ModelArts

Gestión de recursos

Edición 01

Fecha 2024-09-14

HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD.

Marcas registradas y permisos

El logotipo HUAWEI 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

Es posible que la totalidad o parte de los productos, las funcionalidades y/o los servicios que figuran en el presente documento no se encuentren dentro del alcance de un contrato vigente entre Huawei Cloud y el cliente. Las funcionalidades, los productos y los servicios adquiridos se limitan a los estipulados en el respectivo contrato. A menos que un contrato especifique lo contrario, ninguna de las afirmaciones, informaciones ni recomendaciones contenidas en el presente documento constituye garantía alguna, ni expresa ni implícita.

Huawei está permanentemente preocupada por la calidad de los contenidos de este documento; sin embargo, ninguna declaración, información ni recomendación aquí contenida constituye garantía alguna, ni expresa ni implícita. La información contenida en este documento se encuentra sujeta a cambios sin previo aviso.

Huawei Cloud Computing Technologies Co., Ltd.

Dirección: Huawei Cloud Data Center Jiaoxinggong Road
Avenida Qianzhong
Nuevo distrito de Gui'an
Gui Zhou, 550029
República Popular China

Sitio web: https://www.huaweicloud.com/intl/es-us/

Índice

1 Grupo de recursos

Grupos de recursos de ModelArts

Al usar ModelArts para el desarrollo de IA, puede usar cualquiera de los siguientes grupos de recursos:

- **Dedicated resource pool**: Ofrece recursos más controlables y no se puede compartir con otros usuarios. Cree un grupo de recursos dedicado y selecciónelo durante el desarrollo de IA. El grupo de recursos dedicado puede ser un clúster elástico o un BMS elástico.
  - Clúster elástico: puede ser Standard o Lite.
    - En un clúster elástico Standard, se proporcionan recursos de cómputo exclusivos, con los que puede entregar instancias durante el entrenamiento laboral, el despliegue de modelos y el desarrollo del entorno de ModelArts.
    - Un clúster elástico Lite proporciona clústeres alojados de Kubernetes con complementos de desarrollo de IA y complementos de aceleración para usuarios de recursos de Kubernetes. Puede operar los nodos y los clústeres de Kubernetes en el grupo de recursos con los recursos y tareas nativos de IA proporcionados.
  - BMS elástico: Proporciona diferentes modelos de BMS de *x*PU. Puede acceder a un BMS elástico con una EIP e instalar controladores y software relacionados con GPU y NPU en una imagen de SO específica. Para cumplir con los requisitos de entrenamiento de rutina de los ingenieros de algoritmos, SFS y OBS se pueden utilizar para almacenar y leer datos.
- **Public Resource Pool**: proporciona clústeres de cómputo público a gran escala, que se asignan en función de la configuración de parámetros de trabajo. Los recursos se aíslan por trabajo. Puede utilizar los grupos de recursos públicos de ModelArts para ofrecer trabajos de entrenamiento, modelos de desplegar o ejecutar instancias de DevEnviron y se facturará en régimen de pago por uso.

Diferencias entre los grupos de recursos dedicados y los grupos de recursos públicos

- Los grupos de recursos dedicados proporcionan clústeres de cómputo dedicados y recursos de red para los usuarios. Los grupos de recursos dedicados de diferentes usuarios están físicamente aislados, mientras que los grupos de recursos públicos solo están aislados lógicamente. En comparación con los grupos de recursos públicos, los grupos de recursos dedicados ofrecen un mejor rendimiento en aislamiento y seguridad.

- Cuando se utiliza un grupo de recursos dedicado para crear trabajos y los recursos son suficientes, los trabajos no se pondrán en cola. Cuando se utiliza un grupo de recursos público para crear trabajos, hay una alta probabilidad de que los trabajos se pongan en cola.
- La red tiene acceso a un grupo de recursos dedicado. Todos los trabajos en ejecución en el grupo pueden acceder al almacenamiento y a los recursos de la red. Por ejemplo, si selecciona un grupo de recursos dedicado con una red accesible al crear un trabajo de entrenamiento, puede acceder a los datos de SFS después de crear el trabajo de entrenamiento.
- Los grupos de recursos dedicados permiten personalizar el entorno de tiempo de ejecución de los nodos físicos, por ejemplo, puede actualizar los controladores de GPU o de Ascend. Los grupos de recursos públicos no admiten esta función.

2 Clúster elástico

2.1 Actualizaciones integrales a las funciones de gestión de grupo de recursos de ModelArts

Los grupos de recursos dedicados de ModelArts se actualizaron y entraron en vigor a las 00:00 GMT+08:00 de 1 de marzo de 2023. En el nuevo sistema, solo hay grupos de recursos dedicados de ModelArts unificados, que ya no se clasifican como los grupos dedicados al desarrollo/entrenamiento y los grupos dedicados al despliegue de servicios. Los grupos de recursos dedicados de la nueva versión admiten la configuración flexible de tipos de trabajos y le permiten gestionar redes e interconectar las VPC con redes.

La nueva página de gestión de grupo de recursos dedicado proporciona funciones más completas y muestra más información sobre los grupos de recursos. En las secciones siguientes de este documento se proporcionan más detalles sobre cómo usar y gestionar grupos de recursos dedicados. Si es la primera vez que utiliza los grupos de recursos dedicados de ModelArts, pruebe con grupos de recursos dedicados de nueva versión. Si ha

utilizado grupos de recursos dedicados de ModelArts, los grupos de la versión antigua se cambiarán sin problemas a grupos de la nueva versión.

Lea el siguiente contenido para obtener más información sobre los grupos de recursos dedicados de la nueva versión.

Características de los grupos de recursos dedicados de la nueva versión

La nueva versión de gestión de grupos de recursos dedicados es una mejora integral de la tecnología y el producto. Las principales mejoras son las siguientes:

- **Tipo de grupo singular de recurso dedicado para diversos fines**: los grupos de recursos dedicados ya no se clasifican en los de desarrollo/entrenamiento y los de despliegue de servicios. Puede ejecutar cargas de trabajo de entrenamiento e inferencia en un grupo de recursos dedicado. También puede establecer los tipos de trabajo admitidos por un grupo de recursos dedicado en función de sus necesidades.
- **Conexión de red de grupo de recursos dedicado**: Puede crear y gestionar redes de grupos de recursos dedicados en la consola de gestión de ModelArts. Si necesita acceder a los recursos de su VPC para trabajos que se ejecutan en un grupo de recursos dedicado, interconecte la VPC con la red del grupo de recursos dedicado.
- **Más detalles de clúster**: La página de detalles del grupo de recursos dedicados de la nueva versión proporciona más detalles del clúster, como trabajos, nodos y monitoreo de recursos, lo que le ayuda a conocer el estado del cluster y a planificar y utilizar mejor los recursos.
- **Gestión de controlador de clúster de GPU/NPU**: En la página de detalles del grupo de recursos dedicado de la nueva versión, puede seleccionar un controlador de tarjeta aceleradora y realizar cambios al enviarlo o actualizarlo sin problemas según los requisitos de servicio.
- **Asignación de recursos de grano fino (próximamente)**: puede dividir el grupo de recursos dedicado en varios pequeños grupos y asignar diferentes cuotas y permisos a cada uno de ellos para una asignación y gestión de recursos flexibles y refinadas.

Se proporcionarán más funciones en versiones posteriores para una mejor experiencia de usuario.

¿Puedo continuar utilizando los grupos de recursos dedicados existentes después de que la actualización surta efecto?

Si ha creado grupos de recursos dedicados, puede acceder a la página de gestión del grupo de recursos dedicado de la versión anterior (clúster elástico) en la consola de gestión de ModelArts y utilizar los grupos de recursos creados, pero no puede crear un grupo de recursos dedicado en esa página. ModelArts permite migrar los grupos de recursos dedicados existentes a la nueva página de gestión. Se le contactará para completar la migración y esto no requiere que realice ninguna operación. Además, la migración no afecta a las cargas de trabajo que se ejecutan en los grupos de recursos dedicados. Preste atención a las nuevas funciones de gestión fáciles de usar de los grupos de recursos dedicados. No hay cambios en la creación de empleos de entrenamiento o servicios de inferencia.

¿Serán más costosos los grupos de recursos dedicados de la nueva versión?

La unidad de tarificación y el precio unitario de los grupos de recursos dedicados de la nueva versión son los mismos que los de los grupos de recursos dedicados de la antigua versión. Si no amplía o reduce sus grupos de recursos dedicados, la tarifa no cambiará. Además, en

versiones posteriores se ofrecerán más funciones de valor agregado, como la división en subgrupo, la compartición elástica y la aceleración de datos, para asignar mejor los recursos de cómputo y mejorar la rentabilidad.

Diferencias entre los grupos de recursos dedicados nuevos y antiguos

- En la versión anterior, los grupos de recursos dedicados al desarrollo/entrenamiento están separados de los dedicados al despliegue de servicios. Además, los grupos de los dos tipos ofrecen diferentes funciones y su experiencia de usuario varía. En la nueva versión, los grupos de recursos dedicados de los dos tipos están unificados. solo necesita configurar uno o varios tipos de trabajo. Luego, el grupo de recursos dedicado soporta automáticamente el tipo de trabajo configurado.
- Los nuevos grupos de recursos dedicados heredan todas las funciones de los antiguos y han mejorado enormemente la experiencia del usuario en funciones clave, como la compra y el cambio de tamaño de un grupo de recursos. Utilice nuevos grupos de recursos dedicados para disfrutar de una experiencia fluida y transparente.
- Además, los nuevos grupos de recursos dedicados ofrecen funciones mejoradas, por ejemplo, actualizar controladores de GPU o de Ascend, ver detalles sobre la cola de trabajos y usar una red para varios grupos. Próximamente habrá más funciones nuevas de los nuevos grupos de recursos dedicados.

¿Cómo puedo obtener ayuda o proporcionar comentarios si encuentro problemas durante el uso?

Al igual que otras funciones de ModelArts, puede reportar problemas u obtener ayuda en la barra lateral de la consola. Además, se recomienda leer las siguientes secciones de este documento para comprender mejor cómo utilizar los grupos de recursos dedicados de ModelArts. Envíe un ticket de servicio para más requerimientos.

Instrucciones de grupos de recursos dedicados

- Si utiliza grupos de recursos dedicados por primera vez, lea **Grupo de recursos** para empezar.
- Consulte **Creación de un grupo de recursos** para crear un grupo de recursos dedicado.
- Consulte **Consulta de detalles sobre un grupo de recursos** para ver los detalles de un grupo de recursos dedicado creado.
- Si las especificaciones de un grupo de recursos dedicado no cumplen con los requisitos de servicio, consulte **Cambio de tamaño de un grupo de recursos** para ajustar las especificaciones.
- Consulte **Cambio de tipos de trabajos soportados por un grupo de recursos** para establecer o cambiar los tipos de trabajo admitidos por un grupo de recursos dedicado.
- Consulte **Actualización de un controlador de grupo de recursos** para actualizar el controlador de GPU/Ascend de sus grupos de recursos dedicados.
- Si ya no se necesita un grupo de recursos dedicado, consulte **Eliminación de un grupo de recursos** para eliminarlo.
- Si se produce una excepción cuando se utiliza un grupo de recursos dedicado, trate la excepción según **Estado anormal de un grupo de recursos dedicado**.
- Consulte **Red de ModelArts** para gestionar redes de grupos de recursos dedicados o interconectar las VPC con las redes.

2.2 Creación de un grupo de recursos

En esta sección se describe cómo crear un grupo de recursos dedicado.

Procedimiento

1. Inicie sesión en la consola de ModelArts. En el panel de navegación, seleccione **Dedicated Resource Pools** > **Elastic Cluster**.

   NOTA

   Para los nuevos usuarios, solo los clústeres elásticos de nueva versión están disponibles en la consola de ModelArts. Los usuarios que han utilizado grupos de recursos dedicados de versiones antiguas pueden acceder a los clústeres elásticos de versiones antiguas y nuevas.

2. En la ficha **Resource Pools**, haga clic en **Create** y configure los parámetros.

   Tabla 2-1 Parámetros del grupo de recursos dedicados

Parámetro	Subparámetro	Descripción
Nombre	N/A	Nombre de un grupo de recursos dedicado. Solo se permiten letras en minúscula, dígitos y guiones (-). El valor debe comenzar con una minúscula y no puede finalizar con un guion (-).
Description	N/A	Breve descripción de un grupo de recursos dedicado.
Billing Mode	N/A	Puede seleccionar **Pay-per-use**.
Tipo de grupo de recursos	N/A	Puede seleccionar **Physical** o **Logical**. Si no hay ninguna especificación lógica, no se muestra **Logical**.
Job Type	N/A	Seleccione los tipos de trabajo admitidos por el grupo de recursos en función de los requisitos del servicio. ● **Physical**: **DevEnviron**, **Training Job** e **Inference Service** son compatibles. ● **Logical**: solo se admite **Training Job**.

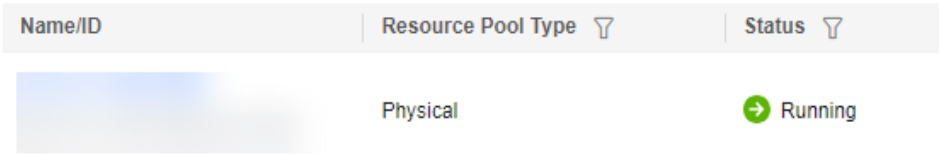
Parámetro	Subparámetro	Descripción
Network	N/A	Red en la que se despliega la instancia de servicio de destino. La instancia puede intercambiar datos con otros recursos de servicios en la nube en la misma red. Seleccione una red en el cuadro de lista desplegable. Si no hay ninguna red disponible, haga clic en **Create** a la derecha para crear una red. Para obtener detalles sobre cómo crear una red, véase **Creación de una red**.
Specification Management	Specifications	Seleccione las especificaciones requeridas. Debido a la pérdida del sistema, los recursos disponibles reales son menores que los especificados en las especificaciones. Después de crear un grupo de recursos dedicado, puede ver los recursos disponibles reales en la ficha **Nodes** de la página de detalles del grupo de recursos dedicado.
	AZ	Puede seleccionar **Automatically allocated** o **Specifies AZ**. Una AZ es una región física donde los recursos usan fuentes de alimentación y redes independientes. Las zonas de disponibilidad están físicamente aisladas pero interconectadas a través de una intranet. ● **Automatically allocated**: las AZ se asignan automáticamente. ● **Specifies AZ**: especifique AZ para los nodos del grupo de recursos. Para garantizar la recuperación ante desastres del sistema, despliegue todos los nodos de la misma AZ. Puede configurar el número de nodos de una AZ.
	Nodes	Seleccione el número de nodos de un grupo de recursos dedicado. Más nodos significan un mayor rendimiento de cómputo. Si **AZ** se establece en **Spec AZ**, no es necesario configurar **Nodes**. **NOTA** Es una buena práctica crear no más de 30 nodos a la vez. De lo contrario, la creación puede fallar debido a limitaciones de tráfico.
	Advanced Configuration	Esto permite configurar el espacio del motor de contenedor. Debe introducir un entero para el espacio del motor de contenedor. No puede ser inferior a 50 GB, que es el valor predeterminado y mínimo. El valor máximo depende de las especificaciones. Para ver los valores válidos, compruebe el indicador de la consola. La personalización del espacio del motor de contenedor no aumenta los costos.
Custom Driver	N/A	Este parámetro solo está disponible cuando se selecciona una variante de GPU o de Ascend. Habilite esta función y seleccione un controlador.
GPU Driver	N/A	Este parámetro solo está disponible cuando el controlador personalizado está habilitado. Seleccione un controlador de acelerador de GPU.

Parámetro	Subparámetro	Descripción
Required Duration	N/A	Seleccione el período de tiempo durante el que desea utilizar el grupo de recursos. Este parámetro es obligatorio solo cuando se selecciona el modo de facturación **Yearly/Monthly**.
Auto-renewal	N/A	Especifica si se debe habilitar la renovación automática. Este parámetro es obligatorio solo cuando se selecciona el modo de facturación **Yearly/Monthly**. ● Las suscripciones mensuales se renuevan cada mes. ● Las suscripciones anuales se renuevan cada año.
Advanced Options	N/A	Seleccione **Configure Now** para configurar la información de etiquetas, el bloque de CIDR y la distribución del nodo controlador.
Tags	N/A	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. Para obtener más información sobre cómo utilizar etiquetas, véase **¿Cómo utiliza ModelArts las etiquetas para gestionar recursos por grupo?** **NOTA** Puede seleccionar una etiqueta de TMS predefinida de la lista desplegable de etiquetas o personalizar una etiqueta. Las etiquetas predefinidas están disponibles para todos los recursos de servicio que admiten etiquetas. Las etiquetas personalizadas solo están disponibles para los recursos de servicio del usuario que las ha creado.
CIDR block	N/A	Puede seleccionar **Default** o **Custom**. ● **Default**: el sistema le asigna aleatoriamente un bloque de CIDR disponible, que no se puede modificar después de crear el grupo de recursos. Para uso comercial, personalice su bloque de CIDR. ● **Custom**: debe personalizar el contenedor de K8S y los bloques de CIDR de servicio K8S. – **K8S Container Network**: utilizado por el contenedor en un clúster, que determina cuántos contenedores puede haber en un clúster. El valor no se puede cambiar después de crear el grupo de recursos. – **K8S Service Network**: se utiliza cuando los contenedores del mismo clúster se acceden entre sí, lo que determina cuántos Service puede haber. El valor no se puede cambiar después de crear el grupo de recursos.

Parámetro	Subparámetro	Descripción
Master Distribution	N/A	Ubicaciones de distribución de nodos de controladores. Puede seleccionar **Random** o **Custom**. ● **Random**: utilice las AZ asignadas aleatoriamente por el sistema. ● **Custom**: seleccione AZ para los nodos del controlador. Distribuya los nodos del controlador en diferentes AZ para la recuperación ante desastres.

3. Haga clic en **Next** y confirme la configuración. Luego, haga clic en **Submit** para crear el grupo de recursos dedicado.
   – Después de crear un grupo de recursos, su estado cambia a **Running**. Solo cuando el número de nodos disponibles es superior a 0, se pueden entregar tareas a este grupo de recursos.

   Figura 2-1 Consulta de un grupo de recursos

   – Pase el cursor sobre **Creating** para ver los detalles sobre el proceso de creación. Haga clic en **View Details**. Aparece en pantalla la página de registro de operaciones.

   Figura 2-2 Creación

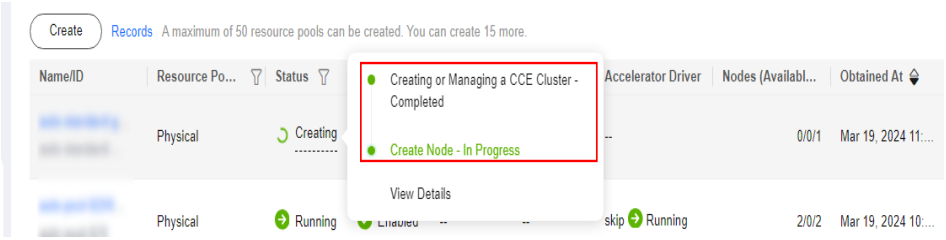
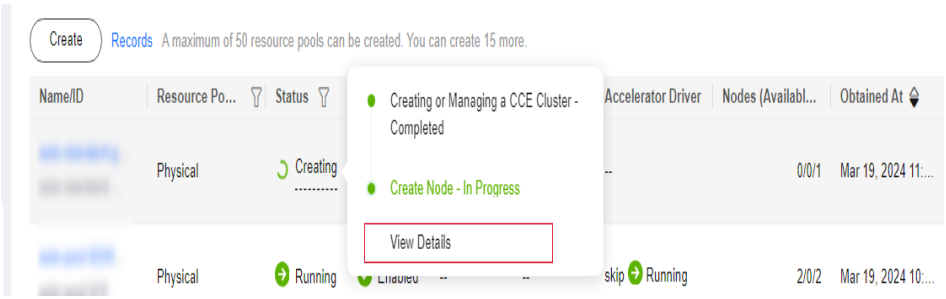
   

   Figura 2-3 Consultar detalles

   – Puede ver los registros de tareas del grupo de recursos haciendo clic en **Records** en la esquina superior izquierda de la lista del grupo de recursos.

