ModelArts

FAQs

Issue	01
Date	2023-11-25

HUAWEI TECHNOLOGIES CO., LTD.

Trademarks and Permissions

Notice

The purchased products, services and features are stipulated by the contract made between Huawei 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 Technologies Co., Ltd.

Address: Huawei Industrial Base
Bantian, Longgang
Shenzhen 518129
People's Republic of China

Website: https://www.huawei.com

Email: support@huawei.com

Security Declaration

Vulnerability

Huawei's regulations on product vulnerability management are subject to "Vul. Response Process". For details about the policy, see the following website:https://www.huawei.com/en/psirt/vul-response-process
For enterprise customers who need to obtain vulnerability information, visit:https://securitybulletin.huawei.com/enterprise/en/security-advisory

Contents

1 General Issues

1.1 What Is ModelArts?

ModelArts is a one-stop AI development platform geared toward developers and data scientists of all skill levels. It enables you to rapidly build, train, and deploy models anywhere (from the cloud to the edge), and manage full-lifecycle AI workflows. ModelArts accelerates AI development and fosters AI innovation with key capabilities, including data preprocessing and auto labeling, distributed training, automated model building, and one-click workflow executing.

The one-stop ModelArts platform covers all stages of AI development, including data processing, AI application creation, and model training and deployment. The underlying layer of ModelArts supports various heterogeneous computing resources. You can flexibly select and use the resources without having to consider the underlying technologies. In addition, ModelArts supports popular open-source AI development frameworks such as TensorFlow and MXNet. Developers can also use self-developed algorithm frameworks to match their usage habits.

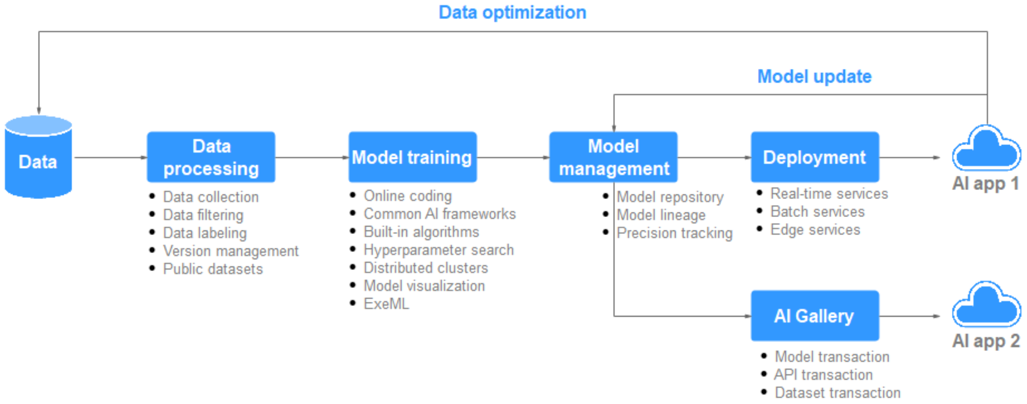
ModelArts aims to achieve simple, convenient AI development. ModelArts is adaptive to the requirements of AI developers of different experience. For example, service developers can use ExeML to quickly build AI applications without coding. AI beginners do not need to pay attention to model development, but directly use built-in algorithms to build AI applications. AI engineers can use multiple development environments to compile code for quick modeling and application development.

Product Architecture

ModelArts is a one-stop AI development platform that supports the entire development process, including data processing, AI application management and deployment, and model training, and provides AI Gallery for sharing models.

ModelArts supports all sorts of AI application scenarios, such as image classification, object detection, video analysis, speech recognition, product recommendation, and exception detection.

Figure 1-1 ModelArts architecture



1.2 What Are the Relationships Between ModelArts and Other Services?

IAM

ModelArts uses Identity and Access Management (IAM) for authentication and authorization. For more information about IAM, see ***Identity and Access Management User Guide***.

OBS

ModelArts uses Object Storage Service (OBS) to securely and reliably store data and models at low costs. For more details, see ***Object Storage Service Console Operation Guide***.

Table 1-1 Relationship between ModelArts and OBS

Function	Subtask	Relationship
ExeML	Data labeling	The data labeled on ModelArts is stored in OBS.
	Auto training	After a training job is completed, the generated model is stored in OBS.
	Service deployment	ModelArts deploys models stored in OBS as real-time services.
AI development lifecycle	Data management	● Datasets are stored in OBS. ● The dataset labeling information is stored in OBS. ● Data can be imported from OBS.

Function	Subtask	Relationship
	Development environment	Data or code files in a notebook instance are stored in OBS.
	Model training	- The datasets used by training jobs are stored in OBS. - The running scripts for training jobs are stored in OBS. - The models generated by training jobs are stored in the specified OBS directories. - The run logs of training jobs are stored in the specified OBS directories.
	AI application management	After a training job is completed, the generated model is stored in OBS. You can import the model from OBS.
	Service deployment	Models stored in OBS can be deployed as services.
Settings	-	Authorizes ModelArts to access OBS (using an agency or access key) so that ModelArts can use OBS to store data and create notebook instances.

CCE

ModelArts uses Cloud Container Engine (CCE) to deploy models as real-time services. CCE enables high concurrency and provides elastic scaling. For more information about CCE, see **Cloud Container Engine User Guide**.

SWR

To use an AI framework that is not supported by ModelArts, use Software Repository for Container (SWR) to customize an image and import the image to ModelArts for training or inference. For details about SWR, see ***Software Repository for Container User Guide***.

1.3 What Are the Differences Between ModelArts and DLS?

Deep Learning Service (DLS) is a one-stop deep learning platform based on the high-performance computing capabilities of HUAWEI CLOUD. With various optimized neural network models, DLS allows you to easily implement model training and evaluation with the flexibility of on-demand scheduling.

However, DLS supports only the deep learning technologies, while ModelArts integrates both the deep learning and machine learning technologies. In addition, ModelArts is a one-stop AI development platform, which manages the AI development lifecycle from data labeling, algorithm development, to model training and deployment. To be specific, ModelArts contains and supports the functions and features of DLS. Currently, DLS is terminated on HUAWEI CLOUD. The functions related to deep learning can be directly used in ModelArts. If you are a DLS user, you can also migrate the data in DLS to ModelArts.

1.4 How Do I Purchase or Enable ModelArts?

ModelArts is an out-of-the-box platform and does not need to be purchased or enabled. You can directly log in to the ModelArts console, complete the global configuration, and use required functions.

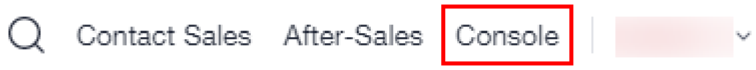
On ModelArts, only the functions that use compute specifications are billed. All public resource pools are billed in pay-per-use mode based on the selected specifications and job running duration. You can purchase a dedicated resource pool on a pay-per-use or yearly/monthly basis. When running a training job or deploying a service, you can use your dedicated resource pool without paying for extra fees.

1.5 How Do I Obtain an Access Key?

Obtaining an Access Key

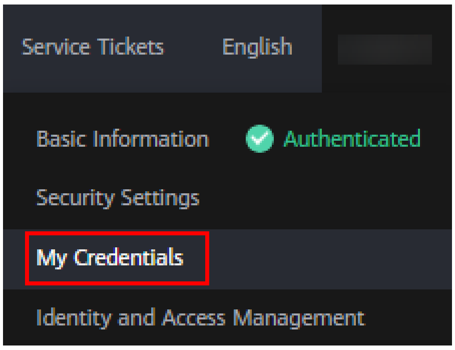
1. Log in to **HUAWEI CLOUD** and click **Console** in the upper right corner of the page to access the HUAWEI CLOUD management console.

   Figure 1-2 Console

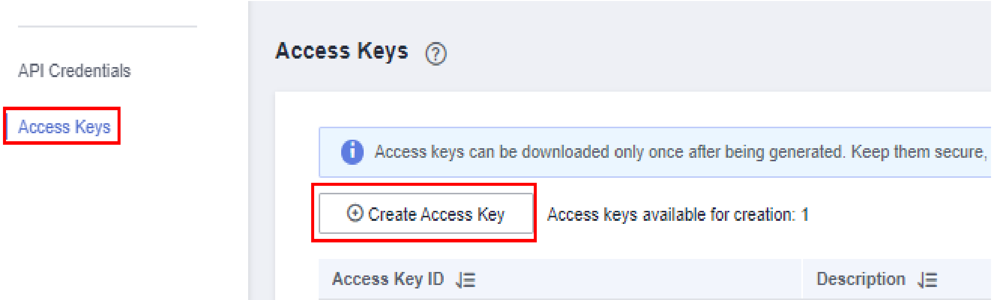
2. Hover the cursor over the account name in the upper right corner of the console and choose **My Credentials** from the drop-down list. The **API Credentials** page is displayed.

Figure 1-3 My Credentials



3. On the **API Credentials** page, choose **Access Keys** > **Create Access Key**.

Figure 1-4 Create Access Key

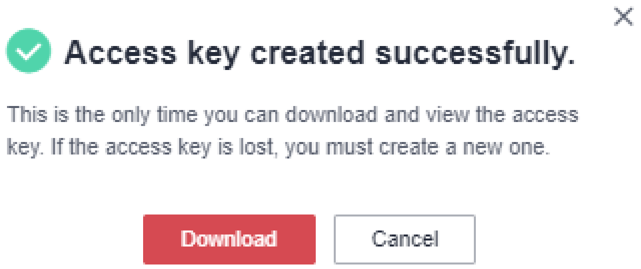


Access keys can be downloaded only once after being generated. Keep them secure,

Access keys available for creation: 1

4. Enter the description of the key and click **OK**. Click **Download** to download the key.

Figure 1-5 Access key created



Access key created successfully.

This is the only time you can download and view the access key. If the access key is lost, you must create a new one.

5. The access key file is saved in the default downloads folder of the browser. Open the **credentials.csv** file to view the access key with an AK and SK.

1.6 How Do I Upload Data to OBS?

Before using ModelArts to develop AI models, data needs to be uploaded to an OBS bucket. You can log in to the OBS console to create an OBS bucket, create a folder in it, and upload data. For details about how to upload data, see ***Object Storage Service Getting Started***.

1.7 What Do I Do If the System Displays a Message Indicating that the AK/SK Pair Is Unavailable?

Issue Analysis

An AK and SK form a key pair required for accessing OBS. Each SK corresponds to a specific AK, and each AK corresponds to a specific user. If the system displays a message indicating that the AK/SK pair is unavailable, it is possible that the account is in arrears or the AK/SK pair is incorrect.

Solution

1. Use the current account to log in to the OBS console and check whether the current account can access OBS.
   - If the account can access OBS, rectify the fault by referring to **2**.
   - If the account cannot access OBS, rectify the fault by referring to **3**.
2. If the account can access OBS, click the username in the upper right corner and select **My Credentials** from the drop-down list. Then, follow the instructions provided in **Access Keys** to check whether the AK/SK pair is created using the current account.
   - If yes, submit a service ticket.
   - If not, replace the AK/SK with those created using the current account. For details, see **Access Keys**.
3. If the account cannot access OBS, check whether it is in arrears.
   - If the account balance is insufficient, top up the account. For details, see **Topping up an Account**.
   - If the account is not in arrears and the system displays a message indicating that the resource reservation is overdue, submit a **service ticket** to apply for OBS resources.

1.8 What Do I Do If a Message Indicating Insufficient Permissions Is Displayed When I Use ModelArts?

If a message indicating insufficient permissions is displayed when you use ModelArts, perform the operations described in this section to grant permissions for related services as needed.

The permissions to use ModelArts depend on OBS authorization. Therefore, ModelArts users require OBS system permissions as well.

- For details about how to grant a user full permissions for OBS and common operations permissions for ModelArts, see **Configuring Common Operations Permissions**.
- For details about how to manage user permissions on OBS and ModelArts in a refined manner and configure custom policies, see **Creating a Custom Policy for ModelArts**.

Configuring Common Operations Permissions

To use the basic functions of ModelArts, assign the **ModelArts CommonOperations** permission on project-level services to users. Since ModelArts depends on OBS permissions, assign the **OBS Administrator** permission on global services to users.

The procedure is as follows:

Step 1 Create a user group.

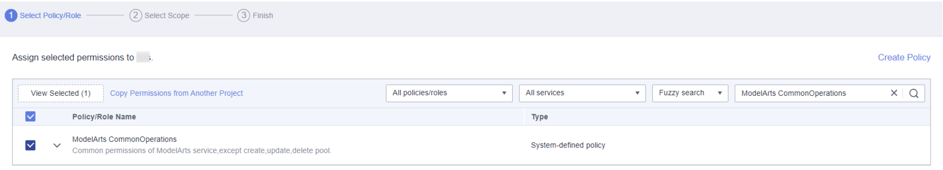
Log in to the IAM console and choose **User Groups** > **Create User Group**. Enter a user group name, and click **OK**.

Step 2 Configure permissions for the user group.

In the user group list, locate the user group created in **step 1**, click **Authorize** , and perform the following operations.

1. Assign the **ModelArts CommonOperations** permission on project-level services to the user group and click **OK**.

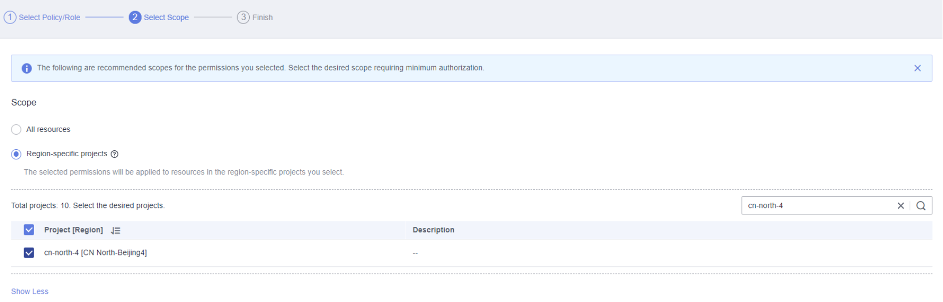
Figure 1-6 Assigning the ModelArts CommonOperations permission



Assign selected permissions to [redacted].

☑	Policy/Role Name	Type
☑	ModelArts CommonOperations Common permissions of ModelArts service,except create,update,delete pool.	System-defined policy

Figure 1-7 Setting Scope to Region-specific projects



The following are recommended scopes for the permissions you selected. Select the desired scope requiring minimum authorization.

Scope

○ All resources

● Region-specific projects

The selected permissions will be applied to resources in the region-specific projects you select.

Total projects: 10. Select the desired projects. cn-north-4

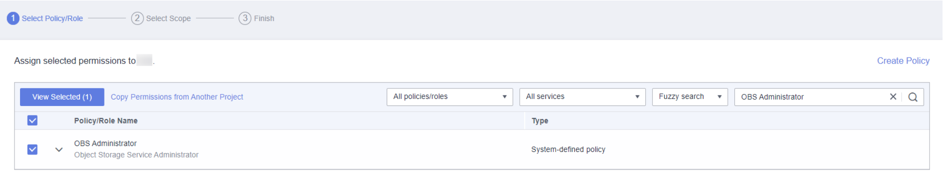
☑	Project [Region]	Description
☑	cn-north-4 [CN North-Beijing4]	--

NOTE

The permission takes effect only in assigned regions. Assign permissions in all regions if the permission is required in all regions.

2. Assign the **OBS Administrator** permission on global services to the user group and click **OK**.

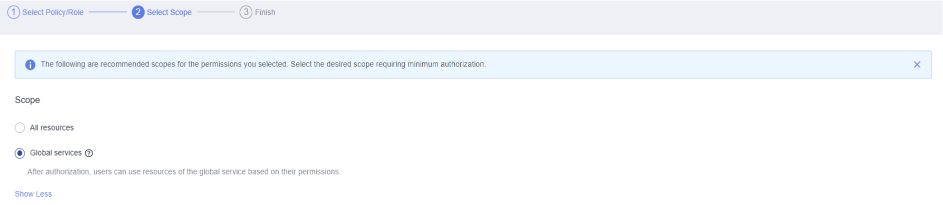
Figure 1-8 Assigning the OBS Administrator permission



Assign selected permissions to [redacted].

☑	Policy/Role Name	Type
☑	OBS Administrator Object Storage Service Administrator	System-defined policy

Figure 1-9 Setting Scope to Global services



The following are recommended scopes for the permissions you selected. Select the desired scope requiring minimum authorization.

Scope

○ All resources

● Global services

After authorization, users can use resources of the global service based on their permissions.

Step 3 **Create a user and add it to the user group**.

Create a user on the IAM console and add the user to the user group created in **step 1**.

Step 4 **Log in** and verify permissions.

Log in to the ModelArts console as the created user, switch to the authorized region, and verify the **ModelArts CommonOperations** and **Tenant Administrator** policies are in effect.

- Choose **Service List** > **ModelArts**. Choose **Dedicated Resource Pools**. On the page that is displayed, select a resource pool type and click **Create**. You should not be able to create a new resource pool.
- Choose any other service in **Service List**. Assuming that the current permissions contain only **ModelArts CommonOperations**, you should get a message indicating that you have insufficient permissions.
- Choose **Service List** > **ModelArts**. On the ModelArts console, choose **Data Management** > **Datasets** > **Create Dataset**. You should be able to access the corresponding OBS path.

----End

Creating a Custom Policy for ModelArts

In addition to the default system policies of ModelArts, you can create custom policies, which can address OBS permissions as well. For more information, see **Creating a Custom Policy**.

You can create custom policies in the visual editor or by creating a JSON file. This section describes how to use a JSON file to configure a custom policy to grant permissions required to use the development environment, and how to configure the minimum OBS permissions for ModelArts users.

NOTE

A custom policy can contain actions for multiple services that are accessible globally or only for region-specific projects.

ModelArts is a project-level service, but OBS is a global service, so you need to create separate policies for the two services and then apply these policies to the users.

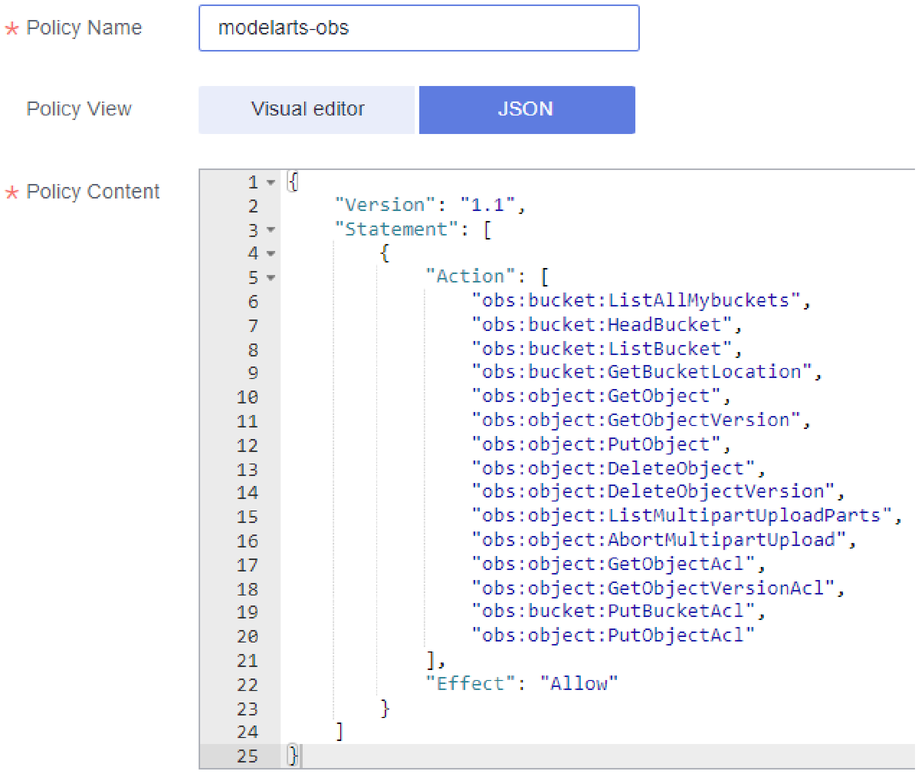
1. Create a custom policy for minimizing permissions for OBS that ModelArts depends on. See **Figure 1-10**.

   Log in to the IAM console, choose **Permissions** > **Policies/Roles**, and click **Create Custom Policy**. Configure the parameters as follows:

   – **Policy Name**: Choose a custom policy name.

- **Policy View**: **JSON**
- **Policy Content**: Follow the instructions in **Example Custom Policies of OBS**. For more information about OBS system permissions, see **OBS Permissions Management**.

Figure 1-10 Minimum permissions for OBS



* Policy Name: modelarts-obs

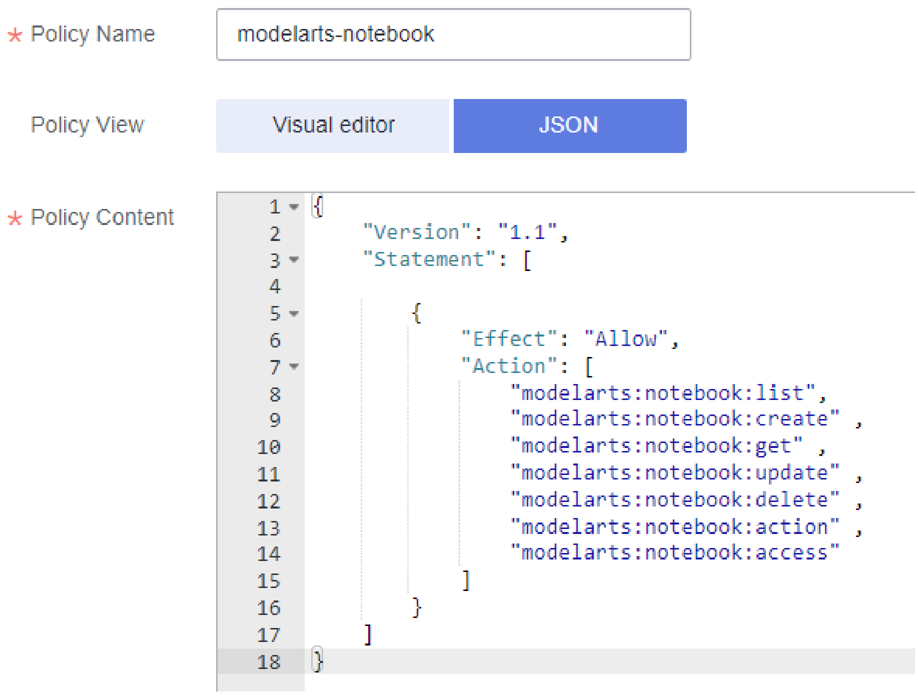
Policy View: Visual editor | JSON

* Policy Content

```
{
    "Version": "1.1",
    "Statement": [
        {
            "Action": [
                "obs:bucket:ListAllMybuckets",
                "obs:bucket:HeadBucket",
                "obs:bucket:ListBucket",
                "obs:bucket:GetBucketLocation",
                "obs:object:GetObject",
                "obs:object:GetObjectVersion",
                "obs:object:PutObject",
                "obs:object:DeleteObject",
                "obs:object:DeleteObjectVersion",
                "obs:object:ListMultipartUploadParts",
                "obs:object:AbortMultipartUpload",
                "obs:object:GetObjectAcl",
                "obs:object:GetObjectVersionAcl",
                "obs:bucket:PutBucketAcl",
                "obs:object:PutObjectAcl"
            ],
            "Effect": "Allow"
        }
    ]
}
```

2. Create a custom policy for the permission to use the ModelArts development environment. See **Figure 1-11**. Configure the parameters as follows:
   - **Policy Name**: Choose a custom policy name.
   - **Policy View**: **JSON**
   - **Policy Content**: Follow the instructions in **Example Custom Policies for Using the ModelArts Development Environment**. For the actions that can be added for custom policies, see **ModelArts API Reference > Permissions Policies and Supported Actions**.

Figure 1-11 Permission to use the development environment



* Policy Name: modelarts-notebook

* Policy Content

```
{
    "Version": "1.1",
    "Statement": [

        {
            "Effect": "Allow",
            "Action": [
                "modelarts:notebook:list",
                "modelarts:notebook:create" ,
                "modelarts:notebook:get" ,
                "modelarts:notebook:update" ,
                "modelarts:notebook:delete" ,
                "modelarts:notebook:action" ,
                "modelarts:notebook:access"
            ]
        }
    ]
}
```

- For the system policies of other services, see **System Permissions**.

3. On the IAM console, **create a user group and grant required permissions.**

   After creating a user group on the IAM console, grant the custom policy created in **1** to the user group.

4. **Create a user and add it to the user group**.

   Create a user on the IAM console and add the user to the group created in **3**.

5. **Log in** and verify permissions.

   Log in to the ModelArts console as the created user, switch to the authorized region, and verify the **ModelArts CommonOperations** and **Tenant Administrator** policies are in effect.

   - Choose **Service List** > **ModelArts**. On the ModelArts console, choose **Data Management** > **Datasets**. If you cannot create a dataset, the permissions (for using the development environment) granted only to ModelArts users have taken effect.
   - Choose **Service List** > **ModelArts**. On the ModelArts console, choose **DevEnviron** > **Notebooks** > **Create**. You should be able to access the OBS path specified in **Storage Path**.

Example Custom Policies of OBS

The permissions to use ModelArts require OBS authorization. The following example shows the minimum OBS required, including the permissions for OBS buckets and objects. After being granted the minimum permissions for OBS, users can access OBS from ModelArts without restrictions.

```
{
    "Version": "1.1",
    "Statement": [
        {
            "Action": [
                "obs:bucket:ListAllMybuckets",
                "obs:bucket:HeadBucket",
                "obs:bucket:ListBucket",
                "obs:bucket:GetBucketLocation",
                "obs:object:GetObject",
                "obs:object:GetObjectVersion",
                "obs:object:PutObject",
                "obs:object:DeleteObject",
                "obs:object:DeleteObjectVersion",
                "obs:object:ListMultipartUploadParts",
                "obs:object:AbortMultipartUpload",
                "obs:object:GetObjectAcl",
                "obs:object:GetObjectVersionAcl",
                "obs:bucket:PutBucketAcl",
                "obs:object:PutObjectAcl"
            ],
            "Effect": "Allow"
        }
    ]
}
```

Example Custom Policies for Using the ModelArts Development Environment

```
{
    "Version": "1.1",
    "Statement": [

        {
            "Effect": "Allow",
            "Action": [
                "modelarts:notebook:list",
                "modelarts:notebook:create" ,
                "modelarts:notebook:get" ,
                "modelarts:notebook:update" ,
                "modelarts:notebook:delete" ,
                "modelarts:notebook:action" ,
                "modelarts:notebook:access"
            ]
        }
    ]
}
```

1.9 How Do I Use ModelArts to Train Models Based on Structured Data?

For most users, ModelArts provides the predictive analytics function of ExeML to train models based on structured data.

For more advanced users, ModelArts provides the notebook creation function of DevEnviron for code development. It allows the users to create training tasks with large volumes of data in training jobs and use the engines such as Scikit_Learn, XGBoost, or Spark_MLlib in the development and training processes.

1.10 What Are Regions and AZs?

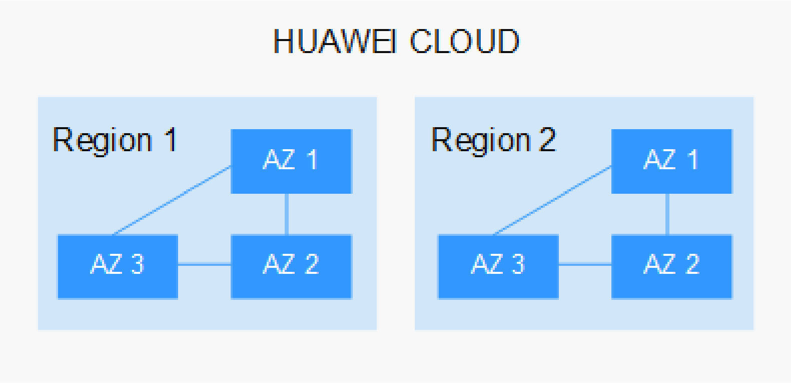
Concept

A region and availability zone (AZ) identify the location of a data center. You can create resources in a specific region and AZ.

- Regions are divided based on geographical location and network latency. Public services, such as Elastic Cloud Server (ECS), Elastic Volume Service (EVS), Object Storage Service (OBS), Virtual Private Cloud (VPC), Elastic IP (EIP), and Image Management Service (IMS), are shared within the same region. Regions are classified into universal regions and dedicated regions. A universal region provides universal cloud services for common tenants. A dedicated region provides specific services for specific tenants.
- An AZ contains one or more physical data centers. Each AZ has independent cooling, fire extinguishing, moisture-proof, and electricity facilities. Within an AZ, computing, network, storage, and other resources are logically divided into multiple clusters. AZs within a region are interconnected using high-speed optical fibers to support cross-AZ high-availability systems.

Figure 1-12 shows the relationship between regions and AZs.

Figure 1-12 Regions and AZs



HUAWEI CLOUD provides services in many regions around the world. Select a region and AZ based on requirements. For more information, see **Global Products and Services**.

Selecting a Region

When selecting a region, consider the following factors:

- Location

  It is recommended that you select the closest region for low network latency and quick access. Regions within the Chinese mainland provide the same infrastructure, BGP network quality, as well as resource operations and configurations. Therefore, if your target users are on the Chinese mainland, you do not need to consider the network latency differences when selecting a region.

  - If your target users are in Asia Pacific (excluding the Chinese mainland), select the **CN-Hong Kong**, **AP-Bangkok**, or **AP-Singapore** region.
  - If your target users are in Africa, select the **AF-Johannesburg** region.
  - If your target users are in Europe, select the **EU-Paris** region.
  - If your target users are in Latin America, select the **LA-Santiago** region.

    NOTE

    The **LA-Santiago** region is located in Chile.

- Resource price

  Resource prices may vary in different regions. For details, see **Product Pricing Details**.

Selecting an AZ

When deploying resources, consider your applications' requirements on disaster recovery (DR) and network latency.

- For high DR capability, deploy resources in different AZs within the same region.
- For lower network latency, deploy resources in the same AZ.

Regions and Endpoints

Before you use an API to call resources, specify its region and endpoint. For more details, see **Regions and Endpoints**.

1.11 How Do I View All Files Stored in OBS on ModelArts?

To view all files stored in OBS when using notebook instances or training jobs, use either of the following methods:

- OBS console

  Log in to OBS console using the current account, and search for the OBS buckets, folders, and files.
- You can use an API to check whether a given directory exists. In an existing notebook instance or after creating a new notebook instance, run the following command to check whether the directory exists:

  ```
  import moxing as mox
  mox.file.list_directory('obs://bucket_name', recursive=True)
  ```

  If there are a large number of files, wait until the final file path is displayed.

1.12 Where Are Datasets of ModelArts Stored in a Container?

Datasets of ModelArts and data in specific data storage locations are stored in OBS.

1.13 Which AI Frameworks Does ModelArts Support?

The AI frameworks and versions supported by ModelArts vary slightly based on the development environment notebook, training jobs, and model inference (AI application management and deployment). The following describes the AI frameworks supported by each module.

Development Environment Notebook

The image and versions supported by development environment notebook instances vary based on runtime environments.

Table 1-2 Images supported by notebook of the new version

Image	Description	Supported Chip	Remote SSH	Online JupyterLab
pytorch1.8-cuda10.2-cudnn7-ubuntu18.04	CPU- or GPU-powered public image for general algorithm development and training, with built-in AI engine PyTorch 1.8	CPU or GPU	Yes	Yes
mindspore1.7.0-cuda10.1-py3.7-ubuntu18.04	CPU- or GPU-powered general algorithm development and training, preconfigured with AI engine MindSpore 1.7.0 and CUDA 10.1	CPU or GPU	Yes	Yes
mindspore1.7.0-py3.7-ubuntu18.04	CPU-powered general algorithm development and training, preconfigured with AI engine MindSpore 1.7.0	CPU	Yes	Yes
pytorch1.10-cuda10.2-cudnn7-ubuntu18.04	CPU- or GPU-powered general algorithm development and training, preconfigured with AI engine PyTorch 1.10 and CUDA 10.2	CPU or GPU	Yes	Yes

Image	Description	Supported Chip	Remote SSH	Online JupyterLab
tensorflow2.1-cuda10.1-cudnn7-ubuntu18.04	CPU- or GPU-powered public image for general algorithm development and training, with built-in AI engine TensorFlow 2.1	CPU or GPU	Yes	Yes
conda3-ubuntu18.04	Clean customized base image only includes Conda	CPU	Yes	Yes
pytorch1.4-cuda10.1-cudnn7-ubuntu18.04	CPU- or GPU-powered public image for general algorithm development and training, with built-in AI engine PyTorch 1.4	CPU or GPU	Yes	Yes
tensorflow1.13-cuda10.0-cudnn7-ubuntu18.04	GPU-powered public image for general algorithm development and training, with built-in AI engine TensorFlow 1.13.1	GPU	Yes	Yes
conda3-cuda10.2-cudnn7-ubuntu18.04	Clean customized base image includes CUDA 10.2, Conda	CPU	Yes	Yes
spark2.4.5-ubuntu18.04	CPU-powered algorithm development and training, preconfigured with PySpark 2.4.5 and can be attached to preconfigured Spark clusters including MRS and DLI	CPU	No	Yes

Image	Description	Supported Chip	Remote SSH	Online JupyterLab
mindspore1.2.0-cuda10.1-cudnn7-ubuntu18.04	GPU-powered public image for algorithm development and training, with built-in AI engine MindSpore-GPU	GPU	Yes	Yes
mindspore1.2.0-openmpi2.1.1-ubuntu18.04	CPU-powered public image for algorithm development and training, with built-in AI engine MindSpore-CPU	CPU	Yes	Yes

Training Jobs

The following table lists the AI engines.

The built-in training engines in the new version are named in the following format:

```
<Training engine name_version>-[cpu | <cuda_version | cann_version >]-<py_version>-<OS name_version>-<
x86_64 | aarch64>
```

Table 1-3 AI engines supported by training jobs of the new version

Runtime Environment	Supported Chip	System Architecture	System Version	AI Engine and Version	Supported CUDA or Ascend Version
TensorFlow	CPU or GPU	x86_64	Ubuntu 18.04	tensorflow_2.1.0-cuda_10.1-py_3.7-ubuntu_18.04-x86_64	CUDA 10.1
PyTorch	CPU or GPU	x86_64	Ubuntu 18.04	pytorch_1.8.0-cuda_10.2-py_3.7-ubuntu_18.04-x86_64	CUDA 10.2

Runtime Environment	Supported Chip	System Architecture	System Version	AI Engine and Version	Supported CUDA or Ascend Version
MPI	GPU	x86_64	Ubuntu 18.04	mindspore_1.3.0-cuda_10.1-py_3.7-ubuntu_1804-x86_64	CUDA 10.1
Horovod	GPU	x86_64	Ubuntu 18.04	horovod_0.20.0-tensorflow_2.1.0-cuda_10.1-py_3.7-ubuntu_18.04-x86_64	CUDA 10.1
				horovod_0.22.1-pytorch_1.8.0-cuda_10.2-py_3.7-ubuntu_18.04-x86_64	CUDA 10.2

Table 1-4 AI engines supported by training jobs of the old version

Runtime Environment	Supported Chip	System Architecture	System Version	AI Engine and Version	Supported CUDA or Ascend Version
TensorFlow	CPU or GPU	x86_64	Ubuntu 16.04	TF-1.8.0-python2.7	-
				TF-1.8.0-python3.6	-
				TF-1.13.1-python2.7	-
				TF-1.13.1-python3.6	-
				TF-2.1.0-python3.6	-
MXNet	CPU or GPU	x86_64	Ubuntu 16.04	MXNet-1.2.1-python2.7	-
				MXNet-1.2.1-python3.6	-
Spark_MLlib	CPU	x86_64	Ubuntu 16.04	Spark-2.3.2-python3.6	-
				Spark-2.3.2-python2.7	-

Runtime Environment	Supported Chip	System Architecture	System Version	AI Engine and Version	Supported CUDA or Ascend Version
Ray	CPU or GPU	x86_64	Ubuntu 16.04	RAY-0.7.4-python3.6	-
PyTorch	CPU or GPU	x86_64	Ubuntu 16.04	PyTorch-1.0.0-python2.7	-
				PyTorch-1.0.0-python3.6	-
				PyTorch-1.3.0-python2.7	-
				PyTorch-1.3.0-python3.6	-
				PyTorch-1.4.0-python3.6	-
Caffe	CPU or GPU	x86_64	Ubuntu 16.04	Caffe-1.0.0-python2.7	CUDA 8.0
MindSpore-GPU	GPU	x86_64	Ubuntu 18.04	MindSpore-1.1.0-python3.7	-
				MindSpore-1.2.0-python3.7	-

Supported AI Engines for ModelArts Inference

If you import a model from a template or OBS to create an AI application, the following AI engines and versions are supported.

NOTE

- Runtime environments marked with **recommended** are unified runtime images, which will be used as mainstream base inference images. The installation packages of unified images are richer. For details, see **Base Inference Images**.
- Images of the old version will be discontinued. Use unified images.
- The base images to be removed are no longer maintained.
- Naming a unified runtime image: <*AI engine name and version*> - <*Hardware and version*: CPU, CUDA, or CANN> - <*Python version*> - <*OS version*> - <*CPU architecture*>

Table 1-5 Supported AI engines and their runtime

Engine	Runtime	Note
TensorFlow	python3.6 python2.7 (unavailable soon) tf1.13-python3.6-gpu tf1.13-python3.6-cpu tf1.13-python3.7-cpu tf1.13-python3.7-gpu tf2.1-python3.7 (unavailable soon) tensorflow_2.1.0-cuda_10.1-py_3.7-ubuntu_18.04-x86_64 (recommended)	● TensorFlow 1.8.0 is used in **python2.7** and **python3.6**. ● **python3.6**, **python2.7**, and **tf2.1-python3.7** indicate that the model can run on both CPUs and GPUs. For other runtime values, if the suffix contains **cpu** or **gpu**, the model can run only on CPUs or GPUs. ● The default runtime is **python2.7**.
Spark_MLlib	python2.7 (unavailable soon) python3.6 (unavailable soon)	● Spark_MLlib 2.3.2 is used in **python2.7** and **python3.6**. ● The default runtime is **python2.7**. ● **python2.7** and **python3.6** can only be used to run models on CPUs.
Scikit_Learn	python2.7 (unavailable soon) python3.6 (unavailable soon)	● Scikit_Learn 0.18.1 is used in **python2.7** and **python3.6**. ● The default runtime is **python2.7**. ● **python2.7** and **python3.6** can only be used to run models on CPUs.
XGBoost	python2.7 (unavailable soon) python3.6 (unavailable soon)	● XGBoost 0.80 is used in **python2.7** and **python3.6**. ● The default runtime is **python2.7**. ● **python2.7** and **python3.6** can only be used to run models on CPUs.

Engine	Runtime	Note
PyTorch	python2.7 (unavailable soon) python3.6 python3.7 pytorch1.4-python3.7 pytorch1.5-python3.7 (unavailable soon) pytorch_1.8.0-cuda_10.2-py_3.7-ubuntu_18.04-x86_64 (recommended)	● PyTorch 1.0 is used in **python2.7**, **python3.6**, and **python3.7**. ● **python2.7**, **python3.6**, **python3.7**, **pytorch1.4-python3.7**, and **pytorch1.5-python3.7** indicate that the model can run on both CPUs and GPUs. ● The default runtime is **python2.7**.
MindSpore	aarch64 (recommended)	AArch64 can run only on Snt3 chips.

1.14 What Are the Functions of ModelArts Training and Inference?

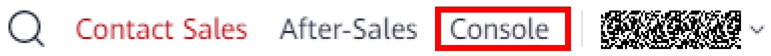
ModelArts training includes ExeML, training management, and dedicated resource pools (for development/training).

ModelArts inference includes AI application management and deployment.

1.15 How Do I View an Account ID and IAM User ID?

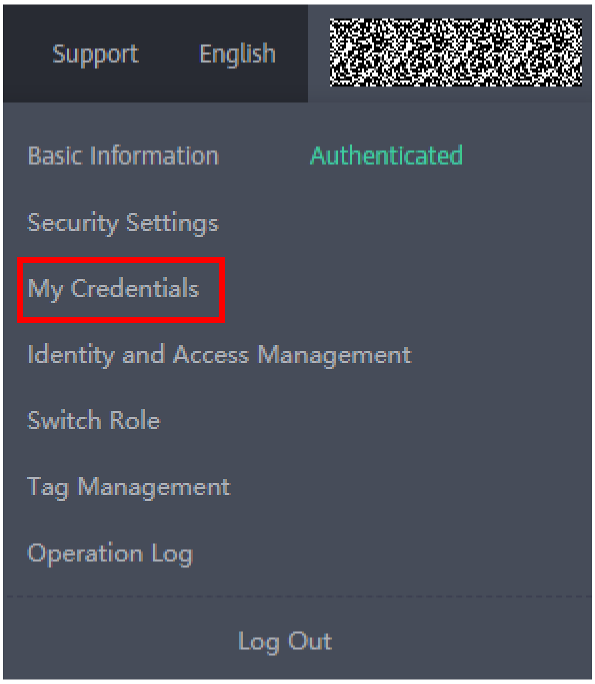
1. Use your IAM account to log in to **HUAWEI CLOUD**.
2. In the upper right corner of the page, click **Console**. The HUAWEI CLOUD management console is displayed.

Figure 1-13 Console



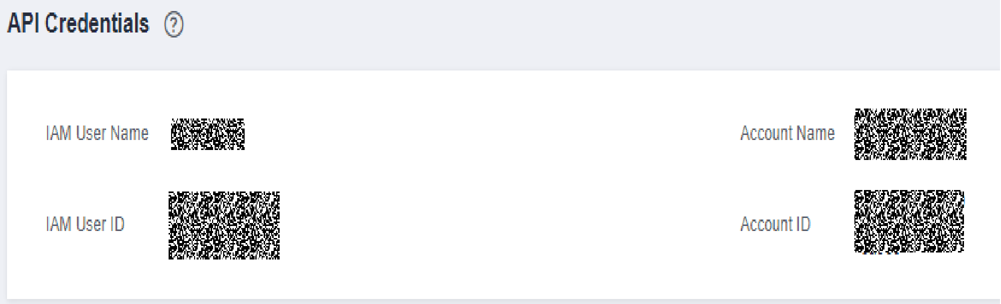
3. Hover the cursor over the account name in the upper right corner of the console and choose **My Credentials** from the drop-down list. The **API Credentials** page is displayed.

Figure 1-14 My Credentials



4. On the **API Credentials** page, obtain the IAM username, user ID, account name, and account ID.

Figure 1-15 Obtaining the credentials



API Credentials

IAM User Name [redacted]

IAM User ID [redacted]

Account Name [redacted]

Account ID [redacted]

1.16 Can AI-assisted Identification of ModelArts Identify a Specific Label?

After a model with multiple labels is trained and deployed as a real-time service, all the labels are identified. If only one type of label needs to be identified, train a model dedicated for identifying the label. To speed up the label identification, select a high flavor for deploying the model.

1.17 How Does ModelArts Use Tags to Manage Resources by Group?

ModelArts can work with Tag Management Service (TMS). When creating resource-consuming tasks in ModelArts, for example, training jobs, configure tags for these tasks so that ModelArts can use tags to manage resources by group.

ModelArts allows you to configure tags when you create training jobs, notebook instances, or real-time inference services.

Operation Process

1. Step 1 Create Predefined Tags on TMS
2. Step 2 Add a Tag to a ModelArts Task
3. Step 3 Obtain ModelArts Resource Usage by Resource Type in TMS

Step 1 Create Predefined Tags on TMS

Log in to the TMS console and create tags on the **Predefined Tags** page. The created tags are global and can be used in all Huawei Cloud regions.

Step 2 Add a Tag to a ModelArts Task

When creating a notebook instance, training job, or real-time inference services in ModelArts, configure a tag for the task.

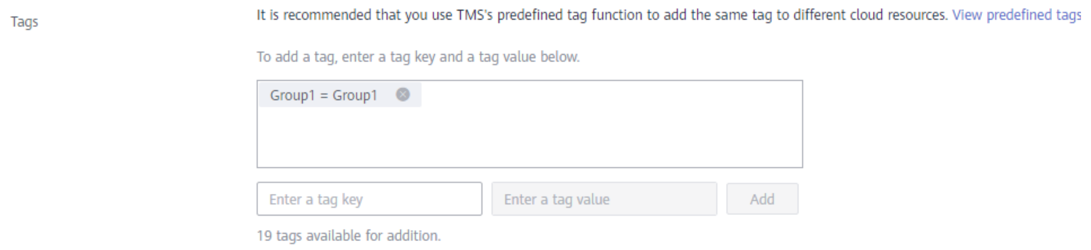
- Add a tag to a ModelArts notebook instance.

  Add a tag when you create a notebook instance. Alternatively, after creating a notebook instance, add a tag on the **Tags** tab on the instance details page.
- Add a tag to a ModelArts training job.

  Add a tag when you create a training job. Alternatively, after creating a training job, add a tag on the **Tags** tab on the job details page.
- Add a tag to a ModelArts real-time service.

  Add a tag when you create a real-time service. Alternatively, after creating a real-time service, add a tag on the **Tags** tab on the service details page.

Figure 1-16 Adding a tag



Tags

It is recommended that you use TMS's predefined tag function to add the same tag to different cloud resources. View predefined tags

To add a tag, enter a tag key and a tag value below.

Group1 = Group1

19 tags available for addition.

NOTE

When adding a tag to a ModelArts task, you can create new tags by specifying the keys and values of the new tags. The tags created here are available only to the current project.

Step 3 Obtain ModelArts Resource Usage by Resource Type in TMS

Log in to the TMS console. On the **Resources Tag** page, view resource tasks in specified regions based on resource types and tags.

- **Region**: one or more Huawei Cloud regions. For details, see **What Are Regions and AZs?**.
- **Resource Type**: **Table 1-6** lists the resource types that can be viewed on ModelArts.
- **Resource Tag**: If no tag is specified, all resources are displayed, regardless of whether the resources are configured with tags. One or multiple tags can be selected to obtain resource usage.

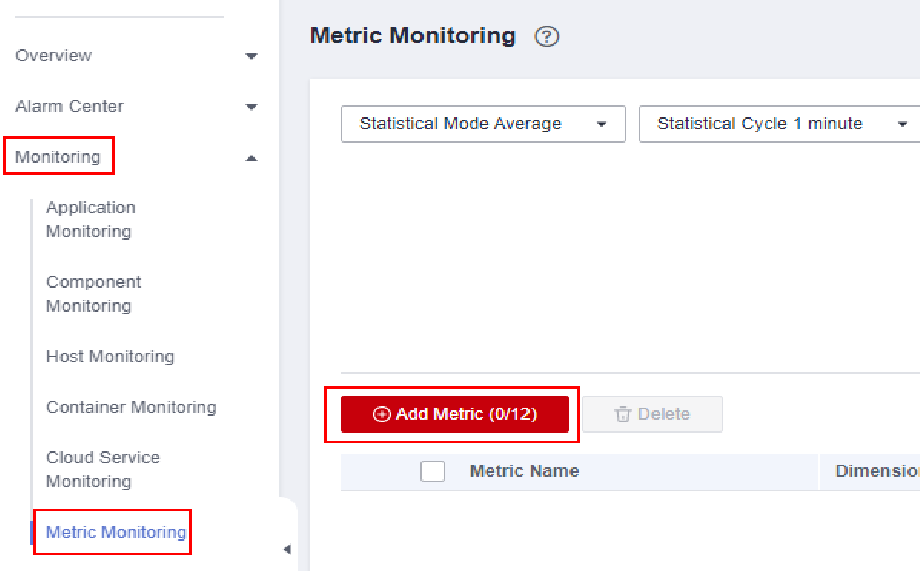
Table 1-6 Resource types that can be viewed on ModelArts

Resource Type	Description
ModelArts-Notebook	Notebook instances in ModelArts DevEnviron
ModelArts-TrainingJob	ModelArts training jobs
ModelArts-RealtimeService	ModelArts real-time inference services
ModelArts-ResourcePool	ModelArts dedicated resource pools

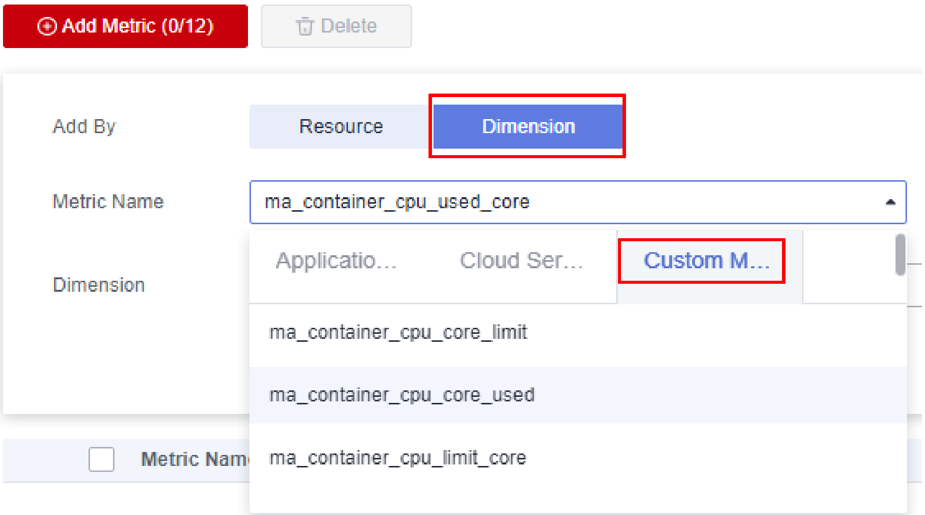
1.18 How Do I View All ModelArts Monitoring Metrics in AOM?

ModelArts periodically collects the usage of key metrics (such as GPUs, NPUs, CPUs, and memory) of each node in a resource pool as well as the usage of key metrics of the development environment, training jobs, and inference services, and reports the data to AOM. You can view the information on AOM.

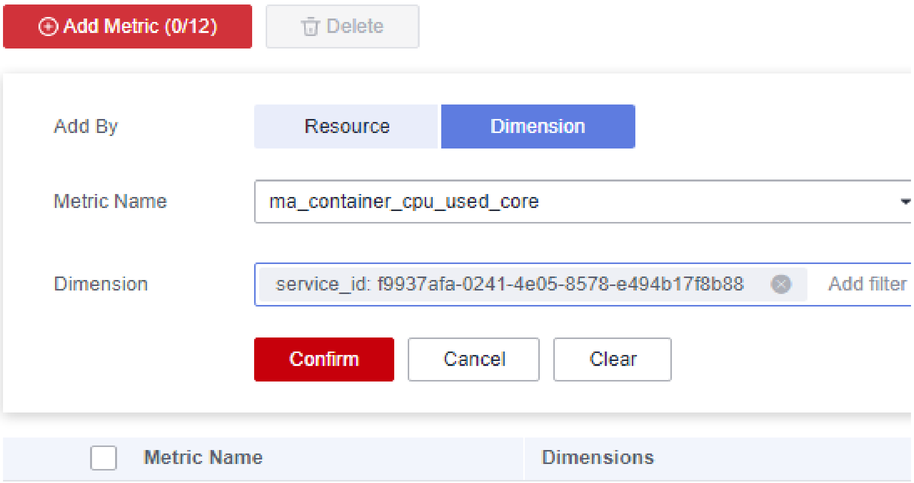
1. Log in to the console and search for **AOM** to go to the AOM console.
2. Choose **Monitoring** > **Metric Monitoring**. On the **Metric Monitoring** page that is displayed, click **Add Metric**.



3. Add metrics and click **Confirm**.



- **Add By**: Select **Dimension**.
- **Metric Name**: Click **Custom Metrics** and select the desired ones for query. For details, see **Table 1-7**, **Table 1-8**, and **Table 1-9**.
- **Dimension**: Enter the tag for filtering the metric. For details, see **Table 1-10**. The following shows an example.



