Fabric

产品介绍

文档版本 01

发布日期 2025-04-28

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 DataArtsFabric

DataArtsFabric（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，帮助企业简单快捷地管理大数据。

- **SQL引擎**

  Fabric提供分布式SQL引擎，实现了元数据服务、计算、缓存和存储的分层解耦和弹性，让每一层动态分配资源而不会影响另一层的性能或可用性。语句级别的弹性扩缩、高性能分布式分析引擎可帮助您在几秒钟内查询TB级别数据，在几分钟内查询PB级别数据。

- **分布式Ray**

  Fabric支持分布式计算框架Ray，来帮助客户解决规模日益增大的数据处理和机器学习/深度学习任务对分布式计算的问题，也为数据工程和机器学习工程提供统一的完整Workflow。Fabric Ray支持Ray-Data、Ray-Train、Ray-Serve模块，分别满足分布式数据预处理、分布式训练、分布式模型推理服务的应用场景。

- **在线推理**

  Fabric提供自研的高性能弹性推理引擎，支持客户基于默认的推理服务下发推理作业，也支持客户独立部署自定义模型。

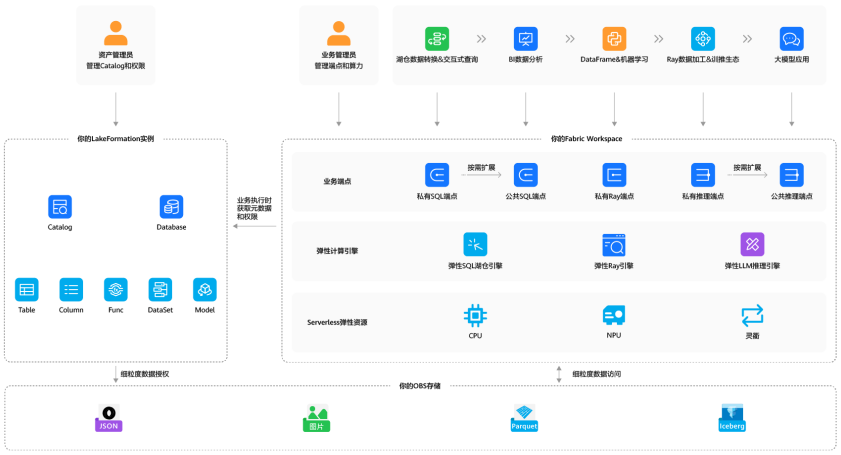
- **异构资源管理**

  Fabric支持CPU+NPU资源统一纳管、统一资源分配；资源调度粒度支持容器级和Actor级，并且支持安全沙箱来实现资源隔离、可靠容错。

- **多语义缓存加速**

  Fabric提供跨引擎、多模态、多语义加速，例如数据缓存、模型缓存、CheckPoint缓存。

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，Fabric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、SQL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，客户可直接调用预置推理服务API下发文本对话等任务，无需购买资源，按需付费。
- 推理服务支持自动扩缩。
- 全托管Ray支持Pod级自动扩缩，应对客户请求波峰压力，实现资源动态分配。
- SQL支持按资源计费与按查询计费两种模式，计算资源支持查询级别快速弹性。

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。
- SQL基于开放湖仓生态，支持ORC、Parquet、Iceberg等数据格式。

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 Fabric SQL 功能介绍

Fabric SQL 介绍

Fabric SQL是一个全托管式数据平台，利用华为云基础设施提供的资源池化和海量存储能力，结合并行执行、元数据解耦、计算持久化分离架构，实现了极致弹性和湖仓一体等能力，提供先进软件及服务（SaaS）技术。全新的无服务器架构可以让您在使用SQL语言处理组织复杂业务时，无需管理基础架构。

Fabric SQL架构基于华为云Fabric平台，主要由服务接入层、计算层与存储层组成。平台实现了元数据服务、计算、缓存和存储的分层解耦和弹性，让每一层动态分配资源而不会影响另一层的性能或可用性。语句级别的弹性扩缩、高性能分布式分析引擎可帮助您在几秒钟内查询TB级别数据，在几分钟内查询PB级别数据。

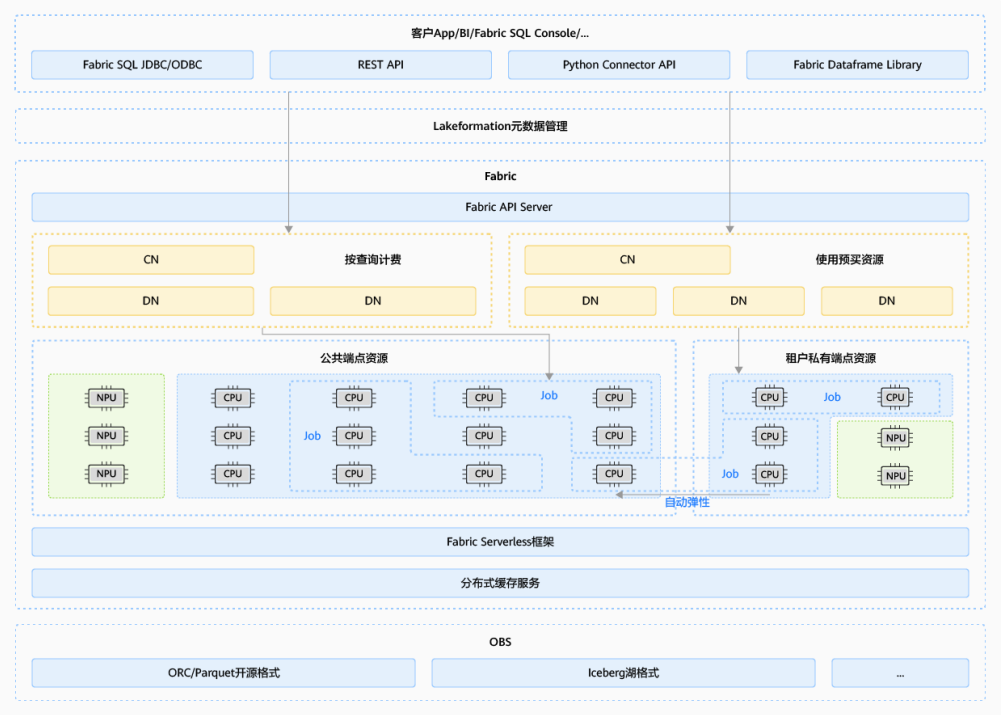
Fabric SQL支持加工和分析Iceberg、ORC、Parquet等开放结构化数据格式，支持开放湖生态，让您可以在使用多个数据湖生态服务时共享同一份数据。

