**网易：大数据工程师实验报告**

云课堂昵称：火影启源 实验日期：2019-02-25

一、实验题目

大数据调度框架（大作业）

二、实验要求

**【第一步】通过flume教学视频中给定的文件地址，启动flume将文件自动同步到HDFS**

同步要求：

1、将文件归档的事件设置为1个小时

2、归档文件存储格式设置lzo

3、HDFS的文件按天分文件夹进行存储，不能全部同步到一个文件夹中

**【第二步】配置mr任务并依赖flume任务**

1、通过教学视频里面的代码，将日志文件解析到hive表中每天对应的分区中（具体的操作步骤不限制有学员自己设计）

2、查看每天分区日志里面是否有其他天的日志

3、假设当天分区存在其他天的日志，请设计一个方案保证当天的分区只能有当天的数据，并且保证所有的数据不会有丢失

（不能每次都扫描所有的日志）

4、选做：将日志解析的文件存储为parquet文件类型

**【第三步】配置sqoop任务，mysql同步到hive**

1、通过视频中的mysql配置，通过sqoop命令行显示mysql里面有多少数据表

2、通过azkaban配置sqoop任务，将product、member、order三表同步到hive的表（或者全库同步）

同步要求：

（1）每天设置全量同步相关数据

（2）每天设置一个新的分区

（3）同步思考（不强制要求作答，强烈希望同学作答，教学知识是可以覆盖的）

如果需要你写一个方案，只要指定数据源数据库，数据源表 和 输出源hive库 输出源表名， 通过你的方案可以实现所有表在一个任务中就可以实现同步，并且每个map的数据是相对均匀的

**【第四步】hive的相关调度**

1、计算每天有多少的pv、uv、订单量和收入,注册用户数（如何一个sql语法全部查出来）

（1）pv uv计算从日志里面出

（2）订单量和收入：金额在数据库order表，订单量在pay事件

（3）注册用户数：register事件的不重复用户数

（4）将结果存入新建hive表，表名字段名，设计格式可以自己定

2、计算访问product页面的用户中，有多少比例在30分钟内下单并且支付成功对应的商品

3、更改第二个sql，需要做到 支持变更不同的页面类型（正则）和目标事件，支持指定时间间隔的转化率分析（需要写明设计思路）

4、通过sql 计算每个商品的每天的pv，uv 并存入新建hive表，表名字段名，设计格式可以自己定义

5、计算每个商品每天的pv，uv的环比情况（今天-昨天/昨天），并且筛选出环比增长最大和环比增长最小的（负数小于正数）

6、计算每天的登录用户数中新老用户占比，并且统计新老用户分别的pv uv

7、设计一个udaf， 将用户当前行为之前30分钟的actionid行为存为一个list，其中list中的序号代表他时间由远到近的时间序（参考collect\_set的udaf）

