ModelArts

Gerenciamento de recursos

Edição 01

Data 2024-09-14

HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD.

Marcas registadas e permissões

e outras marcas registadas da Huawei são marcas registadas da Huawei Technologies Co., Ltd. Todos as outras marcas registadas e os nomes registados mencionados neste documento são propriedade dos seus respectivos detentores.

Aviso

Os produtos, os serviços e as funcionalidades adquiridos são estipulados pelo contrato estabelecido entre a Huawei Cloud e o cliente. Os produtos, os serviços e as funcionalidades descritos neste documento, no todo ou em parte, podem não estar dentro do âmbito de aquisição ou do âmbito de uso. Salvo especificação em contrário no contrato, todas as declarações, informações e recomendações neste documento são fornecidas "TAL COMO ESTÃO" sem garantias ou representações de qualquer tipo, sejam expressas ou implícitas.

As informações contidas neste documento estão sujeitas a alterações sem aviso prévio. Foram feitos todos os esforços na preparação deste documento para assegurar a exatidão do conteúdo, mas todas as declarações, informações e recomendações contidas neste documento não constituem uma garantia de qualquer tipo, expressa ou implícita.

Huawei Cloud Computing Technologies Co., Ltd.

Endereço: Huawei Cloud Data Center, Rua Jiaoxinggong
Avenida Qianzhong
Novo Distrito de Gui'an
Guizhou 550029
República Popular da China

Site: https://www.huaweicloud.com/intl/pt-br/

Índice

1 Pool de recursos

Pools de recursos do ModelArts

Ao usar o ModelArts para desenvolvimento de IA, você pode usar um dos seguintes pools de recursos:

- **Dedicated resource pool**: ele fornece recursos mais controláveis e não pode ser compartilhado com outros usuários. Crie um pool de recursos dedicado e selecione-o durante o desenvolvimento da IA. O pool de recursos dedicados pode ser um cluster elástico ou um BMS elástico.
  - Elastic cluster: pode ser Standard ou Lite.
    - Em um cluster elástico Standard, recursos de computação exclusivos são fornecidos, com os quais você pode fornecer instâncias durante o trabalho de treinamento, implementação de modelo e desenvolvimento de ambiente no ModelArts.
    - Um cluster elástico Lite fornece clusters do Kubernetes hospedados com plug-ins de desenvolvimento de IA convencionais e plug-ins de aceleração para usuários de recursos do Kubernetes. Você pode operar os nós e os clusters do Kubernetes no pool de recursos com os recursos e tarefas de AI Native fornecidos.
  - Elastic BMS: ele fornece diferentes modelos de BMSs de *x*PU. Você pode acessar um BMS elástico por meio de um EIP e instalar drivers e softwares relacionados a GPU e NPU em uma imagem de sistema operacional especificada. Para atender aos requisitos de treinamento de rotina dos engenheiros de algoritmos, o SFS e o OBS podem ser usados para armazenar e ler dados.
- **Public Resource Pool**: fornecem clusters de computação pública em larga escala, que são alocados com base nas configurações de parâmetros de trabalho. Os recursos são isolados por trabalho. Você pode usar pools de recursos públicos do ModelArts para oferecer trabalhos de treinamento, implantar modelos ou executar instâncias do DevEnviron.

Diferenças entre pools de recursos dedicados e pools de recursos públicos

- Os pools de recursos dedicados fornecem clusters de computação dedicados e recursos de rede para os usuários. Os pools de recursos dedicados de diferentes usuários são fisicamente isolados, enquanto os pools de recursos públicos são apenas isolados logicamente. Em comparação com os pools de recursos públicos, os pools de recursos dedicados apresentam melhor desempenho em isolamento e segurança.

- Quando um pool de recursos dedicado é usado para criar trabalhos e os recursos são suficientes, os trabalhos não serão enfileirados. Quando um pool de recursos público é usado para criar trabalhos, há uma alta probabilidade de que os trabalhos sejam enfileirados.
- Um pool de recursos dedicado é acessível à sua rede. Todos os trabalhos em execução no pool podem acessar armazenamento e recursos em sua rede. Por exemplo, se você selecionar um pool de recursos dedicado com uma rede acessível ao criar um trabalho de treinamento, poderá acessar os dados do SFS depois que o trabalho de treinamento for criado.
- Os pools de recursos dedicados permitem que você personalize o ambiente de tempo de execução de nós físicos, por exemplo, você pode atualizar drivers de GPU ou Ascend. Esta função não é suportada por pools de recursos públicos.

2 Cluster elástico

2.1 Atualizações abrangentes das funções de gerenciamento do pool de recursos do ModelArts

Os pools de recursos dedicados do ModelArts foram atualizados e entraram em vigor às 00:00 GMT+08:00 de 1º de março de 2023. No novo sistema, existem apenas pools de recursos dedicados ModelArts unificados, que não são mais classificados como pools dedicados ao desenvolvimento/treinamento e pools dedicados à implementação de serviços. Os pools de recursos dedicados da nova versão oferecem suporte à configuração flexível de tipos de trabalho e permitem que você gerencie redes e interconecte VPCs com as redes.

A nova página dedicada de gerenciamento de pool de recursos fornece funções mais abrangentes e exibe mais informações sobre os pools de recursos. Mais detalhes sobre como usar e gerenciar pools de recursos dedicados são fornecidos nas seções subsequentes deste documento. Se você é novo em pools de recursos dedicados do ModelArts, experimente pools de recursos dedicados de nova versão. Se você usou pools de recursos dedicados do ModelArts, os pools de versão antiga serão facilmente alternados para pools de versão nova.

Leia o conteúdo a seguir para saber mais sobre pools de recursos dedicados de nova versão.

Recursos dos pools de recursos dedicados de nova versão

O gerenciamento de pool de recursos dedicados da nova versão é uma tecnologia abrangente e aprimoramento de produtos. As principais melhorias são as seguintes:

- **Tipo de pool de recursos dedicado único para diversas finalidades**: os pools de recursos dedicados não são mais classificados em desenvolvimento/treinamento e em implementação de serviços. Você pode executar cargas de trabalho de treinamento e inferência em um pool de recursos dedicado. Você também pode definir os tipos de trabalho suportados por um pool de recursos dedicados com base em suas necessidades.
- **Conexão de rede de pool de recursos dedicados**: você pode criar e gerenciar redes de pool de recursos dedicados no console de gerenciamento do ModelArts. Se você precisar acessar recursos em sua VPC para trabalhos executados em um pool de recursos dedicado, interconecte a VPC com a rede do pool de recursos dedicado.
- **Mais detalhes do cluster**: a página de detalhes do pool de recursos dedicados da nova versão fornece mais detalhes do cluster, como trabalhos, nós e monitoramento de

recursos, ajudando você a aprender sobre o status do cluster e planejar e usar melhor os recursos.

- **Gerenciamento de driver de GPU/NPU de cluster**: na página de detalhes do pool de recursos dedicados da nova versão, você pode selecionar um driver de placa aceleradora e executar alterações após o envio ou a atualização suave do driver com base nos requisitos de serviço.
- **Alocação de recursos refinada (em breve)**: você pode dividir seu pool de recursos dedicados em vários pools pequenos e atribuir cotas e permissões diferentes a cada pool pequeno para alocação e gerenciamento de recursos flexíveis e refinados.

Mais recursos serão fornecidos em versões posteriores para uma melhor experiência do usuário.

Posso continuar a usar os pools de recursos dedicados existentes após a atualização entrar em vigor?

Se você tiver criado pools de recursos dedicados, ainda poderá acessar a página de gerenciamento do pool de recursos dedicado (cluster elástico) da versão antiga no console de gerenciamento do ModelArts e usar os pools de recursos criados, mas não poderá criar pool de recursos dedicados nessa página. O ModelArts permite migrar pools de recursos dedicados existentes para a nova página de gerenciamento. Você será contatado para concluir a migração e isso não exige que você execute nenhuma operação. Além disso, a migração não afeta as cargas de trabalho em execução nos pools de recursos dedicados. Preste atenção às novas funções de gerenciamento fáceis de usar dos pools de recursos dedicados. Não há mudança na criação de empregos de treinamento ou serviços de inferência.

Os pools de recursos dedicados de nova versão serão mais caros?

A unidade de cobrança e o preço unitário dos pools de recursos dedicados da nova versão são os mesmos dos pools de recursos dedicados da versão anterior. Se você não escalar dentro ou fora de seus pools de recursos dedicados, a taxa não será alterada. Além disso, mais recursos de valor agregado, como divisão de subpool, compartilhamento elástico e aceleração de dados, serão fornecidos em versões posteriores para melhor alocar recursos de computação e melhorar a relação custo-benefício.

Diferenças entre pools de recursos dedicados novos e anteriores

- Na versão antiga, os pools de recursos dedicados para desenvolvimento/treinamento são separados daqueles dedicados para implementação de serviços. Além disso, os pools dos dois tipos oferecem funções diferentes e sua experiência de usuário varia. Na nova versão, os pools de recursos dedicados dos dois tipos são unificados. Você só precisa configurar um ou vários tipos de trabalho. Em seguida, o pool de recursos dedicado suporta automaticamente o tipo de trabalho configurado.
- Novos pools de recursos dedicados herdam todas as funções dos antigos e melhoraram bastante a experiência do usuário em funções principais, como comprar e redimensionar um pool de recursos. Use novos pools de recursos dedicados para uma experiência suave e transparente.
- Além disso, os novos pools de recursos dedicados oferecem funções aprimoradas, por exemplo, permitindo que você atualize drivers de GPU ou Ascend, visualize detalhes sobre enfileiramento de tarefas e use uma rede para vários pools. Mais novas funções dos novos pools de recursos dedicados estão chegando em breve.

Como obter ajuda ou fornecer feedback se encontrar problemas durante o uso?

Semelhante a outras funções do ModelArts, você pode relatar problemas ou obter ajuda na barra lateral do console. Além disso, é aconselhável ler as seções subsequentes deste documento para entender melhor como usar os pools de recursos dedicados do ModelArts. Envie um tíquete de serviço para mais requisitos.

Instruções de pools de recursos dedicados

- Se você usa pools de recursos dedicados pela primeira vez, comece lendo**Pool de recursos**.
- Crie um pool de recursos dedicado referindo-se a **Criação de um pool de recursos**.
- Exiba os detalhes sobre um pool de recursos dedicado criado fazendo referência a**Exibição de detalhes sobre um pool de recursos**.
- Se as especificações de um pool de recursos dedicados não atenderem aos requisitos de serviço, ajuste as especificações consultando **Redimensionamento de um pool de recursos**.
- Defina ou altere os tipos de trabalho suportados por um pool de recursos dedicados, referindo-se a **Alteração de tipos de trabalho suportados por um pool de recursos**.
- Atualize o driver de GPU/Ascend de seus pools de recursos dedicados referindo-se a **Atualização de um driver de pool de recursos**.
- Se um pool de recursos dedicado não for mais necessário, exclua-o referindo-se a **Exclusão de um pool de recursos**.
- Se qualquer exceção ocorrer quando você usar um pool de recursos dedicados, trate a exceção fazendo referência a **Status anormal de um pool de recursos dedicados**.
- Gerencie redes de pool de recursos dedicados ou interconecte VPCs com as redes fazendo referência a **Rede do ModelArts**.

2.2 Criação de um pool de recursos

Esta seção descreve como criar um pool de recursos dedicado.

Procedimento

1. Efetue logon no console do ModelArts. No painel de navegação, escolha **Dedicated Resource Pools** > **Elastic Cluster**.

   NOTA

   Para novos usuários, somente os clusters elásticos de nova versão estão disponíveis no console do ModelArts. Para os usuários que usaram pools de recursos dedicados de versão antiga, eles podem acessar clusters elásticos de versão antiga e nova.

2. Na guia **Resource Pools**, clique em **Create** e configure parâmetros.

Tabela 2-1 Parâmetros de pool de recursos dedicados

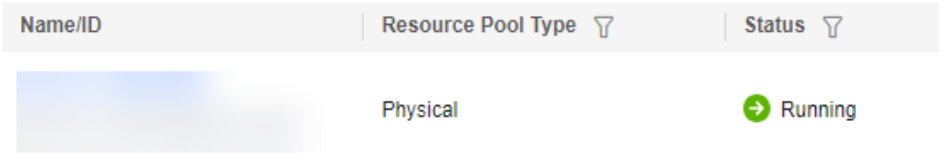
Parâmetro	Subparâmetro	Descrição
Name	N/D	Nome de um pool de recursos dedicado. Somente letras minúsculas, dígitos e hifens (-) são permitidos. O valor deve começar com uma letra minúscula e não pode terminar com um hífen (-).
Description	N/D	Breve descrição de um pool de recursos dedicado.
Billing Mode	N/D	Você pode selecionar **Pay-per-use**.
Resource Pool Type	N/D	Você pode selecionar **Physical** ou **Logical**. Se não houver nenhuma especificação lógica, **Logical** não será exibido.
Job Type	N/D	Selecione os tipos de trabalho suportados pelo pool de recursos com base nos requisitos de serviço. ● **Physical**: **DevEnviron**, **Training Job** e **Inference Service** são suportados. ● **Logical**: somente **Training Job** é suportado.
Network	N/D	Rede na qual a instância do serviço de destino está implementada. A instância pode trocar dados com outros recursos de serviço de nuvem na mesma rede. Selecione uma rede na caixa de listagem suspensa. Se nenhuma rede estiver disponível, clique em **Create** à direita para criar uma rede. Para obter detalhes sobre como criar uma rede, consulte **Criação de uma rede**.
Specification Management	Specifications	Selecione as especificações necessárias. Devido à perda do sistema, os recursos disponíveis reais são menores do que os especificados nas especificações. Depois que um pool de recursos dedicado é criado, você pode exibir os recursos disponíveis reais na página de guia **Nodes** da página de detalhes do pool de recursos dedicado.

Parâmetro	Subparâmetro	Descrição
	AZ	Você pode selecionar **Automatically allocated** ou **Specifies AZ**. Uma AZ é uma região física onde recursos usam fontes de energia e redes independentes. As AZs são fisicamente isoladas, mas interconectadas em uma intranet. ● **Automatically allocated**: as AZs são alocadas automaticamente. ● **Specifies AZ**: especificar AZs para nós do pool de recursos. Para garantir a recuperação de desastres do sistema, implemente todos os nós na mesma AZ. Você pode definir o número de nós em uma AZ.
	Nodes	Selecione o número de nós em um pool de recursos dedicado. Mais nós significam maior desempenho de computação. Se **AZ** estiver definida como **Specifies AZ**, não é necessário configurar **Nodes**. **NOTA** É uma boa prática criar não mais do que 30 nodes por vez. Caso contrário, a criação pode falhar devido à limitação de tráfego.
	Advanced Configuration	Isso permite definir o espaço do mecanismo do contêiner. Você deve inserir um inteiro para o espaço do mecanismo do contêiner. Não pode ser inferior a 50 GB, que é o valor padrão e mínimo. O valor máximo depende das especificações. Para ver os valores válidos, verifique o prompt do console. Personalizar o espaço do mecanismo do contêiner não aumenta os custos.
Custom Driver	N/D	Esse parâmetro está disponível somente quando um flavor de GPU ou Ascend é selecionado. Ative esta função e selecione um driver.
GPU Driver	N/D	Esse parâmetro está disponível somente quando o driver personalizado está habilitado. Selecione um driver de acelerador de GPU.
Required Duration	N/D	Selecione o período de tempo para o qual você deseja usar o pool de recursos. Esse parâmetro é obrigatório somente quando o modo de cobrança **Yearly/Monthly** estiver selecionado.
Auto-renewal	N/D	Especifica se a renovação automática deve ser ativada. Esse parâmetro é obrigatório somente quando o modo de cobrança **Yearly/Monthly** estiver selecionado. ● As assinaturas mensais são renovadas a cada mês. ● As assinaturas anuais são renovadas a cada ano.

Parâmetro	Subparâmetro	Descrição
Advanced Options	N/D	Selecione **Configure Now** para definir as informações de tag, o bloco CIDR e a distribuição do nó do controlador.
Tags	N/D	O ModelArts pode trabalhar com o Tag Management Service (TMS). Ao criar tarefas que consomem recursos no ModelArts, por exemplo, ttabalhos de treinamento, configure tags para que o ModelArts possa usar tags para gerenciar recursos por grupo. Para obter detalhes sobre como usar tags, consulte **Como o ModelArts usa tags para gerenciar recursos por grupo?** **NOTA** Você pode selecionar uma tag do TMS predefinida na lista suspensa de tags ou personalizar uma tag. As tags predefinidas estão disponíveis para todos os recursos de serviço que suportam tags. As tags personalizadas estão disponíveis apenas para os recursos de serviço do usuário que criou as tags.
CIDR block	N/D	Você pode selecionar **Default** ou **Custom**. ● **Default**: o sistema aloca aleatoriamente um bloco CIDR disponível para você, que não pode ser modificado depois que o pool de recursos é criado. Para uso comercial, personalize seu bloco CIDR. ● **Custom**: você precisa personalizar o contêiner de K8S e os blocos CIDR de serviço K8S. – **K8S Container Network**: usado pelo contêiner em um cluster, que determina quantos contêineres podem existir em um cluster. O valor não pode ser alterado após a criação do pool de recursos. – **K8S Service Network**: usado quando os contêineres no mesmo cluster acessam uns aos outros, o que determina quantos serviços podem existir. O valor não pode ser alterado após a criação do pool de recursos.
Master Distribution	N/D	Locais de distribuição dos nós do controlador. Você pode selecionar **Random** ou **Custom**. ● **Random**: use as AZs alocadas aleatoriamente pelo sistema. ● **Custom**: selecione AZs para os nós do controlador. Distribua nós do controlador em diferentes AZs para recuperação de desastres.

3. Clique em **Next** e confirme as configurações. Em seguida, clique em **Submit** para criar o pool de recursos dedicado.
   – Depois que um pool de recursos é criado, seu status muda para **Running**. Somente quando o número de nós disponíveis for maior que 0, as tarefas podem ser entregues a esse pool de recursos.

Figura 2-1 Exibir um pool de recursos



– Passe o cursor sobre **Creating** para exibir os detalhes sobre o processo de criação. Clique em **View Details**. A página de registro da operação é exibida.

Figura 2-2 Criando

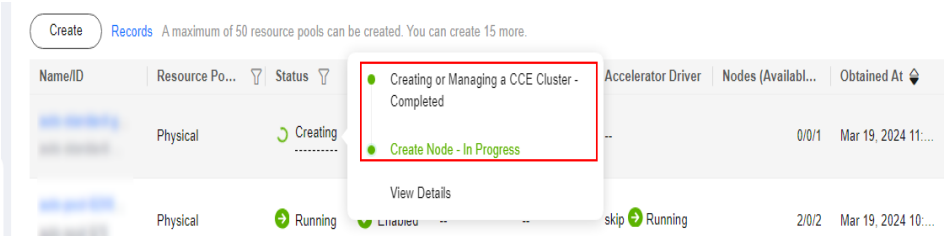
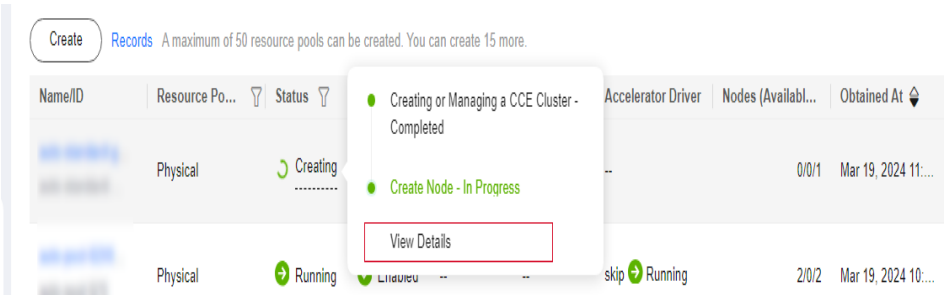


Figura 2-3 Visualizar detalhes



– Você pode exibir os registros de tarefas do pool de recursos clicando em **Records** no canto superior esquerdo da lista do pool de recursos.

Figura 2-4 Registros da operação

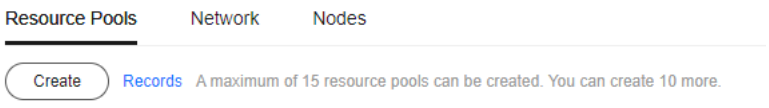
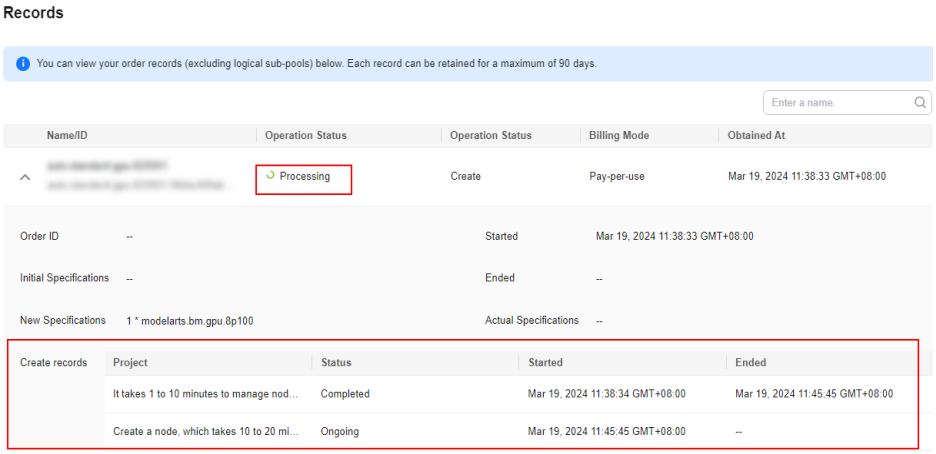


Figura 2-5 Exibir o status do pool de recursos



Perguntas frequentes

E se eu escolher um flavor para um pool de recursos dedicado, mas receber uma mensagem de erro dizendo que nenhum recurso está disponível?

