ALIBABA CLOUD

阿里云

数据湖分析
快速入门

文档版本：20211117

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.使用流程

云原生数据湖分析DLA（Data Lake Analytics）是无服务器（Serverless）化的云上交互式查询分析服务，支持通过Presto和Spark引擎分析多种数据源中的数据。快速入门旨在介绍如何开通DLA、构建数据湖、调用Presto和Spark引擎进行数据分析与计算，帮助您掌握DLA的基本使用流程。

如果您是首次使用云原生数据湖分析DLA的用户，我们建议您先阅读以下部分：

- 产品简介-本内容概述了云原生数据湖分析DLA的产品概念、产品优势及应用场景等内容。
- 产品定价-本内容介绍了云原生数据湖分析DLA的产品定价、计费方式等信息。

云原生数据湖分析DLA入门指南（本指南）-本指南提供了有关使用云原生数据湖分析DLA的基本流程。操作流程概览如下：

1. 开通云原生数据湖分析服务
2. （可选）创建虚拟集群：CU版适用于查询频率高、查询数据量较大的场景，同时也能够给您使用DLA的费用预算带来一定的确定性。推荐您使用CU版本来进行数据分析与计算。

   > **说明** 如果您使用系统默认的扫描量版本来进行数据分析与计算，则无需创建虚拟集群，可以跳过此步骤。CU版与扫描量版的具体差异，请参见扫描量版本与CU版本的差异。

3. 构建数据湖：您可以通过以下多种方式来构建数据湖。例如：
   - 手工上传文件到OSS，然后通过元数据爬取功能创建表来构建数据湖。具体操作请参见上传文件和OSS数据源。
   - 通过其他产品投递文件到OSS（如通过操作审计控制台投递日志文件到OSS），然后通过元数据爬取功能创建表来构建数据湖。具体操作请参见创建单账号跟踪和OSS数据源。
   - 通过一键建仓和多仓合并建仓（仅支持数据库）、实时数据湖（支持数据库和消息日志）功能来构建数据湖。具体操作请参见一键建仓多库合并建仓和实时数据湖。
   - 连接数据源：您也可以通过DLA直接连接其他数据源或者OSS来进行数据分析与计算。具体操作请参见Presto引擎连接数据源和Spark引擎连接数据源。
4. 数据分析与计算：您可以调用Serverless Presto和Spark引擎来进行数据分析与计算。具体请参见Serverless Presto和Serverless Spark。
5. 数据应用：您可以通过DataWorks和DMS来调度DLA Presto和DLA Spark任务，也可以将OSS数据的查询分析结果以BI报表形式进行展示。具体请参见快速搭建Quick BI可视化报表。

