ModelArts

产品介绍

文档版本 01

发布日期 2024-04-22

华为云计算技术有限公司

版权所有 © 华为云计算技术有限公司 2024。 保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为云计算技术有限公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为云计算技术有限公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为云计算技术有限公司

地址： 贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址： https://www.huaweicloud.com/

目 录

 版权所有 © 华为云计算技术有限公司

 版权所有 © 华为云计算技术有限公司

1 图解 ModelArts

 版权所有 © 华为云计算技术有限公司

1.1 初识 ModelArts

版权所有 © 华为云计算技术有限公司

版权所有 © 华为云计算技术有限公司

2 什么是 ModelArts

ModelArts是面向AI开发者的一站式开发平台，提供海量数据预处理及半自动化标注、大规模分布式训练、自动化模型生成及端-边-云模型按需部署能力，帮助用户快速创建和部署模型，管理全周期AI工作流。

“一站式”是指AI开发的各个环节，包括数据处理、算法开发、模型训练、模型部署都可以在ModelArts上完成。从技术上看，ModelArts底层支持各种异构计算资源，开发者可以根据需要灵活选择使用，而不需要关心底层的技术。同时，ModelArts支持Tensorflow、PyTorch、MindSpore等主流开源的AI开发框架，也支持开发者使用自研的算法框架，匹配您的使用习惯。

ModelArts的理念就是让AI开发变得更简单、更方便。

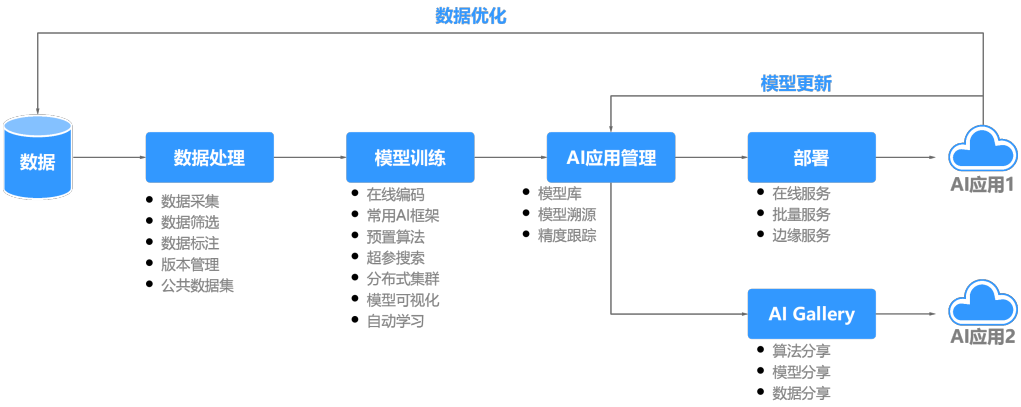
面向不同经验的AI开发者，提供便捷易用的使用流程。例如，面向业务开发者，不需关注模型或编码，可使用自动学习流程快速构建AI应用；面向AI初学者，不需关注模型开发，使用预置算法构建AI应用；面向AI工程师，提供多种开发环境，多种操作流程和模式，方便开发者编码扩展，快速构建模型及应用。

产品架构

ModelArts是一个一站式的开发平台，能够支撑开发者从数据到AI应用的全流程开发过程。包含数据处理、模型训练、模型管理、模型部署等操作，并且提供AI Gallery功能，能够在市场内与其他开发者分享模型。

ModelArts支持应用到图像分类、物体检测、视频分析、语音识别、产品推荐、异常检测等多种AI应用场景。

图 2-1 ModelArts Standard 架构



版权所有 © 华为云计算技术有限公司

产品优势

- **一站式**

  开“箱”即用，涵盖AI开发全流程，包含数据处理、模型开发、训练、管理、部署功能，可灵活使用其中一个或多个功能。

- **易上手**
  - 提供多种预置模型，开源模型想用就用。
  - 模型超参自动优化，简单快速。
  - 零代码开发，简单操作训练出自己的模型。
  - 支持模型一键部署到云、边、端。
- **高性能**
  - 优化深度模型推理中资源的利用率，加速云端在线推理。
  - 可生成在Ascend芯片上运行的模型，实现高效端边推理。
- **灵活**
  - 支持多种主流开源框架(TensorFlow、PyTorch、MindSpore等)。
  - 支持专属资源独享使用。
  - 支持自定义镜像满足自定义框架及算子需求。

首次使用 ModelArts

如果您是首次使用ModelArts的用户，建议您学习并了解如下信息：

- **基础知识了解**

  通过**4 基础知识**章节的内容，了解ModelArts相关的基础知识，包含AI开发的基础流程、AI开发的基础概念，以及ModelArts服务的特有概念和功能的详细介绍。

- **入门使用**

  《**快速入门**》提供了样例的详细操作指导，帮助用户学习并上手使用ModelArts Standard。

- **获取并尝试更多样例**

  ModelArts支持多种开源引擎，基于各类引擎和功能，提供了丰富的样例指导，您可以参考《**最佳实践**》的样例指导，完成相关的模型构建和部署。

- **使用更多的功能，并查看其相关操作指导**
  - 如果您是一个业务开发者，可以使用自动学习功能（无需编码，无需专业的AI基础能力），快速构建模型。详细操作指导可参考《**自动学习**》。
  - 如果您是一个AI工程师，可以使用AI全流程开发，包含使用《**开发环境**》、《**数据准备与分析**》、《**数据标注**》、《**模型训练**》、《**推理部署**》等，您使用一个或多个功能应用到您的AI开发中。
  - 如果您想要直接调用ModelArts的API或SDK完成AI开发，您可以参考《**API参考**》或《**SDK参考**》获取详情。

版权所有 © 华为云计算技术有限公司

3 功能介绍

繁多的AI工具安装配置、数据准备、模型训练慢等是困扰AI工程师的诸多难题。为解决这个难题，将一站式的AI开发平台（ModelArts）提供给开发者，从数据准备到算法开发、模型训练，最后把模型部署起来，集成到生产环境。一站式完成所有任务。

图 3-1 功能总览

ModelArts特色功能如下所示：

- **数据治理**

  支持数据筛选、标注等数据处理，提供数据集版本管理，特别是深度学习的大数据集，让训练结果可重现。
- **极“快”致“简”模型训练**

  自研的MoXing深度学习框架，更高效更易用，有效提升训练速度。
- **多场景部署**

  支持模型部署到多种生产环境，可部署为云端在线推理和批量推理，也可以直接部署到端和边。
- **自动学习**

  支持多种自动学习能力，通过“自动学习”训练模型，用户不需编写代码即可完成自动建模、一键部署。
- **AI Gallery**

版权所有 © 华为云计算技术有限公司

预置常用算法和常用数据集，支持模型在企业内部共享或者公开共享。

 版权所有 © 华为云计算技术有限公司

4 基础知识

4.1 AI 开发基本流程介绍

什么是 AI

AI（人工智能）是通过机器来模拟人类认识能力的一种科技能力。AI最核心的能力就是根据给定的输入做出判断或预测。

AI 开发的目的是什么

AI开发的目的是将隐藏在一大批数据背后的信息集中处理并进行提炼，从而总结得到研究对象的内在规律。

对数据进行分析，一般通过使用适当的统计、机器学习、深度学习等方法，对收集的大量数据进行计算、分析、汇总和整理，以求最大化地开发数据价值，发挥数据作用。

AI 开发的基本流程

AI开发的基本流程通常可以归纳为几个步骤：确定目的、准备数据、训练模型、评估模型、部署模型。

图 4-1 AI 开发流程

版权所有 © 华为云计算技术有限公司

步骤1 确定目的

在开始AI开发之前，必须明确要分析什么？要解决什么问题？商业目的是什么？基于商业的理解，整理AI开发框架和思路。例如，图像分类、物体检测等等。不同的项目对数据的要求，使用的AI开发手段也是不一样的。

步骤2 准备数据

数据准备主要是指收集和预处理数据的过程。

按照确定的分析目的，有目的性的收集、整合相关数据，数据准备是AI开发的一个基础。此时最重要的是保证获取数据的真实可靠性。而事实上，不能一次性将所有数据都采集全，因此，在数据标注阶段你可能会发现还缺少某一部分数据源，反复调整优化。

步骤3 训练模型

俗称“建模”，指通过分析手段、方法和技巧对准备好的数据进行探索分析，从中发现因果关系、内部联系和业务规律，为商业目的提供决策参考。训练模型的结果通常是一个或多个机器学习或深度学习模型，模型可以应用到新的数据中，得到预测、评价等结果。

业界主流的AI引擎有TensorFlow、PyTorch、MindSpore等，大量的开发者基于主流AI引擎，开发并训练其业务所需的模型。

步骤4 评估模型

训练得到模型之后，整个开发过程还不算结束，需要对模型进行评估和考察。经常不能一次性获得一个满意的模型，需要反复的调整算法参数、数据，不断评估训练生成的模型。

一些常用的指标，如准确率、召回率、AUC等，能帮助您有效的评估，最终获得一个满意的模型。

步骤5 部署模型

模型的开发训练，是基于之前的已有数据（有可能是测试数据），而在得到一个满意的模型之后，需要将其应用到正式的实际数据或新产生数据中，进行预测、评价、或以可视化和报表的形式把数据中的高价值信息以精辟易懂的形式提供给决策人员，帮助其制定更加正确的商业策略。

----结束

4.2 AI 开发基本概念

机器学习常见的分类有3种：

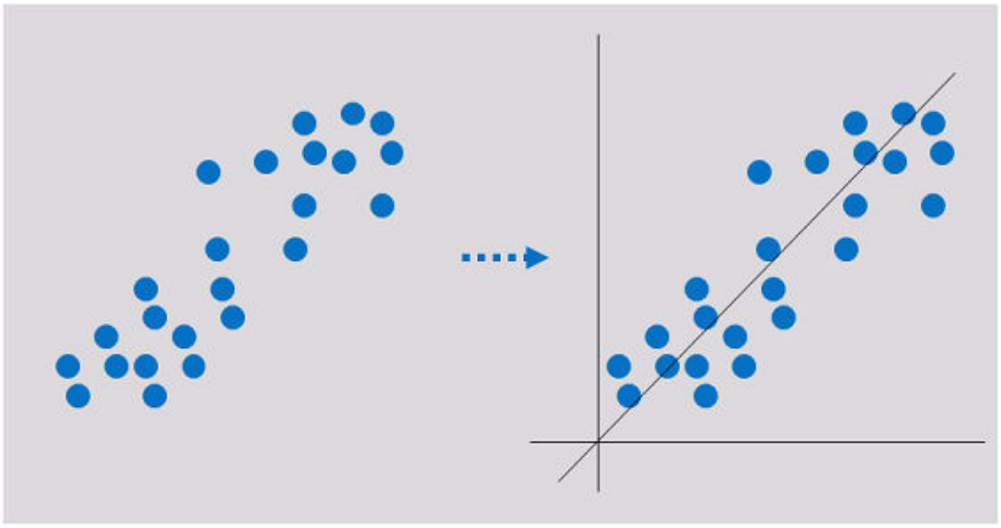
- 监督学习：利用一组已知类别的样本调整分类器的参数，使其达到所要求性能的过程，也称为监督训练或有教师学习。常见的有回归和分类。
- 非监督学习：在未加标签的数据中，试图找到隐藏的结构。常见的有聚类。
- 强化学习：智能系统从环境到行为映射的学习，以使奖励信号（强化信号）函数值最大。

