Alibaba Cloud DataWorks

Best Practices

Issue: 20190115



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.

- 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 disseminat ed 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 due to product version upgrades, adjustment s, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and the updated versions of this document will be occasionally released through Alibaba Cloud-authorized channels. You shall 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 the document in the context that Alibaba Cloud products and services are provided on an "as is", "with all faults" and "as available" basis. 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 bear any liability for any errors or financial losses incurred by any organizations, companies, or individuals arising from their download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, bear responsibility for any indirect, consequential, exemplary, incidental, 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 content of the Alibaba Cloud website, including but not limited to works, products , images, archives, information, materials, website architecture, website graphic layout, and webpage design, are intellectual property of Alibaba Cloud and/or its affiliates. This intellectu

al property includes, but is not limited to, trademark rights, patent rights, copyrights, and trade secrets. No part of the Alibaba Cloud website, product programs, or content 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", "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 contact Alibaba Cloud directly if you discover any errors in this document.

Generic conventions

Table -1: Style conventions

Style	Description	Example
•	This warning information 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.
	This warning information 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 restore business.
	This indicates warning information, supplementary instructions, and other content that the user must understand.	• Notice: Take the necessary precautions to save exported data containing sensitive information.
	This indicates supplemental instructio ns, best practices, tips, and other content that is good to know for the user.	Note: You can use Ctrl + A to select all files.
>	Multi-level menu cascade.	Settings > Network > Set network type
Bold	It is used for buttons, menus, page names, and other UI elements.	Click OK .
Courier font	It is used for commands.	Run the cd /d C:/windows command to enter the Windows system folder.
Italics	It is used for parameters and variables.	bae log listinstanceid Instance_ID
[] or [a b]	It indicates that it is a optional value, and only one item can be selected.	ipconfig [-all -t]
{} or {a b}	It indicates that it is a required value, and only one item can be selected.	<pre>swich { stand slave }</pre>

Contents

Legal disclaimer	I
Generic conventions	I
1 Simple mode and standard mode	1
2 Workshop	3
2.1 Workshop course introduction	3
2.2 Data acquisition: log data upload	4
2.3 Data processing: user portraits	21
2.4 Data quality monitoring	35
3 Best practices for setting scheduling dependencies	44

1 Simple mode and standard mode

The new version of DataWorks introduces both simple and standard modes, this article introduces you to the differences between simple and standard modes.

Simple Mode

A simple mode refers to a DataWorks project that corresponds to a MaxCompute project and cannot set up a development and production environment, you can only do simple data development without strong control over the data development process and table permissions.

The advantage of the simple mode is that the iteration is fast, and the code is submitted without publishing, it will take effect.

The risk of a simple mode is that the development role is too privileged to delete the tables under this project, there is a risk of table permissions.

Standard Mode

Standard mode refers to a DataWorks project corresponding to two MaxCompute projects, which can be set up to develop and produce dual environments, improve code development specifications and be able to strictly control table permissions, the operation of tables in production environments is prohibited, and the data security of production tables is guaranteed.

- All Task edits can be performed only in the development environment, and the Production Environment Code cannot be directly modified, reduce the production environment code modification entry, as much as possible to ensure the production environment code stability.
- The development environment does not turn on task scheduling by default, avoid the development of environmental project cycle operation and production of environmental projects to seize resources, the stability of the operation of production environment tasks is better guaranteed.
- The production environment runs with a default production account, all the tables produced by the production account belong to the main account, you need to use production tables during the development process, all of which need to be applied separately, better control of table permissions.

When creating a project, select **project mode** as the standard mode, fill in the project name and project description, the remaining configuration item select the default value.



The MaxCompute access identity of the production environment cannot be modified to a personal account, otherwise, the data security of the production environment cannot be guaranteed.

Create Project		×
* project name :	DataWorks_DOC	
Display name :		
* Project mode :	In simple mode (single environment) \square	
project description :		
Advanced Settings		
* Enable scheduling Frequency :	on Ø	
Enable Select result downloads in this	on Ø	
project :		
for MaxCompute		
* MaxCompute Project Name:	DataWorks_DOC	0
 MaxCompute access identity: 	Project Leader Account	
★ Quota group:	Pay per view default resource _ \sim	
	Previous	create

2 Workshop

2.1 Workshop course introduction

This module introduces you to the design ideas and core capabilities of DataWorks, to help you gain insight into the ideas and capabilities of Alibaba Cloud DataWorks.

Course Overview

Course duration: Two hours, using an online learning method.

Course object: for all new and old users of DataWorks, such as Java engineer, product operation , HR, etc, as long as you are familiar with standard SQL, you can quickly master the basic skills of DataWorks, you don't need to know much about the principles of data warehouses and MaxCompute. However, it is also recommended that you further study the DataWorks course to gain insight into the basic concepts and functions of DataWorks.

Course objective: Take the common real-world massive log data analysis task as the curriculum background, after completing the course, you will be able to understand the main features of DataWorks, able to demonstrate content according to the course, independently complete data acquisition, data development, task operations and other data jobs common tasks.

This course includes the following:

- Product introduction: You will learn about DataWorks' development history, its overall architectu re, and its modules and their relationships.
- Data Acquisition: Learn How to synchronize data from different data sources to MaxCompute, how to quickly trigger task runs, how to view task logs, and so on.
- Data Processing: learn how to run a data flow chart, how to create a new data table, how to create a data process task node, how to configure periodic scheduling properties for tasks.
- Data quality: Learn how to configure monitoring rules for data quality for tasks, ensure that the task runs quality issues.

DataWorks introduction

DataWorks is a big data research and development platform, using MaxCompute as the main calculation engine, including data integration, data modeling, data development, operations and operations monitoring, data management, data security, data quality, and other product functions

. At the same time, with the algorithm platform PAI to get through, complete link from big data development to Data Mining and machine learning.

Data Collection

For more information on data acquisition, see Data acquisition: log data upload.

Data Processing

For details on data processing, see Data processing: user portraits.

Data quality

For more information on data quality, see *Data quality monitoring*.

Learning to answer questions

If you encounter problems in the learning process, you can add DingTalk groups: 11718465, consulting Alibaba cloud technical support.

2.2 Data acquisition: log data upload

Related Products

The big data products involved in this experiment are *MaxCompute (big data computing services)*. And *DataWorks (data factory, original big data development kit)*.

Prerequisites

Before you begin this lab, you need to make sure you have an Alibaba Cloud account and have a real name.

Activate MaxCompute



Note:

If you have already activated MaxCompute, skip this step to create the project space directly.

- 1. Log in to the *Alibaba Cloud website*, click **Log in** in the upper-right corner to fill in your Alibaba Cloud account and password.
- Select Products > Analytics & Big Data > MaxComputute and go to the MaxCompute product details page.

C Alibaba Cloud Cov Worldwide Cloud Services Partner		Contact Sales	Search	Q	🕲 International - English 🗸
Why Us 🗸 Products 🔨	Solutions V Pricing	Marketplace	Resources 🗸	Support 🗸	Documentation
Elastic Computing	E-MapReduce		Dataphin	(Coming Scon)	
Storage & CDN					
Networking	A fast and fully-hosted		Machine L	earning Platform	n For Al meet your machin
Database Services	DataWorks (Beta)		Elasticsea	rch	
Security	> A full data warehousin				
Monitoring & Management	> Data Integration Real-time and Offline I		Data Lake	Analytics (Con	ing Soon) Id interactive analyt
Domains & Websites					
Analytics & Big Data	Quick BI Intelligent analytics &				
Application Service	DataV				
Media Services					
Middleware	Image Search High-precision visual s	search product solution			
Cloud Communication	>				
Apsara Stack	Chatbot platform for si	xx (1993) mart dialogue interacti.			
Internet of Things					

- 3. Click Start now.
- 4. Select Pay-As-You-Go, click Buy Now.

Create Project

- 1. Log on to the *DataWorks console* by using a primary account.
- 2. You can create a MaxCompute project in two ways.
 - On the console overview page, go to Common FunctionsCreate Project.

	Overview Project List	Schedule Resource List	
🜀 DataWorks Da	ataStudio∙Data Integration∙MaxComp		2 0 0
Fast Entry	Bata Internation	Operation Center	The Data Integration Launch Deport multiple development modes Deport more data channels
Project	units integration	projects	
DataWorks_DOC China East 2	DataWorks演示项目 China East 2	DataWorks說程_简单 China East 2	
Created 2018-08-27 13:32:17 Engine: MaxCompute Service: Data Studio Data Integration Data Management Op	Created 2018-08-20 19:27.18 Engine Mai/Compute Service Data Studio Data Integration Data Management Op	Created 2018-07-26 16:17:55 Engine MaxCompute Service Data Studio Data Integration Data Management Op	
Config Data Studio Data Integration	Config Data Studio Data Integration	Config Data Studio Data Integration	
Common Functions			

 Fill in the configuration items in the Create Project dialog box. Select a region and a calculation engine service.



If you have not purchase the relevant services in the region, it is directly display that there is no service available in the Region. The data development, O&M center, and data management are selected by default.

Configure the basic information and advanced settings for the new project, and click Create project.

