ModelArts

Preguntas frecuentes

Edición	01
Fecha	2025-12-16

HUAWEI TECHNOLOGIES CO., LTD.

Marcas y permisos


y otras marcas registradas de Huawei pertenecen a Huawei Technologies Co., Ltd.
Todas las demás marcas registradas y los otros nombres comerciales mencionados en este documento son propiedad de sus respectivos titulares.

Aviso

Las funciones, los productos y los servicios adquiridos están estipulados en el contrato celebrado entre Huawei y el cliente. Es posible que la totalidad o parte de los productos, las funciones y los servicios descritos en el presente documento no se encuentren dentro del alcance de compra o de uso. A menos que el contrato especifique lo contrario, ninguna de las afirmaciones, informaciones ni recomendaciones contenidas en este documento constituye garantía alguna, ni expresa ni implícita.

La información contenida en este documento se encuentra sujeta a cambios sin previo aviso. En la preparación de este documento se realizaron todos los esfuerzos para garantizar la precisión de sus contenidos. Sin embargo, ninguna declaración, información ni recomendación contenida en el presente constituye garantía alguna, ni expresa ni implícita.

Huawei Technologies Co., Ltd.

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

Sitio web: https://www.huawei.com

Email: support@huawei.com

Índice

1 Cuestiones generales

1.1 ¿Qué es ModelArts?

ModelArts es una plataforma de desarrollo de IA integral dirigida a desarrolladores y científicos de datos de todos los niveles. Le ayuda a crear, entrenar y desplegar modelos rápidamente en cualquier lugar (desde la nube hasta el perímetro) y gestionar flujos de trabajo de IA de todo el ciclo de vida. ModelArts acelera el desarrollo de la IA y fomenta la innovación de la IA con capacidades clave, como el preprocesamiento de datos y el etiquetado automático, el entrenamiento distribuido, la creación automatizada de modelos y la ejecución de flujos de trabajo con solo un clic.

La plataforma de ModelArts integral cubre todas las etapas del desarrollo de IA, incluidos el procesamiento de datos, la creación de aplicaciones de IA y el entrenamiento y despliegue de modelos. La capa subyacente de ModelArts admite varios recursos informáticos heterogéneos. Puede seleccionar y utilizar los recursos de forma flexible sin tener que considerar las tecnologías subyacentes. Además, ModelArts es compatible con marcos de desarrollo de IA de código abierto populares como TensorFlow y MXNet. Los desarrolladores también pueden usar marcos de algoritmos de desarrollo propio para que coincidan con sus hábitos de uso.

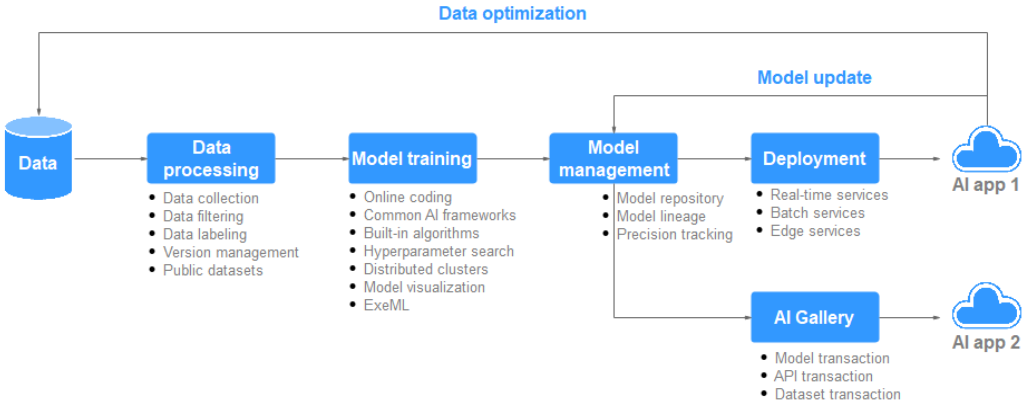
ModelArts tiene como objetivo lograr un desarrollo de IA simple y conveniente.ModelArts es adaptable a los requerimientos de desarrolladores de IA de diferente experiencia. Por ejemplo, los desarrolladores de servicios pueden usar ExeML para crear rápidamente aplicaciones de IA sin codificación. Los principiantes de IA no necesitan prestar atención al desarrollo de modelos, sino que utilizan directamente algoritmos integrados para crear aplicaciones de IA. Los ingenieros de IA pueden usar múltiples entornos de desarrollo para compilar código para un modelado rápido y desarrollo de aplicaciones.

Arquitectura del producto

ModelArts es una plataforma de desarrollo de IA integral que soporta todo el proceso de desarrollo, incluido el procesamiento de datos, la gestión y el despliegue de aplicaciones de IA y el entrenamiento de modelos, y proporciona AI Gallery para compartir modelos.

ModelArts admite todo tipo de escenarios de aplicaciones de IA, como clasificación de imágenes, detección de objetos, análisis de vídeo, reconocimiento de voz, recomendación de productos y detección de excepciones.

Figura 1-1 Arquitectura de ModelArts



1.2 ¿Cuáles son las relaciones entre ModelArts y otros servicios?

IAM

ModelArts utiliza Identity and Access Management (IAM) para la autenticación y autorización. Para obtener más información acerca de IAM, consulte ***Guía del usuario de Identity and Access Management***.

OBS

ModelArts utiliza Object Storage Service (OBS) para almacenar datos y modelos de forma segura y fiable a bajo costo. Para obtener más detalles, consulte ***Guía de la operación de consola de Object Storage Service***.

Tabla 1-1 Relación entre ModelArts y OBS

Función	Subtarea	Relación
ExeML	Etiquetado de datos	Los datos etiquetados en la ModelArts se almacenan en OBS.
	Entrenamiento automático	Después de completar un trabajo de entrenamiento, el modelo generado se almacena en OBS.
	Despliegue de modelos	ModelArts implementa modelos almacenados en OBS como servicios en tiempo real.
Ciclo de vida de desarrollo de la IA	Gestión de datos	● Los conjuntos de datos se almacenan en OBS. ● La información de etiquetado del conjunto de datos se almacena en OBS. ● Los datos se pueden importar desde OBS.
	Entorno de desarrollo	Los archivos de datos o de código de una instancia de bloc de notas se almacenan en OBS.

Función	Subtarea	Relación
	Entrenamiento de modelos	● Los conjuntos de datos utilizados por los trabajos de formación se almacenan en OBS. ● Los scripts que se ejecutan para los trabajos de entrenamiento se almacenan en OBS. ● Los modelos generados por los trabajos de entrenamiento se almacenan en las rutas de OBS especificadas. ● Los registros de ejecución de los trabajos de entrenamiento se almacenan en las rutas de OBS especificadas.
	Gestión de aplicaciones de IA	Después de completar un trabajo de entrenamiento, el modelo generado se almacena en OBS. Puede importar el modelo desde OBS.
	Despliegue del servicio	Los modelos almacenados en OBS se pueden implementar como servicios.
Ajustes	-	Autoriza a ModelArts a acceder a OBS (usando una delegación o clave de acceso) para que ModelArts pueda usar OBS para almacenar datos y crear instancias de bloc de notas.

CCE

ModelArts utiliza Cloud Container Engine (CCE) para desplegar modelos como servicios en tiempo real. CCE permite una alta simultaneidad y proporciona escalado elástico. Para obtener más información acerca de CCE, vea la ***Guía de usuario de Cloud Container Engine***.

SWR

Para utilizar un marco de IA que no es compatible con ModelArts, utilice Software Repository for Container (SWR) para personalizar una imagen e importarla a ModelArts para su entrenamiento o inferencia. Para obtener más información sobre SWR, consulte ***Guía de usuario de Software Repository for Container***.

1.3 ¿Cuáles son las diferencias entre ModelArts y DLS?

Deep Learning Service (DLS) es una plataforma integral de aprendizaje profundo basada en las capacidades informáticas de alto rendimiento de Huawei Cloud. Con varios modelos de redes neuronales optimizados, DLS le permite desplegar fácilmente el entrenamiento y la evaluación del modelo con la flexibilidad de la programación bajo demanda.

Sin embargo, DLS solo admite las tecnologías de aprendizaje profundo, mientras que ModelArts integra tanto las tecnologías de aprendizaje profundo como de aprendizaje automático. ModelArts es una plataforma integral de desarrollo de IA que gestiona el ciclo de vida de desarrollo de IA desde el etiquetado de datos, el desarrollo de algoritmos hasta el entrenamiento de modelos y el despliegue. Para ser específicos, ModelArts contiene y soporta

las funciones y características de DLS. Actualmente, DLS se termina en Huawei Cloud. Las funciones relacionadas con el aprendizaje profundo se pueden utilizar directamente en ModelArts. Si es usuario de DLS, también puede migrar los datos de DLS a ModelArts.

1.4 ¿Cómo puedo comprar o activar ModelArts?

ModelArts es una plataforma lista para usar y no necesita comprarse ni habilitarse. Puede iniciar sesión directamente en la consola de ModelArts, completar la configuración global y usar las funciones requeridas.

En el caso de ModelArts solo se facturan las funciones que utilizan especificaciones informáticas. Todos los grupos de recursos públicos se facturan en modo de pago por uso según las especificaciones seleccionadas y la duración de ejecución del trabajo. Puede comprar un grupo de recursos dedicado sobre una base de pago por uso o anual/mensual. Al ejecutar un trabajo de entrenamiento o desplegar un servicio, puede usar su grupo de recursos dedicado sin pagar tarifas adicionales.

1.5 ¿Qué chips de Ascend son compatibles?

Actualmente, Ascend 310 y Ascend 910 son compatibles.

- **Model training**: Ascend 910 se puede utilizar para entrenar modelos. ModelArts proporciona algoritmos diseñados para el entrenamiento de modelos con Ascend 910.
- **Model inference**: Cuando un modelo se despliega como un servicio en tiempo real de ModelArts puede usar recursos de Ascend 310 para inferencia del modelo.

1.6 ¿Cómo obtengo una clave de acceso?

Obtención de una clave de acceso

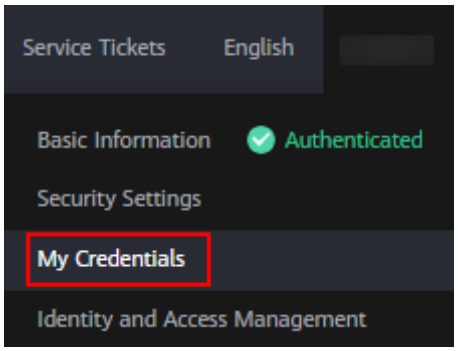
1. Inicie sesión en **Huawei Cloud** y haga clic en **Console** en la esquina superior derecha de la página para acceder a la consola de gestión de Huawei Cloud.

   Figura 1-2 Consola

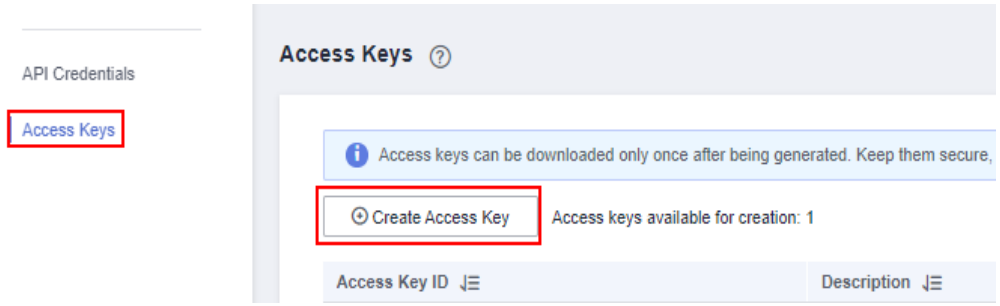
2. Coloque el cursor sobre el nombre de la cuenta en la esquina superior derecha de la consola y elija **My Credentials** en la lista desplegable. Se muestra la página **API Credentials**.

Figura 1-3 Mis credenciales



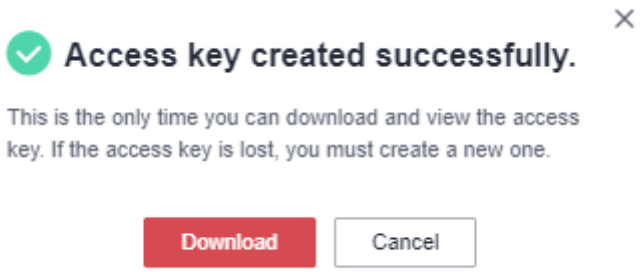
3. En la página **API Credentials**, seleccione **Access Keys** > **Create Access Key**.

Figura 1-4 Crear clave de acceso



4. Ingrese la descripción de la clave y haga clic en **OK**. Haga clic en **Download** para descargar la clave.

Figura 1-5 Clave de acceso creada



5. El archivo de clave de acceso se guarda en la carpeta de descargas predeterminada del navegador. Abra el archivo **credentials.csv** para ver la clave de acceso con una AK y una SK.

1.7 ¿Cómo subo datos a OBS?

Antes de usar ModelArts para desarrollar modelos de IA, los datos deben subirse a un bucket de OBS. Puede iniciar sesión en la consola de OBS para crear un bucket de OBS, crear una carpeta en él y cargar datos. Para obtener más información sobre cómo cargar datos, consulte ***Pasos iniciales de Object Storage Service***.

1.8 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.

1.9 ¿Qué debo hacer si se muestra un mensaje que indica permisos insuficientes cuando utilizo ModelArts?

Si aparece un mensaje que indica que no hay permisos suficientes cuando utiliza ModelArts, realice las operaciones descritas en esta sección para conceder permisos para los servicios relacionados según sea necesario.

Los permisos para usar ModelArts dependen de la autorización de OBS. Por lo tanto, los usuarios de ModelArts también requieren permisos del sistema de OBS.

- Para obtener más información acerca de cómo conceder a un usuario permisos completos para OBS y permisos de operaciones comunes para ModelArts, consulte **Configuración de permisos de operaciones comunes**.
- Para obtener más información acerca de cómo gestionar los permisos de usuario en OBS y ModelArts de forma precisa y configurar las políticas personalizadas, consulte **Creación de una política personalizada para ModelArts**.

Configuración de permisos de operaciones comunes

Para utilizar las funciones básicas de ModelArts es necesario aplicar la política **ModelArts CommonOperations** para los servicios a nivel de proyecto. El uso de ModelArts depende de los permisos de OBS. La política de **Tenant Administrator** debe aplicarse globalmente para todos los usuarios del proyecto.

El procedimiento es el siguiente:

1. Cree un grupo de usuarios.

   Inicie sesión en la consola de IAM y elija **User Groups** > **Create User Group**. Escriba un nombre de grupo de usuarios y haga clic en **OK**.

2. Asigne permisos al grupo de usuarios.

   En la lista de grupos de usuarios, haga clic en **Manage Permissions** en la columna **Operation** de la fila que contiene el grupo de usuarios creado en el paso **1**. En la página de ficha **Permissions**, haga clic en **Assign Permissions**. Configure los siguientes permisos:

   – Establezca **Scope** en **Global service project** y seleccione la política **Tenant Administrator**. Véase **Figura 1-6**.

   – Establezca **Scope** en **Region-specific projects** y seleccione la política **ModelArts CommonOperations**. Véase **Figura 1-7**.

   > NOTA
   >
   > La autorización de un proyecto específico regional solo tiene efecto en la región autorizada. Si la autorización tiene que surtir efecto en todas las regiones, la autorización debe repetirse para cada región involucrada.

   Figura 1-6 Asignación de permisos para el proyecto de servicio global

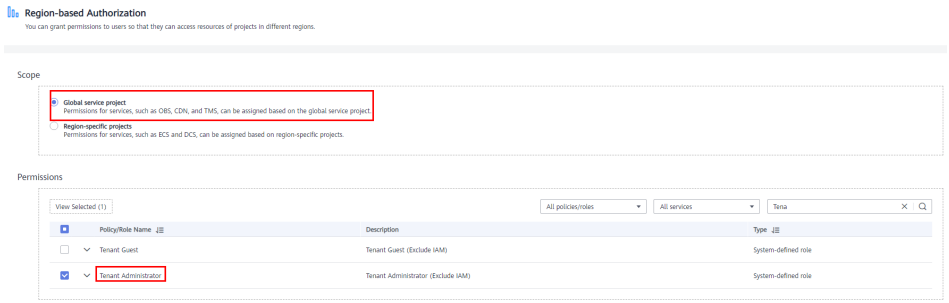
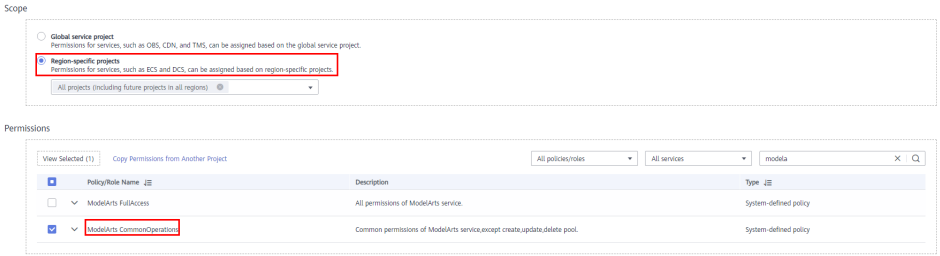
   

   Figura 1-7 Asignación de permisos para proyectos específicos de la región

3. **Cree un usuario y agréguelo al grupo de usuarios**.

   Cree un usuario en la consola de IAM y agregue el usuario al grupo creado en **1**.

4. **Inicie sesión** y verifique los permisos.

Inicie sesión en la consola de ModelArts como el usuario creado, cambie a la región autorizada y compruebe que las políticas **ModelArts CommonOperations** y **Tenant Administrator** están en vigor.

- Elija **Service List** > **ModelArts**. Elija **Dedicated Resource Pools**. En la página que se muestra, seleccione un tipo de grupo de recursos y haga clic en **Create**. No debería poder crear un nuevo grupo de recursos.
- Elija cualquier otro servicio en **Service List**. Suponiendo que los permisos actuales solo contienen **ModelArts CommonOperations**, debe recibir un mensaje que indique que no tiene permisos suficientes.
- Elija **Service List** > **ModelArts**. En la consola de ModelArts, elija **Data Management** > **Datasets** > **Create Dataset**. Debería poder acceder a la ruta de OBS correspondiente.

Creación de una política personalizada para ModelArts

Además de las políticas de sistema predeterminadas de ModelArts, puede crear políticas personalizadas, que también pueden abordar los permisos de OBS. Para obtener más información, vea **Crear una política personalizada**.

Puede crear políticas personalizadas en el editor visual o creando un archivo de JSON. Esta sección describe cómo usar un archivo de JSON para configurar una política personalizada para conceder los permisos necesarios para usar el entorno de desarrollo y cómo configurar los permisos de OBS mínimos para los usuarios de ModelArts.

NOTA

Una política personalizada puede contener acciones para varios servicios a los que se puede acceder globalmente o solo para los proyectos específicos de la región.

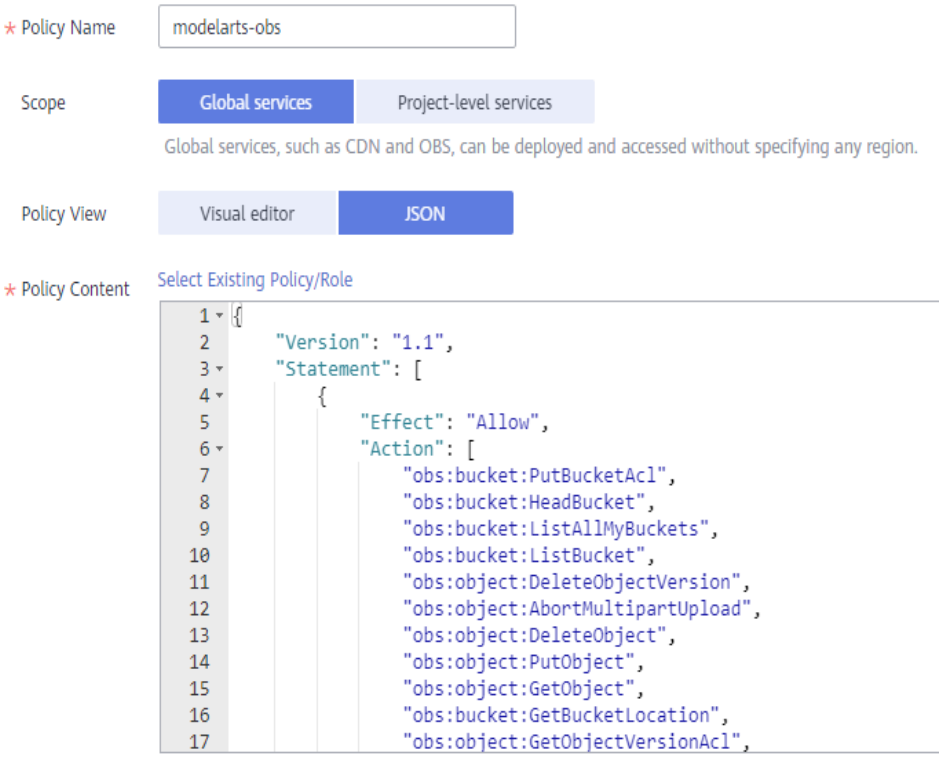
ModelArts es un servicio a nivel de proyecto, pero OBS es un servicio global, por lo que debe crear políticas separadas para los dos servicios y luego aplicar estas políticas a los usuarios.

1. Cree una política personalizada para minimizar los permisos para OBS de la que depende ModelArts. Véase **Figura 1-8**.

   Inicie sesión en la consola de IAM y elija **Permissions** > **Create Custom Policy**. Configure los parámetros de la siguiente manera:

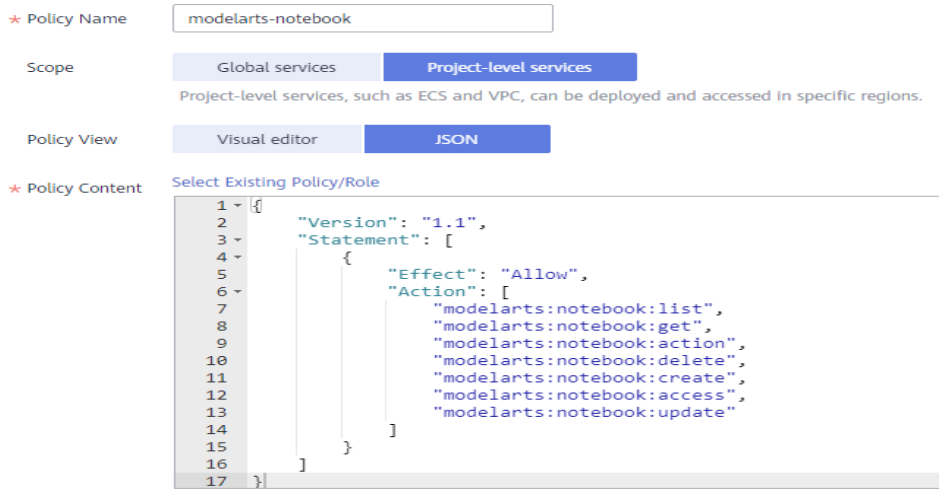
   - **Policy Name**: Seleccione un nombre de la política personalizada.
   - **Scope**: **Global services**.
   - **Policy View**: **JSON**.
   - **Policy Content**: Siga las instrucciones de **Ejemplo de políticas personalizadas de OBS**. Para obtener más información acerca de los permisos del sistema de OBS, consulte **Gestión de permisos de OBS**.

Figura 1-8 Permisos mínimos para OBS



2. Cree una política personalizada para el permiso para usar el entorno de desarrollo de ModelArts. Véase **Figura 1-9**. Configure los parámetros de la siguiente manera:
   - **Policy Name**: Seleccione un nombre de la política personalizada.
   - **Scope: Project-level services**.
   - **Policy View**: **JSON**.
   - **Policy Content**: Siga las instrucciones de **Ejemplo de políticas personalizadas para utilizar el entorno de desarrollo de ModelArts**. Para ver las acciones que se pueden agregar para las políticas personalizadas, consulte **Referencia de la API de ModelArts > Políticas de permisos y acciones admitidas**.

Figura 1-9 Permiso para utilizar el entorno de desarrollo



– Para ver las políticas del sistema de otros servicios, consulte **Permisos del sistema**.

3. En la consola de IAM, **cree un grupo de usuarios y conceda los permisos necesarios.**

   Después de crear un grupo de usuarios en la consola de IAM, conceda la política personalizada creada en **1** al grupo de usuarios.

4. **Cree un usuario y agréguelo al grupo de usuarios**.

   Cree un usuario en la consola de IAM y agregue el usuario al grupo creado en **3**.

5. **Inicie sesión** y verifique los permisos.

   Inicie sesión en la consola de ModelArts como el usuario creado, cambie a la región autorizada y compruebe que las políticas **ModelArts CommonOperations** y **Tenant Administrator** están en vigor.

   – Elija **Service List** > **ModelArts**. En la consola de ModelArts, elija **Data Management** > **Datasets**. Si no puede crear un conjunto de datos, los permisos (para usar el entorno de desarrollo) concedidos solo a los usuarios de ModelArts han surtido efecto.

   – Elija **Service List** > **ModelArts**. En la consola de ModelArts, elija **DevEnviron** > **Notebooks** > **Create**. Debe poder acceder a la ruta de acceso de OBS especificada en **Storage Path**.

Ejemplo de políticas personalizadas de OBS

Los permisos para usar ModelArts requieren la autorización de OBS. En el ejemplo siguiente se muestra el OBS mínimo necesario, incluidos los permisos para los bucket y objetos de OBS. Después de obtener los permisos mínimos para OBS, los usuarios pueden acceder a OBS desde ModelArts sin restricciones.

```
{
    "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"
        }
    ]
}
```

Ejemplo de políticas personalizadas para utilizar el entorno de desarrollo de ModelArts

```
{
    "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"
            ]
        }
    ]
}
```

1.10 ¿Cómo uso ModelArts para entrenar modelos basados en datos estructurados?

Para la mayoría de los usuarios, ModelArts proporciona la función de análisis predictivo de ExeML para entrenar modelos basados en datos estructurados.

Para usuarios más avanzados, ModelArts proporciona la función de creación de notebook de DevEnviron para el desarrollo de código. Permite a los usuarios crear tareas de entrenamiento con grandes volúmenes de datos en trabajos de entrenamiento y utilizar los motores como Scikit_Learn, XGBoost o Spark_MLlib en los procesos de desarrollo y entrenamiento.

1.11 ¿Qué son las Regiones y las AZ?

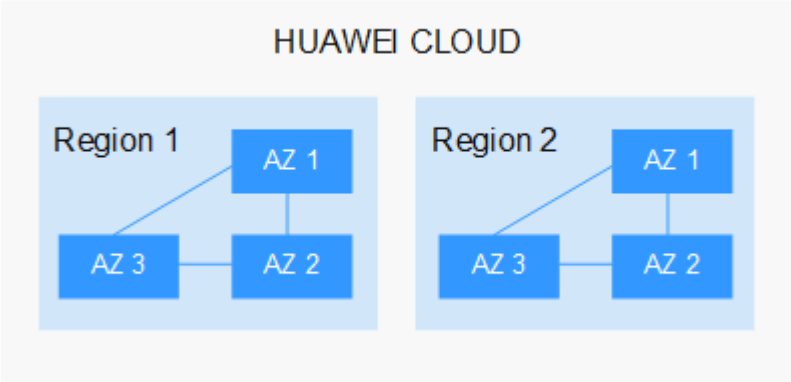
Concepto

Una región y una zona de disponibilidad (AZ) identifican la ubicación de un centro de datos. Puede crear recursos en una región específica y AZ.

- Las regiones se dividen en función de la ubicación geográfica y la latencia de la red. Los servicios públicos, como Elastic Cloud Server (ECS), Elastic Volume Service (EVS), Object Storage Service (OBS), Virtual Private Cloud (VPC), Elastic IP (EIP) y Image Management Service (IMS), se comparten dentro de la misma región. Las regiones se clasifican en regiones universales y regiones dedicadas. Una región universal proporciona servicios en la nube universales para los tenants estándares. Una región dedicada proporciona servicios específicos para tenants específicos.
- Una AZ contiene uno o más centros de datos físicos. Cada AZ cuenta con instalaciones independientes de electricidad, de refrigeración, de extinción de incendios y a prueba de humedad. Dentro de una AZ, los recursos de computación, red, almacenamiento y otros se dividen de forma lógica en múltiples clústeres. Las AZ dentro de una región están interconectadas usando fibras ópticas de alta velocidad para soportar sistemas de alta disponibilidad cruzados.

La **Figura 1-10** muestra la relación entre las regiones y AZ.

Figura 1-10 Regiones y AZ



Huawei Cloud ofrece servicios en muchas regiones de todo el mundo. Seleccione una región y AZ según los requisitos. Para obtener más información, consulte **Productos y servicios globales**.

Selección de una región

Al seleccionar una región, tenga en cuenta los siguientes factores:

- Localización

  Se recomienda seleccionar la región más cercana para una baja latencia de red y un acceso rápido. Las regiones dentro de China continental proporcionan la misma infraestructura, calidad de red BGP, así como operaciones de recursos y configuraciones. Por lo tanto, si sus usuarios objetivo están en China continental, no es necesario tener en cuenta las diferencias de latencia de la red al seleccionar una región.
- Precio del recurso

  Los precios de los recursos pueden variar en diferentes regiones. Para obtener más información, consulte **Detalles de precios del producto**.

Selección de una AZ

Al implementar recursos, tenga en cuenta los requisitos de las aplicaciones en cuanto a la recuperación ante desastres (DR) y la latencia de la red.

- Para una alta capacidad de DR, despliegue recursos en diferentes AZ dentro de la misma región.
- Para una menor latencia de red, despliegue recursos en la misma AZ.

Regiones y puntos de conexión

Antes de usar una API para invocar a recursos, especifique su región y punto de conexión.

1.12 ¿Cómo puedo comprobar si ModelArts y un bucket de OBS están en la misma región?

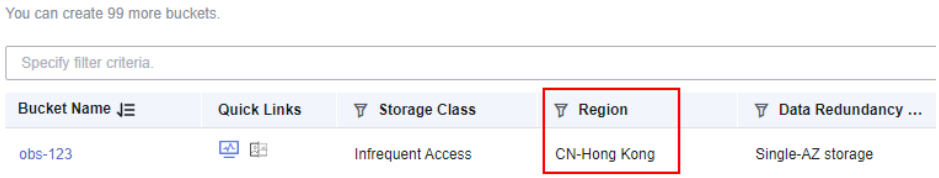
Si es necesario especificar un directorio de OBS para usar funciones de ModelArts, como la creación de trabajos de entrenamiento y conjuntos de datos, asegúrese de que el bucket de OBS y ModelArts estén en la misma región.

Comprobación de si el bucket de OBS y ModelArts están en la misma región

1. Compruebe la región donde reside el bucket de OBS creado.
   a. Inicie sesión en OBS Console.
   b. En la página **Object Storage Service**, para buscar un bucket, introduzca una palabra clave en **Bucket Name**.

      En la columna **Region**, vea la región donde se encuentra el bucket de OBS creado.

      Figura 1-11 Región donde se encuentra un bucket de OBS

2. Compruebe la región donde se despliega ModelArts.

   Inicie sesión en la consola de ModelArts y vea la región donde reside ModelArts en la esquina superior izquierda.
3. Compruebe si la región del bucket de OBS creado es la misma que la de ModelArts. Asegúrese de que el bucket de OBS está en la misma región que ModelArts.

1.13 ¿Cómo puedo ver todos los archivos almacenados en OBS de ModelArts?

Para ver todos los archivos almacenados en OBS al utilizar instancias de notebook o trabajos de entrenamiento, utilice uno de los métodos siguientes:

- Consola de OBS

  Inicie sesión en la consola de OBS con la cuenta actual y busque los bucket, carpetas y archivos de OBS.
- Puede usar una API para comprobar si existe un directorio determinado. En una instancia de notebook existente o después de crear una nueva instancia de notebook, ejecute el siguiente comando para comprobar si el directorio existe:

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

  Si hay un gran número de archivos, espere hasta que se muestre la ruta final del archivo.

1.14 ¿Por qué se muestra el error: 403 Forbidden cuando realizo operaciones en OBS?

Síntoma

El mensaje "Error: stat:403" aparece cuando uso mox.file.copy_parallel en ModelArts para realizar operaciones en OBS.

Figura 1-12 Mensaje de error

```
ERROR:root:
        stat:403
        errorCode:None
        errorMessage:None
        reason:Forbidden
        request-id:000001752610DE67600F295F15304A6C
        retry:0
```

Causas posibles

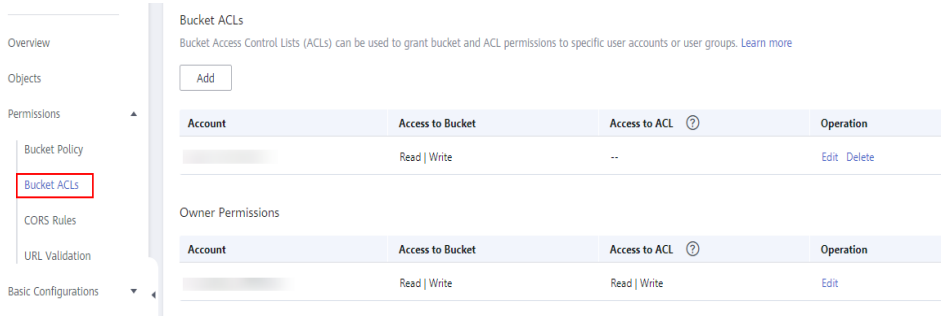
- ModelArts utiliza una AK/SK para la autenticación global, y la AK/SK se ha eliminado y recreado.
- No tiene permiso para acceder a los bucket de OBS.
- La política de bucket de OBS es incorrecta.

Solución

Si accede al bucket de OBS mediante una cuenta de usuario de IAM, póngase en contacto con la cuenta del tenant para realizar las siguientes operaciones:

- Para la causa 1, vaya a la página de configuración global y vuelva a configurar la autorización.
- For cause 2, log in to the OBS console, search for the target OBS bucket, and click the bucket name to go to the **Overview** page. En el panel de navegación de la izquierda, elija **Permissions** > **Bucket ACLs**. En la página **Bucket ACLs** que se muestra, compruebe si la cuenta actual tiene los permisos de lectura y escritura. Si no es así, póngase en contacto con el propietario del bucket para conceder los permisos.

  Figura 1-13 ACL de bucket



- Para la causa 2, inicie sesión en la consola de OBS, busque el bucket de OBS de destino y haga clic en el nombre del bucket para ir a la página **Overview**. En el panel de navegación, elija **Permissions** > **Bucket Policy** y verifique que los usuarios de IAM puedan acceder al bucket de OBS actual.

Si el error persiste, consulte **¿Por qué no puedo acceder a OBS (403 AccessDenied) después de recibir el permiso de acceso OBS?** para la solución de problemas adicionales.

1.15 ¿Dónde se almacenan los conjuntos de datos de ModelArts en un contenedor?

Los conjuntos de datos de ModelArts y datos en ubicaciones de almacenamiento de datos específicas se almacenan en OBS.

1.16 ¿Qué marcos de IA admite ModelArts?

Los marcos y versiones de IA compatibles con ModelArts varían ligeramente según el notebook de entorno de desarrollo, los trabajos de entrenamiento y la inferencia de modelos (gestión y despliegue de aplicaciones de IA). A continuación se describen los marcos de IA soportados por cada módulo.

Notebook de entorno de desarrollo

La imagen y las versiones compatibles con las instancias de notebook de entorno de desarrollo varían según los entornos de tiempo de ejecución.

Tabla 1-2 Imágenes soportadas por el notebook de la nueva versión

Imagen	Descripción	Chip soportado	SSH remoto	Jupyter Lab en línea
pytorch1.8-cuda10.2-cudnn7-ubuntu18.04	Imagen pública impulsada por CPU o GPU para desarrollo y entrenamiento de algoritmos generales, con el motor de IA integrado PyTorch 1.8	CPU o GPU	Sí	Sí
mindspore1.7.0-cuda10.1-py3.7-ubuntu18.04	Desarrollo y entrenamiento de algoritmos generales basados en CPU o GPU, preconfigurados con el motor de IA MindSpore 1.7.0 y CUDA 10.1	CPU o GPU	Sí	Sí
mindspore1.7.0-py3.7-ubuntu18.04	Desarrollo y entrenamiento de algoritmos generales con CPU, preconfigurados con el motor de IA MindSpore 1.7.0	CPU	Sí	Sí
pytorch1.10-cuda10.2-cudnn7-ubuntu18.04	Desarrollo y entrenamiento de algoritmos generales basados en CPU o GPU, preconfigurados con el motor de IA PyTorch 1.10 y CUDA 10.2	CPU o GPU	Sí	Sí
tensorflow2.1-cuda10.1-cudnn7-ubuntu18.04	Imagen pública impulsada por CPU o GPU para el desarrollo y entrenamiento de algoritmos generales, con el motor de IA integrado TensorFlow 2.1	CPU o GPU	Sí	Sí
conda3-ubuntu18.04	Limpiar imagen base personalizada solo incluye Conda	CPU	Sí	Sí

Imagen	Descripción	Chip soportado	SSH remoto	Jupyter Lab en línea
pytorch1.4-cuda10.1-cudnn7-ubuntu18.04	Imagen pública impulsada por CPU o GPU para desarrollo y entrenamiento de algoritmos generales, con el motor de IA integrado PyTorch 1.4	CPU o GPU	Sí	Sí
tensorflow1.13-cuda10.0-cudnn7-ubuntu18.04	Imagen pública impulsada por GPU para el desarrollo y el entrenamiento de algoritmos generales, con el motor de IA integrado TensorFlow 1.13.1	GPU	Sí	Sí
conda3-cuda10.2-cudnn7-ubuntu18.04	La imagen base personalizada limpia incluye CUDA 10.2, Conda	CPU	Sí	Sí
spark2.4.5-ubuntu18.04	Desarrollo y entrenamiento de algoritmos con CPU, preconfigurados con PySpark 2.4.5 y se pueden conectar a clústeres de Spark preconfigurados, incluidos MRS y DLI	CPU	No	Sí
mindspore1.2.0-cuda10.1-cudnn7-ubuntu18.04	Imagen pública impulsada por GPU para desarrollo y entrenamiento de algoritmos, con motor de IA integrado MindSpore-GPU	GPU	Sí	Sí
mindspore1.2.0-openmpi2.1.1-ubuntu18.04	Imagen pública impulsada por CPU para desarrollo y entrenamiento de algoritmos, con motor de IA integrado MindSpore-CPU	CPU	Sí	Sí

Tabla 1-3 Imágenes soportadas por el notebook de la versión antigua

Entorno de tiempo de ejecución	Motor y versión de IA incorporados	Chip soportado
Multi-Engine 1.0 (Python3, Recomendado)	MXNet 1.2.1	CPU o GPU
	PySpark 2.3.2	CPU
	PyTorch 1.0.0	GPU
	TensorFlow 1.13.1	CPU o GPU
	TensorFlow 1.8	CPU o GPU

Entorno de tiempo de ejecución	Motor y versión de IA incorporados	Chip soportado
	XGBoost-Sklearn	CPU
Multi-Engine 1.0 (Python2)	Caffe 1.0.0	CPU o GPU
	MXNet 1.2.1	CPU o GPU
	PySpark 2.3.2	CPU
	PyTorch 1.0.0	GPU
	TensorFlow 1.13.1	CPU o GPU
	TensorFlow 1.8	CPU o GPU
	XGBoost-Sklearn	CPU
Multi-Engine 2.0 (Python3)	PyTorch 1.4.0	GPU
	R-3.6.1	CPU o GPU
	TensorFlow 2.1.0	CPU o GPU

Trabajos de entrenamiento

En la siguiente tabla se enumeran los motores de IA.

Los motores de entrenamiento incorporados en la nueva versión se nombran en el siguiente formato:

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

Tabla 1-4 Motores de IA apoyados por trabajos de entrenamiento de la nueva versión

Entorno de tiempo de ejecución	Chip soportado	Arquitectura del sistema	Versión del sistema	Motor de AI y su versión	Versión compatible con CUDA o Ascend
TensorFlow	CPU o 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 o GPU	x86_64	Ubuntu 18.04	pytorch_1.8.0-cuda_10.2-py_3.7-ubuntu_18.04-x86_64	CUDA 10.2

Entorno de tiempo de ejecución	Chip soportado	Arquitectura del sistema	Versión del sistema	Motor de AI y su versión	Versión compatible con CUDA o Ascend
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

Tabla 1-5 Motores de IA soportados por trabajos de entrenamiento de la versión anterior

Entorno de tiempo de ejecución	Chip soportado	Arquitectura del sistema	Versión del sistema	Motor de AI y su versión	Versión compatible con CUDA o Ascend
TensorFlow	CPU o 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 o GPU	x86_64	Ubuntu 16.04	MXNet-1.2.1-python2.7	-

Entorno de tiempo de ejecución	Chip soportado	Arquitectura del sistema	Versión del sistema	Motor de AI y su versión	Versión compatible con CUDA o Ascend
				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	-
Ray	CPU o GPU	x86_64	Ubuntu 16.04	RAY-0.7.4-python3.6	-
PyTorch	CPU o 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 o 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	-

Motores de IA compatibles para la inferencia de ModelArts

Si importa un modelo desde una plantilla u OBS para crear una aplicación de IA, puede seleccionar los motores y las versiones de IA en la siguiente tabla.

NOTA

- Los entornos de tiempo de ejecución marcados con **recommended** provienen de imágenes unificadas, que se utilizarán como imágenes de inferencia base convencionales. Las imágenes unificadas proporcionan paquetes de instalación completos.
- Las imágenes de la versión anterior serán descontinuadas. Utilice imágenes unificadas.
- Una imagen de tiempo de ejecución unificada se denomina con el siguiente formato: *<AI engine and version> - <Hardware and version: CPU, CUDA, or CANN> - <Python version> - <SO version> - <CPU architecture>*

Tabla 1-6 Motores de IA compatibles y su tiempo de ejecución

Motor	Entorno de ejecución	Nota
TensorFlow	python3.6 python2.7 tf1.13-python2.7-gpu tf1.13-python2.7-cpu tf1.13-python3.6-gpu tf1.13-python3.6-cpu tf1.13-python3.7-cpu tf1.13-python3.7-gpu tf2.1-python3.7 tensorflow_2.1.0-cuda_10.1-py_3.7-ubuntu_18.04-x86_64 (recommended)	● TensorFlow 1.8.0 se utiliza en **python2.7** y **python3.6**. ● **python3.6**, **python2.7**, y **tf2.1-python3.7** indican que el modelo puede ejecutarse tanto en CPU como en GPU. Para otros valores de tiempo de ejecución, si el sufijo contiene **cpu** o **gpu**, el modelo solo puede ejecutarse en CPU o GPU. ● El tiempo de ejecución predeterminado es **python2.7**.
MXNet	python3.7 python3.6 python2.7	● MXNet 1.2.1 se utiliza en **python2.7**, **python3.6** y **python3.7**. ● **python2.7**, **python3.6** y **python3.7** indican que el modelo puede ejecutarse tanto en CPU como en GPU. ● El tiempo de ejecución predeterminado es **python2.7**.
Caffe	python2.7 python3.6 python3.7 python2.7-gpu python3.6-gpu python3.7-gpu python2.7-cpu python3.6-cpu python3.7-cpu	● Caffe 1.0.0 se utiliza en **python2.7**, **python3.6**, **python3.7**, **python2.7-gpu**, **python3.6-gpu**, **python3.7-gpu**, **python2.7-cpu**, **python3.6-cpu** y **python3.7-cpu**. ● **python 2.7**, **python3.7** y **python3.6** solo se pueden usar para ejecutar modelos en CPU. Para otros valores de tiempo de ejecución, si el sufijo contiene **cpu** o **gpu**, el modelo solo puede ejecutarse en CPU o GPU. Utilice el tiempo de ejecución de **python2.7-gpu**, **python3.6-gpu**, **python3.7-gpu**, **python2.7-cpu**, **python3.6-cpu** o **python3.7-cpu**. ● El tiempo de ejecución predeterminado es **python2.7**.

