Fabric

产品介绍

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 什么是 Fabric

DataArts Fabric（简称Fabric）是华为云提供的数据+AI一站式开发平台，提供从数据处理、分析到模型微调、推理、部署上线的全生命周期管理能力，让数据工程师、数据科学家、AI应用开发工程师等多角色使用自己最熟悉的工具，在同一个工作台上工作，实现从开发到生产的高效协同。Fabric可实现自动扩缩，以支持最苛刻的应用程序。根据应用程序的需求以细粒度增量扩展资源，与为峰值负载预置资源池的服务相比，可为客户节省高达 50% 的成本。

Fabric基于Serverless资源池，让数据和AI的多种工作负载共池、CPU和NPU异构资源共池、开发和生产共池，变革客户的资源投资方式，实现在离线混部、训推一体，帮助客户削峰填谷，提升资源使用率。它提供极致体验，客户无需管理集群，零资源门槛启动开发和生产任务，使能客户在快速变化的业务中，低成本试错。

产品架构

Fabric提供高性能、高可靠、低时延、低成本的海量存储系统，与华为云的大数据服务组合使用，可大幅度降低成本，帮助企业简单快捷的管理大数据。

- **工作空间**

  工作空间是Fabric的基本单元，按团队切分工作空间，后续所有的操作都在工作空间中进行，隔离作业。

- **Cell**

  在故障处理上，Cell通过按照故障范围隔离，精准将故障限制在特定区域，避免扩散，保障系统稳定。

  在安全方面，Cell实现账号隔离，保障账号独立运行和权限，防止干扰和泄露，且具备网络隔离，分隔网络区域，增强安全性，防止攻击和入侵。

- **EndPoint**

  按作业类型切分资源组，避免不同类型作业之间的相互影响，减少等待和冲突。同时合理分配资源，提高资源利用率。

- **数据**

  设计规格为99.95%可用性，满足业务连续性的要求。

图 1-1 产品架构图

访问方式

Fabric提供了多种访问方式。

当前提供了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。除此外，LakeFormation也提供SDK客户端，更进一步方便计算引擎的对接集成。

- 控制台方式

  Fabric支持通过管理控制台访问，包含Ray作业、模型部署、模型推理等功能，您可以在管理控制台端到端完成您的数据、AI开发。

- API方式

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

- SDK方式

  如果您需要将Fabric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。Fabric的SDK是对Fabric提供的REST API进行的Python/Java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》。

2 产品优势

Fabric服务具有以下优势：

数智一站式开发，提供统一的开发体验

- 一个工作空间，提供多种工作负载，包含SQL、基于Ray的数据工程、模型推理。
- 基于LakeFormation统一管理结构化、半结构化、非结构化数据，数智开发全流程，一份元数据和一份权限控制。
- 数据+AI共享一份数据，客户无需进行数据复制。

开箱即用，资源弹性，按需使用

- 预置开源主流三方大模型的推理服务，客户可直接调预置推理服务API下发文本对话、文生图等任务，无需购买资源，按需付费。
- 全托管Ray、客户自建模型端点支持min-max自动弹性伸缩，应对客户请求波峰压力，实现资源动态分配。

开源生态

- 基于昇腾生态提供开源Ray的能力，并在开源Ray的能力上提供Redis高可靠。
- Ray dashboard提供可视化监控、故障排查、性能调优以及管理应用运行情况。
- 提供Ray CAP，客户可自定义Ray镜像。