Figura 2-4 Registros de operaciones

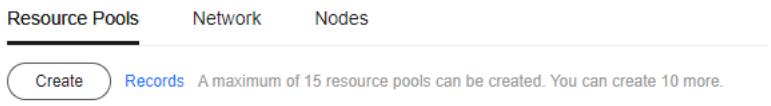
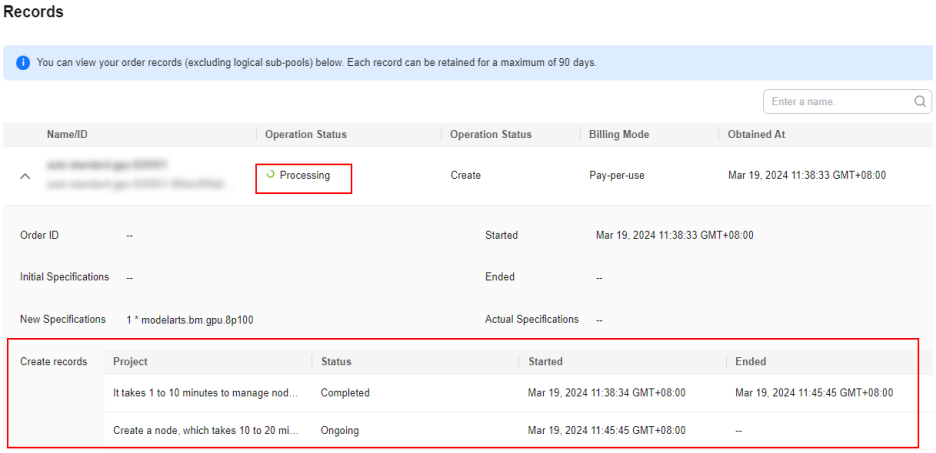


Figura 2-5 Consulta del estado del grupo de recursos



Preguntas frecuentes

¿Qué pasa si elijo una variante para un grupo de recursos dedicado, pero recibo un mensaje de error que dice que no hay ningún recurso disponible?

Las variantes de los recursos dedicados cambian en función de la disponibilidad en tiempo real. A veces, puede elegir una variante en la página de compra, pero se agota antes de pagar y crear el grupo de recursos. Esto hace que la creación del grupo de recursos falle.

Puede probar otra variante en la página de creación y volver a crear el grupo de recursos.

P: ¿Por qué no puedo usar todos los recursos de CPU en un nodo en un grupo de recurso?

Los nodos del grupo de recursos tienen sistemas y componentes instalados en ellos. Estos consumen algunos recursos de CPU. Por ejemplo, si un nodo tiene 8 vCPU, pero algunos de ellos son utilizados por los componentes del sistema, los recursos disponibles serán inferiores a 8 vCPU.

Puede comprobar los recursos de CPU disponibles haciendo clic en la ficha **Nodes** de la página de detalles del grupo de recursos antes de iniciar una tarea.

2.3 Consulta de detalles sobre un grupo de recursos

Página de detalles del grupo de recurso

- Inicie sesión en la consola de ModelArts. En el panel de navegación, seleccione **Dedicated Resource Pools** > **Elastic Cluster**.
- Haga clic en junto al tipo o estado del grupo de recursos en el encabezado de la tabla. En la esquina superior derecha de la lista, seleccione **Name** o **Resource ID** para filtrar los grupos de recursos. Para obtener el ID del recurso, acceda a la página **Billing Center**

> **Orders** > **My Orders** y haga clic en **Details** en la columna **Operation** del orden de destino.

- En la lista del grupo de recursos, haga clic en un grupo de recursos para ir a su página de detalles y ver su información.
  - Si hay varios grupos de recursos, haga clic en ▼ en la esquina superior izquierda de la página de detalles de un grupo de recursos para cambiar de grupo de recursos. Haga clic en **More** en el extremo superior derecho para realizar operaciones como cambiar el tamaño o eliminar el grupo de recursos. Las operaciones disponibles varían según el grupo de recursos.
  - En el área **Network** de **Basic Information**, puede hacer clic en el número de grupos de recursos asociados para ver los grupos de recursos asociados.
  - En el área de información extendida, puede ver la información de monitoreo, trabajos, nodos, especificaciones y eventos. Para obtener más información, véase la siguiente sección.

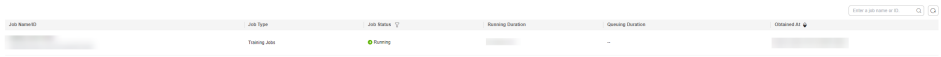
Consulta de trabajos en un grupo de recursos

En la página de detalles del grupo de recursos, haga clic en **Jobs**. Puede ver todos los trabajos que se ejecutan en el grupo de recursos. Si un trabajo está en cola, puede ver su posición en cola.

NOTA

Solo se pueden ver los trabajos de entrenamiento.

Figura 2-6 Trabajos



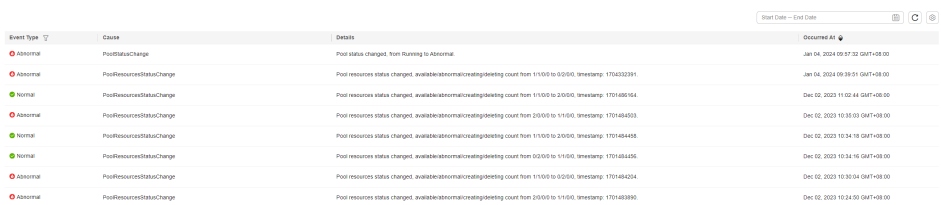
Consulta de eventos de grupo de recursos

En la página de detalles del grupo de recursos, haga clic en **Events**. Puede ver todos los eventos del grupo de recursos. La causa de un evento es **PoolStatusChange** o **PoolResourcesStatusChange**.

En la lista de eventos, haga clic en  a la derecha de **Event Type** para filtrar eventos.

- Cuando un grupo de recursos comienza a crearse o se vuelve anormal, el estado del grupo de recursos cambia y el cambio se registra como un evento.
- Cuando cambia la cantidad de nodos disponibles o anormales o en proceso de creación o eliminación, cambia el estado del nodo del grupo de recursos y el cambio se registrará como un evento.

Figura 2-7 Eventos



Consulta de nodos de grupo de recursos

En la página de detalles del grupo de recursos, haga clic en **Nodes**. Puede ver todos los nodos del grupo de recursos y el uso de recursos de cada nodo.

Algunos recursos están reservados para componentes de clúster. Por lo tanto, **CPUs (Available/Total)** no indica la cantidad de recursos físicos en el nodo. Solo muestra la cantidad de recursos que pueden ser utilizados por los servicios. Los núcleos de CPU se miden en milicores, y 1000 milicores equivalen a 1 núcleo físico.

- Reemplazo de un nodo:

  En la ficha **Nodes**, localice el nodo que se va a reemplazar. En la columna **Operation**, haga clic en **Replace**. No se cobran tarifas por esta operación.

  Verifique los registros de reemplazo de nodos en la página **Records**. **Running** indica que el nodo se está reemplazando. Después del reemplazo, puede verificar el nuevo nodo en la lista de nodos.

  El reemplazo no puede durar más de 24 horas. Si no se encuentra ningún recurso adecuado después de que se agote el tiempo de espera del reemplazo, el estado cambia a **Failed**. Pase el ratón sobre para verificar la causa de la falla.

  NOTA

  - La cantidad de reemplazos por día no puede superar el 20 % del total de nodos en el grupo de recursos. La cantidad de nodos que se reemplazarán no puede superar el 5 % del total de nodos en el grupo de recursos.
  - Asegúrese de que haya recursos de nodo inactivos. De lo contrario, el reemplazo puede fallar.
  - Si hay nodos en el estado **Resetting** en los registros de operación, los nodos del grupo de recursos no se pueden reemplazar.

- Restablecimiento de un nodo

  En la ficha **Nodes**, localice el nodo que desea restablecer. Haga clic en **Reset** en la columna **Operation** para restablecer un nodo. También puede seleccionar varios nodos y hacer clic en **Reset** para restablecer varios nodos.

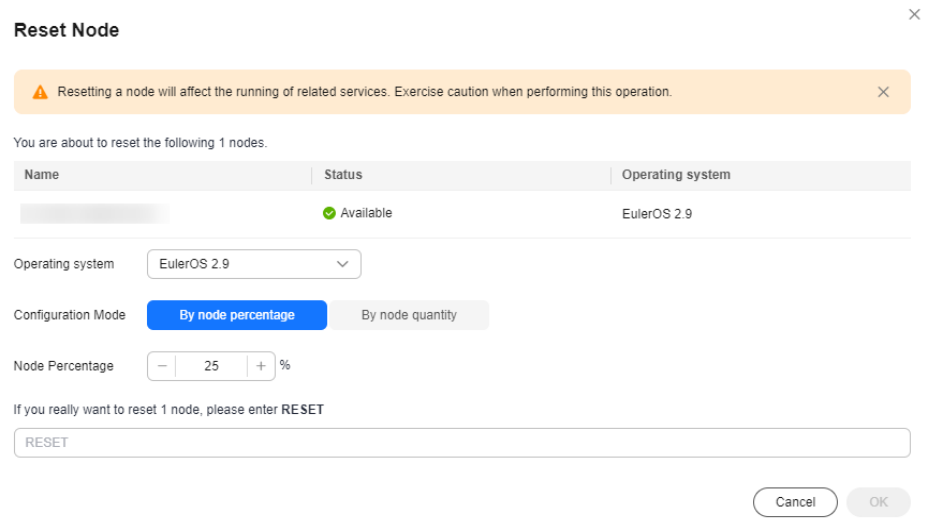
  Configure los parámetros descritos en la siguiente tabla.

  Tabla 2-2 Parámetros

Parámetro	Descripción
Operating System	Seleccione un SO del cuadro de lista desplegable.
Configuration Mode	Seleccione un modo de configuración para restablecer el nodo. ● **By node percentage**: la proporción máxima de nodos que se pueden restablecer si hay varios nodos en la tarea de restablecimiento ● **By node quantity**: el número máximo de nodos que se pueden restablecer si hay varios nodos en la tarea de restablecimiento

  Verifique los registros de reinicio del nodo en la página **Records**. Si se está restableciendo el nodo, su estado será **Resetting**. Una vez finalizado el restablecimiento, el estado del nodo cambia a **Available**. El restablecimiento de un nodo no se cobrará.

Figura 2-8 Restablecimiento de un nodo

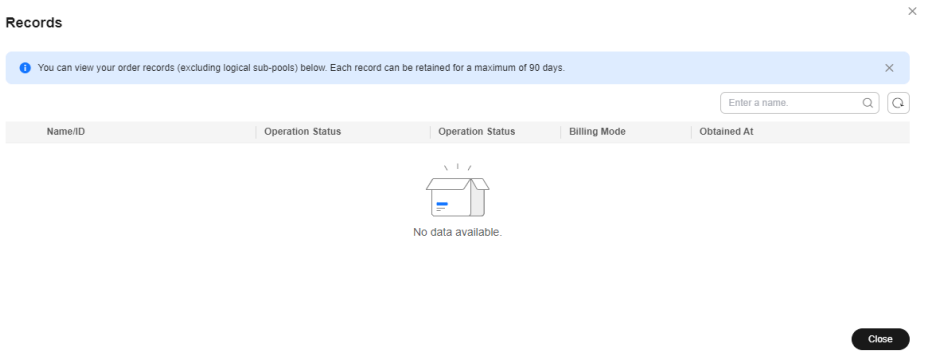


NOTA

- El restablecimiento de un nodo afectará a los servicios en ejecución.
- Solo se pueden restablecer los nodos en estado **Available**.
- Un nodo singular puede estar en una sola tarea de restablecimiento a la vez. No se pueden entregar varias tareas de restablecimiento al mismo nodo a la vez.
- Si hay nodos en estado **Replacing** en los registros de operación, los nodos del grupo de recursos no se pueden restablecer.
- Cuando se está actualizando el controlador de un grupo de recursos, los nodos de este grupo de recursos no se pueden restablecer.
- Para las especificaciones de GPU y NPU, después de reiniciar el nodo, se puede actualizar el controlador del nodo. Espere pacientemente.

Figura 2-9 Nodos

Figura 2-10 Registros de operaciones



- Eliminación, cancelación de suscripción o lanzamiento de un nodo
  - Para un grupo de recursos de pago por uso, haga clic en **Delete** en la columna **Operation**.

Para eliminar nodos por lotes, active las casillas de verificación situadas junto a los nombres de nodo y haga clic en **Delete**.

– Para un grupo de recursos anual/mensual cuyos recursos no han expirado, haga clic en **Unsubscribe** en la columna **Operation**.
– Para un grupo de recursos anual/mensual cuyos recursos han expirado (en el período de gracia), haga clic en **Release** en la columna **Operation**.

Si el botón para eliminar está disponible para un nodo anual/mensual, el nodo es un nodo de inventario, haga clic en **Delete**.

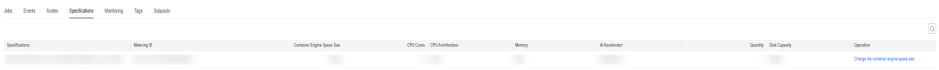
NOTA

- Antes de eliminar, cancelar la suscripción o lanzar un nodo, asegúrese de que no hay trabajos en ejecución en este nodo. De lo contrario, los trabajos se interrumpirán.
- Elimine, cancele la suscripción o lance los nodos anormales en un grupo de recursos y agregue nodos nuevos para su sustitución.
- Si solo hay un nodo, no se puede eliminar, cancelar o lanzar.

Consulta de especificaciones de grupo de recursos

En la página de detalles del grupo de recursos, haga clic en **Specifications**. Puede ver las especificaciones utilizadas por el grupo de recursos y el número de cada especificación.

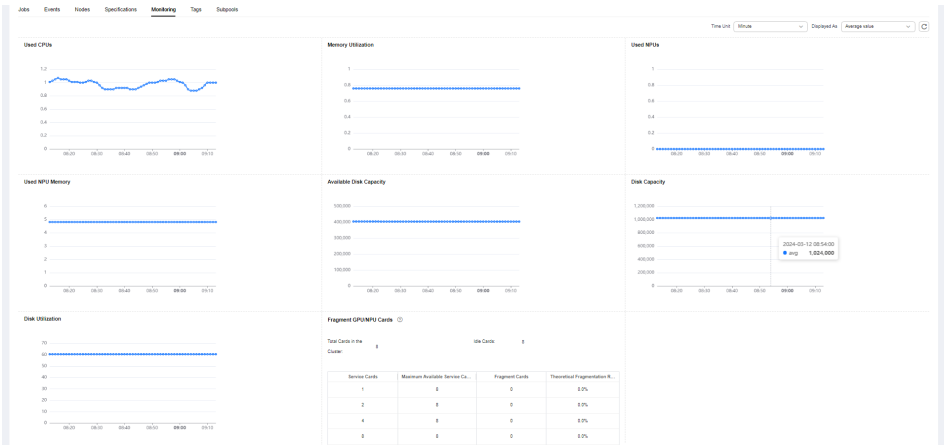
Figura 2-11 Ver especificaciones de grupo de recursos (El tamaño del motor de contenedor se muestra como el valor predeterminado si no está definido)



Consulta de información de control de grupo de recursos

En la página de detalles del grupo de recursos, haga clic en **Monitoring**. Se muestra el uso de recursos, incluidas las CPU usadas, el uso de memoria y la capacidad de disco disponible del grupo de recursos. Si se utilizan aceleradores de IA en el grupo de recursos, también se muestra la información de monitoreo de GPU y NPU.

Figura 2-12 Consulta de vistas de recursos

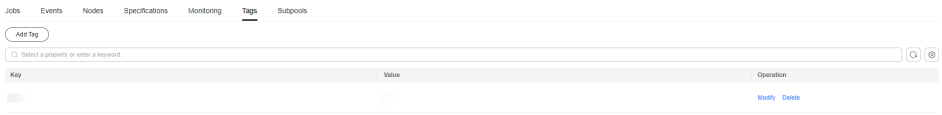


Consulta de etiquetas

Puede agregar etiquetas a un grupo de recursos para una búsqueda rápida.

En la página de detalles del grupo de recursos, haga clic en **Tags**. Puede ver, agregar, modificar y eliminar etiquetas de un grupo de recursos. Para obtener más información sobre cómo utilizar etiquetas, véase **¿Cómo utiliza ModelArts las etiquetas para gestionar recursos por grupo?**

Figura 2-13 Etiquetas



NOTA

Puede agregar hasta 20 etiquetas.

2.4 Cambio de tamaño de un grupo de recursos

Descripción

La demanda de recursos en un grupo de recursos dedicado puede cambiar debido a los cambios de los servicios de desarrollo de IA. En este caso, ModelArts puede cambiar el tamaño del grupo de recursos dedicado.

- Puede agregar nodos para las variantes existentes en el grupo de recursos.
- Puede eliminar nodos de las variantes existentes en el grupo de recursos.

NOTA

Antes de reducir un grupo de recursos, asegúrese de que no haya servicios ejecutándose en el grupo. Alternativamente, vaya a la página de detalles del grupo de recursos, elimine los nodos donde no se están ejecutando servicios para reducir el grupo.

Restricciones

- Solo se puede cambiar el tamaño de los grupos de recursos dedicados en estado **Running**.
- Al reducir un grupo de recursos dedicado, el número de variantes o nodos de una variante no se puede reducir a 0.

Cambio de tamaño de un grupo de recursos dedicado

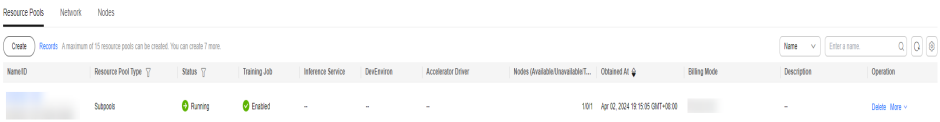
Puede cambiar el tamaño de un grupo de recursos de cualquiera de las siguientes maneras:

- Ajuste de la cantidad de nodos de las especificaciones existentes
- Cambio de tamaño del espacio del motor de contenedor

1. Inicie sesión en la consola de gestión de ModelArts. En el panel de navegación, seleccione **Dedicated Resource Pools** > **Elastic Cluster**.

> **NOTA**
>
> Un grupo de recursos se suspende cuando se migra de la versión anterior a la nueva. No puede ajustar la capacidad de dicho grupo de recursos ni darse de baja de él.

Figura 2-14 Grupos de recursos



2. Agregue o elimine los nodos.

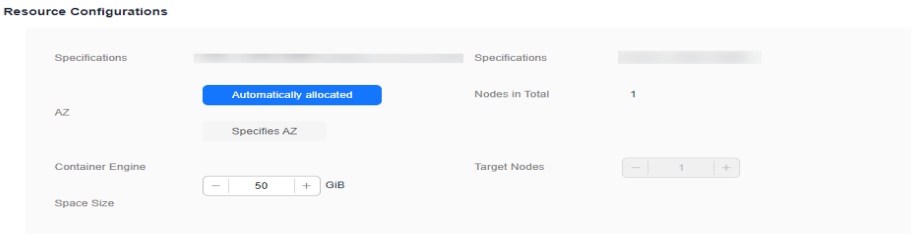
   Haga clic en **Adjust Capacity** en la columna **Operation** del grupo de recursos de destino.

   En el área **Resource Configurations**, configure **AZ** como **Automatically allocated** o **Specifies AZ**. Haga clic en **Submit** y luego en **OK** para guardar los cambios.

   – Si **AZ** se establece en **Automatically allocated**, puede aumentar o disminuir el número de nodos para escalar horizontalmente o en el grupo de recursos. Después del ajuste, los nodos se asignan automáticamente a las AZ.

   – Si selecciona **Specifies AZ**, puede asignar nodos a diferentes AZ.

Figura 2-15 Configuración de recursos



3. Cambie el tamaño del espacio del motor de contenedor.

   Si necesita un motor de contenedor más grande, realice cualquiera de las siguientes operaciones:

   – Para recursos nuevos, puede especificar el espacio del motor de contenedor al crear un grupo de recursos. Para obtener más información, véase las configuraciones avanzadas de **Specification Management** en **Creación de un grupo de recursos**.

   – Para los recursos existentes, se puede modificar el espacio del motor de contenedor.

     ■ Método 1: Haga clic en el grupo de recursos de destino para ver sus detalles. Haga clic en la pestaña **Specifications**, busque las especificaciones de destino y haga clic en **Change the contenedor engine space size** en la columna **Operation**.

     ■ Método 2: Busque el grupo de recursos de destino y haga clic en **Adjust Capacity** en la columna **Operation**.

AVISO

El cambio de tamaño del espacio del motor de contenedor solo se aplica a los nodos nuevos. Además, dockerBaseSize puede variar entre los nodos de esta variante dentro del grupo de recursos. En consecuencia, esto puede conducir a discrepancias en el estado de las tareas distribuidas entre los diferentes nodos.