2.通过元信息发现功能查询并分析OSS数据

本文档将以通过DLA的元信息发现功能查询并分析OSS数据为例，帮助您快速掌握DLA的基本使用流程。

前提条件

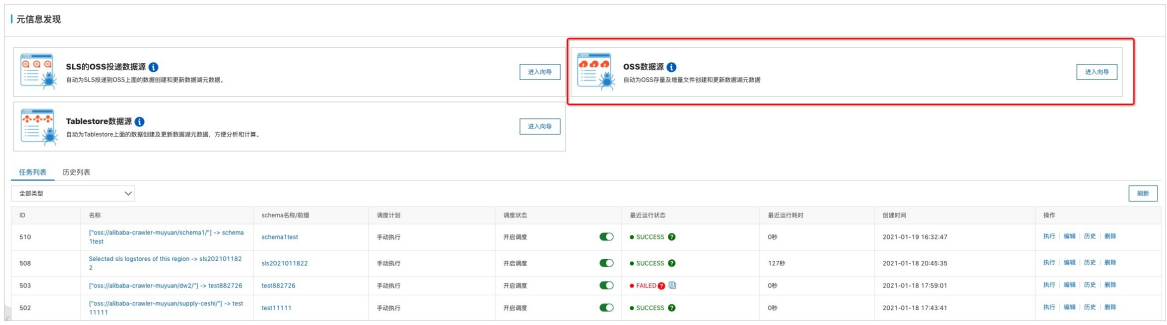
已注册阿里云账号并完成实名认证。

> **说明** 如果您还没有创建阿里云账号，系统会在您开通云原生数据湖分析服务时提示您注册账号。

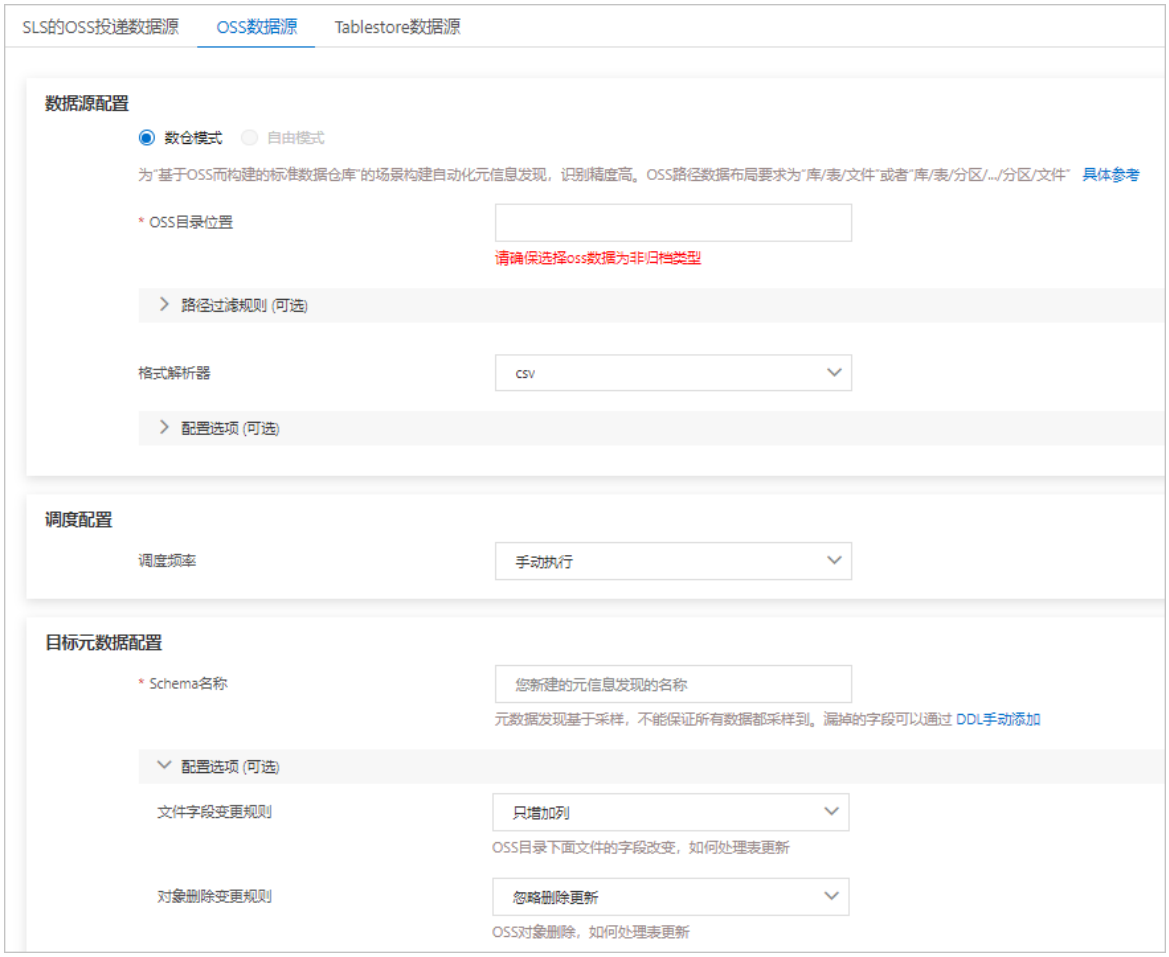
操作步骤

1. 开通云原生数据湖分析服务。
2. 登录OSS管理控制台，上传文件到OSS，具体操作请参见上传文件。

   例如将supplier_with_header.csv文件上传到OSS的指定目录*oss://alibaba-crawler/schema1/supply_ceshi/*。
3. 登录数据湖分析管理控制台，在左侧导航栏单击**数据湖管理 > 元信息发现**。
4. 在**元信息发现**页面的**OSS数据源**区域，单击**进入向导**。



