

Data Lake Compute Getting Started Product Documentation





Copyright Notice

©2013-2025 Tencent Cloud. All rights reserved.

Copyright in this document is exclusively owned by Tencent Cloud. You must not reproduce, modify, copy or distribute in any way, in whole or in part, the contents of this document without Tencent Cloud's the prior written consent.

Trademark Notice

STencent Cloud

All trademarks associated with Tencent Cloud and its services are owned by the Tencent corporate group, including its parent, subsidiaries and affiliated companies, as the case may be. Trademarks of third parties referred to in this document are owned by their respective proprietors.

Service Statement

This document is intended to provide users with general information about Tencent Cloud's products and services only and does not form part of Tencent Cloud's terms and conditions. Tencent Cloud's products or services are subject to change. Specific products and services and the standards applicable to them are exclusively provided for in Tencent Cloud's applicable terms and conditions.

Contents

Getting Started

Complete Process for New User Activation

- DLC Data Import Guide
- Quick Start with Data Analytics in Data Lake Compute
- Quick Start with Permission Management in Data Lake Compute
- Quick Start with Partition Table
- Enabling Data Optimization
- Cross-Source Analysis of EMR Hive Data
- Standard Engine Configuration Guide

Getting Started Complete Process for New User Activation

Last updated : 2025-03-12 18:03:39

Before official and proper use of Data Lake Compute (DLC), you need to configure the initialization parameters and permissions in the target region in advance with the Tencent Cloud root account or DLC administrator account. To avoid unnecessary errors, it is recommended that common users use DLC after the administrator completes the configurations.

Activating the DLC Service for the Root Account(Performed with the Tencent Cloud Root Account)

- 1. Log in to the DLC console.
- 2. Click Go to CAM.
- 3. Authorize the activation of DLC Data Lake Service.



4. In Role Management, click Grant.

Cloud Access Management	Contraction of the second s
🔡 Dashboard	
은 Users 👻	Service Authorization
♣ User Groups	After you agree to grant permissions to Data Lake Compute, a preset role will be created and relevant permissions will be granted to Data Lake Compute
Policies	Role Name DLC_QCSLinkesRaleinCheckDLCResource
Roles	Role Type Service-Linked Role
🔄 Identity Providers 👻	Description The current role is the DLC service role, which will access your other service resources within the scope of the parmissions of the associated policy.
C3 Federated * Account	Authorized Policies Preset Policy QcboudAccessForOLCLinkedRoleinCheckOLCRezource()
(2) Access Key 👻	Crant Cancel

Creating a Tencent Cloud Sub-account and Adding DLC Service Policies Under the Root Account(Performed with the Tencent Cloud Root Account)

To enable multiple accounts to use the DLC service collaboratively, activate the DLC service as follows:



Creating a Tencent Cloud Sub-account

To enable a sub-account to access DLC, go to the Create User page of CAM for configuration.

Adding the DLC Service Policy QcloudDLCFullAccess

- 1. On the list of users in CAM page, select the user to be authorized.
- 2. Click Authorize.
- 3. In the pop-up window, enter DLC and select QcloudDLCFullAccess.

Adding the Sub-account as a DLC Account

Note:

Operation permission: The first time adding a sub-account as a DLC account, use the Tencent Cloud root account. 1. Log in to the DLC console, select Permission Management, and click **Add user**.



Data Lake Compute	Permission management	🖏 Guangzhou 🔻					Use guide 🧭 🛛 Permission manag	ement guide 🗵
E Overview	User Work group							
© Data Explore 프는 Data Management	Both sub-account and coordina here 2.	tor users need to be granted with permiss	ons to use data, engines, and other resou	rces. A user may be associated with one or i	more work groups to inherit all of their permis	sions. An admin user has all resource permiss	ions. For more permission guidelines, see	×
Data Jobs Data Engines	Add user Batch delete						Enter a user ID or name	Q Ø
Global	User ID	Username	User type ①	Description	Added by	Add time 🗘	Operation	
Configuration			Ad	and the second se	1	2023-03-07 16:56:41	Edit Authorize 🔻 Delet	æ
Permission Management	Total items: 1					10 -	/ page 🛛 4 1 / 1 pa	ge 🕨 H
 Storage Configuration 								
Data Operations 👻								

2. Select the user ID of the account you want to add, and specify the user type.

Jser Information	Select Tencent Cloud Users (Create User (i)				
	User ID	Username	Description (Optional)	Operation		
		Click the icon in the upper left corr	ner to add the user to authorize.			
Jser type	O General user					
	Users need to be assigned the com	responding database and table permis	ssions or be bound to a working group to access re	esources.		
		atahases tables data engines and c	ther resources and can edit permissions of all acco	ounts except the root account		
A/	Bind to Working Group					

DLC Account Types and Permission Scopes

Permission & Operation	DLC Administrator	Common User		
Data permissions	All permissions	No permission by default, and permissions need to be authorized by a DLC administrator.		
Standard engine	All permissions	No permission by default, and permissions need to be authorized by a DLC administrator.		
User management	Available	Unavailable		
Work group management	Available	Unavailable		

Authorization scope	All permissions	Authorizable permissions
---------------------	-----------------	--------------------------

Causes of Addition Failures

If an error occurs when you add the account as a DLC account, check whether any of the following issues occur. If not, <u>contact us for help</u>.

- 1. Repeated addition: The account has already been added as a DLC account.
- 2. Account input error: An incorrect account is entered.

Adding a DLC Administrator Account

Note:

Operation permission: The first time adding a sub-account as a DLC administrator account, use the Tencent Cloud root account. For subsequent addition of sub-accounts as DLC administrator accounts, you can use the added DLC administrator account.

The add method is consistent with the previous chapter (adding an account as a DLC user). When selecting the type of user, choose DLC administrator.



DLC Administrator Account Types and Management Scope

Administrator Type	Manageable Platform	Creating a Tencent Cloud Sub- account	Adding a DLC Policy	Adding a Tencent Cloud Sub-account as a DLC User	DLC Computing and Data Permissions
Root account (owner)	CAM/DLC	Allowed	Allowed	Allowed	Allowed
DLC administrator account	DLC	Disallowed	Disallowed	Allowed	Allowed

Configuring the Storage Path of Query Results



Note:

Operation permission: Configured with the root account or DLC administrator account.

Enter Storage Configuration in the DLC console and configure the storage path for requry results on the Overview page or the Storage Configuration page. After the configuration is completed, the query results are stored in the specified Cloud Object Storage (COS) path or the managed storage device of DLC.

Query result storage config	uration
Modify configuration	
Storage path of query results (Internal storage
	The SELECT query results are stored in the internal storage of Data Lake Compute, and the underlying storage service is COS. The results will be retained for 36

Feature description:

1. DLC internal storage: The SELECT query results are stored in the DLC storage. The underlying storage is COS. The results can be stored for 36 hours.

2. User storage: The SELECT query results are stored in the bucket path on COS. You need to check whether the COS-related permissions are granted.

3. Metadata acceleration bucket: The performance of query and analysis in the local region can be improved.

4. For internal tables, the metadata acceleration bucket can be enabled directly. For external tables, you need to check whether the metadata acceleration bucket can be enabled based on the engine permission.

Note: A shared engine cannot be bound to a metadata acceleration bucket. When a user selects the user storage path, the exclusive engine needs to be bound to a metadata acceleration bucket before querying takes effect.

Purchasing an Engine

Note:

Operation permission: Purchased with the root account or an account having the financial permission.

You can purchase different types of engines based on your business requirements. Engines are classified into standard engines and SuperSQL engines. The two types of engines support different SQL syntaxes. The standard engine supports the native syntax and behavior, while the SuperSQL engine supports the DLC-developed SuperSQL syntax.

1. Purchase method: Go to the Standard Engine page.

2. Click **Create resource** to enter the purchase page.

Note:

1. The engine is divided into the standard engine and SuperSQL engine. The difference between them is: They support different SQL syntaxes. The standard engine supports native syntax and behaviors, while the SuperSQL engine supports DLC's self-developed SuperSQL syntax.

2. Engine Specification Purchase Advice: Since a 16-CU cluster is relatively small in scale, it is advisable to use it only for testing scenarios. For real production scenarios, it is recommended to purchase a cluster with 64 CUs or more.

Standard Engine Permission Management

Note:

Operation permission: Configured with the root account or a sub-account having the CAM permissions.

Tencent Cloud CAM controls the DLC standard engine permissions. To ensure that the sub-account can use the DLC standard engine smoothly, you need to use the root account to authorize the sub-account the permissions. After the standard engine is created, all sub-accounts with the **QcloudDLCFullAccess policy (all read/write and access permissions in the DLC console)** have the permissions to use, manage, and monitor the standard engine. To achieve granular permission management of the standard engine, for example, user A only having the permission of engine A, you can create custom policies.

Scenario	Operation
Scenario 1: The sub-account has all standard engine permissions.	Associate the preset QcloudDLCFullAccess policy with the sub-account.
Scenario 2: The sub-account has partial standard engine permissions.	Create a custom policy.

Granting All Standard Engine Permissions to a Sub-account

After you log in to the DLC console by using the root account or a sub-account having CAM operation permissions, find the sub-account in the sub-account list, click Authorize in the **Operation** column, search for and select **QcloudDLCFullAccess**, and click **OK**.

Granting Partial Standard Engine Permissions to a Sub-account

DLC supports resource-level authentication based on CAM tags. You can use tags to manage the existing standard engine resources and engine-related API permissions of DLC by category, achieving multidimensional resource management by category and granular authorization. For details about Tencent Cloud tags, see Tag. Based on Tencent Cloud tags, you can quickly achieve the following effects for DLC standard engine resources: All users in department A can only use the standard engine resources associated with the tag for department A. and

cannot use the standard engines associated with the tag for another department.

When users in department A create DLC standard engine resources, the resources should be associated with the tag for department A. If the resource is not associated with any tag or is associated with the tag for another department (rather than department A), the creation fails (optional).

Operation Steps

Step 1: Creating a Tag

1. Enter the Tag List page and click Create Tag.

2. Set **Tag Key** and **Tag Value**, click **OK**, and then a tag is created. For example, to create a tag with the department name Analyze, enter department for Tag Key and Analyze for Tag Value.

Step 2: Tagging the Standard Engine

Log in to the DLC console and select Standard engine. In the tag option, select a tag you want to bind. For detailed operations on tagging a standard engine, see Associating Tag with Private Engine Resource.

Note:

Once a specific tag is associated with a standard engine, such as "department: Analyze" in the above example, only the users who are associated with that tag can use and manage the standard engine.

Step 3: Creating a Custom Policy

- 1. Log in to the CAM console.
- 2. Click Create Custom Policy. In the pop-up window, select Authorize by Tag.



