalk chatbot and obtain a webhook URL.

2.7.2. Configure monitoring rules for MaxCompute

The Monitoring Rules page is the most important part of Data Quality. On this page, you can configure rules to monitor data in E-MapReduce, Hologres, AnalyticDB for PostgreSQL, MaxCompute, and DataHub. This topic describes how to configure monitoring rules for MaxCompute.

Create a MaxCompute connection

1. Log on to the DataWorks console.
2. In the left-side navigation pane, click **Workspaces**.
3. On the Workspaces page, find the workspace in which you want to create a connection and click **Data Integration** in the Actions column.
4. On the **Data Integration** page, click **Connection** in the left-side navigation pane. The **Data Source** page appears.
5. Click **New data source** in the upper-right corner. In the Add data source dialog box, set the parameters as required to create a MaxCompute connection. For more information, see Configure a MaxCompute connection.

Select the MaxCompute connection

1. On the current page, click the ☰ icon in the upper-left corner and choose **All Products > Data**

governance > Data Quality.

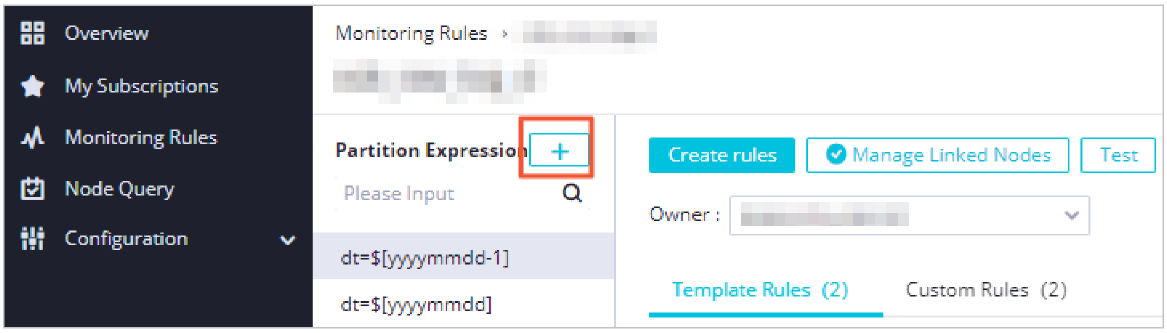
2. On the Data Quality page, click **Monitoring Rules** in the left-side navigation pane.
3. On the Monitoring Rules page, select **MaxCompute** from the **Engine/Data Source** drop-down list and select a MaxCompute project. All the tables in the selected MaxCompute project are displayed.You can search for a table by table name. Fuzzy search based on the initial letters of a table name is supported.
4. Find the table for which you want to configure monitoring rules and click **View Monitoring Rules** in the Actions column.

Configure a partition filter expression

In Data Quality, you must configure rules based on a partition filter expression:

- To configure rules for a non-partitioned table, you can specify NOTAPARTITIONTABLE as the partition filter expression.
- To configure rules for a partitioned table, you can specify a data timestamp expression, such as $[yyyymmdd], or a regular expression as the partition filter expression.

On the rule configuration page of a table, click the **+** icon next to **Partition Expression**.



You can create a partition filter expression or select a recommended partition filter expression.

- Create a partition filter expression

  In the **Add Partition** dialog box, enter a partition filter expression that conforms to the syntax as required. For a non-partitioned table, select **NOTAPARTITIONTABLE** from the recommended partition filter expressions.

  - For a table with only one partition, follow the format: Partition key=Partition value. The partition value can be a constant or a system parameter. You must configure partition filter expressions by using the last partition.
  - For a table with multiple partitions, follow the format: Partition key 1=Partition value/Partition key 2=Partition value/Partition key N=Partition value. Each partition value can be a constant or a system parameter. You must enclose a parameter in brackets [], such as $[yyyymmdd-N].

  The data timestamp that is configured in a partition filter expression also determines the recurrence of the partition filter expression. For example, if the data timestamp is the date of five days ago, the partition filter expression is triggered every five days. The following table describes the supported partition filter expressions.

Partition filter expression	Description
dt=$[yyyymmdd-N]	Indicates N days before.

Partition filter expression	Description
dt=$[yyyymm01-1]	Indicates the first day of each month.
dt=$[yyyymm01-Nm]	Indicates the first day of the month that is N months before the current month.
dt=$[yyyymmld-1]	Indicates the last day of each month.
dt=$[yyyymmld-1m]	Indicates the last day of the month that is N months before the current month.
dt=$[hh24miss-1/24]	Indicates one hour before the hour that is specified by the data timestamp.
dt=$[hh24miss-30/24/60]	Indicates half an hour before the hour that is specified by the data timestamp.
$[yyyymmdd]	Indicates the data timestamp.
$[yyyymmdd-1]	Indicates one day before the data timestamp of the current instance.
$[yyyymmddhh24miss]	Indicates the data timestamp of the current instance. Follow the `yyyymmddhh24miss` format, where: ○ yyyy indicates a four-digit year. ○ mm indicates a two-digit month. ○ dd indicates a two-digit day. ○ hh24 indicates a two-digit hour (24-hour clock). ○ mi indicates two-digit minutes. ○ ss indicates two-digit seconds.
NOTAPARTITIONTABLE	Indicates the partition filter expression of a non-partitioned table.

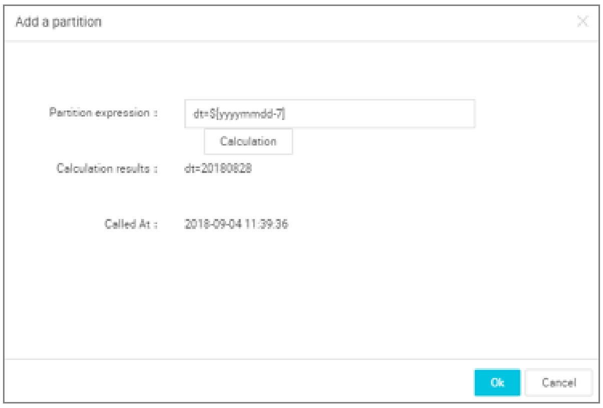
- Select a recommended partition filter expression

  This section describes how to select a recommended partition filter expression. In this example, the dt partition is used. We recommend that you specify a regular expression as the partition filter expression for a dynamic partitioned table.

  i. In the **Add Partition** dialog box, click the Partition Expression field. A drop-down list appears to show you the partition filter expressions that are recommended by Data Quality.

     - Select a recommended partition filter expression that meets your expectation.
     - Specify a custom partition filter expression if no recommended partition filter expressions meet your expectation.

  ii. After you enter a partition filter expression, click **Verify**. Data Quality uses the current time as the data timestamp to calculate data and verify the partition filter expression.



Add a partition

Partition expression : dt=$[yyyymmdd-7]

Calculation results : dt=20180828

Called At : 2018-09-04 11:39:36

iii. Click **OK**.

If you need to delete a configured partition filter expression, move the pointer over the partition filter expression and click the **Delete** icon. After you delete a partition filter expression, all rules that are configured based on the partition filter expression are also deleted.

Link a partition filter expression to a node

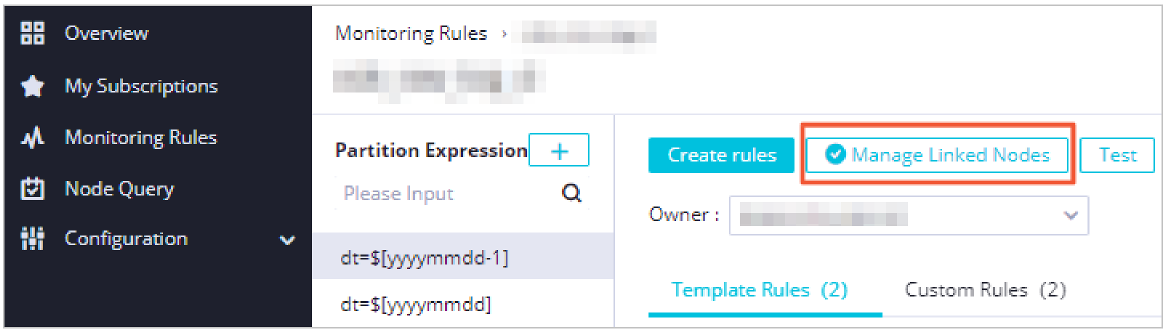
To monitor the quality of data involved in a node, you must link a partition filter expression to the node.

- The Manage Linked Nodes dialog box lists all committed nodes. Data Quality allows you to link a partition filter expression to a node in another workspace.
- Before you link a partition filter expression to a node in another workspace, make sure that you are an administrator, a developer, or an administration expert in the two workspaces.

You can link a partition filter expression to one or more nodes. After nodes are linked, Data Quality can automatically monitor linked nodes.

> **Note** Data Quality allows you to flexibly link a partition filter expression to a node. You can select a node that is not related to your table.

1. On the rule configuration page of a table, click **Manage Linked Nodes**.



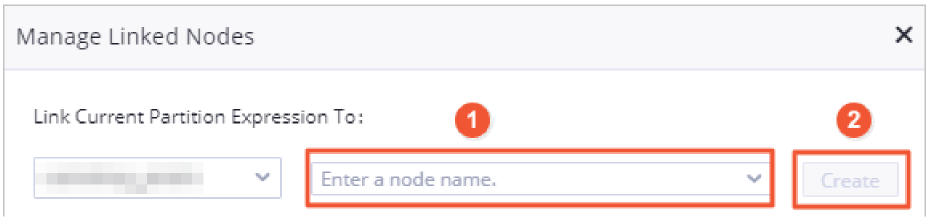
Partition Expression

dt=$[yyyymmdd-1]

dt=$[yyyymmdd]

Owner :

2. In the **Manage Linked Nodes** dialog box, enter the name of the node that you want to link to the partition filter expression.



Manage Linked Nodes

Link Current Partition Expression To:

3. Click **Create**.

Create a rule

The Monitoring Rules page is the most important part of Data Quality, where you can create rules for your tables.

Data Quality allows you to create template rules and custom rules as needed. If you want to create a template rule or a custom rule, you can click **Add Monitoring Rule** or **Quick Create**. For more information, see Configure monitoring rules.

After rules are configured, you can click **Batch Create** to save all the configured rules for the current partition filter expression.

Creation method	Parameter	Description
Add Monitoring Rule	**Rule Name**	The name of the rule.
	Rule Type	The type of the rule. Valid values: • **Rule Type**: If a node reaches the error threshold, Data Quality reports an error alert and determines that the node fails. If a node reaches the warning threshold, Data Quality reports a warning alert and determines that the node is successful. • **Soft**: If a node reaches the error threshold, Data Quality reports an error alert and determines that the node is successful. If a node reaches the warning threshold, Data Quality does not report a warning alert and determines that the node is successful.
	Auto-Generated Threshold	Specifies whether to use dynamic thresholds. You can use the dynamic threshold feature only in DataWorks Enterprise Edition or more advanced editions.
	Rule Source	The source of the rule. Valid values: **Built-in Template** and **Rule Templates**.
	Field	The fields to be monitored. You can select **All Fields in Table** or a specific field. If you select a field, you can apply the rule to the specified field in the table. **Note** In this example, select All Fields in Table and set other parameters for the table-specific rule.

Creation method	Parameter	Description
	Template	• The template that you want to apply to the rule. If you set the **Rule Source** parameter to **Built-in Template**, the built-in table-specific rules are displayed. • If you set the **Rule Source** parameter to **Rule Templates**, you must set parameters such as **Sampling Method** and **Set Flag**. For more information, see Create, manage, and use rule templates.
	Comparison Method	The comparison method of the rule. Valid values: **Absolute Value**, **Raise**, and **Drop**.
	Thresholds	The warning threshold and error threshold of the fluctuation. You can enter thresholds or adjust the slider to specify thresholds.
	Description	The description of the rule.
Quick Create	**Rule Name**	The name of the rule.
	Field	The fields to be monitored. You can select All Fields in Table or a specific field. If you select a field, you can apply the rule to the specified field in the table.
	Trigger	• The trigger condition of the rule. If you select All Fields in Table for the Field parameter, you can set this parameter to **The number of columns is greater than 0** or **Table row number dynamic threshold**. **Notice** You can use the dynamic threshold feature only in DataWorks Enterprise Edition or more advanced editions. • If you select a field for the Field parameter, you can select **The field value already exists**, **Null Field**, **Unique value dynamic threshold**, Summary value dynamic threshold, Average dynamic threshold, Maximum dynamic threshold, or Minimum dynamic threshold. **Notice** You can use the dynamic threshold feature only in DataWorks Enterprise Edition or more advanced editions.

Test rules

After rules are configured for a partition filter expression, you can test all these rules and view the test results.

> ? **Note** You can manually run these rules to test their configurations and notification methods. We recommend that you test rules as required.

1. On the rule configuration page of a table, click **Test**.
2. In the **Test** dialog box, set the **Data Timestamp** parameter.

Parameter	Description
Partition	The partition filter expression for which rules are run. The actual partition key varies based on the data timestamp. For a non-partitioned table, NOPARTITIONTABLE is used as the partition filter expression.
Data Timestamp	The data timestamp for testing rules. The default value is the current time.

3. Click **Test**.
4. In the Test dialog box, click **The test is complete. Click to view the results**. On the **Node Query** page, view the test results. For more information, see View monitoring results.

Manage subscriptions

By default, Data Quality sends notifications to the user who created a partition filter expression. You can add other users so that Data Quality sends notifications to them.

1. On the rule configuration page of a table, click **Manage Subscriptions**.
2. In the **Manage Subscriptions** dialog box, specify the notification method and notification recipient.Data Quality supports the following four methods: **Email**, **Email and SMS**, **DingTalk Chatbot**, and **DingTalk Chatbot @ALL**.

   > ? **Note** Add a DingTalk chatbot and obtain a webhook URL. Then, copy the webhook URL to the Manage Subscriptions dialog box. For more information, see Add a DingTalk chatbot and obtain a webhook URL.

3. Click **Save**.

View operations logs

On the rule configuration page of a table, click **View Operation Log**. In the **Operations Logs** panel, you can view the information about each operation, including the user who performed the operation, the time when the operation was performed, and the operation details.

The **Details** column displays the details of each operation that is performed on the current partition filter expression, including the rule configuration details.

View check results

On the rule configuration page of a table, click **View Check Results** to go to the **Node Query** page. On this page, you can view the check results for all rules under the current partition filter expression.

Clone rules

1. On the rule configuration page of a table, click **Clone Rules**.
2. In the **Clone Rules** dialog box, set the **Target Expression** parameter.
3. Select **Clone Subscribers** and **Change Table Names in Custom Rules** as required.
4. Click **Clone**.

2.7.3. Built-in rule templates for offline data

This topic describes the verification logic of Data Quality and the built-in rule templates that are provided for monitoring offline data.

Terms

- sample: the sample value for the current day. For example, if you want to check the fluctuation of table rows on an SQL node in a day, the sample is the number of table rows on that day.
- baseline: the comparison value from the previous samples.
  - If you want to check the fluctuation of table rows on an SQL node in a day, the baseline is the number of table rows on the previous day.
  - If you want to check the average fluctuation of table rows on an SQL node in seven days, the baseline is the average number of table rows in the last seven days.

Verification logic

Data Quality supports three verification methods: comparison with a fixed value, comparison with thresholds, and dynamic threshold.

Verification method	Verification logic
Comparison with a fixed value	1. Return the Boolean result based on the verification expression. The following comparison operators are supported: `>`, `<`, `>=`, `<=`, and `!=` 2. If the calculation result is true, the data is considered to be normal. If the calculation result is false, an error alert is reported.
Comparison with thresholds	• If the absolute value of the fluctuation does not exceed the warning threshold, the data is considered to be normal. • If the absolute value of the fluctuation does not meet the condition in Case 1 and does not exceed the error threshold, a warning alert is reported. • If the fluctuation does not meet the condition in Case 2, an error alert is reported.

Verification method	Verification logic
Dynamic threshold	You do not need to set thresholds. The system automatically checks the metrics in real time based on algorithm models. If the value of a metric falls outside a reasonable range, an alert is reported. **Notice** You must purchase DataWorks Enterprise Edition or a more advanced edition to use the dynamic threshold feature.

Description of built-in rule templates for offline data

Template name	Description
Fluctuations of the average value of a field compared with that on the previous day, that of seven days ago, and that of one month ago	Data Quality compares the average value of a field with that on the previous day, that of seven days ago, and that of one month ago to obtain the fluctuations. Then, Data Quality compares the obtained fluctuations with thresholds. If a fluctuation exceeds a threshold, an alert is reported.
Fluctuations of the sum of values in a field compared with that on the previous day, that of seven days ago, and that of one month ago	Data Quality compares the sum of values in a field with that on the previous day, that of seven days ago, and that of one month ago to obtain the fluctuations. Then, Data Quality compares the obtained fluctuations with thresholds. If a fluctuation exceeds a threshold, an alert is reported.
Fluctuations of the minimum value of a field compared with that on the previous day, that of seven days ago, and that of one month ago	Data Quality compares the minimum value of a field with that on the previous day, that of seven days ago, and that of one month ago to obtain the fluctuations. Then, Data Quality compares the obtained fluctuations with thresholds. If a fluctuation exceeds a threshold, an alert is reported.
Fluctuations of the maximum value of a field compared with that on the previous day, that of seven days ago, and that of one month ago	Data Quality compares the maximum value of a field with that on the previous day, that of seven days ago, and that of one month ago to obtain the fluctuations. Then, Data Quality compares the obtained fluctuations with thresholds. If a fluctuation exceeds a threshold, an alert is reported.
Number of unique values in a field	Data Quality compares the number of unique values in a field after deduplication with a fixed value.
Fluctuations of the number of unique values in a field compared with that on the previous day, that of seven days ago, and that of one month ago	Data Quality compares the number of unique values in a field after deduplication with that on the previous day, that of seven days ago, and that of one month ago. This is a comparison with a fixed value.
Fluctuations of the number of table rows compared with that on the previous day, that of seven days ago, and that of one month ago	Data Quality compares the number of table rows with that on the previous day, that of seven days ago, and that of one month ago to obtain the fluctuations.
Number of null values in a field	Data Quality compares the number of null values in a field with a fixed value.

Template name	Description
Ratio of the number of null values in a field to the total number of rows	Data Quality compares the ratio of the number of null values in a field to the total number of rows with a fixed value. ⓘ **Note** The fixed value is a decimal.
Ratio of the number of duplicated values in a field to the total number of rows	Data Quality compares the ratio of the number of duplicated values in a field to the total number of rows with a fixed value.
Number of duplicated values in a field	Data Quality subtracts the number of values in a field after deduplication from the total number of rows to obtain the number of duplicated values in the field. Then, Data Quality compares the number of duplicated values with a fixed value.
Ratio of the number of unique values in a field to the total number of rows	Data Quality compares the ratio of the number of unique values in a field to the total number of rows with a fixed value.
Fluctuation of the average value of a field compared with that on the previous day	Data Quality compares the average value of a field with that on the previous day to obtain the fluctuation. Then, Data Quality compares the obtained fluctuation with thresholds.
Fluctuation of the sum of values in a field compared with that on the previous day	Data Quality compares the sum of values in a field with that on the previous day to obtain the fluctuation. Then, Data Quality compares the obtained fluctuation with thresholds.
Fluctuation of the minimum value of a field compared with that on the previous day	Data Quality compares the minimum value of a field with that on the previous day to obtain the fluctuation. Then, Data Quality compares the obtained fluctuation with thresholds.
Fluctuation of the maximum value of a field compared with that on the previous day	Data Quality compares the maximum value of a field with that on the previous day to obtain the fluctuation. Then, Data Quality compares the obtained fluctuation with thresholds.
Fluctuation of the sum of values in a field compared with that in the last cycle	Data Quality compares the sum of values in a field with that in the last cycle to obtain the fluctuation. Then, Data Quality compares the obtained fluctuation with thresholds. If the fluctuation exceeds a threshold, an alert is reported.
Fluctuation of the minimum value of a field compared with that in the last cycle	Data Quality compares the minimum value of a field with that in the last cycle to obtain the fluctuation. Then, Data Quality compares the obtained fluctuation with thresholds. If the fluctuation exceeds a threshold, an alert is reported.
Fluctuation of the maximum value of a field compared with that in the last cycle	Data Quality compares the maximum value of a field with that in the last cycle to obtain the fluctuation. Then, Data Quality compares the obtained fluctuation with thresholds. If the fluctuation exceeds a threshold, an alert is reported.

Template name	Description
Count of each discrete point for grouping in a field	The count of each discrete point for grouping in a field.
Fluctuations of the count of each discrete point for grouping in a field compared with that on the previous day, that of seven days ago, or that of one month ago	The fluctuations of the count of each discrete point for grouping in a field compared with that on the previous day, that of seven days ago, or that of one month ago.
Total number of discrete points for grouping in a field	The total number of discrete points for grouping in a field.
Fluctuation of the total number of discrete points for grouping in a field compared with that on the previous day	The fluctuation of the total number of discrete points for grouping in a field compared with that on the previous day.
Whether the table size, in bytes, remains unchanged, compared with that in the last cycle	The table size, in bytes, remains unchanged, compared with that in the last cycle.
Whether the table size, in bytes, is changed, compared with that in the last cycle	The table size in bytes is changed, compared with that in the last cycle.
Whether the number of table rows is changed, compared with that in the last cycle	The number of table rows is changed, compared with that in the last cycle.
Whether the number of table rows remains unchanged, compared with that in the last cycle	The number of table rows remains unchanged, compared with that in the last cycle.
Difference between the table size, in bytes, and that in the last cycle	Data Quality compares the table size in bytes with that in the last cycle to obtain the difference.
Difference between the number of table rows and that in the last cycle	Data Quality compares the number of table rows collected on the current day with that in the partition generated in the last cycle to obtain the difference.
Number of table rows	The number of table rows.
Table size, in bytes	The table size, in bytes.
Difference between the number of table rows and that on the previous day	Data Quality compares the number of table rows collected on the current day with that in the partition generated on the previous day to obtain the difference.
Difference between the table size, in bytes, and that on the previous day	Data Quality compares the table size in bytes with that on the previous day to obtain the difference.

Template name	Description
Fluctuation of the table size compared with that on the previous day	This template is used to compare the table size with that on the previous day to obtain the fluctuation. For example, you can set the warning threshold to 5% and the error threshold to 10%. If the fluctuation is greater than 5% and less than or equal to 10%, a warning alert is reported. If the fluctuation is greater than 10%, an error alert is reported.
Fluctuation of the table size compared with that of seven days ago	This template is used to compare the table size with that of seven days ago. For example, you can set the warning threshold to 5% and the error threshold to 10%. If the fluctuation is greater than 5% and less than or equal to 10%, a warning alert is reported. If the fluctuation is greater than 10%, an error alert is reported.
Fluctuation of the table size compared with that of one month ago	This template is used to compare the table size with that of one month ago. For example, you can set the warning threshold to 5% and the error threshold to 10%. If the fluctuation is greater than 5% and less than or equal to 10%, a warning alert is reported. If the fluctuation is greater than 10%, an error alert is reported.
Fluctuation of the number of table rows compared with the average number in the last seven days	The average number of table rows in the last seven days is the baseline.
Fluctuation of the number of table rows compared with the average number in the last 30 days	The average number of table rows in the last 30 days is the baseline.
Fluctuation of the number of table rows compared with that of the previous day	Data Quality compares the number of table rows collected on the current day with that in the partition generated on the previous day to obtain the fluctuation.
Fluctuation of the number of table rows compared with that of seven days ago	Data Quality compares the number of table rows collected on the current day with that in the partition generated seven days ago to obtain the fluctuation.
Fluctuation of the number of table rows compared with that of one month ago	Data Quality compares the number of table rows collected on the current day with that in the partition generated one month ago to obtain the fluctuation.
Fluctuations of the number of table rows compared with that on the previous day, that of seven days ago, that of one month ago, and that on the first day of the current month	Data Quality compares the number of table rows with that on the previous day, that of seven days ago, that of one month ago, and that on the first day of the current month to obtain the fluctuations.

Template name	Description
Fluctuation of the number of table rows compared with that in the last cycle	Data Quality compares the number of table rows collected on the current day with that in the partition generated in the last cycle to obtain the fluctuation.

3.Data security guard

3.1. Overview

The Data Security Guard service protects the security of your data. This service provides features such as data recognition, data activities, data risks, data auditing, and rule change. This topic describes how to activate and use Data Security Guard.

Go to the Data Security Guard page

1. Log on to the DataWorks console.
2. In the left-side navigation pane, click **Workspaces**.
3. After you select the region where the required workspace resides, find the workspace and click **Data Analytics**.
4. Click the ☰ icon in the upper-left corner and choose **All Products > Data governance > Data Security Guard**.
5. Click **Try now** to go to the Data Security Guard page.

   > Note
   >
   > - If you have activated Data Security Guard by using your Alibaba Cloud account, the Data Security Guard homepage appears.
   > - If you have not activated Data Security Guard by using your Alibaba Cloud account, the Data Security Guard activation page appears.

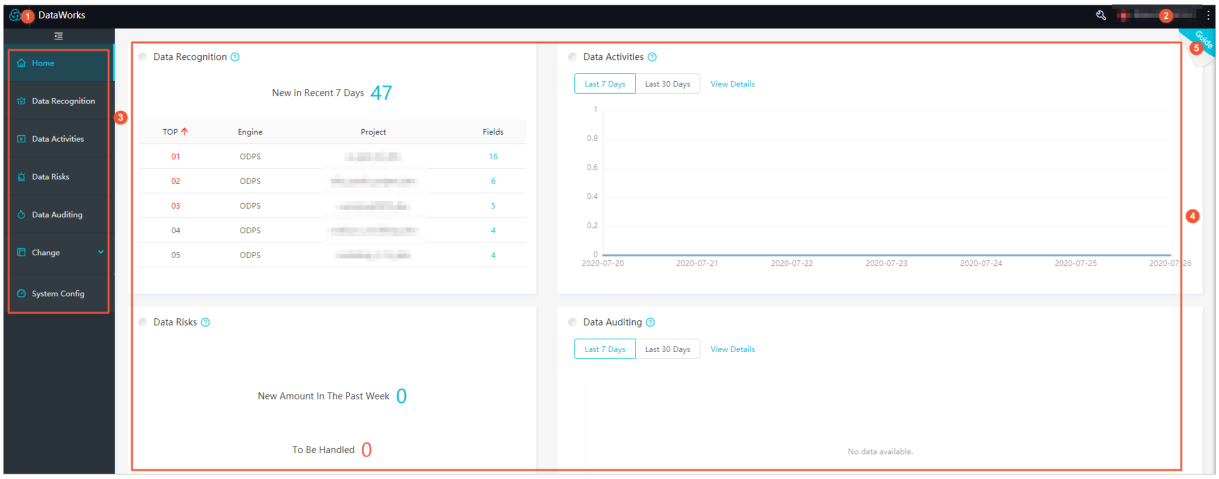
Activate Data Security Guard

Log on with your Alibaba Cloud account. On the **Terms of Service** page, select **I have read and agree to all the preceding terms** and click **Activate**.

> Notice You must use your Alibaba Cloud account to activate Data Security Guard.

Use Data Security Guard

After you activate Data Security Guard, you can use the service.



TOP	Engine	Project	Fields
01	ODPS	[redacted]	16
02	ODPS	[redacted]	6
03	ODPS	[redacted]	5
04	ODPS	[redacted]	4
05	ODPS	[redacted]	4

No.	Name	Description
1	Menu	Provides access to services that you can use, such as **DataStudio**, **Data Integration**, **Operation Center**, and **Data Security Guard**.
2	User information	The logged on user. You can view and modify the user information, including the email address, mobile phone number, AccessKey ID, and AccessKey secret.
3	Navigation pane	The navigation pane for different features of Data Security Guard. For more information, see Discover data, View data activities, Data risk identification, and Set data identification rules.
4	Data Security Guard homepage	- **Data Recognition**: displays the total number of fields that hit the configured data identification rules in the last seven days and sorts the fields by project. - **Data Activities**: displays the number of data access activities that are detected on each day in a specified time range for the fields that hit the configured data identification rules. **Note** The supported time ranges are the last seven days and the last 30 days. You can click **View Details** to go to the **Data Activities** page. - **Data Risks**: displays the number of data risks that are detected in the last seven days and the number of data risks that have not been handled. - **Data Auditing**: displays the number of data risks that are detected and the number of data risks that have been handled in the last seven days and the last 30 days. You can click **View Details** to go to the **Data Auditing** page.
5	Switch to the guide page	Click **Guide** in the upper-right corner to go to the service guide page and view the service information.

3.2. Discover data

After you configure rules, the Data Recognition page helps you identify sensitive data in a workspace.

Prerequisites

Data Security Guard is activated. You must use the Alibaba Cloud account to activate the service. For more information, see Overview.

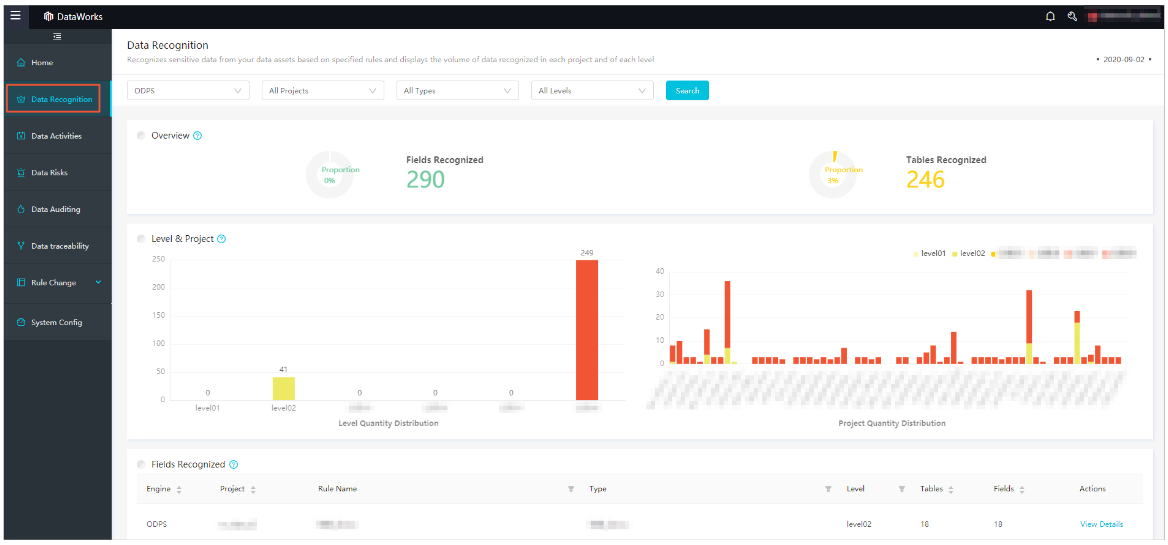
Context

On the next day after you configure data identification rules, you can view the distribution of sensitive data.

> ⓘ **Note** A security administrator can specify the access mode and the users that are permitted to access the Data Recognition page on the **System Config** page.

Procedure

1. Log on to the DataWorks console.
2. In the left-side navigation pane, click **Workspaces**.
3. After you select the region where the required workspace resides, find the workspace and click **Data Analytics**.
4. Click the ☰ icon in the upper-left corner and choose **All Products > Data governance > Data Security Guard**.
5. Click **Try now**. The **Data Recognition** page appears.The Data Recognition page displays data assets from multiple dimensions, such as workspaces and levels. On this page, you can view the total number and proportion of fields and tables that hit data identification rules. You can also view the field list and the number of fields that hit data identification rules at each security level and in each workspace.



Data Recognition

Recognizes sensitive data from your data assets based on specified rules and displays the volume of data recognized in each project and of each level

• 2020-09-02 •

ODPS | All Projects | All Types | All Levels

Fields Recognized 290

Tables Recognized 246

Engine	Project	Rule Name	Type	Level	Tables	Fields	Actions
ODPS	[redacted]	[redacted]	[redacted]	level02	18	18	View Details

3.3. View data activities

The Data Activities page displays the information about sensitive data that is identified based on the configured rules. The information includes the number of data activities that involve sensitive data, trend of data activities, number of exported data entries, and details of data export. The information helps you manage each data activity that involves sensitive data. This page does not display the operational data that is generated based on an E-MapReduce compute engine.

Prerequisites

Data Security Guard is activated for your Alibaba Cloud account. For more information, see Overview.

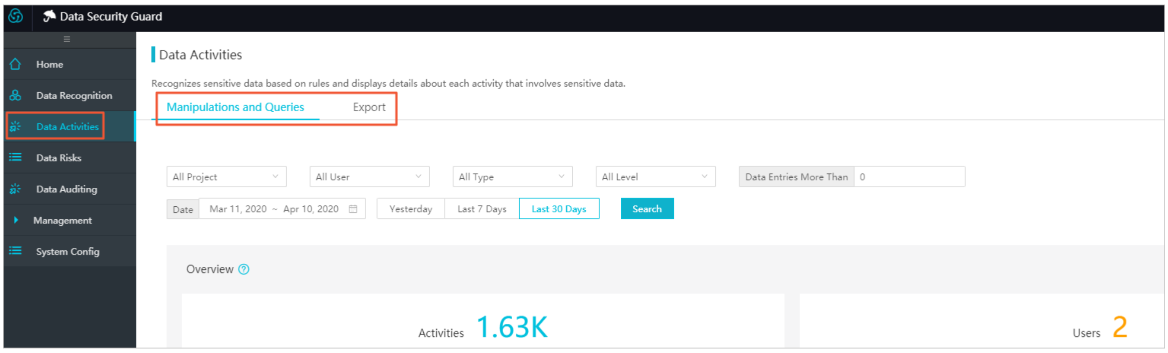
Context

On the next day after you configure data identification rules, you can view the information about data activities and data export.

> **Note** On the **System Config** page, a security administrator can specify the access mode and the users who are permitted to view the information on the **Data Activities** page.

Procedure

1. Log on to the DataWorks console.
2. In the left-side navigation pane, click **Workspaces**.
3. After you select the region where the required workspace resides, find the workspace and click **Data Analytics**.
4. Click the ☰ icon in the upper-left corner and choose **All Products** > **Data governance** > **Data Security Guard**.
5. Click **Try now** to go to the Data Security Guard page.
6. In the left-side navigation pane, click **Data Activities**.



Data Activities

Recognizes sensitive data based on rules and displays details about each activity that involves sensitive data.

The **Data Activities** page consists of the **Manipulations and Queries** and **Export** tabs.

- **Manipulations and Queries**: displays the data activities that involve create and insert operations. Failed data activities are excluded.

  On the next day after you configure data identification rules, you can view the information about data activities on the **Manipulations and Queries** tab. The information includes the overview, trend, and records of data activities.
- **Export**: displays the data activities of exporting data from MaxCompute.

On the next day after you configure data identification rules, you can view the information about data export from MaxCompute on the **Export** tab. The information includes the total number of exported data entries, number of exported data entries per day, and accounts of top 5 exported data entries within the specified time range.

3.4. Data risk identification

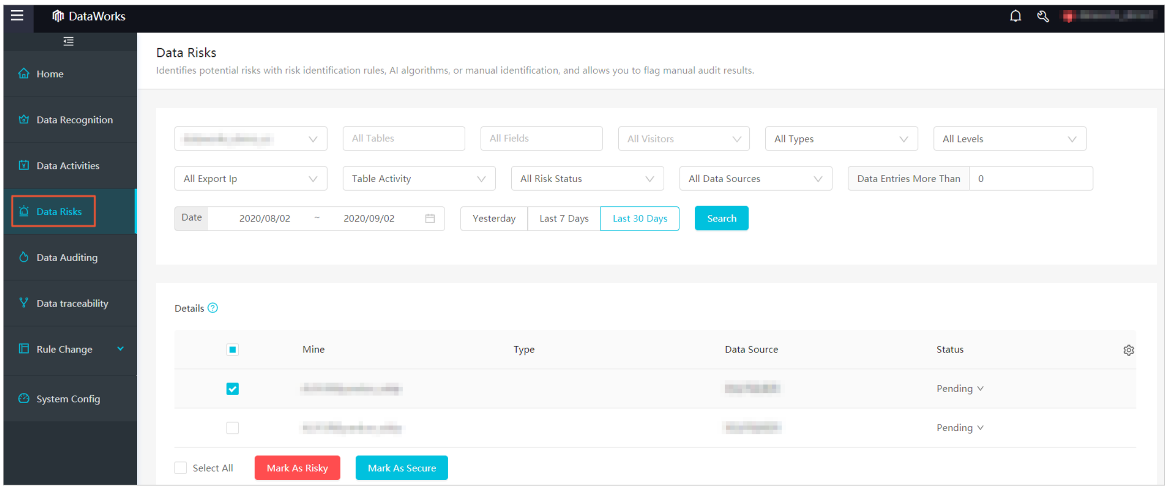
The Data Risks page displays potential risks that are identified by using manual tagging, risk identification rules, or AI algorithms, and allows you to flag manual audit results. This page does not support risk identification for the operation data of E-MapReduce compute engines.

Prerequisites

Data Security Guard is activated by the tenant administrator. For more information, see Overview.

Procedure

1. Log on to the DataWorks console.
2. In the left-side navigation pane, click **Workspaces**.
3. After you select the region where the required workspace resides, find the workspace and click **Data Analytics**.
4. Click the ☰ icon in the upper-left corner and choose **All Products > Data governance > Data Security Guard**.
5. Click **Try now** to go to the Data Security Guard page.
6. In the left-side navigation pane, click **Data Risks**.You can view the details of the data whose risk is identified and perform the **Mark As Risky** or **Mark As Secure** operation.



Data Risks

Identifies potential risks with risk identification rules, AI algorithms, or manual identification, and allows you to flag manual audit results.

All Tables | All Fields | All Visitors | All Types | All Levels

All Export Ip | Table Activity | All Risk Status | All Data Sources | Data Entries More Than 0

Date 2020/08/02 ~ 2020/09/02 | Yesterday | Last 7 Days | Last 30 Days | Search

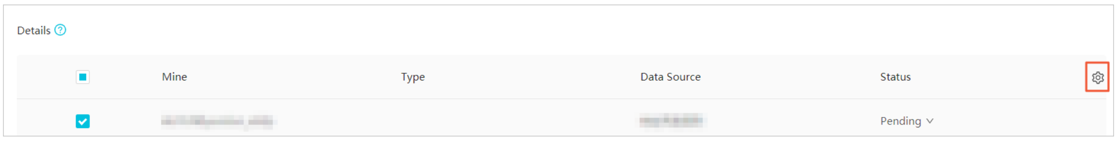
Details

■	Mine	Type	Data Source	Status
☑				Pending
□				Pending

□ Select All | Mark As Risky | Mark As Secure

To update the columns that are displayed in the **Details** section, perform the following steps:

i. In the **Details** section, click the ⚙ icon on the right.



Details

■	Mine	Type	Data Source	Status
☑				Pending

ii. In the **Configure Columns** dialog box, select or clear the columns to be displayed as needed.

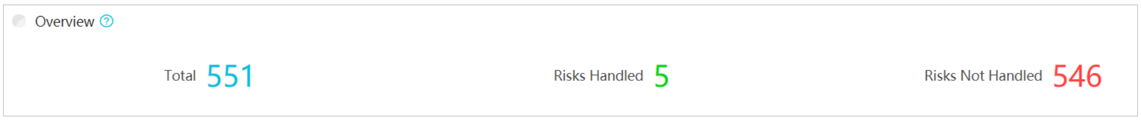
iii. Click **OK**.

3.5. Audit data

The Data Auditing page displays the handling results and distribution of your data risks from multiple dimensions. The data risks of the E-MapReduce compute engine are not displayed.

Procedure

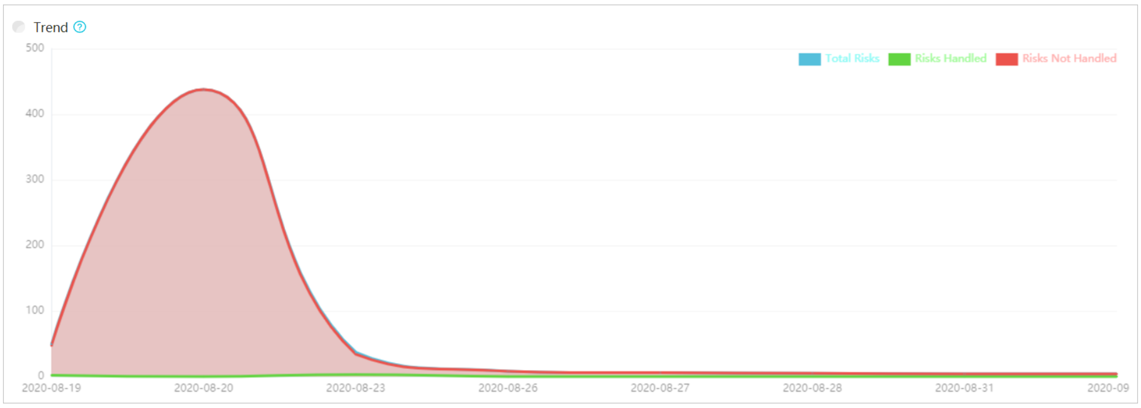
1. Log on to the DataWorks console.
2. In the left-side navigation pane, click **Workspaces**.
3. After you select the region where the required workspace resides, find the workspace and click **Data Analytics**.
4. Click the ☰ icon in the upper-left corner and choose **All Products** > **Data governance** > **Data Security Guard**.
5. Click **Try now** to go to the Data Security Guard page.
6. In the left-side navigation pane, click **Data Auditing**.On this page, you can view a data risk overview, trends in the number of data risks, and risk analysis from multiple dimensions. The statistics are collected yesterday, in the last seven days, and in the last 30 days.
   - **Overview**: displays the total number of data risks, number of data risks that have been handled, and number of data risks that have not been handled within the specified time period.

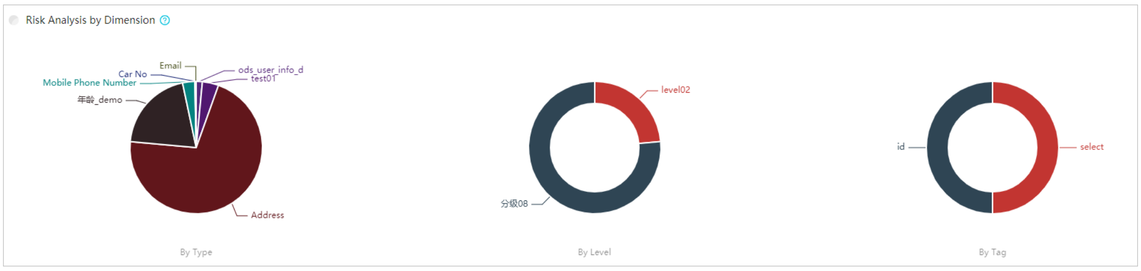
     Overview

     Total 551　Risks Handled 5　Risks Not Handled 546

   - **Trend**: displays the trends in the total number of data risks, number of data risks that have been handled, and number of data risks that have not been handled within the specified time period.

   - **Risk Analysis by Dimension**: displays the percentages of different types of data risks based on the type, level, and tag.

3.6. Set data identification rules

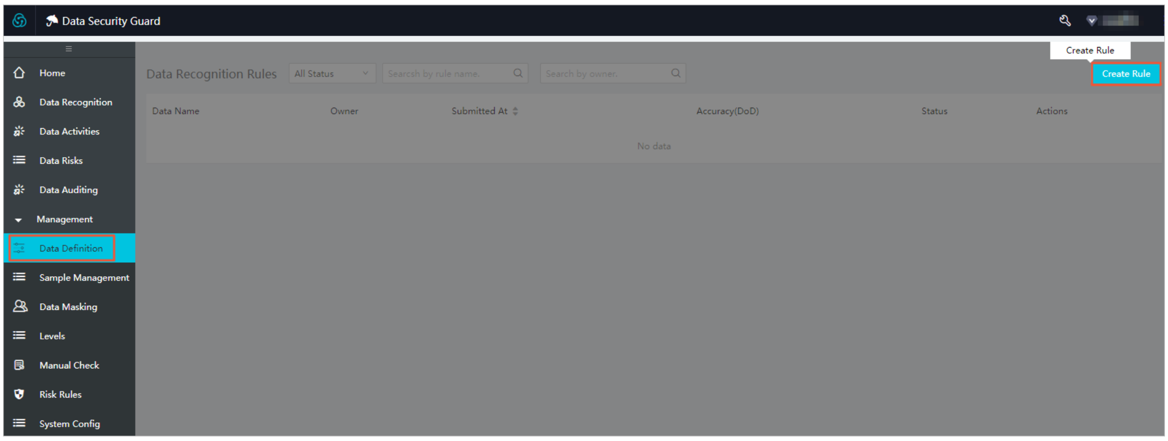
You can create data identification rules to efficiently identify sensitive data under a tenant. This topic describes how to create and set data identification rules.

Go to the Data Recognition Rules page

1. Log on to the DataWorks console. Find the required workspace and click **Data Analytics**.
2. Click the More icon in the upper-left corner and choose **All Products > Data Security Guard**.
3. Click **Try now**. The **Data Recognition** page appears.
4. In the left-side navigation pane, choose **Rule Change > Data Recognition Rules**. On the Data Recognition Rules page, you can create, copy, modify, and delete rules.

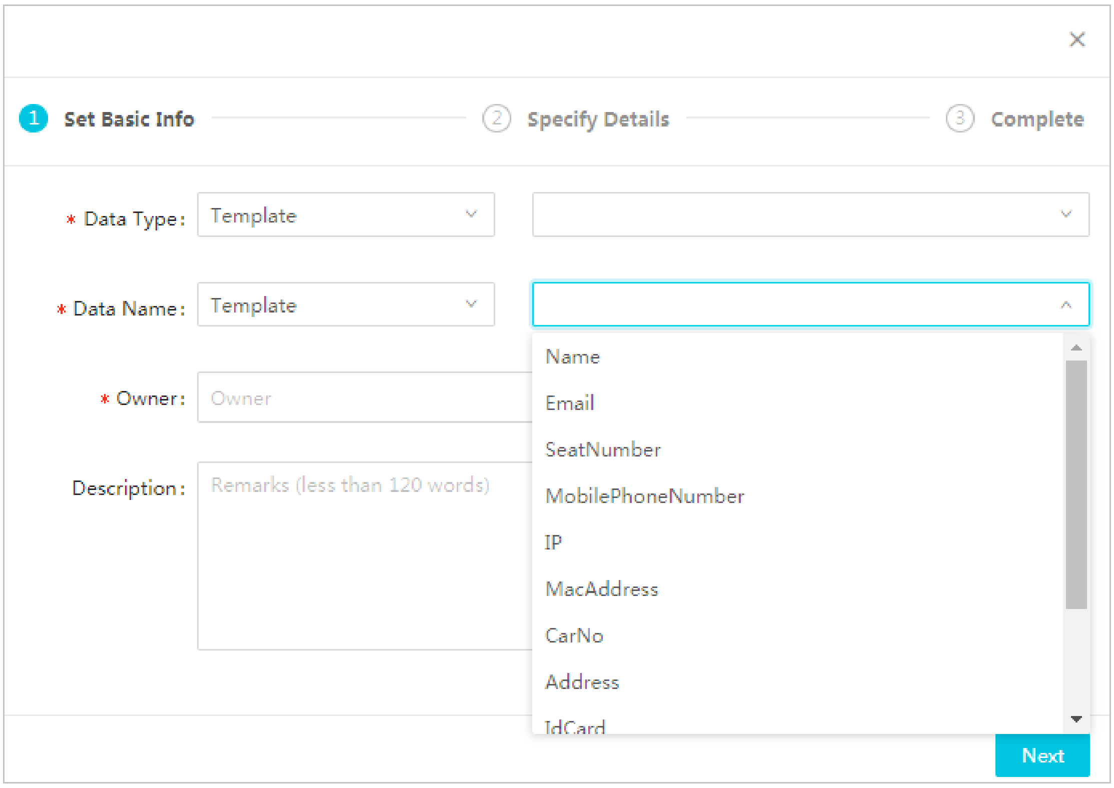
Create a rule

1. Click **Create Rule** in the upper-right corner.



2. In the **Set Basic Info** step, set the parameters as required and click **Next**.

You can create a template-based rule or a custom rule.



