实验题目

结构化查询：Hive

实验要求

1、安装hive客户端

2、通过mr程序将日志中的所有字段解析到额外的HDFS目录

3、通过hive命令行将hive表映射到对应的HDFS文件

4、查询日志条数及排重用户数结果（检查与mr程序结果是否相同）

5、将过程截图写入作业文档，并提交附件

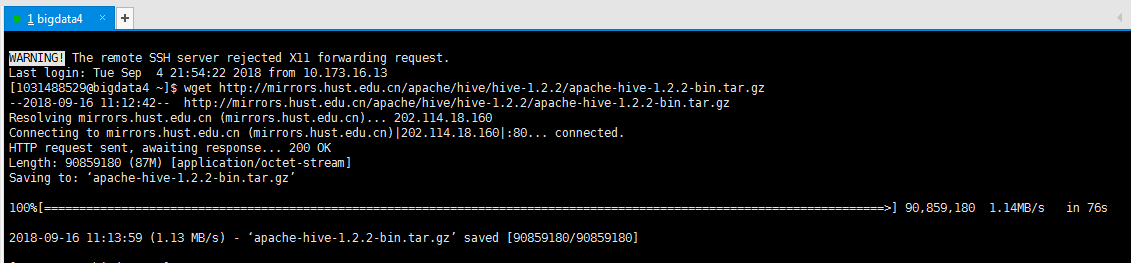
操作步骤

1，下载并解压、配置安装：

主要命令如下：

下载：

[1031488529@bigdata4 ~]$ wget <http://mirrors.hust.edu.cn/apache/hive/hive-1.2.2/apache-hive-1.2.2-bin.tar.gz>



解压：

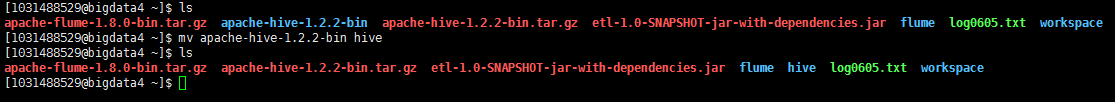
[1031488529@bigdata4 ~]$ tar -xzvf apache-hive-1.2.2-bin.tar.gz

tar -zxvf apache-flume-1.8.0-bin.tar.gz

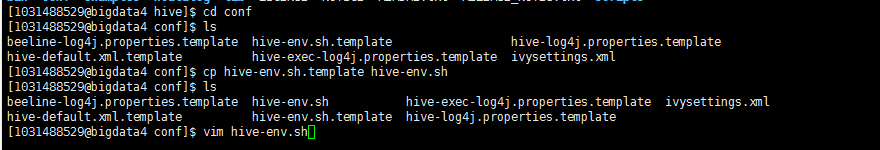


重命名：

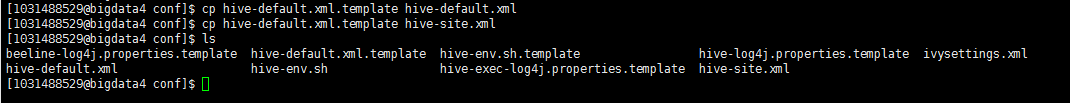
[1031488529@bigdata4 ~]$ mv apache-hive-1.2.2-bin hive



