ALIBABA CLOUD

阿里云

数据湖分析

产品简介

文档版本：20201110

阿里云

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。

2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。

3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。

4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。

5. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含“阿里云”、“Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。

6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

格式	说明	样例
危险	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	危险 重置操作将丢失用户配置数据。
警告	该类警示信息可能会导致系统重大变更甚至故障，或者导致人身伤害等结果。	警告 重启操作将导致业务中断，恢复业务时间约十分钟。
注意	用于警示信息、补充说明等，是用户必须了解的内容。	注意 权重设置为0，该服务器不会再接受新请求。
说明	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	说明 您也可以通过按Ctrl+A选中全部文件。
>	多级菜单递进。	单击**设置**> **网络**> **设置网络类型**。
粗体	表示按键、菜单、页面名称等UI元素。	在**结果确认**页面，单击**确定**。
Courier字体	命令或代码。	执行 `cd /d C:/window` 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	`bae log list --instanceid` *Instance_ID*
[] 或者 [a|b]	表示可选项，至多选择一个。	`ipconfig [-all|-t]`
{} 或者 {a|b}	表示必选项，至多选择一个。	`switch {active|stand}`

目录

1.什么是云原生数据湖分析

云原生数据湖分析（Data Lake Analytics，DLA）是无服务器化（Serverless）的云原生数据湖分析服务，支持按需与保留资源使用，打造最具性价比的云原生数据湖分析平台。提供一站式的云原生数据湖分析与计算服务，支持 ETL、机器学习、流、交互式分析，可以与 OSS、数据库等多种数据源搭配使用。

- 数据湖管理，帮助客户解决高效构建安全的数据湖，功能包括元数据的管理、一键入湖、元数据爬取、增量湖等功能。
- 数据源矩阵，详情请参见数据源与功能的矩阵。

数据源	Serverless SQL(兼容Presto版本)	Serverless Spark
OSS	支持	支持
RDS	支持	支持
PolarDB	支持	支持
Lindorm	待支持	支持
Hbase	待支持	支持
MongoDB	支持	待支持
Tablestore	支持	支持
AnalyticDB MySQL	支持	支持
AnalyticDB MySQL	支持	支持
AnalyticDB PostgreSQL	支持	支持
MaxCompute	支持	支持
Elasticsearch	支持	支持
ECS自建Druid数据库数据	支持	支持

- 任务调度
  - DMS
  - DataWorks

何时使用DLA

DLA主要围绕数据湖存储OSS提供一站式的云原生数据湖分析与计算方案，如果您有如下的痛点可以使用DLA：

- 寻求一站式的数据湖解决方案，从数据高效入湖、数据的ETL、机器学习、交互式分析。DLA提供了数据湖构建、Presto&Spark引擎。
- 寻求安全的数据处理解决方案。DLA所有的库表及存储的数据都有一整套安全的方案，避免数据被误用。
- 寻求低成本的数据处理方案。DLA方案是完全Serverless的解决方案，是阿里云提供的云原生的数据处理方案。

- 从之前Hadoop体系过渡到数据湖方案。DLA提供与Hadoop体系兼容的过渡方案。

为什么同时支持Serverless SQL与Serverless Spark?

DLA Serverless SQL是在开源Apache Presto基础上研发，完全由内存完成计算工作，具备高性能、交互式的分析体验，秒级可返回；DLA Serverless Spark是在开源Apache Spark基础上研发，兼容Apache Spark所有的API。

以下场景推荐您使用DLA Serverless Spark:

- 需要自定义Code，SQL很难表达的，例如编写Java、Scala、Python或者SQL带条件的。
- 需要大规模的清洗，例如1天清洗1次OSS 1TB~1PB的数据。
- 需要算法支持，DLA Spark支持完整的Spark算法库。
- 需要支持Streaming。

基本概念

- DLA是Region化的，不同Region账号体系、元数据体系是完全隔离的；
- DLA提供 扫描量版本与CU版本的 计费模式，其中扫描量版本 支持 SQL，CU版本支持SQL与Spark。
  - 扫描量版本：用户提交SQL查询，按照SQL实际扫描的数据量计费。
  - CU版本：按照实际的资源使用量计费，1CU为1CPU4GB。
- 虚拟集群VC（Virtual Cluster）是对底层资源的抽象，可以针对VC配置网络打通、及一些基本的信息。
  - 需要CU版本计费时需要构建VC集群。
  - 扫描量版本的资源是平台构建一批VC，用户无需直接为资源付费，资源会按照扫描量转化为实际的费用，主要是为了满足用户无需持有资源且能得到立即响应的体验。
- 账号：分为DLA账号、RAM账号，DLA账号与RAM账号可以关联。
- 元数据：支持库、表、列、视图等，每个库只能对应一种数据源，元数据是 SQL引擎、Spark引擎均可安全访问。
- 权限：支持库、表级别的细粒度授权。
- 语法标准：
  - DDL：参考Hive标准。
  - DCL：MySQL数据库标准
  - DML：SQL是兼容Presto标准，Spark SQL是Spark的标准。

2.典型场景

- 阶段1：构建数据湖（需要关联 DLA 的Meta）
  - 数据库 入湖：可以通过DLA提供的 一键建湖 的能力，客户也可以通过其他手段建湖；
  - 文件上传：数据上传后，DLA元数据爬取功能可自动爬取构建好元数据体系；
  - 流式数据：DLA提供了DLA SparkStreaming来对接，并写入OSS之中，提供Hudi格式的支持，自动关联好DLA的Meta；
- 阶段2：DLA Serverless Spark提供强大的数据清洗能力，把ODS层的原始数据清洗为结构化的DW数据。
- 阶段3：DLA可以提供SQL交互式分析或者 通过DLA Spark做进一步的计算等。

联邦分析分析： 同时连接多个数据源做数据的分析

- 联邦分析与轻量级清洗方案：可对接数十种数据源对各种数据源进行查询与轻量数据级清洗。
- 可以支持扫描量版本 与 CU版本混合使用。

3.产品优势

云原生数据湖分析（Data Lake Analytics，DLA）是一种架构的云原生数据湖分析服务，对比常规的分析方案具有如下优势：

对比类目	自建Hadoop系统	阿里云 DLA + OSS方案
产品体系	复杂、组件较多	一体化、端到端（入湖=>管理=>ETL =>分析查询），产品体验好；组件精耕细作 Presto、Spark；
持有成本	高（弹性弱，一直持有固定集群）	低（按照扫描量计费 或者CU收费，按照实际使用计费）
学习与运维成本	高（需要较长时间 搭建、配置、运维、学习）	低（即开即用、零运维成本）
弹性	无	云原生、弹性强、一分钟可弹300节点参与计算
安全、多租户	基于 Kerberos&Ranger，较为复杂	支持数据库模式库、表授权模式，多租户
功能	开源功能，缺乏云连接器的支持，云内部系统对接与优化	针对阿里云OSS & OTS &ADB 等数据源深度优化