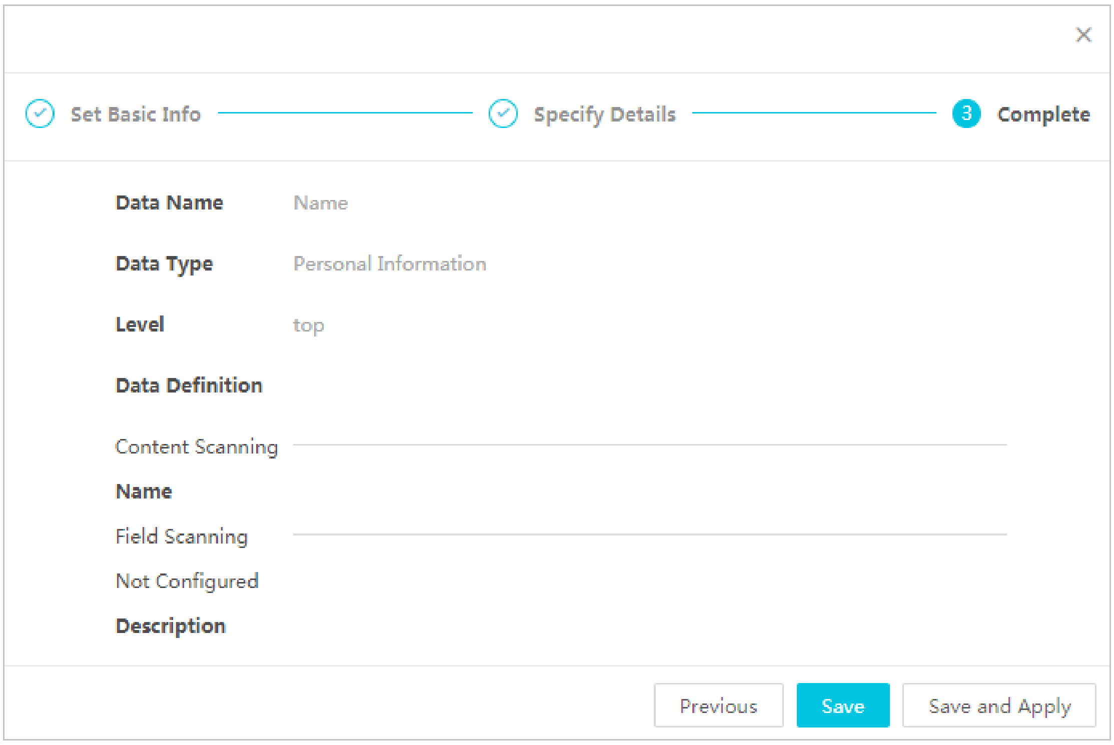
* Data Type: Template
* Data Name: Template
* Owner: Owner
Description: Remarks (less than 120 words)

Parameter	Description
Data Type	The category of the rule. You can select Add By Template or Custom. ◦ If you select **Add By Template**, you can further select **Personal Information**, **Merchant Information**, or **Company Information**. ◦ If you select **Custom**, you can enter a data category.
Data Name	◦ If you select **Add By Template**, you can select a template from the following built-in sensitive data identification templates: **Name**, **Email**, **Seat Number**, **Mobile Phone Number**, **IP**, **Mac Address**, **Car No**, **Address**, **Post Code**, **Id Card**, **Bank Card**, and **Company**. ◦ If you select **Custom**, you can enter a data name. **Note** A custom rule that is used to identify sensitive data must have a unique name.
Owner	The user who sets the rule.
Description	The description of the rule. The description can be up to 128 characters in length.

3. In the **Specify Details** step, set the **Level** and **Data Recognition Rules** parameters and click **Next**.

Parameter	Description
Level	The security level of the data to which the rule is applied. If the existing levels do not meet your needs, go to the **Data Level Management** page and create levels as needed. For more information, see Manage data security levels.
Content Scanning	Specifies whether to enable content scanning. You can use the content that is specified by the selected template or select Regex Express. ○ If you create a template-based rule, you cannot change the content to be scanned. However, you can manually correct the data identification results of the rule. For more information, see Manually correct data. ○ If you select Regex Express, you can customize the identification rule.
Field Scanning	Specifies whether to enable field scanning. You can use exact match or fuzzy match to specify one or more field names to be identified by the rule. The rule is applied if data matches one of the specified field names.

4. After you confirm the settings, click **Save**.



Data Name Name

Data Type Personal Information

Level top

Data Definition

Content Scanning

Name

Field Scanning

Not Configured

Description

> **Note** After the data identification rule is created, the data identification results of the rule are displayed on the next day.

After the data identification rule is created, you can view the data identification results of the rule on the **Data Recognition**, **Data Activities**, and **Data Risks** pages.

Copy a rule

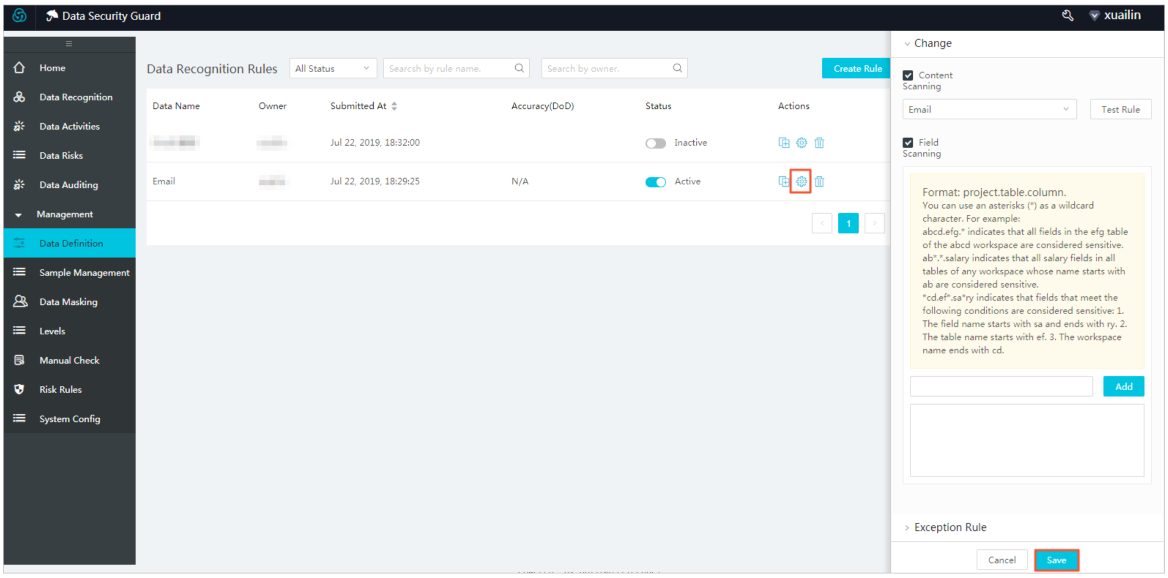
Find a rule and click the **Copy** icon. A new rule with the same settings is created.

By default, the status of the new rule is **Inactive**. You can modify the rule and enable it as needed.

Modify a rule

To modify an existing rule, perform the following steps:

1. Set the status of the rule to **Inactive**.
2. Click the **Rule Configuration** icon.
3. In the panel that appears, modify the parameters in the **Basic Settings**, **Advanced Settings**, **Change**, and **Exception Rule** sections.



Data Name	Owner	Submitted At	Accuracy(DoD)	Status	Actions
[redacted]	[redacted]	Jul 22, 2019, 18:32:00		Inactive	
Email	[redacted]	Jul 22, 2019, 18:29:25	N/A	Active	

☑ Content Scanning

Email

☑ Field Scanning

Format: project.table.column.
You can use an asterisks (*) as a wildcard character. For example:
abcd.efg.* indicates that all fields in the efg table of the abcd workspace are considered sensitive.
ab*.*.salary indicates that all salary fields in all tables of any workspace whose name starts with ab are considered sensitive.
cd.ef.sa*ry indicates that fields that meet the following conditions are considered sensitive: 1. The field name starts with sa and ends with ry. 2. The table name starts with ef. 3. The workspace name ends with cd.

4. Click **Save**.
5. After you confirm the settings, set the status of the rule to **Active**.

Delete a rule

To delete a rule, find the rule and click the **Delete** icon. In the message that appears, click **OK**.

3.7. Customize de-identification rules

This topic describes how to customize de-identification rules in Data Security Guard so that DataWorks can dynamically de-identify the results of ad hoc queries.

Prerequisites

DataWorks Professional Edition or a more advanced edition is activated.

Go to the Data Masking page

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 where your workspace resides, find the workspace, and then click **Data Analytics** in the Actions column.
2. Click the ☰ icon in the upper-left corner and choose **All Products** > **Data governance** > **Data**

Security Guard.

3. Click **Try now** to go to the Data Security Guard homepage.
4. In the left-side navigation pane, choose **Rule Change > Data Masking**.The **Data Masking** page has two tabs: **Data Masking** and **Whitelist**.

Customize de-identification rules in Data Security Guard

1. On the **Data Masking** page, set the **Masking Scene** parameter to **Global Config(_default_scene_code)**.
2. Create a de-identification rule.
   i. On the **Data Masking** tab, click **Create Rule** in the upper-right corner.

ii. In the **Create Rule** dialog box, set the **Masking Rule** and **Method** parameters.

   You can select an existing data identification rule from the **Masking Rule** drop-down list. For more information about data identification rules, see Set data identification rules.

   You can set the **Method** parameter to **Pseudonymisation**, **HASH**, or **Masking Out**. The valid values that are displayed for the **Method** parameter vary based on the data identification rule that you select from the **Masking Rule** drop-down list.

   - **Pseudonymisation**

     This method replaces the text of a data record with an artificial pseudonym of the same data type. If you select this method, you must specify whether to enable **Data watermark** and select a security domain from the **Domain** drop-down list.

     - **Data watermark**: Watermarks allow you to track the source of the data. If your data leaks, you can track the potential source where the data leakage occurs based on the watermark.
     - **Domain**: De-identification policies vary with security domains. In different security domains, different de-identification results are generated for the same data record based on the same de-identification rule. If you do not have a security domain plan for the de-identification rule, randomly select a security domain from the drop-down list.

   - **HASH**

     If you select HASH, you must specify whether to enable **Data watermark** and select a security domain from the **Domain** drop-down list.

     - **Data watermark**: Watermarks allow you to track the source of the data. If your data leaks, you can track the potential source where the data leakage occurs based on the watermark.
     - **Domain**: De-identification policies vary with security domains. In different security domains, different de-identification results are generated for the same data record based on the same de-identification rule. If you do not have a security domain plan for the de-identification rule, randomly select a security domain from the drop-down list.
     -

   - **Masking Out**

     This method uses asterisks (*) to mask specified parts of a data record. This is a commonly used method.

Parameter	Description
Recommended	You can select recommended policies to mask data of common types such as ID card numbers and bank card numbers.
Custom	You can flexibly specify whether to mask the specified number of characters at the first, middle, or last part of a data record.

iii. Click **Save**.

iv. On the **Data Masking** tab, set the status of the created de-identification rule to **Active** or **Inactive** as needed.You can click the 👁 icon in the Actions column of the de-identification rule to test whether it works.

3. Configure a whitelist.
    i. Click the **Whitelist** tab.
    ii. On the **Whitelist** tab, click **Add Account** in the upper-right corner.
    iii. In the **Add Account** dialog box, set the **Rule**, **Account**, and **Effective From** parameters.

    > **Note** If a user queries data beyond the time range that is specified in the whitelist, the query results are de-identified.

Verify the de-identification result in DataWorks

After you create and configure de-identification rules, DataWorks dynamically de-identifies the results of queries in your workspace based on the rules.

> **Note** You must first turn on Mask Data in Page Query Results for your workspace in the DataWorks console. For more information, see Workspace settings.

3.8. Manage data security levels

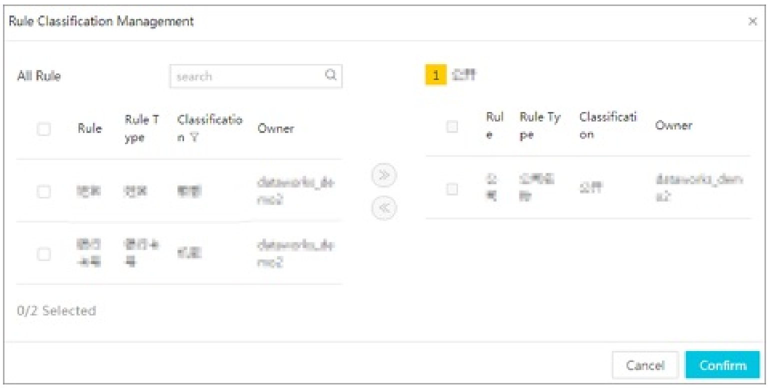
You can manage the data security levels for sensitive data on the Data Level Management page to improve data management efficiency. This topic describes how to create, modify, and delete data security levels, and how to adjust the priority of the levels.

Go to the Data Level Management page

1. Go to the **DataStudio** 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 Analytics** in the Actions column.
2. Click the ☰ icon in the upper-left corner and choose **All Products > Data governance > Data Security Guard**.
3. Click **Try now**. The **Data Security Guard** homepage appears.
4. In the left-side navigation pane, choose **Rule Change > Data Level Management**. On the Data Level Management page, you can create, modify, and delete data security levels, and adjust the priority of the levels.

Manage data security levels

- Create a data security level.
    i. On the **Data Level Management** page, click **Add Level** in the upper-right corner.
    ii. In the **Add Level** dialog box, set the **Level Name** parameter. The **Operated By** parameter cannot be modified.
    iii. Click **OK**.
- Manage a data security level: Find a level and click the ✎ icon. In the **Manage Rules by Level** dialog box, modify the configuration of the level.



- Delete a data security level: You can delete a data security level that you no longer need. Find a level and click the icon. In the message that appears, click **Delete**.
- Adjust the priority of a data security level: Find a level and click the icon to drag the level. You can move up or down the level to raise or lower its priority.

3.9. Manually correct data

This topic describes how to manually correct the data identification results of rules on the Manual Check tab.

> **Note** The manually corrected data identification results are displayed on the next day.

1. Log on to the DataWorks console. Find the required workspace and click **Data Analytics**.
2. Click the More icon in the upper-left corner and choose **All Products** > **Data Security Guard**.
3. Click **Try now**. The **Data Recognition** page appears.
4. In the left-side navigation pane, choose **Rule Change** > **Manual Check**.

   On the Manual Check tab, you can remove incorrectly identified data records, change the rule for identified data records, and remove or recover multiple data records at a time.

   - To remove an incorrectly identified data record, find the data record and click the button in the **Status** column.

     > **Note** You can recover data records that you have removed.

   - To change the rule for an identified data record, click the **Edit** icon next to the rule name and select a required rule.

     > **Note** You can only select a rule that has been configured.

   - To remove or recover multiple data records at a time, select the data records and click **Batch Remove** or **Batch Recovery**.

3.10. Manage risk identification rules

You can configure rules to identify risks in daily access to your data on the Custom Identification Rules page. You can also enable AI-based identification rules to identify data risks.

The **Data Risks** page lists the data activities where risks are identified. You can check these data activities and mark them as secure or risky. On the **Data Activities** page, you can click an activity to view the risk rule that is hit.

1. Log on to the DataWorks console. Find the required workspace and click **Data Analytics**.
2. Click the More icon in the upper-left corner and choose **All Products > Data Security Guard**.
3. Click **Try now**. The **Data Recognition** page appears.
4. In the left-side navigation pane, choose **Rule Change > Custom Identification Rules**. On the Custom Identification Rules page, you can create, copy, modify, and delete risk identification rules. You can also configure AI-based identification rules

Configure risk identification rules

- Create a rule

  Click **Create Rule** in the upper-right corner. In the **Create Rule** dialog box, set the **Rule Name**, **Owner**, and **Description** parameters and click **OK**.

- Copy a rule

  Find a rule and click the **Copy** icon. A new rule with the same settings is created.

  By default, the status of the new rule is **Inactive**. You can modify the rule and enable it as needed.

- Modify a rule

  To modify an existing rule, perform the following steps:

  i. Set the status of the rule to **Inactive**.
  ii. Click the **Edit** icon.
  iii. In the **Change** panel, modify the parameters in the **Basic Settings** and **Rule Settings** sections.
  iv. Click **Save**.
  v. After you confirm the settings, set the status of the rule to **Active**.

- Delete a rule

  To delete a rule, find the rule and click the **Delete** icon. In the message that appears, click **Delete**.

Configure AI-based identification rules

Choose **Custom Identification Rules > AI-based Identification Rules**. On the AI-based Identification Rules tab, only the Similar SQL Query rule is provided.

To enable an AI-based identification rule, set the status of the rule to **Active**.

> **Note**
>
> - After the rule is enabled, SQL queries that hit the rule are displayed on the Data Risks page on the next day.
> - You can disable the rule by changing its status to **Inactive**. Disabling the rule does not remove the SQL queries that have been identified based on the rule.

4.Data Map

4.1. Overview

In addition to metadata management, Data Map also supports the management of data assets in enterprises. Data Map allows you to globally search for data, view metadata details, preview data, view data lineage, and manage data categories. Data Map can help you search for, understand, and use data.

On the homepage of Data Map, you can enter keywords to search for tables by name. You can also click a table in one of the following sections to view the table data: **Recently Viewed Tables**, **Recently Read Tables**, **Most Viewed Tables**, and **Most Read Tables**.

- If you prefer a powerful search engine, go to the homepage to search for data.

  > ⓘ **Note** The homepage appears when you go to the **Data Map** page. To return to the homepage from other pages, click **Data Map** in the upper-left corner.

- If you need to find tables by project, click **All Data** in the top navigation bar.

  On the All Data page, you can view tables in different data stores on different tabs, including the **MaxCompute**, **E-MapReduce**, **OSS**, **AnalyticDB for PostgreSQL**, **MySQL**, **PostgreSQL**, **SQL Server**, and **Oracle** tabs.

  You can add a MaxCompute table to favorites, apply for permissions on a MaxCompute table, view the lineage of a MaxCompute table, and view the data definition language (DDL) statement that is used to create a MaxCompute table.

- If you need to view the overall data of the current tenant, click **Overview** in the top navigation bar.
- If you need to modify tables that are owned by yourself, click **My Data** in the top navigation bar.
- If you are a category administrator or workspace administrator and need to modify the workspace configurations or global categories, click **Configuration Management** in the top navigation bar. For more information, see Manage categories and permissions on MaxCompute tables.

Note on the upgrade from Data Management to Data Map

- Service: Data Map was released to replace Data Management.
- Date:
  - Data Map was released in each region during the period of 20:00:00 to 20:20:00 each day from June 25, 2019 to June 28, 2019.
  - Data Management was brought offline in all regions at 20:00:00 on July 1, 2019.
- Regions where Data Map has been released: China (Shanghai), China (Beijing), China (Hangzhou), China (Shenzhen), China (Hong Kong), Singapore, Australia (Sydney), Malaysia (Kuala Lumpur), India (Mumbai), Indonesia (Jakarta), Germany (Frankfurt), UK (London), US (Virginia), US (Silicon Valley), and UAE (Dubai).

Compared with Data Management, Data Map improves the overall visual interaction and provides features by role. The following table compares Data Management with Data Map.

Item	Data Management	Data Map	Improvement
Page features and roles	Based on the feature types, Data Management provides the following pages: ● Overview page ● Page for searching for data ● Page for managing tables ● Page for managing permissions ● Page for managing settings	● Based on the relationships between roles and features, Data Map provides the following pages: ○ **Overview** ○ Homepage and **All Data**: allow you to search for data. ○ **My Data**: allows you to manage your tables and favorites, apply for permissions, and approve applications. ○ **Configuration Management**: allows you to manage workspaces or categories. ● Data Map enhances the search features for data operators and supports category-based search.	● Data operators who account for the largest proportion of users can search for required data with ease. ● Dedicated pages are provided for data administrators by role, including table owners, workspace administrators, and category administrators. This way, data administrators can understand their responsibilities and use required features with ease.
User operations	Different pages are provided for you to manage tables, manage favorites, manage permissions, and apply for permissions on resources and functions.	● On the **My Data** page, you can apply for permissions on resources and functions and approve permission applications for tables, resources, and functions. ● On the **My Data** page, you can also manage tables and favorites.	Data-related operations are centralized on one page for easy search and use.
Table preview permission	Any user can preview tables. Table owners or workspace administrators cannot control the preview permission.	● Only the owner of a table can preview the table. Other users must apply for the corresponding permission to preview the table. ● A workspace administrator can turn on or off a switch to determine whether to allow other users to preview tables in the production or development environment. ○ By default, users can preview tables in the development environment. ○ By default, users cannot preview tables in the production environment.	● Table owners can manage permissions on their tables in a finer-grained manner. ● Workspace administrators can control the data preview permission of tables in their workspaces with ease.

Item	Data Management	Data Map	Improvement
Table creation permission	You can create a table in Data Management without any permission limit.	• No entry is provided for table creation. • We recommend that you create or modify tables on the Workspace Tables or Manually Triggered Workflows tab of **DataStudio**. This way, you can reuse the permission configurations of the development or O&M role in DataStudio. • You can add a table to a category on the **My Data** page in **Data Map**.	• Tables must be created and modified in **DataStudio**. Permissions are controlled by role. • Only table owners can change the table schema. This prevents unauthorized users from creating tables or adding tables to categories.

Data Map feedback

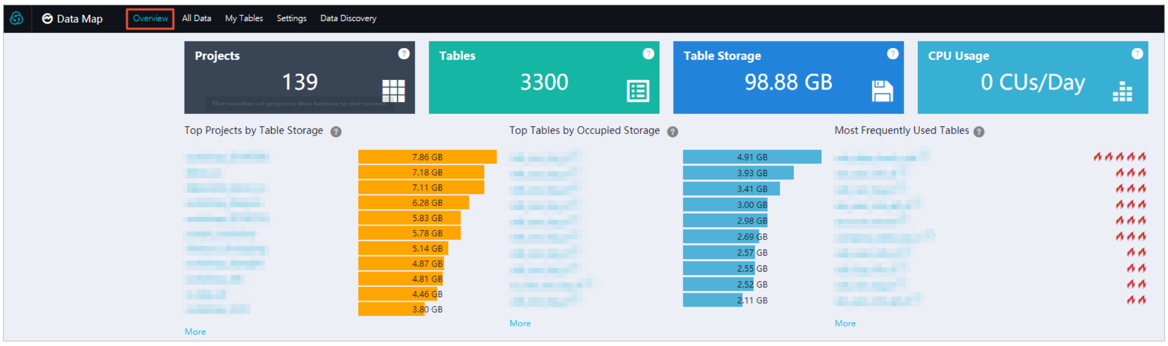
If you have questions about Data Map, scan the following QR code to join the DataWorks DingTalk group for consultation.

4.2. View overall data

This topic describes how to view the overall data of a tenant on the Overview page.

Procedure

1. Log on to the DataWorks console.
2. In the left-side navigation pane, click **Workspaces**.
3. After you select the region where the required workspace resides, find the workspace and click **Data Analytics**.
4. On the DataStudio page, click the ☰ icon in the upper-left corner and choose **All Products > Data governance > DataMap**. The homepage of Data Map appears.
5. In the top navigation bar, click **Overview**.The **Overview** page displays the offline statistics of the current tenant.



Note The Overview page displays statistics for the previous day.

Section	Description
Projects	The total number of MaxCompute projects under the tenant.
Tables	The total number of tables under the tenant.
Occupied Storage	The total storage that is occupied by all tables of the tenant.
Storage trend chart	The offline statistics about the trend of storage usage.
Top Projects by Table Storage	The top projects that occupy the most storage space under the tenant.
Top Tables by Occupied Storage	The top 10 tables that occupy the most storage space under the tenant. You can click a table name to go to the details page of the table. **Note** The logical storage space that is occupied by projects and tables is collected in a T+1 manner. The numbers next to the project and table names indicate the sizes of the occupied logical storage space. The project storage volume includes the table storage volume and the storage volumes of resources, data in the recycle bin, and other system files. Therefore, the project storage volume is larger than the table storage volume. The table storage volume is charged based on the logical storage rather than the physical storage.
Most Frequently Used Tables	The most frequently referenced tables under the tenant. You can click a table name to go to the details page of the table.

4.3. View and manage data and data permissions

You can view and manage data on the Owned by Me, Managed by Me, Managed by Tenant Account, and My Favorites pages. This topic describes how to view and manage data and data permissions.

Context

Data Map updates data one day after the data is generated. If you need to query real-time data, we recommend that you use SQL statements.

Go to the My Data page

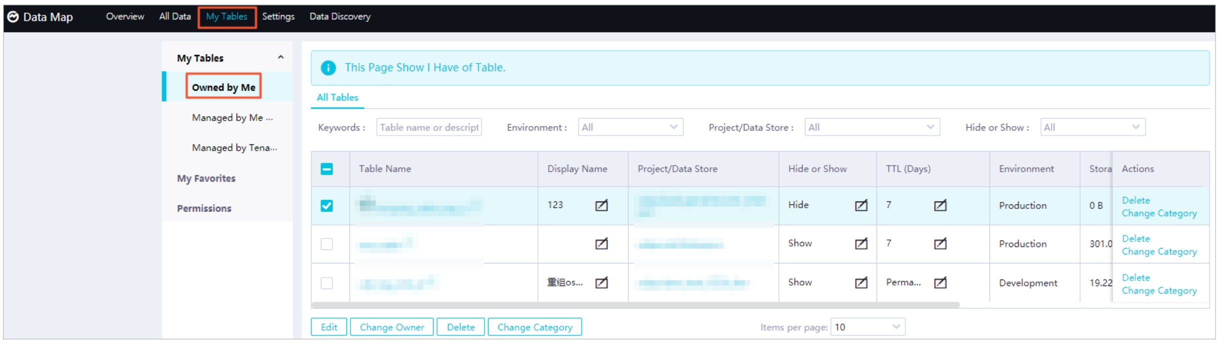
1. Log on to the DataWorks console.
2. In the left-side navigation pane, click **Workspaces**.
3. On the Workspaces page, find the workspace in which you want to view and manage data and data permissions and click **Data Analytics** in the Actions column.
4. On the DataStudio page, click the ☰ icon in the upper-left corner and choose **All Products >**

DataMap. The homepage of Data Map appears.

5. In the top navigation bar, click **My Data**. The **Owned by Me** page appears.

View and manage data on the Owned by Me page

On the **Owned by Me** page, you can search for data by keyword, environment, MaxCompute project or data store, and visibility. You can also view the details about a table and perform relevant operations on the table.



This Page Show I Have of Table.

Keywords : Table name or descript Environment : All Project/Data Store : All Hide or Show : All

	Table Name	Display Name	Project/Data Store	Hide or Show	TTL (Days)	Environment	Stora	Actions
☑	[redacted]	123	[redacted]	Hide	7	Production	0 B	Delete Change Category
□	[redacted]		[redacted]	Show	7	Production	301.0	Delete Change Category
□	[redacted]	重组os...	[redacted]	Show	Perma...	Development	19.22	Delete Change Category

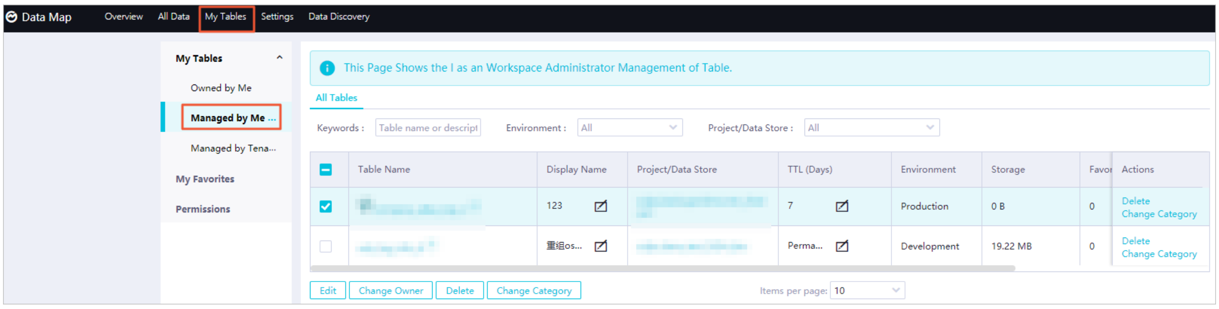
Items per page: 10

GUI element	Description
Table Name	The name of the table. You can click a table name to go to the details page of the table.
Display Name	The display name of the table. You can click the icon in the **Display Name** column of a table to change the display name of the table.
Project/Data Store	The MaxCompute project or data store where the table resides. Different suffixes indicate different environments. For example, _dev indicates the development environment.
Hide or Show	Indicates whether the table is hidden. You can click the icon in the **Hide or Show** column of a table to hide or show the table. **Note** If you hide a table, the details page of the table does not display the **Apply for Permission** button.
TTL (Days)	The time-to-live (TTL) of the table. The value is the same as that you set when you created the table.
Environment	The environment to which the table belongs. Valid values: **Development** and **Production**.
Storage	The volume of data that is stored in the table.
Favorites	The number of times that users add the table to favorites.
Views in Last 30 Days	The number of times that users view the table in the last 30 days.
Created At	The time when the table was created.

GUI element	Description
Actions	The operations that you can perform on the table. You can click **Delete** or **Change Category** in the Actions column of a table to delete the table or change the category of the table.
Edit, Change Owner, Delete, and Change Category	The operations that you can perform on multiple tables at a time. You can select tables and click **Edit**, **Change Owner**, **Delete**, or **Change Category** to modify the tables, change the table owners, delete the tables, or change the categories of the tables.

View and manage data on the Managed by Me page

In the left-side navigation pane, click **Managed by Me**. On the page that appears, you can search for data by keyword, MaxCompute project or data store, and environment. You can also view the details about a table and perform relevant operations on the table.



This Page Shows the I as an Workspace Administrator Management of Table.

Keywords : Table name or descript Environment : All Project/Data Store : All

☐	Table Name	Display Name	Project/Data Store	TTL (Days)	Environment	Storage	Favor	Actions
☑	[redacted]	123 ✎	[redacted]	7 ✎	Production	0 B	0	Delete Change Category
☐	[redacted]	重組os... ✎	[redacted]	Perma... ✎	Development	19.22 MB	0	Delete Change Category

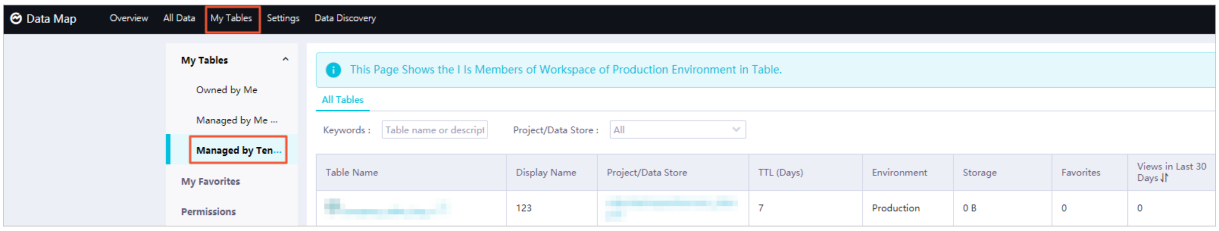
Items per page: 10

GUI element	Description
Table Name	The name of the table. You can click a table name to go to the details page of the table.
Display Name	The display name of the table. You can click the ✎ icon in the **Display Name** column of a table to change the display name of the table.
Project/Data Store	The MaxCompute project or data store where the table resides. Different suffixes indicate different environments. For example, _dev indicates the development environment.
TTL (Days)	The TTL of the table. The value is the same as that you set when you created the table.
Environment	The environment to which the table belongs. Valid values: **Development** and **Production**.
Storage	The volume of data that is stored in the table.
Favorites	The number of times that users add the table to favorites.
Views in Last 30 Days	The number of times that users view the table in the last 30 days.
Created At	The time when the table was created.

GUI element	Description
Actions	The operations that you can perform on the table. You can click **Delete** or **Change Category** in the Actions column of a table to delete the table or change the category of the table.
Edit, Change Owner, Delete, and Change Category	The operations that you can perform on multiple tables at a time. You can select tables and click **Edit**, **Change Owner**, **Delete**, or **Change Category** to modify the tables, change the table owners, delete the tables, or change the categories of the tables.

View data on the Managed by Tenant Account page

In the left-side navigation pane, click **Managed by Tenant Account**. On the page that appears, you can search for data by keyword and MaxCompute project or data store. You can also view the details about a table.



This Page Shows the I Is Members of Workspace of Production Environment in Table.

Keywords : Table name or descript

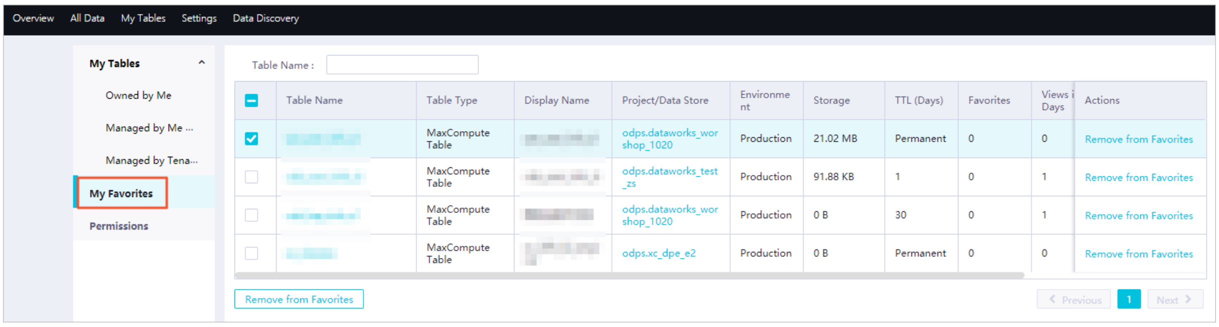
Project/Data Store : All

Table Name	Display Name	Project/Data Store	TTL (Days)	Environment	Storage	Favorites	Views in Last 30 Days
[redacted]	123	[redacted]	7	Production	0 B	0	0

GUI element	Description
Table Name	The name of the table. You can click a table name to go to the details page of the table.
Display Name	The display name of the table.
Project/Data Store	The MaxCompute project or data store where the table resides. Different suffixes indicate different environments. For example, **_dev** indicates the development environment.
TTL (Days)	The TTL of the table. The value is the same as that you set when you created the table.
Environment	The environment to which the table belongs. Valid values: **Development** and **Production**.
Storage	The volume of data that is stored in the table.
Favorites	The number of times that users add the table to favorites.
Views in Last 30 Days	The number of times that users view the table in the last 30 days.
Created At	The time when the table was created.

View data on the My Favorites page

In the left-side navigation pane, click **My Favorites**. On the page that appears, you can view the tables that you have added to favorites.



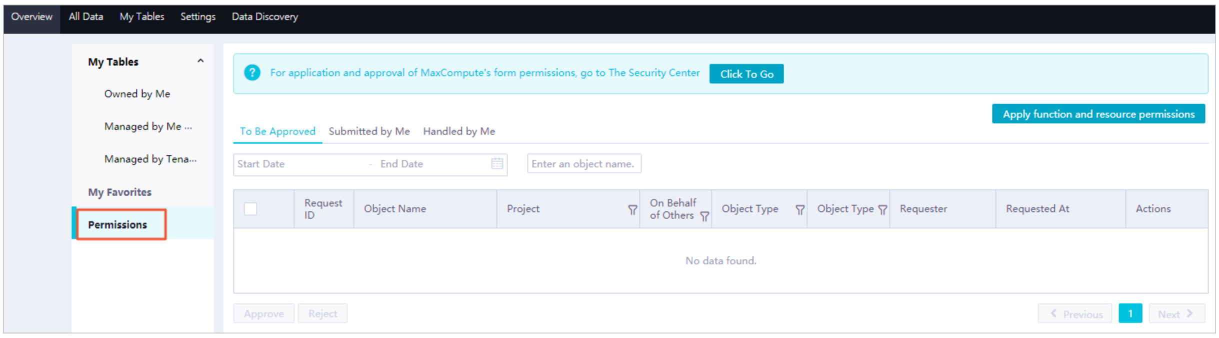
☐	Table Name	Table Type	Display Name	Project/Data Store	Environment	Storage	TTL (Days)	Favorites	Views i Days	Actions
☑	[redacted]	MaxCompute Table	[redacted]	odps.dataworks_workshop_1020	Production	21.02 MB	Permanent	0	0	Remove from Favorites
☐	[redacted]	MaxCompute Table	[redacted]	odps.dataworks_test_zs	Production	91.88 KB	1	0	1	Remove from Favorites
☐	[redacted]	MaxCompute Table	[redacted]	odps.dataworks_workshop_1020	Production	0 B	30	0	1	Remove from Favorites
☐	[redacted]	MaxCompute Table	[redacted]	odps.xc_dpe_e2	Production	0 B	Permanent	0	0	Remove from Favorites

You can click **Remove from Favorites** in the Actions column of a table to remove the table from your favorites.

View and manage data permissions

In the left-side navigation pane, click **Permission Management**. On the Permission Management page, you can view and manage data permissions.

You can click **Apply for Function and Resource Permissions** in the upper-right corner of the **Permission Management** page to request permissions. You can also view permission requests on the **To Be Approved**, **Submitted by Me**, and **Handled by Me** tabs.



For application and approval of MaxCompute's form permissions, go to The Security Center

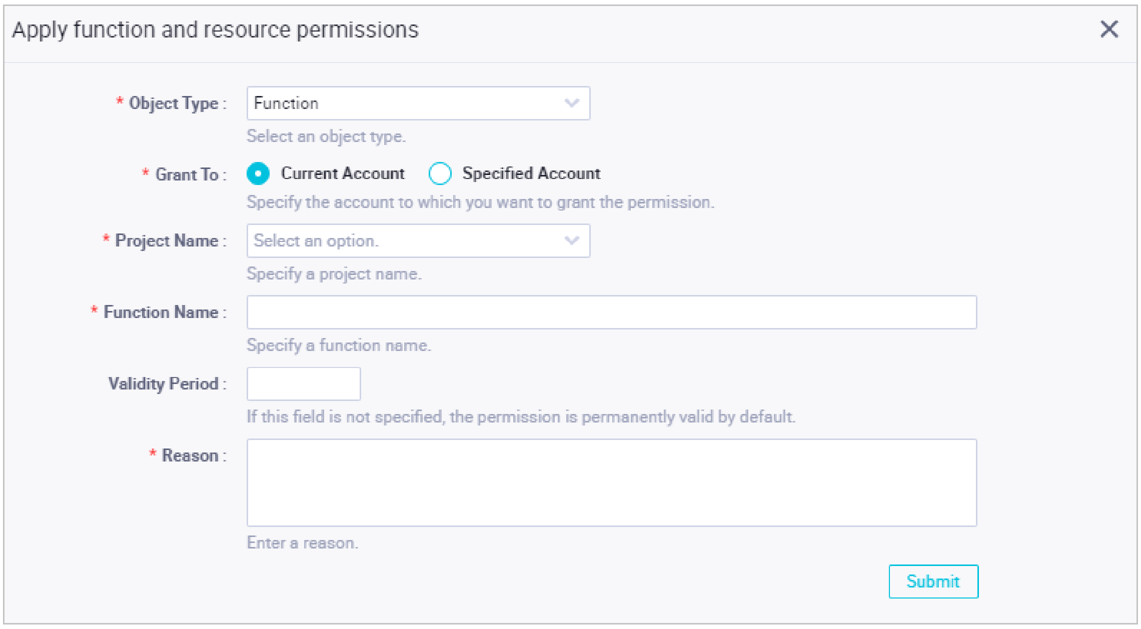
☐	Request ID	Object Name	Project	On Behalf of Others	Object Type	Object Type	Requester	Requested At	Actions

No data found.

- **Apply for Function and Resource Permissions**

  i. On the Permission Management page, click **Apply for Function and Resource Permissions** in the upper-right corner.

  ii. In the **Apply for Function and Resource Permissions** dialog box, set the parameters as required. The following table describes the parameters for requesting permissions.



Apply function and resource permissions

* Object Type : Function
Select an object type.

* Grant To : ● Current Account ○ Specified Account
Specify the account to which you want to grant the permission.

* Project Name : Select an option.
Specify a project name.

* Function Name :
Specify a function name.

Validity Period :
If this field is not specified, the permission is permanently valid by default.

* Reason :
Enter a reason.

Parameter	Description
Object Type	The type of object on which you want to request permissions. Valid values: **Functions** and **Resources**.
Grant To	The account to which the permissions will be granted. Valid values: **Current Account** and **Specified Account**. ■ If you select **Current Account**, the permissions will be granted to you after the request is approved. ■ If you select **Specified Account**, you must also set the **Username** parameter. The permissions will be granted to the specified account after the request is approved.
Project Name	The name of the MaxCompute project that contains the function or resource on which you want to request permissions.
Function Name or **Resource Name**	The full name of the function or resource in the project. If the resource is a file, enter the full name of the file, including the file name extension, for example, my_mr.jar.
Validity Period	The validity period of the permissions, in days. If this parameter is not specified, the permissions are permanently valid. After the validity period expires, the system automatically revokes the permissions.
Reason	The reason why you request the permissions.

- **To Be Approved**

  If you are a workspace administrator, click the **To Be Approved** tab. On the tab, you can view and approve the requests for permissions on all objects such as tables, resources, and functions in the workspace.

- **Submitted by Me**

  On the **Permission Management** page, click the **Submitted by Me** tab.

On the **Submitted by Me** tab, you can view the permission requests that you have submitted.

- **Handled by Me**

  If you are a workspace administrator, click the **Handled by Me** tab on the **Permission Management** page.

  On the **Handled by Me** tab, you can view the permission requests that you have handled for all objects such as tables, resources, and functions in the workspace.

4.4. Manage categories and permissions on MaxCompute tables

This topic describes how to manage categories and permissions on MaxCompute tables that are in your owned or managed workspaces on the Configuration Management page of Data Map.

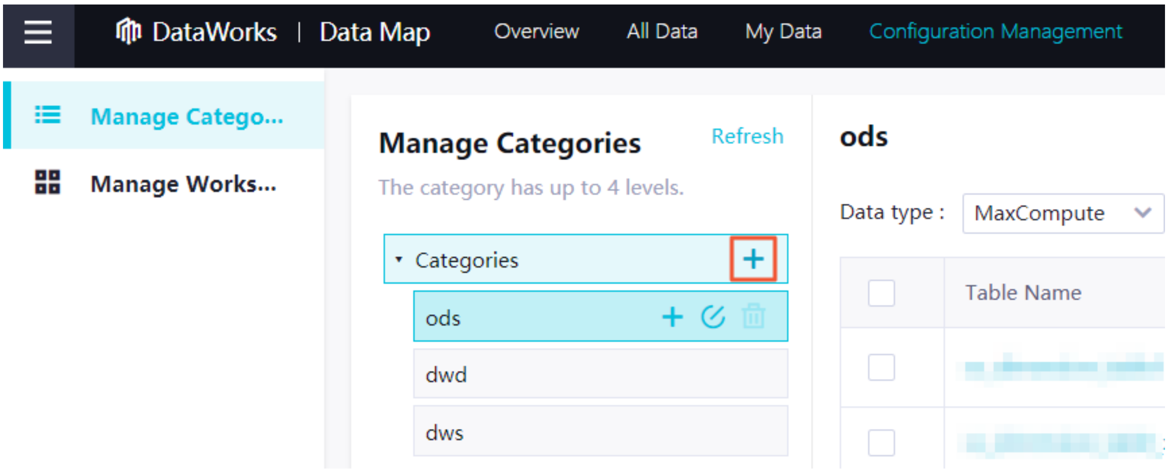
Go to the Configuration Management page

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 where your workspace resides. Find the workspace and click **Data Map** in the Actions column.
4. In the top navigation bar, click **Configuration Management**. The **Manage Categories** page appears.The **Configuration Management** page allows you to manage categories and permissions on MaxCompute tables in workspaces.

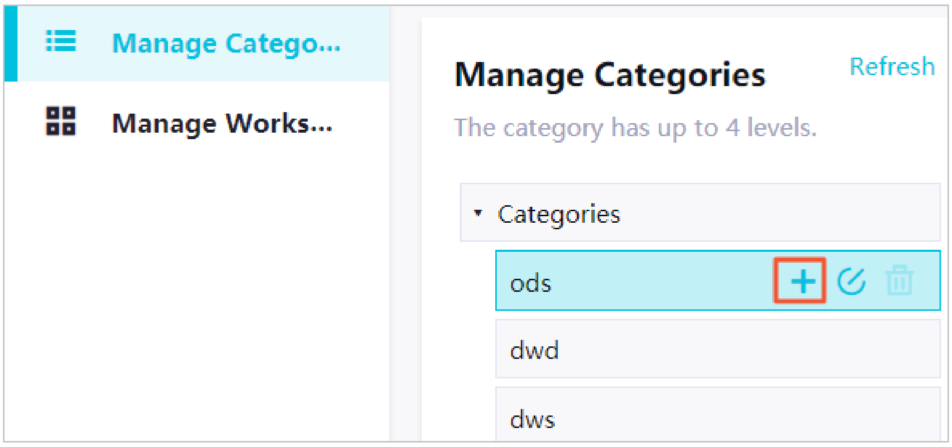
Manage categories

On the **Manage Categories** page, you can perform the following steps to create a category and add tables to and remove tables from the category:

1. On the **Manage Categories** page, move the pointer over **Categories** and click the + icon. In the field that appears, enter a category name and press the Enter key to create a level 1 category.

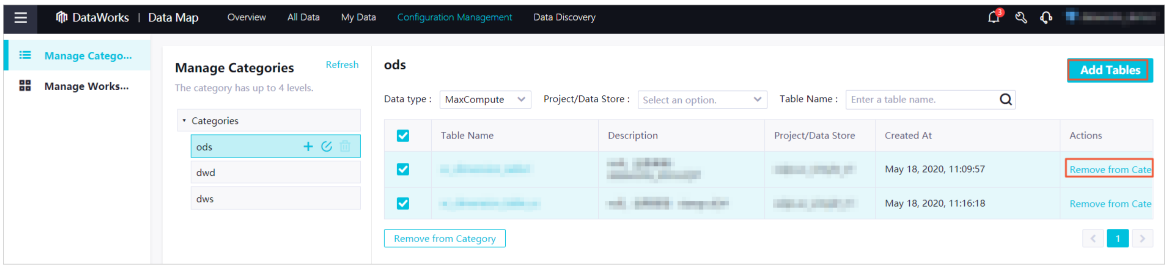


2. Move the pointer over the level 1 category and click the + icon. In the field that appears, enter a category name and press the Enter key to create a level 2 category.



