# ModelArts

# Visão geral de serviço

**Edição** 01

**Data** 17-06-2024

**HUAWEI 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, serviços e funcionalidades adquiridos são estipulados pelo contrato feito entre a Huawei e o cliente. Todos ou parte dos produtos, serviços e funcionalidades descritos neste documento pode 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 ESTÁ" sem garantias, ou representações de qualquer tipo, seja expressa ou implícita.

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 Technologies Co., Ltd.

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

Site: https://www.huawei.com

Email: support@huawei.com

# Índice



# 1 Infográficos

# 1.1 O que é o ModelArts?

# 2 O que é o ModelArts?

O ModelArts é uma plataforma de desenvolvimento de IA projetada para desenvolvedores e cientistas de dados de todos os níveis. Ele permite que você crie, treine e implemente modelos rapidamente em qualquer lugar (da nuvem até a borda) e gerencie fluxos de trabalho de IA de ciclo de vida completo. O ModelArts acelera o desenvolvimento e promove a inovação da IA com recursos distintos, incluindo pré-processamento e rotulagem automática de dados, treinamento distribuído, construção automatizada de modelos e execução de fluxo de trabalho com um clique.

O ModelArts abrange todos os estágios do desenvolvimento de IA, incluindo processamento de dados, desenvolvimento de algoritmos e treinamento e implementação de modelos. As tecnologias subjacentes do ModelArts suportam vários recursos de computação heterogêneos, permitindo que os desenvolvedores selecionem e usem recursos de forma flexível. Além disso, o ModelArts suporta estruturas populares de desenvolvimento de IA de código aberto, como TensorFlow, PyTorch e MindSpore. O ModelArts também permite que você use estruturas de algoritmo personalizadas adaptadas às suas necessidades.

O ModelArts visa simplificar o desenvolvimento da IA.

O ModelArts é adequado para desenvolvedores de IA com diferentes níveis de experiência em desenvolvimento. Os desenvolvedores de serviços podem usar o ExeML para criar rapidamente aplicações de IA sem codificação. Os iniciantes podem usar diretamente algoritmos internos para criar aplicações de IA. Os engenheiros de IA podem usar vários ambientes de desenvolvimento para compilar rapidamente o código para modelagem e desenvolvimento de aplicações.

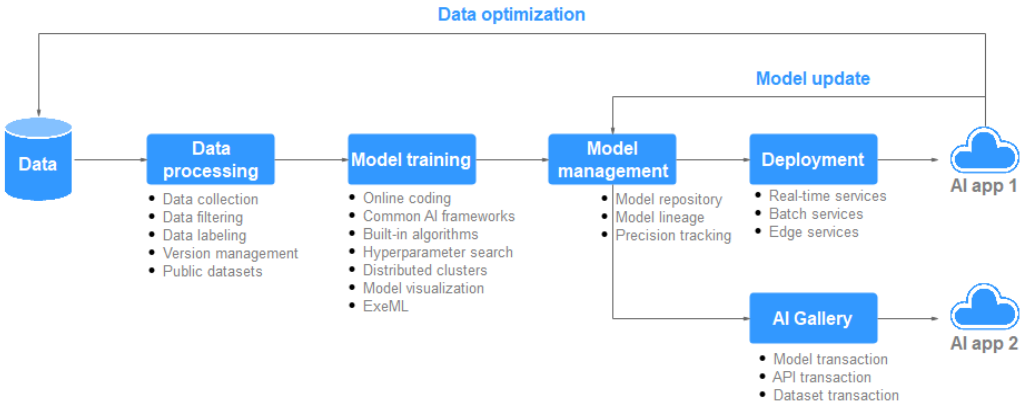
## Arquitetura do produto

O ModelArts oferece suporte a todo o processo de desenvolvimento, incluindo processamento de dados e treinamento, gerenciamento e implementação de modelos. Ele também fornece Galeria de IA para compartilhar modelos.

O ModelArts oferece suporte a vários cenários de aplicações de IA, como classificação de imagens, detecção de objetos, análise de vídeo, reconhecimento de fala, recomendação de produtos e detecção de exceções.

**Figura 2-1** Arquitetura do ModelArts




## Vantagens do produto

- **Plataforma completa**

  A plataforma de desenvolvimento de IA pronta para uso e de ciclo de vida completo fornece processamento de dados completo e desenvolvimento, treinamento, gerenciamento e implementação de modelos.

- **Facilidade de uso**
  - Múltiplos modelos integrados fornecidos e uso gratuito de modelos de código aberto
  - Otimização automática de hiperparâmetros
  - Desenvolvimento sem código e operações simplificadas
  - Implementação de modelos com um clique para a nuvem, borda e dispositivos

- **Alto desempenho**
  - A estrutura de aprendizado profundo MoXing autodesenvolvida acelera o desenvolvimento e o treinamento de algoritmos.
  - A utilização otimizada da GPU acelera a inferência em tempo real.
  - Modelos executados em chips Ascend AI obtêm inferência mais eficiente.

- **Flexível**
  - Principais estruturas de código aberto, como TensorFlow, PyTorch e MindSpore
  - Principais GPUs
  - Chips Ascend
  - Uso exclusivo de recursos dedicados
  - Imagens personalizadas para estruturas e operadores personalizados

## Usar o ModelArts pela primeira vez

Se você é um usuário iniciante, as informações a seguir ajudarão você a se familiarizar com o ModelArts:

- **Conceitos básicos**

  **Conhecimento básico** descreve os conceitos básicos de ModelArts, incluindo o processo básico e conceitos de desenvolvimento de IA e conceitos específicos e funções de ModelArts.

- **Primeiros passos**

  ***Primeiros passos*** fornece amostras com operações detalhadas, ajudando você a começar a usar o ModelArts.

- **Melhores práticas**

  O ModelArts oferece suporte a vários mecanismos de código aberto e fornece casos de uso extensivos com base nos mecanismos e funções. Você pode criar e implantar modelos consultando as **Melhores práticas**.

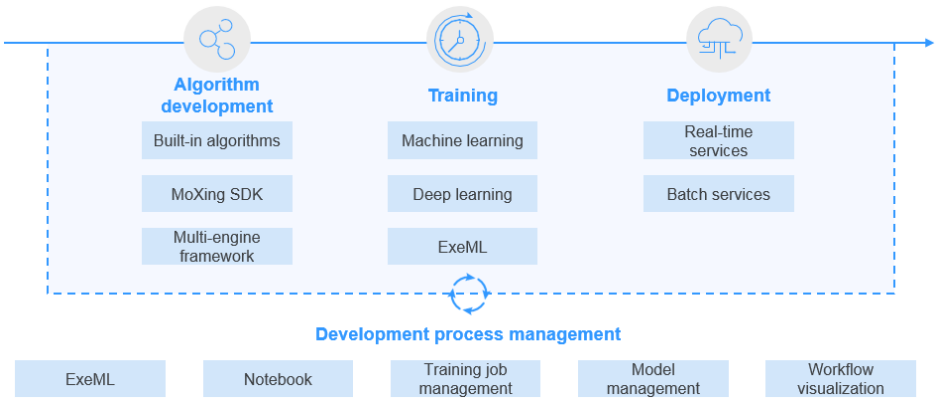
- **Outras funções e guias de operação**
  - Se você é um desenvolvedor de serviços, pode usar o ExeML para criar modelos rapidamente sem codificação. Para obter detalhes, consulte ***Guia de usuário (ExeML)***.
  - Se você é um engenheiro de IA, pode usar uma ou mais funções em seu desenvolvimento de IA, incluindo **DevEnviron**, **preparação de dados**, **rotulagem de dados**, **desenvolvimento de modelos** e **inferência**. Você pode usar uma ou mais funções em seu desenvolvimento de IA.
  - Se você quiser usar APIs ou SDKs do ModelArts para desenvolvimento de IA, consulte ***Referência de API*** ou ***Referência de SDK***.

# 3 Funções

Os engenheiros de IA enfrentam desafios na instalação e configuração de várias ferramentas de IA, preparação de dados e treinamento de modelos. Para enfrentar esses desafios, a plataforma de desenvolvimento de IA ModelArts é fornecida. A plataforma integra preparação de dados, desenvolvimento de algoritmos, treinamento e implementação de modelos no ambiente de produção, permitindo que os engenheiros de IA realizem o desenvolvimento de IA completo.

**Figura 3-1** Visão geral das funções




O ModelArts possui os seguintes recursos:

- **Governança de dados**

  Gerencia a preparação de dados, como filtragem e rotulagem de dados e versões de conjuntos de dados.

- **Treinamento de modelo rápido e simplificado**

  Permite treinamento distribuído de alto desempenho e simplifica a codificação com a estrutura de aprendizado profundo MoXing desenvolvida.

- **Sinergia entre nuvem, borda e dispositivo**

  Implementa modelos em vários ambientes de produção, como dispositivos, borda e nuvem, e suporta inferência em lote e em tempo real.

- **Aprendizado automático**

Permite a criação de modelos sem codificação e suporta classificação de imagens, detecção de objetos e análise preditiva.

# 4 Conhecimento básico

## 4.1 Introdução ao ciclo de vida de desenvolvimento da IA

### O que é IA

A inteligência artificial (IA) é uma tecnologia capaz de simular a cognição humana por meio de máquinas. A capacidade central da IA é fazer um julgamento ou previsão com base em uma determinada entrada.

### Qual é o propósito do desenvolvimento da IA

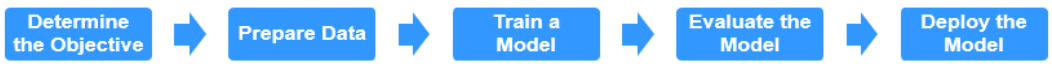
O desenvolvimento da IA visa processar e extrair informações centralmente de volumes de dados para resumir os padrões internos dos objetos de estudo.

Grandes volumes de dados coletados são computados, analisados, resumidos e organizados usando estatísticas apropriadas, aprendizado de máquina e métodos de aprendizado profundo para maximizar o valor dos dados.

### Processo básico de desenvolvimento de IA

O processo básico de desenvolvimento de IA inclui as seguintes etapas: determinação de um objetivo, preparação de dados e treinamento, avaliação e implementação de um modelo.

**Figura 4-1** Processo de desenvolvimento de IA




**Passo 1** **Determine um objetivo.**

Antes de iniciar o desenvolvimento da IA, determine o que analisar. Quais problemas você quer resolver? Qual é o objetivo do negócio? Classifique a estrutura de desenvolvimento de IA e as ideias com base no entendimento do negócio. Por exemplo, classificação de imagens e detecção de objetos. Diferentes projetos têm diferentes requisitos para dados e métodos de desenvolvimento de IA.

**Passo 2** **Prepare os dados.**

A preparação de dados refere-se à coleta e ao pré-processamento de dados.

A preparação de dados é a base do desenvolvimento da IA. Quando você coleta e integra dados relacionados com base no objetivo determinado, o mais importante é garantir a autenticidade e a confiabilidade dos dados obtidos. Normalmente, você não pode coletar todos os dados ao mesmo tempo. Na fase de rotulagem de dados, você pode achar que algumas fontes de dados estão faltando e, em seguida, pode ser necessário ajustar e otimizar os dados repetidamente.

**Passo 3 Prepare os dados.**

A modelagem envolve a análise dos dados preparados para encontrar a causalidade, as relações internas e os padrões regulares, fornecendo referências para a tomada de decisões comerciais. Após o treinamento do modelo, geralmente um ou mais modelos de aprendizado de máquina ou de aprendizado profundo são gerados. Esses modelos podem ser aplicados a novos dados para obter previsões e resultados de avaliação.

Um grande número de desenvolvedores desenvolve e treina modelos exigidos por serviços relevantes baseados em mecanismos populares de IA, como TensorFlow, Spark_MLlib, MXNet, Caffe, PyTorch, XGBoost-Sklearn e MindSpore.

**Passo 4 Avalie o modelo.**

Um modelo gerado pelo treinamento precisa ser avaliado. Normalmente, você não pode obter um modelo satisfatório após a primeira avaliação e pode precisar ajustar repetidamente os parâmetros e dados do algoritmo para otimizar ainda mais o modelo.