回归

回归反映的是数据属性值在时间上的特征，产生一个将数据项映射到一个实值预测变量的函数，发现变量或属性间的依赖关系，其主要研究问题包括数据序列的趋势特

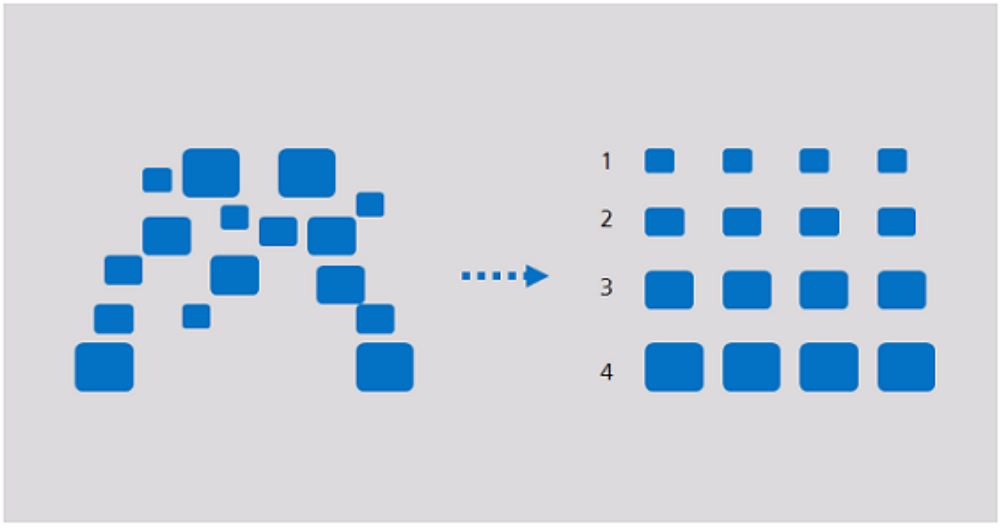
版权所有 © 华为云计算技术有限公司

征、数据序列的预测以及数据间的关系等。它可以应用到市场营销的各个方面，如客户寻求、保持和预防客户流失活动、产品生命周期分析、销售趋势预测及有针对性的促销活动等。



分类

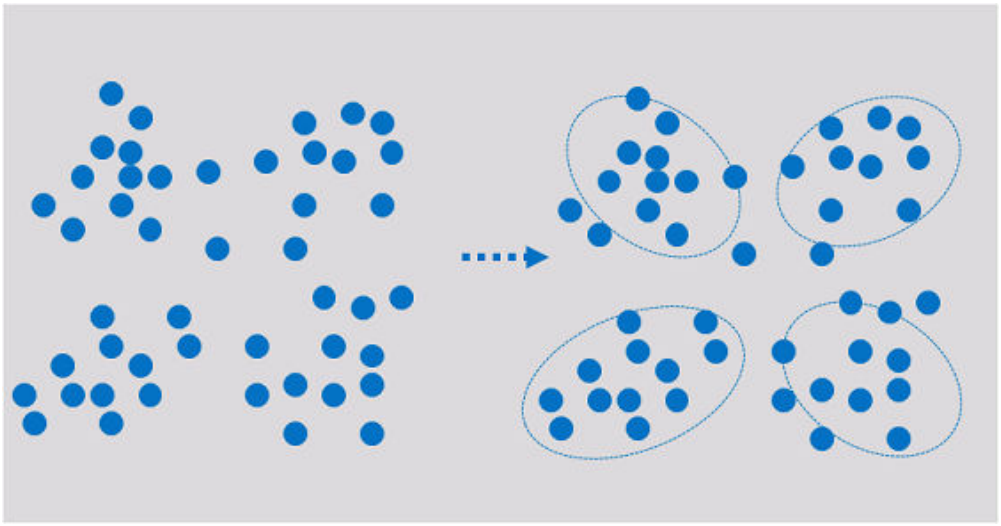
分类是找出一组数据对象的共同特点并按照分类模式将其划分为不同的类，其目的是通过分类模型，将数据项映射到某个给定的类别。它可以应用到客户的分类、客户的属性和特征分析、客户满意度分析、客户的购买趋势预测等。



聚类

聚类是把一组数据按照相似性和差异性分为几个类别，其目的是使得属于同一类别的数据间的相似性尽可能大，不同类别中的数据间的相似性尽可能小。它可以应用到客户群体的分类、客户背景分析、客户购买趋势预测、市场的细分等。

版权所有 © 华为云计算技术有限公司



与分类不同，聚类分析数据对象，而不考虑已知的类标号（一般训练数据中不提供类标号）。聚类可以产生这种标号。对象根据最大化类内的相似性、最小化类间的相似性的原则进行聚类或分组。对象的聚类是这样形成的，使得在一个聚类中的对象具有很高的相似性，而与其他聚类中的对象很不相似。

4.3 ModelArts 中常用概念

自动学习

自动学习功能可以根据标注数据自动设计模型、自动调参、自动训练、自动压缩和部署模型，不需要代码编写和模型开发经验。只需三步，标注数据、自动训练、部署模型，即可完成模型构建。

端-边-云

端-边-云分别指端侧设备、智能边缘设备、公有云。

推理

指按某种策略由已知判断推出新判断的思维过程。人工智能领域下，由机器模拟人类智能，使用构建的神经网络完成推理过程。

在线推理

在线推理是对每一个推理请求同步给出推理结果的在线服务（Web Service）。

批量推理

批量推理是对批量数据进行推理的批量作业。

Ascend 芯片

Ascend芯片是华为设计的高计算力低功耗的AI芯片。

版权所有 © 华为云计算技术有限公司

资源池

ModelArts提供的大规模计算集群，可应用于模型开发、训练和部署。支持公共资源池和专属资源池两种，分别为共享资源池和独享资源池。ModelArts默认提供公共资源池。专属资源池需单独创建，专属使用，不与其他用户共享。

AI Gallery

预置常用模型和算法，您可以直接获取使用。您也可以将自己开发的模型、算法或数据集分享至市场，共享给个人或者公开共享。

MoXing

MoXing是ModelArts自研的组件，是一种轻型的分布式框架，构建于TensorFlow、PyTorch、MXNet、MindSpore等深度学习引擎之上，使得这些计算引擎分布式性能更高，同时易用性更好。MoXing包含很多组件，其中MoXing Framework模块是一个基础公共组件，可用于访问OBS服务，和具体的AI引擎解耦，在ModelArts支持的所有AI引擎(TensorFlow、MXNet、PyTorch、MindSpore等)下均可以使用。

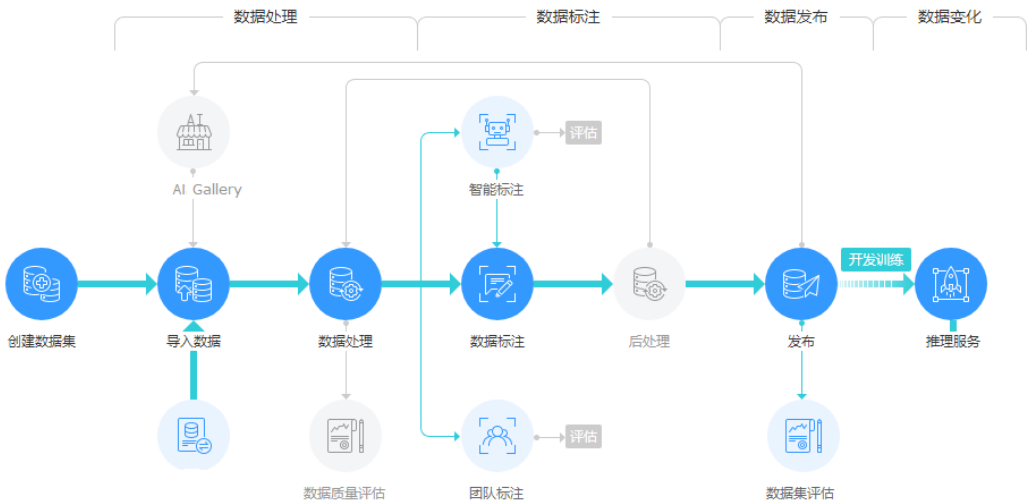
MoXing Framework模块提供了OBS中常见的数据文件操作，如读写、列举、创建文件夹、查询、移动、复制、删除等。

在ModelArts Notebook中使用MoXing接口时，可直接调用接口，无需下载或安装SDK，使用限制比ModelArts SDK和OBS SDK少，非常便捷。

4.4 数据管理

AI开发过程中经常需要处理海量数据，数据准备与标注耗费整体开发一半以上时间。ModelArts数据管理提供了一套高效便捷的管理和标注数据框架。不仅支持图片、文本、语音、视频等多种数据类型，涵盖图像分类、目标检测、音频分割、文本分类等多个标注场景，可适用于各种AI项目，如计算机视觉、自然语言处理、音视频分析等；同时提供数据筛选、数据分析、数据处理、团队标注以及版本管理等功能，AI开发者可基于该框架实现数据标注全流程处理。如图4-2所示。

图 4-2 数据标注全流程



版权所有 © 华为云计算技术有限公司

数据管理平台提供了聚类分析、数据特征分析、数据清洗、数据校验、数据增强、数据选择等分析处理能力，可帮助开发者进一步理解数据和挖掘数据，从而准备出一份满足开发目标或项目要求的高价值数据。

开发者在数据管理平台可以在线完成图像分类、目标检测、音频分割、文本三元组、视频分类等各种标注场景，同时也可以使用ModelArts智能标注方案，通过预置算法或自定义算法代替人工完成数据标注，提升标注效率。

针对大规模协同标注场景，数据管理平台还提供了强大的团队标注，支持标注团队管理、人员管理、角色管理等，实现从项目的创建、数据分配、进度把控、标注、审核、验收全流程。为用户带来标注效率提升的同时，又最小化项目管理开销。

此外，数据管理平台时刻保障用户数据的安全性和隐私性，确保用户数据仅在授权范围内使用。

新版数据管理中将数据集和数据标注功能解耦，更方便用户使用。

4.5 开发环境

软件开发的历史，就是一部降低开发者成本，提升开发体验的历史。在AI开发阶段，ModelArts也致力于提升AI开发体验，降低开发门槛。ModelArts开发环境，以云原生的资源使用和开发工具链的集成，目标为不同类型AI开发、探索、教学用户，提供更好云化AI开发体验。

ModelArts Notebook云上云下，无缝协同

- 代码开发与调测。云化JupyterLab使用，本地IDE+ModelArts插件远程开发能力，贴近开发人员使用习惯
- 云上开发环境，包含AI计算资源，云上存储，预置AI引擎
- 运行环境自定义，将开发环境直接保存成为镜像，供训练、推理使用

亮点特性 1：远程开发 - 支持本地 IDE 远程访问 Notebook

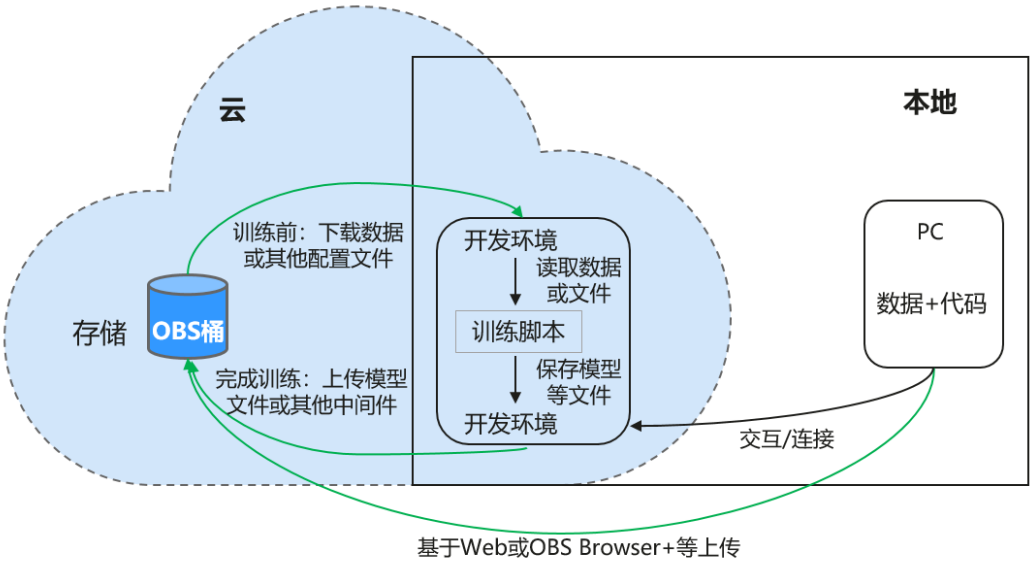
Notebook提供了远程开发功能，通过开启SSH连接，用户本地IDE可以远程连接到ModelArts的Notebook开发环境中，调试和运行代码。

对于使用本地IDE的开发者，由于本地资源限制，运行和调试环境大多使用团队公共搭建的资源服务器，并且是多人共用，这带来一定的环境搭建和维护成本。

而ModelArts的Notebook的优势是即开即用，它预先装好了不同的AI引擎，并且提供了非常多的可选规格，用户可以独占一个容器环境，不受其他人的干扰。只需简单配置，用户即可通过本地IDE连接到该环境进行运行和调试。

版权所有 © 华为云计算技术有限公司

图 4-3 本地 IDE 远程访问 Notebook 开发环境



ModelArts的Notebook可以视作是本地PC的延伸，均视作本地开发环境，其读取数据、训练、保存文件等操作与常规的本地训练一致。

对于习惯使用本地IDE的开发者，使用远程开发方式，不影响用户的编码习惯，并且可以方便快捷的使用云上的Notebook开发环境。

本地IDE当前支持VS Code、PyCharm、SSH工具。还有专门的插件PyCharm Toolkit和VS Code Toolkit，方便将云上资源作为本地的一个扩展。

亮点特性 2：开发环境保存 - 支持一键镜像保存

ModelArts的新版Notebook提供了镜像保存功能。支持一键将运行中的Notebook实例保存为镜像，将准备好的环境保存下来，可以作为自定义镜像，方便后续使用，并且方便进行分享。