5. 在**OSS数据源**页签的**数据源配置**、**调度配置**、**目标元数据配置**区域，根据实际需要进行参数配置。



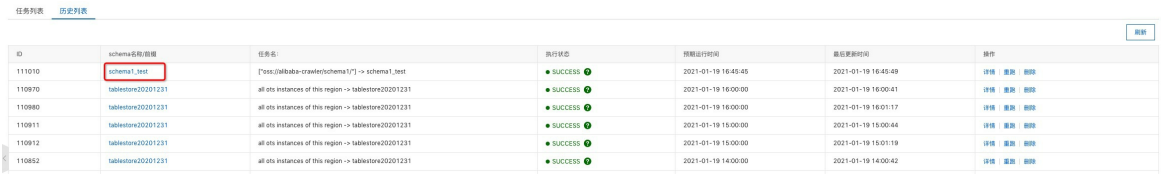
参数配置说明如下表所示：

参数	说明
数仓模式和自由模式	您可以选择数仓模式或自由模式： ○ 数仓模式：为“基于OSS而构建的标准数据仓库”的场景构建自动化元信息发现，识别精度高。OSS路径数据布局要求为“库/表/文件”或者“库/表/分区/.../分区/文件”。 ○ 自由模式：为“探索OSS上的数据进行分析”的场景构建自动化元信息发现。对OSS数据布局没有要求，可能会产生差异化的表。
OSS目录位置	文件在OSS中的存储地址，以/结尾。系统会根据您选择的文件夹路径，自动设置OSS路径。 **说明** 系统会自动拉取与DLA同地域的OSS Bucket，您可以根据业务需要从下拉列表中选择Bucket。选择Bucket后，系统会自动列出该Bucket下所有的Object和文件；选中目标Object和文件后，系统会自动将其添加到右侧的OSS路径处。
格式解析器	默认自动解析，即按照顺序调用所有内置解析器，也可指定特定文件类型的格式解析器，比如json、parquet、avro、orc、csv。
调度频率	您可以根据需要定期计划运行元信息发现任务。

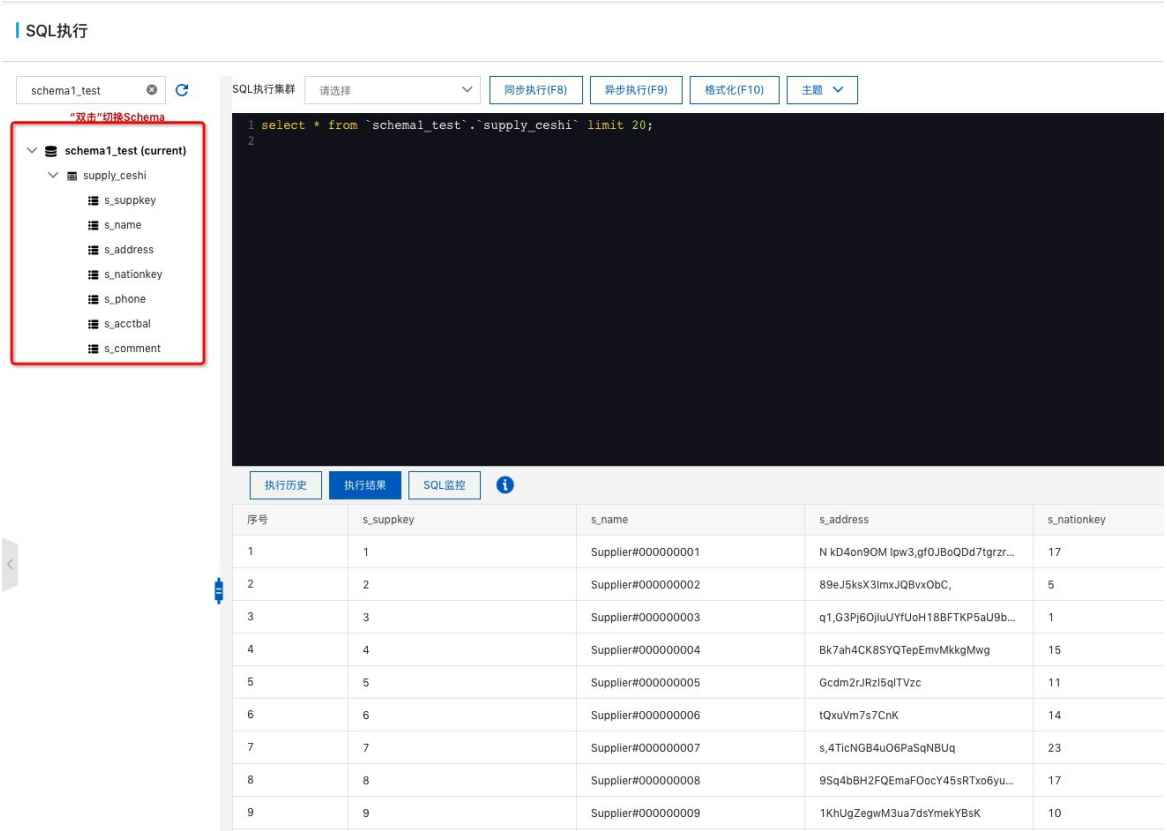
参数	说明
Schema名称	设置Schema名称，即映射到DLA中的数据库名称（默认每个发现任务会新创建一个独立的Schema）。
配置选项	高级自定义设置项，如字段分隔符、引用标识、表头模式、允许单个列字段等。