Os flavors de recursos dedicados mudam com base na disponibilidade em tempo real. Às vezes, você pode escolher um flavor na página de compra, mas ele está esgotado antes de você pagar e criar o pool de recursos. Isso faz com que a criação do pool de recursos falhe.

Você pode tentar um flavor diferente na página de criação e criar o pool de recursos novamente.

P: por que não posso usar todos os recursos da CPU em um nó em um pool de recursos?

Os nós do pool de recursos têm sistemas e plug-ins instalados neles. Estes ocupam alguns recursos da CPU. Por exemplo, se um nó tiver 8 vCPUs, mas algumas delas forem usadas por componentes do sistema, os recursos disponíveis serão menos de 8 vCPUs.

Você pode verificar os recursos de CPU disponíveis clicando na guia **Nodes** na página de detalhes do pool de recursos, antes de iniciar uma tarefa.

2.3 Exibição de detalhes sobre um pool de recursos

Página de detalhes do pool de recursos

- Efetue logon no console do ModelArts. No painel de navegação, escolha **Dedicated Resource Pools** > **Elastic Cluster**.
- Clique em ao lado do tipo de pool de recursos ou status no cabeçalho da tabela. No canto superior direito da lista, selecione **Name** ou **Resource ID** para filtrar pools de recursos. Para obter o ID do recurso, acesse a página **Billing Center** > **Orders** > **My Orders** e clique em **Details** na coluna **Operation** do pedido de destino.
- Na lista de pool de recursos, clique em um pool de recursos para ir para sua página de detalhes e exibir suas informações.
  - Se houver vários pools de recursos, clique em no canto superior esquerdo da página de detalhes de um pool de recursos para alternar entre pools de recursos. Clique em **More** no canto superior direito para executar operações como

redimensionar ou excluir o pool de recursos. As operações disponíveis variam dependendo do pool de recursos.

- Na área **Network** de **Basic Information**, você pode clicar no número de pools de recursos associados para exibir pools de recursos associados.
- Na área de informações estendidas, você pode exibir as informações de monitoramento, trabalhos, nós, especificações e eventos. Para obter detalhes, consulte a seguinte seção.

Exibir trabalhos em um pool de recursos

Na página de detalhes do pool de recursos, clique em **Jobs**. Você pode exibir todos os trabalhos em execução no pool de recursos. Se um trabalho estiver enfileirando, você poderá visualizar sua posição de enfileiramento.

NOTA

Somente os trabalhos de treinamento podem ser visualizados.

Figura 2-6 Trabalhos

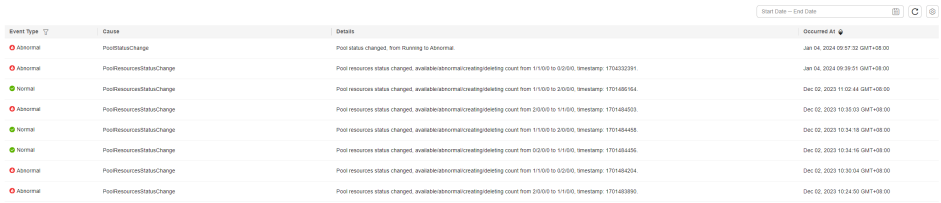
Exibir eventos do pool de recursos

Na página de detalhes do pool de recursos, clique em **Events**. Você pode exibir todos os eventos do pool de recursos. A causa de um evento é **PoolStatusChange** ou **PoolResourcesStatusChange**.

Na lista de eventos, clique em à direita de **Event Type** para filtrar eventos.

- Quando um pool de recursos começa a ser criado ou se torna anormal, o status do pool de recursos muda e a alteração é registrada como um evento.
- Quando o número de nós disponíveis ou anormais ou em processo de criação ou exclusão for alterado, o status do nó do pool de recursos será alterado e a alteração será registrada como um evento.

Figura 2-7 Eventos



Exibir nós do pool de recursos

Na página de detalhes do pool de recursos, clique em **Nodes**. Você pode exibir todos os nós no pool de recursos e o uso de recursos de cada nó.

Alguns recursos são reservados para componentes de cluster. Portanto, **CPUs (Available/Total)** não indica o número de recursos físicos no nó. Ele exibe apenas o número de recursos

que podem ser usados pelos serviços. Os núcleos da CPU são medidos em milinúcleos, e 1000 milinúcleos equivalem a 1 núcleo físico.

- Substituir um nó:

  Na guia **Nodes**, localize o nó a ser substituído. Na coluna **Operation**, clique em **Replace**. Nenhuma taxa é cobrada para esta operação.

  Verifique os registros de substituição de nó na página **Records**. **Running** indica que o nó está sendo substituído. Após a substituição, você pode verificar o novo nó na lista de nós.

  A substituição não pode durar mais de 24 horas. Se nenhum recurso adequado for encontrado após o tempo limite de substituição, o status mudará para **Failed**. Passe o mouse sobre  para verificar a causa da falha.

  NOTA

  - O número de substituições por dia não pode exceder 20% do total de nós no pool de recursos. O número de nós a serem substituídos não pode exceder 5% do total de nós no pool de recursos.
  - Certifique-se de que haja recursos de nó ociosos. Caso contrário, a substituição pode falhar.
  - Se houver nós no estado **Resetting** nos registros de operação, os nós no pool de recursos não poderão ser substituídos.

- Redefinir um nó

  Na guia **Nodes**, localize o nó que deseja redefinir. Clique em **Reset** na coluna **Operation** para redefinir um nó. Você também pode selecionar vários nós e clicar em **Reset** para redefinir vários nós.

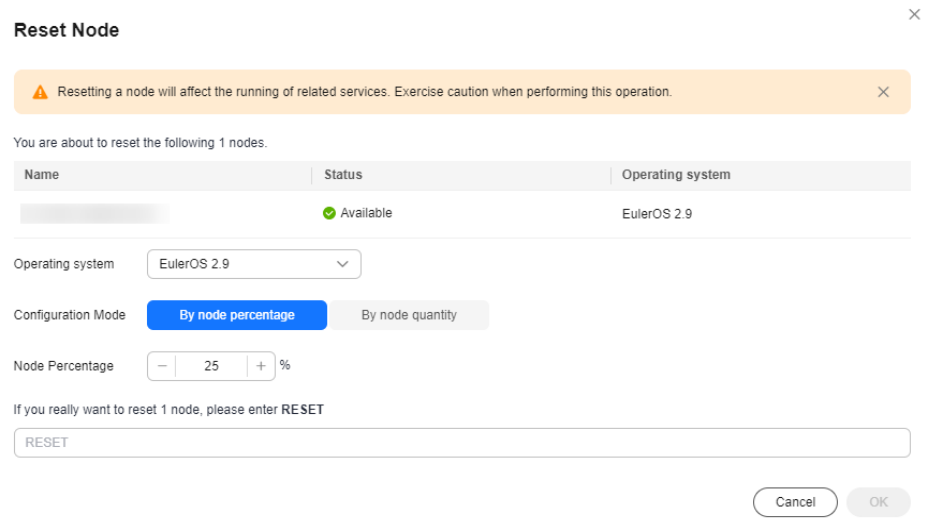
  Configure os parâmetros descritos na tabela abaixo.

  Tabela 2-2 Parâmetros

Parâmetro	Descrição
Operating System	Selecione um sistema operacional na caixa de listagem suspensa.
Configuration Mode	Selecione um modo de configuração para redefinir o nó. ● **By node percentage**: a proporção máxima de nós que podem ser redefinidos se houver vários nós na tarefa de redefinição ● **By node quantity**: o número máximo de nós que podem ser redefinidos se houver vários nós na tarefa de redefinição

  Verifique os registros de redefinição de nó na página **Records**. Se o nó está sendo reiniciado, seu status é **Resetting**. Após a redefinição ser concluída, o status do nó muda para **Available**. A redefinição de um nó não será cobrada.

Figura 2-8 Redefinir um nó

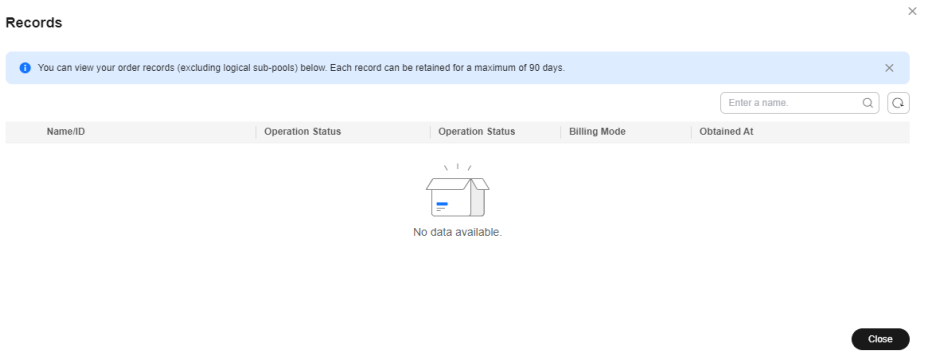


NOTA

- A redefinição de um nó afetará os serviços em execução.
- Somente os nós no estado **Available** podem ser redefinidos.
- Um único nó pode estar em apenas uma tarefa de redefinição por vez. Várias tarefas de redefinição não podem ser entregues ao mesmo nó por vez.
- Se houver nós no estado **Replacing** nos registros de operação, os nós no pool de recursos não poderão ser redefinidos.
- Quando o driver de um pool de recursos está sendo atualizado, os nós nesse pool de recursos não podem ser redefinidos.
- Para especificações de GPU e NPU, após o nó ser redefinido, o driver do nó pode ser atualizado. Espere pacientemente.

Figura 2-9 Nós

Figura 2-10 Registros da operação



- Exclusão, cancelamento de assinatura ou liberação de um nó
  - Para um pool de recursos de pagamento por uso, clique em **Delete** na coluna **Operation**.

Para excluir nós em lotes, marque as caixas de seleção ao lado dos nomes dos nós e clique em **Delete**.

- Para um pool de recursos anual/mensal cujos recursos não estejam expirados, clique em **Unsubscribe** na coluna **Operation**.
- Para um pool de recursos anual/mensal cujos recursos tenham expirado (no período de tolerância), clique em **Release** na coluna **Operation**.

Se o botão de exclusão estiver disponível para um nó anual/mensal, se o nó for um nó de inventário, clique em **Delete**.

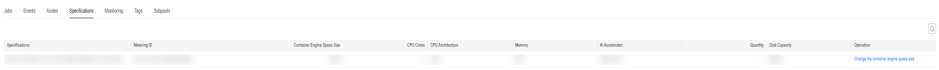
NOTA

- Antes de excluir, cancelar a assinatura ou liberar um nó, verifique se não há tarefas em execução nesse nó. Caso contrário, os trabalhos serão interrompidos.
- Excluir, cancelar a assinatura ou liberar nós anormais em um pool de recursos e adicionar novos para substituição.
- Se houver apenas um nó, ele não pode ser excluído, cancelado ou liberado.

Exibir especificações do pool de recursos

Na página de detalhes do pool de recursos, clique em **Specifications**. Você pode exibir as especificações usadas pelo pool de recursos e o número de cada especificação.

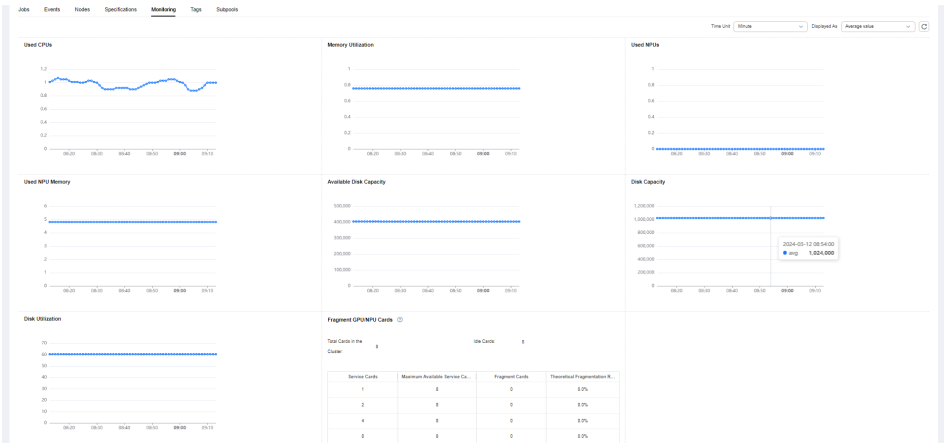
Figura 2-11 Exibir especificações do pool de recursos (O tamanho do mecanismo de contêiner é exibido como o valor padrão se não estiver definido.)



Exibir informações de monitoramento do pool de recursos

Na página de detalhes do pool de recursos, clique em **Monitoring**. O uso de recursos, incluindo CPUs usadas, uso de memória e capacidade de disco disponível do pool de recursos, é exibido. Se aceleradores de IA forem usados no pool de recursos, as informações de monitoramento de GPU e NPU também serão exibidas.

Figura 2-12 Exibir visualizações de recursos

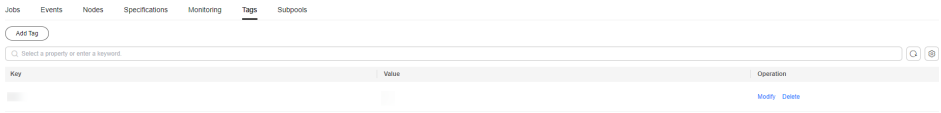


Visualização de tags

Você pode adicionar tags a um pool de recursos para pesquisa rápida.

Na página de detalhes do pool de recursos, clique em **Tags**. Você pode exibir, adicionar, modificar e excluir tags de um pool de recursos. Para obter detalhes sobre como usar tags, consulte **Como o ModelArts usa tags para gerenciar recursos por grupo?**

Figura 2-13 Tags



NOTA

Você pode adicionar até 20 tags.

2.4 Redimensionamento de um pool de recursos

Descrição

A demanda por recursos em um pool de recursos dedicados pode mudar devido às mudanças nos serviços de desenvolvimento de IA. Nesse caso, você pode redimensionar seu pool de recursos dedicados no ModelArts.

- Você pode adicionar nós para flavors existentes no pool de recursos.
- Você pode excluir nós de variações existentes no pool de recursos.

NOTA

Antes de reduzir um pool de recursos, verifique se não há serviços em execução no pool. Como alternativa, vá para a página de detalhes do pool de recursos, exclua os nós onde nenhum serviço está sendo executado para reduzir no pool.

Restrições

- Somente pools de recursos dedicados no status **Running** podem ser redimensionados.
- Ao reduzir um pool de recursos dedicado, o número de flavors ou nós de um flavor não pode ser reduzido para 0.

Redimensionar um pool de recursos dedicado

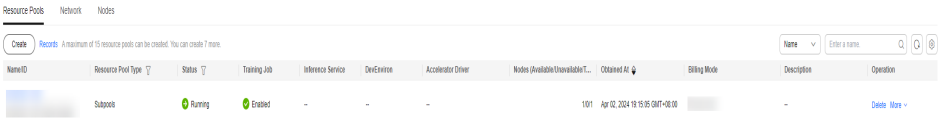
Você pode redimensionar um pool de recursos de qualquer uma das seguintes maneiras:

- Ajustar o número de nós das especificações existentes
- Redimensionar o espaço do mecanismo do contêiner

1. Faça logon no console de gerenciamento do ModelArts. No painel de navegação, escolha **Dedicated Resource Pools** > **Elastic Cluster**.

> **NOTA**
>
> Um pool de recursos é suspenso quando é migrado da versão anterior para a nova versão. Você não pode ajustar a capacidade de tal pool de recursos ou cancelar a assinatura dele.

Figura 2-14 Pools de recursos



2. Adicione ou exclua nós.

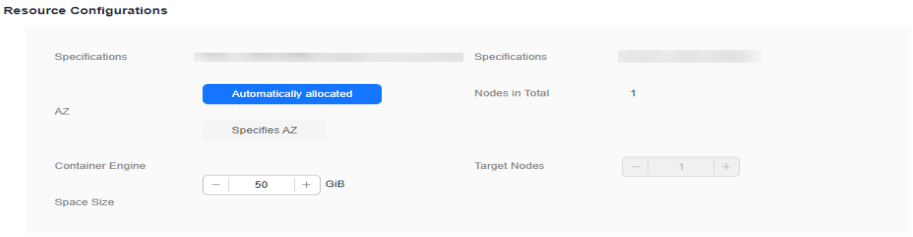
   Clique em **Adjust Capacity** na coluna **Operation** do pool de recursos de destino.

   Na área **Resource Configurations**, defina **AZ** como **Automatically allocated** ou **Specifies AZ**. Clique em **Submit** e depois em **OK** para salvar as alterações.

   – Se **AZ** estiver definida como **Automatically allocated**, você poderá aumentar ou diminuir o número de nós a serem dimensionados ou no pool de recursos. Após o dimensionamento, os nós são automaticamente alocados para as AZs.

   – Se você selecionar **Specifies AZ**, poderá alocar nós a diferentes AZs.

Figura 2-15 Configurações de recurso



3. Redimensione o espaço do mecanismo do contêiner.

   Se você precisar de um tamanho maior de mecanismo de contêiner, execute uma das seguintes operações:

   – Para novos recursos, você pode especificar o espaço do mecanismo de contêiner ao criar um pool de recursos. Para obter detalhes, consulte configurações avançadas do **Gerenciamento de especificações** em **Criação de um pool de recursos**.

   – Para recursos existentes, o espaço do mecanismo de contêiner pode ser modificado.

   ■ Método 1: clique no pool de recursos de destino para exibir seus detalhes. Clique na guia **Specifications**, localize as especificações de destino e clique em **Change the container engine space size** na coluna **Operation**.

   ■ Método 2: localize o pool de recursos de destino e clique em **Adjust Capacity** na coluna **Operation**.

> **AVISO**
>
> O redimensionamento do espaço do mecanismo de contêiner só é aplicável a novos nós. Além disso, dockerBaseSize pode variar entre nós desse flavor dentro do pool de recursos. Consequentemente, isso pode levar a discrepâncias no status das tarefas distribuídas entre os diferentes nós.

Figura 2-16 Redimensionar o espaço do mecanismo de contêiner (guia **Specifications**)

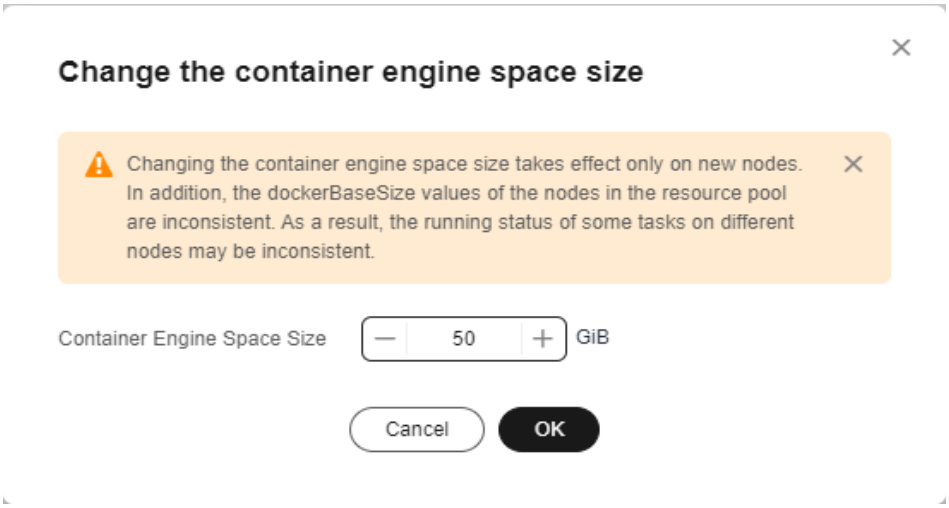
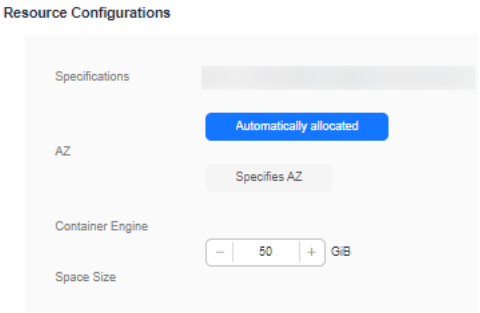


Figura 2-17 Redimensionar o espaço do mecanismo de contêiner (página **Resize**)



2.5 Definição de uma política de renovação

Descrição

O ModelArts permite que você execute as seguintes operações para pools de recursos anuais/mensais:

- Ative a renovação automática.
- Modifique as configurações de renovação automática.
- Renove-as manualmente.

Restrições

O pool de recursos dedicados de destino deve estar em execução.

Procedimento

1. Efetue logon no console do ModelArts. No painel de navegação, escolha **Dedicated Resource Pools** > **Elastic Cluster**.
2. Na lista de pool de recursos, escolha **More** > **Set Renewal Policy** na coluna **Operation** do pool de recursos de destino.

3. Na caixa de diálogo exibida, clique em **OK**. Você verá a página **Renewals** do centro de cobrança.
4. Defina a política de renovação.
   - Para habilitar a renovação automática para um pool de recursos anual/mensal, clique na guia **Manual Renewals**, localize o pool de recursos de destino e escolha **More** > **Enable Auto-Renewal** na coluna **Operation**.
   - Para modificar as configurações de renovação automática de um pool de recursos anual/mensal, clique na guia **Auto Renewals**, localize o pool de recursos de destino e escolha **More** > **Modify Auto-Renew** na coluna **Operation** para modificar as configurações de renovação automática, como o modo de renovação, a duração da renovação e o número de renovações.
   - Para renovar manualmente um pool de recursos anual/mensal, localize-o e clique em **Renew** na coluna **Operation**.

2.6 Modificação da política de expiração

Descrição

O ModelArts permite alterar a política de expiração de um pool de recursos anual/mensal para pagamento por uso ou não renovação após a expiração.

Restrições

O pool de recursos dedicados de destino deve estar em execução.