3. On the Visual Policy Generator tab, select DLC for Service and select All actions (dlc:*) for Action.



1 Edit Policy >	2 Associate User/User Group/Role	
Visual Policy Generator	JSON	
Add Services and Operation	1S Add	
▼ dlc(All actions)		
Service •	DLC O	All Services (*)
	✓ Data Lake Compute (dlc)	
Action *	Select actions	
	✓ All actions (dlc:*) Show More	
	Action Type	
	Write (221 selected) Show More	
	 List (29 selected) Show More 	
	Conters (12 selected) Show More	
	Among your selected actions, 373 API does not support authorization by tag	

Note:

To restrict the permissions to terminate, create, or modify a standard engine, you can deselect the APIs under All actions. The APIs are described in the following table.

Situation	DLC API to be Deselected
Unable to terminate an engine	DeleteDataEngine
Unable to create an engine	CreateDataEngine
Unable to modify an engine	UpdateDataEngine

4. Select the "department: Analyze" tag created previously. By default, resource_tag under Select Condition Key is selected.

You can select request_tag, so that the users in the Analyze department will be required to associate the new DLC standard engine with the tag "department: Analyze" when creating an engine. For more introduction and usage of request_tag, see request_tag.

5. Click **Next** and CAM creates multiple split sub-policies. You can enter a name that is easy to search for a split subcustom policy, such as "DLC-department-analyze-tag-policy". Next to **Authorized Users** or **Authorized User Groups,** select the user/user group that needs to be associated with this custom policy. For example, associate all employee sub-accounts of the Analyze department in this example.

6. Click **Complete** and the custom policy is created. After the above custom policy is created, all DLC users associated with this custom policy can only access the standard engine tagged with "department: Analyze".

Note:

1. To facilitate subsequent user maintenance, it is recommended that a user group be associated.

2. If a user associated with a custom policy has already been associated with the preset QcloudDLCFullAccess policy, the user still has all standard engine permissions.

SuperSQL Engine Permission Management

Note:

Operation permission: Configured by the root account or DLC admin.

Automatic Authorization Of Engine Operation Permissions (Only Supports SuperSQL Engine)

DLC supports enabling SuperSQL engine operation permissions by default. After enabling, all users will have the following permissions for this engine by default:

Usage: Use this engine for task execution.

Operation: Suspend or hang up the engine.

Monitoring: Monitor the usage and Ops of the engine.

Note:

1. After disabling, administrators will continue to have all engine permissions by default, while ordinary users require administrators to add permissions on the permission management page.

2. The original ordinary users' permissions are not affected and can be deleted by going to Permission Management page.

3. Subsequent newly created ordinary users will not have usage permissions and need to be manually added on the Permission Management page.

How To Enable and Disable Automatic Authorization Engine Permissions

There are two permission entries for default enabling/disabling engine operation:

Entry 1: Engine Purchase Page > Advanced Configuration Item.

Advanced configu		
Advanced configu	uration	
IP range of cluster	10.255.0.0/16	Modify
	This option affect source.	s the network interconnection between services. In case of non-federated queries, default configuration is recommended; in federated queries, the IP range of the engine must be different from that of t
Auto-granting of		
engine permissions	OPERATE: Pause of MONITOR: Monitor	abled, all User's are granted the following permissions on this engine: ine to execute tasks or suspend the engine r and maintain the engine based on its usage ermissions, see here f2

Entry 2: Go to the Engine Management page and click to edit authorization engine permissions.

Data Lake Compute	SuperSQL engine	S Beijing 🗸						SuperSQL engine
 Overview Data Exploration 	Data engines include ex mode. For more billing engine is suspended. F	xclusive data engines and shared da information, see <mark>Billing Overview [</mark> or details about the directions and p	ta engines. Public engines 3 . You can configure an aut rrecautions, see Manage E :	(shared engines) are bille tomatic suspension or sch xclusive Data Engine [2].	d by scanning volume. Excl eduled suspension policy f	usive engines are billed in a pay ior a pay-as-you-go data engine	-as-you-go or m . No fees are inc	onthly subscription urred after the data
Data Management	Create resource Bill quer	y 🖪 🛛 Renewal management 🖪			Please se	lect resource tags or enter keyw	ords to filter. Se	parate multiple C
Data Job Data History	Engine name/ID	Cluster description	Auto-granting of en	Engine size	Network configurat	Created at \$	Creator	Operation
 Insight Management 	10 Aug	sion Private engine	Yes 💉	16CU (standard) 1-2 cluster(s)		2025-02-19 14:55:13		Monitor Spec configuration Parameter configuration More ▼
Engine Management SuperSQL Engine Standard Engine	12.0	ision Private engine	Yes 🎤	16CU (standard) 1-2 cluster(s)		2025-02-17 15:32:11		Monitor Spec configuration Parameter configuration More ▼
 Network Connection Configuration 	81 m.	ision Private engine	Yes 🖋	16CU (standard) 1-2 cluster(s)		2025-02-17 15:13:03		Monitor Spec configuration Parameter configuration More ▼
Ops Management of Permission Management		Private engine	Yes 🖋	16CU (standard) 1-2 cluster(s)		2024-03-28 16:11:28	-	Monitor Spec configuration Parameter configuration More ▼
Storage Configuration		Public engine	No 🧳			2022-03-17 17:31:24		Spec configuration Parameter configuration More ▼

After setting the engine permissions, click Yes.

Set engine _l	permissions	×
Engine name		
Resource ID		
Auto- granting of engine permissions	 ✓ ✓	ng
Confirm	Cancel	

Activate Sub-Account Permissions For SuperSQL Engine In DLC

After creating a user or workgroup, click the authorization operation in the list to add permissions to the workgroup. DLC Data Engines are divided into two categories: **SuperSQL Engine** and **Standard Engine**. For detailed differences and application scenarios, please refer to Data Engines. The earlier-launched **SuperSQL Engine** permissions are managed through the DLC console, where you can quickly manage SuperSQL Engine permissions in DLC Console > Permission Management. On the other hand, **Standard Engine's** permission management is uniformly controlled by Tencent Cloud CAM. For information on Standard Engine's permission management, please refer to DLC Permission Overview.



For the SuperSQL Engine, based on the usage scenario of the user or workgroup, you can check the engine's permission policy in **DLC Console > Permission Management > Engine Permissions**.

Note:

Usage: Use this engine for task execution.

Modification: Modify the engine's configuration parameters, such as specification adjustment of the engine.

Operation: Suspend or hang up the engine.

Monitoring: Monitor the usage and Ops of the engine.

Deletion: Delete the engine.

Authorizable: After checking, all members under this Sub-user or workgroup have the authorization permission for the engine.

Add permission		×
Data engine	Enter	
Engine permission	All Use Modify Operation Volume Monitor Delete	
Authorizable	Yes	

Data Permission Management

Note:

Operation permission: Configured with the root account or a DLC administrator account.

The root account or DLC administrator account can be used to configure permissions based on the usage scenario of users or work groups. On the Permission Management page, select the data permission policy. DLC data permissions include:

Data directory permissions

Permission for creating databases under the DataLakeCatalog directory and the permission for creating databases in other data directories.

Permission scope:

- 1. Whether to allow users or work groups to create databases under DataLakeCatalog
- 2. Whether to allow users or work groups to create databases in other data directories

Database and table permissions

Permissions for setting databases, data tables, views, and functions.

Permission scope:

Permission Type	Databases	Data Tables	Views and Functions
Query and analysis permissions	Permissions to query all tables, views, and functions. create data tables in the database.	Query	Query
Data editing permissions	Permissions to modify and delete databases and create tables. All permissions of all tables, views, and functions.	Data query, insertion, update, and deletion. Table modification and deletion.	Query, creation, modification, and deletion
Owner permissions (permission is further granted based on the data editing permission)	Permissions to modify and delete databases and create tables. All permissions of all tables, views, and functions.	Data query, insertion, update, and deletion. Table modification and deletion.	Query, creation, modification, and deletion

Advanced permission settings for databases and tables

If you select a single database, you can further set the query, insertion, update, and deletion permissions of tables, views, and functions under the database. If you select multiple databases, you can only set database permissions. In advanced mode, you can set column-leve permissions. By selecting a single data table, you can add query permission for columns. You can also select one or more columns or all columns to grant permissions.



Add permiss	ion			2
Catalog	DataLakeCata	g 🔹		
Setting mode	Standard	Advanced		
Database	Enter			
	When selecting a selecting more the	single database, you can continue In one databases, you can only se	to set permissions for tables, views, t t permissions at the database level.	functions, and columns; but whe
Authorizable	Yes			

Row-level permissions

Based on database and table permissions, add row-level filter expressions to restrict the access scope.

Add permissic	n	×
Permission type	Row-level filtering Row-level filtering allows you to set row-level permissions for a specified table to filter data that is accessible.	
Catalog	DataLakeCatalog	
Database	Enter	
Data table	Enter v	
Expression	Enter an expression	
	A row filter expression specifies the filter conditions. Example: year > 2010 and country != 'US'	

DLC User Management

Note:

Operation permission: Configured with the root account or a DLC administrator account.

Modifying User Permissions

Select a user in the user list and click Edit to manage the user permissions.

Viewing User Permissions

Click a user ID in the user list to enter the user details page.

Deleting User Permissions



Remove specified permissions: In the user list on the Permission Management page, click Edit to enter the details page for modification.

Delete users: In the user list on the Permission Management page, click Delete.

Work Group

Note:

Operation permission: Configured with the root account or a DLC administrator account.

To achieve better unified management and reduce administrators' management costs, you can create a work group to add users in batch and authorize them in a unified manner.

Creating a Work Group

Click **Work group** on the Permission Management page and click Add work group.

Adding Users to a Work Group

Method 1: When adding a user, select Bind to Working Group.

User Information	Select Tencent Cloud Users	(i) Create User (i)		
	User ID	Username	Description (Optional)	Operation
			0	
		Click the icon in the upper left corr	er to add the user to authorize.	
User type	 General user 			
ooor type			sions or he bound to a working group to access	resources
0001 () po	Users need to be assigned t	the corresponding database and table permise	sions of be bound to a working group to access	000010001

Method 2: On the Work group tab, click Edit to edit a work group.

Permissio	n managemen	t 🔇 Singapore 🔻					Operation Guide 💝 Permission	nanagement guide 🗵
User	Work group							
Adding	Sub-accounts an	d Assigning Permission	s to Databases	, Tables, and Engines				Hide ^
1 /	Add Tencent Clou	d Users as DLC users.			Configure database, table, and	engine permissions.		
1	dd DLC User Mar	nage Tencent Cloud Users			Bind to Working Group Authorize or	n User List Page 🚯		
() Bat	ch add users to a wo	rk group to batch grant them	the permissions of	f data, engines, and other resources of th	s work group. There is no need to add an	admin to a work group. For more p	permission guides, see here 🛂 .	×
Add work	group Batc	h delete				Enter a work group	o name	Q Ø
Work	Iroup ID	Work group name	User count	Description	Added by	Add time \$	Operation	
42069		test0001	4		100	2025-02-24 11:01:32	Edit Delete	
42068		qzzhu	15	qzzhu		2025-02-24 11:00:57	Edit Delete	
38521		test	0			2024-09-10 20:31:00	Edit Delete	
Total items:	3						10 v / page H 4 1	/1 page 🕨 🕨

Method 3: Add a user when creating a work group.

