旁路导入方案

## 需求

一般新项目上线时会从云梯导入上百G数据，甚至上T的数据，etao的很多项目基本都要求每天全量导入数据到OB中，OB在线导入新表数据等需求都需要短时间内向OB导入大量数据，直接从ups中导入数据比较慢，导入100G左右的数据就需要3个多小时，使用极不方便，猜你喜欢项目每天需要全量导入300G数据，从ups快速导入这么多数据几乎不可能，实现旁路导入数据的功能很迫切。当OB加入二级索引功能后，需要OB能加载离线创建的索引表。

当前OB不支持删除表操作，一旦在switch schema时删除了表，就必须清除rs的commit log并重启rs，运维很不方便。

## 功能设计

旁路导入的基本思想是将需要导入的数据转换成OB内部的SSTable文件，并分发到cs的服务器上，然后在每日合并时或者rs统一发起加载操作。

一般情况下，需要导入的数据在云梯中，在云梯中生成OB的SSTable文件，然后cs从云梯中将SSTable文件拖到本地磁盘并加载。

1. 在OB非每日合并期间，从rs获取tablet range信息,或者为新表预分配tablet range，将tablet range转换成云梯做map-reduce的配置文件，在云梯中按照OB的tablet range将数据分片不同的文件。
2. 在云梯的map阶段，按照应用逻辑和tablet range分配，生成与tablet range相对应的中间文件，在reduce阶段，将中间文件转换成SSTable文件，通过jni的方式调用OB生成SSTable文件的动态链接库，这个动态链接库提供一个函数，append（char\* buf， const int64\_t buf\_size, bool is\_first\_row, bool is\_last\_row），buf中的数据以行为单位，buf可以包含一行或多行，列数据用\1分隔，这个动态链接库可以用现在的data\_builder工具稍加修改完成，这个动态链接库将行数据转化成SSTable的数据格式，并将数据写到STDOUT中，云梯使用streaming接口从STDOUT中读取数据并写入HDFS中。Reduce生成的SSTable的命名包括三部分“prefix+table\_id+tablet range”，prefix有云梯的逻辑决定。
3. 从云梯将SSTable文件拷贝到gateway上，禁用rs的tablet balance功能，从rs中获取root table，然后按照root table中tablet range与cs的对应关系，将每个SSTable分发到多个cs上。如果是加新表，按预分配的tablet range均分到多个cs上。
4. 如果添加新表并导入数据，先修改schema并switch schema，rs可以比较简单的加载预分配的tablet range到root table，不需要新建一个root table，减少工作量。如果需要加载在静态数据上创建索引表的SSTable文件，在每日合并前，先删除这些索引表，先修改schema并switch schema，每日合并后，这些索引表的前一个版本的数据都删除了，然后再在静态数据上生成索引文件，并加载。
5. 当SSTable文件已经在cs上准备好， cs在每日合并过程中加载新的SSTable文件，在每日合并开始前，cs扫描磁盘并加载新的数据，每个新的SSTable文件都有一个固定的前缀，比如”new”，便于cs识别，从SSTable文件名中获得table id和tabletrange信息，并为新的SSTable文件分配一个sstable\_id，修改文件名，然后加载该SSTable文件。如果某个表的新tablet数据需要替换旧的tablet或者加入的新表没有动态数据需要合并，新加载的tablet的版本号与当前cs做每日合并的版本号相同，先将没有修改的表按原有逻辑做每日合并，新加载的tablet不做每日合并，其它的每日合并逻辑保持不变；如果导入新表的静态数据并且ups中有该表动态数据，新加载的tablet的版本号与当前cs服务的tablet版本号相同，在每日合并时将新加入的tablet信息加入到当前服务的tablet image中，然后按照原有逻辑进行每日合并，新加的tablet也需要做每日合并。
6. 合并完成后，tablet导入也就完成了，开启rs的tablet balance功能，如果新导入的tablet大小不均匀，可以再发起一次每日合并，使tablet发生分裂，使tablet大小更均匀，tablet迁移会让每个cs上每个table的tablet个数更均衡。转入步骤1开始新的一天的数据导入

