ALIBABA CLOUD

Alibaba Cloud

DataWorks

FAQ

Document Version: 20220712

Alibaba Cloud

Legal disclaimer

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

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

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

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

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

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

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

Document conventions

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

Table of Contents

1.FAQ about Data Integration tasks and Data Integration resource groups

1.1. Overview

Network connectivity and operations on resource groups

- Network connectivity for data synchronization
  - Which information about DataWorks and its network capabilities do I need to take note of before I configure a data synchronization node?
  - How do I make sure the network connectivity between a resource group in DataWorks and a self-managed data source that is hosted on an Elastic Compute Service (ECS) instance when I synchronize data from the self-managed data source?
  - How do I make sure the network connectivity between a resource group in DataWorks and a data source that is deployed in a different region from the resource group when I synchronize data from the data source?
  - When I synchronize data from a data source, the account that I use to access the data source is different from the account that I use to access DataWorks. How do I make sure the network connectivity between DataWorks and the data source?
  - What do I do if the network connectivity test for a data source in a VPC fails?
- Operations on resource groups
  - I cannot find the exclusive resource group for Data Integration that I purchased when I test network connectivity for a data source or run a data synchronization node. What do I do?
  - How can I determine the type of the resource group on which a data synchronization node is run from a log?
  - How do I configure a resource group to wait for gateway resources?

Real-time synchronization

- Precautions for configuring real-time synchronization nodes
  - What types of data sources support real-time synchronization?
  - Why is the Internet not recommended for real-time synchronization?
  - What operation does DataWorks perform on the data records that are synchronized in real time?
  - How do I deal with the TRUNCATE statement during real-time data synchronization?
  - How do I improve the speed and performance of real-time synchronization?
  - Can I directly run a real-time synchronization node on the codeless user interface (UI)?
- Kafka data synchronization in real time

  When I run a node to synchronize data from Kafka in real time, the following error message appears: Startup mode for the consumer set to timestampOffset, but no begin timestamp was specified.. What do I do?
- MySQL data synchronization in real time

  - When I run a node to synchronize data from MySQL in real time, the following error message appears: Cannot replicate because the master purged required binary logs.. What do I do?
  - When I run a node to synchronize data from MySQL, the following error message appears: MysqlBinlogReaderException. What do I do?
  - When I run a node to synchronize data from MySQL, the following error message appears: show master status' has an error!. What do I do?
  - When I run a node to synchronize data from MySQL in real time, the following error message appears: parse.exception.PositionNotFoundException: can't find start position forxxx. What do I do?
  - When I run a node to synchronize data from MySQL in real time, data can be read at the beginning but cannot be read after a period of time. What do I do?
- Hologres data synchronization in real time

  When I run a node to synchronize data from Hologres in real time, the following error message appears: permission denied for database xxx. What do I do?

Batch synchronization

- O&M of batch synchronization nodes
  - Why is the connectivity test of a data source successful, but the corresponding batch synchronization node fails to be run?
  - How do I change the resource group that is used to run a data synchronization node in Data Integration?
  - How do I locate and handle dirty data?
- Common plug-in errors
  - How do I handle a dirty data error that is caused by encoding format configuration issues or garbled characters?
  - What do I do if the error message [TASK_MAX_SLOT_EXCEED]:Unable to find a gateway that meets resource requirements. 20 slots are requested, but the maximum is 16 slots. is returned?
  - What do I do if a server-side request forgery (SSRF) attack is detected in a node?
  - What do I do if the error message OutOfMemoryError: Java heap space is returned when I run a batch synchronization node?
  - What do I do if the same batch synchronization node fails to be run occasionally?
  - Batch synchronization
  - What do I do if the error message Duplicate entry 'xxx' for key 'uk_uk_op' is returned when I run a batch synchronization node?
  - What do I do if the error message plugin xx does not specify column is returned when I run a batch synchronization node?
- Specific plug-in errors
  - What do I do if an error occurs when I add a MongoDB data source as the root user?
  - The authDB database used by MongDB is the admin database. How do I synchronize data from business databases?
  - How do I convert the values of the variables in the query parameter into values in the timestamp format when I synchronize incremental data from a table of a MongDB database?

  - What do I do if the error message AccessDenied The bucket you access does not belong to you. is returned when I read data from an OSS bucket?
  - Is an upper limit configured for the number of OSS objects that can be read?
  - What do I do if the error message Code:[RedisWriter-04], Description:[Dirty data]. - source column number is in valid! is returned when I write data to Redis in hash mode?
  - What do I do if the following error message is returned when I read data from or write data to ApsaraDB RDS for MySQL: Application was streaming results when the connection failed. Consider raising value of 'net_write_timeout/net_read_timeout,' on the server.?
  - What do I do if the error message The last packet successfully received from the server was 902,138 milliseconds ago is returned when I read data from MySQL?
  - What do I do if an error occurs when I read data from PostgreSQL?
  - What do I do if the error message Communications link failure is returned?
  - What do I do if the error message The download session is expired. is returned when I read data from a MaxCompute table?
  - What do I do if the error message Error writing request body to server is returned when I write data to a MaxCompute table?
  - What do I do if data fails to be written to DataHub because the amount of data that I want to write to DataHub at a time exceeds the upper limit?
- Batch synchronization
  - Batch synchronization
  - How do I customize table names in a batch synchronization node?
  - What do I do if the table that I want to select does not appear in the Table drop-down list in the Source section when I configure a batch synchronization node?
  - What are the items that I must take note of when I use the Add feature in a synchronization node that reads data from the MaxCompute table?
  - How do I read data in partition key columns from a MaxCompute table?
  - How do I synchronize data from multiple partitions of a MaxCompute table?
  - What do I do if a synchronization node fails to be run because the name of a column in the source table is a keyword?
  - Why is no data obtained when I read data from a LogHub table whose columns contain data?
  - Why is some data missing when I read data from a LogHub data source?
  - What do I do if the fields that I read based on the field mapping configuration in LogHub are not the expected fields?
  - I configured the endDateTime parameter to specify the end time for reading from a Kafka data source, but some data that is returned is generated at a time point later than the specified end time. What do I do?
  - How do I remove the random strings that appear in the data I write to OSS?
  - How does the system synchronize data from a MySQL data source on which sharding is performed to a MaxCompute table?

1.2. Network connectivity and operations on resource groups

- Network connectivity for data synchronization
  - Which information about DataWorks and its network capabilities do I need to take note of before I configure a data synchronization node?
  - How do I make sure the network connectivity between a resource group in DataWorks and a self-managed data source that is hosted on an Elastic Compute Service (ECS) instance when I synchronize data from the self-managed data source?
  - How do I make sure the network connectivity between a resource group in DataWorks and a data source that is deployed in a different region from the resource group when I synchronize data from the data source?
  - When I synchronize data from a data source, the account that I use to access the data source is different from the account that I use to access DataWorks. How do I make sure the network connectivity between DataWorks and the data source?
  - What do I do if the network connectivity test for a data source in a VPC fails?
- Operations on resource groups
  - I cannot find the exclusive resource group for Data Integration that I purchased when I test network connectivity for a data source or run a data synchronization node. What do I do?
  - How can I determine the type of the resource group on which a data synchronization node is run from a log?
  - How do I configure a resource group to wait for gateway resources?

Which information about DataWorks and its network capabilities do I need to take note of before I configure a data synchronization node?

Before you configure a data synchronization node, take note of the following items:

- The virtual private cloud (VPC), vSwitch, and region that are used for the data source from which you want to synchronize data, and the region in which your DataWorks workspace resides.
- Whether the data source and your DataWorks workspace are deployed in different regions under different accounts.

For more information about how to troubleshoot issues that occur when you configure and run a synchronization node, see Supported data source types, readers, and writers.

If you encounter issues when you test the network connectivity of a data source, we recommend that you troubleshoot the issue by referring to Select a network connectivity solution.

If you use an exclusive resource group for Data Integration, we recommend that you perform the following operations before you configure a data synchronization node: purchase an exclusive resource group for Data Integration, associate the exclusive resource group for Data Integration with the VPC in which the data source resides, evaluate whether you need to add a route, configure a whitelist for the data source, and associate the exclusive resource group for Data Integration with a DataWorks workspace. For more information, see Create and use an exclusive resource group for Data Integration.

How do I make sure the network connectivity between a resource group in DataWorks and a self-managed data source that is hosted on an Elastic Compute Service (ECS) instance when I synchronize data from the self-managed data source?

If you want to use an exclusive resource group for Data Integration to access a self-managed data source that is hosted on an ECS instance over an internal network, you must configure network settings for the exclusive resource group for Data Integration. For more information, see Create and use an exclusive resource group for Data Integration. Take note of the following items:

- If you associate the exclusive resource group for Data Integration with a VPC in which the ECS instance resides, a route that points to the CIDR block of the VPC is automatically added. We recommend that you do not delete the added route. If you delete the added route, you may fail to access other data sources and an error may be reported during data synchronization.
- You must add the CIDR block of the vSwitch to which the exclusive resource group for Data Integration is bound to the whitelist of the data source. For more information, see Add the EIP or CIDR block of an exclusive resource group for Data Integration to the whitelist of a data source.

How do I make sure the network connectivity between a resource group in DataWorks and a data source that is deployed in a different region from the resource group when I synchronize data from the data source?

Before you configure and run a data synchronization node, we recommend that you use a network connectivity solution. For more information, see Select a network connectivity solution. Take note of the following items:

- If you want to synchronize data from a data source that is deployed in a region different from your DataWorks resource group over the Internet, you must add the elastic IP address (EIP) and CIDR block of the exclusive resource group for Data Integration to the whitelist of the data source. For more information, see Add the EIP or CIDR block of an exclusive resource group for Data Integration to the whitelist of a data source.

  > **Note** You are charged for the traffic generated over the Internet. For more information, see Billing of Internet traffic.

- If you want to synchronize data from a data source over an internal network and the account that is used to access the data source is different from the account that you use to access DataWorks, you must perform the following operations:

  i. Use a network connectivity service of Alibaba Cloud to establish a connection between the networks of the two Alibaba Cloud accounts. You can use a network connectivity service such as VPN Gateway or Express Connect.

  ii. Associate an exclusive resource group for Data Integration with the VPC that is connected to the network of the region where the data source is deployed.

  iii. Add a custom route to your data center and add the IP address of the destination data source to your data center.

  iv. Add the CIDR block of the vSwitch to which the exclusive resource group for Data Integration is bound to the whitelist of the data source. For more information, see Add the EIP or CIDR block of an exclusive resource group for Data Integration to the whitelist of a data source.

When I synchronize data from a data source, the account that I use to access the data source is different from the account that I use to access DataWorks. How do I make sure the network connectivity between DataWorks and the data source?

Before you configure and run a data synchronization node, we recommend that you use a network connectivity solution for troubleshooting. For more information, see Select a network connectivity solution.

- If you want to synchronize data over the Internet, you must add the EIP and CIDR block of the exclusive resource group for Data Integration to the whitelist of the data source. For more information, see Add the EIP or CIDR block of an exclusive resource group for Data Integration to the whitelist of a data source.

  Note You are charged for the traffic generated over the Internet. For more information, see Billing of Internet traffic.

- If you want to synchronize data from a data source over an internal network and the account that is used to access the data source is different from the account that you use to access DataWorks, you must perform the following operations:

  i. Use a network connectivity service of Alibaba Cloud to establish a connection between the networks of the two Alibaba Cloud accounts. You can use a network connectivity service such as VPN Gateway or Express Connect.

  ii. Associate an exclusive resource group for Data Integration with the VPC that is connected to the network of the Alibaba Cloud account that is used to access the data source.

  iii. Add a custom route to your data center and add the IP address of the destination data source to your data center.

  iv. Add the CIDR block of the vSwitch to which the exclusive resource group for Data Integration is bound to the whitelist of the data source. For more information, see Add the EIP or CIDR block of an exclusive resource group for Data Integration to the whitelist of a data source.

What do I do if the network connectivity test for a data source in a VPC fails?

- For a data source that is added by using a VPC endpoint:

  i. Make sure that you have associated an exclusive resource group for Data Integration with the VPC in which the data source resides.

  ii. Make sure that you have added the CIDR block of the vSwitch to which the exclusive resource group for Data Integration is bound to the whitelist of the data source. For more information, see Add the EIP or CIDR block of an exclusive resource group for Data Integration to the whitelist of a data source.

- For a data source that is added by using a public endpoint, if the network connectivity test fails, check whether you have added the EIP of the exclusive resource group for Data Integration to the whitelist of the data source. For more information, see Add the EIP or CIDR block of an exclusive resource group for Data Integration to the whitelist of a data source.

  Note You are charged for the traffic generated over the Internet. For more information, see Billing of Internet traffic.

I cannot find the exclusive resource group for Data Integration that I purchased when I test network connectivity for a data source or run a data synchronization node. What do I do?

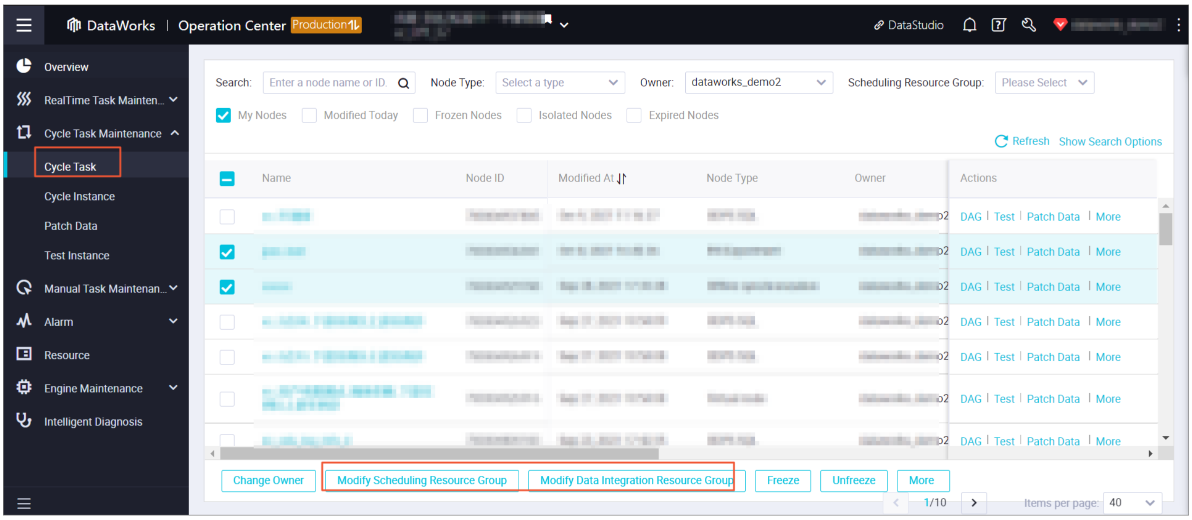
Make sure that you have associated the exclusive resource group for Data Integration with a DataWorks workspace. For more information, see Associate an exclusive resource group with a workspace.

How can I determine the type of the resource group on which a data synchronization node is run from a log?

- If the node is run on the shared resource group, the log contains the following information: `running in Pipeline[basecommon_ group_xxxxxxxxx]` .
- If the node is run on a custom resource group for Data Integration, the log contains the following information: `running in Pipeline[basecommon_xxxxxxxxx]` .
- If the node is run on an exclusive resource group for Data Integration, the log contains the following information: `running in Pipeline[basecommon_S_res_group_xxx]` .

How do I change the type of the resource group on which a data synchronization node is run?

- Change the type of the resource group for scheduling and the type of the resource group for Data Integration on which a data synchronization node is run in the production environment in Operation Center:

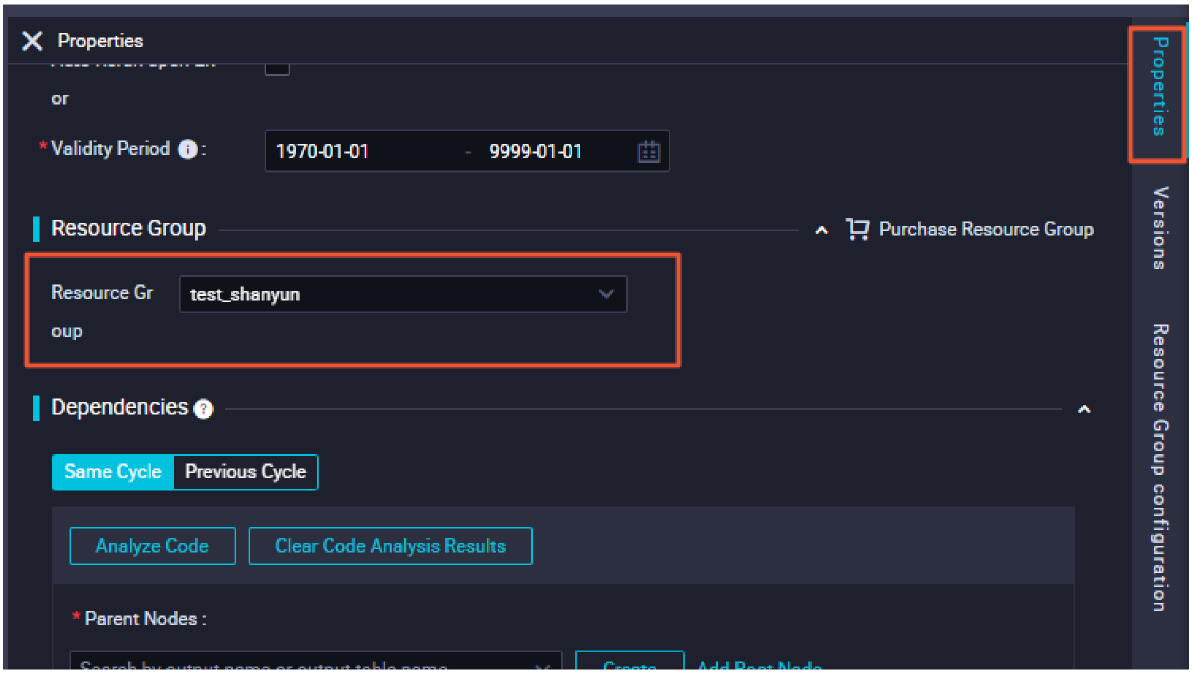


Owner: dataworks_demo2

- Change the type of the resource group on which a data synchronization node is run in the production environment based on the deployment process on the DataStudio page:

> **Note** After you perform the following operations to change the type of the resource group on which a data synchronization node is run, click Deploy to apply the change. For a workspace in standard mode, the change takes effect only in the development environment if you click only Submit. If you want to apply the change to an auto triggered node in the production environment, you must also click Deploy. After the auto triggered node is committed and deployed, you can view the type of the resource group on the Cycle Task page in Operation Center.

1. Change the type of the resource group for scheduling.



Properties

or

* Validity Period ⓘ: 1970-01-01 - 9999-01-01

Resource Group

Purchase Resource Group

Resource Group: test_shanyun

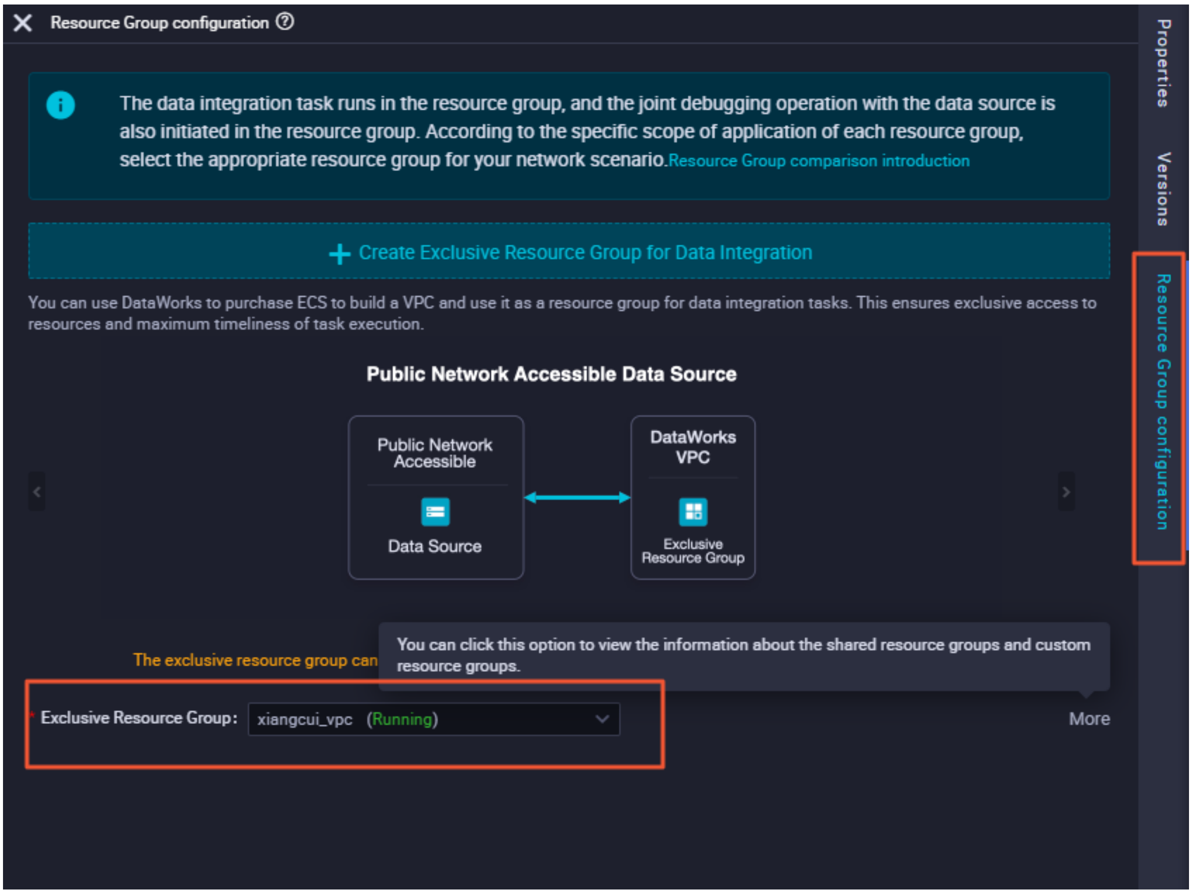
Dependencies

Same Cycle
Previous Cycle
Analyze Code
Clear Code Analysis Results

* Parent Nodes :

Search by output name or output table name
Create
Add Root Node
Properties
Versions
Resource Group configuration

2. Change the type of the resource group for Data Integration.



Resource Group configuration

The data integration task runs in the resource group, and the joint debugging operation with the data source is also initiated in the resource group. According to the specific scope of application of each resource group, select the appropriate resource group for your network scenario.Resource Group comparison introduction

You can use DataWorks to purchase ECS to build a VPC and use it as a resource group for data integration tasks. This ensures exclusive access to resources and maximum timeliness of task execution.

Public Network Accessible Data Source

The exclusive resource group can

You can click this option to view the information about the shared resource groups and custom resource groups.

Exclusive Resource Group: xiangcui_vpc (Running)

How do I configure a resource group to wait for gateway resources?

Log on to the DataWorks console. In the left-side navigation pane, click Resource Groups. The Custom Resource Groups tab appears by default. Find the resource group that is used to run a node and click Deploy in the Actions column. In the Create Deploy Task dialog box, check whether the server is in the Stopped state and whether the server is occupied by other nodes.

If the issue persists, run the following command to restart the service:

```
su - admin /home/admin/alisatasknode/target/alisatasknode/bin/serverctl restart
```

1.3. Real-time synchronization

- Synchronization node configuration
  - What types of data sources support real-time synchronization?
  - Why is the Internet not recommended for real-time synchronization?
  - What operation does DataWorks perform on the data records that are synchronized in real time?
  - How do I deal with the TRUNCATE statement during real-time data synchronization?
  - How do I improve the speed and performance of real-time synchronization?
  - Can I directly run a real-time synchronization node on the codeless user interface (UI)?
- Error for real-time synchronization from Kafka

  When I run a node to synchronize data from Kafka in real time, the following error message appears: Startup mode for the consumer set to timestampOffset, but no begin timestamp was specified.. What do I do?
- Errors for real-time synchronization from MySQL
  - When I run a node to synchronize data from MySQL in real time, the following error message appears: Cannot replicate because the master purged required binary logs.. What do I do?
  - When I run a node to synchronize data from MySQL, the following error message appears: MysqlBinlogReaderException. What do I do?
  - When I run a node to synchronize data from MySQL, the following error message appears: show master status' has an error!. What do I do?
  - When I run a node to synchronize data from MySQL in real time, the following error message appears: parse.exception.PositionNotFoundException: can't find start position forxxx. What do I do?
  - When I run a node to synchronize data from MySQL in real time, data can be read at the beginning but cannot be read after a period of time. What do I do?
- Error for real-time synchronization from Hologres

  When I run a node to synchronize data from Hologres in real time, the following error message appears: permission denied for database xxx. What do I do?

What types of data sources support real-time synchronization?

For more information about the types of data sources that support real-time synchronization, see Plug-ins for data sources that support real-time synchronization.

Why is the Internet not recommended for real-time synchronization?

Real-time synchronization over the Internet has the following disadvantages:

- Packet loss may occur and the performance of data synchronization may be affected due to unstable network connection.
- The security of data synchronization is low.

What operation does DataWorks perform on the data records that are synchronized in real time?

When Data Integration synchronizes data from a data source such as MySQL, Oracle, LogHub, or PolarDB to DataHub or Kafka in real time, Data Integration adds five fields to the data records synchronized to the destination. These fields are used for operations such as metadata management, sorting, and deduplication. For more information, see Fields used for real-time synchronization.

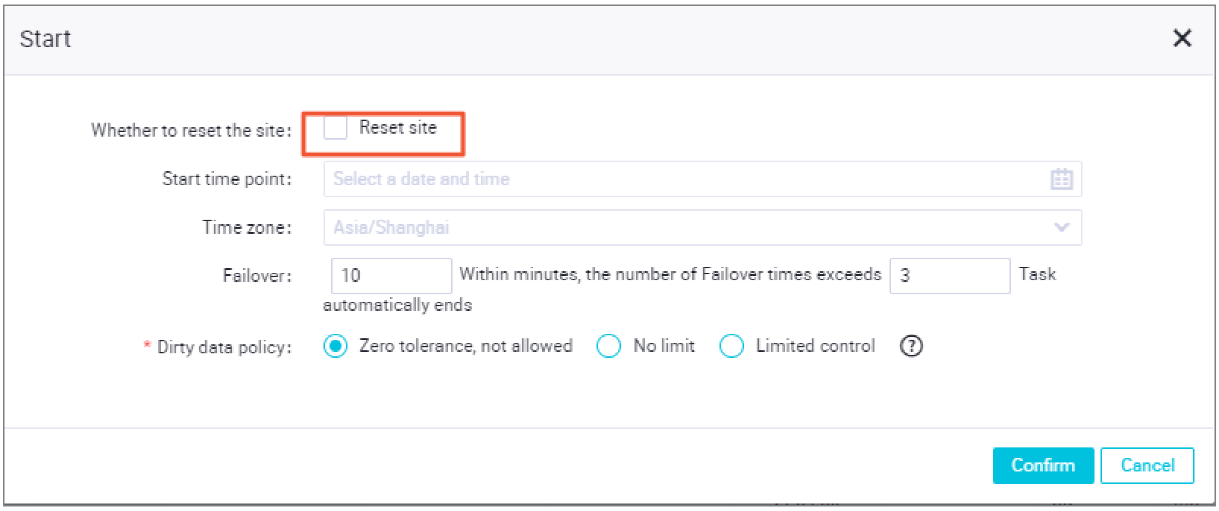
Why does my real-time synchronization node have high latency?

The high latency may be caused by the following reasons:

- The amount of incremental data in the source is small or excessively large.
- The network connection is poor. We recommend that you do not use the Internet for real-time synchronization.
- The offset from which data starts to be synchronized is earlier than the current time. As a result, it takes a period of time to read the historical data before data can be read in real time.

When I run a node to synchronize data from Kafka in real time, the following error message appears: `Startup mode for the consumer set to timestampOffset, but no begin timestamp was specified.`. What do I do?

Specify an offset from which you want to synchronize data.



Start

Whether to reset the site: □ Reset site

Start time point: Select a date and time

Time zone: Asia/Shanghai

Failover: 10 Within minutes, the number of Failover times exceeds 3 Task automatically ends

* Dirty data policy: ● Zero tolerance, not allowed ○ No limit ○ Limited control

When I run a node to synchronize data from MySQL in real time, the following error message appears: `Cannot replicate because the master purged required binary logs.`. What do I do?

Data Integration cannot find the binary logs generated for the offset from which you want to synchronize data. You must check the retention duration of the binary logs of your MySQL data source and specify an offset within the retention duration when you start your synchronization node.

Note If Data Integration cannot find the binary logs, you can reset the offset to the current time.

When I run a node to synchronize data from MySQL, the following error message appears: `MysqlBinlogReaderException`. What do I do?

The binary logging feature is disabled for the secondary MySQL database. If you want to synchronize data from the secondary MySQL database, you must enable this feature for the secondary database. To enable the feature, consult the administrator of the database.

For more information, see Enable the binary logging feature for the MySQL database.

When I run a node to synchronize data from MySQL, the following error message appears: `show master status' has an error!`. What do I do?

If the detailed information of the error is `Caused by: java.io.IOException: message=Access denied; you need (at least one of) the SUPER, REPLICATION CLIENT privilege(s) for this operation, with command: show master status`, the account that you use has no permissions to access the source.

The account that you use to access the source must have the SELECT, REPLICATION SLAVE, and REPLICATION CLIENT permissions on the MySQL database. For more information about how to grant the required permissions on the database to an account, see Create an account and grant the required permissions to the account.

When I run a node to synchronize data from MySQL in real time, the following error message appears: `parse.exception.PositionNotFoundException: can't find start position forxxx`. What do I do?

Data Integration cannot find the binary logs generated for the offset from which you want to synchronize data. You must reset an offset for the node.

When I run a node to synchronize data from MySQL in real time, data can be read at the beginning but cannot be read after a period of time. What do I do?

1. Run the following command on the desired MySQL database to view the binary log files that record the data write operation in the database:

   ```
   show master status
   ```

2. Search for `journalName=mysql-bin.xx,position=xx` in the binary log files of the MySQL database to check whether the binary log files contain data records about the offset specified by the position parameter. For example, you can search for journalName=mysql-bin.000001,position=50.
3. Contact the database administrator if data is being written to the MySQL database but no data write operations are recorded in binary logs.

How do I deal with the TRUNCATE statement during real-time data synchronization?

Real-time synchronization supports the TRUNCATE statement. The TRUNCATE statement takes effect when full and incremental data is merged. If you do not execute the TRUNCATE statement, excessive data may be generated during data synchronization.

How do I improve the speed and performance of real-time synchronization?

If data is written to the destination at a low speed, you can set the number of parallel threads to a larger value and adjust the values of the Java Virtual Machine (JVM) parameters. The values of the JVM parameters affect only the frequency of full heap garbage collection (Full GC). A large JVM heap memory reduces the frequency of full GC and improves the performance of real-time synchronization.

When I run a node to synchronize data from Hologres in real time, the following error message appears: `permission denied for database xxx`. What do I do?

Before you run a node to synchronize data from Hologres in real time, you must obtain the permissions of the <db>_admin user group in the Hologres console for your account. For more information, see Overview.

Can I directly run a real-time synchronization node on the codeless user interface (UI)?

You cannot directly run a real-time synchronization node on the codeless UI. You must commit and deploy the real-time synchronization node and run the node in the production environment. For more information, see Create, configure, commit, and manage real-time sync nodes.

1.4. Batch synchronization

This topic provides answers to some commonly asked questions about batch synchronization.

- O&M of batch synchronization nodes
  - Why is the connectivity test of a data source successful, but the corresponding batch synchronization node fails to be run?
  - How do I change the resource group that is used to run a data synchronization node in Data Integration?
  - How do I locate and handle dirty data?
  - What do I do if a batch synchronization node runs for an extended period of time?
- Common plug-in errors
  - How do I handle a dirty data error that is caused by encoding format configuration issues or garbled characters?
  - What do I do if the error message [TASK_MAX_SLOT_EXCEED]:Unable to find a gateway that meets resource requirements. 20 slots are requested, but the maximum is 16 slots. is returned?
  - What do I do if a server-side request forgery (SSRF) attack is detected in a node?
  - What do I do if the error message OutOfMemoryError: Java heap space is returned when I run a batch synchronization node?
  - What do I do if the same batch synchronization node fails to be run occasionally?
  - What do I do if the error message Duplicate entry 'xxx' for key 'uk_uk_op' is returned when I run a batch synchronization node?

  - What do I do if the error message plugin xx does not specify column is returned when I run a batch synchronization node?
  - What do I do if a synchronization node fails to be run because the name of a column in the source table is a keyword?

- Errors of specific plug-ins
  - What do I do if an error occurs when I add a MongoDB data source as the root user?
  - The authDB database used by MongDB is the admin database. How do I synchronize data from business databases?
  - How do I convert the values of the variables in the query parameter into values in the timestamp format when I synchronize incremental data from a table of a MongDB database?
  - What do I do if the error message AccessDenied The bucket you access does not belong to you. is returned when I read data from an OSS bucket?
  - Is an upper limit configured for the number of OSS objects that can be read?
  - What do I do if the error message Code:[RedisWriter-04], Description:[Dirty data]. - source column number is in valid! is returned when I write data to Redis in hash mode?
  - What do I do if the following error message is returned when I read data from or write data to ApsaraDB RDS for MySQL: Application was streaming results when the connection failed. Consider raising value of 'net_write_timeout/net_read_timeout,' on the server.?
  - What do I do if the error message The last packet successfully received from the server was 902,138 milliseconds ago is returned when I read data from MySQL?
  - What do I do if an error occurs when I read data from PostgreSQL?
  - What do I do if the error message Communications link failure is returned?
  - What do I do if the error message The download session is expired. is returned when I read data from a MaxCompute table?
  - What do I do if the error message Error writing request body to server is returned when I write data to a MaxCompute table?
  - What do I do if data fails to be written to DataHub because the amount of data that I want to write to DataHub at a time exceeds the upper limit?
  - What do I do if the JSON data returned based on the path:[] condition is not of the ARRAY type when I use RestAPI Writer to write data?

- Batch synchronization
  - How do I customize table names in a batch synchronization node?
  - What do I do if the table that I want to select does not appear in the Table drop-down list in the Source section when I configure a batch synchronization node?
  - What are the items that I must take note of when I use the Add feature in a synchronization node that reads data from the MaxCompute table?
  - How do I read data in partition key columns from a MaxCompute table?
  - How do I synchronize data from multiple partitions of a MaxCompute table?
  - What do I do if a synchronization node fails to be run because the name of a column in the source table is a keyword?
  - Why is no data obtained when I read data from a LogHub table whose columns contain data?
  - Why is some data missing when I read data from a LogHub data source?

  - What do I do if the fields that I read based on the field mapping configuration in LogHub are not the expected fields?
  - I configured the endDateTime parameter to specify the end time for reading from a Kafka data source, but some data that is returned is generated at a time point later than the specified end time. What do I do?
  - How do I remove the random strings that appear in the data I write to OSS?
  - How does the system synchronize data from a MySQL data source on which sharding is performed to a MaxCompute table?

Why is the connectivity test of a data source successful, but the corresponding batch synchronization node fails to be run?

- If the data source has passed the connectivity test, you can test the connectivity again to make sure that the resource group that you use is connected to the data source and the data source remains unchanged.
- Check whether the resource group that is connected to the data source is the same as the resource group that you use to run a batch synchronization node.

  Check the resource that is used to run a node:

  - If the node is run on the shared resource group for Data Integration, the log contains the following information: `running in Pipeline[basecommon_ group_xxxxxxxxx]` .
  - If the node is run on a custom resource group for Data Integration, the log contains the following information: `running in Pipeline[basecommon_xxxxxxxxx]` .
  - If the node is run on an exclusive resource group for Data Integration, the log contains the following information: `running in Pipeline[basecommon_S_res_group_xxx]` .
- If the node that is scheduled to run in the early morning occasionally fails but reruns successfully, check the load of the data source at the time when the failure occurred.

How do I change the resource group that is used to run a data synchronization node in Data Integration?

You can go to **DataStudio** or **Operation Center** to change the resource group that is used to run a data synchronization node in Data Integration. For more information, see Associate an exclusive resource group with a workspace.

How do I locate and handle dirty data?

Definition: If an exception occurs when a single data record is written to the destination, the data record is considered as dirty data. Therefore, data records that fail to be written to the destination are considered as dirty data.

Impact: Dirty data fails to be written to the destination. You can control whether dirty data can be generated and the maximum number of dirty data records that can be generated. By default, dirty data is allowed in Data Integration. You can specify the maximum number of dirty data records that can be generated when you configure a synchronization node. For more information, see Configure channel control policies.

- Dirty data is allowed in a synchronization node: If a dirty data record is generated, the synchronization node continues to run. However, the dirty data record is discarded and is not written to the destination.
- The maximum number of dirty data records that can be generated is specified for a synchronization

node:

- ○ If you set the maximum number of dirty data records that can be generated to 0, the synchronization node fails and exits when a dirty data record is generated.
- ○ If you set the maximum number of dirty data records that can be generated to x, the synchronization node fails and exits when the number of dirty data records exceeds x. The synchronization node continues to run if the number of dirty data records is less than x. However, the dirty data records are discarded and are not written to the destination.

Analysis of dirty data generated during data synchronization:

- Problem description: The following error message is returned: `{"message":"Dirty data is found in the data that is to be written to the destination MaxCompute table: The [third] field contains dirty data. Check and correct the data, or increase the threshold value and ignore this dirty data record.","record":[{"byteSize":0,"index":0,"type":"DATE"},{"byteSize":0,"index":1,"type":"DATE"},{"byteSize":1,"index":2,"rawData":0,"type":"LONG"},{"byteSize":0,"index":3,"type":"STRING"},{"byteSize":1,"index":4,"rawData":0,"type":"LONG"},{"byteSize":0,"index":5,"type":"STRING"},{"byteSize":0,"index":6,"type":"STRING"}` .
- The logs show that the third field contains dirty data. You can identify the cause of dirty data based on the following scenarios:
    - ○ If dirty data is reported by a writer, you must check the CREATE TABLE statement of the writer. The data size of the specified field in the destination MaxCompute table is less than the data size of the same field in the source MySQL table.
    - ○ If you want to write data from the source to the destination, the following requirements must be met: 1. The data type in source columns must match the data type in destination columns. For example, data of the VARCHAR type in source columns cannot be written to the destination columns that contain data of the INT type. 2. The size of data defined by the data type of destination columns must be sufficient to receive the data in the mapping columns in the source. For example, you can write data of the LONG, VARCHAR, or DOUBLE type from the source to the columns that contain data of the STRING or TEXT type.
    - ○ If a dirty data error is not clear, you must copy and print out dirty data records, observe the data, and then compare the data type of the records with the data type in destination columns to identify dirty data records.

    Example:

    ```
    {"byteSize":28,"index":25,"rawData":"ohOM71vdGKqXOqtmtriUs5QqJsf4","type":"STRING"}
    ```

    byteSize: the number of bytes. index:25: the 26th field. rawData: a specific value. type: the data type.

What do I do if a batch synchronization node runs for an extended period of time?

Possible cause 1: The batch synchronization node itself takes an extended period of time to complete.

- The SQL statements to be executed before or after synchronization, such as the statements specified by the preSQL or postSQL parameter, take an extended period of time to execute in databases. This prolongs the period of running the batch synchronization node.
- The shard key is not properly configured. As a result, the batch synchronization node is slow.

A batch synchronization node uses the shard key specified by the splitPk parameter to shard data. Then, the node concurrently synchronizes data shards to improve the efficiency of data synchronization. For more information about whether a shard key is required by a plug-in, see the documentation about the plug-in.

Solution:

- When you configure the SQL statements to be executed before or after synchronization, we recommend that you filter data by using fields that have been indexed.
- Properly configure a shard key. For example, when you configure a shard key for MySQL Reader, take note of the following items:
    - We recommend that you set the splitPk parameter to the name of the primary key column of the table. Data can be evenly distributed to different shards based on the primary key column, instead of being intensively distributed only to specific shards.
    - A shard key can be used to shard only data of integer data types. If you set the splitPk parameter for data of other data types, MySQL Reader still uses a single thread to read data.
    - If the splitPk parameter is not provided or is left empty, MySQL Reader uses a single thread to read data.

Possible cause 2: The batch synchronization node waits for resources.

Solution: If the logs show that the node waits for resources for an extended period of time, the number of concurrent nodes on the exclusive resource group for Data Integration to be used reaches the upper limit. For more information about the specific causes and solutions, see Why does a data synchronization node wait for resources for an extended period of time?.

> **Note** Resource groups for scheduling distribute batch synchronization nodes to resource groups for Data Integration to run the batch synchronization nodes. Therefore, a batch synchronization node also occupies resources on a resource group for scheduling. If the resources are occupied for a long period of time, other batch synchronization nodes and nodes of other types may be blocked.

How do I handle a dirty data error that is caused by encoding format configuration issues or garbled characters?

- Problem description:

  If data contains emoticons, a dirty data error message similar to the following error message may be returned when you synchronize the data: `[13350975-0-0-writer] ERROR StdoutPluginCollector - dirty data {"exception":"Incorrect string value: '\\xF0\\x9F\\x98\\x82\\xE8\\xA2...' for column 'introduction' at row 1","record":[{"byteSize":8,"index":0,"rawData":9642,"type":"LONG"},}],"type":"writer"}` .
- Cause:
    - utf8mb4 is not configured for a data source. As a result, an error is reported when data that contains emoticons is synchronized.
    - Data in the source contains garbled characters.
    - The encoding format is different between a data source and a synchronization node.
    - The encoding format of the browser is different from the encoding format of the data source or synchronization node. As a result, the preview fails or the previewed data contains garbled characters.

- Solution:

  The solution varies based on the cause.

  - If data in the source contains garbled characters, process the data before you run a synchronization node.
  - If the encoding format of the data source is different from the encoding format of the synchronization node, modify the configuration for the encoding format of the data source to be the same as the encoding format of the synchronization node.
  - If the encoding format of the browser is different from the encoding format of the data source or synchronization node, modify the configuration for the encoding format of the browser and make sure that the encoding format is the same as the encoding format of the data source and synchronization node. Then, preview the data.

  You can perform the following operations:

  i. If you add a data source by using a Java Database Connectivity (JDBC) URL, add the encoding format utf8mb4 to the JDBC URL. JDBC URL sample: `jdbc:mysql://xxx.x.x.x:3306/database?com.mysql.jdbc.faultInjection.serverCharsetIndex=45` .

  ii. If you add a data source by using an instance ID, suffix the data source name with the encoding format, such as `database?com.mysql.jdbc.faultInjection.serverCharsetIndex=45` .

  iii. Change the encoding format of the data source to utf8mb4. For example, you can change the encoding format of the ApsaraDB RDS data source in the ApsaraDB RDS console.

  > **Note** Run the following command to set the encoding format of the Apsara RDS data source to utf8mb4: `set names utf8mb4` . Run the following command to view the encoding format of the Apsara RDS data source: `show variables like 'char%'` .

What do I do if the error message `[TASK_MAX_SLOT_EXCEED]:Unable to find a gateway that meets resource requirements. 20 slots are requested, but the maximum is 16 slots.` is returned?

- Cause:

  The number of nodes that are run in parallel is set to an excessively large value and the resources are not sufficient to run the nodes.

- Solution:

  Reduce the number of batch synchronization nodes that are run in parallel.

  - If you configure a batch synchronization node by using the codeless user interface (UI), set **Expected Maximum Concurrency** to a smaller value in the Channel step. For more information, see Configure channel control policies.
  - If you configure a batch synchronization node by using the code editor, set the **concurrent** parameter to a smaller value when you configure the channel control policies. For more information, see Configure channel control policies.

What do I do if a server-side request forgery (SSRF) attack is detected in a node?

If the data source is added by using a virtual private cloud (VPC) address, you cannot use the shared resource group for Data Integration to run a node. Instead, you can use an exclusive resource group for Data Integration to run the node. For more information about an exclusive resource group for Data Integration, see Create and use an exclusive resource group for Data Integration.

What do I do if the error message `OutOfMemoryError: Java heap space` is returned when I run a batch synchronization node?

Solution:

1. If you use an exclusive resource group for Data Integration to run a node, you can adjust the values of the Java Virtual Machine (JVM) parameters.
2. If the reader or writer that you use supports the batchsize or maxfilesize parameter, set the batchsize or maxfilesize parameter to a smaller value.

   If you want to check whether a reader or writer supports the batchsize or maxfilesize parameter, see Supported data source types, readers, and writers.
3. Reduce the number of nodes that are run in parallel.
   - If you configure a batch synchronization node by using the codeless user interface (UI), set **Expected Maximum Concurrency** to a smaller value in the Channel step. For more information, see Configure channel control policies.
   - If you configure a batch synchronization node by using the code editor, set the **concurrent** parameter to a smaller value when you configure the channel control policies. For more information, see Configure channel control policies.
4. If you synchronize files, such as Object Storage Service (OSS) files, reduce the number of files that you want to read.

What do I do if the same batch synchronization node fails to be run occasionally?

If a batch synchronization node occasionally fails to be run, a possible cause is that the whitelist configuration of the data source for the node is incomplete.

- Use an exclusive resource group for Data Integration to run the batch synchronization node:
  - If you have added the IP address of the ENI (Elastic Network Interface) of the exclusive resource group for Data Integration to the whitelist of the data source, when the resource group is scaled out, you must add the ENI IP address to the whitelist again to update the whitelist.
  - We recommend that you directly add the CIDR block of the vSwitch to which the exclusive resource group for Data Integration is bound to the whitelist of the data source. Otherwise, you must update the ENI IP address each time the resource group is scaled out. For more information about data marts, see Configure a whitelist.
- Use the shared resource group for Data Integration to run the batch synchronization node:

  Make sure that all the CIDR blocks of the machines that are used for data synchronization in the region where the shared resource group for Data Integration resides are added to the whitelist of the data source. For more information, see Add the IP addresses or CIDR blocks of the servers in the region where the DataWorks workspace resides to the whitelist of a data source.

If the configuration of the whitelist for the data source is complete, check whether the connection between the data source and Data Integration is interrupted due to the heavy load of the data source.

What do I do if an error occurs when I add a MongoDB data source as the root user?

Change the username. You must use the name of the user that has operation permissions on the data source instead of the root user.

For example, if you want to synchronize data of the name table in the test data source, use the name of the user that has operation permissions on the test data source.

The authDB database used by MongoDB is the admin database. How do I synchronize data from business databases?

Enter the name of a business database when you configure a data source to make sure that the user that you use has the required permissions on the business database. If the error message "auth failed" is returned when you test the connectivity of the data source, ignore the error message. If you configure a synchronization node by using the code editor, add the "adthDb":"admin" parameter to the JSON configurations of the synchronization node.

How do I convert the values of the variables in the query parameter into values in the timestamp format when I synchronize incremental data from a table of a MongoDB database?

Use assignment nodes to convert data of the DATE type into data of the TIMESTAMP format and use the timestamp value as an input parameter for data synchronization from MongoDB. For more information, see How do I synchronize incremental data that is in the timestamp format from a table of a MongoDB database?

What do I do if the error message `AccessDenied The bucket you access does not belong to you.` is returned when I read data from an OSS bucket?

The user that is configured for OSS and has the AccessKey pair does not have permissions to access the bucket. Grant the user permissions to access the bucket.

Is an upper limit configured for the number of OSS objects that can be read?

In Data Integration, the number of OSS objects that can be read from OSS is not limited. The maximum number of OSS objects that can be read is determined by the JVM parameters that are configured for a synchronization node. To prevent out of memory (OOM) errors, we recommend that you do not set the Object parameter to an asterisk (*).

What do I do if the error message `Code:[RedisWriter-04], Description:[Dirty data]. - source column number is in valid!` is returned when I write data to Redis in hash mode?

- Cause:

  If you want to store data in Redis in hash mode, make sure that attributes and values are generated in pairs. Example: `odpsReader: "column":[ "id", "name", "age", "address", ]`. In Redis, if RedisWriter: "keyIndexes":[0 ,1] is used, id and name are used as keys, age is used as an attribute, and address is used as a value in Redis. If the source is MaxCompute and only two columns are configured, you cannot store the Redis cache in hash mode, and an error is reported.

- Solution:

  If you want to use only two columns, you must store data in Redis by using the string mode. If you need to store data in hash mode, you must configure at least three columns in the source.

What do I do if the following error message is returned when I read data from or write data to ApsaraDB RDS for MySQL: `Application was streaming results when the connection failed. Consider raising value of 'net_write_timeout/net_read_timeout,' on the server.` ?

- Cause:
  - net_read_timeout: If the error message contains this parameter, the execution time of an SQL statement exceeded the maximum execution time allowed by ApsaraDB RDS for MySQL. The SQL statement is one of the multiple SQL statements that are obtained after a single data acquisition SQL statement is equally split based on the splitpk parameter when you run a synchronization node to read data from the MySQL data source.
  - net_write_timeout: If the error message contains this parameter, the timeout period in which the system waits for a block to be written to a data source is too small.
- Solution:

  Add the net_write_timeout or net_read_timeout parameter to the URL of the ApsaraDB RDS for MySQL database and set the parameter to a larger value. You can also set the net_write_timeout or net_read_timeout parameter to a different value in the ApsaraDB RDS console.
- Suggestion:

  If possible, configure the synchronization node to be rerun automatically.

Example: `jdbc:mysql://192.168.1.1:3306/lizi?useUnicode=true&characterEncoding=UTF8&net_write_timeout=72000`

What do I do if the error message `The last packet successfully received from the server was 902,138 milliseconds ago` is returned when I read data from MySQL?

In this case, the CPU utilization is normal but the memory usage is high. As a result, the data source is disconnected from Data Integration.

If you confirm that the synchronization node can be rerun automatically, we recommend that you configure the node to be automatically rerun if an error occurs. For more information, see Configure time properties.

What do I do if an error occurs when I read data from PostgreSQL?

- Problem description: The error message `org.postgresql.util.PSQLException: FATAL: terminating connection due to conflict with recovery` is returned when I use a batch synchronization tool to synchronize data from PostgreSQL.
- Cause: This error occurs because the system takes a long time to obtain data from the PostgreSQL database. To resolve this issue, specify the max_standby_archive_delay and max_standby_streaming_delay parameters in the code of the synchronization node. For more information, see Standby Server Events.