8、基于udaf 重写第二个作业的sql，对比有何不同

**【第五步】配置sqoop任务将hive任务结果导入mysql数据库**

1、配置第四步的第四道题的数据，将每天的数据同步到mysql的表中

**【第六步】（强烈建议你设计）：**

1、基于你现有所学的所有内容，设计你自己的框架

题目：框架使用方需求如下：

（1）业务方不想知道你怎么同步的，我只想在使用的时候指定数据源数据库，数据源表 和 输出源hive库 输出源表名， 你能在第二天自动帮我同步全量信息

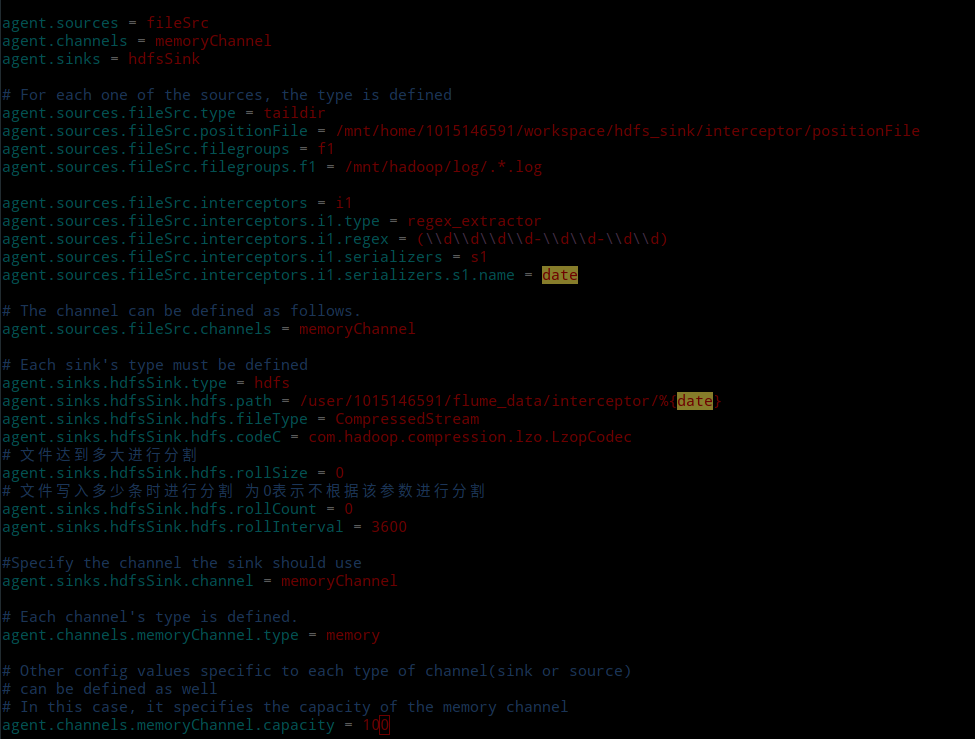
（2）日志解析来说，业务方不想知道你是通过什么引擎实现的，他只希望写个输入路径  输出hive表和分区，还有解析逻辑， 其他过程系统自动帮我完成

（3）业务方不想写add jar这样的方式创建临时函数， 你如何在设计一个规范 保证业务方可以直接使用你生成的udf， 可以注册为常用函数，也可以以其他方式

（4）sqoop导入数据库太慢了， 如何设计一个在任务中，直接将结果写入数据库，业务方不想知道中间细节，只希望hive查询结果可以直接插入数据库（不做转化）

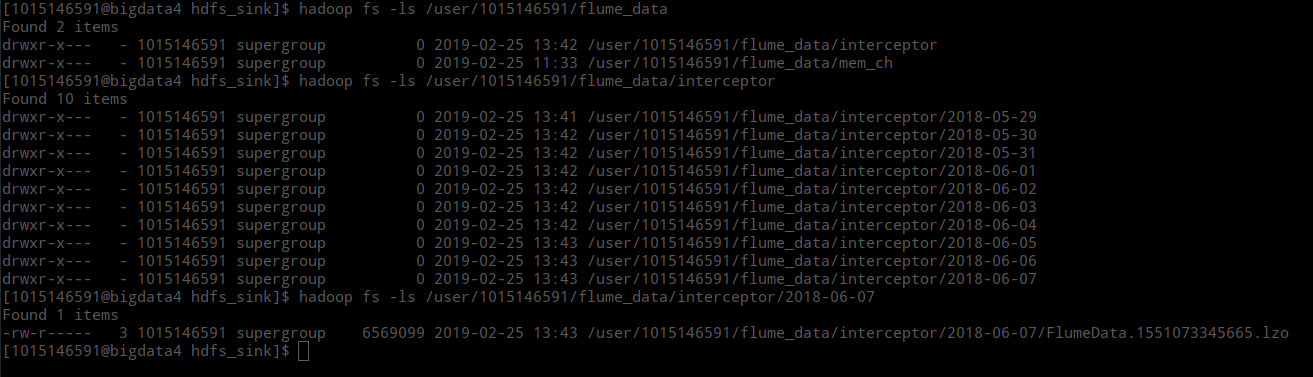
三、操作步骤

**【第一步】通过flume教学视频中给定的文件地址，启动flume将文件自动同步到HDFS.**

1. 编写Flume配置文件flume，配置文件内容如下：2. 执行flume：

./bin/flume-ng agent --conf conf --conf-file conf/flume-homework.properties --name agent