6. 配置完成后，单击**创建**。

   元信息发现任务创建成功后，在**任务列表**中您将能看到创建成功的任务信息。元信息发现任务将根据您设置的**调度频率**，需要您手动执行或者定期自动调度该任务。

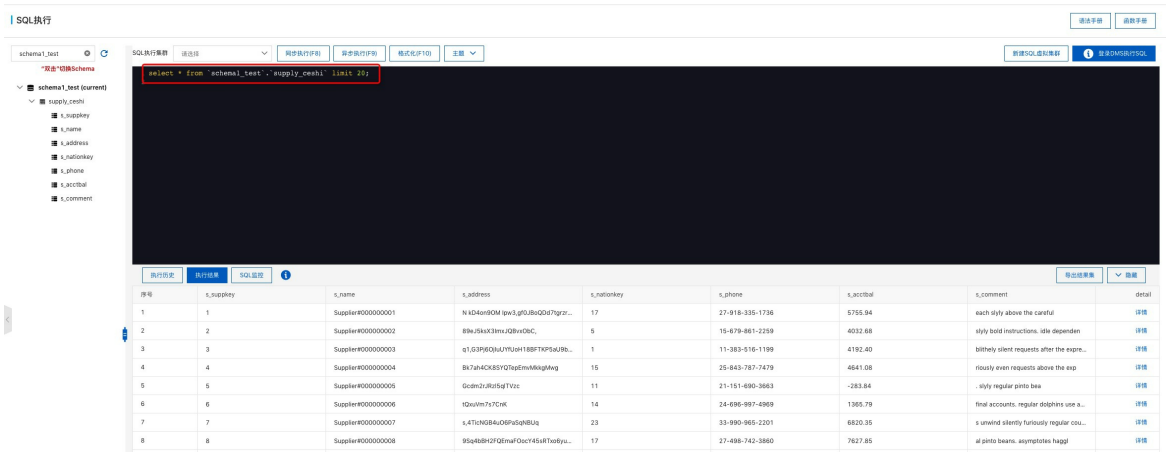


   元信息发现任务执行成功后，单击**schema名称/前缀**列下面的数据库名称链接（如单击**alibaba**），跳转到**Serverless Presto > SQL执行**页面。您可以看到DLA自动发现创建成功的库、表、列信息。



7. 在**Serverless Presto > SQL执行**页面编写SQL语句，单击**同步执行**或者**异步执行**，执行SQL语句。

   例如在**schema1_test**下执行 `select * from `schema1_test`.`supply_ceshi` limit 20;` 。



您可以在**执行结果**中，查看DLA从OSS目录*oss://alibaba-crawler/schema1/supply_ceshi/*下的*supplier_with_header.csv*文件中自动发现的数据信息。

