ALIBABA CLOUD

阿里云

大数据计算服务

常见问题

文档版本：20211214

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.项目管理

本文汇总了项目管理的相关问题。

血缘信息的上下游表的相关信息，多久会更新？

每天更新一次，更新方式有两种：

- 实时更新：建表之后立刻能看到。
- 全量更新：实时更新失败会进行全量更新，每天07:00:00之前更新完毕。

表的上下游表是从用户每天执行的任务中提取的相关关系吗？

是的，是通过调度系统运行的任务提取的相关关系。

使用DataWorks提交任务时，需要设置时间参数${bdp.system.bizdate}，如果想提取一年前、半年前、一个月前或一周前的时间，如何设置？

不支持基于${bdp.system.bizdate}进行自定义设置，详情请参见配置调度参数。

如何删除已经创建的MaxCompute项目？

执行如下步骤：

1. 登录DataWorks控制台，进入DataWork**工作空间列表**页面。
2. 找到需要删除的项目。单击项目后的**更多**，选择**删除项目**。
3. 进入删除项目页面，手动输入验证码。
4. 输入完成后删除成功。

> **说明** RAM用户无删除项目权限，如需操作请联系项目管理员。

如何区分工作空间和MaxCompute项目空间？

您需要基于DataWorks工作空间来创建MaxCompute项目空间。即先创建DataWorks工作空间，再创建MaxCompute项目空间，详情请参见创建MaxCompute项目。区分二者的方式如下：

- DataWorks工作空间：登录DataWorks控制台，在左侧导航栏，单击**工作空间列表**，此处您看到的是DataWorks工作空间。
- MaxCompute项目空间：登录MaxCompute控制台，在**项目管理**页签，您可以查看到具体的MaxCompute项目空间及所属DataWorks工作空间。

cmd_file相当于一个脚本程序，是一系列SQL和MapReduce程序，为了复用该脚本程序，可以用$变量吗？可以在调用时传入参数吗？

不支持传入变量，您可以在*cmd.sh*脚本文件中，动态构造MaxCompute的执行语句。

有类似ptkill的方法批量中止超时任务的操作吗？

不支持批量中止操作，只能执行 `kill instanceid` 命令逐一中止任务。

MapJoin中的大表和小表是否可以互换位置？

MapJoin中的大表和小表是根据表占用空间Size大小区分的。

系统会将您指定的小表全部加载到执行Join操作的程序的内存中，继而加快Join的执行速度。如果将大表和小表互换位置，系统不会报错，但是性能会变差。

MaxCompute不支持访问外网，但分布式处理需要访问外网，该如何实现？是否有云组件支持？

MaxCompute只支持访问资源表和资源文件，不支持访问外部的资源，也没有类似的云组件。建议您把对应的请求封装在资源文件和资源表里，或者直接写在MaxCompute表中。

使用MaxCompute客户端连接服务时，报错ODPS-0410031，怎么处理？

报错信息如下。

```
FAILED: ODPS-0410031:Authentication request expired - the expire time interval exceeds the max limitation: 900000, max_interval_date:900000,expire_date:2015-12-23T10:15:31.000Z,now_date:2015-12-23T02:16:00.000Z
```

- 问题原因：出现上述报错，是因为安装客户端的机器本地时间与MaxCompute服务器上的时间不一致，时间相差超过15分钟就会导致请求被服务器认为超时而拒绝。
- 解决办法：可以将机器的本地时间调整后，重新打开客户端。对于中国站区域，机器本地时间获取当前时间即可。

MaxCompute如何在客户端上查看一个任务的历史信息？

执行以下步骤进行操作：

1. 执行 `show p;` 命令获取任务的InstanceID。

   ```
   odps@ aliyun2014>show p;
   StartTime            RunTime  Status   InstanceID                Owner                          Query

   2016-01-20 00:03:06   1s       Failed   [redacted]                [redacted]                     insert into table test
   _flash partition(ds ='2016-01-20')
   select '2016-01-29',getdate() from dual;
   2016-01-20 09:25:05   61s      Success  [redacted]                [redacted]

   2016-01-20 09:29:42   47s      Success  [redacted]                [redacted]

   3 instances
   ```

   如果是SQL任务的话，可以直接看到SQL的内容。

   如果要查看几天前的Instance，可以执行如下命令。

   ```
   show p from date1 to date2 number;
   ```

   例如，查询2015年5月20日至2015年5月22日之间的任务信息。

   ```
   show p from 2015-5-20 to 2015-5-22 2000;
   ```

2. 通过 `wait` 命令，获得任务的运行日志。

```
odps@ aliyun2014>wait 20160120012942877g7aridx5;

ID = 20160120012942877g7aridx5
Log view:
http://logview.odps.aliyun.com/logview/?h=http://service.odps.aliyun.com/api&p=aliyun2014&i=20160120012942877g7aridx5&token=Z
UdVaEtiR[redacted]10sIk
VmZmVjdC[redacted]JpZHg
ilI19XSwiVmVyc2lvbiI6IjEifQ==
Summary:
resource cost: cpu 0.30 Core * Min, memory 0.45 GB * Min
inputs:
        aliyun2014.iris: 150 (1960 bytes)
outputs:
        aliyun2014.wc_out: 1 (552 bytes)
Job run time: 29.000
Job run mode: fuxi job
M1:
        instance count: 1
        run time: 10.000
        instance time:
                min: 5.000, max: 5.000, avg: 5.000
        input records:
                input: 150  (min: 150, max: 150, avg: 150)
        output records:
                R2_1: 150  (min: 150, max: 150, avg: 150)
        writer dumps:
                R2_1: (min: 0, max: 0, avg: 0)
R2_1:
        instance count: 1
        run time: 29.000
        instance time:
                min: 4.000, max: 4.000, avg: 4.000
        input records:
                input: 150  (min: 150, max: 150, avg: 150)
        output records:
                R2_1FS_DataSink_6: 1  (min: 1, max: 1, avg: 1)
        reader dumps:
                input: (min: 0, max: 0, avg: 0)
```

3. 在浏览器中输入Logview地址，即可获得任务的详细日志。详情请参见使用Logview查看作业运行信息。

MaxCompute支持快照吗？ChangeLog的设置方式是什么？

不支持快照，也没有ChangeLog之类的配置功能。

MaxCompute中可以设置表的过期时间，是否有办法设置分区的过期时间？

不支持。

MaxCompute会有lock-in问题吗？

MaxCompute 2.0在用户接口上兼容开源（还在提高兼容性，因为开源系统API也在变动），所以不会有lock-in的问题。在兼容Hive语法、语义以及开发应用各种基于规则的优化器（Rbo）的前提下，引入和开发了基于统计数据指导下更精确的性能优化组件，增加了全新的优化规则。

MaxCompute是否支持RESTful接口？

支持ResuFul API，MaxCompute仅提供Python和Java两种语言实现该功能。

运维中心补数据功能怎么使用？

补数据功能就是重跑任务，您可以选择日期时间段。详细信息请参见执行补数据并管理补数据实例。

新建的RAM用户无法访问MaxCompute，是什么原因？

此问题由权限问题造成，需要阿里云主账号为RAM用户授权，详情请参见RAM用户使用DataWorks。

如何将开通数据保护的MaxCompute表数据导入另一项目空间？

如果您想要将已经开通数据保护的MaxCompute项目中的表数据导入到另一个MaxCompute项目中，需要执行以下操作：

1. 在源表项目空间执行如下命令。

```
add TrustedProject dest_project_name;
```

2. 在目标项目空间（dest_project_name）中执行如下命令。

```
create table like select * from src_project_name.table_name;
```

即创建一张与源表同结构的表，再把数据以SQL的形式插入，即可完成数据的导入。

如何查看某个MaxCompute项目及每张数据表所使用的磁盘空间？

可以使用 `desc 表名;` 命令查看MaxCompute存放的表大小。MaxCompute暂时不能查看整个项目空间大小，但是您可以从DataWorks的数据管理里看到统计信息，详细请参见如何查看MaxCompute的数据量？。

如何调用Package中的表和函数？

Package所在项目空间的Owner授权给当前项目空间。当前项目空间安装这个Package，然后通过 `project_name.table_name` 访问这个Package中的表；通过 `project_name.function_name` 调用这个Package中的函数。

MaxCompute项目中的Owner能否更换为RAM用户？

项目的Owner不可以更换，创建项目空间的人即是项目Owner。您可以将Admin的角色赋予RAM用户。

与Owner相比，Admin角色有哪些限制？

与Owner相比，Admin角色不能进行如下操作：

- Admin角色不能将Admin权限指派给用户。
- 不能设定项目空间的安全配置。
- 不能修改项目空间的鉴权模型。
- Admin角色所对应的权限不能被修改。

在MaxCompute页面，运行任务类的功能在哪里能看到？

MaxCompute没有这样的对应功能，您可以使用 `show p;` 命令查看历史的任务，并使用 `Kill` 命令中止对应的任务。请参见MaxCompute如何在客户端上查看一个任务的历史信息？或实例操作。

使用Use Project命令进入项目空间时为什么会报错Can't bind xml to class?

详细报错信息如下。

```
FAILED:Can't bind xml to class com.aliyun.odps.Project$ProjectModel.
```

这种情况通常是代理软件导致，请关闭代理软件再次尝试。

如果不调用com.aliyun.odps.Instance中的waitForSuccess()方法，是否会导致数据有遗漏？

建议您以官网标准写法为准，`waitForSuccess()` 方法用于监控任务执行是否成功，建议添加调用。

在整个解决方案中，是如何使用MaxCompute的？

MaxCompute通常作为解决方案的一部分，与其它系统的交互过程如下：

1. 上传数据到MaxCompute。

2. 通过SQL或MapReduce任务进行数据分析和挖掘处理。
3. 数据分析、挖掘结果存储到MaxCompute中的结果表。
4. 把MaxCompute结果表导出到RDS数据库（或其它在线存储方案），以提供在线服务。

MaxCompute节点上下游依赖调度的逻辑是什么？

下游对上游的依赖需要遵循如下原则：

- 下游任务生成的实例会将结束时间离自己最近的一个上游实例作为上游依赖，如果上游依赖实例运行成功，才会触发下游节点实例运行。
- 如果上游节点每天生成多个实例，下游无法识别是哪一个实例的结束时间离它最近，因此必须等上游当天生成的所有实例运行完成后才会触发下游节点运行。

MaxCompute节点依赖的上下游节点如何配置？

您可以通过DataWorks的节点上下文功能配置上下游节点，更多配置操作请参见配置节点上下文。

DataWorks与MaxCompute的区别是什么？

MaxCompute是数据仓库，负责存储数据或者对数据进行一系列的开发和运算。

DataWorks的底层存储依赖于MaxCompute，同时为MaxCompute的一系列功能提供了可视化开发和节点的流程管理等功能。详情请参见什么是DataWorks。

父结点和子结点都是虚拟结点，子结点依赖父结点，父结点为天调度，子结点为月调度。但实际执行时子节点每天运行是什么原因？

月调度即调度任务在每月的特定几天，在特定时间点自动运行一次。

在非指定日期，为保证下游实例正常运行，系统会每天生成实例后直接设置为运行成功，而不会真正执行任何逻辑，也不会占用资源。详情请请参见时间属性配置说明。

数据集成里新建离线同步，数据来源是LogHub，如何配置日志迁移的开始时间和结束时间？

日志迁移的开始时间和结束时间在DataWorks中需要通过参数传参。参数配置请参见配置调度参数。

参数配置中$cyctime参数用于指定任务定时调度时间。如果是[]参数，则以cyctime为基准参与运行，和Oracle的时间运算方式一致。

第一次数据同步是全量同步，之后可以选择增量同步。如果每半个小时同步一次，可以设置任务为分钟调度，设置间隔时间为30分钟。详情请参见数据增量同步。

如何查看某个用户在项目中的操作历史记录？

您可以通过审计日志功能查看用户的操作记录。

如何查看项目中添加的资源包信息？

您可以使用MaxCompute的list resources;命令查看项目中存在的资源包信息。

如何查看MaxCompute项目的存储资源使用量？

您可以通过MaxCompute管家查看项目的存储资源使用情况。您可以查看如下信息：

- 当前存储量：指定配额组下的全部项目在搜索截止时刻的存储资源使用量。
- 存储大小趋势：指定配额组下的全部项目在搜索时间段内的存储使用量。

如果需要查看更详细的信息，可以通过用量明细账单查看存储量。

2.系统安全

本文汇总了系统安全常见问题。

- 用户与授权：
  - RAM用户无法访问DataWorks，提示缺少AccessKey ID，但实际有AccessKey ID，如何处理？
  - MaxCompute授权时，报错lack of account provider，如何处理？
  - RAM用户申请项目空间下生产环境表的权限，审批过程中提示授权失败，如何处理？
  - 如何授予用户操作表的权限？
- 项目数据保护与共享：
  - 如何跨项目空间读取数据？
  - MaxCompute如何保证数据安全？
  - MaxCompute的数据是否可靠？
  - 作业运行报错FAILED: ODPS-0010000，是什么原因？
  - 因涉及数据保护，无法将MaxCompute数据导出至MySQL，如何处理？
  - VPC IP白名单是否支持设置网段？
  - 如何找回被删除的表？
  - 如何查看某用户创建的表？
  - RAM用户如何访问其他阿里云账号创建的项目？

如何跨项目空间读取数据？

您可以基于Package实现跨项目空间读取数据的功能，详情请参见基于Package的跨项目空间资源访问、项目空间操作和Package赋权案例。

MaxCompute如何保证数据安全？

MaxCompute拥有完备的措施来保证用户的数据安全：

- 多用户场景，除项目空间的所有者（Project Owner）或项目管理员之外，未经授权的用户无法访问项目空间。
- MaxCompute提供了多种授权方式，保证只有经过授权的用户才能访问项目空间。详情请参见用户与权限管理。
- MaxCompute的安全沙箱系统，防止用户恶意操作。详情请参见Java沙箱。
- 使用AccessKey ID和AccessKey Secret验证用户身份。如果出现信息泄露，您可以快速禁用AccessKey ID和AccessKey Secret，同时不会影响其它AccessKey ID和AccessKey Secret的使用。
- 出现紧急数据安全风险时，您可以执行 `set ProjectProtection=true;` 命令开启数据保护，禁止导出数据，阻止出现数据进一步泄露的情况。

RAM用户无法访问DataWorks，提示缺少AccessKey ID，但实际有AccessKey ID，如何处理？

您需要在个人信息中绑定AccessKey信息。进入个人信息页面，单击**修改AccessKey信息**，输入**Access Key ID**和**Access Key Secret**。完成配置后，请您重新尝试访问DataWorks。

MaxCompute授权时，报错lack of account provider，如何处理？

出现上述报错时，您可以尝试通过如下方法解决：

- 查看是否出现语法错误。您可以根据具体的授权操作检查对应的语法，详情请参见授权。
- 尝试配置账号域。例如 `abc@aliyun.com` 配置为 `aliyun$abc@aliyun.com` 。

MaxCompute的数据是否可靠？

MaxCompute集群是三个副本存储，提供数据的可靠性。

您在使用MaxCompute期间，如果指定了表的生命周期，满足删除规则后，表会被系统自动删除；如果没有指定生命周期，表默认会永久保存。

RAM用户申请项目空间下生产环境表的权限，审批过程中提示授权失败，如何处理？

报错信息如下。

```
class java.lang.IllegalArgumentException: AccessId should not be empty.
```

进入个人信息页面，确认项目空间的所有者和RAM用户是否都已配置AccessKey ID和AccessKey Secret。

作业运行报错FAILED: ODPS-0010000，是什么原因？

报错信息如下。

```
FAILED: ODPS-0010000:System internal error - fuxi task failed, AllMachineInBlackList.
```

这是由于MaxCompute集群做了安全加固，请您提工单联系MaxCompute团队处理。

因涉及数据保护，无法将MaxCompute数据导出至MySQL，如何处理？

您可以通过关闭数据保护或配置Exception Policy来导出数据，详情请参见项目空间的数据保护。

如何授予用户操作表的权限？

授权详情请参见授权。

VPC IP白名单是否支持设置网段？

支持。详情请参见管理IP白名单。

如何为其他成员授权？权限管理中的“客体（Object）”和“操作（Action）”是什么？

可以使用阿里云账号或具备管理员角色的用户为其他RAM用户执行授权操作。

MaxCompute使用ACL授权、授权涉及到三个要素：主体（Subject，可以是用户也可以是角色）、客体（Object）和操作（Action）。更多要素相关信息，请参见授权。

