DataArts Fabric

User Guide

Issue 01

Date 2025-12-19





Copyright © Huawei Cloud Computing Technologies Co., Ltd. 2025. All rights reserved.

No part of this document may be reproduced or transmitted in any form or by any means without prior written consent of Huawei Cloud Computing Technologies Co., Ltd.

Trademarks and Permissions

HUAWEI 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, quarantees 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/

i

Contents

1 Preparations	1
1.1 Creating an IAM User and Assigning Permissions to Use DataArts Fab	ric1
1.2 Configuring DataArts Fabric Service Agency Permissions	4
1.3 Creating an Access Client	8
1.4 Creating a Workspace	<u>C</u>
2 Ray Scenario	11
2.1 Ray Resource Management	11
2.1.1 Purchasing a Ray Resource	11
2.1.2 Unsubscribing from Ray Resources	12
2.2 Image Management	13
2.3 Managing Ray Clusters	15
2.3.1 Creating a Ray Cluster	15
2.3.2 Viewing the Ray Cluster Overview	17
2.3.3 Creating a Ray Job	17
2.3.4 Running a Ray Job	18
2.3.5 Managing Ray Jobs	19
2.3.6 Viewing the Ray Dashboard	19
2.3.7 Deleting a Ray Cluster	20
2.3.8 Viewing Metrics	20
2.4 Managing Ray Services	21
2.4.1 Creating a Ray Service	21
2.4.2 Upgrading a Ray Service	26
2.4.3 Running an Inference Service	27
2.4.4 Deleting a Ray Service	28
3 DataArts Fabric SQL	29
3.1 DataArts Fabric SQL Usage Process	29
3.2 Managing SQL Endpoints	29
3.2.1 Enabling a Public Endpoint	30
3.2.2 Creating a SQL Endpoint	30
3.2.3 Modifying a SQL Endpoint	31
3.2.4 Deleting a SQL Endpoint	31
3.2.5 Querying Details About a SQL Endpoint	31

3.2.6 Querying SQL Job History	32
3.2.7 Monitoring SQL Endpoints	32
3.3 Managing SQL Sessions	32
3.3.1 Viewing SQL Sessions	32
3.3.2 Checking SQL Execution Status	33
3.4 Using SQL Editor	33
3.5 Interconnection with Ecosystem Components	33
3.5.1 Accessing DataArts Fabric SQL Using DBeaver	33
3.5.2 Accessing DataArts Fabric SQL Using Tableau	35
3.5.3 Obtaining JDBC	37
4 Large Model Inference Scenarios	38
4.1 Introduction to Large Model Inference Scenarios	38
4.2 Large Model Inference Process	39
4.3 Using a Public Inference Service for Inference	39
4.3.1 Viewing a Public Inference Service	39
4.3.2 Enabling an Inference Service	40
4.3.3 Performing Inference in the Playground	41
4.4 Creating My Inference Service for Inference	42
4.4.1 Creating a Model	42
4.4.2 Managing a Model	45
4.4.3 Creating an Inference Endpoint	46
4.4.4 Creating an Inference Service	48
4.4.5 Using an Inference Service for Inference	51
4.4.6 Deleting an Inference Service	52
4.4.7 Deleting an Inference Endpoint	53
4.5 Viewing All Metrics on AOM	53
5 O&M Management	55
5.1 Configuring Message Notifications	55
5.2 Deleting a Notification	56

Preparations

1.1 Creating an IAM User and Assigning Permissions to Use DataArts Fabric

Before using DataArts Fabric functions, prepare the account, configure permissions for the account and its sub-accounts, and create a workspace. This section describes how to create an IAM user and assign permissions to use DataArts Fabric.

Prerequisites

You have a valid Huawei Cloud account.

Procedure

- **Step 1** Log in to the Huawei Cloud console. Click in the upper left corner of the page and choose **Identity and Access Management** from the service list.
- **Step 2** Choose **Permissions** > **Policies/Roles**, click **Create Custom Policy** in the upper right corner, set necessary parameters, and click **OK**. For details about the creation process, see **Creating a Custom Policy**.

Administrators can set different policies for different user groups to control user permissions. Administrators can configure permissions as required. The following lists recommended permission combinations.

Table 1-1 Permissions

Service Role	Policy	Function
System administrator	{ "Version": "1.1", "Statement": [{ "Effect": "Allow", "Action": ["DataArtsFabric:*:*", "obs:bucket;*", "obs:object:*"] }] }	With all DataArts Fabric permissions, this role can perform all DataArts Fabric operations.
Resource administrator	{ "Version": "1.1", "Statement": [{ "Effect": "Allow", "Action": ["DataArtsFabric:workspace:*", "DataArtsFabric:endpoint:*", "lakeformation:instance:*"] }] }	With the permission to manage users' DataArts Fabric resources, this role can create and delete workspaces and endpoints.

Service Role	Policy	Function
Inference service operator	{ "Version": "1.1", "Statement": [{ "Effect": "Allow", "Action": ["DataArtsFabric:workspace:list", "DataArtsFabric:endpoint:list", "DataArtsFabric:endpoint:show", "DataArtsFabric:model:*", "DataArtsFabric:service:*", "obs:object:*", "obs:bucket:ListBucket"] }] }	This role can perform inference-related operations, such as registering models, creating inference services, and running inference.
Job service operator	{ "Version": "1.1", "Statement": [{ "Effect": "Allow", "Action": ["DataArtsFabric:workspace:list", "DataArtsFabric:endpoint:list", "DataArtsFabric:endpoint:show", "DataArtsFabric:job:*", "obs:object:*", "obs:bucket:ListBucket"] }] }	This role can perform job-related services, including creating and executing jobs.

- **Step 3** In the navigation pane on the left, click **User Groups**. Click **Create User Group** in the upper right corner, enter the user group name, and click **OK**.
- **Step 4** In the user group list, select the created user group, click **Authorize**, select the required policies, and click **Next**. Select **Scope** as required and click **OK**. For details, see **Creating a User Group and Assigning Permissions**.

- Step 5 In the navigation pane on the left, click Users. In the upper right corner, click Create User. Enter User Details as required, select Access Type and Credential Type, and click Next.
- **Step 6** In the **Available User Groups** list, select the target user group and click **Create**. For more information, see **Creating an IAM User**.

----End

1.2 Configuring DataArts Fabric Service Agency Permissions

Current cloud services provide multiple functions. Different functions require different agency permissions. For details, see **Table 1-2**.

Prerequisites

You have a valid Huawei Cloud account.

Procedure

- **Step 1** Log in to the DataArts Fabric Workspace Management Console and click **Service Authorization**.
- **Step 2** Authorize an agency on the **Service Authorization** page. You can configure the agency permissions based on the policy as required.

Table 1-2 Agency policy

Agency Policy Name	Permission Item	Ma nda tor y (Ye s/N o)	Function
FABRIC_CO MMON_PO LICY	iam:agencies:listAgencies iam:roles:getRole iam:permissions:listRolesForA gency obs:bucket:ListAllMyBuckets obs:bucket:ListBucket obs:object:GetObjectVersion obs:object:GetObject	Yes	 IAM permissions: Only some read-only permissions are assigned to enable the service to compare the user's agency with the required agency. The user will be prompted to update the agency if necessary. OBS permissions: All services, including jobs and inference, require the permission to read OBS files. With this permission, job files can be pulled from the user's OBS bucket for execution and model files can be deployed. For OBS permissions, the user can manually modify the OBS-related part in the fabric_admin_trust agency on the IAM agency page to restrict the access to OBS resources. For details, see the Example Custom Policies part in IAM Permissions.
FABRIC_LTS _POLICY	lts:groups:create lts:groups:get lts:groups:list lts:topics:create lts:topics:get lts:topics:list	Yes	Permissions required by DataArts Fabric to configure log dumping.

Agency Policy Name	Permission Item	Ma nda tor y (Ye s/N o)	Function
FABRIC_SEL F_POLICY	DataArtsFabric:workspace:list Route DataArtsFabric:workspace:list Route DataArtsFabric:workspace:sh owSession DataArtsFabric:workspace:list MessagePolicy DataArtsFabric:endpoint:sho w DataArtsFabric:endpoint:list DataArtsFabric:job:dropJobIns tance DataArtsFabric:job:listJobInst ance	Yes	Permissions required by DataArts Fabric to help users manage resources.
FABRIC_LAK EFORMATI ON_POLICY	lakeformation:accessTenant:g rant lakeformation:access:delete lakeformation:access:create lakeformation:access:describe lakeformation:agreement:gra nt lakeformation:agreement:des cribe lakeformation:agreement:can cel lakeformation:agency:create lakeformation:agency:describ e	No	Permissions required by DataArts Fabric to use LakeFormation. If LakeFormation needs to be interconnected, enable this policy.
FABRIC_SM N_POLICY	smn:topic:publish	No	Permissions required by DataArts Fabric to use Simple Message Notification (SMN). If the message notification capability is required, enable this policy.

Agency Policy Name	Permission Item	Ma nda tor y (Ye s/N o)	Function
FABRIC_SW R_POLICY	swr:repo:listRepoDomains swr:repo:listRepoTags swr:repo:createRepoDomain	No	Permissions required by DataArts Fabric to use usershared images.
FABRIC_VPC EP_POLICY	vpcep:epservices:get vpcep:connections:update vpcep:permissions:update vpcep:permissions:list	No	Permissions required by DataArts Fabric to connect to the user network.
FABRIC_OB S_POLICY	obs:bucket:PutLifecycleConfiguration obs:bucket:ListBucketMultipartUploads obs:object:GetObject obs:bucket:HeadBucket obs:bucket:DeleteBucket obs:bucket:CreateBucket obs:bucket:ListAllMyBuckets obs:bucket:ListBucket	No	Permissions required by DataArts Fabric to use the OBS bucket.

□ NOTE

All agency permissions except the mandatory ones can be canceled.

Step 3 Add an agency to the bucket policy.

DataArts Fabric uses the **fabric_admin_trust** agency to access files in the OBS bucket of the user. Therefore, the agency needs to access the OBS bucket of the user.