3.DLA Presto CU版本快速入门

本文主要教您如何快速上手阿里云云原生数据湖分析DLA Presto CU版本。

操作步骤

1. **创建虚拟集群**

   与Serverless Presto扫描版不同，在CU版本下执行SQL前，必须要创建一个虚拟集群。具体请参见虚拟集群管理。

   > **说明** 创建虚拟集群时，选项**引擎**选择Presto。

2. **配置数据源网络**

   > **说明** 如果需要连接您VPC内的数据源（如VPC内的RDS、AnalyticDB等），您需要配置数据源网络，如不需要连接，忽略此步骤即可。

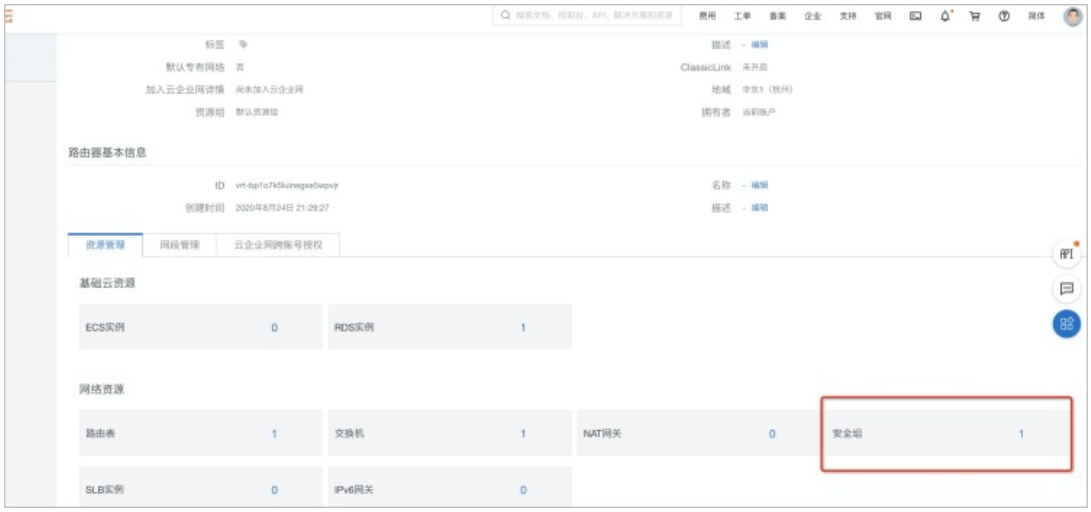
   i. 在配置数据源网络前，您需要授予DLA账户访问您VPC相关API的权限，详细操作步骤请参见配置数据源网络。

