ModelArts

Descripción general del servicio

Edición	01
Fecha	2024-09-20

HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD.

Marcas registradas y permisos

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

Aviso

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

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

Huawei Cloud Computing Technologies Co., Ltd.

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

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

Índice

1 Infografías

1.1 Qué es ModelArts

2 Qué es ModelArts

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

ModelArts cubre todas las etapas del desarrollo de la IA, incluidos el procesamiento de datos, el desarrollo de algoritmos y los entrenamiento y despliegue de modelos. Las tecnologías subyacentes de ModelArts admiten varios recursos informáticos heterogéneos, lo que permite a los desarrolladores seleccionar y usar recursos de forma flexible. Además, ModelArts soporta marcos de desarrollo de IA de código abierto populares como TensorFlow, PyTorch y MindSpore. ModelArts también es compatible con los marcos de algoritmos personalizados adaptados a sus necesidades.

ModelArts pretende simplificar el desarrollo de la IA.

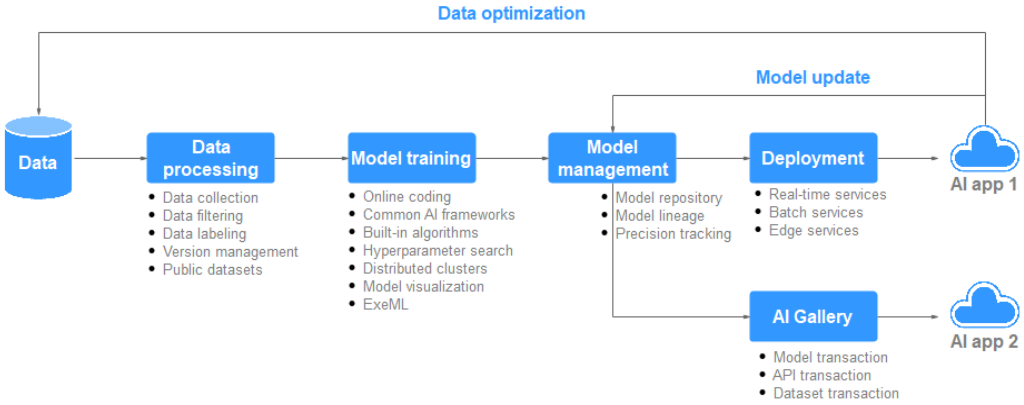
ModelArts es adecuado para desarrolladores de IA con diferentes niveles de experiencia en desarrollo. Los desarrolladores de servicios pueden usar ExeML para crear rápidamente las aplicaciones de IA sin codificación. Los principiantes pueden usar directamente los algoritmos integrados para crear las aplicaciones de IA. Los ingenieros de IA pueden utilizar múltiples entornos de desarrollo para compilar rápidamente los códigos para el modelado y el desarrollo de aplicaciones.

Arquitectura del producto

ModelArts es compatible con todo el proceso de desarrollo, incluidos el procesamiento de datos y el entrenamiento, gestión y despliegue de modelos. También proporciona AI Gallery para compartir modelos.

ModelArts admite varios escenarios de la aplicación de IA, como clasificación de imágenes, detección de objetos, análisis de vídeo, reconocimiento de voz, recomendación de productos y detección de excepciones.

Figura 2-1 Arquitectura de ModelArts



Ventajas del producto

- **Plataforma única**

  La plataforma de desarrollo de IA de ciclo completo y listo para usar proporciona procesamiento de datos y desarrollo, entrenamiento, gestión y despliegue de modelos.

- **Fácil para el uso**
  - Múltiples modelos incorporados proporcionados y el uso gratuito de modelos de código abierto
  - Optimización automática de hiperparámetros
  - Desarrollo sin código y operaciones simplificadas
  - Despliegue con un solo clic de los modelos a la nube, al perímetro y a los dispositivos

- **Alto rendimiento**
  - El marco de aprendizaje profundo MoXing desarrollado por sí mismo acelera el desarrollo y el entrenamiento de algoritmos.
  - La utilización optimizada de la GPU acelera la inferencia en tiempo real.
  - Los modelos que se ejecutan en chips de AI de Ascend logran una inferencia más eficiente.

- **Flexible**
  - Marcos principales de código abierto como TensorFlow, PyTorch y MindSpore
  - GPU principal
  - Chip de Ascend
  - Uso exclusivo de recursos dedicados
  - Imágenes personalizadas para marcos y operadores personalizados

Usar ModelArts por primera vez

Si usted es un usuario primerizo, la siguiente información le ayudará a familiarizarse con ModelArts:

- **Conceptos básicos**

Conocimiento básico describe los conceptos básicos de ModelArts, incluidos los proceso y conceptos básicos del desarrollo de IA, y conceptos y funciones específicos de ModelArts.

- **Pasos iniciales**

  Pasos iniciales ofrece ejemplos con operaciones detalladas para ayudarle a empezar a utilizar ModelArts.

- **Prácticas recomendadas**

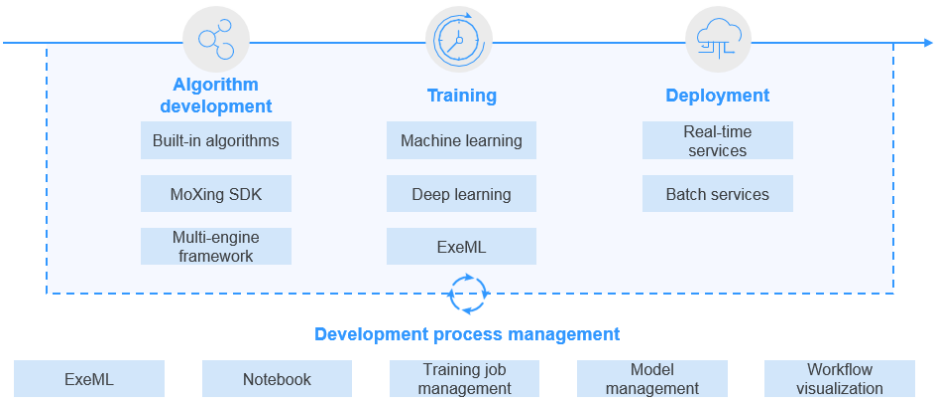
  ModelArts es compatible con múltiples motores de código abierto y proporciona casos de uso extensos basados en los motores y funciones. Puede crear y desplegar modelos consultando **Prácticas recomendadas**.

- **Guías para otras funciones y operaciones**
  - Si usted es un desarrollador de servicios, puede usar ExeML para crear modelos rápidamente sin necesidad de codificación. Consulte la **Guía de usuario (ExeML)** para obtener más detalles.
  - Si es ingeniero de IA, puede utilizar una o más funciones en su desarrollo de IA, como **DevEnviron**, **preparación de datos**, **etiquetado de datos**, **desarrollo de modelos** e **inferencia**. Puede utilizar una o más funciones en su desarrollo de IA.
  - Si quiere usar las API o los SDK de ModelArts para el desarrollo de IA, consulte la ***Referencia de las API*** o la ***Referencia de los SDK***.

3 Funciones

Los ingenieros de IA enfrentan desafíos en la instalación y configuración de diversas herramientas de IA, la preparación de datos y el entrenamiento de modelos. Para hacer frente a estos desafíos, se proporciona la plataforma de desarrollo de IA integral ModelArts. La plataforma integra la preparación de datos, el desarrollo de algoritmos, el entrenamiento de modelos y el despliegue de modelos en el entorno de producción, lo que permite a los ingenieros de IA realizar un desarrollo de IA integral.

Figura 3-1 Descripción de la función



ModelArts tiene las siguientes características:

- **Gobernanza de datos**

  Gestiona la preparación de datos, como el filtrado y etiquetado de datos, y las versiones de conjuntos de datos.

- **Entrenamiento rápido y simplificado del modelo**

  Permite el entrenamiento distribuido de alto rendimiento y simplifica la codificación con el marco de aprendizaje profundo MoXing desarrollado por sí mismo.

- **Sinergía de dispositivos de borde en la nube**

  Despliegue los modelos en diversos entornos de producción, como dispositivos, el perímetro y la nube, y admite inferencias en tiempo real y por lotes.

- **Autoaprendizaje**

Permite la creación de modelos sin codificación y admite la clasificación de imágenes, la detección de objetos y el análisis predictivo.

4 Conocimiento básico

4.1 Introducción al ciclo de vida del desarrollo de la IA

Qué es IA

La inteligencia artificial (IA) es una tecnología capaz de simular la cognición humana a través de máquinas. La capacidad central de la IA es hacer un juicio o predicción basado en una entrada dada.

¿Cuál es el propósito del desarrollo de la IA?

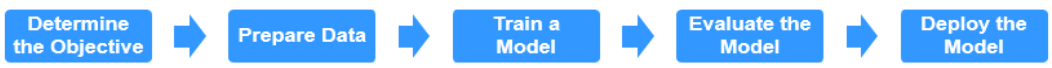
El desarrollo de la IA apunta a procesar y extraer información de volúmenes de datos de manera centralizada para resumir los patrones internos de los objetos de estudio.

Se computan, analizan, resumen y organizan volúmenes masivos de datos recopilados mediante estadísticas, aprendizaje automático y métodos de aprendizaje profundo adecuados para maximizar el valor de los datos.

Proceso básico de desarrollo de la IA

El proceso básico del desarrollo de la IA incluye los siguientes pasos: determinar un objetivo, preparar datos y entrenar, evaluar e implementar un modelo.

Figura 4-1 Proceso de desarrollo de IA



Paso 1 Determinar un objetivo.

Antes de comenzar el desarrollo de IA, determine qué analizar. ¿Qué problemas quiere resolver? ¿Cuál es el objetivo de negocio? Clasifique el marco de desarrollo de IA y las ideas basadas en la comprensión del negocio. Por ejemplo, clasificación de imágenes y detección de objetos. Diferentes proyectos tienen diferentes requisitos para datos y métodos de desarrollo de IA.

Paso 2 Preparar los datos.

La preparación de datos se refiere a la recopilación y el preprocesamiento de datos.

La preparación de datos es la base del desarrollo de IA. Cuando se recopilan e integran datos relacionados en función del objetivo determinado, lo más importante es garantizar la autenticidad y fiabilidad de los datos obtenidos. Por lo general, no puede recopilar todos los datos al mismo tiempo. En la fase de etiquetado de datos, puede encontrar que faltan algunas fuentes de datos y, a continuación, puede que tenga que ajustar y optimizar los datos repetidamente.

Paso 3 Entrenar a un modelo.

El modelado implica analizar los datos preparados para encontrar la causalidad, las relaciones internas y los patrones regulares, proporcionando así referencias para la toma de decisiones comerciales. Después del entrenamiento del modelo, generalmente se generan uno o más modelos de aprendizaje automático o aprendizaje profundo. Estos modelos se pueden aplicar a nuevos datos para obtener predicciones y resultados de evaluación.

Un gran número de desarrolladores desarrollan y entrenan modelos requeridos por servicios relevantes basados en motores de IA populares, como TensorFlow, Spark_MLlib, MXNet, Caffe, PyTorch, XGBoost-Sklearn, y MindSpore.

Paso 4 Evaluar el modelo.

Es necesario evaluar un modelo generado por el entrenamiento. Típicamente, no se puede obtener un modelo satisfactorio después de la primera evaluación, y puede ser necesario ajustar repetidamente los parámetros y datos del algoritmo para optimizar aún más el modelo.

Algunas métricas comunes, como la precisión, el recuerdo y el área bajo la curva (AUC), le ayudan a evaluar y obtener un modelo satisfactorio de manera efectiva.

Paso 5 Desplegar el modelo.

El desarrollo y el entrenamiento del modelo se basan en datos existentes (que pueden ser datos de prueba). Después de obtener un modelo satisfactorio, el modelo debe aplicarse formalmente a los datos reales o a los datos recién generados para la predicción, evaluación y visualización. Los resultados se pueden comunicar a los responsables de la toma de decisiones de forma intuitiva, ayudándoles a desarrollar las estrategias empresariales adecuadas.