Users need to check whether a bucket policy is configured for the OBS bucket used by DataArts Fabric. If a bucket policy is configured, ensure that the agency is not denied by the existing bucket policy and perform the following steps to add the agency to the bucket policy:

- 1. Log in to the OBS console and choose **Resources** > **Buckets** in the navigation pane on the left.
- 2. On the **Buckets** page, click the bucket name to access its **Objects** tab page.
- 3. In the navigation pane on the left, choose **Permissions** > **Bucket Policies**. Then, click **Create**.

4. On the **Create Bucket Policy** page, customize the policy name, set **Principal** to **Other accounts**, and enter the agency account (in the format of Account ID/Agency name). The agency name is **fabric_admin_trust**. Example: s3a7973a07cf4725abf5ba0b6d7*****/fabric_admin_trust.

Step 4 Check whether server-side encryption is configured for OBS.

If server-side encryption is configured for the OBS bucket and the encryption mode is **SSE-KMS**, add the **KMS Administrator** permission to the agency **fabric_admin_trust**. For details, see **Why Cannot an Authorized Account or User Upload or Download KMS Encrypted Objects?**

Due to security management requirements, DataArts Fabric cannot directly configure the **KMS Administrator** permission for users. Users need to perform the following steps to confirm and add the permission:

- 1. Log in to the OBS console and choose **Resources** > **Buckets** in the navigation pane on the left.
- 2. On the **Buckets** page, click the bucket name to access its **Objects** tab page.
- 3. In the navigation pane on the left, choose **Overview**.
- 4. In the **Basic Configurations** area, check whether **Server-Side Encryption** is configured and **Encryption Method** is **SSE-KMS**.
 - If Encryption Mode is not SSE-KMS, skip the following steps.
 - If Encryption Mode is SSE-KMS, go to the next step.
- 5. Configure the **KMS Administrator** permission.
 - a. In the upper right corner of **OBS Console**, hover the mouse over the username and click **Identity and Access Management**.
 - b. In the navigation pane of the **Identity and Access Management** console, click **Agencies**.
 - c. On the **Agencies** page, search for the agency name **fabric_admin_trust** in the text box. On the right of the **fabric_admin_trust** agency, click **Authorize**.
 - d. In the text box in the upper right corner of the **Authorize Agency** page, search for the policy name **KMS Administrator**, select the policy, click **Next**, and click **OK** to complete the authorization.

----End

1.3 Creating an Access Client

After creating an access client, you can use the VPC Endpoint service to access DataArts Fabric APIs with a private domain name or IP address.

Prerequisites

- You have a valid Huawei Cloud account.
- The current account already has sufficient quotas for resources such as VPCEP and DNS private domain names. If the creation fails, the client is automatically rolled back and deleted, and related resources are reclaimed.

Creating a client will incur fees. The actual fees are subject to the bill. For details, see **Billing Modes**.

- Log in to the DataArts Fabric Workspace Management Console. Choose Access Management. On the displayed client list tab page, click Create Client.
- 2. On the **Create Client** page, enter **Client Name**, select **VPC** and **Subnet**, select **I have read, understood, and agreed to the above content**, and click **OK**.

Table 1-3 Parameters for creating a client

Parameter	Description
Client Name	Custom client name, which can contain only letters, numbers, underscores (_), and hyphens (-), with a length of 4 to 32 characters.
VPC	Select VPC in the drop-down list. For details about how to create a VPC, see Creating a VPC with a Subnet .
Subnet	Select the subnet in the drop-down list. For details about how to create a subnet, see Creating a VPC with a Subnet .

When the **Status** of the client changes to **Running**, the client is created.

3. Click the client name. On the displayed client details page, you can view the domain name and IP address of the access connection list.

When accessing the service with the domain name or IP address, set **HOST** in the request header to the domain name.

- Calling with the domain name:
 curl -kv https://fabric-ep.{region}.myhuaweicloud.com/healthcheck -H "host:fabric-ep. {region}.myhuaweicloud.com"
- Calling with the IP address:
 curl -kv https://192.168.0.200/healthcheck -H "host:fabric-ep.{region}.myhuaweicloud.com"

1.4 Creating a Workspace

A workspace is the basic unit of DataArts Fabric. All subsequent operations are performed in the workspace. Therefore, you need to create a workspace after account authorization is configured.

You can create one or more workspaces as required. Each workspace is independent.

Prerequisites

You have a valid Huawei Cloud account.

Step 1 Log in to the DataArts Fabric Workspace Management Console, click **Create Workspace**, set necessary parameters according to **Table 1-4**, and click **Create**.

After the workspace is created, the **Workspace Management Console** page is displayed.

Table 1-4 Parameters for creating a workspace

Parameter	Description
Workspace Name	Enter a workspace name. Cluster names of the same account must be unique.
Workspace Description	(Optional) Describe the workspace.
Enterprise Project	After an enterprise project is selected, clusters and their security groups will be created in that project. To manage clusters and other resources like nodes, elastic load balancers, and node security groups, you can use the Enterprise Project Management Service (EPS).
Tags	(Optional) You can add tags to resources to classify resources.
	You can create predefined tags on the TMS console. The predefined tags are available to all resources that support tags. You can use predefined tags to improve efficiency of the tag creation and resource migration. For details, see Creating Predefined Tags.
	• A tag key can have a maximum of 128 characters, including letters, digits, spaces, and special characters (:=+@). It cannot start or end with a space, or start with _sys A tag key cannot be empty.
	 A tag value can have a maximum of 255 characters, including letters, digits, spaces, and special characters (:=+@). The value can be empty.

Step 2 Click **Access Workspace** in the created workspace. When the user agreement is displayed, you can view the agreement and click **Agree**. Then, you can access the created workspace.

----End

2 Ray Scenario

2.1 Ray Resource Management

2.1.1 Purchasing a Ray Resource

Prerequisites

- You have a valid Huawei Cloud account.
- You have at least one workspace available.

Procedure

Ray is fully managed. Before using Ray, you need to purchase a Ray resource. Perform the following operations:

- **Step 1** Log in to the DataArts Fabric Workspace Management Console.
- Step 2 Select the created workspace, click Access Workspace, and choose Resources and Assets > Ray Resources.
- **Step 3** Click **Buy Ray Resource** in the upper right corner. The **Purchase Ray Resource** page is displayed. Select the DPU or APU specification, quantity, and purchase duration as required. For details, see **Table 2-1**.

Table 2-1 Parameters on the Purchase Ray Resource page

Parameter	Description
Billing Mode	You can select Yearly/Monthly or Pay-per-Use .
Resource Type	You can select DPU or APU as required. DPU : CPU-based compute unit oriented to data analysis scenarios. APU : NPU-based compute unit oriented to AI scenarios.

Parameter	Description
Specifications	The differences between DPU resource specifications, such as fabric.ray.dpu.d1x, fabric.ray.dpu.d2x, and fabric.ray.dpu.d4x, lie in the number of CPUs and memory size.
	The differences between APU resource specifications lie in the number of Ascend cards and server models. You can select resources of different specifications as required.
Purchase Duration	You can select the purchase duration as required.

There are minimum requirements for purchasing Ray resources. At least four **fabric.ray.dpu.d1x** resources are required. In DataArts Fabric, **fabric.ray.dpu.d** $n\mathbf{x} = n \times \mathbf{fabric.ray.dpu.d1x}$.

Step 4 Click **Next**. Confirm the configurations and click **Pay** to go to the payment page. After the payment is done, the purchase is complete. You can view the status of purchased resources in the Ray resource tag.

○ NOTE

- After the purchase is complete, the status of new resources on the **Ray Resources** page is **Preparing**. If the purchase is successful, the status changes to **Running**. Otherwise, the status changes to **Failed**.
- If you purchase resources for the first time, wait for about 15 to 20 minutes. If the purchase list contains APU resources, wait for about 40 to 50 minutes. If resources of other specifications are added, wait for about 5 minutes. If APU resources are added, wait for about 20 minutes. You can manually refresh the resource status to check whether the resources are ready.
- For a specific specification, you can purchase only one resource. After the purchase is successful, you can adjust the number of the resource through scaling. If you need to purchase multiple resources of the same specification at the same time, create a workspace and purchase the resource again.

----End

2.1.2 Unsubscribing from Ray Resources

Prerequisites

- You have a valid Huawei Cloud account.
- You have at least one workspace available.
- You have purchased Ray resources.

Procedure

- **Step 1** Log in to the DataArts Fabric Workspace Management Console.
- Step 2 Select the created workspace, click Access Workspace, and choose Resources and Assets > Ray Resources.

Step 3 The operations for the **Yearly/Monthly** and **Pay-per-Use** modes are different:

- Yearly/Monthly: Locate the target Ray resource and click More > Unsubscribe in the Operation column.
- Pay-per-Use: Click Delete in the Operation column.

∩ NOTE

Deleted or unsubscribed Ray resources cannot be restored and may affect the status of existing Ray clusters.

Step 4 In the displayed dialog box, enter **DELETE** and click **Confirm**.

----End

2.2 Image Management

Prerequisites

- You have a valid Huawei Cloud account.
- You have at least one workspace available.
- You have enabled the image package whitelist function. If you need a trial, choose Service Tickets > Create Service Tickets on the top navigation bar of DataArts Fabric Workspace Management Console to apply for the permission.

Uploading an Image Package to SWR

Log in to the SWR console. In the **Upload Through SWR** dialog box, upload the image package to SWR as prompted. If the file size exceeds 2 GB, use the client to upload the file. For details, see *Uploading an Image*.

Creating an Image Package

- 1. Log in to the DataArts Fabric Workspace Management Console, select the created workspace, and click **Access Workspace**.
- 2. In the navigation pane, choose **O&M Management** > **Image Package Management**. On the displayed page, click **Create Image Package** in the upper right corner.
- 3. Enter the name and version name as prompted, select the path of the image package stored in OBS for the specified version, and click **Confirm** to create an image package.

For details, see Table 2-2.