保存镜像时，安装的依赖包（pip包）不丢失，VS Code远程开发场景下，在Server端安装的插件不丢失。

亮点特性 3：预置镜像 - 即开即用，优化配置，支持主流 AI 引擎

每个镜像预置的AI引擎和版本是固定的，在创建Notebook实例时明确AI引擎和版本，包括适配的芯片。

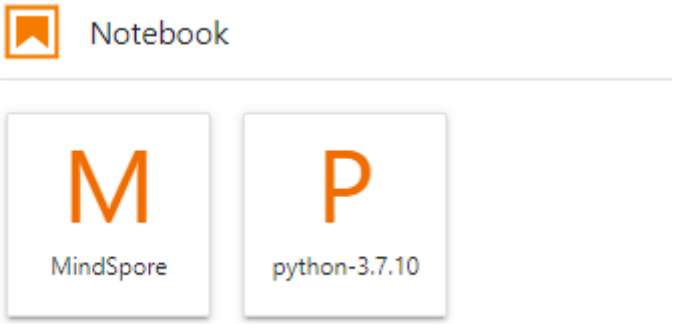
ModelArts开发环境给用户提供了一组预置镜像，主要包括PyTorch、Tensorflow、MindSpore系列。用户可以直接使用预置镜像启动Notebook实例，在实例中开发完成后，直接提交到ModelArts训练作业进行训练，而不需要做适配。

ModelArts开发环境提供的预置镜像版本是依据用户反馈和版本稳定性决定的。当用户的功能开发基于ModelArts提供的版本能够满足的时候，建议用户使用预置镜像，这些镜像经过充分的功能验证，并且已经预置了很多常用的安装包，用户无需花费过多的时间来配置环境即可使用。

ModelArts开发环境提供的预置镜像主要包含：

版权所有 © 华为云计算技术有限公司

- 常用预置包，基于标准的Conda环境，预置了常用的AI引擎，例如PyTorch、MindSpore；常用的数据分析软件包，例如Pandas、Numpy等；常用的工具软件，例如cuda、cudnn等，满足AI开发常用需求。
- 预置Conda环境：每个预置镜像都会创建一个相对应的Conda环境和一个基础Conda环境python（不包含任何AI引擎），如预置Mindspore所对应的Conda环境如下：



  用户可以根据是否使用AI引擎参与功能调试，并选择不同的Conda环境。
- Notebook：是一款Web应用，能够使用户在界面编写代码，并且将代码、数学方程和可视化内容组合到一个文档中。
- JupyterLab插件：插件包括规格切换，分享案例到AI Gallery进行交流，停止实例等，提升用户体验。
- 支持SSH远程连接功能，通过SSH连接启动实例，在本地调试就可以操作实例，方便调试。
- ModelArts开发环境提供的预置镜像支持功能开发后，直接提到ModelArts训练作业中进行训练。

说明

- 为了简化操作，ModelArts的新版Notebook，同一个Notebook实例中不支持不同引擎之间的切换。
- 不同Region支持的AI引擎不一样，请以控制台实际界面为准。

亮点特性 4：提供在线的交互式开发调试工具 JupyterLab

ModelArts集成了基于开源的JupyterLab，可为您提供在线的交互式开发调试。您无需关注安装配置，在ModelArts管理控制台直接使用Notebook，编写和调测模型训练代码，然后基于该代码进行模型的训练。

JupyterLab是一个交互式的开发环境，是Jupyter Notebook的下一代产品，可以使用它编写Notebook、操作终端、编辑MarkDown文本、打开交互模式、查看csv文件及图片等功能。

4.6 模型训练

模型训练中除了数据和算法外，开发者花了大量时间在模型参数设计上。模型训练的参数直接影响模型的精度以及模型收敛时间，参数的选择极大依赖于开发者的经验，参数选择不当会导致模型精度无法达到预期结果，或者模型训练时间大大增加。

为了降低开发者的专业要求，提升开发者模型训练的开发效率及训练性能，ModelArts提供了可视化作业管理、资源管理、版本管理等功能，基于机器学习算法及强化学习

版权所有 © 华为云计算技术有限公司

的模型训练自动超参调优，如learning rate、batch size等自动的调参策略；预置和调优常用模型，简化模型开发和全流程训练管理。

当前大多数开发者开发模型时，为了满足精度需求，模型通常达到几十层，甚至上百层，参数规模达到百兆甚至在GB规格以上，导致对计算资源的规格要求极高，主要体现在对硬件资源的算力及内存、ROM的规格的需求上。端侧资源规格限制极为严格，以端侧智能摄像头为例，通常端侧算力在1TFLOPS，内存在2GB规格左右，ROM空间在2GB左右，需要将端侧模型大小控制在百KB级别，推理时延控制在百毫秒级别。

这就需要借助模型精度无损或微损下的压缩技术，如通过剪枝、量化、知识蒸馏等技术，实现模型的自动压缩及调优，进行模型压缩和重新训练的自动迭代，以保证模型的精度损失极小。无需重新训练的低比特量化技术实现模型从高精度浮点向定点运算转换，多种压缩技术和调优技术实现模型计算量满足端、边小硬件资源下的轻量化需求，模型压缩技术在特定领域场景下实现精度损失<1%。

当训练数据量很大时，深度学习模型的训练将会非常耗时。深度学习训练加速一直是学术界和工业界所关注的重要问题。

分布式训练加速需要从软硬件两方面协同来考虑，仅单一的调优手段无法达到期望的加速效果。所以分布式加速的调优是一个系统工程，需要从硬件角度（芯片、硬件设计）考虑分布式训练架构，如系统的整体计算规格、网络带宽、高速缓存、功耗、散热等因素，充分考虑计算和通信的吞吐量关系，以实现计算和通信时延的隐藏。

软件设计需要结合高性能硬件特性，充分利用硬件高速网络实现高带宽分布式通信，实现高效的数据集本地数据缓存技术，通过训练调优算法，如混合并行，梯度压缩、卷积加速等技术，实现分布式训练系统软硬件端到端的高效协同优化，实现多机多卡分布式环境下训练加速。ModelArts在千级别资源规格多机多卡分布式环境下，典型模型ResNet50 在ImageNet数据集上实现加速比>0.8，是行业领先水平。

衡量分布式深度学习的加速性能时，主要有如下2个重要指标：

- 吞吐量，即单位时间内处理的数据量。
- 收敛时间，即达到一定的收敛精度所需的时间。

吞吐量一般取决于服务器硬件（如更多、更大FLOPS处理能力的AI加速芯片，更大的通信带宽等）、数据读取和缓存、数据预处理、模型计算（如卷积算法选择等）、通信拓扑等方面的优化。除了低bit计算和梯度（或参数）压缩等，大部分技术在提升吞吐量的同时，不会造成对模型精度的影响。为了达到最短的收敛时间，需要在优化吞吐量的同时，对调参方面也做调优。调参不到位会导致吞吐量难以优化，当batch size超参不足够大时，模型训练的并行度就会相对较差，吞吐量难以通过增加计算节点个数而提升。

对用户而言，最终关心的指标是收敛时间，因此ModelArts的MoXing实现了全栈优化，极大缩短了训练收敛时间。在数据读取和预处理方面，MoXing通过利用多级并发输入流水线使得数据IO不会成为瓶颈；在模型计算方面，MoXing对上层模型提供半精度和单精度组成的混合精度计算，通过自适应的尺度缩放减小由于精度计算带来的损失；在超参调优方面，采用动态超参策略（如momentum、batch size等）使得模型收敛所需epoch个数降到最低；在底层优化方面，MoXing与底层华为服务器和通信计算库相结合，使得分布式加速进一步提升。

ModelArts 高性能分布式训练优化点

- 自动混合精度训练（充分发挥硬件计算能力）
- 动态超参调整技术（动态batch size、image size、momentum等）
- 模型梯度的自动融合、拆分

版权所有 © 华为云计算技术有限公司

- 基于BP bubble自适应的计算，通信算子调度优化
- 分布式高性能通信库（nstack、HCCL）
- 分布式数据-模型混合并行
- 训练数据压缩、多级缓存

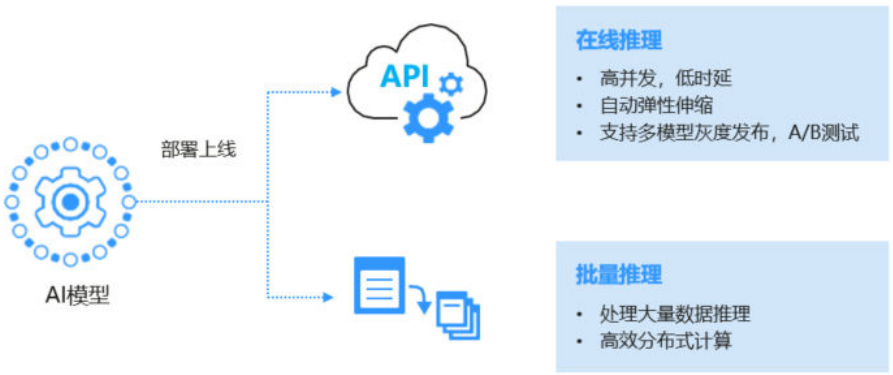
4.7 模型部署

ModelArts提供模型、服务管理能力，支持多厂商多框架多功能的镜像和模型统一纳管。

通常AI模型部署和规模化落地非常复杂。

例如，智慧交通项目中，在获得训练好的模型后，需要部署到云、边、端多种场景。如果在端侧部署，需要一次性部署到不同规格、不同厂商的摄像机上，这是一项非常耗时、费力的巨大工程，ModelArts支持将训练好的模型一键部署到端、边、云的各种设备上和各种场景上，并且还为个人开发者、企业和设备生产厂商提供了一整套安全可靠的一站式部署方式。

图 4-4 部署模型的流程



- 在线推理服务，可以实现高并发，低延时，弹性伸缩，并且支持多模型灰度发布、A/B测试。
- 支持各种部署场景，部署为云端的在线推理服务和批量推理任务。

版权所有 © 华为云计算技术有限公司

5 ModelArts 支持哪些 AI 框架?

ModelArts的开发环境Notebook、训练作业、模型推理（即AI应用管理和部署上线）支持的AI框架及其版本，不同模块的呈现方式存在细微差异，各模块支持的AI框架请参见如下描述。

开发环境 Notebook

开发环境的Notebook，根据不同的工作环境，对应支持的镜像和版本有所不同。

表 5-1 Notebook 支持的镜像

镜像名称	镜像描述	适配芯片	支持SSH远程开发访问	支持在线JupyterLab访问
pytorch1.8-cuda10.2-cudnn7-ubuntu18.04	CPU、GPU通用算法开发和训练基础镜像，预置AI引擎PyTorch1.8	CPU/GPU	是	是
mindspore1.7.0-cuda10.1-py3.7-ubuntu18.04	CPU and GPU general algorithm development and training, preconfigured with AI engine MindSpore1.7.0 and cuda 10.1	CPU/GPU	是	是
mindspore1.7.0-py3.7-ubuntu18.04	CPU general algorithm development and training, preconfigured with AI engine MindSpore1.7.0	CPU	是	是

版权所有 © 华为云计算技术有限公司

镜像名称	镜像描述	适配芯片	支持SSH远程开发访问	支持在线Jupyter Lab访问
pytorch1.10-cuda10.2-cudnn7-ubuntu18.04	CPU and GPU general algorithm development and training, preconfigured with AI engine PyTorch1.10 and cuda10.2	CPU/GPU	是	是
tensorflow2.1-cuda10.1-cudnn7-ubuntu18.04	CPU、GPU通用算法开发和训练基础镜像，预置AI引擎TensorFlow2.1	CPU/GPU	是	是
conda3-ubuntu18.04	Clean user customized base image only include conda	CPU	是	是
pytorch1.4-cuda10.1-cudnn7-ubuntu18.04	CPU、GPU通用算法开发和训练基础镜像，预置AI引擎PyTorch1.4	CPU/GPU	是	是
tensorflow1.13-cuda10.0-cudnn7-ubuntu18.04	GPU通用算法开发和训练基础镜像，预置AI引擎TensorFlow1.13.1	GPU	是	是
conda3-cuda10.2-cudnn7-ubuntu18.04	Clean user customized base image include cuda10.2, conda	CPU	是	是
spark2.4.5-ubuntu18.04	CPU algorithm development and training, prebuilt PySpark 2.4.5 and is able to attach to preconfigured spark cluster including MRS and DLI.	CPU	否	是
mindspore1.2.0-cuda10.1-cudnn7-ubuntu18.04	GPU算法开发和训练基础镜像，预置AI引擎MindSpore-GPU	GPU	是	是

版权所有 © 华为云计算技术有限公司

镜像名称	镜像描述	适配芯片	支持SSH远程开发访问	支持在线Jupyter Lab访问
mindspore1.2.0-openmpi2.1.1-ubuntu18.04	CPU算法开发和训练基础镜像，预置AI引擎MindSpore-CPU	CPU	是	是

