DataArts Fabric

Product Introduction

Issue	01
Date	2025-12-19

HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD.

Trademarks and Permissions


and other Huawei trademarks are the property of Huawei Technologies Co., Ltd.
All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei Cloud and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Huawei Cloud Computing Technologies Co., Ltd.

Address: Huawei Cloud Data Center Jiaoxinggong Road
Qianzhong Avenue
Gui'an New District
Gui Zhou 550029
People's Republic of China

Website: https://www.huaweicloud.com/intl/en-us/

Contents

1 What Is DataArts Fabric?

DataArts Fabric is a comprehensive, data-AI converged computing service offered by Huawei Cloud. It serves as a one-stop development platform that manages the entire lifecycle of data processing—from analysis and model fine-tuning to inference and deployment. This enables diverse roles such as data engineers, data scientists, and AI application developers to collaborate efficiently on a unified workspace using their preferred tools, seamlessly transitioning from development to production. DataArts Fabric automatically scales to meet the demands of even the most resource-intensive applications. By incrementally allocating resources based on specific needs, it eliminates the need for pre-provisioned peak-load capacity pools, reducing customer costs by up to 50%.

Built on a serverless resource pool, DataArts Fabric enables shared workloads across data and AI tasks, heterogeneous CPU and NPU resources, and combined development and production environments. This innovative approach transforms resource investment strategies, facilitating mixed online-offline operations and integrated training-inference workflows. It helps customers optimize resource utilization, balance workload peaks and valleys, and achieve cost efficiency. With its user-friendly design, there is no cluster management required, enabling zero-resource-barrier initiation of development and production tasks. This empowers businesses to experiment with low-cost iterations in rapidly evolving markets.

Architecture

DataArts Fabric offers a high-performance, highly reliable, low-latency, and cost-effective mass storage system. When integrated with Huawei Cloud big data services, it significantly reduces costs and simplifies big data management for enterprises.

- **SQL engine**

  DataArts Fabric's distributed SQL engine features layered decoupling of metadata services, computing, caching, and storage, enabling elastic resource allocation at each layer without impacting performance or availability. Statement-level elastic scaling and high-performance distributed analysis engines facilitate TB-level data queries in seconds and PB-level queries in minutes.

- **Distributed Ray**

  To overcome distributed computing challenges in data processing and machine learning/deep learning workloads, DataArts Fabric supports the Ray

framework. It also provides a unified workflow that integrates data engineering and machine learning engineering. DataArts Fabric Ray's Ray-Data, Ray-Train, and Ray-Serve modules facilitate distributed data preprocessing, model training, and inference services.

- **Online inference**

  DataArts Fabric includes a proprietary, high-performance elastic inference engine. You can run inference jobs using the default inference service or by deploying your own custom models.

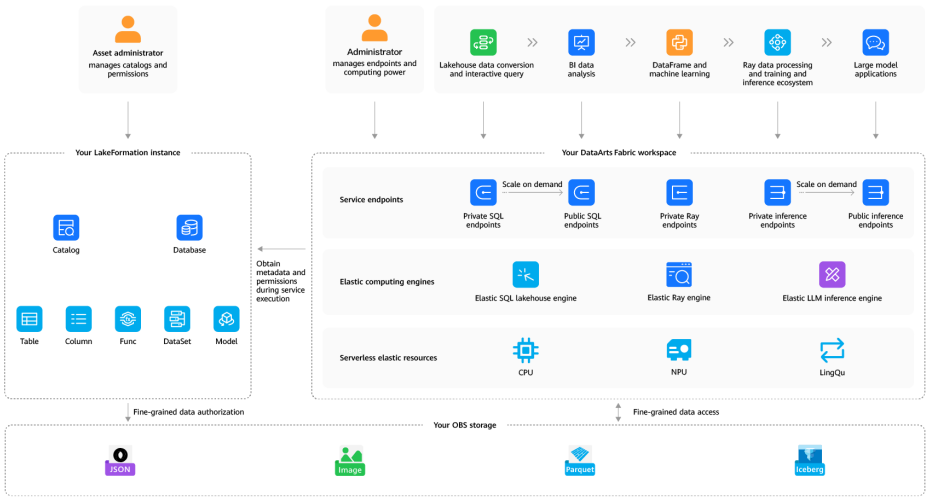
- **Heterogeneous resource management**

  DataArts Fabric offers unified management and allocation of CPU and NPU resources. Resource scheduling is supported at container-level or actor-level granularity. Furthermore, DataArts Fabric provides secure sandboxes for resource isolation and robust fault tolerance.