Add Metric (0/12) | Delete

Add By: Resource | **Dimension**

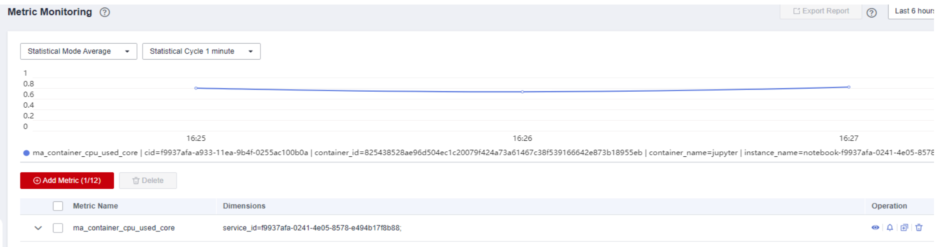
Metric Name: ma_container_cpu_used_core

Dimension: service_id: f9937afa-0241-4e05-8578-e494b17f8b88 | Add filter

Confirm | Cancel | Clear

□	Metric Name	Dimensions

4. View the metrics.



Statistical Mode Average | Statistical Cycle 1 minute

● ma_container_cpu_used_core | cid=f9937afa-a933-11ea-9b4f-0255ac100b0a | container_id=825438528ae96d504ec1c20079f424a73a61467c38f539166642e873b18955eb | container_name=jupyter | instance_name=notebook-f9937afa-0241-4e05-8578

Add Metric (1/12) | Delete

□	Metric Name	Dimensions	Operation
□	ma_container_cpu_used_core	service_id=f9937afa-0241-4e05-8578-e494b17f8b88;	

Table 1-7 Container metrics

Classification	Name	Metric	Description	Unit	Value Range
CPU	CPU Usage	ma_container_cpu_util	CPU usage of a measured object	%	0%–100%
	Used CPU Cores	ma_container_cpu_used_core	Number of CPU cores used by a measured object	Cores	≥ 0
	Total CPU Cores	ma_container_cpu_limit_core	Total number of CPU cores that have been applied for a measured object	Cores	≥ 1

Classification	Name	Metric	Description	Unit	Value Range
Memory	Total Physical Memory	ma_container_memory_capacity_megabytes	Total physical memory that has been applied for a measured object	MB	≥ 0
	Physical Memory Usage	ma_container_memory_util	Percentage of the used physical memory to the total physical memory	%	0%–100%
	Used Physical Memory	ma_container_memory_used_megabytes	Physical memory that has been used by a measured object (**container_memory_working_set_bytes** in the current working set) (Memory usage in a working set = Active anonymous page and cache, and file-baked page ≤ **container_memory_usage_bytes**)	MB	≥ 0

Classification	Name	Metric	Description	Unit	Value Range
Storage	Disk Read Rate	ma_container_disk_read_kilobytes	Volume of data read from a disk per second	KB/s	≥ 0
	Disk Write Rate	ma_container_disk_write_kilobytes	Volume of data written into a disk per second	KB/s	≥ 0
GPU memory	Total GPU Memory	ma_container_gpu_mem_total_megabytes	Total GPU memory of a training job	MB	> 0
	GPU Memory Usage	ma_container_gpu_mem_util	Percentage of the used GPU memory to the total GPU memory	%	0%–100%
	Used GPU Memory	ma_container_gpu_mem_used_megabytes	GPU memory used by a measured object	MB	≥ 0
GPU	GPU Usage	ma_container_gpu_util	GPU usage of a measured object	%	0%–100%

Classification	Name	Metric	Description	Unit	Value Range
	GPU Memory Bandwidth Usage	ma_container_gpu_mem_copy_util	GPU memory bandwidth usage of a measured object For example, the maximum memory bandwidth of NVIDIA GPU V100 is 900 GB/s. If the current memory bandwidth is 450 GB/s, the memory bandwidth usage is 50%.	%	0%–100%
	GPU Encoder Usage	ma_container_gpu_enc_util	GPU encoder usage of a measured object	%	%
	GPU Decoder Usage	ma_container_gpu_dec_util	GPU decoder usage of a measured object	%	%
	GPU Temperature	DCGM_FI_DEV_GPU_TEMP	GPU temperature	°C	Natural number
	GPU Power	DCGM_FI_DEV_POWER_USAGE	GPU power	Watt (W)	> 0
	GPU Memory Temperature	DCGM_FI_DEV_MEMORY_TEMP	GPU memory temperature	°C	Natural number

Classification	Name	Metric	Description	Unit	Value Range
Network I/O	Downlink Rate (BPS)	ma_container_network_receive_bytes	Inbound traffic rate of a measured object	Bytes/s	≥ 0
	Downlink Rate (PPS)	ma_container_network_receive_packets	Number of data packets received by an NIC per second	Packets/s	≥ 0
	Downlink Error Rate	ma_container_network_receive_error_packets	Number of error packets received by an NIC per second	Packets/s	≥ 0
	Uplink Rate (BPS)	ma_container_network_transmit_bytes	Outbound traffic rate of a measured object	Bytes/s	≥ 0
	Uplink Error Rate	ma_container_network_transmit_error_packets	Number of error packets sent by an NIC per second	Packets/s	≥ 0
	Uplink Rate (PPS)	ma_container_network_transmit_packets	Number of data packets sent by an NIC per second	Packets/s	≥ 0

Classification	Name	Metric	Description	Unit	Value Range
Notebook service metrics	Notebook Cache Directory Size	ma_container_notebook_cache_dir_size_bytes	A high-speed local disk is attached to the **/cache** directory for GPU notebook instances. This metric indicates the total size of the directory.	Bytes	≥ 0
	Notebook Cache Directory Utilization	ma_container_notebook_cache_dir_util	A high-speed local disk is attached to the **/cache** directory for GPU notebook instances. This metric indicates the utilization of the directory.	%	0%–100%

Table 1-8 Node metrics (collected only in dedicated resource pools)

Classification	Name	Metric	Description	Unit	Value Range
CPU	Total CPU Cores	ma_node_cpu_limit_core	Total number of CPU cores that have been applied for a measured object	Cores	≥ 1

Classification	Name	Metric	Description	Unit	Value Range
	Used CPU Cores	ma_node_cpu_used_core	Number of CPU cores used by a measured object	Cores	≥ 0
	CPU Usage	ma_node_cpu_util	CPU usage of a measured object	%	0%–100%
	CPU I/O Wait Time	ma_node_cpu_iowait_counter	Disk I/O wait time accumulated since system startup	jiffies	≥ 0
Memory	Physical Memory Usage	ma_node_memory_util	Percentage of the used physical memory to the total physical memory	%	0%–100%
	Total Physical Memory	ma_node_memory_total_megabytes	Total physical memory that has been applied for a measured object	MB	≥ 0
Network I/O	Downlink Rate (BPS)	ma_node_network_receive_rate_bytes_seconds	Inbound traffic rate of a measured object	Bytes/s	≥ 0
	Uplink Rate (BPS)	ma_node_network_transmit_rate_bytes_seconds	Outbound traffic rate of a measured object	Bytes/s	≥ 0

Classification	Name	Metric	Description	Unit	Value Range
Storage	Disk Read Rate	ma_node_disk_read_rate_kilobytes_seconds	Volume of data read from a disk per second (Only data disks used by containers are collected.)	KB/s	≥ 0
	Disk Write Rate	ma_node_disk_write_rate_kilobytes_seconds	Volume of data written into a disk per second (Only data disks used by containers are collected.)	KB/s	≥ 0
	Total Cache	ma_node_cache_space_capacity_megabytes	Total cache of the Kubernetes space	MB	≥ 0
	Used Cache	ma_node_cache_space_used_capacity_megabytes	Used cache of the Kubernetes space	MB	≥ 0
	Total Container Space	ma_node_container_space_capacity_megabytes	Total container space	MB	≥ 0
	Used Container Space	ma_node_container_space_used_capacity_megabytes	Used container space	MB	≥ 0
	Disk Information	ma_node_disk_info	Basic disk information	N/A	≥ 0

Classification	Name	Metric	Description	Unit	Value Range
	Total Reads	ma_node_disk_reads_completed_total	Total number of successful reads	N/A	≥ 0
	Merged Reads	ma_node_disk_reads_merged_total	Number of merged reads	N/A	≥ 0
	Bytes Read	ma_node_disk_read_bytes_total	Total number of bytes that are successfully read	Bytes	≥ 0
	Read Time Spent	ma_node_disk_read_time_seconds_total	Time spent on all reads	Seconds	≥ 0
	Total Writes	ma_node_disk_writes_completed_total	Total number of successful writes	N/A	≥ 0
	Merged Writes	ma_node_disk_writes_merged_total	Number of merged writes	N/A	≥ 0
	Written Bytes	ma_node_disk_written_bytes_total	Total number of bytes that are successfully written	Bytes	≥ 0
	Write Time Spent	ma_node_disk_write_time_seconds_total	Time spent on all write operations	Seconds	≥ 0
	Ongoing I/Os	ma_node_disk_io_now	Number of ongoing I/Os	N/A	≥ 0
	I/O Execution Duration	ma_node_disk_io_time_seconds_total	Time spent on executing I/Os	Seconds	≥ 0

Classification	Name	Metric	Description	Unit	Value Range
	I/O Execution Weighted Time	ma_node_disk_io_time_weighted_seconds_tota	Weighted time spent on executing I/Os	Seconds	≥ 0
GPU	GPU Usage	ma_node_gpu_util	GPU usage of a measured object	%	0%–100%
	Total GPU Memory	ma_node_gpu_mem_total_megabytes	Total GPU memory of a measured object	MB	> 0
	GPU Memory Usage	ma_node_gpu_mem_util	Percentage of the used GPU memory to the total GPU memory	%	0%–100%
	Used GPU Memory	ma_node_gpu_mem_used_megabytes	GPU memory used by a measured object	MB	≥ 0
	Tasks on a Shared GPU	node_gpu_share_job_count	Number of tasks running on a shared GPU	Number	≥ 0
	GPU Temperature	DCGM_FI_DEV_GPU_TEMP	GPU temperature	°C	Natural number
	GPU Power	DCGM_FI_DEV_POWER_USAGE	GPU power	Watt (W)	> 0
	GPU Memory Temperature	DCGM_FI_DEV_MEMORY_TEMP	GPU memory temperature	°C	Natural number

Classification	Name	Metric	Description	Unit	Value Range
InfiniBand or RoCE network	Total Amount of Data Received by an NIC	ma_node_infiniband_port_received_data_bytes_total	The total number of data octets, divided by 4, (counting in double words, 32 bits), received on all VLs from the port.	Double words (32 bits)	≥ 0
	Total Amount of Data Sent by an NIC	ma_node_infiniband_port_transmitted_data_bytes_total	The total number of data octets, divided by 4, (counting in double words, 32 bits), transmitted on all VLs from the port.	Double words (32 bits)	≥ 0

Classification	Name	Metric	Description	Unit	Value Range
NFS mounting status	NFS Getattr Congestion Time	ma_node_mountstats_getattr_backlog_wait	Getattr is an NFS operation that retrieves the attributes of a file or directory, such as size, permissions, owner, etc. Backlog wait is the time that the NFS requests have to wait in the backlog queue before being sent to the NFS server. It indicates the congestion on the NFS client side. A high backlog wait can cause poor NFS performance and slow system response times.	ms	≥ 0

Classification	Name	Metric	Description	Unit	Value Range
	NFS Getattr Round Trip Time	ma_node_mountstats_getattr_rtt	Getattr is an NFS operation that retrieves the attributes of a file or directory, such as size, permissions, owner, etc. RTT stands for Round Trip Time and it is the time from when the kernel RPC client sends the RPC request to the time it receives the reply34. RTT includes network transit time and server execution time. RTT is a good measurement for NFS latency. A high RTT can indicate network or server issues.	ms	≥ 0

Classification	Name	Metric	Description	Unit	Value Range
	NFS Access Congestion Time	ma_node_mountstats_access_backlog_wait	Access is an NFS operation that checks the access permissions of a file or directory for a given user. Backlog wait is the time that the NFS requests have to wait in the backlog queue before being sent to the NFS server. It indicates the congestion on the NFS client side. A high backlog wait can cause poor NFS performance and slow system response times.	ms	≥ 0

Classification	Name	Metric	Description	Unit	Value Range
	NFS Access Round Trip Time	ma_node_mountstats_access_rtt	Access is an NFS operation that checks the access permissions of a file or directory for a given user. RTT stands for Round Trip Time and it is the time from when the kernel RPC client sends the RPC request to the time it receives the reply34. RTT includes network transit time and server execution time. RTT is a good measurement for NFS latency. A high RTT can indicate network or server issues.	ms	≥ 0

Classification	Name	Metric	Description	Unit	Value Range
	NFS Lookup Congestion Time	ma_node_mountstats_lookup_backlog_wait	Lookup is an NFS operation that resolves a file name in a directory to a file handle. Backlog wait is the time that the NFS requests have to wait in the backlog queue before being sent to the NFS server. It indicates the congestion on the NFS client side. A high backlog wait can cause poor NFS performance and slow system response times.	ms	≥ 0

Classification	Name	Metric	Description	Unit	Value Range
	NFS Lookup Round Trip Time	ma_node_mountstats_lookup_rtt	Lookup is an NFS operation that resolves a file name in a directory to a file handle. RTT stands for Round Trip Time and it is the time from when the kernel RPC client sends the RPC request to the time it receives the reply34. RTT includes network transit time and server execution time. RTT is a good measurement for NFS latency. A high RTT can indicate network or server issues.	ms	≥ 0

Classification	Name	Metric	Description	Unit	Value Range
	NFS Read Congestion Time	ma_node_mountstats_read_backlog_wait	Read is an NFS operation that reads data from a file. Backlog wait is the time that the NFS requests have to wait in the backlog queue before being sent to the NFS server. It indicates the congestion on the NFS client side. A high backlog wait can cause poor NFS performance and slow system response times.	ms	≥ 0

Classification	Name	Metric	Description	Unit	Value Range
	NFS Read Round Trip Time	ma_node_mountstats_read_rtt	Read is an NFS operation that reads data from a file. RTT stands for Round Trip Time and it is the time from when the kernel RPC client sends the RPC request to the time it receives the reply34. RTT includes network transit time and server execution time. RTT is a good measurement for NFS latency. A high RTT can indicate network or server issues.	ms	≥ 0

Classification	Name	Metric	Description	Unit	Value Range
	NFS Write Congestion Time	ma_node_mountstats_write_backlog_wait	Write is an NFS operation that writes data to a file. Backlog wait is the time that the NFS requests have to wait in the backlog queue before being sent to the NFS server. It indicates the congestion on the NFS client side. A high backlog wait can cause poor NFS performance and slow system response times.	ms	≥ 0

Classification	Name	Metric	Description	Unit	Value Range
	NFS Write Round Trip Time	ma_node_mountstats_write_rtt	Write is an NFS operation that writes data to a file. RTT stands for Round Trip Time and it is the time from when the kernel RPC client sends the RPC request to the time it receives the reply34. RTT includes network transit time and server execution time. RTT is a good measurement for NFS latency. A high RTT can indicate network or server issues.	ms	≥ 0

Table 1-9 Diagnosis (IB, collected only in dedicated resource pools)

Classification	Name	Metric	Description	Unit	Value Range
InfiniBand or RoCE network	PortXmitData	infiniband_port_xmit_data_total	The total number of data octets, divided by 4, (counting in double words, 32 bits), transmitted on all VLs from the port.	Total count	Natural number
	PortRcvData	infiniband_port_rcv_data_total	The total number of data octets, divided by 4, (counting in double words, 32 bits), received on all VLs from the port.	Total count	Natural number
	SymbolErrorCounter	infiniband_symbol_error_counter_total	Total number of minor link errors detected on one or more physical lanes.	Total count	Natural number
	LinkErrorRecoveryCounter	infiniband_link_error_recovery_counter_total	Total number of times the Port Training state machine has successfully completed the link error recovery process.	Total count	Natural number
	PortRcvErrors	infiniband_port_rcv_errors_total	Total number of packets containing errors that were received on the port including: Local physical errors (ICRC, VCRC, LPCRC, and all physical errors that cause entry into the BAD PACKET or BAD PACKET DISCARD states of the packet receiver state machine) Malformed data packet errors (LVer, length, VL) Malformed link packet errors (operand, length, VL) Packets discarded due to buffer overrun (overflow)	Total count	Natural number

Classification	Name	Metric	Description	Unit	Value Range
	LocalLinkIntegrityErrors	infiniband_local_link_integrity_errors_total	This counter indicates the number of retries initiated by a link transfer layer receiver.	Total count	Natural number
	PortRcvRemotePhysicalErrors	infiniband_port_rcv_remote_physical_errors_total	Total number of packets marked with the EBP delimiter received on the port.	Total count	Natural number
	PortRcvSwitchRelayErrors	infiniband_port_rcv_switch_relay_errors_total	Total number of packets received on the port that were discarded when they could not be forwarded by the switch relay for the following reasons: DLID mapping VL mapping Looping (output port = input port)	Total count	Natural number
	PortXmitWait	infiniband_port_transmit_wait_total	The number of ticks during which the port had data to transmit but no data was sent during the entire tick (either because of insufficient credits or because of lack of arbitration).	Total count	Natural number
	PortXmitDiscards	infiniband_port_xmit_discards_total	Total number of outbound packets discarded by the port because the port is down or congested.	Total count	Natural number

For details about the metrics of an InfiniBand or RoCE network, see **NVIDIA Mellanox documents**.

Table 1-10 Metric names

Classification	Metric	Description
Container metrics	modelarts_service	Service to which a container belongs, which can be **notebook**, **train**, or **infer**

Classification	Metric	Description
	instance_name	Name of the pod to which the container belongs
	service_id	Instance or job ID displayed on the page, for example, **cf55829e-9bd3-48fa-8071-7ae870dae93a** for a development environment **9f322d5a-b1d2-4370-94df-5a87de27d36e** for a training job
	node_ip	IP address of the node to which the container belongs
	container_id	Container ID
	cid	Cluster ID
	container_name	Name of the container
	project_id	Project ID of the account to which the user belongs
	user_id	User ID of the account to which the user who submits the job belongs
	npu_id	Ascend card ID, for example, **davinci0** (to be discarded)
	device_id	Physical ID of Ascend AI processors
	device_type	Type of Ascend AI processors
	pool_id	ID of a resource pool corresponding to a physical dedicated resource pool
	pool_name	Name of a resource pool corresponding to a physical dedicated resource pool
	logical_pool_id	ID of a logical subpool
	logical_pool_name	Name of a logical subpool
	gpu_uuid	UUID of the GPU used by the container
	gpu_index	Index of the GPU used by the container
	gpu_type	Type of the GPU used by the container
	account_name	Account name of the creator of a training, inference, or development environment task
	user_name	Username of the creator of a training, inference, or development environment task

Classification	Metric	Description
	task_creation_time	Time when a training, inference, or development environment task is created
	task_name	Name of a training, inference, or development environment task
	task_spec_code	Specifications of a training, inference, or development environment task
	cluster_name	CCE cluster name
Node metrics	cid	ID of the CCE cluster to which the node belongs
	node_ip	IP address of the node
	host_name	Hostname of a node
	pool_id	ID of a resource pool corresponding to a physical dedicated resource pool
	project_id	Project ID of the user in a physical dedicated resource pool
	npu_id	Ascend card ID, for example, **davinci0** (to be discarded)
	device_id	Physical ID of Ascend AI processors
	device_type	Type of Ascend AI processors
	gpu_uuid	UUID of a node GPU
	gpu_index	Index of a node GPU
	gpu_type	Type of a node GPU
	device_name	Device name of an InfiniBand or RoCE network NIC
	port	Port number of the IB NIC
	physical_state	Status of each port on the IB NIC
	firmware_version	Firmware version of the IB NIC
	filesystem	NFS-mounted file system
	mount_point	NFS mount point
Diagnos	cid	ID of the CCE cluster to which the node where the GPU resides belongs
	node_ip	IP address of the node where the GPU resides

Classification	Metric	Description
	pool_id	ID of a resource pool corresponding to a physical dedicated resource pool
	project_id	Project ID of the user in a physical dedicated resource pool
	gpu_uuid	GPU UUID
	gpu_index	Index of a node GPU
	gpu_type	Type of a node GPU
	device_name	Name of a network device or disk device
	port	Port number of the IB NIC
	physical_state	Status of each port on the IB NIC
	firmware_version	Firmware version of the IB NIC

Table 1-11 Metric names

Classification	Metric	Description
Container metrics	modelarts_service	Service to which a container belongs, which can be **notebook**, **train**, or **infer**
	instance_name	Name of the pod to which the container belongs
	service_id	Instance or job ID displayed on the page, for example, **cf55829e-9bd3-48fa-8071-7ae870dae93a** for a development environment **9f322d5a-b1d2-4370-94df-5a87de27d36e** for a training job
	node_ip	IP address of the node to which the container belongs
	container_id	Container ID
	cid	Cluster ID
	container_name	Name of the container
	project_id	Project ID of the account to which the user belongs
	user_id	User ID of the account to which the user who submits the job belongs

Classification	Metric	Description
	pool_id	ID of a resource pool corresponding to a physical dedicated resource pool
	pool_name	Name of a resource pool corresponding to a physical dedicated resource pool
	logical_pool_id	ID of a logical subpool
	logical_pool_name	Name of a logical subpool
	gpu_uuid	UUID of the GPU used by the container
	gpu_index	Index of the GPU used by the container
	gpu_type	Type of the GPU used by the container
	account_name	Account name of the creator of a training, inference, or development environment task
	user_name	Username of the creator of a training, inference, or development environment task
	task_creation_time	Time when a training, inference, or development environment task is created
	task_name	Name of a training, inference, or development environment task
	task_spec_code	Specifications of a training, inference, or development environment task
	cluster_name	CCE cluster name
Node metrics	cid	ID of the CCE cluster to which the node belongs
	node_ip	IP address of the node
	host_name	Hostname of a node
	pool_id	ID of a resource pool corresponding to a physical dedicated resource pool
	project_id	Project ID of the user in a physical dedicated resource pool
	gpu_uuid	UUID of a node GPU
	gpu_index	Index of a node GPU
	gpu_type	Type of a node GPU
	device_name	Device name of an InfiniBand or RoCE network NIC

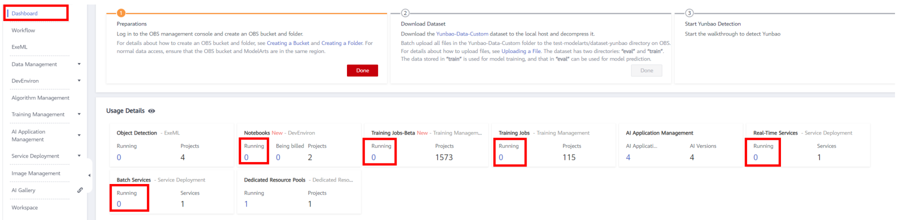
Classification	Metric	Description
	port	Port number of the IB NIC
	physical_state	Status of each port on the IB NIC
	firmware_version	Firmware version of the IB NIC
	filesystem	NFS-mounted file system
	mount_point	NFS mount point
Diagnos	cid	ID of the CCE cluster to which the node where the GPU resides belongs
	node_ip	IP address of the node where the GPU resides
	pool_id	ID of a resource pool corresponding to a physical dedicated resource pool
	project_id	Project ID of the user in a physical dedicated resource pool
	gpu_uuid	GPU UUID
	gpu_index	Index of a node GPU
	gpu_type	Type of a node GPU
	device_name	Name of a network device or disk device
	port	Port number of the IB NIC
	physical_state	Status of each port on the IB NIC
	firmware_version	Firmware version of the IB NIC

1.19 Why Is the Job Still Queued When Resources Are Sufficient?

- If a public resource pool is used, the resources may be used by other users. Please wait or find solutions in **Why Is a Training Job Always Queuing?**.
- If a dedicated resource pool is used, perform the following operations:
  a. Check whether other jobs (including inference jobs, training jobs, and development environment jobs) are running in the dedicated resource pool.

     On the **Dashboard** page, you can go to the details page of the running jobs or instances to check whether the dedicated resource pool is used. You can stop them based on your needs to release resources.

Figure 1-17 Dashboard



1 Preparations
Log in to the OBS management console and create an OBS bucket and folder.
For details about how to create an OBS bucket and folder, see Creating a Bucket and Creating a Folder. For normal data access, ensure that the OBS bucket and ModelArts are in the same region.
Done

2 Download Dataset
Download the Yunbao-Data-Custom dataset to the local host and decompress it.
Batch upload all files in the Yunbao-Data-Custom folder to the test-modelarts/dataset-yunbao directory on OBS.
For details about how to upload files, see Uploading a File. The dataset has two directories: "eval" and "train".
The data stored in "train" is used for model training, and that in "eval" can be used for model prediction.
Done

3 Start Yunbao Detection
Start the walkthrough to detect Yunbao

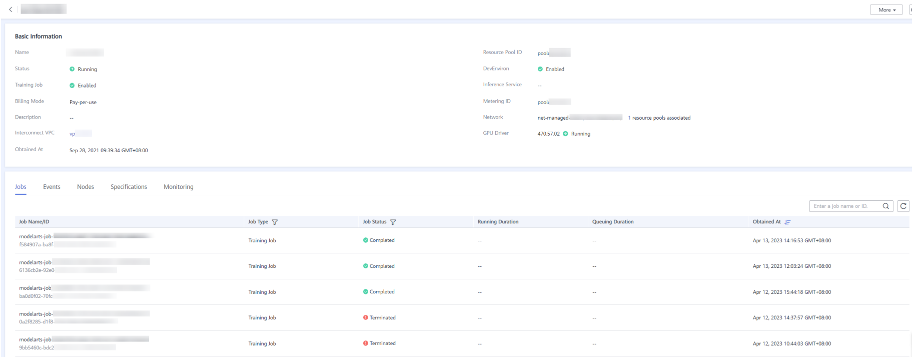
Usage Details

Object Detection - ExeML: Running 0, Projects 4
Notebooks New - DevEnviron: Running 0, Being billed 0, Projects 2
Training Jobs-Beta New - Training Managem...: Running 0, Projects 1573
Training Jobs - Training Management: Running 0, Projects 115
AI Application Management: AI Applicati... 4, AI Versions 4
Real-Time Services - Service Deployment: Running 0, Services 1
Batch Services - Service Deployment: Running 0, Services 1
Dedicated Resource Pools - Dedicated Reso...: Running 1, Projects 1

b. Go to the details page of the dedicated resource pool to check whether there are other queuing jobs.

If yes, the new job needs to be queued.

Figure 1-18 Queuing jobs



Basic Information

Name		Resource Pool ID	pool
Status	Running	DevEnviron	Enabled
Training Job	Enabled	Inference Service	--
Billing Mode	Pay-per-use	Metering ID	pool
Description	--	Network	net-managed- 1 resource pools associated
Interconnect VPC	vp	GPU Driver	470.57.02 Running
Obtained At	Sep 28, 2021 09:39:34 GMT+08:00		

Jobs Events Nodes Specifications Monitoring

Enter a job name or ID.

Job Name/ID	Job Type	Job Status	Running Duration	Queuing Duration	Obtained At
modelarts-job- f584907a-ba8f-	Training Job	Completed	--	--	Apr 13, 2023 14:16:53 GMT+08:00
modelarts-job- 6136cb2e-92e0	Training Job	Completed	--	--	Apr 13, 2023 12:03:24 GMT+08:00
modelarts-job ba0d0f02-70fc	Training Job	Completed	--	--	Apr 12, 2023 15:44:18 GMT+08:00
modelarts-job- 0a2f8285-d1f8-	Training Job	Terminated	--	--	Apr 12, 2023 14:37:57 GMT+08:00
modelarts-job 9bb5460c-bdc2	Training Job	Terminated	--	--	Apr 12, 2023 10:44:03 GMT+08:00

c. Check whether resources are fragmented.

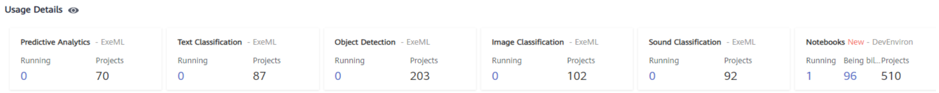
For example, the cluster has two nodes, and there are four idle cards on each node. However, your job requires eight cards on one node. In this case, the idle resources cannot be allocated to your job.

2 Billing

2.1 How Do I View the ModelArts Jobs Being Billed?

Log in to the ModelArts management console. In the navigation pane on the left, click **Dashboard** and view the jobs that are being billed. Go to the management page and stop them based on site requirements. For example, if a notebook instance is being billed, choose **DevEnviron** > **Notebook**, and stop the running notebook instance.

Figure 2-1 Viewing jobs that are being billed



Usage Details

Predictive Analytics - ExeML
Running 0
Projects 70

Text Classification - ExeML
Running 0
Projects 87

Object Detection - ExeML
Running 0
Projects 203

Image Classification - ExeML
Running 0
Projects 102

Sound Classification - ExeML
Running 0
Projects 92

Notebooks New - DevEnviron
Running 1
Being bil... 96
Projects 510

The following items will be billed when you use ModelArts:

- Workflow: A workflow is billed when it is running. After using the workflow, stop it and the training jobs and services created for running the workflow. Additionally, clear the data stored in OBS.
- ExeML: ExeML is billed when it is running. After using ExeML, stop it and the training jobs and services created for running ExeML. Additionally, clear the data stored in OBS.
- Notebook instances:
  - Running notebook instances are billed. To stop billing a notebook instance, stop or delete it. If EVS is used for storage, clear the data stored in EVS.
  - You will be billed for a paid flavor when you try CodeLab. To stop billing the flavor, stop the notebook instance on the JupyterLab page.
- Training jobs: Running training job are billed. To stop billing a training job, stop it. Additionally, clear the data stored in OBS.
- Deployed services: If a model is deployed as a real-time service, the service will be billed. To stop the billing, stop the deployed service. Additionally, clear the data stored in OBS.

- Dedicated resource pools: If you have purchased a dedicated resource pool in ModelArts for AI development and use this resource pool for running ExeML jobs, workflows, notebook instances, training jobs, and deployed services, the compute resources used in these operations will be billed via the dedicated resource pool. A pay-per-use dedicated resource pool is billed continuously since it has been created, so delete it if you do not use it.

⚠ CAUTION

In addition to the billing items displayed on the **Dashboard** page of ModelArts, OBS and EVS storage will be separately billed if they are used.

- To stop billing the OBS resources, go to the OBS management console and clear the data in OBS.
- To stop billing the EVS storage, go to the ModelArts management console and delete the notebook instances with EVS storage. To stop billing the EVS resources, go to the EVS management console and clear the data in EVS.

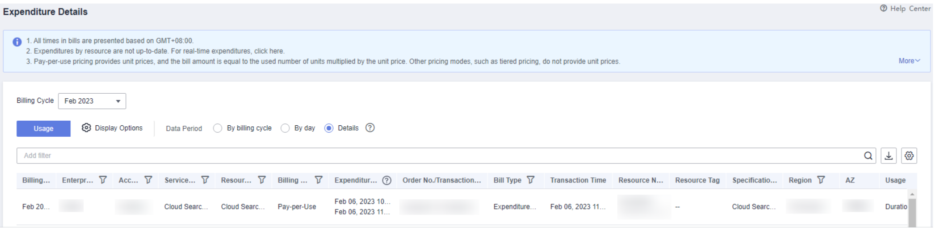
2.2 How Do I View ModelArts Expenditure Details?

In **Billing Center**, the expenditure details of ModelArts are displayed on an hourly basis. You can switch to the **Expenditure Details** page to view the fees consumed for each job.

Procedure:

1. In **Billing Center**, choose **Billing** > **Expenditure Items**. On the **Expenditure Items** page, click the order or transaction number of a record in the list to view its expenditure details.
2. On the **Expenditure Details** page, the resource usage and fees in the order are displayed. Select display options and data period to view expenditure details.

Figure 2-2 Expenditure details



Expenditure Details

1. All times in bills are presented based on GMT+08:00.
2. Expenditures by resource are not up-to-date. For real-time expenditures, click here.
3. Pay-per-use pricing provides unit prices, and the bill amount is equal to the used number of units multiplied by the unit price. Other pricing modes, such as tiered pricing, do not provide unit prices.

Billing Cycle: Feb 2023

Data Period: ○ By billing cycle ○ By day ● Details

Billing...	Enterpr...	Acc...	Service...	Resour...	Billing ...	Expenditur...	Order No./Transaction...	Bill Type	Transaction Time	Resource N...	Resource Tag	Specificatio...	Region	AZ	Usage
Feb 20...	[redacted]	[redacted]	Cloud Searc...	Cloud Searc...	Pay-per-Use	Feb 06, 2023 10... Feb 06, 2023 11...	[redacted]	Expenditure...	Feb 06, 2023 11...	[redacted]	--	Cloud Searc...	[redacted]	[redacted]	Duratio

2.3 Will I Be Charged for Uploading Datasets to ModelArts?

You are not charged for dataset management and labeling in ModelArts. However, datasets are stored in OBS, and you will be billed for the storage on OBS. For details, see **OBS pricing details** and create OBS buckets to store data used by ModelArts.

2.4 What Should I Do to Avoid Unnecessary Billing After I Label Datasets and Exit?

Labeling datasets is free of charge. However, OBS bills you for the storage space used for storing the datasets. You are advised to go to the OBS management console and delete the stored data and OBS buckets to stop billing.

2.5 How Do I Stop Billing for a ModelArts ExeML Project?

- For ExeML jobs created using a **public resource pool**:
  - Log in to the ModelArts management console. In the ExeML job list, delete the target ExeML job. In the training job list, stop the training jobs created for running the ExeML job. In the real-time service list, stop the services created for running the ExeML job. After these operations are complete, the billing is stopped.
  - Log in to the OBS management console, access your OBS bucket, and delete the data stored in the OBS bucket. After these operations are complete, the billing is stopped.
- For ExeML jobs created using a **dedicated resource pool**:
  - Log in to the ModelArts management console. In the ExeML job list, delete the target ExeML job. In the training job list, stop the training jobs created for running the ExeML job. In the real-time service list, stop the services created for running the ExeML job. In the resource pool list, delete the dedicated resource pool that is running the target ExeML job. After these operations are complete, the billing is stopped.
  - Log in to the OBS management console, access your OBS bucket, and delete the data stored in the OBS bucket. After these operations are complete, the billing is stopped.

2.6 How Do I Stop Billing If I Do Not Use ModelArts?

The cost of AI development in ModelArts mainly includes the storage fee and resource fee. If ModelArts is no longer used, stop or delete the services running in ModelArts and delete the data stored in OBS and EVS.

Clearing Storage Data

ModelArts data is stored in OBS. To stop storage billing, switch to the OBS console and delete the data and directories.

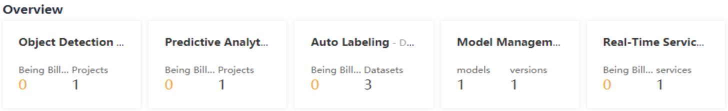
Clearing Resources

To stop resource billing, check running jobs in ModelArts, and stop or delete the jobs.

Procedure

Log in to the ModelArts console. In the navigation pane, choose **Dashboard**. On the **Dashboard** area, view the jobs that are being billed. Then, stop the jobs as needed.

Figure 2-3 Viewing jobs that are being billed



Overview

Object Detection ...
Being Bill... Projects
0 1

Predictive Analyt...
Being Bill... Projects
0 1

Auto Labeling - D...
Being Bill... Datasets
0 3

Model Managem...
models versions
1 1

Real-Time Servic...
Being Bill... services
0 1

- In the navigation pane of the ModelArts console, choose **Workflow** and check whether there are any running workflows. If so, click **Delete** in the **Operation** column. Then, the billing will be stopped accordingly.
- In the navigation pane of the ModelArts console, choose **ExeML** and check whether there are any running projects. If so, click **Delete** in the **Operation** column. Then, the billing will be stopped accordingly.
- In the navigation pane of the ModelArts console, choose **DevEnviron** > **Notebook** and check whether there are running notebook instances. If so, click **Stop** in the **Operation** column. Then, the billing will be stopped accordingly. Check whether there are notebook instances using EVS storage. If so, stop and delete the notebook instances. Then, the EVS billing will be stopped accordingly.
- In the navigation pane of the ModelArts console, choose **Training Management** > **Training Jobs**, and check whether there are running jobs. If so, click **Stop** in the **Operation** column. Then, the billing will be stopped accordingly.
- In the navigation pane of the ModelArts console, choose **Training Management** > **Training Jobs**, click the **Visualization Jobs** tab, and check whether there are running jobs. If so, click **Stop** in the **Operation** column. Then, the billing will be stopped accordingly.
- In the navigation pane of the ModelArts console, choose **Service Deployment** > **Real-Time Services** and check whether there are running jobs. If so, click **Stop** in the **Operation** column. Then, the billing will be stopped accordingly.
- In the navigation pane of the ModelArts console, choose **Service Deployment** > **Batch Services** and check whether there are running jobs. If so, click **Stop** in the **Operation** column. Then, the billing will be stopped accordingly.

2.7 How Are Training Jobs Billed?

- If you use a public resource pool, you are charged based on the selected flavor, number of nodes, and running duration. The billing rule is as follows: Unit price of the flavor x Number of nodes x Running duration (accurate to seconds).
- If you use dedicated resource pool, training jobs are not billed separately. You are charged for the dedicated resource pool.

2.8 Why Does Billing Continue After All Projects Are Deleted?

Even if ExeML projects, notebook instances, training jobs, or services of ModelArts are stopped, and there is no charging item is displayed on the **Dashboard** page, the account may still be billed for the OBS storage being used.

The possible causes are as follows:

1. You have uploaded data to OBS for storage when using ModelArts, and the OBS storage is billed. In this case, go to the OBS management console and delete the data, folders, and OBS buckets that are no longer needed.
2. You have selected EVS storage when creating a notebook instance, and the storage is separately billed even after the notebook instance is stopped. In this case, delete the notebook instance.
3. You have switched the flavor to a charging one when experiencing CodeLab.

   In this case, go to the CodeLab page and click ⏻ in the upper right corner to stop the notebook instance.

3 ExeML

3.1 Functional Consulting

3.1.1 What Is ExeML?

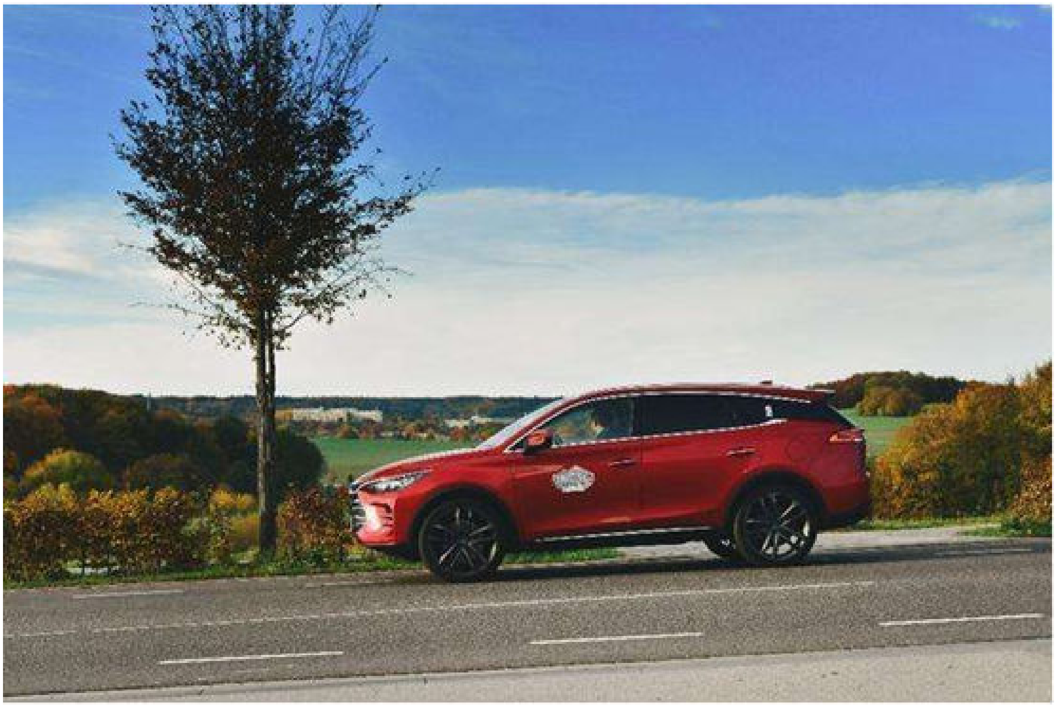
ExeML is the process of automating model design, parameter tuning, and model training, compression, and deployment with the labeled data. The process is free of coding and does not require developers' experience in model development.

Users who do not have encoding capability can use the labeling, one-click model training, and model deployment functions of ExeML to build AI models.

3.1.2 What Are Image Classification and Object Detection?

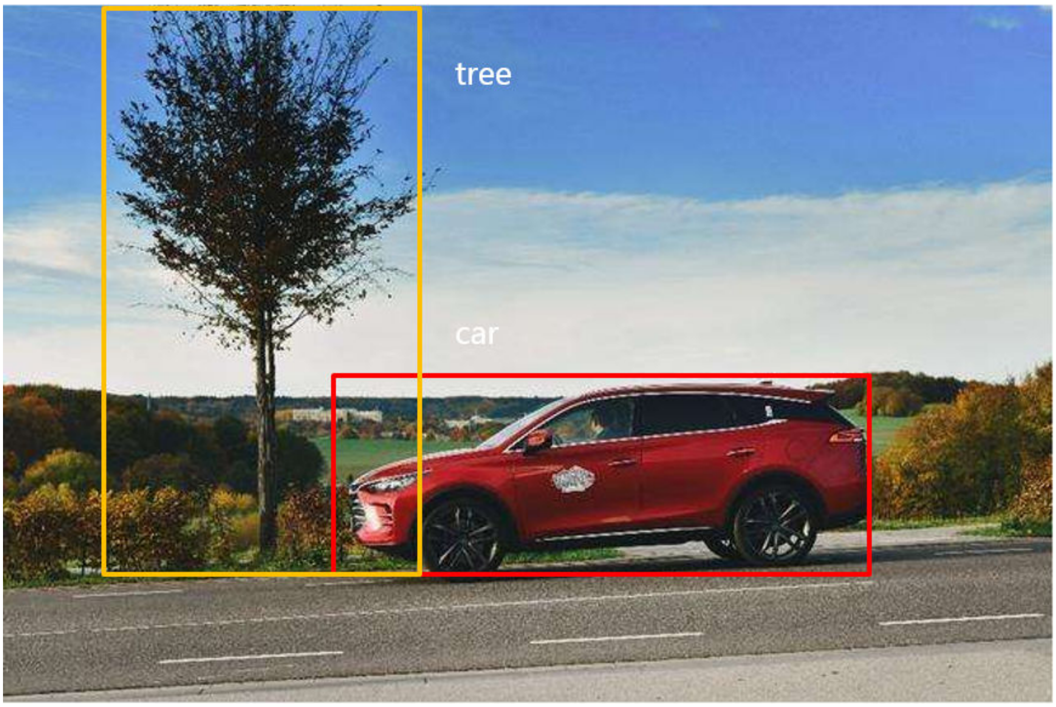
Image classification is an image processing method that separates different classes of targets according to the features reflected in the images. With quantitative analysis on images, it classifies an image or each pixel or area in an image into different categories to replace human visual interpretation. In general, image classification aims to identify a class, status, or scene in an image. It is applicable to scenarios where an image contains only one object. **Figure 3-1** shows an example of identifying a car in an image.

Figure 3-1 Image classification



Object detection is one of the classical problems in computer vision. It intends to label objects with frames and identify the object classes in an image. Generally, if an image contains multiple objects, object detection can identify the location, quantity, and name of each object in the image. It is suitable for scenarios where an image contains multiple objects. **Figure 3-2** shows an example of identifying a tree and a car in an image.

Figure 3-2 Object detection



3.1.3 What Are the Differences Between ExeML and Subscribed Algorithms?

ModelArts provides different AI development modes for new and experienced developers.

- For new developers, you can use ExeML to develop models without coding. When you use ExeML, the system automatically selects appropriate algorithms and parameters for model training.
- For experienced AI developers, you can select subscribed algorithms for model training. In addition, you can customize the parameters required for training.

3.2 Preparing Data

3.2.1 What Are the Requirements for Training Data When You Create a Predictive Analytics Project in ExeML?

Requirements on Datasets

- Dataset consists of letters, digits, hyphens (-), and underscores (_), and must be in CSV format. Data files cannot be stored in the root directory of an OBS bucket, but in a folder in the OBS bucket, for example, **/obs-xxx/data/input.csv**.
- Use newline characters (\n or LF) to separate lines and commas (,) to separate columns in the file content. The file content cannot include non-English symbols (for example, Chinese characters). The column content cannot contain special characters such as commas, line breaks, or quotation marks. It is recommended that the column content consist of only letters and numbers.
- Data training
  - The number of columns in the training data must be the same, and there has to be at least 100 data records (a feature with different values is considered as different data records).
  - The training columns cannot contain timestamp formats (such as yy-mm-dd and yyyy-mm-dd).
  - If a column has only one value, the column is considered invalid. Ensure that there are at least two values in the label column and no data is missing.

    NOTE

    The label column is the training target specified in a training task. It is the output (prediction item) for the model trained using the dataset.
  - In addition to the label column, the dataset must contain at least two valid feature columns. Ensure that there are at least two values in each feature column and that the percentage of missing data must be lower than 10%.
  - The CSV file cannot contain a table header, or the training will fail.

3.2.2 What Formats of Images Are Supported by Object Detection or Image Classification Projects?

Images in JPG, JPEG, PNG, or BMP format are supported.

3.3 Creating a Project

3.3.1 Is There a Limit on the Number of ExeML Projects That Can Be Created?

ModelArts ExeML supports image classification, object detection, predictive analytics, sound classification, and text classification projects. Up to 100 ExeML projects can be created.

3.3.2 Why Is There No Data Available in the Dataset Input Path When I Create a Project?

Possible Cause

1. The created OBS bucket and project are not in the same region.
2. Global authorization is not configured for the account.
3. The format of data in the OBS bucket does not meet service requirements.

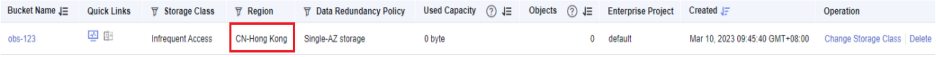
Solution

Check whether the ModelArts project and the created OBS bucket are in the same region.

1. Check the region where the created OBS bucket is located.
   a. Log in to OBS Console.
   b. On the **Object Storage Service** page, to search for a bucket, enter a keyword in **Bucket Name**.

      In the **Region** column, view the region where the created OBS bucket is located.

      Figure 3-3 Region where an OBS bucket is located

Bucket Name	Quick Links	Storage Class	Region	Data Redundancy Policy	Used Capacity	Objects	Enterprise Project	Created	Operation
obs-123		Infrequent Access	CN-Hong Kong	Single-AZ storage	0 byte	0	default	Mar 10, 2023 09:45:40 GMT+08:00	Change Storage Class | Delete

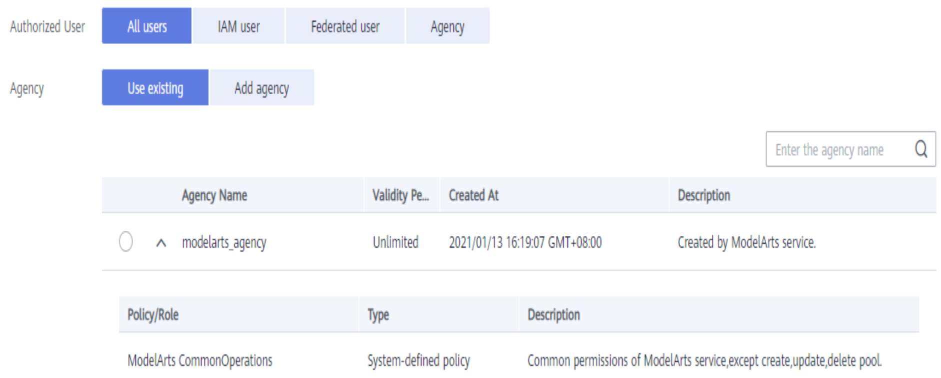
2. Check the region where ModelArts is deployed.

   Log in to the ModelArts management console and view the region where ModelArts is located in the upper left corner.
3. Check whether the region of the created OBS bucket is the same as that of ModelArts. Ensure that they are the same.

Configuring Access Authorization (Global Configuration)

1. Log in to the ModelArts management console. In the left navigation pane, choose **Settings**. The **Global Configuration** page is displayed.
2. Click **Add Authorization**. On the **Add Authorization** page that is displayed, configure the parameters.

Figure 3-4 Viewing permissions



Authorized User: All users | IAM user | Federated user | Agency

Agency: Use existing | Add agency

Agency Name	Validity Pe...	Created At	Description
modelarts_agency	Unlimited	2021/01/13 16:19:07 GMT+08:00	Created by ModelArts service.

Policy/Role	Type	Description
ModelArts CommonOperations	System-defined policy	Common permissions of ModelArts service,except create,update,delete pool.

3. Select **I have read and agree to the ModelArts Service Statement** and click **Create**.

3.4 Labeling Data

3.4.1 Can I Add Multiple Labels to an Image for an Object Detection Project?

Yes. You can add multiple labels to an image.

3.4.2 Why Are Some Images Displayed as Unlabeled After I Upload Labeled Images in an Object Detection Job?

Check whether the labeling files of the images displayed as unlabeled are correct. If the coordinates of the bounding box files exceed those of the images, the images are treated as unlabeled by default in ExeML.

3.5 Training Models

3.5.1 What Should I Do When the Train Button Is Unavailable After I Create an Image Classification Project and Label the Images?

The **Train** button turns to be available when the training images for an image classification project are classified into at least two categories, and each category contains at least five images.

3.5.2 How Do I Perform Incremental Training in an ExeML Project?

Each round of training generates a training version in an ExeML project. If a training result is unsatisfactory (for example, if the precision is not good enough), you can add high-quality data or add or delete labels, and perform training again.

NOTE

- Currently, incremental training is only supported for the following types of ExeML projects: image classification, object detection, and sound classification.
- For better training results, use high-quality data for incremental training to improve data labeling performance.

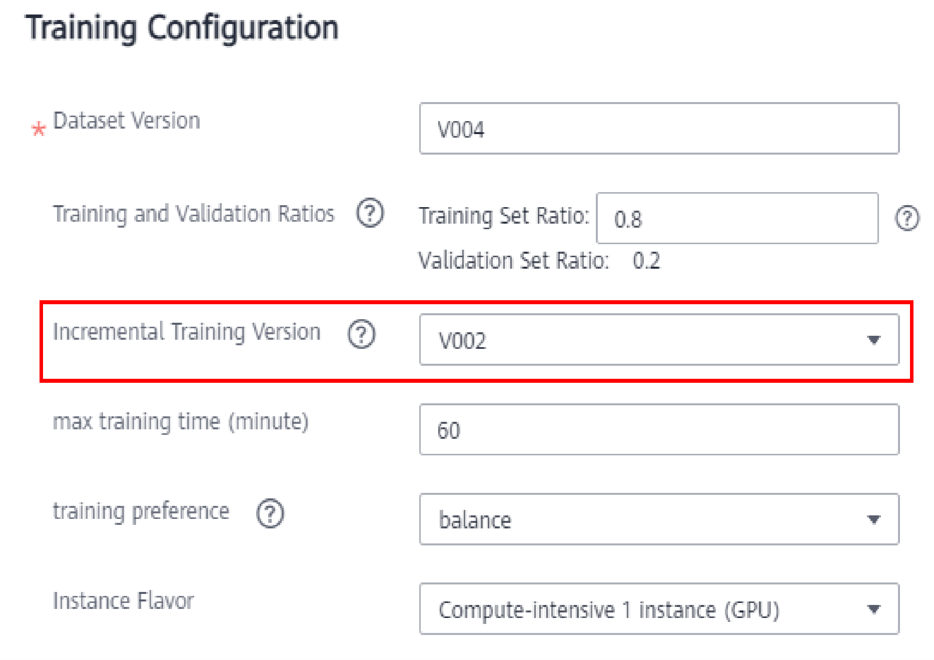
Incremental Training Procedure

1. Log in to the ModelArts console, and click **ExeML** in the left navigation pane.
2. On the **ExeML** page, click a project name. The ExeML details page of the project is displayed.
3. On the **Label Data** page, click the **Unlabeled** tab. On the **Unlabeled** tab page, you can add images or add or delete labels.

   If you add images, label the added images again. If you add or delete labels, check all images and label them again. You also need to check whether new labels need to be added for the labeled data.
4. After all images are labeled, click **Train** in the upper right corner. In the **Training Configuration** dialog box that is displayed, set **Incremental Training Version** to the training version that has been completed to perform incremental training based on this version. Set other parameters as prompted.

   After the settings are complete, click **Yes** to start incremental training. The system automatically switches to the **Train Model** page. After the training is complete, you can view the training details, such as training precision, evaluation result, and training parameters.

Figure 3-5 Selecting an incremental training version



Training Configuration

* Dataset Version	V004
Training and Validation Ratios	Training Set Ratio: 0.8 Validation Set Ratio: 0.2
Incremental Training Version	V002
max training time (minute)	60
training preference	balance
Instance Flavor	Compute-intensive 1 instance (GPU)

3.5.3 Can I Download a Model Trained Using ExeML?

No. The model cannot be downloaded. You can view the model or deploy the model as a real-time service on the **AI Application Management** page.

3.5.4 Why Does ExeML Training Fail?

If the training of an ExeML project fails, perform the following steps to rectify the fault:

1. Access **Billing Center** and check whether the account is in arrears.
   a. If the account is in arrears, **top up the account**.
   b. If the account is not in arrears, go to **2**.
2. Check whether the OBS path for storing image data complies with the following requirements:
   - The OBS path does not contain other folders.
   - The file name does not contain the following special characters: ~`@#$%^&*{}[]:;+=<>/

   If the OBS path meets the requirements, go to **3**.
3. The failure cause may vary depending on the ExeML project.
   - If image recognition training fails, check whether there are damaged images. If there are damaged images, replace or delete them.
   - If object detection training fails, check whether the labeling mode of the dataset is correct. Currently, ExeML supports only rectangle-based labeling.
   - If predictive analytics training fails, check the label column. Currently, the label column supports discrete and continuous data. Only one column can be selected.

- If sound classification training fails, check whether the audio files are 16-bit WAV files.

If the fault persists, submit a **service ticket** for technical support.

3.5.5 What Do I Do If an Image Error Occurred During Model Training Using ExeML?

If an image classification or object detection algorithm of ExeML is used, after the labeled data is trained, the training result is an image error. **Table 3-1** lists solutions to different errors.

Table 3-1 Image errors in image classification and object detection of ExeML

No.	Error Parameter	Error Description	Solution Parameter	Solution Description
1	load failed	The image cannot be decoded or restored.	ignore	The system has ignored this image. No manual operation is required.
2	tf-decode failed	The image cannot be decoded by TensorFlow or restored.	ignore	The system has ignored this image. No manual operation is required.
3	size over	The image size exceeded 5 MB.	resize to small	The system has compressed the image size to be less than 5 MB. No manual operation is required.
4	mode illegal	The image is not in RGB format.	convert to rgb	The system has converted the image to the RGB format. No manual operation is required.
5	type illegal	The file is not an image but can be converted to JPG.	convert to jpg	The system has converted the image to the JPG format. No manual operation is required.

3.5.6 What Do I Do If Error ModelArts.0010 Occurred When I Use ExeML to Start Training as an IAM User?

Use the ACL permission assigned by the tenant account for the OBS bucket used by ModelArts.

3.5.7 What Is the Training Speed of Each Parameter in ExeML Training Preference Settings?

Preference settings are as follows:

performance_first: performance first. The training duration is short and the generated model is small. The training speed is 10 ms for TXT or image training.

balance: balanced performance and precision. The training speed is 14 ms for TXT or image training.

accuracy_first: precision first. The training duration is long and the generated model is large. The training speed is 16 ms for TXT or image training.

3.5.8 What Do I Do If ERROR:input key sound is not in model Occurred When I Use ExeML for Sound Classification Prediction?

According to the error log of the real-time service, the audio file used for prediction is empty. Use a large audio file for prediction.

3.6 Deploying Models

3.6.1 What Type of Service Is Deployed in ExeML?

Models created in ExeML are deployed as real-time services. You can add images or compile code to test the services, as well as call the APIs using the URLs.

After model development is successful, you can choose **Service Deployment** > **Real-Time Services** in the left navigation pane of the ModelArts console to view running services, and stop or delete services.

4 Data Management

4.1 Are There Size Limits for Images to be Uploaded?

For data management, there are limits on the image size when you upload images to the datasets whose labeling type is object detection or image classification. The size of an image cannot exceed 8 MB, and only JPG, JPEG, PNG, and BMP formats are supported.

Note that for ExeML, the size of an image to be uploaded cannot exceed 5 MB.

Solutions:

- Import the images from OBS. Upload images to any OBS directory and import the images from the OBS directory to an existing dataset.
- Use data source synchronization. Upload images to the input directory or its subdirectory of a dataset, and click **Synchronize Data Source** on the dataset details page to add new images. Note that synchronizing a data source will delete the files deleted from OBS from the dataset. Exercise caution when performing this operation.
- Create a dataset. Upload images to any OBS directory. You can directly use the image directory as the input directory to create the dataset.

4.2 What Do I Do If Images in a Dataset Cannot Be Displayed?

Symptom

Images in a created dataset cannot be displayed during labeling, and they cannot be viewed by clicking them. Alternatively, the system displays a message indicating that an error occurred in image loading.

Possible Cause

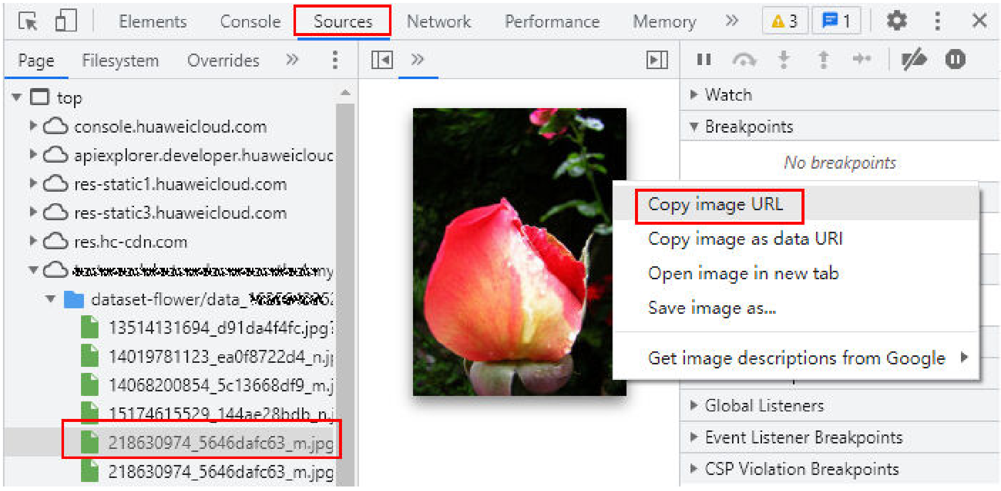
- The local network may be faulty. As a result, OBS cannot be accessed and images cannot be loaded.

- You are not allowed to access the target OBS bucket.
- The OBS bucket or file may be encrypted.
- The OBS storage class does not allow the parallel file system to process images. Therefore, the thumbnails cannot be displayed.

Solution

1. The following uses Google Chrome as an example. Press **F12** to open the browser console, locate the image, and copy the image URL.

Figure 4-1 Obtaining the image URL


Elements
Console
Sources
Network
Performance
Memory
»
⚠3
1
Page
Filesystem
Overrides
»
top
console.huaweicloud.com
apiexplorer.developer.huaweicloud
res-static1.huaweicloud.com
res-static3.huaweicloud.com
res.hc-cdn.com
[redacted]
dataset-flower/data_[redacted]
13514131694_d91da4f4fc.jpg
14019781123_ea0f8722d4_n.j
14068200854_5c13668df9_m.j
15174615529_144ae28bdb_n.j
218630974_5646dafc63_m.jpg
218630974_5646dafc63_m.jpg
Watch
Breakpoints
No breakpoints
Copy image URL
Copy image as data URI
Open image in new tab
Save image as...
Get image descriptions from Google
Global Listeners
Event Listener Breakpoints
CSP Violation Breakpoints

2. Enter the URL in a new browser. The "Your Connection Is Not Private" message is displayed. Click **Advanced** on the page and choose **Proceed to <link> (unsafe)** to go to the target website.
3. After the image is successfully accessed, return to the ModelArts console to access the dataset. The image is displayed.

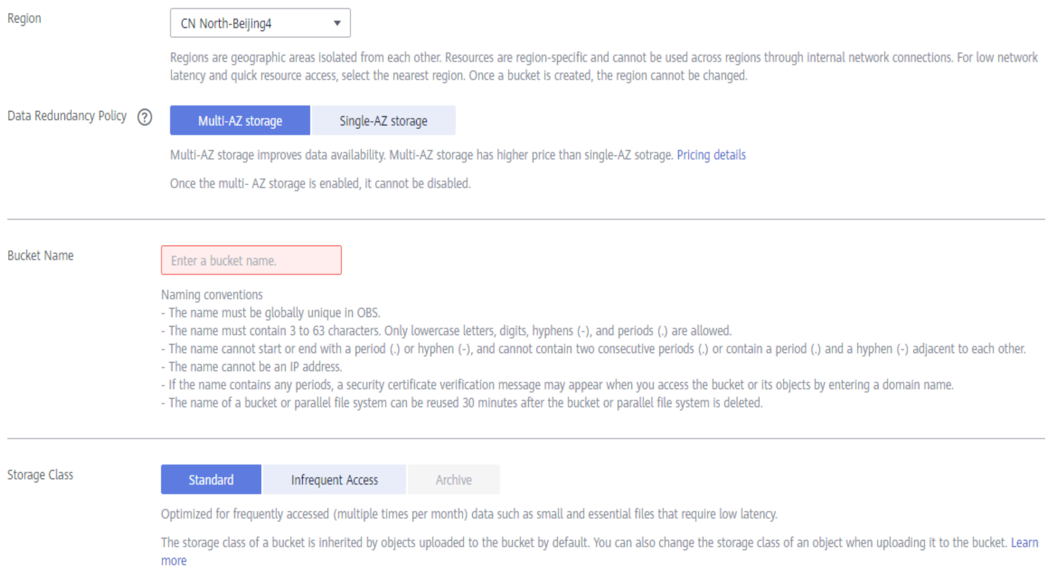
4.3 How Do I Integrate Multiple Object Detection Datasets into One Dataset?

Create a parent directory in an OBS bucket, in the directory add the same number of folders as that of datasets, export one dataset to one folder, and use the parent directory to create a dataset.

Log in to the ModelArts management console and choose **Data Management** > **Datasets**. Click the target dataset to switch to its **Dashboard** page. Then, click **Export** in the upper right corner of the page to export the dataset to a folder in the OBS parent directory.

4.4 What Do I Do If Importing a Dataset Failed?

The possible cause is that the storage class of the target OBS bucket is incorrect. In this case, select a bucket of the standard storage class to import data.



Region: CN North-Beijing4

Regions are geographic areas isolated from each other. Resources are region-specific and cannot be used across regions through internal network connections. For low network latency and quick resource access, select the nearest region. Once a bucket is created, the region cannot be changed.

Data Redundancy Policy: Multi-AZ storage

Multi-AZ storage improves data availability. Multi-AZ storage has higher price than single-AZ sotrage. Pricing details

Once the multi- AZ storage is enabled, it cannot be disabled.

Bucket Name: Enter a bucket name.

Naming conventions
- The name must be globally unique in OBS.
- The name must contain 3 to 63 characters. Only lowercase letters, digits, hyphens (-), and periods (.) are allowed.
- The name cannot start or end with a period (.) or hyphen (-), and cannot contain two consecutive periods (.) or contain a period (.) and a hyphen (-) adjacent to each other.
- The name cannot be an IP address.
- If the name contains any periods, a security certificate verification message may appear when you access the bucket or its objects by entering a domain name.
- The name of a bucket or parallel file system can be reused 30 minutes after the bucket or parallel file system is deleted.

Storage Class: Standard

Optimized for frequently accessed (multiple times per month) data such as small and essential files that require low latency.

The storage class of a bucket is inherited by objects uploaded to the bucket by default. You can also change the storage class of an object when uploading it to the bucket. Learn more

4.5 What Do I Do to Import Locally Labeled Data to ModelArts?

ModelArts allows you to import data by importing datasets. Locally labeled data can be imported from an OBS directory or the manifest file. After the import, you can label the data again or modify the labels in ModelArts Data Management.

For details about how to import data from an OBS directory or manifest file, see **Import Operation**.

4.6 Why Does Data Fail to Be Imported Using the Manifest File?

Symptom

Failed to use the manifest file of the published dataset to import data again.

Possible Cause

Data has been changed in the OBS directory of the published dataset, for example, images have been deleted. Therefore, the manifest file is inconsistent with data in the OBS directory. As a result, an error occurs when the manifest file is used to import data again.

Solution

- Method 1 (recommended): Publish a new version of the dataset again and use the new manifest file to import data.
- Method 2: Modify the manifest file on your local PC, search for data changes in the OBS directory, and modify the manifest file accordingly. Ensure that the manifest file is consistent with data in the OBS directory, and then import data using the new manifest file.

4.7 Where Are Labeling Results Stored?

The ModelArts console provides data visualization capabilities, which allows you to view detailed data and labeling information on the console. To learn more about the path for storing labeling results, see the following description.