如何找回被删除的表？

MaxCompute提供的备份恢复功能可以帮助您恢复表数据，更多信息，请参见备份与恢复。

如何查看某用户创建的表？

MaxCompute提供Information Schema功能，您可以通过Information Schema的TABLES视图指定owner_name来查看创建的表。

RAM用户如何访问其他阿里云账号创建的项目？

场景：例如现有两个阿里云账号A和B，A账号下有一个RAM用户账号C（ram_user_1），账号C需要访问账号B创建的MaxCompute项目。

实现方法：账号B将账号A加入到账号B创建的项目中，同时账号B为账号A授予MaxCompute Admin角色。然后用账号A登录账号B的项目，通过 `add user ram$A:ram_user_1;` 命令添加账号C到B的项目中。

3.数据上传下载

本文汇总了数据上传下载时遇到的常见问题。

- Tunnel命令常见问题：
  - DataWorks的最大屏显行数是多少？
  - Odpscmd Tunnel目录文件支持中文吗？
  - Tunnel是否支持多并发？
  - Tunnel是否支持ASCII字符的分隔符？
  - 文件大小是否有限制？记录大小是否有限制？是否要使用压缩？
  - 同一个表或分区是否可以并行上传？
  - 是否支持不同字符编码？为什么会出现乱码？
  - 导入后的脏数据如何处理？
  - 上传下载的文件路径是否可以有空格？
  - 导入数据最后一列为什么多出\r符号？
  - Tunnel上传下载正常速度范围是多少？
  - Tunnel域名是什么？
  - 无法上传下载如何处理？
  - 上传下载速度缓慢如何处理？
  - Tunnel需注意的分隔符问题有哪些？
  - Tunnel Upload数据的行为是追加还是覆盖？
  - MaxCompute使用Tunnel Upload命令上传文件数据报错，是否有类似MySQL的-f参数，可以强制跳过错误数据继续进行上传的命令？
  - MaxCompute使用Tunnel Upload命令行上传数据，对数据大小有限制吗？
  - MaxCompute使用Tunnel Upload上传是否支持引用一个表的配置？
  - Tunnel中的history命令信息会保存多久？
- Tunnel上传/下载常见问题：
  - 如何使用Tunnel下载部分指定数据？
  - MaxCompute数据导出分别有几种格式？
  - Tunnel上传数据如何实现覆盖重写的功能？
  - MaxCompute使用Tunnel Upload命令上传数据，如何实现批量上传一个目录下的多个文件到同一张表，并且每个文件放在不同的分区内？
  - MaxCompute使用Tunnel Upload命令行上传数据，如果数据使用空格作为列分隔符，或需要对数据做正则表达式过滤时该如何处理？
  - Tunnel路由功能是什么原因？
  - 什么是MaxCompute Tunnel？
  - BlockId是否可以重复？
  - Block大小是否存在限制？
  - Session是否可以共享使用，是否存在生命周期？
  - 遇到读写超时或IOException时如何处理？
  - MaxCompute Tunnel目前支持哪些语言的SDK？

- MaxCompute Tunnel是否支持多个客户端同时上传同一张表？
- MaxCompute Tunnel适合批量上传还是流式上传？
- MaxCompute Tunnel上传数据时一定要先存在分区吗？
- Dship与MaxCompute Tunnel的关系？
- 用MaxCompute Tunnel上传数据时，一个Block的数据量大小多大比较合适？
- 使用MaxCompute Tunnel下载，总是提示Timeout，是什么原因？
- Tunnel上传时每个Session的生命周期是一天，因源表数据太大，导致Session超时任务失败，如何处理？
- 上传数据Session太多导致上传速度慢，应该如何解决？
- 利用Tunnel命令行工具上传数据时，共分为50个Block，开始一切正常，但是在第22个Block时，出现Upload Fail，重试直接跳过开始上传第23个Block，为什么会发生这种情况？
- 本地服务器每天采集的网站日志有10 GB，需要上传至MaxCompute，在使用Tunnel Upload命令上传时速度约为300 KB/S，如何提升上传速度？
- 如何在Shell脚本中将一个TXT文件中的数据上传到MaxCompute的表中？
- MaxCompute使用Tunnel Upload命令上传数据，如果数据里面有回车或空格为什么上传失败？
- MaxCompute使用TunnelUpload命令上传数据，使用逗号进行列分割，但是数据中有逗号，这种情况如何分割？
- MaxCompute使用Tunnel Upload命令上传数据。Tunnel Upload命令默认使用逗号分割的，但数据CSV文件也是用逗号分割的。文件中的一列数据里本身就含有用引号引起来的逗号。这种情况如何处理？
- MaxCompute使用Tunnel Upload命令上传数据，需要上传很多个数据文件到一个表中，是否有方法写一个脚本就可以把文件夹下的所有数据文件循环上传上去？
- MaxCompute使用Tunnel Upload命令上传两个文件，第一个文件上传结束之后，第二个文件没有上传且没有报错信息，是什么原因？
- MaxCompute使用Tunnel Upload命令行上传CSV文件，如何跳过第一行表头上传其他数据？
- MaxCompute使用TunnelUpload命令行上传CSV文件，为什么导入成功后原文本中有很大一部分内容莫名消失？
- MaxCompute使用Tunnel Upload命令行上传数据，设置了经典网络的Endpoint，但为什么会连接到外网的Tunnel Endpoint？
- MaxCompute使用Tunnel Upload命令行上传数据是否支持限速？
- MaxCompute使用Tunnel Upload命令行上传数据太慢如何处理？
- MaxCompute使用Tunnel Upload命令行上传数据，是按照数据压缩前还是压缩后的大小计费？
- 为什么MaxCompute控制台下载数据返回JSON文件？

- Tunnel SDK常见问题：
  - 使用Tunnel Java SDK上传数据，上传的数据可以自动分配到各个分区吗？
  - 使用Tunnel Java SDK上传数据，如果是分区表，SDK能够动态根据数据创建不同的分区吗？
  - MaxCompute使用Tunnel SDK上传数据时，编写完UDF打成JAR包后上传，对JAR包大小有要求吗？
  - 使用Tunnel批量数据通道SDK来导入MaxCompute数据库是否有分区限制？
  - 如何使用TunnelBufferedWriter规避使用Tunnel SDK进行批量数据上传出错的问题？
  - Tunnel SDK如何一次下载分区表里的所有分区？
- 报错处理常见问题：
  - 通过MaxCompute Tunnel下载，报错You have NO privilege如何处理 ？
  - 报错Java heap space FAILED，如何处理？

什么是MaxCompute Tunnel?

MaxCompute Tunnel是MaxCompute的数据通道，您可以通过Tunnel向MaxCompute中上传或者下载数据。Tunnel仅支持表（不包括视图View）数据的上传下载。

Odpscmd Tunnel目录文件支持中文吗？

支持中文。

Tunnel上传是否支持通配符或正则表达式？

使用Tunnel命令行工具上传数据，当前不支持通配符或正则表达式 。

Tunnel是否支持ASCII字符的分隔符？

命令行方式不支持，配置文件可以用十六进制表示。例如 `\u000A` ，表示回车。

Tunnel需注意的分隔符问题有哪些？

Tunnel 需要注意的分隔符问题，如下所示：

- 行分隔符 `rd` 、列分隔符 `fd` 。
- 列分隔符 `fd` 不能包含行分隔符 `rd` 。
- 默认值为 `\r\n` （windows）和 `\n` （linux）。
- 上传开始的时候会打印提示信息，告知本次上传所使用的行分隔符（0.21.0版本及以后）供用户查看和确认。

文件大小是否有限制？记录大小是否有限制？是否要使用压缩？

文件大小没有限制，但一次上传无法超过2小时，根据实际上传速度和时间可以估算能够上传的数据量。

记录大小不能超过200 MB。

默认会使用压缩，如果带宽允许的情况下，可以关掉压缩。

同一个表或分区是否可以并行上传？

可以并行上传。

是否支持不同字符编码？为什么会出现乱码？

支持不同的编码格式参数，带Bom的标识文件不需要指定编码。

可能是上传文件的字符编码和工具指定的编码不符。

导入后的脏数据如何处理？

导入结束后，如果有脏数据可以通过 `tunnel show bad [sessionid]` 命令查看脏数据。

上传下载的文件路径是否可以有空格？

可以有空格，参数需要用双引号括起来。

导入数据最后一列为什么多出\r符号？

Windows的换行符是 `\r\n` ，Mac OS和Linux的换行符是 `\n` ，Tunnel命令使用系统换行符作为默认列分隔符，所以从Mac OSX或Linux上传Windows编辑保存的文件会把 `\r` 作为数据内容导进去。

Tunnel上传下载正常速度范围是多少？

Tunnel上传下载受网络因素影响较大，正常网络情况下速度范围在1 MB/s~20 MB/s区间内。

Tunnel域名是什么？

不同Region对应不同的域名，详情请参见Endpoint。

无法上传下载如何处理？

找到配置中Tunnel域名，通过 `curl -i 域名` （例如 `curl -i http://dt.odps.aliyun.com` ）测试网络是否连通，若无法连通请检查机器网络或更换为正确的域名。

上传下载速度缓慢如何处理？

您可以从以下几方面进行检查：

- 检查机器网络状态，通过 `ping tunnel_endpoint域名` 检查网络延迟是否异常。
- 检查流量状态， 通过 `ifstat` 等命令检查客户端机器的流量是否满载。
- 若为ECS机器，请检查是否使用的公网域名而不是跨域或ECS域名，若使用公网域名，请检查ECS的带宽使用情况是否打满或更换域名。

报错Java heap space FAILED，如何处理？

报错信息如下。

```
Java heap space FAILED: error occurred while running tunnel command
```

您可以从以下几方面进行解决：

- 如果是上传数据，通常是单行数据太大导致，与整体文件的大小无关。
  i. 首先确认是否是分隔符错误，导致所有数据都进入同一行记录，导致单行数据太大。

ii. 如果分隔符正确，文件中的单行数据的确很大，则为客户端程序的内存不够用，需要调整客户端进程的启动参数。打开odpscmd脚本，适当增加Java进程启动选项中的内存值。如 `java -Xms64m -Xmx512m -classpath "${clt_dir}/lib/*:${clt_dir}/conf/" com.aliyun.openservices.odps.console.ODPSConsole "$@"` 中将 `-Xms64m -Xmx512m` 的值增大即可。

- 如果下载数据，通常是数据量太大，客户端程序的内存不够用。打开odpscmd脚本，适当增加Java进程启动选项中的内存值。如 `java -Xms64m -Xmx512m -classpath "${clt_dir}/lib/*:${clt_dir}/conf/" com.aliyun.openservices.odps.console.ODPSConsole "$@"` 中将 `-Xms64m -Xmx512m` 的值增大即可。

MaxCompute数据导出分别有几种格式？

一般使用Tunnel Download导出数据，格式有TXT、CSV，详情请参见Tunnel命令。

使用 Tunnel SDK 上传数据时，报错StatusConflict，如何处理？

报错信息如下。

```
RequestId=20170116xxxxxxx, ErrorCode=StatusConflict, ErrorMessage=You cannot complete the specified operation under the current upload or download status. java.io.IOException: RequestId=20170116xxxxxxx, ErrorCode=StatusConflict, ErrorMessage=You cannot complete the specified operation under the current upload or download status.at com.aliyun.odps.tunnel.io.TunnelRecordWriter.close(TunnelRecordWriter.java:93)
```

- 问题原因：由上述报错可见，此问题是在准备Close这个Writer时出现的。可能有以下几种情况：
  - 对一个已经关闭的Writer做了关闭操作。
  - 这个Writer对应的Session已经关闭。
  - Session已经被提交过。
- 解决方法：您可以针对上述可能出现的原因进行排查，比如打印一些日志、在提交前打印一些当前Writer的状态与Session的状态。

使用Tunnel Java SDK上传数据，上传的数据可以自动分配到各个分区吗？

目前Tunnel是无法自动上传数据并自动分配到各个分区的。每一次上传只支持数据上传到一张表或表的一个分区，有分区的表一定要指定上传的分区，多级分区一定要指定到末级分区。Java SDK详情请参见Java SDK

使用Tunnel Java SDK上传数据，如果是分区表，SDK能够动态根据数据创建不同的分区吗？

首先需要创建好分区，在使用SDK上传数据时指定分区。或者您也可以先把数据上传到MaxCompute上的表中，再用SQL语句动态分区。

MaxCompute使用Tunnel SDK上传数据时，编写完UDF打成JAR包后上传，对JAR包大小有要求吗？

JAR包不能超过10 MB，如果JAR超过10 MB，建议转用MaxCompute Tunnel Upload命令行上传数据。

使用Tunnel批量数据通道SDK来导入MaxCompute数据库是否有分区限制？

目前支持6万个分区。分区数量过多，会给统计和分析带来极大的不便。MaxCompute会限制单个作业中Instance的数量。作业的Instance和用户输入的数据量及分区数量是密切相关的，因此建议先评估下业务，选择合适的分区策略，避免分区过多带来的影响。

关于分区表的更多信息请参见分区。

BlockId是否可以重复？

同一个UploadSession中的BlockId不能重复。对于同一个UploadSession，用一个BlockId打开RecordWriter，写入一批数据后，调用Close，写入成功后不可以重新再用该BlockId打开另一个RecordWriter写入数据。 Block默认最多20000个，即取值范围为0~19999。

Block大小是否存在限制？

每次上传至Tunnel的数据块大小默认为100 MiB。一个Block大小上限为100 GB，强烈建议为大于64 MB的数据，每一个Block对应一个文件，小于64 MB的文件统称为小文件，小文件过多将会影响使用性能。

使用新版BufferedWriter可以更简单的进行上传且可以避免小文件等问题，详情请参见 Tunnel-SDK-BufferedWriter。

Session是否可以共享使用，是否存在生命周期？

每个Session在服务端的生命周期为24小时，创建后24小时内均可使用，也可以跨进程、线程共享使用，但是必须保证同一个BlockId没有重复使用，分布式上传的操作步骤如下：

1. 创建Session。
2. 估算数据量。
3. 分配Block（例如线程1使用0~100，线程2使用100~200）。
4. 准备数据。
5. 上传数据。
6. Commit所有写入成功的Block。

遇到读写超时或IOException时如何处理？

上传数据时，Writer每写入8KB数据会触发一次网络动作，如果120秒内没有网络动作，服务端将主动关闭连接，届时Writer将不可用，请重新打开一个新的Writer写入。

建议使用Tunnel-SDK-BufferedWriter。

下载数据时，Reader也有类似机制，若长时间没有网络IO会被断开连接，建议Reader过程连续进行，中间不穿插其他系统的接口。

MaxCompute Tunnel目前支持哪些语言的SDK？

MaxCompute Tunnel目前有Java及C++版的SDK。

MaxCompute Tunnel是否支持多个客户端同时上传同一张表？

支持。

MaxCompute Tunnel适合批量上传还是流式上传？

MaxCompute Tunnel用于批量上传，不适合流式上传，流式上传可以使用 DataHub高速流式数据通道。

MaxCompute Tunnel上传数据时一定要先存在分区吗？

是的，Tunnel不会自动创建分区。

Dship与MaxCompute Tunnel的关系？

Dship是一个工具，通过MaxCompute Tunnel来进行上传和下载。

Tunnel Upload数据的行为是追加还是覆盖？

追加模式。

Tunnel路由功能是什么原因？

路由功能指的是Tunnel SDK通过设置MaxCompute获取Tunnel Endpoint的功能。因此，SDK可以只设置MaxCompute的Endpoint来正常工作。

用MaxCompute Tunnel上传数据时，一个Block的数据量大小多大比较合适？

需要综合考虑网络情况、实时性要求、数据如何使用以及集群小文件等因素。通常，如果数量较大且是持续上传模式，Block的数据量在64 MB~256 MB之间， 如果是每天传一次的批量模式，Block可以设置为1 GB左右。

使用MaxCompute Tunnel下载，总是提示Timeout，是什么原因？

通常是Endpoint错误，请检查Endpoint配置。简单的判断方法是通过Telnet等方法检测网络连通性。

通过MaxCompute Tunnel下载，报错You have NO privilege如何处理？

报错信息如下。

```
You have NO privilege ‘odps:Select‘ on {acs:odps:*:projects/XXX/tables/XXX}. project ‘XXX‘ is protected.
```

该项目空间开启了数据保护功能，如果您需要把一个项目中的数据导向另一个项目，需要该项目空间所有者进行操作。

Tunnel上传时异常FlowExceeded如何处理？

报错信息如下。

```
ErrorCode=FlowExceeded,ErrorMessage=Your flow quota is exceeded
```

Tunnel对请求的并发进行了控制，默认上传和下载的并发资源组为2000，任何相关的请求发出到结束过程中均会占用一个Quota单位 。若出现类似错误，解决方案有如下几种：