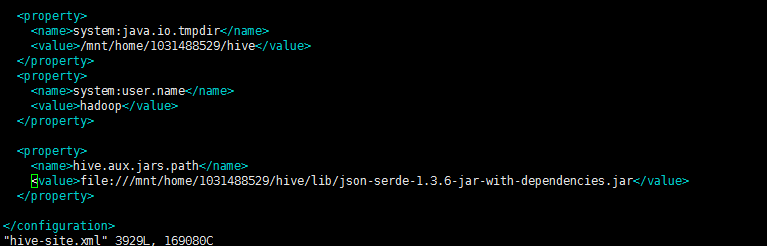
配置hive-env.sh：





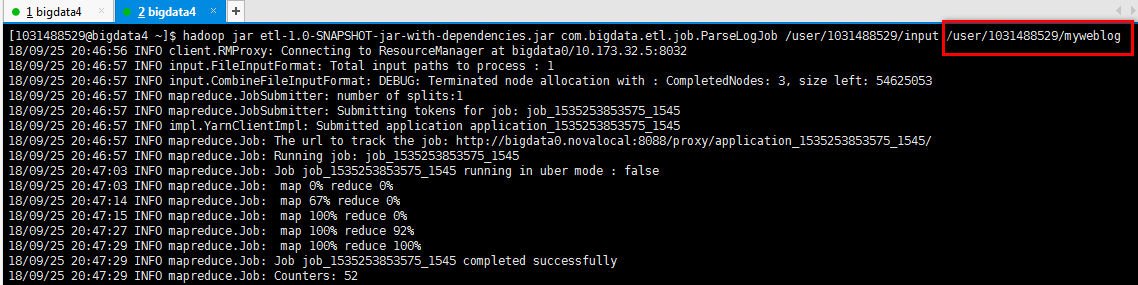


编辑hive-site.xml，在末尾增加属性，**注意tempdir要配置到自己的目录（用户名，也要改为自己的账号，否则建内部表时权限有问题）**



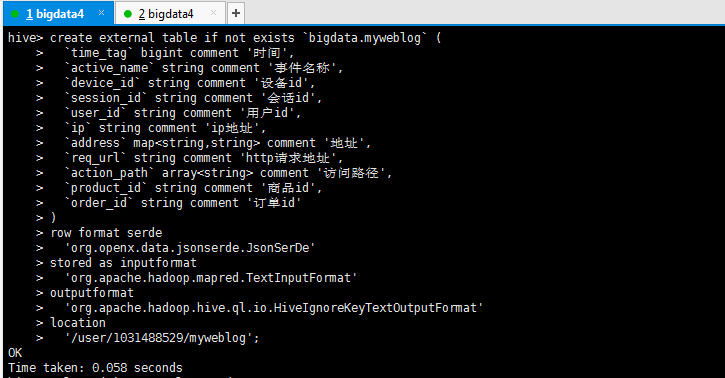
2，通过MR程序将/user/1031488529/input的日志文件解析到/user/1031488529/myweblog目录，解析日志总记录数为227785，去重用户数为10085，其中3条解析错误：

hadoop jar etl-1.0-SNAPSHOT-jar-with-dependencies.jar com.bigdata.etl.job.ParseLogJob /user/1031488529/input **/user/1031488529/myweblog**



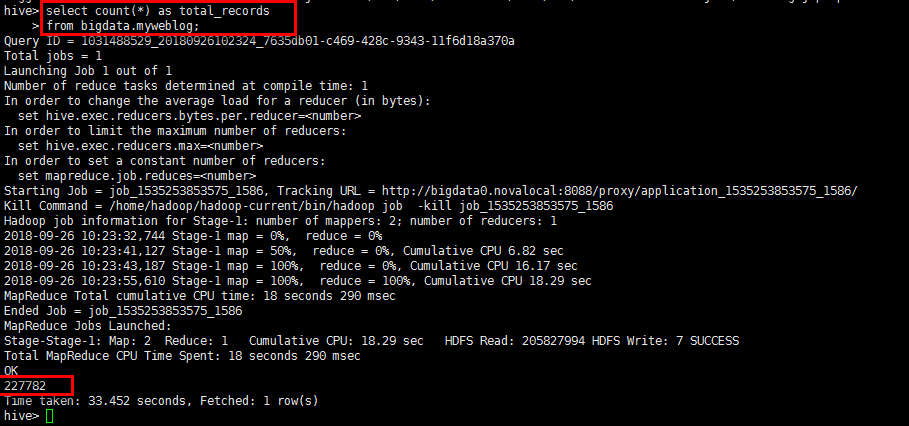


3，建立日志表bigdata.myweblog：映射到上面的日志文件（未使用分区）



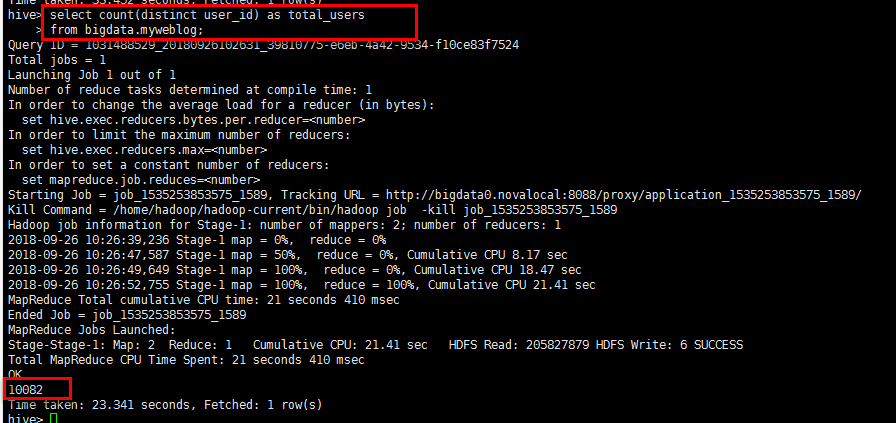
4，查询日志表总记录数：结果为227782

select count(\*) as total\_records from bigdata.myweblog;



查询日志表去重用户数：结果为10082

select count(distinct user\_id) as total\_users from bigdata.myweblog;