What do I do if the error message `Communications link failure` is returned?

- Read data from a data source:
  - Problem description:

    The following error message is returned when data is read from a data source: `Communications link failure The last packet successfully received from the server was 7,200,100 milliseconds ago. The last packet sent successfully to the server was 7,200,100 milliseconds ago. - com.mysql.jdbc.exceptions.jdbc4.CommunicationsException: Communications link failure`.
  - Cause:

    Slow SQL queries result in timeout when you read data from MySQL.
  - Solution:
    - Check whether the **WHERE** clause is specified to make sure that an index is added for the filter field.
    - Check whether a large amount of data exists in the source table. If a large amount of data exists in the source table, we recommend that you run multiple nodes to execute the SQL queries.
    - Check the database logs to find which SQL queries are delayed and contact the database administrator to resolve the issue.
- Write data to a data source:
  - Problem description:

    The following error message is returned when data is written to a data source: `Caused by: java.util.concurrent.ExecutionException: ERR-CODE: [TDDL-4614][ERR_EXECUTE_ON_MYSQL] Error occurs when execute on GROUP 'xxx' ATOM 'dockerxxxxx_xxxx_trace_shard_xxxx': Communications link failure The last packet successfully received from the server was 12,672 milliseconds ago. The last packet sent successfully to the server was 12,013 milliseconds ago. More...`.
  - Cause:

    A socket timeout occurred due to slow SQL queries. The default value of the SocketTimeout parameter of Taobao Distributed Data Layer (TDDL) connections is 12 seconds. If the execution time of an SQL statement on a MySQL client exceeds 12 seconds, a TDDL-4614 error is returned. This error occasionally occurs when the data volume is large or the server is busy.
  - Solution:
    - We recommend that you rerun the synchronization node after the database becomes stable.
    - Contact the database administrator to adjust the value of the SocketTimeout parameter.

What do I do if the error message `Duplicate entry 'xxx' for key 'uk_uk_op'` is returned when I run a batch synchronization node?

- Problem description: The following error message is returned: `Error updating database. Cause: com.mysql.jdbc.exceptions.jdbc4.MySQLIntegrityConstraintViolationException: Duplicate entry 'cfc68cd0048101467588e97e83ffd7a8-0' for key 'uk_uk_op'`.
- Cause: In Data Integration, different instances of the same synchronization node cannot be run at the same time. Therefore, multiple synchronization instances that are configured based on the same JSON configurations cannot be run at the same time. For a synchronization node whose instances are run at 5-minute intervals, the instance that is scheduled to run at 00:00 and the instance that is scheduled to run at 00:05 are both run at 00:05 due to a delay caused by the ancestor node of the synchronization node. As a result, one of the instances fails to be run. This issue may occur if you backfill data for or rerun a synchronization node that is running.

- Solution: Stagger the running time of instances. We recommend that you configure nodes that are scheduled to run by hour to depend on their instances in the last cycle. For more information, see Scenario 2: Configure scheduling dependencies for a node that depends on last-cycle instances.

What do I do if the error message `plugin xx does not specify column` is returned when I run a batch synchronization node?

A possible cause is that the field mapping for the batch synchronization node is incorrect or the column parameter is incorrectly configured in a reader or writer.

1. Check whether the mapping between the source fields and the destination fields is configured.
2. Check whether the column parameter is configured in a reader or writer based on your business requirements.

What do I do if the error message `The download session is expired.` is returned when I read data from a MaxCompute table?

- Problem description:

  ```
  Code:DATAX_R_ODPS_005:Failed to read data from a MaxCompute table, Solution:[Contact the administrator of MaxCompute]. RequestId=202012091137444331f60b08cda1d9, ErrorCode=StatusConflict, ErrorMessage=The download session is expired.
  ```

- Cause:

  If you want to read data from a MaxCompute table, you must run a Tunnel command in MaxCompute to upload and download data. On the server, the lifecycle for each Tunnel session spans 24 hours after the session is created. If a batch synchronization node is run for more than 24 hours, it fails to be run and exits. For more information about the Tunnel service, see Usage notes.

- Solution:

  You can increase the number of batch synchronization nodes that can be run in parallel or configure the volume of data to be synchronized to make sure that the volume of data can be synchronized within 24 hours.

What do I do if the error message `Error writing request body to server` is returned when I write data to a MaxCompute table?

- Problem description:

  ```
  Code:[OdpsWriter-09], Description:[Failed to write data to the destination MaxCompute table.]. - Failed to write Block 0 to the destination MaxCompute table, uploadId=[202012081517026537dc0b0160354b]. Contact the administrator of MaxCompute. - java.io.IOException: Error writing request body to server.
  ```

- Cause:
  - Possible cause 1: The data type is incorrect. The source data does not comply with MaxCompute data type specifications. For example, the value 4.2223 cannot be written to the destination MaxCompute table in the format of DECIMAL(precision,scale), such as DECIMAL(18,10).
  - Possible cause 2: The MaxCompute block is abnormal or the communication is abnormal.

- Solution:

  Convert the data type of the data that is to be synchronized into a data type that is supported by the destination. If an error is still reported after you convert the data type, you can submit a ticket for troubleshooting.

What do I do if data fails to be written to DataHub because the amount of data that I want to write to DataHub at a time exceeds the upper limit?

- Problem description:

  ```
   ERROR JobContainer - Exception when job runcom.alibaba.datax.common.exception.DataXException: Code:[DatahubWriter-04], Description:[Failed to write data to DataHub.]. - com.aliyun.datahub.exception.DatahubServiceException: Record count 12498 exceed max limit 10000 (Status Code: 413; Error Code: TooLargePayload; Request ID: 20201201004200a945df0bf8e11a42)
  ```

- Cause:

  The amount of data that you want to write to DataHub at a time exceeds the upper limit that is allowed by DataHub. The following parameters specify the maximum amount of data that can be written to DataHub:

  - **maxCommitSize**: specifies the maximum amount of the buffered data that Data Integration can accumulate before it commits the data to the destination. Unit: MB. The default value is 1048576, in bytes, which is 1 MB.
  - **batchSize**: specifies the maximum number of the buffered data records that a single synchronization task can accumulate before it commits the data records to the destination.

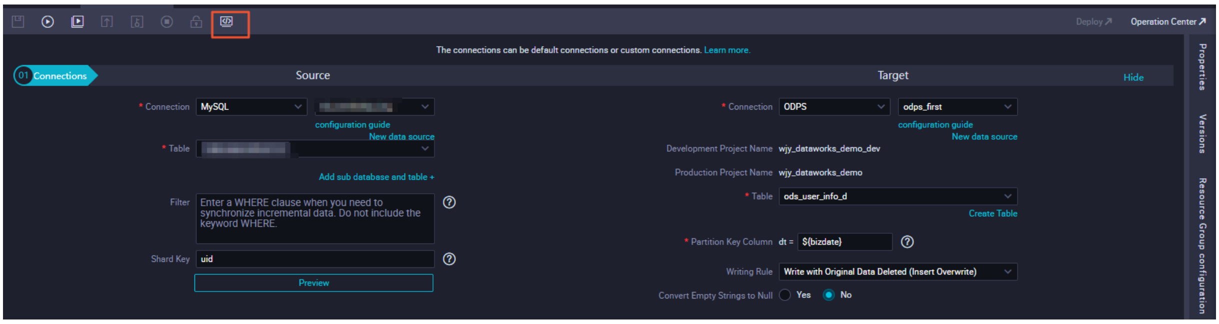
- Solution:

  Set the **maxCommitSize** and **batchSize** parameters to smaller values.

How do I customize table names in a batch synchronization node?

The tables from which you want to synchronize data are named in a consistent format. For example, the tables are named by date and the table schema is consistent, such as **orders_20170310**, **orders_20170311**, and **orders_20170312**. You can specify custom table names by using the scheduling parameters specified in Create a synchronization node by using the code editor. This way, the synchronization node automatically reads table data of the previous day from the source every morning.

For example, if the current day is March 15, 2017, the synchronization node can automatically read data of the **orders_20170314** table from the source.



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

Source:
- * Connection: MySQL [redacted]
- * Table: [redacted]
- Filter: Enter a WHERE clause when you need to synchronize incremental data. Do not include the keyword WHERE.
- Shard Key: uid

Target:
- * Connection: ODPS odps_first
- Development Project Name: wjy_dataworks_demo_dev
- Production Project Name: wjy_dataworks_demo
- * Table: ods_user_info_d
- * Partition Key Column: dt = ${bizdate}
- Writing Rule: Write with Original Data Deleted (Insert Overwrite)
- Convert Empty Strings to Null: ○ Yes ● No

In the code editor, use a variable to specify the name of a source table, such as **orders_${tablename}**. The tables are named by date. If you want the synchronization node to read data of the previous day from the source every day, assign the value ${yyyymmdd} to the ${tablename} variable in the parameter configurations of the synchronization node.

Note

For more FAQ about how to use scheduling parameters, see Overview of scheduling parameters.

What do I do if the table that I want to select does not appear in the Table drop-down list in the Source section when I configure a batch synchronization node?

When you configure a batch synchronization node, the Table drop-down list in the **Source** section displays only the first 25 tables in the selected data source by default. If the selected data source contains more than 25 tables and the table that you want to select does not appear in the Table drop-down list, enter the name of the table in the Table field. You can also configure the batch synchronization node in the code editor.

What are the items that I must take note of when I use the Add feature in a synchronization node that reads data from the MaxCompute table?

1. You can enter constants. Each constant must be enclosed in a pair of single quotation marks ('), such as *'abc'* and *'123'*.
2. You can use the Add feature together with scheduling parameters, such as *'${bizdate}'*. For more information about how to use scheduling parameters, see Overview of scheduling parameters.
3. You can specify the partition key columns from which you want to read data, such as the partition key column pt.
4. If the field that you entered cannot be parsed, the value of the Type parameter for the field is displayed as *Custom*.
5. MaxCompute functions are not supported.
6. If the value of Type for the fields that you manually added, such as the partition key columns of MaxCompute tables, is Custom, synchronization nodes can still be run although the partition key columns cannot be previewed in LogHub.

How do I read data in partition key columns from a MaxCompute table?

Add a data record in the field mapping configuration area, and specify the name of a partition key column, such as pt.

How do I synchronize data from multiple partitions of a MaxCompute table?

Locate the partitions from which you want to read data.

- You can use Linux Shell wildcards to specify the partitions. An asterisk (*) indicates zero or multiple characters, and a question mark (?) indicates a single character.
- The partitions that you specify must exist in the source table. Otherwise, the system reports an error for the synchronization node. If you want the synchronization node to be successfully run even if the partitions that you specify do not exist in the source table, use the code editor to modify the code of the node. In addition, you must add `"successOnNoPartition": true` to the configuration of MaxCompute Reader.

For example, the partitioned table test contains four partitions: pt=1,ds=hangzhou, pt=1,ds=shanghai, pt=2,ds=hangzhou, and pt=2,ds=beijing. In this case, you can set the partition parameter based on the following instructions:

- To read data from the partition pt=1,ds=hangzhou, specify `"partition":"pt=1,ds=hangzhou"` .
- To read data from all the ds partitions in the pt=1 partition, specify `"partition":"pt=1,ds=*"` .
- To read data from all the partitions in the test table, specify `"partition":"pt=*,ds=*"` .

You can also perform the following operations in the code editor to specify other conditions based on which data is read from partitions:

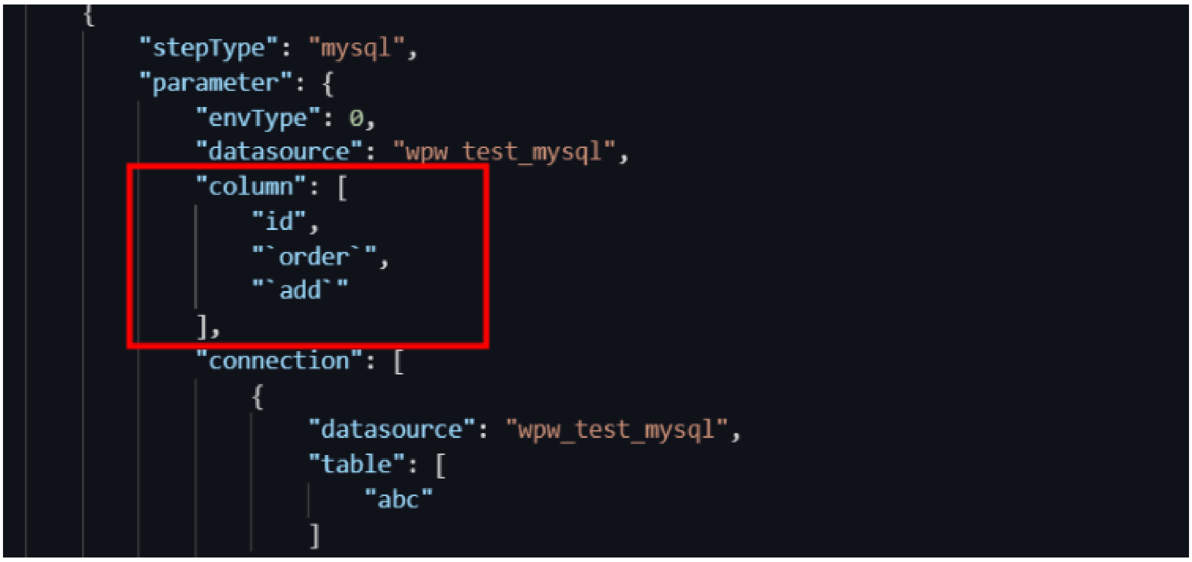
- To read data from the partition that stores the largest amount of data, add `/*query*/ ds=(select MAX(ds) from DataXODPSReaderPPR)` to the configuration of MaxCompute Reader.
- To filter data based on filter conditions, add `/*query*/ pt+Expression` to the configuration of MaxCompute Reader. For example, `/*query*/ pt>=20170101 and pt<20170110` indicates that you want to read the data that is generated from January 1, 2017 to January 9, 2017 from all the pt partitions in the test table.

> **Note**
>
> MaxCompute Reader processes the content that follows `/*query*/` as a WHERE clause.

What do I do if a synchronization node fails to be run because the name of a column in the source table is a keyword?

- Cause: The column parameter contains reserved fields or fields whose names start with a number.
- Solution: Use the code editor to configure a synchronization node in Data Integration and escape special fields in the configuration of the column parameter. For more information about how to use the code editor to configure a synchronization node, see Create a synchronization node by using the code editor.
  - MySQL uses grave accents (`) as escape characters to escape keywords in the following format: `` `Keyword` `` .
  - Oracle and PostgreSQL use double quotation marks (") as escape characters to escape keywords in the following format: `"Keyword"` .
  - SQL Server uses brackets ([]) as escape characters to escape keywords in the following format: `[Keyword]` .

  MySQL escape character example



```
    {
        "stepType": "mysql",
        "parameter": {
            "envType": 0,
            "datasource": "wpw_test_mysql",
            "column": [
                "id",
                "`order`",
                "`add`"
            ],
            "connection": [
                {
                    "datasource": "wpw_test_mysql",
                    "table": [
                        "abc"
                    ]
```

- A MySQL data source is used in the following example:
    1. Execute the following statement to create a table named aliyun, which contains a column named table: `` create table aliyun (`table` int ,msg varchar(10)); ``
    2. Execute the following statement to create a view and assign an alias to the table column: `` create view v_aliyun as select `table` as col1,msg as col2 from aliyun; ``

        Note

        - MySQL uses table as a keyword. If the name of a column in the source table is table, an error is reported during data synchronization. In this case, you must create a view to assign an alias to the table column.
        - We recommend that you do not use a keyword as the name of a column.
    3. You can execute the preceding statement to assign an alias to the column whose name is a keyword. When you configure a synchronization node, use the v_aliyun view to replace the aliyun table.

Why is no data obtained when I read data from a LogHub table whose columns contain data?

In LogHub Reader, column names are case-sensitive. Check for the column name configuration in LogHub Reader.

Why is some data missing when I read data from a LogHub data source?

In Data Integration, a synchronization node reads data from a LogHub data source at the time when the data is generated in LogHub. Check whether the value of the metadata field receive_time, which is configured for reading data, is within the time range specified for the synchronization node in the LogHub console.

What do I do if the fields that I read based on the field mapping configuration in LogHub are not the expected fields?

Manually modify the configuration of the column parameter in the LogHub console.

I configured the endDateTime parameter to specify the end time for reading from a Kafka data source, but some data that is returned is generated at a time point later than the specified end time. What do I do?

Kafka Reader reads data from a Kafka data source in batches. If data that is generated later than the time specified by endDateTime is found in a batch of read data, Kafka Reader stops reading data. However, the data generated later than the end time is also written to the destination.

- You can set the skipExceedRecord parameter to specify whether to write such data to the destination. For more information, see Kafka Reader. To prevent data loss, we recommend that you set the skipExceedRecord parameter to false to ensure that the data generated later than the end time is not skipped.
- You can use the max.poll.records parameter in Kafka to specify the amount of data to poll for at the same time. Configure this parameter and the number of synchronization nodes that can be run in parallel to control the excess data volume that is allowed. The allowed excess volume of data is calculated based on the following formula: Allowed excess data volume < max.poll.records × Number of synchronization nodes that can be run in parallel.

How do I remove the random strings that appear in the data I write to OSS?

The prefix for the names of the files that you want to write to OSS. OSS simulates the directory effect by adding delimiters to file names. such as "object": "datax". This way, the names of the files start with datax and end with random strings. The number of files determines the number of tasks that a synchronization node is split into.

If you do not want to use a random universally unique identifier (UUID) as the suffix, we recommend that you set the writeSingleObject parameter to true. For more information, see the description of the writeSingleObject parameter in OSS Writer.

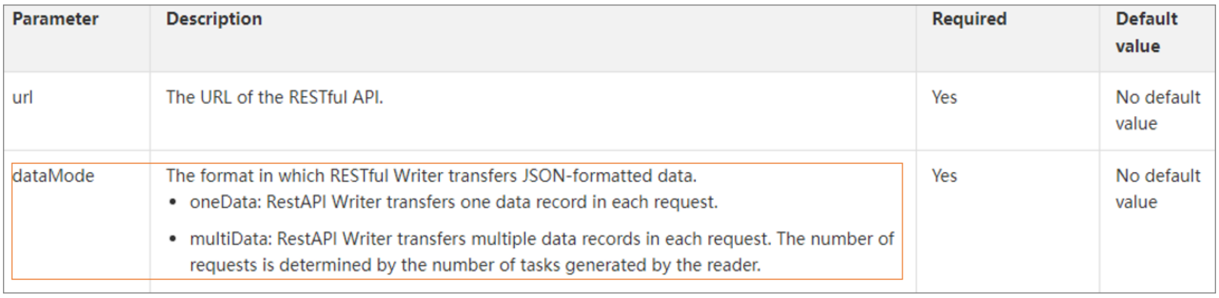
For more information, see OSS Writer.

How does the system synchronize data from a MySQL data source on which sharding is performed to a MaxCompute table?

For more information about how to configure MySQL Reader to read data from a MySQL data source, see MySQL Reader.

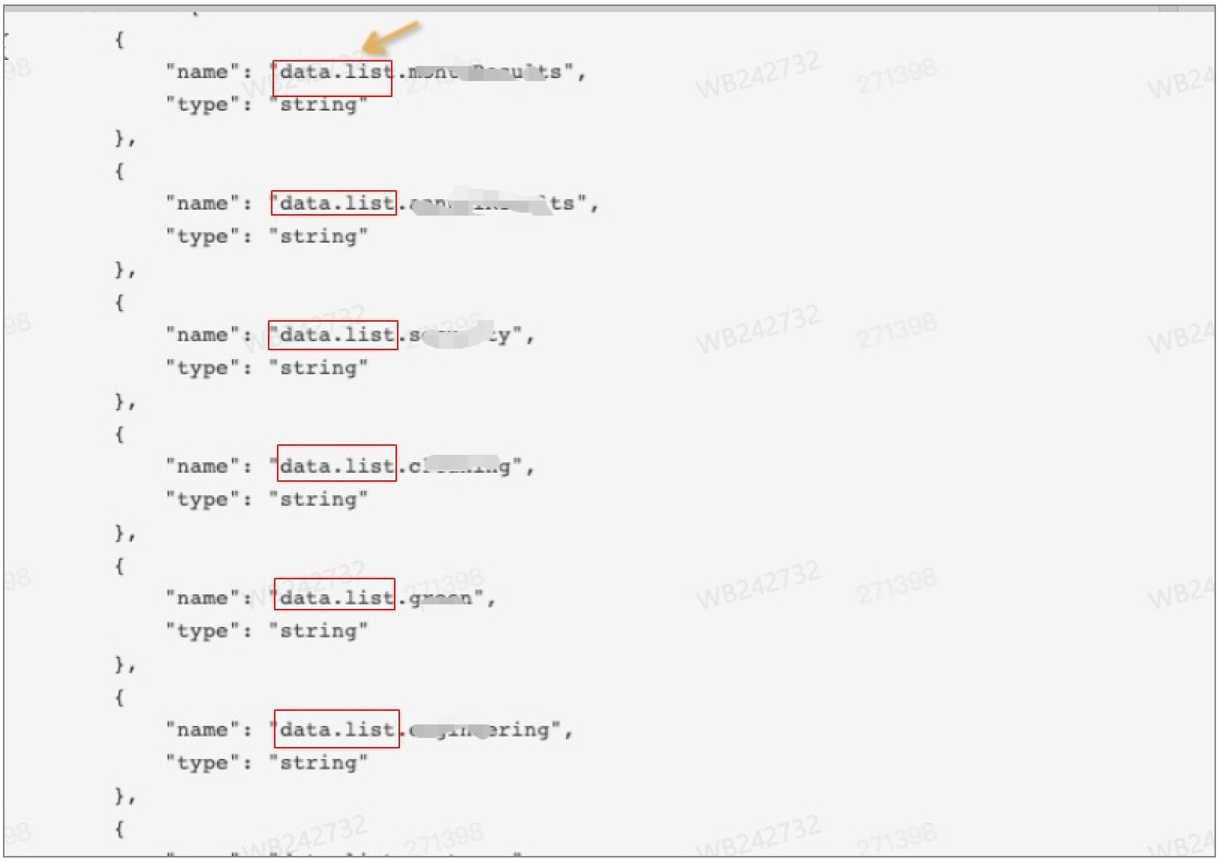
What do I do if the JSON data returned based on the path:[] condition is not of the ARRAY type when I use RestAPI Writer to write data?

The dataMode parameter can be set to oneData or multiData for RestAPI Writer. If you want to use RestAPI Writer to write multiple data records, set dataMode to multiData. For more information, see RestAPI Writer. You must also add the dataPath:"data.list" parameter to the script of RestAPI Reader.



Parameter	Description	Required	Default value
url	The URL of the RESTful API.	Yes	No default value
dataMode	The format in which RESTful Writer transfers JSON-formatted data. - oneData: RestAPI Writer transfers one data record in each request. - multiData: RestAPI Writer transfers multiple data records in each request. The number of requests is determined by the number of tasks generated by the reader.	Yes	No default value

> **Notice** Do not prefix a column name with data.list when you configure the column parameter. The following figure shows column names that are prefixed by data.list.



```
        {
            "name": "data.list.[redacted]ts",
            "type": "string"
        },
        {
            "name": "data.list.[redacted]ts",
            "type": "string"
        },
        {
            "name": "data.list.s[redacted]y",
            "type": "string"
        },
        {
            "name": "data.list.c[redacted]g",
            "type": "string"
        },
        {
            "name": "data.list.g[redacted]",
            "type": "string"
        },
        {
            "name": "data.list.[redacted]ering",
            "type": "string"
        },
        {
```

2.Properties

2.1. Scheduling parameters

This topic provides answers to some frequently asked questions about scheduling parameters.

- Typical scenarios of scheduling parameters
  - I run an instance of a node at 00:00 on the current day to analyze the data in the partition that corresponds to 23:00 on the previous day. However, the data in the partition that corresponds to 23:00 on the current day is analyzed. What do I do?
  - How do I specify a table partition in a format that contains a space, such as pt=yyyy-mm-dd hh24:mi:ss?
  - A node is scheduled to run at the time specified by the $cyctime or $[yyyymmddhh24miss] variable. The node is scheduled to run at 20:00 every day, but the ancestor node of the node fails to run as scheduled. As a result, the node is delayed and runs at 00:00 on the next day. In this case, is the value of the $cyctime or $[yyyymmddhh24miss] variable 20:00 or 00:00?
  - How do I configure the time properties of an ODPS Spark node?
  - How can I reprocess the return values of the scheduling parameters for a node if the node cannot process the return values?
  - What are the differences between the return values of a MaxCompute date function and a scheduling parameter?
- Testing of scheduling parameters
  - How do I test the configurations of the scheduling parameters on the DataStudio page?
  - FAILED: ODPS-0130161:[1,84] Parse exception - invalid token '$'
  - What do I do if the params format error, please check your params(key=values) error is reported?
- Differences in the value assignment logic of scheduling parameters

  What are the differences in the value assignment logic of scheduling parameters among the Run, Run with Parameters, and Perform Smoke Testing in Development Environment modes?
- O&M of scheduling parameters and check of the configurations of the scheduling parameters
  - How do I check the validity of the values of scheduling parameters in the production environment?
  - How do I check whether the values of the scheduling parameters of an instance are valid by viewing logs?
  - How are node instances generated on the day when daylight saving time begins and ends?
  - I configure a scheduling parameter for a node and commit and deploy the node, but the return value of the scheduling parameter remains unchanged. What do I do?

I run an instance of a node at 00:00 on the current day to analyze the data in the partition that corresponds to 23:00 on the previous day. However, the data in the partition that corresponds to 23:00 on the current day is analyzed. What do I do?

- Problem description: The table partition format is day=yyyymmdd,hour=hh24. The $[yyyymmdd]$[hh24-1/24] variable is used to specify the date and time of a partition. If I run an instance at 00:00, the custom variable datetime=$[yyyymmdd] specifies the current day instead of the previous day. As a result, the data in the partition that corresponds to 23:00 on the current day is analyzed.

- Solution: Change the value of datetime to $[yyyymmdd-1/24] and retain the value $[hh24-1/24] for hour.

  How to configure:

  - In the code: `day=datetime, hour={hour},`
  - Scheduling parameters that are configured for nodes: `datetime=[yyyymmdd-1/24],hour=[hh24-1/24]`

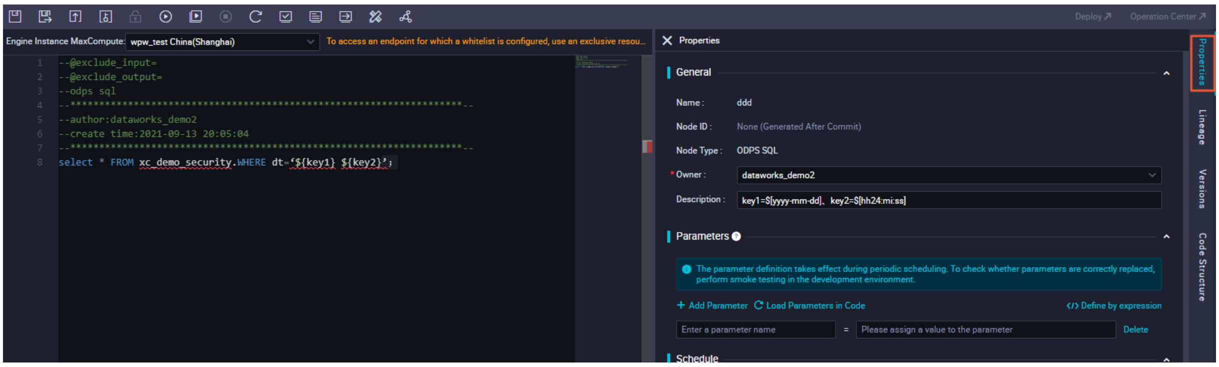
  Scenarios:

  - For an instance that is scheduled to run at 2021-07-21 00:00:00, the value of $[yyyymmdd-1/24] is 20210720, and the value of $[hh24-1/24] is 23. This happens because 1 hour before 2021-07-21 00:00:00 is a point in time on the previous day.
  - For an instance that is scheduled to run at 2021-07-21 01:00:00, the value of $[yyyymmdd-1/24] is 20210721, and the value of $[hh24-1/24] is 00. This happens because 1 hour before 2021-07-21 01:00:00 is still a point in time on the current day.

How do I specify a table partition in a format that contains a space, such as pt=yyyy-mm-dd hh24:mi:ss?

> **Notice** Spaces are not allowed in scheduling parameters.

Solution: Use the custom variable datetime=$[yyyy-mm-dd] to obtain the date and the custom variable hour=$[hh24:mi:ss] to obtain the time. Then, join the variables with a space to form pt=${datetime} ${hour}.



Engine Instance MaxCompute: wpw_test China(Shanghai)

To access an endpoint for which a whitelist is configured, use an exclusive resou...

```
--@exclude_input=
--@exclude_output=
--odps sql
--*********************************************************************--
--author:dataworks_demo2
--create time:2021-09-13 20:05:04
--*********************************************************************--
select * FROM xc_demo_security.WHERE dt='${key1} ${key2}';
```

Properties

General

Name : ddd

Node ID : None (Generated After Commit)

Node Type : ODPS SQL

* Owner : dataworks_demo2

Description : key1=$[yyyy-mm-dd]、key2=$[hh24:mi:ss]

Parameters

The parameter definition takes effect during periodic scheduling. To check whether parameters are correctly replaced, perform smoke testing in the development environment.

+ Add Parameter Load Parameters in Code Define by expression

Enter a parameter name = Please assign a value to the parameter Delete

Schedule

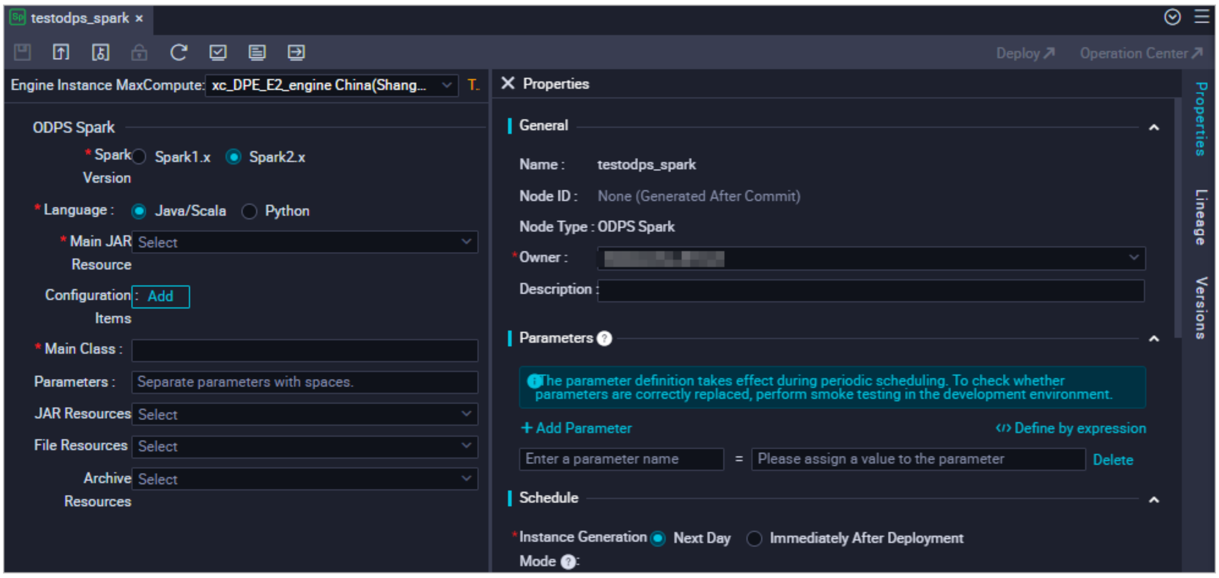
A node is scheduled to run at the time specified by the $cyctime or $[yyyymmddhh24miss] variable. The node is scheduled to run at 20:00 every day, but the ancestor node of the node fails to run as scheduled. As a result, the node is delayed and runs at 00:00 on the next day. In this case, is the value of the $cyctime or $[yyyymmddhh24miss] variable 20:00 or 00:00?

If the resources are insufficient, the time at which an instance actually runs may be different from the time at which the instance is scheduled to run. The scheduled time of an instance is fixed at the time when the instance is generated and does not change even if the time at which the instance runs changes. Therefore, scheduling parameters are configured based on the fixed scheduled time and do not change even if the time at which the instance runs changes.

How do I configure the time properties of an ODPS Spark node?

After you create an ODPS Spark node, you must configure variables in the Parameters field on the node configuration tab.

After you configure the variables, click the **Properties** tab in the right-side navigation pane of the node configuration tab. In the Properties panel, assign values for the variables. You can assign values based on the description of scheduling parameters in this topic.



Engine Instance MaxCompute: xc_DPE_E2_engine China(Shang...

ODPS Spark

* Spark Version: ○ Spark1.x ● Spark2.x

* Language: ● Java/Scala ○ Python

* Main JAR Resource: Select

Configuration Items: Add

* Main Class:

Parameters: Separate parameters with spaces.

JAR Resources: Select

File Resources: Select

Archive Resources: Select

Properties

General

Name: testodps_spark

Node ID: None (Generated After Commit)

Node Type: ODPS Spark

* Owner: [redacted]

Description:

Parameters

The parameter definition takes effect during periodic scheduling. To check whether parameters are correctly replaced, perform smoke testing in the development environment.

+ Add Parameter ‹/› Define by expression

Enter a parameter name = Please assign a value to the parameter Delete

Schedule

* Instance Generation Mode: ● Next Day ○ Immediately After Deployment

How can I reprocess the return values of the scheduling parameters for a node if the node cannot process the return values?

After you configure scheduling parameters for some nodes, such as batch synchronization nodes, the nodes cannot process the return values of the scheduling parameters unless you reprocess the return values. You can configure assignment nodes as the ancestor nodes of these nodes. This way, you can use the assignment nodes to reference the scheduling parameters and reprocess the return values of scheduling parameters. Then, you can use context-based parameters to pass the reprocessed values to the required descendant nodes. For more information about how to configure assignment nodes, see Configure an assignment node. For more information about how to configure context-based parameters, see Configure input and output parameters.

What are the differences between the return values of a MaxCompute date function and a scheduling parameter?

- If you use a MaxCompute date function, the return value is the system time when the instance runs. If the instance runs at different points in time, the return values are different.
- If you use a scheduling parameter, the return result is a calculated result of the scheduled time. If the instance runs at different points in time, the return values remain the same.

How do I test the configurations of the scheduling parameters on the DataStudio page?

The values of the scheduling parameters are automatically replaced in the scheduling system only after you deploy the scheduling parameters in the production environment. If you want to check whether the values of the scheduling parameters are valid on the DataStudio page, click the Perform Smoke Testing in Development Environment icon in the top toolbar of the node configuration tab.

> Note For a data integration node, you cannot check whether the values of the scheduling parameters are valid in the development environment. If you want to perform such a test, you must create an SQL node and then test the configurations of the scheduling parameters by clicking the Perform Smoke Testing in Development Environment icon. If the scheduling parameters pass the test, you can use the configurations of these parameters in the data integration node.

```
FAILED: ODPS-0130161:[1,84] Parse exception - invalid token '$'
```

Cause: The scheduling parameters are not specified or the values of the scheduling parameters are invalid.

Solution:

1. Check whether the scheduling parameters are specified.
2. Check whether the values of the scheduling parameters are valid. For more information, see Overview of scheduling parameters.

> Notice After you modify the scheduling parameters of a node, you must commit and deploy them. After the scheduling parameters are deployed, go to the Cycle Task page in Operation Center and check whether the values of the scheduling parameters are updated on the General tab of the node.

What do I do if the `params format error, please check your params(key=values)` error is reported?

1. Check whether values are assigned to variables.
2. Check whether spaces are used in the scheduling parameters.
3. Check whether a node name contains periods (.) and Chinese characters at the same time.

time1=2$[yyyymmdd3hh24:mi:ss] and time1=$[yyyymmdd]4time2=$[hh24:mi:ss]. Symbols 1, 2, 3, and 4 represent the positions where spaces may be added.

Do not add a space before or after the equal sign (=) in a scheduling parameter. In this example, do not add spaces in the positions specified by Symbols 1 and 2.

Do not include a space in the value of a scheduling parameter. In this example, do not add a space in the position specified by Symbol 3.

Separate two scheduling parameters with a space. In this case, add a space in the position specified by Symbol 4.

What are the differences in the value assignment logic of scheduling parameters among the Run, Run with Parameters, and Perform Smoke Testing in Development Environment modes?

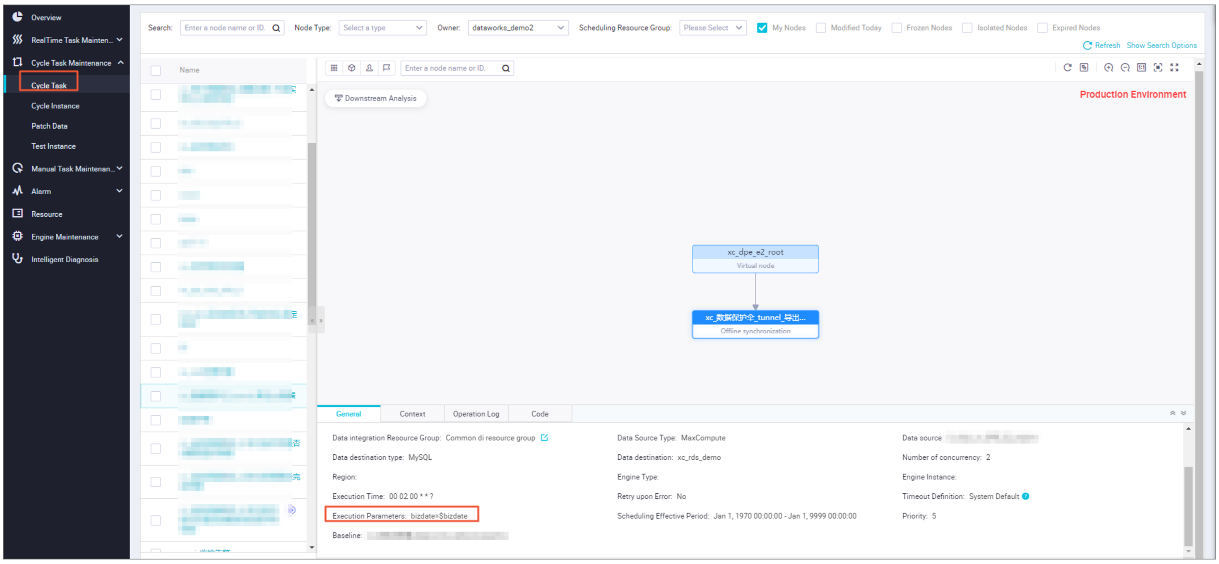
- Run: The first time you click the Run icon, you must manually assign constants to variables in the node code. The constants are recorded by DataWorks. If you modify the code, the variables still use the constants that you assigned.
- Run with Parameters: If you use the Run with Parameters mode, you must manually assign constants to variables in the code. If you modify the variables in the code, you must use the Run with Parameters mode to reassign constants to the variables.
- Perform Smoke Testing in Development Environment: You can enter a data timestamp to simulate

automatic node scheduling and obtain the new values of the scheduling parameters at the specified data timestamp.

> **Note** If you want to change the resource group that is used by a node, click the Run with Parameters icon.

How do I check the validity of the values of scheduling parameters in the production environment?

After you modify the scheduling parameters of a node on the DataStudio page and commit and deploy the node, you can check whether the scheduling parameters are specified based on your requirements on the General tab of the Cycle Task page in Operation Center. If the configurations do not meet your requirements, check whether the deployment package of the node is generated as expected on the deployment page.



Data integration Resource Group: Common di resource group

Data destination type: MySQL

Region:

Execution Time: 00 02 00 * * ?

Execution Parameters: bizdate=$bizdate

Baseline: [redacted]

Data Source Type: MaxCompute

Data destination: xc_rds_demo

Engine Type:

Retry upon Error: No

Scheduling Effective Period: Jan 1, 1970 00:00:00 - Jan 1, 9999 00:00:00

Data source [redacted]

Number of concurrency: 2

Engine Instance:

Timeout Definition: System Default

Priority: 5

Check whether the values of the scheduling parameters of a single instance are valid on the General tab of the Cycle Instance page.

> **Notice** When you modify the scheduling parameters of an auto triggered node for which a single instance is generated, the configurations of the scheduling parameters of the single instance are updated in real time. The real-time update is performed no matter whether the instance is run.

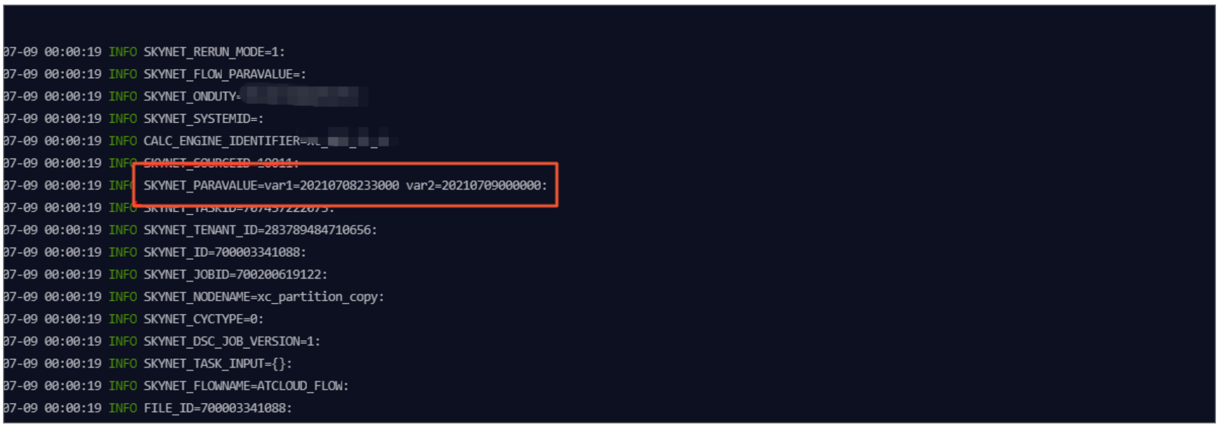
Scenarios:

- For example, you assign $bizdate to the time1 scheduling parameter of Instance A1 of Node A. If Instance A1 is run successfully on the current day, the time1 scheduling parameter is set to the data timestamp specified by bizdate in the code.
- If you change the value of the time1 scheduling parameter from $bizdate to $cyctime at a point in time on the current day, Instance A1 is run at the scheduled time that is specified by cyctime on the current day.
- If you rerun Instance A1, the latest configuration of time1 is used. In this example, time1=$cyctime.
- If you want to view the values of the scheduling parameters of the instance before the value changes, check logs. For more information about how to check whether the values of the scheduling parameters of an instance are valid by viewing logs, see How do I check whether the values of the

scheduling parameters of an instance are valid by viewing logs?.

How do I check whether the values of the scheduling parameters of an instance are valid by viewing logs?

Find `SKYNET_PARAVALUE` in the code.



```
07-09 00:00:19 INFO SKYNET_RERUN_MODE=1:
07-09 00:00:19 INFO SKYNET_FLOW_PARAVALUE=:
07-09 00:00:19 INFO SKYNET_ONDUTY=[redacted]
07-09 00:00:19 INFO SKYNET_SYSTEMID=:
07-09 00:00:19 INFO CALC_ENGINE_IDENTIFIER=[redacted]
07-09 00:00:19 INFO SKYNET_SOURCEID=10011:
07-09 00:00:19 INFO SKYNET_PARAVALUE=var1=20210708233000 var2=20210709000000:
07-09 00:00:19 INFO SKYNET_TASKID=707437222075:
07-09 00:00:19 INFO SKYNET_TENANT_ID=283789484710656:
07-09 00:00:19 INFO SKYNET_ID=700003341088:
07-09 00:00:19 INFO SKYNET_JOBID=700200619122:
07-09 00:00:19 INFO SKYNET_NODENAME=xc_partition_copy:
07-09 00:00:19 INFO SKYNET_CYCTYPE=0:
07-09 00:00:19 INFO SKYNET_DSC_JOB_VERSION=1:
07-09 00:00:19 INFO SKYNET_TASK_INPUT={}:
07-09 00:00:19 INFO SKYNET_FLOWNAME=ATCLOUD_FLOW:
07-09 00:00:19 INFO FILE_ID=700003341088:
```

How are node instances generated on the day when daylight saving time begins and ends?

DataWorks supports the immediate instance generation and daylight saving time-based parameter computing features. This way, nodes can be run as expected when daylight saving time begins or ends. For example, the time zone is UTC-8.

- When daylight saving time begins, 10 minutes before 03:00 is 01:50, and 23 instances are generated on that day. The system does not run the instance that is scheduled to run at 02:00 on that day.
- When daylight saving time ends, 10 minutes before 03:00 is 02:50, and 24 instances are generated on that day.

If a node scheduled by day, week, or month is scheduled to run within the period that is skipped when daylight saving time begins, a node instance is generated and run at 00:00 on that day.

I configure a scheduling parameter for a node and commit and deploy the node, but the return value of the scheduling parameter remains unchanged. What do I do?

Check whether the scheduling parameter is overwritten by a workflow parameter with the same name. For more information, see Use workflow parameters. If a workflow parameter with the same name exists, you can delete the workflow parameter based on your business requirements. If you need to retain the workflow parameter, you must change the name of the scheduling parameter for the node.

2.2. Scheduling dependencies

This topic provides answers to some frequently asked questions about scheduling dependencies.

- Introductions to scheduling dependencies
  - What are scheduling dependencies?
  - Why are scheduling dependencies required?
  - How do I configure scheduling dependencies for a node?

  - Which scenarios do not support scheduling dependencies?
  - How do I delete a table on which a node does not depend?

- How to configure scheduling dependencies
  - The system automatically adds an output name to Parent Nodes for my node based on the automatic parsing feature, but an error message appears, indicating that the output represented by the output name does not exist. What do I do?
  - The name and ID of the descendant node of my node are empty and cannot be specified in the output of my node. Why does this happen?
  - How do I delete the tables on which my node does not depend?
  - What rules are used when a node needs to depend on its ancestor nodes to run?
  - What is the output name of a node used for?
  - Can a node have multiple output names?
  - Can multiple nodes have the same output name?
  - How do I prevent DataWorks from parsing temporary tables when DataWorks parses the scheduling dependencies of a node?
  - How do I configure an ancestor node for the start node of a workflow?

- Node deletion or changes
  - Why do I find a non-existent output name of Node B when I enter an output name to search for the ancestor nodes of Node A?
  - When I undeploy a node, the system displays an error message indicating that the node has descendant nodes and cannot be undeployed. However, no descendant nodes can be found for the node in the Properties panel. Why does this happen?

- How to configure cross-cycle scheduling dependencies in different scenarios
  - Why do some scheduling dependencies of nodes appear as dotted lines in Operation Center?
  - I configure the instance of a node scheduled by hour in the current cycle to depend on the instance of the node in the previous cycle. What are the impacts on this node and its descendant node?
  - How do I configure a dependency in which a node scheduled by day depends on a node scheduled by hour?
  - When does a node scheduled by day start to run if I configure a node scheduled by hour as the ancestor node of the node scheduled by day?
  - How do I configure a node scheduled by day to depend on a specific instance that is generated on the current day for a node scheduled by hour?
  - How do I configure a node scheduled by day to depend on all the instances that are generated on the previous day instead of the current day for a node scheduled by hour?
  - In which scenarios do I need to configure the instance of a node in the current cycle to depend on the instance of the node in the previous cycle?
  - How do I configure dependencies for a node that needs to depend on multiple nodes?
  - Node B scheduled by day depends on Node A scheduled by hour, and Node B starts to run only after all the instances that are generated on the current day for Node A are successful. Will the execution of Node B be affected if Node A still runs on the next day?
  - Node A runs every hour on the hour, and Node B runs once every day. How do I configure Node B to automatically run after the first instance of Node A is run every day?

- How to configure scheduling dependencies in different scenarios
  - How do I configure Node A, Node B, and Node C to run in sequence once per hour?
  - How do I configure dependencies between nodes that reside in the same region but belong to different workspaces and workflows?
- Other frequently asked questions
  - I have configured rerun properties for my node, but the node does not rerun after it fails. In addition, the error message "Task Run Timed Out, Killed by System!!!" appears. What do I do?

What are scheduling dependencies?

Scheduling dependencies define the relationships between nodes. After you configure scheduling dependencies for a node, the node is run only after its ancestor nodes are successful.

> **Note**
>
> After scheduling dependencies are configured for a node, one of the prerequisites to run the node is that its ancestor nodes are successful.

Why are scheduling dependencies required?

Scheduling dependencies ensure that a node can obtain the required data for its execution from its ancestor nodes. A node obtains the required data only after DataWorks detects that the ancestor nodes are successful and generate the latest table data. This prevents a node from obtaining invalid data or obtaining no data before the table data of its ancestor nodes is generated as expected.

How do I configure scheduling dependencies for a node?

Use the output of a node as the input of another node to establish a dependency between the two nodes.

> **Note**
>
> - The system automatically configures an input or an output for an SQL node by using the following methods:
>   - If a table is specified in the SELECT statement of the code for a node, the system adds the table name to Parent Nodes for the node based on the automatic parsing feature.
>   - If a table is specified in the INSERT or CREATE statement of the code for a node, the system adds the table name to Output for the node based on the automatic parsing feature.
> - You must manually add the output of a Data Integration node to Output for the node in the format of Project name.Table name. This way, the system can find the node that generates the output for its descendant node based on the automatic parsing feature.
> - The output name of a node must be unique. This way, the system can find the node that generates the output based on the unique output name.

