# Alibaba Cloud DataWorks

**Best Practices** 

Issue: 20190228

MORE THAN JUST CLOUD | **[-]** Alibaba Cloud

# <u>Legal disclaimer</u>

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 due to product version upgrades , adjustments, 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 intellectual 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. 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.	<b>Danger:</b> 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 informatio n, 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 instructions, 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 list instanceid Instance_ID
[] or [a b]	It indicates that it is a optional value, and only one item can be selected.	ipconfig [-all -t]

Style	Description	Example
{} 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 Workshop1
1.1 Workshop course introduction1
1.2 Data acquisition: log data upload2
1.3 Data processing: user portraits23
1.4 Data quality monitoring38
2 Data migration
2.1 Migrate data from Hadoop to MaxCompute52
2.2 Migrate JSON data from OSS to MaxCompute
2.3 Migrate JSON data from MongoDB to MaxCompute
3 Data development81
3.1 Best practices for setting scheduling dependencies
3.2 Use Eclipse to develop a Java-based UDF87
3.3 Use MaxCompute to analyze IP sources102
3.4 Performing a task at a specific time with branch node108
4 Data security117
4.1 Log on to DataWorks through a specific IP address with a RAM user117

# 1 Workshop

### 1.1 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 architecture, 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.

### 1.2 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.
- 2. Select Products > Analytics & Big Data > MaxComputute and go to the MaxCompute product details page.



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

### Create workspace

1. Log on to the DataWorks console by using a primary account.

- 2. You can create a workspace in two ways.
  - On the console Overview page, go to Commonly Used Features > Create Workspace.

			Overview	Workspaces	Resources	Comp
🕟 Da	ataWorks	Data	aStudio∙Data Inte	gration · MaxCor	npute	
Shortcuts						
Data Analytics		Da	ata Integration	Maintenance C	enter	
Workspaces						
DTplus_DOC	China East 1		100.000	China East 1	1.00	100
Created At:Jan 30, 2019, 10:18 Compute Engines:MaxComput Services:Data Studio Data Inte	Created At:Jan 30, 2019, 10:18:52 Compute Engines:MaxCompute PAI calculation engine Services:Data Studio Data Integration Data Management D		Created At:Jan 10, 2019, 13:46 Compute Engines:MaxComput Services:Data Studio Data Inte	:08 e PAI calculation engine gration Data Management D	Created / Compute Services:	At:Sep 02, 20 Engines:Ma Data Studio
Workspace Settings	Data Analytics		Workspace Settings	Data Analytics	Wor	kspace Setti
Data Service	Data Integration		Data Service	Data Integration		Data Service
Commonly Used Feature	2S X Activate CDN					

• On the console Workspace page, select region, and then click Create Workspace in the upper right corner.

		Overview Work	spaces Resources Con	npute Engines		
China East 1 China East 2 China Asia Pacific NE 1 Middle East 1	a South 1 China North 2 Hong Kor Asia Pacific SOU 1 Asia Pacific SE 5 Search	g US West 1 Asia Pacific SE 1 UK	US East 1 EU Central 1 Asia Pac	ific SE 2 Asia Pacific S	E3	Create Workspace Refresh
Workspace/Display Name	Mode	Created At	Administrator	Status	Service	Actions
10.m.30	Single Environment	Feb 21, 2019, 17:21:10	- Apple and Appl	Enabled	∞ 🔨	Workspace Settings Data Analytics Change Services Data Integration Data Service More —
	Development and Production En vironments	Feb 21, 2019, 17:14:17	1941.000	Enabled	~	Workspace Settings Data Analytics Change Services Data Service More -
						Total 1 Pages 🛛 🕹 🚺

3. Fill in the configuration items in the Create Workspace 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 analytics, O&M, and administration are selected by default.

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

Create Workspace	×
Basic Information	
* Workspace Name :	
Display Name :	
* Mode :	Single Environment ②
Description :	
Advanced Settings	
* Task Recurrence :	
* SELECT Result Download :	
Information of MaxCompute	
* MaxCompute Project Name :	0
* Identity to Access MaxCompute:	Norkspace Owner 🔀 😨
* Resource Group:	Pay per view default resource group $\sim$
	Previous Create Workspace

Note:

• The workspace name needs to begin with a letter or underline, and can only contain letters, underscores, and numbers.

- The workspace name is globally unique, it is recommended that you use your own easy-to-distinguish name as the project space name for this lab.
- 5. Once the workspace has been created successfully, you can select the Workspace page to Data Analytics after viewing the workspace space.

		Overview Works	spaces Resources	Compute Engines		
China East 1 China East 2 China So Asia Pacific NE 1 Middle East 1 As	uth 1 China North 2 Hong Kor ia Pacific SOU 1 Asia Pacific SE 5	ng US West 1 Asia Pacific SE 1 UK	US East 1 EU Central 1	Asia Pacific SE 2 Asia Pacific SE	3	Create Workspace Refresh
	Search					
Workspace/Display Name	Mode	Created At	Administrator	Status	Service	Actions
10.00.00	Single Environment	Feb 21, 2019, 17:21:10	100.0.000	Enabled	∞ 🔥	Workspace Settings Data Analytics Change Services Data Integration Data Service More —
100.000 100.000	Development and Production En vironments	Feb 21, 2019, 17:14:17	100.00	Enabled	~	Workspace Settings Data Analytics Change Services Data Service More -

### Create data source



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
Uverview	Data Source	Dete Source : All Type	Deta Source :     Name			Add Data Source
Resource Consumptio	Data Source Name	Deta Source Type	Link Information	Description	Created At	Actions
<ul> <li>Synchronization Reso</li> <li>Date Source</li> </ul>	odps_first	COPS	ODPS Endpoint. http://service.cdps.aliyun.com/api ODPS Project Name: DataWorks_DOC Access M: UTAkwCi7pgJJSeQ	connection from odps celc e ngine 61155	2018-08-27 13:32:26	
Resource Group						

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: lXnzUngTSebt3SfLYxZxoSjGAK6IaF
- 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 :	ApsaraDB for RDS ~	
* Data Source Name :	rds_workshop_log	
Description :	rds log synchronization	
* RDS Instance ID :	rm-bp1z69dodhh85z9qa	?
Primary Account of :	1156529087455811	0
RDS Instance		
* Database Name :	turight, may	
* Username :	here, who	
* 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.	

### **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 Analytics, select Create Business Flow.
- 2. Fill in the Business Flow name and description.

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

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

*	workshop ×						
۵	n o 🗉 🗐						
	<ul> <li>Data Integration</li> </ul>	Development	Blood			୯ଡେଇ	ର ପ 🖬
	Data Sync						
	<ul> <li>Data Development</li> </ul>						
	🔄 ODPS SQL						
	Sh Shell						
	M ODPS MR						
	VI Virtual Node						
	Py PyODPS						
	SQL Component Node						
	M OPEN MR						

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 : Viri	tual Node ~		
Node Name . wo	iksnop_start		
Destination Folder : Bus	siness Flow/workshop ✓	Cancel	
Create Node		×	
Node Type · De	ta Sunc 🗸 🗸	n'	
Hode Type . Da			
Node Name : rds	s_datasync		
Destination Folder : Bu	siness Flow/workshop 🛛 🗸 🗸		
	Submit	Cancel	
Create Node			×
Node Ty	vpe : Data Sync	~	
Node Nar	me : oss_datasync		
Destination Fold	der : Business Flow/workshop	) ~	
		Submit	Cancel

5. Drag the connection to set the workshop\_start node to the upstream of both data synchronization nodes.



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.

DI rds_datasync ×	vn workshop_start 🔵						
U D &	⊕ C :						
1	X Depend on	Last Interval : 🔲					
	Resources ⑦						
	Resource Group : default_re						
	Dependencies ⑦						
	Auto Parse : 💿 Yes 🔿 No	Parse I/0					
	Upstream Node Enter an		+ Use the proje	ect Root Node			
	Upstream Node Output Ne me	Upstream Node Output Table N ame	Node Name	Upstream Node ID	Owner	Source	Actions
	DataWorks_DOC_root		dataworks_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	
	Submit	Cancel

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

Basics							
	Table Alias :						
	Level 1 Topic :	Select ~	Leve	el 2 Topic : Select	Create Topic	C	
	Description :						
Physical Mod	рI						
	Partition :	Partitioned Table      No	on- L	.ife Cycle : 🛛 🔽	Days :	0	
		Partitioned Table					
	Table Level :	Select ~	Table	Category : Select	Create Level	C	
	Table type :	• Internal table C Extern	al table				

4. 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	
pt	STRING ~					

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

```
logs
// Create
                target
                         table
                                 for
                                       oss
            а
                                       ods_raw_lo g_d
CREATE
         TABLE
                  ΙĒ
                       NOT
                             EXISTS
                                                       (
         STRING
   col
)
PARTITIONE D
                 BY (
       STRING
   dt
);
// Creates
                target table
                                  for
                                        RDS
           а
CREATE
        TABLE
                      NOT
                IF
                            EXISTS
                                       ods_user_i
                                                  nfo_d (
        STRING COMMENT 'User ID',
er STRING COMMENT 'Gender',
  uid
   gender STRING
                                        ,
  age_range STRING COMMENT 'Age
                                       range ',
                    COMMENT ' Zodiac '
   zodiac
            STRING
)
PARTITIONE D
                 BY
                    (
      STRING
   dt
```

- );
- 5. Click Submit to Development Environment and Submit to Production Environment. You can configure both of the tables in this way.

🗮 ods_raw_log_d	× 🛃 worksh	op ×	🌐 tftp		Sq ten_precent_movies $\times$	Di OSS_ratings	× 🛃 Movies_0	DS 🗙 嚞	DataWorks_Test ×	
DDL Mode			t Environmer	1t	Submit to Development Environ	ment Load f			Submit to Prod	uction Environment
		Tal	ole Name	ods_raw	_log_d					
		Business	Process	worksho	p					
Basics										
	Table Alias :	ods_rav	N							
	Level 1 Topic :	worksh	op_table		Level 2 T	opic : table1		Create T	opic C	
	Description :									
Physical Model	_									
	Partition :	💿 Part	itioned Table		ŀ Life C	Sycle :				
		Partition	ed Table							
	Table Level :	Select			Table Cate	gory : Select			evel C	
	Table type :									

### 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}.

[	Destination				Hide
reated by you. Click here to	o check the supported data so	urce types.			
* Data Source :	ODPS ~	odps_first	~	?	
* Table :	ods_raw_log_d	This must be specified.	~		
		Generate Destination	Table		
* Partition :	dt = \${bizdate}	?			
Clearance Rule :	Clear Existing Data Before W	riting (Insert Overwrite)	*		
Compression :	💿 Disable 🔵 Enable				
Consider Empty . String as Null <sup>.</sup>	💽 Yes 🔵 No				

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

Di oss_iii	嚞 workshop 🛛 🗙					
						Admir
02 Mappings		Source Table		Destination Table		
	Location/Value	Туре	0	Field	Туре	Map Fields with the Same Name
	Column 0	string	•	≿o col	STRING	Map Fields in the Same Line

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

03 Channel		
You can control the data sy	mchronization process through the transmission rate and the number of allo	wed dirty data records. See data synchronization documents.
P 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 ~	

- 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 🌑	sq creat_table_ddl ×	o write_result x			
	ি ঠে বি	ê <b>2</b>			
01 Data Source		Source			De
	The data so	ources can be default da	te sources o	or data sources o	created by you. Click here to
Data Source :	MySQL	rds_workshop_k	og v	0	* Data Source :
* Table :	'ods_user_info_d' >				
		Add De	rta Source +		
Data Filtering :	Enter SQL WHERE a incremental data ay the keyword "WHER	statements, which are us inchronization. Do not in E."	ed for clude	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}.

🔃 rds datasync 💿 🥤	creat table ddl X N write result X			
Con restanting to				
	1 i 🗉 🗃 🖬			
01 Data Source	Source		Destination	Hide
	The data sources can be default data sources o	or data sources created by you. Click her	re to check the supported data source types.	
* Data Source :	MySQL v rds_workshop_log v	Deta Source :	ODPS v odps_first v	0
* Table :	`ods_user_info_d' × ∽	* Table :	ods_user_info_d v	
	Add Data Source +		Generate Destination Tabl	
Data Filtering :	Enter SQL WHERE statements, which are used for incremental data synchronization. Do not include the keyword "WHERE."	Partition :	dt = \$(bizdene)	
		Clearance Rule :	Clear Existing Data Before Writing (Insert Overwrit_ $$	
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 rđ	t_datasy	nc 🌒	Se cre	at_table,	,ddl ×		vrite_result	×					
۳	۲	D	ſ	6			ø						
									500	ig as Null			
02	Mappi	ing				Sour	ce Table			Desti	ination Table		
			Fiel				ype	Ċ			Field	Туре	Map of the same name
			uid				ARCHAR	•		•	uid	STRING	Enable Same-Line Mapping
			gen	der			ARCHAR	•		•	gender	STRING	Cancel mapping
			age	range			ARCHAR	•		•	age_range	STRING	
			zod	liac			ARCHAR	•		•	zodiac	STRING	
			Add	1+									

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 alk	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 r@ber of dirty data records. Dirty data is allowed by default. task ends.	dirty data records, the
Task's Resource Group :	Default resource group ~	

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 ×	Di write_result ×				
	া 🐻 🔍	- -				
01 Data Source		Source		Destination		Hide
	The data sou	rces can be default data source	s or data sources created by you. Click he	ere to check the supported da	ta source types.	
* Data Source :	MySQL	rds_workshop_log	⑦ * Data Source	: ODPS v	odps_first	?
* 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.
- 2. 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			
ବ୍ଦ 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 rds\_datasync task and select View Log.

When the following words appear in the log, it indicates that the synchronization task runs successfully and synchronizes a batch of data.

Di oss_datasync 🗙 👫 worksho		Ξ
✓ Data Integration	୯ ଡ ୧ ୧ ୮	Paramet
Di Data Sync		
Sql       Sql       Sql       Sql       Sql       Component       PyODPS       Virtual Node	Di rds_datasync IDi oss_datasync IDi	Operation Records Version
2002rec/s 20028 201902 26 15 58 55 NFO	of the Shell command 0  on of Shell command completed  context cleanup temp file success.  thronization ended with return code: [0].  of the Shell command 0  of Shell command 0  of Shell command completed	

3. Right-click the oss\_datasync task and select View Log. The confirmation method is consistent with the rds\_datasync task.

### Check if the data is successfully imported into MaxCompute

- 1. Click temporary query in the left-hand navigation bar.
- 2. Select New > ODPS SQL.

