# ModelArts

# 产品介绍

**文档版本** 01

**发布日期** 2024-08-14

**华为云计算技术有限公司**

### 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



# 1 图解 ModelArts

## 1.1 初识 ModelArts

# 1.2 初识 Workflow

# 2 什么是 ModelArts

ModelArts是华为云提供的一站式AI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。

## 产品形态

ModelArts提供多种产品形态，如下表所示。

**表 2-1** ModelArts 产品形态介绍

| 产品形态 | 产品定位 | 使用场景 |
| --- | --- | --- |
| ModelArts Standard | 面向AI开发者的一站式开发平台， 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、模型管理、部署上线等端到端的AI开发工具链，实现AI全流程生命周期管理。 | 面向有AI开发平台诉求的用户。 |
| ModelArts MaaS | 提供端到端的大模型生产工具链和昇腾算力资源，并预置了当前主流的第三方开源大模型。支持大模型数据生产、微调、提示词工程、应用编排等功能。 | 用户无需自建平台，可以基于MaaS平台开箱即用，对预置大模型进行二次开发，用于生产商用。 |
| ModelArts Lite-Server | 面向云主机资源型用户，基于裸金属服务器进行封装，可以通过弹性公网IP直接访问操作服务器。 | 适用于已经自建AI开发平台，仅有算力需求的用户，提供高性价比的AI算力，并预装主流AI开发套件以及自研的加速插件。 |

| 产品形态 | 产品定位 | 使用场景 |
| --- | --- | --- |
| ModelArts Lite-Cluster | 面向k8s资源型用户，提供k8s原生接口，用户可以直接操作资源池中的节点和k8s集群。 | 适用于已经自建AI开发平台，仅有算力需求的用户。要求用户具备k8s基础知识和技能。 |
| ModelArts Edge | 为客户提供了统一边缘部署和管理能力，支持统一纳管异构边缘设备，提供AI应用部署、AI应用和节点管理、资源池与负载均衡、应用商用保障等能力，帮助客户快速构建高性价比的边云协同AI解决方案。 | 适用于边缘部署场景。 |
| AI Gallery | AI Gallery百模千态社区，为用户提供优质的昇腾云AI模型开发体验和丰富的社区资源。 | 适用于AI开发探索。 |

## 产品架构

ModelArts产品架构请参考图2-1。

**图 2-1** ModelArts 产品架构


- 算力层提供全系列昇腾硬件，万卡级大规模集群管理能力，提供资源负载调度管理能力，兼容业界主流AI开发调试、训练推理框架。
- AI平台层提供端到端的AI开发工具链，支持开发者一站式完成模型开发和上线，并提供高效的资源管理能力，支持自动化故障恢复，提升AI模型开发、训练、上线全流程效率。

- AI开发工具链层提供端到端的大模型开发工具链，支持主流优质开源大模型“开箱即用”，提供大模型开发套件，提升大模型开发效率并缩短开发周期。

## 访问方式

ModelArts基于不同的产品形态提供了多种访问方式。

- **管理控制台方式**

  ModelArts Standard支持通过管理控制台访问，包含自动学习、数据管理、开发环境、模型训练、AI应用管理、部署上线等功能，您可以在管理控制台端到端完成您的AI开发。

  ModelArts MAAS可以通过管理控制台访问，包括大模型数据生产、微调、提示词工程、应用编排等功能。

- **SDK方式**

  如果您需要将ModelArts Standard功能集成到第三方系统，用于二次开发，可选择调用SDK方式完成目的。ModelArts的SDK是对ModelArts Standard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ModelArts Standard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 Standard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**》。

- **云原生方式**

  如果您使用的是ModelArts Lite Server形态，您可以通过弹性公网IP直接访问云主机，详情请参见《**ModelArts Lite用户指南**》。

  如果您使用的是ModelArts Lite Cluster形态，您可以通过k8s原生接口操作集群，详情请参见《**ModelArts Lite用户指南**》。

# 3 产品优势

ModelArts服务具有以下产品优势。

**稳定安全的算力底座，极快至简的模型训练**

- 支持万节点计算集群管理
- 大规模分布式训练能力，加速大模型研发
- 提供高性价比国产算力
- 多年软硬件经验沉淀，AI场景极致优化
- 加速套件，训练、推理、数据访问多维度加速

**一站式端到端生产工具链，一致性开发体验**

- 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。
- 支持本地 IDE+ModelArts 插件远程开发能力，线上线下协同开发，开发训练一体化架构，支持大模型分布式部署及推理
- 统一管理 AI 开发全流程，提升开发效率，记录模型构建实验全流程

**多场景部署，灵活满足业务需求**

- 支持云端/边端部署等多种生产环境
- 支持在线推理、批量推理、边缘推理多形态部署

**AI工程化能力，支持AI全流程生命周期管理**

- 支持MLOps能力，提供数据诊断、模型监测等分析能力，训练智能日志分析与诊断

**容错能力强，故障恢复快**

- 提供机柜、节点、加速卡、任务多场景故障感知和检测
- 提供节点级、作业级、容器级，多级故障恢复，保障千卡作业稳定训练

**多种资源形态**

- 集群模式，开箱即提供好Kubernetes集群，直接使用，方便高效
- 节点模式，客户可采用开源或自研框架，自行构建集群，更强的掌控力和灵活性

**零改造迁移**

- 提供业界通用的k8s接口使用资源，业务跨云迁移无压力
- SSH直达节点和容器，一致体验

# 4 应用场景

本节介绍ModelArts服务的主要应用场景。

## 大模型

支持三方开源大模型，实现智能回答、聊天机器人、自动摘要、机器翻译、文本分类等任务。

## AIGC

提供AIGC场景化解决方案，辅助创作文案、图像、音视频等数字内容。

## 自动驾驶

实现车辆自主感知环境、规划路径和控制行驶。支持自动驾驶场景PB级数据下模型高效训练，助力自动驾驶特有的感知、规控、仿真生成等全链路相关算法深度优化并快速迭代。

## 内容审核

深入业务场景，提供完备成熟的内容审核/CV场景快速昇腾迁移的方案，高效解决业务内容审核的算力/国产化需求，助力企业业务稳健发展。

## 政府

提高公共服务的效率和质量，加强公共安全，优化政策方案和决策过程等。

## 金融

为金融机构带来更加高效、智能、精准的服务。

## 矿山

提供端到端AI生产线能力和高性能AI算力，提升大模型推理效率，为矿山行业带来更高效、智能、安全和可持续的生产方案。

## 铁路

实现列车智能调度、设备故障预测、铁路线路安全监控等功能。

## 医疗

报告智能解读、互联网检验以及居民全周期健康管理等领域的应用，为用户提供更加多元化、智慧化、精益化的服务。

# 5 功能介绍

## 5.1 Standard 功能介绍

### 5.1.1 Standard 自动学习

ModelArts通过机器学习的方式帮助不具备算法开发能力的业务开发者实现算法的开发，基于迁移学习、自动神经网络架构搜索实现模型自动生成，通过算法实现模型训练的参数自动化选择和模型自动调优的自动学习功能，让零AI基础的业务开发者可快速完成模型的训练和部署。

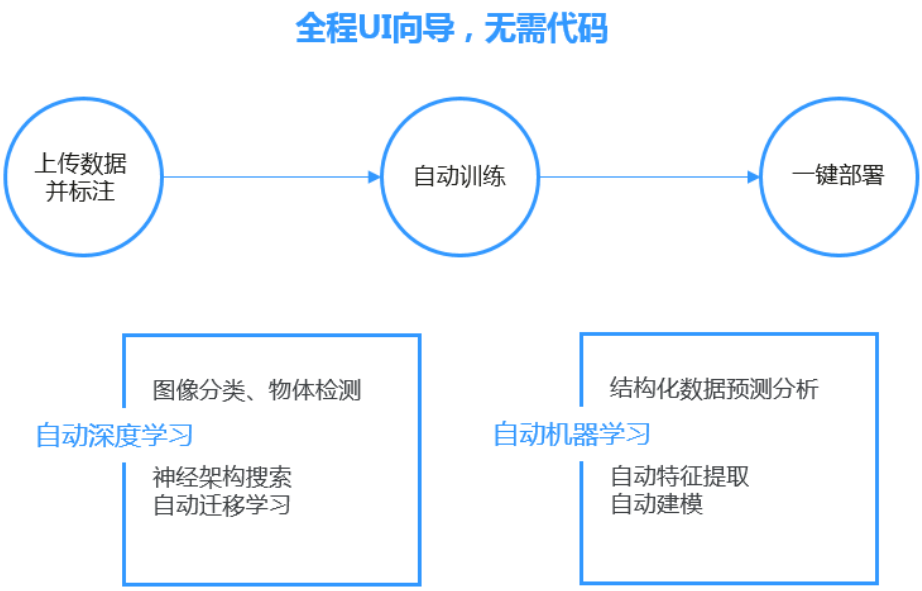
**ModelArts自动学习，为入门级用户提供AI零代码解决方案**

- 支持图片分类、物体检测、预测分析、声音分类场景
- 自动执行模型开发、训练、调优和推理机器学习的端到端过程
- 根据最终部署环境和开发者需求的推理速度，自动调优并生成满足要求的模型