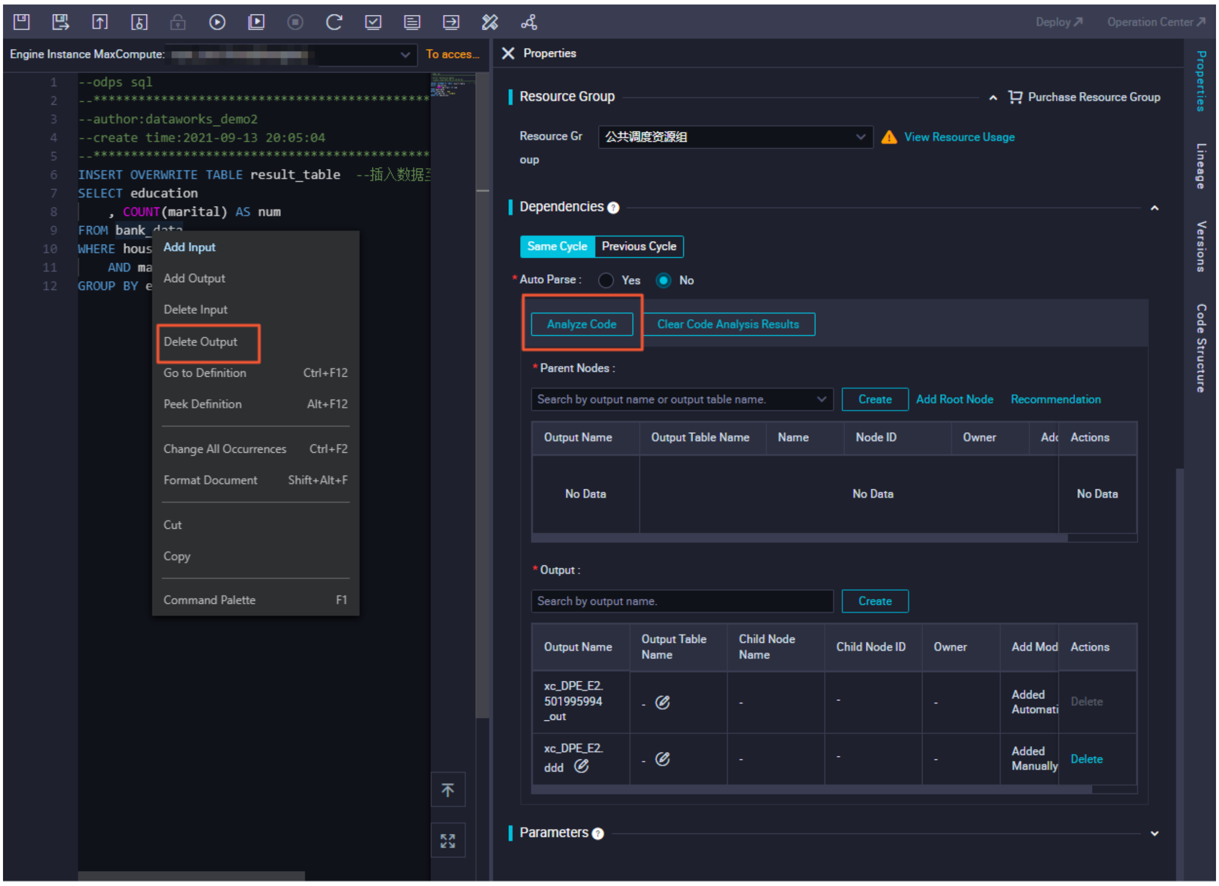
Which scenarios do not support scheduling dependencies?

Scheduling dependencies ensure that a node can obtain the table data generated by its ancestor node that is scheduled to run. However, if the ancestor node of a node is not scheduled to run, the system cannot monitor the generation of the latest table data by the ancestor node. If a node uses a SELECT statement to query data of a table that is not generated by an auto-triggered node, you must manually delete the dependency of the node that is automatically generated by the SELECT statement. Tables that are not generated by auto triggered nodes include the following types:

- Tables uploaded from on-premises machines to DataWorks
- Dimension tables
- Tables that are not generated by nodes scheduled by DataWorks
- Tables generated by manually triggered nodes

How do I delete a table on which a node does not depend?

On the configuration tab of the node, find the table name in the code for the node, right-click the table name, and then select Delete Input. In the Dependencies section of the Properties panel, set the Auto Parse parameter to Yes.



Engine Instance MaxCompute: [redacted]

```
--odps sql
--*******************************************
--author:dataworks_demo2
--create time:2021-09-13 20:05:04
--*******************************************
INSERT OVERWRITE TABLE result_table  --插入数据
SELECT education
    , COUNT(marital) AS num
FROM bank_data
WHERE hous
    AND ma
GROUP BY e
```

Resource Group: 公共调度资源组

Auto Parse : ○ Yes ● No

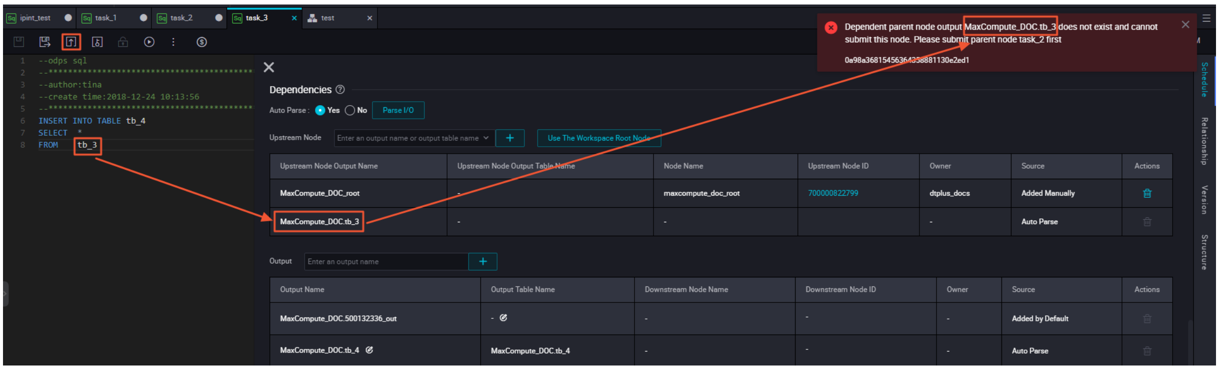
Parent Nodes : Search by output name or output table name.

Output Name	Output Table Name	Name	Node ID	Owner	Adc	Actions
No Data	No Data					No Data

Output : Search by output name.

Output Name	Output Table Name	Child Node Name	Child Node ID	Owner	Add Mod	Actions
xc_DPE_E2.501995994_out	-	-	-	-	Added Automati	Delete
xc_DPE_E2.ddd	-	-	-	-	Added Manually	Delete

The system automatically adds an output name to Parent Nodes for my node based on the automatic parsing feature, but an error message appears, indicating that the output represented by the output name does not exist. What do I do?



```
--odps sql
--*****************************************
--author:tina
--create time:2018-12-24 10:13:56
--*****************************************
INSERT INTO TABLE tb_4
SELECT  *
FROM    tb_3
```

Dependent parent node output MaxCompute_DOC.tb_3 does not exist and cannot submit this node. Please submit parent node task_2 first

0a98a36815456364358881130e2ed1

Dependencies

Auto Parse : ● Yes ○ No Parse I/O

Upstream Node: Enter an output name or output table name + Use The Workspace Root Node

Upstream Node Output Name	Upstream Node Output Table Name	Node Name	Upstream Node ID	Owner	Source	Actions
MaxCompute_DOC_root	-	maxcompute_doc_root	700000822799	dtplus_docs	Added Manually	
MaxCompute_DOC.tb_3	-	-		-	Auto Parse	

Output: Enter an output name +

Output Name	Output Table Name	Downstream Node Name	Downstream Node ID	Owner	Source	Actions
MaxCompute_DOC.500132336_out	-	-	-	-	Added by Default	
MaxCompute_DOC.tb_4	MaxCompute_DOC.tb_4	-	-	-	Auto Parse	

The system fails to find the node that generates the output based on the output name.

This error may be caused by the following reasons:

- The node that generates the output is not committed. Commit the node and try again.
- The node that generates the output is committed, but the output name of the node is different from the output name that is automatically added by the system.

Note

- If tb_2 in the preceding figure is the output table of a node, you must add tb_2 to Output for the node in the format of `Project name.Table name`. For more information, see Logic of same-cycle scheduling dependencies.
- If tb_2 is a table that is not generated by an auto triggered node, you must right-click the table name in the code and select Delete Input to delete the table. In the Dependencies section of the Properties panel, set the Auto Parse parameter to Yes.

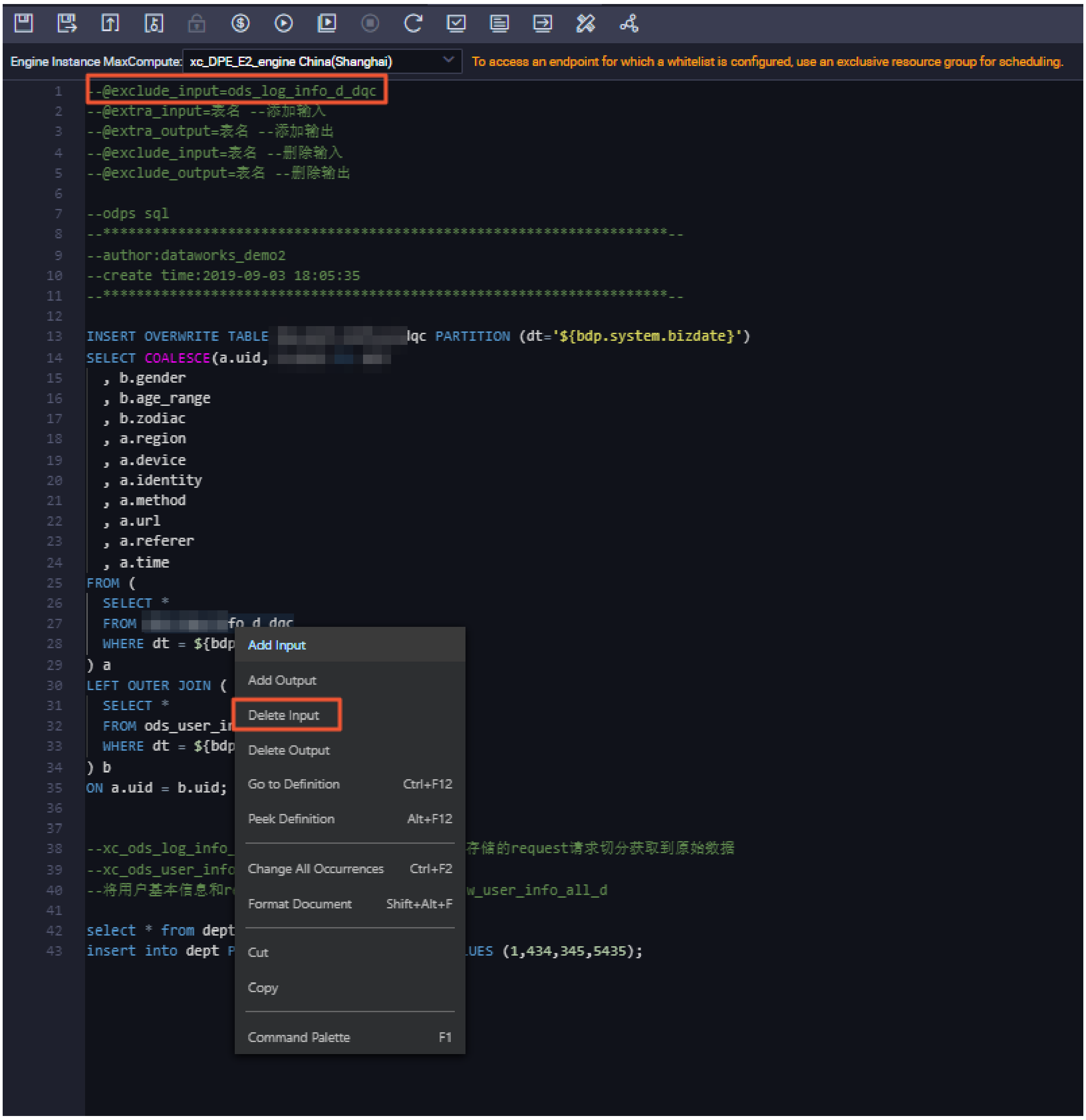
For more information, see Which scenarios do not support scheduling dependencies?.

The name and ID of the descendant node of my node are empty and cannot be specified in the output of my node. Why does this happen?

After you configure the output of a node as the input of another node, scheduling dependencies are established between the two nodes. If a node has no descendant node, the name and ID of the descendant node are empty. After you configure a descendant node for your node, the name and ID of the descendant node are automatically displayed.

How do I delete the tables on which my node does not depend?

On the configuration tab of the node, find the table name in the code for the node, right-click the table name, and then select Delete Input. In the Dependencies section of the Properties panel, set the Auto Parse parameter to Yes.



Engine Instance MaxCompute: xc_DPE_E2_engine China(Shanghai)

To access an endpoint for which a whitelist is configured, use an exclusive resource group for scheduling.

```
--@exclude_input=ods_log_info_d_dqc
--@extra_input=表名 --添加输入
--@extra_output=表名 --添加输出
--@exclude_input=表名 --删除输入
--@exclude_output=表名 --删除输出

--odps sql
--**********************************************************************--
--author:dataworks_demo2
--create time:2019-09-03 18:05:35
--**********************************************************************--

INSERT OVERWRITE TABLE [redacted]lqc PARTITION (dt='${bdp.system.bizdate}')
SELECT COALESCE(a.uid, [redacted]
  , b.gender
  , b.age_range
  , b.zodiac
  , a.region
  , a.device
  , a.identity
  , a.method
  , a.url
  , a.referer
  , a.time
FROM (
  SELECT *
  FROM [redacted]fo_d_dqc
  WHERE dt = ${bdp
) a
LEFT OUTER JOIN (
  SELECT *
  FROM ods_user_i
  WHERE dt = ${bdp
) b
ON a.uid = b.uid;


--xc_ods_log_info_                       存储的request请求切分获取到原始数据
--xc_ods_user_info
--将用户基本信息和r                      w_user_info_all_d

select * from dept
insert into dept P                       UES (1,434,345,5435);
```

What rules are used when a node needs to depend on its ancestor nodes to run?

In the scheduling system of DataWorks, dependencies are configured to ensure that a node can obtain the required data generated by another node. You can determine whether to configure dependencies between nodes based on the data lineage of the tables generated by the nodes. For more information, see Logic of same-cycle scheduling dependencies.

What is the output name of a node used for?

The output name of a node is used to establish a dependency with another node. For example, if the output name of Node A is ABC and Node B uses ABC as its input name, a dependency is established between Node A and Node B.

Can a node have multiple output names?

Yes, a node can have multiple output names. The output name of a node defines the node. If a node (Node A) needs to depend on another node (Node B), Node A can reference an output name of Node B as its input name. This way, a dependency is established between Node A and Node B.

Can multiple nodes have the same output name?

No, multiple nodes cannot have the same output name. The output name of each node must be unique. This way, if a node references the output of another node, the system can find the node that generates the output based on the unique output name and the automatic parsing feature, and a dependency can be established between the two nodes. If multiple nodes generate data to the same table, you must determine the last node that generates data to the table. This ensures that another node can obtain the required data from the table. In addition, you must change the output names of the remaining nodes to ensure that the output names of all nodes are unique.

How do I prevent DataWorks from parsing temporary tables when DataWorks parses the scheduling dependencies of a node?

On the configuration tab of the node, right-click a temporary table name in the SQL code for the node and select **Delete Input** or **Delete Output**. In the Dependencies section of the Properties panel, set the Auto Parse parameter to Yes and click Parse I/O to parse the input and output for the node.

How do I configure an ancestor node for the start node of a workflow?

If you want to configure an ancestor node for the start node of a workflow, you can create a zero load node in the workflow and use the zero load node as the start node of the workflow. Then, you can configure the root node of the workspace as the ancestor node of the zero load node.

For more information about how to use a zero load node, see Create a zero-load node.

Why do I find a non-existent output name of Node B when I enter an output name to search for the ancestor nodes of Node A?

DataWorks searches for the ancestor nodes of a node among the output names of nodes that are committed and deployed to the scheduling system based on the automatic parsing feature. After Node B is committed, if you delete the output name of Node B and do not commit Node B to the scheduling system again, the deleted output name of Node B can still be found.

When I undeploy a node, the system displays an error message indicating that the node has descendant nodes and cannot be undeployed. However, no descendant nodes can be found for the node in the Properties panel. Why does this happen?

You can undeploy a node only after no nodes depend on the node in the development and production environments.

You can go to the **development environment** and **production environment** separately to check whether some nodes still depend on the node.

Why do some scheduling dependencies of nodes appear as dotted lines in Operation Center?

If the scheduling dependencies of a node appear as dotted lines, cross-cycle scheduling dependencies are configured for the node. For more information about cross-cycle scheduling dependencies, see Scenario 2: Configure scheduling dependencies for a node that depends on last-cycle instances.

I configure the instance of a node scheduled by hour in the current cycle to depend on the instance of the node in the previous cycle. What are the impacts on this node and its descendant node?

- Impact on the current node: The instance of the node in the current cycle can be run only after the instance of the node in the previous cycle is successful.

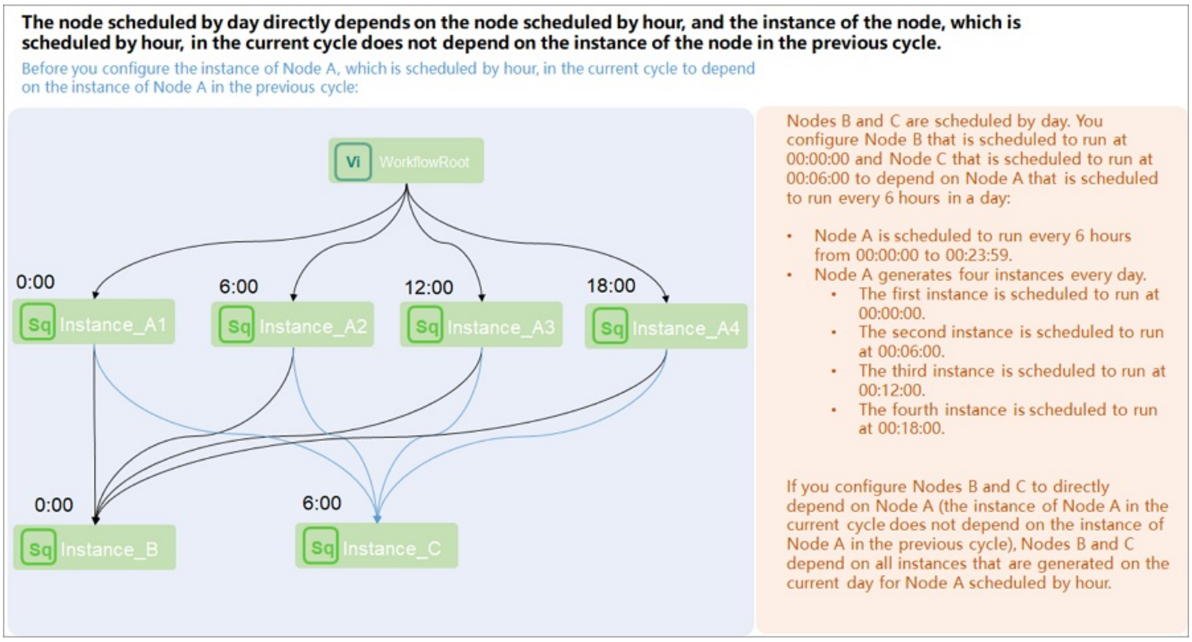
  Scenario: If a node that is scheduled by hour starts to run at 00:00, the instance of the node in the second cycle can be run only after the instance of the node in the first cycle is successful.

- Impact on the descendant node of the current node: If the current node has a descendant node that is scheduled by day, the descendant node no longer directly depends on multiple instances of the current node but instead directly depends only on a specific instance of the current node. In this case, the descendant node indirectly depends on multiple instances of the current node.

How do I configure a dependency in which a node scheduled by day depends on a node scheduled by hour?

- Scenario 1: Configure a node scheduled by day to depend on all the instances that are generated on the **current day** for a node scheduled by hour.

  Configure the node scheduled by day to directly depend on the node scheduled by hour. This way, the node scheduled by day depends on all instances that are generated on the current day for the node scheduled by hour.



The node scheduled by day directly depends on the node scheduled by hour, and the instance of the node, which is scheduled by hour, in the current cycle does not depend on the instance of the node in the previous cycle.

Before you configure the instance of Node A, which is scheduled by hour, in the current cycle to depend on the instance of Node A in the previous cycle:

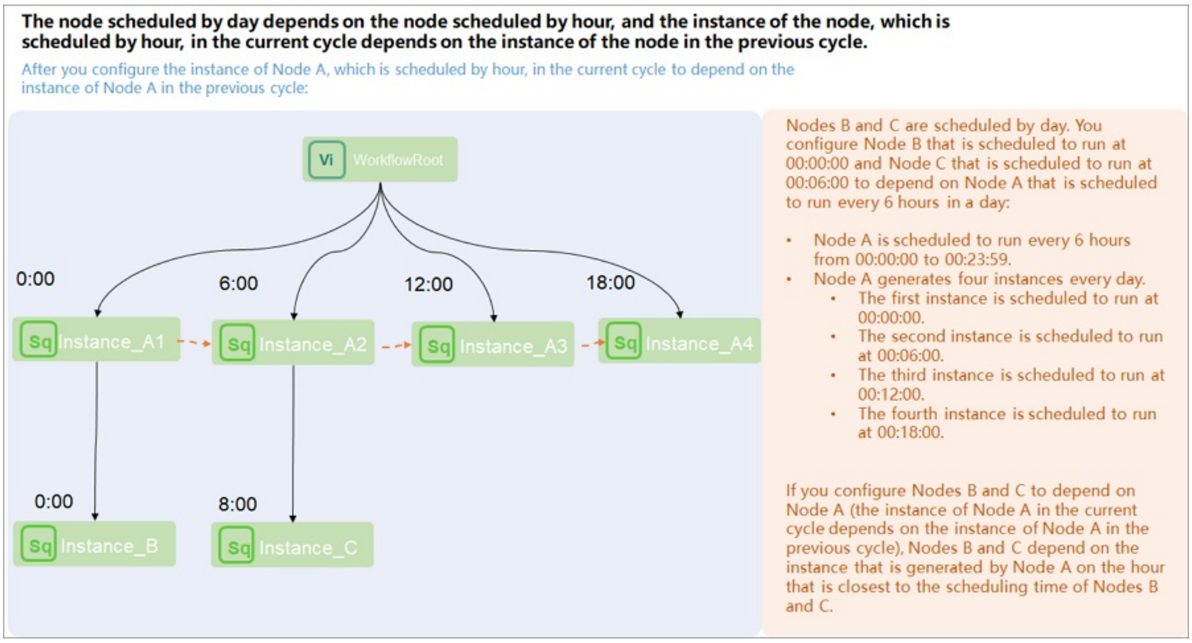
Nodes B and C are scheduled by day. You configure Node B that is scheduled to run at 00:00:00 and Node C that is scheduled to run at 00:06:00 to depend on Node A that is scheduled to run every 6 hours in a day:

- Node A is scheduled to run every 6 hours from 00:00:00 to 00:23:59.
- Node A generates four instances every day.
  - The first instance is scheduled to run at 00:00:00.
  - The second instance is scheduled to run at 00:06:00.
  - The third instance is scheduled to run at 00:12:00.
  - The fourth instance is scheduled to run at 00:18:00.

If you configure Nodes B and C to directly depend on Node A (the instance of Node A in the current cycle does not depend on the instance of Node A in the previous cycle), Nodes B and C depend on all instances that are generated on the current day for Node A scheduled by hour.

- Scenario 2: Configure a node scheduled by day to depend on a specific instance that is generated on the **current day** for a node scheduled by hour.

- For the node scheduled by hour, configure the instance of the node in the current cycle to depend on the instance of the node in the previous cycle. This indicates that you must set the Depend On parameter to Instances of Current Node on the Previous Cycle tab in the Dependencies section of the Properties tab in the DataWorks console.
- For the node scheduled by day, configure the node as the ancestor node of the node scheduled by day. This indicates that you must add the output name of the node scheduled by hour to Parent Nodes in the Dependencies section of the Properties panel for the node scheduled by day.



The node scheduled by day depends on the node scheduled by hour, and the instance of the node, which is scheduled by hour, in the current cycle depends on the instance of the node in the previous cycle.

After you configure the instance of Node A, which is scheduled by hour, in the current cycle to depend on the instance of Node A in the previous cycle:

Nodes B and C are scheduled by day. You configure Node B that is scheduled to run at 00:00:00 and Node C that is scheduled to run at 00:06:00 to depend on Node A that is scheduled to run every 6 hours in a day:

- Node A is scheduled to run every 6 hours from 00:00:00 to 00:23:59.
- Node A generates four instances every day.
  - The first instance is scheduled to run at 00:00:00.
  - The second instance is scheduled to run at 00:06:00.
  - The third instance is scheduled to run at 00:12:00.
  - The fourth instance is scheduled to run at 00:18:00.

If you configure Nodes B and C to depend on Node A (the instance of Node A in the current cycle depends on the instance of Node A in the previous cycle), Nodes B and C depend on the instance that is generated by Node A on the hour that is closest to the scheduling time of Nodes B and C.

- Scenario 3: Configure a node scheduled by day to depend on all instances that are generated on the **previous day** for a node scheduled by hour.
  - On the **Previous Cycle** tab in the Dependencies section of the Properties panel for the node scheduled by day, set the Depend On parameter to **Other Nodes** and enter the ID of the node scheduled by hour in the field that appears.
  - On the Same Cycle tab in the Dependencies section of the Properties panel for the node scheduled by day, remove the output name of the node scheduled by hour from **Parent Nodes** for the node scheduled by day.

Note If you configured a node scheduled by day to depend on a node scheduled by hour on the Previous Cycle tab, you must remove the output name of the node scheduled by hour from Parent Nodes for the node scheduled by day on the Same Cycle tab. Otherwise, the node scheduled by day depends on all instances that are generated on the previous day and the current day for the node scheduled by hour.

When does a node scheduled by day start to run if I configure a node scheduled by hour as the ancestor node of the node scheduled by day?

Principle: If a node scheduled by hour is configured as the ancestor node of a node scheduled by day, the node scheduled by day depends on all instances that are generated on the current day for the node scheduled by hour. This indicates that the node scheduled by day starts to run only after the last instance that is generated on the current day for the node scheduled by hour is successful.

Scenario:

- The node scheduled by hour starts to run at 00:00 and runs every hour. In this case, the node scheduled by day starts to run only after all the 24 instances of the node scheduled by hour are successful.
- View the dependencies of the node scheduled by day in Operation Center: Find the node scheduled by day in Operation Center, open the directed acyclic graph (DAG) of the node, right-click the node name in the DAG, and then select Show Ancestor Nodes to view all the 24 instances that are generated on the current day for the node scheduled by hour. The dependencies of the node scheduled by day in the DAG appear as solid lines.

How do I configure a node scheduled by day to depend on a specific instance that is generated on the current day for a node scheduled by hour?

Principle: If you want to configure a node scheduled by day to depend on a specific instance that is generated on the current day for a node scheduled by hour, you must configure the instance of the node scheduled by hour in the current cycle to depend on the instance of the node scheduled by hour in the previous cycle and set the scheduled time of the node scheduled by day to the scheduled time of a specific instance of the node scheduled by hour.

Scenario: Configure a node scheduled by day to depend on an instance that is generated on the current day for a node scheduled by hour and starts to run at 12:00.

- Dependency configuration:
  - For the node scheduled by hour: On the **Previous Cycle** tab in the **Dependencies** section of the **Properties** panel, set the Depend On parameter to **Instances of Current Node**.
  - For the node scheduled by day: Set the time when the node starts to run to 12:00.
- View dependencies in Operation Center:
  - Find the node scheduled by day in Operation Center, open the DAG of the node, right-click the node name in the DAG, and then select Show Ancestor Nodes to view the instance that is generated on the current day for the node scheduled by hour and starts to run at 12:00. The dependencies of the node scheduled by day in the DAG appear as solid lines.
  - Find the node scheduled by hour in Operation Center, open the DAG of the node, right-click the node name in the DAG, and then select Show Ancestor Nodes to view the instance that starts to run at 11:00. The instance that starts to run at 12:00 depends on the instance that starts to run at 11:00. The dependency of the node scheduled by hour appears as a dotted line. This is because the following configuration is performed for the node scheduled by hour: The instance of the node in the current cycle depends on the instance of the node in the previous cycle.

How do I configure a node scheduled by day to depend on all the instances that are generated on the previous day instead of the current day for a node scheduled by hour?

Principle: If you want to configure a node scheduled by day to depend on all the instances that are generated on the previous day for a node scheduled by hour, you must configure a cross-cycle dependency on the node scheduled by hour for the node scheduled by day.

Scenario: Configure a node scheduled by day to depend on all the instances that are generated on the previous day for the node scheduled by hour.

- Dependency configuration:

  - For the node scheduled by day: On the **Previous Cycle** tab in the **Dependencies** section of the **Properties** panel, set the Depend On parameter to **Other Nodes** and enter the ID of the node scheduled by hour in the field that appears.
  - For the node scheduled by hour: You do not need to configure dependencies.

- View dependencies in Operation Center:

  Find the node scheduled by day in Operation Center, open the DAG of the node, right-click the node name in the DAG, and then select Show Ancestor Nodes to view all the instances that are generated on the previous day for the node scheduled by hour. The dependencies of the node scheduled by day appear as dotted lines. This is because that this node is configured with a cross-cycle dependency on the node scheduled by hour.

In which scenarios do I need to configure the instance of a node in the current cycle to depend on the instance of the node in the previous cycle?

Scenario: If a node needs to use data that is generated by the node itself in the previous cycle, you can configure the node to depend on its own instance in the previous cycle. In this case, the instance of the node in the current cycle runs only after the instance in the previous cycle is successful. This ensures that the instance in the current cycle can obtain data from the instance in the previous cycle.

- A node needs to use data generated by the node itself in the previous cycle. For this node, you must set the Depend On parameter to **Instances of Current Node** on the **Previous Cycle** tab in the **Dependencies** section of the **Properties** panel for the node.
- A node scheduled by hour depends on a node scheduled by day. After the instance that is generated on a day for the node scheduled by day is successful, the execution time of all the instances that are generated on this day for the node scheduled by hour arrives. As a result, all the instances of the node scheduled by hour are concurrently run. In this case, set the Depend On parameter to **Instances of Current Node** on the **Previous Cycle** tab in the **Dependencies** section of the **Properties** panel for the node scheduled by hour.

How do I configure dependencies for a node that needs to depend on multiple nodes?

If a node needs to depend on multiple nodes, you must determine whether to configure dependencies between the node and these nodes. If the node strongly depends on the table data generated by these nodes, we recommend that you configure dependencies between the node and these nodes. For more information about how to determine whether to configure dependencies between nodes, see Why are scheduling dependencies required?.

For example, Node A is scheduled by hour and generates Table A, and Node B is scheduled by day and generates Table B. Node C depends on Node A and Node B and needs to use data in Table A and Table B.

If you add the output name of Node A to **Parent Nodes** for Node C, but do not add the output name of Node B to **Parent Nodes** for Node C, Node C may start to run even if Node B is still running. As a result, Node C fails to obtain data in Table B, and an error occurs on Node C. To resolve this issue, you must add the output names of both Node A and Node B to **Parent Nodes** for Node C.

If a node does not strongly depend on the table data generated by another node and the node can obtain the data even if the latest data is not generated by another node, you do not need to configure a dependency between the two nodes.

Node B scheduled by day depends on Node A scheduled by hour, and Node B starts to run only after all the instances that are generated on the current day for Node A are successful. Will the execution of Node B be affected if Node A still runs on the next day?

Node B depends on all the instances that are generated on the current day for Node A. Node B automatically runs every day after all the instances of Node A are successful. If the last instance of Node A is successful on the next day, Node B still runs, but at a time that is different from the specified time. Scheduling parameters can be replaced as expected.

Node A runs every hour on the hour, and Node B runs once every day. How do I configure Node B to automatically run after the first instance of Node A is run every day?

When you configure time properties for Node A in the Dependencies section of the Properties panel, click the **Previous Cycle** tab and set the Depend On parameter to **Instances of Current Node**. You must set the Run At parameter to 00:00 for Node B in the Schedule section of the Properties panel. This way, Node B depends only on the first instance that is generated every day for Node A. The first instance is generated at 00:00.

How do I configure Node A, Node B, and Node C to run in sequence once per hour?

1. Dependencies: Configure the output of Node A as the input of Node B and the output of Node B as the input of Node C.
2. Scheduling cycle: Configure Node A, Node B, and Node C to be scheduled by hour.

How do I configure dependencies between nodes that reside in the same region but belong to different workspaces and workflows?

Principle: Use the output of a node as the input of another node to establish a dependency between the two nodes.

Add the output name of a node to Parent Nodes for another node to establish a dependency between the two nodes. The two nodes can belong to different workspaces and workflows.

I have configured rerun properties for my node, but the node does not rerun after it fails. In addition, the error message "Task Run Timed Out, Killed by System!!!" appears. What do I do?

- Problem description: The **Rerun** parameter in the **Schedule** section of the **Properties** panel is set to **Allow Regardless of Running Status** or **Allow upon Failure Only** for the node. However, the node does not rerun after it fails, and the error message `Task Run Timed Out, Killed by System!!!` appears.
- Cause: The **Timeout Definition** parameter is configured in the **Schedule** section of the **Properties** panel. If the execution duration of the node exceeds the value of the Timeout Definition parameter, the node automatically stops and does not rerun. A node that fails to be run due to a timeout cannot be rerun.
- Solution: Manually enable the node to rerun in this scenario.

3.Operation Center

3.1. Overview

Data backfill

- Feature of generating retroactive data for nodes
- Why do the retroactive instances of a node that is scheduled by hour or minute not run in parallel after I enable the parallelism feature for the node?
- The retroactive instances of a node are not run after I specify the data timestamp for retroactive data generation. The retroactive instances are in the Pending (Schedule) state and are highlighted in yellow in the DAG. Why does this happen?
- Why is a retroactive instance of an auto triggered node in the Pending (Schedule) state after I specify the last day and the current day for the Data Timestamp parameter?
- Why are multiple retroactive instances generated for a node if I set the data timestamp to 00:00:00 to 01:00:00?
- If a large number of retroactive instances are generated for a node, the retroactive instances are in the Pending (Resources) state and are highlighted in yellow in the DAG. Why does this happen?
- Why do I receive the error message which indicates that the scheduled runtime of a node is not within the specified data timestamp range?
- Why cannot retroactive instances be generated for a node after I enable retroactive data generation for the node?

Wait for resources

- Why does a node wait for resources?
- Why does a node wait for gateway resources for an extended period of time?
- Why does a data synchronization node wait for resources for an extended period of time?

Dry-run instances

- What is a dry-run instance?
- Why does a dry-run instance exist?
    i. Scenario 1: An instance is scheduled to run on a specific day every week or every month
    ii. Scenario 2: An instance is generated in real time but is deprecated
    iii. Scenario 3: The status of an instance is set to successful
    iv. Scenario 4: The property of an instance is dry run
    v. Scenario 5: An instance is not selected for a temporary workflow

  Troubleshoot dry runs for nodes that are scheduled on a daily basis

Nodes that are not run

- What are the conditions that are required for a node to successfully run?
- Why is an auto triggered instance not run after its scheduling time arrives?

Node instance status

- What do I do if I cannot find the desired auto triggered node on the Cycle Task page in Operation Center?
- Why cannot I find even one instance of an auto triggered node?
- What do I do if I can find instances of other auto triggered nodes but cannot find even one instance of the desired auto triggered node?
- What do I do if the desired auto triggered node has instances but the instances are not run?
- What are the conditions that must be met to run a node?

Nodes that are successfully run but have no data generated

- Scenario 1: An auto triggered node is successfully run and has operational logs
- Scenario 2: An auto triggered node is successfully run but has no operational logs

Nodes that fail to be run

- Errors for nodes that fail to be rerun
  - I have configured rerun properties for my node, but the node does not rerun after it fails, and the following error message appears: Task Run Timed Out, Killed by System!!!. What do I do?
  - I set the Auto Rerun Times upon Error parameter to 1 for my node, but the node does not rerun after it fails. What do I do?
- Errors for MaxCompute nodes
  - What do I do if the error message ODPS-0420095: Access Denied - Authorization Failed [4093], You have NO privilege to do the restricted operation on {acs:odps:*:projects/xxxx}. Access Mode is AllDenied. appears?
  - What do I do if the error message ODPS-0420061: Invalid parameter in HTTP request - Fetched data is larger than the rendering limitation. Please try to reduce your limit size or column number appears?
  - What do I do if the data synchronized by using multiple threads is out of order?
- Errors for AnalyticDB for MySQL nodes

  What do I do if my synchronization node that uses an AnalyticDB for MySQL data source and runs on a shared resource group fails to run?
- Errors for database nodes

  When I run a data synchronization node that uses a MySQL data source, the system displays an error message indicating that the Java Database Connectivity (JDBC) driver for the MySQL node is not supported. What do I do?
- Errors for general nodes
  - How do I view the logs of a for-each node, do-while node, or PAI node in Operation Center?
  - What do I do if the error message error in your condition run fail appears when I run a branch node?
  - What do I do if the error message None Ftp connection info!! appears when I run an FTP Check node?
  - What do I do if the error message Connect Failed appears when I run an FTP Check node?
  - What do I do if the error message The current time has exceeded the end-check time point! appears when I run an FTP Check node?
  - What do I do if the error message File not Exists or exceeded the end-check time point! appears when I run an FTP Check node?

- Errors for resource groups

  What do I do if the error message no available machine resources under the task resource group appears for my resource group for scheduling?

Node undeployment

- How do I undeploy a node?
- What do I do if a node that has subnodes fails to be undeployed?
- How do I check whether a node is undeployed from the production environment?
- How do I recover an undeployed node?

Isolated nodes

- What is an isolated node?
- How can I fix an isolated node?

Others

- Resource groups
  - How long are the logs of resource groups for scheduling and node instances that are run on such resource groups retained?
  - Why am I unable to perform big data computing on a resource group for scheduling?
- Others

  How do I adjust the priority of a node instance?

3.2. Nodes that are waiting for resources

This topic provides answers to some frequently asked questions about nodes that are waiting for resources.

- Why does a node wait for resources?
- Why does a node wait for gateway resources for an extended period of time?
- Why does a data synchronization node wait for resources for an extended period of time?

Why does a node wait for resources?

- Problem description

  After the scheduling system commits a node to a compute engine, the node may wait for resources due to the following reasons:

  - The node waits for a resource group for scheduling.
  - If the node is a data synchronization node in Data Integration, the node waits for a resource group for Data Integration.
  - If the node is a computing node, the node waits for computing resources.
- Cause

The resources in a resource group are limited. If the resources are occupied by a node for an extended period of time, other nodes cannot be run until the occupied resources are released. For more information about how the scheduling system commits nodes, see Overview.

Why does a node wait for gateway resources for an extended period of time?

- Problem description

  The log of a node shows that the node is waiting for gateway resources.

- Cause

  The number of nodes that are running in parallel on the related resource group for scheduling reaches the upper limit.

- Solution

  Wait until the nodes that are running release the occupied resources or scale out your exclusive resource group for scheduling. You can use one of the following methods to view resource usage in the resource group:

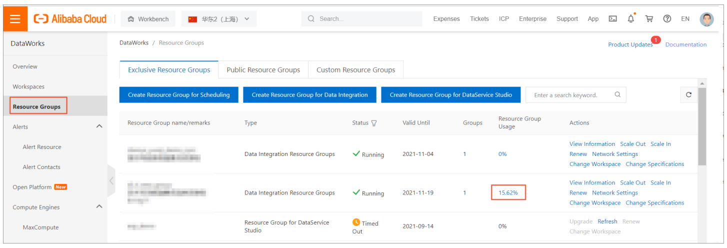
  - If you use the shared resource group for scheduling, you can view the resource usage in the resource group on the **Overview** page of **Operation Center**.

    > Note The peak hours for DataWorks nodes are from 00:00 to 09:00 every day. During this period, resources in the shared resource group for scheduling may be insufficient, and nodes may wait for resources. You can change the scheduling time of the node or purchase an exclusive resource group for scheduling in the DataWorks console.

  - If you use an exclusive resource group for scheduling, you can log on to the DataWorks console, go to the Exclusive Resource Groups tab of the Resource Groups page or go to the Resource page of **Operation Center**, and then view the nodes that are running on the exclusive resource group for scheduling and the resource usage in the resource group.

    - View resource usage in the DataWorks console.

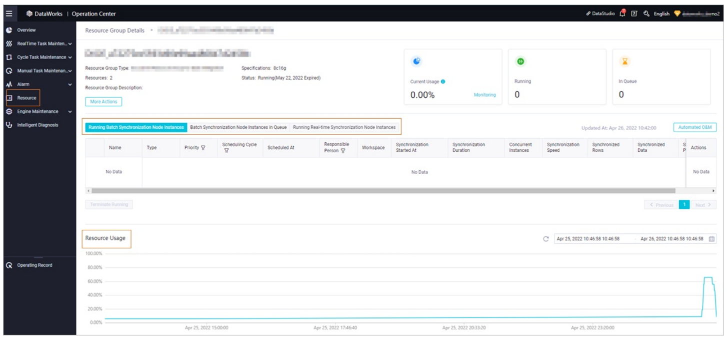
      Log on to the DataWorks console and click **Resource Groups** in the left-side navigation pane. On the **Exclusive Resource Groups** tab of the **Resource Groups** page, find the exclusive resource group and view the resource usage in the **Resource Group Usage** column. Then, you can click the percentage value in the **Resource Group Usage** column, and view the details of the resource group and the resource usage in the resource group.



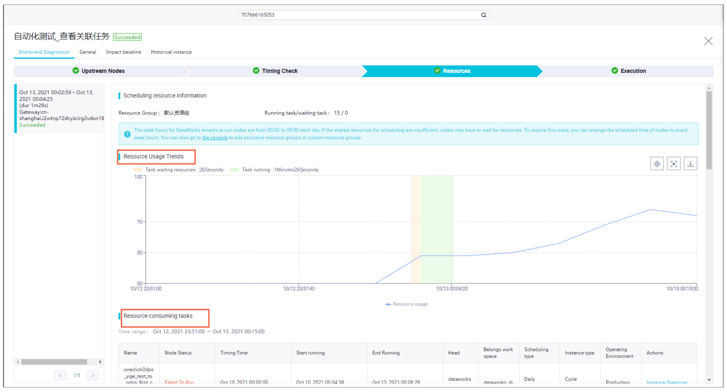
Resource Group name/remarks	Type	Status	Valid Until	Groups	Resource Group Usage	Actions
[redacted]	Data Integration Resource Groups	✓ Running	2021-11-04	1	0%	View Information Scale Out Scale In Renew Network Settings Change Workspace Change Specifications
[redacted]	Data Integration Resource Groups	✓ Running	2021-11-19	1	15.62%	View Information Scale Out Scale In Renew Network Settings Change Workspace Change Specifications
[redacted]	Resource Group for DataService Studio	Timed Out	2021-09-14		0%	Upgrade Refresh Renew Change Workspace

    - View resource usage in Operation Center.

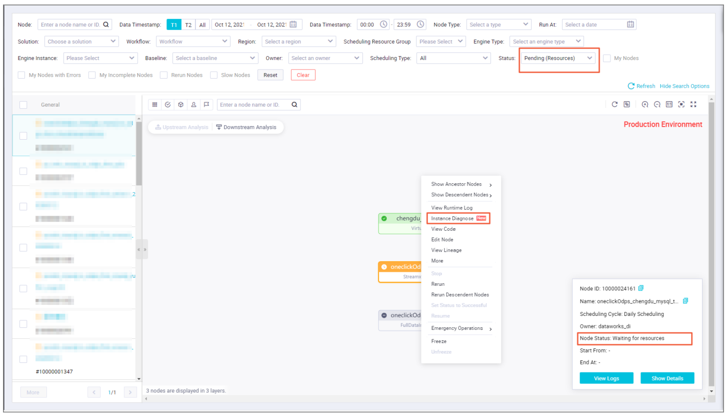
Go to **Operation Center**, click **Resource** in the left-side navigation pane, and then view the information about resource groups, such as the resource usage.



On the **Intelligent Diagnosis** page, you can view the information about nodes that are running and waiting in a queue. You can also view the nodes that are occupying resources when the current node is waiting for resources.



To view the nodes that are running when the current node is waiting for resources, perform the following steps: On the **Overview** page of Operation Center, click **Instance waiting for resource**. On the page that appears, find the current node, right-click the node in the directed acyclic graph (DAG), and then select **Instance Diagnose**. On the page that appears, click the **Resources** tab and view the nodes that are displayed in the **Resource-consuming tasks** section.



Note The maximum number of nodes that can be run in parallel on an exclusive resource group for scheduling varies based on the specifications of the resource group. For more information, see Billing of exclusive resource groups for scheduling (subscription).

Why does a data synchronization node wait for resources for an extended period of time?

- Problem description

  The log of a data integration node shows that the node is waiting for resources.

- Cause

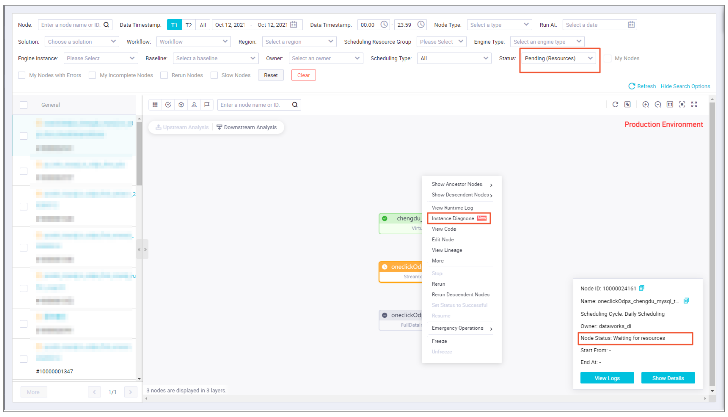
  This issue occurs because the number of nodes that are running in parallel on the current resource group for Data Integration exceeds the upper limit for the resource group. As a result, the node keeps waiting for resources.

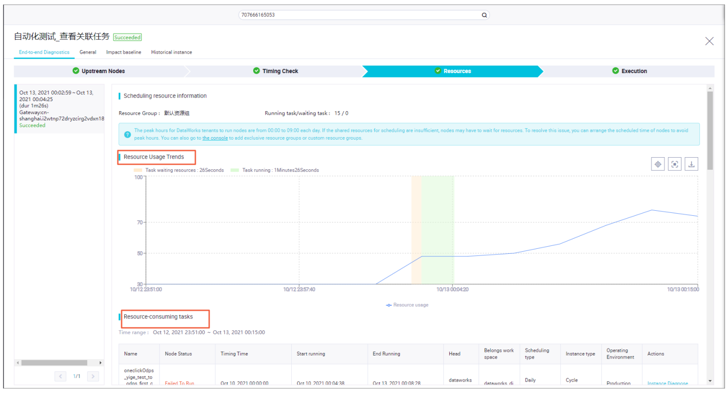
- Solution

  - Check whether the value of the concurrent parameter that is specified when you configure the node is excessively large. If the value is too large, you must set the parameter based on the upper limit for the resource group for Data Integration and the number of nodes that are running in parallel.

- You may set the concurrent parameter of the node to a value that exceeds the upper limit for the resource group for Data Integration. In this case, stop the node and change the value of the concurrent parameter before you run the node again.

  Go to **Operation Center**. The **Overview** page appears. On the **Overview** page, click the **Data Integration** tab and view the details of data synchronization nodes in the **Synchronization task execution details** section.

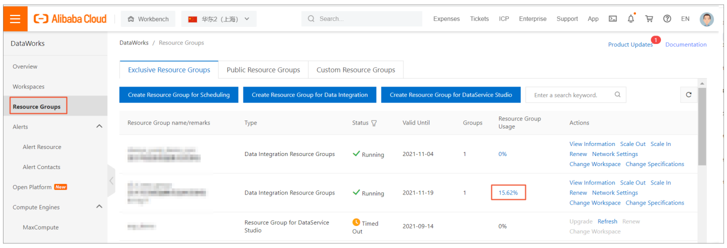




> **Note**
>
> - You must set the concurrent parameter for a node based on the maximum number of nodes that can be run in parallel on an exclusive resource group for Data Integration. In addition, you must also consider the number of nodes that you want to run in parallel and the sum of the values of the concurrent parameter for each node.
> - A data synchronization node occupies the resources of a resource group for Data Integration. If resources are occupied by the node for an extended period of time, other nodes to be run on the resource group cannot be run.

- Wait until the nodes that are running release the occupied resources or scale out your exclusive resource group for Data Integration. You can use one of the following methods to view resource usage in the resource group:
  - View resource usage in the DataWorks console.

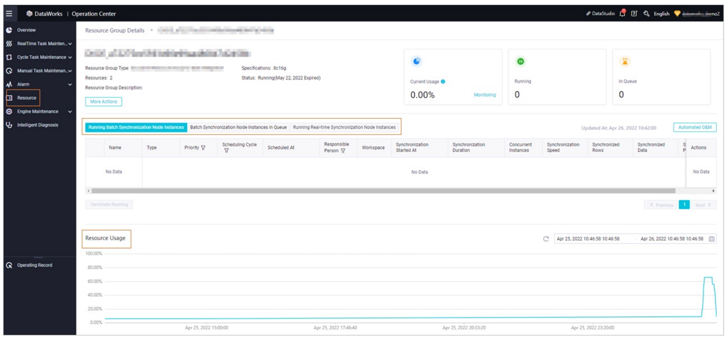
    Log on to the DataWorks console and click **Resource Groups** in the left-side navigation pane. On the **Exclusive Resource Groups** tab of the **Resource Groups** page, find the exclusive resource group and view the resource usage in the **Resource Group Usage** column. Then, you can click the percentage value in the **Resource Group Usage** column, and view the details of the resource group and the resource usage in the resource group.