3. 执行完毕后查看HDFS上的文件：

可以看到Flume已经将日志按日期存储到HDFS上了。

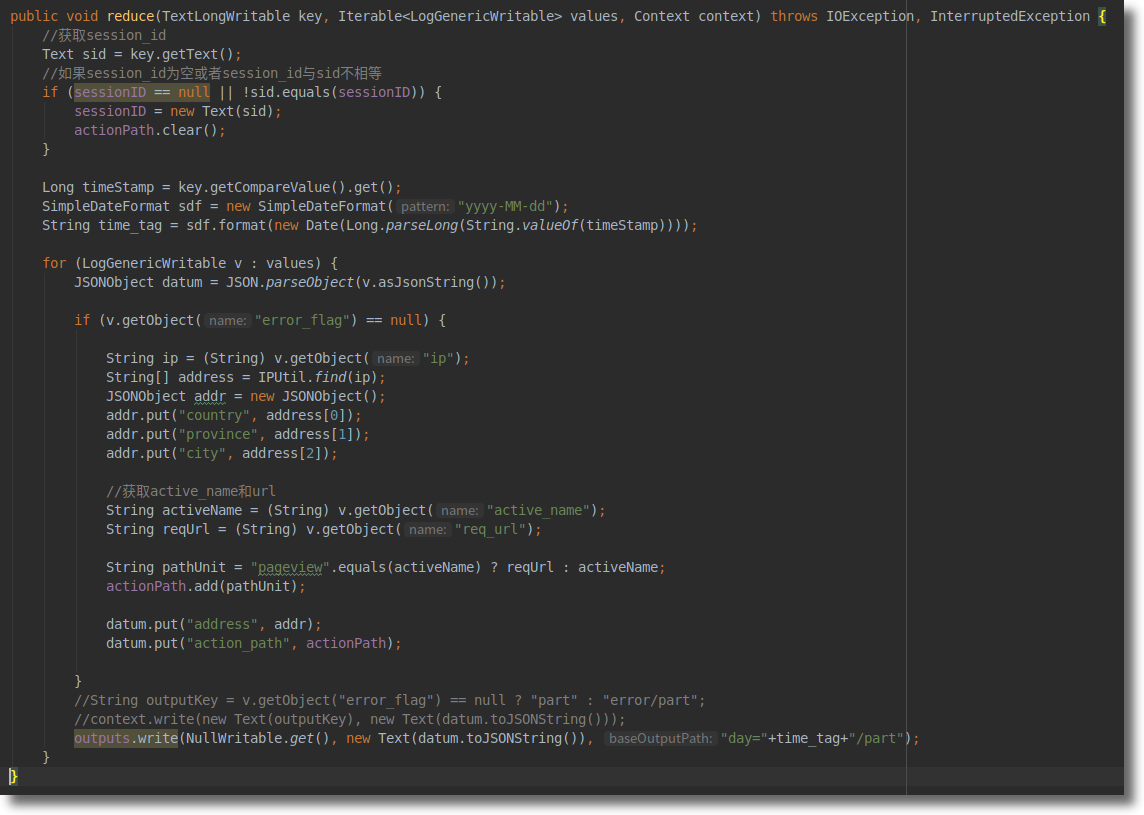
**【第二步】配置mr任务并依赖flume任务**

**1. 通过教学视频里面的代码，将日志文件解析到hive表中每天对应的分区中（具体的操作步骤不限制有学员自己设计）**

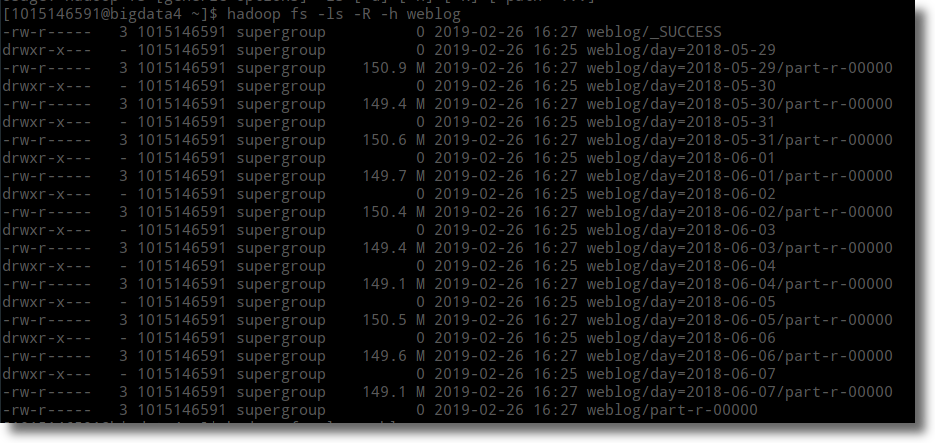
1.1 修改MR程序，使其能够处理/user/1015146591/flume\_data/interceptor目录下的所有文件，并以日期为文件夹名称输出相关日期数据的处理结果到对应的文件夹中。

使用org.apache.hadoop.mapreduce.lib.input.MultipleInputs和org.apache.hadoop.mapreduce.lib.output.MultipleOutputs分别处理多文件输入和多文件输出，部分代码如下。