Fabric SQL积极拥抱Data+AI生态并提供了Python UDF特性，支持用户在SQL中直接运行Python脚本，使能一站式AI数据处理；

Fabric SQL提供可视化界面，并提供JDBC驱动，方便与现有应用和第三方工具交互。同时提供REST API接口和Python API接口，方便开发者使用熟悉的编程语言转换和管理数据。

产品架构图

图 4-1 Fabric SQL 产品架构图



功能介绍

下表介绍了Fabric SQL的关键功能。

表 4-1 Fabric SQL 产品功能

特性	特性及规格描述
支持标准SQL语法	基于ANSI标准SQL规范扩展，支持GBK、UTF-8、SQL ASCII以及Latin-1字符集。
DDL	支持SCHEMA、TABLE的Create/Alter/Drop/Show/Describe。
数据类型	支持smallint, int, bigint, float, double, numeric, timestamp, date, varchar, char, bool, binary, string。
应用程序接口	标准JDBC 4.0、Restful API、Python Connector API。
事务能力	支持分区级别事务能力，支持并发控制。 ICEBERG支持完整的事务能力。
多租户	不同租户通过不同的CN/DN隔离，CN/DN通过独占POD来进行隔离。
数据导入导出	数据导入通过INSERT INTO导入，导出也通过insert into。从外表导出外表，格式转换支持：Parquet/ORC -> Parquet/ORC。

特性	特性及规格描述
弹性能力	支持两级弹性，资源池内根据查询特征秒级弹性计算单元（节点弹性速度秒级<2s），根据资源水位线触发额外计算资源扩缩容。 **说明** Beta特性，扩容操作需要通过管控面触发fabric服务购买容器。
SMP	提供节点内SMP并行计算能力，充分利用多核CPU资源，默认开启query_dop。
向量化执行	向量化执行引擎，提升OLAP性能。
统计信息搜集	Analyze完成统计信息搜集，提升优化器评估精度，确保数据库性能稳定高效。
存储格式支持	支持PARQUET/ORC/Iceberg格式。
DML支持	支持Insert Into/Insert Overwrite。
分区表	PARQUET/ORC/Iceberg均支持分区表。
视图	支持视图。
用户自定义函数(UDF)	支持用户以自定义函数的方式扩展SQL并统一执行，当前编程语言仅支持Python。
弹性计算规模	单查询最大规模256 弹性计算节点。
细粒度访问控制	表元数据由LakeFormation管理，使用IAM权限管理，当前由Lakeformation负责整体的权限控制。

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

FabricSQL 产品规格

表 5-3 SQL 产品规格

类型	规格	算力
DPU	fabric.sql.dcu.std	提供约1CPU4G内存算力。

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 • KMS Administrator（可选）
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 • IAM PolicyFullAccess • KMS Administrator（可选）
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	√	√	√
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	√	√	×

操作	DataArtsFabricConsoleFullPolicy	DataArtsFabricFullPolicy	DataArtsFabricReadOnlyPolicy
查询服务列表	√	√	√
创建服务	√	√	×
修改服务	√	√	×
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	√	√	√

操作	DataArtsFabricConsoleFullPolicy	DataArtsFabricFullPolicy	DataArtsFabricReadOnlyPolicy
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

7.1 Ray、XDS 约束限制

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。

7.2 Fabric SQL 约束限制

FabricSQL 约束与限制

技术指标	最大值
并发Session数	500
异步创建Session等待队列最大长度	1000

Fabric SQL 服务使用限制

表 7-2 Fabric SQL 服务使用限制

事项	说明
开通服务	一个账号支持开通一个Fabric SQL服务，并且该账号各子用户共用一个服务。
连接服务	提供了SQL编辑器，JDBC，SDK，API接口等多种连接方式。
运维操作	Serverless形态，不涉及扩容、升级、备份恢复、容灾等运维操作。
超时限制	由于用户授权Token存在时效性（8小时），单次请求时间如果超过Token有效期会导致语句执行失败。
SQL语法	参见**数据库操作使用限制**。

数据库操作使用限制

表 7-3 数据库操作

类别	语法	是否支持
基本功能	CREATE EXTERNAL TABLE	是
	DROP TABLE	是
	CREATE VIEW	是
	DROP VIEW	是
	INSERT	是
	SELECT	是
	TRUNCATE	是

类别	语法	是否支持
	EXPLAIN	是
	ANALYZE	是
	ALTER TABLE DROP PARTITIONS	是
	ALTER TABLE SET TABLEPROPERTIES	是
	ALTER TABLE UNSET TABLEPROPERTIES	是
	ALTER TABLE DROP COLUMNS	否（仅Iceberg支持）
	ALTER TABLE ADD COLUMNS	否（仅Iceberg支持）
	ALTER TABLE COLUMN RENAME	否（仅Iceberg支持）
	CREATE EXTERNAL TABLE AS	是