Resource Group name/remarks	Type	Status	Valid Until	Groups	Resource Group Usage	Actions
[redacted]	Data Integration Resource Groups	✓ Running	2021-11-04	1	0%	View Information Scale Out Scale In Renew Network Settings Change Workspace Change Specifications
[redacted]	Data Integration Resource Groups	✓ Running	2021-11-19	1	15.62%	View Information Scale Out Scale In Renew Network Settings Change Workspace Change Specifications
[redacted]	Resource Group for DataService Studio	Timed Out	2021-09-14		0%	Upgrade Refresh Renew Change Workspace

  - View resource usage in Operation Center.

    Go to **Operation Center**. Click **Resource** in the left-side navigation pane and view the information about resource groups, such as the resource usage and the nodes that are running on the resource groups.



Resource Group Type: [redacted]

Resources: 2

Resource Group Description:

Specifications: 8c16g

Status: Running(May 22, 2022 Expired)

Current Usage 0.00%

Running 0

In Queue 0

Running Batch Synchronization Node Instances | Batch Synchronization Node Instances in Queue | Running Real-time Synchronization Node Instances

Updated At: Apr 26, 2022 10:42:00

Name	Type	Priority	Scheduling Cycle	Scheduled At	Responsible Person	Workspace	Synchronization Started At	Synchronization Duration	Concurrent Instances	Synchronization Speed	Synchronized Rows	Synchronized Data	Actions
No Data	No Data												No Data

Apr 25, 2022 10:46:58 10:46:58 - Apr 26, 2022 10:46:58 10:46:58

> **Note** The maximum number of nodes that can be run in parallel on an exclusive resource group for Data Integration varies based on the specifications of the resource group. For more information, see Billing of exclusive resource groups for Data Integration (subscription).

3.3. Nodes that are not run

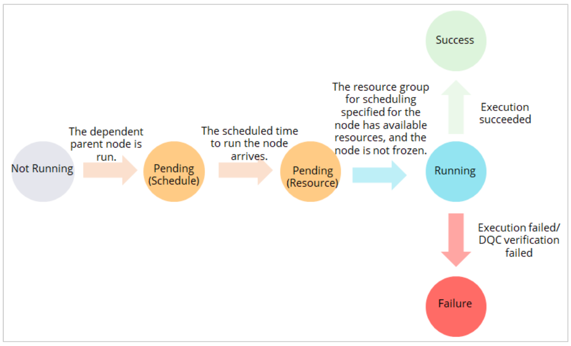
This topic provides answers to some frequently asked questions about nodes that are not run.

- What are the conditions that are required for a node to successfully run?
- Why is an auto triggered instance not run after its scheduling time arrives?

What are the conditions that are required for a node to successfully run?

In Operation Center, auto triggered instances that are in different states are marked in different colors.

To ensure that a node can be scheduled as expected, the following conditions must be met: The scheduling resources for the node are sufficient. The scheduling time of the node has arrived. The ancestor nodes of the node are successfully run. The node is not frozen.

No.	Status	Icon	Flowchart
1	Successful	✓	 Success The resource group for scheduling specified for the node has available resources, and the node is not frozen. Execution succeeded The dependent parent node is run. The scheduled time to run the node arrives. Not Running Pending (Schedule) Pending (Resource) Running Execution failed/ DQC verification failed Failure
2	Not running	⊖	
3	Failed	⊗	
4	Running	➡	
5	Waiting	🕒	
6	Suspended or frozen	✱	

- If an auto triggered instance is marked in purple, the instance is frozen. The node that generates the instance is not run and descendant nodes of the node are blocked from running. To view operation records of the instance, go to the Cycle Instance page in Operation Center, click **Show Details** in the directed acyclic graph (DAG) of the instance, and then click **Operation Log** tab.
- If an auto triggered instance is marked in yellow, the scheduling time of the instance has not arrived. To view the information about the instance, go to the Cycle Instance page in Operation Center, click **Show Details** in the DAG of the instance, and then click **General** tab.

> Note
>
> - If the instance is in the **Waiting for resources** state, the number of nodes that are run on the current resource group in the current workspace has reached the upper limit. You can right-click the instance in the DAG on the Cycle Instance page, and select Instance Diagnose. In the Resources step, you can view the nodes that occupy the resources in the current resource group. For more information about resource waiting, see Nodes that are waiting for resources.
> - If the instance is in the **Waiting time** state, the scheduling time of the instance has not arrived.

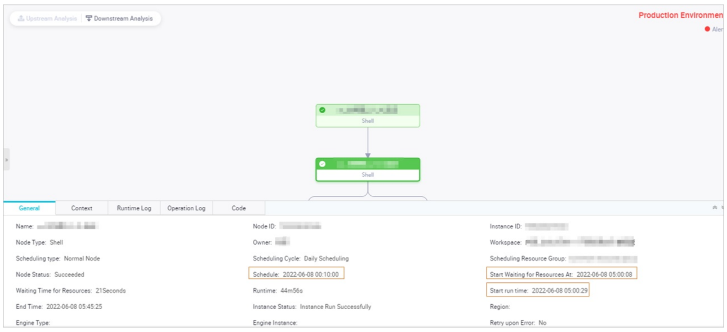
- If an auto triggered instance is marked in gray, the instance is not run as expected. You can right-click an instance that is marked in gray in the DAG on the Cycle Instance page and select Show Ancestor Nodes to view the status of the ancestor nodes of the node that generates the instance. We recommend that you use the **intelligent diagnosis** and **ancestor node analysis** features to view the status of the ancestor nodes of the node that generates the instance.
- If an auto triggered instance is marked in blue, the instance is running. If the instance is in the running state for a long period of time, the scheduling resources may be insufficient. For more information about why an instance waits for resources, see Nodes that are waiting for resources.

> Note If the ancestor nodes of the node that generates the instance are not in the preceding states and all nodes in the workflow to which the node belongs are marked in gray, the dependency between the node and its ancestor nodes is changed and the workflow is isolated. For more information, see Isolated nodes.

Why is an auto triggered instance not run after its scheduling time arrives?

- Problem description

  For an auto triggered instance, the scheduling time and the time at which the instance starts to run may be different. For example, the instance may not start to run after its scheduling time arrives.



Name: [redacted]
Node Type: Shell
Scheduling type: Normal Node
Node Status: Succeeded
Waiting Time for Resources: 21Seconds
End Time: 2022-06-08 05:45:25
Engine Type:

Node ID: [redacted]
Owner: [redacted]
Scheduling Cycle: Daily Scheduling
Schedule: 2022-06-08 00:10:00
Runtime: 44m56s
Instance Status: Instance Run Successfully
Engine Instance:

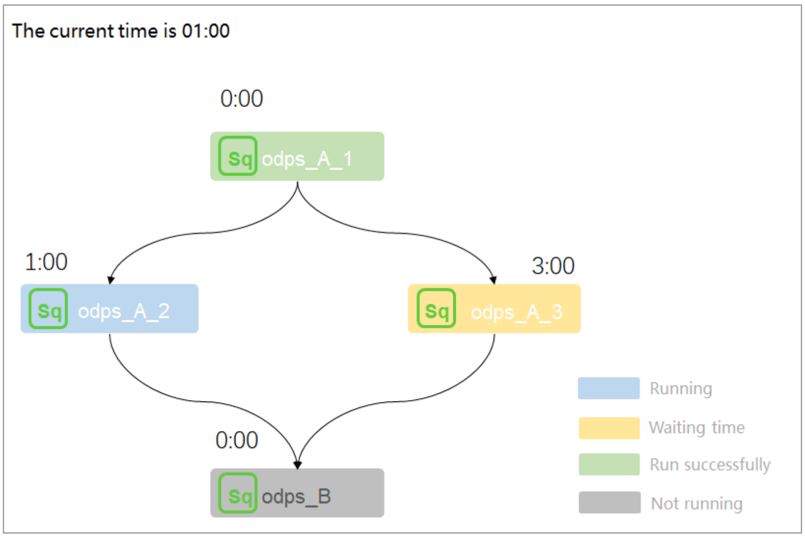
Instance ID: [redacted]
Workspace: [redacted]
Scheduling Resource Group: [redacted]
Start Waiting for Resources At: 2022-06-08 05:00:08
Start run time: 2022-06-08 05:00:29
Region:
Retry upon Error: No

- Troubleshooting

  You can use the ancestor node analysis feature to identify the instance that blocks the current instance from running. Then, you can use the intelligent diagnosis feature to quickly identify the cause of the issue.

  The following scenarios show the reasons why an auto triggered instance in DataWorks is not run as expected. In the following figures, a three-layer dependency is used. A dependency of more than three layers may exist in actual scenarios. The dependency logic is the same regardless of the number of layers.

- Scenario 1: The ancestor nodes of the auto triggered node that generates the current instance are not successfully run. The instance is in the **Not running** state and is marked in gray.



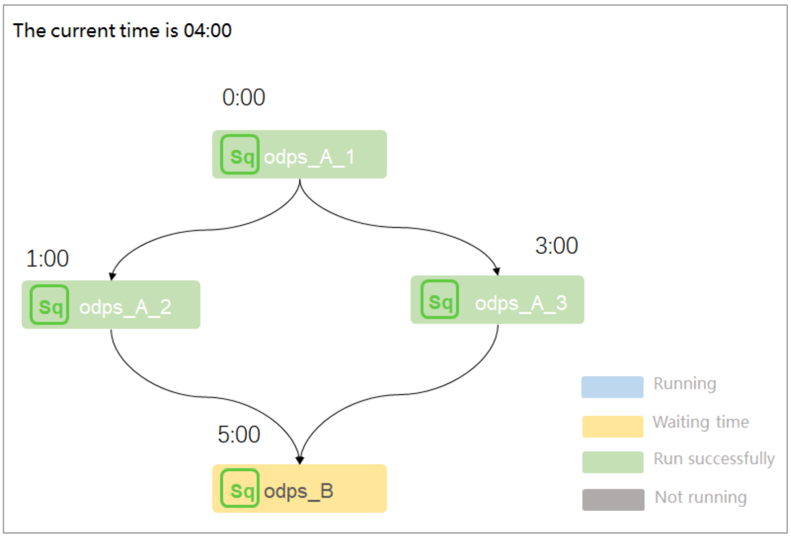
Example: The current time is 01:00. Instance A2 is scheduled to run at 01:00, Instance A3 is scheduled to run at 03:00, and Instance B is scheduled to run at 00:00. Instance B depends on Instance A2 and Instance A3.

Analysis: The scheduling time of Instance B is 00:00 and has arrived. However, the scheduling time of Instance A2 and Instance A3 has not arrived. In this case, Instance B can start to run only after Instance A2 and Instance A3 are successfully run. If Instance A2 or Instance A3 fails to be run, Instance B cannot be run as expected. In this case, Instance B enters the **Not running** state.

Conclusion: The current instance can start to run only after its ancestor instances are successfully run.

> **Note** If the ancestor instances of the current instance are in the running state for a long period of time, you can use one of the following methods to fix the issue:
>
> - If the ancestor instances are generated by non-batch synchronization nodes, you can submit a ticket to the technical support team of the compute engine that is used to process data in the non-batch synchronization nodes.
> - If the ancestor instances are generated by batch synchronization nodes, one possible cause is that the ancestor instances are in the state of waiting for resources for a long period of time. Another possible cause is that the speed at which the logic of some code is processed is slow during the node running. For more information, see How to troubleshoot the issue that the execution duration of a batch synchronization node is long?

- Scenario 2: The scheduling time of an auto triggered node that generates the current instance has not arrived. The instance is in the **Waiting time** state and is marked in yellow.

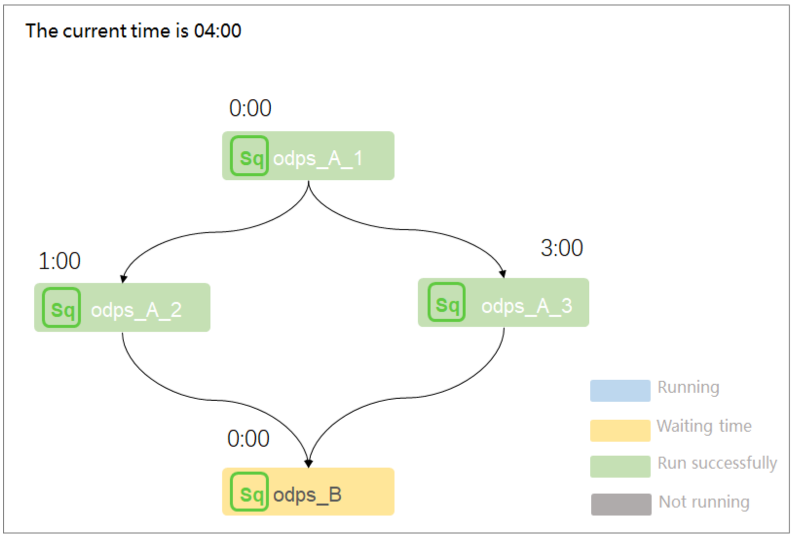


Example: The current time is 04:00. Instance A2 is scheduled to run at 01:00, Instance A3 is scheduled to run at 03:00, and Instance B is scheduled to run at 05:00. Instance B depends on Instance A2 and Instance A3.

Analysis: Instance A2 and Instance A3 are successfully run before the scheduling time of Instance B arrives. In this case, Instance B is in the **Waiting time** state before its scheduling time arrives.

Conclusion: The current instance can be run after its scheduling time arrives.

- Scenario 3: Scheduling resources in the current workspace are not sufficient. Therefore, the current auto triggered node instance cannot be run. The instance is in the **Waiting for resources** state and is marked in yellow.



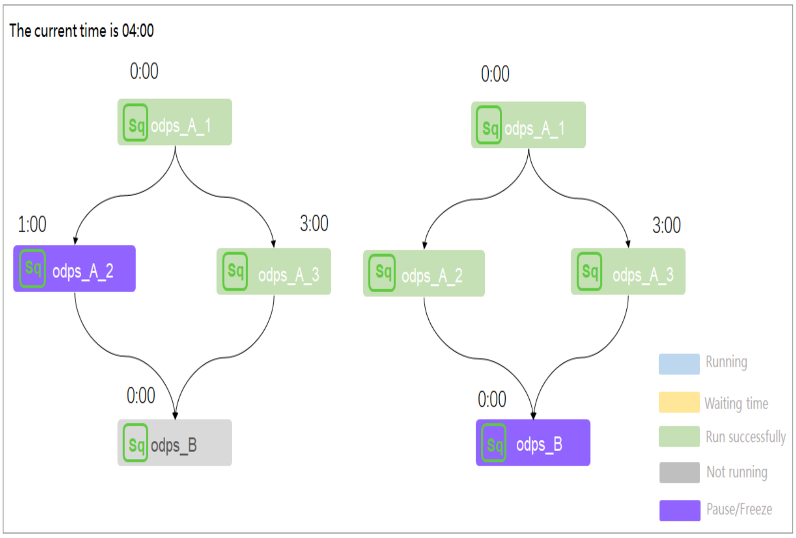
Example: The current time is 04:00. Instance A2 is scheduled to run at 01:00, Instance A3 is scheduled to run at 03:00, and Instance B is scheduled to run at 00:00. Instance B depends on instances A2 and A3.

Analysis: Instance A2 and Instance A3 are successfully run before the scheduling time of Instance B arrives. However, the scheduling resources in the resource group for scheduling that is used to run Instance B are insufficient. As a result, the status of Instance B is **Waiting for resources**.

Conclusion: The current instance can be run only when the scheduling resources in the current workspace are sufficient. If an instance is waiting for resources, the log information shows that the number of nodes that are run on the current resource group in the current workspace has reached the upper limit, and the instance is waiting for gateway resources.

Note If Instance B is run on an exclusive resource group for scheduling and you want to view the resource usage of the resource group, you can use one of the following methods: 1. Log on to the DataWorks console, and click Resource Groups in the left-side navigation pane. On the Resource Groups page, view the nodes that are run on the exclusive resource group for scheduling and view the usage of the resource group. 2. Right-click Instance B in the DAG and select Instance Diagnose to check the nodes that occupy resources in the exclusive resource group for scheduling. For more information, see Nodes that are waiting for resources.

- Scenario 4: The current instance is frozen. The instance is in the Frozen state and is marked in purple.



Example: The current time is 04:00. Instance A2 is scheduled to run at 01:00 and is in the suspended state, Instance A3 is scheduled to run at 03:00, and Instance B is scheduled to run at 00:00. Instance B depends on instances A2 and A3.

Analysis: Scenario 1 shows that an instance can start to run only after all of its ancestor instances are successfully run. The DAG on the left in the preceding figure shows that Instance A2 is frozen. As a result, Instance B cannot be run. The DAG on the right in the preceding figure shows that all the ancestor instances of Instance B are successfully run. However, Instance B cannot be run as scheduled because Instance B is frozen.

Conclusion: If the current instance is frozen or the ancestor instances of the current instance are frozen, the current instance cannot be run as scheduled.

3.4. Node instance status

This topic provides answers to some frequently asked questions about node instances.

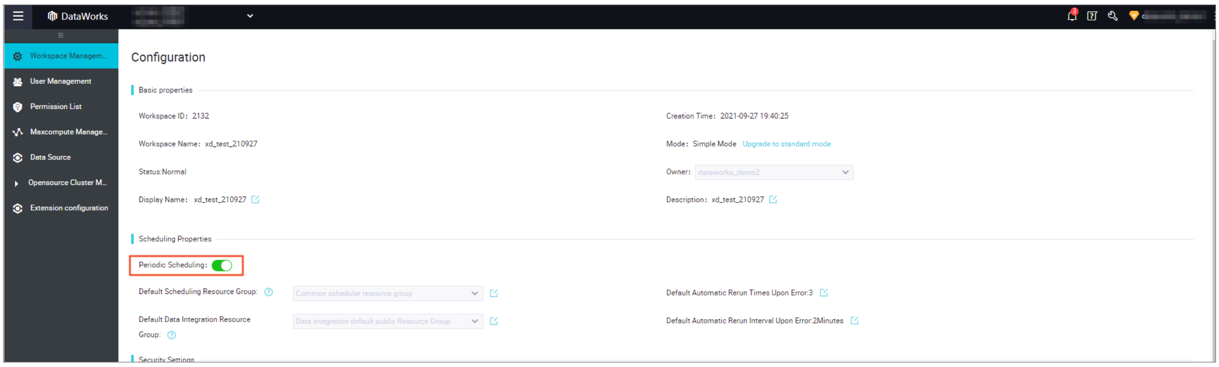
- What do I do if I cannot find the desired auto triggered node on the Cycle Task page in Operation Center?
- Why cannot I find even one instance of an auto triggered node?
- What do I do if I can find instances of other auto triggered nodes but cannot find even one instance of the desired auto triggered node?
- What do I do if the desired auto triggered node has instances but the instances are not run?
- What are the conditions that must be met to run a node?

What do I do if I cannot find the desired auto triggered node on the Cycle Task page in Operation Center?

The auto triggered node is not deployed to the production environment. Check whether the deployment fails.

Why cannot I find even one instance of an auto triggered node?

[Troubleshooting] Check whether the Periodic Scheduling switch is turned on in the Scheduling Properties section on the Configuration page. If you want to turn on the switch, use the Alibaba Cloud account to go to the Configuration page and turn on the Periodic Scheduling switch in the Scheduling Properties section.



Configuration

Basic properties

Workspace ID: 2132

Creation Time: 2021-09-27 19:40:25

Workspace Name: xd_test_210927

Mode: Simple Mode Upgrade to standard mode

Status:Normal

Owner: dataworks_demo2

Display Name: xd_test_210927

Description: xd_test_210927

Scheduling Properties

Periodic Scheduling: ●

Default Scheduling Resource Group: Common scheduler resource group

Default Automatic Rerun Times Upon Error:3

Default Data Integration Resource Group: Data integration default public Resource Group

Default Automatic Rerun Interval Upon Error:2Minutes

Security Settings

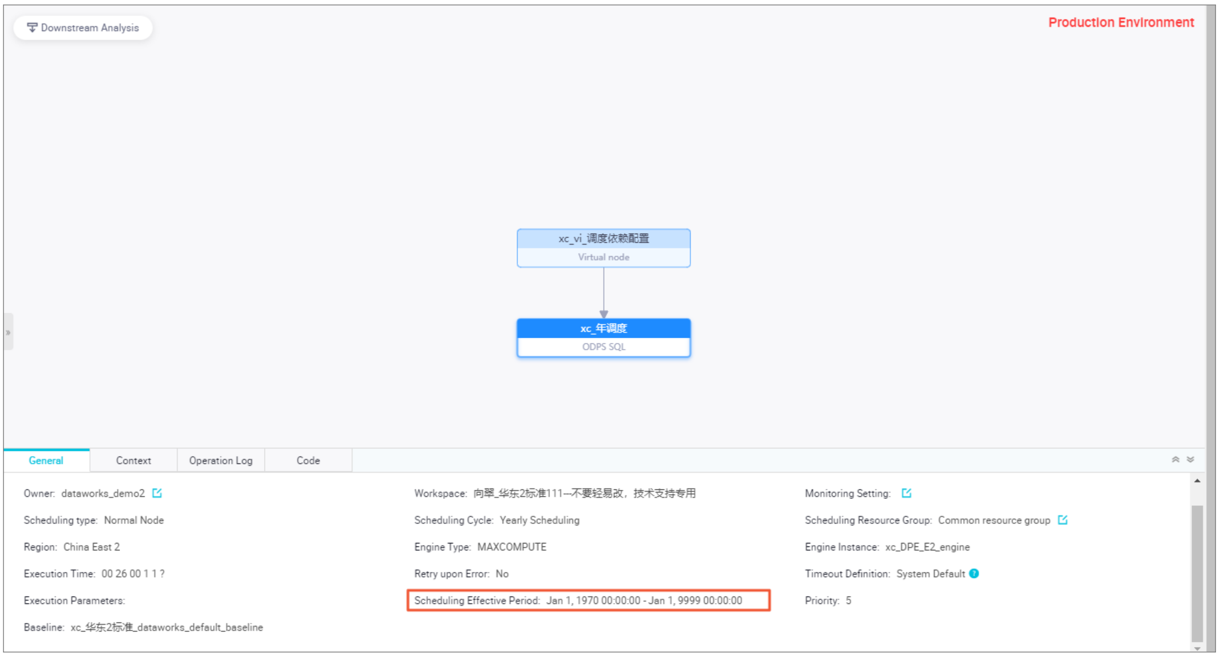
What do I do if I can find instances of other auto triggered nodes but cannot find even one instance of the desired auto triggered node?

[Troubleshooting] Scenario 1: Check whether the desired auto triggered node is deployed after 23:30.

[Troubleshooting] Scenario 2: Right-click the **desired auto triggered node** in the canvas and select Show Ancestor Nodes to check whether the auto triggered node has ancestor nodes. If the auto triggered node has no ancestor nodes and becomes an isolated node, after you receive a node isolation alert, you must handle the alert at the earliest opportunity.

[Solution] The node dependency is changed. The auto triggered node does not have ancestor nodes. Submit the dependency of the auto triggered node again.

[Troubleshooting] Scenario 3: Check whether the auto triggered node and its ancestor nodes are within the validity period. A deprecated auto triggered node does not generate instances. On the Properties tab of the DataStudio page, change the value of the Validity Period parameter in the Schedule section.



Owner: dataworks_demo2

Workspace: 向翠_华东2标准111--不要轻易改，技术支持专用

Monitoring Setting:

Scheduling type: Normal Node

Scheduling Cycle: Yearly Scheduling

Scheduling Resource Group: Common resource group

Region: China East 2

Engine Type: MAXCOMPUTE

Engine Instance: xc_DPE_E2_engine

Execution Time: 00 26 00 1 1 ?

Retry upon Error: No

Timeout Definition: System Default

Execution Parameters:

Scheduling Effective Period: Jan 1, 1970 00:00:00 - Jan 1, 9999 00:00:00

Priority: 5

Baseline: xc_华东2标准_dataworks_default_baseline

What do I do if the desired auto triggered node has instances but the instances are not run?

[Troubleshooting] Right-click **a dimmed instance of the desired auto triggered node** and select Show Ancestor Nodes to check whether the ancestor nodes are in the state of running, run failed, pending, or unfrozen. The auto triggered node is not run if its ancestor nodes fail to be run.

> **Note** For more information, see Nodes that are not run.

[Ancestor node status]

1. Frozen (purple)

   An instance that is frozen is marked in purple. In this case, the auto triggered node and its descendant nodes are not run. Click the auto triggered node. In the window that appears in the lower-right corner of the page, click Show Details, and view the operation log of the node on the Operation Log tab.

2. Pending (yellow)

   A node that is waiting for its scheduled time to arrive is marked in yellow. Click the node. In the window that appears in the lower-right corner of the page, click Show Details, and view the scheduled time of the node on the General tab.

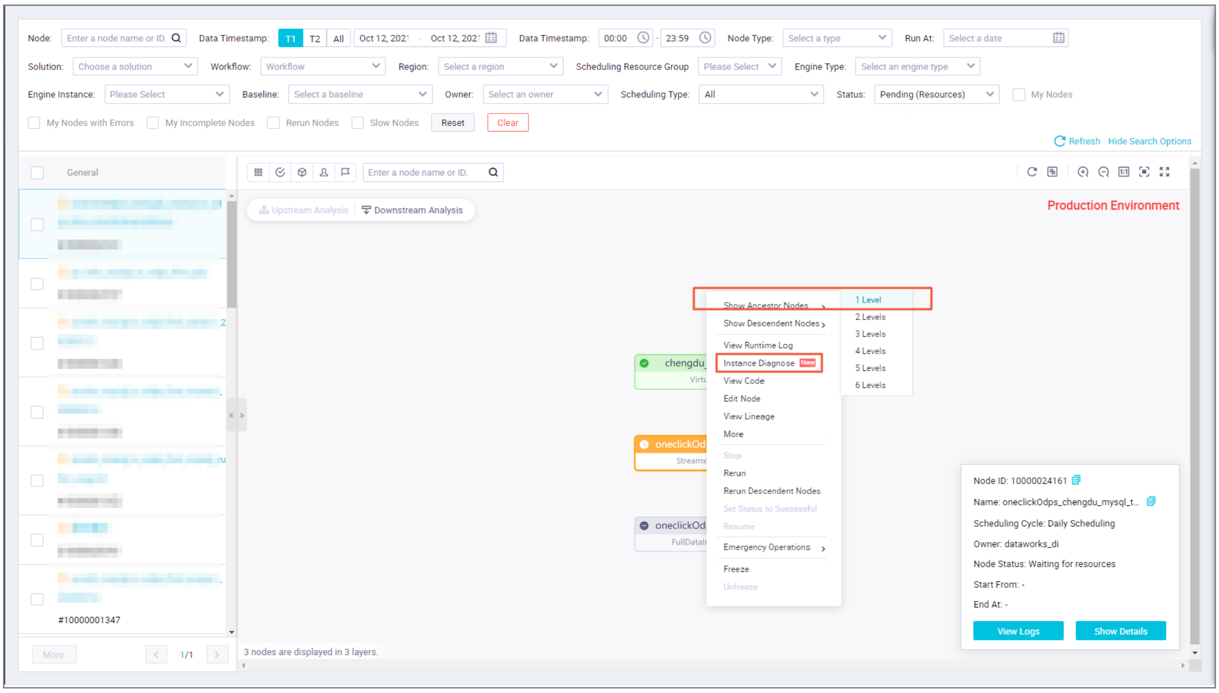
   An auto triggered node that is waiting for scheduling resources is marked in yellow. If a note about waiting for scheduling resources appears in a log, the number of auto triggered nodes that are run in the current workspace has reached the upper limit. Right-click the auto triggered node and select Instance Diagnose. On the Intelligent Diagnosis page, check which nodes are running while the current auto triggered node is waiting for scheduling resources.

3. To be run (dimmed)

   Right-click a dimmed instance of the auto triggered node and select Show Ancestor Nodes to check the status of each ancestor node.

   If all the instances are dimmed, check whether the auto triggered node is an isolated node by referring to the question "What do I do if I can find instances of other auto triggered nodes but cannot find even one instance of the desired auto triggered node?" in this topic.

> **Note** We recommend that you use the intelligent diagnosis and ancestor node analysis features.



Node: Enter a node name or ID. Data Timestamp: T1 T2 All Oct 12, 2021 Oct 12, 2021 Data Timestamp: 00:00 - 23:59 Node Type: Select a type Run At: Select a date

Solution: Choose a solution Workflow: Workflow Region: Select a region Scheduling Resource Group Please Select Engine Type: Select an engine type

Engine Instance: Please Select Baseline: Select a baseline Owner: Select an owner Scheduling Type: All Status: Pending (Resources) □ My Nodes

□ My Nodes with Errors □ My Incomplete Nodes □ Rerun Nodes □ Slow Nodes Reset Clear

[redacted]

#10000001347

Node ID: 10000024161

Name: oneclickOdps_chengdu_mysql_t...

Scheduling Cycle: Daily Scheduling

Owner: dataworks_di

Node Status: Waiting for resources

Start From: -

End At: -

3 nodes are displayed in 3 layers.

[Cause]

1. An auto triggered node is run as scheduled if the node is not frozen and the following information is provided: scheduling resources, scheduled time, and the status of each ancestor node.
2. If the ancestor nodes are not in the preceding states and the entire workflow is dimmed, the ancestor node dependency is changed and the entire workflow is isolated.

What are the conditions that must be met to run a node?

1. The scheduled time of the node has arrived. The node that is waiting for its scheduled time to arrive is marked in yellow.
2. The running of all the ancestor nodes of the current auto triggered node is complete. An ancestor node that is successfully run is marked in green. You can view the status of each ancestor node of the auto triggered node in a directed acyclic graph (DAG) in Operation Center.
3. The resources in a workspace are sufficient to run nodes. An auto triggered node that is waiting for scheduling resources is marked in yellow. The operation log shows that the number of auto triggered nodes that are run in parallel in the workspace has reached the upper limit and the current node is waiting for gateway resources.
4. The node is run as scheduled. A node that is frozen is marked in purple.

3.5. Dry-run instances

This topic provides answers to some frequently asked questions about a dry-run instance.

- What is a dry-run instance?
- Why does a dry-run instance exist?
  1. Scenario 1: An instance is scheduled to run on a specific day every week or every month
  2. Scenario 2: An instance is generated in real time but is deprecated
  3. Scenario 3: The status of an instance is set to successful
  4. Scenario 4: The property of an instance is dry run

What is a dry-run instance?

A dry-run instance refers to an instance that is normally scheduled and successfully run but has no operational logs and execution duration. A dry-run instance does not process data.

Why does a dry-run instance exist?

Scenario 1: An instance is scheduled to run on a specific day every week or every month

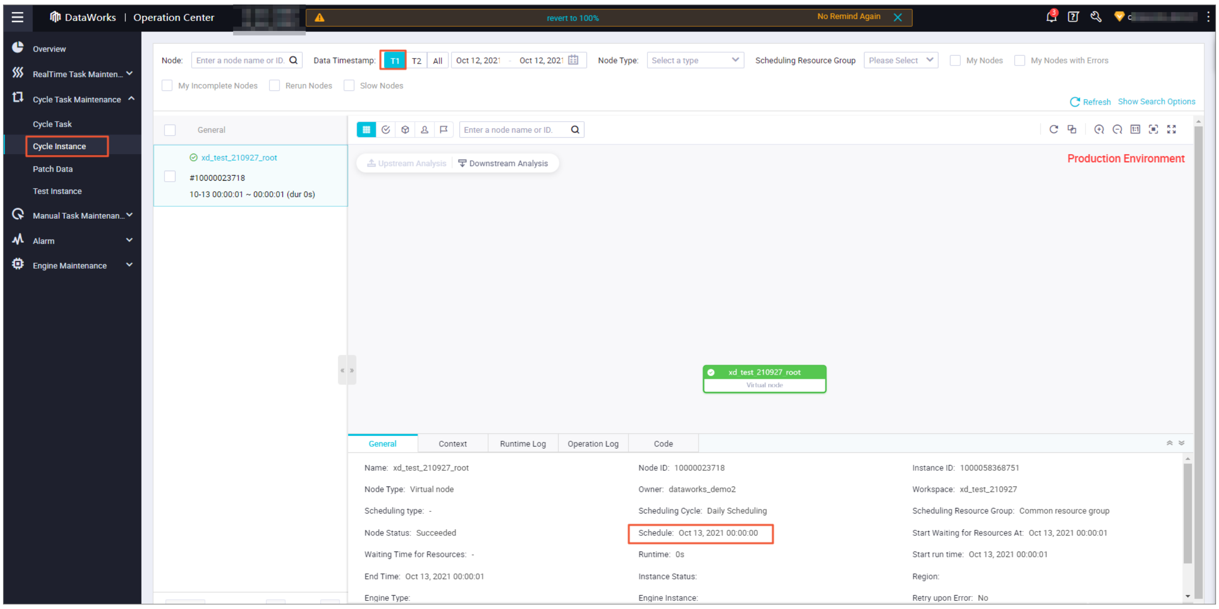
For a node that is scheduled to run on **a specific day every week** or on **a specific day every month**, the scheduling system runs the node only on that day every week or month. On the other days, dry-run instances are generated but the scheduling system does not actually run the node. You must confirm the specific day on which a node is scheduled to run.

> **Note** If you want the scheduling system to actually run an auto triggered instance, you must set the scheduled time of the corresponding node to a point in time that is more than 10 minutes after the node is deployed. Alternatively, you can specify yesterday for the data timestamp and use a data backfill node to backfill data. This way, an auto triggered instance can run today as scheduled.

Solution to specifying the data timestamp when you use a data backfill node to backfill data for a node scheduled by week or month to prevent from generating dry-run instances

If a node is scheduled to run on the first day of every month, we recommend that you set the data timestamp of a data backfill node to the end of every month. If a node is scheduled to run on Monday of every week, we recommend that you set the data timestamp of a data backfill node to Sunday of every week.

View the scheduled time and data timestamp of the auto triggered instance on the current day.



Node: Enter a node name or ID. Data Timestamp: T1 T2 All Oct 12, 2021 - Oct 12, 2021 Node Type: Select a type Scheduling Resource Group Please Select □ My Nodes □ My Nodes with Errors

□ My Incomplete Nodes □ Rerun Nodes □ Slow Nodes

□ General

□ xd_test_210927_root
#10000023718
10-13 00:00:01 ~ 00:00:01 (dur 0s)

Name: xd_test_210927_root
Node Type: Virtual node
Scheduling type: -
Node Status: Succeeded
Waiting Time for Resources: -
End Time: Oct 13, 2021 00:00:01
Engine Type:

Node ID: 10000023718
Owner: dataworks_demo2
Scheduling Cycle: Daily Scheduling
Schedule: Oct 13, 2021 00:00:00
Runtime: 0s
Instance Status:
Engine Instance:

Instance ID: 1000058368751
Workspace: xd_test_210927
Scheduling Resource Group: Common resource group
Start Waiting for Resources At: Oct 13, 2021 00:00:01
Start run time: Oct 13, 2021 00:00:01
Region:
Retry upon Error: No

Scenario 2: An instance is generated in real time but is deprecated

In this example, in the Schedule section on the Properties tab, Start Instantiation is set to **Immediately After Deployment** to generate auto triggered instances for a node. The scheduling system runs only the instances of the node whose scheduled time is more than 10 minutes after the node is deployed. For the instances of the node whose scheduled time is within 10 minutes after the deployment time of the node, the scheduling system does not actually run these instances but generates dry-run instances. The status of the instances is Deprecated real time generated task. For more information, see Configure time properties for a node to immediately generate an instance.

Scenario 3: The status of an instance is set to successful

After you set the status of a failed instance to successful, the scheduling system does not actually run the instance and continues to run the instances of the descendant node of the current node.**Succeeded** The status of the instance is Instance Set Successfully.

Scenario 4: The property of an instance is dry run

In the Schedule section on the Properties tab in DataStudio, check whether Recurrence is set to **Dry Run** for a node.

Scenario 5: An instance is not selected for a temporary workflow

In this example, Node C depends on Node B, and Node B depends on Node A. If you want to backfill data for Nodes A and C, the status of Node B is Unselected instance in temporary workflow.

Troubleshoot dry runs for nodes that are scheduled on a daily basis

If a node is scheduled on a daily basis, check whether Recurrence is set to Dry Run for the node in the Schedule section on the Properties tab.

> **Notice** T+1: indicates that the scheduling system runs nodes on the second day by using the data that is generated on the current day.

3.6. Node freezing and unfreezing

This topic provides answers to some frequently asked questions about node freezing and unfreezing.

- What happens after I freeze or unfreeze an auto triggered node or an auto triggered instance?
- What happens to data backfill node instances and test node instances after I freeze or unfreeze an auto triggered node?
- How do I rerun an unfrozen auto triggered instance?
- Why is a frozen auto triggered node run as scheduled?
- How do I check the operations that are performed on a node and who performed the operations?

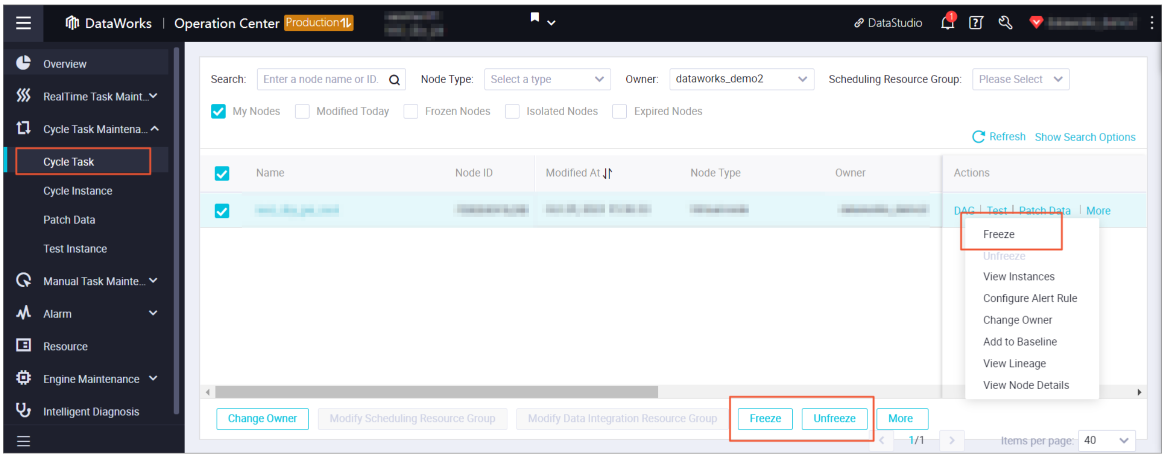
What happens after I freeze or unfreeze an auto triggered node or an auto triggered instance?

- Freeze or unfreeze an auto triggered node

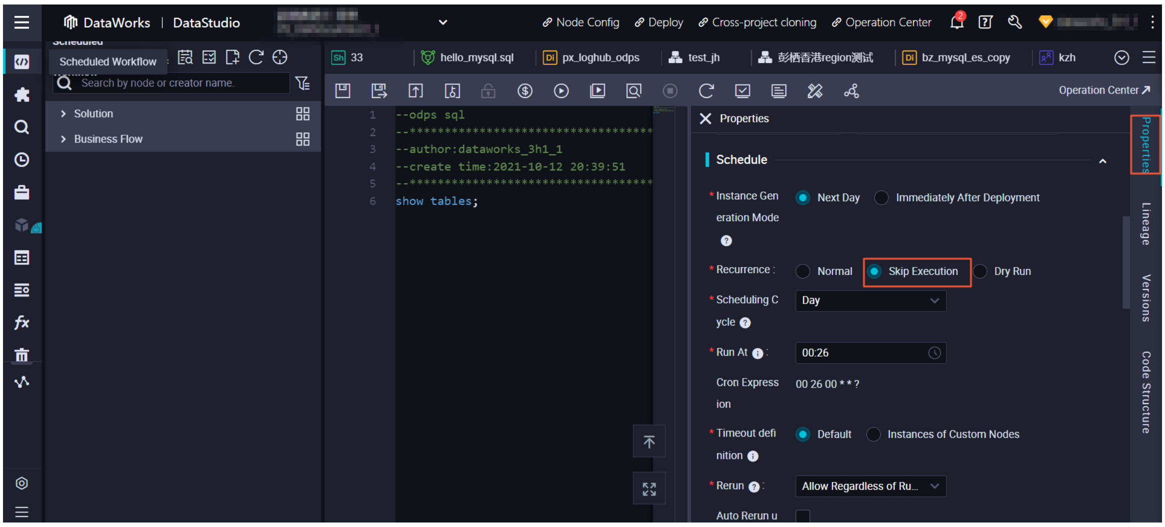
  Every night, the system generates auto triggered instances that are scheduled to run on the next day based on an auto triggered node. If you freeze an auto triggered node, the auto triggered instances that are generated after the freeze operation are also frozen and the descendant instances that are run based on the auto triggered instances cannot be run.

Note The auto triggered instances that are scheduled to run on the same day when you freeze the auto triggered node can be run as expected. Auto triggered instances that are scheduled to run on the next day after the freeze operation are frozen and all the descendant nodes are blocked from running.

- Freeze or unfreeze auto triggered nodes in **Operation Center**.



- Freeze or unfreeze an auto triggered node in DataStudio



```
--odps sql
--**********************************
--author:dataworks_3h1_1
--create time:2021-10-12 20:39:51
--**********************************
show tables;
```

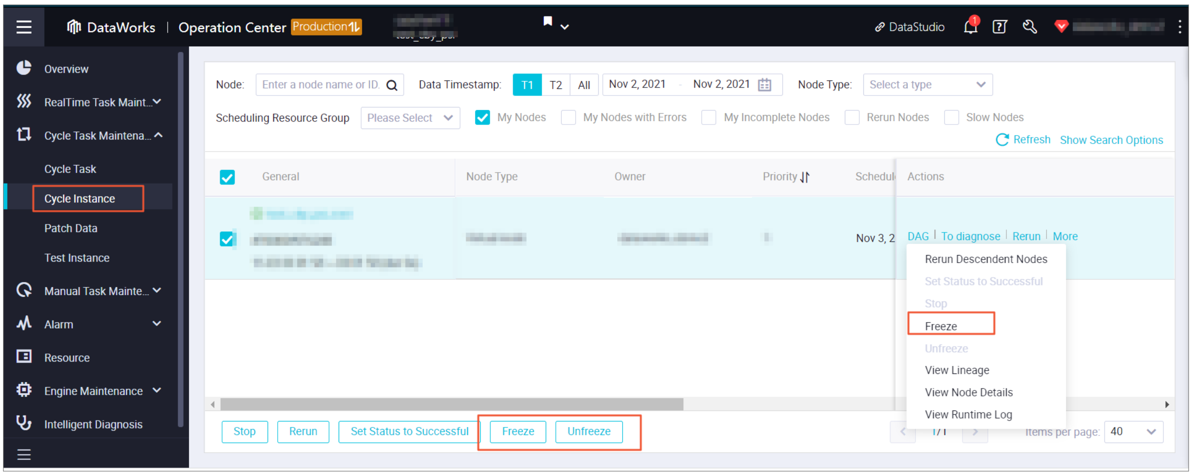
Schedule

- * Instance Generation Mode: ● Next Day ○ Immediately After Deployment
- * Recurrence : ○ Normal ● Skip Execution ○ Dry Run
- * Scheduling Cycle: Day
- * Run At : 00:26
- Cron Expression: 00 26 00 * * ?
- * Timeout definition: ● Default ○ Instances of Custom Nodes
- * Rerun : Allow Regardless of Ru...
- Auto Rerun u

Note In the Schedule section of the Properties tab on the DataStudio page, set Recurrence to Skip Execution or Normal. Then, commit and deploy the node again. This way, the auto triggered node can be frozen or unfrozen in the production environment.

- Freeze or unfreeze an auto triggered instance

The freeze or unfreeze operation on an auto triggered instance does not have an impact on the status of the auto triggered node to which the instance belongs. If an auto triggered node is frozen and you unfreeze an auto triggered instance that belongs to the node, other auto triggered instances that are scheduled to run on the next day are still frozen.



Notice Freeze and unfreeze operations are manual operations. For any questions, view operation logs.

What happens to data backfill node instances and test node instances after I freeze or unfreeze an auto triggered node?

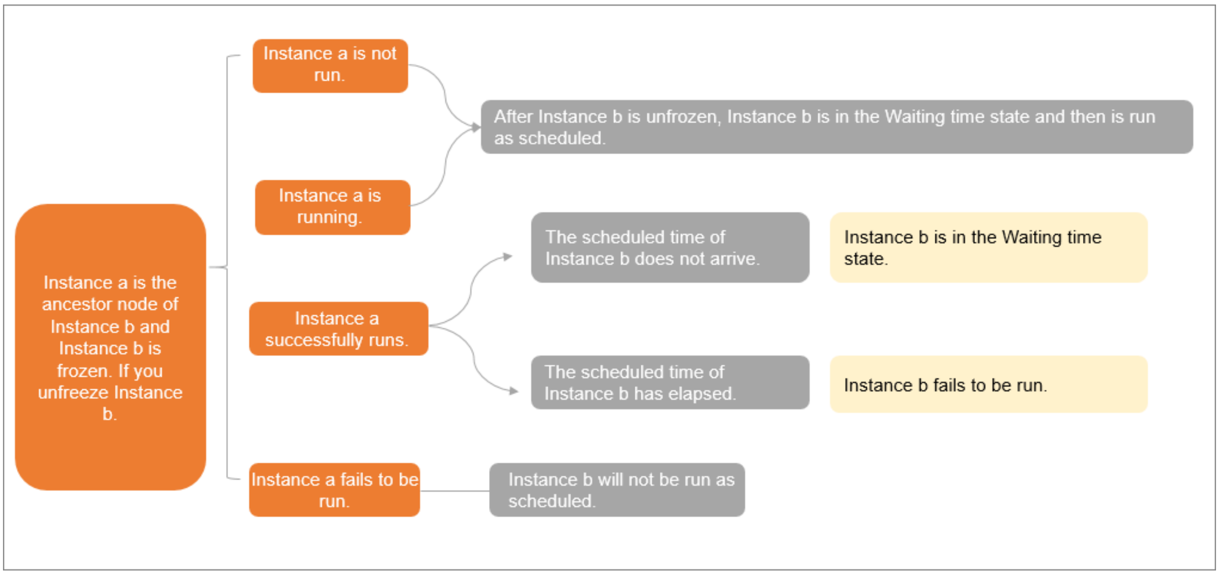
A data backfill node instance and a test node instance are snapshots that are generated for an auto triggered node. If an auto triggered node is frozen, the generated data backfill node instance and test node instance are also frozen.

How do I rerun an unfrozen auto triggered instance?

An unfrozen auto triggered instance is run based on the scheduled time that is specified in the Schedule section of the Properties tab, and the status of the ancestor node of the instance.

For example, Instance a is the ancestor node of Instance b and Instance b is frozen. If you unfreeze Instance b:

- Scenario 1: If Instance a is not run, after Instance b is unfrozen, Instance b is in the Waiting time state and then is run as scheduled.
- Scenario 2: If Instance a is running, after Instance b is unfrozen, Instance b is in the Waiting time state and then is run as scheduled.
- Scenario 3: Instance a successfully runs.
    - If the scheduled time of Instance b does not arrive, Instance b is in the Waiting time state.
    - If the scheduled time of Instance b has elapsed, Instance b fails to be run. If you want to rerun Instance b, click **Rerun** in the **Actions** column of Instance b. After Instance b successfully runs, the descendant instances of Instance b are run as scheduled.
- Scenario 4: If Instance a fails to be run, Instance b is not run. For more information about how to troubleshoot an instance that fails to be run, see Nodes that are not run.


Instance a is the ancestor node of Instance b and Instance b is frozen. If you unfreeze Instance b.
Instance a is not run.
Instance a is running.
After Instance b is unfrozen, Instance b is in the Waiting time state and then is run as scheduled.
Instance a successfully runs.
The scheduled time of Instance b does not arrive.
Instance b is in the Waiting time state.
The scheduled time of Instance b has elapsed.
Instance b fails to be run.
Instance a fails to be run.
Instance b will not be run as scheduled.

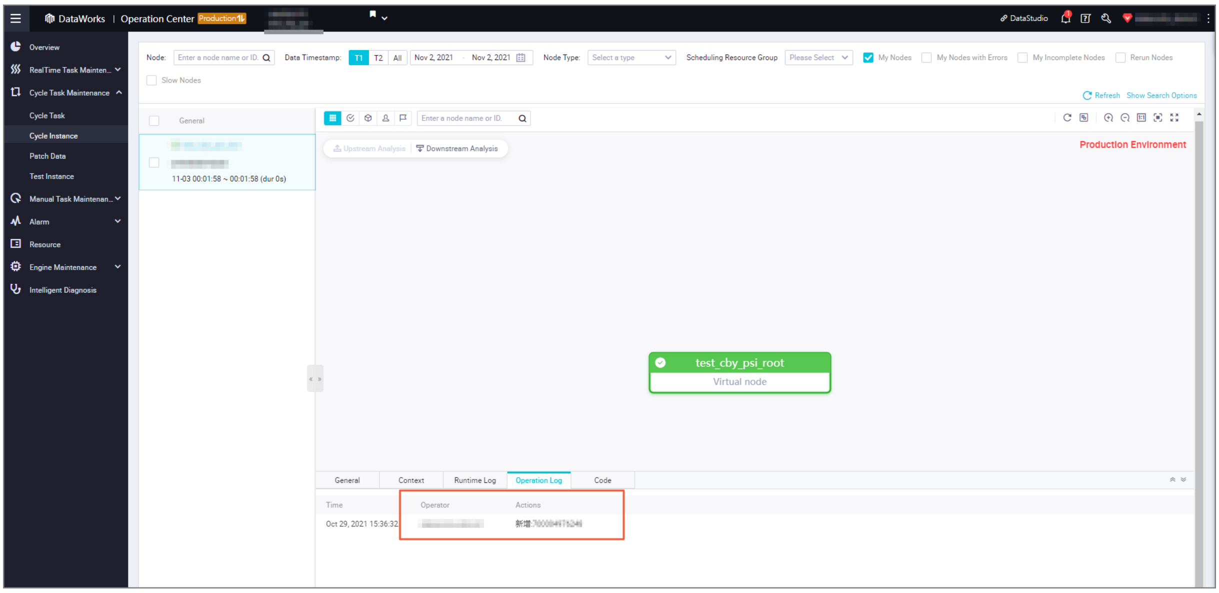
Why is a frozen auto triggered node run as scheduled?