Algumas métricas comuns, como a precisão, a recuperação e a área sob a curva (AUC), ajudam a avaliar e obter um modelo satisfatório.

**Passo 5 Implemente o modelo.**

O desenvolvimento e o treinamento do modelo são baseados em dados existentes (que podem ser dados de teste). Depois que um modelo satisfatório é obtido, o modelo precisa ser formalmente aplicado a dados reais ou dados recém-gerados para previsão, avaliação e visualização. As descobertas podem então ser relatadas aos tomadores de decisão de maneira intuitiva, ajudando-os a desenvolver as estratégias de negócios corretas.

**----Fim**

# 4.2 Conceitos básicos de desenvolvimento de IA

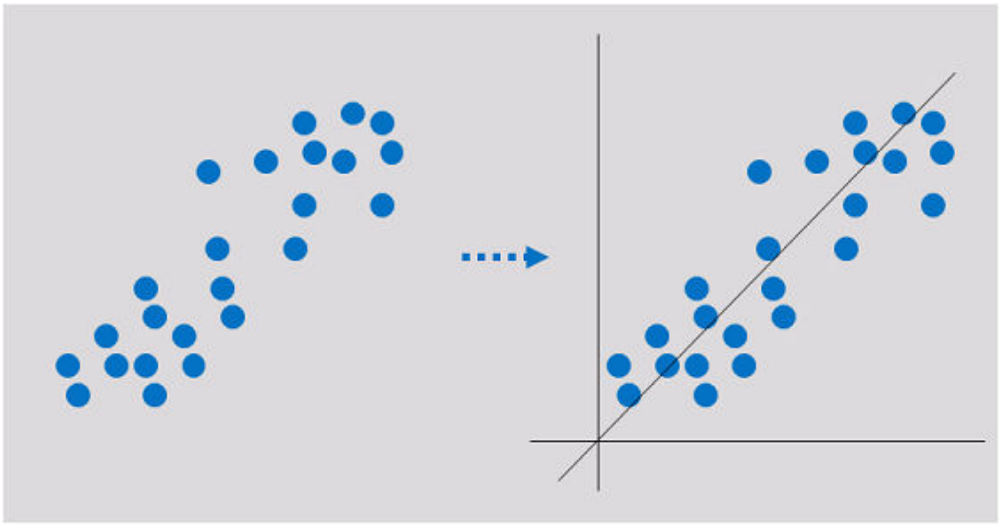
O aprendizado de máquina é classificado em aprendizado supervisionado, não supervisionado e por reforço.

- O aprendizado supervisionado usa amostras rotuladas para ajustar os parâmetros dos classificadores para alcançar o desempenho necessário. Pode ser considerado como um aprendizado com um professor. O aprendizado supervisionado comum inclui regressão e classificação.
- O aprendizado não supervisionado é usado para encontrar estruturas ocultas em dados não rotulados. Clustering é uma forma de aprendizado não supervisionado.
- Aprendizado por reforço é uma área de aprendizado de máquina preocupada com a forma como os agentes de software devem agir em um ambiente para maximizar alguma noção de recompensa cumulativa.

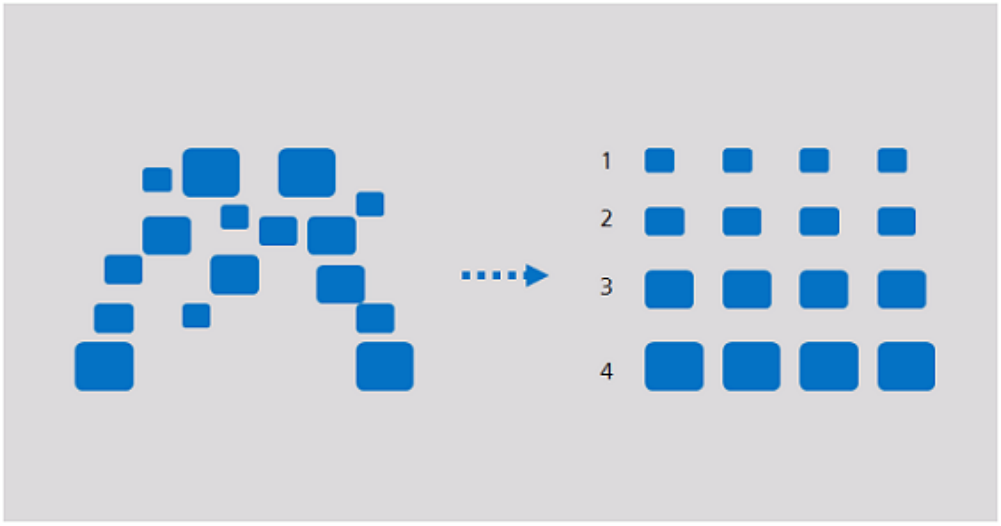
## Regressão

A regressão reflete o recurso de tempo dos atributos de dados e gera uma função que mapeia um atributo de dados para uma previsão de variável real para encontrar a dependência entre a variável e o atributo. A regressão analisa principalmente dados e prevê dados e relacionamento de dados. A regressão pode ser usada para desenvolvimento de clientes, retenção, prevenção de rotatividade de clientes, análise do ciclo de vida da produção, previsão de tendências de vendas e promoção direcionada.



## Classificação

A classificação envolve definir um conjunto de categorias com base nas características comuns dos objetos e identificar a qual categoria um objeto pertence. A classificação pode ser usada para classificação de clientes, propriedades de clientes, análise de recursos, análise de satisfação do cliente e previsão de tendências de compra do cliente.

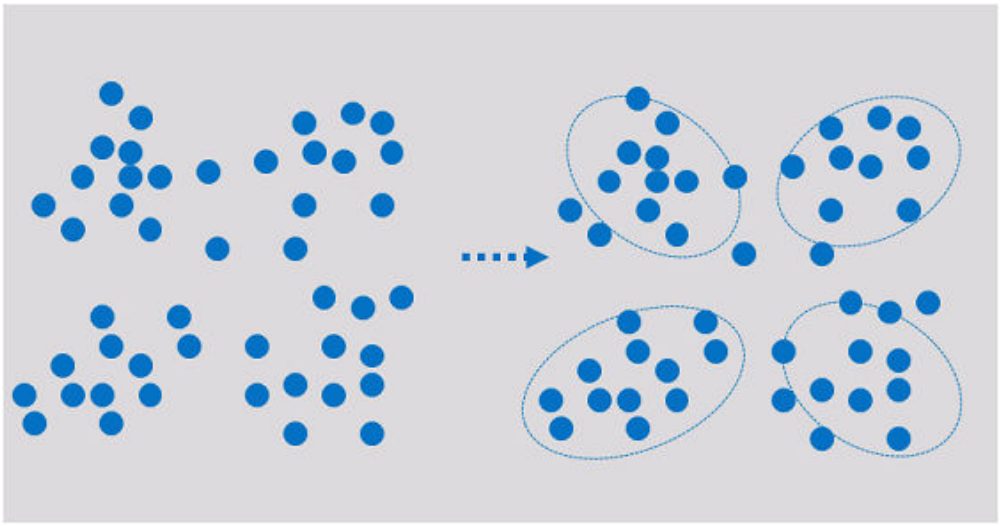



## Clustering

Clustering envolve o agrupamento de um conjunto de objetos de tal forma que os objetos no mesmo grupo são mais semelhantes uns aos outros do que aqueles em outros grupos.

Clustering pode ser usado para segmentação de clientes, análise de características do cliente, previsão de tendência de compra do cliente e segmentação de mercado.



Clustering analisa objetos de dados e produz rótulos de classe. Os objetos são agrupados com base nas semelhanças maximizadas e minimizadas para formar clusters. Desta forma, os objetos no mesmo cluster são mais semelhantes entre si do que aqueles em outros clusters.

# 4.3 Conceitos comuns do ModelArts

## ExeML

O ExeML é o processo de automatização do projeto de modelo, ajuste de parâmetros e treinamento de modelo, compactação de modelo e implementação de modelo com os dados rotulados. O processo é livre de código e não requer que os desenvolvedores tenham experiência em desenvolvimento de modelos. Um modelo pode ser construído em três etapas: rotular dados, treinar um modelo e implementar o modelo.

## Device-Edge-Cloud

Device-Edge-Cloud indica dispositivos, nós de borda inteligentes e a nuvem pública.

## Inferência

Inferência é o processo de derivar um novo julgamento de um julgamento conhecido de acordo com uma determinada estratégia. Na IA, as máquinas simulam a inteligência humana e completam a inferência baseada em redes neurais.

## Inferência em tempo real

A inferência em tempo real especifica um serviço da Web que fornece um resultado de inferência para cada solicitação de inferência.

## Inferência em lote

A inferência em lote especifica uma tarefa em lote que processa dados em lote para inferência.

### Chip Ascend

Os chips Ascend são uma série de chips de IA desenvolvidos pela Huawei com alto desempenho de computação e baixo consumo de energia.

## 4.4 Introdução às ferramentas de desenvolvimento

NOTA

Este documento descreve as funções do notebook DevEnviron da nova versão.

O desenvolvimento de software é um processo de reduzir os custos do desenvolvedor e melhorar a experiência de desenvolvimento. No desenvolvimento de IA, o ModelArts dedica-se a melhorar a experiência de desenvolvimento de IA e simplificar o processo de desenvolvimento. O DevEnviron do ModelArts usa recursos da nuvem nativa e integra a cadeia de ferramentas de desenvolvimento para fornecer uma melhor experiência de IA na nuvem para desenvolvimento, exploração e ensino de IA.

Notebook de ModelArts para colaboração perfeita na nuvem e no local

- Plug-ins de JupyterLab na nuvem, IDE local e ModelArts para desenvolvimento e depuração remotos, adaptados às suas necessidades
- Ambiente de desenvolvimento em nuvem com recursos de computação de IA, armazenamento em nuvem e mecanismos de IA integrados
- Ambiente de tempo de execução personalizado salvo como uma imagem para treinamento e inferência

### Recurso 1: desenvolvimento remoto, permitindo acesso remoto ao notebook a partir de um IDE local

O notebook da nova versão oferece desenvolvimento remoto. Depois de ativar o SSH remoto, você pode acessar remotamente o ambiente de desenvolvimento do notebook de ModelArts para depurar e executar código de um IDE local.

Devido a recursos locais limitados, os desenvolvedores que usam um IDE local executam e depuram o código normalmente em um servidor de CPU ou GPU compartilhado entre os membros da equipe. Construir e manter o servidor de CPU ou GPU é caro.

As instâncias de notebook de ModelArts estão prontas para uso com vários mecanismos e flavors internos para você selecionar. Você pode usar um ambiente de contêiner dedicado. Somente após configurações simples, você pode acessar remotamente o ambiente para executar e depurar o código do IDE local.

O notebook de ModelArts pode ser considerado como uma extensão de um ambiente de desenvolvimento local. As operações como leitura de dados, treinamento e salvamento de arquivos são as mesmas realizadas em um ambiente local.

O notebook de ModelArts permite que você use recursos na nuvem com os hábitos de codificação locais inalterados.

Um IDE local suporta o código do Visual Studio (VS), o PyCharm e o SSH.

## Recurso 2: imagens predefinidas prontas para uso com configurações otimizadas e suporte a mecanismos de IA convencionais

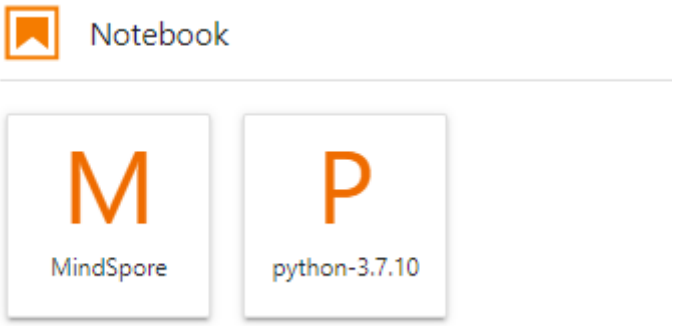
Os mecanismos de IA e as versões pré-configuradas em cada imagem são fixos. Ao criar uma instância de notebook, especifique um mecanismo de IA e uma versão, incluindo o tipo de chip.