Create Project		×
* project name :	DataWorks_DOC	
Display name :		
* Project mode :	In simple mode (single environment) 🗹	
project description :		
Advanced Settings		
* Enable scheduling Frequency :	on Ø	
Enable Select result downloads in this		
project :		
for MaxCompute		
* MaxCompute Project Name:	DataWorks_DOC	
 MaxCompute access identity: 	Project Leader Account	
* Quota group:	Pay per view default resource _ $ \smallsetminus $	
	Previous	create



Note:

- The project name needs to begin with a letter or underline, and can only contain letters, underscores, and numbers.
- The project name is globally unique, it is recommended that you use your own easy-todistinguish name as the project space name for this lab.

 Once the project has been created successfully, you can select the Project List page to Data Studio after viewing the project space.

Overview Project List Schedule Resource List						
China North 2 China East 1 Ohina Asia Pacific NE 1 Middle East 1 A	East 2 Asia Pacific SE 2 China sia Pacific SU 1 Asia Pacific SE 5 Search	South 1 Hong Kong US West 1	Asia Pacific SE 1 US East 1 EU C	entral 1 Asia Pacific	: SE 3	Create Project Refresh
Project / display name	Project mode	Create time	administrator	status	Subscribed service	operation
nodi nodi	In simple mode (single environ ment)	2018-09-10 10:48:11	dataworks_demo2	normal	~	Config Data Studio Modify service More 🗸

Create data source

Note:

Based on the scenario simulated by this lab, you need to distribute to create both the OSS data source and the RDS data source.

- Create a new OSS data source
 - 1. Select the Data Integration > Data Source Page, and click Add Data Source.

🙆 Data Integrati	On DataWorks_DOC	¥			Project Space	English
= • Overview	Data Source	Data Source : All	Deta Source :			Add Data Source
Teska		iype	Name			
Resource Consumptio	Data Source Name	Data Source Type	Link Information	Description	Created At	Actions
 Synchronization Reso. 	odps_first	CDPS	00PS Endpoint: http://service.odps.aliyun.com 00PS Project Name: DataWorks_DOC Access M: UTAbuCi7pq.USeQ	api connection from odos celo e ngine 61155	2018-08-27 13:32:26	
Resource Group						
Client Data Collection						

2. Select the data source type as OSS, with other configuration items as follows.

* Data Source Name :	oss_workshop_log	
Description :		
* Endpoint :	http://oss-cn-shanghai-internal.aliyuncs.com	0
* Bucket:	dataworks-workshop	0
* Access Id :	LTAINEhd4MZ8pX64	0
* Access Key:	•••	
Test Connectivity:	Test Connectivity	

Parameters:

- Endpoint: http://oss-cn-shanghai-internal.aliyuncs.com
- bucket: dataworks-workshop
- AK ID: LTAINEhd4MZ8pX64
- AK Key: IXnzUngTSebt3SfLYxZxoSjGAK6IaF
- **3.** Click **Test Connectivity**, and after the connectivity test passes, click **Finish** to save the configuration.

Note:

If the test connectivity fails, check your AK and the region in which the item is located. It is recommended to create the project in East China 2, and other regions do not guarantee network access.

- Add RDS Data Source
 - 1. Select the Data Integration > Data SourcePage, and click Add Data Source.
 - 2. Select the data source type as MySQL, and fill in the configuration information.

* Data Source Type :	ApparaDB for RDS V	
* Data Source Name :	rds_workshop_log	
Description :	rds log synchronization	
* RDS Instance ID :	rm-bp1z69dodhh85z9qa	0
* Primary Account of :	1156529087455811	0
RDS Instance		
* Database Name :	tungha, myy	
* Username :	heads	
* Password :		
Test Connectivity:	Test Connectivity	
0	The connectivity test can be passed only after the data source is added to the	
	RDS whitelist. Click here to see how to add a data source to the whitelist.	
	Ensure that the database is available.	
	Ensure that the firewall allows the data sent from or to the database to pass by.	
	Designe	Finis

Parameters:

- Data source type: ApsaraDB for RDS
- Data source name: rds_workshop_log
- Data source description: RDS log data synchronization
- RDS instance name: rm-bp1z69dodhh85z9qa
- RDS instance buyer ID: 1156529087455811
- Database name: workshop
- Username/Password: workshop/workshop#2017
- **3.** Click **Test Connectivity**, and after the connectivity test passes, click **Finish** to save the configuration.

Create a Business Flow

- 1. Right-click Business Flow under Data Development, select Create Business Flow.
- 2. Fill in the Business Flow name and description.

Create Business Flo	w	×
Business Name : Description :	workshop finish the DataWorks tutorial	
	Create	Cencel

3. Click Create to complete the creation of the Business Flow.



4. Enter the Business Flow Development Panel and drag a virtual node and two data sync nodes (oss_datasync and rds_datasync) into the Panel.

Create Node		×		
Node Type : Virtual Nod	le ~ start			
Destination Folder : Business F	low/workshop ~	Cancel		
Create Node		×		
Node Type : Data Sync				
Node Name : rds_datasy	nc			
Destination Folder : Business F	low/workshop ~			
	Submit	Cancel		
Create Node				×
Node Type :	Data Sync		~	
Node Name :	oss_datasync			
Destination Folder :	Business Flow/workshop		~	
			Submit	Cancel

5. Drag the connection to set the workshop_start node to the upstream of both data synchronization nodes.

D	Data Sync
	Data Development
S	ODPS SQL
Ľ	
P	/ PyODPS
s	Shell
5	SQL Component
	Noue
	Control
	Cross-tenant node
	Inspection

Configure workshop_start task

Since the new version sets the input and output nodes for each node, you need to set an input for the workshop_start node, the virtual node in the Business Flow can be set to the upstream node as the project root node, the project root node is generally named project name _ root.

You can configure it by clicking **Schedule**. When the task configuration is complete, click **Save**.

D rds_d	atasync ×	🕅 workshop_start 🔵						
	f) [8	⊕ C :						
1		X Depend on Last	Interval :					
		Resources ⑦						
		Resource Group : default_resource						
		Dependencies ⑦						
		Auto Parse : 💿 Yes 🔿 No	Parse I/O					
		Upstream Node Enter an outp		H Use the proje	ct Root Node			
		Upstream Node Output Na me	Upstream Node Output Table N ame	Node Name	Upstream Node ID	Owner	Source	Actions
		DataWorks_DOC_root		detaworks_doc_r oot		dataworks_dem o2	Added Manua Ily	

Create Table

1. Right-click Table and choose Create Table.



2. Type in Table Name(ods_raw_log_d and ods_user_info_d) for oss logs and RDS respectively.

Create Table				×
Database Type :	• MaxCompute			
Table Name :	Enter a table name			
		Subr	nit	Cancel

3. Type in your Table Alias and choose Partitioned Table.

Basics					
Table A	lias :				
Level 1 To	opic: Select ~	Level 2 Topic :	Select ~	Create Topic C	
Descrip	tion :				
Physical Model					
Parti	tion : • Partitioned Table No Partitioned Table	on-Life Cycle :		Days : 0	
Table L	evel : Select ~	Table Category :	Select ~	Create Level	
Table 1	ype : 💽 Internal table 🔵 Extern	nal table			

 Type in the field and partition information, click Submit to Development Environment and Submit to Production Environment.

Table Structure								
Add Field Move Up								
Field English Name	Field Alias	Field Type	Length/Set	Description	Primary Key ⑦			
col		STRING ~						
Add Partition								
Field English Name	Field Type	Length	Description	Partition Date Format	Partition Date Granularity			
рţ	STRING ~							

You can also click **DDL Mode**, use the following SQL statements to create tables.

```
//Creates a target table for oss logs
```

```
CREATE TABLE IF NOT EXISTS ods_raw_log_d (
 Col_string
)
PARTITIONED BY (
 dt STRING
);
//Creates a target table for RDS
CREATE TABLE IF NOT EXISTS ods_user_info_d (
 uid STRING COMMENT 'User ID',
  gender STRING COMMENT 'Gender',
  age_range STRING COMMENT 'Age range',
  zodiac STRING COMMENT 'Zodiac'
)
PARTITIONED BY (
  dt STRING
);
```

5. Click Submit to Development Environment and Submit to Production Environment. You

🗰 ods_raw_log_d × 🗸 workshop × 🖽 tftp	x Sq ten_precent_movies x DI OSS_ra	atings x 🛃 Movies_ODS x 🗸 DataWorks_Test x
DDL Mode Load from Development Environm	ent Submit to Development Environment	Load from Production Environment Submit to Production Environment
Table Name	ods_raw_log_d	
Business Process	workshop	
Basics		
Table Alias : ods_raw		
Level 1 Topic : workshop_table	✓ Level 2 Topic : tat	ple1
Description :		
Physical Model		
Partition : • Partitioned Tab Partitioned Table	le 🔿 Non-Life Cycle : 🗌	
Table Level : Select	✓ Table Category : Se	lect
Table type : (*) Internal table (External table	

can configure both of the tables in this way.

Configure the data synchronization task

- Configure the oss_datasync node
 - 1. Double-click the oss_datasync node node to go to the node configuration page.
 - 2. Select a data source.

Select the data source as the maid in the oss data source.

* Data Source :	OSS v oss_workshop_log	~	?
* Object Prefix :	user_log.txt		
	Add +		
* File Type :	text	~	
* Column Separator :	I		
Encoding :	UTF-8		
Null String :	Enter the sting that represents null		
* Compression :	None	~	
Format			
* Include Header:	Νο	~	

Parameters:

- Data source: oss_workshop_log
- Object Prefix: /user_log.txt
- Column Separator: |
- **3.** Select data destination

Select the data destination is ods_raw_log_d in the odps_first data source. Both partition information and cleanup rules take the system default, the default configuration of the partition is \${bizdate}.