**ModelArts自动学习，为资深级用户提供模板化开发能力**

- 提供“自动学习白盒化”能力，开放模型参数、自动生成模型，实现模板化开发，提高开发效率
- 采用自动深度学习技术，通过迁移学习（只通过少量数据生成高质量的模型），多维度下的模型架构自动设计（神经网络搜索和自适应模型调优），和更快、更准的训练参数自动调优自动训练
- 采用自动机器学习技术，基于信息熵上限近似模型的树搜索最优特征变换和基于信息熵上限近似模型的贝叶斯优化自动调参，从企业关系型（结构化）数据中，自动学习数据特征和规律，智能寻优特征&ML模型及参数，准确性甚至达到专家开发者的调优水平

**图 5-1** 自动学习流程



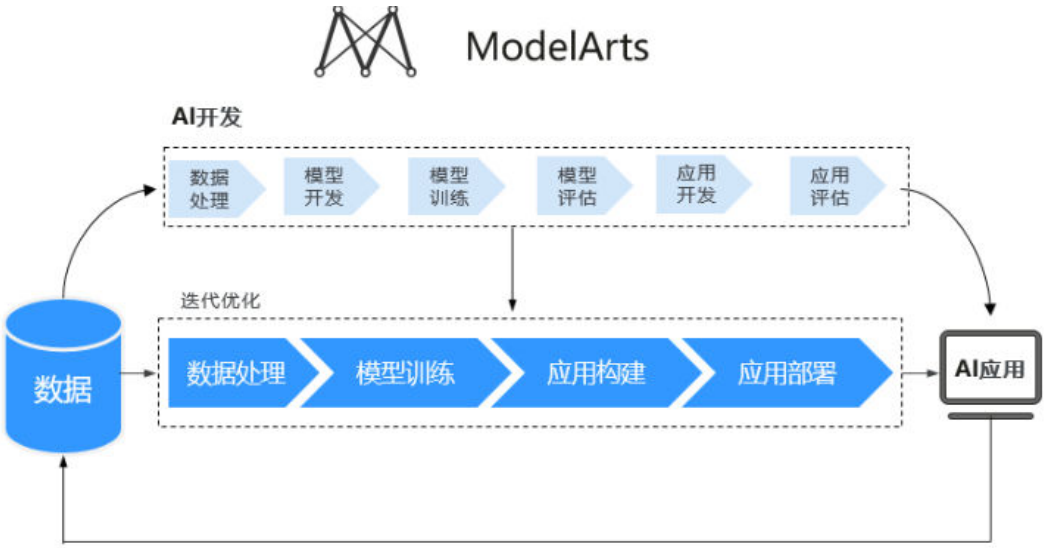

## 5.1.2 Standard Worflow

Workflow是开发者基于实际业务场景开发用于部署模型或应用的流水线工具，核心是将完整的机器学习任务拆分为多步骤工作流，每个步骤都是一个可管理的组件，可以单独开发、优化、配置和自动化。Workflow有助于标准化机器学习模型生成流程，使团队能够大规模执行AI任务，并提高模型生成的效率。

**ModelArts Workflow提供标准化MLOps解决方案，降低模型训练成本**

- 支持数据标注、数据处理、模型开发/训练、模型评估、应用开发、应用评估等步骤
- 自动协调工作流步骤之间的所有依赖项，提供运行记录、监控、持续运行等功能
- 针对工作流开发，Workflow提供流水线需要覆盖的功能以及功能需要的参数描述，供用户使用SDK对步骤以及步骤之间的关系进行定义
- 针对工作流复用，用户可以在开发完成后将流水线固化下来，提供下次或其他人员使用，同时无需关注流水线中包含什么算法或如何实现

**图 5-2** Workflow 流程




## 5.1.3 Standard 数据管理

ModelArts Standard数据管理提供了一套高效便捷的管理和标注数据框架。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，适用于计算机视觉、自然语言处理、音视频分析等AI项目场景。

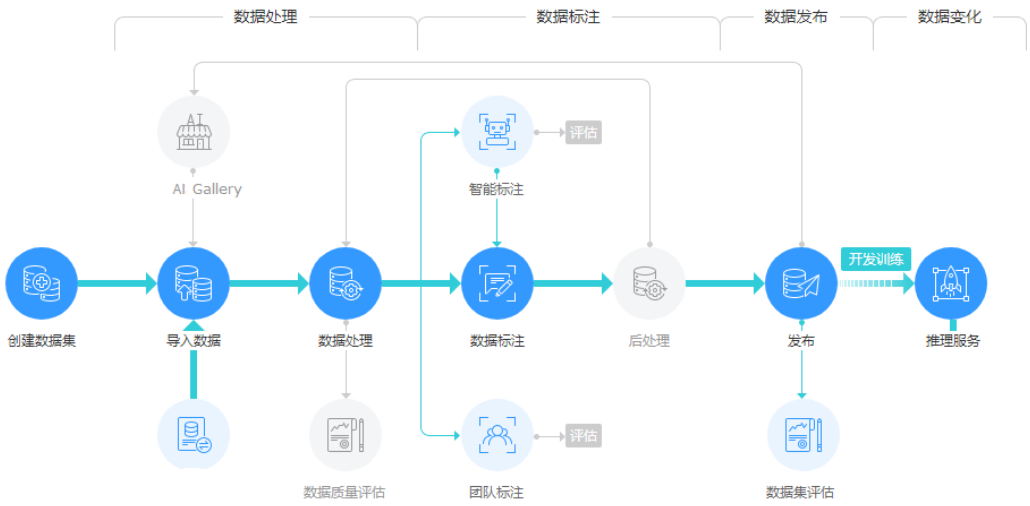
**说明**

ModelArts Standard数据管理模块重构中，当前能力不做演进，将结合大模型时代能力进行全新升级，敬请期待。

**ModelArts Standard数据管理支持多维度数据管理能力**

- 数据集管理：提供数据集创建、数据预览、数据集版本管理等能力
- 数据标注：提供在线标注能力，包含图像分类、目标检测、音频分割、文本三元组等标注场景；提供图片智能标注方案，提升标注效率；提供团队标注能力，支持多人协同标注与标注任务的审核验收
- 数据处理：提供数据清洗、数据校验、数据增强、数据选择等分析处理能力

**图 5-3** 数据标注全流程




## 5.1.4 Standard 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 Standard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 Standard 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

ModelArts CodeLab（JupyterLab），让AI探索&教学更简单

- 云原生Notebook，案例内容秒级接入与分享
- Serverless化实例管理，资源自动回收
- 免费算力，规格按需切换

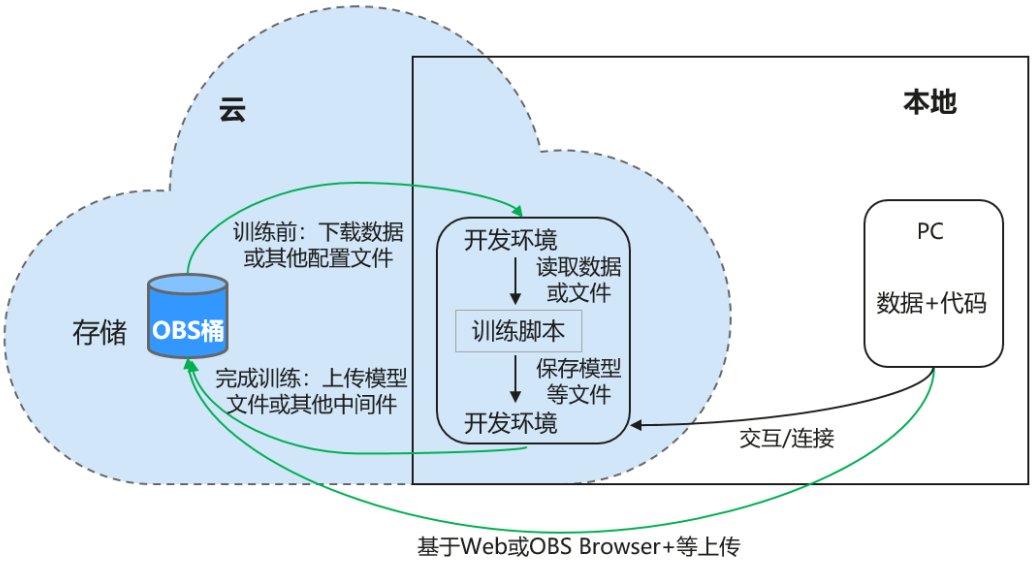
### 亮点特性 1：远程开发 - 支持本地 IDE 远程访问 Notebook

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。

**图 5-4** 本地 IDE 远程访问 Notebook 开发环境




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：开发环境支持一键镜像保存

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

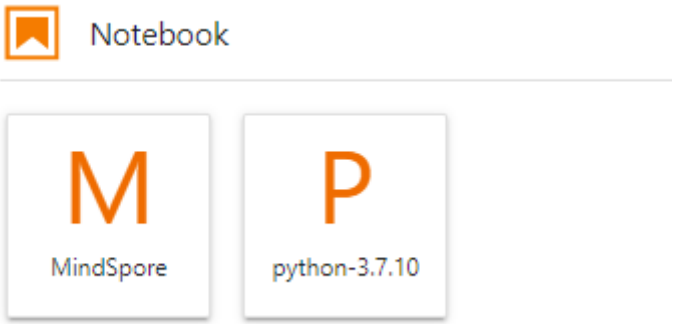
每个镜像预置的AI引擎和版本是固定的，在创建Notebook实例时明确AI引擎和版本，包括适配的芯片。