Adding Permissions to a Work Group

On the Work group tab, click Edit and add permissions.

Editing Work Group Members or Permissions

On the Work group tab, click Edit and manage the work group members or permissions.

Deleting a Work Group

On the Work group tab, click Delete for the work group.

FAQs

Why a failure occurs when I add the account as a DLC account?

If an error occurs when you add the account as a DLC account, first check whether any of the following issues occur. If not, please contact aftersales personnel for help.

- 1. Repeated addition: The account has already been added as a DLC account.
- 2. Account input error: An incorrect account is entered.

Why it displays "unavailable feature" or "no permission" when I enter the DLC console?

Your account may not have been added as a DLC account, or the administrator has not configured permissions for your account. You can contact your administrator to add your account as a DLC account and configure the relevant permissions for your account.

Which operations must be performed with the root account during the use of DLC?

The root account is responsible for activating the DLC service, creating Tencent Cloud sub-accounts, authorizing the policy QcloudDLCFullAccess for sub-accounts, designating the first DLC administrator, and granting financial permissions.

How are the permissions determined if the permissions of a user are inconsistent with those of the work group to which the user belongs?

The permissions are the union of user and work group permissions.

DLC Data Import Guide

Last updated : 2024-07-31 17:23:10

External Table Data Import via COS

DLC supports querying and analyzing data directly on COS without migrating the data. Therefore, you only need to import the data into COS to start using DLC for seamless data analysis, achieving complete decoupling of data storage and computation. Currently, it supports uploading in multiple formats such as orc, parquet, avro, json, csv, and text files.

Currently, COS offers a variety of data import methods. You can choose from the following methods based on your situation.

log in to COS and proceed with file upload directly. For related operating steps, see Uploading an Object.

Import data using various upload tools provided by COS. For a list of supported tools, see Tool Overview.

Import data using SDKs or APIs provided by the COS service. For service-related instructions, see Upload Interface Documentation.

If you need to analyze logs from CLS, you can directly deliver logs to COS by partition and then analyze and query directly through DLC. For related operations, see Using DLC (Hive) to Analyze CLS Logs.

If you need to import data from other cloud services (such as database CDB, etc.) into COS, you can use DataInLong to perform the import. When creating a data synchronization link, select the cloud service to export from as the data source and choose COS as the destination to complete the data import.

If you encounter any issues during data import, you can consult us for a solution by Submitting a Ticket. After importing data into COS, you can perform SQL queries through the DLC console, API, or SDKs, enabling table creation, analysis, and export of results. For detailed operations, see Quick Start with Data Analytics in Data Lake Compute.

Data import into native tables

To provide better data query performance, DLC also supports importing data into native tables for query analysis. DLC native tables are arranged in the Iceberg table format, optimizing data during the import process. If you have the following use cases, it is recommended to use native tables for data query analysis.

In data warehouse analysis scenarios, aiming to leverage the Iceberg index for better analytical performance. If there's a need to update data, the DLC service supports performing UPSERT operations through SQL or data jobs. Data is written or updated in real-time through DataInLong, Flink, SCS, Spark Streaming, with concurrent reads and writes, requiring transactional guarantees for data processing business. Wishing to utilize Iceberg table features, such as time travel, multi-version snapshots, hidden partitions, partition evolution, and other advanced data lake features.

If you need to import data into a native table, you can choose one of the following methods based on your situation. Directly import through the DLC console.

Caution

When importing data through the console, there are certain restrictions, mainly for rapid testing and it's not recommended for production use.

If your original data is in services like MySQL or Kafka and you need to write or update MySQL binlog and message middleware data to DLC in near real-time, this can be achieved through DataInLong DataInlong's real-time import capability. Or through SCS, Flink writing. For operational guidance, you can contact us through a Work Order. If the original data is in data services such as MySQL, Kafka, MongoDB, etc., offline synchronization tasks by DataInLong DataInLong can be used to transfer data to native tables. During the data warehouse modeling process, external tables are used as the source layer of original data. In the process of transferring data to native tables, business-specific data distributions can be reorganized through building sparse indexes, etc., to achieve excellent query analysis performance of native tables. If guidance is needed, you can Contact Us.

Use SQL statements SELECT INSERT to query the data from the external table and then write it into the native table. For example, after creating a native table in DLC with the same table structure as the external table, the transfer can be completed by executing SQL syntax with the SparkSQL engine. Syntax example is as follows:

--- External table name: outtertable, Native table name: innertable insert into innertable select * from outtertable

If you encounter any issues during data import, you can consult us for solutions by submitting a work order.

Multiple data sources federated query analysis

If you do not wish to export data to the native tables of COS or DLC, DLC also offers the capability of data federation query analysis. It supports rapid association and analysis of data from multiple data sources through SQL without relocating data. Currently, it supports a variety of data sources including MySQL, SQLServer, clickhouse, PostgreSQL, EMR on HDFS, and EMR on COS.

When using federated analysis, it is necessary for the data source and data engine to be on the same network, ensuring network connectivity. Management can refer to Engine Network Configuration.

When querying EMR data through DLC federated analysis, the query performance will be on par with or even exceed that of EMR, making it suitable for production environments. It allows for the full utilization of DLC's fully-managed elastic capabilities to reduce costs and increase efficiency without relocating EMR services.

Federated analysis enables quick unification and analysis of data from multiple data sources, providing a convenient method for data insights and rapid analysis. With the support of DLC's fully-managed elastic capabilities, it effectively

reduces the cost of use. It also supports the use of INSERT INTO/INSERT OVERWRITE syntax to write federated data into DLC native tables, completing data import.

When analyzing data from other data sources through federated analysis, since the computation process involves synchronizing data to the DLC for analysis, there is some performance loss compared to directly querying the original data sources. If high query performance is required, data can be imported into native tables for analysis. The operation can be seen in Data import into native tables.

Quick Start with Data Analytics in Data Lake Compute

Last updated : 2024-07-17 15:19:00

Data Lake Compute allows you to quickly query and analyze COS data. Currently, CSV, ORC, Parquet, JSON, Avro, and text files are supported.

With Data Lake Compute, you can complete data analysis queries on COS in just a minute. It currently supports multiple formats including CSV, ORC, PARQUET, JSON, ARVO, and text files.

Preliminary Preparations

Before initiating a query, you need to activate the internal permissions of Data Lake Compute and configure the path for query results.

Step 1: Establish the necessary internal permissions for Data Lake Compute.

Note

If the user already has the necessary permissions, or if they are the root account administrator, this step can be disregarded.

If you are logging in as a sub-account for the first time, in addition to the necessary CAM authorization, you also need to request any Data Lake Compute admin or root account admin to grant you the necessary Data Lake Compute permissions from the **Permission Management** menu on the left side of the Data Lake Compute console (for a detailed explanation of permissions, please refer to DLC Permission Overview).

1. Table Permissions: Grant read and write operation permissions to the corresponding catalog, database, table, and view.

2. Engine Permissions: These can grant usage, monitoring, and modification rights to the computation engine.

Note

The system will automatically provide each user with a shared public-engine based on the Presto kernel, allowing you to quickly try it out without the need to purchase a private cluster first.

For detailed steps on granting permissions, please refer to Sub-account Permission Management.

Step 2: Configure the path for query results.

Upon initial use of Data Lake Compute, you must first configure the path for query results. Once configured, the query results will be saved to this COS path.

1. Log in to the Data Lake Compute DLC console and select the service region.

2. Navigate to **Data Exploration** via the left sidebar menu.



3. Under the **Database and Tables** page, click on **Storage Configuration** to set the path for query results.

_	-		
٢	Data Explore S Guangzhou *		SQL syntax reference [d] Data explore
	Database Query ϕ +	Query-2023-12-07 • + •	🗘 Storage
Q	Catalog DataLakeCatalog 🔻	O Running Image: Base C: Refresh C: Refresh C: Refresh C	😂 - Select a default database 🔻 📑 Select a data engine
≣	Select a target database 🔹	1	
표			
E	▶ ⊜ a		
Ø			
۵			
	• 8.		
	•		
	•		
	• 🗧 -		
) 🛢 (
	• 🛢 :		
	· A.		
E		Query result	Run history Download his

Specify the COS path for storage. If there are no available COS buckets in your account, you can create one through the Object Storage Console.

anaged storage	✓ Enable		
aged storage type	General bucket		
ery result storage path 🛈	Managed storage	User-defined storage	
	Select a COS path		
	Storage on COS requires	COS permissions, and fees in	curred will be determined by COS.

Analysis Steps

Step 1. Create a database

If you are familiar with SQL statements, write the CREATE DATABASE statement in the query and skip the creation wizard.



- 1. Log in to the Data Lake Compute console and select the service region.
- 2. Select **Data Explore** on the left sidebar.
- 3. Select **Database & table**, click "+", and select **Create a database** as shown below:

Data Lake Compute	Data Explore S Guangzhou *	1 Click the "+" button SQL priter reference (2 Data exp
B Overview	Database Query 🗘 +	Query 2023-12-19 6 + + \$\$ \$2019
🕲 Data Explore	Catalog DataLakeCatalog Cre	🛤 3 database 🗋 Sene 🖸 Fernent 🛱 Formet 🛱
📃 Data Tasks	Select a target database	2. Click "Create a database"
∃‡ Data Management	* 8	
Data Jobs	🕨 🌐 Table	
🗘 Data Engines 🛛 🔹 •	► 88 View	
(a) Global ~	Function	
Configuration		
Data Operations *	 B lable Nove 	
	> lo view	

Enter the database name and description.