- 休眠一下再重试。
- 增大Project的Tunnel并发资源组。此方法需要联系管理员评估流量压力。
- 请Project Owner控制占用了大量并发资源组的任务。

Tunnel上传时每个Session的生命周期是一天，因源表数据太大，导致Session超时任务失败，如何处理？

建议将源表拆分成2个任务执行。

上传数据Session太多导致上传速度慢，应该如何解决？

应合理设置Block大小。Block ID最大为20000，Session的时间根据具体业务需求设置，Session 提交以后数据才可见。建议您创建Session的频率不要太高，建议最多5分钟一个Session，Session里的Block值应该设置的较大一些，建议每个Block超过64 MB。

为什么使用Tunnel命令行在DataIDE上进行分区上传时报错？

报错信息如下。

```
FAILED: error occurred while running tunnel command.
```

DataIDE不支持MaxCompute Tunnel命令行工具的Upload语句。

利用Tunnel命令行工具上传数据时，共分为50个Block，开始一切正常，但是在第22个Block时，出现Upload Fail，重试直接跳过开始上传第23个Block，为什么会发生这种情况？

一个Block对应一个HTTP Request，多个Block的上传可以并发且是原子的，一次同步请求要么成功要么失败，不会影响其他的Block。

重传Retry有次数的限制，当重传的次数超过了这个限制，就会继续上传下一个Block。上传完成后，可以通过 `select count(*) from table;` 语句，检查是否有数据丢失。

本地服务器每天采集的网站日志有10 GB，需要上传至MaxCompute，在使用Tunnel Upload命令上传时速度约为300 KB/S，如何提升上传速度？

Tunnel Upload命令上传是不设速度限制的。上传速度的瓶颈在网络带宽以及服务器性能。为了提升性能，可以考虑在上传时分区分表，或在多台ECS上传或下载数据。

如何在Shell脚本中将一个TXT文件中的数据上传到MaxCompute的表中？

请参见MaxCompute客户端（odpscmd）设置命令行的启动参数，在Shell中的启动命令如下。

```
/odpscmd/bin/odpscmd -e “tunnel upload "$FILE" project.table”
```

MaxCompute使用Tunnel Upload命令上传数据，如果数据里面有回车或空格为什么上传失败？

如果数据里有回车或空格，可以给数据设置不同于回车或空格的分隔符后，用 `-rd` 和 `-fd` 指定对应的分隔符实现数据的上传。如果无法更换数据中的分隔符，可以将数据作为单独一行上传，然后使用UDF解析。例如下面示例数据中包含回车，使用 “,” 作为列分隔符 `-rd` ，使用 “@” 作为行分隔符 `-fd` ，可以正常上传。

```
shopx,x_id,100@
shopy,y_id,200@
shopz,z_id,300@
```

上传命令

```
odps@ MaxCompute_DOC>tunnel u d:\data.txt sale_detail/sale_date=201312,region=hangzhou -s false -fd ",
" -rd "@";
```

上传结果

```
+-----------+-------------+-------------+-----------+--------+
| shop_name | customer_id | total_price | sale_date | region |
+-----------+-------------+-------------+-----------+--------+
| shopx     | x_id        | 100.0       | 201312    | hangzhou |
| shopy     | y_id        | 200.0       | 201312    | hangzhou |
| shopz     | z_id
d       | 300.0       | 201312    | hangzhou |
+-----------+-------------+-------------+-----------+--------+
```

MaxCompute使用TunnelUpload命令上传数据，使用逗号进行列分割，但是数据中有逗号，这种情况如何分割？

如果数据描述字段内本身有逗号，可以考虑转换数据的分隔符为其他符号，再通过 `-fd` 指定为其他分隔符进行上传。

MaxCompute使用Tunnel Upload命令上传数据。Tunnel Upload命令默认使用逗号分割的，但数据CSV文件也是用逗号分割的。文件中的一列数据里本身就含有用引号引起来的逗号。这种情况如何处理？

CSV文件使用其他分隔符，可以通过 `-fd` 参数指定。

通常，如果数据中有很多符号，可能与分隔符发生冲突，可以自定义数据中的分隔符来避免冲突，比如 `$#@$@ 或者)$*#@$@$` 。

MaxCompute使用Tunnel Upload命令上传数据时失败，报错java.lang.OutOfMemoryError是什么原因？

数据上传时内存溢出了。目前TunnelUpload命令是支持海量数据的上传的，如果出现内存溢出，可能是因为数据的行分隔符和列分隔符设置错误，导致整个文本会被认为是同一条数据，全部缓存至内存里导致内存溢出报错。

这种情况下可以先用少量的数据进行测试，当 `-td` 及 `-fd` 调试成功后再上传全量数据。

MaxCompute使用Tunnel Upload命令上传数据，需要上传很多个数据文件到一个表中，是否有方法写一个脚本就可以把文件夹下的所有数据文件循环上传上去？

Tunnel Upload命令上传支持文件或目录（指一级目录）的上传，详情请参见使用说明。

例如下述命令，上传数据为文件夹*d:\data*。

```
odps@ MaxCompute_DOC>tunnel u d:\data sale_detail/sale_date=201312,region=hangzhou -s false;
```

导入文件夹报错，字段不匹配，但是这个文件夹下的文件单独导入时是可以导入的，是因为文件太大吗？

在Upload命令后加上 `-dbr=false -s true` 对数据格式进行验证。

出现 `column mismatch` 通常是由于列数不匹配导致的。例如列分隔符设置的不对或者文件最后有空行，导致空行通过分隔符进行分割时列数不对。

MaxCompute使用Tunnel Upload命令上传两个文件，第一个文件上传结束之后，第二个文件没有上传且没有报错信息，是什么原因？

当使用老版本MaxCompute客户端，上传参数有 `--scan` 时，续跑模式的参数传递存在问题，将 `--scan=true` 去掉重试即可。

MaxCompute使用Tunnel Upload命令把一个目录下的所有文件上传到一个表里，并且想要自动建立分区，执行报错为acp FAILE，是什么原因？

报错信息如下。

```
Unrecognized option: -acp FAILED: error occurred while running tunnel comman
```

出现这种报错通常是因为使用了不支持的命令或字符。MaxCompute使用Tunnel Upload命令上传不支持通配符及正则表达式。

MaxCompute使用Tunnel Upload命令上传文件数据报错，是否有类似MySQL的-f参数，可以强制跳过错误数据继续进行上传的命令？

使用 `-dbr true` 参数忽略脏数据（多列、少列及列数据类型不匹配等情况）。 `-dbr` 参数默认值为false，表示不忽视脏数据，当值为true时，将不符合表定义的数据全部忽略。详情请参见使用说明。

MaxCompute使用Tunnel Upload命令上传文件数据报错如下是为什么？

```
java.io.IOException: RequestId=XXXXXXXXXXXXXXXXXXXXXXXXX, ErrorCode=StatusConflict, ErrorMessage=You cannot complete the specified operation under the current upload or download status.
    at com.aliyun.odps.tunnel.io.TunnelRecordWriter.close(TunnelRecordWriter.java:93)
    at com.xgoods.utils.aliyun.maxcompute.OdpsTunnel.upload(OdpsTunnel.java:92)
    at com.xgoods.utils.aliyun.maxcompute.OdpsTunnel.upload(OdpsTunnel.java:45)
    at com.xeshop.task.SaleStatFeedTask.doWork(SaleStatFeedTask.java:119)
    at com.xgoods.main.AbstractTool.excute(AbstractTool.java:90)
    at com.xeshop.task.SaleStatFeedTask.main(SaleStatFeedTask.java:305)java.io.IOException: RequestId=XXXXXXXXXXXXXXXXXXXXXXXXX, ErrorCode=StatusConflict, ErrorMessage=You cannot complete the specified operation under the current upload or download status.
```

此错误表示当前文件已经在上传或下载中，无法重复操作。

MaxCompute使用Tunnel Upload命令行上传数据，对数据大小有限制吗？

Tunnel Upload命令行通常不会限制需上传的数据大小。

MaxCompute使用Tunnel Upload命令行上传CSV文件，如何跳过第一行表头上传其他数据？

建议使用 `-h true` 参数，跳过第一行表头。

一次性上传8000万条数据时，在执行odps tunnel recordWriter.close()时报错StatusConflict，是什么原因？

报错原因如下。

```
ErrorCode=StatusConflict, ErrorMessage=You cannot complete the specified operation under the current upload or download status.
```

此报错说明Session状态错误，建议重新创建一个Session再上传一次数据。

从报错上看，前面的操作里已经关闭或者提交了这个Session。对于不同的分区，每个分区需要单独的一个Session。对于多次提交导致报错，请先检查数据是否已经上传成功，如果失败，请重新上传一次。请参见多线程上传示例

如何使用TunnelBufferedWriter规避使用Tunnel SDK进行批量数据上传出错的问题？

MaxCompute Java SDK在0.21.3-public版本之后新增了BufferredWriter的SDK，简化了数据上传，并且提供了容错功能。BufferedWriter从用户角度看，就是在Session上打开一个Writer然后进行写记录即可。具体实现时，BufferedWriter先将记录缓存在客户端的缓冲区中，并在缓冲区填满之后打开一个HTTP连接进行上传。

BufferedWriter会尽最大可能容错，保证数据上传上去。使用方法请参见BufferedWriter使用指南

MaxCompute使用TunnelUpload命令行上传CSV文件，为什么导入成功后原文本中有很大一部分内容莫名消失？

这种情况很可能是因为数据编码格式错误或者是分隔符使用错误导致上传到表的数据错误。建议规范原始数据后上传。

MaxCompute使用Tunnel Upload上传是否支持引用一个表的配置？

可以使用Shell脚本执行Tunnel Upload命令行实现上传。可通过 `/odpscmd/bin/odpscmd -e` 执行脚本，并在脚本内粘贴表格配置。

MaxCompute使用Tunnel Upload命令行上传数据，如果数据使用空格作为列分隔符，或需要对数据做正则表达式过滤时该如何处理？

Tunnel Upload命令行不支持正则表达式。如果数据使用空格作为列分隔符，或需要对数据做正则表达式过滤时可借助MaxCompute的UDF自定义函数功能。

首先，将数据作为单列数据上传。本例中原始数据如下，列分割符为空格，行分隔符为回车，并且需要取的部分数据在引号内，部分数据例如”-“需要被过滤。这种复杂的需求可通过正则表达式实现。

```
10.21.17.2 [24/Jul/2018:00:00:00 +0800] - "GET https://help.aliyun.com/document_detail/73477.html" 200 0 81615 81615 "-" "iphone" - HIT - - 0_0_0 001 - - - -
10.17.5.23 [24/Jul/2018:00:00:00 +0800] - "GET https://help.aliyun.com/document_detail/73478.html" 206 0 49369 49369 "-" "huawei" - HIT - - 0_0_0 002 - - - -
10.24.7.16 [24/Jul/2018:00:00:00 +0800] - "GET https://help.aliyun.com/document_detail/73479.html" 206 0 83821 83821 "-" "vivo" - HIT - - 0_0_0 003 - - - -
```

1. 为使数据单列上传，首先在MaxCompute项目空间内创建一个单列的表格用于接收数据。

   ```
   odps@ bigdata_DOC>create table userlog1(data string);
   ```

2. 使用一个不存在的列分隔符 `\u0000` 上传数据，从而达到不分割列的效果。

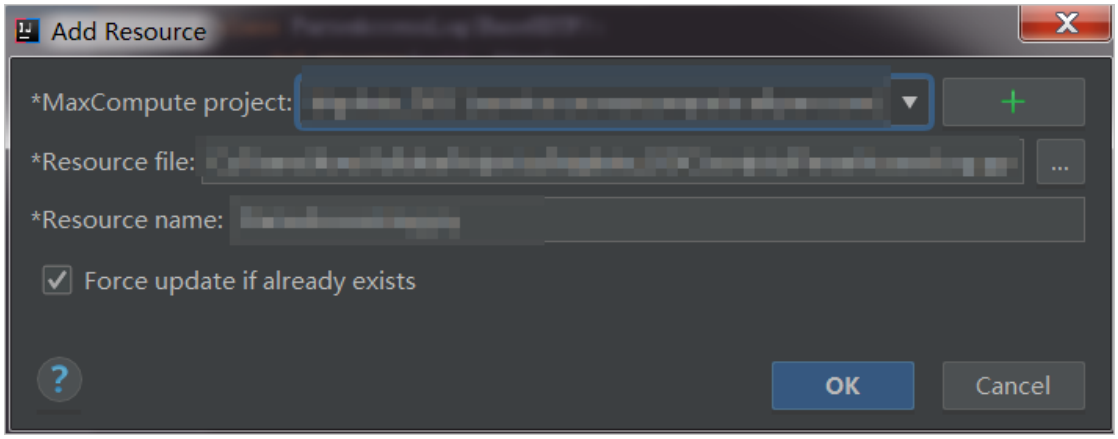
   ```
   odps@ bigdata_DOC>tunnel upload C:\userlog.txt userlog1 -s false -fd "\u0000" -rd "\n";
   ```

3. 完成原始数据上传后，使用MaxCompute IntelliJ IDEA编写一个Python UDF（您也可以使用JAVA UDF），详情可参见配置Python开发环境。

   i. 使用代码如下。

```
from odps.udf import annotate
from odps.udf import BaseUDTF
import re       #此处引入正则函数
regex = '([(\d\.)]+) \[(.*?)\] - "(.*?)" (\d+) (\d+) (\d+) (\d+) "-" "(.*?)" - (.*?) - - (.*?) (.*?) - - - -'       #使用的正则表达式
# line -> ip,date,request,code,c1,c2,c3,ua,q1,q2,q3
@annotate('string -> string,string,string,string,string,string,string,string,string,string,string') #请注意string数量和真实数据保持一致，本例中有11列。
class ParseAccessLog(BaseUDTF):
  def process(self, line):
    try:
      t = re.match(regex, line).groups()
      self.forward(t[0], t[1], t[2], t[3], t[4], t[5], t[6], t[7], t[8], t[9], t[10])
    except:
      pass
```

   ii. 完成函数的编写后，上传代码。完成函数的编写后，选择上传代码。



   iii. 完成上传后，注册函数并填写函数名称，本例中函数名称为ParseAccessLog。

4. 函数上传完成后，就可以使用编写的UDF函数处理上传到表格userlog1的原始数据了，注意不要写错列的名称，本例中为data。您可以使用正常的SQL语法，新建一个表格userlog2用于存放处理后的数据。

```
odps@ bigdata_DOC>create table userlog2 as select ParseAccessLog(data) as (ip,date,request,code,c1,c2,c3,ua,q1,q2,q3) from userlog1;
```

   完成处理后，可以观察到目标表已创建，数据成功分列。

```
+----+------+---------+------+----+----+----+----+----+----+----+
| ip | date | request | code | c1 | c2 | c3 | ua | q1 | q2 | q3 |
+----+------+---------+------+----+----+----+----+----+----+----+
| 10.21.17.2 | 24/Jul/2018:00:00:00 +0800 | GET https://help.aliyun.com/document_detail/73477.html | 20
0 | 0 | 81615 | 81615 | iphone | HIT | 0_0_0 | 001 |
| 10.17.5.23 | 24/Jul/2018:00:00:00 +0800 | GET https://help.aliyun.com/document_detail/73478.html | 20
6 | 0 | 4936 | 4936 | huawei | HIT | 0_0_0 | 002 |
| 10.24.7.16 | 24/Jul/2018:00:00:00 +0800 | GET https://help.aliyun.com/document_detail/73479.html | 20
6 | 0 | 83821 | 83821 | vivo | HIT | 0_0_0 | 003 |
+----+------+---------+------+----+----+----+----+----+----+----+
```

MaxCompute使用Tunnel Upload命令上传数据，如何实现批量上传一个目录下的多个文件到同一张表，并且每个文件放在不同的分区内？

可以巧妙的利用Shell脚本实现上述功能，本章节中以在Windows环境下配合odpscmd客户端使用Shell脚本举例，Linux环境下原理相同。Shell脚本内容如下。

```
#!/bin/sh
C:/odpscmd_public/bin/odpscmd.bat  -e "create table user(data string) partitioned by (dt int);" //首先创建一个分区表user，分区关键字为dt，本例中odpscmd客户端的安装路径为C:/odpscmd_public/bin/odpscmd.bat，您可以根据您的实际环境调整路径。
dir=$(ls C:/userlog)  //定义变量dir，为存放文件的文件夹下所有文件的名称。
pt=0 //变量pt用于作为分区值，初始为0，每上传好一个文件+1，从而实现每个文件都存放在不同的分区。
for i in $dir  //定义循环，遍历文件夹C:/userlog下的所有文件。
do
  let pt=pt+1  //每次循环结束，变量pt+1。
  echo $i  //显示文件名称。
  echo $pt //显示分区名称。
  C:/odpscmd_public/bin/odpscmd.bat  -e "alter table user add partition (dt=$pt);tunnel upload C:/userlog/$i user/dt=$pt -s false -fd "%" -rd "@";" //利用odpscmd首先添加分区，然后向分区中上传文件。
done
```