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。

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。

开发环境提供的预置镜像主要包含：

- 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。

## 5.1.5 Standard 模型训练

ModelArts Standard模型训练提供容器化服务和计算资源管理能力，负责建立和管理机器学习训练工作负载所需的基础设施，减轻用户的负担，为用户提供灵活、稳定、易用和极致性能的深度学习训练环境。通过ModelArts Standard模型训练，用户可以专注于开发、训练和微调模型。

**ModelArts Standard模型训练支持大规模训练作业，提供高可用的训练环境**

- 支持单机多卡、多机多卡的分布式训练，有效加速训练过程
- 支持训练作业的故障感知、故障诊断与故障恢复，包含硬件故障与作业卡死故障，并支持进程级恢复、容器级恢复与作业级恢复，提供容错与恢复能力，保障用户训练作业的长稳运行
- 提供训练作业断点续训与增量训练能力，即使训练因某些原因中断，也可以基于checkpoint接续训练，保障需要长时间训练的模型的稳定性和可靠性，避免重头训练耗费的时间与计算成本
- 支持训练数据使用SFS Turbo文件系统进行数据挂载，训练作业产生的中间和结果等数据可以直接高速写入到SFS Turbo缓存中，并可被下游业务环节继续读取并处理，结果数据可以异步方式导出到关联的OBS对象存储中进行长期低成本存储，从而加速训练场景下加速OBS对象存储中的数据访问

**ModelArts Standard模型训练提供便捷的作业管理能力，提升用户模型训练的开发效率**

- 提供算法资产的管理能力，支持通过算法资产、自定义算法、AI Gallery订阅算法创建训练作业，使训练作业的创建更灵活、易用
- 提供实验管理能力，用户通常需要调整数据集、调整超参等进行多轮作业从而选择最理想的作业，模型训练支持统一管理多个训练作业，方便用户选择最优的模型
- 提供训练作业的事件信息（训练作业生命周期中的关键事件点）、训练日志（训练作业运行过程和异常信息）、资源监控（资源使用率数据）、Cloud Shell（登录训练容器的工具）等能力，方便用户更清楚得了解训练作业运行过程，并在遇到任务异常时更加准确的排查定位问题

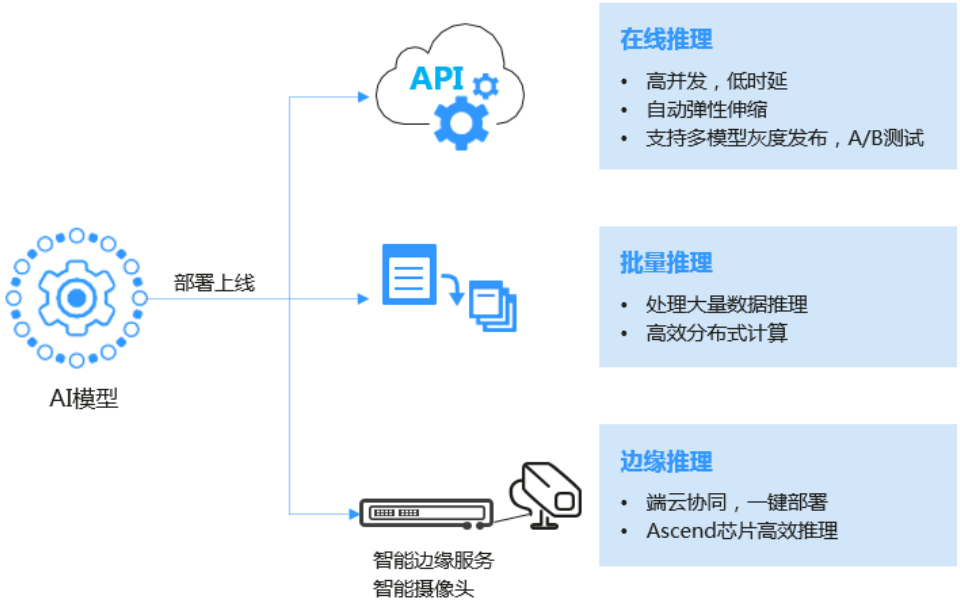
## 5.1.6 Standard 模型部署

ModelArts Standard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 5-5** 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，既能部署为云端的在线推理服务和批量推理任务，也能部署到端，边等各种设备。
- 一键部署，可以直接推送部署到边缘设备中，选择智能边缘节点，推送模型。
- ModelArts基于Snt3高性能AI推理芯片的深度优化，具有PB级别的单日推理数据处理能力，支持发布云上推理的API百万个以上，推理网络时延毫秒。

## 5.1.7 Standard 资源管理

在使用ModelArts进行AI开发时，您可以选择使用如下两种资源池：

**专属资源池**：专属资源池不与其他用户共享，资源更可控。在使用专属资源池之前，您需要先创建一个专属资源池，然后在AI开发过程中选择此专属资源池。其中专属资源池分为弹性集群和弹性裸金属。

- 弹性集群又分为Standard弹性集群与Lite弹性集群。
  - Standard弹性集群提供独享的计算资源，使用ModelArts Standard开发平台的训练作业、部署模型以及开发环境时，通过Standard弹性集群的计算资源进行实例下发。
  - Lite弹性集群面向k8s资源型用户，提供托管式k8s集群，并预装主流AI开发插件以及自研的加速插件，以云原生方式直接向用户提供AI Native的资源、任务等能力，用户可以直接操作资源池中的节点和k8s集群。请参见**弹性集群k8s Cluster**。
- 弹性裸金属：弹性裸金属提供不同型号的xPU裸金属服务器，您可以通过弹性公网IP进行访问，在给定的操作系统镜像上可以自行安装GPU&NPU相关的驱动和其他软件，使用SFS或OBS进行数据存储和读取相关的操作，满足算法工程师进行日常训练的需要。请参见**弹性裸金属DevServer**。

**公共资源池：**公共资源池提供公共的大规模计算集群，根据用户作业参数分配使用，资源按作业隔离。 用户下发训练作业、部署模型、使用开发环境实例等，均可以使用ModelArts提供的公共资源池完成，按照使用量计费，方便快捷。

**专属资源池和公共资源池的能力差异**

- 专属资源池为用户提供独立的计算集群、网络，不同用户间的专属资源池物理隔离，公共资源池仅提供逻辑隔离，专属资源池的隔离性、安全性要高于公共资源池。
- 专属资源池用户资源独享，在资源充足的情况下，作业是不会排队的；而公共资源池使用共享资源，在任何时候都有可能排队。
- 专属资源池支持打通用户的网络，在该专属资源池中运行的作业可以访问打通网络中的存储和资源。例如，在创建训练作业时选择打通了网络的专属资源池，训练作业创建成功后，支持在训练时访问SFS中的数据。
- 专属资源池支持自定义物理节点运行环境相关的能力，例如GPU/Ascend驱动的自助升级，而公共资源池暂不支持。

**专属资源池有什么能力?**

新版专属资源池是一个全面的技术和产品的改进，主要能力提升如下：

- 专属资源池类型归一：不再区分训练、推理专属资源池。如果业务允许，您可以在一个专属资源池中同时跑训练和推理的Workload。同时，也可以通过“设置作业类型”来开启/关闭专属资源池对特定作业类型的支持。
- 自助专属池网络打通：可以在ModelArts管理控制台自行创建和管理专属资源池所属的网络。若需要在专属资源池的任务中访问自己VPC上的资源，可通过“打通VPC”来实现。
- 更加完善的集群信息：全新改版的专属资源池详情页面中，提供了作业、节点、资源监控等更加全面的集群信息，可帮助您及时了解集群现状，更好的规划使用资源。
- 自助管理集群GPU/NPU驱动：每个用户对集群的驱动要求不同，在新版专属资源池列表页中，可自行选择加速卡驱动，并根据业务需要进行立即变更或平滑升级。

## 5.1.8 Standard 支持的 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。

### 统一镜像列表

ModelArts提供了ARM+Ascend规格的统一镜像，包括MindSpore、PyTorch。适用于开发环境，模型训练，服务部署，请参考**统一镜像列表**。**表5-1**、**表5-2所示镜像仅发布在西南-贵阳一区域**。

**表 5-1** MindSpore

| 预置镜像 | 适配芯片 | 适用范围 |
| --- | --- | --- |
| mindspore_2.2.0-cann_7.0.1-py_3.9-euler_2.10.7-aarch64-snt9b | Ascend snt9b | Notebook、训练、推理部署 |

| 预置镜像 | 适配芯片 | 适用范围 |
| --- | --- | --- |
| mindspore_2.1.0-cann_6.3.2-py_3.7-euler_2.10.7-aarch64-snt9b | Ascend snt9b | Notebook、训练、推理部署 |
| mindspore_2.2.10-cann_7.0.0-py_3.9-hce_2.0.2312-aarch64-snt9b | Ascend snt9b | Notebook、训练、推理部署 |