O DevEnviron do ModelArts fornece um grupo de imagens predefinidas, incluindo imagens PyTorch, TensorFlow e MindSpore. Você pode usar uma imagem predefinida para iniciar a instância do notebook. Após o desenvolvimento na instância, envie um trabalho de treinamento sem qualquer adaptação.

As versões de imagem predefinidas no ModelArts são determinadas com base no feedback do usuário e na estabilidade da versão. Se o seu desenvolvimento pode ser feito usando as versões predefinidas no ModelArts, por exemplo, o MindSpore 1.5, use imagens predefinidas. Essas imagens foram totalmente verificadas e têm muitos pacotes de instalação comumente usados embutidos. Elas estão prontas para usar, aliviando você de configurar o ambiente.

As imagens predefinidas no DevEnviron do ModelArts são:

- Pacotes predefinidos comuns: mecanismos de IA comuns, como PyTorch e MindSpore, baseados no Conda padrão, pacotes de software de análise de dados comuns, como Pandas e Numpy, e software de ferramenta comum, como CUDA e CUDNN, atendendo aos requisitos comuns de desenvolvimento de IA.
- Ambientes Conda predefinidos: um ambiente Conda e Conda Python básico (excluindo qualquer mecanismo de IA) são criados para cada imagem predefinida. A figura a seguir mostra o ambiente Conda para o MindSpore predefinido.

  Selecione um ambiente Conda com base em se o mecanismo de IA é usado para depuração.
- Notebook: uma aplicação da Web que permite codificar na GUI e combinar o código, as equações matemáticas e o conteúdo visualizado em um documento.
- Plug-ins do JupyterLab: permitem mudança de flavor, compartilhamento de casos na Galeria de IA para comunicação e interrupção de instâncias para melhorar a experiência do usuário.
- SSH remoto: permite depurar remotamente uma instância de notebook a partir de um PC local.
- Depois que as imagens predefinidas no DevEnviron de ModelArts suportam o desenvolvimento, os trabalhos de treinamento podem ser executados no ModelArts.

NOTA

- Para simplificar as operações, o notebook de ModelArts da nova versão não suporta alternância entre mecanismos de IA em uma instância de notebook.
- Os mecanismos de IA variam de acordo com as regiões. Para obter detalhes sobre os mecanismos de IA disponíveis em uma região, consulte os mecanismos de IA exibidos no console de gerenciamento.

### Recurso 3: JupyterLab, uma ferramenta de desenvolvimento e depuração interativa on-line

O ModelArts integra JupyterLab de código aberto para desenvolvimento e depuração interativos on-line. Você pode usar o notebook no console de gerenciamento do ModelArts para compilar e depurar código e treinar modelos baseados no código, sem necessidade de instalação ou configuração de ambiente.

O JupyterLab é um ambiente de desenvolvimento interativo. É o produto da próxima geração do Jupyter Notebook. O JupyterLab permite compilar notebooks, operar terminais, editar texto Markdown, ativar interação e visualizar arquivos e imagens CSV.

## 4.5 Treinamento de modelos

Além de dados e algoritmos, os desenvolvedores gastam muito tempo configurando os parâmetros de treinamento do modelo. Os parâmetros de treinamento do modelo determinam a precisão e o tempo de convergência do modelo. A seleção de parâmetros depende muito da experiência dos desenvolvedores. A seleção inadequada de parâmetros afetará a precisão do modelo ou aumentará significativamente o tempo necessário para o treinamento do modelo.

Para simplificar o desenvolvimento de IA e melhorar a eficiência do desenvolvimento e o desempenho do treinamento, o ModelArts oferece gerenciamento de tarefas visualizadas, gerenciamento de recursos e gerenciamento de versões e realiza otimização de hiperparâmetros com base em aprendizado de máquina e aprendizado por reforço. Ele fornece políticas automáticas de ajuste de hiperparâmetros, como taxa de aprendizado e tamanho de lote, e integra modelos comuns.

Atualmente, quando a maioria dos desenvolvedores constrói modelos, os modelos geralmente têm dezenas de camadas ou até centenas de camadas e parâmetros de nível MB ou GB para atender aos requisitos de precisão. Como resultado, as especificações dos recursos de computação são extremamente altas, especialmente o poder de computação dos recursos de hardware, memória e ROM. As especificações de recursos no lado do dispositivo são estritamente limitadas. Por exemplo, o poder de computação no lado do dispositivo é de 1 TFLOPS, o tamanho da memória é de cerca de 2 GB e o espaço da ROM é de cerca de 2 GB, então o tamanho do modelo no lado do dispositivo deve ser limitado a 100 KB e o atraso de inferência deve ser limitado a 100 milissegundos.

Portanto, tecnologias de compressão com precisão de modelo sem perdas ou quase sem perdas, como poda, quantização e destilação de conhecimento, são usadas para implementar compressão e otimização automática de modelos e iteração automática de compressão e reciclagem de modelos para controlar a perda de precisão do modelo. A tecnologia de quantização de bits baixos, que elimina a necessidade de reciclagem, converte o modelo de um ponto flutuante de alta precisão para uma operação de ponto fixo. Várias tecnologias de compactação e otimização são usadas para atender aos requisitos leves dos recursos de hardware de dispositivos e de borda. A tecnologia de compactação do modelo reduz a precisão em menos de 1% em cenários específicos.

Quando o volume de dados de treinamento é grande, o treinamento do modelo de aprendizado profundo é demorado. Em tecnologia de visão computacional, ImageNet-1k (um conjunto de dados de classificação contendo 1.000 classes de imagens, conhecido como ImageNet) é um conjunto de dados comumente usado. Se você usar uma GPU P100 para treinar um modelo ResNet-50 no conjunto de dados, levará quase uma semana. Isso dificulta o rápido desenvolvimento de aplicações de aprendizado profundo. Portanto, a aceleração do treinamento de aprendizado profundo sempre foi uma preocupação importante para a academia e a indústria.

A aceleração do treinamento distribuído precisa ser considerada em termos de software e hardware. Um único método de otimização não pode atender às expectativas. Portanto, a otimização da aceleração distribuída é um projeto de sistema. A arquitetura de treinamento distribuído precisa ser considerada em termos de design de hardware e chip. Para minimizar atrasos de computação e comunicação, muitos fatores precisam ser considerados, incluindo especificações gerais de computação, largura de banda da rede, cache de alta velocidade, consumo de energia e dissipação de calor do sistema e a relação entre a taxa de transferência de comunicação e computação.

O design do software precisa combinar recursos de hardware de alto desempenho para usar totalmente a rede de hardware de alta velocidade e implementar comunicação distribuída de alta largura de banda e cache de dados local eficiente. Usando algoritmos de otimização de treinamento, como paralelo híbrido, compressão de gradiente e aceleração de convolução, o software e o hardware do sistema de treinamento distribuído podem ser eficientemente coordenados e otimizados de ponta a ponta, e a aceleração de treinamento pode ser implementada em um ambiente distribuído de vários hosts e cartões. O ModelArts oferece uma aceleração líder do setor de mais de 0,8 para o ResNet50 no conjunto de dados ImageNet no ambiente distribuído com milhares de hosts e cartões.

Para medir o desempenho de aceleração do aprendizado profundo distribuído, os dois indicadores principais a seguir são usados:

- Taxa de transferência, ou seja, a quantidade de dados processados em uma unidade de tempo
- Tempo de convergência, ou seja, o tempo necessário para atingir certa precisão

A taxa de transferência depende do hardware do servidor (por exemplo, mais chips de aceleração de IA com maior capacidade de processamento FLOPS e maior largura de banda de comunicação alcançam maior taxa de transferência) leitura e cache de dados, pré-processamento de dados, computação de modelos (por exemplo, seleção de algoritmos de convolução) e otimização de topologia de comunicação. Exceto a computação de baixo bits e a compactação de gradiente (ou parâmetro), a maioria das tecnologias melhora a taxa de transferência sem afetar a precisão do modelo. Para alcançar o menor tempo de convergência, você precisa otimizar a taxa de transferência e ajustar os parâmetros. Se os parâmetros não forem ajustados corretamente, a taxa de transferência não poderá ser otimizada. Se o tamanho do lote for definido para um valor pequeno, o desempenho paralelo do treinamento do modelo será relativamente ruim. Como resultado, a taxa de transferência não pode ser melhorada mesmo que o número de nós de computação seja aumentado.

Os usuários estão mais preocupados com o tempo de convergência. A estrutura MoXing implementa otimização de pilha completa e reduz significativamente o tempo de convergência do treinamento. Para leitura e pré-processamento de dados, o MoXing usa pipelines de entrada simultânea de vários níveis para evitar que I/Os de dados se tornem um gargalo. Em termos de computação de modelos, o MoXing fornece cálculo de precisão híbrida, que combina semiprecisão e precisão única para os modelos de camada superior e reduz a perda causada pelo cálculo de precisão por meio de dimensionamento adaptativo. Políticas de

hiperparâmetro dinâmico (como momento e tamanho do lote) são usadas para minimizar o número de épocas necessárias para a convergência do modelo. MoXing também trabalha com servidores e bibliotecas de computação subjacentes da Huawei para melhorar ainda mais a aceleração distribuída.

### Otimização de treinamento distribuído de alto desempenho do ModelArts

- Precisão híbrida automática para utilizar totalmente os recursos de computação de hardware
- Tecnologias de ajuste de hiperparâmetro dinâmico (tamanho dinâmico do lote, tamanho da imagem e momento)
- Mesclagem e divisão automáticas do gradiente do modelo
- Otimização de agendamento do operador de comunicação com base na computação adaptativa de bolhas da BP
- Bibliotecas distribuídas de comunicação de alto desempenho (NStack e HCCL)
- Modelo de dados distribuídos paralelo híbrido
- Treinamento de compactação de dados e cache de vários níveis

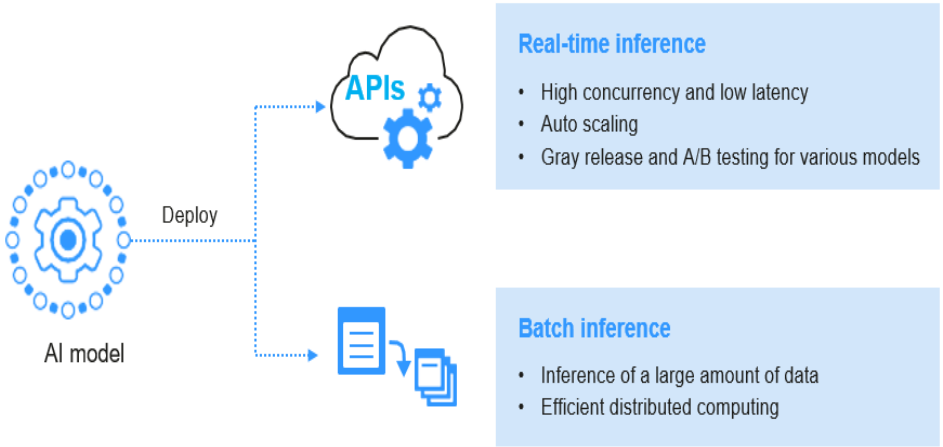
## 4.6 Implementação de modelos

O ModelArts é capaz de gerenciar modelos e serviços. Isso permite que as imagens e modelos de estrutura mainstream de vários fornecedores sejam gerenciados de maneira unificada.

Geralmente, a implantação do modelo de IA e a implementação em larga escala são complexas.

Por exemplo, em um projeto de transporte inteligente, o modelo treinado precisa ser implementado na nuvem, nas bordas e nos dispositivos. É preciso tempo e esforço para implementar o modelo nos dispositivos, por exemplo, implementar um modelo em câmeras de diferentes especificações e fornecedores. O ModelArts oferece suporte à implementação com um clique de um modelo treinado em vários dispositivos para diferentes cenários de aplicações. Além disso, ele fornece um conjunto de modos de implementação one-stop seguros e confiáveis para desenvolvedores individuais, empresas e fabricantes de dispositivos.