Table 2-2 Parameters for creating an image package

Parameter	Description
Name	Image package name.
Descriptio n	Enter the description of the image package as required.

Parameter	Description
Туре	Image package type. Select RAY_CLUSTER for the Ray cluster scenario and RAY_SERVICE for the Ray service scenario.
Version Name	An image package can have multiple versions. Enter a version name based on the current information.
Version Descriptio n	Description of the version to be created.
Version Type	Currently, only OBS is supported.
Path	OBS path of the version to be created.

□ NOTE

If a **RAY_CLUSTER** or **RAY_SERVICE** Cap is created, the Cap name and version name must be the same as the package name.

For example, if the OBS path is **obs:**//xxx/ray-cap/files/ray-cluster-2.34.0.tgz and the package name is ray-cluster-2.34.0, the Cap name must be ray-cluster and the version must be **2.34.0**. Otherwise, the verification on the page fails.

After a custom version is created, you can view **My Ray image package** when creating a Ray cluster or **My Ray service image package** when creating a Ray service.

Adding an Image Package Version

- 1. On the **Image Package Management** page, locate the target image package and click **View Versions** in the **Operation** column.
- 2. On the Current Image Package Versions page, click Add Version.
- On the displayed page, set related parameters and click Confirm.
 For details, see Table 2-3.

Table 2-3 Parameters for creating an image package version

Parameter	Description
Version Name	An image package can have multiple versions. Enter a version name based on the current information. The image package version must be the same as the version of the selected OBS file package.
Version Description	Description of the version to be created.
Version Type	Currently, only OBS is supported.
Path	OBS path of the version to be created. Select the parent directory that contains the metadata.yaml file.

Deleting an Image Package Version

After an image package version is deleted, all related data will be cleared. Exercise caution when performing this operation.

- 1. On the **Image Package Management** page, locate the target image package and click **View Versions** in the **Operation** column.
- 2. On the **Current Image Package Versions** page, locate the target version and click **Delete** in the **Operation** column.
- 3. In the displayed dialog box, enter **DELETE** or click **One-click input**, and then click **Confirm**.

Deleting an Image Package

Deleted image packages cannot be restored, and all related data will be cleared. Exercise caution when performing this operation.

- 1. On the **Image Package Management** page, locate the target image package and click **Delete** in the **Operation** column.
- 2. In the displayed dialog box, enter **DELETE** or click **One-click input**, and then click **Confirm**.

2.3 Managing Ray Clusters

2.3.1 Creating a Ray Cluster

Ray is a high-performance distributed execution framework that provides distributed computing abstractions using an architecture different from traditional distributed computing systems.

Ray clusters are fully managed and exclusively used. You do not need to worry about background resource management. Ray clusters provide Ray-based distributed job execution capabilities and are fully compatible with open-source versions, so you can use Ray clusters without complex script adaptation. In addition, Ray clusters natively support dashboard capabilities that are user-friendly. Compared with open-source Ray, DataArts Fabric incorporates a series of security hardening measures to ensure user data security, such as gRPC channel encryption and dashboard access authentication.

Prerequisites

- You have a valid Huawei Cloud account.
- You have at least one workspace available.
- You have purchased the required Ray resources.

Procedure

- **Step 1** Log in to the DataArts Fabric Workspace Management Console.
- Step 2 Select the created workspace and click Access Workspace. In the navigation pane on the left, choose Resources and Assets > Ray Clusters. Click Create Ray Cluster in the upper right corner.

Step 3 On the displayed page, set **Head Specifications** and **Worker Specifications** as required by referring to **Table 2-4**. Then, click **Create Now**.

Table 2-4 Parameters for creating a Ray cluster

Parameter	Description
Cluster Name	Name of the Ray cluster to be created.
Ray Image Package Type	Select a public Ray image package.
Ray Image Package	You can select different Ray versions as required. The version number is the same as that of the Ray community.
Head Specifications	Head node specifications of the Ray cluster to be created. Set this parameter as required. All specifications are displayed in the specification list. The selected specification can be downward compatible with the created Ray resource. For example, if the fabric.ray.dpu.d4x
	resource is created, you can select fabric.ray.dpu.d1x , fabric.ray.dpu.d2x , or fabric.ray.dpu.d4x for Head Specifications . That is, a large resource specification can be split into multiple smaller resource specifications.
Worker Specifications	Worker group specifications of the Ray cluster to be created. Multiple worker groups can be created. Select a specification from the resource specification list for worker node deployment, and set the minimum and maximum number of worker nodes. The minimum number must be at least 1, and the maximum number can be set based on workloads. When the Ray cluster is initialized, the minimum number of worker nodes are created. The number of worker nodes is dynamically scaled to the maximum number based on workloads. You can also add worker nodes of different specifications. The selected worker node specification must be also downward compatible with the existing resource. For example, if the purchased Ray resource is fabric.ray.dpu.d4x and fabric.ray.dpu.d1x is selected for Head Specifications, you can also select fabric.ray.dpu.d1x for Worker Specifications and set the maximum number of worker nodes to 3.

----End

You can manually refresh the page to monitor the Ray cluster creation progress. The creation takes approximately 3 to 5 minutes.

If a Ray cluster fails to be created, delete the failed cluster before creating another one. This prevents the failed cluster from occupying resources.

2.3.2 Viewing the Ray Cluster Overview

Prerequisites

- You have a valid Huawei Cloud account.
- You have at least one workspace available.
- You have at least one Ray cluster.

Procedure

- **Step 1** Log in to the DataArts Fabric Workspace Management Console.
- **Step 2** Select the created workspace and click **Access Workspace**. In the navigation pane on the left, choose **Resources and Assets** > **Ray Clusters**. Click a Ray cluster to view its details.

Table 2-5 Parameters

Parameter	Description
Cluster Name	User-defined Ray cluster name
Cluster ID	Unique ID of a cluster
Status	Status of the current cluster
Description	Custom description of the cluster
Created By	Creator of the cluster
Created On	Time when the cluster is created
Cluster Version	Version of the deployed Ray cluster
Ray Resources	Specifications and quantity of resources required for cluster deployment
Link	Link for accessing the Ray dashboard

----End

2.3.3 Creating a Ray Job

Prerequisites

- You have a valid Huawei Cloud account.
- You have at least one workspace available.
- You have at least one Ray cluster available.
- Job code has been developed based on service requirements and uploaded to OBS. For details about how to create an OBS bucket and upload files, see Creating a Bucket.

- **Step 1** Log in to the DataArts Fabric Workspace Management Console.
- **Step 2** Select the created workspace, click **Access Workspace**, and choose **Development** and **Production** > **Jobs**.
- **Step 3** Click **Create Job** in the upper right corner. Set required parameters by referring to **Table 2-6**. Set **Job Type** to **Ray** and other parameters as required. The Ray main file is the main entry file of the job you have developed.

Table 2-6 Parameters for creating a job

Parameter	Description
Job Name	Name of the job to be created.
Job Type	The default value is Ray .
Code Directory	Select the OBS directory where the job code is stored.
Ray Main File	Select the main entry Python file of the job running code in the code directory. Ensure that you have selected the main entry file for running the job. Otherwise, the job running may not meet your expectation. NOTE Do not enter sensitive information in the script or print sensitive information through the script.
Ray Job Parameters	Parameters required for executing the Ray main file. The following is an example: ["class","org.ray.examples.rayTest","10","model_2","20"] NOTE Do not enter sensitive information in the script or print sensitive information through the script.
Dependencies	Software and its version on which the Ray job depends. Before the Ray job is executed, the dependency is installed using pip. The format is specified in the requirements.txt file. Example: numpy==1.24.3
Ray Clusters	Target Ray cluster where the job is executed.
Version	Job version.
Version Description	Version description, which contains a maximum of 1000 characters.

----End

2.3.4 Running a Ray Job

Prerequisites

• You have a valid Huawei Cloud account.

- You have at least one workspace available.
- You have at least one Ray cluster available.
- You have at least one job available.

- **Step 1** Log in to the DataArts Fabric Workspace Management Console.
- **Step 2** Select the created workspace, click **Access Workspace**, and choose **Development** and **Production** > **Jobs**.
- **Step 3** Select the target job in the list, specify the endpoint where the job runs, and click **Start** in the **Operation** column.

----End

2.3.5 Managing Ray Jobs

Prerequisites

- You have a valid Huawei Cloud account.
- You have at least one workspace available.
- You have at least one Ray cluster available.
- You have at least one job available.

Procedure

- **Step 1** Log in to the DataArts Fabric Workspace Management Console.
- **Step 2** Select the created workspace, click **Access Workspace**, and choose **Development** and **Production** > **Jobs**.
- **Step 3** You can select **Start**, **View**, or **Delete** in the **Operation** column as required. You can filter jobs by job name, status, endpoint name, and type.
- **Step 4** Click **View Dashboard** in the **Operation** column to access the dashboard provided by Ray and view the job running details.

----End

2.3.6 Viewing the Ray Dashboard

After creating a Ray cluster or running a Ray job, you can access the dashboard provided by Ray to view the Ray cluster details or the job running details.

Prerequisites

- You have a valid Huawei Cloud account.
- You have at least one workspace available.
- You have at least one Ray cluster available.
- You have at least one job available.

You can use either of the following methods to view the path:

- Method 1: Access the dashboard from the Ray Clusters page.
 - a. Log in to the DataArts Fabric Workspace Management Console.
 - Select the created workspace and click Access Workspace. In the navigation pane on the left, choose Resources and Assets > Ray Clusters.
 - c. Click the target Ray cluster.
 - d. Click the link at the bottom.
- Method 2: Access the dashboard from the Job History page.
 For details, see Managing Ray Jobs.

2.3.7 Deleting a Ray Cluster

Prerequisites

- You have a valid Huawei Cloud account.
- You have at least one workspace available.
- You have at least one Ray cluster.
- If the current Ray cluster has records of running jobs, you need to delete the jobs before deleting the Ray cluster.