   ii. 在虚拟集群管理页面单击**详情**，进入集群详情页面。

iii. 单击**新增数据源网络**，选择您想要连接的数据源对应的**虚拟网络**、**交换机Id**和**安全组Id**。

- 交换机ID可以在您VPC内数据源的基本信息页面获取，以RDS为例。

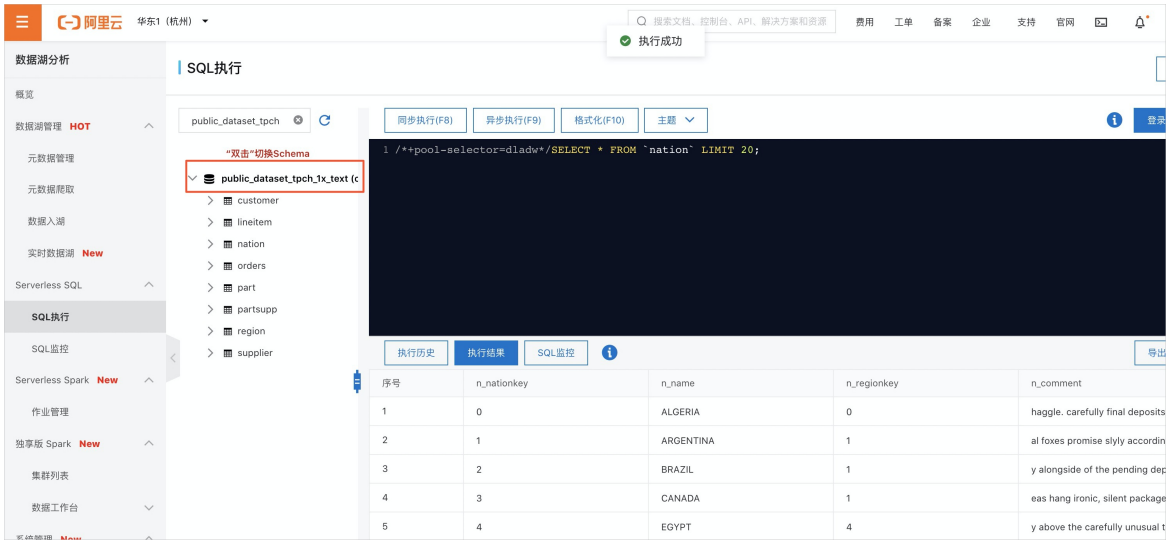
- 安全组ID可以登录VPC控制台获取。



> **说明** 这里安全组可能会有多个，选择一个可以访问您数据源的安全组即可。

3. **执行SQL**

在CU版本下，当您创建好虚拟集群，首次开始执行SQL时，您会发现系统已经默认自动生成一个公共数据集。



在左边列表中选中public_dataset_xxx这个库，并且执行SQL语句，例如：

```
/*+cluster=dladw*/SELECT * FROM `nation` LIMIT 20;
```

> 说明 cluster是您之前创建的虚拟集群实例名称。

更多关于DLA SQL用法的详细操作，请参见常用SQL。

4.DLA Spark快速入门

熟悉Spark的开发者都了解SparkPi，它相当于Spark引擎的”Hello World!“。本文介绍如何在DLA控制台跑通SparkPi。

准备事项

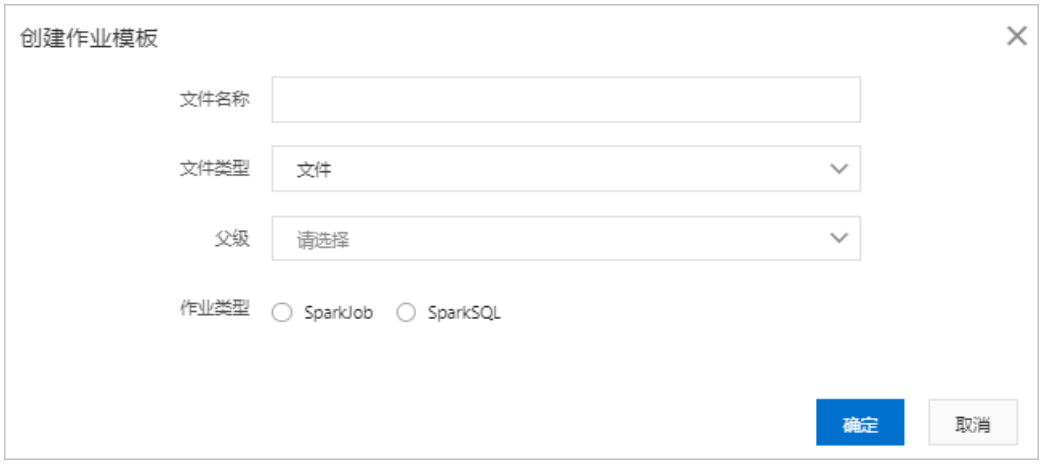
1. 您需要在提交作业之前首先创建虚拟集群，具体操作请参考创建虚拟集群。

   > **说明** 创建虚拟集群时注意选择引擎类型为Spark。

2. 如果您是子账号登录，需要配置子账号提交作业的权限，具体操作请参考快速配置子账号权限。

操作步骤

1. 登录云原生数据湖分析管理控制台。
2. 在**概览**页面的左上角，选择虚拟集群所在地域。
3. 单击左侧导航栏的**Serverless Spark > 作业管理**。
4. 在**作业编辑**页面，单击**创建作业模板**，填写以下作业信息：



5. 新创建的作业中包含了SparkPi作业的默认配置，在**作业编辑**页面，单击**执行**即可。

   > **说明** 关于作业提交的详细说明，请参见创建和执行Spark作业。