----Fin

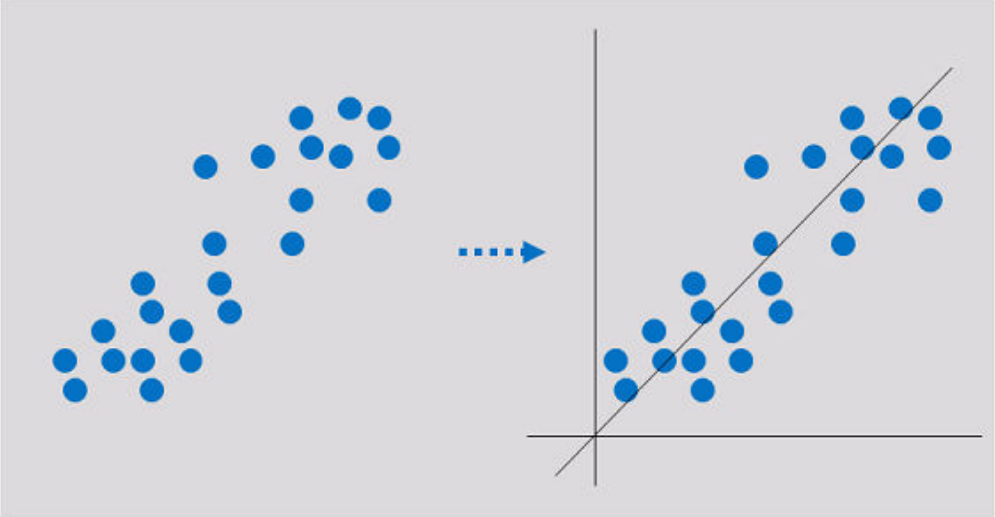
4.2 Conceptos básicos del desarrollo de IA

El aprendizaje automático se clasifica en aprendizaje supervisado, no supervisado y de refuerzo.

- El aprendizaje supervisado utiliza muestras etiquetadas para ajustar los parámetros de los clasificadores para lograr el rendimiento requerido. Se puede considerar como un aprendizaje con un maestro. El aprendizaje supervisado común incluye regresión y clasificación.
- El aprendizaje no supervisado se utiliza para encontrar las estructuras ocultas en datos sin etiquetar. El agrupamiento es una forma de aprendizaje no supervisado.
- El aprendizaje reforzado es un área del aprendizaje automático que se ocupa de cómo los agentes de software deben tomar acciones en un entorno para maximizar alguna noción de recompensa acumulada.

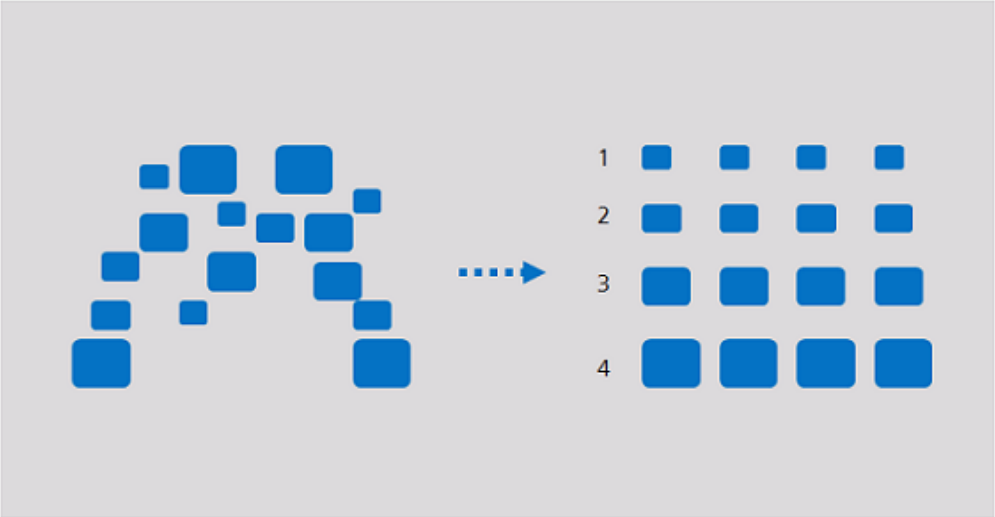
Regresión

La regresión refleja la característica de tiempo de los atributos de datos y genera una función que asigna un atributo de datos a una predicción de variable real para encontrar la dependencia entre la variable y el atributo. La regresión analiza principalmente los datos y predice la relación entre los datos y los datos. La regresión se puede utilizar para el desarrollo de clientes, la retención, la prevención de la rotación de clientes, el análisis del ciclo de vida de la producción, la predicción de tendencias de ventas y la promoción dirigida.



Clasificación

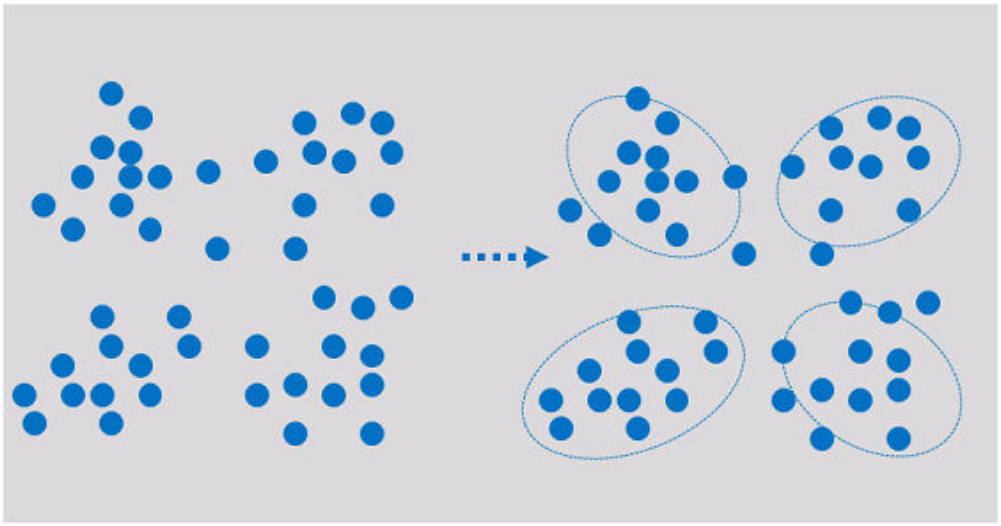
La clasificación implica definir un conjunto de categorías basadas en las características comunes de los objetos e identificar a qué categoría pertenece un objeto. La clasificación se puede utilizar para la clasificación de clientes, propiedades de clientes, análisis de características, análisis de satisfacción de clientes y predicción de tendencias de compra de clientes.



Clusterización

La clusterización consiste en agrupar un conjunto de objetos de manera que los objetos de un mismo grupo sean más similares entre sí que los de otros grupos. La clusterización se puede

utilizar para la segmentación de clientes, análisis de características de clientes, predicción de tendencias de compra de clientes y segmentación de mercado.



La clusterización analizan los objetos de datos y produce etiquetas de clase. Los objetos se agrupan según las similitudes maximizadas y minimizadas para formar clústeres. De esta manera, los objetos en el mismo clúster son más similares entre sí que los de otros clústeres.

4.3 Conceptos comunes de ModelArts

ExeML

ExeML es el proceso de automatizar el diseño de modelos, el ajuste de parámetros y el entrenamiento de modelos, la compresión de modelos y el despliegue de modelos con los datos etiquetados. El proceso es libre de código y no requiere que los desarrolladores tengan experiencia en el desarrollo de modelos. Un modelo se puede construir en tres pasos: etiquetar datos, entrenar un modelo e implementar el modelo.

Dispositivo-Frontera-Nube

Dispositivo-Frontera-Nube indica dispositivos, nodos perimetrales inteligentes y la nube pública.

Inferencia

Inference es el proceso de derivar un nuevo juicio a partir de un juicio conocido de acuerdo con una determinada estrategia. En IA, las máquinas simulan la inteligencia humana y la inferencia completa basada en redes neuronales.

Inferencia en tiempo real

Real-Time Inference especifica un servicio web que proporciona un resultado de inferencia para cada solicitud de inferencia.

Inferencia por lotes

Batch inference especifica un trabajo por lotes que procesa datos por lotes para inferencia.

Chip de Ascend

Los chips de Ascend son una serie de chips de IA desarrollados por Huaweicon alto rendimiento informático y bajo consumo de energía.

4.4 Introducción a las herramientas de desarrollo

NOTA

Este documento describe las funciones del notebook de DevEnviron de la nueva versión.

El desarrollo de software es un proceso de reducir los costos de los desarrolladores y mejorar la experiencia de desarrollo. En el desarrollo de IA, el ModelArts se dedica a mejorar la experiencia de desarrollo de IA y a simplificar el proceso de desarrollo. ModelArts DevEnviron utiliza los recursos nativos en la nube e integra la cadena de herramientas de desarrollo para proporcionar una mejor experiencia de desarrollo de IA en la nube para el desarrollo, la exploración y la enseñanza de IA.

Notebook de ModelArts para una perfecta colaboración en la nube y en las instalaciones

- JupyterLab en la nube, IDE local y complementos de ModelArts para desarrollo y depuración remotos, adaptados a sus necesidades
- Entorno de desarrollo en la nube con recursos informáticos de IA, almacenamiento en la nube y motores de IA integrados
- Entorno de tiempo de ejecución personalizado guardado como una imagen para entrenamiento e inferencia

Característica 1: Desarrollo remoto, que permite el acceso remoto al notebook desde un IDE local

El notebook de la nueva versión proporciona el despliegue remoto. Después de habilitar SSH remoto, puede acceder de forma remota al entorno de despliegue de notebooks de ModelArts para depurar y ejecutar el código desde un IDE local.

Debido a los recursos locales limitados, los desarrolladores que usan un IDE local ejecutan y depuran código normalmente en un servidor de CPU o GPU compartido entre los miembros del equipo. La construcción y el mantenimiento del servidor de CPU o GPU son costosos.

Las instancias de notebooks de ModelArts están listas para usar con varios motores y variantes integrados para que seleccione. Puede utilizar un entorno de contenedor dedicado. Solo después de configuraciones simples, puede acceder remotamente al entorno para ejecutar y depurar código desde su IDE local.

El notebook de ModelArts puede considerarse como una extensión de un entorno de desarrollo local. Las operaciones como la lectura de datos, el entrenamiento y el almacenamiento de archivos son las mismas que las realizadas en un entorno local.

El notebook de ModelArts le permite usar recursos en la nube sin cambios en los hábitos de codificación locales.

Un IDE local admite Visual Studio (VS) Code, PyCharm y SSH.

Característica 2: Imágenes preestablecidas que están listas para usar con configuraciones optimizadas y compatibles con los motores de IA convencionales

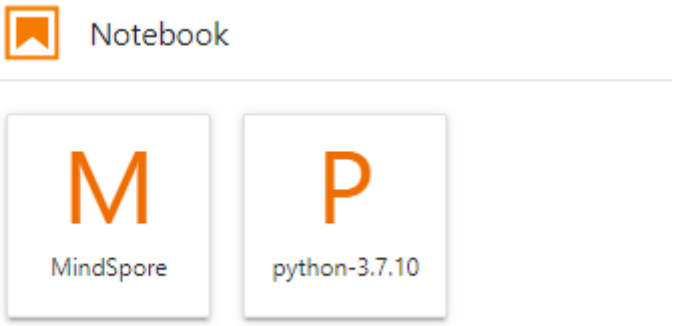
Los motores de IA y las versiones preestablecidas en cada imagen son fijos. Al crear una instancia de notebook, especifique un motor y una versión de IA, incluido el tipo de chip.

ModelArts DevEnviron proporciona un grupo de imágenes preestablecidas, incluidas las imágenes de PyTorch, de TensorFlow y de MindSpore. Puede utilizar una imagen preestablecida para iniciar la instancia del notebook. Después del despliegue en la instancia, presenta un trabajo de entrenamiento sin ninguna adaptación.

Las versiones de imagen preestablecidas de ModelArts se determinan en función de los comentarios del usuario y la estabilidad de la versión. Si su despliegue se puede llevar a cabo utilizando las versiones preestablecidas de ModelArts, por ejemplo, MindSpore 1.5, utilice las imágenes preestablecidas. Estas imágenes han sido completamente verificadas y tienen muchos paquetes de instalación de uso común incorporados. Están listos para usar, lo que le libera de configurar el entorno.