1	Destination	Hide
eated by you. Click here t	o check the supported data source types.	
* Data Source :	ODPS ~ odps_first ~ ?	
* Table :	This must be specified. ods_raw_log_d	
	Generate Destination Table	
* Partition :	dt = \${bizdate}	
Clearance Rule :	Clear Existing Data Before Writing (Insert Overwrite) 🗸 🗸	
Compression :	💿 Disable 🔵 Enable	
Consider Empty . String as Null	● Yes ○ No	

4. Configure the field mapping, connect the fields that you want to synchronize.

Sq creet_table_ddl x	El ftp_datasync	write_result	×				
••••	• J •	6 0					
02 Mapping		Source Table			Desti	nation Table	
	Location/Value	Туре	Ċ	0		Field	Туре
	Column 0	string		•	•	col	STRING
	Column 1	string					
	Column 2	string					
	Column 3	string					
	Column 4	string					

5. Configure Transmission Rate with a maximum operating rate of 10 Mb/s.

03 Channel		
You can control the data s	ynchronization process through the transmission rate and the number of allo	owed dirty data records. See data synchronization documents.
* DMU :	1 ~	0
* Number of Concurrent Jobs :	1 ~ ⑦	
* Transmission Rate :	O Unlimited 💿 Limited 10 MB/s	
If there are more than :	Maximum n@ber of dirty data records. Dirty data is allowed by default. task ends.	dirty data records, the
Task's Resource Group :	Default resource group V	

- 6. Verify that the current task is configured and can be modified. After the confirmation is correct, click **Save** in the upper left corner.
- 7. Closes the current task and returns to the Business Flow configuration panel.
- Configure the rds_datasync node Node
 - 1. Double-click the rds_datasync node node to go to the node configuration page.
 - 2. Select a data source.

Select the data source that is located in the MySQL data source rds_workshop_log, and the table is named as ods_user_info_d, the split key uses the default to generate columns.

Di rds_datasync 🌒 [Eq creat_table_ddl ×	write_result ×		
B 💿 🖻	• 1 E	÷ 🛛		
01 Data Source		Source		De
	The data so	urces can be default data sour	ces or data sources o	reated by you. Click here to
* Deta Source :	MySQL	✓ rds_workshop_log	~ ?	* Deta Source :
* Table :	'ods_user_info_d' ×			
		Add Data Sour		
Data Filtering :	Enter SQL WHERE a incremental data sy the keyword "WHER	tatements, which are used for nchronization. Do not include E."	0	
Sharding Key:	uid		0	
		Preview		

3. Select data destination

Select the data destination ods_user_info_d in the data source named odps_first. Both partition information and cleanup rules take the system default, the default configuration of the partition is \${bizdate}.

Image: Source Image: Source Destination Hide Image: Source MySQL Image: Source of data sources created by you. Click here to check the supported data source types. Image: Source of data source types. Image: Image	🕅 ede datasuna 🍙 🚺	and anot table del at . 🕅 units anoth . M			
Image: Source Source Source Destination Hide The data sources can be default data sources or data sources created by you. Click here to check the supported data source types. Image: MySOL I I I I I I I I I I I I I I I I I I I	En res_eatesync •	a creat_table_ool X B write_result X			
Image: Data Source Source Definition Definition Hide Deters Source: MySQL Inde_workshop_Jog Bets Source: ODPS Inde_workshop_Jog Indeworkshop_Jog Indeworkshop_Jog Indeworkshop_Jog Indeworkshop_Jog Inde	•••	1 1 🗉 🗇 🖾			
The data sources can be default data sources or data sources created by you. Click here to check the supported data source types. • Data Source: • Data Filtering: • Enter SQL WHERE: statements, which are used for incremental data synchronization. Do not include the keyword WHERE! • Partinion: • the keyword WHERE! • Sharding Key: • ud • Compression: • Data Source:	01 Data Source	Source		Destination	Hide
Posta Source: MySQL v dis_workshop_log v ? Posta Source: ODPS v ddps_first v ? Table: ods_user_info_d × v * Table: ods_user_info_d v * Partition: dt = \$\begin{subarray}{c} & *		The data sources can be default data sources	or data sources created by you. Click her	re to check the supported data source types.	
* Table: iods_user_info_d × v * Table: ods_user_info_d v Add Data Source + Cenerate Destination Table Data Filtering: Enser SOL WHERE statements, which are used for incremental data synchronization. Do not include the keyword "WHERE" @ Partition: dt = \$bizdene) @ Partition: dt = \$bizdene) @ Clearance Rule: Clearence Rule: Clearence Rule: Clearence Rule: Clearence Rule: Compression: © Disable © Enable Preview Consider Empty: © Yes © No	* Data Source :	MySQL v rds_workshop_log v	Oeta Source :	ODPS V odps_first V	0
Add Data Source + Generate Destination Table Data Filtering: Enter SQL WHERE statements, which are used for incremental data synchronization. Do not include the keyword WHERE.* ? * Partition: dt = \$\bickloadsec ? Sharding Key: uid ? Clearance Rule: Clear Existing Data Before Writing (Insert Overwrit \vicession - Compression: © Diaable ① Enable Preview ? Compression: ?	* Table :	`ods_user_info_d' × ∽	* Table :	ods_user_info_d v	
Data Filtering: Enser SQL WHERE statements, which are used for incremental data synchronization. Do not include the keyword "WHERE."		Add Data Source		Generate Destination Table	
Clearance Rule : Clear Existing Data Before Writing (Insert Overwrit ~ Sharding Key : uid ⑦ Compression : ③ Diaable 〇 Enable Preview Consider Empty . String as Null : ④ Yes 〇 No	Data Filtering :	Enter SQL WHERE statements, which are used for incremental data synchronization. Do not include the keyword "WHERE."	⑦ Partition:	dt = \$(bizdete)	
Sherding Key: uid Compression: Disable Enable Preview Consider Empty. Yes No String as Null:			Clearance Rule :	Clear Existing Data Before Writing (Insert Overwrit_ \vee	
Preview Consider Empty. O Yes No String as Null	Sharding Key :	uid	Compression :	💿 Disable 🔵 Enable	
		Preview	Consider Empty String as Null	😑 Yes 🔿 No	

4. Configure the field mapping, default in association with the name mapping.

Di rda	_datasyn	c •	50 en	et_table	_ddl ×		write_resu	t ×						
۳	\odot	Þ	ſ.	b			ø							
									St	ning as Null '	-)		
02	Mappin	9				Sou	rce Table			Des	tinatio	n Table		
			Fie	ld			Туре	C			Fiel	ы	Туре	Map of the same name
			uid				VARCHAR	•			uid 🌔		STRING	Enable Same-Line Mapping
			ge	nder			VARCHAR	•			gen	ıder	STRING	Cancel mapping
			89	e_range			VARCHAR	•			age	_range	STRING	
			zo	fiac			VARCHAR	۰			zod	fiec	STRING	
			Ad	d +										

5. Configure Transmission Rate with a maximum operating rate of 10 Mb/s.

03 Channel		
You can control the data a	ynchronization process through the transmission rate and the number of alk	owed dirty data records. See data synchronization documents.
* DMU :	1 *	0
* Number of Concurrent Jobs :	1 ~ ⑦	
* Transmission Rate :	O Unlimited O Limited 10 MB/s	
If there are more than :	Maximum r@ber of dirty data records. Dirty data is allowed by default. task ends.	dirty data records, the
Tesk's Resource Group :	Default resource group \lor	

6. Verify that the current task is configured and can be modified. After the confirmation is correct, click **Save** in the upper left corner.

Di rds_datasync 🔵 [a creat_table_ddl 🗙 🖸	write_result ×				
	1 I I I I I	j @				
01 Data Source		Source		Destination		Hide
* Data Source :	The data source	es can be default data sources o rds_workshop_log	or data sources created by you. Click her	e to check the supported dat ODPS ~	a source types. odps_first v	0
* Table :	`ods_user_info_d` ×		* Table :	ods_user_info_d		
		Add Data Source +	÷.		Generate Destination Table	

7. Closes the current task and returns to the Business Flow configuration panel.

Submit Business Flow tasks

- 1. Click **Submit** to submit the current Business Flow.
- Select the nodes in the submit dialog box, and check the Ignore Warning on I/O Inconsistency, click Submit.

Run workflow task

1. Click Run.

✓ Data Integration	Development	Blood	
Di Data Sync			
 Data Development 			
Sq ODPS SQL			
Sh Shell			
Mr ODPS MR			
Vi Virtual Node			Vi workshop_start
Py PyODPS			

During a task run, you can view the run status.

- 2. Right-click the SQL task and select view log.
- 3. Right-click the OSS _ Data Synchronization task and select view log.
- 4. Right-click the RDS _ Data Synchronization task and select view log.

Check if the data is successfully imported into MaxCompute

1. Click temporary query in the left-hand navigation bar.

2. Select New > ODPS SQL.



3. Write and execute SQL statement to check the entries imported into ods_raw_log_d.

Sq select_01 • 🚑 workshop X	
<pre>1odps sql 2***********************************</pre>	
Runtime Log Result[1] ×	
A 1 _c0 ~ 2 0	

4. Also write and execute SQL statements to view the number of imported ods_user_info_d records.

Note:

The SQL statement is as follows, where the partition columns need to be updated to the

business date, if the task runs on a date of 20180717, the business date is 20180716.

```
Check that data was written to MaxCompute
select count(*) from ods_raw_log_d where dt=business date;
select count(*) from ods_user_info_d where dt=business date;
```

Next step

Now that you've learned how to synchronize the log data, complete the data acquisition, you can continue with the next tutorial. In this tutorial, you will learn how to calculate and analyze the collected data. For more information, see *Data processing: user portraits*.