Data Lake Compute	Data Explore S Hong Kong 🔻		Create database
	Database Query 🗘 +	Query-2024 Draft × + •	Database name a
Data Explore			Database name * Enter a database name
			Description Optional
🧮 Data Scheduling	Select a target database	1	
≟⊧ Data Management	▶ ■a0		
🗐 Data Job	▶ ₿a1		
D -	Eat_database_f71a094a53ca4df4a83		
E lask History	Bauto_database_mrlk9		
 Insight Management 	eauto_database_obnrc		
	Coercion		
Engine Management	Edatabase1		
5 SuperSQL Engine	Edatabase10		
Standard Engine BETA	Edatabase100		
Engine Network	▶ ⊜database11		
Configuration	Edatabase12		
Ops Management	Edatabase13		
J Permission	Edatabase14		
Management	Edatabase15		
Storage	Edatabase16		
Configuration	Edatabase18		
İ Audit Log	Edatabase19		
Monitoring &	▶ ⊜database2		
	▶ ⊜database20		
	Edatabase21		
E	Adatahaca99	Query result	Confirm Cancel



4. After selecting an execution engine in the top-right corner, run the CREATE DATABASE statement.

	_		
Data Lake Compute	Data Explore S Hong Kong 🔻		Data explore gu
Cverview	Database Query 🗘 +	Que Draft x + •	🗘 Storage cor
E Data Explore	Catalog		🖲 Default database 🔻 F Select a data engine 🄻 📐 SuperSQL Syntax
≣ Data Scheduling	Select a target database	1 select 1	-
∃≑ Data Management			
🗐 Data Job	↓ €		
Task History) E		
	→ €		
Manayement	▶ € ∩		
Engine Management			
59 SuperSQL Engine			
Standard Engine BETA			
Engine Network	→ Ê I∎ ■		
Configuration	▶ ₽ I		

As shown in the picture below:

Databa	ase Query	φ+	Query	-2024	Draft 🗙	+	•				🗘 Storage confi
Catalog	DataLakeCatalog	▼	۲	B	G E) C	R N		Aa 🔻	■Default database ▼ Select a data engine ▼ SuperSQL Syntax ▼
Select a	target database	•	1	CREATE	DATABA	SE IF	NOT	EXISTS	"Data	aLakeCatalo	g"."demo2" COMMENT 'for demo_test'
• •											
► SE											
• 6											

For details, see Table Management.

Step 2. Create an external table

If you are familiar with SQL statements, write the CREATE TABLE statement in the query and skip the creation wizard.

- 1. Log in to the Data Lake Compute console and select the service region.
- 2. Select **Data Explore** on the left sidebar.



3. Select **Database & table**, select the created table, and right-click to select **Create external table**.

Note:

An external table generally refers to a data file stored in a COS bucket under your account. It can be directly created in Data Lake Compute for analysis with no need to load additional data. It is external, so only its metadata will be deleted when you run DROP TABLE, while your original data will remain.

Data Lake Compute	Data Explore S Guangzhou 🔻		SQL syntax reference 🖾 🛛 Data explore
B Overview	Database Query Ø	+ Query-2023-12-07 • + •	🗘 Storage
) Data Explore	Catalog DataLakeCatalog	▼ Partial run 🖾 Save 🗘 Refresh 🖨 Format 💀	🛢 Select a default database 🔻 🔚 Select a data engine
📃 Data Tasks	Select a target database	CREATE DATABASE IF NOT EXISTS 'DataLakeCatalog'.'test'	
Ξ‡ Data Management	* 🛢 :		
E Data Jobs	🕨 🌐 Table		
😚 Data Engines 🛛 🗸 •	▶ 88 View	reate native table	
Global	Function	reate external table	
Configuration	· 8.		
델 Data Operations 🔹	•		

4. Generate the table creation statement based on the wizard, and then complete the steps of setting the basic information, selecting the data format, editing the column, and editing the partition.

Step 1. Select the COS path of the data file (which must be a directory in a COS bucket but not a bucket itself). There is also a quick method to upload a file to COS. The operations require relevant COS permissions.

Step 2. Select the data file format. In the **Advanced options**, you can select automatic inference, and then the backend will parse the file format and automatically generate the table column information for fast column inference.

0	Data Explore 🕲 Guangzhou 🔻		Create databas	e
88	Database Query Ø +	Query-2023-12-07 • + •	Database name *	Enter a database name
Q	Catalog DataLakeCatalog 👻	O Running Image: Comparison of the state o	Description	Optional
≣	Select a target database	1		
H				
Ξ				
\odot				
٨				
) B.			
	• 🛢 :			
	· 🛢 :			
	• 🛢			
	* 🛢 z 💦 r			
	• 8			
	▶ ⊜ .			
	▶ ⊜ ¢ .			
	* 🛢 (🖈)			
_			_	
E	-	Query result	Confirm	Cancel

Note:

Structure inference is an auxiliary tool for table creation and may not be 100% accurate. You need to check and modify the field names and types as needed.



Create external	table			>
Data path *	Select a data path		Select a COS path	
Data format	Select a data format			
ata table name	Text file (Log, TXT, and others)			
escription	NOSL			
	PARQUET			
	AVRO			
ield info	Infer structure			
	Automatically infer the data structor modify the data structure.	ure based on the selected f	ile. Please confirm the data structure	info, or manually
	Field name	Field type	Field configuration	Operation
		No da	ta	
artitioning				
Confirm	Cancel			Show SQL

Step 3. Skip this step if there is no partition. Proper partitioning helps improve the analysis performance. For more information on partitioning, see <u>Querying Partition Table</u>.

Partitioning			
	Partition field	Partition type	Operation
	Enter	Select 💌	Insert Delete
	Add		

5. Click **Complete** to generate the SQL statement for table creation. Then, select a data engine and run the statement to create a table.



Database Query 🗘 +	Query-2023-12-07 • + •	🗘 Storage configurati
Catalog DataLakeCatalog 💌	Partial run 🗇 Save 🗘 Refresh 🛱 Format	🛢 Select a default database 🔻 🔚 Select a data engine 🍷 🗍 🚥
Select a target database	<pre>1 CREATE EXTERNAL TABLE IF HOT EXISTS 'DataLakeCatalog'.'ubi'.'demotest' (2 'C0' string, 3 'C1' string, 4 'C2' string, 5 'C3' string, 6 'C4' string, 6 'C4' string, 8) ROW FORMAT SERDE 'org.apache.haddoop.hive.serde2.OpenCSVSerde' WITH SERDEPROPERTIES ('separatorChar' - ',', 'quoteChar' - ''') STORED AS TE 8) ROW FORMAT SERDE 'org.apache.haddoop.hive.serde2.OpenCSVSerde' WITH SERDEPROPERTIES ('separatorChar' - ',', 'quoteChar' - ''') STORED AS TE </pre>	XTFILE LOCATION *cosn://amana *** >>

Step 3. Run the SQL analysis

After the data is prepared, write the SQL analysis statement, select an appropriate compute engine, and start data analysis.

Database Query Ø +	Query-2023-12-07	7				\$ Storage configuration
Catalog DataLakeCatalog *	Partial run	Complet 🔻 🛅 Save 🖸 Refresh	🛱 Format 🔤		😫 Select a default database 🔻 🔚 publi	c-engine(SuperSQL-P 1.0-public) *
Select a target database 💌	1 SELECT	* FROM `DataLakeCatalog`.`demo`.`tes	t_1`-LIMIT-10;			
• •						
▶ 🖨a						
▼ III Table						
⊞ test_1						
i test_2						
▶ 88 View						
Function						
• 8	Query result	Statistics				Run history Download history
• 🛢 :	Task ID SQL deta	ails Export Suggestions 🗹				
	Query time 3.05	is Scanned data volume 2.7 KB Billable scanned vol	ume 34.0 MB 🚯			
• • • •	10 entries in total	I (up to 1,000 entries shown in the console)Copy I				
	id		pro_name	price	pro_date	
	12		product12	13.3	20230712	A
	6		product6	15.3	20230712	
	10		product10	14.3	20230712	

Sample

Write a SQL statement with all data query results being SUCCESS and run the statement after selecting a compute engine.

```
select * from `DataLakeCatalog`.`demo2`.`demo_audit_table` where _c5 =
'SUCCESS'
```

Quick Start with Permission Management in Data Lake Compute

Last updated : 2024-09-18 18:02:02

During the utilization of Data Lake Compute (DLC), if you need to establish varying access permissions for employees within your organization to achieve isolation of authority among them, you can employ the permissions management feature for meticulous management of user and workgroup permissions.

Note:

1. The policy of permissions is highly correlated with the usage of the product. It is recommended that administrators configure the policies for roles such as workgroups and sub-users in advance before officially utilizing the product features.

2. In different regions, administrators are required to reconfigure the member management and permissions management for DLC in that specific region.

CAM Authorization

Data Lake Compute (DLC) possesses a comprehensive data access permission mechanism. If you have sub-account management requirements, please grant the corresponding sub-account with the QcloudDLCFullAccess (Full read-write access to Data Lake Compute (DLC)) policy in the Access Management Console. For specific steps on creating sub-accounts and authorizing policies.

Data Lake Compute (DLC) offers permissions refined to the granularity of row and column levels in data tables, ensuring that you need not worry about overstepping authority with this operation.

Cloud Access Management	Policies					CAM Policy Instr
E Dashboard						
은 Users 🗸	 Associate users or user groups with policies to grant permissi 	ions.				
ዲ User Groups	Create Custom Policy Delete			All Policies Preset Policy	Custom Policies dlc	© Q
Dolicies						
[4] Roles	Policy Name	Service Type T	Description		Last Modified	Operation
🖻 Identity Providers 👻	QcloudDLCFullAccess	Data Lake Compute	Full read-write access to Data Lake Compute (DLC)		2021-09-22 16:37:49	Associate User/User Group/Ro
Federated Account					2023-10-18 16:31:33	Associate User/User Group/Ro
(ছ) Access Key 👻			A-1-1		2023-10-18 16:31:23	Associate User/User Group/Rc
	0 selected, 3 in total				10 👻 / page	H 4 1 /1page ►

Users and Workgroups

DLC manages user permissions through two methods: user authorization and workgroup binding authorization. **User:** Refers to users in CAM, including administrators, sub-accounts, and collaborator accounts. **Workgroup:** DLC allows a group of users to be bound to a workgroup, granting the group access to data, engines, and other resources. This enables batch management of user permissions, ensuring that all users within the same workgroup have the same level of access.

Note:

When a user's individual permissions differ from the permissions of the workgroup they belong to, the combined permissions will be the union of both sets.

By default, regular users created by an administrator do not have any permissions. To grant permissions, users should be added to a workgroup, and appropriate permission policies should be assigned to the workgroup, allowing the users within it to acquire the necessary permissions.

Adding a User

Data Lake Compute utilizes the Tencent Cloud account ID as the default user ID. It distinguishes between two user types: administrators and ordinary users. Administrators inherently possess all resource permissions, while ordinary users must be granted specific permissions or be associated with a work group to acquire permissions.

1. Incorporate a user and associate them with a work group.

Log into the DLC console, select Permission Management, and click on Users > Add User to incorporate a new user.

0	Permission management	🔇 Guangzhou 👻					Use guide Ø Permission managemen
	User Work group						
Q 7-1	Both sub-account and coordin	nator users need to be granted with permissio	ons to use data, engines, and other resources. A use	r may be associated with one or more work gr	oups to inherit all of their permissions. An admin us	er has all resource permissions. For more permission	nguidelines, see here .
Ē	Add user Batch delete						Enter a user ID or name
Ŷ	User ID	Username	User type 🚯	Description	Added by	Add time \$	Operation
()	•					2023-12-06 17:35:24	Edit Authorize 🔻 Delete
						2023-03-07 16:56:41	Edit Authorize 🔻 Delete
	Total items: 2					10	▼ / page H < 1 /1 page

2. Enter the basic information: Provide the user ID, user name, and description, and select the user type.

Note:

When selecting the user type as "Ordinary User", permissions can be obtained through individual authorization or by acquiring all permissions of a specified work group. When selecting "Administrator" as the user type, there is no need to associate with a work group to gain all permissions.