**Figura 4-2** Processo de implementação de um modelo



- O serviço de inferência em tempo real apresenta alta simultaneidade, baixa latência e dimensionamento elástico, e suporta versão cinza multimodelo e testes A/B.
- Os modelos podem ser implementados como serviços de inferência em tempo real e tarefas de inferência em lote na nuvem.

# 5 Serviços relacionados

## IAM

ModelArts usa o Identity and Access Management (IAM) para autenticação e autorização. Para obter mais informações sobre o IAM, consulte ***Guia de usuário do Identity and Access Management***.

## OBS

O ModelArts usa o Object Storage Service (OBS) para armazenar dados e modelos de forma segura e confiável a baixo custo. Para obter mais detalhes, consulte ***Guia de operação do console do Object Storage Service***.

**Tabela 5-1** Relação entre ModelArts e OBS

| Função | Subtarefa | Relação |
| --- | --- | --- |
| ExeML | Rotulagem de dados | Os dados rotulados no ModelArts são armazenados no OBS. |
| | Treinamento automático | Depois que um trabalho de treinamento é concluído, o modelo gerado é armazenado no OBS. |
| | Implementação de modelo | O ModelArts implementa modelos armazenados no OBS como serviços em tempo real. |
| Ciclo de vida do desenvolvimento de IA | Gerenciamento de dados | ● Os conjuntos de dados são armazenados no OBS.<br>● As informações de rotulagem do conjunto de dados são armazenadas no OBS.<br>● Os dados podem ser importados do OBS. |
| | Ambiente de desenvolvimento | Os dados ou arquivos de código em uma instância de bloco de anotações são armazenados no OBS. |

| Função | Subtarefa | Relação |
| --- | --- | --- |
| | Treinamento de modelo | ● Os conjuntos de dados usados pelos trabalhos de treinamento são armazenados no OBS.<br>● Os scripts em execução para trabalhos de treinamento são armazenados no OBS.<br>● Os modelos gerados pelos trabalhos de treinamento são armazenados nos caminhos especificados do OBS.<br>● Os logs de execução dos trabalhos de treinamento são armazenados nos caminhos do OBS especificados. |
| | Gerenciamento de aplicações de IA | Depois que um trabalho de treinamento é concluído, o modelo gerado é armazenado no OBS. Você pode importar o modelo do OBS. |
| | Implementação de serviços | Os modelos armazenados no OBS podem ser implementados como serviços. |
| Configurações | - | Autoriza o ModelArts a acessar o OBS (usando uma agência ou chave de acesso) para que o ModelArts possa usar o OBS para armazenar dados e criar instâncias de notebook. |

## EVS

O ModelArts usa o Elastic Volume Service (EVS) para armazenar instâncias de notebook criadas. Para obter mais detalhes, consulte ***Guia de usuário do Elastic Volume Service***.

## CCE

O ModelArts usa o Cloud Container Engine (CCE) para implementar modelos como serviços em tempo real. O CCE permite alta simultaneidade e proporciona dimensionamento elástico. Para obter mais informações sobre o CCE, consulte ***Guia de usuário do Cloud Container Engine***.

## SWR

Use o Software Repository for Container (SWR) para personalizar uma imagem e importá-la ao ModelArts para treinamento ou inferência. Para obter detalhes sobre o SWR, consulte ***Guia de usuário do Software Repository for Container***.

## Cloud Eye

O ModelArts monitoriza serviços online e carregamentos de modelos em tempo real e envia alarmes e notificações automaticamente. Para obter detalhes sobre o Cloud Eye, consulte ***Guia de usuário do Cloud Eye***.

# 6 Como acessar o ModelArts?

Você pode acessar o ModelArts por meio do console de gerenciamento baseado na Web ou usando interfaces de programação de aplicações (APIs) baseadas em HTTPS.

- **Usar o console de gerenciamento**

  O ModelArts possui um console de gerenciamento simples e fácil de usar e oferece uma série de funções, incluindo ExeML, gerenciamento de dados, ambiente de desenvolvimento, treinamento de modelos, gerenciamento de aplicativos de IA, Galeria de IA e implementação de serviços. Você pode concluir o desenvolvimento de IA de ponta a ponta no console de gerenciamento.

  Para usar o console de gerenciamento do ModelArts, você precisa se registrar na HUAWEI CLOUD primeiro. Se você criou uma conta da Huawei Cloud, escolha **AI** > **ModelArts** no site oficial e faça logon no console de gerenciamento.

- **Usar SDKs**

  Se quiser integrar o ModelArts a um sistema de terceiros para desenvolvimento secundário, chame SDKs para concluir o desenvolvimento. Os SDKs do ModelArts encapsulam APIs RESTful fornecidas pelo ModelArts para simplificar o desenvolvimento secundário. Para obter detalhes sobre os SDKs e as operações, consulte ***Referência de SDK do ModelArts***.

  Além disso, você pode chamar diretamente os SDKs do ModelArts ao escrever código em um notebook no console de gerenciamento.

- **Usar APIs**

  Para acessar o ModelArts, use APIs para integrar o ModelArts a um sistema de terceiros. Para obter detalhes sobre as APIs e operações, consulte ***Referência de API do ModelArts***.

# 7 Gerenciamento de permissões

O ModelArts permite configurar permissões refinadas para o gerenciamento refinado de recursos e permissões. Isso é comumente usado por grandes empresas, mas é complexo para usuários individuais. Recomenda-se que usuários individuais configurem permissões para usar ModelArts consultando **Atribuição de permissões a usuários individuais para usar ModelArts**.

NOTA

Se você atender a qualquer uma das seguintes condições, leia este documento.

- Você é um usuário corporativo e
  - Existem vários departamentos em sua empresa e você precisa controlar as permissões dos usuários para que os usuários em diferentes departamentos possam acessar apenas seus recursos e funções dedicados.
  - Existem várias funções (como administradores, desenvolvedores de algoritmos e pessoal de O&M de aplicações) em sua empresa. Você precisa deles para usar apenas funções específicas.
  - Há logicamente vários ambientes (como o ambiente de desenvolvimento, ambiente de pré-produção e ambiente de produção) e estão isolados uns dos outros. Você precisa controlar as permissões dos usuários em diferentes ambientes.
  - Você precisa controlar as permissões de usuários ou grupos de usuários específicos do IAM.
- Você é um usuário individual e criou vários usuários do IAM. Você precisa atribuir diferentes permissões do ModelArts a diferentes usuários do IAM.
- Você precisa entender os conceitos e operações do gerenciamento de permissões do ModelArts.

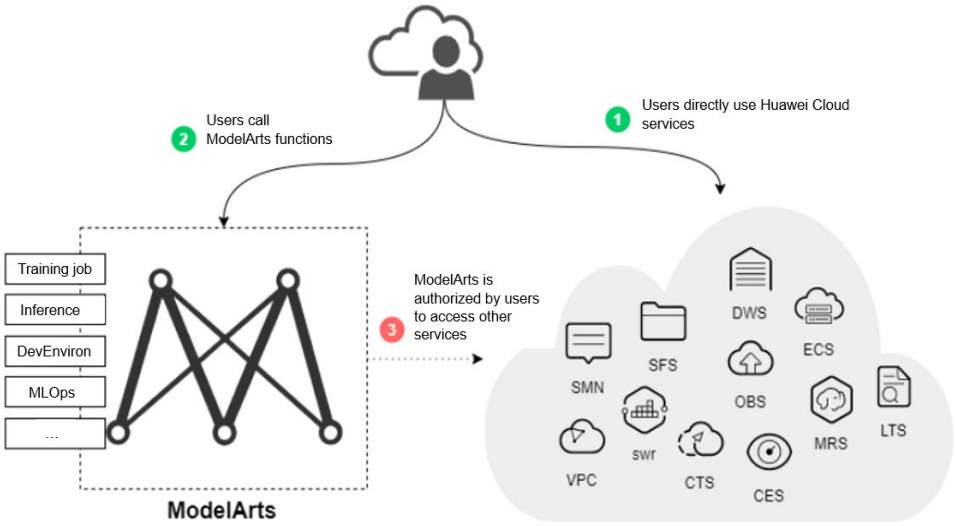
O ModelArts usa o Identity and Access Management (IAM) para a maioria das funções de gerenciamento de permissões. Antes de ler abaixo, aprenda sobre **Conceitos Básicos**. Isso ajuda você a entender melhor este documento.

Para implementar o gerenciamento refinado de permissões, o ModelArts fornece controle de permissões, autorização de agência e espaço de trabalho. A seguir descreve os detalhes.

## Permissões e agências do ModelArts

**Figura 7-1** Gerenciamento de permissões




As funções expostas do ModelArts são controladas por meio de permissões do IAM. Por exemplo, se você, como usuário do IAM, precisa criar uma tarefa de treinamento no ModelArts, deve ter a permissão **modelarts:trainJob:create**. Para obter detalhes sobre como atribuir permissões a um usuário (você precisa adicionar o usuário a um grupo de usuários e, em seguida, atribuir permissões ao grupo de usuários), consulte **Gerenciamento de permissões**.

O ModelArts deve acessar outros serviços para computação de IA. Por exemplo, o ModelArts deve acessar o OBS para ler seus dados para treinamento. Para fins de segurança, o ModelArts deve estar autorizada a acessar outros serviços em nuvem. Esta é a autorização da agência.

O seguinte resume o gerenciamento de permissões:

- Seu acesso a qualquer serviço de nuvem é controlado por meio do IAM. Você deve ter as permissões do serviço de nuvem. (As permissões de serviço necessárias variam de acordo com as funções que você usa.)
- Para usar as funções do ModelArts, você precisa conceder permissões por meio do IAM.
- O ModelArts deve ser autorizado por você para acessar outros serviços em nuvem para computação de IA.

## Gerenciamento de permissões do ModelArts

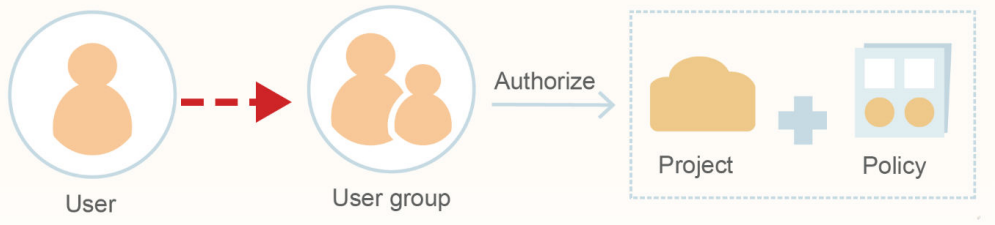
Por padrão, os novos usuários do IAM não têm nenhuma permissão atribuída. Você precisa adicionar o usuário a um grupo de usuários e conceder políticas ao grupo de usuários, para que os usuários do grupo possam herdar as permissões. Após a autorização, os usuários podem executar operações no ModelArts com base em permissões.

**⚠ CUIDADO**

O ModelArts é um serviço de nível de projeto implementado e acessado em regiões físicas específicas. Ao autorizar uma agência, você pode definir o escopo das permissões selecionadas para todos os recursos, projetos empresariais ou projetos específicos da região. Se você especificar projetos específicos da região, as permissões selecionadas serão aplicadas aos recursos nesses projetos.

Para obter detalhes, consulte **Criação de um grupo de usuários e atribuição de permissões**.




Ao atribuir permissões a um grupo de usuários, o IAM não atribui permissões específicas diretamente ao grupo de usuários. Em vez disso, o IAM precisa adicionar as permissões a uma política e, em seguida, atribuir a política ao grupo de usuários. Para facilitar o gerenciamento de permissões de usuário, cada serviço de nuvem fornece algumas políticas predefinidas para você usar diretamente. Se as políticas predefinidas não puderem atender aos seus requisitos de gerenciamento de permissões refinado, você poderá personalizar as políticas.