Las imágenes preestablecidas en el DevEnviron de ModelArts mantiene:

- Paquetes preestablecidos comunes: motores comunes de IA como PyTorch y MindSpore basados en el estándar Conda, paquetes comunes de software de análisis de datos como Pandas y Numpy, y software de herramientas comunes como CUDA y CUDNN, que cumplen con los requisitos comunes de despliegue de IA.
- Entornos de Conda preestablecidos: Se crean un entorno de Conda y Conda Python básico (excluyendo cualquier motor de IA) para cada imagen preestablecida. La siguiente figura muestra el entorno de Conda para MindSpore preestablecido.

  Seleccione un entorno de Conda en función de si se utiliza el motor AI para la depuración.
- Notebook: una aplicación web que permite codificar en la GUI y combinar el código, las ecuaciones matemáticas y el contenido visualizado en un documento.
- Complementos de JupyterLab: permite cambiar de variante, compartir casos en AI Gallery para la comunicación y detener la instancia para mejorar la experiencia del usuario.
- SSH remoto: le permite depurar remotamente una instancia de notebook desde un PC local.
- Después de las imágenes preestablecidas en el despliegue de soporte de ModelArts DevEnviron, los trabajos de entrenamiento se pueden ejecutar en ModelArts.

NOTA

- Para simplificar las operaciones, el notebook de ModelArts de la nueva versión no admite la conmutación entre motores de IA en una instancia de notebook.
- Los motores de IA varían según las regiones. Para obtener más información sobre los motores de IA disponibles en una región, consulte los motores de IA que se muestran en la consola de gestión.

Característica 3: JupyterLab, una herramienta interactiva de despliegue y depuración en línea

ModelArts integra JupyterLab de código abierto para el despliegue interactivo en línea y la depuración. Puede utilizar el notebook en la consola de gestión de ModelArts para compilar y depurar código y entrenar modelos basados en el código, sin importar la instalación o configuración del entorno.

JupyterLab es un entorno de despliegue interactivo. Es el producto de última generación de Jupyter Notebook. JupyterLab le permite compilar notebook, operar terminales, editar texto Markdown, habilitar la interacción y ver archivos e imágenes CSV.

4.5 Entrenamiento de modelos

Además de datos y algoritmos, los desarrolladores dedican mucho tiempo a configurar los parámetros de entrenamiento del modelo. Los parámetros de entrenamiento del modelo determinan la precisión y el tiempo de convergencia del modelo. La selección de parámetros depende en gran medida de la experiencia de los desarrolladores. La selección incorrecta de los parámetros afectará a la precisión del modelo o aumentará significativamente el tiempo requerido para el entrenamiento del modelo.

Para simplificar el desarrollo de IA y mejorar la eficiencia del desarrollo y el rendimiento del entrenamiento, ModelArts ofrece las gestiones visualizadas de trabajos, recursos y versiones, también realiza automáticamente la optimización de hiperparámetros basada en el aprendizaje automático y el aprendizaje por refuerzo. Proporciona políticas automáticas de ajuste de hiperparámetros, como la tasa de aprendizaje y el tamaño de lote, e integra los modelos comunes.

Actualmente, cuando la mayoría de los desarrolladores construyen modelos, los modelos suelen tener docenas de capas o incluso cientos de capas y parámetros de nivel MB o GB para cumplir con los requisitos de precisión. Como resultado, las especificaciones de los recursos de computación son extremadamente altas, especialmente la potencia de computación de los recursos de hardware, la memoria y la ROM. Las especificaciones de recursos del lado del dispositivo son estrictamente limitadas. Por ejemplo, la potencia de computación del lado del dispositivo es de 1 TFLOPS, el tamaño de la memoria es de aproximadamente 2 GB y el espacio de ROM es de aproximadamente 2 GB, por lo que el tamaño del modelo en el lado del dispositivo debe estar limitado a 100 KB y el retardo de inferencia debe estar limitado a 100 milisegundos.

Por lo tanto, las tecnologías de compresión con precisión de modelo sin pérdida o casi sin pérdida, como la poda, la cuantificación y la destilación de conocimientos, se utilizan para implementar la compresión y optimización automáticas de modelos, y la iteración automática de compresión y reentrenamiento de modelos para controlar la pérdida de precisión de los modelos. La tecnología de cuantificación de bits bajos, que elimina la necesidad de reentrenamiento, convierte el modelo de punto flotante de alta precisión en una operación de punto fijo. Se utilizan múltiples tecnologías de compresión y optimización para satisfacer los

requisitos de ligereza de los recursos de hardware de dispositivos y bordes. La tecnología de compresión del modelo reduce la precisión en menos del 1% en valores específicos.

Cuando el volumen de datos de entrenamiento es grande, el entrenamiento del modelo de aprendizaje profundo consume mucho tiempo. En tecnología de visión artificial, ImageNet-1k (un conjunto de datos de clasificación que contiene 1,000 clases de imagen, denominado ImageNet) es un conjunto de datos de uso común. Si utiliza una GPU P100 para entrenar un modelo ResNet-50 en el conjunto de datos, tardará casi una semana. Esto dificulta el rápido desarrollo de aplicaciones de aprendizaje profundo. Por lo tanto, la aceleración del entrenamiento de aprendizaje profundo siempre ha sido una preocupación importante para la academia y la industria.

La aceleración distribuida del entrenamiento debe considerarse en términos de software y de hardware. Un único método de optimización no puede responder a las expectativas. Por lo tanto, la optimización de la aceleración distribuida es un proyecto de sistema. La arquitectura de entrenamiento distribuido debe considerarse en términos de hardware y diseño de chips. Para minimizar los retrasos de computación y de comunicación, hay que tener en cuenta muchos factores, como las especificaciones generales de cálculo, el ancho de banda de la red, la caché de alta velocidad, el consumo de energía y la disipación térmica del sistema, así como la relación entre el rendimiento de cálculo y el de comunicación.

El diseño del software debe combinar funciones de hardware de alto rendimiento para aprovechar al máximo la red de hardware de alta velocidad e implantar una comunicación distribuida de gran ancho de banda y un almacenamiento en caché local de datos eficiente. Mediante el uso de algoritmos de optimización del entrenamiento, como el paralelo híbrido, la compresión de gradiente y la aceleración de convolución, el software y el hardware del sistema de entrenamiento distribuido pueden coordinarse y optimizarse eficientemente de extremo a extremo, y la aceleración del entrenamiento puede implementarse en un entorno distribuido de múltiples hosts y tarjetas. ModelArts ofrece una aceleración líder en la industria de más de 0.8 para ResNet50 en el conjunto de datos de ImageNet en un entorno distribuido con miles de hosts y tarjetas.

Para medir el rendimiento de aceleración del aprendizaje profundo distribuido, se utilizan los siguientes dos indicadores clave:

- Rendimiento, es decir, la cantidad de datos procesados en una unidad de tiempo
- Tiempo de convergencia, es decir, el tiempo necesario para lograr cierta precisión

El rendimiento depende del hardware del servidor (por ejemplo, más chips de aceleración de IA con mayor capacidad de procesamiento de FLOPS y mayor ancho de banda de comunicación logran un mayor rendimiento), la lectura y el almacenamiento en caché de datos, el preprocesamiento de datos, la computación de modelos (por ejemplo, la selección del algoritmo de convolución) y la optimización de la topología de comunicación. Excepto la computación de bajo bit y la compresión de gradiente (o parámetros), la mayoría de las tecnologías mejoran el rendimiento sin afectar la precisión del modelo. Para lograr el menor tiempo de convergencia, es necesario optimizar el rendimiento y ajustar los parámetros. Si los parámetros no se ajustan correctamente, no se puede optimizar el rendimiento. Si el tamaño del lote se establece en un valor pequeño, el rendimiento paralelo del entrenamiento del modelo será relativamente pobre. Como resultado, el rendimiento no puede mejorarse incluso si se aumenta el número de nodos de computación.

Lo que más preocupa a los usuarios es el tiempo de convergencia. El marco de MoXing implementa la organización completa y reduce significativamente el tiempo de convergencia del entrenamiento. Para la lectura y el preprocesamiento de datos, MoXing utiliza las canalizaciones de entrada simultáneas de varios niveles para evitar que las E/S de datos se conviertan en cuellos de botella. En cuanto a la computación de modelos, MoXing ofrece

cálculo de precisión híbrido, que combina semiprecisión y precisión única para los modelos de capas superiores y reduce la pérdida causada por el cálculo de precisión mediante escalado adaptativo. Las políticas de hiperparámetros dinámicos (como el impulso y el tamaño del lote) se utilizan para minimizar el número de épocas necesarias para la convergencia del modelo. MoXing también funciona con servidores y bibliotecas informáticas subyacentes de Huawei para mejorar aún más la aceleración distribuida.

Optimización del entrenamiento distribuido de alto rendimiento de ModelArts

- Precisión híbrida automática para utilizar plenamente las capacidades informáticas de hardware
- Tecnologías de ajuste dinámico de hiperparámetros (tamaño dinámico de lote, tamaño de imagen e impulso)
- Combinación y separación automáticas de gradiente del modelo
- Optimización de la programación de operadores de comunicaciones basada en la computación adaptativa de BP bubble
- Bibliotecas de comunicación distribuidas de alto rendimiento (NStack y HCCL)
- Paralelo híbrido de modelo de datos distribuido
- Compresión de datos de entrenamiento y caché multinivel

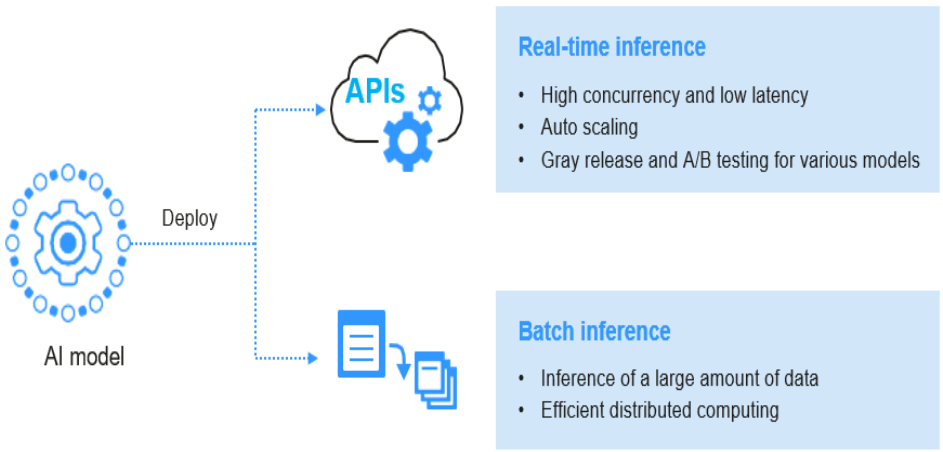
4.6 Despliegue de modelos

ModelArts es capaz de gestionar modelos y servicios. Esto permite gestionar de forma unificada las imágenes y los modelos de las principales estructuras de varios proveedores.

En general, el despliegue del modelo de IA y la implementación a gran escala son complejos.

Por ejemplo, en un proyecto de transporte inteligente, el modelo capacitado debe desplegarse en la nube, los bordes y los dispositivos. Desplegar el modelo en los dispositivos requiere tiempo y esfuerzo, por ejemplo, en las cámaras de diferentes especificaciones y vendedores. ModelArts admite el despliegue con un solo clic de un modelo entrenado en varios dispositivos para diferentes escenarios de aplicación. Además, ofrece un conjunto de modos de despliegue seguros y fiables de ventanilla única para desarrolladores individuales, empresas y fabricantes de dispositivos.

Figura 4-2 Proceso de desplegar un modelo



- El servicio de inferencia en tiempo real cuenta con alta concordancia, baja demora y escalamiento. Se soporta el lanzamiento de varios modelos y pruebas A/B.
- Los modelos se pueden desplegar como servicios de inferencia en tiempo real y tareas de inferencia por lotes en la nube.

5 Servicios relacionados

IAM

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

OBS

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

Tabla 5-1 Relación entre ModelArts y OBS

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

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

EVS

ModelArts utiliza Elastic Volume Service (EVS) para almacenar instancias de notebook creadas. Para obtener más información, consulte la ***Guía de usuario de Elastic Volume Service***.