Figura 2-16 Cambio de tamaño del espacio del motor de contenedor (ficha **Specifications**)

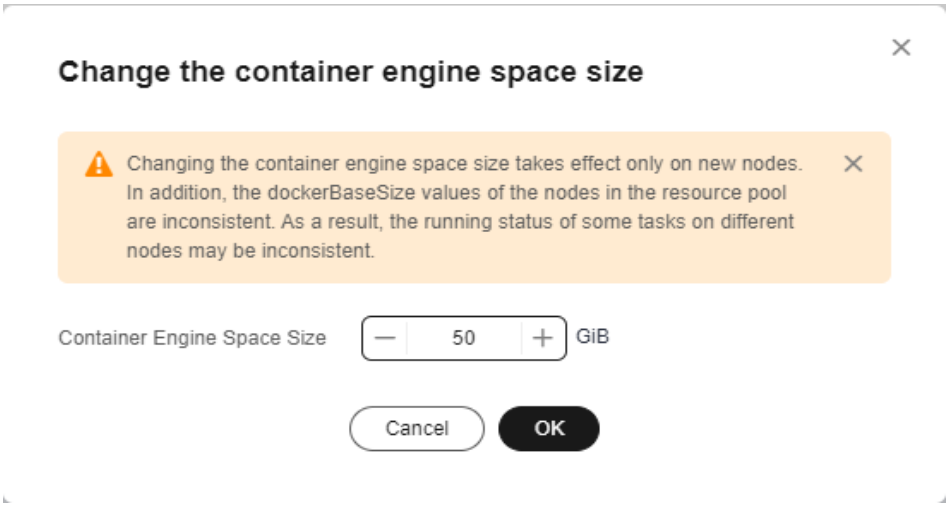
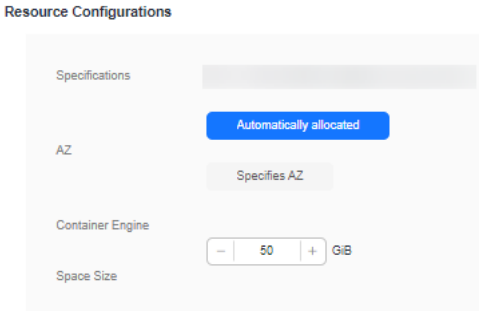


Figura 2-17 Cambio de tamaño del espacio del motor contenedor (página **Resize**)



2.5 Establecimiento de una política de renovación

Descripción

ModelArts permite realizar las siguientes operaciones para grupos de recursos anuales/mensuales:

- Habilitar la renovación automática.
- Modificar la configuración de la renovación automática.
- Renuévelos manualmente.

Restricciones

El grupo de recursos dedicado de destino debe estar ejecutándose.

Procedimiento

1. Inicie sesión en la consola de ModelArts. En el panel de navegación, seleccione **Dedicated Resource Pools** > **Elastic Cluster**.
2. En la lista de grupos de recursos, seleccione **More** > **Set Renewal Policy** en la columna **Operation** del grupo de recursos de destino.
3. En el cuadro de diálogo que aparece, haga clic en **OK**. Verá la página **Renewals** del centro de facturación.
4. Establezca la política de renovación.
   - Para habilitar la renovación automática de un grupo de recursos anual/mensual, haga clic en la pestaña **Manual Renewals**, busque el grupo de recursos de destino y seleccione **More** > **Enable Auto-Renewal** en la columna **Operation**.
   - Para modificar la configuración de renovación automática de un grupo de recursos anual/mensual, haga clic en la pestaña **Auto Renewals**, busque el grupo de recursos de destino y elija **More** > **Modify Auto-Renew** en la columna **Operation** para modificar la configuración de renovación automática, como el modo y la duración de la renovación y número de las renovaciones.
   - Para renovar manualmente un grupo de recursos anual/mensual, búsquelo y haga clic en **Renew** en la columna **Operation**.

2.6 Modificación de la política de caducidad

Descripción

ModelArts le permite cambiar la política de caducidad de un grupo de recursos anual/mensual a pago por uso o no renovación después de la caducidad.

Restricciones

El grupo de recursos dedicado de destino debe estar ejecutándose.

Procedimiento

1. Inicie sesión en la consola de ModelArts. En el panel de navegación, seleccione **Dedicated Resource Pools** > **Elastic Cluster**.
2. En la lista de grupos de recursos, seleccione **More** > **Change Billing Mode** en la columna **Operation** del grupo de recursos de destino.
3. En el cuadro de diálogo que aparece, haga clic en **OK**. Verá la página **Renewals** del centro de facturación.
4. Modifique la política de caducidad.
   - Si no se ha habilitado la renovación automática para el grupo de recursos de destino, haga clic en la ficha **Manual Renewals**, y seleccione **More** > **Change to Pay-per-Use After Expiration** o **More** > **Cancel Renewal** en la columna **Operation** del grupo de recursos de destino.

– Si se ha habilitado la renovación automática para el grupo de recursos de destino, haga clic en la pestaña **Auto Renewals** y elija **More** > **Cancel Renewal** en la columna **Operation** del grupo de recursos de destino.

2.7 Migración del espacio de trabajo

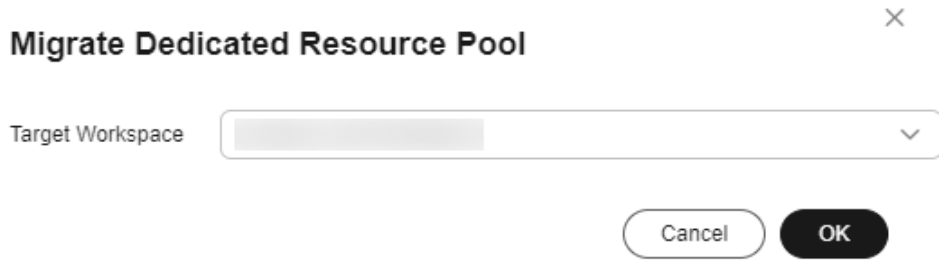
Contexto

El espacio de trabajo de un grupo de recursos dedicado está asociada con un proyecto de empresa, que implica la recopilación de facturas. ModelArts proporciona espacios de trabajo para aislar los permisos de operación de recursos de distintos usuarios de IAM. La migración de espacios de trabajo incluye la migración de grupos de recursos y la migración de redes. Para obtener más detalles, véase las siguientes secciones.

Migración del espacio de trabajo para un grupo de recursos

1. Inicie sesión en la consola de gestión de ModelArts. En el panel de navegación, seleccione **Dedicated Resource Pools** > **Elastic Cluster**.
2. En la lista de grupos de recursos, seleccione **More** > **Migrate Workspace** en la columna **Operation** del grupo de recursos de destino.
3. En el cuadro de diálogo **Migrate Dedicated Resource Pool** que aparece, seleccione el espacio de trabajo de destino y haga clic en **OK**.

Figura 2-18 Migración del espacio de trabajo



Migración del espacio de trabajo de una red

1. Inicie sesión en la consola de gestión de ModelArts. En el panel de navegación, seleccione **Dedicated Resource Pools** > **Elastic Cluster**. Luego, haga clic en la pestaña **Network**.
2. En la lista de redes, seleccione **More** > **Migrate Workspace** en la columna **Operation** de la red de destino.
3. En el cuadro de diálogo que aparece, seleccione el espacio de trabajo de destino y haga clic en **OK**.

Figura 2-19 Migración del espacio de trabajo

2.8 Cambio de tipos de trabajos soportados por un grupo de recursos

Descripción

ModelArts admite muchos tipos de trabajos. Algunos de ellos pueden ejecutarse en grupos de recursos dedicados, incluidos trabajos de entrenamiento, servicios de inferencia y entornos de desarrollo de notebook.

Puede cambiar los tipos de trabajo admitidos por un grupo de recursos dedicado. Las opciones disponibles para **Job Type** son **Training Job**, **Inference Service** y **DevEnviron**.

Solo los tipos de trabajos seleccionados se pueden entregar al grupo de recursos dedicado correspondiente.

ATENCIÓN

Para soportar diferentes tipos de trabajos, se realizan diferentes operaciones en el backend, como la instalación de componentes y la configuración del entorno de red. Algunas operaciones utilizan recursos del grupo de recursos. Como resultado, los recursos disponibles para usted disminuyen. Por lo tanto, seleccione solo los tipos de trabajo que necesita para evitar el desperdicio de recursos.

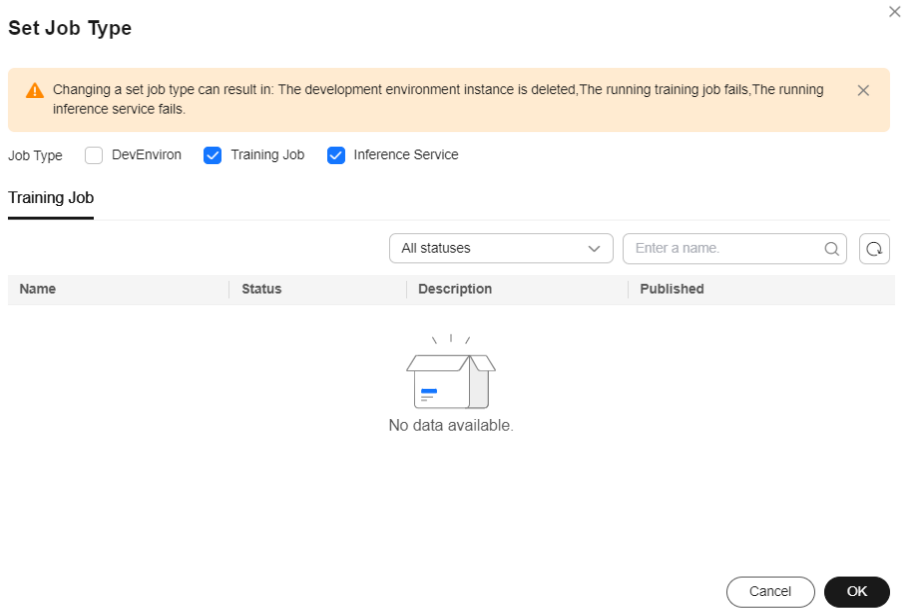
Restricciones

El grupo de recursos dedicado de destino debe estar ejecutándose.

Procedimiento

1. Inicie sesión en la consola de gestión de ModelArts. En el panel de navegación, seleccione **Dedicated Resource Pools** > **Elastic Cluster**.
2. En la columna **Operation** de un grupo de recursos, seleccione **More** > **Set Job Type**.
3. En el cuadro de diálogo **Set Job Type**, seleccione los tipos de trabajo.

Figura 2-20 Establecer el tipo de trabajo



4. Haga clic en **OK**.

2.9 Actualización de un controlador de grupo de recursos

Descripción

Si se utilizan GPU o recursos de Ascend en un grupo de recursos dedicado, es posible que deba personalizar los controladores de GPU o de Ascend. ModelArts le permite actualizar los controladores de GPU o de Ascend de sus grupos de recursos dedicados.

Hay dos modos de actualización de controlador: actualización segura y actualización forzosa.

NOTA

- Actualización segura: los servicios en ejecución no se ven afectados. Una vez iniciada la actualización, los nodos se aíslan (no se pueden entregar nuevos trabajos). Una vez completadas las tareas existentes en los nodos, se realiza la actualización. La actualización segura puede tardar mucho tiempo porque primero se deben completar los trabajos.
- Actualización forzosa: los controladores se actualizan directamente, independientemente de si hay trabajos en ejecución.

Restricciones

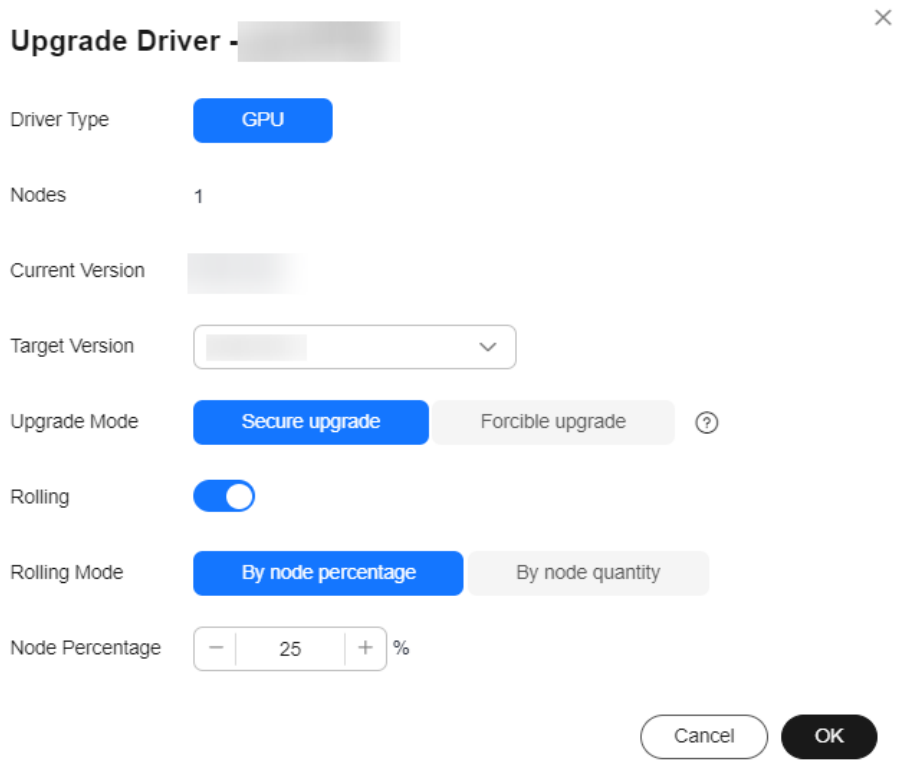
- El grupo de recursos dedicado de destino debe estar ejecutándose y el grupo de recursos contiene los recursos de GPU o de Ascend.
- Para un grupo de recursos lógico, el controlador solo se puede actualizar después de habilitar la vinculación de nodos. Para habilitar la vinculación de nodos, envíe un ticket de servicio para comunicarse con los ingenieros de Huawei.

Actualización del controlador

1. Inicie sesión en la consola de gestión de ModelArts. En el panel de navegación, seleccione **Dedicated Resource Pools** > **Elastic Cluster**.

2. En la columna **Operation** del grupo de recursos de destino, seleccione **More** > **Upgrade Driver**.
3. En el cuadro de diálogo **Upgrade Driver**, se muestran el tipo de controlador, la cantidad de nodos, la versión actual, la versión de destino y el modo de actualización del grupo de recursos dedicados.
   - **Target Version**: seleccione una versión de controlador de destino en la lista desplegable.
   - **Upgrade Mode**: seleccione **Secure upgrade** o **Forcible upgrade**.
   - **Rolling Mode**: Una vez habilitado, puede actualizar el controlador en modo de rotación. Actualmente, se admite la rotación por porcentaje de nodo y por cantidad de nodo. Si se selecciona **By node percentage**, el número de nodos que se van a actualizar en cada lote es la relación de nodos multiplicada por el número total de nodos del grupo de recursos. Si se selecciona **By node quantity**, el número de nodos que se van a actualizar en cada lote es el configurado.

Figura 2-21 Actualizar un controlador



4. Haga clic en **OK** para iniciar la actualización del controlador.

2.10 Eliminación de un grupo de recursos

Si ya no se necesita un grupo de recursos dedicado para el desarrollo de servicios de IA, puede eliminar el grupo de recursos para lanzar recursos.

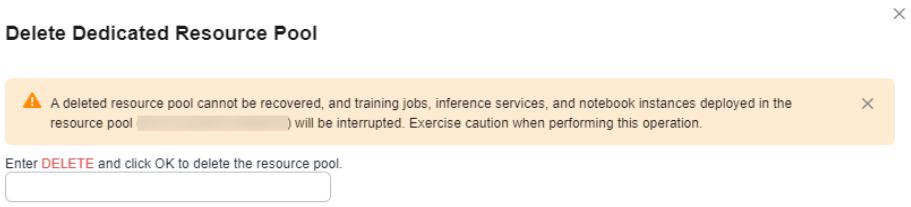
NOTA

Después de eliminar un grupo de recursos dedicado, los entornos de desarrollo, los trabajos de entrenamiento y los servicios de inferencia que dependen del grupo de recursos no están disponibles. No se puede restaurar un grupo de recursos dedicado después de haber sido eliminado.

1. Inicie sesión en la consola de gestión de ModelArts. En el panel de navegación, seleccione **Dedicated Resource Pools** > **Elastic Cluster**.
2. Busque la fila que contiene el grupo de recursos de destino y seleccione **More** > **Delete** en la columna **Operation**.
3. En el cuadro de diálogo **Delete Dedicated Resource Pool**, escriba **DELETE** en el cuadro de texto y haga clic en **OK**.

   Puede alternar entre pestañas en la página de detalles para ver los trabajos de entrenamiento y las instancias de notebook creadas con el grupo de recursos y los servicios de inferencia desplegados en el grupo de recursos.

Figura 2-22 Eliminación de un grupo de recursos



2.11 Estado anormal de un grupo de recursos dedicado

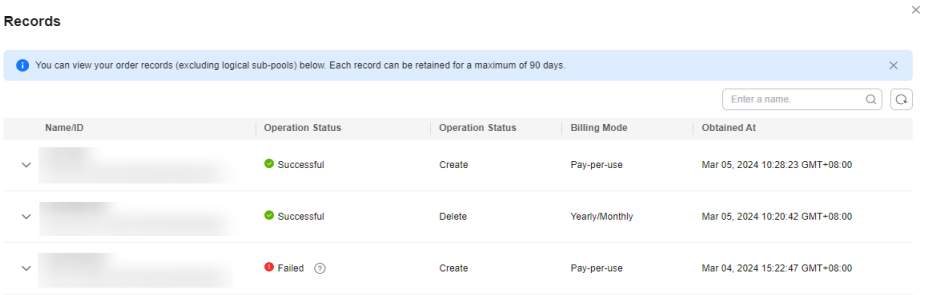
Límite de cuota de recursos

Cuando utiliza un grupo de recursos dedicado (por ejemplo, escalar recursos, crear una VPC, crear una VPC y una subred o interconectar una VPC), si el sistema muestra un mensaje que indica que la cuota de recursos es limitada, **envíe un ticket de servicio**.

Error de creación/cambio

1. Inicie sesión en la consola de gestión de ModelArts. En el panel de navegación, seleccione **Dedicated Resource Pools** > **Elastic Cluster**.
2. Haga clic en **Records** a la derecha de **Create**. En el cuadro de diálogo **Records**, consulte los registros de tareas fallidas.

Figura 2-23 Error al crear un grupo de recursos



3. Pase el cursor sobre para ver la causa de las fallas de las tareas.

   NOTA

   De forma predeterminada, los registros de tareas fallidas se ordenan por tiempo de aplicación. Se pueden mostrar y conservar un máximo de 500 registros de tareas fallidas durante tres días.

Localización del nodo defectuoso

ModelArts agregará una mancha en un nodo defectuoso de K8S detectado para que los trabajos no se vean afectados ni programados en el nodo manchado. La siguiente tabla enumera las fallas que se pueden detectar. Puede localizar la falla consultando el código de aislamiento y el método de detección.

Tabla 2-3 Código de aislamiento

Código de aislamiento	Categoría	Subcategoría	Descripción	Método de detección
A050101	GPU	Memoria de la GPU	Existe un error de ECC de la GPU.	Ejecute el comando **nvidia-smi -a** y verifique si **Pending Page Blacklist** es **Yes** o si el valor de **multi-bit Register File** es mayor que 0. Para las GPU de Ampere, verifique si existe el siguiente contenido: ● Error de SRAM incorregible ● Reasignación de registros de fallas ● Xid 95 sucesos en **dmsg** (Para obtener más detalles, véase **Gestión de errores de memoria de GPU de NVIDIA**) La arquitectura de Ampere tiene los siguientes niveles de errores de memoria de GPU: ● L1: Se trata de errores de ECC de un solo bit que se pueden corregir. No afectan los servicios en ejecución. Para verificar estos errores, ejecute el comando **nvidia-smi -a** y busque **Volatile Correctable**. ● L2: Se trata de errores de ECC de varios bits que no se pueden corregir. Provocan fallas en los servicios en ejecución y requieren un reinicio del proceso para recuperarse. Para verificar estos errores, ejecute el comando **nvidia-smi -a** y busque **Volatile Uncorrectable**. ● L3: Estos son errores no suprimidos y pueden afectar a otros servicios. Requieren un restablecimiento de la tarjeta o un reinicio del nodo para borrarlos. Para comprobar estos errores, busque los eventos de Xid que contengan el número 95. (Los registros pendientes de reasignación son solo para referencia. Es necesario restablecer las tarjetas cuando el servicio está inactivo para activar el proceso de reasignación.)

Código de aislamiento	Categoría	Subcategoría	Descripción	Método de detección
				● L4: Estos son errores que requieren un reemplazo de tarjeta. Para verificar estos errores, busque el campo **SRAM Uncorrectable** que sea mayor que 4 o el campo **Remapped Failed** que no sea cero.
A050102	GPU	Otra	El resultado de **nvidia-smi** contiene ERR.	Ejecute **nvidia-smi -a** y verifique si el resultado contiene ERR. Por lo general, el hardware, como la fuente de alimentación o el ventilador, presenta fallas.
A050103	GPU	Otra	Se agota el tiempo de espera de ejecución de **nvidia-smi** o no existe.	Verifique que el código de salida de **nvidia-smi** no sea **0**.
A050104	GPU	Memoria de la GPU	El error ECC se ha producido 64 veces.	Ejecute el comando **nvidia-smi -a**, localice **Retired Pages** y verifique si la suma de **Single Bit** y **Double Bit** es mayor que 64.
A050148	GPU	Otra	Se produjo una alarma de infoROM.	Ejecute el comando **nvidia-smi** y verifique si el resultado contiene la alarma "infoROM is corrupted".
A050109	GPU	Otra	Otros errores de GPU	Verifique si existe otro error de GPU. Por lo general, el hardware presenta fallas. Póngase en contacto con el ingeniero técnico.
A050147	IB	Enlace	El estado de la NIC de IB es anormal.	Ejecute el comando **ibstat** y verifique si la NIC no se encuentra en estado activo.
A050121	NPU	Otra	El DCMI de la NPU detecta una excepción de controlador.	El entorno del controlador de NPU no es normal.
A050122	NPU	Otra	El dispositivo de DCMI de la NPU no funciona correctamente.	El dispositivo de NPU no funciona correctamente. La interfaz de Ascend DCMI devuelve una alarma importante o urgente.