实际运行Shell脚本效果如下，本例中以两个文件userlog1及userlog2举例。

```
C:\Program Files\Git>sh new2.sh

ID = 20180930101843619gmgpe62m
OK
userlog1.txt
1

ID = 20180930101847179ghpnz192
OK
Upload session: 20180930181848c2dbdb0b1e78e146
Start upload:C:\userlog\userlog1.txt
Using @ to split records
Upload in strict schema mode: true
Total bytes:31   Split input to 1 blocks
2018-09-30 18:18:40     upload block: '1'
2018-09-30 18:18:41     upload block complete, blockid=1
OK
userlog2.txt
2

ID = 20180930101852483gux1r292
OK
Upload session: 20180930181853c3dcdb0b1f57b5c4
Start upload:C:\userlog\userlog2.txt
Using @ to split records
Upload in strict schema mode: true
Total bytes:34   Split input to 1 blocks
2018-09-30 18:18:46     upload block: '1'
2018-09-30 18:18:46     upload block complete, blockid=1
OK
```

完成上传后，您可以在odpscmd客户端查看表数据。

```
odps@ MaxCompute_DOC>select * from user where dt < 100;

ID = 20180930102040449nn5f392
Log view:
http://logview.odps.aliyun.com/logview/?h=http://service.cn.maxcompute.aliyun.co
m/api&p=MaxCompute_DOC&i=20180930102040449nn5f392&token=NzRrNDAxaGx2RDRsV2QwN08z
TGdJSUp4ZEdJPSxPRFBTX09CTzoxMDc5OTI2ODk2OTk5NDIxLDE1Mzg5MDc2NDAseyJTdGF0ZW11bnQi
Olt7IkFjdGlvbiI6WyJvZHBzOlJlYWQiXSwiRWZmZWN0IjoiQWxsb3ciLCJSZXNvdXJjZSI6WyJhY3M6
b2RwczoqOnByb2p1Y3RzL21heGNvbXB1dGVfZG9jL2luc3RhbmNlcy8yMDE4MDkzMDEwMjA0MDQ0Z25u
NWYzOTIiXX1dLCJWZXJzaW9uIjoiMSJ9
Job Queueing...
+------------+------------+
| data       | dt         |
+------------+------------+
| we123#asd  | 1          |
| we1234#asd | 1          |
| we1235#asd | 1          |
| waa123#asd | 2          |
| waa1234#asd | 2          |
| waa1235#asd | 2          |
+------------+------------+
6 records (at most 10000 supported) fetched by instance tunnel.
```

MaxCompute使用Tunnel Upload命令上传数据时为什么报错java.io.IOException?

详细报错信息如下。

```
java.io.IOException: Error writing request body to server
```

这是一个上传数据到服务器时的异常，通常是因为上传过程中的网络连接断开/超时导致的。

- 当您的数据源并非是本地文件，需要从数据库等地方获取，因此数据在写入的过程中还需要等待数据获取而导致超时。目前UploadSession在上传数据的过程中，如果600秒没有数据上传，则被认为超时。
- 用户通过公网的Endpoint进行数据上传，由于公网网络质量不稳定导致超时。

解决方法如下：

- 在上传的过程中，先获取数据，再调用Tunnel SDK上传数据。
- 一个Block可以上传64MB~1GB的数据，最好不要超过1万条数据以免因重试导致超时。一个Session可以拥有最多2万个Block。如果您的数据在ECS上，可以参见Endpoint。

MaxCompute使用Tunnel Upload命令行上传数据，设置了经典网络的Endpoint，但为什么会连接到外网的Tunnel Endpoint?

配置文件*odps_config.ini*中除了Endpoint之外还需要配置Tunnel_Endpoint。请参见Endpoint进行配置。目前只有华东2（上海）区域不需要设置Tunnel Endpoint。

MaxCompute使用Tunnel Upload命令行上传数据是否支持限速?

目前MaxCompute使用Tunnel Upload命令行不支持限速，需要通过SDK单独处理。

MaxCompute使用Tunnel Upload命令行上传数据太慢如何处理?

如果上传数据太慢，可以考虑使用 `-threads` 参数将数据切片上传，例如将文件切分为10片上传。

```
odps@ bigdata_DOC>tunnel upload C:\userlog.txt userlog1 -threads 10 -s false -fd "\u0000" -rd "\n";
```

MaxCompute使用Tunnel Upload命令行上传数据，是按照数据压缩前还是压缩后的大小计费?

按照Tunnel压缩后的大小进行计费。

DataWorks的最大屏显行数是多少?

DataWorks默认显示10000行，显示行数目前在DataWorks上不可配置。如果您需要下载数据，请使用Tunnel Download。

Tunnel目录文件支持中文吗?

支持中文。

Tunnel是否支持.dbf后缀非加密数据库文件?

Tunnel支持文本文件，不支持二进制的文件。

Tunnel是否支持多并发?

支持，命令如下。

```
tunnel upload E:/1.txt tmp_table_0713 --threads 5;
```

Tunnel导入数据时报错分区不存在，如何处理？

- 问题现象：在执行一些操作时，例如Tunnel传数据，报错提示 `ErrorCode=NoSuchPartition, ErrorMessage=The specified partition does not exist` 。
- 问题原因：数据要插入的分区不存在导致此错误。
- 解决办法：

  您可以使用 `SHOW PARTITIONS TABLE_NAME;` 命令来判断分区是否存在，并通过 `ALTER TABLE TABLE_NAME ADD [IF NOT EXISTS] PARTITION partition_spec` 来创建对应的分区。

  如果问题还未解决，请提工单。

Tunnel上传数据报错Blocks Not Match，是什么原因？

- 问题现象：使用Tunnel SDK上传数据的时候，报错信息如下。

  ```
  ErrorCode=Local Error, ErrorMessage=Blocks not match, server: 0, tunnelServiceClient: 1
  at com.aliyun.odps.tunnel.TableTunnel$UploadSession.commit(TableTunnel.java:814)
  ```

- 问题原因：从报错上看，原因是服务器收到的Block个数和Commit时候参数里的个数不一致。
- 解决办法：
  - 在代码中查看 `uploadSession.openRecordWriter(i)` 打开的Writer个数和Commit的时候的Block的数组是否能对应上。
  - 代码中写入执行完成后，是否调用 `recordWriter.close();` 。如果直接执行Commit，可能导致服务器端的Block个数不符合预期。

如问题还未解决，请提工单。

Tunnel SDK如何一次下载分区表里的所有分区？

- 问题现象：使用Tunnel SDK下载分区表，返回如下报错信息。

  ```
  ErrorCode=MissingPartitionSpec, ErrorMessage=You need to specify a partitionspec along with the specified table.
  ```

- 问题原因：使用Tunnel SDK下载分区表，需要指定分区列的列值，否则会报错。
- 解决方法：
  - 如果您使用客户端工具里的Tunnel命令行进行导出，客户端支持分区表整个导出，其结果会导出到一个文件夹里。
  - 如果您使用Tunnel SDK进行导出，可以先使用SDK获取分区表的所有分区，如下。

    ```
    odps.tables().get(tablename) t.getPartitions()
    ```

如问题还未解决，请提工单。

如何使用Tunnel下载部分指定数据？

目前Tunnel不支持数据的计算或者过滤。如果需要实现此功能，您可以考虑以下两种方法：

- 先运行SQL任务，将需要下载的数据保存成一张临时表，下载结束后再删除此临时表。

- 如果您所需要的数据量比较小，可以使用SQL命令直接查询需要的数据，无需下载。

Tunnel上传数据如何实现覆盖重写的功能？

目前Tunnel只提供追加的插入方式，如果用户需要覆盖重写，请先删除分区里的数据后再插入数据。

- 如果表是分区表，可以使用 `ALTER TABLE TABLE_NAME DROP [IF EXISTS] PARTITION partition_spec;` 命令。
- 如果是非分区表，可以使用 `TRUNCATE TABLE table_name;` 命令。

为什么MaxCompute控制台下载数据返回JSON文件？

- 问题现象：通过DataWorks下载数据，却得到名为 `getTableDataCsv.json` 的文件，内容如下：

```
{"code":"-100","message":"needLogin","success":"false"}
```

- 问题原因：出现上述问题，是因为未获取用户的登录信息。
- 解决方法：
  - 可能是会话超时，请重新登录后再试。
  - 可能是一个浏览器多个页签中同时登录了多个账号，导致被系统认为没有登录。请先关闭其他页签并重新登录后再试。
  - 可能是使用了类似迅雷的下载软件，因为下载软件并没有保存用户的登录信息，导致被系统认为没有登录。请直接下载数据即可，不需使用下载软件。

MaxCompute导出的数据有几种格式？

一般使用Tunnel Download导出数据，格式有TXT、CSV，详情请参见使用说明。

同一Region内使用Tunnel下载数据为什么会产生费用？

同一Region内使用Tunnel下载数据，必须配置经典网络/VPC类型的Tunnel Endpoint，否则数据可能路由到其他Region，从公网下载从而产生费用。

使用Tunnel SDK上传为何提示重复提交？

- 问题现象：使用Tunnel SDK上传数据时，报错如下。

```
RequestId=20170116xxxxxxx, ErrorCode=StatusConflict, ErrorMessage=You cannot complete the specified operation under the current upload or download status. java.io.IOException: RequestId=20170116xxxxxxx, ErrorCode=StatusConflict, ErrorMessage=You cannot complete the specified operation under the current upload or download status.
at com.aliyun.odps.tunnel.io.TunnelRecordWriter.close(TunnelRecordWriter.java:93)
```

- 问题原因：由上述报错可见，此问题是在准备关闭此Writer时出现的。可能有以下几种情况：
  - 对一个已经关闭的Writer执行关闭操作。
  - 这个Writer对应的Session已经关闭。
  - Session已经被提交过。
- 解决方法：请对上述可能出现的原因进行排查，例如打印日志、在提交前打印当前Writer与Session的状态。

报错Unauthorized是什么原因？

可能原因为：

- AccessKey ID或AccessKey Secret有误。

- 本地机器时间与服务端时间相差15分钟以上。
- 问题现象：报错信息如下。

  ```
  ErrorCode=Unauthorized, ErrorMessage=The request authorization header is invalid or missing.
  ```

- 问题原因：可能原因如下：
  - AccessKey ID或AccessKey Secret有误。
  - 本地机器时间与服务端时间相差15分钟以上。
- 解决方法：
  - 检查AccessKey ID和AccessKey Secret是否有误。
  - 将机器的本地时间调整后，重新打开客户端。对于国内Region，设备获取当前时间即可。

使用MaxCompute Tunnel或者Dship下载数据时，如何设置Endpoint更合理？

您可以通过公网或者阿里云内网环境对MaxCompute Tunnel进行访问。当您在阿里云内网环境中，使用MaxCompute Tunnel内网连接下载数据时，MaxCompute不会将该操作产生的流量计入计费。

不同网路环境Tunnel Endpoint配置请参见Endpoint。

Tunnel地址支持自动路由。

> **说明** 当使用Tunnel时，MaxCompute及Tunnel的公网及内网地址需要配对使用。

MaxCompute不支持删除部分数据，该如何删除部分脏数据？

建议一个单表（没有分区的表）或一个分区尽量一次性写完（所有数据WRITE完毕后，只调用一次COMPLETE），不要多次去写同一个分区，否则容易出现脏数据。一旦出现脏数据，可以通过以下方法进行删除：

- 删除整个表或该分区，重新上传数据。
- 如果脏数据可以通过WHERE条件过滤出来，也可以通过INSERT+WHERE条件，把需要的数据导入到另一张新表或就地更新（源和目的分区/表名相同的更新方式）。

Fluented插件报错ShardNotReady，如何处理？

- 问题现象：使用Fluentd插件fluent-plugin-aliyun-odps，运行几天后突然报错如下。

  ```
  ShardNotReady, Message: write failed because Shard may wait to be loaded now
  ```

- 问题原因：出现上述报错的原因，如下所示：
  - 可能系统正在升级，会短暂出现这个问题，会很快恢复。
  - Fluentd存在多个进程，请查看配置项Shard_Num是否都配置为相同的值，如果值不同则会导致这个问题。
  - 存在其他方式（例如SDK）执行了Loadshard/Unloadshard的操作导致此问题。

调用StreamWriter向MaxCompute中写入数据时，报错“ErrorCode=MalformedDataStream”，如何处理？

多个线程调用同一个StreamWriter就会出现此错误。建议您使用多个线程，每个线程建立一个StreamWriter，Session是可以多线程用一个。

上传数据的流程是什么？

根据具体场景，流程会有所不同。通常，流程如下：

1. 准备源数据，例如源文件或数据表。
2. 设计表结构和分区定义，进行数据类型转换，然后在MaxCompute上创建表。
3. 在MaxCompute表上添加分区（没有分区时忽略此步骤）。假如使用日期作分区，则添加分区，如：“20140312”。
4. 把数据上传到指定分区或表上。

如何将OSS上存储的数据文件上传至MaxCompute？

将数据文件从OSS下载到ECS，再通过内网或外网连接上传到MaxCompute。

如果数据存储在阿里云的ECS、RDS 或OSS上，将数据上传至MaxCompute的速度如何？

数据的上传速度分以下两种情况：

- 如果数据存储的节点（ECS等）在杭州，则可以进行内网数据传输，效率非常高，单线程能够达到50000条/秒或20 MB/s。
- 如果数据存储的节点在青岛或北京，需要采用公网方式上传或下载，速度取决于公网的速度。

若数据不在云上存储而是在本地存储，是否可以上传到MaxCompute中？

可以，MaxCompute支持通过公网方式上传或下载数据。

> **说明**
>
> - 运行效率取决于网络带宽。
> - 因为公网存在不稳定情况，如果数据量较大，请在代码中实现并发和断点续传逻辑。
> - 如果是长期运行，建议把服务器迁移到阿里云。内网环境下上传或下载效率要比公网环境下高百倍左右。
> - 如果历史数据量非常大，建议您提交工单，以协商数据传输方案。

使用DataWorks导入一个包含有中文的TXT文件，导入和查看均正常，但在MaxCompute客户端查看表详情时，中文显示为乱码是什么原因？

请保证您的TXT文件为UTF-8格式。若不是，您可以用记事本打开文件，单击 **另存为**，将其保存为UTF-8格式。

使用Tunnel上传数据无法导入表中，为何报Java异常Column Mismatch？

通常情况下，可能是数据源文件的行分隔符有问题，将多条记录当成一条记录了。您需要查看是否存在此问题，并重新设置 `-rd` 参数。

如何从MaxCompute上批量导入数据到OCS业务场景（需要定时从MaxCompute上将Key-Value的Value值批量导入OCS）？

您可以定时调用MaxCompute的Tunnel SDK读取MaxCompute的数据，处理后再用OCS的SDK把数据写入OCS内。更多详细信息，请参见DownloadSession。

HubTable数据上传必须用Java SDK吗？能否仅使用客户端完成？

不能，目前只支持在客户端执行 `hub load 3 shards on test_project.test_table` 命令加载Shard，之后还是要用SDK来上传数据。

MaxCompute里存储的数据可以导出到E-MapReduce中吗？

目前E-MapReduce支持HDFS和OSS。但还没有工具可以直接从MaxCompute导入到HDFS。

可以通过DataWorks数据集成将MaxCompute中的数据导入到OSS中，然后使用SDK导入到HDFS，也可以在OSS上长期保存，计算时再读取使用。

DataHub 的应用场景是什么？

DataHub用于实时上传数据的场景，主要用于流式计算。数据上传后会保存到实时表里，后续会在几分钟内通过定时任务的形式同步到离线表里，供离线计算使用。

DataHub表和MaxCompute中创建的表，是否是一个表？

MaxCompute的表分为离线表和在线表，DataHub的表是在线表。只有在线表才能处理DataHub数据，以及作为流式计算的输入源。

在线表的数据，会每隔几分钟开启一个定时任务归档到离线表里，所以在线表的数据和离线表的数据是有几分钟的延迟。

MaxCompute中，DataHub是否有流量限制？

在MaxCompute中上传实时数据，将数据通过某个数据通道（Shard）写入MaxCompute表中时，多个客户端可以同时往一个Shard中写。

目前的流量限制为一个Shard数据包5000 pack/s，流量10 MB/s。您可以根据数据表实际的写入流量，配置一个或者多个Shard。

使用DataHub，为何持续报错“ErrorCode:ShardNotReady”？

DataHub写入数据的分区不存在因此报错，请手动创建缺少的分区重试即可。

Tunnel命令上传数据时，需要设置行分隔符和列分隔符，但是如果源数据里也包含这样的字符，会导致Tunnel命令解析数据的时候出现异常，报错提示列数不匹配，如何处理？

