数据湖计算

快速入门

产品文档

腾讯云

文档目录

快速入门
一分钟入门 DLC 数据分析

最近更新时间：2022-09-20 15:02:34

使用数据湖 DLC，您仅需一分钟即可完成对象存储 COS 上的数据分析查询，目前支持 CSV、ORC、PARQUET、JSON、ARVO、文本文件等多个格式。

前置准备

设置必要 DLC 内部权限

> 说明：
> 如果用户已经有权限，或者为主账户管理员，可忽略此步骤。

若您是首次登录的子账号，除了必要的 CAM 授权，还需要请任意 DLC 管理员或主账号管理员在 DLC 控制台左侧**权限管理**菜单，为您授予必要的 DLC 权限（详细权限说明参见 DLC 权限概述：

1. 库表权限：可授予对应的 catalog、database、table，view 等读写操作权限。
2. 引擎权限：可授予计算引擎的使用、监控、修改等权限。

> 说明：
> 系统会默认为每个用户开通基于 presto 内核的共享 public-engine，方便您可以快速试用，无需先购买独享集群。

详细权限授予步骤参见 子账号权限管理。

分析步骤

步骤1：创建数据库

如果您对 SQL 语句熟悉，可直接在查询中编写 create database 语句，跳过创建向导。

1. 登录 数据湖计算 DLC 控制台，选择**服务地域**。
2. 左侧导航菜单进入**数据探索**。

3. 选择**库表**，单击“+”，选择**创建数据库**进行数据库新建。如下图所示：

4. 右上角选择执行引擎后，执行生成的 create database 语句，完成建库。

步骤2：创建外表

如果您对 SQL 语句熟悉，可直接在查询中编写 create table 语句，跳过创建向导。

1. 登录 数据湖计算 DLC 控制台，选择服务地域。
2. 左侧导航菜单进入**数据探索**。
3. 选择库表，选中当前创建的表后，右键单击，选择**创建外表向导**。

> 说明：
> 外表一般指数据文件放到您自己账号下的 COS 桶，DLC 可以直接建立外表进行分析，无需额外加载数据。基于外表的特性，例如在执行 drop table 等动作时，DLC 并不会删除您的原始数据，只会删除 table 的元信息。

4. 按照向导生成创表语句，按照**基本信息 > 数据格式 > 编辑列 > 编辑分区**，完成各个步骤。
   - step1：选择数据文件存放的 COS 路径（路径必须是 COS 桶下的目录，不能直接建立到 COS 桶），此处也提供快速上传文件到 COS 的快捷方式。操作需具备 COS 相关的权限。
   - step2: 选择数据文件的格式，**高级选项**中可选择自动推断格式，后端将解析文件格式，自动生成表的列信息，快速完成列信息推断。

> 说明：
> 结构推断为建表辅助工具，不能保证100%正确，仍需您进行复查核对字段名、类型是否符合预期，根据实际情况编辑修改为正确的信息。

- step3: 如果没有分区可以跳过此步骤，合理的分区可以帮助提升分析性能。详细分区信息可参见 查询分区表。

5. 单击**完成**，会生成 SQL 建表语句，选择数据引擎后执行生成的语句即可完成建表。

步骤3：执行 sql 分析

数据准备完备后，您就可以开始书写 SQL 分析语句，选择合适的计算引擎，开始数据分析。

示例

编写数据查询所有结果为 SUCCESS 记录的 SQL 语句，选择计算引擎后执行。

```
select * from `DataLakeCatalog`.`demo2`.`demo_audit_table` where _c5 = 'SUCCESS'
```

一分钟入门 DLC 权限管理

最近更新时间：2022-08-16 09:41:59

用户与工作组

DLC 通过对用户授权和绑定工作组授权两种方式来管理用户权限。

- 工作组：DLC 可以将一批用户绑定到工作组，并授予该组数据、引擎等资源权限，来批量管理用户权限，在同一个工作组的用户具有相同的权限。
- 用户：CAM 中的用户，包括子账号、协作者账号。

> 说明：
> 当用户被赋予的权限与所在工作组权限不同时，两者权限取并集。

操作步骤

1. DLC 左侧导航栏选择**权限管理**。

2. 创建工作组
   单击**工作组 > 添加工作组**来创建用户的工作组，创建工作组时可以选择用户进行绑定或创建一个空工作组。详细操作可参见 用户和用户组。

3. 工作组授权
   创建工作组后单击列表中的**授权**操作，为工作组添加权限，包括**数据权限**和**引擎权限**。

   i. 数据权限
   - 数据目录权限：包括在数据目录下创建数据库和创建数据目录两种权限。

- 数据库表权限：可授予库表级别的细粒度权限，包括对库，表，视图，函数的查询、编辑等权限。

ii. 引擎权限

选择数据引擎并授予使用、修改、删除等权限。

4. 创建用户

添加用户并绑定工作组：单击**用户 > 添加用户**，添加新用户。选择用户类型为”普通用户“后绑定工作组并获取该工作组的所有权限，选择用户类型为 **DLC 管理员**不需要绑定工作组。