Procedure

- **Step 1** Log in to the DataArts Fabric Workspace Management Console.
- **Step 2** Select the created workspace and click **Access Workspace**. In the navigation pane on the left, choose **Resources and Assets** > **Ray Clusters**.
- **Step 3** Select the target Ray cluster and click **Delete** in the upper right corner.



Once a Ray cluster is deleted, all its records will be cleared and cannot be restored. Exercise caution when performing this operation.

Step 4 In the displayed dialog box, enter **DELETE** and click **Confirm**.

----End

2.3.8 Viewing Metrics

For you to query the usage of Ray cluster resources, the cloud service platform reports metrics to AOM.

Prerequisites

• You have a valid Huawei Cloud account.

- You have at least one workspace available.
- You have at least one Ray cluster.

- **Step 1** Log in to the AOM console.
- **Step 2** In the navigation pane, choose **Metric Browsing** and set **Metric Sources** to **Prometheus_AOM_Default**.
- **Step 3** Select **All metrics** and enter a metric name for query.

Table 2-7 Monitoring metrics

Metric	Description
fabric_dpu_cpu_usage	CPU usage of the head and worker nodes in a Ray cluster Unit: %
fabric_dpu_mem_usag e	Memory usage of the head and worker nodes in a Ray cluster Unit: %

----End

2.4 Managing Ray Services

2.4.1 Creating a Ray Service

Prerequisites

- You have a valid Huawei Cloud account. For details, see Creating an IAM
 User and Assigning Permissions to Use DataArts Fabric and Configuring
 DataArts Fabric Service Agency Permissions.
- You have at least one workspace available. For details, see Creating a
 Workspace.
- You have purchased the required Ray resources. For details, see Purchasing a Ray Resource.
- You have created a Ray service image package version and an inference deployment file. For details, see **Creating an Image Package**.
- If you use Log Tank Service (LTS) to view logs, you need to obtain the LTS permissions. For details, see **Granting LTS Permissions to IAM Users**.

Creating a Ray Service

Step 1 Log in to the DataArts Fabric Workspace Management Console.

- **Step 2** Select the created workspace and click **Access Workspace**. In the navigation pane, choose **Resources and Assets** > **Ray Services**, and click **Create Ray Service** in the upper right corner.
- Step 3 On the displayed page, set required parameters, including Basic Settings, Log Settings, Ray Cluster Settings, Data, and Ray Serve Settings.

For details, see **Table 2-8**.

Table 2-8 Parameters for creating a Ray service

Parameter		Description
Basic Settings	Ray Service Name	Name of the Ray service to be created.
	Add Descriptio n	Click Add Description and enter the introduction to the Ray service in the text box. It can contain a maximum of 1,000 characters.
	Image Package Source	The options are Public Ray image package and My Ray service image package . Select My Ray service image package .
		Public Ray image package: Public image packages provided by DataArts Fabric. They are open-source Ray images and support enhanced DataArts Fabric features such as channel encryption, secure dashboard access, and key encryption and decryption.
		My Ray service image package: Image packages customized by tenants. Tenants can customize Ray images and use DataArts Fabric's image package management to create and deploy image packages.
	Image Package Name	Name of the service image package to be used.
	lmage Package Version	Select a Ray service version as required.
Log Settings	Enabling LTS	Whether to store Ray service run logs in the log service provided by Huawei Cloud LTS.
		After this function is enabled, logs in the following paths are collected:
		/tmp/ray/session_latest/logs/**/*/var/log/service-log/**/*
	Log Group	Select a log group of Huawei Cloud LTS. You can create a log group on the LTS console. For details, see Creating a Log Group.

Parameter		Description
	Log Stream	Select a log stream of Huawei Cloud LTS. You can create a log stream on the LTS console. For details, see Creating a Log Stream.
Ray Cluster	Head Specificati	Head node specifications of the Ray cluster to be created. Set this parameter as required.
Settings	ons	All specifications are displayed in the specification list. The selected specification can be downward compatible with the created Ray resource. For example, if the fabric.ray.dpu.d4x resource is created, you can select fabric.ray.dpu.d1x, fabric.ray.dpu.d2x, or fabric.ray.dpu.d4x for Head Specifications. That is, a large resource specification can be split into multiple smaller resource specifications.
	Worker Specificati ons	Worker group specifications of the Ray cluster to be created. You can click Add Worker Group to create multiple worker groups of different specifications.
		Select a specification from the resource specification list for worker node deployment, and set the minimum and maximum number of worker nodes. The minimum number must be at least 1, and the maximum number can be set based on workloads.
		When the Ray cluster is initialized, the minimum number of worker nodes are created. The number of worker nodes is dynamically scaled to the maximum number based on workloads.
		The selected worker node specification must be also downward compatible with the existing resource. For example, if the purchased Ray resource is fabric.ray.dpu.d4x and fabric.ray.dpu.d1x is selected for Head Specifications, you can also select fabric.ray.dpu.d1x for Worker Specifications and set the maximum number of worker nodes to 3.
Data	Data Input	Model path used for running the inference service. After the Ray service is created, the model files in this path are copied to the Ray service cluster.
Ray Serve Settings	Add Applicatio n	You can click Add Application to configure and customize deployment files, running environments, and scheduling parameters. A maximum of five applications can be added.
	Applicatio n Name	Name of the application to be created.
	Code Directory	Code directory required for inference. You can select OBS, Image Path, or Other.

Parameter		Description
	eployme File oth	Path of the inference instance in the code.
	outing efix	Routing prefix for inference. The routing prefix of each application must be unique.
nt	ivironme ariables	Select Environment Variables as required and click Add to configure environment variables. For details, see Managing Environment Variables of a Training Container .
De nt	eployme	Inference instance corresponding to the application. Select Deployment and set this parameter based on the specifications of each application.
		Multiple deployments can be created in a single application. Configurations of the Ray actor, automatic scaling, and inference can be customized for each deployment.
		Resources required by the Ray actor of each deployment can be configured separately. However, the total number of resources required for deployments in a single application cannot exceed the worker specifications in the basic settings.
		You can configure the fixed and maximum number of replicas, as well as the automatic scaling range for a deployment. If the fixed number of replicas has been configured for a deployment, automatic scaling cannot be configured.

----End

Viewing Ray Service Details

- **Step 1** Log in to the DataArts Fabric Workspace Management Console.
- **Step 2** Select the created workspace and click **Access Workspace**. In the navigation pane on the left, choose **Resources and Assets** > **Ray Services**.
- **Step 3** On the displayed page, click the name of the target Ray service to access its details page.

On the displayed page, you can view the Ray service overview and Ray Serve settings. For details, see **Table 2-9** and **Table 2-10**.

Table 2-9 Parameters on the Overview tab

Parameter	Description
Ray Service Name	User-defined Ray service name.

Parameter	Description
Ray Service ID	Unique ID of the Ray service.
Status	Status of the current Ray service.
Description	Custom description of the Ray service.
Created By	Creator of the Ray service.
Created	Time when the Ray service is created.
Image Package Version	Version of the Ray service image required in the current Ray service deployment.
Head Specifications	Resource specifications and quantity required by head nodes in the Ray service deployment.
Worker Specifications	Resource specifications and quantity required by worker nodes in the Ray service deployment.
Dashboard	Link for accessing the Ray dashboard.
Data	Path and environment variables generated based on the user-defined input path.
Log Transfer to LTS	You can select Yes or No . If you enable LTS in the log settings when creating the Ray service, set this parameter to Yes .
View LTS Logs	If Log Transfer to LTS is enabled, you can click the link to go to the LTS log stream to view logs.

Table 2-10 Parameters on the Ray Serve Settings tab

Parameter	Description
Application name	Name of the created application.
Inference Address	Address for calling the inference service. For details, see Running an Inference Service.
Code Directory	Directory of the code required for inference.
Deployment File Path	Path of the inference instance in the code.
Routing Prefix	Routing prefix for inference. The routing prefix of each application must be unique.
Environment Variables	Environment variables in the container, which are generated based on the code directory and model directory.

Parameter	Description
Deployment	Inference instance corresponding to the application.
	Multiple deployments can be created in a single application. Configurations of the Ray actor, automatic scaling, and inference can be customized for each deployment.
	Resources required by the Ray actor of each deployment can be configured separately. However, the total number of resources required for deployments in a single application cannot exceed the worker specifications in the basic settings.
	You can configure the fixed and maximum number of replicas, as well as the automatic scaling range for a deployment. If the fixed number of replicas has been configured for a deployment, automatic scaling cannot be configured.

----End

2.4.2 Upgrading a Ray Service

Prerequisites

- You have a valid Huawei Cloud account. For details, see Creating an IAM
 User and Assigning Permissions to Use DataArts Fabric and Configuring
 DataArts Fabric Service Agency Permissions.
- You have at least one workspace available. For details, see Creating a
 Workspace.
- You have purchased the required Ray resources. For details, see Purchasing a Ray Resource.
- You have created the Ray service image package version and inference deployment file (if required). For details, see **Creating an Image Package**.

Upgrading a Ray Service

∩ NOTE

You can select an image package version or modify the Ray service configurations as required. The version upgrade does not interrupt your existing services.

- **Step 1** Log in to the DataArts Fabric Workspace Management Console.
- **Step 2** Select the created workspace, click **Access Workspace**, choose **Resources and Assets** > **Ray Services**, and search for the Ray service to be upgraded.
- **Step 3** Click **Upgrade Version** in the **Operation** column.
- **Step 4** On the displayed page, select required values for **Image Package Version**, **Ray Cluster Settings**, **Data**, and **Ray Serve Settings**.

For details, see Table 2-8.

□ NOTE

The waiting time for upgrading the Ray service is 3,000s. If it times out, the upgrade fails.

----End

Rolling Back the Ray Service

If the Ray service upgrade fails due to incorrect upgrade configurations or other reasons, you need to roll back the Ray service.

- **Step 1** On the **Ray Services** page, search for the Ray service that fails to be upgraded, and click **Roll back to the previous version** in the **Operation** column.
- **Step 2** In the displayed dialog box, confirm the target version and click **Confirm**.