Código de aislamiento	Categoría	Subcategoría	Descripción	Método de detección
A050123	NPU	Enlace	La red de DCMI de la NPU no funciona correctamente.	La conexión de red de la NPU no es normal.
A050129	NPU	Otra	Otros errores de NPU	Verifique si existe otro error de NPU. No puede rectificar la falla. Póngase en contacto con el ingeniero técnico.
A050149	NPU	Enlace	Verifique si el puerto de red de la herramienta hccn está desconectado intermitentemente.	La red de NPU es inestable y se desconecta intermitentemente. Ejecute el comando **hccn_tool-i ${device_id} -link_stat -g** y la red se desconectará más de cinco veces en 24 horas.
A050951	NPU	Memoria de la GPU	La cantidad de ECC de NPU alcanza el umbral de mantenimiento.	El valor del recuento de páginas aisladas de doble bit de HBM de la NPU es mayor o igual que 64.
A050146	Runtime	Otra	NTP no es normal.	El servicio ntpd o chronyd no funciona correctamente.
A050202	Runtime	Otra	El nodo no está listo.	El nodo no está disponible. El nodo de K8S contiene una de las siguientes manchas: ● node.kubernetes.io/unreachable ● node.kubernetes.io/not-ready
A050203	Runtime	Desconexión	La cantidad de tarjetas de IA normales no coincide con la capacidad real.	La GPU o la NPU están desconectadas.
A050206	Runtime	Otra	El disco duro de Kubelet es de solo lectura.	El directorio **/mnt/paas/kubernetes/kubelet** es de solo lectura.
A050801	Gestión de nodos	O&M del nodo	El recurso está reservado.	El nodo se marca como nodo en standby y contiene una mancha.
A050802	Gestión de nodos	O&M del nodo	Se produjo un error desconocido.	El nodo está marcado con una mancha desconocida.

Código de aislamiento	Categoría	Subcategoría	Descripción	Método de detección
A200001	Gestión de nodos	Actualización de controladores	Se está actualizando la GPU.	Se está actualizando la GPU.
A200002	Gestión de nodos	Actualización de controladores	Se está actualizando la NPU.	Se está actualizando la NPU.
A200008	Gestión de nodos	Admisión de nodo	Se está examinando la admisión.	Se está examinando la admisión, incluida la verificación de la configuración básica del nodo y la verificación simple del servicio.
A050933	Gestión de nodos	Failover de tolerancia a fallas	Se migrará el servicio de Failover en el nodo contaminado.	Se migrará el servicio de Failover en el nodo contaminado.
A050931	Entrenamiento de kit de herramientas	Contenedor de comprobación previa	Se detecta un error de GPU en el contenedor de comprobación previa.	Se detecta un error de GPU en el contenedor de comprobación previa.
A050932	Entrenamiento de kit de herramientas	Contenedor de comprobación previa	Se detecta un error de IB en el contenedor de comprobación previa.	Se detecta un error de IB en el contenedor de comprobación previa.

2.12 Red de ModelArts

Red de ModelArts y VPC

Las redes de ModelArts son respaldadas por las VPC y que se utilizan para interconectar nodos en un grupo de recursos de ModelArts. Solo puede configurar el nombre y el bloque

CIDR para una red. Para garantizar que no haya ningún segmento de direcciones IP en el bloque de CIDR superpuesto con el de la VPC a la que se va a acceder, hay varios bloques de CIDR disponibles para que los seleccione.

Una VPC proporciona una red virtual aislada lógicamente para sus instancias. Puede configurar y gestionar la red según sea necesario. La VPC proporciona redes virtuales lógicamente aisladas, configurables y gestionables para servidores en la nube, contenedores en la nube y bases de datos en la nube. Le ayuda a mejorar la seguridad del servicio en la nube y a simplificar el despliegue de la red.

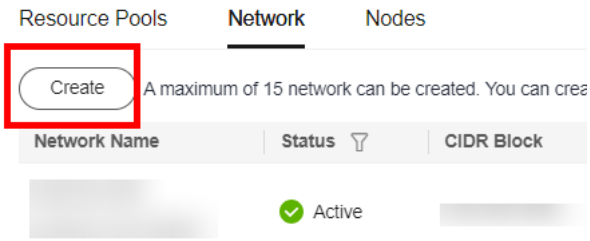
Requisitos previos

- Una VPC está disponible.
- Hay una subred disponible.

Creación de una red

1. Inicie sesión en la consola de gestión de ModelArts. En el panel de navegación, seleccione **Dedicated Resource Pools** > **Elastic Cluster**.
2. Haga clic en **Network** y luego en **Create**.

   Figura 2-24 Lista de redes

3. En el cuadro de diálogo **Create Network**, configure los parámetros.
   - **Network Name**: nombre personalizado
   - **CIDR Block**: Puede seleccionar **Preset** o **Custom**.

   NOTA

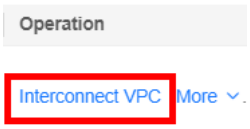
   - Cada usuario puede crear un máximo de 15 redes.
   - Asegúrese de que no haya ningún segmento de dirección IP en el bloque de CIDR que se solape con el de la VPC a la que se va a acceder. El bloque de CIDR no se puede cambiar una vez creada la red. Los posibles bloques de CIDR de conflicto son los siguientes:
     - Su bloqueo de CIDR de VPC
     - Bloque de CIDR de contenedores (consistentemente en 172.16.0.0/16)
     - Bloque del CIDR del servicio (consistentemente en 10.247.0.0/16)
4. Confirme las configuraciones y haga clic en **OK**.

(Opcional) Interconexión de una VPC con una red de ModelArts

La interconexión de VPC le permite utilizar recursos en todas las VPC, lo que mejora la utilización de los recursos.

1. En la página **Network**, haga clic en **Interconnect VPC** en la columna **Operation** de la red de destino.

Figura 2-25 Interconectar la VPC



2. En el cuadro de diálogo que aparece en pantalla, haga clic en el botón situado a la derecha de **Interconnect VPC** y seleccione una VPC y una subred disponibles en las listas desplegables.

   NOTA

   La red del otro extremo a interconectar no puede superponerse con el bloque de CIDR actual.

Figura 2-26 Parámetros para interconectar una VPC con una red

- Si no hay ninguna VPC disponible, haga clic en **Create VPC** a la derecha para crear una VPC.
- Si no hay ninguna subred disponible, haga clic en **Create Subnet** a la derecha para crear una subred.
- Se pueden interconectar varias subredes en una VPC. Puede hacer clic en + para agregar hasta 10 subredes.

Habilitación de un grupo de recursos dedicado para acceder a Internet

Para habilitar un grupo de recursos dedicado para acceder a Internet, siga estos pasos:

Paso 1 Interconecte una VPC. Para obtener más detalles, véase **(Opcional) Interconexión de una VPC con una red de ModelArts**.

Paso 2 Para obtener detalles sobre cómo configurar un servidor de SNAT para una VPC, véase **Configuración de un servidor de SNAT**.

----Fin

Eliminación de una red

Si ya no se necesita una red para el desarrollo de servicios de IA, puede eliminarla.

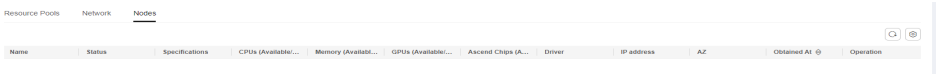
1. Vaya a la ficha **Network** y haga clic en **Delete** en la columna **Operation** de una red.

2. Confirme la información y haga clic en **OK**.

2.13 Nodos de ModelArts

Los nodos que no son gestionados por el grupo de recursos se consideran nodos libres. Para ver la información sobre los nodos libres, inicie sesión en la consola de gestión de ModelArts, seleccione **Dedicated Resource Pools** > **Elastic Cluster** y haga clic en la pestaña **Nodes**.

Figura 2-27 Nodos



Lance los recursos de nodos libres de acuerdo con el siguiente contenido:

- Para un nodo de pago por uso, haga clic en **Delete** en la columna **Operation**.
- Para un nodo anual/mensual cuyos recursos no han expirado, haga clic en **Unsubscribe** en la columna **Operation**.
- Para un nodo anual/mensual cuyos recursos han expirado (en el período de gracia), haga clic en **Release** en la columna **Operation**.

Si el botón de eliminación está disponible para un nodo anual/mensual, haga clic en el botón para eliminar el nodo.

NOTA

Las operaciones de eliminación, cancelación de suscripción y lanzamiento no se pueden deshacer. Realice esta operación con precaución.

3 Logs de auditoría

3.1 Operaciones de clave registradas por CTS

Con CTS, puede obtener operaciones asociadas con ModelArts para consultas, auditorías y operaciones de backtrack posteriores.

Requisitos previos

Se ha habilitado CTS.

Operaciones de gestión de datos clave rastreadas por CTS

Tabla 3-1 Operaciones de gestión de datos clave rastreadas por CTS

Operación	Tipo de recurso	Rastro
Creación de un conjunto de datos	Dataset	createDataset
Eliminación de un conjunto de datos	Dataset	deleteDataset
Actualización de un conjunto de datos	Dataset	updateDataset
Publicación de una versión de conjunto de datos	Dataset	publishDatasetVersion
Eliminación de una versión de conjunto de datos	Dataset	deleteDatasetVersion
Sincronización del origen de datos	Dataset	syncDataSource

Operación	Tipo de recurso	Rastro
Exportación de un conjunto de datos	Dataset	exportDataFromDataset
Creación de una tarea de etiquetado automático	Dataset	createAutoLabelingTask
Creación de una tarea de agrupación automática	Dataset	createAutoGroupingTask
Creación de una tarea de despliegue automático	Dataset	createAutoDeployTask
Importación de muestras a un conjunto de datos	Dataset	importSamplesToDataset
Creación de una etiqueta de conjunto de datos	Dataset	createLabel
Actualización de una etiqueta de conjunto de datos	Dataset	updateLabel
Eliminación de una etiqueta de conjunto de datos	Dataset	deleteLabel
Eliminación de una etiqueta de conjunto de datos y sus muestras etiquetadas	Dataset	deleteLabelWithSamples
Adición de muestras	Dataset	uploadSamples
Eliminación de muestras	Dataset	deleteSamples
Detención de una tarea de etiquetado automático	Dataset	stopTask
Creación de una tarea de etiquetado de equipo	Dataset	createWorkforceTask
Eliminación de una tarea de etiquetado de equipo	Dataset	deleteWorkforceTask
Inicio de la aceptación de una tarea de etiquetado de equipo	Dataset	startWorkforceSampling-Task
Aprobar, rechazar o cancelar la aceptación de una tarea de etiquetado de equipo	Dataset	updateWorkforceSampling-Task
Enviar comentarios de revisión de muestra para una tarea de aceptación	Dataset	acceptSamples
Adición de una etiqueta a una muestra	Dataset	updateSamples

Operación	Tipo de recurso	Rastro
Envío de un correo electrónico a los miembros del equipo	Dataset	sendEmails
Comenzar una tarea de etiquetado de equipo como gerente de equipo	Dataset	startWorkforceTask
Actualización de una tarea de etiquetado de equipo	Dataset	updateWorkforceTask
Adición de una etiqueta a una muestra con etiqueta de equipo	Dataset	updateWorkforceTaskSamples
Revisión de los resultados del etiquetado del equipo	Dataset	reviewSamples
Creación de un miembro del equipo de etiquetado	Workforce	createWorker
Actualización de miembros del equipo de etiquetado	Workforce	updateWorker
Eliminación de un miembro del equipo de etiquetado	Workforce	deleteWorker
Eliminación de miembros del equipo de etiquetado por un lote	Workforce	batchDeleteWorker
Creación de un equipo de etiquetado	Workforce	createWorkforce
Actualización de un equipo de etiquetado	Workforce	updateWorkforce
Eliminación de un equipo de etiquetado	Workforce	deleteWorkforce
Creación automática de una delegación de IAM	IAM	createAgency
Iniciar sesión en la consola de etiquetado como miembro de etiquetado del equipo	labelConsoleWorker	workerLoginLabelConsole
Cerrar sesión en la consola de etiquetado como miembro de etiquetado del equipo	labelConsoleWorker	workerLogOutLabelConsole

Operación	Tipo de recurso	Rastro
Cambio de la contraseña para iniciar sesión en la consola de etiquetado como miembro de etiquetado del equipo	labelConsoleWorker	workerChangePassword
Solucionar el problema de la pérdida de la contraseña para iniciar sesión en la consola de etiquetado como miembro del equipo de etiquetado.	labelConsoleWorker	workerForgetPassword
Restablecimiento de la contraseña para iniciar sesión en la consola de etiquetado con la URL como miembro de etiquetado del equipo	labelConsoleWorker	workerResetPassword

Operaciones clave de DevEnviron rastreadas por CTS

Tabla 3-2 Operaciones clave de DevEnviron rastreadas por CTS

Operación	Tipo de recurso	Rastro
Creación de una instancia de notebook	Notebook	createNotebook
Supresión de una instancia de notebook	Notebook	deleteNotebook
Apertura de una instancia de notebook	Notebook	openNotebook
Inicio de una instancia de notebook	Notebook	startNotebook
Detención de una instancia de Notebook	Notebook	stopNotebook
Actualización de una instancia de notebook	Notebook	updateNotebook
Eliminación de un NotebookApp	NotebookApp	deleteNotebookApp
Cambio de especificaciones de CodeLab	NotebookApp	updateNotebookApp

Operaciones de trabajo de entrenamiento clave rastreadas por CTS

Tabla 3-3 Operaciones de trabajo de entrenamiento clave rastreadas por CTS

Operación	Tipo de recurso	Rastro
Creación de un trabajo de entrenamiento	ModelArtsTrainJob	createModelArtsTrainJob
Creación de una versión de trabajo de entrenamiento	ModelArtsTrainJob	createModelArtsTrainVersion
Detención de un trabajo de entrenamiento	ModelArtsTrainJob	stopModelArtsTrainVersion
Modificación de la descripción de un trabajo de entrenamiento	ModelArtsTrainJob	updateModelArtsTrainDesc
Eliminación de una versión de trabajo de entrenamiento	ModelArtsTrainJob	deleteModelArtsTrainVersion
Eliminación de un trabajo de entrenamiento	ModelArtsTrainJob	deleteModelArtsTrainJob
Configuración de trabajo de entrenamiento	ModelArtsTrainConfig	createModelArtsTrainConfig
Modificación de la configuración de un trabajo de entrenamiento	ModelArtsTrainConfig	updateModelArtsTrainConfig
Eliminación de una configuración de trabajo de entrenamiento	ModelArtsTrainConfig	deleteModelArtsTrainConfig
Creación de un trabajo de visualización	ModelArtsTensorboardJob	createModelArtsTensorboardJob
Eliminación de un trabajo de visualización	ModelArtsTensorboardJob	deleteModelArtsTensorboardJob
Modificación de la descripción de un trabajo de visualización	ModelArtsTensorboardJob	updateModelArtsTensorboardDesc
Detención de un trabajo de visualización	ModelArtsTensorboardJob	stopModelArtsTensorboardJob
Reinicio de un trabajo de visualización	ModelArtsTensorboardJob	restartModelArtsgTensorboardJob

Operaciones clave de gestión de aplicaciones de IA rastreadas por CTS

Tabla 3-4 Operaciones clave de gestión de aplicaciones de IA rastreadas por CTS

Operación	Tipo de recurso	Rastro
Creación de una aplicación de IA	Model	addModel
Actualización de una aplicación de IA	Model	updateModel
Eliminación de una aplicación de IA	Model	deleteModel
Creación de una tarea de conversión de modelo	Convert	addConvert
Actualización de una tarea de conversión de modelo	Convert	updateConvert
Eliminación de una tarea de conversión de modelo	Convert	deleteConvert

Operaciones de gestión de servicios clave rastreadas por CTS

Tabla 3-5 Operaciones de gestión de servicios clave rastreadas por CTS

Operación	Tipo de recurso	Rastro
Despliegue de un servicio	Service	addService
Eliminación de un servicio	Service	deleteService
Actualización de un servicio	Service	updateService
Inicio o detención de un servicio	Service	startOrStopService
Adición de una clave de acceso de usuario	Service	addAkSk
Eliminación de una clave de acceso de usuario	Service	deleteAkSk
Creación de un grupo de recursos dedicado	Cluster	createCluster
Eliminación de un grupo de recursos dedicado	Cluster	deleteCluster
Adición de un nodo a un grupo de recursos dedicado	Cluster	addClusterNode

Operación	Tipo de recurso	Rastro
Eliminación de un nodo de un grupo de recursos dedicado	Cluster	deleteClusterNode
Obtención de un resultado de la creación de un grupo de recursos dedicado	Cluster	createClusterResult

Operaciones clave de AI Gallery rastreadas por CTS

Tabla 3-6 Operaciones clave de AI Gallery rastreadas por CTS

Operación	Tipo de recurso	Rastro
Publicación de un activo	ModelArts_Market	create_content
Modificación de la información de activos	ModelArts_Market	modify_content
Publicación de una versión de activos	ModelArts_Market	add_version
Suscripción a un activo	ModelArts_Market	subscription_content
Extracción de un activo de los favoritos	ModelArts_Market	cancel_star_content
Dar Me gusta a un activo	ModelArts_Market	like_content
Cancelar Me gusta a un activo	ModelArts_Market	cancel_like_content
Publicación de una actividad	ModelArts_Market	publish_activity
Registro de una actividad	ModelArts_Market	regist_activity
Modificación de la información de usuario	ModelArts_Market	update_user

Operaciones de gestión de recursos clave rastreadas por CTS

Tabla 3-7 Operaciones de gestión de recursos clave rastreadas por CTS

Operación	Tipo de recurso	Rastro
Creación de un grupo de recursos	PoolV2	CreatePoolV2
Eliminación de un grupo de recursos	PoolV2	DeletePoolV2

Operación	Tipo de recurso	Rastro
Actualización de un grupo de recursos	PoolV2	UpdatePoolV2
Creación de una red	NetworksV1	CreateNetworksV1
Eliminación de una red	NetworksV1	DeleteNetworksV1
Actualización de una red	NetworksV1	UpdateNetworksV1

3.2 Consulta de logs de auditoria

Una vez habilitado CTS, el servicio inicia las operaciones de grabación relacionadas con ModelArts. La consola de gestión de CTS almacena los últimos siete días de registros de operación. Esta sección describe cómo consultar los registros de operación de los últimos siete días en la consola de gestión de CTS.

Procedimiento

1. Inicie sesión en la consola de gestión de CTS.
2. Haga clic en  en la esquina superior izquierda de la página y seleccione una región.
3. En el panel de navegación izquierdo, haga clic en **Trace List**.
4. Especifique los criterios de filtro utilizados para consultar trazas. Los siguientes cuatro criterios de filtro están disponibles:
   - **Trace Source**, **Resource Type** y **Search By**

     Seleccione un criterio de filtro en la lista desplegable.

     Si selecciona **Trace name** para **Search By**, también debe seleccionar un nombre de seguimiento específico.

     Si selecciona **Resource ID** para **Search By**, debe introducir un ID de recurso específico.

     Si selecciona **Resource name** para **Search By**, debe seleccionar o introducir un nombre de recurso específico.
   - **Operator**: Seleccione un operador específico (un usuario en lugar de una cuenta).
   - **Trace Status**: Las opciones disponibles incluyen **All trace statuses**, **normal**, **warning** e **incident**. Solo se puede habilitar una de ellas.
   - **Time Range**: Puede ver los rastros generados durante cualquier rango de tiempo de los últimos siete días.
5. Haga clic en ⌄ a la izquierda de un rastro para expandir sus detalles.
6. Haga clic en **View Trace** en la columna **Operation**. En el cuadro de diálogo **View Trace** mostrado, se muestran los detalles de la estructura de rastro.

   Para obtener más información sobre los campos clave de la estructura de rastro de CTS, consulte la **Guía de usuario de Cloud Trace Service**.

4 Recursos de monitoreo

4.1 Descripción general

Todas las métricas informadas por ModelArts se almacenan en AOM, lo que permite consumir métricas. Puede ver las alarmas de umbral de métrica y las alarmas reportadas en la consola de AOM o usar herramientas de visualización como Grafana para ver y analizar las alarmas. Grafana proporciona diferentes vistas y plantillas para el monitoreo, que le permiten ver el uso de recursos en tiempo real en los paneles de control claramente.

4.2 Uso de Grafana para consultar métricas de monitoreo de AOM

4.2.1 Procedimiento

Grafana admite varias vistas y plantillas de monitoreo, lo que satisface sus diversos requisitos. Después de agregar el origen de datos en Grafana, puede ver todas las métricas de monitoreo de ModelArts almacenadas en AOM con Grafana.