2.3 Data processing: user portraits

This article shows you how to process log data that has been collected into MaxCompute through dataworks.

Note:

Before you begin this experiment, please complete the operation in *Data acquisition: log data upload*.

Create data tables

You can refer to Data acquisition: log data upload to create data tables.

- Create ods_log_info_d table
 - Right click **Table** in the workshop business flow. Click **Create Table** and enter the table's name ods_log_info_d. You can then click DDL Mode to type in the table creation SQL statements.

The following are table creation statements:

```
CREATE TABLE IF NOT EXISTS ods_log_info_d (
 IP string comment 'IP address ',
 uid STRING COMMENT 'User ID',
 Time string comment 'time': MI: ss ',
 Status string comment 'server return status code ',
 Bytes string comment 'the number of bytes returned to the Client
 ۰,
 Region string comment ', get' from IP ',
 Method string comment 'HTTP request type ',
 URL string comment 'urle ',
 Protocol string comment 'HTTP Protocol version number ',
 Referer string comment 'source ures ',
 Device string comment 'terminal type ',
  Identity string comment 'Access type crawler feed user unknown'
)
PARTITIONED BY (
 dt STRING
);
```

- 2. Click Submit to Development Environment and Submit to Production Environment.
- Create dw_user_info_all_d table

The method of creating a new report table is identical to that of a table statement as follows:

```
-- Create a copy table

CREATE TABLE IF NOT EXISTS dw_user_info_all_d (

uid STRING COMMENT 'User ID',

gender STRING COMMENT 'Gender',

age_range STRING COMMENT 'Age range',

zodiac STRING COMMENT 'Zodiac sign'

Region string comment ', get' from IP ',

Device string comment 'terminal type ',
```

```
Identity string comment 'Access type crawler feed user unknown ',
Method string comment 'HTTP request type ',
URL string comment 'url ',
Referer string comment 'source url ',
Time string comment 'time': MI: ss'
)
PARTITIONED BY (
DT string
);
```

Create rpt_user_info_d table

The following are table creation statements:

```
-- Create a copy table
Create Table if not exists rpt_user_info_d(
   uid STRING COMMENT 'User ID',
   Region string comment ', get' from IP ',
   Device string comment 'terminal type ',
   PV bigint comment 'cv ',
   gender STRING COMMENT 'Gender',
   age_range STRING COMMENT 'Age range',
   zodiac STRING COMMENT 'Zodiac sign'
)
PARTITIONED BY (
   DT string
);
```

Business Flow Design

Open the Workshop Business Flow and drag three ODPS SQL nodes amed as

"ods_log_info_d、dw_user_info_all_d、rpt_user_info_d" into the canvas, n, and configure dependencies.



Creating user-defined functions

- 1. Download ip2region.jar.
- 2. Right-click **Resource**, and select **Create Resource** > jar.



3. Click Select File, select ip2region. jar that has been downloaded locally, and click OK.

Create Resource		×						
* Resource Name :	ip2region.jar							
Destination Folder :	Destination Folder : Business Flow/workshop/Resource ~							
Resource Type :								
	Upload to ODPS The resource will also be uploaded to ODPS.							
File :	ip2region.jar (4.62M)							
	ок	Cancel						

4. After the resource has been uploaded to dataworks, click Submit.



5. Right-click a function and select Create Function.



6. Enter the function name getregion, select the Business Flow to which you want to belong, and click **Submit**.

Create Function			×
Function Name :	getregion		
Destination Folder :			
		Submit	Cancel

 Enter the function configuration in the Registry Function dialog box, specify the class name, description, command format, and parameter description.

🕫 getregion 🗙 Ja ip2region.jar 🗙 📲	workshop ×
E 🗈 C	
Registry Function	
Function Name	
* Class Name	crg.alidata.odps.udf.lp2Region
* Resources	: ip2region.jar
Description	IP address convert to region
Command Format	: getregion("ip")
Parameters	ip address

Parameters:

- Function Name: getregion
- Class Name: org. alidata. ODPS. UDF. ip2region
- · Resource list: ip2region. Jar
- · Description: IP address translation area
- Command Format: getregion ('IP ')
- Parameter description: IP Address
- 8. Click Save and submit.

Configure ODPS SQL nodes

- Configure ods_log_info_d Node
 - Double-click the ods_log_info_d node to go to the node configuration page and write the processing logic.



The SQL logic is as follows:

```
INSERT OVERWRITE TABLE ods_log_info_d PARTITION (dt=${bdp.system.
bizdate})
SELECT ip
  , uid
   Time
  , Status
  , Bytes-use a custom UDF to get a locale over IP
  , Getregion (IP) as region -- the request difference is divided
into three fields through the regular
  , Regexp_substr (request, '(^ [^] +)') as Method
), Regexp_extract (request, '^ [^] + (. *) [^] + $ ') As URL
  ), FIG (request, '([^] + $ )') as protocol-get more precise URLs
 with regular clear refer
  , (Referer, '^ [^/] +: /([^/] +) {1 }') as Referer-Get terminal
information and access form through agent
  , Case
    When tolower (agent) rlike 'android 'then 'android'
    WHEN TOLOWER(agent) RLIKE 'iphone' THEN 'iphone'
    WHEN TOLOWER (agent) RLIKE 'ipad' THEN 'ipad'
    WHEN TOLOWER(agent) RLIKE 'macintosh' THEN 'macintosh'
    WHEN TOLOWER(agent) RLIKE 'windows phone' THEN 'windows_phone'
    WHEN TOLOWER(agent) RLIKE 'windows' THEN 'windows_pc'
    ELSE 'unknown'
  End as Device
  , Case
    WHEN TOLOWER(agent) RLIKE '(bot|spider|crawler|slurp)' THEN '
crawler'
    WHEN TOLOWER(agent) RLIKE 'feed'
    OR regexp_extract(request, '^[^ ]+ (. *) [^ ]+$') RLIKE 'feed
' THEN 'feed'
    WHEN TOLOWER(agent) NOT RLIKE '(bot|spider|crawler|feed|slurp
) '
    AND agent RLIKE '^[Mozilla Opera]'
```

```
AND regexp_extract(request, '^[^ ]+ (. *) [^ ]+$') NOT RLIKE '
feed' THEN 'user'
ELSE 'unknown'
END AS identity
FROM (
    SELECT SPLIT(col, '##@@')[0] AS ip
   , SPLIT(col, '##@@')[1] AS uid
   , SPLIT(col, '##@@')[2] AS time
   , SPLIT(col, '##@@')[3] AS request
   , SPLIT(col, '##@@')[4] AS status
   , SPLIT(col, '##@@')[5] AS bytes
   , Split (cola, '# @') [6] As Referer
   , SPLIT(col, '##@@')[7] AS agent
FROM ods_raw_log_d
Where dt = $ {BDP. system. Date}
) a;
```

2. Click Save.



- Configure dw_user_info_all_d Node
 - Double-click the dw_user_info_all_d node to go to the node configuration page and write the processing logic.

The SQL logic is as follows:

```
INSERT OVERWRITE TABLE dw_user_info_all_d PARTITION (dt='${bdp.
system.bizdate}')
SELECT COALESCE(a.uid, b.uid) AS uid
, b.gender
, b.age_range
, B. flavdiac
```

```
, a.region
  , a.device
  , a.identity
  , a.method
  , a.url
  , a.referer
  , a.time
FROM (
 SELECT *
 From fig
 WHERE dt = ${bdp.system.bizdate}
) a
LEFT OUTER JOIN (
  SELECT *
  FROM ods_user_info_d
 WHERE dt = ${bdp.system.bizdate}
) b
ON a.uid = b.uid; WHERE dt = ${bdp.system.bizdate}
) a;
```

- 2. Click Save.
- Configure a rpt_user_info_d Node
 - Double-click the fig node to go to the node configuration page and write the processing logic.

The SQL logic is as follows:

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

2. Click Save.

Submitting Business Flows

- 1. Click Submit to submit the node tasks that have been configured in the Business Flow.
- Select the nodes that need to be submitted in the Submitdialog box, and check the Ignore Warnings on I/O Inconsistency, click Submit.

Submit					×
	Node	 <			
	Note	worksho	dw_user_info_all_d op user portrait part is written logical Warnings on I/O Inconsistency	lly.	
				Submit	Cancel

Running Business Flows

1. Click **Run** to verify the code logic.



- 2. Click Queries in the left-hand navigation bar.
- 3. Select New > ODPS SQL.