CCE

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

SWR

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

Cloud Eye

ModelArts utiliza Cloud Eye para monitorear los servicios en línea y modelar cargas en tiempo real y enviar alarmas y notificaciones automáticamente. Consulte la ***Guía de usuario de Cloud Eye*** para obtener más información sobre Cloud Eye.

6 Cómo accedo a ModelArts

Puede acceder a ModelArts a través de la consola de gestión basada en web o mediante interfaces de programación de aplicaciones (API) basadas en HTTPS.

- **Con la consola de gestión**

  ModelArts cuenta con una consola de gestión sencilla y fácil de usar, y ofrece una serie de funciones como ExeML, gestión de datos, entorno de desarrollo, entrenamiento de modelos, gestión de aplicaciones de IA, AI Gallery y despliegue de servicios. Puede completar el desarrollo de IA de extremo a extremo en la consola de gestión.

  Para usar la consola de gestión de ModelArts, primero debe registrarse en Huawei Cloud. Si ha creado una cuenta de Huawei Cloud, seleccione **AI** > **ModelArts** en el sitio web oficial e inicie sesión en la consola de gestión.

- **Con los SDK**

  Si desea integrar ModelArts en un sistema de terceros para desarrollo secundario, llame a SDK para completar el desarrollo. Los SDK de ModelArts encapsulan las API RESTful proporcionadas por ModelArts para simplificar el desarrollo secundario. Para obtener más información acerca de los SDK y las operaciones, consulte ***Referencia del SDK de ModelArts***.

  Además, puede invocar directamente a los SDK de ModelArts cuando escriba código en un notebook en la consola de gestión.

- **Con las API**

  Si quiere integrar ModelArts en un sistema de terceros para desarrollo secundario, usa APIs para acceder a ModelArts. Consulte la ***Referencia de las API de ModelArts*** para obtener detalles sobre las API y las operaciones.

7 Gestión de permisos

ModelArts le permite configurar permisos de grano fino para una gestión refinada de recursos y permisos. Esto lo utilizan habitualmente las grandes empresas, pero es complejo para los usuarios individuales. Se recomienda que los usuarios individuales configuren los permisos para utilizar ModelArts según **Asignación de permisos a los usuarios individuales para usar ModelArts**.

NOTA

Si cumple alguna de las siguientes condiciones, lea este documento.

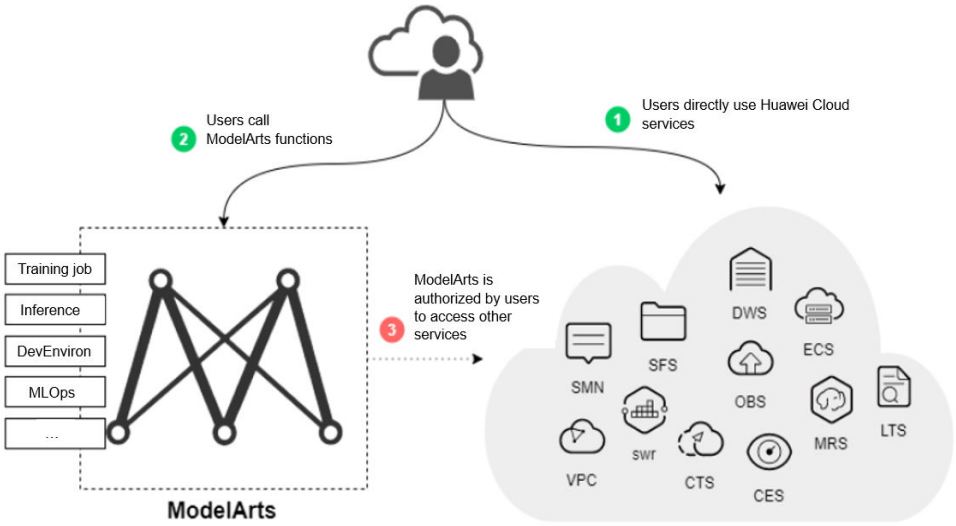
- Usted es un usuario empresarial y
  - En su empresa hay varios departamentos y necesita controlar los permisos de los usuarios para que los usuarios de los distintos departamentos sólo puedan acceder a los recursos y funciones que les corresponden.
  - Existen múltiples roles en su empresa, incluidos administradores, desarrolladores de algoritmos e ingenieros de O&M de aplicaciones. Solo es necesario que utilicen las funciones específicas.
  - Lógicamente, existen varios entornos aislados, como el entorno de desarrollo, el entorno de preproducción y el entorno de producción. Es necesario controlar los permisos de los usuarios en diferentes entornos.
  - Es necesario controlar los permisos de usuarios o grupos de usuarios específicos de IAM.
- Usted es un usuario individual y ha creado varios usuarios de IAM. Debe asignar los permisos diferentes de ModelArts a diferentes usuarios de IAM.
- Debe comprender los conceptos y operaciones de la gestión de permisos de ModelArts.

ModelArts utiliza Identity and Access Management (IAM) para la mayoría de los permisos de las funciones de gestión. Antes de leer a continuación, conozca los **Conceptos básicos**. Esto le ayudará a comprender mejor este documento.

Para implementar una gestión de permisos detallada, ModelArts proporciona control de permisos, autorización de agencias y espacio de trabajo. A continuación se describen los detalles.

Permisos y delegaciones de ModelArts

Figura 7-1 Gestión de permisos



Las funciones expuestas de ModelArts se controlan mediante permisos de IAM. Por ejemplo, si usted, como usuario de IAM, necesita crear un trabajo de entrenamiento en ModelArts, debe tener el permiso **ModelArts:trainJob:create**. Para obtener detalles sobre cómo asignar permisos a un usuario (necesita agregar el usuario a un grupo de usuarios y luego asignar los permisos a este grupo), consulte **Gestión de permisos**.

ModelArts debe acceder a otros servicios para el cómputo de IA. Por ejemplo, ModelArts debe acceder al OBS para leer los datos para el entrenamiento. Por razones de seguridad, ModelArts debe estar autorizado para acceder a otros servicios en la nube. Esta es la autorización de la delegación.

A continuación se presenta un resumen de la gestión de permisos:

- El acceso a cualquier servicio en la nube se controla con IAM. Debe tener los permisos del servicio en la nube. (Los permisos de servicios requeridos varían según las funciones que utilice)
- Para utilizar las funciones de ModelArts, debe otorgar permisos con IAM.
- ModelArts debe estar autorizado por usted para acceder a otros servicios en la nube para el cómputo de IA.

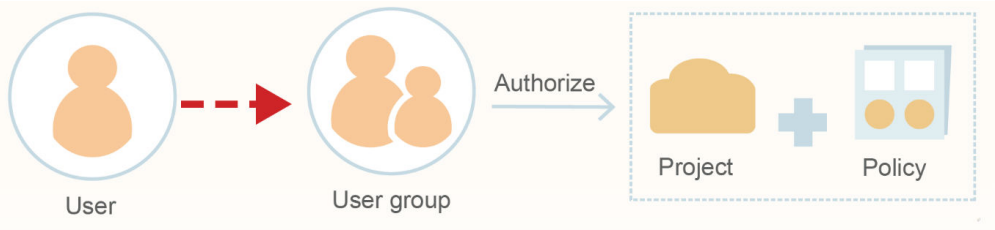
Gestión de permisos de ModelArts

De forma predeterminada, los nuevos usuarios de IAM no tienen ningún permiso asignado. Es necesario agregar el usuario a un grupo de usuarios y otorgar políticas al grupo de usuarios, de modo que los usuarios del grupo puedan heredar los permisos. Después de la autorización, los usuarios pueden realizar operaciones de ModelArts basadas en permisos.

⚠ ATENCIÓN

ModelArts es un servicio a nivel de proyecto desplegado y al que se accede en regiones físicas específicas. Cuando autoriza una delegación, puede establecer el alcance de los permisos seleccionados para todos los recursos, proyectos empresariales o proyectos específicos de la región. Si especifica proyectos específicos para cada región, los permisos seleccionados se aplicarán a los recursos de estos proyectos.

Para obtener más detalles, consulte **Creación de un grupo de usuarios y asignación de permisos**.



Al asignar permisos a un grupo de usuarios, IAM no asigna directamente los permisos específicos al grupo de usuarios. En su lugar, IAM debe agregar los permisos a una política y luego asignar la política al grupo de usuarios. Para facilitar la gestión de los permisos de usuarios, cada servicio en la nube ofrece algunas políticas preestablecidas para que usted las utilice directamente. Si las políticas preestablecidas no pueden satisfacer sus requisitos de gestión de permisos de grano fino, puede personalizar las políticas.

Tabla 7-1 enumera todas las políticas preestablecidas definidas por el sistema que ModelArts admite.

Tabla 7-1 Políticas definidas por el sistema compatibles con ModelArts

Política	Descripción	Tipo
ModelArts FullAccess	Administrator permissions for ModelArts. Los usuarios a los que se les conceden estos permisos pueden operar y usar ModelArts.	Política definida por el sistema
ModelArts CommonOperations	Permisos de usuario comunes para ModelArts. Los usuarios a los que se conceden estos permisos pueden operar y usar ModelArts pero no pueden gestionar grupos de recursos dedicados.	Política definida por el sistema
ModelArts Dependency Access	Permisos en los servicios dependientes para ModelArts	Política definida por el sistema

Por lo general, ModelArts FullAccess solo se asigna a los administradores. Si no se requiere una gestión detallada, asignar ModelArts CommonOperations a todos los usuarios cumplirá con los requisitos de desarrollo de la mayoría de los equipos pequeños. Si desea personalizar políticas para una gestión detallada, consulte **IAM**.

> **NOTA**
>
> Cuando se asigna permisos de ModelArts a un usuario, el sistema no asigna automáticamente los permisos de otros servicios al usuario. Esto garantiza la seguridad y evita las operaciones no autorizadas inesperadas. En este caso, sin embargo, debe asignar por separado permisos de diferentes servicios a los usuarios para que puedan realizar algunas operaciones de ModelArts.
>
> Por ejemplo, si un usuario de IAM necesita utilizar datos de OBS para entrenamiento y el permiso de entrenamiento de ModelArts ha sido configurado para el usuario de IAM, el usuario de IAM aún necesita ser asignado con los permisos de OBS de lectura, escritura y lista. El permiso de lista de OBS permite seleccionar la ruta de datos de entrenamiento en ModelArts. El permiso de lectura se utiliza para obtener una vista previa de los datos y leer datos para el entrenamiento. El permiso de escritura se utiliza para guardar los resultados y logs del entrenamiento.
>
> - Para usuarios individuales o pequeñas organizaciones, es una buena práctica configurar la política **Tenant Administrator** que se aplica a los servicios globales para usuarios de IAM. De esta manera, los usuarios de IAM pueden obtener todos los permisos de usuario excepto IAM. Sin embargo, esto puede causar problemas de seguridad. (Para un usuario individual, su usuario de IAM predeterminado pertenece al grupo de usuarios **admin** y tiene el permiso **Tenant Administrator**.)
> - Si desea restringir las operaciones de usuarios, configure los permisos mínimos de OBS para los usuarios de ModelArts. Para obtener más detalles, consulte **Gestión de permisos de OBS**. Para obtener más información sobre la gestión de permisos detallados de otros servicios en la nube, consulte los documentos de los servicios en la nube correspondientes.

Autorización de delegación de ModelArts

ModelArts debe estar autorizado por los usuarios para acceder a otros servicios en la nube para el cómputo de IA. En el sistema de permisos de IAM, dicha autorización se realiza con delegaciones.

Para obtener más información sobre los conceptos básicos y las operaciones de delegaciones, consulte **Delegación de servicios en la nube**.