Data Lake Compute	÷	Add use	er	
Overview				
E Data Explore	1	Basic i	nfo > (2) Bind	d work group
Ξ‡ Data Management				
🗐 Data Jobs	User	r ID		
🕅 Data Engines 🗸 •	User	name	Enter a username	0
Global ^ Configuration	User	type		•
L. Permission			Admin s fe	for all resources (including data and engines), and can manage other admins except the root account user. A general user needs to be granted with relevant permissions or associated with a work group to access corresponding resources.
Management	Desc	ription	General user	
 Storage Configuration 				
델 Data Operations 🔹				

3. Associate with a work group: Select a work group for association (optional).

Add user				Bind work g	group						
				Select a work	group			:	Selected (0)	
Basic info > 2 Bind	work group			Enter an ID o	or name		Q,		ID	Name	Description
Pind work aroun	Binding a user to a work group grants the user all permissions on the group			D ID	Name	Description					
bild work group						No data yet					
Work group ID	Work group name	Description A	ld tim								
								↔			
		_									
			lata								
				4			•				

User authorization

In the user list, authorize each user individually. The authorization includes "Data Permissions" and "Engine Permissions", and the permission policy is consistent with the work group's permission policy.

ser Work group						,
- Work group						
Both sub-account and coord	rdinator users need to be granted with permissi	ons to use data, engines, and other resources. A us	er may be associated with one or more work gr	oups to inherit all of their permissions. An admin u	ser has all resource permissions. For more permiss	ion guidelines, see here 🗹 .
dd user Batch delete						Enter a user ID or name
User ID	Username	User type 🚯	Description	Added by	Add time 🗘	Operation
1					2023-12-06 17:35:24	Edit Authorize 🔻 Del
			12011		2023-03-07 16:56:41	Edit Authorize 🔻 Del

Add Work Group

1. In the Data Lake Compute DLC, select Permission Management from the left sidebar, and click on Work Group > Add Work Group to create a work group for the user. When creating a work group, you can choose to bind it to a user



or create an empty work group. For detailed operations, refer to Users and User Groups.

Batch remove Batch remove Vork group ID Work group ID	Permission management
Add work group Easter remove Enter a work Work group ID Work group name User count Description Added by Add time ‡ Operation	
Work group ID Work group name User count Description Added by Add time \$ Operation	group name
■ No data	

2. Enter the basic information: Provide the work group name and description.

Add work g	roup
1 Basic info	> 2 Bind user
Work group name	Enter a work group name
Description	Enter a description

3. Associate a user: The associated user will acquire all permissions under the respective work group.

← Add work group					
Sasic info >	2 Bind user				
Bind user Batch remov	e An associated user will obtain all permission of this work group				
Username	User type	Description	Add time 🕈	Added by	Operation
			No data		
Total items: 0				10 🔻 / page 🛛 H	✓ 1 /1 page → H

Granting permissions to a work group



After creating the work group, click on the Authorize operation in the list to add permissions to the work group, including Data Permissions and Engine Permissions.

Permission manager	nent 🕲 Guangzhou	Ŧ				Use guide 🧭 🛛 Permission manageme
User Work grou	2					
 Batch add users to 	a work group to batch grant the	m the permissions o	f data, engines, and other resources of th	is work group. There is no need to add an admin to a work group. For mo	re permission guides, see here 😰 .	
Add work group	Batch remove					Enter a work group name
Work group ID	Work group name	User count	Description	Added by	Add time \$	Operation
30635	test	0		100006728148	2023-12-07 15:37:30	Edit Authorize 💌 Remove
Total items: 1						10 v Data permission / 1 page Engine permission
Total items: 1						10 v Engine permission

Data permission

Data permissions include:

Data Catalog Permissions: These include two types of permissions under the data catalog, namely, the ability to Create Database and Create Data Catalog.

0	÷	Grant data permissions						Add permis	sion
		Basic info Work group name test Description Catalog/Database/Table Add permission Batch repostes Permission type Ca	talog	Database	Table/View/Function	0	Column	Permission typ	C Catalog Database & table The catalog option covers permissions to create databases under DataLakeCatalog and other catalogs, while the data table option covers permissions of databases, data tables, views, and functions. Create database under DataLakeCatalog Create catalog Ves
		Total items: 0							
		Add permission Eatch reposses Permission type Permission type	Catalog	ſ	Jatabase		Data table		
E	4					=	No data	Confirm	Gined

Database Table Permissions: Fine-grained permissions at the database table level can be granted, including query and edit permissions for databases, tables, views, and functions.

0	÷	Grant data pern	nissions						Add permiss	ion
# © !!! # D		Basic info Work group name 1231 Description 1231	2 23						Permission type Catalog Setting mode	Catalog ODstabase & table The catalog option covers permissions to create databases under DataLakeCatalog and other catalogs, while the da table option covers permissions of databases, data tables, views, and functions. DataLakeCatalog T Standard Advanced
E (> (6)		Catalog/Database/Tal	ble Batch repossess		Databara			Column	Database	Select a database/view/function Selected (3) Enter a database name
삗							0	No data		▼ ● ● ■
		Total items: 0 Row-level permission	5							 . ⊂ e³ . ⊂ e⁴ . ⊂ e⁴ . ⊂ e⁴ . ⊂ e⁴
		Add permission		Catalog		Database		Data table	Permission	Covery analysis Covery analysis Covery analysis Covery data Covery data
E	4						=	No data	Confirm	Cancel

Engine permission

Select a data engine and grant the permissions to use, modify, or delete it.

Grant engine permissions			Add permission	Add permission		
asic info			Data engine	All 🕲		
fork group name 12312			Engine permission	All		
escription 123123				✓ Use Modify Operation Monif Delete		
ermission info			Authorizable	Yes		
Add permission Batch repossess						
Name	Permission (j)	Authorizable 🛈				

Engine operation permissions are granted automatically

DLC supports default enablement of engine operation class permissions. Once enabled, all users will by default have the following permissions for that engine: Utilize: Execute tasks using this engine. Operation: Initiation of engine suspension or standby. Monitoring: Administration of engine usage monitoring. **Note:**

©2013-2025 Tencent Cloud International Pte. Ltd.

1. Upon termination, administrators inherently maintain all engine privileges. Ordinary users require an administrator to add permissions on the permission management page.

2. Existing ordinary user permissions will remain intact and can be deleted on the Permission Management page.

3. Subsequent newly created ordinary users have no usage rights, which should be manually added on the Permission Management page.

How do I enable or disable the self-delivery authorization engine

By default, the engine enables/disables two operation permission entries:

Access 1: Engine Purchase page > Advanced Configuration Items



Access 2: Go to the SuperSQL engine page and click Edit Auto-granting of engine permissions.

හි	Tencent Cloud Overview	Products •					M Ticket 🔻	Billing (Center ▼ English ▼
٢	SuperSQL engine	Hong Kong 🔻							SuperSQL engine
∷	 Data Lake Compute offers can be billed on a pay-as-y no fees charged on it after 	both public and p you-go basis or su suspension. For c	rivate data engines. A public data en ibscribed monthly. For more billing in operations and notes, see Managing	igine is managed by Data i ifo, see Billing Overview I Private Data Engines 옵	Lake Compute and billed by s Z . A pay-as-you-go data eng	scanned data volum gine can be configur	e, with no operation or permission red ed with the auto-suspension or sched	quired; a p luled susp	rivate data engine vension policy, with
	Create resource Bill query	Z Renewal mar	agement 🖸				Select a resource tag or enter keyw	ord(s) (ser	parate two
e	Engine Name/ID	to-renewal	Start and stop policy	Cluster description	Auto-granting of en	Engine Size	Network configuration	Created	Operation
A 	dī ⊿dī	No	Manual start, Manual suspension	Private engine	No di	16CU Standard 1-2 cluster(s)		2024-07	Monitor Spec configuration Parameter Configuration More ▼
9 (ும் பட	-	Auto-start, Manual suspension	Private engine	No /	16CU Standard 1-5 cluster(s)		2024-06	Monitor Spec configuration Parameter Configuration More ▼
 o*	-n		Manual start, Manual suspension	Public engine	No 🎤	10 M		2022-08	Monitor Spec configuration Parameter Configuration More ▼
	Total items: 3						10 v / page	H	1 /1 page

After setting engine permissions, click Confirm.

Tencent Cloud	Overview Products • +						☑ Ticket ▼ Billing Center ▼ English ▼		
Data Lake Compute	SuperSQL engine	Hong Kong 🔻				Set engine	permissions		
 Overview Data Explore Data Scheduling 	Data Lake Compute offers required; a private data en the auto-suspension or sch	both public and priv gine can be billed on neduled suspension p	ate data engines. A public data engin a pay-as-you-go basis or subscribe policy, with no fees charged on it afte	he is managed by Data Lak d monthly. For more billing r suspension. For operatio	e Compute and t info, see Billing (ns and notes, see	Engine name Resource ID Auto-			
∃⊧ Data Management	Create resource Bill query	Z Renewal manag	ement 🗹			granting of engine permissions	If this option is disabled, admin users can use this engining eneral users can use the engine only after being added		
🗐 Data Job	Engine Name/ID	Auto-renewal	Start and stop policy	Cluster description	Auto-grantin		Permission Management page. The permissions of general users existing before the disc operation are not affected, and these users can be deleti-		
 Task History Insight Management 	自动化专用常稳拔测_勿用 币 DataEngine-iwxhwnud 币	裁测_勿用		Private engine No 🧨			the Permission Management page. General users created later have no permission to use th engine, and need to be added on the Permission Manag page to use the engine.		
Engine Management SuperSQL Engine Standard BETA	at_data_engine_presto li DataEngine-p3d2xfq1 fi		Auto-start, Manual suspension	Private engine	No 🥜				
Engine Network Configuration	public-engine [] DataEngine-public-1313074 []	-	Manual start, Manual suspension	Public engine	No 🌶				
Ops Management	Total itoms: 3								
	iota iteliis. 3					Confirm	Cancel		

Quick Start with Partition Table

Last updated : 2024-07-17 15:25:14

Data Lake Compute Partition Table

With the partition catalog feature, you can store data with different characteristics in different catalogs. In this way, when exploring data, you can filter data by partition through the where condition. This greatly reduces the scanned data volume and improves the query efficiency.

Note:

Partitions in the same table should adopt the same data type and format.

Internal tables in Data Lake Compute are implemented as implicit partitions, so you don't need to care about the partition catalog structure.