Use the same method to create more categories. DataWorks allows you to create categories at a maximum of four levels. You can click the ✎ icon to edit a category or click the 🗑 icon to delete a category.

3. Add tables to and remove tables from a category.

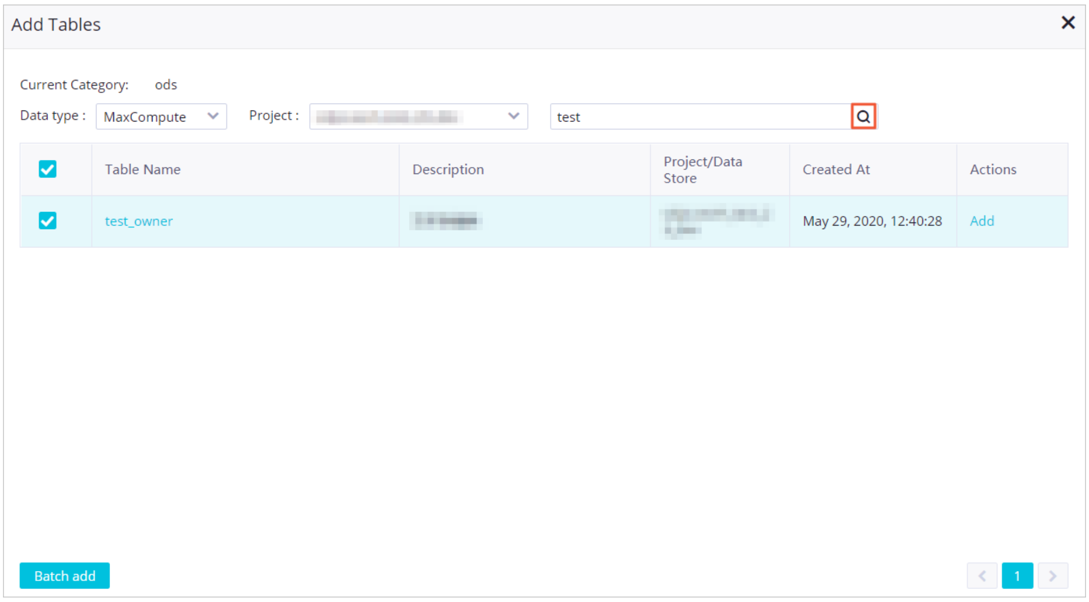


ods

Data type : MaxCompute Project/Data Store : Select an option. Table Name : Enter a table name.

☑	Table Name	Description	Project/Data Store	Created At	Actions
☑	[redacted]	[redacted]	[redacted]	May 18, 2020, 11:09:57	Remove from Cate
☑	[redacted]	[redacted]	[redacted]	May 18, 2020, 11:16:18	Remove from Cate

- Add tables to a category
  a. Select the category and click **Add Tables** in the upper-right corner.
  b. In the **Add Tables** dialog box, specify the table type and project, enter a table name or keyword, and then click the 🔍 icon to search for tables.



Current Category: ods

Data type : MaxCompute Project : [redacted] test

☑	Table Name	Description	Project/Data Store	Created At	Actions
☑	test_owner	[redacted]	[redacted]	May 29, 2020, 12:40:28	Add

c. If you want to add a table to the category, find the table and click **Add** in the Actions column.

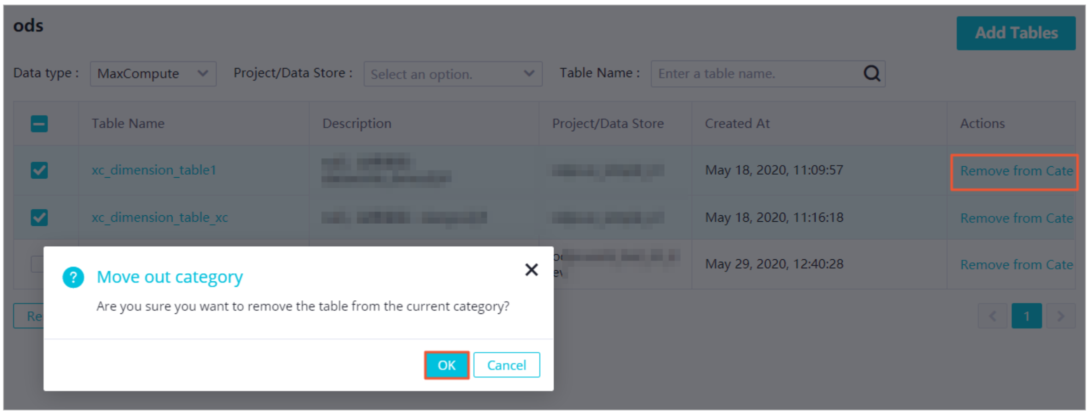
   If you want to add multiple tables at a time, select the tables and click **Batch add**.

- Remove tables from a category

  a. Select the category. If you want to remove a table from the category, find the table and click **Remove** in the Actions column.

     If you want to remove multiple tables at a time, select the tables and click **Remove from Category**.

  b. In the **Move out category** message, click **OK**.



ods

Add Tables

Data type : MaxCompute Project/Data Store : Select an option. Table Name : Enter a table name.

☐	Table Name	Description	Project/Data Store	Created At	Actions
☑	xc_dimension_table1	[redacted]	[redacted]	May 18, 2020, 11:09:57	Remove from Cate
☑	xc_dimension_table_xc	[redacted]	[redacted]	May 18, 2020, 11:16:18	Remove from Cate
☐			[redacted]	May 29, 2020, 12:40:28	Remove from Cate

Move out category

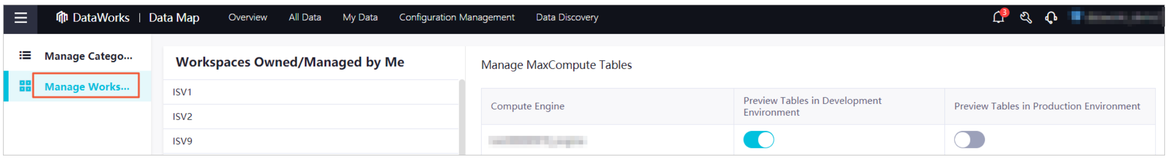
Are you sure you want to remove the table from the current category?

OK Cancel

Manage permissions on MaxCompute tables

On the **Manage Workspaces** page, you can specify whether MaxCompute tables can be previewed in the compute engine in the development and production environments.

1. In the left-side navigation pane, click **Manage Workspaces**.
2. In the **Workspaces Owned/Managed by Me** section, click the workspace for which you want to manage permissions on MaxCompute tables.
3. In the **Manage MaxCompute Tables** section, turn on or off the switch in the **Preview Tables in Development Environment** or **Preview Tables in Production Environment** column.



Workspaces Owned/Managed by Me

ISV1

ISV2

ISV9

Manage MaxCompute Tables

Compute Engine	Preview Tables in Development Environment	Preview Tables in Production Environment
[redacted]	●	○

4. If you turn on the switch in the **Preview Tables in Production Environment** column, the **Attention** message appears. To turn on the switch, you must click **I already know that I am sure to open it**.

> **Note**
>
> - If the workspace is in basic mode, you can turn on or off only the switch in the **Preview Tables in Production Environment** column.
> - After the switch in the **Preview Tables in Production Environment** column is turned on, all members of the workspace can preview MaxCompute tables in the production environment without requesting permissions. This may cause the leak of sensitive data. Therefore, use caution before you turn on the switch.

4.5. Table details

4.5.1. View the details of a table

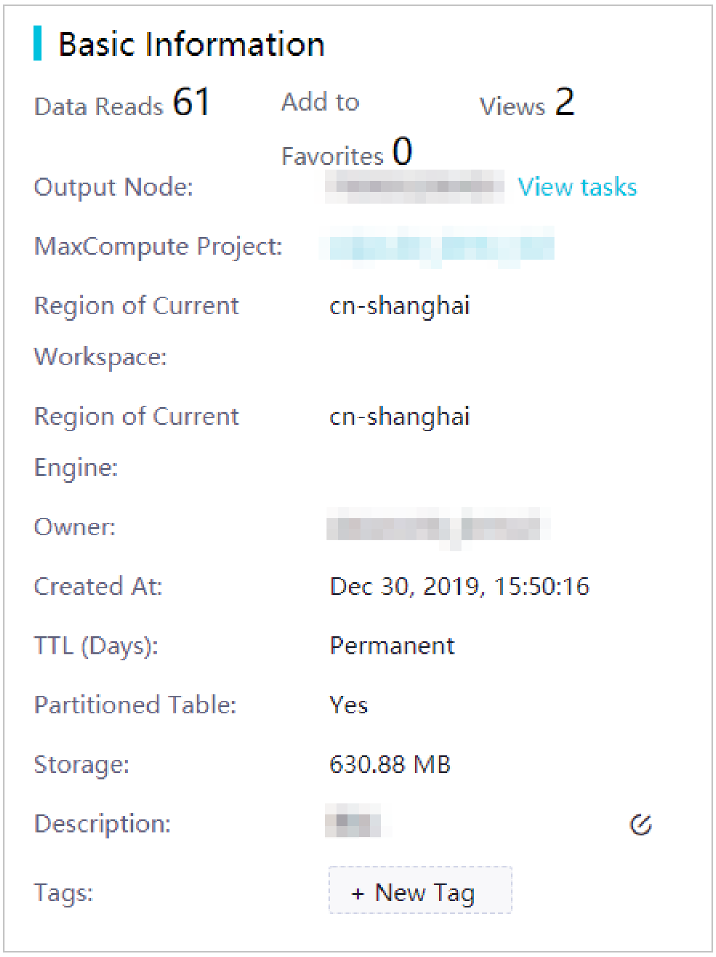
This topic describes how to go to the details page of a table and view the details about the table, such as the basic information, output information, and lineage information.

Go to the details page of a table

1. Log on to the DataWorks console.
2. In the left-side navigation pane, click **Workspaces**.
3. On the Workspaces page, find the workspace in which you want to view the details of a table and click **Data Analytics** in the Actions column.
4. On the DataStudio page, click the ☰ icon in the upper-left corner and choose **All Products > Data governance > DataMap**. The homepage of Data Map appears.
5. In the top navigation bar, click **All Data**.
6. On the All Data page, click a tab as required, such as MaxCompute.
7. On the tab that appears, click the name of the table that you want to view.On the details page that appears, you can view the basic information, business information, permission information, technical information, detailed information, output information, lineage information, reference records, and usage notes of the table. You can also preview and detect data in the table.

View basic information

In the **Basic table information** section, you can view the numbers of reads, favorites, and views. You can also check the output nodes, MaxCompute project name, region to which the current workspace belongs, region to which the engine belongs, owner, creation time, time-to-live (TTL), storage capacity, description, and tags of the table, and whether the table is partitioned.



Basic Information

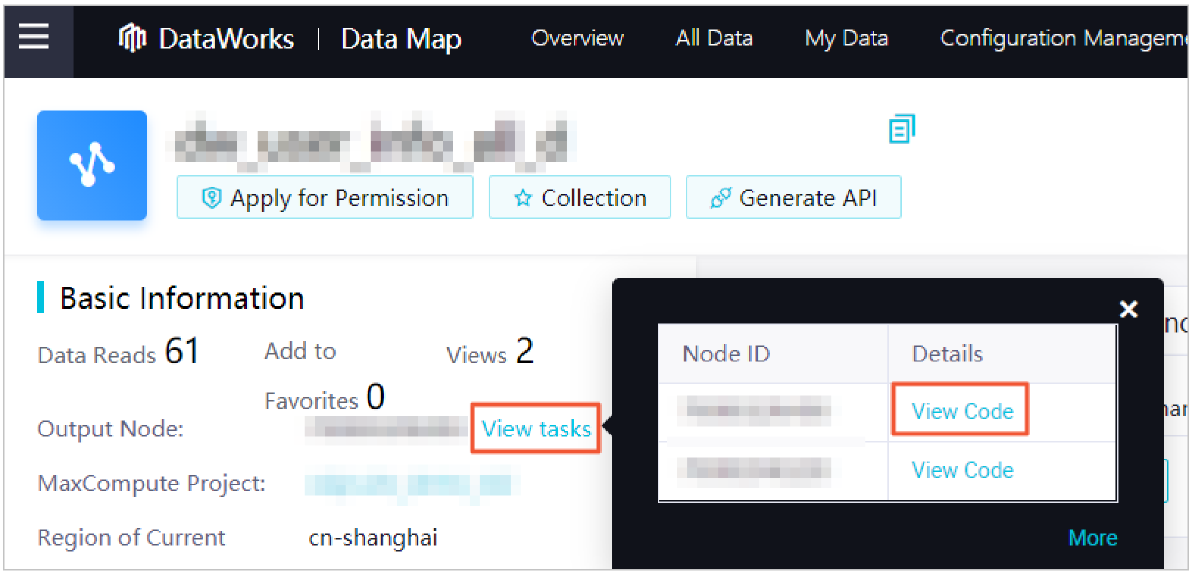
Data Reads 61 Add to Favorites 0 Views 2

Output Node:	[redacted] View tasks
MaxCompute Project:	[redacted]
Region of Current Workspace:	cn-shanghai
Region of Current Engine:	cn-shanghai
Owner:	[redacted]
Created At:	Dec 30, 2019, 15:50:16
TTL (Days):	Permanent
Partitioned Table:	Yes
Storage:	630.88 MB
Description:	[redacted]
Tags:	+ New Tag

You can perform the following operations in the **Basic table information** section:

- View the code of the output node of the table: Click **View Code** next to **Output Node**. On the **Operation Center** page, view the node code.

If the table has multiple output nodes, move the pointer over **View tasks** next to **Output Node**. In the pop-up window, find the output node whose code you want to view and click **View Code** in the Details column.



[redacted]

Basic Information

Data Reads 61
Add to Favorites 0
Views 2
Output Node: [redacted]
MaxCompute Project: [redacted]
Region of Current cn-shanghai

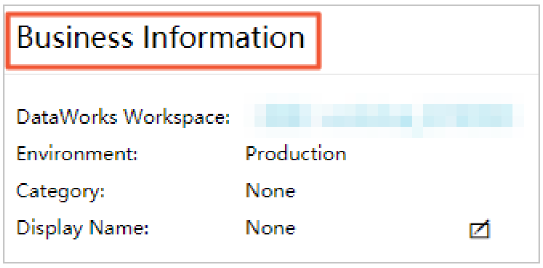
Node ID	Details
[redacted]	View Code
[redacted]	View Code

- View the details about the MaxCompute project: Click the MaxCompute project name. On the page that appears, view the details about the MaxCompute project to which the table belongs.
- Edit the description of the table: Click the ✎ icon next to **Description**, enter a description in the field that appears, and then click the ✔ icon.
- Add a tag to or remove a tag from the table: Click + New Tag next to **Tags**, enter a tag name in the field that appears, and then press the **Enter** key.

  To remove a tag from the table, move the pointer over the tag and click the **X** icon.

View business information

In the **Business Information** section, you can view the DataWorks workspace name, environment type, category, and display name of the table.



Business Information

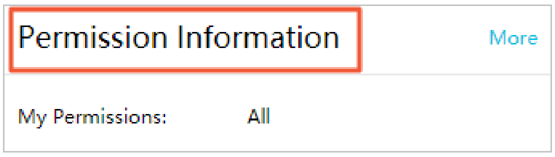
DataWorks Workspace: [redacted]
Environment: Production
Category: None
Display Name: None

You can perform the following operations in the **Business Information** section:

- View the details about the workspace: Click the DataWorks workspace name. On the page that appears, view the details about the DataWorks workspace to which the table belongs.
- Edit the display name of the table: Click the ✎ icon next to **Display Name**, enter a name in the field that appears, and then click the ✔ icon.

View permission information

In the **Permission Information** section, you can view your permissions on the table.



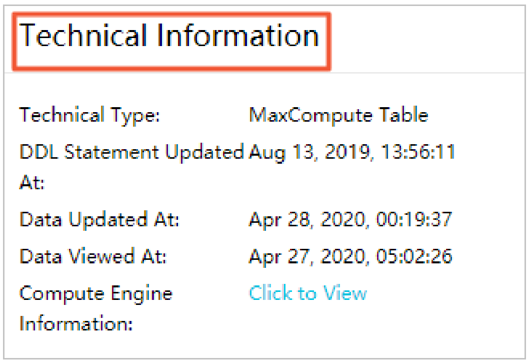
Permission Information

My Permissions: All

In the **Permission Information** section, you can click **More** in the upper-right corner. In the **Apply for Permission** dialog box, you can request more permissions on the table.

View technical information

In the **Technical Information** section, you can view the technical type, last time when the data definition language (DDL) statement was modified, last time when the data was modified, last time when the data was viewed, and compute engine information.



Technical Information

Technical Type: MaxCompute Table

DDL Statement Updated At: Aug 13, 2019, 13:56:11

Data Updated At: Apr 28, 2020, 00:19:37

Data Viewed At: Apr 27, 2020, 05:02:26

Compute Engine Information: Click to View

In the **Technical Information** section, you can click **View** next to **Compute Engine Information**. In the **Compute Engine Information** dialog box, you can view or copy the information about the compute engine.

Note By default, the time format yyyy-MM-dd HH:mm:ss is used to describe the compute engine.

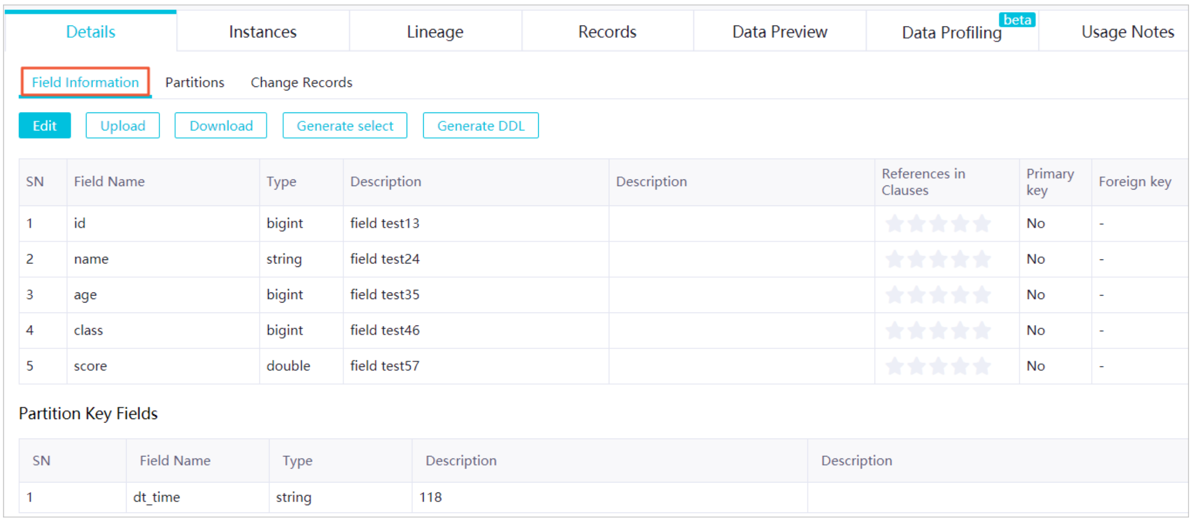
View detailed information

The **Details** tab contains the following tabs: **Field Information**, **Partitions**, and **Change Records**.

- **Field Information** tab

  On the **Field Information** tab, you can view the name, data type, description, business description, and popularity of each field. You can also check whether a field is a primary key or foreign key.

If the table is a partitioned table, you can also view the information about partition fields, including the field name, data type, description, and business description.



SN	Field Name	Type	Description	Description	References in Clauses	Primary key	Foreign key
1	id	bigint	field test13		☆☆☆☆☆	No	-
2	name	string	field test24		☆☆☆☆☆	No	-
3	age	bigint	field test35		☆☆☆☆☆	No	-
4	class	bigint	field test46		☆☆☆☆☆	No	-
5	score	double	field test57		☆☆☆☆☆	No	-

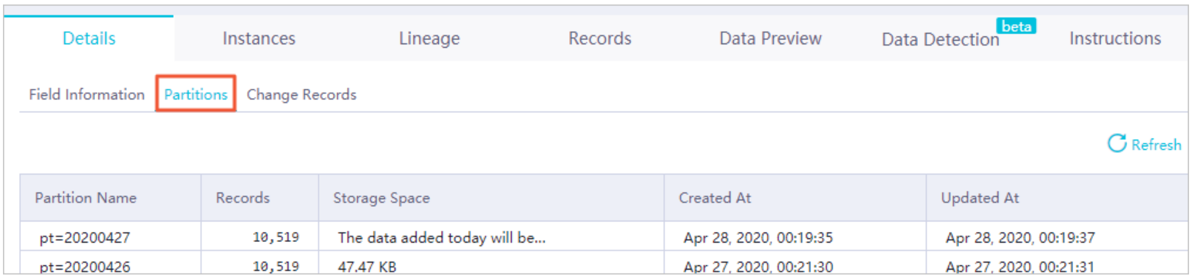
Partition Key Fields

SN	Field Name	Type	Description	Description
1	dt_time	string	118	

Action	Description
Edit	Click this button. You can modify the field description and business description, specify the security level, and specify whether a field is a primary key. Then, click **Save** or **Cancel** as needed. You can specify a security level for multiple fields at a time. **Note** You can specify security levels for fields only in a MaxCompute project with LabelSecurity enabled.
Upload	Click this button. In the **Batch Upload Field Information** dialog box, drag and upload a local file. Only .xlsx files created in Excel 2007 are supported. You can also click **Download Template**.
Download	Click this button to download the field information of the current table.
Generate select	Click this button. In the **Generate SELECT Statement** dialog box, view or copy the `SELECT` statement that you can use to query data in the current table.
Generate DDL.	Click this button. In the **Generate DDL Statement** dialog box, view or copy the DDL statement that is used to create the current table.

- **Partitions** tab

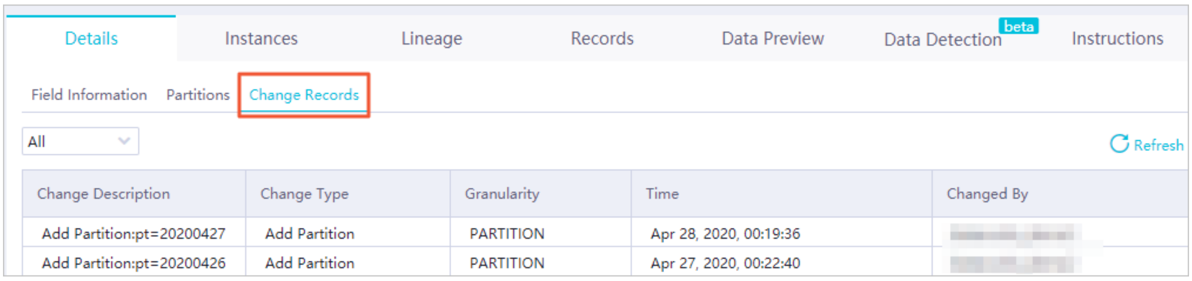
On the **Partitions** tab, you can view the name, number of records, storage capacity, creation time, and last update time of each partition in the current table.



Partition Name	Records	Storage Space	Created At	Updated At
pt=20200427	10,519	The data added today will be...	Apr 28, 2020, 00:19:35	Apr 28, 2020, 00:19:37
pt=20200426	10,519	47.47 KB	Apr 27, 2020, 00:21:30	Apr 27, 2020, 00:21:31

- **Change Records** tab

  On the **Change Records** tab, you can view the description, type, granularity, time, and operator of changes that are performed on the current table.



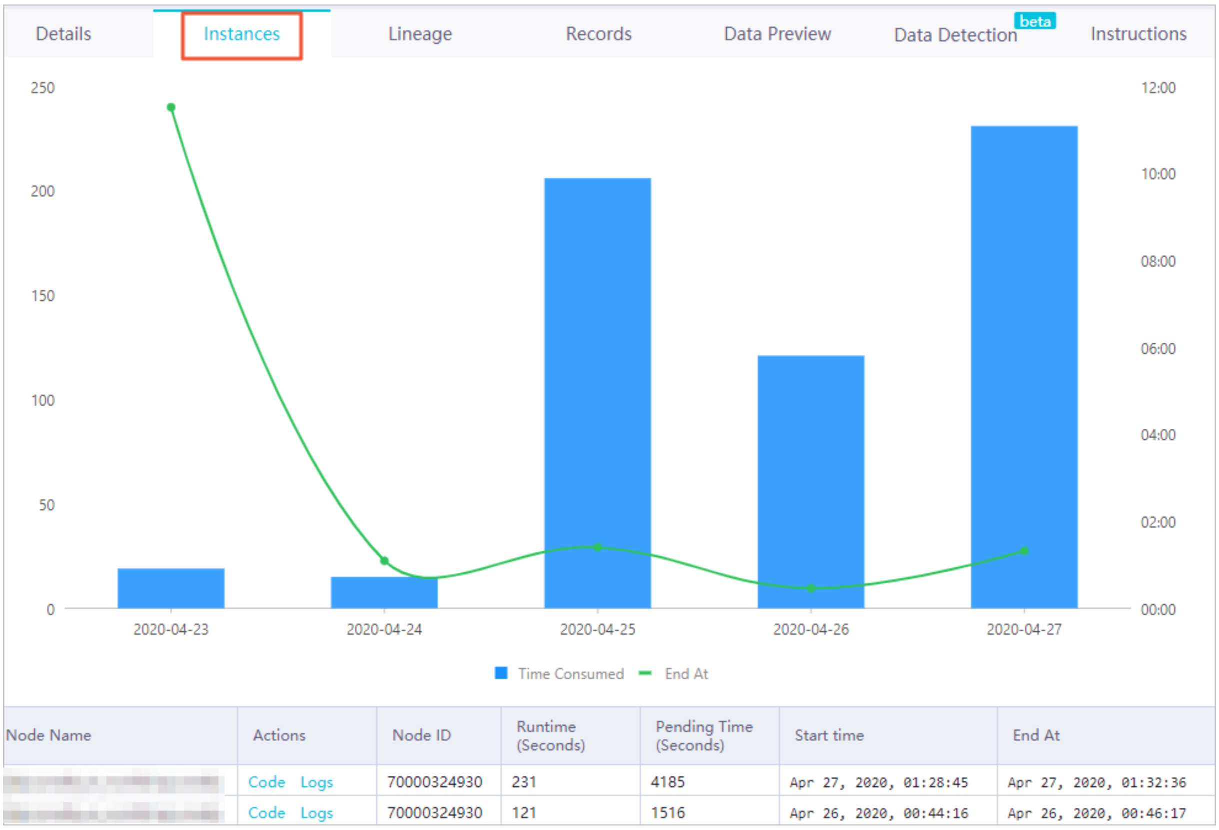
Change Description	Change Type	Granularity	Time	Changed By
Add Partition:pt=20200427	Add Partition	PARTITION	Apr 28, 2020, 00:19:36	[redacted]
Add Partition:pt=20200426	Add Partition	PARTITION	Apr 27, 2020, 00:22:40	[redacted]

  On the **Change Records** tab, you can also select a change type from the drop-down list in the upper-left corner to filter the table changes.

  Change types include **Create Table**, **Modify Table**, **Delete Table**, **Create Partition**, **Delete Partition**, **Change Owner**, and **Change TTL**.

View output information

If the table data periodically changes with the corresponding node, you can view the change status and data that is continuously updated on the **Instances** tab.



Node Name	Actions	Node ID	Runtime (Seconds)	Pending Time (Seconds)	Start time	End At
[redacted]	Code Logs	70000324930	231	4185	Apr 27, 2020, 01:28:45	Apr 27, 2020, 01:32:36
[redacted]	Code Logs	70000324930	121	1516	Apr 26, 2020, 00:44:16	Apr 26, 2020, 00:46:17

On this tab, you can also click **View Code** or **View Log** in the **Actions** column of a node to view the code or logs of the node.

View lineage information

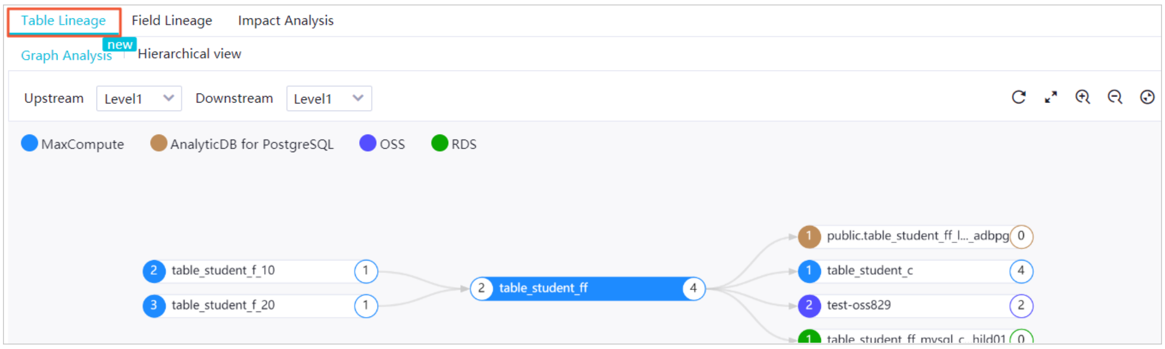
On the Lineage tab, you can view the source and destination of data and manage the lineage information with ease.

> **Note** To use this feature, you must activate a DataWorks advanced edition. For example, if the compute engine is MaxCompute, this feature is available only in DataWorks Standard Edition or more advanced editions. If the compute engine is E-MapReduce, this feature is available only in DataWorks Professional Edition or more advanced editions.

The **Lineage** tab contains the following tabs: **Table Lineage**, **Field Lineage**, and **Impact Analysis**.

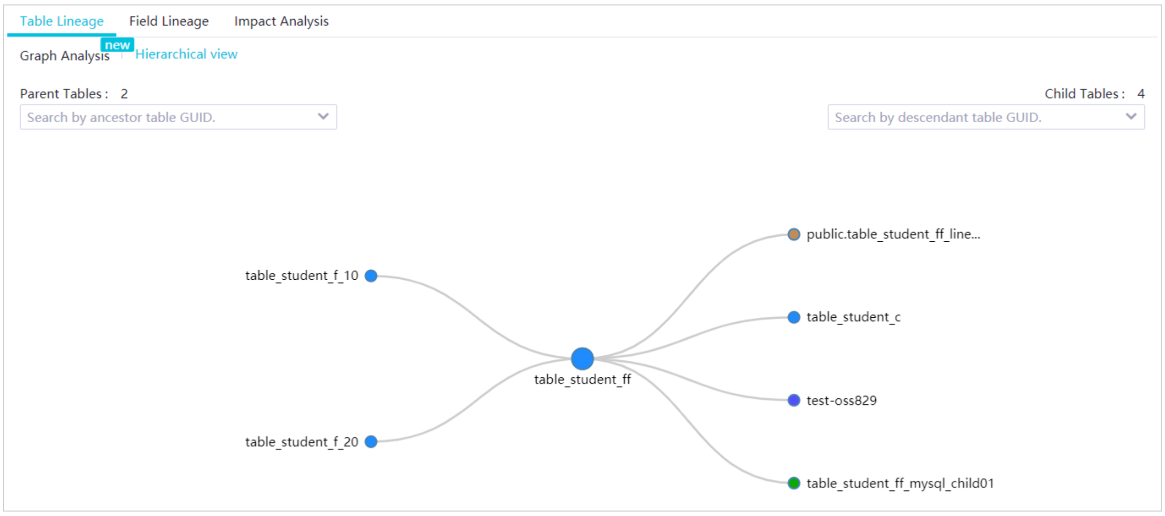
- The **Table Lineage** tab consists of the **Graph Analysis** and **Hierarchical view** tabs.

- **Graph Analysis** tab: displays the ancestor and descendant tables of a specified level for the current table and the number of ancestor and descendant tables for each table.



Upstream Level1 Downstream Level1

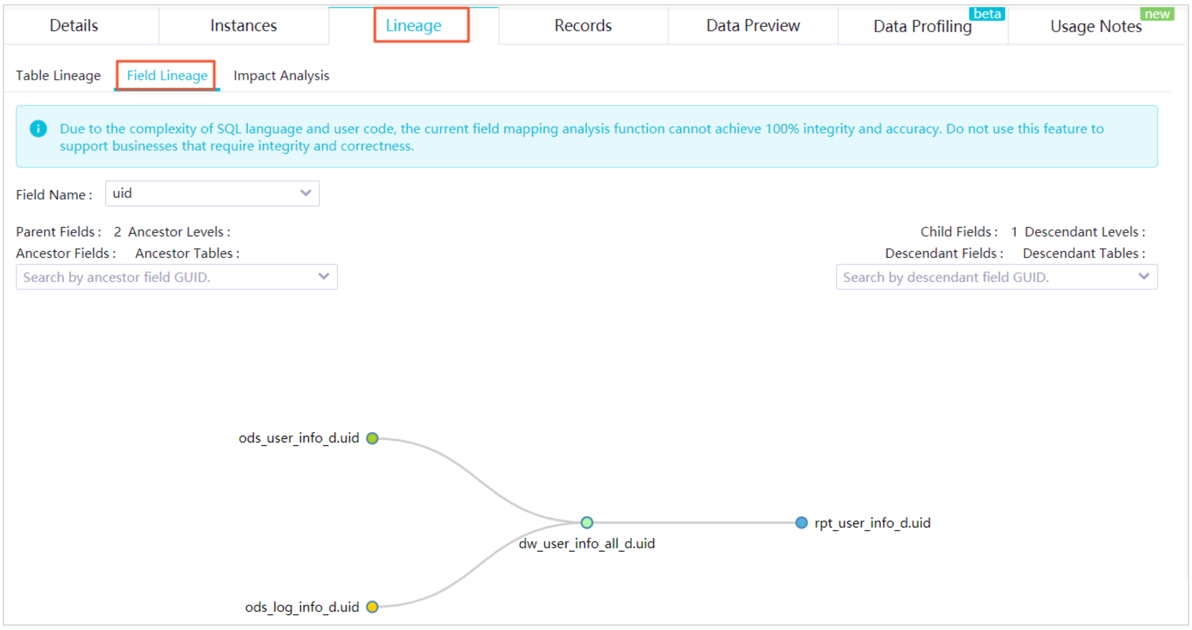
- **Hierarchical view** tab: displays the parent and child tables of the current table by default. You can search for the parent and child tables based on the globally unique identifier (GUID).



Parent Tables : 2

Child Tables : 4

- On the **Field Lineage** tab, you can select a field from the **Field Name** drop-down list to view the lineage information of the field.



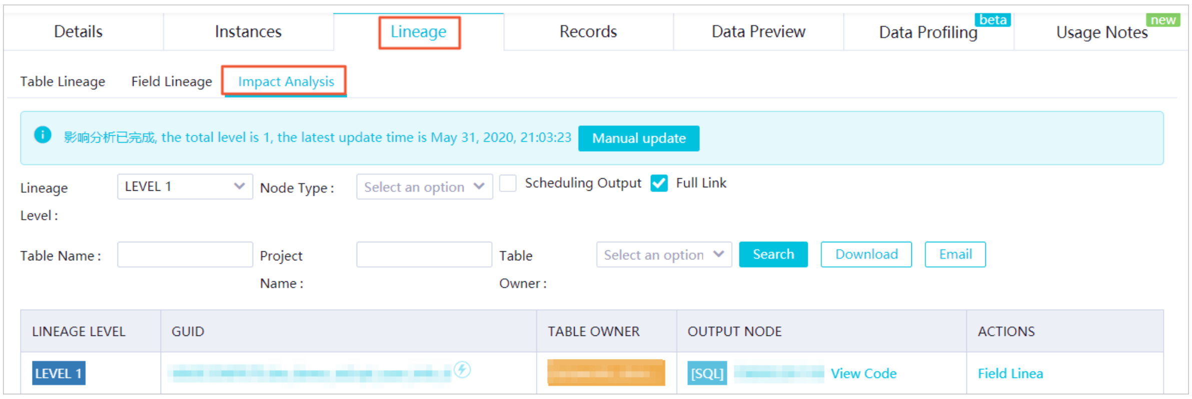
Due to the complexity of SQL language and user code, the current field mapping analysis function cannot achieve 100% integrity and accuracy. Do not use this feature to support businesses that require integrity and correctness.

Field Name : uid

Parent Fields : 2 Ancestor Levels :
Ancestor Fields : Ancestor Tables :

Child Fields : 1 Descendant Levels :
Descendant Fields : Descendant Tables :

- On the **Impact Analysis** tab, you can query the node that generates a lineage and the full link of the lineage based on information such as the lineage level, field, node type, table name, workspace name, and table owner.



影响分析已完成, the total level is 1, the latest update time is May 31, 2020, 21:03:23

Lineage Level : LEVEL 1 Node Type : Select an option □ Scheduling Output ☑ Full Link

Table Name : Project Name : Table Owner : Select an option

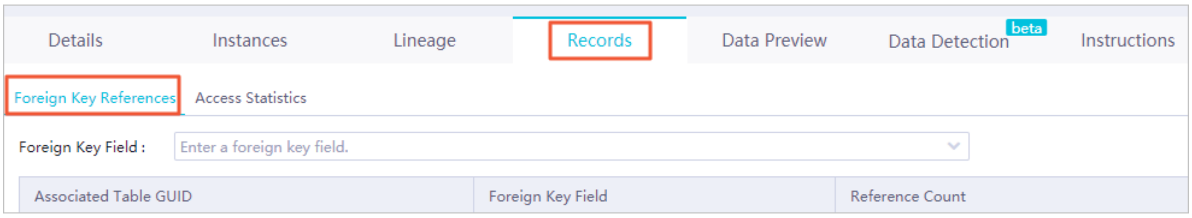
LINEAGE LEVEL	GUID	TABLE OWNER	OUTPUT NODE	ACTIONS
LEVEL 1	[redacted]	[redacted]	[SQL] [redacted] View Code	Field Linea

You can click **Manual update** to rerun the impact analysis. You can also download the impact analysis result or send the impact analysis result to the owners of descendant tables of the current table by sending emails.

View reference records

The **Records** tab contains the following tabs: **Foreign Key References** and **Access Statistics**.

- **Foreign Key References** tab: On this tab, you can check the number of users who reference the current table.



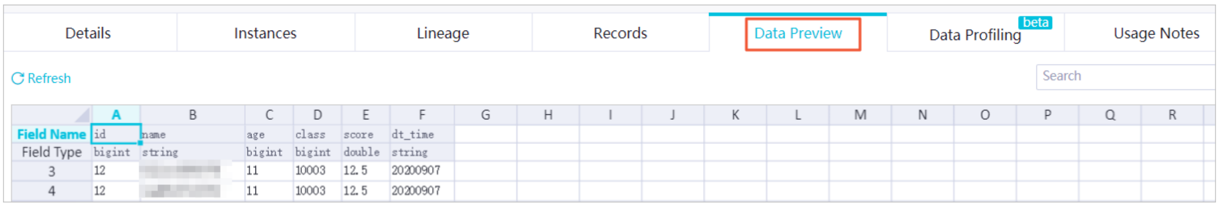
Foreign Key Field : Enter a foreign key field.

Associated Table GUID	Foreign Key Field	Reference Count

- **Access Statistics** tab: On this tab, you can view the reference records in a line chart.

Preview data

On the **Data Preview** tab, you can preview the data of the current table.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
Field Name	id	name	age	class	score	dt_time												
Field Type	bigint	string	bigint	bigint	double	string												
3	12	[redacted]	11	10003	12.5	20200907												
4	12	[redacted]	11	10003	12.5	20200907												

> **Notice** Only authorized users can preview tables in the production environment. If you do not have the required permissions, click **Apply Now**.

Detect data

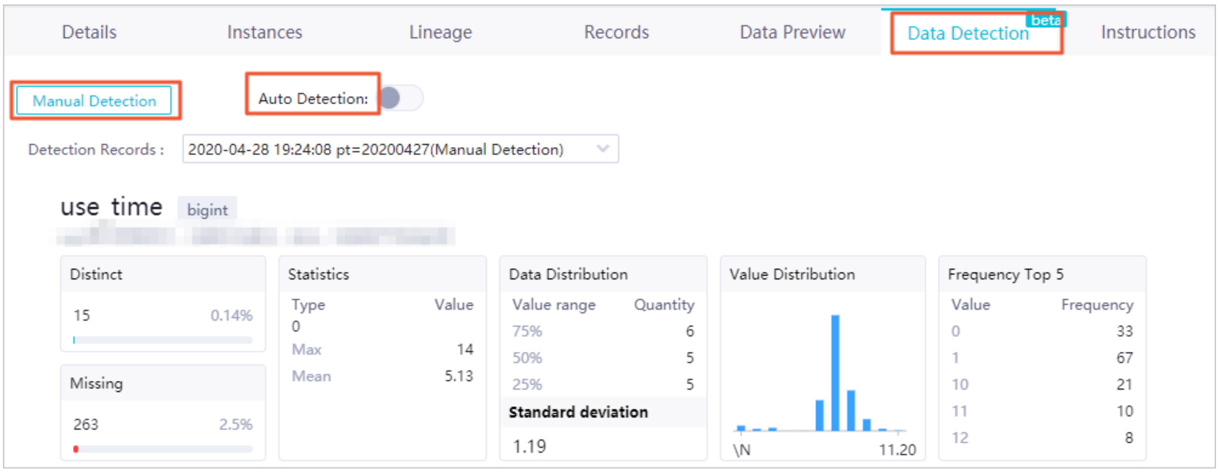
> **Notice** The data detection feature is in public preview and is supported only by the MaxCompute compute engine in the China (Shanghai) region.

DataWorks detects the data of a table based on the schema and a partition key value. The data detection results include basic statistics and data distribution.

The data detection feature has the following limits:

- You can detect only partitioned tables.
- You can detect only tables in the production environment.
- Only the table owner has the permission to enable automatic detection.

On the **Data Profiling** tab, you can set the detection mode and view detection records.



Detection Records : 2020-04-28 19:24:08 pt=20200427(Manual Detection)

use time bigint

[redacted]

Distinct	
15	0.14%

Missing	
263	2.5%

Statistics

Type	Value
0	
Max	14
Mean	5.13

Data Distribution

Value range	Quantity
75%	6
50%	5
25%	5

Standard deviation

1.19

Frequency Top 5

Value	Frequency
0	33
1	67
10	21
11	10
12	8

You can detect a table in a manual or an automatic manner:

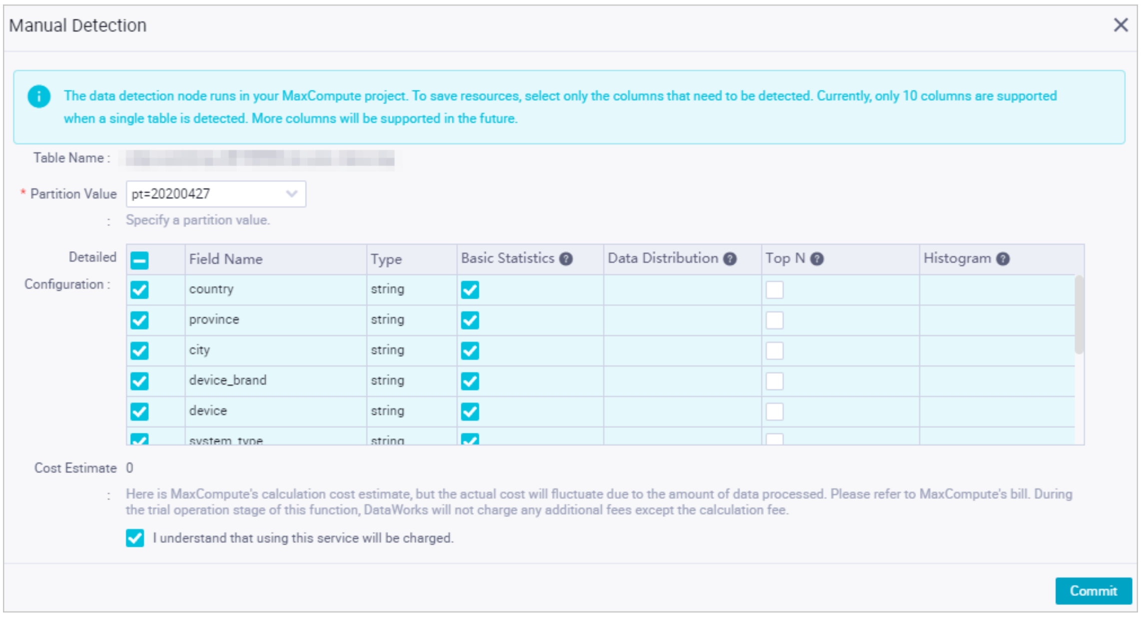
- Manual detection

> ? **Note** Data detection nodes are run in the MaxCompute project to which a detected table belongs. You can detect a maximum of 10 columns in a single table at a time. To save resources, select only the columns that need to be detected when you configure a manual detection node.

If you want to configure a manual detection node, perform the following steps:

i. On the **Data Profiling** tab, click **Manual Profiling**.

ii. In the **Manual Profiling** dialog box, set the parameters as required.



Manual Detection

The data detection node runs in your MaxCompute project. To save resources, select only the columns that need to be detected. Currently, only 10 columns are supported when a single table is detected. More columns will be supported in the future.

Table Name : [redacted]

* Partition Value : pt=20200427

Specify a partition value.

Detailed Configuration :

☐	Field Name	Type	Basic Statistics	Data Distribution	Top N	Histogram
☑	country	string	☑		☐	
☑	province	string	☑		☐	
☑	city	string	☑		☐	
☑	device_brand	string	☑		☐	
☑	device	string	☑		☐	
☑	system_type	string	☑		☐	

Cost Estimate : 0

Here is MaxCompute's calculation cost estimate, but the actual cost will fluctuate due to the amount of data processed. Please refer to MaxCompute's bill. During the trial operation stage of this function, DataWorks will not charge any additional fees except the calculation fee.

☑ I understand that using this service will be charged.

Parameter	Description
Table Name	The name of the table, in the format of Workspace name.Table name. The value of this parameter is generated by the system and cannot be modified.
Partition Value	The key value of the partition to be detected. Select a partition from the drop-down list.
Detailed Configuration	The columns to be detected. Select columns as needed.
Estimated Cost	The estimated cost of the detection node. The cost is estimated based on the preceding settings. **Notice** MaxCompute SQL statements may be executed to detect data. In this case, you will be charged for using the MaxCompute service. Note that the estimated cost is for reference only. The actual cost may vary with the volume of the detected data. Check your MaxCompute bill for the actual cost.

iii. Select **I understand that using this service will be charged**.

iv. Click **Commit**.

v. After the detection node is run, view the data detection results on the **Data Profiling** tab.