**Tabela 7-1** lista todas as políticas predefinidas definidas pelo sistema suportadas pelo ModelArts.

**Tabela 7-1** Políticas definidas pelo sistema suportadas pelo ModelArts

| Política | Descrição | Tipo |
|---|---|---|
| ModelArts FullAccess | Permissões de administrador para ModelArts. Os usuários que recebem essas permissões podem operar e usar o ModelArts. | Política definida pelo sistema |
| ModelArts CommonOperations | Permissões comuns de usuário para ModelArts. Os usuários com essas permissões podem operar e usar o ModelArts, mas não podem gerenciar pools de recursos dedicados. | Política definida pelo sistema |
| ModelArts Dependency Access | Permissões em serviços dependentes para ModelArts | Política definida pelo sistema |

Geralmente, ModelArts FullAccess é atribuída apenas aos administradores. Se o gerenciamento refinado não for necessário, atribuir ModelArts CommonOperations a todos os usuários atenderá aos requisitos de desenvolvimento da maioria das equipes pequenas. Se você quiser personalizar políticas para o gerenciamento de permissões refinado, consulte **IAM**.

NOTA

Quando você atribui permissões do ModelArts a um usuário, o sistema não atribui automaticamente as permissões de outros serviços ao usuário. Isso garante a segurança e evita operações inesperadas não autorizadas. Nesse caso, no entanto, você deve atribuir permissões de diferentes serviços aos usuários para que eles possam executar algumas operações do ModelArts.

Por exemplo, se um usuário do IAM precisar usar dados do OBS para treinamento e a permissão de treinamento do ModelArts tiver sido configurada para o usuário do IAM, o usuário do IAM ainda precisará receber as permissões de leitura, gravação e lista do OBS. A permissão de lista do OBS permite que você selecione o caminho de dados de treinamento no ModelArts. A permissão de leitura é usada para visualizar dados e ler dados para treinamento. A permissão de gravação é usada para salvar resultados de treinamento e logs.

- Para usuários individuais ou pequenas organizações, é uma boa prática configurar a política **Tenant Administrator** que se aplica a serviços globais para usuários do IAM. Dessa forma, os usuários do IAM podem obter todas as permissões de usuário, exceto o IAM. No entanto, isso pode causar problemas de segurança. (Para um usuário individual, o usuário padrão do IAM pertence ao grupo de usuários **admin** e tem a permissão **Tenant Administrator**.)
- Se você quiser restringir as operações do usuário, configure as permissões mínimas do OBS para usuários do ModelArts. Para obter detalhes, consulte **Gerenciamento de permissões do OBS**. Para obter detalhes sobre o gerenciamento refinado de permissões de outros serviços de nuvem, consulte os documentos de serviço de nuvem correspondentes.

## Autorização da agência do ModelArts

O ModelArts deve ser autorizado pelos usuários a acessar outros serviços em nuvem para computação de IA. No sistema de permissões do IAM, essa autorização é realizada por meio de agências.

Para obter detalhes sobre os conceitos básicos e as operações das agências, consulte **Delegação de serviço de nuvem**.

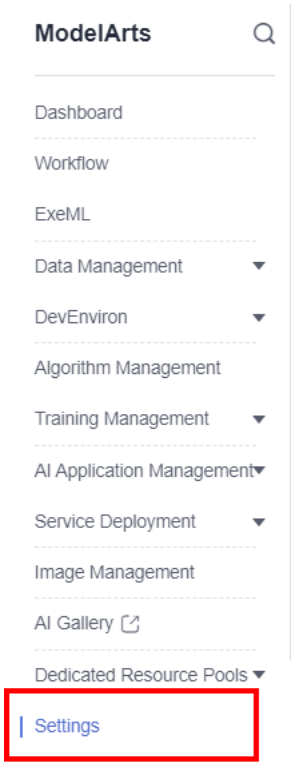
Para simplificar a autorização de agência, o ModelArts oferece suporte à configuração automática de autorização de agência. Você só precisa configurar uma agência para você ou usuários especificados na página **Global Configuration** do console do ModelArts.

NOTA

- Somente os usuários com a permissão de gerenciamento de agência do IAM podem executar essa operação. Geralmente, os membros do grupo de usuários administradores do IAM têm essa permissão.
- A autorização da agência do ModelArts é específica da região, o que significa que você deve executar a autorização da agência em cada região que você usa.

**Figura 7-2** Configurações

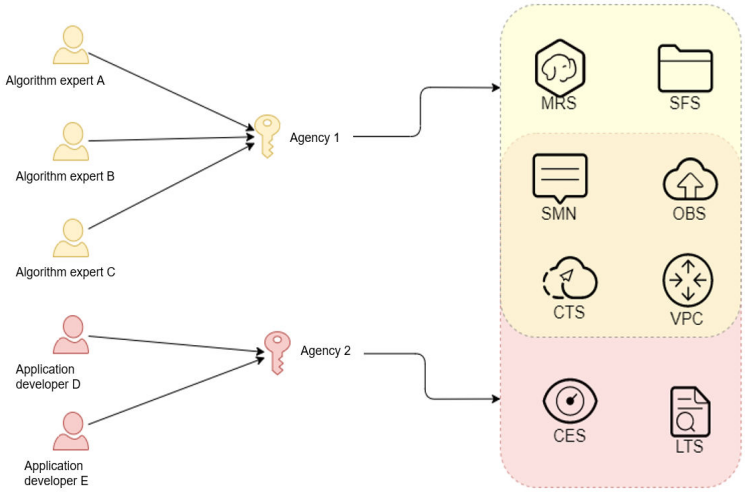



Na página **Global Configuration** do console do ModelArts, após clicar em **Add Authorization**, você pode configurar uma agência para um usuário específico ou todos os usuários. Geralmente, uma agência chamada **modelarts_agency_<*Username*>_*Random ID*** é criada por padrão. Na área **Permissions**, você pode selecionar a configuração de permissão predefinida ou selecionar as políticas necessárias. Se as duas opções não atenderem aos seus requisitos, crie uma agência na página de gerenciamento do IAM (é necessário delegar ModelArts para acessar seus recursos) e use uma agência existente em vez de adicionar uma agência na página **Add Authorization**.

O ModelArts associa vários usuários a uma agência. Isso significa que, se dois usuários precisarem configurar a mesma agência, não será necessário criar uma agência para cada usuário. Em vez disso, você só precisa configurar a mesma agência para os dois usuários.

**Figura 7-3** Mapeamento entre usuários e agências




NOTA

Cada usuário pode usar o ModelArts somente depois de estar associado a uma agência. No entanto, mesmo que as permissões atribuídas à agência sejam insuficientes, nenhum erro é relatado quando a API é chamada. Um erro ocorre apenas quando o sistema utiliza funções não autorizadas. Por exemplo, você habilita a notificação de mensagem ao criar um trabalho de treinamento. A notificação de mensagem requer autorização de SMN. No entanto, um erro ocorre somente quando as mensagens precisam ser enviadas para o trabalho de treinamento. O sistema ignora alguns erros e outros erros podem causar falhas de trabalho. Ao implementar a minimização de permissões, verifique se você ainda terá permissões suficientes para as operações necessárias no ModelArts.

## Autorização estrita

No modo de autorização estrita, é necessária autorização explícita do administrador da conta para que os usuários do IAM acessem o ModelArts. O administrador pode adicionar as permissões necessárias do ModelArts para usuários comuns por meio de políticas de autorização.

No modo de autorização não restrita, os usuários do IAM podem usar o ModelArts sem autorização explícita. O administrador precisa configurar a política de negação para usuários do IAM para impedi-los de usar algumas funções do ModelArts.

O administrador pode alterar o modo de autorização na página **Global Configuration**.

AVISO

Recomenda-se o modo de autorização estrita. Nesse modo, os usuários do IAM devem estar autorizados a usar as funções do ModelArts. Dessa forma, o escopo de permissão dos usuários do IAM pode ser controlado com precisão, minimizando as permissões concedidas aos usuários do IAM.

## Gerenciar o acesso a recursos usando espaços de trabalho

O espaço de trabalho permite que os clientes corporativos dividam seus recursos em vários espaços logicamente isolados e gerenciem o acesso a diferentes espaços. Como um usuário

corporativo, você pode enviar a solicitação para ativar a função do espaço de trabalho ao seu gerente de suporte técnico.

Depois que o espaço de trabalho é ativado, um espaço de trabalho padrão é criado. Todos os recursos que você criou estão neste espaço de trabalho. Um espaço de trabalho é como um gêmeo do ModelArts. Você pode alternar entre áreas de trabalho no canto superior esquerdo do console do ModelArts. Trabalhos em espaços de trabalho diferentes não se afetam mutuamente.

Ao criar um espaço de trabalho, você deve vinculá-lo a um projeto corporativo. Vários espaços de trabalho podem ser vinculados ao mesmo projeto da empresa, mas um espaço de trabalho não pode ser vinculado a vários projetos corporativo. Você pode usar espaços de trabalho para restrições refinadas no acesso a recursos e permissões de diferentes usuários. As restrições são as seguintes:

- Os usuários devem ser autorizados a acessar espaços de trabalho específicos (isso deve ser configurado nas páginas para criar e gerenciar espaços de trabalho). Isso significa que o acesso a ativos de IA, como conjuntos de dados e algoritmos, pode ser gerenciado usando espaços de trabalho.
- Nas operações de autorização de permissão anteriores, se você definir o escopo para projetos da empresa, a autorização terá efeito somente para espaços de trabalho vinculados aos projetos selecionados.

NOTA

- Restrições em espaços de trabalho e autorização de permissão entram em vigor ao mesmo tempo. Ou seja, um usuário deve ter a permissão para acessar o espaço de trabalho e a permissão para criar trabalhos de treinamento (a permissão se aplica a esse espaço de trabalho) para que o usuário possa enviar trabalhos de treinamento nesse espaço de trabalho.
- Se você ativou um projeto corporativo, mas não ativou um espaço de trabalho, todas as operações serão executadas no projeto da empresa padrão. Certifique-se de que as permissões nas operações necessárias se aplicam ao projeto corporativo padrão.
- As restrições anteriores não se aplicam a usuários que não ativaram nenhum projeto corporativo.

## Resumo

Principais recursos do gerenciamento de permissões do ModelArts:

- Se você for um usuário individual, não precisa considerar o gerenciamento de permissões refinado. Sua conta tem todas as permissões para usar o ModelArts por padrão.
- Todas as funções do ModelArts são controladas pelo IAM. Você pode usar a autorização do IAM para implementar o gerenciamento de permissões refinado para usuários específicos.
- Todos os usuários (incluindo usuários individuais) podem usar funções específicas somente após a autorização da agência no ModelArts (**Settings** > **Add Authorization**). Caso contrário, podem ocorrer erros inesperados.
- Se você ativou a função de projeto corporativo, também pode ativar o espaço de trabalho do ModelArts e usar a autorização básica e o espaço de trabalho para o gerenciamento de permissões refinado.

# 8 Segurança

## 8.1 Responsabilidades compartilhadas

Huawei garante que seu compromisso com a segurança cibernética nunca será superado pela consideração de interesses comerciais. Para lidar com os desafios emergentes de segurança na nuvem e ameaças e ataques à segurança na nuvem, a Huawei Cloud constrói um sistema abrangente de garantia de segurança de serviços em nuvem para diferentes regiões e indústrias com base nas vantagens exclusivas de software e hardware, leis, regulamentos, padrões da indústria e ecossistema de segurança da Huawei.