Background

When creating a dataset in ModelArts, set both **Input Dataset Path** and **Output Dataset Path** to OBS.

- **Input Dataset Path**: OBS path where the raw data is stored.
- **Output Dataset Path**: Under this path, directories are generated based on the dataset version after data is labeled in ModelArts and datasets are published. The manifest files (containing data and labeling information) used in ModelArts are also stored in this path. For details about the files, see **Directory Structure of Dataset Versions**.

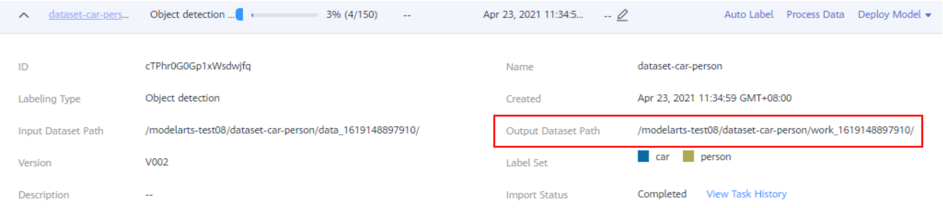
Procedure

1. Log in to the ModelArts console and choose **Data Management** > **Datasets**.
2. Select your desired dataset and click the triangle icon on the left of the dataset name to expand the dataset details. You can obtain the OBS path set for **Output Dataset Path**.

   NOTE

   Before obtaining labeling results, ensure that at least one dataset version is available.

   Figure 4-2 Dataset details

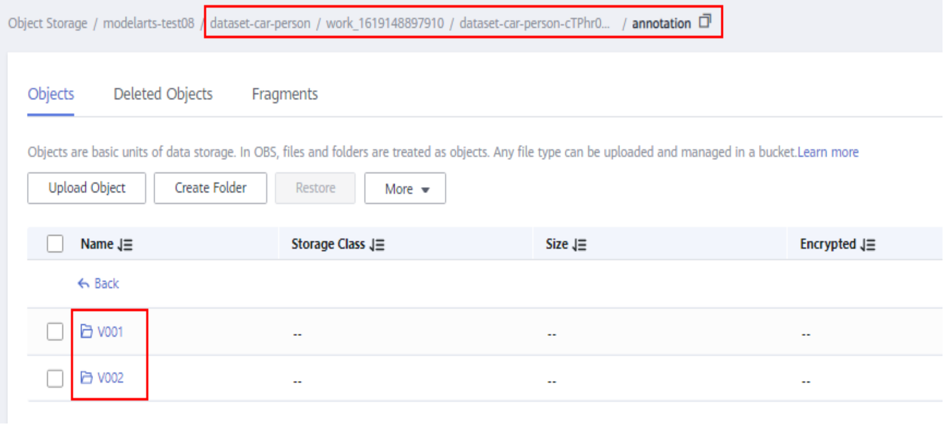
   

   dataset-car-pers... Object detection ... 3% (4/150) -- Apr 23, 2021 11:34:5... --

ID	cTPhr0G0Gp1xWsdwjfq	Name	dataset-car-person
Labeling Type	Object detection	Created	Apr 23, 2021 11:34:59 GMT+08:00
Input Dataset Path	/modelarts-test08/dataset-car-person/data_1619148897910/	Output Dataset Path	/modelarts-test08/dataset-car-person/work_1619148897910/
Version	V002	Label Set	■ car ■ person
Description	--	Import Status	Completed View Task History

3. Log in to the OBS console and locate the directory of the corresponding dataset version from the OBS path obtained in 2 to obtain the labeling result of the dataset.

Figure 4-3 Obtaining the labeling result


Object Storage / modelarts-test08 / dataset-car-person / work_1619148897910 / dataset-car-person-cTPhr0... / annotation
Objects
Deleted Objects
Fragments
Upload Object
Create Folder
Restore
More
Back

Objects are basic units of data storage. In OBS, files and folders are treated as objects. Any file type can be uploaded and managed in a bucket.Learn more

Name	Storage Class	Size	Encrypted
V001	--	--	--
V002	--	--	--

4.8 How Do I Download Labeling Results to a Local PC?

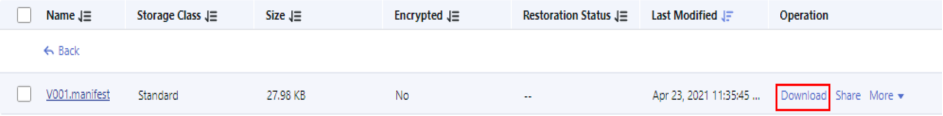
After being published, the labeling information and data in ModelArts datasets are stored as manifest files in the OBS path set for **Output Dataset Path**.

To obtain the OBS path, do as follows:

1. Log in to the ModelArts management console and choose **Data Management** > **Datasets**.
2. Locate the target dataset and click the triangle icon on the left of the dataset name to expand the dataset details. You can obtain the OBS path set for **Output Dataset Path**.
3. Log in to the OBS management console and locate the version directory from the obtained OBS path to obtain the labeling result of the dataset.

To download the labeling results to a local PC, go to the OBS path where the manifest files are stored and click **Download**.

Figure 4-4 Downloading labeling results



Name	Storage Class	Size	Encrypted	Restoration Status	Last Modified	Operation
V001.manifest	Standard	27.98 KB	No	--	Apr 23, 2021 11:35:45 ...	Download Share More

4.9 Why Cannot Team Members Receive Emails for a Team Labeling Task?

The possible causes are as follows:

- All dataset data has been labeled. An email can be sent to team members only if there is unlabeled data in the dataset when the team labeling task is created.
- Team members receive emails for team labeling tasks. No email will be sent when you create a labeling team or add members to a labeling team.

- Your email address has not been configured or has been incorrectly configured. For details about how to configure an email address, see **Managing Team Members**.
- Team members' email addresses are blocked.

4.10 How Data Is Distributed Between Team Members During Team Labeling?

Data is evenly assigned to all members in a labeling team, but not assigned to certain members.

- If the data cannot be evenly distributed because the number of images is not in proportion to the number of team members, the excessive images will be randomly distributed to team members.
- If the number of samples is less than the number of team members, some members may not be allocated with samples. Samples are allocated only to annotators. For example, there are 10,000 samples to be labeled and all the five members in the team are annotators, each annotator will be allocated with 2000 samples.

4.11 How Do I Merge Two Datasets?

Datasets cannot be merged.

However, you can perform the following operations to merge the data of two datasets into one dataset.

For example, to merge datasets A and B, do the following:

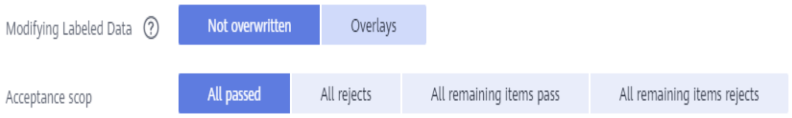
1. Publish datasets A and B.
2. Obtain the manifest files of the two datasets from the OBS path set for **Output Dataset Path**.
3. Create empty dataset C and select an empty OBS folder for **Input Dataset Path**.
4. Import the manifest files of datasets A and B to dataset C.

   After the import is complete, data in datasets A and B is merged into dataset C. To use the merged dataset, publish dataset C.

4.12 Does Auto Labeling Support Polygons?

No. Polygons cannot be used in auto labeling. Only rectangles can be used in auto labeling. If a sample is labeled using other bounding boxes, the sample will not be trained.

4.13 What Do the Options for Accepting a Team Labeling Task Mean?



All passed: All items, including the rejected ones will pass the review.

All rejects: All items, including the ones that have passed the review will be rejected. In this case, the passed items must be labeled and reviewed again in the next acceptance.

All remaining items pass: The rejected items are still rejected, and the remaining items will automatically pass the review.

All remaining items rejects: The selected items that have passed the review do not need to be labeled. All the selected items that have been rejected and the items that have not been selected must be labeled again for acceptance.

4.14 Why Are Images Displayed in Different Angles Under the Same Account?

There are rotation angles of certain images, and the rules of processing such images vary depending on browsers. The following figures show compatibility with browsers.

- L indicates the latest version. L3 indicates the latest three stable browser versions when the product is released.
- If your browser is of an earlier version, the page display will be adversely affected, and the system will prompt you to upgrade your browser.
- If your browser is not compatible with the management console, the system will advise you to upgrade your browser or install a desired browser.

Table 4-1 Compatibility with PC browsers

Browser	Version	OS	Compatibility
Internet Explorer	11	Windows 7	Not guaranteed
Microsoft Edge	L3	Windows 10	Fully compatible
	< 79	Windows 10	Not guaranteed
Mozilla Firefox	L3	Windows 10	Fully compatible

Browser	Version	OS	Compatibility
	L3	CentOS 7+	Partially compatible You can use this version to perform basic interactive operations, but visual and interactive effects may be affected.
	L3	Ubuntu 14.04 LTS+	Partially compatible You can use this version to perform basic interactive operations, but visual and interactive effects may be affected.
	L3	macOS 10+	Partially compatible You can use this version to perform basic interactive operations, but visual and interactive effects may be affected.
Google Chrome	L3	Windows 10	Fully compatible
	L3	CentOS 7+	Partially compatible You can use this version to perform basic interactive operations, but visual and interactive effects may be affected.
	L3	Ubuntu 14.04 LTS+	Partially compatible You can use this version to perform basic interactive operations, but visual and interactive effects may be affected.
	L3	macOS 10+	Partially compatible You can use this version to perform basic interactive operations, but visual and interactive effects may be affected.
Safari	L2	macOS 10+	Partially compatible You can use this version to perform basic interactive operations, but visual and interactive effects may be affected.

Table 4-2 Compatibility with mobile browsers

Browser	Version	OS	Compatibility
Chrome	L3	Android	Fully compatible
Safari	L3	IOS	Fully compatible
UC Browser	L3	Android	Fully compatible
QQ Browser	L3	Android	Fully compatible
360 Secure Browser	L3	Android	Fully compatible
Baidu Browser	L3	Android	Fully compatible

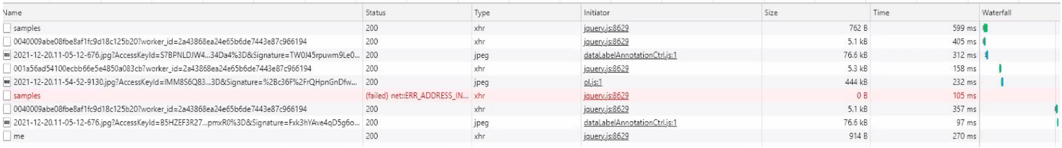
4.15 Do I Need to Train Data Again If New Data Is Added After Auto Labeling Is Complete?

After auto labeling is complete, confirm the labeled data. If you add new data before confirming the labeled data, all unlabeled data will be automatically labeled again. If you add new data after confirming the labeled data, the data must be trained again.

4.16 Why Does the System Display a Message Indicating My Label Fails to Save on ModelArts?

Symptom

Take the Google Chrome browser as an example. When an image is labeled for the first time, the system displays a message in the upper right corner, indicating that the label fails to save. But when the same image is labeled the second time, a message is displayed, indicating that the label is saved. This issue occurs occasionally. When this issue occurs, the request status is **(failed)net::ERR_*ADDRESS*_IN_USE**, which is obtained by pressing **F12** on the Google Chrome browser and clicking **Network**.



Name	Status	Type	Initiator	Size	Time	Waterfall
samples	200	xhr	jquery.js:8629	762 B	599 ms	
0040009abe08fbe8af1fc9d18c125b20?worker_id=2a43868ea24e65b6de7443e87c966194	200	xhr	jquery.js:8629	5.1 kB	405 ms	
2021-12-20.11-05-12-676.jpg?AccessKeyId=S7BPNLDJW4...34Da4%3D&Signature=TW0J45rpuwm9Le0...	200	jpeg	dataLabelAnnotationCtrl.js:1	76.6 kB	312 ms	
001a56ad54100ecbb66e5e4850a083cb?worker_id=2a43868ea24e65b6de7443e87c966194	200	xhr	jquery.js:8629	5.3 kB	158 ms	
2021-12-20.11-54-52-9130.jpg?AccessKeyId=IMM8S6Q83...3D&Signature=%2Bc36F%2FrQHpnGnDfw...	200	jpeg	ol.js:1	444 kB	232 ms	
samples	(failed) net::ERR_ADDRESS_IN...	xhr	jquery.js:8629	0 B	105 ms	
0040009abe08fbe8af1fc9d18c125b20?worker_id=2a43868ea24e65b6de7443e87c966194	200	xhr	jquery.js:8629	5.1 kB	357 ms	
2021-12-20.11-05-12-676.jpg?AccessKeyId=B5HZEF3R27...pmxR0%3D&Signature=Fxk3hYAve4qD5g6o...	200	jpeg	dataLabelAnnotationCtrl.js:1	76.6 kB	97 ms	
me	200	xhr	jquery.js:8629	914 B	270 ms	

Possible Cause

The local network is faulty, for example, the network is unstable, or the network configuration is incorrect.

Solution

- Switch to a stable network and try again.
- Initialize the network configuration. To do so, start **cmd** as the administrator, run the **netsh winsock reset** command, and restart the computer. Then, log in to the data labeling platform again.

4.17 Can One Label By Identified Among Multiple Labels?

After a model is trained with multiple labels and deployed as a real-time service, all the labels are identified. If only one type of label needs to be identified, train a model dedicated for identifying the label. To speed up the label identification, select a high flavor for deploying the model.

4.18 Why Are Newly Added Images Not Automatically Labeled After Data Amplification Is Enabled?

After data amplification is enabled, images newly added in image classification datasets cannot be automatically labeled, but those added in object detection datasets can be.

4.19 How Do I Use Soft-NMS to Reduce Bounding Box Overlapping?

YOLOv3 algorithms subscribed to in Huawei Cloud AI Gallery can use Soft-NMS to reduce overlapped bounding boxes. No official information has been released to show that YOLOv5 algorithms support this function. Use this function in custom algorithms.

4.20 How Do I Add Images to a Validation or Training Dataset?

You are not allowed to manually add images to a training or validation dataset, but can only set a training and validation ratio. Then, the system randomly allocates the images to the training and validation datasets based on the ratio.

Setting a Training and Validation Ratio

When you publish a dataset, only the dataset of the image classification, object detection, text classification, or sound classification type supports data splitting.

By default, data splitting is disabled. After this function is enabled, set a training and validation ratio.

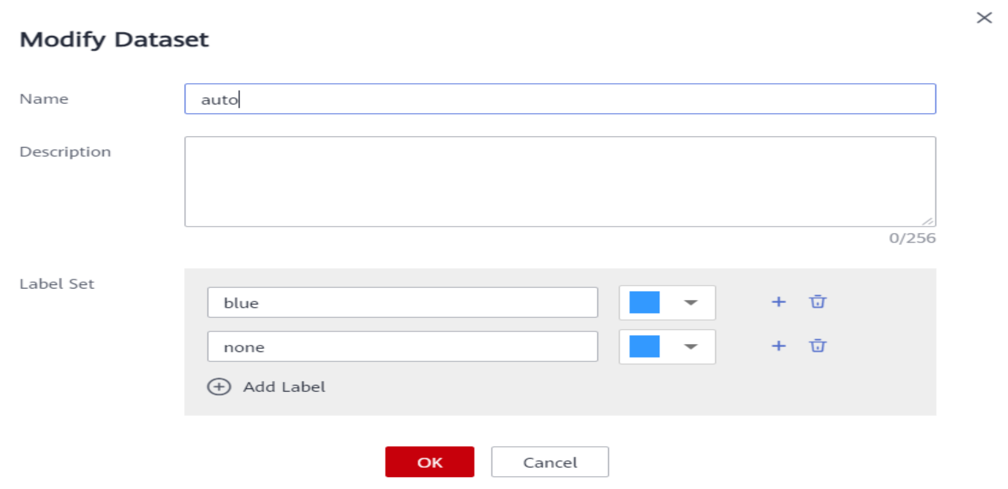
Enter a value ranging from 0 to 1 for the training set ratio. After the training set ratio is set, the validation set ratio is determined. The sum of the training set ratio and the validation set ratio is 1.

The training set ratio is the ratio of sample data used for model training. The validation set ratio is the ratio of the sample data used for model validation. The training and validation ratios affect the performance of training templates.

4.21 Can I Customize Labels for an Object Detection Dataset?

Yes. You can add custom labels to the label set of an object detection dataset by modifying the dataset.

Figure 4-5 Modify Dataset



Modify Dataset

Name: auto

Description:

0/256

Label Set

blue	🟦	+ 🗑
none	🟦	+ 🗑

⊕ Add Label

4.22 What ModelArts Data Management Can Be Used for?

The functions provided ModelArts data management vary depending on the type of the dataset.

Dataset Type	Labeling Type	Creating a Dataset	Importing Data	Exporting Data	Publishing a Dataset	Modifying a Dataset	Managing Dataset Versions	Auto Grouping	Data Feature Engineering
Files	Image classification	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported

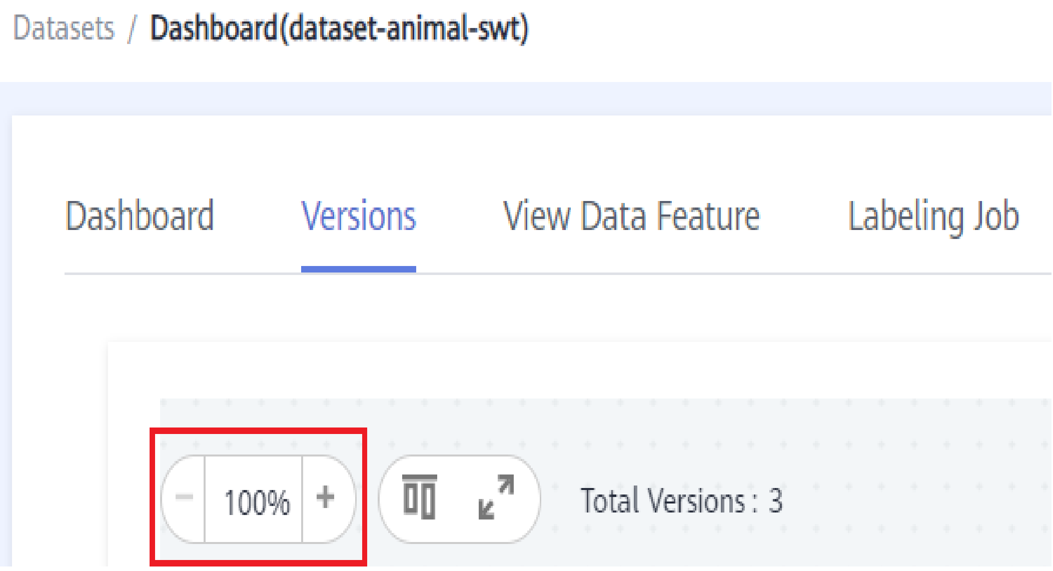
Dataset Type	Labeling Type	Creating a Dataset	Importing Data	Exporting Data	Publishing a Dataset	Modifying a Dataset	Managing Dataset Versions	Auto Grouping	Data Feature Engineering
	Object detection	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported
	Image segmentation	Supported	Supported	Supported	Supported	Supported	Supported	Supported	N/A
	Sound classification	Supported	Supported	N/A	Supported	Supported	Supported	N/A	N/A
	Speech labeling	Supported	Supported	N/A	Supported	Supported	Supported	N/A	N/A
	Speech paragraph labeling	Supported	Supported	N/A	Supported	Supported	Supported	N/A	N/A
	Text classification	Supported	Supported	N/A	Supported	Supported	Supported	N/A	N/A
	Named entity recognition	Supported	Supported	N/A	Supported	Supported	Supported	N/A	N/A
	Text triplet	Supported	Supported	N/A	Supported	Supported	Supported	N/A	N/A
	Videos	Supported	Supported	N/A	Supported	Supported	Supported	N/A	N/A

Dataset Type	Labeling Type	Creating a Dataset	Importing Data	Exporting Data	Publishing a Dataset	Modifying a Dataset	Managing Dataset Versions	Auto Grouping	Data Feature Engineering
	Free format	Supported	N/A	Supported	Supported	Supported	Supported	N/A	N/A
Tables	Tables	Supported	Supported	N/A	Supported	Supported	Supported	N/A	N/A

4.23 Why Is My New Dataset Version Unavailable in Versions?

The version list can be zoomed in or out. Zoom out the page before searching.

Click the name of the target dataset to go to the dataset overview page. Then, zoom out the **Versions** page.

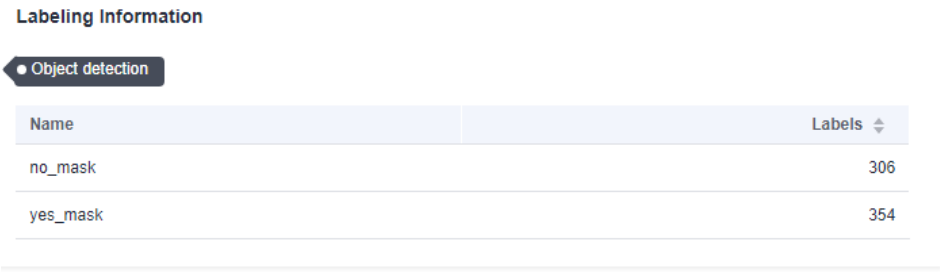


4.24 How Do I View the Size of a Dataset?

Only the number of samples in a dataset is collected in data management. There is no entrance to view the dataset size.

4.25 How Do I View Labeling Details of a New Dataset?

1. Log in to the ModelArts management console and choose **Data Management** > **Datasets** from the navigation pane on the left.
2. Locate the target dataset by name and click its name. The **Dashboard** tab page is displayed.
3. On the **Dashboard** tab page, click **View Details** in the **Labeling Information** area.



Labeling Information

● Object detection

Name	Labels
no_mask	306
yes_mask	354

4.26 How Do I Export Labeled Data?

Only datasets of image classification, object detection, and image segmentation types can be exported.

- For image classification datasets, only the label files in TXT format can be exported.
- For object detection datasets, only XML label files in Pascal VOC format can be exported.
- For image segmentation datasets, only XML label files in Pascal VOC format and mask images can be exported.

For other types of datasets, use **data version publishing** to publish the datasets.

4.27 How Do I Split a Dataset?

When you publish a dataset, only the dataset of the image classification, object detection, text classification, or sound classification type supports data splitting.

By default, data splitting is disabled. After this function is enabled, set the training and validation ratios.

Enter a value ranging from 0 to 1 for the training set ratio. After the training set ratio is set, the validation set ratio is determined. The sum of the training set ratio and the validation set ratio is 1.

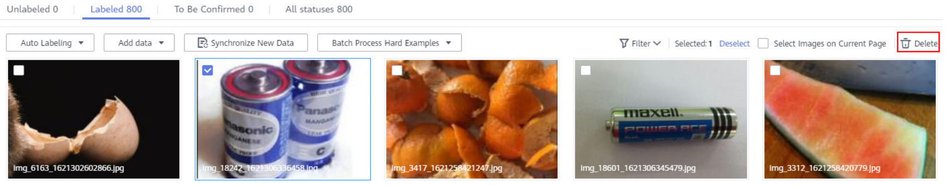
The training set ratio is the ratio of sample data used for model training. The validation set ratio is the ratio of the sample data used for model validation. The training and validation ratios affect the performance of training templates.

4.28 How Do I Delete a Dataset Image?

1. Log in to the ModelArts management console. In the navigation pane, choose **Data Management** > **Label Data**. The data labeling list is displayed. Click the dataset from which you want to delete images. The labeling details page is displayed.
2. On the **All statuses**, **Unlabeled**, or **Labeled** tab page, select the images to be deleted or click **Select Images on Current Page**, and click 🗑 to delete them. In the displayed dialog box, select or deselect **Delete the source files from OBS** as required. After confirmation, click **Yes** to delete the images.

   If a tick is displayed in the upper left corner of an image, the image is selected. If no image is selected on the page, 🗑 is unavailable.

   Figure 4-6 Deleting a dataset image

   
   Unlabeled 0
Labeled 800
To Be Confirmed 0
All statuses 800
Auto Labeling
Add data
Synchronize New Data
Batch Process Hard Examples
Filter
Selected: 1
Deselect
Select Images on Current Page
Delete
img_6163_1621302602866.jpg
img_18242_1621306336458.jpg
img_3417_1621258421247.jpg
img_18601_1621306345479.jpg
img_3312_1621258420779.jpg

4.29 Why Is There No Sample in the ModelArts Dataset Downloaded from AI Gallery and Then an OBS Bucket?

Check the format of the data downloaded from AI Gallery. For example, compressed packages and Excel files will be ignored. The following table lists the supported formats.

Dataset Type	Labeling Type	Creating a Dataset	Importing Data	Exporting Data	Publishing a Dataset	Modifying a Dataset	Managing Dataset Versions	Auto Grouping	Data Feature Engineering
Files	Image classification	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported
	Object detection	Supported	Supported	Supported	Supported	Supported	Supported	Supported	Supported

Dataset Type	Labeling Type	Creating a Dataset	Importing Data	Exporting Data	Publishing a Dataset	Modifying a Dataset	Managing Dataset Versions	Auto Grouping	Data Feature Engineering
	Image segmentation	Supported	Supported	Supported	Supported	Supported	Supported	Supported	N/A
	Sound classification	Supported	Supported	N/A	Supported	Supported	Supported	N/A	N/A
	Speech labeling	Supported	Supported	N/A	Supported	Supported	Supported	N/A	N/A
	Speech paragraph labeling	Supported	Supported	N/A	Supported	Supported	Supported	N/A	N/A
	Text classification	Supported	Supported	N/A	Supported	Supported	Supported	N/A	N/A
	Named entity recognition	Supported	Supported	N/A	Supported	Supported	Supported	N/A	N/A
	Text triplet	Supported	Supported	N/A	Supported	Supported	Supported	N/A	N/A
	Videos	Supported	Supported	N/A	Supported	Supported	Supported	N/A	N/A
	Free format	Supported	N/A	Supported	Supported	Supported	Supported	N/A	N/A
Tables	Tables	Supported	Supported	N/A	Supported	Supported	Supported	N/A	N/A

5 Notebook (New Version)

5.1 Constraints

5.1.1 Is sudo Privilege Escalation Supported?

For security purposes, notebook instances do not support sudo privilege escalation.

5.1.2 Does ModelArts Support apt-get?

Terminal in ModelArts DevEnviron does not support **apt-get**. You can use a **custom image** to support it.

5.1.3 Is the Keras Engine Supported?

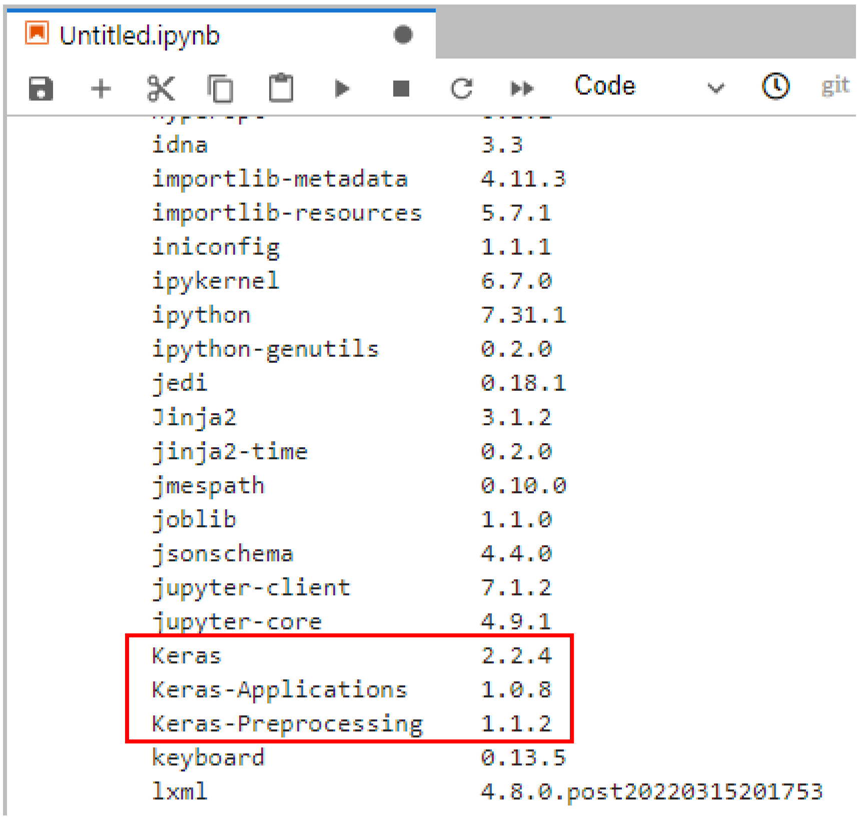
Notebook instances in **DevEnviron** support the Keras engine. The Keras engine is not supported in job training and model deployment (inference).

Keras is an advanced neural network API written in Python. It is capable of running on top of TensorFlow, CNTK, or Theano. Notebook instances in **DevEnviron** support **tf.keras**.

How Do I View Keras Versions?

1. On the ModelArts management console, create a notebook instance with image **TensorFlow-1.13** or **TensorFlow-1.15**.
2. Access the notebook instance. In JupyterLab, run **!pip list** to view Keras versions.

Figure 5-1 Viewing Keras versions



```
idna                     3.3
importlib-metadata       4.11.3
importlib-resources      5.7.1
iniconfig                1.1.1
ipykernel                6.7.0
ipython                  7.31.1
ipython-genutils         0.2.0
jedi                     0.18.1
Jinja2                   3.1.2
jinja2-time              0.2.0
jmespath                 0.10.0
joblib                   1.1.0
jsonschema               4.4.0
jupyter-client           7.1.2
jupyter-core             4.9.1
Keras                    2.2.4
Keras-Applications       1.0.8
Keras-Preprocessing      1.1.2
keyboard                 0.13.5
lxml                     4.8.0.post20220315201753
```

5.1.4 Does ModelArts Support the Caffe Engine?

The Python 2 environment of ModelArts supports Caffe, but the Python 3 environment does not support it.

5.1.5 Can I Install MoXing in a Local Environment?

No. MoXing can be used only on ModelArts.

5.1.6 Can Notebook Instances Be Remotely Logged In?

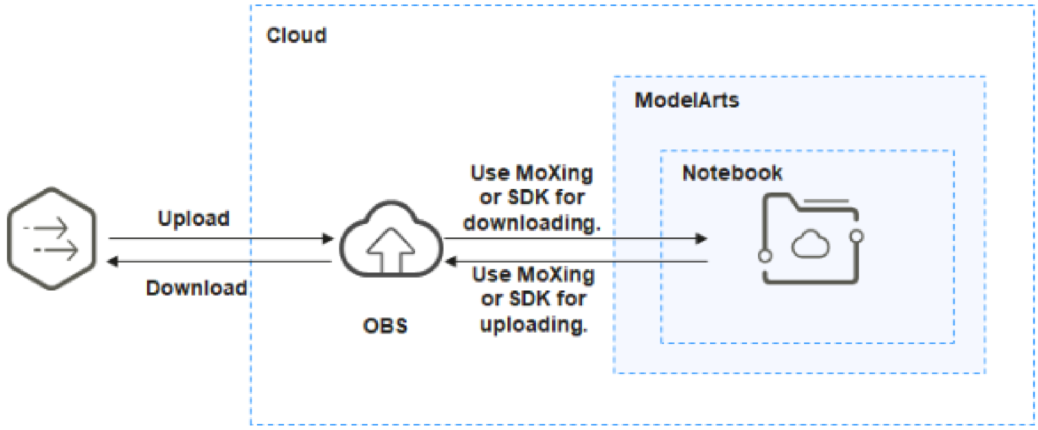
The notebook instances of the new version can be remotely logged in. To do so, enable remote SSH when you create the notebook instances. Remotely log in to a notebook instance from a local IDE through **PyCharm professional** or **VS Code**.

5.2 Data Upload or Download

5.2.1 How Do I Upload a File from a Notebook Instance to OBS or Download a File from OBS to a Notebook Instance?

In a notebook instance, you can call the ModelArts MoXing API or SDK to exchange data with OBS for uploading a file to OBS or downloading a file from OBS to the notebook instance.

Figure 5-2 Uploading or downloading a file


Cloud
ModelArts
Notebook
Upload
Download
OBS
Use MoXing or SDK for downloading.
Use MoXing or SDK for uploading.

For details about how to upload files using OBS Browser, see **Uploading and Downloading Files Through OBS Browser+**.

Method 1: Using MoXing to Upload and Download a File

Developed by the ModelArts team, MoXing is a distributed training acceleration framework built on open-source deep learning engines such as TensorFlow and PyTorch. MoXing makes model coding easier and more efficient.

MoXing provides a set of file object APIs for reading and writing OBS files.

For details about the mapping between MoXing APIs and native APIs and how to call APIs, see **MoXing File Operations**.

Sample code:

```
import moxing as mox

# Download the OBS folder sub_dir_0 from OBS to a notebook instance.
mox.file.copy_parallel('obs://bucket_name/sub_dir_0', '/home/ma-user/work/sub_dir_0')
# Download the OBS file obs_file.txt from OBS to a notebook instance.
mox.file.copy('obs://bucket_name/obs_file.txt', '/home/ma-user/work/obs_file.txt')

# Upload the OBS folder sub_dir_0 from a notebook instance to OBS.
mox.file.copy_parallel('/home/ma-user/work/sub_dir_0', 'obs://bucket_name/sub_dir_0')
# Upload the OBS file obs_file.txt from a notebook instance to OBS.
mox.file.copy('/home/ma-user/work/obs_file.txt', 'obs://bucket_name/obs_file.txt')
```

Method 2: Using SDK to Upload and Download a File

Call the ModelArts SDK for downloading a file from OBS.

Sample code: Download **file1.txt** from OBS to **/home/ma-user/work/** in the notebook instance. All the bucket name, folder name, and file name are customizable.

```
from modelarts.session import Session
session = Session()
session.obs.download_file(src_obs_file="obs://bucket-name/dir1/file1.txt", dst_local_dir="/home/ma-user/
work/")
```

Call the ModelArts SDK for downloading a folder from OBS.

Sample code: Download **dir1** from OBS to **/home/ma-user/work/** in the notebook instance. The bucket name and folder name are customizable.

```
from modelarts.session import Session
session = Session()
session.obs.download_dir(src_obs_dir="obs://bucket-name/dir1/", dst_local_dir="/home/ma-user/work/")
```

Call the ModelArts SDK for uploading a file to OBS.

Sample code: Upload **file1.txt** in the notebook instance to OBS bucket **obs://bucket-name/dir1/**. All the bucket name, folder name, and file name are customizable.

```
from modelarts.session import Session
session = Session()
session.obs.upload_file(src_local_file='/home/ma-user/work/file1.txt', dst_obs_dir='obs://bucket-name/dir1/')
```

Call the ModelArts SDK for uploading a folder to OBS.

Sample code: Upload **/work/** in the notebook instance to **obs://bucket-name/dir1/work/** of **bucket-name**. The bucket name and folder name are customizable.

```
from modelarts.session import Session
session = Session()
session.obs.upload_dir(src_local_dir='/home/ma-user/work/', dst_obs_dir='obs://bucket-name/dir1/')
```

Error Handling

If you download a file from OBS to your notebook instance and the system displays error message "Permission denied", perform the following operations for troubleshooting:

- Ensure that the target OBS bucket and notebook instance are in the same region, for example, **CN North-Beijing4**. If the OBS bucket and notebook instance are in different regions, the access to OBS is denied. For details, see **Check whether the OBS bucket and ModelArts are in the same region**
- Ensure that the notebook account has the permission to read data in the OBS bucket. If the account does not have the permission, see **How Do I Access the OBS Bucket of Another Account from a Notebook Instance?**

5.2.2 How Do I Upload Local Files to a Notebook Instance?

For details about how to upload files to JupyterLab in notebook of the new version, see **Uploading Files to JupyterLab**.

5.2.3 How Do I Import Large Files to a Notebook Instance?

- **Large files (files larger than 100 MB)**

  Use OBS to upload large files. To do so, use OBS Browser to upload a local file to an OBS bucket and use ModelArts SDK to download the file from OBS to a notebook instance.

  To upload files using OBS Browser, see **Uploading and Downloading Files Through OBS Browser+**.

  For details about how to use ModelArts SDK or MoXing to download files from OBS, see **How Do I Upload a File from a Notebook Instance to OBS or Download a File from OBS to a Notebook Instance?**

- **Folders**

  Compress a folder into a package and upload the package in the same way as uploading a large file. After the package is uploaded to a notebook instance, decompress it on the **Terminal** page.

  ```
  unzip xxx.zip # Directly decompress the package in the path where the package is stored.
  ```

  For more details, search for the decompression command in mainstream search engines.

5.2.4 Where Will the Data Be Uploaded to?

Data may be stored in OBS or EVS, depending on which kind of storage you have configured for your Notebook instances:

- OBS

  After you click **upload**, the data is directly uploaded to the target OBS path specified when the notebook instance was created.

- EVS

  After you click **upload**, the data is uploaded to the instance container, that is, the **~/work** directory on the **Terminal** page.

5.2.5 How Do I Download Files from a Notebook Instance to a Local Computer?

For details about how to download files from JupyterLab in notebook of the new version, see **Downloading a File from JupyterLab to a Local Path**.

5.2.6 How Do I Copy Data from Development Environment Notebook A to Notebook B?

Data cannot be directly copied from notebook A to notebook B. To copy data, do as follows:

1. Upload the data of notebook A to OBS.
2. Download data from OBS to notebook B.

For details about how to upload and download files, see **How Do I Upload a File from a Notebook Instance to OBS or Download a File from OBS to a Notebook Instance?**

5.2.7 What Do I Do If Uploading a File Failed?

Symptom

When a file is uploaded to a notebook instance, the uploading is consistently in progress on the GUI. When a file is uploaded through MoXing, an error occurred. When an OBS file is uploaded, no bucket is displayed, and the message "Obtaining data" is displayed.

Possible Cause

OBS access is not authorized.

Solution

For details, see **Configuring Access Authorization**.

5.3 Data Storage

5.3.1 How Do I Rename an OBS File?

OBS files cannot be renamed on the OBS console. To rename an OBS file, call a MoXing API in an existing or newly created notebook instance.

The following shows an example:

Rename **obs_file.txt obs_file_2.txt**.

```
import moxing as mox
mox.file.rename('obs://bucket_name/obs_file.txt', 'obs://bucket_name/obs_file_2.txt')
```

5.3.2 Do Files in /cache Still Exist After a Notebook Instance is Stopped or Restarted? How Do I Avoid a Restart?

Temporary files are stored in the **/cache** directory and will not be saved after the notebook instance is stopped or restarted. Data stored in the **/home/ma-user/work** directory will be retained after the notebook instance is stopped or restarted.

To avoid a restart, do not train heavy-load jobs that consume large amounts of CPU, GPU, or memory resources in the development environment.

5.3.3 How Do I Use the pandas Library to Process Data in OBS Buckets?

Step 1 Download data from OBS to a notebook instance. For details, see **Downloading a File from JupyterLab to a Local Path**.

Step 2 Process pandas data by following the instructions provided in ***pandas User Guide***.

----End

5.3.4 How Do I Access the OBS Bucket of Another Account from a Notebook Instance?

To access OBS files of another account from a notebook instance, you must have read and write permissions on the target OBS bucket.

Contact the OBS bucket creator to grant read and write permissions on the current HUAWEI CLOUD account's OBS bucket by referring to **Granting Read and Write Permissions on a Bucket to Other HUAWEI CLOUD Accounts**. If your account is an IAM account or in other scenarios, see section "Configuration Cases in Typical

Permission Control Scenarios" in ***Object Storage Service Permissions Configuration Guide*** for instructions about how to grant OBS bucket permissions.

After obtaining read and write permissions on the OBS bucket, you can use the MoXing API in your notebook instance to access the OBS bucket and read data.

5.3.5 What Is the Default Working Directory on JupyterLab?

- **OBS**

  For this type of notebook instances, the files uploaded to JupyterLab are stored in the OBS path specified during the instance creation by default.

  All read and write operations are performed on the files selected in the OBS path and are irrelevant to the current instance space. To synchronize data from an OBS path to the instance space, use **JupyterLab upload and download functions**.
- **Notebook instance with EVS storage**

  For this type of notebook instances, the files uploaded to JupyterLab are by default stored in the EVS space automatically allocated during the instance creation.

  All read and write operations are performed on the files selected in the EVS space. You can mount your data to the **~/work** directory.

5.4 Environment Configurations

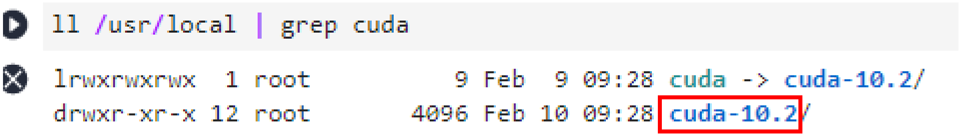
5.4.1 How Do I Check the CUDA Version Used by a Notebook Instance?

Run the following command to view the CUDA version of the target notebook instance:

```
ll /usr/local | grep cuda
```

The following shows an example.

Figure 5-3 Checking the CUDA version in the current environment



```
ll /usr/local | grep cuda
```

```
lrwxrwxrwx  1 root          9 Feb  9 09:28 cuda -> cuda-10.2/
drwxr-xr-x 12 root       4096 Feb 10 09:28 cuda-10.2/
```

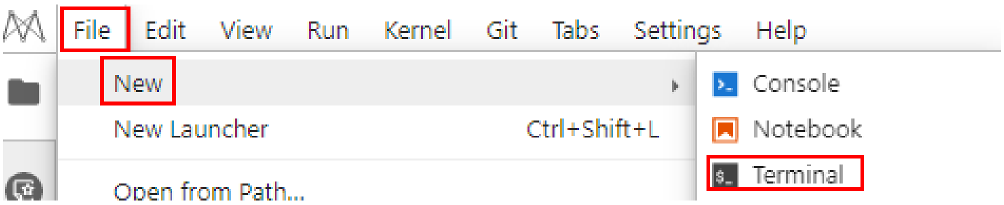
In the preceding example, the CUDA version is 10.2.

5.4.2 How Do I Enable the Terminal Function in DevEnviron of ModelArts?

1. Log in to the ModelArts management console, and choose **DevEnviron** > **Notebooks**.
2. Create a notebook instance. When the instance is running, click **Open** in the **Operation** column. The **JupyterLab** page is displayed.

3. Choose **File** > **New** > **Terminal**. The **Terminal** page is displayed.

Figure 5-4 Going to the **Terminal** page



5.4.3 How Do I Install External Libraries in a Notebook Instance?

Multiple environments such as Jupyter and Python have been integrated into ModelArts notebook to support many frameworks, including TensorFlow, MindSpore, PyTorch, and Spark. You can use **pip install** to install external libraries in Jupyter Notebook or on the **Terminal** page.

Installing External Libraries in Jupyter Notebook

You can use JupyterLab to install Shapely in the **TensorFlow-1.8** environment.

1. Open a notebook instance and access the **Launcher** page.
2. In the **Notebook** area, click **TensorFlow-1.8** and create an IPYNB file.
3. In the new notebook instance, enter the following command in the code input bar:

   !pip install Shapely

Installing External Libraries on the Terminal Page

You can use **pip** to install external libraries in the **TensorFlow-1.8** environment on the **Terminal** page. For example, to install Shapely:

1. Open a notebook instance and access the **Launcher** page.
2. In the **Other** area, click **Terminal** and create a terminal file.
3. Enter the following commands in the code input box to obtain the kernel of the current environment and activate the Python environment on which the installation depends:

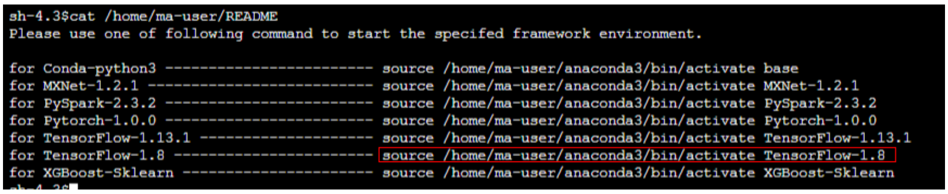
   cat /home/ma-user/README

   source /home/ma-user/anaconda3/bin/activate TensorFlow-1.8

   NOTE

   To install TensorFlow in another Python environment, replace **TensorFlow-1.8** in the command with the target engine.

Figure 5-5 Activating the environment



```
sh-4.3$cat /home/ma-user/README
Please use one of following command to start the specifed framework environment.

for Conda-python3 ----------------------- source /home/ma-user/anaconda3/bin/activate base
for MXNet-1.2.1 ------------------------- source /home/ma-user/anaconda3/bin/activate MXNet-1.2.1
for PySpark-2.3.2 ----------------------- source /home/ma-user/anaconda3/bin/activate PySpark-2.3.2
for Pytorch-1.0.0 ----------------------- source /home/ma-user/anaconda3/bin/activate Pytorch-1.0.0
for TensorFlow-1.13.1 ------------------- source /home/ma-user/anaconda3/bin/activate TensorFlow-1.13.1
for TensorFlow-1.8 ---------------------- source /home/ma-user/anaconda3/bin/activate TensorFlow-1.8
for XGBoost-Sklearn --------------------- source /home/ma-user/anaconda3/bin/activate XGBoost-Sklearn
sh-4.3$
```

4. Run the following command in the code input box to install Shapely:

   pip install Shapely

5.4.4 How Do I Obtain the External IP Address of My Local PC?

Search for "IP address lookup" in a mainstream search engine.

Figure 5-6 IP address lookup

5.4.5 How Can I Resolve Abnormal Font Display on a ModelArts Notebook Accessed from iOS?

Symptom

When a ModelArts notebook is accessed from iOS, the font is displayed abnormally.

Solution

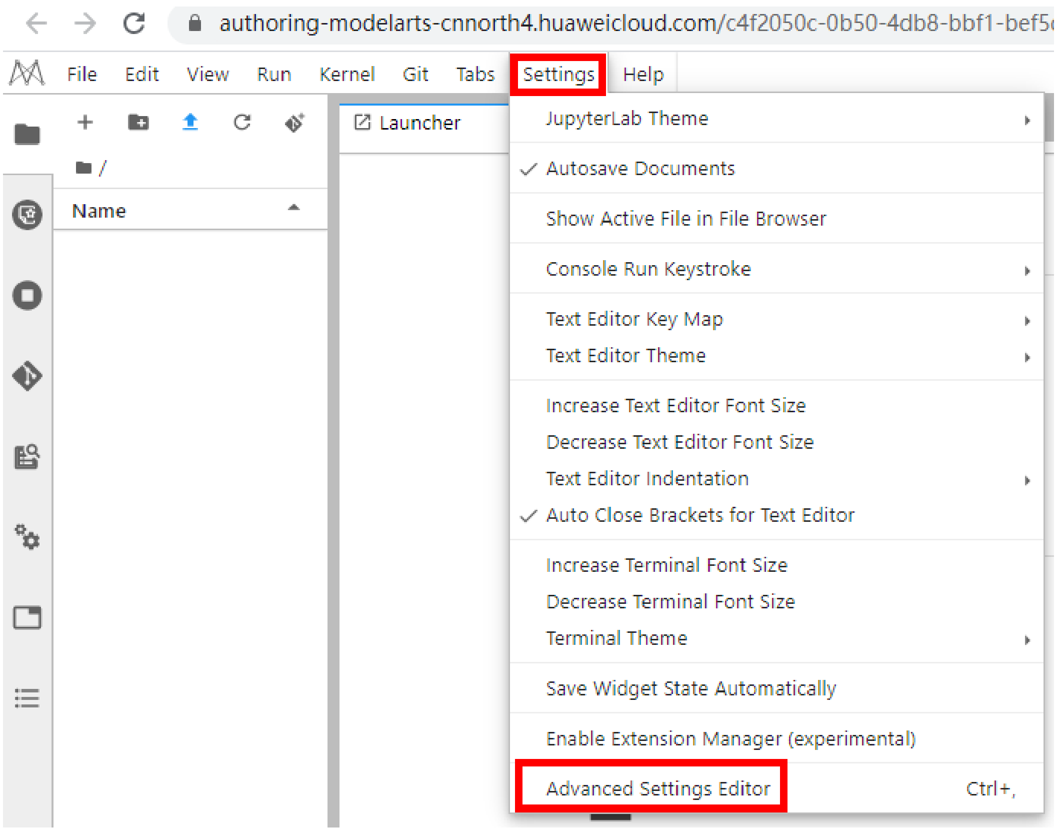
Set **fontFamily** of **Terminal** to **Menlo**.

Procedure

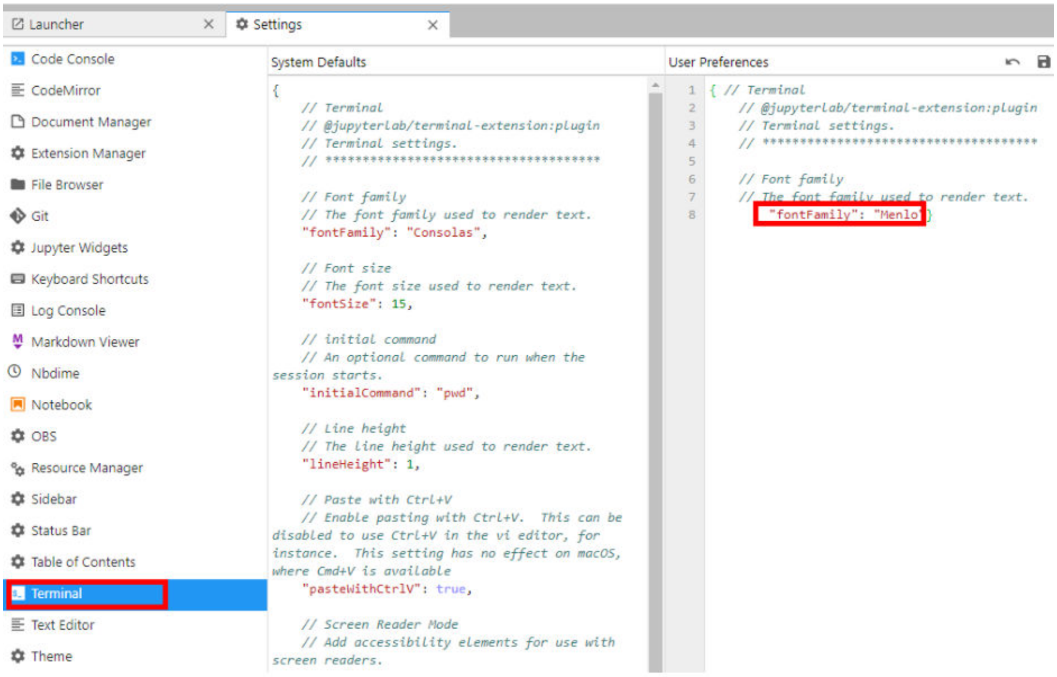
Step 1 Log in to the ModelArts management console and choose **DevEnviron** > **Notebook**.

Step 2 Locate the row containing the target notebook instance and click **Open** in the **Operation** column. The **JupyterLab** page is displayed.

Step 3 On the **JupyterLab** page, choose **Settings** > **Advanced Settings Editor**. The **Settings** tab page is displayed.


authoring-modelarts-cnnorth4.huaweicloud.com/c4f2050c-0b50-4db8-bbf1-bef5
File
Edit
View
Run
Kernel
Git
Tabs
Settings
Help
Launcher
/
Name
JupyterLab Theme
Autosave Documents
Show Active File in File Browser
Console Run Keystroke
Text Editor Key Map
Text Editor Theme
Increase Text Editor Font Size
Decrease Text Editor Font Size
Text Editor Indentation
Auto Close Brackets for Text Editor
Increase Terminal Font Size
Decrease Terminal Font Size
Terminal Theme
Save Widget State Automatically
Enable Extension Manager (experimental)
Advanced Settings Editor
Ctrl+,

Step 4 Choose **Terminal** in the navigation pane on the left and set **fontFamily** to **Menlo**.



System Defaults

```
{
    // Terminal
    // @jupyterlab/terminal-extension:plugin
    // Terminal settings.
    // ************************************

    // Font family
    // The font family used to render text.
    "fontFamily": "Consolas",

    // Font size
    // The font size used to render text.
    "fontSize": 15,

    // initial command
    // An optional command to run when the
session starts.
    "initialCommand": "pwd",

    // Line height
    // The line height used to render text.
    "lineHeight": 1,

    // Paste with Ctrl+V
    // Enable pasting with Ctrl+V.  This can be
disabled to use Ctrl+V in the vi editor, for
instance.  This setting has no effect on macOS,
where Cmd+V is available
    "pasteWithCtrlV": true,

    // Screen Reader Mode
    // Add accessibility elements for use with
screen readers.
```

User Preferences

```
{ // Terminal
    // @jupyterlab/terminal-extension:plugin
    // Terminal settings.
    // ************************************

    // Font family
    // The font family used to render text.
      "fontFamily": "Menlo"}
```

----End

5.4.6 Is There a Proxy for Notebook? How Do I Disable It?

There is a proxy for Notebook.

Run the **env|grep proxy** command to obtain the notebook proxy.

Run the **unset https_proxy unset http_proxy** command to disable the proxy.

5.5 Notebook Instances

5.5.1 What Do I Do If I Cannot Access My Notebook Instance?

Troubleshoot the issue based on error code.

A Black Screen Is Displayed When a Notebook Instance Is Opened

A black screen is displayed after a notebook instance is opened, which is caused by a proxy issue. Change the proxy to rectify the fault.

A Blank Page Is Displayed When a Notebook Instance Is Opened

- If a blank page is displayed after a notebook instance is opened, clear the browser cache and open the notebook instance again.
- Check whether the ad filtering component is installed for the browser. If yes, disable the component.

Error 404

If this error is reported when an IAM user creates an instance, the IAM user does not have the permissions to access the corresponding storage location (OBS bucket).

Solution

1. Log in to the OBS console using the primary account and grant access permissions for the OBS bucket to the IAM user. For details about the operation, see **Principal**.
2. After the IAM user obtains the permissions, log in to the ModelArts console, delete the instance, and use the OBS path to create a notebook instance.

Error 503

If this error is reported, it is possible that the instance is consuming too many resources. If this is the case, stop the instance and restart it.

Error 504

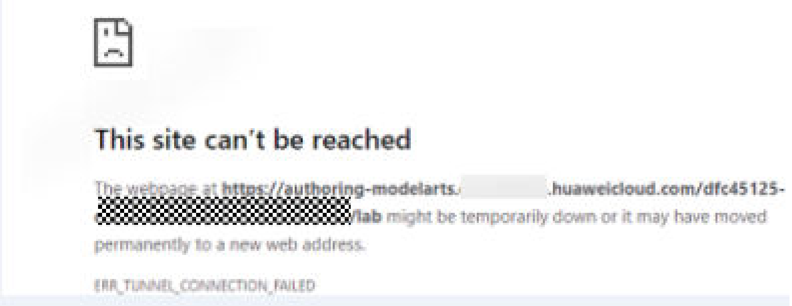
If this error is reported, **submit a service ticket** or contact customer service.

Error 500

Notebook JupyterLab cannot be opened, and error 500 is reported. The possible cause is that the disk space in the **work** directory is used up. In this case, identify the fault cause and clear the disk by referring to **Disk Space Used Up**.

Error "This site can't be reached"

After a notebook instance is created, click **Open** in the **Operation** column. The error message shown in the following figure is displayed.



This site can't be reached

The webpage at **https://authoring-modelarts.[redacted].huaweicloud.com/dfc45125-[redacted]/lab** might be temporarily down or it may have moved permanently to a new web address.

ERR_TUNNEL_CONNECTION_FAILED

Do as follows to resolve this issue: Copy the domain name of the page , add it to the **Do not use proxy server for addresses beginning with** text box, and save the settings.

5.5.2 What Should I Do When the System Displays an Error Message Indicating that No Space Left After I Run the pip install Command?

Symptom

In the notebook instance, error message "No Space left..." is displayed after the **pip install** command is run.

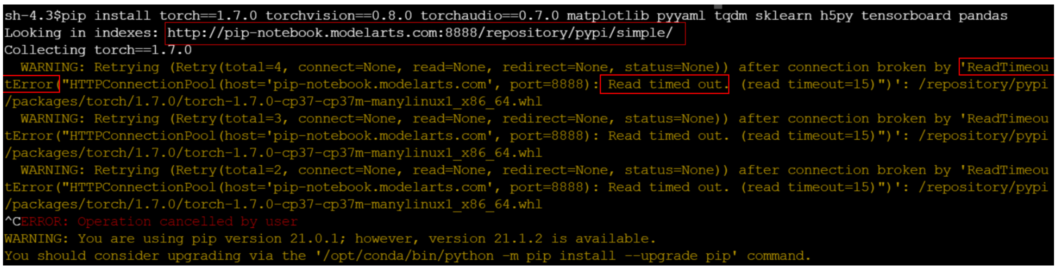
Solution

You are advised to run the **pip install --no-cache **** command instead of the **pip install **** command. Adding the **--no-cache** parameter can solve such problem.

5.5.3 What Do I Do If "Read timed out" Is Displayed After I Run pip install?

Symptom

After I run **pip install** in a notebook instance, the system displays error message "ReadTimeoutError..." or "Read timed out...".



```
sh-4.3$pip install torch==1.7.0 torchvision==0.8.0 torchaudio==0.7.0 matplotlib pyyaml tqdm sklearn h5py tensorboard pandas
Looking in indexes: http://pip-notebook.modelarts.com:8888/repository/pypi/simple/
Collecting torch==1.7.0
  WARNING: Retrying (Retry(total=4, connect=None, read=None, redirect=None, status=None)) after connection broken by 'ReadTimeou
tError("HTTPConnectionPool(host='pip-notebook.modelarts.com', port=8888): Read timed out. (read timeout=15)")': /repository/pypi
/packages/torch/1.7.0/torch-1.7.0-cp37-cp37m-manylinux1_x86_64.whl
  WARNING: Retrying (Retry(total=3, connect=None, read=None, redirect=None, status=None)) after connection broken by 'ReadTimeou
tError("HTTPConnectionPool(host='pip-notebook.modelarts.com', port=8888): Read timed out. (read timeout=15)")': /repository/pypi
/packages/torch/1.7.0/torch-1.7.0-cp37-cp37m-manylinux1_x86_64.whl
  WARNING: Retrying (Retry(total=2, connect=None, read=None, redirect=None, status=None)) after connection broken by 'ReadTimeou
tError("HTTPConnectionPool(host='pip-notebook.modelarts.com', port=8888): Read timed out. (read timeout=15)")': /repository/pypi
/packages/torch/1.7.0/torch-1.7.0-cp37-cp37m-manylinux1_x86_64.whl
^CERROR: Operation cancelled by user
WARNING: You are using pip version 21.0.1; however, version 21.1.2 is available.
You should consider upgrading via the '/opt/conda/bin/python -m pip install --upgrade pip' command.
```

Solution

Run **pip install --upgrade pip** and then **pip install**.