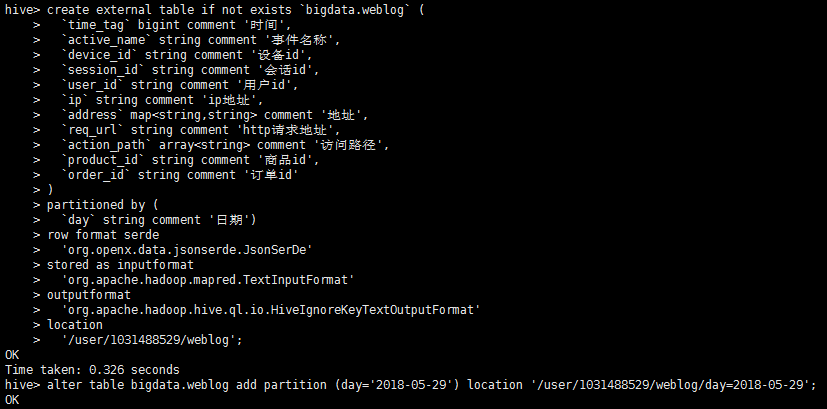


**除去解析错误的记录，日志表的总行数和去重用户数与日志文件一致；**

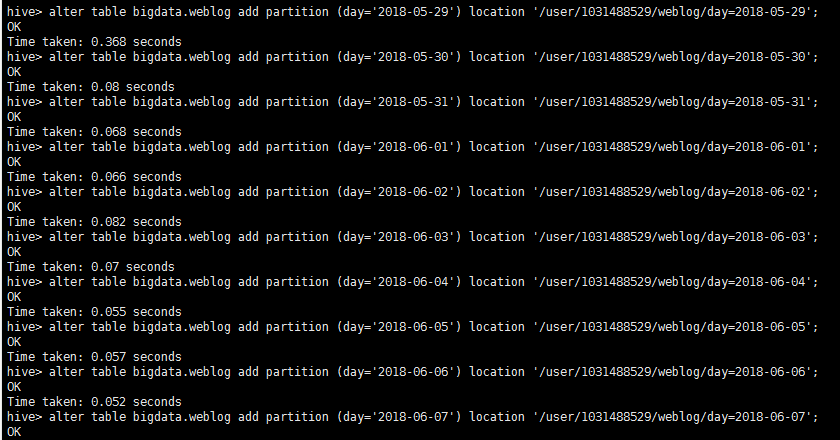
**以下为教学视频操作，仅供参考：**

将user/hadoop/weblog/下的文件复制到自己的hdfs目录/user/1031488529/weblog

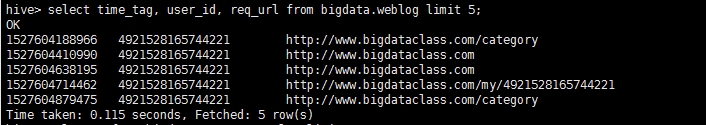
然后，创建日志表



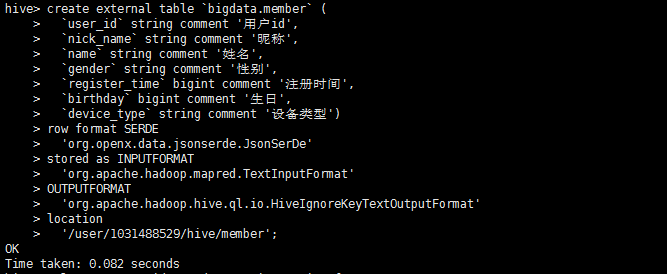
创建日期分区