Every night, the system generates auto triggered instances that are scheduled to run on the next day based on an auto triggered node. You can manually create a data backfill node and a test node for the auto triggered node, and a data backfill node instance and a test node instance can be generated based on snapshot information. The following items describe why a frozen auto triggered node is still run as scheduled:

- Check whether an auto triggered node is frozen.
- The freeze operation that is performed on an auto triggered node does not take effect on the auto triggered instances that are generated before the freeze operation.
  - The auto triggered instances that are scheduled to run on the same day when you freeze the auto triggered node can be run as expected.
  - The freeze operation does not take effect on data backfill node instances and test node instances that are generated before the freeze operation.

How do I check the operations that are performed on a node and who performed the operations?

You can view operation logs on the Cycle Task or Cycle Instance page in Operation Center.



[redacted]

[redacted]

11-03 00:01:58 ~ 00:01:58 (dur 0s)

Time	Operator	Actions
Oct 29, 2021 15:36:32	[redacted]	新增:[redacted]

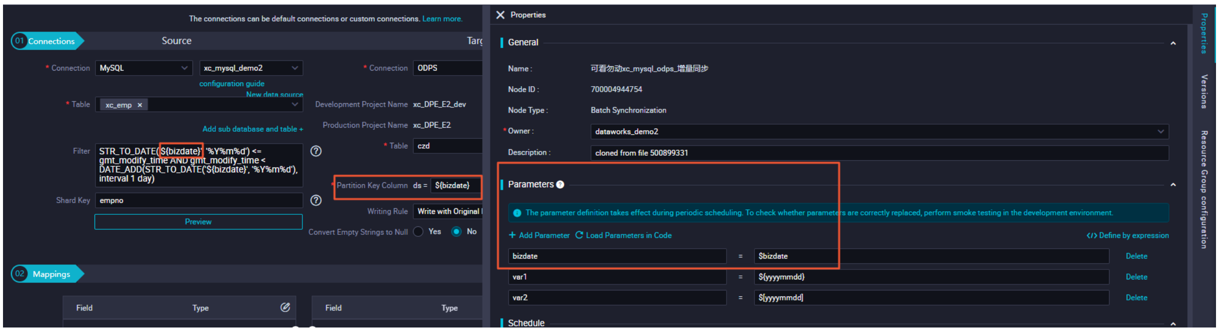
3.7. Data backfill

This topic provides answers to some frequently asked questions about the Data backfill.

- Feature of generating retroactive data for nodes
- Why do the retroactive instances of a node that is scheduled by hour or minute not run in parallel after I enable the parallelism feature for the node?
- The retroactive instances of a node are not run after I specify the data timestamp for retroactive data generation. The retroactive instances are in the Pending (Schedule) state and are highlighted in yellow in the DAG. Why does this happen?
- Why is a retroactive instance of an auto triggered node in the Pending (Schedule) state after I specify the last day and the current day for the Data Timestamp parameter?
- Why are multiple retroactive instances generated for a node if I set the data timestamp to 00:00:00 to 01:00:00?
- If a large number of retroactive instances are generated for a node, the retroactive instances are in the Pending (Resources) state and are highlighted in yellow in the DAG. Why does this happen?
- Why do I receive the error message which indicates that the scheduled runtime of a node is not within the specified data timestamp range?
- Why cannot retroactive instances be generated for a node after I enable retroactive data generation for the node?

Feature of generating retroactive data for nodes

DataWorks allows you to generate retroactive data for nodes for a specified time range in the past or the future. The scheduling parameters of the nodes are automatically replaced with specific values based on the data timestamps that you specify for retroactive data generation. The following figure shows how to write incremental data from a MySQL database to a specified time partition in MaxCompute.



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

01 Connections

Source

* Connection: MySQL, xc_mysql_demo2

* Table: xc_emp

Filter: STR_TO_DATE(${bizdate}' '%Y%m%d') <= gmt_modify_time AND gmt_modify_time < DATE_ADD(STR_TO_DATE('${bizdate}', '%Y%m%d'), interval 1 day)

Shard Key: empno

Targ

* Connection: ODPS

Development Project Name: xc_DPE_E2_dev

Production Project Name: xc_DPE_E2

* Table: czd

* Partition Key Column: ds = ${bizdate}

Writing Rule: Write with Original I

Convert Empty Strings to Null: ○ Yes ● No

02 Mappings

Properties

General

Name: 可看勿动xc_mysql_odps_增量同步

Node ID: 700004944754

Node Type: Batch Synchronization

* Owner: dataworks_demo2

Description: cloned from file 500899331

Parameters

The parameter definition takes effect during periodic scheduling. To check whether parameters are correctly replaced, perform smoke testing in the development environment.

bizdate = $bizdate

var1 = ${yyyymmdd}

var2 = $[yyyymmdd]

Schedule

Why do the retroactive instances of a node that is scheduled by hour or minute not run in parallel after I enable the parallelism feature for the node?

The parallelism feature allows you to run multiple retroactive instances of a daily scheduled node in parallel to generate retroactive data for a number of days based on the data timestamp. However, if a node is scheduled by hour or minute, whether all the retroactive instances that are generated for the node on a day can be run in parallel is not controlled by the parallelism feature. Instead, the retroactive instances that are generated for the node on a day can be run in parallel only if you do not configure the node to depend on its instance in the last cycle. For more information, see Scenario 2: Configure scheduling dependencies for a node that depends on last-cycle instances.

1. If you disable the parallelism feature, one retroactive instance is run multiple times in sequence based on the data timestamp.

   In other words, the retroactive instance can be run again only after the retroactive data is generated for the last cycle.

2. If you enable the parallelism feature, you can set the Number of Concurrent nodes parameter to a value that is allowed by the resource groups as needed. In this case, multiple retroactive instances are generated.

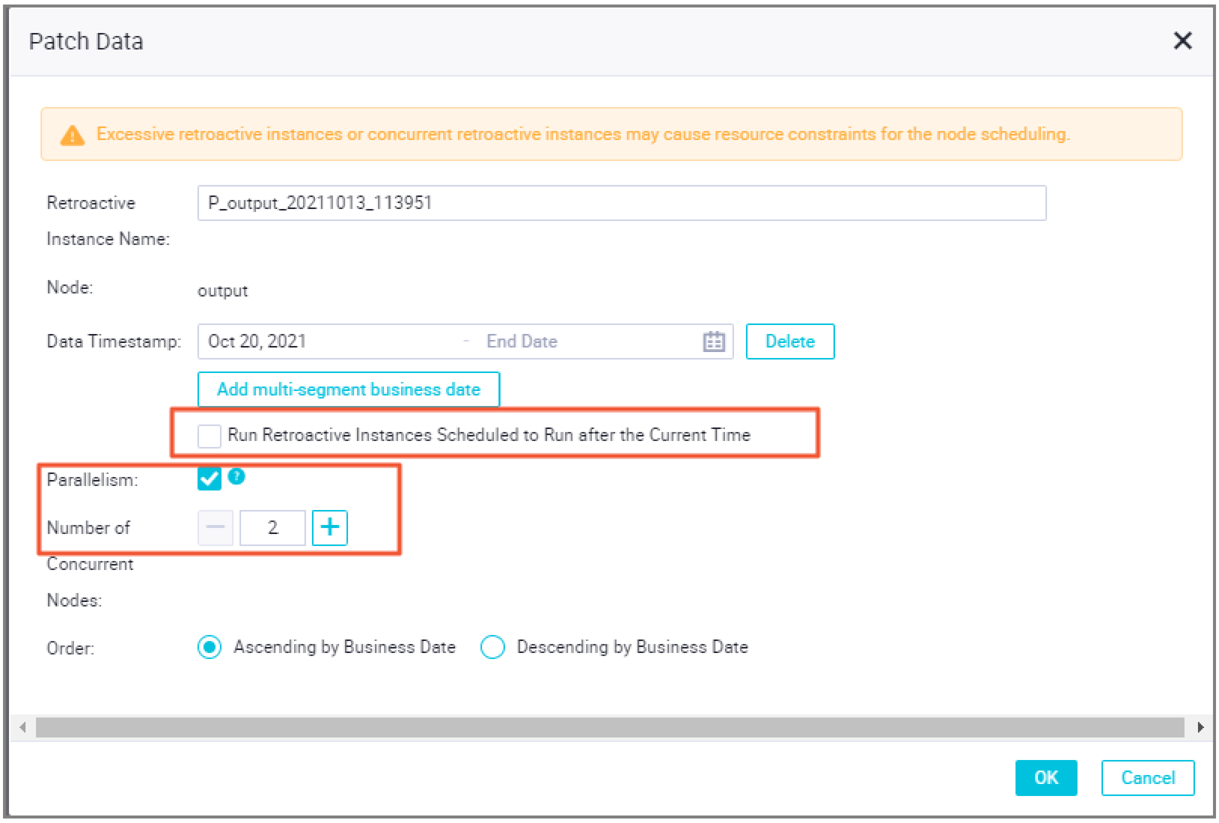
   The retroactive instances are run in parallel based on the data timestamp.

Scenario: You want to generate retroactive data for one week for a node that is scheduled by hour or minute.

- If you configure the node to depend on its instance in the last cycle, one retroactive instance is run multiple times in sequence on each day based on the data timestamp.
- If you do not configure the node to depend on its instance in the last cycle, multiple retroactive instances are run in parallel on each day based on the data timestamp.

The retroactive instances of a node are not run after I specify the data timestamp for retroactive data generation. The retroactive instances are in the Pending (Schedule) state and are highlighted in yellow in the DAG. Why does this happen?

When you generate retroactive data for a node, if you set the Data Timestamp parameter to a future time range that is later than the current time, the retroactive instances of the node are in the Pending (Schedule) state. You can specify whether to immediately run the retroactive instances.


Patch Data

Excessive retroactive instances or concurrent retroactive instances may cause resource constraints for the node scheduling.

Retroactive Instance Name: P_output_20211013_113951

Node: output

Data Timestamp: Oct 20, 2021 - End Date

□ Run Retroactive Instances Scheduled to Run after the Current Time

Parallelism: ☑

Number of Concurrent Nodes: 2

Order: ● Ascending by Business Date ○ Descending by Business Date

Specify whether to select Run Retroactive Instances Scheduled to Run after the Current Time based on your business requirements:

- If you set the Data Timestamp parameter to a future time range and do not select this parameter, the retroactive instances are in the Pending (Schedule) state and are highlighted in yellow in the directed acyclic graph (DAG).
- If you set the Data Timestamp parameter to a future time range and select this parameter, the retroactive instances are immediately run.

Why is a retroactive instance of an auto triggered node in the Pending (Schedule) state after I specify the last day and the current day for the Data Timestamp parameter?

DataWorks runs an auto triggered node on the current day based on the data whose data timestamp is of the last day. The process of generating retroactive data for the last day for an auto triggered node is the same as that of running the auto triggered node on the current day.

Note To query the instance that is generated by the auto triggered node for the current day, set the Data Timestamp parameter to T1 on the Cycle Instance page. The data timestamp of the instance is of the last day, and the scheduled runtime of the instance is of the current day.

Why are multiple retroactive instances generated for a node if I set the data timestamp to 00:00:00 to 01:00:00?

The number of retroactive instances that are generated for a node depends on the scheduled runtime that you specify for the node.

- Scenario 1: You configure a node to be scheduled by hour from 00:00:00 to 23:59:00. If you set the data timestamp to 00:00:00 to 01:00:00, two retroactive instances are generated and scheduled at 00:00:00 and 01:00:00.
- Scenario 2: You configure a node to be scheduled every 30 minutes from 00:00:00 to 23:59:00. If you set the data timestamp to 00:00:00 to 01:00:00, three instances are generated and scheduled at 00:00:00, 00:30:00, and 01:00:00.

If a large number of retroactive instances are generated for a node, the retroactive instances are in the Pending (Resources) state and are highlighted in yellow in the DAG. Why does this happen?

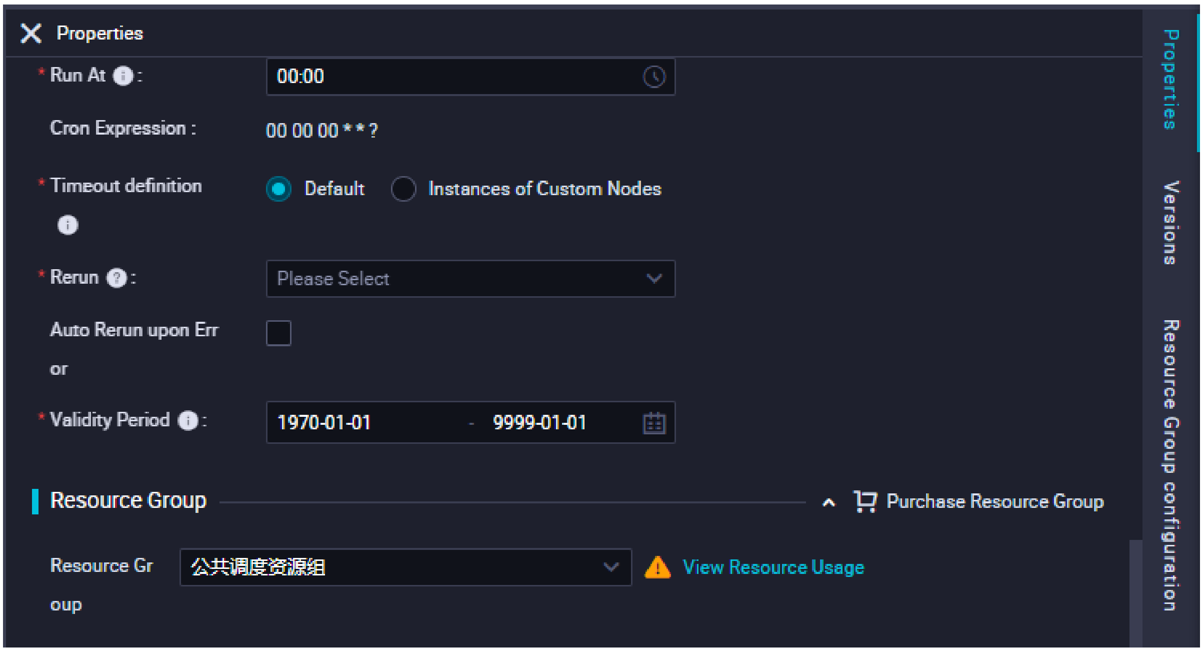
The maximum number of concurrent instances is limited for a resource group for scheduling. If the number of concurrent instances of a node exceeds the upper limit of the resource group for scheduling, the instances are in the Pending (Resources) state. For more information about how to troubleshoot this issue, see Nodes that are waiting for resources.

Why do I receive the error message which indicates that the scheduled runtime of a node is not within the specified data timestamp range?

You must specify a time range for a node that is scheduled by hour or minute. Otherwise, retroactive instances cannot be generated for the node.

Why cannot retroactive instances be generated for a node after I enable retroactive data generation for the node?

Retroactive instances can be generated for nodes whose scheduled runtime is within the specified data timestamp range. Make sure that the scheduled runtime of the node meets this requirement.



Properties

* Run At: 00:00

Cron Expression : 00 00 00 * * ?

* Timeout definition: ● Default ○ Instances of Custom Nodes

* Rerun: Please Select

Auto Rerun upon Error □

* Validity Period: 1970-01-01 - 9999-01-01

Resource Group

Resource Group 公共调度资源组

3.8. Nodes that are successfully run but have no data generated

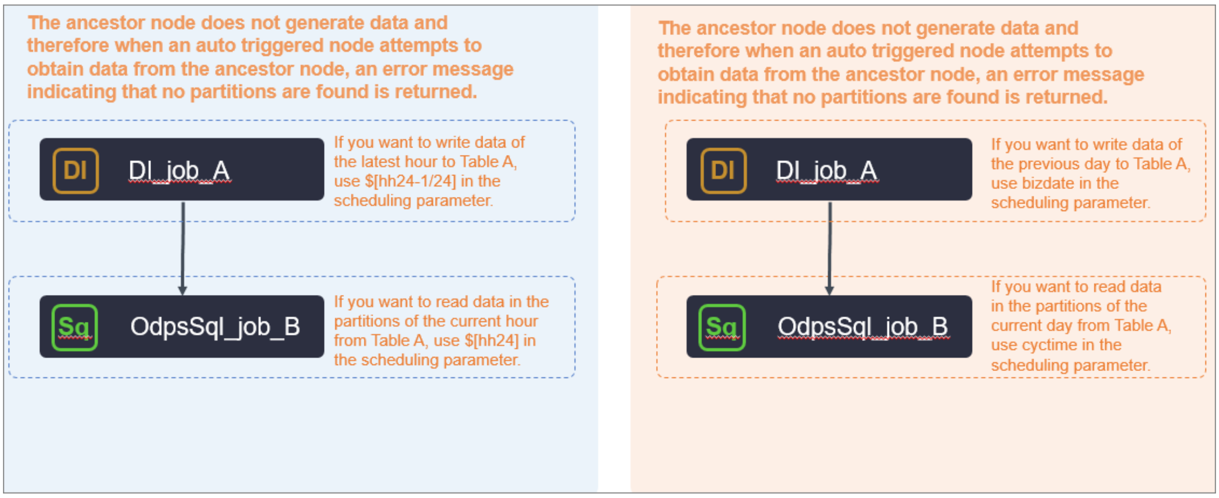
This topic provides answers to some frequently asked questions about the nodes that are successfully run but have no data generated.

- Scenario 1: An auto triggered node is successfully run and has operational logs
- Scenario 2: An auto triggered node is successfully run but has no operational logs

Scenario 1: An auto triggered node is successfully run and has operational logs

If an auto triggered node is successfully run, the code logic of the node is executed as expected. However, when the instances of the auto triggered node attempt to be run as scheduled, the node fails to obtain data of its ancestor nodes or an error is reported, indicating that the partitions of the output table of its ancestor nodes do not exist. You can manually rerun the instances to obtain data of the ancestor nodes.

1. The auto triggered node does not depend on its ancestor nodes that generates an output table.
2. The auto triggered node depends on its ancestor nodes that generates an output table. However, the partitions of the output table are not what are expected. This indicates that the cycle of the instances on which the auto triggered node depends is incorrect. On the General tab of the auto triggered node or on the page that appears after you click the auto triggered node and click View Logs, you can view the values of the parameters for the ancestor and descendant nodes of the auto triggered node in a day.



The ancestor node does not generate data and therefore when an auto triggered node attempts to obtain data from the ancestor node, an error message indicating that no partitions are found is returned.

If you want to write data of the latest hour to Table A, use $[hh24-1/24] in the scheduling parameter.

If you want to read data in the partitions of the current hour from Table A, use $[hh24] in the scheduling parameter.

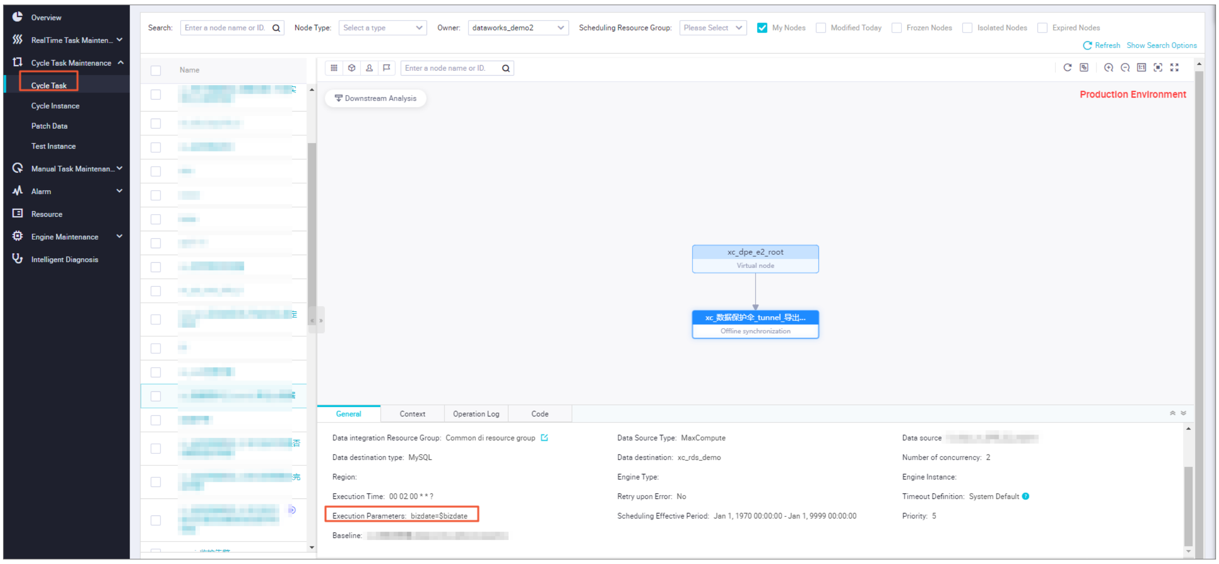
The ancestor node does not generate data and therefore when an auto triggered node attempts to obtain data from the ancestor node, an error message indicating that no partitions are found is returned.

If you want to write data of the previous day to Table A, use bizdate in the scheduling parameter.

If you want to read data in the partitions of the current day from Table A, use cyctime in the scheduling parameter.

Reconfigure dependencies between nodes.

Check the values of the parameters for the ancestor and descendant nodes of the auto triggered node



Search: Enter a node name or ID. Node Type: Select a type Owner: dataworks_demo2 Scheduling Resource Group: Please Select ☑ My Nodes □ Modified Today □ Frozen Nodes □ Isolated Nodes □ Expired Nodes

Data integration Resource Group: Common di resource group

Data Source Type: MaxCompute

Data source [redacted]

Data destination type: MySQL

Data destination: xc_rds_demo

Number of concurrency: 2

Region:

Engine Type:

Engine Instance:

Execution Time: 00 02 00 * * ?

Retry upon Error: No

Timeout Definition: System Default

Execution Parameters: bizdate=$bizdate

Scheduling Effective Period: Jan 1, 1970 00:00:00 - Jan 1, 9999 00:00:00

Priority: 5

Baseline: [redacted]

Scenario 2: An auto triggered node is successfully run but has no operational logs

View the status of the auto triggered node on the General tab. For more information, see Dry-run instances.

3.9. Node failures

This topic provides answers to some frequently asked questions about node failures.

- Errors for node rerun failures
  - I have configured rerun properties for my node, but the node does not rerun after it fails, and the following error message appears: Task Run Timed Out, Killed by System!!!. What do I do?
  - I set the Auto Rerun Times upon Error parameter to 1 for my node, but the node does not rerun after it fails. What do I do?
- Errors for ODPS nodes
  - What do I do if the error message ODPS-0420095: Access Denied - Authorization Failed [4093], You have NO privilege to do the restricted operation on {acs:odps:*:projects/xxxx}. Access Mode is AllDenied. appears?
  - What do I do if the error message ODPS-0420061: Invalid parameter in HTTP request - Fetched data is larger than the rendering limitation. Please try to reduce your limit size or column number appears?
  - What do I do if the data synchronized by using multiple threads is out of order?
- Errors for AnalyticDB for MySQL nodes

  What do I do if my synchronization node that uses an AnalyticDB for MySQL data source and runs on a shared resource group fails to run?
- Errors for general nodes
  - How do I view the logs of a for-each node, do-while node, or PAI node in Operation Center?
  - What do I do if the error message error in your condition run fail appears when I run a branch node?

  - What do I do if the error message None Ftp connection info!! appears when I run an FTP Check node?
  - What do I do if the error message Connect Failed appears when I run an FTP Check node?
  - What do I do if the error message The current time has exceeded the end-check time point! appears when I run an FTP Check node?
  - What do I do if the error message File not Exists or exceeded the end-check time point! appears when I run an FTP Check node?

- Error for a resource group

  What do I do if the error message no available machine resources under the task resource group appears for my resource group for scheduling?

I have configured rerun properties for my node, but the node does not rerun after it fails, and the following error message appears: `Task Run Timed Out, Killed by System!!!`. What do I do?

- Problem description:

  The **Rerun** parameter in the **Schedule** section of the **Properties** tab is set to **Allow Regardless of Running Status** or **Allow upon Failure Only** for the node. However, the node does not rerun after it fails, and the error message `Task Run Timed Out, Killed by System!!!` is displayed.

- Possible cause:

  The **Timeout Period** parameter in the **Schedule** section of the **Properties** tab is configured for the node. If the running duration of the node exceeds the value of the Timeout Period parameter, the node automatically stops running and does not rerun.

- Solution:

  Manually rerun the node.

I set the Auto Rerun Times upon Error parameter to 1 for my node, but the node does not rerun after it fails. What do I do?

- Problem description:

  The **Auto Rerun Times upon Error** parameter is set to **1** for the node in the **Schedule** section of the **Properties** tab. However, the node does not rerun after it fails.

- Possible cause:

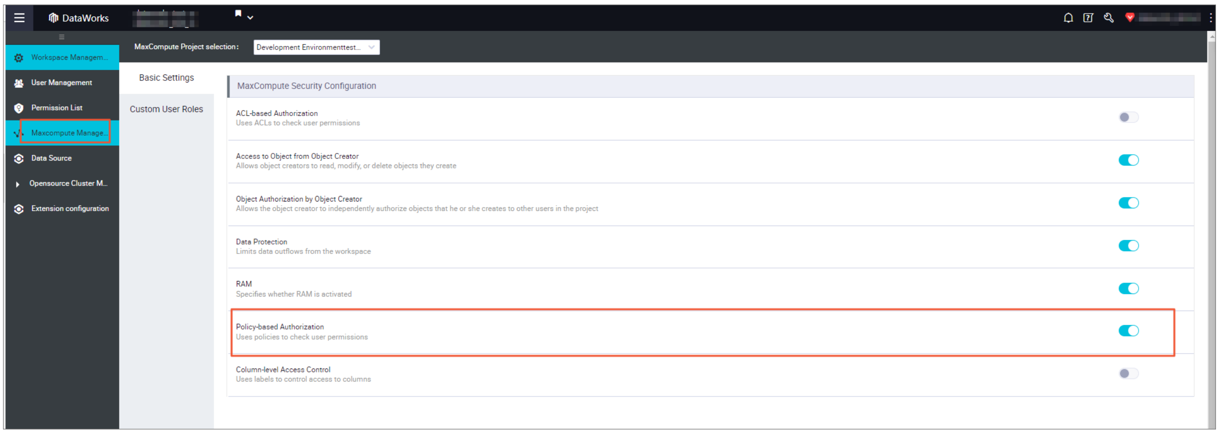
  The value that you specified for the **Auto Rerun Times upon Error** parameter is invalid. If you set the **Auto Rerun Times upon Error** parameter to **n**, the node actually reruns **n - 1** times. The default value of the **Auto Rerun Times upon Error** parameter is **3**, the minimum value of this parameter is **1**, and the maximum value is **10**. The value 1 indicates that the node does not rerun after it fails, and the value 10 indicates that the node reruns nine times after it fails. You can configure this parameter based on your business requirements.

- Solution:

  To enable the node to rerun **once** after it fails, set the **Auto Rerun Times upon Error** parameter to **2**.

What do I do if the error message ODPS-0420095: Access Denied - Authorization Failed [4093], You have NO privilege to do the restricted operation on {acs:odps:*:projects/xxxx}. Access Mode is AllDenied. appears?

1. Check whether the MaxCompute compute engine has overdue payments.
2. On the **MaxCompute Management** page, check whether Policy-based Authorization is turned on.



MaxCompute Project selection: Development Environmenttest...

MaxCompute Security Configuration

ACL-based Authorization
Uses ACLs to check user permissions

Access to Object from Object Creator
Allows object creators to read, modify, or delete objects they create

Object Authorization by Object Creator
Allows the object creator to independently authorize objects that he or she creates to other users in the project

Data Protection
Limits data outflows from the workspace

RAM
Specifies whether RAM is activated

Policy-based Authorization
Uses policies to check user permissions

Column-level Access Control
Uses labels to control access to columns

What do I do if the error message ODPS-0420061: Invalid parameter in HTTP request - Fetched data is larger than the rendering limitation. Please try to reduce your limit size or column number appears?

Specify a threshold for the number of data records that can be returned in your SQL statement. If you want to obtain more data records, you can export the data records. If you want to obtain more than 10,000 data records, use a Tunnel command to export the data records.

What do I do if the data synchronized by using multiple threads is out of order?

A synchronization node reads data from MaxCompute tables in random order. If you do not configure order settings, the data returned by the synchronization node is also out of order.

By default, the data synchronized from MaxCompute is stored in random order. If you want to obtain sorted data, configure order settings for the synchronized data. For example, you can configure **order by xx limit n** in the SQL statement of the synchronization node to sort data.

What do I do if my synchronization node that uses an AnalyticDB for MySQL data source and runs on a shared resource group fails to run?

Purchase an exclusive resource group for scheduling, bind the resource group to a virtual private cloud (VPC) that can be connected to the AnalyticDB for MySQL data source, and then run the synchronization node on the exclusive resource group for scheduling. For more information, see Test data source connectivity.

When I run a data synchronization node that uses a MySQL data source, the system displays an error message indicating that the Java Database Connectivity (JDBC) driver for the MySQL node is not supported. What do I do?

- The error message `sql execute failed! The JDBC driver is not supported.` appears because the MySQL data source that you selected is not added by using the connection string mode.
- Select a MySQL data source that is added by using the connection string mode. You can refer to the operations in Add a MySQL data source to go to the **Data Source** page, find the desired data source, and then click **Edit** in the **Operation** column to view the mode that is used to add the data source.

How do I view the logs of a for-each node, do-while node, or PAI node in Operation Center?

Find your node in Operation Center and open the directed acyclic graph (DAG) of the node. Then, right-click the node name in the DAG and select View Internal Nodes.

What do I do if the error message `error in your condition run fail` appears when I run a branch node?

- Check whether the branch conditions that you specified for the branch node comply with the specifications of the Python syntax.
- If the output of the ancestor assignment node of the branch node is strings, you must enclose variables in single quotation marks (') for the branch node to reference the variables.

What do I do if the error message `None Ftp connection info!!` appears when I run an FTP Check node?

- Problem description: The FTP Check node that is used to check whether the FTP data source contains the Done file fails to run, and the error message `None Ftp connection info!!` appears.
- Possible cause: The FTP data source is incorrectly configured. As a result, the FTP Check node fails to obtain information about the FTP data source.
- Solution: Go to the **Data Source** page to check whether the configurations of the data source that you use are correct. For more information about how to go to the page, see Manage connections. If no FTP data source is available on the Data Source page, you must add an FTP data source. For more information, see Add an FTP data source.

What do I do if the error message `Connect Failed` appears when I run an FTP Check node?

- Problem description: The FTP Check node that is used to check whether the FTP data source contains the Done file fails to run, and the error message `Connect Failed` appears.
- Possible cause: The FTP data source fails to connect to the FTP server.
- Solution: Run the **telnet IP address Port number** command to check whether the FTP server is normally running. Replace the IP address and port number in this command with the IP address and port number of the FTP data source. You can go to the **Data Source** page of the DataWorks console to view the IP address and port number of the FTP data source. For more information about how to go to the **Data Source** page, see Manage connections.

What do I do if the error message `The current time has exceeded the end-check time point!` appears when I run an FTP Check node?

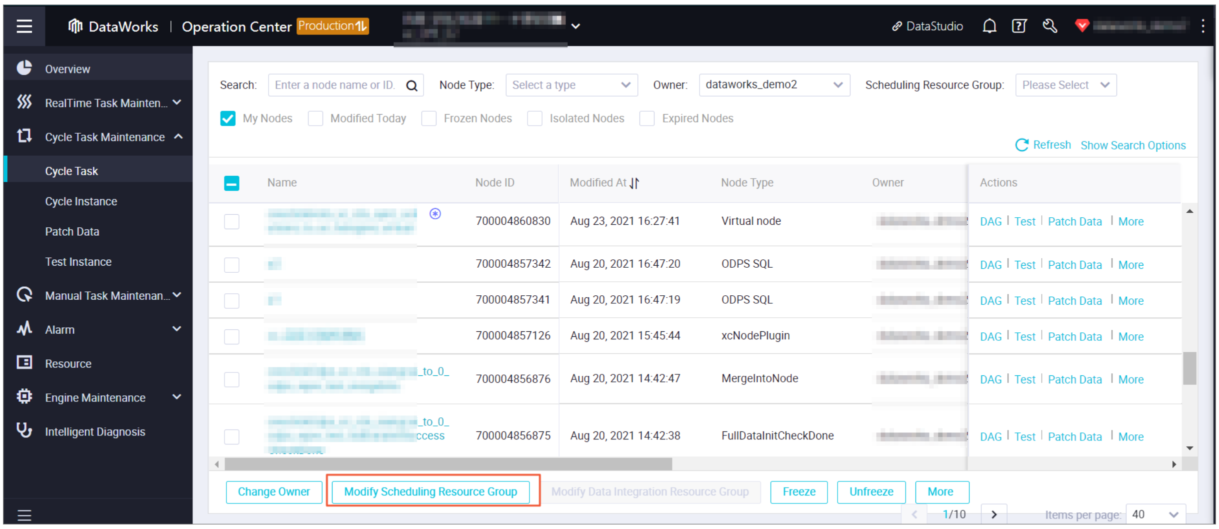
- Problem description: The FTP Check node that is used to check whether the FTP data source contains the Done file fails to run, and the error message `The current time has exceeded the end-check time point!` appears.
- Cause: The time specified by **Check stop time** has elapsed.
- Solution: Modify the value of **Check stop time** based on your business requirements on the configuration tab of the FTP Check node. For more information, see Configure a detection policy.

What do I do if the error message `File not Exists or exceeded the end-check time point!` appears when I run an FTP Check node?

- Problem description: The FTP Check node that is used to check whether the FTP data source contains the Done file fails to run, and the error message `File not Exists or exceeded the end-check time point!` appears.
- Cause: One possible cause is that the FTP Check node does not find the Done file before the time specified by **Check stop time** arrives. Another possible cause is that the time specified by **Check stop time** has elapsed.
- Solution: This is an expected error. If this error message appears, DataWorks does not trigger the descendant node of the FTP Check node to run.

What do I do if the error message `no available machine resources under the task resource group` appears for my resource group for scheduling?

- Problem description: The system displays the following error message for the resource group for scheduling: `no available machine resources under the task resource group`.
- Solution: Change the resource group for scheduling that you use to run your node. To change the resource group, perform the following operations: Log on to the DataWorks console and go to Operation Center. In the left-side navigation pane of the **Operation Center** page, choose **Cycle Task Maintenance > Cycle Task**. On the Cycle Task page, find the auto triggered node for which you want to change the resource group and click Modify Scheduling Resource Group in the lower part of the page.



Search: Enter a node name or ID. Node Type: Select a type Owner: dataworks_demo2 Scheduling Resource Group: Please Select

☑ My Nodes □ Modified Today □ Frozen Nodes □ Isolated Nodes □ Expired Nodes

	Name	Node ID	Modified At	Node Type	Owner	Actions
□	[redacted]	700004860830	Aug 23, 2021 16:27:41	Virtual node	[redacted]	DAG | Test | Patch Data | More
□	[redacted]	700004857342	Aug 20, 2021 16:47:20	ODPS SQL	[redacted]	DAG | Test | Patch Data | More
□	[redacted]	700004857341	Aug 20, 2021 16:47:19	ODPS SQL	[redacted]	DAG | Test | Patch Data | More
□	[redacted]	700004857126	Aug 20, 2021 15:45:44	xcNodePlugin	[redacted]	DAG | Test | Patch Data | More
□	[redacted]_to_0_ [redacted]	700004856876	Aug 20, 2021 14:42:47	MergeIntoNode	[redacted]	DAG | Test | Patch Data | More
□	[redacted]_to_0_ [redacted]ccess	700004856875	Aug 20, 2021 14:42:38	FullDataInitCheckDone	[redacted]	DAG | Test | Patch Data | More

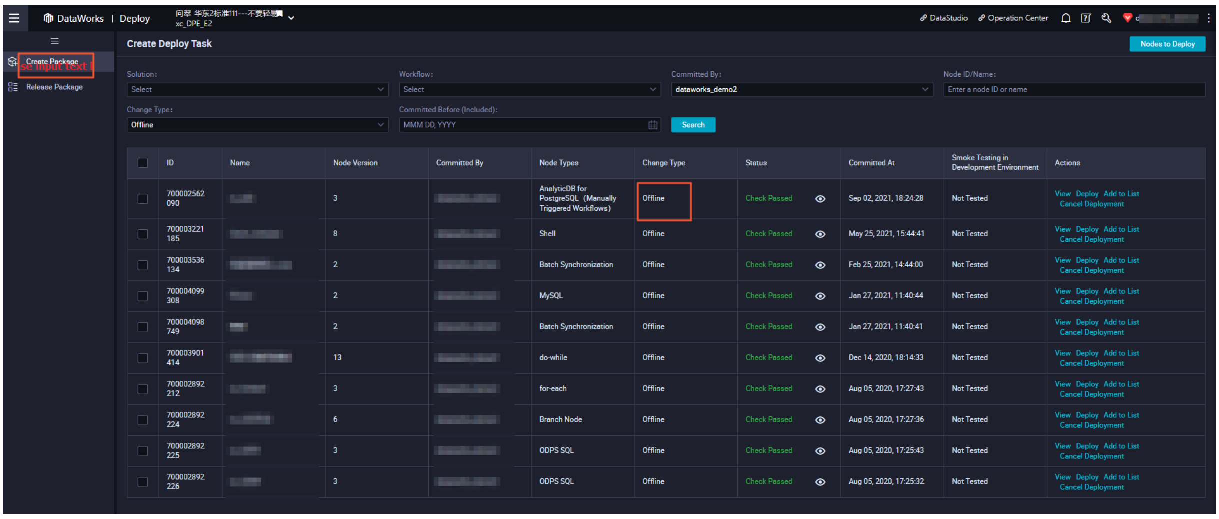
3.10. Node undeployment

This topic provides answers to some frequently asked questions about node undeployment.

- How do I undeploy a node?
- What do I do if a node that has subnodes fails to be undeployed?
- How do I check whether a node is undeployed from the production environment?
- How do I recover an undeployed node?

How do I undeploy a node?

After you delete a node in the development environment on the DataStudio page, a node undeployment record is generated on the Create Package page. Click Deploy in the Actions column of the record to deploy the undeployment operation. If the deployment operation is successful, the node is undeployed from the production environment.



Create Deploy Task

Solution: Select

Workflow: Select

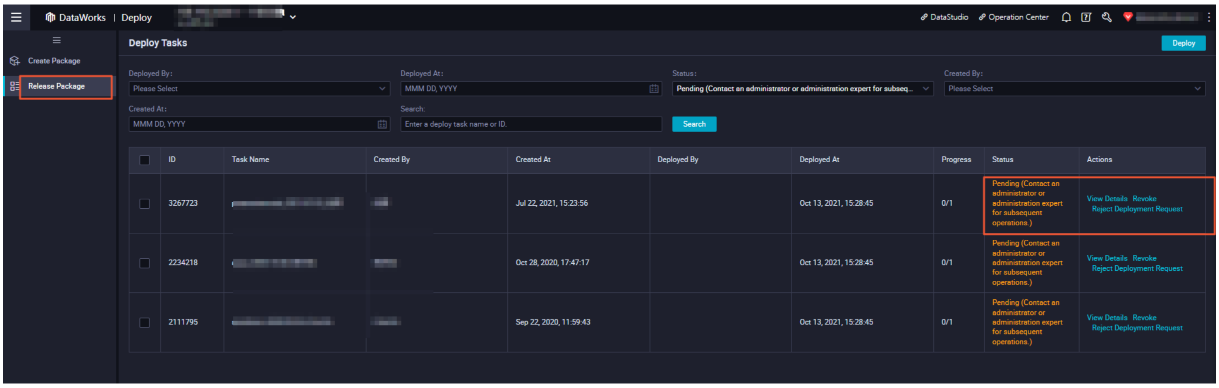
Committed By: dataworks_demo2

Node ID/Name: Enter a node ID or name

Change Type: Offline

Committed Before (Included): MMM DD, YYYY

☐	ID	Name	Node Version	Committed By	Node Types	Change Type	Status	Committed At	Smoke Testing in Development Environment	Actions
☐	700002562090	[redacted]	3	[redacted]	AnalyticDB for PostgreSQL (Manually Triggered Workflows)	Offline	Check Passed	Sep 02, 2021, 18:24:28	Not Tested	View Deploy Add to List Cancel Deployment
☐	700003221185	[redacted]	8	[redacted]	Shell	Offline	Check Passed	May 25, 2021, 15:44:41	Not Tested	View Deploy Add to List Cancel Deployment
☐	700003536134	[redacted]	2	[redacted]	Batch Synchronization	Offline	Check Passed	Feb 25, 2021, 14:44:00	Not Tested	View Deploy Add to List Cancel Deployment
☐	700004099308	[redacted]	2	[redacted]	MySQL	Offline	Check Passed	Jan 27, 2021, 11:40:44	Not Tested	View Deploy Add to List Cancel Deployment
☐	700004098749	[redacted]	2	[redacted]	Batch Synchronization	Offline	Check Passed	Jan 27, 2021, 11:40:41	Not Tested	View Deploy Add to List Cancel Deployment
☐	700003901414	[redacted]	13	[redacted]	do-while	Offline	Check Passed	Dec 14, 2020, 18:14:33	Not Tested	View Deploy Add to List Cancel Deployment
☐	700002892212	[redacted]	3	[redacted]	for-each	Offline	Check Passed	Aug 05, 2020, 17:27:43	Not Tested	View Deploy Add to List Cancel Deployment
☐	700002892224	[redacted]	6	[redacted]	Branch Node	Offline	Check Passed	Aug 05, 2020, 17:27:36	Not Tested	View Deploy Add to List Cancel Deployment
☐	700002892225	[redacted]	3	[redacted]	ODPS SQL	Offline	Check Passed	Aug 05, 2020, 17:25:43	Not Tested	View Deploy Add to List Cancel Deployment
☐	700002892226	[redacted]	3	[redacted]	ODPS SQL	Offline	Check Passed	Aug 05, 2020, 17:25:32	Not Tested	View Deploy Add to List Cancel Deployment



Deploy Tasks

Deployed By: Please Select

Deployed At: MMM DD, YYYY

Status: Pending (Contact an administrator or administration expert for subseq...

Created By: Please Select

Created At: MMM DD, YYYY

Search: Enter a deploy task name or ID.

☐	ID	Task Name	Created By	Created At	Deployed By	Deployed At	Progress	Status	Actions
☐	3267723	[redacted]	[redacted]	Jul 22, 2021, 15:23:56		Oct 13, 2021, 15:28:45	0/1	Pending (Contact an administrator or administration expert for subsequent operations.)	View Details Revoke Reject Deployment Request
☐	2234218	[redacted]	[redacted]	Oct 28, 2020, 17:47:17		Oct 13, 2021, 15:28:45	0/1	Pending (Contact an administrator or administration expert for subsequent operations.)	View Details Revoke Reject Deployment Request
☐	2111795	[redacted]	[redacted]	Sep 22, 2020, 11:59:43		Oct 13, 2021, 15:28:45	0/1	Pending (Contact an administrator or administration expert for subsequent operations.)	View Details Revoke Reject Deployment Request

Whether the deployment operation is successful depends on the permissions of the role of a user that performs this operation and the specified workflow. If the deployment operation fails, check the status of the deployment package on the Release Package page.

> **Note** Click Workspace Manage in the upper-right corner. On the User Management page, view the O&M personnel.

What do I do if a node that has subnodes fails to be undeployed?

You can undeploy a node only after no nodes depend on the node in the **development** and **production** environments.

> ? **Note** If you undeploy a node, exceptions may occur on the nodes that depend on this node. We recommend that you contact the owner of each node before you undeploy the current node.

Operation guidance:

1. The node you want to undeploy has a subnode that you created.

   Go to the **DataStudio** page, find the desired subnode, and then click the Properties tab on the right side. In the Dependencies section, delete the dependency between the subnode and its ancestor node that you want to undeploy and click **Save** and **Submit** (the dependency between the subnode and its ancestor node is deleted in the development environment). Then, click **Deploy** (the dependency between the subnode and its ancestor node is deleted in the production environment).

2. The node that you want to undeploy has a subnode that is not created by you.

   Search for the subnode in Operation Center in the **development environment** and **production environment**, and contact the owner of the subnode to change its ancestor node to a different node. Click **Submit** (the dependency between the subnode and its ancestor node is deleted in the development environment) and **Deploy** (the dependency between the subnode and its ancestor node is deleted in the production environment). Then, you can delete the node that you want to undeploy.

How do I check whether a node is undeployed from the production environment?

After you undeploy a node, find the node ID on the Cycle Task page in Operation Center in the development environment and production environment to check whether the node still exists.

> **Warning** Node undeployment is a high-risk operation. Proceed with caution.

How do I recover an undeployed node?

A deleted node is placed in a recycle bin. If you want to recover a deleted node, go to the recycle bin and recover the node.

> ? **Note** Before you recover a deleted node, make sure that the node undeployment operation is successful in both the development and production environments. If the error message shown in the preceding figure appears, you can refer to the operations described in the **How do I undeploy a node?** section in this topic to fix the error.

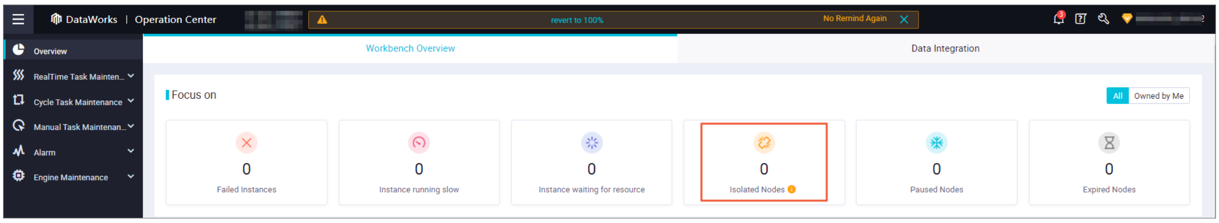
3.11. Isolated nodes

This topic provides answers to some frequently asked questions about isolated nodes.

- What is an isolated node?
- How can I fix an isolated node?

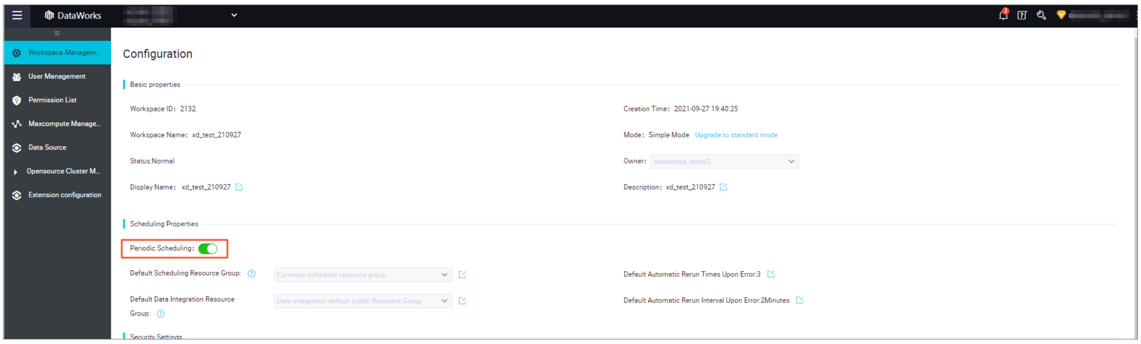
What is an isolated node?

An isolated node is a node that does not depend on any ancestor node. You can view its ancestor nodes on the Cycle Task or Cycle Instance page after you right-click the node and select Show Ancestor Nodes. This type of node cannot be run as scheduled. If an isolated node has too many descendant nodes, serious risks may arise. A node isolation alert is automatically reported if an isolated node is identified. We recommend that you handle the alert at the earliest opportunity.



How can I fix an isolated node?

1. Possible causes and solutions:
    - An isolated auto triggered node has no ancestor nodes and you cannot find its ancestor nodes on the Cycle Task or Cycle Instance page in the production environment in Operation Center. In this case, configure ancestor nodes for the isolated auto triggered node again, deploy the ancestor nodes, and then check whether the node is successfully deployed on the Cycle Task page in Operation Center.
    - Ancestor nodes are out of the validity period for scheduling. To resolve this issue, specify Validity Period in the Schedule section on the Properties tab of the DataStudio page.
    - The desired auto triggered node and its ancestor node reside in different workspaces and the Periodic Scheduling switch is turned off for the workspace in which the ancestor node resides. Contact the owner of the workspace to turn on the Periodic Scheduling switch or delete the cross-workspace node dependency.

      Configuration

      Basic properties

      Workspace ID：2132

      Creation Time：2021-09-27 19:40:25

      Workspace Name：xd_test_210927

      Mode：Simple Mode Upgrade to standard mode

      Status:Normal

      Owner：dataworks_demo2

      Display Name：xd_test_210927

      Description：xd_test_210927

      Scheduling Properties

      Periodic Scheduling：

      Default Scheduling Resource Group： Common scheduler resource group

      Default Automatic Rerun Times Upon Error:3

      Default Data Integration Resource Group： Data integration default public Resource Group

      Default Automatic Rerun Interval Upon Error:2Minutes

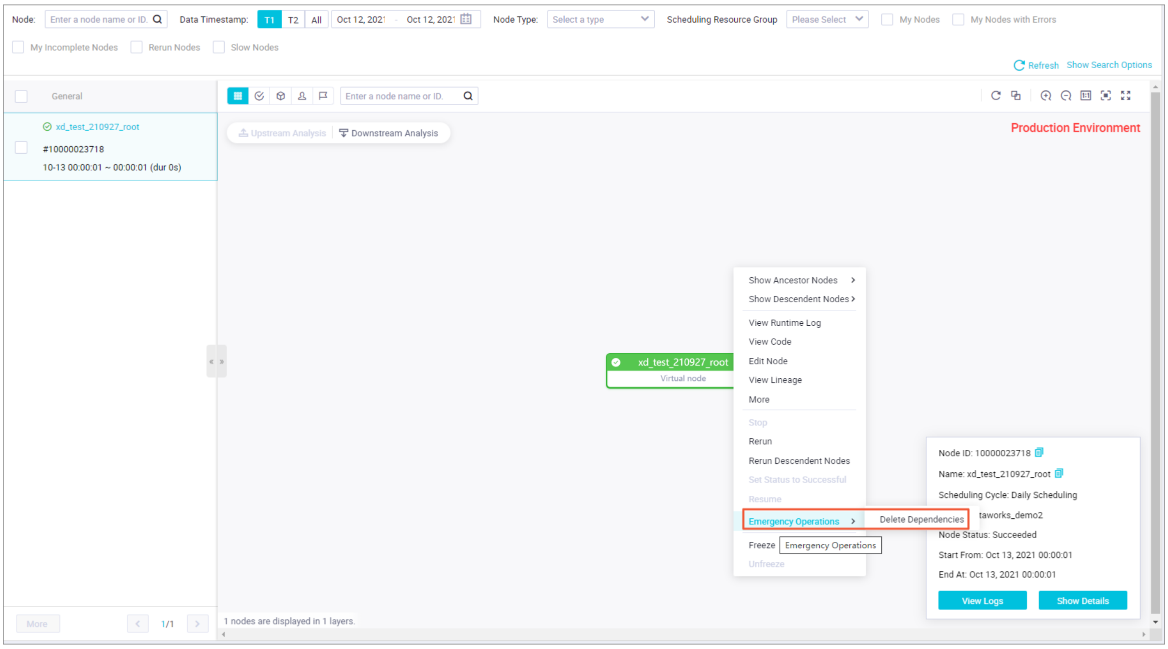
      Security Settings

    - The output name of the ancestor node is changed or deleted.