5.5.4 What Do I Do If the Code Can Be Run But Cannot Be Saved, and the Error Message "save error" Is Displayed?

If the notebook instance can run the code but cannot save it, the error message "save error" is displayed when you save the file. In most cases, this error is caused by a security policy of Web Application Firewall (WAF).

On the current page, some characters in your input or output of the code are intercepted because they are considered to be a security risk. Submit a service ticket and contact customer service to check and handle the problem.

5.5.5 Why Is a Request Timeout Error Reported When I Click the Open Button of a Notebook Instance?

When a notebook container breaks down due to memory overflow or other reasons, if you click the **Open** button of the notebook instance, a request timeout error is displayed.

In this case, wait for about half a minute or so until the container is restored, and then click **Open** again.

5.5.6 When the SSH Tool Is Used to Connect to a Notebook Instance, Server Processes Are Cleared, but the GPU Usage Is Still 100%

This fault occurs because code execution is suspended and the GPU memory is not released. Alternatively, the program is cleared due to memory overflow during code execution. In this case, you need to release the GPU memory and restart the instance. To avoid unsaved code caused by the end of processes, you are advised to periodically save the code to an OBS bucket or the **./work** directory of the container.

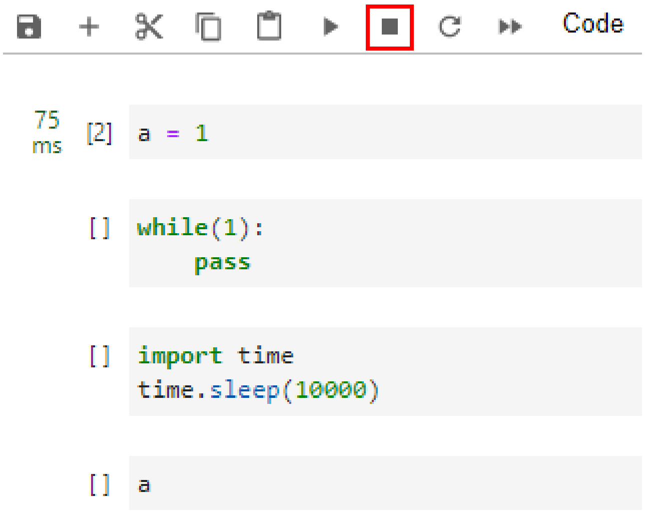
5.6 Code Execution

5.6.1 What Do I Do If a Notebook Instance Won't Run My Code?

If a notebook instance fails to execute code, you can locate and rectify the fault as follows:

1. If the execution of a cell is suspended or lasts for a long time (for example, the execution of the second and third cells in **Figure 5-7** is suspended or lasts for a long time, causing execution failure of the fourth cell) but the notebook page still responds and other cells can be selected, click **interrupt the kernel** highlighted in a red box in the following figure to stop the execution of all cells. The notebook instance retains all variable spaces.

Figure 5-7 Stopping all cells



```
[2] a = 1
```

```
[ ] while(1):
        pass
```

```
[ ] import time
    time.sleep(10000)
```

```
[ ] a
```

2. If the notebook page does not respond, close the notebook page and the ModelArts console. Then, open the ModelArts console and access the notebook instance again. The notebook instance retains all the variable spaces that exist when the notebook instance is unavailable.
3. If the notebook instance still cannot be used, access the **Notebook** page on the ModelArts console and stop the notebook instance. After the notebook instance is stopped, click **Start** to restart the notebook instance and open it. The instance will have preserved all the spaces for the variables that were unable to run.

5.6.2 Why Does the Instance Break Down When dead kernel Is Displayed During Training Code Running?

The notebook instance breaks down during training code running due to insufficient memory caused by large data volume or excessive training layers.

After this error occurs, the system automatically restarts the notebook instance to fix the instance breakdown. In this case, only the breakdown is fixed. If you run the training code again, the failure will still occur. To solve the problem of insufficient memory, you are advised to create a new notebook instance and use a resource pool of higher specifications, such as a GPU or dedicated resource pool, to run the training code. An existing notebook instance that has been successfully created cannot be scaled up using resources with higher specifications.

5.6.3 What Do I Do If cudaCheckError Occurs During Training?

Symptom

The following error occurs when the training code is executed in a notebook:

```
cudaCheckError() failed : no kernel image is available for execution on the device
```

Possible Cause

Parameters **arch** and **code** in **setup.py** have not been set to match the GPU compute power.

Solution

For Tesla V100 GPUs, the GPU compute power is **-gencode arch=compute_70,code=[sm_70,compute_70]**. Set the compilation parameters in **setup.py** accordingly.

5.6.4 What Should I Do If DevEnviron Prompts Insufficient Space?

If space is insufficient, use notebook instances of the EVS type.

Upload code and data to an OBS bucket for the original notebook instance by referring to **How Do I Upload a File from a Notebook Instance to OBS or Download a File from OBS to a Notebook Instance?**. Then, create a notebook instance of the EVS type, and download files from OBS to the new notebook instance.

5.6.5 Why Does the Notebook Instance Break Down When opencv.imshow Is Used?

Symptom

When opencv.imshow is used in a notebook instance, the notebook instance breaks down.

Possible Causes

The cv2.imshow function in OpenCV malfunctions in a client/server environment such as Jupyter. However, Matplotlib does not have this problem.

Solution

Display images by referring to the following example. Note that OpenCV displays BGR images while Matplotlib displays RGB images.

Python:

```
from matplotlib import pyplot as plt
import cv2
img = cv2.imread('Image path')
plt.imshow(cv2.cvtColor(img, cv2.COLOR_BGR2RGB))
plt.title('my picture')
plt.show()
```

5.6.6 Why Cannot the Path of a Text File Generated in Windows OS Be Found In a Notebook Instance?

Symptom

When a text file generated in Windows is used in a notebook instance, the text content cannot be read and an error message may be displayed indicating that the path cannot be found.

Possible Causes

The notebook instance runs Linux and its line feed format (CRLF) differs from that (LF) in Windows.

Solution

Convert the file format to Linux in your notebook instance.

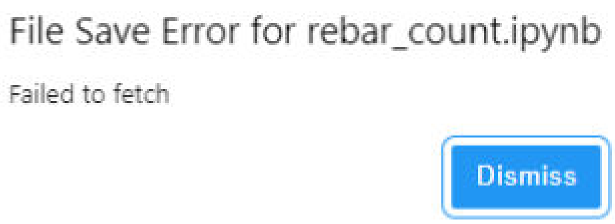
Shell:

```
dos2unix  File name
```

5.6.7 What Do I Do If Files Fail to Be Saved in JupyterLab?

Symptom

When a file is saved in JupyterLab, an error message is displayed.



File Save Error for rebar_count.ipynb

Failed to fetch

Possible Cause

A third-party plug-in has been installed on the browser, and the proxy intercepts the request. As a result, the file cannot be saved.

Solution

Disable the plug-in and save file again.

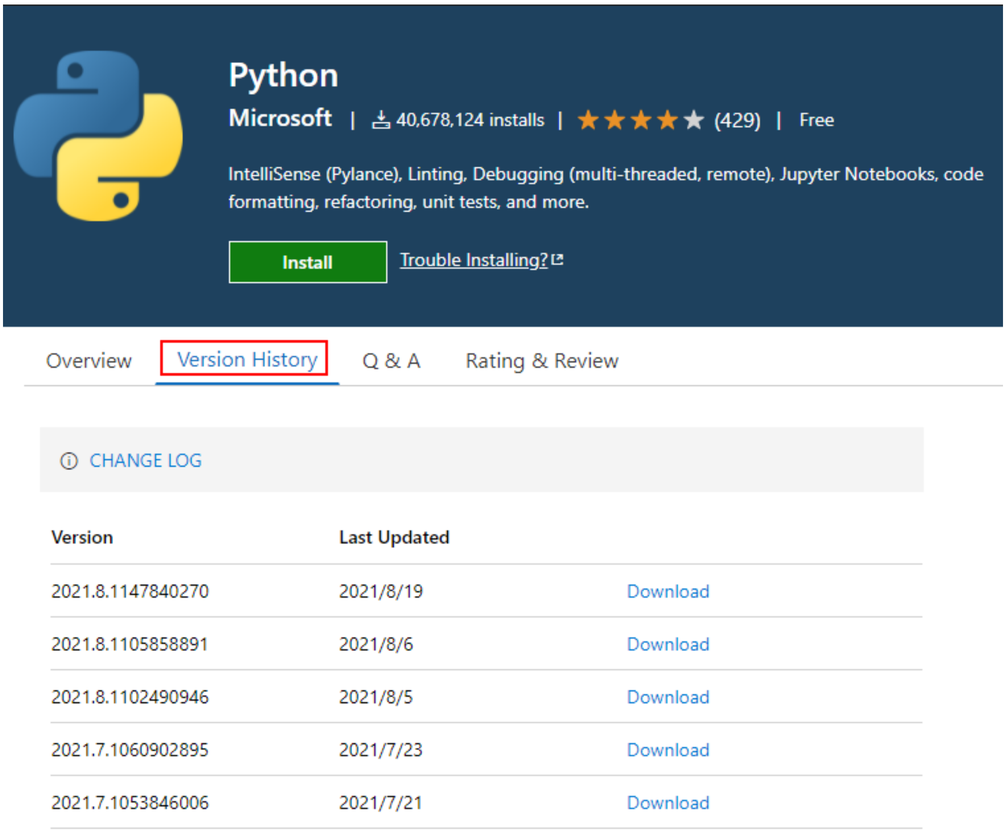
5.7 VS Code

5.7.1 What Do I Do If Installing a Remote Plug-in Failed?

Method 1 (recommended): Using an offline package

1. Log in at the **official VS Code website** and search for the target Python plug-in.
2. Click the **Version History** tab of the plug-in and download the offline installation package.

Figure 5-8 Offline installation package of the Python plug-in



IntelliSense (Pylance), Linting, Debugging (multi-threaded, remote), Jupyter Notebooks, code formatting, refactoring, unit tests, and more.

Version	Last Updated	
2021.8.1147840270	2021/8/19	Download
2021.8.1105858891	2021/8/6	Download
2021.8.1102490946	2021/8/5	Download
2021.7.1060902895	2021/7/23	Download
2021.7.1053846006	2021/7/21	Download

3. In local VS Code, drag the downloaded VSIX file to the remote notebook.
4. Right-click the file and choose **Install Extension VSIX** from the shortcut menu.

Method 2: Setting the default remote plug-in

Set the default remote plug-in in VS Code by following the instructions provided in **How Can I Set the Default Remote Plug-in in VS Code?** This enables automatic plug-in installation when the plug-in is connected.

Method 3: Taking measures provided at **official VS Code website**

Tips (adjust parameter settings as needed):

```
 "remote.SSH.connectTimeout": 10,
 "remote.SSH.maxReconnectionAttempts": null,
 "remote.downloadExtensionsLocally": true,
 "remote.SSH.useLocalServer": false,
"remote.SSH.localServerDownload": "always",
```

5.7.2 What Do I Do If a Restarted Notebook Instance Can Be Connected Only After I Locally Delete known_hosts?

To resolve this issue, set notebook parameters **StrictHostKeyChecking no** and **UserKnownHostsFile=/dev/null** in the local **ssh config** file.

```
Host roma-local-cpu
  HostName x.x.x.x # IP address
  Port 22522
  User ma-user
```

```
IdentityFile C:/Users/my.pem
StrictHostKeyChecking no
ForwardAgent yes
```

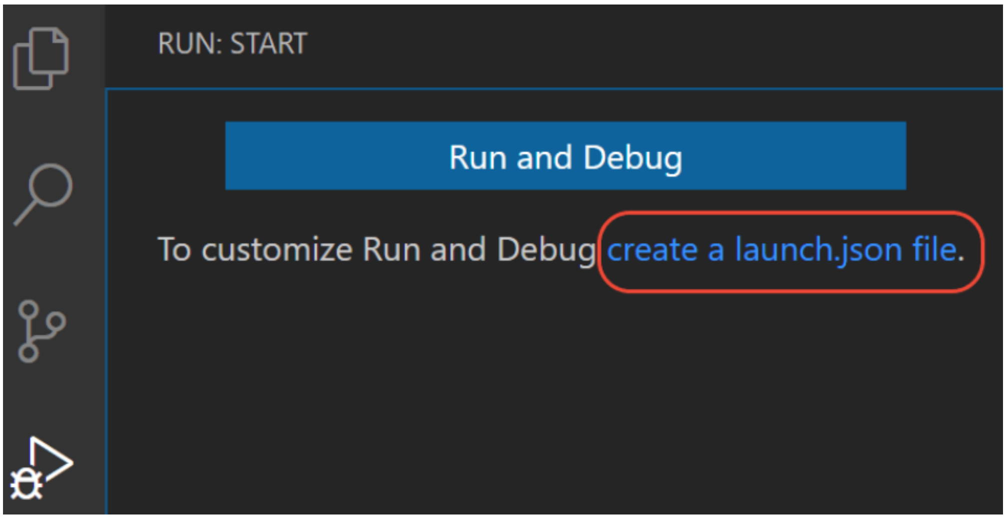
Note: SSH logins are insecure because the **known_hosts** file will be ignored during the logins.

5.7.3 What Do I Do If the Source Code Cannot Be Accessed When I Use VS Code for Debugging?

If the **launch.json** file already exists, go to step 3.

Step 1: Open launch.json.

- Method 1: Click **Run** (**Ctrl+Shift+D**) in the menu bar on the left and click **create a launch.json file**.

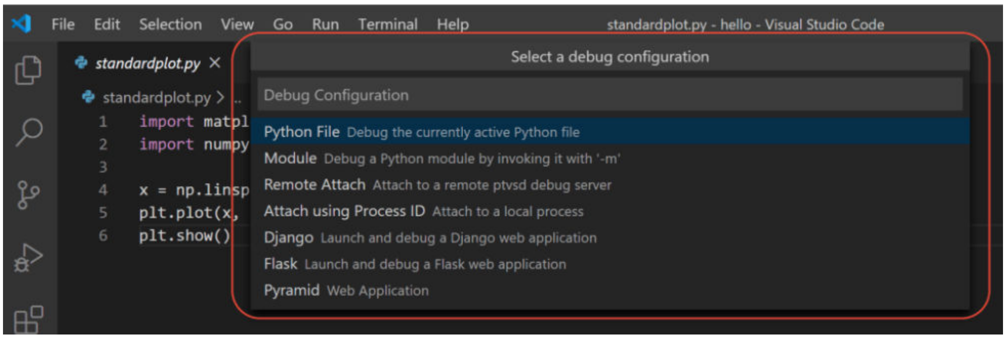


To customize Run and Debug create a launch.json file.

- Method 2: In the menu bar, choose **Run** > **Open configurations**.

Step 2: Select a language.

To set a Python language, select **Python File** in **Select a debug configuration**. The operations for setting other languages are similar.

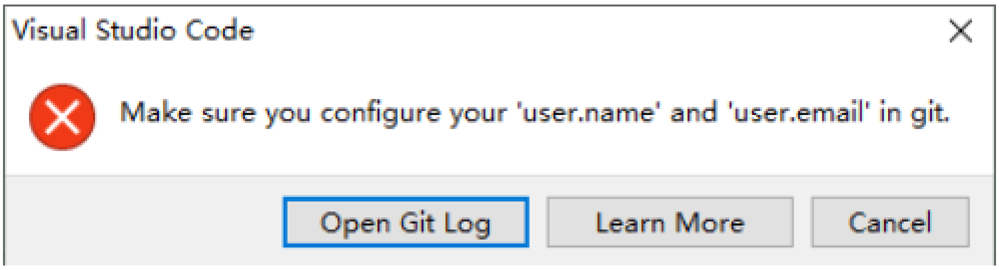

File
Edit
Selection
View
Go
Run
Terminal
Help
standardplot.py - hello - Visual Studio Code
standardplot.py
standardplot.py
Select a debug configuration
Debug Configuration
Python File Debug the currently active Python file
Module Debug a Python module by invoking it with '-m'
Remote Attach Attach to a remote ptvsd debug server
Attach using Process ID Attach to a local process
Django Launch and debug a Django web application
Flask Launch and debug a Flask web application
Pyramid Web Application

```
import matpl
import numpy

x = np.linsp
plt.plot(x,
plt.show()
```

Step 3: Set justMyCode to False in launch.json.

```
{
    // Use IntelliSense to learn about possible attributes.
    // Hover to view descriptions of existing attributes.
    // For more information, visit: https://go.microsoft.com/fwlink/?linkid=830387
    "version": "0.2.0",
    "configurations": [
            {
                "name": "Python: Current file",
                "type": "python",
                "request": "launch",
                "program": "${file}",
                "console": "integratedTerminal",
                "justMyCode": false
            }
        ]
    }
```

5.7.4 What Do I Do If a Message Is Displayed Indicating an Incorrect Username or Email Address When I Use VS Code to Submit Code?



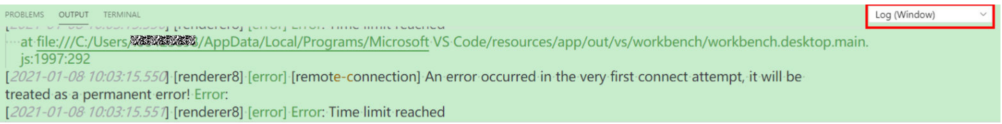
Make sure you configure your 'user.name' and 'user.email' in git.

1. In VS Code, press **Ctrl+Shift+P**.
2. Search for **Python: Select Interpreter** and select the target Python.
3. Choose **Terminal** > **New Terminal**. The CLI of the remote container is displayed.
4. On the VS Code terminal, run the following commands and submit the code again:
   ```
   git config --global user.email xxx@xxx.com # Change the email address to yours.
   git config --global user.name xxx # Change the username to yours.
   ```

5.7.5 How Can I View Remote Logs in VS Code?

1. In VS Code, press **Ctrl+Shift+P**.
2. Search for **show logs**.
3. Choose **Remote Server**.

Alternatively, switch logs in the red box shown in the following figure.



```
    at file:///C:/Users/[redacted]/AppData/Local/Programs/Microsoft VS Code/resources/app/out/vs/workbench/workbench.desktop.main.
    js:1997:292
[2021-01-08 10:03:15.550] [renderer8] [error] [remote-connection] An error occurred in the very first connect attempt, it will be
treated as a permanent error! Error:
[2021-01-08 10:03:15.551] [renderer8] [error] Error: Time limit reached
```

5.7.6 How Can I Open the VS Code Configuration File settings.json?

1. In VS Code, press **Ctrl+Shift+P**.
2. Search for **Open Settings (JSON)**.

5.7.7 How Can I Change the VS Code Background Color to Light Green?

Add the following settings to the VS Code configuration file **settings.json**:

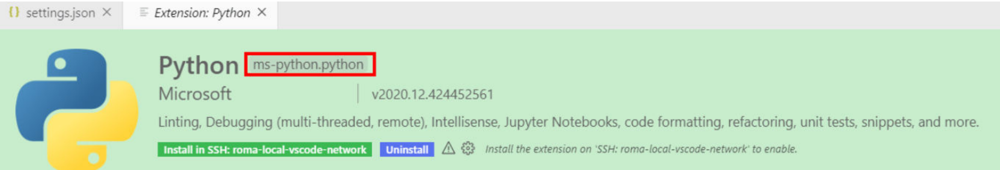
```
"workbench.colorTheme": "Atom One Light",
"workbench.colorCustomizations": {
"[Atom One Light]": {
 "editor.background": "#C7EDCC",
"sideBar.background": "#e7f0e7",
"activityBar.background": "#C7EDCC",
     },
},
```

5.7.8 How Can I Set the Default Remote Plug-in in VS Code?

Add **remote.SSH.defaultExtensions**, for example, for automatically installing Python and Maven plug-ins, to the VS Code configuration file **settings.json**.

```
"remote.SSH.defaultExtensions": [
    "ms-python.python",
    "vscjava.vscode-maven"
  ],
```

To obtain a plug-in name, click the plug-in in VS Code.



Linting, Debugging (multi-threaded, remote), Intellisense, Jupyter Notebooks, code formatting, refactoring, unit tests, snippets, and more.

Install the extension on 'SSH: roma-local-vscode-network' to enable.

5.7.9 How Can I Install a Local Plug-in on the Remote End or a Remote Plug-in on the Local End Through VS Code?

1. In VS Code, press **Ctrl+Shift+P**.
2. Search for **install local** and select the plug-in as required.

5.8 Failures to Access the Development Environment Through VS Code

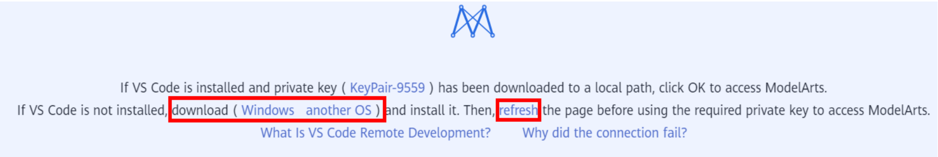
5.8.1 What Do I Do If the VS Code Window Is Not Displayed?

Possible Cause

VS Code is not installed or the installed version is outdated.

Solution

Download and install VS Code. (Windows users click **Windows**. Users of other operating systems click **another OS**.) After the installation, click **refresh** to complete the connection.



If VS Code is installed and private key (KeyPair-9559) has been downloaded to a local path, click OK to access ModelArts.
If VS Code is not installed, download (Windows another OS) and install it. Then, refresh the page before using the required private key to access ModelArts.
What Is VS Code Remote Development? Why did the connection fail?

5.8.2 What Do I Do If a Remote Connection Failed After VS Code Is Opened?

NOTICE

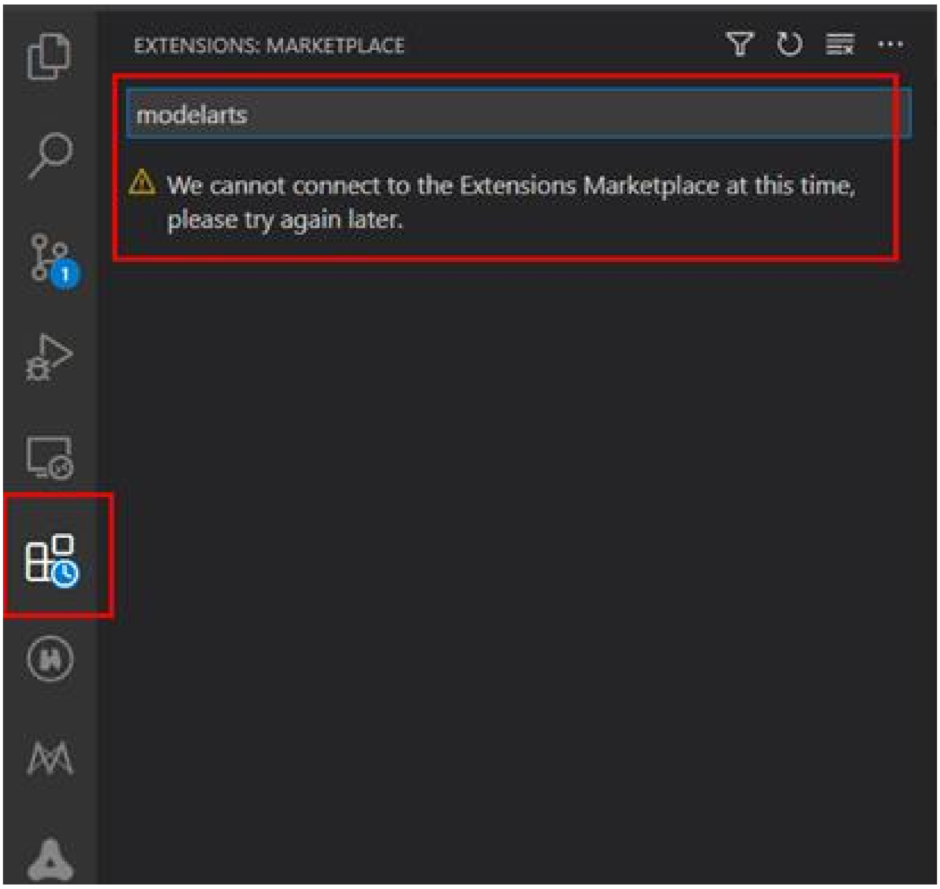
If your local PC runs Linux, see possible cause 2.

Possible Cause 1

Automatically installing the VS Code plug-in ModelArts-HuaweiCloud failed.

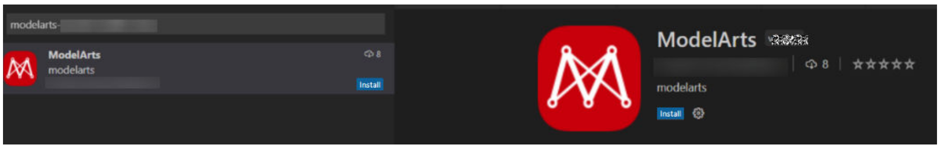
Solution

Method 1: Verify that the VS Code network is accessible. Search for **ModelArts-HuaweiCloud** in the VS Code marketplace. If the following information is displayed, a network error occurred. In this case, switch to another proxy or use another network.

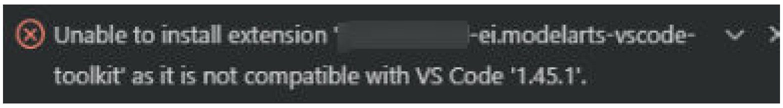


⚠ We cannot connect to the Extensions Marketplace at this time, please try again later.

Search for **ModelArts-HuaweiCloud** again. If the following information is displayed, the network is normal. Then, switch back to the ModelArts management console and try to access VS Code again.



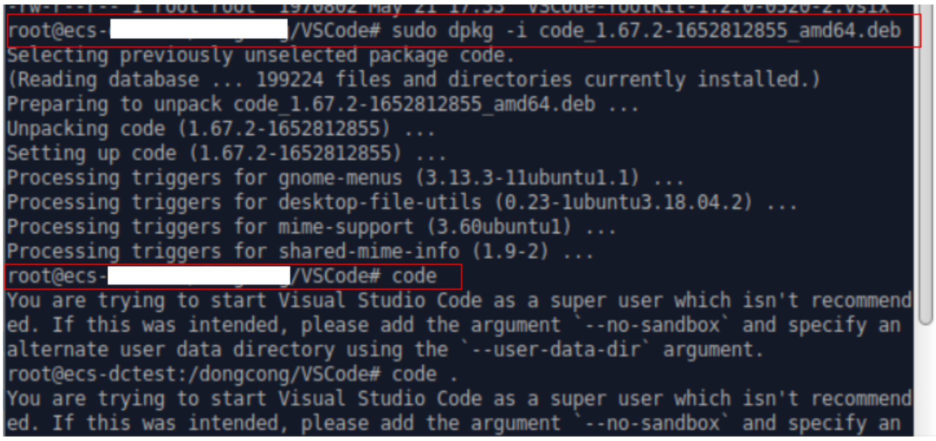
Method 2: If the error message shown in the following figure is displayed, the VS Code version is outdated. Upgrade the VS Code to 1.57.1 or the latest version.



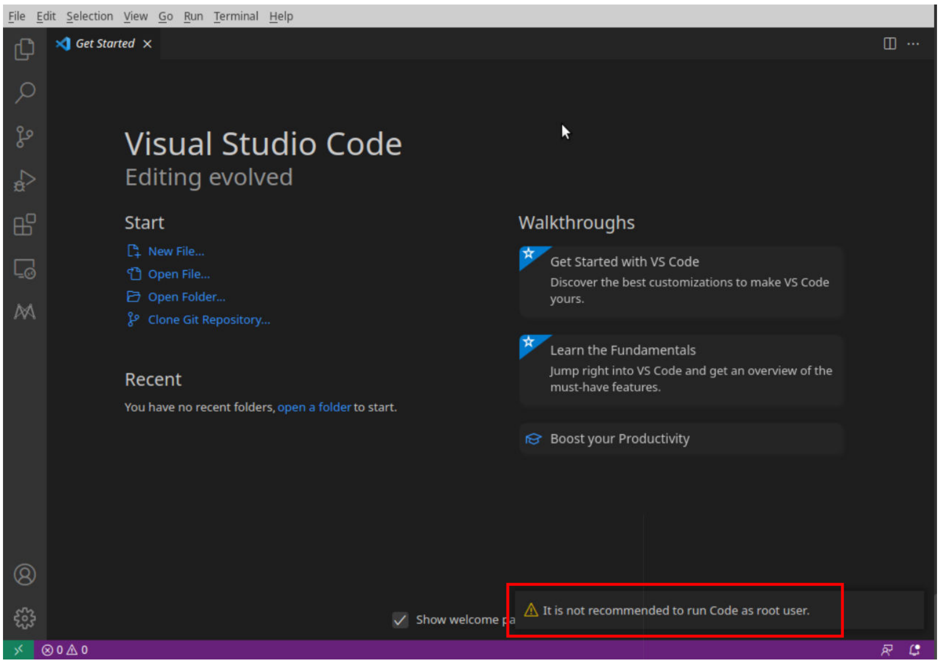
Unable to install extension '[redacted]-ei.modelarts-vscode-toolkit' as it is not compatible with VS Code '1.45.1'.

Possible Cause 2

The local PC runs Linux, and VS Code is installed as user **root**. When you access VS Code, the information "It is not recommended to run Code as root user" is displayed.



```
-rw-r--r-- 1 root root  1970802 May 21 17:33  vscode-toolkit-1.2.0-0520-2.vsix
root@ecs-[redacted]/VSCode# sudo dpkg -i code_1.67.2-1652812855_amd64.deb
Selecting previously unselected package code.
(Reading database ... 199224 files and directories currently installed.)
Preparing to unpack code_1.67.2-1652812855_amd64.deb ...
Unpacking code (1.67.2-1652812855) ...
Setting up code (1.67.2-1652812855) ...
Processing triggers for gnome-menus (3.13.3-11ubuntu1.1) ...
Processing triggers for desktop-file-utils (0.23-1ubuntu3.18.04.2) ...
Processing triggers for mime-support (3.60ubuntu1) ...
Processing triggers for shared-mime-info (1.9-2) ...
root@ecs-[redacted]/VSCode# code
You are trying to start Visual Studio Code as a super user which isn't recommend
ed. If this was intended, please add the argument `--no-sandbox` and specify an
alternate user data directory using the `--user-data-dir` argument.
root@ecs-dctest:/dongcong/VSCode# code .
You are trying to start Visual Studio Code as a super user which isn't recommend
ed. If this was intended, please add the argument `--no-sandbox` and specify an
```



Discover the best customizations to make VS Code yours.

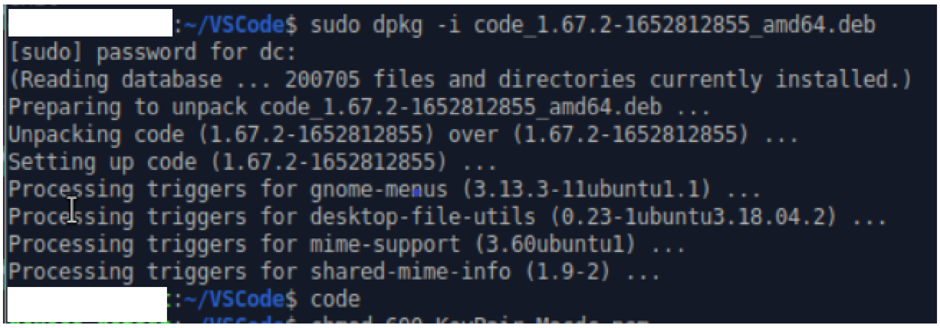
Jump right into VS Code and get an overview of the must-have features.

You have no recent folders, open a folder to start.

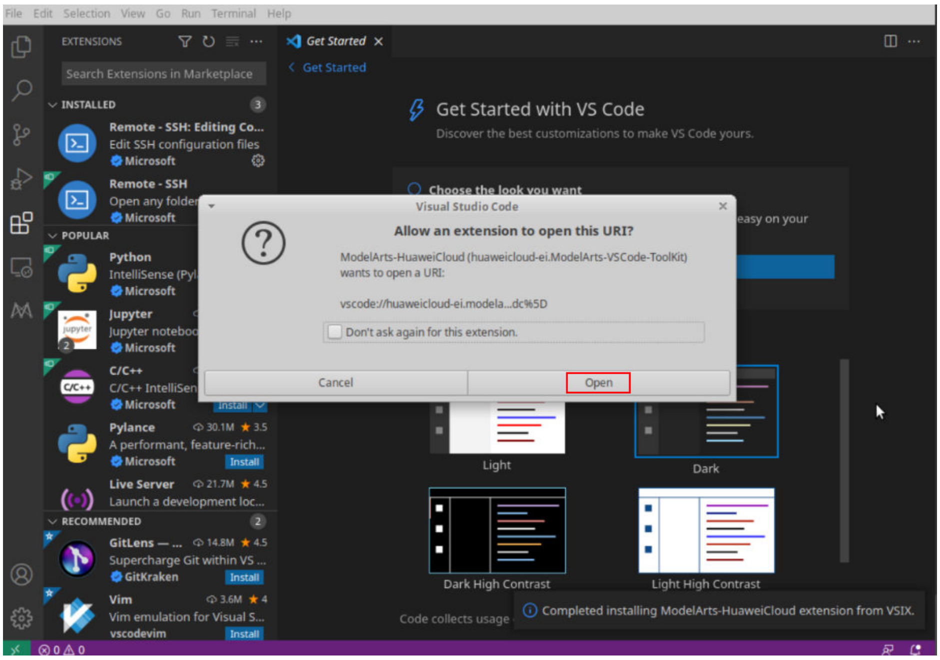
It is not recommended to run Code as root user.

Solution

Install VS Code as a non-**root** user, return to the ModelArts management console, and click **Access VS Code**.



```
[redacted]:~/VSCode$ sudo dpkg -i code_1.67.2-1652812855_amd64.deb
[sudo] password for dc:
(Reading database ... 200705 files and directories currently installed.)
Preparing to unpack code_1.67.2-1652812855_amd64.deb ...
Unpacking code (1.67.2-1652812855) over (1.67.2-1652812855) ...
Setting up code (1.67.2-1652812855) ...
Processing triggers for gnome-menus (3.13.3-11ubuntu1.1) ...
Processing triggers for desktop-file-utils (0.23-1ubuntu3.18.04.2) ...
Processing triggers for mime-support (3.60ubuntu1) ...
Processing triggers for shared-mime-info (1.9-2) ...
[redacted]:~/VSCode$ code
```



Visual Studio Code

Allow an extension to open this URI?

ModelArts-HuaweiCloud (huaweicloud-ei.ModelArts-VSCode-ToolKit) wants to open a URI:

vscode://huaweicloud-ei.modela...dc%5D

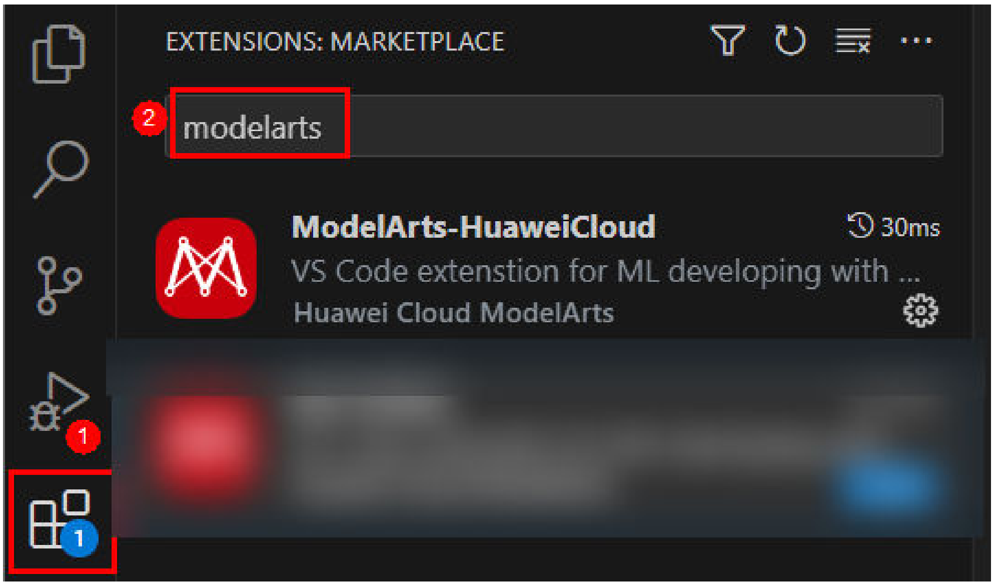
□ Don't ask again for this extension.

Completed installing ModelArts-HuaweiCloud extension from VSIX.

5.8.3 Basic Problems Causing the Failures to Access the Development Environment Through VS Code

If the VS Code fails to connect to the development environment, perform the following steps:

Step 1 Check whether the plug-in package is of the latest version. Search for the plug-in in extensions and check whether it needs to be upgraded.



ModelArts-HuaweiCloud

VS Code extenstion for ML developing with ...

Huawei Cloud ModelArts

Step 2 Check whether the instance is running. If yes, go to the next step.

Step 3 Run the following command in VS Code's Terminal to connect to the remote development environment:

```
ssh -tt -o StrictHostKeyChecking=no -i ${IdentityFile} ${User}@${HostName} -p ${Port}
```

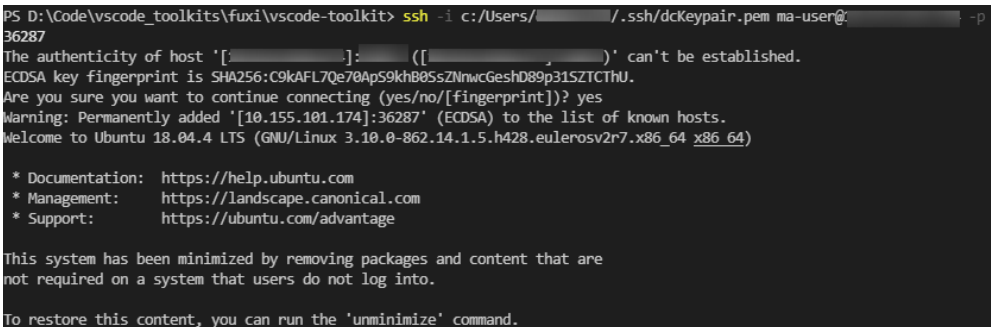
Parameters:

- **IdentityFile**: path to the local key

- **User**: username, for example, **ma-user**

- **HostName**: IP address

- **Port**: port number

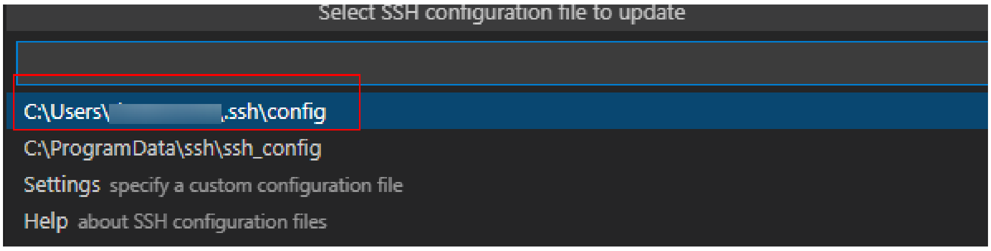


```
PS D:\Code\vscode_toolkits\fuxi\vscode-toolkit> ssh -i c:/Users/[redacted]/.ssh/dcKeypair.pem ma-user@[redacted] -p
36287
The authenticity of host '[[redacted]]:[redacted] ([[redacted])' can't be established.
ECDSA key fingerprint is SHA256:C9kAFL7Qe70ApS9khB0SsZNnwcGeshD89p31SZTCThU.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added '[10.155.101.174]:36287' (ECDSA) to the list of known hosts.
Welcome to Ubuntu 18.04.4 LTS (GNU/Linux 3.10.0-862.14.1.5.h428.eulerosv2r7.x86_64 x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

This system has been minimized by removing packages and content that are
not required on a system that users do not log into.

To restore this content, you can run the 'unminimize' command.
```

If the connection is successful, go to the next step.

Step 4 Check whether the configuration is correct. If yes, go to the next step.

Check the **config** file.



Select SSH configuration file to update
C:\Users\[redacted].ssh\config
C:\ProgramData\ssh\ssh_config
Settings specify a custom configuration file
Help about SSH configuration files

```
HOST remote-dev
    hostname <instance connection host>
    port <instance connection port>
    user ma-user
    IdentityFile ~/.ssh/test.pem
    StrictHostKeyChecking no
    UserKnownHostsFile /dev/null
    ForwardAgent yes
```

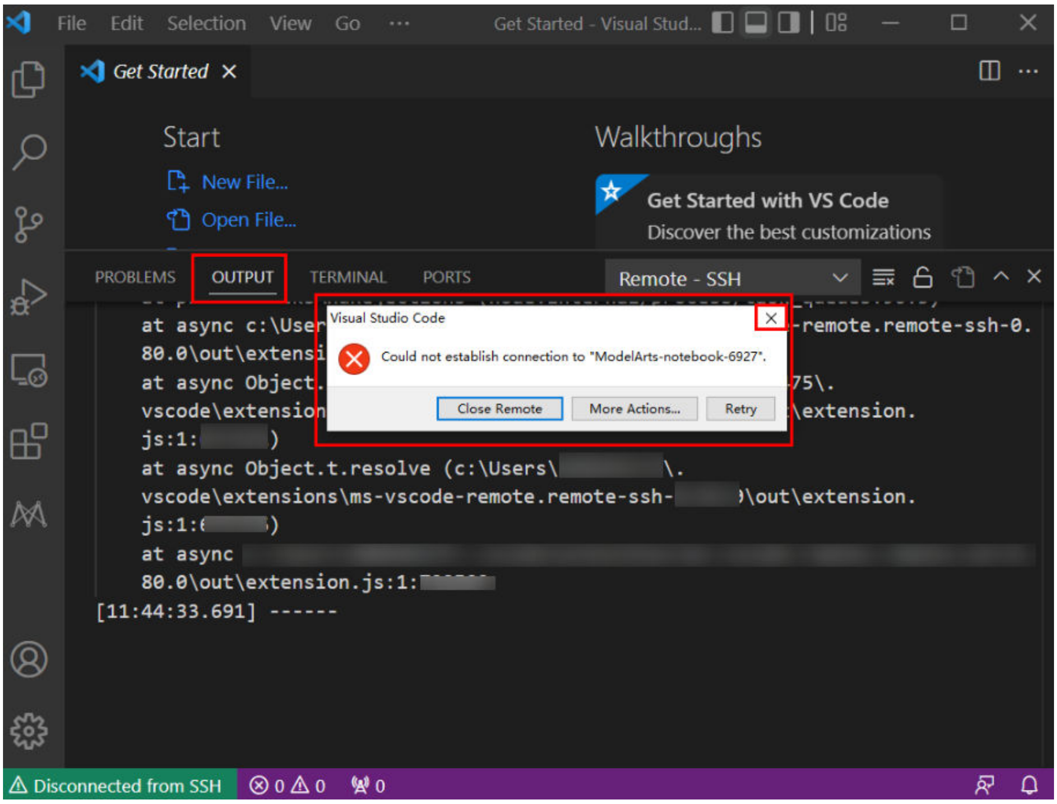
Step 5 Check the key file. You are advised to save the key file in **C:\Users\xx.ssh** and ensure that the file does not contain Chinese characters.

Step 6 If the fault persists, rectify it by referring to the FAQs in **follow-up sections**.

----End

5.8.4 What Do I Do If Error Message "Could not establish connection to xxx" Is Displayed During a Remote Connection?

Symptom



Visual Studio Code

Could not establish connection to "ModelArts-notebook-6927".

```
    at async c:\User                                            -remote.remote-ssh-0.
    80.0\out\extensi
    at async Object.                                            75\.
    vscode\extension                                            \extension.
    js:1:[redacted])
    at async Object.t.resolve (c:\Users\[redacted]\.
    vscode\extensions\ms-vscode-remote.remote-ssh-[redacted]\out\extension.
    js:1:[redacted])
    at async [redacted]
    80.0\out\extension.js:1:[redacted]
[11:44:33.691] ------
```

Possible Cause

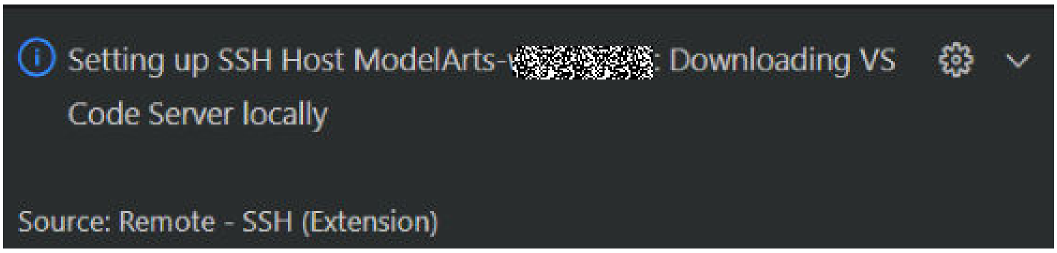
Establishing a remote SSH connection to an instance through VS Code failed.

Solution

Close the displayed dialog box, view the error information in **OUTPUT**, and rectify the fault by referring to the troubleshooting methods provided in the following sections.

5.8.5 What Do I Do If the Connection to a Remote Development Environment Remains in "Setting up SSH Host xxx: Downloading VS Code Server locally" State for More Than 10 Minutes?

Symptom



Setting up SSH Host ModelArts-[redacted]: Downloading VS Code Server locally

Source: Remote - SSH (Extension)

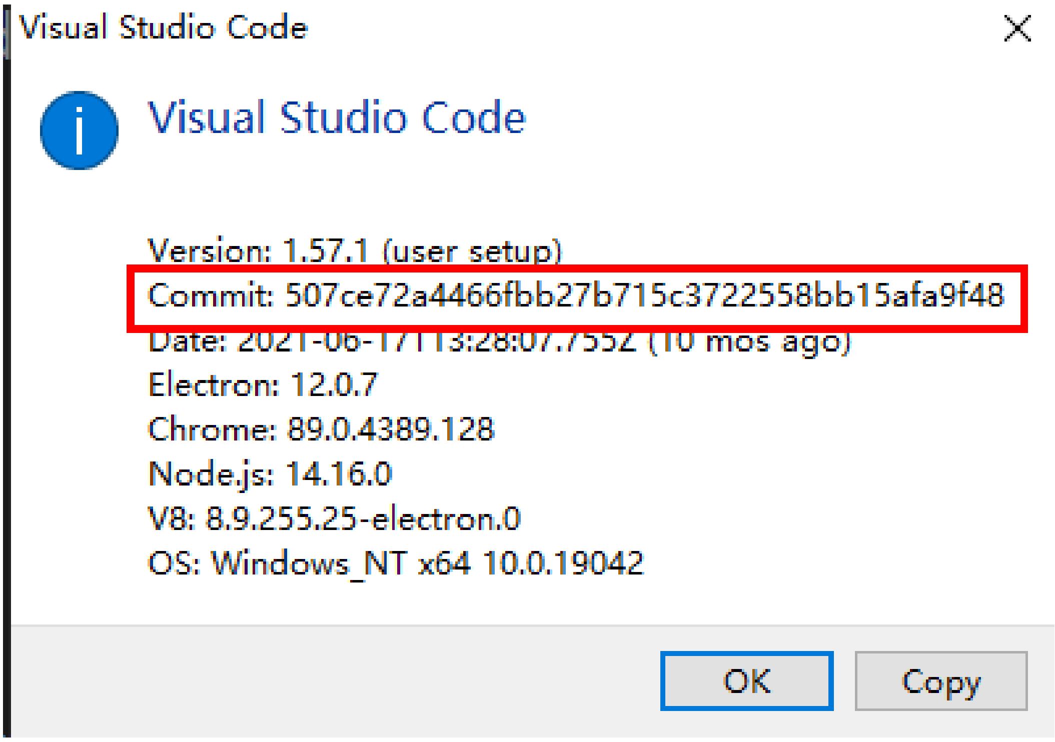
Possible Cause

The local network is faulty. As a result, it takes a long time to automatically install the VS Code server remotely.

Solution

Manually install the VS Code server.

Step 1 Obtain the VS Code commit ID.



Visual Studio Code

Visual Studio Code

Version: 1.57.1 (user setup)
Commit: 507ce72a4466fbb27b715c3722558bb15afa9f48
Date: 2021-06-17T13:28:07.755Z (10 mos ago)
Electron: 12.0.7
Chrome: 89.0.4389.128
Node.js: 14.16.0
V8: 8.9.255.25-electron.0
OS: Windows_NT x64 10.0.19042

Step 2 Download the VS Code server package of the required version. Select Arm or x86 based on the CPU architecture of the development environment.

NOTE

Replace *${commitID}* in the following link with the commit ID obtained in **Step 1**.

- For Arm, download **vscode-server-linux-arm64.tar.gz**.

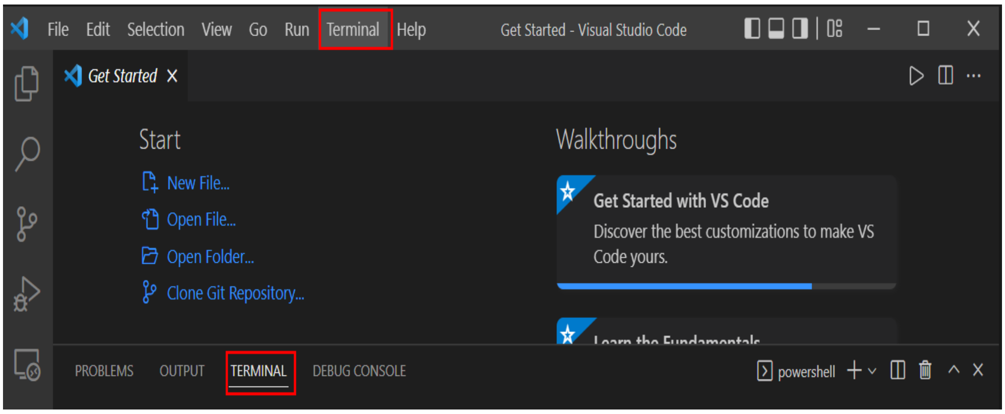
  https://update.code.visualstudio.com/commit:${commitID}/server-linux-arm64/stable

- For x86, download **vscode-server-linux-x64.tar.gz**.

  https://update.code.visualstudio.com/commit:${commitID}/server-linux-x64/stable

Step 3 Access the remote environment.

Switch to **Terminal** in VS Code.



Run the following command in VS Code Terminal to access the remote development environment:

```
ssh -tt -o StrictHostKeyChecking=no -i ${IdentityFile} ${User}@${HostName} -p ${Port}
```

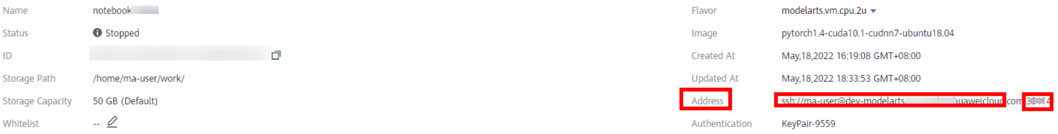
Parameters:

- **IdentityFile**: Path to the local key

- **User**: Username, for example, **ma-user**

- **HostName**: IP address

- **Port**: Port number



Name	notebook[redacted]	Flavor	modelarts.vm.cpu.2u ▾
Status	❶ Stopped	Image	pytorch1.4-cuda10.1-cudnn7-ubuntu18.04
ID	[redacted]	Created At	May,18,2022 16:19:08 GMT+08:00
Storage Path	/home/ma-user/work/	Updated At	May,18,2022 18:33:53 GMT+08:00
Storage Capacity	50 GB (Default)	Address	ssh://ma-user@dev-modelarts[redacted]uaweicloud.com[redacted]4
Whitelist	--	Authentication	KeyPair-9559

Step 4 Manually install the VS Code server.

Run the following commands on the VS Code terminal to clear the residual data (replace *${commitID}* in the commands with the commit ID obtained in **Step 1**):

```
rm -rf /home/ma-user/.vscode-server/bin/${commitID}/*
mkdir -p /home/ma-user/.vscode-server/bin/${commitID}
```

Upload the VS Code server package to the development environment.

```
exit
scp -i xxx.pem -P 31205 Local path to the VS Code server package ma-user@xxx:/home/ma-user/.vscode-server/bin
ssh -tt -o StrictHostKeyChecking=no -i ${IdentityFile} ${User}@${HostName} -p ${Port}
```

Parameters:

- **IdentityFile**: Path to the local key

- **User**: Username, for example, **ma-user**

- **HostName**: IP address

- **Port**: Port number

Take Arm as an example. Decompress the VS Code server package to **$HOME/.vscode-server/bin**. Replace *${commitID}* in the command with the commit ID obtained in **Step 1**.

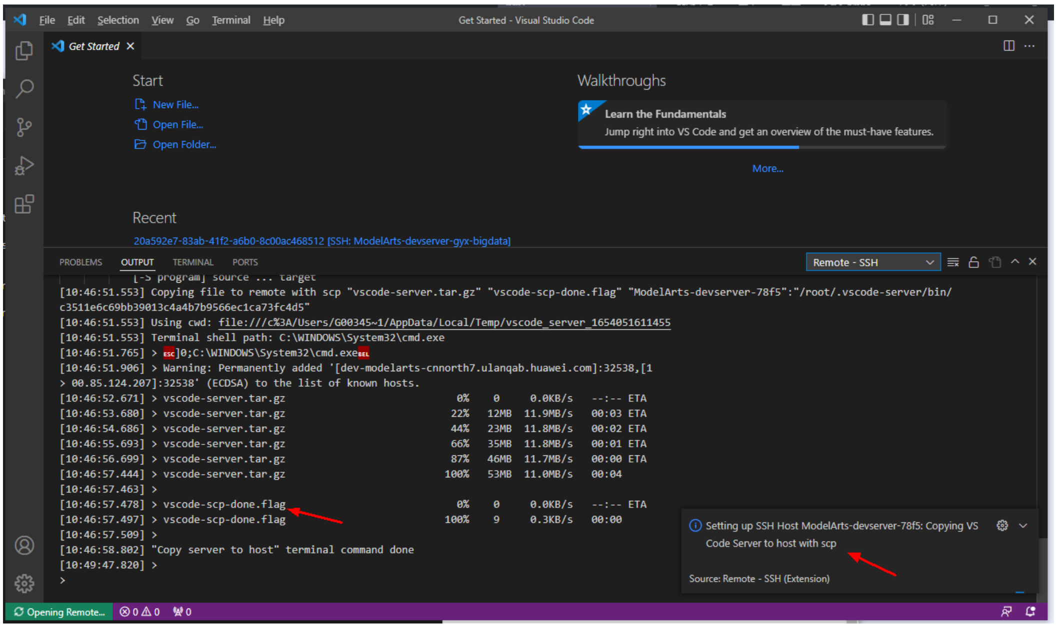
```
cd /home/ma-user/.vscode-server/bin
tar -zxf vscode-server-linux-arm64.tar.gz
mv vscode-server-linux-arm64/* ${commitID}
```

Step 5 Establish the remote connection again.

----End

5.8.6 What Do I Do If the Connection to a Remote Development Environment Remains in the State of "Setting up SSH Host xxx: Downloading VS Code Server locally" for More Than 10 Minutes?

Symptom


File Edit Selection View Go Terminal Help
Get Started - Visual Studio Code
Get Started
Start
New File...
Open File...
Open Folder...
Walkthroughs
Learn the Fundamentals
Jump right into VS Code and get an overview of the must-have features.
More...
Recent
20a592e7-83ab-41f2-a6b0-8c00ac468512 [SSH: ModelArts-devserver-gyx-bigdata]
PROBLEMS
OUTPUT
TERMINAL
PORTS
Remote - SSH
Opening Remote...
0 0 0

```
        [-S program] source ... target
[10:46:51.553] Copying file to remote with scp "vscode-server.tar.gz" "vscode-scp-done.flag" "ModelArts-devserver-78f5":"/root/.vscode-server/bin/
c3511e6c69bb39013c4a4b7b9566ec1ca73fc4d5"
[10:46:51.553] Using cwd: file:///c%3A/Users/G00345~1/AppData/Local/Temp/vscode_server_1654051611455
[10:46:51.553] Terminal shell path: C:\WINDOWS\System32\cmd.exe
[10:46:51.765] > ESC]0;C:\WINDOWS\System32\cmd.exeBEL
[10:46:51.906] > Warning: Permanently added '[dev-modelarts-cnnorth7.ulanqab.huawei.com]:32538,[1
> 00.85.124.207]:32538' (ECDSA) to the list of known hosts.
[10:46:52.671] > vscode-server.tar.gz                        0%    0     0.0KB/s   --:-- ETA
[10:46:53.680] > vscode-server.tar.gz                       22%   12MB  11.9MB/s   00:03 ETA
[10:46:54.686] > vscode-server.tar.gz                       44%   23MB  11.8MB/s   00:02 ETA
[10:46:55.693] > vscode-server.tar.gz                       66%   35MB  11.8MB/s   00:01 ETA
[10:46:56.699] > vscode-server.tar.gz                       87%   46MB  11.7MB/s   00:00 ETA
[10:46:57.444] > vscode-server.tar.gz                      100%   53MB  11.0MB/s   00:04
[10:46:57.463] >
[10:46:57.478] > vscode-scp-done.flag                        0%    0     0.0KB/s   --:-- ETA
[10:46:57.497] > vscode-scp-done.flag                      100%    9     0.3KB/s   00:00
[10:46:57.509] >
[10:46:58.802] "Copy server to host" terminal command done
[10:49:47.820] >
>
```

Setting up SSH Host ModelArts-devserver-78f5: Copying VS Code Server to host with scp

Source: Remote - SSH (Extension)

Possible Cause

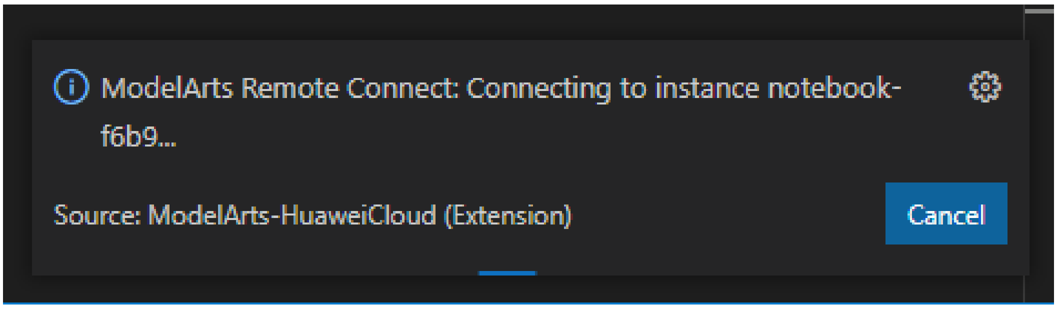
Logs show that **vscode-scp-done.flag** has been uploaded locally, but it is not received on the remote end.