Motor	Entorno de ejecución	Nota
Spark_MLlib	python2.7 python3.6	● Spark_MLlib 2.3.2 se utiliza en **python2.7** y **python3.6**. ● El tiempo de ejecución predeterminado es **python2.7**. ● **python2.7** y **python3.6** solo se pueden usar para ejecutar modelos en CPU.
Scikit_Learn	python2.7 python3.6	● Scikit_Learn 0.18.1 se utiliza en **python2.7** y **python3.6**. ● El tiempo de ejecución predeterminado es **python2.7**. ● **python2.7** y **python3.6** solo se pueden usar para ejecutar modelos en CPU.
XGBoost	python2.7 python3.6	● XGBoost 0.80 se utiliza en python2.7 y python3.6. ● El tiempo de ejecución predeterminado es **python2.7**. ● **python2.7** y **python3.6** solo se pueden usar para ejecutar modelos en CPU.
PyTorch	python2.7 python3.6 python3.7 pytorch1.4-python3.7 pytorch1.5-python3.7 pytorch_1.8.0-cuda_10.2-py_3.7-ubuntu_18.04-x86_64 (recomendado)	● PyTorch 1.0 se utiliza en **python2.7**, **python3.6** y **python3.7**. ● **python2.7**, **python3.6**, **python3.7**, **pytorch1.4-python3.7** y **pytorch1.5-python3.7** indican que el modelo puede ejecutarse tanto en CPU como en GPU. ● El tiempo de ejecución predeterminado es **python2.7**.
MindSpore	aarch64 mindspore_1.7.0-cpu-py_3.7-ubuntu_18.04-x86_64 (recommended)	AArch64 solo puede ejecutarse en chips D310.

1.17 ¿Cuáles son las funciones del entrenamiento y la inferencia de ModelArts?

El entrenamiento de ModelArts incluye ExeML, gestión de entrenamiento y grupos de recursos dedicados (para desarrollo/entrenamiento).

La inferencia de ModelArts incluye despliegue y gestión de aplicaciones de IA.

1.18 ¿Cómo puedo ver un ID de cuenta y un ID de usuario de IAM?

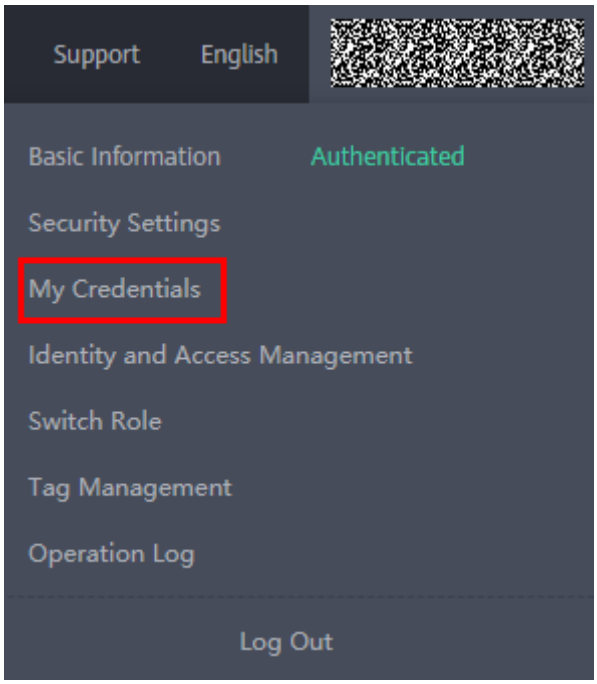
1. Utilice su cuenta de IAM para iniciar sesión en **Huawei Cloud**.
2. En la esquina superior derecha de la página, haga clic en **Console**. Se muestra la consola de gestión de Huawei Cloud.

   Figura 1-14 Consola

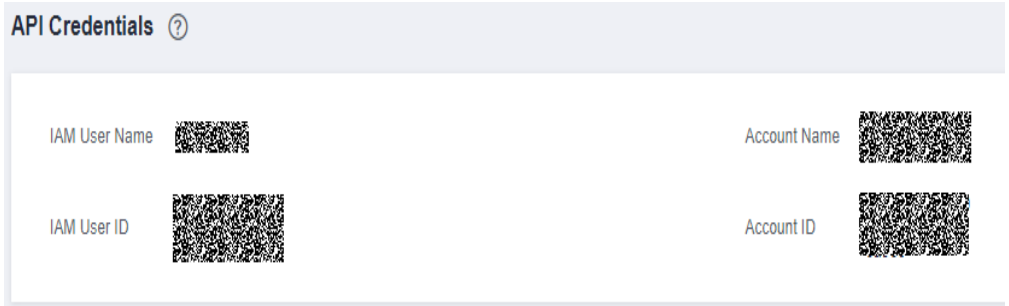
3. Coloque el cursor sobre el nombre de la cuenta en la esquina superior derecha de la consola y elija **My Credentials** en la lista desplegable. Se muestra la página **API Credentials**.

   Figura 1-15 Mis credenciales

4. En la página **API Credentials**, obtenga el nombre de usuario, el ID de usuario, el nombre de cuenta y el ID de cuenta de IAM.

   Figura 1-16 Obtención de las credenciales

1.19 ¿Puede la identificación asistida por IA de ModelArts identificar una etiqueta específica?

Después de que un modelo con varias etiquetas es entrenado y desplegado como un servicio en tiempo real, todas las etiquetas son identificadas. Si solo se necesita identificar un tipo de etiqueta, entrenar un modelo dedicado a identificar la etiqueta. Para acelerar la identificación de la etiqueta, seleccione una variante alta para desplegar el modelo.

1.20 ¿Cómo utiliza ModelArts las etiquetas para gestionar recursos por grupo?

ModelArts puede trabajar con Tag Management Service (TMS). Al crear tareas que consumen recursos de ModelArts, por ejemplo, trabajos de entrenamiento, configure etiquetas para estas tareas de modo que ModelArts pueda usar etiquetas para gestionar recursos por grupo.

ModelArts le permite configurar etiquetas cuando crea trabajos de entrenamiento, instancias de notebook o servicios de inferencia en tiempo real.

Procedimiento operativo

1. **Paso 1 Crear etiquetas predefinidas en TMS**
2. **Paso 2 Agregar una etiqueta a una tarea de ModelArts**
3. **Paso 3 Obtener el uso de recursos de ModelArts por tipo de recurso en TMS**

Paso 1 Crear etiquetas predefinidas en TMS

Inicie sesión en la consola de TMS y cree etiquetas en la página **Predefined Tags**. Las etiquetas creadas son globales y se pueden usar en todas las regiones de Huawei Cloud.

Paso 2 Agregar una etiqueta a una tarea de ModelArts

Al crear una instancia de notebook, trabajo de entrenamiento o servicios de inferencia en tiempo real de ModelArts, configure una etiqueta para la tarea.

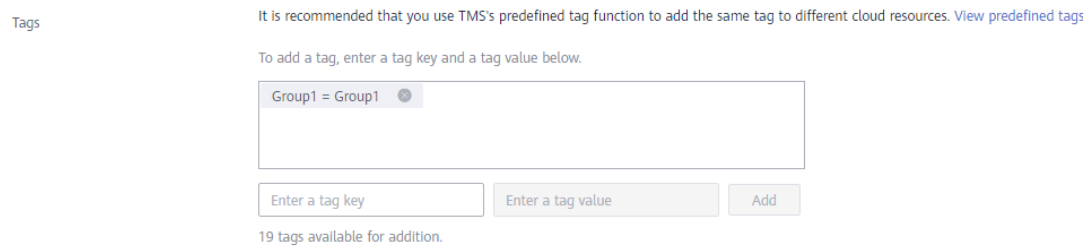
- Agregar una etiqueta a una instancia de notebook de ModelArts.

  Agregue una etiqueta al crear una instancia de notebook. Como alternativa, después de crear una instancia de notebook, agregue una etiqueta en la ficha **Tags** de la página de detalles de la instancia.
- Agregue una etiqueta a un trabajo de entrenamiento de ModelArts.

  Agregue una etiqueta cuando cree un trabajo de entrenamiento. Alternativamente, después de crear un trabajo de entrenamiento, agregue una etiqueta en la ficha **Tags** de la página de detalles del trabajo.
- Agregue una etiqueta a un servicio en tiempo real de ModelArts.

  Agregue una etiqueta cuando cree un servicio en tiempo real. Alternativamente, después de crear un servicio en tiempo real, agregue una etiqueta en la ficha **Tags** de la página de detalles del servicio.

Figura 1-17 Adición de una etiqueta



NOTA

Al agregar una etiqueta a una tarea de ModelArts, puede crear nuevas etiquetas especificando las claves y los valores de las nuevas etiquetas. Las etiquetas creadas aquí solo están disponibles para el proyecto actual.

Paso 3 Obtener el uso de recursos de ModelArts por tipo de recurso en TMS

Inicie sesión en la consola de TMS. En la página **Resources Tag**, vea las tareas de recursos en regiones especificadas según los tipos de recursos y las etiquetas.

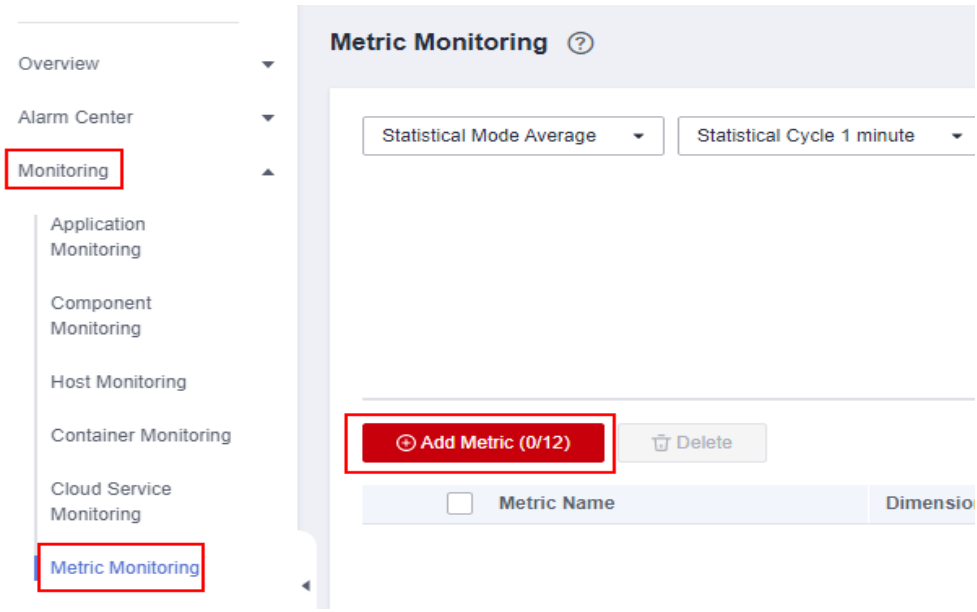
- **Region**: una o más regiones de Huawei Cloud. Para obtener más información, véase **¿Qué son las Regiones y las AZ?**.
- **Resource Type**: Tabla 1-7 muestra una lista de los tipos de recursos que se pueden ver en ModelArts.
- **Resource Tag**: Si no se especifica ninguna etiqueta, se muestran todos los recursos, independientemente de si los recursos están configurados con etiquetas. Se pueden seleccionar una o varias etiquetas para obtener el uso de recursos.

Tabla 1-7 Tipos de recursos que se pueden ver en ModelArts

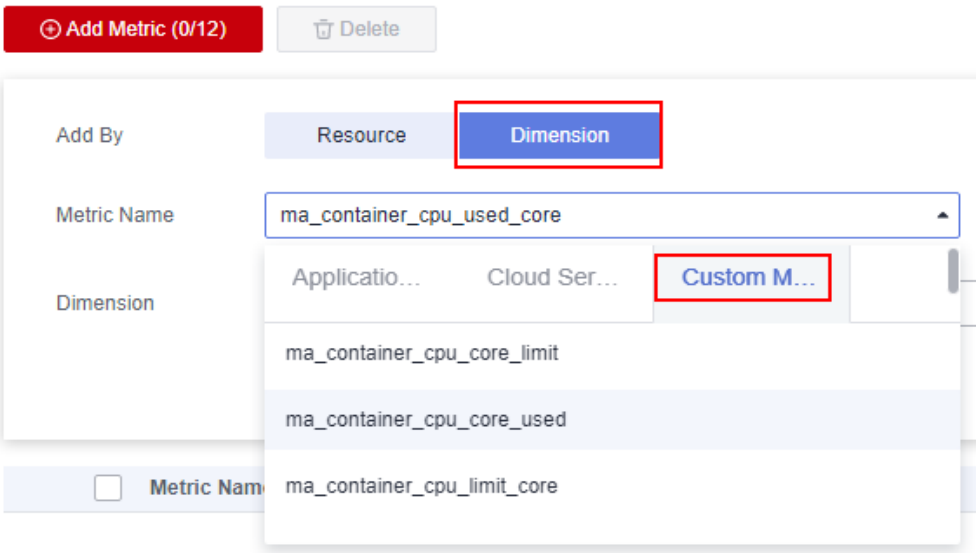
Tipo de recurso	Descripción
ModelArts-Notebook	Instancias de notebook de ModelArts DevEnviron
ModelArts-TrainingJob	Trabajos de entrenamiento de ModelArts
ModelArts-RealtimeService	Servicios de inferencia de ModelArts en tiempo real
ModelArts-ResourcePool	Grupos de recursos dedicados de ModelArts

1.21 ¿Cómo puedo ver todas las métricas de supervisión de ModelArts?

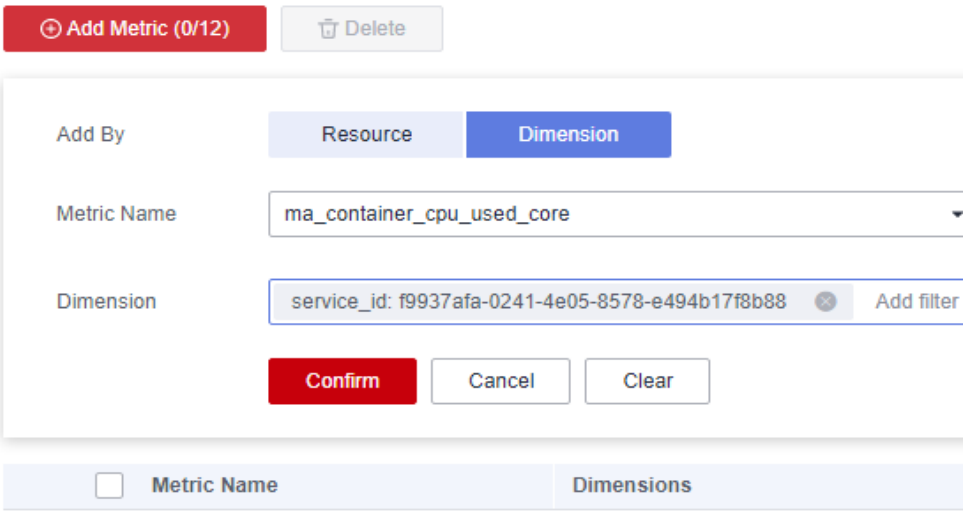
1. Inicie sesión en la consola de Huawei Cloud y busque **AOM** para ir a la consola de AOM.
2. Elija **Monitoring** > **Metric Monitoring**. En la página **Metric Monitoring** que se muestra, haga clic en **Add Metric**.



3. Agregue una métrica para la consulta.



- **Add By**: Seleccione **Dimension**.
- **Metric Name**: Haga clic en **Custom Metrics** y seleccione los que desee para la consulta. Para obtener más información, consulte **Tabla 1-8**, **Tabla 1-9** y **Tabla 1-10**.
- **Dimension**: Ingrese la etiqueta para filtrar la métrica. Para obtener más información, véase **Tabla 1-11**. A continuación se muestra un ejemplo.



4. Haga clic en **Confirm**. Se muestra la información de la métrica.

Tabla 1-8 Métricas de contenedores

Clasificación	Nombre	Métrica	Descripción	Unidad	Rango de valores
CPU	Uso de CPU	ma_container_cpu_util	Uso de CPU de un objeto medido	%	0%–100%
	Núcleos de CPU usados	ma_container_cpu_used_core	Número de núcleos de CPU utilizados por un objeto medido	Núcleos	≥ 0
	Total de núcleos de CPU	ma_container_cpu_limit_core	Número total de núcleos de CPU que se han aplicado a un objeto medido	Núcleos	≥ 1
Memoria	Memoria física total	ma_container_memory_capacity_megabytes	Memoria física total aplicada a un objeto medido	MB	≥ 0

Clasificación	Nombre	Métrica	Descripción	Unidad	Rango de valores
	Uso de la memoria física	ma_container_memory_util	Porcentaje de la memoria física utilizada en relación con la memoria física total	%	0%–100%
	Memoria física usada	ma_container_memory_used_megabytes	Memoria física que ha sido utilizada por un objeto medido (**container_memory_working_set_bytes** en el conjunto de trabajo actual) (Uso de memoria en un conjunto de trabajo = página anónima y cache activos, y la página file-baked ≤ **container_memory_usage_bytes**)	MB	≥ 0
Almacenamiento	Velocidad de lectura de los discos	ma_container_disk_read_kilobytes	Volumen de datos leídos de un disco por segundo	KB/s	≥ 0
	Velocidad de escritura del disco	ma_container_disk_write_kilobytes	Volumen de datos escritos en un disco por segundo	KB/s	≥ 0
Memoria de la GPU	Memoria total de la GPU	ma_container_gpu_mem_total_megabytes	Memoria total de la GPU de un trabajo de entrenamiento	MB	> 0
	Uso de la memoria de GPU	ma_container_gpu_mem_util	Porcentaje de la memoria de la GPU utilizada con respecto a la memoria total de la GPU	%	0%–100%
	Memoria de GPU usada	ma_container_gpu_mem_used_megabytes	Memoria de GPU utilizada por un objeto medido	MB	≥ 0
GPU	Uso de GPU	ma_container_gpu_util	Uso de GPU de un objeto medido	%	0%–100%

Clasificación	Nombre	Métrica	Descripción	Unidad	Rango de valores
	Uso del ancho de banda de la memoria de la GPU	ma_container_gpu_mem_copy_util	Uso del ancho de banda de memoria de la GPU de un objeto medido. Por ejemplo, el ancho de banda de memoria máximo de la NVIDIA GPU V100 es de 900 GB/s. Si el ancho de banda de memoria actual es de 450 GB/s, el uso del ancho de banda de memoria es del 50%.	%	0%–100%
	Uso del codificador de GPU	ma_container_gpu_enc_util	Uso del codificador de GPU de un objeto medido	%	%
	Uso del decodificador de GPU	ma_container_gpu_dec_util	Uso del decodificador de GPU de un objeto medido	%	%
E/S de red	Velocidad de enlace descendente (BPS)	ma_container_network_receive_bytes	Tasa de tráfico entrante de un objeto medido	Bytes/s	≥ 0
	Velocidad de enlace descendente (PPS)	ma_container_network_receive_packets	Número de paquetes de datos recibidos por una NIC por segundo	Paquetes/s	≥ 0
	Tasa de error de enlace descendente	ma_container_network_receive_error_packets	Número de paquetes de error recibidos por una NIC por segundo	Paquetes/s	≥ 0
	Velocidad de enlace ascendente (BPS)	ma_container_network_transmit_bytes	Tasa de tráfico saliente de un objeto medido	Bytes/s	≥ 0
	Tasa de error de enlace ascendente	ma_container_network_transmit_error_packets	Número de paquetes de error enviados por una NIC por segundo	Paquetes/s	≥ 0

Clasificación	Nombre	Métrica	Descripción	Unidad	Rango de valores
	Velocidad de enlace ascendente (PPS)	ma_container_network_transmit_packets	Número de paquetes de datos enviados por una NIC por segundo	Paquetes/s	≥ 0
NPU	Uso de NPU	ma_container_npu_util	Uso de NPU de un objeto medido	0%–100%	%
	Uso de memoria de NPU	ma_container_npu_memory_util	Porcentaje de la memoria de NPU usada respecto a la memoria total de NPU	0%–100%	%
	Memoria usada de NPU	ma_container_npu_memory_used_megabytes	Memoria utilizada de NPU por un objeto medido	≥ 0	MB
	Memoria total de NPU	ma_container_npu_memory_total_megabytes	Memoria total de NPU de un objeto medido	> 0	MB

Tabla 1-9 Métricas de nodo (recogidas solo en grupos de recursos dedicados)

Clasificación	Nombre	Métrica	Descripción	Unidad	Rango de valores
CPU	Total de núcleos de CPU	ma_node_cpu_limit_core	Número total de núcleos de CPU que se han aplicado a un objeto medido	Núcleos	≥ 1
	Núcleos de CPU usados	ma_node_cpu_used_core	Número de núcleos de CPU utilizados por un objeto medido	Núcleos	≥ 0
	Uso de CPU	ma_node_cpu_util	Uso de CPU de un objeto medido	%	0%–100%
Memoria	Uso de la memoria física	ma_node_memory_util	Porcentaje de la memoria física utilizada en relación con la memoria física total	%	0%–100%

Clasificación	Nombre	Métrica	Descripción	Unidad	Rango de valores
	Memoria física total	ma_node_memory_total_megabytes	Memoria física total aplicada a un objeto medido	MB	≥ 0
E/S de red	Velocidad de enlace descendente (BPS)	ma_node_network_receive_rate_bytes_seconds	Tasa de tráfico entrante de un objeto medido	Bytes/s	≥ 0
	Velocidad de enlace ascendente (BPS)	ma_node_network_transmit_rate_bytes_seconds	Tasa de tráfico saliente de un objeto medido	Bytes/s	≥ 0
Almacenamiento	Velocidad de lectura de los discos	ma_node_disk_read_rate_kilobytes_seconds	Volumen de datos leídos de un disco por segundo (Solo se recopilan los discos de datos utilizados por contenedores.)	KB/s	≥ 0
	Velocidad de escritura del disco	ma_node_disk_write_rate_kilobytes_seconds	Volumen de datos escritos en un disco por segundo (Solo se recopilan los discos de datos utilizados por contenedores.)	KB/s	≥ 0
	Caché total	ma_node_cache_space_capacity_megabytes	Caché total del espacio de Kubernetes	MB	≥ 0
	Caché Usada	ma_node_cache_space_used_capacity_megabytes	Caché usada del espacio de Kubernetes	MB	≥ 0
	Espacio total del contenedor	ma_node_container_space_capacity_megabytes	Espacio total del contenedor	MB	≥ 0

Clasificación	Nombre	Métrica	Descripción	Unidad	Rango de valores
	Espacio usado de contenedor	ma_node_container_space_used_capacity_megabytes	Espacio usado de contenedor	MB	≥ 0
GPU	Uso de GPU	ma_node_gpu_util	Uso de GPU de un objeto medido	%	0%–100%
	Memoria total de la GPU	ma_node_gpu_mem_total_megabytes	Memoria total de la GPU de un objeto medido	MB	> 0
	Uso de la memoria de GPU	ma_node_gpu_mem_util	Porcentaje de la memoria de la GPU utilizada con respecto a la memoria total de la GPU	%	0%–100%
	Memoria de GPU usada	ma_node_gpu_mem_used_megabytes	Memoria de GPU utilizada por un objeto medido	MB	≥ 0
	Tareas en una GPU compartida	node_gpu_share_job_count	Número de tareas que se ejecutan en una GPU compartida	Número	≥ 0
NPU	Uso de NPU	ma_node_npu_util	Uso de NPU de un objeto medido	%	0%–100%
	Uso de memoria de NPU	ma_node_npu_memory_util	Porcentaje de la memoria de NPU usada respecto a la memoria total de NPU	%	0%–100%
	Memoria usada de NPU	ma_node_npu_memory_used_megabytes	Memoria utilizada de NPU por un objeto medido	MB	≥ 0
	Memoria total de NPU	ma_node_npu_memory_total_megabytes	Memoria total de NPU de un objeto medido	MB	> 0

Clasificación	Nombre	Métrica	Descripción	Unidad	Rango de valores
InfiniBand o RoCE network	Cantidad total de datos recibidos por una NIC	ma_node_infiniband_port_received_data_bytes_total	Número total de octetos de datos, dividido por 4, (contando en palabras dobles, 32 bits), recibidos en todos los VL desde el puerto.	contando en palabras dobles, 32 bits	≥ 0
	Cantidad total de datos enviados por una NIC	ma_node_infiniband_port_transmitted_data_bytes_total	El número total de octetos de datos, dividido por 4, (contando en palabras dobles, 32 bits), transmitidos en todos los VL desde el puerto.	contando en palabras dobles, 32 bits	≥ 0

Tabla 1-10 Perfilado y diagnóstico (GPU | IB, recopilados solo en grupos de recursos dedicados)

Clasificación	Nombre	Métrica	Descripción	Unidad	Rango de valores
GPU	Temperatura de la GPU	DCGM_FI_DEV_GPU_TEMP	Temperatura de la GPU	°C	Número natural
	Potencia de la GPU	DCGM_FI_DEV_POWER_USAGE	Potencia de la GPU	W	Número natural
	Temperatura de la memoria	DCGM_FI_DEV_MEMORY_TEMP	Temperatura de la memoria	°C	Número natural

Clasificación	Nombre	Métrica	Descripción	Unidad	Rango de valores
	Actividad del motor de gráficos	DCGM_FI_PROF_GR_ENGINE_ACTIVE	Porcentaje del tiempo cuando el motor gráfico o de computación está en el estado activo dentro de un período de tiempo. Este es un valor promedio de todos los motores gráficos o de computación. Un motor gráfico o de computación activo indica que el contexto gráfico o de computación está asociado con un subproceso y que el contexto gráfico o de computación está ocupado.	Porcentaje (fracción)	0-1.0
	Ocupación de SM	DCGM_FI_PROF_SM_OCCUPANCY	Relación entre el número de haces de hilos que residen en el SM y el número máximo de haces de hilos que pueden residir en el SM dentro de un período de tiempo Este es un valor promedio de todos los SM dentro de un periodo de tiempo. Un valor alto no significa un uso alto de GPU. Solo cuando el ancho de banda de memoria de la GPU es limitado, un alto valor de cargas de trabajo (**DCGM_FI_PROF_DRAM_ACTIVE**) indica un uso más eficiente de la GPU.	Porcentaje (fracción)	0-1.0

Clasificación	Nombre	Métrica	Descripción	Unidad	Rango de valores
	Actividad de Tensor	DCGM_FI_PROF_PIPE_TENSOR_ACTIVE	Fracción del período durante el cual el tubo tensor (HMMA/IMMA) está activo Este es un valor promedio dentro de un período de tiempo, no un valor instantáneo. Un valor más alto indica una mayor utilización de núcleos tensores. El valor 1 (100%) indica que se envía una instrucción de tensor cada ciclo de instrucción en todo el período (una instrucción se completa en dos ciclos). Si el valor es 0.2 (20%), las posibles causas son las siguientes: Durante todo el período, el 20% de los núcleos tensores de SM funcionan al 100% de utilización. Durante todo el período, todos los núcleos tensores de SM funcionan con una utilización del 20%. Durante 1/5 de todo el período, todos los núcleos tensores de SM funcionan al 100% de utilización. Otras combinaciones	Porcentaje (fracción)	0-1.0

Clasificación	Nombre	Métrica	Descripción	Unidad	Rango de valores
	Uso de BW de memoria	DCGM_FI_PROF_DRAM_ACTIVE	Porcentaje del tiempo para enviar o recibir datos desde la memoria del dispositivo en un período de tiempo Este es un valor promedio dentro de un período de tiempo, no un valor instantáneo. Un valor más alto indica una mayor utilización de la memoria del dispositivo. El valor 1 (100%) indica que una instrucción de DRAM se ejecuta una vez por ciclo a lo largo de un período (el valor máximo puede alcanzarse en un pico de aproximadamente 0.8). Si el valor es 0.2 (20%), indicando que los datos se leen o se escriben en la memoria del dispositivo durante el 20% del ciclo dentro de un periodo.	Porcentaje (fracción)	0-1.0

Clasificación	Nombre	Métrica	Descripción	Unidad	Rango de valores
	Actividad del motor FP16	DCGM_FI_PROF_PIPE_FP16_ACTIVE	Fracción del período durante el cual el tubo FP16 (media precisión) está activo Este es un valor promedio dentro de un período de tiempo, no un valor instantáneo. Un valor mayor indica un mayor uso de núcleos de FP16. El valor 1 (100%) indica que la instrucción de FP16 se ejecuta cada dos ciclos (por ejemplo, tarjetas Volta) en un periodo. Si el valor es 0.2 (20%), las posibles causas son las siguientes: During the entire period, 20% of the SM FP16 cores run at 100% utilization. Durante todo el período, todos los núcleos de SM de FP16 funcionan con una utilización del 20%. Durante 1/5 de todo el período, todos los núcleos de SM de FP16 funcionan al 100% de utilización. Otras combinaciones	Porcentaje (fracción)	0-1.0

Clasificación	Nombre	Métrica	Descripción	Unidad	Rango de valores
	Actividad del motor FP32	DCGM_FI_PROF_PIPE_FP32_ACTIVE	Fracción del período durante el cual el tubo de adición múltiple fusionado (FMA) está activo. Multiplicar-añadir se aplica a FP32 (precisión simple) y enteros. Este es un valor promedio dentro de un período de tiempo, no un valor instantáneo. Un valor mayor indica un mayor uso de núcleos de FP32. El valor 1 (100%) indica que la instrucción de FP32 se ejecuta cada dos ciclos (por ejemplo, tarjetas Volta) en un periodo. Si el valor es 0.2 (20%), las posibles causas son las siguientes: Durante todo el período, el 20% de los núcleos de SM de FP32 funcionan al 100% de utilización. Durante todo el período, todos los núcleos de SM de FP32 funcionan con una utilización del 20%. Durante 1/5 de todo el período, todos los núcleos de SM de FP32 funcionan al 100% de utilización. Otras combinaciones	Porcentaje (fracción)	0-1.0

Clasificación	Nombre	Métrica	Descripción	Unidad	Rango de valores
	Actividad del motor FP64	DCGM_FI_PROF_PIPE_FP64_ACTIVE	Fracción del período durante el cual el tubo FP64 (doble precisión) está activo Este es un valor promedio dentro de un período de tiempo, no un valor instantáneo. Un valor mayor indica un mayor uso de núcleos de FP64. El valor 1 (100%) indica que la instrucción de FP64 se ejecuta cada cuatro ciclos (por ejemplo, tarjetas Volta) en un período. Si el valor es 0.2 (20%), las posibles causas son las siguientes: Durante todo el período, el 20% de los núcleos de SM de FP64 funcionan al 100% de utilización. Durante todo el período, todos los núcleos de SM de FP64 funcionan con una utilización del 20%. Durante 1/5 de todo el período, todos los núcleos de SM de FP64 funcionan al 100% de utilización. Otras combinaciones	Porcentaje (fracción)	0-1.0

Clasificación	Nombre	Métrica	Descripción	Unidad	Rango de valores
	Actividad de SM	DCGM_FI_PROF_SM_ACTIVE	Fracción del tiempo durante el cual al menos un haga de hilos está activo en un SM dentro de un periodo de tiempo. Este es un valor promedio de todos los SM y es insensible al número de hilos en cada bloque. Un paquete de subprocesos está activo después de ser programado y asignado con recursos. El conjunto de hilos puede estar en el estado informático o en un estado no informático (por ejemplo, esperando una solicitud de memoria). Si el valor es inferior a 0.5, las GPU no se utilizan de manera eficiente. El valor debe ser mayor que 0.8. Por ejemplo, una GPU tiene N SM: Una función de núcleo utiliza N bloques de subproceso para ejecutarse en todos los SM en un periodo. En este caso, el valor es 1 (100%). Una función del núcleo ejecuta N/5 bloques de subprocesos en un periodo. En este caso, el valor es 0.2. Una función del núcleo utiliza N bloques de subprocesos y ejecuta solo 1/5 de ciclos en un	Porcentaje (fracción)	0-1.0

Clasificación	Nombre	Métrica	Descripción	Unidad	Rango de valores
			periodo. En este caso, el valor es 0.2.		
	Ancho de banda de PCIe	DCGM_FI_PROF_PCIE_TX_BYTES DCGM_FI_PROF_PCIE_RX_BYTES	Velocidad de datos transmitidos o recibidos a través del bus de PCIe, incluido el encabezado del protocolo y la carga útil de datos Este es un valor promedio dentro de un período de tiempo, no un valor instantáneo. La tasa se promedia durante el período. Por ejemplo, si se transmite 1 GB de datos dentro de 1 segundo, la velocidad de transmisión es 1 GB/s independientemente de si los datos se transmiten a una velocidad o ráfaga constante. Teóricamente, el ancho de banda máximo PCIe Gen3 es de 985 MB/s por canal.	Bytes/s	≥ 0

Clasificación	Nombre	Métrica	Descripción	Unidad	Rango de valores
	Ancho de banda de NVLink	DCGM_FI_PROF_NVLINK_RX_BYTES DCGM_FI_PROF_NVLINK_TX_BYTES	Velocidad a la que se transmiten o reciben datos con NVLink, excluido el encabezado del protocolo Este es un valor promedio dentro de un período de tiempo, no un valor instantáneo. La tasa se promedia durante el período. Por ejemplo, si se transmite 1 GB de datos dentro de 1 segundo, la velocidad de transmisión es 1 GB/s independientemente de si los datos se transmiten a una velocidad o ráfaga constante. Teóricamente, el ancho de banda máximo de NVLink Gen2 es de 25 GB/s por enlace en cada dirección.	Bytes/s	≥ 0
InfiniBand o RoCE network	PortXmitData	infiniband_port_xmit_data_total	El número total de octetos de datos, dividido por 4, (contando en palabras dobles, 32 bits), transmitidos en todos los VL desde el puerto.	Recuento total	Número natural
	PortRcvData	infiniband_port_rcv_data_total	Número total de octetos de datos, dividido por 4, (contando en palabras dobles, 32 bits), recibidos en todos los VL desde el puerto.	Recuento total	Número natural
	SymbolErrorCounter	infiniband_symbol_error_counter_total	Número total de errores de enlace menores detectados en uno o más carriles físicos.	Recuento total	Número natural

Clasificación	Nombre	Métrica	Descripción	Unidad	Rango de valores
	LinkErrorRecoveryCounter	infiniband_link_error_recovery_counter_total	Número total de veces que la máquina de estado de entrenamiento de puerto ha completado con éxito el proceso de recuperación de error de enlace.	Recuento total	Número natural
	PortRcvErrors	infiniband_port_rcv_errors_total	Número total de paquetes que contienen errores recibidos en el puerto, incluido: Errores físicos locales (ICRC, VCRC, LPCRC y todos los errores físicos que provocan la entrada en los estados BAD PACKET o BAD PACKET DISCARD de la máquina de estado del receptor de paquetes) Errores mal formados del paquete de datos (LVer, longitud, VL) Errores de paquetes de enlace mal formados (operando, longitud, VL) Paquetes descartados debido al desbordamiento de búfer (desbordamiento)	Recuento total	Número natural
	LocalLinkIntegrityErrors	infiniband_local_link_integrity_errors_total	Este contador indica el número de reintentos iniciados por un receptor de capa de transferencia de enlace.	Recuento total	Número natural
	PortRcvRemotePhysicalErrors	infiniband_port_rcv_remote_physical_errors_total	Número total de paquetes marcados con el delimitador EBP recibidos en el puerto.	Recuento total	Número natural

Clasificación	Nombre	Métrica	Descripción	Unidad	Rango de valores
	PortRcvSwitchRelayErrors	infiniband_port_rcv_switch_relay_errors_total	Número total de paquetes recibidos en el puerto que fueron descartados cuando no pudieron ser reenviados por el switch relay por las siguientes razones: Asignación de DLID Asignación de VL Bucle (puerto de salida = puerto de entrada)	Recuento total	Número natural
	PortXmitWait	infiniband_port_transmit_wait_total	El número de ticks durante los cuales el puerto tenía datos para transmitir, pero no se envió ningún dato durante todo el tick (ya sea por falta de créditos o por falta de arbitraje).	Recuento total	Número natural
	PortXmitDiscards	infiniband_port_xmit_discards_total	Número total de paquetes salientes descartados por el puerto porque el puerto está inactivo o congestionado.	Recuento total	Número natural

Tabla 1-11 Nombres de las métricas

Clasificación	Métrica	Descripción
Métricas de contenedores	modelarts_service	Servicio al que pertenece un contenedor, que puede ser **notebook**, **train** o **infer**
	instance_name	Nombre del pod al que pertenece el contenedor
	service_id	ID de instancia o trabajo que se muestra en la página, por ejemplo, **cf55829e-9bd3-48fa-8071-7ae870dae93a** para un entorno de desarrollo **9f322d5a-b1d2-4370-94df-5a87de27d36e** para un trabajo de entrenamiento
	node_ip	Dirección IP del nodo al que pertenece el contenedor

Clasificación	Métrica	Descripción
	container_id	ID del contenedor
	cid	ID del clúster
	container_name	Nombre del contenedor
	project_id	ID de proyecto de la cuenta a la que pertenece el usuario
	npu_id	Identificación de la tarjeta Ascend, por ejemplo **davinci0**
	gpu_uuid	UUID de la GPU utilizada por el contenedor
Métricas de nodos	cid	ID del clúster de CCE al que pertenece el nodo
	node_ip	Dirección IP del nodo
	pool_id	ID de un grupo de recursos correspondiente a un grupo de recursos dedicado físico
	project_id	ID de proyecto del usuario en un grupo de recursos físico dedicado
	npu_id	Identificación de la tarjeta Ascend, por ejemplo **davinci0**
	gpu_uuid	UUID de una GPU de nodo
	device_name	Nombre del dispositivo de una NIC de red InfiniBand o RoCE
Perfilado y diagnóstico	cid	ID del clúster de CCE al que pertenece el nodo donde reside la GPU
	node_ip	Dirección IP del nodo donde reside la GPU
	pool_id	ID de un grupo de recursos correspondiente a un grupo de recursos dedicado físico
	project_id	ID de proyecto del usuario en un grupo de recursos físico dedicado
	gpu_uuid	UUID de GPU
	device_name	Nombre del dispositivo de una NIC de red InfiniBand o RoCE

1.22 ¿Por qué el trabajo sigue en cola cuando los recursos son suficientes?

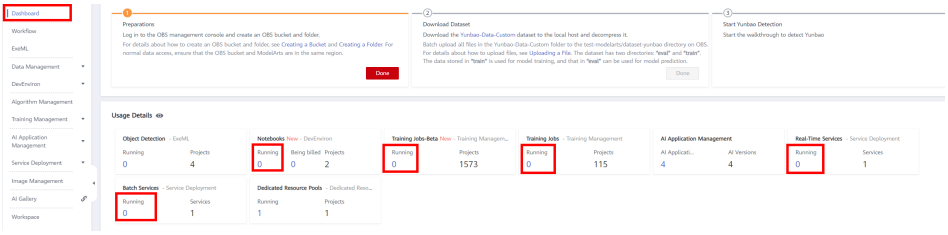
- Si se utiliza un grupo de recursos públicos, los recursos pueden ser utilizados por otros usuarios. Por favor, espere o encuentre soluciones en **¿Por qué un trabajo de entrenamiento siempre está en cola?**.

- Si se utiliza un grupo de recursos dedicado, realice las siguientes operaciones:

  a. Compruebe si se están ejecutando otros trabajos (incluidos los trabajos de inferencia, los trabajos de entrenamiento y los trabajos de entorno de desarrollo) en el grupo de recursos dedicado.

  En la página **Dashboard**, puede ir a la página de detalles de los trabajos o instancias en ejecución para comprobar si se utiliza el grupo de recursos dedicado. Puede detenerlos en función de sus necesidades para liberar recursos.

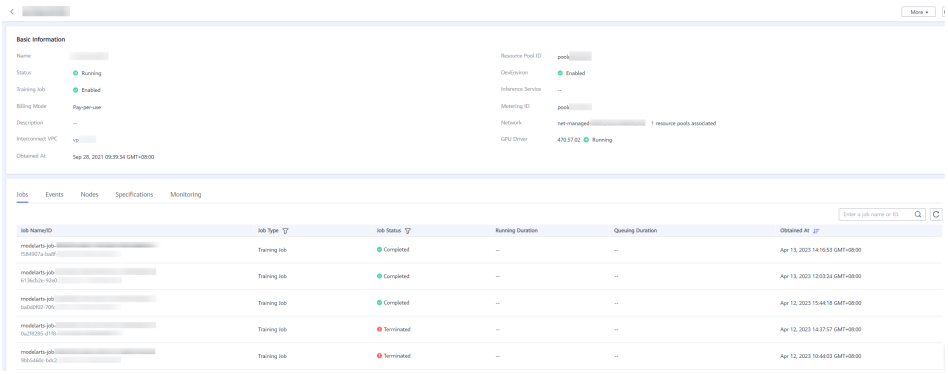
  Figura 1-18 Panel

  b. Vaya a la página de detalles del grupo de recursos dedicado para comprobar si hay otros trabajos de cola.

  En caso afirmativo, el nuevo trabajo debe estar en cola.

  Figura 1-19 Trabajos de cola

  c. Compruebe si los recursos están fragmentados.

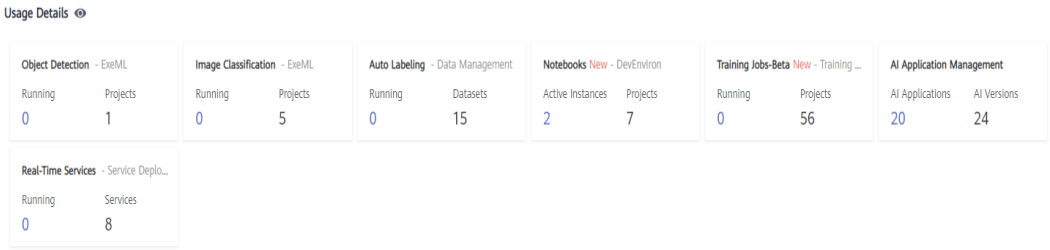
  Por ejemplo, el clúster tiene dos nodos y hay cuatro tarjetas inactivas en cada nodo. Sin embargo, su trabajo requiere ocho tarjetas en un nodo. En este caso, los recursos inactivos no se pueden asignar a su trabajo.

2 Facturación

2.1 ¿Cómo puedo ver los trabajos de ModelArts que se están facturando?

Inicie sesión en la consola de gestión de ModelArts. En el panel de navegación de la izquierda, haga clic en **Dashboard** y vea los trabajos que se están facturando. Si el número de instancias activas es mayor que 0, hay instancias que se están facturando. Vaya a la página de lista de instancias y deténgalas según los requisitos del sitio. Por ejemplo, si se está facturando una instancia de notebook, elija **DevEnviron** > **Notebook** y detenga la instancia de notebook en ejecución.

