ALIBABA CLOUD

阿里云

Databricks 数据洞察

管理元数据

文档版本：20211230

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.独立RDS元数据库

本文介绍如何配置独立的阿里云RDS，作为Databricks数据洞察集群的元数据。

前提条件

已购买RDS，详情请参见创建RDS MySQL实例。

> 说明
>
> 本文以MySQL 5.7版本为例介绍。

元数据库准备

1. 创建hivemeta的数据库

   详情请参见创建数据库和账号中的创建数据库。

2. 创建用户并授权读写权限

3. 获取数据库内网地址

   i. 在实例详细页面，单击左侧导航栏中的数据库连接。

   ii. 在数据库连接页面，单击内网地址进行复制。

创建Databricks数据洞察集群

在创建集群的页面，配置以下参数，其他参数的配置请参见创建集群。

参数	描述
元数据选择	选择独立RDS MySQL
数据库连接	数据库连接填写格式为jdbc:mysql://rm-xxxxxx.mysql.rds.aliyuncs.com/<数据库名称>。 • rm-xxxxxx.mysql.rds.aliyuncs.com为hive元数据库所在RDS实例的内网地址。 • <数据库名称>为hive元数据库名称。
数据库用户名	填写hive元数据库中账号的用户名。
数据库密码	填写hive元数据库中账号的密码。

Metastore初始化

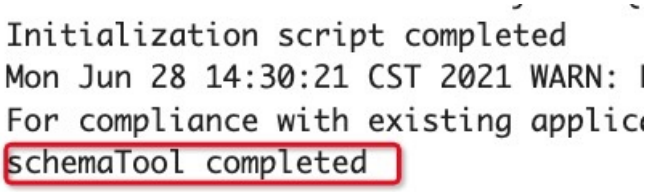
1. 连接RDS元数据库，首先需要打通Databricks集群与MySQL实例的VPC和VSwitch网络。详情参见RDS(MySQL数据源打通)。

2.配置安全组白名单，并访问Notebook，详情参见安全组白名单。

3.登录Notebook，执行Metastore初始化命令。

```
%sh
schematool -initSchema -dbType mysql
```

待初始化成功



> **说明**
>
> 在初始化之前，Hive的Hive MetaStore、HiveServer2和Spark的ThriftServer可能会出现异常，待初始化之后会恢复正常。

2.数据湖元数据管理

Dat abricks 数据洞察DBR 7.3, Spark 3.0.1, Scala 2.12及之后版本，在创建集群选择元数据类型时支持数据湖元数据作为Hive数据库。数据湖元数据是服务化高可用并且可扩展的元数据库，您无需额外购买独立的元数据库，就可以实现多个引擎计算，例如同时使用Dat abricks 数据洞察和E-MapReduce。多个Dat abricks 数据洞察集群可以共享统一元数据库。

前提条件

- 已在数据湖构建（Dat a Lake Formation）控制台开通数据湖构建。

> **注意**
>
> 数据湖元数据产品目前只支持华北2（北京）、华东2（上海）和华东1（杭州）三个地域。

- 进入RAM访问 控制台给AliyunDDIAccessingOSSRole角色添加一个AliyunDDIAccessingDLFRolePolicy自定义策略，策略详情如下：

```
{
    "Version": "1",
    "Statement": [
        {
            "Action": [
                "dlf:BatchCreatePartitions",
                "dlf:BatchCreateTables",
                "dlf:BatchDeletePartitions",
                "dlf:BatchDeleteTables",
                "dlf:BatchGetPartitions",
                "dlf:BatchGetTables",
                "dlf:BatchUpdatePartitions",
                "dlf:BatchUpdateTables",
                "dlf:CreateDatabase",
                "dlf:CreateFunction",
                "dlf:CreatePartition",
                "dlf:CreateTable",
                "dlf:DeleteDatabase",
                "dlf:DeleteFunction",
                "dlf:DeletePartition",
                "dlf:DeleteTable",
                "dlf:GetDatabase",
                "dlf:GetFunction",
                "dlf:GetPartition",
                "dlf:GetTable",
                "dlf:ListCatalogs",
                "dlf:ListDatabases",
                "dlf:ListFunctionNames",
                "dlf:ListFunctions",
                "dlf:ListPartitionNames",
                "dlf:ListPartitions",
                "dlf:ListPartitionsByExpr",
                "dlf:ListPartitionsByFilter",
                "dlf:ListTableNames",
                "dlf:ListTables",
                "dlf:RenamePartition",
                "dlf:RenameTable",
                "dlf:UpdateDatabase",
                "dlf:UpdateFunction",
                "dlf:UpdateTable",
                "dlf:UpdateTableColumnStatistics",
                "dlf:GetTableColumnStatistics",
                "dlf:DeleteTableColumnStatistics",
                "dlf:UpdatePartitionColumnStatistics",
                "dlf:GetPartitionColumnStatistics",
                "dlf:DeletePartitionColumnStatistics",
                "dlf:BatchGetPartitionColumnStatistics"
            ],
            "Resource": "*",
            "Effect": "Allow"
        }
    ]
}
```

背景信息

数据湖元数据已适配Dat abricks 数据洞察的Spark SQL。

适用场景

数据湖元数据具有高可用和易维护的特点，因此适合在如下场景下使用数据湖元数据：

- Dat abricks 数据洞察集群的生产环境，您无需维护独立的元数据库。
- 横向使用多种大数据计算引擎，例如Dat abricks 数据洞察、MaxCompute、EMR等，元数据可以集中管理。
- 多个Dat abricks 数据洞察集群，可以统一管理元数据。

创建集群

创建Dat abricks 数据洞察集群时，如图**元数据选择**为**数据湖元数据**方式，创建详情请参见创建集群。

> **说明**
>
> 如果需要迁移数据库的元数据信息，请提交工单处理。