Fabric

用户指南

文档版本 01

发布日期 2024-12-31

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 Fabric 使用流程

Fabric平台提供了一个serverless化的从数据到模型部署的AI全流程开发体验，针对每个环节，其使用是相对独立自由的。本章节梳理了Fabric使用流程详解，您可以选择其中一种方式完成AI开发。

表 1-1 使用流程说明

流程	说明	详细指导
创建工作空间	创建一个工作空间，后续所有的能力都承载在工作空间中。	**创建工作空间**
创建端点	创建一个端点，根据业务类型不同，创建不同类型的端点。	**创建推理端点**
注册模型	用户可以将存储在OBS的微调模型文件，在模型管理的界面注册为自己的微调模型。	**创建模型**
部署服务	Fabric支持部署用户基于基模型微调的微调模型	**创建推理服务**
访问服务	微调模型部署完成后，用户可以使用Fabric提供的推理接口直接进行推理。	**使用推理服务进行推理**

2 准备工作

2.1 创建 IAM 用户并授权使用 Fabric

在使用Fabric相关功能之前，您需要提前做好准备工作，包括开通账号、开通与配置账号子账号权限、创建工作空间等。本章节详细介绍创建IAM用户并授权使用Fabric操作步骤。

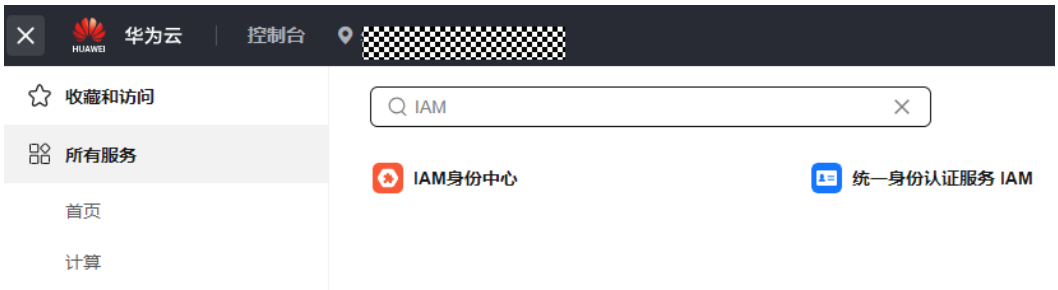
前提条件

已有可正常使用的华为云账号。

操作步骤

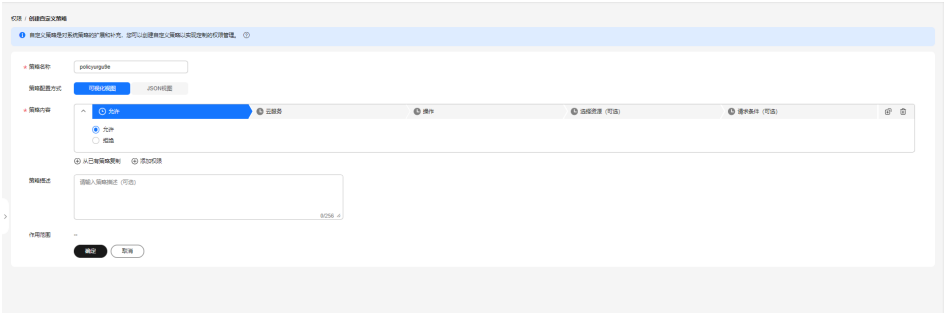
步骤1 登录华为云控制台，在页面左上角单击☰，在服务列表中选择“统一身份认证服务IAM”。

图 2-1 选择服务列表



步骤2 单击“权限管理>权限”，单击右上角创建“自定义策略”，输入必要参数后单击“确定”。详细创建流程参考**创建自定义策略**。

图 2-2 创建自定义策略



管理员可以通过为不同的用户组设置不同的策略，对不同的用户设置不同的用户组来实现用户权限控制，管理员可以根据自己的需求配置权限，下面给出一些建议的权限组合供管理员参考。

表 2-1 权限介绍

业务角色	策略	功能

系统管理员	{
	"Version": "1.1",
	"Statement": [
	{
	"Effect": "Allow",
	"Action": [
	"DataArtsFabric:*:*",
	"obs:bucket;*",
	"obs:object:*"
	]
	}
	]
	}
		拥有Fabric所有权限，可以进行所有Fabric操作。

业务角色	策略	功能

资源管理员	{
	"Version": "1.1",
	"Statement": [
	{
	"Effect": "Allow",
	"Action": [
	"DataArtsFabric:workspace:*",
	"DataArtsFabric:endpoint:*",
	"lakeformation:instance:*"
	]
	}
	]
	}
		用户Fabric资源的管理权限，可以进行工作空间，端点的创建删除等操作。

推理业务操作员	{
	"Version": "1.1",
	"Statement": [
	{
	"Effect": "Allow",
	"Action": [
	"DataArtsFabric:workspace:list",
	"DataArtsFabric:endpoint:list",
	"DataArtsFabric:endpoint:show",
	"DataArtsFabric:model:*",
	"DataArtsFabric:service:*",
	"obs:object:*",
	"obs:bucket:ListBucket"
	]
	}
	]
	}
		可以执行推理相关的业务，包括注册模型，创建推理服务，进行推理。

业务角色	策略	功能

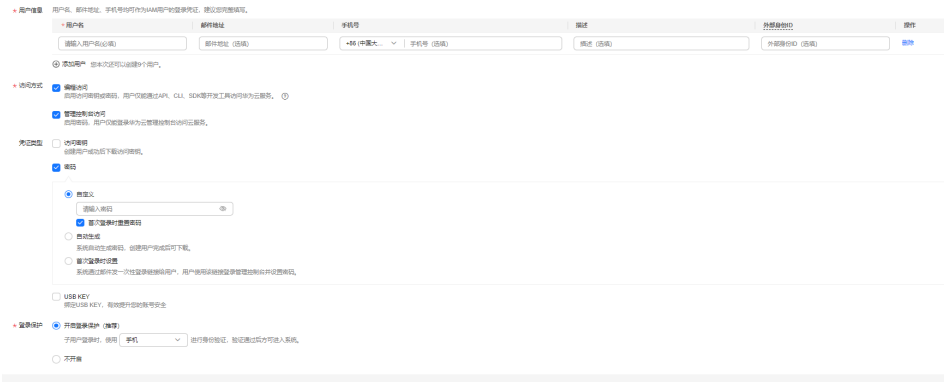
作业业务操作员	{
	"Version": "1.1",
	"Statement": [
	{
	"Effect": "Allow",
	"Action": [
	"DataArtsFabric:workspace:list",
	"DataArtsFabric:endpoint:list",
	"DataArtsFabric:endpoint:show",
	"DataArtsFabric:job:*",
	"obs:object:*",
	"obs:bucket:ListBucket"
	]
	}
	]
	}
		可以执行作业相关的业务，包括创建作业执行作业。

步骤3 单击“用户组”，单击右上角“创建用户组”，输入用户组名称后单击“确定”。

步骤4 在用户组列表中，选择创建的用户组，单击“授权”，选择必要的策略，单击“确定”。详细流程参考**创建用户组并授权**。

步骤5 单击“用户”，右上角“创建用户”，输入必要参数后，单击“确定”。创建的详细流程参考**创建IAM用户**。

图 2-3 示例图片



----结束

2.2 配置 Fabric 服务委托权限

当前云服务提供多种功能，不同的功能需要不同的委托权限。详细见**表2-2**。在创建好的工作空间中，可以进行服务授权。

前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。

操作步骤

步骤1 登录**Fabric工作空间管理台**，单击“服务授权”。

图 2-4 服务授权界面

步骤2 在服务授权页面配置授权委托。用户可以根据实际需要参照委托策略进行配置委托权限。

图 2-5 服务授权配置

表 2-2 委托策略

委托策略名称	权限项	是否必须	功能
FABRIC_COMMON_POLICY	iam:tokens:assume iam:groups:listGroups iam:users:listUsers iam:roles:listRoles iam:groups:listGroupsForUser iam:agencies:listAgencies iam:roles:getRole iam:permissions:listRolesForAgency obs:bucket:ListAllMyBuckets obs:bucket:GetLifecycleConfigurationd obs:bucket:GetBucketLocation obs:bucket:ListBucket obs:object:GetObjectVersion obs:object:GetObject DataArtsFabric:workspace:list DataArtsFabric:endpoint:list DataArtsFabric:endpoint:show DataArtsFabric:endpoint:listRoute	是	• IAM相关权限：仅委托部分只读权限，保证服务能够比较当前用户的委托和服务需要的委托，用于提示用户进行委托更新。 • OBS相关权限：服务所有业务，包括作业，推理，都需要OBS文件的读取权限，保证后续能够从用户的OBS桶拉取到作业文件进行执行，模型文件进行部署。针对OBS的权限，用户可以在IAM的委托界面手动修改fabric_admin_trust委托中OBS相关的部分，限制服务可以访问的OBS资源，具体如何设置参考**IAM权限**，OBS自定义策略样例。
FABRIC_AOM_POLICY	aom:alarm:put	否	Fabric服务使用运维管理服务所需的权限。如果需要指标监控和告警能力，需要开启。

委托策略名称	权限项	是否必须	功能
FABRIC_LAKEFORMATION_POLICY	lakeformation:accessTenant:grant lakeformation:access:delete lakeformation:access:create lakeformation:access:describe lakeformation:access:describe lakeformation:agreement:grant lakeformation:agreement:describe lakeformation:agreement:cancel lakeformation:agency:create lakeformation:agency:drop lakeformation:agency:describe	否	Fabric服务使用LakeFormation服务所需的权限。如果需要对接LakeFormation，则需要开启。
FABRIC_SMN_POLICY	smn:topic:publish	否	Fabric服务使用消息通知服务所需的权限。如果需要消息通知能力，则需要开启。

说明

除必选的委托，其他委托权限都支持取消。

----结束

2.3 创建工作空间

工作空间是Fabric的基本单元，后续所有的操作都在工作空间中进行。因此在账号授权配置完成，需要首先创建工作空间。

用户可根据实际需要创建一个或多个工作空间，各个工作空间是单独隔离的。

前提条件

已有可正常使用的华为云账号。

操作步骤

步骤1 登录华为云控制台后，在页面左上角单击，在服务列表中选择“Fabric”。

图 2-6 Fabric 服务

步骤2 单击“创建工作空间”，参照**创作工作空间填写页面参数说明**输入必要参数后，单击“直接创建”。创建工作空间完成后会返回工作空间管理台界面。

图 2-7 创建工作空间

表 2-3 创作工作空间填写页面参数说明

参数	说明
工作空间名称	请输入工作空间名称，同一账号下集群不可重名。
工作空间描述	可选，请输入工作空间描述。
Metastore	可选，需要绑定的lakeformation实例。

参数	说明
企业项目	选择某个企业项目后，集群和集群安全组将会创建在该企业项目下。您可以通过企业项目服务（EPS）管理集群及其他资源（节点、ELB、以及节点的安全组等）。
标签	可选，通过为资源添加标签，可以对资源进行自定义标记，实现资源的分类。 您可以在TMS中创建“预定义标签”，预定义标签对所有支持标签功能的服务资源可见，通过使用预定义标签可以提升标签创建和迁移效率。具体请参见**创建预定义标签**。 • 标签键只能包含中文、英文字母、数字、空格和特殊字符(-_.:=+@)，且首尾不能包含空格，不能以_sys_开头，长度不超过128个字符。资源标签键不可以为空。 • 标签值只能包含中文，英文字母、数字、空格和特殊字符(-_.:/=+@)，长度不超过255个字符。资源标签值可以为空。