Figura 2-1 Consultar trabajos que se están facturando



Cuando utilice ModelArts se facturarán los siguientes conceptos:

- ExeML: Se le facturará por ejecutar un trabajo ExeML. Para detener la facturación, detenga el trabajo.
- Instancias del notebook:
  - Se facturan las instancias de notebook en ejecución. Para detener la facturación de una instancia de notebook, deténgala o elimínela.
  - El almacenamiento de EVS que seleccionó al crear una instancia de notebook se factura por separado. Se factura continuamente incluso después de que se detenga la instancia del notebook adjunto. Para dejar de facturar el almacenamiento de EVS, elimine la instancia de notebook de destino.
  - Se le cobrará por una variante de pago cuando pruebe CodeLab. Para dejar de facturar la variante, detenga la instancia del notebook en la página de JupyterLab.

- Trabajos de entrenamiento: Se facturan los trabajos de entrenamiento en ejecución. Para dejar de facturar un trabajo de entrenamiento, deténgalo.
- Despliegue de modelos: si un modelo se despliega como servicio en tiempo real, por lotes o perimetral, se facturará el servicio. Para detener la facturación, detenga el servicio desplegado.
- Evaluación de modelo: Se le facturará por crear un trabajo de evaluación de modelo. Para detener la facturación, detenga el trabajo.
- OBS: OBS se utiliza para almacenar datos y se facturará por separado por gestión de datos, entorno de desarrollo, entrenamiento, despliegue de modelos y ExeML. Para dejar de facturar los recursos de OBS, vaya a la consola de gestión de OBS y borre los datos en OBS.

⚠ ATENCIÓN

Además de los conceptos de facturación que se muestran en la página **Dashboard** de ModelArts, el almacenamiento de OBS y de EVS se facturará por separado si se utilizan.

- Para dejar de facturar los recursos de OBS, vaya a la consola de gestión de OBS y borre los datos en OBS.
- Para dejar de facturar el almacenamiento de EVS, vaya a la consola de gestión de ModelArts y elimine las instancias de notebook con almacenamiento de EVS.

2.2 ¿Cómo puedo ver los detalles de consumo de ModelArts?

En **Billing Center**, los datos de consumo de ModelArts se muestran cada hora. Puede cambiar a la página **Expenditures** para ver las tarifas consumidas para cada trabajo.

Método de consulta:

1. En la página **Billing Center**, seleccione **Bills > Expenditures**. Haga clic en la ficha **Yearly/Monthly** o **Pay-per-Use** según el tipo de paquete.
2. A continuación, puede ver los gastos de cada tarea de todos los servicios de la lista. Busque la fila donde reside la transacción de destino y haga clic en **Details** en la columna **Operation**.
3. Vea información relacionada en la página **Transaction Overview** mostrada, como **Consumed On**, **Name/ID**, **Specifications** y **Amount**.

2.3 ¿Se me cobrará por cargar conjuntos de datos a ModelArts?

No se le cobrará por la gestión de conjuntos de datos y el etiquetado de ModelArts, pero los conjuntos de datos se almacenan en OBS, y la gestión de conjuntos de datos de ModelArts utiliza los datos almacenados en OBS. No se le facturará por cargar los conjuntos de datos, pero se le facturará por el almacenamiento en OBS. Para obtener más información, consulte el calculador de **detalles de precios de OBS** y cree bucket de OBS para almacenar datos utilizados por ModelArts.

2.4 ¿Qué debo hacer para evitar la facturación innecesaria después de etiquetar conjuntos de datos y salir?

El etiquetado de conjuntos de datos es gratuito. Sin embargo, OBS le factura el espacio de almacenamiento utilizado para almacenar los conjuntos de datos. Para evitar una facturación innecesaria, se recomienda acceder a la OBS Console y eliminar los datos y los bucket de OBS en la ruta de acceso de OBS para almacenar los conjuntos de datos después del etiquetado de datos.

2.5 ¿Cómo dejo de facturar un proyecto ExeML de ModelArts?

Elimine el proyecto ExeML creado.

Para ser específicos, en la lista de proyectos ExeML de la consola de ModelArts, busque la fila donde reside el proyecto ExeML que desea detener y haga clic en **Delete** en la columna **Operation**.

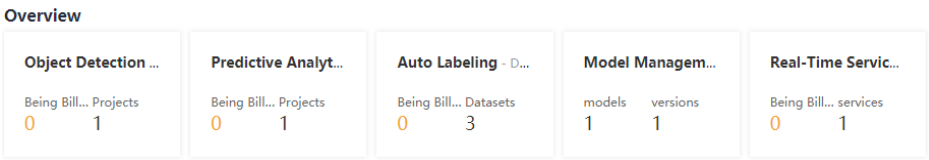
2.6 ¿Cómo dejo de facturar si no uso ModelArts?

Detenga o elimine los trabajos en ejecución creados en ModelArts, especialmente los trabajos en instancias de notebook, los trabajos de visualización y los trabajos para desplegar un modelo como servicios en tiempo real. Además, cambie a la consola de OBS y elimine los datos y directorios correspondientes porque los datos de ModelArts se almacenan en OBS.

Solución:

Inicie sesión en la consola de ModelArts. En el panel de navegación izquierdo, haga clic en **Dashboard**. En el área **Overview**, puede ver los trabajos que se están facturando. Luego, detenga los trabajos relacionados según sea necesario.

Figura 2-2 Consultar trabajos que se están facturando



- En el panel de navegación izquierdo de la consola de ModelArts, elija **DevEnviron** > **Notebook** para cambiar a la página **Notebook** y comprobar si hay instancias de notebook en el estado **Running**. Si es así, haga clic en **Stop** en la columna **Operation**. A continuación, la facturación se detiene en consecuencia. Compruebe si hay instancias de notebook que utilizan el almacenamiento de EVS. Si es así, detenga y elimine las instancias del notebook. A continuación, la facturación del EVS se detendrá en consecuencia.
- Seleccione **Training Management** > **Training Jobs** y compruebe si hay trabajos en ejecución. Si hay trabajos en ejecución, haga clic en **Stop** en la columna **Operation** para detener los trabajos.

- Seleccione **Training Management** > **Training Jobs**, haga clic en la ficha **Visualization Jobs** y compruebe si hay trabajos en ejecución. Si hay trabajos en ejecución, haga clic en **Stop** en la columna **Operation** para detener los trabajos.
- En el panel de navegación izquierdo de la consola de ModelArts, elija **Service Deployment** > **Real-Time Services** y compruebe si hay trabajos en estado **Running**. Si hay trabajos en ejecución, haga clic en **Stop** en la columna **Operation** para detener los trabajos.
- En el panel de navegación izquierdo de la consola de ModelArts, seleccione **Service Deployment** > **Batch Services** y compruebe si hay trabajos en estado **Running**. Si hay trabajos en ejecución, haga clic en **Stop** en la columna **Operation** para detener los trabajos.

2.7 ¿Cómo se facturan los trabajos de entrenamiento?

Los trabajos de entrenamiento de ModelArts se facturan sobre una base de pago por uso. El precio varía según el tipo de grupo de recursos. Un trabajo de entrenamiento se factura en función de los recursos consumidos durante cada ejecución. Cuando un trabajo de entrenamiento está en el estado **Successful** o **Failed**, se detiene la facturación. Se está facturando un trabajo de entrenamiento para correr.

2.8 ¿Por qué continúa la facturación después de que se eliminen todos los proyectos?

Incluso si los proyectos de ExeML, las instancias de notebook, los trabajos de entrenamiento o los servicios de ModelArts se detienen y no se muestra ningún elemento de cargo en la página **Dashboard**, es posible que se siga facturando a la cuenta por el almacenamiento de OBS que se esté utilizando.

Las causas posibles son:

1. Ha cargado datos a OBS para su almacenamiento cuando usa ModelArts y se factura el almacenamiento de OBS. En este caso, vaya a la consola de gestión de OBS y elimine los datos, carpetas y bucket de OBS que ya no sean necesarios.
2. Ha seleccionado el almacenamiento de EVS al crear una instancia de notebook y el almacenamiento se factura por separado incluso después de detener la instancia de notebook. En este caso, elimine la instancia del notebook.
3. Ha cambiado la variante a uno de carga cuando experimenta CodeLab. En este caso, vaya a la página CodeLab y haga clic en  en la esquina superior derecha para detener la instancia del notebook.

2.9 ¿Necesito comprar recursos de pago por uso?

Los recursos de pago por uso no necesitan ser comprados. Se le cobrará en función del uso de recursos.

3 ExeML

3.1 Consultoría funcional

3.1.1 ¿Qué es ExeML?

ExeML es el proceso de automatizar el diseño de modelos, ajuste de parámetros y entrenamiento de modelos, compresión y despliegue con los datos etiquetados. El proceso está libre de codificación y no requiere la experiencia de los desarrolladores en el desarrollo de modelos.

Los usuarios que no tienen capacidad de codificación pueden usar las funciones de etiquetado, entrenamiento de modelo con un solo clic y despliegue de modelos de ExeML para crear modelos de IA.

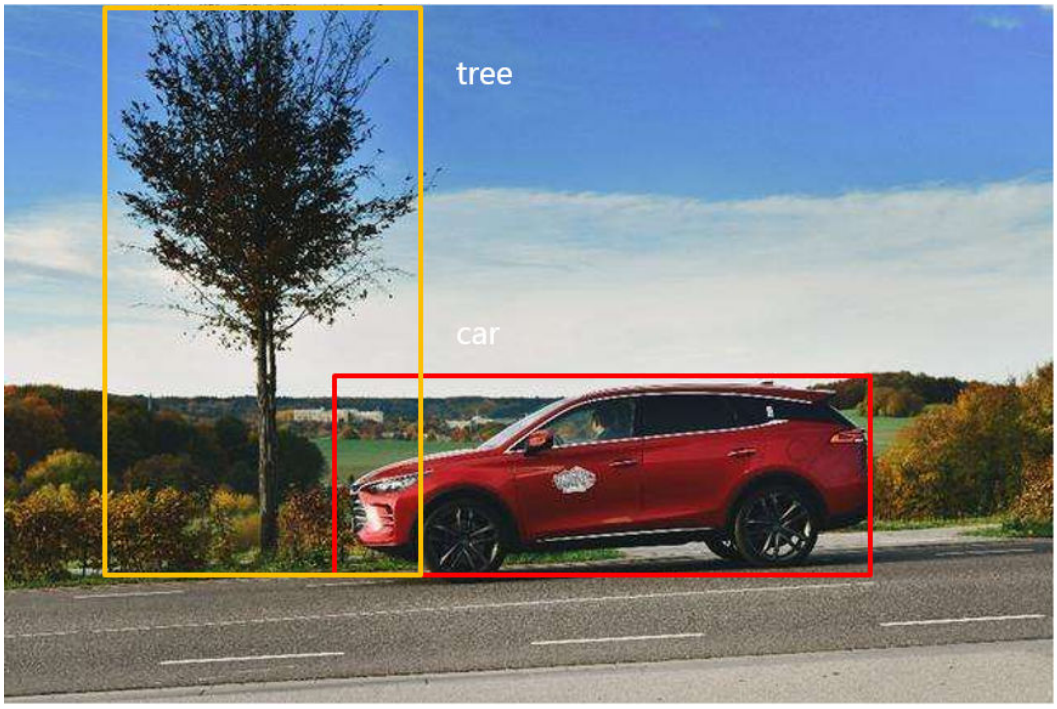
3.1.2 ¿Qué son la clasificación de imágenes y la detección de objetos?

La clasificación de imágenes es un método de procesamiento de imágenes que separa diferentes clases de objetivos de acuerdo con las características reflejadas en las imágenes. Con el análisis cuantitativo de imágenes, clasifica una imagen o cada píxel o área de una imagen en diferentes categorías para reemplazar la interpretación visual humana. En general, la clasificación de imágenes tiene como objetivo identificar una clase, estado o escena en una imagen. Es aplicable a escenarios en los que una imagen contiene solo un objeto. **Figura 3-1** muestra un ejemplo de identificación de un coche en una imagen.

Figura 3-1 Clasificación de imágenes

La detección de objetos es uno de los problemas clásicos en la visión por computadora. Tiene la intención de etiquetar objetos con marcos e identificar las clases de objetos en una imagen. En general, si una imagen contiene varios objetos, la detección de objetos puede identificar la ubicación, la cantidad y el nombre de cada objeto en la imagen. Es adecuado para escenarios en los que una imagen contiene varios objetos. **Figura 3-2** muestra un ejemplo de identificación de un árbol y un coche en una imagen.

Figura 3-2 Detección de objetos



3.1.3 ¿Cuáles son las diferencias entre ExeML y los algoritmos suscritos?

ModelArts ofrece diferentes modos de desarrollo de IA para desarrolladores nuevos y experimentados.

- Para los nuevos desarrolladores, puede usar ExeML para desarrollar modelos sin codificación. Cuando utiliza ExeML, el sistema selecciona automáticamente los algoritmos y parámetros adecuados para el entrenamiento del modelo.
- Para los desarrolladores de IA experimentados, puede seleccionar algoritmos suscritos para el entrenamiento de modelos. Además, puede personalizar los parámetros necesarios para el entrenamiento.

3.2 Preparación de datos

3.2.1 ¿Cuáles son los requisitos para los datos de entrenamiento cuando crea un proyecto de análisis predictivo en ExeML?

Requisitos sobre conjuntos de datos

- El conjunto de datos consta de letras, dígitos, guiones (-) y guiones bajos (_), y debe estar en formato CSV. Los archivos de datos no se pueden almacenar en el directorio raíz de un bucket de OBS, sino en una carpeta del bucket de OBS, por ejemplo **/obs-xxx/data/input.csv**.
- Utilice caracteres de nueva línea (\n o LF) para separar líneas y comas (,) para separar columnas en el contenido del archivo. El contenido del archivo no puede incluir símbolos que no sean inglés (por ejemplo, caracteres chinos). El contenido de la columna no puede contener caracteres especiales como comas, saltos de línea o comillas. Se recomienda que el contenido de la columna consista únicamente en letras y números.
- Entrenamiento de datos
  - El número de columnas en los datos de entrenamiento debe ser el mismo, y debe haber al menos 100 registros de datos (una característica con valores diferentes se considera como registros de datos diferentes).
  - Las columnas de entrenamiento no pueden contener formatos de marca de tiempo (como aa-mm-dd y aaaa-mm-dd).
  - Si una columna tiene solo un valor, la columna se considera no válida. Asegúrese de que hay al menos dos valores en la columna de etiqueta y de que no faltan datos.

    NOTA

    La columna de etiqueta es el destino de entrenamiento especificado en una tarea de entrenamiento. Es la salida (elemento de predicción) para el modelo entrenado usando el conjunto de datos.
  - Además de la columna de etiqueta, el conjunto de datos debe contener al menos dos columnas de elemento válidas. Asegúrese de que hay al menos dos valores en cada columna de elemento y de que el porcentaje de datos que faltan debe ser inferior al 10%.
  - El archivo CSV no puede contener un encabezado de tabla o el entrenamiento fallará.

3.2.2 ¿Qué formatos de imágenes son compatibles con los proyectos de detección de objetos o clasificación de imágenes?

Se admiten imágenes en formato JPG, JPEG, PNG o BMP.

3.3 Creación de un proyecto

3.3.1 ¿Hay un límite en el número de proyectos de ExeML que se pueden crear?

ModelArts ExeML es compatible con proyectos de clasificación de imágenes, detección de objetos, análisis predictivo, clasificación de sonido y clasificación de texto. Se pueden crear hasta 100 proyectos de ExeML.

3.3.2 ¿Por qué no hay datos disponibles en la ruta de entrada del conjunto de datos cuando creo un proyecto?

Causa posible

1. El bucket de OBS y el proyecto creados no están en la misma región.
2. La autorización global no está configurada para la cuenta.
3. El formato de los datos en el bucket de OBS no cumple con los requisitos de servicio.

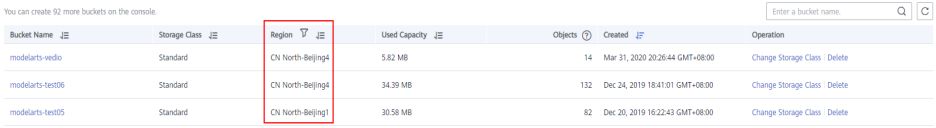
Solución

Compruebe si el proyecto de ModelArts y el bucket de OBS creado están en la misma región.

1. Compruebe la región donde se encuentra el bucket de OBS creado.
   a. Inicie sesión en OBS Console.
   b. En la página **Object Storage Service**, para buscar un bucket, introduzca una palabra clave en **Bucket Name**.

      En la columna **Region**, vea la región donde se encuentra el bucket de OBS creado.

      Figura 3-3 Región donde se encuentra un bucket de OBS

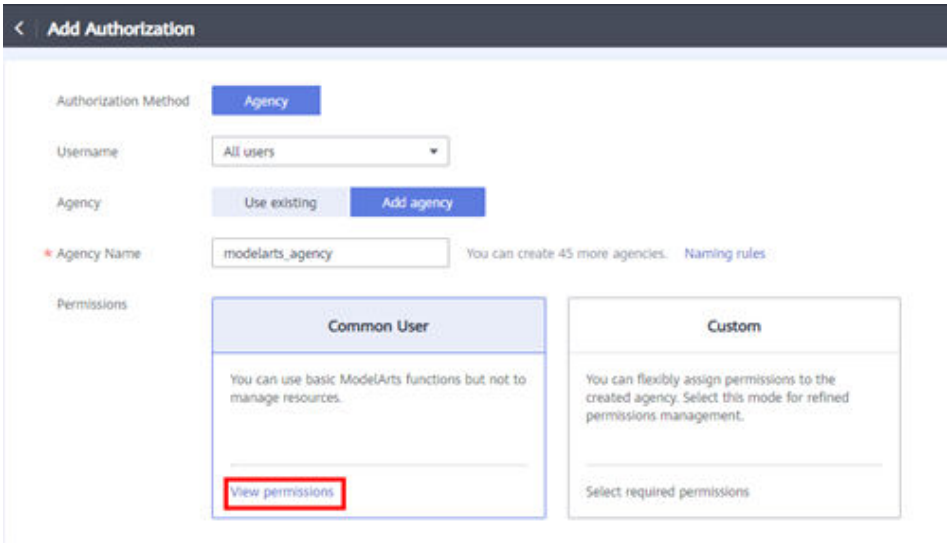
      

2. Compruebe la región donde se despliega ModelArts.

   Inicie sesión en la consola de gestión de ModelArts y vea la región donde se encuentra ModelArts en la esquina superior izquierda.
3. Compruebe si la región del bucket de OBS creado es la misma que la de ModelArts. Asegúrese de que sean iguales.

Configuración de la autorización de acceso (configuración global)

1. Inicie sesión en la consola de gestión de ModelArts. En el panel de navegación de la izquierda, elija **Settings**. Se muestra la página **Global Configuration**.
2. Haga clic en **Add Authorization**. En la página **Add Authorization** que se muestra, configure los parámetros.



3. Seleccione **I have read and agree to the ModelArts Service Statement** y haga clic en **Create**.

3.4 Etiquetado de datos

3.4.1 ¿Puedo agregar varias etiquetas a una imagen para un proyecto de detección de objetos?

Sí. Puede agregar varias etiquetas a una imagen.

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 ¿Qué debo hacer cuando el botón Train no está disponible después de crear un proyecto de clasificación de imágenes y etiquetar las imágenes?

El botón **Train** está disponible cuando las imágenes de entrenamiento para un proyecto de clasificación de imágenes se clasifican en al menos dos categorías, y cada categoría contiene al menos cinco imágenes.

3.5.2 ¿Cómo realizo entrenamiento incremental en un proyecto ExeML?

Cada ronda de entrenamiento genera una versión de entrenamiento en un proyecto ExeML. Si el resultado de un entrenamiento no es satisfactorio (por ejemplo, si la precisión no es lo suficientemente buena), puede agregar datos de alta calidad o agregar o eliminar etiquetas y volver a realizar el entrenamiento.

NOTA

- Actualmente, el entrenamiento incremental solo es compatible con los siguientes tipos de proyectos ExeML: clasificación de imágenes, detección de objetos y clasificación de sonido.
- Para obtener mejores resultados de entrenamiento, utilice datos de alta calidad para el entrenamiento incremental con el fin de mejorar el rendimiento del etiquetado de datos.

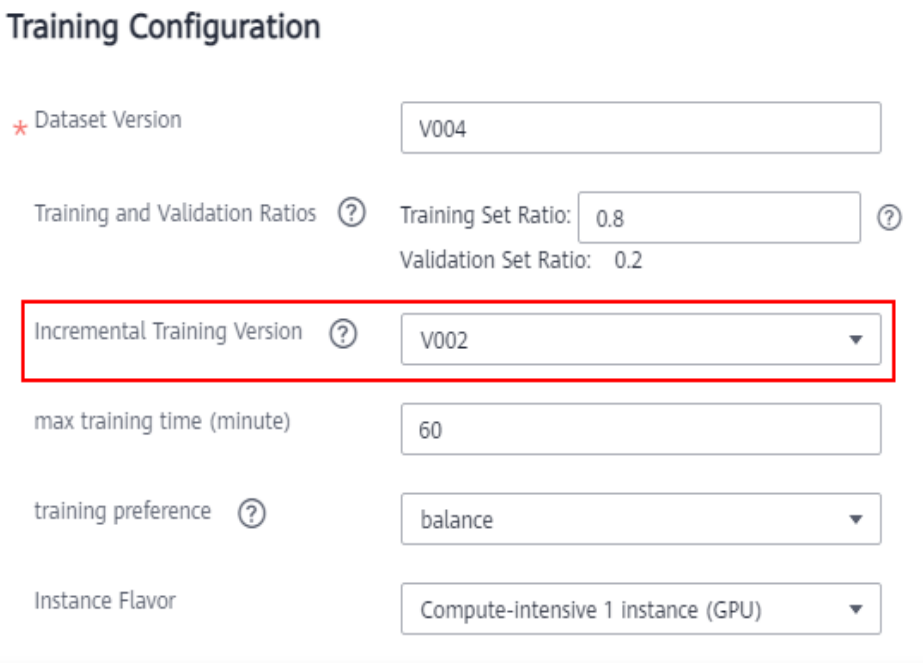
Procedimiento de entrenamiento Incremental

1. Inicie sesión en la consola de ModelArts y haga clic en **ExeML** en el panel de navegación izquierdo.
2. En la página **ExeML**, haga clic en un nombre de proyecto. Se muestra la página de detalles de ExeML del proyecto.
3. En la página **Label Data**, haga clic en la ficha **Unlabeled**. En la página de fichas **Unlabeled**, puede agregar imágenes, o agregar o eliminar etiquetas.

   Si agrega imágenes, vuelva a etiquetar las imágenes agregadas. Si agrega o elimina etiquetas, compruebe todas las imágenes y las etiquete de nuevo. También debe comprobar si es necesario agregar nuevas etiquetas para los datos etiquetados.
4. Una vez etiquetadas todas las imágenes, haga clic en **Train** en la esquina superior derecha. En el cuadro de diálogo **Training Configuration** que se muestra, establezca **Incremental Training Version** en la versión de entrenamiento que se ha completado para realizar entrenamiento incremental basado en esta versión. Establezca otros parámetros como se le solicite.

   Una vez completada la configuración, haga clic en **Yes** para iniciar el entrenamiento incremental. El sistema cambia automáticamente a la página **Train Model**. Una vez completado el entrenamiento, puede ver los detalles del entrenamiento, como la precisión del entrenamiento, el resultado de la evaluación y los parámetros de entrenamiento.

Figura 3-4 Selección de una versión de entrenamiento incremental



3.5.3 ¿Puedo descargar un modelo entrenado usando ExeML?

No. El modelo no se puede descargar. Puede ver el modelo o desplegar el modelo como un servicio en tiempo real en la página **AI Application Management**.

3.5.4 ¿Por qué falla el entrenamiento de ExeML?

Si el entrenamiento de un proyecto ExeML falla, realice los siguientes pasos para rectificar la falla:

1. Acceda a **Billing Center** y compruebe si la cuenta está en mora.
   a. Si la cuenta está en mora, **recargue la cuenta**.
   b. Si la cuenta no está en mora, vaya a **2**.
2. Compruebe si la ruta de OBS para almacenar datos de imagen cumple con los siguientes requisitos:
   – La ruta de acceso de OBS no contiene otras carpetas.
   – El nombre del archivo no contiene los siguientes caracteres especiales: ~`@#$ %^&*{}[]:;+=<>/

   Si la ruta de OBS cumple con los requisitos, vaya a **3**.
3. La causa de la falla puede variar dependiendo del proyecto de ExeML.
   – Si el entrenamiento de reconocimiento de imágenes falla, compruebe si hay imágenes dañadas. Si hay imágenes dañadas, reemplácelas o elimínelas.
   – Si el entrenamiento de detección de objetos falla, compruebe si el modo de etiquetado del conjunto de datos es correcto. Actualmente, ExeML solo admite las etiquetas basadas en rectángulos.
   – Si el entrenamiento de análisis predictivo falla, revise la columna de etiqueta. Actualmente, la columna de etiquetas soporta datos discretos y continuos. Solo se puede seleccionar una columna.

– Si el entrenamiento de clasificación de sonido falla, compruebe si los archivos de audio son archivos WAV de 16 bits.

Si el error persiste, envíe un **ticket de servicio** para obtener soporte técnico.

3.5.5 ¿Qué hago si se produjo un error de imagen durante el entrenamiento del modelo con ExeML?

Si se utiliza un algoritmo de clasificación de imágenes o de detección de objetos de ExeML, después de entrenar los datos etiquetados, el resultado de entrenamiento es un error de imagen. **Tabla 3-1** enumera soluciones a diferentes errores.

Tabla 3-1 Errores de imagen en la clasificación de imágenes y detección de objetos de ExeML

N.º	Parámetro de error	Descripción del error	Parámetro de la solución	Descripción de la solución
1	load failed	La imagen no se puede decodificar o restaurar.	ignore	El sistema ha ignorado esta imagen. No se necesita ninguna operación manual.
2	tf-decode failed	La imagen no puede ser decodificada por TensorFlow ni restaurada.	ignore	El sistema ha ignorado esta imagen. No se necesita ninguna operación manual.
3	size over	El tamaño de la imagen superó los 5 MB.	resize to small	El sistema ha comprimido el tamaño de la imagen a menos de 5 MB. No se necesita ninguna operación manual.
4	mode illegal	La imagen no está en formato RGB.	convert to rgb	El sistema ha convertido la imagen al formato RGB. No se necesita ninguna operación manual.
5	type illegal	El archivo no es una imagen, pero se puede convertir a JPG.	convert to jpg	El sistema ha convertido la imagen al formato JPG. No se necesita ninguna operación manual.

3.5.6 ¿Qué hago si se produjo el error de ModelArts.0010 cuando uso ExeML para iniciar el entrenamiento como usuario de IAM?

Utilice el permiso de ACL asignado por la cuenta de tenant para el bucket de OBS utilizado por ModelArts.

3.5.7 ¿Cuál es la velocidad de entrenamiento de cada parámetro en la configuración de preferencias de entrenamiento de ExeML?

Los ajustes de preferencia son los siguientes:

performance_first: la primera actuación. La duración del entrenamiento es corta y el modelo generado es pequeño. La velocidad de entrenamiento es de 10 ms para TXT o entrenamiento de imagen.

balance: rendimiento equilibrado y precisión. La velocidad de entrenamiento es de 14 ms para TXT o entrenamiento de imagen.

accuracy_first: precisión al primero. La duración del entrenamiento es larga y el modelo generado es grande. La velocidad de entrenamiento es de 16 ms para TXT o entrenamiento de imagen.

3.5.8 ¿Qué hago si "ERROR:input key sound is not in model" ocurre cuando uso ExeML para la predicción de clasificación de sonido?

De acuerdo con el log de errores del servicio en tiempo real, el archivo de audio utilizado para la predicción está vacío. Utilice un archivo de audio grande para la predicción.

3.6 Despliegue de modelos

3.6.1 ¿Qué tipo de servicio se despliega en ExeML?

Los modelos creados en ExeML se despliegan como servicios en tiempo real. Puede agregar imágenes o compilar código para probar los servicios, así como invocar a las API usando las URL.

Después de que el desarrollo del modelo se realice correctamente, puede elegir **Service Deployment** > **Real-Time Services** en el panel de navegación izquierdo de la consola de ModelArts para ver los servicios en ejecución y detener o eliminar servicios.

4 Notebook (Nueva Versión)

4.1 Restricciones

4.1.1 ¿Es compatible el motor Keras?

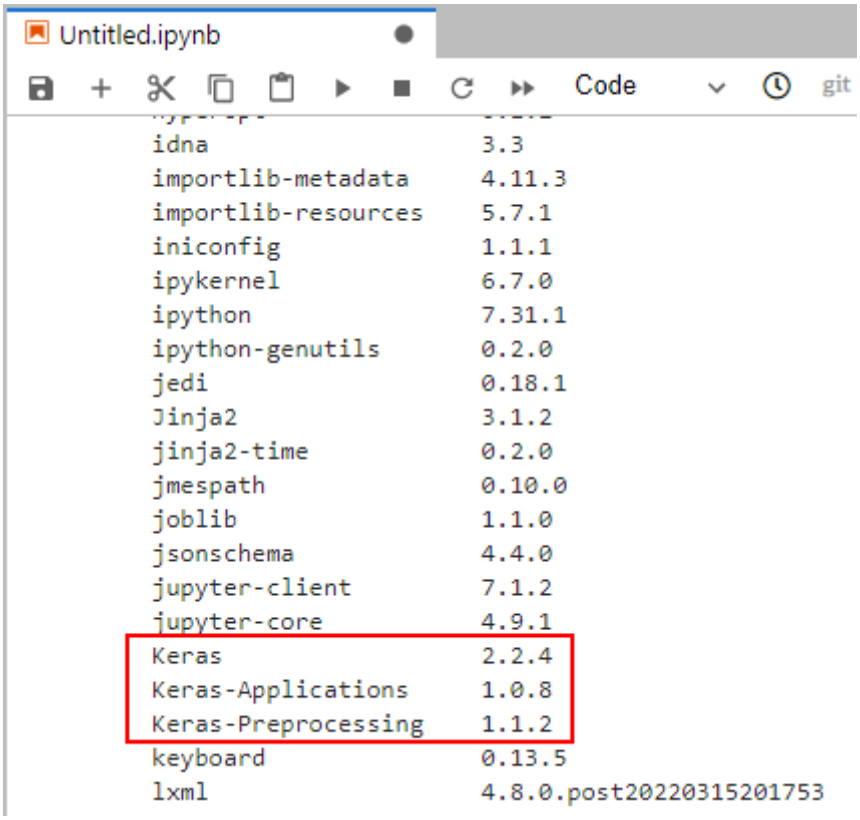
Las instancias de notebook en **DevEnviron** son compatibles con el motor Keras. El motor Keras no es compatible con el entrenamiento laboral y el despliegue de modelos (inferencia).

Keras es una API de red neuronal avanzada escrita en Python. Es capaz de funcionar encima de TensorFlow o CNTK o Theano. Las instancias de notebook de **DevEnviron** son compatibles con **tf.keras**.

¿Cómo consulto las versiones de Keras?

1. En la consola de gestión de ModelArts, cree una instancia de notebook con la imagen **TensorFlow-1.13** o **TensorFlow-1.15**.
2. Acceda a la instancia del notebook. En el JupyterLab, ejecute **!pip list** para ver las versiones de Keras.

Figura 4-1 Consulta de versiones de Keras



4.1.2 ¿ModelArts es compatible con el motor de Caffe?

El entorno de Python 2 de ModelArts soporta Caffe, pero el entorno de Python 3 no lo soporta.

4.1.3 ¿Puedo instalar MoXing en un entorno local?

No. MoXing solo se puede utilizar en ModelArts.

4.1.4 ¿Se pueden iniciar sesión de forma remota en las instancias de notebook?

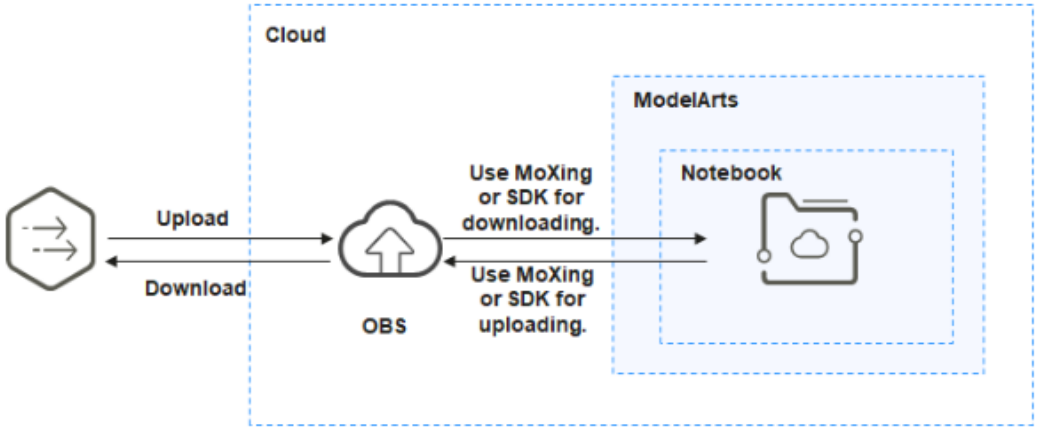
Las instancias de notebook de la nueva versión se pueden iniciar sesión de forma remota. Para ello, habilite el SSH remoto cuando cree las instancias del notebook. Inicie sesión de forma remota en una instancia de notebook desde un PyCharm profesional o VS Code.

4.2 Carga o descarga de datos

4.2.1 ¿Cómo cargo un archivo desde una instancia de Notebook a OBS o descargo un archivo desde OBS a una instancia de Notebook?

En una instancia de notebook, puede invocar a la API de MoXing de ModelArts o al SDK para intercambiar datos con OBS para cargar un archivo en OBS o descargar un archivo desde OBS a la instancia de notebook.

Figura 4-2 Cargar o descargar un archivo



Para obtener más información sobre cómo cargar archivos mediante OBS Browser, consulte **Carga y descarga de archivos mediante OBS Browser +**.

Método 1: Usar MoXing para cargar y descargar un archivo

Desarrollado por el equipo de ModelArts, MoXing es un marco de aceleración de entrenamiento distribuido basado en motores de aprendizaje profundo de código abierto como TensorFlow y PyTorch. MoXing hace que la codificación de modelos sea más fácil y eficiente.

MoXing proporciona un conjunto de API de objetos de archivo para leer y escribir archivos de OBS.

Para obtener detalles sobre la asignación entre las API de MoXing y las API nativas y cómo invocar a las API, consulta **Operaciones de archivos de MoXing**.

Código de ejemplo:

```
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')
```


Método 2: Uso del SDK para cargar y descargar un archivo

Invoque al SDK de ModelArts para descargar un archivo de OBS.

Ejemplo de código: Descargue **file1.txt** de OBS a **/home/ma-user/work/** en la instancia de notebook. Todo el nombre del bucket, el nombre de la carpeta y el nombre del archivo son personalizables.

```
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/")
```

Invoque al SDK de ModelArts para descargar una carpeta desde OBS.

Código de ejemplo: Descargue **dir1** de OBS a **/home/ma-user/work/** en la instancia de notebook. El nombre del bucket y el nombre de la carpeta son personalizables.

```
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/")
```

Invoque al SDK de ModelArts para cargar un archivo en OBS.

Ejemplo de código: Cargue **file1.txt** en la instancia del notebook a **obs://bucket-name/dir1/** del bucket de OBS. Todo el nombre del bucket, el nombre de la carpeta y el nombre del archivo son personalizables.

```
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/')
```

Invoque al SDK de ModelArts para cargar una carpeta en OBS.

Ejemplo de código: Suba **/work/** en la instancia del notebook a **obs://bucket-name/dir1/work/** del **bucket-name**. El nombre del bucket y el nombre de la carpeta son personalizables.

```
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/')
```

Manejo de errores

Si descarga un archivo de OBS a la instancia de su notebook y el sistema muestra el mensaje de error "Permission denied", realice las siguientes operaciones para solucionar problemas:

- Asegúrese de que el bucket de OBS y la instancia del notebook de destino estén en la misma región, por ejemplo, en **CN North-Beijing4**. Si el bucket de OBS y la instancia de notebook están en diferentes regiones, se deniega el acceso a OBS. Para obtener más información, consulte **¿Cómo puedo comprobar si ModelArts y un bucket de OBS están en la misma región?**
- Asegúrese de que la cuenta del notebook tenga permiso para leer datos en el bucket de OBS. Si la cuenta no tiene el permiso, consulte **¿Cómo accedo al bucket de OBS de otra cuenta desde una instancia de notebook?**

4.2.2 ¿Cómo cargo archivos locales a una instancia de Notebook?

Para obtener más información sobre cómo subir archivos a JupyterLab en el notebook de la nueva versión, consulte **Carga de archivos a JupyterLab**.

4.2.3 ¿Cómo puedo importar archivos grandes a una instancia de notebook?

- **Archivos grandes (archivos más de 100 MB)**

  Utilice OBS para subir archivos de gran tamaño. Para ello, utilice OBS Browser para cargar un archivo local a un bucket de OBS y utilice el SDK de ModelArts para descargar el archivo desde OBS a una instancia de notebook.

  Para cargar archivos mediante OBS Browser, consulte **Carga y descarga de archivos con OBS Browser+**.

  Para obtener más información sobre cómo usar el SDK o MoXing de ModelArts para descargar archivos de OBS, consulte **¿Cómo cargo un archivo desde una instancia de Notebook a OBS o descargo un archivo desde OBS a una instancia de Notebook?**

- **Carpeta**

  Comprimir una carpeta en un paquete y cargar el paquete de la misma manera que cargar un archivo grande. Después de cargar el paquete en una instancia de notebook, descomprima en la página **Terminal**.

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

  Para obtener más detalles, busque el comando de descompresión en los motores de búsqueda principales.

4.2.4 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.

4.2.5 ¿Cómo descargo archivos de una instancia de Notebook a un equipo local?

Para obtener más información sobre cómo descargar archivos de JupyterLab en el notebook de la nueva versión, consulte **Descargar un archivo de JupyterLab a una ruta local**.

4.2.6 ¿Cómo puedo copiar datos del entorno de desarrollo del notebook A al notebook B?

Los datos no se pueden copiar directamente del notebook A al notebook B. Para copiar datos, haga lo siguiente:

1. Suba los datos del notebook A a OBS.
2. Descargue datos de OBS al notebook B.

Para obtener más información sobre cómo cargar y descargar archivos, consulte **¿Cómo cargo un archivo desde una instancia de Notebook a OBS o descargo un archivo desde OBS a una instancia de Notebook?**

4.3 Almacenamiento de datos

4.3.1 ¿Cómo cambio el nombre de un archivo de OBS?

No se puede cambiar el nombre de los archivos de OBS en la consola de OBS. Para cambiar el nombre de un archivo de OBS, invoque a una API de MoXing en una instancia de notebook existente o recién creada.