reduce函数中，使用MultipleOutputs根据日志时间，将解析后的日志存输出到不同的文件夹中：

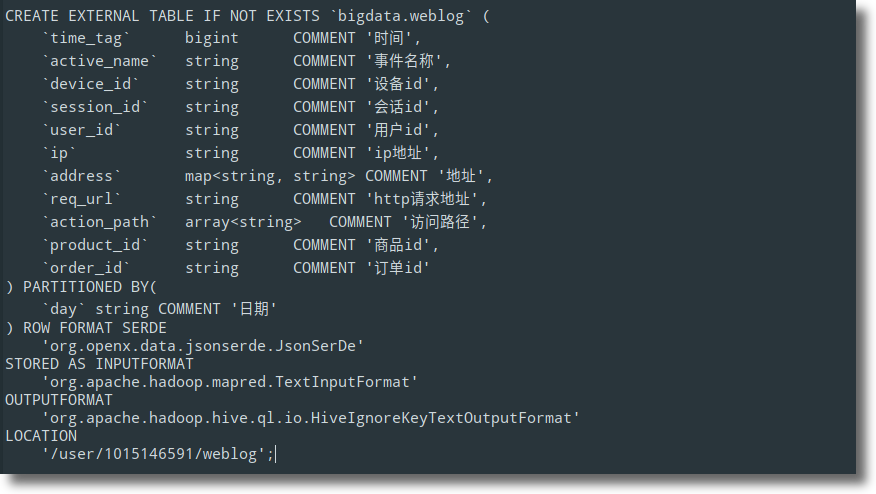
配置多个输入文件：

运行MR程序，查看输出目录weblog是否有正确的日志输出：

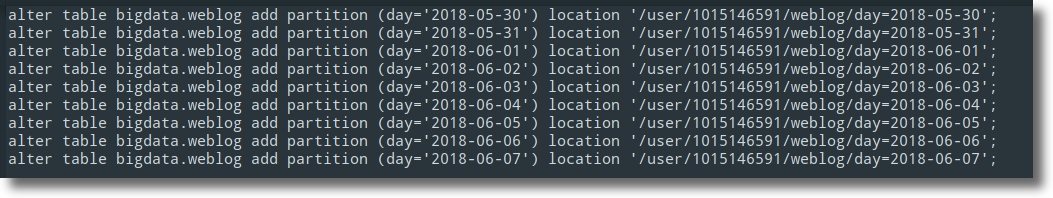