> Warning If you forcibly change the value of the Parent Node Output Name parameter, all descendant nodes may not be run as scheduled. Exercise caution when you perform this operation. Before you undeploy a node, delete the dependencies between this node and all its descendant nodes.

2. Workaround for the case that an isolated node has descendant nodes:

Descendant nodes of an isolated node cannot run. In an emergency situation, you can delete the dependency of your node on an isolated node if you confirm that your node can generate data as expected even if its dependency on the isolated node is deleted.



Node ID: 10000023718

Name: xd_test_210927_root

Scheduling Cycle: Daily Scheduling

taworks_demo2

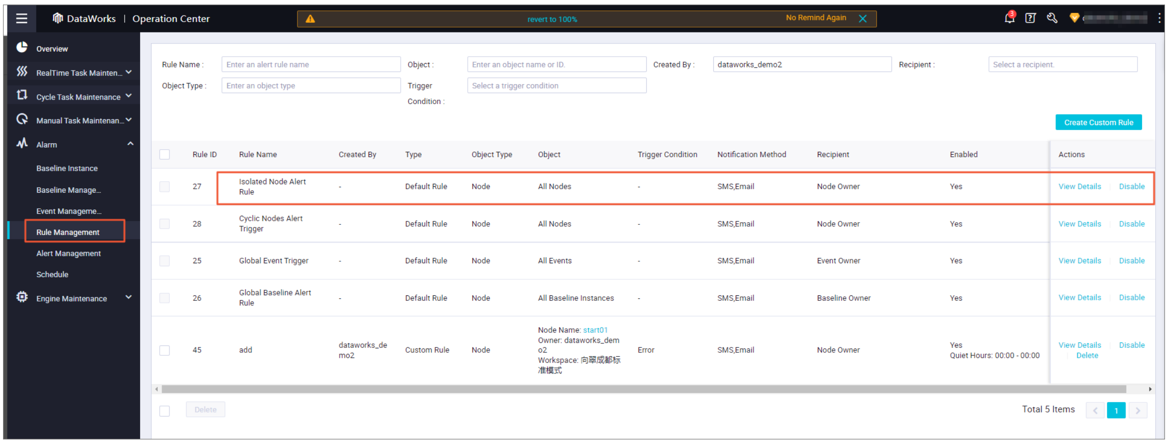
Node Status: Succeeded

Start From: Oct 13, 2021 00:00:01

End At: Oct 13, 2021 00:00:01

3. Change the node isolation alert recipient:

   Isolated Node Alert Rule is a default rule. By default, a node isolation alert is sent to the owner of the node. You can use the Alibaba Cloud account to enable or disable the rule. You can click View Details in the Actions column that corresponds to the node isolation alert rule to change the alert notification method and add other alert recipients.



Rule Name : Enter an alert rule name
Object : Enter an object name or ID.
Created By : dataworks_demo2
Recipient : Select a recipient.
Object Type : Enter an object type
Trigger Condition : Select a trigger condition

Rule ID	Rule Name	Created By	Type	Object Type	Object	Trigger Condition	Notification Method	Recipient	Enabled	Actions
27	Isolated Node Alert Rule	-	Default Rule	Node	All Nodes	-	SMS,Email	Node Owner	Yes	View Details Disable
28	Cyclic Nodes Alert Trigger	-	Default Rule	Node	All Nodes	-	SMS,Email	Node Owner	Yes	View Details Disable
25	Global Event Trigger	-	Default Rule	Node	All Events	-	SMS,Email	Event Owner	Yes	View Details Disable
26	Global Baseline Alert Rule	-	Default Rule	Node	All Baseline Instances	-	SMS,Email	Baseline Owner	Yes	View Details Disable
45	add	dataworks_demo2	Custom Rule	Node	Node Name: start01 Owner: dataworks_demo2 Workspace: 向翠成都标准模式	Error	SMS,Email	Node Owner	Yes Quiet Hours: 00:00 - 00:00	View Details Disable Delete

Total 5 Items 1

3.12. Other FAQ

This topic provides answers to other frequently asked questions about Operation Center.

- FAQ about nodes and node instances
  - What is the relationship between auto triggered nodes and auto triggered node instances, data backfill instances, or test instances?
- FAQ about resource groups

  - How long are the logs of resource groups for scheduling and node instances that are run on such resource groups retained?
  - Why am I unable to perform big data computing on a resource group for scheduling?
- FAQ about other items
  - How do I adjust the priority of a node instance?
  - How do I view the priority of a node instance?

What is the relationship between auto triggered nodes and auto triggered node instances, data backfill instances, or test instances?

DataWorks generates instances that are scheduled to run for auto triggered nodes every night based on the time when the auto triggered nodes are committed and the instance generation modes that you configure for the auto triggered nodes. You can perform operations on the auto triggered nodes to generate and run data backfill instances and test instances for the nodes.

> ? **Note** You can set the Instance Generation Mode parameter to Immediately After Deployment for an auto triggered node. After you deploy the auto triggered node, DataWorks generates an auto triggered node instance for the node, and you can view the instance on the Cycle Instance page of Operation Center.

How long are the logs of resource groups for scheduling and node instances that are run on such resource groups retained?

The logs of the shared resource group for scheduling are retained for one week, and the node instances that are run on this type of resource group are retained for one month.

The logs of an exclusive resource group for scheduling and the node instances that are run on this type of resource group are retained for one month.

> ? **Note** If the size of logs generated for node instances that are run every day exceeds 3 MB, the system clears the logs.

Why am I unable to perform big data computing on a resource group for scheduling?

Resource groups for scheduling are used to schedule nodes and provide only limited resources. Therefore, resource groups for scheduling are not suitable for big data computing. MaxCompute can process large amounts of data. We recommend that you use MaxCompute for big data computing.

How do I adjust the priority of a node instance?

To adjust the priority of a node instance, adjust the priority of the baseline with which the node instance is associated. The priority of a baseline can be set to 1, 3, 5, 7, or 8. A larger value indicates a higher priority. The higher the priority of a baseline, the faster DataWorks schedules the node instance that is associated with the baseline.

Adjust the priority of a baseline with which node instances are associated to adjust the priorities of the node instances.

1. Go to the **Operation Center** page.

2. In the left-side navigation pane, choose **Alarm > Baseline Management**.
3. On the Baseline Management page, find the baseline whose priority you want to adjust and click **Change** in the Actions column.
4. In the **Change Baseline** dialog box, change the value of the **Priority** parameter.
5. Click **Complete**.

How do I view the priority of a node instance?

To view the priority of a node instance, perform the following steps: Go to the Operation Center page. In the left-side navigation pane, choose **Cycle Task Maintenance > Cycle Instance**. On the page that appears, find the node instance whose priority you want to view and click the node instance name. In the directed acyclic graph (DAG) of the node instance, click Show Details. On the General tab, view the priority of the node instance.

4.DataStudio

This topic provides answers to some frequently asked questions about DataStudio.

- Resources
  - Which kind of resource group can I use when I reference a third-party package in a PyODPS node?
  - How do I reference a resource in a node?
  - How do I download a resource that is uploaded to DataWorks?
  - How do I upload a resource whose size is greater than 30 MB?
  - How do I use a resource that is uploaded to DataWorks by using odpscmd?
  - How do I upload a JAR package on my on-premises machine to DataWorks as a JAR resource and reference the uploaded resource in a node?
  - How do I use a MaxCompute table in DataWorks?
- PyODPS
  - Can a Python resource call another Python resource?
  - Can PyODPS call custom functions to use third-party packages?
  - When I call a pickle file in a PyODPS 3 node, the following error message appears: _pickle.UnpicklingError: invalid load key, '\xef.. What do I do?
  - How do I delete a MaxCompute resource?
- Nodes and workflows
  - How do I recover a node that is deleted?
  - What is the impact on the instances of a node after the node is deleted?
  - How do I view the versions of a node?
  - How do I check whether a node is committed?
  - After a node is modified and committed and deployed to the production environment, is the existing faulty node in the production environment overwritten?
  - How do I export the code of a node?
  - Can I configure properties for all nodes in a workflow at a time?
  - How do I clone a workflow?
- Tables
  - How do I create a table in a visualized manner?
  - How do I add fields to a table that is in the production environment?
  - How do I delete a table?
  - How do I upload data from my on-premises machine to a MaxCompute table?
  -
  - How do I query data that is in the production environment from the development environment on the DataStudio page?
  - How do I control whether the queried table data can be downloaded?
  - How do I download more than 10,000 data records?
  -

- Operational logs and retention period of operational logs
  - How do I query historical operational logs on the DataStudio page?
  - How long are operational logs on the DataStudio page retained?
- Batch operations
  - How do I perform operations on multiple nodes, resources, or functions at a time?
  - How do I change resource groups for scheduling for multiple nodes in a workflow at a time on the DataStudio page?
- Power BI connection to MaxCompute

  What do I do if an error is reported when I connect Power BI to MaxCompute?
- API calls
  - When I call a DataWorks API operation, the following error message appears: access is forbidden. Please first activate DataWorks Enterprise Edition or Flagship Edition. What do I do?
  -
- Other items
  - How do I disable the MaxCompute Query Acceleration (MCQA) feature if I want to obtain the instance ID that is used to download more than 10,000 data records?

Which kind of resource group can I use when I reference a third-party package in a PyODPS node?

Use an exclusive resource group for scheduling. For more information, see Use a PyODPS node to reference a third-party package.

How do I control whether the queried table data can be downloaded?

Before you download data from DataWorks, you must enable the Download SELECT Query Result feature. If no download entry point is available, the Download SELECT Query Result feature is disabled for your workspace. If you use a RAM user and need to use this feature, contact the owner of the Alibaba Cloud account or the workspace administrator to enable this feature in the Workspace Settings panel or on the Workspace Management page.

After you query data in DataStudio, the download entry point is displayed in the lower-right corner of the query result tab, as shown in the following figure.

You can download only a maximum of 10,000 data records from DataStudio due to the limits of the compute engine.

How do I download more than 10,000 data records?

Use a Tunnel command of MaxCompute. For more information, see Use SQLTask and Tunnel to export a large amount of data.

How do I reference a resource in a node?

Find the resource that you want to reference in the node in the Scheduled Workflow pane, right-click the resource name, and then select **Insert Resource Path**.

How do I download a resource that is uploaded to DataWorks?

Find the resource that you want to download in the Scheduled Workflow pane, right-click the resource name, and then select **View Versions**. In the Versions dialog box, click Download in the Actions column.

How do I upload a resource whose size is greater than 30 MB?

Use a Tunnel command to upload the resource. Then, add the resource to DataStudio in the MaxCompute Resources pane for future use. For more information, see How do I use a resource that is uploaded to DataWorks by using odpscmd?.

How do I use a resource that is uploaded to DataWorks by using odpscmd?

If you want to use a resource that is uploaded to DataWorks by using odpscmd, add the resource to DataStudio in the MaxCompute Resources pane.

How do I upload a JAR package on my on-premises machine to DataWorks as a JAR resource and reference the uploaded resource in a node?

Upload the JAR package to DataWorks on the DataStudio page as a JAR resource. If you want to reference the resource in a node, find the resource in the Scheduled Workflow pane, right-click the resource name, and then select **Insert Resource Path**. A comment is automatically added at the beginning of the code for the node, and the node can directly reference the resource in its code based on the resource name.

For example, you want to reference the resource test.jar in a Shell node. After you select Insert Resource Path, the comment `##@resource_reference{"test.jar"}` is automatically added at the beginning of the code for the Shell node.

How do I use a MaxCompute table in DataWorks?

You cannot use the codeless user interface (UI) to upload a MaxCompute table to DataWorks. If you want to use a MaxCompute table in DataWorks, perform the following steps:

1. On the DataStudio page of DataWorks, create a file resource that has the same name as the MaxCompute table and upload the file. In this example, the userlog3.txt file is uploaded.

   > **Note**
   >
   > Do not select Upload to MaxCompute.

2. After you upload the file, execute a statement on odpscmd to add the MaxCompute table resource to DataWorks. In this example, the statement `add table userlog3 -f;` is executed.

3. Select the uploaded file resource to use the resource.

Can a Python resource call another Python resource?

A Python resource can call another Python resource in the same workspace.

Can PyODPS call custom functions to use third-party packages?

If you do not want to use the **map** method of DataFrame to call the `test` function, you can use PyODPS to call custom functions to use third-party packages. For more information, see Reference a third-party package in a PyODPS node.

When I call a pickle file in a PyODPS 3 node, the following error message appears: `_pickle.UnpicklingError: invalid load key, '\xef.`. What do I do?

Check whether the code of your PyODPS 3 node contains special characters. If the code contains special characters, compress the code into a ZIP package, upload the package to DataWorks, and then decompress the package to call the pickle file.

How do I delete a MaxCompute resource?

To delete a MaxCompute resource in a workspace in basic mode, right-click the resource name and select Delete to delete it. To delete a MaxCompute resource in a workspace in standard mode, you must delete the resource in the development environment and then delete it in the production environment again. The following example shows how to delete a MaxCompute resource in the development and production environments.

> **Note**
>
> In a DataWorks workspace in standard mode, the development environment is isolated from the production environment. If you delete a resource on the DataStudio page, the resource is deleted only from the development environment. The same resource is deleted from the production environment only after you deploy the delete operation to the production environment.

1. Delete a resource from the development environment. In the desired workflow, choose **MaxCompute > Resource**, right-click the resource name that you want to delete, and then select **Delete**. In the Delete dialog box, click **OK**.

2. Delete a resource from the production environment. A resource can be deleted from the production environment only after the delete operation of the resource is deployed to the production environment. On the **DataStudio** page, click **Deploy** in the upper-right corner. On the Create Deploy Task page, set Change Type to Offline, find the package of the resource that is deleted in the previous step, and click **Deploy** in the **Actions** column. In the Create Deploy Task dialog box, click **Deploy**.

   After you click Deploy, the resource is deleted from the production environment.

How do I recover a node that is deleted?

On the DataStudio page, click the Recycle Bin icon in the left-side navigation pane. In the Recycle Bin pane, find the node that you want to recover, right-click the node name, and then select Restore.

How do I view the versions of a node?

Find the node whose versions you want to view in the Scheduled Workflow pane and double-click the node name to go to the configuration tab of the node. Then, click Versions in the right-side navigation pane. On the Versions tab, you can view the versions of the node.

> **Notice**
>
> A version is generated only after you commit the code.

How do I clone a workflow?

Use a node group. For more information, see Create and reference a node group.

How do I export the code of a node?

Use Migration Assistant. For more information, see Overview.

How do I check whether a node is committed?

If you want to check whether a node is committed, find the desired workflow in the **Scheduled Workflow** pane and expand the workflow to view the status of each node in this workflow. If the icon is displayed on the left side of a node, the node is committed. Otherwise, the node is not committed.

Can I configure properties for all nodes in a workflow at a time?

No, you cannot configure properties for all nodes in a workflow. In DataWorks, you are not allowed to configure properties for a workflow. If a workflow contains multiple nodes, you must configure properties for the nodes one by one. For example, if a workflow contains 20 nodes, you must configure properties for these nodes one by one.

What is the impact on the instances of a node after the node is deleted?

The scheduling system generates one or more instances for a node every day based on the time properties of the node. If the node is deleted after it is run for a period of time, its instances are retained. However, the instances will fail to run after the node is deleted. This is because the required code is unavailable.

After a node is modified and committed and deployed to the production environment, is the existing faulty node in the production environment overwritten?

No, the existing faulty node is not overwritten. The updated code is used to run new node instances that are not run, and the existing node instances are retained. If scheduling properties are modified, the modified configurations apply only to the new node instances.

How do I create a table in a visualized manner?

Go to the DataStudio page and click the Workspace Tables icon in the left-side navigation pane. In the Workspace Tables pane, create a table.

How do I add fields to a table that is in the production environment?

If you use an Alibaba Cloud account, add fields to the table in the **Workspace Tables** pane of the DataStudio page and commit the table to the production environment.

If you use a RAM user, you must request the permissions of the O&M engineer or workspace administrator role for the RAM user, use the RAM user to add fields to the table in the **Workspace Tables** pane of the DataStudio page, and then commit the table to the production environment.

How do I delete a table?

You can delete a table from the development environment on the DataStudio page.

To delete a table from the production environment, use one of the following methods:

- Go to Data Map and delete the table on the My Data tab.
- Create an ODPS SQL node, and enter and execute the DROP statement on the configuration tab of the node. For more information about how to create an ODPS SQL node, see Create an ODPS SQL node. For more information about the syntax of the DROP statement, see Table operations.

How do I upload data from my on-premises machine to a MaxCompute table?

Go to the DataStudio page and use the Import feature in the Scheduled Workflow pane to import the data.

When I create a table in a workspace with which an E-MapReduce (EMR) compute engine instance is associated, the following error message appears: `call emr exception`. What do I do?

- Possible cause:

  Security settings are not configured for the security group to which your EMR cluster belongs. Before you associate an EMR compute engine instance with your workspace, add the following rules to the security group of the ECS instance that hosts your EMR cluster. Otherwise, the preceding error message may appear.

  - Action: Allow
  - Protocol type: Custom TCP
  - Port range: 8898/8898
  - Authorization object: 100.104.0.0/16

- Solution:

  Check the security settings of the security group of the ECS instance that hosts your EMR cluster. If the security settings do not include the preceding rules, add the rules to the security group.

How do I query data that is in the production environment from the development environment on the DataStudio page?

In a workspace in standard mode, if you want to query data that is in the production environment from the development environment on the DataStudio page, specify the table whose data you want to query in the **Project name.Table name** format.

In a workspace that is upgraded from the basic mode to the standard mode, if you want to query data that is in the production environment from the development environment on the DataStudio page, you must request the permissions of the producer role first and specify the table whose data you want to query in the **Project name.Table name** format. For more information about how to request the permissions, see Request permissions on tables.

How do I query historical operational logs on the DataStudio page?

Click the **Operational history** icon in the left-side navigation pane of the DataStudio page. In the Operational history pane, you can view the historical operational logs.

How long are operational logs on the DataStudio page retained?

By default, operational logs on the DataStudio page are retained for three days.

> **Note** For more information about the retention period of logs and instances in Operation Center of the production environment, see How long are the logs of resource groups for scheduling and node instances that are run on such resource groups retained?.

How do I perform operations on multiple nodes, resources, or functions at a time?

Go to the **DataStudio** page and click the **Batch Operation** icon in the Scheduled Workflow pane. On the Batch Operation-Data Development tab, you can perform the desired operation on multiple nodes, resources, or functions at a time. Then, you can commit the objects on which you perform the operation at a time and deploy the objects on the Create Deploy Task page to make the modifications take effect.

How do I change resource groups for scheduling for multiple nodes in a workflow at a time on the DataStudio page?

Find the desired workflow on the DataStudio page, move the pointer over the workflow name, and then click the icon on the right side of the workflow name. On the tab that appears, select the nodes for which you want to change resource groups for scheduling and click Switch Resource Groups. After you change the resource groups for the nodes, click the Submit icon in the top toolbar to commit the nodes at a time. Then, deploy the nodes on the Create Deploy Task page at a time to make the modifications take effect in the production environment.

What do I do if an error is reported when I connect Power BI to MaxCompute?

MaxCompute cannot be connected to Power BI. We recommend that you connect Hologres instead of Power BI to MaxCompute. For more information, see Endpoints for connecting to Hologres.

When I call a DataWorks API operation, the following error message appears: `access is forbidden. Please first activate DataWorks Enterprise Edition or Flagship Edition`. What do I do?

Activate DataWorks Enterprise Edition or Ultimate Edition. For more information, see Overview.

How do I disable the MaxCompute Query Acceleration (MCQA) feature if I want to obtain the instance ID that is used to download more than 10,000 data records?

To obtain the instance ID that is required to download the data records, you must disable the MCQA feature.

> **Note** DataWorks allows you to download only a total of 10,000 data records. If you want to download more than 10,000 data records of an ODPS SQL node, you must use a Tunnel command.

Add `set odps.mcqa.disable=true;` to the code of the ODPS SQL node and execute this statement together with other SELECT statements.

5.Accounts and permissions

5.1. User permission management

This topic provides answers to some frequently asked questions about user permission management.

- Why are no workspaces found after I log on to the DataWorks console by using a RAM user?
- How do I add a RAM user to a workspace?
- How do I grant a RAM user the permissions to create a DataWorks workspace?
- How do I use an Alibaba Cloud account to attach the AliyunDataWorksFullAccess policy to a RAM user?
- How do I create a custom MaxCompute role that has only the query permissions?
- How do I configure the phone number and email address for a RAM user?
- What do I need to take note of if I want to remove a RAM user?

Why are no workspaces found after I log on to the DataWorks console by using a RAM user?

The DataWorks console displays only the workspaces to which your RAM user is added. To view a workspace by using a RAM user, add the RAM user to the workspace first.

How do I add a RAM user to a workspace?

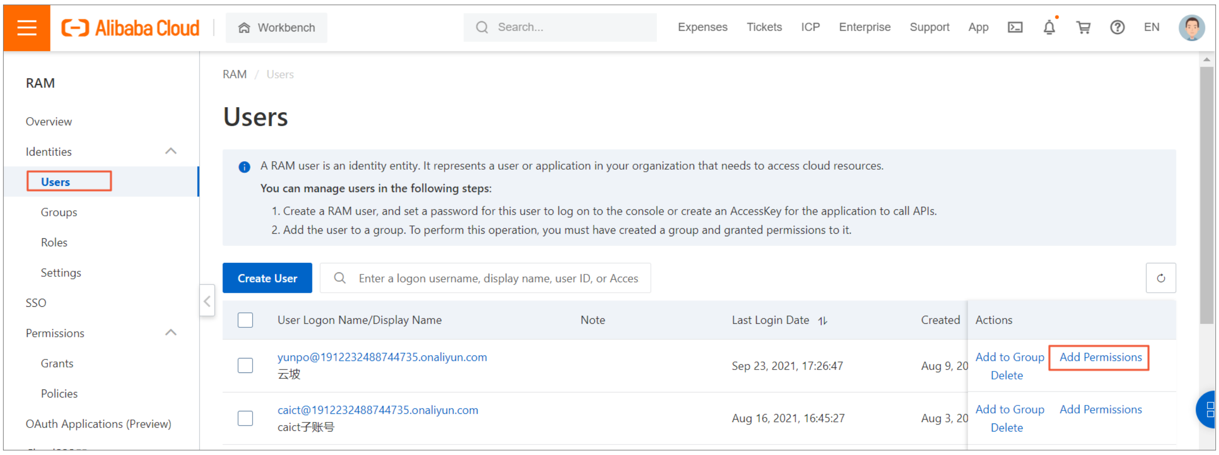
A workspace administrator can add a RAM user to a workspace on the Workspace Management page. For more information, see Add workspace members and assign roles to them.

How do I grant a RAM user the permissions to create a DataWorks workspace?

Use your Alibaba Cloud account to attach the AliyunDataWorksFullAccess policy to a RAM user on the Users page in the RAM console.

How do I use an Alibaba Cloud account to attach the AliyunDataWorksFullAccess policy to a RAM user?

Perform the steps shown in the following figures to attach the AliyunDataWorksFullAccess policy to a RAM user on the Users page in the RAM console.



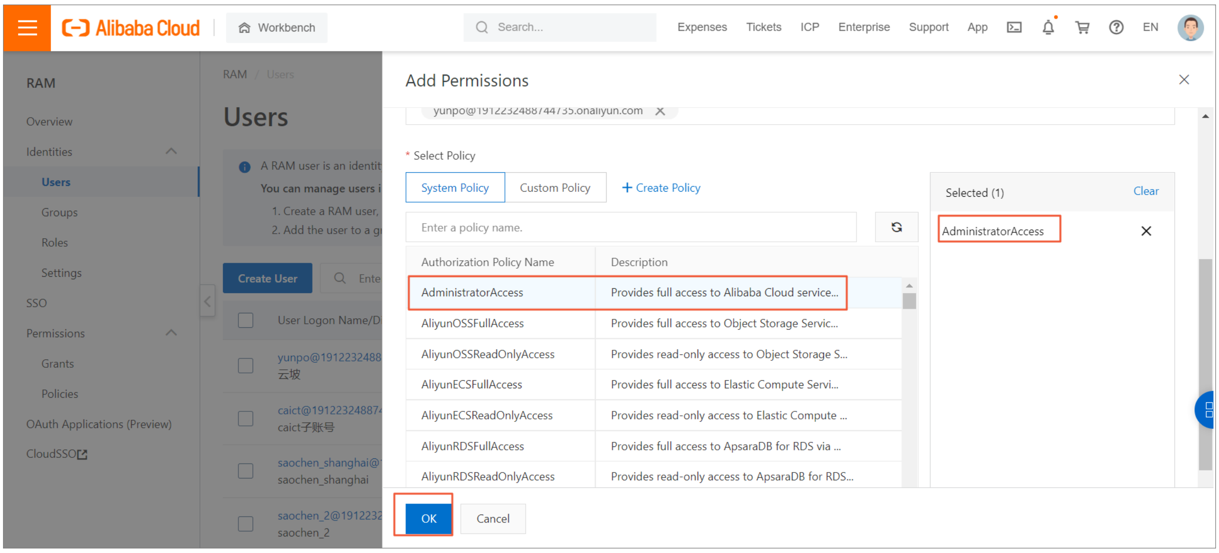
Users

A RAM user is an identity entity. It represents a user or application in your organization that needs to access cloud resources.

You can manage users in the following steps:

1. Create a RAM user, and set a password for this user to log on to the console or create an AccessKey for the application to call APIs.
2. Add the user to a group. To perform this operation, you must have created a group and granted permissions to it.

□	User Logon Name/Display Name	Note	Last Login Date	Created	Actions
□	yunpo@1912232488744735.onaliyun.com 云坡		Sep 23, 2021, 17:26:47	Aug 9, 20	Add to Group Add Permissions Delete
□	caict@1912232488744735.onaliyun.com caict子账号		Aug 16, 2021, 16:45:27	Aug 3, 20	Add to Group Add Permissions Delete



* Select Policy

Authorization Policy Name	Description
AdministratorAccess	Provides full access to Alibaba Cloud service...
AliyunOSSFullAccess	Provides full access to Object Storage Servic...
AliyunOSSReadOnlyAccess	Provides read-only access to Object Storage S...
AliyunECSFullAccess	Provides full access to Elastic Compute Servi...
AliyunECSReadOnlyAccess	Provides read-only access to Elastic Compute ...
AliyunRDSFullAccess	Provides full access to ApsaraDB for RDS via ...
AliyunRDSReadOnlyAccess	Provides read-only access to ApsaraDB for RDS...

How do I create a custom MaxCompute role that has only the query permissions?

For more information, see Create a custom MaxCompute role that has only the query permissions.

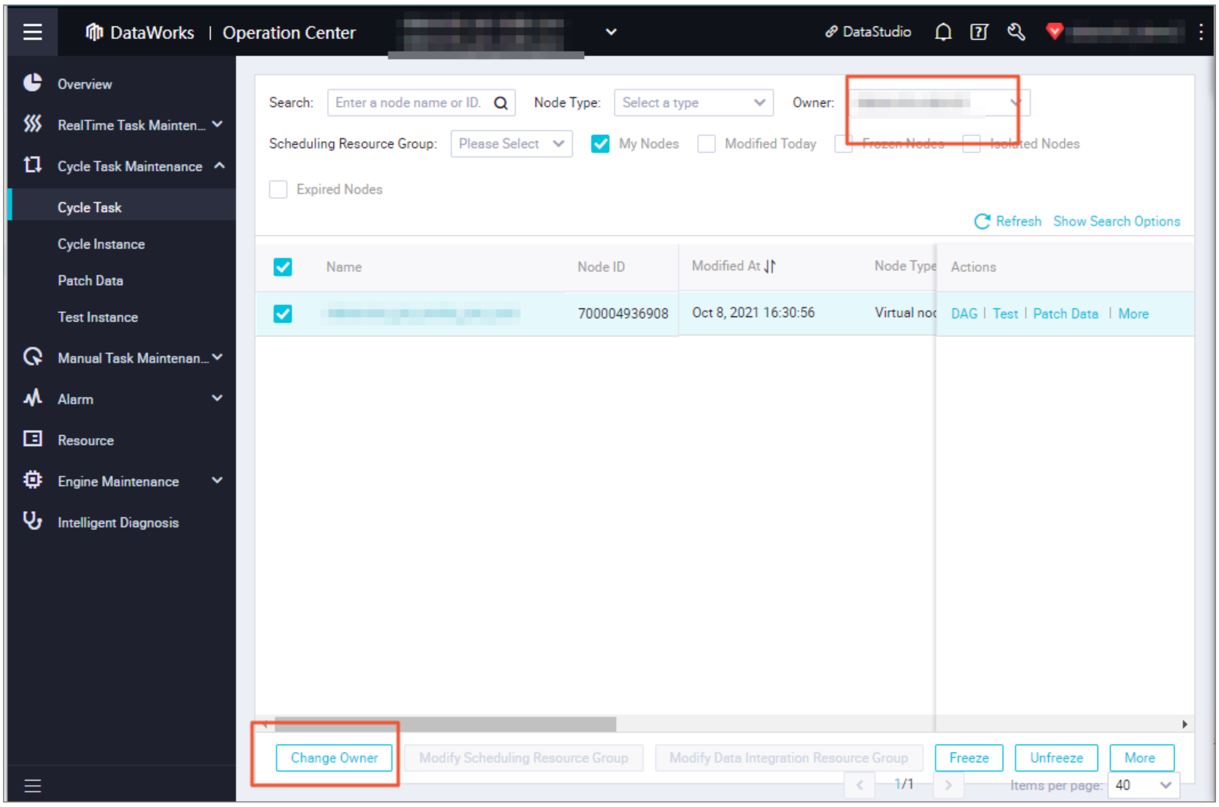
How do I configure the phone number and email address for a RAM user?

If you want to receive alert notifications by using your RAM user, use your Alibaba Cloud account to configure the phone number and email address for the RAM user on the Alert Contacts page in the DataWorks console. For more information, see Configure and view alert contacts.

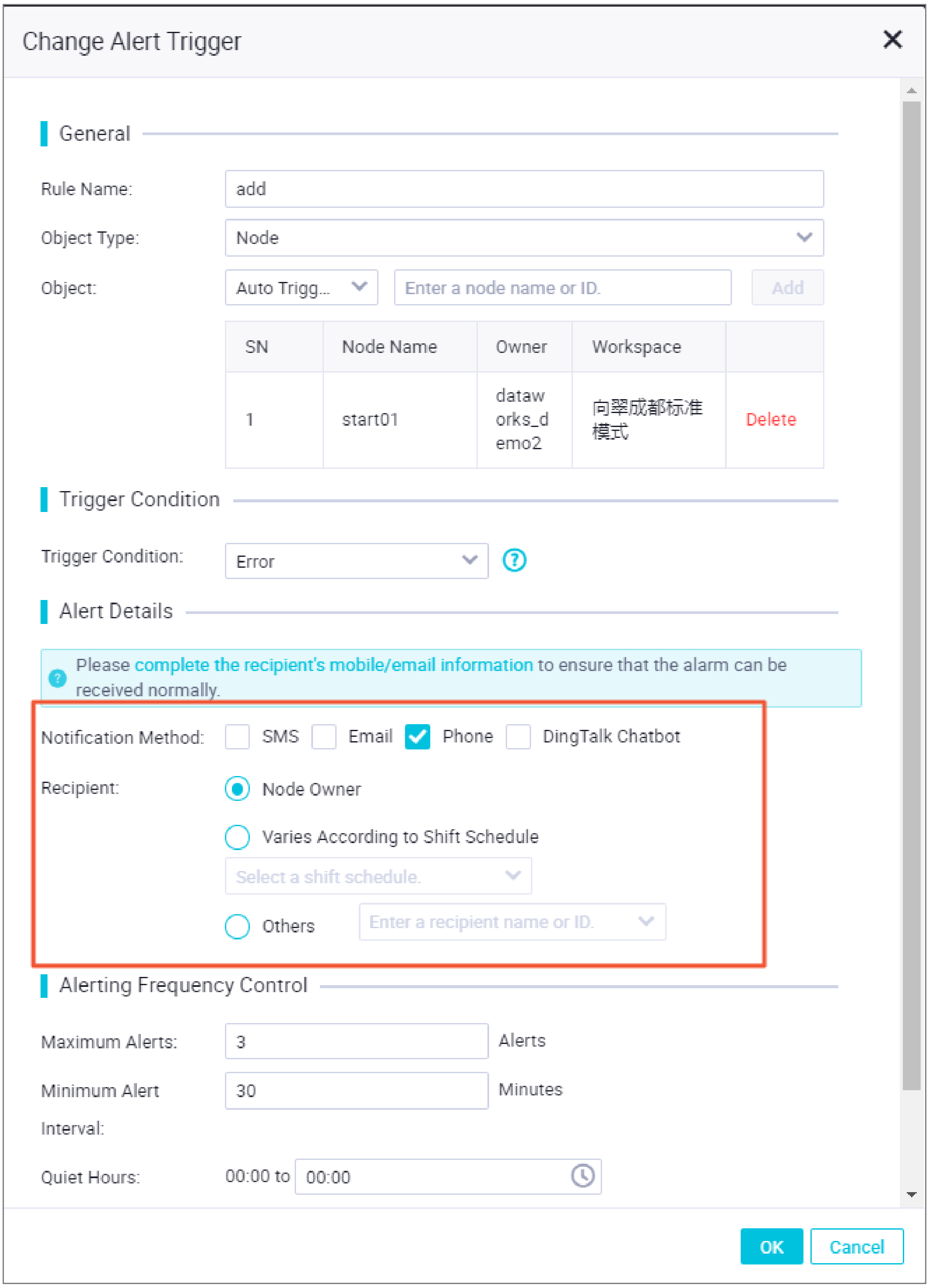
What do I need to take note of if I want to remove a RAM user?

Before a RAM user is removed, change the owner of related nodes in the DataWorks console and then remove the RAM user in the RAM console. If you do not change the owner first, errors may occur.

The following figure shows how to change the owner of multiple nodes at a time.



Warning After you change the owner, update the recipient configurations of related alert rules at the earliest opportunity. The configurations include the shift schedule.



Change Alert Trigger

General

Rule Name: add

Object Type: Node

Object: Auto Trigg... Enter a node name or ID. Add

SN	Node Name	Owner	Workspace	
1	start01	dataworks_demo2	向翠成都标准模式	Delete

Trigger Condition

Trigger Condition: Error

Alert Details

Please complete the recipient's mobile/email information to ensure that the alarm can be received normally.

Notification Method: □ SMS □ Email ☑ Phone □ DingTalk Chatbot

Recipient: ● Node Owner

○ Varies According to Shift Schedule

Select a shift schedule.

○ Others Enter a recipient name or ID.

Alerting Frequency Control

Maximum Alerts: 3 Alerts

Minimum Alert Interval: 30 Minutes

Quiet Hours: 00:00 to 00:00

OK Cancel

5.2. Operation permission management

This topic provides answers to some frequently asked questions about operation permission management.

How do I grant users in a workspace the permissions on service modules?

You can assign built-in roles to RAM users to control their permissions on service modules based on your business scenarios. You can also assign custom workspace-level roles to the RAM users to control their read/write permissions on service modules. For more information about the permissions of each built-in role, see Permissions of built-in workspace-level roles. For more information about custom workspace-level roles, see Manage workspace-level roles and members.

How do I grant users in a workspace the operation permissions on compute engine instances?

After you assign a workspace-level role to a user, the operation permissions granted to the user are based on the compute engine type and compute engine configurations.

- Logic of operation permissions on MaxCompute compute engine instances:
    - The DataWorks built-in roles and the roles in a MaxCompute project in the development environment have a permission mapping. By default, a DataWorks built-in role has all the permissions its mapped MaxCompute project role has on MaxCompute compute engine instances in the development environment.

- The DataWorks built-in roles and the roles in a MaxCompute project in the production environment do not have a permission mapping. A DataWorks built-in role cannot directly manage resources of a MaxCompute project in the production environment.

  > **Note** For example, a user that is assigned the Workspace Manager or Development role has permissions on most service modules and all the permissions on a workspace in the development environment (a MaxCompute project in the development environment). By default, the user that is assigned the Workspace Manager or Development role does not have the permissions on the same workspace in the production environment (the same MaxCompute project in the production environment). If a RAM user wants to access a table in the production environment from the development environment, you must apply for the operation permissions on the table for the RAM user in Data Map. For more information, see Request permissions.

Node running environment	Scenario
The node is run in DataStudio (in the development environment).	■ Scenario 1: Use an Alibaba Cloud account or a RAM user to run the `select col1 from tablename` command to access a table in the development environment. Specify the table name in the following format: *projectname_dev.tablename*. ■ Scenario 2: Use an Alibaba Cloud account or a RAM user to run the `select col1 from projectname.tablename` command to access a table in the production environment. Specify the table name in the following format: *projectname.tablename*. **Note** By default, a RAM user that is not selected when you associate a MaxCompute compute engine instance with a workspace does not have permissions to access data in the production environment. If you want to use the RAM user to access data in the production environment, you must apply for permissions in Data Map.
The node is run in Operation Center (in the production environment).	Scenario: Use the account that is selected when you associate a MaxCompute compute engine instance with a workspace to run the `select col1 from tablename` command to access a table in the production environment. Specify the table name in the following format: *projectname.tablename*.

- Logic and description of operation permissions on E-MapReduce (EMR) compute engine instances:

  - Logic: If your workspace is associated with an EMR compute engine instance, the permissions of a built-in role on DataWorks service modules depend on the permissions of the role. The permissions of the built-in role on the compute engine instance are the same as the permissions of the account that is selected when the compute engine instance is associated with the workspace.

Mode	Environment	Account in use	How it works
Shortcut mode	The node is run in DataStudio (in the development environment).	Hadoop user	
	The node is run in Operation Center (in the production environment).		
Security mode	The node is run in DataStudio (in the development environment).	The account that you selected for the development environment when you configure the compute engine	You can configure the Lightweight Directory Access Protocol (LDAP) permission mapping for members in a DataWorks workspace to manage the permissions of a RAM user on EMR features when the RAM user uses DataWorks. When you use an Alibaba Cloud account or a RAM user to commit code in DataWorks, the user that has the same name in EMR will run the node. **Note** For more information about the permission mapping between DataWorks members and EMR users, see Associate an EMR cluster with a DataWorks workspace.
	The node is run in Operation Center (in the production environment).	The account that you selected for the production environment when you configure the compute engine	

  - Permission control: You can use EMR Ranger to manage the permissions of each user in an EMR compute engine instance. This ensures that Alibaba Cloud accounts, node owners, or RAM users have different data permissions when they run EMR nodes in DataWorks.

- Logic of operation permissions on other compute engine instances:

If you associate a workspace with a compute engine instance other than a MaxCompute or EMR compute engine instance, whether the node that you want to run in DataStudio can use the compute engine resources is determined by the account that is selected when you associate the compute engine instance with the workspace.

How do I allow access to the DataWorks console only from the internal network of an enterprise?

If you want to allow access to the DataWorks console only from the internal network of an enterprise, log on to the RAM console and configure a security policy to allow access only from the public IP addresses that are mapped to the private IP addresses of the enterprise.

For more information, see Configure security policies for RAM users.

5.3. ActionTrail

This topic describes the answers to the frequently asked questions about ActionTrail.

- How do I retrieve events that record the operations performed by users in DataWorks, such as the operation to download data from the DataWorks console?
- What settings can I configure for important data in advance to trace data leaks?
- How can I audit permissions that are granted to a MaxCompute table?
- 表数据，How do I restore a deleted node?
- How do I compare node versions and roll back to an earlier version?

How do I retrieve events that record the operations performed by users in DataWorks, such as the operation to download data from the DataWorks console?

DataWorks is integrated with ActionTrail. This allows you to view and retrieve DataWorks behavioral events of your Alibaba Cloud account over the last 90 days in the ActionTrail console. 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. For more information, see Use ActionTrail to query behavior events.

What settings can I configure for important data in advance to trace data leaks?

To protect important data, enable the data watermark feature on the Data Masking page of Data Security Guard so that you can trace the users who may cause the data leaks. For more information, see Create a data masking rule.

How can I audit permissions that are granted to a MaxCompute table?

Go to Security Center. On the Permission audit tab of the Data access control page, check the members who have the permissions on a table and the validity period of the permissions. You can also revoke the permissions.Audit permissions

How do I restore a deleted node?

You can restore a node that is recently deleted in the Recycle Bin pane in DataStudio. Take note that DataWorks generates a new ID for the restored node.

How do I compare node versions and roll back to an earlier version?

Double-click the node on the DataStudio page. In the right-side pane, click **Versions** to compare versions and roll back to an earlier version. For more information, see Versions.

6.Resource groups

6.1. Exclusive resource groups

This topic provides answers to some frequently asked questions about exclusive resources groups of DataWorks.

- Use scenarios
  - In which scenario do I need to use an exclusive resource group for scheduling?
  - In which scenario do I need to use an exclusive resource group for Data Integration?
- Network environments
  - What conditions must be met before an exclusive resource group can access a data source that resides in a VPC?
  - How do I view the network environment of a data source?
  - What information about an exclusive resource group must be added to the whitelist of a data source that the exclusive resource group needs to access?
- Renewal and configuration modification
  - How do I renew a resource group?
  - How do I change the specifications of an exclusive resource group? How do I scale in or out an exclusive resource group?
- Use instructions
  - How do I change the resource group for scheduling that is used to run a node?
  - How do I change the resource group for Data Integration that is used to run a Data Integration node?
  - How do I associate an exclusive resource group with or disassociate an exclusive resource group from a workspace?

In which scenario do I need to use an exclusive resource group for scheduling?

If a node that is not used for data integration needs to access a data source that resides in a virtual private cloud (VPC) or a data source that is configured with a whitelist, you must use an exclusive resource group for scheduling. For more information, see Create and use an exclusive resource group for scheduling.

In which scenario do I need to use an exclusive resource group for Data Integration?

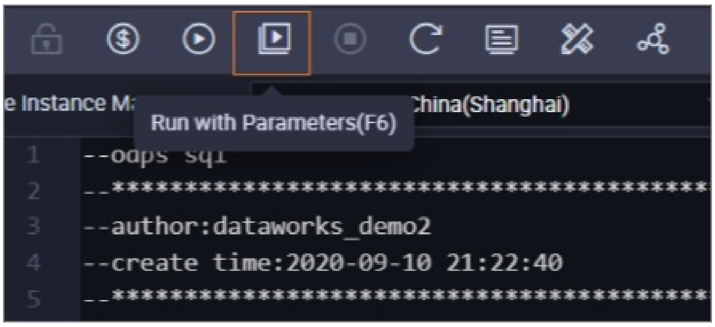
If you want to synchronize data from a data source that resides in a VPC, you must use an exclusive resource group for Data Integration. For more information, see Create and use an exclusive resource group for Data Integration.

How do I change the resource group for scheduling that is used to run a node?

1. If you want to change the resource group for scheduling used to test the running of a node that is

in the development environment, perform the following operation:

Click **Run with Parameters** in the top toolbar on the configuration tab of the node. In the Parameters dialog box, select the resource group that you want to use.

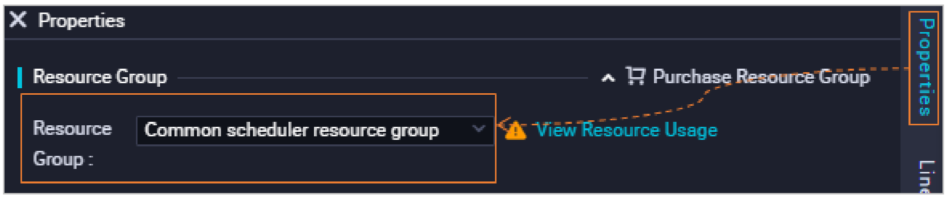


```
--odps sql
--*****************************************
--author:dataworks_demo2
--create time:2020-09-10 21:22:40
--*****************************************
```

2. If you want to change the resource group for scheduling used to run a node that is in the production environment, use one of the following methods:

    - Change the resource group for scheduling on the configuration tab of the node. Then, commit and deploy the node for the change to take effect.

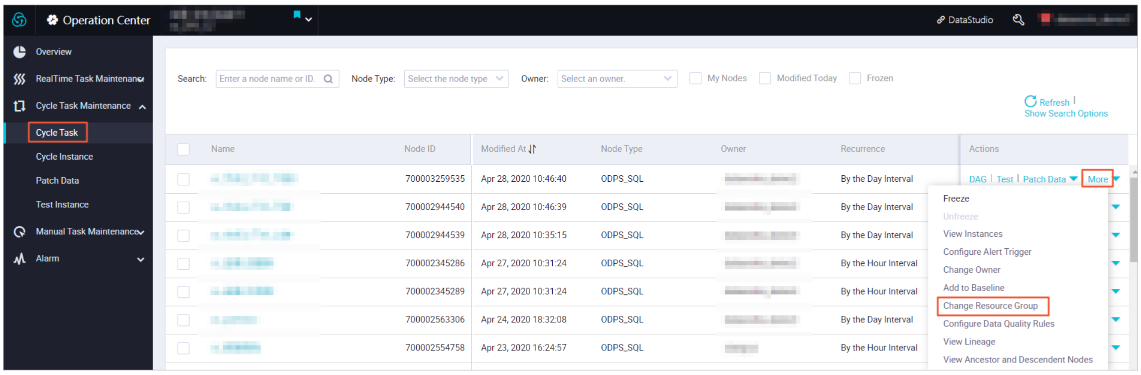
      On the configuration tab of the node, click the **Properties** tab in the right-side navigation pane. In the **Resource Group** section of the Properties tab, select the resource group that you want to use. Then, save the change, and commit and deploy the node for the change to take effect. For more information, see Configure a resource group.



Resource Group : Common scheduler resource group

    - Change the resource group for scheduling in Operation Center of the production environment.

      On the DataStudio page, click the ☰ icon in the upper-left corner and choose **All Products > Task Operation > Operation Center** to navigate to the **Operation Center** page of the production environment. On the Operation Center page, choose Cycle Task Maintenance > **Cycle Task** in the left-side navigation pane. On the Cycle Task page, find the auto triggered node and choose More > Modify Scheduling Resource Group in the Actions column.



Search: Enter a node name or ID. Node Type: Select the node type Owner: Select an owner. □ My Nodes □ Modified Today □ Frozen

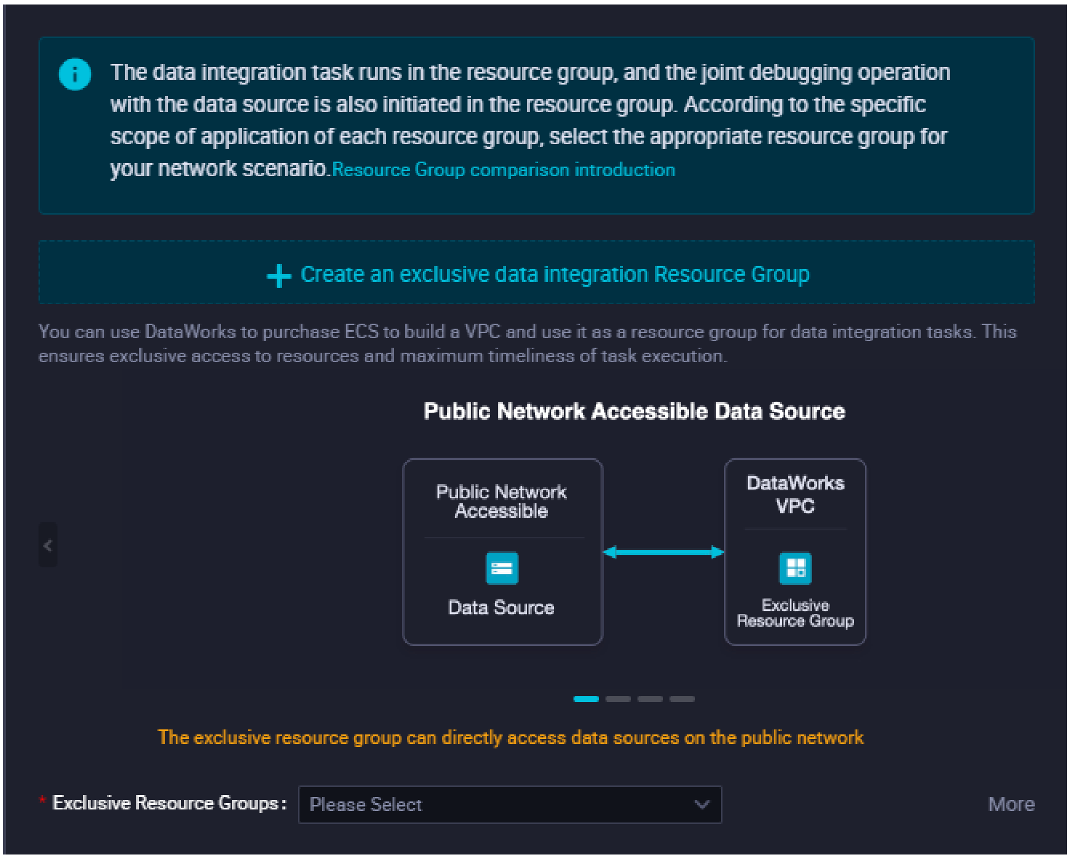
□	Name	Node ID	Modified At	Node Type	Owner	Recurrence	Actions
□	[redacted]	700003259535	Apr 28, 2020 10:46:40	ODPS_SQL	[redacted]	By the Day Interval	DAG | Test | Patch Data More
□	[redacted]	700002944540	Apr 28, 2020 10:46:39	ODPS_SQL	[redacted]	By the Day Interval	
□	[redacted]	700002944539	Apr 28, 2020 10:35:15	ODPS_SQL	[redacted]	By the Day Interval	
□	[redacted]	700002345286	Apr 27, 2020 10:31:24	ODPS_SQL	[redacted]	By the Hour Interval	
□	[redacted]	700002345289	Apr 27, 2020 10:31:24	ODPS_SQL	[redacted]	By the Hour Interval	
□	[redacted]	700002563306	Apr 24, 2020 18:32:08	ODPS_SQL	[redacted]	By the Day Interval	
□	[redacted]	700002554758	Apr 23, 2020 16:24:57	ODPS_SQL	[redacted]	By the Hour Interval	

> **Note** Zero load nodes do not occupy resources. Therefore, you do not need to or cannot change the resource group for scheduling for a zero load node.