3 应用场景

本节介绍Fabric服务的主要应用场景。

- **数据工程**

  高效处理大规模数据，通过并行计算加速数据处理过程，例如数据清洗、转换和聚合。
- **分布式机器学习**

  Ray支持分布式训练和调优，可以用于处理大规模数据集和模型，使得模型训练更加高效。
- **大模型**

  使用大模型实现智能对话、自动摘要、机器翻译、文本分类、图像生成等任务。
- **数据实时分析**

  提供标准SQL接口，用户仅需使用SQL便可实现海量数据查询分析。

4 计费说明

4.1 计费项

Fabric服务根据RAY、推理业务场景有不同的策略进行计费。详细的计费项及说明请参考**表4-1**。每个计费项的详细价格请参考**产品价格详情**。

表 4-1 计费项信息

计费项	*计费说明*
RAY资源	此处根据您创建的RAY资源规格和数量按照使用时间进行计费，不同的数据处理单元或AI计算单元规格的价格不同，支持包周期和按需付费两种模式。
模型算力单元时	此处根据您创建推理端点后部署模型实例所消耗的推理模型单元时长进行收费，支持按需付费。按照推理端点下实际的**模型实例数量* 算力单元资源数量 * 使用时长**按照秒级上报使用量，不同基模型对应的算力单元要求参考**公共模型**。

4.2 计费模式

Fabric服务提供包年包月、按需计费两种计费模式供您灵活选择。

- 包年包月：一种预付费模式，即先付费再使用，按照订单的购买周期进行结算。购买周期越长，享受的折扣越大。一般适用于计算资源需求量长期稳定的成熟业务。
- 按需：一种后付费模式，即先使用再付费，按照ModelArts计算资源的实际使用时长计费，秒级计费，按小时结算。按需计费模式允许您根据实际业务需求灵活地调整资源使用，无需提前预置资源，从而降低预置过多或不足的风险。一般适用于资源需求波动的场景，可以即开即停。

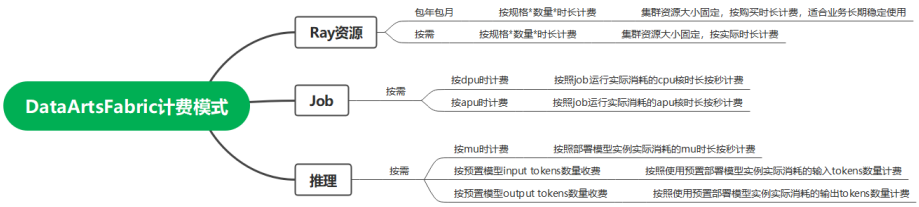
详细的计费区别请参考**表4-2**。

表 4-2 Fabric 服务计费模式

计费模式	付费方式	计费周期	适用计费项
包年包月	预付费 按照订单的购买周期结算。	按订单的购买周期计费。	RAY资源。
按需	后付费 按照云服务器实际使用时长计费。	按照资源实际使用量，每小时出话单扣费。	RAY资源、MU时。

Fabric服务不同业务场景的计费模式如图4-1所示。

图 4-1 Fabric 计费模式



4.3 计费样例

1. 样例1：Fabric服务RAY资源-计费说明

   RAY资源的“按需计费”模式都是秒级计费，Fabric产品价格详情中标出了每小时价格，您可以将每小时价格除以3600，即得到每秒价格。

   示例，某一RAY资源按需实例，fabric.ray.dpu.d1x规格价格为0.2元/小时，购买数量为5的按需实例根据规格数量 * 实际使用时长、按秒计费。

   - 使用30分钟，根据实际使用时长按秒计费：(0.2/3600)*5*30*60=0.5元。
   - 使用1小时，根据实际使用时长按秒计费：(0.2/3600)*5*60*60=1元。

2. 样例2：Fabric服务模型算力单元MU时-计费说明

   MU时的“按需计费”模式都是秒级计费，Fabric产品价格详情中标出了每小时价格，您可以将每小时价格除以3600，即得到每秒价格。

   示例，某一基模型为LLAMA3_8B推理端点实例，每个实例部署消耗2MU算力，假设MU时价格为30.0元/小时。根据部署模型实例数量 * MU换算比例 * 实际使用时长，按秒计费。

   - 使用30分钟，部署1个模型服务实例且数量无变化，根据实际使用时长按秒计费：(1/3600)*1*2*30*60=30元。
   - 使用1小时，其中一段15分钟时间内，服务实例数为2，剩余时间内实例数为1，根据实际使用时长按秒计费：(30/3600)*1*2*45*60 + (30/3600)*2*2*15*60=75元。

