12 AI DataLake 常见问题

文档版本 01
发布日期 2026-07-10

华为技术有限公司

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。

安全声明

漏洞处理流程

华为公司对产品漏洞管理的规定以“漏洞处理流程”为准，该流程的详细内容请参见如下网址：
https://www.huawei.com/cn/psirt/vul-response-process
如企业客户须获取漏洞信息，请参见如下网址：
https://securitybulletin.huawei.com/enterprise/cn/security-advisory

目 录

1 服务咨询类

什么是 AI DataLake?

智能数据湖（AI DataLake）是华为云构建的企业级多模态智能数据分析与管理平台，为您提供构建AI时代数据基础设施的完整能力。提供**多模数据引擎Aura**、**AI计算引擎Ray**、**批处理引擎Spark**和**流处理引擎Flink**四大核心计算引擎，聚焦多模数据处理、异构算力混合调度，开放湖仓处理，构建新一代多模湖仓架构，促进Data+AI协同创新。

AI DataLake以工作空间为载体，采用数据、资源、引擎三层解耦架构实现灵活高效的企业级数据管理：

- **数据与引擎解耦：**所有引擎共享同一份数据，通过多模数据管理平台统一治理数据资产，避免数据孤岛和数据冗余，确保数据一致性。
- **资源与引擎解耦：**支持多引擎共用资源池，细粒度管理通算、智算资源，通过资源的灵活配置，提升资源利用率。
- **资源与数据解耦：**数据跨智算与通算资源流通，灵活支撑交互式分析、多模分析、批处理、实时计算、AI计算等多样化数据业务场景。

了解更多AI DataLake产品介绍，请参见**什么是AI DataLake**。

AI DataLake 提供了哪些计算引擎?

AI DataLake提供了多种计算引擎，满足不同业务场景的需求，您可根据实际负载灵活选择。

- 多模数据引擎Aura - 专为多模态数据处理设计。
- AI计算引擎Ray - 专注于AI计算处理。
- 批处理引擎Spark - 大规模批量数据处理。
- 流处理引擎Flink - 高吞吐实时流计算。

更多信息，请参见**计算引擎**。

2 工作空间类

什么是工作空间?

使用AI DataLake进行数据分析场景，管理员首先需要考虑的就是项目隔离、权限隔离与管理、资源分配问题。AI DataLake的工作空间正是为解决这些问题而设计，从系统层面为管理者提供对使用AI DataLake的用户（成员）权限、资源、底层计算引擎配置的管理能力。它通过环境隔离、资源绑定的基本逻辑，为您提供一个独立的开发环境，并基于此环境创建计算资源、选择引擎、配置端点、开发作业，从而实现高效有序、安全隔离的数据开发。

工作空间作为独立的数据开发和管理单元，提供了以下核心功能：

- 每个工作空间拥有独立的计算资源，不同工作空间之间资源完全隔离，互不影响。所有数据分析任务均在工作空间内完成。
- 一个工作空间绑定一个LakeFormation实例，使用LakeFormation管理元数据和数据访问权限。
- 一个工作空间即可满足不同的数据处理需求，在工作空间下创建计算资源，并根据业务需求选择引擎创建端点，开发作业，每个空间可以新建不同的端点，用于连接计算引擎和计算资源，且支持专属预留资源和按需弹性资源。
- 工作空间提供轻量级的作业开发环境，支持数据开发和调试。
- 通过基于IAM的细粒度权限控制不同的IAM用户对空间的访问权限。
- 不同工作空间之间的数据、权限、作业相互隔离，保障项目数据安全。

您只有在加入工作空间并被分配权限后，才可具备AI DataLake的相关操作权限。了解更多AI DataLake工作空间使用约束与限制，请参见**工作空间约束限制**。

工作空间可以迁移吗?