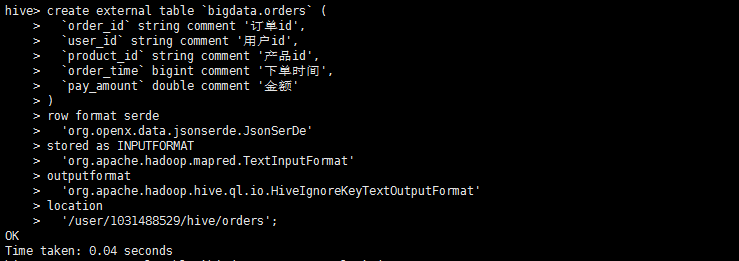
测试日志数据：



创建用户信息表



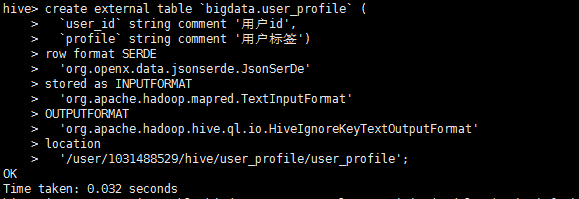
创建订单表：



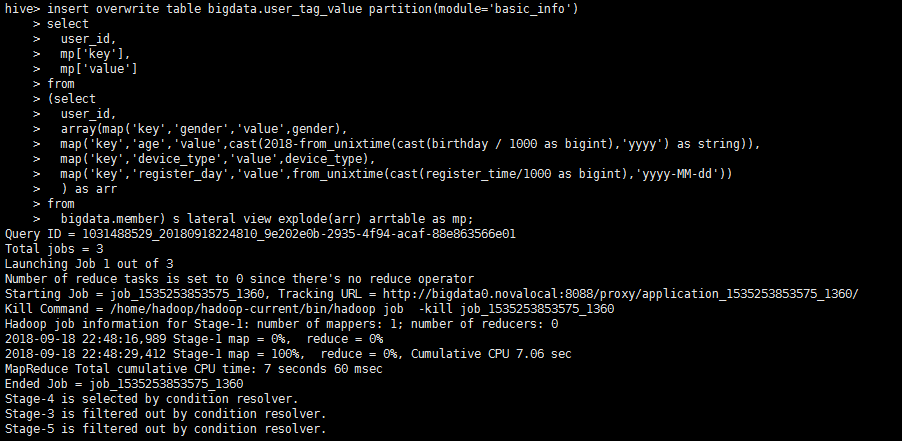
创建中间表：

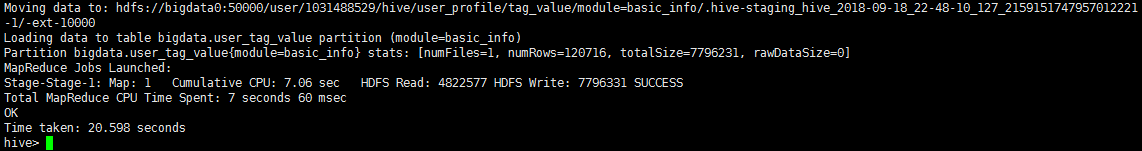


创建大宽表：

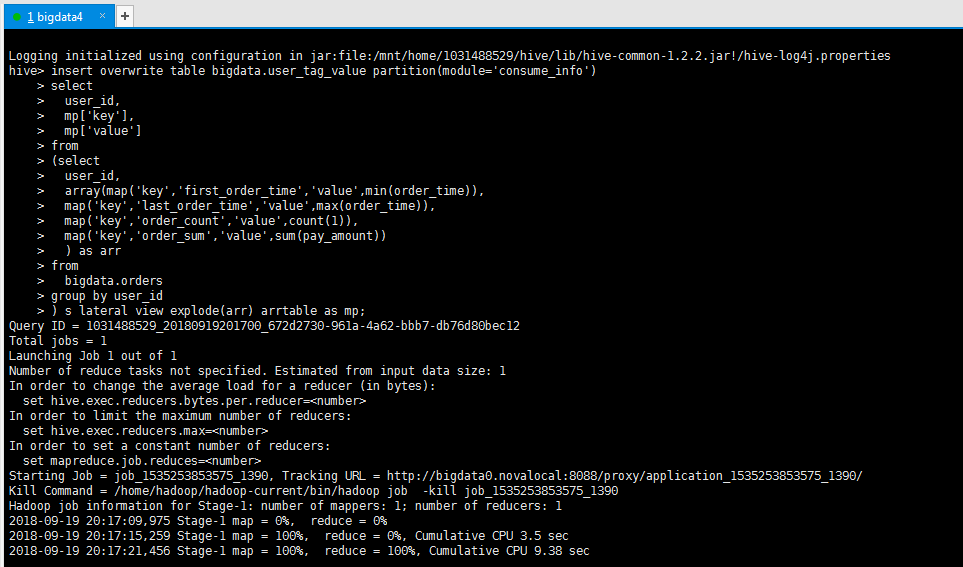


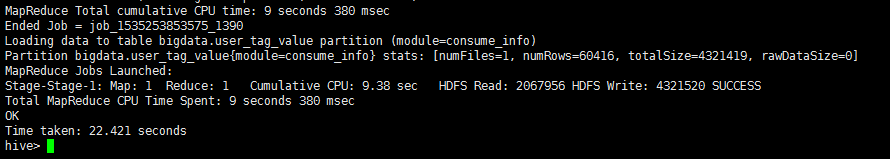