建议您使用如下办法解决：

- 如果可以修改原文件的分隔符，将原文件中的分隔符修改为一些比较特殊的，数据里不会包含的字符或者字符串，然后通过设置 `-rd` 和 `-fd` 修改上传时的分隔符设置，从而正确解析。详情请参见使用说明。
- 如果有特定的逻辑可以解析原文件，您可以考虑编写Tunnel SDK代码读取原文件后解析字符串并上传。详情请参见概述。

如问题还未解决，请提工单。

当一份数据源中的数据对应MaxCompute分区表的多个分区，如何根据数据的内容导入到不同的分区里？

数据导入时只能将数据导入到非分区表或者是分区表的指定分区里。如果需要将数据导入到不同的分区中，可以先创建一张表临时存放所有的导入数据，然后使用SQL进行动态分区。更多动态分区语法请参见插入或覆写动态分区数据（DYNAMIC PARTITION）。

- 如果通过Tunnel等方法实现数据的导入，可以在完成导入后，自行运行SQL进行动态分区。
- 如果通过DataWorks的同步任务执行每天的定时任务，可以在同步任务结束后再运行SQL任务，并把SQL任务的父任务设置成同步任务。

DataHub和Tunnel应用场景的区别是什么？

DataHub用于实时上传数据的场景，主要用于流式计算的场景。数据上传后会保存到实时表里，后续会在几分钟内通过定时任务的形式同步到离线表里，供离线计算使用。

Tunnel用于批量上传数据到离线表里，适用于离线计算的场景。

DataHub上传数据时，对数据大小有哪些限制？

每个字段的大小不能超过这个字段本身的限制，详情请参见数据类型版本说明，例如STRING类型的长度不能超过8 MB。

目前上传的过程，是将多条数据打包成一个Package后上传。

多线程上传数据时报错ODPS-0110061，如何处理？

- 问题现象：多线程上传数据时，报错如下。

  ```
  FAILED: ODPS-0110061: Failed to run ddltask - Modify DDL meta encounter exception : ODPS-0010000:System internal error - OTS transaction exception - Start of transaction failed. Reached maximum retry times because of OTSStorageTxnLockKeyFail(Inner exception: Transaction timeout because cannot acquire exclusive lock.)
  ```

- 问题原因：上传数据时高并发写入同一个表，频繁并发操作导致报错。
- 解决方法：请适当减少并发数，在请求之间加入延迟时间，并且在出错的时候重试。

MaxCompute如何通过Sharding-JDBC抽取和回流数据？

目前MaxCompute不支持通过Sharding-JDBC抽取和回流数据。JDBC的更多信息请参见概述。

Tunnel中的history命令信息会保存多久？

与时间无关，默认保存500条。

4.SQL

4.1. SQL语句

本文为您介绍SQL语句的常见问题。

- 功能说明：
  - MaxCompute与关系型数据库有什么区别？
  - MaxCompute与标准SQL的主要区别是什么？如何解决？
  - MaxCompute能否像MySQL一样灵活使用用户变量（即MySQL的@变量名）？
  - 如何关闭复制和下载功能？
  - 如何查看SQL执行费用？
  - MaxCompute是否支持ORDER BY FIELD NULLS LAST语法？
  - MaxCompute客户端支持并行下载吗？
  - 如何关闭查询加速模式？
  - 如何收集一个月内访问MaxCompute的用户信息？
  - 如何在SQL中实现循环？
  - 如何在SQL中调用赋值节点？
- 分区：
  - 如果源表没有分区字段，是否可以增加或更改分区？
  - 如何更新MaxCompute表或分区中的数据？
  - 如何删除MaxCompute表或分区中的数据？
  - 在MaxCompute中，一张表的分区的数量是否越多越好？
  - 分区和分区列的区别是什么？
  - 在执行MaxCompute SQL时使用动态分区，将GMT格式化作为分区字段，产生大量分区和记录数，一直没有运行完成，是什么原因？
  - SQL语句支持一次添加多个分区吗？
  - 设置表的生命周期为3天，每个表的分区存储量很大，如何清理分区表旧数据？
  - 如何能提高查询效率？分区设置能调整吗？
  - 如果一个表有很多分区，如何清空表的所有分区？
  - 如何对表的分区数据做LEFT JOIN操作？
  - 如何查看指定的分区是否存在？
  - 如何查看分区数量？
- 表：
  - 如何查看MaxCompute表的最近访问时间？
  - 除UUID函数外，如何设置MaxCompute表的主键，实现唯一性索引？
  - 是否可以添加或删除列？
  - 如何添加列？
  - 如何删除列？
  - 如何查看MaxCompute的数据量？

- 如何用MAPJOIN缓存多张小表？
- 如何向MaxCompute表中插入数据？
- MaxCompute表如何设置自增长列？
- 如果表数据量较大，如何删除非分区表中的重复数据？
- MaxCompute单表可以存放的最大列数是多少？
- MaxCompute查询得到的数据是根据什么排序的？
- MaxCompute支持虚拟表吗？例如MySQL中的DUAL表？
- 是否能将RDS中的表一次性导入到MaxCompute中？如果导入成功为什么物理存储显示是0？
- 当目标表的字段类型为VARCHAR(10)，插入数据溢出时会报错吗？
- MaxCompute的表有无索引？
- 如何快速查看项目空间下哪些表是分区表？
- 如何将开发环境的表数据同步至生产环境的表中？
- 通过JDBC方式访问MaxCompute可以向MaxCompute中插入数据吗？
- 因误操作删除的表可以恢复吗？
- 除使用UDF外，如何合并两个没有任何关联关系的表？
- 如何将一行数据拆分为多行数据？
- 如何删除生产环境的表？
- 如何修改表的Hash Clustering属性？
- 如何查看指定的表是否存在？
- 是否支持将非分区表修改为分区表？
- 如何创建视图？
- 如何查询某个用户创建的表？
- 待创建表的字段名与关键字相同，如何处理？
- 如何查看表的行数？

- 报错处理：
  - 执行MaxCompute SQL过程中，报错ODPS-0130071，如何处理？
  - 在执行MaxCompute SQL过程中，报错Table xx has n columns，but query has m columns，如何处理？
  - 使用COALESCE函数时，只要超过一个Expression，会报错ODPS-0130071，如何处理？
  - 执行TO_DATE函数时，报错没有分钟部分，如何处理？
  - 隐式类型转换时，报错ODPS-0121035，如何处理？
  - 在执行MaxCompute SQL过程中，报错输入表过多，如何处理？
  - 在执行MaxCompute SQL过程中，报错输出表的分区过多，如何处理？
  - 在执行MaxCompute SQL过程中，报错ODPS-0010000，如何处理？
  - 在执行MaxCompute SQL过程中，报错Repeated key in GROUP BY，如何处理？
  - 向MaxCompute表中插入FLOAT类型的数据报错，如何处理？
  - 在执行MaxCompute SQL过程中，报错ODPS-0130089，如何处理？
  - 删除分区时，报错ODPS-0130161，如何处理？
  - 已经指定了分区条件，为何提示禁止全表扫描？
  - 执行查询SQL时，报错ValidateJsonSize error，如何处理？

- 插入动态分区报错，如何处理？
- 执行MaxCompute SQL过程中，报错Expression not in GROUP BY key，如何处理？
- 执行MaxCompute SQL过程中，报错ODPS-0130071，如何处理？
- 在执行MaxCompute SQL过程中，报错ODPS-0121145 ，什么原因？
- 在执行MaxCompute SQL过程中，报错Semantic analysis exception，如何处理？
- MaxCompute SQL设置过滤条件后，为什么还报错提示输入的数据超过100 GB？
- MaxCompute客户端的SQL语句执行成功，为什么会打印出异常信息？
- 查询分区表的WHERE条件是add_months('yyyy-mm-dd',x)，报错is full scan with all partitions，please specify partition predicates，如何处理？
- 查询数据时，报错Semanticanalysisexception-XXXtypeisnotenabled incurrentmode，如何处理？
- 在DataWorks里执行需要传参的SQL时报错，是什么原因？
- 执行SELECT * FROM XXX ORDER BY XXX;命令报错，如何处理？
- 执行MaxCompute SQL过程中，报错Parse exception - invalid token 'cost'，如何处理？
- 获取项目空间下的所有表名称报错，如何处理？
- 在执行MaxCompute SQL过程中，报错ODPS-0121096，什么原因？

- 数据类型：
  - DOUBLE类型数据精度问题
  - 如何解决DECIMAL数据类型精度溢出问题？
  - MaxCompute的时间类型字段是否可以不带时分秒？
  - 新创建的项目空间不支持数据类型自动隐式转换，如何处理？
- 函数：
  - MySQL支持的SUBSTRING_INDEX函数在MaxCompute中支持吗？
  - REGEXP_COUNT函数的参数pattern是否支持嵌入查询语句？
  - 多路输出的情况下，能否在REDUCE函数中获取到每一个Label输出表的表结构？
  - 使用ROUND函数对DOUBLE类型数据四舍五入，为何结果存在偏差？
  - 如何判断一个字段是否为空？
  - 如何连接相同字段？
- 作业：
  - 执行INSERT操作过程中出现错误，会损坏原有数据吗？
  - 在执行MaxCompute SQL过程中，使用NOT IN后面接子查询，子查询返回的结果是上万级别的数据量，但当IN和NOT IN后面的子查询返回的是分区时，返回的数量上限为1000。在必须使用NOT IN的情况下，该如何实现此查询？
  - 如何查看MaxCompute日执行的所有SQL？
  - 如何处理单字段大于8 MB的限制？
  - 在执行MaxCompute SQL过程中，对DOUBLE类型的数据进行等值比较，为什么结果不符合预期？
  - 外关联后发现数据条数比原表多，如何处理？
  - 对相同数据执行INSERT SELECT操作和SELECT操作的结果为什么不一致？
  - 补数据时，误选择INSERT OVERWRITE操作导致原数据库中的30 GB数据被清理，可以恢复吗？
  - 运行SQL语句查询有1万条数据的表的数据，查询一直处于Job Quening状态，如何处理？
  - MaxCompute客户端使用-e参数执行SQL时，是否有长度限制？

如何查看MaxCompute表的最近访问时间？

您可以在DataWorks数据地图中查询表，进入表详情页面查看技术信息，获取表的最近访问时间。

除UUID函数外，如何设置MaxCompute表的主键，实现唯一性索引？

MaxCompute应用于海量数据的批量计算场景，不支持设置表的主键或者唯一性索引。

如果源表没有分区字段，是否可以增加或更改分区？

MaxCompute不支持在源表上直接增加或修改分区字段，分区字段一旦创建就无法修改。您可以重新创建一张分区表，使用动态分区SQL将源表数据导入至新分区表，详情请参见插入或覆写动态分区数据（DYNAMIC PARTITION）。

如何更新MaxCompute表或分区中的数据？

MaxCompute不支持直接对表数据执行更新（UPDATE）操作。

您需要把源分区或源表中的数据导入到新分区或新表中，在导入过程中执行相应的更新逻辑操作。新分区或新表可以与源分区或源表相同，即就地更新。

如何删除MaxCompute表或分区中的数据？

MaxCompute不支持直接对表数据执行删除（DELETE）操作。您可以通过如下方法删除：

- 删除（DROP）表，达到删除数据的目的。
- 如果是非分区表，您可以执行 `TRUNCATE TABLE table_name;` 命令清空表数据或通过 `INSERT OVERWRITE` 命令实现类似的功能。
  - 示例一：删除TableA表中Col=1的数据，命令示例如下。

    ```
    INSERT OVERWRITE TABLE TableA SELECT a,b,c.... FROM TableA WHERE Col <> 1;
    ```

  - 示例二：执行如下命令删除全部数据。

    ```
    INSERT OVERWRITE TABLE TableA SELECT a,b,c.... FROM TableA WHERE 1=2;
    ```

- 如果是分区表，您可以执行 `ALTER TABLE table_name DROP IF EXISTS PARTITION（分区名=‘具体分区值’）` 命令，删除对应的分区，即可删除分区对应的数据。

  例如，表testtable的分区列为ds，执行如下命令删除 `ds='20170520'` 的分区。

  ```
  ALTER TABLE testtable DROP IF EXISTS PARTITION (ds='20170520');
  ```

- 使用INSERT和WHERE条件，将需要的数据导入到至另一个新分区或新表中。INSERT支持源表和目标表相同。

  ```
  INSERT OVERWRITE TABLE sale_detail SELECT * FROM sale_detail WHERE name=“mengyonghui”;
  ```

在MaxCompute中，一张表的分区的数量是否越多越好？

在MaxCompute中，一张表最多允许有60000个分区，同时每个分区的容量没有上限。但是分区数量过多，不便于统计和分析。

MaxCompute限制单个作业中最多不能超过一定数量的Instance，而作业中的Instance数量和输入的数据量以及分区数量是密切相关的，所以您需要根据业务需要，选择合适的分区策略。

如何用MAPJOIN缓存多张小表？

您可以在MAPJOIN中填写表的别名。

示例为鸢尾花的数据，表名为iris，表结构如下。

```
+------------------------------------------------------------------------------+
| Field       | Type    | Label | Comment                                     |
+------------------------------------------------------------------------------+
| sepal_length | double  |       |                                             |
| sepal_width  | double  |       |                                             |
| petal_length | double  |       |                                             |
| petal_width  | double  |       |                                             |
| category     | string  |       |                                             |
+------------------------------------------------------------------------------+
```

执行命令示例如下。

```
SELECT
 /*+ MAPJOIN(b,c) */
 a.,
 b.cnt AS cnt_category,
 c.cnt AS cnt_all
FROM iris a
JOIN
(
 SELECT COUNT() AS cnt,category FROM iris GROUP BY category
) b
ON a.category = b.category
JOIN
(
 SELECT COUNT(*) AS cnt FROM iris
) c;
```

是否可以添加或删除列？

在MaxCompute中，可以添加列，但不可以删除列。

如何添加列？

添加列的命令示例如下。

```
ALTER TABLE table_name ADD COLUMNS (col_name1 type1, col_name2 type2…);
```

如果表中已经存在一部分数据，则新添加列的值为NULL。

如何删除列？

MaxCompute不支持删除表的列。如果您有删除列的需求，可以通过如下方法实现：

1. 创建一张新表，命令示例如下。

   ```
   CREATE TABLE new_table_name AS SELECT c1,c2,c3 FROM table_name;
   ```

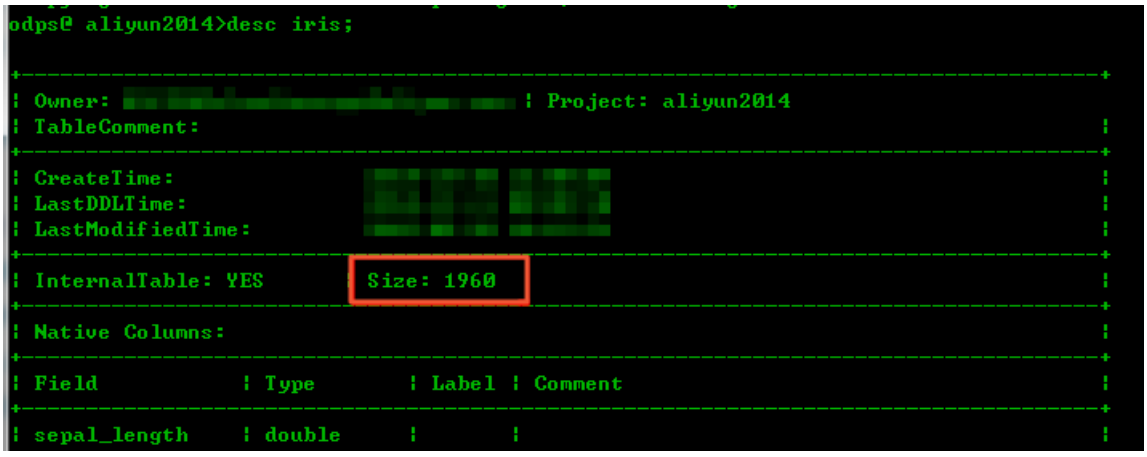
2. 删除旧的表，并重命名新表，命令示例如下。

   ```
   ALTER TABLE new_table_name RENAME TO table_name;
   ```

如何查看MaxCompute的数据量？

查看MaxCompute的数据量包含查看数据条数和占用的物理空间大小。

- 您可以使用 desc 命令查看全量表的物理空间。



- 使用SQL语句查看表的数据条数。例如 SELECT COUNT() AS cnt FROM iris; 。
- 查看分区表单个分区的数据示例如下。 area='N',pdate='1976' 是partition_table2表的一个分区。