Para ver las métricas de supervisión de AOM mediante los complementos de Grafana, realice los siguientes pasos:

1. NOTA

   Puede instalar y configurar Grafana de cualquiera de las siguientes maneras: **Instalación y configuración de Grafana en Windows**, **Instalación y configuración de Grafana en Linux** y **Instalación y configuración de Grafana en una instancia de notebook**.

2. **Configuración de un origen de datos de Grafana**
3. **Uso de Grafana para configurar paneles y consultar datos de métrica**

4.2.2 Instalación y configuración de Grafana

4.2.2.1 Instalación y configuración de Grafana en Windows

Escenario de aplicación

Esta sección describe cómo instalar y configurar Grafana en un sistema operativo de Windows.

Procedimiento

1. Descargue el paquete de instalación de Grafana.

   Vaya al **enlace de descarga**, haga clic en **Download the installer** y espere hasta que la descarga se realice correctamente.

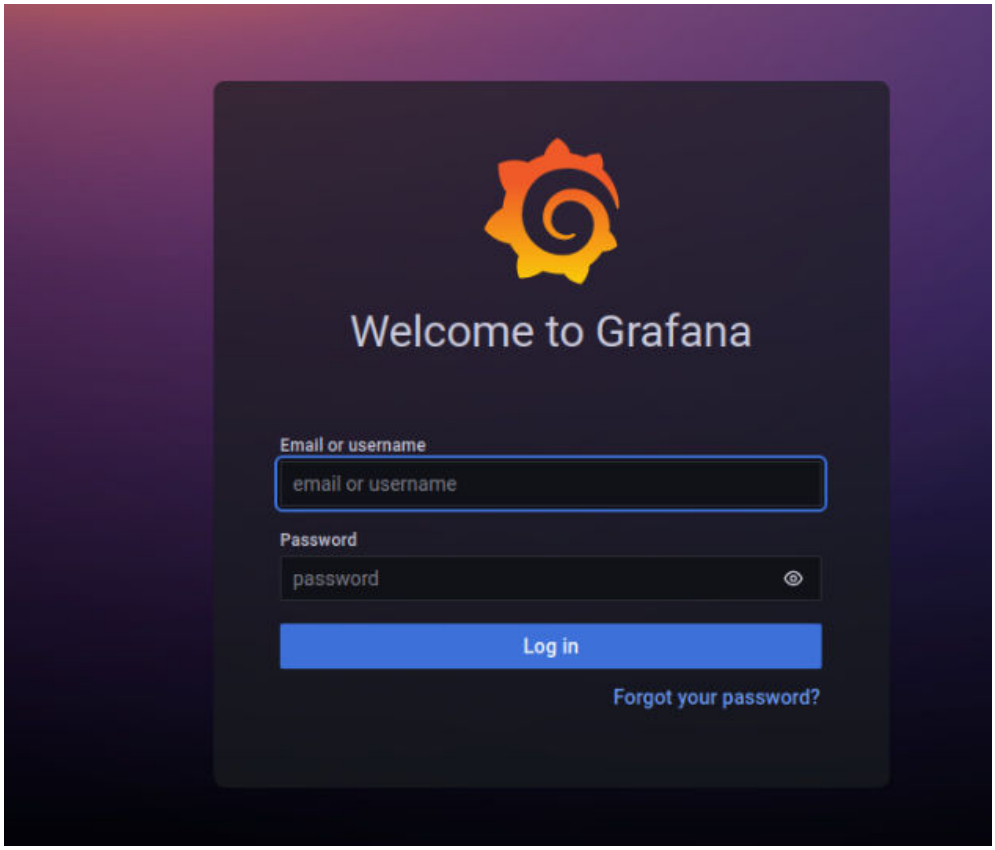
2. Instale Grafana.

   Haga doble clic en el paquete de instalación e instale Grafana como se indica.

3. En Windows Services Manager, habilite Grafana.

4. Inicie sesión en Grafana.

   Grafana se ejecuta en el puerto 3000 de forma predeterminada. Después de abrir http://localhost:3000, aparece la página de inicio de sesión de Grafana. El nombre de usuario y la contraseña predeterminados para el primer inicio de sesión es **admin**. Una vez que el inicio de sesión sea exitoso, cambie la contraseña según se le solicite.



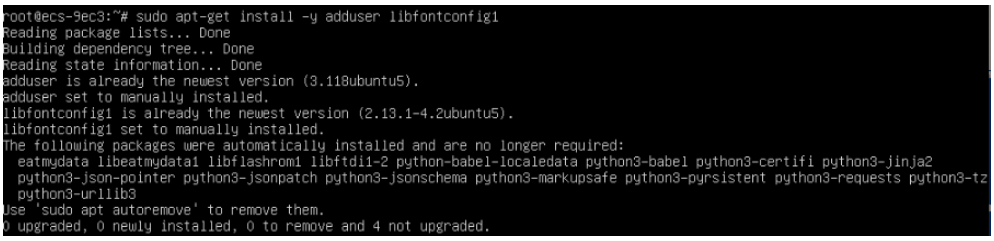
4.2.2.2 Instalación y configuración de Grafana en Linux

Requisitos previos

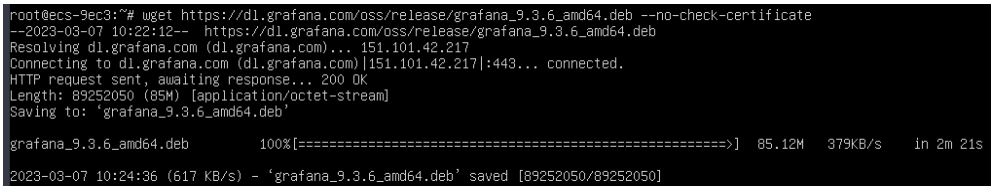
- Existe un servidor de Ubuntu accesible al Internet. En caso negativo, deben cumplirse las siguientes condiciones:
- Ha obtenido un ECS. (Se recomienda seleccionar 8 vCPU o superior, imagen de Ubuntu de la versión 22.04 y almacenamiento local de 100 GB). Para obtener más detalles, véase **Compra de un ECS**.
- Ha adquirido una EIP y la ha vinculado al ECS. Para obtener más detalles, véase **Asignación de una EIP y su vinculación a un ECS**.

Procedimiento

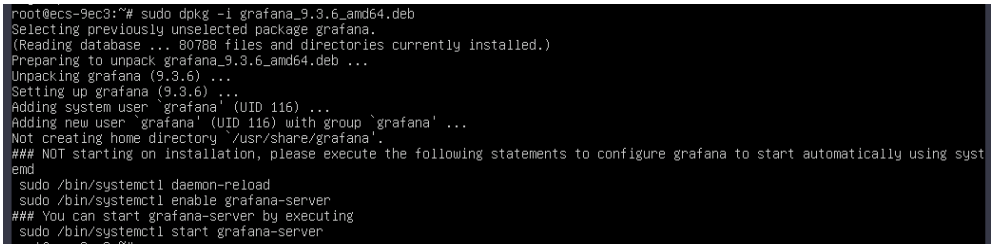
1. Inicie sesión en el ECS. Seleccione un método de inicio de sesión. Para obtener más detalles, véase .
2. Ejecute el siguiente comando para instalar libfontconfig1:
   ```
   sudo apt-get install -y adduser libfontconfig1
   ```
   La operación se realiza correctamente si se muestra la siguiente información:

3. Ejecute el siguiente comando para descargar el paquete de instalación de Grafana:
   ```
   wget https://dl.grafana.com/oss/release/grafana_9.3.6_amd64.deb --no-check-certificate
   ```
   Descarga completada:

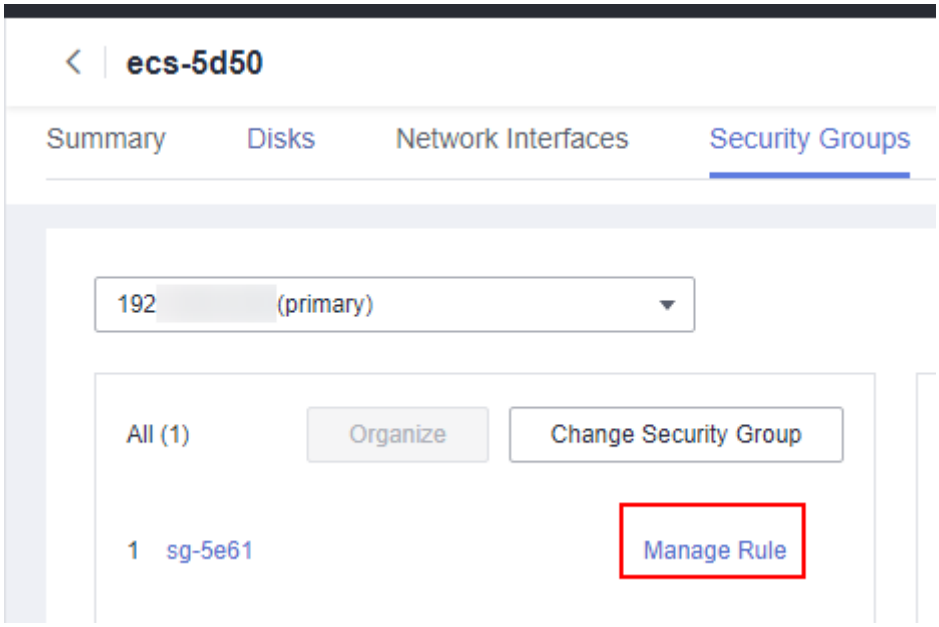
4. Ejecute el siguiente comando para instalar Grafana:
   ```
   sudo dpkg -i grafana_9.3.6_amd64.deb
   ```

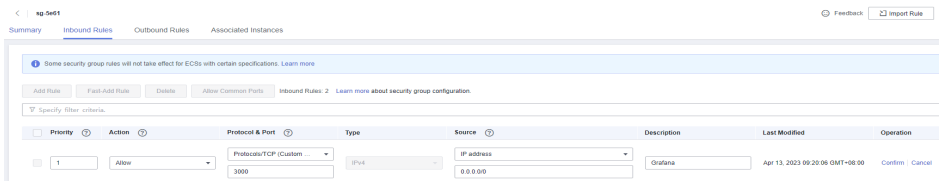
5. Ejecute el siguiente comando para iniciar Grafana:
   ```
   sudo /bin/systemctl start grafana-server
   ```
6. Acceda a las configuraciones de Grafana en su PC local.

Asegúrese de que se haya asociado una EIP al ECS y de que la configuración del **grupo de seguridad** sea correcta (se permite el tráfico entrante desde el puerto de TCP 3000 y todo el tráfico saliente). Proceso de configuración:

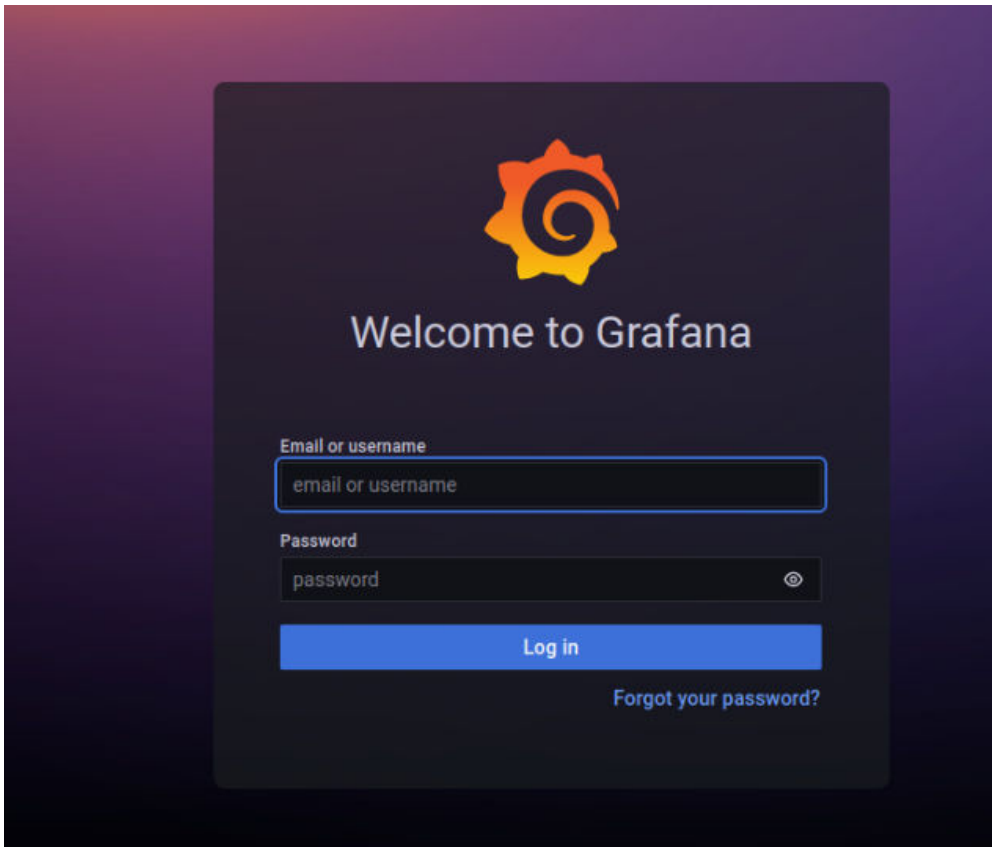
a. Haga clic en el nombre del ECS para ir a la página de detalles del ECS. Luego, haga clic en la ficha **Security Groups** y luego en **Manage Rule**.



b. Haga clic en **Inbound Rules** y permita el tráfico entrante desde el puerto de TCP 3000. De forma predeterminada, se permite todo el tráfico saliente.



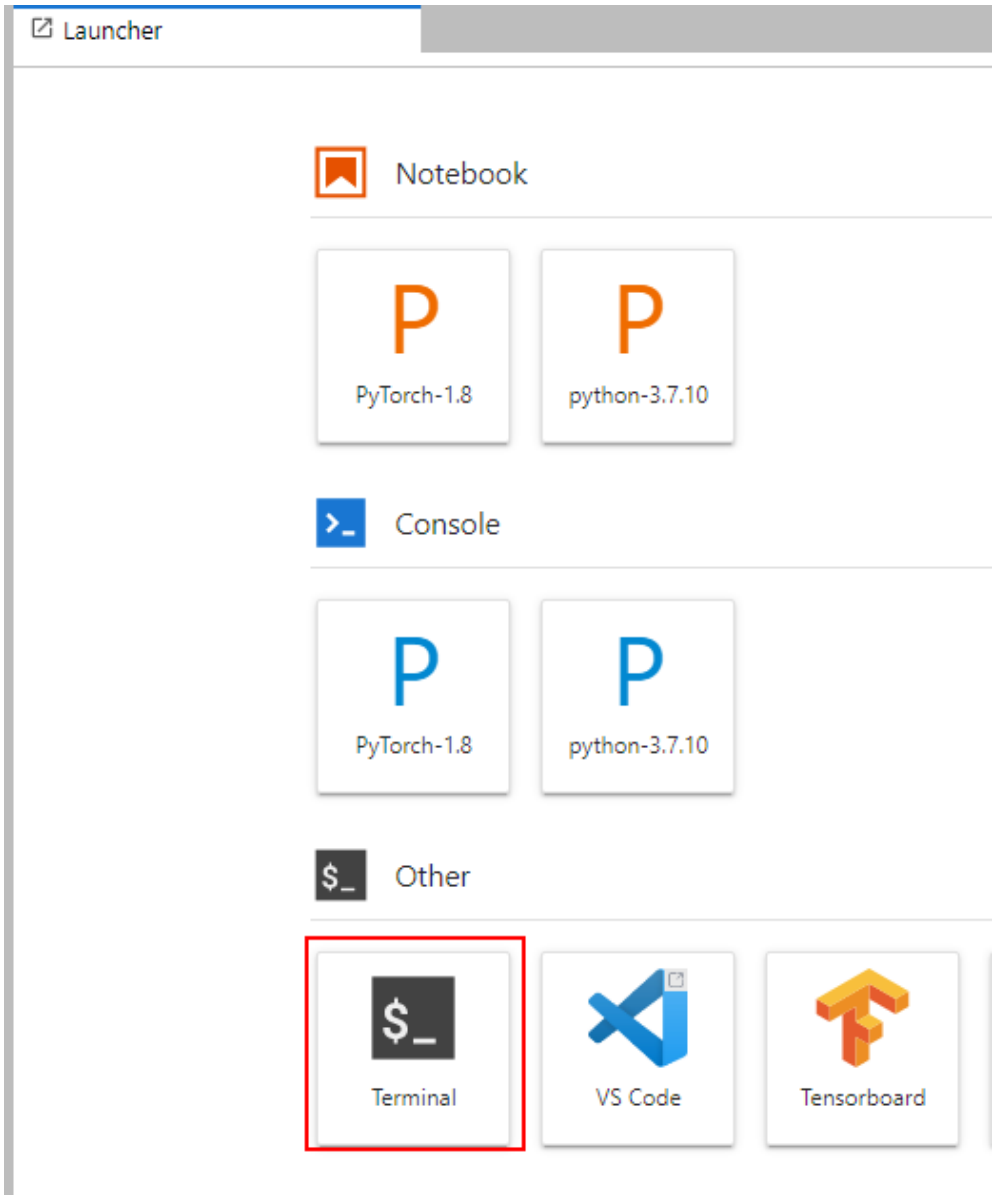
7. Acceda a **http://{*EIP*}:3000** en un navegador. El nombre de usuario y la contraseña predeterminados para el primer inicio de sesión es **admin**. Una vez que el inicio de sesión sea exitoso, cambie la contraseña según se le solicite.



4.2.2.3 Instalación y configuración de Grafana en una instancia de notebook

Requisitos previos

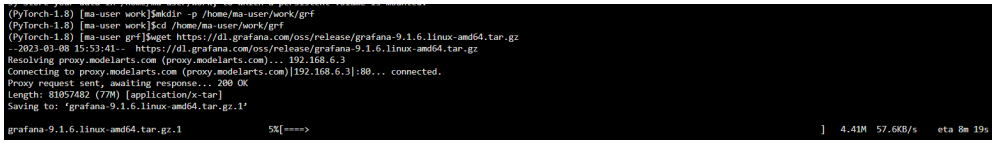
- Hay disponible una instancia de notebook basada en CPU o GPU en ejecución.
- Se abre un terminal.



Procedimiento

1. Ejecute los siguientes comandos en secuencia en su terminal para descargar e instalar Grafana:

```
mkdir -p /home/ma-user/work/grf
cd /home/ma-user/work/grf
wget https://dl.grafana.com/oss/release/grafana-9.1.6.linux-amd64.tar.gz
tar -zxvf grafana-9.1.6.linux-amd64.tar.gz
```



2. Registre Grafana con jupyter-server-proxy.
   a. Ejecute los siguientes comandos en su terminal:

```
mkdir -p /home/ma-user/.local/etc/jupyter
```



```
vi /home/ma-user/.local/etc/jupyter/jupyter_notebook_config.py
```

```
(PyTorch-1.8) [ma-user grf]$mkdir -p /home/ma-user/.local/etc/jupyter
(PyTorch-1.8) [ma-user grf]$vi /home/ma-user/.local/etc/jupyter/jupyter_notebook_config.py
```

b. Agregue el siguiente código en **jupyter_notebook_config.py**, pulse **Esc** para salir y escriba **:wq** para guardar los cambios:

```
c.ServerProxy.servers = {
  'grafana': {
    'command': ['/home/ma-user/work/grf/grafana-9.1.6/bin/
grafana-server', '--homepath', '/home/ma-user/work/grf/
grafana-9.1.6', 'web'],
    'timeout': 1800,
    'port': 3000
  }
}
```

NOTA

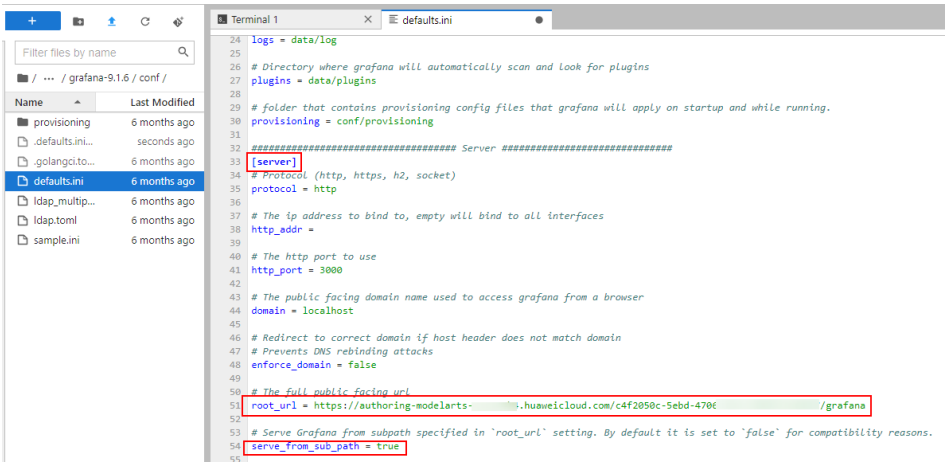
Si **jupyter_notebook_config.py** (ruta de acceso: **/home/ma-user/.local/etc/jupyter/jupyter_notebook_config.py**) contiene el archivo **c.ServerProxy.servers**, agregue el par de clave y valor correspondiente.