插入用户标签数据：



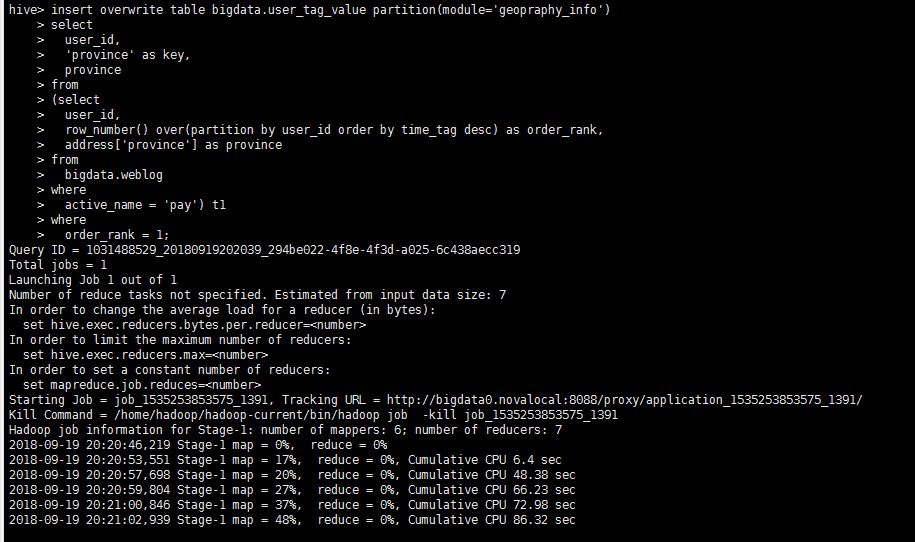


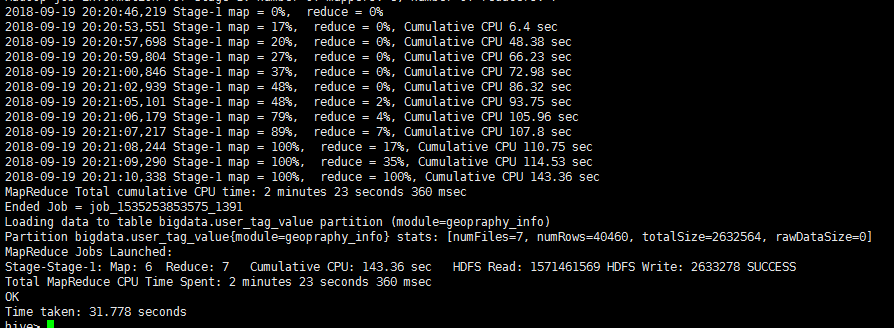
插入订单类信息：



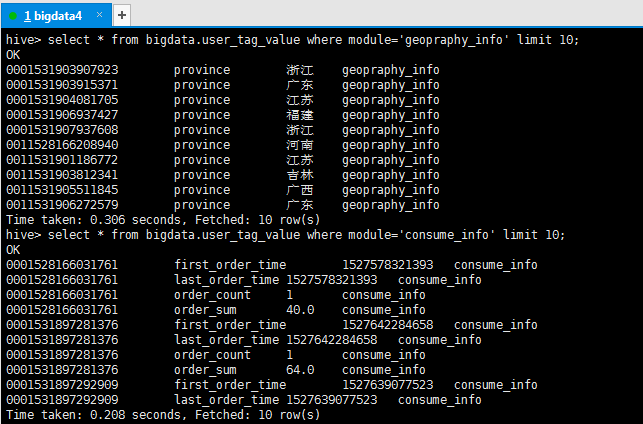


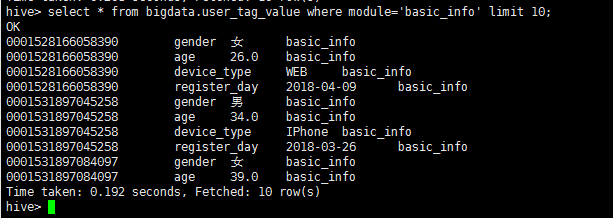
插入地理位置信息：



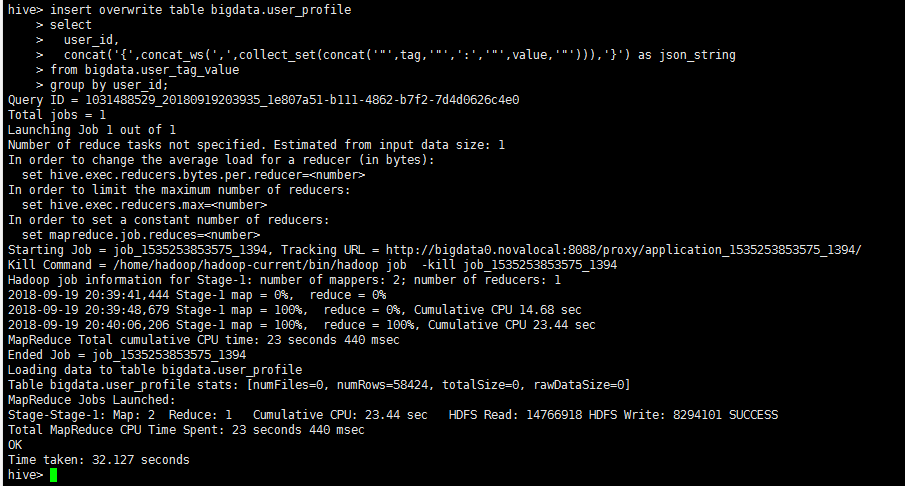


查看以上插入结果：

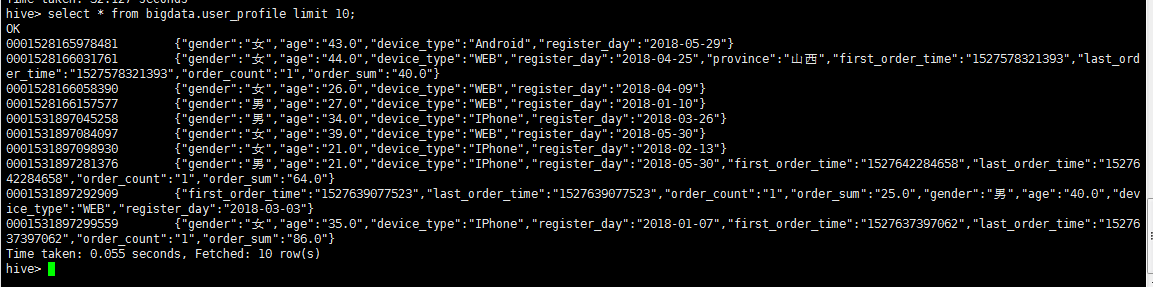