Procedimento

1. Efetue logon no console do ModelArts. No painel de navegação, escolha **Dedicated Resource Pools** > **Elastic Cluster**.
2. Na lista de pool de recursos, escolha **More** > **Change Billing Mode** na coluna **Operation** do pool de recursos de destino.
3. Na caixa de diálogo exibida, clique em **OK**. Você verá a página **Renewals** da central de cobrança.
4. Modifique a política de expiração.
   - Se a renovação automática não tiver sido ativada para o pool de recursos de destino, clique na guia **Manual Renewals** e escolha **More** > **Change to Pay-per-Use After Expiration** or **More** > **Cancel Renewal** na coluna **Operation** do pool de recursos de destino.
   - Se a renovação automática tiver sido ativada para o pool de recursos de destino, clique na guia **Auto Renewals** e escolha **More** > **Cancel Renewal** na coluna **Operation** do pool de recursos de destino.

2.7 Migração do espaço de trabalho

Contexto

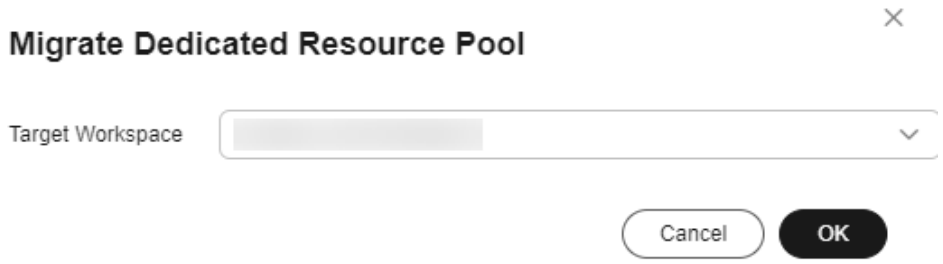
O espaço de trabalho de um pool de recursos dedicado está associado a um projeto empresarial, que envolve a coleta de faturas. O ModelArts fornece espaços de trabalho para

isolar permissões de operação de recursos de diferentes usuários do IAM. A migração do espaço de trabalho inclui a migração do pool de recursos e a migração de rede. Para obter detalhes, consulte as seguintes seções.

Migrar o espaço de trabalho para um pool de recursos

1. Faça logon no console de gerenciamento do ModelArts. No painel de navegação, escolha **Dedicated Resource Pools** > **Elastic Cluster**.
2. Na lista de pool de recursos, escolha **More** > **Migrate Workspace** na coluna **Operation** do pool de recursos de destino.
3. Na caixa de diálogo **Migrate Dedicated Resource Pool** exibida, selecione o espaço de trabalho de destino e clique em **OK**.

Figura 2-18 Migrar o espaço de trabalho



Migrar o espaço de trabalho para uma rede

1. Faça logon no console de gerenciamento do ModelArts. No painel de navegação, escolha **Dedicated Resource Pools** > **Elastic Cluster**. Em seguida, clique na guia **Network**.
2. Na lista de redes, escolha **More** > **Migrate Workspace** na coluna **Operation** da rede de destino.
3. Na caixa de diálogo exibida, selecione a área de trabalho de destino e clique em **OK**.

Figura 2-19 Migrar o espaço de trabalho

2.8 Alteração de tipos de trabalho suportados por um pool de recursos

Descrição

O ModelArts oferece suporte a muitos tipos de trabalhos. Alguns deles podem ser executados em pools de recursos dedicados, incluindo trabalhos de treinamento, serviços de inferência e ambientes de desenvolvimento de notebooks.

Você pode alterar os tipos de cargo suportados por um pool de recursos dedicado. As opções disponíveis para **Job Type** são **Training Job**, **Inference Service** e **DevEnviron**.

Somente tipos selecionados de trabalhos podem ser entregues ao pool de recursos dedicado correspondente.

⚠ CUIDADO

Para suportar diferentes tipos de trabalho, diferentes operações são executadas no back-end, como a instalação de plug-ins e a configuração do ambiente de rede. Algumas operações usam recursos no pool de recursos. Como resultado, os recursos disponíveis para você diminuem. Portanto, selecione apenas os tipos de trabalho necessários para evitar o desperdício de recursos.

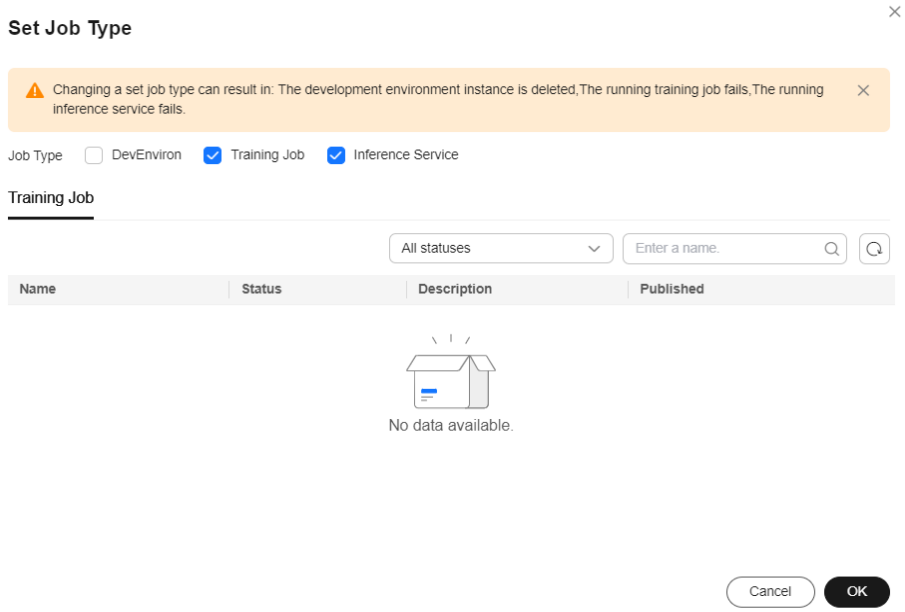
Restrições

O pool de recursos dedicados de destino deve estar em execução.

Procedimento

1. Faça logon no console de gerenciamento do ModelArts. No painel de navegação, escolha **Dedicated Resource Pools** > **Elastic Cluster**.
2. Na coluna **Operation** de um pool de recursos, escolha **More** > **Set Job Type**.
3. Na caixa de diálogo **Set Job Type**, selecione os tipos de trabalho.

Figura 2-20 Configurar o tipo de trabalho



4. Clique em **OK**.

2.9 Atualização de um driver de pool de recursos

Descrição

Se GPUs ou recursos de Ascend forem usados em um pool de recursos dedicado, talvez seja necessário personalizar os drivers da GPU ou Ascend. O ModelArts permite que você atualize drivers de GPU ou Ascend de seus pools de recursos dedicados.

Existem dois modos de atualização de driver: atualização segura e atualização forçada.

NOTA

- Atualização segura: os serviços em execução não são afetados. Após o início da atualização, os nós são isolados (novos trabalhos não podem ser entregues). Depois que os trabalhos existentes nos nós são concluídos, a atualização é realizada. A atualização segura pode levar muito tempo porque os trabalhos devem ser concluídos primeiro.
- Atualização forçada: os drivers são atualizados diretamente, independentemente de haver trabalhos em execução.

Restrições

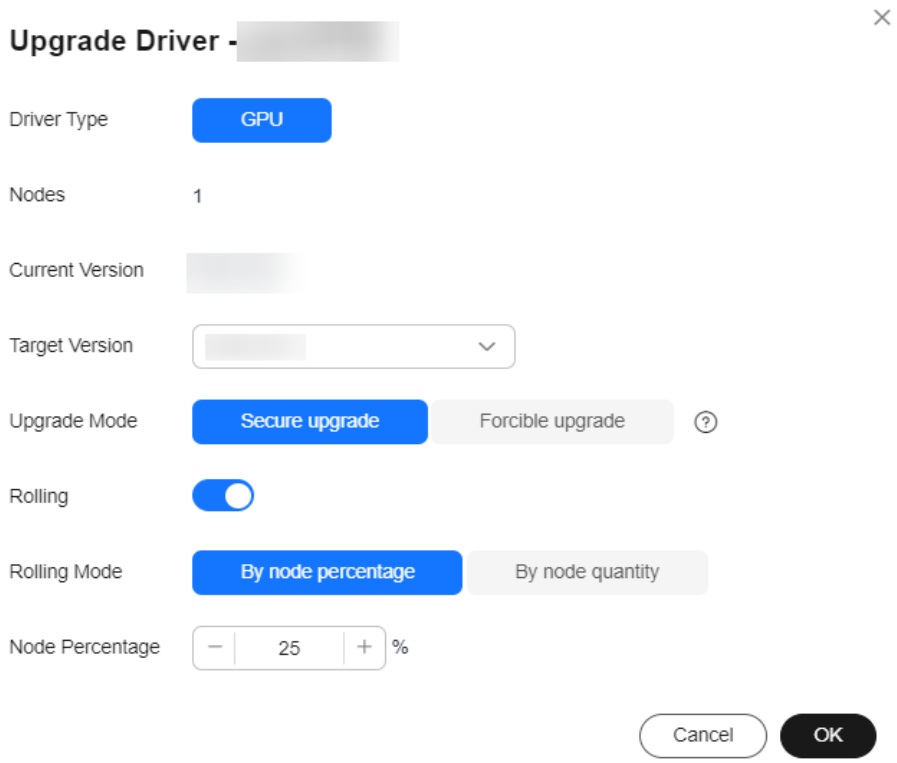
- O pool de recursos dedicados de destino deve estar em execução e o pool de recursos contém recursos de GPU ou Ascend.
- Para um pool de recursos lógicos, o driver pode ser atualizado somente após a vinculação de nó ser ativada. Para ativar a vinculação de nó, envie um tíquete de serviço para entrar em contato com os engenheiros da Huawei.

Atualizar o driver

1. Faça logon no console de gerenciamento do ModelArts. No painel de navegação, escolha **Dedicated Resource Pools** > **Elastic Cluster**.

2. Na coluna **Operation** do pool de recursos de destino, escolha **More** > **Upgrade Driver**.
3. Na caixa de diálogo **Upgrade Driver**, o tipo de driver, o número de nós, a versão atual, a versão de destino e o modo de atualização do pool de recursos dedicados são exibidos.
   - **Target Version**: selecione uma versão do driver de destino na lista suspensa.
   - **Upgrade Mode**: selecione **Secure upgrade** ou **Forcible upgrade**.
   - **Rolling Mode**: uma vez ativado, você pode atualizar o driver no modo contínuo. Atualmente, a rolagem por porcentagem de nó e por quantidade de nó são suportadas. Se **By node percentage** for selecionada, o número de nós a serem atualizados em cada lote será a proporção de nós multiplicada pelo número total de nós no pool de recursos. Se **By node quantity** for selecionada, o número de nós a serem atualizados em cada lote é o que você configurou.

Figura 2-21 Atualizar um driver



4. Clique em **OK** para iniciar a atualização do driver.

2.10 Exclusão de um pool de recursos

Se um pool de recursos dedicado não for mais necessário para o desenvolvimento de serviços de IA, você poderá excluir o pool de recursos para liberar recursos.

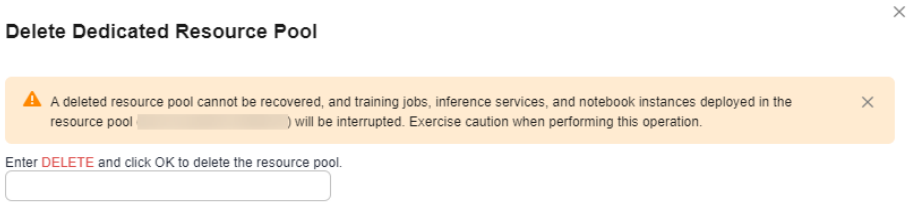
NOTA

Depois que um pool de recursos dedicado é excluído, os ambientes de desenvolvimento, trabalhos de treinamento e serviços de inferência que dependem do pool de recursos ficam indisponíveis. Um pool de recursos dedicado não pode ser restaurado após ser excluído.

1. Faça logon no console de gerenciamento do ModelArts. No painel de navegação, escolha **Dedicated Resource Pools** > **Elastic Cluster**.
2. Localize a linha que contém o pool de recursos de destino e escolha **More** > **Delete** na coluna **Operation**.
3. Na caixa de diálogo **Delete Dedicated Resource Pool**, insira **DELETE** na caixa de texto e clique em **OK**.

   Você pode alternar entre guias na página de detalhes para exibir os trabalhos de treinamento e as instâncias do bloco de anotações criadas usando o pool de recursos e serviços de inferência implementados no pool de recursos.

Figura 2-22 Exclusão de um pool de recursos



2.11 Status anormal de um pool de recursos dedicados

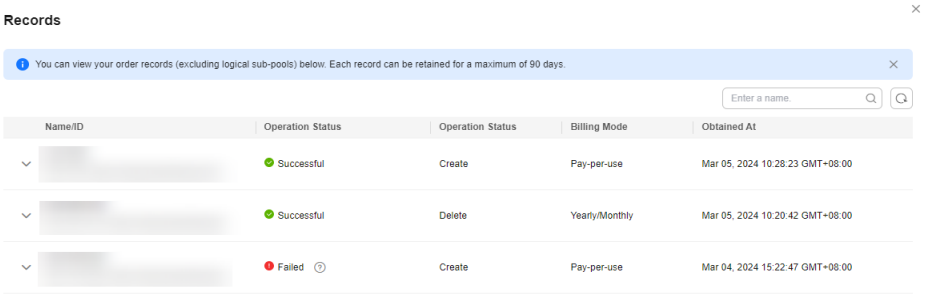
Limite de cota de recursos

Quando você usa um pool de recursos dedicados (por exemplo, dimensionamento de recursos, criação de uma VPC, criação de uma VPC e sub-rede ou interconexão de uma VPC), se o sistema exibir uma mensagem indicando que a cota de recursos é limitada, **envie um tíquete de serviço**.

Falhou na criação/falhou na alteração

1. Faça logon no console de gerenciamento do ModelArts. No painel de navegação, escolha **Dedicated Resource Pools** > **Elastic Cluster**.
2. Clique em **Records** à direita de **Create**. Na caixa de diálogo **Records**, visualize registros de tarefas com falha.

Figura 2-23 Falhou ao criar um pool de recursos



3. Passe o cursor sobre , veja a causa das falhas da tarefa.

> **NOTA**
>
> Por padrão, os registros de tarefas com falha são classificados por hora da aplicação. Um máximo de 500 registros de tarefas com falha podem ser exibidos e mantidos por três dias.

Localizar nó defeituoso

O ModelArts adicionará uma mancha em um nó defeituoso do K8S detectado para que os trabalhos não sejam afetados ou agendados para o nó contaminado. A tabela a seguir lista as falhas que podem ser detectadas. Você pode localizar a falha consultando o código de isolamento e o método de detecção.

Tabela 2-3 Código de isolamento

Código de isolamento	Categoria	Sub-categoria	Descrição	Método de detecção
A050101	GPU	Memória de GPU	Erro de ECC da GPU.	Execute o comando **nvidia-smi -a** e verifique se **Pending Page Blacklist** é **Yes** ou se o valor de **multi-bit Register File** é maior que 0. Para GPUs Ampere, verifique se o seguinte conteúdo existe: ● Erro de SRAM incorrigível ● Registros de falha de remapeamento ● Eventos Xid 95 em **dmsg** (Para obter detalhes, consulte **Gerenciamento de erros de memória da GPU de NVIDIA**.) A arquitetura Ampere tem os seguintes níveis de erros de memória da GPU: ● L1: estes são erros ECC de bit único que podem ser corrigidos. Eles não afetam os serviços em execução. Para verificar esses erros, execute o comando **nvidia-smi -a** e procure por **Volatile Correctable**. ● L2: estes são erros ECC de vários bits que não podem ser corrigidos. Eles fazem com que os serviços em execução falhem e exigem uma reinicialização do processo para recuperar. Para verificar esses erros, execute o comando **nvidia-smi -a** e procure por **Volatile Uncorrectable**. ● L3: estes são erros não suprimidos e podem afetar outros serviços. Eles exigem uma reinicialização da placa ou uma reinicialização do nó para limpar. Para verificar esses erros, procure os eventos Xid que contêm o número 95. (Os registros de Remapeados pendentes são apenas para referência. Você precisa redefinir as placas quando o serviço estiver ocioso para acionar o processo de remapeamento.) ● L4: são erros que exigem a substituição da placa. Para verificar

Código de isolamento	Categoria	Subcategoria	Descrição	Método de detecção
				esses erros, procure o campo **SRAM Uncorrectable** que é maior que 4 ou o campo **Remapped Failed** que não é zero.
A050102	GPU	Outros	A saída de **nvidia-smi** contém ERR.	Execute **nvidia-smi -a** e verifique se a saída contém ERR. Normalmente, o hardware, como a fonte de alimentação ou o ventilador, está com defeito.
A050103	GPU	Outros	A execução de **nvidia-smi** expira ou não existe.	Verifique se o código de saída de **nvidia-smi** não é **0**.
A050104	GPU	Memória de GPU	Erro de ECC ocorreu 64 vezes.	Execute o comando **nvidia-smi -a**, localize **Retired Pages** e verifique se a soma de **Single Bit** e **Double Bit** é maior que 64.
A050148	GPU	Outros	Um alarme de infoROM ocorre.	Execute o comando **nvidia-smi** e verifique se a saída contém o alarme "infoROM is corrupted".
A050109	GPU	Outros	Outros erros de GPU	Verifique se existe outro erro de GPU. Normalmente, há um hardware defeituoso. Entre em contato com o engenheiro técnico.
A050147	IB	Ligação	A NIC IB é anormal.	Execute o comando **ibstat** e verifique se a NIC não está no estado ativo.
A050121	NPU	Outros	Uma exceção de driver é detectada pela NPU DCMI.	O ambiente do driver da NPU é anormal.
A050122	NPU	Outros	O dispositivo de DCMI da NPU é anormal.	O dispositivo da NPU é anormal. A interface Ascend DCMI retorna um alarme importante ou urgente.
A050123	NPU	Ligação	A rede de DCMI da NPU é anormal.	A conexão de rede da NPU é anormal.
A050129	NPU	Outros	Outros erros da NPU	Verifique se existe outro erro da NPU. Você não pode corrigir a falha. Entre em contato com o engenheiro técnico.

Código de isolamento	Categoria	Sub-categoria	Descrição	Método de detecção
A050149	NPU	Ligação	Verifique se a porta de rede da ferramenta hccn está desconectada intermitentemente.	A rede da NPU é instável e desconectada intermitentemente. Execute o comando **hccn_tool-i ${device_id} -link_stat -g** e a rede será desconectada mais de cinco vezes em 24 horas.
A050951	NPU	Memória de GPU	O número de ECCs da NPU atinge o limite de manutenção.	O valor de HBM Double Bit Isolated Pages Count da NPU é maior ou igual a 64.
A050146	Tempo de execução	Outros	O NTP está anormal.	O serviço ntpd ou chronyd é anormal.
A050202	Tempo de execução	Outros	O nó não está pronto.	O nó não está disponível. O nó do K8S contém uma das seguintes manchas: ● node.kubernetes.io/unreachable ● node.kubernetes.io/not-ready
A050203	Tempo de execução	Desconexão	O número de placas de IA normais não corresponde à capacidade real.	A GPU ou NPU está desconectada.
A050206	Tempo de execução	Outros	O disco rígido do Kubelet é somente leitura.	O diretório **/mnt/paas/kubernetes/kubelet** é somente leitura.
A050801	Gerenciamento de nó	O&M do nó	O recurso é reservado.	O nó é marcado como o nó em espera e contém uma mancha.
A050802	Gerenciamento de nó	O&M do nó	Ocorre um erro desconhecido.	O nó está marcado com uma mancha desconhecida.
A200001	Gerenciamento de nó	Atualização do driver	A GPU está sendo atualizada.	A GPU está sendo atualizada.

Código de isolamento	Categoria	Subcategoria	Descrição	Método de detecção
A200002	Gerenciamento de nó	Atualização do driver	A NPU está sendo atualizada.	A NPU está sendo atualizada.
A200008	Gerenciamento de nó	Admissão de nó	A admissão está sendo examinada.	A admissão está sendo examinada, incluindo verificação básica de configuração de nó e verificação de serviço simples.
A050933	Gerenciamento de nó	Tolerância a falhas de Failover	O serviço de Failover no nó contaminado será migrado.	O serviço de Failover no nó contaminado será migrado.
A050931	Toolkit de treinamento	Contêiner de pré-verificação	Um erro de GPU é detectado no contêiner de pré-verificação.	Um erro de GPU é detectado no contêiner de pré-verificação.
A050932	Toolkit de treinamento	Contêiner de pré-verificação	Um erro de IB é detectado no contêiner de pré-verificação.	Um erro de IB é detectado no contêiner de pré-verificação.

2.12 Rede do ModelArts

Rede do ModelArts e da VPC

As redes do ModelArts são suportadas por VPCs e usadas para interconectar nós em um pool de recursos do ModelArts. Você só pode configurar o nome e o bloco CIDR para uma rede. Para garantir que não haja nenhum segmento de endereço IP no bloco CIDR sobreposto ao da VPC a ser acessada, vários blocos CIDR estão disponíveis para você selecionar.

Uma VPC fornece uma rede virtual isolada logicamente para suas instâncias. Você pode configurar e gerenciar a rede conforme necessário. A VPC fornece redes virtuais logicamente isoladas, configuráveis e gerenciáveis para servidores em nuvem, contêineres em nuvem e bancos de dados em nuvem. Ela ajuda você a melhorar a segurança do serviço de nuvem e simplificar a implementação de rede.

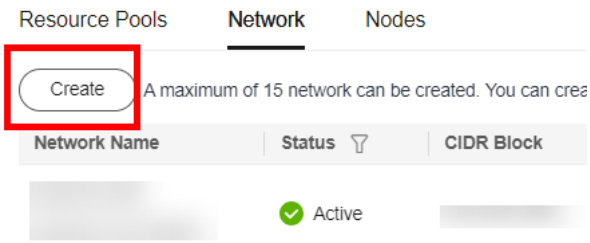
Pré-requisitos

- Uma VPC está disponível.
- Uma sub-rede está disponível.