3. Modifique la URL para acceder a Grafana en JupyterLab.

a. En el panel de navegación de la izquierda, abra el archivo **vi /home/ma-user/work/grf/grafana-9.1.6/conf/defaults.ini**.

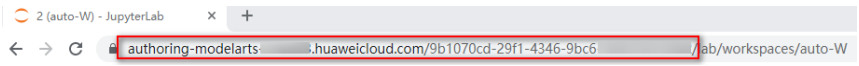
b. Cambie los campos **root_url** y **serve_from_sub_path** de **[server]**.

Figura 4-1 Modificación del archivo **defaults.ini**



En el archivo:

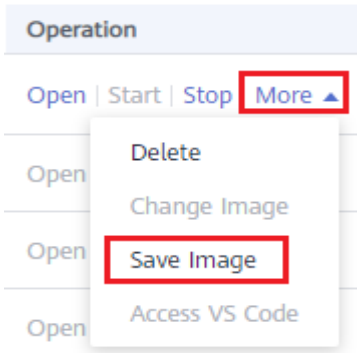
- El valor de **root_url** tiene el formato **https:**{*Jupyterlab domain name*}/{*Instance ID*}**/grafana**. Puede obtener el nombre de dominio y el ID de instancia en el cuadro de direcciones de la página de JupyterLab.



- Establezca **Serve_from_sub_path** en **true**.

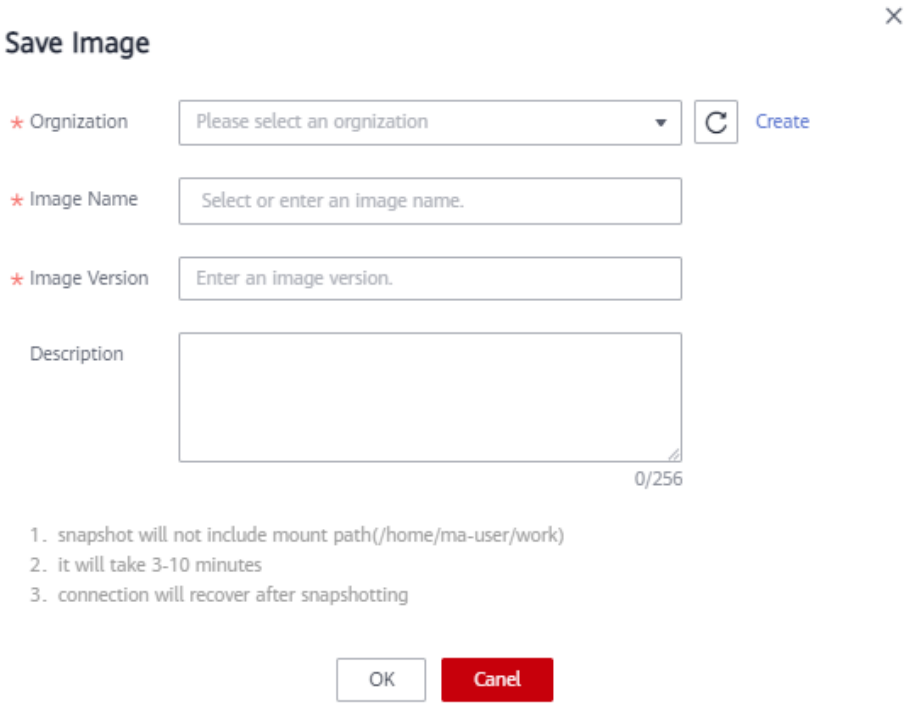
4. Guarde la imagen de la instancia de notebook.

a. Inicie sesión en la consola de ModelArts y seleccione **DevEnviron** > **Notebook**. En la lista de instancias de notebook, seleccione **More** > **Save Image** en la columna **Operation** de la instancia de destino.



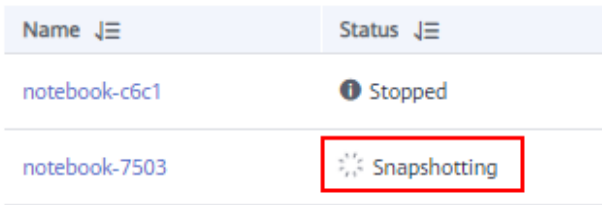
b. En el cuadro de diálogo **Save Image**, configure los parámetros. Haga clic en **OK** para guardar la imagen.

Figura 4-2 Guardar una imagen



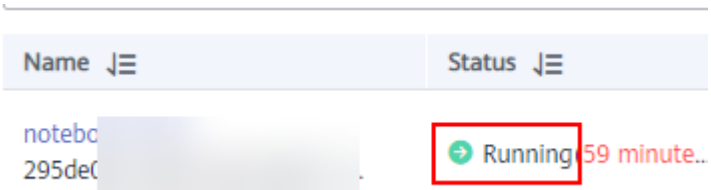
c. La imagen se guardará como una instantánea y tardará unos 5 minutos. Durante este período de tiempo, no realice ninguna operación en la instancia.

Figura 4-3 Instantáneas



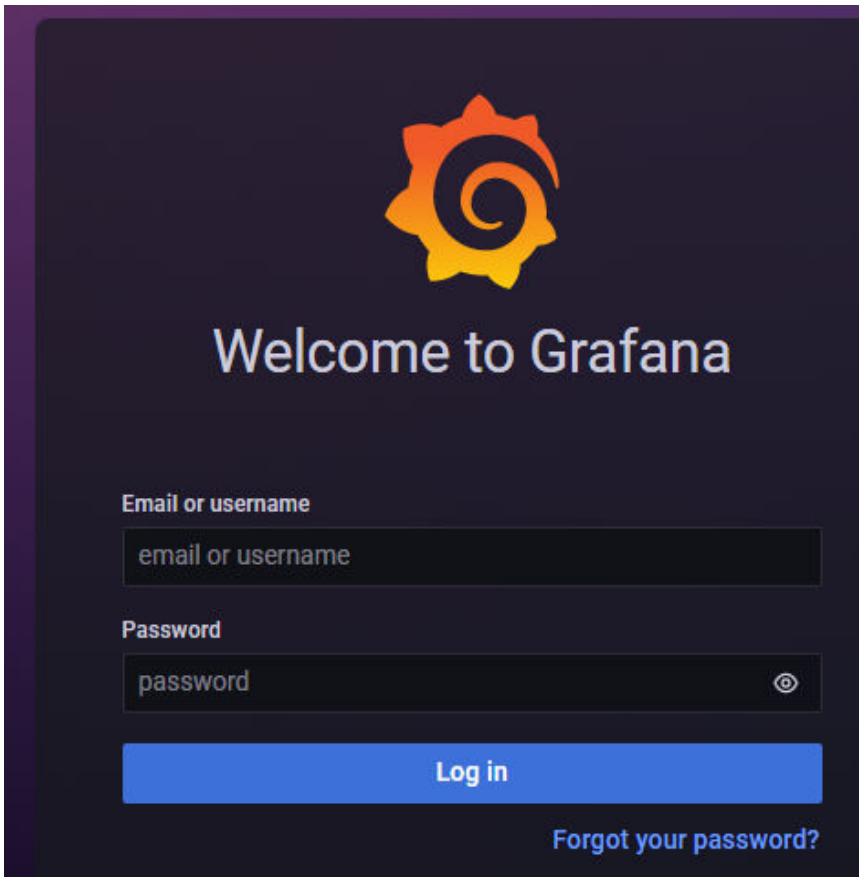
d. Después de guardar la imagen, el estado de la instancia cambia a **Running**. Luego, reinicie la instancia del notebook.

Figura 4-4 Imagen guardada



5. Abra la página de Grafana.

   Abra una ventana del explorador y escriba el valor de **root_url** configurado en **3** en el cuadro de direcciones. Si se muestra la página de inicio de sesión de Grafana, Grafana se instala y configura en la instancia de notebook. El nombre de usuario y la contraseña predeterminados para el primer inicio de sesión es **admin**. Una vez que el inicio de sesión sea exitoso, cambie la contraseña según se le solicite.



4.2.3 Configuración de un origen de datos de Grafana

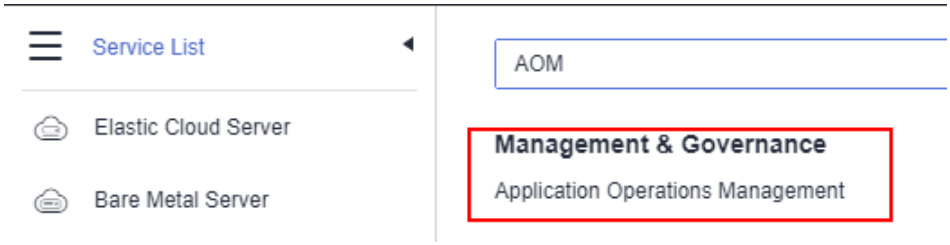
Antes de ver los datos de monitoreo de ModelArts en Grafana, configure el origen de datos.

Requisitos previos

- Se ha instalado Grafana.

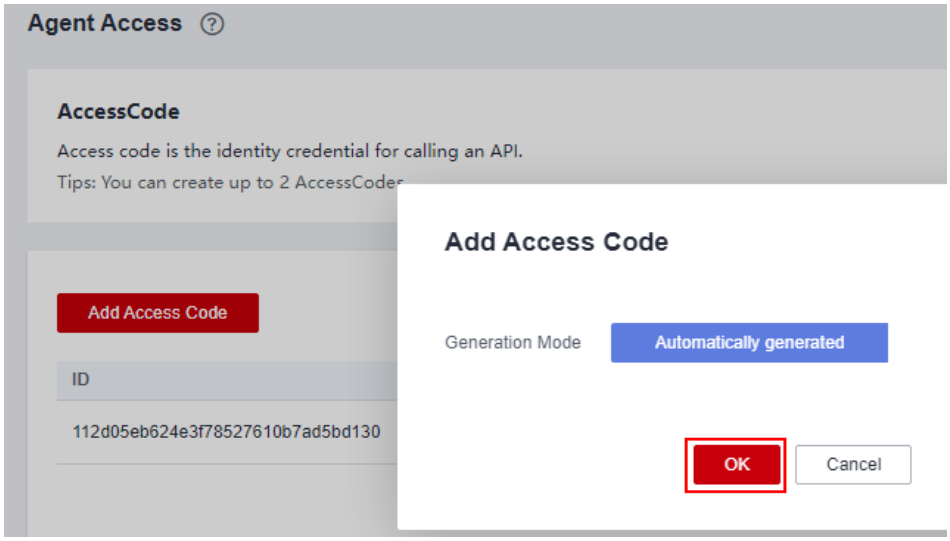
Procedimiento

1. Agregar código de acceso.
   a. Inicie sesión en la consola de AOM.



   b. En el panel de navegación de la izquierda, seleccione **Configuration Management** > **Agent Access** y haga clic en **Add Access Code** para generar un código de acceso.

Figura 4-5 Generación de un código de acceso



   c. Haga clic en para ver el código de acceso generado.

Figura 4-6 Consulta del código de acceso

2. Obtener la dirección de URL del origen de datos.

   La dirección de URL tiene el formato **https://**{*Endpoint*}**/v1/**{*project_id*}.

   – Puede obtener la información del punto de conexión de AOM desde Regiones y puntos de conexión.
   – Establezca **project_id** en el ID de proyecto de la región correspondiente. Puede obtener el ID del proyecto en **My Credentials**.

Figura 4-7 Mis credenciales

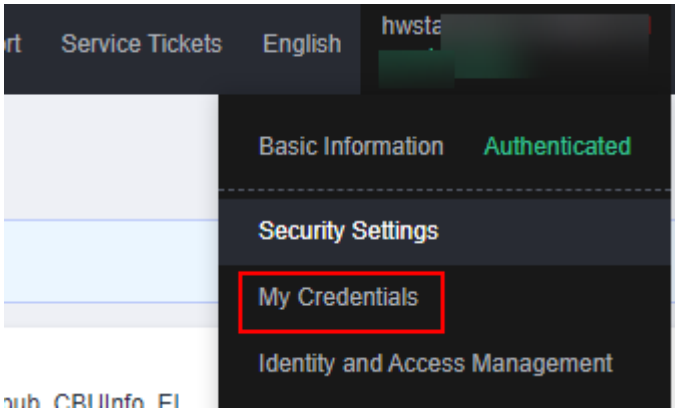
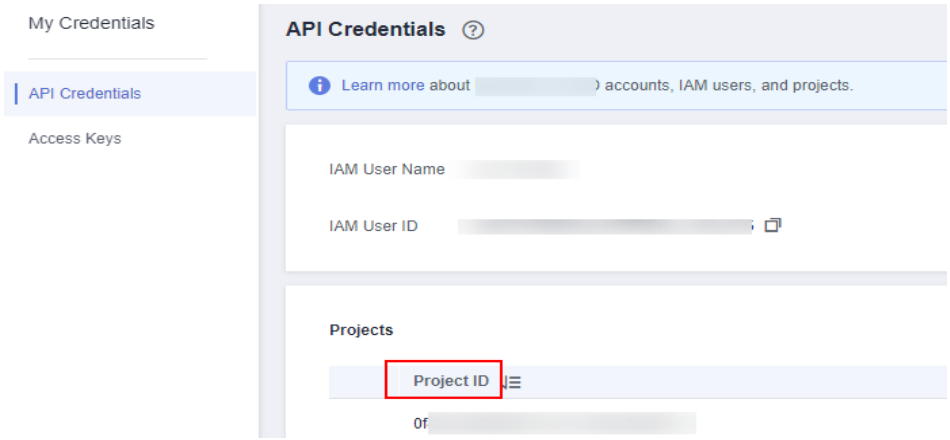
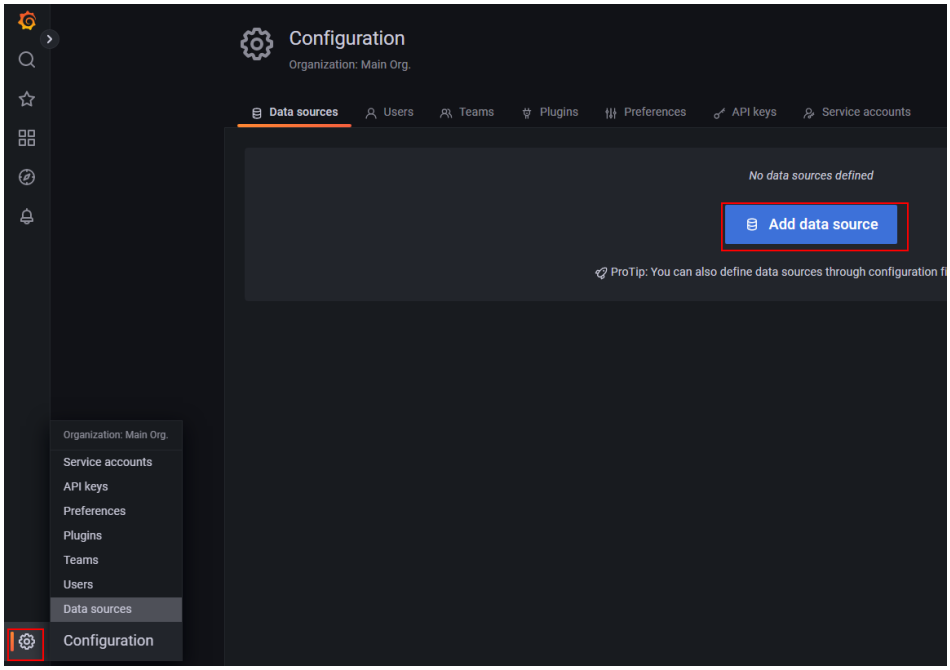


Figura 4-8 Obtención del ID del proyecto



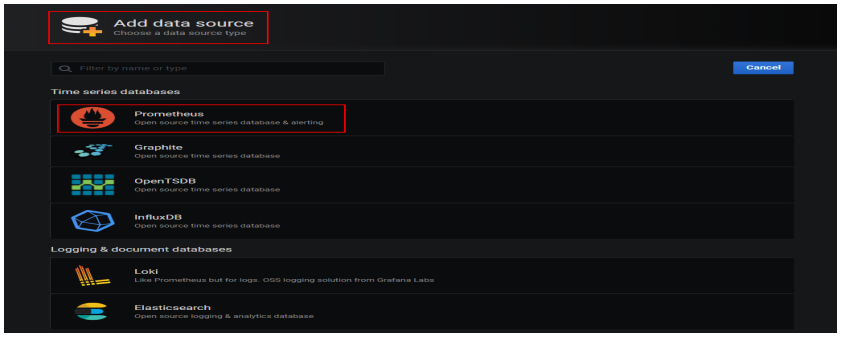
3. Agregue una fuente de datos a Grafana.
   a. Inicie sesión en Grafana. El nombre de usuario y la contraseña predeterminados para el primer inicio de sesión es **admin**. Una vez que el inicio de sesión sea exitoso, cambie la contraseña según se le solicite.
   b. En el panel de navegación, seleccione **Configuration** > **Data Sources**. Luego, haga clic en **Add data source**.

Figura 4-9 Configuración de Grafana



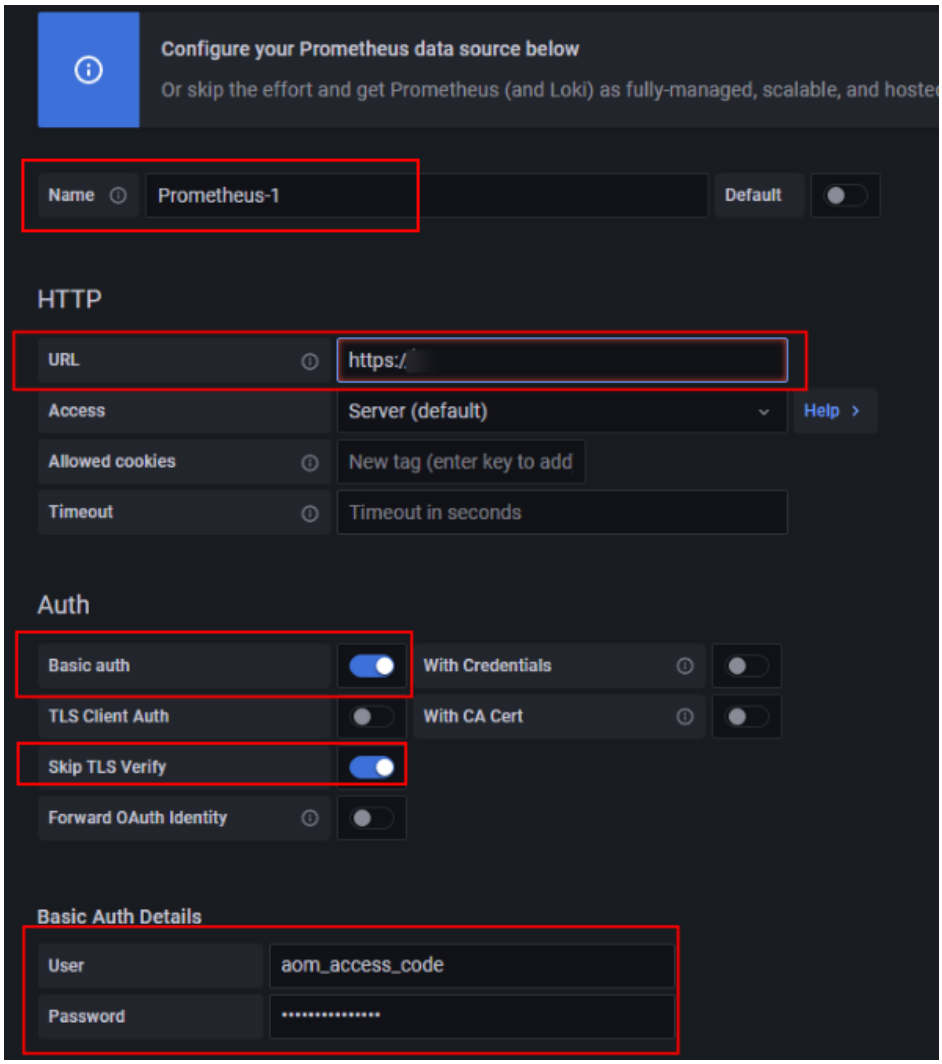
c. Haga clic en **Prometheus** para acceder a la página de configuración.

Figura 4-10 Acceso a la página de configuración de Prometheus



d. Configure los parámetros como se muestra en la siguiente figura.

Figura 4-11 Configuración de un origen de datos de Grafana



NOTA

La versión real de Grafana varía en función del método de instalación. **Figura 4-11** es solo un ejemplo.