Data	DataStudio DataW	/orks_DOC					Cr	oss-project cloning	Operation Center	٩	wangdan	English
Ш	Queries 🔉 🔒 🖪	2 C 🕀	Sa select_01	🕨 🚠 workshop								
	Enter a file or creator (2)	Folder	m e		С	22						
*	✓ Queries	Create >	ODPS SQL	3								
民	1 1 test birdbd(002 08-29		Shell ODPS	QL angdan								
ü	50 select_01 Mel002 08-3			te time:2018-0								

4. Write and execute SQL statements, Query Task for results, and confirm data output.

Sa sele	ct_01	• [Sa rpt_u	ser_info	x b_d	Sa dw_u	ser_info.	x b_lla	Sq 🗙	ds_log_info	_d ×	🖪 getregi	on	× 📠 ip2regio	n.jør	× 🛔	workshop	×	
•		۲	Þ		С	88													
<pre>1odps sql 2***********************************</pre>																			
9 10			from r	pt_us	er_in	fo_d whe	re dt-	201807	22 lim	it 10;									
																			不
																			53 52
运行	日志		结果	[1]	×	结果[2]	×												
1	d	A	✓ reg	jion	8	✓ device	с ;	× 1	pv	D	• gen	E der	~	F age_range	¥ 2	odiac	G	✓ dt	н

The query statement is as follows:

```
--- View the data in the data box
Select * From glaswhere DT ''business day'' limit 10;
```

Publishing Business Flow

After the Business Flow is submitted, it indicates that the task has entered the development environment, but the task of developing an environment does not automatically schedule, so the tasks completed by the configuration need to be published to the production environment (before publishing to the production environment, test this task code).

1. Click **Publish** To Go To The publish page.



- 2. Select the task to publish and click Add To Be-Published List.
- 3. Enter the list of pending releases, and click Pack and publish all.
- 4. View published content on the Publish Package List page.

Run tasks in production

- 1. After the task has been published successfully, click **Operation center**.
- 2. Select Workshop Business Flows in the Task List.

Operation Cent	ter Dat	taWorks_DOC2 💎 🗸 🗸			DataStudio	dataworks_demo
≡ (§ 0&M Overview	Search:	Node Name/Node ID Q	Solutions: Plea	ect Business Flow: Please select Node Type: Please select O	wner:: Select an owner	~
🚽 Task List	Baseline	Please select V	Nodes	ified Today Paused (Frozen) Node Reset Clear		
Cycle Task						C Refresh
🚯 Manual Task		Name:	Node ID	Testing Envrionment. Please be	cautious.	
▶ Task O&M		workshopstart	7000005641			
▶ Alerm		rds_数据同步	7000005641			
		ftp_数据同步	7000005641			
		ods_log_info_d	7000005641			
		dw_user_info_all_d	7000005641	dataworks_doc2_root Virtual Node		
		rpt_user_info_d	7000005641			
		dataworks_doc2_root	7000005641	workshopstart		
				Virtual Node		
				ftp.数据同步 rds_数; Data Integration Data Inte	据同步 sgration	
					Node ID:	
					Node Name:	
					Schedule Type::	Day Schedule
					Owner:: Description:	
	More	▼ < 1/1 >				

3. Right-click the workshop_start node in the DAG graph and select **Patch Data > Current and downstream nodes**.



4. Check the task that needs to fill the data, enter the business date, and click OK.

Retroactive Insertion		×
* Retroactive Insertion Name:	P_workshop_start_20180831_105048	
* Select Business Date:	2018-08-30 - 2018-08-30	
* Allow Parallel:	Not Parallel 🗸	
* Select the node for retro	active insertion.:	
Task Name	Search by name Q	Task Type 🍸
DataWorks_D	OC(79023)	
workshop_sta	irt	Virtual Node
create_table_c	ddl	ODPS_SQL
ftp_sync		Data Integratio n
rds_sync		Data Integratio n
ods_log_info_	d	ODPS_SQL
dw_user_info	_all_d	ODPS_SQL
rpt_user_info_	b,	ODPS_SQL
		OK Cancel

When you click **OK**, you automatically jump to the patch data task instance page.

5. Click Refresh until the SQL task runs successfully.

() OSM Overview	Search: 700000461343 Q	Retroactive Insertion N	leme: Please select	Y Node Type: Please a	oloct Y Owner.	Select an owner	Y
Task List	Run Date: 2018-08-31	Business Date: Selec	t date 🔲	Baseline: Please select	✓ My Nodes	Reset Clear	
Cycle Task							C Refresh Hide Search
(2) Manual Task	Instance Name	Status	Task Type	Owner	Timer	Business Date	Actions
 Tesk OSM 	 P_workshop_start_20180831_105048 	Running					✓ Batch Terminate
R Cycle Instance	✓ 2018-08-30	Running				2018-08-30	×
Manual Instance	workshop_start	⊗ Ran	Virtual Node	wangdan	2018-08-31 00:05:00	2018-08-30	DAG I Terminete I Rerun I More 🔻
Testing Instance							
PetchDeta							
» Alerm							

Next step

Now that you 've learned how to create SQL tasks, how to handle raw log data, you can continue with the next tutorial. In this tutorial, you will learn how to set up data quality monitoring for tasks completed by your development, ensures the quality of tasks running. For more information, see*Data quality monitoring*.

2.4 Data quality monitoring

The paper mainly discusses how to monitor the data quality in the process of using the data workshop, set up quality monitoring rules, monitor alerts and tables.

Prerequisites

Please complete the experiment*Data acquisition: log data upload* and *Data processing: user portraits*before proceeding with this experiment.

Data quality

Data quality (DQC), is a one-stop platform that supports quality verification, notification, and management services for a wide range of heterogeneous data sources. Currently, Data Quality supports monitoring of MaxCompute data tables and DataHub real-time data streams. When the offline MaxCompute data changes, the Data Quality verifies the data, and blocks the production links to avoid spread of data pollution. Furthermore, Data Quality provides verification of historical results. Thus, you can analyze and quantify data quality. In the streaming data scenario, Data Quality can monitor the disconnections based on the DataHub data tunnel. Data Quality also provides orange and red alarm levels, and supports alarm frequency settings to minimize redundant alarms.

The process of using data quality is to configure monitoring rules for existing tables. After you configure a rule, you can run a trial to verify the rule. When the trial is successful, you can associate this rule with the scheduling task. Once the association is successful, every time the scheduling task code is run, the data quality validation rules are triggered to improve task accuracy. Once the subscription is successful, the data quality of this table will be notified by mail or alarm whenever there is a problem.

Note:

The data quality will result in additional costs.

Add Table Rule Configuration

If you have completed the log data upload and user portrait experiments, you will have the following table: ods_raw_log_d, ods_user_info_d, ods_log_info_d, dw_user_info_all_d, rpt_user_i nfo_d.

The most important thing in data quality is the configuration of table rules, so how to configure table rules is reasonable? Let's take a look at how the tables above be configured with table rules.

ods_raw_log_d

You can see all the table information under the item in the *data quality*, now you are going to configure the data quality monitoring rules for the ods_raw_log_d data sheet.

=		List of tables				
 DQC Monitoring 	ODPS data source 🗸 🗸					
88 Overview	Q	oda_raw_log_d 🔍	Responsible : All res	ponsible persons 🗸 🗸		
Hy Subscription	dataworks_doc	Table Name	deta source	Application name	Responsible	operating
💼 Rule Configuration		ods_rew_log_d	ODPS	dataworks_doc	dataworks_demo2	Configure monitoring rules

Select the ods_raw_log_d table and click **Rule Configuration** to go to the following page.

DQC Monitoring Overview	Rule configuration Application name : datawork	Rule configuration Application name : dataworks_doc > Table Name : ods_raw_log_d > Partition expression :										
B My Subscription	+ Added partition expression	Template rules (0)	Self-help rules (0)	Strong trand	Oranna thrashold	Red thrashold	Comparison	Evacted value	Configurator	onersting		
E Rule Configuration	Currently there is no operational partition	rieid name	remplate name	strong trend	Grange threshold	neu threshold	method	Expected value	Configurator	operating		
Mission Inquiries	expression. Please first add the partition expression					No data						

You can review the data sources for this ods_raw_log_d table. The data for ods_raw_log_d table is from ftp. Its partition is \${bdp.system.bizdate} format and is written into the table ("dbp. system.bizdate" is the date to get to the day before).

For this type of daily log data, you can configure the partition expression for the table. There are several kinds of partition expressions, and you can select dt = \$ [yyyymmdd-1]. Refer to the documentation *Parameter configuration* for detailed interpretation of scheduling expressions.

	Template rules (0)	Self-help rules (0	0)					
Ided partition expression	Add a partition	n			×	20	Expected value	Configu
mently there is no erational partition pression. Please first add the rition expression	Parition	expression 1	Please enter the partition expression Calculation	Ok Can	cel			

Note:

If there is no partition columns in the table, you can configure it as no partition. Depending on the real partition value, you can configure the corresponding partition expression.

After confirm, you can see the interface below and choose to create rules.

E DOC Monitoring	Rule configuration Application name : datawork	Rule configuration Application name : dataworks_doc > Table Name : ods_raw_log_d > Partition expression : dt=S(yyyymmdd=1]											
My Subscription	+	Responsible : wangde	an .			Trial run Subscription Management	Create rules						
Rule Configuration	Added partition expression	Template rules (0)	Self-help rules (0)										
Mission Inquiries	- dt+S[yyyymmdd-1]	Field name	Template name	Strong trend	Orange threshold Red threshold	Comparison Expected value Configurator method	operating						
					No deta								

When you select to create a rule, the following interface appears.

= DOC Monitoring	Rule configuration			Template rules Self-help rules						
BE Overview	Application name : datawork	ks_doc > Table Name	: ods_raw_log_d > Pa	+ Add monitoring rules + Quick add						
Hy Subscription	+ Added partition expression	Template rules (0)	Self-help rules (0)							
Rule Configuration	- dt+\$[yyyymmdd-1]	Field name	Template name	Field name : Plesse select.						