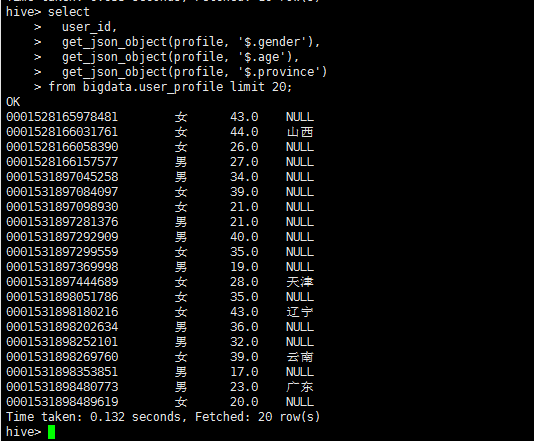
插入大宽表：



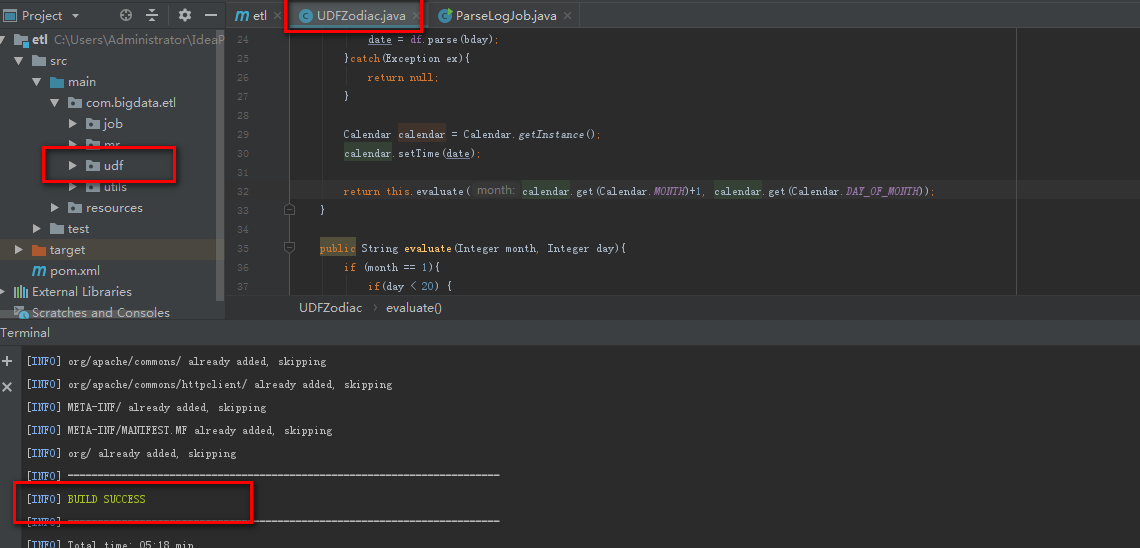
查看结果：



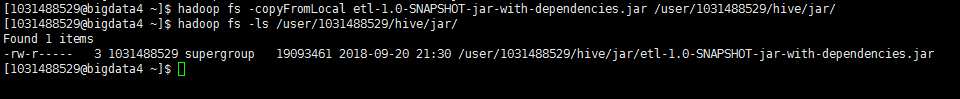
大宽表使用：



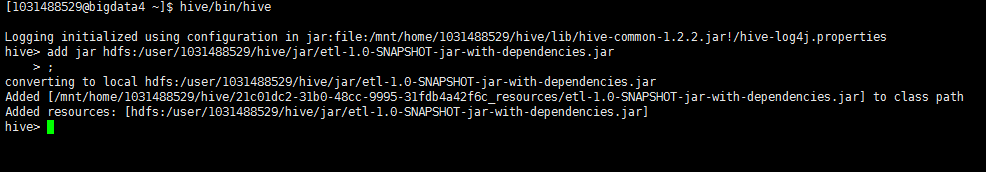
编辑MR自定义函数：



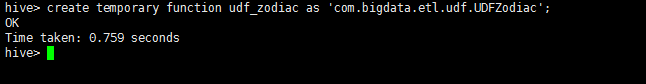
上传到bigdata4，并拷贝到hdfs目录：



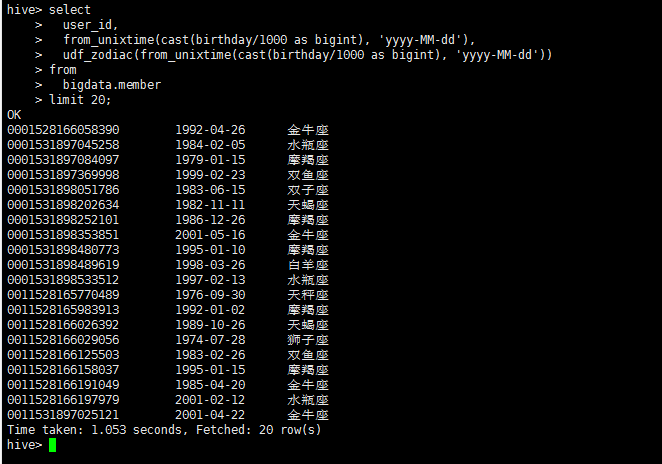