Data	DataStudio	DataWorks_DOC	v				Cross-project cloning	Operation Center	٩	wangdan	English
ш	Queries 8	BCO	Se select_01	🛃 workshop							
	Enter a file or creator r	2 Folder	<u>™                                    </u>		C 2	22					
*	✓ Queries	Create >	ODPS SQL	3							
民	1 test birdboll00		Shell ODPS SO	ll. angdan							
B	Select_01 Md	他定 08-30 14:01									

3. Write and execute SQL statement to check the entries imported into ods\_raw\_log\_d.

Sq selec	t_01 🔵	🛃 work	shop 🤉	*
	60	) 🕨		C 🗱
1 2 3 4 5 6 7 8	odps ***** autho creat ***** select select	sql ******** count(*) count(*)	******* 018-08 ****** from ( from (	<pre>-30 14:01:29 -30 14:01:29 -30 raw_log_d where dt=20180830; ods_user_info_d where dt=20180830;</pre>
Runtime	e Log	Resu	k[1] 🤇 🤇	ĸ
1 _c0 2 0	A	~		

4. Also write and execute SQL statements to view the number of imported ods\_user\_info\_d records.



date ;

```
select count (*) from ods_user_i nfo_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*.

### 1.3 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 inData 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
  - 1. 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
                  ΙF
                       NOT
                             EXISTS
                                       ods_log_in
                                                   fo d
                                                          (
                 comment ' IP
        string
                                  address ',
   ip
                            ' User
                                    ID ',
   uid
         STRING
                  COMMENT
                            ' time : yyyymmddhh : mi : ss
   time
          string
                   comment
                               ' server
                                                             code ',
   status
            string
                     comment
                                         return
                                                    status
                              ' the
           string
  bytes
                    comment
                                      number
                                               of
                                                    bytes
 returned
                the
                       Client '
            to
                                region , get
                                                from
                                                        IP ',
   region
            string
                     comment
                               ' HTTP
   method
            string
                     comment
                                        request
                                                  type ',
                  comment 'url'
   url
         string
                                 ' HTTP
                                          Protocol
                                                     version
   protocol
              string
                       comment
number '
                      comment ' source
                                           ures '
   referer
             string
                     comment ' terminal
            string
   device
                                            type
                       comment ' Access
                                                              feed
   identity
             string
                                            type
                                                   crawler
        unknown '
user
)
PARTITIONE
            D
                     (
                 ΒY
  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
            а
                сору
                       table
         TABLE
CREATE
                 ΙF
                      NOT
                            EXISTS
                                     dw_user_in fo_all_d (
                          ' User
                 COMMENT
                                   ID '
  uid
        STRING
                    COMMENT ' Gender
  gender
          STRING
  age_range STRING COMMENT ' Age
                                        range ',
                            ' Zodiac
                                        sign '
           STRING COMMENT
  zodiac
                             ' region ,
                                                     IP ',
  region
           string
                    comment
                                         get
                                               from
                             ' terminal
                    comment
  device
           string
                                          type '
                                                          feed
  identity string
                      comment
                              ' Access
                                          type
                                                 crawler
       unknown '
user
                ,
                    comment ' HTTP
  method string
                                      request
                                               type ',
  url string comment 'url',
  referer string comment ' source
                                        url '
  time string comment ' time : yyyymmddhh : mi : ss '
)
PARTITIONE
            D
                BY (
  dt
       string
);
```

· Create rpt\_user\_info\_d table

The following are table creation statements:

```
Create
                  copy
                          table
___
              а
 Create
          Table
                   if
                        not
                               exists
                                         rpt_user_i
                                                     nfo_d (
                                      ID ',
                             ' User
   uid
         STRING
                   COMMENT
                                  region ,
                                             get
   region
            string
                      comment
                                                    from
                                                           IP ',
                                ' terminal
            string
   device
                      comment
                                              type ',
                  comment ' pv ',
NG COMMENT ' Gender ',
        bigint
   pv
   gender STRING
                                  ' Age
                         COMMENT
   age_range
              STRING
                                            range ',
                      COMMENT ' Zodiac
                                            sign '
   zodiac
            STRING
)
 PARTITIONE D
                  ΒY
                     (
   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 :			
Resource Type :	JAR	~	
	Upload to ODPS The resource will also be uploaded to ODPS.		
File :	ip2region.jar (4.62M)	×	
		OK Cance	

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

111	Data Developn 🖉 🗟 📮 Ċ 🕀 🖆	🜆 ip2region.jar 🗙 🟯 workshop 🗙
	Enter a file or creator name	ចា 🗈 🗄
*	> Solution	
民	✓ Business Flow	Upload Resource
а	> 🚣 base_cdp	Saved Files : ip2region.jar
	> 🚣 works	Unique Resource Identifier: OSS-KEY-oyhaqmgvvtsrgbo7yeqh1vz3
	Y 🚣 workshop	
Ħ	V 🔁 Data Integration	Upload to ODPS. The resource will also be uploaded to ODPS.
R	• [0] rds_sync Mel凤注 08-30 10:0	Re-upload : Upload
_	Deta Devalopment	
f.	Sal create table ddl Mellinii 08	
Ť	• W workshop_start Mel位定 08-	
	• Sq ods_log_info_d Me锁定 08-3	
	• Sq dw_user_info_all_d Me锁宽	
	• Sq rpt_user_info_d Me做定 08-3	
	> 🧾 Table	
	Y 💋 Resource	
	• 🜆 ip2region.jar Me微定 08-30	
	> 📴 Function	
-	> 🔚 Algorithm	



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
	-		

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

Fx getregion 🗙 Ja ip2region.jar 🗙 嚞 v	vorkshop X
≞ t t t t C	
Registry Function	
Function Name :	
* Class Name :	org.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
  - 1. Double-click the ods\_log\_info\_d node to go to the node configuration page and write the processing logic.

Sq ods	_log_info_d	Fx getregion	×	J∎ ip2region.jar	× 🛔	workshop	×	
	≞♪	լջ 🗄	€					
	create t:	ime:2018-08 ******	-30 15 *****	:51:50 **********				
	INSERT OVER	RWRITE TABL	E ods_	log_info_d 🛛	PARTIT	ION (dt=	\${bd	p.system.bizdate})
	SELECT ip							
	, uid							
	, time							
	, status							
	, bytes	Use costom	UDF to	get region	by ip			
	, getregio	n(ip) AS re	gion -	-Request is	divid	ed into t	thre	e fields by regularization.
	, regexp_s	ubstr(reque		^[^]+ )') /	s met	hod		
	, regexp e	xtract(requ	est, '	^[^]+(.*)	[^]+	\$') AS ur	rl	
	, regexp_s	ubstr(reque	st, '(	[^ ]+\$)') As	prot	ocolGe		ore precise URL through regular
	, regexp ex	xtract(refe	rer, '	^[^/]+://(['	·/]+){	1}') AS I	refe	rerGet terminal information a
	, CASE							

The SQL logic is as follows:

```
INSERT
         OVERWRITE
                      TABLE
                                ods_log_in fo_d
                                                    PARTITION
                                                                ( dt =
${ bdp . system . bizdate })
SELECT
          ip
    uid
  ,
     time
  ,
     status
  ,
            -- use a
                           custom
                                     UDF
                                           to
                                                 get
                                                            locale
     bytes
                                                       а
        IΡ
 over
     getregion ( ip ) as
                               region -- the
                                                  request
 difference is
                  divided
                                      three
                                               fields
                                                                   the
                               into
                                                        through
   regular
    regexp_sub str ( request , '(^[^ ]+ )') AS method
regexp_ext ract ( request , '^[^ ]+ (.*) [^ ]+$') AS
regexp_sub str ( request , '([^ ]+$)') AS protocol
  ,
                                                                  url
  ,
                         urls with regular clear
get
     more precise
                                                            refer
    regexp_ext ract ( referer , '^[^/]+://([^/]+){ 1 }') AS
 referer
                                 informatio n
         -- Get
                     terminal
                                                  and
                                                        access
                                                                  form
   through
             agent
     CASE
WHEN
       TOLOWER ( agent ) RLIKE ' android ' THEN ' android '
                                RLIKE ' iphone ' THEN ' iphone '
     WHEN
            TOLOWER ( agent )
                                        'ipad ' THEN 'ipad '
     WHEN
                                 RLIKE
            TOLOWER ( agent )
            TOLOWER ( agent )
                                       ' macintosh '
                                                       THEN
     WHEN
                                 RLIKE
macintosh
                                                     phone ' THEN
                                                                     - 1
     WHEN
            TOLOWER ( agent )
                                RLIKE
                                       ' windows
windows_ph one '
     WHEN
            TOLOWER ( agent )
                                RLIKE
                                       ' windows ' THEN
                                                            . 1
windows_pc '
     ELSE
           ' unknown '
   END
       AS
              device
     CASE
     WHEN
            TOLOWER ( agent )
                                RLIKE '( bot | spider | crawler |
slurp )'
           THEN ' crawler
```

```
TOLOWER ( agent ) RLIKE ' feed '
      WHEN
      OR regexp_ext ract ( request , '^[^ ]+ (.*) [^ ]+$')
 RLIKE ' feed ' THEN ' feed '
                                             RLIKE '( bot | spider |
      WHEN
              TOLOWER ( agent ) NOT
 crawler | feed | slurp )'
AND agent RLIKE '^[ Mozilla | Opera ]'
             regexp_ext ract ( request , '^[^ ]+ (.*) [^ ]+$')
' feed ' THEN ' user '
                                                                                NOT
      AND
   RLIKE
      ELSE ' unknown '
   END AS
                 identity
   FROM
           (
        I (
ELECT SPLIT ( col , '##@@'
SPLIT ( col , '##@@')[ 1 ]
SPLIT ( col , '##@@')[ 2 ]
      SELECT
                                  '##@@')[ 0 ] AS
                                                            ip
                                            AS
                                                   uid
     ,
                                            AS
                                                   time
     ,
                          '##@@')[3
                                         ]
                                            AS
        SPLIT
                ( col ,
                                                   request
     ,
                          '##@@')[ 4 ]
'##@@')[ 5 ]
'##@@')[ 6 ]
                                            AS
        SPLIT
                ( col ,
                                                   status
     ,
                ( col ,
                                            AS
                                                   bytes
        SPLIT
        SPLIT ( col , '##@@')[ 6 ]
SPLIT ( col , '##@@')[ 7 ]
1 ods_raw_lo g_d
                                            AS
                                                   referer
     ,
                                            AS
                                                   agent
   FROM
           dt = ${ bdp . system . bizdate }
   WHERE
)
  а;
```

2. Click Save.



- · Configure dw\_user\_info\_all\_d Node
  - 1. 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_in
                                           fo_all_d
                                                       PARTITION
                                                                  (
dt ='${ bdp . system . bizdate }')
SELECT
          COALESCE ( a . uid ,
                                b.uid)
                                            AS
                                                 uid
     b . gender
  ,
     b . age_range
  ,
     b . zodiac
  ,
     a . region
  ,
     a . device
  ,
     a . identity
  ,
     a . method
  ,
    a . url
  ,
    a . referer
  ,
         time
     а
      •
 FŔOM
  SELECT
          *
         ods_log_in fo_d
  FROM
           dt = ${ bdp . system . bizdate }
  WHERE
)
  а
LEFT
       OUTER
                JOIN (
  SELECT
           *
          ods_user_i nfo_d
  FROM
           dt = ${ bdp . system . bizdate }
  WHERE
)
  b
ON
      a . uid = b . uid;
```