训练作业

创建训练作业时，训练支持的AI引擎及对应版本如下所示。

预置引擎命名格式如下：

```
<训练引擎名称_版本号>-[cpu | <cuda_版本号 | cann_版本号 >]-<py_版本号>-<操作系统名称_版本号>-<x86_64 | aarch64>
```

表 5-2 训练作业支持的 AI 引擎

工作环境	系统架构	系统版本	AI引擎与版本	支持的cuda或Ascend版本
TensorFlow	x86_64	Ubuntu18.04	tensorflow_2.1.0-cuda_10.1-py_3.7-ubuntu_18.04-x86_64	cuda10.1
PyTorch	x86_64	Ubuntu18.04	pytorch_1.8.0-cuda_10.2-py_3.7-ubuntu_18.04-x86_64	cuda10.2
MPI	x86_64	Ubuntu18.04	mindspore_1.3.0-cuda_10.1-py_3.7-ubuntu_1804-x86_64	cuda_10.1
Horovod	x86_64	ubuntu_18.04	horovod_0.20.0-tensorflow_2.1.0-cuda_10.1-py_3.7-ubuntu_18.04-x86_64	cuda_10.1
			horovod_0.22.1-pytorch_1.8.0-cuda_10.2-py_3.7-ubuntu_18.04-x86_64	cuda_10.2

说明

不同区域支持的AI引擎有差异，请以实际环境为准。

版权所有 © 华为云计算技术有限公司

推理支持的 AI 引擎

在ModelArts创建AI应用时，若使用预置镜像“从模板中选择”或“从OBS中选择”导入模型，则支持如下常用引擎及版本的模型包。

说明

- 标注“推荐”的Runtime来源于统一镜像，后续统一镜像将作为主流的推理基础镜像。统一镜像中的安装包更齐全，详细信息可以参见**推理基础镜像列表**。
- 推荐将旧版镜像切换为统一镜像，旧版镜像后续将会逐渐下线。
- 待下线的基本镜像不再维护。
- 统一镜像Runtime的命名规范：<AI引擎名字及版本> - <硬件及版本：cpu或cuda或cann> - <python版本> - <操作系统版本> - <CPU架构>

表 5-3 支持的常用引擎及其 Runtime

模型使用的引擎类型	支持的运行环境（Runtime）	注意事项
TensorFlow	python3.6 python2.7（待下线） tf1.13-python3.6-gpu tf1.13-python3.6-cpu tf1.13-python3.7-cpu tf1.13-python3.7-gpu tf2.1-python3.7（待下线） tensorflow_2.1.0-cuda_10.1-py_3.7-ubuntu_18.04-x86_64（推荐）	• python2.7、python3.6的运行环境搭载的TensorFlow版本为1.8.0。 • python3.6、python2.7、tf2.1-python3.7，表示该模型可同时在CPU或GPU运行。其他Runtime的值，如果后缀带cpu或gpu，表示该模型仅支持在CPU或GPU中运行。 • 默认使用的Runtime为python2.7。
Spark_MLlib	python2.7（待下线） python3.6（待下线）	• python2.7以及python3.6的运行环境搭载的Spark_MLlib版本为2.3.2。 • 默认使用的Runtime为python2.7。 • python2.7、python3.6只能用于运行适用于CPU的模型。
Scikit_Learn	python2.7（待下线） python3.6（待下线）	• python2.7以及python3.6的运行环境搭载的Scikit_Learn版本为0.18.1。 • 默认使用的Runtime为python2.7。 • python2.7、python3.6只能用于运行适用于CPU的模型。

版权所有 © 华为云计算技术有限公司

模型使用的引擎类型	支持的运行环境（Runtime）	注意事项
XGBoost	python2.7（待下线） python3.6（待下线）	• python2.7以及python3.6的运行环境搭载的XGBoost版本为0.80。 • 默认使用的Runtime为python2.7。 • python2.7、python3.6只能用于运行适用于CPU的模型。
PyTorch	python2.7（待下线） python3.6 python3.7 pytorch1.4-python3.7 pytorch1.5-python3.7（待下线） pytorch_1.8.0-cuda_10.2-py_3.7-ubuntu_18.04-x86_64（推荐）	• python2.7、python3.6、python3.7的运行环境搭载的PyTorch版本为1.0。 • python2.7、python3.6、python3.7、pytorch1.4-python3.7、pytorch1.5-python3.7，表示该模型可同时在CPU或GPU运行。 • 默认使用的Runtime为python2.7。
MindSpore	aarch64（推荐）	aarch64只能用于运行在Snt3芯片上。

版权所有 © 华为云计算技术有限公司

6 与其他服务的关系

与统一身份认证服务的关系

ModelArts使用统一身份认证服务（Identity and Access Management，简称IAM）实现认证功能。IAM的更多信息请参见**《统一身份认证服务用户指南》**。

与对象存储服务的关系

ModelArts使用对象存储服务（Object Storage Service，简称OBS）存储数据和模型，实现安全、高可靠和低成本的存储需求。OBS的更多信息请参见**《对象存储服务控制台指南》**。

表 6-1 ModelArts 各环节与 OBS 的关系

功能	子任务	ModelArts与OBS的关系
自动学习	数据标注	ModelArts标注的数据存储在OBS中。
	自动训练	训练作业结束后，其生成的模型存储在OBS中。
	部署上线	ModelArts将存储在OBS中的模型部署上线为在线服务。
AI全流程开发	数据管理	• 数据集存储在OBS中。 • 数据集的标注信息存储在OBS中。 • 支持从OBS中导入数据。
	开发环境	Notebook实例中的数据或代码文件存储在OBS中。
	训练模型	• 训练作业使用的数据集存储在OBS中。 • 训练作业的运行脚本存储在OBS中。 • 训练作业输出的模型存储在指定的OBS中。 • 训练作业的过程日志存储在指定的OBS中。
	AI应用管理	训练作业结束后，其生成的模型存储在OBS中，创建AI应用时，从OBS中导入已有的模型文件。

版权所有 © 华为云计算技术有限公司

功能	子任务	ModelArts与OBS的关系
	部署上线	将存储在OBS中的模型部署上线。
全局配置	-	获取访问授权（使用委托或访问密钥授权），以便ModelArts可以使用OBS存储数据、创建Notebook等操作。

与云硬盘的关系

ModelArts使用云硬盘服务（Elastic Volume Service，简称EVS）存储创建的Notebook实例。EVS的更多信息请参见《**云硬盘用户指南**》。

与云容器引擎的关系

ModelArts使用云容器引擎（Cloud Container Engine，简称CCE）部署模型为在线服务，支持服务的高并发和弹性伸缩需求。CCE的更多信息请参见**《云容器引擎用户指南》**。

与容器镜像服务的关系

当使用ModelArts不支持的AI框架构建模型时，可通过构建的自定义镜像导入ModelArts进行训练或推理。您可以通过容器镜像服务（Software Repository for Container，简称SWR）制作并上传自定义镜像，然后再通过容器镜像服务导入ModelArts。SWR的更多信息请参见**《容器镜像服务用户指南》**。

与云监控的关系

ModelArts使用云监控服务（Cloud Eye Service， 简称CES）监控在线服务和对应模型负载，执行自动实时监控、告警和通知操作。CES的更多信息请参见**《云监控服务用户指南》**。

版权所有 © 华为云计算技术有限公司

7 如何访问 ModelArts

云服务平台提供了Web化的服务管理平台，即管理控制台和基于HTTPS请求的API（Application programming interface）管理方式。

- **管理控制台方式**

  ModelArts Standard提供了简洁易用的管理控制台，包含自动学习、数据管理、开发环境、模型训练、AI应用管理、部署上线、AI Gallery等功能，您可以在管理控制台端到端完成您的AI开发。

  使用ModelArts管理控制台，需先注册华为云。如果您已注册华为云，可从主页选择“ 人工智能 > AI开发平台ModelArts”直接登录管理控制台。

- **SDK方式**

  如果您需要将ModelArts集成到第三方系统，用于二次开发，可选择调用SDK方式完成目的。ModelArts的SDK是对ModelArts服务提供的REST API进行的Python封装，简化用户的开发工作。具体操作和SDK详细描述，请参见《**SDK参考**》。

  除此之外，在管理控制台的Notebook中编写代码时，也可直接调用ModelArts SDK。

- **API方式**

  如果您需要将ModelArts集成到第三方系统，用于二次开发，请使用API方式访问ModelArts，具体操作和API详细描述，请参见《**API参考**》。

版权所有 © 华为云计算技术有限公司

8 计费说明

8.1 计费概述

ModelArts是面向AI开发者的一站式开发平台，提供海量数据预处理及半自动化标注、大规模分布式训练、自动化模型生成及端-边-云模型按需部署能力，帮助用户快速创建和部署AI应用，管理全周期AI工作流。

ModelArts服务的计费方式简单、灵活，您既可以选择按实际使用时长计费，也可以选择更经济的按包周期（包年/包月）计费方式。详细的费用价格请参见**产品价格详情**。

更多详细的计费介绍，请参见**《计费说明》**文档。

8.2 计费项

在ModelArts中进行AI全流程开发时，涉及到计费项主要包括存储费用、资源费用。

- **存储费用**：使用OBS存储产生的费用、EVS存储（仅适用于Notebook）产生的费用。
- **资源费用**：使用ModelArts计算资源产生的费用。

版权所有 © 华为云计算技术有限公司

存储费用

表 8-1 存储计费项说明

计费项	说明
对象存储（Object Storage Service，OBS）	ModelArts使用对象存储服务，存储数据和模型，会产生相应的费用，具体费用可参见**对象存储价格详情**。

资源费用

ModelArts服务可根据用户的使用区域和所需要的业务类型，选择合适的计算资源，完成相应的AI开发。具体资源详情请参考**产品价格详情**。

在使用ModelArts时，不同场景的计算资源使用详情可参见**按需付费使用**。

表 8-2 计算资源计费项说明

计费项	说明
AI全流程开发	面向有AI基础的开发者，提供机器学习和深度学习的算法开发及部署全功能，包含数据处理、模型开发、模型训练、AI应用管理和部署上线流程。 涉及计费项包含： • 开发环境（Notebook） • 模型训练（训练作业） • 部署上线（在线服务）
自动学习	面向AI基础能力弱的开发者，根据标注数据、自动设计、调优、训练模型和部署服务，根据开发者零编码实现模型定制化开发。此计费资源仅适用于自动学习作业的训练和部署。 涉及计费项包括： • 自动学习-训练作业 • 自动学习-服务部署 **说明** 当前仅支持按需计费模式，不支持包年包月计费模式。

8.3 计费模式

本文主要介绍ModelArts中使用的计算资源的计费模式，包括按需计费和按包周期（包年/包月）计费，供您灵活选择。

- **按需计费**：这种购买方式比较灵活，可以即开即停。在创建开发环境、创建训练作业、部署模型服务等页面中选择相应资源规格时购买。
- **按包周期（包年/包月）**：ModelArts提供包年和包月购买资源的模式。这种购买方式相对于按需付费则能够提供更大的折扣。

版权所有 © 华为云计算技术有限公司

说明

- **“公共资源池”**只支持按需计费模式。
- 目前只有“专属资源池”支持包周期购买模式。不同区域支持的专属资源池功能及购买方式有差异，请以控制台实际界面为准。

  专属资源池购买入口在“ModelArts控制台>专属资源池>创建”页面中。若ModelArts控制台无专属资源池入口或者专属资源池的购买页面中无包年/包月购买方式，说明该区域不支持包周期的购买方式。

表 8-3 计费模式

计费模式	包年/包月	按需计费
付费方式	预付费 按照订单的购买周期结算。	后付费 按照资源的实际使用时长计费。
计费周期	按订单的实际购买时长计费。	秒级计费，按小时结算。
更改计费方式	支持变更为按需资源（仅旧版“开发环境/训练专用”的资源池） 但包年/包月计费模式到期后，按需的计费模式才会生效。	支持变更为包年/包月资源（仅旧版“开发环境/训练专用”的资源池） 但按需计费的模式有消费记录后，才能变更计费方式。
使用场景	适用于可预估资源使用周期的场景，价格比按需计费模式更优惠。这种计费模式更推荐长期使用者购买。	适用于资源需求波动的场景，可以即开即停。

8.4 变更配置

在使用ModelArts时，您可根据业务需要选择合适的计算资源。当作业启动后，您可以使用如下变更配置的方式。若ModelArts提供的变更配置方式不满足您的要求，您也可以通过重建作业，做数据迁移的方式实现配置变更。

更改计费模式

ModelArts支持对专属资源池进行计费模式变更，具体操作请参见资源池。

说明

ModelArts仅旧版“开发环境/训练专用”的专属资源池支持按需与包周期互转。

计费模式的变更生效限制如下：

- 按需计费转包年/包月：按需计费的模式有消费记录后，才能变更计费方式。变更完后，包年/包月资源立即生效。
- 包年/包月变更为按需：包年/包月转按需，需包年/包月资费模式到期后，按需的计费模式才会生效。