步骤3 单击已创建的工作空间中的“进入工作空间”，弹出用户协议时，用户可查看声明协议，确认后单击“同意授权”，后续即可正常进入创建好的工作空间。

----结束

3 Ray 场景

Ray是一款高性能分布式执行框架，它使用了和传统分布式计算系统不一样的架构，提供了分布式计算的抽象方式。

Ray集群采用全托管独享模式，用户无需关心后台的资源管理，提供基于Ray的分布式作业执行能力，完全兼容开源版本，用户无需对脚本进行复杂的适配就可以使用，并且开放原生的dashboard能力，保证用户的使用习惯。相比开源ray，Fabric服务做了一些列的安全加固，保证用户数据安全，如grpc通道加密、dashboard认证访问等。

3.1 购买 Ray 资源

前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。

操作步骤

由于Ray是全托管模式，在使用Ray前，先需要购买Ray资源。操作步骤如下：

步骤1 登录**Fabric工作空间管理台**。

步骤2 选择已创建的工作空间，单击“进入工作空间”，选择“资源与资产 > Ray集群”。

步骤3 右上角单击“购买Ray资源”，进入购买页面。用户根据需求选择合适的DPU或者APU规格、数量、购买时长等内容。详细参数说明请参见**购买Ray页面参数说明**。

图 3-1 购买 Ray 资源

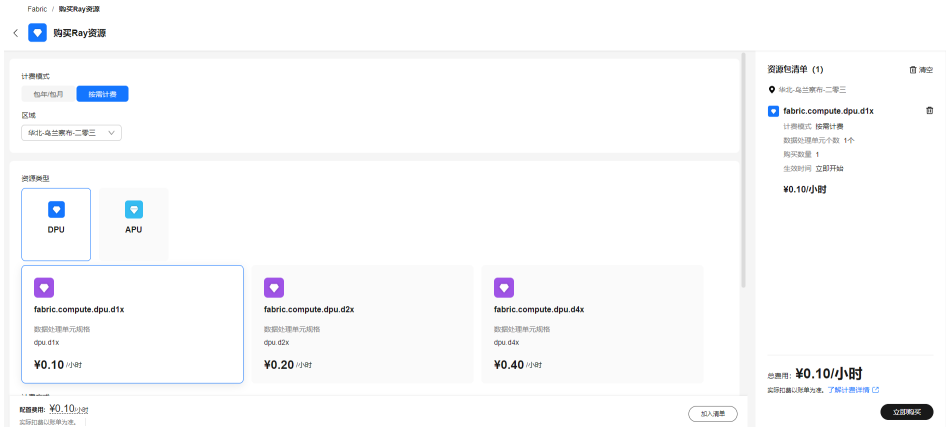


表 3-1 购买 Ray 页面参数说明

参数	参数说明
计费模式	可选择：包年/包月或者按需计费。
资源类型	分为DPU和APU两种，可根据实际需要勾选。 DPU：面向数据分析场景，基于CPU的计算单元。 APU：面向AI场景，基于NPU的计算单元。
规格大小	DPU资源规格fabric.ray.dpu.d1x、fabric.ray.dpu.d2x、fabric.ray.dpu.d4x等规格之间的区别主要体现在cpu数量及内存大小。 APU资源规格之间的区别主要体现在昇腾卡数量和机型差异，可根据需求选择不同规格的资源创建。 具体资源大小参考**产品规格**。
购买时长	可根据实际需要选择购买时长。

说明

购买Ray资源有最低资源要求，最低需要4个fabric.ray.dpu.d1x的资源总量，Fabric服务中fabric.ray.dpu.dnx = n * fabric.ray.dpu.d1x。

步骤4 选择完成后，单击“加入清单”。在资源包清单中确认无误后，单击“立即购买”。确认配置详情完成后，单击“去支付”跳转付款页面，付款完成即可完成购买。可在Ray资源标签中查看购买的资源状态。

说明

- 购买完成后“资源与资产-Ray集群-Ray资源”页面中处于“准备中”，如果购买成功则变为“运行中”否则会变为“失败”状态。
- 如果是首次购买资源，则需要等待15-20分钟。如果购买清单中包含APU资源，则需要40-50分钟；如果是新增其他规格资源，则需要等待5分钟左右；如果新增的是APU资源，则需要等待20分钟左右。可手动刷新资源状态查看资源是否已准备好。
- 同一种规格只能购买一次，购买成功后可通过扩缩容调整数量。如果需要同时多次购买同一种规格，可新创建一个workspace再购买。

----结束

3.2 管理镜像包

前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。
- 请确保您已开通镜像包操作白名单功能。如果有试用需求，请提工单申请权限。

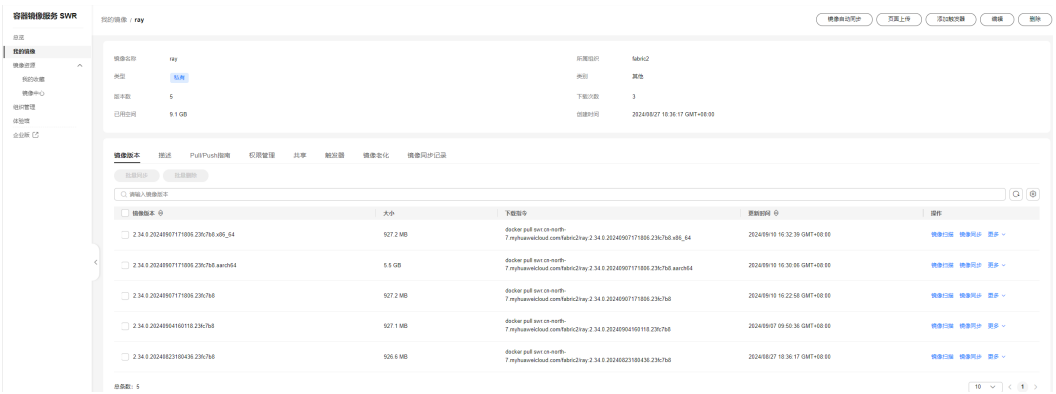
上传镜像压缩包到 SWR

登录容器镜像服务SWR控制台，在“页面上传”对话框参照上传提示信息，上传镜像压缩包到SWR。如果文件大小超过2GB，请使用客户端上传。具体操作，请参见**页面上传镜像**。

图 3-2 上传镜像

上传完成后，镜像压缩包如下图所示。

图 3-3 镜像上传成功



创建镜像包

1. 登录Fabric工作空间管理台，选择已创建的工作空间，单击“进入工作空间”。
2. 在左侧导航栏，选择“运维管理 > 镜像包管理”，单击右上角的“创建镜像包”。
3. 根据提示输入名称和版本名称，为指定版本选择存储在OBS的镜像包路径，配置完成后，单击“确认”创建镜像包。

界面参数说明请参见**创建镜像包参数说明**。

图 3-4 创建镜像包

表 3-2 创建镜像包参数说明

参数	参数说明
名称	镜像包名称。
描述	根据需求填写该镜像包的描述信息。
版本名称	镜像包可有多个版本，根据当前创建信息填入一个版本名称。
版本描述	当前创建版本的描述信息。
路径	当前创建版本所在的OBS路径。

说明

如果新建RAY_CLUSTER类型的Cap，Cap的名称、版本名称需要和包名严格一致。

例如：OBS路径下包名obs://xxx/ray-cap/files/ray-cluster-2.34.0.tgz，包名为ray-cluster-2.34.0。则名称必须为ray-cluster，版本必须为2.34.0，否则页面会校验不通过。

新增镜像包版本

1. 在“镜像包管理”页面的“操作”列，单击目标镜像包对应的“查看版本列表”。
2. 在“当前镜像包版本列表”页面，单击“新增版本”。
3. 在新增镜像包版本页面，配置相关信息，然后单击“确认”。

   界面参数说明请参见**创建镜像包版本参数说明**。

图 3-5 创建镜像包版本

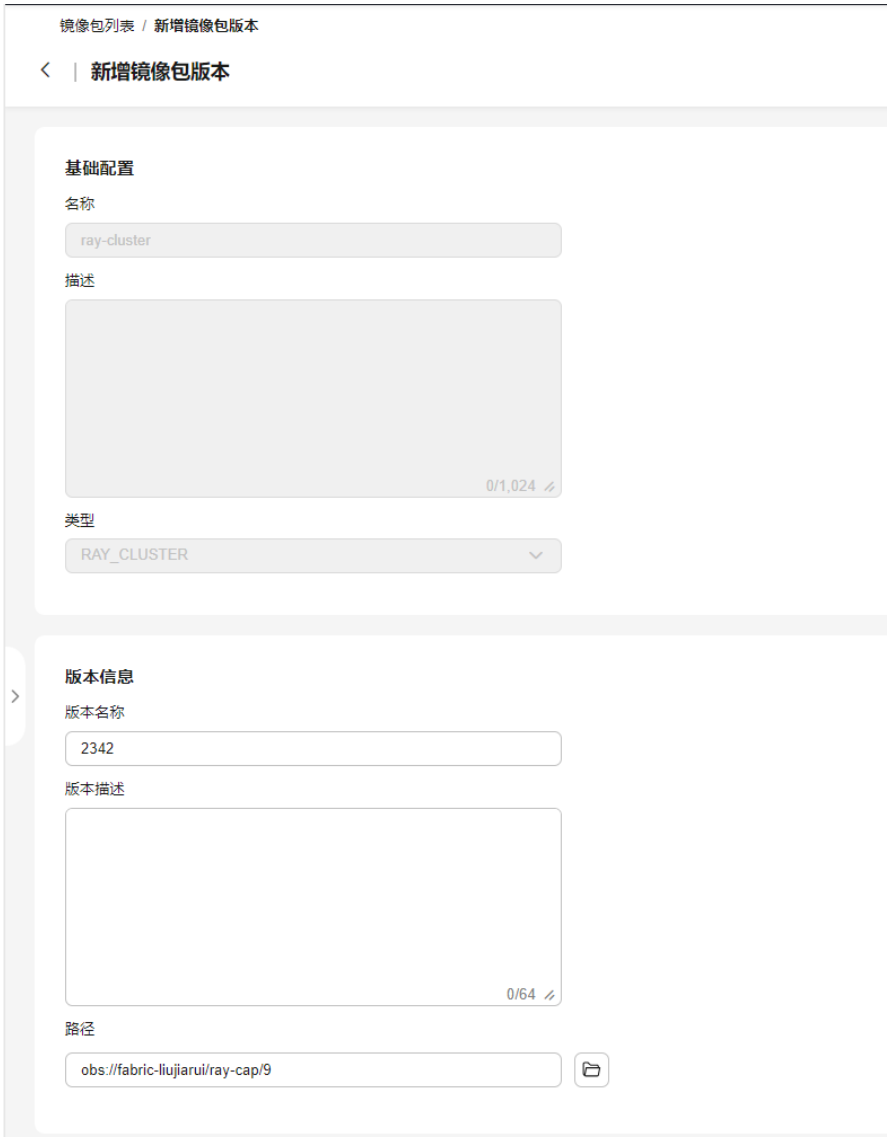


表 3-3 创建镜像包版本参数说明

参数	参数说明
版本名称	镜像包支持有多个版本，请根据当前创建信息填入一个版本名称。镜像包版本需要和选择的OBS文件的包版本号一致。
路径	当前创建版本所在的OBS路径。请选择到包含metadata.yaml文件的父级目录。