2. Click Save.

· Configure a rpt\_user\_info\_d Node

1. 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_i
                                           nfo_d
                                                   PARTITION
                                                               ( dt
='${ bdp . system . bizdate }')
SELECT
        uid
    MAX ( region )
    MAX ( device )
    COUNT (0) AS
                       pv
 ,
    MAX ( gender )
 ,
    MAX ( age_range )
 ,
    MAX ( zodiac )
       dw_user_in fo_all_d
FROM
        dt = ${ bdp . system . bizdate }
WHERE
GROUP
        ΒY
             uid ;
```

2. Click Save.

# Submitting Business Flows

- 1. Click Submit to submit the node tasks that have been configured in the Business Flow.
- 2. 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		Node Name	
		rds_datasync	
		oss_datasync	
		ods_log_info_d	
Description	worksho	p user portrait part is written logically.	
	🔽 Ignore I	I/O Inconsistency Alerts	
		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.



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



The query statement is as follows:

View in \_\_\_ the data the data box day '' '' business limit 10 select \* From glaswhere dt ;

**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.

- Operation Center DataWorks\_DOC2 💎 DataStudio 🔌 dataworks\_der 🕑 0&M Overview Q, ✓ Business Flow: Please select Node Type: Please select Solutions: 
   ✓
   Nodes
   Modified Today
   Paused (Frozen) Node
   Reset
   Clear
   🚽 Task List Baseline: Please select C Refresh Name: Node ID Testing Envrionment. Please be cautious. C⊕∈ 🚯 Manual Task workshopstart 700000564174 Task 0&M rds\_数据同步 700000564176 Alem oss\_数据同步 700000564175 700000564180 ods\_log\_info\_d dataworks\_doc2\_root dw\_user\_info\_all\_d 700000564181 700000564182 rpt\_user\_info\_d dataworks\_doc2\_root 700000564173 shopstart rds\_数据同步 More 💌 < 1/1 >
- 2. Select Workshop Business Flows in the Task List.

3. Right-click the workshop\_start node in the DAG graph and select Patch Data > Current and downstream nodes.

dataworks_doc2 Virtual Node	_root				
workshopsta Virtual Noc OSS_数据同 Data Integrat	rt Show Parent Node> Show Child Node > Node Details View Code Edit Node View Instances View Lineage	rds_数据同步 Data Integration		Node ID: Node Name: Schedule Type::	7000005641 workshopsta Day Schedul
	Patch Data > Freeze Unfreeze	Current node	wnstr ent a	eam nodes	dataworks_d More m nodes

Retroactive Insertion				×				
* Retroactive Insertion Name:	P_workshop_start_	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)							
vorkshop_sta	art			Virtual Node				
create_table_	ddl			ODPS_SQL				
OSS_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_	_d			ODPS_SQL				
				OK Cancel				

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

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	lame: Please select	Y Node Type: Please	select Y Owner	Select an owner	¥
<ul> <li>Task List</li> </ul>	Run Dete: 2018-08-31	Business Date: Sele	ct date 🛛	Baseline: Please select	Y My Nodes	Reset Clear	
Cycle Task							C Refresh Hide Search
🚯 Manual Task	Instance Name	Status	Task Type	Owner	Timer	Business Date	Actions
- Tesk OSM	<ul> <li>P_workshop_start_20180831_105048</li> </ul>	Running					✓ Batch Terminate
Cycle Instance	✓ 2018-08-30	Running				2018-08-30	v
	workshop_start	⊗ Ran	Virtual Node	wangdan	2018-08-31 00:05:00	2018-08-30	DAG   Terminate   Rerun   More 🔻
DR Manual Instance							
Testing Instance							
PatchData							
<ul> <li>Alerm</li> </ul>							

#### 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*.

# 1.4 Data quality monitoring

This topic 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, notificati on, 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\_info\_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.

≡	MaxCompute V	Tables		
DQC Monitoring	Search by keyword.	ods_raw_log_d		
My Subscription	dqc_test	Table Name	Owner	Actions
Rule Configuration		ods_raw_log_d	And a second	Configure Monitoring Rules
🛱 Task Query				< Previous 1 Next >

Select the ods\_raw\_log\_d table and click Rule Configuration to go to the following page.

≡	⇔ Rules								
E Overview	Rules > Project Name : dqc_test_0221 > Table Name : ods_raw_log_d > Partition Expression : Bind to Schedules A								
My Subscription	Existing Partition Expression	Rule Templates ( 0 ) Custom Ru	ules ( 0 )						
Task Query	No partition expression is found. Add a partition express first.	Rules name Rules Field	Strong Rules Template	Dynamic Comparis Threshold Method	on Warning Threshold	Error Threshold Expected Value	Configured By	Actions	
				Ν	o data				

You can review the data sources for this ods\_raw\_log\_d table. The data for ods\_raw\_log\_d table is from OSS. 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.





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 Rule.

= DQC Monitoring	S Rules Rules > Project Name : dqc_	> Table Name : ods_raw_log_d > Partition Expression : dt=\$[yyymmdd-1] Bind to Schedules]	٨								
EB Overview											
Hy Subscription	Existing Partition Expression	Rule Templates (0) Custom Rules (0) Owner :	Test Run Subscriptions Create Rule More •								
Rule Configuration	- dt=\$[yyyymmdd-1]	Rules name Rules Field Strong Rules Template Dynamic Comparison Threehold Mathod	Warning Error Threshold Expected Configured By Actions								
		плоятой тепои	The shou								
		No data									

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

≡ → DQC Monitoring	S Rules		Rule Template	Custom Rules			
EE Overview	Rules > Project Name : dqc_to > Table Name	: ods_raw_log_d > Partition Expression		+ Add Monitoring Rules	+ Add		
My Subscription	Existing Partition Expression Rule Templates (	Custom Rules ( 0 ) Owner :					
Rule Configuration     Task Query	dt=Slyyyymmdd-1      Rules name	Rules Field Strong Rules Tem					

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

E DQC Monitoring	S Section		_	Rule Template Custom Rules							
88 Overview	Rules > Project Name : dqc_te	> Table Name : ods_raw_log_d	> Partition Expression	+	Add Monitoring Rules		+ Add				
Hy Subscription	Existing Partition Expression	Rule Templates (0) Custom Rule	les ( 0 ) Owner		·						
E Rule Configuration	+	Rules name Rules Field	Strong Rules Temp	* Rules name :	Please input rules name		* Strong or Week : 💿 Strong 🔿 Wee	ak			
Task Query	ar offffinning (			* Rules Field :	Table Level Rules (table)	$\sim$					
				* Rules Template :	Table rows			^			
				* Compare the style	Table size changed from the	e previous cycle		^			
					Table rows unchanged from	the previous cycle					
				Description :	Table size shrinks from the Table rows decrease from the	previous cycle he previous cycle					
				[	<ul> <li>Table rows</li> </ul>						
					Table space (bytes) The difference between the	current table rows ar	nd rows of 1 day ago	. 1			
					The difference between the	current table space,	in bytes, and the table space of 1 day ago	v			

The data in this table comes from the log file that is uploaded by OSS 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:

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.

≡ → DQC Monitoring	5 Rules	Rule Template Custom Rules
88 Overview	Rules > Project Name : dqc_t > Table Name : ods_raw_log_d > Partition Expression	+ Add Monitoring Rules + Add
My Subscription	Existing Partition Expression Rule Templates (0) Custom Rules (0) Owner	
Rule Configuration	tdt=Slyyymmdd-1]     Rules name Rules Field Strong Rules Tet	Rules name : Please input rules name     Strong or Weak:      Strong     Weak
112 Inan durið		* Rules Field : Table Level Rules (table)
		Rules Template: Table rows
		Compare the style Greater Than
		: Description :
	1	
		Save
-		
0		

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 Test Run button that can be used to verify configured rules. The Test Run button can immediately trigger the validation rules for data quality.



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

_ DQC Monitoring     _ Overview	S Rules Rules > Project Name : dqc_t	> Ta	ble Name:ods_raw_log_d 〉	Partition Express	sion : dt=S[yyyyr	mmdd-1] Bind	to Schedules						
My Subscription	Existing Partition Expression	Rule Tr	Test Run					×		Test Run	Subscriptions	5 Crea	nte Rule More *
🛱 Task Query	- dt=S[yyyymmdd-1]	Rule							rror Threshold	Expected Value	Configured By	y Action	IS
		DQC	Test Run Partition	dt=S[yyyymmo						0	- 144	Chanç	e   Delete   Log
		×	Scheduling Time	: 2019-02-22 Test Run The test run is o	13:24:47	to view the result.	]	Close					
DQC Monitoring	Instance Details												
E Overview	Project :dqc_100_0000 Table Na	me :ods_raw_loş	_d > dt=S[yyyymmdd-1]		13:24:49 Mor	re							Refresh
My Subscription	Rules name Rules Field	Strong/Weak	Sampling Filter Style Conditions	Check type	Check Style	Comparison Method	Warning Threshold	Error Threshold	Expected Value	History Results	Sampling Result	Status	Actions
Rule Configuration	DQC_rule_01 ·	Strong	table_count	Numeric		Greater Than			0		570386 1	Normal	View History Results



Note:

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.

Bind to schedules

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.

Associate table rules in operation center

In Operation Center, locate the OSS\_datasync task in Cycle Task, and right-click to select Configure Quality Monitoring.

Operation Cen	ter DQC_test_0221 ~ Da	taS
≡		
(C) 0&M Overview	Search: Search by node name or II Q. Solution: Please Select V Business Flow: Please Select V Node Type: Please Select V Owner:	~
🚽 Task List	Baseline: Please Select V V My Nodes Modified Today Nodes Paused Reset Clear	
Cycle Task		
式 Manual Task	Name         Node ID           CAUTION: You are working in the production environment.	
🚽 Task O&M	oss_there 700002197850	
Cycle Instance	ods_log_info_d 700002197982	
	rpt_user_info_d 700002197984	
	dw_user_info_all_d 700002197983 workshoostart	
Testing Instance	rds700002197849	
PatchData	workshopstart 700002197848	
▶ Alarm	s show Parent Nodes	
	View Node Details	
	View Code	
	ods_kEdit Node	
	View Instances	
	View Lineage	
	Test	
	Patch Data >	
	Pause	
	More v < 1/1 >	
	Compute Quality 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].

After the configuration is completed, as shown in the figure below.

Configure Quality Monitoring			×
Current Node:oss_			
ODPS Project name: DQC_test_0221	V Table Name: ods_user_	.info_d 🗸	
Partitioning Expression: dt=\$[yyyymmd	ld-1] Add		
ODPS Project Name	Table Name	Partitioning Expression	Actions
DQC_test_0221	ods_user_info_d	dt=\$[yyyymmdd-1]	Configurations   Delete
			Close

# Click Configurations to quickly go to the rule configuration interface.

- DQC Monitoring	S Rules	S         Rules           Brules         Project Name - dec test 0221. > Table Name - ods user info d. > Partition Francescion - dts/Succement/4-11         Dates - December - dts/Succement/4-11									
믑 Overview					bind to ochedules	J					
My Subscription	Existing Partition Expression	Rule Templates ( 0 ) Custom Rules	(0) Owner:					Test Run	Subscriptions	Create Rule	More *
Task Query	- dt=S[yyyymmdd-1]	Rules name Rules Field	Strong Rules Template	Dynamic Threshold	Comparison Method	Warning Threshold	Error Threshold	Expected Value	Configured By	Actions	
					No da	a					

# Configure task subscriptions

After the associated scheduling, every time the scheduling task is run, the data quality verification 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, notifications are made based on the configured alarm policy.

Click Subscriptions to set up subscription methods. Email notifications, Email and SMS notifications are currently supported.

DQC Monitoring	Rules     Rules > Project Name : dqc_test_0221 > Table Name : ods_traw_log_d > Partition Expression : dtr\$lypyymmdd:1]     Bind to Schedules
My Subscription	Existing Partition Expression + Test Run Subscriptions Create Rule More +
Task Query	- dt-Slyyyymmdd-1] Rules name Rules Field Strong Rules Template Dynamic Comparison Warning Error Threshold Error Threshold Configured By Actions
	DQC_rule_01 Table Level Rules Strong Table rows No Greater Than 0 Change   Delete   Log
DQC Monitoring     Overview	Hules         Rules       > Project Name : dqc_test_0221 > Table Name : ods_raw_log_d > Partition Expression : dt+S(yyyymmdd+1)         Bind to Schedules
My Subscription	Existing Partition Expression Rule Templates (1) Custom Rules (0) Owner Test Run Subscriptions Create Rule More
Rule Configuration     Task Query	T         T           •dts[tyyymmdd1]         Rules name         Rules Field         Strong         Rules Template         Dynamic         Comparison         Warning         Error Threshold         Error Threshold         Value         Comparison         Value         Value         Configured By         Actions
	DOC.rul Subscriptions X - O Change   Delete   Log
	Notification Method Recipient Actions
	Email Notification > Save
	All members of this project will be shown here.
DQC Monitoring	Rules       > Project Name : dqc_test_0/221       > Table Name : ods_raw_log_d       > Partition Expression : dtr\$hyyyymmdd-1]       Bind to Schedules
My Subscription	Existing Partnion Expression Rule Templates (1) Custom Rules (0) Owner: Test Run Subscriptions Create Rule Nore •
Rule Configuration	+ Bules name Bules Field Strong Bules Template Dynamic Comparison Warning Error Traveshuld Expected Configured By Actions
🕂 Task Query	DOC Rule Subscriptions     V
	Notification Method Recipient Actions
	Email Notification ∧ Save
	Email and SMS Notification DingTalk Charbot-based Notification
	DingTalk Chatbot-based @ALL Notification

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

DQC Monitoring	My Subscriptions		Partition Expression	Owner	Actions
믬 Overview	MaxCompute $\checkmark$ Enter a table name. C	Clear	dt=\$[yyyymmdd-1]		Check Results   Notification Method  Cancel Subscription
My Subscription	MaxCompute Project	Table Name			
Rule Configuration	dqc_test_0221	ods_raw_log_d			
🛱 Task Query	DQC_test_0221	ods_user_info_d			

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.

<ul> <li>DQC Monitoring</li> </ul>	S Sector Sec									
BB Overview	Rules > Project Name : dqc_test_	es > Project Name : dqc_test_0221 > Table Name : ods_user_info_d > Partition Expression : dt=Slyyyymmdd-1] Bind to Schedules 🛆								
My Subscription	Existing Partition Expression	Rule Templates ( 0 ) Custom Rules ( 0 ) Owner :	Test Run Subscriptions Create Rule More •							
Task Query	- dt=S[yyyymmdd-1]	Rules name Rules Field Strong Rules Template Dynamic Comparison Warning Error Threshold Method Threshold	d Expected Configured By Actions Value							
		No data								

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.

S S S S S S S S S S S S S S S S S S S		Rule Template Custom Rules
Rules > Project Name : dqc_test	L0221 > Table Name : ods_user_info_d > Partition Expression	+ Add Monitoring Rules + Add
Existing Partition Expression	Rule Templates (0)         Custom Rules (0)         Owner	
- dt=S[yyyymmdd-1]	Rules name Rules Field Strong Rules Temp	Rules name: DQC_rule_02     Strong or Weak:      Strong     Weak
		Rules Field : Table Level Rules (table)
		Rules Templete : Table rows
		Compare the style     Greater Than     Fxpected Value:     0
		- Description :

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 Save button.

S Rules		Rule Template Cu	istom Rules		
Rules > Project Name : dqc_test_	0221 > Table Name : ods_user_info_d > Partition Expression	* Rules name :	DQC_rule_03	* Strong or Weak : Stron	ıg 💿 Weak
Existing Partition Expression	Rule Templates (0) Custom Rules (0) Owner :	* Rules Field :	uid (string)		
+	+		Repeated values		~
- dt=S[yyyymmdd-1]	Rules name Rules Field Strong Rules Tem;	* Compare the	Less Than $\checkmark$	* Expected Value : 1	
		style :			
		Description :			
		* Rules name :	DQC_rule_02	* Strong or Weak : 💿 Stron	ig 🔿 Weak
		* Rules Field :	Table Level Rules (table)		
		* Rules Template :	Table rows		$\sim$
		* Compare the	Greater Than $\checkmark$	* Expected Value : 0	
		style :			
		Description :			
				Sa	ve Cancel



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

Pay attention: don't forget to try Test Run -> Bind to Schedules -> Subscription.

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	S Rules	0221 \ Table Name Lode log info d	Derrition Expression 1 dr-9	issummedd 1		•					
BB Overview	Rules / Project Marie , oquitest	Jozzi / Table Name : ous_log_info_u	Partition Expression : ut-a	yyyynniod-1	Bind to Schedules	₫\$					
My Subscription	Existing Partition Expression	Rule Templates ( 0 ) Custom Rule	es ( 0 ) Owner :					Test Run	Subscriptions	Create Rule	More •
Task Query	- dt=S[yyyymmdd-1]	Rules name Rules Field	Strong Rules Template	Dynamic Threshold	Comparison Method	Warning Threshold	Error Threshold	Expected Value	Configured By	Actions	
					No dat	a					

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.

= DQC Monitoring	S S S S S S S S S S S S S S S S S S S			Rule Template Cus	tom Rules				
EE Overview	Rules > Project Name : dqc_test_	.0221 > Table Name : ods_log_info_d > Partition Expre	ession	+	Add Monitoring Rules		+	Add	
Hy Subscription	Existing Partition Expression	Rule Templates (0) Custom Rules (0) Owner	r internet						
Rule Configuration	+	Rules name Rules Field Strong Rule	s Temr	* Rules name :	Please input rules name		* Strong or Weak : 🔘	Strong Wea	ık
Task Query				* Rules Field :	Table Level Rules (table)	~			
				* Rules Template :	Table rows				~
				* Compare the style	Unequal To	$\sim$	* Expected Value : 0		
				1					
		(		Description :					
								Save	ancel

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	S Rules			
BB Overview	Rules / Project Name : oqc_test	Juzzi / Table Name : tp_usel_mito_u / Partubon expression : dt=stypygninitud=1		
Hy Subscription	Existing Partition Expression	Rule Templates ( 0 ) Custom Rules ( 0 ) Owner :	Test Run Subscriptions	Create Rule More *
Rule Configuration Task Query	+ - dt=S[yyyymmdd-1]	Rules name Rules Field Strong Rules Template Dynamic Comparison Warning Error Thr Threshold Method Threshold Error Thr	eshold Expected Configured By Act	tions
		No data		

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 style is set to field repeat values less than 1.

- DOC Monitoring	S Rules		Rule Template Custom Rules					
日日 Overview	Rules > Project Name : dqc_test	.0221 > Table Name : rpt_user_info_d > Partition Expression	+	+ Add Monitoring Rules		Add		
My Subscription	Existing Partition Expression	Rule Templates ( 0 ) Custom Rules ( 0 ) Owner :						
Rule Configuration	+							
🛱 Task Query	- dt=S[yyyymmdd-1]	Rules name Rules Field Strong Rules Tem	* Rules name :	Please input rules name	* Strong or Weak :	Strong  Weak		
			* Rules Field :	uid (string) $\checkmark$				
			* Rules Template :	Repeated values		~		
			* Compare the style	Less Than $\checkmark$	* Expected Value : 1			
			Description :					

Continue to add monitoring rules.

My Subscription       Existing Partition Expression               Hule Templates (0) Custom Rules (0) Owner : main rules Field Strong Rules Templates                Rules name	≡	Rules Rules > Project Name : dqc_test_0221 > Table Name : rpt_user_info_d > Parition Expressiv	Rule Template         Custom Rules           on         * Aud Wollioning Rules         * Aud
Tack Query      Tack Query      Tack Query      Advant (Type Index 1 and	My Subscription Rule Configuration	Existing Partition Expression +  Public Templates (0) Custom Rules (0) Owner :  Public Templates (0) Custom Rules (0) Owner :	Rules name : Please input rules name     Strong or Week : Strong @ Week     Rules Field : Table Level Rules (table) ~
	n Task Query		Rules Template : Compare the current rows returned by the SQL statement with those of 7 days ago     Compare the Absolute Value     style :     Fluctuation 0% 25% 50% 75% 100     Werning Threshold: 0 Error Threshold : 50     Description :
• Rulea name: Pease input rules name     • Strong or Weak: Strong @ Weak      Save Cance			Rules name : Please input rules name     Strong or Weak : Strong @ Weak      Save     Cancel

Note:

Here you monitor the number of table rows mainly to view the daily UV fluctuations in order to keep abreast of application dynamics.

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.

# 2 Data migration

# 2.1 Migrate data from Hadoop to MaxCompute

This topic describes how to use the data synchronization feature of DataWorks to migrate data from Hadoop to Alibaba Cloud MaxCompute.

Prepare the environment

1. Build a Hadoop cluster.

Before data migration, you must ensure that your Hadoop cluster works properly. You can use Alibaba Cloud E-MapReduce to automatically build a Hadoop cluster.

The version information of E-MapReduce Hadoop is as follows:

E-MapReduce version: EMR-3.10.1 or 3.11.0

Cluster type: Hadoop

Software(for EMR-3.11.0): HDFS2.7.2 / YARN2.7.2 / Hive2.3.3 / Ganglia3.7.2 / Spark2.2.1 / HUE4.1.0 / Zeppelin0.7.3 / Tez0.9.1 / Sqoop1.4.6 / Pig0.14.0 / ApacheDS2.0.0 / Knox0.13.0

The network type of the Hadoop cluster is classic. The region is China East 1 ( Hangzhou). The ECS compute resource of the master instance group is configured with an Internet IP address and an intranet IP address. The high availability mode is set to No (a non-HA mode). The following figure shows the configuration for EMR -3.10.1.

Software Configuration	Hardware	Configuration		Basic Co	nfigura	ation		ОК
Version Configuration								
EMR Version:	EMR-3.10.1						~	
Cluster Type:	• Hadoop	Kafka						
Required Services:	ApacheDS (2.0.0)	Knox (0.13.0)	Hadoop YAR	N (2.7.2)	Hado	op HDFS (2.7.2		
	Ganglia (3.7.2)	Zepplin (0.7.1)	HUE (4.1.0)	Sqoop (1.4	4.6)	Tez (0.9.1)	Pig (0.14.0)	
	Spark (2.2.1) H	ive (2.3.2)						
Optional Services:	Ranger (0.7.1)	Flink (1.4.0)	npala (2.10.0)	HAS (1.1.:	L)	Phoenix (4.10.0		
	Zookeeper (3.4.11)	Oozie (4.2.0)	Storm (1.1.2)	Presto	o (0.188	B) HBase (1	.1.1)	
	Click to Choose							
High Security Mode: 👔								
Enable Custom Setting: 👔								

#### 2. MaxCompute

For more information, see Activate MaxCompute.

Activate MaxCompute and create a project. In this topic, create a project named bigdata\_DOC in China East 1 (Hangzhou) and enable the related DataWorks services for this project.

#### **Prepare data**

1. Create test data on the Hadoop cluster.

In the E-MapReduce console, go to the Hadoop cluster page and use Notebook to create a notebook task. The table creation Hive statements in this example are as follows:

```
CREATE TABLE IF NOT
EXISTS hive_doc_g ood_sale (
create_tim e timestamp,
category STRING,
brand STRING,
buyer_id STRING,
trans_num BIGINT,
```

```
trans_amou nt
                 DOUBLE ,
click_cnt
            BIGINT
)
PARTITIONE D
                ΒY
                    (pt
                          string) ROW
                                          FORMAT
           FIELDS
                    TERMINATED
                                 BY ','
                                         lines
                                                 terminated
                                                              by
DELIMITED
 '\ n '
```

Click run. The test table hive\_doc\_g ood\_sale is then successfully created on the E-MapReduce Hadoop cluster.

Insert the test data. You can select data from OSS or other data sources, or manually insert a small amount of test data. The following data can be manually inserted:

```
insert into
hive_doc_g ood_sale PARTITION ( pt = 1 ) values (' 2018 - 08
- 21 ',' Coat ',' Brand A ',' lilei ', 3 , 500 . 6 , 7 ),(' 2018
- 08 - 22 ',' Fresh food ',' Brand B ',' lilei ', 1 , 303 , 8
),(' 2018 - 08 - 22 ',' Coat ',' Brand C ',' hanmeimei ', 2 , 510
, 2 ),( 2018 - 08 - 22 ,' Toiletries ',' Brand A ',' hanmeimei ',
1 , 442 . 5 , 1 ),(' 2018 - 08 - 22 ',' Fresh food ',' Brand D
',' hanmeimei ', 2 , 234 , 3 ),(' 2018 - 08 - 23 ',' Coat ',' Brand
B ',' jimmy ', 9 , 2000 , 7 ),(' 2018 - 08 - 23 ',' Fresh food
',' Brand A ',' jimmy ', 5 , 45 . 1 , 5 ),(' 2018 - 08 - 23 ','
Coat ',' Brand E ',' jimmy ', 5 , 100 . 2 , 4 ),(' 2018 - 08 - 24
',' Fresh food ',' Brand G ',' peiqi ', 10 , 5560 , 7 ),(' 2018
- 08 - 24 ',' Sanitary ware ',' Brand F ',' peiqi ', 1 , 445 .
6 , 2 ),(' 2018 - 08 - 24 ',' Coat ',' Brand A ',' ray ', 3 , 777
, 3 ),(' 2018 - 08 - 24 ',' Sanitary ware ',' Brand G ',' ray
', 3 , 122 , 3 ),(' 2018 - 08 - 24 ',' Coat ',' Brand C ',' ray
', 1 , 62 , 7 );
```

After inserting the data, you can use the select \* from hive\_doc\_g ood\_sale where pt = 1 ; statement to check whether the data exists in the Hadoop cluster table for migration.

2. Use DataWorks to create a destination table.

In the DataWorks console, click the MaxCompute project, and choose Data Development > New > Create Table.

In the displayed window, enter the following table creation SQL statements:

```
CREATE
        TABLE
                    NOT
                          EXISTS
                                   hive_doc_g ood_sale (
                TF
  create_tim e string,
  category STRING,
          STRING
  brand
  buyer_id
           STRING
  trans_num BIGINT
                   DOUBLE ,
  trans_amou nt
  click_cnt
              BIGINT
 )
```

```
PARTITIONE D BY ( pt string );
```

When creating a table, pay attention to the mapping between the Hive data type and the MaxCompute data type.

The data synchronization feature of DataWorks does not support timestamp data. Therefore, in the DataWorks table creation statements, create\_time is set to a string value. You can also use the odpscmd Command Line (CLI) tool to create a table. For more information about how to install and configure the tool, see *Install* and configure a client. The table creation process is as follows.

```
odps@ bigdata_DOC>CREATE TABLE IF NOT EXISTS hive_doc_good_sale(
   create_time timestamp,
   category STRING,
   brand STRING,
   buyer_id STRING,
   trans_num BIGINT.
   trans_amount DOUBLE,
   click_cnt BIGINT
   PARTITIONED BY (pt string) ;
                  >
                                     >
                                                       >
         >
                            >
                                               >
ID = 20180906110540873gev1bpim
ΟK
odps@ bigdata_DOC>drop table hive_doc_good_sale;
Confirm to "drop table hive_doc_good_sale;" (yes/no)? yes
ID = 20180906110825180gxh66292
```

# Note:

Considering the compatibility of the Hive and MaxCompute data types, we recommend that you run the following command on the odpscmd client:

set odps . sql . type . system . odps2 = true ; set

```
odps . sql . hive . compatible = true ;
```

After the table is created, you can choose Data Development > Table Query in the DataWorks console to view the table created in MaxCompute, as shown in the following figure.



#### Synchronize data

1. Create a custom resource group.

In most cases, the network between the project data node of MaxCompute and the data node of the Hadoop cluster is not connected. You can customize a resource group to execute the synchronization task of DataWorks on the master node of the Hadoop cluster. (In general, the network between the master node and the data node on the Hadoop cluster is connected).

a. View the data node of the Hadoop cluster.

On the home page of the E-MapReduce console, choose Cluster Management > Cluster > Hosts. You can view the data node of the Hadoop cluster. As shown in the following figure, the host name of the master node on the E-MapReduce Hadoop cluster (non-HA mode) is emr-header-1, and the host name of the data node is emr-worker-X.

E-MapReduce	Overview Cluster	Management	Data Platform	New Alert	Operation	Logs He
8 Cluster Overview	Home Page > Cluster Ma	inagement > C	luster ( C-D3706C57	2667999E) > ⊢	losts	
Clusters and Serv	Hosts					
😚 Hosts	ECS InstanceID	E	nter a hostname		Enter an internal	IP address.
Cluster Scripts	ECS ID	Ho	stname	IP Information		Role
$\%$ Access Links and $\cdots$						
A User Management	i e se s	r 🗗 en	nr-worker-2	Intranet IP:192.16	58.1.154	CORE
🖒 Scaling						
	i	2 en	nr-header-1	Intranet IP:192.16	58.1.152	MASTER
				Public Network:		
	i	<b>d'</b> en	nr-worker-1	Intranet IP:19216	58.1.153	CORF

You can also click the ECS ID of the master node, click Connect on the displayed ECS details page, and run the hadoop dfsadmin –report command to view the data node, as shown in the following figure.

DFS Used:: 0.05% Under replicated blocks: 0 Blocks with corrupt replicas: 0 Missing blocks: 0 Missing blocks (with replication factor 1): 0 Live datanodes (2): Name: 10.31.122.189:50010 (emr-worker-1.cluster-74503) Hostname: emr-worker-1.cluster-74503 Decommission Status : Normal Configured Capacity: 333373341696 (310.48 GB) DFS Used: 155725824 (148.51 MB) Non DFS Used: 325541888 (310.46 MB) DFS Remaining: 332892073984 (310.03 GB) DFS Used: 0.05% DFS Remaining%: 99.86% Configured Cache Capacity: 0 (0 B) Cache Used: 0 (0 B) Cache Remaining: 0 (0 B) Cache Used%: 100.00% Cache Remaining%: 0.00% Xceivers: 1 Last contact: Thu Sep 06 19:41:01 CST 2018 Name: 10.81.78.209:50010 (emr-worker-2.cluster-74503) Hostname: emr-worker-2.cluster-74503 Decommission Status : Normal Configured Capacity: 333373341696 (310.48 GB) DFS Used: 155725824 (148.51 MB) Non DFS Used: 325451776 (310.38 MB) DFS Remaining: 332892164096 (310.03 GB) DFS Used:: 0.05% DFS Remaining%: 99.86% Configured Cache Capacity: 0 (0 B) Cache Used: 0 (0 B) Cache Remaining: 0 (0 B) Cache Used%: 100.00% Cache Remaining%: 0.00% kceivers: 1 Last contact: Thu Sep 06 19:41:02 CST 2018

As shown in the preceding figure, the data node has only an intranet address and cannot communicate with the default resource group of DataWorks. Therefore, you need to customize a resource group and set the master node to a node that executes the synchronization task of DataWorks.

b. Create a custom resource group.

In the DataWorks console, go to the Data Integration page, select Resource Group, and click New Resource Groups, as shown in the following figure.

resources.

 Outstanding
 Operation Mode
 Data Hindgraded
 Data

For more information about how to customize a resource group, see Add task

When you add a server, you need to enter the information such as the ECS UUID and machine IP address. (For a classic network, enter the sever name. For a VPC network, enter the server UUID. You can add scheduling resources only for classic networks in China East 2 (Shanghai) in DataWorks V2.0. In other regions, select the VPC network type when you add a scheduling resource group regardless of whether your network type is classic or VPC.) Set the machine IP address to the Internet IP address of the master node because the intranet IP address may be inaccessible. The ECS UUID must be connected to the master node management terminal, which can be obtained by running the dmidecode | grep UUID command, as shown in the following figure. (You can use the same method if your Hadoop cluster is not built on E-MapReduce.)

```
[root@emr-header-1 logs]# dmidecode ¦ grep UUID
UUID: F631D86C-
```

After adding the server, make sure that the network between the master node and DataWorks is connected. If you are using an ECS server, you must set the server security group. If you are using an intranet IP address for communication, set the server security group. For more information, see *Adding security groups*.

If you are using an Internet IP address, you can directly set the Internet ingress and egress under Security Group Rules(In practical application scenarios, we recommend that you set detailed bypass rules for your data security.)

After completing the preceding steps, install the custom resource group agent as prompted. If the state is available, the custom resource group is added successfully.

If the state is unavailable, you can log on to the master node, and run the tail - f / home / admin / alisataskn ode / logs / heartbeat . log command to check whether the heartbeat message between DataWorks and the master node has timed out, as shown in the following figure.

2. Create a data source.

For more information about how to create a data source in DataWorks, see *Configuring Data Source*.

After you create a project in DataWorks, the data source is set to odps\_first by default. Therefore, you only need to add a Hadoop cluster data source. To do so, perform the following steps: On the Data Integration page of DataWorks, choose Data Source > New Source, and select HDFS.

In the displayed window, enter the data source name and defaultFS. If the E-MapReduce Hadoop cluster is an HA cluster, the address is IP:8020 of hdfs://emr -header-1. If the E-MapReduce Hadoop cluster is a non-HA cluster, the address is IP:9000 of hdfs://emr-header-1. In this topic, emr-header-1 is connected to DataWorks through the Internet. Therefore, enter the Internet IP address and open the security group.

C DataWorks	bigdata_DOC	→ Data Inte	egration Data Deve	lopment	Data Management	Operation Center	Projec	t Management	Data Service	dtplus_docs	→ Eng	glish <del>-</del>
=	Type: All	V Name	New HDFS Data Sources	s				×			New Sou	rce
<ul> <li>Overview</li> </ul>	Type:											معت
Consume	Name	Туре	* Name	HDFS1					Description		Act	tion
<ul> <li>Project Space</li> </ul>	odps first	ODPS	Description						connection from odps calc engine 1			
<ul> <li>Project Overview</li> </ul>			* defaultFS :		fs://Serverip: port			0	144			
Consume	odps_es	ODPS	Test Connectivity	Test Conne	ectivity				test		Edit Del	lete
<ul> <li>Offline Sync</li> </ul>	HDFS_data_source	HDFS							Elasticsearch 测试		Edit Del	lete
8 Sync Tasks	HDFS1	HDFS				Pr	revious				Edit Del	lete
Data Sources										Previous	1 Next	t >

After the configuration is completed, click Test Connectivity. If Test connectivity successfully is displayed, the data source is added successfully.



If the network type of the E-MapReduce Hadoop cluster is VPC, the connectivity test is not supported.

3. Configure the data synchronization task.

On the Data Integration page of DataWorks, click Sync Tasks and create a script mode. In the displayed window, select a data source, as shown in the following figure.

=							
- Overview	✓ Image: Control C	Untitled-0 × @E	s_nais_test ×				
Consume	> 📔 HDFS2MC						
	ES_hdfs_test i'm locking 20						
<ul> <li>Project Space</li> </ul>	• 🕢 ES_test_pro i'm locking 201						
<ul> <li>Project Overview</li> </ul>	• 🕢 es_test_pro02 i'm locking 2						
Consume	Mdfs2mc1 i'm locking 2018-		import tomplate			X	
E consume	• 🕢 hdfs2mc2 i'm locking 2018-		import template				
	• 🕢 odps2hadoop i'm locking 20		* Source type :	Hdfs	~ 🕐		
Sync Tasks	• 🕢 odps_es_test i'm locking 20						
Data Sources			* data sources .	additional data sources	~		
B Data Sources							Script Model
- Resource			Type of objective :	ODPS	~ (?)	1	
தீ Resource			* data sources :	odps_first (odps)	$\sim$	c a	atile and efficient
				additional data sources		đ	able depth optimization,
						ľ	ort all data sources
					confirmation Ca	ancel	

After the template is imported, the synchronization task is converted to the script mode. For more information, see *Script Mode*.

When you configure the data synchronization task script, the data types of the DataWorks synchronization task and the Hive table are as follows.

Data type in the Hivetable	Data type in DataX / DataWorks

TINYINT,SMALLINT,INT,BIGINT	Long
FLOAT,DOUBLE,DECIMAL	Double
String,CHAR,VARCHAR	String
BOOLEAN	Boolean
Date,TIMESTAMP	Date
Binary	Binary

The code details are as follows:

```
{
           " configurat ion ": {
    " reader ": {
        " plugin ": " hdfs ",
                                       " parameter ": {
    " path ": "/ user / hive / warehouse / hive_doc_g ood_sale
      /",
                                                      " datasource ": " HDFS1 ",
                                                      " column ": [
                                                                   {
                                                                               " index ": 0 ,
" type ": " string "
                                                                  },
                                                                  {
                                                                               " index ": 1 ,
" type ": " string "
                                                                  },
{
                                                                               " index ": 2 ,
" type ": " string "
                                                                  },
{
                                                                               " index ": 3 ,
" type ": " string "
                                                                  },
{
                                                                              " index ": 4 ,
" type ": " long "
                                                                  },
{
                                                                             " index ": 5 ,
" type ": " double "
                                                                   },
                                                                   {
                                                                             " index ": 6 ,
" type ": " long "
                                                                   }
                                                   ],

" defaultFS ": " hdfs :// 121 . 199 . 11 . 138 : 9000 ",

" fieldDelim iter ": ",",

" encoding ": " UTF - 8 ",

" fileType ": " text "
                                       }
                        },
" writer ": {
    " plugin ": " odps ",
    " plugin ": {
    " plugin ": {
    " plugin ": {
    " plugin ": {
    " plugin ": " odps ";
    " plugin ": " plugin ": " odps ";
    " plugin ": " od
                                                       " partition ": " pt = 1 ",
```

```
" truncate ": false ,
       " datasource ": " odps_first ",
       " column ": [
         " create_tim e ",
         " category ",
         ....
           brand "
         н
           buyer_id "
         11
           trans_num "
         " trans_amou nt ",
         " click_cnt "
       ],
" table ": " hive_doc_g ood_sale "
    }
  },
    setting ": {
" errorLimit ": {
       " record ": " 1000 "
    },
       speed ": {
       " throttle ": false ,
         concurrent ": 1 ,
       " mbps ": " 1 ",
       " dmu ":
                  1
    }
  }
},
" type ": " job ",
" version ": " 1 . 0 "
```

The path parameter indicates the place where the data is stored in the Hadoop cluster. You can log on to the master node and run the hdfs dfs - ls / user / hive / warehouse / hive\_doc\_g ood\_sale command to confirm the place. For a partition table, you do not need to specify the partitions. The data synchronization feature of DataWorks can automatically recurse to the partition path, as shown in the following figure.

After the configuration is completed, click Run. If a message is displayed indicating that the task is executed successfully, the synchronization task is completed. If a message is displayed indicating that the task failed to be executed, copy the logs for further troubleshooting.

Verify the results

}

In the DataWorks console, choose Data Development > Table Query and select the hive\_doc\_g ood\_sal e table. You can check whether the Hive data has been synchronized to MaxCompute. You can also create a table query task, enter the
select \* FROM hive\_doc\_g ood\_sale where pt = 1 ; script in the task, and click Run to query the results.

You can also enter select \* FROM hive\_doc\_g ood\_sale where pt = 1; in the odpscmd CLI tool to query the table results.

#### Migrate data from MaxCompute to Hadoop

To migrate data from MaxCompute to Hadoop, perform the preceding steps but exchange the reader and writer objects in the synchronization script. The following is an example:

```
{
  ...
    configurat ion ": {
    reader ": {
      " plugin ": " odps ",
       " parameter ": {
      " partition ": " pt = 1 ",
" isCompress ": false ,
       " datasource ": " odps_first ",
" column ": [
         " create_tim e ",
         " category ",
         " brand "
       " buyer_id "
       " trans_num ",
" trans_amou nt ",
       " click_cnt "
    ],
" table ": " hive_doc_g ood_sale "
    }
  " plugin ": " hdfs ",
    " parameter ": {
    " path ": "/ user / hive / warehouse / hive_doc_g ood_sale ",
    " fileName ": " pt = 1 ",
" datasource ": " HDFS_data_ source ",
    " column ": [
       Ł
         " name ": " create_tim e ",
         " type ": " string "
       },
       {
         " name ": " category ",
         " type ": " string "
      },
       {
         " name ": " brand "
         " type ": " string "
      },
       {
         " name ": " buyer_id ",
         " type ": " string "
      },
       {
         " name ": " trans_num ",
```

```
" type ": " BIGINT "
        },
         {
           " name ": " trans_amou nt ",
           " type ": " DOUBLE "
        },
         {
           " name ": " click_cnt ",
           " type ": " BIGINT "
        }
     ],
     " defaultFS ": " hdfs :// 47 . 99 . 162 . 100 : 9000 ",
" writeMode ": " append ",
     " fieldDelim iter ": ",",
" encoding ": " UTF - 8 ",
" fileType ": " text "
      }
   },
     setting ": {
" errorLimit ": {
        " record ": " 1000 "
      speed ": {
      " throttle ": false ,
      " concurrent ": 1,
     " mbps ": " 1 ",
" dmu ": 1
   }
   }
},
" type ": " job ",
" version ": " 1 . 0 "
```

Before executing the preceding synchronization task, you must set the Hadoop cluster. For more information, see *Configure HDFS Writer*. After executing the synchronization task, you need to manually copy the synchronized files.

## 2.2 Migrate JSON data from OSS to MaxCompute

This topic describes how to use the data integration feature of DataWorks to migrate JSON data from OSS to MaxCompute and use the built-in string function GET\_JSON\_OBJECT of MaxCompute to extract JSON information.

Preparations

• Upload data to OSS.

Convert your JSON file to a TXT file and upload it to OSS. The following is a JSON file example:

```
{
                                 " category ": " reference ",
" author ": " Nigel Rees ",
" title ": " Sayings of the
                                                                                                         Century ",
                                 " price ": 8 . 95
                           },
{
                                 " category ": " fiction ",
" author ": " Evelyn Waugh ",
" title ": " Sword of Honour ",
                                 " price ": 12 . 99
                           },
{
                                   " category ": " fiction ",
" author ": " J . R . R . Tolkien ",
" title ": " The Lord of the Ri
" isbn ": " 0 - 395 - 19395 - 8 ",
" price ": 22 . 99
                                                                                                                Rings ",
                           }
                    ],
" bicycle ": {
" color ": " red ",
" color ": 19 . 95
                             " price ": 19 . 95
                     }
        },
" expensive ": 10
}
```

Upload the *applog*. *txt* file to OSS. In this example, the OSS bucket is located in China (Shanghai).

#### Use DataWorks to migrate JSON data from OSS to MaxCompute

• 1. Add an OSS data source.

In the DataWorks console, go to the *Data Integration* page and add an OSS data source. For more information, see *Configure OSS data source*.

Data Integrat	tion DTplus_DOC	>								٩,	quwenjie English
≡ ✔ Overview	Data Source	Data Source Type :	Add Data Source		. <u> </u>			×		C Refresh	Add Data Source
👑 Tasks	Data Source Nam	e Data Sou	MySQL	SQL Server	PostgreSQL	Oracle	DM		Status	Connected At	Actions
Monitoring			00	$\overline{\mathbf{v}}$	میگر ا	$\otimes$					
🚽 Sync Resources	odps_first	ODPS	DRDS	POLARDB	HybridDB for MySQL	HybridDB for		- 1			
🛧 Data Source			Big data storage			PostgrebyL		- 1			
📦 Resource Group			MaxCompute (ODPS)	<b>%</b>	AnalyticDB (ADS)	45					
			Semi-structuredstorage	e HDFS	FTP	Lynning					
			NoSQL	Memcache (OCS)	Redis	Table Store (OTS)					
			Message queue								

The parameters are shown in the following figure. Click Complete after the connectivity test is successful. The endpoints in this topic include http :// oss

```
- cn - shanghai . aliyuncs . com and http :// oss - cn - shanghai -
internal . aliyuncs . com .
```



Because the OSS and DataWorks projects are located in the same region, the intranet endpoint http://oss - cn - shanghai - internal . aliyuncs . com is used.

Add Data Source OSS		×
* Data Source Name :	OSS_userlog	
Description :		
* Endpoint :	http://oss-cn-shanghai-internat.aliyuncs.com	0
* Bucket :	dcogood2	0
* AccessKey ID :	• <sup>1</sup> C.	0
* AccessKey Secret :		
Test Connectivity:	Test Connectivity	
	Previous	Complete

• 2. Create a data synchronization task.

In the DataWorks console, create a data synchronization node. For more information, see *Configure OSS Reader*. At the same time, create a table named

mqdata in DataWorks to store the JSON data. For more information, see Create a

table.

	Data Analytics 🖉 🛱 📮 Ċ 🕀 🗄	Ja AliSpark-2.x-quicks	start-1 × 🏢 mqdata 🛛 🗙 Di json2ma	x × 😝 Data
())				
*				
Q	✓ Business Flow			
Ŀ	<ul> <li>✓ ♣ works</li> <li>&gt; ➡ Data Integration</li> </ul>			
	> 🕢 Data Analytics			(?) *
≕	<ul><li>✓ Ⅲ Table</li><li>&gt; Ø Resource</li></ul>			
=0	> 🛃 Function	02 Mappings	Source Table	
fx		Create Table		×
	> 🞯 Control	Database Type :	• MaxCompute	
Σ				
亩		Table Name :	mqdata	
			Submit	Cancel
				d

You can set the table parameters on the graphical interface. The mqdata table has only one column, which is named MQ data. The data type is string.

Table Alias :	MQ data store					
Level 1 Topic :	Select ~	Level 2 Topic :	Select		C	
Description :						
Physical Model						
Partition :	O Partitioned Table 💿 Non- Partitioned Table	Life Cycle :				
Table Level :	Select ~	Table Category :	Select		C	
Table type :						
Table Structure						
Add Field Move Up						
Field English Name	Field Alias Field Type		Length/Set	Description	Primary Key 🕜	
mqdata	string		string		No	

• 3. Set the parameters.

After creating a table, you can set the data synchronization task parameters on the graphical interface, as shown in the following figure. First, set the destination data

source to odps\_first and the destination table to mqdata. Then, set the original data source to OSS and enter the file path and name as the object prefix.

01 Data Source		ource		Destination		
	The data sou	rces can be default data sourc	es or data sources created by you. Click here	to check the supported data sou	irce types.	
* Data Source :	OSS v	OSS_userlog ~	Data Source :	ODPS ~	odps_first	
* Object Prefix :	applog.txt		* Table :	mqdata		
	Add +					
* File Type :	csv		Partition :	None		
* Column Separator :			Clearance Rule :	Clear Existing Data Before Wri	iting (Insert Overwrite) 🛛 🗸	
Encoding :	UTF-8		Compression :	💿 Disable 🔵 Enable		
Null String :			Consider Empty String as Null	🔵 Yes 💿 No		
* Compression :	None					
Format						
* Include Header:	No					

## Note:

You can set the column delimiter to caret (^) or any other character that is not contained in the TXT file. DataWorks supports multiple column delimiters for the TXT data sources in OSS. Therefore, you can use characters such as %&%#^\$\$^% to separate the data into a column.

Select Enable Same Line Mapping.

02 Mapping		Source Table			Destinatio	n Table		
	Location/Value	Туре	Ø	$\bigcirc$		Field	Туре	Map of the same name
	Column 0	string	•	•	P	mqdata	STRING	Enable Same-Line Mapping
	Column 1	string						
	Column 2	string						
	Column 3	string						
	Column 4	string						

Click the script switching button in the upper-left corner to switch to the script mode. Set fileFormat to " fileFormat ":" binary ". The following is an example of the code in script mode:

```
" nullFormat ": "",
                " compress ": "",
                " datasource ": " OSS_userlo g ",
                " column ": [
                      {
                           " name ": 0 ,
" type ": " string ",
" index ": 0
                      }
                ],

" skipHeader ": " false ",

" encoding ": " UTF - 8 ",

" fieldDelim iter ": "^",

" fileFormat ": " binary ",
                " object ": [
                      " applog . txt "
                ٦
           },
" name ": " Reader ",
" reader
           " category ": " reader "
     },
{
           " stepType ": " odps ",
           " parameter ": {
    " partition ": "",
    " isCompress ": false,
    " truncate ": true,
                " datasource ": " odps_first ",
                " column ": [
" mqdata "
                ],
"emptyAsNul l": false,
                " table ": " mqdata "
           },
" name ": " Writer ",
" write
           " category ": " writer "
     }
],
"version ": " 2 . 0 ",
" order ": {
     " hops ": [
           {
                " from ": " Reader ",
                " to ": " Writer "
           }
     ]
" errorLimit ": {
" record ": ""
     " concurrent ": 2,
           " throttle ": false ,
           " dmu ": 1
     }
}
 Note:
```

}

In this step, after the JSON file is synchronized from OSS to MaxCompute, data in the file is saved in the same row. That is, data in the JSON file shares the same field. You can use the default values for other parameters.

After completing the preceding settings, click run.

#### Verify the result

1. Create an ODPS SQL node in your Business Flow.

➤ Business Flow			
🗸 🛃 test		* Include Header:	No
> 😑 Data	Integration		
✓ 400 F	Create Data DevelopmentNo	ode ID > ODPS SQL	

2. Enter the SELECT \* from mqdata ; statement to view the data in the mqdata table.

Note:					
You can also run the	SELECT	*	from	mqdata	; command on the MaxCompute
client to view the data	and perfo	orm	subsequ	ient steps	3.

3. Verify that the data imported to the table is correct and use SELECT

GET\_JSON\_O BJECT ( mqdata . MQdata ,'\$. expensive ') FROM mqdata

; to obtain the value of expensive in the JSON file.



Additional information

To verify the result, you can also use the built-in string function *GET\_JSON\_OBJECT* in MaxCompute to obtain the JSON data as needed.

## 2.3 Migrate JSON data from MongoDB to MaxCompute

This topic describes how to use the data integration feature of DataWorks to extract JSON fields from MongoDB to MaxCompute.

#### Preparations

1. Prepare an account.

Create a user in the database in advance to add data sources in DataWorks. In this example, you can run the db . createUser ({ user :" bookuser ", pwd :"

123456 ", roles :[" root "]}) command to create a user named bookuser. The password of the user is 123456, and the permission is root.

2. Prepare data.

Upload data to your MongoDB. In this example, Alibaba Cloud ApsaraDB for MongoDB is used. The network type is VPC. (An Internet IP address is required for MongoDB to communicate with the default resource group of DataWorks.) The test data is as follows:

```
{
      " store ": {
             " book ": [
                    {
                         " category ": " reference ",
" author ": " Nigel Rees "
" title ": " Sayings of
                                                            of the
                                                                               Century ",
                         " price ": 8.95
                    },
{
                         " category ": " fiction ",
" author ": " Evelyn Waugh ",
                         " title ": " Sword
" price ": 12 . 99
                                                         of Honour ",
                    },
{
                          " category ": " fiction ",
" author ": " J . R . R . Tolkien ",
" title ": " The Lord of the R
                                                                                     Rings ",
                          " isbn ": " 0 - 395 - 19395 - 8 ",
" price ": 22 . 99
                    }
               ],
" bicycle ": {
                      " color ": " red ".
                      " price ": 19 . 95
                }
      },
" expensive ":
                               10
}
```

Log on to the DMS console of MongoDB. In this example, the database name is admin and the collection is userlog. You can run the db . userlog . find (). limit ( 10 ) command in the query window to view the uploaded data.

#### Use DataWorks to extract data to MaxCompute

• 1. Add a MongoDB data source.

In the DataWorks console, go to the *Data Integration* page and add a *MongoDB* data source.

Data Integration	n DTplus_DOC 💎 🗸	4 quwe
≡ ✓ Overview ﷺ Tasks	MySQL     SQL Server     PostgreSQL     Image: Sql Server       MySQL     SQL Server     PostgreSQL     Oracle	C Refresh Add Data Source
Monitoring Sync Resources Data Source	Deti Nan DRDS POLARDB POLARDB for MySQL HybridDB for PostgreSQL	Status Conne At
📦 Resource Group	odp     Big data storage       Image: MaxCompute (ODPS)     Image: Datahub         AnalyticDB (ADS)     Lightning	
	Semi-structuredstorage	
	NoSQL MongoDB Memcache (OCS) Redis Table Store (OTS)	
	Message queue	

The parameters are shown in the following figure. Click Complete after the connectivity test is successful. In this example, the network type of MongoDB is VPC. Therefore, the Data Source Type must be set to Public IP Address Available.

	Public IP Address Available 🗸 🗸
ata Source Name :	mongodb_userlog
Description :	
* Address :	dds-uf -pub.mongodb.rds.aliyuncs.com:3717
	Add Address
* Database Name :	admin
* Username :	bookuser
* Password :	
Test Connectivity :	Test Connectivity
	For MongoDB data sources:
0	Date Integration only supports longer to your MangeDR replice set using the
0	bata integration only supports logon to your wongood replica set using the
0	corresponding account.

To obtain the IP address and the port number, log on to the and click the target instance. Example parameters are shown in the following figure.

• 2. Create a data synchronization task.

In the DataWorks console, create a data synchronization node. For more information, see *Configure OSS Reader*.

Datal	DataStudio DTplus_DOC 💎		
Ш		📑 Data Integration 🗴 🜌 Resource 🛛 🗙	
(/)			
÷.			
Q	▶ Business Flow		
0			
0	Y 📑 Data Integration		
×	> VI> Data Analytics		
=	> III Table		
=0	> fx Function		
£.	> 🚼 Algorithm	Create Node	×
	> 🞯 Control		
		Node Type: Data Sync	
Σ			
亩		Node Name : jeon2max	
		Destination Folder: Business Flow/works/Data Integration ~	
		Submit	ancel

At the same time, create a table named mqdata in DataWorks to store JSON data. For more information, see *Create a table*.

✓ Business Flow	Data Source		300				
🗸 📥 works							
Y 📑 Data Integration							
• Di json2max Loc							
> 🕢 Data Analytics							
🗸 🥅 Table							
> 🧭 Resource	02 Mappings						
> 🛃 Function							
> 🚼 Algorithm				Create Table			×
> 🞯 Control							
				Database Type :	<ul> <li>MaxCompute</li> </ul>		
	03 Channel			Table Name :	mqdata		
						Submit	

You can set the table parameters on the graphical interface. The mqdata table has only one column, which is named MQ data. The data type is string.

Table Alias :	MQ data store				
Level 1 Topic :	Select ~	Level 2 Topic : Select		C	
Description :					
Physical Model					
Partition :	Partitioned Table      Non- Partitioned Table	Life Cycle :			
Table Level :	Select ~	Table Category : Select		C	
Table type :					
Table Structure					
Add Field Move Up					
Field English Name	Field Alias Field Type	Length/Set	Description	Primary Key ⑦	
mqdata	string	string		No	

· 3. Set the parameters.

After creating a table, you can set the data synchronization task parameters on the graphical interface. First, set the destination data source to odps\_first and the destination table to mqdata. Then, set the original data source to MongoDB and select mongodb\_userlog. After completing the preceding settings, click Switch to script mode. The following is an example of the code in script mode:

```
{
    " type ": " job ",
    " steps ": [
        {
            " stepType ": " mongodb ",
            .....
              parameter ": {
                " datasource ": " mongodb_us erlog ",
// Data
                    name
           source
                  column ": [
                11
                    {
                         " name ": " store . bicycle . color ", //
JSON
        field
                path .
                         In this example, the
                                                     value
                                                               of
color
         is
              extracted
                         " type ": " document . document . string
" // The
                          fields in this
            number
                     of
                                                 line
                                                        must be
                 that in
the same as
                                      preceding
                              the
                                                   line ( the
                                                                 name
                       JSON
                               field
                                                 level - 1
   line ). If
                                                             field
                the
                                       is
                                           а
                           expensive
          example ,
                                        field
   for
                    the
                                                 in
                                                      this
                                                             topic ,
enter
         the
               string .
                ],
" collection Name // Collection
                                                      name ": "
userlog "
            },
" name ": " Reader ",
                             "." reader
            " category ": " reader "
        },
{
```

```
" stepType ": " odps ",
               parameter ": {
             ...
                 " partition ": "",
" isCompress ": false ,
                 " truncate ": true
                 " datasource ": " odps_first ",
                   column ": [
                 ....
                            " mqdata " // Table
                                                   column
                                                                     in
                                                             name
 MaxCompute
                 ],
"emptyAsNul l": false,
                 " table ": " mqdata "
             },
" name ": " Writer ",
" write
             " category ": " writer "
        }
    ],
"version ": " 2 . 0 ",
     order ":
                {
        " hops ":
                   Γ
             {
                 " from ": " Reader ",
                 " to ": " Writer "
             }
        ]
    " errorLimit ": {
            " record ": ""
        },
" speed ": {
             " concurrent ": 2,
             " throttle ": false ,
             " dmu ": 1
        }
    }
}
```

After completing the preceding settings, click Run. If the following information is displayed, the code has run successfully.

#### Verify the result





Enter the SELECT \* from mqdata ; statement to view the data in the mqdata table. You can also run the SELECT \* from mqdata ; command on the *MaxCompute client* to view the data.

# 3 Data development

## 3.1 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).

Sq test_sqL01 🌒 🚣 test 🛛 🗙									> ≡
" ", F & - • • §									0&M
<pre>1odps sql 2</pre>	Dependencies (?) was Parse : (?) Yes (?) No Parse 1/0 Jostream Node Enter an output name or output table name v + Use The Workspace Root Node							Schedule Relation	
	Upstream Node Output Name project_b_name.pm_table_b						Source Auto Parse		ship Version
	Output Enter an output name								Struct
			Downstrea	am Node Name					ure
	MaxCompute_DOC.500143227_out	- @					Added by Default		
	MaxCompute_DOC.test_sql_01 @	- @					Added Manually		
	MaxCompute_DOC.pm_table_a @	MaxCompute_DOC.pm_table_a	-		•	-	Auto Parse	Û	



#### 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.

Sq Tes	Task_1 x 🖂 project_b_neme pm_table. x 🔄 test_us[.01 🗣 驫 test x											
凹												
1Opto:sql 2												
5												
6 7 8	INSERT OVERNRITE TABLE pm_table_a SELECT * HexCompute_DOC_root - EDRM				maxcompute_doc_root	700000822799	diplot, dana	Added Manually				
9	8 FROM project_b_name.pm_table_b 9 ;	Output test										
	MaxCompute_DOC.500143227_out								Added by Default			
			MaxCompute_DOC.test_sql_01		- 0	-				Added Manually	Û	icture

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 🗙 🐻 test_aqL01 🌑 🚜	🛛 project/unite privités. x 🖂 ten x									
" " h & ÷ • : •										
1odps sql author:t[mm 4create time:2018-12-27 10:25:30 	Auto Parse :         • Yes         No         Parse IO           Upstream Node         Enter an output name or output table name								Schedule Rela	
8 FROM project_b_name.pm_table_b 9 ;										
	project_b_name.pm_table_b						Auto Parse			
	Output test				Invalid upstream of	depender				
									Structu	
	MexCompute_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	i.	
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 : • • 5 • 5									08M
1odps sql 2 3author:15mm 4create time:2018-12-27 10:25:30	Contraction Contra	3 After you find	the out	put name, er	nter it here.				
5 6 INSERT OVERWRITE TABLE pm_table_a 7 SELECT * 8 SEROM _ product b pame pm_table b	Upstream Node Enter an output name or output	ut table name Y + Use The We							Relatio
9 ;	Upstream Node Output Name								
	project_b_name.pm_table_b						Auto Parse		Ver
	Find the output task for the	he table and view the o	utput na	ame of the ou	utput 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 🛛 🗙									
E E F & C : 6									
<ol> <li>odps sql</li> <li>author;tibm</li> <li>author;tibm</li> <li>create time:2018-12-27 10:25:30</li> <li>INSERT OVERWRITE TABLE pm_table_a</li> <li>5 ELECT</li> <li>FROM project_b_name.pm_table_b</li> </ol>	X     When you do not know what the upstream node is, click it.     Peendancies ()     Average of the second sec								
9 ;		Upstream Node Output Table Nan	ie	Node Name					
	MexCompute_DOC_root			maxcompute_doc_root		digitar, dana	Added Manually		
	Output Enter an output name +								
3	MaxCompute_DDC.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 🛛 🗙								≡	
1dps sel 2	Acto Perse  Yes ONo Perse 10 Upstream Node Contra and concept table name   Upstream Node Contra and concept table name   + Use The Workspace Rest Node								
FROM project_b_name.pm_table_b     ;     Upstream Node Output Name									
	b -					Auto Parse			
Output Enter an output na								on Si	
MaxCompute DOC 50110	8 · 8					Added by Default			
MaxCompute_DOC.pm_tab	ole_a 🏈 MaxCompute_DC	OC.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\_na me . 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\_ is recognized as a temporary table. Auto parsing does not resolve the temporary table. If the table with a name starting with t\_ is not a temporary table, contact your project administrator to modify it in the project configuration.

骿	Configuration Center
Þ	Project Configuration
ī	Templates
<b>+</b>	Theme Management
	Table Levels
9	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  $t_-$ , 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.



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

🕤 DataStudio	MaxCompute_DOC	: ~				Cro	ss-project cloning	Operation Center	4 =	English	
Data Developm 🖉 🗄	ដៃC⊕ស	Sq test_sqLC	01 × 🚣 test 🛛 ×							=	
Enter a file or creator nar										08M	
> Solution			x								
✓ Business Flow			Dependencies ②	ancies (0							
🗸 📲 test											
> 🔁 Data Integr	ation									R	
🗸 🕢 Data Devel	opment		Upstream Node Enter an output name or outp								
Sq abc Lo											
Sq Create_	Table Me Locked 12-2										
• Sq JAVA_t	est Locked by dtplus_c		MaxCompute_DOC_root		maxcompute_doc_root		diplos_dece	Added Manually			
Py Pytest											
Vi start M			Output Enter an output name								
• Sq Task_1											
Sq Task_2	Me Locked 12-25 10										
Sq Task_3			N=-C DOC 5001 10007					Added to Defect			
• Sp test Lo			MaxCompute_DOC:500143227_out					Added by Default			
• Mr testMR			MaxCompute_DOC.test_sql_01 @	- 6				Added Manually			
● Sq test_sq	LO1 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 :								
		<pre>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</pre>	As long as yo the dependen continue to po submission.	u confirm that cies are correct, erform the					
	<ul> <li>I confirm to proceed with the submission.</li> <li>2 Tick it, write a note about the change, and click confirm button.</li> </ul>								
Tip If y the	s: ou don't write a no confirm button.	ote, you can't click	Cancel						

## 3.2 Use Eclipse to develop a Java-based UDF

This topic describes how to develop a Java-based user-defined function (UDF) by using the Eclipse-integrated ODPS plug-in.

#### Preparations

Before developing a Java-based UDF using Eclipse, you need to make the following preparations:

1. Use Eclipse to install the ODPS plug-in.

#### 2. Create an ODPS project.

In Eclipse, choose File > New > ODPS Project, enter the project name, and click Config ODPS console installation path to configure the installation path of the *odpscmd client*.

New ODPS Project Wizard	
Create ODPS project	
Project name: ODPS JAVA UDF Use default location Location: C:\Users\furui.fr\eclipse-workspace\OD Config ODPS console installation path	DPS JAVA UDF Browse
<ul> <li>Use default ODPS console installation path</li> <li>Specify ODPS console installation path</li> <li>Version: 0.29.4</li> </ul>	Config ODPS console installation path Browse
? < Back	Next > Finish Cancel

Enter the installation package path and click Apply. The ODPS plug-in automatically parses the version of the odpscmd client.

Preferences		
ODPS Settings	Config ODPS console installation path	
	Config ODPS console installation path C:\odpscmd_public Browse	
	Version: 0.29.4	
	Run Mode Local   Remote	
	limit record count of downloaded 100 (0~10000)	
	Retain local job temp directory	
	Restore Defaults Apply	
?	Apply and Close Cancel	

Click Finish.

#### Procedure

• 1. Create a Java-based UDF in the ODPS project.

One the Package Explorer pane, right-click the ODPS Java-based UDF project you have created, and choose New > UDF.



Set the UDF package to com.aliyun.example.udf and name to Upper2Lower, and click Finish.

CNew UDF		
New UDF Create a new l	UDF implementation.	
Source folder:	ODPS JAVA UDF/src	Browse
Package:	Package: com.aliyun.example.udf	
Name:	Upper2Lower	
Superclass:	com.aliyun.odps.udf.UDF	Browse
Interfaces:		Add
		Remove
?	Finish	Cancel

An automatic Java code is generated after you create a UDF. Do not change the name of the evaluate() function.



• 2. Implement the evaluate() function contained in the UDF file.

Write the function code to be implemented into the evaluate() function. Do not change the name of the evaluate() function. The following is an example of how to convert uppercase letters to lowercase letters.