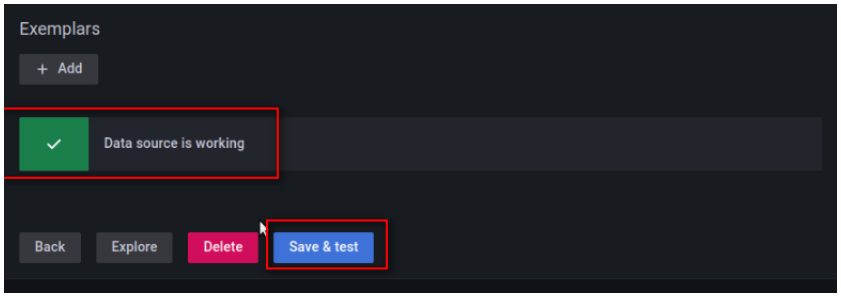
Tabla 4-1 Parámetros

Parámetro	Descripción
Name	Nombre personalizado
URL	Dirección de URL **https://**{*Endpoint*}**/v1/**{*project_id*} combinada en **Obtener la dirección URL del origen de datos.**
Basic auth	Habilitada
Skip TLS Verify	Habilitada

Parámetro	Descripción
User	aom_access_code
Password	Código de acceso generado en **Agregar un código de acceso**.

e. Después de la configuración, haga clic en **Save & test**. Si aparece el mensaje **Data source is working**, se configura el origen de datos.

Figura 4-12 Origen de datos agregado



4.2.4 Uso de Grafana para configurar paneles y consultar datos de métrica

En Grafana, puede personalizar los paneles para diversas vistas. ModelArts también proporciona plantillas de configuración para clústeres. En esta sección se describe cómo configurar un panel con una plantilla de ModelArts o por crear un panel. Para obtener más información, véase los **tutoriales de Grafana**.

Preparación

ModelArts proporciona plantillas para la vista de clústeres, la vista de nodos, la vista de usuarios, la vista de tareas y la vista de detalles de tareas. Estas plantillas se pueden descargar de los documentos oficiales de Grafana. Puede importarlos y utilizarlos en Paneles.

Tabla 4-2 URL para descarga de plantillas

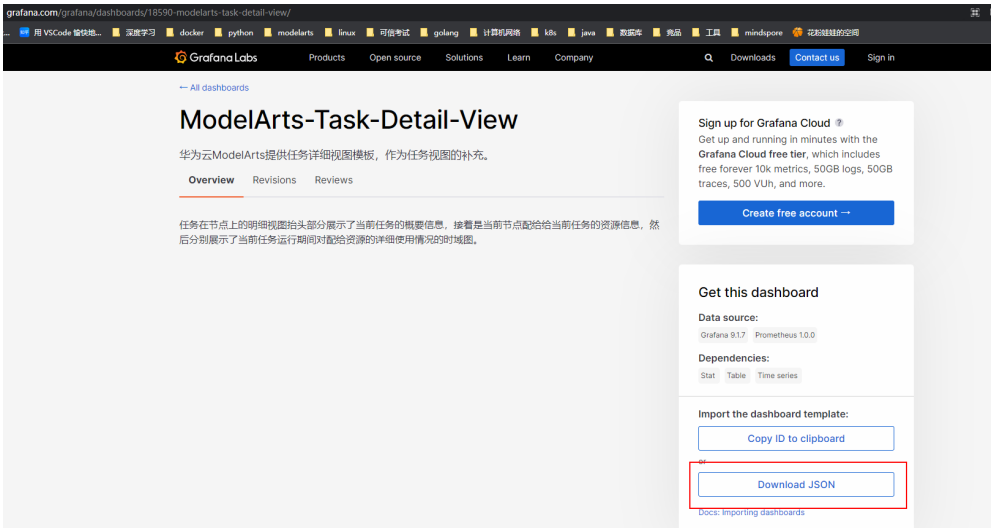
Nombre de la plantilla	URL para descarga
Vista de clúster	https://grafana.com/grafana/dashboards/18582-modelarts-cluster-view/
Vista de nodo	https://grafana.com/grafana/dashboards/18583-modelarts-node-view/
Vista de usuario	https://grafana.com/grafana/dashboards/18588-modelarts-user-view/
Vista de tareas	https://grafana.com/grafana/dashboards/18604-modelarts-task-view/

Nombre de la plantilla	URL para descarga
Vista de detalles de tareas	https://grafana.com/grafana/dashboards/18590-modelarts-task-detail-view/

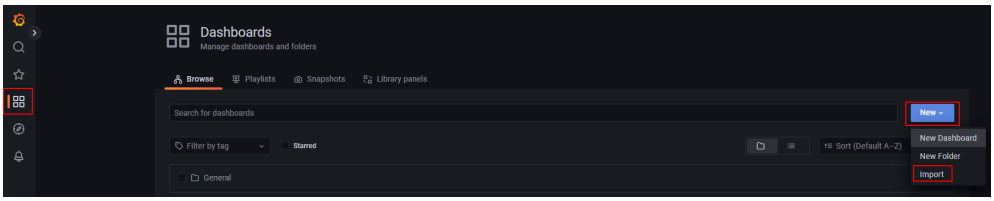
Uso de una plantilla de ModelArts para ver métricas

1. (Opcional) Seleccione la plantilla que desea utilizar. **Preparación** muestra las direcciones de descarga de todas las plantillas. Abra la dirección de destino y haga clic en **Download JSON**.

   Figura 4-13 Downloading the template for the task details view

2. Abra **Dashboards** y seleccione **New** > **Import**.

3. Importe una plantilla de panel de cualquiera de las siguientes formas:
   - Método 1: Cargue el archivo de JSON descargado en **1**, como se muestra en **Figura 4-14**.
   - Método 2: Copie la dirección de descarga de la plantilla proporcionada en **Preparación** y haga clic en **Load** como se muestra en **Figura 4-15**.

Figura 4-14 Carga de un archivo de JSON para importar una plantilla de panel

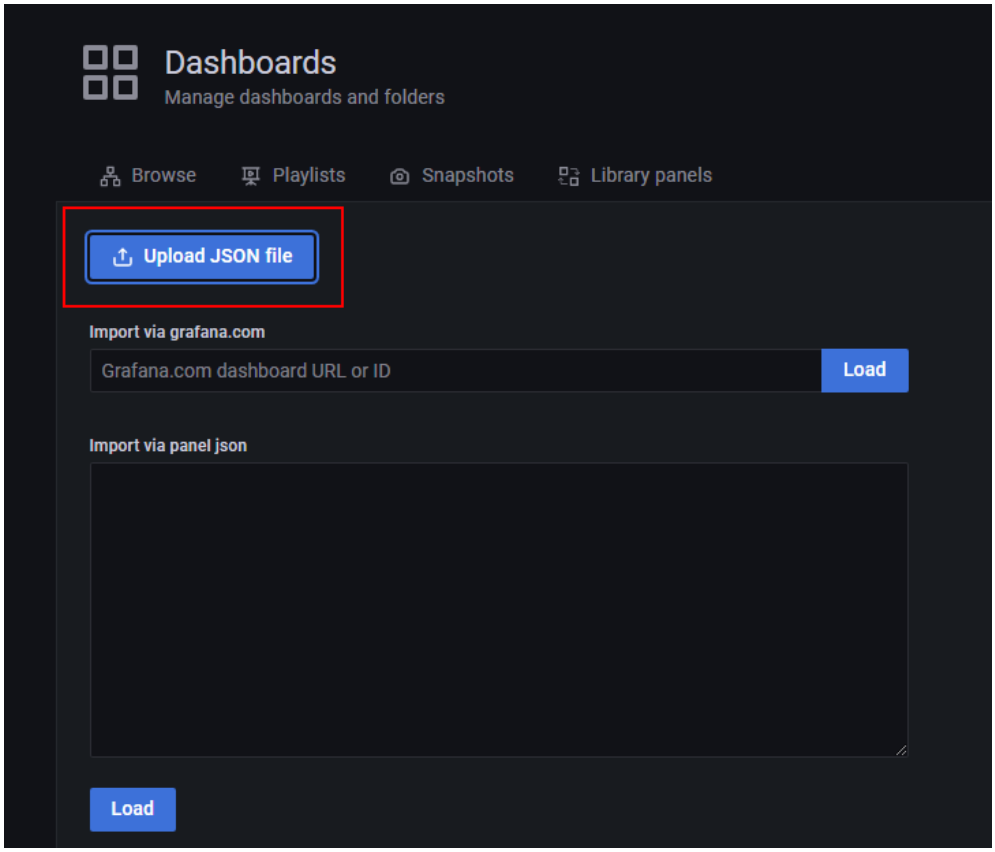
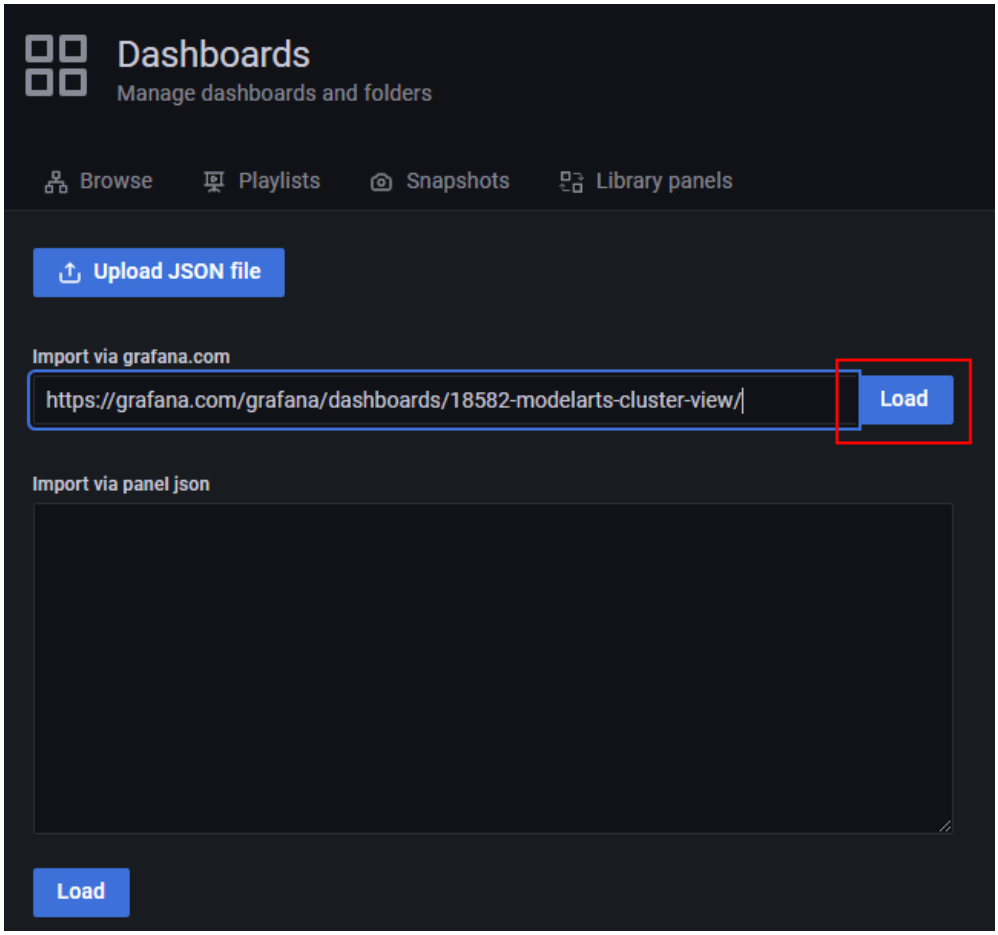
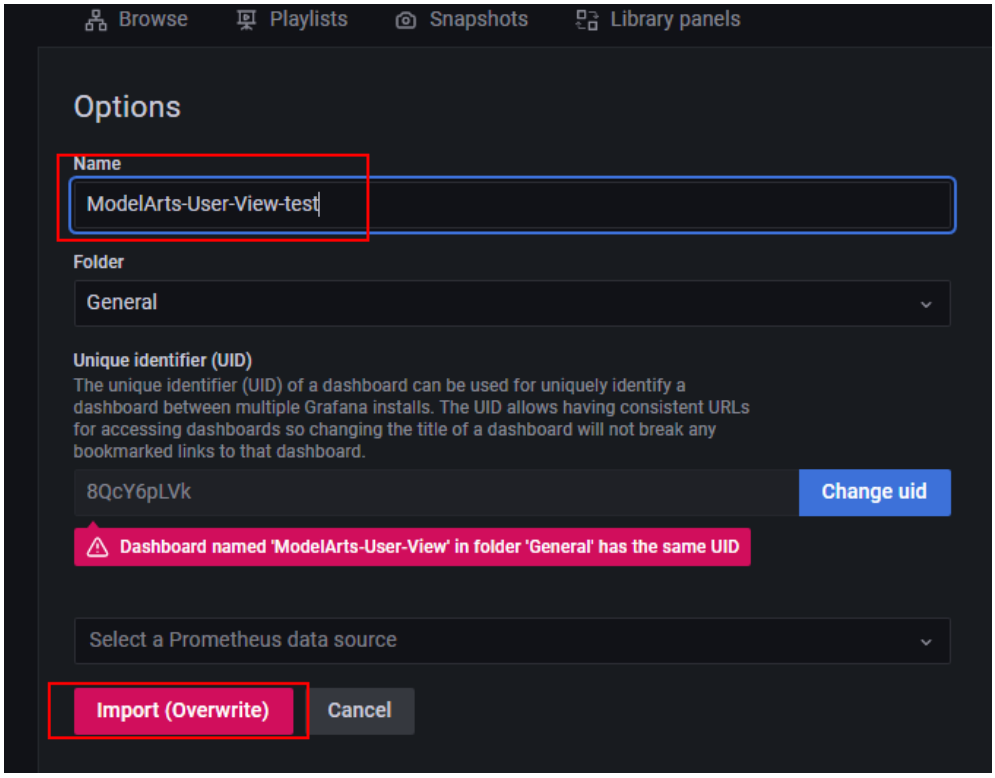


Figura 4-15 Copiar la dirección de la plantilla e importar la plantilla del panel



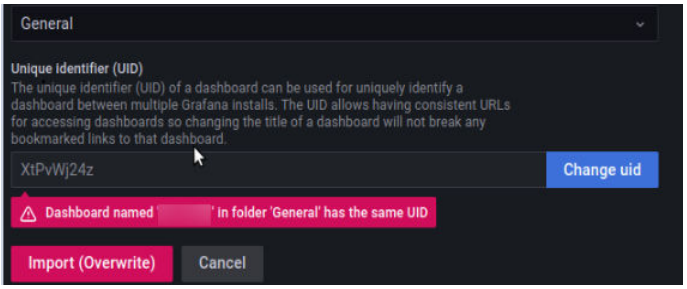
4. Cambie el nombre de la vista y haga clic en **Import**.

Figura 4-16 Cambio del nombre de la vista

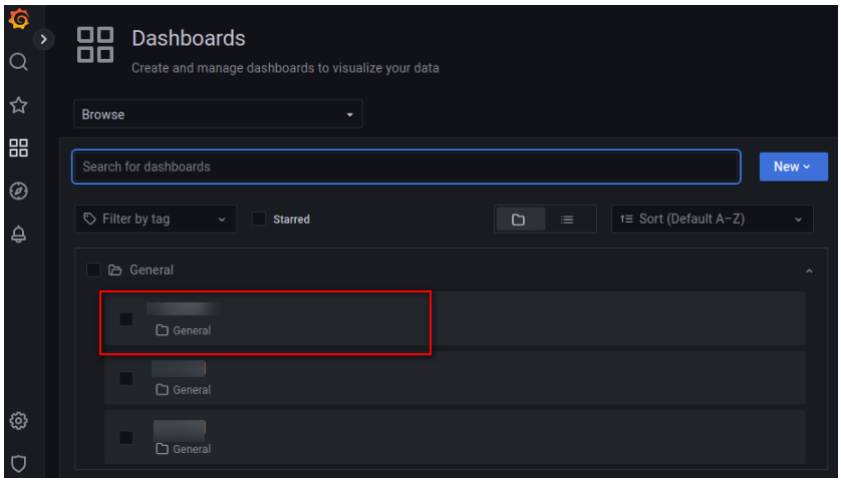


Nota: Si aparece un mensaje que indica que el UID está duplicado, cambie el UID en el archivo de JSON y haga clic en **Import**.

Figura 4-17 Cambio del UID



5. Después de la importación, vea las vistas importadas en **Dashboards**. A continuación, haga clic en una vista para abrir la página de supervisión.

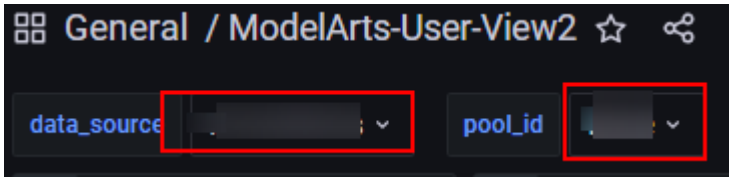


6. Utilice la plantilla.

   Una vez realizada la importación correctamente, puede hacer clic en la plantilla para ver sus detalles. En esta sección se presentan algunas funciones comunes.

   – Cambio del origen de datos y del grupo de recursos

   Figura 4-18 Cambio del origen de datos y del grupo de recursos

   Haga clic en el área marcada por el cuadro rojo. Aparecerá una lista desplegable. Desde allí, puede cambiar el origen de datos y el grupo de recursos.

   – Actualización de datos

   Haga clic en el botón para actualizar situado en la esquina superior derecha para actualizar todos los datos del panel. También se actualizan los datos de cada panel.

   – Cambio del tiempo de actualización automática

   Figura 4-19 Changing the automatic refresh time

   El intervalo de actualización predeterminado de una plantilla es de 15 minutos. Si necesita actualizar el intervalo, cambie el valor del cuadro de lista desplegable situado en la esquina superior derecha.

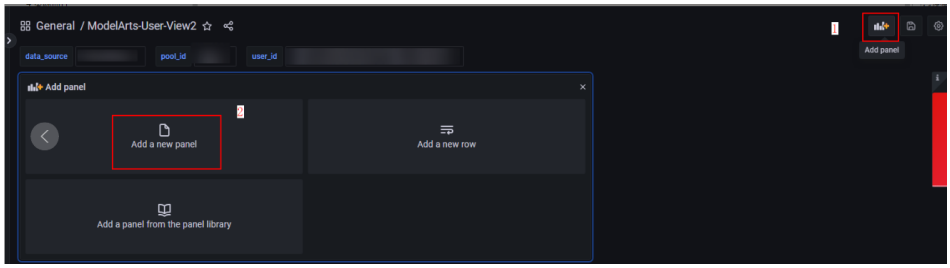
   – Cambio del rango de tiempo para obtener datos del panel

Figura 4-20 Cambio del rango de tiempo para obtener datos

Haga clic en el botón situado en la esquina superior derecha para cambiar el intervalo de tiempo para obtener datos. Este intervalo de tiempo afecta a todos los paneles excepto a aquellos con un tiempo fijo.

– Adición de un panel

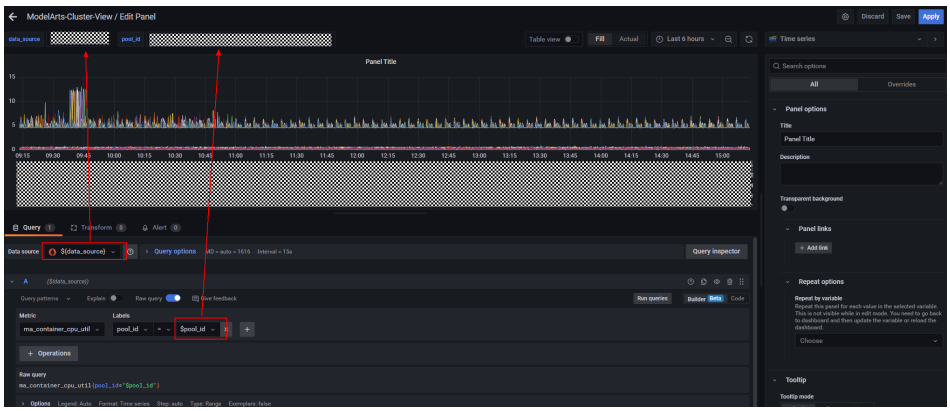
Figura 4-21 Adición de un panel



Haga clic en el ícono + situado en la esquina superior derecha para agregar un panel.

Después de agregar un panel, puede obtener los datos del mismo. Configure el origen de datos y el grupo de recursos de la siguiente manera para utilizar la configuración actual del panel.