Criação de uma rede

1. Faça logon no console de gerenciamento do ModelArts. No painel de navegação, escolha **Dedicated Resource Pools** > **Elastic Cluster**.
2. Clique em **Network** e depois em **Create**.

   Figura 2-24 Lista de redes

3. Na caixa de diálogo **Create Network**, defina os parâmetros.
   - **Network Name**: nome personalizável
   - **CIDR Block**: você pode selecionar **Preset** ou **Custom**.

   NOTA

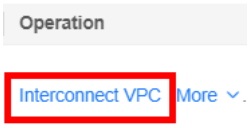
   - Cada usuário pode criar no máximo 15 redes.
   - Verifique se não há nenhum segmento de endereço IP no bloco CIDR que se sobreponha ao da VPC a ser acessada. O bloco CIDR não pode ser alterado após a criação da rede. Possíveis conflitos de blocos CIDR são os seguintes:
     - Seu bloco CIDR da VPC
     - Bloco CIDR do contêiner (consistentemente para ser 172.16.0.0/16)
     - Bloco CIDR de serviço (consistentemente 10.247.0.0/16)
4. Confirme as configurações e clique em **OK**.

(Opcional) Interconexão de uma VPC com uma rede do ModelArts

A interconexão VPC permite que você use recursos entre VPCs, melhorando a utilização de recursos.

1. Na página **Network**, clique em **Interconnect VPC** na coluna **Operation** da rede de destino.

   Figura 2-25 Interconectar VPC

2. Na caixa de diálogo exibida, clique no botão à direita de **Interconnect VPC** e selecione uma VPC e uma sub-rede disponíveis nas listas suspensas.

NOTA

A rede de par a ser interconectada não pode sobrepor-se com o bloco CIDR atual.

Figura 2-26 Parâmetros para interconexão de uma VPC com uma rede

- Se nenhuma VPC estiver disponível, clique em **Create VPC** à direita para criar uma VPC.
- Se nenhuma sub-rede estiver disponível, clique em **Create Subnet** à direita para criar uma sub-rede.
- Várias sub-redes numa VPC podem ser interconectadas. Você pode clicar em + para adicionar até 10 sub-redes.

Ativar um pool de recursos dedicados para acessar a Internet

Para ativar um conjunto de recursos dedicados para aceder à Internet, siga estes passos:

Passo 1 Interconecte uma VPC. Para mais detalhes, consulte **(Opcional) Interconexão de uma VPC com uma rede do ModelArts**.

Passo 2 Para obter detalhes sobre como configurar um servidor SNAT para uma VPC, consulte **Configuração de um servidor SNAT**.

----Fim

Exclusão de uma rede

Se uma rede não for mais necessária para o desenvolvimento de serviços de IA, você poderá excluir a rede.

1. Vá para a página de guia **Network** e clique em **Delete** na coluna **Operation** de uma rede.
2. Confirme as informações e clique em **OK**.

2.13 Nós do ModelArts

Os nós que não são gerenciados pelo pool de recursos são considerados nós livres. Para exibir as informações sobre nós livres, faça logon no console de gerenciamento do ModelArts, escolha **Dedicated Resource Pools** > **Elastic Cluster** e clique na guia **Nodes**.

Figura 2-27 Nós

Libere os recursos de nós livres de acordo com o seguinte conteúdo:

- Para um nó de pagamento por uso, clique em **Delete** na coluna **Operation**.
- Para um nó anual/mensal cujos recursos não estejam expirados, clique em **Unsubscribe** na coluna **Operation**.
- Para um nó anual/mensal cujos recursos tenham expirado (no período de tolerância), clique em **Release** na coluna **Operation**.

Se o botão de excluir estiver disponível para um nó anual/mensal, clique no botão para excluir o nó.

NOTA

As operações de exclusão, cancelamento de assinatura e liberação não podem ser desfeitas. Tenha cuidado ao realizar esta operação.

3 Logs de auditoria

3.1 Principais operações gravadas pelo CTS

Com o CTS, você pode registrar operações associadas ao ModelArts para consulta posterior, auditoria e operações de retrocesso.

Pré-requisitos

O CTS foi habilitado.

Principais operações de gerenciamento de dados rastreadas pelo CTS

Tabela 3-1 Principais operações de gerenciamento de dados rastreadas pelo CTS

Operação	Tipo de recurso	Rastreamento
Criar um conjunto de dados	Dataset	createDataset
Excluir um conjunto de dados	Dataset	deleteDataset
Atualizar um conjunto de dados	Dataset	updateDataset
Publicar uma versão do conjunto de dados	Dataset	publishDatasetVersion
Excluir uma versão do conjunto de dados	Dataset	deleteDatasetVersion
Sincronizar a fonte de dados	Dataset	syncDataSource
Exportar um conjunto de dados	Dataset	exportDataFromDataset
Criar uma tarefa de rotulagem automática	Dataset	createAutoLabelingTask

Operação	Tipo de recurso	Rastreamento
Criar uma tarefa de agrupamento automático	Dataset	createAutoGroupingTask
Criar uma tarefa de implementação automática	Dataset	createAutoDeployTask
Importar amostras para um conjunto de dados	Dataset	importSamplesToDataset
Criar um rótulo de conjunto de dados	Dataset	createLabel
Atualizar um rótulo de conjunto de dados	Dataset	updateLabel
Excluir um rótulo de conjunto de dados	Dataset	deleteLabel
Excluir um rótulo de conjunto de dados e suas amostras rotuladas	Dataset	deleteLabelWithSamples
Adicionar amostras	Dataset	uploadSamples
Excluir amostras	Dataset	deleteSamples
Interromper uma tarefa de rotulagem automática	Dataset	stopTask
Criar uma tarefa de rotulagem da equipe	Dataset	createWorkforceTask
Excluir uma tarefa de rotulagem da equipe	Dataset	deleteWorkforceTask
Iniciar a aceitação uma tarefa de rotulagem da equipe	Dataset	startWorkforceSampling-Task
Aprovar, rejeitar ou cancelar a aceitação de uma tarefa de rotulagem de equipe	Dataset	updateWorkforceSampling-Task
Enviar comentários de revisão de amostra para uma tarefa de aceitação	Dataset	acceptSamples
Adicionar um rótulo a uma amostra	Dataset	updateSamples
Enviar um e-mail para os membros de rotulagem da equipe	Dataset	sendEmails

Operação	Tipo de recurso	Rastreamento
Iniciar uma tarefa de rotulagem de equipe como gerente de equipe	Dataset	startWorkforceTask
Atualizar uma tarefa de rotulagem de equipe	Dataset	updateWorkforceTask
Adicionar um rótulo a uma amostra com rótulo de equipe	Dataset	updateWorkforceTaskSamples
Revisar resultados de rotulagem da equipe	Dataset	reviewSamples
Criar um membro da equipe de rotulagem	Workforce	createWorker
Atualizar membros da equipe de rotulagem	Workforce	updateWorker
Excluir um membro da equipe de rotulagem	Workforce	deleteWorker
Excluir membros da equipe de rotulagem em um lote	Workforce	batchDeleteWorker
Criar uma equipe de rotulagem	Workforce	createWorkforce
Atualizar uma equipe de rotulagem	Workforce	updateWorkforce
Excluir uma equipe de rotulagem	Workforce	deleteWorkforce
Criar automaticamente de uma agência do IAM	IAM	createAgency
Efetuar logon no console de rotulagem como um membro de rotulagem da equipe	labelConsoleWorker	workerLoginLabelConsole
Efetuar logoff do console de rotulagem como um membro de rotulagem da equipe	labelConsoleWorker	workerLogOutLabelConsole
Alterar a senha para efetuar logon no console de rotulagem como um membro de rotulagem da equipe	labelConsoleWorker	workerChangePassword

Operação	Tipo de recurso	Rastreamento
Tratar o problema de que a senha para fazer logon no console de rotulagem como um membro de rotulagem da equipe é perdida	labelConsoleWorker	workerForgetPassword
Redefinir a senha para fazer logon no console de rotulagem por meio do URL como um membro de rotulagem da equipe	labelConsoleWorker	workerResetPassword

Principais operações do DevEnviron rastreadas pelo CTS

Tabela 3-2 Principais operações do DevEnviron rastreadas pelo CTS

Operação	Tipo de recurso	Rastreamento
Criar uma instância de notebook	Notebook	createNotebook
Excluir uma instância de notebook	Notebook	deleteNotebook
Abrir uma instância de notebook	Notebook	openNotebook
Iniciar uma instância de notebook	Notebook	startNotebook
Interromper uma instância de notebook	Notebook	stopNotebook
Atualizar uma instância de notebook	Notebook	updateNotebook
Excluir uma NotebookApp	NotebookApp	deleteNotebookApp
Alternar especificações do CodeLab	NotebookApp	updateNotebookApp

Principais operações de trabalho de treinamento rastreadas pelo CTS

Tabela 3-3 Principais operações de trabalho de treinamento rastreadas pelo CTS

Operação	Tipo de recurso	Rastreamento
Criar um trabalho de treinamento	ModelArtsTrainJob	createModelArtsTrainJob

Operação	Tipo de recurso	Rastreamento
Criar uma versão de trabalho de treinamento	ModelArtsTrainJob	createModelArtsTrainVersion
Interromper um trabalho de treinamento	ModelArtsTrainJob	stopModelArtsTrainVersion
Modificar a descrição de um trabalho de treinamento	ModelArtsTrainJob	updateModelArtsTrainDesc
Excluir uma versão de trabalho de treinamento	ModelArtsTrainJob	deleteModelArtsTrainVersion
Excluir um trabalho de treinamento	ModelArtsTrainJob	deleteModelArtsTrainJob
Configurar um trabalho de treinamento	ModelArtsTrainConfig	createModelArtsTrainConfig
Modificar a configuração de um trabalho de treinamento	ModelArtsTrainConfig	updateModelArtsTrainConfig
Excluir uma configuração de trabalho de treinamento	ModelArtsTrainConfig	deleteModelArtsTrainConfig
Criar um trabalho de visualização	ModelArtsTensorboardJob	createModelArtsTensorboardJob
Excluir um trabalho de visualização	ModelArtsTensorboardJob	deleteModelArtsTensorboardJob
Modificar a descrição de um trabalho de visualização	ModelArtsTensorboardJob	updateModelArtsTensorboardDesc
Interromper um trabalho de visualização	ModelArtsTensorboardJob	stopModelArtsTensorboardJob
Reiniciar um trabalho de visualização	ModelArtsTensorboardJob	restartModelArtsgTensorboardJob

Principais operações de gerenciamento de aplicações de IA rastreadas pelo CTS

Tabela 3-4 Principais operações de gerenciamento de aplicações de IA rastreadas pelo CTS

Operação	Tipo de recurso	Rastreamento
Criar uma aplicação de IA	Model	addModel
Atualizar uma aplicação de IA	Model	updateModel
Excluir uma aplicação de IA	Model	deleteModel

Operação	Tipo de recurso	Rastreamento
Criar uma tarefa de conversão de modelo	Convert	addConvert
Atualizar uma tarefa de conversão de modelo	Convert	updateConvert
Excluir uma tarefa de conversão de modelo	Convert	deleteConvert

Principais operações de gerenciamento de serviços rastreadas pelo CTS

Tabela 3-5 Principais operações de gerenciamento de serviços rastreadas pelo CTS

Operação	Tipo de recurso	Rastreamento
Implementar um serviço	Service	addService
Excluir um serviço	Service	deleteService
Atualizar um serviço	Service	updateService
Iniciar ou parar um serviço	Service	startOrStopService
Adicionar uma chave de acesso de usuário	Service	addAkSk
Excluir uma chave de acesso de usuário	Service	deleteAkSk
Criar um pool de recursos dedicados	Cluster	createCluster
Excluir um pool de recursos dedicados	Cluster	deleteCluster
Adicionar um nó a um pool de recursos dedicados	Cluster	addClusterNode
Excluir um nó de um pool de recursos dedicados	Cluster	deleteClusterNode
Obter um resultado da criação de um pool de recursos dedicados	Cluster	createClusterResult

Principais operações da Galeria de IA rastreadas pelo CTS

Tabela 3-6 Principais operações da Galeria de IA rastreadas pelo CTS

Operação	Tipo de recurso	Rastreamento
Publicar um ativo	ModelArts_Market	create_content
Modificar informações de ativo	ModelArts_Market	modify_content
Publicar uma versão de ativo	ModelArts_Market	add_version
Inscrever a um ativo	ModelArts_Market	subscription_content
Remover um ativo dos favoritos	ModelArts_Market	cancel_star_content
Gostar de um ativo	ModelArts_Market	like_content
Desgostar de um ativo	ModelArts_Market	cancel_like_content
Publicar uma atividade	ModelArts_Market	publish_activity
Inscrever-se em uma atividade	ModelArts_Market	regist_activity
Modificar informações de usuário	ModelArts_Market	update_user

Principais operações de gerenciamento de recursos rastreadas pelo CTS

Tabela 3-7 Principais operações de gerenciamento de recursos rastreadas pelo CTS

Operação	Tipo de recurso	Rastreamento
Criar um pool de recursos	PoolV2	CreatePoolV2
Excluir um pool de recursos	PoolV2	DeletePoolV2
Atualizar um pool de recursos	PoolV2	UpdatePoolV2
Criar uma rede	NetworksV1	CreateNetworksV1
Excluir uma rede	NetworksV1	DeleteNetworksV1
Atualizar uma rede	NetworksV1	UpdateNetworksV1

3.2 Visualização de logs de auditoria

Depois que o CTS é habilitado, o CTS começa a gravar operações relacionadas ao ModelArts. O console de gerenciamento CTS armazena os últimos sete dias de registros de operação. Esta

seção descreve como consultar os registros de operação dos últimos 7 dias no console do gerenciamento do CTS.

Procedimento

1. Faça logon no console de gerenciamento do CTS.
2. Clique em  no canto superior esquerdo da página e selecione uma região.
3. No painel de navegação esquerdo, clique em **Trace List**.
4. Especifique os critérios de filtro usados para consultar rastreamentos. Os quatro critérios de filtro a seguir estão disponíveis:
   - **Trace Source**, **Resource Type** e **Search By**

     Selecione um critério de filtro na lista suspensa.

     Se você selecionar **Trace name** para **Search By**, será necessário selecionar um nome de rastreamento específico.

     Se você selecionar **Resource ID** para **Search By**, precisa inserir um ID de recurso específico.

     Se você selecionar **Resource name** para **Search By**, precisa selecionar ou inserir um nome de recurso específico.
   - **Operator**: selecione um operador específico (um usuário em vez de uma conta).
   - **Trace Status**: as opções disponíveis incluem **All trace statuses**, **Normal**, **Warning** e **Incident**. Você só pode selecionar um deles.
   - **Time Range**: você pode exibir rastreamentos gerados durante qualquer intervalo de tempo dos últimos sete dias.
5. Clique em ⌄ à esquerda de um traço para expandir seus detalhes.
6. Clique em **View Trace** na coluna **Operation**. Na caixa de diálogo **View Trace** exibida, os detalhes da estrutura de rastreamento são exibidos.

   Para obter detalhes sobre os campos-chave na estrutura de rastreamento CTS, consulte **Guia de usuário do Cloud Trace Service**.

4 Monitoramento de recursos

4.1 Visão geral

Todas as métricas relatadas pelo ModelArts são armazenadas no AOM, o que permite consumir métricas. Você pode visualizar alarmes de limite de métricas e alarmes relatados no console do AOM ou usar ferramentas de visualização como o Grafana para visualizar e analisar os alarmes. O Grafana fornece diferentes visualizações e modelos para monitoramento, que permitem que você veja o uso de recursos em tempo real nos painéis de controle com clareza.

4.2 Uso do Grafana para exibir as métricas de monitoramento do AOM

4.2.1 Procedimento

O Grafana suporta várias visualizações e modelos de monitoramento, atendendo aos seus diversos requisitos. Depois de adicionar a fonte de dados no Grafana, você pode visualizar todas as métricas de monitoramento do ModelArts armazenadas no AOM usando o Grafana.

Para visualizar as métricas de monitoramento do AOM usando plug-ins do Grafana, execute as seguintes etapas:

1. **Instalação e configuração do Grafana**

   NOTA

   Você pode instalar e configurar o Grafana usando qualquer uma das seguintes maneiras: **Instalação e configuração do Grafana no Windows**, **Instalação e configuração do Grafana no Linux** e **Instalação e configuração do Grafana em uma instância de notebook**.

2. **Configuração de uma fonte de dados do Grafana**
3. **Uso do Grafana para configurar painéis e visualizar dados métricos**

4.2.2 Instalação e configuração do Grafana

4.2.2.1 Instalação e configuração do Grafana no Windows

Cenário de aplicação

Esta seção descreve como instalar e configurar o Grafana em um sistema operacional Windows.

Procedimento

1. Baixe o pacote de instalação do Grafana.

   Vá para o **link de download**, clique em **Download the installer** e aguarde até que o download seja bem-sucedido.

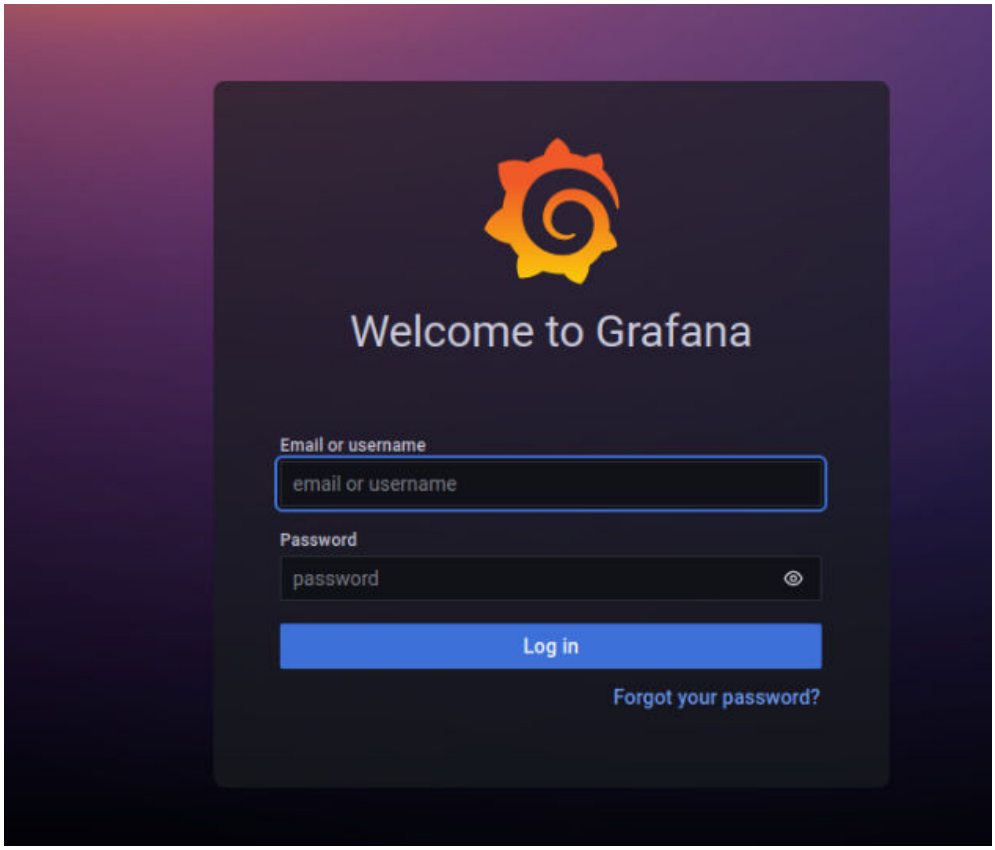
2. Instale o Grafana.

   Clique duas vezes no pacote de instalação e instale o Grafana como instruído.

3. No Windows Services Manager, ative o Grafana.

4. Faça logon no Grafana.

   O Grafana é executado na porta 3000 por padrão. Depois de abrir http://localhost:3000, é exibida a página de logon do Grafana. O nome de usuário e a senha padrão para o primeiro logon são **admin**. Depois que o logon for bem-sucedido, altere a senha conforme solicitado.



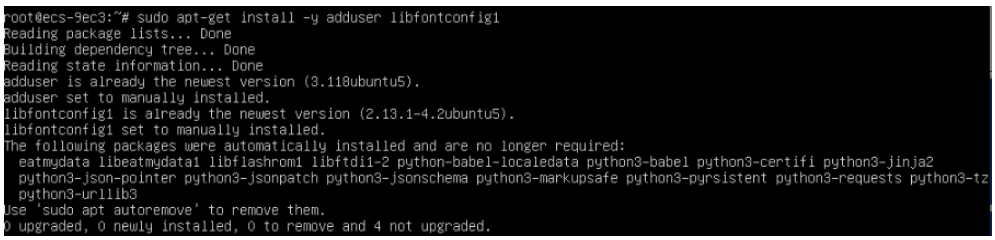
4.2.2.2 Instalação e configuração do Grafana no Linux

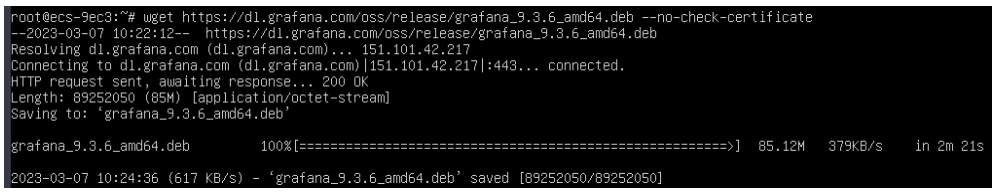
Pré-requisitos

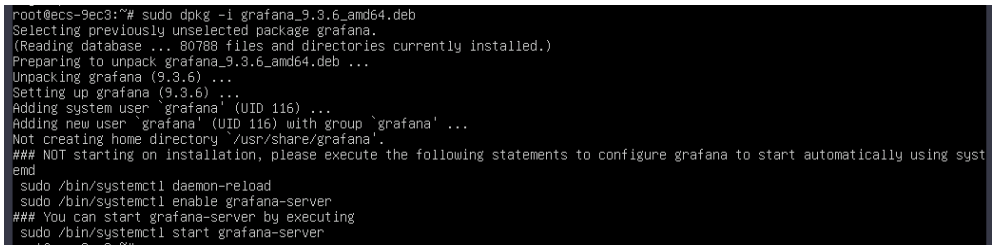
- Um servidor Ubuntu que é acessível à Internet está disponível. Se não, devem ser satisfeitas as seguintes condições:
- Você obteve um ECS. (É aconselhável selecionar 8 vCPUs ou superior, imagem do Ubuntu da versão 22.04 e 100 GB de armazenamento local.) Para obter detalhes, consulte **Compra de um ECS**.
- Você comprou um EIP e vinculou-o ao ECS. Para obter detalhes, consulte **Atribuição de um EIP e sua vinculação a um ECS**.