4.4 变更配置

当前Fabric服务计费项仅支持修改RAY资源，其他业务场景都是按使用量按需计费，不涉及订单变更流程，且暂时不支持计费方式变更，因此变更配置只涉及Ray资源大小变更场景。

- 修改Ray资源大小对费用影响如**表4-3**所示：

表 4-3 费用影响

当前计费模式	变更场景	对费用的影响
按需	RAY资源数量变更（升配/降配）	变更成功后，新的计费方式将立即生效。
包年包月	RAY资源数量增加（补差价升配）	升配后新资源数量将在原来已有的时间周期内立即生效。需按照与原规格的价格差异，结合已使用的时间周期，补上差价。 例如：（以下价格仅作示例，实际价格以**价格详情**为准） 客户于2024/11/1 购买了数量为1，规格为fabric.compute.dpu.d1x的RAY资源，购买时长为1个月，此时价格为18.4元/月，客户使用余额支付18.4元，实付金额为18.4元。 客户在2018/11/24 将Ray资源数量升级为5，价格为92元/月。 这时，剩余天数为 30 - 24 = 6天，升配费用=92 / 30 * 6 - 18.4 / 30 * 6 = 14.72元。 了解更多变更资源计费信息，请参见**变更资源费用说明**。
包年包月	RAY资源数量减少（即时降配）	降配成功后新的资源大小将在原来已有的时间周期内立即生效。按照与原规格的价格差异，结合已使用的时间周期，退款差价。 例如：（以下价格仅作示例，实际价格以**价格详情**为准） 客户于2024/11/1 购买了数量为5，规格为fabric.compute.dpu.d1x的RAY资源，购买时长为1个月，此时价格为18.4元/月，客户使用余额支付92元，实付金额为92元。 客户在2018/11/24 将Ray资源数量降级为4，价格为18.4元/月。 这时，剩余天数为 30 - 24 = 6天，降配退差价=92 / 30 * 6 - 18.4 / 30 * 6 = 14.72元。 了解更多变更资源计费信息，请参见**变更资源费用说明**。

4.5 费用账单

- **账单上报周期**

  按需计费模式的资源按照固定周期上报使用量到计费系统进行结算。按需计费模式产品根据使用量类型的不同，分为按小时、按天、按月三种周期进行结算，具体扣费规则可以参考**按需产品周期结算说明**。

  示例：按小时结算的云服务器在8:30删除资源，但是8:00~9:00期间产生的费用，通常会在10:00左右才进行扣费。在“费用中心 > 账单管理 > 费用账单 > 流水账单”中“消费时间”即按需产品的实际使用时间。

- **怎样查看完整的费用账单。**

  华为云支持按月查看费用账单汇总账单和账单详情。

  - 汇总账单：汇总数据可以展示不同汇总维度下的应付金额、扣费明细等数据，每个产品只展示一条汇总数据。当月最终汇总账单将在次月3日生成，在次月4日10点后可查看和导出。
  - 账单详情：根据需要选择不同的维度查看账单明细，包括流水账单、和自定义账单。自定义账单支持按使用量查看账单、按资源查看账单、按产品查看账单等。

- **怎样查看指定资源的账单**

  - 查询RAY资源的账单：

    i. 在云服务控制台RAY资源页面获取资源ID。

    ii. 根据资源ID在费用中心查看资源账单。详细操作参考费用中心的关于“根据资源ID/资源名称查询账单明细”操作指导。

  - 查询推理端点的算力单元时账单：

    Fabric服务推理端点ID与账单中上报的资源ID不一致，一个推理端点对应MU时计费资源ID的拼接规则为mu.{端点ID}。例如，假设推理端点ID为32de36ea-26c0-4876-ae48-fdbbb03cd455，其上报到账单中的MU时的资源ID为mu.32de36ea-26c0-4876-ae48-fdbbb03cd455。