----End

2.4.3 Running an Inference Service

Prerequisites

- You have a valid Huawei Cloud account. For details, see Creating an IAM
 User and Assigning Permissions to Use DataArts Fabric and Configuring
 DataArts Fabric Service Agency Permissions.
- You have at least one workspace available. For details, see Creating a
 Workspace.
- You have at least one Ray service. For details, see Creating a Ray Service.

Running an Inference Service

- **Step 1** Log in to the DataArts Fabric Workspace Management Console.
- Step 2 Select the created workspace, click Access Workspace, and choose Resources and Assets > Ray Services.
- **Step 3** On the displayed page, obtain the inference address of the target Ray service from the **Inference Address** column.
- **Step 4** Call the inference address using the API tool or other methods to query the inference result.

You can use curl for inference as shown in the following:

curl -s -k --location -X POST 'https://fabric-inference-url/v1/workspaces/{workSpaceId}/endpoints/ {endPointId}/rayservice/fruit' --header "X-Auth-Token: \$(cat test.json)" --header 'Content-Type: application/json' --data-raw '["MANGO", 3]'

The inference result is **9**.

----End

Viewing the Ray Dashboard

- **Step 1** Log in to the DataArts Fabric Workspace Management Console.
- Step 2 Select the created workspace, click Access Workspace, and choose Resources and Assets > Ray Services.

- **Step 3** On the displayed page, click the target Ray service name.
- **Step 4** On the displayed details page, choose the **Overview** tab and click **View Now** on the right of **Dashboard**. The Ray dashboard is displayed, where you can view details about the inference service.

----End

2.4.4 Deleting a Ray Service

Prerequisites

You have at least one Ray service. For details, see Creating a Ray Service.

Procedure



Once a Ray service is deleted, all its records will be cleared and cannot be restored. Exercise caution when performing this operation.

- **Step 1** Log in to the DataArts Fabric Workspace Management Console.
- **Step 2** Select the created workspace, click **Access Workspace**, and choose **Resources and Assets** > **Ray Services**.
- **Step 3** Locate the target Ray service and choose **More** > **Delete** on the right. In the displayed dialog box, enter **DELETE** and click **Confirm**.

----End

3 DataArts Fabric SQL

3.1 DataArts Fabric SQL Usage Process

Table 3-1 describes the DataArts Fabric SQL usage process.

Table 3-1 Operation process

Procedure	Description
Preparations	Sign up for a HUAWEI ID and enable Huawei Cloud services, complete real-name authentication, top up your account, enable LakeFormation and OBS permissions, and confirm the agency.
Creating a workspace	Create a workspace. You can skip this step if you have created a workspace.
Creating a SQL endpoint	Create a SQL endpoint. You can skip this step if you use a public endpoint.
Planning and creating an OBS bucket and importing data	Create an OBS bucket and folders for data storage.
Planning and creating catalogs and databases	Create catalogs and databases on the LakeFormation page and specify the OBS bucket directory.
Querying data	Query data on the SQL Editor page.

3.2 Managing SQL Endpoints

3.2.1 Enabling a Public Endpoint

Public endpoints are accessible to all users but must be enabled before use. To enable a public endpoint:

- **Step 1** Log in to the Huawei Cloud DataArts Fabric console and click **Access Workspace**.
- Step 2 In the navigation pane on the left, choose Resources and Assets > SQL Endpoints. On the displayed page, click the Public Endpoints tab. On this tab page, click Enable Endpoint.



This process takes about 1 minute.

----End

3.2.2 Creating a SQL Endpoint

In addition to using public endpoints, you can also create your own endpoints when using DataArts Fabric SQL. These endpoints are personal and not visible to other users.

- **Step 1** Log in to the Huawei Cloud DataArts Fabric console and click **Access Workspace**.
- **Step 2** In the navigation pane on the left, choose **Resources and Assets** > **SQL Endpoints**.
- Step 3 Click Create Endpoint. Set Endpoint Name, Description, and Pre-warmed Resources and select Enable Public Endpoint.

Basic information for creating a SQL endpoint

Paramet er	Description
Endpoint Name	(Mandatory) The name of the SQL endpoint. The name contains 1 to 128 characters and must be unique. Only letters, digits, underscores (_), hyphens (-), periods (.), and spaces are allowed.
Descripti on	(Optional) A brief description of the SQL endpoint. The value can contain up to 512 characters. The following special characters are not allowed: !<>='&"
Pre- warmed Resource s	(Mandatory) The number of pre-warmed resources allocated to the SQL endpoint. The value ranges from 50 to 5000.

Paramet er	Description
Enable Public Endpoint	When enabled, SQL jobs will be automatically scheduled to a public endpoint if pre-warmed resources are insufficient.

----End

3.2.3 Modifying a SQL Endpoint

When using DataArts Fabric SQL, you can modify the endpoints you created.

- **Step 1** Log in to the Huawei Cloud DataArts Fabric console and click **Access Workspace**.
- **Step 2** In the navigation pane on the left, choose **Resources and Assets** > **SQL Endpoints**.
- **Step 3** Click the desired endpoint card. On the displayed details page, click **Edit SQL Endpoint**, modify the content, and save the modification.

----End

3.2.4 Deleting a SQL Endpoint

- **Step 1** Log in to the Huawei Cloud DataArts Fabric console and click **Access Workspace**.
- Step 2 In the navigation pane on the left, choose Resources and Assets > SQL Endpoints.
- **Step 3** On the displayed page, locate the desired endpoint card and click the deletion button in its upper right corner. In the dialog box that appears, confirm the deletion and click **OK**.

----End

3.2.5 Querying Details About a SQL Endpoint

- **Step 1** Log in to the Huawei Cloud DataArts Fabric console and click **Access Workspace**.
- Step 2 In the navigation pane on the left, choose Resources and Assets > SQL Endpoints.
- **Step 3** On the displayed page, click the desired endpoint card to view its details.

----End

3.2.6 Querying SQL Job History

- **Step 1** Log in to the Huawei Cloud DataArts Fabric console and click **Access Workspace**.
- **Step 2** In the navigation pane on the left, choose **Resources and Assets** > **SQL Endpoints**.
- **Step 3** Click the endpoint card. On the endpoint details page, click the **SQL Job History** tab and query SQL job history.
- **Step 4** Click **Export** and select a date to export SQL job history.

Ⅲ NOTE

Only SQL job history from T-90 to T-1 can be exported.

----End

3.2.7 Monitoring SQL Endpoints

- **Step 1** Log in to the Huawei Cloud DataArts Fabric console and click **Access Workspace**.
- **Step 2** In the navigation pane on the left, choose **Resources and Assets** > **SQL Endpoints**.
- **Step 3** Click the endpoint card. On the endpoint details page, click the **Endpoint Monitor** tab.

Resource usage data is collected every 15 seconds. However, if SQL execution times are short, they may not be accurately reflected in the monitoring charts.

The resource usage metrics are updated every 5 minutes, while active session concurrency data is refreshed on a 1-minute cycle.

----End

3.3 Managing SQL Sessions

3.3.1 Viewing SQL Sessions

- **Step 1** Log in to the DataArts Fabric Workspace Management Console. On the displayed page, locate your desired workspace card and click **Access Workspace**.
- **Step 2** In the navigation pane on the left, choose **Development and Production > SQL Session**.
- **Step 3** (Optional) On the displayed page, locate the session you want to close and click **Close** in its **Operation** column. In the **Close Session** dialog box that appears, enter **YES** and click **OK**.

----End

3.3.2 Checking SQL Execution Status

- **Step 1** Log in to the DataArts Fabric Workspace Management Console. On the displayed page, locate your desired workspace card and click **Access Workspace**.
- **Step 2** In the navigation pane on the left, choose **Development and Production > SQL Session**.
- **Step 3** On the displayed page, locate the session for which you wish to check SQL status and click **SQL Execution Status** in its **Operation** column. Here, you can check the query statement, status, and submission time.
- **Step 4** (Optional) On the displayed page, locate the session you want to cancel SQL execution and click **Cancel Execution** in its **Operation** column. In the **Cancel Execution** dialog box that appears, enter **YES** and click **OK**.

----End

3.4 Using SQL Editor

You can use the SQL Editor to connect to DataArts Fabric SQL and perform SQL operations.

- **Step 1** Log in to the Huawei Cloud DataArts Fabric console and click **Access Workspace**.
- Step 2 In the navigation pane on the left, choose Development and Production > SQL Editor. Select a LakeFormation instance, a LakeFormation catalog, and a SQL endpoint to run the SQL statements. You can also enable Session Mode to retain the execution session. In scenarios where SQL statements need to be executed frequently and intermittently, this function can save the time for creating a session.
- **Step 3** The query results support the **Overwrite** and **Append** modes. **Overwrite** clears previous query results, and **Append** retains them. Query results can be displayed in tables.
- **Step 4** Query data using SQL statements by referring to *Developer Guide* and *SQL Syntax*.

----End

3.5 Interconnection with Ecosystem Components

3.5.1 Accessing DataArts Fabric SQL Using DBeaver

DBeaver is a SQL client and database management tool. For relational databases, you can use JDBC APIs to interact with a database through the JDBC driver.

Obtaining DBeaver

You can obtain the DBeaver of the required version from **the DBeaver community** based on the OS.

Interconnecting DataArts Fabric SQL with JDBC

- **Step 1** Obtain the Maven coordinates of JDBC. For details, see **Obtaining JDBC**.
- **Step 2** Open DBeaver, choose **Database** > **Driver Manager** from the menu bar, and add a custom driver.
- **Step 3** In the **Driver Manager** dialog box, click **New** to open a new driver dialog box.
- Step 4 Switch to the Libraries tab. Click Add Artifact, copy the Maven coordinates obtained in Step 1 to Dependency Declaration, and click OK. Click Find Class. On the displayed page, click Download, select org.postgresql.Driver, and click OK.
- **Step 5** Switch to the **Settings** tab page and set the parameters. The **Driver Name** can be customized. Set **Driver Type** to **Generic**. After the JAR file of the driver is imported, the class name is automatically loaded.