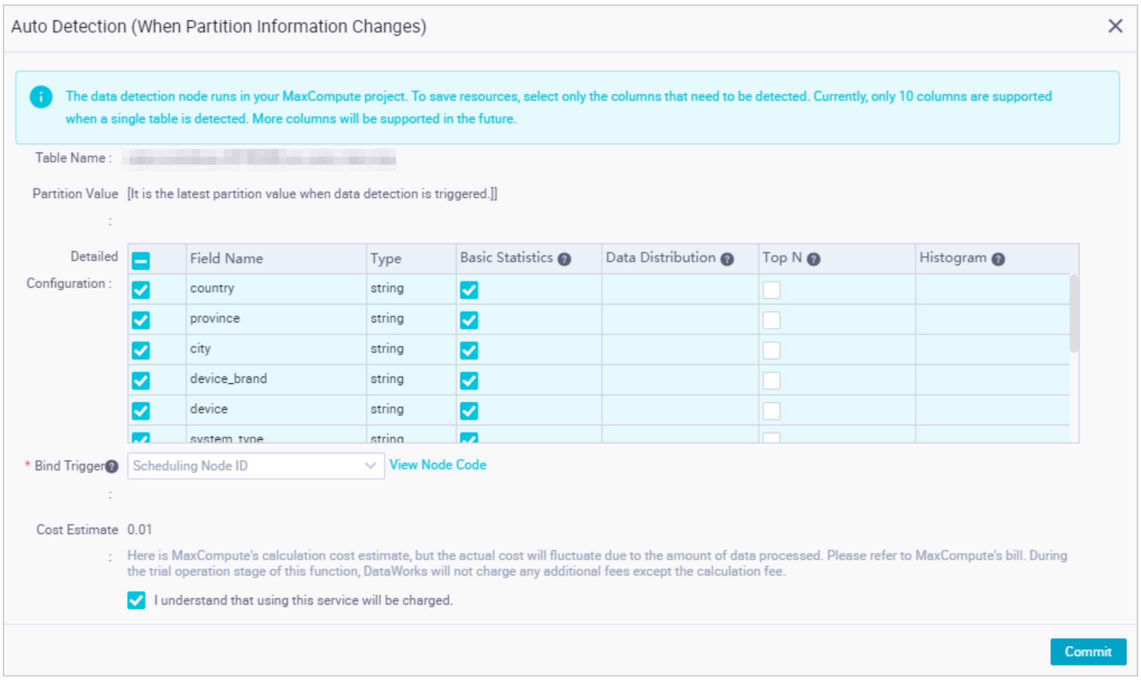
You can select a detection record from the **Profiling Records** drop-down list to view the data detection results.

- Automatic detection

  If you want to configure an automatic detection node, perform the following steps:

  i. Turn on **Auto Profiling**.

  ii. In the **Auto Profiling (When Partition Information Changes)** dialog box, set the parameters as required.



Auto Detection (When Partition Information Changes)

The data detection node runs in your MaxCompute project. To save resources, select only the columns that need to be detected. Currently, only 10 columns are supported when a single table is detected. More columns will be supported in the future.

Table Name : [redacted]

Partition Value : [It is the latest partition value when data detection is triggered.]]

Detailed Configuration :

☐	Field Name	Type	Basic Statistics	Data Distribution	Top N	Histogram
☑	country	string	☑		□	
☑	province	string	☑		□	
☑	city	string	☑		□	
☑	device_brand	string	☑		□	
☑	device	string	☑		□	
☑	system_type	string	☑		□	

* Bind Trigger : Scheduling Node ID View Node Code

Cost Estimate : 0.01

Here is MaxCompute's calculation cost estimate, but the actual cost will fluctuate due to the amount of data processed. Please refer to MaxCompute's bill. During the trial operation stage of this function, DataWorks will not charge any additional fees except the calculation fee.

☑ I understand that using this service will be charged.

Parameter	Description
Table Name	The name of the table, in the format of Workspace name.Table name. The value of this parameter is generated by the system and cannot be modified.
Partition Value	The latest partition key value when data detection is triggered. The value of this parameter is generated by the system and cannot be modified.
Detailed Configuration	The columns to be detected. Select columns as needed.
Bind Trigger	The scheduling node that triggers the detection node. Select a scheduling node from the drop-down list. You can view the IDs of scheduling nodes in **Operation Center**. We recommend that you select the output node of the current table. If you bind a scheduling node to an automatic detection node and commit the detection node, the system detects the latest partition after the scheduling node is run.
Estimated Cost	The estimated cost of the detection node. The cost is estimated based on the preceding settings. Notice MaxCompute SQL statements may be executed to detect data. In this case, you will be charged for using the MaxCompute service. Note that the estimated cost is for reference only. The actual cost may vary with the volume of the detected data. Check your MaxCompute bill for the actual cost.

iii. Select **I understand that using this service will be charged**.

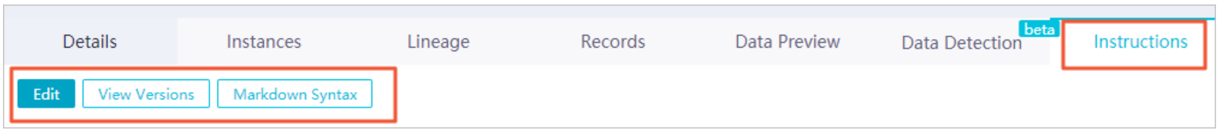
iv. Click **Commit**.

v. After the detection node is run, view the data detection results on the **Data Profiling** tab.

You can select a detection record from the **Profiling Records** drop-down list to view the data detection results.

View usage notes

On the **Usage Notes** tab, you can edit usage notes, check the historical versions of the usage notes, and view the Markdown syntax. You can also learn the relevant information based on the description of the data.



4.5.2. Apply for table permissions

This topic describes how to apply for table permissions on the Data Map or Security Center page.

Context

For tables in the workspaces that reside in China (Shanghai) or China (Beijing), you can apply for permissions on these tables on the **Security Center** page. For tables in the workspaces that reside in other regions, you can apply for permissions on these tables on the **Data Map** page.

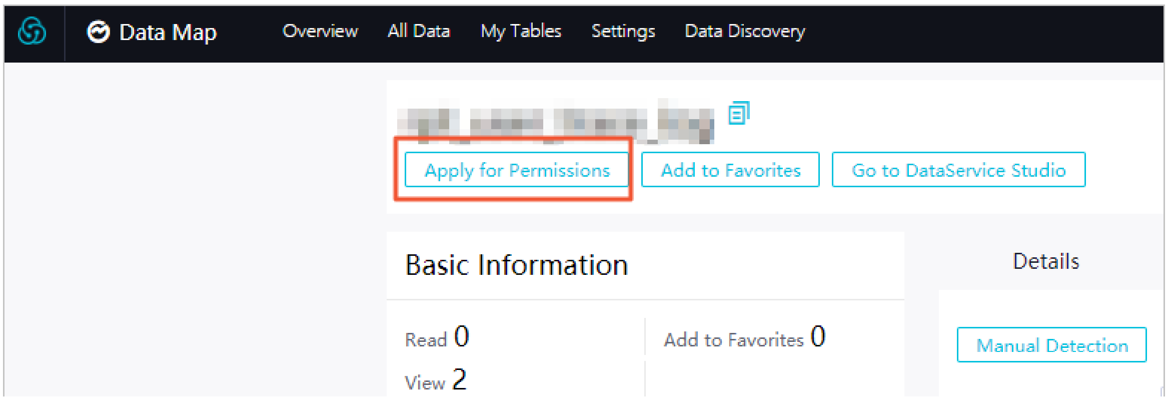
Go to the details page of a table

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 where the target workspace resides. Find the target workspace and click **Data Analytics** in the Actions column.
4. On the DataStudio page, click in the upper-left corner and choose **All Products > DataMap**.

   The homepage of Data Map appears.
5. In the top navigation bar, click **All Data**.
6. On the All Data page, click a tab as required, such as MaxCompute.
7. On the tab that appears, click the name of the table on which you want to apply for permissions.

Apply for table permissions on the Security Center page

1. On the table details page, click **Apply for Permission**.



Detection Records :

> **Note** If a table is hidden, the **Apply for Permissions** button does not appear.

2. On the **Table Permission Request** page that appears, set the parameters as required. The following table describes the configuration parameters.

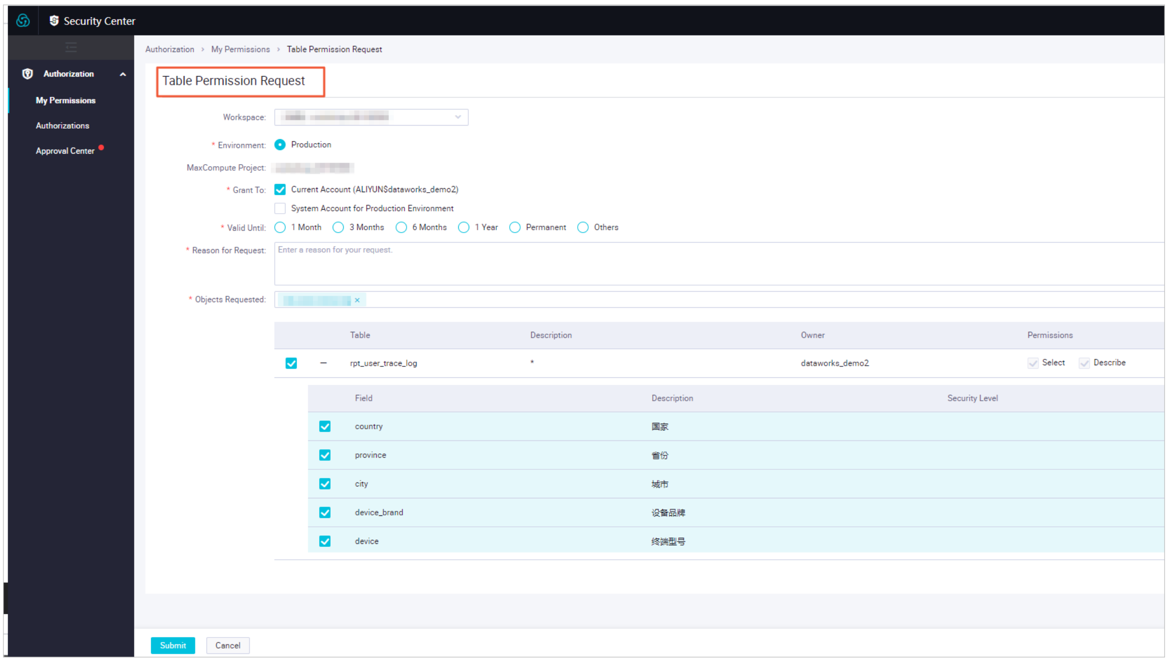


Table Permission Request

Workspace: [redacted]

* Environment: ● Production

MaxCompute Project: [redacted]

* Grant To: ☑ Current Account (ALIYUN$dataworks_demo2)
□ System Account for Production Environment

* Valid Until: ○ 1 Month ○ 3 Months ○ 6 Months ○ 1 Year ○ Permanent ○ Others

* Reason for Request: Enter a reason for your request.

* Objects Requested: [redacted] ×

		Table	Description	Owner	Permissions
☑	–	rpt_user_trace_log	*	dataworks_demo2	✓ Select ✓ Describe

	Field	Description	Security Level
☑	country	国家	
☑	province	省份	
☑	city	城市	
☑	device_brand	设备品牌	
☑	device	终端型号	

Parameter	Description
Workspace	The workspace where the target table resides.
Environment	The environment in which the permissions are used. For a workspace in the standard mode, the valid values are **Development** and **Production** . For a workspace in the basic mode, the valid value is **Production**.
MaxCompute Project	The name of the MaxCompute project associated with the DataWorks workspace. The value is automatically generated and you cannot change the value.
Grant To	The account to which the permissions are granted. Valid values: **Current Account** and **System Account for Production Environment**.
Valid Until	The validity period of the permissions on the table. Valid values: **1 Month**, **3 Months**, **6 Months**, **1 Year**, **Permanent**, and **Others**.
Reason for Request	The reason for applying for permissions. Enter a brief reason for faster approval.
Objects Requested	The table on which you want to apply for permissions and the permissions you apply for.

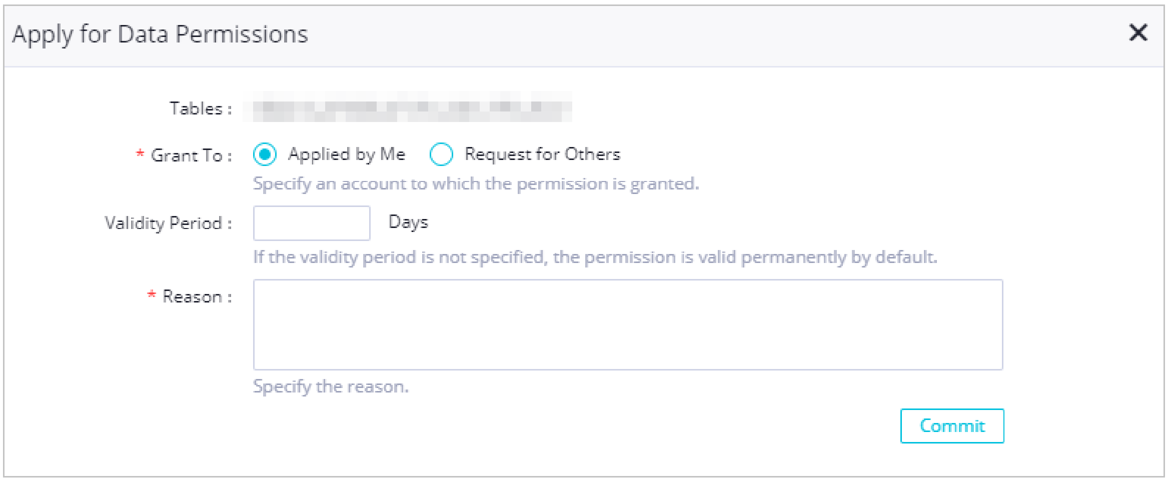
3. Click **Submit**.

Apply for table permissions on the Data Map page

1. On the table details page, click **Apply for Permission**.

   > **Note** If a table is hidden, the **Apply for Permissions** button does not appear.

2. In the **Apply for Permission** dialog box that appears, set the parameters as required. The following table describes the configuration parameters.



Apply for Data Permissions

Tables : [redacted]

* Grant To : ● Applied by Me ○ Request for Others

Specify an account to which the permission is granted.

Validity Period : Days

If the validity period is not specified, the permission is valid permanently by default.

* Reason :

Specify the reason.

Parameter	Description
Table	The name of the table on which you want to apply for permissions. The value is automatically generated and you cannot change the value.
Grant To	The account to which the permissions are granted. Valid values: **Current Account** and **Specified Account**.
Username	The account for which you want to apply for table permissions. **Note** This parameter is valid only when you set **Grant To** to **Specified Account**.
Validity Period	The validity period of the permissions on the table. If you do not set this parameter, the permissions are permanently valid.
Reason	The reason for applying for permissions. Enter a brief reason for faster approval.

3. Click **Commit**.

View the application status

1. On the DataStudio page, click ☰ in the upper-left corner and choose **All Products** > **DataMap**.

   The homepage of Data Map appears.

2. In the top navigation bar, click **My Data**.
3. On the My Data page, click **Permission Management** in the left-side navigation pane.

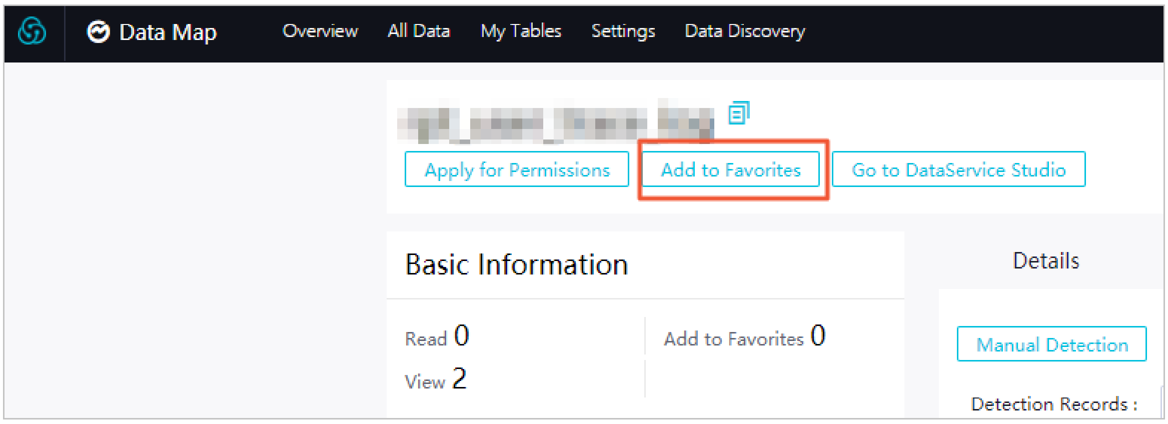
4. On the page that appears, click the **Submitted by Me** tab.
5. Find the target application record and click **View** in the Actions column. The application details appears.

4.5.3. Add a table to favorites

This topic describes how to add a table to or remove it from favorites, and view the tables added to favorites.

Procedure

1. Go to the details page of a table.
   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 where the target workspace resides. Find the target workspace and click **Data Analytics** in the Actions column.
   iv. On the DataStudio page, click in the upper-left corner and choose **All Products > DataMap**. The homepage of Data Map appears.
   v. In the top navigation bar, click **All Data**.
   vi. On the **All Data** page, click a tab as required.
   vii. On the tab that appears, click the name of the table you want to add to favorites.
2. On the table details page, click **Add to Favorites**.



3. In the top navigation bar, click **My Data**.
4. On the **My Data** page, click **My Favorites** in the left-side navigation pane. On the page that appears, you can view all the tables that you have added to favorites and remove tables from favorites. To remove a table from favorites, find the target table and click **Remove from Favorites** in the Actions column.

4.5.4. Go to DataService Studio to create API operations

This topic describes how to go to DataService Studio from the details page of a table and create API operations.

Procedure

1. Go to the details page of a table.
    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 where the target workspace resides. Find the target workspace and click **Data Analytics** in the Actions column.
    iv. On the DataStudio page, click in the upper-left corner and choose **All Products > DataMap**. The homepage of Data Map appears.
    v. In the top navigation bar, click **All Data**.
    vi. On the **All Data** page, click a tab as required.
    vii. On the tab that appears, click the name of the table based on which you want to create API operations in DataService Studio.
2. On the details page of the table, click **Create API in DataService Studio**.
3. On the **DataService Studio** page that appears, create API operations based on tables or register existing API operations as required. For more information, see Overview.

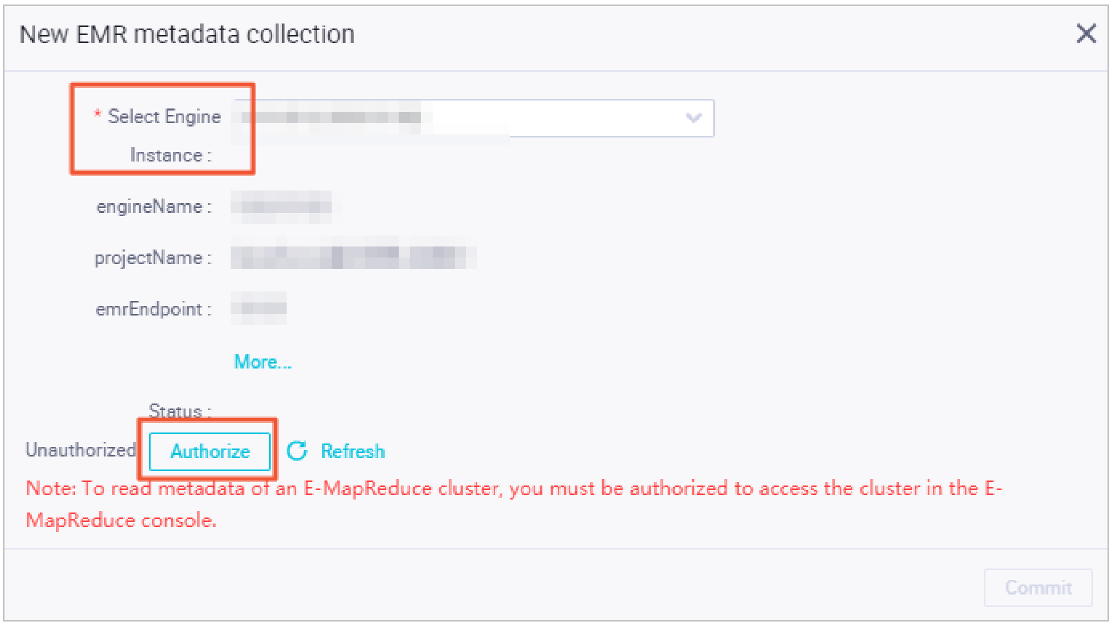
4.6. Data discovery

4.6.1. Collect metadata from an E-MapReduce data store

This topic describes how to create a crawler to collect metadata from an E-MapReduce data store to DataWorks. You can view the collected metadata on the Data Map page.

Procedure

1. Go to the **Data Discovery** page.
    i. Log on to the DataWorks console.
    ii. In the left-side navigation pane, click **Workspaces**.
    iii. In the top navigation bar, select a region as required, find the workspace where you want to create a crawler, and then click **Data Analytics** in the Actions column.
    iv. On the DataStudio page, click the ☰ icon in the upper-left corner and choose **All Products > Data governance > DataMap**. The homepage of Data Map appears.
    v. In the top navigation bar, click **Data Discovery**.
2. In the left-side navigation pane, click E-MapReduce. On the **Obtain Metadata from E-MapReduce** page, click **Create Crawler**.
3. In the **Create Crawler** dialog box, select an engine instance from the **Select Engine Instance** drop-down list and click **Authorize**.



New EMR metadata collection

* Select Engine Instance : [redacted]

engineName : [redacted]

projectName : [redacted]

emrEndpoint : [redacted]

More...

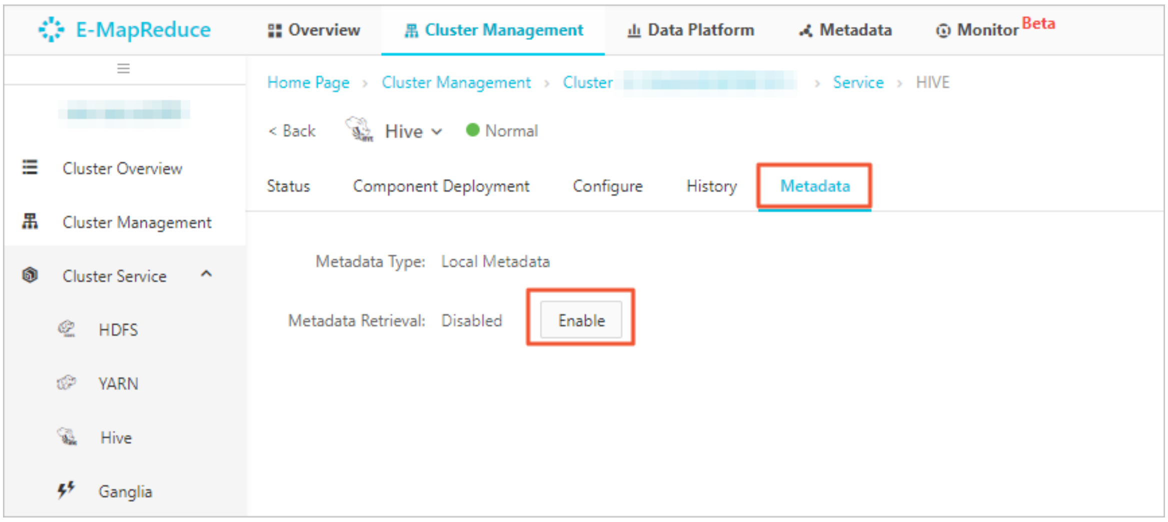
Status :

Unauthorized Authorize Refresh

Note: To read metadata of an E-MapReduce cluster, you must be authorized to access the cluster in the E-MapReduce console.

Commit

4. On the page that appears, click the **Metadata** tab and click **Enable**.



Metadata Type: Local Metadata

Metadata Retrieval: Disabled Enable

5. In the **Confirm Operation** message, click **OK**.
6. Return to the **Create Crawler** dialog box on the **Data Map** page and click **Refresh**.
7. After the authorization status changes to **Authorized**, click **Commit**.
8. On the **Obtain Metadata from E-MapReduce** page, find the created crawler and click **Obtain All** in the Actions column.Click **Refresh** in the upper-right corner of the page and verify that the value in the **Running Status** column of the created crawler changes to **The data has been collected**.

   > Note After full metadata is collected from the E-MapReduce data store, the system automatically synchronizes new metadata from the data store.

   If you want to delete the created crawler, click **Delete** in the Actions column. In the **Delete Instance** message, click **OK**.

9. View the metadata collected from the E-MapReduce data store.

   i. In the top navigation bar, click **All Data**

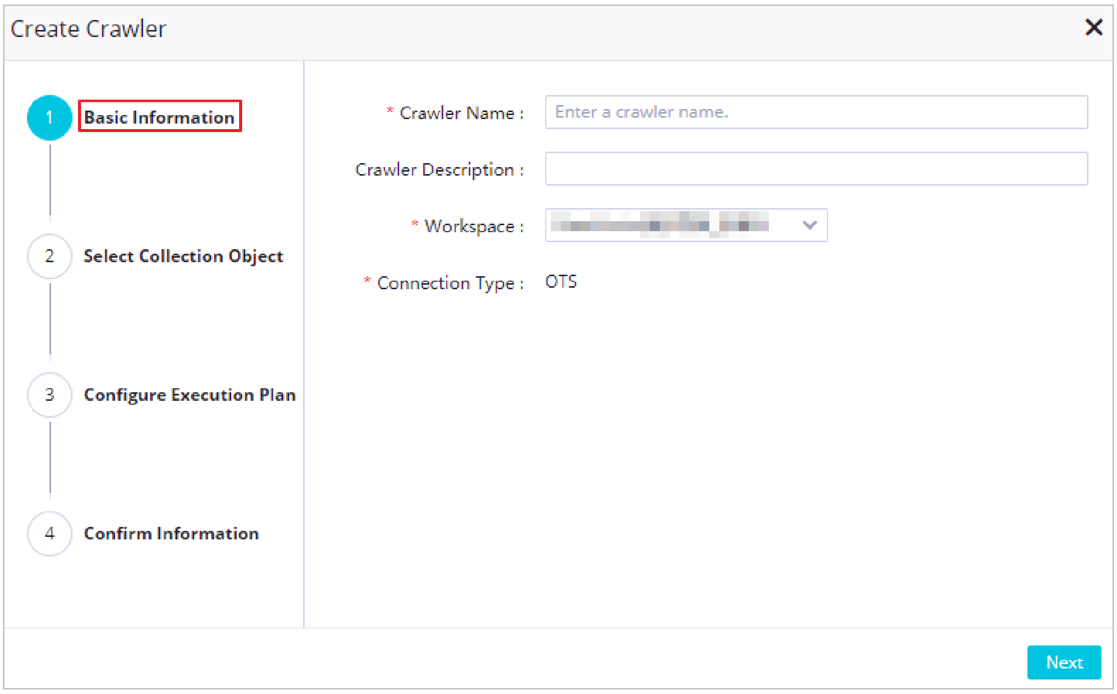
   i. In the top navigation bar, click **All Data**.
   ii. Click the **E-MapReduce** tab.
   iii. On the **E-MapReduce** tab, click the name of the table that stores the collected metadata and view the table details.

4.6.2. Collect Tablestore metadata

You can collect the table schema and lineage to DataMap, which displays the inner structure and association relationships of a table. This topic describes how to create a crawler and collect Tablestore metadata to DataWorks. You can view the collected metadata on the DataMap page.

1. Go to the **Data Discovery** page.
   i. Log on to the DataWorks console.
   ii. In the left-side navigation pane, click **Workspaces**.
   iii. In the top navigation bar, select a region as required, find the workspace where you want to create a crawler, and then click **Data Analytics** in the Actions column.
   iv. On the DataStudio page, click the ☰ icon in the upper-left corner and choose **All Products > Data governance > DataMap**. The homepage of Data Map appears.
   v. In the top navigation bar, click **Data Discovery**.
2. In the left-side navigation pane, click **OTS**.
3. On the **OTSMetadata Crawler** page, click **Create Crawler**.
4. In the **Create Crawler** dialog box, perform the following steps:

i. In the **Basic Information** step, configure the parameters.



Create Crawler

1 Basic Information

2 Select Collection Object

3 Configure Execution Plan

4 Confirm Information

* Crawler Name : Enter a crawler name.

Crawler Description :

* Workspace : [redacted]

* Connection Type : OTS

Parameter	Description
Crawler Name	Required. This parameter specifies the name of the crawler. You must specify a unique name.
Crawler Description	The description of the crawler.
Workspace	The workspace of the data source from which you want to collect metadata.
Connection Type	The type of data source from which you want to collect metadata. The default value is **OTS** and cannot be changed.

ii. Click **Next**.

iii. In the **Select Collection Object** step, select a data source from the **Connection** drop-down list.If the drop-down list does not contain the data source that you need, click **Create Connection** to go to the **data source** page to create a data source. For more information, see Configure a Tablestore connection.

iv. Click **Test Crawler Connectivity**.If the database is configured with a whitelist, you must add the IP address of the workspace based on the region where the workspace resides to the whitelist.

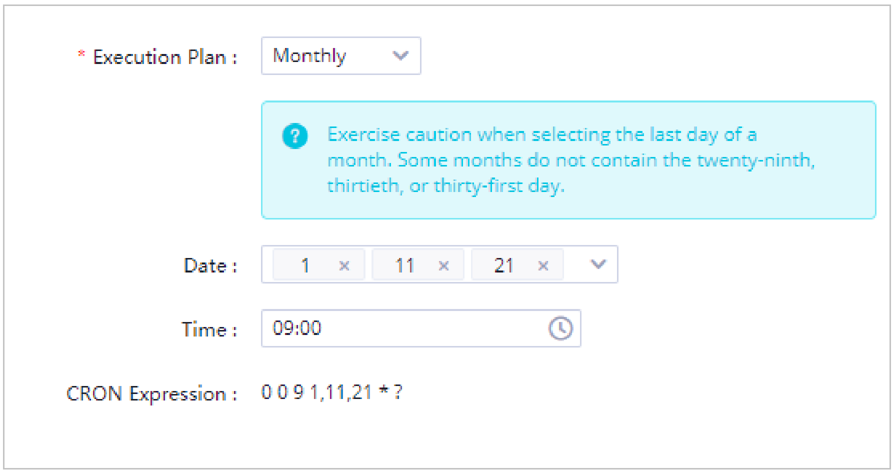
v. If the message **The connectivity test has been passed** appears, click **Next**.If the message **The connectivity test failed** appears, check whether you have configured a valid data source.

vi. In the **Configure Execution Plan** step, specify **Execution Plan**.Valid values of the **Execution Plan** parameter: **On-demand Execution**, **Monthly**, **Weekly**, **Daily**, and **Hourly**. Different running plans are generated based on different running cycles. The system collects Tablestore metadata of the data source based on the running cycle that you specify. The following descriptions explain each value and provide examples:

- On-demand Execution: The system collects Tablestore metadata based on your business requirements.
- Monthly: The system automatically collects Tablestore metadata each day on several specific days of each month.

  > **Notice** Some months do not have the 29th, 30th, or 31st day, so the system does not collect Tablestore metadata on these dates. We recommend that you do not select dates at the end of months.

  The following figure shows that the system collects Tablestore metadata once a day at 09:00 on the 1st, 11th, and 21th day of each month. **CRON Expression** is automatically generated based on the values of Date and Time.

  * Execution Plan : Monthly

  Exercise caution when selecting the last day of a month. Some months do not contain the twenty-ninth, thirtieth, or thirty-first day.

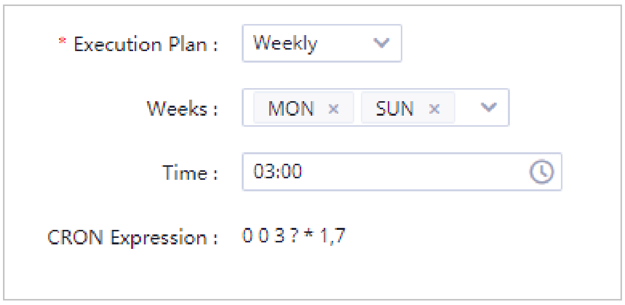
  Date : 1 × 11 × 21 ×

  Time : 09:00

  CRON Expression : 0 0 9 1,11,21 * ?

- Weekly: The system automatically collects Tablestore metadata once a day at a specific time point several days a week.

  The following figure shows that the system automatically collects Tablestore metadata at 03:00 on Sunday and Monday of each week.

  * Execution Plan : Weekly

  Weeks : MON × SUN ×

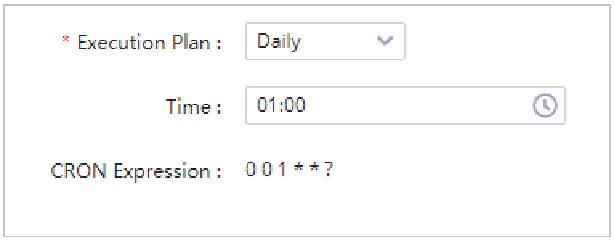
  Time : 03:00

  CRON Expression : 0 0 3 ? * 1,7

  If the **Time** parameter is not specified, the system automatically collects Tablestore metadata at 00:00:00 on the specific days of each week.

- Daily: The system automatically collects Tablestore metadata at a specific time point of each day.

  The following figure shows that the system automatically collects Tablestore metadata at 01:00 each day.

  * Execution Plan : Daily

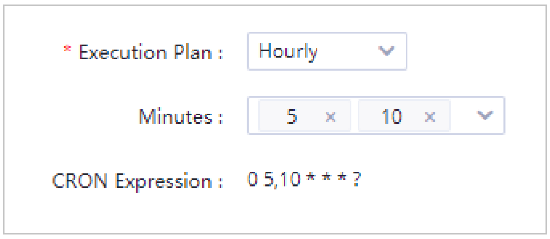
  Time : 01:00

  CRON Expression : 0 0 1 * * ?

- Hourly: The system automatically collects Tablestore metadata at `N × 5 minutes` each hour.

  > **Note** For a Tablestore metadata collection task that runs each hour, you can select a time point that is a multiple of five minutes.

  The following figure shows that the system automatically collects Tablestore metadata at the 5th and 10th minute of each hour.

  * Execution Plan : Hourly

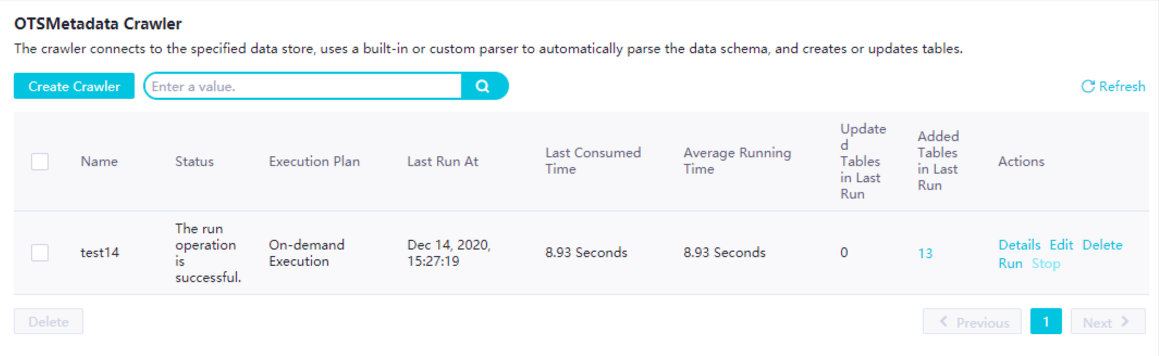
  Minutes : 5 × 10 ×

  CRON Expression : 0 5,10 * * * ?

vii. Click **Next**.

viii. In the **Confirm Information** step, check the information that you specified and click **OK**.

5. On the **OTSMetadata Crawler** page, you can view the information about your crawler and manage your crawler.



OTSMetadata Crawler

The crawler connects to the specified data store, uses a built-in or custom parser to automatically parse the data schema, and creates or updates tables.

□	Name	Status	Execution Plan	Last Run At	Last Consumed Time	Average Running Time	Updated Tables in Last Run	Added Tables in Last Run	Actions
□	test14	The run operation is successful.	On-demand Execution	Dec 14, 2020, 15:27:19	8.93 Seconds	8.93 Seconds	0	13	Details Edit Delete Run Stop

The following description shows the information that you can view and the operations that you can perform:

- You can view **Status**, **Execution Plan**, **Last Run At**, **Last Consumed Time**, **Average Running Time**, Updated Tables in Last Run, and Added Tables in Last Run of your crawler.
- You can click **Details**, **Edit**, **Delete**, **Run**, or **Stop** in the **Actions** column to perform the required operations.
  - **Details**: View **Crawler Name**, **Connection Type**, and **Execution Plan** of the crawler.

- **Edit**: Modify the configurations of the crawler.
- **Delete**: Delete the crawler.
- **Run**: **Run** the task to collect Tablestore metadata. The **Run** entry is available only when **Execution Plan** is set to **On-demand Execution**.
- **Stop**: Stop running the crawler.

Result

After Tablestore metadata is collected, switch back to the previous page and click **All Data** in the top navigation bar. On the page that appears, click the **OTS** tab in the upper part. On the OTS tab, you can view the table that stores the collected Tablestore metadata.

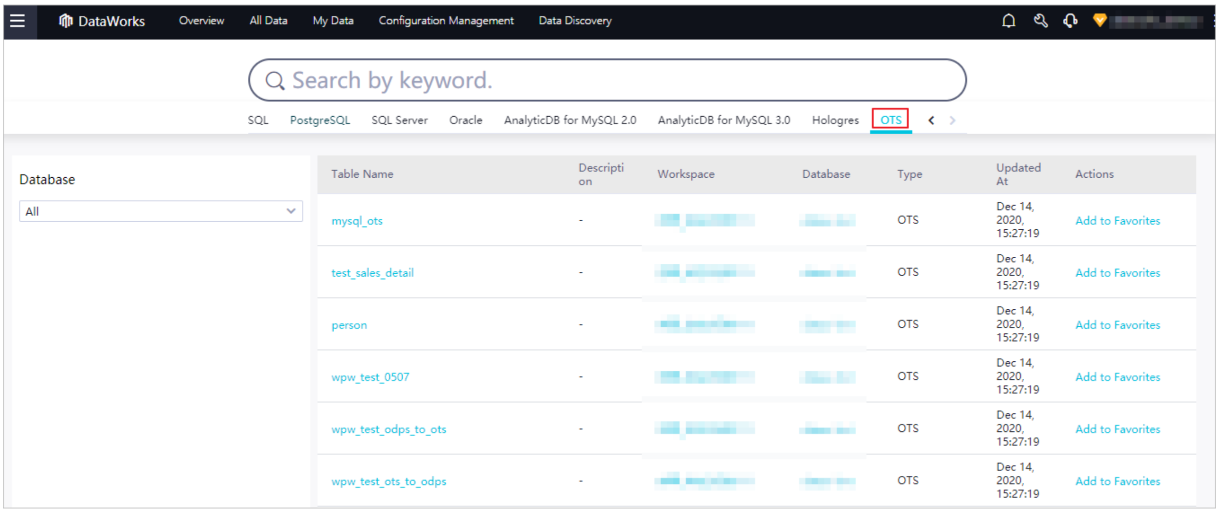
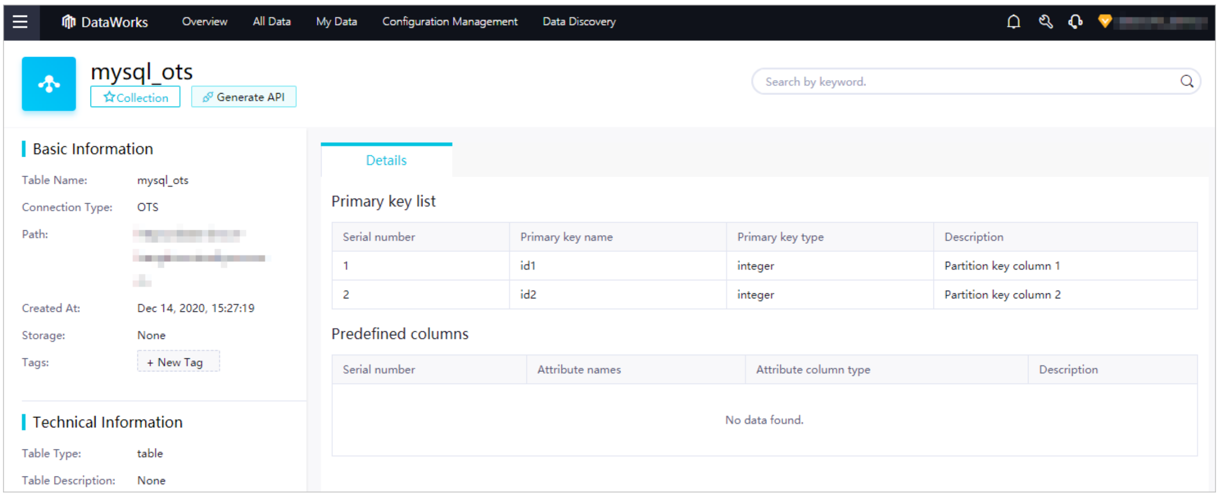


Table Name	Description	Workspace	Database	Type	Updated At	Actions
mysql_ots	-	[redacted]	[redacted]	OTS	Dec 14, 2020, 15:27:19	Add to Favorites
test_sales_detail	-	[redacted]	[redacted]	OTS	Dec 14, 2020, 15:27:19	Add to Favorites
person	-	[redacted]	[redacted]	OTS	Dec 14, 2020, 15:27:19	Add to Favorites
wpw_test_0507	-	[redacted]	[redacted]	OTS	Dec 14, 2020, 15:27:19	Add to Favorites
wpw_test_odps_to_ots	-	[redacted]	[redacted]	OTS	Dec 14, 2020, 15:27:19	Add to Favorites
wpw_test_ots_to_odps	-	[redacted]	[redacted]	OTS	Dec 14, 2020, 15:27:19	Add to Favorites

Click **table name**, **workspace**, or **database** to view the related details.

Example 1: View the details of the *mysql_ots* table.



Basic Information

Table Name: mysql_ots
Connection Type: OTS
Path: [redacted]
Created At: Dec 14, 2020, 15:27:19
Storage: None
Tags: + New Tag

Technical Information

Table Type: table
Table Description: None

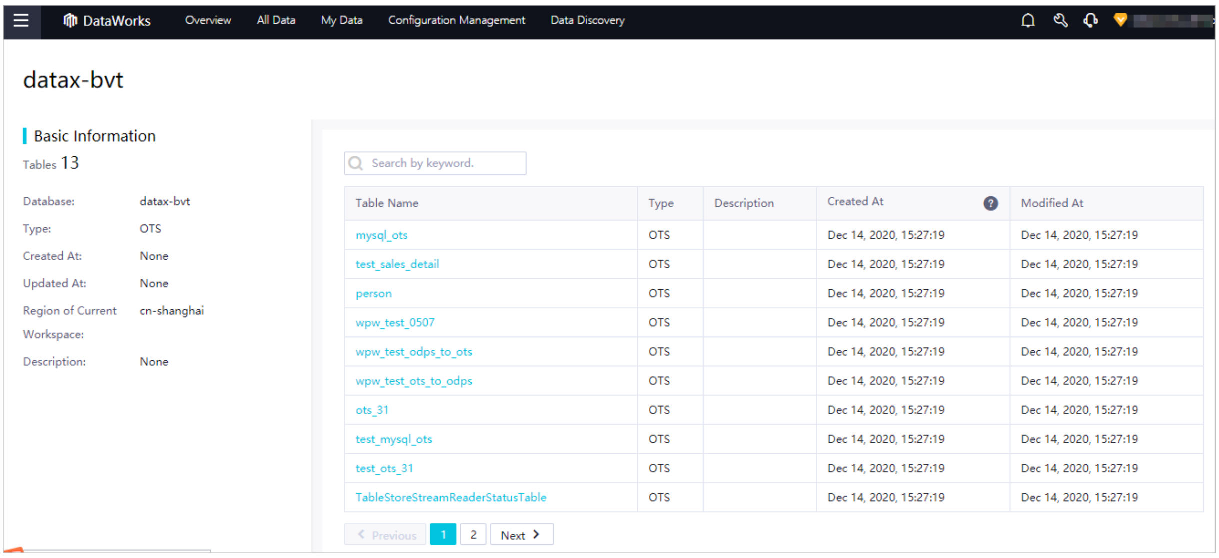
Primary key list

Serial number	Primary key name	Primary key type	Description
1	id1	integer	Partition key column 1
2	id2	integer	Partition key column 2

Predefined columns

Serial number	Attribute names	Attribute column type	Description
No data found.			

Example 2: View all tables in the *datax-bvt* database.



datax-bvt

Basic Information

Tables 13

Database:	datax-bvt
Type:	OTS
Created At:	None
Updated At:	None
Region of Current Workspace:	cn-shanghai
Description:	None

Search by keyword.

Table Name	Type	Description	Created At	Modified At
mysql_ots	OTS		Dec 14, 2020, 15:27:19	Dec 14, 2020, 15:27:19
test_sales_detail	OTS		Dec 14, 2020, 15:27:19	Dec 14, 2020, 15:27:19
person	OTS		Dec 14, 2020, 15:27:19	Dec 14, 2020, 15:27:19
wpw_test_0507	OTS		Dec 14, 2020, 15:27:19	Dec 14, 2020, 15:27:19
wpw_test_odps_to_ots	OTS		Dec 14, 2020, 15:27:19	Dec 14, 2020, 15:27:19
wpw_test_ots_to_odps	OTS		Dec 14, 2020, 15:27:19	Dec 14, 2020, 15:27:19
ots_31	OTS		Dec 14, 2020, 15:27:19	Dec 14, 2020, 15:27:19
test_mysql_ots	OTS		Dec 14, 2020, 15:27:19	Dec 14, 2020, 15:27:19
test_ots_31	OTS		Dec 14, 2020, 15:27:19	Dec 14, 2020, 15:27:19
TableStoreStreamReaderStatusTable	OTS		Dec 14, 2020, 15:27:19	Dec 14, 2020, 15:27:19

< Previous 1 2 Next >

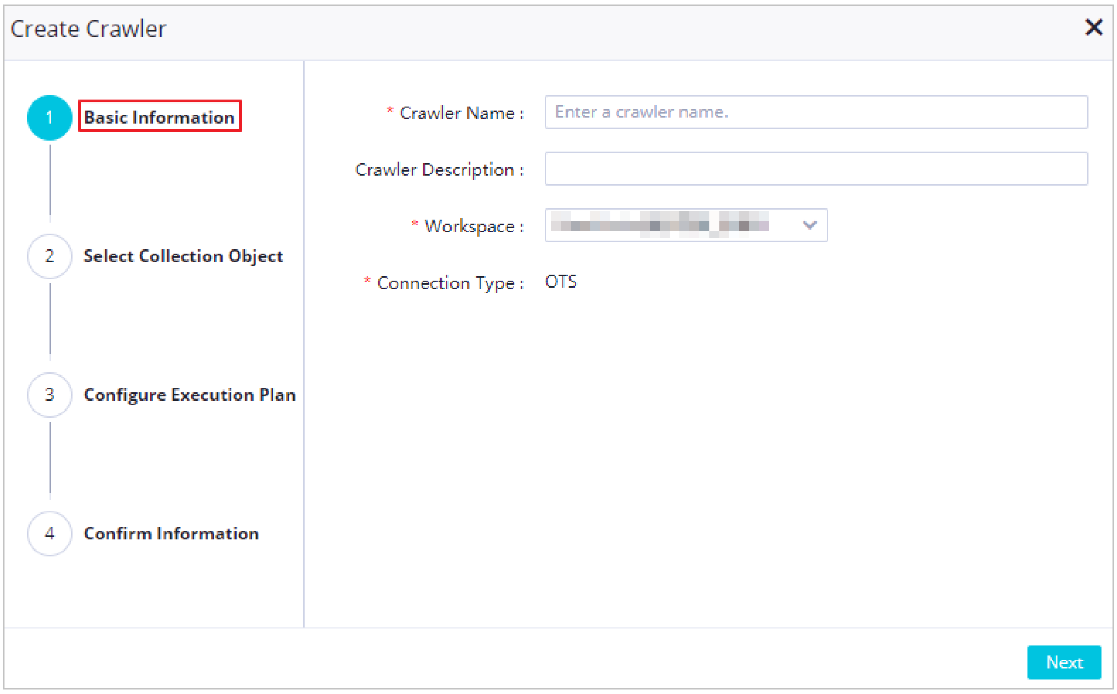
4.6.3. Collect metadata from a MySQL data store

This topic describes how to create a crawler to collect metadata from a MySQL data store to DataWorks. You can view the collected metadata on the Data Map page.