Click Add monitoring rule and a prompt window appears for you to configure the rule.

. DOC Monitoring	Rule configuration			Template rules Self-hel	p rules					
() Overview	+	Responsible : wanpti	: odt_rax_log_d > Pa	+ A0	s monitoring rules		+ Quick add			
Hy Subscription	Added partition expression	Template rules (0)	Setting rules (0)							
[]] Note Configuration	- dt=0[cccommit#1]	Faldname	Template name	Field Type :	Table level rules 🛛 🗸	Strong and weak :	🔿 Strong 📀	week		
Masion Inquiries				Template type :	SQL taok table rows, 7 de.	trend :	Absolute value	~		
				Comparison of	BT load line number, U.C. 20 days 1	ive number, 1,7,30 days fluctuation test.				
				volatility (ST denials of rows repectations of					
				Crange threshold :	ET desirate of the number of rows, desirate of the number of news of 2 desirate of CUTMODE model do RM DT CUTMODE model expectations	12,38 days fluctuation test 07/dt and the head office of 78 expectations checksown closelmeres	te calibration of expe	cariors		
				Field name i Rule type i	ET load line collection of expected SQ, task table rows, 7 days even SQ, task table rows, 30 days even	privaletility detection age volatility detection		-		

The data in this table comes from the log file that is uploaded by FTP as the source table. You need to determine whether there is data in this table partition as soon as possible. If there is no data in this table, you need to stop the subsequent tasks from running as if the source table does not have data, the subsequent task runs without meaning.

Note:

Only under strong rules does the red alarm cause the task to block, setting the instance state to failure.

When configuring rules, you need to select the template type as the number of table rows, sets the strength of the rule to strong. Click the **Save** button after the settings are completed.

Note:

This configuration is primarily to avoid the situation that there is no data in the partition, which causes the data source for the downstream task to be empty.

Rules test

In the upper-right corner, there is a node test button that can be used to verify configured rules . The test button can immediately trigger the validation rules for data quality.

Rule configuration Application name : datawork	Rule configuration Application name : dataworks_doc > Table Name : ods_raw_log_d > Partition expression : dt=S[yyyymmdd-1]											
+	Responsible : wangda	n					Trial	un Subscr	iption Management	Create rules More *		
Added partition expression	Template rules (1)	Self-help rules (0)										
+ dt=S[yyyymmdd-1]	Field name	Template name	Strong	trend	Orange threshold	Red threshold	Comparison method	Expected value	Configurator	operating		
	table_count	SQL task table rows, 1,7, 30 days fluctuation text	Strong	Absolute value	10%	50%	-	-	wangdan	modify delete Log 9		

When you click the test button, you are prompted for a window to confirm the test date. After a run is clicked, there will be a prompt information below telling you to jump to the test results by clicking prompt information.

•	Rule configuration												
 DQC Monitoring 	Lastration name / datas	orba dar a Table N	ame i oda mar	In d a Partic	a more stars to	a theorem	a.11						
2 Overview	Approximation marrie . Calcon	and a contraction of the second	arre : ousjaw,	Jog.o - Fartos	in expression 1.		0.11						
III Mathematica	+	Responsible : w	engden						Trial run	Subscription Manag	ement	Create rules	More *
a an and the	Added partition expression	Trial run											
[]] Rule Configuration	· dt=\${									endudus Conference			
Mission Inquiries										contrast company			
			Test run district :						-	wangdan		modify delete	Log
			and the second se	2018-06-31 1	20933				_				
			Called At 1	2010/00/31		<u> </u>							
				Trial run									
			I	The runs succes	intury Crock to your	Sect run results							- 4
							-						
								shut down					
=	Example details												
 DQC Monitoring 	example details		d - deflere		8 2018 08 21 12.1	2.27 Marca						Refr	esh
88 Overview	approximent paraworks_coc 1	sole mente coscrevo	ofte > analikkiu	mpp-1 in	8 2010/00/01 121	LLI MORE							
Hy Subscription	Fields descriptio	on Statistical function	Strong/weak	Comparison method	Expected value	Onange threshold	Red threshold	Conditions 201	results	Sampling result	status	operating	
Rule Configuration		table count	Strong	-	-	10%	50%	-	0,0%	570386	normal	See historical	
Mission Inquiries									0,0%			results	
~													
CT													
	9:												

According to the test results, the data of the Mission output can be confirmed to be in line with the expectations. It is recommended that once each table rule is configured, a trial operation should be carried out to verify the applicability of the table rules.

When the rules are configured and the trial runs are successful, you need to associate the table with its output task. In this way, every time the output task of the table is run, the validation of the data quality rules is triggered to ensure the accuracy of the data.

Associated Scheduling

Data quality support being associated with scheduling tasks. After the table rules and scheduling tasks are bound, when the task instance is run, the data quality check is triggered. There are two ways to schedule table ruless:

- Perform table rule associations in operations center tasks.
- Association in the regular configuration interface for data quality.

Operations Center Association Table rules

In care center, in cycle tasks, locate the **ftp_datasync** task, and in **more**, select **configure data monitoring**.

Enter the monitored table name in the burst window, as well as the partition expression. The table entered here is named as ods_user_info_d and the partition expression is dt = \$ [yyyymmdd-1].

Click **Configure** to quickly go to the rule configuration interface.

=	Rule configuration	s. doc. > Table Name	: ods.raw.log.d ≽ Parti	tion expr	ression : dt-	Slwwmmdd-1]					
B Overview	+	Responsible : wangdan Trial run Subscription Management Create roles More *									
Rule Configuration	Added partition expression	Template rules (1)	Self-help rules (0)								
Mission Inquiries	 dt=\$[yyyymmdd-1] 	Field name	Template name	Strong	trend	Orange threshold	Red threshold	Comparison method	Expected value	Configurator	operating
		table_count	SQL task table rows, 1,7, 30 days fluctuation text	Strong	Absolute value	10%	50%	-	-	wangden	modify delete Log s

Configure task subscriptions

After the associated scheduling, every time the scheduling task is run, the data quality verificati on is triggered. Data quality supports setting up rule subscriptions, and you can set up subscriptions for important tables and their rules, set up your subscription to alert you based on the results of the data quality check. If the data quality check results are abnormal, notificati ons are made based on the configured alarm policy. Click **Subscription Management** to set up subscription methods. Mail notifications, email and SMS notifications are currently supported.

DOC Monitoring Overview My Subscription Num Configuration Massion Inquiries	Rule configuration Application name : dataword + Added partition expression + dt=3[yyymmdd=1]	ks, doc ≻ Table Name : Responsible : wangdan Templaterules (1) Field name	ods_raw_Jog_d > Partitio	n expression : dt	-S(yyyymmdd-1) Orange threshold	d Red threshold G	Trial run iomperison enhod	Subscription M	anagement Cre	ne rules More *
		table_count	SQL task table rows, 1,7, 30 days fluctuation S test	trong Absolute value	10%	50% -		- wangd	an e	fily delete Log
Image: Configuration Image: Configuration	+ Added partition expres - dt=\$[yyyymmdd-1]	sion Template	subscription / nules me deteworks, wing bunt	Management demo2 The	Q ere will pr catio	Add subscribers resent all i	operating member delete	Comparimethod	Trial run Su ¹⁰⁰¹ Expected vo r oject.	bacription Manage viue Configurator wangdan
= _ DQC Monitoring	Rule confi Application	guration name : datawor	ks_doc > Table Na	ame : ods_ra	w_log_d > F	Partition expre	ession : d	lt=S[yyyymmdc	I-1]	
BB Overview	+		Responsible : v	Subscriptio	on Managem	ient		7		
 Rule Configuration Mission Inquiries 	• dt=S[yyyym	mdd-1]	Field name	Subscrib	"aemoz er Subso	cription method	× 1	vop subscribers	operating	Compo metho
			table_count	umpie) () E	-mail notificatio mail and SMS n	n otifications		delete	-

After the subscription management settings are set up, you can view and modify them in **My Subscription**.

DOC Monitoring	my subscription			Partition expression	Responsible	operating
88 Overview	COPS data source 🗸 🗸			dt=\$[yyyymmdd-1]	wangdan	Last check result method to informe * Subscribed
My Subscription	Search by tablename. Q Empty					
Rule Configuration	Table Name	data source	application			
Alission Inquiries	0_gol_war_sbo	odps	dataworks_doc			

It is recommended that you subscribe to all rules so that the verification results are not notified in a timely manner.

ods_user_info_d

The data in the ods_user_info_d table is from RDS database. When you configure rules, you need to configure the table to check the number of rows and the unique validation of the primary key to avoid duplication of data.

Similarly, you need to configure a monitoring rule for a partition field first, and the monitoring time expression is: dt = \$[yyyymmdd-1]. After successful configuration, you can see a successful partition configuration record in the partition expression that has been added.

	Rule configuration Application name : datawork	ks_doc > Table Name	: ods_user_info_d >	Partition expression :	dt=S[yyyymmdd-1]		
My Subscription	+ Added partition expression	Responsible : wangde	self-help rules (0)			Trial run Subscription Management	Create rules More *
Kule Configuration	• dt=S[yyyymmdd-1]	Field name	Template name	Strong trend	Orange threshold Red threshold	Comparison Expected value Configurator method	operating

After the partition expression is configured, click **Create Rule** on the right to configure the validation rules for data quality. Add monitoring rules for table rows, rule intensity is set to strong, comparison mode is set to expectations greater than 0.