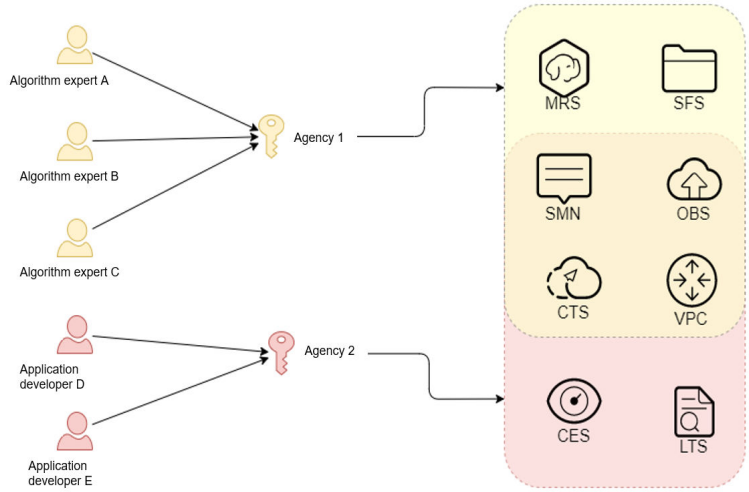
Para simplificar la autorización de delegaciones, ModelArts admite la configuración automática de autorizaciones de delegaciones. Solo necesita configurar una delegación para usted o para los usuarios especificados en la página **Global Configuration** de la consola de ModelArts.

> **NOTA**
>
> - Solo los usuarios con permiso de gestión de agencias de IAM pueden realizar esta operación. Por lo general, los miembros del grupo de usuarios admin de IAM tienen este permiso.
> - La autorización de delegación de ModelArts es específica de cada región, lo que significa que debe realizar la autorización de delegación en cada región que utilice.

En la página **Global Configuration** de la consola de ModelArts, después de hacer clic en **Add Authorization**, puede configurar una delegación para un usuario específico o para todos los usuarios. Por lo general, se crea una delegación denominada **modelarts_agency_<*Username*>_*Random ID*** de forma predeterminada. En el área **Permissions**, puede seleccionar la configuración de permisos preestablecida o seleccionar las políticas necesarias. Si ambas opciones no satisfacen sus requisitos, puede crear una delegación en la página de gestión de IAM (necesita delegar ModelArts para acceder a sus recursos) y luego usar una delegación existente en lugar de agregar una delegación en la página **Add Authorization**.

ModelArts asocia varios usuarios con una delegación. Esto significa que si dos usuarios necesitan configurar la misma delegación, no es necesario crear una delegación para cada usuario. En su lugar, solo necesita configurar la misma delegación para los dos usuarios.

Figura 7-2 Mapeo entre usuarios y agencias



NOTA

Cada usuario puede utilizar ModelArts solo después de asociarse con una delegación. Sin embargo, aunque los permisos asignados a la delegación sean insuficientes, no se reporta ningún error cuando se invoca a la API. Solo se produce un error cuando el sistema utiliza funciones no autorizadas. Por ejemplo, puede activar la notificación de mensajes al crear un trabajo de entrenamiento. La notificación de mensajes requiere autorización de SMN. Sin embargo, solo se produce un error cuando es necesario enviar mensajes para el trabajo de entrenamiento. El sistema ignora algunos errores y otros pueden causar fallas en el trabajo. Cuando implemente la minimización de permisos, asegúrese de que aún tendrá suficientes permisos para las operaciones requeridas en ModelArts.

Autorización estricta

En el modo de autorización estricta, se requiere la autorización explícita por el administrador de la cuenta para que los usuarios de IAM puedan acceder a ModelArts. El administrador puede agregar los permisos de ModelArts requeridos a los usuarios comunes mediante las políticas de autorización.

En el modo de autorización no estricta, los usuarios de IAM pueden usar ModelArts sin autorización explícita. El administrador necesita configurar la política de denegación para usuarios de IAM para evitar que utilicen algunas funciones de ModelArts.

El administrador puede cambiar el modo de autorización en la página **Global Configuration**.

AVISO

Se recomienda el modo de autorización estricta. En este modo, los usuarios de IAM deben estar autorizados a utilizar funciones de ModelArts. De este modo, se puede controlar con precisión el alcance de los permisos de los usuarios de IAM, minimizando los permisos concedidos a los usuarios de IAM.

Gestión del acceso a los recursos con espacios de trabajo

El espacio de trabajo permite a los clientes empresariales dividir sus recursos en varios espacios lógicamente aislados y gestionar el acceso a los distintos espacios. Como usuario de

empresa, puede enviar la solicitud de habilitación de la función de espacio de trabajo a su responsable de asistencia técnica.

Una vez que se habilita el espacio de trabajo, se crea un espacio de trabajo predeterminado. Todos los recursos que ha creado se encuentran en este espacio de trabajo. Un espacio de trabajo es como un gemelo de ModelArts. Puede cambiar de un espacio de trabajo a otro en la esquina superior izquierda de la consola de ModelArts. Los trabajos en espacios de trabajo diferentes no se afectan entre sí.

Al crear un espacio de trabajo, debe vincularlo a un proyecto empresarial. Pueden vincularse múltiples espacios de trabajo a un mismo proyecto empresarial, pero no puede vincularse un espacio de trabajo a varios proyectos empresariales. Puede utilizar espacios de trabajo para refinar las restricciones de acceso a los recursos y los permisos de los distintos usuarios. Las restricciones son las siguientes:

- Los usuarios deben estar autorizados para acceder a espacios de trabajo específicos (esto debe configurarse en las páginas de creación y gestión de espacios de trabajo). Esto significa que el acceso a activos de IA, como conjuntos de datos y algoritmos, puede gestionarse mediante espacios de trabajo.
- En las operaciones de autorización de permisos anteriores, si establece el ámbito en proyectos de empresa, la autorización solo tendrá efecto para los espacios de trabajo vinculados a los proyectos seleccionados.

NOTA

- Las restricciones sobre los espacios de trabajo y la autorización de permisos surten efecto al mismo tiempo. Es decir, un usuario debe tener los permisos para acceder al espacio de trabajo y para crear trabajos de entrenamiento (el permiso se aplica a este espacio de trabajo) para que el usuario pueda enviar trabajos de entrenamiento en este espacio de trabajo.
- Si ha habilitado un proyecto de empresa pero no ha habilitado un espacio de trabajo, todas las operaciones se realizarán en el proyecto de empresa predeterminado. Asegúrese de que los permisos de las operaciones requeridas se aplican al proyecto de empresa por defecto.
- Las restricciones anteriores no se aplican a los usuarios que no hayan habilitado ningún proyecto de empresa.

Resumen

Características principales de la gestión de permisos de ModelArts:

- Si usted es un usuario individual, no es necesario que considere la gestión detallada de permisos. Su cuenta tiene todos los permisos para utilizar ModelArts de forma predeterminada.
- IAM controla todas las funciones de ModelArts. Puede utilizar la autorización de IAM para implementar una gestión detallada de permisos para los usuarios específicos.
- Todos los usuarios (incluidos los usuarios individuales) pueden usar funciones específicas solo después de la autorización de la delegación en ModelArts (**Settings** > **Add Authorization**). De lo contrario, pueden producirse errores inesperados.
- Si ha habilitado la función de proyecto empresarial, también puede habilitar el espacio de trabajo de ModelArts y utilizar tanto la autorización básica como el espacio de trabajo para una gestión de permisos refinada.

8 Seguridad

8.1 Responsabilidades compartidas

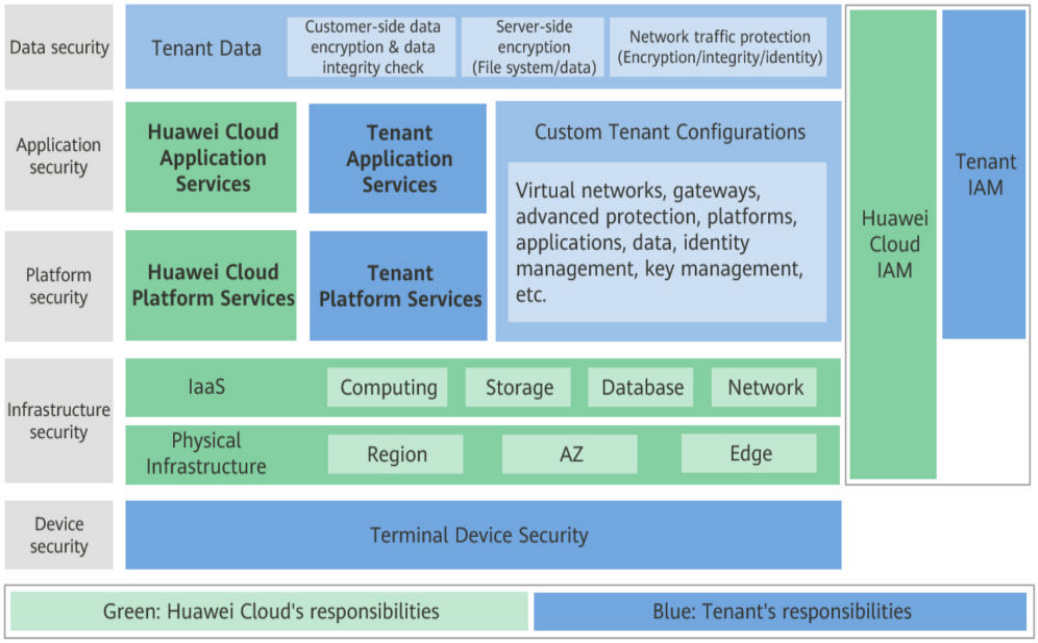
Huawei garantiza que su compromiso con la seguridad cibernética nunca se verá compensado por la consideración de intereses comerciales. Para hacer frente a los desafíos emergentes de seguridad en la nube y a las amenazas y ataques generalizados a la seguridad en la nube, Huawei Cloud construye un sistema integral de garantía de seguridad de servicios en la nube para diferentes regiones e industrias basado en las ventajas únicas de software y hardware de Huawei, las leyes, las regulaciones, los estándares de la industria y el ecosistema de seguridad.

Figura 8-1 ilustra las responsabilidades compartidas por Huawei Cloud y los usuarios.

- **Huawei Cloud**: Garantizar la seguridad de los servicios en la nube y proporcionar nubes seguras. Las responsabilidades de seguridad de Huawei Cloud incluyen garantizar la seguridad de nuestros servicios de IaaS, PaaS y SaaS, así como los entornos físicos de los centros de datos de Huawei Cloud, donde nuestros servicios de IaaS, PaaS, y SaaS operan. Huawei Cloud es responsable no solo de las funciones de seguridad y el rendimiento de nuestra infraestructura, servicios en la nube y tecnologías, sino también de la seguridad general de la nube y, en el sentido más amplio, de la certificación de seguridad de nuestra infraestructura y servicios.
- **Tenant**: Utilizar la nube de forma segura. Los tenants de Huawei Cloud son responsables de la gestión segura y efectiva de las configuraciones personalizadas por el tenant de los servicios en la nube, incluidos IaaS, PaaS y SaaS. Esto incluye, entre otros, redes virtuales, el sistema operativo de los hosts e invitados de máquinas virtuales, firewalls virtuales, API Gateway, servicios de seguridad avanzados, todo tipo de servicios en la nube, datos del tenant, cuentas de identidad, y gestión de claves.

Libro blanco de seguridad de Huawei Cloud elabora las ideas y medidas para construir la seguridad en Huawei Cloud, incluidas las estrategias de seguridad en la nube, el modelo de responsabilidad compartida, el cumplimiento y la privacidad, las organizaciones y el personal de seguridad, la seguridad de la infraestructura, el servicio y la seguridad del tenant, la seguridad de ingeniería, seguridad de O&M y seguridad del ecosistema.

Figura 8-1 Modelo de responsabilidad de seguridad compartida de Huawei Cloud



8.2 Identificación y gestión de activos

Identificación de activos

Sus activos en AI Gallery incluyen sus activos de IA publicados y su información personal.

Los activos de IA incluyen, entre otros, textos, gráficos, datos, artículos, fotos, imágenes, ilustraciones, código, algoritmos de IA y modelos de IA.

Su información personal incluye:

- Apodo, foto de perfil y correo electrónico para registrar la cuenta
- Nombre, número de celular y correo electrónico para participar en las prácticas
- Información empresarial para convertirse en socio
- Nombre de contacto, número de celular y correo electrónico para publicar activos

Gestión de activos

AI Gallery gestiona de forma centralizada los activos publicados por los usuarios.