URL template:

idbc:fabricsql://{host}[:{port}]/[{database}]

Click **OK** to add the DataArts Fabric SQL driver.

- **Step 6** After the creation is complete, click **New Database Connection**, select the driver added in the previous step, and click **Next**.
- **Step 7** On the **Main** tab page, enter the host name and database name. You do not need to set the username and password. Switch to the **Driver properties** tab page and set the parameters described in **Table 3-2**. Click **Finish**.
- **Step 8** Set connection properties. Switch to **Driver properties** tab page.

Table 3-2 DataArts Fabric SQL connection parameters

Property Name	Description	Mandatory or Optional	How to Obtain
AccessKeyID	Authentication ID	Mandatory	Creating a
SecretAccessKey	Authentication key	Mandatory	Permanent Access Key or Obtaining Temporary Access Keys and Security Tokens of an IAM User
securityToken	STSToken	Optional (required when a temporary AK/SK pair is used.)	
workspaceId	Workspace ID	Mandatory	Click View Details on the Workspace Management Console page.

Property Name	Description	Mandatory or Optional	How to Obtain
endpointId	Endpoint ID	Mandatory	Querying Details About a SQL Endpoint
lakeformation_inst ance_id	LakeFormation instance ID	Mandatory	LakeFormation data
PGDBNAME	LakeFormation catalog name	Mandatory	

Step 9 After the database connection is added, click the drop-down arrow to display the schema list in the database.

----End

3.5.2 Accessing DataArts Fabric SQL Using Tableau

Tableau is a popular BI tool in the industry. For relational databases, you can use JDBC APIs to interact with a database through the JDBC driver.

Obtaining Tableau

You can obtain the latest Tableau version from the Tableau official website.

Interconnecting DataArts Fabric SQL with JDBC

- Step 1 Obtain JDBC. For details, see Obtaining JDBC.
- **Step 2** Install JDBC as instructed. After the installation is complete, find the **Drivers** folder in the Tableau installation directory and copy the JDBC JAR file to the folder.

The following is an example of the directory on Windows. For details, see **Tableau** and JDBC.

C:\Program Files\Tableau\Drivers

Step 3 Open Tableau. Click Other Databases (JDBC).

If this option is not displayed, click **More** and then click **Other Databases (JDBC)**.

Step 4 Enter the JDBC URL in **URL** and select **PostgreSQL** for **Dialect**.

JDBC URL template:

jdbc:fabricsql://<host>[:<port>]/<database>

Example:

jdbc:fabricsql://example.com:1234/database

Parameter	Description
host	DataArts Fabric SQL address

Parameter	Description
port	(Optional) Port number
database	Name of the LakeFormation catalog created in Planning and Creating a LakeFormation Instance, Catalog, and Database.

Additionally, you need to configure **Properties file** for authentication. Click **Browse** and select the compiled property file. The extension name of the property file is **.properties**. In this example, the property file is named **serverless.properties**. Then, click **Sign in**.

The following shows an example of the property file:

AccessKeyID=YOUR_AK
SecretAccessKey=YOUR_AK
securityToken=YOUR_STOKEN
workspaceId=YOUR_WORKSPACE
endpointId=YOUR_ENDPIONT_ID
lakeformation_instance_id=YOUR_LF_ID
PGDBNAME=YOUR_CATALOG

You do not need to set the username and password.

Property Name	Description	Mandatory or Optional	How to Obtain
AccessKeyID	Authentication ID	Mandatory	Creating a Permanent
SecretAccessKey	Authentication key	Mandatory	Access Key or Obtaining Temporary
securityToken	STSToken	Optional (required when a temporary AK/SK pair is used.)	Access Keys and Security Tokens of an IAM User
workspaceId	Workspace ID	Mandatory	Click View Details on the Workspace Management Console page.
endpointId	Endpoint ID	Mandatory	Querying Details About a SQL Endpoint
lakeformation_inst ance_id	LakeFormation instance ID	Mandatory	LakeFormation data

Property Name	Description	Mandatory or Optional	How to Obtain
PGDBNAME	LakeFormation catalog name	Mandatory	

----End

3.5.3 Obtaining JDBC

JDBC drivers are used to connect to DataArts Fabric SQL. You can obtain JDBC in the following way:

Obtaining It from the Maven Repository

Copy the Maven repository information to the pom.xml file.

Add the following Maven coordinates to the **pom.xml** file:

<dependency>
 <groupId>com.huaweicloud.dws</groupId>
 <artifactId>huaweicloud-dws-jdbc</artifactId>
 <version>8.5.1</version>
</dependency>

4 Large Model Inference Scenarios

4.1 Introduction to Large Model Inference Scenarios

Common large models include large language models (LLMs), multimodal large models, and text-to-image large models. LLMs support text generation and can perform inference based on your prompts. LLMs can be widely used in the following fields:

- Q&A system: LLMs can process natural languages, understand your intents, and answer your questions.
- Content production: LLMs can generate coherent articles, stories, and dialogues based on given texts or topics.
- Text summarization: LLMs can summarize long texts and extract key information, helping you quickly understand text content.
- Machine translation: LLMs can process translation tasks between multiple languages to implement cross-language communication.

Currently, DataArts Fabric provides the following two methods for inference:

- Using a Public Inference Service for Inference: DataArts Fabric provides a
 public inference service based on open-source LLMs (such as Qwen2 and
 GLM4). You can view inference endpoints on the Inference Endpoint page,
 and select a desired endpoint to enable it. Then, you can use the public
 inference service in the playground. In case of this method, common opensource large models can be used for inference after being enabled, and you
 do not need to deploy them.
- Creating My Inference Service for Inference: DataArts Fabric allows you to create your own inference services. You can upload your own LLMs or use public LLMs to deploy the inference services. Models created on the Model page of DataArts Fabric are visible only to yourself. You can view and delete models, and manage model versions, including adding, viewing, and deleting model versions.

4.2 Large Model Inference Process

DataArts Fabric offers an end-to-end AI development process in serverless mode, covering everything from data preparation to model deployment. Each stage of the process can be used independently. This section describes the DataArts Fabric usage process. You can choose one of the available methods to complete AI development.

Table 4-1 Process description

Process	Description	Reference
Creating a workspa ce	Create a workspace. All subsequent operations are performed in the workspace.	Creating a Workspace
Creating an endpoint	Create an endpoint. Create endpoints of different types based on service types.	Creating an Inference Endpoint
Registeri ng a model	You can register the fine-tuning model file stored in OBS as your fine-tuning model on the model management page.	Creating a Model
Deployin g a service	DataArts Fabric supports the deployment of models that have been fine-tuned based on base models.	Creating an Inference Service
Accessin g the service	After the fine-tuned model is deployed, you can use the inference API provided by DataArts Fabric to perform inference.	Using an Inference Service for Inference

4.3 Using a Public Inference Service for Inference

4.3.1 Viewing a Public Inference Service

During the trial period of an inference endpoint, you can directly use a public inference service for inference. The current public inference service is deployed based on the open-source large model. The service list is as follows (subject to the actual inference service).

Nam Description Free Quota Pro Ma Ma mp xim e χi mu t um Out Te m Co mp put lat Tok nt ext е ens Le Len ng gth th 16, OWE With 72 billion parameters, During the OBT, a 23 16.3 N 2 Qwen2 outperforms most quota of 1 million 00 60 72B previous open-weight models in tokens is provided multiple benchmark tests in for free. After the terms of language quota is used up, understanding, generation, the service is multilingual capabilities, coding, unavailable and mathematics, and inference. It is there is no option also competitive with proprietary to purchase more. The validity period models. is 90 days after the service is enabled. If the validity period expires, the service becomes invalid.

Table 4-2 Public inference service

4.3.2 Enabling an Inference Service

You need to apply for and enable a public inference service before using it. After the public inference service is enabled, you will obtain a certain free quota, which is valid within a certain period of time. If the period of time expires, the service cannot be used. If you want to continue to use the service, you are advised to deploy the inference service.

Prerequisites

- You have a valid Huawei Cloud account.
- You have at least one workspace available.

- **Step 1** Log in to the DataArts Fabric Workspace Management Console.
- Step 2 Select a created workspace, click Access Workspace, and choose Inference Endpoint > Public Endpoint. The Public Inference Endpoint page is displayed.

Step 3 On the **Public Endpoint** page, you can view the public inference endpoints that have been enabled. Endpoints within the trial period are labeled with a "Running" status.

----End

4.3.3 Performing Inference in the Playground

DataArts Fabric provides a playground where you can select inference services and run inference. The playground supports streaming inference, allows you to configure different inference parameters such as **max_tokens**, and supports comparison of different inference services.

Notes and Constraints

The common constraints on using public inference services are as follows:

- Token quota: Each public inference service has a free quota. After the quota is used up, the service is unavailable and there is no option to purchase more. You can share the quota of each public inference service in all of your workspaces at the current site.
- Time: The validity period is 90 days from the date when the service is enabled.
 If the validity period expires, the service becomes invalid. If the same
 inference service is enabled in different workspaces, the time when the service
 was first enabled is used.
- Different models have different context length constraints. For details, see **Table 4-2**.
- Service level agreement (SLA) is not guaranteed. You can create your own inference services for enhanced inference performance.

Prerequisites

- You have a valid Huawei Cloud account.
- You have at least one workspace available.
- You have enabled the public inference service. For details, see **Enabling an Inference Service**.

Procedure

- **Step 1** Log in to the DataArts Fabric Workspace Management Console.
- **Step 2** Select the created workspace and click **Access Workspace**.
- **Step 3** In the navigation pane on the left, choose **Inference Services** > **Public Inference Services**.
- **Step 4** Click **Playgrounds**. The **Playgrounds** page is displayed for inference.
- Step 5 (Optional) Adjust inference parameters.

You can click **Advanced Configuration** to adjust certain inference parameters like **max_tokens**. The following table lists the parameters.