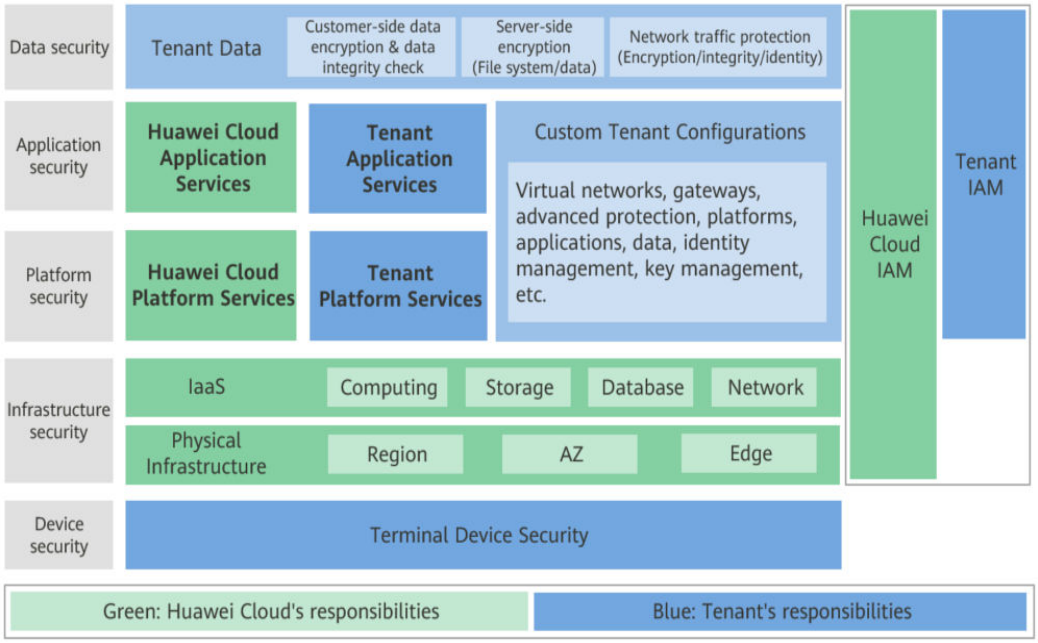
**Figura 8-1** ilustra as responsabilidades partilhadas pela Huawei Cloud e pelos usuários.

- **Huawei Cloud**: garanta a segurança dos serviços de nuvem e fornecer nuvens seguras. As responsabilidades de segurança da Huawei Cloud incluem garantir a segurança de nossos serviços de IaaS, PaaS e SaaS, bem como os ambientes físicos dos data centers da Huawei Cloud onde nossos serviços de IaaS, PaaS e SaaS operam. A Huawei Cloud é responsável não apenas pelas funções de segurança e pelo desempenho de nossa infraestrutura, serviços de nuvem e tecnologias, mas também pela segurança geral de O&M na nuvem e, no sentido mais amplo, pela certificação de segurança de nossa infraestrutura e serviços.
- **Locatário**: use a nuvem com segurança. Os locatários da Huawei Cloud são responsáveis pelo gerenciamento seguro e eficaz das configurações personalizadas dos serviços em nuvem, incluindo IaaS, PaaS e SaaS. Isso inclui, mas não se limita a redes virtuais, o SO de hosts e convidados de máquinas virtuais, firewalls virtuais, API Gateway, serviços avançados de segurança, todos os tipos de serviços em nuvem, dados de locatários, contas de identidade e gerenciamento de chaves.

**O livro branco de segurança da Huawei Cloud** elabora as ideias e medidas para a construção da segurança da Huawei Cloud, incluindo estratégias de segurança na nuvem, o modelo de responsabilidade compartilhada, conformidade e privacidade, organizações e pessoal de segurança, segurança de infraestrutura, serviço e segurança de locatários, segurança de engenharia, segurança de O&M e segurança do ecossistema.

**Figura 8-1** Modelo de responsabilidade de segurança compartilhada da Huawei Cloud




# 8.2 Identificação e gerenciamento de ativos

## Identificação de ativos

Seus ativos na Galeria de IA incluem seus ativos de IA publicados e suas informações pessoais.

Os ativos de IA incluem, mas não estão limitados a textos, gráficos, dados, artigos, fotos, imagens, ilustrações, código, algoritmos de IA e modelos de IA.

Suas informações pessoais incluem:

- Apelido, foto de perfil e e-mail para registro da conta
- Nome, número de celular e e-mail para participar das práticas
- Informações corporativas para se tornar um parceiro
- Nome do contato, número de celular e e-mail para publicação de ativos

## Gerenciamento de ativos

A Galeria de IA gerencia de forma centralizada os ativos publicados pelos usuários.

- A Galeria de IA armazena ativos de arquivos em baldes oficiais do OBS.
- A Galeria de IA armazena ativos de imagem em repositórios oficiais do SWR.

A Galeria de IA armazena informações pessoais dos usuários em bancos de dados. A Galeria de IA criptografa informações pessoais confidenciais, como números de celular e e-mails, em bancos de dados.

Para obter mais informações sobre Galeria de IA, consulte **Galeria de IA**.

# 8.3 Autenticação de identidade e controle de acesso

## Autenticação de identificação

Você pode usar os serviços do ModelArts por meio do console, das APIs ou dos SDKs. Essencialmente, as solicitações de acesso são enviadas por meio de APIs REST do ModelArts.

As APIs do ModelArts podem ser acessadas após a autenticação bem-sucedida. As solicitações enviadas pelo console podem ser autenticadas usando tokens, e as solicitações de chamadas de APIs podem ser autenticadas usando tokens ou AK/SK. Para detalhes, veja **Autenticação**.

## Controle de acesso

O ModelArts permite configurar permissões refinadas para o gerenciamento refinado de recursos e permissões. Para fazer isso, o ModelArts fornece controle de permissão do IAM, autorização de agência e espaço de trabalho.

- Controle de permissão de IAM

  Para usar as funções do ModelArts, você precisa conceder permissões por meio do IAM. Por exemplo, se você precisar criar um trabalho de treinamento no ModelArts, deve ter a permissão **modelarts:trainJob:create**.

  Se nenhuma diretiva de autorização refinada estiver configurada para um usuário criado pelo administrador, o usuário terá todas as permissões do ModelArts por padrão. Para controlar as permissões do usuário, o administrador precisa adicionar o usuário a um grupo de usuários no IAM e configurar políticas de autorização refinadas para o grupo de usuários. Desta forma, o usuário obtém as permissões definidas nas políticas antes de executar operações em recursos de serviço de nuvem. Durante a autorização baseada em política, o administrador pode selecionar o escopo de autorização com base nos tipos de recurso do ModelArts. Para obter detalhes sobre permissões de recursos, consulte **Políticas de permissões e ações suportadas**.

- Autorização da agência

  O ModelArts precisa acessar outros serviços para computação de IA. Por exemplo, o ModelArts precisa acessar o OBS para ler seus dados para treinamento. Para fins de segurança, o ModelArts deve estar autorizada a acessar outros serviços em nuvem. Esta é a autorização da agência.

  O ModelArts não salva suas credenciais de autenticação de token. Antes de executar operações em seus recursos (como buckets do OBS) em um trabalho de back-end, é necessário autorizar explicitamente o ModelArts por meio de uma agência do IAM. O ModelArts usará a agência para obter uma credencial de autenticação temporária para executar operações em seus recursos. Para obter detalhes, consulte **Configuração de autorização de acesso (configuração global)**.

- Espaço de trabalho

  O espaço de trabalho permite que os clientes que habilitaram **projetos empresariais** para dividir seus recursos em vários espaços logicamente isolados e controlar o acesso a diferentes espaços.

  Depois que o espaço de trabalho é ativado, um espaço de trabalho padrão é criado. Todos os recursos que você criou estão neste espaço de trabalho. Um espaço de trabalho é como

um gêmeo do ModelArts. Você pode alternar entre espaços de trabalho no canto superior esquerdo do painel de navegação. Trabalhos em espaços de trabalho diferentes não se afetam mutuamente. O ModelArts permite criar vários espaços de trabalho para desenvolver algoritmos e gerenciar e implementar modelos para diferentes objetivos de serviço. Desta forma, as saídas de desenvolvimento de diferentes aplicativos são gerenciadas em diferentes espaços de trabalho para uso.

### Gerenciamento de acesso remoto

Quando você usa um IDE local para acessar remotamente o ambiente de desenvolvimento do notebook do ModelArts por SSH, o par de chaves é necessário para autenticação. Você também pode adicionar os endereços IP para acessar remotamente a instância de notebook à lista branca.

## 8.4 Proteção de dados

O ModelArts toma medidas diferentes para manter os dados armazenados no ModelArts seguros e confiáveis.

| Medida | Descrição |
| --- | --- |
| Proteção estática de dados | A AI Gallery criptografa informações pessoais confidenciais, como números de celular e e-mails, em bancos de dados. O algoritmo de criptografia AES é usado. |
| Proteção na transmissão de dados | Ao importar aplicações de IA no ModelArts, ele suporta HTTP e HTTPS, mas o HTTPS é recomendado para transmissão de dados mais segura. |
| Verificação da integridade dos dados | Quando você carrega arquivos de modelo ou ativos da AI Gallery para implementação de inferência, os dados podem se tornar inconsistentes devido ao sequestro de rede, cache e outros motivos. O ModelArts verifica a consistência dos dados calculando o valor SHA256 quando os dados são carregados ou baixados. |
| Mecanismo de isolamento de dados | Quando uma instância de notebook é criada, o armazenamento de dados de locatários diferentes é isolado, para que locatários diferentes não podem ver dados de outros locatários. |

## 8.5 Auditoria e registro em logs

### Auditoria

Cloud Trace Service (CTS) registra as operações nos recursos em nuvem em sua conta. Você pode usar os logs gerados pelo CTS para realizar análises de segurança, rastrear alterações de recursos, auditar a conformidade e localizar falhas.

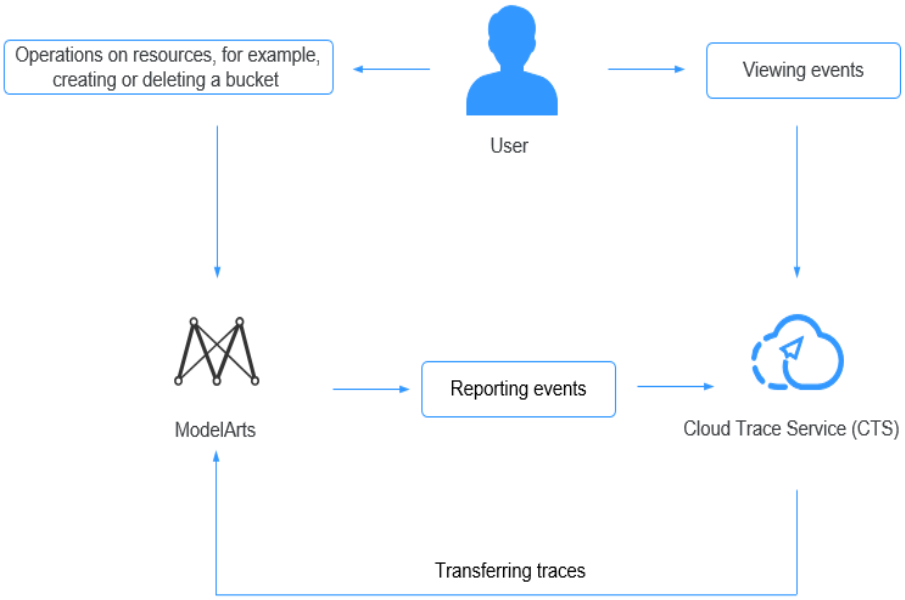
Depois de habilitar o CTS e configurar um rastreador, o CTS pode registrar gerenciamento e rastros de dados do ModelArts para auditoria.

Para obter detalhes sobre como habilitar e configurar o CTS, consulte **Habilitação do CTS**.

Para obter detalhes sobre o gerenciamento do ModelArts e rastros de dados que podem ser rastreados pelo CTS, consulte **Principais operações registradas para gerenciamento de dados**, **Principais operações de DevEnviron registradas pelo CTS**, **Principais operações de trabalho de treinamento registradas pelo CTS**, **Principais operações de gerenciamento de aplicações de IA registradas pelo CTS** e **Principais operações de gerenciamento de serviço registradas pelo CTS**.