**表 5-2** PyTorch

| 预置镜像 | 适配芯片 | 适用范围 |
| --- | --- | --- |
| pytorch_1.11.0-cann_6.3.2-py_3.7-euler_2.10.7-aarch64-snt9b | Ascend snt9b | Notebook、训练、推理部署 |
| pytorch_2.1.0-cann_7.0.0-py_3.9-hce_2.0.2312-aarch64-snt9b | Ascend snt9b | Notebook、训练、推理部署 |
| pytorch_1.11.0-cann_7.0.0-py_3.9-hce_2.0.2312-aarch64-snt9b | Ascend snt9b | Notebook、训练、推理部署 |

## 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 5-3** 新版 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 |

| 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 |
| --- | --- | --- | --- | --- |
| pytorch1.10-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 |
| 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-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 |
| conda3-cuda10.2-cudnn7-ubuntu18.04 | Clean user customized base image include cuda10.2, conda | CPU | 是 | 是 |
| tensorflow1.15-mindspore1.7.0-cann5.1.0-euler2.8-aarch64 | Ascend+ARM算法开发和训练基础镜像，AI引擎预置TensorFlow和MindSpore | Ascend | 是 | 是 |
| modelbox1.3.0-tensorrt7.1.3-cuda10.2-cudnn8-euler2.9.6 | AI应用开发基础镜像，预置AI应用编排引擎ModelBox、AI引擎TensorRT，仅支持SSH连接 | GPU | 是 | 否 |
| modelbox1.3.0-libtorch1.9.1-cuda10.2-cudnn8-euler2.9.6 | AI应用开发基础镜像，预置AI应用编排引擎ModelBox、AI引擎LibTorch，仅支持SSH连接 | GPU | 是 | 否 |

| 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 |
| --- | --- | --- | --- | --- |
| 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 |
| mlstudio-pyspark2.3.2-ubuntu16.04 | CPU算法开发和训练基础镜像，包含可以图形化机器学习算法开发和调测MLStudio工具，并预置PySpark2.3.2 | CPU | 否 | 是 |
| mindspore_1.10.0-cann_6.0.1-py_3.7-euler_2.8.3 | Ascend+ARM algorithm development and training. MindSpore is preset in the AI engine. | Ascend | 是 | 是 |
| mindspore_1.9.0-cann_6.0.0-py_3.7-euler_2.8.3 | Ascend+ARM algorithm development and training. MindSpore is preset in the AI engine. | Ascend | 是 | 是 |
| mindspore1.7.0-cann5.1.0-py3.7-euler2.8.3 | Ascend+ARM算法开发和训练基础镜像，AI引擎预置MindSpore | Ascend | 是 | 是 |
| tensorflow1.15-cann5.1.0-py3.7-euler2.8.3 | Ascend+ARM算法开发和训练基础镜像，AI引擎预置TensorFlow | Ascend | 是 | 是 |
| mlstudio-pyspark2.4.5-ubuntu18.04 | CPU算法开发和训练基础镜像，包含可以图形化机器学习算法开发和调测MLStudio工具，并预置PySpark2.4.5 | 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 |
| rlstudio1.0.0-ray1.3.0-cuda10.1-ubuntu18.04 | CPU、GPU强化学习算法开发和训练基础镜像，预置AI引擎 | CPU/GPU | 是 | 是 |

| 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 |
| --- | --- | --- | --- | --- |
| mindquantum0.9.0-mindspore2.0.0-cuda11.6-ubuntu20.04 | MindSpore2.0.0 and MindQuantum0.9.0 | CPU | 是 | 是 |
| 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 |
| cylp0.91.4-cbcpy2.10-ortools9.0-cplex20.1.0-ubuntu18.04 | CPU运筹优化求解器开发基础镜像，预置cylp，cbcpy，ortools及cplex | 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 5-4** 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 |
| **Ascend-Powered-Engine** | aarch64 | Euler2.8 | mindspore_1.7.0-cann_5.1.0-py_3.7-euler_2.8.3-aarch64 | cann 5.1.0 |
| | | | tensorflow_1.15-cann_5.1.0-py_3.7-euler_2.8.3-aarch64 | cann 5.1.0 |
| **MPI** | x86_64 | Ubuntu18.04 | mindspore_1.3.0-cuda_10.1-py_3.7-ubuntu_1804-x86_64 | cuda_10.1 |

| 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 |
| --- | --- | --- | --- | --- |
| **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。

## 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 5-5** 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<br>python2.7（待下线）<br>tf1.13-python3.6-gpu<br>tf1.13-python3.6-cpu<br>tf1.13-python3.7-cpu<br>tf1.13-python3.7-gpu<br>tf2.1-python3.7（待下线）<br>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。<br>• 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。<br>• 默认使用的Runtime为python2.7。 |

| 模型使用的引擎类型 | 支持的运行环境（Runtime） | 注意事项 |
| --- | --- | --- |
| Spark_MLlib | python2.7（待下线）<br>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。<br>• 默认使用的Runtime为python2.7。<br>• python2.7、python3.6只能用于运行适用于CPU的模型。 |
| Scikit_Learn | python2.7（待下线）<br>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。<br>• 默认使用的Runtime为python2.7。<br>• python2.7、python3.6只能用于运行适用于CPU的模型。 |
| XGBoost | python2.7（待下线）<br>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。<br>• 默认使用的Runtime为python2.7。<br>• python2.7、python3.6只能用于运行适用于CPU的模型。 |
| PyTorch | python2.7（待下线）<br>python3.6<br>python3.7<br>pytorch1.4-python3.7<br>pytorch1.5-python3.7（待下线）<br>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。<br>• 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。<br>• 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。 |

# 5.2 MaaS 大模型即服务平台功能介绍

对于普通企业来说，大模型开发不仅需要强大的算力，还需要学习训练、部署的相关参数配置和规格选择等专业知识。MaaS作为一个面向客户的大模型服务化平台，提供简单易用的模型开发工具链，支持大模型定制开发，让模型应用与业务系统无缝衔接，显著降低了企业AI落地的成本与难度。

- **业界主流开源大模型覆盖全**

  MaaS集成了业界主流开源大模型，含Llama、Baichuan、Yi、Qwen、AIGC等模型系列，所有的模型均基于昇腾AI云服务进行全面适配和优化，使得精度和性能显著提升。开发者无需从零开始构建模型，只需选择合适的预训练模型进行微调或直接应用，大大减轻模型集成的负担。

- **零代码、免配置、免调优模型开发**

  平台结合与100+客户适配、调优开源大模型的行业实践经验，沉淀了大量适配昇腾，和调优推理参数的最佳实践。通过为客户提供一键式训练、自动超参调优等能力，和高度自动化的参数配置机制，使得模型优化过程不再依赖于手动尝试，显著缩短了从模型开发到部署的周期，确保了模型在各类应用场景下的高性能表现，让客户能够更加聚焦于业务逻辑与创新应用的设计。

- **资源易获取，按需收费，按需扩缩，支撑故障快恢与断点续训**

  企业在具体使用大模型接入企业应用系统的时候，不仅要考虑模型体验情况，还需要考虑模型具体的精度效果，和实际应用成本。

  MaaS提供灵活的模型开发能力，同时基于昇腾云的算力底座能力，提供了若干保障客户商业应用的关键能力。

  保障客户系统应用大模型的成本效率，按需收费，按需扩缩的灵活成本效益资源配置方案，有效避免了资源闲置与浪费，降低了进入AI领域的门槛。

  架构强调高可用性，多数据中心部署确保数据与任务备份，即使遭遇故障，也能无缝切换至备用系统，维持模型训练不中断，保护长期项目免受时间与资源损耗，确保进展与收益。

- **大模型应用开发，帮助开发者快速构建智能Agents**

  在企业中，项目级复杂任务通常需要理解任务并拆解成多个问题再进行决策，然后调用多个子系统去执行。MaaS基于多个优质昇腾云开源大模型，提供优质 Prompt 模板，让大模型准确理解业务意图，分解复杂任务，沉淀出丰富的多个智能Agent，帮助企业快速智能构建和部署大模型应用。

# 5.3 Lite 功能介绍

ModelArts Lite基于软硬件深度结合、垂直优化，构建开放兼容、极致性价比、长稳可靠、超大规模的云原生AI算力集群，提供一站式开通、网络互联、高性能存储、集群管理等能力，满足AI高性能计算等场景需求。目前其已在大模型训练推理、自动驾驶、AIGC、内容审核等领域广泛得到应用。

ModelArts Lite又分以下2种形态：

- ModelArts Lite Server提供不同型号的xPU裸金属服务器，您可以通过弹性公网IP进行访问，在给定的操作系统镜像上可以自行安装加速卡相关的驱动和其他软件，使用SFS或OBS进行数据存储和读取相关的操作，满足算法工程师进行日常训练的需要。
- ModelArts Lite Cluster面向k8s资源型用户，提供托管式k8s集群，并预装主流AI开发插件以及自研的加速插件，以云原生方式直接向用户提供AI Native的资源、任务等能力，用户可以直接操作资源池中的节点和k8s集群。