- **Multi-semantic cache acceleration**

  DataArts Fabric delivers cross-engine, multi-modal, and multi-semantic acceleration through various caching mechanisms, including data, model, and checkpoint caches.

Figure 1-1 Product architecture



Access Methods

DataArts Fabric offers multiple access methods:

A web-based management console, HTTPS-based APIs, and SDK clients for seamless compute engine integration.

- Management console

  DataArts Fabric can be accessed via **its management console** for managing Ray jobs, SQL jobs, model deployment, and model inference. This enables end-to-end data and AI development directly from the console.

- APIs

To integrate DataArts Fabric into third-party systems or for secondary development, use the provided HTTPS APIs. For specific operations and detailed descriptions of the APIs, refer to **API Reference**.

- SDKs

  To integrate DataArts Fabric functionalities into third-party systems for secondary development, use the SDKs. DataArts Fabric SDKs encapsulate the REST APIs in Python and Java, simplifying development. For specific operations and detailed descriptions of the SDKs, refer to **SDK Reference**.

2 Advantages

DataArts Fabric has the following advantages:

One-stop data intelligence development with unified experience

- A single workspace supports multiple workloads, including SQL, Ray-based data engineering, and model inference.
- Structured, semi-structured, and unstructured data are uniformly managed through LakeFormation, ensuring consistent metadata and permission control across the entire data intelligence development process.
- Data and AI workloads share a single copy of data, eliminating the need for data replication.

Out-of-the-box, elastic, and on-demand resources

- Mainstream open-source third-party large model inference services are pre-configured. Customers can directly call preset inference service APIs for tasks like text dialogues without acquiring additional resources. Pay-per-use billing is supported.
- The inference service automatically scales to meet demand.
- Ray workloads benefit from pod-level automatic scaling, dynamically allocating resources to handle peak loads.
- SQL supports both resource-based and query-based billing, with rapid, query-level elasticity for computing resources.

Open-source ecosystem

- Open-source Ray capabilities are provided within the Ascend ecosystem, with high-reliability Redis integration.
- The Ray dashboard offers visualized monitoring, troubleshooting, performance tuning, and application runtime management.
- SQL supports data formats such as ORC, Parquet, and Iceberg, leveraging the open lakehouse ecosystem.

3 Scenario

This section describes the main application scenarios of DataArts Fabric.

- **Data engineering**

  Efficiently process data through parallel computing, accelerating tasks such as data cleansing, conversion, and aggregation.
- **Distributed machine learning**

  Leverage Ray for distributed training and tuning, enabling efficient model training with large-scale datasets and models.
- **Foundation models**

  Utilize large models to implement intelligent dialogues, automatic summarization, machine translation, text classification, and image generation.
- **Real-time data analysis**

  Perform queries and analysis on massive datasets using standard SQL APIs.

4 Features

4.1 DataArts Fabric SQL Features

Introduction to DataArts Fabric SQL

DataArts Fabric SQL is a fully managed data platform designed for superb elasticity and lakehouse capabilities. It leverages Huawei Cloud's robust infrastructure, featuring resource pooling and massive storage. This, combined with its unique architecture—including parallel execution, metadata decoupling, and compute-storage separation—delivers advanced Software as a Service (SaaS) technologies. Its serverless architecture empowers you to process complex business logic using SQL, eliminating the need for infrastructure management.

Built upon the Huawei Cloud DataArts Fabric platform, DataArts Fabric SQL's architecture comprises a service access layer, compute layer, and storage layer. This design ensures hierarchical decoupling and elasticity across metadata services, computing, cache, and storage. Each layer can dynamically allocate resources without impacting the performance or availability of others. Statement-level elastic scaling and high-performance distributed analysis engines facilitate TB-level data queries in seconds and PB-level queries in minutes.