删除镜像包版本

删除镜像包版本后，相关数据将被全部清除，请您谨慎操作。

1. 在“镜像包管理”页面的“操作”列，单击目标镜像包对应的“查看版本列表”。
2. 在“当前镜像包版本列表”页面的“操作”列，单击目标版本对应的“删除”。
3. 在“删除当前镜像包版本”对话框，输入“DELETE”或者单击“一键输入”，然后单击“确认”。

图 3-6 删除镜像包版本

删除镜像包

删除镜像包后无法恢复，相关数据将被全部清除，请您谨慎操作。

1. 在“镜像包管理”页面的“操作”列，单击目标镜像包对应的“删除”。
2. 在“删除当前镜像包”对话框，输入“DELETE”或者单击“一键输入”，然后单击“确认”。

图 3-7 删除当前镜像包

3.3 创建 Ray 集群

前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。
- 已购买相应的Ray资源。

操作步骤

步骤1 登录**Fabric工作空间管理台**。

步骤2 选择已创建的工作空间，单击“进入工作空间”，选择“资源与资产 > Ray集群”。单击右上角的创建Ray集群。

步骤3 在创建Ray集群界面，参照**创建Ray集群参数说明**根据需求选择合适的head以及worker规格以及数量，参数填写完成后，单击“创建”即可创建Ray集群。

图 3-8 创建 Ray 集群

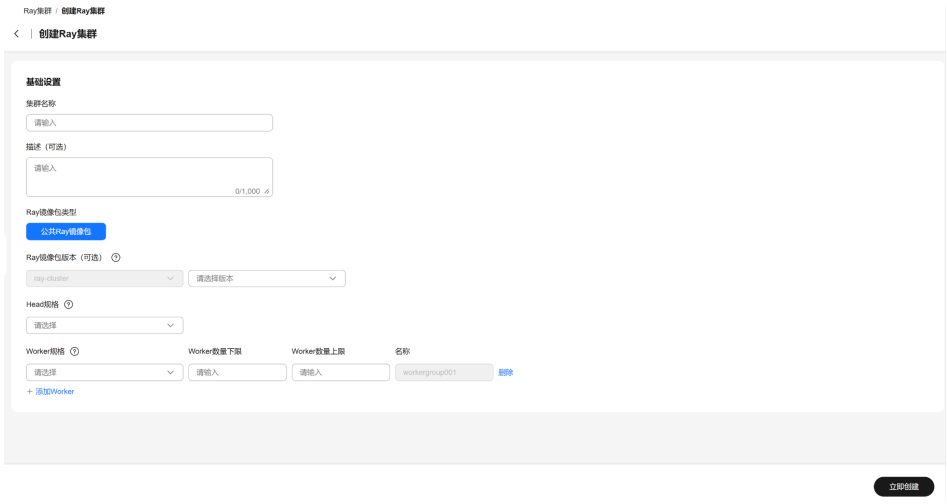


表 3-4 创建 Ray 集群参数说明

参数	参数说明
集群名称	创建Ray集群的名称。
Ray类型	选择公共Ray镜像包。
Ray镜像包版本	可根据需求选择不同的Ray版本，版本号与Ray社区的版本一致。
Head规格	创建Ray集群的head节点规格，可根据业务需求选择。 规格选择列表中可以看到所有的规格，选择的规格可根据创建的Ray资源向下兼容，比如创建了一个fabric.ray.dpu.d4x的资源，那么在选择head规格的时候可以选择fabric.ray.dpu.d1x、fabric.ray.dpu.d2x、fabric.ray.dpu.d4x，即一个大的资源规格可以被拆分为多个小的资源规格。

参数	参数说明
Worker规格	创建Ray集群的worker group规格，可创建多个worker group。 从资源规格列表中选择一个规格部署Worker节点，同时配置worker节点的数量上/下限，worker节点下限至少需要填1，上限请根据业务压力填写。Ray集群初始化创建下限数量的worker规格，根据负载压力动态弹性扩缩到上限数量。也可添加多种不同规格的worker节点。worker节点的规格选择也遵循已有资源向下兼容拆分的规则。例如，当前购买的Ray资源为fabric.ray.dpu.d4x，其中head节点规格选择了fabric.ray.dpu.d1x，那么worker节点也可以选择fabric.ray.dpu.d1x，同时数量上限设置为3。

----结束

说明

您可以手动刷新查看Ray集群创建状态，创建过程约需要3-5分钟。

如果创建Ray集群失败，再次创建之前需要先删除创建失败的Ray集群，避免失败的集群继续占用资源。

3.4 查看 Ray 集群概览

前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。
- 已有至少一个Ray集群。

操作步骤

步骤1 登录**Fabric工作空间管理台**。

步骤2 选择已创建的工作空间，单击“进入工作空间”，选择“资源与资产 > Ray集群”。单击任意一个Ray集群可查看详情页面。

图 3-9 Ray 详情

表 3-5 参数说明

参数	说明
集群名称	自定义的Ray集群名称。
集群ID	集群唯一标识ID。
状态	当前集群状态。
描述	对集群的自定义描述信息。
创建人	集群的创建者。
创建时间	创建集群的时间。
集群版本	集群当前部署的Ray集群版本信息。
Ray资源	集群部署所占资源规格及数量。
访问链接	Ray dashboard的访问地址。

----结束

3.5 创建 Ray Job

前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。

- 已有至少一个可用的Ray集群。
- 已根据业务需求开发Job相关代码，并将代码上传至OBS（创建OBS桶及上传文件请参考**OBS创建桶**）。

操作步骤

步骤1 登录**Fabric工作空间管理台**。

步骤2 选择已创建的工作空间，单击“进入工作空间”，选择“开发与生产 > Job定义”。

步骤3 单击右上角“创建作业”。根据**创建Job参数说明**填写必要的信息，作业类型选择Ray，其他内容根据情况填写后创建作业。其中，Ray主文件为您开发的Job主入口文件。

图 3-10 创建 Job

表 3-6 创建 Job 参数说明

参数	参数说明
Job名称	创建Job定义的名称。
Job类型	默认为Ray。

参数	参数说明
代码目录	选择您存储在OBS的Job定义目录。
Ray主文件	选择代码目录中的Job运行代码的主入口Python文件。请确保您选择的主文件为整个Job运行的主入口文件，否则运行Job可能与您的预期不符。 **说明** 请不要在脚本中输入敏感信息，也不要通过脚本打印敏感信息。
Ray作业参数	Ray主文件执行时所需的参数，示例如下： `["--class","org.ray.examples.rayTest","10","model_2","20"]` **说明** 请不要在脚本中输入敏感信息，也不要通过脚本打印敏感信息。
依赖库	Ray作业运行前所依赖的软件及版本，Ray作业运行前会先通过pip安装此依赖。格式与requirements.txt一致。示例如下： `numpy==1.24.3`
Ray集群	指定在目标Ray集群上执行。
版本名称	作业版本。
版本描述	版本描述，字数为1000字以内。

----结束

3.6 运行 Ray Job

前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。
- 已有至少一个可用的Ray集群。
- 已有至少一个可用的Job作业。

操作步骤

步骤1 登录**Fabric工作空间管理台**。

步骤2 选择已创建的工作空间，单击“进入工作空间”，选择“开发与生产 > Job定义”。

步骤3 在作业列表中选择一个作业，指定其运行的Endpoint后，单击操作列“启动”，即可启动一个Job。

图 3-11 启动作业

----结束

3.7 管理 Ray Job

前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。
- 已有至少一个可用的Ray集群。
- 已有至少一个可用的Job作业。

操作步骤

步骤1 登录**Fabric工作空间管理台**。

步骤2 选择已创建的工作空间，单击“进入工作空间”，选择“开发与生产 > Job定义”。

步骤3 可根据需要在操作列选择Job的启动、查看、删除等操作。可根据Job名称、状态、运行端点名称、类型过滤不同的Job。

步骤4 通过操作列“查看Dashboard”，打开Ray自带的dashboard工具，查看Job的运行情况详情。

图 3-12 示例图片

----结束

3.8 查看 Ray dashboard

创建Ray集群后，运行Ray Job，如果需要查看Job的运行情况，或者查看Ray集群的详细信息，可通过打开Ray自带的dashbaord查看。

前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。
- 已有至少一个可用的Ray集群。
- 已有至少一个可用的Job作业。

操作步骤

查看路径的方法有以下两种：

- 方法一：通过Ray集群页面进入dashboard。
  a. 登录F**Fabric工作空间管理台**。
  b. 选择已创建的工作空间，单击“进入工作空间”，选择“资源与资产 > Ray集群”。
  c. 单击想要查看dashboard的Ray集群。
  d. 单击最下方的访问链接。

  图 3-13 示例图片

- 方法二：通过Job运行页面进入dashboard。
  请参见**管理Ray Job**中通过Job进入dashboard查看。

3.9 删除 Ray 集群

前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。
- 已有至少一个Ray集群。
- 如果当前Ray集群有运行Job记录，则需要先删除Job才能删除Ray集群。