```
Upper2Lower.java
                                      TestUpper2Lower
                                                          🛽 Upper2Lower.java 🛛
WordCount.java
 1 package com.aliyun.example.udf;
 2
 3 import com.aliyun.odps.udf.UDF;
 4
 5 public class Upper2Lower extends UDF {
 6⊜
        public String evaluate(String s) {
 7
            if (s == null) { return null; }
 8
            return s.toLowerCase();
 9
        }
 10 }
```

```
package
          com . aliyun . example . udf ;
import
         com . aliyun . odps . udf . UDF ;
public
                                          UDF {
         class
                 Upper2Lowe r
                                extends
    public
            String
                      evaluate ( String
                                         s ) {
        if (s ==
                     null ) { return null ; }
                s . toLowerCas e ();
        return
   }
}
```

Save the code.

Test the Java-based UDF code

Before testing the Java-based UDF code, store some uppercase letters on MaxCompute. Create a test table named upperABC using the create table upperABC (upper string); SQL statement on the odpscmd client.



Use the insert into upperABC values (' ALIYUN '); SQL statement to insert the string of uppercase letters 'ALIYUN'.

Choose Run > Run Configurations to set the test parameters.



Set the test parameters. Set Project to the name of the Java ODPS project you have created, and set Select ODPS project to the MaxCompute project name. Note that the project name needs to match the name of that connected to the odpscmd client. Set Table to upperABC. After completing all the settings, click Run.

Run Configurations			
Create, manage, and run config	urations		
Image: Second system       Image: Second system         Image: Secon	Name: Upper2Lower  G UDF UDTF UDAF A JRE Classpat Project: ODPS JAVA UDF UDF UDTF UDAF class: com.aliyun.example.udf.Upper2Lower Select ODPS project MaxCompute_DOC example_project Input Table Table: upperABC Partitions: Columns:	h Tenvironment Common	Browse Search ie: p1=1,p2=1 (default all partitions) ie: c1,c2,c3 (default all columns)
Filter matched 10 of 10 items			Revert Apply
0			Run Close

You can view the test result in the Console pane, as shown in the following figure.

## Note:

Eclipse obtains the string of uppercase letters from the table and converts them to a string of lowercase letters, which is 'aliyun'. However, the uppercase letters stored on MaxCompute are not converted.



#### Use the Java-based UDF

You can use the Java-based UDF after the test is successful. The procedure is as follows:

#### 1. Export the JAR package.

Right-click the ODPS project you have created and select Export.

eclipse-workspace	- C	DPS JAVA UDF/src/com/al	iyun/example/udf/Upper2Lower.ja	va - Eclipse			
File Edit Source I	Refa	actor Navigate Search	Project Run Window Help				
🗖 🗕 🖩 🕼 🖗 🏄 🤅	≥ ₹		<b>♀</b> ♥ ♂ ▼ ≫ ⇔ <i>∢</i> ▼ ∦ ▼	- 종 - ◆ ◆ ◆ → → →			
😫 Package Explorer	83	₽\$ \$ ▼ □	🛿 WordCount.java 🔹 Upper2L	.ower.java 📄 TestUpper2Lower	🛽 Upper2Lower.java 🛛	- 0	문
MaxCompute_I	DOG	C	1 package com.aliyun.exam	ple.udf;		*	
A 🖉 ODPS JAVA UD	F	Now	2	df UDE:			
▲ src		Go Into	,				1
Dependent		Open in New Window		• extends UDF {			
🛛 🛋 JRE System I		Open Type Hierarchy	F4	return null; }			
Referenced		Show In	Alt+Shift+W ►	Case();			
examples	D	Сору	Ctrl+C				
verip	Þ	Copy Qualified Name					
		Paste	Ctrl+V				
	<u></u>	Delete					
	_9	Remove from Context	Ctri+Ait+Shirt+Down				
		Source	Alt+Shift+S ►				
		Refactor	Alt+Shift+T ▸				
	പ്പ	Import					
	4	Export					
	S	Refresh	F5				
		Close Project					
		Close Unrelated Projects	i				
		Carrange As					
		Coverage As Run As	•				
		Debug As	•				
		Validate					
		Restore from Local Histo	ory				
		Team	•				
		Compare With	F	ation El Consola M	(	<b>x</b> %	B. 5
		conngure	, 	LIDELIDTELIDAEL C:\Program Eiles	lava\ire1.8.0.192\bin\iavaw	ava (2018	在12日
	_	Properties	AIt+Enter	Dle scheme : MaxCompute DOC.up	perABC>C:\Users\furui	i.fr\ecl	ipse-
			[INFO]Start to download tak	ole: 'MaxCompute_DOC.upperABC'	, download mode:AUTO		
			[INFO]Tunnel DownloadSessic [INFO]Start to write table	on ID is : 2018121417544782dcd MaxCompute DOC upperARC>C	b0b0†817516 \Users\furui_fr\eclinse	-worksn	ace\(
			Limejotare to write table.	- ancompute_bot.upperAbt-70.			

On the displayed page, select JAR file and click Next.

Export	_ <b>—</b> X
<b>Select</b> Export resources into a JAR file on the local file system.	Ż
Select an export wizard:	
type filter text	
🕨 🗁 Install	•
<ul> <li>▲ Java</li> <li>↓ JAR file</li> <li>ℚ Javadoc</li> <li>↓ Runnable JAR file</li> </ul>	
Run/Debug	
De Casks	
▷ ➢ XML	-
? < Back Next > Finish	Cancel

Enter the JAR package name and click Finish. Then, the JAR package is exported to your workspace directory.

<b>JAR File Specification ①</b> The export destination will be re	elative to your workspace.	.0
Select the resources to export:		
MaxCompute_DOC  MaxCompute_DOC  ODPS JAVA UDF	<ul> <li>✓ I.classpath</li> <li>✓ I.project</li> </ul>	
<ul> <li>Export generated class mes and</li> <li>Export all output folders for che</li> <li>Export Java source files and res</li> <li>Export refactorings for checked</li> <li>Select the export destination:</li> </ul>	ecked projects ources d projects. <u>Select refactorings</u>	0
JAR file: upper.jar	•	Browse
JAR file: upper.jar Options:	<b>-</b>	Browse
JAR file: upper.jar Options: Compress the contents of the J	▼ _	Browse

2. Upload the JAR package to DataWorks.

Log on to the DataWorks console, find the MaxCompute\_DOC project, and go to the *Data Studio* page. Choose Business Flow > Resource > Create Resource > JAR and create a *JAR resource*.



On the displayed page, upload the JAR package you have exported.

Create Resource			×
* Resource Name :	upper.jar		
Folder :			
Resource Type :	JAR	~	
	Upload to ODPS The resource will also be uploaded to ODPS.		
File :	Upload		
		ОК	Cancel

The JAR package is uploaded to DataWorks. To upload it to MaxCompute, click the JAR package and click Submit and Unlock.

f b 🗊	
Upload Resource	
Saved Files :	upperjar
Unique Resource Identifier :	OSS-KEY-inhap5h3ptc6dbwfhs9rq3nr
	✓ Upload to ODPS The resource will also be uploaded to ODPS.
Re-Upload :	Upload

You can run the list resources command on the odpscmd client to view the uploaded JAR package.
#### 3. Create a resource function.

After uploading the JAR package to your MaxCompute project, choose Business Flow > Function > Create Function and create a *function* named upperlower\_Java. After completing these settings, click Save and Submit and Unlock.

Search by file or creator name.	u d 🛛 🖓 🖓	
> Solution	Register Function	
Business Flow		i: (#4=*
> 🏯 base_cdp	Function Name	
<ul> <li>Lione_uatabase_ua_workshop_log</li> <li>Liost</li> </ul>	* Class Name	: com.aliyun.example.udf.Upper2Lower
<ul> <li>Data Integration</li> <li>Data Studio</li> </ul>	* Resources	:: upperjer
> 🛄 Table	Description	
Resource     Function		
Fx upperlower_java     ocked by	Command Format	
> C Control	Parameters	£

You can run the list functions command on the odpscmd client to view the registered function. Then, the upperlower\_Java Java-based UDF registered using Eclipse can be used.

Check the Java-based UDF

In the odpscmd CLI, run the select upperlower \_Java (' ABCD ') from dual ; command. In the following figure, the output is abcd, indicating that the function has converted a string of uppercase letters to lowercase letters.



#### Additional information

For more information about how to develop Java-based UDFs, see Java UDF.

To use IntelliJ IDEA to develop a Java-based UDF, see Use IntelliJ IDEA to develop a Javabased UDF.

## 3.3 Use MaxCompute to analyze IP sources

This topic describes how to use MaxCompute to analyze IP sources. The procedure includes downloading and uploading data from an IP address library, writing a user-defined function (UDF), and writing a SQL statement.

#### Background

The query APIs of *Taobao IP address library* are *IP address strings*. The following is an example.



HTTP requests are not directly allowed in MaxCompute. However, you can query IP addresses in MaxCompute using one of the following methods:

- 1. Run a SQL statement and then initiate an HTTP request. This method is inefficient. The request will be rejected if the query frequency is lower than 10 QPS.
- 2. Download the IP address library to the local server. This method is inefficient and will affect the data analysis in data warehouses.
- 3. Maintain the IP address library regularly and upload it to MaxCompute. This method is relatively effective. However, you need to maintain the IP address library regularly.

The following further describes the third method.

Download an IP address library

- 1. You need to obtain data from an IP address library. This section provides a *demo* of *an incomplete UTF-8 IP address library*.
- 2. Download the UTF-8 IP address library and check the data format, as shown in the following figure.

0,16777215,"0.0.0.0","0.255.255","","Intranet IP","Intranet IP","Intranet IP" 16777216,16777471,"1.0.0.0","1.0.0.255","Australia","","","","","" 16777472,16778239,"1.0.1.0","1.0.3.255","China","Fujian","Fuzhou","Telecom"

The first four strings of data are the starting and ending IP addresses, among which the first two are decimal integers and the second two are expressed in dot-decimal notation. The decimal integer format is used to check whether an IP address belongs to the target network segment.

#### Upload data from the IP address library

1. Create a table data definition language (DDL) on the MaxCompute client, or create a table on the GUI in DataWorks.

```
DROP
        TABLE
                IF
                      EXISTS
                               ipresource ;
CREATE
          TABLE
                  ΙF
                       NOT
                              EXISTS
                                       ipresource
(
                BIGINT
     start_ip
    , end_ip
               BIGINT
     start_ip_a rg
                       string
      end_ip_arg string
                STRING
      country
             STRING
      area
             STRING
      city
               STRING
      county
            STRING
      isp
);
```

2. Run the Tunnel commands to upload the ipdata.txt.utf8 file, which is stored on the D

drive.

```
odps @ workshop_d emo > tunnel
                                   upload
                                            D:/ ipdata . txt .
utf8
       ipresource ;
```

You can use the select count (\*) from ipresource; SQL statement to view the uploaded data. Generally, the quantity of data increases in the library due to regular updates and maintenance.

3. Use the select count (\*) from ipresource limit 0, 10; SQL statement to view the first 10 pieces of data in the ipresource table, as shown in the following figure.

J∘	b Queueing	•		
+-   +-	start_ip	end_ip	start_ip_arg   end_ip_arg   country   area   city   county   isp	
	3395369026	3395369026	″202. 97. 56. 66″   ″202. 97. 56. 66″   "China"   <sub>"Hunan"</sub>   "Changsha"   <sub>""</sub>   "Telecom"	
	3395369027	3395369028	″202. 97. 56. 67″   ″202. 97. 56. 68″   *China"   *Heilongjiang"   **   **   *Telecom"	
	3395369029	3395369029	″202. 97. 56. 69″   ″202. 97. 56. 69″   "China"   "Anhui"   "Hefei"   🖏   "Telecom"	
	3395369030	3395369030	″202.97.56.70″   ″202.97.56.70″   "China"   "Hunan"   "Changsha"   ",   "Telecom"	
	3395369031	3395369033	″202. 97. 56. 71″   ″202. 97. 56. 73″   "China"   "Heilongjiang"   **   **   *Telecom"	
	3395369034	3395369034	"202. 97. 56. 74"   "202. 97. 56. 74"   "China"   "Hunan"   "Changsha"	
	3395369035	3395369036	″202. 97. 56. 75″   ″202. 97. 56. 76″   "China"   "Heilongjiang"   🖏   🖏   "Telecom"	
	3395369037	3395369037	″202. 97. 56. 77″   ″202. 97. 56. 77″   "China"   "Jiangsu"   "Nanjing"   <sub>""</sub>   "Telecom"	
	3395369038	3395369038	″202. 97. 56. 78″   ″202. 97. 56. 78″   "China"   "Hunan"   "Changsha"   ""   "Telecom"	
	3395369039	3395369040	″202. 97. 56. 79″   ″202. 97. 56. 80″   "China"   "Heilongjiang"   <sub>""</sub>   <sub>""</sub>   "Telecom"	

#### Write a UDF

1. Choose Data Studio > Business Flow > Resource. Right-click Resource and choose

Create Resource > Python. In the displayed dialog box, enter the name of the

Python resource, select Upload to ODPS and click OK, as shown in the following figure.



2. Write code for the Python resource. The following is an example:

```
from
       odps . udf
                   import
                            annotate
@ annotate (" string -> bigint ")
        ipint ( object ):
class
       evaluate (self, ip):
 def
  try :
   return
            reduce ( lambda
                           x, y:(x << 8)+y,
                                                         map (
int , ip . split ('.')))
  except :
   return
            0
```

Click Submit and Unlock.



3. Choose Data Studio > Business Flow > Function. Right-click Function and select Create Function.

Set the function class name to ipint . ipint , and the folder to the resource name, and click Submit and Unlock.

Fx ipint.ipint	🕘 🗗 ipint.py	/ × 🛔	test	×		
면 더	ه 🕞	C				
Register Fur	nction					
					~	
1	Function Name :					
	* Class Name :	ipint.ipint				
	* Resources :	ipint.py				
	Description :					
Cor	nmand Format :					

4. Create an ODPS SQL node and run the SQL statement to check whether the ipint function works as expected. The following is an example.



You can also create a local *ipint* . *py* file and use the *MaxCompute client* to upload the resource.

odps @ MaxCompute \_DOC > add py D :/ ipint . py ; OK : Resource ' ipint . py ' have been created .

After uploading the resource, use the client to register the function.

```
odps @ MaxCompute _DOC > create function ipint as ipint .
ipint using ipint . py ;
Success : Function ' ipint ' have been created .
```

The function can be used after registration. You can use select ipint (' 1 . 2 . 24 . 2 '); on the client to test the function.



You can perform *cross-project authorization* to share the UDF with other projects under the same Alibaba Cloud account.

1. Create a package named ipint.

```
odps @ MaxCompute _DOC > create package ipint ;
OK
```

2. Add the UDF to the package.

```
odps @ MaxCompute _DOC > add function ipint to package
ipint;
OK
```

3. Allow a bigdata\_DOC project to install the package.

odps @ MaxCompute \_DOC > allow project bigdata\_DO C to install package ipint; OK

4. Switch to a bigdata\_DOC project that needs to use the UDF and install the package.

```
odps @ MaxCompute _DOC > use bigdata_DO C ;
odps @ bigdata_DO C > install package MaxCompute _DOC .
ipint ;
OK
```

5. Then, the UDF can be used. If a user (such as Bob) of the bigdata\_DOC project

wants to access the resource, the administrator can grant the access permission to the user by using the ACL.