How do I change the resource group for Data Integration that is used to run a Data Integration node?

1. If you want to change the resource group for Data Integration used to test the running of a Data Integration node that is in the development environment, perform the following operation:

   On the configuration tab of the node, click **Resource Group configuration** in the right-side navigation pane. On the Resource Group configuration tab, select the exclusive resource group for Data Integration that you want to use from the Exclusive Resource Group drop-down list.

   The data integration task runs in the resource group, and the joint debugging operation with the data source is also initiated in the resource group. According to the specific scope of application of each resource group, select the appropriate resource group for your network scenario.Resource Group comparison introduction

   + Create an exclusive data integration Resource Group

   You can use DataWorks to purchase ECS to build a VPC and use it as a resource group for data integration tasks. This ensures exclusive access to resources and maximum timeliness of task execution.

   Public Network Accessible Data Source

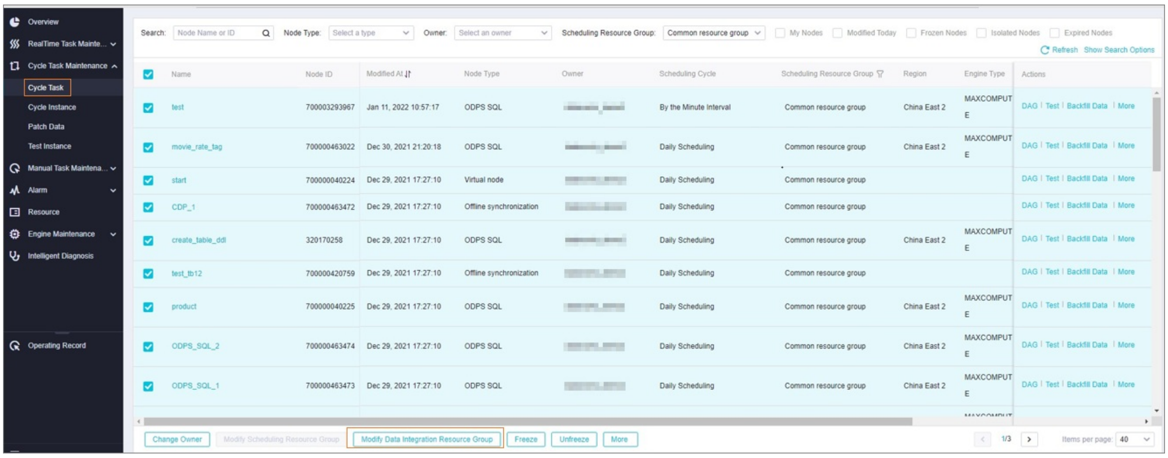
   The exclusive resource group can directly access data sources on the public network

   * Exclusive Resource Groups: Please Select

   Note If this exclusive resource group is also required in the production environment, you must select this resource group for this node and commit and deploy the node to the production environment.

2. If you want to change the resource group for Data Integration used to run a Data Integration node that is in the production environment, perform the following operations:

   On the DataStudio page, click the ☰ icon in the upper-left corner and choose **All Products** > **Task Operation** > **Operation Center** to navigate to the **Operation Center** page of the production environment. On the Operation Center page, choose Cycle Task Maintenance > **Cycle Task** in the left-side navigation pane. On the Cycle Task page, find the auto triggered node for which you want to change the resource group and click Modify Data Integration Resource Group in the lower part of the page.



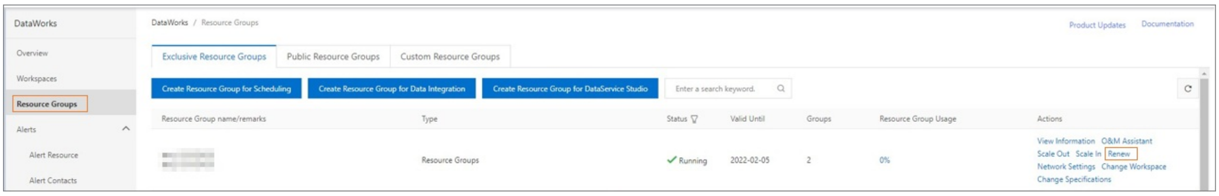
Search: Node Name or ID Node Type: Select a type Owner: Select an owner Scheduling Resource Group: Common resource group □ My Nodes □ Modified Today □ Frozen Nodes □ Isolated Nodes □ Expired Nodes

☑	Name	Node ID	Modified At	Node Type	Owner	Scheduling Cycle	Scheduling Resource Group	Region	Engine Type	Actions
☑	test	700003293967	Jan 11, 2022 10:57:17	ODPS SQL	[redacted]	By the Minute Interval	Common resource group	China East 2	MAXCOMPUTE	DAG | Test | Backfill Data | More
☑	movie_rate_tag	700000463022	Dec 30, 2021 21:20:18	ODPS SQL	[redacted]	Daily Scheduling	Common resource group	China East 2	MAXCOMPUTE	DAG | Test | Backfill Data | More
☑	start	700000040224	Dec 29, 2021 17:27:10	Virtual node	[redacted]	Daily Scheduling	Common resource group			DAG | Test | Backfill Data | More
☑	CDP_1	700000463472	Dec 29, 2021 17:27:10	Offline synchronization	[redacted]	Daily Scheduling	Common resource group			DAG | Test | Backfill Data | More
☑	create_table_ddl	320170258	Dec 29, 2021 17:27:10	ODPS SQL	[redacted]	Daily Scheduling	Common resource group	China East 2	MAXCOMPUTE	DAG | Test | Backfill Data | More
☑	test_tb12	700000420759	Dec 29, 2021 17:27:10	Offline synchronization	[redacted]	Daily Scheduling	Common resource group			DAG | Test | Backfill Data | More
☑	product	700000040225	Dec 29, 2021 17:27:10	ODPS SQL	[redacted]	Daily Scheduling	Common resource group	China East 2	MAXCOMPUTE	DAG | Test | Backfill Data | More
☑	ODPS_SQL_2	700000463474	Dec 29, 2021 17:27:10	ODPS SQL	[redacted]	Daily Scheduling	Common resource group	China East 2	MAXCOMPUTE	DAG | Test | Backfill Data | More
☑	ODPS_SQL_1	700000463473	Dec 29, 2021 17:27:10	ODPS SQL	[redacted]	Daily Scheduling	Common resource group	China East 2	MAXCOMPUTE	DAG | Test | Backfill Data | More

1/3 Items per page: 40

How do I renew a resource group?

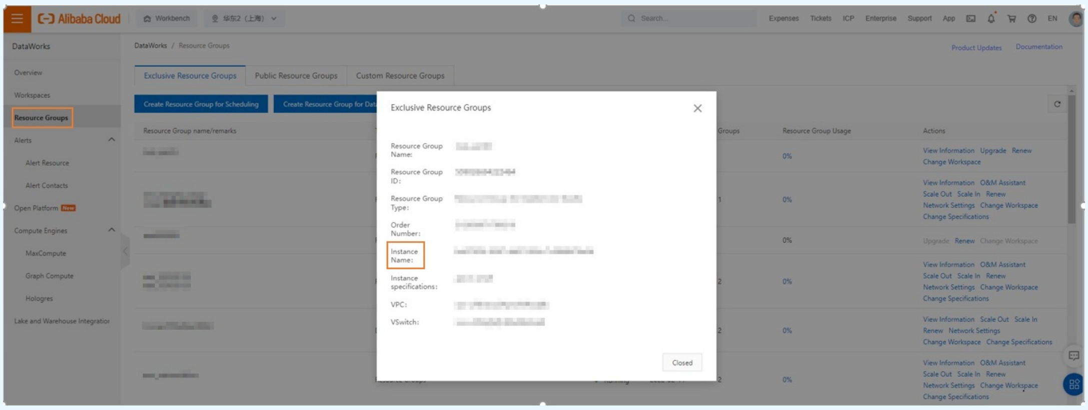
Log on to the DataWorks console. Click Resource Groups in the left-side navigation pane. On the Resource Groups page, find the resource group that you want to renew and click Renew in the Actions column.

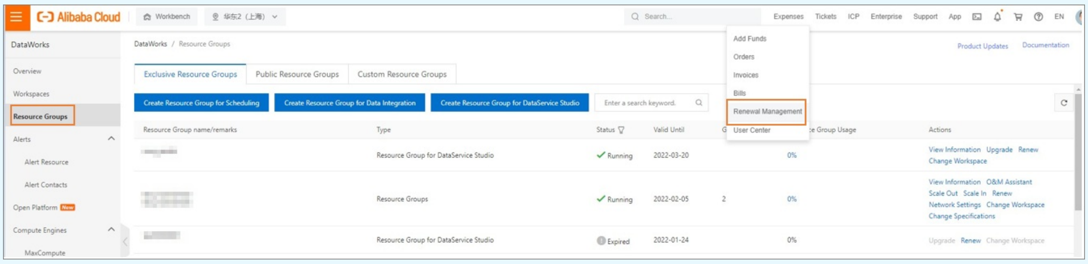

DataWorks
Overview
Workspaces
Resource Groups
Alerts
Alert Resource
Alert Contacts
DataWorks / Resource Groups
Product Updates
Documentation
Exclusive Resource Groups
Public Resource Groups
Custom Resource Groups
Create Resource Group for Scheduling
Create Resource Group for Data Integration
Create Resource Group for DataService Studio
Enter a search keyword.

Resource Group name/remarks	Type	Status	Valid Until	Groups	Resource Group Usage	Actions
[redacted]	Resource Groups	✓ Running	2022-02-05	2	0%	View Information O&M Assistant Scale Out Scale In Renew Network Settings Change Workspace Change Specifications

Note

- If you want to use a RAM user to renew a resource group, you must grant the required permissions to the RAM user. For more information, see User permission management.
- If you want to switch the renewal method of the resource group between manual renewal and auto renewal, you can perform the operation on the Renewal Management page based on the instance name of the resource group.

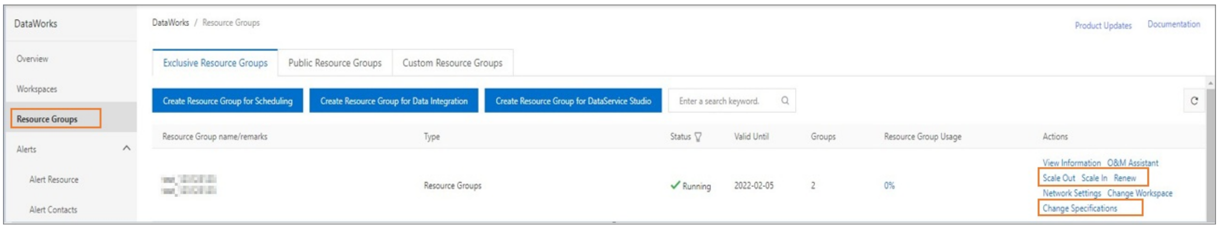




Resource Group name/remarks	Type	Status	Valid Until	G	...e Group Usage	Actions
[illegible]	Resource Group for DataService Studio	✓ Running	2022-03-20		0%	View Information Upgrade Renew Change Workspace
[illegible]	Resource Groups	✓ Running	2022-02-05	2	0%	View Information O&M Assistant Scale Out Scale In Renew Network Settings Change Workspace Change Specifications
[illegible]	Resource Group for DataService Studio	Expired	2022-01-24		0%	Upgrade Renew Change Workspace

How do I change the specifications of an exclusive resource group? How do I scale in or out an exclusive resource group?

Log on to the DataWorks console. Click Resource Groups in the left-side navigation pane. The Exclusive Resource Groups tab appears. On the Exclusive Resource Groups tab, find the desired exclusive resource group and click Scale Out, Scale In, or Change Specifications in the Actions column to perform the related operation.


DataWorks
Overview
Workspaces
Resource Groups
Alerts
Alert Resource
Alert Contacts
DataWorks / Resource Groups
Product Updates
Documentation
Exclusive Resource Groups
Public Resource Groups
Custom Resource Groups
Create Resource Group for Scheduling
Create Resource Group for Data Integration
Create Resource Group for DataService Studio
Enter a search keyword.

Resource Group name/remarks	Type	Status	Valid Until	Groups	Resource Group Usage	Actions
[illegible]	Resource Groups	✓ Running	2022-02-05	2	0%	View Information O&M Assistant Scale Out Scale In Renew Network Settings Change Workspace Change Specifications

Note

- If you want to use a RAM user to renew a resource group, you must grant the required permissions to the RAM user. For more information, see User permission management.
- If you scale in or out a resource group, the number of resources in the resource group is decreased or increased.

What conditions must be met before an exclusive resource group can access a data source that resides in a VPC?

If you want to use an exclusive resource group to access a data source that resides in a VPC in DataWorks, the following conditions must be met:

- The exclusive resource group resides in the same zone as the data source and is associated with a VPC.
- The exclusive resource group is associated with the same VPC and vSwitch as the data source.
- If the data source is configured with a whitelist or security group, the elastic IP address (EIP) of the exclusive resource group or the Classless Inter-Domain Routing (CIDR) block of the VPC or vSwitch with which the exclusive resource group is associated is added to the whitelist or security group. For more information, see Configure a whitelist and Configure a security group for an ECS instance where a self-managed data store resides.

How do I view the network environment of a data source?

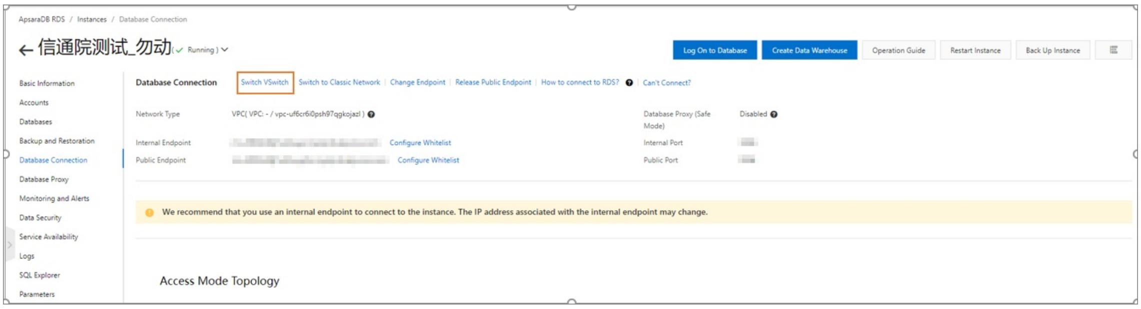
- If you want to view the network environment of an ApsaraDB RDS data source, perform the following steps:

  i. Log on to the ApsaraDB RDS console.

  ii. In the left-side navigation pane, click **Instances**. On the Instances page, find the instance whose network environment you want to view and click the instance name to go to the details page of the instance.

  iii. In the left-side navigation pane of the instance details page, click **Database Connection**.

  iv. On the Database Connection page, click **Switch VSwitch**.



Network Type: VPC(VPC: - / vpc-uf6cr6i0psh97qgkojazl)

Database Proxy (Safe Mode): Disabled

Internal Endpoint: [redacted]

Internal Port: [redacted]

Public Endpoint: [redacted]

Public Port: [redacted]

We recommend that you use an internal endpoint to connect to the instance. The IP address associated with the internal endpoint may change.

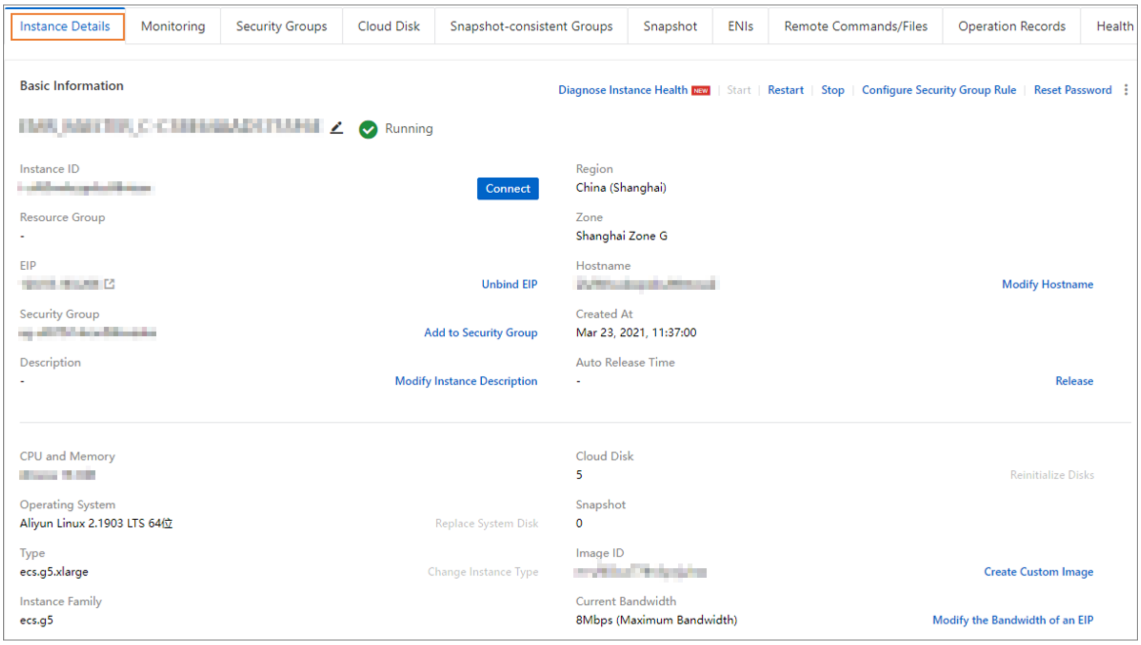
  v. In the **Switch VSwitch** dialog box, view the vSwitch with which the data source is associated.

- If you want to view the network environment of a self-managed data source hosted on an Elastic Compute Service (ECS) instance, perform the following steps:

  i. Log on to the ECS console.

  ii. In the left-side navigation pane, choose Instances & Images > Instances. On the **Instances** page, find the instance whose network environment you want to view and click the instance name to go to the details page of the instance.

  iii. On the **instance details** page, view the **configuration information** of the instance.



Basic Information

[redacted] Running

Instance ID
[redacted]

Resource Group
-

EIP
[redacted]

Security Group
[redacted]

Description
-

Region
China (Shanghai)

Zone
Shanghai Zone G

Hostname
[redacted]

Created At
Mar 23, 2021, 11:37:00

Auto Release Time
-

CPU and Memory
[redacted]

Operating System
Aliyun Linux 2.1903 LTS 64位

Type
ecs.g5.xlarge

Instance Family
ecs.g5

Cloud Disk
5

Snapshot
0

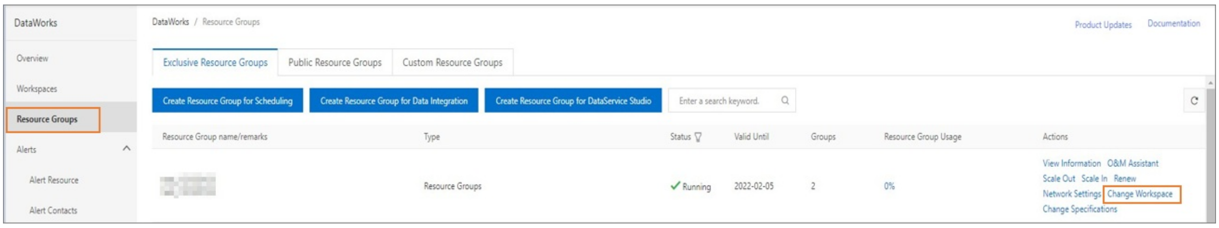
Image ID
[redacted]

Current Bandwidth
8Mbps (Maximum Bandwidth)

What information about an exclusive resource group must be added to the whitelist of a data source that the exclusive resource group needs to access?

How do I associate an exclusive resource group with or disassociate an exclusive resource group from a workspace?

Log on to the DataWorks console. Click Resource Groups in the left-side navigation pane. The Exclusive Resource Groups tab appears. On the Exclusive Resource Groups tab, find the exclusive resource group for which you want to change the workspace and click Change Workspace in the Actions column. In the Modify home workspace dialog box, you can associate the resource group with or disassociate the resource group from a workspace.


DataWorks
Overview
Workspaces
Resource Groups
Alerts
Alert Resource
Alert Contacts
DataWorks / Resource Groups
Product Updates
Documentation
Exclusive Resource Groups
Public Resource Groups
Custom Resource Groups
Create Resource Group for Scheduling
Create Resource Group for Data Integration
Create Resource Group for DataService Studio
Enter a search keyword.

Resource Group name/remarks	Type	Status	Valid Until	Groups	Resource Group Usage	Actions
[redacted]	Resource Groups	✓ Running	2022-02-05	2	0%	View Information O&M Assistant Scale Out Scale In Renew Network Settings Change Workspace Change Specifications

> **Note** You must obtain the administrator permissions of the workspace before you can perform the preceding operations.

6.2. Custom resource groups

This topic describes the answers to the frequently asked questions about custom resource groups.

- How do I install a monitor?
- What can I do if I fail to add an ECS instance to a custom resource group for scheduling?
- What can I do if an ECS instance is suddenly stopped and fails to be restarted?
- What can I do if a node that is running on a custom resource group for scheduling is waiting for

How do I install a monitor?

If an error occurs when you use a custom resource group for scheduling, check whether a monitor is installed for the agent by performing the following steps:

1. Log on to each Elastic Compute Service (ECS) instance and switch to the root user.
2. Run the following command:

   ```
   wget https://alisaproxy.shuju.aliyun.com/install_monitor.sh --no-check-certificate
   ```

3. If no monitor is installed, run the following script to install one:

   ```
   sh install_monitor.sh
   ```

What can I do if I fail to add an ECS instance to a custom resource group for scheduling?

If you fail to add an ECS instance to a custom resource group for scheduling and the status of the instance is always **Stopped**, consider the following reasons:

- The hostname or UUID you entered on the ECS instance registration page is different from the actual one.

  Methods that you can use to check the hostname or UUID:

  - If you set the network type to classic network, check whether the hostname and IP address you entered on the registration page are the same as those returned after you run the `hostname` and `hostname -i` commands on the ECS instance. Take note that you can set the network type to classic network only if the ECS instance is in the China (Shanghai) region.

    > **Note** Check whether you have changed the hostname. If you have changed the hostname, go to the */etc/hosts* directory and check whether the instance is bound to a host. If the instance is bound to a host, enter the name of the bound host on the registration page.

  - If you set the network type to Virtual Private Cloud (VPC), check whether the UUID you entered on the registration page is the same as that returned after you run the `dmidecode | grep UUID` command on the ECS instance.

    > **Note**
    > - If you do not install dmidecode, install it first.
    > - The UUIDs returned by different versions of dmidecode are case-sensitive.
    > - The hostnames are case-sensitive.

  If the issue is caused by this reason, resolve it by performing the following steps:

  i. Remove the original instance.

  ii. Enter the valid IP address and hostname or UUID and register the instance again.

- The initialization commands are invalid.

  To check whether the initialization commands are valid, perform the following steps:

  i. Log on to the ECS instance and run the following command:

  ```
  cat /home/admin/alisatasknode/target/alisatasknode/conf/config.properties | grep driv
  er
  ```

  ii. Log on to the DataWorks console.

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

  iv. On the Resource Groups page, click the **Custom Resource Groups** tab.

  v. Find the required resource group and click **Initialize Server** in the Actions column.

  vi. Check whether the username in the output of the preceding command is the same as that in the initialization dialog box.

  If the issue is caused by this reason, run the valid commands listed in the initialization dialog box to re-initialize the instance.

Note

- After an instance is registered, you can initialize the instance on the **Custom Resource Groups** tab of the **Resource Groups** page.

  The initialization commands for different resource groups are different and cannot be mixed up.
- Copy the commands in the **Initialize Server** dialog box and run them in sequence.
- For instances in a VPC, use the initialization commands for instances on the classic network.

- The difference between the time of the ECS instance and time in UTC+8 is greater than 5 minutes.

  To check the time difference, perform the following steps:

  i. Log on to the ECS instance.

  ii. Run the `date` command and check whether the difference between the returned time and time in UTC+8 is greater than 5 minutes.

  If the issue is caused by this reason, confirm that a time adjustment does not affect your business and then adjust the time of the ECS instance to that in UTC+8.

- The permissions on relevant directories are invalid.

  To check the permissions, perform the following steps:

  i. Log on to the ECS instance and run the `ps -ef | grep zoo | grep -v cdp` command.

  ii. Check whether the returned processes are owned by the admin user.

  If the processes are owned by the admin user, check whether the admin user has permissions on the */home/admin/alisatasknode* directory and its subdirectories.

  If the root permission is required, perform the following steps:

  i. Switch to the root user and run the `chown admin:admin /home/admin -R` command.

  ii. Switch back to the admin user and run the `/home/admin/alisatasknode/target/alisatasknode/bin/serverctl restart` command to restart the agent.

- An error occurs when the `install.sh` script is run.

  To check whether an error occurs when the install.sh script is run, perform the following steps:

  i. Run the `install.sh` script.

  ii. Check whether a log file is generated in the `/home/admin/alisatasknode/logs` directory. If no log file is generated, the agent is not installed.

  If the issue is caused by this reason, resolve it by performing the following steps:

  i. Check whether the operating system of the ECS instance is CentOS 5, CentOS 6, or CentOS 7. If the ECS instance does not run one of the preceding operating systems, change the operating system and re-initialize the instance.

  ii. Run the `/opt/taobao/java/bin/java -V` command to check whether the version of the Java Development Kit (JDK) is 1.8.

  iii. Run the `ls -al /opt/taobao` command to check whether the admin user has permissions on the /opt/taobao directory. If the root permission is required, switch to the root user and run the `chown admin:admin /opt/taobao -R` command. Then, switch back to the admin user and run the initialization commands.

What can I do if an ECS instance is suddenly stopped and fails to be restarted?

If an instance in a custom resource group instance is suddenly stopped after a period of use and fails to be restarted or the issue persists after the restart, consider the following reasons:

- Different users have started the agent. This results in inconsistent permissions on relevant directories.

  To check whether different users have started the agent, perform the following steps:

  i. Log on to the ECS instance and switch to the root user.

  ii. Run the `ps -ef | grep zoo | grep -v cdp` command.

  If two processes are returned, different users have started the agent. In this case, perform the following steps:

  i. Log on to the ECS instance and run the `kill -9` command to end the two processes returned by the preceding `ps` command.

  ii. Switch to the root user and run the `chown admin:admin /home/admin/ -R` command.

  iii. Switch back to the admin user.

  iv. Run the `/home/admin/alisatasknode/target/alisatasknode/bin/serverctl restart` command to restart the agent.

- Relevant processes occupy too many handles.

  To check whether relevant processes occupy too many handles, perform the following steps:

  - Log on to the ECS instance and run the `grep "temporarily unavailable" /home/admin/alisatasknode/logs/alisatasknode.log` command. If a result is returned, relevant processes occupy too many handles.
  - Restart the agent. If you fail to restart the agent and the error message `Caused by: java.io.IOException: error=11, Resource temporarily unavailable` is returned, relevant processes occupy too many handles.

  If the issue is caused by this reason, resolve it by performing the following steps:

  i. Switch to the root user and run the `ps -ef | grep zoo | grep -v cdp` command.

  ii. Run the `kill -9` command to end all the processes returned by the preceding `ps` command.

  iii. Run the `chown admin:admin /home/admin/ -R` command.

  iv. Switch back to the admin user.

  v. Run the `/home/admin/alisatasknode/target/alisatasknode/bin/serverctl restart` command to restart the agent.

- The ECS instance is in a VPC and the UUID of the ECS instance is changed.

  i. Log on to the ECS instance and run the `dmidecode | grep UUID` command. If the letters in the UUID were in uppercase, check whether the returned letters are in lowercase.

  ii. Compare the returned UUID with the one in the **Manage Server** dialog box.

If the issue is caused by this reason, remove the original instance on the **Custom Resource Groups** tab and add the instance with the new UUID.

> **Note** If the instance cannot be removed and the error message `remove node failed, exception: [3006:ERROR_GATEWAY_EXIST_TASKS]:gateway tasks not empty` is returned, record the region where the instance resides, copy the error message, and then submit a ticket to seek technical help from Alibaba Cloud.

What can I do if a node that is running on a custom resource group for scheduling is waiting for resources for an extended period of time?

If a node that is running on a custom resource group for scheduling is waiting for resources for an extended period of time, consider the following reasons:

- The ECS instance that processes the node is stopped.

  To check whether the ECS instance is stopped, perform the following steps:

  i. Log on to the DataWorks console.

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

  iii. On the Resource Groups page, click the **Custom Resource Groups** tab.

  iv. Find the instance to be checked and click **Manage Server** in the Actions column to check whether the status of the ECS instance is Stopped.

  If the ECS instance is stopped, log on to the instance and start the agent.

- The ECS instance is temporarily unavailable.
  - To check whether the ECS instance is temporarily unavailable, perform the following steps:

    a. Log on to the ECS instance.

    b. View logs in the */home/admin/alisatasknode/logs/alisatasknode_status.log* file.

    The logs display the status of the instance in real time. If the instance status is **BUSY** or **HANGUP**, a node that is running on the instance occupies many resources.

  - To resolve the issue, perform the following steps:

    a. Run the `ps -ef | grep taskexec` command to view the relevant processes of nodes.

    b. Check logs to find the node that occupies many resources.

    If the node is abnormal, terminate it in the DataWorks console. Wait 2 minutes. Then the instance automatically works again.

- The agent is abnormal.

  To check whether the agent is abnormal, perform the following steps:

  - Run the `df -h` command to check whether the disk usage is 100%.
  - Check whether the CPU utilization and memory usage are too high.

  If the issue is caused by this reason, resolve the issue of the instance and then restart the agent.

How do I temporarily disable or initialize the agent?

To temporarily disable the agent, select one of the following methods:

- If you add the agent on the **Custom Resource Groups** tab of the **Resource Groups** page, find the instance for which you want to disable the agent and click **Manage Server** in the Actions column. In the **Manage Server** dialog box, click **Freeze**.
- If you added the agent on the **Custom Resource Groups** page in **Data Integration**, the agent cannot be stopped. You can submit a ticket to seek technical help from Alibaba Cloud.

To initialize the agent, perform the following steps:

1. Switch to the root user and run the `ps -ef | grep zoo | grep alisa` command.
2. Run the `kill -9` command to end the processes returned by the preceding `ps` command.
3. Delete the */home/admin/alisatasknode* directory.
4. Run the `install.sh` script in an empty directory.

   > ⓘ **Note** Download the install.sh script in the region where the instance resides.

How do I enable the agent to automatically work after I restart an ECS instance?

After you restart an ECS instance, you can perform the following steps to enable the agent to automatically work:

1. Log on to the ECS instance and switch to the root user.
2. Run the `wget https://alisaproxy.shuju.aliyun.com/install_monitor.sh --no-check-certificate` command.
3. Run the `sh install_monitor.sh` command.

What can I do if I fail to start a custom resource group?

If you enter the hostname instead of the UUID to register an ECS instance whose network type is VPC and fail to start a custom resource group, run the `tail -f /home/admin/alisatasknode/logs/alisatasknode.log` command to view the operational logs to determine the cause.

If you enter the UUID to register an ECS instance whose network type is VPC and fail to start a custom resource group, the initialization commands may be invalid and you must change the commands to the valid ones.

To resolve the issue, perform the following steps:

1. Log on to the DataWorks console.
2. In the left-side navigation pane, click **Resource Groups**.
3. On the Resource Groups page, click the **Custom Resource Groups** tab.
4. Find the instance for which you want to change initialization commands to the valid ones and click **Initialize Server** in the Actions column.
5. Perform the steps listed in the **Initialize Server** dialog box.

   > ⓘ **Note** In Step 3, change `enable_uuid=false` to `enable_uuid=true` in the commands.

What are the advantages of custom resource groups?

- Ensure enough resources: If all tenants share shared resource groups, high resource usage may lead to an extended period of time for waiting for resources. If you have high requirements for resource usage, you can select a custom resource group to run your node when you create the node.
- Connect to databases in various network environments: Shared resource groups cannot connect to databases in a VPC. Therefore, you can use a custom resource group to connect to databases in the VPC.
- Be used for scheduling: If resources for scheduling are insufficient, you can use a custom resource group.
- Improve concurrency: Shared resource groups contain a limited number of slots. You can add slots by creating custom resource groups to concurrently run more nodes.

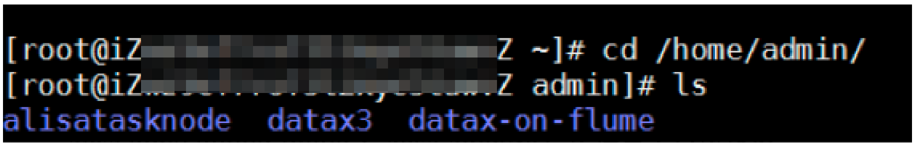
What are the limits on custom resource groups?

- You can add an ECS instance to only one custom resource group, but you can add multiple ECS instances to a custom resource group.
- If you set the network type to classic network when you register an ECS instance, you must enter the hostname of the instance. If you set the network type to VPC, you must enter the UUID of the instance.
- You can select only one network type for each custom resource group.
- You cannot run manually triggered node instances on custom resource groups.
- ECS instances must be able to access the Internet. You can configure a public IP address, an Elastic IP Address (EIP), and a SNAT IP address of the NAT gateway for an ECS instance.

What information do I need to view after I install a custom resource group?

After you install a custom resource group based on the instructions in the DataWorks console, log on to an ECS instance and view the following information about the agent:

- Default directory: */home/admin/*. This directory usually contains the following folders:
  - *alisatasknode*: stores agent-related configurations and commands.
  - *datax* and *datax-on-flume*: store the synchronization wrapper library and configurations.



```
[root@iZ[redacted]Z ~]# cd /home/admin/
[root@iZ[redacted]Z admin]# ls
alisatasknode  datax3  datax-on-flume
```

- Agent-related commands: You can run commands such as stop, start, and restart on the agent process.

```
/home/admin/alisatasknode/target/alisatasknode/bin/serverctl start/stop/restart
```

- Operational logs: The operational logs of the agent are stored in the following directories:
  - */home/admin/alisatasknode/taskinfo/*: stores the operational logs of shell scripts. The logs are the same as the operational logs of DataWorks nodes.
  - */home/admin/alisatasknode/logs*: The *alisatasknode.log* file stores the running information of the agent, such as the node operations and the heartbeat status of the agent.

- */home/admin/datax3/log*: stores detailed operational logs of data integration nodes. If a node fails, you can view the logs for troubleshooting.

How do I monitor the status of the agent process?

You can perform the following steps to monitor the agent process. If the agent process exits, you can recover it at the earliest opportunity.

1. Log on to the ECS instance as the root user.
2. Run the `wget https://alisaproxy.shuju.aliyun.com/install_monitor.sh --no-check-certificate` command.
3. Run the `sh install_monitor.sh` command. By default, monitoring logs are stored in the */home/admin/alisatasknode/monitor/monitor.log* file.

What types of resources does the DataWorks scheduling system provide?

Custom resource groups are used in the DataWorks scheduling system. The DataWorks scheduling system provides level-1 scheduling resources and level-2 running resources.

- Level-1 scheduling resources: Go to the **Operation Center** page and choose **Cycle Task Maintenance > Cycle Instance** in the left-side navigation pane. On the Cycle Instance page, right-click the specified instance in the directed acyclic graph (DAG) on the right and select **More**. On the **General** tab, you can view the level-1 scheduling resources.
- Level-2 running resources: Go to the **Data Integration** page and click **Custom Resource Group** in the left-side navigation pane. On the Custom Resource Groups page, you can view the level-2 running resources.

How do I use custom resource groups?

- Configure level-1 scheduling resources

  Log on to the DataWorks console. In the left-side navigation pane, click **Resource Groups**. On the Resource Groups page, click the **Custom Resource Groups** tab and create a custom resource group on this tab.

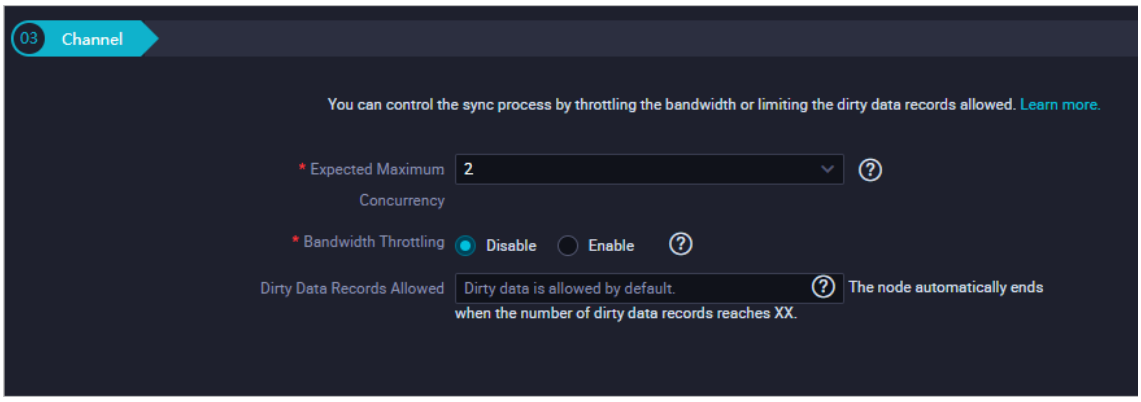
  > **Note** The resource groups that you create on this tab are applicable to shell nodes, and the resources that you configure on this tab are level-1 scheduling resources.

- Configure level-2 running resources

  i. Log on to the DataWorks console. In the left-side navigation pane, click Workspaces. On the Workspaces page, find the specified workspace and click **Data Integration** in the Actions column. On the page that appears, click **Custom Resource Group** in the left-side navigation pane. On the Custom Resource Groups page, click **Add Resource Group** in the upper-right corner to create a custom resource group. For more information, see Create and use a custom resource group for Data Integration.

  > **Note** The resource groups that you create on this page are applicable only to synchronization nodes, and the resources that you configure on this page are level-2 running resources.

  ii. After you create the custom resource group, go to the configuration tab of the specified node in **DataStudio**. In the right-side navigation pane, click the **Resource Group configuration** tab. In the Resource Group configuration panel, select the custom resource group that you created.



03 Channel

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

* Expected Maximum Concurrency: 2

* Bandwidth Throttling: ● Disable ○ Enable

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

What can I do if an error message is returned when I add an ECS instance to a custom resource group?

If the error message `gateway already exists` is returned when you add an ECS instance to a custom resource group, perform the following steps to resolve this issue:

1. Check whether an ECS instance with the same hostname or UUID exists on the Custom Resource Groups tab of the **Resource Groups** page and the **Custom Resource Groups** page in **Data Integration**. This is because the error message indicates that the ECS instance has been registered in the gateway and an ECS instance can be added to only one custom resource group.
2. If you do not find such an ECS instance in your workspace, provide the request ID for Alibaba Cloud engineers for consultation.

What can I do if the custom resource group I create is unavailable?

- View logs in the *alisatasknode.log* file to check whether the heartbeat status code 302 is returned. If the heartbeat status code 302 is returned, check the following items:
  - Check whether the UUID on the Custom Resource Groups page is the same as that returned after you run the `dmidecode | grep UUID` command on the ECS instance.

    > **Note** The UUID is case-sensitive.

  - If the UUIDs are different, enter the valid UUID and reinstall the agent.

    > **Note** For dmidecode 3.0.5 or earlier, letters in a UUID are in uppercase. If you upgrade dmidecode to 3.1.2 or later, the letters in the UUID change to lowercase ones. This leads to an abnormal heartbeat. In this case, you must reinstall the agent.

  - Check whether the username and password in the *config.properties* file are the same as those that appear when you install the agent on the custom resource group page. If not, reinstall the agent by running the commands listed in the agent installation dialog box.
  - If the UUID, username, and password are valid, check the node.uuid.enable parameter in the *config.properties* file. For an ECS instance in a VPC, the value of this parameter must be true. If node.uuid.enable is set to false for an ECS instance in a VPC, change the value to true and restart the agent process.
- View logs in the *alisatasknode.log* file to check whether the logs contain information related to

connection timeout. If such information exists, perform the following steps:

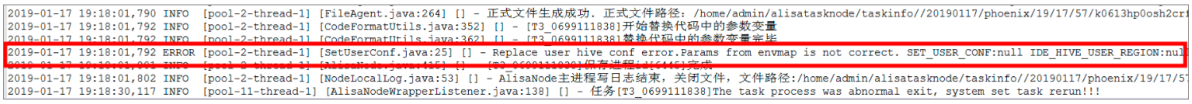
i. Check whether the ECS instance can access the Internet, for example, whether the ECS instance is configured with a public IP address, an EIP, or a SNAT IP address of the NAT gateway. You can run the `ping www.taobao.com` command and check whether www.taobao.com can be reached by PING messages to determine whether the ECS instance can access the Internet.

ii. If the ECS instance can access the Internet, check whether access control is enabled in the outbound rule of the security group for traffic over the Internet or internal network. If access control is enabled, add the IP address and port number of the gateway to the outbound rule.

What can I do if an ECS instance is normal but a shell node fails?

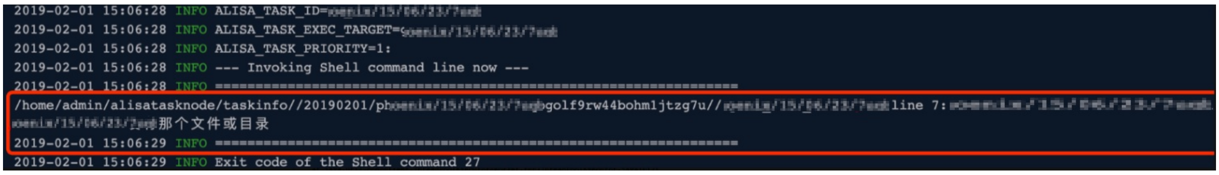
- Use the keyword T3_0699121848 to search for detailed error information in the *alisatasknode.log* file.
- Log on to the ECS instance, switch to the admin user, and then run the `python -V` command to check whether the Python version is 2.7 or 2.6.

  The agent supports Python V2.7 or V2.6. If the Python version is not V2.7 or V2.6, the error message replace user hive conf error is returned.



```
2019-01-17 19:18:01,790 INFO  [pool-2-thread-1] [FileAgent.java:264] [] - 正式文件生成成功. 正式文件路径: /home/admin/alisatasknode/taskinfo//20190117/phoenix/19/17/57/k0613hp0osh2crf
2019-01-17 19:18:01,792 INFO  [pool-2-thread-1] [CodeFormatUtils.java:352] [] - [T3_0699111838]开始替换代码中的参数变量
2019-01-17 19:18:01,792 INFO  [pool-2-thread-1] [CodeFormatUtils.java:362] [] - [T3_0699111838]替换代码中的参数变量完毕
2019-01-17 19:18:01,792 ERROR [pool-2-thread-1] [SetUserConf.java:25] [] - Replace user hive conf error.Params from envmap is not correct. SET_USER_CONF:null IDE_HIVE_USER_REGION:nul
2019-01-17 19:18:01,801 INFO  [pool-2-thread-1] [AlisaNode.java:415] [] - [T3_0699111838]保存进程id[6445]完成
2019-01-17 19:18:01,802 INFO  [pool-2-thread-1] [NodeLocalLog.java:53] [] - AlisaNode主进程写日志结束, 关闭文件, 文件路径:/home/admin/alisatasknode/taskinfo//20190117/phoenix/19/17/57
2019-01-17 19:18:30,117 INFO  [pool-11-thread-1] [AlisaNodeWrapperListener.java:138] [] - 任务[T3_0699111838]The task process was abnormal exit, system set task rerun!!!
```

What can I do if I fail to find a specific operational log file of DataWorks?

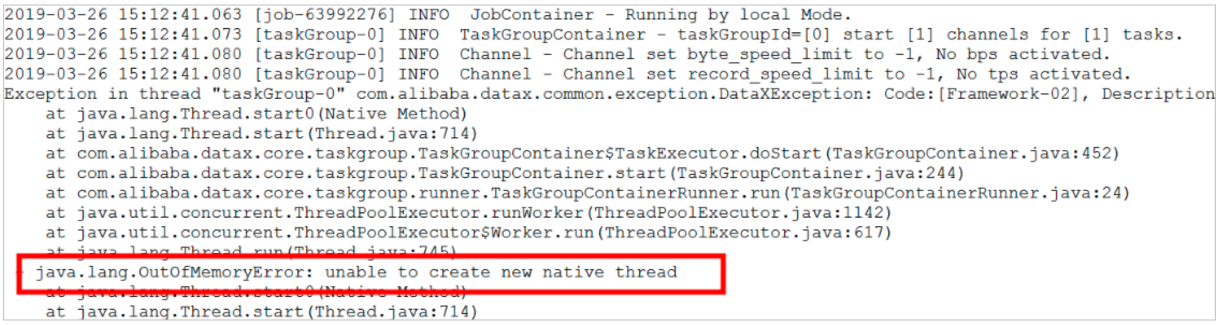


```
2019-02-01 15:06:28 INFO ALISA_TASK_ID=[redacted]
2019-02-01 15:06:28 INFO ALISA_TASK_EXEC_TARGET=[redacted]
2019-02-01 15:06:28 INFO ALISA_TASK_PRIORITY=1:
2019-02-01 15:06:28 INFO --- Invoking Shell command line now ---
2019-02-01 15:06:28 INFO ===============================================================
/home/admin/alisatasknode/taskinfo//20190201/ph[redacted]gbgolf9rw44bohmljtzg7u//[redacted]line 7:[redacted]
[redacted]那个文件或目录
2019-02-01 15:06:29 INFO ===============================================================
2019-02-01 15:06:29 INFO Exit code of the Shell command 27
```

Log on to the ECS instance, switch to the admin user, and then run the `sh -x Script name` command. Check whether the command can be run. If the command fails to be run, resolve the issue based on the returned error message.

What can I do if an OOM error occurs and I fail to allocate memory to relevant threads when I run a node on a custom resource group?

Problem description: An out of memory (OOM) error occurs when a node is run on a custom resource group. The operational logs shown in the following figure indicate that memory cannot be allocated to the relevant threads.



```
2019-03-26 15:12:41.063 [job-63992276] INFO  JobContainer - Running by local Mode.
2019-03-26 15:12:41.073 [taskGroup-0] INFO  TaskGroupContainer - taskGroupId=[0] start [1] channels for [1] tasks.
2019-03-26 15:12:41.080 [taskGroup-0] INFO  Channel - Channel set byte_speed_limit to -1, No bps activated.
2019-03-26 15:12:41.080 [taskGroup-0] INFO  Channel - Channel set record_speed_limit to -1, No tps activated.
Exception in thread "taskGroup-0" com.alibaba.datax.common.exception.DataXException: Code:[Framework-02], Description
    at java.lang.Thread.start0(Native Method)
    at java.lang.Thread.start(Thread.java:714)
    at com.alibaba.datax.core.taskgroup.TaskGroupContainer$TaskExecutor.doStart(TaskGroupContainer.java:452)
    at com.alibaba.datax.core.taskgroup.TaskGroupContainer.start(TaskGroupContainer.java:244)
    at com.alibaba.datax.core.taskgroup.runner.TaskGroupContainerRunner.run(TaskGroupContainerRunner.java:24)
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
    at java.lang.Thread.run(Thread.java:745)
  java.lang.OutOfMemoryError: unable to create new native thread
    at java.lang.Thread.start0(Native Method)
    at java.lang.Thread.start(Thread.java:714)
```

Cause: The memory size that you set when you create a custom resource group determines the slot capability of the resource group. The system and agent processes in a resource group occupy a part of memory. Therefore, the memory of an ECS instance cannot be all used for slots, and an OOM error may occur when too many nodes are concurrently run.

Solution: We recommend that you decrease the memory size for the custom resource group and reserve 2 GB of memory for the system and agent processes. If other processes exist, we recommend that you reserve more memory.

How do I fix the log4j exception of DataX when I use a custom resource group for Data Integration?

To fix the log4j exception, perform the following steps:

1. **Download the Apache log4j-core file**

   i. Download the Apache log4j-core file.

   ii. Decompress the downloaded *Apache log4j-core file* named `tar zxvf apache-log4j-2.17.1-bin.tar.gz` and find the `log4j-core-2.17.1.jar` JAR package in the directory.

2. **Fix the log4j exception**

   i. Upload the JAR package to a temporary directory.

   Upload the `log4j-core-2.17.1.jar` JAR package to the temporary directory on the ECS instance, such as `/tmp/` .