Rule configuration			Template rules Self-he	Ip rules				
Application name : datawork	ks_doc > Table Name	: ods_user_info_d > F						
+	Responsible : wangde	in	+ Ad	id monitoring rules	+ Quick add			
Added partition expression	Template rules (0)	Self-help rules (0)						
- dt=\$[yyyymmdd-1]	Field name	Template name	Field Type : Template type :	Table level rules ∨ SQL task table rows, 1,7, ∨	Strong and weak : trend :	Strong w Absolute value	eak V	
			Comparison of volatility :	25%	75%	75%	100%	
			Orange threshold :	10 %	Red threshold :	50	5	

Add column-level rules and set primary key columns to monitor columns. The template type is: the number of repeated values in the field is verified, and the rule is set to weak, the comparison mode is set to a field where the number of duplicate values is less than 1. After the setting is completed, click the bulk **Save** button.

Rule configuration			Template rules Self-he	lp rules			
Application name : datawor	ks_doc > Table Name	: ods_user_info_d > F					
+	Responsible : wangde	10	+ Ad		+ Quick add		
Added partition expression	Template rules (0)	Self-help rules (0)					
- dt:\$[yyyymmdd-1]	Field name	Template name	Field Type :	Table level rules 🔍	Strong and weak (Strong O v	veak
	_		Template type :	SQL task table rows, 1,7, $ \lor$	trend :	Absolute value	\sim
			Comparison of	25%	75%	75%	100%
			volatility :		0		
			Orange threshold -	10 3	Rad threshold -	50	
			change internet i	M	NEW EINERALUNG I		

Note:

This configuration is primarily designed to avoid duplication of data which may result in contamination of downstream data.

ods_log_info_d

The data of this ods_log_info_d table mainly is the analysis of the data in the table. Because the data in the log cannot be configured for excessive monitoring, you only need to configure the validation rules that is not empty for the table data. The partition expression for the first configuration table is: dt = \$[yyyymmdd-1]

•	DQC Monitoring	Application pame - datawork	re doc o Table Name	: ode log info d > E	artition expression	dt_Chosemmdd_1		
88	Overview	Application name : datawork	table Name	, ous_loy_inio_d > P	annon expression .	ut-s(yyy)mmuu-1]		
	My Subscription	+	Responsible : wangda	n			Trial run Subscription Management	Create rules More *
	P. I. Carferrie	Added partition expression	Template rules (0)	Self-help rules (0)				
LE	Rule Configuration	- dt=S[vvvvmmdd-1]	Eield name	Templete neme	Stroog trand	Orange threshold. Red threshold	Comparison Expansed value Configurator	operation
Ø	Mission Inquiries		Piero name	remplate name	Strong trend	orange oneshold . Ned sineshold	method Expected value Comparator	operating

The configuration table data is not an empty calibration rule, and the rule strength should be set to strong. The comparison is set to an expected value of not equal to 0, and after the setup is complete, click the **Save** button.

E DOC Monitoring	Rule configuration	a des - Table Name a sta los info d - Da	Template rules Self-help rules	
B Overview	+	Responsible : wangdan	+ Add monitoring rules	+ Quick add
Rule Configuration	Added partition expression - dt=\${yyymmdd-1}	Template rules (0) Self-belp rules (0) Field name Template name	Field Type : Table level rules V	Strong and weak : Strong weak
T. museefact			Template type : SQL task table rows, 1,7, Comparison of 0%, 25%	75% 75% 100%
			Grange threshold : 10 %	Red threshold : 50 N

dw_user_info_all_d

This dw_user_info_all_d table is a summary of data for both the ods_user_info_d table and the ods_log_info_d table, because the process is relatively simple, the ODS layer is also configured with a rule that the number of table rows is not empty, so the table does not have the data quality monitoring rules configured to save on computing resources.

rpt_user_info_d

The rpt_user_info_d table is the result table after the data aggregation. Based on the data in this table, you can monitor the number of table rows for fluctuations, and verify the unique values for primary keys. Partition expression for the first configuration table: dt = \$[yyyymmdd-1

]

•	= DQC Monitoring Overview	Rule configuration Application name : datawork	s_doc > Table Name	: rpt_user_info_d > P	artition expression : d	sS[yyyymmdd-1]		
	Mc Subarristian	+	Responsible : wangde	in			Trial run Subscription Management	Create rules Mo
	Rule Configuration	Added partition expression	Template rules (0)	Self-help rules (0)				
đ	Mission Inquiries	- d1=S[yyyymmdd=1]	Field name	Template name	Strong trend	Orange threshold Red threshold	Comparison Expected value Configurator method	operating

Then you may configure the monitoring rules: Click **Create rule** on the right, and click **Add Monitoring Rules** to monitor columns. The number of repeated values in the field is verified, and the rule is set to weak. The comparison mode is set to field repeat values less than 1.

E DVC Maximiza	Rule configuration			Template rules Self-help rules
8 Overview	Application name : datawork	ts_doc > Table Name : rpt	t_user_info_d > Pa	
Hy Subscription	+	Responsible : wangdan		+ Add monitoring rules + Quick add
(ii) Rule Configuration	Added partition expression	Template rules (0) Sel	if-help rules (0)	Field Type : Teble level nules V Strong and weak : Strong weak
Mission Inquiries	- 01-03933 (minute) (1	Field name Ti	emplete name	Template type : SQL task table rows, 1,7, \lor trend : Mosolute value \lor
				Comparison of 0% 25% 75% 75% 100%
				Orange threshold : 10 % Red threshold : 50 %

Continue to add monitoring and table rules.

cation name : dataworks_d	doc > Table Name :	rpt_user_info_d > Pi					
	lesponsible : wangdan		+ Add	monitoring rules		+ Quick add	
I partition expression	(emplate rules (0)	Self-help rules (0)					
(gyyymredd-1)	Field name	Template name	Field name : Ruletype :	vid V Pield null velue O Pi	rid repetition value		
			Field Type : Template type :	Table level rules \lor SQL task table rows, 17, \lor	Strong and weak (trend :	Strong 💌 v Absolute value	wesk
			Comparison of p volatility :	25%	75%	75%	100%
			Orange threshold :	0 S	Red threshold :	50	8
				Grange Breakfold :	Orange threshold : 0 ×	Orange threshold : 0 N Red threshold :	Congethreehold : 0 N Red threehold : 50

As you may notice, the lower are the tables in the data warehouse, the more times the strong rules are set. That's because the data in the ODS layer is used as the raw data in the warehouse and you need to ensure the accuracy of its data, avoiding poor data quality in the ODS layer, and stop it in time.

Data quality also provides an interface for task queries on which you can view the validation results for configured rules.

3 Best practices for setting scheduling dependencies

In DataWorks V2.0, when configuring scheduling dependencies, dependencies between tasks need to be set according to the output name of the current node as an associated item. This article details how to configure the input and output of task scheduling dependencies.

How to configure the node input of a task

There are two ways to configure the node input: one is to use the automatic code parsing function to resolve the dependency of the task, the other is to manually enter the task dependency (manually entering the **Upstream Node Output Name**).

Sig test_sqL01 ● 🛔 test 🛛 🗙						<.>	Ξ
" " T & 🙃 📀 : (9)						0	8M
1odps sql 2	X Dependencies ③ Auto Parse: ● Yes ◯ No Parse 1/0						
<pre>SELECT * SELECT * FROM project_b_name.pm_table_b ;</pre>	Upstream Node Enter an output name or outp						elationshi
			Node Name				
	project_b_name.pm_table_b				Auto Parse		Version
	Output Enter an output name						
			am Node Name				
	MaxCompute_DOC.500143227_out	- @			Added by Default		
	MaxCompute_DOC.test_sql_01 @	- Ø			Added Manually		
	MaxCompute_DOC.pm_table_a @	MaxCompute_DOC.pm_table_a			Auto Parse		



Note:

When manually entering an upstream node, the input is **Output Name** of the parent node. If the parent node task name does not match the parent node's output name, be sure to enter the node output name correctly.

When configuring an upstream node, you may encounter problems with the upstream node parsed automatically is an invalid upstream dependency. A method of identifying whether dependencies are valid: view the parsed upstream dependencies and check if the value is displayed in the **Upstream Node ID** column, as shown in the following figure.

Se Task_1 x Se project_b_name.pm_table x Se test_sql	_01 💿 🏯 test 🛛 🗙					Ξ
변 🛱 여 🐻 🔂 🕤 : 🌀						
1odps sql 2***********************************	Upstream Node Enter an output name or output tab					
4create time:2018-12-27 10:25:30						
5 INSERT OVERWRITE TABLE pm_table_a 7 SELECT * 8 FROM project b name.pm table b	MaxCompute_DOC_root	maxcompute_doc_root	700000822799	diplot, dana	Added Manually	Relation
9 ;	Ourput test					
	MexCompute_DOC.500143227_out				Added by Default	
	MexCompute_DOC.test_sql_01 @				Added Manually	

The configuration of task dependencies is essentially to set the dependencies between two nodes. Only the nodes that exist will be able to set up valid dependencies, task dependencies can be set successfully.

Invalid upstream dependency

Invalid upstream dependencies are usually in two cases.