如果使用SQL语句，您可以使用WHERE条件过滤分区，例如 SELECT COUNT() AS cnt FROM partition_table2 WHERE area='N' AND pdate='1976'; 。

执行INSERT操作过程中出现错误，会损坏原有数据吗？

不会损坏原有数据。MaxCompute满足原子性，INSERT操作执行成功则更新数据，INSERT操作执行失败则回滚数据。

在执行MaxCompute SQL过程中，报错Table xx has n columns，but query has m columns，如何处理？

执行INSERT INTO或INSERT OVERWRITE操作插入数据时，需要保证SELECT得到的字段和插入的表的字段匹配，匹配内容包括顺序、字段类型和总的字段数量。MaxCompute不支持插入表的指定字段，其他字段为NULL或者其他默认值时，您可以在SELECT时设置为NULL，例如 SELECT 'a',NULL FROM XX 。

执行MaxCompute SQL过程中，报错ODPS-0130071，如何处理？

- 问题现象：执行MaxCompute SQL的过程中，出现类似如下报错。

```
FAILED: ODPS-0130071:Semantic analysis exception - Both left and right aliases encountered in JOIN : line 3:3 'xx' : . If you really want to perform this join, try mapjoin
```

- 问题原因：
  - SQL关联条件ON中包含非等值连接，例如 table1.c1>table2.c3 。

- SQL中JOIN条件的某一侧数据来自两张表，例如 `table1.col1 = concat(table1.col2,table2.col3)` 。

- 解决方法：
  - 修改SQL语句。
  - 如果其中一张表比较小，您可以使用MAPJOIN方法。

分区和分区列的区别是什么？

MaxCompute中的表可以分区，分区表有分区列。您可以通过分区列创建分区。

例如分区 `ds=20150101` ，此处 `ds` 是一个分区列，而 `ds=20150101` 是一个分区。

如何向MaxCompute表中插入数据？

向MaxCompute表中数据的示例如下：

1. 执行如下命令创建一个有1条记录的DUAL表。

```
CREATE TABLE DUAL(cnt BIGINT);
INSERT INTO TABLE DUAL SELECT COUNT(*) AS cnt FROM DUAL;
```

2. 执行如下语句，向MaxCompute表中插入记录。

```
INSERT INTO TABLE xxxx SELECT 1,2,3 FROM DUAL;
```

在执行MaxCompute SQL过程中，使用NOT IN后面接子查询，子查询返回的结果是上万级别的数据量，但当IN和NOT IN后面的子查询返回的是分区时，返回的数量上限为1000。在必须使用NOT IN的情况下，该如何实现此查询？

您可以使用 `LEFT OUTER JOIN` 命令查询。

```
SELECT * FROM a WHERE a.ds NOT IN (SELECT ds FROM b);
改成如下语句。
SELECT a.* FROM a LEFT OUTER JOIN (SELECT DISTINCT ds FROM b) bb ON a.ds=bb.ds WHERE bb.ds is null;
```

如何处理单字段大于8 MB的限制？

由于存储机制限制，MaxCompute表中单个字段的最大长度不能超过8 MB。对于超过8 MB的字段，建议您拆分成多个字段。具体的拆分逻辑您可以根据业务特性设计，保证每个字段不超过8 MB即可。

由于复杂结构的超大字段在数据开发和分析中会严重影响计算性能，因此建议根据数据仓库建设规范来设计您的数据架构，避免出现超大字段：

- 具有复杂结构的原始数据，作为ODS层，最好以压缩的方式归档。
- 定时（例如每天）对ODS层的增量数据做数据清洗，复杂字段拆分为多个简单字段，然后存储在CDM层的表中，便于统计和分析数据。

MaxCompute表如何设置自增长列？

MaxCompute不支持自增长列功能，如果您有此需求，且数据量比较小，建议使用ROW_NUMBER实现。

如何查看MaxCompute日执行的所有SQL？

详情请参见如何在MaxCompute客户端查看一个作业的历史信息？ 。

使用COALESCE函数时，只要超过一个Expression，会报错ODPS-0130071，如何处理？

- 问题现象：使用COALESCE函数只要超过一个Expression，就会报错，报错信息如下。

  ```
  FAILED: ODPS-0130071:Semantic analysis exception - Expression not in GROUP BY key : line 8:9 "$.table"
  ```

  报错的SQL如下。

  ```
  SELECT
  MD5(CONCAT(aid,bid)) AS id
  ,aid
  , bid
  , SUM(amountdue) AS amountdue
  , COALESCE(
  SUM(REGEXP_COUNT(GET_JSON_OBJECT(extended_x, '$.table.tableParties'), '{')),
  DECODE(GET_JSON_OBJECT(extended_x, '$.table'), NULL, 0, 1)
  ) AS tableparty
  , DECODE(SUM(headcount),null,0,SUM(headcount) ) AS headcount
  , 'a' AS pt
  FROM e_orders
  WHERE pt='20170425'
  GROUP BY aid, bid;
  ```

- 问题原因：GROUP BY后面缺少分组字段，因此报错。
- 解决方法：如下表达式的返回值实际上是字段，需要把整个表达式写在GROUP BY后面。

  ```
  COALESCE(
  SUM(REGEXP_COUNT(GET_JSON_OBJECT(extended_x, '$.table.tableParties'), '{')),
  DECODE(GET_JSON_OBJECT(extended_x, '$.table'), NULL, 0, 1)
  ) AS tableparty
  , DECODE(SUM(headcount),null,0,SUM(headcount) ) AS headcount
  ```

DOUBLE类型数据精度问题

- 问题现象：海量数据量场景，对DOUBLE类型的数据执行类似SUM的操作，真实结果和预期结果有所偏差。
- 问题原因：由于数据精度造成的结果偏差。DOUBLE类型是8字节双精度的浮点数。
- 解决方法：您可以考虑先用STRING类型存放数据，然后编写UDF处理数据从而满足任意精度的计算需求。

执行TO_DATE函数时，报错没有分钟部分，如何处理？

- 问题现象：执行SQL语句 `TO_DATE（‘2016-07-18 18:18:18’，‘yyyy-MM-dd HH:mm:ss’）` 时，报错如下。

  ```
  FAILED: ODPS-0121095:Invalid arguments - format string has second part, but doesn’t have minute part : yyyy-MM-dd HH:mm:ss
  ```

- 问题原因：日期格式有误。 `mm` 和 `MM` 都表示月份，分钟需要使用 `mi` 。

隐式类型转换时，报错ODPS-0121035，如何处理？

- 问题现象：执行SQL时报错如下。

```
FAILED: ODPS-0121035:Illegal implicit type cast - in function to_char, in function cast, string datetime’s format must be yyyy-mm-dd hh:mi:ss, input string is xxx。
```

- 问题原因：如果您调用一个需要传入DATETIME类型参数的函数，却传入一个STRING类型的字段，系统会对此字段进行隐式类型转换。MaxCompute仅支持 `yyyy-mm-dd hh:mi:ss` 类型的字符串隐式转换。如果字符串中的数据不是此格式，就会报错。
- 解决方法：建议您使用TO_DATE函数把STRING类型的数据转换为DATETIME类型的数据。如果字符串的类型和TO_DATE要求的不一样，请使用UDF进行解析。

在执行MaxCompute SQL过程中，对DOUBLE类型的数据进行等值比较，为什么结果不符合预期？

由于MaxCompute中DOUBLE类型的数值存在一定的精度差，因此不建议直接使用等于号（=）对两个DOUBLE类型的数据进行比较。

请对两个DOUBLE类型数据相减，然后取绝对值，当绝对值足够小时，系统判定两个DOUBLE类型的数据数值相等。

在执行MaxCompute SQL过程中，报错输入表过多，如何处理？

- 问题现象：在执行MaxCompute SQL过程中，报错如下。

```
FAILED: ODPS-0123065:Join exception - Maximum 16 join inputs allowed
```

- 问题原因：MaxCompute SQL最多支持6张小表的MAPJOIN，并且连续JOIN的表不能超过16张。
- 解决方法：把部分小表JOIN成一张临时表作为输入表，减少输入表的个数。

在执行MaxCompute SQL过程中，报错输出表的分区过多，如何处理？

- 问题现象：在执行MaxCompute SQL过程中，报错如下。

```
FAILED: ODPS-0123031:Partition exception - a single instance cannot output data to more than 10000 partitions
```

- 问题原因：虽然单个MaxCompute表允许有6万个分区，但是单个作业涉及的输出表分区数量只允许有10000个。出现这个错误，通常是因为分区字段设置有误，例如根据ID字段分区造成分区过多。
- 解决方法：一般作业输出动态分区数达到几千已经很大，超过10000可能存在业务逻辑或SQL语法问题。如无逻辑或语法问题，建议修改分区表的分区字段，或将业务逻辑拆分为多个作业，避免出现该错误。

在执行MaxCompute SQL过程中，报错Repeated key in GROUP BY，如何处理？

- 问题现象：在执行MaxCompute SQL过程中，报错如下。

```
FAILED: ODPS-0130071:Semantic analysis exception - Repeated key in GROUP BY。
```

- 问题原因：SELECT DISTINCT后不能跟常量。
- 解决方法：将SQL拆分为两层，内层处理没有常量的DISTINCT逻辑，外层加入常量数据。

在执行MaxCompute SQL过程中，报错ODPS-0010000，如何处理？

- 问题现象：在执行MaxCompute SQL过程中，报错如下。

```
FAILED: ODPS-0010000:System internal error - OTS filtering exception - Ots read range partitions exceeds the specified limit:10000： tableName:xxxx , please check hive conf key
```

- 问题原因：MaxCompute单张表支持6万个分区，但是单次查询最多只支持1万个分区。该报错常见原因如下：
  - 分区未写分区条件。
  - 使用类似用户ID的字段作为分区字段，导致分区数量过多。
- 解决方法：
  - 如果未写分区条件，补上分区条件即可。
  - 如果分区列不合适，导致分区数量太多，请考虑更改分区列。

  如果问题还未解决，请提工单。

在执行MaxCompute SQL过程中，报错ODPS-0130089，如何处理？

- 问题现象：在执行MaxCompute SQL调用UDF时，报错如下。

  ```
  FAILED:ODPS-0130089 Invalid UDF reference - class not ‘xxx’ found for function
  ```

- 问题原因：未找到您定义的类。
- 解决方法：
  - 请在代码中核对路径名和类名是否正确，例如大小写是否正确。
  - 使用解压工具解压JAR包，查看类文件是否存在。函数使用的JAR包名称，您可以在客户端执行 `LIST FUNCTIONS;` 命令获取。

外关联后发现数据条数比原表多，如何处理？

- 问题现象：执行如下MaxCompute SQL语句。

  ```
  SELECT COUNT(*) FROM table1 a LEFT OUTER JOIN table2 b ON a.ID = b.ID;
  ```

  执行完毕后，查询返回结果的条数大于*table1*的数据条数。
- 问题原因：上述的SQL是table1通过ID字段和table2的ID字段做左外关联，所以会出现以下情况：
  - 如果table2表中找不到关联数据，table1也会返回一条数据。
  - 如果table1找不到但是table2能找到关联数据，则不返回结果。
  - 如果table1和table2都能找到关联数据，该关联逻辑和普通的内关联一样。如果同样的ID字段在table2中能找到数据，返回结果为table1和table2的笛卡尔积。

  示例如下。

  table1的数据如下。

id	values
1	a
1	b
2	c

  table2的数据如下。

id	values
1	A
1	B
3	D

执行 `SELECT * FROM t1 LEFT OUTER JOIN t2 ON t1.id = t2.id;` 返回的结果如下。

id1	values1	id2	values2
1	b	1	B
1	b	1	A
1	a	1	B
1	a	1	A
2	c	NULL	NULL

- id=1的数据两边都有，执行笛卡尔积，返回4条数据。
- id=2的数据只有table1有，返回了1条数据。
- id=3的数据只有table2有，table1里没数据，不返回数据。

- 解决方法：首先确认出现数据条数增加是否是因为table2的数据导致。

```
Select id,COUNT() AS cnt FROM table2 GROUP BY id having cnt>1 LIMIT 10;
```

此处增加 `LIMIT 10` 是考虑到如果table2中的数据条数很多，会刷屏。如果只是确认问题，验证前几条数据即可。如果是在重复的情况下不希望执行笛卡尔积，希望有类似SQL里IN的功能，可以改写SQL为如下语句。

```
SELECT * FROM table1 a LEFT OUTER JOIN (SELECT DISTINCT id FROM table2) b ON a.id = b.id;
```

删除分区时，报错ODPS-0130161，如何处理？

- 问题现象：执行如下MaxCompute SQL语句。

```
ALTER TABLE pol_self_overall_2016_part DROP PARTITION;
```

返回如下报错。

```
FAILED: ODPS-0130161:Parse exception - line 1:44 mismatched input '<EOF>' expecting ( near 'partition' in alter table statement
```

- 问题原因：MaxCompute不支持批量删除分区，需要指定并逐个删除具体分区。
- 解决方法：删除分区时，指定具体的分区。按照如下语法格式修改SQL语句。

```
ALTER TABLE TABLE_NAME DROP [IF EXISTS] PARTITION partition_spec;
  partition_spec:
    : (partition_col1 = partition_col_value1, partition_col2 = partition_col_value2, ...)
```

如果表数据量较大，如何删除非分区表中的重复数据？

如果每一列都一样，您可以对所有列执行GROUP BY操作。例如，非分区表*table1*的列为c1，c2和c3，您可以执行如下语句。

```
INSERT OVERWRITE TABLE table1 SELECT c1, c2, c3 FROM table1 GROUP BY c1, c2, c3;
```

> **说明** 建议您在执行此操作前，做好数据备份工作并根据数据量评估此方式的代价是否比重新导入的代价低。

向MaxCompute表中插入FLOAT类型的数据报错，如何处理？

MaxCompute 2.0支持的基本数据类型请参见数据类型版本说明。其中：FLOAT数据类型没有常量定义，若要插入该类型数据，可以使用CAST函数转换数据类型。例如 `CAST(5.1 AS FLOAT)` 将字符串 `‘5.1’` 转为FLOAT类型 `5.1` 。

MaxCompute SQL中使用到新数据类型（TINYINT、SMALLINT、INT、FLOAT、VARCHAR、TIMESTAMP或BINARY）时，需要执行如下set语句开启新数据类型开关：

- Session级别：如果使用新数据类型，您需要在SQL语句前加上set语句 `set odps.sql.type.system.odps2=true;` ，并与SQL语句一起提交执行。
- Project级别：执行set命令 `setproject odps.sql.type.system.odps2=true;` 打开Project级别的新数据类型。该命令需要项目Owner执行。

对相同数据执行INSERT SELECT操作和SELECT操作的结果为什么不一致？

- 问题现象：对相同的STRING类型字段分别执行SQL语句，出现小数位不统一的现象。执行SELECT操作保留2位小数，执行INSERT SELECT操作，结果显示多个小数位。
- 问题原因：对于INSERT SELECT操作，原始字段类型是STRING，在隐式转换为目标类型DECIMAL的过程中，先转换为DOUBLE类型，然后在DOUBLE类型数据的基础上执行ROUND操作。由于DOUBLE类型本身是不精确的，虽然执行了ROUND操作，但是依然可能显示多个小数位。
- 解决方法：建议使用显示转换方式，增加如下语句通过CASE显示转换为DECIMAL类型。

```
CASE WHEN pcm.abc IS NULL THEN 0
            ELSE ROUND(CAST(pcm.abc as decimal) ,2)
        END abc
```

