

阿里云

数据湖分析

AnalyticDB for MySQL 3.0

文档版本：20210630

## 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<br>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<br>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<br>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<br>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 <code>cd /d C:/window</code> 命令，进入Windows系统文件夹。  |
| 斜体   | 表示参数、变量。                           | <code>bae log list --instanceid</code><br><code>Instance_ID</code>  |
| [ ] 或者 [a b]   | 表示可选项，至多选择一个。                      | <code>ipconfig [-all -t]</code>   |
| { } 或者 {a b}   | 表示必选项，至多选择一个。                      | <code>switch {active stand}</code>  |

# 目录

|                                      |    |
|--------------------------------------|----|
| 1.读写AnalyticDB for MySQL 3.0数据       | 05 |
| 1.1. 背景信息以及准备工作                      | 05 |
| 1.2. 创建AnalyticDB for MySQL Schema和表 | 06 |

# 1.读写AnalyticDB for MySQL 3.0数据

## 1.1. 背景信息以及准备工作

您可以在DLA中创建OSS对应的Schema和文件表，对OSS中不同类型的文件数据进行清洗，然后将得到的结果数据写入AnalyticDB for MySQL 3.0，使得您可以在毫秒级针对千亿级数据进行即时的多维分析透视和业务探索。

- 在同一地域开通[数据湖分析](#) DLA（Data Lake Analytics）服务、[AnalyticDB for MySQL 3.0](#)以及[对象存储服务](#) OSS（Object Storage Service）服务。
- 在AnalyticDB for MySQL中创建数据库和表，用于存储OSS数据，请参见[CREATE DATABASE](#)、[CREATE TABLE](#)。

表结构信息如下所示。

```
CREATE TABLE shipping (
  id bigint NOT NULL COMMENT '',
  origin_state varchar NOT NULL COMMENT '',
  origin_zip varchar NOT NULL COMMENT '',
  destination_state varchar NOT NULL COMMENT '',
  destination_zip varchar NOT NULL COMMENT '',
  package_weight int NOT NULL COMMENT '',
  PRIMARY KEY (id)
)
DISTRIBUTED BY HASH (id)
COMMENT ''
```

```
CREATE TABLE order_table (
  customer_id bigint NOT NULL COMMENT '',
  order_id varchar NOT NULL COMMENT '',
  order_time timestamp NOT NULL COMMENT '',
  order_amount double NOT NULL COMMENT '',
  order_type varchar NOT NULL COMMENT '',
  address varchar NOT NULL COMMENT '',
  city varchar NOT NULL COMMENT '',
  order_season bigint COMMENT '',
  PRIMARY KEY (customer_id,order_time)
)
DISTRIBUTED BY HASH(customer_id)
PARTITION BY VALUE (DATE_FORMAT(order_time, '%Y%m%d')) LIFECYCLE 30
COMMENT ''
```

### 注意事项

- AnalyticDB for MySQL为主键覆盖逻辑，若整个INSERT FROM SELECT的ETL任务失败，需要整体重试。
- 在AnalyticDB for MySQL端查询写入数据时，会有一定的延迟，具体延迟时间取决于AnalyticDB for MySQL的集群配置。
- 建议将ETL任务尽量切分为小的单位分批执行。例如200GB的OSS数据，在业务允许的情况下，200GB的数据切分为100个文件夹，每个文件夹2GB数据。在DLA中建100张表，100张表分别做ETL，单个ETL任务失败，只重试单个ETL任务。

# 1.2. 创建AnalyticDB for MySQL Schema和表

您要先在DLA中创建AnalyticDB for MySQL Schema和对应的外表，然后才能通过INSERT INTO SELECT将OSS中的数据写入AnalyticDB for MySQL。

## 创建Schema

1. 登录DLA控制台。
2. 单击左侧导航栏的访问点管理，然后，登录DMS，在DLA中创建AnalyticDB for MySQL Schema。

```
CREATE SCHEMA ads_database_schema WITH DBPROPERTIES
(
  CATALOG = 'adb3',
  LOCATION = 'jdbc:mysql://am-bp1m76i0ivxpc60v313****.ads.aliyuncs.com:3306/adb_demo',
  VPC_ID='your-adb-vpc_id',
  INSTANCE_ID='your-adb-instance_id',
  USER='account1',
  PASSWORD='Account1'
);
```

| 参数               | 说明                                     |
|------------------|--|
| CATALOG = 'adb3' | 指定创建的是AnalyticDB for MySQL 3.0 Schema。 |
| LOCATION         | AnalyticDB for MySQL的连接信息。             |
| VPC_ID           | AnalyticDB for MySQL的VPC ID。           |
| INSTANCE_ID      | AnalyticDB for MySQL的集群ID。             |
| USER             | AnalyticDB for MySQL中的数据库账号。           |
| PASSWORD         | USER 对应的密码。                            |

## 创建表

在DLA中创建AnalyticDB for MySQL，分为两种情况：

- 表名、字段名或者字段类型与AnalyticDB for MySQL中的表信息不一致。
- 表名、字段名或者字段类型与AnalyticDB for MySQL的表信息完全一致。

## 表名和字段信息一致（建议一致）

以下示例在DLA中创建shipping表。

```
CREATE EXTERNAL TABLE shipping (  
  id bigint NOT NULL COMMENT "",  
  origin_state varchar NOT NULL COMMENT "",  
  origin_zip varchar NOT NULL COMMENT "",  
  destination_state varchar NOT NULL COMMENT "",  
  destination_zip varchar NOT NULL COMMENT "",  
  package_weight int NOT NULL COMMENT "",  
  PRIMARY KEY (id)  
)
```

表名或字段信息不一致

以下示例在DLA中创建与order\_table对应的表order\_table1。其中，order\_table1中部分字段名与order\_table不同，order1\_id的数据类型与order\_id不同。

```
CREATE EXTERNAL TABLE order_table1 (  
  customer1_id bigint NOT NULL COMMENT "",  
  order1_id bigint NOT NULL COMMENT "",  
  order1_time date NOT NULL COMMENT "",  
  order1_amount double NOT NULL COMMENT "",  
  order1_type varchar NOT NULL COMMENT "",  
  address1 varchar NOT NULL COMMENT "",  
  city1 varchar NOT NULL COMMENT "",  
  order1_season bigint COMMENT "",  
  PRIMARY KEY (customer1_id)  
)  
tblproperties (  
  table_mapping = 'adb_demo.order_table',  
  column_mapping = 'customer1_id,customer_id; order1_id,order_id; order1_time:order_time,  
  order1_amount:order_amount, order1_type:order_type, address1:address,  
  city1:city,order1_season:order_season'  
);
```

| 参数             | 说明              |
|----------------|-----------------|
| tblproperties  | DLA中的表与源表的映射关系。 |
| table_mapping  | 表名映射。           |
| column_mapping | 字段（列）映射。        |