5. 授权用户

在用户列表为用户单独进行授权，授权包含”数据权限“和”引擎权限“，同工作组权限。

更多详细操作参见 子账号权限管理。

一分钟入门分区表

最近更新时间：2022-08-16 09:41:59

DLC 分区表

用户可以将数据按照分区目录的方式进行存储，将不同特征的数据存放在不同的目录下，在进行数据探索时，通过 where 条件按照分区进行过滤，DLC 的数据扫描量将大幅减少，提高查询效率。

注意：

- 同一个表的分区应使用相同的数据类型及格式。
- DLC 内表采用隐藏式分区实现，可不用关注分区目录结构。

创建分区表

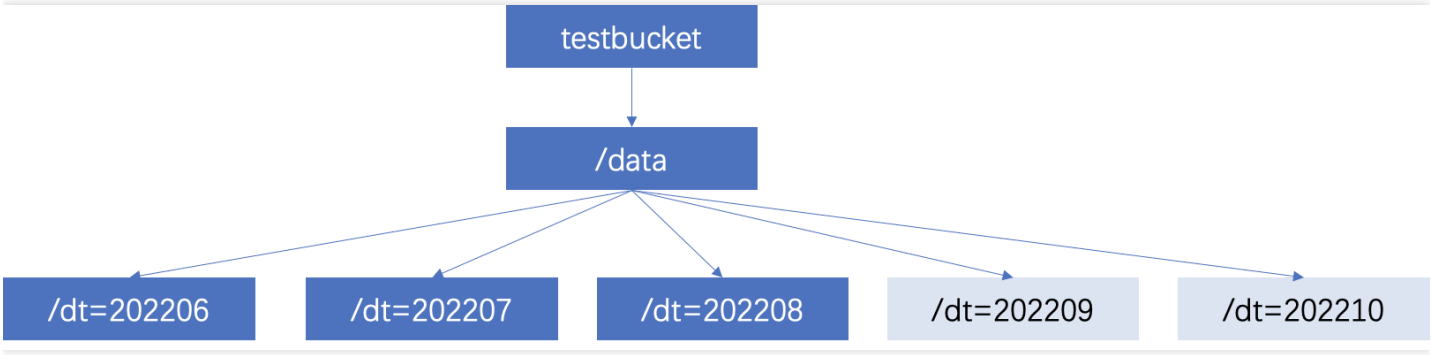
创建表格语句中通过 PARTITIONED BY 指定分区字段。
示例：创建分区表 test_part

```
CREATE EXTERNAL TABLE IF NOT EXISTS `DataLakeCatalog`.`test_a_db`.`test_part` (
`_c0` int,
`_c1` int,
`_c2` string,
`dt` string
) USING PARQUET PARTITIONED BY (dt) LOCATION 'cosn://testbucket/data/';
```

添加分区

通过 alter table add partition 添加分区

如果用户的数据分区目录为 Hive 的分区规则：分区列名=分区列值，可采用这种方式添加分区，目录组织方式如图。




```
ALTER TABLE `DataLakeCatalog`.`test_a_db`.`test_part` add PARTITION (dt = '202206
')
ALTER TABLE `DataLakeCatalog`.`test_a_db`.`test_part` add PARTITION (dt = '202207
')
ALTER TABLE `DataLakeCatalog`.`test_a_db`.`test_part` add PARTITION (dt = '202208
')
ALTER TABLE `DataLakeCatalog`.`test_a_db`.`test_part` add PARTITION (dt = '202209
')
ALTER TABLE `DataLakeCatalog`.`test_a_db`.`test_part` add PARTITION (dt = '202210
')
```

通过 alter table 指定 location 添加分区

如果用户的数据组织为普通的 cos 目录（非“分区列名=分区列值”组织方式），可以在 add partition 时指定目录。
SQL 参考：

```
ALTER TABLE `DataLakeCatalog`.`test_a_db`.`test_part` add PARTITION (dt = '20221
1') LOCATION='cosn://testbucket/data2/202211'
ALTER TABLE `DataLakeCatalog`.`test_a_db`.`test_part` add PARTITION (dt = '20221
2') LOCATION='cosn://testbucket/data2/202212'
```