补数据时，误选择INSERT OVERWRITE操作导致原数据库中的30 GB数据被清理，可以恢复吗？

INSERT OVERWRITE操作相当于执行了先删除后插入的操作，不能恢复，需要重新插入数据。

已经指定了分区条件，为何提示禁止全表扫描？

- 问题现象：在两个项目里执行如下同一段代码，一个项目中成功，一个项目中失败。

```
SELECT t.stat_date
FROM fddev.tmp_001 t
LEFT OUTER JOIN (SELECT '20180830' AS ds FROM fddev.dual ) t1
ON t.ds = 20180830
GROUP BY t.stat_date;
```

失败报错如下。

```
Table(fddev，tmp_001) is full scan with all partisions,please specify partition predicates.
```

- 问题原因：在执行SELECT操作时，如果需要指定分区请使用WHERE子句。使用ON属于非标准用法。

  执行成功的项目设置了允许非标准SQL的行为，即执行了 `set odps.sql.outerjoin.supports.filters=true` 命令，该配置会把ON里的条件转为换成过滤条件，可用于兼容HIVE语法，但不符合SQL标准。

- 解决方法：建议将分区过滤条件置于WHERE子句。

运行SQL语句查询有1万条数据的表的数据，查询一直处于Job Quening状态，如何处理？

请排查任务运行状态，可能有任务运行完了所有的作业，请先中止此任务。

执行查询SQL时，报错ValidateJsonSize error，如何处理？

- 问题现象：执行包含200个Union All的SQL语句 `select count(1) as co from client_table union all ...` ，出现如下报错。

```
FAILED: build/release64/task/fuxiWrapper.cpp(344): ExceptionBase: Submit fuxi Job failed, {
  "ErrCode": "RPC_FAILED_REPLY",
  "ErrMsg": "exception: ExceptionBase:build/release64/fuxi/fuximaster/fuxi_master.cpp(1018): Exception Base: StartAppFail: ExceptionBase:build/release64/fuxi/fuximaster/app_master_mgr.cpp(706): ExceptionBase: ValidateJsonSize error: the size of compressed plan is larger than 1024KB\nStack
```

- 问题原因：
  - SQL语句转化为执行计划后，超过了底层架构限制的1024 KB，导致SQL执行报错。执行计划的长度与SQL语句长度没有直接换算关系，暂时无法预估。
  - 由于分区量过大导致执行计划超过限制。
  - 由于小文件比较多导致SQL运行失败。
- 解决方法：
  - 对于过长的SQL语句，建议拆分成多次运行，避免触发长度限制。
  - 如果分区过大，需要调整分区个数，详情请参见分区。
  - 如果是由于小文件较多导致，请参见小文件优化及作业诊断常见问题。

MySQL支持的SUBSTRING_INDEX函数在MaxCompute中支持吗？

支持，详情请参见SUBSTRING_INDEX。

插入动态分区报错，如何处理？

- 问题现象：执行MaxCompute SQL插入动态分区时，报错如下。

```
FAILED: ODPS-0123031:Partition exception - invalid dynamic partition value: province=上海
```

- 问题原因：使用了非法的动态分区。动态分区是根据指定字段进行分区，不支持特殊字符和中文动态分区字段。

  插入动态分区时，如下情况会返回异常：

  - 在分布式环境下执行动态分区SQL时，单个进程最多只能输出512个动态分区，否则会返回异常。
  - 任意动态分区SQL不允许生成超过2000个动态分区，否则会返回异常。
  - 动态生成的分区值不允许为NULL，否则会返回异常。

- 如果目标表有多级分区，在执行INSERT操作时，允许指定部分分区为静态，但是静态分区必须是高级分区，否则会返回异常。

执行MaxCompute SQL过程中，报错Expression not in GROUP BY key，如何处理？

- 问题现象：执行MaxCompute SQL时，报错如下。

```
FAILED: ODPS-0130071:Semantic analysis exception - Expression not in GROUP BY key : line 1:xx ‘xxx’
```

- 问题原因：在GROUP BY子句中，SELECT查询的列，必须是GROUP BY中的列或聚合函数（例如SUM或COUNT）加工过的列。不支持直接引用非GROUP BY的列。详情请参见SELECT语法。

MaxCompute客户端使用-e参数执行SQL时，是否有长度限制？

有长度限制，SQL语句长度不能超过2 MB。

MaxCompute客户端支持并行下载吗？

支持并行下载。

在并行下载时，请注意本地服务器配置、CPU、网络带宽或服务器负载等的情况，以免影响并行下载功能。

在MaxCompute客户端执行SQL时，可以使用自建的ECS调度资源吗？

如果执行SQL语句时，使用的是公共资源，可能会出现等待的情况。添加自建调度资源详情请参见新增自定义数据集成资源组。

MaxCompute单表可以存放的最大列数是多少？

MaxCompute单表可以存放的最大列数为1200列。如果您的列数超过限制，可以参考如下方式处理：

- 对数据进行降维，缩减到1200列以内。
- 修改数据的保存方式，例如设备证书、稀疏或稠密矩阵。

MaxCompute查询得到的数据是根据什么排序的？

MaxCompute中表的读取是无序的。如果您没有进行自定义设置，查询获取的结果也是无序的。

如果您对数据的顺序有要求，需要对数据进行排序。例如，在SQL中需要加上 `ORDER BY xx LIMIT n` 对数据进行排序。

如果您需要对数据进行全排序，只需要将 `LIMIT` 面的 `n` 设置为 `数据总条数+1` 即可。

> **说明** 海量数据的全排序，对性能的影响非常大，而且很容易造成内存溢出问题，请尽量避免执行该操作。

MaxCompute如何非交互式运行MaxCompute SQL？

在操作系统中，您可以通过Shell非交互式运行MaxCompute SQL：

- 使用 `odps -f filename` 方式，读取并处理SQL文件。

  如果运行SQL，Filename文件的第一行是 `SQL` 表示已经进入SQL模式。

```
SQL
 SELECT FROM table_name WHERE xxx;
```

- 如果只运行一个SQL语句，您可以使用MaxCompute SQL中的 `sqltext` 方法，命令示例如下。

  ```
  ./odpscmd -e "SELECT FROM DUAL;"
  ```

  您可以通过 `odps -help` 获得更多的信息。

  此功能可以配合 `crontab` 命令定时执行SQL，建议您使用DataWorks的周期任务功能，详情请参见查看周期任务。

使用MaxCompute SQL自定义函数查询时，为什么提示内存不够？

因为数据量太大并且有倾斜，SQL作业超出默认设置的内存。

执行 `set odps.sql.udf.joiner.jvm.memory=xxxx;` 命令增大内存。

MaxCompute与关系型数据库有什么区别？

- MaxCompute适合海量存储和大数据分析，不适合在线服务。
- MaxCompute SQL的语法是ANSI SQL92的一个子集，并有自己的扩展，与Oracle或MySQL类似。
- MaxCompute表不支持主键、索引和字段约束。
- MaxCompute表不支持UPDATE操作。
- MaxCompute表不支持DELETE操作，只能DROP整个分区或表。在MaxCompute中创建表时，不允许指定字段默认值。
- SELECT操作输出屏显的数据行数受限制，最大为10000条。不支持通过SELECT下载数据，不同于ODBC或JDBC的ResultSet方式。
- 在MaxCompute中需要通过Tunnel、Dship工具或MaxCompute Tunnel SDK导出数据。

MaxCompute支持虚拟表吗？例如MySQL中的DUAL表？

不支持虚拟表，您可以手动创建DUAL表。

在执行MaxCompute SQL时使用动态分区，将GMT格式化作为分区字段，产生大量分区和记录数，一直没有运行完成，是什么原因？

动态分区涉及的分区比较多，数据分发花费时间较多。

MaxCompute能否像MySQL一样灵活使用用户变量（即MySQL的@变量名）？

不支持参数化的SQL。

REGEXP_COUNT函数的参数pattern是否支持嵌入查询语句？

不支持，您可以改写为支持的语法，例如JOIN。

在执行MaxCompute SQL过程中，报错Semantic analysis exception，如何处理？

SQL语句如下。

```
SELECT a.id as id > , IFNULL(CONCAT('phs\xxx', a.insy, '\xxxb\xxx', IFNULL()))
```

报错信息如下。

```
Semantic analysis exception - Invalid function : line 1:41 'ifnull'
```

MaxCompute没有提供 IFNULL 函数导致报错。您需要使用 CASE WHEN 表达式或 COALESCE 命令。SQL调试详情请参见其他函数。

SQL能将MaxCompute的配置转移到另外一个阿里云账号上吗？

您可以通过Package方法实现，详情请参见MaxCompute多团队协同数据开发项目管理最佳实践。

MaxCompute SQL设置过滤条件后，为什么还报错提示输入的数据超过100 GB？

先过滤分区，再取数据。取数据后，再过滤其他非分区字段。输入表的大小是取决于过滤分区过滤后，过滤其他字段前表的大小。

如何处理外部表执行SQL慢的问题？

- OSS外部表中的GZ压缩文件读取慢
  - 问题现象：用户创建了一个OSS外部表，数据源为OSS中的GZ压缩文件，大小为200 GB。在读取数据过程中执行缓慢。
  - 解决方法：此类情况可能是由于Map端执行计算的Mapper数量过少，所以SQL处理慢。
    - 对于结构化数据，您可以设置以下参数调整单个Mapper读取数据量的大小，加速SQL执行。

      ```
      set odps.sql.mapper.split.size=256; #调整每个Mapper读取table数据的大小，单位是MB。
      ```

    - 对于非结构化数据，您需要查看OSS外部表路径下的OSS文件是否只有1个。如果只有1个，由于压缩方式下的非结构化数据不支持拆分，所以只能生产1个Mapper，导致处理速度较慢。建议您在OSS对应的外部表路径下，将OSS大文件拆分为小文件，从而增加读取外部表生成的Mapper数量，提升读取速度。
- 使用SDK搜索MaxCompute外部表数据速度慢
  - 问题描述：使用SDK搜索MaxCompute外部表数据速度慢。
  - 解决方法：外部表仅支持全量搜索，所以较慢，建议您改用MaxCompute内部表。
- 查询外部表Tablestore数据慢
  - 问题现象：查询外部表Tablestore的数据慢，同样的业务数据，1个实时写入Tablestore，1个定时写入MaxCompute，两个表结构和数据量一样。查询MaxCompute内部表耗时远小于查询Tablestore外部表。
  - 解决方法：这种情况可能是对1份数据进行了多次计算，导致速度慢。相比每次从Tablestore远程读取数据，更高效快速的方法是先一次性把需要的数据导入到MaxCompute内部，转为MaxCompute内部表，再进行查询。

SQL语句支持一次添加多个分区吗？

不支持，需要分多次逐个添加。

设置表的生命周期为3天，每个表的分区存储量很大，如何清理分区表旧数据？

设置了生命周期的表超过设定时间没有修改，系统会自动回收。

通过 desc table_name partition(pt_spec) 命令查看旧的分区修改时间是否在生命周期内修改过。通过 desc nginx_log 命令查看生命周期时间，MaxComptue每天17:00点进行回收，DataWorks上的数据显示有延迟，一般会延迟一天。

如何能提高查询效率？分区设置能调整吗？

当利用分区字段对表进行分区时，新增分区、更新分区和读取分区数据均不需要做全表扫描，可以提高处理效率。详情请参见表操作和MaxCompute的分区配置和使用。

是否能将RDS中的表一次性导入到MaxCompute中？如果导入成功为什么物理存储显示是0？

物理存储显示并不是实时同步的，通常次日才可以看到。您可以在DataWorks中使用SQL查看表数据是否正常同步。

如何关闭复制和下载功能？

在工作空间配置中关闭能下载Select结果开关。工作空间配置详情请参见工作空间配置。

使用SQLTask执行SQL查询时，如果查询结果条数大于限制的1000条，该如何获取所有数据？

您可以将SQL查询的结果集写入一张表中，通过Tunnel下载所有数据。详情请参见导出SQL运行结果的方法总结。

SQLTask中，按照如下方法返回结果集的数据量是否有限制？如果有限制，最大返回结果集大小是多少？

```
Instance instance = SQLTask.run(odps, "sql语句");
instance.waitForSuccess();
List<Record> records = SQLTask.getResult(instance);
```

有限制，您可以最多调整到5000。如果数据量比较大，建议您使用Tunnel SDK导出数据。

SQLTask查询数据和DownloadSession在使用及功能上，有什么不同？

SQLTask运行SQL并返回结果，返回条数有限制，默认是1000条。

DownloadSession下载某个存在的表里的数据，结果条数无限制。

如何下载超过一万行的数据？

如果您使用MaxCompute客户端，可以先把SQL结果插入到一张表中，然后使用Tunnel下载表中的数据。详情请参见导出SQL运行结果的方法总结和使用说明。

如何查看SQL执行费用？

您可以使用 COST SQL 命令查询费用，详情请参见计费方式。

MaxCompute SQL中LIKE模糊查询的WHERE条件是否支持正则表达式？

支持，例如 SELECT * FROM user_info WHERE address RLIKE '[0-9]{9}'; ，表示查找9位数字组成的ID。

在执行MaxCompute SQL过程中，报错ODPS-0121145 ，什么原因？

报错信息如下。

```
--报错1。
ODPS-0121145:Data overflow - param convert to Double result is nan, input param is NaN
--报错2。
ODPS-0121145:Data overflow - Div result is inf, two params are 19.000000 and 0.000000
```

报错1产生原因：原始数据中有空值。

报错2产生原因：数据溢出，超出数据类型的值域范围。需要确认数据类型是否正确。

多路输出的情况下，能否在REDUCE函数中获取到每一个Label输出表的表结构？

您可以通过 `result1.getColumns()` 方法获取表的字段信息。

最新版本的SDK信息请参见ODPS SDK Core API，代码需要您自行调试。

使用ROUND函数对DOUBLE类型数据四舍五入，为何结果存在偏差？

- 问题现象：使用 `ROUND` 函数对DOUBLE类型的数据进行四舍五入，发现4.515四舍五入结果为4.51。

  ```
  SELECT ROUND(4.515, 2),ROUND(125.315, 2) FROM DUAL;
  ```

- 问题原因：DOUBLE类型是8字节双精度浮点数，存在一定的精度差。示例中， `4.515` 的DOUBLE类型表示结果为 `4.514999999...` ，因此四舍五入时被计算为4.51。

如果只同步100条数据，如何在过滤条件WHERE中通过LIMIT实现？

LIMIT不支持在过滤条件中使用。您可以先在数据库中使用SQL筛选出100条数据，再执行同步操作。

对表A执行GROUP BY生成表B，表B比表A的行数少，但表B的物理存储量是表A的10倍，是什么原因造成的？

数据在MaxCompute中是列式压缩存储的，如果同一列的前后数据的内容是相似的，压缩比会比较高。当 `odps.sql.groupby.skewindata=true` 打开时，使用SQL写入数据，数据比较分散，压缩比较小。如果希望数据的压缩比较高，您可以在使用SQL写入数据时进行局部排序。

如果一个表有很多分区，如何清空表的所有分区？

删除分区语法如下。您需要逐个删除分区。如果要删除大量分区，建议重建一个新表。

```
ALTER TABLE TABLE_NAME DROP [IF EXISTS] PARTITION partition_spec;
```

MaxCompute客户端的SQL语句执行成功，为什么会打印出异常信息？

- 问题现象：在使用MaxCompute客户端执行MaxCompute SQL语句时，在SQL执行成功的情况下，打印出如下错误信息。

  ```
  com.aliyun.openservices.odps.console.ODPSConsoleException
  ```

- 问题原因：本地计算机开启了网络代理软件。
- 解决方法：退出或者关闭您的网络代理软件。

当目标表的字段类型为VARCHAR(10)，插入数据溢出时会报错吗？

对 VARCHAR(10) 数据类型的字段插入数据时，数据长度溢出时会截断并不报错。

在DataWorks里执行需要传参的SQL时报错，是什么原因？

在开发环境运行需要传参的SQL时，需要单击**高级运行**。

MaxCompute与标准SQL的主要区别是什么？如何解决？

MaxCompute与标准SQL的主要区别及解决方法，详情请参见与标准SQL的主要区别及解决方法。

MaxCompute是否支持ORDER BY FIELD NULLS LAST语法？

MaxCompute不支持此语法。MaxCompute支持的语法请参见与其他SQL语法的差异。

SQL作业运行过慢，如何优化？

SQL作业可以通过Logview进行定位，定位方法请参见使用Logview查看作业运行信息。

优化SQL作业，详情请参见计算优化最佳实践。

MaxCompute的时间类型字段是否可以不带时分秒？

时间类型字段使用DATE数据类型。使用该数据类型时，您需要打开MaxCompute 2.0数据类型开关。详情请参见2.0数据类型版本。

MaxCompute的表有无索引？

没有索引，Hash Clustering可以提供类似数据库中Cluster index的效果，详情请参见表操作。

如何快速查看项目空间下哪些表是分区表？

执行如下命令查看项目空间下的分区表信息。

```
SELECT table_name FROM information_schema.columns WHERE is_partition_key = true GROUP BY table_name;
```