- AI Gallery almacena activos de archivos en bucket oficiales de OBS.
- AI Gallery almacena activos de imagen en repositorios de SWR oficiales.

AI Gallery almacena la información personal de los usuarios en bases de datos. AI Gallery cifra información personal sensible, como números de celular y correos electrónicos, en bases de datos.

Para obtener más información sobre AI Gallery, consulte **AI Gallery**.

8.3 Autenticación de identidad y control de acceso

Autenticación de identidad

Puede utilizar los servicios de ModelArts con la consola, las API o los SDK. Esencialmente, las solicitudes de acceso se envían con las API de REST de ModelArts.

Se puede acceder a las API de ModelArts tras la autenticación exitosa. Las solicitudes enviadas con la consola se pueden autenticar mediante tokens y las solicitudes de invocaciones a las API se pueden autenticar mediante tokens o AK/SK. Consulte **Autenticación** para obtener más detalles.

Control de acceso

ModelArts le permite configurar permisos de grano fino para una gestión refinada de recursos y permisos. Para ello, ModelArts proporciona control de permisos de IAM, autorización de agencias y espacio de trabajo.

- Control de permisos de IAM

  Para utilizar las funciones de ModelArts, debe otorgar permisos con IAM. Por ejemplo, si necesita crear un trabajo de entrenamiento para ModelArts, debe obtener el permiso **ModelArts:trainJob:create**.

  Si no se configura una política de autorización detallada para un usuario creado por el administrador, el usuario tiene todos los permisos de ModelArts por defecto. Para controlar los permisos de usuario, el administrador debe agregar al usuario a un grupo de usuarios en IAM y configurar políticas de autorización detalladas para este grupo. De esta manera, el usuario obtiene los permisos definidos en las políticas antes de realizar operaciones sobre los recursos del servicio en la nube. Durante la autorización basada en políticas, el administrador puede seleccionar el ámbito de autorización en función de los tipos de recursos de ModelArts. Para obtener más detalles sobre los permisos de recurso, consulte **Políticas de permisos y acciones admitidas**.

- Autorización de delegación

  ModelArts necesita acceder a otros servicios para la computación de IA. Por ejemplo, ModelArts debe acceder a OBS para leer los datos para el entrenamiento. Por razones de seguridad, ModelArts debe estar autorizado para acceder a otros servicios en la nube. Esta es la autorización de la delegación.

  ModelArts no guarda sus credenciales de autenticación de token. Antes de realizar operaciones en sus recursos (como buckets de OBS) en un trabajo de backend, es necesario que autorice explícitamente a ModelArts con una delegación de IAM. ModelArts utilizará la delegación para obtener una credencial de autenticación temporal para realizar operaciones con sus recursos. Para obtener más detalles, consulte **Configuración de la autorización de acceso (Configuración global)**.

- Espacio de trabajo

  Espacio de trabajo permite a los clientes que han habilitado los **proyectos empresariales** a dividir sus recursos en múltiples espacios aislados lógicamente y controlar el acceso a diferentes espacios.

  Una vez que se habilita el espacio de trabajo, se crea un espacio de trabajo predeterminado. Todos los recursos que ha creado se encuentran en este espacio de

trabajo. Un espacio de trabajo es como un gemelo de ModelArts. Puede pasar de un espacio de trabajo a otro en la esquina superior izquierda del panel de navegación. Los trabajos en diferentes espacios de trabajo no se afectan entre sí. ModelArts le permite crear múltiples espacios de trabajo para desarrollar algoritmos y gestionar y desplegar modelos para diferentes objetivos de servicio. De este modo, los resultados del desarrollo de las distintas aplicaciones se gestionan en diferentes espacios de trabajo para su uso.

Gestión de acceso remoto

Cuando utiliza un IDE local para acceder remotamente al entorno de desarrollo de notebook de ModelArts con SSH, se requiere el par de claves para la autenticación. También puede agregar a la lista blanca las direcciones IP para acceder de forma remota a la instancia de notebook.

8.4 Protección de datos

ModelArts toma diferentes medidas para mantener los datos almacenados en ModelArts seguros y fiables.

Medida	Descripción
Protección de datos estáticos	AI Gallery cifra información personal sensible, como números de celular y correos electrónicos, en bases de datos. Se usa el algoritmo de encriptación AES.
Protección para la transmisión de datos	Al importar aplicaciones de IA en ModelArts, está compatible con HTTP y HTTPS, pero se recomienda HTTPS para una transmisión de datos más segura.
Comprobación de integridad de datos	Al cargar archivos de modelo o activos de AI Gallery para el despliegue de inferencia, los datos pueden ser incoherentes debido al secuestro de la red, el almacenamiento en caché y otras razones. ModelArts verifica la coherencia de los datos calculando el valor de SHA256 cuando se cargan o se descargan datos.
Mecanismo de aislamiento de datos	Cuando se crea una instancia de bloc de notas, se aísla el almacenamiento de datos de los distintos tenants, de modo que estos no puedan ver los datos de otros tenants.

8.5 Auditoría y registro

Auditoría

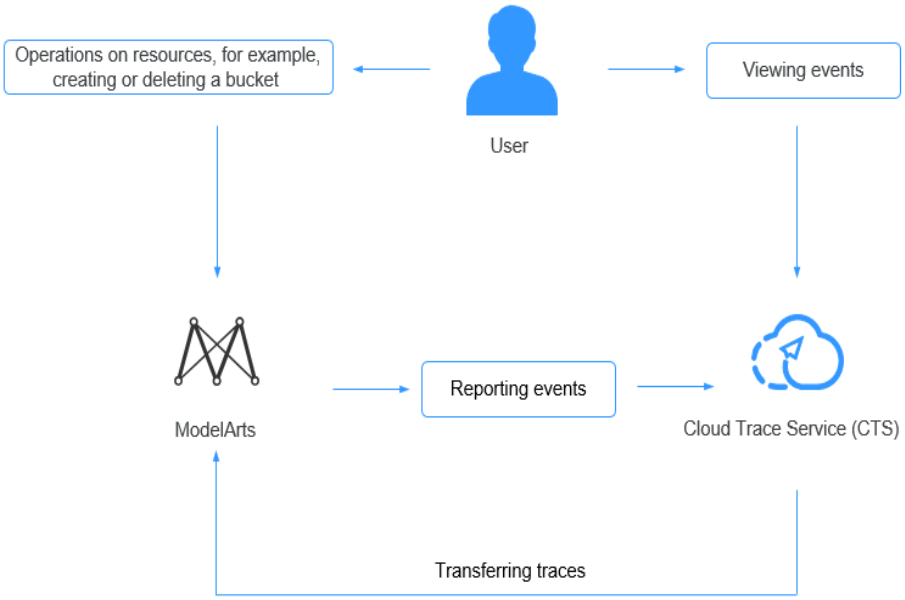
Cloud Trace Service (CTS) registra las operaciones en los recursos de la nube de su cuenta. Puede utilizar los registros generados por CTS para realizar análisis de seguridad, rastrear cambios en los recursos, auditar el cumplimiento y localizar fallos.

Después de habilitar CTS y configurar un seguidor, CTS puede registrar la gestión y el seguimiento de datos de ModelArts para su auditoría.

Para obtener más información acerca de cómo habilitar y configurar CTS, consulte **Habilitación de CTS**.

Para obtener detalles sobre la gestión de ModelArts y los seguimientos de datos que CTS puede rastrear, consulte **Operaciones clave registradas para la gestión de datos**, **Operaciones clave de DevEnviron registradas por CTS**, **Operaciones clave de trabajo de entrenamiento registradas por CTS**, **Operaciones clave de gestión de aplicaciones de IA registradas por CTS** y **Operaciones clave de gestión de servicios registradas por CTS**.

Figura 8-2 CTS



Operaciones clave de gestión de datos registradas por CTS

Tabla 8-1 Operaciones clave de gestión de datos registradas por CTS

Operación	Tipo de recurso	Traza
Creación de un conjunto de datos	dataset	createDataset
Eliminación de un conjunto de datos	dataset	deleteDataset
Actualización de un conjunto de datos	dataset	updateDataset
Publicación de una versión de conjunto de datos	dataset	publishDatasetVersion
Eliminación de una versión de conjunto de datos	dataset	deleteDatasetVersion
Sincronización del origen de datos	dataset	syncDataSource
Exportación de un conjunto de datos	dataset	exportDataFromDataset

Operación	Tipo de recurso	Traza
Creación de una tarea de etiquetado automático	dataset	createAutoLabelingTask
Creación de una tarea de agrupación automática	dataset	createAutoGroupingTask
Creación de una tarea de despliegue automática	dataset	createAutoDeployTask
Importación de muestras a un conjunto de datos	dataset	importSamplesToDataset
Creación de una etiqueta de conjunto de datos	dataset	createLabel
Modificación de una etiqueta de conjunto de datos	dataset	updateLabel
Eliminación de una etiqueta de conjunto de datos	dataset	deleteLabel
Eliminación de una etiqueta de conjunto de datos y las muestras correspondientes	dataset	deleteLabelWithSamples
Adición de muestras	dataset	uploadSamples
Supresión de muestras	dataset	deleteSamples
Detención de una tarea de etiquetado automático	dataset	stopTask
Creación de un trabajo de etiquetado de equipo	dataset	createWorkforceTask
Eliminación de un trabajo de etiquetado de equipo	dataset	deleteWorkforceTask
Inicio de la aceptación del etiquetado del equipo	dataset	startWorkforceSampling-Task
Aprobación/rechazo/cancelación de la aceptación	dataset	updateWorkforceSam-plingTask
Envío de comentarios de revisión de muestras para su aceptación	dataset	acceptSamples
Adición de una etiqueta a una muestra	dataset	updateSamples
Envío de un correo electrónico a los miembros del equipo de etiquetado	dataset	sendEmails
Inicio de trabajo de etiquetado de equipo como persona de contacto	dataset	startWorkforceTask

Operación	Tipo de recurso	Traza
Actualización de un trabajo de etiquetado de equipo	dataset	updateWorkforceTask
Adición de una etiqueta a una muestra con etiqueta de equipo	dataset	updateWorkforceTask-Samples
Revisión de los resultados del etiquetado del equipo	dataset	reviewSamples
Creación de un miembro del equipo de etiquetado	workforce	createWorker
Actualización de un miembro del equipo de etiquetado	workforce	updateWorker
Eliminación de un miembro del equipo de etiquetado	workforce	deleteWorker
Eliminación por lotes de los miembros del equipo de etiquetado	workforce	batchDeleteWorker
Creación de un equipo de etiquetado	workforce	createWorkforce
Actualización de un equipo de etiquetado	workforce	updateWorkforce
Eliminación de un equipo de etiquetado	workforce	deleteWorkforce
Creación automática de una delegación de IAM	IAM	createAgency
Inicio de sesión en la consola de etiquetado como miembro del equipo de etiquetado	labelConsoleWorker	workerLoginLabelConsole
Cierra de sesión en la consola de etiquetado como miembro del equipo de etiquetado	labelConsoleWorker	workerLogOutLabelConsole
Cambio de la contraseña de la consola de etiquetado como miembro del equipo de etiquetado	labelConsoleWorker	workerChangePassword
Olvidar la contraseña de la consola de etiquetado como miembro del equipo de etiquetado	labelConsoleWorker	workerForgetPassword
Restablecimiento de la contraseña de la consola de etiquetado a través del URL como miembro del equipo de etiquetado	labelConsoleWorker	workerResetPassword

Operaciones clave de DevEnviron registradas por CTS

Tabla 8-2 Operaciones clave de DevEnviron registradas por CTS

Operación	Tipo de recurso	Nombre del rastro
Creación de una instancia de notebook	Notebook	createNotebook
Supresión de una instancia de notebook	Notebook	deleteNotebook
Apertura de una instancia de notebook	Notebook	openNotebook
Inicio de una instancia de notebook	Notebook	startNotebook
Stopping a notebook instance	Notebook	stopNotebook
Actualización de una instancia de notebook	Notebook	updateNotebook
Eliminación de un NotebookApp	NotebookApp	deleteNotebookApp
Cambio de especificaciones de CodeLab	NotebookApp	updateNotebookApp