Procedimento

1. Faça logon no ECS. Selecione um método de logon. Para obter detalhes, consulte .
2. Execute o seguinte comando para instalar libfontconfig1:
   ```
   sudo apt-get install -y adduser libfontconfig1
   ```
   A operação será bem-sucedida se as seguintes informações forem exibidas:

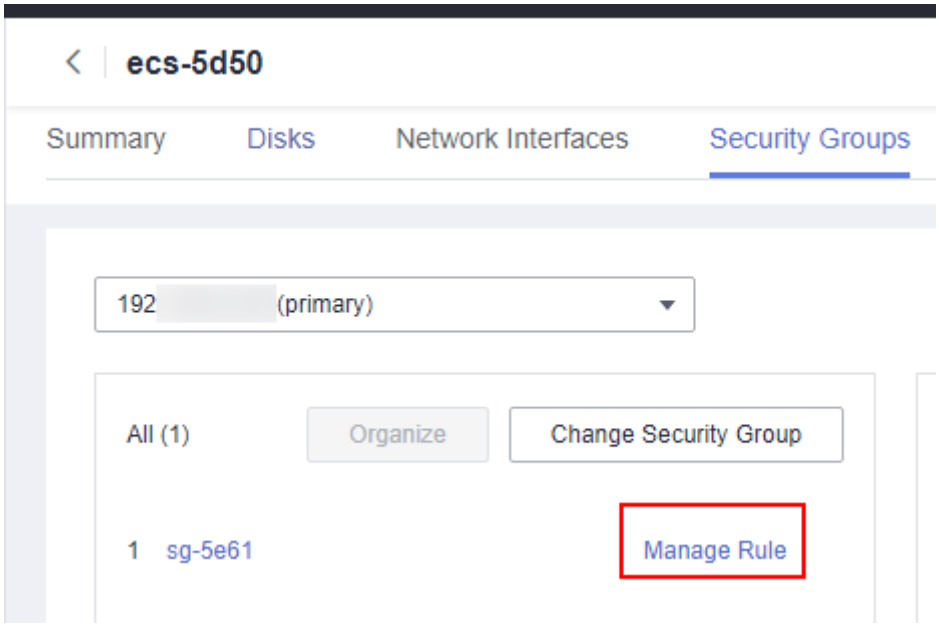
   
   ```
   root@ecs-9ec3:~# sudo apt-get install -y adduser libfontconfig1
   Reading package lists... Done
   Building dependency tree... Done
   Reading state information... Done
   adduser is already the newest version (3.118ubuntu5).
   adduser set to manually installed.
   libfontconfig1 is already the newest version (2.13.1-4.2ubuntu5).
   libfontconfig1 set to manually installed.
   The following packages were automatically installed and are no longer required:
     eatmydata libeatmydata1 libflashrom1 libftdi1-2 python-babel-localedata python3-babel python3-certifi python3-jinja2
     python3-json-pointer python3-jsonpatch python3-jsonschema python3-markupsafe python3-pyrsistent python3-requests python3-tz
     python3-urllib3
   Use 'sudo apt autoremove' to remove them.
   0 upgraded, 0 newly installed, 0 to remove and 4 not upgraded.
   ```
3. Execute o seguinte comando para baixar o pacote de instalação de Grafana:
   ```
   wget https://dl.grafana.com/oss/release/grafana_9.3.6_amd64.deb --no-check-certificate
   ```
   Download concluído:

   
   ```
   root@ecs-9ec3:~# wget https://dl.grafana.com/oss/release/grafana_9.3.6_amd64.deb --no-check-certificate
   --2023-03-07 10:22:12--  https://dl.grafana.com/oss/release/grafana_9.3.6_amd64.deb
   Resolving dl.grafana.com (dl.grafana.com)... 151.101.42.217
   Connecting to dl.grafana.com (dl.grafana.com)|151.101.42.217|:443... connected.
   HTTP request sent, awaiting response... 200 OK
   Length: 89252050 (85M) [application/octet-stream]
   Saving to: 'grafana_9.3.6_amd64.deb'

   grafana_9.3.6_amd64.deb        100%[=====================================================>]  85.12M   379KB/s    in 2m 21s

   2023-03-07 10:24:36 (617 KB/s) - 'grafana_9.3.6_amd64.deb' saved [89252050/89252050]
   ```
4. Execute o seguinte comando para instalar o Grafana:
   ```
   sudo dpkg -i grafana_9.3.6_amd64.deb
   ```

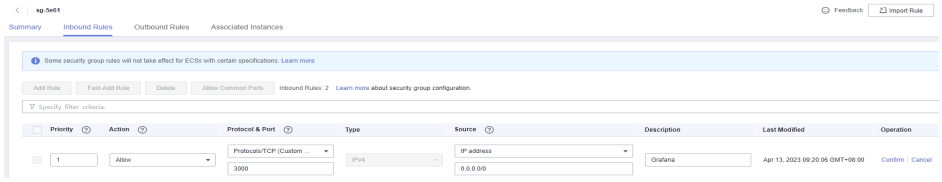
   
   ```
   root@ecs-9ec3:~# sudo dpkg -i grafana_9.3.6_amd64.deb
   Selecting previously unselected package grafana.
   (Reading database ... 80788 files and directories currently installed.)
   Preparing to unpack grafana_9.3.6_amd64.deb ...
   Unpacking grafana (9.3.6) ...
   Setting up grafana (9.3.6) ...
   Adding system user `grafana' (UID 116) ...
   Adding new user `grafana' (UID 116) with group `grafana' ...
   Not creating home directory `/usr/share/grafana'.
   ### NOT starting on installation, please execute the following statements to configure grafana to start automatically using syst
   emd
    sudo /bin/systemctl daemon-reload
    sudo /bin/systemctl enable grafana-server
   ### You can start grafana-server by executing
    sudo /bin/systemctl start grafana-server
   ```
5. Execute o seguinte comando para iniciar o Grafana:
   ```
   sudo /bin/systemctl start grafana-server
   ```
6. Acesse as configurações do Grafana em seu PC local.

   Verifique se um EIP foi vinculado ao ECS e se a configuração do **grupo de segurança** está correta (o tráfego de entrada da porta TCP 3000 e todo o tráfego de saída são permitidos). Processo de configuração:

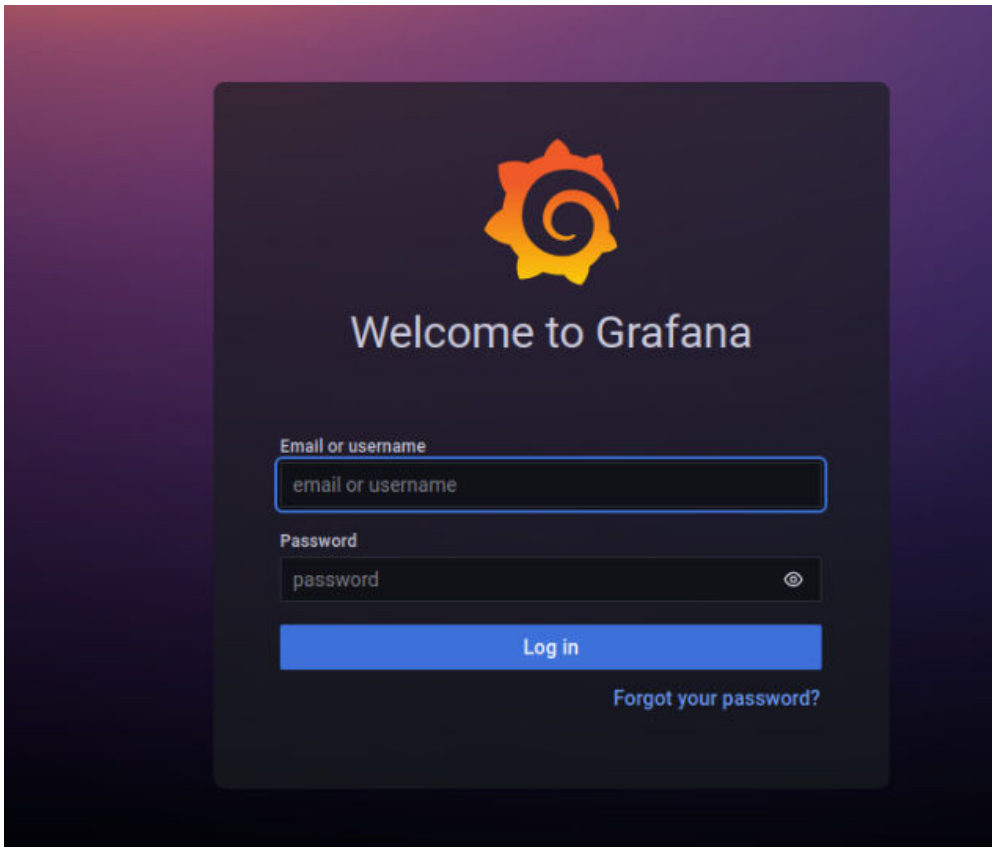
a. Clique no nome do ECS para ir para a página de detalhes do ECS. Em seguida, clique na guia **Security Groups** e clique em **Manage Rule**.



b. Clique em **Inbound Rules** e permita tráfego de entrada da porta TCP 3000. Por padrão, todo o tráfego de saída é permitido.



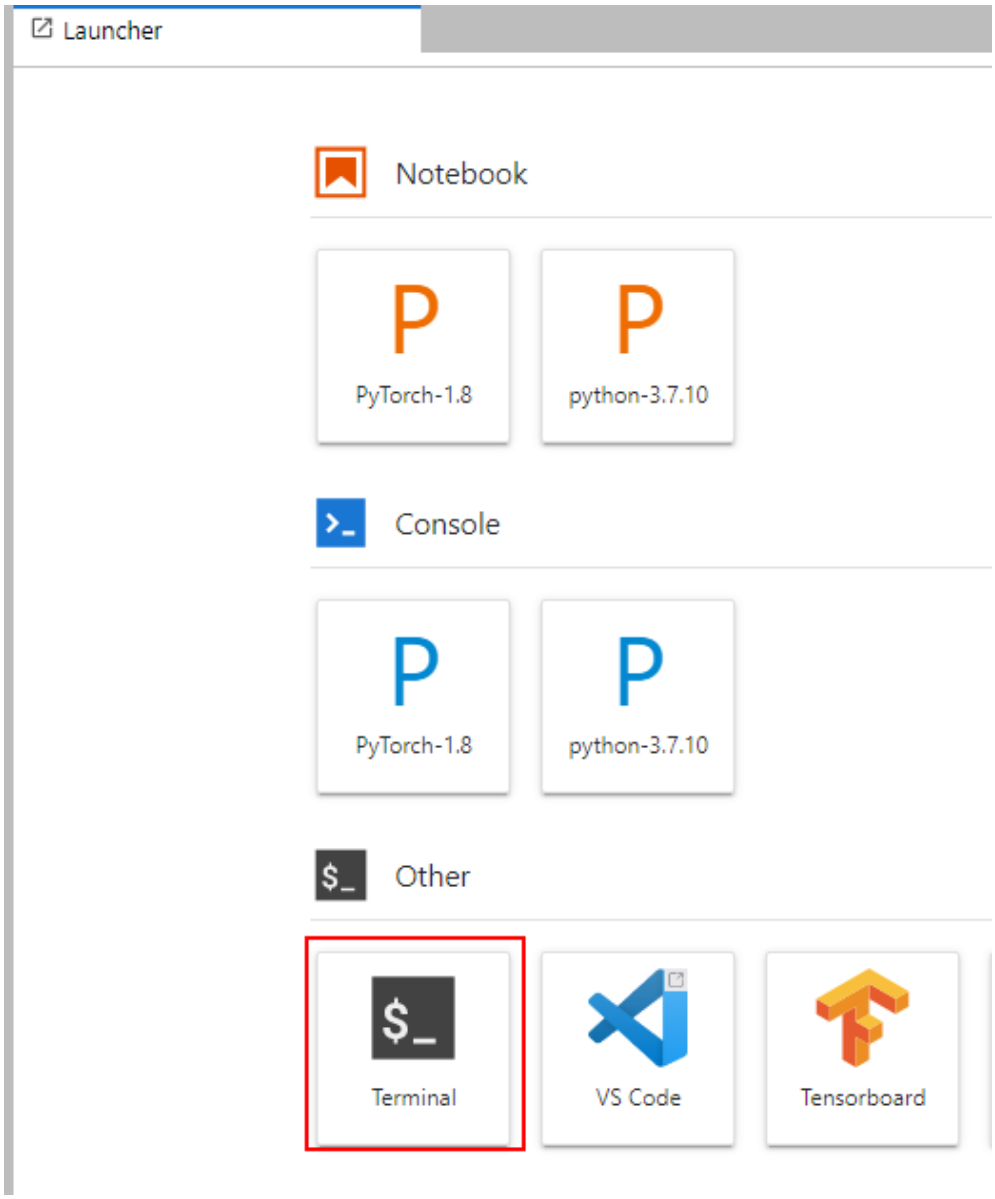
7. Acesse **http://{*EIP*}:3000** em um navegador. O nome de usuário e a senha padrão para o primeiro logon são **admin**. Depois que o logon for bem-sucedido, altere a senha conforme solicitado.



4.2.2.3 Instalação e configuração do Grafana em uma instância de notebook

Pré-requisitos

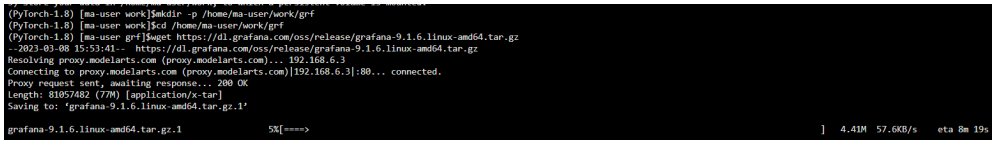
- Uma instância de notebook baseada em CPU ou GPU em execução está disponível.
- Um terminal é aberto.