版权所有 © 华为云计算技术有限公司

专属资源池扩缩容

ModelArts支持对“运行中”的专属资源池进行扩缩容操作，具体请参见**扩缩容资源池**。

说明

若您使用的是旧版的专属资源池，可参见**资源池**对资源池进行扩缩容操作。专属资源池的扩缩容限制如下：

- 包周期（包年/包月）专属资源池支持扩容，不支持缩容。支持在包周期到期后，将专属资源池设置为按需计费。
- 按需计费的专属资源池，可以手动扩缩容，计费会按照修改后的节点数量进行收费。

8.5 续费

目前ModelArts提供按需和包年/包月购买方式。按需是每小时扣费，如果余额不足会导致欠费。包年/包月是超出当前包年、包月的额度后，系统会自动以按需计费的方式进行结算，只要您的帐户上有足够余额，则不会影响您的使用，如果余额不足会导致欠费。如果您未能续费，华为云不会立即停止您的业务，订单转入保留期，此时将终止服务，数据仍然保留。

- 保留期的时长由客户等级而定，具体请参见**保留期**。
- 如需续费，请进入**续费管理**页面进行续费操作。

8.6 欠费与到期

ModelArts专属资源池欠费与到期说明如下：

- 按需计费模式的资源，没有到期的概念。按需资源按每小时扣费，当余额不足，无法对上一个小时的费用进行扣费，就会导致欠费。欠费后资源会被冻结，您可以续费后解冻资源，方可继续正常使用。欠费后，会陆续进入宽限期和保留期。
- 包年/包月模式的资源到期后，会陆续进入宽限期和保留期。

在宽限期内客户可正常访问和使用此资源池。如果您在宽限期内仍未续费资源池，那么就会进入保留期，资源状态变为“已冻结”，您将无法对处于保留期的资源执行任何操作。保留期到期后，若资源池仍未续费，那么资源池将被自动删除。

注意

在保留期进行的续费，是以原到期时间作为生效时间，您应当支付从进入保留期开始到续费时的服务费用。

欠费受限

您购买的资源欠费后，会导致部分操作受限，建议您尽快续费。具体受限操作如**表8-4**所示：

版权所有 © 华为云计算技术有限公司

表 8-4 欠费受限操作

功能	受限操作
Workflow	订阅workflow、模型训练、部署上线
自动学习	模型训练、部署上线
开发环境-Notebook	创建Notebook、启动Notebook
训练管理-训练作业	创建训练作业
部署上线-在线服务、批量服务、边缘服务	部署在线服务、批量服务、边缘服务
专属资源池	创建专属资源池

 版权所有 © 华为云计算技术有限公司

9 权限管理

ModelArts作为一个完备的AI开发平台，支持用户对其进行细粒度的权限配置，以达到精细化资源、权限管理之目的。这类特性在大型企业用户的使用场景下很常见，但对个人用户则显得复杂而意义不足，所以建议个人用户在使用ModelArts时，参照**个人用户快速配置ModelArts访问权限**来进行初始权限设置。

说明

您是否需要阅读本文档?

如果下述问题您的任何一个回答为“是”，则需要阅读此文档

- 您是企业用户，且
  - 存在多个部门，且需要限定不同部门的用户只能访问其专属资源、功能
  - 存在多种角色（如管理员、算法开发者、应用运维），希望限制不同角色只能使用特定功能
  - 逻辑上存在多套“环境”且相互隔离（如开发环境、预生产环境、生产环境），并限定不同用户在不同环境上的操作权限
  - 其他任何需要对特定子用户（组）做出特定权限限制的情况
- 您是个人用户，但已经在IAM创建多个子用户，且期望限定不同子用户所能使用的ModelArts功能、资源不同。
- 希望了解ModelArts的权限控制能力细节，期望理解其概念和实操方法。

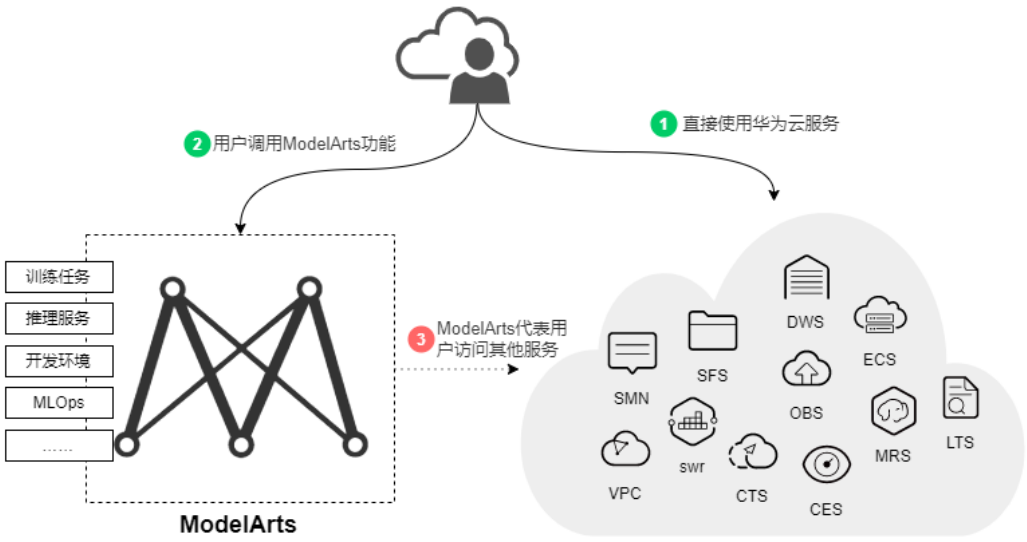
ModelArts的大部分权限管理能力均基于统一身份认证服务（Identity and Access Management，简称IAM）来实现，在您继续往下阅读之前，强烈建议您先行熟悉**IAM基本概念**，如果能完整理解IAM的所有概念，将更加有助于您理解本文档。

为了支持客户对ModelArts的权限做精细化控制，提供了3个方面的能力来支撑，分别是：权限、委托和工作空间。下面分别讲解。

版权所有 © 华为云计算技术有限公司

理解 ModelArts 的权限与委托

图 9-1 权限管理抽象



ModelArts与其他服务类似，对外暴露的每个功能，都通过IAM的权限来进行控制。比如，用户（此处指IAM子用户，而非租户）希望在ModelArts创建训练作业，则该用户必须拥有 "modelarts:trainJob:create" 的权限才可以完成操作（无论界面操作还是API调用）。关于如何给一个用户赋权（准确讲是需要先将用户加入用户组，再面向用户组赋权），可以参考IAM的文档《**权限管理**》。

而ModelArts还有一个特殊的地方在于，为了完成AI计算的各种操作，AI平台在任务执行过程中需要访问用户的其他服务，典型的就是训练过程中，需要访问OBS读取用户的训练数据。在这个过程中，就出现了ModelArts“代表”用户去访问其他云服务的情形。从安全角度出发，ModelArts代表用户访问任何云服务之前，均需要先获得用户的授权，而这个动作就是一个“委托”的过程。用户授权ModelArts再代表自己访问特定的云服务，以完成其在ModelArts平台上执行的AI计算任务。

综上，对于**图1 权限管理抽象**可以做如下解读：

- 用户访问任何云服务，均是通过标准的IAM权限体系进行访问控制。用户首先需要具备相关云服务的权限（根据您具体使用的功能不同，所需的相关服务权限多寡亦有差异）。
- **权限**：用户使用ModelArts的任何功能，亦需要通过IAM权限体系进行正确权限授权。
- **委托**：ModelArts上的AI计算任务执行过程中需要访问其他云服务，此动作需要获得用户的委托授权。

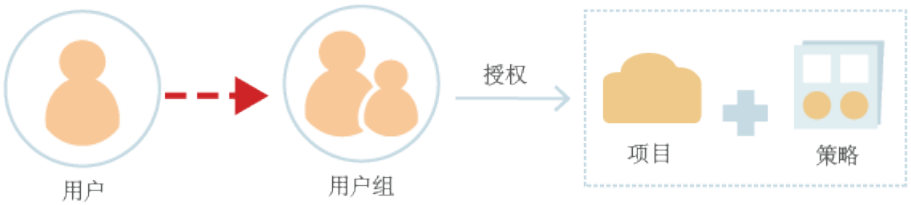
ModelArts 权限管理

默认情况下，管理员创建的IAM用户没有任何权限，需要将其加入用户组，并给用户组授予策略，才能使得用户组中的用户获得对应的权限，这一过程称为授权。授权后，用户就可以基于授予的权限对云服务进行操作。

版权所有 © 华为云计算技术有限公司

⚠ 注意

- ModelArts部署时通过物理区域划分，为项目级服务，授权时“选择授权范围方案”可以选择“指定区域项目资源”，如果授权时指定了区域对应的项目，则该权限仅对此项目生效；简单的做法是直接选择“所有资源”。
- ModelArts也支持企业项目，所以选择授权范围方案时，也可以指定企业项目。具体操作参见《**创建用户组并授权**》。



IAM在对用户组授权的时候，并不是直接将具体的某个权限进行赋权，而是需要先将权限加入到“策略”当中，再把策略赋给用户组。为了方便用户的权限管理，各个云服务都提供了一些预置的“系统策略”供用户直接使用。如果预置的策略不能满足您的细粒度权限控制要求，则可以通过“自定义策略”来进行精细控制。

表9-1列出了ModelArts的所有预置系统策略。

表 9-1 ModelArts 系统策略

策略名称	描述	类型
ModelArts FullAccess	ModelArts管理员用户，拥有所有ModelArts服务的权限	系统策略
ModelArts CommonOperations	ModelArts操作用户，拥有所有ModelArts服务操作权限除了管理专属资源池的权限	系统策略
ModelArts Dependency Access	ModelArts服务的常用依赖服务的权限	系统策略

通常来讲，只给管理员开通“ModelArts FullAccess”，如果不需要太精细的控制，直接给所有用户开通“ModelArts CommonOperations”即可满足大多数小团队的开发场景诉求。如果您希望通过自定义策略做深入细致的权限控制，请阅读**ModelArts的IAM权限控制详解**。

版权所有 © 华为云计算技术有限公司

说明

ModelArts的权限不会凌驾于其他服务的权限之上，当您给用户进行ModelArts赋权时，系统不会自动对其他相关服务的相关权限进行赋权。这样做的好处是更加安全，不会出现预期外的“越权”，但缺点是，您必须同时给用户赋予不同服务的权限，才能确保用户可以顺利完成某些ModelArts操作。

举例，如果用户需要用OBS中的数据进行训练，当已经为IAM用户配置ModelArts训练权限时，仍需同时为其配置对应的OBS权限（读、写、列表），才可以正常使用。其中OBS的列表权限用于支持用户从ModelArts界面上选择要进行训练的数据路径；读权限主要用于数据的预览以及训练任务执行时的数据读取；写权限则是为了保存训练结果和日志。

- 对于个人用户或小型组织，一个简单做法是为IAM用户配置“作用范围”为“全局级服务”的“Tenant Administrator”策略，这会使用户获得除了IAM以外的所有用户权限。在获得便利的同时，由于用户的权限较大，会存在相对较大的安全风险，需谨慎使用。（对于个人用户，其默认IAM账号就已经属于admin用户组，且具备Tenant Administrator权限，无需额外操作）
- 当您需要限制用户操作，仅为ModelArts用户配置OBS相关的最小化权限项，具体操作请参见**OBS权限管理**。对于其他云服务，也可以进行精细化权限控制，具体请参考对应的云服务文档。

ModelArts 委托授权

前文已经介绍，ModelArts在执行AI计算任务过程中，需要“代表”用户去访问其他云服务，而此动作需要提前获得用户的授权。在IAM权限体系下，此类授权动作是通过“委托”来完成。

关于委托的基本概念及操作可以参考对应的IAM文档《**委托其他云服务管理资源**》。

为了简化用户的委托授权操作，ModelArts增加了自动配置委托授权的支持，用户仅需在ModelArts控制台的“全局配置”页面中，为自己或特定用户配置委托即可。