如何将开发环境的表数据同步至生产环境的表中？

执行如下命令。

```
INSERT INTO project.table SELECT * FROM project_dev.table;
```

如果没有生产环境中表的读写权限，需要完成账号授权，详情请参见授权。

查询分区表的WHERE条件是add_months('yyyy-mm-dd',x)，报错is full scan with all partitions，please specify partition predicates，如何处理？

您可以通过 EXPLAIN 命令查看SQL中的分区剪裁是否生效。详情请参见分区剪裁合理性评估。

通过JDBC方式访问MaxCompute可以向MaxCompute中插入数据吗？

您可以通过 INSERT 操作插入数据，详情请参见使用说明。

如何对表的分区数据做LEFT JOIN操作？

详情请参见JOIN。

使用GROUP BY分组查询100亿条数据会不会影响性能？GROUP BY对数据量有没有限制？

无影响。无限制。GROUP BY分组查询详情请参见SELECT语法。

MaxCompute SQL支持WITH AS语句吗？

支持，MaxCompute支持SQL标准的CTE，提高SQL语句的可读性与执行效率。详情请参见COMMON TABLE EXPRESSION（CTE）。

是否可以在DataWorks执行set命令打开2.0数据类型开关？

可以。使用2.0数据类型时，您可以在DataWorks的临时查询模式或ODPS SQL节点下执行 `set odps.sql.type.system.odps2=true;` （Session级别）或 `setproject odps.sql.type.system.odps2=true;` （Project级别）命令。

如何解决DECIMAL数据类型精度溢出问题？

执行 `set odps.sql.decimal.odps2=true;` 命令，打开2.0数据类型开关。

如何判断一个字段是否为空？

您可以通过MaxCompute SQL的运算符判断字段是否为空，详情请参见运算符。

因误操作删除的表可以恢复吗？

不可以。在MaxCompute客户端（odpscmd）和Intellij IDEA中删除表为不可逆操作。请谨慎操作。

如何查看已执行的所有SQL作业？

您可以通过Information Schema服务的TASKS_HISTORY视图查看已执行的所有SQL作业。详情请参见Information Schema概述。

MaxCompute SQL如何实现多行注释？

您可以使用快捷键Ctrl+/实现多行注释，详情请参见编辑器快捷键列表。

查询数据时，报错Semanticanalysisexception-XXXtypeisnotenabled incurrentmode，如何处理？

执行 `set odps.sql.decimal.odps2=true;` 命令，打开2.0数据类型开关。

是否可以通过DataWorks的Shell节点调取MaxCompute SQL？

不可以。Shell节点仅支持标准Shell语法，不支持交互性语法。如果作业较多，您可以使用ODPS SQL节点执行作业，详情请参见创建ODPS SQL节点。

MaxCompute支持修改表字段的数据类型吗？

不支持。只能添加字段列，生产环境中的表不允许删除表字段、修改表字段和分区字段。如果必须修改，请删除表之后重新建表。您也可以创建外部表，删除并重建表后，可以重新加载数据。数据类型详情请参见数据类型版本说明。

除使用UDF外，如何合并两个没有任何关联关系的表？

您可以通过 UNION ALL 运算完成纵向合并。横向合并可以通过 ROW_NUMBER 函数实现，两个表都新加一个ID列，进行ID关联，然后取两个张表的字段。

执行SELECT * FROM XXX ORDER BY XXX;命令报错，如何处理？

ORDER BY必须与LIMIT共同使用。详情请参见SELECT语法。

如何将一行数据拆分为多行数据？

结合使用Lateral View和表生成函数（例如Split和Explode），将一行数据拆成多行数据，并对拆分后的数据进行聚合。详情请参见Lateral View。

执行MaxCompute SQL过程中，报错Parse exception - invalid token 'cost'，如何处理？

使用Java SDK中的SQLCostTask接口查询单条SQL费用。接口使用方式请参见SQLCostTask。

如何删除生产环境的表？

在MaxCompute客户端（odpscmd）或DataWorks的数据开发（DataStudio）中执行如下命令删除生产环境的表。

```
DROP TABLE project_name.table_name;
```

如何连接相同字段？

可以使用WM_CONCAT函数连接相同字段，详情请参见WM_CONCAT。

如何修改表的Hash Clustering属性？

增加表的Hash Clustering属性：`ALTER TABLE table_name [CLUSTERED BY (col_name [, col_name, ...])] [SORTED BY (col_name [ASC | DESC] [, col_name [ASC | DESC] ...])] INTO number_of_buckets BUCKETS;`。

去除表的Hash Clustering属性：`ALTER TABLE table_name NOT CLUSTERED;`。

如何查看指定的表是否存在？

可以使用函数TABLE_EXISTS查看指定的表是否存在，详情请参见TABLE_EXISTS。

如何查看指定的分区是否存在？

可以使用函数PARTITION_EXISTS查看指定的分区是否存在，详情请参见PARTITION_EXISTS。

获取项目空间下的所有表名称报错，如何处理？

使用MaxCompute的元数据服务，详情请参见Information Schema概述。

如何升级表的数据类型？

在SQL语句前增加 `set odps.sql.type.system.odps2=true;` 命令，与SQL语句一并提交。

新创建的项目空间不支持数据类型自动隐式转换，如何处理？

确认是否开启MaxCompute 2.0，关闭MaxCompute 2.0之后可进行隐式转换。详情请参见数据类型转换。

是否支持将非分区表修改为分区表？

非分区表不支持更改为分区表，也不支持增加分区列。您可以重新创建一张分区表，详情请参见表操作。

如何创建视图？

根据创建视图语法编写SQL语句即可，更多创建视图语法信息，请参见视图操作。

如何关闭查询加速模式？

在SQL语句前添加 `set odps.mcqa.disable=true;` 命令，与SQL语句一起提交执行，即可关闭查询加速模式。

在执行MaxCompute SQL过程中，报错ODPS-0121096，什么原因？

- 问题现象：执行MaxCompute SQL的过程中，出现类似如下报错。

  ```
  Failed to run ddltask - Modify DDL meta encounter exception : ODPS-0121096:MetaStore transaction conflict - Reached maximum retry times because of OTSStorageTxnLockKeyFail(Inner exception: Transaction timeout because cannot acquire exclusive lock.)
  ```

- 问题原因：MaxCompute对正在操作的表没有锁机制，该错误是由元数据产生竞争导致。您需要检查是否存在同时多次对表或表分区执行读写操作的情况。
- 解决方法：在MaxCompute还没有锁机制的情况下，不要同时对一张表或表分区执行多次读写操作。

在客户端的odps_config.ini文件中设置use_instance_tunnel=false，instance_tunnel_max_record=10，为什么Select还是能输出很多记录？

需要修改 `use instance tunnel=false` 为 `use_instance_tunnel=true` ，才能通过 `instance_tunnel_max_record` 控制输出记录数。

在MaxCompute SQL中使用JOIN时，分区裁剪条件放在ON中分区裁剪会生效，还是放在WHERE中才会生效？

如果分区剪裁条件置于WHERE语句中，分区剪裁会生效。如果分区剪裁条件置于ON语句中，从表的分区剪裁会生效，主表的分区剪裁不会生效即会全表扫描。更多分区裁剪信息，请参见分区剪裁合理性评估。

如何查询某个用户创建的表？

只查询某个用户创建的表可以使用元数据视图TABLES，通过owner_name字段过滤。更多TABLES信息，请参见TABLES。

如何用正则表达式判断是否为中文？

命令示例如下：

```
select '汉字' rlike '[\\x{4e00}-\\x{9fa5}]+';
```

如何判断给定的表是否存在？

您可以使用 `TABLE_EXISTS` 函数，更多信息，请参见TABLE_EXISTS。

如何收集一个月内访问MaxCompute的用户信息？

您可以使用审计日志功能来查看。MaxCompute会完整地记录用户的各项操作行为，并通过阿里云ActionTrail服务将用户行为日志实时推送至ActionTrail。更多信息，请参见审计日志。

如何在SQL中实现循环？

您可以通过DataWorks的do-while节点实现。

如何在SQL中调用赋值节点？

您可以通过DataWorks的for-each节点实现。

待创建表的字段名与关键字相同，如何处理？

在对表、列或分区命名时如果使用关键字，需要给关键字加 `` 符号进行转义，否则会报错。

如何查看表的行数？

您可以执行 `select count(*) from table_name;` 命令查看分区表或非分区表的行数。

如何查看分区数量？

您可以通过Information Schema的PARTITIONS视图，获取到分区名，进而获取到分区数量。

4.2. UDF

本文为您介绍UDF的常见问题。

- 如何开启编写UDF的权限？
- 自定义函数中用到了FastJson，打包Extract为fastjson-1.2.8.jar，使用该自定义函数时报错java.lang.NoClassDefFoundError是什么原因？
- 编写UDAF时，报错Resolve Annotation Not Found如何处理？
- MaxCompute中是否有函数可以把2017-01-23转化为20170123？
- MaxCompute表的DECIMAL类型如何设置为保留2位小数？
- MaxCompute有类似Group_concat的函数吗？
- 执行定时任务时某个节点运行失败，日志报错skynet_packageid is null，是什么原因造成的？
- MaxCompute是否支持Scipy？
- 如何通过自定义日志打印对UDAF进行线上调试？
- UDAF函数的参数是否支持任意参数类型？
- MaxCompute是否支持Unicode编码？
- UDF可以读取云上资源，是否可以创建写入的方法？UDF是否可以缓存数据？并让其他UDF直接获取缓存数据？
- UDF初始化时，系统是否会分配一个新的JVM来运行此UDF？是否有办法让所有UDF都运行在同一个JVM中？

UDF初始化时，系统是否会分配一个新的JVM来运行此UDF？是否有办法让所有UDF都运行在同一个JVM中？

MaxCompute SQL执行时内部有调度系统，会分配资源。资源分配不支持手动设置，但您可以设置JVM的内存。

UDF可以读取云上资源，是否可以创建写入的方法？UDF是否可以缓存数据？并让其他UDF直接获取缓存数据？

目前还不支持。

如何开启编写UDF的权限？

- 问题现象：编写UDF，注册函数，执行时报错 `FAILED: Do not allow java UDF in project: xxx` 。
- 问题原因：出现上述报错，是因为您没有UDF编写权限，目前编写UDF的权限不是默认开放的。
- 解决办法：如果您需要获得此权限，请通过工单进行申请。

自定义函数中用到了FastJson，打包Extract为fastjson-1.2.8.jar，使用该自定义函数时报错java.lang.NoClassDefFoundError是什么原因？

- 问题现象：报错信息如下。

  ```
  java.lang.NoClassDefFoundError: java/io/Fi
  ```

- 问题原因：FastJson被沙箱拦截了，您可以使用Gson包来实现，详情请参见Java沙箱。

编写UDAF时，报错Resolve Annotation Not Found如何处理？

- 问题现象：编写UDAF的过程中，报错如下。

  ```
  FAILED: ODPS-0140051:Invalid function - com.aliyun.odps.udf.impl.AnnotationParser$ParseError: @Resolve annotation not found.
  ```

- 解决办法：UDAF需要设置 `@Resolve` 。例如， `@Resolve({"double->double"})` 表示这个UDAF传入和传出的参数都是DOUBLE类型。

MaxCompute中是否有函数可以把2017-01-23转化为20170123？

有，请参见字符串函数中的REGEXP_REPLACE函数。

MaxCompute表的DECIMAL类型如何设置为保留2位小数？

建议您使用STRING数据类型，然后使用UDF实现数据的运算。或者您也可以考虑在计算结束后对数据执行 `round()` 命令。

MaxCompute有类似Group_concat的函数吗？

有，请参见聚合函数中的WM_CONCAT。

执行定时任务时某个节点运行失败，日志报错skynet_packageid is null，是什么原因造成的？

主要原因是没有获得具体运行SQL的AccessKey信息，存在以下几种可能性：

- 如果是在数据开发界面直接运行时报错，请检查个人的AccessKey是否存在。
- 如果是在运维中心运行时报错，请检查此任务对应的任务责任人的AccessKey是否存在。
- 如果是在DataWorks的生产环境运行时报错，请检查主账号的AccessKey是否存在，如果不存在请更新主账号的AccessKey。

MaxCompute是否支持Scipy？

目前MaxCompute暂不支持Scipy，但您可以在MaxCompute UDF中运行Scipy，详情请参见在MaxCompute UDF中运行Scipy。

如何通过自定义日志打印对UDAF进行线上调试？

在日常使用中会出现代码在本地IDE环境里调试成功，但是在线上调试结果不符合预期的情况。主要是因为本地IDE里无法模拟多个Worker进行分布式调试UDAF，所以一些缺陷在线上测试时才会暴露。

建议您使用手工打印日志，针对UDAF进行调试，详情请参见基于自定义日志打印的UDAF调试。

UDAF函数的参数是否支持任意参数类型？

目前UDF的参数必须指定类型，不支持任意参数类型。

Hive的GenericUDAF支持任意参数类型，MaxCompute 2.0兼容Hive UDF，示例请参见兼容Hive UDF。

MaxCompute是否支持Unicode编码？

MaxCompute不支持类似 `\u0001` 格式的Unicode编码。如果您需要使用Unicode编码，可以将字符串写入MaxCompute表，通过 `select xxx from table` 的方式将字段传至UDF函数中处理。

如何使用UDF并行处理千万级数据？

您可以通过设置odps.stage.mapper.split.size参数控制Worker的数量，增加作业处理的并发数以提高效率。相同数据量下，该参数设置的越小，Worker数量越多，效率越高。详情请参见SET操作。

5.PyODPS

本文为您介绍PyODPS的常见问题。

PyODPS节点是否支持Python 3?

支持。详情请参见创建PyODPS 3节点。

使用PyODPS统计表中某个字段的空值率时，用EXECUTE_SQL还是DataFrame，哪个性能更高?

DataFrame聚合性能更高，推荐您使用DataFrame执行聚合操作，详情请参见聚合操作。

如何使用PyODPS下载全量数据?

PyODPS默认不限制从Instance读取的数据量。但是对于受保护的项目，您通过Tunnel下载数据将受限。此时，如果未设 `options.tunnel.limit_instance_tunnel` ，系统会自动打开数据量限制，可下载的数据量受项目限制，通常为10000条。如果您需要手动限制下载的数据量，可以通过 `open_reader` 方法增加LIMIT选项，或设置 `options.tunnel.limit_instance_tunnel=True` 。

通过PyODPS的DataFrame处理数据时，资源是如何使用的?

只有MaxCompute的对象才会调用MaxCompute资源。通过PyODPS的DataFrame处理数据时，使用MaxCompute的分布式计算能力，系统会将数据提交至MaxCompute集群，调用MaxCompute资源进行计算。执行 `df` 操作会使用到内存，例如下载数据。

如何确定PyODPS运行在服务端还是客户端?

在使用PyODPS会出现对内存和数据大小的限制，但这些并非是PyODPS的限制，其本质是对客户端DataWorks PyODPS节点的限制，而PyODPS运行在服务端对内存和数据的大小是没有限制的，但是很多人无法确定PyODPS到底是运行在服务端还是客户端。

- 发生在客户端的行为

  PyODPS在客户端运行有以下三种情况：

  - 情况一：PyODPS DataFrame转化成pandas的DataFrame。
  - 情况二：SQL执行时在服务端，将结果进行遍历时在客户端。

    代码示例如下。

    ```
    result = o.execute_sql('select * from my_new_table;',hints={'odps.sql.allow.fullscan': 'true'})
    with result.open_reader() as reader:
      for record in reader:
        print record[0],record[1]
    ```

  - 情况三：head()、tail()等查看结果的函数。

- 发生在服务端的行为

判断PyODPS运行在服务端的标准是代码是否可以编译成SQL运行。如果代码最终编译成SQL运行，则PyODPS运行在服务端。

PyODPS在服务端运行有以下三种情况：

- 情况一：直接使用SQL语句执行。

```
result = o.execute_sql('select * from my_new_table;',hints={'odps.sql.allow.fullscan': 'true'})
```

- 情况二：代码转化成PyODPS的DataFrame并使用PyODPS DataFrame的算子。PyODPS DataFrame做运算的算子最终会编译成SQL执行。
- 情况三：DataFrame使用用户自定义函数处理数据。用户自定义函数最终也会编译成SQL执行。