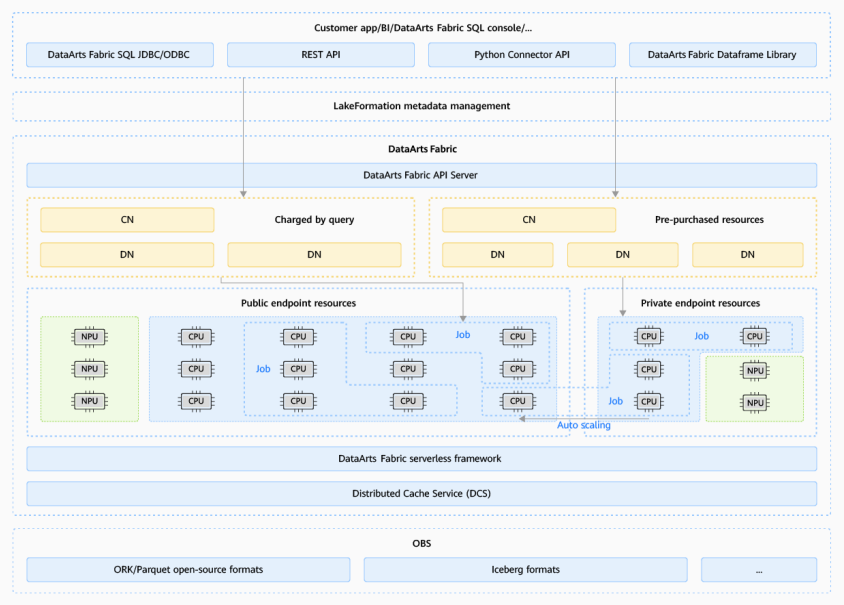
DataArts Fabric SQL supports processing and analyzing open structured data formats, including Iceberg, ORC, and Parquet. By aligning with the open lake ecosystem, it enables seamless data sharing when working with multiple data lake ecosystem services.

Aligned with the Data+AI ecosystem, DataArts Fabric SQL offers a Python UDF feature that enables users to execute Python scripts directly within SQL, supporting one-stop AI data processing.

DataArts Fabric SQL provides a visualized interface and a JDBC driver for easy interaction with existing applications and third-party tools. Additionally, it offers REST and Python APIs, allowing developers to manage and transform data using familiar programming languages.

Product Architecture

Figure 4-1 DataArts Fabric SQL architecture



Functions

The following table describes the key functions of DataArts Fabric SQL.

Table 4-1 DataArts Fabric SQL features

Feature	Description
Standard SQL statements	Supports ANSI standard SQL, extended to include GBK, UTF-8, SQL ASCII, and Latin-1 character sets.
DDL	Supports CREATE, ALTER, DROP, SHOW, and DESCRIBE operations for schemas and tables.
Data types	Supports smallint, int, bigint, float, double, numeric, timestamp, date, varchar, char, bool, binary, and string.
APIs	Supports standard JDBC 4.0, RESTful APIs, and Python Connector APIs.
Transaction capabilities	Supports partition-level transactions and concurrency control. Iceberg offers full transaction capabilities.
Multi-tenant management	Tenants are isolated via dedicated CNs/DNs. Each CN/DN occupies an exclusive POD for isolation.
Importing and exporting data	Data import and export use INSERT INTO. Data export from foreign tables supports format conversion (Parquet <-> ORC).

Feature	Description
Scalability	Supports two-level elasticity. Elastic compute units (nodes) scale within seconds (less than 2s) in a resource pool based on query characteristics. Additional compute resources scale based on resource usage. **NOTE** This is a beta feature. Capacity expansion operations require the DataArts Fabric service to provision containers via the management plane.
SMP	Provides intra-node Symmetric Multi-Processing (SMP) for full utilization of multi-core CPUs. **query_dop** is enabled by default.
Vectorized execution	The vectorized execution engine enhances OLAP performance.
Statistics collection	ANALYZE collects statistics to improve optimizer accuracy and ensure stable, efficient database performance.
Storage formats	Supports Parquet, ORC, and Iceberg.
DML	Supports INSERT INTO and INSERT OVERWRITE.
Partitioned tables	Supports partitioned tables for Parquet, ORC, and Iceberg.
Views	Supports views.
User-defined functions (UDFs)	Extends SQL statements with user-defined functions for unified execution. Currently, only Python is supported.
Elastic computing scale	A single query supports up to 256 elastic compute nodes.
Fine-grained access control	Table metadata is managed by LakeFormation, utilizing IAM permissions. LakeFormation currently oversees overall permission control.