Creating a Partition Table

Specify the partition field through the PARTITIONED BY parameter in the table creation statement. Example: Creating the test_part partition table

```
CREATE EXTERNAL TABLE IF NOT EXISTS `DataLakeCatalog`.`test_a_db`.`test_part` (
`_c0` int,
`_c1` int,
`_c2` string,
`dt` string
) USING PARQUET PARTITIONED BY (dt) LOCATION 'cosn://testbucket/data/';
```

Adding a Partition

Adding a partition through ALTER TABLE ADD PARTITION

If your data partition catalog uses the Hive partitioning rule (partition column name=partition column value), the rule can be used to add partitions. The catalog is organized as follows:



```
ALTER TABLE `DataLakeCatalog`.`test_a_db`.`test_part` add PARTITION (dt =
'202206')
ALTER TABLE `DataLakeCatalog`.`test_a_db`.`test_part` add PARTITION (dt =
'202207')
ALTER TABLE `DataLakeCatalog`.`test_a_db`.`test_part` add PARTITION (dt =
'202208')
ALTER TABLE `DataLakeCatalog`.`test_a_db`.`test_part` add PARTITION (dt =
'202209')
ALTER TABLE `DataLakeCatalog`.`test_a_db`.`test_part` add PARTITION (dt =
'202210')
```

Adding a partition by specifying the location through **ALTER TABLE**

If your data adopts a general COS catalog (not in the "partition column name=partition column value" format), you can specify a catalog when adding a partition.

Sample SQL:

```
ALTER TABLE `DataLakeCatalog`.`test_a_db`.`test_part` add PARTITION (dt = '202211') LOCATION='cosn://testbucket/data2/202211'
ALTER TABLE `DataLakeCatalog`.`test_a_db`.`test_part` add PARTITION (dt = '202212') LOCATION='cosn://testbucket/data2/202212'
```

Automatically adding a partition through MSCK REPAIR TABLE

Use the MSCK REPAIR TABLE statement to scan the data catalog specified during table creation. If there is a new partition catalog, the system will automatically add the partitions to the metadata of the data table. Sample SQL:

```
MSCK REPAIR TABLE `DataLakeCatalog`.`test_a_db`.`test_part`
```

We recommend you use ALTER TABLE to add a partition preferably, as automatic adding through MSCK REPAIR TABLE has the following restraints:

MSCK REPAIR TABLE only adds partitions to the metadata of the data table but does not delete them.



MSCK REPAIR TABLE is not recommended if the data volume is large, as it will scan all the data, which may

cause a timeout.

If your partition catalog doesn't use the Hive partitioning rule (partition column name=partition column value), MSCK

REPAIR TABLE cannot be used.

Enabling Data Optimization

Last updated : 2024-07-31 17:23:30

In big data scenarios, frequent fragmented writes generate a large number of small files, which significantly slow down performance. Based on extensive production practice experience, DLC offers you efficient, simple, and flexible data optimization capabilities that can handle near real-time scenarios with large data volumes.

Note:

1. In Upsert scenarios, a large number of small files and snapshots will be generated. You need to configure data optimization before writing to avoid the need for extensive resource processing of historical backlog of small files after writing.

2. Currently, data optimization capability only supports DLC native tables.

3. The initial execution of data optimization tasks may be slow, depending on the stock data volume size and the selected engine resource specifications.

4. It is recommended to separate the data optimization engine from the business engine to avoid the situation where data optimization tasks and business tasks compete for resources, causing delays in business tasks.

Configure data optimization through the DLC console

DLC data optimization strategies can also be set in the data directory, database, and data table. When data optimization strategies are not specifically set for a database or data table, they will inherit the optimization strategy from the previous level. When configuring data optimization, users need to select an engine. To execute data optimization tasks, if the user currently does not have a data engine, they may refer to Purchasing Dedicated Data Engine to make a purchase. DLC data governance supports Spark SQL Engine and Spark Job Engine.

1. If a user chooses the Spark Job Engine as the data optimization resource, DLC will create data optimization tasks on that engine. Depending on the size of the cluster, the optimized task data created will vary. For instance, if the cluster size is smaller than 32 CU, one data optimization tasks will be created to execute all optimization tasks. If the cluster size is larger than 32 CU, two data optimization tasks will be created to separately execute write optimization and data deletion optimization.

2. When choosing a Spark Job as a data optimization resource, some resources need to be reserved. If the optimization tasks queue exceeds 50, DLC will launch temporary data optimization tasks to quickly process the backlog of optimization tasks.

Data Directory Configuration Steps

You can use DLC's Data Catalog Editing Feature to configure data optimization capabilities for your data directory.



1. Go to the Data Management Module in the DLC Console, enter the **Data Management** page, and click **Data Optimization**.

Data Lake Compute	Data management	🔇 Hong Kong 👻							
# Overview	Catalog Database	Bucket list Function							
E) Data Explore									
📃 Data Scheduling	 Perform creation, edit, de 	eletion, or other management operations o	n catalogs.						
	Create catalog Select a	connection type	All Last 7 days	Last 30 days Select	date Select date 苗			Enter a	name
🗉 Data Job	Catalog name	Connection type	Connection info	Status	Creator	Created at \$	Update time \$	Connectivity status	Operation
Task History	DataLakeCatalog	DataLakeCatalog	-	Created successfully		-	-	-	Data optimization
									Function Management Test
Engine Management	Total items: 1							10 v / page	 ✓ 1 /1 page

2. Open the **Data Optimization** page of the data directory, configure the corresponding data optimization resources and policies. Once confirmed, the data optimization feature will automatically apply to that data directory.

Edit database									
Governance * S resources Only	elect a da y SparkSC	ta engine QL engines a	are supp	▼ oorted					
Governance	Smart (i)	🔵 Cus	tom (j)						
Governance rules 🔺									
Verge small files 🛈									
∕lin file count	_	5	+		Max size of file to	-	128	+	М
Scheduling interval	_	60	+	Minute	merge				
Data files threshold (i)	_	20	+		Delete files threshold (i)	_	20	+	
Equality delete files hreshold (j)	-	1000	+		Position delete files threshold (i)	-	1000	+	
Delete expired anapshots (j)									
Retained snapshot count	-	5	+		Deletion time slot	-	2	+	Day
									.

Note:

Only supports configuring data optimization for the DataLakeCatalog data directory.



Database Configuration Steps

If you want to configure a data optimization strategy for a specific database individually, you can use the database editing capabilities of DLC to configure data optimization capabilities for the database.

1. Enter the DLC console Data Management Module, enter the **Database** page, enter the database list under DataLakeCatalog.

Data Lake Compute	Data management 🔇 Gu	angzhou 🔻				Use guide 🧭	Data Managen
E Overview	Catalog Database Buc	ket list					
Data Explore	 This module allows you to mana 	oe databases under different cataloos. You can click the name of a database to manage its ta	ibles, views, and other data objects, or to manually import	t data to tables. Learn more 😫 . Data operations require relevant data permissions. For more per	mission auide, see	a here 🖾 .	
🔲 Data Tasks							
E‡ Data Management	Create database DataLakeCatal	og v				Enter a name	
🗉 Data Jobs	Database name ‡	Created at \$	Description	Creator	Operation		
🗇 Data Engines 🔹 🔹 •		2023-12-19 14:45:31			Edit Delete		
Global * Configuration		2023-10-26 17:51:31		-	Edit Delete		
Data Operations *	a marke fa	2023-10-26 17:51:30			Edit Delete		

2. Open the database page, click **Data Optimization Configuration**. Once confirmed, the data optimization strategy will automatically apply to that database.

Note:

When creating a database and editing data, the default to show data optimization strategy inherits the data optimization strategy of the superior data directory. If you want to customize the data optimization strategy, you need to select **Custom Configuration** and configure data optimization resources and policies.

Data Table Configuration Steps

If you want to configure a data optimization strategy for a specific data table individually, you can use the data table editing capabilities of DLC to configure data optimization capabilities for the data table.

1. Enter the DLC console Data Management Module, enter the **Database** page, select a database, then enter the **Data Table** list page, and click **Create Native Table**.

Data Lake Compute	Catabase / mv.test3	
III Overview	Data table View Function	Task history Storage
Data Explore Data Tasks	O Data tables under the database. You can manage basic info, fields, and other info in the native and external tables, and import data from local system or COS and export data to COS asynchronously. You can view the task running details in the task history. For the billing mode of the	native table, see Billing Overview 🛂 .
표 Data Management	Create native table Create external table Select a table type 💌 Update time All Last 7 days Last 30 days Select date Select date 🗂 Ratch deele	Enter a name
E Data Jobs	Data table name \$ Table type Rows \$ Table size \$ Governance status Created at \$ Update time \$ Creator De	scription Operation
🕅 Data Engines 🔹 •		
In the second		
Data Operations *		
	Total items: 0	10 💌 / page 🛛 4 🔄 1 🔢 / 1 page

2. Open the Create Native Table page, configure the corresponding optimization resources, and once confirmed, the data optimization strategy will automatically apply to that data table.

3. For already created tables, you can click **Data Optimization Configuration** to edit the existing data table's data optimization strategy.

Note:

When creating or editing a data table, the default data optimization strategy displayed inherits from the parent data table's data optimization strategy. If you want to customize the data optimization strategy, you need to select **Custom Configuration** and configure data optimization resources and policies.

Optimize data through attribute field configuration

Besides the above visualization method for configuring data optimization, you can also manually specify library and table field attributes for configuration. For example:

```
// for table govern policy
ALTER TABLE
   `DataLakeCatalog`.`wd_db`.`wd_tb`
SET
TBLPROPERTIES (
    'smart-optimizer.inherit' = 'none',
    'smart-optimizer.written.enable' = 'enable'
)
// for database govern policy
ALTER DATABASE
    `DataLakeCatalog`.`wd_db`
SET
DBPROPERTIES (
    'smart-optimizer.inherit' = 'none',
    'smart-optimizer.written.enable' = 'enable'
)
```

The attribute values for data optimization can be modified via the ALTER statement. The attribute value definitions are as follows:

Attribute Value	Meaning	Default Value	Value Description
smart-optimizer.inherit	Whether to Inherit from the Parent Strategy	default	none: Does not inherit default: Inherit
smart-optimizer.written.enable	Whether Write Optimization is Enabled	disable	disable: Not Enabled enable: Enabled

smart- optimizer.written.advance.compact- enable	(Optional) Advanced Write Optimization Parameters, Whether to Start Small File Merge	enable	disable: Not Enabled enable: Enabled
smart- optimizer.written.advance.delete- enable	(Optional) Advanced Write Optimization Parameters, Whether to Start Data Cleanup	enable	disable: Not Enabled enable: Enabled
smart- optimizer.written.advance.min- input-files	(Optional) Merge Minimum Number of Input Files	5	When the number of files in a table or partition exceeds the minimum number of files, the platform will automatically check and initiate file optimization merge. File optimization merge can effectively improve analyze query performance. The larger the minimum number of files, the higher the resource load. The smaller the minimum number of files, the more flexible the execution, and tasks will be more frequent. It is recommended to set the value to 5.
smart- optimizer.written.advance.target- file-size-bytes	(Optional) Merge Target Size	134217728 (128 MB)	During file optimization merge, files will be combined to meet the target size as much as possible. It is recommended to set the value to 128M.
smart- optimizer.written.advance.retain- last	(Optional) Snapshot Expiration Time, Unit Days	5	When the snapshot retention time exceeds this value, the platform will mark the snapshot as expired. The longer the snapshot expiration time, the slower the snapshot cleanup speed, and the more storage space is occupied.