Procedure

1. Go to the **Data Discovery** page.
    i. Log on to the DataWorks console.
    ii. In the left-side navigation pane, click **Workspaces**.
    iii. In the top navigation bar, select a region as required, find the workspace where you want to create a crawler, and then click **Data Analytics** in the Actions column.
    iv. On the DataStudio page, click the ☰ icon in the upper-left corner and choose **All Products > Data governance > DataMap**. The homepage of Data Map appears.
    v. In the top navigation bar, click **Data Discovery**.
2. In the left-side navigation pane, click **MySQL**.
3. On the **MySQLMetadata Crawler** page, click **Create Crawler**.
4. In the **Create Crawler** dialog box, perform the following steps:

i. In the **Basic Information** step, set the basic parameters.



* Crawler Name : Enter a crawler name.

Crawler Description :

* Workspace : [redacted]

* Connection Type : OTS

Parameter	Description
Crawler Name	Required. The name of the crawler. You must specify a unique name.
Crawler Description	The description of the crawler.
Workspace	The workspace of the data store from which metadata will be collected.
Connection Type	The type of the data store from which metadata will be collected. The default value is **MySQL** and cannot be changed.

ii. Click **Next**.

iii. In the **Select Collection Object** step, select a connection from the **Connection** drop-down list.If the required connection does not exist, click **Create Connection** to go to the **Data Source** page in **Workspace Management** and create the connection. For more information, see Configure a MySQL data source.

> **Note** You can select a connection to an Apsara RDS for MySQL database or a MySQL database that is accessible from the Internet by using a Java Database Connectivity (JDBC) connection string.

iv. Click **Test Crawler Connectivity**.If the database is configured with a whitelist, you must add the IP address of DataWorks based on the region where the workspace resides to the whitelist.

v. When the message **The connectivity test has been passed** appears, click **Next**.

vi. In the **Configure Execution Plan** step, specify the execution plan.Valid values of the **Execution Plan** parameter: **On-demand Execution**, **Monthly**, **Weekly**, **Daily**, and **Hourly**.

vii. Click **Next**.

viii. In the **Confirm Information** step, verify that the configuration of the crawler is correct and click **OK**.

5. On the **MySQLMetadata Crawler** page, find the created crawler and click **Run** in the Actions column.After the crawler is run, click the number in the **Updated Tables in Last Run** or **Added Tables in Last Run** column to view the details about the updated or added tables.

   You can also perform the following operations on the OSSMetadata Crawler page:

   - Click **Details** in the Actions column of a crawler. In the **Crawler Details** dialog box, view the detailed information about the crawler.
   - Click **Edit** in the Actions column of a crawler. In the **Edit Crawler** dialog box, modify the configuration of the crawler.
   - Click **Delete** in the Actions column of a crawler. In the **Confirm** message, click **OK** to delete the crawler.
   - Click **Stop** in the **Actions** column of a running crawler to stop the crawler.

Result

After the crawler is run, you can go to the **MySQL** tab of the **All Data** page and view the table that stores the collected metadata.

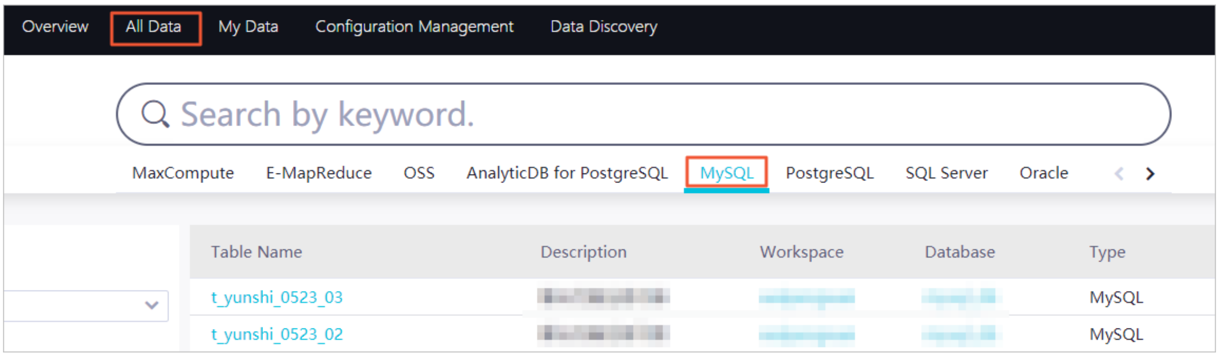
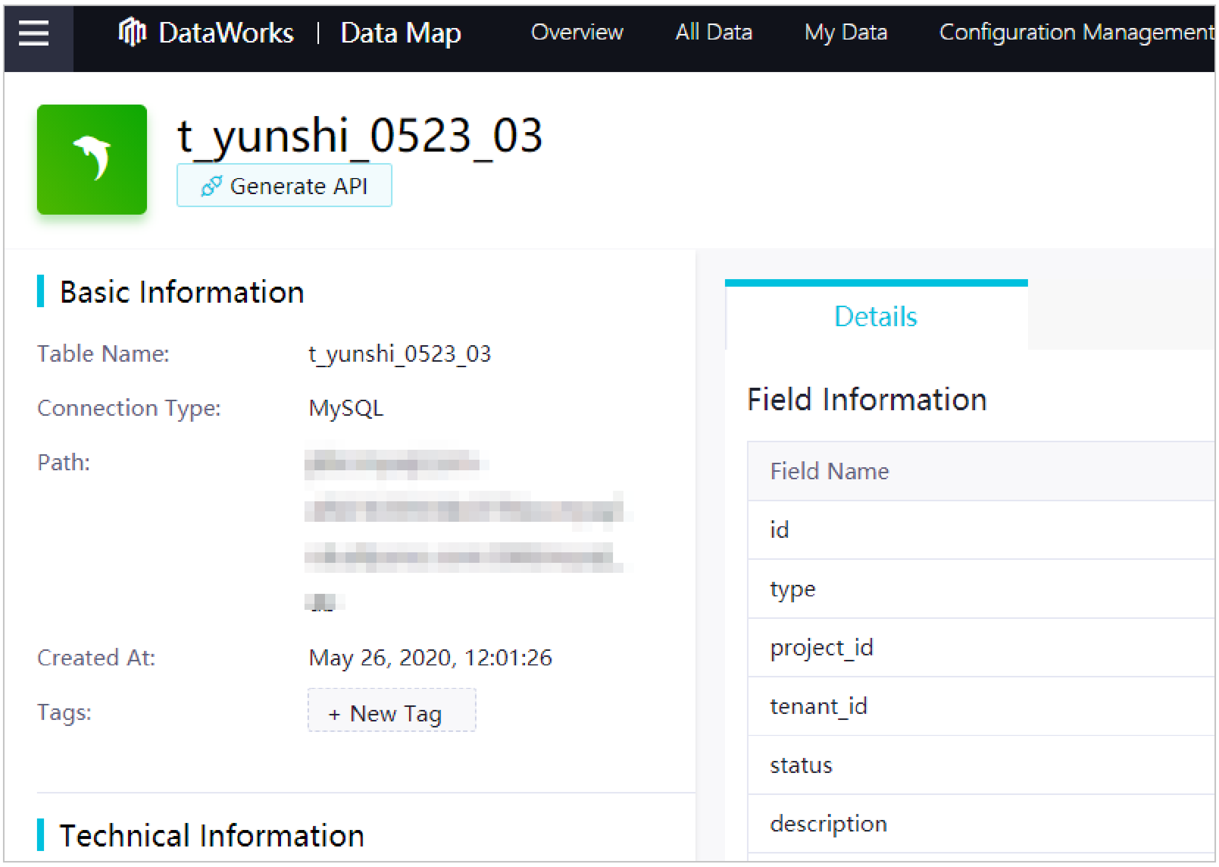


Table Name	Description	Workspace	Database	Type
t_yunshi_0523_03	[redacted]	[redacted]	[redacted]	MySQL
t_yunshi_0523_02	[redacted]	[redacted]	[redacted]	MySQL

Then, you can click the name of the table to view the table details.



t_yunshi_0523_03

Basic Information

Table Name: t_yunshi_0523_03

Connection Type: MySQL

Path: [redacted]

Created At: May 26, 2020, 12:01:26

Tags: + New Tag

Technical Information

Field Information

Field Name
id
type
project_id
tenant_id
status
description

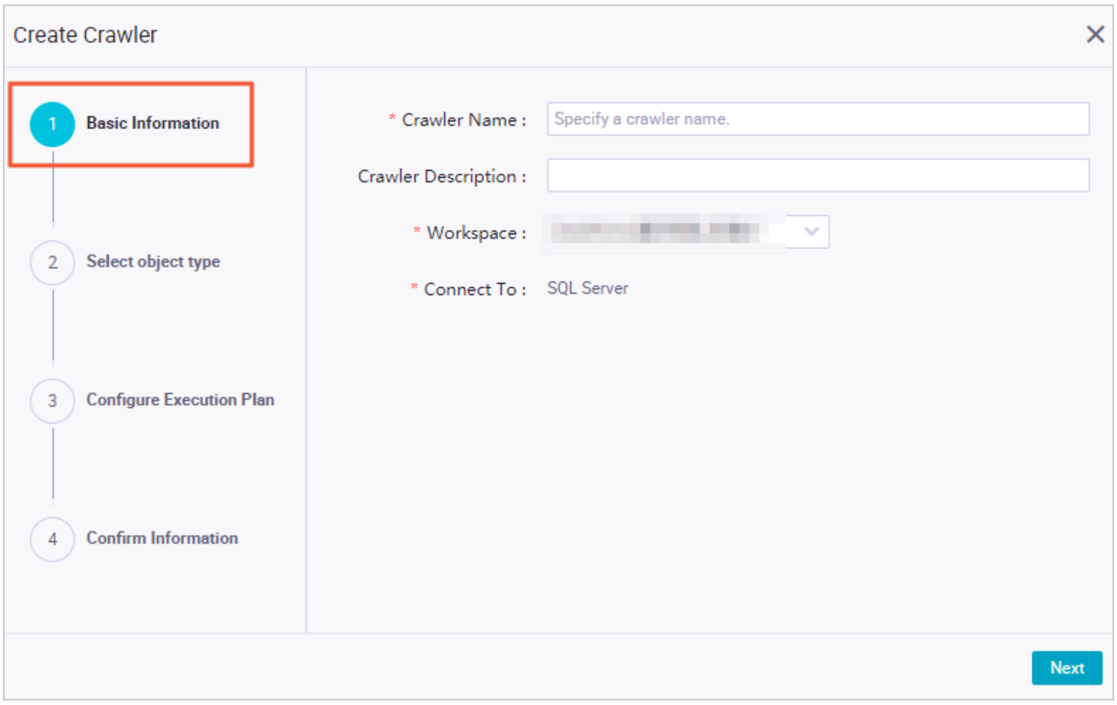
4.6.4. Collect metadata from an SQL Server data store

This topic describes how to create a crawler to collect metadata from an SQL Server data store to DataWorks. You can view the collected metadata on the Data Map page.

Procedure

1. Go to the **Data Discovery** page.
   i. Log on to the DataWorks console.
   ii. In the left-side navigation pane, click **Workspaces**.
   iii. In the top navigation bar, select a region as required, find the workspace where you want to create a crawler, and then click **Data Analytics** in the Actions column.
   iv. On the DataStudio page, click the ≡ icon in the upper-left corner and choose **All Products > Data governance > DataMap**. The homepage of Data Map appears.
   v. In the top navigation bar, click **Data Discovery**.
2. In the left-side navigation pane, click **SQL Server**.
3. On the **SQL ServerMetadata Crawler** page, click **Create Crawler**.
4. In the **Create Crawler** dialog box, perform the following steps:

i. In the **Basic Information** step, set the basic parameters.



* Crawler Name : Specify a crawler name.

Crawler Description :

* Workspace : [redacted]

* Connect To : SQL Server

Parameter	Description
Crawler Name	Required. The name of the crawler. You must specify a unique name.
Crawler Description	The description of the crawler.
Workspace	The workspace of the data store from which metadata will be collected.
Connection Type	The type of the data store from which metadata will be collected. The default value is **SQL Server** and cannot be changed.

ii. Click **Next**.

iii. In the **Select Collection Object** step, select a connection from the **Connection** drop-down list.If the required connection does not exist, click **Create Connection** to go to the **Data Source** page in **Workspace Management** and create the connection. For more information, see Configure an SQL Server connection.

iv. Click **Test Crawler Connectivity**.If the database is configured with a whitelist, you must add the IP address of DataWorks based on the region where the workspace resides to the whitelist.

v. When the message **The connectivity test has been passed** appears, click **Next**.

vi. In the **Configure Execution Plan** step, specify the execution plan.Valid values of the **Execution Plan** parameter: **On-demand Execution**, **Monthly**, **Weekly**, **Daily**, and **Hourly**.

vii. Click **Next**.

viii. In the **Confirm Information** step, verify that the configuration of the crawler is correct and click **OK**.

5. On the **SQL ServerMetadata Crawler** page, find the created crawler and click **Run** in the Actions column.After the crawler is run, click the number in the **Updated Tables in Last Run** or **Added Tables in Last Run** column to view the details about the updated or added tables.

   You can also perform the following operations on the OSSMetadata Crawler page:

   - Click **Details** in the Actions column of a crawler. In the **Crawler Details** dialog box, view the detailed information about the crawler.
   - Click **Edit** in the Actions column of a crawler. In the **Edit Crawler** dialog box, modify the configuration of the crawler.
   - Click **Delete** in the Actions column of a crawler. In the **Confirm** message, click **OK** to delete the crawler.
   - Click **Stop** in the **Actions** column of a running crawler to stop the crawler.

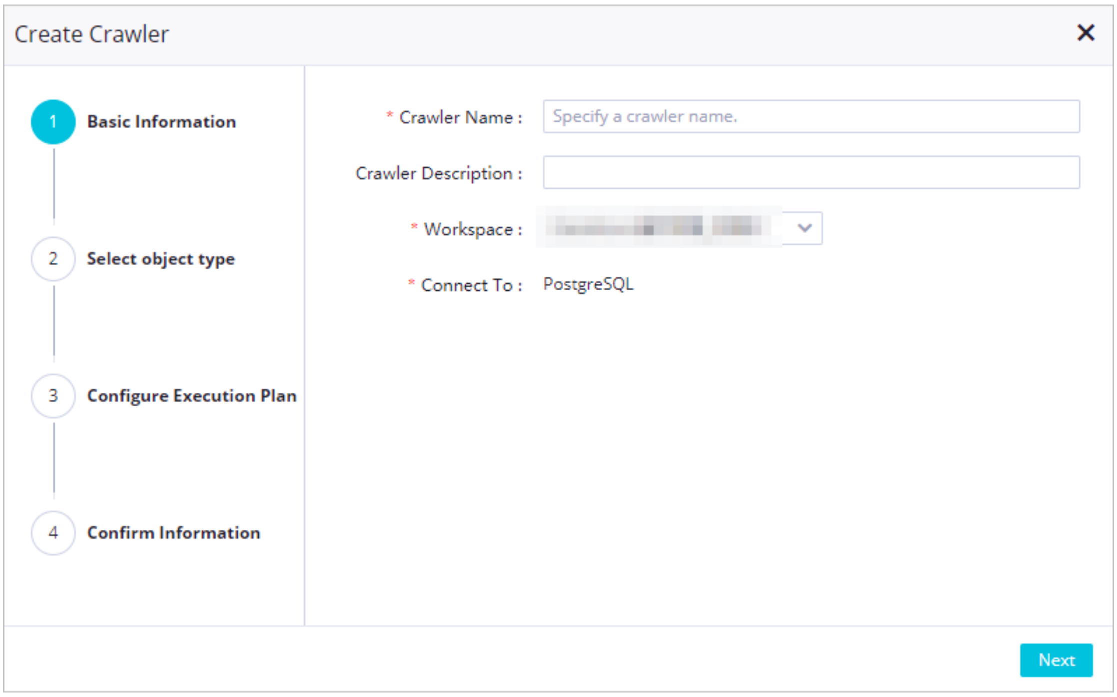
4.6.5. Collect metadata from a PostgreSQL data store

This topic describes how to create a crawler to collect metadata from a PostgreSQL data store to DataWorks. You can view the collected metadata on the Data Map page.

Procedure

1. Go to the **Data Discovery** page.
   i. Log on to the DataWorks console.
   ii. In the left-side navigation pane, click **Workspaces**.
   iii. In the top navigation bar, select a region as required, find the workspace where you want to create a crawler, and then click **Data Analytics** in the Actions column.
   iv. On the DataStudio page, click the ☰ icon in the upper-left corner and choose **All Products > Data governance > DataMap**. The homepage of Data Map appears.
   v. In the top navigation bar, click **Data Discovery**.
2. In the left-side navigation pane, click **PostgreSQL**.
3. On the **PostgreSQLMetadata Crawler** page, click **Create Crawler**.
4. In the **Create Crawler** dialog box, perform the following steps:

i. In the **Basic Information** step, set the basic parameters.



Create Crawler

1 Basic Information

2 Select object type

3 Configure Execution Plan

4 Confirm Information

* Crawler Name : Specify a crawler name.

Crawler Description :

* Workspace : [redacted]

* Connect To : PostgreSQL

Parameter	Description
Crawler Name	Required. The name of the crawler. You must specify a unique name.
Crawler Description	The description of the crawler.
Workspace	The workspace of the data store from which metadata will be collected.
Connection Type	The type of the data store from which metadata will be collected. The default value is **PostgreSQL** and cannot be changed.

ii. Click **Next**.

iii. In the **Select Collection Object** step, select a connection from the **Connection** drop-down list.If the required connection does not exist, click **Create Connection** to go to the **Data Source** page in **Workspace Management** and create the connection.

iv. Click **Test Crawler Connectivity**.If the database is configured with a whitelist, you must add the IP address of DataWorks based on the region where the workspace resides to the whitelist.

v. When the message **The connectivity test has been passed** appears, click **Next**.

vi. In the **Configure Execution Plan** step, specify the execution plan.Valid values of the **Execution Plan** parameter: **On-demand Execution**, **Monthly**, **Weekly**, **Daily**, and **Hourly**.

vii. Click **Next**.

viii. In the **Confirm Information** step, verify that the configuration of the crawler is correct and click **OK**.

5. On the **PostgreSQLMetadata Crawler** page, find the created crawler and click **Run** in the Actions column.After the crawler is run, click the number in the **Updated Tables in Last Run** or **Added**

Tables in Last Run column to view the details about the updated or added tables.

You can also perform the following operations on the OSSMetadata Crawler page:

- Click **Details** in the Actions column of a crawler. In the **Crawler Details** dialog box, view the detailed information about the crawler.
- Click **Edit** in the Actions column of a crawler. In the **Edit Crawler** dialog box, modify the configuration of the crawler.
- Click **Delete** in the Actions column of a crawler. In the **Confirm** message, click **OK** to delete the crawler.
- Click **Stop** in the **Actions** column of a running crawler to stop the crawler.

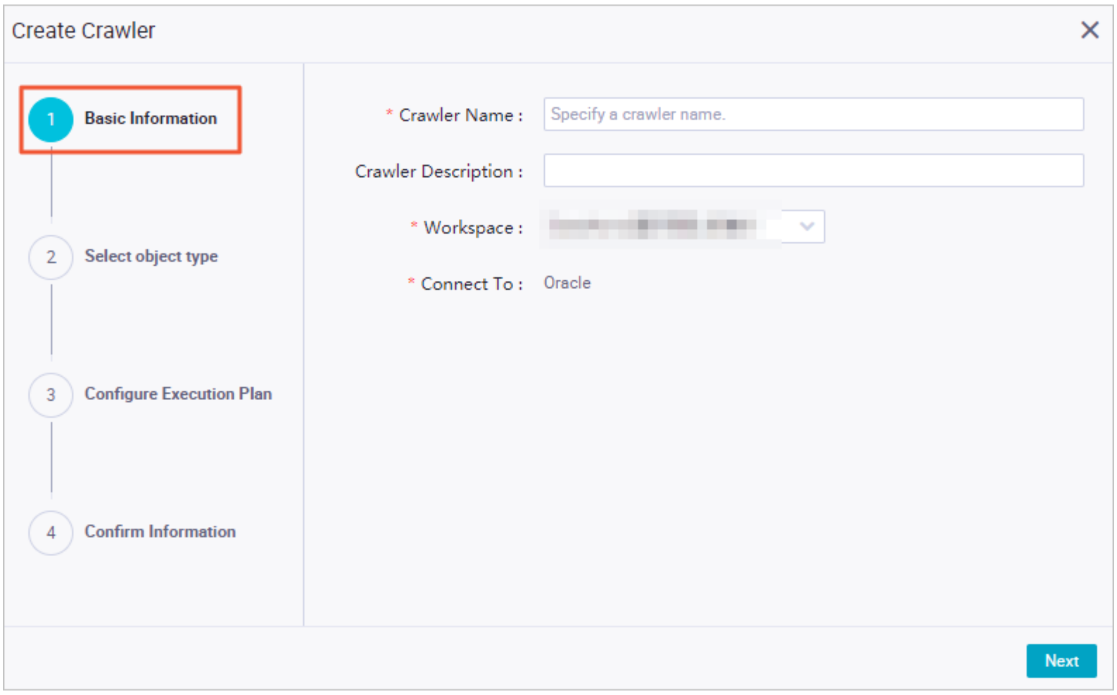
4.6.6. Collect metadata from an Oracle data store

This topic describes how to create a crawler to collect metadata from an Oracle data store to DataWorks. You can view the collected metadata on the Data Map page.

Procedure

1. Go to the **Data Discovery** page.
   i. Log on to the DataWorks console.
   ii. In the left-side navigation pane, click **Workspaces**.
   iii. In the top navigation bar, select a region as required, find the workspace where you want to create a crawler, and then click **Data Analytics** in the Actions column.
   iv. On the DataStudio page, click the ☰ icon in the upper-left corner and choose **All Products > Data governance > DataMap**. The homepage of Data Map appears.
   v. In the top navigation bar, click **Data Discovery**.
2. In the left-side navigation pane, click **Oracle**.
3. On the **OracleMetadata Crawler** page, click **Create Crawler**.
4. In the **Create Crawler** dialog box, perform the following steps:

i. In the **Basic Information** step, set the basic parameters.



Create Crawler

1 Basic Information

2 Select object type

3 Configure Execution Plan

4 Confirm Information

* Crawler Name : Specify a crawler name.

Crawler Description :

* Workspace : [redacted]

* Connect To : Oracle

Parameter	Description
Crawler Name	Required. The name of the crawler. You must specify a unique name.
Crawler Description	The description of the crawler.
Workspace	The workspace of the data store from which metadata will be collected.
Connection Type	The type of the data store from which metadata will be collected. The default value is **Oracle** and cannot be changed.

ii. Click **Next**.

iii. In the **Select Collection Object** step, select a connection from the **Connection** drop-down list.If the required connection does not exist, click **Create Connection** to go to the **Data Source** page in **Workspace Management** and create the connection.

iv. Click **Test Crawler Connectivity**.If the database is configured with a whitelist, you must add the IP address of DataWorks based on the region where the workspace resides to the whitelist.

v. When the message **The connectivity test has been passed** appears, click **Next**.

vi. In the **Configure Execution Plan** step, specify the execution plan.Valid values of the **Execution Plan** parameter: **On-demand Execution**, **Monthly**, **Weekly**, **Daily**, and **Hourly**.

vii. Click **Next**.

viii. In the **Confirm Information** step, verify that the configuration of the crawler is correct and click **OK**.

5. On the **OracleMetadata Crawler** page, find the created crawler and click **Run** in the Actions column.After the crawler is run, click the number in the **Updated Tables in Last Run** or **Added Tables in Last Run** column to view the details about the updated or added tables.

You can also perform the following operations on the OSSMetadata Crawler page:

- Click **Details** in the Actions column of a crawler. In the **Crawler Details** dialog box, view the detailed information about the crawler.
- Click **Edit** in the Actions column of a crawler. In the **Edit Crawler** dialog box, modify the configuration of the crawler.
- Click **Delete** in the Actions column of a crawler. In the **Confirm** message, click **OK** to delete the crawler.
- Click **Stop** in the **Actions** column of a running crawler to stop the crawler.

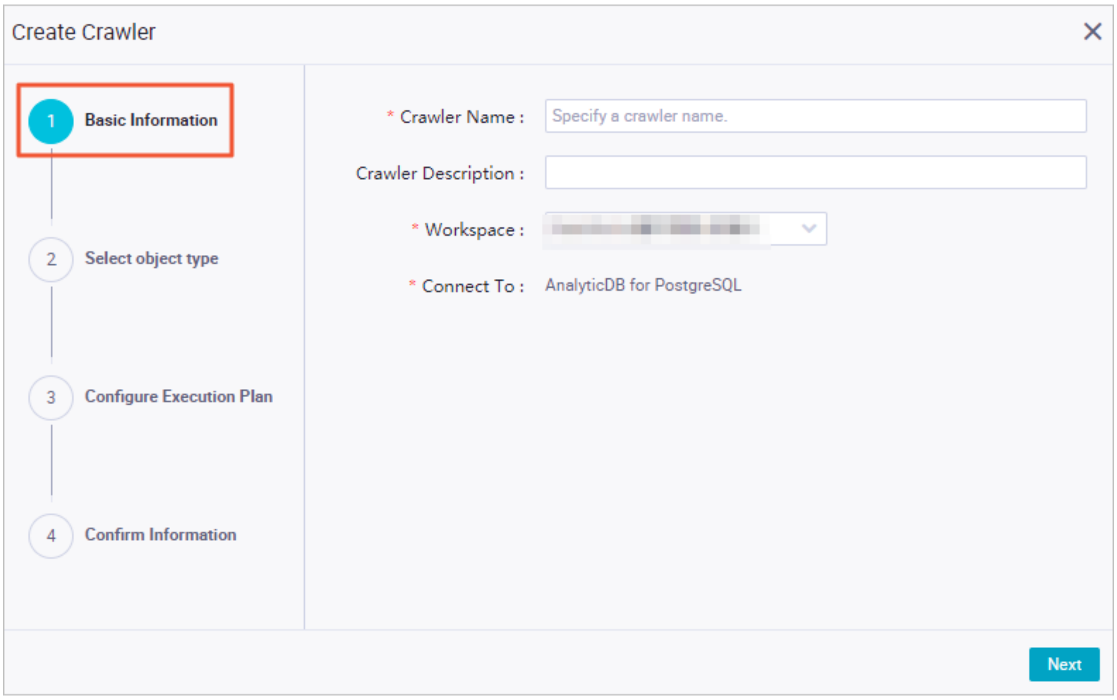
4.6.7. Collect metadata from an AnalyticDB for PostgreSQL data store

This topic describes how to create a crawler to collect metadata from an AnalyticDB for PostgreSQL data store to DataWorks. After the metadata is collected, you can manage tables in the AnalyticDB for PostgreSQL data store on the Data Map page.

Procedure

1. Go to the **Data Discovery** page.
   i. Log on to the DataWorks console.
   ii. In the left-side navigation pane, click **Workspaces**.
   iii. In the top navigation bar, select a region as required, find the workspace where you want to create a crawler, and then click **Data Analytics** in the Actions column.
   iv. On the DataStudio page, click the ☰ icon in the upper-left corner and choose **All Products > Data governance > DataMap**. The homepage of Data Map appears.
   v. In the top navigation bar, click **Data Discovery**.
2. In the left-side navigation pane, click **AnalyticDB for PostgreSQL**.
3. On the **AnalyticDB for PostgreSQLMetadata Crawler** page, click **Create Crawler**.
4. In the **Create Crawler** dialog box, perform the following steps:

i. In the **Basic Information** step, set the basic parameters.



* Crawler Name : Specify a crawler name.

Crawler Description :

* Workspace :

* Connect To : AnalyticDB for PostgreSQL

Parameter	Description
Crawler Name	Required. The name of the crawler. You must specify a unique name.
Crawler Description	The description of the crawler.
Workspace	The workspace of the data store from which metadata will be collected.
Connection Type	The type of the data store from which metadata will be collected. The default value is **AnalyticDB for PostgreSQL** and cannot be changed.

ii. Click **Next**.

iii. In the **Select Collection Object** step, select a connection from the **Connection** drop-down list.If the required connection does not exist, click **Create Connection** to go to the **Data Source** page in **Workspace Management** and create the connection.

iv. Click **Test Crawler Connectivity**. If the message **The connectivity test has been passed** appears, the DataWorks metadata service can access the AnalyticDB for PostgreSQL data store.

v. Click **Next**.

5. On the **AnalyticDB for PostgreSQLMetadata Crawler** page, find the created crawler and click **Run** in the Actions column.

After the crawler is run, click the number in the **Updated Tables in Last Run** or **Added Tables in Last Run** column to view the details about the updated or added tables.

> **Notice** The **Run** button is displayed only in the Actions column of a crawler that needs to be manually triggered.

You can also perform the following operations on the AnalyticDB for PostgreSQLMetadata Crawler page:

- Click **Details** in the Actions column of a crawler. In the **Crawler Details** dialog box, view the detailed information about the crawler.
- Click **Edit** in the Actions column of a crawler. In the **Edit Crawler** dialog box, modify the configuration of the crawler.
- Click **Delete** in the Actions column of a crawler. In the **Confirm** message, click **OK** to delete the crawler.
- Click **Stop** in the **Actions** column of a running crawler to stop the crawler.

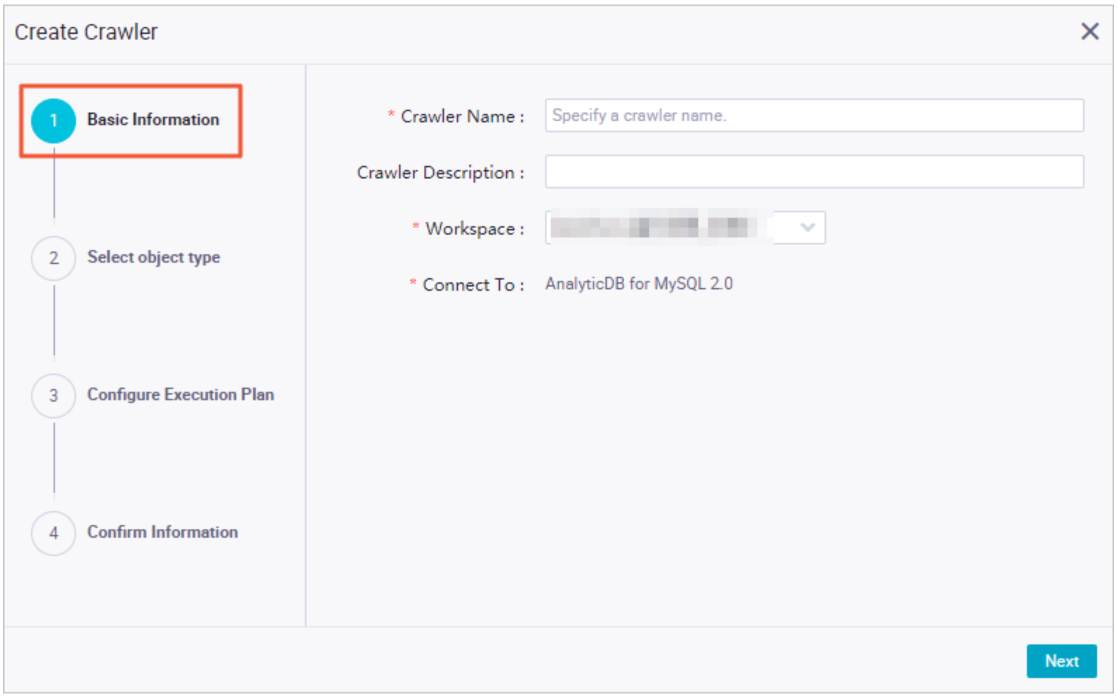
4.6.8. Collect metadata from an AnalyticDB for MySQL 2.0 data store

This topic describes how to create a crawler to collect metadata from an AnalyticDB for MySQL 2.0 data store to DataWorks. You can view the collected metadata on the Data Map page.

Procedure

1. Go to the **Data Discovery** page.
   i. Log on to the DataWorks console.
   ii. In the left-side navigation pane, click **Workspaces**.
   iii. In the top navigation bar, select a region as required, find the workspace where you want to create a crawler, and then click **Data Analytics** in the Actions column.
   iv. On the DataStudio page, click the ☰ icon in the upper-left corner and choose **All Products > Data governance > DataMap**. The homepage of Data Map appears.
   v. In the top navigation bar, click **Data Discovery**.
2. In the left-side navigation pane, click **AnalyticDB for MySQL 2.0**.
3. On the **AnalyticDB for MySQL 2.0Metadata Crawler** page, click **Create Crawler**.
4. In the **Create Crawler** dialog box, perform the following steps:

i. In the **Basic Information** step, set the basic parameters.



* Crawler Name : Specify a crawler name.

Crawler Description :

* Workspace : [redacted]

* Connect To : AnalyticDB for MySQL 2.0

Parameter	Description
Crawler Name	Required. The name of the crawler. You must specify a unique name.
Crawler Description	The description of the crawler.
Workspace	The workspace of the data store from which metadata will be collected.
Connection Type	The type of the data store from which metadata will be collected. The default value is **AnalyticDB for MySQL 2.0** and cannot be changed.

ii. Click **Next**.

iii. In the **Select Collection Object** step, select a connection from the **Connection** drop-down list.If the required connection does not exist, click **Create Connection** to go to the **Data Source** page in **Workspace Management** and create the connection.

iv. Click **Test Crawler Connectivity**.If the database is configured with a whitelist, you must add the IP address of DataWorks based on the region where the workspace resides to the whitelist.

v. When the message **The connectivity test has been passed** appears, click **Next**.

vi. In the **Configure Execution Plan** step, specify the execution plan.Valid values of the **Execution Plan** parameter: **On-demand Execution**, **Monthly**, **Weekly**, **Daily**, and **Hourly**.

vii. Click **Next**.

viii. In the **Confirm Information** step, verify that the configuration of the crawler is correct and click **OK**.

5. On the **AnalyticDB for MySQL 2.0Metadata Crawler** page, find the created crawler and click **Run** in the Actions column.After the crawler is run, click the number in the **Updated Tables in**

Last Run or **Added Tables in Last Run** column to view the details about the updated or added tables.

You can also perform the following operations on the OSSMetadata Crawler page:

- Click **Details** in the Actions column of a crawler. In the **Crawler Details** dialog box, view the detailed information about the crawler.
- Click **Edit** in the Actions column of a crawler. In the **Edit Crawler** dialog box, modify the configuration of the crawler.
- Click **Delete** in the Actions column of a crawler. In the **Confirm** message, click **OK** to delete the crawler.
- Click **Stop** in the **Actions** column of a running crawler to stop the crawler.

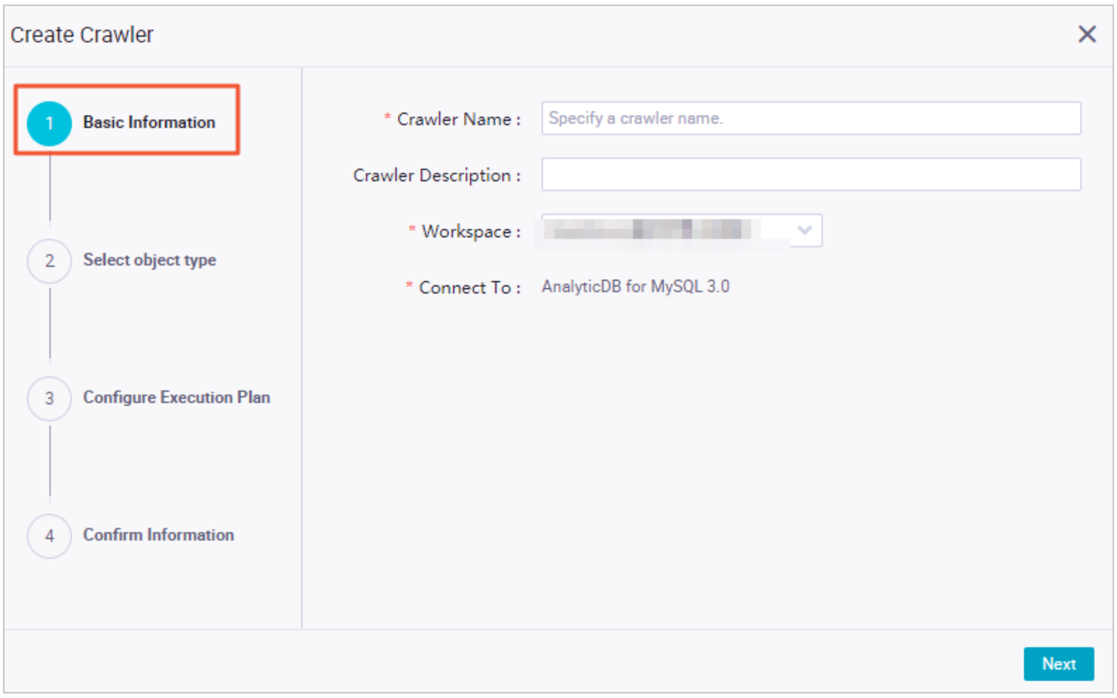
4.6.9. Collect metadata from an AnalyticDB for MySQL 3.0 data store

This topic describes how to create a crawler to collect metadata from an AnalyticDB for MySQL 3.0 data store to DataWorks. You can view the collected metadata on the Data Map page.

Procedure

1. Go to the **Data Discovery** page.
    i. Log on to the DataWorks console.
    ii. In the left-side navigation pane, click **Workspaces**.
    iii. In the top navigation bar, select a region as required, find the workspace where you want to create a crawler, and then click **Data Analytics** in the Actions column.
    iv. On the DataStudio page, click the ☰ icon in the upper-left corner and choose **All Products** > **Data governance** > **DataMap**. The homepage of Data Map appears.
    v. In the top navigation bar, click **Data Discovery**.
2. In the left-side navigation pane, click **AnalyticDB for MySQL 3.0**.
3. On the **AnalyticDB for MySQL 3.0Metadata Crawler** page, click **Create Crawler**.
4. In the **Create Crawler** dialog box, perform the following steps:

i. In the **Basic Information** step, set the basic parameters.



Create Crawler

* Crawler Name : Specify a crawler name.

Crawler Description :

* Workspace : [redacted]

* Connect To : AnalyticDB for MySQL 3.0

Parameter	Description
Crawler Name	Required. The name of the crawler. You must specify a unique name.
Crawler Description	The description of the crawler.
Workspace	The workspace of the data store from which metadata will be collected.
Connection Type	The type of the data store from which metadata will be collected. The default value is **AnalyticDB for MySQL 3.0** and cannot be changed.

ii. Click **Next**.

iii. In the **Select Collection Object** step, select a connection from the **Connection** drop-down list.If the required connection does not exist, click **Create Connection** to go to the **Data Source** page in **Workspace Management** and create the connection.

iv. Click **Test Crawler Connectivity**.If the database is configured with a whitelist, you must add the IP address of DataWorks based on the region where the workspace resides to the whitelist.

v. When the message **The connectivity test has been passed** appears, click **Next**.

vi. In the **Configure Execution Plan** step, specify the execution plan.Valid values of the **Execution Plan** parameter: **On-demand Execution**, **Monthly**, **Weekly**, **Daily**, and **Hourly**.

vii. Click **Next**.

viii. In the **Confirm Information** step, verify that the configuration of the crawler is correct and click **OK**.

5. On the **AnalyticDB for MySQL 3.0Metadata Crawler** page, find the created crawler and click **Run** in the Actions column.After the crawler is run, click the number in the **Updated Tables in**

Last Run or **Added Tables in Last Run** column to view the details about the updated or added tables.

You can also perform the following operations on the OSSMetadata Crawler page:

- Click **Details** in the Actions column of a crawler. In the **Crawler Details** dialog box, view the detailed information about the crawler.
- Click **Edit** in the Actions column of a crawler. In the **Edit Crawler** dialog box, modify the configuration of the crawler.
- Click **Delete** in the Actions column of a crawler. In the **Confirm** message, click **OK** to delete the crawler.
- Click **Stop** in the **Actions** column of a running crawler to stop the crawler.

4.6.10. Collect metadata from an OSS data store

This topic describes how to create a crawler to collect metadata from an Object Storage Service (OSS) data store to DataWorks. You can view the collected metadata on the Data Map page.

Context

DataWorks allows you to collect metadata from OSS data stores only in the China (Shanghai) region. This feature is in invitational preview.

Procedure

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

   i. Log on to the DataWorks console.

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

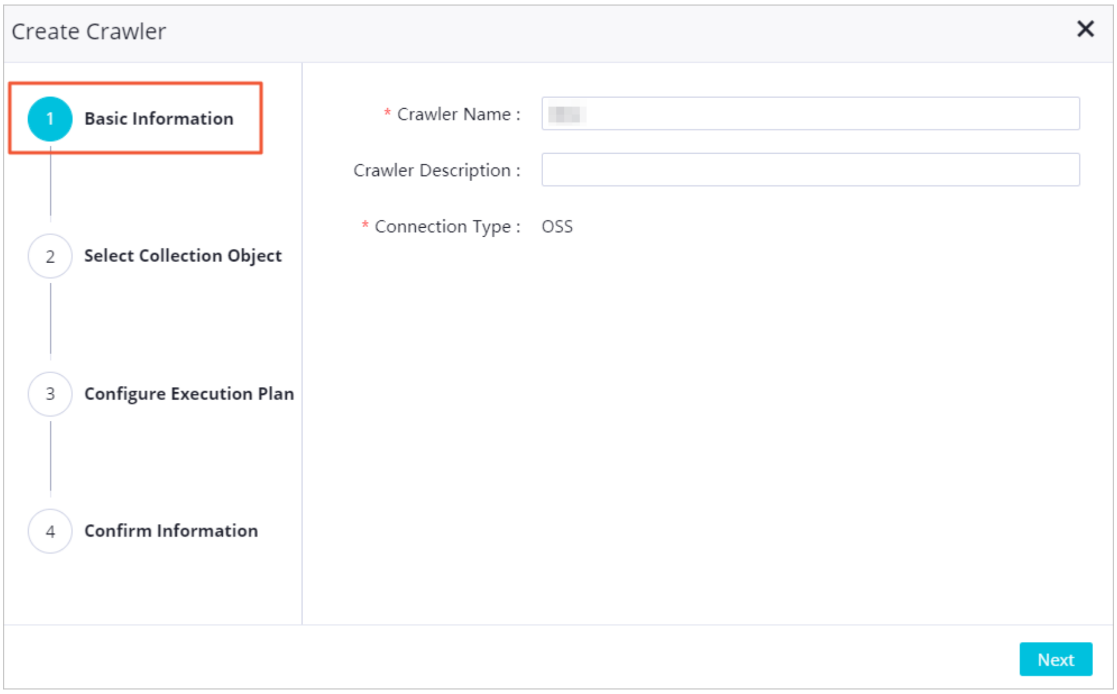
   iii. In the top navigation bar, select a region as required, find the workspace where you want to create a crawler, and then click **Data Analytics** in the Actions column.

   iv. On the DataStudio page, click the ☰ icon in the upper-left corner and choose **All Products** > **Data governance** > **DataMap**. The homepage of Data Map appears.

   v. In the top navigation bar, click **Data Discovery**.

2. In the left-side navigation pane, click **OSS**.
3. On the **OSSMetadata Crawler** page, click **Create Crawler**.
4. In the **Create Crawler** dialog box, perform the following steps:

i. In the **Basic Information** step, set the basic parameters.



Create Crawler

* Crawler Name :

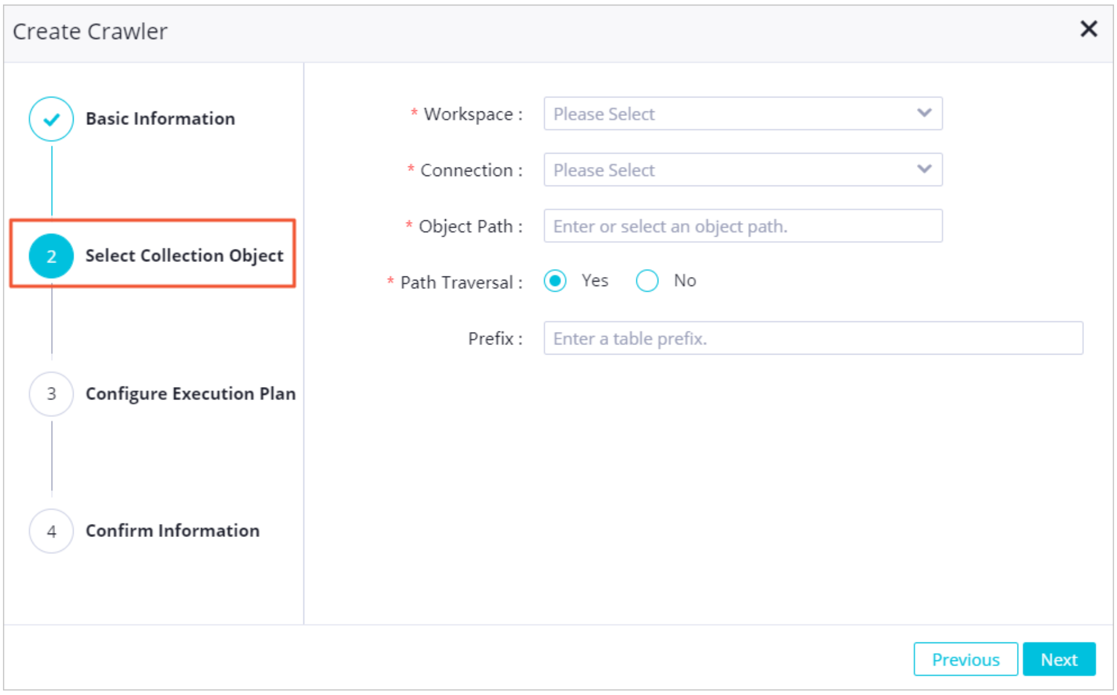
Crawler Description :

* Connection Type : OSS

Parameter	Description
Crawler Name	Required. The name of the crawler. You must specify a unique name.
Crawler Description	The description of the crawler.
Connection Type	The type of the data store from which metadata will be collected. The default value is OSS and cannot be changed.

ii. Click **Next**.

iii. In the **Select Collection Object** step, set the parameters of the object from which metadata will be collected.