操作步骤

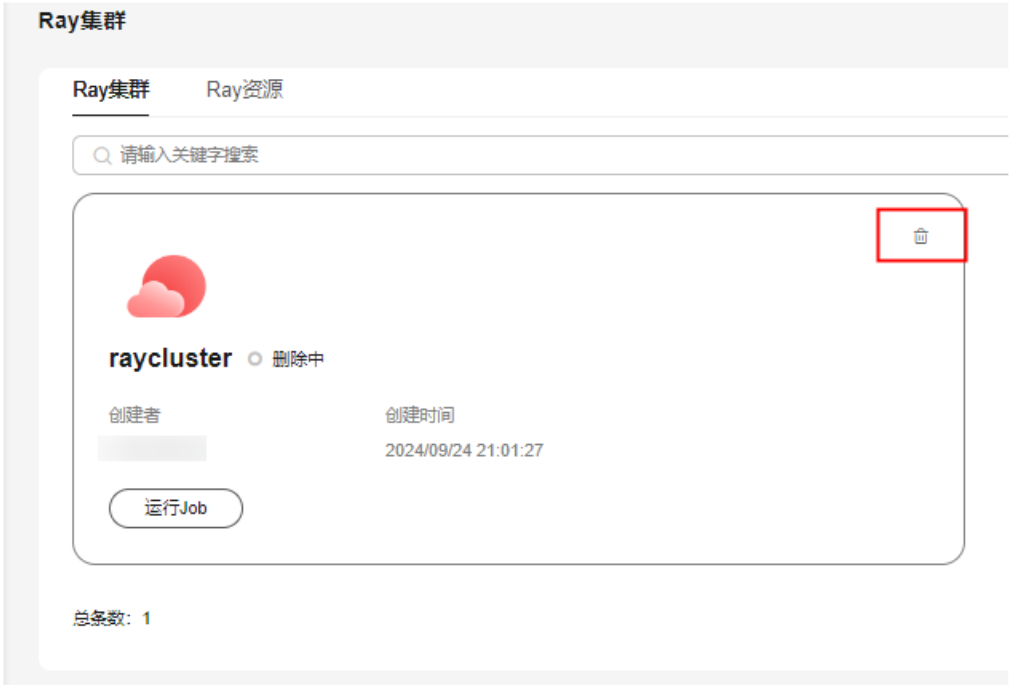
步骤1 登录**Fabric工作空间管理台**。

步骤2 选择已创建的工作空间，单击“进入工作空间”，选择“资源与资产 > Ray集群”。

步骤3 选择需要删除的Ray集群，单击右上角的“删除”按钮即可删除对应的Ray集群。

> **⚠ 注意**
>
> Ray集群一旦删除所有记录都会被清理掉，且无法恢复。请谨慎操作。

图 3-14 删除示例



步骤4 在弹出的二次确认界面确认后，输入“DELETE”后单击“确认”，即可删除Ray集群。

----结束

3.10 退订 Ray 资源

前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。
- 已购买Ray资源。

操作步骤

步骤1 登录**Fabric工作空间管理台**。

步骤2 选择已创建的工作空间，单击“进入工作空间”，选择“资源与资产 > Ray集群”。

步骤3 包周期与按需的操作有所不同，分别如下：

- 包周期：在对应Ray资源的操作列单击“更多 > 退订”。

- 按需：直接单击操作列“删除”。

图 3-15 退订 Ray 资源

说明

删除/退订Ray资源后无法恢复，且可能影响已存在的Ray集群状态。

步骤4 在弹出的二次确认界面确认后，输入“DELETE”后单击“确认”，即可删除已订购的Ray资源。

----结束

3.11 查看指标

为使用户更好地掌握Ray集群资源的使用情况，云服务平台将指标上报到了应用运维管理AOM，用户可以通过应用运维管理AOM查询资源使用情况。

前提条件

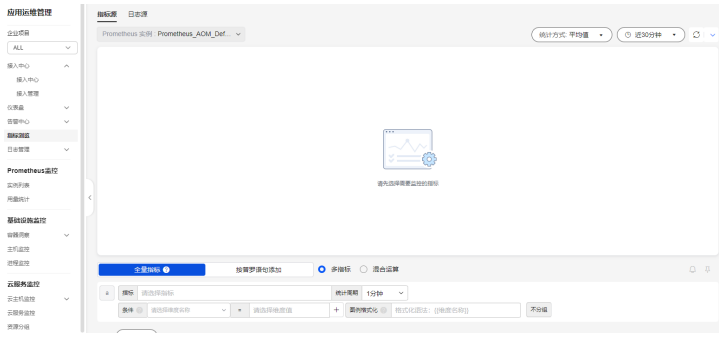
- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。
- 已有至少一个Ray集群。

操作步骤

步骤1 登录应用运维管理平台。

步骤2 选择指标预览，指标源选择Prometheus_AOM_Default。

图 3-16 配置指标源



步骤3 全量指标中输入指标名称进行查询。

表 3-7 监控指标

指标名称	描述
fabric_dpu_cpu_usage	该指标用于统计Ray集群head和worker的cpu资源使用率。 单位：百分比。
fabric_dpu_mem_usage	该指标用于统计Ray集群head和worker的内存资源使用率。 单位：百分比。

----结束

4 大模型推理场景

4.1 大模型推理场景介绍

常见的大模型包括大语言模型、多模态大模型、文生图大模型等，其中大语言模型支持文本生成，可以根据用户输入的提示词（prompt）进行推理，可广泛应用于以下领域：

- 问答系统：大语言模型可以处理自然语言，理解用户的意图，回答用户提出的问题。
- 内容生产：大语言模型可以基于给定的文本或主题生成连贯的文章、故事、对话等。
- 文本摘要：大语言模型可以对长文本进行摘要，提取关键信息，方便用户快速了解文本内容。
- 机器翻译：大语言模型可以处理多种语言之间的翻译任务，实现跨语言交流。

当前Fabric提供以下两种方式进行推理：

- **用公共推理服务进行推理**：Fabric提供基于开源大语言模型（Qwen2、GLM4等）的公共推理服务，用户可以在推理端点查看公共端点，选择自己想用的端点进行开通，然后就可以在试验场使用公共推理服务。该方式无需部署，开通后即可使用常见的开源大模型进行推理。
- **创建我的推理服务进行推理**：Fabric支持用户创建自己专属的推理服务进行部署，用户可以上传自己的大语言模型，也可以使用公共的大语言模型进行部署。在Fabric模型页面创建的模型是仅自己可见，其他用户不可见。用户可以查看和删除模型，也可以对模型版本进行管理，包括新增、查看和删除模型版本。

4.2 用公共推理服务进行推理

4.2.1 查看公共推理服务

推理端点试用期内，可以直接使用公共推理服务进行推理。目前的公共推理服务是基于开源大模型部署的，列表如下（实际的推理服务以服务为准）：

表 4-1 公共推理服务

名称	描述	免费额度	最大上下文长度	prompt模板长度	最大输出token
QWEN_2_72B	Qwen2在包括语言理解、生成、多语言能力、编码、数学和推理在内的多个基准测试中，超越了大多数以前的开放权重模型，与专有模型表现出竞争力。该模型参数规模为720亿。	公测期间提供100万token免费配额，超过配额不可用，也没办法再购买；有效期为服务开通90天内，超过时间则失效。	16k	23	16360

4.2.2 开通推理服务

对于公共推理服务，用户需要先申请开通，开通后才可以使用。开通公共推理服务之后用户会获得一定的免费配额，并在一定的时间内有效，超过将无法使用。如果用户想继续使用，建议部署推理服务使用。

前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。

操作步骤

步骤1 登录**Fabric工作空间管理台**。

步骤2 选择已创建的工作空间，单击“进入工作空间”，选择“推理端点 > 公共端点”，进入“公共推理端点”页面。

步骤3 试用期内的公共推理端点，会有运行中标志。公共端点页面可查看已经开通的公共推理端点。

图 4-1 查看开通的推理端点

----结束

4.2.3 在试验场进行推理

Fabric提供了试验场，方便用户在页面上选择推理服务进行推理。试验场支持流式推理，支持用户配置max_tokens等不同的推理参数，还支持不同的推理服务对比。

约束与限制

使用公共推理服务时的通用约束限制如下：

- Token配额约束：每种公共推理服务都有免费配额限制，超过配额不可用，也无法再购买。每种公共推理服务的配额为当前用户在当前局点下所有工作空间共享；
- 时间约束：有效期为服务开通90天内，超过时间则失效。同一个推理服务在不同工作空间下面开通，以首次开通为准。
- 不同的模型有不同的上下文长度约束，请见表**公共推理服务**。
- 不保证SLA，如果想要更高的性能，建议创建自己的推理服务进行推理；

前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。
- 已开通公共推理服务，开通流程请参见**开通推理服务**。

操作步骤

步骤1 登录**Fabric工作空间管理台**。

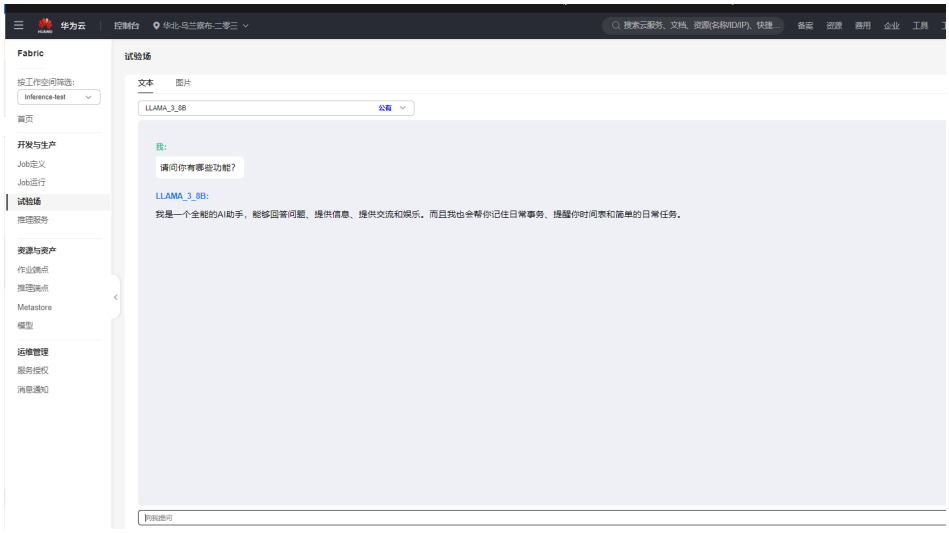
步骤2 选择已创建的工作空间，单击“进入工作空间”。

步骤3 在左侧菜单栏中选择“推理服务> 公共推理服务”，进入“公共推理服务”页面。

图 4-2 查看公共推理服务

步骤4 单击“试验场”，进入“试验场”页面，进行推理操作。

图 4-3 试验场推理界面



步骤5 调节推理参数（可选）。

如果想调节推理的一些参数，可以单击高级配置来调节推理的max_tokens等参数。参数列表如下。

表 4-2 推理参数说明

名称	说明
max_tokens	要在聊天完成中生成的最大token数。不同公共推理服务支持的最大max_tokens不一样，具体参考公共推理服务介绍。
temperature	Temperature是用于调整随机程度的数字。介于0和2之间。较高的值（如0.8）将使输出更随机，而较低的值（如0.2）将使输出更集中和确定性。
top_p	核心采样，用于控制AI模型根据累积概率考虑的标记范围。
frequency_penalty	数字介于-2.0和2.0之间。频率惩罚，控制文本中词汇的重复度，避免生成文本中某些词汇或短语出现过于频繁。正值会根据它们在文本中的现有频率惩罚新令牌，从而降低模型逐字重复同一行的可能性。
presence_penalty	数字介于-2.0和2.0之间。存在惩罚，控制文本中话题的重复度，避免在对话或文本中反复讨论相同的主题或观点。正值会根据到目前为止它们是否出现在文本中来惩罚新令牌，从而增加模型谈论新主题的可能性。

图 4-4 配置推理参数

步骤6 对比多个推理服务（可选）。

如果您想对比多个推理服务，Fabric也提供了推理服务的对比功能。您可以单击右上角的“新增对比”按钮进行新增，最多支持3个推理服务进行对比。

图 4-5 推理服务对比

----结束

4.3 创建我的推理服务进行推理

4.3.1 创建模型

在Fabric部署推理服务的时候除了使用公共模型，用户也可以自己创建模型。用户可以在Fabric模型页面创建模型，这些模型是属于用户个人，其他用户不可见。

约束与限制

创建模型的通用约束如下：

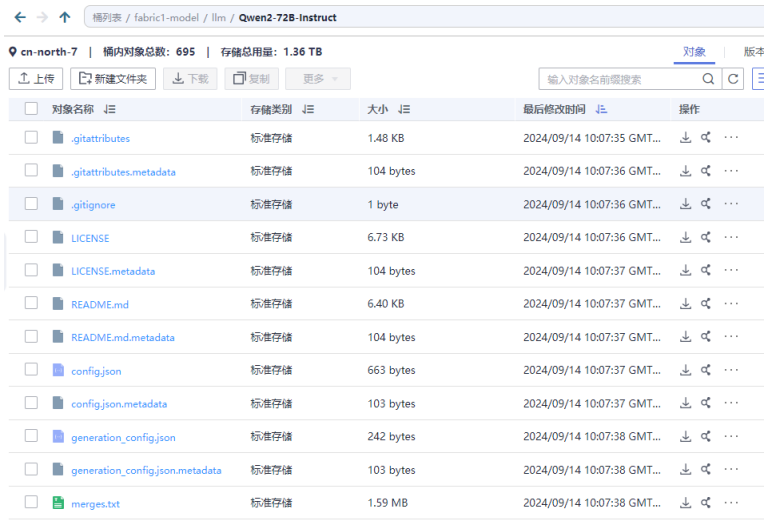
- 需要是Fabric支持的基模型，否则不支持，基模型列表如下：

  表 4-3 基模型列表

基模型类型	描述
QWEN_2_72B	Qwen2在包括语言理解、生成、多语言能力、编码、数学和推理在内的多个基准测试中，超越了大多数以前的开放权重模型，与专有模型表现出竞争力，参数规模为720亿。
GLM_4_9B	GLM-4-9B是智谱AI推出的最新一代预训练模型GLM-4系列中的开源版本。在语义、数学、推理、代码和知识等多方面的数据集测评中表现出较高的性能，参数规模为90亿。