A continuación se muestra un ejemplo:

Cambie el nombre de **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')
```

4.3.2 ¿Todavía existen archivos en /cache después de que se detenga o reinicie una instancia de notebook? ¿Cómo puedo evitar un reinicio?

Los archivos temporales se almacenan en el directorio **/cache** y no se guardarán después de que se detenga o reinicie la instancia del notebook. Los datos almacenados en el directorio **/home/ma-user/work** se conservarán después de que se detenga o reinicie la instancia del notebook.

Para evitar un reinicio, no entrene trabajos de carga pesada que consumen grandes cantidades de recursos de CPU, GPU o memoria en el entorno de desarrollo.

4.3.3 ¿Cómo uso la biblioteca de pandas para procesar datos en los bucket de OBS?

Paso 1 Descargue datos de OBS a una instancia de notebook. Para obtener más información, consulte **Descargar un archivo de JupyterLab a una ruta local**.

Paso 2 Procesar los datos de los pandas siguiendo las instrucciones proporcionadas en la ***Guía de usuario de pandas***.

----Fin

4.3.4 ¿Cómo accedo al bucket de OBS de otra cuenta desde una instancia de Notebook?

Para acceder a los archivos de OBS de otra cuenta desde una instancia de notebook, debe tener permisos de lectura y escritura en el bucket de OBS de destino.

Póngase en contacto con el creador del bucket de OBS para conceder permisos de lectura y escritura en el bucket de OBS de la cuenta actual de Huawei Cloud, consultando **Otorgar permisos de lectura y escritura en un bucket a otras cuentas de Huawei Cloud**. Si su cuenta es una cuenta de IAM o en otros escenarios, consulte la sección "Casos de

configuración en escenarios típicos de control de permisos" de la ***Guía de configuración de permisos de Object Storage Service*** para obtener instrucciones sobre cómo otorgar permisos de bucket de OBS.

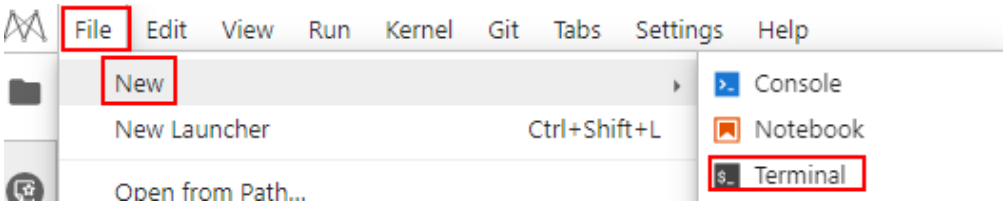
Después de obtener permisos de lectura y escritura en el bucket de OBS, puede usar la API MoXing en su instancia de notebook para acceder al bucket de OBS y leer datos.

4.4 Configuraciones de entorno

4.4.1 ¿Cómo puedo activar la función de terminal en DevEnviron de ModelArts?

1. Inicie sesión en la consola de gestión de ModelArts y seleccione **DevEnviron > Notebooks**.
2. Cree una instancia de notebook. Cuando la instancia se esté ejecutando, haga clic en **Open** en la columna **Operation**. Se muestra la página **JupyterLab**.
3. Elija **File** > **New** > **Terminal**. Se muestra la página **Terminal**.

Figura 4-3 Ir a la página **Terminal**



4.4.2 ¿Cómo instalo las bibliotecas externas en una instancia de notebook?

Múltiples entornos como Jupyter y Python se han integrado en el notebook de ModelArts para soportar muchos marcos, incluidos TensorFlow, MindSpore, PyTorch y Spark. Puede utilizar **pip install** para instalar las bibliotecas externas en Jupyter Notebook o en la página **Terminal**.

Instalación de bibliotecas externas en Jupyter Notebook

Puede utilizar JupyterLab para instalar Shapely en el entorno **TensorFlow-1.8**.

1. Abra una instancia de notebook y acceda a la página **Launcher**.
2. En el área **Notebook**, haga clic en **TensorFlow-1.8** y cree un archivo IPYNB.
3. En la nueva instancia del cuaderno, introduzca el siguiente comando en la barra de entrada de código:

   !pip install Shapely

Instalación de las bibliotecas externas en la página Terminal

Puede usar **pip** para instalar las bibliotecas externas en el entorno **TensorFlow-1.8** en la página **Terminal**. Por ejemplo, para instalar Shapely:

1. Abra una instancia de notebook y acceda a la página **Launcher**.

2. En el área **Other**, haga clic en **Terminal** y cree un archivo de terminal.
3. Ingrese los siguientes comandos en el cuadro de entrada de código para obtener el núcleo del entorno actual y activar el entorno de Python del que depende la instalación:

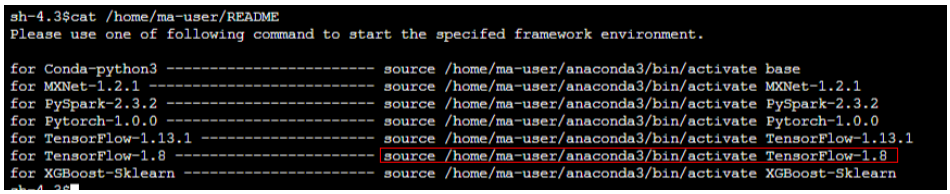
   cat /home/ma-user/README

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

   NOTA

   Para instalar TensorFlow en otro entorno de Python, reemplace **TensorFlow-1.8** en el comando con el motor de destino.

   Figura 4-4 Activar el entorno

4. Ejecute el siguiente comando en el cuadro de entrada de código para instalar Shapely:

   pip install Shapely

4.4.3 ¿Cómo puedo resolver la visualización de fuentes anormales en un notebook de ModelArts al que se accede desde iOS?

Síntoma

Cuando se accede a un notebook de ModelArts desde iOS, la fuente se muestra de forma anormal.

Solución

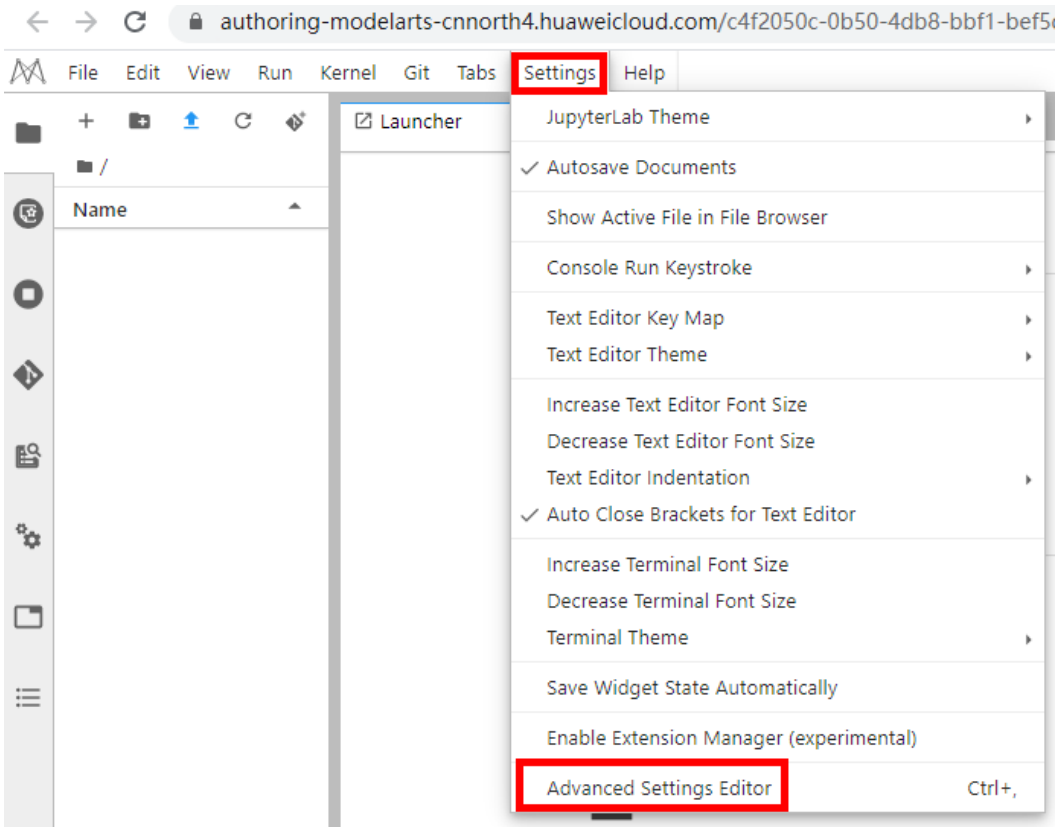
Establezca **fontFamily** de **Terminal** a **Menlo**.

Procedimiento

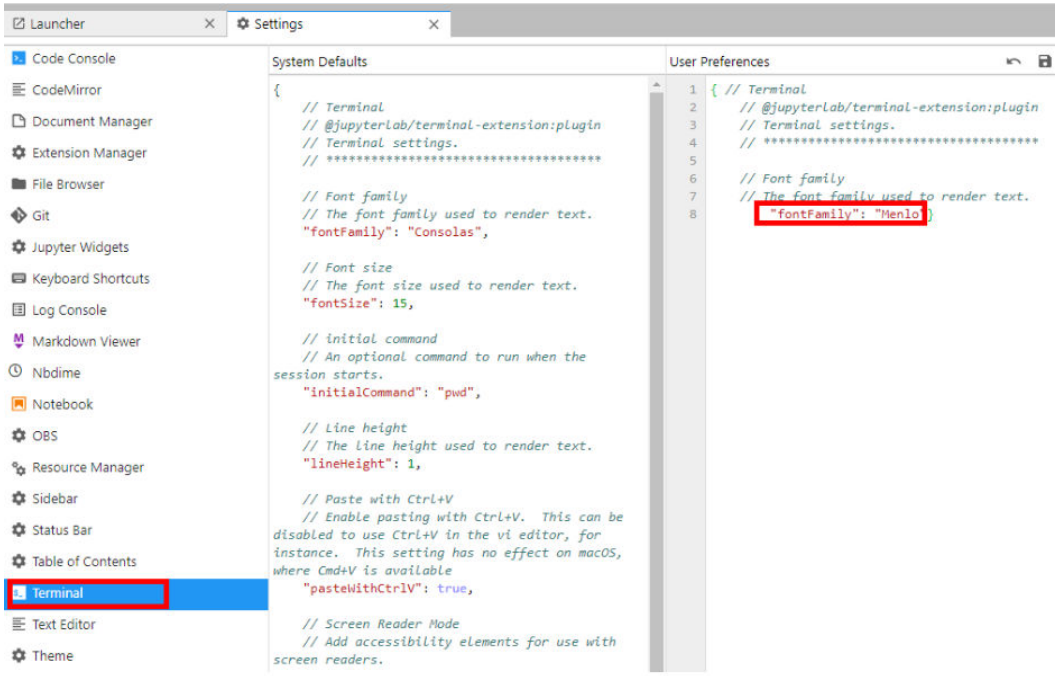
Paso 1 Inicie sesión en la consola de gestión de ModelArts y seleccione **DevEnviron > Notebook**.

Paso 2 Busque la fila que contiene la instancia del notebook de destino y haga clic en **Open** en la columna **Operation**. Se muestra la página **JupyterLab**.

Paso 3 En la página **JupyterLab**, seleccione **Settings > Advanced Settings Editor**. Se muestra la página de ficha **Settings**.



Paso 4 Elija **Terminal** en el panel de navegación de la izquierda y establezca **fontFamily** en **Menlo**.



----Fin

4.5 Instancias de notebook

4.5.1 ¿Qué hago si no puedo acceder a mi instancia de notebook?

Solucionar el problema basándose en el código de error.

Error 404

Si se informa de este error cuando un usuario de IAM crea una instancia, el usuario de IAM no tiene los permisos para acceder a la ubicación de almacenamiento correspondiente (bucket de OBS).

Solución

1. Inicie sesión en la consola de OBS con la cuenta principal y conceda permisos de acceso para el bucket de OBS al usuario de IAM. Para obtener más información sobre la operación, consulte **Principal**.
2. Después de que el usuario de IAM obtenga los permisos, inicie sesión en la consola de ModelArts, elimine la instancia y use la ruta de acceso de OBS para crear una instancia de notebook.

Error 503

Si se informa de este error, es posible que la instancia esté consumiendo demasiados recursos. Si este es el caso, detenga la instancia y reiníciela.

Error 504

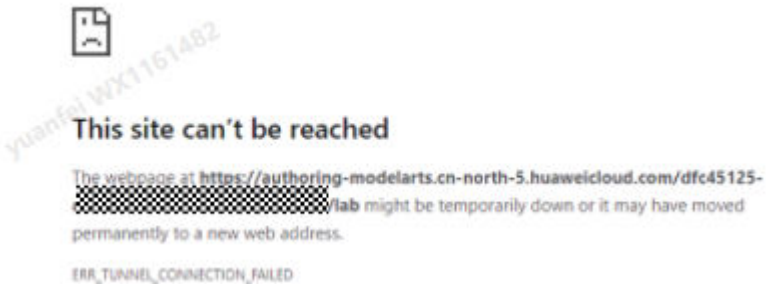
Si se informa de este error, **envíe un ticket de servicio** o póngase en contacto con el servicio de atención al cliente.

Error 500

No se puede abrir el JupyterLab del notebook y se notifica el error 500. La posible causa es que se agota el espacio en disco en el directorio **work**. En este caso, identifique la causa de la falla y borre el disco haciendo referencia a **Espacio en disco usado**.

Error "This site can't be reached"

Después de crear una instancia de notebook, haga clic en **Open** en la columna **Operation**. Se muestra el mensaje de error que se muestra en la siguiente figura.



Haga lo siguiente para resolver este problema: Copie el nombre de dominio de la página (**authoring-modelarts.cn-xxx-5.huaweicloud.com** en la figura anterior), agréguelo al cuadro de texto **Do not use proxy server for addresses beginning with** y guarde la configuración.

4.5.2 ¿Qué debo hacer cuando el sistema muestra un mensaje de error que indica que no queda espacio después de ejecutar el comando pip install?

Síntoma

En la instancia del notebook, aparece el mensaje de error "No Space left..." después de ejecutar el comando **pip install**.

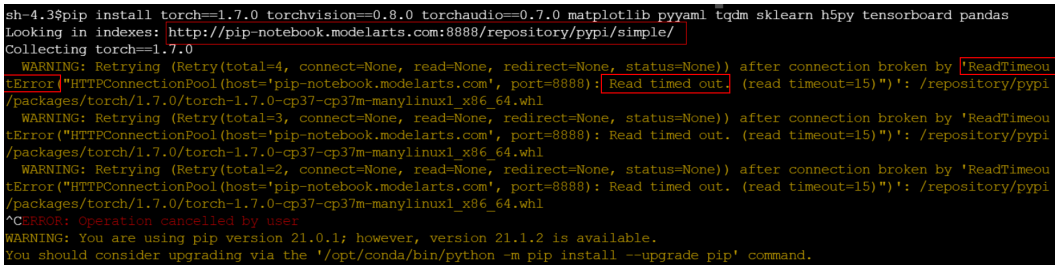
Solución

Se recomienda ejecutar el comando **pip install --no-cache **** en lugar del comando **pip install ****. La adición del parámetro **--no-cache** puede resolver este problema.

4.5.3 ¿Qué hago si se muestra "Read timed out" después de ejecutar pip install?

Síntoma

Después de ejecutar **pip instalar** en una instancia de notebook, el sistema muestra el mensaje de error "ReadTimeoutError..." o "Read timed out...".



Solución

Ejecute **pip install --upgrade pip** y luego **pip install**.

4.5.4 ¿Qué hago si el código se puede ejecutar pero no se puede guardar y se muestra el mensaje de error "save error"?

Si la instancia del notebook puede ejecutar el código pero no puede guardarlo, aparecerá el mensaje de error "save error" al guardar el archivo. En la mayoría de los casos, este error es causado por una política de seguridad del Web Application Firewall (WAF).

En la página actual, algunos caracteres de entrada o salida del código son interceptados porque se consideran un riesgo de seguridad. Envíe un ticket de servicio y póngase en contacto con el servicio de atención al cliente para verificar y manejar el problema.

4.5.5 ¿Por qué se notifica un error de tiempo de espera de solicitud cuando hago clic en el botón Open de una instancia de Notebook?

Cuando un contenedor de notebook se descompone debido a un desbordamiento de memoria u otras razones, si hace clic en el botón **Open** de la instancia de notebook, se muestra un error de tiempo de espera de solicitud.

En este caso, espere aproximadamente medio minuto hasta que se restaure el contenedor y, a continuación, haga clic en **Open** de nuevo.

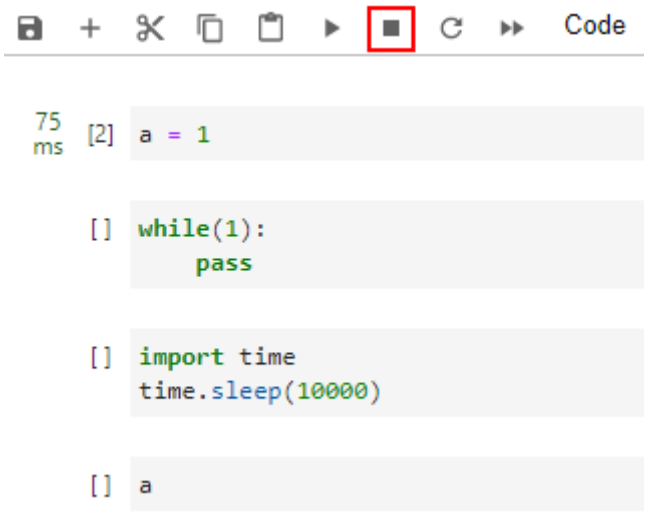
4.6 Code Execution

4.6.1 ¿Qué hago si una instancia de notebook no ejecuta mi código?

Si una instancia de notebook no puede ejecutar código, puede localizar y rectificar el error de la siguiente manera:

1. Si la ejecución de una celda se suspende o dura mucho tiempo (por ejemplo, la ejecución de la segunda y tercera celdas de **Figura 4-5** se suspende o dura mucho tiempo, provocando una falla de ejecución de la cuarta celda). pero la página del notebook todavía responde y se pueden seleccionar otras celdas, haga clic en **interrupt the kernel** resaltado en un cuadro rojo en la siguiente figura para detener la ejecución de todas las celdas. La instancia del notebook conserva todos los espacios variables.

   Figura 4-5 Detener todas las celdas

2. Si la página del notebook no responde, cierre la página del notebook y la consola de ModelArts. A continuación, abra la consola de ModelArts y vuelva a acceder a la instancia del notebook. La instancia de notebook conserva todos los espacios variables que existen cuando la instancia de notebook no está disponible.
3. Si aún no se puede utilizar la instancia del notebook, acceda a la página **Notebook** de la consola de ModelArts y detenga la instancia del notebook. Una vez que se detenga la instancia de notebook, haga clic en **Start** para reiniciar la instancia de notebook y abrirla.

La instancia habrá conservado todos los espacios para las variables que no pudieron ejecutarse.

4.6.2 ¿Por qué se descompone la instancia cuando se muestra el núcleo muerto durante la ejecución del código de entrenamiento?

La instancia del notebook se descompone durante la ejecución del código de entrenamiento debido a la memoria insuficiente causada por el gran volumen de datos o el exceso de capas de entrenamiento.

Después de que se produzca este error, el sistema reinicia automáticamente la instancia del notebook para solucionar la falla de la instancia. En este caso, solo se fija la avería. Si vuelve a ejecutar el código de entrenamiento, el error seguirá ocurriendo. Para resolver el problema de la falta de memoria, se recomienda crear una nueva instancia de notebook y utilizar un grupo de recursos de especificaciones más altas, como una GPU o un grupo de recursos dedicado, para ejecutar el código de entrenamiento. Una instancia de notebook existente que se haya creado correctamente no se puede ampliar con recursos con especificaciones más altas.

4.6.3 ¿Qué hago si cudaCheckError ocurre durante el entrenamiento?

Síntoma

El siguiente error se produce cuando el código de entrenamiento se ejecuta en un notebook:

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

Causa posible

Los parámetros **arch** y **code** de **setup.py** no se han establecido para que coincidan con la potencia de procesamiento de la GPU.

Solución

Para las GPU Tesla V100, la potencia de cómputo de la GPU es de **-gencode arch=compute_70,code=[sm_70,compute_70]**. Establezca los parámetros de compilación de **setup.py** en consecuencia.

4.6.4 ¿Qué debo hacer si DevEnviron genera espacio insuficiente?

Si el espacio es insuficiente, utilice instancias de notebook del tipo EVS.

Sube el código y los datos a un bucket de OBS para la instancia original del notebook haciendo referencia a **¿Cómo cargo un archivo desde una instancia de Notebook a OBS o descargo un archivo desde OBS a una instancia de Notebook?**. A continuación, cree una instancia de notebook del tipo EVS y descargue archivos de OBS a la nueva instancia de notebook.

4.6.5 ¿Por qué se descompone la instancia del notebook cuando se utiliza opencv.imshow?

Síntoma

Cuando se utiliza opencv.imshow en una instancia de notebook, la instancia de notebook se descompone.

Causas posibles

La función cv2.imshow en OpenCV funciona mal en un entorno cliente/servidor como Jupyter. Sin embargo, Matplotlib no tiene este problema.

Solución

Muestra las imágenes haciendo referencia al siguiente ejemplo. Tenga en cuenta que OpenCV muestra imágenes de BGR mientras que Matplotlib muestra imágenes de RGB.

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()
```

4.6.6 ¿Por qué no se puede encontrar la ruta de acceso de un archivo de texto generado en el sistema operativo Windows en una instancia de notebook?

Síntoma

Cuando se utiliza un archivo de texto generado en Windows en una instancia de notebook, no se puede leer el contenido de texto y se puede mostrar un mensaje de error que indica que no se puede encontrar la ruta de acceso.

Causas posibles

La instancia del notebook ejecuta Linux y su formato de alimentación de línea (CRLF) difiere de ese (LF) en Windows.

Solución

Convierta el formato de archivo a Linux en su instancia de notebook.

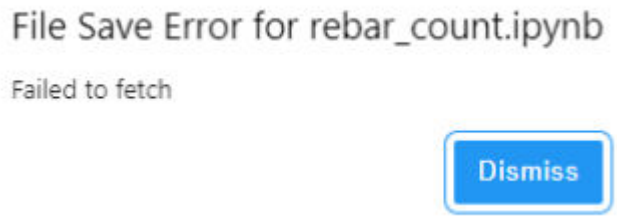
Shell:

```
dos2unix  File name
```


4.6.7 ¿Qué debo hacer si JupyterLab no se guarda ningún archivo?

Síntoma

Cuando se guarda un archivo en el JupyterLab se muestra un mensaje de error.



Causa posible

Se ha instalado un complemento de terceros en el navegador, y el proxy intercepta la solicitud. Como resultado, el archivo no se puede guardar.

Solución

Desactivar el complemento y guardar el archivo de nuevo.

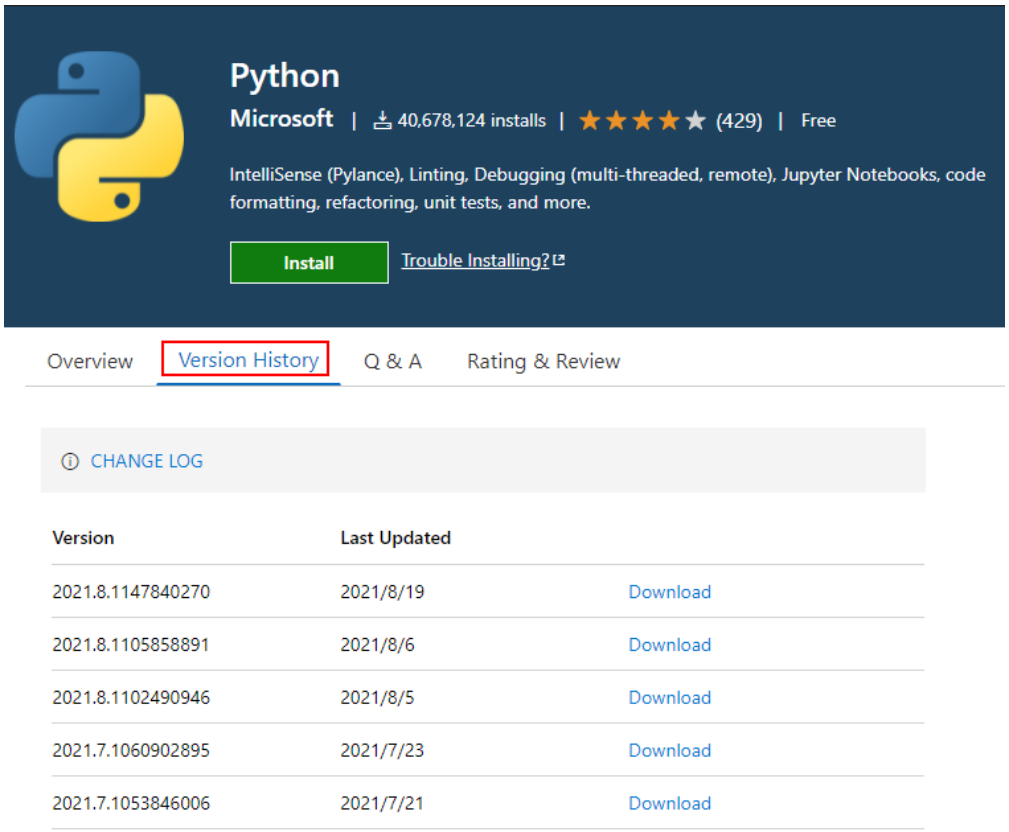
4.7 VS Code

4.7.1 ¿Qué hago si falló la instalación de un complemento remoto?

Método 1 (recomendado): Usar un paquete sin conexión

1. Inicie sesión en el **sitio web oficial de VS Code** y busque el complemento de Python de destino.
2. Haga clic en la pestaña **Version History** del complemento y descargue el paquete de instalación sin conexión.

Figura 4-6 Paquete de instalación fuera de línea del complemento de Python



3. En el VS Code local, arrastre el archivo VSIX descargado al notebook remoto.
4. Haga clic con el botón derecho en el archivo y elija **Install Extension VSIX** en el menú contextual.

Método 2: Configuración del complemento remoto por defecto

Ajuste el complemento remoto por defecto en VS Code siguiendo las instrucciones proporcionadas en **How Can I Set the Default Remote Plug-in in VS Code?**. Esto permite la instalación automática del complemento cuando el complemento está conectado.

Método 3: Tomar medidas proporcionadas en el **sitio web oficial del VS Code**

Sugerencias (ajuste la configuración de los parámetros según sea necesario):

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

4.7.2 ¿Qué hago si solo se puede conectar una instancia de notebook reiniciada después de eliminar localmente 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
```

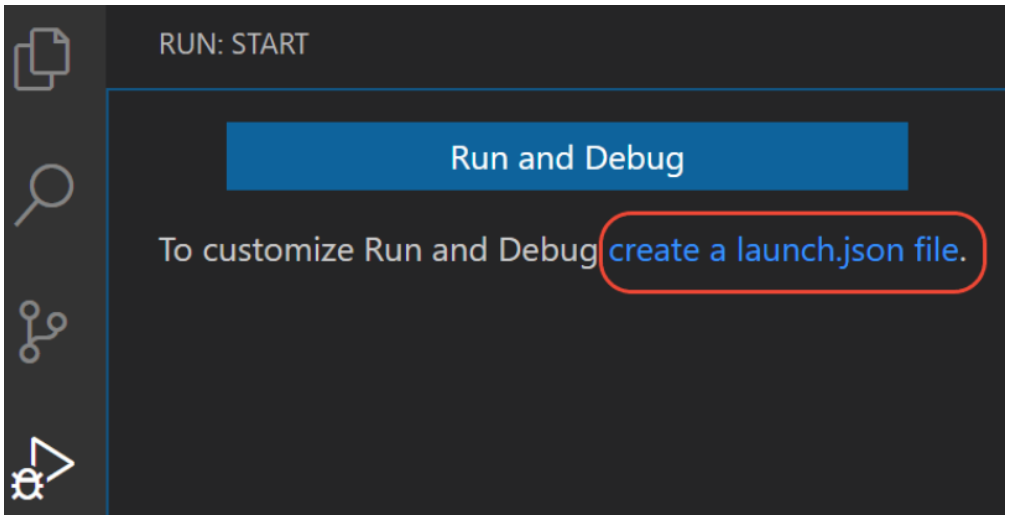
Note: SSH logins are insecure because the **known_hosts** file will be ignored during the logins.

4.7.3 ¿Qué hago si no se puede acceder al código fuente cuando uso VS Code para la depuración?

Si el archivo **launch.json** ya existe, vaya al paso 3.

Paso 1: Abrir launch.json.

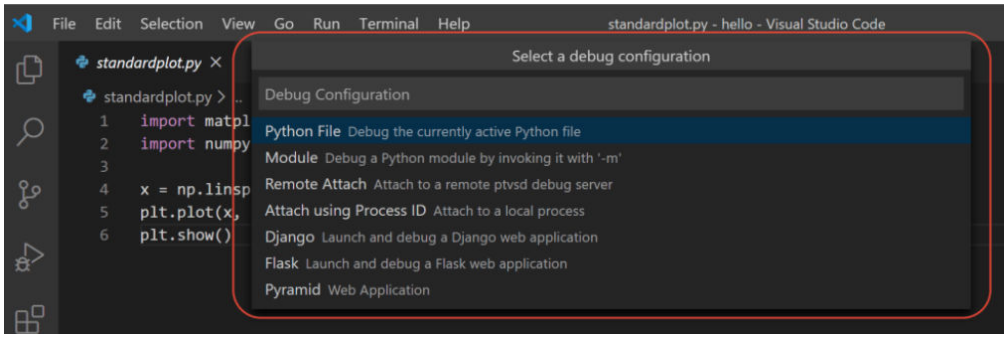
- Método 1: Haga clic en **Run** (**Ctrl+Shift+D**) en la barra de menú de la izquierda y haga clic en **create a launch.json file**.



- Método 2: En la barra de menús, elija **Run** > **Open configurations**.

Paso 2: Seleccionar una lengua.

Para establecer un lenguaje Python, seleccione **Python File** en **Select a debug configuration**. Las operaciones para configurar otros idiomas son similares.



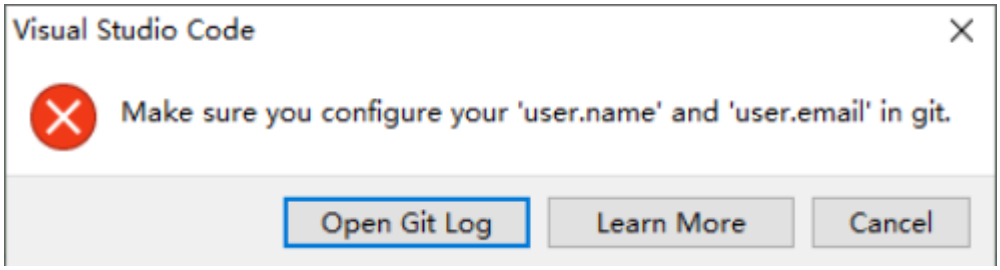
Paso 3: Establecer justMyCode en False en launch.json.

```
{
    // 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
                }
            ]
        }
```

4.7.4 ¿Qué hago si se muestra un mensaje que indica un nombre de usuario o una dirección de correo electrónico incorrectos cuando uso VS Code para enviar el código?



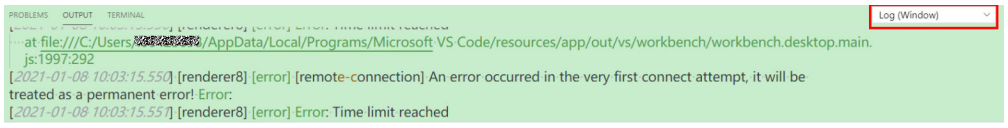
1. En VS Code, pulse **Ctrl+Shift+P**.
2. Encuentre **Python: Select Interpreter** y seleccione el Python de destino.
3. Elija **Terminal > New Terminal**. Se muestra CLI del contenedor remoto.
4. En el terminal de VS Code, ejecute los siguientes comandos y vuelva a enviar el código:

```
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.
```

4.7.5 ¿Cómo puedo ver los logs remotos en VS Code?

1. En VS Code, pulse **Ctrl+Shift+P**.
2. Encuentre **show logs**.
3. Elija **Remote Server**.

Alternativamente, cambie los logs en el cuadro rojo que se muestra en la siguiente figura.



4.7.6 ¿Cómo puedo abrir el archivo de configuración de VS Code settings.json?

1. En VS Code, pulse **Ctrl+Shift+P**.
2. Búsqueda de **Open Settings (JSON)**.

4.7.7 ¿Cómo cambio el color de fondo del VS Code al verde claro?

Agregue la siguiente configuración al archivo de configuración de VS Code **settings.json**:

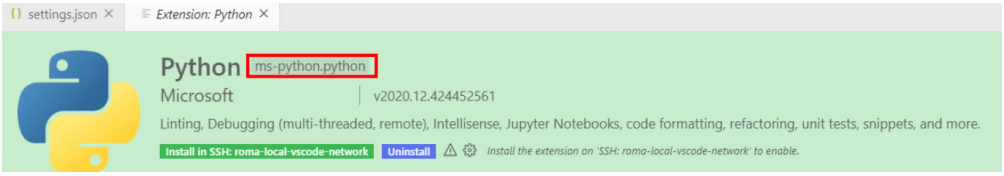
```
"workbench.colorTheme": "Atom One Light",
"workbench.colorCustomizations": {
"[Atom One Light]": {
 "editor.background": "#C7EDCC",
"sideBar.background": "#e7f0e7",
"activityBar.background": "#C7EDCC",
     },
},
```

4.7.8 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.



4.7.9 ¿Cómo puedo instalar un complemento local en el extremo remoto o un complemento remoto en el extremo local con VS Code?

1. En VS Code, pulse **Ctrl+Shift+P**.
2. Busque **install local** y seleccione el complemento según sea necesario.

4.8 Fallas en el acceso al entorno de desarrollo con VS Code

4.8.1 ¿Cuándo lo hago si no se muestra la ventana de VS Code?

Causa posible

VS Code no está instalado o la versión instalada está desactualizada.

Solución

Descargar e instalar VS Code.

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

AVISO

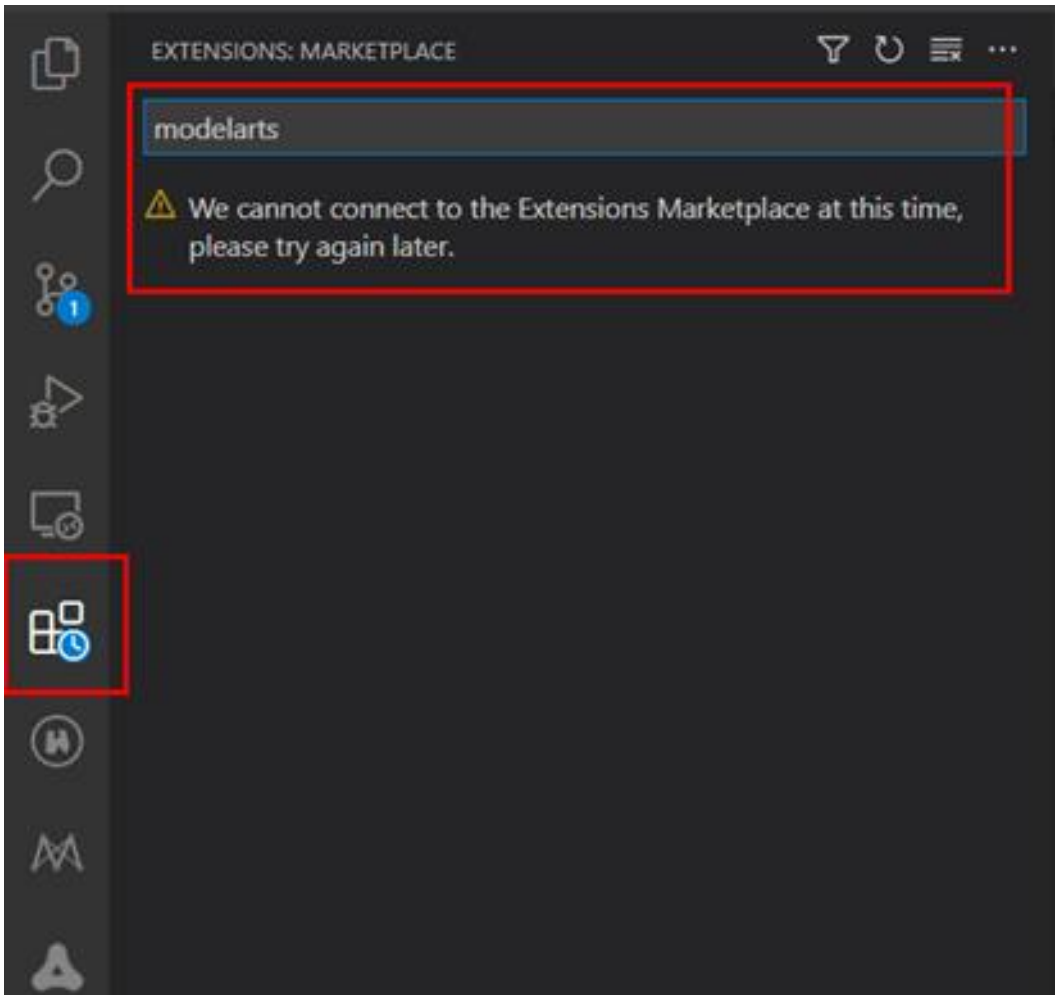
If your local PC runs Linux, see possible cause 2.

Possible Cause 1

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

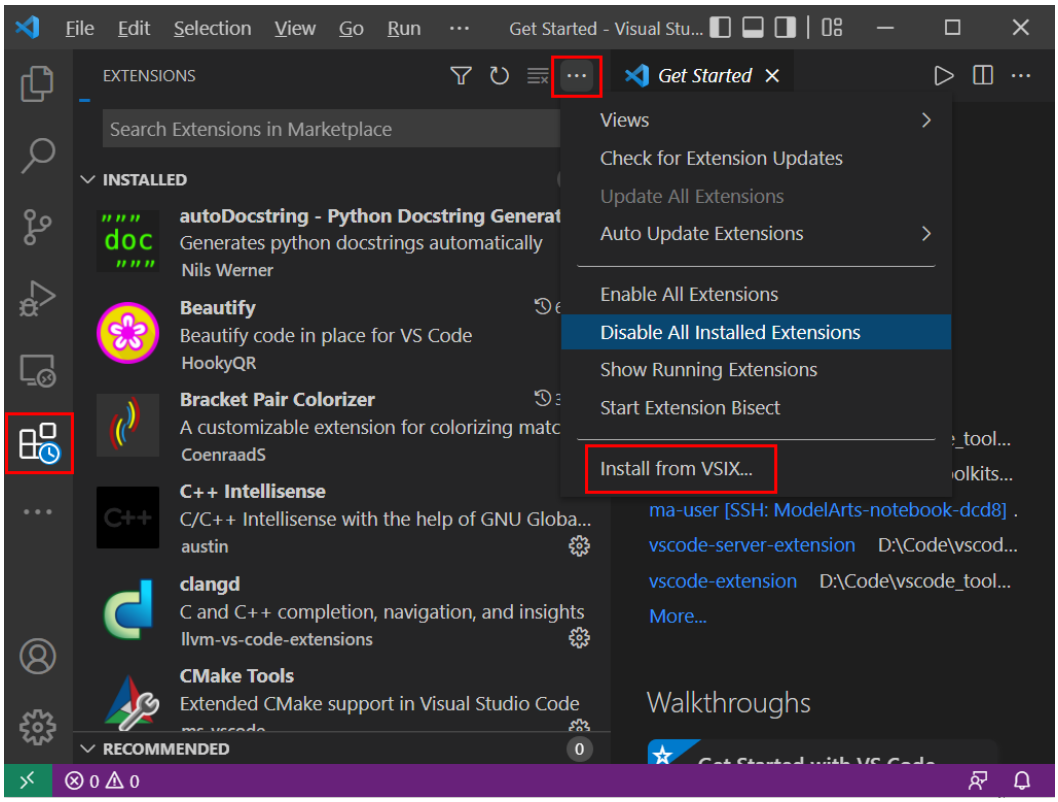
Solution

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.

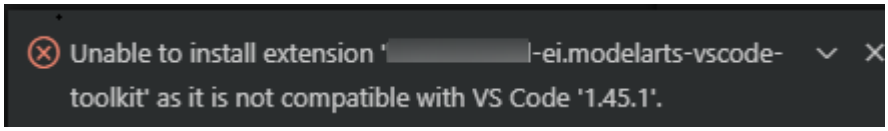


Search for **ModelArts-HuaweiCloud** again. If the following information is displayed, the network is normal. Then, switch back to the ModelArts management console and try to access VS Code again.

Method 2: If the VS Code marketplace cannot be accessed, manually download and install **ModelArts-HuaweiCloud** and **Remote-SSH**.

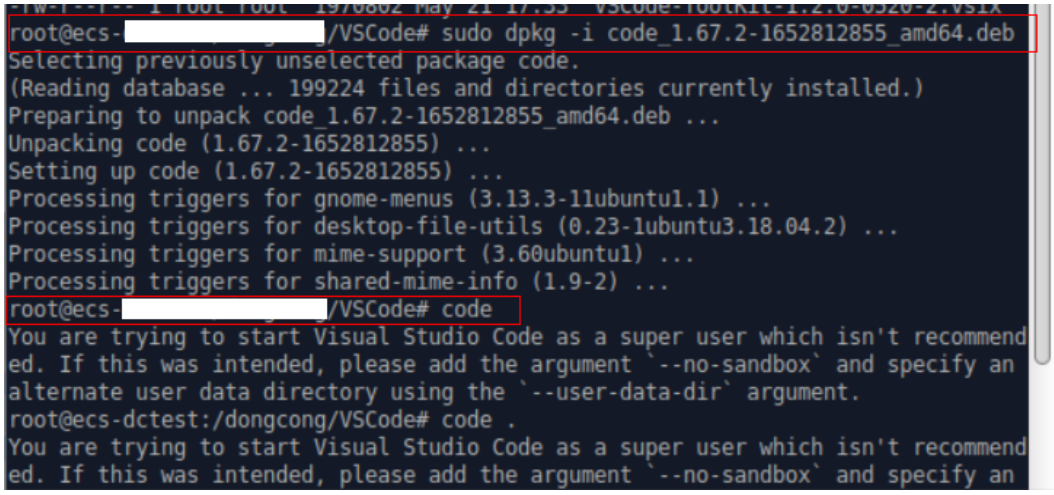


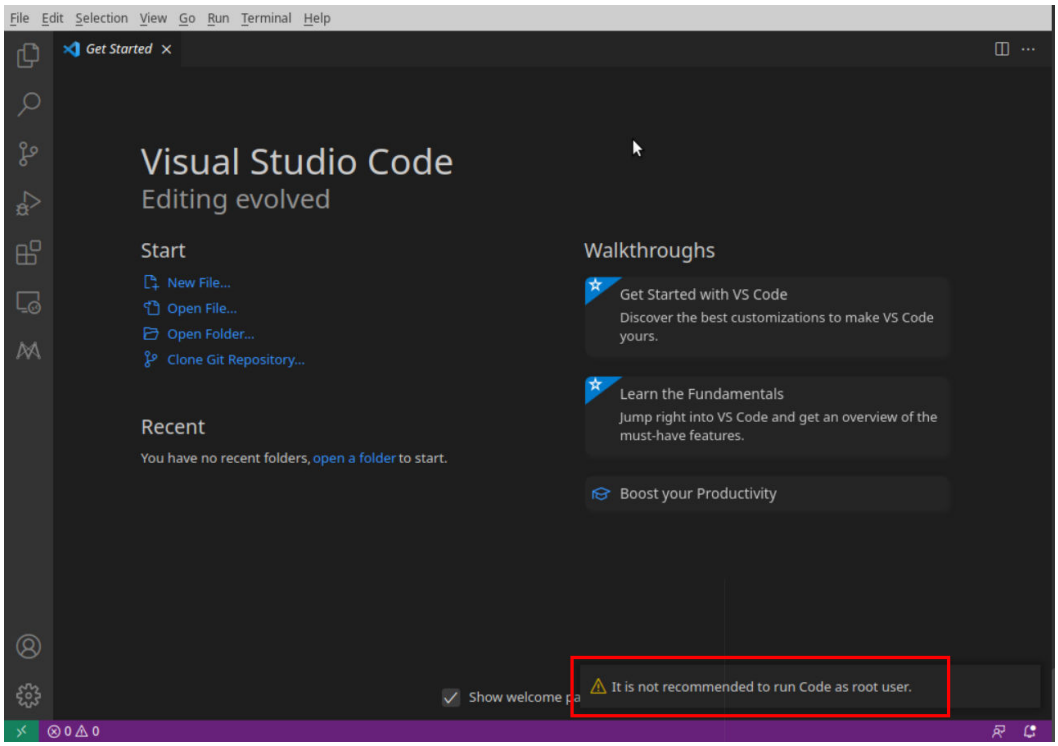
Method 3: If the error message shown in the following figure is displayed, the VS Code version is outdated. Upgrade the VS Code to 1.57.1 or the latest version.



Possible Cause 2

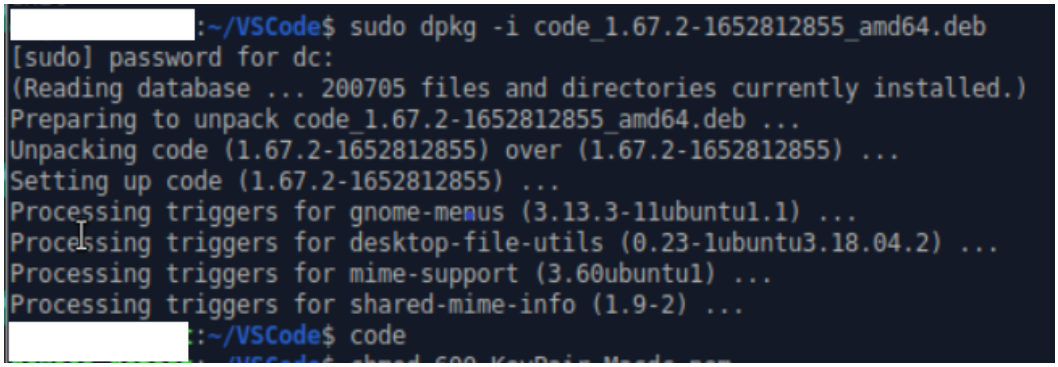
The local PC runs Linux, and VS Code is installed as user **root**. When you access VS Code, the information "It is not recommended to run Code as root user" is displayed.

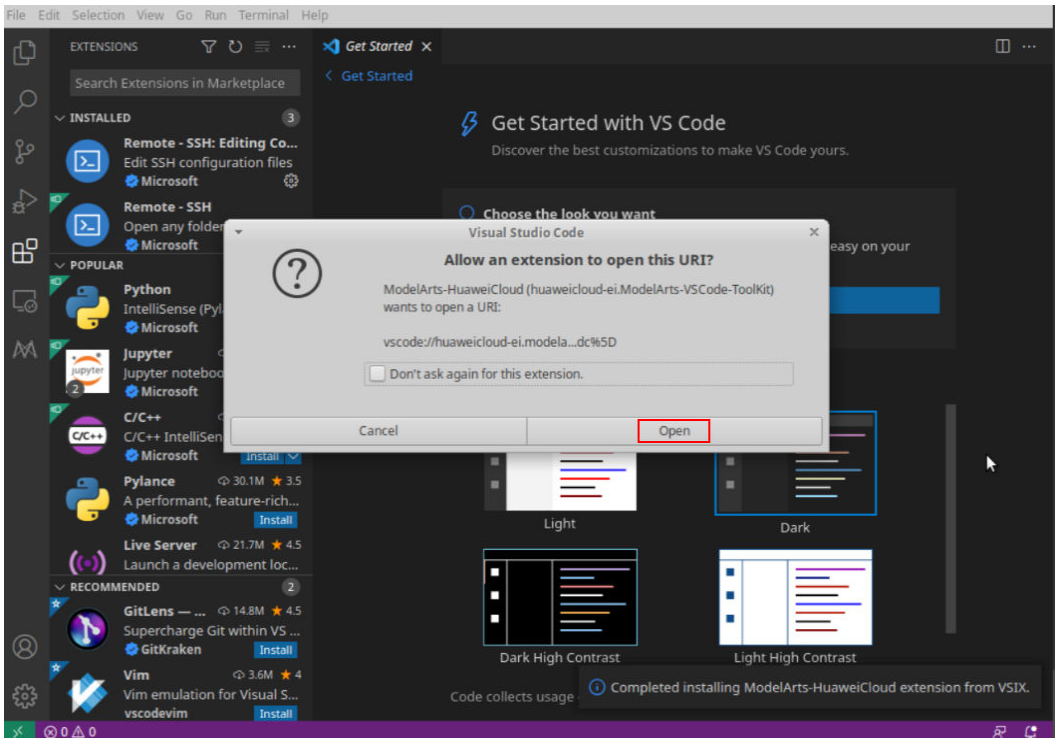




Solution

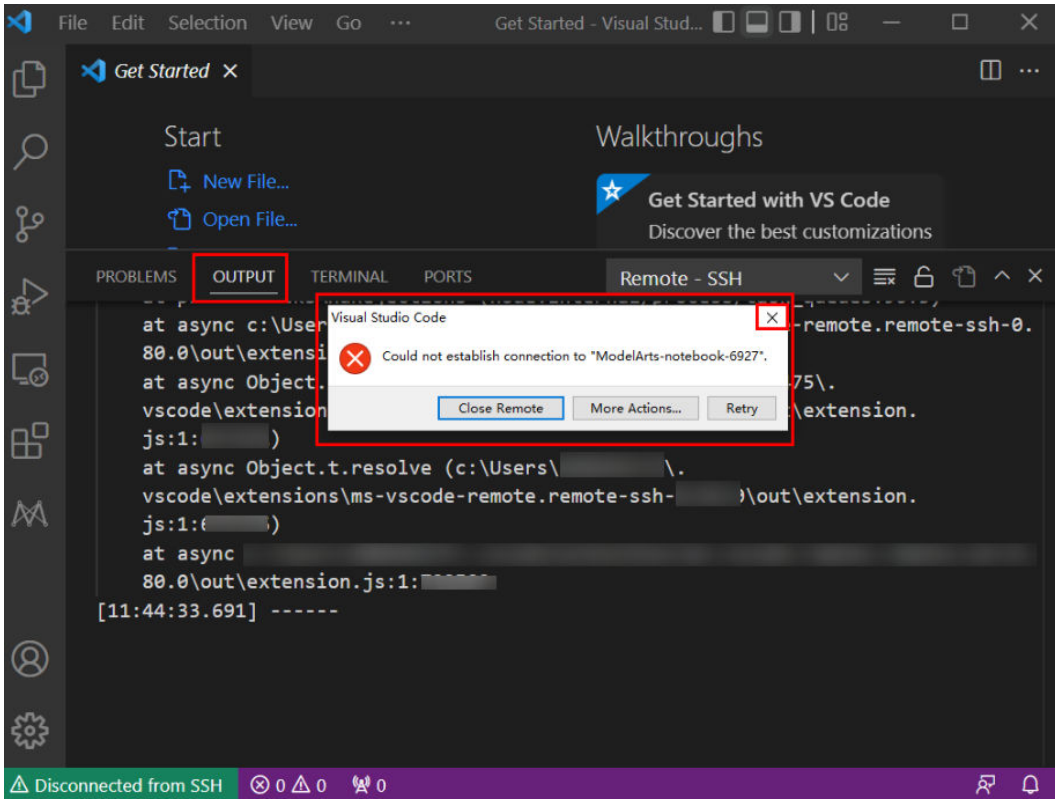
Install VS Code as a non-**root** user, return to the ModelArts management console, and click **Access VS Code**.





4.8.3 ¿Qué hago si se muestra el mensaje de error "Could not establish connection to xxx" durante una conexión remota?

Síntoma



Causa posible

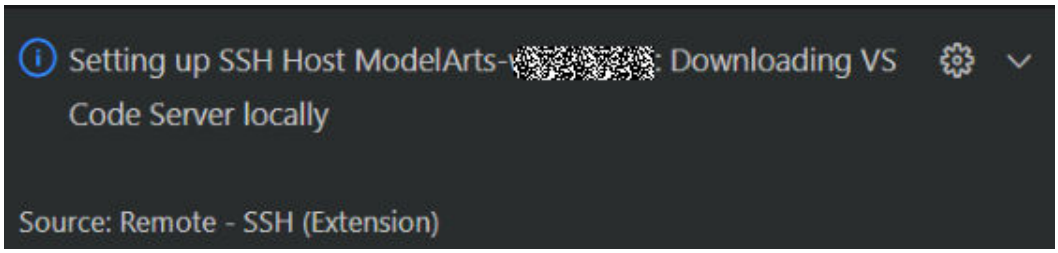
Error al establecer una conexión de SSH remota a una instancia con VS Code.

Solución

Cierre el cuadro de diálogo mostrado, vea la información de error de **OUTPUT** y rectifique el error consultando los métodos de solución de problemas que se proporcionan en las siguientes secciones.

4.8.4 ¿Qué hago si la conexión a un entorno de desarrollo remoto permanece en estado "Setting up SSH Host xxx: Downloading VS Code Server locally" por más de 10 minutos?

Síntoma



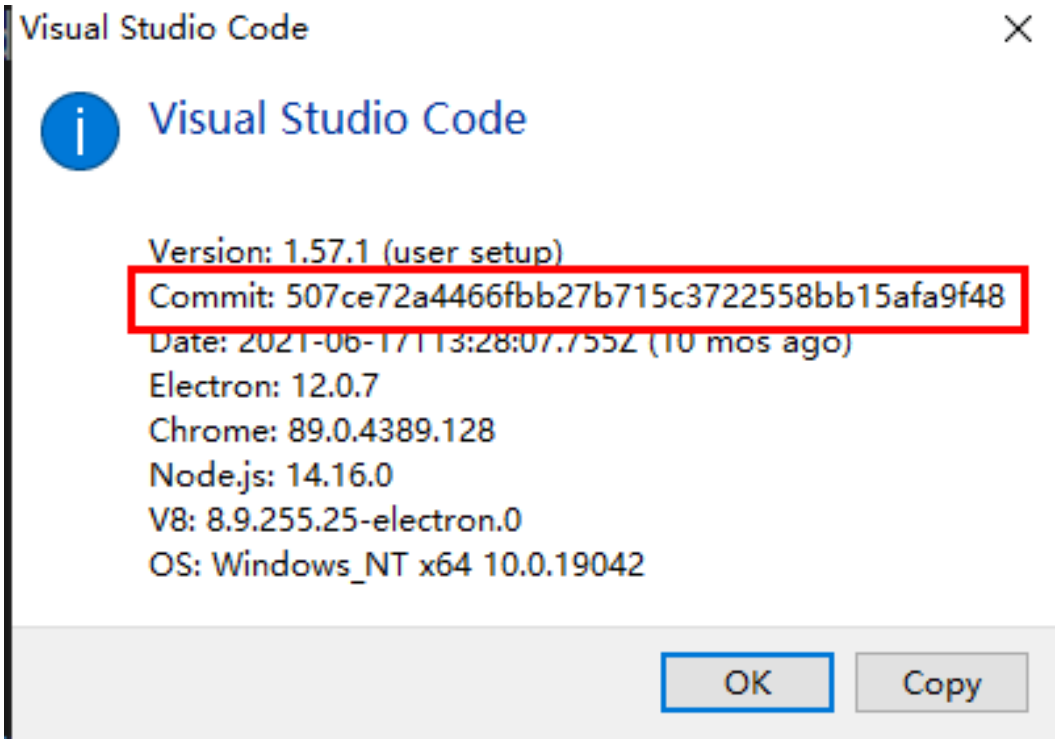
Causa posible

La red local está defectuosa. Como resultado, se necesita mucho tiempo para instalar automáticamente el servidor de VS Code de forma remota.

Solución

Instale manualmente el servidor de VS Code.

Paso 1 Obtenga el ID de confirmación de VS Code.



Paso 2 Descargue el paquete del servidor de VS Code de la versión requerida. Seleccione Arm o x86 según la arquitectura de CPU del entorno de desarrollo.

NOTA

Reemplace *${commitID}* en el siguiente enlace con el ID de confirmación obtenido en la versión de **Paso 1**.

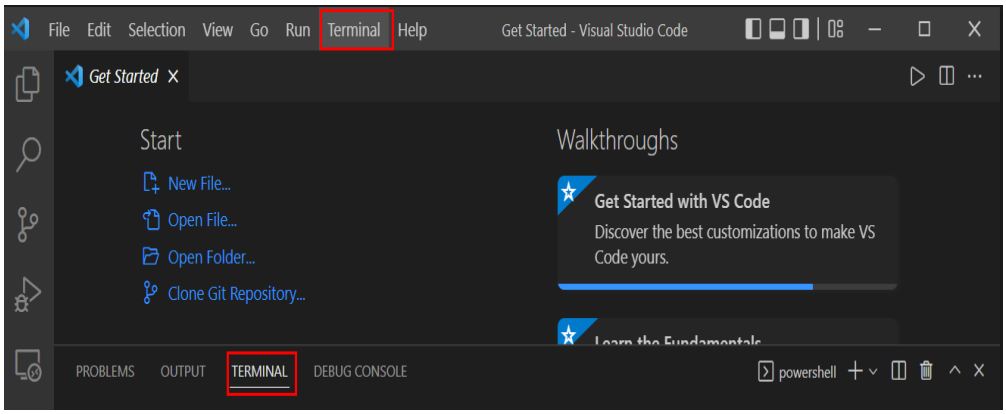
- Para Arm, descargue **vscode-server-linux-arm64.tar.gz**.

  https://update.code.visualstudio.com/commit:${commitID}/server-linux-arm64/stable
- Para x86, descargue **vscode-server-linux-x64.tar.gz**.

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

Paso 3 Acceda al entorno remoto.

Cambie a **Terminal** en VS Code.



Ejecute el siguiente comando en VS Code Terminal para acceder al entorno de desarrollo remoto:


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

Parámetros:

- **IdentityFile**: Ruta de acceso a la clave local

- **User**: Nombre de usuario, por ejemplo, **ma-user**

- **HostName**: Dirección IP

- **Port**: Número de puerto

Paso 4 Instale manualmente el servidor de VS Code.

Ejecute los siguientes comandos en el terminal de VS Code para borrar los datos residuales (reemplace *${commitID}* en los comandos con el ID de confirmación obtenido en **Paso 1**):

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

Cargue el paquete de servidor de VS Code al entorno de desarrollo.