    i. 在云服务控制台推理端点页面获取端点ID。

    ii. 拼接对应的MU时资源ID。

    iii. 根据上报账单的资源ID在费用中心查看资源账单。详细操作参考费用中心的关于“根据资源ID/资源名称查询账单明细”操作指导。

更多关于费用账单的详细描述请参考**账单介绍**。

4.6 停止计费

在查看账单后，如果您需要对某些资源停止计费可参考以下步骤：

1. 在账单中获取资源ID或资源名称等其他资源信息。
2. 根据上一步的信息，在云服务的控制台找到云服务资源。
3. 将资源停止计费。

具体操作如下：

- RAY资源：停止RAY资源的计费，需要删除/退订RAY资源，删除后可能导致已有的RAY集群不可用。

- 推理MU时：删除推理端点下的推理服务实例或者删除推理端点后，则不会产生费用。

4.7 到期与续费

客户欠费后，可以查看欠费详情。为防止相关资源被停止或者释放，需要客户及时进行充值，账号将进入欠费状态，需要在约定时间内支付欠款，详细操作请参考**欠费还款**。

如果没有及时的进行续费或充值，将进入宽限期。如宽限期满仍未续费或充值，将进入保留期。在保留期内资源将停止服务。保留期满仍未续费或充值，存储在云服务中的数据将被删除、云服务资源将被释放。详细说明请参考**资源停止服务或逾期释放说明**。宽限期与保留期的具体规则请参考**宽限期保留期**。

资源到期

如果账号欠费，会根据“客户等级”定义不同的保留期时长。进入保留期后您在Fabric服务中创建的Ray资源及模型实例会予以保留，账号会处于受限状态。在受限状态下，您无法通过控制台创建端点和使用端点，但仍然可以执行其他操作。保留期满仍未缴清欠款，存储在Fabric中的数据将被删除且无法恢复。

关于保留期时长等更多详细介绍，详见**保留期**。

续费

包年包月RAY资源支持续费，如需续费，请在管理控制台**续费管理**页面进行续费操作。您还可以设置到期自动续费。续费相关操作请参考**续费管理**。

5 产品规格

模型推理产品规格

表 5-1 模型推理产品规格

类型	规格	算力
MU	mu.llama3.8b	为llama3.8b模型，提供短token场景约400RPM算力。
	mu.llama3.70b	为llama3.70b模型，提供短token场景约100RPM算力。
	mu.llama3.1.8b	为llama3.1.8b模型，提供短token场景约190RPM算力。
	mu.llama3.1.70b	为llama3.1.70b模型，提供短token场景约130RPM算力。
	mu.qwen2.72b	为qwen2.72b模型，提供短token场景约1700RPM算力。
	mu.glm4.9b	为glm4.9b模型，提供短token场景约110RPM算力。

Ray 集群产品规格

表 5-2 ray 产品规格

类型	规格	算力
DPU	fabric.ray.dpu.d1x	提供约4CPU16G内存算力。
	fabric.ray.dpu.d2x	提供约8CPU32G内存算力。
	fabric.ray.dpu.d4x	提供约16CPU64G内存算力。
	fabric.ray.dpu.d8x	提供约32CPU128G内存算力。

类型	规格	算力
	fabric.ray.dpu.d16x	提供约64CPU256G内存算力。
	fabric.ray.dpu.d32x	提供约128CPU512G内存算力。
APU	fabric.ray.apu.b1.1x	提供昇腾AI加速型(B1)1卡算力
	fabric.ray.apu.b2.1x	提供昇腾AI加速型(B2)1卡算力
	fabric.ray.apu.b3.1x	提供昇腾AI加速型(B3)1卡算力
	fabric.ray.apu.b1.8x	提供昇腾AI加速型(B1)8卡算力
	fabric.ray.apu.b2.8x	提供昇腾AI加速型(B2)8卡算力
	fabric.ray.apu.b3.8x	提供昇腾AI加速型(B2)8卡算力