Solution

Close all VS Code windows, return to the ModelArts management console, and click **Access VS Code**.

5.8.7 What Do I Do If the Connection to a Remote Development Environment Remains in the State of "ModelArts Remote Connect: Connecting to instance xxx..." for More Than 10 Minutes?

Symptom



ModelArts Remote Connect: Connecting to instance notebook-f6b9...

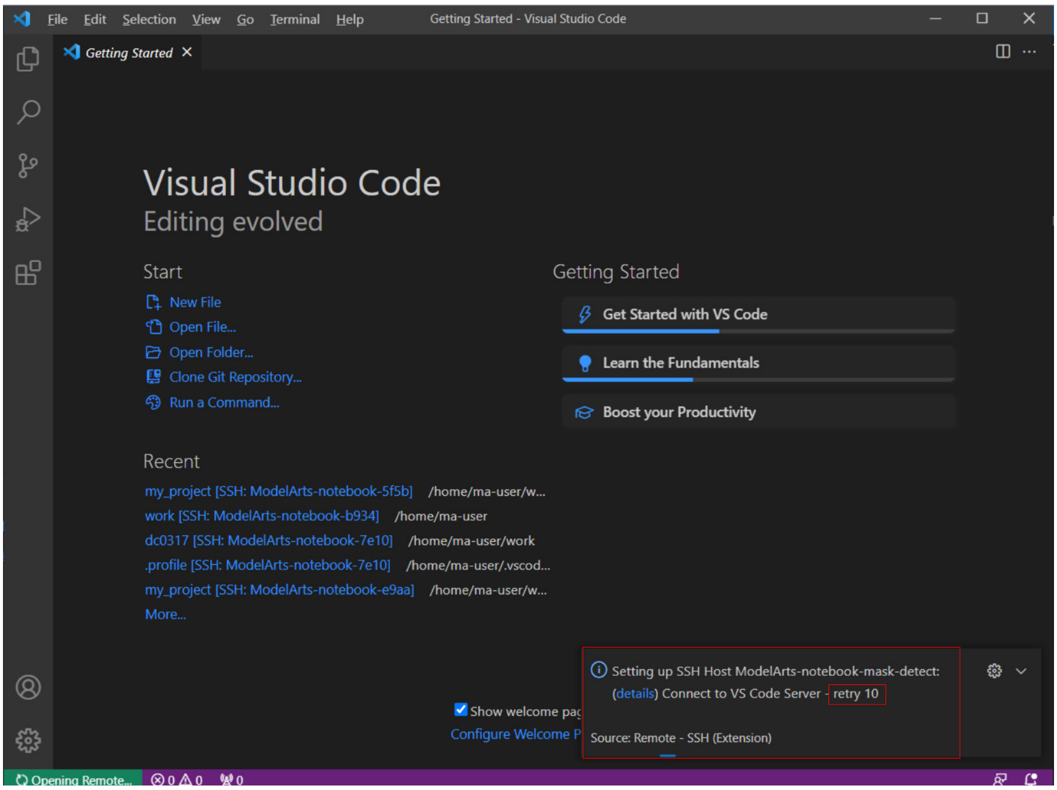
Source: ModelArts-HuaweiCloud (Extension)

Solution

Click **Cancel**, return to the ModelArts management console, and click **Access VS Code**.

5.8.8 What Do I Do If a Remote Connection Is in the Retry State?

Symptom


File Edit Selection View Go Terminal Help
Getting Started - Visual Studio Code
Getting Started
Visual Studio Code
Editing evolved
Start
New File
Open File...
Open Folder...
Clone Git Repository...
Run a Command...
Getting Started
Get Started with VS Code
Learn the Fundamentals
Boost your Productivity
Recent
my_project [SSH: ModelArts-notebook-5f5b] /home/ma-user/w...
work [SSH: ModelArts-notebook-b934] /home/ma-user
dc0317 [SSH: ModelArts-notebook-7e10] /home/ma-user/work
.profile [SSH: ModelArts-notebook-7e10] /home/ma-user/.vscod...
my_project [SSH: ModelArts-notebook-e9aa] /home/ma-user/w...
More...
Show welcome pag
Configure Welcome P
Opening Remote...
⊗ 0 ⚠ 0
0

Setting up SSH Host ModelArts-notebook-mask-detect: (details) Connect to VS Code Server - retry 10

Source: Remote - SSH (Extension)

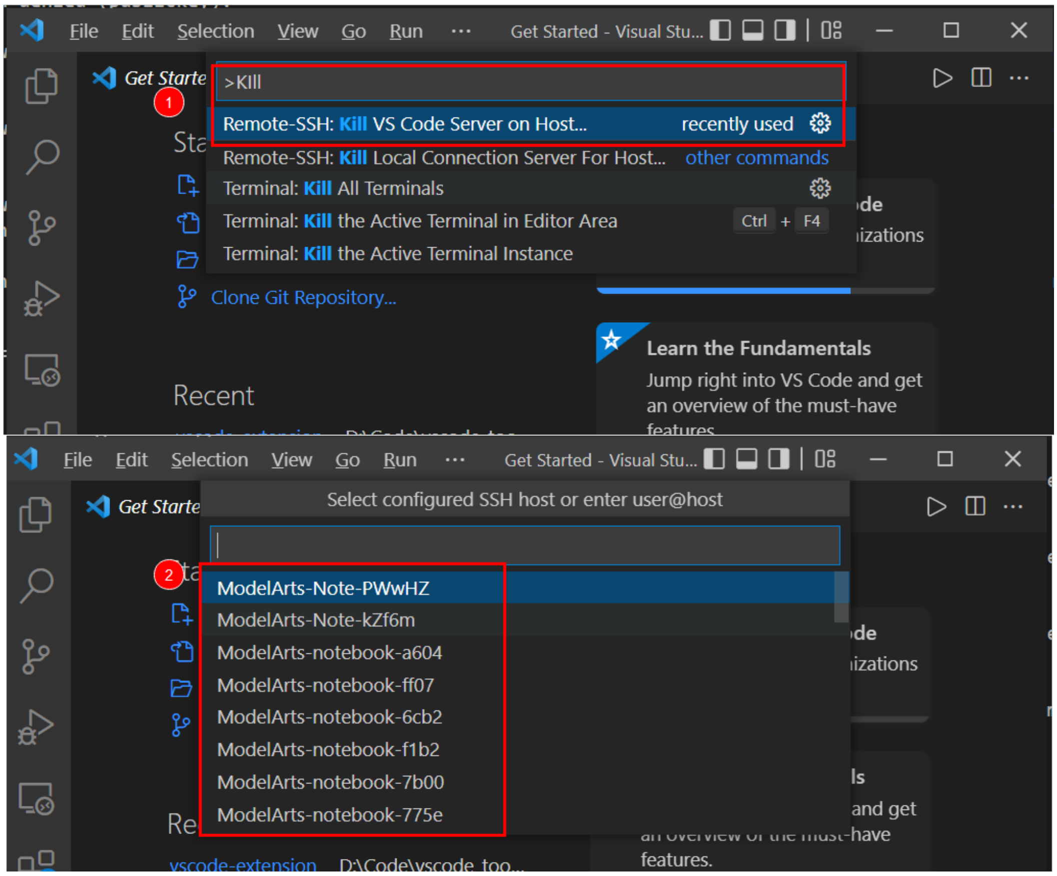
Possible Cause

Downloading the VS Code server failed before, leading to residual data. As a result, new download cannot be performed.

Solution

Method 1 (performed locally): Open the command panel (**Ctrl+Shift+P** for Windows and **Cmd+Shift+P** for macOS), search for **Kill VS Code Server on Host**, and locate the affected instance, which will be automatically cleared. Then, establish the connection again.

Figure 5-9 Clearing the affected instance


File Edit Selection View Go Run ··· Get Started - Visual Stu...
>Kill
1
Remote-SSH: Kill VS Code Server on Host... recently used
Remote-SSH: Kill Local Connection Server For Host... other commands
Terminal: Kill All Terminals
Terminal: Kill the Active Terminal in Editor Area Ctrl + F4
Terminal: Kill the Active Terminal Instance
Clone Git Repository...
Recent
Learn the Fundamentals
Jump right into VS Code and get an overview of the must-have features.
Select configured SSH host or enter user@host
2
ModelArts-Note-PWwHZ
ModelArts-Note-kZf6m
ModelArts-notebook-a604
ModelArts-notebook-ff07
ModelArts-notebook-6cb2
ModelArts-notebook-f1b2
ModelArts-notebook-7b00
ModelArts-notebook-775e
vscode-extension D:\Code\vscode_too...

Method 2 (performed remotely): Delete the files that are being used in **/home/ma-user/.vscode-server/bin/** on the VS Code terminal. Then, establish the connection again.

```
ssh -tt -o StrictHostKeyChecking=no -i ${IdentityFile} ${User}@${HostName} -p ${Port}
rm -rf /home/ma-user/.vscode-server/bin/
```

Parameters:

- **IdentityFile**: Path to the local key

- **User**: Username, for example, **ma-user**

- **HostName**: IP address

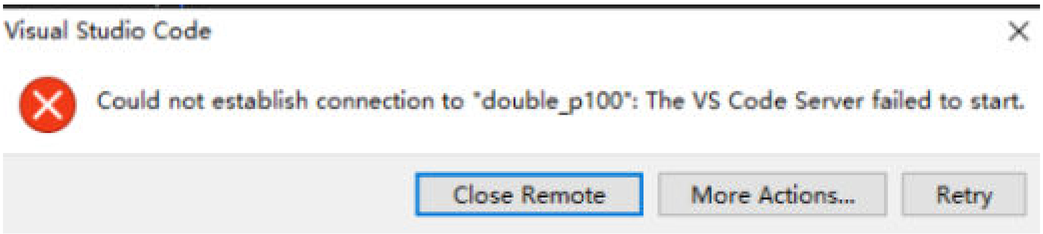
- **Port**: Port number

NOTE

The preceding methods can also be used to resolve issues related to the VS Code server.

5.8.9 What Do I Do If Error Message "The VS Code Server failed to start" Is Displayed?

Symptom

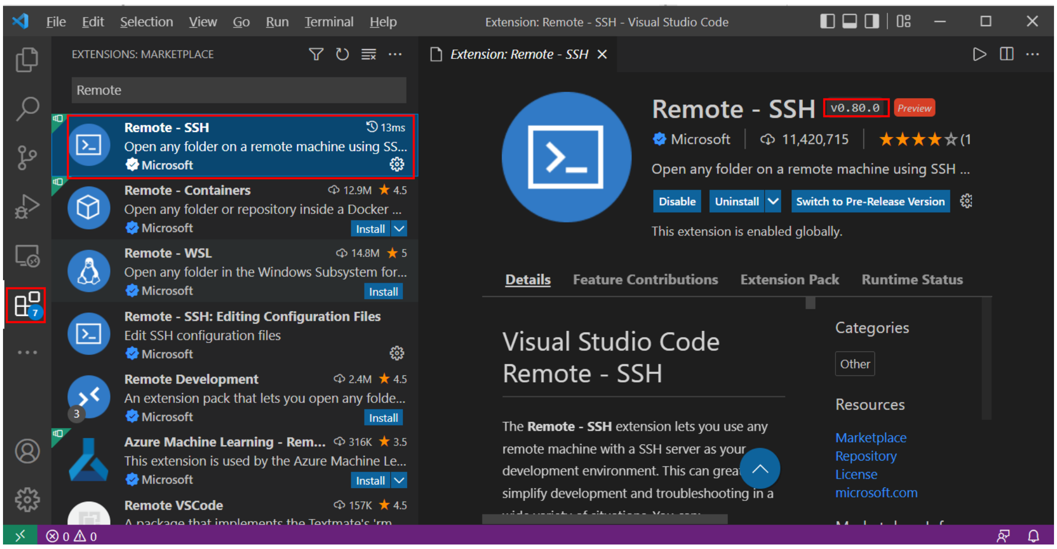


Visual Studio Code

Could not establish connection to "double_p100": The VS Code Server failed to start.

Solution

Step 1 Check whether the VS Code version is 1.65.0 or later. If so, check the Remote-SSH version. If the version is earlier than 0.76.1, upgrade Remote-SSH.

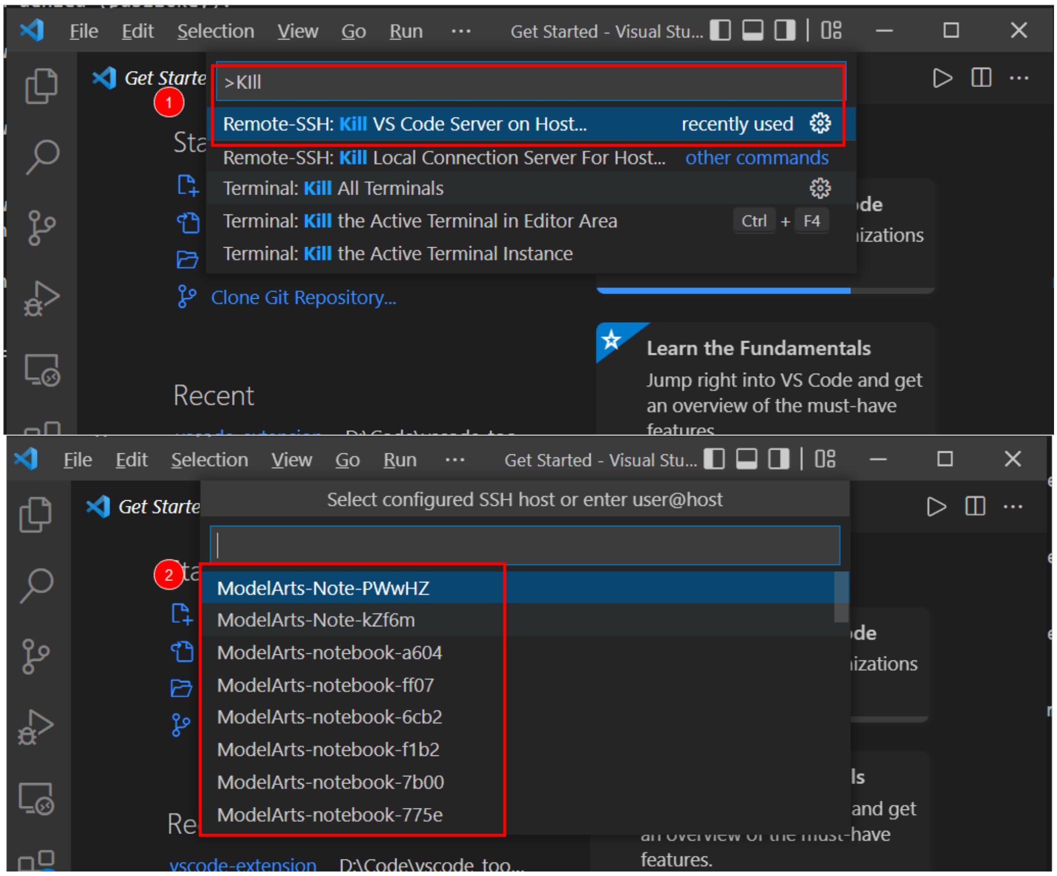


File Edit Selection View Go Run Terminal Help
Extension: Remote - SSH - Visual Studio Code
EXTENSIONS: MARKETPLACE
Remote
Remote - SSH 13ms
Open any folder on a remote machine using SS...
Microsoft
Remote - Containers 12.9M 4.5
Open any folder or repository inside a Docker ...
Microsoft Install
Remote - WSL 14.8M 5
Open any folder in the Windows Subsystem for...
Microsoft Install
Remote - SSH: Editing Configuration Files
Edit SSH configuration files
Microsoft
Remote Development 2.4M 4.5
An extension pack that lets you open any folde...
Microsoft Install
Azure Machine Learning - Rem... 316K 3.5
This extension is used by the Azure Machine Le...
Microsoft Install
Remote VSCode 157K 4.5
Extension: Remote - SSH
Remote - SSH v0.80.0 Preview
Microsoft 11,420,715 ★★★★☆(1
Open any folder on a remote machine using SSH ...
Disable Uninstall Switch to Pre-Release Version
This extension is enabled globally.
Details Feature Contributions Extension Pack Runtime Status
Visual Studio Code Remote - SSH
Categories
Other
Resources
Marketplace
Repository
License
microsoft.com
⊗ 0 ⚠ 0

The **Remote - SSH** extension lets you use any remote machine with a SSH server as your development environment. This can grea simplify development and troubleshooting in a

Step 2 Open the command panel (**Ctrl+Shift+P** for Windows and **Cmd+Shift+P** for macOS), search for **Kill VS Code Server on Host**, and locate the affected instance, which will be automatically cleared. Then, establish the connection again.

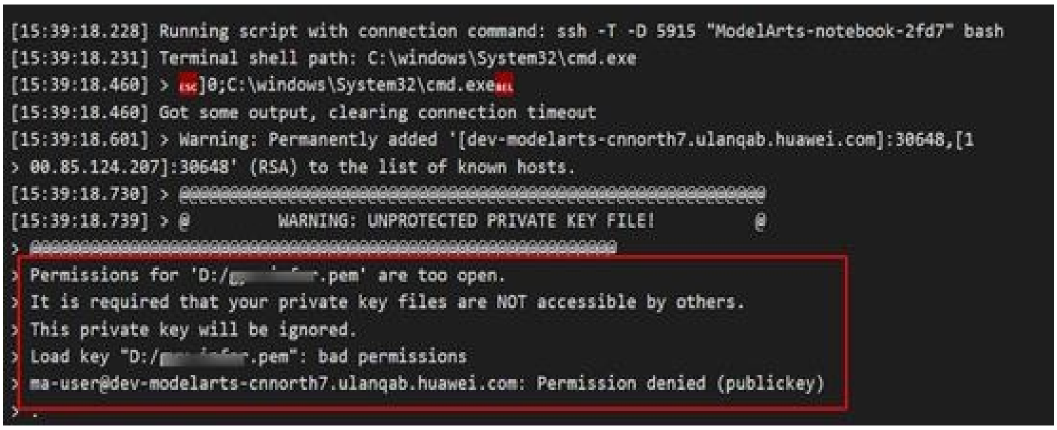
Figure 5-10 Clearing the affected instance


File Edit Selection View Go Run ···
Get Started - Visual Stu...
Get Starte
1
>Kill
Remote-SSH: Kill VS Code Server on Host...
recently used
Remote-SSH: Kill Local Connection Server For Host...
other commands
Terminal: Kill All Terminals
Terminal: Kill the Active Terminal in Editor Area
Ctrl + F4
Terminal: Kill the Active Terminal Instance
Clone Git Repository...
Recent
Learn the Fundamentals
Jump right into VS Code and get an overview of the must-have features.
Select configured SSH host or enter user@host
2
ModelArts-Note-PWwHZ
ModelArts-Note-kZf6m
ModelArts-notebook-a604
ModelArts-notebook-ff07
ModelArts-notebook-6cb2
ModelArts-notebook-f1b2
ModelArts-notebook-7b00
ModelArts-notebook-775e
vscode-extension
D:\Code\vscode_too...

----End

5.8.10 What Do I Do If Error Message "Permissions for 'x:/xxx.pem' are too open" Is Displayed?

Symptom



```
[15:39:18.228] Running script with connection command: ssh -T -D 5915 "ModelArts-notebook-2fd7" bash
[15:39:18.231] Terminal shell path: C:\windows\System32\cmd.exe
[15:39:18.460] > ␛]0;C:\windows\System32\cmd.exe␇
[15:39:18.460] Got some output, clearing connection timeout
[15:39:18.601] > Warning: Permanently added '[dev-modelarts-cnnorth7.ulanqab.huawei.com]:30648,[1
> 00.85.124.207]:30648' (RSA) to the list of known hosts.
[15:39:18.730] > @@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
[15:39:18.739] > @         WARNING: UNPROTECTED PRIVATE KEY FILE!          @
> @@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
> Permissions for 'D:/[redacted].pem' are too open.
> It is required that your private key files are NOT accessible by others.
> This private key will be ignored.
> Load key "D:/[redacted].pem": bad permissions
> ma-user@dev-modelarts-cnnorth7.ulanqab.huawei.com: Permission denied (publickey)
> .
```

Possible Cause

Possible cause 1: The key file is not stored in the specified path. For details, see the **security restrictions** or **VS Code document**. Resolve this issue by referring to solution 1.

Possible cause 2: For macOS or Linux, the permission on the key file or the folder where the key is stored may be incorrect. Resolve this issue by referring to solution 2.

Solution

Solution 1:

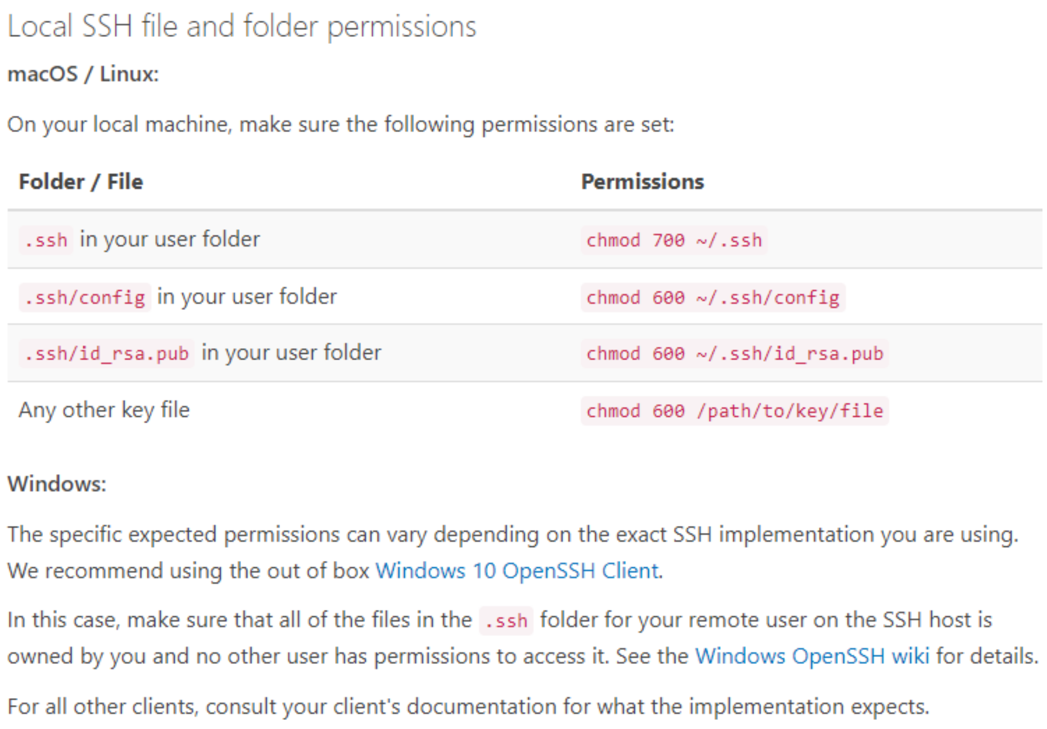
Place the key file in a specified path or its sub-path:

Windows: **C:\Users\{{user}}**

macOS or Linux: **Users/{{user}}**

Solution 2:

Check the file and folder permissions.



Local SSH file and folder permissions

macOS / Linux:

On your local machine, make sure the following permissions are set:

Folder / File	Permissions
`.ssh` in your user folder	`chmod 700 ~/.ssh`
`.ssh/config` in your user folder	`chmod 600 ~/.ssh/config`
`.ssh/id_rsa.pub` in your user folder	`chmod 600 ~/.ssh/id_rsa.pub`
Any other key file	`chmod 600 /path/to/key/file`

Windows:

The specific expected permissions can vary depending on the exact SSH implementation you are using. We recommend using the out of box Windows 10 OpenSSH Client.

In this case, make sure that all of the files in the `.ssh` folder for your remote user on the SSH host is owned by you and no other user has permissions to access it. See the Windows OpenSSH wiki for details.

For all other clients, consult your client's documentation for what the implementation expects.

5.8.11 What Do I Do If Error Message "Bad owner or permissions on C:\Users\Administrator/.ssh/config" or "Connection permission denied (publickey)" Is Displayed?

Symptom

The following error message is displayed: "Bad owner or permissions on C:\Users\Administrator/.ssh/config" or "Connection permission denied (publickey). Please make sure the key file is correctly selected and the file permission is correct. You can view the instance keypair information on ModelArts console."

Possible Causes

The permission to the SSH folder has been granted to other users, not only to the current Windows user, or the current user does not have the permission. In these cases, you only need to modify the permission.

Solution

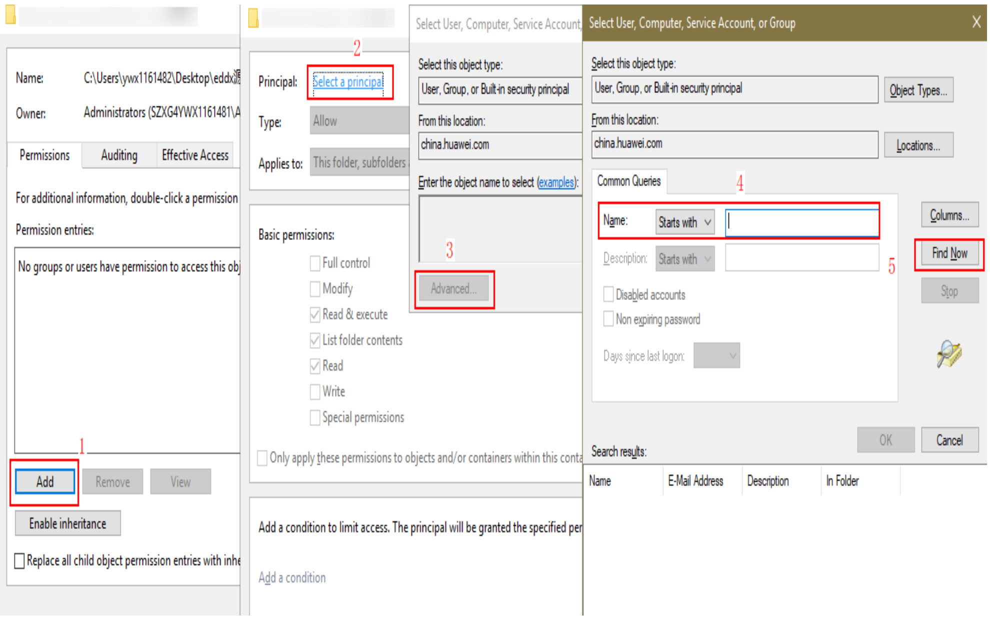
1. Find the SSH folder, which is typically located in **C:\Users**, for example, **C:\Users\xxx**.

   NOTE

   The file name in **C:\Users** must be the same as the Windows login username.

2. Right-click the folder and choose **Properties**. Then, click the **Security** tab.
3. Click **Advanced**. In the displayed window, click **Disable inheritance**. Then, in the **Block Inheritance** dialog box, click **Remove all inherited permissions from this object**. In this case, all users will be deleted.
4. Add an owner. In the same window, click **Add**. In the displayed window, click **Select a principal** next to **Principal**. In the displayed **Select User, Computer, Service Account, or Group** dialog box, click **Advanced**, enter the username, and click **Find Now**. Then, the search results will be displayed. Select your account and click **OK** to close all windows.

   Figure 5-11 Adding an owner

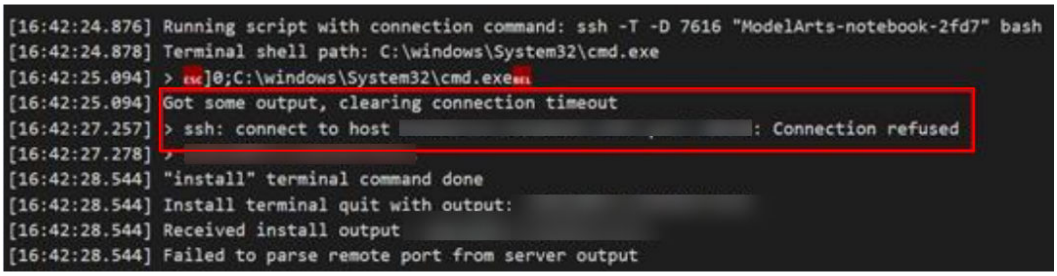


Name: C:\Users\ywx1161482\Desktop\eddx
Owner: Administrators (SZXG4YWX1161481\A
For additional information, double-click a permission
Permission entries:
No groups or users have permission to access this obj
Principal: Select a principal
Type: Allow
Applies to: This folder, subfolders
Basic permissions:
□ Full control
□ Modify
☑ Read & execute
☑ List folder contents
☑ Read
□ Write
□ Special permissions
□ Only apply these permissions to objects and/or containers within this conta
Add a condition to limit access. The principal will be granted the specified per
Select this object type: User, Group, or Built-in security principal
From this location: china.huawei.com
Enter the object name to select (examples):
Name: Starts with
Description: Starts with
□ Disabled accounts
□ Non expiring password
Days since last logon:
Search results:
□ Replace all child object permission entries with inhe

5. Close and open VS Code again and try to remotely access the SSH host. Ensure that the target key is stored in the SSH folder.

5.8.12 What Do I Do If Error Message "ssh: connect to host xxx.pem port xxxxx: Connection refused" Is Displayed?

Symptom



```
[16:42:24.876] Running script with connection command: ssh -T -D 7616 "ModelArts-notebook-2fd7" bash
[16:42:24.878] Terminal shell path: C:\windows\System32\cmd.exe
[16:42:25.094] > ESC]0;C:\windows\System32\cmd.exeBEL
[16:42:25.094] Got some output, clearing connection timeout
[16:42:27.257] > ssh: connect to host [redacted]: Connection refused
[16:42:27.278] > [redacted]
[16:42:28.544] "install" terminal command done
[16:42:28.544] Install terminal quit with output: [redacted]
[16:42:28.544] Received install output [redacted]
[16:42:28.544] Failed to parse remote port from server output
```

Possible Cause

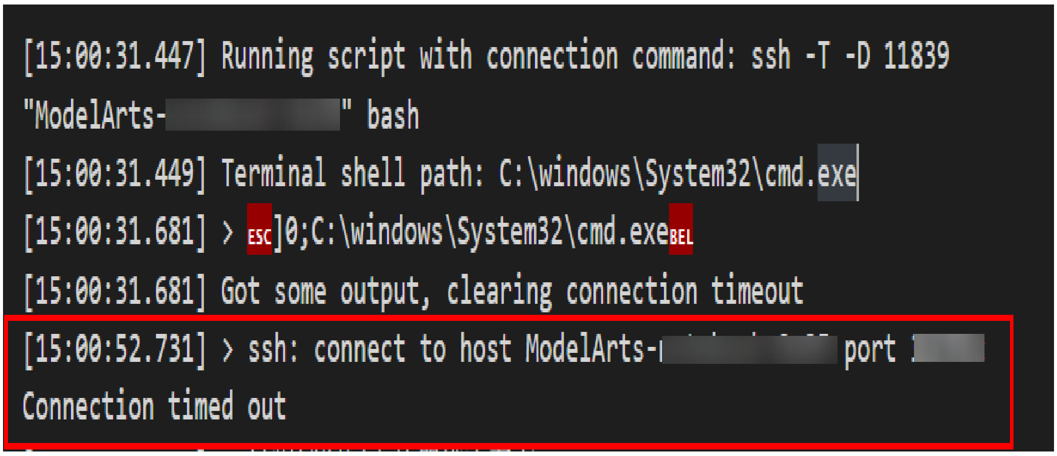
The target instance is not running.

Solution

Log in to the ModelArts management console and check the status of the instance. If the instance is stopped, start it. If the instance is in other states, such as **Error**, stop and then start it. After the instance status changes to **Running**, establish the remote connection again.

5.8.13 What Do I Do If Error Message "ssh: connect to host ModelArts-xxx port xxx: Connection timed out" Is Displayed?

Symptom



```
[15:00:31.447] Running script with connection command: ssh -T -D 11839
"ModelArts-[redacted]" bash
[15:00:31.449] Terminal shell path: C:\windows\System32\cmd.exe
[15:00:31.681] > ESC]0;C:\windows\System32\cmd.exeBEL
[15:00:31.681] Got some output, clearing connection timeout
[15:00:52.731] > ssh: connect to host ModelArts-[redacted] port [redacted]
Connection timed out
```

Possible Cause

Possible cause 1: The whitelisted IP addresses configured for the instance are different from the ones used in the local network.

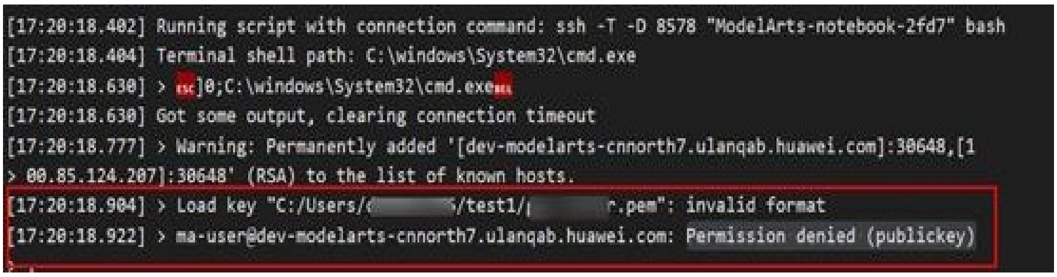
Change the whitelist so that the whitelisted IP addresses are the same as those used in the local network or disable the whitelist.

Possible cause 2: The local network is inaccessible.

Solution: Check the local network and network restrictions.

5.8.14 What Do I Do If Error Message "Load key "C:/Users/xx/test1/xxx.pem": invalid format" Is Displayed?

Symptom



```
[17:20:18.402] Running script with connection command: ssh -T -D 8578 "ModelArts-notebook-2fd7" bash
[17:20:18.404] Terminal shell path: C:\windows\System32\cmd.exe
[17:20:18.630] > ESC]0;C:\windows\System32\cmd.exeBEL
[17:20:18.630] Got some output, clearing connection timeout
[17:20:18.777] > Warning: Permanently added '[dev-modelarts-cnnorth7.ulanqab.huawei.com]:30648,[1
> 00.85.124.207]:30648' (RSA) to the list of known hosts.
[17:20:18.904] > Load key "C:/Users/[redacted]/test1/[redacted].pem": invalid format
[17:20:18.922] > ma-user@dev-modelarts-cnnorth7.ulanqab.huawei.com: Permission denied (publickey)
```

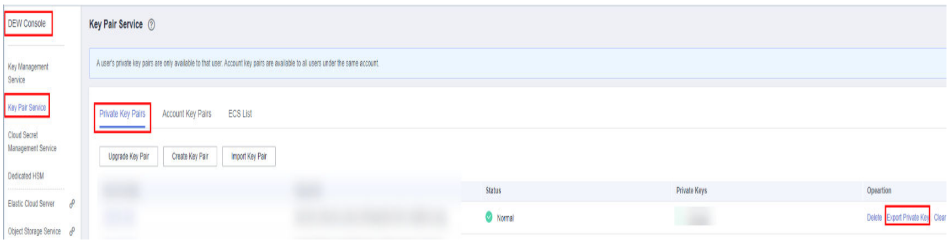
Possible Cause

The content or format of the key file is incorrect.

Solution

Use the correct key file for remote access. If there is no correct key file locally or the file is damaged, perform the following operations:

1. Log in to the console, search for **DEW**. On the DEW management console, choose **Key Pair Service** and click **Private Key Pairs**. Then, view and download the correct key file.



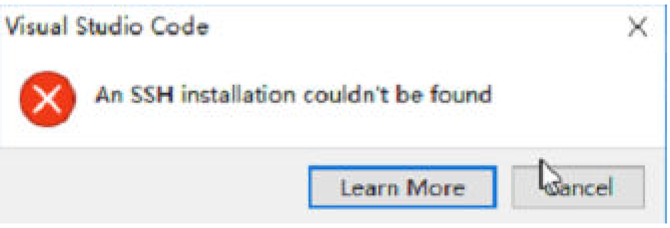
A user's private key pairs are only available to that user. Account key pairs are available to all users under the same account.

[redacted]	[redacted]	Status	Private Keys	Operation
[redacted]	[redacted]	Normal	[redacted]	Delete Export Private Key Clear

2. If the key cannot be downloaded and the originally downloaded key was lost, create a new development environment instance and a new key file. Replacing a key file in a running development environment will be supported later.

5.8.15 What Do I Do If Error Message "An SSH installation couldn't be found" or "Could not establish connection to instance xxx: 'ssh' ..." Is Displayed?

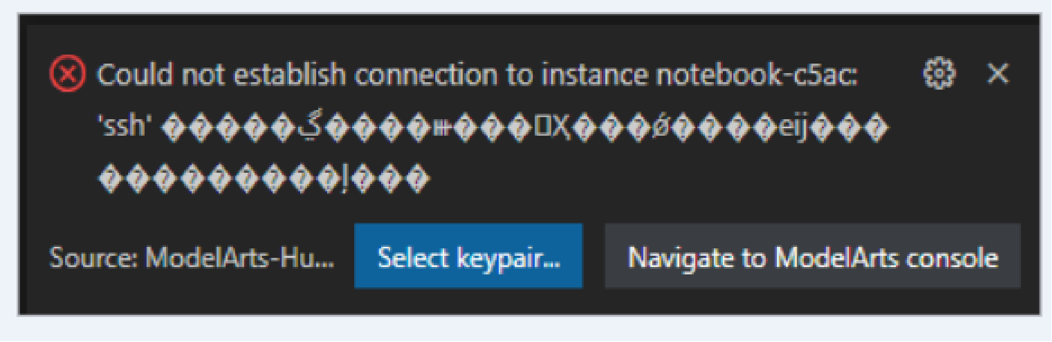
Symptom



Visual Studio Code

An SSH installation couldn't be found

Or



Could not establish connection to instance notebook-c5ac: 'ssh' �����ڴ����ⵌ���□X����ǿ�����eij��� ���������!���

Source: ModelArts-Hu...

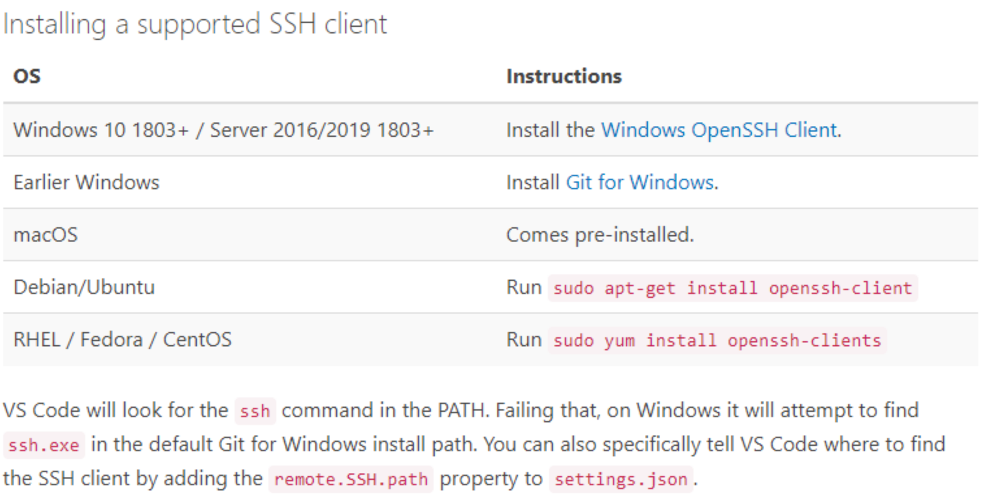
When VS Code attempts to access a notebook instance, the system always prompts you to select a certificate, and the message, excepting the title, consists of garbled characters. After the certificate is selected, the system still does not respond and the connection failed.

Possible Cause

OpenSSH is not installed in the current environment or is not installed in the default path. For details, see the **VS Code document**.

Solution

- If OpenSSH is not installed in the current environment, **download and install it**.



Installing a supported SSH client

OS	Instructions
Windows 10 1803+ / Server 2016/2019 1803+	Install the Windows OpenSSH Client.
Earlier Windows	Install Git for Windows.
macOS	Comes pre-installed.
Debian/Ubuntu	Run `sudo apt-get install openssh-client`
RHEL / Fedora / CentOS	Run `sudo yum install openssh-clients`

VS Code will look for the `ssh` command in the PATH. Failing that, on Windows it will attempt to find `ssh.exe` in the default Git for Windows install path. You can also specifically tell VS Code where to find the SSH client by adding the `remote.SSH.path` property to `settings.json`.

If OpenSSH fails to be installed, manually **download the OpenSSH installation package** and perform the following operations:

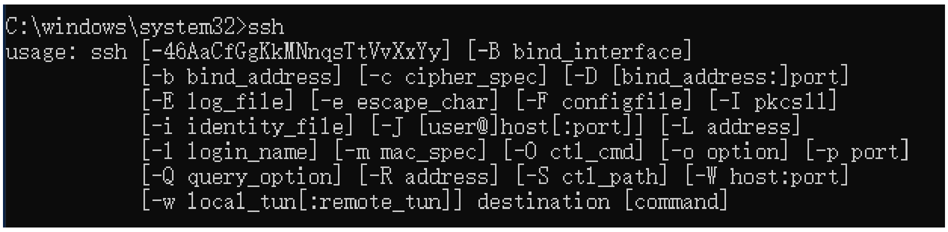
Step 1 Download the .zip package and decompress it into **C:\Windows\System32**.

Step 2 In **C:\Windows\System32\OpenSSH-xx**, open CMD as the administrator and run the following command:

```
powershell.exe -ExecutionPolicy Bypass -File install-sshd.ps1
```

Step 3 Add **C:\Program Files\OpenSSH-xx** (in which the SSH executable .exe file is stored) to environment system variables.

Step 4 Open CMD again and run **ssh**. If the following information is displayed, the installation is successful. Otherwise, go to **Step 5** and **Step 6**.



```
C:\windows\system32>ssh
usage: ssh [-46AaCfGgKkMNnqsTtVvXxYy] [-B bind_interface]
           [-b bind_address] [-c cipher_spec] [-D [bind_address:]port]
           [-E log_file] [-e escape_char] [-F configfile] [-I pkcs11]
           [-i identity_file] [-J [user@]host[:port]] [-L address]
           [-l login_name] [-m mac_spec] [-O ctl_cmd] [-o option] [-p port]
           [-Q query_option] [-R address] [-S ctl_path] [-W host:port]
           [-w local_tun[:remote_tun]] destination [command]
```

Step 5 Enable port 22 (default OpenSSH port) on the firewall and run the following command in Command Prompt:

```
netsh advfirewall firewall add rule name=sshd dir=in action=allow protocol=TCP localport=22
```

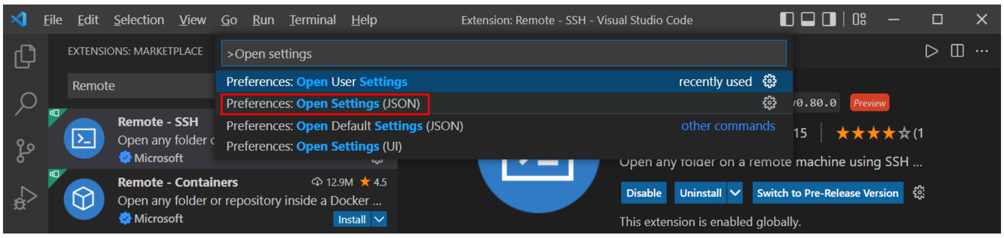
Step 6 Run the following command to start OpenSSH:

```
Start-Service sshd
```

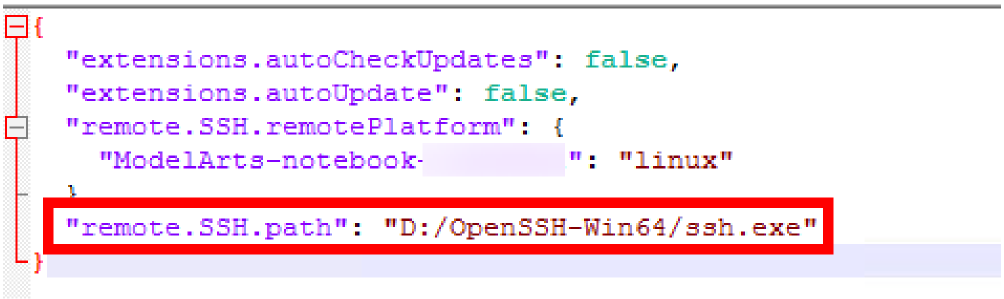
----End

- If OpenSSH is not installed in the default path, open the command panel (**Ctrl+Shift+P** for Windows and **Cmd+Shift+P** for macOS).

  Search for **Open settings**.


File Edit Selection View Go Run Terminal Help
Extension: Remote - SSH - Visual Studio Code
EXTENSIONS: MARKETPLACE
Remote
>Open settings
Preferences: Open User Settings
recently used
Preferences: Open Settings (JSON)
Preferences: Open Default Settings (JSON)
other commands
Preferences: Open Settings (UI)
Remote - SSH
Open any folder c
Microsoft
Remote - Containers
12.9M ★ 4.5
Open any folder or repository inside a Docker ...
Microsoft
Install
v0.80.0
Preview
Open any folder on a remote machine using SSH ...
Disable
Uninstall
Switch to Pre-Release Version
This extension is enabled globally.

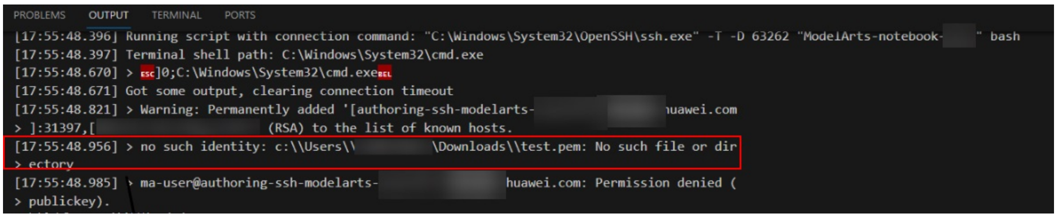
Add **remote.SSH.path** to **settings.json**, for example, **"remote.SSH.path": "*Installation path of the local OpenSSH*"**.



```
{
    "extensions.autoCheckUpdates": false,
    "extensions.autoUpdate": false,
    "remote.SSH.remotePlatform": {
      "ModelArts-notebook-[redacted]": "linux"
    }
    "remote.SSH.path": "D:/OpenSSH-Win64/ssh.exe"
}
```

5.8.16 What Do I Do If Error Message "no such identity: C:/ Users/xx /test.pem: No such file or directory" Is Displayed?

Symptom



```
[17:55:48.396] Running script with connection command: "C:\Windows\System32\OpenSSH\ssh.exe" -T -D 63262 "ModelArts-notebook-[redacted]" bash
[17:55:48.397] Terminal shell path: C:\Windows\System32\cmd.exe
[17:55:48.670] > ESC]0;C:\Windows\System32\cmd.exeBEL
[17:55:48.671] Got some output, clearing connection timeout
[17:55:48.821] > Warning: Permanently added '[authoring-ssh-modelarts-[redacted]huawei.com
> ]:31397,[[redacted] (RSA) to the list of known hosts.
[17:55:48.956] > no such identity: c:\\Users\\[redacted]\Downloads\\test.pem: No such file or dir
> ectory
[17:55:48.985] > ma-user@authoring-ssh-modelarts-[redacted]huawei.com: Permission denied (
> publickey).
```

Possible Cause

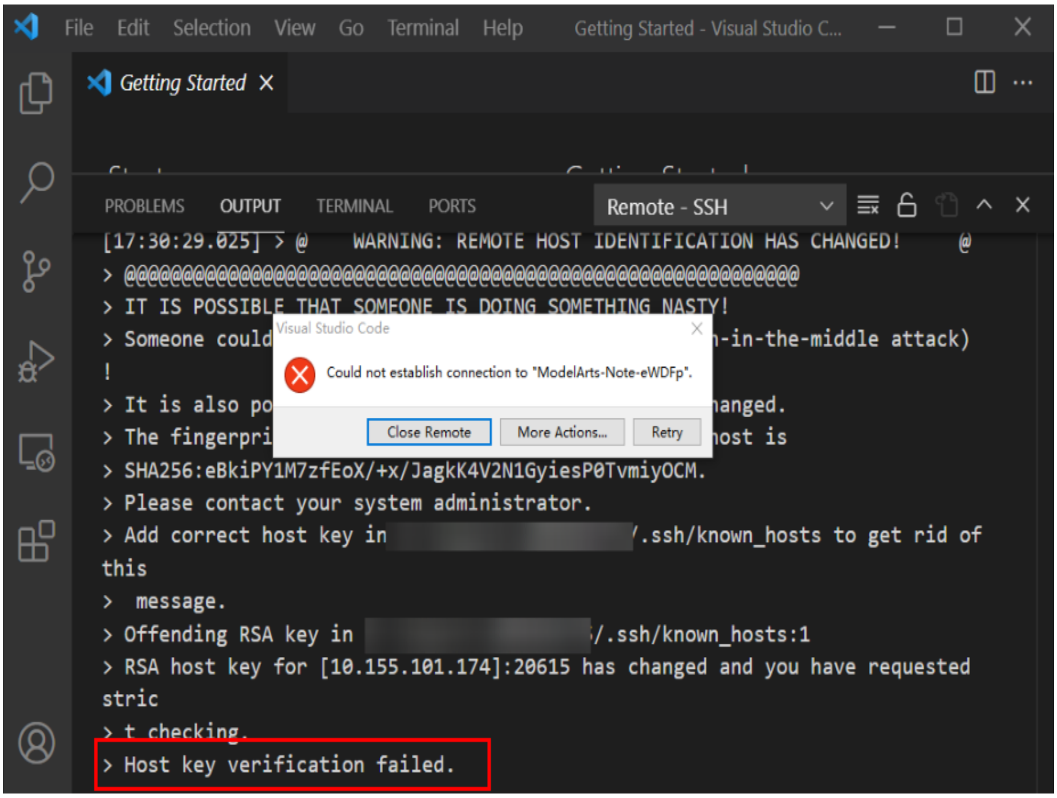
The key file is not in the path, or the name of the key file in the path has been changed.

Solution

Select the key path again.

5.8.17 What Do I Do If Error Message "Host key verification failed" or "Port forwarding is disabled" Is Displayed?

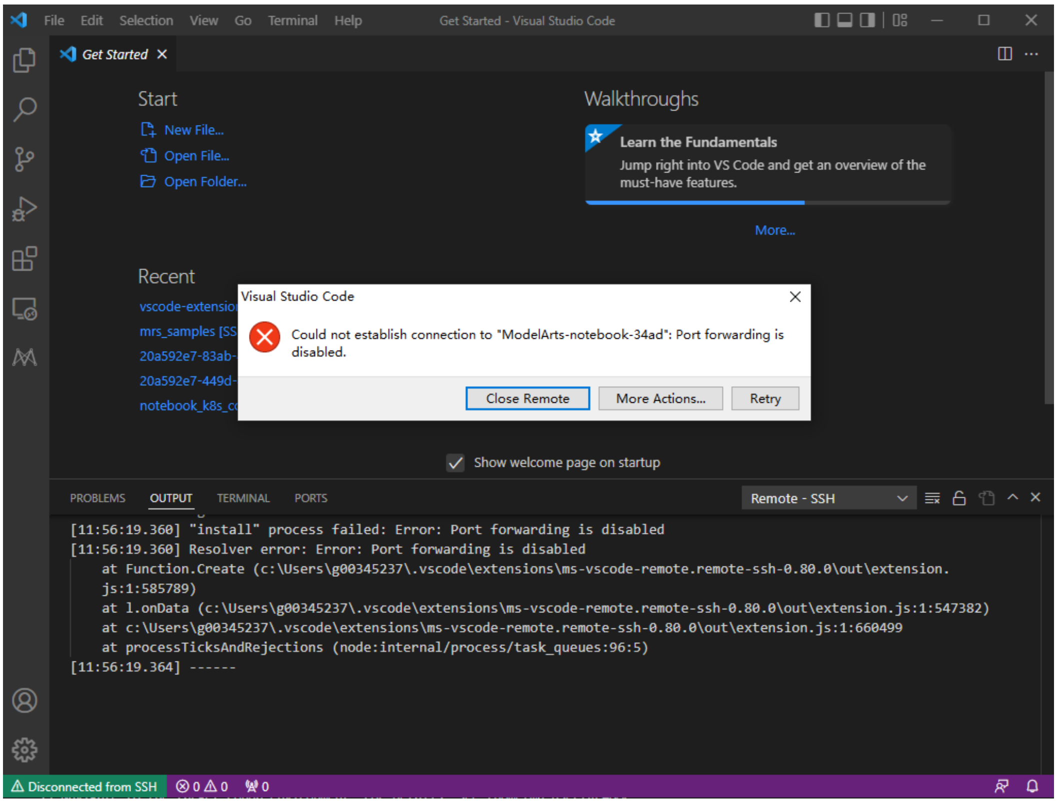
Symptom



Could not establish connection to "ModelArts-Note-eWDFp".

```
[17:30:29.025] > @    WARNING: REMOTE HOST IDENTIFICATION HAS CHANGED!     @
> @@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
> IT IS POSSIBLE THAT SOMEONE IS DOING SOMETHING NASTY!
> Someone could                                        n-in-the-middle attack)
!
> It is also po                                       anged.
> The fingerpri                                       host is
> SHA256:eBkiPY1M7zfEoX/+x/JagkK4V2N1GyiesP0TvmiyOCM.
> Please contact your system administrator.
> Add correct host key in [redacted]/.ssh/known_hosts to get rid of
this
>  message.
> Offending RSA key in [redacted]/.ssh/known_hosts:1
> RSA host key for [10.155.101.174]:20615 has changed and you have requested
stric
> t checking.
> Host key verification failed.
```

Or



Visual Studio Code

Could not establish connection to "ModelArts-notebook-34ad": Port forwarding is disabled.

Close Remote | More Actions... | Retry

```
[11:56:19.360] "install" process failed: Error: Port forwarding is disabled
[11:56:19.360] Resolver error: Error: Port forwarding is disabled
    at Function.Create (c:\Users\g00345237\.vscode\extensions\ms-vscode-remote.remote-ssh-0.80.0\out\extension.
    js:1:585789)
    at l.onData (c:\Users\g00345237\.vscode\extensions\ms-vscode-remote.remote-ssh-0.80.0\out\extension.js:1:547382)
    at c:\Users\g00345237\.vscode\extensions\ms-vscode-remote.remote-ssh-0.80.0\out\extension.js:1:660499
    at processTicksAndRejections (node:internal/process/task_queues:96:5)
[11:56:19.364] ------
```

Possible Cause

After the notebook instance is restarted, its public key changes. The alarm is generated when OpenSSH detected the key change.

Solution

- Add **-o StrictHostKeyChecking=no** for remote access through the CLI in VS Code.

  ```
  ssh -tt -o StrictHostKeyChecking=no -i ${IdentityFile} ${User}@${HostName} -p ${Port}
  ```

  Parameters:

  - **IdentityFile**: Path to the local key

  - **User**: Username, for example, **ma-user**

  - **HostName**: IP address

  - **Port**: Port number

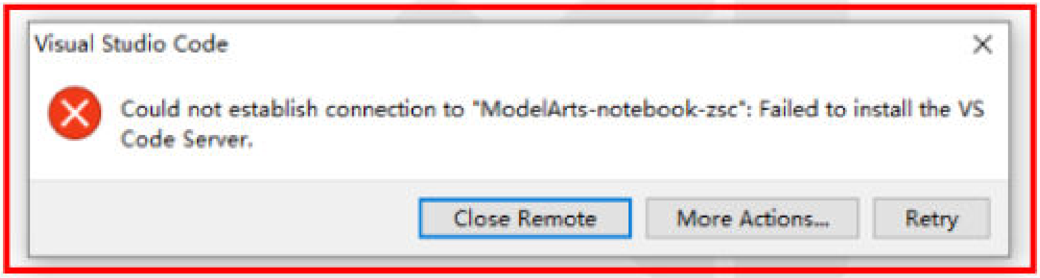
- Add **StrictHostKeyChecking no** and **UserKnownHostsFile=/dev/null** to the local **ssh config** file for manual configuration of remote access in VS Code.

  ```
  Host xxx
      HostName x.x.x.x # IP address
      Port 22522
      User ma-user
      IdentityFile C:/Users/my.pem
      StrictHostKeyChecking no
      UserKnownHostsFile=/dev/null
      ForwardAgent yes
  ```

Note that SSH logins will be insecure after the preceding parameters are added because the **known_hosts** file will be ignored during the logins.

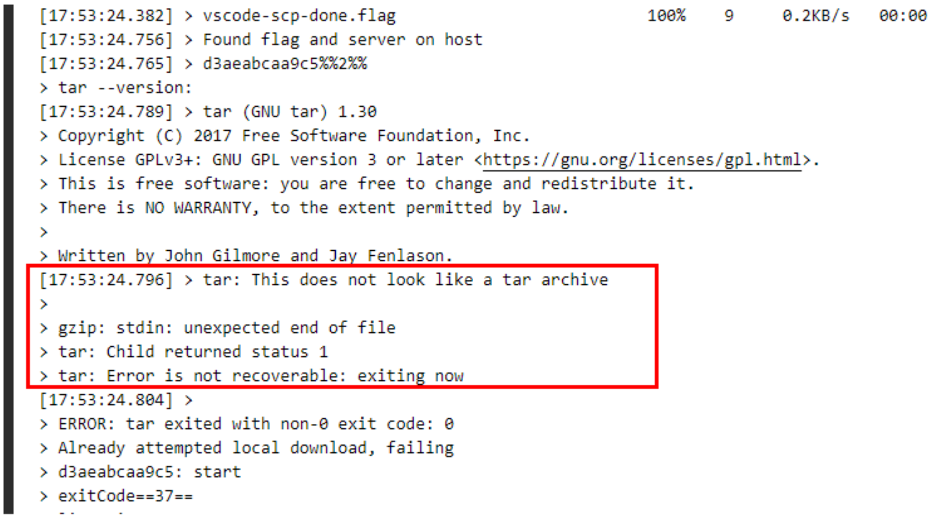
5.8.18 What Do I Do If Error Message "Failed to install the VS Code Server" or "tar: Error is not recoverable: exiting now" Is Displayed?

Symptom



Could not establish connection to "ModelArts-notebook-zsc": Failed to install the VS Code Server.

Or



```
[17:53:24.382] > vscode-scp-done.flag                               100%    9     0.2KB/s   00:00
[17:53:24.756] > Found flag and server on host
[17:53:24.765] > d3aeabcaa9c5%%2%%
> tar --version:
[17:53:24.789] > tar (GNU tar) 1.30
> Copyright (C) 2017 Free Software Foundation, Inc.
> License GPLv3+: GNU GPL version 3 or later <https://gnu.org/licenses/gpl.html>.
> This is free software: you are free to change and redistribute it.
> There is NO WARRANTY, to the extent permitted by law.
>
> Written by John Gilmore and Jay Fenlason.
[17:53:24.796] > tar: This does not look like a tar archive
>
> gzip: stdin: unexpected end of file
> tar: Child returned status 1
> tar: Error is not recoverable: exiting now
[17:53:24.804] >
> ERROR: tar exited with non-0 exit code: 0
> Already attempted local download, failing
> d3aeabcaa9c5: start
> exitCode==37==
```

Possible Cause

The disk space of **/home/ma-user/work** is insufficient.