- 模型格式需要为**safetensors**的格式。safetensors是Huggingface推出的一种可靠、易移植的机器学习模型存储格式，用于安全地存储Tensor，而且速度快。样例如下：

图 4-6 模型文件样例



前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。
- 已创建用于存储模型的OBS桶及文件夹，上传好符合要求的模型文件，并且模型存储的OBS桶与Fabric在同一区域。具体请参见**创建OBS桶**。

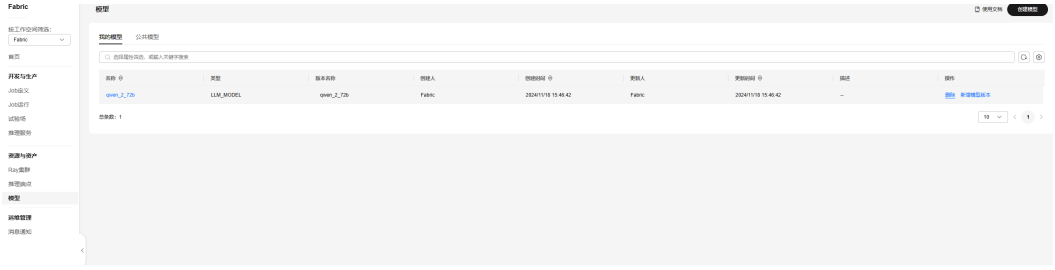
操作步骤

步骤1 登录**Fabric工作空间管理台**。

步骤2 选择已创建的工作空间，单击“进入工作空间”。

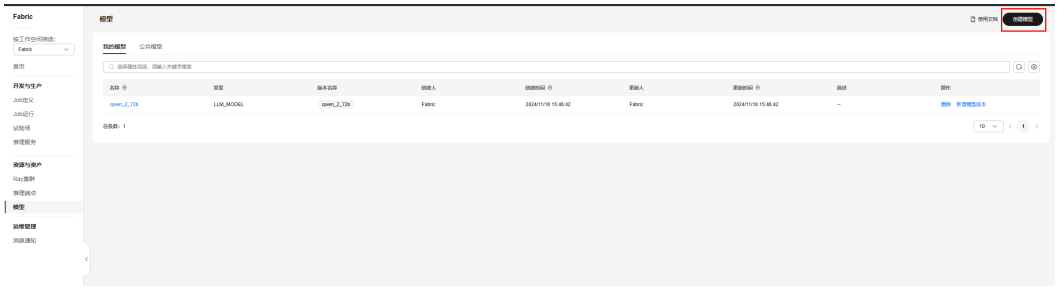
步骤3 在左侧菜单栏中选择“资源与资产 > 模型”，进入“模型”管理页面。

图 4-7 进入模型管理页面



步骤4 单击“创建模型”，进入“创建模型”页面。

图 4-8 进入创建模型页面



步骤5 填写模型基本信息，包括名称、描述等，并选择模型文件的OBS路径，然后单击“立即创建”，详细描述请见：

表 4-4 创建模型的基本信息

参数名称	说明
模型名称	必填，模型的名称。 长度为1-64，不支持重复名称。 只能包含中文、字母、数字、下划线、中划线、点、空格。
模型描述	可选，模型的描述信息。 长度为0-1024。不支持^!<>=&"'等特殊字符。
版本名称	必填，版本的名称。 长度为1-64，不支持重复名称。 只能包含中文、字母、数字、下划线、中划线、点、空格。
版本描述	可选，版本的描述信息。 长度为0-1024。不支持^!<>=&"'等特殊字符
基模型类型	必选，基模型的类型，描述具体请见**基模型列表**。
模型文件路径	必填，模型文件路径。目前支持OBS路径，该路径需要当前用户有读取的权限。

图 4-9 创建我的模型

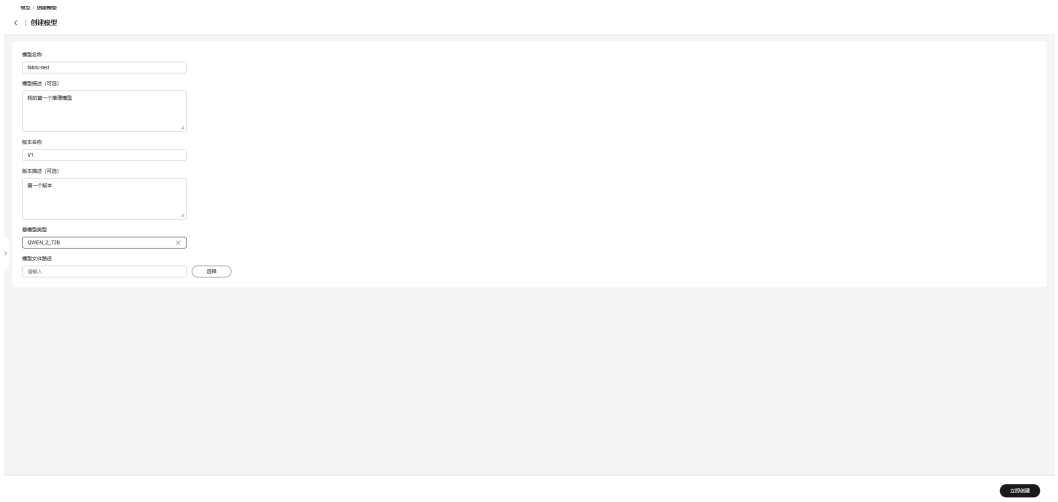
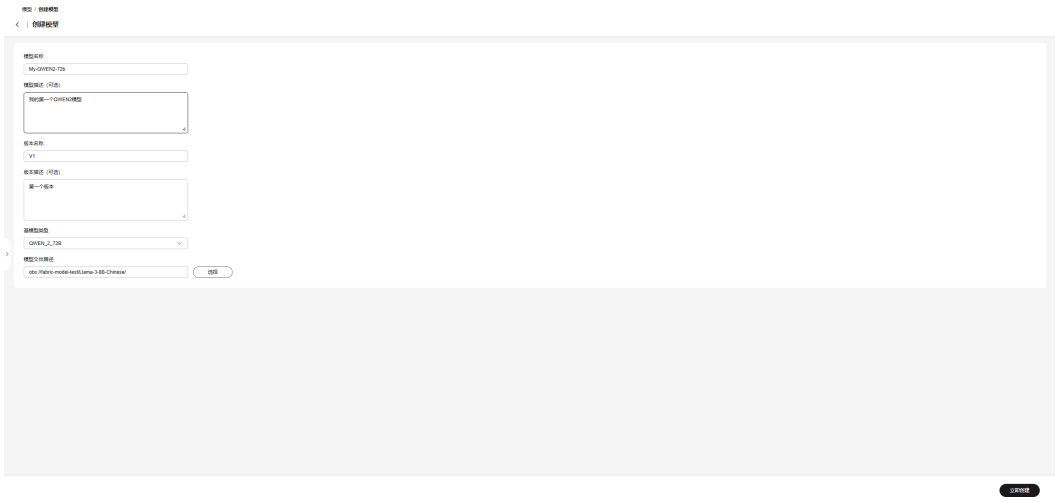


图 4-10 选择模型的 OBS 路径

步骤6 再次单击“我的模型”，即可在模型列表中看见刚创建的模型。

图 4-11 在模型列表中查看创建的模型



----结束

4.3.2 管理模型

在Fabric创建模型后，用户可以查看和删除模型，也可以对模型版本进行管理，包括新增、查看和删除模型版本。

前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。
- 已创建用于存储模型的OBS桶及文件夹，上传好符合要求的模型文件，并且模型存储的OBS桶与Fabric在同一区域。具体请参见**创建OBS桶**。

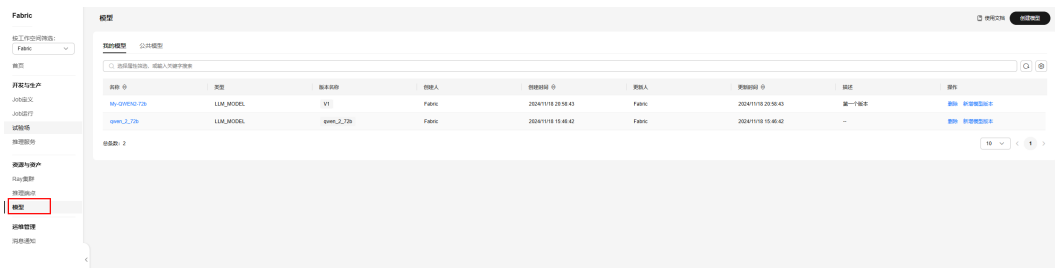
操作步骤

步骤1 登录**Fabric工作空间管理台**。

步骤2 选择已创建的工作空间，单击“进入工作空间”。

步骤3 在左侧菜单栏中选择“资源与资产> 模型”，进入“模型”管理页面。

图 4-12 进入模型管理页面



步骤4 查看当前模型下面的版本列表；您可以使用该版本，即设置为当前版本。

图 4-13 模型版本列表入口

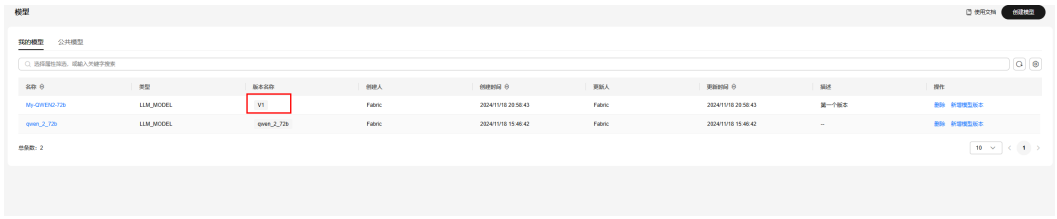
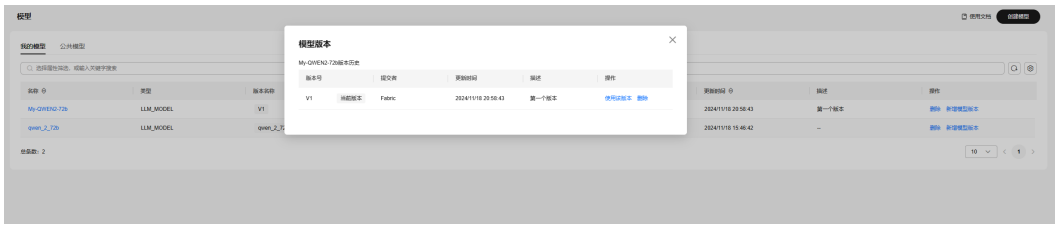


图 4-14 查看模型版本列表



步骤5 （可选）新增模型版本。

如果您的模型有迭代更新，可以选择新增模型版本。

在我的模型页面，单击操作列“新增模型版本”，填写基本信息后，单击“新增版本”即可完成新增。

模型版本新增后不支持修改。新增模型的基本信息如下：

表 4-5 创建模型版本的基本信息

参数名称	说明
版本名称	必填，版本的名称。 长度为1-64，不支持重复名称。 只能包含中文、字母、数字、下划线、中划线、点、空格。
版本描述	可选，版本的描述信息。 长度为0-1024。不支持^!<>=&"'等特殊字符。
模型文件路径	必填，模型文件路径。目前支持OBS路径，该路径需要当前用户有读取的权限。