6 权限管理

如果您需要对华为云上购买的Fabric资源，为企业中的员工设置不同的访问权限，以达到不同员工之间的权限隔离，您可以使用统一身份认证服务（Identity and Access Management，简称IAM）进行精细的权限管理。该服务提供用户身份认证、权限分配、访问控制等功能，可以帮助您安全的控制华为云资源的访问。如果华为账号已经能满足您的要求，不需要通过IAM对用户进行权限管理，您可以跳过本章节，不影响您使用Fabric服务的其它功能。

IAM是华为云提供权限管理的基础服务，无需付费即可使用，您只需要为您账号中的资源进行付费。

通过IAM，您可以通过授权控制他们对华为云资源的访问范围。例如您的员工中有负责软件开发的人员，您希望他们拥有Fabric的使用权限，但是不希望他们拥有删除Fabric等高危操作的权限，那么您可以使用IAM进行权限分配，通过授予用户仅能使用Fabric，但是不允许删除Fabric实例的权限，控制他们对Fabric资源的使用范围。

目前IAM支持角色与策略授权。

表 6-1 角色与策略授权说明

名称	核心关系	涉及的权限	授权方式	适用场景
角色与策略授权	用户-权限-授权范围	• 系统角色 • 系统策略 • 自定义策略	为主体授予角色或策略	核心关系为“用户-权限-授权范围”，每个用户根据所需权限和所需授权范围进行授权，无法直接给用户授权，需要维护更多的用户组，且支持的条件键较少，难以满足细粒度精确权限控制需求，更适用于对细粒度权限管控要求较低的中小企业用户。

例如：如果需要对IAM用户授予可以创建华北-北京四区域的ECS和华南-广州区域的OBS的权限，基于角色与策略授权的场景中，管理员需要创建两个自定义策略，并且为IAM用户同时授予这两个自定义策略才可以实现权限控制。在基于身份策略授权的场景中，管理员仅需要创建一个自定义身份策略，在身份策略中通过条件键“g:RequestedRegion”的配置即可达到身份策略对于授权区域的控制。将身份策略附加主体或为主体授予该身份策略即可获得相应权限，权限配置方式更细粒度更灵活。

关于IAM的详细介绍，请参见**IAM产品介绍**。

角色与策略权限管理

Fabric服务支持角色与策略授权。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或角色，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，用户就可以基于被授予的权限对云服务进行操作。

Fabric部署时通过物理区域划分，为项目级服务。授权时，“授权范围”需要选择“指定区域项目资源”，然后在指定区域（如华北-北京四）对应的项目（cn-north-4）中设置相关权限，并且该权限仅对此项目生效；如果“授权范围”选择“所有资源”，则该权限在所有区域项目中都生效。访问Fabric时，需要先切换至授权区域。

下表列出了Fabric所有的系统权限。

表 6-2 Fabric 系统权限

系统角色/策略名称	描述	类别	依赖关系
DataArtsFabricFullPolicy	Fabric服务的所有权限。	系统策略	• IAM Agency Management FullAccess • OBS OperateAccess • LakeFormation ReadOnlyAccess
DataArtsFabricConsoleFullPolicy	在控制台页面使用Fabric服务的所有权限，包含DataArtsFabricFullPolicy的全部权限，以及部分在控制台页面需要的权限。	系统策略	• IAM Agency Management FullAccess • OBS OperateAccess • LakeFormation ReadOnlyAccess
DataArtsFabricReadOnlyPolicy	Fabric服务的只读访问权限。	系统策略	LakeFormation ReadOnlyAccess

下表列出了Fabric常用操作与系统权限的授权关系，您可以参照该表选择合适的系统权限。

表 6-3 Fabric 常用操作与系统权限的授权关系

操作	DataArtsFabricConsoleFullPolicy	DataArtsFabricFullPolicy	DataArtsFabricReadOnlyPolicy
查询Workspace列表	√	√	√