smart- optimizer.written.advance.before- days	(Optional) Number of Expired Snapshots to Retain	2	Expired snapshots exceeding the retention count will be cleaned up. The more expired snapshots retained, the more storage space is occupied. It is recommended to set the value to 5.
smart- optimizer.written.advance.expired- snapshots-interval-min	(Optional) Snapshot Expiration Execution Cycle	600(10 hour)	The platform will periodically scan snapshots and expire them. The shorter the execution cycle, the more sensitive the snapshot expiration will be, but it may consume more resources.
smart- optimizer.written.advance.cow- compact-enable	(Optional) Enable Merge for COW Tables (V1 Table or V2 Non-Upsert Table)	disable	Once this configuration item is enabled, the system will automatically generate file merge tasks for COW tables. Note: COW tables usually have a large data volume, and file merging may consume a lot of resources. You can choose whether to enable file merging for COW tables based on resource availability and table size.
smart- optimizer.written.advance.strategy	(Optional) File Merge Strategy	binpack	binpack (default merge strategy): Merges data files that meet the merge conditions into larger data files using the append method. sort: The sort strategy merges files based on specified fields. You can choose query condition fields that are frequently used in your business scenarios as the sorting fields. Merging in this way can improve query performance.
smart- optimizer.written.advance.sort- order	(Optional) When the file merge strategy is sort, the configured sort collation	-	If you haven't configured a sorting strategy, the Upsert Table will sort using the configured upsert key values (by default, the first two key values) in an ASC NULLS LAST manner. If a sorting strategy cannot be found for COW Table during a sort merge,



			the binpack default merge strategy will be used.
smart- optimizer.written.advance.remove- orphan-interval-min	(Optional) Period for Removing Orphan Files	1440(24 hour)	The platform will periodically scan and clean up orphan files. The shorter the execution cycle, the more sensitive the cleanup of orphan files will be, but it may consume more resources.

Optimization Suggestions

The DLC backend regularly statistics native table metric items and combine these metrics with best practices to provide optimization suggestions for native tables. There are four categories of optimization suggestion items, including basic configuration for table usage scenarios, data optimization recommendations, and recommendations for data storage distribution items.

Optimization recommendation check items	Sub-check item	Meaning	Business Scenario	Optimization Suggestions
	Metadata governance enabled	Check whether metadata governance is enabled to prevent metadata volume expansion due to frequent table writes	append/merger into/upsert	Recommended to enable
Basic attribute configuration check of the table	Bloom filter set	Check if the bloom filter is set. After enabling the bloom filter for MOR tables, it quickly filters the deletes files, speeding up MOR table queries and deletes file merges	upsert	Must enable
	Metrics key attributes configured	Check if metrics are set to full. Once this attribute is enabled, it will record all metrics information, preventing incomplete metrics information recording due to excessively long table locations	append/merger into/upsert	Must enable



Data optimization configuration check	Small File Merge	Check if small file merging is enabled	merge into/upsert	Must enable
	Snapshot Expiration	Check if snapshot expiration is enabled	append/merge into/upsert	Recommended to enable
	Remove orphaned files	Check if removing orphaned files is enabled	append/merge into/upsert	Recommended to enable
Recent governance task check items	Recent governance task check items	If data governance is enabled, the system will track the execution of data governance tasks. If multiple tasks in a row time out or fail, it will be deemed in need of optimization	append/merger into/upsert	Recommended to enable
Data Storage Distribution	Average File Size	Collect summary information from snapshots, calculate the average file size, and if the average file size is less than 10MB, it will be deemed in need of optimization	append/merger into/upsert	Recommended to enable
	MetaData Meta File Size	Collect table metadata.json Meta File Size, if the file size exceeds 10MB, it will be deemed in need of optimization	append/merger into/upsert	Recommended to enable
	Number of Table Snapshots	Collect Number of Table Snapshots, if the number of snapshots exceeds 1000, it will be deemed in need of optimization	append/merger into/upsert	Recommended to enable

Optimization Suggestions for Basic Configuration Items of Table Attributes

Check and configure Metadata Governance Method

Step1 Inspection Method

Use 'show TBLPROPERTIES' to view table attributes and check if "write.metadata.delete-after-commit.enabled",

"write.metadata.previous-versions-max" are configured.

Step2 Configuration Method

If Step1 finds that it's not configured, you can configure it using the following Alter table DDL, with the method referenced below.

```
ALTER TABLE
  `DataLakeCatalog`.`axitest`.`upsert_case`
SET
  TBLPROPERTIES(
    'write.metadata.delete-after-commit.enabled' = 'true',
    'write.metadata.previous-versions-max' = '100'
);
```

Note:

To enable automatic metadata governance, "write.metadata.delete-after-commit.enabled" should be set to true. The number of historical metadata to retain can be set according to the actual situation, for example, setting "write.metadata.previous-versions-max" to 100 will retain up to 100 historical metadata.

Inspecting and Setting Bloom Filter Method

Step1 Inspection Method

Use show TBLPROPERTIES to view table attributes, and check if "write.parquet.bloom-filter-enabled.column. {column}" is set to true.

Step2 Configuration Method

If Step1 finds that it's not configured, you can configure it using the following Alter table DDL, with the method referenced below.

```
ALTER TABLE
  `DataLakeCatalog`.`axitest`.`upsert_case`
SET
  TBLPROPERTIES(
    'write.parquet.bloom-filter-enabled.column.id' = 'true'
);
```

Note:

It is recommended to enable bloom in upsert scenarios, and configure it based on the upsert primary key. If there are multiple primary keys, it is advisable to set it for the first two primary key fields.

After updating the bloom fields, if there are upstream writes from inlong/oceans/flink, you must restart the upstream import job.

Check and configure table key attributes metrics

Step1 Inspection Method

View table properties using `show TBLPROPERTIES` and check if "write.metadata.metrics.default" is configured as "full".

Step2 Configuration Method

If Step1 finds that it's not configured, you can configure it using the following Alter table DDL, with the method referenced below.

```
ALTER TABLE
  `DataLakeCatalog`.`axitest`.`upsert_case`
SET
  TBLPROPERTIES('write.metadata.metrics.default' = 'full');
```

Data Optimization Configuration Recommendations

Step1 Inspection Method

Check using SQL

View table properties using `show TBLPROPERTIES` and check if data optimization is configured. Refer to DLC Native Table Core Capabilities for the attribute configuration values for data optimization.

Visual inspection through the DLC Console

Go to the Data Management Module in the DLC Console, enter the **Database** page, select a database to access the **Data Table** list page, choose the table to inspect, and proceed to **Data Optimization Configuration**.

Step2 Configuration Method

Follow the guidance to enable data optimization.

Recent recommendations for data governance optimization task items

Check if data governance is functioning properly

Step1 Inspection Method

Enter the DLC Console Data Management Module, enter the **Database** page, select a database and then enter the **Data Table** list page, click on the data table name, enter **Optimized Monitoring**, choose **Optimization Task** then select **Today's Optimization**, check for tasks that failed in the last three hours, if there are any, the check is not passed. Select the failed task, in **View Details** look at the **Execution Results**.

Step2 Fix Methods

Summary of Reasons and Solutions for Failed Scenario Data Optimization Tasks.

1. Data Governance Configuration Error led to failure.

Sort Merge Strategy was enabled, but the collation was incorrectly configured, or a nonexistent rule was set.

The configuration for the data governance engine has changed, leading to the inability to find an appropriate engine when running governance tasks.

2. Task Execution Timed Out.

Note:



After repairing the recent data optimization task performance, it is necessary to wait three hours before checking if it has recovered.

Data Storage Distribution Item Optimization Suggestions

Note:

Failure in this scenario check is usually due to large data volume. It's recommended to handle it manually before considering addition to Data Optimization Governance.

It is recommended to use the more efficient Spark job engine.

When manually merging small files, configure the target-file-size-bytes parameter based on the business scenario. For upsert operations, it is advised not to exceed 134217728, i.e., 128M. For append/merge into operations, it is advised not to exceed 536870912, i.e., 512M.

When using the Spark job engine to handle snapshot expiration, the max_concurrent_deletes parameter can be increased.

Average Data File Size Check Failure Handling Method

Step1 Summary of Reasons

The average size of data files is too small, usually occurring in the following scenarios:

The table is partitioned too finely, resulting in each partition having only a small amount of data.

When tables are written using the Insert into/overwrite method, the upstream data is dispersed, such as when the upstream data is also from a partitioned table with little data within partitions.

The table is written to the MOR Table using the merge into method, but small file merging has not been performed. The table is written using the upsert method, but small file merging has not been performed.

Step2 Fix Methods

Refer to the following SQL to manually perform small file merging.

```
CALL `DataLakeCatalog`.`system`.`rewrite_data_files`(
  `table` => 'test_db.test_tb',
  `options` => map(
    'delete-file-threshold',
    '10',
    'max-concurrent-file-group-rewrites', --Subject to actual resource conditions,
    '5',
    'partial-progress.enabled',
    'true',
    'partial-progress.max-commits',
    '10',
    'max-file-group-size-bytes',
    '10737418240',
    'min-input-files',
    '30',
    'target-file-size-bytes',
    '134217728'
```

)

MetaData Meta File Size Check Failure Handling Method

Step1 Summary of Reasons

MetaData file size is too large, usually caused by an excessive number of data files, mainly due to the following reasons:

The table has been written to using the append method for a long time, and each write involves a large number of scattered files.

The table has the attributes of an MOR table and has been written to long-term using the merge into method, but small file merging is not enabled.

The table has not undergone snapshot expiration for an extended period, maintaining a large number of historical snapshot data files.

The table partitions are large, and each partition contains a large number of small files.

Step2 Fix Methods

Refer to manually perform small file merging.

Refer to the following SQL to manually execute the expired snapshot SQL and clean up historical snapshots.

```
CALL DataLakeCatalog.system.rewrite_data_files(
  table => 'test_db.test_tb',
 options => map(
    'delete-file-threshold',
    '10',
    'max-concurrent-file-group-rewrites', --The higher the concurrency, and the fa
    '5',
    'partial-progress.enabled',
    'true',
    'partial-progress.max-commits',
    '10',
    'max-file-group-size-bytes',
    '10737418240',
    'min-input-files',
    '30',
    'target-file-size-bytes',
    '134217728'
  )
)
```

Based on the service scenario, the written files are aggregated to a certain extent to avoid scattered files.