```
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}
```

Parámetros:

- **IdentityFile**: Ruta de acceso a la clave local

- **User**: Nombre de usuario, por ejemplo, **ma-user**

- **HostName**: Dirección IP

- **Port**: Número de puerto

Tomemos Arm como ejemplo. Descomprima el paquete de servidor de VS Code a **$HOME/.vscode-server/bin**. Reemplace *${commitID}* en el comando con el ID de confirmación obtenido en **Paso 1**.

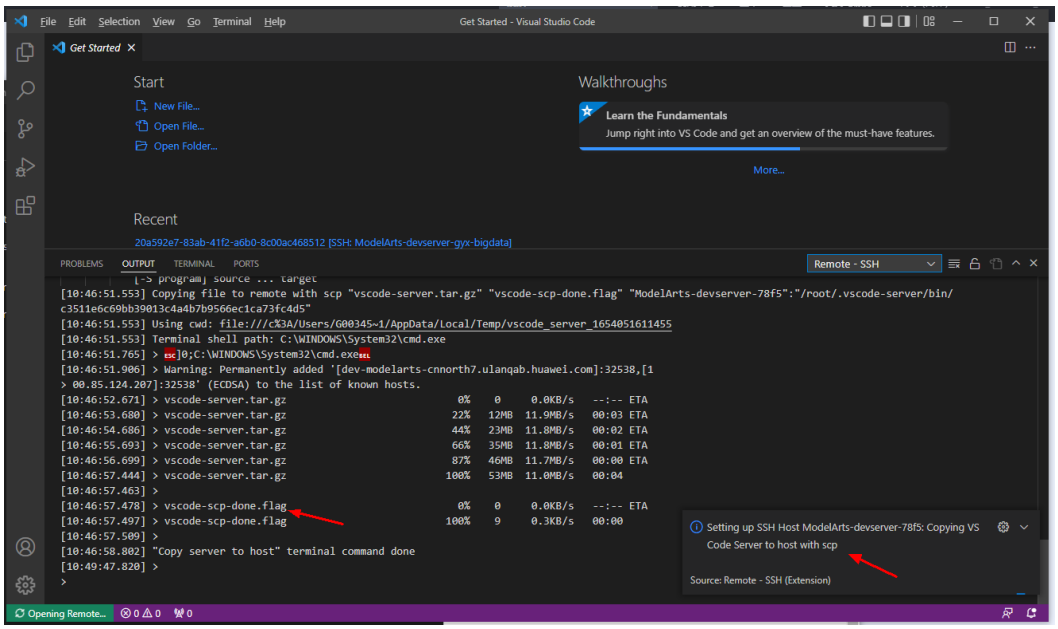
```
cd /home/ma-user/.vscode-server/bin
tar -zxf vscode-server-linux-arm64.tar.gz
mv vscode-server-linux-arm64/* ${commitID}
```

Paso 5 Vuelva a establecer la conexión remota.

----Fin

4.8.5 ¿Qué debo hacer si la conexión a un entorno de desarrollo remoto permanece en el estado de "Setting up SSH Host xxx: Downloading VS Code Server locally" por más de 10 minutos?

Síntoma



Causa posible

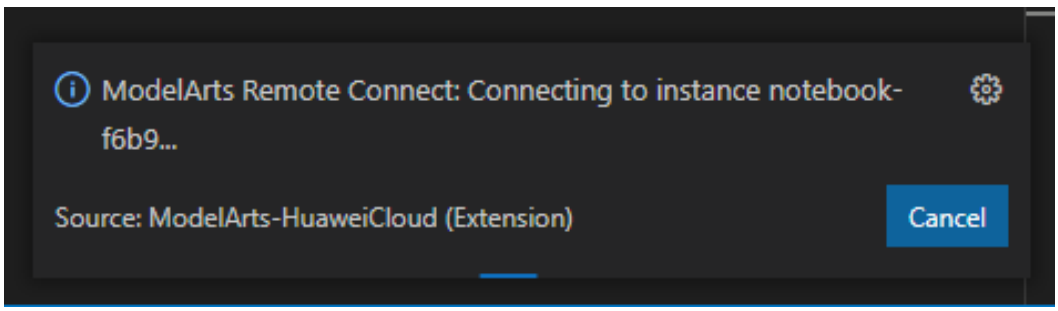
Los logs muestran que **vscode-scp-done.flag** se ha cargado localmente, pero no se recibe en el extremo remoto.

Solución

Cierre todas las ventanas de VS Code, vuelva a la consola de gestión del ModelArts y haga clic en **Access VS Code**.

4.8.6 ¿Qué hago si la conexión a un entorno de desarrollo remoto permanece en el estado de "ModelArts Remote Connect: Connecting to instance xxx..." durante más de 10 minutos?

Síntoma

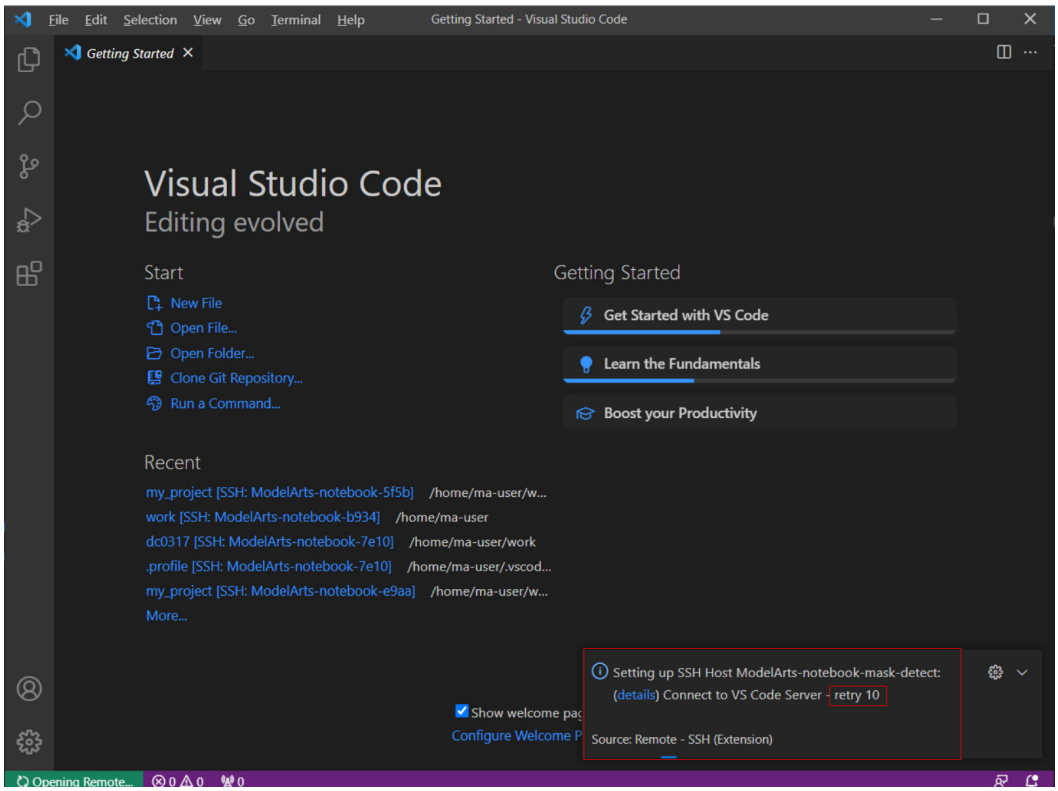


Solución

Haga clic en **Cancel**, vuelva a la consola de gestión del ModelArts y haga clic en **Access VS Code**.

4.8.7 ¿Qué hago si una conexión remota está en el estado de reintento?

Síntoma



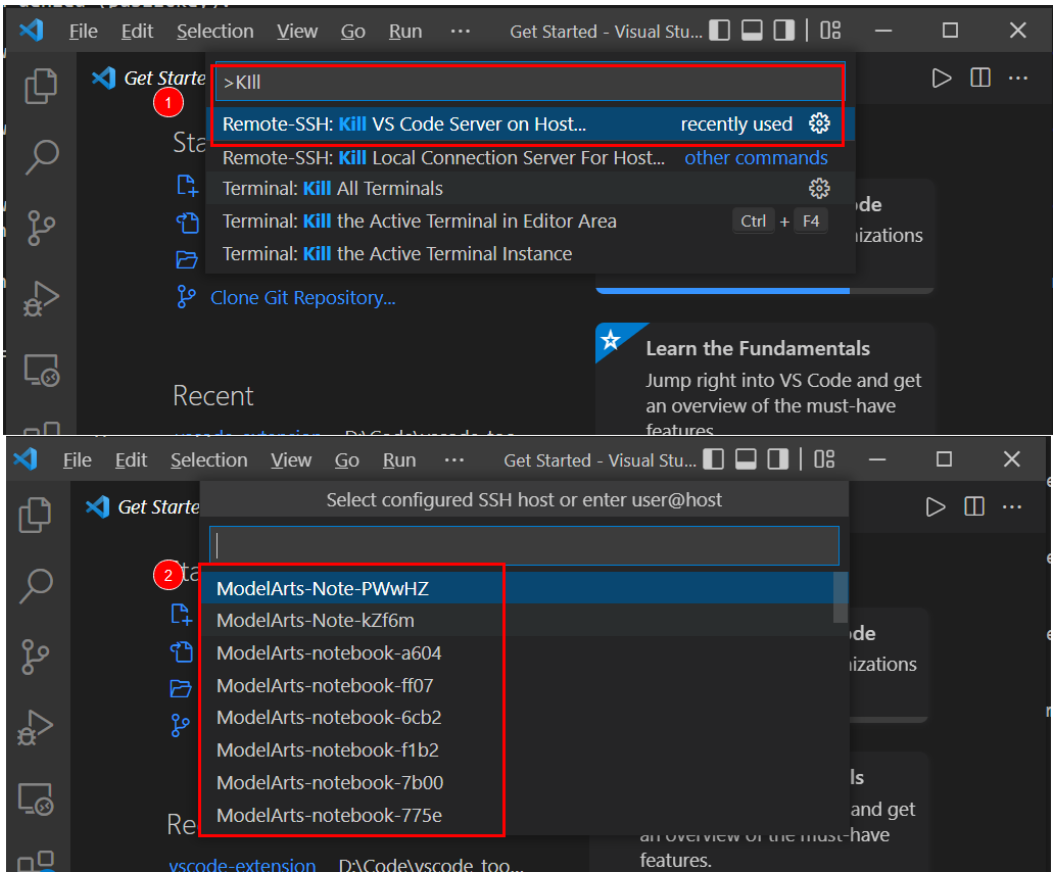
Causa posible

La descarga del servidor de VS Code falló antes, lo que llevó a datos residuales. Como resultado, no se puede realizar una nueva descarga.

Solución

Método 1 (realizado localmente): Abra el panel de comandos (**Ctrl+Shift+P** para Windows y **Cmd+Shift+P** para Mac), encuentre **Kill VS Code Server on Host** y localice la instancia afectada, que se borrará automáticamente. Luego, vuelva a establecer la conexión.

Figura 4-7 Borrar la instancia afectada



Método 2 (realizado de forma remota): Elimine los archivos que se están utilizando en /**home/ma-user/.vscode-server/bin/** en el terminal de VS Code. Luego, vuelva a establecer la conexión.

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

Parámetros:

- **IdentityFile**: Ruta de acceso a la clave local

- **User**: Nombre de usuario, por ejemplo, **ma-user**

- **HostName**: Dirección IP

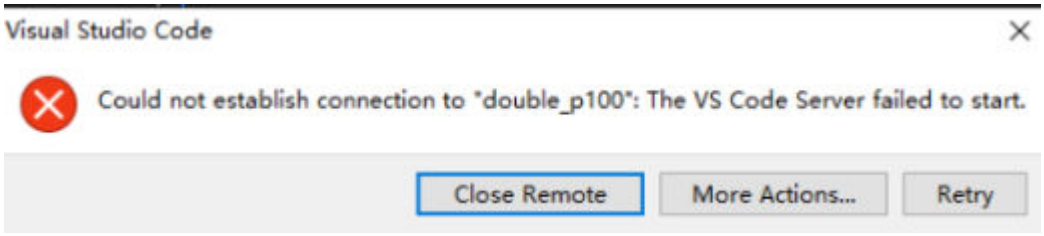
- **Port**: Número de puerto

NOTA

Los métodos anteriores también se pueden utilizar para resolver problemas relacionados con el servidor de VS Code.

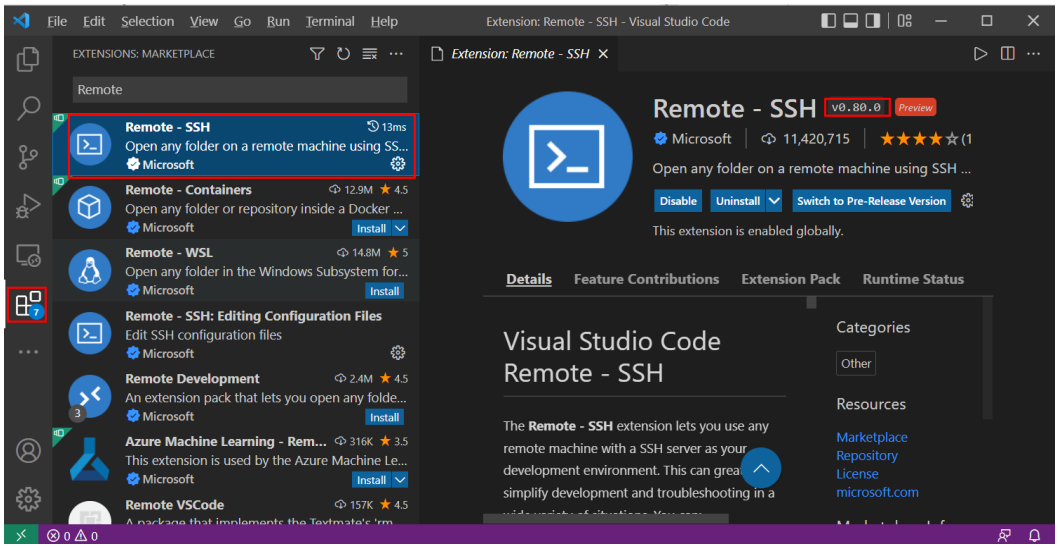
4.8.8 ¿Qué hago si se muestra el mensaje de error "The VS Code Server failed to start"?

Síntoma



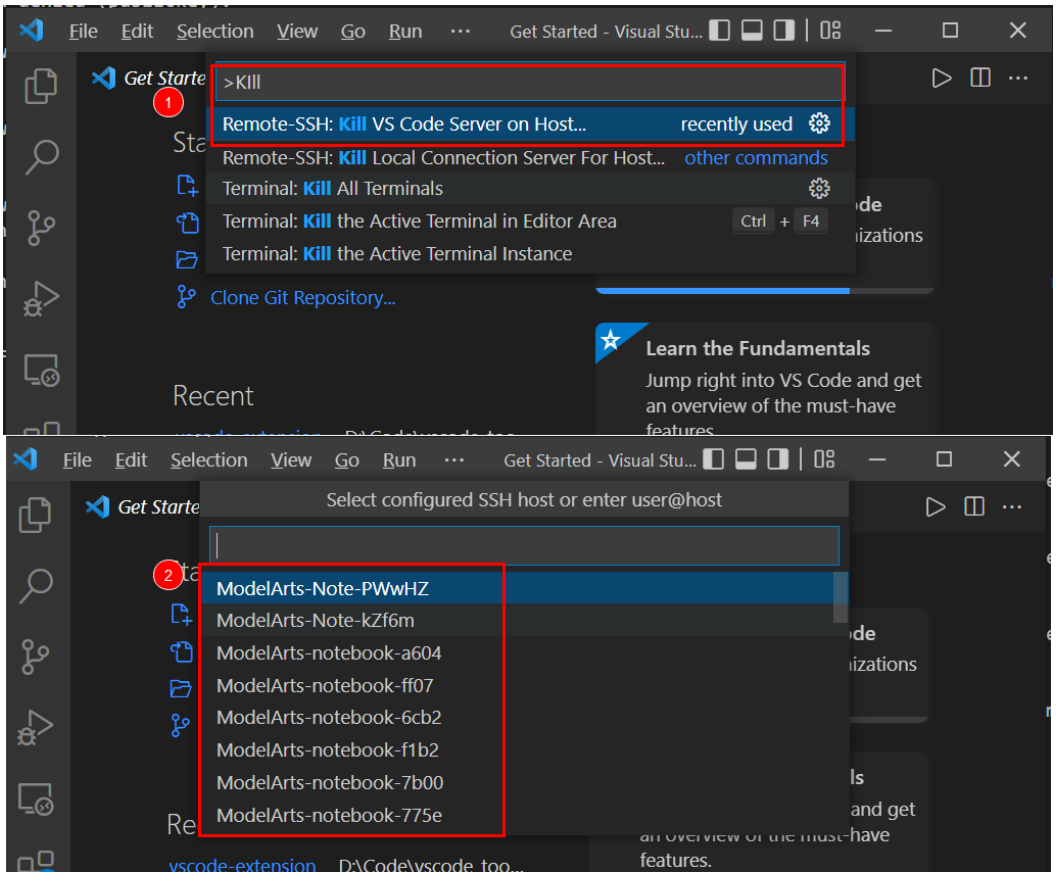
Solución

Paso 1 Compruebe si la versión de VS Code es 1.65.0 o posterior. Si es así, compruebe la versión de Remote-SSH. Si la versión es anterior a 0.76.1, actualice Remote-SSH.



Paso 2 Abra el panel de comandos **Ctrl+Shift+P** para Windows y **Cmd+Shift+P** para Mac, busque **Kill VS Code Server on Host** y localice la instancia afectada, que se borrará automáticamente. Luego, vuelva a establecer la conexión.

Figura 4-8 Borrar la instancia afectada



----Fin

4.8.9 ¿Qué hago si se muestra el mensaje de error "Permissions for 'x:/xxx.pem' are too open"?

Síntoma

Causa posible

Posible causa 1: El archivo de clave no se almacena en la ruta de acceso especificada. Para obtener más información, consulte las **restricciones de seguridad** o el **documento de VS Code**. Resuelva este problema haciendo referencia a la solución 1.

Posible causa 2: Para Mac o Linux, el permiso en el archivo de clave o la carpeta donde se almacena la clave puede ser incorrecto. Resuelva este problema haciendo referencia a la solución 2.

Solución

Solución 1:

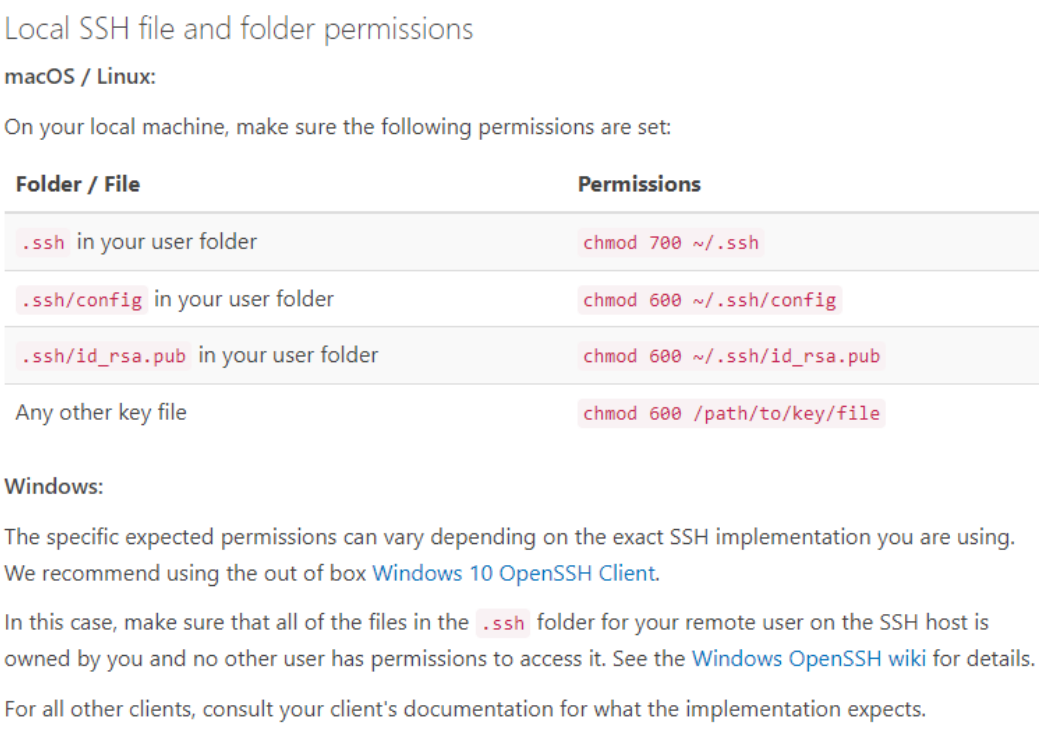
Coloque el archivo clave en una ruta de acceso especificada o su subruta:

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

Mac o Linux: **Users/{{user}}**

Solución 2:

Compruebe los permisos de archivo y carpeta.



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.

4.8.10 ¿Qué hago si se muestra un mensaje de error Bad owner or permissions on C:\Users\Administrator/.ssh/config" o "Connection permission denied (publickey)"?

Síntoma

Aparece el siguiente mensaje de error: Bad owner or permissions on C:\Users \Administrator/.ssh/config" o "Connection permission denied (publickey)". Asegúrese de que

el archivo de clave está seleccionado correctamente y que el permiso del archivo es correcto. You can view the instance keypair information on ModelArts console."

Causas posibles

El permiso para la carpeta SSH se ha concedido a otros usuarios, no solo al usuario actual de Windows, o el usuario actual no tiene el permiso. En estos casos, solo necesita modificar el permiso.

Solución

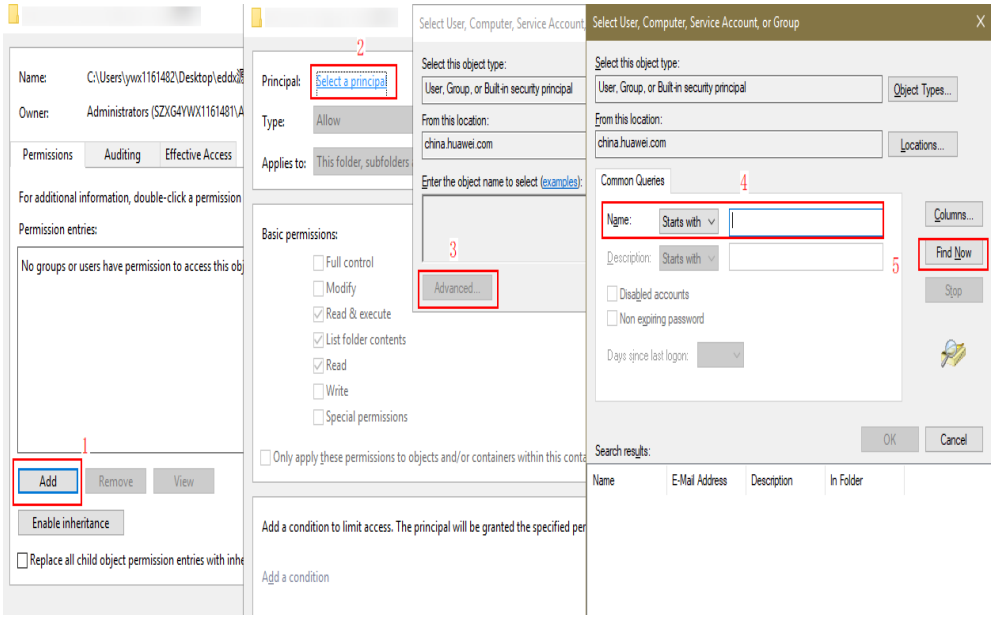
1. Encuentre la carpeta SSH, que normalmente se encuentra en la **C:\Users** por ejemplo, **C:\Users\xxx**.

   NOTA

   El nombre de archivo de **C:\Users** debe ser el mismo que el nombre de usuario de inicio de sesión de Windows.

2. Haga clic con el botón derecho del ratón en la carpeta y elija **Properties**. A continuación, haga clic en la ficha **Security**.
3. Haga clic en **Advanced**. En la ventana que se muestra, haga clic en **Disable inheritance**. A continuación, en el cuadro de diálogo **Block Inheritance**, haga clic en **Remove all inherited permissions from this object**. En este caso, todos los usuarios serán eliminados.
4. Agregue un propietario. En la misma ventana, haga clic en **Add**. En la ventana que se muestra, haga clic en **Select a principal** junto a **Principal**. En el cuadro de diálogo **Select User, Computer, Service Account, or Group**, haga clic en **Advanced**, escriba el nombre de usuario y haga clic en **Find Now**. A continuación, se mostrarán los resultados de la búsqueda. Seleccione su cuenta y haga clic en **OK** para cerrar todas las ventanas.

   Figura 4-9 Adición de un propietario

5. Cierre y abra VS Code de nuevo e intente acceder de forma remota al host SSH. Asegúrese de que la clave de destino está almacenada en la carpeta SSH.

4.8.11 ¿Qué hago si se muestra el mensaje de error "ssh: connect to host xxx.pem port xxxxx: Connection refused"?

Síntoma

Causa posible

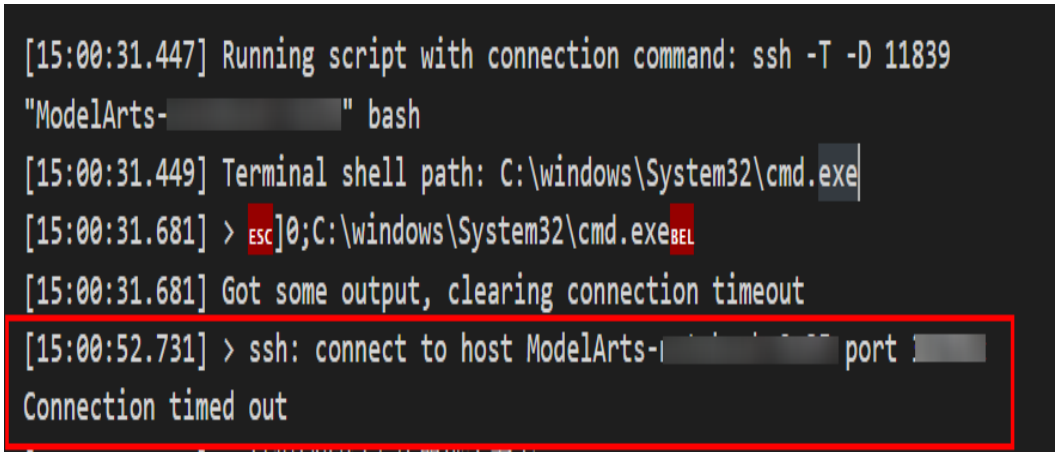
La instancia de destino no se está ejecutando.

Solución

Inicie sesión en la consola de gestión de ModelArts y compruebe el estado de la instancia. Si se detiene la instancia, iníciela. Si la instancia está en otros estados, como **Error**, deténgala y luego iníciela. Después de que el estado de la instancia cambie a **Running**, vuelva a establecer la conexión remota.

4.8.12 ¿Qué hago si se muestra el mensaje de error "ssh: connect to host ModelArts-xxx port xxx: Connection timed out"?

Síntoma



Causa posible

Causa posible 1: Las direcciones IP de la lista blanca configuradas para la instancia son diferentes de las utilizadas en la red local.

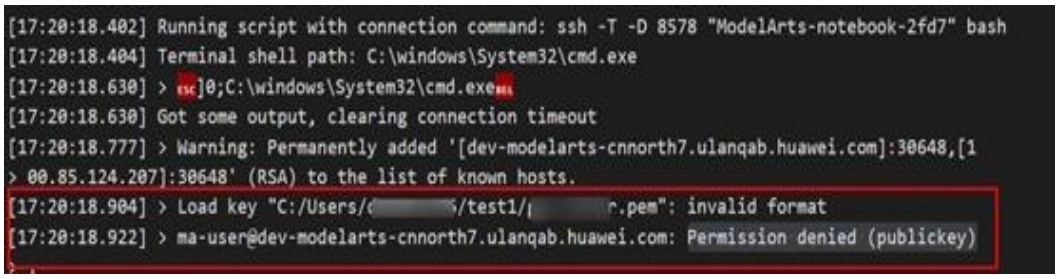
Cambiar la lista blanca para que las direcciones IP de la lista blanca sean las mismas que las usadas en la red local o deshabilitar la lista blanca.

Posible causa 2: La red local es inaccesible.

Solution: Check the local network and network restrictions.

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

Symptom



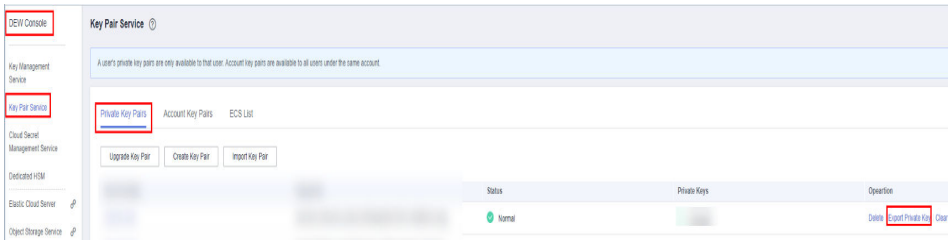
Possible Cause

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:

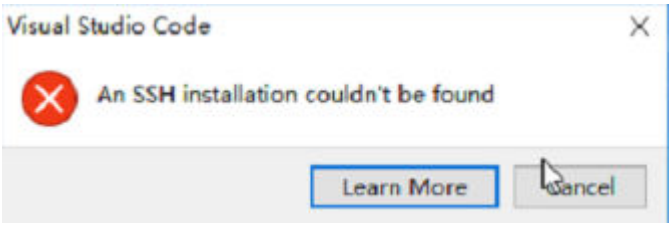
1. Log in to the HUAWEI CLOUD console, search for **DEW**. On the DEW management console, choose **Key Pair Service** and click **Private Key Pairs**. Then, view and download the correct key file.



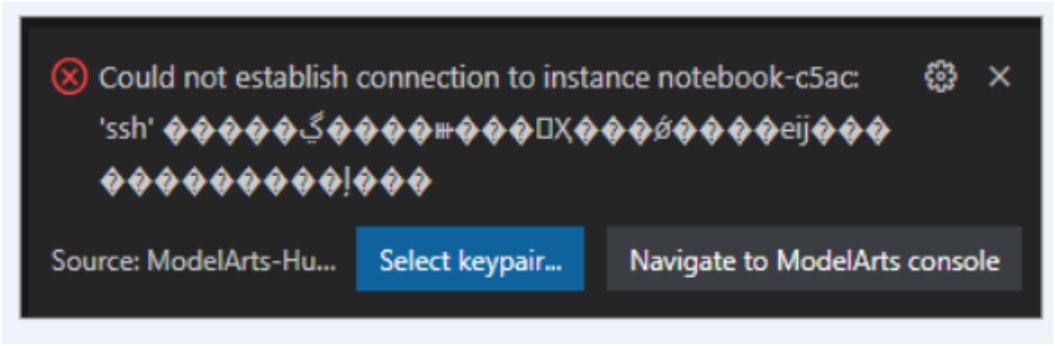
2. If the key cannot be downloaded and the originally downloaded key was lost, create a new development environment instance and a new key file. Replacing a key file in a running development environment will be supported later.

4.8.14 ¿Qué hago si se muestra el mensaje de error "An SSH installation couldn't be found" o "Could not establish connection to instance xxx: 'ssh' ..."?

Síntoma



O



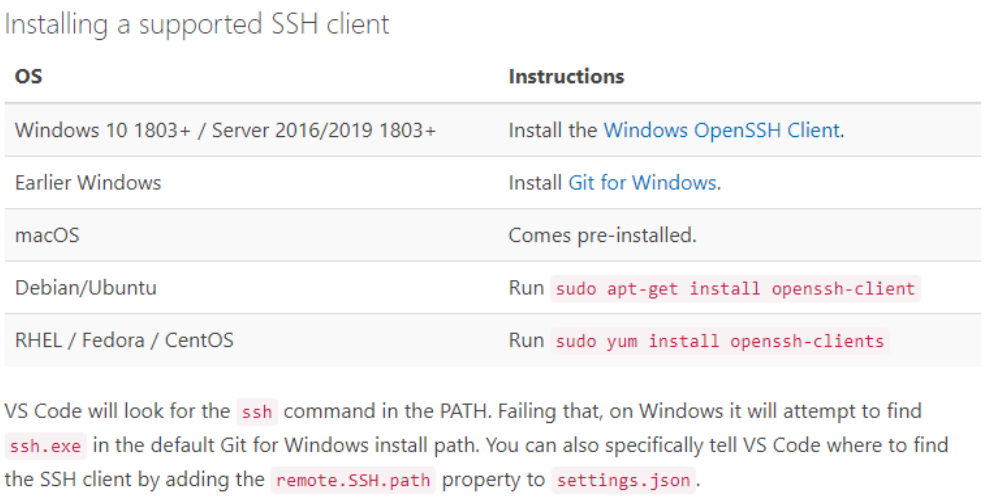
Cuando VS Code intenta tener acceso a una instancia de notebook, el sistema siempre le pide que seleccione un certificado y el mensaje, excepto el título, consta de caracteres confusos. Después de seleccionar el certificado, el sistema sigue sin responder y la conexión falló.

Causa posible

OpenSSH no está instalado en el entorno actual o no está instalado en la ruta de acceso predeterminada. Para obtener más información, consulte el **documento de VS Code**.

Solución

- Si OpenSSH no está instalado en el entorno actual, **descárguelo e instálelo**.



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`.

Si OpenSSH no se instala, **descargue el paquete de instalación de OpenSSH** manualmente y realice las siguientes operaciones:

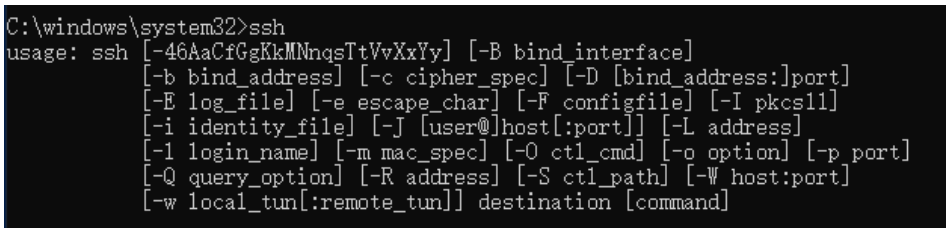
Paso 1 Descargue el paquete .zip y descomprima en **C:\Windows\System32**.

Paso 2 En el **C:\Windows\System32**, abra CMD como administrador y ejecute el siguiente comando:

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

Paso 3 Agregue **C:\Program Files\OpenSSH-xx** (en el que se almacena el archivo ejecutable .exe SSH) a las variables del sistema de entorno.

Paso 4 Abra CMD de nuevo y ejecute **ssh**. Si se muestra la siguiente información, la instalación se realiza correctamente. De lo contrario, vaya a **Paso 5** y **Paso 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]
```

Paso 5 Habilite el puerto 22 (puerto OpenSSH predeterminado) en el firewall y ejecute el siguiente comando en el símbolo del sistema:

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

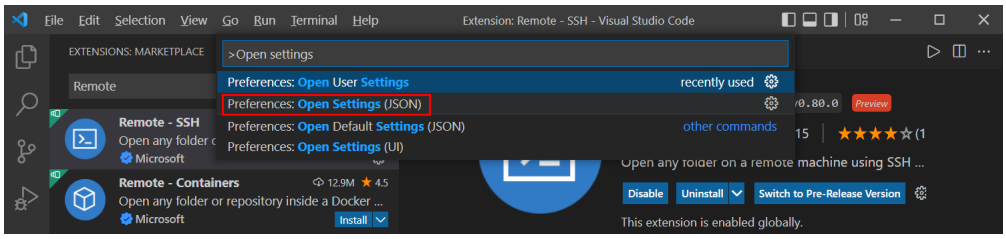
Paso 6 Ejecute el siguiente comando para iniciar OpenSSH:

```
Start-Service sshd
```

----Fin

- Si OpenSSH no está instalado en la ruta predeterminada, abra el panel de comandos **Ctrl+Shift+P** para Windows y **Cmd+Shift+P** para Mac.