操作	DataArtsFabricConsoleFullPolicy	DataArtsFabricFullPolicy	DataArtsFabricReadOnlyPolicy
创建Workspace	√	√	×
修改Workspace	√	√	×
修改Workspace监控配置	√	√	×
删除Workspace	√	√	×
查询计算资源	√	√	√
创建计算资源	√	√	×
修改计算资源	√	√	×
删除计算资源	√	√	×
查询Workspace的Endpoint列表	√	√	√
创建Workspace的Endpoint	√	√	×
查询Workspace的Endpoint详情	√	√	√
修改Workspace的Endpoint	√	√	×
删除Workspace的Endpoint	√	√	×
查询作业列表	√	√	√
创建作业	√	√	×
查询作业	√	√	√
修改作业	√	√	×
删除作业	√	√	×
查询服务列表	√	√	√
创建服务	√	√	×
修改服务	√	√	×

操作	DataArtsFabricConsoleFullPolicy	DataArtsFabricFullPolicy	DataArtsFabricReadOnlyPolicy
查询服务	√	√	√
删除服务	√	√	×
创建模型	√	√	×
查询模型列表	√	√	√
查询模型	√	√	√
删除模型	√	√	×
修改模型	√	√	×
创建标签	√	√	×
删除标签	√	√	×
获取标签列表	√	√	√
查询指定资源标签	√	√	√
标签查询资源列表	√	√	√
创建消息通知策略	√	√	×
查询消息通知策略列表	√	√	√
删除消息通知策略	√	√	×
查询运行作业列表	√	√	√
运行作业	√	√	×
查询运行作业	√	√	√
删除运行作业	√	√	×
取消运行作业	√	√	×
调用推理服务实例	√	√	×
查询路由列表	√	√	√
查询Session信息	√	√	√
订阅公共端点	√	√	×

Fabric 控制台功能依赖的角色或策略

表 6-4 Fabric 控制台依赖服务的角色或策略

控制台功能	依赖服务	需配置角色/策略
服务授权	统一身份认证管理 IAM	IAM用户设置了IAM Agency Management FullAccess权限后才能在服务授权界面进行授权。
创建工作空间	湖仓构建服务 LakeFormation	设置了DataArtsFabricFullPolicy的用户可以创建工作空间，配置了LakeFormation ReadOnlyAccess后可以在创建工作空间时指定metastore为lakeformation metastore。
创建模型	对象存储服务OBS	IAM用户设置了DataArtsFabricFullPolicy之后，还需要设置OBS OperateAccess才能在模型管理界面创建模型并指定模型文件所在的OBS路径。
创建消息通知策略	统一身份认证管理 IAM 消息通知服务SMN	IAM用户设置了DataArtsFabricFullPolicy之后，还需要设置IAM Agency Management ReadOnly权限和SMN ReadOnlyAccess权限才能在消息通知页面创建消息通知策略。

相关链接

- **IAM产品介绍**
- **创建IAM用户并授权使用Fabric**
- **权限及授权项说明**

7 约束与限制

大模型 LICENSE 约束

不同的开源大模型有不同的LICENSE约束，详细请见下表：

表 7-1 大模型 LICENSE 约束

模型名称	LICENSE地址
Llama 3 8B Chinese Instruct	https://github.com/meta-llama/llama/blob/main/LICENSE
Llama 3 70B	https://github.com/meta-llama/llama/blob/main/LICENSE
Llama 3.1 8B Chinese Chat	https://huggingface.co/meta-llama/Meta-Llama-3.1-8B/blob/main/LICENSE
Llama 3.1 70B	https://huggingface.co/meta-llama/Meta-Llama-3.1-8B/blob/main/LICENSE
Qwen 2 72B Instruct	https://huggingface.co/Qwen/Qwen2-72B-Instruct/blob/main/LICENSE
Glm 4 9B Chat	https://huggingface.co/THUDM/glm-4-9b-chat/blob/main/LICENSE

公共推理服务约束与限制

- 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。