Figura 4-22 Uso de la configuración actual del panel de control



Creación de un panel para ver métricas

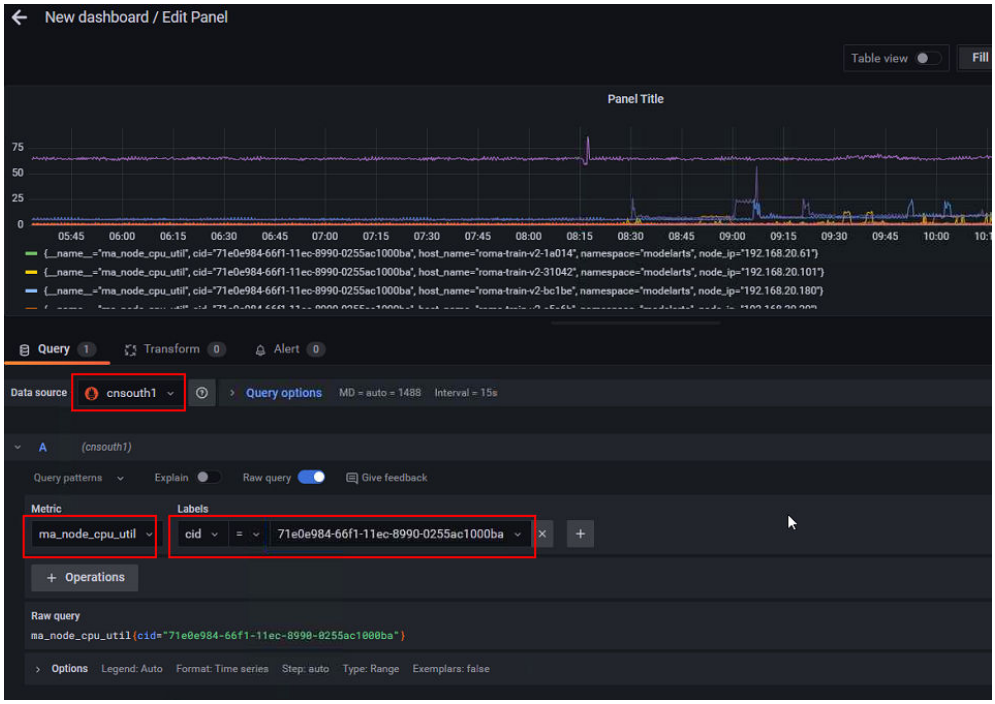
1. Abra **Dashboards**, haga clic en **New** y seleccione **New Panel**.
2. Haga clic en **Add a new panel**.
3. En la página **New dashboard / Edit Panel**, configure los siguientes parámetros:

   Data source: **Origen de datos de Grafana configurado**

   Metric: Nombre de la métrica. Se puede obtener la métrica que se desea consultar consultando **Tabla 4-3**, **Tabla 4-4** y **Tabla 4-5**.

   Labels: Se utiliza para filtrar la métrica. Para obtener más detalles, véase **Tabla 4-6**.

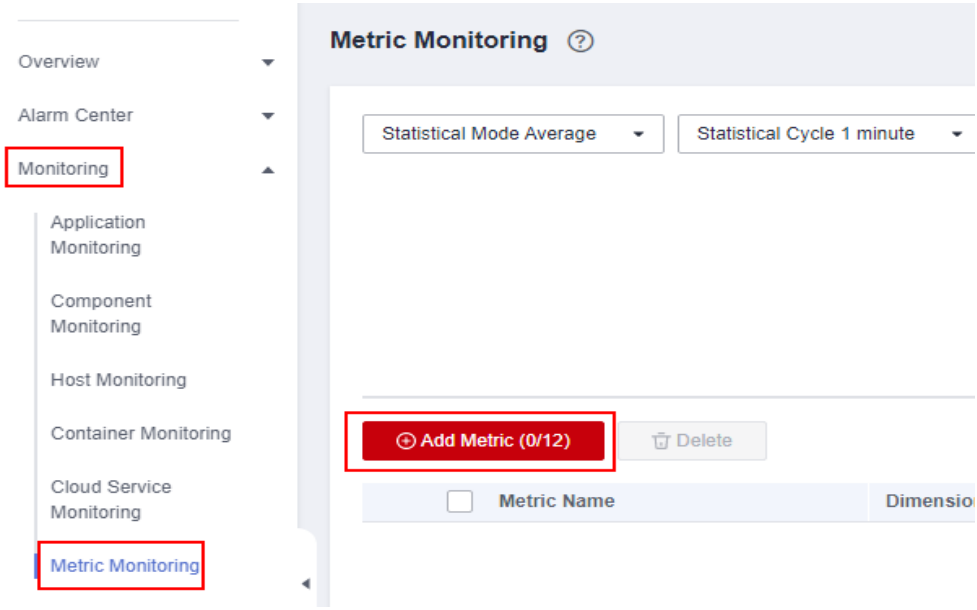
Figura 4-23 Creación de un panel para ver métricas



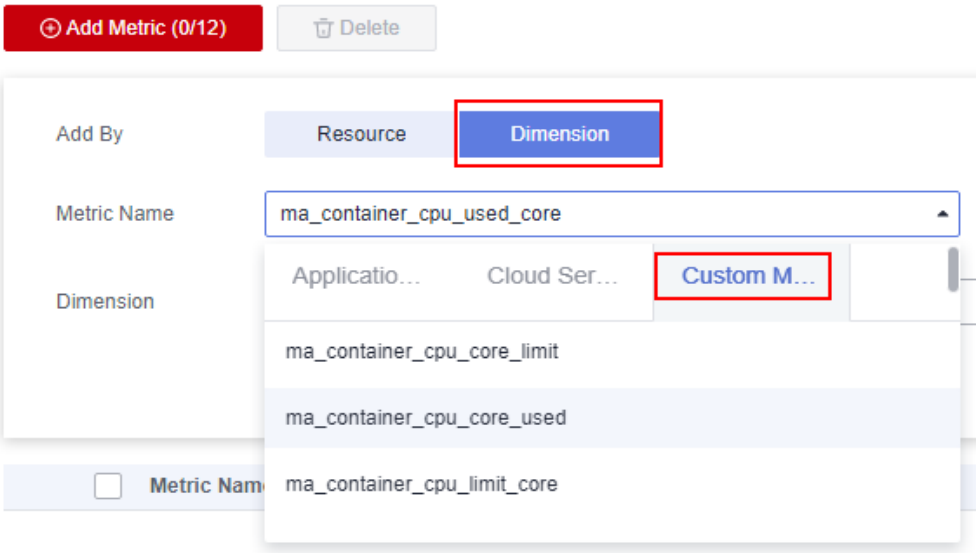
4.3 Consulta de todas las métricas de control de ModelArts en la consola de AOM

ModelArts periódicamente recopila el uso de métricas clave (como GPU, NPU, CPU y memoria) de cada nodo en un grupo de recursos, así como el uso de métricas clave de entorno de desarrollo, de trabajos de entrenamiento y de servicios de inferencia y luego reporta los datos a AOM. Puede ver la información en AOM.

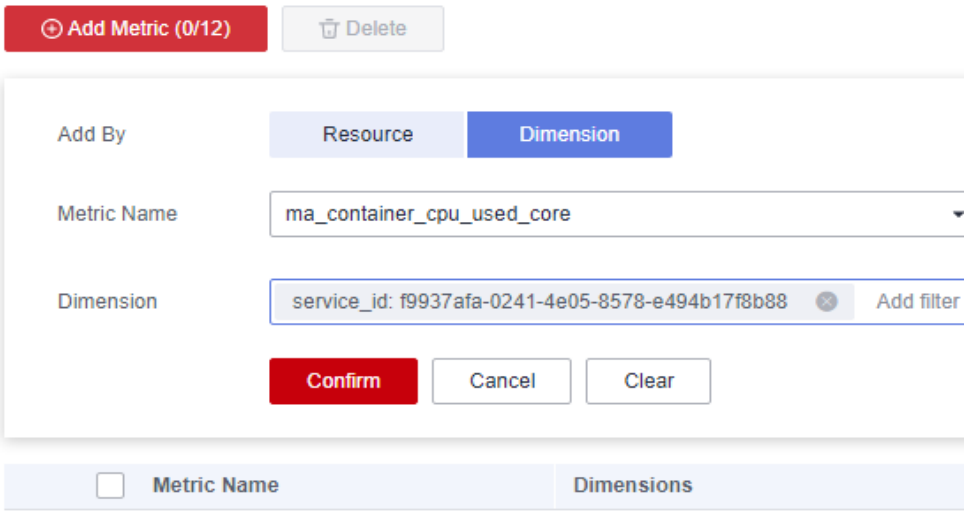
1. Inicie sesión en la consola y busque **AOM** para ir a la consola de AOM.
2. Seleccione **Monitoring** > **Metric Monitoring**. En la página **Metric Monitoring** que aparece en pantalla, haga clic en **Add Metric**.



3. Agregue métricas y haga clic en Confirm.



- **Add By**: seleccione Dimension.
- **Metric Name**: Haga clic en **Custom Metrics**. Seleccione los deseados para la consulta. Para obtener más información, véase **Tabla 4-3**, **Tabla 4-4** y **Tabla 4-5**.
- Dimension: introduzca la etiqueta para filtrar la métrica. Para más detalles, véase **Tabla 4-6**. A continuación se muestra un ejemplo.



4. Consulte las métricas.

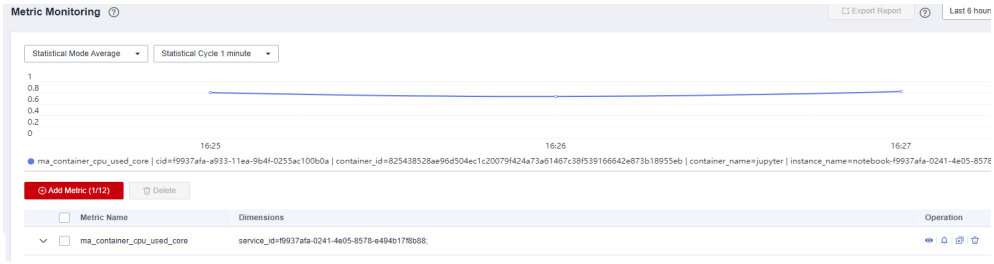


Tabla 4-3 Métricas de contenedores

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
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

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
	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 utilizada por un objeto medido (**container_memory_working_set_bytes** en el espacio de trabajo actual) (Uso de memoria en un conjunto de trabajo = página anónima activa y caché, y página horneada en archivos ≤ **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%

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
	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	%	%
	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	Watt (W)	> 0
	Temperatura de memoria de GPU	DCGM_FI_DEV_MEMORY_TEMP	Temperatura de memoria de GPU	°C	Número natural
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

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
	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
	Velocidad de enlace ascendente (PPS)	ma_container_network_transmit_packets	Número de paquetes de datos enviados por una NIC por segundo	Paquetes/s	≥ 0
Métricas de servicio de notebook	Tamaño de directorio de caché de notebook	ma_container_notebook_cache_dir_size_bytes	Se conecta un disco local de alta velocidad al directorio **/cache** para instancias de notebook de GPU. Esta métrica indica el tamaño total del directorio.	Bytes	≥ 0
	Uso de directorio de caché de notebook	ma_container_notebook_cache_dir_util	Se conecta un disco local de alta velocidad al directorio **/cache** para instancias de notebook de GPU. Esta métrica indica la utilización del directorio.	%	0%–100%

Tabla 4-4 Métricas de nodo (recogidas solo en grupos de recursos dedicados)

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
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

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
	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%
	Tiempo de espera de E/S de CPU	ma_node_cpu_iowait_counter	Tiempo de espera de E/S de disco acumulado desde el inicio del sistema	jiffies	≥ 0
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%
	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

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
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
	Espacio usado de contenedor	ma_node_container_space_used_capacity_megabytes	Espacio usado de contenedor	MB	≥ 0

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
	Información del disco	ma_node_disk_info	Información básica del disco	N/A	≥ 0
	Total de lecturas	ma_node_disk_reads_completed_total	Número total de lecturas exitosas	N/A	≥ 0
	Lecturas combinadas	ma_node_disk_reads_merged_total	Número de lecturas combinadas	N/A	≥ 0
	Bytes leídos	ma_node_disk_read_bytes_total	Número total de bytes que se leen correctamente	Bytes	≥ 0
	Tiempo dedicado para lectura	ma_node_disk_read_time_seconds_total	Tiempo dedicado a todas las lecturas	Segundos	≥ 0
	Total de escrituras	ma_node_disk_writes_completed_total	Número total de escrituras exitosas	N/A	≥ 0
	Escrituras combinadas	ma_node_disk_writes_merged_total	Número de escrituras combinadas	N/A	≥ 0
	Bytes escritos	ma_node_disk_written_bytes_total	Número total de bytes que se escriben correctamente	Bytes	≥ 0
	Tiempo dedicado para escritura	ma_node_disk_write_time_seconds_total	Tiempo dedicado en todas las operaciones de escritura	Segundos	≥ 0
	E/S en curso	ma_node_disk_io_now	Cantidad de E/S en curso	N/A	≥ 0

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
	Duración de ejecución de E/S	ma_node_disk_io_time_seconds_total	Time spent on executing I/Os	Segundos	≥ 0
	Tiempo ponderado de ejecución de E/S	ma_node_disk_io_time_weighted_seconds_tota	Número ponderado de segundos dedicados a E/S	Segundos	≥ 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
	Temperatura de la GPU	DCGM_FI_DEV_GPU_TEMP	Temperatura de la GPU	°C	Número natural

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
	Potencia de la GPU	DCGM_FI_DEV_POWER_USAGE	Potencia de la GPU	Watt (W)	> 0
	Temperatura de memoria de GPU	DCGM_FI_DEV_MEMORY_TEMP	Temperatura de memoria de GPU	°C	Número natural
InfiniBand o red de RoCE	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

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
Estado de montaje de NFS	Tiempo de congestión de Getattr de NFS	ma_node_mountstats_getattr_backlog_wait	Getattr es una operación de NFS que recupera los atributos de un archivo o directorio, como tamaño, permisos, propietario, etc. La espera de retraso es el tiempo que las solicitudes de NFS tienen que esperar en la cola de retrasos antes de ser enviadas al servidor de NFS. Indica la congestión del lado del cliente de NFS. Una espera atrasada alta puede ocasionar un rendimiento deficiente de NFS y tiempos de respuesta lentos del sistema.	ms	≥ 0

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
	Tiempo de ida y vuelta de Getattr de NFS	ma_node_mountstats_getattr_rtt	Getattr es una operación de NFS que recupera los atributos de un archivo o directorio, como tamaño, permisos, propietario, etc. RTT significa tiempo de ida y vuelta y es el tiempo desde que el cliente de RPC del kernel envía la petición de RPC hasta el momento en que recibe la reply34. RTT incluye el tiempo de tránsito de la red y el tiempo de ejecución del servidor. RTT es una buena medida para la latencia de NFS. Un RTT alto puede indicar problemas	ms	≥ 0

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
			de red o de servidor.		
	Tiempo de congestión de acceso de NFS	ma_node_mountstats_access_backlog_wait	El acceso es una operación de NFS que comprueba los permisos de acceso de un archivo o directorio para un usuario determinado. La espera de retraso es el tiempo que las solicitudes de NFS tienen que esperar en la cola de retrasos antes de ser enviadas al servidor de NFS. Indica la congestión del lado del cliente de NFS. Una espera atrasada alta puede ocasionar un rendimiento deficiente de NFS y tiempos de respuesta lentos del sistema.	ms	≥ 0

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
	Tiempo de ida y vuelta de acceso de NFS	ma_node_mountstats_access_rtt	El acceso es una operación de NFS que comprueba los permisos de acceso de un archivo o directorio para un usuario determinado. RTT significa tiempo de ida y vuelta y es el tiempo desde que el cliente de RPC del kernel envía la petición de RPC hasta el momento en que recibe la reply34. RTT incluye el tiempo de tránsito de la red y el tiempo de ejecución del servidor. RTT es una buena medida para la latencia de NFS. Un RTT alto puede indicar problemas	ms	≥ 0

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
			de red o de servidor.		
	Tiempo de congestión de búsqueda de NFS	ma_node_mountstats_lookup_backlog_wait	La búsqueda es una operación de NFS que resuelve un nombre de archivo en un directorio en un controlador de archivo. La espera de retraso es el tiempo que las solicitudes de NFS tienen que esperar en la cola de retrasos antes de ser enviadas al servidor de NFS. Indica la congestión del lado del cliente de NFS. Una espera atrasada alta puede ocasionar un rendimiento deficiente de NFS y tiempos de respuesta lentos del sistema.	ms	≥ 0

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
	Tiempo de ida y vuelta de búsqueda de NFS	ma_node_mountstats_lookup_rtt	La búsqueda es una operación de NFS que resuelve un nombre de archivo en un directorio en un controlador de archivo. RTT significa tiempo de ida y vuelta y es el tiempo desde que el cliente de RPC del kernel envía la petición de RPC hasta el momento en que recibe la reply34. RTT incluye el tiempo de tránsito de la red y el tiempo de ejecución del servidor. RTT es una buena medida para la latencia de NFS. Un RTT alto puede indicar problemas	ms	≥ 0

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
			de red o de servidor.		
	Tiempo de congestión de lectura de NFS	ma_node_mountstats_read_backlog_wait	Leer es una operación de NFS que lee datos de un archivo. La espera de retraso es el tiempo que las solicitudes de NFS tienen que esperar en la cola de retrasos antes de ser enviadas al servidor de NFS. Indica la congestión del lado del cliente de NFS. Una espera atrasada alta puede ocasionar un rendimiento deficiente de NFS y tiempos de respuesta lentos del sistema.	ms	≥ 0

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
	Tiempo de ida y vuelta de lectura de NFS	ma_node_mountstats_read_rtt	Leer es una operación de NFS que lee datos de un archivo. RTT significa tiempo de ida y vuelta y es el tiempo desde que el cliente de RPC del kernel envía la petición de RPC hasta el momento en que recibe la reply34. RTT incluye el tiempo de tránsito de la red y el tiempo de ejecución del servidor. RTT es una buena medida para la latencia de NFS. Un RTT alto puede indicar problemas de red o de servidor.	ms	≥ 0

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
	Tiempo de congestión de escritura de NFS	ma_node_mountstats_write_backlog_wait	Write es una operación de NFS que escribe datos en un archivo. La espera de retraso es el tiempo que las solicitudes de NFS tienen que esperar en la cola de retrasos antes de ser enviadas al servidor de NFS. Indica la congestión del lado del cliente de NFS. Una espera atrasada alta puede ocasionar un rendimiento deficiente de NFS y tiempos de respuesta lentos del sistema.	ms	≥ 0

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
	Tiempo de ida y vuelta de escritura de NFS	ma_node_mountstats_write_rtt	Write es una operación de NFS que escribe datos en un archivo. RTT significa tiempo de ida y vuelta y es el tiempo desde que el cliente de RPC del kernel envía la petición de RPC hasta el momento en que recibe la reply34. RTT incluye el tiempo de tránsito de la red y el tiempo de ejecución del servidor. RTT es una buena medida para la latencia de NFS. Un RTT alto puede indicar problemas de red o de servidor.	ms	≥ 0

Tabla 4-5 Diagnóstico (InfiniBand, recopilado solo en los grupos de recursos dedicados)

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
InfiniBand o red de RoCE	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
	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

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
	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
	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: Mapeo de DLID Mapeo de VL Bucle (puerto de salida = puerto de entrada)	Recuento total	Número natural

Categoría	Nombre	Métrica	Descripción	Unidad	Rango de valor
	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 4-6 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
	container_id	ID de contenedor
	cid	ID de clúster
	container_name	Nombre del contenedor
	project_id	ID de proyecto de la cuenta a la que pertenece el usuario
	user_id	ID de usuario de la cuenta a la que pertenece el usuario que envía el trabajo
	pool_id	ID de un grupo de recursos correspondiente a un grupo de recursos dedicado físico

Clasificación	Métrica	Descripción
	pool_name	Nombre de un grupo de recursos correspondiente a un grupo de recursos dedicado físico
	logical_pool_id	ID de un subgrupo lógico
	logical_pool_name	Nombre de un subgrupo lógico
	gpu_uuid	UUID de la GPU utilizada por el contenedor
	gpu_index	Índice de la GPU utilizada por el contenedor
	gpu_type	Tipo de GPU utilizada por el contenedor
	account_name	Nombre de la cuenta del creador de una tarea de entrenamiento, de inferencia o de entorno de desarrollo
	user_name	Nombre de usuario del creador de una tarea de entrenamiento, de inferencia o de entorno de desarrollo
	task_creation_time	Hora en la que se crea una tarea de entrenamiento, de inferencia o de entorno de desarrollo
	task_name	Nombre de una tarea de entrenamiento, de inferencia o de entorno de desarrollo
	task_spec_code	Especificaciones de una tarea de entrenamiento, de inferencia o de entorno de desarrollo
	cluster_name	Nombre del clúster de CCE
Métricas de nodos	cid	ID del clúster de CCE al que pertenece el nodo
	node_ip	Dirección IP del nodo
	host_name	Nombre de host de un 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
	gpu_uuid	UUID de una GPU de nodo
	gpu_index	Índice de una GPU de nodo
	gpu_type	Tipo de GPU de nodo

Clasificación	Métrica	Descripción
	device_name	Nombre del dispositivo de una NIC de InfiniBand o de red de RoCE
	port	Número de puerto de la NIC de InfiniBand
	physical_state	Estado de cada puerto de la NIC de InfiniBand
	firmware_version	Versión de firmware de la NIC de InfiniBand
	filesystem	Sistema de archivos montado en NFS
	mount_point	Punto de montaje de NFS
Diagnos	cid	ID del clúster de CCE al que pertenece el nodo con la GPU equipada
	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
	gpu_index	Índice de una GPU de nodo
	gpu_type	Tipo de GPU de nodo
	device_name	Nombre de un dispositivo de red o de disco
	port	Número de puerto de la NIC de InfiniBand
	physical_state	Estado de cada puerto de la NIC de InfiniBand
	firmware_version	Versión de firmware de la NIC de InfiniBand