Hive加载jar包：



创建临时函数：

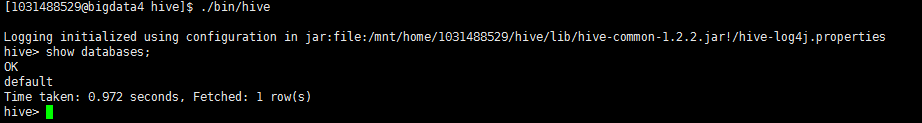


测试自定义函数：

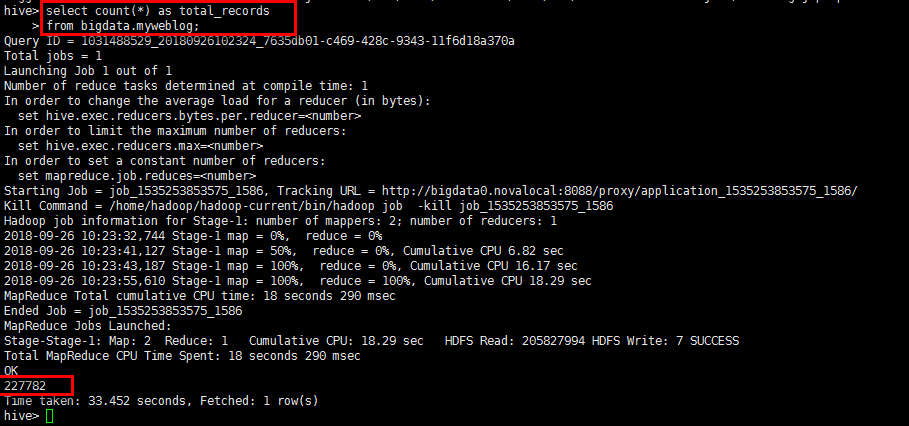


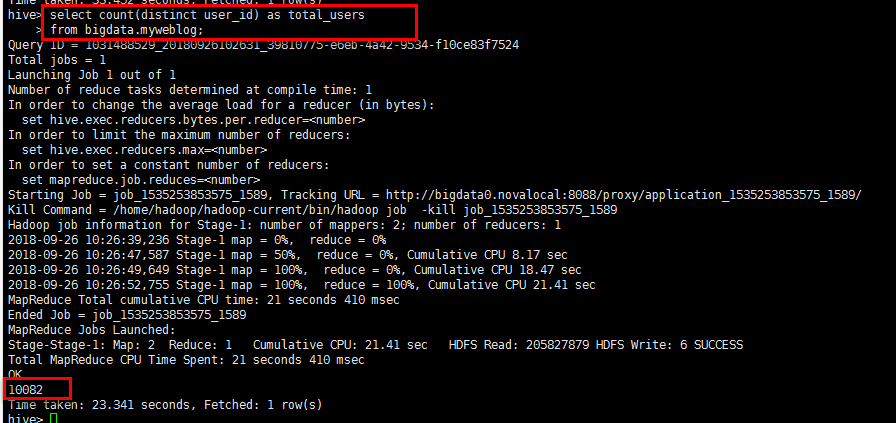
实验结果

启动hive成功



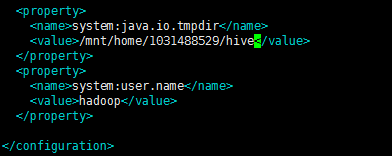
查询自定义日志表：



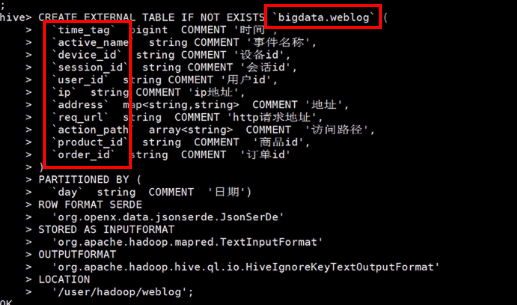


总结反思