图 4-15 新增模型版本入口

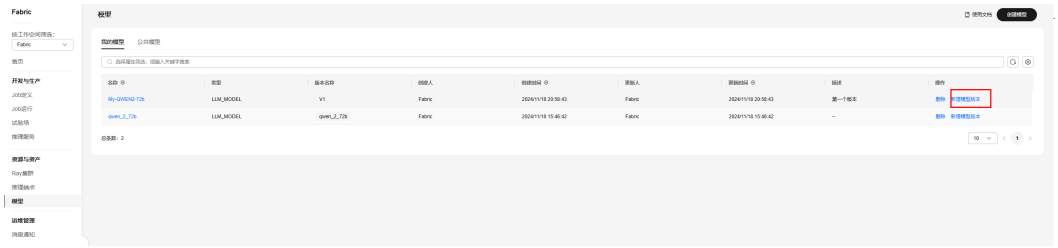
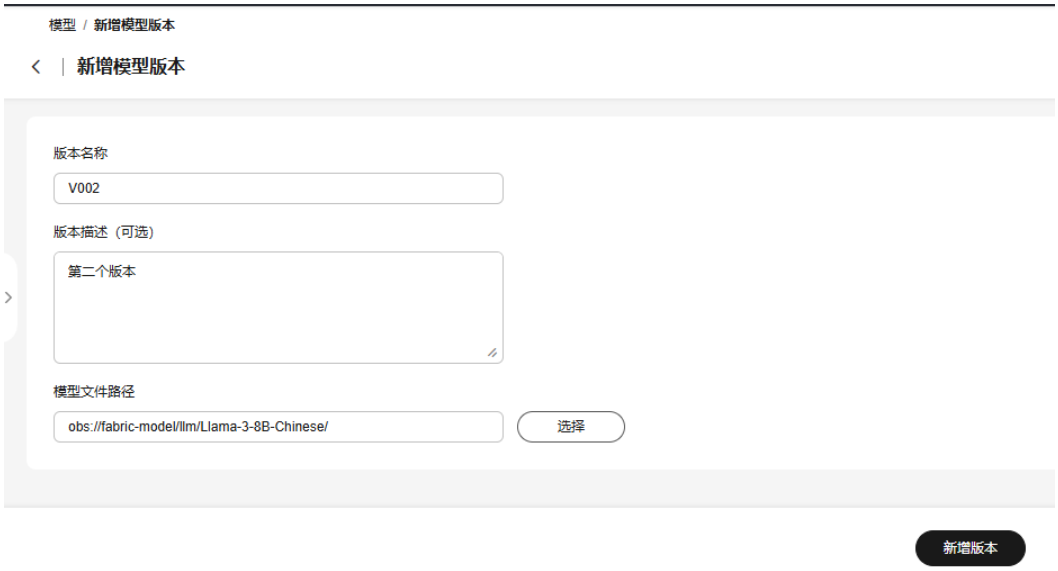


图 4-16 新增模型版本



步骤6 （可选）删除模型版本。

您也可以删除您不想要的模型版本。

单击页面操作列的“删除”按钮，再次确认后进行删除。

图 4-17 删除模型入口

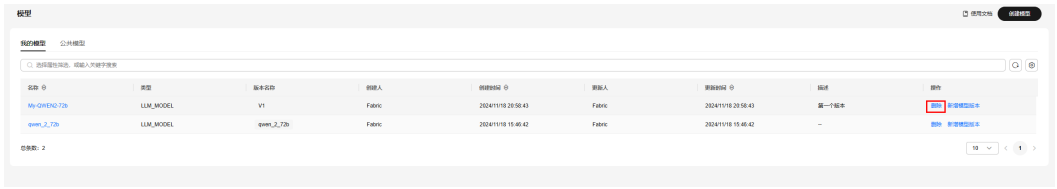
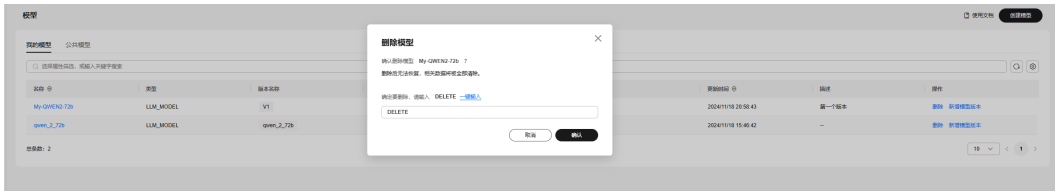


图 4-18 删除模型



----结束

4.3.3 创建推理端点

用户在创建推理服务之前，需要先创建推理端点。创建推理端点的时候可以配置最大资源数，然后在推理端点之上创建推理服务，推理端点上的所有推理服务的总资源数不能超过推理端点的最大资源数，方便用户控制推理端点的资源使用量；

前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。

操作步骤

步骤1 登录**Fabric工作空间管理台**。

步骤2 选择已创建的工作空间，单击“进入工作空间”，选择“资源与资产->推理端点”。

步骤3 单击右上角的“创建推理端点”。参照**创建推理端点的基本信息**填写端点的名称、描述、资源规格和数量等基本信息，单击“创建”。

图 4-19 进入推理端点创建页面

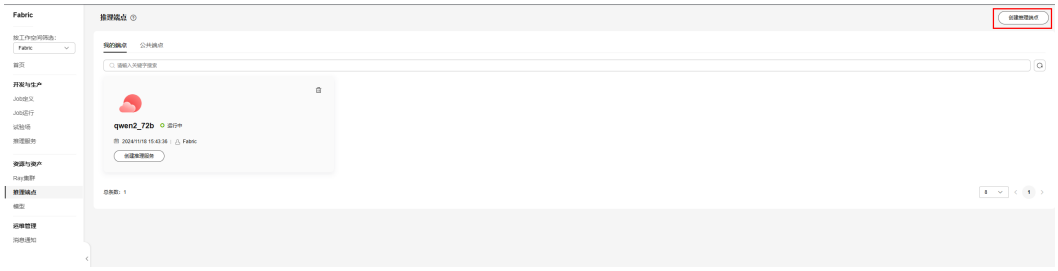
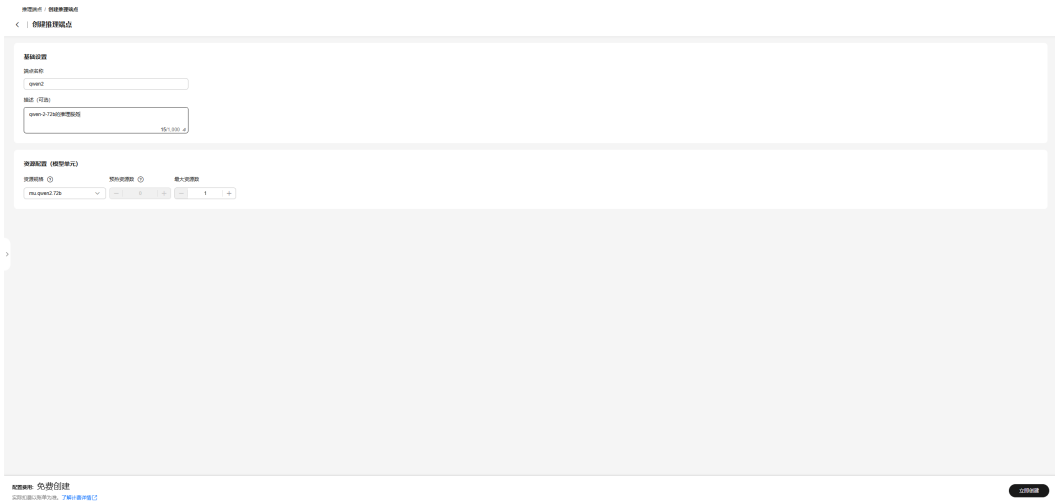


表 4-6 创建推理端点的基本信息

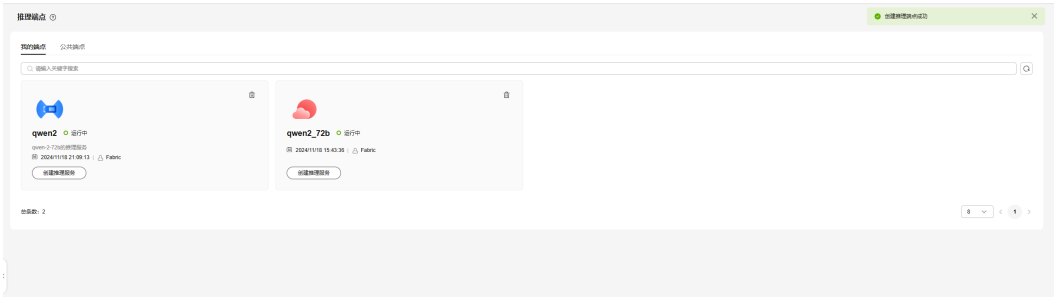
参数名称	说明
名称	必填，推理端点名称。 长度为1-64，不支持重复名称。 只能包含中文、字母、数字、下划线、中划线、点、空格。
描述	可选，推理服务的描述信息。 长度为0-1024。不支持^!<>=&"'等特殊字符。
资源规格	必填，资源规格，不同的资源规格支持不同的模型。
预热资源数	目前只支持0，推理端点的预热资源数。
最大资源数	必填，推理端点的最大资源数。最大值不能小于1，最大为1000。同时最大资源数不能小于预热资源数。

图 4-20 创建推理端点



步骤4 返回“资源与资产 > 推理端点”页面，选择“我的端点”即可查看已创建的端点。

图 4-21 推理端点



----结束

4.3.4 创建推理服务

在Fabric进行推理的时候，除了选择已有的公共推理服务进行推理，用户也可以部署自己的推理服务进行推理。

在Fabric部署推理服务的时候需要先有模型，您可以使用前面自己创建的模型，为了方便您操作，Fabric也默认提供了一些开源的公共模型，相关列表如下：

表 4-7 公共模型

模型名称	简介	基模型类型	算力要求(MU)	最大上下文长度	prompt模板长度	最大输出token
Qwen 2 72B Instruct	Qwen2在包括语言理解、生成、多语言能力、编码、数学和推理在内的多个基准测试中，超越了大多数以前的开放权重模型，与专有模型表现出竞争力。该模型参数规模为720亿。	QWEN_2_72B	8	16k	23	16360
Glm 4 9B Chat	GLM-4-9B是智谱AI推出的最新一代预训练模型GLM-4系列中的开源版本。在语义、数学、推理、代码和知识等多方面的数据集测评中表现出较高的性能。该模型参数规模为90亿。	GLM_4_9B	2	32k	16	32751

Prompt模板长度为系统prompt，不管用户输入什么，系统都会将prompt模板加入到输入中。最大上下文长度包括prompt模板长度、用户最大输入token长度和最大输出token之和。