Procedimento

1. Execute os seguintes comandos em sequência no seu terminal para baixar e instalar o 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 o Grafana com jupyter-server-proxy.
   a. Execute os seguintes comandos no seu 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. Em **jupyter_notebook_config.py**, adicione o seguinte código, pressione **Esc** para sair e digite **:wq** para salvar as alterações:

```
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

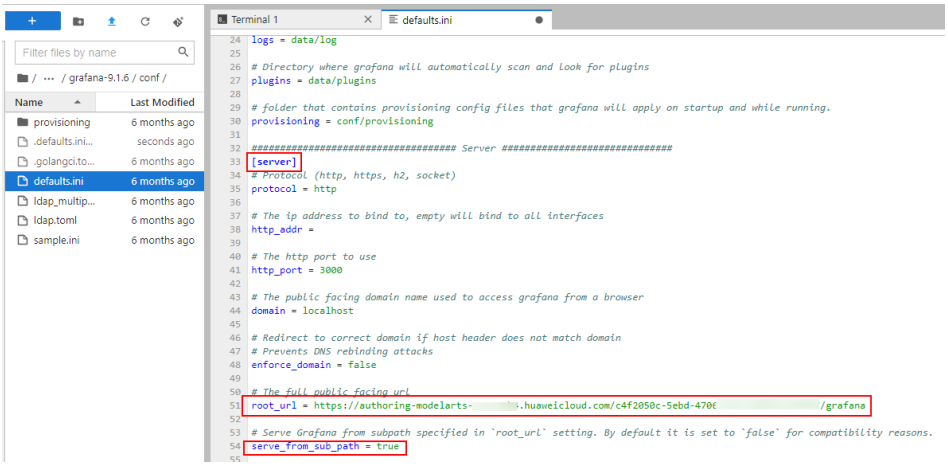
Se **jupyter_notebook_config.py** (caminho: **/home/ma-user/.local/etc/jupyter/jupyter_notebook_config.py**) contém o campo **c.ServerProxy.servers**, adicione o par chave-valor correspondente.

3. Modifique o URL para acessar o Grafana no JupyterLab.

a. No painel de navegação à esquerda, abra o arquivo **vi /home/ma-user/work/grf/grafana-9.1.6/conf/defaults.ini**.

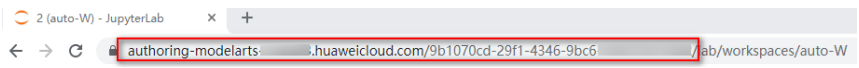
b. Altere os campos **root_url** e **serve_from_sub_path** em **[server]**.

Figura 4-1 Modificar o arquivo **defaults.ini**



No arquivo:

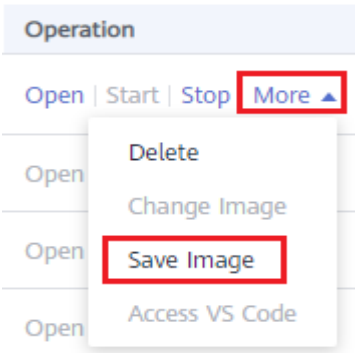
- O valor de **root_url** está no formato de **https:**{*Jupyterlab domain name*}/{*Instance ID*}**/grafana**. Você pode obter o nome de domínio e o ID da instância na caixa de endereço da página de JupyterLab.



- Defina **Serve_from_sub_path** como **true**.

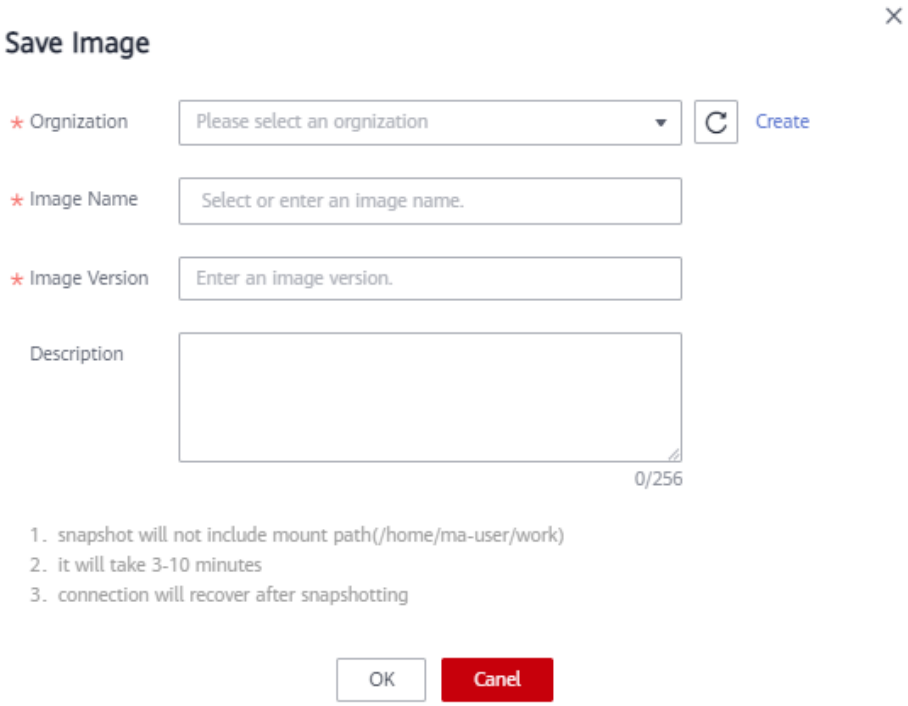
4. Salve a imagem da instância do notebook.

a. Faça logon no console do ModelArts e escolha **DevEnviron** > **Notebook**. Na lista de instâncias do notebook, escolha **More** > **Save Image** na coluna **Operation** da instância de destino.



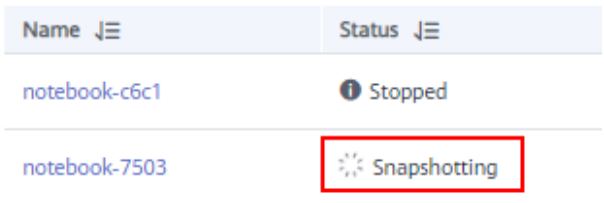
b. Na caixa de diálogo **Save Image**, configure os parâmetros. Clique em **OK** para salvar a imagem.

Figura 4-2 Salvar uma imagem



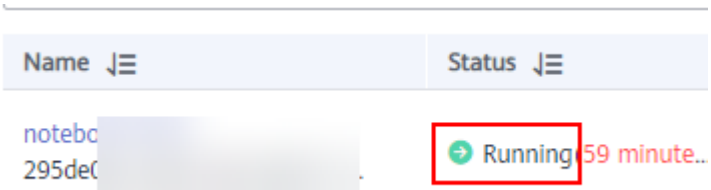
c. A imagem será salva como um snapshot e levará cerca de 5 minutos. Durante esse período de tempo, não execute nenhuma operação na instância.

Figura 4-3 Captação de snapshot



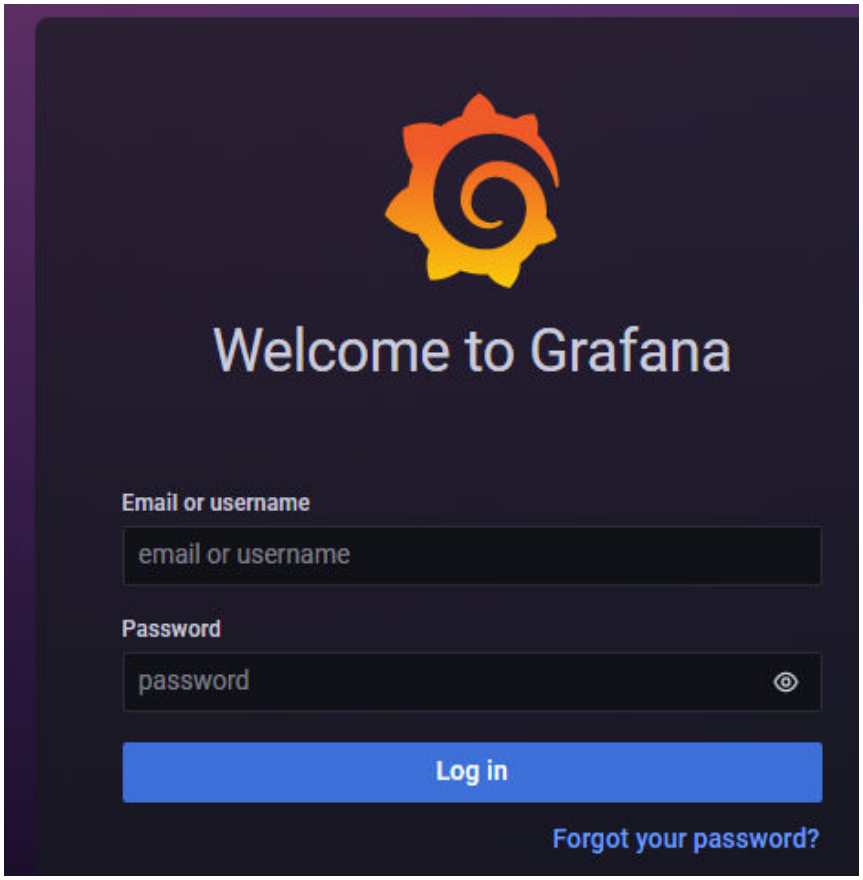
d. Depois que a imagem é salva, o status da instância muda para **Running**. Em seguida, reinicie a instância do notebook.

Figura 4-4 Imagem salva



5. Abra a página Grafana.

   Abra uma janela do navegador e digite o valor de **root_url** configurado em **3** na caixa de endereço. Se a página de logon do Grafana for exibida, o Grafana será instalado e configurado na instância do notebook. O nome de usuário e a senha padrão para o primeiro logon são **admin**. Depois que o logon for bem-sucedido, altere a senha conforme solicitado.



4.2.3 Configuração de uma fonte de dados do Grafana

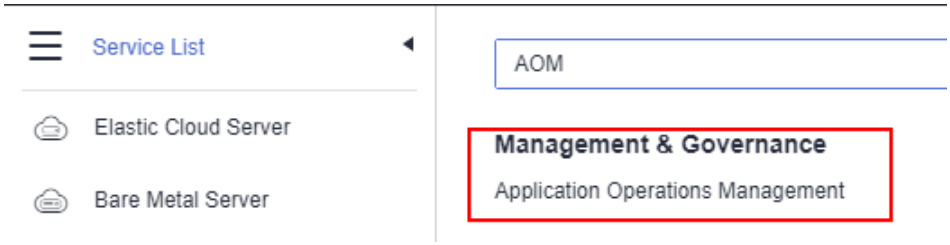
Antes de visualizar os dados de monitoramento do ModelArts no Grafana, configure a fonte de dados.

Pré-requisitos

- Grafana foi instalado.

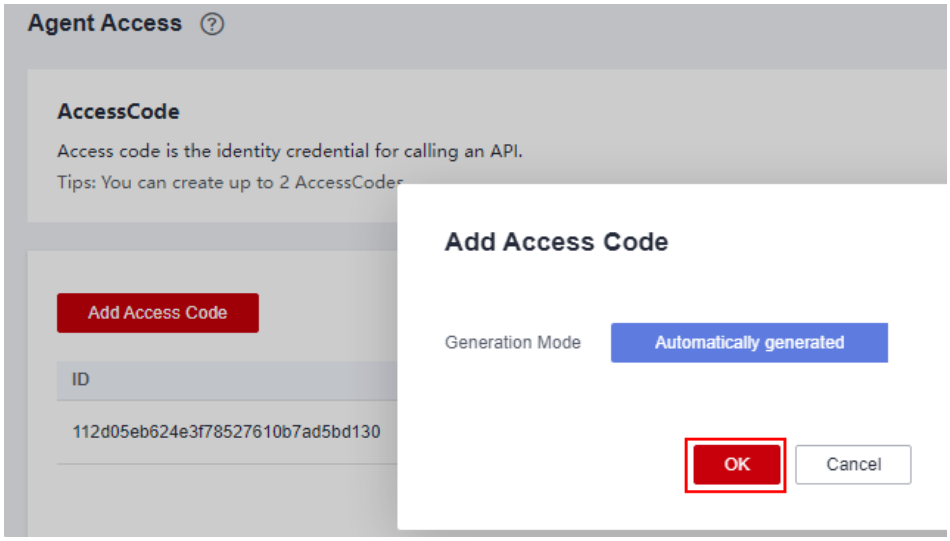
Procedimento

1. Adicione um código de acesso.
   a. Efetue logon no console do AOM.



   b. No painel de navegação à esquerda, escolha **Configuration Management** > **Agent Access** e clique em **Add Access Code** para gerar um código de acesso.

Figura 4-5 Gerar um código de acesso



   c. Clique em para exibir o código de acesso gerado.

Figura 4-6 Visualizar o código de acesso

2. Obtenha o URL da fonte de dados.

   O URL está no formato de **https://**{*Endpoint*}**/v1/**{*project_id*}.

   – Você pode obter as informações do ponto de extremidade do AOM em Regiões e pontos de extremidade.
   – Defina **project_id** como o ID do projeto da região correspondente. Você pode obter o ID do projeto de **My Credentials**.

Figura 4-7 Minhas credenciais

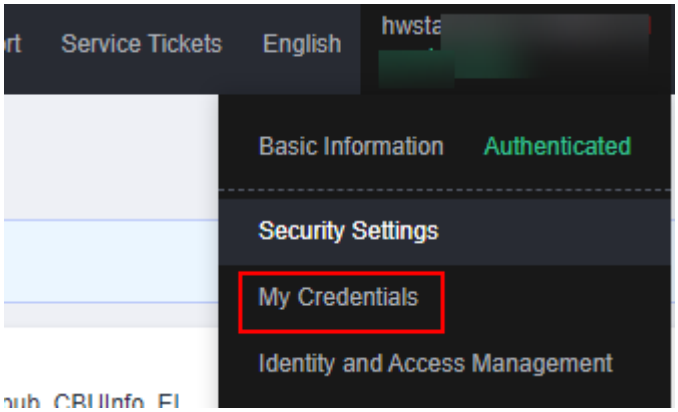
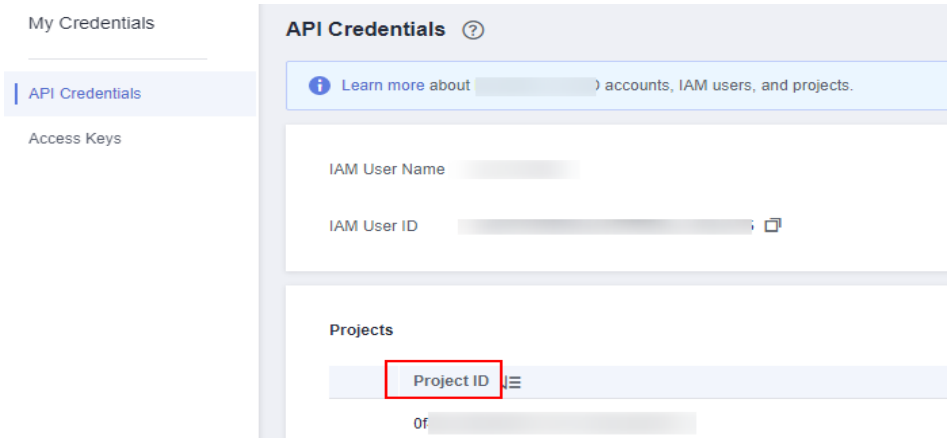
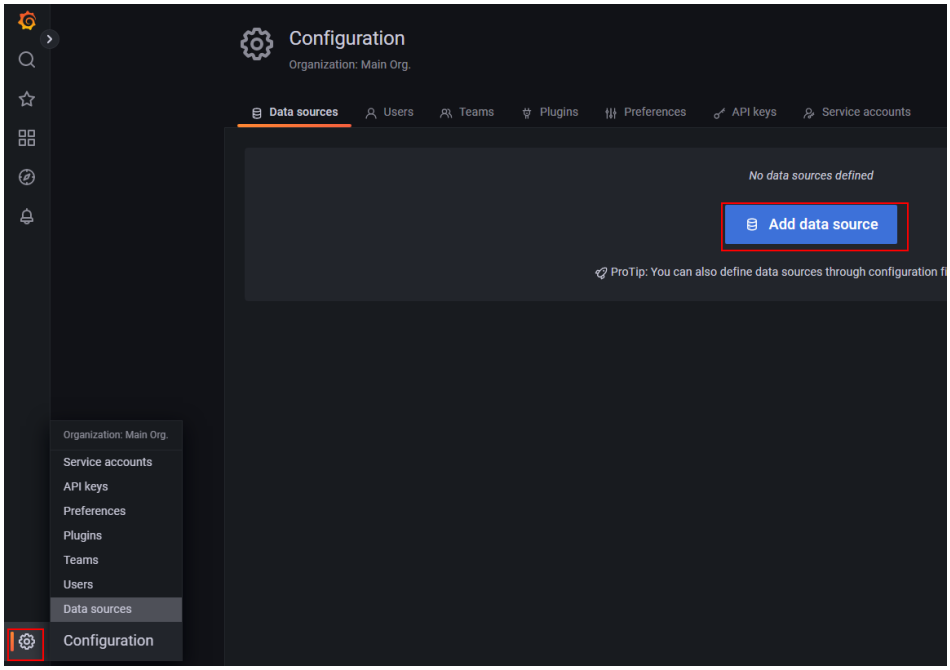


Figura 4-8 Obtenção do ID do projeto



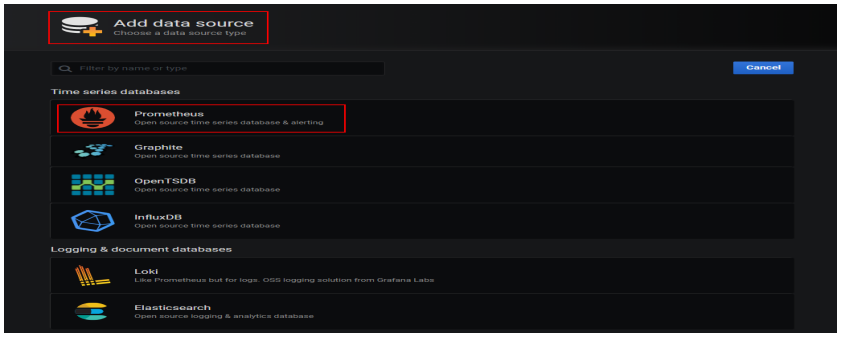
3. Adicione uma fonte de dados ao Grafana.
   a. Faça logon no Grafana. O nome de usuário e a senha padrão para o primeiro logon são **admin**. Depois que o logon for bem-sucedido, altere a senha conforme solicitado.
   b. No painel de navegação, escolha **Configuration** > **Data Sources**. Em seguida, clique em **Add data source**.

Figura 4-9 Configurar o Grafana



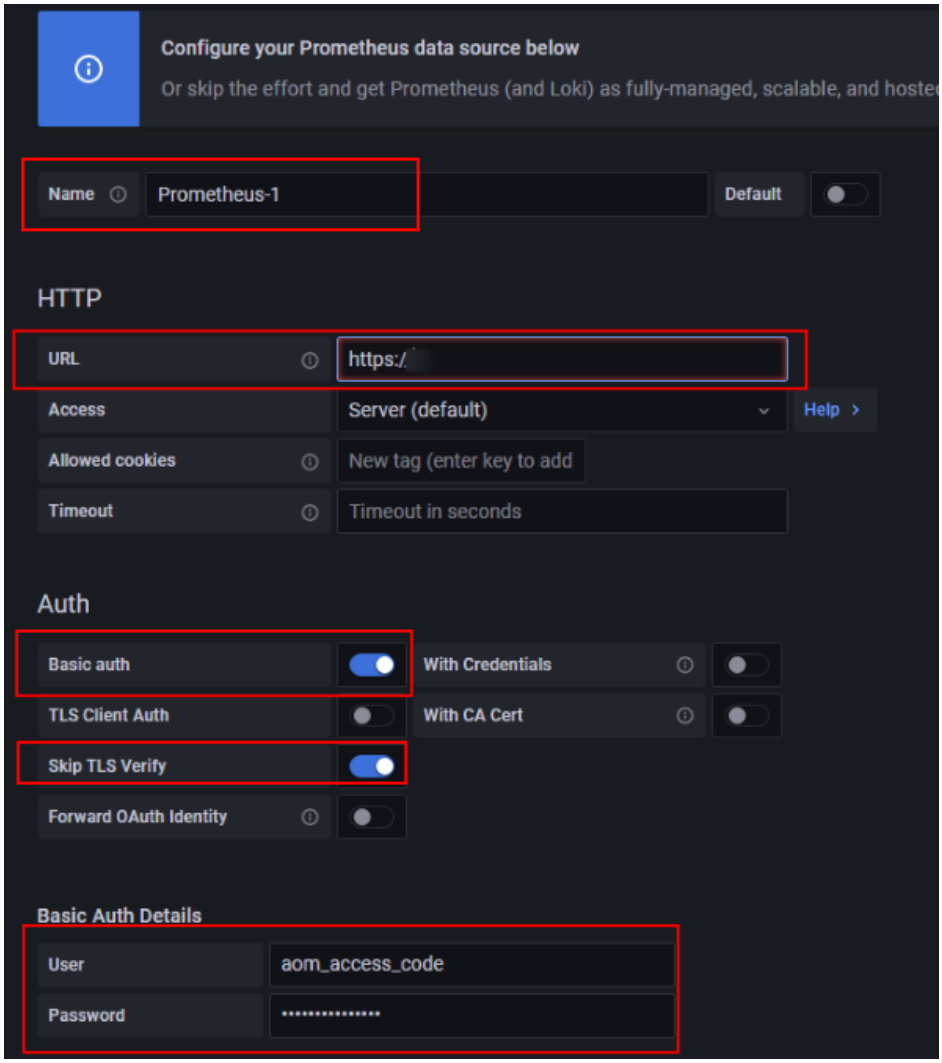
c. Clique em **Prometheus** para acessar a página de configuração.

Figura 4-10 Entrar na página de configuração do Prometheus



d. Configure os parâmetros conforme mostrado na figura a seguir.

Figura 4-11 Configurar uma fonte de dados do Grafana



NOTA

A versão atual do Grafana varia dependendo do método de instalação. **Figura 4-11** é apenas um exemplo.

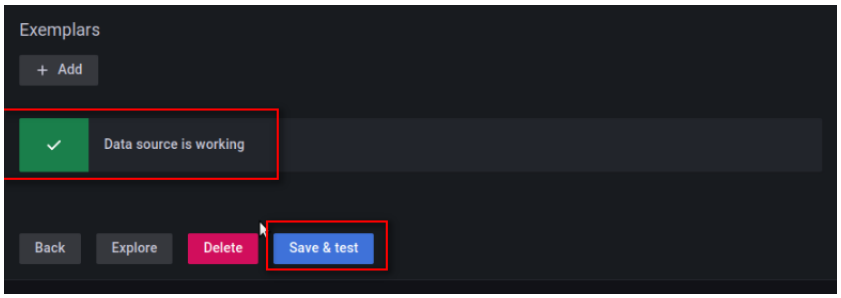
Tabela 4-1 Parâmetros

Parâmetro	Descrição
Name	Nome personalizável
URL	URL **https://**{*Endpoint*}**/v1/**{*project_id*} combinado em **Obter o URL da fonte de dados**
Basic auth	Ativado
Skip TLS Verify	Ativado
User	aom_access_code

Parâmetro	Descrição
Password	Código de acesso gerado em **Adicionar um código de acesso.**

e. Após a configuração, clique em **Save & test**. Se a mensagem **Data source is working** for exibida, a fonte de dados está configurada.

Figura 4-12 Fonte de dados adicionada



4.2.4 Uso do Grafana para configurar painéis e visualizar dados métricos

No Grafana, você pode personalizar painéis para várias visualizações. O ModelArts também fornece modelos de configuração para clusters. Esta seção descreve como configurar um painel usando um modelo do ModelArts ou criando um painel. Para mais informações, veja **tutoriais do Grafana**.

Preparativos

O ModelArts fornece modelos para exibição de cluster, exibição de nó, exibição de usuário, exibição de tarefa e exibição de detalhes da tarefa. Esses modelos podem ser baixados dos documentos oficiais do Grafana. Você pode importá-los e usá-los em **Dashboards**.

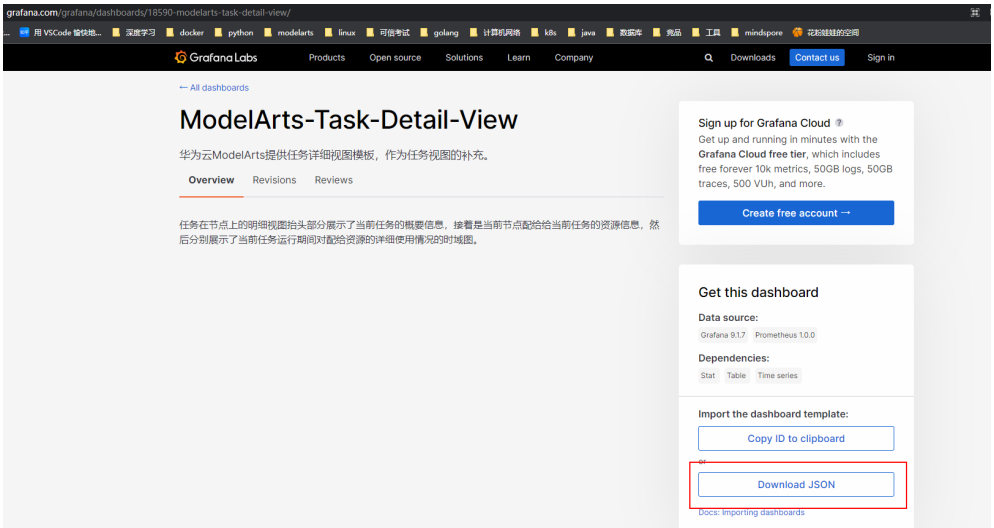
Tabela 4-2 URLs de download do modelo

Nome do modelo	URL de download
Exibição de cluster	https://grafana.com/grafana/dashboards/18582-modelarts-cluster-view/
Exibição de nó	https://grafana.com/grafana/dashboards/18583-modelarts-node-view/
Exibição de usuário	https://grafana.com/grafana/dashboards/18588-modelarts-user-view/
Exibição de tarefa	https://grafana.com/grafana/dashboards/18604-modelarts-task-view/
Exibição de detalhes da tarefa	https://grafana.com/grafana/dashboards/18590-modelarts-task-detail-view/

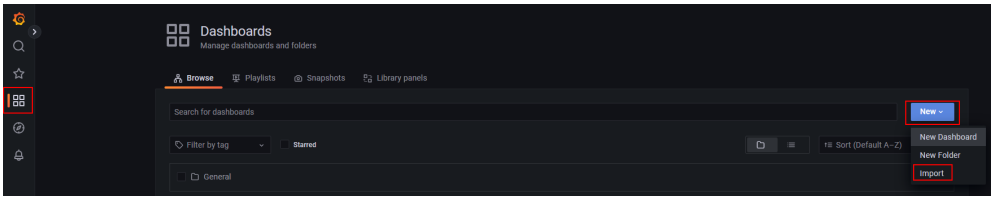
Usar um modelo do ModelArts para exibir métricas

1. (Opcional) Selecione o modelo que deseja usar. **Preparativos** exibe os endereços de download de todos os modelos. Abra o endereço de destino e clique em **Download JSON**.

   Figura 4-13 Baixar o modelo para a exibição de detalhes da tarefa

2. Abra **Dashboards** e escolha **New** > **Import**.

3. Importe um modelo de painel de controle de uma das seguintes maneiras:
   - Método 1: carregue o arquivo JSON baixado em **1**, como mostrado em **Figura 4-14**.
   - Método 2: copie o endereço de download do modelo fornecido em **Preparativos** e clique em **Load**, conforme mostrado em **Figura 4-15**.

Figura 4-14 Carregar um arquivo JSON para importar um modelo de painel

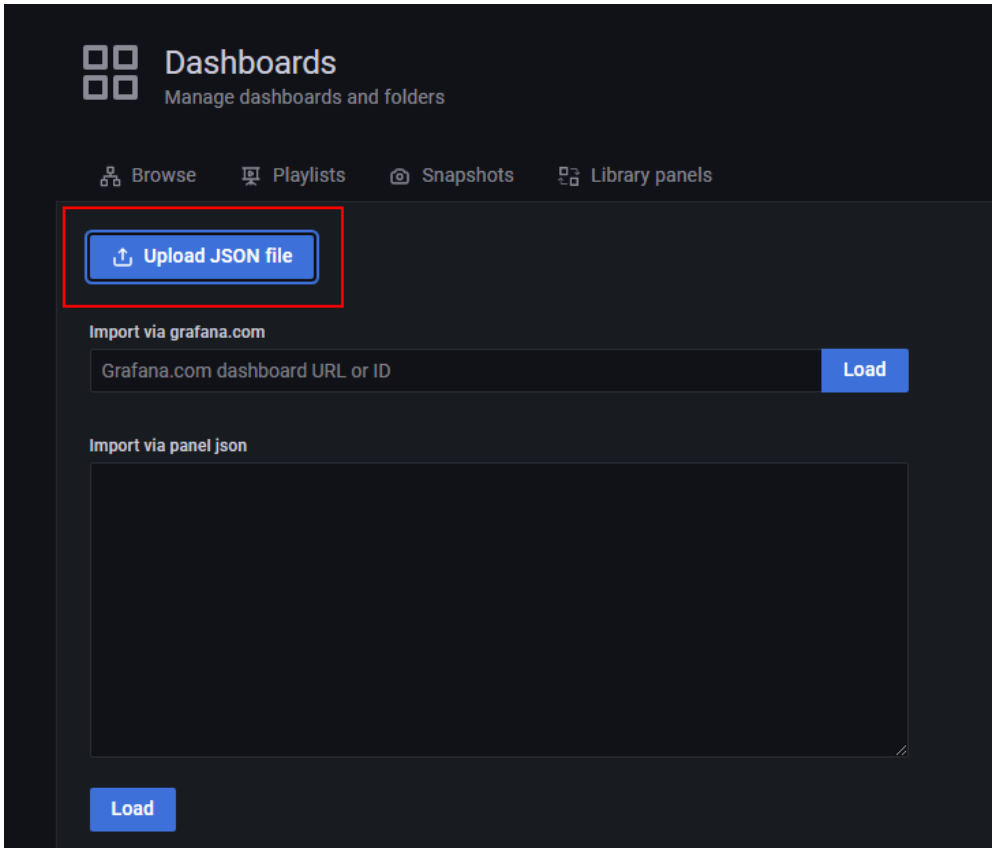
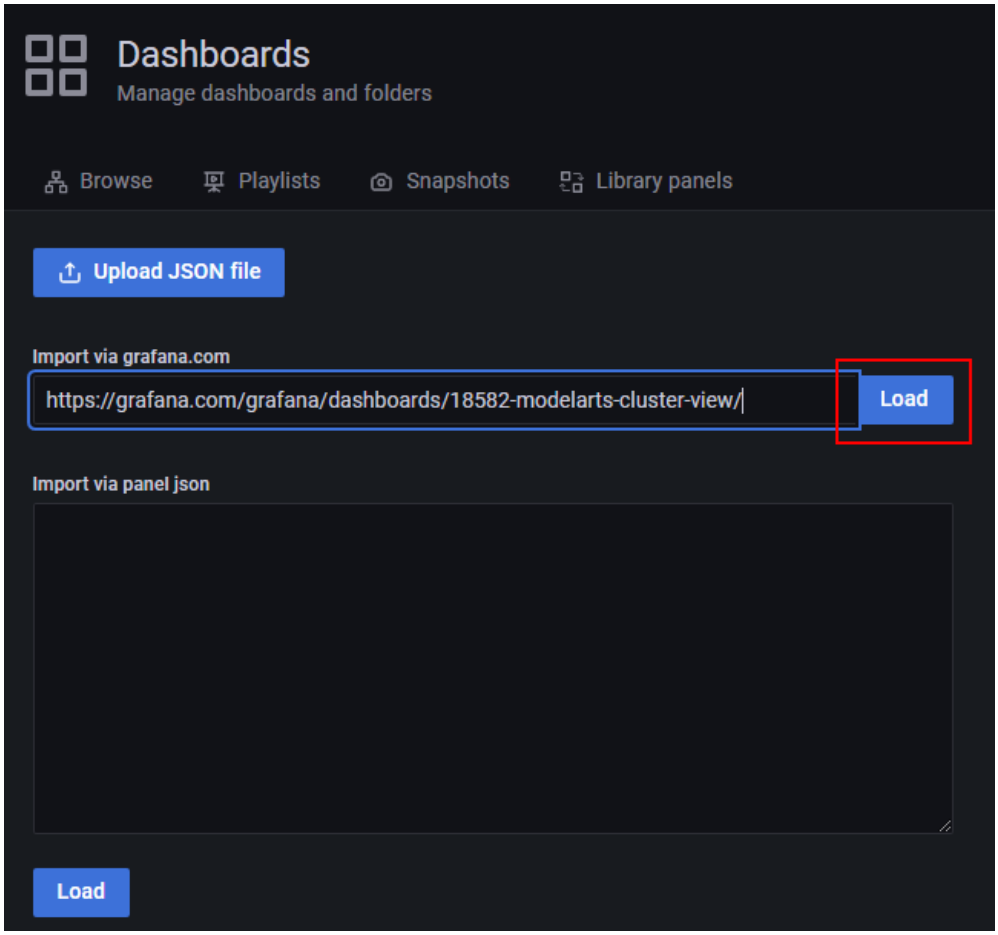
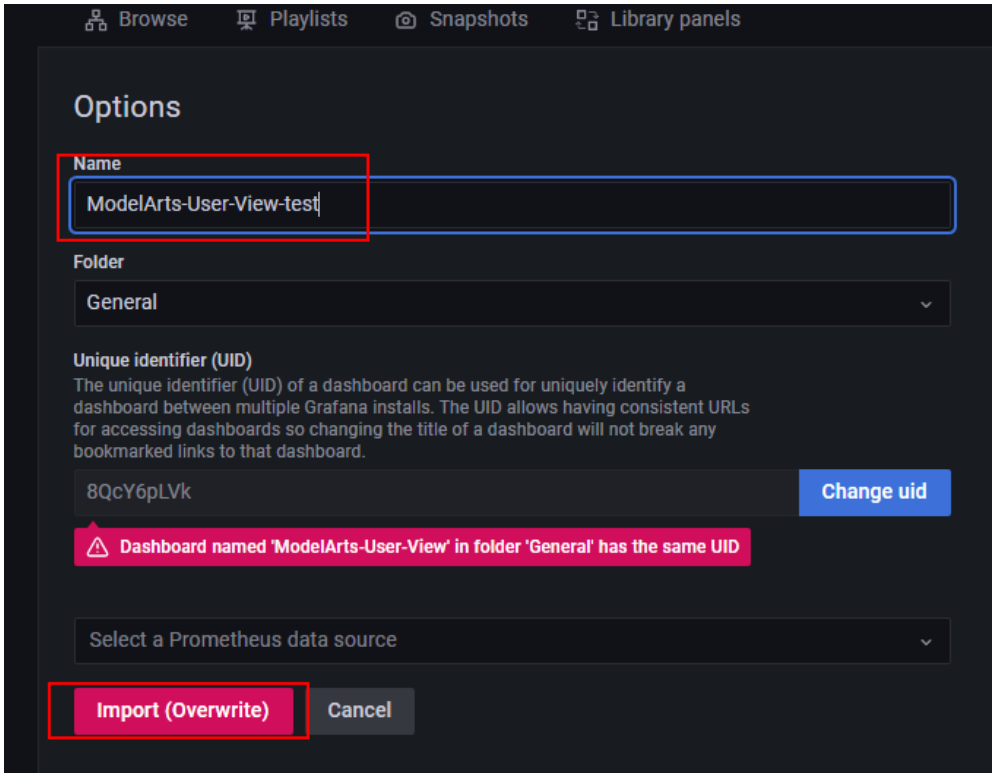


Figura 4-15 Copiar o endereço do modelo e importar o modelo do painel



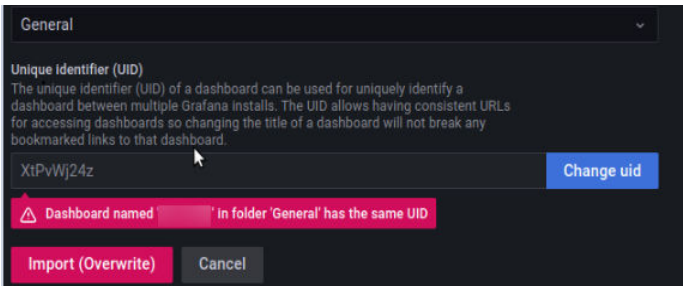
4. Altere o nome da vista e clique em **Import**.

Figura 4-16 Alterar o nome da exibição

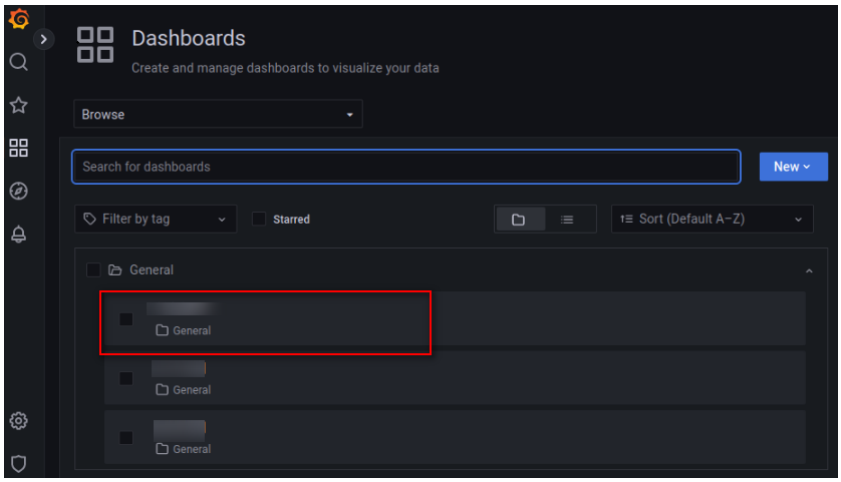


Observação: se uma mensagem for exibida indicando que o UID é duplicado, altere o UID no arquivo JSON e clique em **Import**.

Figura 4-17 Mudar o UID



5. Após a importação, visualize as exibições importadas em **Dashboards**. Em seguida, clique em uma exibição para abrir a página de monitoramento.

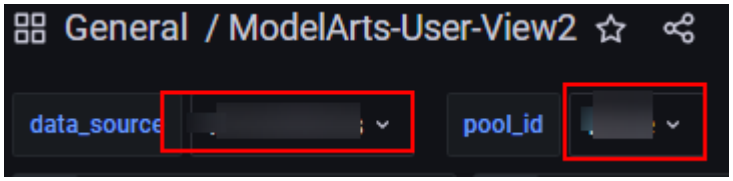


6. Use o modelo.

   Depois que a importação for bem-sucedida, você poderá clicar no modelo para exibir seus detalhes. Esta seção apresenta algumas funções comuns.

   – Alterar a origem de dados e o pool de recursos

   Figura 4-18 Alterar a origem de dados e o pool de recursos

   Clique na área marcada pela caixa vermelha. Uma lista suspensa aparecerá. A partir daí, você pode alterar a fonte de dados e o pool de recursos.

   – Atualizar dados

   Clique no botão de atualizar no canto superior direito para atualizar todos os dados no painel. Os dados de cada painel também são atualizados.

   – Alterar o tempo de atualização automática

   Figura 4-19 Alterar o tempo de atualização automática

   O intervalo de atualização padrão de um modelo é de 15 minutos. Se você precisar atualizar o intervalo, altere o valor na caixa de listagem suspensa no canto superior direito.

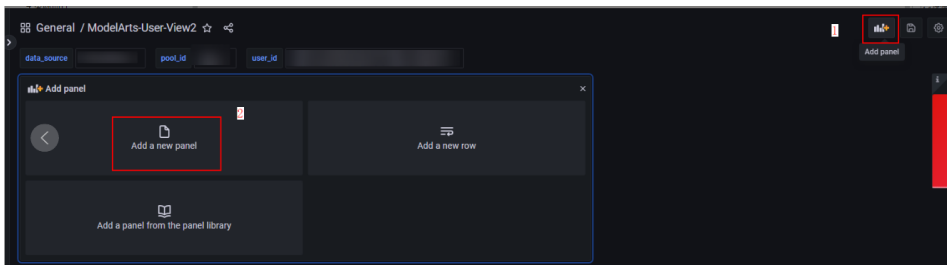
   – Alterar o intervalo de tempo para a obtenção de dados do painel

Figura 4-20 Alterar o intervalo de tempo para obtenção de dados

Clique no botão no canto superior direito para alterar o intervalo de tempo para obter dados. Esse intervalo de tempo afeta todos os painéis, exceto aqueles com um tempo fixo.

– Adicionar um painel

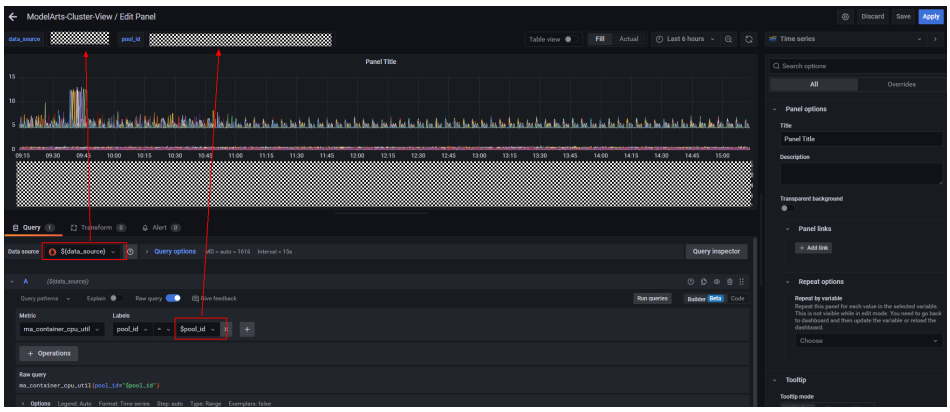
Figura 4-21 Adicionar um painel



Clique no ícone do botão no canto superior direito para adicionar um painel.

Depois que um painel é adicionado, você pode obter os dados no painel. Configure a fonte de dados e o pool de recursos da seguinte forma para usar as configurações atuais do painel.

Figura 4-22 Usar as configurações atuais do painel



Criar um painel para exibir métricas

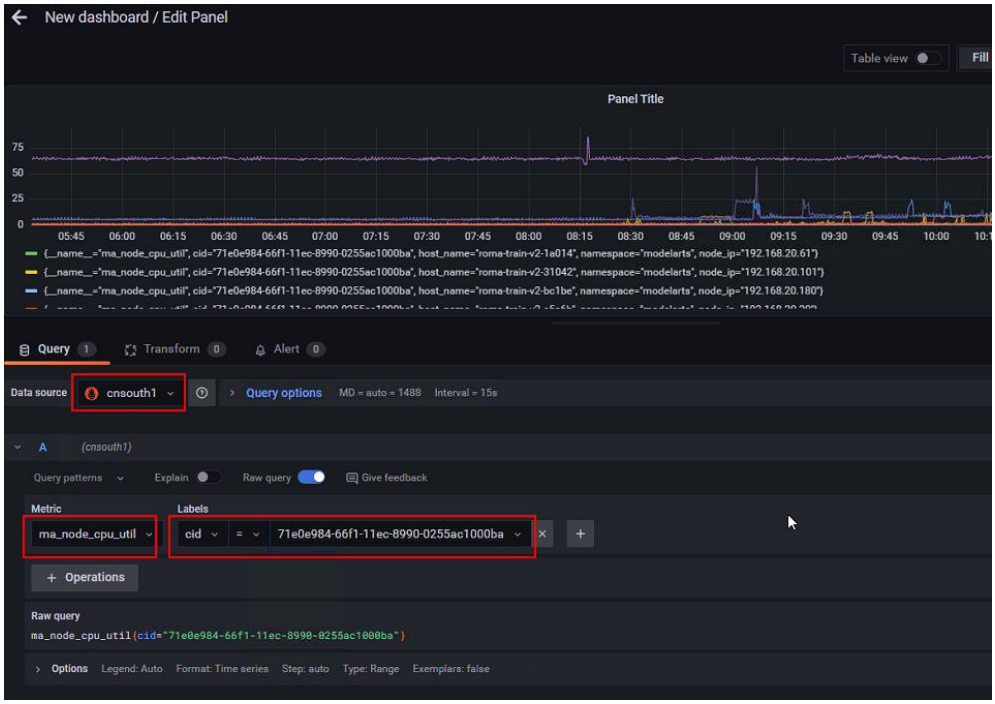
1. Abra **Dashboards**, clique em **New** e escolha **New Dashboard**.
2. Clique em **Add a new panel**.
3. Na página **New dashboard / Edit Panel**, defina os seguintes parâmetros:

   Data source: **fonte de dados do Grafana configurada**

   Metric: nome da métrica. Você pode obter a métrica a ser consultada referindo-se a **Tabela 4-3**, **Tabela 4-4** e **Tabela 4-5**.

   Labels: usado para filtrar a métrica. Para mais detalhes, consulte **Tabela 4-6**.

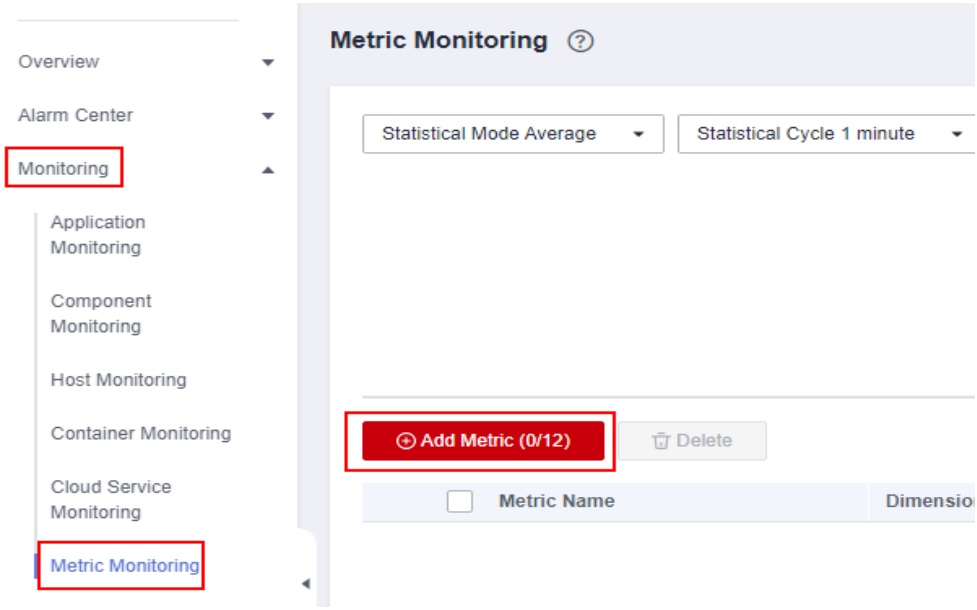
Figura 4-23 Criar um painel para visualizar métricas



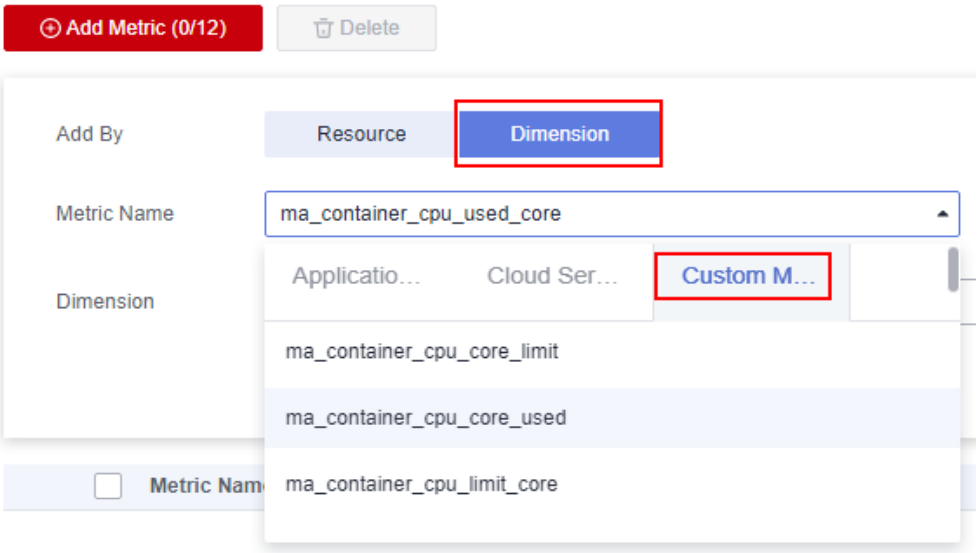
4.3 Exibição de todas as métricas de monitoramento do ModelArts no console do AOM