ModelArts Lite Cluster主要支持以下功能：

- **支持专属备机（高可用冗余节点）**

  在ModelArts Console支持用户购买备机也就是“高可用冗余节点”，通过主备节点倒换的方式顶替主节点运行，缩短主备倒换的时间，提升故障切换的成功率，以确保业务的连续性，以解决当前平台仅提供了后台创建“专属备机”的能力，存在用户与SRE沟通效率低、资源恢复不及时、无法自动故障恢复等问题。

- **同一昇腾算力资源池中，支持存在不同订购周期的服务器**

  同一昇腾算力资源池中，支持资源池中订购不同计费类型/计费周期的资源，解决如下用户的使用场景：

  - 用户在包长周期的资源池中无法扩容短周期的节点。
  - 用户无法在包周期的资源池中扩容按需的节点（包括AutoScaler场景）。

- **支持SFS产品权限划分**

  支持SFS权限划分特性，可以实现训练场景中，挂载的SFS的文件夹能够权限控制，避免出现所有人都可以挂载使用，导致某用户误删所有数据的情况。

- **支持选择资源池的驱动版本**

  通过选择资源池的驱动版本，解决资源池所有节点驱动版本一致的时候，并且没有指定驱动版本，会导致后续加入资源池的节点并不能自动升级到该版本情况，优化了当前需手工处理，增加运维成本问题。

- **支持节点新进入集群，默认启用准入检测，以能够拉起真实的GPU/NPU检测任务**

  支持集群扩容时，扩容的节点默认开启准入检测，该准入检测也可关闭，以提升拉起真实的GPU/NPU检测任务成功率。

# 5.4 AI Gallery 功能介绍

面向开发者提供了AI Gallery大模型开源社区，通过大模型为用户提供服务，普及大模型行业。AI Gallery提供了大量基于昇腾云底座适配的三方开源大模型，同步提供了可以快速体验模型的能力、极致的开发体验，助力开发者快速了解并学习大模型。

- **构建零门槛线上模型体验，零基础开发者开箱即用，初学者三行代码使用所有模型**

  通过AI Gallery的AI应用在线模型体验，可以实现模型服务的即时可用性，开发者无需经历繁琐的环境配置步骤，即可直观感受模型效果，快速尝鲜大模型，真正达到“即时接入，即时体验”的效果。

  当开发者对希望对模型进行开发和训练，AI Gallery为零基础开发者，提供无代码开发工具，快速推理、部署AI应用；为具备基础代码能力的开发者，AI Gallery将复杂的模型、数据及算法策略深度融合，构建了一个高效协同的模型体验环境，让开发者仅需几行代码即可调用任何模型，大幅度降低了模型开发门槛。

- **充足澎湃算力，最佳实践算力推荐方案，提升实践效率和成本**

  AI Gallery深谙开发者在人工智能项目推进过程中面临的实际困难，尤其是高昂的模型训练与部署成本，这往往成为创意落地的阻碍。通过大量开发者实践，针对主流昇腾云开源大模型，沉淀最佳的算力组合方案，为开发者在开发模型的最后一步，提供最佳实践的算力方案、实践指南和文档，节省开发者学习和试错资金成本，提升学习和开发效率。