用户可以在模型导航栏下查看公共模型信息，可以使用公共模型部署推理服务，但是不允许删除公共模型。

约束与限制

部署推理服务时的通用约束限制如下：

- 推理服务资源规格最小值为1，最大值为100
- 部署推理服务的时候选择的推理端点下的推理服务资源最大值不能超过推理端点的最大资源数。

前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。
- 已创建推理端点。
- 已创建用于推理的模型。

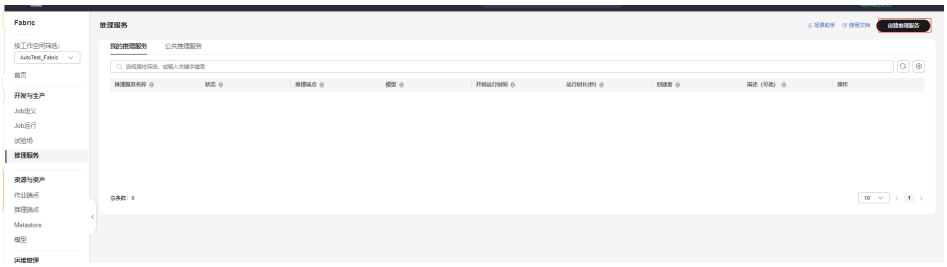
操作步骤

步骤1 登录**Fabric工作空间管理台**。

步骤2 选择已创建的工作空间，单击“进入工作空间”，在左侧菜单栏中选择“推理服务 > 我的推理服务”。

步骤3 进入“我的推理服务”页面。选择右上角的“创建推理服务”，进入创建页面。

图 4-22 进入推理服务创建页面



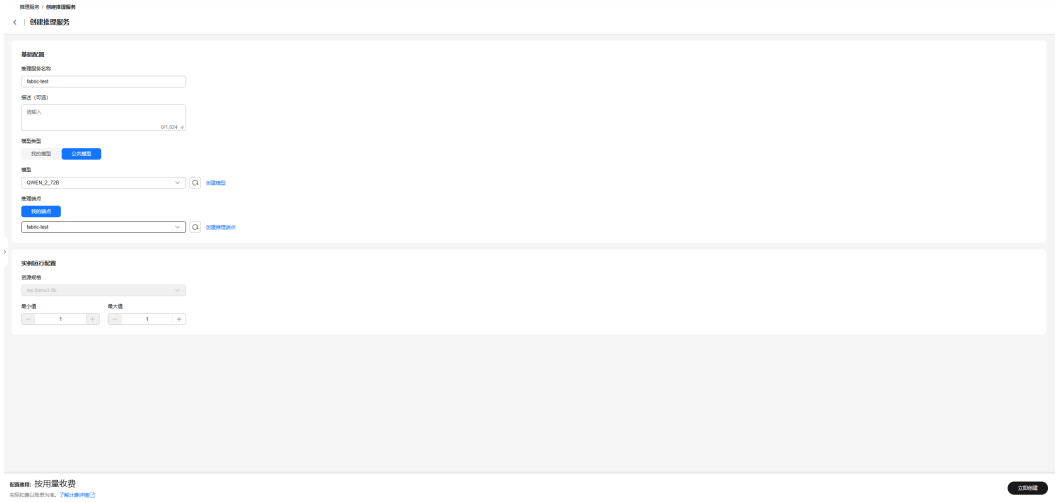
步骤4 填写创建推理服务的名称、描述等基本信息，并选择推理端点和模型。模型可以选择公共模型或者我的模型。然后配置资源最小值和最大值。详细描述请见下表。

表 4-8 创建推理服务的基本信息

参数名称	说明
名称	必填，推理服务名称。 长度为1-64，不支持重复名称。 只能包含中文、字母、数字、下划线、中划线、点、空格。
描述	可选，推理服务的描述信息。 长度为0-1024。不支持^!<>=&"'等特殊字符
推理端点	必填，用户自己创建的推理端点。
模型	必填，模型名称，可以是公共模型，也可以是用户自己创建的模型。
模型版本	必填，模型的版本。

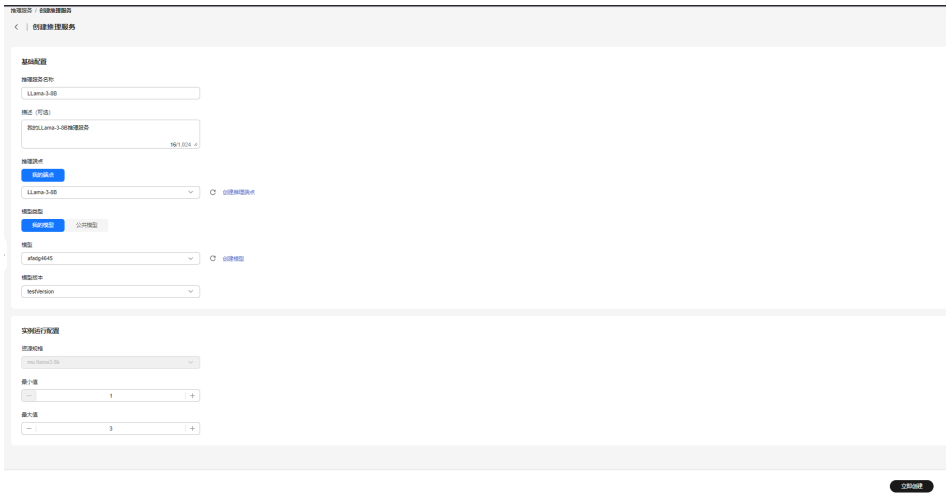
参数名称	说明
资源规格	必填，资源规格，需要与推理端点的规格保持一致，否则不支持。
最小值	必填，推理服务的最小实例数，即使没有请求，也会创建最小的实例数。最小值不能小于1，最大为1000。推理服务会根据不同的请求负载，在最小实例数和最大实例数之间进行自动扩缩。
最大值	必填，推理服务的最大实例数。最大值不能小于1，最大为1000。同时最大值不能小于最小值，并且最大值应该小于等于所选推理端点的最大资源数。同一推理端点下的所有推理服务的最大资源数之和应该小于等于所选推理端点的最大资源数。推理服务会根据不同的请求负载，在最小实例数和最大实例数之间进行自动扩缩。请求增加后，推理服务的实例数量不会超过最大值。

图 4-23 创建推理服务-公共模型



如果您想使用自己的模型，可以选择我的模型。这个模型需要您提前创建好，可以参考：**创建模型**。

图 4-24 创建推理服务-我的模型



步骤5 填写完成后，单击“立即创建”即可。

步骤6 在推理服务页面可查看已创建的推理服务。

----结束

4.3.5 使用推理服务进行推理

部署完推理服务之后，用户可以在试验场选择已有的推理服务进行推理，也可以调用API进行推理，具体请参考API文档（API链接到API参考）。下面是使用试验场进行推理的步骤：

前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。
- 已创建推理服务。

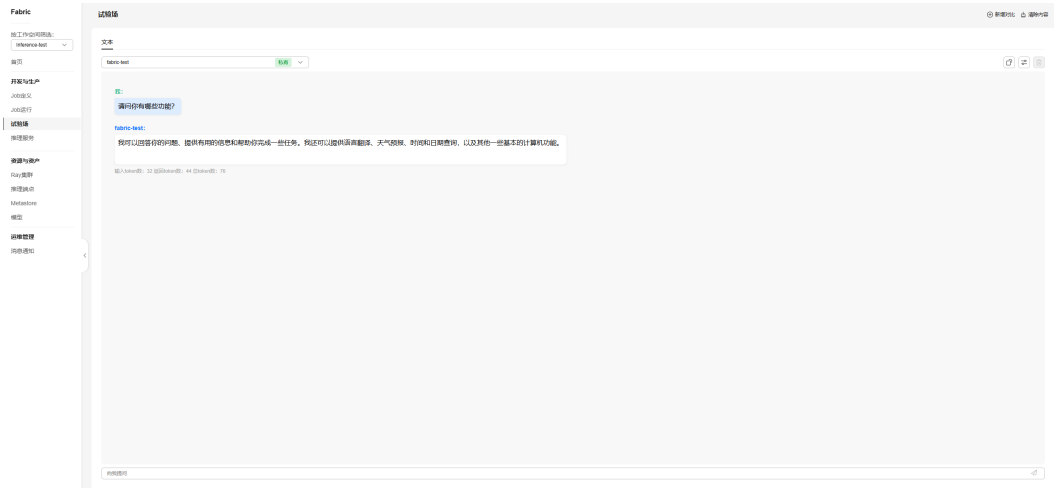
操作步骤

步骤1 登录Fabric工作空间管理台。

步骤2 选择已创建的工作空间，单击“进入工作空间”，在左侧导航栏选择“开发与生产 > 试验场”。

步骤3 单击“试验场”，进入“试验场”页面，进行推理。

图 4-25 选择公共推理服务进行推理



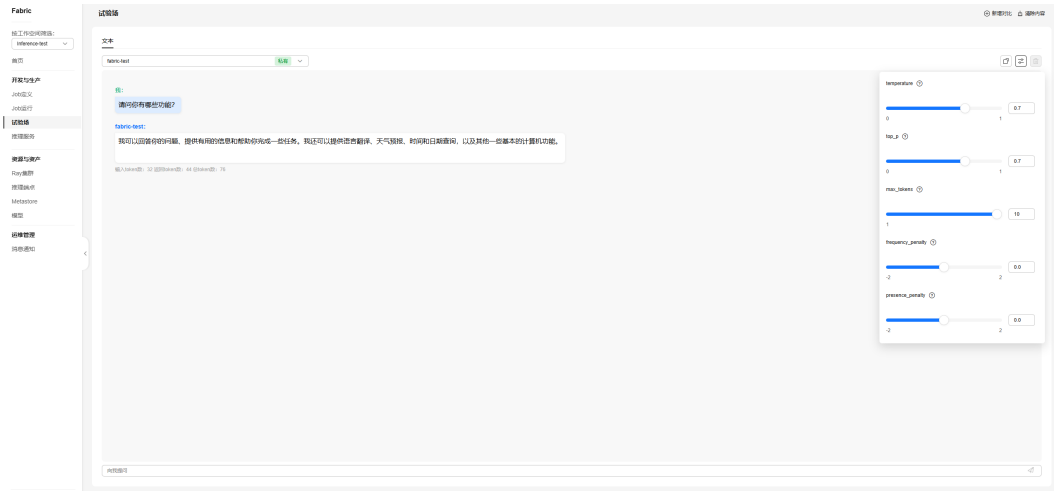
步骤4 （可选）参数调节。

如果需要调节推理的一些参数，可以单击高级配置来调节推理的max_tokens等参数。参数说明如下：

表 4-9 推理参数说明

名称	说明
max_tokens	要在聊天完成中生成的最大token数。不同公共推理服务支持的最大max_tokens不一样，具体参考公共推理服务介绍。
temperature	Temperature是用于调整随机程度的数字。介于0和2之间。较高的值（如0.8）将使输出更随机，而较低的值（如0.2）将使输出更集中和确定性。
top_p	核心采样，用于控制AI模型根据累积概率考虑的标记范围。
frequency_penalty	数字介于-2.0和2.0之间。频率惩罚，控制文本中词汇的重复度，避免生成文本中某些词汇或短语出现过于频繁。正值会根据它们在文本中的现有频率惩罚新令牌，从而降低模型逐字重复同一行的可能性。
presence_penalty	数字介于-2.0和2.0之间。存在惩罚，控制文本中话题的重复度，避免在对话或文本中反复讨论相同的主题或观点。正值会根据到目前为止它们是否出现在文本中来惩罚新令牌，从而增加模型谈论新主题的可能性。