O ModelArts coleta periodicamente o uso de métricas criticais (como GPUs, NPUs, CPUs e memória) de cada nó em um pool de recursos, bem como o uso de métricas criticais do ambiente de desenvolvimento, trabalhos de treinamento e serviços de inferência, e reporta os dados ao AOM. Você pode ver as informações no AOM.

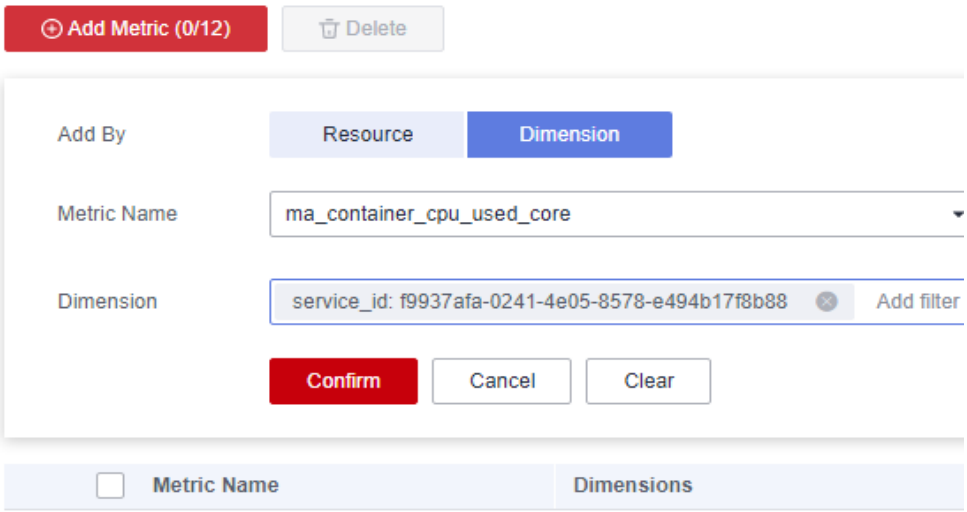
1. Faça logon no console e procure por **AOM** para ir para o console do AOM.
2. Escolha **Monitoring** > **Metric Monitoring**. Na página **Metric Monitoring** exibida, clique em **Add Metric**.



3. Adicione métricas e clique em Confirm.



- **Add By**: selecione Dimension.
- **Metric Name**: clique em **Custom Metrics**. Selecione as desejadas para consulta. Para obter detalhes, consulte **Tabela 4-3**, **Tabela 4-4** e **Tabela 4-5**.
- Dimension: insira a tag para filtrar a métrica. Para mais detalhes, consulte **Tabela 4-6**. O seguinte mostra um exemplo.



4. Veja as métricas.

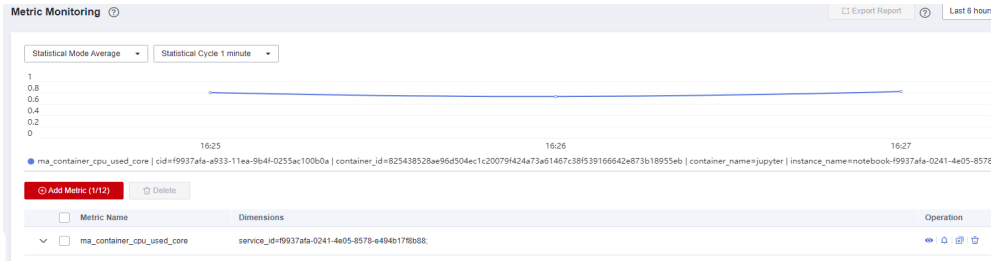


Tabela 4-3 Métricas de contêiner

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
CPU	CPU Usage	ma_container_cpu_util	Uso da CPU de um objeto medido	%	0%–100%
	Used CPU Cores	ma_container_cpu_used_core	Número de núcleos de CPU usados por um objeto medido	Núcleos	≥ 0
	Total CPU Cores	ma_container_cpu_limit_core	Número total de núcleos de CPU que foram aplicados a um objeto medido	Núcleos	≥ 1
Memória	Total Physical Memory	ma_container_memory_capacity_megabytes	Memória física total que foi aplicada para um objeto medido	MB	≥ 0
	Physical Memory Usage	ma_container_memory_util	Porcentagem da memória física usada em relação à memória física total	%	0%–100%

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
	Used Physical Memory	ma_container_memory_used_megabytes	Memória física que foi usada por um objeto medido (**container_memory_working_set_bytes** no conjunto de trabalho atual) (Uso de memória em um conjunto de trabalho = página anônima ativa e cache, e a página de arquivo ≤ **container_memory_usage_bytes**)	MB	≥ 0
Armazenamento	Disk Read Rate	ma_container_disk_read_kilobytes	Volume de dados lidos de um disco por segundo	KB/S	≥ 0
	Disk Write Rate	ma_container_disk_write_kilobytes	Volume de dados gravados em um disco por segundo	KB/S	≥ 0
Memória de GPU	Total GPU Memory	ma_container_gpu_mem_total_megabytes	Memória total da GPU de um trabalho de treinamento	MB	> 0
	GPU Memory Usage	ma_container_gpu_mem_util	Porcentagem da memória da GPU usada em relação à memória total da GPU	%	0%–100%
	Used GPU Memory	ma_container_gpu_mem_used_megabytes	Memória da GPU usada por um objeto medido	MB	≥ 0
GPU	GPU Usage	ma_container_gpu_util	Uso da GPU de um objeto medido	%	0%–100%

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
	GPU Memory Bandwidth Usage	ma_container_gpu_mem_copy_util	Uso da largura de banda da memória da GPU de um objeto medido. Por exemplo, a largura de banda máxima da memória da GPU de NVIDIA V100 é de 900 GB/s. Se a largura de banda da memória atual for de 450 GB/s, o uso da largura de banda da memória será de 50%.	%	0%–100%
	GPU Encoder Usage	ma_container_gpu_enc_util	Uso do codificador de GPU de um objeto medido	%	%
	GPU Decoder Usage	ma_container_gpu_dec_util	Uso do decodificador de GPU de um objeto medido	%	%
	GPU Temperature	DCGM_FI_DEV_GPU_TEMP	Temperatura da GPU	°C	Número natural
	GPU Power	DCGM_FI_DEV_POWER_USAGE	Potência da GPU	Watt (W)	> 0
	GPU Memory Temperature	DCGM_FI_DEV_MEMORY_TEMP	Temperatura da memória da GPU	°C	Número natural
I/O de rede	Downlink Rate (BPS)	ma_container_network_receive_bytes	Taxa de tráfego de entrada de um objeto medido	Bytes/s	≥ 0
	Downlink Rate (PPS)	ma_container_network_receive_packets	Número de pacotes de dados recebidos por uma NIC por segundo	Pacotes/s	≥ 0
	Downlink Error Rate	ma_container_network_receive_error_packets	Número de pacotes de erro recebidos por uma NIC por segundo	Pacotes/s	≥ 0

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
	Uplink Rate (BPS)	ma_container_network_transmit_bytes	Taxa de tráfego de saída de um objeto medido	Bytes/s	≥ 0
	Uplink Error Rate	ma_container_network_transmit_error_packets	Número de pacotes de erro enviados por uma NIC por segundo	Pacotes/s	≥ 0
	Uplink Rate (PPS)	ma_container_network_transmit_packets	Número de pacotes de dados enviados por uma NIC por segundo	Pacotes/s	≥ 0
Métricas de serviço de notebook	Notebook Cache Directory Size	ma_container_notebook_cache_dir_size_bytes	Um disco local de alta velocidade é anexado ao diretório **/cache** para instâncias de notebook de GPU. Essa métrica indica o tamanho total do diretório.	Bytes	≥ 0
	Notebook Cache Directory Utilization	ma_container_notebook_cache_dir_util	Um disco local de alta velocidade é anexado ao diretório **/cache** para instâncias de notebook de GPU. Essa métrica indica a utilização do diretório.	%	0%–100%

Tabela 4-4 Métricas de nó (coletadas apenas em pools de recursos dedicados)

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
CPU	Total CPU Cores	ma_node_cpu_limit_core	Número total de núcleos de CPU que foram aplicados a um objeto medido	Núcleos	≥ 1

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
	Used CPU Cores	ma_node_cpu_used_core	Número de núcleos de CPU usados por um objeto medido	Núcleos	≥ 0
	CPU Usage	ma_node_cpu_util	Uso da CPU de um objeto medido	%	0%–100%
	CPU I/O Wait Time	ma_node_cpu_iowait_counter	Tempo de espera de I/O de disco acumulado desde a inicialização do sistema	jiffies	≥ 0
Memória	Physical Memory Usage	ma_node_memory_util	Porcentagem da memória física usada em relação à memória física total	%	0%–100%
	Total Physical Memory	ma_node_memory_total_megabytes	Memória física total que foi aplicada para um objeto medido	MB	≥ 0
I/O de rede	Downlink Rate (BPS)	ma_node_network_receive_rate_bytes_seconds	Taxa de tráfego de entrada de um objeto medido	Bytes/s	≥ 0
	Uplink Rate (BPS)	ma_node_network_transmit_rate_bytes_seconds	Taxa de tráfego de saída de um objeto medido	Bytes/s	≥ 0

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
Armazenamento	Disk Read Rate	ma_node_disk_read_rate_kilobytes_seconds	Volume de dados lidos de um disco por segundo (somente discos de dados usados por contêineres são coletados.)	KB/S	≥ 0
	Disk Write Rate	ma_node_disk_write_rate_kilobytes_seconds	Volume de dados gravados em um disco por segundo (somente discos de dados usados por contêineres são coletados.)	KB/S	≥ 0
	Total Cache	ma_node_cache_space_capacity_megabytes	Cache total do espaço do Kubernetes	MB	≥ 0
	Used Cache	ma_node_cache_space_used_capacity_megabytes	Cache usado do espaço do Kubernetes	MB	≥ 0
	Total Container Space	ma_node_container_space_capacity_megabytes	Espaço total do contêiner	MB	≥ 0
	Used Container Space	ma_node_container_space_used_capacity_megabytes	Espaço de contêiner usado	MB	≥ 0

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
	Disk Information	ma_node_disk_info	Informações básicas do disco	N/D	≥ 0
	Total Reads	ma_node_disk_reads_completed_total	Número total de leituras bem-sucedidas	N/D	≥ 0
	Merged Reads	ma_node_disk_reads_merged_total	Número de leituras mescladas	N/D	≥ 0
	Bytes Read	ma_node_disk_read_bytes_total	Número total de bytes lidos com sucesso	Bytes	≥ 0
	Read Time Spent	ma_node_disk_read_time_seconds_total	Tempo gasto em todas as leituras	Segundos	≥ 0
	Total Writes	ma_node_disk_writes_completed_total	Número total de gravações bem-sucedidas	N/D	≥ 0
	Merged Writes	ma_node_disk_writes_merged_total	Número de gravações mescladas	N/D	≥ 0
	Bytes gravados	ma_node_disk_written_bytes_total	Total number of bytes that are successfully written	Bytes	≥ 0
	Write Time Spent	ma_node_disk_write_time_seconds_total	Tempo gasto em todas as operações de gravação	Segundos	≥ 0

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
	Ongoing I/Os	ma_node_disk_io_now	Número de I/Os em andamento	N/D	≥ 0
	I/O Execution Duration	ma_node_disk_io_time_seconds_total	Tempo gasto na execução de I/Os	Segundos	≥ 0
	I/O Execution Weighted Time	ma_node_disk_io_time_weighted_seconds_tota	O número ponderado de segundos gastos em I/Os	Segundos	≥ 0
GPU	GPU Usage	ma_node_gpu_util	Uso da GPU de um objeto medido	%	0%–100%
	Total GPU Memory	ma_node_gpu_mem_total_megabytes	Memória total da GPU de um objeto medido	MB	> 0
	GPU Memory Usage	ma_node_gpu_mem_util	Porcentagem da memória da GPU usada em relação à memória total da GPU	%	0%–100%
	Used GPU Memory	ma_node_gpu_mem_used_megabytes	Memória da GPU usada por um objeto medido	MB	≥ 0
	Tasks on a Shared GPU	node_gpu_share_job_count	Número de tarefas em execução em uma GPU compartilhada	Número	≥ 0
	GPU Temperature	DCGM_FI_DEV_GPU_TEMP	Temperatura da GPU	°C	Número natural

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
	GPU Power	DCGM_FI_DEV_POWER_USAGE	Potência da GPU	Watt (W)	> 0
	GPU Memory Temperature	DCGM_FI_DEV_MEMORY_TEMP	Temperatura da memória da GPU	°C	Número natural
Rede InfiniBand ou RoCE	Total Amount of Data Received by a NIC	ma_node_infiniband_port_received_data_bytes_total	O número total de dados octetos, dividido por 4, (contando em palavras duplas, 32 bits), recebidos em todos os VLs da porta.	(contando em palavras duplas, 32 bits)	≥ 0
	Total Amount of Data Sent by a NIC	ma_node_infiniband_port_transmitted_data_bytes_total	O número total de dados octetos, dividido por 4, (contando em palavras duplas, 32 bits), transmitidos em todos os VLs a partir da porta.	(contando em palavras duplas, 32 bits)	≥ 0

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
Status de montagem do NFS	NFS Getattr Congestion Time	ma_node_mountstats_getattr_backlog_wait	Getattr é uma operação do NFS que recupera os atributos de um arquivo ou diretório, como tamanho, permissões, proprietário, etc. A espera de lista de pendências é o tempo que as solicitações do NFS precisam esperar na fila de lista de pendências antes de serem enviadas para o servidor do NFS. Indica o congestionamento no lado do cliente do NFS. Uma alta espera de lista de pendências pode causar desempenho ruim do NFS e tempos de resposta lentos do sistema.	ms	≥ 0

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
	NFS Getattr Round Trip Time	ma_node_mountstats_getattr_rtt	Getattr é uma operação do NFS que recupera os atributos de um arquivo ou diretório, como tamanho, permissões, proprietário, etc. RTT significa Round Trip Time e é o momento a partir do momento em que o cliente RPC do kernel envia a solicitação RPC até o momento em que recebe a resposta34. O RTT inclui o tempo de trânsito da rede e o tempo de execução do servidor. O RTT é uma boa medida para a latência do NFS. Um RTT alto pode indicar problemas	ms	≥ 0