不支持。

创建工作空间时请您根据业务场景切换到对应的区域后再创建工作空间，一旦创建后暂不支持迁移工作空间至其他区域。

选择区域的建议：区域是云服务器的物理数据中心所在的位置，区域不同即云服务器物理数据中心距离用户的物理距离不同，网络延迟不同。为了降低访问时延、提高访问速度，请就近选择靠近您业务的区域。

3 计算资源池类

AI DataLake 计算资源有哪些使用方式?

AI DataLake提供的计算资源分为预留资源池和弹性资源两种类型，满足不同业务场景的算力需求。使用这两种计算资源有三种方式：

- **预留资源：**仅使用预留资源，您可以提前购买实例资源，资源类型覆盖CPU、NPU、GPU三种类型的算力，您可根据业务负载特点选择合适的计算资源组合。

  预留资源池为核心业务提供独占资源保障，性能稳定可预测，适用于长期稳定运行的批处理任务、AI计算场景。
- **按需弹性**：仅使用按需弹性资源，无需提前购买，按实际使用量计费，自动响应业务需求。
- **混合模式**：如果在配置端点时您选择的资源模式是“混合模式”，那么系统将为您开通弹性资源，在预留资源不足时自动调度弹性资源补充资源。既保障了业务平稳运行，又兼顾了资源的成本优化。

不同端点类型的计算资源使用模式分别是什么?

表 3-1 不同端点类型的计算资源使用模式

引擎类型	端点类型	资源使用模式	说明
Aura	• AuraJob • AuraJob V2	预留资源	使用预留资源池，资源独享。适合负载稳定、持续运行的业务场景，如生产项目、关键任务。
		混合模式	使用预留资源池和弹性资源，保障核心业务的同时，可应对突发负载。适用于日常有稳定负载，有突发的波峰波谷的场景。
Ray	RayCluster	预留资源	使用预留资源池，资源独享。适合负载稳定、持续运行的业务场景，如生产项目、关键任务。
	RayJob	预留资源	使用预留资源池，资源独享。适合负载稳定、持续运行的业务场景，如生产项目、关键任务。

引擎类型	端点类型	资源使用模式	说明
		混合模式	使用预留资源池和弹性资源，保障核心业务的同时，可应对突发负载。适用于日常有稳定负载，有突发的波峰波谷的场景。
		按需弹性	按作业实际运行时长计费。适合短期或偶发性的业务需求，如开发测试或临时任务。
Spark	• SparkSQL • SparkJob	预留资源	使用预留资源池，资源独享。适合负载稳定、持续运行的业务场景，如生产项目、关键任务。
		混合模式	使用预留资源池和弹性资源，保障核心业务的同时，可应对突发负载。适用于日常有稳定负载，有突发的波峰波谷的场景。
		按需弹性	按作业实际运行时长计费。适合短期或偶发性的业务需求，如开发测试或临时任务。

了解更多，请参见**计算资源池概述**。

4 端点类

什么是端点?

端点提供访问AI DataLake服务的入口，通过端点可以连接计算引擎与计算资源、 进行数据开发与查询、配置端点使用资源的Min/Max控制资源弹性范围。

在AI DataLake服务中，端点是各类计算引擎对外暴露的唯一标准化访问入口，端点的本质是“计算引擎”和“计算资源”的绑定，使得作业可以访问AI DataLake的引擎能力并使用AI DataLake的计算资源。

您可以把端点比喻为AI DataLake的“业务窗口”，您可以按需设置不同的窗口，这些窗口对应不同的业务需求，每个“窗口”都绑定了对应的“引擎”和“计算资源”。

端点的优势：

- 自动绑定引擎和计算资源池：创建端点时，AI DataLake自动将引擎和计算资源池绑定，在后续提交作业时无需再手动配置作业与引擎和资源的映射关系。
- 灵活配置分配给端点的计算资源：您可以在创建端点时根据业务需求配置所需的资源池的计算规格，即资源池分配对当前端点的**保障配额（min）**、**最大配额（max）**。

公测期间开放了哪些端点类型?