**Figura 8-2** CTS




## Principais operações de gerenciamento de dados registradas pelo CTS

**Tabela 8-1** Principais operações de gerenciamento de dados registradas pelo CTS

| Operação | Tipo de recurso | Rastreamento |
| --- | --- | --- |
| Criar um conjunto de dados | Conjunto de dados | 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 |

| Operação | Tipo de recurso | Rastreamento |
| --- | --- | --- |
| Criar uma tarefa de rotulagem automática | dataset | createAutoLabelingTask |
| 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 |
| Modificar 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 as amostras correspondentes | dataset | deleteLabelWithSamples |
| Adicionar amostras | dataset | uploadSamples |
| Excluir amostras | dataset | deleteSamples |
| Interromper uma tarefa de rotulagem automática | dataset | stopTask |
| Criar um trabalho de rotulagem de equipe | dataset | createWorkforceTask |
| Excluir um trabalho de rotulagem de equipe | dataset | deleteWorkforceTask |
| Iniciar a aceitação da rotulagem da equipe | dataset | startWorkforceSampling-Task |
| Aprovar/rejeitar/cancelar a aceitação | dataset | updateWorkforceSam-plingTask |
| Enviar comentários de revisão de amostra para aceitação | dataset | acceptSamples |
| Adicionar um rótulo a uma amostra | dataset | updateSamples |
| Enviar um e-mail para rotular membros da equipe | dataset | sendEmails |
| Iniciar o trabalho de rotulagem da equipe como a pessoa de contato | dataset | startWorkforceTask |
| Atualizar um trabalho de rotulagem de equipe | dataset | updateWorkforceTask |

| Operação | Tipo de recurso | Rastreamento |
| --- | --- | --- |
| 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 um membro da equipe de rotulagem | workforce | updateWorker |
| Excluir um membro da equipe de rotulagem | workforce | deleteWorker |
| Excluir membros da equipe de rotulagem em 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 da equipe de rotulagem | labelConsoleWorker | workerLoginLabelConsole |
| Efetuar logout do console de rotulagem como um membro da equipe de rotulagem | labelConsoleWorker | workerLogOutLabelConsole |
| Alterar a senha do console de rotulagem como um membro da equipe de rotulagem | labelConsoleWorker | workerChangePassword |
| Esquecer a senha do console de rotulagem como membro da equipe de rotulagem | labelConsoleWorker | workerForgetPassword |
| Redefinir a senha do console de rotulagem por meio do URL como um membro da equipe de rotulagem | labelConsoleWorker | workerResetPassword |

## Principais operações de DevEnviron registradas pelo CTS

**Tabela 8-2** Principais operações do DevEnviron registradas pelo CTS

| Operação | Tipo de recurso | Nome do rastreamento |
|---|---|---|
| Criar uma instância de notebook | Notebook | createNotebook |
| Excluir uma instância de bloco 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 registradas pelo CTS

**Tabela 8-3** Principais operações de trabalho de treinamento registradas pelo CTS

| Operação | Tipo de recurso | Rastreamento |
|---|---|---|
| Criar um trabalho de treinamento | ModelArtsTrainJob | createModelArtsTrainJob |
| 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 |
| Criar uma configuração de trabalho de treinamento | ModelArtsTrainConfig | createModelArtsTrainConfig |

| Operação | Tipo de recurso | Rastreamento |
| --- | --- | --- |
| 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 |
| Deletar 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 registradas pelo CTS

**Tabela 8-4** Principais operações de gerenciamento de aplicações de IA registradas pelo CTS

| Operação | Tipo de recurso | Rastreamento |
| --- | --- | --- |
| Criar uma aplicação de AI | model | addModel |
| Atualizar uma aplicação de AI | model | updateModel |
| Excluir uma aplicação de AI | model | deleteModel |
| 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 serviço registradas pelo CTS

**Tabela 8-5** Principais operações de gerenciamento de serviços registradas pelo CTS

| Operação | Tipo de recurso | Rastreamento |
| --- | --- | --- |
| Implementar um serviço | service | addService |

| Operação | Tipo de recurso | Rastreamento |
| --- | --- | --- |
| Excluir um serviço | service | deleteService |
| Atualizar um serviço | service | updateService |
| Iniciar ou parar um serviço | service | startOrStopService |
| Adicionar uma chave de acesso | service | addAkSk |
| Excluir uma chave de acesso | 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 do pool de recursos dedicados | cluster | createClusterResult |

## Principais operações de Galeria de IA gravadas pelo CTS

**Tabela 8-6** Principais operações da Galeria de IA gravadas 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 uma atividade | ModelArts_Market | regist_activity |
| Modificar informações de usuário | ModelArts_Market | update_user |

### Registro em logs

Você pode ativar o log do ModelArts para análise ou auditoria. Depois que o CTS é habilitado, o CTS começa a gravar operações no ModelArts. O console de gerenciamento CTS armazena os últimos sete dias de registros de operação. Esta seção descreve como exibir os registros de operação dos últimos 7 dias no console do gerenciamento do CTS.

Para obter detalhes sobre como exibir logs de auditoria no CTS, consulte **Exibição de logs de auditoria**.

## 8.6 Resiliência de serviço

Resiliência refere-se à resiliência de segurança dos serviços em nuvem após ataques, excluindo confiabilidade e disponibilidade. Este capítulo descreve os recursos de ModelArts de defesa e detecção contra invasões, defesa contra jitter, uso adequado de nomes de domínio e detecção de segurança de conteúdo.

### Suítes de segurança e Cloud Bastion Host para defesa e detecção aprimoradas contra invasões

Suítes de segurança foram implementados no ModelArts no host, aplicação, rede e camadas de dados para detectar invasões prontamente.

- O ModelArts usa componentes seguros para prevenir riscos de aplicações da Web implementados e usa o WAF para proteção de segurança.
- Os produtos do Host Security Service (HSS) foram implementados em todos os hosts que transportam serviços do ModelArts. Esses produtos incluem, entre outros, HSS e Compute Security Platform (CSP) desenvolvidos pela Huawei.
- O Vulnerability Scan Service (VSS) foi implementado no ModelArts e executa varredura de rotina para detectar e corrigir vulnerabilidades rapidamente.
- O ModelArts realiza O&M de segurança em recursos de nuvem por meio de uma plataforma de gerenciamento de segurança.
- Situation Awareness (SA) foi implementado no ModelArts para entender a situação de segurança, consultar históricos de ataques e detectar rapidamente riscos de conformidade e responder a alarmes de ameaças.
- Anti-DDoS avançado (AAD) foi implementado nos EIPs que carregam os principais serviços do ModelArts para evitar tempestades de tráfego.
- O Database Security Service (DBSS) foi implementado em bancos de dados do ModelArts que armazenam dados importantes.

### Políticas de prevenção de jitter, resposta de emergência e restauração contra ataques

ModelArts isola recursos de diferentes locatários, para que os ataques aos recursos de um locatário não afetem os recursos dos outros.

- O ModelArts fornece pools de recursos dedicados fisicamente isolados, para que os ataques aos recursos de um locatário não afetem os recursos de outros.
- O ModelArts define e mantém suas especificações de desempenho para defender ataques, por exemplo, configurando o controle de tráfego no acesso à API.

- O ModelArts fornece relatórios de alarme e autoproteção contra ataques.
- O ModelArts detecta o comportamento anormal do serviço, por exemplo, detectando dados da plataforma de operações anormais e integrando registros de segurança.
- O ModelArts fornece controle de risco e resposta de emergência contra ataques. Por exemplo, o ModelArts identifica rapidamente locatários mal-intencionados e endereços IP mal-intencionados.
- O ModelArts restaura rapidamente os serviços depois que os ataques de tráfego param.

### Especificações de uso de nome de domínio e políticas de segurança de conteúdo de locatário dos serviços de nuvem

Os nomes de domínio do ModelArts cumprem determinados requisitos de segurança para evitar riscos de conformidade e ataques de phishing.

Nomes de domínio visíveis aos locatários: nomes de domínio acessíveis aos locatários, que exigem mais atenção à segurança e conformidade.

Nomes de domínio invisíveis para os locatários: nomes de domínio usados pelos serviços da Huawei Cloud para ligar uns aos outros na intranet, caso em que os usuários externos não conseguem acessar os servidores DNS autoritários; ou nomes de domínio que só podem ser acessados por funcionários da Huawei, funcionários parceiros, e pessoal terceirizado em zonas amarelas e verdes através da rede de escritórios da Huawei (ou seja, esses nomes de domínio não podem ser acessados pela Internet).

- Os nomes de domínio básicos da Huawei Cloud não são diretamente alocados aos locatários, mas são usados com segurança.
- Os nomes de domínio externos que foram licenciados não são usados pelos serviços da Huawei Cloud para chamar uns aos outros na intranet.

## 8.7 Monitoramento de riscos

O ModelArts monitora automaticamente seus serviços e modelos em tempo real e gerencia alarmes e notificações, para que você possa acompanhar as métricas de desempenho dos serviços e modelos. Para obter detalhes, consulte **Métricas do ModelArts**.

## 8.8 Recuperação de falhas

A infraestrutura global do ModelArts foi criada para regiões e AZs da Huawei Cloud. Uma região da Huawei Cloud fornece várias AZs fisicamente independentes e isoladas que são conectadas por meio de redes com baixa latência, alta taxa de transferência e alta redundância. Você pode projetar e operar aplicações e bancos de dados defeituosos migrados automaticamente entre AZs sem interromper os serviços. Em comparação com a infraestrutura tradicional de um único data center ou vários data centers, as AZs oferecem maior disponibilidade, tolerância a falhas e escalabilidade.

O ModelArts faz backup de seus dados de banco de dados para recuperação em caso de falha de serviço ou danos aos dados originais.

### Recuperação do ambiente de falha

Se um nó de computação usado por uma instância de notebook estiver com defeito, a instância será migrada automaticamente para outro nó disponível. Em seguida, a instância é

restaurada. O ModelArts permite que você monte um disco EVS em uma instância. O EVS da Huawei Cloud fornece armazenamento em bloco escalável que apresenta alta confiabilidade, alto desempenho e uma variedade de especificações para servidores. A durabilidade de dados atinge 99,9999999%.

### Recuperação automática de uma falha de treinamento

Durante o treinamento do modelo, uma falha de treinamento pode ocorrer devido a uma falha de hardware. Para falhas de hardware, o ModelArts fornece verificação de tolerância a falhas para isolar nós defeituosos e melhorar a experiência do usuário no treinamento.

A verificação de tolerância a falhas envolve a pré-verificação do ambiente e a verificação periódica do hardware. Se alguma falha for detectada durante uma das verificações, o ModelArts isola automaticamente o hardware defeituoso e emite o trabalho de treinamento novamente. No treinamento distribuído, a verificação de tolerância a falhas será executada em todos os nós de computação usados pelo trabalho de treinamento.

### Recuperação de uma falha de implementação de inferência

Durante a execução do serviço, se uma instância de inferência apresentar defeito devido a uma falha de hardware, o ModelArts detecta automaticamente a falha e migra a instância defeituosa para outro nó disponível. Depois que a instância for reiniciada, ela será restaurada. O nó defeituoso é automaticamente isolado e não agendado para instâncias de inferência em execução.

## 8.9 Gerenciamento de atualização

### Atualização de serviço em tempo real do ModelArts

Para um serviço implementado, você pode alterar a versão da aplicação de IA para atualizá-lo.

Os serviços podem ser atualizados em três modos: atualização completa, atualização contínua (aumentar instâncias) e atualização contínua (diminuir instâncias). Para obter detalhes sobre os três modos de atualização, consulte **Figura 8-3**.