说明

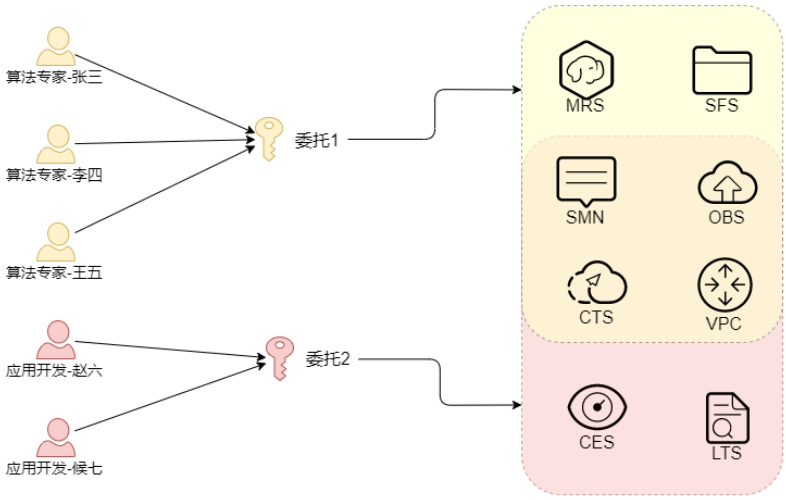
- 只有具备IAM委托管理权限的用户才可以进行此项操作，通常是IAM admin用户组的成员才具备此权限。
- 目前ModelArts的委托授权操作是分区域操作的，这意味着您需要在每个您所用到的区域均执行委托授权操作。

在ModelArts控制台的“全局配置”页面，单击“添加授权”后，系统会引导您为特定用户或所有用户进行委托配置，通常默认会创建一个名为“modelarts_agency_<用户名>_随机ID”的委托条目。在权限配置的区域，您可以选择ModelArts提供的预置配置，也可以自定义选择您所授权的策略。当然如果这两种形态对于您的诉求均过于粗犷，您也可以直接在IAM管理页面里创建完全由您进行精细化配置的委托（需要委托给ModelArts服务），然后在此页面的委托选择里使用“已有委托”“”（而非“新增委托”）。

至此，您应该已经发现了一个细节，ModelArts在使用委托时，是将其与用户进行关联的，用户与委托的关系是多对1的关系。这意味着，如果两个用户需要配置的委托一致，那么不需要为每个用户都创建一个独立的委托项，只需要将两个用户都“指向”同一个委托项即可。

版权所有 © 华为云计算技术有限公司

图 9-2 用户与委托对应关系



说明

每个用户必须关联委托才可以使用ModelArts，但即使委托所赋之权限不足，在API调用之初也不会报错，只有到系统具体使用到该功能时，才会发生问题。例如，用户在创建训练任务时打开了“消息通知”，该功能依赖SMN委托授权，但只有训练任务运行过程中，真正需要发送消息时，系统才会“出错”，而有些错误系统会选择“忽略”，另一些错误则可能导致任务直接失败。当您做深入的“权限最小化”限制时，请确保您在ModelArts上将要执行的操作仍旧有足够的权限。

严格授权模式

严格授权模式是指在IAM中创建的子用户必须由账号管理员显式在IAM中授权，才能访问ModelArts服务，管理员用户可以通过授权策略为普通用户精确添加所需使用的ModelArts功能的权限。

相对的，在非严格授权模式下，子用户不需要显式授权就可以使用ModelArts，管理员需要在IAM上为子用户配置Deny策略来禁止子用户使用ModelArts的某些功能。

账号的管理员用户可以在“全局配置”页面修改授权模式。

须知

如无特殊情况，建议优先使用严格授权模式。在严格授权模式下，子用户要使用ModelArts的功能都需经过授权，可以更精确的控制子用户的权限范围，达成权限最小化的安全策略。

用工作空间限制资源访问

工作空间是ModelArts面向企业客户提供的一个高阶功能，用于进一步将用户的资源划分在多个**逻辑隔离**的空间中，并支持以空间维度进行访问的权限限定。目前工作空间功能是“受邀开通”状态，作为企业用户您可以通过您对口的技术支持经理申请开通。

在开通工作空间后，系统会默认为您创建一个“default”空间，您之前所创建的所有资源，均在该空间下。当您创建新的工作空间之后，相当于您拥有了一个新的

版权所有 © 华为云计算技术有限公司

“ModelArts分身”，您可以通过菜单栏的左上角进行工作空间的切换，不同工作空间中的工作互不影响。

创建工作空间时，必须绑定一个企业项目。多个工作空间可以绑定到同一个企业项目，但一个工作空间**不可以**绑定多个企业项目。借助工作空间，您可以对不同用户的资源访问和权限做更加细致的约束，具体为如下两种约束：

- 只有被授权的用户才能访问特定的工作空间（在创建、管理工作空间的页面进行配置），这意味着，像数据集、算法等AI资产，均可以借助工作空间做访问的限制。
- 在前文提到的权限授权操作中，如果“选择授权范围方案”时设定为“指定企业项目资源”，那么该授权仅对绑定至该企业项目的工作空间生效。

说明

- 工作空间的约束与权限授权的约束是叠加生效的，意味着对于一个用户，必须同时拥有工作空间的访问权和训练任务的创建权限（且该权限覆盖至当前的工作空间），他才可以在这个空间里提交训练任务。
- 对于已经开通企业项目但没有开通工作空间的用户，其所有操作均相当于在“default”企业项目里进行，请确保对应权限已覆盖了名为default的企业项目。
- 对于未开通企业项目的用户，不受上述约束限制。

本章小结

对于ModelArts的权限管理，总结了如下几条关键点：

- 如果您是个人用户，则不需要考虑细粒度权限问题，您的账户默认具备使用ModelArts的所有权限。
- ModelArts平台的所有功能均通过IAM体系进行了权限管控，您可以通过标准的IAM**授权**动作，来对特定用户进行精细化的权限管控。
- 对于所有用户（包括个人用户），需要完成对ModelArts的**委托**授权（ModelArts > 全局配置 > 添加授权），才能使用特定的功能，否则会造成您的操作出现不可预期的错误。
- 对于开通了企业项目的用户，可以进一步申请开通ModelArts的**工作空间**，通过组合使用基础授权和工作空间，来达成更加复杂的权限控制目的。

版权所有 © 华为云计算技术有限公司

10 安全

10.1 责任共担

华为云秉承“将公司对网络和业务安全性保障的责任置于公司的商业利益之上”。针对层出不穷的云安全挑战和无孔不入的云安全威胁与攻击，华为云在遵从法律法规业界标准的基础上，以安全生态圈为护城河，依托华为独有的软硬件优势，构建面向不同区域和行业的完善云服务安全保障体系。

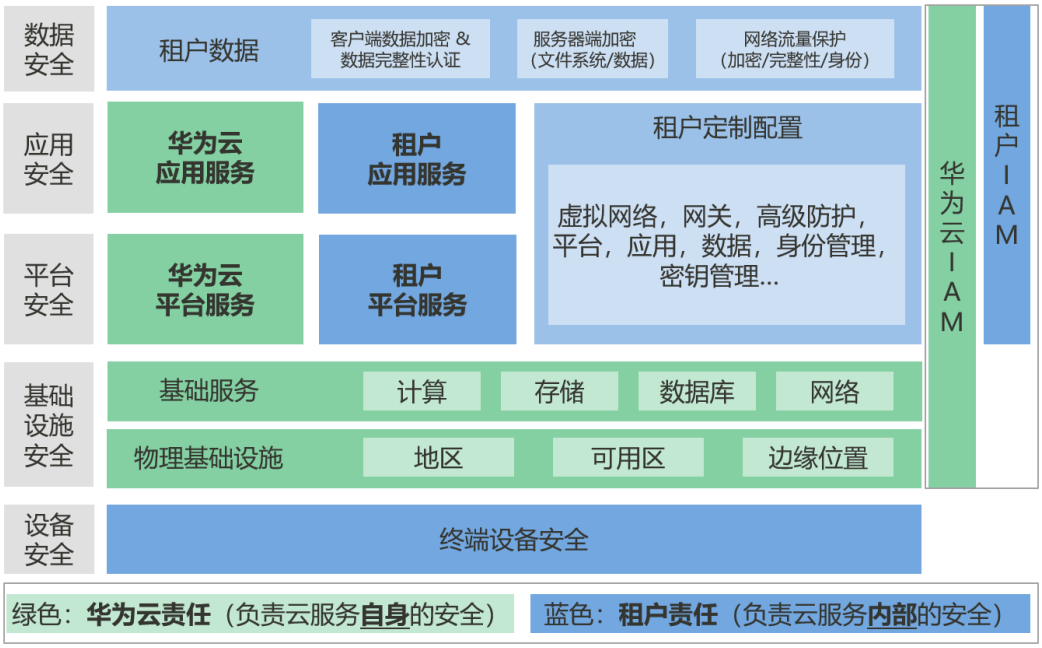
安全性是华为云与您的共同责任，如**图10-1**所示。

- **华为云**：负责云服务**自身**的安全，提供安全的云。华为云的安全责任在于保障其所提供的IaaS、PaaS和SaaS各类各项云服务自身的安全，涵盖华为云数据中心的物理环境设施和运行其上的基础服务、平台服务、应用服务等。这不仅包括华为云基础设施和各项云服务技术的安全功能和性能本身，也包括运维运营安全，以及更广义的安全合规遵从。
- **租户**：负责云服务**内部**的安全，安全地使用云。 华为云租户的安全责任在于对使用的IaaS、PaaS和SaaS类各项云服务内部的安全以及对租户定制配置进行安全有效的管理，包括但不限于虚拟网络、虚拟主机和访客虚拟机的操作系统，虚拟防火墙、API网关和高级安全服务，各项云服务，租户数据，以及身份账号和密钥管理等方面的安全配置。

 版权所有 © 华为云计算技术有限公司

《华为云安全白皮书》详细介绍华为云安全性的构建思路与措施，包括云安全战略、责任共担模型、合规与隐私、安全组织与人员、基础设施安全、租户服务与租户安全、工程安全、运维运营安全、生态安全。

图 10-1 华为云安全责任共担模型



10.2 资产识别与管理

资产识别

用户在AI Gallery中的资产包括用户发布的AI资产以及用户提供的一些个人信息。

AI资产包括但不限于文本、图形、数据、文章、照片、图像、插图、代码、AI算法、AI模型等。

用户的个人信息包括：

- 用户注册时提供的昵称、头像、邮箱。
- 用户参加实践时提供的姓名、手机号、邮箱。
- 用户伙伴注册时提供的企业信息。
- 用户发布资产时提供的联系人姓名、手机号、邮箱。

资产管理

对于用户发布在AI Gallery中的资产，AI Gallery会做统一的保存管理。

- 对于文件类型的资产，AI Gallery会将资产保存在AI Gallery官方的OBS桶内。
- 对于镜像类型的资产，AI Gallery会将资产保存在AI Gallery官方的SWR仓库内。

对于用户提供的一些个人信息，AI Gallery会保存在数据库中。个人信息中的敏感信息，如手机，邮箱等，AI Gallery会在数据库中做加密处理。

AI Gallery的更多介绍请参见《AI Gallery》。

版权所有 © 华为云计算技术有限公司

10.3 身份认证与访问控制

身份认证

用户访问ModelArts的方式有多种，包括ModelArts控制台、API、SDK，无论访问方式封装成何种形式，其本质都是通过ModelArts提供的REST风格的API接口进行请求。

ModelArts的接口均需要进行认证鉴权以此来判断是否通过身份认证。通过控制台发出的请求需要通过Token认证鉴权，调用API接口**认证鉴权**支持Token认证和AK/SK认证两种方式。

访问控制

ModelArts作为一个完备的AI开发平台，支持用户对其进行细粒度的权限配置，以达到精细化资源、权限管理之目的。为了支持客户对ModelArts的权限做精细化控制，提供了3个方面的能力来支撑，分别是：IAM权限控制、委托授权和工作空间。

- IAM权限控制

  用户使用ModelArts的任何功能，都需要通过IAM权限体系进行正确的权限授权。例如：用户希望在ModelArts创建训练作业，则该用户必须拥有"modelarts:trainJob:create" 的权限才可以完成操作（无论界面操作还是API调用）。

  管理员新创建的用户在没有配置细粒度授权策略时，默认具有ModelArts所有权限。如果需要控制用户的详细权限，管理员可以通过IAM为用户组配置细粒度授权策略，使用户获得策略定义的权限，操作对应云服务的资源。基于策略授权时，管理员可以按ModelArts的资源类型选择授权范围。详细的资源权限项可以参见API参考中的**权限策略和授权项**章节。

- 委托授权

  为了完成AI计算的各种操作，ModelArts在AI计算任务执行过程中需要访问用户的其他服务，例如训练过程中，需要访问OBS读取用户的训练数据。在这个过程中，就出现了ModelArts“代表”用户去访问其他云服务的情形。从安全角度出发，ModelArts代表用户访问任何云服务之前，均需要先获得用户的授权，而这个动作就是一个“委托”的过程。用户授权ModelArts再代表自己访问特定的云服务，以完成其在ModelArts平台上执行的AI计算任务。

  ModelArts服务不会保存用户的Token认证凭据，在后台作业中操作用户的资源（如OBS桶）前，需要用户通过IAM委托向ModelArts显式授权，ModelArts在需要时使用用户的委托获取临时认证凭据用于操作用户资源，具体配置见**配置访问授权**章节。