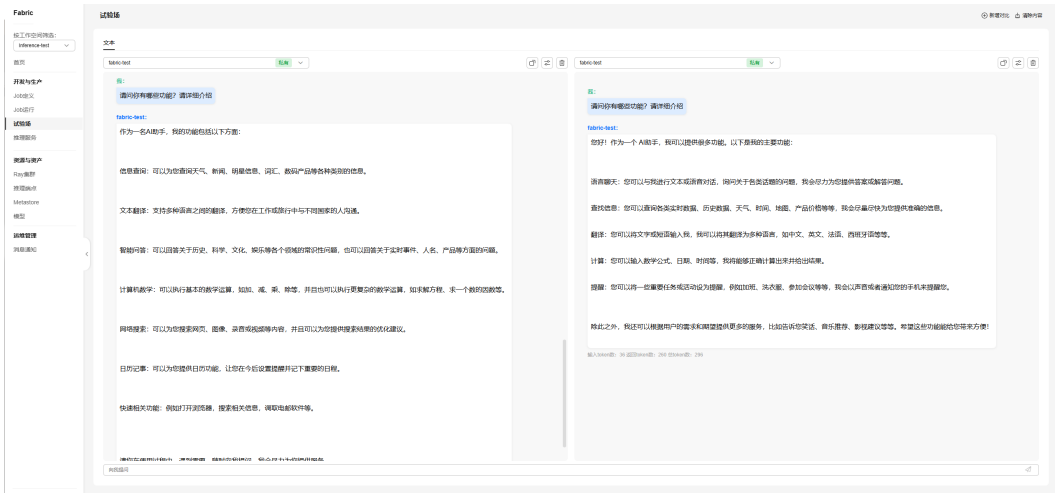
图 4-26 配置推理参数



步骤5 （可选）多个推理对比。

如果需要对比多个推理服务时，可以单击右上角的“新增对比”按钮进行新增，最多支持3个推理服务进行对比。

图 4-27 对比



----结束

4.3.6 删除推理服务

当您不想使用推理服务的时候，您可以删除自己创建的推理服务。

前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。
- 已创建推理服务。

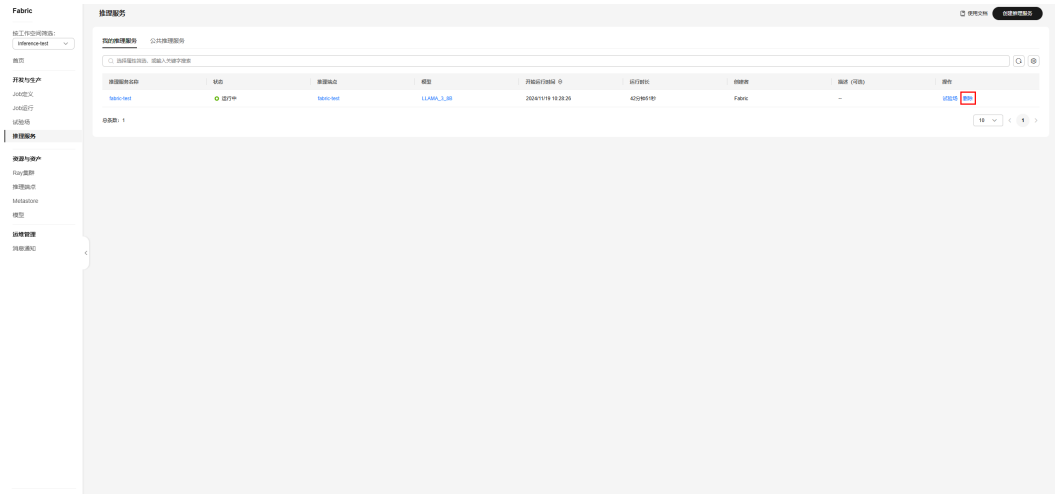
操作步骤

步骤1 登录**Fabric工作空间管理台**。

步骤2 选择已创建的工作空间，单击“进入工作空间”，选择“开发与生产 > 推理服务”。

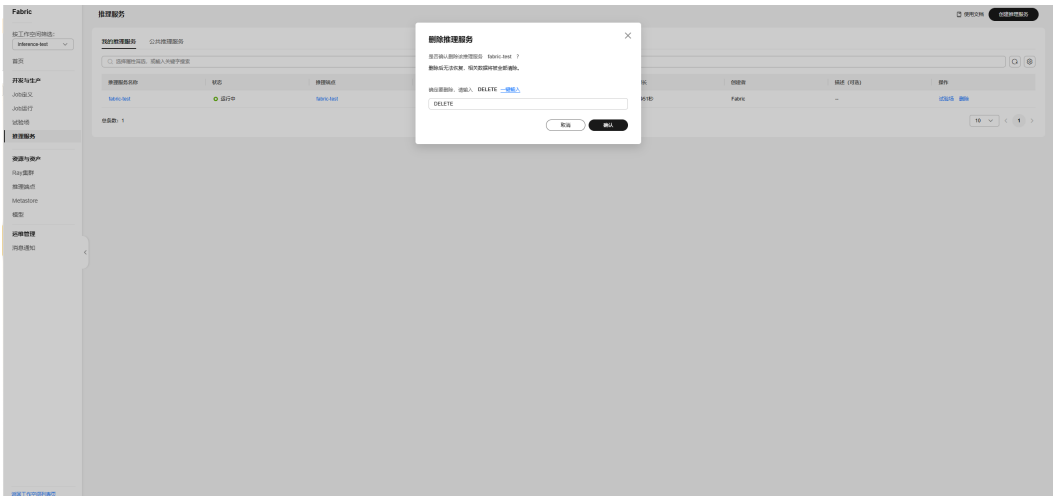
步骤3 选择想要删除的推理服务，单击其操作栏的“删除”按钮进行删除。

图 4-28 触发推理服务删除



步骤4 在弹出的二次确认界面确认后，输入“DELETE”后单击“确认”，即可完成删除。

图 4-29 确认删除推理服务



----结束

4.3.7 删除推理端点

前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。
- 已创建推理端点。

操作步骤

步骤1 登录**Fabric工作空间管理台**。

步骤2 选择已创建的工作空间，单击“进入工作空间”，选择“资源与资产->推理端点”。

步骤3 单击想要删除的推理端点右上角的垃圾桶标记，确认后删除推理端点。

图 4-30 推理端点

----结束

4.4 通过 AOM 查看全量指标

为使用户更好地掌握推理实例资源的使用情况，云服务平台将指标上报到了应用运维管理AOM，用户可以通过应用运维管理AOM查询资源使用情况。

前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。
- 已有至少一个推理实例。

操作步骤

步骤1 登录应用运维管理平台。

步骤2 选择指标预览，指标源选择Prometheus_AOM_Default。

步骤3 全量指标中输入指标名称进行查询。

表 4-10 监控指标

指标名称	描述
mu_usage	该指标用于展示当前推理实例的实际MU使用量 单位：个数。

----结束

5 运维管理

5.1 设置消息通知

消息通知功能用于通知用户其作业的执行情况。

前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。
- 需要配置FABRIC_SMN_POLICY委托，具体操作参考**配置Fabric云服务委托权限**。

操作步骤

步骤1 登录**Fabric工作空间管理台**。

步骤2 选择已创建的工作空间，单击“进入工作空间”，选择“运维管理->消息通知”。

步骤3 单击右上角的创建通知，参照**创建通知参数说明**设置参数，后单击“立即创建”。

图 5-1 创建通知

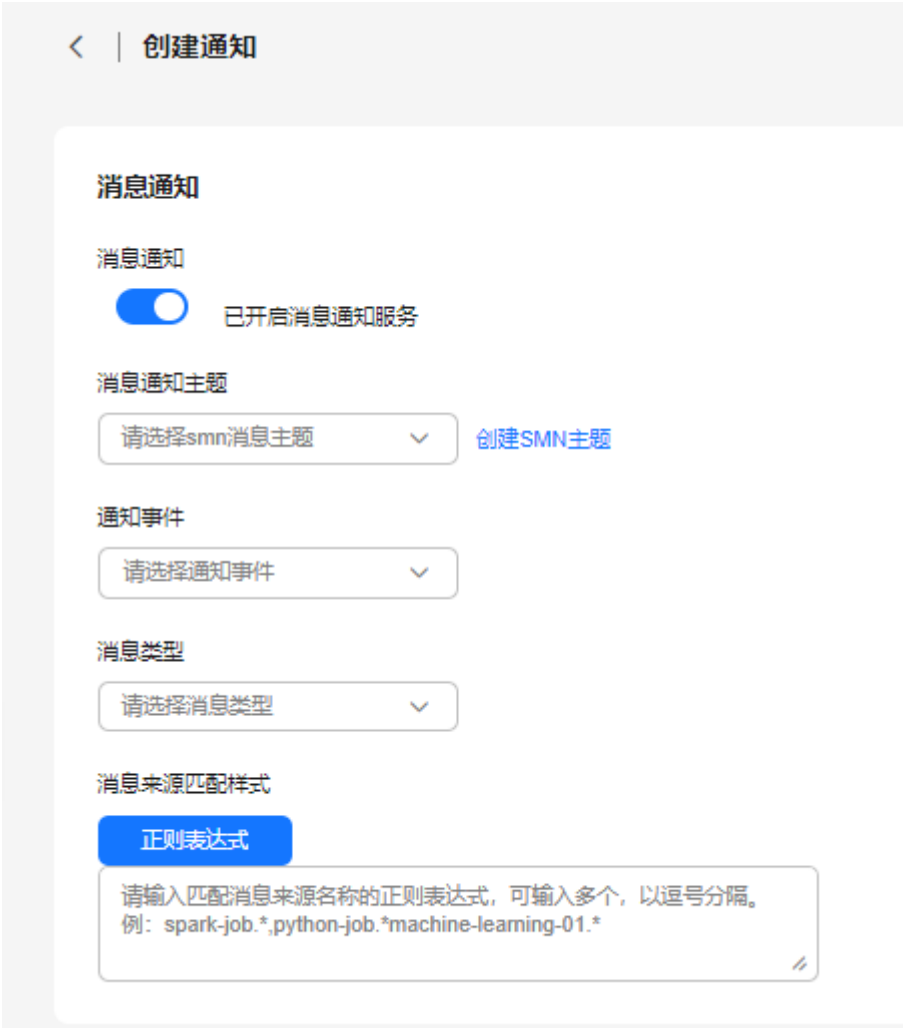


表 5-1 创建通知参数说明

参数	是否必选	说明
消息通知主题	是	在SMN中创建的主题，最终的消息会发送到对应的主题中。
通知事件	是	何时进行消息通知，分为： ● 成功时通知 ● 失败时通知 可以全选
消息类型	是	当前只支持选择作业，会对作业执行结果进行通知。

参数	是否必选	说明
消息来源匹配样式	是	需要匹配的消息来源，支持正则表达式 作业场景：作业名称的正则匹配。 例如：存在作业名称为test-job的作业，则可以填写test-j.*，test-job等方式进行匹配。

步骤4 创建成功后，可以在消息通知列表中看到已经创建的消息通知，单击消息来源后的数字可以看到当前配置的消息来源。

说明

相同的主题，通知事件，通知类型的不同消息来源会合并到一条记录中。

步骤5 删除通知（可选）。

如果需要删除消息通知，通过单击消息来源后的数字，在弹框中选择需要删除的消息类型，单击“删除”，确认后即可删除通知。

图 5-2 删除通知

----结束

5.2 删除消息通知

消息通知功能用于通知用户其作业的执行情况。入不需要时，可以通过删除操作删除通知。

前提条件

- 已有可正常使用的华为云账号。
- 已有至少一个正常可用的工作空间。
- 需要配置FABRIC_SMN_POLICY委托，具体操作参考**配置Fabric云服务委托权限**。
- 已有至少一个消息通知。

操作步骤

步骤1 登录**Fabric工作空间管理台**。

步骤2 选择已创建的工作空间，单击“进入工作空间”，选择“运维管理->消息通知”。

步骤3 单击消息来源后的数字，在弹框中选择需要删除的消息类型，单击“删除”，确认后即可删除通知。

图 5-3 删除通知

----结束