# 6 AI 开发基础知识

## 6.1 AI 开发基本流程介绍

### 什么是 AI 开发

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 6-1** AI 开发流程


**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**

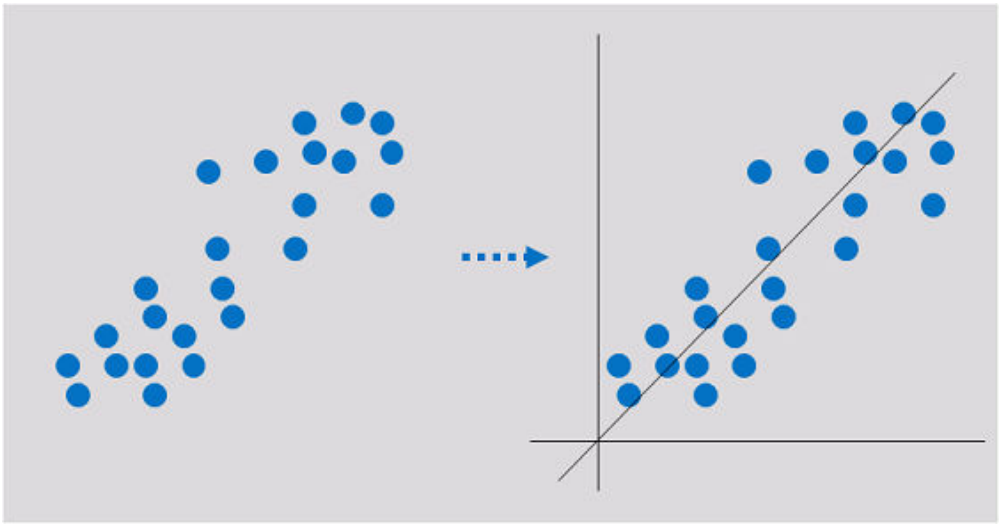
# 6.2 AI 开发基本概念

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。

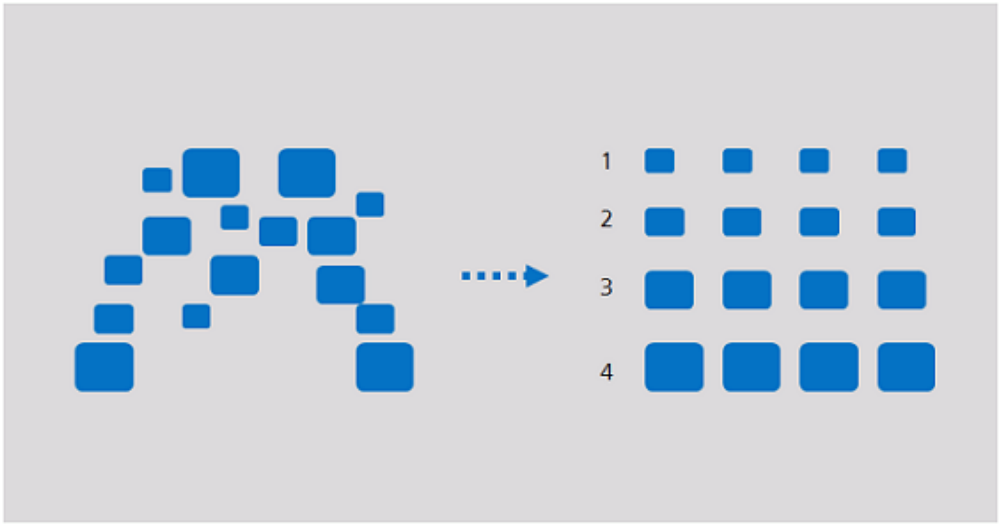
## 回归

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。



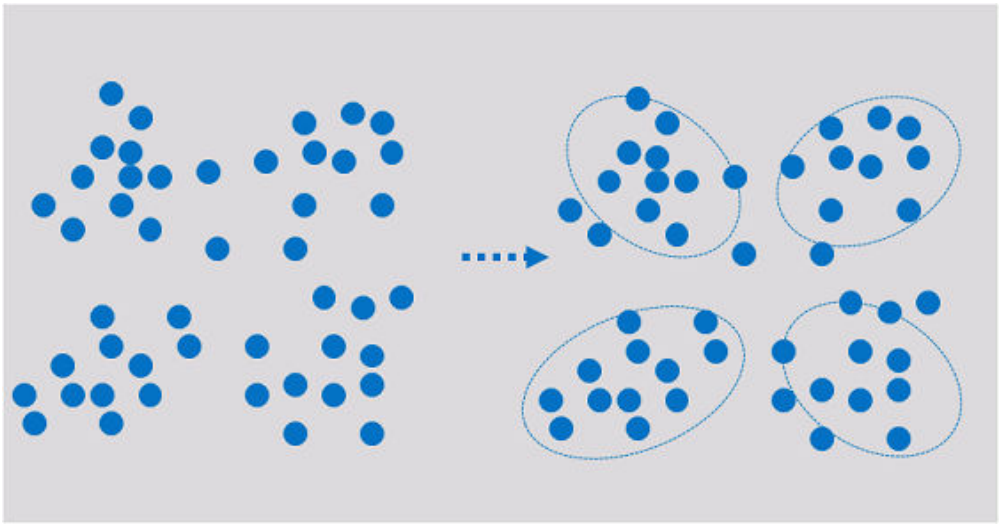
## 分类

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。



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。

# 6.3 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。

## 昇腾芯片

昇腾芯片又叫Ascend芯片，是华为自主研发的高计算力低功耗的AI芯片。

## 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 Standard默认提供公共资源池。ModelArts Standard专属资源池需单独创建，专属使用，不与其他用户共享。

ModelArts Lite Server和ModelArts Lite Cluster使用的都是专属资源池。

## 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。

# 7 安全

## 7.1 责任共担

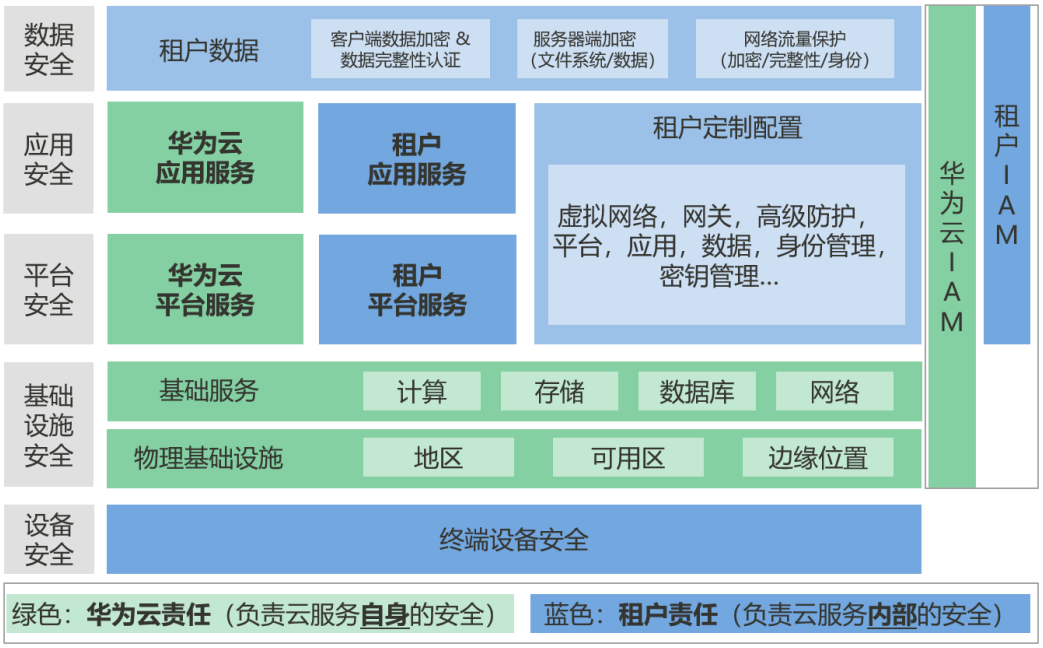
华为云秉承“将公司对网络和业务安全性保障的责任置于公司的商业利益之上”。针对层出不穷的云安全挑战和无孔不入的云安全威胁与攻击，华为云在遵从法律法规业界标准的基础上，以安全生态圈为护城河，依托华为独有的软硬件优势，构建面向不同区域和行业的完善云服务安全保障体系。

安全性是华为云与您的共同责任，如图7-1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。

《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 7-1** 华为云安全责任共担模型




# 7.2 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**》。

# 7.3 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。

## 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。

# 7.4 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。 |

# 7.5 审计与日志

## 审计

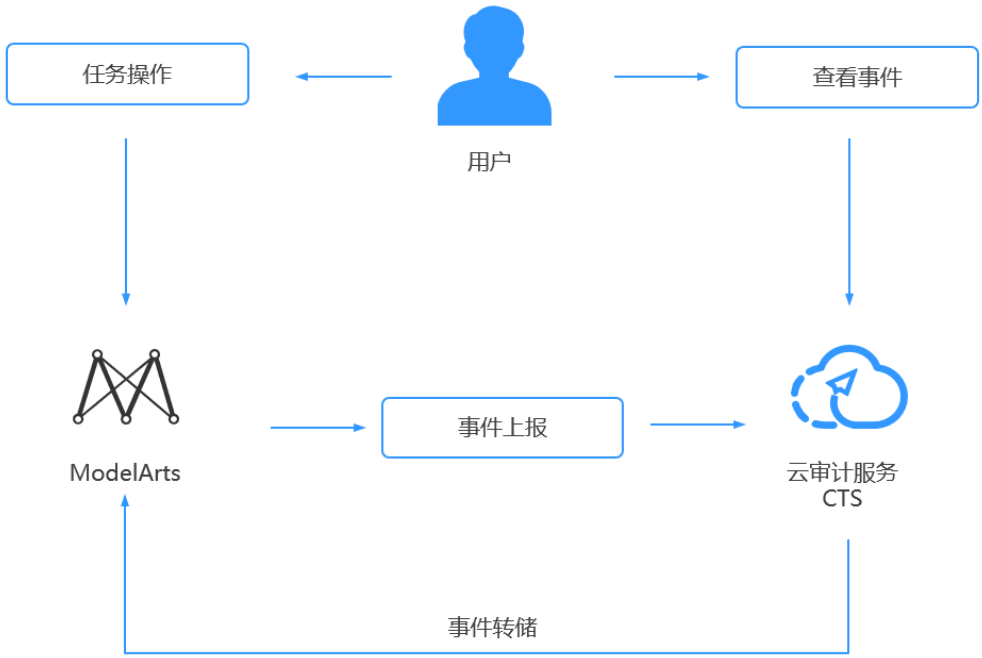
云审计服务（Cloud Trace Service，CTS），是华为云安全解决方案中专业的日志审计服务，提供对各种云资源操作记录的收集、存储和查询功能，可用于支撑安全分析、合规审计、资源跟踪和问题定位等常见应用场景。

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**。

**图 7-2** 云审计服务




## 数据管理支持审计的关键操作列表

**表 7-1** 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 |

| 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 |

| 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 7-2** 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 7-3** 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 |

| 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 7-4** 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 |

## 服务管理支持审计的关键操作列表

**表 7-5** 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 |
| 启停边缘服务节点 | service | startOrStopNodesService |
| 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 7-6** 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 |
| 收藏资产 | ModelArts_Market | star_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 |

### 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。

## 7.6 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。

### 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。
- 所有中国大陆境内下沉POD区服务使用的域名已完成备案。
- 所有中国大陆境内下沉POD区的服务均遵守国家**《互联网信息服务管理办法》**要求。

## 7.7 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**。

## 7.8 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。

## 7.9 更新管理

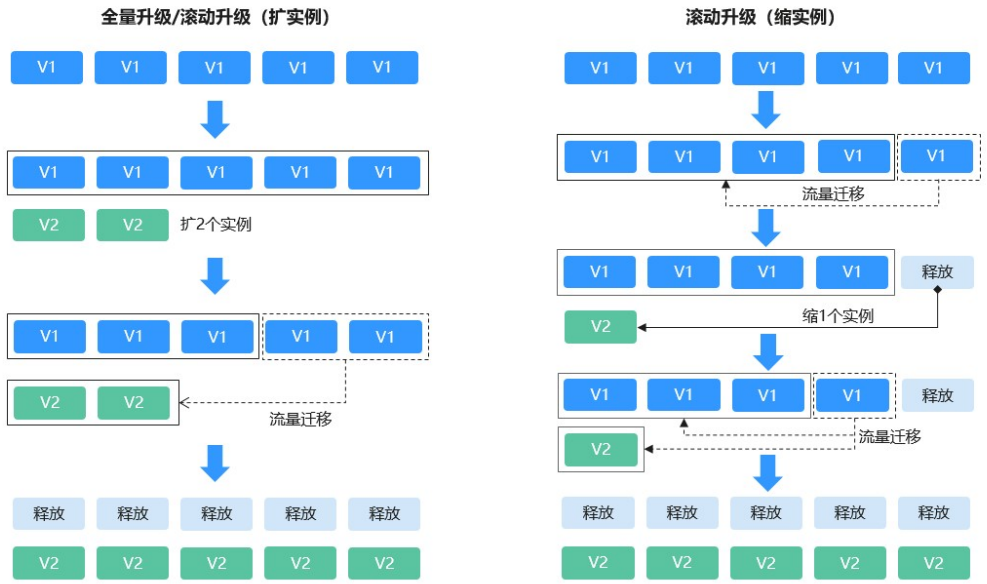
### ModelArts 在线服务更新

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7-3。

- 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 7-3** 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

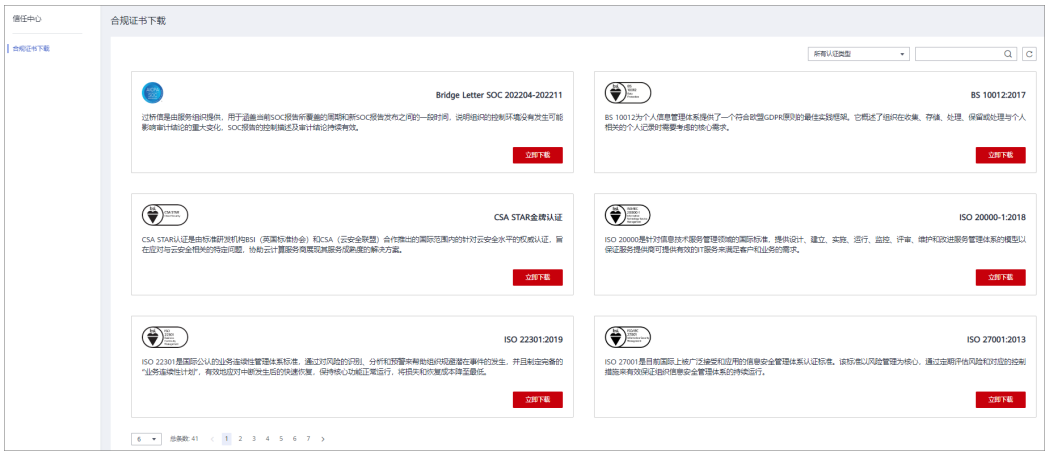
ModelArts包含开发环境、训练管理、推理部署三个功能模块，三个模块采用统一的流程提供基础镜像。这些镜像会不定期更新升级，修复已知漏洞。