- 工作空间

  工作空间是ModelArts面向已经开通**企业项目**的企业客户提供的一个高阶功能，用于进一步将用户的资源划分在多个**逻辑隔离**的空间中，并支持以空间维度进行访问的权限限定。

  在开通工作空间后，系统会默认为您创建一个“default”空间，您之前所创建的所有资源，均在该空间下。当您创建新的工作空间之后，相当于您拥有了一个新的“ModelArts分身”，您可以通过菜单栏的左上角进行工作空间的切换，不同工作空间中的工作互不影响。ModelArts的用户需要为不同的业务目标开发算法、管理和部署模型，此时可以创建多个工作空间，把不同应用开发过程的输出内容划分到不同工作空间中，便于管理和使用。

版权所有 © 华为云计算技术有限公司

远程接入管理

使用本地IDE远程SSH连接ModelArts的Notebook开发环境时，需要用到密钥对进行鉴权认证。同时支持白名单访问控制，即设置允许远程接入访问这个Notebook的IP地址。

10.4 数据保护技术

ModelArts通过多种数据保护手段和特性，保障存储在ModelArts中的数据安全可靠。

数据保护手段	说明
静态数据保护	对于AI Gallery收集的用户个人信息中的敏感信息，如用户邮箱和手机号，AI Gallery在数据库中做了加密处理。其中，加密算法采用了国际通用的AES算法。
传输中的数据保护	在ModelArts中导入AI应用时，支持用户自己选择HTTP和HTTPS两种传输协议，为保证数据传输的安全性，推荐用户使用更加安全的HTTPS协议。
数据完整性检查	推理部署功能模块涉及到的用户模型文件和发布到AIGallery的资产在上传过程中，有可能会因为网络劫持、数据缓存等原因，存在数据不一致的问题。ModelArts提供通过计算SHA256值的方式对上传下载的数据进行一致性校验。
数据隔离机制	在ModelArts的开发环境中创建Notebook实例时，数据存储是按照租户隔离，租户之间互相看不到数据。

10.5 审计与日志

审计

云审计服务（Cloud Trace Service，CTS），是华为云安全解决方案中专业的日志审计服务，提供对各种云资源操作记录的收集、存储和查询功能，可用于支撑安全分析、合规审计、资源跟踪和问题定位等常见应用场景。

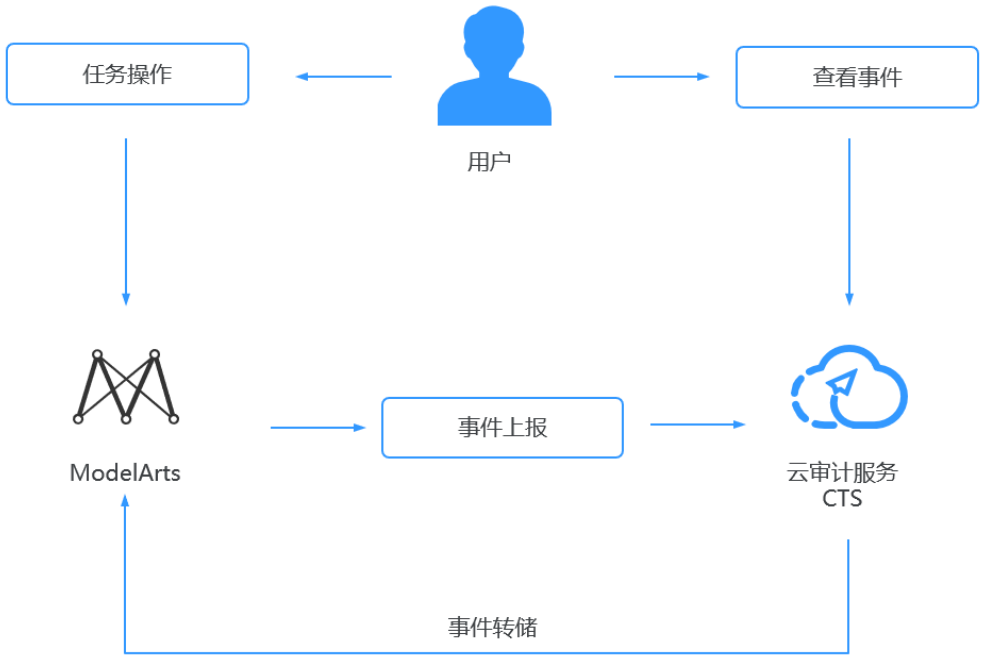
用户开通云审计服务并创建和配置追踪任务后，CTS可记录ModelArts的管理事件和数据事件用于审计。

CTS的详细介绍和开通配置方法，请参见**CTS快速入门**。

CTS支持追踪的ModelArts管理事件和数据事件列表，请参见**支持云审计的关键操作**、**开发环境支持审计的关键操作列表**、**训练作业支持审计的关键操作列表**、**AI应用管理支持审计的关键操作列表**、**服务管理支持审计的关键操作列表**。

版权所有 © 华为云计算技术有限公司

图 10-2 云审计服务



数据管理支持审计的关键操作列表

表 10-1 数据管理支持审计的关键操作列表

操作名称	资源类型	事件名称
创建数据集	dataset	createDataset
删除数据集	dataset	deleteDataset
更新数据集	dataset	updateDataset
发布数据集版本	dataset	publishDatasetVersion
删除数据集版本	dataset	deleteDatasetVersion
同步数据源	dataset	syncDataSource
导出数据集	dataset	exportDataFromDataset
创建自动标注任务	dataset	createAutoLabelingTask
创建自动分组任务	dataset	createAutoGroupingTask
创建自动部署任务	dataset	createAutoDeployTask
导入样本到数据集	dataset	importSamplesToDataset
创建数据集标签	dataset	createLabel
更新数据集标签	dataset	updateLabel

版权所有 © 华为云计算技术有限公司

操作名称	资源类型	事件名称
删除数据集标签	dataset	deleteLabel
删除数据集标签和对应的样本	dataset	deleteLabelWithSamples
添加样本	dataset	uploadSamples
删除样本	dataset	deleteSamples
停止自动标注任务	dataset	stopTask
创建团队标注任务	dataset	createWorkforceTask
删除团队标注任务	dataset	deleteWorkforceTask
启动团队标注验收的任务	dataset	startWorkforceSamplingTask
通过/驳回/取消验收任务	dataset	updateWorkforceSamplingTask
提交验收任务的样本评审意见	dataset	acceptSamples
给样本添加标签	dataset	updateSamples
发送邮件给团队标注任务的成员	dataset	sendEmails
接口人启动团队标注任务	dataset	startWorkforceTask
更新团队标注任务	dataset	updateWorkforceTask
给团队标注样本添加标签	dataset	updateWorkforceTaskSamples
团队标注审核	dataset	reviewSamples
创建标注成员	workforce	createWorker
更新标注成员	workforce	updateWorker
删除标注成员	workforce	deleteWorker
批量删除标注成员	workforce	batchDeleteWorker
创建标注团队	workforce	createWorkforce
更新标注团队	workforce	updateWorkforce
删除标注团队	workforce	deleteWorkforce
自动创建IAM委托	IAM	createAgency
标注成员登录labelConsole标注平台	labelConsoleWorker	workerLoginLabelConsole

版权所有 © 华为云计算技术有限公司

操作名称	资源类型	事件名称
标注成员登出labelConsole标注平台	labelConsoleWorker	workerLogOutLabelConsole
标注成员修改labelConsole平台密码	labelConsoleWorker	workerChangePassword
标注成员忘记labelConsole平台密码	labelConsoleWorker	workerForgetPassword
标注成员通过url重置labelConsole标注密码	labelConsoleWorker	workerResetPassword

开发环境支持审计的关键操作列表

表 10-2 开发环境支持审计的关键操作列表

操作名称	资源类型	事件名称
创建Notebook	Notebook	createNotebook
删除Notebook	Notebook	deleteNotebook
打开Notebook	Notebook	openNotebook
启动Notebook	Notebook	startNotebook
停止Notebook	Notebook	stopNotebook
更新Notebook	Notebook	updateNotebook
删除NotebookApp	NotebookApp	deleteNotebookApp
切换CodeLab规格	NotebookApp	updateNotebookApp

训练作业支持审计的关键操作列表

表 10-3 训练作业支持审计的关键操作列表

操作名称	资源类型	事件名称
创建训练作业	ModelArtsTrainJob	createModelArtsTrainJob
创建训练作业版本	ModelArtsTrainJob	createModelArtsTrainVersion
停止训练作业	ModelArtsTrainJob	stopModelArtsTrainVersion
更新训练作业描述	ModelArtsTrainJob	updateModelArtsTrainDesc

版权所有 © 华为云计算技术有限公司

操作名称	资源类型	事件名称
删除训练作业版本	ModelArtsTrainJob	deleteModelArtsTrainVersion
删除训练作业	ModelArtsTrainJob	deleteModelArtsTrainJob
创建训练作业参数	ModelArtsTrainConfig	createModelArtsTrainConfig
更新训练作业参数	ModelArtsTrainConfig	updateModelArtsTrainConfig
删除训练作业参数	ModelArtsTrainConfig	deleteModelArtsTrainConfig
创建可视化作业	ModelArtsTensorboardJob	createModelArtsTensorboardJob
删除可视化作业	ModelArtsTensorboardJob	deleteModelArtsTensorboardJob
更新可视化作业描述	ModelArtsTensorboardJob	updateModelArtsTensorboardDesc
停止可视化作业	ModelArtsTensorboardJob	stopModelArtsTensorboardJob
重启可视化作业	ModelArtsTensorboardJob	restartModelArtsgTensorboardJob

AI 应用管理支持审计的关键操作列表

表 10-4 AI 应用管理支持审计的关键操作列表

操作名称	资源类型	事件名称
创建AI应用	model	addModel
更新AI应用	model	updateModel
删除AI应用	model	deleteModel
添加转换任务	convert	addConvert
更新转换任务	convert	updateConvert
删除转换任务	convert	deleteConvert

版权所有 © 华为云计算技术有限公司

服务管理支持审计的关键操作列表

表 10-5 服务管理支持审计的关键操作列表

操作名称	资源类型	事件名称
部署服务	service	addService
删除服务	service	deleteService
更新服务	service	updateService
启停服务	service	startOrStopService
添加用户访问密钥	service	addAkSk
删除用户访问密钥	service	deleteAkSk
创建专属资源池	cluster	createCluster
删除专属资源池	cluster	deleteCluster
添加专属资源池节点	cluster	addClusterNode
删除专属资源池节点	cluster	deleteClusterNode
获取专属资源池创建结果	cluster	createClusterResult

AI Gallery 支持审计的关键操作列表

表 10-6 AI Gallery 支持审计的关键操作列表

操作名称	资源类型	事件名称
发布资产	ModelArts_Market	create_content
修改资产信息	ModelArts_Market	modify_content
发布资产新版本	ModelArts_Market	add_version
订阅资产	ModelArts_Market	subscription_content
取消收藏资产	ModelArts_Market	cancel_star_content
点赞资产	ModelArts_Market	like_content
取消点赞资产	ModelArts_Market	cancel_like_content
发布实践	ModelArts_Market	publish_activity
报名实践	ModelArts_Market	regist_activity
修改个人资料	ModelArts_Market	update_user

版权所有 © 华为云计算技术有限公司

日志

出于分析或审计等目的，用户可以开启ModelArts的日志记录功能。在您开启了云审计服务后，系统会记录ModelArts的相关操作，且控制台保存最近7天的操作记录。本节介绍如何在云审计服务管理控制台查看最近7天的操作记录。

对接云审计服务的配置方法请参见**查看审计日志**章节。

10.6 服务韧性

韧性特指安全韧性，即云服务受攻击后的韧性，不含可靠性、可用性。本章主要阐述ModelArts服务受入侵的检测响应能力、防抖动的能力、域名合理使用、内容安全检测等能力。

安全防护套件覆盖和使用堡垒机，增强入侵检测和防御能力

ModelArts服务部署主机层、应用层、网络层和数据层的安全防护套件。及时检测主机层、应用层、网络层和数据层的安全入侵行为。

- ModelArts服务涉及对互联网开放的Web应用，采用了统一推荐的Web安全组件防范Web安全风险，并且通过WAF进行安全防护。
- 所有承载ModelArts服务的主机部署了主机安全防护产品。包括不限于华为自研HSS或计算安全平台CSP。
- ModelArts服务部署了漏洞扫描服务并自行进行例行扫描，能快速发现漏洞并能及时修复。
- ModelArts服务通过统一的安全管控平台对云上资源进行安全运维。
- ModelArts服务部署了态势感知服务，以感知攻击现状，还原攻击历史，同时及时发现合规风险，对威胁告警及时响应。
- ModelArts承载关键业务的对外开放EIP部署了高防服务，以防大流量攻击。
- ModelArts对存放关键数据的数据库部署了数据库安全服务。