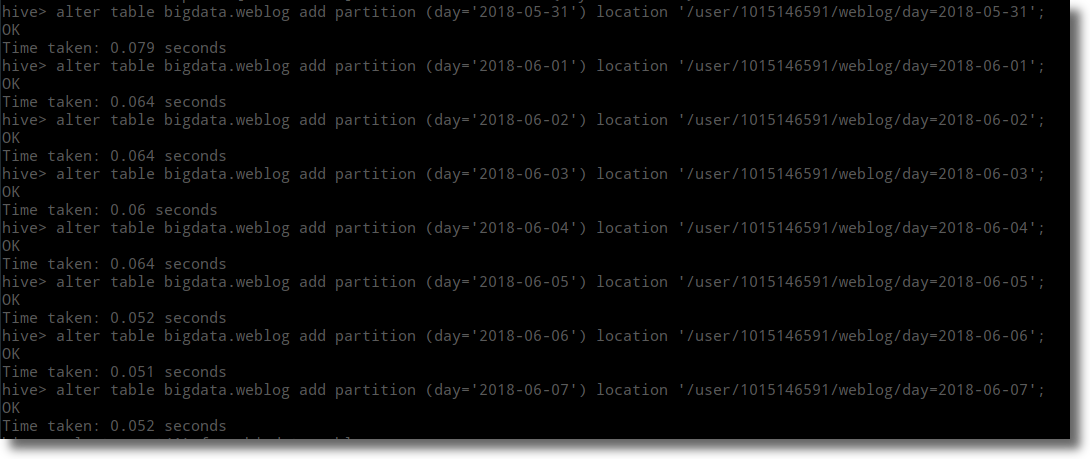
1.2. 创建hive数据库和数据表

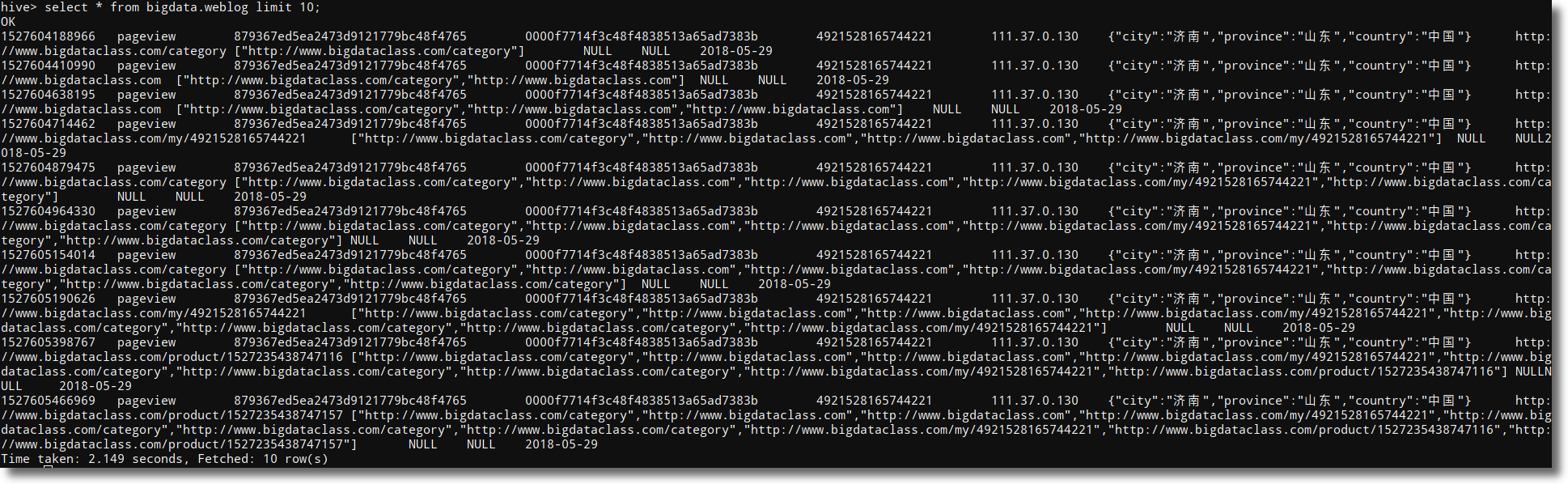
在本地/mnt/home/1015146591/workspace/homework\_hive目录下启动hive，创建bigdata数据库和weblog数据表：