# 7.10 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 7-4** 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 7-5** 资源中心


## 销售许可证&软件著作权证书

另外，华为云还提供了以下销售许可证及软件著作权证书，供用户下载和参考。具体请查看**合规资质证书**。

**图 7-6** 销售许可证&软件著作权证书


# 7.11 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

## 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。

# 8 约束与限制

本节介绍ModelArts服务在使用过程中的约束和限制。

## 规格限制

**表 8-1** 规格说明

| 资源类型 | 规格 | 说明 |
| --- | --- | --- |
| 计算资源 | 所有按需计费、包年/包月、套餐包中的计算资源规格，包括CPU、GPU和NPU | 购买的所有类型的计算资源均不支持跨Region使用。 |
| 计算资源 | 套餐包 | 套餐包仅用于公共资源池，不能用于专属资源池。 |

## 配额限制

查看每个配额项目支持的默认配额，请参考**怎样查看我的配额？**，登录控制台查询您的配额详情。

**表 8-2** 配额

| 资源类型 | 默认配额限制 | 是否支持调整 | 说明 |
| --- | --- | --- | --- |
| Standard Notebook | 一个账号最多创建10个Notebook。 | 否 | 更多信息，请参见**创建Notebook实例**。 |
| Standard推理部署在线服务 | 单个账号最多可创建20个在线服务。 | 是<br>**提交工单**申请提升配额 | 更多信息，请参见**部署在线服务**。 |

| 资源类型 | 默认配额限制 | 是否支持调整 | 说明 |
| --- | --- | --- | --- |
| Standard推理部署批量服务 | 单个账号最多可创建1000个批量服务。 | 否 | 更多信息，请参见**部署批量服务**。 |
| Standard推理部署边缘服务 | 单个账号最多可创建1000个边缘服务。 | 否 | 更多信息，请参见**部署边缘服务**。 |
| Standard专属资源池 | 一个账号最多创建50个专属资源池。 | 是<br>**提交工单**申请提升配额 | 更多信息，请参见**创建专属资源池**。 |
| Standard标签 | 1个训练作业、Notebook实例或在线服务任务最多支持20个标签配额。 | 否 | 更多信息，请参见**标签**。 |

## 功能限制

**表 8-3** 功能约束与限制

| 功能 | 使用限制 |
| --- | --- |
| Standard专属资源池 | • 单次创建Standard专属资源池时，节点数建议不大于30，否则可能触发限流导致创建失败。更多信息请参见，**创建专属资源池**。<br>• 只支持对状态为“运行中”的Standard专属资源池进行扩缩容，且不能缩容到0。<br>• Standard专属资源池状态处于“运行中”时，才能修改资源池的作业类型。<br>• Standard专属资源池状态处于“运行中”，且专属池中的节点需要含有GPU/Ascend资源，才能升级专属资源池的驱动。<br>• 对于Standard逻辑资源池，需要开启节点绑定后才能进行驱动升级，请提交工单联系华为工程师开启节点绑定。 |

| 功能 | 使用限制 |
| --- | --- |
| Standard Notebook | • Notebook实例删除后不可恢复，实例删除后，挂载目录下的数据也将一并删除，请谨慎操作。<br>• Notebook实例状态必须在“停止”中，才能变更Notebook实例镜像。<br>• Notebook实例状态只有处于“停止”、“运行中”和“启动失败”时，才能变更Notebook实例规格。<br>• Notebook实例的存储配置采用的是云硬盘EVS。云硬盘EVS存储容量最大支持4096GB，达到4096GB时，不允许再扩容。单次最大可以扩容100GB。<br>• Notebook实例停止后，扩容后的EVS容量仍然有效。EVS计费也是按照扩容后的容量进行计费。云硬盘EVS只要使用就会计费，请在停止Notebook实例后，确认不使用EVS就及时删除数据，释放资源，避免产生费用。<br>• Notebook中保存的镜像大小不超过35G，镜像层数不能超过125层。否则镜像会保存失败。 |
| Standard训练作业 | • 训练日志仅保留30天，超过30天会被清理。如果用户需要永久保存日志，请在创建训练作业时，打开永久保存日志开关设置作业日志路径即可将日志转存至OBS路径。Ascend训练场景下，默认要求填写作业日志在OBS的存放路径，其他资源的训练场景下，永久保存日志开关需要用户手动开启。<br>• 仅专属资源池支持使用Cloud Shell登录训练容器，且训练作业必须处于“运行中”状态。<br>• 在训练管理的“创建算法”页面，来源于AI Gallery中订阅的算法不支持另存为新算法。<br>• 训练作业卡死检测目前仅支持资源类型为GPU的训练作业。<br>• 仅使用新版专属资源池训练时才支持设置训练作业优先级。公共资源池和旧版专属资源池均不支持设置训练作业优先级。<br>• 仅支持PyTorch和MindSpore框架的分布式训练和调测，如果MindSpore要进行多机分布式训练调试，则每台机器上都必须有8张卡。<br>• 使用自定义镜像创建训练作业时，镜像大小推荐15GB以内，最大不要超过资源池的容器引擎空间大小的一半。镜像过大会直接影响训练作业的启动时间。ModelArts公共资源池的容器引擎空间为50G，专属资源池的容器引擎空间的默认为50G，支持在创建专属资源池时自定义容器引擎空间。<br>• 用于训练的自定义镜像的默认用户必须为“uid”为“1000”的用户。 |

| 功能 | 使用限制 |
| --- | --- |
| Standard推理的创建AI应用 | • 创建AI应用时导入OBS文件，最大支持20GB。更多信息，请参见**创建AI应用**。<br>• 创建AI应用时，系统文件超过容器引擎空间大小时，会提示镜像内空间不足。当前，公共资源池容器引擎空间的大小最大支持50G，专属资源池容器引擎空间的默认为50G，专属资源池容器引擎空间可在创建资源池时自定义设置，设置专属资源池容器引擎空间不会造成额外费用增加。更多信息，请参见**导入AI应用对镜像大小的约束限制**。<br>• 自动学习项目中，在完成模型部署后，其生成的模型也将自动上传至AI应用列表中。但是自动学习生成的AI应用无法下载，只能用于部署上线。 |
| Standard推理服务部署 | • 只支持使用专属资源池部署的在线服务使用CloudShell访问推理容器，且在线服务必须处于“运行中”状态。<br>• Standard推理中，将AI应用部署为批量服务时，只支持使用公共资源池，暂不支持使用专属资源池。<br>• 对于同步请求模式的AI应用，如果预测请求时延超过60s，会造成请求失败，甚至会有服务业务中断的风险，预测请求时延超过60s时，建议制作异步请求模式的镜像。 |
| Lite Server | • ModelArts Lite Server使用裸金属服务器时，如果升级/修改操作系统内核或者驱动，很可能导致驱动和内核版本不兼容，从而导致OS无法启动，或者基本功能不可用。如果需要升级/修改，请**联系华为云技术支持**。<br>• ModelArts Lite Server使用ECS服务器时不支持重装操作系统，部分区域使用裸金属服务器时也不支持重装操作系统，若您想重装操作系统，您可通过切换操作系统的方式解决。更多信息，请参见**Server使用前须知**。<br>• ModelArts Lite Server服务器重装或者切换操作系统后，对应的EVS系统盘ID发生变化，和下单时订单中的EVS ID已经不一致， 因此EVS系统盘无法扩容，并显示信息："当前订单已到期，无法进行扩容操作，请续订"。建议通过挂载数据盘EVS或挂载SFS盘等方式进行存储扩容。 |
| Lite Cluster | • 只支持对状态为“运行中”的Lite Cluster资源池进行扩缩容，且不能缩容到0。<br>• 对于新建的Lite Cluster资源池，支持在新建时资源池指定容器引擎空间大小。<br>• 对于存量的Lite Cluster资源池，可设置容器引擎空间大小应用于新增的节点，存量节点不支持修改容器引擎空间大小，且会导致资源池内该规格下节点的dockerBaseSize不一致，可能会使得部分任务在不同节点的运行情况不一致。<br>• Lite Cluster资源池状态处于“运行中”，且资源池中的节点需要含有GPU/Ascend资源时，才可以升级Lite Cluster资源池的驱动。<br>• 对于的ite Cluster逻辑资源池，需要开启节点绑定后才能进行驱动升级，请提交工单联系华为工程师开启节点绑定。 |

| 功能 | 使用限制 |
| --- | --- |
| ModelArts与OBS交互 | • ModelArts不支持从加密的OBS桶中读取数据，创建OBS桶时，请勿开启桶加密。<br>• ModelArts不支持跨区域访问OBS桶，请确保使用的OBS与ModelArts在同一区域。 |