Table 4-3 Inference parameters

Parameter	Description
max_tokens	The maximum number of tokens to be generated during the chat. The value varies depending on different public inference services. For details, see the introduction to public inference services.
temperature	A number used to adjust randomness. The value ranges from 0 to 2. A larger value (for example, 0.8) leads to a more random output, while a smaller value (for example, 0.2) results in a more centralized and deterministic output.
top_p	The nucleus sampling strategy, which is used to control the range of tokens the AI model considers based on the cumulative probability.
frequency_penalty	The frequency penalty, which is used to control the repetition of words in the text to avoid frequent occurrence of some words or phrases in the generated text. The value ranges from –2.0 to 2.0. Positive values penalize new tokens based on how often they have appeared in the existing text, reducing the likelihood that the model repeats the same line verbatim.
presence_penalty	The presence penalty, which is used to control the repetition of topics in the text, avoiding repeated discussions of the same topic or viewpoint in the dialogue or text. The value ranges from -2.0 to 2.0. Positive values penalize new tokens based on whether they have appeared in the text so far, increasing the model's likelihood of talking about new topics.

Step 6 (Optional) Compare multiple inference services.

DataArts Fabric allows you to compare multiple inference services. You can click **Compare** in the upper right corner to add a comparison of up to three inference services.

----End

4.4 Creating My Inference Service for Inference

4.4.1 Creating a Model

In addition to using a public model, you can also create your own model when deploying an inference service on DataArts Fabric. You can create models on the **Model** page of DataArts Fabric. These models belong to you and are invisible to other users.

Notes and Constraints

The common constraints for creating a model are as follows:

• The model you create must be a base model supported by DataArts Fabric. The following table lists the base models.

Table 4-4 Base models

Base Model Type	Description
QWEN_ 2_72B	With 72 billion parameters, Qwen2 outperforms most previous open-weight models in multiple benchmark tests in terms of language understanding, generation, multilingual capabilities, coding, mathematics, and inference. It is also competitive with proprietary models.
GLM_4_ 9B	GLM-4-9B is an open-source version of the latest-generation pre- trained GLM-4 series models launched by Zhipu AI. With 9 billion parameters, it has high performance in datasets evaluation in terms of semantics, mathematics, inference, code, and knowledge.
LLAMA_ 3_8B	With 8 billion parameters, it is the third-generation model of the Llama series. Llama3 has achieved leading performance in multiple benchmark tests. The model uses a large amount of Chinese data for pre-training, expanding the coverage of the Chinese character sets.
LLAMA_ 3_70B	With 70 billion parameters, it is the third-generation model of the Llama series. Llama3 has achieved leading performance in multiple benchmark tests.
LLAMA_ 3.1_8B	Llama3.1 is the first publicly available model and performs well in common sense, steerability, mathematics, tool usage, and multilingual translation. It supports advanced use cases, such as long text summarization, multilingual dialogue agents, and coding assistants. With 8 billion parameters, the model uses a large amount of Chinese data for pre-training, expanding the coverage of the Chinese character sets.
LLAMA_ 3.1_70B	Llama3.1 is the first publicly available model and is close to top AI models in terms of common sense, steerability, mathematics, tool usage, and multilingual translation. It supports advanced use cases, such as long text summarization, multilingual dialogue agents, and coding assistants. The model has 70 billion parameters.

 The model must be in safetensors format. The safetensors format developed by Huggingface is a reliable and easy-to-port storage format of machine learning models. It is used to securely store tensors at a high speed.
 Click the following URLs to view the format requirements for specific model examples.

Base Model Type	Model Example Name	Model Source
LLAMA_3_ 8B	Llama 3 8B Chinese Instruct	https://www.modelscope.cn/models/ FlagAlpha/Llama3-Chinese-8B- Instruct
LLAMA_3_ 70B	Llama 3 70B	https://huggingface.co/meta-llama/ Meta-Llama-3-70B-Instruct
LLAMA_3.1 _8B	Llama 3.1 8B Chinese Chat	https://modelscope.cn/models/XD_AI/ Llama3.1-8B-Chinese-Chat
LLAMA_3.1 _70B	Llama 3.1 70B	https://huggingface.co/meta-llama/ Llama-3.1-70B-Instruct
QWEN_2_7 2B	Qwen 2 72B Instruct	https://huggingface.co/Qwen/ Qwen2-72B
GLM_4_9B	Glm 4 9B Chat	https://huggingface.co/THUDM/ glm-4-9b-chat

Prerequisites

- You have a valid Huawei Cloud account.
- You have at least one workspace available.
- You have created an OBS bucket and folders for storing models, uploaded model files that meet the requirements, and ensured that the OBS bucket for storing models is in the same region as DataArts Fabric. For details, see Creating an OBS Bucket.

- **Step 1** Log in to the DataArts Fabric Workspace Management Console.
- **Step 2** Select the created workspace and click **Access Workspace**.
- **Step 3** In the navigation pane on the left, choose **Resources and Assets** > **Model**. The **Model** page is displayed.
- **Step 4** Click **Create Model**. The **Create Model** page is displayed.
- **Step 5** Enter basic model information, including the name and description, select the OBS path of the model file, and click **Create Now**. For details, see the following table.

Paramet Description er Model Indicates the model name, which is mandatory. Name The name contains 1 to 64 characters and must be unique. Only letters, digits, underscores (_), hyphens (-), periods (.), and spaces are allowed. Model Indicates the model description, which is optional. Descripti The value contains 0 to 1,024 characters. Special characters such as on ^!<>=&"' are not supported. Version Indicates the version name, which is mandatory. Name The name contains 1 to 64 characters and must be unique. Only letters, digits, underscores (_), hyphens (-), periods (.), and spaces are allowed. Version Indicates the version description, which is optional. Descripti The value contains 0 to 1,024 characters. Special characters such as on ^!<>=&"' are not supported. Base Indicates the base model type, which is mandatory. For details, see Model **Table 4-4.** Type Model Indicates the model file path, which is mandatory. Currently, the OBS

Table 4-5 Basic information about the model to be created

Step 6 Click **My Models** again. The created model is displayed in the model list.

----End

File Path

4.4.2 Managing a Model

After creating a model on DataArts Fabric, you can view or delete the model and manage its versions, including adding, viewing, and deleting model versions.

path is supported. The current user must have the read permission

Prerequisites

• You have a valid Huawei Cloud account.

on the path.

- You have at least one workspace available.
- You have created an OBS bucket and folders for storing models, uploaded model files that meet the requirements, and ensured that the OBS bucket for storing models is in the same region as DataArts Fabric. For details, see Creating an OBS Bucket.

Procedure

Step 1 Log in to the DataArts Fabric Workspace Management Console.

- **Step 2** Select the created workspace and click **Access Workspace**.
- **Step 3** In the navigation pane on the left, choose **Resources and Assets** > **Model**. The **Model** page is displayed.
- **Step 4** View the version list of the current model. You can select a version as the current version for use.
- **Step 5** (Optional) Add a model version.

If your model is iteratively updated, you can add a model version.

On the **My Models** page, click **Add Model Version** in the **Operation** column, enter basic information, and click **Add Version**.

A model version cannot be modified after being added. The following table lists the basic information about the new model.

Table 4-6 Basic information about the model to be created

Paramet er	Description
Version Name	Indicates the version name, which is mandatory. The name contains 1 to 64 characters and must be unique. Only letters, digits, underscores (_), hyphens (-), periods (.), and spaces are allowed.
Version Descripti on	Indicates the version description, which is optional. The value contains 0 to 1,024 characters. Special characters such as ^!<>=&"' are not supported.
Model File Path	Indicates the model file path, which is mandatory. Currently, the OBS path is supported. The current user must have the read permission on the path.

Step 6 (Optional) Delete a model version.

You can also delete unnecessary model versions.

Click **Delete** in the **Operation** column and confirm the deletion.

----End

4.4.3 Creating an Inference Endpoint

Before creating an inference service, you need to create an inference endpoint. When creating an inference endpoint, you can configure the maximum number of resources. Then, you can create inference services on the inference endpoint. The total number of resources of all inference services on the inference endpoint cannot exceed the maximum number of resources of the inference endpoint. This helps you control the resource usage of the inference endpoint.

Prerequisites

- You have a valid Huawei Cloud account.
- You have at least one workspace available.

Procedure

- **Step 1** Log in to the DataArts Fabric Workspace Management Console.
- Step 2 Select a created workspace, click Access Workspace, and choose Resources and Assets > Inference Endpoint.
- **Step 3** Click **Create Inference Endpoint** in the upper right corner. Enter the endpoint name, description, resource specifications, and quantity by referring to **Table 4-7**, and click **Create Now**.

Table 4-7 Basic information about creating an inference endpoint

Paramet er	Description
Endpoint Name	Indicates the name of an inference endpoint, which is mandatory. The name contains 1 to 64 characters and must be unique. Only letters, digits, underscores (_), hyphens (-), periods (.), and spaces are allowed.
Descripti on	Indicates the description of an inference service, which is optional. The value contains 0 to 1,024 characters. Special characters such as ^!<>=&"' are not supported.
Comput e Unit Type	This parameter is used to filter resource specifications.
Resource Specifica tions	Indicates the resource specifications, which is mandatory. Different resource specifications support different models.
Pre- warmed Resource s	Currently, only 0 is supported, which is the number of pre-warmed resources of the inference endpoint.
Maximu m Number of Resource s	Indicates the maximum number of resources of an inference endpoint, which is mandatory. The value ranges from 1 to 1,000. In addition, the maximum number of resources cannot be less than the number of pre-warmed resources.

Step 4 Choose **Resources and Assets** > **Inference Endpoint** > **My Endpoint** to view the created endpoint.

----End

4.4.4 Creating an Inference Service

When performing inference on DataArts Fabric, you can select an existing public inference service or deploy your own inference service.

Before deploying an inference service on DataArts Fabric, you need to have a model available. You may use a model you previously created. For convenience, DataArts Fabric also provides several open-source public models by default. The following table lists these models.

