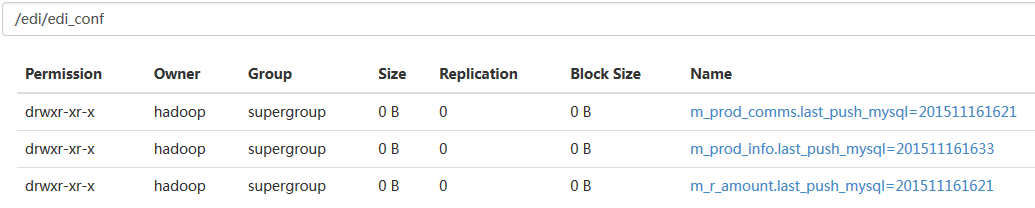
# EasyDataInsight项目

#### Hive数据表同步到Mysql的实现

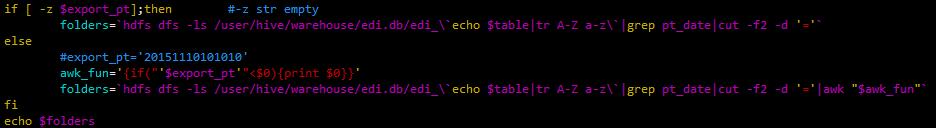
1. 前提
   1. Mysql表名若为M\_A，则对应hive表名应为EDI\_A，Mysql 临时表名为T\_A；
   2. M\_A字段类型和个数与SELECT \* FROM EDI\_A一致（包含分区字段），而T\_A字段个数与类型是EDI\_A中除分区字段以外的字段。
   3. 执行同步脚本的主机能访问mysql DB server.
2. 步骤如下：
3. 检查哪些需要同步过去
   1. 情况一：从hdfs://edi/edi\_conf中取到要导出表对应的最后同步时间last\_sync\_dt；用该时间在hdfs的该表数据路径下找分区日期在last\_sync\_dt之后的所有分区，这些分区的数据都是需要同步到mysql的。
      1. 配置列表



* + 1. 若存在，取得最后一次同步的分区名保存到export\_pt

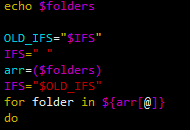


* + 1. 取到需要导出的分区到folder变量

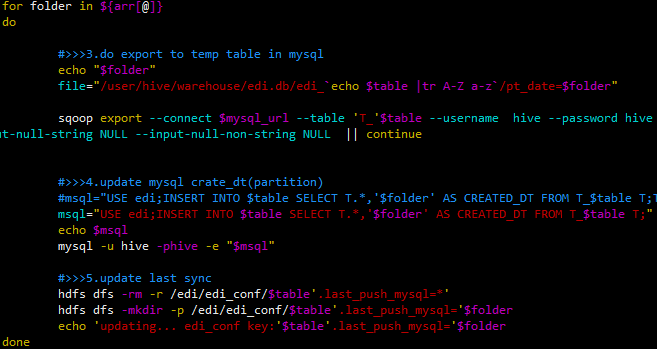
图一

* 1. 情况二：从hdfs://edi/edi\_conf中没有取到last\_sync\_dt，表示从来没同步过该表，所以要同步的数据就是该表在hdfs路径下所有分区。如图一。

1. 开始同步
   1. 循环所有分区的文件列表，每次同步一个分区的数据;
   2. 读取当前操作hive table需要同步到mysql的分区列表folders，遍历folders;



* 1. 导出一个分区
     1. 第一步：取到需要导出的分区文件
     2. 第二步：用sqoop1.4的sqoop export命令导出该分区文件到mysql对应的T\_XX临时表（勿删）；
     3. 第三步：命令行执行mysql命令，将T\_XX表的数据加PT\_DATE字段后转移到XX表，并清空T\_XX表；
     4. 第四步：更新当前同步表的最后同步分区为刚成功导出的分区编号。
  2. 继续循环下一个分区，直到当前表分区都已导出；
  3. 继续循环下一个需导出的hive表。



1. 结束。打印成功提示及执行时长。