Create Crawler

* Workspace : Please Select

* Connection : Please Select

* Object Path : Enter or select an object path.

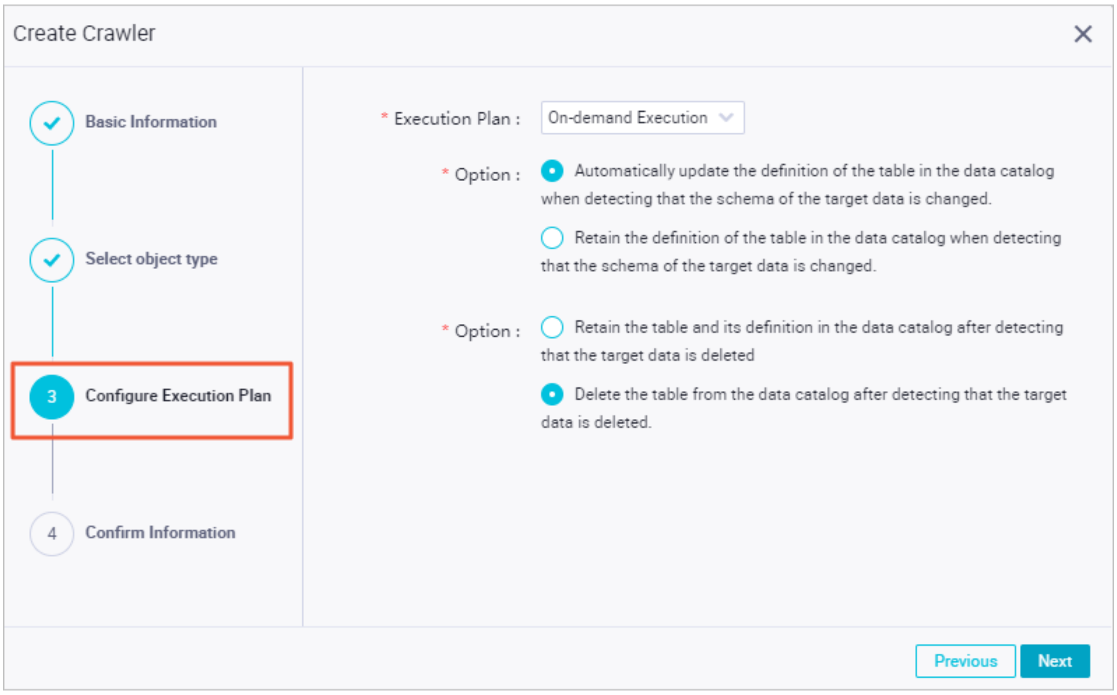
* Path Traversal : ● Yes ○ No

Prefix : Enter a table prefix.

Parameter	Description
Workspace	The workspace of the OSS data store from which metadata will be collected.
Connection	The connection to the OSS data store from which metadata will be collected. If the required connection does not exist, go to the **Data Source** page in **Workspace Management** and create the connection. For more information, see Configure an OSS connection.
Object Path	The path of the OSS object from which metadata will be collected.
Path Traversal	Specifies whether to traverse sub-paths in the specified path.
Prefix	The prefix of the names of tables that the crawler automatically generates. By default, a generated table is named after the corresponding OSS object.

iv. Click **Next**.

v. In the **Configure Execution Plan** step, set the scheduling parameters.



Create Crawler

* Execution Plan : On-demand Execution

* Option : ● Automatically update the definition of the table in the data catalog when detecting that the schema of the target data is changed.
○ Retain the definition of the table in the data catalog when detecting that the schema of the target data is changed.

* Option : ○ Retain the table and its definition in the data catalog after detecting that the target data is deleted
● Delete the table from the data catalog after detecting that the target data is deleted.

Parameter	Description
Execution Plan	The execution plan. Valid values: **On-demand Execution**, **Monthly**, **Weekly**, **Daily**, **Hourly**, and **Custom**.
Update Options	The policy for updating the table that stores the collected metadata.
Delete Options	The policy for deleting the table that stores the collected metadata.

vi. Click **Next**.

vii. In the **Confirm Information** step, verify that the configuration of the crawler is correct and click **OK**.

5. On the **OSSMetadata Crawler** page, find the created crawler and click **Run** in the Actions column.After the crawler is run, click the number in the **Updated Tables in Last Run** or **Added Tables in Last Run** column to view the details about the updated or added tables.

   You can also perform the following operations on the OSSMetadata Crawler page:

   - Click **Details** in the Actions column of a crawler. In the **Crawler Details** dialog box, view the detailed information about the crawler.
   - Click **Edit** in the Actions column of a crawler. In the **Edit Crawler** dialog box, modify the configuration of the crawler.
   - Click **Delete** in the Actions column of a crawler. In the **Confirm** message, click **OK** to delete the crawler.
   - Click **Stop** in the **Actions** column of a running crawler to stop the crawler.

6. View the metadata collected from the OSS data store.

   i. In the top navigation bar, click **All Data**.

ii. Click the **OSS** tab.

iii. On the **OSS** tab, click the name of the table that stores the collected metadata and view the table details.

5.Data opening feature

5.1. Install and remove the data opening package

DataWorks provides the data opening feature. Before you use the feature to collect metadata, you must install the data opening package in your workspace. This topic describes how to install the data opening package and how to view the installation result after the package is installed.

Limits

- The data opening feature is in public preview and is available only for DataWorks Enterprise Edition and other advanced editions.
- The data opening feature can be used only if you select MaxCompute as the computing engine.

Install the data opening package

1. Obtain the required permissions.Before you install the data opening package in a workspace, you must submit a ticket to provide the workspace information to the DataWorks technical support personnel. Then, the technical support personnel can grant the required permissions to the workspace. This way, you can install the data opening package in the workspace and use the data opening feature to collect metadata. Make sure that the workspace meets the following requirements:
    - The workspace has a well-established permission management mechanism. This can prevent data leaks caused by unlimited sharing of the metadata that is collected by using the data opening feature.
    - The workspace can be shared within an enterprise or a team. This way, the metadata that is collected by using the data opening feature can be shared among different data development teams.
2. Use the account of the workspace owner to install the data opening package.In this topic, the data opening package is installed in a workspace on the DataStudio page of the workspace. Go to the DataStudio page of the workspace that is authorized and run the following commands to install the data opening package on the MaxCompute node:

    > Note
    >
    > - If the authorized workspace is a standard-mode workspace, the data opening package must be installed in both the development environment and production environment. This is because that the production environments of a standard-mode workspace are isolated from the development environments.
    > - You can change the names of both the workspace and the data opening package in the commands used to install the data opening package based on your business requirements. The following commands are used to install the data opening package in the workspace that resides in the China (Hangzhou) region:

```
-- Install the data opening package in the workspace that resides in the China (Hangzhou) region in a de
velopment environment (work_test_2_dev).
INSTALL PACKAGE u_meta_hangzhou.systables;
-- Install the data opening package in a workspace that resides in the China (Hangzhou) region in a prod
uction environment (work_test_2).
USE work_test_2;
INSTALL PACKAGE u_meta_hangzhou.systables;
```

In the preceding commands:

- *u_meta_hangzhou*: specifies the name of the workspace in which you want to install the data opening package. This workspace resides in the China (Hangzhou) region. Alibaba Cloud provides different data opening packages for different regions. The name of the workspace varies based on the region. For more information about the names of the workspaces in different regions, see Appendix 2: Available data opening packages. You can change the name of the workspace based on your business requirements.
- *systables*: specifies the name of the data opening package that you want to install in a workspace in the China (Hangzhou) region. Alibaba Cloud provides different data opening packages for different regions. The name of the data opening package varies based on the region. For information about the names of the data opening packages in different regions, see Appendix 2: Available data opening packages. You can change the name of the data opening package based on your business requirements.

3. View the installation result of the data opening package.In the list of the installed data opening packages, find the data opening package that you installed. If the **Status** of the data opening package is **OK**, the data opening package is installed.

```
-- Check whether the u_meta_hangzhou.systables data opening package is included in the data openin
g packages installed in the workspace.
SHOW PACKAGES;
-- Output example
+-------------+-----------------+-------------------------+--------+
| PackageName | SourceProject   | InstallTime             | Status |
+-------------+-----------------+-------------------------+--------+
| systables   | u_meta_hangzhou | 2020-11-26T15:25:22+0800 | OK     |
+-------------+-----------------+-------------------------+--------+
```

View the tables or views provided by the data opening package

In most cases, when you use the data opening feature to query a table or view, you need to specify the full name of the table or view. The table or view name that you need to specify varies based on the data opening packages of different versions. You can run the following command to query the tables and views provided by the installed data opening package and view the names of and permissions on the tables or views:

```
DESCRIBE PACKAGE u_meta_hangzhou.systables;
```

In the following example, a command is run in DataStudio of DataWorks to query the tables provided by the installed data opening package, and the query results are also provided.

Note The names of the tables and views provided by the data opening package also contain the version information of the data opening package. The version number of the data opening package changes with the iteration and releases of new features. When you use the data opening feature, the actual version number of the data opening package in the package specified by the systables parameter prevails. For example, in the **raw_v_tenant_user_v1_1** view, **v1_1** is the version number.

```
-- View the tables or views contained in the u_meta_hangzhou.systables package.
DESCRIBE PACKAGE u_meta_hangzhou.systables;
-- Output example
CreateTime:       2020-11-18T20:17:24+0800
PackageName:      systables
SourceProject:    u_meta_hangzhou
Object List
+------------+---------------------------------+-----------------+
| ObjectType | ObjectName                      | ObjectPrivileges |
+------------+---------------------------------+-----------------+
| TABLE      | raw_v_meta_biz_table_wiki_v1_1  | Describe,Select |
+------------+---------------------------------+-----------------+
| TABLE      | raw_v_meta_column_usage_v1_1    | Describe,Select |
+------------+---------------------------------+-----------------+
| TABLE      | raw_v_meta_column_v1_1          | Describe,Select |
+------------+---------------------------------+-----------------+
| TABLE      | raw_v_meta_database_v1_1        | Describe,Select |
+------------+---------------------------------+-----------------+
| TABLE      | raw_v_meta_partition_v1_1       | Describe,Select |
+------------+---------------------------------+-----------------+
| TABLE      | raw_v_meta_table_detail_log_v1_1 | Describe,Select |
+------------+---------------------------------+-----------------+
| TABLE      | raw_v_meta_table_join_map_v1_1  | Describe,Select |
+------------+---------------------------------+-----------------+
| TABLE      | raw_v_meta_table_lineage_v1_1   | Describe,Select |
+------------+---------------------------------+-----------------+
| TABLE      | raw_v_meta_table_output_v1_1    | Describe,Select |
+------------+---------------------------------+-----------------+
| TABLE      | raw_v_meta_table_usage_v1_1     | Describe,Select |
+------------+---------------------------------+-----------------+
| TABLE      | raw_v_meta_table_v1_1           | Describe,Select |
+------------+---------------------------------+-----------------+
| TABLE      | raw_v_meta_view_v1_1            | Describe,Select |
+------------+---------------------------------+-----------------+
| TABLE      | raw_v_schedule_di_resgroup_v1_1 | Describe,Select |
+------------+---------------------------------+-----------------+
| TABLE      | raw_v_schedule_node_relation_v1_1 | Describe,Select |
+------------+---------------------------------+-----------------+
| TABLE      | raw_v_schedule_node_v1_1        | Describe,Select |
+------------+---------------------------------+-----------------+
| TABLE      | raw_v_schedule_task_v1_1        | Describe,Select |
+------------+---------------------------------+-----------------+
| TABLE      | raw_v_tenant_res_group_v1_1     | Describe,Select |
+------------+---------------------------------+-----------------+
| TABLE      | raw_v_tenant_user_v1_1          | Describe,Select |
+------------+----------------------------------+------------------+
| TABLE      | raw_v_tenant_workspace_user_v1_1 | Describe,Select |
+------------+----------------------------------+------------------+
| TABLE      | raw_v_tenant_workspace_v1_1      | Describe,Select |
+------------+----------------------------------+------------------+
| TABLE      | rpt_v_meta_ind_table_core_v1_1   | Describe,Select |
+------------+----------------------------------+------------------+
| TABLE      | rpt_v_meta_ind_table_extra_v1_1  | Describe,Select |
+------------+----------------------------------+------------------+
```

View the schema of a table or view and its field descriptions

Run the following command to query the schema of a table or view and its field descriptions:

```
DESCRIBE u_meta_hangzhou.rpt_v_meta_ind_table_core_v1_0;
```

In the preceding command, *rpt_v_meta_ind_table_core_v1_0* is the name of the table or view whose schema and field descriptions you want to query. You can obtain the actual table or view name based on the instructions provided in view tables or views provided by the data opening package.

> **Note** DataWorks provides data from multiple dimensions, such as detail data, metric data, and dimensional data. You can obtain and store data to a DataWorks workspace based on your business requirements to save storage space. The data can be used as the historical data for subsequent data governance or O&M.

In the following example, a command is run to query the **rpt_v_meta_ind_table_core** table in DataStudio of DataWorks, and the query results are also provided.

```
-- View the descriptions of the fields in the rpt_v_meta_ind_table_core table.
DESCRIBE u_meta_hangzhou.rpt_v_meta_ind_table_core_v1_0;
-- Output example
+------------------------------------------------------------------------------------+
| Owner: ALIYUN$dataworks-datagovernance | Project: u_meta_hangzhou                  |
| TableComment: core metrics of the table in the metadata module                     |
+------------------------------------------------------------------------------------+
| CreateTime:          2020-12-07 20:02:53                          |
| LastDDLTime:         2020-12-07 20:02:53                          |
| LastModifiedTime:    2020-12-07 20:02:53                          |
+------------------------------------------------------------------------------------+
| VirtualView  : YES  | ViewText: CREATE OR REPLACE VIEW rpt_v_meta_ind_table_core_v1_1 (@param_biz_da
te STRING)
RETURNS @ret_result TABLE (
  tenant_id        BIGINT       COMMENT    'Specifies the ID of the DataWorks tenant.',
  project_id       BIGINT       COMMENT    'Specifies the ID of the DataWorks workspace.',
  catalog_name      STRING       COMMENT    'Specifies the catalog to which the table belongs. The catalog n
ame for a project in MaxCompute is odps.',
  database_name      STRING       COMMENT    'Specifies the name of the database or MaxCompute project.',
  table_name       STRING       COMMENT    'Specifies the name of the table.',
  table_uuid       STRING       COMMENT    'Specifies the ID of the table.',
  owner_yun_acct      STRING       COMMENT    'Specifies the Alibaba Cloud account used by the table owner
.',
  dim_life_cycle     BIGINT       COMMENT    'Specifies whether to configure the lifecycle for the table. The va
lue 0 indicates that the lifecycle is not configured for the table, and other values indicate the specific values
of the lifecycle.'
  is_partition_table  BOOLEAN       COMMENT    'Specifies whether the table is a partitioned table. The value
true indicates that the table is a partitioned table.',
  entity_type       BIGINT        COMMENT    'Specifies the entity type. The value 0 indicates a table, and the va
lue 1 indicates a view.',
  categories        STRING        COMMENT    'Specifies the categories in the table.',
  last_access_time    BIGINT        COMMENT    'Specifies the time at which the table was last accessed. This ti
me is a 10-digit UNIX timestamp.',
  `size`           BIGINT        COMMENT    'Specifies the size of the table, which indicates the logical storage sp
ace occupied by data in the table. The volume of data stored in a view is NULL.',
  column_count      BIGINT        COMMENT    'Specifies the number of fields in the table, including the partiti
on key column.',
  partition_count     BIGINT        COMMENT    'Specifies the number of partitions in the table. If the table is a
non-partitioned table, the value of this parameter is NULL.',
  detail_view_count   BIGINT        COMMENT    'Specifies the number of times that the details of the table ar
e viewed on the page.',
  favorite_count      BIGINT        COMMENT    'Specifies the number of times that the table is added to favorit
e.',
  biz_date          STRING        COMMENT    'Specifies the date on which business data was generated.'
) COMMENT 'Core metrics of the table in the metadata module' AS
SELECT * FROM u_meta_hangzhou.rpt_v_meta_ind_table_core_proxy(@param_biz_date) |
+------------------------------------------------------------------------------------+
| Native Columns:                                                  |
+------------------------------------------------------------------------------------+
| Field         | Type       | Label | Comment                         |
+------------------------------------------------------------------------------------+
| tenant_id       | bigint     |       | Indicates the ID of the DataWorks tenant.                    |
| project_id      | bigint     |       | Indicates the ID of the DataWorks workspace.                 |
| catalog_name    | string     |       | Indicates the catalog to which the table belongs. The catalog name for a proje
ct in MaxCompute is odps.;       |
| database_name   | string     |       | Indicates the name of the database or MaxCompute project.                    |
| table_name      | string     |       | Indicates the name of the table.                           |
| table_uuid      | string     |       l Indicates the ID of the table.                           |
| owner_yun_acct  | string     |       | Indicates the Alibaba Cloud account used by the table owner.
|
| dim_life_cycle  | bigint     |       | Indicates whether the lifecycle is configured for the table. The value 0 indicate
s that the lifecycle is not configured for the table, and other values indicate the specific values of the lifecycl
e.          |
| is_partition_table | boolean |       | Indicates whether the table is a partitioned table. The value true indicat
es that the table is a partitioned table.                 |
| entity_type     | bigint     |       | Indicates the entity type. The value 0 indicates a table, and the value 1 indicate
s a view.               |
| categories      | string     |       | Indicates the categories in the table.                       |
| last_access_time | bigint    |       | Indicates the time at which the table was last accessed. This time is a 10-digi
t UNIX timestamp.            |
| size          | bigint     |       | Indicates the size of the table, which indicates the logical storage space occupied by
data in the table. The volume of data stored in a view is NULL.           |
| column_count    | bigint     |       | Indicates the number of fields in the table, including the partition key colum
n.                      |
| partition_count | bigint     |       | Indicates the number of partitions in the table. If the table is a non-partition
ed table, the value of this parameter is NULL.                    |
| detail_view_count | bigint    |       | Indicates the number of times that the details of the table are viewed on t
he page.                |
| favorite_count  | bigint     |       | Indicates the number of times that the table is added to favorite.
```

```
|
| biz_date     | string   |     | Indicates the date on which business data was generated.                 |
+------------------------------------------------------------------------------------+
```

Remove the data opening package

In this section, the data opening package is removed from a workspace in DataStudio of DataWorks by running one of the following commands:

> **Note** If the authorized workspace is a standard-mode workspace, the data opening package must be removed from both the development environment and production environment. This is because that the production environments of a standard-mode workspace are isolated from the production environments.

```
-- Remove the data opening package from a workspace in a development environment (work_test_2_dev).
UNINSTALL PACKAGE u_meta_hangzhou.systables;
-- Remove the data opening package from a workspace in a production environment (work_test_2).
USE work_test_2;
UNINSTALL PACKAGE u_meta_hangzhou.systables;
```

What's next

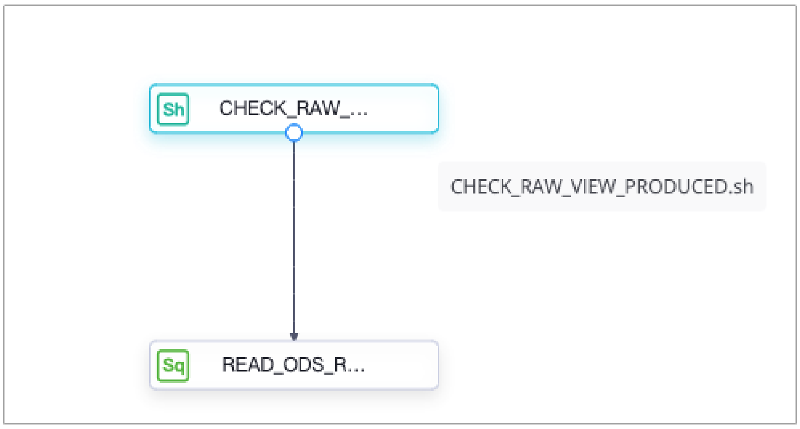
After the data opening package is installed, you can use the data opening feature to collect metadata for subsequent data governance or O&M. For more information about how to use the data opening feature, see Use the data opening feature.

5.2. Check whether the data opening feature has obtained the metadata of a node

After the data opening package is installed, you can use the data opening feature to obtain the metadata of a node. Before you perform this operation, make sure that the node has generated the latest metadata. Otherwise, you may fail to obtain the metadata. This topic describes how to use an ancestor node to check whether the data opening feature has obtained the metadata of a node.

Context

Based on the dependency principles of nodes in DataWorks, the system runs a node only after the ancestor node of the node is successfully run. To check whether the data opening feature has obtained the metadata of a node (Node A), create a node (Node B) and configure Node B as the ancestor node of Node A. If Node B detects that the data opening feature has obtained the metadata of Node A, the running of Node B is complete. The following figure shows the process.



The process involves the following two nodes:

- The `CHECK_RAW_VIEW_PRODUCED.sh` node is created to check whether the data opening feature has obtained the metadata of the READ_ODS_RAW_DATA.sql node. If the CHECK_RAW_VIEW_PRODUCED.sh node detects that the metadata has been obtained, the running of the CHECK_RAW_VIEW_PRODUCED.sh node is complete. Then, the system starts to run the READ_ODS_RAW_DATA.sql node.
- `READ_ODS_DATA.sql` is the node for which you want to check whether the data opening feature has obtained the metadata.

This topic only describes how to use the data opening feature to check whether the data opening feature has obtained the metadata of a node. For more information about how to use this feature, see Use the data opening feature.

- You can use a Shell node as an ancestor node to check whether the data opening feature has obtained the metadata of a node.
- You must use an Alibaba Cloud account that has permissions to access the data opening package and create task instances in the current MaxCompute project.

  To ensure both data and node security, we recommend that you use a RAM user of the Alibaba Cloud account. In addition, assign the **visitor** role and grant only the read permissions on the data opening package to the RAM user.

You can refer to the following steps to check whether the data opening feature has obtained the metadata of a node:

1. Create a RAM user and grant permissions to the RAM user
2. Check whether the data opening package has obtained the metadata of a node
3. Configure dependencies for the Shell node

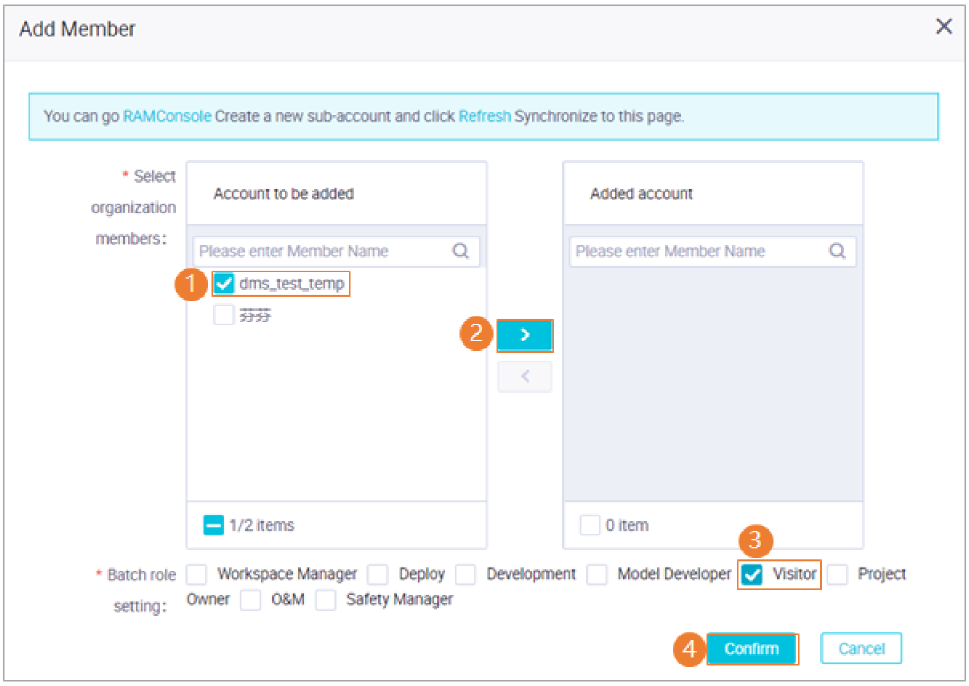
Create a RAM user and grant permissions to the RAM user

This section describes how to create a RAM user and grant permissions to the RAM user.

1. Create a RAM user. To ensure data security, we recommend that you do not authorize the RAM user to log on to the Alibaba Cloud Management Console but only create an AccessKey pair for the RAM user. For more information about how to create a RAM user, see Prepare a RAM user. In this

example, a RAM user named `dw_odps_test` is created.

2. Assign the **visitor** role to the RAM user. Add the RAM user to the required DataWorks workspace and assign the **visitor** role that has the minimum permissions to the RAM user. For more information, see Add workspace members.



Add Member

You can go RAMConsole Create a new sub-account and click Refresh Synchronize to this page.

* Select organization members:

Account to be added

Please enter Member Name

1 ☑ dms_test_temp

□ 芬芬

1/2 items

2 >

<

Added account

Please enter Member Name

□ 0 item

* Batch role setting: □ Workspace Manager □ Deploy □ Development □ Model Developer 3 ☑ Visitor □ Project Owner □ O&M □ Safety Manager

4 Confirm Cancel

3. Grant the read permissions on the data opening package to the RAM user. To enable the RAM user to read the data provided by the data opening package of DataWorks, you must grant the read permissions on the data opening package to the RAM user. In addition, you must grant the CreateInstance permission to the RAM user. This way, the RAM user can create task instances in the MaxCompute project when it reads data from the data opening package. For more information, see Grant a role or user.

```
-- Authorize the RAM user to create task instances in the MaxCompute project
GRANT CreateInstance ON PROJECT {Name of the MaxCompute project for which the DataWorks data o
pening package is installed} TO USER RAM$ {The Alibaba Cloud account of the RAM user}: dw_odps_test
;
-- Authorize the RAM user to read data from the data opening package (In this example, the RAM user is
authorized to read data from a data opening package installed for a MaxCompute project that resides in
the China (Hangzhou) region. For data in the data opening package installed for a MaxCompute project
that resides in another region, change the region information in the project name.)
GRANT READ ON PACKAGE u_meta_hangzhou.systables TO USER RAM$ (The Alibaba Cloud account of t
he RAM user): dw_odps_test;
-- View the authorization result
show grants for RAM${The Alibaba Cloud account of the RAM user}: dw_odps_test;
```

The following code provides an authorization example:

```
-- Authorization example
[roles]
role_project_guest
Authorization Type: ACL
[user/RAM${The Alibaba Cloud account of the RAM user}: dw_odps_test]
A    projects/{Name of the MaxCompute project for which the DataWorks data opening package is install
ed}: CreateInstance
A    projects/{Name of the MaxCompute project for which the DataWorks data opening package is install
ed}/packages/u_meta_hangzhou.systables: Read
```

Check whether the data opening package has obtained the metadata of a node

You can create a Shell node and configure it as the ancestor node of the node for which you want to check whether the data opening feature has obtained the metadata.

1. Create a Shell node. For more information about how to create a Shell node, see Create a Shell node.
2. Compile code for the Shell node. In the following example, the raw_v_meta_database_v1_1 view in the data opening package installed for a MaxCompute project that resides in the China (Hangzhou) region is used to query whether the data opening feature has obtained the metadata of a node. Where:
    - **u_meta_hangzhou** specifies the name of the MaxCompute project for which the data opening package is installed. This project resides in the China (Hangzhou) region. You can change the region information in the project name based on your business requirements. For more information about u_meta project names in different regions, see Appendix 2: Available data opening packages.
    - **raw_v_meta_database_v1_1** is a view that is provided by the data opening package for querying the metadata of a node. You can change the view name in the following code to the name of the view that you want to use. For more information about the views provided by the data opening package, see Appendix 1: List and structure details of tables and views.

```
## check if specified view had been produced already
# $1 view name to check
# $2 bizdate to check
# $3 endpoint for this odps project
# $4 name of this odps project
# $5 AccessKey id being used
# $6 AccessKey secret being used
function checkIfSpecifiedViewProduced()
{
  CHECK_SQL="SELECT CASE WHEN COUNT(*) > 0 THEN 'PRODUCED_ALREADY' ELSE 'NOT_PRODUCED_YET' END AS PRODUCE_FLAG FROM u_meta_hangzhou.$1('$2')"
  /opt/taobao/tbdpapp/odpswrapper/odpsconsole/bin/odpscmd --endpoint=$3 --project=$4 -u $5 -p $6 -e "$CHECK_SQL" | grep --color "PRODUCED_ALREADY"
  return $?
}
## check if view raw_v_meta_database_v1_1 had been produced already
checkIfSpecifiedViewProduced "raw_v_meta_database_v1_1" $1 $2 $3 $4 $5
RET_VAL=$?
while [ $RET_VAL -ne  0 ]
do
  echo "DataWorks open data was NOT produced yet, sleep for 300 seconds"
  sleep 300
  checkIfSpecifiedViewProduced "raw_v_meta_database_v1_1" $1 $2 $3 $4 $5
  RET_VAL=$?
done
echo "DataWorks opend data was produced already."
```

> **Note** In the preceding code, `sleep` specifies the interval between two checks (unit: seconds) if no metadata is obtained by using the raw_v_meta_database_v1_1 view. You can change the value of sleep based on the actual situation of the project.

Configure dependencies for the Shell node

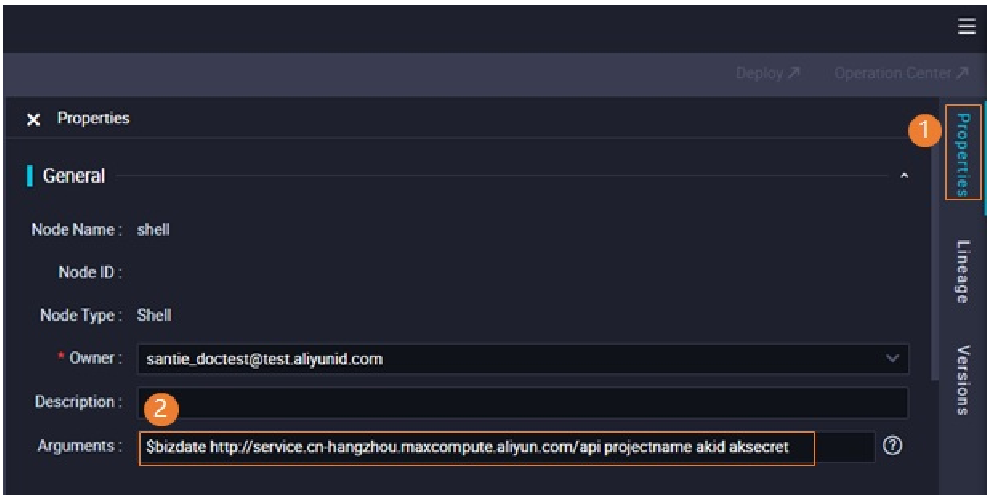
The following items must be configured for the Shell node:

- Scheduling dependency

  The Shell node must be configured as the ancestor node of the node for which you want to check whether the data opening feature has obtained the metadata. Therefore, you must configure the **output** of the Shell node as the **input** of the node. This way, a dependency is established between the node and Shell node. For more information, see Instructions to configure scheduling dependencies.

- Parameters

The code of the Shell node contains the following custom parameters. You must also configure these custom parameters in the General section of the Properties tab. Separate these parameters with spaces. For more information, see Configure scheduling parameters.



Properties

General

Node Name : shell

Node ID :

Node Type : Shell

* Owner : santie_doctest@test.aliyunid.com

Description :

Arguments : $bizdate http://service.cn-hangzhou.maxcompute.aliyun.com/api projectname akid aksecret

- Parameter 1: **$bizdate**, which specifies the date on which the business was performed. This parameter is equivalent to ${yyyymmdd}.
- Parameter 2: a character constant parameter, which can be set to the endpoint of MaxCompute in a specific region.

  For example, your MaxCompute service is activated in the China (Hangzhou) region. In this case, set this parameter to `http://service.cn-hangzhou.maxcompute.aliyun.com/api`. For more information about the endpoints of MaxCompute in other regions, see Endpoints.
- Parameter 3: a character constant parameter, which can be set to the name of the MaxCompute project for which the data opening package is installed.
- Parameter 4: a character constant parameter, which can be set to the AccessKey ID of the RAM user. For more information about how to obtain an AccessKey ID, see Obtain an AccessKey pair.
- Parameter 5: a character constant parameter, which can be set to the AccessKey secret of the RAM user. For more information about how to obtain an AccessKey secret, see Obtain an AccessKey pair.

What to do next

After the Shell node is configured, you can refer to the instructions in Use the data opening feature to create a node for which you want to check whether the data opening feature has obtained the metadata and configure the node. After you commit the node, the Shell node starts to check whether the data opening feature has obtained the metadata of the node. If the Shell node detects that the data opening feature has obtained the metadata of the node, the running of the Shell node is complete. Then, the system starts to run the node. This ensures that you can obtain your desired metadata from the data opening package.

5.3. Use the data opening feature

After the data opening package of DataWorks is installed, you can use the data opening feature to collect metadata in DataWorks and use the metadata for subsequent data governance and O&M. This topic describes the scenarios of the data opening feature and commands required to use the feature.

Prerequisites

The data opening package is installed. For more information, see Install and remove the data opening package.

Instructions

The following sections describe the commands that are used to obtain various types of metadata from a MaxCompute node of DataWorks. Before you use these commands, go to the code editing page of the MaxCompute node.

1. Go to the DataStudio page.

    i.

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

    iii. On the Workspaces page, find the desired workspace and click **Data Analytics** in the Actions column.

2. Go to the code editing page of the MaxCompute node.In the left-side navigation pane, click Business Flow, find your workflow, and then click the workflow name. Then, right-click MaxCompute and choose Create > ODPS SQL. In the Create Node dialog box, configure the parameters required to create a MaxCompute SQL node.

After the MaxCompute SQL node is configured, the code editing page of the node appears. On this page, you can use the data opening feature to query the following types of metadata:

- View the created databases in a MaxCompute project
- View the tables in a MaxCompute project
- View the statistical information of a table
- View the node details of an output table
- View the ancestor and descendant nodes of a node
- Query the owner details of a table or node

View the created databases in a MaxCompute project

The **raw_v_meta_database** view provided by the data opening feature can be used to query the created databases in a MaxCompute project. You can run the following command to perform the operation:

```
SELECT * FROM u_meta_hangzhou.raw_v_meta_database_Version('Business date');
```

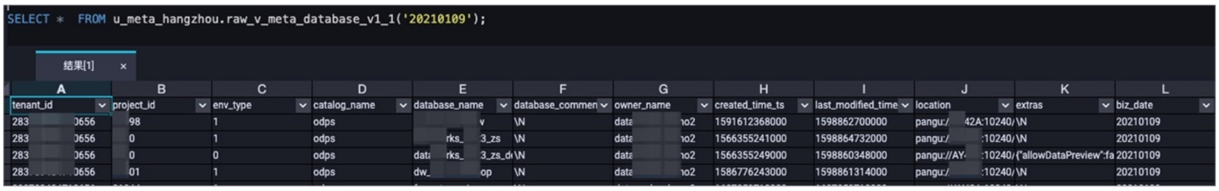
In the preceding command:

- *Version*: specifies the version of the data opening package that you install and use. The format of the version is similar to *v1_1*. You can run the DESCRIBE PACKAGE command to query the actual version of the data opening package that you use. For more information, see View the tables or views provided by the data opening package.
- *Business date*: represents the specific business date on which you want to view the metadata information. The date is presented in the *yyyymmdd* format. If you do not specify a specific business date, the metadata information on all business dates is queried.

The following code provides a command example:

```
-- Query the databases created on January 9, 2021.
SELECT * FROM u_meta_hangzhou.raw_v_meta_database_v1_1('20210109');
```

The following figure shows the query result.



```
SELECT *  FROM u_meta_hangzhou.raw_v_meta_database_v1_1('20210109');
```

结果[1] ×

A	B	C	D	E	F	G	H	I	J	K	L
tenant_id	project_id	env_type	catalog_name	database_name	database_commen	owner_name	created_time_ts	last_modified_time	location	extras	biz_date
283[redacted]0656	[redacted]98	1	odps	[redacted]v	\N	data[redacted]o2	1591612368000	1598862700000	pangu:/[redacted]42A:10240/	\N	20210109
283[redacted]0656	[redacted]0	1	odps	[redacted]rks_[redacted]3_zs	\N	data[redacted]o2	1566355241000	1598864732000	pangu:/[redacted]:10240/	\N	20210109
283[redacted]0656	[redacted]0	0	odps	data[redacted]rks_[redacted]3_zs_d	\N	data[redacted]o2	1566355249000	1598860348000	pangu://AY[redacted]:10240/	{"allowDataPreview":fa	20210109
283[redacted]0656	[redacted]01	1	odps	dw_[redacted]op	\N	data[redacted]o2	1586776243000	1598861314000	pangu:/[redacted]:10240/	\N	20210109

For more information about the fields in the query result, see Metrics in the raw_v_meta_database table.

View the tables in a MaxCompute project

The **raw_v_meta_table** view provided by the data opening feature can be used to query the tables in a MaxCompute project. You can run the following command to perform the operation:

```
SELECT *
 FROM u_meta_hangzhou.raw_v_meta_table_Version('Business date')
 WHERE catalog_name = 'your_catalog_name'
  AND database_name = 'your_database_name'
  AND table_name = 'your_table_name'
;
```

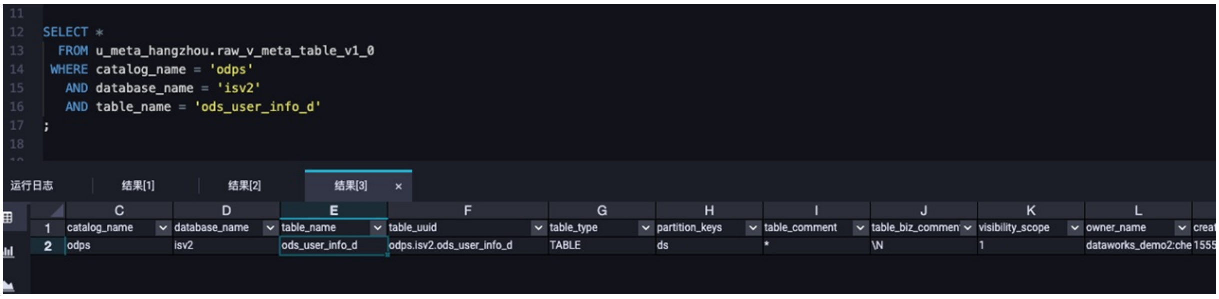
In the preceding command:

- *Version*: specifies the version of the data opening package that you install and use. The format of the version is similar to *v1_1*. You can run the DESCRIBE PACKAGE command to query the actual version of the data opening package that you use. For more information, see View the tables or views provided by the data opening package.
- *Business date*: represents the specific business date on which you want to view the metadata information. The date is presented in the *yyyymmdd* format. If you do not specify a specific business date, the metadata information on all business dates is queried.
- *your_catalog_name*: specifies the computing engine whose metadata you want to view. Set the value to **odps**.
- *your_database_name*: specifies the name of the project in which you want to view metadata information. You can specify this parameter based on your business requirements.
- *your_table_name*: specifies the name of the table whose metadata information you want to view. You can specify this parameter based on your business requirements.

The following code provides a command example:

```
-- Query the data of the ods_user_info_d table under the MaxCompute project isv2 on January 9, 2021.
SELECT *
 FROM u_meta_hangzhou.raw_v_meta_table_v1_1('20210109')
 WHERE catalog_name = 'odps'
  AND database_name = 'isv2'
  AND table_name = 'ods_user_info_d'
;
```

The following figure shows the query result.



```
SELECT *
  FROM u_meta_hangzhou.raw_v_meta_table_v1_0
 WHERE catalog_name = 'odps'
   AND database_name = 'isv2'
   AND table_name = 'ods_user_info_d'
;
```

	C	D	E	F	G	H	I	J	K	L	
1	catalog_name	database_name	table_name	table_uuid	table_type	partition_keys	table_comment	table_biz_commen	visibility_scope	owner_name	creat
2	odps	isv2	ods_user_info_d	odps.isv2.ods_user_info_d	TABLE	ds	*	\N	1	dataworks_demo2:che	1555

For more information about the fields in the query result, see Metrics in the raw_v_meta_table table.

View the statistical information of a table

The **rpt_v_meta_ind_table_core** and **rpt_v_meta_ind_table_extra** views provided by the data opening feature can be used to query the statistical information of a table, such as the tenant ID and lifecycle. You can run the following command to perform the operation:

```
SELECT c.tenant_id, c.table_uuid, c.dim_life_cycle, c.is_partition_table, c.entity_type, c.categories, c.last_access_time, c.partition_count, c.favorite_count, e.output_task_count
FROM u_meta_hangzhou.rpt_v_meta_ind_table_core_Version('Business date') c
LEFT OUTER JOIN u_meta_hangzhou.rpt_v_meta_ind_table_extra_Version('Business date') e
ON c.table_uuid = e.table_uuid AND c.tenant_id = e.tenant_id
WHERE c.catalog_name = 'your_catalog_name'
 AND c.database_name = 'your_database_name'
 AND c.table_name = 'your_table_name'
;
```

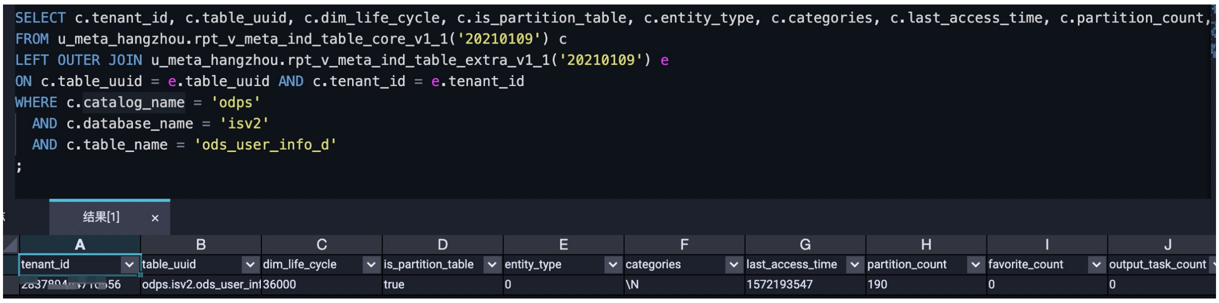
In the preceding command:

- *Version*: specifies the version of the data opening package that you install and use. The format of the version is similar to *v1_1*. You can run the DESCRIBE PACKAGE command to query the actual version of the data opening package that you use. For more information, see View the tables or views provided by the data opening package.
- *Business date*: represents the specific business date on which you want to view the metadata information. The date is presented in the *yyyymmdd* format. If you do not specify a specific business date, the metadata information on all business dates is queried.
- *your_catalog_name*: specifies the computing engine whose metadata you want to view. Set the value to **odps**.
- *your_database_name*: specifies the name of the project in which you want to view metadata information. You can specify this parameter based on your business requirements.
- *your_table_name*: specifies the name of the table whose metadata information you want to view. You can specify this parameter based on your business requirements.
-

The following code provides a command example:

```
-- Query the statistical information of the ods_user_info_d table under the MaxCompute project isv2 on January 9, 2021.
SELECT c.tenant_id, c.table_uuid, c.dim_life_cycle, c.is_partition_table, c.entity_type, c.categories, c.last_access_time, c.partition_count, c.favorite_count, e.output_task_count
FROM u_meta_hangzhou.rpt_v_meta_ind_table_core_v1_1('20210109') c
LEFT OUTER JOIN u_meta_hangzhou.rpt_v_meta_ind_table_extra_v1_1('20210109') e
ON c.table_uuid = e.table_uuid AND c.tenant_id = e.tenant_id
WHERE c.catalog_name = 'odps'
 AND c.database_name = 'isv2'
 AND c.table_name = 'ods_user_info_d'
;
```

The following figure shows the query result.



```
SELECT c.tenant_id, c.table_uuid, c.dim_life_cycle, c.is_partition_table, c.entity_type, c.categories, c.last_access_time, c.partition_count,
FROM u_meta_hangzhou.rpt_v_meta_ind_table_core_v1_1('20210109') c
LEFT OUTER JOIN u_meta_hangzhou.rpt_v_meta_ind_table_extra_v1_1('20210109') e
ON c.table_uuid = e.table_uuid AND c.tenant_id = e.tenant_id
WHERE c.catalog_name = 'odps'
  AND c.database_name = 'isv2'
  AND c.table_name = 'ods_user_info_d'
;
```

A	B	C	D	E	F	G	H	I	J
tenant_id	table_uuid	dim_life_cycle	is_partition_table	entity_type	categories	last_access_time	partition_count	favorite_count	output_task_count
2837[redacted]56	odps.isv2.ods_user_inf	36000	true	0	\N	1572193547	190	0	0

For more information about the fields in the query result, see Core metrics in the rpt_v_meta_ind_table_core table and Additional metrics in the rpt_v_meta_ind_table_extra table.

View the node details of an output table

The **raw_v_meta_table_output** and **raw_v_schedule_node** views provided by the data opening feature can be used to query the node details of an output table. You can run the following command to perform the operation:

```
SELECT s.*, o.schedule_instance_id, execute_time
 FROM u_meta_hangzhou.raw_v_meta_table_output_Version('Business date') o
 LEFT OUTER JOIN u_meta_hangzhou.raw_v_schedule_node_Version('Business date') s
 ON o.schedule_task_id = s.node_id
WHERE o.type = 'your_table_type'
 AND o.database = 'your_database_name'
 AND o.table = 'your_table_name'
 AND s.project_env = 'your_project_environment'
;
```

In the preceding command:

- *Version*: specifies the version of the data opening package that you install and use. The format of the version is similar to *v1_1*. You can run the DESCRIBE PACKAGE command to query the actual version of the data opening package that you use. For more information, see View the tables or views provided by the data opening package.
- *Business date*: represents the specific business date on which you want to view the metadata information. The date is presented in the *yyyymmdd* format. If you do not specify a specific business date, the metadata information on all business dates is queried.
- *your_table_type*: specifies the type of the output table whose metadata you want to view. You can

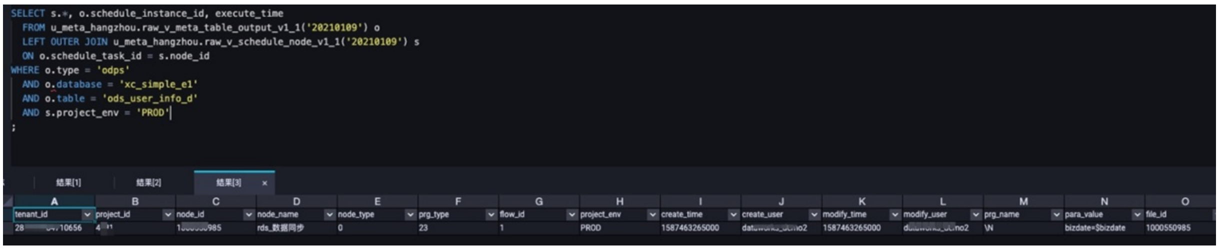
specify only the MaxCompute type. Set the value to **odps**.