云服务防抖动和遭受攻击后的应急响应/恢复策略

ModelArts服务具备租户资源隔离能力，避免单租户资源被攻击导致爆炸半径大，影响其他租户。

- ModelArts服务具备资源池和隔离能力，避免单租户资源被攻击导致爆炸半径过大风险。
- ModelArts服务定义并维护了性能规格用于自身的抗攻击性。例如：设置API访问限制，防止恶意接口调用等场景。
- ModelArts服务在攻击场景下，具备告警能力及自我保护能力。
- ModelArts服务提供了业务异常行为感知能力。例如运营平台异常数据感知，安全日志集成等。
- ModelArts服务具备遭受攻击时的风险控制和应急响应能力。例如快速识别恶意租户，恶意IP。
- ModelArts服务具备攻击流量停止后，快速恢复业务的能力。

版权所有 © 华为云计算技术有限公司

云服务域名使用安全及租户内容安全策略

ModelArts服务使用的租户可见域名、租户不可见域名均满足如下安全相关要求，避免了域名使用过程中的合规和钓鱼风险。其中：

租户可见域名：指租户可访问的域名，需要格外重视安全性和合规性。

租户不可见域名：指华为云服务在内网相互调用使用的域名，外部用户无法访问到对应的权威DNS服务器；或者Internet受限访问域名，只允许华为办公网络黄&绿区华为员工及合作方或外包人员访问的域名。

- 华为云基础域名安全使用，避免直接为租户分配基础域名。
- 华为云服务在内网互相调用使用的域名，避免使用外部已备案域名。

10.7 监控安全风险

ModelArts支持监控ModelArts在线服务和对应模型负载，执行自动实时监控、告警和通知操作，帮助用户更好地了解服务和模型的各项性能指标。详细内容请参见**ModelArts支持的监控指标**。

10.8 故障恢复

ModelArts全球基础设施围绕华为云区域和可用区构建。华为云区域提供多个在物理上独立且隔离的可用区，这些可用区通过延迟低、吞吐量高且冗余性高的网络连接在一起。利用可用区，您可以设计和操作在可用区之间无中断地自动实现故障转移的应用程序和数据库。与传统的单个或多个数据中心基础设施相比，可用区具有更高的可用性、容错性和可扩展性。

ModelArts通过对DB的数据进行备份，保证在原数据被破坏或损坏的情况下可以恢复业务。

开发环境故障恢复

针对用户创建的Notebook计算实例，后台计算节点故障后会立即自动迁移到其他可用节点上，实例状态会自动恢复。针对数据存储部分，提供了云硬盘存储挂载方式，华为云云硬盘提供高可靠、高性能、规格丰富并且可弹性扩展的块存储服务，数据持久性高达99.9999999%。

训练故障自动恢复

用户在训练模型过程中，存在因硬件故障而产生的训练失败场景。针对硬件故障场景，ModelArts提供容错检查功能，帮助用户隔离故障节点，优化用户训练体验。

容错检查包括两个检查项：环境预检测与硬件周期性检查。当环境预检查或者硬件周期性检查任一检查项出现故障时，隔离故障硬件并重新下发训练作业。针对于分布式场景，容错检查会检查本次训练作业的全部计算节点。

推理部署故障恢复

用户部署的在线推理服务运行过程中，如发生硬件故障导致推理实例故障，ModelArts会自动检测到并迁移受影响实例到其它可用节点，实例启动后恢复推理请求处理能力。故障的硬件节点会自动隔离不再调度和运行推理服务实例。

版权所有 © 华为云计算技术有限公司

10.9 更新管理

ModelArts 在线服务更新

对于已部署的推理服务，ModelArts支持通过更换AI应用的版本号，实现服务升级。

推理服务有三种升级模式：全量升级、滚动升级（扩实例）和滚动升级（缩实例）。了解三种升级模式的流程，请参见图10-3。

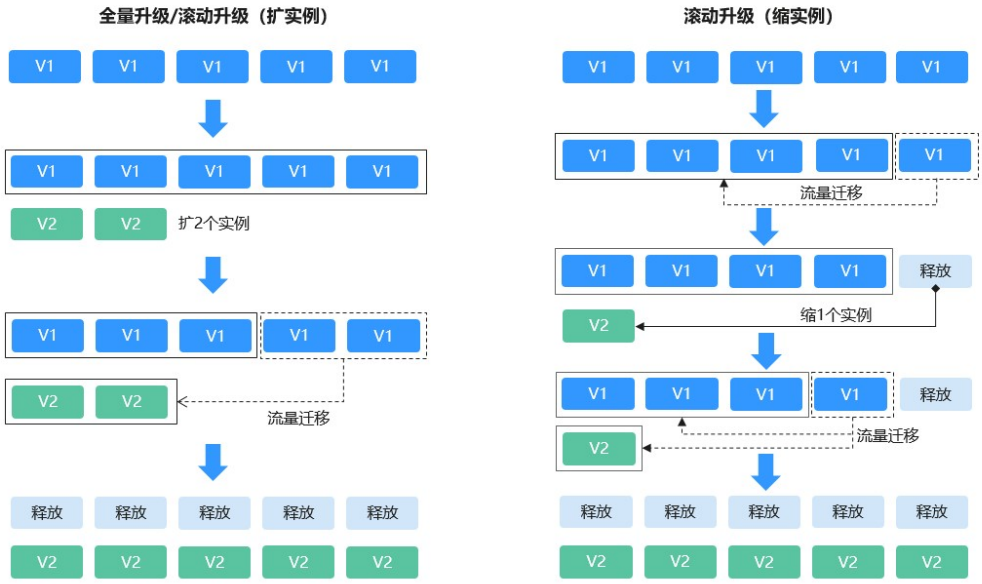
- 全量升级

  需要额外的双倍的资源，先全量创建新版本实例，然后再下线旧版本实例。
- 滚动升级（扩实例）

  需额外消耗部分实例资源用于滚动升级，扩实例越大，升级速度越快。
- 滚动升级（缩实例）

  通过腾出部分实例资源用于滚动升级，缩实例数越大，升级速度越快，造成业务中断可能性越大。

图 10-3 推理服务升级流程



推理服务更新升级的具体操作请参见升级服务。

镜像更新升级

ModelArts包含开发环境、训练管理、推理部署三个功能模块，三个模块采用统一的流程提供基础镜像。这些镜像会不定期更新升级，修复已知漏洞。

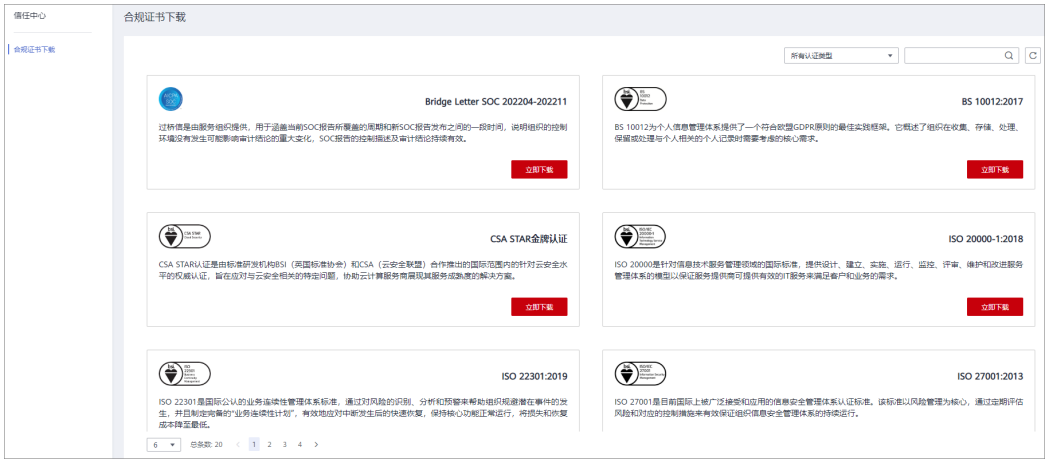
版权所有 © 华为云计算技术有限公司

10.10 认证证书

合规证书

华为云服务及平台通过了多项国内外权威机构（ISO/SOC/PCI等）的安全合规认证，用户可自行**申请下载**合规资质证书。

图 10-4 合规证书下载



资源中心

华为云还提供以下资源来帮助用户满足合规性要求，具体请查看**资源中心**。

图 10-5 资源中心

版权所有 © 华为云计算技术有限公司

10.11 安全边界

云服务的责任共担模型是一种合作方式，其中云服务提供商和云服务客户共同承担云服务的安全和合规性责任。这种模型是为了确保云服务的安全性和可靠性而设计的。

根据责任共担模型，云服务提供商和云服务客户各自有一些责任。云服务提供商负责管理云基础架构，提供安全的硬件和软件基础设施，并确保云基础架构的可用性。而云服务客户则需要负责保护自己的数据和应用程序，以及遵守相关的合规性要求。

具体而言，云服务提供商应该提供以下服务和功能：

- 建立和维护安全的基础设施，包括网络、服务器和存储设备等。
- 提供安全的底层基础平台，保证底层环境的运行时安全。
- 提供安全的身份验证和访问控制机制，以确保只有授权用户可以访问云服务，保证租户之前的相互隔离。
- 提供可靠的备份和灾难恢复机制，以确保数据不会因为硬件故障或自然灾害等原因而丢失。
- 提供透明的安全监控和事件响应服务，及时的安全更新和漏洞修补。

而云服务客户则需要执行以下任务：

- 将数据和应用程序加密，以保护数据的机密性和完整性。
- 确保AI应用的相关软件都得到及时的安全更新和漏洞修补。
- 遵守相关的合规性要求，如GDPR、HIPAA、PCI DSS等。
- 进行适当的访问控制，以确保只有授权用户可以访问管理在线服务等相关资源。
- 监控和报告任何异常活动，并及时采取措施。

推理部署安全责任

- 提供商
  - 底层ecs相关的系统补丁修复
  - k8s的版本更新和漏洞修复
  - 虚拟机OS的版本生命周期维护
  - ModelArts推理平台自身的安全合规性
  - 容器应用服务加固
  - 模型运行环境的版本更新和漏洞定期修复
- 客户侧
  - 资源的授权，访问控制
  - 保证应用的供应链安全，依赖和自身的安全性，安全扫描、审计和准入校验机制，保证制品源头的安全性
  - 权限配置和凭证下发权限最小化
  - AI应用运行时（自定义镜像，OBS模型和依赖）的安全性
  - 及时更新修复安全问题
  - 凭证等敏感数据的安全存储

版权所有 © 华为云计算技术有限公司

推理部署安全最佳实践

- 外部依赖服务

  ModelArts推理使用中需要用到一些其他的云服务，当您需要授权时，可以根据实际所需的权限范围进行自定义授权，其中模型管理依赖OBS相关权限，租户可以细化权限到具体ModelArts使用的桶。

- 内部资源授权

  ModelArts推理当前已支持细粒度授权，租户可以根据实际的权限要求对子用户进行相应的权限配置，限制某些资源的管理，实现权限最小化。

- AI应用管理

  使用从训练或者从OBS中选择创建AI应用，推荐用户使用动态加载的方式导入，动态加载实现了模型和镜像的解耦，便于进行模型资产的保护。用户需要及时更新AI应用的相关依赖包，解决开源或者第三方包的漏洞。AI应用相关的敏感信息，需要解耦开，在“在线服务”部署时进行相应配置。请选择ModelArts推荐的运行时环境，旧的运行环境官方已停止维护，可能存在安全漏洞。

  使用从容器镜像中选择创建AI应用时，在构建镜像环节，需要采用业界公开的可信基础镜像，例如来自OpenEuler，Ubuntu等的发布镜像，镜像运行用户需要创建非root普通用户，不能采用root用户直接运行。镜像中只安装运行时依赖的安全包，减少镜像的大小，同时安装包需要更新到最新的无漏洞版本。敏感信息和镜像解耦，可以在服务部署时配置，不能直接硬编码在Dockerfile中。定期针对镜像进行安全扫描，及时安装补丁修复漏洞。增加健康检查接口，确保健康检查可以正常返回业务状态，便于告警和故障恢复。容器应该采用https的安全传输通道，并使用业界推荐的加密套件保证业务数据的安全性。

- 部署上线

  部署服务时，需要注意为服务设置合适计算节点规格，防止服务因资源不足而过载或者资源过大而浪费。尽量避免在容器中监听其他端口，有本地内部需要访问的其他端口，监听在localhost上。避免通过环境变量传递敏感信息，需要通过加密组件进行加密后再通过环境变量配置。

  部署在线服务，当打开APP认证时，app认证密钥是在线服务的另一个访问凭据，需要妥善保存app密钥，防止泄露。

版权所有 © 华为云计算技术有限公司

11 配额说明

本服务应用的基础设施如下：

- 弹性云服务器
- 云硬盘
- 虚拟私有云
- 云容器引擎

其配额查看及修改请参见**关于配额**。

版权所有 © 华为云计算技术有限公司