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
			de rede ou servidor.		

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
	NFS Access Congestion Time	ma_node_mountstats_access_backlog_wait	Access é uma operação do NFS que verifica as permissões de acesso de um arquivo ou diretório para um determinado usuário. A espera de lista de pendências é o tempo que as solicitações do NFS precisam esperar na fila de lista de pendências antes de serem enviadas para o servidor do NFS. Indica o congestionamento no lado do cliente do NFS. Uma alta espera de lista de pendências pode causar desempenho ruim do NFS e tempos de resposta lentos do sistema.	ms	≥ 0

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
	NFS Access Round Trip Time	ma_node_mountstats_access_rtt	Access é uma operação do NFS que verifica as permissões de acesso de um arquivo ou diretório para um determinado usuário. RTT significa Round Trip Time e é o momento a partir do momento em que o cliente RPC do kernel envia a solicitação RPC até o momento em que recebe a resposta34. O RTT inclui o tempo de trânsito da rede e o tempo de execução do servidor. O RTT é uma boa medida para a latência do NFS. Um RTT alto pode indicar problemas de rede ou servidor.	ms	≥ 0

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
	NFS Lookup Congestion Time	ma_node_mountstats_lookup_backlog_wait	Lookup é uma operação do NFS que resolve um nome de arquivo em um diretório para um identificador de arquivo. A espera de lista de pendências é o tempo que as solicitações do NFS precisam esperar na fila de lista de pendências antes de serem enviadas para o servidor do NFS. Indica o congestionamento no lado do cliente do NFS. Uma alta espera de lista de pendências pode causar desempenho ruim do NFS e tempos de resposta lentos do sistema.	ms	≥ 0

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
	NFS Lookup Round Trip Time	ma_node_mountstats_lookup_rtt	Lookup é uma operação do NFS que resolve um nome de arquivo em um diretório para um identificador de arquivo. RTT significa Round Trip Time e é o momento a partir do momento em que o cliente RPC do kernel envia a solicitação RPC até o momento em que recebe a resposta34. O RTT inclui o tempo de trânsito da rede e o tempo de execução do servidor. O RTT é uma boa medida para a latência do NFS. Um RTT alto pode indicar problemas de rede ou servidor.	ms	≥ 0

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
	NFS Read Congestion Time	ma_node_mountstats_read_backlog_wait	Read é uma operação do NFS que lê dados de um arquivo. A espera de lista de pendências é o tempo que as solicitações do NFS precisam esperar na fila de lista de pendências antes de serem enviadas para o servidor do NFS. Indica o congestionamento no lado do cliente do NFS. Uma alta espera de lista de pendências pode causar desempenho ruim do NFS e tempos de resposta lentos do sistema.	ms	≥ 0

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
	NFS Read Round Trip Time	ma_node_mountstats_read_rtt	Read é uma operação do NFS que lê dados de um arquivo. RTT significa Round Trip Time e é o momento a partir do momento em que o cliente RPC do kernel envia a solicitação RPC até o momento em que recebe a resposta34. O RTT inclui o tempo de trânsito da rede e o tempo de execução do servidor. O RTT é uma boa medida para a latência do NFS. Um RTT alto pode indicar problemas de rede ou servidor.	ms	≥ 0

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
	NFS Write Congestion Time	ma_node_mountstats_write_backlog_wait	Write é uma operação do NFS que grava dados em um arquivo. A espera de lista de pendências é o tempo que as solicitações do NFS precisam esperar na fila de lista de pendências antes de serem enviadas para o servidor do NFS. Indica o congestionamento no lado do cliente do NFS. Uma alta espera de lista de pendências pode causar desempenho ruim do NFS e tempos de resposta lentos do sistema.	ms	≥ 0

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
	NFS Write Round Trip Time	ma_node_mountstats_write_rtt	Write é uma operação do NFS que grava dados em um arquivo. RTT significa Round Trip Time e é o momento a partir do momento em que o cliente RPC do kernel envia a solicitação RPC até o momento em que recebe a resposta34. O RTT inclui o tempo de trânsito da rede e o tempo de execução do servidor. O RTT é uma boa medida para a latência do NFS. Um RTT alto pode indicar problemas de rede ou servidor.	ms	≥ 0

Tabela 4-5 Diagnóstico (InfiniBand, coletado somente em pools de recursos dedicados)

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
Rede InfiniBand ou RoCE	PortXmitData	infiniband_port_xmit_data_total	O número total de dados octetos, dividido por 4, (contando em palavras duplas, 32 bits), transmitidos em todos os VLs a partir da porta.	Contagem total	Número natural
	PortRcvData	infiniband_port_rcv_data_total	O número total de dados octetos, dividido por 4, (contando em palavras duplas, 32 bits), recebidos em todos os VLs da porta.	Contagem total	Número natural
	SymbolErrorCounter	infiniband_symbol_error_counter_total	Número total de pequenos erros de link detectados em uma ou mais pistas físicas.	Contagem total	Número natural
	LinkErrorRecoveryCounter	infiniband_link_error_recovery_counter_total	Número total de vezes que a máquina de estado Port Training concluiu com êxito o processo de recuperação de erro de link.	Contagem total	Número natural

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
	PortRcvErrors	infiniband_port_rcv_errors_total	Número total de pacotes contendo erros que foram recebidos na porta, incluindo: Erros físicos locais (ICRC, VCRC, LPCRC e todos os erros físicos que causam entrada nos estados BAD PACKET ou BAD PACKET DISCARD da máquina de estado do receptor de pacotes) Erros de pacotes de dados mal formados (operand, length, VL) Erros de pacotes de link mal formados (operand, length, VL) Pacotes descartados devido ao excesso do buffer (overflow)	Contagem total	Número natural
	LocalLinkIntegrityErrors	infiniband_local_link_integrity_errors_total	Esse contador indica o número de novas tentativas iniciadas por um receptor de camada de transferência de link.	Contagem total	Número natural
	PortRcvRemotePhysicalErrors	infiniband_port_rcv_remote_physical_errors_total	Número total de pacotes marcados com o delimitador EBP recebidos na porta.	Contagem total	Número natural
	PortRcvSwitchRelayErrors	infiniband_port_rcv_switch_relay_errors_total	Número total de pacotes recebidos na porta que foram descartados quando não puderam ser encaminhados pelo relé do interruptor pelos seguintes motivos: Mapeamento de DLID Mapeamento de VL Looping (porta de saída = porta de entrada)	Contagem total	Número natural

Categoria	Nome	Métrica	Descrição	Unidade	Intervalo de valor
	PortXmitWait	infiniband_port_transmit_wait_total	O número de ticks durante os quais a porta tinha dados para transmitir, mas nenhum dado foi enviado durante todo o tick (seja por insuficiência de créditos ou por falta de arbitragem).	Contagem total	Número natural
	PortXmitDiscards	infiniband_port_xmit_discards_total	Número total de pacotes de saída descartados pela porta porque a porta está inativa ou congestionada.	Contagem total	Número natural

Tabela 4-6 Nomes de métrica

Classificação	Métrica	Descrição
Métricas de contêiner	modelarts_service	Serviço ao qual um contêiner pertence, que pode ser **notebook**, **train** ou **infer**
	instance_name	Nome do pod ao qual o contêiner pertence
	service_id	ID da instância ou do trabalho exibido na página, por exemplo, **cf55829e-9bd3-48fa-8071-7ae870dae93a** para um ambiente de desenvolvimento **9f322d5a-b1d2-4370-94df-5a87de27d36e** para um trabalho de treinamento
	node_ip	Endereço IP do nó ao qual o contêiner pertence
	container_id	ID do contêiner
	cid	ID do cluster
	container_name	Nome do contêiner
	project_id	ID do projeto da conta à qual o usuário pertence
	user_id	ID do usuário da conta à qual pertence o usuário que submete o trabalho
	pool_id	ID de um pool de recursos correspondente a um pool de recursos dedicados físicos

Classificação	Métrica	Descrição
	pool_name	Nome de um pool de recursos correspondente a um pool de recursos dedicados físicos
	logical_pool_id	ID de um subpool lógico
	logical_pool_name	Nome de um subpool lógico
	gpu_uuid	UUID da GPU usada pelo contêiner
	gpu_index	Índice da GPU usada pelo contêiner
	gpu_type	Tipo da GPU usada pelo contêiner
	account_name	Nome da conta do criador de uma tarefa de treinamento, inferência ou ambiente de desenvolvimento
	user_name	Nome de usuário do criador de uma tarefa de treinamento, inferência ou ambiente de desenvolvimento
	task_creation_time	Momento em que uma tarefa de treinamento, inferência ou ambiente de desenvolvimento é criada
	task_name	Nome de uma tarefa de treinamento, inferência ou ambiente de desenvolvimento
	task_spec_code	Especificações de uma tarefa de treinamento, inferência ou ambiente de desenvolvimento
	cluster_name	Nome do cluster do CCE
Métricas de nó	cid	ID do cluster do CCE ao qual o nó pertence
	node_ip	Endereço IP do nó
	host_name	Nome de host de um nó
	pool_id	ID de um pool de recursos correspondente a um pool de recursos dedicados físicos
	project_id	ID do projeto do usuário em um pool de recursos dedicados físicos
	gpu_uuid	UUID de uma GPU de nó
	gpu_index	Índice de uma GPU de nó
	gpu_type	Tipo de uma GPU de nó
	device_name	Nome do dispositivo de uma NIC de rede RoCE ou InfiniBand
	port	Número da porta da NIC InfiniBand

Classificação	Métrica	Descrição
	physical_state	Status de cada porta na NIC InfiniBand
	firmware_version	Versão de firmware da NIC InfiniBand
	filesystem	Sistema de arquivos montado no NFS
	mount_point	Ponto de montagem do NFS
Diagnósticos	cid	ID do cluster do CCE ao qual pertence o nó com a GPU equipada
	node_ip	Endereço IP do nó em que a GPU reside
	pool_id	ID de um pool de recursos correspondente a um pool de recursos dedicados físicos
	project_id	ID do projeto do usuário em um pool de recursos dedicados físicos
	gpu_uuid	UUID da GPU
	gpu_index	Índice de uma GPU de nó
	gpu_type	Tipo de uma GPU de nó
	device_name	Nome de um dispositivo de rede ou de um dispositivo de disco
	port	Número da porta da NIC InfiniBand
	physical_state	Status de cada porta na NIC InfiniBand
	firmware_version	Versão de firmware da NIC InfiniBand