1. The parent node does not exist.

🔄 project_b_name.pm_table x 🔄 test_sqL01 🌖 🛔	test X					
E E I & & • • • •						
1odps sql author:tlmm 4create tlmm: 5 6 INSERT OVERWRITE TABLE pm_table_m 5 cscret	Yes No Purse IO Auto Purse : • Yes No Purse IO Upstream Node Enter an output name or output table name					
<pre>8 FROM project_b_name.pm_table_b 9 ;</pre>						
	project_b_name.pm_table_b				Auto Parse	
	Output test			Invalid upstream of		
)	MaxCompute_DOC.500143227_out				Added by Default	
	MaxCompute_DOC.pm_table_a	MaxCompute_DOC.pm_table_a			Auto Parse	

2. The parent node output does not exist.

Submit	×	
You submitted 2 nodes. You can only submit your nodes	β.	
Node ID "test_sql_01"	An error occurred while submitting.	
Node ID "project_b_name.pm_table_b*	 Dependent parent node output project_b_name.pm_table_b does not exist and cannot submit this node. Please submit parent node first 	×

Invalid upstream dependencies typically occur because the parsed parent node output name does not exist. In this case, it may be due to the fact that the table "project_b_name.pm_table_b" does

not output task, or the node output is configured incorrectly for the table output task and can't be parsed. There are two solutions:

- **1.** Confirm that the table has an output task.
- **2.** Confirm what the output name of this table's output task is, and manually enter the node output name into the dependent upstream node.

🖾 test_sqL01 🗙 🛔 test 🛛 🗙								Ξ
0 : • f & • • • •								08M
1odps sql 2 3author;tlinu 4create time:2018-12-27 10:25:30 5	X Dependencies ⑦ Auto Parse : • Yes No Parse I/0	3 After you find	the output name,	enter it here.				
6 INSERT OVERWRITE TABLE pm_table_a 7 SELECT * 8 ERGM project b pame pm_table_b	Upstream Node Enter an output name or output	ut table name 👻 🕂 Use The We						Relation
9 ;	Upstream Node Output Name							
	project_b_name.pm_table_b					Auto Parse		
	Find the output task for the	ne table and view the o	utput name of the	output task.				
	Output Enter an output name							
The output name is here.	Output Name 2							
	MaxCompute_DOC.500143227_out	- Ø				Added by Default		
	MaxCompute_DOC.pm_table_a	MaxCompute_DOC.pm_table_a		•	-	Auto Parse	Ê	



Note:

When you enter an upstream node manually, you enter the parent node's output name. If the parent node task name does not match the parent node's output name, be sure to enter the node output name correctly.

For example, the output name of the upstream node A is A1, and downstream node B depends on node A. At this point, enter A1 in the input box of the upstream node, and click the plus sign on the right to add it.

How to configure upstream dependencies

If your table is extracted from the source library and there is no upstream, you can click **Use The Workspace Root Node** to obtain upstream dependencies.

🕞 test_sqL01 🌑 🚠 test 🛛 🗙								
≝ ≝ F & ⊖ : §								
1odps sql 	X Dependencies ③ Auto Parse: ○ Yes ④ No Parse //3	W	/hen you do not knov	v what the upstream node	is, click it.			
<pre>8 FROM project_b_name.pm_table_b 9 ;</pre>	Upstream Node Enter an output name or output table		- Use The Workspace Root No	de Automatic recommended				
	MaxCompute_DOC_root			maxcompute_doc_root		digitar, dana	Added Manually	
2	MexCompute_DOC.500143227_out						Added by Default	

How to configure the node output of a task

The simplest way to efficiently configure the node output is: the node name, the node output name and the node output table name share the same name and three in one. The advantages are as follows.

- 1. You can quickly know which table this task is operating on.
- 2. It is possible to quickly know how far this task will impact if it fails.
- **3.** When you use auto parsing to configure task dependencies, as long as the node output is consistent with the three-in-one rule, the precision performance of automatic parsing is greatly improved.

Automatic parsing

Automatic parsing: refers to automatically parse scheduling dependencies by the code. Implementation principle: only table names can be obtained in the code, and the automatic parsing function can parse the corresponding output task according to the table name.

For example, the type node code is shown below.

```
INSERT OVERWRITE TABLE pm_table_a SELECT * FROM project_b_name.
pm_table_b;
```

The dependencies parsed are as follows.

🔄 test_sqL01 💿 👗 test 🛛 🗙									
" " h i 🗄 📀 : 🕲									
<pre>1odps sql 2</pre>	the node output name Dependencies Auto Perse : Yes No Perse 10								
7 SELECT * 8 FROM project_b_name.pm_table_b 9 ;	Upstream Node Output Name project_b_name.pm_table_b						Source Auto Parse		
upstream node output name	Comput Enter an output neme								
	Output Name	Output Table Name		Node Name	Downstream Node ID				
	MarComputer DOC 500140237 cost						Added by Default		
	MaxCompute_DOC.pm_table_a 🗭	MaxCompute_DOC.pm_table_a	-		-	-	Auto Parse	÷	

DataWorks can automatically parse the node which this node needs to be dependent on project_b_name to output pm_table_b, and the final output of the node pm_table_a . Therefore, the resolution is that the parent node output name is project_b_name. pm_table_b, and the node output name is project_name.pm_table_a(The project name is MaxCompute DOC).

- If you do not want to use dependencies that are parsed from the code, select No.
- If there are many tables in the code that are temporary tables: For example, the table beginning with t_ is a temporary table. Then the table is not parsed as the schedule dependency. The

definition of temporary tables is that you can define which form the table begins with is a temporary table by project configuration.

- If a table in the code is both the output table and referenced table (depended table), it is parsed only as the output table.
- If a table in the code is referenced or output for multiple times, only one scheduling dependency is parsed.

Note:

By default, a table with a name starting with t_{i} is recognized as a temporary table. Auto parsing does not resolve the temporary table. If the table with a name starting with t_{i} is not a temporary table, contact your project administrator to modify it in the project configuration.

镨	Configuration Center
	Project Configuration
ī	Templates
\$	Theme Management
	Table Levels
3	Backup and Restore

How to delete the input and output of a table

When you're in the process of data development, you often use static tables (data is uploaded to a table from a local file), this static data does not actually output task. At this time, when configuring dependencies, you need to delete the input of the static table: if the static table does not satisfy the form of $\mathbf{t}_{,}$, it will not be processed as a temporary table, in which case you need to delete the input of the static table.

You select the table name in the code, click **Remove Input**.

Sq test	_sql_01 🔵 🛔	test							Ξ
•		ي 🖯	\odot	: (\$			08	M
1 2 3 4 5	odps sql ********* author create ti ********								Schedule
6 7 8 9	INSERT OVER SELECT * FROM pro	RWRITE TA	BLE pm_t me.pm_ta	able_b	a Add Input Add Output				Relationship
>					Remove Input Remove Output				Version
					Go to Definition Peek Definition Change All Occurrences	Ctrl+F12 Alt+F12 Ctrl+F2			Structure
					Cut Copy		-		
					Command Palette	F1	<u>۲</u>		

If you are upgrading from DataWorks to DataWorks V2.0, we set the node output for the migrated DataWorks task to ProjectName.NodeName for you by default.

DataStudio MaxCompute_DO		Cross-project cloning	Operation Center	🔍 💷	Englist
Data Developm 🖉 🗟 📑 Ċ 🕀 🖆	test_sqL01 × 👗 test 🛛 ×				
Enter a file or creator name					M80
> Solution 🔡					
✓ Business Flow	3 Dependencies ()				
Y 📇 test	4 Auto Parse: O Yes No. Parse/0				
> 🚍 Data Integration					
 Data Development 	7 SEI Upstream Node Enter an output name or output table name 👻 🕂 Use The Workspace Root Node Automatic recommended				
Sq abc Locked by xuailin 12-18					
Sq Create_Table Me Locked 12-					
Sq JAVA_test Locked by dtplus_	MaxCompute_DOC_root - maxcompute_doc_root NVMMX2VV	diplos_dece	Added Manually		
Py Pytest Me Locked 12-24 13:					
VI stort Me Locked 12-25 09:53	Output Enter an output name				
• Sq Task_1 Me Locked 12-26 15					
Sq Task_2 Me Locked 12-2510					
Sq Task_3 Me Locked 12-2510	Hard Commune DAG 6001 (2022) and (2		Added by Defende		
• Sp test Locked by dtplus_docs	MaxCompute_DOC.S00145227_out		Added by Default		
Mr testMR Locked by dtplus_dos	MaxCompute_DOC test_sql_01 @ - @ -		Added Manually		
• Sq test_sqL01 Me Locked 12-			,		

Attentions

When the task dependency configuration is complete, the submitted window shows an option: whether confirm to proceed with the submission when the input and output does not match the code blood analysis.

The premise of this option is that you have confirmed that the dependencies are correct. If you cannot confirm, you can confirm the dependencies as described above.

Submit New Version		×
Note :		
	input and output and blood analysis does not match the code. n user submitted input: MaxCompute_DOC_root n blood analysis of input: project_b_name.pm_table_b n submitted by the user of the output: MaxCompute_DOC.test_sql_O1 n blood analysis of output: MaxCompute_DOC.pm_table_a	As long as you confirm that the dependencies are correct, continue to perform the submission.
2 Tick it, v Tips: If you don't write a no the confirm button] I confirm to proceed with the submission. Write a note about the change, a Ote, you can't click	and click confirm button.