If the data is written into insert into/insert overwrite, you can automatically add a repartition in either of the following ways.

1. This parameter takes effect when both of the following parameters are true. In this case, you can use the preceding parameters to control the number or size of automatically adapted partitions after repartition.

spark.sql.adaptive.enabled : This parameter must be true. The default value is true for cluster creation.

spark.sql.adaptive.insert.repartition : This parameter must be true. The default value is false for cluster creation.

2. Specify the following parameters to take effect. This case repartition spark. The partition number after SQL. The adaptive. Insert. The repartition. ForceNum the specified value.

spark.sql.adaptive.insert.repartition.forceNum : This parameter specifies the value of the partition to be partitioned. It is left blank by default when the cluster is created.

Check the number of snapshots. This operation fails to pass the check

Step1 Cause summary

Snapshots do not expire for a long time.

The upsert writes data to the checkpoint interval improperly, resulting in a large number of snapshots.

Step2 Repair method

See Snapshot expiration SQL to perform snapshot expiration operations.

Adjust the flink write checkpoint interval. It is recommended that the checkpoint interval of DLC native table upsert be 3 to 5 minutes.

Cross-Source Analysis of EMR Hive Data

Last updated : 2024-07-17 15:27:21

Data Lake Compute allows you to configure an EMR Hive data source for multi-source federated data analysis.

Preparations

Get the EMR Hive address.

Use an account with the permission to create data catalogs. For more information on permissions, see Permission Overview.

Creating an EMR Hive data source

- 1. Log in to the Data Lake Compute console and select the service region.
- 2. Select Data Explore on the left sidebar, click + in the Database & table column, and select Create data catalog.

Data Explore		ore 🔇 Gu	angzhou 🔻		
Data	oase	Query	¢+	Query-2023-12-1	1 •
Catalog	DataL	akeCatalog	Crea	te a database	
Select a	a target	database	Crea	te catalog	

3. Select **EMR Hive (HDFS)** for **Connection type** and select the target EMR instance. The VPC information will be populated by default after the instance is selected. **EMR versions supported by EMR Hive are 2.3.5, 2.3.7, 3.1.1, and 3.1.2.**

Note:

Relevant permissions are required for you to select the EMR Hive instance.

Create catalog		
1 Catalog configuration	> 2 Network configuration	
Connection type *	EMR Hive(HDFS)	
Connection name *	hdfs_demo	
Description	hdfs_demo	
EMR instance *	· · · · · · · · · · · · · · · · · · ·	
Data source VPC *	φ	
	vailable	
Ha setting *	HA Non-HA	
Hive version *	2.3.5	
Hive access address *		
Cluster name 🛞	Example: thrift://ip:port, metastore. The address can be queried in the EMR console	
Node 🛈 🔹 *		
Back		

4. Select the **Run cluster**. Currently, you can only select a private data engine of Presto. If there is no engine, create one on the **Data engine** page. For more information on the purchase process, see **Purchasing Private Data Engine**. **Note:**

The IP range of the selected data engine cannot be the same as that of the EMR instance; otherwise, a network conflict will occur, and you cannot query or analyze data. 5. Click **Confirm**.

Querying the EMR Hive data

After the data catalog is created, you can switch to it from the **Data catalog** menu on the **Data Explore** page.



At this point, you can query and analyze the data catalog with SQL statements.

Select the data engine bound when the data catalog is created and click **Run** to get the query result.

Note:

You can only query the data catalog with its bound data engine. To change the bound engine, click the set icon next to the data catalog.

Data	Explo	Guar	ngzhou 🔻		
Datab	oase	Query	¢ +		
Catalog	Wé		→		
Select a target database Modify configuration					
) 🛢 i		_	Delete		

Standard Engine Configuration Guide

Last updated : 2024-09-04 11:11:02

DLC offers two types of engines: the Standard Engine and the SuperSQL Engine. For a detailed comparison, see the table below or see the Data Engine Introduction. You can select the appropriate engine based on your specific business needs. If you choose the Standard Engine, you can follow the instructions in this document for configuration and usage.

Engine Types	Available Types	Main Features	Usage Requirements	Purchase Recommendations
Standard Engine	Spark Presto	Integrated Spark: The Standard Spark Engine supports native syntax from the Spark/Presto community, making it easy to learn and migrate. Flexible usage: Supports both Hive JDBC and Presto JDBC. Integrated Spark: The Standard Spark Engine can execute SQL and Spark batch tasks.	The free Gateway specification is 2 CU.	 Requires the use of Spark/Presto native syntax. Prefer to purchase a Spark engine for batch jobs and offline SQL tasks. Prefer to use Hive JDBC and Presto JDBC.
SuperSQL Engine	SparkSQL\\nSpark Jobs\\nPresto	Unified syntax: A single syntax is applicable to both Spark and Presto engines. Supports federated queries.	Requires learning the SuperSQL unified syntax.\\nFor SQL/batch tasks, it is recommended to purchase the corresponding engine type.	 Prefer to use Spark + Presto unified syntax. Federated queries are required.

Note:

1. Before purchasing, you should ensure that your account has been granted financial permissions in CAM.

2. Resources cannot be used across regions, so confirm that the current region is correct before purchasing.

Standard Engine Configuration Guide

After completing the purchase and configuration of the Standard Engine, you can use it within DLC's **Data Exploration**. Additionally, for the Spark Standard Engine, if you have multi-tenant or task isolation requirements, you can also configure **Resource Group** for resource allocation and isolation. The detailed guide is as follows:

Step 1: Purchasing the Engine

Note:

1. Engines cannot be used across regions.

2. Engine specification recommendation: Since a 16 CU cluster is relatively small, it is recommended only for testing scenes. For real production environments, it is recommended to choose a cluster with a specification of 64 CUs or more.

3. Engine network configuration: Custom network configurations can be set during the initial purchase. If you need to make changes later, please Submit Ticket to apply for modifications.





Data Lak	e Compute Back					Documentation Billing Co
Engine edition	SuperSQL engine Standard engine If you are more accustomed to the communit purchase and use the SuperSQL version. For	Beta ine y's syntax and behavior, y details, see Introduction t	ou are advised to purcha o Data Engines.	se and use a standard e	ngine. To ensure unified ser	mantics between different engines, you are advised to
Billing mode	Pay-as-you-go Monthly subso In this mode, a cluster is billed based on the loads and irregular task cycles.	ription Detailed con	nparison bended when no task is ir	progress. A suspended	cluster incurs no cost. It is	suitable for data compute applications with certain task
Region	-Hong Kong/Macao/TaiWan (China Region)- Hong Kong Cloud products in different regions are not in region nearest to your customers to reduce a	Southeast Asia Singapore terconnected over private ccess latency.	Eastern U.S Virginia networks and the region	Europe Frankfurt cannot be changed afte	- Southeast Asia Pacific- Jakarta r you purchase the service	Please proceed with caution.We recommend you select
Engine confi	gurations					
Engine type	Presto Spark Presto applies to interactive query and analy	is.				

Step 2: Using Data Exploration

Selecting the Standard Engine for Queries

Note:

Depending on the type of Standard Engine, you may need to switch to the corresponding syntax for queries.

If you select the Standard Spark Engine in Data Exploration, you can allocate task resources by using the DLC default resource group, a created resource group, or a one-time resource group (custom configuration).

Data Explore	Hong Kong 🔻					Data explore
Database Query	¢ +	Query-2024 Draft 🗙	+ •			🌣 Storage co
Catalog DataLakeCatalog	•		3 C	ĸ」	🛢 Default database 🔻 🗄 test 🄻 📐 Sta	andard - Presto Synta
Select a target database	•	1 2 SELECT * ERM) Datal akeCata	Data engine	Refresh	BET I THE TRANSPORT IN THE OWNER
Etest				test	Standard-Presto T	
				The engine supportsStandard - P	Presto Syntaxquery.	
				Engine (kernel version) Difference rules. For details, see Kernel Version	fferent kernel versions support different SQL syntax /ersions.	
				Standard Presto (Standard-P 1	(.0)	
				① Create engine		
				Advanced settings A	Configuration description	
		Query result		+ Select configuration. Mor	re 🔻	ownload history
		フ の Task ID Export				

Retrieving Full Results

Currently, the Standard Engine only supports returning up to 1,000 query results in the console. To retrieve the full results, you can see the following methods:

Engine	Retrieval Method
Standard Spark Engine	 Users can configure the engine to automatically save query results to a COS path or view them in DLC's managed storage. Results can be downloaded locally for review.
Standard Presto Engine	Retrieve full results via JDBC.

Step 3: Configuring Resource Groups (Optional)

Resource groups provide a secondary queue division of computing resources within the Spark Standard Engine. For a detailed introduction, see Resource Group Introduction. The computing units (CUs) of the DLC Spark Standard Engine can be allocated across multiple resource groups as needed. You can set the minimum and maximum CU limits for each resource group, along with start/stop policies, concurrency levels, and dynamic/static parameters, ensuring resource isolation and efficient workload management in complex scenes such as multi-tenancy and multi-tasking.

When you purchase a Standard Spark Engine, DLC provides a default resource group and also allows you to create multiple custom resource groups based on your specific business needs for flexible usage.



Note:

An engine can have a one-to-many relationship with resource groups. For example, Engine A can have several resource groups.

Managing and Configuring Resource Groups

1. Click to enter the resource group management of the corresponding engine.

2. Enter the Resource Management Group interface, and click **Create Resource Group** to configure a custom resource group. Alternatively, you can view and use the DLC default-configured resource group (no configuration required).

Appendix

Recommendations for Selecting Gateway Specifications

Gateway Specification	Spark Batch Instant Concurrency (Submitted/Running Tasks)	Concurrent Spark SQL/Presto SQL Queries	Number of Presto Engines Managed	Number of Spark Resource Groups Managed	Gateway HA
2 CU	30/50	100	4	50	No
16 CU	80/150	250	12	150	Yes
32 CU	220/400	600	35	400	Yes
64 CU	400/600	1000	70	700	Yes

Note:

The gateway is provided by default with a 2 CU specification (free of charge). If you need to upgrade the specifications, you can click Gateway details \rightarrow select Specification Configuration to adjust and purchase.



Gateway					
The gateThroughIn the test	 The gateway is a gateway service that helps users build connections between the local database and the DLC standard engine. Through the gateway, you can use the console, JDBC, or other methods to submit SQL queries, analyses, and other tasks to the standard engine. Learn more In the test period, the gateway of 2 CUs is free of charge. If you have any questions, submit a ticket. 				
Spec confi	guration Start Suspend Monitor				
Gateway Name	default-gateway-mszysnf6				
Resource ID	DataEngine-4nllqhmf				
Spec	2CU				
Status	Running				
Tag	No tag <i>P</i> Tags are used to categorize resources. To learn more, see Tag Documentation ☐				