Solution

Delete unnecessary files in **/home/ma-user/work**.

5.8.19 What Do I Do If Error Message "XHR failed" Is Displayed When a Remote Notebook Instance Is Accessed Through VS Code?

Possible Cause

The network of the environment may be faulty.

Solution

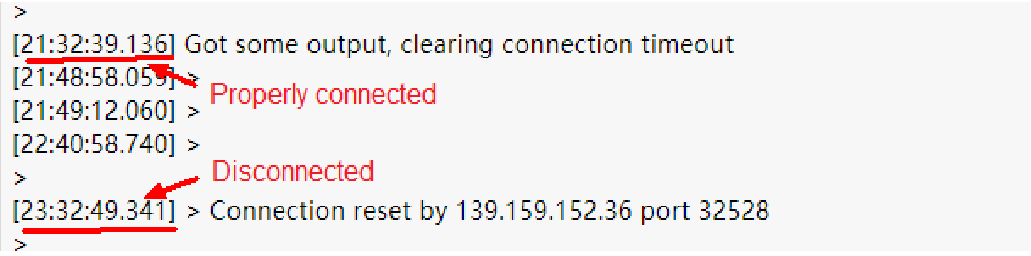
Rectify the fault by referring to **Troubleshooting Failed XHR**.

5.8.20 What Do I Do for an Automatically Disconnected VS Code Connection If No Operation Is Performed for a Long Time?

Symptom

After an SSH connection is set up through VS Code, no operation is performed for a long time and the window retains open. When the connection is used again, it is found that the connection is disconnected and no error message is displayed.

According to VS Code Remote-SSH logs, the connection was disconnected about two hours after the setup.



```
>
[21:32:39.136] Got some output, clearing connection timeout
[21:48:58.059] >
[21:49:12.060] >
[22:40:58.740] >
>
[23:32:49.341] > Connection reset by 139.159.152.36 port 32528
>
```

Possible Cause

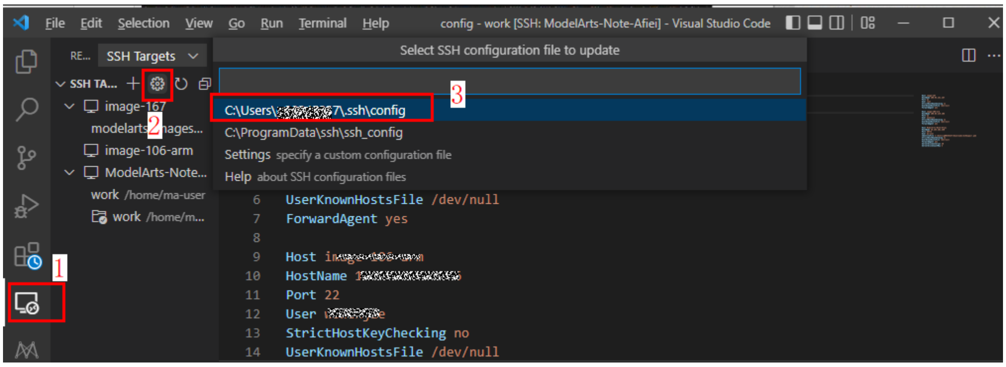
After SSH interaction stops for a period of time, the firewall disconnects idle connections (http://bluebiu.com/blog/linux-ssh-session-alive.html). The default SSH configuration does not lead to a proactive disconnection upon timeout. Since the instance runs stably on the backend, set up the connection again to resolve this issue.

Solution

To retain connections if no operation is performed for a long time, configure periodic message sending through SSH. In this way, the connection will not become idle on the firewall.

- Configure the client as needed. If the client is not configured, no heartbeat packet will be sent to the server by default.

Figure 5-12 Opening the VS Code SSH configuration file

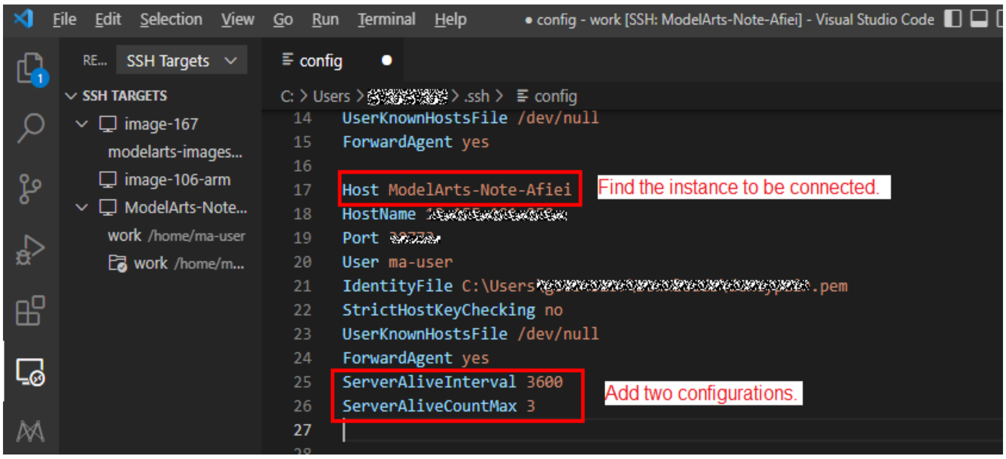


Select SSH configuration file to update

- C:\Users\[redacted]\.ssh\config
- C:\ProgramData\ssh\ssh_config
- Settings specify a custom configuration file
- Help about SSH configuration files

```
UserKnownHostsFile /dev/null
ForwardAgent yes

Host [redacted]
HostName [redacted]
Port 22
User [redacted]
StrictHostKeyChecking no
UserKnownHostsFile /dev/null
```

Figure 5-13 Adding configurations


File Edit Selection View Go Run Terminal Help
● config - work [SSH: ModelArts-Note-Afiei] - Visual Studio Code
SSH Targets
SSH TARGETS
image-167
modelarts-images...
image-106-arm
ModelArts-Note...
work /home/ma-user
work /home/m...
config
C: > Users > [redacted] > .ssh > config

```
UserKnownHostsFile /dev/null
ForwardAgent yes

Host ModelArts-Note-Afiei
HostName [redacted]
Port [redacted]
User ma-user
IdentityFile C:\Users[redacted].pem
StrictHostKeyChecking no
UserKnownHostsFile /dev/null
ForwardAgent yes
ServerAliveInterval 3600
ServerAliveCountMax 3
```

Find the instance to be connected.

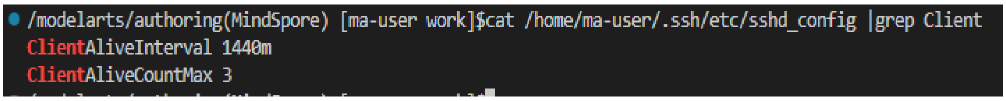
Add two configurations.

The configuration is as follows:

```
Host ModelArts-xx
   ...
   ServerAliveInterval 3600  # Add this configuration in the unit of second, indicating that the client will actively send a heartbeat packet to the server every hour.
   ServerAliveCountMax 3  # Add this configuration, indicating that if the server does not respond after the heartbeat packet is sent for three times, the connection will be disconnected.
```

For example, if the firewall is configured to disconnect a connection if the connection is idle for two hours, set **ServerAliveInterval** to a value less than two hours (for example, one hour) on the client to prevent the firewall from disconnecting the connection.

- Configure the server in **/home/ma-user/.ssh/etc/sshd_config**. (Notebook has been configured, and 24 hours is longer than the time configured on the firewall for disconnecting connections. This configuration does not need to be manually modified. It is only used to help understand the SSH configuration.)



```
/modelarts/authoring(MindSpore) [ma-user work]$cat /home/ma-user/.ssh/etc/sshd_config |grep Client
ClientAliveInterval 1440m
ClientAliveCountMax 3
```

The preceding configuration shows that the server actively sends a heartbeat packet to the client every 24 hours, and the connection will be disconnected if the client does not respond after the heartbeat packet is sent for three times.

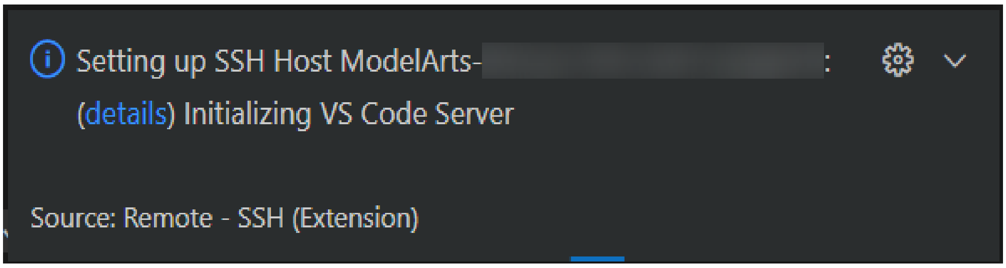
For details, see https://unix.stackexchange.com/questions/3026/what-do-options-serveraliveinterval-and-clientaliveinterval-in-sshd-config-d.

- If a connection must be consistently retained, it is a good practice to write logs in a separate log file and run the script on the backend. For example:

```
nohup train.sh > output.log 2>&1 & tail -f output.log
```

5.8.21 What Do I Do If It Takes a Long Time to Set Up a Remote Connection After VS Code Is Automatically Upgraded?

Symptom



Setting up SSH Host ModelArts-[redacted]:
(details) Initializing VS Code Server

Source: Remote - SSH (Extension)

Possible Cause

VS Code is automatically upgraded. As a result, download the new VS Code server to set up a new connection.

Solution

Disable automatic VS Code upgrade. To do so, click **Settings** in the lower left corner, search for **Update: Mode**, and set it to **none**.

Figure 5-14 Settings

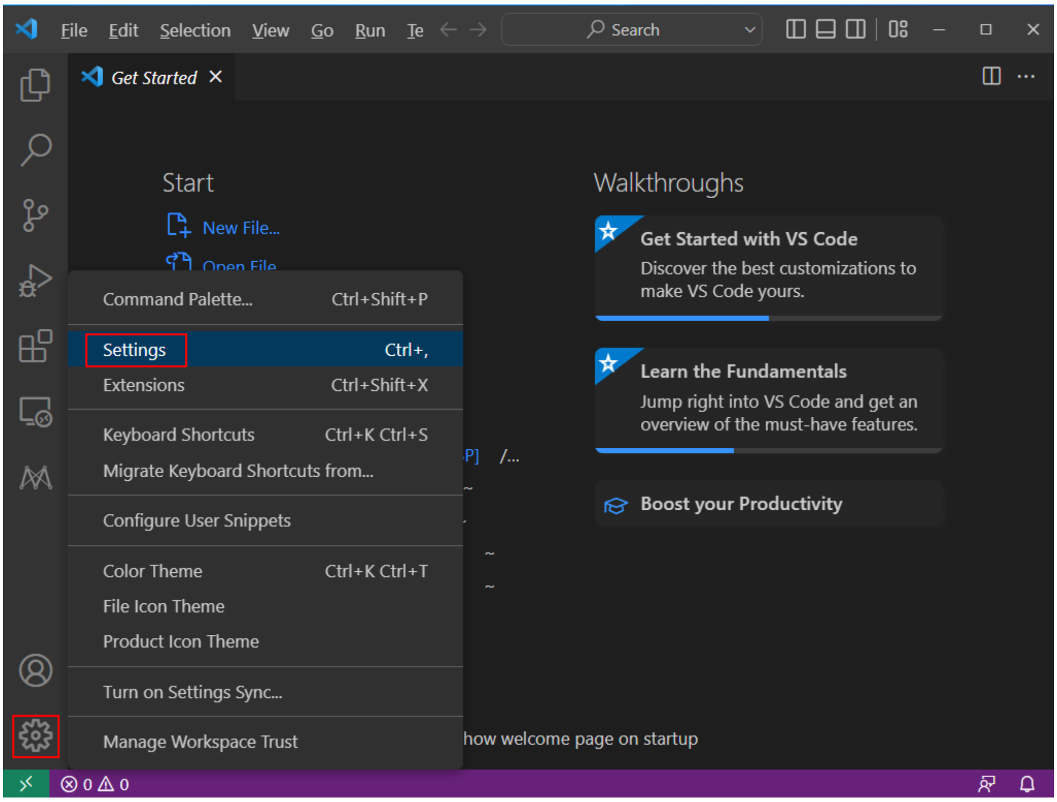
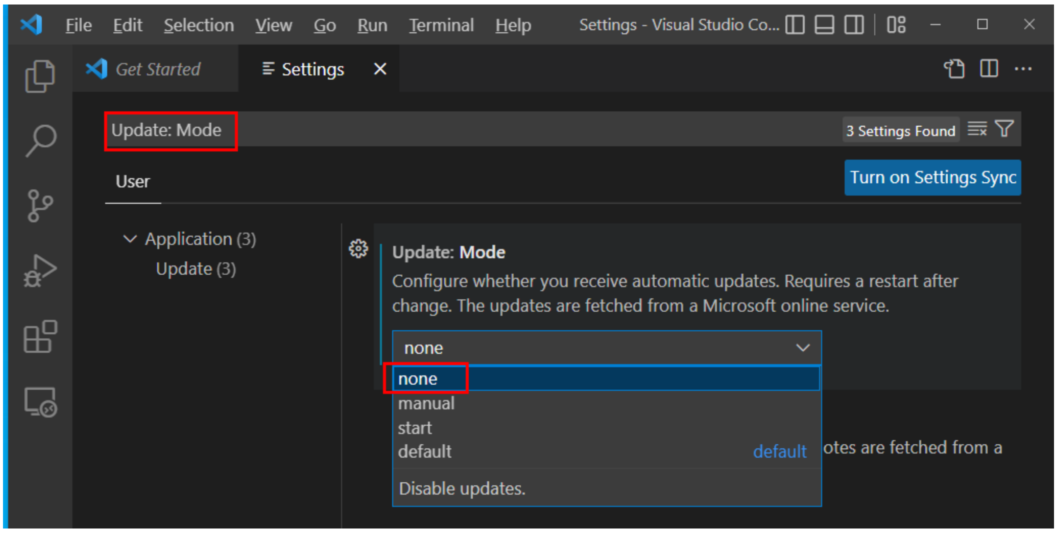


Figure 5-15 Setting the update mode to none



Update: Mode

Update: Mode
Configure whether you receive automatic updates. Requires a restart after change. The updates are fetched from a Microsoft online service.

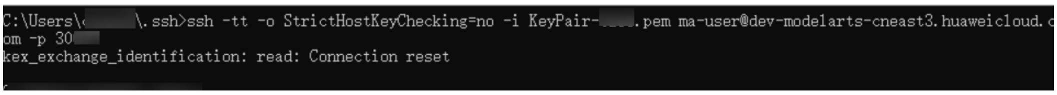
none

- none
- manual
- start
- default — default

Disable updates.

5.8.22 What Do I Do If Error Message "Connection reset" Is Displayed During an SSH Connection?

Symptom



```
C:\Users\[redacted]\.ssh>ssh -tt -o StrictHostKeyChecking=no -i KeyPair-[redacted].pem ma-user@dev-modelarts-cneast3.huaweicloud.c
om -p 30[redacted]
kex_exchange_identification: read: Connection reset
```

Possible Causes

The user network is restricted. For example, SSH is disabled by default on some enterprise networks.

Solution

Apply for the SSH permission.

5.8.23 What Can I Do If a Notebook Instance Is Frequently Disconnected or Stuck After I Use MobaXterm to Connect to the Notebook Instance in SSH Mode?

Symptom

After MobaXterm is connected to a development environment, it is disconnected after a period of time.

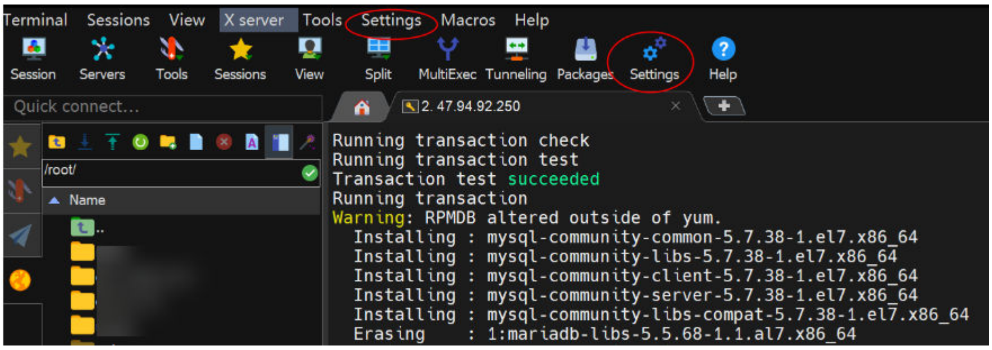
Possible Cause

When MobaXterm is configured, **SSH keepalive** is not selected or **Stop server after** of MobaXterm Professional is set to a value that is too small.

Solution

Step 1 Open MobaXterm and click **Settings** on the menu bar.

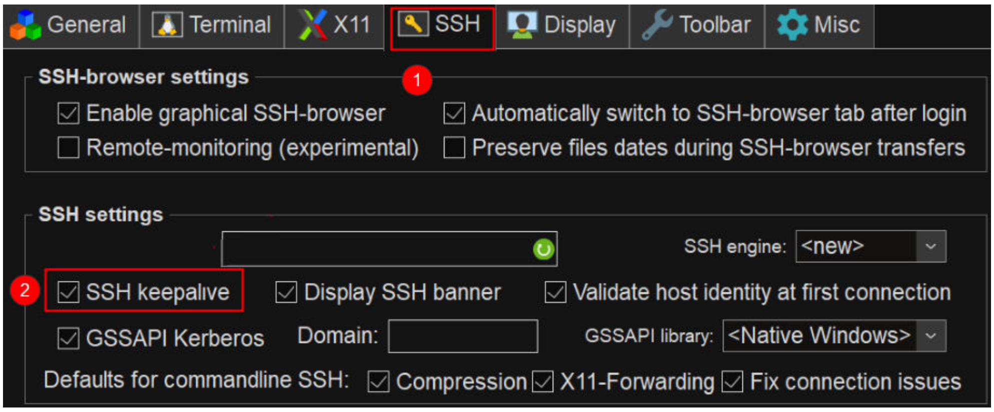
Figure 5-16 Settings



```
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
Warning: RPMDB altered outside of yum.
  Installing : mysql-community-common-5.7.38-1.el7.x86_64
  Installing : mysql-community-libs-5.7.38-1.el7.x86_64
  Installing : mysql-community-client-5.7.38-1.el7.x86_64
  Installing : mysql-community-server-5.7.38-1.el7.x86_64
  Installing : mysql-community-libs-compat-5.7.38-1.el7.x86_64
  Erasing    : 1:mariadb-libs-5.5.68-1.1.al7.x86_64
```

Step 2 On the MobaXterm configuration page, click the **SSH** tab and select **SSH keepalive**.

Figure 5-17 Selecting SSH keepalive



SSH-browser settings

☑ Enable graphical SSH-browser ☑ Automatically switch to SSH-browser tab after login

□ Remote-monitoring (experimental) □ Preserve files dates during SSH-browser transfers

SSH settings

SSH engine: <new>

☑ SSH keepalive ☑ Display SSH banner ☑ Validate host identity at first connection

☑ GSSAPI Kerberos Domain: GSSAPI library: <Native Windows>

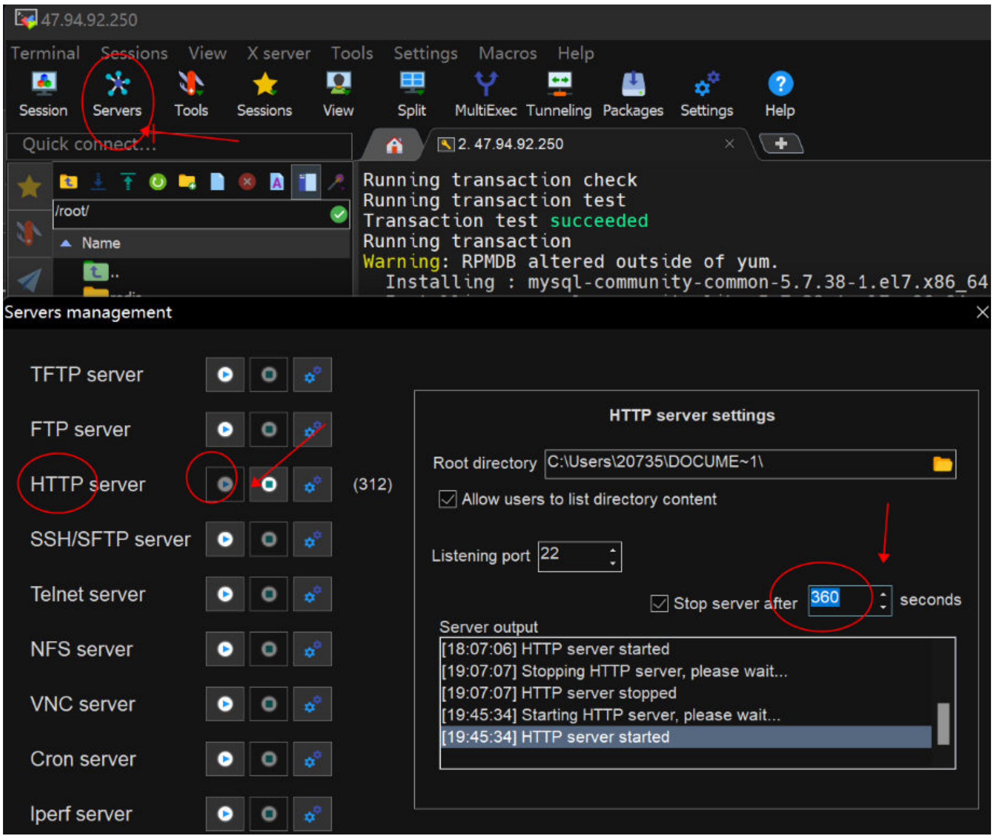
Defaults for commandline SSH: ☑ Compression ☑ X11-Forwarding ☑ Fix connection issues

NOTE

If MobaXterm Professional is used, go to **step 3**.

Step 3 Change the default value **360 seconds** to **3600 seconds** or a larger value for **Stop server after**.

Figure 5-18 Setting Stop server after



```
Running transaction check
Running transaction test
Transaction test succeeded
Running transaction
Warning: RPMDB altered outside of yum.
  Installing : mysql-community-common-5.7.38-1.el7.x86_64
```

HTTP server settings

Root directory C:\Users\20735\DOCUME~1\

☑ Allow users to list directory content

Listening port 22

☑ Stop server after 360 seconds

Server output

```
[18:07:06] HTTP server started
[19:07:07] Stopping HTTP server, please wait...
[19:07:07] HTTP server stopped
[19:45:34] Starting HTTP server, please wait...
[19:45:34] HTTP server started
```

----End

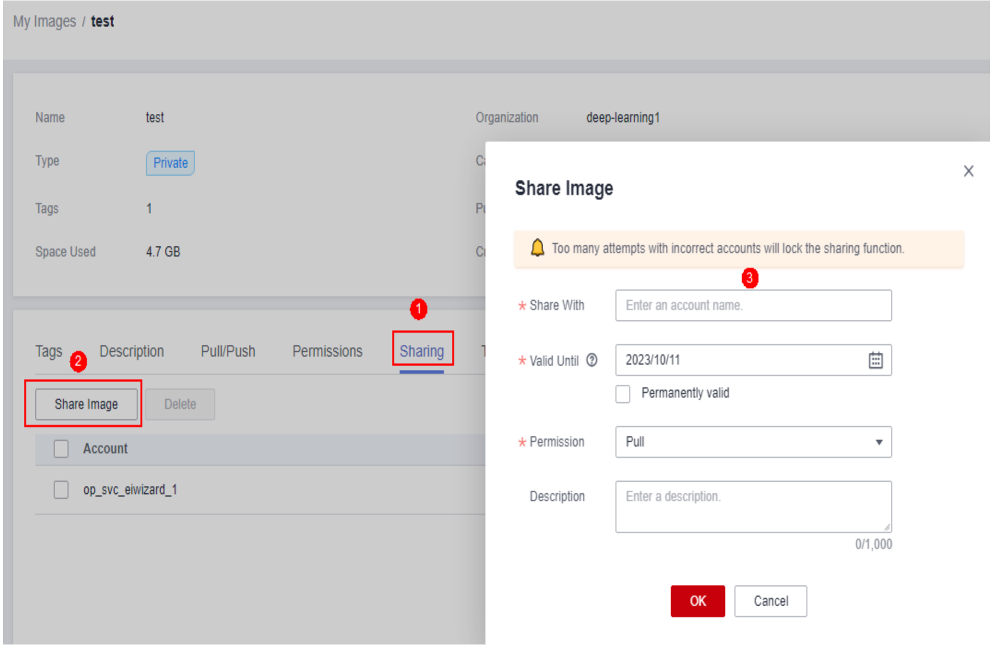
5.9 Using Custom Images in Notebook

5.9.1 How Do I Use the Image Customized by a User Under a Different Master Account to Create a Notebook Instance?

Two users belong to different master accounts. If user A needs to use user B's custom image to create a notebook instance, user B needs to share the image with user A. Then user A pulls the shared image and registers it before using it in the notebook instance. The procedure is as follows:

Operations performed by user B:

1. Log in to the SWR console and choose **My Images**.
2. Click the name of the image to be shared to access its details page.
3. On the **Sharing** tab, click **Share Image**. In the displayed dialog box, set required parameters such as the account and click **OK**.



My Images / test

Name test
Organization deep-learning1
Type Private
Tags 1
Space Used 4.7 GB

Share Image

Too many attempts with incorrect accounts will lock the sharing function.

* Share With: Enter an account name.

* Valid Until: 2023/10/11

□ Permanently valid

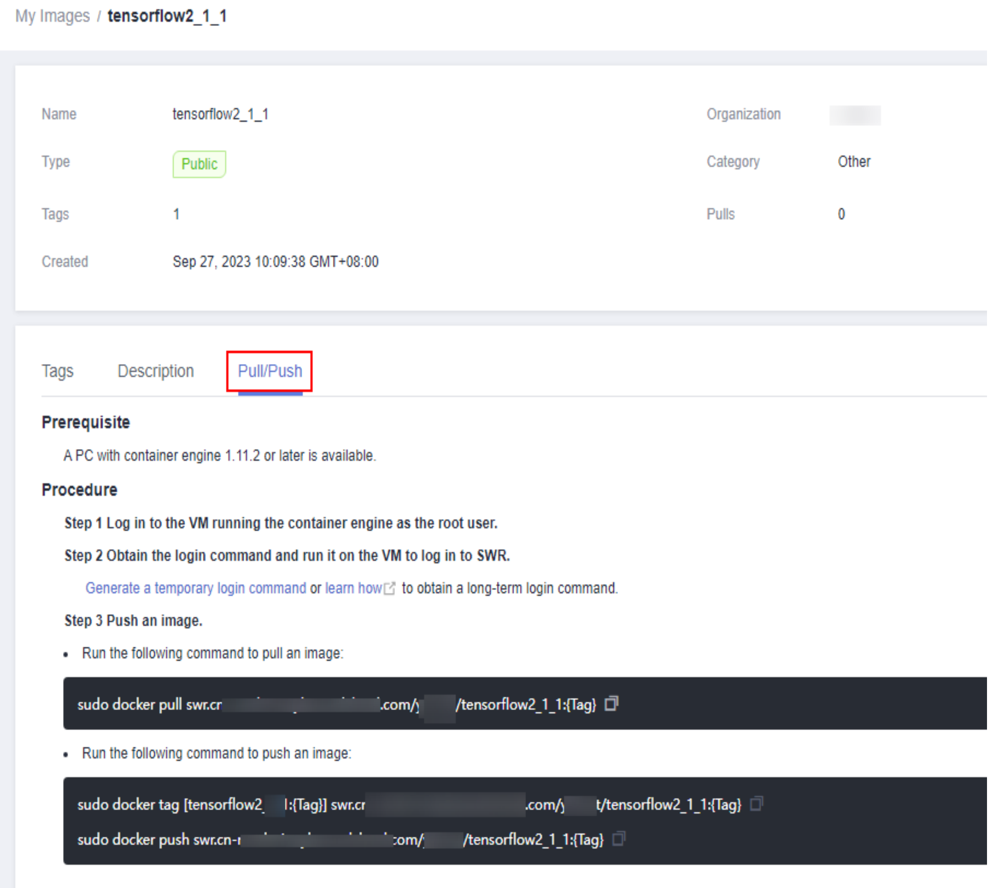
* Permission: Pull

Description: Enter a description.

0/1,000

Operations performed by user A:

1. Log in to the SWR console, choose **My Images** > **Shared Images**, view the image shared by user B, and click the image name to go to its details page.
2. Pull the image shared by user B as your own image by following the instructions provided on the **Pull/Push** tab.



My Images / tensorflow2_1_1

Name tensorflow2_1_1
Organization [redacted]
Type Public
Category Other
Tags 1
Pulls 0
Created Sep 27, 2023 10:09:38 GMT+08:00

Prerequisite

A PC with container engine 1.11.2 or later is available.

Procedure

Step 1 Log in to the VM running the container engine as the root user.

Step 2 Obtain the login command and run it on the VM to log in to SWR.

Generate a temporary login command or learn how to obtain a long-term login command.

Step 3 Push an image.

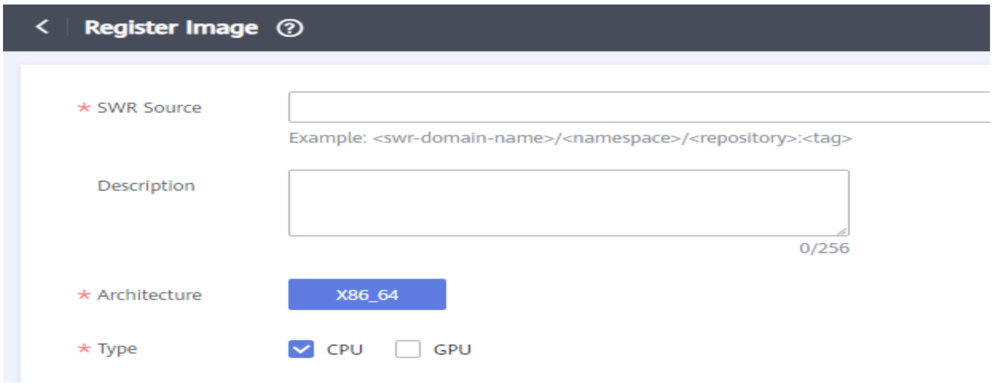
- Run the following command to pull an image:

```
sudo docker pull swr.cr[redacted].com/[redacted]/tensorflow2_1_1:{Tag}
```

- Run the following command to push an image:

```
sudo docker tag [tensorflow2_[redacted]l:{Tag}] swr.cr[redacted].com/[redacted]t/tensorflow2_1_1:{Tag}
sudo docker push swr.cn-r[redacted]:om/[redacted]/tensorflow2_1_1:{Tag}
```

3. Log in to the ModelArts console, select the pulled image, and register it. After the registration is successful, you can use the image on the notebook instance.



* SWR Source

Example: <swr-domain-name>/<namespace>/<repository>:<tag>

Description

0/256

* Architecture: X86_64

* Type: ☑ CPU □ GPU

5.10 Others

5.10.1 How Do I Use Multiple Ascend Cards for Debugging in a Notebook Instance?

An Ascend multi-card training job runs in multi-process, multi-card mode. The number of cards is equal to the number of Python processes. The Ascend underlayer reads the environment variable **RANK_TABLE_FILE**, which has been configured in the development environment, without requiring manual configuration. For example, to run a job on eight cards, the code is as follows:

```
export RANK_SIZE=8
current_exec_path=$(pwd)
echo 'start training'
for((i=0;i<=$RANK_SIZE-1;i++));
do
echo 'start rank '$i
mkdir ${current_exec_path}/device$i
cd ${current_exec_path}/device$i
echo $i
export RANK_ID=$i
dev=`expr $i + 0`
echo $dev
export DEVICE_ID=$dev
python train.py > train.log 2>&1 &
done
```

Set the environment variable **DEVICE_ID** in **train.py**.

```
devid = int(os.getenv('DEVICE_ID'))
 context.set_context(mode=context.GRAPH_MODE, device_target="Ascend", device_id=devid)
```

5.10.2 Why Is the Training Speed Similar When Different Notebook Flavors Are Used?

If your training job is single-process in code, the training speed is basically the same no matter when the notebook flavor of 8 vCPUs and 64 GB of memory or the flavor of 72 vCPUs and 512 GB of memory is used. For example, if your training job uses 2 vCPUs and 4 GB of memory, the training speed is similar no matter when you use the notebook flavor of 4 vCPUs and 8 GB of memory or the flavor of 8 vCPUs and 64 GB of memory.

If your training job is multi-process in code, the training speed backed by the notebook flavor of 72 vCPUs and 512 GB of memory is higher than that backed by the notebook flavor of 8 vCPUs and 64 GB of memory.

5.10.3 How Do I Perform Incremental Training When Using MoXing?

If you are not satisfied with training results when using MoXing to build a model, you can perform incremental training after modifying some data and label information.

Adding Incremental Training Parameters to mox.run

After modifying labeling data or datasets, you can modify the **log_dir** parameter in and add the **checkpoint_path** parameter to **mox.run**. Set **log_dir** to a new directory and **checkpoint_path** to the output path of the previous training results. If the output path is an OBS directory, set the path to a value starting with **obs://**.

If labels are changed for label data, perform operations in **If Labels Are Changed** before running **mox.run**.

```
mox.run(input_fn=input_fn,
        model_fn=model_fn,
        optimizer_fn=optimizer_fn,
        run_mode=flags.run_mode,
        inter_mode=mox.ModeKeys.EVAL if use_eval_data else None,
        log_dir=log_dir,
        batch_size=batch_size_per_device,
        auto_batch=False,
        max_number_of_steps=max_number_of_steps,
        log_every_n_steps=flags.log_every_n_steps,
        save_summary_steps=save_summary_steps,
        save_model_secs=save_model_secs,
        checkpoint_path=flags.checkpoint_url,
        export_model=mox.ExportKeys.TF_SERVING)
```

If Labels Are Changed

If the labels in a dataset have changed, execute the following statement. The statement must be executed before running **mox.run**.

In the statement, the **logits** variable indicates classification layer weights in different networks, and different parameters are configured. Set this parameter to the corresponding keyword.

```
mox.set_flag('checkpoint_exclude_patterns', 'logits')
```

If the built-in network of MoXing is used, the corresponding keyword needs to be obtained by calling the following API. In this example, the **Resnet_v1_50** keyword is the value of **logits**.

```
import moxing.tensorflow as mox

model_meta = mox.get_model_meta(mox.NetworkKeys.RESNET_V1_50)
logits_pattern = model_meta.default_logits_pattern
print(logits_pattern)
```

You can also obtain a list of networks supported by MoXing by calling the following API:

```
import moxing.tensorflow as mox
print(help(mox.NetworkKeys))
```

The following information is displayed:

```
Help on class NetworkKeys in module
moxing.tensorflow.nets.nets_factory:

class NetworkKeys(builtins.object)
 |  Data descriptors defined here:
 |
 |  __dict__
 |      dictionary for instance variables (if defined)
 |
 |  __weakref__
 |      list of weak references to the object (if defined)
 |
 |  ----------------------------------------------------------------------
 |  Data and other attributes defined here:
 |
 |  ALEXNET_V2 = 'alexnet_v2'
 |
 |  CIFARNET = 'cifarnet'
 |
 |  INCEPTION_RESNET_V2 = 'inception_resnet_v2'
 |
 |  INCEPTION_V1 = 'inception_v1'
 |
 |  INCEPTION_V2 = 'inception_v2'
 |
 |  INCEPTION_V3 = 'inception_v3'
 |
 |  INCEPTION_V4 = 'inception_v4'
 |
 |  LENET = 'lenet'
 |
 |  MOBILENET_V1 = 'mobilenet_v1'
 |
 |  MOBILENET_V1_025 = 'mobilenet_v1_025'
 |
 |  MOBILENET_V1_050 = 'mobilenet_v1_050'
 |
 |  MOBILENET_V1_075 = 'mobilenet_v1_075'
 |
 |  MOBILENET_V2 = 'mobilenet_v2'
 |
 |  MOBILENET_V2_035 = 'mobilenet_v2_035'
 |
 |  MOBILENET_V2_140 = 'mobilenet_v2_140'
 |
 |  NASNET_CIFAR = 'nasnet_cifar'
 |
 |  NASNET_LARGE = 'nasnet_large'
 |
 |  NASNET_MOBILE = 'nasnet_mobile'
 |
 |  OVERFEAT = 'overfeat'
 |
 |  PNASNET_LARGE = 'pnasnet_large'
 |
 |  PNASNET_MOBILE = 'pnasnet_mobile'
 |
 |  PVANET = 'pvanet'
 |
 |  RESNET_V1_101 = 'resnet_v1_101'
 |
 |  RESNET_V1_110 = 'resnet_v1_110'
```

```
|
| RESNET_V1_152 = 'resnet_v1_152'
|
| RESNET_V1_18 = 'resnet_v1_18'
|
| RESNET_V1_20 = 'resnet_v1_20'
|
| RESNET_V1_200 = 'resnet_v1_200'
|
| RESNET_V1_50 = 'resnet_v1_50'
|
| RESNET_V1_50_8K = 'resnet_v1_50_8k'
|
| RESNET_V1_50_MOX = 'resnet_v1_50_mox'
|
| RESNET_V1_50_OCT = 'resnet_v1_50_oct'
|
| RESNET_V2_101 = 'resnet_v2_101'
|
| RESNET_V2_152 = 'resnet_v2_152'
|
| RESNET_V2_200 = 'resnet_v2_200'
|
| RESNET_V2_50 = 'resnet_v2_50'
|
| RESNEXT_B_101 = 'resnext_b_101'
|
| RESNEXT_B_50 = 'resnext_b_50'
|
| RESNEXT_C_101 = 'resnext_c_101'
|
| RESNEXT_C_50 = 'resnext_c_50'
|
| VGG_16 = 'vgg_16'
|
| VGG_16_BN = 'vgg_16_bn'
|
| VGG_19 = 'vgg_19'
|
| VGG_19_BN = 'vgg_19_bn'
|
| VGG_A = 'vgg_a'
|
| VGG_A_BN = 'vgg_a_bn'
|
| XCEPTION_41 = 'xception_41'
|
| XCEPTION_65 = 'xception_65'
|
| XCEPTION_71 = 'xception_71'
```

5.10.4 How Do I View GPU Usage on the Notebook?

If you select GPU when creating a notebook instance, perform the following operations to view GPU usage:

1. Log in to the ModelArts management console, and choose **DevEnviron > Notebooks**.
2. In the **Operation** column of the target notebook instance in the notebook list, click **Open** to go to the **Jupyter** page.
3. On the **Files** tab page of the **Jupyter** page, click **New** and select **Terminal**. The **Terminal** page is displayed.
4. Run the following command to view GPU usage:

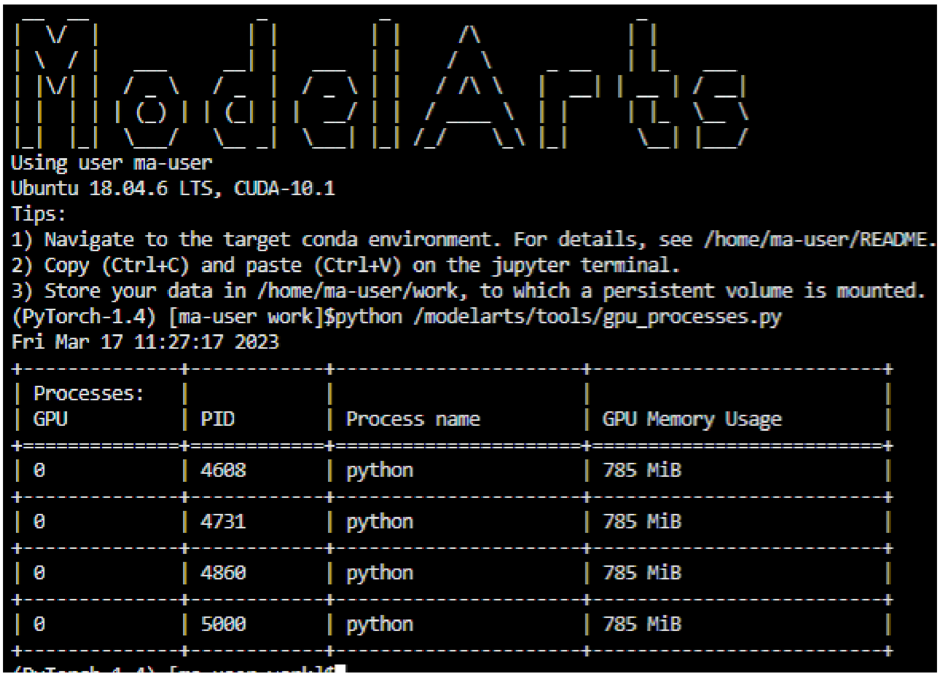
   ```
   nvidia-smi
   ```

5. Check which processes in the current notebook instance use GPUs.

   Method 1:

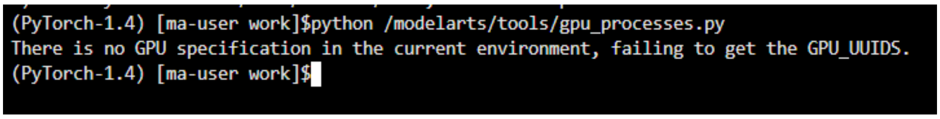
```
python /modelarts/tools/gpu_processes.py
```

   The following figure shows the case that the current process is using GPUs.



```
Using user ma-user
Ubuntu 18.04.6 LTS, CUDA-10.1
Tips:
1) Navigate to the target conda environment. For details, see /home/ma-user/README.
2) Copy (Ctrl+C) and paste (Ctrl+V) on the jupyter terminal.
3) Store your data in /home/ma-user/work, to which a persistent volume is mounted.
(PyTorch-1.4) [ma-user work]$python /modelarts/tools/gpu_processes.py
Fri Mar 17 11:27:17 2023
+-------------+----------+--------------------+------------------------+
| Processes:  |          |                    |                        |
| GPU         | PID      | Process name       | GPU Memory Usage       |
+=============+==========+====================+========================+
| 0           | 4608     | python             | 785 MiB                |
+-------------+----------+--------------------+------------------------+
| 0           | 4731     | python             | 785 MiB                |
+-------------+----------+--------------------+------------------------+
| 0           | 4860     | python             | 785 MiB                |
+-------------+----------+--------------------+------------------------+
| 0           | 5000     | python             | 785 MiB                |
+-------------+----------+--------------------+------------------------+
(PyTorch-1.4) [ma-user work]$
```

   The following figure shows the case that the current process is not using GPUs.



```
(PyTorch-1.4) [ma-user work]$python /modelarts/tools/gpu_processes.py
There is no GPU specification in the current environment, failing to get the GPU_UUIDS.
(PyTorch-1.4) [ma-user work]$
```

   Method 2:

   Open **/resource_info/gpu_usage.json** and view the processes that are using GPUs.

```
{
    <notebook name>: {
        <GPU0 UUID>: [
            {
                "pid": 2263,
                "processName": "python",
                "gpuMemoryUsage": "4935Mi"
            },
            {...}
        ]
        <GPU1 UUID>: [...]
    }
}
```

   If no process is using GPUs, the file may be unavailable or empty.

5.10.5 How Can I Obtain GPU Usage Through Code?

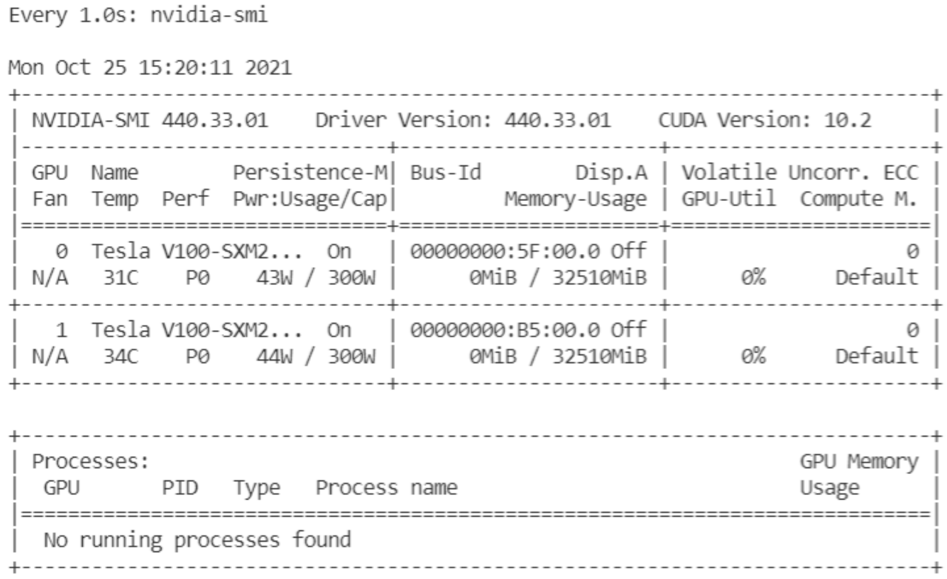
Run the shell or python command to obtain the GPU usage.

Using the shell Command

1. Run the **nvidia-smi** command.

   This operation relies on CUDA NVCC.

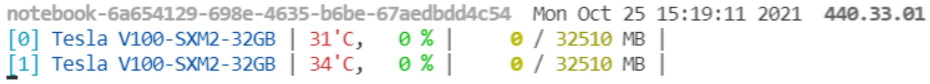
```
watch -n 1 nvidia-smi
```



```
Every 1.0s: nvidia-smi

Mon Oct 25 15:20:11 2021
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 440.33.01    Driver Version: 440.33.01    CUDA Version: 10.2     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:5F:00.0 Off |                    0 |
| N/A   31C    P0    43W / 300W |      0MiB / 32510MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla V100-SXM2...  On   | 00000000:B5:00.0 Off |                    0 |
| N/A   34C    P0    44W / 300W |      0MiB / 32510MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

2. Run the **gpustat** command.

```
pip install gpustat
gpustat -cp -i
```



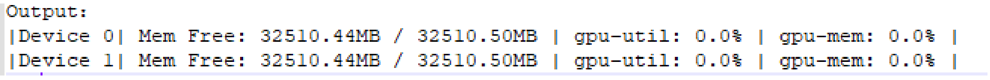
```
notebook-6a654129-698e-4635-b6be-67aedbdd4c54  Mon Oct 25 15:19:11 2021  440.33.01
[0] Tesla V100-SXM2-32GB | 31'C,   0 % |     0 / 32510 MB |
[1] Tesla V100-SXM2-32GB | 34'C,   0 % |     0 / 32510 MB |
```

   To stop the command execution, press **Ctrl+C**.

Using the python Command

1. Run the **nvidia-ml-py3** command (commonly used).

```
!pip install nvidia-ml-py3
import nvidia_smi
nvidia_smi.nvmlInit()
deviceCount = nvidia_smi.nvmlDeviceGetCount()
for i in range(deviceCount):
    handle = nvidia_smi.nvmlDeviceGetHandleByIndex(i)
    util = nvidia_smi.nvmlDeviceGetUtilizationRates(handle)
    mem = nvidia_smi.nvmlDeviceGetMemoryInfo(handle)
    print(f"|Device {i}| Mem Free: {mem.free/1024**2:5.2f}MB / {mem.total/1024**2:5.2f}MB | gpu-util: {util.gpu:3.1%} | gpu-mem: {util.memory:3.1%} |")
```

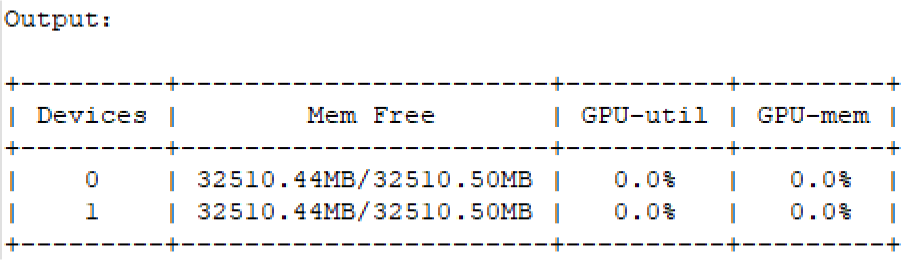


```
Output:
|Device 0| Mem Free: 32510.44MB / 32510.50MB | gpu-util: 0.0% | gpu-mem: 0.0% |
|Device 1| Mem Free: 32510.44MB / 32510.50MB | gpu-util: 0.0% | gpu-mem: 0.0% |
```

2. Run the **nvidia_smi**, **wapper**, and **prettytable** commands.

   Use the decorator to obtain the GPU usage in real time during model training.

```
def gputil_decorator(func):
    def wrapper(*args, **kwargs):
        import nvidia_smi
        import prettytable as pt

        try:
            table = pt.PrettyTable(['Devices','Mem Free','GPU-util','GPU-mem'])
            nvidia_smi.nvmlInit()
            deviceCount = nvidia_smi.nvmlDeviceGetCount()
            for i in range(deviceCount):
                handle = nvidia_smi.nvmlDeviceGetHandleByIndex(i)
                res = nvidia_smi.nvmlDeviceGetUtilizationRates(handle)
                mem = nvidia_smi.nvmlDeviceGetMemoryInfo(handle)
                table.add_row([i, f"{mem.free/1024**2:5.2f}MB/{mem.total/1024**2:5.2f}MB",
f"{res.gpu:3.1%}", f"{res.memory:3.1%}"])

        except nvidia_smi.NVMLError as error:
            print(error)

        print(table)
        return func(*args, **kwargs)
    return wrapper
```

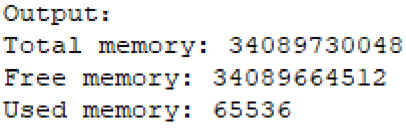


```
Output:

+----------+------------------------+-----------+----------+
| Devices  |        Mem Free        | GPU-util  | GPU-mem  |
+----------+------------------------+-----------+----------+
|    0     | 32510.44MB/32510.50MB  |   0.0%    |   0.0%   |
|    1     | 32510.44MB/32510.50MB  |   0.0%    |   0.0%   |
+----------+------------------------+-----------+----------+
```

3. Run the **pynvml** command.

   Run **nvidia-ml-py3** to directly obtain the nvml c-lib library, without using **nvidia-smi**. Therefore, this command is recommended.

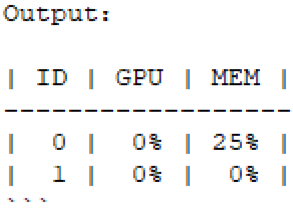
```
from pynvml import *
nvmlInit()
handle = nvmlDeviceGetHandleByIndex(0)
info = nvmlDeviceGetMemoryInfo(handle)
print("Total memory:", info.total)
print("Free memory:", info.free)
print("Used memory:", info.used)
```

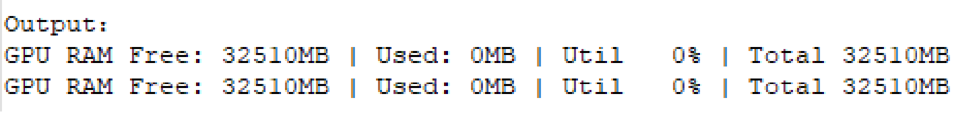


```
Output:
Total memory: 34089730048
Free memory: 34089664512
Used memory: 65536
```

4. Run the **gputil** command.