查看数据表结构：

1.3. 将存储于HDFS上解析后的数据根据partition分区加载到数据库中：

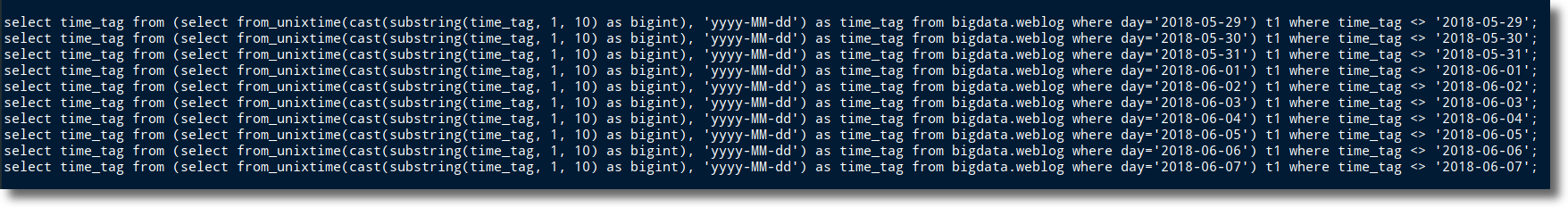


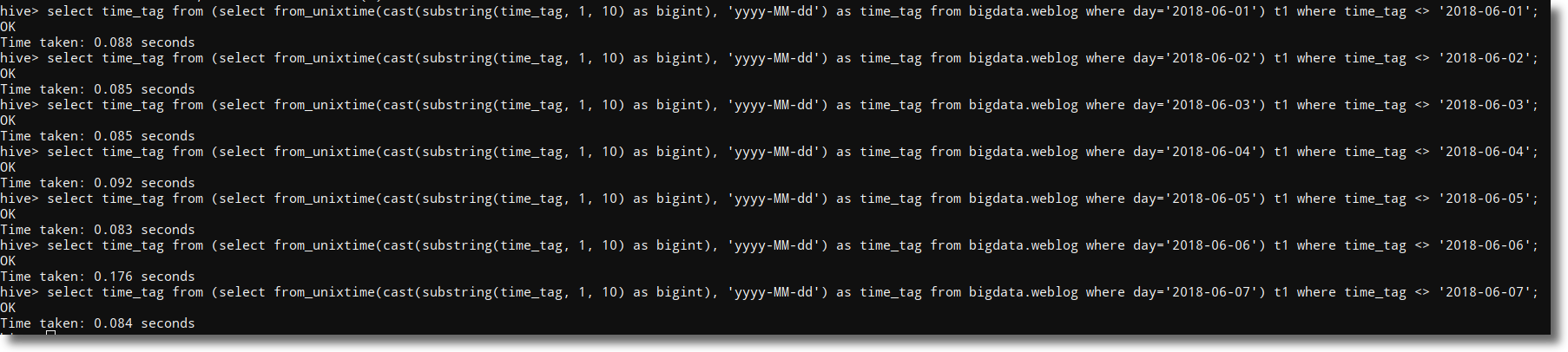
查看数据路中是否已经加载到了数据：

可以看到，hive中已经正确加载到了数据。

**2. 查看每天分区日志里面是否有其他天的日志**

SQL语句如下：

查询结果如下：

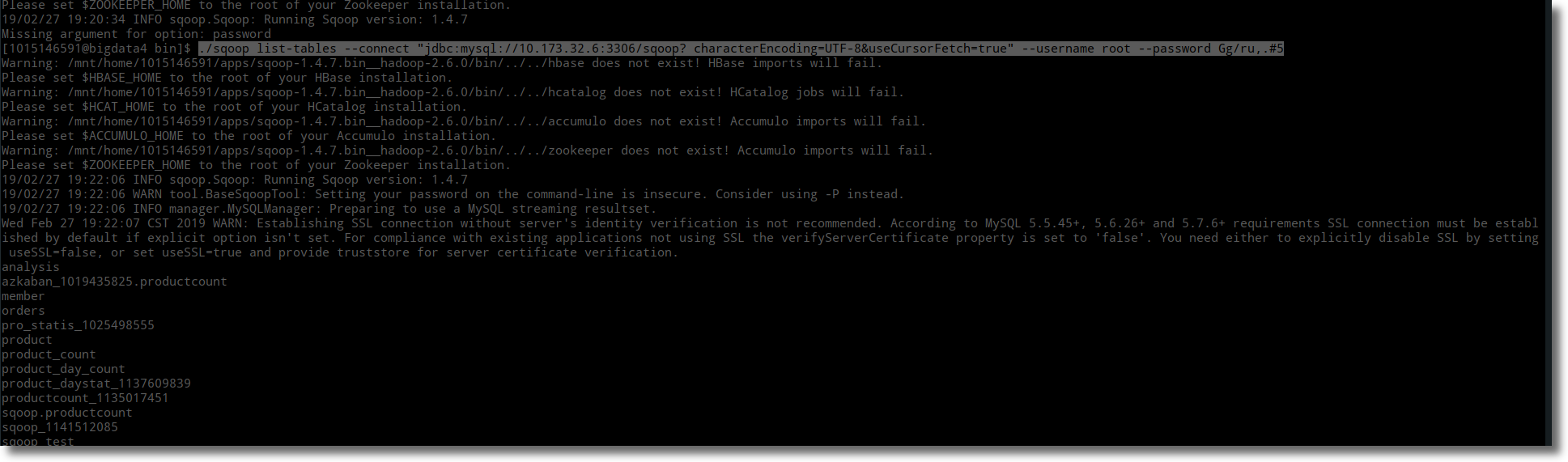
可以看到每天分区日志里面没有其他天的日志。

**3. 假设当天分区存在其他天的日志，请设计一个方案保证当天的分区只能有当天的数据，并且保证所有的数据不会有丢失。**

可以通过编写sql代码，首先检测出各个分区中属于其他天的日志数据，然后将当天分区内的其他天的日志转移到对应的分区，这样的话就能保证数据的不丢失，同时当天分区只存有当天的数据。

**【第三步】配置sqoop任务，mysql同步到hive**

1. 通过sqoop命令查看mysql里有多少数据表：



**【第四步】hive的相关调度**

**【第五步】配置sqoop任务将hive任务结果导入mysql数据库**

**【第六步】（强烈建议你设计）：基于你现有所学的所有内容，设计你自己的框架**

四、实验结果

（此处填写你的程序运行结果）

五、总结反思

（此处填写你的收获、遇到的问题及反思）

六、作业提交

1. 请将本文档按照 【云课堂昵称+网易大数据—课程名称】的方式命名；

（比如：沐沐+网易大数据—数据存储：HDFS）

2. 请将本文档、数据文件打包以附件形式上传到课程作业部分