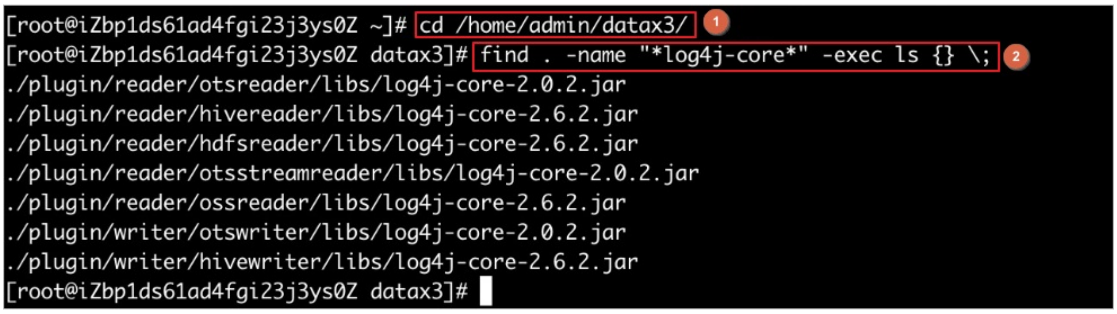
   ii. Go to the DataX installation directory.

   On the specified ECS instance in the custom resource group for Data Integration, run the `cd /home/admin/datax3/` command to go to the `/home/admin/datax3/` installation directory of `DataX 3` .

   iii. Confirm the JAR packages to be replaced.

   Run the `find . -name "*log4j-core*" -exec ls {} \;` command to query the `log4j-core` JAR packages.

   The following figure shows the JAR packages to be replaced.



```
[root@iZbp1ds61ad4fgi23j3ys0Z ~]# cd /home/admin/datax3/
[root@iZbp1ds61ad4fgi23j3ys0Z datax3]# find . -name "*log4j-core*" -exec ls {} \;
./plugin/reader/otsreader/libs/log4j-core-2.0.2.jar
./plugin/reader/hivereader/libs/log4j-core-2.6.2.jar
./plugin/reader/hdfsreader/libs/log4j-core-2.6.2.jar
./plugin/reader/otsstreamreader/libs/log4j-core-2.0.2.jar
./plugin/reader/ossreader/libs/log4j-core-2.6.2.jar
./plugin/writer/otswriter/libs/log4j-core-2.0.2.jar
./plugin/writer/hivewriter/libs/log4j-core-2.6.2.jar
[root@iZbp1ds61ad4fgi23j3ys0Z datax3]#
```

   iv. Back up the `log4j-core` JAR packages to be replaced.

   Run the `find . -name "*log4j-core*" -exec mv {} /tmp \;` command to back up the `log4j-core` JAR packages to be replaced to the `/tmp` directory.

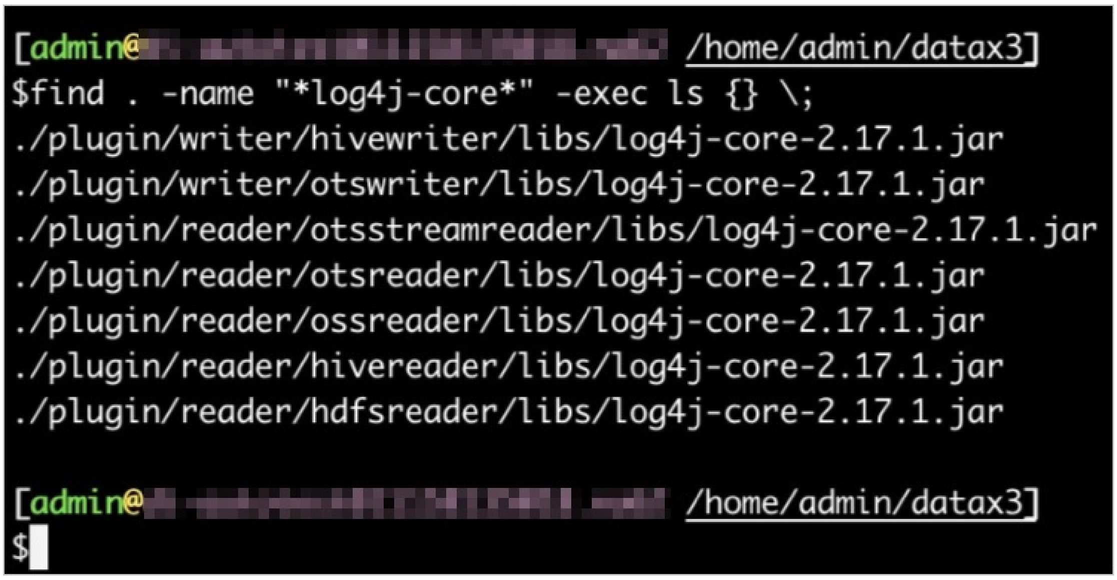
v. Replace the `log4j-core` JAR packages.

Delete the JAR packages to be replaced from their directory. Copy the `/tmp/ log4j-core-2.17.1.jar` JAR package and paste it into this directory. The following example provides sample commands:

```
cp /tmp/log4j-core-2.17.1.jar  ./plugin/writer/hivewriter/libs/
cp /tmp/log4j-core-2.17.1.jar  ./plugin/writer/otswriter/libs/
cp /tmp/log4j-core-2.17.1.jar  ./plugin/reader/otsstreamreader/libs/
cp /tmp/log4j-core-2.17.1.jar  ./plugin/reader/otsreader/libs/
cp /tmp/log4j-core-2.17.1.jar  ./plugin/reader/ossreader/libs/
cp /tmp/log4j-core-2.17.1.jar  ./plugin/reader/hivereader/libs/
cp /tmp/log4j-core-2.17.1.jar  ./plugin/reader/hdfsreader/libs/
```

vi. View the replacement result.

Run the `find . -name "*log4j-core*" -exec ls {} \;` command to view the replacement result.



```
[admin@[redacted] /home/admin/datax3]
$find . -name "*log4j-core*" -exec ls {} \;
./plugin/writer/hivewriter/libs/log4j-core-2.17.1.jar
./plugin/writer/otswriter/libs/log4j-core-2.17.1.jar
./plugin/reader/otsstreamreader/libs/log4j-core-2.17.1.jar
./plugin/reader/otsreader/libs/log4j-core-2.17.1.jar
./plugin/reader/ossreader/libs/log4j-core-2.17.1.jar
./plugin/reader/hivereader/libs/log4j-core-2.17.1.jar
./plugin/reader/hdfsreader/libs/log4j-core-2.17.1.jar

[admin@[redacted] /home/admin/datax3]
$
```

Note After the replacement is complete, pay attention to the execution process of the scheduling task. If an exception occurs, submit a ticket.

If you want to roll back to the original version of JAR packages, move the JAR packages that you backed up to the directory from which they are deleted, and delete the `log4j-core-2.17.1.jar` JAR package. The following example provides sample commands:

```
rm -rf ./plugin/reader/otsreader/libs/log4j-core-2.17.1.jar
rm -rf ./plugin/reader/hivereader/libs/log4j-core-2.17.1.jar
rm -rf ./plugin/reader/hdfsreader/libs/log4j-core-2.17.1.jar
rm -rf ./plugin/reader/otsstreamreader/libs/log4j-core-2.17.1.jar
rm -rf ./plugin/reader/ossreader/libs/log4j-core-2.17.1.jar
rm -rf ./plugin/writer/otswriter/libs/log4j-core-2.17.1.jar
rm -rf ./plugin/writer/hivewriter/libs/log4j-core-2.17.1.jar
cp /tmp/log4j-core-2.0.2.jar ./plugin/reader/otsreader/libs/
cp /tmp/log4j-core-2.6.2.jar ./plugin/reader/hivereader/libs/
cp /tmp/log4j-core-2.6.2.jar ./plugin/reader/hdfsreader/libs/
cp /tmp/log4j-core-2.0.2.jar ./plugin/reader/otsstreamreader/libs/
cp /tmp/log4j-core-2.6.2.jar ./plugin/reader/ossreader/libs/
cp /tmp/log4j-core-2.0.2.jar ./plugin/writer/otswriter/libs/
cp /tmp/log4j-core-2.6.2.jar ./plugin/writer/hivewriter/libs/
```

7.Data Source

7.1. Management of permissions on data sources

This topic provides answers to some commonly asked questions about management of permissions on data sources.

- Which roles can manage permissions on data sources?
- Can I share a data source that has been shared?
- Can the creator of a data source control the data source?
- What is a private data source?
- Which users can revoke the share permissions on a data source?
- How do I check the share relationship of a data source?
- Is the connectivity status of the resource group to which the data source you want to share belongs shared after the data source is shared?
- Is the information of the task that has a dependency relationship with a data source shared after the data source is shared?

Which roles can manage permissions on data sources?

The roles such as tenant owner, tenant administrator, workspace administrator, and project owner are authorized to manage permissions on data sources.

Can I share a data source that has been shared?

DataWorks does not allow you to share a data source that has been shared.

Can the creator of a data source control the data source?

No. A data source is not controlled by the creator.

The creator may be removed or degraded to a low-privilege role such as developer due to incomplete information of the creator. To ensure security of a data source, only the authorized roles are allowed to control this data source.

What is a private data source?

A private data source is shared with a single user. Only this user has permissions to view and use this data source.

Other users cannot view the data source even if they have permissions to control the data source.

Which users can revoke the share permissions on a data source?

- If users have permissions to manage a data source, they can modify the share permissions on the **permission management** page of the data source.
- If users have the **edit** share permissions on a data source, they can **revoke** the share permissions.
- If users have the **read-only** share permissions on a data source, they cannot **revoke** the share permissions.

How do I check the share relationship of a data source?

- After a data source is shared with a user, the user can control the share relationship by using the original data source, for example, cancel the sharing as needed. In this case, DataWorks does not check the task dependency of the shared data source.
- After a data source is shared with a user that has the **edit** share permission and the user **cancels the sharing**, DataWorks checks the task dependency of the shared data source. If the task dependency exists, the user must delete the task that has a dependency relationship with this data source before the user deletes the data source.

Is the connectivity status of the resource group to which the data source you want to share belongs shared after the data source is shared?

No. The shared data source is considered a new data source that may have a different resource group from the original data source. Therefore, you must perform a connectivity test for the new data source.

Is the information of the task that has a dependency relationship with a data source shared after the data source is shared?

No. The shared data source is considered a new data source. It has no relationship with the task that has a dependency relationship with the original data source.

7.2. Troubleshooting for connections

This topic describes how to troubleshoot issues related to connectivity, parameters, and permissions when you create connections in DataWorks.

Connectivity

> Note
>
> - If you use an RDS data store, we recommend that you configure a whitelist for the RDS data store. For more information, see Configure a whitelist.
> - If you use a user-created data store on an Elastic Compute Service (ECS) instance, we recommend that you configure a security group for the ECS instance. For more information, see Configure a security group for an ECS instance where a self-managed data store resides.

Connectivity test failures are major issues.

- Problem description: When I create a connection to a MySQL data store whose network type is the classic network, the connectivity test fails. The following error message is returned: `Connection failed, data store connectivity test failed, database connection failed, database connection string: ... error message: Communications link failure. The last packet sent successfully to the server was 0 milliseconds ago. The dirver has not received any packets from the server` .

  Solution: The error is usually caused by network connectivity issues. We recommend that you check whether your network is accessible, whether the firewall has limits on the specified IP address or port, and whether the security group has been configured to allow traffic for the specified IP address or port.

- Problem description: When I create a connection to an ApsaraDB for MongoDB data store, the connectivity test fails. The following error message is returned:

```
error message: Timed out after 5000 ms while waiting for a server that matches ReadPrefer
enceServerSelector{readPreference=primary}. Client view of cluster state is {type=UNKNOWN
, servers=[..] error with code: PROJECT_DATASOURCE_CONN_ERROR
```

Solution: First, check the region where your DataWorks workspace resides. Then, check the network type of the ApsaraDB for MongoDB data store. If the network type is virtual private cloud (VPC), the ApsaraDB for MongoDB data store does not support connectivity tests in a VPC. In this case, use Method 1 to avoid this issue.

To synchronize data from or to an ApsaraDB for MongoDB data store in a VPC, you can use one of the following methods:

- Method 1: Synchronize data by using the Internet.
  a. Create a connection to the ApsaraDB for MongoDB data store and set the connection type to **Connection string mode**.
  b. Enable the ApsaraDB for MongoDB data store in a VPC to access the Internet.
  c. Add relevant IP addresses to the whitelist of the ApsaraDB for MongoDB data store. For more information, see Configure a whitelist.
  d. Test the connectivity of the MongoDB connection.
- Method 2: Configure a custom resource group and synchronize data by using the internal network.
  a. Create a custom resource group on an ECS instance that is in the same region and VPC as the ApsaraDB for MongoDB data store. For more information, see Add a MongoDB data source.
  b. Add the IP address of the ECS instance to the whitelist or security group of the ApsaraDB for MongoDB data store.
  c. Save the connection settings without testing the connectivity when you create a MongoDB connection. The ApsaraDB for MongoDB data store does not support connectivity tests in a VPC.
  d. Use the custom resource group to run a sync node for synchronizing data from or to the ApsaraDB for MongoDB data store and test the sync node.

> **Note** Be sure to add relevant IP addresses to the whitelist of the corresponding data store.

- Problem description: When I create a connection to a user-created MongoDB data store, the connectivity test fails.

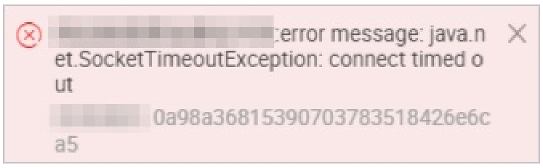
  Solution:

  i. Create a connection to the user-created MongoDB data store and set the connection type to **Connection string mode**.
  ii. If the user-created MongoDB data store is deployed on an ECS instance in a VPC, enable the MongoDB data store to access the Internet.
  iii. Check the connectivity between the network and the specified port. Check the firewall and security group settings of the ECS instance.
  iv. Check the access control, permissions, and remote logon of the user-created MongoDB data store.
  v. Confirm that the endpoint in the `host:port` format, database name, and username are

correct for the user-created MongoDB data store.

> **Note**
>
> When you create a MongoDB connection, you must use the username that is created for the database where the table to be synchronized resides. Do not use root as the username.
>
> For example, to import the name table from the test database, enter test as the database name.
>
> Enter a username that is created for the specified database, instead of root. For example, use a username that is created for the test database.

- Problem description: When I create a connection to a Redis data store in a VPC, the connectivity test fails. An error message is returned, as shown in the following figure.

  [redacted]:error message: java.net.SocketTimeoutException: connect timed out
  [redacted] 0a98a36815390703783518426e6ca5

  Solution: If no public IP address is available, ensure that the Redis data store is in the same region as your DataWorks workspace. You can add scheduling resources to support the connectivity of the Redis connection.

- Problem description: I create a MongoDB connection and a whitelist is configured for the MongoDB data store. However, the connectivity test still fails and the following error message is returned:

  ```
  error message: Timed out after 5000 ms while waiting for a server that matches ReadPrefer
  enceServerSelector{readPreference=primary}
  ```

  Solution: The MongoDB data store in a VPC cannot be connected to default resource groups of DataWorks on the internal network. Therefore, you cannot use default resource groups to run sync nodes for the MongoDB connection. You can enable the MongoDB data store to access the Internet or create custom resource groups to support the connectivity of the MongoDB connection.

- Problem description: When I create a connection to a MySQL data store in a Docker container, the connectivity test fails.

  Solution: The MySQL data store in a Docker container cannot be connected by using a Java Database Connectivity (JDBC) URL that is composed of the public IP address of the server. As a result, the connectivity test fails. You must map the port of the MySQL data store to the host of the Docker container and use the mapped link of the port to connect to the MySQL data store.

- Problem description: When I create a Redis connection, the connectivity test fails. The following error message is returned:

  ```
  error message: java.net.SocketTimeoutException: connect timed out
  ```

  Solution: DataWorks does not support creating a connection to a Redis data store by using the internal network. We recommend that you enable the Redis data store to access the Internet. When you create a Redis connection, set the connection type to **Connection string mode** to connect to the Redis data store by using the Internet.

- Problem description: When I create a connection to an ApsaraDB for RDS data store, the connectivity test fails.

Solution:

i. When the connectivity test of an RDS connection fails, you must add the IP addresses of servers involved in data synchronization to the whitelist of your ApsaraDB for RDS data store. For more information, see Configure a whitelist.

> **Note** If you use custom resource groups to run sync nodes that synchronize data from or to the ApsaraDB for RDS data store, you must add the IP addresses of ECS instances where the custom resource groups are configured to the whitelist of the ApsaraDB for RDS data store.

ii. When you create the RDS connection, ensure that the ID of the ApsaraDB for RDS instance, ID of the Alibaba Cloud account used to purchase the ApsaraDB for RDS instance, username, password, and database name are correct.

- Problem description: When I create a connection to a user-created MySQL data store on an ECS instance, the connectivity test fails.

  Solution:

  i. Check the connectivity between the network and the specified port. Check the firewall and security group settings of the ECS instance.

  ii. Check the access control, permissions, and remote logon of the user-created MySQL data store.

  iii. Confirm that the username, password, and IP address and port number in the specified JDBC URL are correct for the user-created MySQL data store.

  iv. If the ECS instance is in a VPC, you can configure sync nodes for the MySQL connection only in the code editor. The connectivity test of the MySQL connection also fails. In this case, you can create custom resource groups to run these sync nodes.

Parameters

- Problem description: When I create a MySQL connection, the connectivity test fails. The following error message is returned:

```
Connection failed, data store connectivity test failed, database connection failed... err
or message: No suitable direver found for...
```

  Solution: The error may be caused by an invalid format of the JDBC URL. When you enter the JDBC URL, do not add spaces or special characters to the URL. The correct format is `jdbc:mysql://ServerIP:Port/Database`.

- Problem description: When I create a MongoDB connection, root is used as the username for connecting to the database. An error message is returned.

  Solution: When you create a MongoDB connection, you must use the username that is created for the database where the table to be synchronized resides. Do not use root as the username. For example, to import the name table from the test database, enter test as the database name. Enter a username that is created for the specified database, instead of root. For example, use a username that is created for the test database.

- Problem description: When I create an RDS connection, the database cannot be connected.

  Solution: Check the account ID that you enter for the RDS connection. You must enter the ID of the Alibaba Cloud account used to purchase the ApsaraDB for RDS instance. Do not enter the account ID of a Resource Access Management (RAM) user.

- Problem description: When I create the default MaxCompute connection, the connectivity test fails.

  Solution: The default MaxCompute connection named odps_fisrt is created by the system. You do not need to create it again.

- Problem description: I want to create a HybridDB for PostgreSQL connection in DataWorks.

  Solution: You can select PostgreSQL to create a HybridDB for PostgreSQL connection.

- Problem description: A Distributed Relational Database Service (DRDS) instance does not have a public endpoint. I want to map the internal endpoint of the DRDS instance to a custom domain name and create a DRDS connection.

  Solution: The domain name mapping method is not supported. You cannot use this method to create a DRDS connection.

- Problem description: When I create a connection to an RDS data store for which a whitelist is configured, the error message `user not exist ip white list reference` is still returned.

  Solution: The error occurs because the entered username is invalid. Check the username that you enter for the RDS connection.

Permissions

- Problem description: When I create an AnalyticDB for MySQL connection, the connectivity test fails. The following error message is returned:

  ```
  Database connection failed, database connection string: ${jdbcUrl}, username: XXXXXX, err
  or message: You don't have privilege for connecting database 'dw', userId=RAM$XXX, schema
  Id=XX
  ```

  Solution: Check whether the RAM user, whose AccessKey ID and AccessKey secret are entered for the AnalyticDB for MySQL connection, has the permission to access the AnalyticDB for MySQL data store.

  Users who can access an AnalyticDB for MySQL data store are authenticated based on the Alibaba Cloud account. You can authorize RAM users under your Alibaba Cloud account to access your AnalyticDB for MySQL data store.

- Problem description: After a RAM user logs on to the DataWorks console, the user has no permission to view or create connections. An error message is returned.

  Solution: Only RAM users who have the workspace administrator permissions can create, delete, or modify connections.

8.Data governance

- Data Map
  - Why is the value of the Storage parameter quite different from the statistics shown in the storage trend chart on the Overview page of Data Map?
  - Why does Data Map not display the real-time lineage information?
  - What do I do if no search results are returned when I search for a new table in Data Map?
  - How do I notify the owners of descendant nodes of the changes to the business logic?
- Data Security Guard

  Why does the de-identification feature fail to take effect for the query results sometimes after I configure de-identification rules in Data Security Guard?

Why is the value of the Storage parameter quite different from the statistics shown in the storage trend chart on the Overview page of Data Map?

The value of the Storage parameter is updated in real time, while the storage trend chart displays the storage statistics at specific points in time. If a small number of temporary tables are generated after a specific point in time, the statistics at the specific point in time in the storage trend chart may be inconsistent with the value of the Storage parameter.

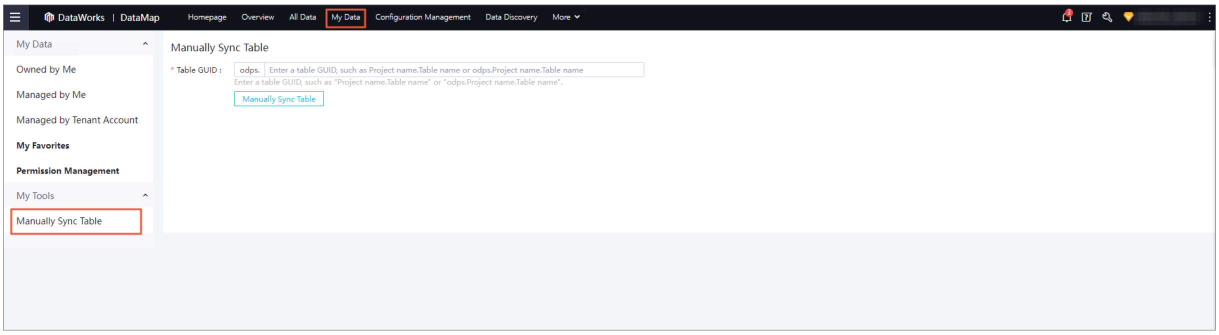
Why does Data Map not display the real-time lineage information?

The lineage information is updated with a latency of at least one day.

What do I do if no search results are returned when I search for a new table in Data Map?

You can use the manual synchronization feature in Data Map if the new table cannot be found in Data Map or the data of the table is inconsistent with that in Data Map. Manual synchronization is required in the following scenarios:

- The schema changes of the tables in the Workspace Tables pane are not updated to Data Map.
- No search results are returned when you search for a new table in Data Map.
- No search results are returned when you search for a new table in Data Integration.

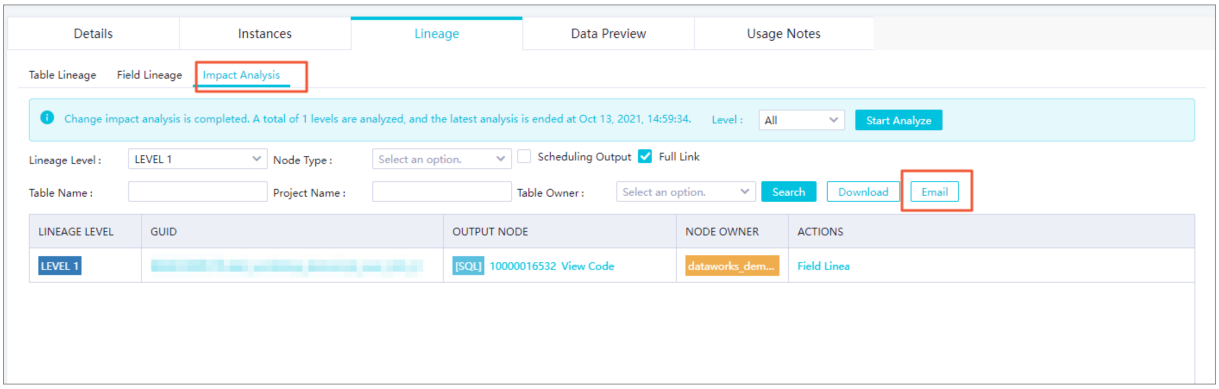


* Table GUID : odps. Enter a table GUID, such as Project name.Table name or odps.Project name.Table name

Enter a table GUID, such as "Project name.Table name" or "odps.Project name.Table name".

How do I notify the owners of descendant nodes of the changes to the business logic?

On the **Data Map** page, find and click the table that you want to view. On the table details page that appears, choose **Lineage > Impact Analysis**. Then, perform operations as shown in the following figure to inform multiple owners by sending email notifications.



Change impact analysis is completed. A total of 1 levels are analyzed, and the latest analysis is ended at Oct 13, 2021, 14:59:34.

Level : All

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	OUTPUT NODE	NODE OWNER	ACTIONS
LEVEL 1	[redacted]	[SQL] 10000016532 View Code	dataworks_dem...	Field Linea

Why does the de-identification feature fail to take effect for the query results sometimes after I configure de-identification rules in Data Security Guard?

The de-identification feature works only if 80% of the records that are queried match the de-identification rules.

9.Intelligent monitoring

This topic provides answers to frequently asked questions about the intelligent monitoring module.

- What can I do if I am unable to receive alert notifications after I configure an alert in Operation Center?
- What can I do if I want to disable alerting for a node?
- Why is a baseline in the Empty Baseline state?
- Why is no alert notification sent for a baseline in the Overtime state?
- Can I disable alerting for a node that slows down?
- Why am I unable to receive an alert notification for a node error?
- What can I do if I receive an alert notification at night?

What can I do if I am unable to receive alert notifications after I configure an alert in Operation Center?

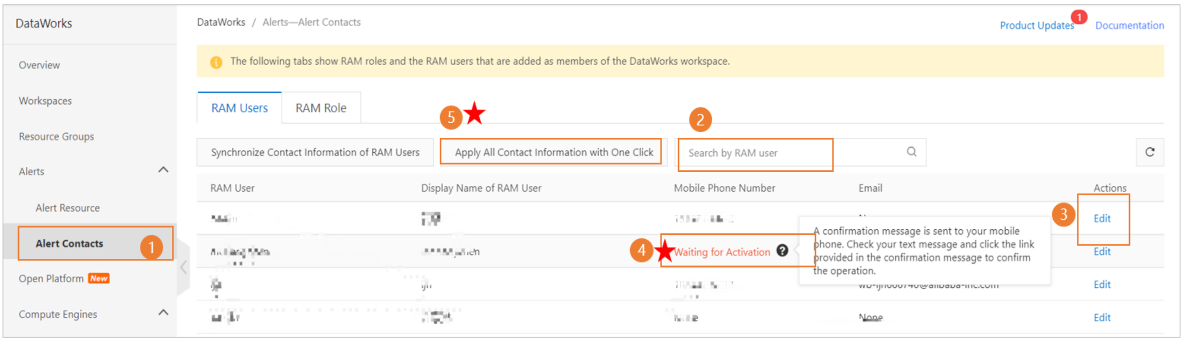
Check whether the alert is triggered. If the alert is triggered but you cannot receive alert notifications, troubleshoot the issue based on the notification method that you configure. The notification methods include **text message**, **email**, and **DingTalk group message**.

- **Check whether the alert is triggered**
  - If the alert is configured for an auto triggered node, check the status of the node instances on the **Cycle Instance** page in Operation Center and whether the alert can be triggered for the node.

    For more information about the conditions for triggering a custom alert, see Manage custom alert rules. For more information about the conditions for triggering a baseline alert, see Monitor.
  - If the alert is configured for a real-time synchronization node, check the status of the real-time synchronization node. To do so, go to **Operation Center** and choose **RealTime Task Maintenance** > **Real Time DI** in the left-side navigation pane.
- **Failed to receive alert notifications in text messages or emails after the alert is triggered**

  Check whether the phone numbers and email addresses of alert contacts are properly configured in DataWorks.

  On the homepage of the DataWorks console, choose **Alerts** > **Alert Contacts** in the left-side navigation pane. On the Alert Contacts page, you can view and configure alert contacts. The following figure shows the steps.



The following tabs show RAM roles and the RAM users that are added as members of the DataWorks workspace.

A confirmation message is sent to your mobile phone. Check your text message and click the link provided in the confirmation message to confirm the operation.

  If the specified alert contacts cannot receive alert notifications after the alert is triggered, perform the following checks on the **Alert Contacts** page:

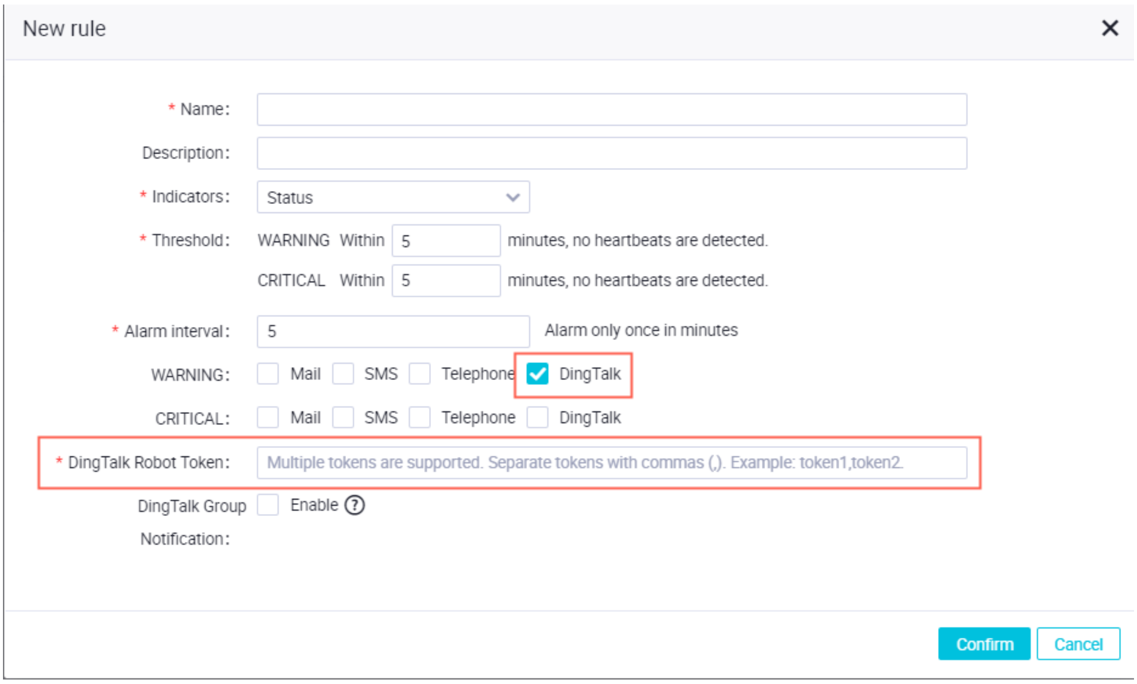
  - Check whether the phone numbers and email addresses of the alert contacts are configured.
  - Check whether the alert contacts activate the phone numbers and email addresses that have been configured.

> **Note**
>
> - Alibaba Cloud accounts and RAM users that are granted the AliyunDataWorksFullAccess permission can configure contact information for RAM users. For more information, see Configure and view alert contacts.
> - If the phone numbers or email addresses of the alert contacts are not properly configured, the system sends alert notifications to the recipients that are listed on the Common Settings page. As a result, the specified alert contacts cannot receive the alert notifications.

- **Failed to receive alert notifications in DingTalk groups after the alert is triggered**

  Perform the following checks:

  - **Check whether the webhook URL of the DingTalk chatbot is correct on the alert configuration page**
    - If the alert is configured for an auto triggered node, check whether the webhook URL is valid. For example, check for extra spaces.
    - If the alert is configured for a real-time synchronization node, check whether the token information of the DingTalk chatbot is correct.



New rule ✕

* Name:

Description:

* Indicators: Status

* Threshold: WARNING Within 5 minutes, no heartbeats are detected.

CRITICAL Within 5 minutes, no heartbeats are detected.

* Alarm interval: 5 Alarm only once in minutes

WARNING: □ Mail □ SMS □ Telephone ☑ DingTalk

CRITICAL: □ Mail □ SMS □ Telephone □ DingTalk

* DingTalk Robot Token: Multiple tokens are supported. Separate tokens with commas (,). Example: token1,token2.

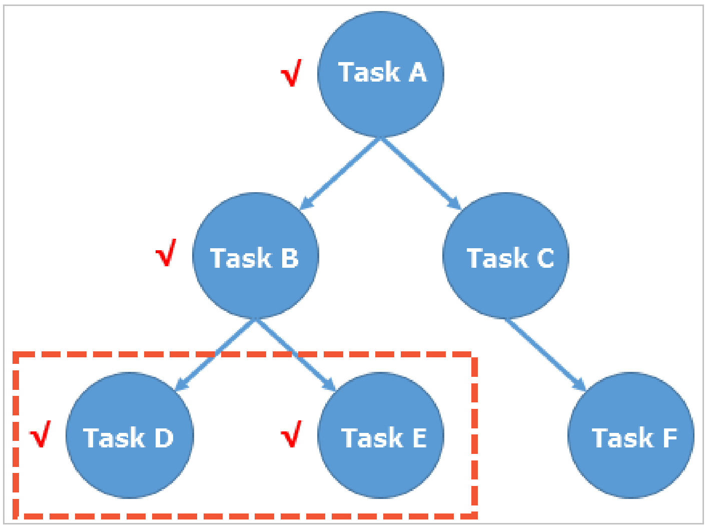
DingTalk Group Notification: □ Enable ?

  - **Check whether the DingTalk chatbot is correctly configured**

    When you add a chatbot to the DingTalk group for receiving alert notifications, set the **Security Settings** parameter to **Custom Keywords** and make sure that the keywords include DataWorks. For more information, see the "Send alert notifications to a DingTalk group" section of the Manage custom alert rules topic.

What can I do if I want to disable alerting for a node?

After a baseline is created and enabled, the intelligent monitoring module monitors all nodes in the baseline and their ancestor nodes. If a node in the baseline or an ancestor node of the baseline affects data generation of the monitored nodes in the baseline, the intelligent monitoring module sends an alert notification to the node owner. For more information, see Monitor.



In the example shown in the preceding figure, DataWorks has six nodes, and Nodes D and E belong to a baseline. The intelligent monitoring module monitors Nodes D and E and all their ancestor nodes. In this case, the intelligent monitoring module detects errors or slowdowns on Node A, B, D, or E. Nodes C and F are not monitored by the intelligent monitoring module.

- If you want to disable alerting for Nodes D and E, contact the baseline owner to remove Nodes D and E from the baseline.
- Nodes A and B are ancestor nodes of Nodes D and E and may affect data generation of the monitored nodes in the baseline. If an error or a slowdown occurs on Node A or B, the intelligent monitoring module sends an alert notification to the node owner.

  If you want to disable alerting for Node A or B, contact the owners of Nodes D and E to delete the dependency of Nodes D and E on Node A or B.

Why is a baseline in the Empty Baseline state?

In the following scenarios, a baseline may enter the Empty Baseline state:

- Scenario 1: A node can belong to only one baseline. If you add a node to another baseline, the system removes the node from the current baseline and adds it to the specified baseline. If all nodes are removed from a baseline, the baseline enters the Empty Baseline state.
- Scenario 2: On the day when a baseline is created, the baseline is in the Empty Baseline state. After you enable the baseline, a baseline instance is generated on the next day.
- Scenario 3: You specify an invalid point in time as the baseline time for an auto triggered node instance in an hour-level baseline.

> **Note** For example, the node is scheduled to run at 6:00 and 18:00 every day. However, you specify 6:00 and 18:00 as the baseline time when you create a baseline and add the node to the baseline.

Why is no alert notification sent for a baseline in the Overtime state?

Baseline monitoring is controlled by the baseline switch and enabled for nodes. Overtime is a baseline state, which indicates that the nodes in a baseline are not complete when the baseline time is reached. If all nodes in a baseline are run as expected, no alert is triggered even if the baseline enters the Overtime state. This is because the intelligent monitoring module cannot determine which node has an error.

If the baseline enters the Overtime state when all nodes are run as expected, consider the following reasons:

- The baseline time is improper.
- The node dependency is improper.

Can I disable alerting for a node that slows down?

The intelligent monitoring module notifies you of a node slowdown only if a node meets both of the following conditions:

- The node is an ancestor node of an important baseline.
- Compared with its historical performance, the node does slow down.

You can view the descendant baseline affected by the node on the **Event management** tab in Operation Center. Then, you can confirm the impact with the party whose baseline contains descendant nodes of your node.

- If the node slowdown has a minor impact, you can disable alerting.
- If the node slowdown has a major impact, maintain your node properly.

Why am I unable to receive an alert notification for a node error?

The intelligent monitoring module notifies you of a node error only if a node meets one of the following conditions:

- The node is an ancestor node of a baseline that is enabled. For more information about baselines, see 基线管理.
- A custom alert rule is configured. For more information about how to configure a custom alert rule, see Manage custom alert rules.

What can I do if I receive an alert notification at night?

1. Go to the **DataStudio** page.
    i. Log on to the DataWorks console.
    ii. In the left-side navigation pane, click **Workspaces**.
    iii. In the top navigation bar, select the region in which the workspace that you want to manage resides. Find the workspace and click **DataStudio** in the Actions column.
2. On the DataStudio page, click the ☰ icon in the upper-left corner and choose **All Products > Data Development And Task Operation > Operation Center**.
3. In the left-side navigation pane, choose Alarm > **Smart Baseline > Event management**.
4. On the **Event management** tab, disable alerting. You can disable alerting in one of the following ways:
    - Handle the event that triggers the alert. Then, alerting is temporarily disabled for the event.

a. Find the event and click **Handle** in the Operation column.

b. In the **Handle Event** dialog box, set the **Handling Time** parameter.

c. Click **OK**.

> ? **Note** DataWorks records the event handling operation and pauses alerting for the event when the event is being handled.

- Ignore the event that triggers the alert. Then, alerting is permanently disabled for the event.

a. Find the event and click **Ignore** in the Operation column.

b. In the **Ignore Event** message, click **OK**.

> ? **Note** DataWorks records the event ignoring operation and permanently stops alerting for the event.

10.DataService Studio

This topic describes the FAQ about the DataService Studio service.

- Why does DataService Studio fail to access a data store?
- Why does DataService Studio fail to connect to a user-created database hosted on an ECS instance?
- Do I need to activate the API Gateway service?
- How can I configure a connection?
- What is the difference between the codeless UI and the code editor for creating API operations?
- What is the role of an API group in DataService Studio? Is it the same as that in API Gateway?
- How can I configure an API group appropriately?
- How many API groups can I create?
- When do I need to enable the pagination feature to display return results on multiple pages?
- Do API operations created in DataService Studio support POST requests?
- Do API operations created in DataService Studio support the HTTPS protocol?

Why does DataService Studio fail to access a data store?

The data store is configured with a whitelist. To properly connect DataService Studio to the data store, add the IP addresses of DataService Studio in the corresponding region to the whitelist.

Region	CIDR block of DataService Studio	Public IP address
China (Beijing)	11.193.100.0/24,11.193.199.0/24	39.106.244.50,47.95.63.101,47.95.63.93,39.106.244.48
China (Zhangjiakou)	11.112.227.0/24	-
China (Shanghai)	11.193.96.0/24,11.193.48.0/24,11.193.108.0/24	101.132.31.146,106.15.14.240,106.15.14.75,101.132.31.221
China (Hangzhou)	11.197.246.0/24,11.193.55.0/24	101.37.74.122,114.55.197.231,114.55.198.83,101.37.74.206
China (Shenzhen)	11.193.103.0/24 and 11.193.94.0/24	120.78.45.154,120.78.46.137,120.78.46.107,120.78.45.140
China (Chengdu)	11.195.52.0/24	-
Japan (Tokyo)	11.199.250.0/24	-
US (Silicon Valley)	11.193.216.0/24	-
Singapore	11.197.188.0/24,11.197.227.0/24	-
China (Hong Kong)	11.193.200.0/24,11.193.12.0/24	-
Germany (Frankfurt)	11.199.93.0/24	-
Indonesia (Jakarta)	11.194.50.0/24	-

Why does DataService Studio fail to connect to a user-created database hosted on an ECS instance?

DataService Studio cannot connect to user-created databases hosted on Elastic Compute Service (ECS) instances by using an internal endpoint.

Do I need to activate the API Gateway service?

API Gateway provides you with high-performance and highly available API hosting services. If you want to make your API operations available to others, activate the API Gateway service first.

How can I configure a connection?

To configure a connection, follow these steps: Go to the Workspace Management page of the target workspace. On the **Workspace Management** page, click Data Source in the left-side navigation pane. On the **Data Source** page, click Add a Connection or Add Connections in the upper-right corner to create and configure one or more connections. DataService Studio automatically reads data from the connections that you have configured.

Does DataService Studio support Lightning connections?

No, DataService Studio currently does not support Lightning connections. To use Lightning data, purchase a Hologres instance.

Note Currently, you cannot create API operations in DataService Studio for Hologres.

What is the difference between the codeless UI and the code editor for creating API operations?

The code editor provides more powerful features. For more information, see Create an API in the code editor.

What is the role of an API group in DataService Studio? Is it the same as that in API Gateway?

An API group is a set of API operations specific to a feature or scenario. It is the smallest organization unit in DataService Studio, which is similar to an API group in API Gateway. After you publish an API operation created in DataService Studio to API Gateway, API Gateway automatically creates an API group with the same name.

How can I configure an API group appropriately?

Typically, an API group includes API operations that provide similar features or resolve a specific issue. For example, a weather API group can include API operations that are used to check the weather by city and by longitude and latitude.

How many API groups can I create?

You can create a maximum of 100 API groups with an Alibaba Cloud account.

When do I need to enable the pagination feature to display return results on multiple pages?

By default, an API call returns a maximum of 2,000 records. If an API call may return more than 2,000 records, enable the pagination feature. If you do not specify any request parameters, an API call usually returns a large number of records. In this case, the system automatically enables the pagination feature.

Do API operations created in DataService Studio support POST requests?

Currently, API operations created in DataService Studio support GET and POST requests.

Do API operations created in DataService Studio support the HTTPS protocol?

Currently, API operations created in DataService Studio support both HTTP and HTTPS protocols.

11.Security Center

This topic describes the FAQ about Security Center.

- What permissions can I request in Security Center?
- What is the relationship between Data Management and Security Center?
- Why cannot I select fields when I request permissions?
- Who will handle my request?
- Why do I find two requests on the My Requests tab after I submit only one request?
- I request permissions on a field for one month only. Why does the validity period of the permissions become permanent after my request is approved?
- Why do I have permissions on some tables and fields on which I have not requested any permissions?
- Why does a request disappear from the Pending My Approval tab before I handle it?
- What can I do if the message "An error occurred in the MaxCompute project" appears when I specify the workspace and environment?
- Why do I fail to revoke permissions on a field?
- Why do I fail to request permissions by using my Alibaba Cloud account?
- In Security Center, can I view the permission requesting and approval records of Data Management?
- Can I revoke permissions based on the requesting records in Security Center?
- A permission request submitted in Data Management has not been approved yet. Do I need to submit it again in Security Center?
- How do I specify the LabelSecurity parameter for fields?

What permissions can I request in Security Center?

In Security Center, you can request permissions on tables in DataWorks workspaces in the development environment and production environment.

What is the relationship between Data Management and Security Center?

Security Center is a service that upgrades and replaces the permission and security features in Data Management. You can click **My Permissions** in the left-side navigation pane of the **Security Center** page to view the permissions requested or granted by using the `odpscmd grant` command in **Data Management**.

If you want to request other permissions and handle permission requests on the graphical user interface (GUI), go to **Security Center** and perform operations as required. The **Data Management** service no longer supports permission requesting or approval.

Why cannot I select fields when I request permissions?

If LabelSecurity is enabled for a workspace, you can request permissions on fields in this workspace. If LabelSecurity is disabled for a workspace, you can request permissions only on tables in this workspace.

Who will handle my request?

Your request will be handled by a workspace administrator or a table owner. After either of them approves or rejects your request, the request is closed.

Why do I find two requests on the My Requests tab after I submit only one request?

The tables in your request belong to two owners. In this case, Security Center automatically splits your request into two by table owner.

I request permissions on a field for one month only. Why does the validity period of the permissions become permanent after my request is approved?

The security level of this field is zero or not higher than the security level of your account.

Why do I have permissions on some tables and fields on which I have not requested any permissions?

The possible reasons are as follows:

- An administrator has granted the permissions to you by running commands in the DataWorks console.
- After your request is approved in Security Center, Security Center also grants you the permissions on fields whose security level is zero or not higher than the security level of your account, even though you have not requested the permissions.

Why does a request disappear from the Pending My Approval tab before I handle it?

Another workspace administrator or table owner has approved the request before you handle it. The approved request is closed and no longer appears on the **Pending My Approval** tab.

What can I do if the message "An error occurred in the MaxCompute project" appears when I specify the workspace and environment?

Send the error message and error code to the workspace administrator for troubleshooting.

Why do I fail to revoke permissions on a field?

You can only revoke permissions on the fields whose security level is higher than the security level of your account.

Why do I fail to request permissions by using my Alibaba Cloud account?

By default, an Alibaba Cloud account has all permissions. Therefore, you do not need to request permissions for your Alibaba Cloud account. Entry points to unnecessary operations such as permission requesting are hidden for an Alibaba Cloud account. This does not affect the use of the account.

In Security Center, can I view the permission requesting and approval records of Data Management?

Security Center and Data Management have not synchronized permission requesting or approval records yet. You need to go to **Data Management** to view the permission requesting and approval records of Data Management.

Can I revoke permissions based on the requesting records in Security Center?

Currently, Security Center is not the only service that provides authorization. To facilitate permission revocation, the Authorizations page in Security Center provides an access control list (ACL) of all users, regardless of the authorization channel. You can revoke any granted permissions without using the requesting records.

A permission request submitted in Data Management has not been approved yet. Do I need to submit it again in Security Center?

Security Center and Data Management have not synchronized permission requesting and approval records yet. You need to submit the permission request again in Security Center.

How do I specify the LabelSecurity parameter for fields?

You need to go to **Data Map** to set the LabelSecurity parameter for fields.

12.App Studio

This topic describes the FAQ about App Studio.

- What issue may occur if I set a breakpoint in a main function?
- What can I do if the startup time of a program is too long?
- Can I add a variable in the Watch section?

What issue may occur if I set a breakpoint in a main function?

```
public class Main {
    public static void main(String[] args) {
        SpringApplication.run(Main.class,args) ; // Set a breakpoint in this line and then
start debugging.
    }
}
```

After you set a breakpoint in a main function, if you select **Step Over** under Debug when the thread suspends at the breakpoint, the thread may fail to be resumed.

Currently, DataWorks does not support multi-thread programming or debugging. Therefore, do not set a breakpoint in a main function.

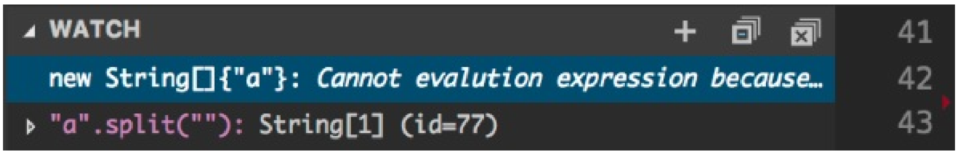
What can I do if the startup time of a program is too long?

If you set a breakpoint in the line that defines a function and then select Start Debugging under Debug, it will take a long time to start the program.

This symptom is normal. We recommend that you use line breakpoints instead of method breakpoints.

Can I add a variable in the Watch section?

You cannot add some variables in the Watch section.



```
new String[]{"a"}: Cannot evalution expression because…
"a".split(""): String[1] (id=77)
```

As shown in the preceding figure, DataWorks throws an exception for the first variable and fails to execute the code with the second variable.

13.Stream Studio

This topic describes the FAQ about Stream Studio.

- What computing engine do I need to activate before I can use Stream Studio?
- Which modes of Realtime Compute does Stream Studio support?
- Are there differences between Realtime Compute in the shared mode and that in the exclusive mode for Stream Studio?
- Where can I create a Realtime Compute project and bind it to a DataWorks workspace?
- What are the advantages of the directed acyclic graph (DAG) mode in Stream Studio? What are the similarities and differences between the DAG mode and SQL mode?
- What types of SQL does Stream Studio support?
- What can I do if I cannot create a node on the Stream Studio page?

What computing engine do I need to activate before I can use Stream Studio?

Before you can use Stream Studio, you must activate Realtime Compute because Stream Studio is a development platform based on Alibaba Cloud Realtime Compute.

Which modes of Realtime Compute does Stream Studio support?

Stream Studio supports Realtime Compute in the shared mode and exclusive mode.

Are there differences between Realtime Compute in the shared mode and that in the exclusive mode for Stream Studio?

Yes, there are differences. For security purposes, Realtime Compute in the shared mode does not support UDFs, whereas that in the exclusive mode supports. If you have high requirements on the performance and features, we recommend that you use Realtime Compute in the exclusive mode.

Where can I create a Realtime Compute project and bind it to a DataWorks workspace?

You can create a Realtime Compute project in the Realtime Compute console after logon.

To bind the created Realtime Compute project to a DataWorks workspace, follow these steps: Log on to the DataWorks console and click Workspaces in the left-side navigation pane. On the **Workspaces** page that appears, find the target workspace and bind the Realtime Compute project to the workspace. You can also bind the Realtime Compute project to a new workspace when you create the workspace. After you bind the Realtime Compute project to a workspace, you can go to Stream Studio.

What are the advantages of the directed acyclic graph (DAG) mode in Stream Studio? What are the similarities and differences between the DAG mode and SQL mode?

Stream Studio supports both the DAG mode and the SQL mode to develop real-time computing nodes. In the DAG mode, you can perform drag-and-drop operations on components to configure real-time computing nodes without writing code. You can freely switch between the DAG mode and the SQL mode.

What types of SQL does Stream Studio support?

Realtime Compute is based on Apache Flink. Therefore, Stream Studio supports Flink SQL.

What can I do if I cannot create a node on the Stream Studio page?

Before you can properly use Stream Studio, you must activate Realtime Compute, create a project, and then bind the project to a DataWorks workspace.

If you still cannot create a node on the Stream Studio page after completing the preceding steps, follow these steps:

- Clear the browser cache.
- Verify that the AccessKey ID and AccessKey secret are bound to Stream Studio.
- Verify that the workspace to which the Realtime Compute project is bound has the permission to access Realtime Compute.