5 Product Specifications

Model Inference Product Specifications

Table 5-1 Model inference product specifications

Type	Specification	Computing Power
MU	mu.llama3.8b	It offers the computing power of approximately 400 requests per minute (RPM) for the Llama-3-8B model to process short tokens.
	mu.llama3.70b	It offers the computing power of approximately 100 RPM for the Llama-3-70B model to process short tokens.
	mu.llama3.1.8b	It offers the computing power of approximately 190 RPM for the Llama-3.1-8B model to process short tokens.
	mu.llama3.1.70b	It offers the computing power of approximately 130 RPM for the Llama-3.1-70B model to process short tokens.
	mu.qwen2.72b	It offers the computing power of approximately 1,700 RPM for the Qwen2-72B model to process short tokens.
	mu.glm4.9b	It offers the computing power of approximately 110 RPM for the GLM-4-9B model to process short tokens.

Ray Cluster Specifications

Table 5-2 Ray cluster specifications

Type	Specification	Computing Power
DPU	fabric.ray.dpu.d1x	It offers approximately 4 vCPUs and 16 GB of memory.
	fabric.ray.dpu.d2x	It offers approximately 8 vCPUs and 32 GB of memory.
	fabric.ray.dpu.d4x	It offers approximately 16 vCPUs and 64 GB of memory.
	fabric.ray.dpu.d8x	It offers approximately 32 vCPUs and 128 GB of memory.
	fabric.ray.dpu.d16x	It offers approximately 64 vCPUs and 256 GB of memory.
	fabric.ray.dpu.d32x	It offers approximately 128 vCPUs and 512 GB of memory.
APU	fabric.ray.apu.b1.1x	If offers the computing power of one AI accelerator card (B1).
	fabric.ray.apu.b2.1x	If offers the computing power of one AI accelerator card (B2).
	fabric.ray.apu.b3.1x	If offers the computing power of one AI accelerator card (B3).
	fabric.ray.apu.b1.8x	If offers the computing power of eight AI accelerator cards (B1).
	fabric.ray.apu.b2.8x	If offers the computing power of eight AI accelerator cards (B2).
	fabric.ray.apu.b3.8x	If offers the computing power of eight AI accelerator cards (B2).

DataArts Fabric SQL Specifications

The following table lists the DataArts Fabric SQL specifications.

Table 5-3 SQL specifications

Type	Specification	Computing Power
DCU	fabric.sql.dcu.std	It offers approximately 1 vCPU and 4 GB of memory.

6 Permission Management

If you need to assign different permissions for employees in your organization to access DataArts Fabric resources, IAM is a good choice for fine-grained permission management. IAM provides user authentication, permission assignment, and access control, enabling secure management of access to your cloud resources. If the HUAWEI ID already meets your permission requirements and you do not need IAM for user permission management, you may skip this section.

IAM can be used free of charge. You pay only for the resources in your account.

With IAM, you can control the scope of access to specific Huawei Cloud resources. For example, if certain employees are involved in software development and need access to DataArts Fabric resources, but you want to restrict high-risk operations like resource deletion, you can create IAM users and grant them usage permissions only, while withholding deletion privileges.

DataArts Fabric supports role/policy-based authorization.

Table 6-1 Role/Policy-based authorization

Policy	Core Relationship	Permission	Authorization Method	Application Scenario
Role/Policy-based authorization	User-permission-authorization scope	• System-defined roles • System-defined policies • Custom policies	Assigning roles or policies to principals	To authorize a user, you need to add it to a user group first and then specify the scope of authorization. It provides a limited number of condition keys and cannot meet the requirements of fine-grained permissions control. This method is suitable for small- and medium-sized enterprises.

For example, if you need to grant an IAM user permission to create ECSs in the **CN North-Beijing4** region A and OBS buckets in the **CN South-Guangzhou** region B, under role/policy-based authorization, you must create two custom policies and assign both to the IAM users. With identity policy-based authorization, however, you only need to create one custom identity policy and configure the condition key **g:RequestedRegion** for the policy, and then attach the policy to the users or grant the users the access permissions to the specified regions. Identity policy-based authorization is more flexible than role/policy-based authorization.