使用 MSCK REPAIR 自动添加分区

使用 MSCK REPAIR TABLE 语句，扫描在建表时指定的数据目录。若存在新的分区目录，则系统会自动将这些分区添加到数据表的元数据信息中。
SQL参考：

```
MSCK REPAIR TABLE `DataLakeCatalog`.`test_a_db`.`test_part`
```

建议优先选择 alter table 的方式添加分区，如果采用 msck repair 自动添加分区，有如下约束条件：

- MSCK REPAIR TABLE 仅向数据表元数据添加分区，不会删除分区。
- 数据数据量较大时，不推荐使用 MSCK REPAIR TABLE 的方式，该方式会扫描全部数据量，可能会导致超时。

- 如果分区目录不为 Hive 的分区规则：分区列名=分区列值，不能采用 MSCK REPAIR TABLE 方式。

一分钟入门 UDF 函数

最近更新时间：2022-08-16 09:41:59

UDF 说明

用户可通过编写 UDF 函数，打包为 JAR 文件后，在数据湖计算定义为函数在查询分析中使用。目前数据湖计算 DLC 的 UDF 为 HIVE 格式，继承 org.apache.hadoop.hive.ql.exec.UDF，实现 evaluate 方法。
示例：简单数组 UDF 函数

```
public class MyDiff extends UDF {
public ArrayList<Integer> evaluate(ArrayList<Integer> input) {
ArrayList<Integer> result = new ArrayList<Integer>();
result.add(0, 0);
for (int i = 1; i < input.size(); i++) {
result.add(i, input.get(i) - input.get(i - 1));
}
return result;
}
}
```

pom 文件参考

```
<dependencies>
<dependency>
<groupId>org.slf4j</groupId>
<artifactId>slf4j-log4j12</artifactId>
<version>1.7.16</version>
<scope>test</scope>
</dependency>
<dependency>
<groupId>org.apache.hive</groupId>
<artifactId>hive-exec</artifactId>
<version>1.2.1</version>
</dependency>
</dependencies>
```

创建函数

若您了解 SQL 语法，可通过**数据探索**执行 CREATE FUNCTION语法完成函数创建，或通过可视化界面创建，流程如下：

> 说明：
> 数据湖计算 DLC 的数据管理页目前处于邀测阶段，如需免费体验可 提交工单 进行申请。

1. 登录 数据湖计算控制台 ，选择服务地域。
2. 通过左侧导航菜单进入**数据管理**，选择需要创建的函数的数据库。
3. 单击**函数**进入函数管理页面。
4. 单击**创建函数**进行创建。

UDF 的程序包支持本地上传或选择 COS 路径（需具备 COS 相关权限），示例为选择 cos 路径创建。
函数类名包含“包信息”及“函数的执行类名”。

函数使用

1. 登录 数据湖计算控制台，选择服务地域。
2. 通过左侧导航菜单进入数据探索，选择计算引擎后即可使用 SQL 调用函数。

跨源分析 EMR Hive 数据

最近更新时间：2022-08-16 09:41:59

数据湖计算 DLC 支持配置 EMR Hive 的数据源进行跨源联合分析。

使用前准备

- 获取 EMR Hive 地址。
- 使用具备创建数据目录权限的账号，详细权限请参见 DLC 权限概述。

创建 EMR Hive 数据源

1. 登录 数据湖计算 DLC 控制台，选择服务地域。
2. 通过左侧导航栏进入**数据探索**，单击库表栏的**+**按钮，选择**新建数据目录**。

3. 选择连接类型为 EMR Hive（HDFS），选择 EMR 的对应实例，VPC 信息将在实例选择后默认填充。**EMR Hive 支持 EMR 的版本：2.3.5，2.3.7，3.1.1，3.1.2**。

> 注意：
> 需具备 EMR Hive 实例的相关权限才可进行选择。

4. 选择运行集群，目前仅支持选择Presto的独享数据引擎，如无对应引擎可至数据引擎页进行数据引擎创建。购买流程请参见 购买独享数据引擎。

> 注意：
> 所选数据引擎网段不可与 EMR 实例网段相同，否则将导致网络冲突，无法进行数据查询分析。

5. 单击**确认**按钮即可完成数据目录创建。

查询 EMR Hive 数据

完成数据目录创建之后，即可在**数据探索页**的数据目录菜单进行数据目录切换。

此时您可通过 SQL 语句对该数据目录进行查询分析。
选择创建数据目录时绑定的数据引擎即可单击**运行**按钮，获得查询结果。

> 注意：
> 仅绑定的数据引擎可查询该数据目录，其他数据引擎将无法进行查询。如需变更绑定的引擎，可单击数据目录判的设置按钮就行编辑修改。