```
package
odps @ bigdata_DO C > grant
                               Read
                                     on
                                                    MaxCompute
                          aliyun $ bob @ aliyun . com ; -- Use
_DOC . ipint
            to
                   user
the
     ACL
           to grant the
                             package
                                       access
                                                permission
                                                             to
 Bob .
```

Use the IP address library in SQL

### Note:

This section uses the IP address 1.2.254.2 as an example. You can use a specific field to query an IP address as needed.

₹ ۲

🗸 isp

#### You can use the following SQL code to view the test result:

start\_ip\_arg

```
select * from ipresource
WHERE ipint (' 1 . 2 . 24 . 2 ') >= start_ip
AND ipint (' 1 . 2 . 24 . 2 ') <= end_ip</pre>
```

To ensure the data accuracy, you can regularly obtain data from the Taobao IP address library to maintain the ipresource table.

country

## 3.4 Performing a task at a specific time with branch node

✓ end\_ip\_arg

Background of branch nodes

During the daily use of DataWorks, you may often encounter the following problem: I have a node that needs to be executed on the last day of each month. How should I set it up?

Answer: Before the *branch node* appears, the Cron expression can not express this scene, so it is temporarily unavailable to support.

Now, DataWorks officially supports branch nodes. With branch nodes, we can apply the switch-case programming model to perfectly meet the above requirements.

Branch nodes and other control nodes

On the Data Development page, you can see the various control nodes currently supported by DataWorks, including assignment nodes, branch nodes, merge nodes and so on.



The functions of various types of control nodes:

· Pass its own results to the downstream assignment node:

The *assignment node* reuses the characteristics that the *node context* depends. Based on the two existing constant/variable node context, the assignment node comes with a custom context output. DataWorks captures the select result or the print result of the assignment node. This result is used as the value of the context output parameter in the form of outputs for reference by downstream nodes.

• Determine which downstream branch nodes are normally executed:

The *branch node* reuse the characteristics of the *input and output* set in the dependencies on DataWorks.

For common nodes, the output of the node is only a globally unique string. When the downstream needs to set dependencies, searching for this globally unique string as input to the node can be hung into the list of downstream nodes.

However, for a branch node, we can associate a condition for each output: when the downstream set dependency, we can selectively use the output of a certain condition as the output of the branch node. In this way, when the node becomes the downstream of the branch node, it is also associated with the condition of the branch node: that is, the condition is satisfied, and the downstream corresponding to the output is executed normally; the other downstream nodes corresponding to the output that does not meet the criteria are set to run empty.

The merge node that is normally scheduled regardless of whether the upstream performs normally:

For branches that are not selected by the branch node, DataWorks puts all node instances on this branch link as empty run instances. That is to say, once an

upstream of a certain instance is running empty , this instance itself becomes empty.

Dataworks can currently prevent this empty run attribute from being passed without restriction by *merge node*: for a merge node instance, no matter how many empty run instances its upstream has, it will succeed directly and will no longer leave the downstream empty running.

From the figure below, you can see the logical relationship of the dependency tree in the presence of a branch node.



• ASN: an *assignment node*, which is used to calculate more complex situations to prepare for branch node conditional selection.

- X/Y: branch nodes, they are downstream of the assignment node ASN, and make branch selection according to the output of the assignment node. As shown in the green line in the figure, the node X selects the left branch, the node Y selects the two branches on the left:
  - The node A/C are executed normally because they are downstream of the output selected by the node X/Y.
  - Although the node B is downstream of the branch selected by the node Y, since the node X does not select this output, the node B is set to run empty.
  - The node E is not selected by node Y, so even if there is an ordinary upstream named node Z, it is also set to run empty.
  - The upstream node E of the nodeG runs empty, so even if the node C/F are both executed normally, the node E also runs empty.
  - How can the empty running properties no longer be passed down?

JOIN node is a merge node. Its special function is to stop the transfer of empty run properties. You can see that because the node D is downstream of the JOIN node, the empty run attribute of the node B is blocked, and the node D can start running normally.

By using branch nodes to cooperate with other control nodes, we can satisfy the requirement scenario where a node only runs on the last day of each month.

#### Use a branch node

Define task dependencies

First you need to define a set of task dependencies:

- 1. The root assignment node calculates whether the current day is the last day of the month by timing SKYNET\_CYC TIME. If it is, the output is 1, and if it is not, the output is 0. This output is captured by DataWorks and passed to the downstream.
- 2. The branch node defines the branch according to the output of assignment node.
- 3. The two shell nodes are hung under the branch node and perform different branch logic.



#### Define assignment nodes

The assignment node comes with an outputs when it is new, the code for the assignment node supports three kinds of SQL/SHELL/Python.

- For SQL types, DataWorks captures the SQL of the last SELECT statement as the value of outputs.
- For SHELL/Python types, DataWorks captures the last line of standard output as the value of outputs.

In this article, the Python type is used as the code for the assignment node, and the scheduling properties and code settings are as follows.

#### • The code is as follows:

👗 Brar	nch_JudgeDownRun × 👍 Assign_IfisLast × 📇test ×
	T L C
	Please select assignment language: Python v
	import os
	import time
	<pre>dueTimeStr = os.environ['SKYNET CYCTIME'] #20190104000200</pre>
	dueTime = time.strptime(dueTimeStr[:8]), "%Y%M%D")
	<pre>dueMonth = time.strftime("%m",nextDueTime)</pre>
	<pre>nextDueTimeStamp = int(time.mktime(dueTime))+3600*24</pre>
	<pre>nextDueTime = time.localtime(nextDueTimeStamp)</pre>
	<pre>nextDueMonth = time.strftime("%m",nextDueTime)</pre>
	<pre>print ('Current month: %s, next month: %s') % (dueMonth, nextDueMonth)</pre>
	if nextDueMonth == dueMonth:
	print 0
	else:
	print 1

• Schedule configuration

👗 Bra	nch_JudgeDov	vnRun × (	🕂 Assign_IfisLast 🗙 嚞 🛄 test	×					Ξ
	[J] [J]	ê (	2						0&M
1	Ple:	X Output							
2	import t	Outpu	t Name	Output Table Name	Descendant Node Name	Descendant Node ID	Owner		
4 5 6	dueTimeS dueTime dueMonth	-	_test_analyticdb.30229872_out	- Ø	Branch_JudgeDownRun			Added Automatically	
7 8	nextDueT		est_analyticdb.Assign_lfisLast @	- Ø				Added Manually	
10 11 12 13 14 15	nextDue print (' if nextE print	Parame	eters ⑦ ameters Add						
16 17			Parameter Name Val		Description	Parent Node ID	Filter		
		Output Pa	arameters Add						
			Parameter Name Type	Value outputs: the supe	Description Prower of the assign	iment node.			
		1	outputs Variab	le \${outputs}	BRITISHING, BRITIS	A	dded Automatic	cally Change D	

#### **Define branches**

Branch nodes can define conditions with simple Python syntax expressions, each of which is bound to an output. This means that the downstream node under this output is executed when the condition is met, and the other nodes are set to run empty.

#### · Schedule configuration

🝌 Bra	nch_JudgeDo	vnRun 🕘 🚑 Assign_IfisLast 🗙 嚞 💷 test 🛛 🗴								≡
•	<u>ئ</u> ] (۲	€ C							0&	м
	Branch Lo	×								Sche
				RunOnLast						edule
B r a		_test_analyticdb.Branch_JudgeDownstrRun 🖉	- Ø	RunExceptLast			Added Manually			Lineage
		last_day_cond.not_last	- @				Added Automatically			Vers
		last_day_cond.is_last	- Ø				Added Automatically			sion
		Parameters @								
		Input Parameters Add depend on the	output of assignme	ent node: isl ast is 1 on	the last day isl ast i	is () in oth	ier cases			
		No. Parameter Name Value Source		Description	Parent Node ID	Filter				
		1 isLast test_analyticdb.3	0229872_out:outputs	Mariamon, American	• •	Added	Manually Change Dele	ete		
		Output Parameters Add								
		No. Parameter Name Type								
		1 outputs Variable	\${outputs}	COLUMN A. BRENCH	Added	Automatically				

#### • Branch configuration

🚠 Bra	nch_JudgeDownRu	n 🗙 👍 Assign_IfisLast :	× 🛃 📃 test 🗙		
		C			
	Branch Logic del	inition (?)			
		Different brand	h conditions are defined accord	ling to isLast.	
	Branch	Conditions		Branch describe	
	1	\$(isLast)==0	last_day_cond_not_last	If current day is not the last day, use this branch condition.	
	2	\$(isLast)==1	last_day_cond.is_last	If current day is the last day, use this branch condition.	

· Schedule configuration generates output of conditional bindings

🚠 Bra	inch_Jud	geDownRun	×	🕭 Assign_IfisLast 🗴 🛃							≡
면		6		( )							O&M
	Branch Add I	I Logic defin Branch	>	Coutput Search by output national Search by output national Search by output national Search by							Schedule
											Line
		Cond itions		test_analyticdb.30229	9873_out	- @				Added Automatically	age \
							RunOnLast				/ersion
ł		S(ISL ast)= =0		test_analyticdb.Branc	:h_JudgeDownstrRun 🥝	- Ø	RunExceptLast			Added Manually	
		\$(isL ast)= =1		last_day_cond.not_last	Here two outputs	- © with the same na	-	-	-	Added Automatically	
				last_day_cond.is_last		- C	-	-	-	Added Automatically	

Hang the execution task nodes under different branches

Finally, it is important to note when setting dependencies on the nodes that actually perform tasks: you can see that the branch node already has three outputs, according to the logic of setting dependencies in the past, any one of these three outputs can be regarded as input; however, since the output of the branch node is now associated with the condition, it should be carefully selected. · Node dependencies performed on the last day of each month



· Node dependencies performed at other times of each month

Image: Second
Image: Second state sta
1 #I/bin/ 2 #************************************
6 Parent Node Search by output name or output table nam V + Use The Workspace Root Node
Parent Node Output Name Parent Node Output Table Name Node Name Parent Node ID Owner Filter Actions
last_day_cond not_last - Branch_JudgeDownRun Added Manually

#### **Result verification**

Once completing all of the above configuration, you can submit and publish the task. After publishing, you can perform patch data to test the effect: the business date 2018-12-30 and 2018-12-31 are selected, that is, the timing is 2018-12-31 and 2019-01-01 respectively, so that the first batch of patch data should trigger the logic of "last day", the second batch triggers the logic of "non-last day". We look at the difference between the two.

Business date 2018-12-30 (timing 2018-12-31)

· Branch selection results of branch node

»	(	Branch_Judge Branch M	DownRun Iode	
H4. 170	Running Log		15.6	
<ul> <li>⊘ 01-04 23:26:19 ~ 01-04</li> <li>23:26:25</li> <li>Gateway:</li> </ul>	2019-01-04 23:26:1 2019-01-04 23:26:2 python output: Fal 2019-01-04 23:26:2 2019-01-04 23:26:2	9.867 INF0 1.239 INF0 se 1.419 INF0 2.424 INF0	<ul> <li>The foll</li> <li>Started</li> <li></li> <li>not meet</li> </ul>	owing profiles are active: dev ControllerWrapper in 2.072 seconds (JVM runni the condition! condition:1==0
	python output: Tru 2019-01-04 23:26:2 2019-01-04 23:26:2 2019-01-04 23:26:2 2019-01-04 23:26:2 2019-01-04 23:26:2	e 2.435 INFO 3.436 INFO 3.437 INFO 4.070 INFO 4.071 INFO	- meet the - ===>Outp - cost lim - job fini	e condition. condition:1==1 but Result: last_day_cond.is_last me: 2 shed!

· The node "RunOnLast" is executed normally.

• The node "RunExceptLast" is set to run empty.

#### Business date 2018-12-31 (timing 2019-01-01)

· Branch selection results of branch node

>	😔 Bra	Branch Node	n	
R8 173	Running Log	156		
<ul> <li>⊘ 01-04 23:26:41 ~ 01-04</li> <li>23:26:52</li> </ul>	2019-01-04 23:26:48.414 python output: True	INFO - Star	ted ControllerWrapper in 3.4	85 seconds (JVM runni
	2019-01-04 23:26:48.614	INFO	the condition condition:0-	0
Gateway:	python output: False	INIO - meet		
	2019-01-04 23 26 49.634 2019-01-04 23 26 50.635	INFO INFO - not	meet the condition! condition	on:0==1
	2019-01-04 23:26:50.636	INFO - ===> INFO - cost	Output Result: Last	_day_cond.not_last
	2019-01-04 23:26:50.981 2019-01-04 23:26:51 INFO	INFO - job	finished!	

- The node "RunOnLast" is set to run empty.
- The node "RunExceptLast" is executed normally.

#### Summary

Based on the branch node, you have achieved the goal which execute on last day of each month. Of course, this is the easiest way to use a branch node. By using an assignment node with a branch node, you can combine a variety of conditions to meet your business needs.

Finally, the main points of using branch nodes are reviewed.

- DataWorks captures the last SELECT statement or the last line of the standard output stream of the assignment node as the output of an assignment node for downstream references.
- Each output of the branch node is associated with the condition, and the downstream branch node is used as the upstream. It is necessary to understand the meaning of the conditions associated with each output before selecting.
- Unselected branches are set to run empty, and the empty run properties are passed down until the merge node is encountered.
- In addition to blocking the empty run properties, the merge node has more powerful features to wait for your mining.

## 4 Data security

# 4.1 Log on to DataWorks through a specific IP address with a RAM user

In the process of data development , some users with strict permission control require that the RAM user can only log in through a specific IP address in the company. This article describes how to use a RAM user to log on to DataWorks through a specific local IP address.

#### Prerequisites

First, you need to refer to *Create a RAM user* to complete the creation of RAM user and authorize it. The permission AliyunDataworksFullAccess is the default system permission and cannot be modified. You need to create an additional custom authorization policy.

#### Configure a custom permission policy

Log on to the *RAM console*, click Create Policy in the Policies column to enter the edit page. In this case, the policy name is dataworksIPlimit1.

RAM		RAM / Policies			
Overview		Policies			
Identities	^				
Groups		RAM supports two types of policies: system polici	a simple Policy language specification to describe permission sets. For more information, se les managed by Alibaba Cloud and custom policies managed by you.	e Policy syntax.	
Users		- You can use but cannot modify the system policies mana - You can create, modify, or delete the custom policies. In	ed by Alibaba Cloud. Alibaba Cloud maintains and updates the system policy versions. ddition, you need to maintain the policy versions by yourself.		
Settings					
Permissions	^	Create Policy Enter a policy name or note	Q Policy Type All		
Grants		Policy Name	Note	Policy Type Us	ed Times
Policies	«	AdministratorAccess	Provides full access to Alibaba Cloud services and resources.	System Policy 8	
RAM Roles		AliyunOSSFullAccess	Provides full access to Object Storage Service(OSS) via Management Console.	System Policy 0	
OAuth Applications		AliyunOSSReadOnlyAccess	Provides read-only access to Object Storage Service(OSS) via Management Cons	sole. System Policy 0	

#### Select Script for the option of configuration mode, and enter your custom policy.

RAM		RAM / Policies / Create Custom Policy				
Overview		← Create Custom Policy				
Identities	^					
Groups		Policy Name				
Users		dataworksIPlimit1				
Settings		Note				
Permissions	^					
Grants		Configuration Mode				
Policies		Visualized  Script				
RAM Roles	«	Policy Document				
OAuth Applications		Import an existing system policy				
		1          "Version": "1",          3       "Statement":         4       [{         5       "Effect": "Deny",         6       "Action": ["dataworks:*"],         7       "Resource": ["acs:dataworks:*:*;*"],         8       "Condition":         9       {         10       "NotIpAddress":         11       Image: SourceTo". "100 1 1 1/30"				

The complete content of the custom policy is shown in the following figure. "acs: SourceIp" is the IP address that you allow access to DataWorks. In this example, it is 100.1.1.1/32. After entering the information, click OK to create the authorization.

```
{
    " Version ": " 1 ",
    " Statement ":
        [{
            " Effect ": " Deny ",
            " Action ": [" dataworks :*"],
            " Action ": [" acs : dataworks :*:*:*"],
        " Condition ":
            {
            " Condition ":
            {
            " NotIpAddre ss ":
            {
            " acs : SourceIp ": " 100 . 1 . 1 . 1 / 32 "
            }
        }]
}
```

#### Add custom permissions

On the RAM console, click Identities > Users, choose the RAM user you want to control, and clickAdd Permissions.

RAM	RAM / Users						
Overview	Users						
Identities ^							
Groups	A row user is an identity entity, it represents a user of application in your organization mat needs to access double resources.  X 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 agroup. To perform this operation, you must have created a user group and granted permissions to it.						
Users							
Settings							
Permissions	Create User Logon Name V Enter Q			G			
Grants	User Logon Name/Display Name Note Cre	reated	Actions				
Policies	Jar	n 23, 2019, 15:54:32	Add to Group Add Permissions Delete				
RAM Roles							
OAuth Applications	Jar	n 2, 2019, 16:40:32	Add to Group Add Permissions Delete				

Select Custom Policy, add the custom policy you just created to the Selected, and clickOK.

Add Permissions									
Principal hui	aliyun.com ×								
Select Policy									
Custom Policy V Enter		Q	Selected (1) Clear						
Policy Name	Note		dataworksIPlimit1 ×						
StreamComputerAccessPolicy	100000000								
StreamComputeAccessRole									
StreamComputeWriteAccess									
ADE									
dataworksIPlimit1									
Ok Cancel									

Verification

Log on to the DataWorks console using an IP address different from 100 . 1 . 1 .

1 / 32 and find that the login failed.