For more information about IAM, see **IAM Service Overview**.

Role/Policy-based Authorization

DataArts Fabric supports role- and policy-based authorization. By default, new IAM users do not have any permissions. To grant permissions, you must add them to one or more groups and attach policies or roles to these groups. This process is known as authorization. Once authorized, users can use cloud services based on the granted permissions.

DataArts Fabric is a project-level service that can be deployed and accessed within specific physical regions. When you set **Scope** to **Region-specific projects** and select the specified projects (for example, **cn-north-4**) in the specified regions (for example, **CN North-Beijing4**), the users only have permissions for resources in the selected projects. If you set **Scope** to **All resources**, the users have permissions for resources in all region-specific projects. When accessing DataArts Fabric, the users need to switch to a region where they have been authorized to use this service.

The following table lists all system-defined permissions of DataArts Fabric.

Table 6-2 System-defined permissions supported by DataArts Fabric

Role/Policy Name	Description	Category	Dependency
DataArtsFabricFullPolicy	Full permissions for DataArts Fabric.	System-defined policy	• IAM Agency Management FullAccess • OBS Administrator • LakeFormation ReadOnlyAccess • KMS Administrator (optional)

Role/Policy Name	Description	Category	Dependency
DataArtsFabricConsoleFullPolicy	All permissions for using DataArts Fabric on the console, including all permissions of **DataArts FabricFullPolicy** and certain console-specific permissions.	System-defined policy	• IAM Agency Management FullAccess • OBS Administrator • LakeFormation ReadOnlyAccess • IAM PolicyFullAccess • KMS Administrator (optional)
DataArtsFabricReadOnlyPolicy	Read-only permissions for DataArts Fabric.	System-defined policy	LakeFormation ReadOnlyAccess

The following table lists the common operations supported by system-defined permissions for DataArts Fabric. Refer to this table to select the appropriate permissions as needed.

Table 6-3 Common operations supported by system-defined permissions for DataArts Fabric

Operation	DataArtsFabricConsoleFullPolicy	DataArtsFabricFullPolicy	DataArtsFabricReadOnlyPolicy
Listing workspaces	√	√	√
Creating a workspace	√	√	×
Modifying a workspace	√	√	×
Modifying workspace monitoring configuration	√	√	×
Deleting a workspace	√	√	×
Querying compute resources	√	√	√

Operation	DataArtsFabricConsoleFullPolicy	DataArtsFabricFullPolicy	DataArtsFabricReadOnlyPolicy
Creating a compute resource	√	√	×
Modifying a compute resource	√	√	×
Deleting a compute resource	√	√	×
Listing the endpoints of a workspace	√	√	√
Creating an endpoint for a workspace	√	√	×
Querying the endpoint details of a workspace	√	√	√
Modifying an endpoint of a workspace	√	√	×
Deleting an endpoint of a workspace	√	√	×
Listing jobs	√	√	√
Creating a job	√	√	×
Querying jobs	√	√	√
Modifying a job	√	√	×
Deleting a job	√	√	×
Listing services	√	√	√
Creating a service	√	√	×
Modifying a service	√	√	×

Operation	DataArtsFabricConsoleFullPolicy	DataArtsFabricFullPolicy	DataArtsFabricReadOnlyPolicy
Querying services	√	√	√
Deleting a service	√	√	×
Creating a model	√	√	×
Listing models	√	√	√
Querying models	√	√	√
Deleting a model	√	√	×
Modifying a model	√	√	×
Creating a tag	√	√	×
Deleting a tag	√	√	×
Listing tags	√	√	√
Querying tags of a specific resource	√	√	√
Listing resources by tag	√	√	√
Creating a notification policy	√	√	×
Listing notification policies	√	√	√
Deleting a notification policy	√	√	×
Listing running jobs	√	√	√
Running a job	√	√	×

Operation	DataArtsFabricConsoleFullPolicy	DataArtsFabricFullPolicy	DataArtsFabricReadOnlyPolicy
Querying running jobs	√	√	√
Deleting a running job	√	√	×
Canceling a running job	√	√	×
Invoking an inference service instance	√	√	×
Listing routes	√	√	√
Querying session information	√	√	√
Subscribing to a public endpoint	√	√	×
Querying SQL endpoints	√	√	√
Creating a SQL endpoint	√	√	×
Deleting a SQL endpoint	√	√	×
SQL editor	√	√	√

Roles or Policies Required by the DataArts Fabric Console

Table 6-4 Roles or policies required by the DataArts Fabric console

Console Function	Dependency	Role/Policy Required
Granting service permissions	IAM	Granting permissions on the authorization page requires the IAM user to have the **IAM Agency Management FullAccess** policy.