# 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**配置访问授权**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

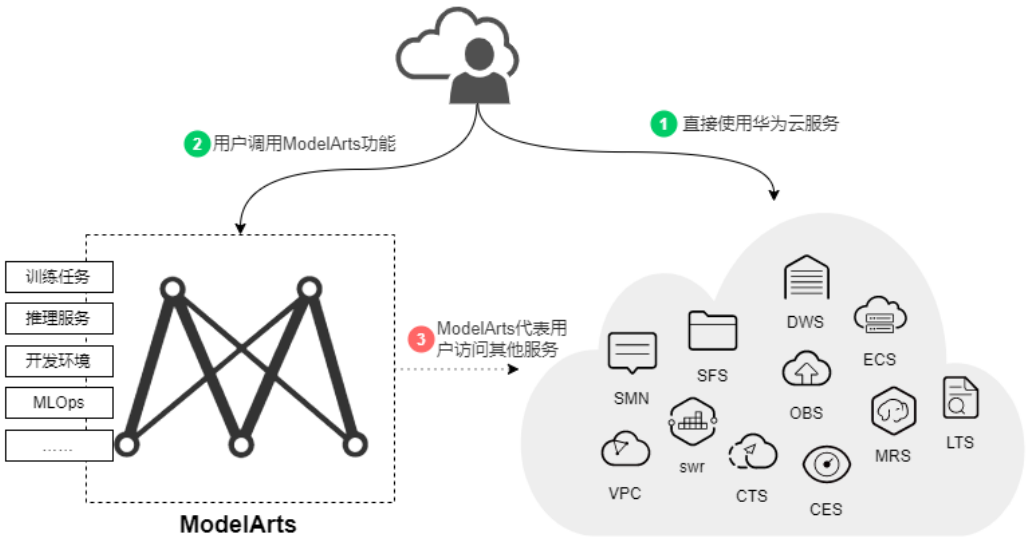
- 您是企业用户，且
  - 存在多个部门，且需要限定不同部门的用户只能访问其专属资源、功能
  - 存在多种角色（如管理员、算法开发者、应用运维），希望限制不同角色只能使用特定功能
  - 逻辑上存在多套“环境”且相互隔离（如开发环境、预生产环境、生产环境），并限定不同用户在不同环境上的操作权限
  - 其他任何需要对特定子用户（组）做出特定权限限制的情况
- 您是个人用户，但已经在IAM创建多个子用户，且期望限定不同子用户所能使用的ModelArts功能、资源不同
- 希望了解ModelArts的权限控制能力细节，期望理解其概念和实操方法

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。

## 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。

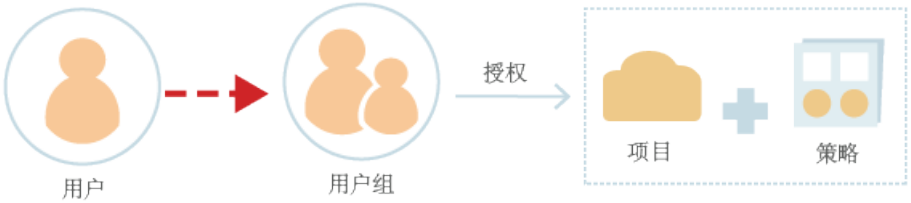
## ModelArts 权限管理

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。

**⚠ 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（如华北-北京4）对应的项目（cn-north-4），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**。

**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。

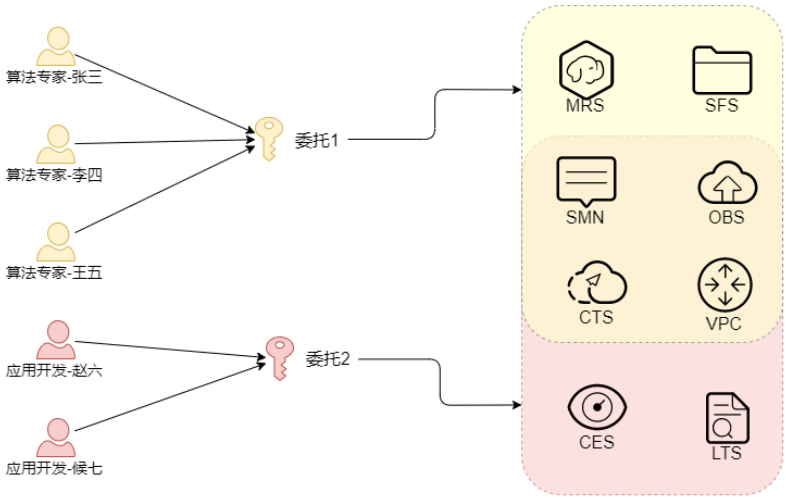
**说明**

- 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。

**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

“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。

# 10 计费说明

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。

# 11 配额与限制

本节介绍ModelArts涉及的相关云服务的配额限制，帮助用户查看和管理自己的配额。

## 什么是配额

配额是在某一区域下最多可同时拥有的某种资源的数量。

华为云为防止资源滥用，对云服务每个区域的用户资源数量和容量做了配额限制。

如果当前资源配额限制无法满足使用需要，您可以申请扩大配额。

## 怎样查看配额

如需查看每个配额项目支持的默认配额，请参考**怎样查看我的配额？**，登录控制台查询您的配额详情。

## 申请扩大配额

如需扩大资源配额，请在华为云管理控制台**申请扩大配额**。

## 配额项说明

使用ModelArts Lite Cluster或Lite Server时，所需的ECS实例数、内存大小、CPU核数和EVS硬盘大小等等资源会超出华为云默认提供的资源配额，因此需要申请扩大配额。具体配额项如下。

**表 11-1** ModelArts Lite 涉及的资源配额

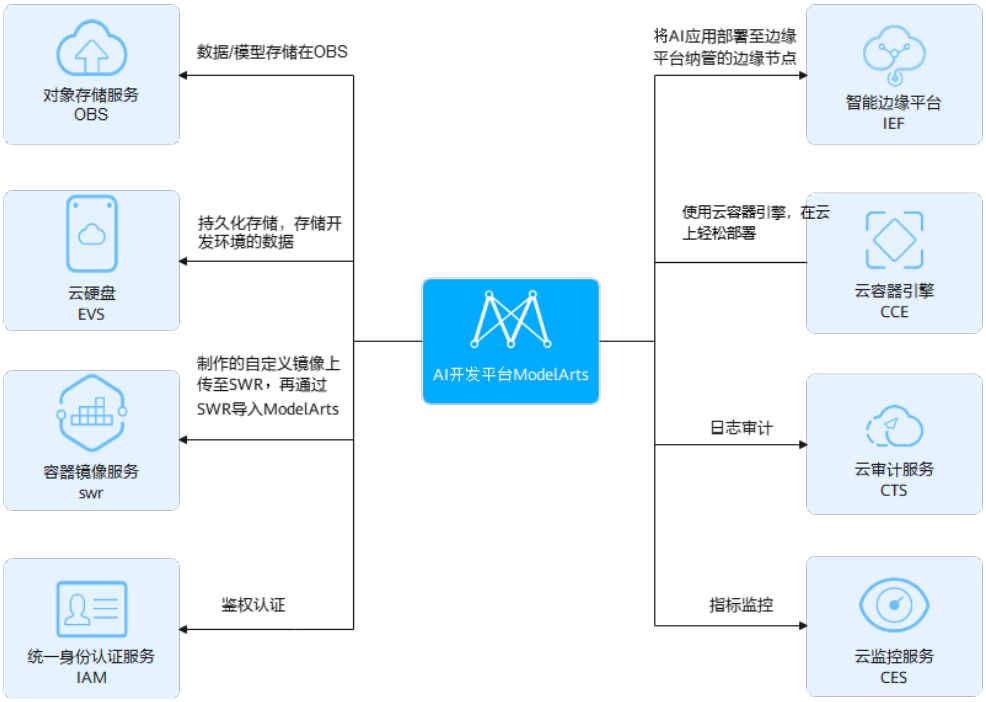
| 服务 | 资源类型 |
| --- | --- |
| ECS资源类型 | ECS实例数 |
| | CPU核心数 |
| | RAM容量（MB） |
| 弹性公网IP EIP资源 | 伸缩带宽策略 |
| 云硬盘EVS SFS资源 | 磁盘数 |

| 服务 | 资源类型 |
| --- | --- |
| | 磁盘容量（GB） |
| | 快照数 |
| 弹性文件服务SFS资源 | 容量配额 |

# 12 与其他云服务的关系

**图 12-1** ModelArts 与其他服务的关系示意图




## 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**《统一身份认证服务产品文档》**。

## 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**《对象存储服务产品文档》**。

**表 12-1** 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。<br>• 数据集的标注信息存储在OBS中。<br>• 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。<br>• 训练作业的运行脚本存储在OBS中。<br>• 训练作业输出的模型存储在指定的OBS中。<br>• 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。 |
| | 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》**。

## 与智能边缘平台的关系

ModelArts可将模型部署至智能边缘平台（Intelligent EdgeFabric，简称IEF）纳管的边缘节点。IEF的更多信息请参见**《智能边缘平台用户指南》**。

## 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》**。

## 与云审计的关系

ModelArts使用云审计服务（Cloud Trace Service，简称CTS）记录ModelArts相关的操作事件，便于日后的查询、审计和回溯。CTS的更多信息请参见**《云审计服务指南》**。