  Búsqueda de **Open settings**.

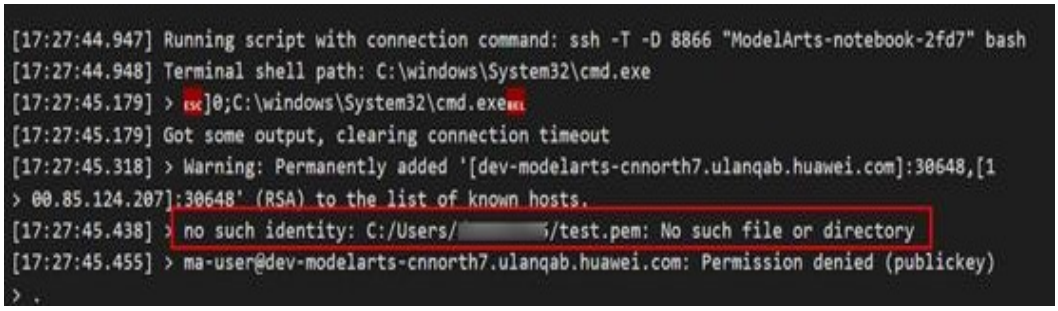


Agregue **remote.SSH.path** a **settings.json**. Por ejemplo, **"remote.SSH.path": "*Installation path of the local OpenSSH*"**.

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

4.8.15 ¿Qué hago si se muestra un mensaje de error "no such identity: C:/Users/xx /test.pem: No such file or directory"?

Síntoma



Causa posible

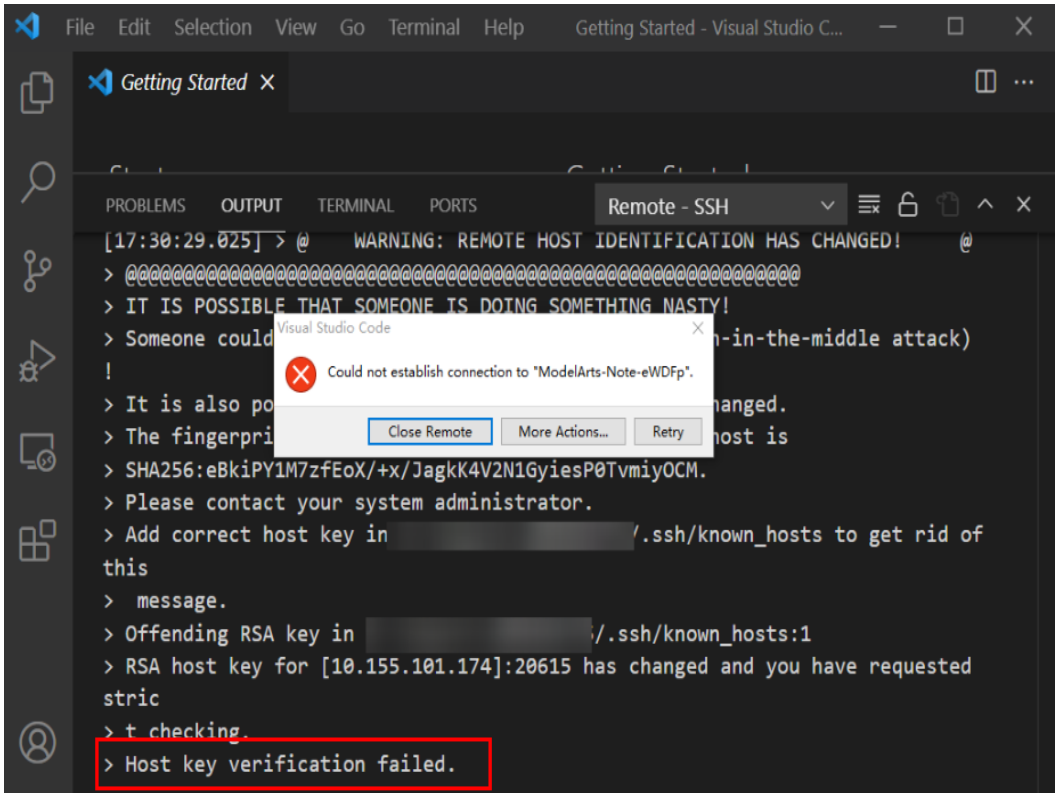
El archivo de clave no está en la ruta o se ha cambiado el nombre del archivo de clave en la ruta.

Solución

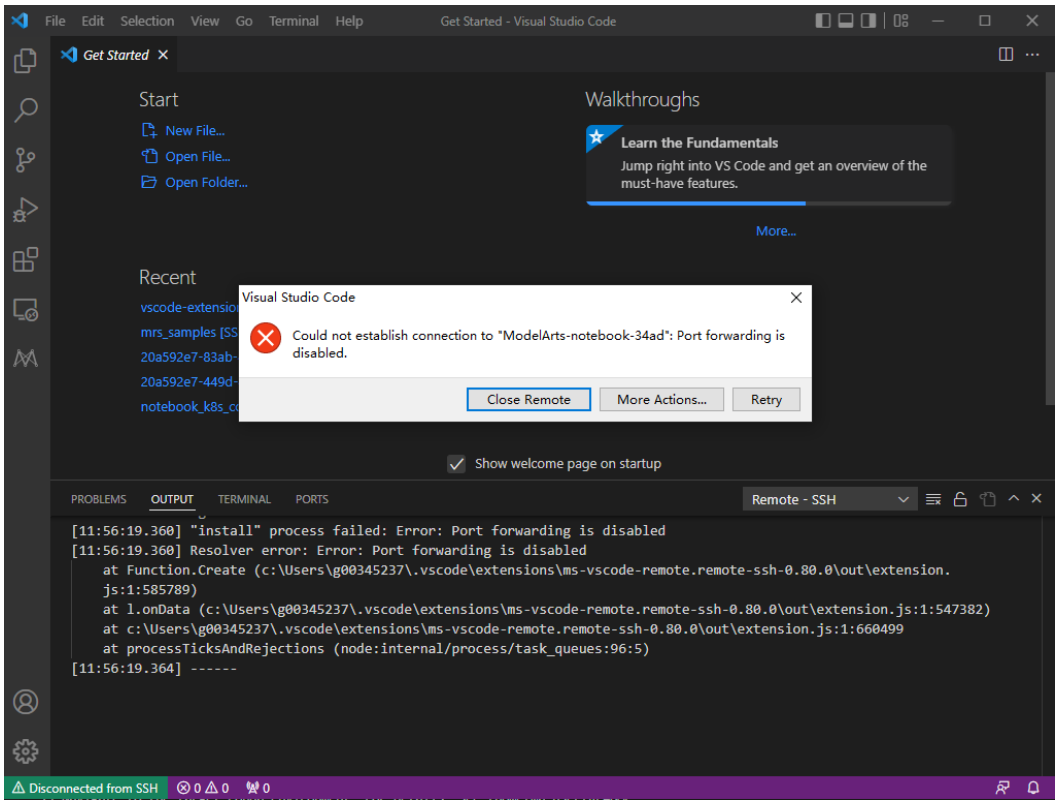
Vuelva a seleccionar la ruta de la clave.

4.8.16 ¿Qué hago si se muestra el mensaje de error "Host key verification failed" o "Port forwarding is disabled"?

Síntoma



O



Causa posible

Después de reiniciar la instancia del notebook, su clave pública cambia. La alarma se genera cuando OpenSSH detectó el cambio de clave.

Solución

- Agregue **-o StrictHostKeyChecking=no** para el acceso remoto con CLI en VS Code.

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

  Parámetros:

  - **IdentityFile**: Ruta de acceso a la clave local
  - **User**: Nombre de usuario, por ejemplo, **ma-user**
  - **HostName**: Dirección IP
  - **Port**: Número de puerto

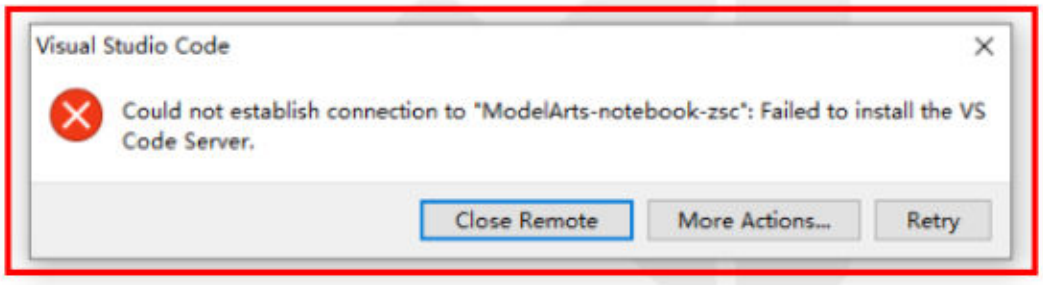
- Agregue **StrictHostKeyChecking no** y **UserKnownHostsFile=/dev/null** al archivo **ssh config** local para la configuración manual del acceso remoto en VS Code.

  ```
  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
  ```

Tenga en cuenta que los inicios de sesión SSH serán inseguros después de que se agreguen los parámetros anteriores porque el archivo **known_hosts** será ignorado durante los inicios de sesión.

4.8.17 ¿Qué hago si se muestra el mensaje de error "Failed to install the VS Code Server" o "tar: Error is not recoverable: exiting now"?

Síntoma



O

```
[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==
```

Causa posible

El espacio en disco de **/home/ma-user/work** es insuficiente.

Solución

Elimine archivos innecesarios de **/home/ma-user/work**.

4.8.18 ¿Qué hago si se muestra el mensaje de error "XHR failed" cuando se accede a una instancia de notebook remota a través de VS Code?

Causa posible

La red del entorno puede ser defectuosa.

Solución

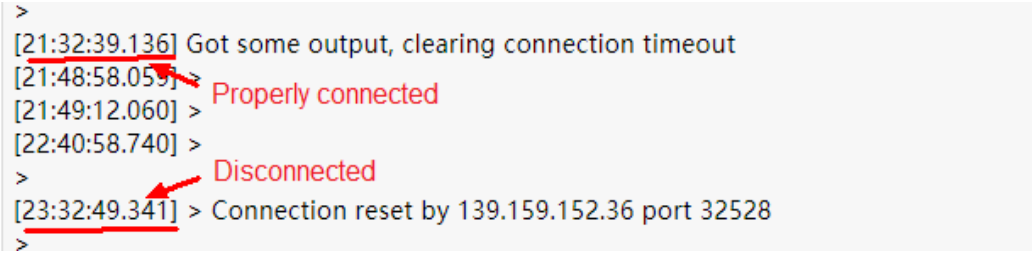
Rectifique el error haciendo referencia a **Solución de problemas de XHR fallido**.

4.8.19 ¿Qué hago para una conexión de VS Code desconectada automáticamente si no se realiza ninguna operación durante mucho tiempo?

Síntoma

Después de establecer una conexión de SSH a través del VS Code, no se realiza ninguna operación durante mucho tiempo y la ventana se mantiene abierta. Cuando se utiliza de nuevo la conexión, se encuentra que la conexión está desconectada y no se muestra ningún mensaje de error.

De acuerdo con los logs de VS Code Remote-SSH, la conexión se desconectó aproximadamente dos horas después de la configuración.



Causa posible

Después de que la interacción de SSH se detenga durante un período de tiempo, el servidor de seguridad desconecta las conexiones inactivas (http://bluebiu.com/blog/linux-ssh-session-alive.html). La configuración de SSH predeterminada no conduce a una desconexión proactiva tras el tiempo de espera. Dado que la instancia se ejecuta de forma estable en el backend, configure la conexión de nuevo para resolver este problema.

Solución

Para conservar las conexiones si no se realiza ninguna operación durante mucho tiempo, configure el envío periódico de mensajes con SSH. De esta manera, la conexión no quedará inactiva en el firewall.

- Configure el cliente según sea necesario. Si el cliente no está configurado, no se enviará ningún paquete de latidos al servidor de forma predeterminada.

Figura 4-10 Abrir el archivo de configuración de SSH de VS Code

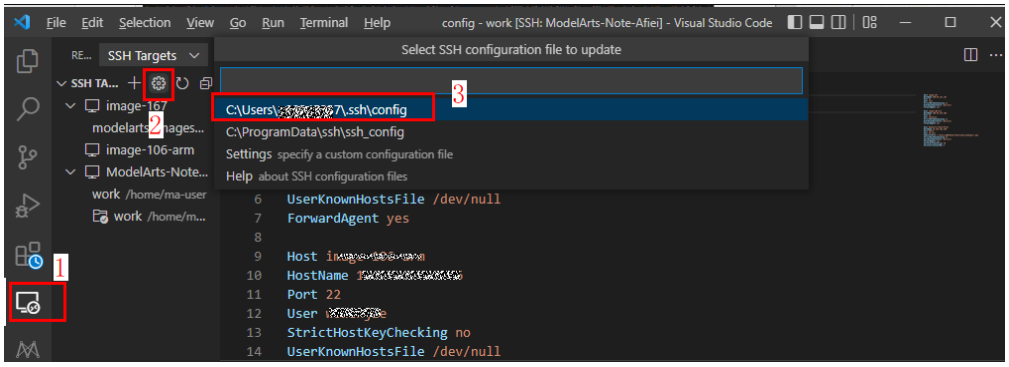
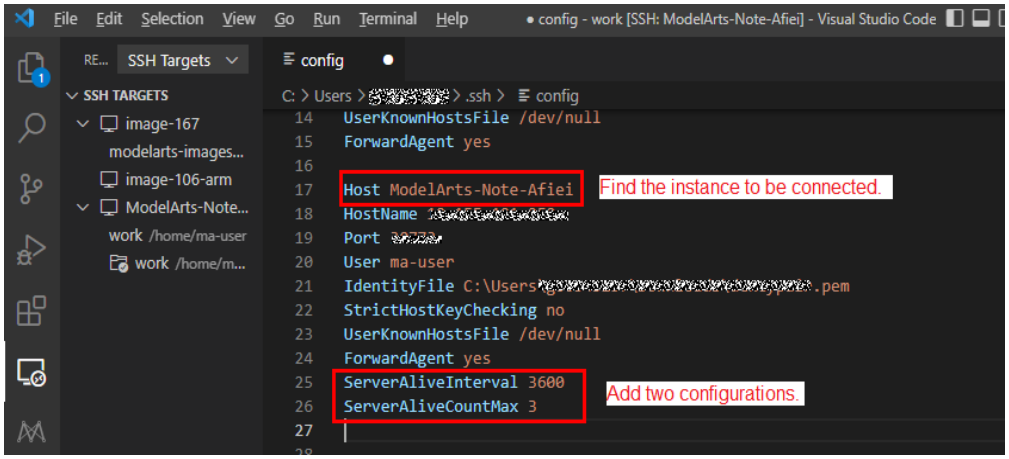


Figura 4-11 Adición de configuraciones



La configuración es la siguiente:

```
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.
```

Por ejemplo, si el firewall está configurado para desconectar una conexión si la conexión está inactiva durante dos horas, establezca **ServerAliveInterval** en un valor inferior a dos horas (por ejemplo, una hora) en el cliente para evitar que el firewall desconecte la conexión.

- Configure el servidor de **/home/ma-user/.ssh/etc/sshd_config** (Se ha configurado el notebook y 24 horas es más que el tiempo configurado en el firewall para desconectar las conexiones. Esta configuración no necesita modificarse manualmente. Solo se utiliza para ayudar a entender la configuración de SSH.)

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

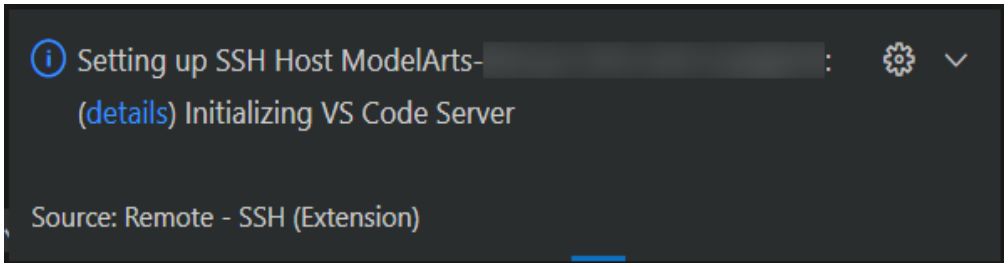
La configuración anterior muestra que el servidor envía activamente un paquete de latido al cliente cada 24 horas, y la conexión se desconectará si el cliente no responde después de que el paquete de latido se envíe tres veces.

Para obtener más información, véase https://unix.stackexchange.com/questions/3026/what-do-options-serveraliveinterval-and-clientaliveinterval-in-sshd-config-d.

- Si una conexión debe conservarse consistentemente, es una buena práctica escribir logs en un archivo de log separado y ejecutar el script en el backend. Por ejemplo:
  ```
  nohup train.sh > output.log 2>&1 & tail -f output.log
  ```

4.8.20 ¿Qué hago si toma mucho tiempo configurar una conexión remota después de actualizar automáticamente VS Code?

Síntoma



Causa posible

VS Code se actualiza automáticamente. Como resultado, descargue el nuevo servidor de VS Code para configurar una nueva conexión.

Solución

Deshabilitar la actualización automática de VS Code. Para ello, haga clic en **Settings** en la esquina inferior izquierda, busque **Update: Mode** y configúrelo en **none**.

Figura 4-12 Ajustes

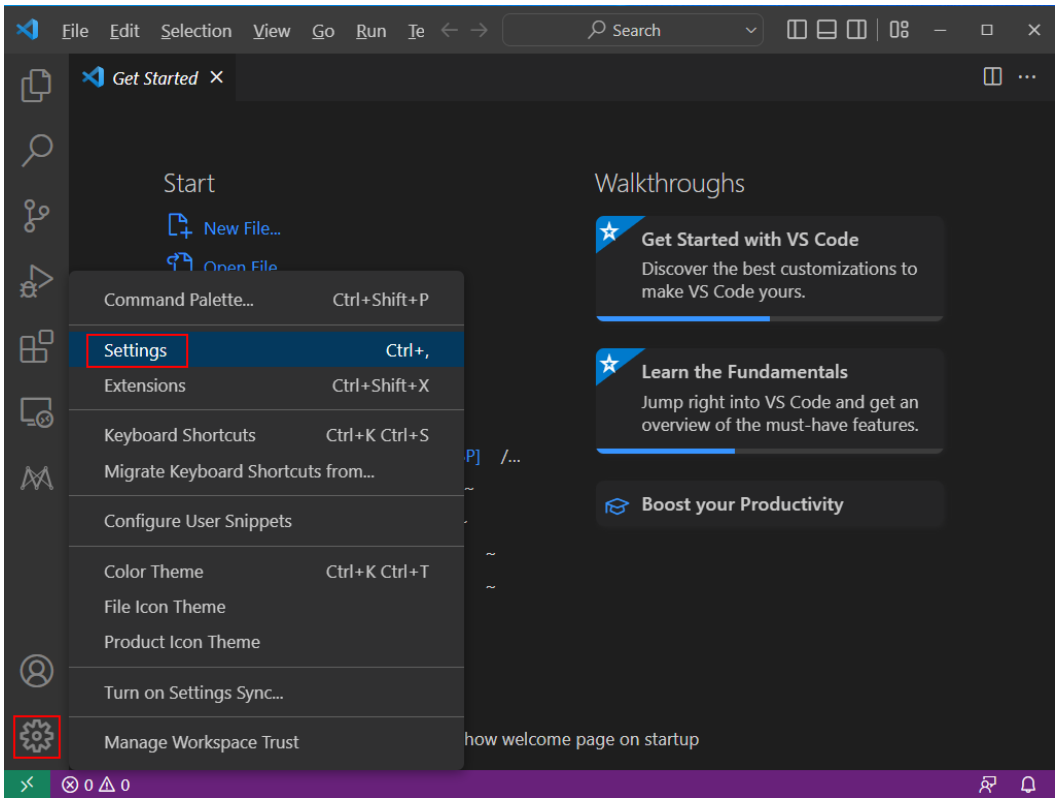
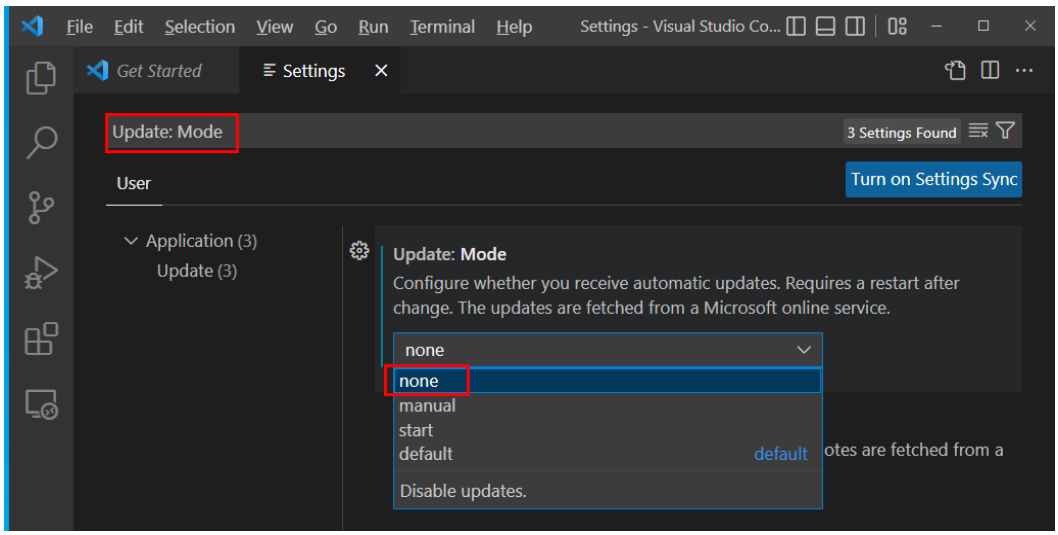
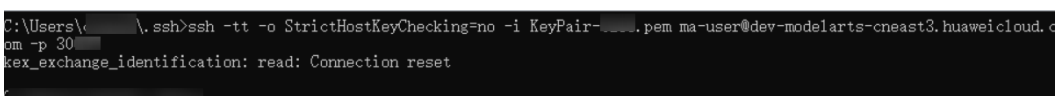


Figura 4-13 Establecer el modo de actualización en ninguno



4.8.21 ¿Qué hago si se muestra el mensaje de error "Connection reset" durante una conexión de SSH?

Síntoma



Causas posibles

La red de usuario está restringida. Por ejemplo, SSH está deshabilitado de forma predeterminada en algunas redes de empresa.

Solución

Solicite el permiso de SSH.

4.8.22 ¿Qué puedo hacer si una instancia de Notebook se desconecta o se atasca con frecuencia después de usar MobaXterm para conectarme a la instancia de Notebook en modo SSH?

Síntoma

Una vez que MobaXterm se conecta a un entorno de desarrollo, se desconecta después de un período de tiempo.

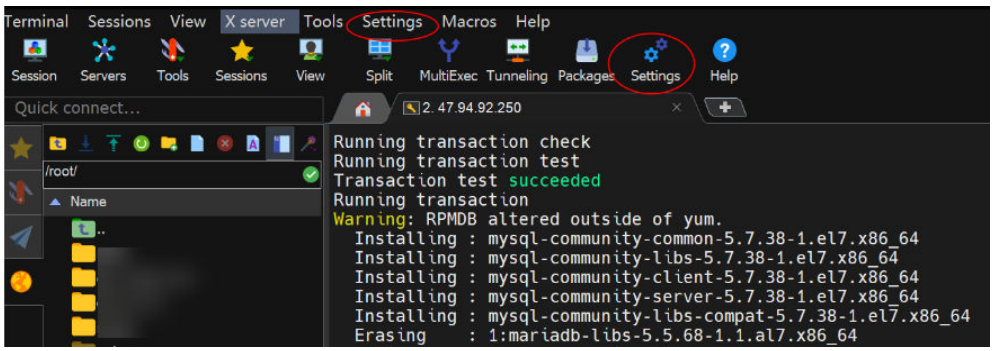
Causa posible

Cuando MobaXterm está configurado, **SSH keepalive** no está seleccionado o **Stop server after** de MobaXterm se establece en un valor que es demasiado pequeño.

Solución

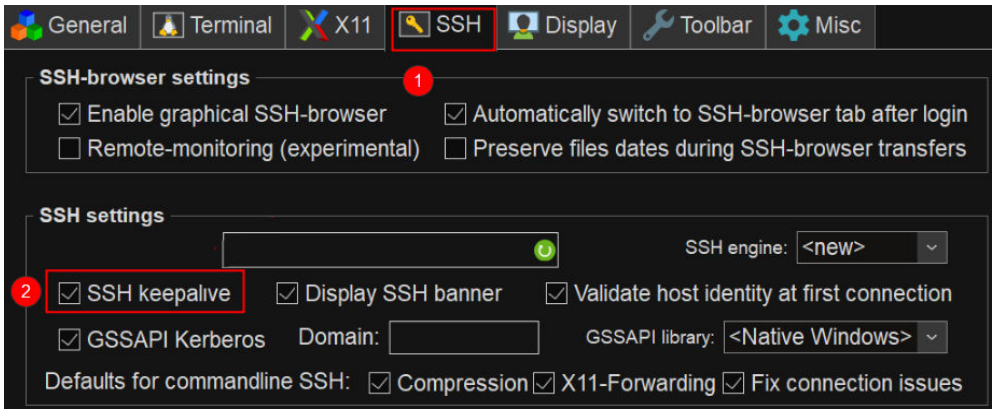
Paso 1 Abra MobaXterm y haga clic en **Settings** en la barra de menús.

Figura 4-14 Ajustes (Settings)



Paso 2 En la página de configuración MobaXterm, haga clic en la ficha **SSH** y seleccione **SSH keepalive**.

Figura 4-15 Selección de la función de mantenimiento de SSH

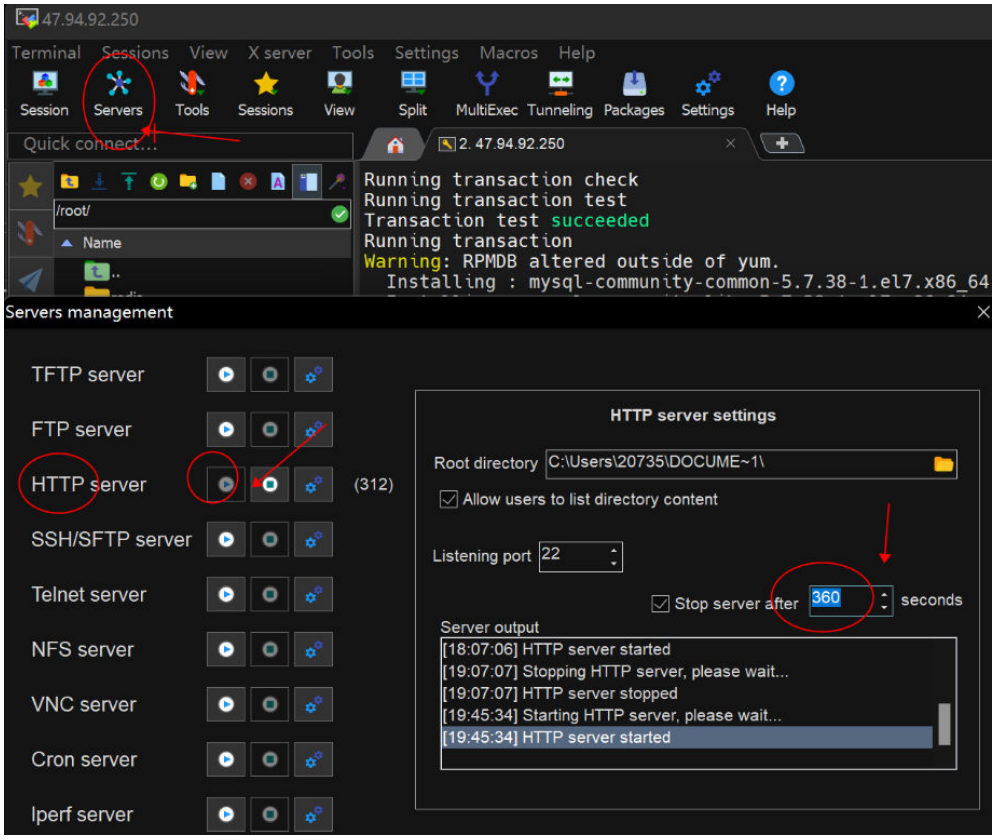


NOTA

Si se utiliza MobaXterm Professional, vaya al **paso 3**.

Paso 3 Cambie el valor predeterminado **360 seconds** a **3600 seconds** o un valor mayor para **Stop server after**.

Figura 4-16 Configuración de Stop server after



----Fin

4.9 Otros

4.9.1 ¿Cómo uso varias tarjetas de Ascend para la depuración en una instancia de notebook?

Un trabajo de entrenamiento multi-tarjeta de Ascend se ejecuta en modo de multi-proceso y multi-tarjeta. El número de tarjetas es igual al número de procesos de Python. El subcapa de Ascend lee la variable de entorno **RANK_TABLE_FILE**, que se ha configurado en el entorno de desarrollo, sin necesidad de configuración manual. Por ejemplo, para ejecutar un trabajo en ocho tarjetas, el código es el siguiente:

```
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)
```

4.9.2 ¿Por qué la velocidad de entrenamiento es similar cuando se usan diferentes variantes para notebook?

Si su trabajo de entrenamiento es de un solo proceso en código, la velocidad de entrenamiento es básicamente la misma sin importar cuándo se use la variante de notebook de 8 vCPU y 64 GB de memoria o la variante de 72 vCPU y 512 GB de memoria. Por ejemplo, si su trabajo de entrenamiento utiliza 2 vCPU y 4 GB de memoria, la velocidad de entrenamiento es similar independientemente de que utilice la variante de notebook de 4 vCPU y 8 GB de memoria o la variante de 8 vCPU y 64 GB de memoria.

Si su trabajo de entrenamiento es multiproceso en código, la velocidad de entrenamiento respaldada por la variante de notebook de 72 vCPU y 512 GB de memoria es mayor que la respaldada por la variante de notebook de 8 vCPU y 64 GB de memoria.

4.9.3 ¿Cómo realizo entrenamiento incremental cuando uso MoXing?

Si no está satisfecho con los resultados del entrenamiento al usar MoXing para crear un modelo, puede realizar un entrenamiento incremental después de modificar algunos datos y la información de etiquetas.

Adición de parámetros de entrenamiento incrementales a mox.run

Después de modificar los datos de etiquetado o los conjuntos de datos, puede modificar el parámetro **log_dir** y agregar el parámetro **checkpoint_path** a **mox.run**. Establezca **log_dir**

en un nuevo directorio y **checkpoint_path** en la ruta de salida de los resultados de entrenamiento anteriores. Si la ruta de salida es un directorio de OBS, establezca la ruta en un valor que comience por **obs://**.

Si se cambian las etiquetas para los datos de etiquetas, realice operaciones de **Si se cambian las etiquetas** antes de ejecutar **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)
```

Si se cambian las etiquetas

Si las etiquetas de un conjunto de datos han cambiado, ejecute la siguiente instrucción. La instrucción debe ejecutarse antes de ejecutar **mox.run**.

En la instrucción, la variable **logits** indica los pesos de capa de clasificación en diferentes redes y se configuran diferentes parámetros. Establezca este parámetro en la palabra clave correspondiente.

```
mox.set_flag('checkpoint_exclude_patterns', 'logits')
```

Si se utiliza la red integrada de MoXing, la palabra clave correspondiente debe obtenerse llamando a la siguiente API. En este ejemplo, la palabra clave **Resnet_v1_50** es el valor de **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)
```

También puede obtener una lista de redes compatibles con MoXing llamando a la siguiente API:

```
import moxing.tensorflow as mox
print(help(mox.NetworkKeys))
```

Se muestra la siguiente información:

```
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'
```

4.9.4 ¿Cómo puedo ver el uso de la GPU en el notebook?

Si selecciona GPU al crear una instancia de notebook, realice las siguientes operaciones para ver el uso de la GPU:

1. Inicie sesión en la consola de gestión de ModelArts y seleccione **DevEnviron > Notebooks**.
2. En la columna **Operation** de la instancia del notebook de destino de la lista del notebook, haga clic en **Open** para ir a la página **Jupyter**.
3. En la ficha **Files** de la página **Jupyter**, haga clic en **New** y seleccione **Terminal**. Se muestra la página **Terminal**.
4. Ejecute el siguiente comando para ver el uso de la GPU:
   ```
   nvidia-smi
   ```
5. Compruebe qué procesos de la instancia actual de notebook utilizan GPU.

   Abra **/resource_info/gpu_usage.json** para ver los procesos que usan las GPU.

   ```
   {
       <notebook name>: {
           <GPU0 UUID>: [
               {
                   "pid": 2263,
                   "processName": "python",
                   "gpuMemoryUsage": "4935Mi"
               },
               {...}
           ]
           <GPU1 UUID>: [...]
       }
   }
   ```

   Si ningún proceso utiliza GPU, es posible que el archivo no exista o esté vacío.

4.9.5 ¿Cómo puedo obtener el uso de GPU con el código?

Ejecute el comando shell o python para obtener el uso de la GPU.

Uso del comando shell

1. Ejecute el comando **nvidia-smi**.

   Esta operación se basa en 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. Ejecute el comando **gpustat**.

   ```
   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 |
   ```

   Para detener la ejecución del comando, presione **Ctrl+C**.

Uso del comando python

1. Ejecute el comando **nvidia-ml-py3** (comúnmente usado).

   ```
   !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. Ejecute los comandos **nvidia_smi**, **wapper** y **prettytable**.

   Utilice el decorador para obtener el uso de la GPU en tiempo real durante el entrenamiento de modelos.


```
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. Ejecute el comando **pynvml**.

   Ejecute **nvidia-ml-py3** para obtener directamente la biblioteca de c-lib de nvml, sin usar **nvidia-smi**. Por lo tanto, se recomienda este comando.

   ```
   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. Ejecute el comando **gputil**.

   ```
   !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
```

Al usar un framework de aprendizaje profundo como PyTorch o TensorFlow también puede usar las API proporcionadas por el framework para consultas.

4.9.6 ¿Qué indicadores de rendimiento en tiempo real de un chip Ascend puedo ver?

El indicador de rendimiento en tiempo real que se puede ver es **npu-smi**, que es similar al **nvidia-smi** de un chip GPU.

4.9.7 ¿El sistema detiene o elimina automáticamente una instancia de notebook si no habilito la parada automática?

La respuesta a esta pregunta difiere dependiendo de las especificaciones de recursos seleccionadas.

- Si utiliza especificaciones gratuitas, la instancia de su notebook se detiene automáticamente después de ejecutarse durante una hora. Si la instancia del notebook no se inicia de nuevo en 72 horas, se eliminará. Por lo tanto, al usar especificaciones gratuitas, preste atención al tiempo de ejecución y haga una copia de respaldo de sus archivos correctamente.
- Si utiliza un grupo de recursos públicos de pago y no habilita la parada automática, la instancia del notebook no se detiene automáticamente o se elimina.
- Si utiliza un grupo de recursos dedicado, la instancia del notebook no se detiene automáticamente. Sin embargo, si se elimina el grupo de recursos dedicado, la instancia del notebook dejará de estar disponible.

4.9.8 ¿Cuáles son las relaciones entre los archivos almacenados en el JupyterLab, Terminal y OBS?

- Los archivos almacenados en el JupyterLab son los mismos que los del directorio de trabajo de la página **Terminal**. Es decir, los archivos se crean en las instancias del notebook o se sincronizan desde OBS.
- Las instancias de notebook con almacenamiento de OBS montado pueden sincronizar archivos de OBS a JupyterLab mediante la función de Sync OBS. Los archivos de la página **Terminal** son los mismos que los de JupyterLab.
- Las instancias de notebook con almacenamiento de EVS montado pueden leer archivos de OBS a JupyterLab mediante la API MoXing o SDK. Los archivos de la página **Terminal** son los mismos que los de JupyterLab.

4.9.9 ¿Cómo puedo migrar datos de una instancia de notebook de versión antigua a una de versión nueva?

El notebook de la versión anterior se discontinuará. Utilice la nueva versión. En esta sección se describe cómo migrar datos de una instancia de notebook de la versión anterior a una instancia de notebook de la nueva versión.

Diferencias de almacenamiento entre las versiones antiguas y nuevas

Tabla 4-1 Almacenamiento compatible con el notebook de las versiones antiguas y nuevas

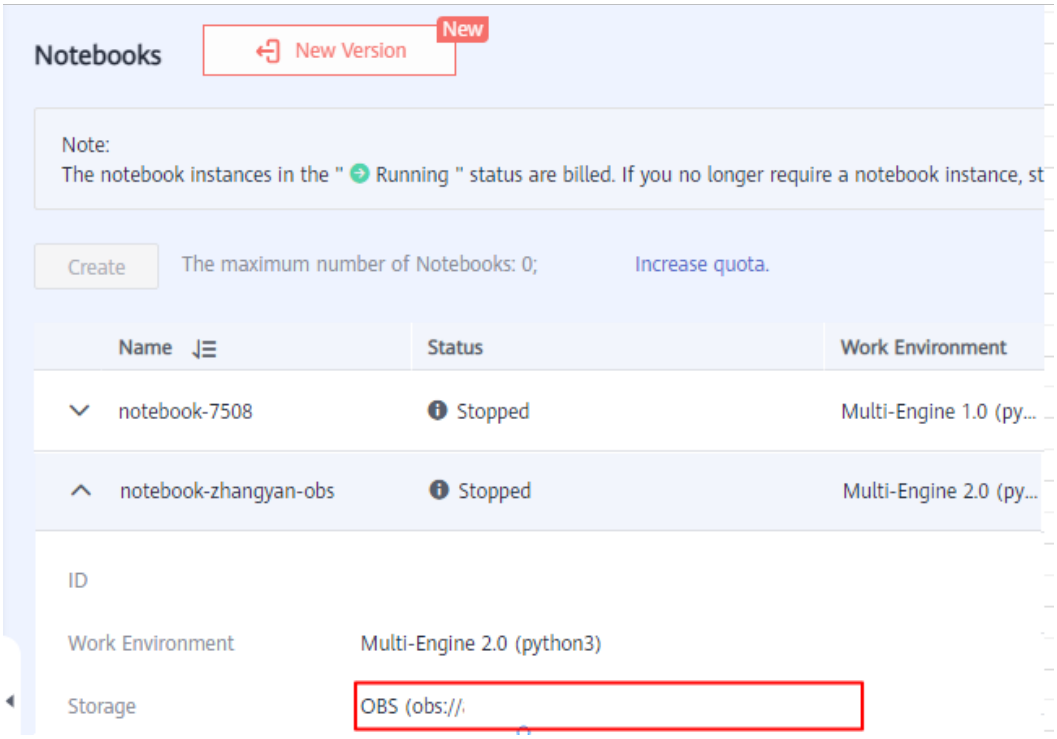
Almacenamiento	Notebook de versión antigua	Notebook de versión nueva	Descripción
OBS	Se admite	No se admite	OBS es un sistema de almacenamiento, no un sistema de archivos. En el notebook de versión antigua, la replicación remota y la replicación local de datos OBS pueden confundirse, lo que lleva a problemas en el control de las operaciones en los datos. Por lo tanto, el montaje de OBS se quita del notebook de la nueva versión. Puede obtener y operar datos de OBS de forma flexible utilizando código.
Sistema de archivos paralelo de OBS	No se admite	Se admite	El notebook de nueva versión permite el montaje dinámico de sistemas de archivos paralelos de OBS. Puede montar el almacenamiento en la página de detalles de una instancia de notebook en ejecución. La migración de datos de la versión anterior a la nueva no está involucrada.
EVS	Se admite	Se admite	Los discos de EVS se pueden conectar a instancias de portátiles de las versiones antiguas y nuevas. Los datos almacenados en la versión anterior deben migrarse a la nueva versión.
SFS	No se admite	Se admite	SFS se utiliza en grupos de recursos dedicados. Esta función ha sido descontinuada en el notebook de la versión anterior. Por lo tanto, la migración de datos no está involucrada.
EFS	No se admite	Se admite	EFS solo se utiliza en el notebook de la nueva versión.

OBS utilizado en notebook de la versión antigua

Cuando las instancias de notebook de la versión anterior usan OBS para almacenamiento, los datos se almacenan en OBS y no es necesario migrarlos. Después de crear una instancia de

notebook de nueva versión, utilice directamente los datos del directorio de OBS. Para obtener más información, consulte **¿Cómo leo y escribo archivos de OBS en una instancia de notebook?**

Figura 4-17 OBS utilizado en notebook de la versión antigua



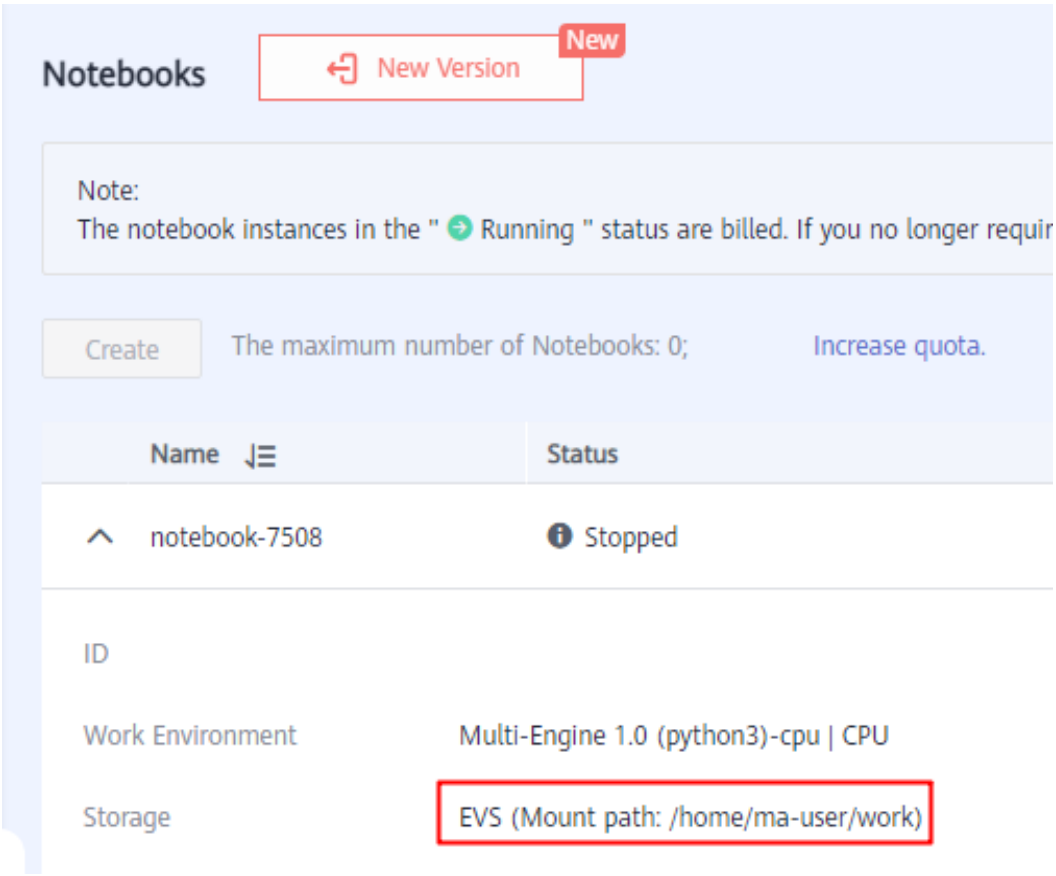
EVS utilizado en notebook de la versión antigua

Si los discos de EVS están conectados a una instancia de notebook de la versión antigua para almacenar datos, realice una copia de respaldo y migre los datos de EVS a una instancia de notebook de la nueva versión.

- Si el volumen de datos almacenados en EVS es pequeño, descargue los datos a un directorio local, cree una instancia de notebook de la nueva versión y cargue los datos a la nueva instancia de notebook.
- Si se almacena una gran cantidad de datos en EVS, cargue los datos a un bucket de OBS. Después de crear una instancia de notebook de la nueva versión, lea los datos del bucket de OBS.

Para obtener más información, consulte **Carga y descarga de datos de Notebook**.

Figura 4-18 Almacenamiento de EVS utilizado en el notebook de la versión antigua



4.9.10 ¿Cómo uso los conjuntos de datos creados en ModelArts en una instancia de notebook?

Los conjuntos de datos creados en ModelArts se almacenan en OBS. Para utilizar estos conjuntos de datos en una instancia de notebook, descárguelos de OBS a la instancia de notebook.

Para más detalles, véase **¿Cómo cargo un archivo desde una instancia de Notebook a OBS o descargo un archivo desde OBS a una instancia de Notebook?**

4.9.11 pip y comandos comunes

pip es una herramienta común de gestión de paquetes de Python. Le permite buscar, descargar, instalar y desinstalar paquetes de Python.

Comandos comunes de pip:

```
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.
```

Para otros comandos, ejecute el comando **pip --help**.

4.9.12 ¿Cuáles son los tamaños de los directorios /cache para diferentes especificaciones de notebook de DevEnviron?

Al crear una instancia de notebook, puede seleccionar CPU, GPU o Ascend según el volumen de datos.

ModelArts monta los discos en **/cache**. Puede utilizar este directorio para almacenar archivos temporales. El directorio **/cache** comparte recursos con el directorio de código. El tamaño del directorio varía según las especificaciones de los recursos.

No se pueden montar discos en **/cache** para CPU. Cuando solo se utiliza una GPU o una tarjeta de Ascend, el tamaño del directorio **/cache** está limitado a 500 GB. Si se utilizan varias GPU o tarjetas de Ascend, el tamaño del directorio **/cache** se limita a 3 TB y se calcula mediante la siguiente fórmula: Tamaño del directorio **/cache** = Número de tarjetas x 500 GB. Para obtener más información, véase **Tabla 4-2**.

Tabla 4-2 tamaños de directorio /cache para diferentes especificaciones de notebook

Especificación	Tamaño del directorio /cache
GPU, 0.25 tarjetas	500 GB x 0.25
GPU, 0.5 tarjetas	500 GB x 0.5
GPU, 1 tarjeta	500 GB
GPU, tarjetas duales	500 GB x 2
GPU, cuatro tarjetas	500 GB x 4
GPU, ocho tarjetas	3 TB
Ascend, tarjeta única	500 GB
Ascend, tarjetas dobles	500 GB x 2
Ascend, cuatro tarjetas	500 GB x 4
Ascend, ocho tarjetas	3 TB
CPU	N/A

5 Trabajos de entrenamiento

5.1 Consultoría funcional

5.1.1 ¿Cuáles son los requisitos de formato para los algoritmos importados desde un entorno local?

ModelArts admite la importación de algoritmos desarrollados localmente. Los requisitos de formato son los siguientes:

- Se admite cualquier lenguaje de programación.
- El archivo de arranque debe tener el formato **.py** o **.pyc**.
- El número de archivos (incluidos los archivos y las carpetas) no puede ser superior a 1,024.
- El tamaño total del archivo no puede superar los 5 GB.

5.1.2 ¿Cuáles son las soluciones para el underfitting?

1. Aumento de la complejidad del modelo
   - Para un algoritmo, agregue más elementos de orden alto al modelo de regresión, mejore la profundidad del árbol de decisiones o aumente el número de capas ocultas y unidades ocultas de la red neuronal para aumentar la complejidad del modelo.
   - Descartar el algoritmo original y utilizar un algoritmo o modelo más complejo. Por ejemplo, use la red neuronal para reemplazar la regresión lineal y use el bosque aleatorio para reemplazar el árbol de decisiones.
2. Adición de más características para hacer que los datos de entrada sean más expresivos
   - La minería de características es muy importante. Específicamente, las características con capacidades de expresión fuertes pueden superar a un gran número de características con capacidades de expresión débiles.
   - La calidad de las características es el enfoque.
   - Para explorar características con capacidades de expresión sólidas, debe tener una comprensión profunda de los escenarios de datos y aplicaciones, que depende de la experiencia.

3. Ajuste de parámetros e hiperparámetros
   - Red neuronal: tasa de aprendizaje, tasa de atenuación de aprendizaje, número de capas ocultas, número de unidades en una capa oculta, parámetros β1 y β2 en el algoritmo de optimización de Adam y batch_size
   - Otros algoritmos: número de árboles en el bosque aleatorio, número de clústeres en los medios de *k* y parámetro de regularización λ
4. Adición de datos de entrenamiento (no recomendado)

   Por lo general, el underfitting es causado por las capacidades débiles de aprendizaje de modelos. La adición de datos no puede aumentar significativamente el efecto del entrenamiento.
5. Reducción de las restricciones de regularización

   La regularización tiene como objetivo prevenir el overfitting del modelo. Si un modelo es underfitting en lugar de overfitting, reduzca el parámetro de regularización **λ** o quite directamente el elemento de regularización.

5.1.3 ¿Cuáles son las precauciones para cambiar los trabajos de entrenamiento de la versión antigua a la nueva?

Las diferencias entre la nueva versión y la versión anterior radican en:

- **Diferencias en la creación de trabajo de entrenamiento**
- **Diferencias en la adaptación del código de entrenamiento**
- **Diferencias en motores de entrenamiento incorporados**

Diferencias en la creación de trabajo de entrenamiento

En la versión anterior, puede usar algoritmos personalizados, marcos comunes e imágenes personalizadas para crear trabajos de entrenamiento.

En la nueva versión, puede utilizar los algoritmos personalizados para crear trabajos de entrenamiento.

En la nueva versión, los algoritmos se pueden seleccionar por categoría al crear un trabajo de entrenamiento. Esto no afecta a los puestos de entrenamiento existentes.

Si utiliza algoritmos personalizados o marcos comunes para crear trabajos de entrenamiento en la versión anterior, puede utilizar un script personalizado para hacerlo en la nueva versión.

Diferencias en la adaptación del código de entrenamiento

En la versión anterior, se requiere que configure la entrada y salida de datos de la siguiente manera:

```
# 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)
```

En la nueva versión, solo necesita configurar las entradas y salidas de entrenamiento. En el código, se utilizan **arg.data_url** y **arg.train_url** como rutas locales.

```
# 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)
```

Diferencias en motores de entrenamiento incorporados

- En la nueva versión, se instala MoXing 2.0.0 o posterior de forma predeterminada para los motores de entrenamiento integrados.
- En la nueva versión, Python 3.7 o posterior se utiliza para los motores de entrenamiento incorporados.
- En la nueva imagen, el directorio principal predeterminado se ha cambiado de **/home/work** a **/home/ma-user**. Comprueba si el código de entrenamiento contiene una codificación dura de **/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.

  Para utilizar un motor de entrenamiento en la versión anterior, cambie a la versión anterior. Tabla 5-1 enumera las diferencias entre los motores de entrenamiento incorporados en las versiones antiguas y nuevas.

Tabla 5-1 Diferencias entre los motores de entrenamiento incorporados en las versiones antiguas y nuevas

Entorno de tiempo de ejecución	Motor y versión de entrenamiento incorporados	Versión anterior	Versión nueva
TensorFlow	TensorFlow-1.8.0	√	x
	TensorFlow-1.13.1	√	Próximamente
	TensorFlow-2.1.0	√	√
MXNet	MXNet-1.2.1	√	x
Caffe	Caffe-1.0.0	√	x

Entorno de tiempo de ejecución	Motor y versión de entrenamiento incorporados	Versión anterior	Versión nueva
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

5.1.4 ¿Cómo obtengo un modelo de ModelArts entrenado?

Los modelos generados con ExeML de ModelArts solo se pueden desplegar en ModelArts y no se pueden descargar a su PC local.

Los modelos entrenados mediante un algoritmo personalizado o de suscripción se almacenan en rutas de OBS especificadas para que los descargue.

5.1.5 ¿Deben ser categóricos los hiperparámetros optimizados usando un algoritmo de TPE?

Los algoritmos de TPE no imponen requisitos sobre los tipos de hiperparámetros optimizados. Sin embargo, para reducir la utilización de recursos para usuarios comunes, la consola de ModelArts requiere que los hiperparámetros de TPE sean de tipo flotante. Para usar parámetros discretos y continuos, invoque a la API de REST.

5.1.6 ¿Para qué se utiliza TensorBoard en los trabajos de visualización de modelos?

Los trabajos de visualización son impulsados por TensorBoard. Para obtener más información acerca de las funciones de TensorBoard, consulte el **sitio web oficial de TensorBoard** .

5.1.7 ¿Cómo obtengo RANK_TABLE_FILE en ModelArts para el entrenamiento distribuido?

ModelArts proporciona automáticamente el archivo **RANK_TABLE_FILE** para usted. Obtenga la ubicación del archivo con variables de entorno.

- Abra el terminal del notebook y ejecute el siguiente comando para ver **RANK_TABLE_FILE**:
  ```
  env | grep RANK
  ```
- En un trabajo de entrenamiento, agregue el siguiente código a la primera línea del script de inicio de entrenamiento para imprimir el valor de **RANK_TABLE_FILE**:
  ```
  os.system('env | grep RANK')
  ```

5.1.8 ¿Cómo obtengo las versiones CUDA y cuDNN de una imagen personalizada?

Obtenga una versión de CUDA:

```
cat /usr/local/cuda/version.txt
```

Obtenga una versión de cuDNN:

```
cat /usr/local/cuda/include/cudnn.h | grep CUDNN_MAJOR -A 2
```

5.1.9 ¿Cómo obtengo un archivo de instalación de MoXing?

Los usuarios no pueden descargar ni instalar archivos de instalación de MoXing. El paquete de instalación de MoXing está preestablecido en las imágenes de trabajo de entrenamiento y notebook de ModelArts y se puede utilizar directamente.

5.1.10 En un entrenamiento con multinodo, el nodo de PS TensorFlow que funciona como un servidor se suspenderá continuamente. ¿Cómo determina ModelArts si el entrenamiento está completo? ¿Qué nodo es un trabajador?

En un entrenamiento distribuido impulsado por TensorFlow, la tarea PS y la tarea de trabajo se inician. La tarea del trabajador es una tarea clave. ModelArts utilizará un código de salida de proceso de la tarea de trabajador para determinar si el trabajo de entrenamiento está completo.

Se usará un nombre de tarea para determinar qué nodo es un trabajador. Un trabajo de Volcano se emite para el entrenamiento, que contiene una tarea de PS y una tarea de trabajador. Los comandos de inicio de las dos tareas son diferentes. El hiperparámetro **task_name** se generará automáticamente, que es **ps** para la tarea PS y **worker** para la tarea de trabajo.

5.1.11 ¿Cómo instalo MoXing para una imagen personalizada?

Para evitar que la instalación automática de MoXing afecte al entorno del paquete en la imagen personalizada, instale manualmente MoXing para la imagen personalizada. MoXing se almacena en el directorio **/home/ma-user/modelarts/package/** después de iniciar el trabajo. Antes de usar MoXing ejecute el siguiente código para instalarlo:


```
import os
os.system("pip install /home/ma-user/modelarts/package/moxing_framework-*.whl")
```

5.2 Lectura de datos durante el entrenamiento

5.2.1 ¿Cómo configuro los datos de entrada y salida para los modelos de entrenamiento de ModelArts?

ModelArts le permite cargar un algoritmo personalizado para crear trabajos de entrenamiento. Crear el algoritmo y subirlo a un bucket de OBS. Para obtener más información sobre cómo crear un algoritmo, consulte Creación de un algoritmo. Para obtener más información sobre cómo crear un trabajo de entrenamiento, consulte Creación de un trabajo de entrenamiento.

Análisis de rutas de entrada y de salida

Cuando un modelo de ModelArts lee datos almacenados en OBS o envía datos a una ruta de OBS especificada, realice las siguientes operaciones para configurar los datos de entrada y de salida:

1. Analice las rutas de entrada y de salida en el código de entrenamiento. Se recomienda el siguiente método:

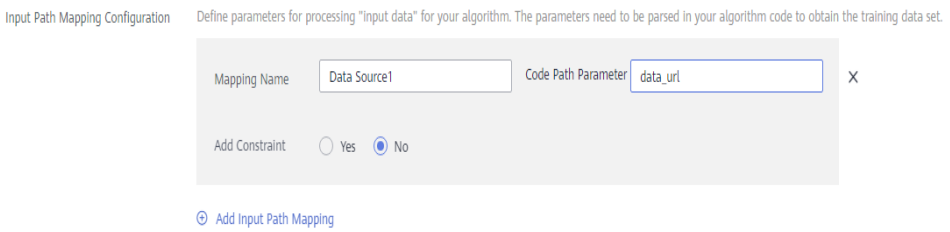
   ```
   import argparse
   # Create a parsing task.
   parser = argparse.ArgumentParser(description="train mnist",

   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()
   ```

   Después de analizar los parámetros, use **data_url** y **train_url** para reemplazar las rutas a la fuente de datos y la salida de datos, respectivamente.

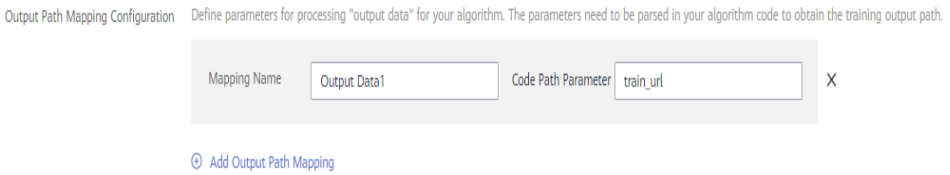
2. Cuando utilice una imagen preestablecida para crear un algoritmo personalizado, configure los parámetros de entrada y salida en la página **Create Algorithm** según la configuración del código.
   - Los datos de entrenamiento son una necesidad para el despliegue de algoritmos. De forma predeterminada, los datos de entrada son **Data Source** y el parámetro de ruta de código es **data_url** (personalizable).

     Figura 5-1 Análisis del parámetro de ruta de entrada **data_url**

   - Después de completar el entrenamiento del modelo, el modelo entrenado y la información de salida deben almacenarse en una ruta de OBS. Por defecto, el dato de salida es **Output Data** y el parámetro de ruta de código es **train_url** (personalizable).

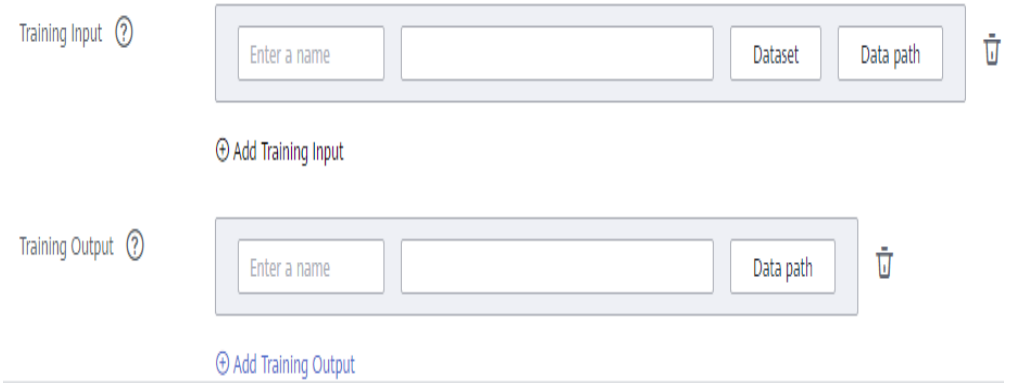
Figura 5-2 Análisis del parámetro de ruta de salida **train_url**



3. Al crear un trabajo de entrenamiento, establezca las rutas de entrada y salida.

   Seleccione la ruta de OBS o la ruta del conjunto de datos como la entrada de entrenamiento y la ruta de OBS como la salida.

Figura 5-3 Ajuste de entrada y salida de entrenamiento



5.2.2 ¿Cómo mejoro la eficiencia del entrenamiento reduciendo la interacción con OBS?

Descripción del escenario

Cuando se utiliza ModelArts para el entrenamiento de aprendizaje profundo personalizado, los datos de entrenamiento generalmente se almacenan en OBS. Si el volumen de datos de entrenamiento es grande (por ejemplo, más de 200 GB), se requiere un grupo de recursos de GPU para el entrenamiento cada vez, lo que resulta en una baja eficiencia del entrenamiento.

Para mejorar la eficiencia del entrenamiento y reducir la interacción con OBS, realice las siguientes operaciones para su optimización.

Principios de optimización

Para el grupo de recursos de GPU proporcionado por ModelArts se conectan SSD NVMe de 500 GB a cada nodo de entrenamiento de forma gratuita. Los SSD están conectados al directorio **/cache**. El ciclo de vida de los datos en el directorio **/cache** es el mismo que el de un trabajo de entrenamiento. Una vez completado el trabajo de entrenamiento, se borra todo el contenido del directorio **/cache** para liberar espacio para el siguiente trabajo de entrenamiento. Por lo tanto, puede copiar datos de OBS al directorio **/cache** durante el entrenamiento para que los datos se puedan leer desde el directorio **/cache** cada vez hasta que se complete el entrenamiento. Una vez completada el entrenamiento, el contenido del directorio **/cache** se borrará automáticamente.

Métodos de optimización

El código de TensorFlow se utiliza como ejemplo.

El siguiente es código antes de la optimización:

```
...
tf.flags.DEFINE_string('data_url', '', 'dataset directory.')
FLAGS = tf.flags.FLAGS
mnist = input_data.read_data_sets(FLAGS.data_url, one_hot=True)
```

El siguiente es un ejemplo del código optimizado. Los datos se copian en el directorio **/cache**.

```
...
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)
```

5.2.3 ¿Por qué la eficiencia de lectura de datos es baja cuando se leen un gran número de archivos de datos durante el entrenamiento?

Si un conjunto de datos contiene un gran número de archivos de datos (pequeños archivos masivos) y los datos se almacenan en OBS, los archivos deben leerse repetidamente desde OBS durante el entrenamiento. Como resultado, el proceso de entrenamiento está a la espera de la lectura de archivos, lo que resulta en una baja eficiencia de lectura.

Solución

1. Comprima los archivos pequeños masivos en un paquete en su PC local, por ejemplo, un paquete .zip.
2. Sube el paquete a OBS.
3. Durante el entrenamiento, descargue directamente este paquete desde OBS al directorio **/cache** de su PC local. Realice esta operación solo una vez.

   Por ejemplo, puede usar mox.file.copy_parallel para descargar el paquete .zip en el directorio **/cache**, descomprimir el paquete y, a continuación, leer los archivos para el entrenamiento.

   ```
   ...
   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)
   ```

5.3 Compilación del código de entrenamiento

5.3.1 ¿Cómo creo un trabajo de entrenamiento cuando el modelo que se va a entrenar hace referencia a un paquete de dependencia?

Almacene el archivo **pip-requirements.txt** en el directorio de código de entrenamiento.

> **NOTA**
>
> Se puede utilizar cualquiera de los siguientes nombres de archivo. Esta sección utiliza **pip-requirements.txt** como ejemplo.
>
> - pip-requirement.txt
> - pip-requirements.txt
> - requirement.txt
> - requirements.txt

Antes de ejecutar el archivo de arranque de entrenamiento, el sistema ejecuta automáticamente el siguiente comando para instalar los paquetes de Python especificados:

```
pip install -r pip-requirements.txt
```

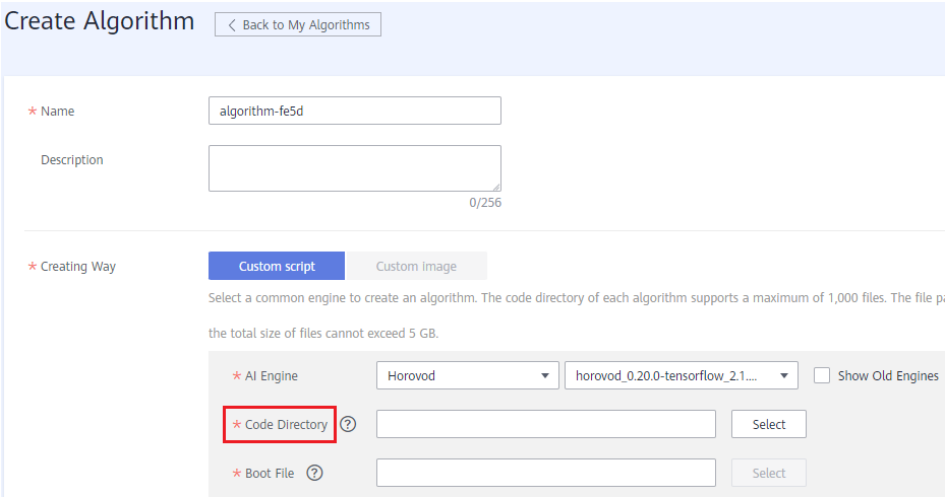
- Para obtener más información sobre el directorio de código, consulte **Almacenamiento del archivo de instalación en el directorio de código**.
- Para obtener más información acerca de las especificaciones de **pip-requirements.txt**, consulte **Especificaciones del archivo de instalación**.

Almacenamiento del archivo de instalación en el directorio de código

ModelArts le permite instalar paquetes de dependencias de terceros durante el entrenamiento de modelos de cualquiera de las siguientes maneras:

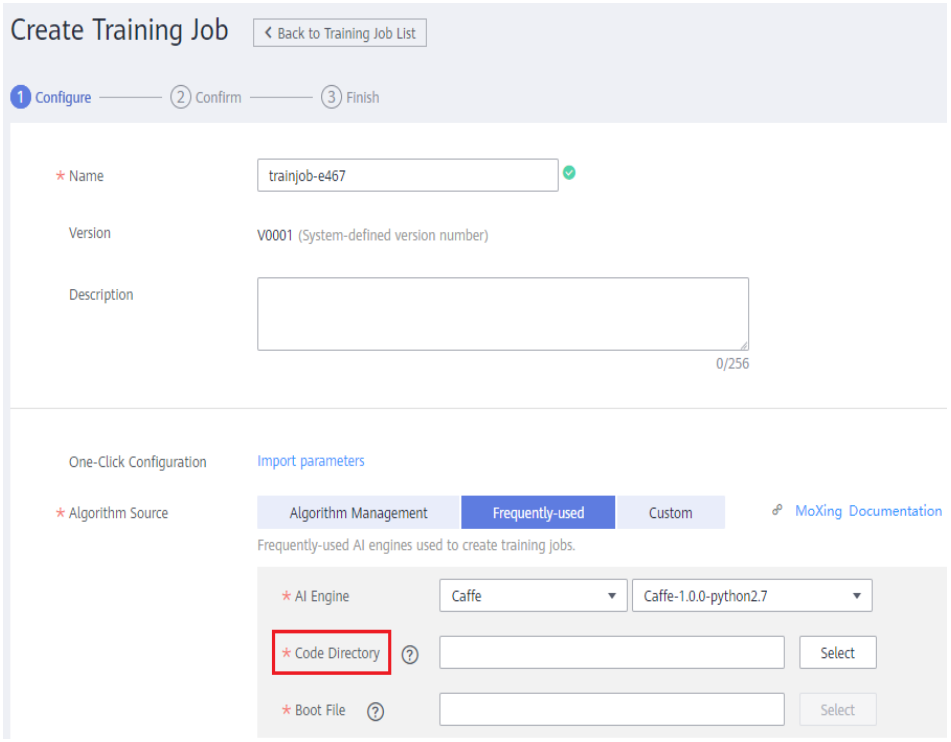
- Método 1 (recomendado): Antes de crear un algoritmo, **almacene los archivos necesarios o paquetes de instalación** en el directorio de código.

  Figura 5-4 Creación de un algoritmo

- Método 2: Antes de utilizar un marco común para crear un trabajo de entrenamiento, **almacene los archivos necesarios o los paquetes de instalación** en el directorio de código. (Esta función dejará de estar disponible en breve.)

Figura 5-5 Usar un framework común para crear un algoritmo



Especificaciones del archivo de instalación

El archivo de instalación varía según el tipo de paquete de dependencia.

- **Paquetes de instalación de código abierto**

  NOTA

  No se admite la instalación con el código fuente de GitHub.

  Cree un archivo llamado **pip-requirements.txt** en el directorio de código y especifique el nombre y el número de versión del paquete de dependencias en el archivo. El formato es *[Package name]==[Version]*.

  Tomemos, por ejemplo, una ruta de OBS especificada por **Code Dir** que contiene archivos de modelo y el archivo **pip-requirements.txt**. La estructura de directorios de código sería la siguiente:

  ```
  |---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
  ```

  A continuación se muestra el contenido del archivo **pip-requirements.txt**:

  ```
  alembic==0.8.6
  bleach==1.4.3
  click==6.6
  ```

- **Paquetes de WHL**

  Si los antecedentes de entrenamiento no admiten la descarga de paquetes de instalación de código abierto o el uso de paquetes de WHL compilados por el usuario, el sistema no puede descargar e instalar automáticamente el paquete. En este caso, coloque el paquete de WHL en el directorio de código, cree un archivo llamado **pip-requirements.txt** y especifique el nombre del paquete WHL en el archivo. El paquete de dependencias debe ser un archivo **.whl**.

Tomemos, por ejemplo, una ruta de acceso de OBS especificada por **Code Dir** que contiene archivos de modelo, el archivo **.whl** y el archivo **pip-requirements.txt**. La estructura de directorios de código sería la siguiente:

```
|---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
```

A continuación se muestra el contenido del archivo **pip-requirements.txt**:

```
numpy-1.15.4-cp36-cp36m-manylinux1_x86_64.whl
tensorflow-1.8.0-cp36-cp36m-manylinux1_x86_64.whl
```

5.3.2 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**.

5.3.3 ¿Cómo instalo una biblioteca de la que depende C++?

Se puede usar una biblioteca de terceros durante el entrenamiento laboral. A continuación se utiliza C++ como ejemplo para describir cómo instalar una biblioteca de terceros.

1. Descargue el código fuente a un PC local y súbalo a OBS. Para obtener más información sobre cómo cargar un archivo mediante OBS Browser, consulte **Carga de un archivo**.
2. Utilice MoXing para copiar el código fuente cargado en OBS a una instancia de notebook en el entorno de desarrollo.

   A continuación se muestra un ejemplo de código para copiar datos a una instancia de notebook en un entorno de desarrollo que se ejecuta en un 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. En la ficha **Files** de la página **Jupyter**, haga clic en **New** y seleccione **Terminal**. Ejecute el siguiente comando para ir a la ruta de destino y compruebe si el código fuente se ha descargado, es decir, si el archivo **data** existe.

   ```
   cd /home/ma-user/work
   ls
   ```
4. Compile código de **Terminal** en función de los requisitos de servicio.
5. Utilice MoXing para copiar los resultados de la compilación en OBS. El siguiente es un ejemplo de código.

   ```
   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. Durante el entrenamiento, utilice MoXing para copiar el resultado de la compilación de OBS al contenedor. El siguiente es un ejemplo de código.

   ```
   import moxing as mox
   mox.file.make_dirs('/cache/data')
   mox.file.copy_parallel('obs://bucket-name/data', '/cache/data')
   ```

5.3.4 ¿Cómo puedo comprobar si una copia de carpeta está completa durante el entrenamiento laboral?

En la secuencia de comandos para el archivo de arranque de trabajo de entrenamiento, ejecute los siguientes comandos para obtener el tamaño de las carpetas copiadas y las carpetas que se

van a copiar. A continuación, determine si la copia de carpeta está completa en función de la salida del comando.

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

get_size indica el tamaño del archivo o carpeta que se va a obtener. **recursive=True** indica que el tipo es carpeta. **True** indica que el tipo es carpeta y **False** indica que el tipo es archivo.

Si el resultado del comando es consistente, la copia de la carpeta está completa. Si el resultado del comando es inconsistente, la copia de la carpeta no está completa.

5.3.5 ¿Cómo cargo algunos parámetros bien entrenados durante el entrenamiento laboral?

Durante el entrenamiento de trabajo, algunos parámetros necesitan ser cargados desde un modelo pre-entrenado para inicializar el modelo actual. Puede utilizar los siguientes métodos para cargar los parámetros:

1. Consulte todos los parámetros mediante el siguiente código.
   ```
   from moxing.tensorflow.utils.hyper_param_flags import mox_flags
   print(mox_flags.get_help())
   ```
2. Especifique los parámetros que se van a restaurar durante la carga del modelo. **checkpoint_include_patterns** es el parámetro que necesita ser restaurado, y **checkpoint_exclude_patterns** es el parámetro que no necesita ser restaurado.
   ```
   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. Especifique una lista de parámetros que se van a entrenar. **trainable_include_patterns** es una lista de parámetros que necesitan ser entrenados, y **trainable_exclude_patterns** es una lista de parámetros que no necesitan ser entrenados.
   ```
   --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.
   ```

5.3.6 ¿Cómo obtengo los parámetros del trabajo de entrenamiento del archivo de arranque del trabajo de entrenamiento?

Los parámetros del trabajo de entrenamiento se pueden generar automáticamente en segundo plano o se pueden introducir manualmente. Para obtener los parámetros del trabajo de entrenamiento:

1. Cuando se crea un trabajo de entrenamiento, **train_url** en los parámetros de ejecución del trabajo de entrenamiento indica dónde se emiten los resultados de entrenamiento, y **data_url** indica un origen de datos. El parámetro **test** se introduce manualmente.

Figura 5-6 Creación de un trabajo de entrenamiento de la nueva versión

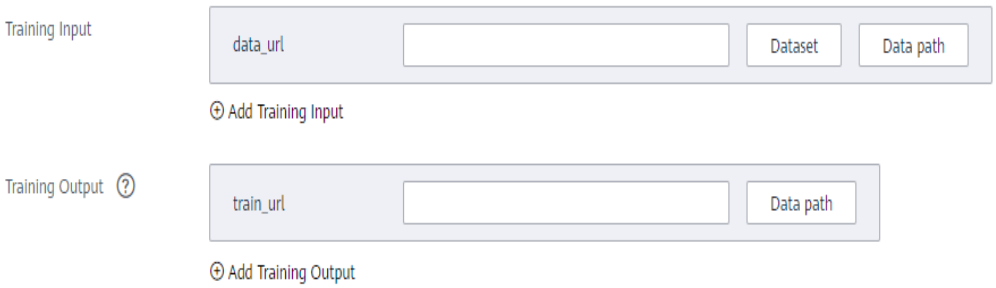
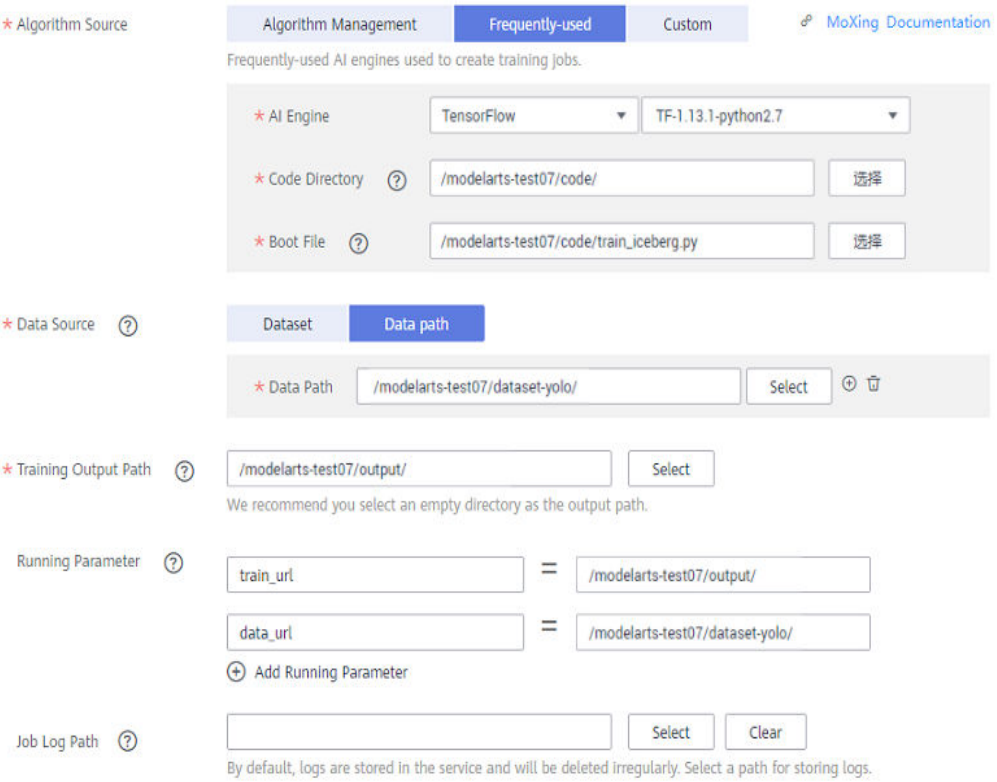


Figura 5-7 Creación de un trabajo de entrenamiento de la versión anterior



2. Después de ejecutar el trabajo de entrenamiento, puede hacer clic en el nombre del trabajo en la lista de trabajos de entrenamiento para ver sus detalles. Puede obtener el modo de entrada de parámetros de los logs, como se muestra en **Figura 5-8**.

Figura 5-8 Consulta de 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. Para obtener los valores de **train_url**, **data_url** y **test** durante el entrenamiento, agregue el siguiente código al archivo de arranque del trabajo de entrenamiento:


```
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')
```

5.3.7 ¿Por qué no puedo usar os.system ('cd xxx') para acceder a la carpeta correspondiente durante el entrenamiento laboral?

Si no puede acceder a la carpeta correspondiente usando **os.system('cd xxx')** en el script de arranque del trabajo de entrenamiento, se recomienda utilizar el siguiente método:

```
import os
os.chdir('/home/work/user-job-dir/xxx')
```

5.3.8 ¿Cómo invoco un script de Shell en un trabajo de entrenamiento para ejecutar el archivo .sh?

ModelArts le permite invocar un script de shell, y puede usar Python para invocar a **.sh**. El procedimiento es el siguiente:

1. Suba el script **.sh** a un bucket de OBS. Por ejemplo, cargue la secuencia de comandos **.sh** en **/bucket-name/code/test.sh**.
2. Cree el archivo **.py** en un PC local, por ejemplo, **test.py**. El fondo descarga automáticamente el directorio de código en el directorio **/home/work/user-job-dir/** del contenedor. Por lo tanto, puede invocar el archivo **.sh** en el archivo de arranque **test.py** de la siguiente manera:
   ```
   import os
   os.system('bash /home/work/user-job-dir/code/test.sh')
   ```
3. Suba **test.py** a OBS. A continuación, la ruta de almacenamiento de archivos es **/bucket-name/code/test.py**.
4. Cuando cree un trabajo de entrenamiento, establezca el directorio de código en **/bucket-name/code/** y el directorio de archivo de arranque en **/bucket-name/code/test.py**.

Después de crear el trabajo de entrenamiento, puede usar Python para invocar el archivo **.sh**.

5.3.9 ¿Cómo obtengo la ruta para almacenar el archivo de dependencia en el código de entrenamiento?

El código desarrollado localmente debe cargarse en el backend de ModelArts. En el código de entrenamiento, es propenso a errores establecer la ruta para almacenar el archivo de dependencia.

Se recomienda la siguiente solución general: Utilice la API del SO para obtener la ruta absoluta del archivo de dependencia.

Por ejemplo:

```
|---project_root                # Root directory for code
     |---bootfile.py            # Boot file
   |---otherfileDirectory       # Directory of other dependency files
     |---otherfile.py           # Other dependency files
```

Haga lo siguiente para obtener la ruta del archivo de dependencia **otherfile_path** en este ejemplo, en el archivo de arranque:

```
import os
current_path = os.path.dirname(os.path.realpath(__file__)) # Directory where the
```



```
boot file is located
project_root = os.path.dirname(current_path) # Root directory of the project,
which is the code directory set on the ModelArts training console
otherfile_path = os.path.join(project_root, "otherfileDirectory", "otherfile.py")
```

5.3.10 ¿Cuál es la ruta de acceso del archivo si se hace referencia a un archivo del directorio modelo en un paquete personalizado de Python?

Para obtener la ruta real de un archivo en un contenedor utilice Python.