Table 4-8 Public models

Model Name	Overview	Base Model Type	Co m pu te (M U)	Ma xi mu m Co nt ext Le ng th	Prompt t Templat e Length	Ma xim um Out put Tok ens
Qwen 2 72B Instruct	With 72 billion parameters, Qwen2 outperforms most previous openweight models in multiple benchmark tests in terms of language understanding, generation, multilingual capabilities, coding, mathematics, and inference. It is also competitive with proprietary models.	QWEN _2_72 B	8	16, 00 0	23	16,3 60
Glm 4 9B Chat	GLM-4-9B is an open-source version of the latest-generation pre-trained GLM-4 series models launched by Zhipu Al. With 9 billion parameters, it has high performance in datasets evaluation in terms of semantics, mathematics, inference, code, and knowledge.	GLM_4 _9B	2	32, 00 0	16	32,7 51

The prompt template length is that of the system prompt. No matter what you enter, the system adds the prompt template to the input. The maximum context length is the sum of the prompt template length, maximum input token length, and maximum output token length.

You can view public model information in the model navigation pane. You can use public models to deploy inference services, but cannot delete public models.

Notes and Constraints

The common constraints on deploying inference services are as follows:

- The resource specifications of the inference service range from 1 to 100.
- The maximum number of inference service resources selected for the inference endpoint cannot exceed the maximum number of resources of the inference endpoint.

Prerequisites

- You have a valid Huawei Cloud account.
- You have at least one workspace available. For details, see Creating a Workspace.
- You have created an inference endpoint. For details, see Creating an Inference Endpoint.
- You have created a model for inference. For details, see Creating a Model.

- **Step 1** Log in to the DataArts Fabric Workspace Management Console.
- **Step 2** Select the created workspace and click **Access Workspace**. In the navigation pane on the left, choose **Development and Production** > **Inference Services**.
- **Step 3** In the upper right corner of the **My Inference Services** tab page on the **Inference Services** page, click **Create Inference Service**.
- **Step 4** Enter the basic information such as the name and description of the inference service to be created, and select the inference endpoint and model. You can click **Public models** or **My models** to select a model. Then, configure the minimum and maximum values of resources. For details, see the following table.

Table 4-9 Parameters for creating an inference service

Parameter Ma nda tor y (Ye s/N o)		nda tor y (Ye s/N	Description
Basic Settin gs	Name	Yes	Indicates the inference service name. The name contains 1 to 64 characters and must be unique. Only letters, digits, underscores (_), hyphens (-), periods (.), and spaces are allowed.
	Descri ption	No	Indicates the description of an inference service. The value contains 0 to 1,024 characters. Special characters such as ^!<>=&"' are not supported.

no to y (Y s/		Ma nda tor y (Ye s/N o)	Description
	Model Type	Yes	You can select My models or Public models .
	Model Yes s		If Model Type is set to My models, select a model you have created from the drop-down list. For details about how to create a model, see Creating a Model.
			If Model Type is set to Public models, select a public inference service from the drop-down list.
	Model Versio n	Yes	If Model Type is set to My models , select a version of a model you have created from the drop-down list.
	Endpoi nt	Yes	Select an inference endpoint you have created from the drop-down list box. For details about how to create an inference endpoint, see Creating an Inference Endpoint.
Instan Resour Yes ce ce Runni Specifi ng cation Settin s		Yes	Indicates the resource specifications, which must be the same as those of the inference endpoint. Otherwise, the specifications are not supported.
gs	Minim um Value	Yes	Indicates the minimum number of instances of an inference service, which is created even if there is no request. The value ranges from 1 to 100. The inference service automatically scales in or out between the minimum and maximum number of instances based on the request load.

Parameter Ma nda tor y (Ye s/N o)		nda tor y (Ye s/N	Description
	Maxim um Value	Yes	Indicates the maximum number of instances of an inference service. The value ranges from 1 to 100. In addition, the maximum value cannot be less than the minimum value, and the maximum value must be less than or equal to the maximum number of resources of the selected inference endpoint. The total maximum number of resources of all inference services under the same inference endpoint must be less than or equal to the maximum number of resources of the selected inference endpoint. The inference service automatically scales in or out between the minimum and maximum number of instances based on the request load. After the request is submitted, the number of instances of the inference service does not exceed the maximum value.

- **Step 5** After the configuration is complete, click **Create Now**.
- **Step 6** On the **Inference Services** page, you can view the created inference service.

----End

4.4.5 Using an Inference Service for Inference

After an inference service is deployed, you can select an existing inference service in the playground for inference or call APIs for inference. For details, see the API document (API reference). The following describes how to use the playground for inference:

Prerequisites

- You have a valid Huawei Cloud account.
- You have at least one workspace available.
- You have created an inference service.

- **Step 1** Log in to the DataArts Fabric Workspace Management Console.
- **Step 2** Select the created workspace and click **Access Workspace**. In the navigation pane on the left, choose **Development and Production** > **Playgrounds**.
- **Step 3** Click **Playgrounds**. The **Playgrounds** page is displayed.
- Step 4 (Optional) Adjust parameters.

If you need to adjust some inference parameters, click **Advanced Configuration** to adjust parameters such as **max_tokens**. The following lists the parameters.

Table 4-10 Inference parameters

Parameter	Description
max_tokens	The maximum number of tokens to be generated during the chat. The value varies depending on different public inference services. For details, see the introduction to public inference services.
temperature	A number used to adjust randomness. The value ranges from 0 to 2. A larger value (for example, 0.8) leads to a more random output, while a smaller value (for example, 0.2) results in a more centralized and deterministic output.
top_p	The nucleus sampling strategy, which is used to control the range of tokens the AI model considers based on the cumulative probability.
frequency_penalty	The frequency penalty, which is used to control the repetition of words in the text to avoid frequent occurrence of some words or phrases in the generated text. The value ranges from –2.0 to 2.0. Positive values penalize new tokens based on how often they have appeared in the existing text, reducing the likelihood that the model repeats the same line verbatim.
presence_penalty	The presence penalty, which is used to control the repetition of topics in the text, avoiding repeated discussions of the same topic or viewpoint in the dialogue or text. The value ranges from -2.0 to 2.0. Positive values penalize new tokens based on whether they have appeared in the text so far, increasing the model's likelihood of talking about new topics.

Step 5 (Optional) Compare multiple inferences.

You can click **Compare** in the upper right corner to compare multiple inference services. A maximum of three inference services can be compared.

----End

4.4.6 Deleting an Inference Service

You can delete an inference service you have created as required.

Prerequisites

- You have a valid Huawei Cloud account.
- You have at least one workspace available.

You have created an inference service.

Procedure

- **Step 1** Log in to the DataArts Fabric Workspace Management Console.
- Step 2 Select the created workspace and click Access Workspace. Choose Development and Production > Inference Services.
- **Step 3** Select the inference service to be deleted and click **Delete** in the **Operation** column.
- **Step 4** In the displayed dialog box, enter **DELETE** and click **OK**.

----End

4.4.7 Deleting an Inference Endpoint

Prerequisites

- You have a valid Huawei Cloud account.
- You have at least one workspace available.
- You have created an inference endpoint.

Procedure

- **Step 1** Log in to the DataArts Fabric Workspace Management Console.
- **Step 2** Select a created workspace, click **Access Workspace**, and choose **Resources and Assets** > **Inference Endpoint**.
- **Step 3** Click the trash can icon in the upper right corner of the inference endpoint to be deleted and confirm the deletion.

----End

4.5 Viewing All Metrics on AOM

To enable you to query the usage of inference instance resources, the cloud service platform reports metrics to AOM.

Prerequisites

- You have a valid Huawei Cloud account.
- You have at least one workspace available.
- At least one inference instance is available.

- **Step 1** Log in to the AOM console.
- **Step 2** Choose **Metric Browsing** and set **Metric Sources** to **Prometheus_AOM_Default**.

Step 3 Enter a metric name in the **All metrics** tab to query the metric.

Table 4-11 Monitoring metrics

Metric	Description
mu_usage	This metric indicates the actual MU usage of the current inference instance. Unit: number.

----End

5 O&M Management

5.1 Configuring Message Notifications

Message notifications are used to notify you of the job execution status.

Prerequisites

- You have a valid Huawei Cloud account.
- You have at least one workspace available.
- You have configured the FABRIC_SMN_POLICY agency. For details, see Configuring DataArts Fabric Service Agency Permissions.

- **Step 1** Log in to the DataArts Fabric Workspace Management Console.
- Step 2 Select the created workspace, click Access Workspace, and choose O&M Management > Message Notifications.
- **Step 3** Click **Create Notification** in the upper right corner, set parameters by referring to **Table 5-1**, and click **Create Now**.

Table 5-1 Parameters for creating a notification

Parameter	Ma nda tor y (Ye s/N o)	Description
SMN Topic	Yes	Topic created in SMN. Messages will be sent to the corresponding topic.

Parameter	Ma nda tor y (Ye s/N o)	Description
Event	Yes	Notification time. The options include:
		Success Notification
		Failure Notification
		You can select both.
Message Type	Yes	Currently, only the job type is supported. The job execution result will be notified.
Message Source Matching Style	Yes	Message source to be matched. Regular expressions are supported.
		Job scenario: regular expression match of the job name.
		For example, if a job named is test-job , you can enter test-j.* or test-job for matching.

Step 4 After the creation is successful, you can view the created message notification in the list. You can click the number next to the message source to view the configured message source.

□ NOTE

For the same topic, different message sources of notification events and message types are combined into one record.

----End

5.2 Deleting a Notification

Message notifications are used to notify you of the job execution status. You can delete a notification that is not needed.

Prerequisites

- You have a valid Huawei Cloud account.
- You have at least one workspace available.
- You have configured the FABRIC_SMN_POLICY agency. For details, see Configuring DataArts Fabric Service Agency Permissions.
- You have at least one notification.

Procedure

Step 1 Log in to the DataArts Fabric Workspace Management Console.

- Step 2 Select the created workspace, click Access Workspace, and choose O&M Management > Message Notifications.
- **Step 3** Click the number next to the message source. In the displayed dialog box, select the target message type and click **Delete**.

----End