- *your_database_name*: specifies the name of the project in which you want to view metadata information. You can specify this parameter based on your business requirements.
- *your_table_name*: specifies the name of the table whose metadata information you want to view. You can specify this parameter based on your business requirements.
- *your_project_environment*: specifies the environment of the MaxCompute project whose metadata you want to view. If you want to view the metadata of a MaxCompute project in a development environment, set the value to **DEV**. If you want to view the metadata of a MaxCompute project in a production environment, set the value to **PROD**.

The following code provides a command example:

```
-- Query the details of the MaxCompute table ods_user_info_d under the xc_simple_e1 project on January 9,
2021.
SELECT s.*, o.schedule_instance_id, execute_time
 FROM u_meta_hangzhou.raw_v_meta_table_output_v1_1('20210109') o
 LEFT OUTER JOIN u_meta_hangzhou.raw_v_schedule_node_v1_1('20210109') s
 ON o.schedule_task_id = s.node_id
WHERE o.type = 'odps'
 AND o.database = 'xc_simple_e1'
 AND o.table = 'ods_user_info_d'
 AND s.project_env = 'PROD'
;
```

The following figure shows the query result.



```
SELECT s.*, o.schedule_instance_id, execute_time
  FROM u_meta_hangzhou.raw_v_meta_table_output_v1_1('20210109') o
  LEFT OUTER JOIN u_meta_hangzhou.raw_v_schedule_node_v1_1('20210109') s
  ON o.schedule_task_id = s.node_id
WHERE o.type = 'odps'
  AND o.database = 'xc_simple_e1'
  AND o.table = 'ods_user_info_d'
  AND s.project_env = 'PROD'
;
```

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
tenant_id	project_id	node_id	node_name	node_type	prg_type	flow_id	project_env	create_time	create_user	modify_time	modify_user	prg_name	para_value	file_id
28[redacted]10656	4[redacted]	1[redacted]985	rds_数据同步	0	23	1	PROD	1587463265000	data[redacted]o2	1587463265000	d[redacted]no2	\N	bizdate=$bizdate	1000550985

For more information about the fields in the query result, see Metrics in the raw_v_meta_table_output table and Metrics in the raw_v_schedule_node table.

View the ancestor and descendant nodes of a node

The **raw_v_schedule_node** and **raw_v_schedule_node_relation** views provided by the data opening feature can be used to query the ancestor and descendant nodes of a node. You can run the following command to perform the operation:

```
-- Query the ancestor nodes of a node.
SELECT *
  FROM u_meta_hangzhou.raw_v_schedule_node_Version('Business date') t
 WHERE t.project_env = 'your_project_environment'
   AND t.node_id IN (
     SELECT parent_node_id
       FROM u_meta_hangzhou.raw_v_schedule_node_relation_Version('Business date') r
      WHERE r.child_node_id = your_child_node_id
        AND r.project_env = 'your_project_environment'
 )
;
-- Query the descendant nodes of a node.
SELECT *
  FROM u_meta_hangzhou.raw_v_schedule_node_Version('Business date') t
 WHERE t.project_env = 'your_project_environment'
   AND t.node_id IN (
     SELECT child_node_id
       FROM u_meta_hangzhou.raw_v_schedule_node_relation_Version('Business date') r
      WHERE r.child_node_id = your_child_node_id
        AND r.project_env = 'your_project_environment'
 )
;
```

In the preceding command:

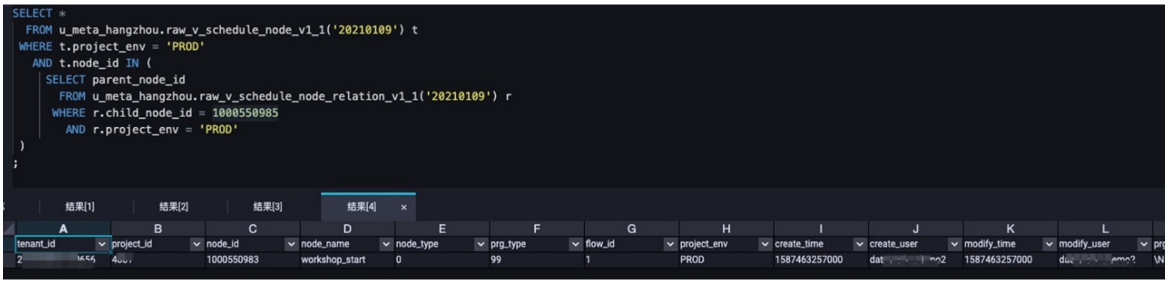
- *Version*: specifies the version of the data opening package that you install and use. The format of the version is similar to *v1_1*. You can run the DESCRIBE PACKAGE command to query the actual version of the data opening package that you use. For more information, see View the tables or views provided by the data opening package.
- *Business date*: represents the specific business date on which you want to view the metadata information. The date is presented in the *yyyymmdd* format. If you do not specify a specific business date, the metadata information on all business dates is queried.
- *your_project_environment*: specifies the environment of the MaxCompute project whose metadata you want to view. If you want to view the metadata of a MaxCompute project in a development environment, set the value to **DEV**. If you want to view the metadata of a MaxCompute project in a production environment, set the value to **PROD**.
- *your_child_node_id*: specifies the ID of the node whose metadata you want to view.

The following code provides command examples:

1. Query the ancestor nodes of a node in a project in a production environment.

```
-- Query the ancestor nodes of the 1000550985 node under a project in a production environment on January 9, 2021.
SELECT *
 FROM u_meta_hangzhou.raw_v_schedule_node_v1_1('20210109') t
 WHERE t.project_env = 'PROD'
  AND t.node_id IN (
    SELECT parent_node_id
      FROM u_meta_hangzhou.raw_v_schedule_node_relation_v1_1('20210109') r
     WHERE r.child_node_id = 1000550985
       AND r.project_env = 'PROD'
 )
;
```

The following figure shows the query result.



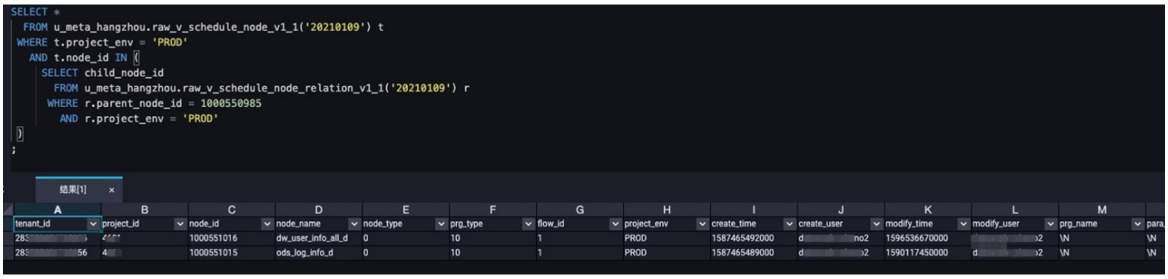
```
SELECT *
  FROM u_meta_hangzhou.raw_v_schedule_node_v1_1('20210109') t
 WHERE t.project_env = 'PROD'
   AND t.node_id IN (
     SELECT parent_node_id
       FROM u_meta_hangzhou.raw_v_schedule_node_relation_v1_1('20210109') r
      WHERE r.child_node_id = 1000550985
        AND r.project_env = 'PROD'
 )
;
```

A	B	C	D	E	F	G	H	I	J	K	L	
tenant_id	project_id	node_id	node_name	node_type	prg_type	flow_id	project_env	create_time	create_user	modify_time	modify_user	prg
[redacted]	[redacted]	1000550983	workshop_start	0	99	1	PROD	1587463257000	[redacted]	1587463257000	[redacted]	\N

2. Query the descendant nodes of a node in a project in a production environment.

```
-- Query the descendant nodes of the 1000550985 node under a project in a production environment on January 9, 2021.
SELECT *
 FROM u_meta_hangzhou.raw_v_schedule_node_v1_1('20210109') t
 WHERE t.project_env = 'PROD'
  AND t.node_id IN (
    SELECT child_node_id
      FROM u_meta_hangzhou.raw_v_schedule_node_relation_v1_1('20210109') r
     WHERE r.parent_node_id = 1000550985
       AND r.project_env = 'PROD'
 )
;
```

The following figure shows the query result.



```
SELECT *
  FROM u_meta_hangzhou.raw_v_schedule_node_v1_1('20210109') t
 WHERE t.project_env = 'PROD'
   AND t.node_id IN (
     SELECT child_node_id
       FROM u_meta_hangzhou.raw_v_schedule_node_relation_v1_1('20210109') r
      WHERE r.parent_node_id = 1000550985
        AND r.project_env = 'PROD'
)
;
```

A	B	C	D	E	F	G	H	I	J	K	L	M	
tenant_id	project_id	node_id	node_name	node_type	prg_type	flow_id	project_env	create_time	create_user	modify_time	modify_user	prg_name	para
[redacted]	[redacted]	1000551016	dw_user_info_all_d	0	10	1	PROD	1587465492000	[redacted]	1596536670000	[redacted]	\N	\N
[redacted]	[redacted]	1000551015	ods_log_info_d	0	10	1	PROD	1587465489000	[redacted]	1590117450000	[redacted]	\N	\N

For more information about the fields in the query result, see Metrics in the raw_v_schedule_node table and Metrics in the raw_v_schedule_node_relation table.

Query the owner details of a table or node

The **raw_v_meta_table** and **raw_v_tenant_user** views provided by the data opening feature can be used to query the owner details of a table or node. You can run the following command to perform the operation:

- Query the owner details of a table.

  ```
  SELECT
    c.catalog_name,
    c.database_name,
    c.table_name,
    c.owner_name,
    u.account_name,
    u.nick
  FROM u_meta_hangzhou.raw_v_meta_table_Version('Business date') c
   LEFT OUTER JOIN u_meta_hangzhou.raw_v_tenant_user_Version('Business date') u
    ON c.owner_name = TOLOWER(u.yun_account)
   WHERE c.catalog_name = 'your_catalog_name'
    AND c.database_name = 'your_database_name'
    AND c.table_name = 'your_table_name'
  ;
  ```

  In the preceding command:

  - *Version*: specifies the version of the data opening package that you install and use. The format of the version is similar to *v1_1*. You can run the DESCRIBE PACKAGE command to query the actual version of the data opening package that you use. For more information, see View the tables or views provided by the data opening package.
  - *Business date*: represents the specific business date on which you want to view the metadata information. The date is presented in the *yyyymmdd* format. If you do not specify a specific business date, the metadata information on all business dates is queried.
  - *your_catalog_name*: specifies the computing engine whose metadata you want to view. Set the value to **odps**.
  - *your_database_name*: specifies the name of the project in which you want to view metadata information. You can specify this parameter based on your business requirements.
  - *your_table_name*: specifies the name of the table whose metadata information you want to view. You can specify this parameter based on your business requirements.
  -

- Query the owner details of a node.

  ```
  SELECT t.project_id, t.node_id, t.node_name,
    t.create_user, u.account_name AS create_user_name, u.nick as create_user_nick,
    t.modify_user, m.account_name AS modify_user_name, m.nick as modify_user_nick
   FROM u_meta_hangzhou.raw_v_schedule_node_Version('Business date') t
  LEFT OUTER JOIN u_meta_hangzhou.raw_v_tenant_user_Version('Business date') u ON t.create_user = u.yun_account
  LEFT OUTER JOIN u_meta_hangzhou.raw_v_tenant_user_Version('Business date') m ON t.modify_user = m.yun_account
   WHERE t.node_id = your_node_id
    AND t.project_env = 'your_project_environment'
  ;
  ```

  In the preceding command:

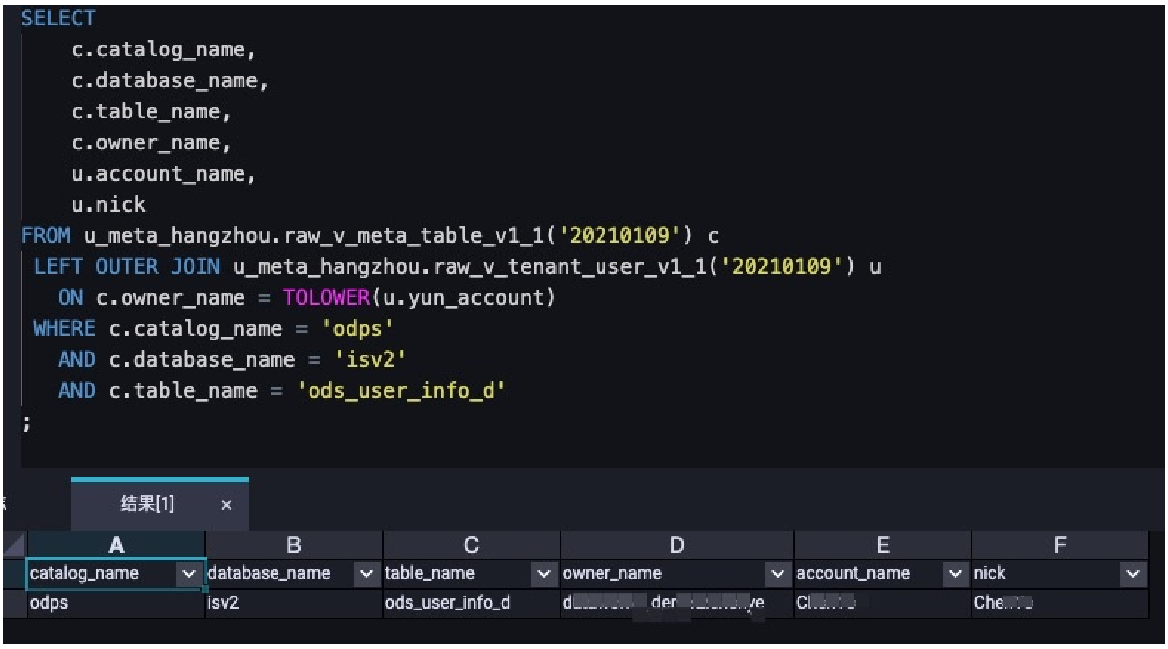
- *Version*: specifies the version of the data opening package that you install and use. The format of the version is similar to *v1_1*. You can run the DESCRIBE PACKAGE command to query the actual version of the data opening package that you use. For more information, see View the tables or views provided by the data opening package.
- *Business date*: represents the specific business date on which you want to view the metadata information. The date is presented in the *yyyymmdd* format. If you do not specify a specific business date, the metadata information on all business dates is queried.
- *your_project_environment*: specifies the environment of the MaxCompute project whose metadata you want to view. If you want to view the metadata of a MaxCompute project in a development environment, set the value to **DEV**. If you want to view the metadata of a MaxCompute project in a production environment, set the value to **PROD**.
- *your_node_id*: specifies the ID of the node whose metadata information you want to view.

The following code provides command examples:

1. Query the owner details of a table on January 9, 2021.

```
SELECT
  c.catalog_name,
  c.database_name,
  c.table_name,
  c.owner_name,
  u.account_name,
  u.nick
FROM u_meta_hangzhou.raw_v_meta_table_v1_1('20210109') c
 LEFT OUTER JOIN u_meta_hangzhou.raw_v_tenant_user_v1_1('20210109') u
  ON c.owner_name = TOLOWER(u.yun_account)
 WHERE c.catalog_name = 'odps'
  AND c.database_name = 'isv2'
  AND c.table_name = 'ods_user_info_d'
;
```

The following figure shows the query result.



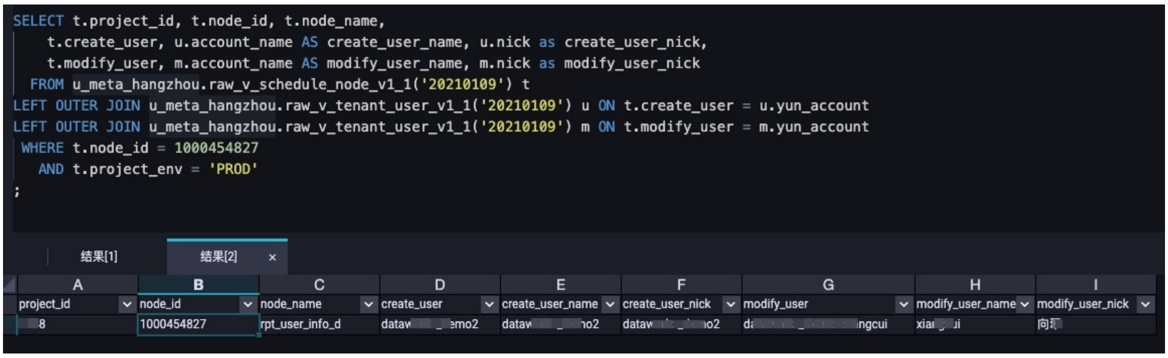
```
SELECT
    c.catalog_name,
    c.database_name,
    c.table_name,
    c.owner_name,
    u.account_name,
    u.nick
FROM u_meta_hangzhou.raw_v_meta_table_v1_1('20210109') c
 LEFT OUTER JOIN u_meta_hangzhou.raw_v_tenant_user_v1_1('20210109') u
   ON c.owner_name = TOLOWER(u.yun_account)
 WHERE c.catalog_name = 'odps'
   AND c.database_name = 'isv2'
   AND c.table_name = 'ods_user_info_d'
;
```

A	B	C	D	E	F
catalog_name	database_name	table_name	owner_name	account_name	nick
odps	isv2	ods_user_info_d	[redacted]	[redacted]	[redacted]

2. Query the owner details of an auto triggered node and the details of the user who modifies the auto triggered node on January 9, 2021.

```
SELECT t.project_id, t.node_id, t.node_name,
  t.create_user, u.account_name AS create_user_name, u.nick as create_user_nick,
  t.modify_user, m.account_name AS modify_user_name, m.nick as modify_user_nick
 FROM u_meta_hangzhou.raw_v_schedule_node_v1_1('20210109') t
LEFT OUTER JOIN u_meta_hangzhou.raw_v_tenant_user_v1_1('20210109') u ON t.create_user = u.yun_account
LEFT OUTER JOIN u_meta_hangzhou.raw_v_tenant_user_v1_1('20210109') m ON t.modify_user = m.yun_account
 WHERE t.node_id = 1000454827
  AND t.project_env = 'PROD'
;
```

The following figure shows the query result.



```
SELECT t.project_id, t.node_id, t.node_name,
    t.create_user, u.account_name AS create_user_name, u.nick as create_user_nick,
    t.modify_user, m.account_name AS modify_user_name, m.nick as modify_user_nick
  FROM u_meta_hangzhou.raw_v_schedule_node_v1_1('20210109') t
LEFT OUTER JOIN u_meta_hangzhou.raw_v_tenant_user_v1_1('20210109') u ON t.create_user = u.yun_account
LEFT OUTER JOIN u_meta_hangzhou.raw_v_tenant_user_v1_1('20210109') m ON t.modify_user = m.yun_account
 WHERE t.node_id = 1000454827
   AND t.project_env = 'PROD'
;
```

A	B	C	D	E	F	G	H	I
project_id	node_id	node_name	create_user	create_user_name	create_user_nick	modify_user	modify_user_name	modify_user_nick
[redacted]8	1000454827	rpt_user_info_d	dataw[redacted]emo2	dataw[redacted]o2	dataw[redacted]o2	d[redacted]ngcui	xia[redacted]i	向[redacted]

For more information about the fields in the query result, see Metrics in the raw_v_schedule_node table and Metrics in the raw_v_schedule_node_relation table.

What's next

The views provided by the data opening feature support date parameters in the yyyyMMdd format. You can create partitions based on dates and query historical data over the last 30 days. If you have requirements such as metric trend analysis, you can query data by business date. Then, you can save the data to a project and use the data as the historical data required to perform metric analysis.

5.4. Appendix 1: List and structure details of tables and views

The data opening feature of DataWorks provides tables and views in various dimensions for you to collect metadata. This topic provides a list of such tables and views and describes the structures of these tables and views.

- Metadata
  - RPT metrics
    - Core metrics in the rpt_v_meta_ind_table_core table
    - Additional metrics in the rpt_v_meta_ind_table_extra table

  - Metrics that are related to metadata details
    - Metrics in the raw_v_meta_database table
    - Metrics in the raw_v_meta_table table
    - Metrics in the raw_v_meta_view table
    - Metrics in the raw_v_meta_column table
    - Metrics in the raw_v_meta_partition table
    - Metrics in the raw_v_meta_table_lineage table
    - Metrics in the raw_v_meta_table_output table
    - Metrics in the raw_v_meta_table_usage table
    - Metrics in the raw_v_meta_column_usage table
    - Metrics in the raw_v_meta_biz_table_wiki table
    - Metrics in the raw_v_meta_table_join_map table
    - Metrics in the raw_v_meta_table_detail_log table
    - Metrics in the raw_v_meta_category table
- Scheduling metadata
  - Metrics in the raw_v_schedule_node table
  - Metrics in the raw_v_schedule_task table
  - Metrics in the raw_v_schedule_node_relation table
  - Metrics in the raw_v_schedule_di_resgroup table
- Tenant metadata
  - Metrics in the raw_v_tenant_res_group table
  - Metrics in the raw_v_tenant_user table
  - Metrics in the raw_v_tenant_workspace table
  - Metrics in the raw_v_tenant_workspace_user table

Core metrics in the rpt_v_meta_ind_table_core table

Metric name	Data type	Description
tenant_id	bigint	The ID of the DataWorks tenant.
project_id	bigint	The ID of the DataWorks workspace.
catalog_name	string	The catalog to which the table belongs. This metric is set to odps for MaxCompute projects.
database_name	string	The name of the database or MaxCompute project.
table_name	string	The name of the table.
table_uuid	string	The unique ID of the table.
owner_yun_acct	string	The Alibaba Cloud account of the table owner.

Metric name	Data type	Description
dim_life_cycle	bigint	The time to live (TTL). Unit: days. • 0: indicates that no TTL is configured. • Other values: indicate a specific period of time.
is_partition_table	boolean	Specifies whether the table is a partitioned table. • true: The table is a partitioned table. • false: The table is a non-partitioned table.
entity_type	bigint	The entity type. • 0: table • 1: view
categories	string	The detailed information about the categories.
last_access_time	bigint	The last time when the table was accessed. The metric value is a 10-digit UNIX timestamp.
size	bigint	The size of the table, which indicates the logical storage space that is occupied by data in the table. Unit: byte. This metric is set to NULL for a view.
column_count	bigint	The number of fields in the table. Partition key columns are included.
partition_count	bigint	The number of partitions in the table. This metric is set to NULL for a non-partitioned table.
detail_view_count	bigint	The number of times that table details are viewed on the page.
favorite_count	bigint	The number of times that the table is added to favorites.

Additional metrics in the rpt_v_meta_ind_table_extra table

Metric name	Data type	Description
tenant_id	bigint	The ID of the DataWorks tenant.
table_uuid	string	The unique ID of the table.
read_count	bigint	The number of times that data is read by using SQL statements. The data includes that of non-scheduled nodes.
read_count_30d	bigint	The number of times that data is read within 30 days by using SQL statements. The data includes that of non-scheduled nodes.

Metric name	Data type	Description
write_count	bigint	The number of times that data is written by using SQL statements. The data includes that of non-scheduled nodes.
join_count	bigint	The number of times that the table is joined.
direct_upstream_count	bigint	The number of parent tables in the lineage.
direct_downstream_count	bigint	The number of child tables in the lineage.
output_task_count	bigint	The number of nodes that generate the data in the table.

Metrics in the raw_v_meta_database table

Metric name	Data type	Description
tenant_id	bigint	The ID of the DataWorks tenant.
project_id	bigint	The ID of the DataWorks workspace.
env_type	bigint	The environment type. ● 0: development environment ● 1: production environment
catalog_name	string	The catalog to which the table belongs. This metric is set to odps for MaxCompute projects.
database_name	string	The name of the database or MaxCompute project.
database_comment	string	The description of the database or MaxCompute project.
owner_name	string	The name of the owner.
created_time_ts	bigint	The creation time. The metric value is a 13-digit timestamp.
last_modified_time_ts	bigint	The last modification time. The metric value is a 13-digit timestamp.
location	string	The storage path of the table in the database.

Metric name	Data type	Description
extras	string	The additional information about the database, which is a JSON string. If the table preview and table visibility range attributes are configured for a MaxCompute project, you can use the allowDataPreview and projectVisibility keys to obtain the values of the attributes. • allowDataPreview: specifies whether tables in a MaxCompute project can be previewed. ○ true: Tables in a MaxCompute project can be previewed. ○ Other values or NULL: Tables in a MaxCompute project cannot be previewed. • projectVisibility: specifies the visibility range of tables in a MaxCompute project. ○ 0: hidden. Tables are visible only for table owners, project administrators, and project owners. ○ 1: visible for tenants. ○ 2: visible for project members.
biz_date	string	The data timestamp.

Metrics in the raw_v_meta_table table

Metric name	Data type	Description
tenant_id	bigint	The ID of the DataWorks tenant.
project_id	string	The ID of the DataWorks workspace.
table_uuid	string	The unique ID of the table.
table_name	string	The name of the table.
table_type	string	The type of the table.
catalog_name	string	The catalog to which the table belongs. This metric is set to odps for MaxCompute projects.
database_name	string	The name of the database or MaxCompute project.
partition_keys	string	The partition keys in the table. Multi-level partitions are separated by commas (,). This metric is set to an empty string for a non-partitioned table.
table_comment	string	The description of the table.

Metric name	Data type	Description
table_biz_comment	string	The business description of the table.
visibility_scope	bigint	The visibility range of the table. • 0: hidden. Tables are visible only for table owners, project administrators, and project owners. • 1: visible for tenants. • 2: visible for project members. •
owner_name	string	The name of the owner.
created_time_ts	bigint	The creation time. The metric value is a 13-digit timestamp.
last_modified_time_ts	bigint	The last time when data was modified. The metric value is a 13-digit timestamp.
last_meta_modified_time_ts	bigint	The last time when table metadata was modified. The metric value is a 13-digit timestamp.
location	string	The storage path of the table.
life_cycle	bigint	The TTL of the table. Unit: days.
data_size	bigint	The logical storage volume of the table. Unit: byte. If the table is a partitioned table, this metric is set to NULL. You must collect statistics on the storage volume based on the partition list.
biz_date	string	The data timestamp.

Metrics in the raw_v_meta_view table

Metric name	Data type	Description
tenant_id	bigint	The ID of the DataWorks tenant.
project_id	string	The ID of the DataWorks workspace.
table_uuid	string	The unique ID of the table.
table_name	string	The name of the table.
catalog_name	string	The catalog to which the table belongs. This metric is set to odps for MaxCompute projects.
database_name	string	The name of the database or MaxCompute project.
table_comment	string	The description of the table.

Metric name	Data type	Description
table_biz_comment	string	The business description of the table.
visibility_scope	bigint	The visibility range of the table. - 0: hidden. Tables are visible only for table owners, project administrators, and project owners. - 1: visible for tenants. - 2: visible for project members. -
owner_name	string	The name of the owner.
created_time_ts	bigint	The creation time. The metric value is a 13-digit timestamp.
last_ddl_time_ts	bigint	The last time when the view was modified by using data definition language (DDL) statements. The metric value is a 13-digit timestamp.
view_text	string	The SQL statement that is used to create a view.
biz_date	string	The data timestamp.

Metrics in the raw_v_meta_column table

Metric name	Data type	Description
tenant_id	bigint	The ID of the DataWorks tenant.
project_id	bigint	The ID of the DataWorks workspace.
catalog_name	string	The catalog to which the table belongs. This metric is set to odps for MaxCompute projects.
database_name	string	The name of the database or MaxCompute project.
table_name	string	The name of the table.
column_name	string	The name of the field.
column_comment	string	The description of the field.
column_biz_comment	string	The business description of the field.
column_type	string	The data type of the field.
column_sequence	bigint	The sequence number of the field, which starts from 1.
is_partition_key	boolean	Specifies whether the field is a partition key.

Metric name	Data type	Description
is_primary_key	boolean	Specifies whether the field is a primary key.
biz_date	string	The data timestamp.

Metrics in the raw_v_meta_partition table

Metric name	Data type	Description
tenant_id	bigint	The ID of the DataWorks tenant.
project_id	bigint	The ID of the DataWorks workspace.
catalog_name	string	The catalog to which the table belongs. This metric is set to odps for MaxCompute projects.
database_name	string	The name of the database or MaxCompute project.
table_name	string	The name of the table.
partition_name	string	The name of the partition.
size	bigint	The logical size of the partition. Unit: byte.
record_number	bigint	The number of records in the partition.
created_time_ts	bigint	The creation time. The metric value is a 13-digit timestamp.
last_modified_time_ts	bigint	The last modification time. The metric value is a 13-digit timestamp.
biz_date	string	The data timestamp.

Metrics in the raw_v_meta_table_lineage table

Note The lineage feature cannot achieve 100% data integrity and accuracy due to the complexity of SQL statements and code. We recommend that you do not use this feature for the business that has integrity and accuracy requirements.

Metric name	Data type	Description
tenant_id	bigint	The ID of the DataWorks tenant.
project_id	bigint	The ID of the DataWorks workspace.
src_type	string	The type of the data source.
src_data_source_id	string	The ID of the data source.

Metric name	Data type	Description
src_database	string	The source database.
src_table	string	The source table.
dest_type	string	The type of the data destination.
dest_data_source_id	string	The ID of the data destination.
dest_database	string	The destination database.
dest_table	string	The destination table.
schedule_task_id	string	The ID of the scheduled node.
schedule_instance_id	string	The instance ID of the scheduled node.
schedule_task_owner	string	The owner of the scheduled node.
job_start_time_ts	bigint	The start time of the node, which is a 13-digit timestamp.
job_end_time_ts	bigint	The end time of the node, which is a 13-digit timestamp.
execute_time	bigint	The time that is required to run the node. Unit: seconds.
input_record_number	bigint	The number of records that were read from the source table.
biz_date	string	The data timestamp.

Metrics in the raw_v_meta_table_output table

Data Map displays the tables whose data is generated by ODPS nodes. The metadata includes the tables whose data is generated by ODPS nodes and data integration nodes.

Note The output information is generated based on lineage.

Metric name	Data type	Description
tenant_id	bigint	The ID of the DataWorks tenant.
project_id	bigint	The ID of the DataWorks workspace in which scheduled nodes are run.
type	string	The type of the data source.
data_source_id	string	The ID of the data source.
database	string	The database.

Metric name	Data type	Description
table	string	The name of the table.
schedule_task_id	string	The ID of the scheduled node.
schedule_instance_id	string	The instance ID of the scheduled node.
schedule_task_owner	string	The owner of the scheduled node.
job_start_time_ts	bigint	The start time of the node, which is a 13-digit timestamp.
job_end_time_ts	bigint	The end time of the node, which is a 13-digit timestamp.
execute_time	bigint	The time that is required to run the node. Unit: seconds.
biz_date	string	The data timestamp.

Metrics in the raw_v_meta_table_usage table

Metric name	Data type	Description
tenant_id	bigint	The ID of the DataWorks tenant.
project_id	bigint	The ID of the DataWorks workspace in which scheduled nodes are run.
catalog_name	string	The catalog to which the table belongs. This metric is set to odps for MaxCompute projects.
database_name	string	The name of the database or MaxCompute project.
table_name	string	The name of the table.
schedule_task_id	string	The ID of the scheduled node.
schedule_task_owner	string	The owner of the scheduled node. If the node is not scheduled in DataWorks, this metric is set to NULL.
job_id	string	The node ID, which may not be the instance ID of the node that is scheduled in DataWorks. You can use this metric to count the number of times that data is read from the table and the number of times that data is written to the table.
op_type	string	The operation type, which can be READ, WRITE, or UNKNOWN.

Metric name	Data type	Description
extras	string	The additional information, which is a JSON string. If a MaxCompute node is run to perform operations on a table, you can use the task_name key to obtain the name of the MaxCompute node. If the ID of a node that is scheduled in DataWorks is not empty, you can use the schedule_task_name key to obtain the name of the scheduled node. Example: `{ "task_name": "console_query_task_1605629400000", "schedule_task_name": "Test SQL node" }`.
biz_date	string	The data timestamp.

Metrics in the raw_v_meta_column_usage table

Metric name	Data type	Description
tenant_id	bigint	The ID of the DataWorks tenant.
project_id	bigint	The ID of the DataWorks workspace in which scheduled nodes are run.
catalog_name	string	The catalog to which the table belongs. This metric is set to odps for MaxCompute projects.
database_name	string	The name of the database or MaxCompute project.
table_name	string	The name of the table.
column_name	string	The name of the field.
schedule_task_id	string	The ID of the scheduled node.
schedule_task_owner	string	The owner of the scheduled node. If the node is not scheduled in DataWorks, this metric is set to NULL.
inst_id	string	The node ID, which may not be the instance ID of the node that is scheduled in DataWorks.
op_type	string	The operation type, which can be SELECT, JOIN, GROUP BY, or WHERE.

Metric name	Data type	Description
extras	string	The additional information, which is a JSON string. If a MaxCompute node is run to perform operations on a table, you can use the task_name key to obtain the name of the MaxCompute node. If the ID of a node that is scheduled in DataWorks is not empty, you can use the schedule_task_name key to obtain the name of the scheduled node. Example: `{ "task_name": "console_query_task_1605629400000", "schedule_task_name": "Test SQL node" }`.
biz_date	string	The data timestamp.

Metrics in the raw_v_meta_biz_table_wiki table

Metric name	Data type	Description
tenant_id	bigint	The ID of the DataWorks tenant.
project_id	bigint	The ID of the DataWorks workspace in which scheduled nodes are run.
catalog_name	string	The catalog to which the table belongs. This metric is set to odps for MaxCompute projects.
database_name	string	The name of the database or MaxCompute project.
table_name	string	The name of the table.
version	string	The version number of Wiki.
operator	string	The final operator, which may be an owner of the table.
content	string	The content of Wiki, which is written by using the Markdown syntax.
update_time_ts	bigint	The modification time. The metric value is a 13-digit timestamp.
biz_date	string	The data timestamp.

Metrics in the raw_v_meta_table_join_map table

Metric name	Data type	Description
tenant_id	bigint	The ID of the DataWorks tenant.

Metric name	Data type	Description
catalog_name	string	The catalog to which the table belongs. This metric is set to odps for MaxCompute projects.
database_name	string	The name of the database or MaxCompute project.
table_name	string	The name of the table.
column_name	string	The name of the field.
join_database_name	string	The name of the associated database or MaxCompute project.
join_table_name	string	The name of the associated table.
join_column_name	string	The name of the associated field.
join_type	string	The type of the JOIN operation, which can be left, right, or inner.
schedule_task_id	string	The ID of the scheduled node.
schedule_task_owner	string	The owner of the scheduled node.
job_id	string	The ID of the node at the engine layer.
extras	string	The additional information, which is a JSON string. If a MaxCompute node is run to perform operations on a table, you can use the task_name key to obtain the name of the MaxCompute node.
biz_date	string	The data timestamp.

Metrics in the raw_v_meta_table_detail_log table

Metric name	Data type	Description
tenant_id	bigint	The ID of the DataWorks tenant.
catalog_name	string	The catalog to which the table belongs. This metric is set to odps for MaxCompute projects.
database_name	string	The name of the database or MaxCompute project.
table_name	string	The name of the table.
operator	string	The user who views table details.
view_time_ts	bigint	The time when table details are viewed. The metric value is a 13-digit timestamp.
biz_date	string	The data timestamp.

Metrics in the raw_v_meta_category table

Metric name	Data type	Description
tenant_id	bigint	The ID of the DataWorks tenant.
category_id	bigint	The ID of the category.
category_name	string	The name of the category.
category_pid	bigint	The ID of the parent category. This metric is set to 0 or NULL for a level 1 category.
depth	bigint	The level of the category. This metric is set to 1 for a level 1 category.
sort_field	double	The field based on which the categories are sorted.
creator_account	string	The account that creates the category.
created_time_ts	bigint	The creation time. The metric value is a 13-digit timestamp.
last_modified_time_ts	bigint	The last modification time. The metric value is a 13-digit timestamp.
biz_date	string	The data timestamp.

Metrics in the raw_v_schedule_node table

Metric name	Data type	Description
tenant_id	bigint	The ID of the tenant.
project_id	bigint	The ID of the DataWorks workspace.
node_id	bigint	The ID of the node.
node_name	string	The name of the node.
node_type	bigint	The scheduling type of the node. • 0: auto triggered node • 1: manually triggered node • 2: paused node • 3: dry-run node
prg_type	bigint	The type of the node. • 10: ODPS SQL node • 23: data integration node

Metric name	Data type	Description
flow_id	bigint	The ID of the workflow.
project_env	string	The environment type. - PROD: production environment - DEV: development environment
create_time	bigint	The creation time. The metric value is a 13-digit timestamp.
create_user	string	The creator.
modify_time	bigint	The last modification time. The metric value is a 13-digit timestamp.
modify_user	string	The user who modifies the node.
prg_name	string	The name of the node type.
para_value	string	The execution parameter.
file_id	bigint	The ID of the file.
file_version	bigint	The file version.
owner	string	The owner of the node.
resgroup_id	bigint	The ID of the resource group.
baseline_id	bigint	The ID of the baseline.
cycle_type	bigint	The recurrence. - 0: daily, weekly, or monthly - Other values: hourly or minutely
repeatable	bigint	The rerun identifier. - 0: Only failed nodes can be rerun. - 1: All nodes can be rerun. - 2: No nodes can be rerun.
connection	string	The connection string of the data source.
dqc_type	bigint	Specifies whether the node uses the Data Quality service. - 0: The node uses the Data Quality service. - 1: The node does not use the Data Quality service.
dqc_description	string	The Data Quality rule.

Metric name	Data type	Description
task_rerun_time	bigint	The number of times that the task can be rerun.
task_rerun_interval	bigint	The rerun interval. Unit: milliseconds.
cron_express	string	The CRON expression that specifies the scheduling frequency of the node.
priority	bigint	The priority of the task. Valid values: 1, 3, 5, 7, and 8. A greater value indicates a higher priority.
start_effect_date	bigint	The time when the node takes effect. The metric value is a 13-digit timestamp.
end_effect_date	bigint	The time when the node loses effect. The metric value is a 13-digit timestamp.
biz_date	string	The data timestamp.

Metrics in the raw_v_schedule_task table

Metric name	Data type	Description
tenant_id	bigint	The ID of the tenant.
project_id	bigint	The ID of the DataWorks workspace.
node_id	bigint	The ID of the node.
node_name	string	The name of the node.
task_id	bigint	The name of the task.
dag_id	bigint	The directed acyclic graph (DAG) ID of the workflow.
task_type	bigint	The scheduling type of the task. • 0: auto triggered task • 1: manually triggered task • 2: paused task • 3 or 5: dry-run task
dag_type	bigint	The DAG type. • 0: for auto triggered nodes • 1: for manually triggered nodes • 3: for retroactive data generation
prg_type	bigint	The type of the node. • 10: ODPS SQL node • 23: data integration node

Metric name	Data type	Description
flow_id	bigint	The ID of the workflow.
create_time	bigint	The creation time. The metric value is a 13-digit timestamp.
modify_time	bigint	The last modification time. The metric value is a 13-digit timestamp.
cycle_time	bigint	The scheduling time, which is a 13-digit timestamp.
in_group_id	bigint	The serial number of the task.
prg_name	string	The name of the node type.
para_value	string	The execution parameter.
file_id	bigint	The ID of the file.
file_version	bigint	The file version.
owner	string	The owner of the node.
resgroup_id	bigint	The ID of the resource group.
baseline_id	bigint	The ID of the baseline.
cycle_type	bigint	The recurrence. ● 0: daily, weekly, or monthly ● Other values: hourly or minutely ●
repeatable	bigint	The rerun identifier. ● 0: Only failed nodes can be rerun. ● 1: All nodes can be rerun. ● 2: No nodes can be rerun. ●
connection	string	The connection string of the data source.
dqc_type	bigint	Specifies whether the node uses the Data Quality service. ● 0: The node uses the Data Quality service. ● 1: The node does not use the Data Quality service. ●
dqc_description	string	The Data Quality rule.
task_rerun_time	bigint	The number of times that the task can be rerun.

Metric name	Data type	Description
task_rerun_interval	bigint	The rerun interval. Unit: milliseconds.
begin_waittime_time	bigint	The time when the node starts to wait for scheduling. The metric value is a 13-digit timestamp.
finish_time	bigint	The time when the running is complete. The metric value is a 13-digit timestamp.
begin_waitres_time	bigint	The time when the node starts to wait for resource allocation. The metric value is a 13-digit timestamp.
begin_run_time	bigint	The time when the node starts to run. The metric value is a 13-digit timestamp.
rerun_times	bigint	The number of times that the task is rerun.
priority	bigint	The priority of the task. Valid values: 1, 3, 5, 7, and 8. A greater value indicates a higher priority.
task_key	string	The unique identifier of the task.
error_msg	string	The reason why the task failed.
status	bigint	The status of the task. • NOT_RUN(1, "Not all ancestor instances are successful.") • WAIT_TIME(2, "The task is waiting for the scheduling time that is specified by dueTime or cycleTime to arrive.") • WAIT_RESOURCE(3, "The task is delivered to the execution engine Alisa and is waiting for scheduling in a queue.") • RUNNING(4, "The task is being run.") • CHECKING(7, "The task is run by using Alisa, and data is delivered to Data Quality for verification.") • CHECKING_CONDITION(8, "The task is run by using Alisa, and branch conditions are being checked.") • FAILURE(5, "The task failed.") • SUCCESS(6, "The task is successful.")
biz_date	string	The data timestamp.

Metrics in the raw_v_schedule_node_relation table

Metric name	Data type	Description
tenant_id	bigint	The ID of the tenant.

Metric name	Data type	Description
child_node_id	bigint	The ID of the descendant node.
parent_node_id	bigint	The ID of the ancestor node.
step_type	bigint	The dependency type. ● 0: common ● 3: cross-cycle
child_flow_id	bigint	The ID of the workflow.
project_env	string	The environment type. ● PROD: production environment ● DEV: development environment
create_time	bigint	The creation time. The metric value is a 13-digit timestamp.
create_user	string	The creator.
modify_time	bigint	The last modification time. The metric value is a 13-digit timestamp.
modify_user	string	The user who modifies the node.
biz_date	string	The data timestamp.

Metrics in the raw_v_schedule_di_resgroup table

Metric name	Data type	Description
tenant_id	bigint	The ID of the tenant.
project_id	bigint	The ID of the DataWorks workspace.
node_id	bigint	The ID of the node.
project_env	string	The environment of the workspace.
res_group_identifier	string	The ID of the resource group for Data Integration.
src_type	string	The type of the data source.
dst_type	string	The type of the data destination.
src_datasource	string	The data source.
dst_datasource	string	The data destination.
config_concurrent	bigint	The number of concurrent nodes.

Metric name	Data type	Description
biz_date	string	The data timestamp.

Metrics in the raw_v_tenant_res_group table

Metric name	Data type	Description
tenant_id	bigint	The ID of the tenant.
res_group_id	bigint	The ID of the resource group.
res_group_identifier	string	The identifier of the resource group.
res_group_type	bigint	The type of the resource group. • 1: resource group for scheduling • 2: MaxCompute quota group • 4: resource group for Data Integration
res_group_mode	bigint	The billing method of the resource group. • 1: subscription • 2: pay-as-you-go • 3: Developer Edition (available only for MaxCompute)
status	bigint	The status of the resource group. • 0: The resource group is normal. • 1: The resource group is frozen. • 2: The resource group is deleted. • 3: The resource group is being created. • 4: The resource group fails to be created. • 5: The resource group is being updated. • 6: The resource group fails to be updated. • 7: The resource group is being deleted. • 8: The resource group fails to be deleted.
biz_ext_key	string	The extension field of the resource group. A value of single indicates an exclusive resource group.
biz_date	string	The data timestamp.

Metrics in the raw_v_tenant_user table

Metric name	Data type	Description
tenant_id	bigint	The ID of the tenant.
yun_account	string	The Alibaba Cloud account.

Metric name	Data type	Description
account_name	string	The name of the account.
nick	string	The display name of the account.
full_yun_account	string	The Alibaba Cloud account that contains the account provider information.
biz_date	string	The data timestamp.

Metrics in the raw_v_tenant_workspace table

Metric name	Data type	Description
tenant_id	bigint	The ID of the tenant.
project_id	bigint	The ID of the workspace.
project_name	string	The name of the workspace.
project_identifier	string	The identifier of the workspace.
project_desc	string	The description of the workspace.
project_owner	string	The owner of the workspace.
status	bigint	The status of the workspace. • 0: The workspace is normal. • 1: The workspace is deleted. • 2: The workspace is being initialized. • 3: The workspace fails to be initialized. • 4: The workspace is manually disabled. • 5: The workspace is being deleted. • 6: The workspace fails to be deleted. • 7: The workspace is frozen due to overdue payments.
biz_date	string	The data timestamp.

Metrics in the raw_v_tenant_workspace_user table

Metric name	Data type	Description
tenant_id	bigint	The ID of the DataWorks tenant.
project_id	bigint	The ID of the DataWorks workspace.
base_id	string	The base ID of the user.

Metric name	Data type	Description
status	bigint	The status of the user. • 0: The user is normal. • 1: The user is disabled. • 2: The user is deleted.
gmt_create_ts	bigint	The creation time. The metric value is a 13-digit timestamp.
gmt_modified_ts	bigint	The last modification time. The metric value is a 13-digit timestamp.
biz_date	string	The data timestamp.

5.5. Appendix 2: Available data opening packages

Region ID	Region name	u_meta project name	Data opening package name
cn-beijing	China (Beijing)	u_meta_beijing	systables
cn-chengdu	China (Chengdu)	u_meta_chengdu	systables
cn-hangzhou	China (Hangzhou)	u_meta_hangzhou	systables
cn-hongkong	China (Hong Kong)	u_meta_hongkong	systables
cn-shanghai	China (Shanghai)	u_meta_shanghai	systables
cn-shenzhen	China (Shenzhen)	u_meta_shenzhen	systables
cn-zhangjiakou	China (Zhangjiakou)	u_meta_zhangjiakou	systables
ap-northeast-1	Japan (Tokyo)	u_meta_tokyo	systables
ap-south-1	India (Mumbai)	u_meta_mumbai	systables
ap-southeast-1	Singapore (Singapore)	u_meta_singapore	systables
ap-southeast-2	Australia (Sydney)	u_meta_sydney	systables
ap-southeast-3	Malaysia (Kuala Lumpur)	u_meta_kualalumpur	systables
ap-southeast-5	Indonesia (Jakarta)	u_meta_jakarta	systables
eu-central-1	Germany (Frankfurt)	u_meta_frankfurt	systables
eu-west-1	UK (London)	u_meta_london	systables

Region ID	Region name	u_meta project name	Data opening package name
me-east-1	UAE (Dubai)	u_meta_dubai	systables
us-east-1	US (Virginia)	u_meta_virginia	systables
us-west-1	US (Silicon Valley)	u_meta_siliconvalley	systables

6.Resource Optimization

6.1. Overview

The Resource Optimization service scans data for optimizable tables and nodes in data storage, data computing, and data collection. You can optimize the tables and nodes accordingly to improve the efficiency of running nodes in DataWorks.