```
os.getcwd() # Obtain the current work directory (absolute path) of the file.
os.path.realpath(__ file __) # Obtain the absolute path of the file.
```

También puede utilizar otros métodos para obtener una ruta de archivo a través del motor de búsqueda y utilizar la ruta obtenida para leer y escribir el archivo.

5.4 Creación de un trabajo de entrenamiento

5.4.1 ¿Qué puedo hacer si se muestra el mensaje "Object directory size/quantity exceeds the limit" al crear un trabajo de entrenamiento?

Análisis de problemas

El directorio de código para crear un trabajo de entrenamiento tiene límites en el tamaño y el número de archivos.

Solución

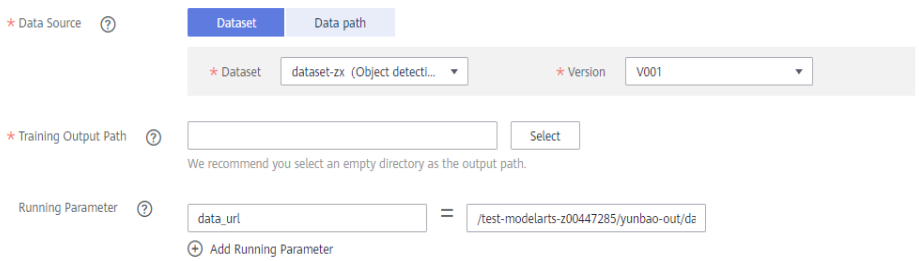
Elimine los archivos excepto el código del directorio de código o guarde los archivos en otros directorios. Asegúrese de que el tamaño del directorio de código no supere los 128 MB y que el número de archivos no supere los 4,096.

5.4.2 ¿Cuáles son las precauciones para establecer parámetros de entrenamiento?

Preste atención a lo siguiente al establecer los parámetros de entrenamiento:

- Si se han configurado el origen del algoritmo y el origen de datos, se establece el parámetro **data_url** automáticamente en función del objeto seleccionado y no se puede modificar directamente cambiando los parámetros de ejecución.

  Figura 5-9 Ajuste automático de los parámetros de ejecución



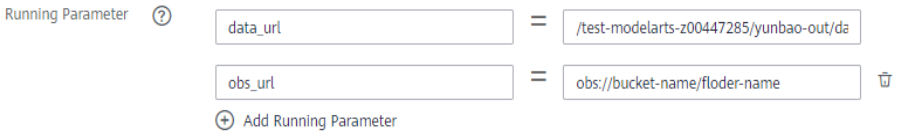
- Al establecer los parámetros de ejecución para crear un trabajo de entrenamiento, solo es necesario establecer los nombres y valores de los parámetros correspondientes. Consulte **Figura 5-10**.

  Figura 5-10 Configuración de los parámetros de operación

- Si un valor de parámetro es una ruta de bucket de OBS, utilice la ruta (comenzando con **obs:**//) a los datos. Véase **Figura 5-11**.

  Figura 5-11 Configuración de una ruta de OBS

- Cuando cree una carpeta de OBS en código, invoque a una API de MoXing de la siguiente manera:

```
import moxing as mox
mox.file.make_dirs('obs://bucket_name/sub_dir_0/sub_dir_1')
```

5.4.3 ¿Cuáles son los tamaños de los directorios /cache para diferentes especificaciones de recursos en el entorno de entrenamiento?

Al crear un trabajo de entrenamiento, puede seleccionar recursos de CPU, GPU o Ascend según el tamaño del trabajo de entrenamiento.

ModelArts monta un disco en **/cache**. Puede utilizar este directorio para almacenar archivos temporales. El directorio **/cache** comparte recursos con el directorio de código. El directorio tiene diferentes capacidades para diferentes especificaciones de recursos.

- Recursos de GPU

  Tabla 5-2 Capacidades de los directorios de caché para recursos de GPU

Especificaciones de la GPU	Capacidad de directorios de cache
V100	800 GB
8*V100	3 TB
P100	800 GB

- Recursos de CPU

Tabla 5-3 Capacidades de los directorios de cache para recursos de CPU

Especificaciones de la CPU	Capacidad de directorios de cache
2 vCPUs | 8 GiB	50 GB
8 vCPUs | 32 GiB	50 GB

- Recursos de Ascend

Tabla 5-4 Capacidades de los directorios de caché para los recursos de Ascend

Especificaciones de Ascend	Capacidad de directorios de cache
Ascend 910	3 TB

5.4.4 ¿Es seguro el directorio /cache de un trabajo de entrenamiento?

El programa de un trabajo de entrenamiento de ModelArts se ejecuta en un contenedor. La dirección de un directorio en el que está montado el contenedor es única, y solo puede accederse por el contenedor en ejecución. Por lo tanto, el directorio **/cache** del trabajo de entrenamiento es seguro.

5.4.5 ¿Por qué un trabajo de entrenamiento siempre está en cola?

Si el trabajo de entrenamiento está siempre en cola, los recursos seleccionados están limitados en el grupo de recursos y el trabajo debe estar en cola. En este caso, espere recursos. Para acelerar la obtención de recursos, haga lo siguiente:

1. Si utiliza un grupo de recursos públicos:

   Los recursos en un grupo de recursos públicos son limitados. Durante las horas pico, los recursos pueden ser insuficientes si el tráfico de servicio es abundante. Trate de tomar las siguientes medidas:

   – Si se usó una variante libre, cámbiela por una cargada. Se proporcionan pocos recursos para las variantes libres, lo que conduce a una alta probabilidad de cola.
   – El menor número de cartas en la variante seleccionada conduce a la menor probabilidad de cola. Por ejemplo, la probabilidad de cola cuando se selecciona una variante de 1 tarjeta es mucho menor que la de cola cuando se selecciona una variante de 8 tarjetas.
   – Cambie a otra región.
   – Si los recursos se utilizarán durante un largo plazo, compre un grupo de recursos dedicado.
2. Si utiliza un grupo de recursos dedicado:
   – Si hay varios grupos de recursos dedicados disponibles, cambie a uno inactivo.
   – Libere los recursos en el grupo de recursos actual, por ejemplo, detener las instancias de notebook que no se utilizan durante mucho tiempo.

- Envíe un trabajo de entrenamiento durante las horas no pico.
- Póngase en contacto con el administrador de cuentas del grupo de recursos dedicado para ampliar el grupo de recursos en función del uso de recursos.

5.5 Gestión de versiones de trabajos de entrenamiento

5.5.1 ¿Un trabajo de entrenamiento apoya llamadas programadas o periódicas?

Los trabajos de entrenamiento de ModelArts no admiten llamadas programadas o periódicas. Cuando su trabajo está en el estado **Running**, puede invocar al trabajo en función de los requisitos de servicio.

5.6 Consulta de detalles de trabajo

5.6.1 ¿Cómo puedo comprobar el uso de recursos de un trabajo de entrenamiento?

En el panel de navegación izquierdo de la consola de gestión ModelArts, elija **Training Management > Training Jobs** para ir a la página **Training Jobs**. En la lista de trabajos de entrenamiento, haga clic en un nombre de trabajo para ver los detalles del trabajo. Puede ver las siguientes métricas en la página de ficha **Resource Usages**.

- **CPU**: Porcentaje (Percent) de uso de la CPU (cpuUsage)
- **MEM** Porcentaje (Percent) de uso de memoria física (memUsage)
- **GPU**: Porcentaje (Percent) de uso de GPU (gpuUtil)
- **GPU_MEM**: Porcentaje (Percent) de uso de memoria de la GPU (gpuMemUsage)

5.6.2 ¿Cómo accedo a los antecedentes de un trabajo de entrenamiento?

ModelArts no admite el acceso a los antecedentes de un trabajo de entrenamiento.

5.6.3 ¿Hay algún conflicto cuando los modelos de dos trabajos de entrenamiento se guardan en el mismo directorio de un contenedor?

Los directorios de almacenamiento de los trabajos de entrenamiento de ModelArts no se afectan entre sí. Los entornos están aislados entre sí y los datos de otros trabajos no se pueden ver.

5.6.4 Solo se conservan tres dígitos válidos en un log de salida del entrenamiento. ¿Se puede cambiar el valor de loss?

En un trabajo de entrenamiento, solo se conservan tres dígitos válidos en un log de salida de entrenamiento. Cuando el valor de **loss** es demasiado pequeño, el valor se muestra como **0.000**. El contenido del log es el siguiente:

```
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
```

Actualmente, el valor de **loss** no se puede cambiar. Puede multiplicar el valor de **loss** por 1000 para evitar este problema.

5.6.5 ¿Se puede descargar o migrar un modelo entrenado a otra cuenta? ¿Cómo obtengo la ruta de descarga?

Puede descargar el modelo entrenado por un trabajo de entrenamiento y subir el modelo descargado a OBS en la región correspondiente a la cuenta de destino.

Obtención de una ruta de descarga de modelo

1. Inicie sesión en la consola de ModelArts. En el panel de navegación izquierdo, elija **Training Management** > **Training Jobs**. Se muestra la página **Training Jobs**.
2. En la lista de trabajos de entrenamiento, haga clic en un nombre de trabajo para ver los detalles del trabajo.
3. En la página de ficha **Configurations**, obtenga la ruta especificada para **Training Output Path**, es decir, la ruta de descarga del modelo de entrenamiento.

Migración del modelo a otra cuenta

Hay dos formas de migrar un modelo entrenado a otra cuenta:

- Descarga el modelo entrenado y luego súbelo al bucket de OBS en la región correspondiente a la cuenta de destino.
- Configure una política para la carpeta o bucket donde se almacena el modelo para autorizar a otras cuentas a realizar operaciones de lectura y escritura. Para obtener más información, consulte **Configuración de una política de bucket personalizada**.

6 Gestión de modelos

6.1 Importación de modelos

6.1.1 ¿Cómo edito los parámetros de dependencia del paquete de instalación en un archivo de configuración de modelo al importar un modelo?

Síntoma

Al importar un modelo desde OBS o una imagen de contenedor, edite un archivo de configuración de modelo. El archivo de configuración del modelo describe el uso del modelo, el marco informático, la precisión, el paquete de dependencia del código de inferencia y la API del modelo. El archivo de configuración debe estar en formato JSON. **dependencies** en el archivo de configuración especifica el paquete de dependencias necesario para configurar el código de inferencia del modelo. Configure el nombre del paquete, el modo de instalación y las restricciones de versión. Para obtener más información, véase **Parámetros**. En la siguiente sección se describe cómo editar **dependencies** en el archivo de configuración del modelo durante la importación del modelo.

Solución

Los paquetes de instalación deben instalarse en secuencia. Por ejemplo, antes de instalar **mmcv-full**, instale **Cython**, **pytest-runner** y **pytest**. En el archivo de configuración **Cython**, **pytest-runner** y **pytest** están por delante de **mmcv-full**.

Por ejemplo:

```
"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"
        }
      ]
   }
]
```

Si la instalación de **mmcv-full** falla, la posible causa es que GCC no se instaló en la imagen básica, lo que provoca un error de compilación. En este caso, utilice el paquete de ruedas en las instalaciones para instalar **mmcv-full**.

Por ejemplo:

```
"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 en el archivo de configuración del modelo admite múltiples matrices de estructura de dependencias en formato de lista.

Por ejemplo:

```
"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"
        },
      ]
    }
]
```

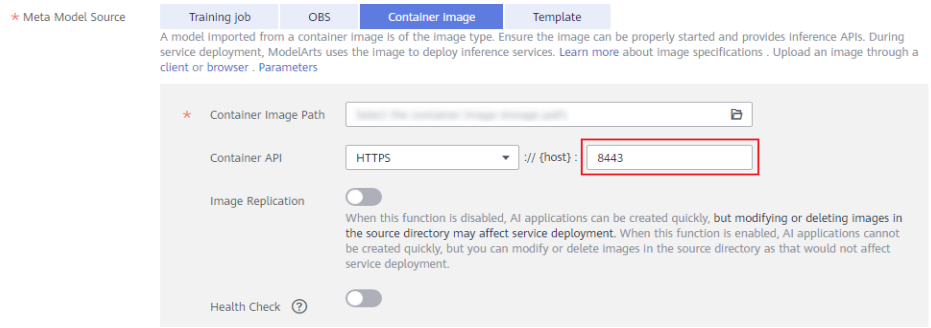
6.1.2 ¿Cómo cambio el puerto predeterminado para crear un servicio en tiempo real usando una imagen personalizada?

Se ha especificado un número de puerto (por ejemplo, 8443) en un archivo de configuración de modelo. Si no especifica un puerto (entonces se usará el puerto 8080 predeterminado) o especifica otro puerto durante la creación de la aplicación de AI, se producirá un error al desplegar la aplicación de AI como servicio. En este caso, establezca el edición de puerto en 8443 en la aplicación de AI para resolver este problema.

Para cambiar el puerto predeterminado, haga lo siguiente:

1. Inicie sesión en la consola de gestión de ModelArts. En el panel de navegación, elija **AI Application Management** > **AI Applications**.
2. Haga clic en **Create**. En la página para crear una aplicación de IA, establezca **Meta Model Source** en **Container image** y seleccione una imagen personalizada.
3. Configure la API de contenedor y el número de puerto. Asegúrese de que el número de puerto es el mismo que el especificado en el archivo de configuración del modelo.

Figura 6-1 Cambio del puerto



4. Después de la configuración, haga clic en **Create now**. Espere hasta que la aplicación de IA se ejecute correctamente.
5. Despliegue de nuevo la aplicación de IA como un servicio en tiempo real.

6.2 ¿Qué hago si se produce una excepción de modelo al desplegar un modelo de imagen personalizado?

Síntoma

Un modelo no se puede desplegar como un servicio en tiempo real. En la página de ficha **Events** de la página de detalles del servicio en tiempo real, se muestra el mensaje "Failed to pull image. Retry later.". Además, no se muestra ninguna información en la página de ficha **Logs**.

Solución

Este problema suele estar relacionado con un modelo que era demasiado grande. Realice los siguientes pasos:

- Simplifique el modelo, vuelva a importarlo y lo despliegue como un servicio en tiempo real.
- Compre un grupo de recursos dedicado y utilícelo para desplegar el modelo como un servicio en tiempo real.

7 Despliegue del servicio

7.1 Consultoría funcional

7.1.1 ¿Qué tipos de servicios se pueden desplegar modelos en ModelArts?

Los modelos se pueden desplegar como servicios en tiempo real o servicios por lotes.

7.1.2 ¿Cuáles son las diferencias entre los servicios en tiempo real y los servicios por lotes?

- Servicios en tiempo real

  Los modelos se despliegan como servicios web. Puede acceder a los servicios con la consola de gestión o las API.

- Servicios por lotes

  Un servicio por lotes realiza la inferencia de los datos por lotes y se detiene automáticamente una vez que se completa el procesamiento de los datos.

Un servicio por lotes procesa datos por lotes a la vez. Un servicio en tiempo real proporciona APIs para que invoque.

7.1.3 ¿Por qué no puedo seleccionar los recursos de Ascend 310?

Los recursos de Ascend 310 son limitados. Si los recursos están agotados, no puede seleccionar recursos de Ascend 310 (en el grupo de recursos público) para inferencia durante despliegue. En la página **Deploy**, el recurso **ARMCPU: 3 vCPUs | 6 GiB Ascend: 1 x Ascend 310** aparecerá atenuado y no se podrá seleccionar.

Soluciones:

- Método 1: Si desea usar Ascend 310 en el grupo de recursos públicos, puede esperar a que otros usuarios liberen los recursos. Si se detienen otros servicios que utilizan los recursos de Ascend 310, puede seleccionar los recursos de despliegue.
- Método 2: Si tiene un grupo de recursos dedicado con recursos de Ascend 310, puede crear un grupo de recursos dedicado de Ascend 310.

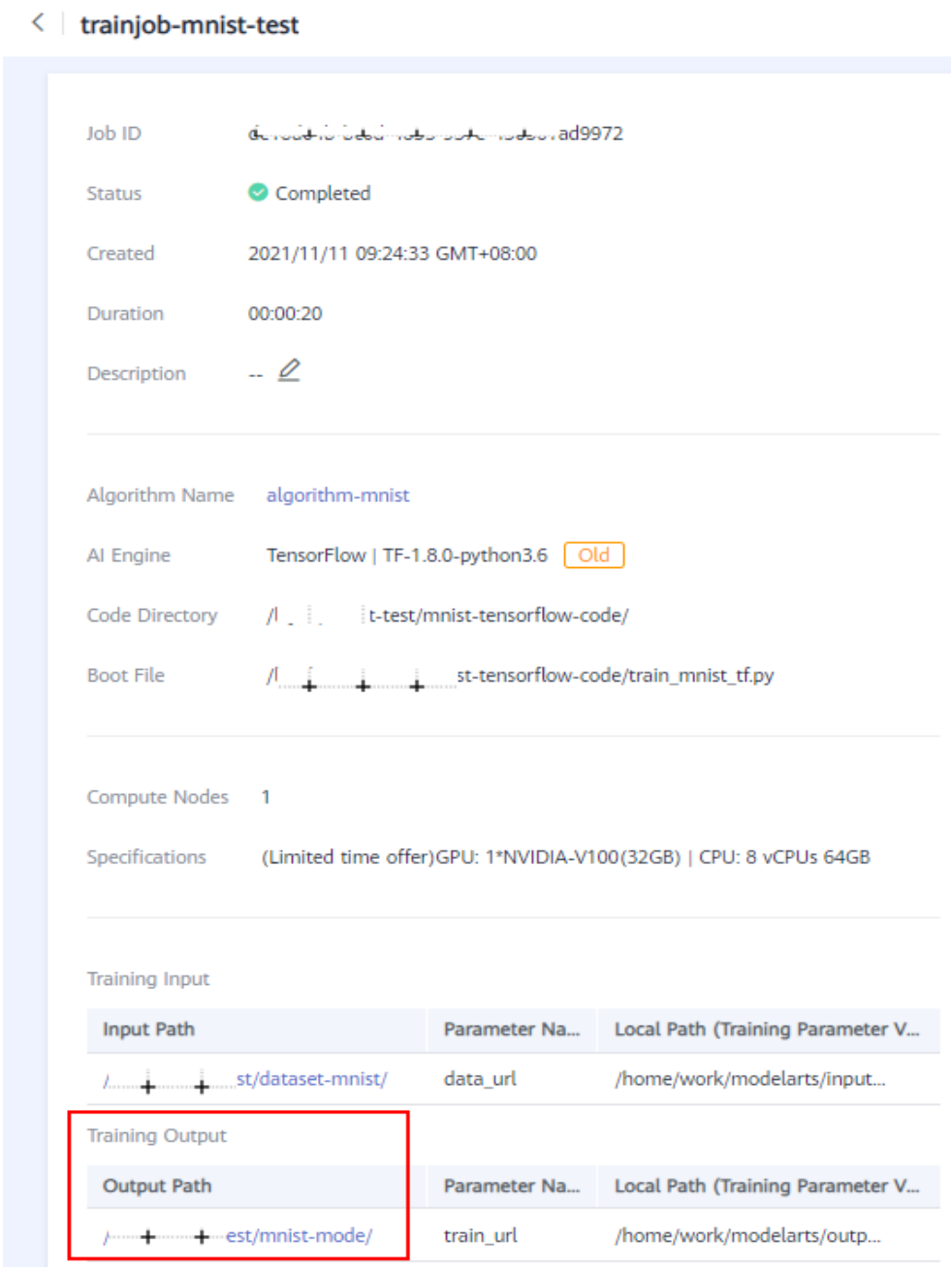
- Método 3: Si los recursos de Ascend 310 en el grupo de recursos dedicado están agotados, puede crear un grupo de recursos dedicado de Ascend 310 después de que otros usuarios hayan eliminado sus instancias de Ascend 310.

7.1.4 ¿Pueden desplegarse localmente los modelos entrenados por ModelArts?

Los modelos entrenados con algoritmos integrados de ModelArts se almacenan en los bucket de OBS y se pueden descargar a un directorio local.

1. En la lista de trabajos de entrenamiento, haga clic en el nombre del trabajo de entrenamiento de destino para ir a su página de detalles, en la que puede obtener la ruta de salida de entrenamiento.

Figura 7-1 Ruta de salida de entrenamiento



2. Haga clic en la ruta para ir a la ruta del objeto de OBS. A continuación, descargue el modelo de OBS.
3. Despliegue el modelo descargado localmente.

   Para obtener más información, consulte **Creación de un modelo local** y **Depuración de un servicio**.

7.1.5 ¿Cuál es el tamaño máximo de un organismo de solicitud de inferencia?

Síntoma

Después de desplegar y ejecutar un servicio, puede enviar una solicitud de inferencia al servicio. El contenido solicitado puede ser imágenes, voz o vídeos, dependiendo del modelo del servicio. Sin embargo, algunas solicitudes fueron interceptadas.

Causa posible

Si utiliza la dirección de solicitud de inferencia (URL de APIG de Huawei Cloud) que se muestra en la pestaña **Usage Guides** de la página de detalles del servicio para la predicción, el tamaño máximo del cuerpo de la solicitud es de 12 MB. Si el cuerpo de la solicitud está sobredimensionado, la solicitud será interceptada.

Si realiza la predicción en la ficha **Prediction** de la página de detalles del servicio, el tamaño del cuerpo de la solicitud no puede superar los 8 MB. El límite de tamaño varía entre las dos fichas porque utilizan enlaces de red diferentes.

Solución

1. Asegúrese de que el tamaño de un cuerpo de solicitud no exceda el límite superior.
2. Si hay solicitudes de inferencia de alta concurrencia y tráfico pesado, envíe un ticket de servicio al soporte de servicio profesional.

Resumen y Sugerencias

Nada

7.1.6 ¿Se pueden facturar los servicios en tiempo real sobre una base anual/mensual?

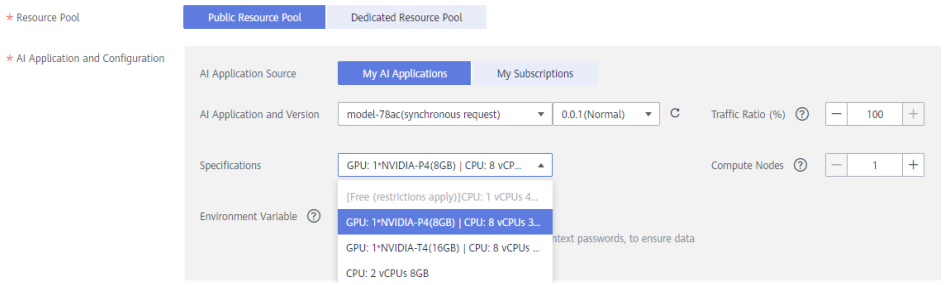
Los servicios en tiempo real no se pueden facturar anualmente/mensualmente.

7.1.7 ¿Cómo selecciono las especificaciones del nodo informático para desplegar un servicio?

Antes de desplegar un servicio, especifique las especificaciones de nodo. Las especificaciones de nodo mostradas en la GUI se calculan mediante ModelArts basándose en la aplicación de AI de destino y las especificaciones de nodo disponibles en el grupo de recursos. Puede seleccionar las especificaciones proporcionadas por ModelArts o personalizar las especificaciones (soportadas solo en grupos de recursos dedicados).

Selección de especificaciones de nodo de cómputo basadas en los recursos requeridos por su aplicación de IA. Por ejemplo, si una aplicación de IA requiere 3 CPU y 10 GB de memoria, seleccione especificaciones de nodo de cómputo superiores a 3 CPU y 10 GB de memoria. Esto garantiza que el servicio se pueda desplegar correctamente y ejecutar correctamente.

Figura 7-2 Especificaciones del nodo informático



Cuando utilice las especificaciones del nodo de cómputo, preste atención a lo siguiente:

Permission control

Las especificaciones de nodo informático de propósito general están disponibles públicamente, por ejemplo **modelarts.vm.cpu.2u**. Puede seleccionar las especificaciones siempre que haya recursos inactivos en el grupo de recursos. ModelArts ofrece dos especificaciones de forma predeterminada, **ModelArts.vm.cpu.2u** alimentado por CPU y **ModelArts.vm.gpu.p4** alimentado por GPU.

Para algunas especificaciones especiales, póngase en contacto con el administrador del sistema para solicitar permisos.

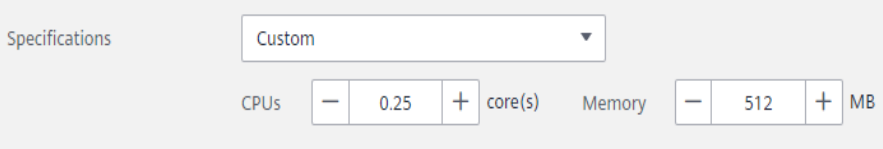
Specifications sold out in a public resource pool

Los recursos en un grupo de recursos públicos son limitados. Si una especificación se muestra como agotada, los recursos de la especificación actual se han agotado. En este caso, seleccione otras especificaciones o cree su propio grupo de recursos dedicado.

Custom specifications

Puede personalizar las especificaciones de recursos solo cuando se utiliza un grupo de recursos dedicado. Las especificaciones no se pueden personalizar en grupos de recursos públicos.

Figura 7-3 Especificaciones personalizadas



Free specifications

Solo los grupos de recursos públicos proporcionan especificaciones gratuitas. Las especificaciones gratuitas están restringidas en cantidad y duración de uso. Las especificaciones gratuitas solo están disponibles en la región **CN North-Beijing4** ahora.

7.1.8 ¿Qué es la versión de CUDA para desplegar un servicio en GPU?

CUDA 10.2 es compatible de forma predeterminada. Si se requiere una versión posterior, envíe un ticket de servicio para solicitar soporte técnico.

7.2 Servicios en tiempo real

7.2.1 ¿Qué hago si se produce un conflicto en el paquete de dependencia de Python de un script de predicción personalizado cuando despliego un servio en tiempo real?

Antes de importar un modelo, guarde el código de inferencia y el archivo de configuración en la carpeta del modelo. Al codificar con Python, importe paquetes personalizados en modo de importación relativa (importación de Python).

Si hay paquetes con nombres duplicados en el código de marco de inferencia de ModelArts y no se importan en modo de importación relativa, se producirá un conflicto, lo que provocará un despliegue de servicio o un fallo de predicción.

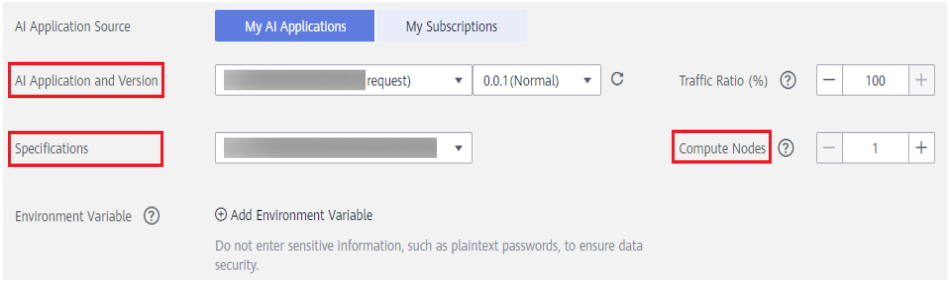
7.2.2 ¿Cómo acelero la predicción en tiempo real?

- Al desplegar un servicio en tiempo real, seleccione los nodos de procesamiento con especificaciones más altas para un mejor rendimiento. Por ejemplo, use GPU en lugar de CPU.
- Cuando despliegue un servicio en tiempo real, agregue el número de nodos informáticos.

  Si establece **Compute Nodes** en **1**, se utilizará el cómputo independiente. Si establece **Compute Nodes** en un valor mayor que 1, se utiliza el cómputo distribuido. Configure este parámetro en función de los requisitos del sitio.
- La velocidad de inferencia está estrechamente relacionada con la complejidad del modelo. Trate de optimizar el modelo para una predicción más rápida.

  ModelArts ofrece gestión de versiones de modelo para facilitar el seguimiento de fuentes y el ajuste repetido del modelo.

Figura 7-4 Despliegue de un servicio en tiempo real



7.2.3 ¿Cuál es el formato de una API de servicio en tiempo real?

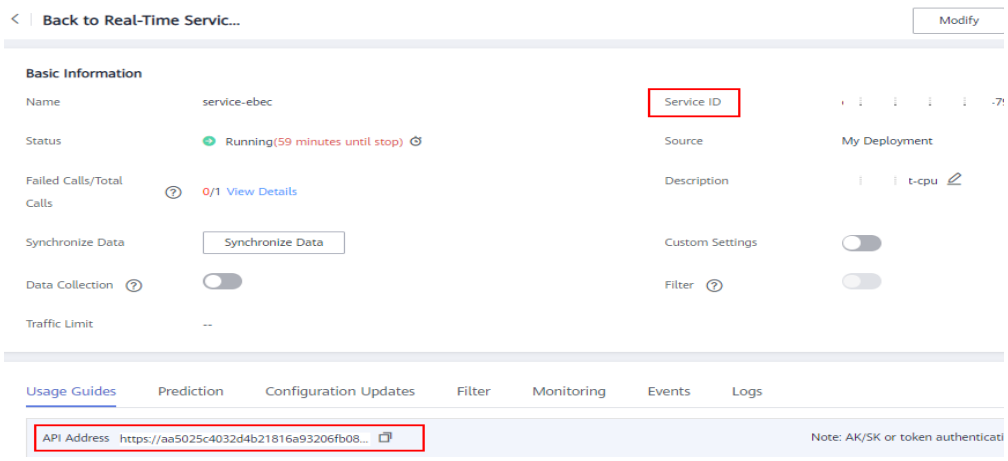
Después de desplegar un modelo como servicio en tiempo real, puede usar la API para inferencia.

```
https://Domain name/Version/infer/service ID
```

Por ejemplo:

```
https://6ac81cdfac4f4a30be95xxxbb682.apig.xxx.xxx.com/v1/infers/
468d146d-278a-4ca2-8830-0b6fb37d3b72
```


Figura 7-5 API



7.2.4 ¿Cómo puedo comprobar si un modelo causa un error cuando se ejecuta un servicio en tiempo real pero la predicción ha fallado?

Síntoma

Se utiliza un servicio en tiempo real en ejecución para la predicción. Después de que se inicia una solicitud de predicción, la respuesta recibida no cumple con la expectativa. Es difícil determinar si el problema es causado por el modelo.

Causa posible

Después de iniciar un servicio en tiempo real, se puede usar cualquiera de los siguientes métodos para la predicción:

- Método 1: Realice la predicción en la pestaña **Prediction** de la página de detalles del servicio.
- Método 2: Obtenga el URL de la API en la pestaña **Usage Guides** de la página de detalles del servicio y use cURL o Postman para la predicción.

Este problema puede producirse después de iniciar una solicitud de inferencia, independientemente de si se utiliza el método 1 o 2.

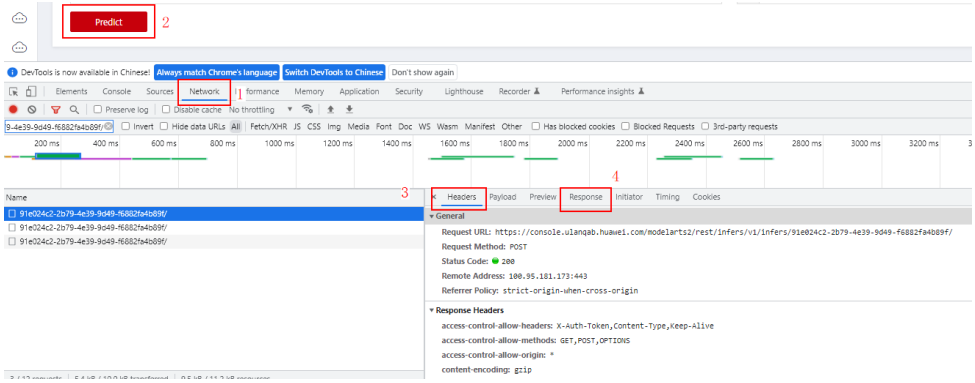
Finalmente se envía una solicitud de inferencia al modelo. El problema puede ser causado por un error ocurrido cuando el modelo procesó la solicitud de inferencia. Determine si el problema es causado por el modelo, lo que facilita la localización rápida de fallos.

Solución

Independientemente de si se usa el método 1 o 2, obtener la cabecera de respuesta y el cuerpo de la solicitud de inferencia.

- Si se utiliza el método 1, obtener la respuesta a la solicitud de inferencia con la herramienta de desarrollador del navegador. Tome Google Chrome como ejemplo. Presione **F12** para abrir la herramienta de desarrollo, haga clic en la ficha **Network** y luego en **Predict**. La respuesta a la solicitud de inferencia se muestra en la página de ficha **Network**.

Figura 7-6 Respuesta a una solicitud de inferencia



Busque la solicitud de inferencia en el panel **Name**. El URL de la solicitud de inferencia contiene la palabra clave **/v1/infers**. Vea el URL completo en el panel **Headers**. Obtenga la respuesta en **Headers** y **Response**.

- Si se utiliza el método 2, obtener la cabecera y el cuerpo de la respuesta con diferentes herramientas. Por ejemplo, ejecute el comando cURL y use **-I** para obtener el encabezado de respuesta.

Si **Server** en el encabezado de respuesta obtenido es **ModelArts** y el cuerpo de la respuesta no contiene un código de error de ModelArts.XXXX, el modelo devuelve la respuesta. Si la respuesta no es la esperada, el problema es causado por el modelo.

Resumen y Sugerencias

Un modelo se puede importar desde una imagen de contenedor, OBS o AI Gallery. A continuación se proporcionan métodos comunes de solución de problemas para cada origen de modelo:

- Para un modelo importado desde una imagen de contenedor, la causa del problema varía según la personalización de la imagen. Compruebe los logs del modelo para identificar la causa.
- Para un modelo importado de OBS, si la respuesta que recibió contiene un código de error MR, por ejemplo, MR.0105, consulte los logs en la ficha **Logs** de la página de detalles del servicio en tiempo real para identificar la causa.
- Para un modelo importado de AI Gallery, consulte al editor del modelo para conocer la causa.

7.2.5 ¿Cómo relleno el encabezado de solicitud y el cuerpo de solicitud de una solicitud de inferencia cuando se está ejecutando un servicio en tiempo real?

Síntoma

Después de desplegar un servicio en tiempo real, puede obtener su dirección de solicitud de inferencia en la ficha **Usage Guides** de la página de detalles del servicio cuando el servicio se está ejecutando. Sin embargo, no hay ninguna instrucción para rellenar la cabecera y el cuerpo de una solicitud de inferencia.

Causa posible

La dirección de solicitud de inferencia en la pestaña **Usage Guides** de la página de detalles del servicio puede ser llamada para inferencia. Por motivos de seguridad, ModelArts toma medidas de autenticación y autorización para evitar llamadas no autorizadas al servicio en tiempo real. Por lo tanto, la cabecera de una solicitud de predicción contiene la información de identidad del iniciador de solicitud, y el cuerpo contiene el contenido a predecir.

El encabezado debe autenticarse siguiendo las reglas de autenticación de Huawei Cloud. El cuerpo debe configurarse en función de los requisitos del modelo, como los requisitos de los scripts de preprocesamiento o las imágenes personalizadas.

Solución

- Encabezado:

  En la pestaña **Usage Guides** de la página de detalles del servicio, puede obtener un máximo de dos direcciones API, una para la autenticación de IAM o AK/SK y la otra para la autenticación de aplicaciones. La estructura de encabezado varía dependiendo del modo de autenticación.

  – Autenticación de IAM o AK/SK: En el encabezado, introduzca el token de nivel de dominio del tenant en la región de destino en el campo **X-Auth-Token**. Para obtener más información, consulte **Obtención de un token de usuario con la autenticación de contraseña**.

  – Autenticación de aplicaciones: la autenticación de aplicaciones se puede clasificar como autenticación de AppCode y autenticación de firma de aplicaciones.

    ■ Para la autenticación de AppCode, introduzca el AppCode de la aplicación asociada al servicio en tiempo real en el campo **X-Apig-AppCode** del encabezado.

    ■ Para la autenticación de firma de aplicación, en el encabezado, introduzca los valores **X-Sdk-Date** y **Authorization** generados mediante AppKey y AppSecret de la aplicación asociada al servicio en tiempo real a través del SDK o herramienta para autenticar la firma de la solicitud. Para obtener más información, consulte **Acceso autenticado mediante una aplicación**.

- Cuerpo:

  El cuerpo varía dependiendo de la fuente del modelo.

  – Si el modelo se importa desde una imagen de contenedor, el cuerpo debe configurarse en función de los requisitos de imagen personalizada. Para obtener más información, póngase en contacto con el creador de la imagen.

  – Si el modelo se importa desde OBS, los requisitos en el cuerpo se reflejan en el preprocesamiento del código de inferencia, que convertirá el cuerpo HTTP de entrada en la entrada requerida por el modelo. Para obtener más información, consulte **Especificaciones para codificación de inferencia de modelo**.

  – Si el modelo se obtiene de AI Gallery, compruebe la descripción de la invocación en AI Gallery o consulte al proveedor del modelo.

Resumen y Sugerencias

Nada

7.2.6 ¿Por qué no puedo acceder a la dirección de solicitud de inferencia obtenida desde el cliente iniciador?

Síntoma

Después de desplegar un servicio en tiempo real, puede obtener la dirección del servidor llamado en la ficha **Usage Guides** de la página de detalles del servicio cuando el servicio se está ejecutando. Sin embargo, esta dirección es inaccesible desde el cliente del iniciador de solicitud. Como resultado, la conexión no se pudo configurar y el nombre de dominio no se puede resolver.

Causa posible

Las direcciones que se muestran en la página de pestañas **Usage Guides** son las direcciones de Huawei Cloud API Gateway (APIG). La red entre el cliente del iniciador de solicitud y Huawei Cloud está desconectada.

Solución

Si el cliente está fuera de la red de Huawei Cloud, asegúrese de que el cliente puede acceder a Internet.

Si el cliente está en la red de Huawei Cloud, se puede acceder a la dirección en la configuración de red predeterminada. No configure configuraciones especiales de red, como reglas de firewall.

Resumen y Sugerencias

Nada

7.2.7 ¿Qué hago si no se extrae una imagen cuando se despliega, inicia, actualiza o modifica un servicio en tiempo real?

Causa posible

El espacio disponible en disco del nodo es menor que el tamaño de la imagen.

Solución

1. Reduzca el tamaño de la imagen.
2. Si el problema persiste después de reducir el tamaño de la imagen, póngase en contacto con el administrador del sistema.

7.2.8 ¿Qué hago si una imagen se reinicia repetidamente cuando se despliega, inicia, actualiza o modifica un servicio en tiempo real?

Causa posible

Hay un error en el código de imagen de contenedor.

Solución

Depure el código de imagen de contenedor basado en los logs de contenedor, crear la aplicación de IA de nuevo y desplegar la aplicación como un servicio en tiempo real.

7.2.9 ¿Qué hago si falló la comprobación del estado de un contenedor cuando se despliega, inicia, actualiza o modifica un servicio en tiempo real?

Causa posible

Error al invocar a la API de comprobación de estado de contenedor.

Solución

Localice el error basado en los registros de contenedor, depure el código, cree una aplicación de IA nuevamente y despliegue la aplicación como un servicio en tiempo real.

7.2.10 ¿Qué hago si los recursos son insuficientes cuando se despliega, inicia, actualiza o modifica un servicio en tiempo real?

Causa posible

Las especificaciones de instancia configuradas están más allá de las especificaciones proporcionadas por el grupo de recursos.

Solución

Cuando los recursos son insuficientes, ModelArts vuelve a intentarlo tres veces. Si se liberan recursos durante estos reintentos, el servicio se puede desplegar correctamente.

Si los recursos siguen siendo insuficientes después de tres reintentos, el despliegue de servicio falla. En este caso, realice las siguientes operaciones para resolver este problema:

- Si el servicio se va a desplegar en un grupo de recursos públicos, espere hasta que otros usuarios liberen recursos.
- Si el servicio se va a desplegar en un grupo de recursos dedicado, seleccione especificaciones de contenedor inferiores o especificaciones personalizadas para desplegar el servicio con la premisa de que se cumplen los requisitos del modelo.
- Amplíe la capacidad del grupo de recursos actual antes de desplegar el servicio.

7.2.11 ¿Qué hago si falló el despliegue de un servicio debido a una cuota insuficiente?

7.2.12 ¿Por qué falló el despliegue de mi servicio con el tiempo de espera deel despliegue adecuado configurado?

Un modelo puede iniciarse correctamente después de desplegar un servicio. El estado de inicio de un modelo se puede detectar con una comprobación de estado.

Compruebe si un servicio se despliega mediante una API de comprobación de estado para las imágenes personalizadas. Cuando cree una aplicación de IA, configure un retraso de comprobación de estado para garantizar la inicialización de contenedores.

Es una buena práctica configurar un retraso de comprobación de estado adecuado para el despliegue de servicio.

8 Grupos de recursos

8.1 ¿Puedo usar ECS para crear un grupo de recursos dedicado para ModelArts?

No. No se permite esta operación. Al crear un grupo de recursos, solo puede seleccionar los variantes de nodo disponibles proporcionados en la consola. Estos tipos de nodo en grupos de recursos dedicados provienen de ECS. Sin embargo, los ECS adquiridos con la cuenta no pueden ser utilizados por los grupos de recursos dedicados para ModelArts.

8.2 ¿Puedo desplegar varios servicios en un nodo de grupo de recursos dedicado?

Sí. Esta operación está permitida.

Cuando despliegue servicios, seleccione un grupo de recursos dedicado y personalice la variante de nodo informático. Seleccione el nodo con una variante baja. Cuando el nodo del grupo de recursos permite múltiples variantes de nodo de servicio, se pueden desplegar múltiples servicios. Si utiliza este método para desplegar un modelo para inferencia, asegúrese de que la variante seleccionada cumple con los requisitos mínimos del modelo para inferencia.Otherwise, the deployment or prediction may fail.

8.3 ¿Cómo se factura un nodo recién agregado a un grupo de recursos dedicado?

Recibirá una nueva factura con el nodo recién agregado incluido. Pague la factura y use el nodo.

8.4 ¿Cuáles son las diferencias entre un grupo de recursos públicos y un grupo de recursos dedicado?

- Todos los usuarios de ModelArts comparten un grupo de recursos públicos. Si los recursos son limitados, es posible que deba unirse a la cola.

- Un grupo de recursos dedicado está dedicado a usted y accesible para su 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 ¿Cómo se ponen en cola los trabajos de entrenamiento?

El primero en entrar, el primero en salir (FIFO) se aplica a los trabajos de entrenamiento. Los trabajos posteriores solo se pueden ejecutar después de que se haya completado el trabajo anterior. Esto puede conducir a la inanición de trabajos pequeños.

NOTA

La falta de trabajo es la siguiente: Por ejemplo, un trabajo de entrenamiento de 64 tarjetas está en cola, y un trabajo de entrenamiento de 1 tarjeta sigue al de 64 tarjetas. El trabajo de entrenamiento de 1 tarjeta solo se puede ejecutar después de que los recursos de 64 tarjetas estén inactivos. Incluso si los recursos de 30 tarjetas están disponibles, el trabajo de entrenamiento de 1 tarjeta no se puede ejecutar.

8.7 ¿Qué hago si los recursos son insuficientes para mirar un nuevo servicio en tiempo real después de detener un servicio en tiempo real en un grupo de recursos dedicado?

Espere varios minutos hasta que se liberen los recursos del servicio en tiempo real detenido.

8.8 ¿Se puede utilizar un grupo de recursos público para la conexión de red entre ModelArts y el servicio de autenticación para ejecutar algoritmos?

No. Los grupos de recursos públicos no se pueden utilizar para la conexión de red entre ModelArts y el servicio de autenticación. La conexión de red se puede configurar utilizando un grupo de recursos dedicado.

8.9 ¿Por qué un grupo de recursos dedicado que no se crea todavía se muestra en la consola después de que se elimina?

Después de eliminar un grupo de recursos dedicado en la consola, el backend libera los recursos utilizados por el grupo. Se tarda varios minutos en liberar los recursos, durante los cuales el grupo todavía se muestra en la consola. Para crear un grupo de recursos dedicado de nuevo, espere 5 minutos después de la eliminación. Además, no utilice el nombre del grupo de recursos dedicado que no se puede crear para nombrar el nuevo grupo de recursos dedicado. Para realizar una prueba automatizada en la interfaz de usuario, es una buena práctica utilizar una cadena aleatoria como el nombre del grupo de recursos dedicado creado.

8.10 ¿Cómo agrego una interconexión de VPC entre un grupo de recursos dedicado y un SFS?

Para agregar una interconexión de VPC, haga lo siguiente:

1. Inicie sesión en la consola de gestión de ModelArts y elija **Dedicated Resource Pools** en el panel de navegación.
2. En la lista del grupo de recursos dedicado, haga clic en el ID o el nombre del grupo de recursos de destino para ir a su página de detalles.
3. Haga clic en **Configure NAS VPC** en la esquina superior derecha. En el cuadro de diálogo **Configure NAS VPC**, habilite **NAS VPC Connection** y configure la VPC NAS y la subred de NAS para que sean las mismas que las configuradas para el SFS.
4. Haga clic en **OK**. Cuando cree un trabajo de entrenamiento, la opción de SFS estará disponible.

8.11 ¿Qué debo hacer si un trabajo de entrenamiento siempre está esperando en una cola de recursos?

Cuando ejecuta una instancia de notebook y crea un trabajo de entrenamiento al mismo tiempo, la instancia de notebook y el trabajo de entrenamiento utilizan los mismos recursos. Si solo hay un nodo en el grupo de recursos dedicado, el trabajo se pondrá en cola debido a la insuficiencia de recursos. En este caso, puede detener la instancia del notebook y, a continuación, crear el trabajo de entrenamiento.

9 API/SDK

9.1 ¿Se pueden usar las API o los SDK de ModelArts para descargar modelos a una PC local?

Las API o los SDK de ModelArts no se pueden usar para descargar modelos a una PC local. Sin embargo, los modelos de salida de los trabajos de entrenamiento se almacenan en OBS. Puede utilizar las API o los SDK de OBS para descargar los modelos. Para obtener más información, consulte **Descargar un objeto**.

9.2 ¿Qué entornos de instalación admiten los SDK de ModelArts?

Los SDK de ModelArts pueden ejecutarse en notebook o entornos locales. Sin embargo, los entornos compatibles varían dependiendo de las arquitecturas. Para obtener más información, véase **Tabla 9-1**.

Tabla 9-1 Entornos de instalación del SDK

Entorno de desarrollo	Arquitectura	Compatible
Notebook	Arm	Sí
	x86	Sí
Entorno local	Arm	No
	x86	Sí

9.3 ¿Utiliza ModelArts la API de OBS para acceder a archivos de OBS por una intranet o Internet?

En la misma región de ModelArts utiliza la API de OBS para acceder a los archivos almacenados en OBS a través de una intranet y no consume tráfico de red pública.

Si descarga datos de OBS por Internet, se le cobrará por el tráfico de la red pública de OBS. Para obtener más información sobre la facturación de OBS, consulte **Concepto de facturación**.

9.4 ¿Cómo obtengo una curva de uso de recursos de trabajo después de enviar un trabajo de entrenamiento llamando a una API?

Después de enviar un trabajo de entrenamiento llamando a una API, inicie sesión en la consola de ModelArts, elija **Training Management** > **Training Jobs** y haga clic en el nombre o ID del trabajo de entrenamiento de destino para ir a su página de detalles. En el área **Resource Usages**, vea la curva de uso de recursos del trabajo.