- Atualização completa

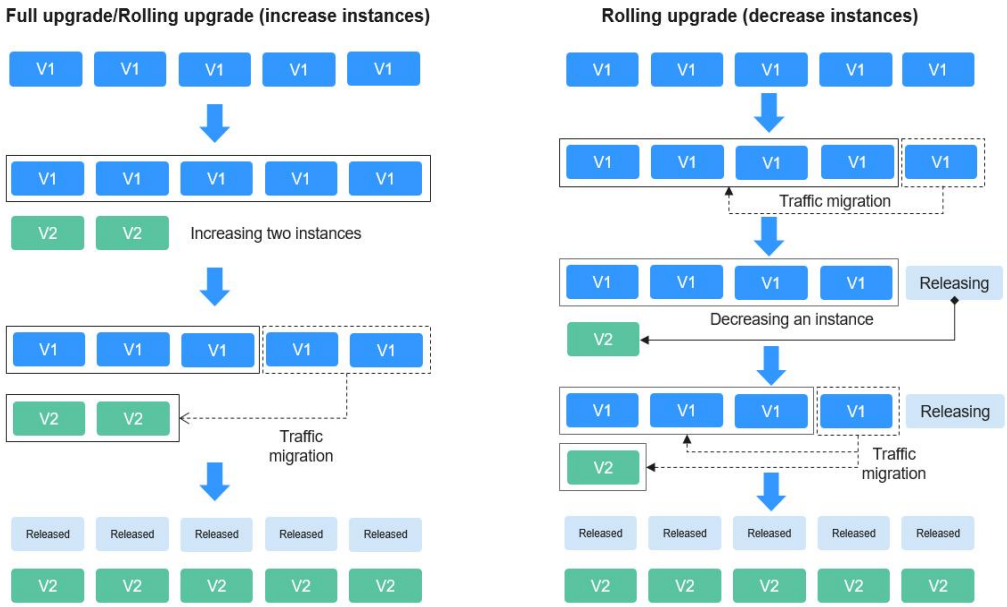
  Os recursos que são duas vezes mais do que os usados pelo serviço serão usados para criar instâncias de nova versão no modo completo.
- Atualização contínua (aumentar instâncias)

  Recursos extras do que os usados pelo serviço serão usados para uma atualização contínua. Um número maior de instâncias a serem aumentadas levará a uma atualização mais rápida.
- Atualização contínua (diminuir instâncias)

  Certos nós destinados a executar serviços serão usados para uma atualização contínua. Um número maior de instâncias a serem reduzidas levará a uma atualização mais rápida, mas a uma maior probabilidade de interrupção do serviço.

**Figura 8-3** Processo de atualização do serviço




Para obter detalhes sobre como atualizar um serviço de inferência, consulte **Atualização de um serviço**.

### Atualização de imagem

O ModelArts fornece três módulos de função: DevEnviron, gerenciamento de treinamento e implementação de inferência. Os três módulos fornecem imagens de base pelo mesmo processo. Essas imagens são atualizadas irregularmente para corrigir vulnerabilidades.

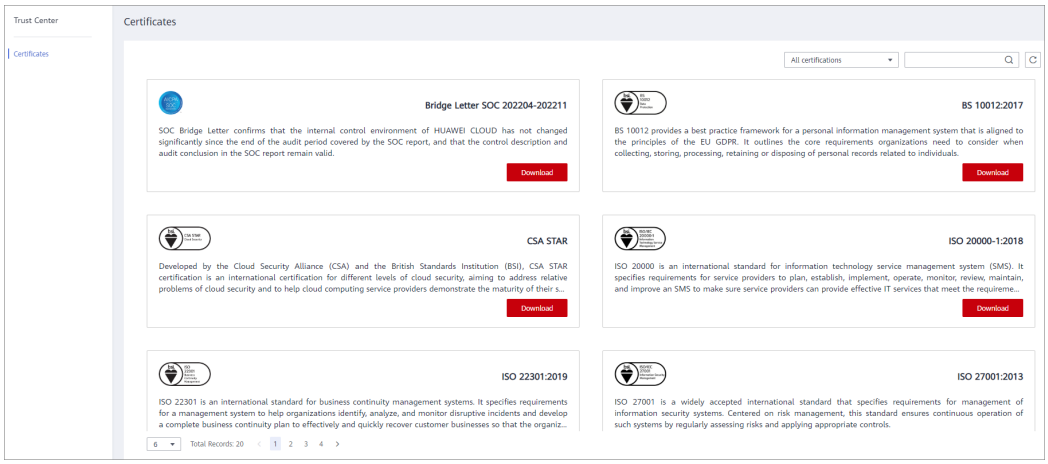
## 8.10 Certificados

### Certificados de conformidade

Os serviços e plataformas da Huawei Cloud obtiveram várias certificações de segurança e conformidade de organizações autorizadas, como a Organização Internacional de Normalização (ISO). Você pode **baixá-los** do console.

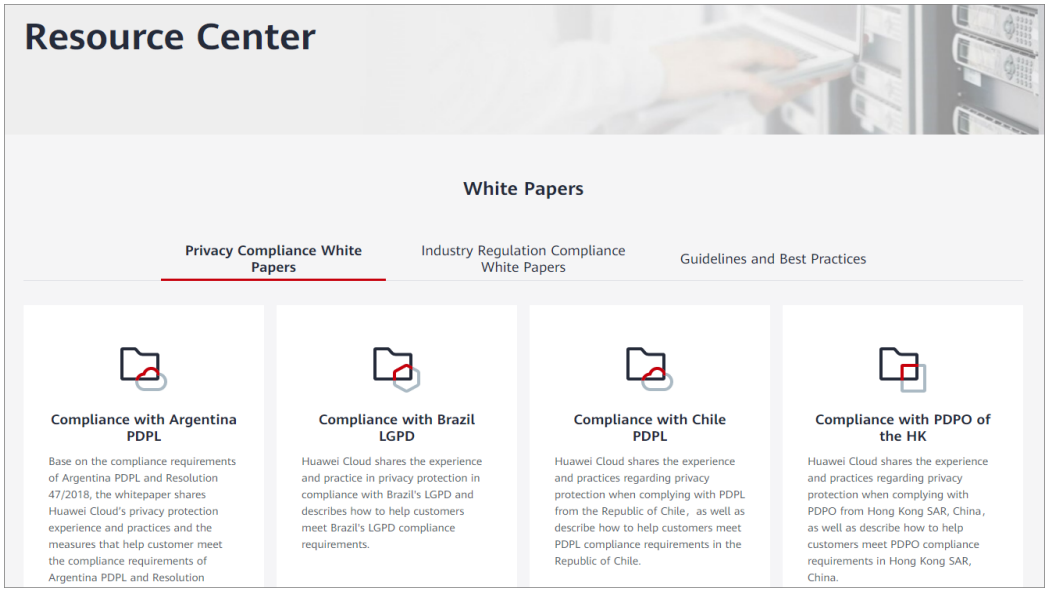
**Figura 8-4** Download de certificações de conformidade




## Central de recursos

A Huawei Cloud também fornece os seguintes recursos para ajudar os usuários a atender aos requisitos de conformidade. Para obter detalhes, consulte **Central de recursos**.

**Figura 8-5** Central de recursos




# 8.11 Fronteira da segurança

O modelo de responsabilidade compartilhada é um modo de cooperação em que tanto provedores quanto clientes assumem responsabilidades de segurança e conformidade dos serviços em nuvem.

Os provedores gerenciam a infraestrutura de nuvem e fornecem hardware e software seguros para garantir a disponibilidade do serviço. Os clientes protegem seus dados e aplicações, ao mesmo tempo em que atendem aos requisitos de conformidade relacionados.

Os prestadores são responsáveis pelos serviços e funções e devem:

- Estabelecer e manter uma infraestrutura segura, incluindo redes, servidores e dispositivos de armazenamento.
- Fornecer plataformas subjacentes confiáveis para garantir a segurança em tempo de execução para o ambiente.
- Fornecer autenticação de identidade e controle de acesso para garantir que apenas usuários autorizados possam acessar os serviços de nuvem e os locatários sejam isolados uns dos outros.
- Fornecer backup e recuperação de desastres confiáveis para evitar a perda de dados devido a falhas de hardware ou desastres naturais.
- Fornecer serviços transparentes de monitoramento e resposta a incidentes, atualizações de segurança e patches de vulnerabilidades.

Os clientes devem:

- Criptografar dados e aplicações para confidencialidade e integridade.
- Certificar-se de que o software da aplicação de IA seja atualizado com segurança e que as vulnerabilidades sejam corrigidas.
- Cumprir os regulamentos relacionados, como GDPR, HIPAA e PCI DSS.
- Controlar o acesso para garantir que somente usuários autorizados possam acessar e gerenciar recursos, como serviços on-line.
- Monitorar e relatar qualquer atividade anormal e tomar medidas em tempo hábil.

## Responsabilidades de segurança de implementação de inferência

- Provedores
  - Corrija os patches relacionados aos ECSs subjacentes.
  - Atualize o K8S e corrija vulnerabilidades.
  - Opere a manutenção do ciclo de vida do sistema operacional da VM.
  - Garanta a segurança e a conformidade da plataforma de inferência ModelArts.
  - Melhore a segurança dos serviços de aplicações de contêiner.
  - Atualize o ambiente de tempo de execução do modelo e corrija vulnerabilidades periodicamente.
- Clientes
  - Autorize o uso de recursos e controlar o acesso.
  - Garanta a segurança de aplicações, sua cadeia de suprimentos e dependências por meio de verificação de segurança, auditoria e verificação de acesso.
  - Minimize as permissões e limite a entrega de credenciais.
  - Garanta a segurança das aplicações de IA (imagens personalizadas, modelos OBS e dependências) durante o tempo de execução.
  - Atualize e corrija vulnerabilidades em tempo hábil.
  - Armazene com segurança dados confidenciais, como credenciais.

## Melhores práticas para segurança de implementação de inferência

- Autorização de serviço externo

A inferência do ModelArts requer autorização de outros serviços em nuvem. Você pode conceder apenas as permissões necessárias com base em suas necessidades. Por exemplo, você pode conceder permissão de acesso em um bucket do OBS a um locatário para gerenciamento de modelo.

- Autorização de recursos internos

  A inferência do ModelArts suporta controle de permissão refinado. Você pode configurar as permissões para os usuários com base nas necessidades reais para restringir as permissões em alguns recursos.

- Gerenciamento de aplicações de IA

  Para dissociar modelos de imagens e proteger ativos de modelo, você pode importar dinamicamente aplicações de IA de treinamentos ou OBS. Você precisa atualizar os pacotes de dependência de aplicações de IA e corrigir vulnerabilidades em pacotes de código aberto ou de terceiros. As informações sensíveis relacionadas às aplicações de IA precisam ser desacopladas e configuradas durante a implementação. Selecione o ambiente de tempo de execução recomendado pelo ModelArts. Os ambientes anteriores podem ter vulnerabilidades de segurança.

  Você pode selecionar imagens confiáveis abertas ao criar aplicações de IA a partir de uma imagem de contêiner, por exemplo, imagens do OpenEuler, Ubuntu e NVIDIA. Crie usuários não raiz em vez de usuários raiz para executar uma imagem. Somente o pacote de segurança necessário durante o tempo de execução é instalado na imagem. Reduza o tamanho da imagem e atualize o pacote de instalação para a versão mais recente livre de vulnerabilidades. Desacople informações confidenciais de imagens durante a implementação do serviço. Não use diretamente as informações no Dockerfile. Realize varreduras de segurança em imagens periodicamente e instale patches para corrigir vulnerabilidades. Para facilitar o relatório de alarme e a retificação de falhas, adicione a interface de verificação de integridade e certifique-se de que o status do serviço possa ser retornado corretamente. Para garantir a segurança dos dados do serviço, use fluxos de transmissão HTTPS e pacotes de criptografia confiáveis para contêineres.

- Implementação de modelo

  Para evitar que os serviços sejam sobrecarregados ou desperdiçados, defina as especificações de nó de computação adequadas durante a implementação. Não ouça outras portas no contêiner. Se outras portas precisarem ser acessadas localmente, ouça-as em localhost. Não transfira diretamente informações confidenciais por meio de variáveis de ambiente. Criptografe informações confidenciais com componente de criptografia antes da transmissão de dados.

  A chave de autenticação da aplicação é uma credencial de acesso para serviços em tempo real. Você deve manter a chave da aplicação corretamente.

# 9 Cotas

O ModelArts utiliza os seguintes recursos de infra-estrutura:

- ECS
- EVS
- VPC

Para obter detalhes sobre como exibir e modificar a cota, consulte **Cotas**.