```
!pip install gputil
import GPUtil as GPU
GPU.showUtilization()
```



```
Output:

| ID | GPU | MEM |
------------------
|  0 |  0% | 25% |
|  1 |  0% |  0% |
```
```

```
import GPUtil as GPU
GPUs = GPU.getGPUs()
for gpu in GPUs:
    print("GPU RAM Free: {0:.0f}MB | Used: {1:.0f}MB | Util {2:3.0f}% | Total
{3:.0f}MB".format(gpu.memoryFree, gpu.memoryUsed, gpu.memoryUtil*100, gpu.memoryTotal))
```



```
Output:
GPU RAM Free: 32510MB | Used: 0MB | Util   0% | Total 32510MB
GPU RAM Free: 32510MB | Used: 0MB | Util   0% | Total 32510MB
```
```

When using a deep learning framework such as PyTorch or TensorFlow, you can also use the APIs provided by the framework for query.

5.10.6 Which Real-Time Performance Indicators of an Ascend Chip Can I View?

The real-time performance indicator that can be viewed is **npu-smi**, which is similar to **nvidia-smi** of a GPU chip.

5.10.7 Does the System Automatically Stop or Delete a Notebook Instance If I Do Not Enable Automatic Stop?

The answer to this question differs depending on the selected resource specifications.

- If you use free specifications, your notebook instance automatically stops after running for one hour. If the notebook instance is not started again within 72 hours, it will be deleted. Therefore, when using free specifications, pay attention to the running time and back your files up properly.
- If you use a paid public resource pool and do not enable automatic stop, the notebook instance does not automatically stop or is deleted.
- If you use a dedicated resource pool, the notebook instance does not automatically stop. However, if the dedicated resource pool is deleted, the notebook instance will be unavailable.

5.10.8 What Are the Relationships Between Files Stored in JupyterLab, Terminal, and OBS?

- Files stored in JupyterLab are the same as those in the work directory on the **Terminal** page. That is, the files are created on your notebook instances or synchronized from OBS.
- Notebook instances with OBS storage mounted can synchronize files from OBS to JupyterLab using the JupyterLab upload and download functions. The files on the **Terminal** page are the same as those in JupyterLab.
- Notebook instances with EVS storage mounted can read files from OBS to JupyterLab using the MoXing API or SDKs. The files on the **Terminal** page are the same as those in JupyterLab.

5.10.9 How Do I Migrate Data from an Old-Version Notebook Instance to a New-Version One?

The notebook of the old version will be discontinued. Use the new version. This section describes how to migrate data from a notebook instance of the old version to a notebook instance of the new version.

Storage Differences Between the Old and New Versions

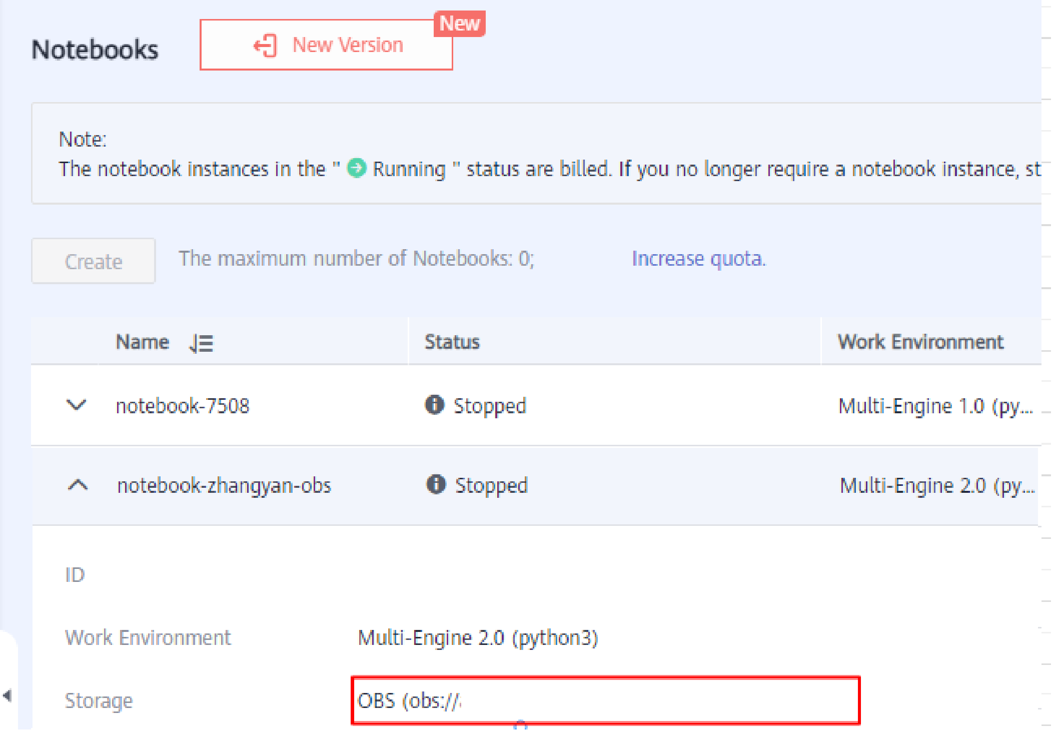
Table 5-1 Storage supported by notebook of the old and new versions

Storage	Old-Version Notebook	New-Version Notebook	Description
OBS	Supported	Not supported	OBS is a storage system, not a file system. In old-version notebook, remote replication and local replication of OBS data may be confused, leading to issues in controlling operations on data. Therefore, OBS mounting is removed from notebook of the new version. You can flexibly obtain and operate OBS data using code.
OBS parallel file system	Not supported	Supported	The new-version notebook allows dynamic mounting of OBS parallel file systems. You can mount storage on the details page of a running notebook instance. Data migration from the old version to the new version is not involved.
EVS	Supported	Supported	EVS disks can be attached to notebook instances of both the old and new versions. Data stored in the old version needs to be migrated to the new version.
SFS	Not supported	Supported	SFS is used in dedicated resource pools. This function has been discontinued in notebook of the old version. Therefore, data migration is not involved.

OBS Used in Notebook of the Old Version

When notebook instances of the old version use OBS for storage, data is stored in OBS and does not need to be migrated. After a new-version notebook instance is created, directly use the data in the OBS directory. For details, see **How Do I Read and Write OBS Files in a Notebook Instance?**

Figure 5-19 OBS used in notebook of the old version



Notebooks

New Version New

Note:
The notebook instances in the " Running " status are billed. If you no longer require a notebook instance, st

Create The maximum number of Notebooks: 0; Increase quota.

Name	Status	Work Environment
notebook-7508	Stopped	Multi-Engine 1.0 (py...
notebook-zhangyan-obs	Stopped	Multi-Engine 2.0 (py...

ID

Work Environment Multi-Engine 2.0 (python3)

Storage OBS (obs://

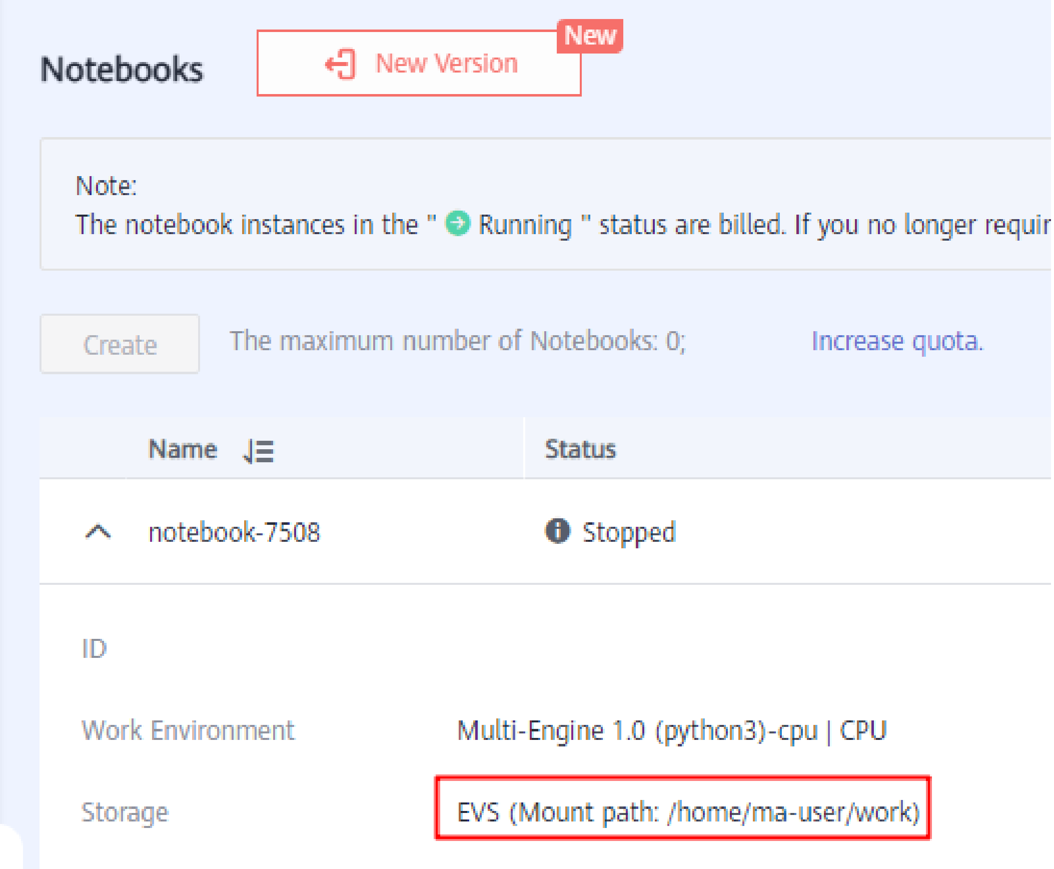
EVS Used in Notebook of the Old Version

If EVS disks are attached to a notebook instance of the old version for storing data, back up and migrate the EVS data to a notebook instance of the new version.

- If the volume of data stored in EVS is small, download the data to a local directory, create a notebook instance of the new version, and upload the data to the new notebook instance.
- If a large amount of data is stored in EVS, upload the data to an OBS bucket. After a notebook instance of the new version is created, read data from the the OBS bucket.

For more details, see **Uploading and Downloading Data in Notebook**.

Figure 5-20 EVS storage used in notebook of the old version



Notebooks

Note:
The notebook instances in the " Running " status are billed. If you no longer requir

The maximum number of Notebooks: 0;

Name	Status
notebook-7508	Stopped

ID

Work Environment Multi-Engine 1.0 (python3)-cpu | CPU

Storage EVS (Mount path: /home/ma-user/work)

5.10.10 How Do I Use the Datasets Created on ModelArts in a Notebook Instance?

Datasets created on ModelArts are stored in OBS. To use these datasets in a notebook instance, download them from OBS to the notebook instance.

For details, see **How Do I Upload a File from a Notebook Instance to OBS or Download a File from OBS to a Notebook Instance?**

5.10.11 pip and Common Commands

pip is a common Python package management tool. It allows you to search for, download, install, and uninstall Python packages.

Common pip commands:

```
pip --help # Obtain help information.
pip install SomePackage==XXXX # Install a specified version.
pip install SomePackage # Install the latest version.
pip uninstall SomePackage # Uninstall a software version.
```

For other commands, run the **pip --help** command.

5.10.12 What Are Sizes of the /cache Directories for Different Notebook Specifications in DevEnviron?

When creating a notebook instance, you can select CPUs, GPUs, or Ascend based on the data volume.

ModelArts mounts disks to **/cache**. You can use this directory to store temporary files. The **/cache** directory shares resources with the code directory. The directory size varies depending on resource specifications.

No disks can be mounted to **/cache** for CPUs. When only one GPU or Ascend card is used, the **/cache** directory size is limited to 500 GB. If multiple GPUs or Ascend cards are used, the **/cache** directory size is limited to 3 TB and calculated using the following formula: **/cache** directory size = Number of cards x 500 GB. For details, see **Table 5-2**.

Table 5-2 /cache directory sizes for different notebook specifications

Specification	/cache Directory Size
GPU, 0.25 cards	500 GB x 0.25
GPU, 0.5 cards	500 GB x 0.5
GPU, 1 card	500 GB
GPU, dual cards	500 GB x 2
GPU, four cards	500 GB x 4
GPU, eight cards	3 TB
Ascend, single card	500 GB
Ascend, dual cards	500 GB x 2
Ascend, four cards	500 GB x 4
Ascend, eight cards	3 TB
CPU	N/A

5.10.13 How Do I Isolate IAM Users for Using Development Environments?

There are two methods to prevent IAM users from viewing, modifying, or deleting notebook instances created by others:

- Solution 1: Delete the modelarts:notebook:listAllNotebooks permission.
- Solution 2: Use **workspaces** to isolate resources. As an enterprise user, you can submit the request for enabling the workspace function to your technical support.

5.10.14 What Is the Impact of Resource Overcommitment on Notebook Instances?

Notebook overcommitment refers to the sharing of GPUs and memory within a node. To fully utilize resources, they are overcommitted in dedicated pools.

Example: A dedicated pool has one CPU node with 8 vCPUs and 64 GB memory. If you create a notebook instance with 2 vCPUs and 8 GB memory, a maximum of 6.67 notebook instances (8 vCPUs/(2 vCPUs x 0.6)) can be started due to overcommitment with an overcommitment ratio of 0.6. In this case, at least 1.2 vCPUs are required for starting the notebook instance, and a maximum of 2 vCPUs are used for running the notebook instance. Similarly, at least 4.8 GB memory is required, and a maximum of 8 GB memory is used for running the notebook instance.

Instances may be forcibly terminated due to overcommitment. For example, if six instances with 2 vCPUs are started on an 8 vCPUs node and the CPU usage of one instance exceeds the upper limit (8 vCPUs) of the node, Kubernetes forcibly terminates the instance that uses the most resources.

Do not overcommit resources as it may result in instance restart.

6 Training Jobs

6.1 Functional Consulting

6.1.1 What Are the Format Requirements for Algorithms Imported from a Local Environment?

ModelArts supports the import of locally developed algorithms. The format requirements are as follows:

- Any programming language is supported.
- The boot file must be in the format of **.py** or **.pyc**.
- The number of files (including files and folders) cannot exceed 1,024.
- The total file size cannot exceed 5 GB.

6.1.2 What Are the Solutions to Underfitting?

1. Increasing model complexity
   - For an algorithm, add more high-order items to the regression model, improve the depth of the decision tree, or increase the number of hidden layers and hidden units of the neural network to increase model complexity.
   - Discard the original algorithm and use a more complex algorithm or model. For example, use the neural network to replace the linear regression, and use the random forest to replace the decision tree.
2. Adding more features to make input data more expressive
   - Feature mining is very important. Specifically, features with strong expression capabilities can outperform a large number of features with weak expression capabilities.
   - Feature quality is the focus.
   - To explore features with strong expression capabilities, you must have an in-depth understanding of data and application scenarios, which depends on experience.

3. Adjusting parameters and hyperparameters
   - Neural network: learning rate, learning attenuation rate, number of hidden layers, number of units in a hidden layer, β1 and β2 parameters in the Adam optimization algorithm, and batch_size
   - Other algorithms: number of trees in the random forest, number of clusters in *k*-means, and regularization parameter λ
4. Adding training data (not recommended)

   Underfitting is usually caused by weak model learning capabilities. Adding data cannot significantly increase the training effect.
5. Reducing regularization constraints

   Regularization aims to prevent model overfitting. If a model is underfitting instead of overfitting, reduce the regularization parameter **λ** or directly remove the regularization item.

6.1.3 What Are the Precautions for Switching Training Jobs from the Old Version to the New Version?

The differences between the new version and the old version lie in:

- **Differences in Training Job Creation**
- **Differences in Training Code Adaptation**
- **Differences in Built-in Training Engines**

Differences in Training Job Creation

- In earlier versions, you can create a training job using **Algorithm Management**, **Frequently-used**, and **Custom**.
- In the new version, you can create a training job using **Custom algorithm**or **My algorithm**.

The new version reorganizes the algorithms to help you find them more easily. Existing training jobs are not affected.

- The saved algorithms in **Algorithm Management** in the old version are in **My algorithm** in the new version.
- The **Frequently-used** in the old version is the **Custom algorithm** in the new version. Select **Preset image** for **Boot Mode** when you create jobs using the new version.
- The **Custom** in the old version is the **Custom algorithm** in the new version. Select **Custom image** for **Boot Mode** when you create jobs using the new version.

Differences in Training Code Adaptation

In the old version, you are required to configure data input and output as follows:

```
# Parse CLI parameters.
import argparse
parser = argparse.ArgumentParser(description='MindSpore Lenet Example')
parser.add_argument('--data_url', type=str, default="./Data",
                    help='path where the dataset is saved')
parser.add_argument('--train_url', type=str, default="./Model", help='if is test, must provide\
                    path where the trained ckpt file')
```

```
args = parser.parse_args()
...
# Download data to your local container. In the code, local_data_path specifies the training input path.
mox.file.copy_parallel(args.data_url, local_data_path)
...
# Upload the local container data to the OBS path.
mox.file.copy_parallel(local_output_path, args.train_url)
```

In the new version, you only need to configure training input and output. In the code, **arg.data_url** and **arg.train_url** are used as local paths. For details, see **Developing a Custom Script**.

```
# Parse CLI parameters.
import argparse
parser = argparse.ArgumentParser(description='MindSpore Lenet Example')
parser.add_argument('--data_url', type=str, default="./Data",
                    help='path where the dataset is saved')
parser.add_argument('--train_url', type=str, default="./Model", help='if is test, must provide\
                    path where the trained ckpt file')
args = parser.parse_args()
...
# The downloaded code does not need to be set. Use data_url and train_url for data training and output.
# Download data to your local container. In the code, local_data_path specifies the training input path.
#mox.file.copy_parallel(args.data_url, local_data_path)
...
# Upload the local container data to the OBS path.
#mox.file.copy_parallel(local_output_path, args.train_url)
```

Differences in Built-in Training Engines

- In the new version, MoXing 2.0.0 or later is installed by default for built-in training engines.
- In the new version, Python 3.7 or later is used for built-in training engines.
- In the new image, the default home directory has been changed from **/home/work** to **/home/ma-user**. Check whether the training code contains hard coding of **/home/work**.
- Built-in training engines are different between the old and new versions. Commonly used built-in training engines have been upgraded in the new version.

  To use a training engine in the old version, switch to the old version. **Table 6-1** lists the differences between the built-in training engines in the old and new versions.

Table 6-1 Differences between the built-in training engines in the old and new versions

Runtime Environment	Built-in Training Engine and Version	Old Version	New Version
TensorFlow	TensorFlow-1.8.0	√	x
	TensorFlow-1.13.1	√	Coming soon
	TensorFlow-2.1.0	√	√
MXNet	MXNet-1.2.1	√	x

Runtime Environment	Built-in Training Engine and Version	Old Version	New Version
Caffe	Caffe-1.0.0	√	x
Spark MLlib	Spark-2.3.2	√	x
Ray	Ray-0.7.4	√	x
XGBoost with scikit-learn	XGBoost-0.80-Sklearn-0.18.1	√	x
PyTorch	PyTorch-1.0.0	√	x
	PyTorch-1.3.0	√	x
	PyTorch-1.4.0	√	x
	PyTorch-1.8.0	x	√
MPI	MindSpore-1.3.0	x	√
Horovod	Horovod_0.20.0-TensorFlow_2.1.0	x	√
	horovod_0.22.1-pytorch_1.8.0	x	√
MindSpore-GPU	MindSpore-1.1.0	√	x
	MindSpore-1.2.0	√	x

6.1.4 How Do I Obtain a Trained ModelArts Model?

Models generated using ModelArts ExeML can be deployed only on ModelArts and cannot be downloaded to your local PC.

Models trained using a custom or subscription algorithm are stored in specified OBS paths for you to download.

6.1.5 How Do I Set the Runtime Environment of the AI Engine Scikit_Learn 0.18.1?

On the **Algorithm Management** page of ModelArts, click **Create**. On the **Create Algorithm** page, select **Show Old Engines** for **AI Engine**. Then, select XGBoost-Sklearn.

For details about how to create an algorithm in ModelArts, see **Creating an Algorithm**.

For details about how to create a training job, see **Creating a Training Job**.

6.1.6 Must the Hyperparameters Optimized Using a TPE Algorithm Be Categorical?

TPE algorithms do not impose requirements on the types of optimized hyperparameters. However, to reduce resource utilization for common users, ModelArts console requires that TPE hyperparameters must be of the floating type. To use both discrete and continuous parameters, call the REST API.

6.1.7 What Is TensorBoard Used for in Model Visualization Jobs?

Visualization jobs are powered by TensorBoard. For details about TensorBoard functions, see the **TensorBoard official website**.

6.1.8 How Do I Obtain RANK_TABLE_FILE on ModelArts for Distributed Training?

ModelArts automatically provides the **RANK_TABLE_FILE** file for you. Obtain the file location through environment variables.

- Open the notebook terminal and run the following command to view **RANK_TABLE_FILE**:
  ```
  env | grep RANK
  ```
- In a training job, add the following code to the first line of the training startup script to print the value of **RANK_TABLE_FILE**:
  ```
  os.system('env | grep RANK')
  ```

6.1.9 How Do I Obtain the CUDA and cuDNN Versions of a Custom Image?

Obtain a CUDA version:

```
cat /usr/local/cuda/version.txt
```

Obtain a cuDNN version:

```
cat /usr/local/cuda/include/cudnn.h | grep CUDNN_MAJOR -A 2
```

6.1.10 How Do I Obtain a MoXing Installation File?

MoXing installation files cannot be downloaded or installed by users. The MoXing installation package is preset in ModelArts notebook and training job images, and can be directly used.

6.1.11 In a Multi-Node Training, the TensorFlow PS Node Functioning as a Server Will Be Continuously Suspended. How Does ModelArts Determine Whether the Training Is Complete? Which Node Is a Worker?

In a TensorFlow-powered distributed training, the PS task and worker task are started. The worker task is a key task. ModelArts will use a process exit code of the worker task to determine whether the training job is complete.

A task name will be used to determine which node is a worker. A Volcano job is issued for training, which contains a PS task and a worker task. The startup commands of the two tasks are different. The hyperparameter **task_name** will be automatically generated, which is **ps** for the PS task and **worker** for the worker task.

6.1.12 How Do I Install MoXing for a Custom Image of a Training Job?

To prevent automatic installation of MoXing from affecting the package environment in the custom image, manually install MoXing for the custom image. MoXing is stored in the **/home/ma-user/modelarts/package/** directory after the job is started. Before using MoXing, run the following code to install it:

```
import os
os.system("pip install /home/ma-user/modelarts/package/moxing_framework-*.whl")
```

NOTE

This case applies only to the training environment.

6.1.13 An IAM User Cannot Select an Existing SFS Turbo File System When Using a Dedicated Resource Pool to Create a Training Job

The IAM user cannot view existing SFS Turbo file systems due to insufficient permissions. To grant access, the user group to which the IAM user belongs must be given SFS FullAccess and SFS Turbo FullAccess permissions.

6.2 Reading Data During Training

6.2.1 How Do I Configure the Input and Output Data for Training Models on ModelArts?

ModelArts allows you to upload a custom algorithm for creating training jobs. **Create the algorithm** and upload it to an OBS bucket. For details about how to create an algorithm, see **Creating an Algorithm**. For details about how to create a training job, see **Creating a Training Job**.

Parsing Input and Output Paths

When a ModelArts model reads data stored in OBS or outputs data to a specified OBS path, perform the following operations to configure the input and output data:

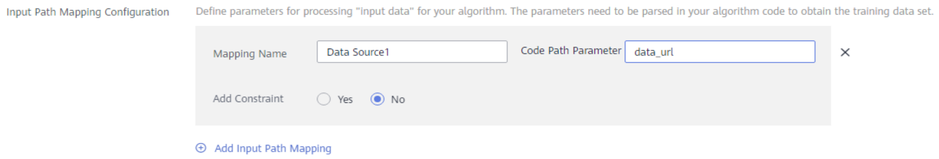
1. Parse the input and output paths in the training code. The following method is recommended:

```
import argparse
# Create a parsing task.
parser = argparse.ArgumentParser(description="train mnist",
                                 formatter_class=argparse.ArgumentDefaultsHelpFormatter)
# Add parameters.
parser.add_argument('--train_url', type=str,
                   help='the path model saved')
parser.add_argument('--data_url', type=str, help='the training data')
# Parse the parameters.
args, unknown = parser.parse_known_args()
```

After the parameters are parsed, use **data_url** and **train_url** to replace the paths to the data source and the data output, respectively.

2. When using a preset image to create a custom algorithm, configure the input and output parameters on the **Create Algorithm** page based on code settings.
   - Training data is a must for algorithm development. It is a good practice to set the input parameter to **data_url**, which is the data input source. You can also customize the input parameter based on the algorithm code in the previous step.

     Figure 6-1 Parsing the input path parameter **data_url**

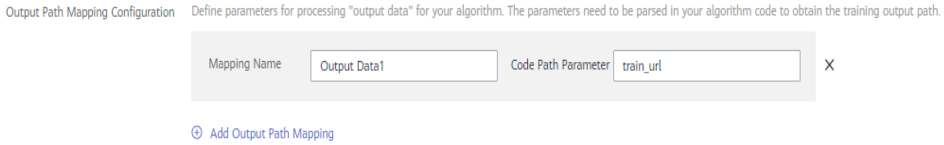
     Input Path Mapping Configuration

     Define parameters for processing "input data" for your algorithm. The parameters need to be parsed in your algorithm code to obtain the training data set.

     Mapping Name: Data Source1
     Code Path Parameter: data_url
     Add Constraint: ○ Yes ● No

   - After model training is complete, the trained model and the output information must be stored in an OBS path. By default, the output data is **Output Data** and the code path parameter is **train_url** (customizable).

     Figure 6-2 Parsing the output path parameter **train_url**

     Output Path Mapping Configuration

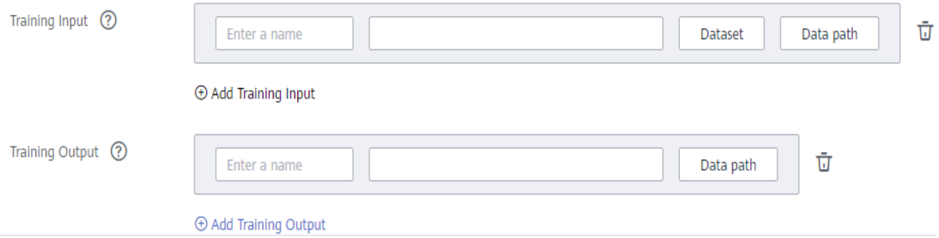
     Define parameters for processing "output data" for your algorithm. The parameters need to be parsed in your algorithm code to obtain the training output path.

     Mapping Name: Output Data1
     Code Path Parameter: train_url

3. When creating a training job, configure the input and output paths.

   Select the OBS path or dataset path as the training input, and the OBS path as the output.

Figure 6-3 Setting training input and output



Training Input ?: Enter a name

Training Output ?: Enter a name

6.2.2 How Do I Improve Training Efficiency While Reducing Interaction with OBS?

Scenario Description

When ModelArts is used for custom deep learning training, training data is usually stored in OBS. If the volume of training data is large (for example, greater than

200 GB), a GPU resource pool is required for training each time, resulting in low training efficiency.

To improve training efficiency while reducing interaction with OBS, perform the following operations for optimization.

Optimization Principles

For the GPU resource pool provided by ModelArts, 500 GB NVMe SSDs are attached to each training node for free. The SSDs are attached to the **/cache** directory. The lifecycle of data in the **/cache** directory is the same as that of a training job. After the training job is complete, all content in the **/cache** directory is cleared to release space for the next training job. Therefore, you can copy data from OBS to the **/cache** directory during training so that data can be read from the **/cache** directory each time until the training is complete. After the training is complete, content in the **/cache** directory will be automatically cleared.

Optimization Methods

TensorFlow code is used as an example.

The following is code before optimization:

```
...
tf.flags.DEFINE_string('data_url', '', 'dataset directory.')
FLAGS = tf.flags.FLAGS
mnist = input_data.read_data_sets(FLAGS.data_url, one_hot=True)
```

The following is an example of the optimized code. Data is copied to the **/cache** directory.

```
...
tf.flags.DEFINE_string('data_url', '', 'dataset directory.')
FLAGS = tf.flags.FLAGS
import moxing as mox
TMP_CACHE_PATH = '/cache/data'
mox.file.copy_parallel('FLAGS.data_url', TMP_CACHE_PATH)
mnist = input_data.read_data_sets(TMP_CACHE_PATH, one_hot=True)
```

6.2.3 Why the Data Read Efficiency Is Low When a Large Number of Data Files Are Read During Training?

If a dataset contains a large number of data files (massive small files) and data is stored in OBS, files need to be repeatedly read from OBS during training. As a result, the training process is waiting for reading files, resulting in low read efficiency.

Solution

1. Compress the massive small files into a package on your local PC, for example, a .zip package.
2. Upload the package to OBS.
3. During training, directly download this package from OBS to the **/cache** directory of your local PC. Perform this operation only once.

   For example, you can use mox.file.copy_parallel to download the .zip package to the **/cache** directory, decompress the package, and then read files for training.

```
...
tf.flags.DEFINE_string('<obs_file_path>/data.zip', '', 'dataset directory.')
FLAGS = tf.flags.FLAGS
import os
import moxing as mox
TMP_CACHE_PATH = '/cache/data'
mox.file.copy_parallel('FLAGS.data_url', TMP_CACHE_PATH)
zip_data_path = os.path.join(TMP_CACHE_PATH, '*.zip')
unzip_data_path = os.path.join(TEMP_CACHE_PATH, 'unzip')
# You can also decompress .zip Python packages.
os.system('unzip '+ zip_data_path + ' -d ' + unzip_data_path))
mnist = input_data.read_data_sets(unzip_data_path, one_hot=True)
```

6.3 Compiling the Training Code

6.3.1 How Do I Create a Training Job When a Dependency Package Is Referenced by the Model to Be Trained?

Store the **pip-requirements.txt** file in the training code directory.

NOTE

Any one of the following file names can be used. This section uses **pip-requirements.txt** as an example.

- pip-requirement.txt
- pip-requirements.txt
- requirement.txt
- requirements.txt

Before the training boot file is executed, the system automatically runs the following command to install the specified Python packages:

```
pip install -r pip-requirements.txt
```

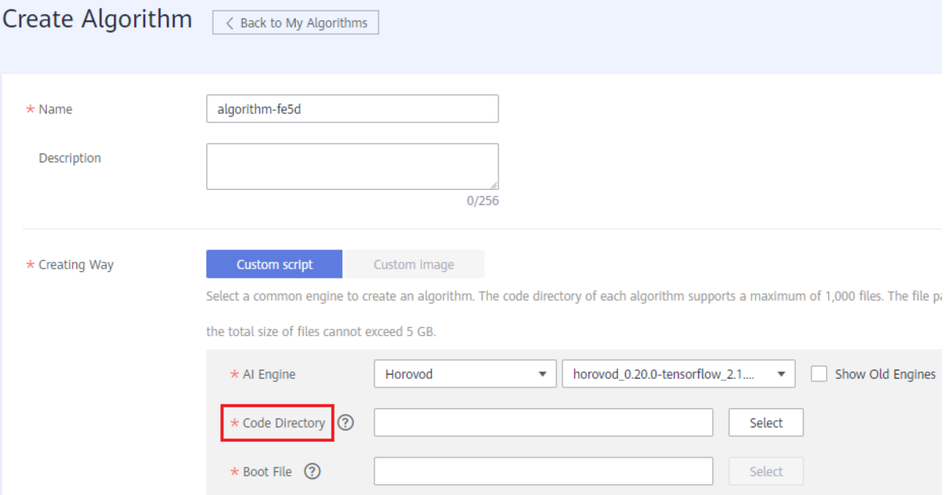
- For details about the code directory, see **Storing the Installation File in the Code Directory**.
- For details about the specifications of **pip-requirements.txt**, see **Installation File Specifications**.

Storing the Installation File in the Code Directory

ModelArts allows you to install third-party dependency packages during model training in either of the following ways:

- Method 1 (recommended): **Before creating an algorithm**, **store the required files or installation packages** in the code directory.

Figure 6-4 Creating an algorithm



Create Algorithm

< Back to My Algorithms

* Name: algorithm-fe5d

Description: 0/256

* Creating Way: Custom script | Custom image

Select a common engine to create an algorithm. The code directory of each algorithm supports a maximum of 1,000 files. The file p

the total size of files cannot exceed 5 GB.

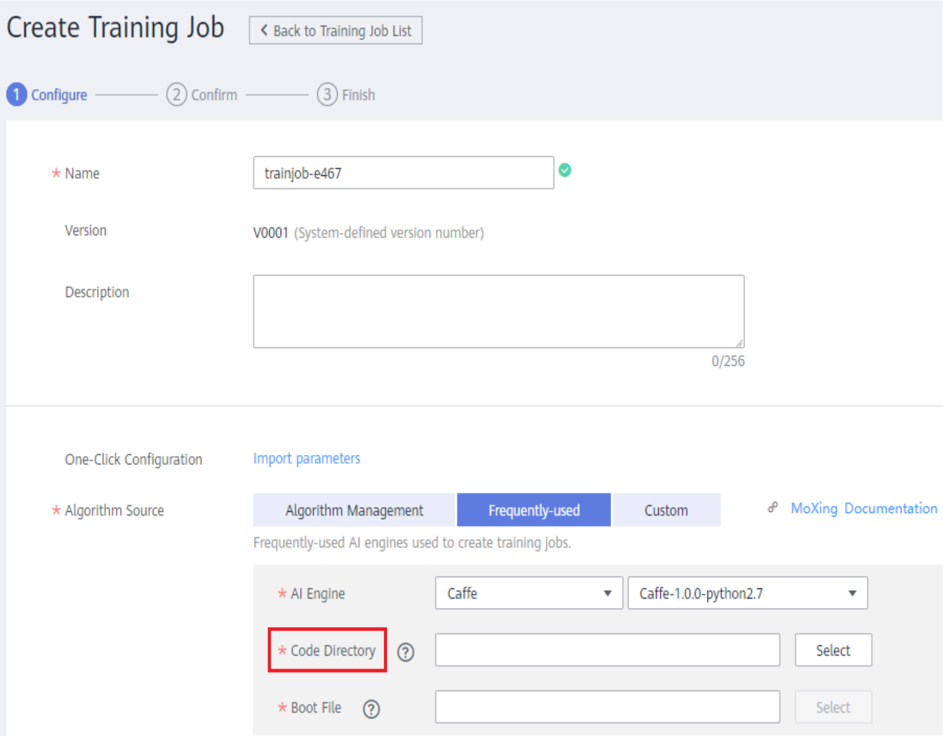
* AI Engine: Horovod | horovod_0.20.0-tensorflow_2.1.... □ Show Old Engines

* Code Directory ⓘ Select

* Boot File ⓘ Select

- Method 2: Before using a common framework to create a training job, **store the required files or installation packages** in the code directory. (This function will be unavailable soon.)

Figure 6-5 Using a common framework to create an algorithm



Create Training Job

< Back to Training Job List

1 Configure —— 2 Confirm —— 3 Finish

* Name: trainjob-e467

Version: V0001 (System-defined version number)

Description: 0/256

One-Click Configuration: Import parameters

* Algorithm Source: Algorithm Management | Frequently-used | Custom   MoXing Documentation

Frequently-used AI engines used to create training jobs.

* AI Engine: Caffe | Caffe-1.0.0-python2.7

* Code Directory ⓘ Select

* Boot File ⓘ Select

Installation File Specifications

The installation file varies depending on the dependency package type.

- **Open-source installation packages**

NOTE

Installation using the source code from GitHub is not supported.

Create a file named **pip-requirements.txt** in the code directory, and specify the name and version number of the dependency package in the file. The format is *[Package name]*==*[Version]*.

Take for example, an OBS path specified by **Code Dir** that contains model files and the **pip-requirements.txt** file. The code directory structure would be as follows:

```
|---OBS path to the model boot file
    |---model.py              #Model boot file
    |---pip-requirements.txt   #Defined configuration file, which specifies the name and version of the
dependency package
```

The following shows the content of the **pip-requirements.txt** file:

```
alembic==0.8.6
bleach==1.4.3
click==6.6
```

- **WHL packages**

  If the training background does not support the download of open source installation packages or use of user-compiled WHL packages, the system cannot automatically download and install the package. In this case, place the WHL package in the code directory, create a file named **pip-requirements.txt**, and specify the name of the WHL package in the file. The dependency package must be a **.whl** file.

  Take for example, an OBS path specified by **Code Dir** that contains model files, the **.whl** file, and the **pip-requirements.txt** file. The code directory structure would be as follows:

```
|---OBS path to the model boot file
    |---model.py              #Model boot file
    |---XXX.whl               #Dependency package. If multiple dependencies are required, place multiple
dependency packages here.
    |---pip-requirements.txt   #Defined configuration file, which specifies the name of the dependency
package
```

  The following shows the content of the **pip-requirements.txt** file:

```
numpy-1.15.4-cp36-cp36m-manylinux1_x86_64.whl
tensorflow-1.8.0-cp36-cp36m-manylinux1_x86_64.whl
```

6.3.2 What Is the Common File Path for Training Jobs?

The path to the training environment and the code directory in the container are generally obtained using the environment variable **${MA_JOB_DIR}**, which is **/home/ma-user/modelarts/user-job-dir**.

6.3.3 How Do I Install a Library That C++ Depends on?

A third-party library may be used during job training. The following uses C++ as an example to describe how to install a third-party library.

1. Download source code to a local PC and upload it to OBS. For details about how to upload a file using OBS Browser, see **Uploading a File**.
2. Use MoXing to copy the source code uploaded to OBS to a notebook instance in the development environment.

   The following is a code example for copying data to a notebook instance in a development environment running on an EVS:

```
import moxing as mox
mox.file.make_dirs('/home/ma-user/work/data')
mox.file.copy_parallel('obs://bucket-name/data', '/home/ma-user/work/data')
```

3. On the **Files** tab page of the **Jupyter** page, click **New** and select **Terminal**. Run the following command to go to the target path, and check whether the source code has been downloaded, that is, whether the **data** file exists.
   ```
   cd /home/ma-user/work
   ls
   ```
4. Compile code in **Terminal** based on service requirements.
5. Use MoXing to copy the compilation results to OBS. The following is a code example.
   ```
   import moxing as mox
   mox.file.make_dirs('/home/ma-user/work/data')
   mox.file.copy_parallel('/home/ma-user/work/data', 'obs://bucket-name/file)
   ```
6. During training, use MoXing to copy the compilation result from OBS to the container. The following is a code example.
   ```
   import moxing as mox
   mox.file.make_dirs('/cache/data')
   mox.file.copy_parallel('obs://bucket-name/data', '/cache/data')
   ```

6.3.4 How Do I Check Whether a Folder Copy Is Complete During Job Training?

In the script for training job boot file, run the following commands to obtain the sizes of the copied folders and the folders to be copied. Then determine whether folder copy is complete based on the command output.

```
import moxing as mox
mox.file.get_size('obs://bucket_name/obs_file',recursive=True)
```

get_size indicates the size of the file or folder to be obtained. **recursive=True** indicates that the type is folder. **True** indicates that the type is folder, and **False** indicates that the type is file.

If the command output is consistent, the folder copy is complete. If the command output is inconsistent, the folder copy is not complete.

6.3.5 How Do I Load Some Well Trained Parameters During Job Training?

During job training, some parameters need to be loaded from a pre-trained model to initialize the current model. You can use the following methods to load the parameters:

1. View all parameters by using the following code.
   ```
   from moxing.tensorflow.utils.hyper_param_flags import mox_flags
   print(mox_flags.get_help())
   ```
2. Specify the parameters to be restored during model loading. **checkpoint_include_patterns** is the parameter that needs to be restored, and **checkpoint_exclude_patterns** is the parameter that does not need to be restored.
   ```
   checkpoint_include_patterns: Variables names patterns to include when restoring checkpoint. Such as: conv2d/weights.
   checkpoint_exclude_patterns: Variables names patterns to include when restoring checkpoint. Such as: conv2d/weights.
   ```
3. Specify a list of parameters to be trained. **trainable_include_patterns** is a list of parameters that need to be trained, and **trainable_exclude_patterns** is a list of parameters that do not need to be trained.
   ```
   --trainable_exclude_patterns: Variables names patterns to exclude for trainable variables. Such as: conv1,conv2.
   ```

```
--trainable_include_patterns: Variables names patterns to include for trainable variables. Such as: logits.
```

6.3.6 How Do I Obtain Training Job Parameters from the Boot File of the Training Job?

Training job parameters can be automatically generated in the background or you can enter them manually. To obtain training job parameters:

1. When a training job is created, **train_url** in the running parameters of the training job indicates where the training results are output to, and **data_url** indicates a data source. The **test** parameter is entered manually.

Figure 6-6 Creating a training job of the new version

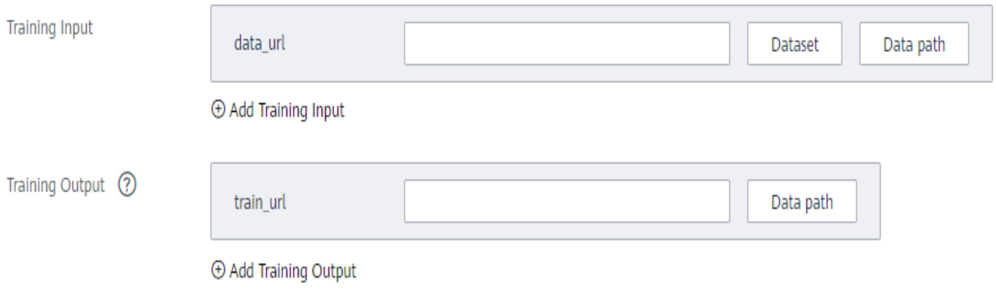
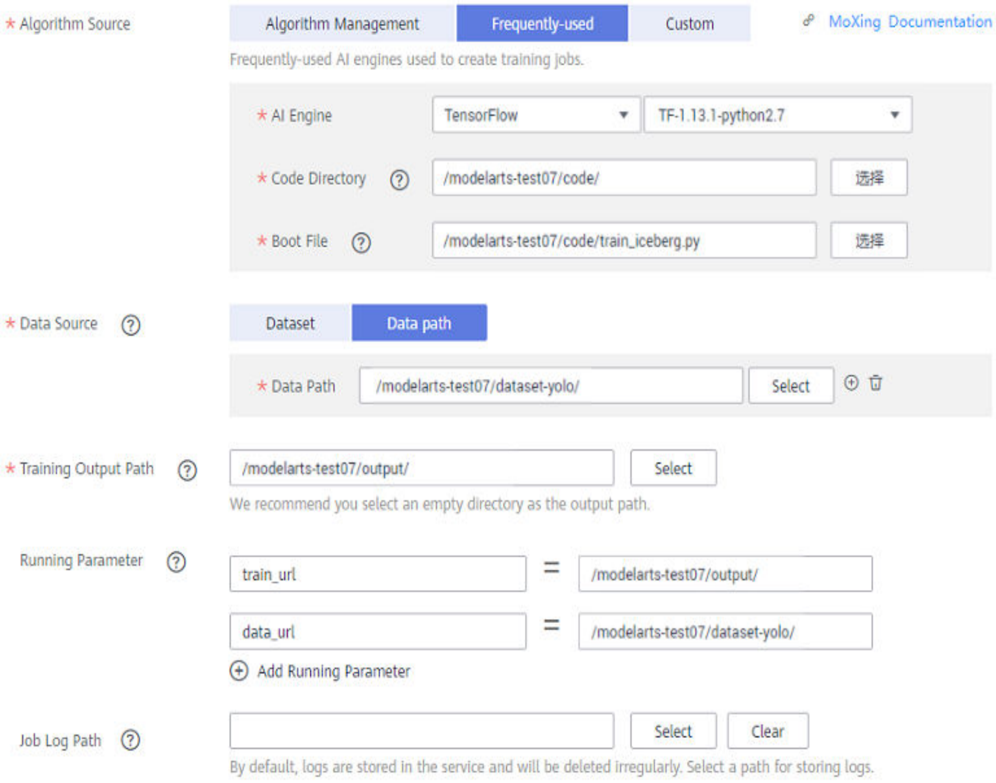


Figure 6-7 Creating a training job of the old version



* Algorithm Source: Frequently-used

Frequently-used AI engines used to create training jobs.

* AI Engine: TensorFlow TF-1.13.1-python2.7

* Code Directory: /modelarts-test07/code/

* Boot File: /modelarts-test07/code/train_iceberg.py

* Data Source: Data path

* Data Path: /modelarts-test07/dataset-yolo/

* Training Output Path: /modelarts-test07/output/

We recommend you select an empty directory as the output path.

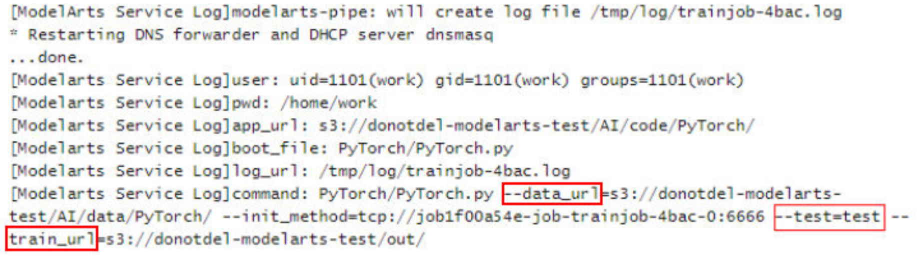
Running Parameter:
train_url = /modelarts-test07/output/
data_url = /modelarts-test07/dataset-yolo/

Job Log Path:

By default, logs are stored in the service and will be deleted irregularly. Select a path for storing logs.

2. After the training job is executed, you can click the job name in the training job list to view its details. You can obtain the parameter input mode from logs, as shown in **Figure 6-8**.

Figure 6-8 Viewing logs



```
[ModelArts Service Log]modelarts-pipe: will create log file /tmp/log/trainjob-4bac.log
* Restarting DNS forwarder and DHCP server dnsmasq
...done.
[Modelarts Service Log]user: uid=1101(work) gid=1101(work) groups=1101(work)
[Modelarts Service Log]pwd: /home/work
[Modelarts Service Log]app_url: s3://donotdel-modelarts-test/AI/code/PyTorch/
[Modelarts Service Log]boot_file: PyTorch/PyTorch.py
[Modelarts Service Log]log_url: /tmp/log/trainjob-4bac.log
[Modelarts Service Log]command: PyTorch/PyTorch.py --data_url=s3://donotdel-modelarts-
test/AI/data/PyTorch/ --init_method=tcp://job1f00a54e-job-trainjob-4bac-0:6666 --test=test --
train_url=s3://donotdel-modelarts-test/out/
```

3. To obtain the values of **train_url**, **data_url**, and **test** during training, add the following code to the boot file of the training job:

```
import argparse
parser = argparse.ArgumentParser()
parser.add_argument('--data_url', type=str, default=None, help='test')
parser.add_argument('--train_url', type=str, default=None, help='test')
parser.add_argument('--test', type=str, default=None, help='test')
```

6.3.7 Why Can't I Use os.system ('cd xxx') to Access the Corresponding Folder During Job Training?

If you cannot access the corresponding folder by using **os.system('cd xxx')** in the boot script of the training job, you are advised to use the following method:

```
import os
os.chdir('/home/work/user-job-dir/xxx')
```

6.3.8 How Do I Invoke a Shell Script in a Training Job to Execute the .sh File?

ModelArts enables you to invoke a shell script, and you can use Python to invoke **.sh**. The procedure is as follows:

1. Upload the **.sh** script to an OBS bucket. For example, upload the **.sh** script to **/bucket-name/code/test.sh**.
2. Create the **.py** file on a local PC, for example, **test.py**. The background automatically downloads the code directory to the **/home/work/user-job-dir/** directory of the container. Therefore, you can invoke the **.sh** file in the **test.py** boot file as follows:

```
import os
os.system('bash /home/work/user-job-dir/code/test.sh')
```

3. Upload **test.py** to OBS. Then the file storage path is **/bucket-name/code/test.py**.
4. When creating a training job, set the code directory to **/bucket-name/code/**, and the boot file directory to **/bucket-name/code/test.py**.

After the training job is created, you can use Python to invoke the **.sh** file.

6.3.9 How Do I Obtain the Dependency File Path to be Used in Training Code?

Since locally developed code must be uploaded to the ModelArts backend, you may set an invalid dependency file path. A recommended general solution to this problem is that you to use the OS API to obtain the absolute path of the dependency files.

The following shows an example of obtaining the path of dependency files in other folders using the OS API.

File directory structure:

```
project_root                #Root directory of code
 └─bootfile.py              #Boot file
 └─otherfileDirectory        #Directory of dependency files
      └─otherfile.py        #Dependency files
```

Add the following code to the boot file to obtain the path (**otherfile_path**) of dependency files:

```
import os
current_path = os.path.dirname(os.path.realpath(__file__)) # Obtain the path of the boot file bootfile.py.
project_root = os.path.dirname(current_path) # Obtain the root directory of the project using the path of the boot file, which is the code directory set on ModelArts console.
otherfile_path = os.path.join(project_root, "otherfileDirectory", "otherfile.py")  # Obtain the path of the dependency files using the root directory of the project.
```

6.3.10 What Is the File Path If a File in the model Directory Is Referenced in a Custom Python Package?

To obtain the actual path to a file in a container, use Python.

```
os.getcwd() # Obtain the current work directory (absolute path) of the file.
os.path.realpath(__ file __) # Obtain the absolute path of the file.
```

You can also use other methods of obtaining a file path through the search engine and use the obtained path to read and write the file.

6.4 Creating a Training Job

6.4.1 What Can I Do If the Message "Object directory size/quantity exceeds the limit" Is Displayed When I Create a Training Job?

Issue Analysis

The code directory for creating a training job has limits on the size and number of files.

Solution

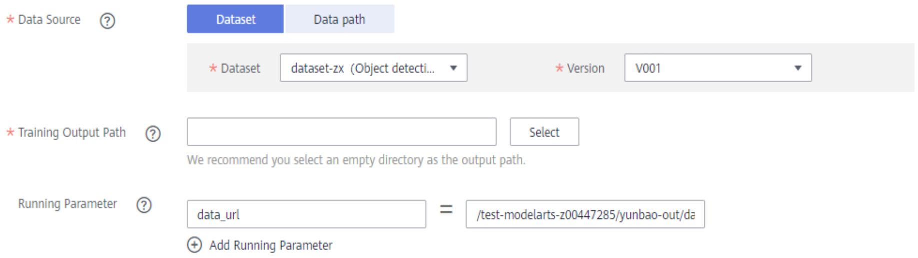
Delete the files except the code from the code directory or save the files in other directories. Ensure that the size of the code directory does not exceed 128 MB and the number of files does not exceed 4,096.

6.4.2 What Are Precautions for Setting Training Parameters?

Pay attention to the following when setting training parameters:

- If the algorithm source and data source have been configured, the **data_url** parameter is automatically set based on the selected object and cannot be directly modified by changing the running parameters.

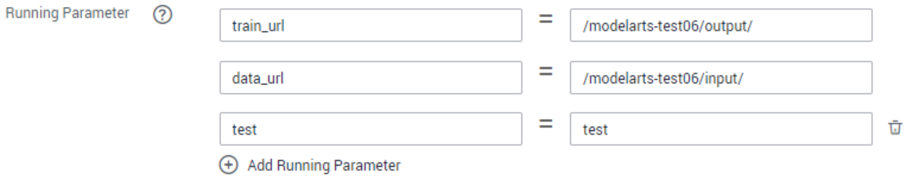
Figure 6-9 Running parameters automatically set



* Dataset: dataset-zx (Object detecti...
* Version: V001
* Training Output Path
Select
We recommend you select an empty directory as the output path.
Running Parameter: data_url = /test-modelarts-z00447285/yunbao-out/da

- When setting running parameters for creating a training job, you only need to set the corresponding parameter names and values. See **Figure 6-10**.

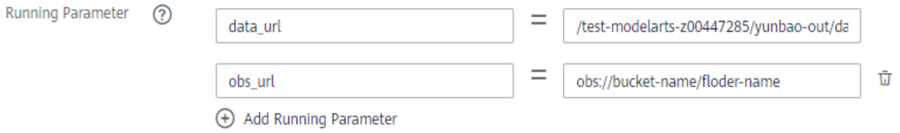
Figure 6-10 Setting running parameters



Running Parameter
train_url = /modelarts-test06/output/
data_url = /modelarts-test06/input/
test = test

- If a parameter value is an OBS bucket path, use the path (starting with **obs://**) to the data. See **Figure 6-11**.

Figure 6-11 Configuring an OBS path



Running Parameter
data_url = /test-modelarts-z00447285/yunbao-out/da
obs_url = obs://bucket-name/floder-name

- When creating an OBS folder in code, call a MoXing API as follows:

```
import moxing as mox
mox.file.make_dirs('obs://bucket_name/sub_dir_0/sub_dir_1')
```

6.4.3 What Are Sizes of the /cache Directories for Different Resource Specifications in the Training Environment?

When creating a training job, you can select CPU, GPU, or Ascend resources based on the size of the training job.

ModelArts mounts a disk to **/cache**. You can use this directory to store temporary files. The **/cache** directory shares resources with the code directory. The directory has different capacities for different resource specifications.

- GPU resources

Table 6-2 Capacities of the cache directories for GPU resources

GPU Specifications	cache Directory Capacity
V100	800 GB
8*V100	3 TB

GPU Specifications	cache Directory Capacity
P100	800 GB

- CPU resources

Table 6-3 Capacities of the cache directories for CPU resources

CPU Specifications	cache Directory Capacity
2 vCPUs | 8 GiB	50 GB
8 vCPUs | 32 GiB	50 GB

6.4.4 Is the /cache Directory of a Training Job Secure?

The program of a ModelArts training job runs in a container. The address of a directory to which the container is mounted is unique, and can be accessed only by the running container. Therefore, the **/cache** directory of the training job is secure.

6.4.5 Why Is a Training Job Always Queuing?

If the training job is always queuing, the selected resources are limited in the resource pool, and the job needs to be queued. In this case, wait for resources. To speed up resource obtaining, do as follows:

1. If you use a public resource pool:

   Resources in a public resource pool are limited. During peak hours, resources may be insufficient if service traffic is heavy. Try to take the following measures:

   – If a free flavor was used, change it to a charged one. Few resources are provided for free flavors, leading to a high queuing probability.

   – The less number of cards in the selected flavor leads to the lower queuing probability. For example, the probability of queuing when selecting a 1-card flavor is much less than that of queuing when selecting an 8-card flavor.

   – Switch to another region.

   – If resources will be used for a long term, purchase a dedicated resource pool.

2. If you use a dedicated resource pool:

   – If there are multiple available dedicated resource pools, switch to an idle one.

   – Release resources in the current resource pool, for example, stop notebook instances that are not used for a long time.

   – Submit a training job during off-peak hours.

   – Contact the account administrator of the resource pool to expand the resource pool based on the usage.

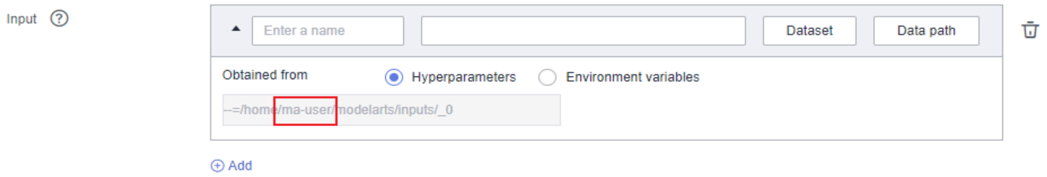
Helpful link: **Why Is the Job Still Queued When Resources Are Sufficient?**

6.4.6 What Determines the Hyperparameter Directory (/work or /ma-user) When Creating a Training Job?

Symptom

The hyperparameter directory for the input and output parameters varies between **/work** and **/ma-user** when creating a training job.

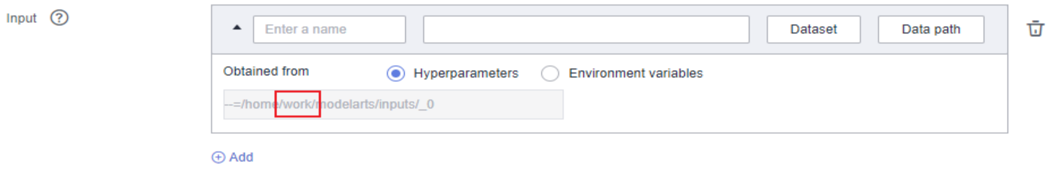
Figure 6-12 /ma-user directory



Obtained from ● Hyperparameters ○ Environment variables

--=/home/ma-user/modelarts/inputs/_0

Figure 6-13 /work directory



Obtained from ● Hyperparameters ○ Environment variables

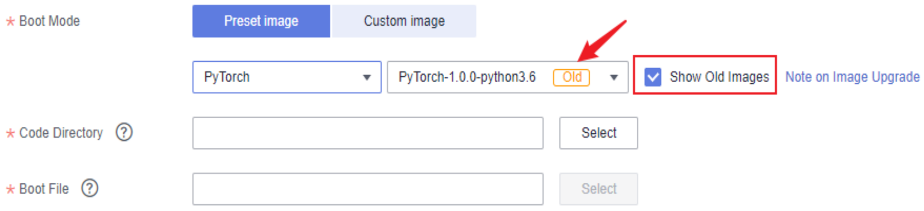
--=/home/work/modelarts/inputs/_0

Solution

The directory varies depending on the selected algorithm for the training job.

- If the selected algorithm is created using an old-version image, the hyperparameter directory of the input and output parameters is **/work**.

  Figure 6-14 Creating an algorithm

  * Boot Mode: Preset image
  PyTorch | PyTorch-1.0.0-python3.6 Old | ☑ Show Old Images
  * Code Directory
  * Boot File

- If the selected algorithm is not created using an old-version image, the hyperparameter directory of the input and output parameters is **/ma-user**.

6.5 Managing Training Job Versions

6.5.1 Does a Training Job Support Scheduled or Periodic Calling?

ModelArts training jobs do not support scheduled or periodic calling. When your job is in the **Running** state, you can call the job based on service requirements.

6.6 Viewing Job Details

6.6.1 How Do I Check Resource Usage of a Training Job?

In the left navigation pane of the ModelArts management console, choose **Training Management > Training Jobs** to go to the **Training Jobs** page. In the training job list, click a job name to view job details. You can view the following metrics on the **Resource Usages** tab page.

- **CPU**: CPU usage (cpuUsage) percentage (Percent)
- **MEM**: Physical memory usage (memUsage) percentage (Percent)
- **GPU**: GPU usage (gpuUtil) percentage (Percent)
- **GPU_MEM**: GPU memory usage (gpuMemUsage) percentage (Percent)

6.6.2 How Do I Access the Background of a Training Job?

ModelArts does not support access to the background of a training job.

6.6.3 Is There Any Conflict When Models of Two Training Jobs Are Saved in the Same Directory of a Container?

Storage directories of ModelArts training jobs do not affect each other. Environments are isolated from each other, and data of other jobs cannot be viewed.

6.6.4 Only Three Valid Digits Are Retained in a Training Output Log. Can the Value of loss Be Changed?

In a training job, only three valid digits are retained in a training output log. When the value of **loss** is too small, the value is displayed as **0.000**. Log content is as follows:

```
INFO:tensorflow:global_step/sec: 0.382191
INFO:tensorflow:step: 81600(global step: 81600) sample/sec: 12.098 loss: 0.000
INFO:tensorflow:global_step/sec: 0.382876
INFO:tensorflow:step: 81700(global step: 81700) sample/sec: 12.298 loss: 0.000
```

Currently, the value of **loss** cannot be changed. You can multiply the value of **loss** by 1000 to avoid this problem.

6.6.5 Can a Trained Model Be Downloaded or Migrated to Another Account? How Do I Obtain the Download Path?

You can download the model trained by a training job and upload the downloaded model to OBS in the region corresponding to the target account.

Obtaining a Model Download Path

1. Log in to the ModelArts console. In the left navigation pane, choose **Training Management** > **Training Jobs**. The **Training Jobs** page is displayed.
2. In the training job list, click a job name to view job details.
3. On the **Configurations** tab page, obtain the path specified for **Training Output Path**, that is, the download path of the training model.

Migrating the Model to Another Account

There are two ways to migrate a trained model to another account:

- Download the trained model and then upload it to the OBS bucket in the region corresponding to the target account.
- Configure a policy for the folder or bucket where the model is stored to authorize other accounts to perform read and write operations. For details, see **Creating a Custom Bucket Policy (Visual Editor)**.

7 Service Deployment

7.1 Model Management

7.1.1 Importing Models

7.1.1.1 How Do I Import the .h5 Model of Keras to ModelArts?

ModelArts does not support the import of models in .h5 format. You can convert the models in .h5 format of Keras to the TensorFlow format and then import the models to ModelArts.

For details about how to convert the Keras format to the TensorFlow format, see the **Keras official website**.

7.1.1.2 How Do I Edit the Installation Package Dependency Parameters in a Model Configuration File When Importing a Model?

Symptom

When importing a model from OBS or a container image, edit a model configuration file. The model configuration file describes the model usage, computing framework, precision, inference code dependency package, and model API. The configuration file must be in JSON format. **dependencies** in the model configuration file specifies the dependencies required for configuring the model inference code. This parameter requires the package name, installation method, and version constraints. For details, see **Specifications for Editing a Model Configuration File** The following section describes how to edit **dependencies** in the model configuration file during model import.

Solution

The installation packages must be installed in sequence. For example, before installing **mmcv-full**, install **Cython**, **pytest-runner**, and **pytest**. In the configuration file, **Cython**, **pytest-runner**, and **pytest** are ahead of **mmcv-full**.

Example:

```
"dependencies": [
   {
   "installer": "pip",
   "packages": [
       {
           "package_name": "Cython"
       },
       {
           "package_name": "pytest-runner"
       },
       {
           "package_name": "pytest"
       },
       {
           "restraint": "ATLEAST",
           "package_version": "5.0.0",
           "package_name": "Pillow"
       },
       {
           "restraint": "ATLEAST",
           "package_version": "1.4.0",
           "package_name": "torch"
       },
       {
           "restraint": "ATLEAST",
           "package_version": "1.19.1",
           "package_name": "numpy"
       },
       {
           "package_name": "mmcv-full"
       }
     ]
  }
]
```

If installing **mmcv-full** failed, the possible cause is that GCC was not installed in the base image, leading to a compilation failure. In this case, use the wheel package on premises to install **mmcv-full**.

Example:

```
"dependencies": [
   {
   "installer": "pip",
   "packages": [
       {
           "package_name": "Cython"
       },
       {
           "package_name": "pytest-runner"
       },
       {
           "package_name": "pytest"
       },
       {
           "restraint": "ATLEAST",
           "package_version": "5.0.0",
           "package_name": "Pillow"
       },
       {
           "restraint": "ATLEAST",
           "package_version": "1.4.0",
           "package_name": "torch"
       },
       {
           "restraint": "ATLEAST",
           "package_version": "1.19.1",
           "package_name": "numpy"
       },
       {
           "package_name": "mmcv_full-1.3.9-cp37-cp37m-manylinux1_x86_64.whl"
       }
     ]
   }
]
```

dependencies in the model configuration file supports multiple dependency structure arrays in list format.

Example:

```
"dependencies": [
    {
    "installer": "pip",
    "packages": [
        {
            "package_name": "Cython"
        },
        {
            "package_name": "pytest-runner"
        },
        {
            "package_name": "pytest"
        },
        {
            "package_name": "mmcv_full-1.3.9-cp37-cp37m-manylinux1_x86_64.whl"
        }
      ]
    },
    {
    "installer": "pip",
    "packages": [
        {
            "restraint": "ATLEAST",
            "package_version": "5.0.0",
            "package_name": "Pillow"
        },
        {
            "restraint": "ATLEAST",
            "package_version": "1.4.0",
            "package_name": "torch"
        },
        {
            "restraint": "ATLEAST",
            "package_version": "1.19.1",
            "package_name": "numpy"
        },
      ]
    }
]
```

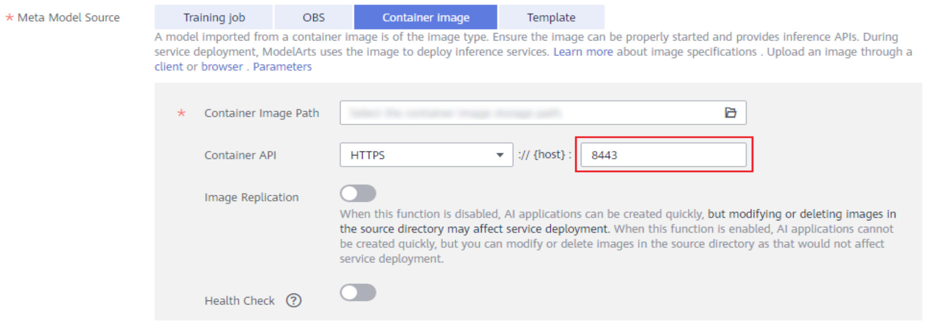
7.1.1.3 How Do I Change the Default Port to Create a Real-Time Service Using a Custom Image?

A port number (for example, 8443) has been specified in a model configuration file. If you do not specify a port (default port 8080 will be used then) or specify another port during AI application creation, deploying the AI application as a service will fail. In this case, set the port number to 8443 in the AI application to resolve this issue.

To change the default port, do as follows:

1. Log in to the ModelArts management console. In the navigation pane, choose **AI Application Management** > **AI Applications**.
2. Click **Create**. On the page for creating an AI application, set **Meta Model Source** to **Container image** and select a custom image.
3. Configure the container API and port number. Ensure that the port number is the same as that specified in the model configuration file.

Figure 7-1 Changing the port



* Meta Model Source: Container image

A model imported from a container image is of the image type. Ensure the image can be properly started and provides inference APIs. During service deployment, ModelArts uses the image to deploy inference services. Learn more about image specifications . Upload an image through a client or browser . Parameters

* Container Image Path

Container API: HTTPS :// {host} : 8443

Image Replication

When this function is disabled, AI applications can be created quickly, but modifying or deleting images in the source directory may affect service deployment. When this function is enabled, AI applications cannot be created quickly, but you can modify or delete images in the source directory as that would not affect service deployment.

Health Check

4. After the configuration, click **Create now**. Wait until the AI application runs properly.
5. Deploy the AI application as a real-time service again.

7.1.1.4 Does ModelArts Support Multi-Model Import?

Importing a model package from OBS to ModelArts applies to single-model scenarios. If multiple models are required, you are advised to import custom images from SWR to create AI applications and deploy services. For details about how to create a custom image, see **Creating a Custom Image and Using It to Create an AI Application**.

7.1.1.5 Restrictions on the Size of an Image for Importing an AI Application

ModelArts uses containers for deploying services. There are size limitations during container runtime. If the size of your model file, custom file, or system file exceeds the Docker size, a message will be displayed, indicating that the image space is insufficient.

The maximum Docker size for a container in a public resource pool is 10 GB, and the maximum Docker size for a container in a dedicated resource pool is 30 GB.

If the AI application is imported from OBS or a training job, the total size of the base image, model files, code, data files, and software packages cannot exceed the limit.

If the AI application is imported from a custom image, the total size of the decompressed image and image dependencies cannot exceed the limit.

7.2 Service Deployment

7.2.1 Functional Consulting

7.2.1.1 What Types of Services Can Models Be Deployed as on ModelArts?

Models can be deployed as real-time services or batch services.

7.2.1.2 What Are the Differences Between Real-Time Services and Batch Services?

- Real-Time Services

  Models are deployed as web services. You can access the services through the management console or APIs.

- Batch Services

  A batch service performs inference on batch data and automatically stops after data processing is completed.

A batch service processes batch data at a time. A real-time service provides APIs for you to call.

7.2.1.3 Why Cannot I Select Ascend 310 Resources?

Ascend 310 resources are limited. If resources are sold out, you cannot select Ascend 310 resources (in the public resource pool) for inference during deployment. On the **Deploy** page, the **Ascend: 1*Ascend-D310(8 GB) | ARM: 3 vCPUs, 6 GB** resource will be unavailable.

Solutions:

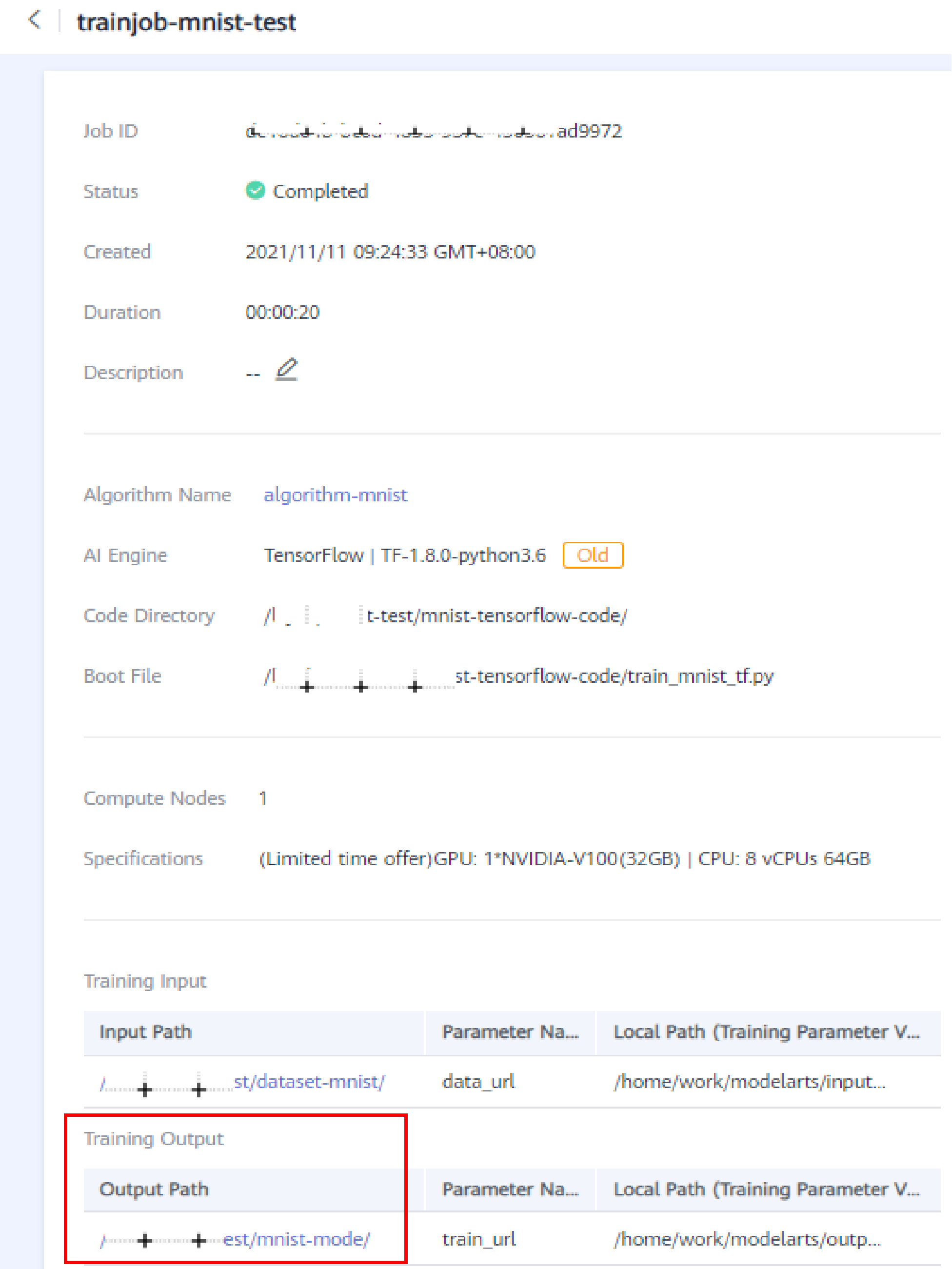
- Method 1: If you want to use Ascend 310 in the public resource pool, you can wait for other users to release the resources. If other services using the Ascend 310 resources stop, you can select the resources for deployment.
- Method 2: If you have a dedicated resource pool with Ascend 310 resources, you can create an Ascend 310 dedicated resource pool.
- Method 3: If Ascend 310 resources in the dedicated resource pool are sold out, you can create an Ascend 310 dedicated resource pool after other users have deleted their Ascend 310 instances.

7.2.1.4 Can Models Trained on ModelArts Be Deployed Locally?

Models trained using ModelArts built-in algorithms are stored in OBS buckets and can be downloaded to a local directory.

1. In the training job list, click the name of the target training job to go to its details page, on which you can obtain the training output path.

Figure 7-2 Training output path



trainjob-mnist-test

Job ID [redacted]ad9972

Status Completed

Created 2021/11/11 09:24:33 GMT+08:00

Duration 00:00:20

Description --

Algorithm Name algorithm-mnist

AI Engine TensorFlow | TF-1.8.0-python3.6 Old

Code Directory /l[redacted]t-test/mnist-tensorflow-code/

Boot File /l[redacted]st-tensorflow-code/train_mnist_tf.py

Compute Nodes 1

Specifications (Limited time offer)GPU: 1*NVIDIA-V100(32GB) | CPU: 8 vCPUs 64GB

Training Input

Input Path	Parameter Na...	Local Path (Training Parameter V...
/[redacted]st/dataset-mnist/	data_url	/home/work/modelarts/input...

Training Output

Output Path	Parameter Na...	Local Path (Training Parameter V...
/[redacted]est/mnist-mode/	train_url	/home/work/modelarts/outp...

2. Click the path to go to the OBS object path. Then, download the model from OBS.

3. Deploy the downloaded model locally.

   For details, see **Creating a Local Model** and **Debugging a Service**.

7.2.1.5 What Is the Maximum Size of a Prediction Request Body?

After a service is deployed and running, you can send an inference request to the service. The requested content can be text, images, voice, or videos, depending on the model of the service.

If you use the inference request address (URL of HUAWEI CLOUD APIG) displayed on the **Usage Guides** tab of the service details page for prediction, the maximum size of the request body is 12 MB. If the request body is oversized, the request will be intercepted.

If you perform the prediction on the **Prediction** tab of the service details page, the size of the request body cannot exceed 8 MB. The size limit varies between the two tab pages because they use different network links.

Ensure that the size of a request body does not exceed the upper limit. If there are high-concurrency and heavy-traffic inference requests, submit a service ticket to professional service support.

7.2.1.6 Can Real-Time Services Be Billed on a Yearly/Monthly Basis?

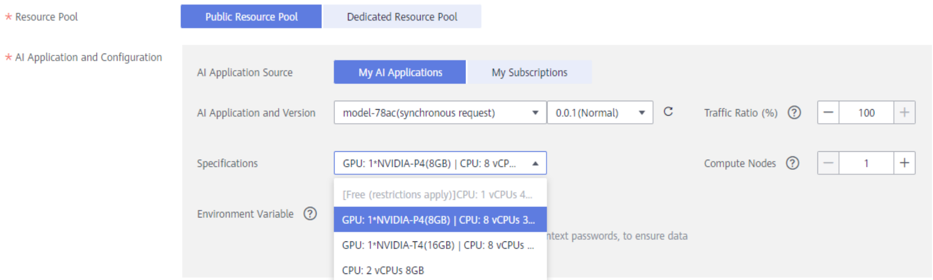
Real-time services cannot be billed on a yearly/monthly basis.

7.2.1.7 How Do I Select Compute Node Specifications for Deploying a Service?

Before deploying a service, specify node specifications. The node specifications displayed on the GUI are calculated by ModelArts based on the target AI application and the node specifications available in the resource pool. You can select the specifications provided by ModelArts or customize the specifications (supported only in dedicated resource pools).

Selecting compute node specifications based on the resources required by your AI application. For example, if an AI application requires 3 CPUs and 10 GB of memory, select compute node specifications higher than 3 CPUs and 10 GB of memory. This ensures that the service can be successfully deployed and run properly.

Figure 7-3 Compute node specifications



* Resource Pool: Public Resource Pool | Dedicated Resource Pool

* AI Application and Configuration

AI Application Source: My AI Applications | My Subscriptions

AI Application and Version: model-78ac(synchronous request) 0.0.1(Normal)

Traffic Ratio (%): 100

Specifications: GPU: 1*NVIDIA-P4(8GB) | CPU: 8 vCP...

- [Free (restrictions apply)]CPU: 1 vCPUs 4...
- GPU: 1*NVIDIA-P4(8GB) | CPU: 8 vCPUs 3...
- GPU: 1*NVIDIA-T4(16GB) | CPU: 8 vCPUs ...
- CPU: 2 vCPUs 8GB

Compute Nodes: 1

Environment Variable

ntext passwords, to ensure data

When using compute node specifications, pay attention to the following:

Permission control

Permissions on general-purpose compute node specifications, for example, **modelarts.vm.cpu.2u** are not controlled. You can select the specifications as long as there are idle resources in the resource pool. ModelArts provides two specifications by default, CPU-powered **modelarts.vm.cpu.2u** and GPU-powered **modelarts.vm.gpu.p4**.

For some special specifications, contact the system administrator to request for permissions.

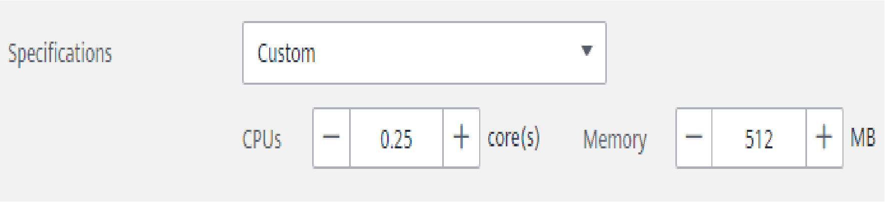
Specifications sold out in a public resource pool

Resources in a public resource pool are limited. If a specification is displayed as sold out, resources of the current specification have been used up. In this case, select other specifications or create your own dedicated resource pool.

Custom specifications

You can customize resource specifications only when a dedicated resource pool is used. Specifications cannot be customized in public resource pools.

Figure 7-4 Custom specifications



Specifications: Custom

CPUs: 0.25 core(s)

Memory: 512 MB

7.2.1.8 What Is the CUDA Version for Deploying a Service on GPUs?

CUDA 10.2 is supported by default. If a later version is required, submit a service ticket to apply for technical support.

7.2.2 Real-Time Services

7.2.2.1 What Do I Do If a Conflict Occurs in the Python Dependency Package of a Custom Prediction Script When I Deploy a Real-Time Service?

Before importing a model, save the inference code and configuration file in the model folder. When coding with Python, import custom packages in relative import (Python import) mode.

If there are packages with duplicate names in the ModelArts inference framework code and they are imported not in relative import mode, a conflict will occur, leading to a service deployment or prediction failure.

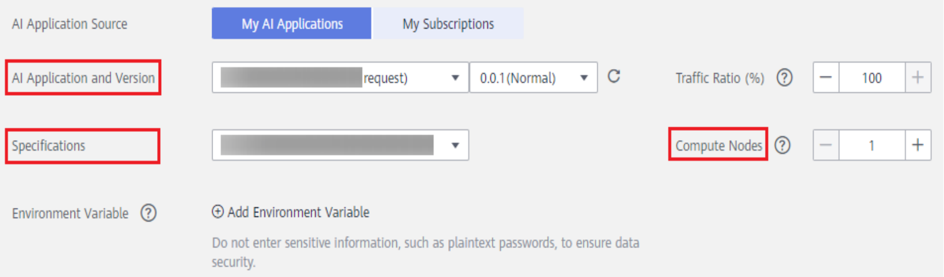
7.2.2.2 How Do I Speed Up Real-Time Prediction?

- When deploying a real-time service, select the compute nodes with higher specifications for better performance. For example, use GPUs instead of CPUs.
- When deploying a real-time service, add the number of compute nodes.

  If you set **Compute Nodes** to **1**, standalone computing is used. If you set **Compute Nodes** to a value greater than 1, distributed computing is used. Configure this parameter based on site requirements.
- The inference speed is closely related to the model complexity. Try to optimize the model for faster prediction.

  ModelArts provides model version management to facilitate source tracing and repeated model tuning.

Figure 7-5 Deploying a real-time service



AI Application Source: My AI Applications | My Subscriptions

AI Application and Version: [redacted] request) | 0.0.1(Normal) — Traffic Ratio (%): 100

Specifications: [redacted] — Compute Nodes: 1

Environment Variable: Add Environment Variable

Do not enter sensitive information, such as plaintext passwords, to ensure data security.

7.2.2.3 Can a New-Version AI Application Still Use the Original API?

ModelArts supports multiple model versions and flexible traffic policies to smoothly gray upgrade model versions. After a service is modified to deploy a new-version model or its model version is upgraded, the original service prediction API remains unchanged.

To adjust a model version, perform the operations described in this section.

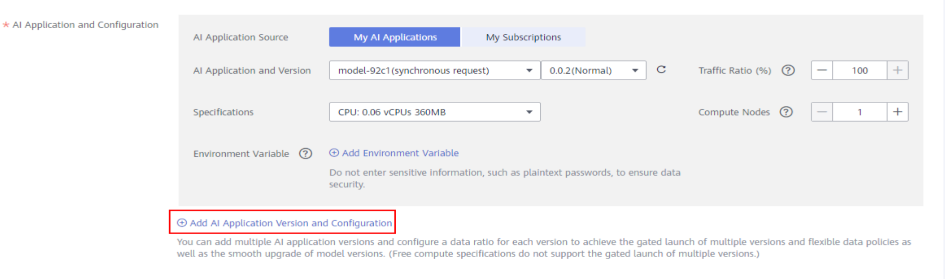
Prerequisites

- A service has been deployed.
- A **new-version AI application** has been created.

Procedure

1. Log in to the ModelArts management console. In the navigation pane on the left, choose **Service Deployment** > **Real-Time Services**. The **Real-Time Services** page is displayed.
2. Locate the target service and click **Modify** in the **Operation** column. The **Modify Service** page is displayed.
3. In the **AI Application and Configuration** area, click **Add AI Application Version and Configuration** to add a new version.

Figure 7-6 Add AI Application Version and Configuration



* AI Application and Configuration

AI Application Source: My AI Applications | My Subscriptions

AI Application and Version: model-92c1(synchronous request) | 0.0.2(Normal) — Traffic Ratio (%): 100

Specifications: CPU: 0.06 vCPUs 360MB — Compute Nodes: 1

Environment Variable: Add Environment Variable

Do not enter sensitive information, such as plaintext passwords, to ensure data security.

Add AI Application Version and Configuration

You can add multiple AI application versions and configure a data ratio for each version to achieve the gated launch of multiple versions and flexible data policies as well as the smooth upgrade of model versions. (Free compute specifications do not support the gated launch of multiple versions.)

4. Set the traffic proportion of the two versions. Service calling requests are allocated based on the proportion. For details about other settings, see **Parameters**. After the setting, click **Next**.
5. Confirm the information and click **Submit**.

7.2.2.4 What Is the Format of a Real-Time Service API?

After an AI application is deployed as a real-time service, you can use the API for inference.

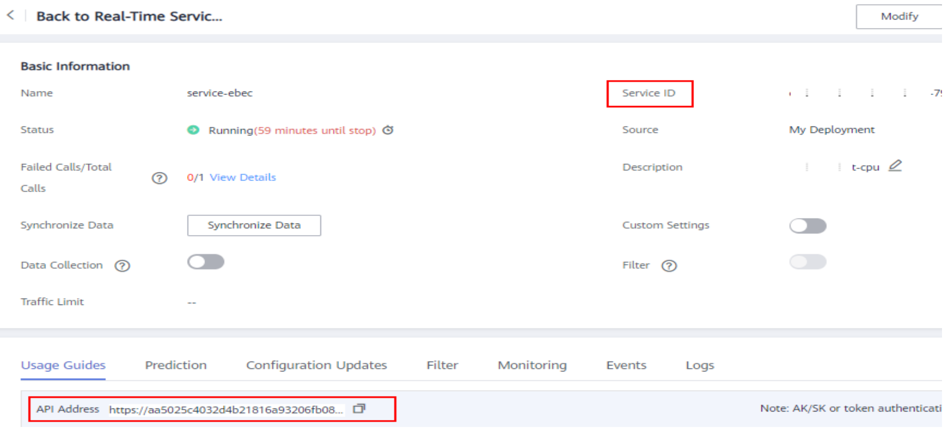
The format of an API is as follows:

```
https://Domain name/Version/infer/service ID
```

Example:

```
https://6ac81cdfac4f4a30be95xxxbb682.apig.xxx.xxx.com/v1/infers/
468d146d-278a-4ca2-8830-0b6fb37d3b72
```

Figure 7-7 API



Back to Real-Time Servic...

Basic Information

Name: service-ebec

Service ID: -7

Status: Running(59 minutes until stop)

Source: My Deployment

Failed Calls/Total Calls: 0/1 View Details

Description: t-cpu

Traffic Limit: --

API Address https://aa5025c4032d4b21816a93206fb08...

Note: AK/SK or token authenticati

7.2.2.5 How Do I Check Whether an Error Is Caused by a Model When a Real-Time Service Is Running But Prediction Failed?

Symptom

A running real-time service is used for prediction. After a prediction request is initiated, the received response does not meet the expectation. It is difficult to determine whether the issue is caused by the model.

Possible Cause

After a real-time service is started, either of the following methods can be used for prediction:

- Method 1: Perform prediction on the **Prediction** tab of the service details page.
- Method 2: Obtain the API URL on the **Usage Guides** tab of the service details page, and use cURL or Postman for prediction.

This issue may occur after an inference request is initiated, regardless of whether method 1 or 2 is used.

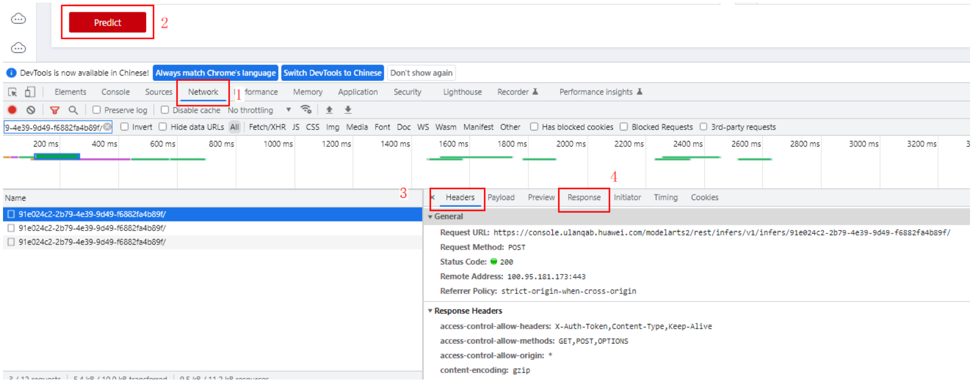
An inference request is finally sent to the model. The issue may be caused by an error occurred when the model processed the inference request. Determine whether the issue is caused by the model, which facilitates rapid fault locating.

Solution

No matter whether method 1 or 2 is used, obtain the response header and body of the inference request.

- If method 1 is used, obtain the response to the inference request through the developer tool of the browser. Take Google Chrome as an example. Press **F12** to open the developer tool, click the **Network** tab and then **Predict**. The response to the inference request is displayed on the **Network** tab page.

  Figure 7-8 Response to an inference request

  Name
  91e024c2-2b79-4e39-9d49-f6882fa4b89f/
  91e024c2-2b79-4e39-9d49-f6882fa4b89f/
  91e024c2-2b79-4e39-9d49-f6882fa4b89f/

  ▾ General
  Request URL: https://console.ulanqab.huawei.com/modelarts2/rest/infers/v1/infers/91e024c2-2b79-4e39-9d49-f6882fa4b89f/
  Request Method: POST
  Status Code: ● 200
  Remote Address: 100.95.181.173:443
  Referrer Policy: strict-origin-when-cross-origin

  ▾ Response Headers
  access-control-allow-headers: X-Auth-Token,Content-Type,Keep-Alive
  access-control-allow-methods: GET,POST,OPTIONS
  access-control-allow-origin: *
  content-encoding: gzip

  Find the inference request in the **Name** pane. The URL of the inference request contains keyword **/v1/infers**. View the complete URL in the **Headers** pane. Obtain the response in **Headers** and **Response**.

- If method 2 is used, obtain the response header and body through different tools. For example, run the cURL command and use **-I** to obtain the response header.

If **Server** in the obtained response header is **ModelArts** and the response body does not contain a ModelArts.XXXX error code, the response is returned by the model. If the response is not as expected, the issue is caused by the model.

Summary and Suggestions

A model can be imported from a container image, OBS, or AI Gallery. The following provides common troubleshooting methods for each model source:

- For a model imported from a container image, the cause of the issue varies depending on image customization. Check model logs to identify the cause.
- For a model imported from OBS, if the response you received contains an MR error code, for example, MR.0105, view logs on the **Logs** tab of the real-time service details page to identify the cause.
- For a model imported from AI Gallery, consult the publisher of the model for the cause.

7.2.2.6 How Do I Fill in the Request Header and Request Body of an Inference Request When a Real-Time Service Is Running?

Symptom

After a real-time service is deployed, you can obtain its inference request address on the **Usage Guides** tab of the service details page when the service is running. However, there is no instruction for filling in the header and body of an inference request.

Possible Cause

The inference request address on the **Usage Guides** tab of the service details page can be called for inference. For security purposes, ModelArts takes authentication and authorization measures to prevent unauthorized calling of the real-time service. Therefore, the header of a prediction request contains the identity information of the request initiator, and the body contains the content to be predicted.

The header must be authenticated by following HUAWEI CLOUD authentication rules. The body must be configured based on model requirements, such as the requirements of pre-processing scripts or custom images.

Solution

- Header:

  On the **Usage Guides** tab of the service details page, you can obtain a maximum of two API addresses, one for IAM or AK/SK authentication and the other for application authentication. The header structure varies depending on the authentication mode.

  - IAM or AK/SK authentication: In the header, enter the domain-level token of the tenant in the target region in the **X-Auth-Token** field. For details, see **Obtaining a User Token Through Password Authentication**.
  - Application authentication: Application authentication can be further classified as AppCode authentication and application signature authentication.

    - For AppCode authentication, enter the AppCode of the application associated with the real-time service in the **X-Apig-AppCode** field of the header.

    - For application signature authentication, in the header, enter the **X-Sdk-Date** and **Authorization** values generated using the AppKey and AppSecret of the application associated with the real-time service through the SDK or tool to authenticate the signature of the request. For details, see **Access Authenticated Using an Application**.

- Body:

  The body varies depending on the model source.

  - If the model is imported from a container image, the body must be configured based on the custom image requirements. For details, contact the image creator.

- If the model is imported from OBS, the requirements on the body are reflected in inference code preprocessing, which will convert the input HTTP body into the input required by the model. For details, see **Specifications for Model Inference Coding**.
- If the model is obtained from AI Gallery, check the calling description in AI Gallery or consult the model provider.

Summary and Suggestions

None

7.2.2.7 Why Cannot I Access the Obtained Inference Request Address from the Initiator Client?

Symptom

After a real-time service is deployed, you can obtain the address of the called server on the **Usage Guides** tab of the service details page when the service is running. However, this address is inaccessible from the client of the request initiator. As a result, the connection failed to set up and the domain name cannot be resolved.

Possible Cause

The addresses displayed on the **Usage Guides** tab page are the addresses of HUAWEI CLOUD API Gateway (APIG). The network between the client of the request initiator and HUAWEI CLOUD is disconnected.

Solution

If the client is out of the HUAWEI CLOUD network, ensure that the client can access the Internet.

If the client is on the HUAWEI CLOUD network, the address is accessible in the default network configuration. Do not configure special network settings, such as firewall rules.

Summary and Suggestions

None

7.2.2.8 What Do I Do If Deploying a Service Failed Due to Insufficient Quota?

During service deployment, error message "A maximum of XXX real-time services are allowed" is returned, indicating that the quota is insufficient.

A maximum of 20 real-time services can be deployed by a user generally. Perform the following operations to resolve this issue:

- Delete malfunctional services.
- Delete services that are not used for a long time.
- Submit a service ticket to apply for a larger quota.

7.2.2.9 Why Did My Service Deployment Fail with Proper Deployment Timeout Configured?

A model can properly start after a service is deployed. The startup status of a model can be detected through a health check.

Check whether a service is deployed using a health check API for custom images. When creating an AI application, configure a health check delay to ensure the initialization of containers.

It is a good practice to configure a proper health check delay for service deployment.

8 Resource Pools

8.1 Can I Use ECSs to Create a Dedicated Resource Pool for ModelArts?

No. This operation is not allowed. When creating a resource pool, you can only select available node flavors provided on the console. These node flavors in dedicated resource pools are from ECSs. However, the ECSs purchased under the account cannot be used by the dedicated resource pools for ModelArts.

8.2 Can I Deploy Multiple Services on One Dedicated Resource Pool Node?

Yes. This operation is allowed.

When you deploy services, select a dedicated resource pool and customize the compute node flavor. Select the node with a low flavor. When the resource pool node allows multiple service node flavors, multiple services can be deployed. If you use this method to deploy a model for inference, ensure that the selected flavor complies with the minimum model requirements for inference. Otherwise, the deployment or prediction may fail.

8.3 How Is a Node Newly Added to a Dedicated Resource Pool Billed?

You will receive a new bill with the newly added node included. Pay for the bill and use the node.

8.4 What Are the Differences Between a Public Resource Pool and a Dedicated Resource Pool?

- A public resource pool is shared between all ModelArts users. If resources are limited, you may need to join the queue.

- A dedicated resource pool is dedicated to you and accessible to your VPC.

8.5 How Do I Log In to a Dedicated Resource Pool Node Through SSH?

Dedicated resource pool nodes of ModelArts cannot be logged in through SSH.

8.6 How Are Training Jobs Queued?

First in first out (FIFO) applies to training jobs. Subsequent jobs can be executed only after the preceding job is complete. This may lead to starvation of small jobs.

NOTE

Job starvation is as follows: For example, a 64-card training job is queuing, and a 1-card training job follows the 64-card one. The 1-card training job can be executed only after the resources of 64 cards are idle. Even if the resources of 30 cards are available, the 1-card training job cannot be executed.

8.7 What Do I Do If Resources Are Insufficient for Staring a New Real-Time Service After I Stop a Real-Time Service in a Dedicated Resource Pool?

Wait for several minutes until the resources of the stopped real-time service are released.

8.8 Can a Public Resource Pool Be Used for Network Connection Between ModelArts and the Authentication Service for Running Algorithms?

No. Public resource pools cannot be used for network connection between the authentication service and ModelArts. The network connection can be set up using a dedicated resource pool.

8.9 Why Is a Dedicated Resource Pool That Fails to Be Created Still Displayed on the Console After It Is Deleted?

After a dedicated resource pool is deleted on the console, the backend releases the resources used by the pool. It takes several minutes to release the resources, during which the pool is still displayed on the console. To create a dedicated resource pool again, wait for 5 minutes after the deletion. Additionally, do not use the name of the dedicated resource pool that fails to be created to name the new dedicated resource pool. To perform an automated test on the UI, it is a good practice to use a random string as the name of the created dedicated resource pool.

8.10 How Do I Add a VPC Peering Connection Between a Dedicated Resource Pool and an SFS?

To add a VPC peering connection, do as follows:

1. Log in to the ModelArts management console and choose **Dedicated Resource Pools** in the navigation pane.
2. In the dedicated resource pool list, click the ID or name of the target resource pool to go to its details page.
3. Click **Configure NAS VPC** in the upper right corner. In the **Configure NAS VPC** dialog box, enable **NAS VPC Connection** and configure the NAS VPC and NAS subnet to be the same as those configured for the SFS.
4. Click **OK**. When you create a training job, the SFS option will be available.

9 API/SDK

9.1 Can ModelArts APIs or SDKs Be Used to Download Models to a Local PC?

ModelArts APIs or SDKs cannot be used to download models to a local PC. However, the output models of training jobs are stored in OBS. You can use OBS APIs or SDKs to download the models. For details, see **Downloading an Object** .

9.2 What Installation Environments Do ModelArts SDKs Support?

ModelArts SDKs can run in notebook or local environments. However, the supported environments vary depending on architectures. For details, see **Table 9-1**.

Table 9-1 SDK installation environments

Development Environment	Architecture	Supported
Notebook	Arm	Yes
	x86	Yes
Local environment	Arm	No
	x86	Yes

9.3 Does ModelArts Use the OBS API to Access OBS Files over an Intranet or the Internet?

In the same region, ModelArts uses the OBS API to access files stored in OBS over an intranet and does not consume public network traffic.

If you download data from OBS through the Internet, you will be charged for the OBS public network traffic. For details about OBS billing, see **Billing Items**.

9.4 How Do I Obtain a Job Resource Usage Curve After I Submit a Training Job by Calling an API?

After submitting a training job by calling an API, log in to the ModelArts console, choose **Training Management** > **Training Jobs**, and click the name or ID of the target training job to go to its details page. In the **Resource Usages** area, view the resource usage curve of the job.

9.5 How Do I View the Old-Version Dedicated Resource Pool List Using the SDK?

You can view the old-version dedicated resource pool list by referring to the following code:

```
from modelarts.session import Session
from modelarts.estimator import Estimator
algo_info = Estimator(modelarts_session=Session()).get_job_pool_list()print(algo_info)
```

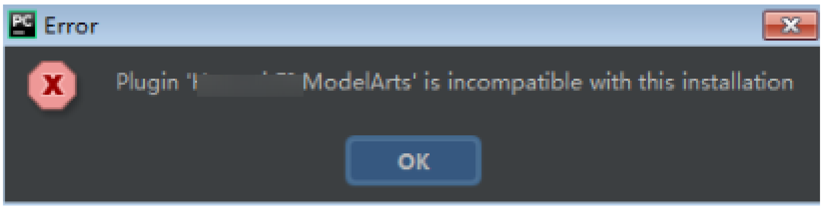
10 Using PyCharm Toolkit

10.1 What Should I Do If an Error Occurs During Toolkit Installation?

Issue

The following error message is displayed during Toolkit installation.

Figure 10-1 Error



Plugin '[redacted] ModelArts' is incompatible with this installation

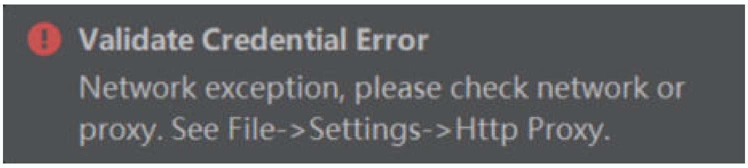
Solution

This issue occurs because the plug-in version is inconsistent with the PyCharm version. You need to obtain the plug-in of the same version as the PyCharm version, that is, version 2019.2 or later.

10.2 What Should I Do If an Error Occurs When I Edit a Credential in PyCharm Toolkit?

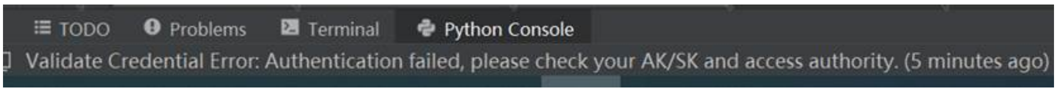
Symptom

When you edit a credential in PyCharm Toolkit, the message "Validate Credential error" is displayed.



Validate Credential Error
Network exception, please check network or proxy. See File->Settings->Http Proxy.

Or



Validate Credential Error: Authentication failed, please check your AK/SK and access authority. (5 minutes ago)

Possible Causes

- Possible cause 1: Information such as the region is incorrectly configured.
- Possible cause 2: The **hosts** file is not configured or is incorrectly configured.
- Possible cause 3: The network proxy settings are incorrect.
- Possible cause 4: The AK/SK is incorrect.
- Possible cause 5: The computer time is incorrectly set.

Solution

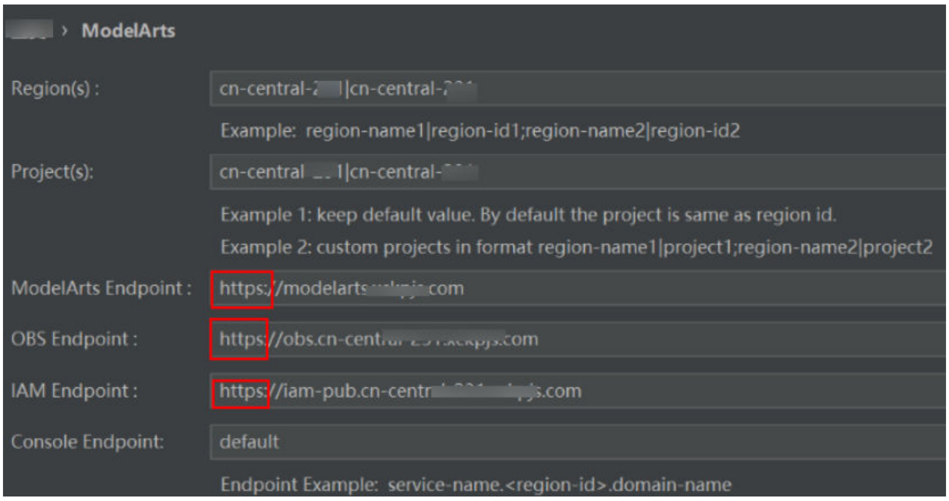
1. Information such as the region is incorrectly configured.

Correctly configure the region, projects, and endpoint.

For example, if the endpoint is incorrect, the authentication fails.

Incorrect example: The endpoint is preceded by **https**.

Figure 10-2 Configuring PyCharm Toolkit



[redacted] › ModelArts

Region(s) : cn-central-[redacted]|cn-central-[redacted]

Example: region-name1|region-id1;region-name2|region-id2

Project(s): cn-central[redacted]|cn-central-[redacted]

Example 1: keep default value. By default the project is same as region id.

Example 2: custom projects in format region-name1|project1;region-name2|project2

ModelArts Endpoint : https://modelarts[redacted]com

OBS Endpoint : https://obs.cn-cent[redacted]com

IAM Endpoint : https://iam-pub.cn-centr[redacted]s.com

Console Endpoint: default

Endpoint Example: service-name.<region-id>.domain-name

2. The hosts file is not configured or is incorrectly configured.

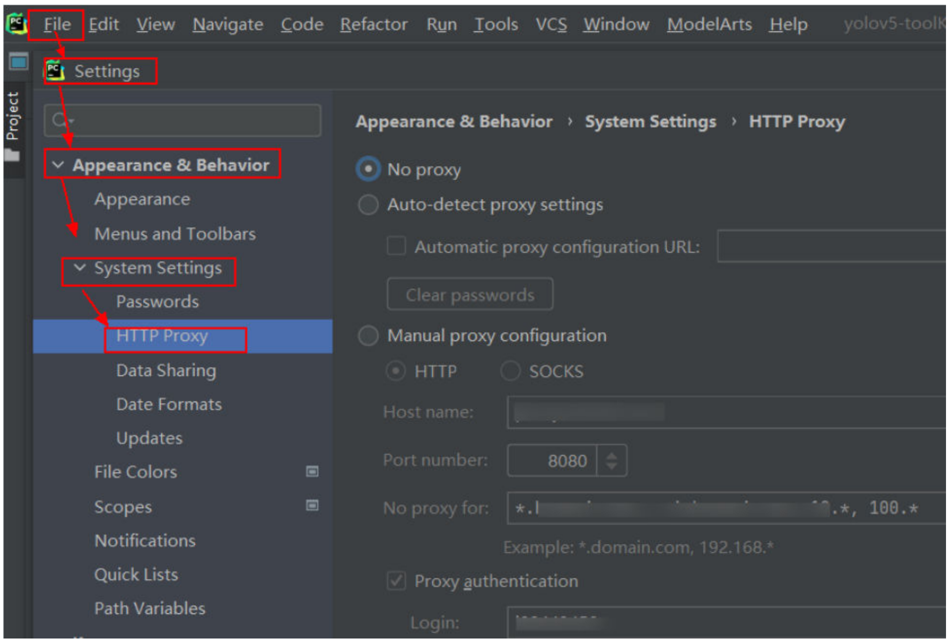
Configure the domain names and IP addresses in the **hosts** file on the local PC.

3. Network proxy settings are incorrect.

If the network requires proxy settings, check whether the proxy settings are correct. You can also use the mobile hotspot to test.

Check whether the proxy settings are correct.

Figure 10-3 PyCharm network proxy settings



● No proxy

○ Auto-detect proxy settings

□ Automatic proxy configuration URL:

Clear passwords

○ Manual proxy configuration

● HTTP ○ SOCKS

Host name: [redacted]

Port number: 8080

No proxy for: *.[redacted].*, 100.*

Example: *.domain.com, 192.168.*

☑ Proxy authentication

Login: [redacted]

4. The AK/SK is incorrect.

Obtained correct AK/SK and try again. For details, see **How Do I Obtain an Access Key?**

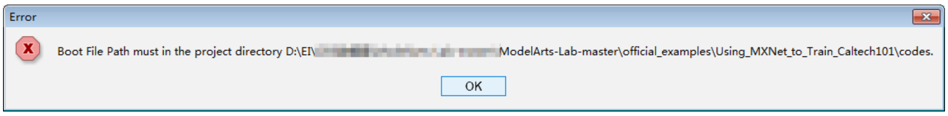
5. The computer time is incorrectly set.

Set the computer time to the correct time.

10.3 Why Cannot I Start Training?

If code that does not belong to the used project is selected in a boot script, training cannot be started. The following figure shows error information. You are advised to add the boot script to the project or open the project where the boot script is located, and then start the training job.

Figure 10-4 Error



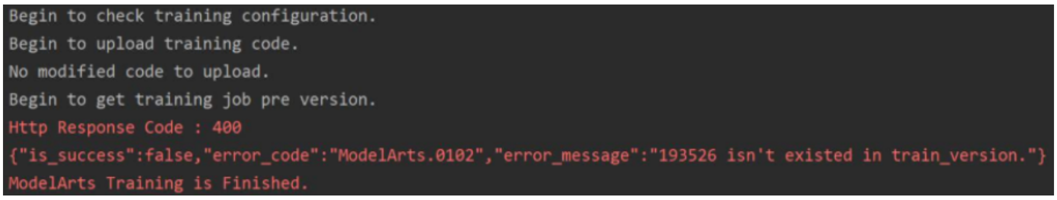
Boot File Path must in the project directory D:\EI\[redacted]ModelArts-Lab-master\official_examples\Using_MXNet_to_Train_Caltech101\codes.

10.4 What Should I Do If Error "xxx isn't existed in train_version" Occurs When a Training Job Is Submitted?

Symptom

Error "xxx isn't existed in train_version" occurs when a training job is submitted. See the following figure.

Figure 10-5 Error "xxx isn't existed in train_version"



```
Begin to check training configuration.
Begin to upload training code.
No modified code to upload.
Begin to get training job pre version.
Http Response Code : 400
{"is_success":false,"error_code":"ModelArts.0102","error_message":"193526 isn't existed in train_version."}
ModelArts Training is Finished.
```

Possible Causes

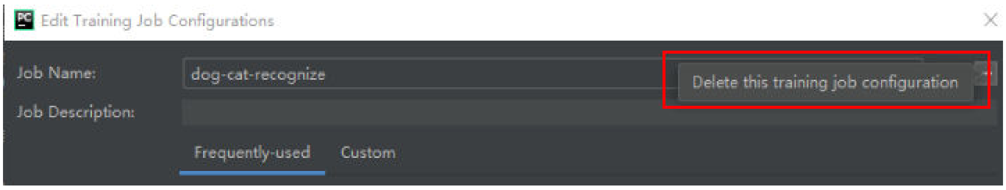
The preceding error occurs because the user logs in to the ModelArts management console and deletes the training job after submitting the training job using PyCharm Toolkit.

PyCharm Toolkit records the training job IDs of ModelArts on the cloud. If you manually delete the job on the ModelArts management console, a message is displayed indicating that the job with the ID cannot be found when you submit the job locally.

Solution

If you have deleted a job on the ModelArts management console, you also need to delete the local configuration from Toolkit. To delete the local configuration, click **Edit Training Configuration**, find the job name, click the minus sign in the upper right corner, and confirm the deletion.

Figure 10-6 Deleting the local configuration



Job Name: dog-cat-recognize

Job Description:

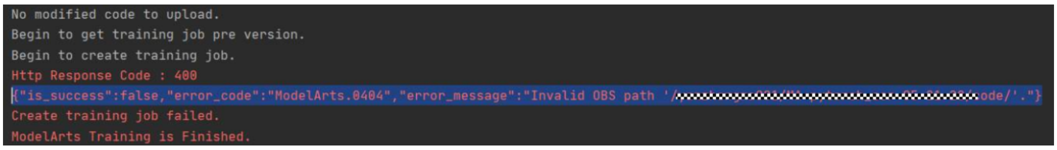
Edit Training Job Configurations
Delete this training job configuration
Frequently-used
Custom

In the displayed confirmation dialog box, confirm the information and click **Yes** to delete the configuration. After the deletion, you can create a training job configuration and submit the training job.

10.5 What Should I Do If Error "Invalid OBS path" Occurs When a Training Job Is Submitted?

When a training job is running, the "Invalid OBS path" error is reported.

Figure 10-7 "Invalid OBS path" error



```
No modified code to upload.
Begin to get training job pre version.
Begin to create training job.
Http Response Code : 400
{"is_success":false,"error_code":"ModelArts.0404","error_message":"Invalid OBS path '/[redacted]code/'."}
Create training job failed.
ModelArts Training is Finished.
```

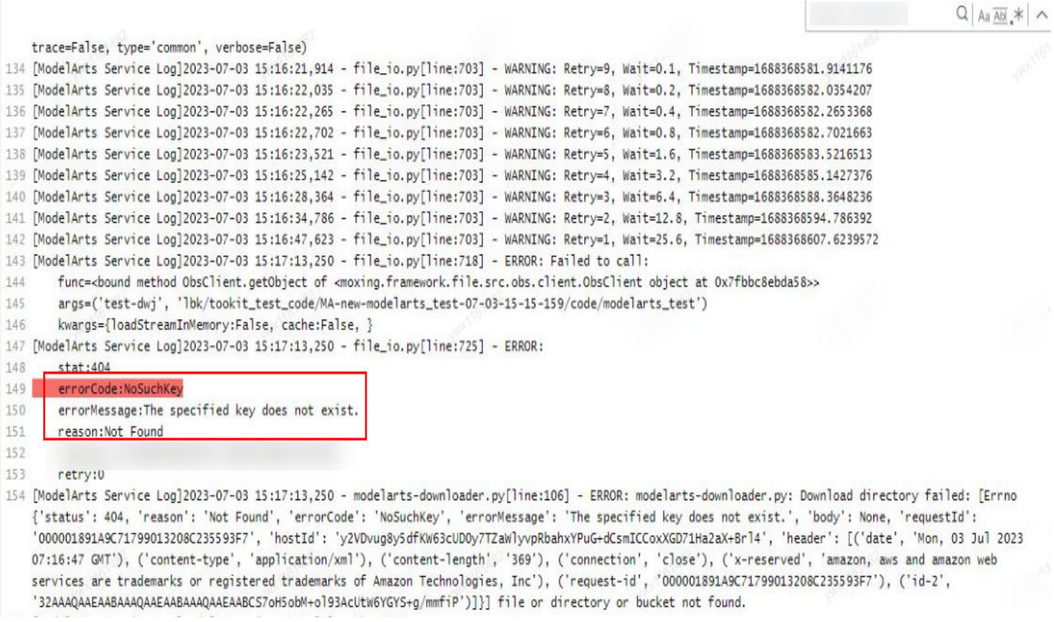
To locate the fault, perform the following operations:

- If you are using ModelArts for the first time, log in to the ModelArts management console and complete access authorization configuration. The agency authorization mode is recommended. After the global configuration is complete, submit the job again.
- Check whether the configured **Data Path in OBS** exists and whether data files exist in the directory. If the directory does not exist, create a directory on OBS and upload the training data to the directory.

10.6 What Should I Do If Error "NoSuchKey" Occurs When PyCharm Toolkit Is Used to Submit a Training Job?

Symptom

When PyCharm Toolkit is used to submit a training job, an error is reported. The log is as follows.



```
  trace=False, type='common', verbose=False)
[ModelArts Service Log]2023-07-03 15:16:21,914 - file_io.py[line:703] - WARNING: Retry=9, Wait=0.1, Timestamp=1688368581.9141176
[ModelArts Service Log]2023-07-03 15:16:22,035 - file_io.py[line:703] - WARNING: Retry=8, Wait=0.2, Timestamp=1688368582.0354207
[ModelArts Service Log]2023-07-03 15:16:22,265 - file_io.py[line:703] - WARNING: Retry=7, Wait=0.4, Timestamp=1688368582.2653368
[ModelArts Service Log]2023-07-03 15:16:22,702 - file_io.py[line:703] - WARNING: Retry=6, Wait=0.8, Timestamp=1688368582.7021663
[ModelArts Service Log]2023-07-03 15:16:23,521 - file_io.py[line:703] - WARNING: Retry=5, Wait=1.6, Timestamp=1688368583.5216513
[ModelArts Service Log]2023-07-03 15:16:25,142 - file_io.py[line:703] - WARNING: Retry=4, Wait=3.2, Timestamp=1688368585.1427376
[ModelArts Service Log]2023-07-03 15:16:28,364 - file_io.py[line:703] - WARNING: Retry=3, Wait=6.4, Timestamp=1688368588.3648236
[ModelArts Service Log]2023-07-03 15:16:34,786 - file_io.py[line:703] - WARNING: Retry=2, Wait=12.8, Timestamp=1688368594.786392
[ModelArts Service Log]2023-07-03 15:16:47,623 - file_io.py[line:703] - WARNING: Retry=1, Wait=25.6, Timestamp=1688368607.6239572
[ModelArts Service Log]2023-07-03 15:17:13,250 - file_io.py[line:718] - ERROR: Failed to call:
    func=<bound method ObsClient.getObject of <moxing.framework.file.src.obs.client.ObsClient object at 0x7fbbc8ebda58>>
    args=('test-dwj', 'lbk/tookit_test_code/MA-new-modelarts_test-07-03-15-15-159/code/modelarts_test')
    kwargs={loadStreamInMemory:False, cache:False, }
[ModelArts Service Log]2023-07-03 15:17:13,250 - file_io.py[line:725] - ERROR:
    stat:404
    errorCode:NoSuchKey
    errorMessage:The specified key does not exist.
    reason:Not Found
    [redacted]
    retry:0
[ModelArts Service Log]2023-07-03 15:17:13,250 - modelarts-downloader.py[line:106] - ERROR: modelarts-downloader.py: Download directory failed: [Errno
{'status': 404, 'reason': 'Not Found', 'errorCode': 'NoSuchKey', 'errorMessage': 'The specified key does not exist.', 'body': None, 'requestId':
'000001891A9C71799013208C235593F7', 'hostId': 'y2VDvug8y5dfKW63cUD0y7TZaWlyvpRbahxYPuG+dCsmICCoxXGD71Ha2aX+Brl4', 'header': [('date', 'Mon, 03 Jul 2023
07:16:47 GMT'), ('content-type', 'application/xml'), ('content-length', '369'), ('connection', 'close'), ('x-reserved', 'amazon, aws and amazon web
services are trademarks or registered trademarks of Amazon Technologies, Inc'), ('request-id', '000001891A9C71799013208C235593F7'), ('id-2',
'32AAAQAAEAABAAAQAAEAABAAAQAAEAABCS7oH5obM+o193AcUtW6YGYS+g/mmfiP')]}] file or directory or bucket not found.
```

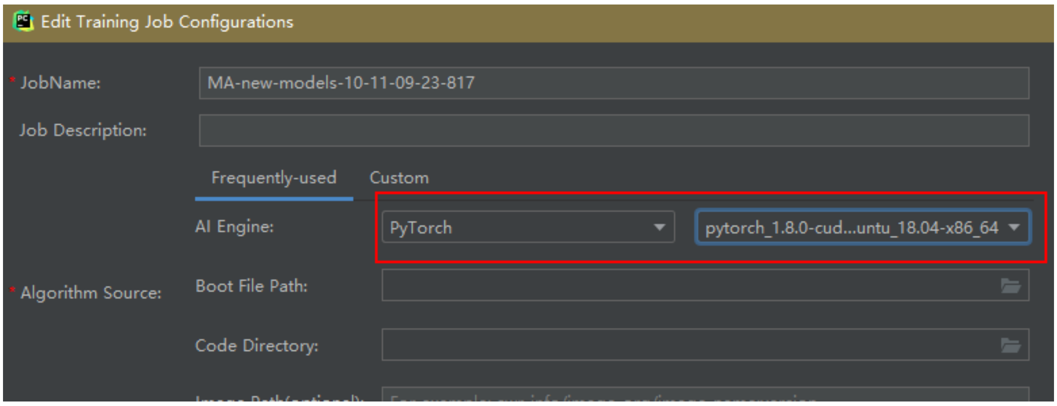
Possible Causes

The image version is too old and the image is incompatible with the new-version training job.

Solution

When using PyCharm Toolkit to submit a training job, select an image version supported by the new-version training job. For details about the supported versions, see **AI engines supported by training jobs of the new version**. Do not select PyTorch-1.0.0, PyTorch-1.3.0, or PyTorch-1.4.0.

Figure 10-8 Selecting an AI engine supported by training jobs of the new version



Edit Training Job Configurations

* JobName: MA-new-models-10-11-09-23-817

Job Description:

Frequently-used | Custom

AI Engine: PyTorch | pytorch_1.8.0-cud...untu_18.04-x86_64

* Algorithm Source: Boot File Path:

Code Directory:

10.7 What Should I Do If an Error Occurs During Service Deployment?

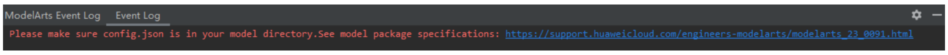
Before deploying a model as a service, edit the configuration file and inference code based on the trained model.

If the **confi.json** configuration file or the **customize_service.py** inference code is missing in the model storage path, an error is displayed, as shown in the following figure.

Solutions:

Write the configuration file and inference code, and save them to the OBS directory where the model to be deployed resides. For details, see **Introduction to Model Package Specifications**.

Figure 10-9 Error



```
Please make sure config.json is in your model directory.See model package specifications: https://support.huaweicloud.com/engineers-modelarts/modelarts_23_0091.html
```

10.8 How Do I View Error Logs of PyCharm Toolkit?

The error logs of PyCharm Toolkit are recorded in the **idea.log** file of PyCharm. For example, in the Windows operating system, the path of the **idea.log** file is **C:\Users\xxx\.IdeaIC2019.2\system\log\idea.log**.

Search for **modelarts** in the log file to view all logs related to PyCharm Toolkit.

10.9 How Do I Use PyCharm ToolKit to Create Multiple Jobs for Simultaneous Training?

PyCharm ToolKit supports only one job at a time. To run another job, you must manually stop the current one.

10.10 What Should I Do If "Error occurs when accessing to OBS" Is Displayed When PyCharm ToolKit Is Used?

Symptom

The PyCharm ToolKit log showed "Error occurs when accessing to OBS".

Possible Causes

You do not have OBS permissions.

Solution

Check whether you have the OBS permissions.

Step 1 Log in to the ModelArts console, choose **Data Management** > **Datasets**, and click **Create**. You have the OBS permissions if you can access the OBS path. If you do not have the OBS permissions, go to **Configure the OBS permis...** to configure the OBS permissions.

Step 2 **Configure the OBS permissions**.

----End

11 Change History

Released On	Description
2023-11-22	Added **What Determines the Hyperparameter Directory (/work or /ma-user) When Creating a Training Job?**.
2023-10-12	Added the following sections: **When the SSH Tool Is Used to Connect to a Notebook Instance, Server Processes Are Cleared, but the GPU Usage Is Still 100%** **Basic Problems Causing the Failures to Access the Development Environment Through VS Code**
2023-09-30	Added cache directory alarm reporting metrics and Ascend metrics in **How Do I View All ModelArts Monitoring Metrics in AOM?**.
2023-9-7	Moved the following sections to **Service Deployment** in *Troubleshooting*: "What Do I Do If an Image Fails to Be Pulled When a Service Is Deployed, Started, Upgraded, or Modified?", "What Do I Do If an Image Restarts Repeatedly When a Service Is Deployed, Started, Upgraded, or Modified?", "What Do I Do If a Container Health Check Fails When a Service Is Deployed, Started, Upgraded, or Modified?", and "What Do I Do If Resources Are Insufficient When a Service Is Deployed, Started, Upgraded, or Modified?". Moved "What Are the Events and Their Types for an AI application?" and "What Are the Events and Their Types for a Service?" to **Managing AI Applications** and **Deploying an AI Application as a Service** in *Model Inference*. Added **Restrictions on the Size of an Image for Importing an AI Application**.

Released On	Description
2023-08-30	Deleted "Data Management" > "Why Can't I Find My Created OBS Bucket After I Select an OBS Path in ModelArts?" and merged OBS documentation into **General Issues > Incorrect OBS Path on ModelArts**.
2023-04-11	Added the FAQs in the *PyCharm Tool Guide* to **FAQs**.
2023-03-30	Deleted the information about old-version notebook that has been terminated.
2022-11-10	Added **How Do I Change the Default Port to Create a Real-Time Service Using a Custom Image?** Added 8.3.2.8-What Do I Do If an Image Fails to Be Pulled When a Real-Time Service Is Deployed, Started, Upgraded, or Modified? Added 8.3.2.9-What Do I Do If an Image Restarts Repeatedly When a Real-Time Service Is Deployed, Started, Upgraded, or Modified? Added 8.3.2.10-What Do I Do If a Container Health Check Failed When a Real-Time Service Is Deployed, Started, Upgraded, or Modified? Added 8.3.2.11-What Do I Do If Resources Are Insufficient When a Real-Time Service Is Deployed, Started, Upgraded, or Modified?
2022-04-28	Added **How Can I Resolve Abnormal Font Display on a ModelArts Notebook Accessed from iOS?** Optimized **How Do I Upload Local Files to a Notebook Instance?** and **How Do I Download Files from a Notebook Instance to a Local Computer?**
2021-10-20	Added **What Are the Differences Between Real-Time Services and Batch Services?**
2021-09-16	Added an FAQ on notebook instances. **What Do I Do If "Read timed out" Is Displayed After I Run pip install?**
2021-04-19	● Added an FAQ on common issues. **How Do I Purchase or Enable ModelArts?** ● Added FAQs on data management. **How Data Is Distributed Between Team Members During Team Labeling?** **How Do I Merge Two Datasets?** ● Added FAQs on notebook instances. **How Do I Access the OBS Bucket of Another Account from a Notebook Instance?** **What Are the Relationships Between Files Stored in JupyterLab, Terminal, and OBS?**

Released On	Description
2021-04-15	● Added an FAQ on training jobs. **Why the Data Read Efficiency Is Low When a Large Number of Data Files Are Read During Training?** ● Added FAQs on notebook instances. **Can Notebook Instances Be Remotely Logged In?** **Does the System Automatically Stop or Delete a Notebook Instance If I Do Not Enable Automatic Stop?**
2021-02-24	● Optimized FAQs on ExeML. ● Optimized FAQs on training jobs. ● Optimized FAQs on model management. ● Optimized FAQs on model deployment.
2021-01-21	Added an FAQ on APIs and SDKs. **Can ModelArts APIs or SDKs Be Used to Download Models to a Local PC?**
2021-01-12	● Added an FAQ on ExeML. **What Are the Differences Between ExeML and Subscribed Algorithms?** ● Added an FAQ on training jobs. **What Are the Solutions to Underfitting?** ● Added an FAQ on model management.
2020-09-29	Added an FAQ on permission policies. ● **What Do I Do If a Message Indicating Insufficient Permissions Is Displayed When I Use ModelArts?**
2020-02-26	Added the following section: ● **Billing**
2019-08-30	Added the following section: **How Do I Create a Training Job When a Dependency Package Is Referenced by the Model to Be Trained?**

Released On	Description
2019-08-20	Added the following sections: **Why Does ExeML Training Fail?** **What Should I Do When the System Displays an Error Message Indicating that No Space Left After I Run the pip install Command?** **How Do I Upload Local Files to a Notebook Instance?** **Where Will the Data Be Uploaded to?** **Why Does the Instance Break Down When dead kernel Is Displayed During Training Code Running?** **What Do I Do If the Code Can Be Run But Cannot Be Saved, and the Error Message "save error" Is Displayed?** **Can I Install MoXing in a Local Environment?** **What Do I Do If a Notebook Instance Won't Run My Code?**
2019-05-09	This is the first official release.