每日合并后，在静态数据上创建索引表的SSTable文件，采用map-reduce的方式在对应的磁盘上生成SSTable文件，然后在rs中加入新索引表，switch schema并加载新索引表的tablet range信息到root table中，然后rs通知cs加载索引表的tablet文件，cs将新加入的tablet信息加入到当前服务器的tablet image中，新创建的索引表的tablet版本号采用当前cs服务的tablet版本号，这个过程中需要将新加入的tablet信息加入原有的tablet index文件中。

### 2.1全新系统上线导入数据

使用旁路导入功能实现全新系统上线是可行的，需要先预分配tablet range，在云梯上按照预分配的tablet range进行数据分片，然后采用上面的步骤几乎全自动化的将数据导入OB中。在静态数据导入的过程中，ups可以提供写服务器，在静态数据分布到cs后，可以发起每日合并，经过几次每日合并后，tablet会发生分裂迁移，每个cs的tablet数据相对均匀，tablet的大小也相对均匀。

### 2.2在线建表并导入新表的静态数据和动态数据

1. 修改schema，并switch schema完成后，如果是从其它系统迁移到OB，这时动态数据开始写入，需要在这一时间点开始dump动态数据写入前的全量数据到云梯。

2. 原始数据已在云梯上，使用上面提到的旁路导入流程导入新表数据，在整个导入数据过程中，新表不能提供读服务，只能提供写服务。在导入数据的过程中，如果向OB读取数据只能读取到动态修改的数据。所有的CS也不是在同一个时刻加载新表的tablet信息，在导入数据过程中每个客户端读取同一个rowkey的数据可能不一致，所以导入数据过程中只能人为限制读服务。一般的业务也不会数据还没有导入完成就开始读服务，但对于OB内部加载索引表的情况，需要OB逻辑保证在索引表没有导入完成之前，索引表不能提供读服务。

3. 导入新表数据完成后，OB开始对该新表提供读服务。

### 2.3在线将已存在的表的新版本数据替换旧版本数据

在非每日合并期间采用上面提到的旁路导入流程在云梯中生成SSTable文件并拷贝到各个cs上，然后发起每日合并将tablet数据导入到OB中，在整个导入的过程中不能修改schema，不能做tablet迁移，在每日合并后才能修改schema和生效rs的tablet balance功能。

### 2.4删除表

删除表操作在cs和ups当前已经支持，当switch schema后，ups获得最新的schema并生效，那么已删除的表就不能再更新了，但原有的数据仍然在memtable中，而且还会转储。Cs在每日合并时，会从rs获取最新的schema，如果表已经删除，每日合并就不会合并相应的表。删除的表的旧数据会被回收掉。

现在删除表功能主要问题出在rs上，rs在switch schema后，并没有将root table中已经删除的表的tablet range数据删除，导致了cs合并后汇报的tablet信息没有包含删除的表的tablet信息，rs误认为合并不成功。Rs修改了这个逻辑，删除表功能就可以了，运维就更容易了。

Rs删除表的tablet range信息后，客户端和ms可能还缓存了该表的tablet range信息，如果读取该表的数据，ups和cs都会正常返回数据，这个行为有点奇怪。但一般删除表时，基本不会再访问该表。

## 主要工作

1. Rs实现一个接口能接受外界命令，然后通知cs导入磁盘上准备好的新tablet数据，并告知cs新tablet数据的table\_id和版本号，这个主要用于在每日合并后在静态数据上创建索引表并加载，暂时还用不到这个接口。
2. Cs能识别rs发过来的数据加载命令
3. Cs能够在每日合并过程中，扫描磁盘，发现并加载新tablet数据，对新版本数据替换旧版本数据和加入的新表没有动态数据需要合并的情况，测试能跳过每日合并，其它的表仍然进行每日合并。对于加入新表tablet数据并有动态数据需要合并，能将tablet信息加入到当前服务的tablet image中，tablet index也需要加入新的tablet信息。Cs能在每日合并后加载索引表的tablet。
4. Rs修复删除表的bug，在switch schema时发现需要删除表，需要将该表对应的tablet range信息删除。
5. 修改data builder工具，提供一个append接口，并编译成动态链接库给hadoop的格式转换使用。
6. 实现从云梯拷贝数据到gateway和从gate way拷贝数据到cs的自动化工具。