> **Note** Currently, the Resource Optimization service is in invitational preview. The supported region includes China (Shanghai) only. If you want to use this service, submit a ticket to apply for the permission.

The Resource Optimization service displays optimizable items in lists. You can optimize these items based on the interpretation provided for the corresponding optimizable items.

The Resource Optimization service provides optimization suggestions on personal resources and workspace resources. The description is as follows:

- **Personal Resource Optimization**: This page displays different data under the logged on personal account, including the **Total Tasks**, **Total Tables**, **Optimization Trends**, and **Personal Resource Optimization** sections.
- **Workspace Resource Optimization**: This page displays different data under the specified workspace for an administrator, including the **Total Tasks**, **Total Tables**, **Optimization Trends**, **Optimizable Computing Node Rankings**, **Optimizable Storage Rankings**, and **Workspace Resource Optimization** sections.

  Based on the information in the **Optimizable Computing Node Rankings** and **Optimizable Storage Rankings** sections, the administrator can inform the corresponding asset owners to optimize the assets.

> **Note** The Total Tasks and Total Tables statistics on the Personal Resource Optimization and Workspace Resource Optimization pages are not updated in real time. The statistics will be updated on the next day after the data is generated.

6.2. Analyze resources

DataWorks provides the resource analysis feature for data developers and administrators to view and analyze their own resources or all resources in a workspace. You can view and analyze the resource usage of tables and nodes and the status of nodes. This helps you optimize the overall resource usage.

Prerequisites

DataWorks Professional Edition or a more advanced edition is activated.

Usage notes

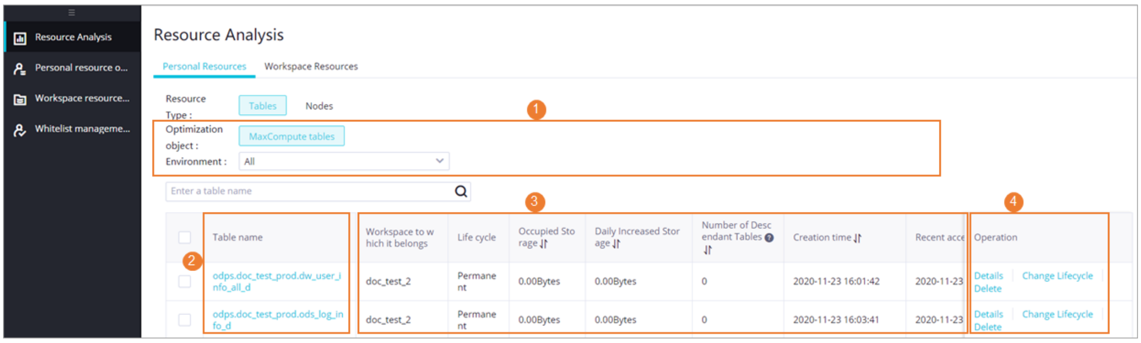
- Different accounts may have different roles or permissions. Therefore, the tables and nodes that you can view on the Resource Analysis page may vary with the account that you use to access the page. Only the administrator of a workspace can view the details of all resources in the workspace.
- You can view the resource usage of MaxCompute tables, MaxCompute nodes, and Data Integration nodes, and the status of MaxCompute nodes and Data Integration nodes.

Procedure

1. Log on to the DataWorks console.
2. In the left-side navigation pane, click **Workspaces**.
3. On the Workspaces page that appears, find the target workspace and click **Data Analytics** in the Actions column.
4. Click the icon in the upper-left corner and choose **All Products > Data governance > Resource Optimization**.

   The **Resource Optimization** page appears.
5. In the left-side navigation pane, click **Resource Analysis**. On the Resource Analysis page, you can view the resource usage of your own tables and nodes or all tables and nodes in the workspace. You can also check the status of your own nodes or all nodes in the workspace.You can choose to view personal resources or workspace resources based on your optimization requirements.
   - The **Personal Resources** tab displays the tables and nodes of the current account.
   - The **Workspace Resources** tab displays all tables and nodes in the current workspace. Only the administrator of a workspace can view the details of all resources in the workspace. The Workspace Resources tab is displayed only when you log on as a workspace administrator.

   The following section describes the perspectives from which tables and nodes are analyzed. You can view the details in the resource list.
   - Resource Type: **Tables**



Resource Type : Tables Nodes

Optimization object : MaxCompute tables

Environment : All

□	Table name	Workspace to which it belongs	Life cycle	Occupied Storage	Daily Increased Storage	Number of Descendant Tables	Creation time	Recent acce	Operation
□	odps.doc_test_prod.dw_user_info_all_d	doc_test_2	Permanent	0.00Bytes	0.00Bytes	0	2020-11-23 16:01:42	2020-11-23	Details | Change Lifecycle Delete
□	odps.doc_test_prod.ods_log_info_d	doc_test_2	Permanent	0.00Bytes	0.00Bytes	0	2020-11-23 16:03:41	2020-11-23	Details | Change Lifecycle Delete

Tables are analyzed from the following perspectives:

- **Occupied Storage**: the total amount of storage space that is occupied by the table.
- **Daily Increased Storage**: the amount of storage space that was increased on the day before the current date, compared with the amount of storage space that was occupied two days before the current date.
- **Number of Descendant Tables**: the number of descendant tables of the table.
- **Output Node**: the ID of the node that generates the table. This information indicates whether a node continuously generates data for the table.
  - If the Output Node column is empty, the table is not an output table of a DataWorks node and may be a temporary table or a dimension table that is seldom updated. Generally, you can manually maintain the table.
  - If the Output Node column has data, the table is an output table of a node. The table may be a table that requires regular updates.

You can plan the optimization operations to be performed on a table based on the analysis results that are displayed on the page and your business requirements. For example, if a table has a long lifecycle, occupies a large amount of storage space, does not have descendant tables, is not accessed from a long period of time, and does not have a node that generates data for it, you can check the details of the table. If the table is an unnecessary table, you can shorten its lifecycle or delete it.

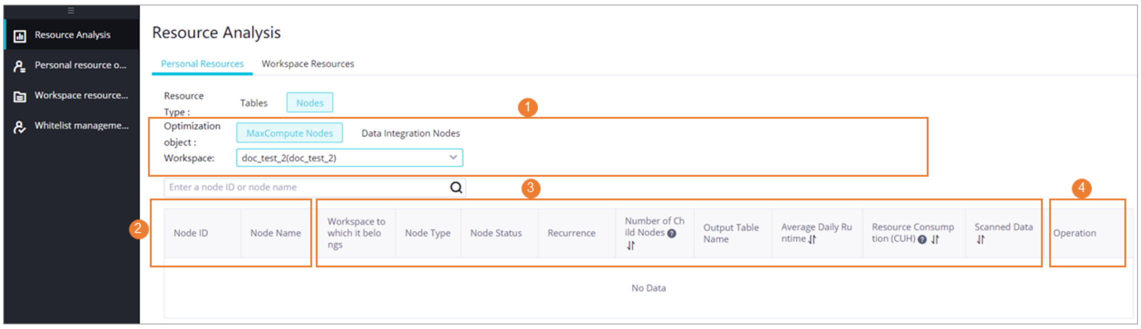
On the Resource Analysis page, you can perform the following optimization operations on a table: **Details**, **Change Lifecycle**, and **Delete**.

> ⓘ **Note** If you change the lifecycle of a table or delete a table, the operation immediately takes effect and you cannot undo the operation. Exercise caution when you perform these operations.

- Resource Type: **Nodes**

> ⓘ **Note** The resource analysis results of nodes show the status and resource usage of the nodes on the day before the current date.

You can view MaxCompute nodes and Data Integration nodes. In the following example, MaxCompute nodes are used.



Resource Type : Nodes
Optimization object : MaxCompute Nodes
Workspace: doc_test_2(doc_test_2)

Node ID	Node Name	Workspace to which it belongs	Node Type	Node Status	Recurrence	Number of Child Nodes	Output Table Name	Average Daily Runtime	Resource Consumption (CUH)	Scanned Data	Operation

No Data

Nodes are analyzed from the following perspectives:

- **Number of Child Nodes**: the number of child nodes of the node. This information is important and helps you determine the dependencies of the node. If the value is not 0, the node has child nodes. In this case, exercise caution when you optimize the node because the optimization may affect the child nodes.
- **Output Table Name**: If the node writes data to MaxCompute tables, the names of these tables are displayed in this column. If the Output Table Name column is empty, the node does not write data to MaxCompute tables.

You can plan the optimization operations to be performed on a node based on the analysis results that are displayed on the page and your business requirements. For example, if a node failed to run, does not have child nodes or output tables, and consumes a large amount of resources, you can check the details of the node. If the node is an unnecessary node, you can pause the node.

On the Resource Analysis page, you can perform the following optimization operations on a node: **Details** and **Pause node**.

> Note If you pause a node, the node instances that have been generated are not affected, whereas newly generated node instances are paused. After you pause a node, you may also need to optimize the output tables of the node.

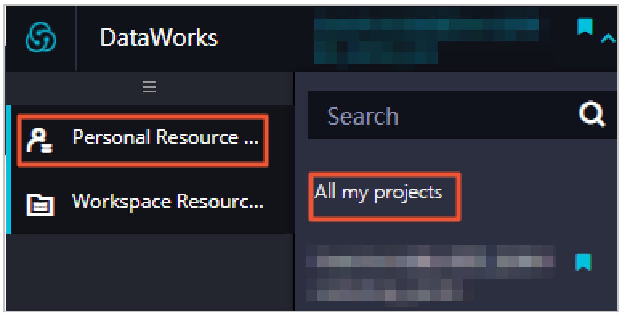
6.3. View optimizable personal resources

The Personal Resource Optimization page lists the optimizable nodes and tables under the logged on personal account.

Procedure

1. Log on to the DataWorks console.
2. In the left-side navigation pane, click **Workspaces**.
3. On the Workspaces page that appears, find the target workspace and click **Data Analytics** in the Actions column.
4. On the DataStudio page that appears, click ◎ in the upper-left corner and choose **All Products > Resource Optimization**. The **Personal Resource Optimization** page appears by default.

   To view the optimization information of other workspaces, select one from the top drop-down list. You can also click **All my projects**.

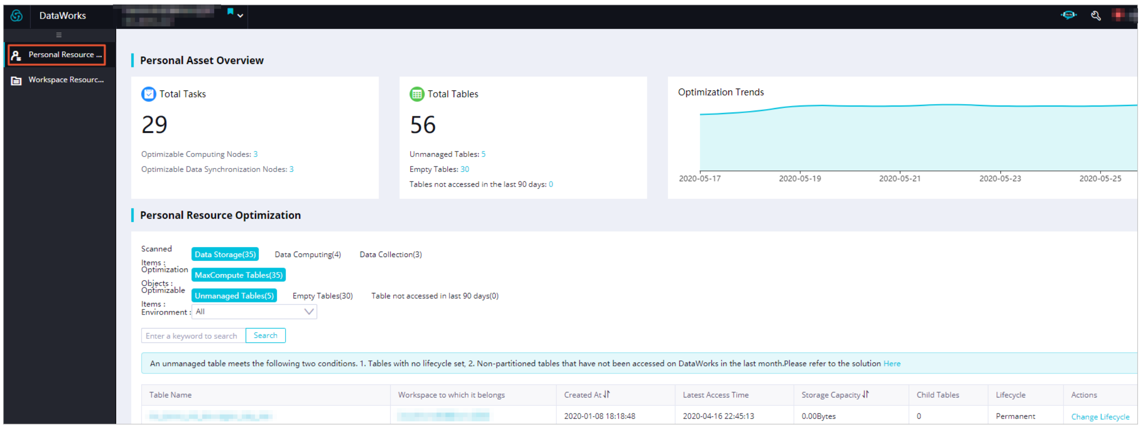
   The **Personal Resource Optimization** page consist of the **Personal Asset Overview** and **Personal Resource Optimization** sections.

   - You can view the total number of nodes and total number of tables under your personal account in the **Personal Asset Overview** section.

     The **Optimization Trends** chart displays the variation trend of the optimizable items in the last 10 days. You can view the number of optimizable items detected on different days.

     > Note The data in the Optimization Trends chart is not generated in real time. The date on the rightmost side of the chart indicates the latest date when the statistics are updated.

   - The Personal Resource Optimization section lists optimizable items of the **Data Storage**, **Data Computing**, and **Data Collection** types. They refer to MaxCompute tables, MaxCompute nodes, and data synchronization nodes that write data to MaxCompute, respectively.



Personal Asset Overview

Total Tasks

29

Optimizable Computing Nodes: 3

Optimizable Data Synchronization Nodes: 3

Total Tables

56

Unmanaged Tables: 5

Empty Tables: 30

Tables not accessed in the last 90 days: 0

Optimization Trends

Personal Resource Optimization

Scanned Items : Data Storage(35) Data Computing(4) Data Collection(3)

Optimization Objects : MaxCompute Tables(35)

Optimizable Items : Unmanaged Tables(5) Empty Tables(30) Table not accessed in last 90 days(0)

Environment : All

Enter a keyword to search Search

An unmanaged table meets the following two conditions. 1. Tables with no lifecycle set, 2. Non-partitioned tables that have not been accessed on DataWorks in the last month.Please refer to the solution Here

Table Name	Workspace to which it belongs	Created At	Latest Access Time	Storage Capacity	Child Tables	Lifecycle	Actions
[redacted]	[redacted]	2020-01-08 18:18:48	2020-04-16 22:45:13	0.00Bytes	0	Permanent	Change Lifecycle

Note The standard workspace mode provided by DataWorks isolates the development environment from the production environment. That is, a DataWorks workspace corresponds to a MaxCompute project in the development environment and a MaxCompute project in the production environment. You can filter projects by **Environment**.

Scanned item	Optimization object	Optimizable item	Description
Data Storage	**MaxCompute Tables**	**Unmanaged Tables**	An unmanaged table refers to a table that meets both of the following conditions: ■ The lifecycle of the table is not specified. ■ The table is a non-partitioned table that has not been accessed for the last 30 days in DataWorks. Tables that meet the preceding conditions are recognized as unmanaged tables. You can set a lifecycle for each unmanaged table to complete the optimization. For more information about the lifecycle, see Lifecycle. **Note** When the lifecycle of a table expires, data in the table will become invalid. We recommend that you exercise caution when performing this operation.
		Empty Tables	An empty table refers to a table with no data. We recommend that you do not delete empty tables directly. You can audit tables that were created a long time ago based on the table creation time to determine whether to delete the tables.

Scanned item	Optimization object	Optimizable item	Description
Data Computing	**MaxCompute Nodes**	**Conflict Task**	When you write the data of multiple nodes to the same table, unexpected results may be returned. We recommend that you do not write the data of multiple nodes to the same partition of the same table in the data development process. This helps avoid data quality issues caused by any failed node when you perform retroactive executions. The idempotence of data must be taken into account. We recommend that you pause one of the ancestor nodes and adjust the dependencies of its descendant nodes based on the number of descendant nodes of each ancestor node.
		Data Tilt	Data skew occurs on nodes where some node instances process more data and take much more time than the others. This prolongs the overall execution time of the nodes, leading to latency. For more information about the solution to data skew, see Long-tail computing optimization.
Data Collection	**Data Synchronization Nodes**	**Empty Import**	The volume of the data imported by certain data synchronization nodes is always 0. We recommend that you pause these nodes or bring them offline.
		Consistent Import	Certain data synchronization nodes have imported the same volume of data for 15 consecutive days. Check whether the source data is no longer updated. Also, check whether the nodes have been paused. We recommend that you stop scheduling any computing and storage resources for the paused nodes.
		Same Origin Import	Certain data synchronization nodes share the same data stores and import duplicate data to MaxCompute. This leads to waste of storage and scheduling resources. You can resolve this issue by merging nodes.

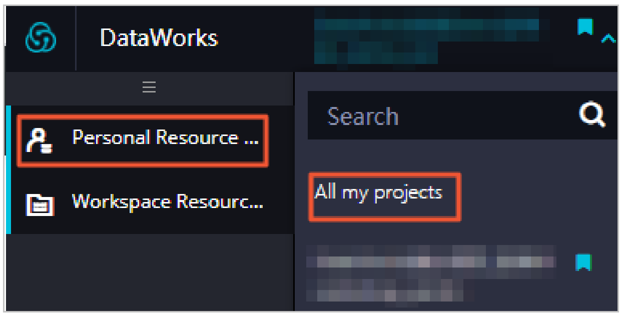
Scanned item	Optimization object	Optimizable item	Description
		OSS Synchronization Optimization	Certain data synchronization nodes transmit data to Object Storage Service (OSS) over the Internet. This consumes Internet traffic and charges you an additional fee. We recommend that you change the endpoint of the data store to an internal IP address. To perform this operation, click the DataWorks icon in the DataWorks console, click **Data Integration** in the left-side navigation pane, and then click **Data Sources**. This cuts the consumption of Internet traffic and improves the data transmission speed. For more information, see Configure endpoints.

6.4. View optimizable workspace resources

If you log on to the DataWorks console as an administrator, you can view the optimizable items of your workspace on the Workspace Resource Optimization page. The page also provides the Optimizable Computing Node Rankings and Optimizable Storage Rankings sections.

Procedure

1. Log on to the DataWorks console.
2. In the left-side navigation pane, click **Workspaces**.
3. On the Workspaces page that appears, find the target workspace and click **Data Analytics** in the Actions column.
4. On the DataStudio page that appears, click in the upper-left corner and choose **All Products > Resource Optimization**.
5. On the Resource Optimization page that appears, click **Workspace Resource Optimization** in the left-side navigation pane. To view the optimization information of other workspaces, select one from the top drop-down list. You can also click **All my projects**.



The **Workspace Resource Optimization** page consists of the **Workspace Asset Overview** and **Workspace Resource Optimization** sections.

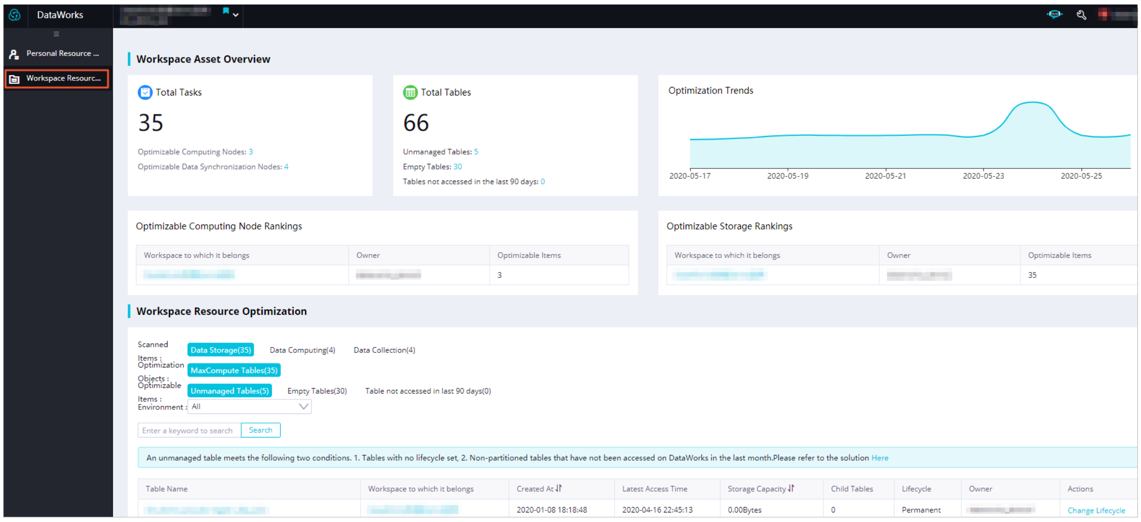
- You can view the total number of nodes and total number of tables in the workspace in the **Workspace Asset Overview** section.

  The **Optimization Trends** chart displays the variation trend of the optimizable items in the last 10 days. You can view the number of optimizable items detected on different days.

  > **Note** The data in the Optimization Trends chart is not generated in real time. The date on the rightmost side of the chart indicates the latest date when the statistics are updated.

  A maximum of 10 items can appear in the **Optimizable Computing Node Rankings** and **Optimizable Storage Rankings** sections, respectively.

- The information provided on the **Workspace Resource Optimization** and **Personal Resource Optimization** pages is basically the same. The difference is that only the administrator can view the owners of optimizable items on the Workspace Resource Optimization page.



Workspace Asset Overview

Total Tasks
35
Optimizable Computing Nodes: 3
Optimizable Data Synchronization Nodes: 4

Total Tables
66
Unmanaged Tables: 5
Empty Tables: 30
Tables not accessed in the last 90 days: 0

Optimizable Computing Node Rankings

Workspace to which it belongs	Owner	Optimizable Items
[redacted]	[redacted]	3

Optimizable Storage Rankings

Workspace to which it belongs	Owner	Optimizable Items
[redacted]	[redacted]	35

Workspace Resource Optimization

Scanned Items : Data Storage(35) Data Computing(4) Data Collection(4)
Optimization Objects : MaxCompute Tables(35)
Optimizable Items : Unmanaged Tables(5) Empty Tables(30) Table not accessed in last 90 days(0)
Environment : All

Enter a keyword to search Search

An unmanaged table meets the following two conditions. 1. Tables with no lifecycle set, 2. Non-partitioned tables that have not been accessed on DataWorks in the last month.Please refer to the solution Here

Table Name	Workspace to which it belongs	Created At	Latest Access Time	Storage Capacity	Child Tables	Lifecycle	Owner	Actions
[redacted]	[redacted]	2020-01-08 18:18:48	2020-04-16 22:45:13	0.00Bytes	0	Permanent	[redacted]	Change Lifecycle

Scanned item	Optimization object	Optimizable item	Description

Scanned item	Optimization object	Optimizable item	Description
Data Storage	**MaxCompute Tables**	**Unmanaged Tables**	An unmanaged table refers to a table that meets both of the following conditions: - The lifecycle of the table is not specified. - The table is a non-partitioned table that has not been accessed for the last 30 days in DataWorks. Tables that meet the preceding conditions are recognized as unmanaged tables. You can set a lifecycle for each unmanaged table to complete the optimization. For more information about the lifecycle, see Lifecycle. **Note** When the lifecycle of a table expires, data in the table will become invalid. We recommend that you exercise caution when performing this operation.
		Empty Tables	An empty table refers to a table with no data. We recommend that you do not delete empty tables directly. You can audit tables that were created a long time ago based on the table creation time to determine whether to delete the tables.
Data Computing	**MaxCompute Nodes**	**Conflict Task**	When you write the data of multiple nodes to the same table, unexpected results may be returned. We recommend that you do not write the data of multiple nodes to the same partition of the same table in the data development process. This helps avoid data quality issues caused by any failed node when you perform retroactive executions. The idempotence of data must be taken into account. We recommend that you pause one of the ancestor nodes and adjust the dependencies of its descendant nodes based on the number of descendant nodes of each ancestor node.

Scanned item	Optimization object	Optimizable item	Description
		Data Tilt	Data skew occurs on nodes where some node instances process more data and take much more time than the others. This prolongs the overall execution time of the nodes, leading to latency. For more information about the solution to data skew, see Long-tail computing optimization.
Data Collection	**Data Synchronization Nodes**	**Empty Import**	The volume of the data imported by certain data synchronization nodes is always 0. We recommend that you pause these nodes or bring them offline.
		Consistent Import	Certain data synchronization nodes have imported the same volume of data for 15 consecutive days. Check whether the source data is no longer updated. Also, check whether the nodes have been paused. We recommend that you stop scheduling any computing and storage resources for the paused nodes.
		Same Origin Import	Certain data synchronization nodes share the same data stores and import duplicate data to MaxCompute. This leads to waste of storage and scheduling resources. You can resolve this issue by merging nodes.
		OSS Synchronization Optimization	Certain data synchronization nodes transmit data to Object Storage Service (OSS) over the Internet. This consumes Internet traffic and charges you an additional fee. We recommend that you change the endpoint of the data store to an internal IP address. To perform this operation, click the DataWorks icon in the DataWorks console, click **Data Integration** in the left-side navigation pane, and then click **Data Sources**. This cuts the consumption of Internet traffic and improves the data transmission speed. For more information, see Configure endpoints.

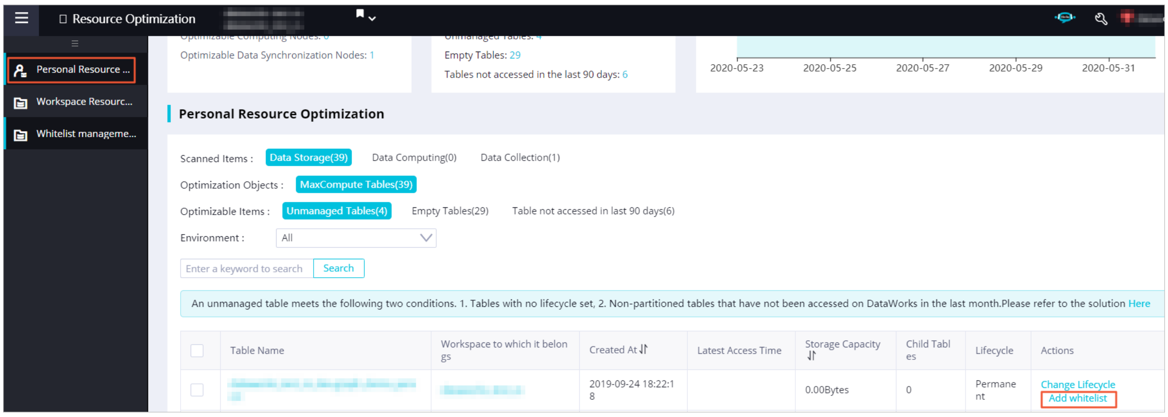
6.5. Manage whitelists

If you need to optimize resources, you can add the governance items with high management costs and low response speed to a whitelist. This topic describes how to add, view, revoke, and close a whitelist.

Add a whitelist

On the **Personal Resource Optimization** page, you can process the governance items under your account. For the business items that are difficult to govern, you can add a whitelist.

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 where your workspace resides, find the workspace, and then click **Data Analytics** in the Actions column.
2. On the DataStudio page, click ◙ in the upper-left corner and choose **All Products** > **Resource Optimization**. The **Personal Resource Optimization** page appears.For more information about the **Personal Resource Optimization** page, see View optimizable personal resources.
3. Find a table or node and click **Add whitelist** in the Actions column.



☰ Resource Optimization [redacted]

Personal Resource ...

Workspace Resourc...

Whitelist manageme...

Optimizable Data Synchronization Nodes: 1

Empty Tables: 29

Tables not accessed in the last 90 days: 6

Personal Resource Optimization

Scanned Items : Data Storage(39) Data Computing(0) Data Collection(1)

Optimization Objects : MaxCompute Tables(39)

Optimizable Items : Unmanaged Tables(4) Empty Tables(29) Table not accessed in last 90 days(6)

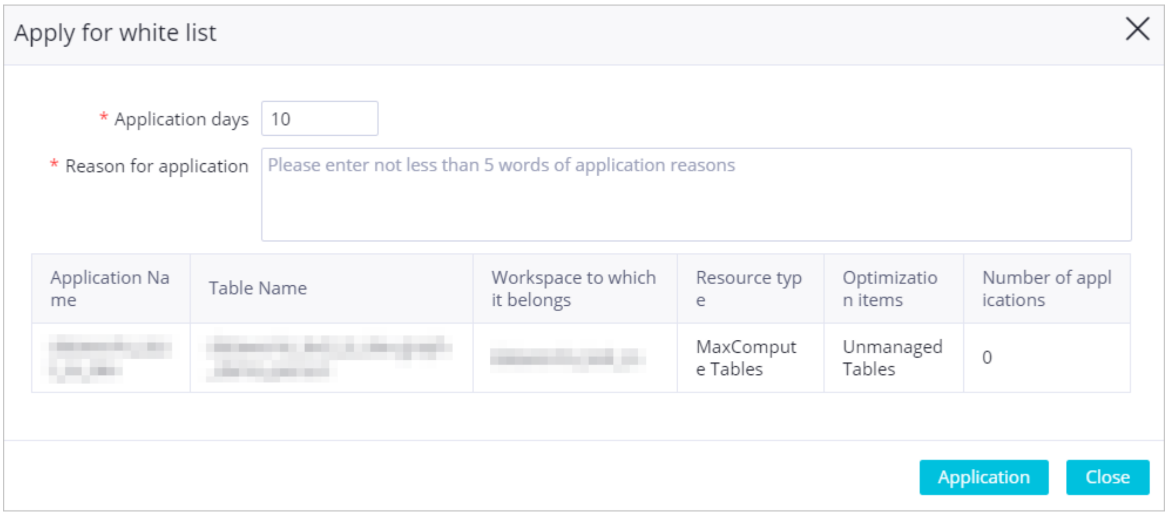
Environment : All

Enter a keyword to search Search

An unmanaged table meets the following two conditions. 1. Tables with no lifecycle set, 2. Non-partitioned tables that have not been accessed on DataWorks in the last month.Please refer to the solution Here

□	Table Name	Workspace to which it belongs	Created At	Latest Access Time	Storage Capacity	Child Tables	Lifecycle	Actions
□	[redacted]	[redacted]	2019-09-24 18:22:18		0.00Bytes	0	Permanent	Change Lifecycle Add whitelist

4. In the **Apply for white list** dialog box, set **Application days** and **Reason for application**.



Apply for white list

* Application days: 10

* Reason for application: Please enter not less than 5 words of application reasons

Application Name	Table Name	Workspace to which it belongs	Resource type	Optimization items	Number of applications
[redacted]	[redacted]	[redacted]	MaxCompute Tables	Unmanaged Tables	0

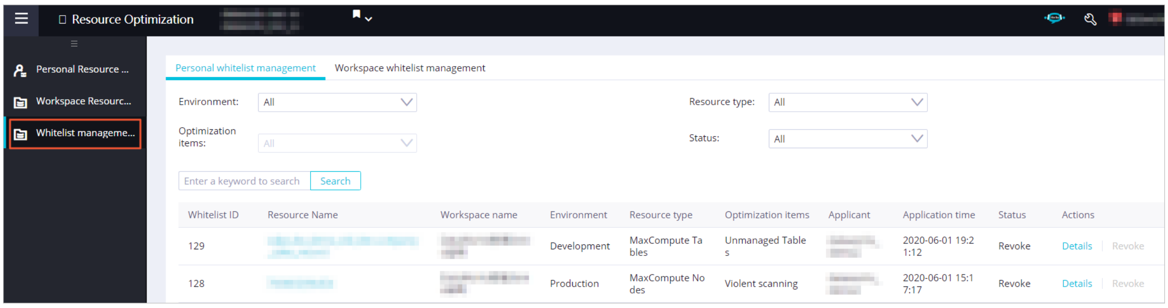
> **Note**
> - The reason for applying for a whitelist must be at least five characters.
> - Unless under special circumstances, do not repeatedly add a governance item to a whitelist.

5. Click **Application**.

View a whitelist

After you add a governance item to a whitelist, you can choose **Whitelist management** > **Personal whitelist management** and view the whitelist.

1. In the left-side navigation pane, click **Whitelist management**. The **Personal whitelist management** tab appears.The **Personal whitelist management** tab displays the whitelists that you have applied for. After you apply for a whitelist, the whitelist application is automatically approved.
2. Find a whitelist and click **Details** in the Actions column.



Environment: All

Resource type: All

Optimization items: All

Status: All

Enter a keyword to search

Whitelist ID	Resource Name	Workspace name	Environment	Resource type	Optimization items	Applicant	Application time	Status	Actions
129	[redacted]	[redacted]	Development	MaxCompute Tables	Unmanaged Tables	[redacted]	2020-06-01 19:21:12	Revoke	Details | Revoke
128	[redacted]	[redacted]	Production	MaxCompute Nodes	Violent scanning	[redacted]	2020-06-01 15:17:17	Revoke	Details | Revoke

3. In the **Whitelist details** dialog box, view the details of the whitelist.

Revoke a whitelist

You can choose **Whitelist management** > **Personal whitelist management** and revoke a whitelist.

1. In the left-side navigation pane, click **Whitelist management**. The **Personal whitelist management** tab appears.
2. Find a whitelist and click **Revoke** in the Actions column.
3. In the **Revoke whitelist** dialog box, enter the reason for revoking the whitelist.

> Note
>
> - The reason for revoking a whitelist must be at least five characters.
> - After the whitelist is revoked, the governance item is restored. We recommend that you govern the item in a timely manner.

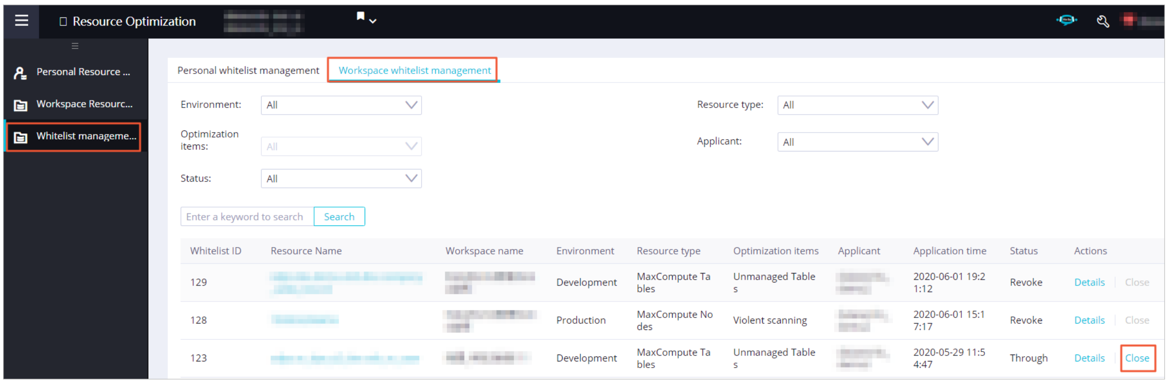
4. Click **Revoke**.

Close a whitelist

As a developer, you can add, view, and revoke whitelists under your account. The workspace administrator can manage whitelists in the current workspace on the **Workspace whitelist management** tab.

The workspace administrator can view the details about the whitelist that a developer adds and close the whitelist.

1. In the left-side navigation pane, click **Whitelist management**.
2. Click **Workspace whitelist management**.
3. Find a whitelist and click **Close** in the Actions column.



Environment: All

Optimization items: All

Status: All

Resource type: All

Applicant: All

Enter a keyword to search

Whitelist ID	Resource Name	Workspace name	Environment	Resource type	Optimization items	Applicant	Application time	Status	Actions
129	[redacted]	[redacted]	Development	MaxCompute Tables	Unmanaged Tables	[redacted]	2020-06-01 19:21:12	Revoke	Details | Close
128	[redacted]	[redacted]	Production	MaxCompute Nodes	Violent scanning	[redacted]	2020-06-01 15:17:17	Revoke	Details | Close
123	[redacted]	[redacted]	Development	MaxCompute Tables	Unmanaged Tables	[redacted]	2020-05-29 11:54:47	Through	Details Close

4. In the **Close whitelist** dialog box, enter the reason for closing the whitelist.

> Note The reason for closing a whitelist must be at least five characters.

5. Click **Confirm**.

7.Use ActionTrail to query behavior events

DataWorks is integrated with ActionTrail. This allows you to query ActionTrail for DataWorks behavior events of your Alibaba Cloud account over the last 90 days. You can use ActrionTrail to deliver the events to a Logstore in Log Service or a specific Object Storage Service (OSS) bucket for monitoring and alerting. This meets the requirements for timely auditing, problem backtracking, and problem analysis. This topic describes how to query DataWorks behavior events in ActionTrail.

Context

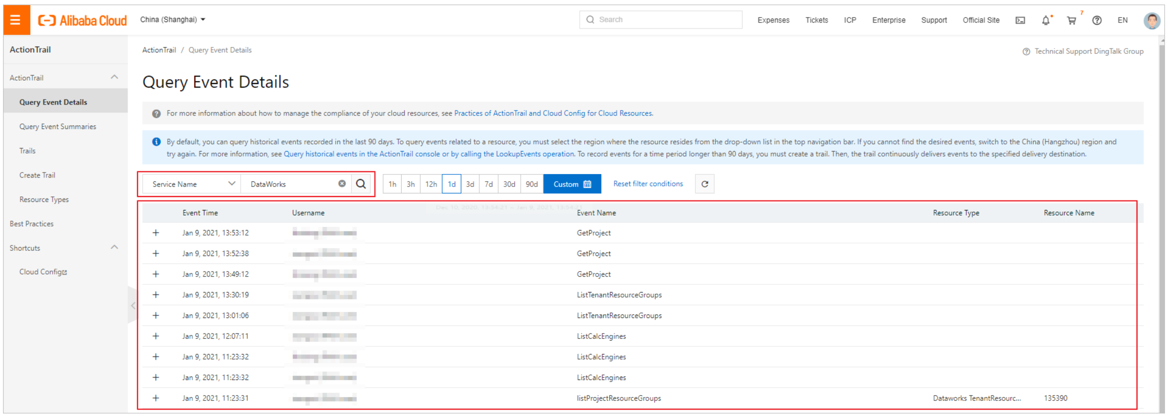
Alibaba Cloud ActionTrail is a service that monitors and records the actions of your Alibaba Cloud account. The actions include the access to and use of cloud products and services through the Alibaba Cloud Management Console, API operations, and SDKs. ActionTrail records these actions as events. You can download these events from the ActionTrail console or configure ActionTrail to deliver these events to Log Service Logstores or OSS buckets. Then, you can perform behavior analysis, security analysis, resource change tracking, and compliance auditing based on the events. For more information, see What is ActionTrail?

Precautions

- After you perform an operation in DataWorks, ActionTrail records the operation in 5 minutes to 10 minutes.
- You can configure tracking alerts for important events to detect and handle anomalous activities in a timely manner.

Query DataWorks behavior events

1. Log on to the ActionTrail console.
2. In the left-side navigation pane, click **Event Detail Query**. Then, select a region in the top navigation bar.
3. On the **Event Detail Query** page, select **Service Name** from the drop-down list and enter **DataWorks** in the search box to query DataWorks events that are recorded.



For more information about how to manage the compliance of your cloud resources, see Practices of ActionTrail and Cloud Config for Cloud Resources.

By default, you can query historical events recorded in the last 90 days. To query events related to a resource, you must select the region where the resource resides from the drop-down list in the top navigation bar. If you cannot find the desired events, switch to the China (Hangzhou) region and try again. For more information, see Query historical events in the ActionTrail console or by calling the LookupEvents operation. To record events for a time period longer than 90 days, you must create a trail. Then, the trail continuously delivers events to the specified delivery destination.

	Event Time	Username	Event Name	Resource Type	Resource Name
+	Jan 9, 2021, 13:53:12	[redacted]	GetProject		
+	Jan 9, 2021, 13:52:38	[redacted]	GetProject		
+	Jan 9, 2021, 13:49:12	[redacted]	GetProject		
+	Jan 9, 2021, 13:30:19	[redacted]	ListTenantResourceGroups		
+	Jan 9, 2021, 13:01:06	[redacted]	ListTenantResourceGroups		
+	Jan 9, 2021, 12:07:11	[redacted]	ListCalcEngines		
+	Jan 9, 2021, 11:23:32	[redacted]	ListCalcEngines		
+	Jan 9, 2021, 11:23:32	[redacted]	ListCalcEngines		
+	Jan 9, 2021, 11:23:31	[redacted]	listProjectResourceGroups	Dataworks TenantResourc...	135390

The query results contain the following information: **Event Time**, **Username**, **Event Name**, **Resource Type**, and **Resource Name**.You can use **Event Name** to determine whether an event is recorded for an API call and query the event meaning.

Note An API operation can be called by using a codeless user interface (UI) or code editor.

- The event is recorded for an API call.

  The **event name** is consistent with the API operation name. You can use the **event name** to query the event meaning from the DataWorks API operation list.

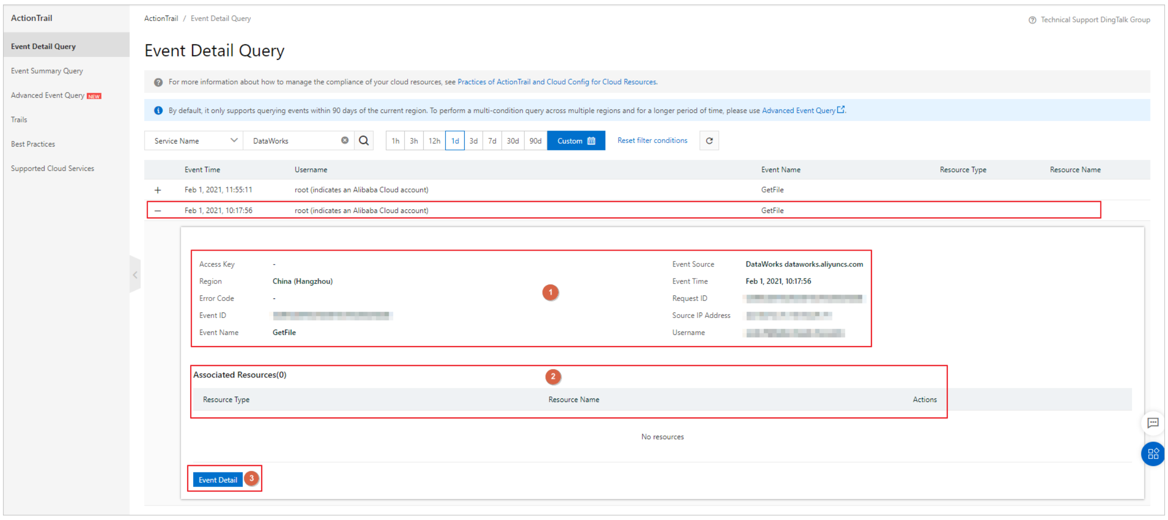
- The event is not recorded for an API call.

  You can query the event meaning from the following table.

Event name	Description	Service module
DownloadExecutionResultDataStudio	Downloads query results.	**DataStudio**
CreateBusiness	Creates a workflow.	
DestroyRelationTableFromBusiness	Deletes all tables from a workflow.	
DeleteBusiness	Deletes a workflow.	
ExecuteFile	Runs a file as a temporary task.	
LockFile	Locks a file.	
UnlockFile	Steals the lock of a file.	
RecoverFile	Recovers files in the recycle bin.	
CloneFile	Clones a file.	
DeleteFolder	Deletes a folder.	
DeleteDeployment	Deletes a deployment task.	
ListCodingProjects	Queries code-based projects.	**AppStudio**

Note If the meaning of an event cannot be obtained by using one of the preceding methods, submit a ticket to query the details of the event.

4. Expand an event and click **Event Detail** to view the details of the event.



ActionTrail

- Event Detail Query
- Event Summary Query
- Advanced Event Query NEW
- Trails
- Best Practices
- Supported Cloud Services

ActionTrail / Event Detail Query

Technical Support DingTalk Group

Event Detail Query

For more information about how to manage the compliance of your cloud resources, see Practices of ActionTrail and Cloud Config for Cloud Resources.

By default, it only supports querying events within 90 days of the current region. To perform a multi-condition query across multiple regions and for a longer period of time, please use Advanced Event Query.

Service Name: DataWorks | 1h | 3h | 12h | 1d | 3d | 7d | 30d | 90d | Custom | Reset filter conditions

	Event Time	Username	Event Name	Resource Type	Resource Name
+	Feb 1, 2021, 11:55:11	root (indicates an Alibaba Cloud account)	GetFile		
–	Feb 1, 2021, 10:17:56	root (indicates an Alibaba Cloud account)	GetFile		

1

Access Key	-	Event Source	DataWorks dataworks.aliyuncs.com
Region	China (Hangzhou)	Event Time	Feb 1, 2021, 10:17:56
Error Code	-	Request ID	[redacted]
Event ID	[redacted]	Source IP Address	[redacted]
Event Name	GetFile	Username	[redacted]

2

Associated Resources(0)

Resource Type	Resource Name	Actions

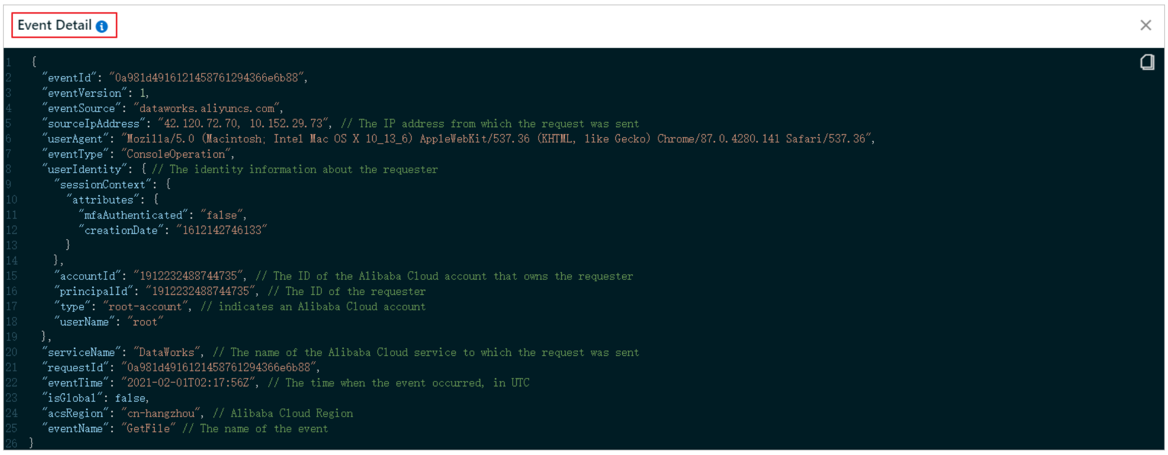
No resources

Event Detail 3

The following table describes the event details.

No.	Description
1	The details of the event. Move the pointer over the username and click **detail** to go to the **RAM console**. Then, you can view the details of the user.
2	The **resource type**, **resource name**, and **operation** involved in the event.
3	You can click **Event Detail** to view the code record of the event.

The following figure shows the code record of the listProjectResourceGroups event.



Event Detail

```
{
  "eventId": "0a981d4916121458761294366e6b88",
  "eventVersion": 1,
  "eventSource": "dataworks.aliyuncs.com",
  "sourceIpAddress": "42.120.72.70, 10.152.29.73", // The IP address from which the request was sent
  "userAgent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/87.0.4280.141 Safari/537.36",
  "eventType": "ConsoleOperation",
  "userIdentity": { // The identity information about the requester
    "sessionContext": {
      "attributes": {
        "mfaAuthenticated": "false",
        "creationDate": "1612142746133"
      }
    },
    "accountId": "1912232488744735", // The ID of the Alibaba Cloud account that owns the requester
    "principalId": "1912232488744735", // The ID of the requester
    "type": "root-account", // indicates an Alibaba Cloud account
    "userName": "root"
  },
  "serviceName": "DataWorks", // The name of the Alibaba Cloud service to which the request was sent
  "requestId": "0a981d4916121458761294366e6b88",
  "eventTime": "2021-02-01T02:17:56Z", // The time when the event occurred, in UTC
  "isGlobal": false,
  "acsRegion": "cn-hangzhou", // Alibaba Cloud Region
  "eventName": "GetFile" // The name of the event
}
```

In the **Event Detail** dialog box, click the icon in the upper-right corner to copy the code record.

What's next

You can use the queried event details to perform behavior analysis, security analysis, resource change tracking, and compliance auditing.