Operaciones clave de trabajo de entrenamiento registradas por CTS

Tabla 8-3 Operaciones clave de trabajo de entrenamiento registradas por CTS

Operación	Tipo de recurso	Traza
Creación de un trabajo de entrenamiento	ModelArtsTrainJob	createModelArtsTrainJob
Creación de una versión de trabajo de entrenamiento	ModelArtsTrainJob	createModelArtsTrainVersion
Detención de un trabajo de entrenamiento	ModelArtsTrainJob	stopModelArtsTrainVersion
Modificación de la descripción de un trabajo de entrenamiento	ModelArtsTrainJob	updateModelArtsTrainDesc
Eliminación de una versión de trabajo de entrenamiento	ModelArtsTrainJob	deleteModelArtsTrainVersion
Supresión de un trabajo de entrenamiento	ModelArtsTrainJob	deleteModelArtsTrainJob

Operación	Tipo de recurso	Traza
Creación de una configuración de trabajo de entrenamiento	ModelArtsTrainConfig	createModelArtsTrainConfig
Modificación de la configuración de un trabajo de entrenamiento	ModelArtsTrainConfig	updateModelArtsTrainConfig
Supresión de una configuración de trabajo de entrenamiento	ModelArtsTrainConfig	deleteModelArtsTrainConfig
Creación de un trabajo de visualización	ModelArtsTensorboardJob	createModelArtsTensorboardJob
Supresión de un trabajo de visualización	ModelArtsTensorboardJob	deleteModelArtsTensorboardJob
Modificación de la descripción de un trabajo de visualización	ModelArtsTensorboardJob	updateModelArtsTensorboardDesc
Detención de un trabajo de visualización	ModelArtsTensorboardJob	stopModelArtsTensorboardJob
Reinicio de un trabajo de visualización	ModelArtsTensorboardJob	restartModelArtsgTensorboardJob

Operaciones clave de gestión de aplicaciones de IA registradas por CTS

Tabla 8-4 Operaciones clave de gestión de aplicaciones de IA registradas por CTS

Operación	Tipo de recurso	Traza
Creación de una aplicación de IA	model	addModel
Actualización de una aplicación de IA	model	updateModel
Eliminación de una aplicación de IA	model	deleteModel
Creación de una tarea de conversión de modelo	convert	addConvert
Actualización de una tarea de conversión de modelo	convert	updateConvert
Eliminación de una tarea de conversión de modelo	convert	deleteConvert

Operaciones clave de gestión de servicios registradas por CTS

Tabla 8-5 Operaciones clave de gestión de servicios registradas por CTS

Operación	Tipo de recurso	Traza
Despliegue de un servicio	service	addService
Eliminación de un servicio	service	deleteService
Actualización de un servicio	service	updateService
Inicio o detención de un servicio	service	startOrStopService
Adición de una clave de acceso	service	addAkSk
Eliminación de una clave de acceso	service	deleteAkSk
Creación de un grupo de recursos dedicado	cluster	createCluster
Eliminación de un grupo de recursos dedicado	cluster	deleteCluster
Adición de un nodo a un grupo de recursos dedicado	cluster	addClusterNode
Eliminar un nodo de un grupo de recursos dedicado	cluster	deleteClusterNode
Obtención de un resultado de la creación del grupo de recursos dedicados	cluster	createClusterResult

Operaciones clave de AI Gallery registradas por CTS

Tabla 8-6 Operaciones clave de AI Gallery registradas por CTS

Operación	Tipo de recurso	Traza
Publicación de un activo	ModelArts_Market	create_content
Modificación de la información de activos	ModelArts_Market	modify_content
Publicación de una versión de activos	ModelArts_Market	add_version
Suscripción a un activo	ModelArts_Market	subscription_content
Extracción de un activo de los favoritos	ModelArts_Market	cancel_star_content
Dar Me gusta a un activo	ModelArts_Market	like_content
Cancelar Me gusta a un activo	ModelArts_Market	cancel_like_content

Operación	Tipo de recurso	Traza
Publicación de una actividad	ModelArts_Market	publish_activity
Registro de una actividad	ModelArts_Market	regist_activity
Modificación de la información de usuario	ModelArts_Market	update_user

Registro

Puede habilitar ModelArts para realizar análisis o auditorías. Una vez que se habilita CTS, CTS inicia las operaciones de grabación en ModelArts. La consola de gestión de CTS almacena los últimos siete días de registros de operación. Esta sección describe cómo consultar los registros de operaciones de los últimos 7 días en la consola de gestión de CTS.

Para obtener detalles sobre cómo consultar logs de auditoría en CTS, véase **Consulta de logs de auditoría**.

8.6 Resiliencia del servicio

La resiliencia se refiere a la resiliencia de seguridad de los servicios en la nube después de los ataques, excluidas la confiabilidad y la disponibilidad. Este capítulo describe las capacidades de ModelArts de defensa y detección contra intrusiones, defensa contra jitter, uso adecuado de nombres de dominio y detección de seguridad de contenidos.

Paquete de seguridad y host bastión en la nube para mejorar la defensa y la detección contra intrusiones

Paquete de seguridad se han desplegado en ModelArts en las capas de host, de aplicación, de red y de datos detectar rápidamente las intrusiones.

- ModelArts utiliza componentes web de seguridad para prevenir los riesgos de seguridad de las aplicaciones web desplegadas en él y utiliza WAF para la protección de la seguridad.
- Host Security Service (HSS) se han desplegado en todos los hosts que transportan servicios de ModelArts. Estos productos incluyen, pero no se limitan a HSS desarrollado por Huawei y Compute Security Platform (CSP).
- Vulnerability Scan Service (VSS) se ha desplegado en ModelArts y realiza exploraciones rutinarias para detectar y corregir rápidamente las vulnerabilidades.
- ModelArts realiza O&M de seguridad en los recursos de la nube con una plataforma de gestión de seguridad.
- Situation Awareness (SA) se ha desplegado en ModelArts para comprender la situación de seguridad, consultar historiales de ataques y detectar rápidamente los riesgos de cumplimiento y responder a las alarmas de amenaza.
- Advanced Anti-DDoS (AAD) se ha desplegado en las EIP que transportan servicios clave de ModelArts para evitar tormentas de tráfico.
- Database Security Service (DBSS) se ha desplegado en las bases de datos de ModelArts que almacenan datos importantes.

Políticas de prevención de fluctuaciones y respuesta de emergencia y restauración frente a ataques

ModelArts aísla los recursos de diferentes tenants para que los ataques a los recursos de un tenant no afecten los recursos de otros.

- ModelArts ofrece grupos de recursos dedicados aislados físicamente para que los ataques a los recursos de un tenant no afecten los recursos de otros.
- ModelArts define y mantiene sus especificaciones de rendimiento para defenderse de los ataques, por ejemplo, configurando el control del tráfico en el acceso a la API.
- ModelArts proporciona informes de alarmas y autoprotección contra ataques.
- ModelArts detecta comportamientos anormales de los servicios, por ejemplo, detectando datos anormales de la plataforma de operaciones e integrando registros de seguridad.
- ModelArts proporciona control de riesgos y respuesta de emergencia ante ataques. Por ejemplo, ModelArts identifica rápidamente los tenants y las direcciones IP maliciosas.
- ModelArts restaura rápidamente los servicios después de que cesan los ataques de tráfico.

Especificaciones de uso de nombres de dominio y políticas de seguridad de contenido de tenants de servicios en la nube

Los nombres de dominio de ModelArts cumplen con ciertos requisitos de seguridad para evitar riesgos de cumplimiento y ataques de suplantación de identidad (phishing).

Nombres de dominio visibles para los tenants: nombres de dominio accesibles para los tenants, que requieren mayor atención a la seguridad y el cumplimiento.

Nombres de dominio invisibles para los tenants: nombres de dominio utilizados por los servicios de Huawei Cloud para llamarse entre sí en la intranet, en cuyo caso los usuarios externos no pueden acceder a los servidores de DNS autoritativos; o nombres de dominio a los que sólo pueden acceder los empleados de Huawei, el personal de los socios y el personal subcontratado en las zonas amarilla y verde a través de la red de oficinas de Huawei (es decir, no se puede acceder a estos nombres de dominio con Internet).

- Los nombres de dominio básicos de Huawei Cloud no se asignan directamente a los tenants, sino que se utilizan de manera segura.
- Los nombres de dominio externos para los que se ha concedido una licencia no son utilizados por los servicios de Huawei Cloud para llamarse entre sí en la intranet.

8.7 Monitoreo de riesgos

ModelArts monitorea automáticamente sus servicios y las cargas de modelos en tiempo real y gestiona alarmas y notificaciones, para que pueda realizar un seguimiento del rendimiento de los servicios y de los modelos. Para obtener más detalles, consulte **Métricas de ModelArts**.

8.8 Recuperación de fallas

La infraestructura global de ModelArts está diseñada para las regiones y zonas de disponibilidad (AZ) de Huawei Cloud. Una región en Huawei Cloud ofrece múltiples AZ físicamente independientes y aisladas que se conectan con redes con baja demora, alto

rendimiento y alta redundancia. Puede diseñar y operar aplicaciones y bases de datos con fallas que se migran automáticamente entre las AZ sin interrumpir los servicios. En comparación con la infraestructura tradicional de un singular centro de datos o múltiples centros de datos, las AZ ofrecen mayor disponibilidad, tolerancia a fallas y escalabilidad.

ModelArts realiza copias de seguridad de los datos de su base de datos para recuperarlos en caso de falla del servicio o de daños en los datos originales.

Recuperación del entorno de fallas

Si un nodo de cómputo utilizado por una instancia de notebook presenta fallas, la instancia se migrará automáticamente a otro nodo disponible. Luego, se restaura la instancia. ModelArts permite montar un disco de EVS en una instancia. Huawei Cloud EVS ofrece almacenamiento en bloques escalable con alta confiabilidad, alto rendimiento y una variedad de especificaciones para servidores. La durabilidad de los datos alcanza el 99.9999999%.

Recuperación automática de un fallo de entrenamiento

Durante el entrenamiento del modelo, es posible que se produzca una falla en el entrenamiento debido a una falla de hardware. En caso de fallas de hardware, ModelArts ofrece verificación de tolerancia a fallas para aislar los nodos defectuosos y mejorar la experiencia del usuario durante el entrenamiento.

La verificación de tolerancia a fallas incluye una verificación previa del entorno y una verificación periódica del hardware. Si se detecta alguna falla durante cualquiera de las comprobaciones, ModelArts aísla automáticamente el hardware defectuoso y vuelve a emitir el trabajo de entrenamiento. En el entrenamiento distribuido, la verificación de tolerancia a fallas se realizará en todos los nodos de cómputo utilizados por el trabajo de entrenamiento.

Recuperación de una falla de despliegue de la inferencia

Durante la ejecución del servicio, si una instancia de inferencia presenta fallas debido a un error de hardware, ModelArts detecta automáticamente la falla y migra esa instancia defectuosa a otro nodo disponible. Después de reiniciar la instancia, se restaurará. El nodo defectuoso se aísla automáticamente y no se programa para ejecutar instancias de inferencia.

8.9 Gestión de actualizaciones

Actualización del servicio en tiempo real de ModelArts

Para un servicio desplegado, puede cambiar la versión de la aplicación de IA para actualizarla.

Los servicios se pueden actualizar en tres modos: actualización completa, actualización continua (aumento de instancias) y actualización continua (disminución de instancias). Consulte **Figura 8-3** para obtener detalles sobre los tres modos de actualización.