Console Function	Dependency	Role/Policy Required
Creating a workspace	LakeFormation	Users with the **DataArtsFabricFullPolicy** policy can create workspaces. Specifying a LakeFormation metastore during workspace creation requires the **LakeFormation ReadOnlyAccess** policy.
Creating a model	OBS	To create a model and specify its OBS file path on the model management page, an IAM user must have the **DataArtsFabricFullPolicy** and **OBS OperateAccess** policies.
Creating a notification policy	IAM SMN	Creating a notification policy requires an IAM user to have the **DataArtsFabricFullPolicy**, **IAM Agency Management ReadOnly**, and **SMN ReadOnlyAccess** policies.

7 Constraints

7.1 Notes and Constraints on Ray and XDS

Large Model License Constraints

Open-source large models are subject to varying license restrictions. Refer to the table below for details.

Table 7-1 Large model license constraints

Model Name	License Address
Llama 3 8B Chinese Instruct	https://github.com/meta-llama/llama/blob/main/LICENSE
Llama 3 70B	https://github.com/meta-llama/llama/blob/main/LICENSE
Llama 3.1 8B Chinese Chat	https://huggingface.co/meta-llama/Meta-Llama-3.1-8B/blob/main/LICENSE
Llama 3.1 70B	https://huggingface.co/meta-llama/Meta-Llama-3.1-8B/blob/main/LICENSE
Qwen 2 72B Instruct	https://huggingface.co/Qwen/Qwen2-72B-Instruct/blob/main/LICENSE
Glm 4 9B Chat	https://huggingface.co/THUDM/glm-4-9b-chat/blob/main/LICENSE

Constraints on Public Inference Services

- Token quota: Each public inference service has a free token quota. After the quota is used up, the service becomes unavailable, and there is no option to purchase more. The quota is shared across all workspaces under the current user at the same site.

- Time limit: Services remain valid for 90 days after activation. Expired services cannot be used. If the same inference service is activated across multiple workspaces, the initial activation date applies.
- Context length variations: Different models have specific context length limits. Refer to **Public Inference Service** for details.
- No service level agreement (SLA) guarantee: For higher performance, you are advised to create your own dedicated inference service.

7.2 DataArts Fabric SQL Restrictions

DataArts Fabric SQL Service Usage Restrictions

Table 7-2 DataArts Fabric SQL service usage restrictions

Item	Description
Service enabling	An account can enable only one DataArts Fabric SQL service, which is shared among all sub-users.
Connection operations	The service supports multiple connection modes, including SQL editor, JDBC, SDK, and API.
O&M operations	The service operates in serverless mode, eliminating the need for O&M tasks such as capacity expansion, upgrades, backup and restoration, or disaster recovery.
Timeout limit	The user authorization token lasts for 8 hours. Requests made after this period will fail to execute.
SQL syntax	For details, see **Database Operation Restrictions**.

Database Operation Restrictions

Table 7-3 Database operation restrictions

Type	Syntax	Supported
Basic functions	CREATE EXTERNAL TABLE	Yes
	DROP TABLE	Yes
	CREATE VIEW	Yes
	DROP VIEW	Yes
	INSERT	Yes
	SELECT	Yes
	TRUNCATE	Yes
	EXPLAIN	Yes

Type	Syntax	Supported
	ANALYZE	Yes
	ALTER TABLE DROP PARTITIONS	Yes
	ALTER TABLE SET TABLEPROPERTIES	Yes
	ALTER TABLE UNSET TABLEPROPERTIES	Yes
	ALTER TABLE DROP COLUMNS	No (supported only by Iceberg)
	ALTER TABLE ADD COLUMNS	No (supported only by Iceberg)
	ALTER TABLE COLUMN RENAME	No (supported only by Iceberg)
	CREATE EXTERNAL TABLE AS	Yes