表 4-1 公测期间开放的端点类型和资源使用模式

功能	端点类型	端点资源使用模式
多模数据引擎Aura	• AuraJob • AuraJobV2	• 预留资源 • 混合模式
AI计算引擎Ray	RayCluster	预留资源
	RayJob	• 预留资源 • 混合模式 • 按需弹性

功能	端点类型	端点资源使用模式
批处理引擎Spark	• SparkSQL • SparkJob	• 预留资源 • 混合模式 • 按需弹性

5 计算引擎类

5.1 在 DataArts Studio 上执行 Aura 作业时报错，如何解决?

问题现象

在DataArts Studio执行作业时，报错为{"error_code":"common.01010003","error_msg":"No permissions for any one of the roles: [te_admin],or action DataArtsFabric.job.run"}。

根因分析

执行作业的用户未获得运行权限，或授权范围未包含所有资源。

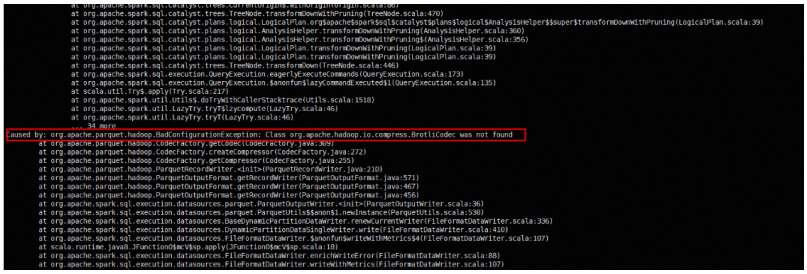
解决方案

1. 登录**IAM管理控制台**。
2. 单击右上方登录的用户名，在下拉列表中选择“统一身份认证”。
3. 在“用户”页面，搜索当前登录的用户名，并单击用户名。
4. 在“所属用户组”页签下，单击用户组名称。
5. 在“授权记录”页签下，请检查是否在授权时选择了运行作业的权限且授权所有资源。
   a. 在“权限”列查看是否有“DataArtsFabricConsoleFullPolicy”策略，或者单击自定义策略名称，在Action中查看是否包含“DataArtsFabric.job.run”。
   b. 在“项目[所属区域]”列，查看策略是否为“所有资源 [包含未来新增项目]”。
   c. 如果在授权时没有选择运行作业的权限且没有授权所有资源，则推荐直接授权“DataArtsFabricConsoleFullPolicy”策略。具体操作如下：
      i. 在“授权记录”页签下，单击“授权”。
      ii. 在“授权”页面，搜索“DataArtsFabricConsoleFullPolicy”，并选中该策略。
      iii. 单击“下一步”，然后授权范围方案选择“所有资源”后，单击“确定”。

5.2 Spark 中 Parquet 的 brotli 压缩格式无法使用怎么办?

问题现象

代码里显示支持brotli压缩格式，但实际使用Spark SQL类型的任务时，发现不支持。



根因分析

Parquet依赖声明中引用了org.apache.hadoop.io.compress.BrotliCodec，但该类在Hadoop中不存在。

解决方案

- 方案1：使用ORC格式时，其本身提供了BrotliCodec的支持。
- 方案2：使用Spark Job任务，并需引入包含org.apache.hadoop.io.compress.BrotliCodec实现的第三方包。

6 作业开发类

作业有哪些开发方式?

AI DataLake提供了多种作业开发方式，包括SQL编辑器、Notebook、DataArts Studio提交作业。

- **SQL编辑器（公测期间，尚未开放）：** 适用于SQL查询场景，执行结果可以在线预览，并支持下载至OBS。SQL编辑器适用于数据分析师日常查数据或简单的数据分析场景。
- **Notebook（公测期间，尚未开放）：** 基于JupyterLab，支持Python、SQL、Markdown混合编辑。适合数据科学家与需要富交互的场景。
- **DataArts Studio数据开发：** 基于可视化的作业编排流程，适用于生产环境的批量作业、复杂ETL开发场景，是企业级数据治理与自动化流程的主要开发工具。更多信息，请参见使用**DataArts Studio开发AI DataLake作业**。