- Actualización completa

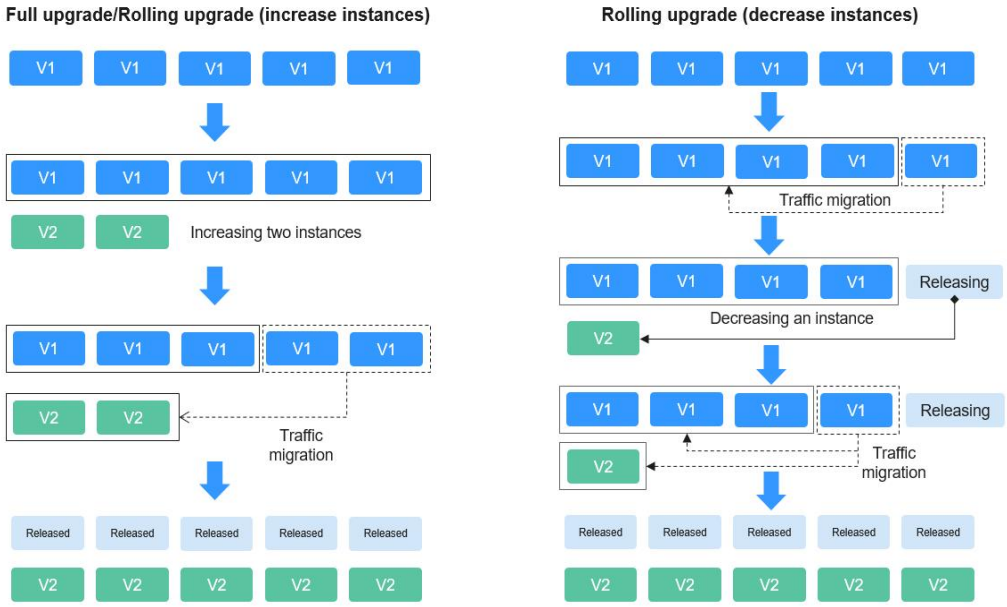
  Los recursos que sean el doble de los utilizados por el servicio se utilizarán para crear instancias de nueva versión en modo completo.
- Actualización continua (aumento de instancias)

Se utilizarán recursos adicionales a los utilizados por el servicio para una actualización continua. Un mayor número de instancias que se van a aumentar llevará a una actualización más rápida.

- Actualización continua (disminución de instancias)

  Algunos nodos destinados a ejecutar servicios se utilizarán para una actualización continua. Un mayor número de instancias a reducir conllevará una actualización más rápida, pero una mayor probabilidad de interrupción del servicio.

Figura 8-3 Proceso de actualización del servicio



Para obtener detalles sobre cómo actualizar un servicio de inferencia, consulte **Actualización de un servicio**.

Actualización de imágenes

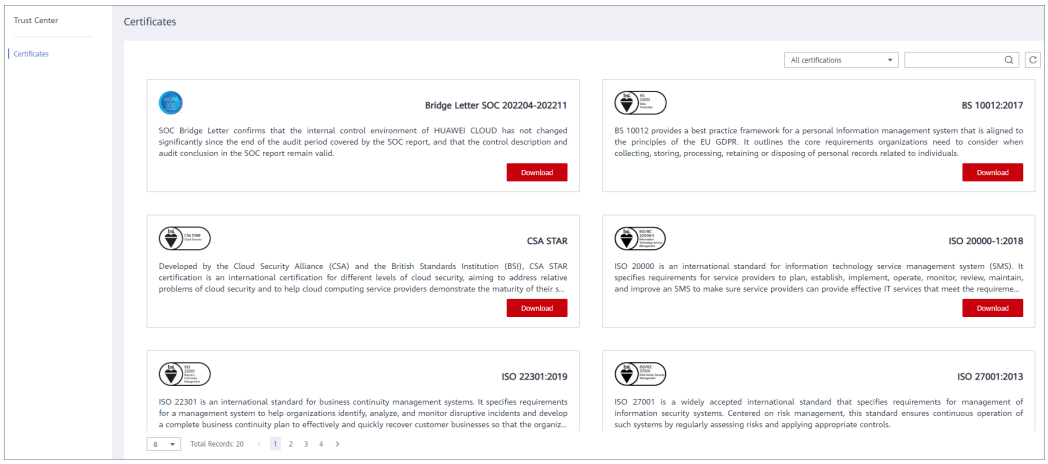
ModelArts ofrece tres módulos de función: DevEnviron, gestión de entrenamiento y despliegue de inferencia. Los tres módulos proporcionan imágenes con el mismo proceso. Estas imágenes se actualizan de forma irregular para corregir vulnerabilidades.

8.10 Certificados

Certificados de cumplimiento

Los servicios y plataformas de Huawei Cloud han obtenido diversas certificaciones de seguridad y cumplimiento de organizaciones autorizadas, como la Organización Internacional de Normalización (ISO). Puede **descargarlos** desde la consola.

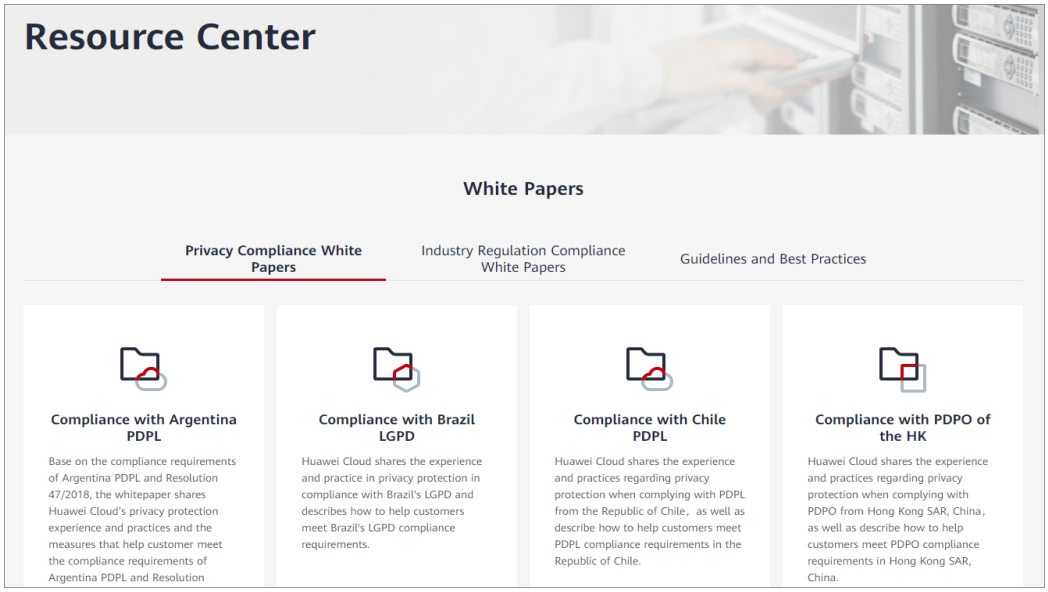
Figura 8-4 Descarga de certificados de cumplimiento



Centro de recurso

Huawei Cloud también proporciona los siguientes recursos para ayudar a los usuarios a cumplir con los requisitos de cumplimiento. Para obtener más información, consulte **Centro de recursos**.

Figura 8-5 Centro de recurso



8.11 Límite de seguridad

El modelo de responsabilidad compartida es un modo de cooperación en el que tanto proveedores como clientes asumen las responsabilidades de seguridad y cumplimiento de los servicios en nube.

Los proveedores gestionan la infraestructura en la nube y proporcionan hardware y software seguros para garantizar la disponibilidad del servicio. Los clientes protegen sus datos y aplicaciones, al tiempo que cumplen los requisitos de conformidad correspondientes.

Los proveedores son responsables de los servicios y funciones y deben:

- Establecer y mantener una infraestructura segura, incluidas redes, servidores y dispositivos de almacenamiento.
- Proporcionar plataformas subyacentes confiables para garantizar la seguridad del tiempo de ejecución para el entorno.
- Proporcionar autenticación de identidades y control de acceso para garantizar que sólo los usuarios autorizados puedan acceder a los servicios en nube y que los inquilinos estén aislados entre sí.
- Proporcionar respaldos de seguridad y recuperación de desastres fiables para evitar la pérdida de datos debida a fallos del hardware o desastres naturales.
- Proporcionar servicios transparentes de supervisión y respuesta a incidentes, actualizaciones de seguridad y parches de vulnerabilidad.

Los clientes deben:

- Encriptar datos y aplicaciones para garantizar la confidencialidad y la integridad.
- Asegurarse de que el software de la aplicación de IA se actualice de forma segura y de que se corrijan las vulnerabilidades.
- Cumplir con las regulaciones relacionadas, como GDPR, HIPAA y PCI DSS.
- Controle el acceso para garantizar que solo los usuarios autorizados puedan acceder y gestionar recursos como los servicios en línea.
- Monitorear y reportar cualquier actividad anormal y tomar acciones oportunas.

Responsabilidades de seguridad en el despliegue de inferencias

- Proveedores
  - Corregir los parches relacionados con los ECS subyacentes.
  - Actualizar el K8S y corregir vulnerabilidades.
  - Realizar el mantenimiento del ciclo de vida del sistema operativo de VM.
  - Garantizar la seguridad y el cumplimiento de la plataforma de inferencia de ModelArts.
  - Mejorar la seguridad de los servicios de aplicaciones de contenedor.
  - Actualizar el entorno de tiempo de ejecución del modelo y corregir las vulnerabilidades periódicamente.
- Clientes
  - Autorizar el uso de recursos y controlar el acceso.
  - Garantizar la seguridad de las aplicaciones, su cadena de suministro y sus dependencias mediante el escaneo, la auditoría y la verificación de acceso.
  - Minimizar los permisos y limitar la entrega de credenciales.
  - Garantizar la seguridad de las aplicaciones de IA (imágenes personalizadas, modelos de OBS y dependencias) durante el tiempo de ejecución.
  - Actualizar y corregir vulnerabilidades de manera oportuna.
  - Almacenar datos confidenciales de forma segura, como credenciales.

Prácticas recomendadas para la seguridad de despliegue de inferencia

- Autorización de servicio externo

  La inferencia de ModelArts requiere autorización de otros servicios en la nube. Usted solo puede otorgar los permisos requeridos en función de sus necesidades. Por ejemplo, puede otorgar permiso de acceso en un bucket de OBS a un tenant para la gestión del modelo.

- Autorización de recursos internos

  La inferencia de ModelArts admite un control de permisos detallado. Puede configurar los permisos para los usuarios en función de las necesidades reales para restringir los permisos en algunos recursos.

- Gestión de aplicaciones de IA

  Para desacoplar los modelos de las imágenes y proteger los activos del modelo, puede importar aplicaciones de IA de forma dinámica desde entrenamientos u OBS. Es necesario actualizar los paquetes de dependencia de las aplicaciones de IA y corregir las vulnerabilidades de los paquetes de código abierto o de terceros. La información sensible relacionada con las aplicaciones de IA debe desacoplarse y configurarse durante el despliegue. Seleccione el entorno de tiempo de ejecución recomendado por ModelArts. Los entornos anteriores pueden presentar vulnerabilidades de seguridad.

  Puede seleccionar imágenes de confianza abiertas al crear aplicaciones AI a partir de una imagen de contenedor, por ejemplo, imágenes de OpenEuler, Ubuntu y NVIDIA. Cree usuarios no root en lugar de usuarios root para ejecutar una imagen. En la imagen solo se instala el paquete de seguridad necesario durante el tiempo de ejecución. Reduzca el tamaño de la imagen y actualice el paquete de instalación a la última versión libre de vulnerabilidades. Desacople la información sensible de las imágenes durante el despliegue de servicio. No utilice directamente la información en Dockerfile. Realice análisis de seguridad en las imágenes periódicamente e instale parches para corregir vulnerabilidades. Para facilitar el reporte de alarmas y la rectificación de fallas, agregue una interfaz de verificación de estado y asegúrese de que el estado del servicio se pueda devolver correctamente. Para garantizar la seguridad de los datos del servicio, utilice flujos de transmisión de HTTPS y suites de encriptaciones confiables para contenedores.

- Despliegue de modelos

  Para evitar que los servicios se sobrecarguen o se desaprovechen, establezca las especificaciones adecuadas de los nodos de cálculo durante el despliegue. No escuche otros puertos del contenedor. Si es necesario acceder a otros puertos localmente, escúchelos en localhost. No transfiera directamente información confidencial con variables de entorno. Cifrar la información confidencial con el componente de encriptación antes de la transmisión de datos.

  La clave de autenticación de aplicaciones es una credencial de acceso para servicios en tiempo real. Debe conservar correctamente la clave de la aplicación.

9 Cuotas

ModelArts utiliza los siguientes recursos de infraestructura:

- ECS
- EVS
- VPC

Para obtener más información sobre cómo ver y modificar la cuota, consulte **Cuotas**.