1，配置hive-site.xml文件，**注意tempdir要配置到自己的目录**

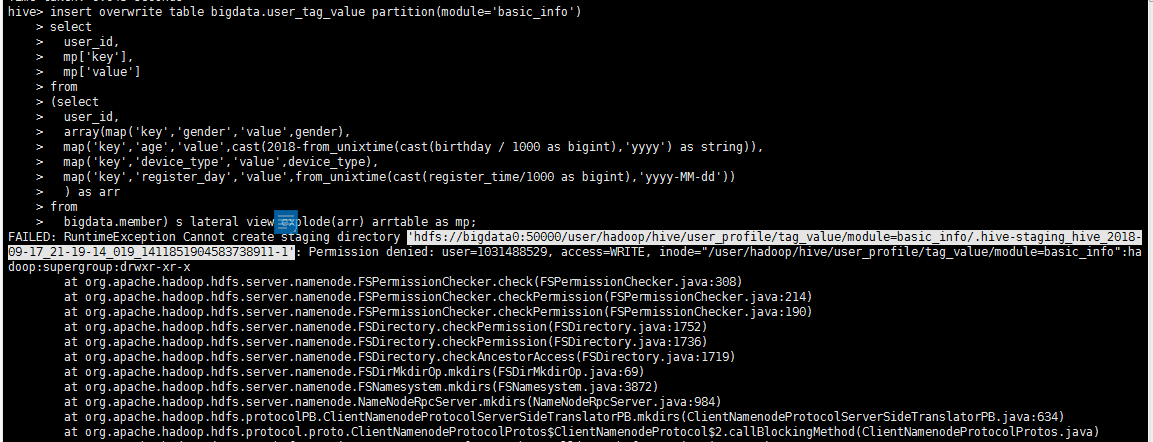


2，建表语句名、字段名**不是用单引号，是用键盘上1左边那个`，否则会报错**



3，hive中执行脚本，**不能带tab**，格式化 要用空格；

4，插入中间表异常：原因是建表的时候location用错了，**应该用自己的hdfs目录；**

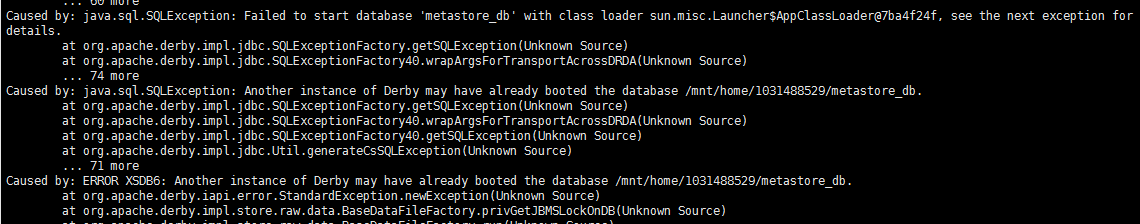




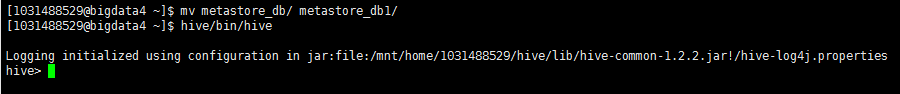
5，意外断网离线后，hive/bin/hive启动hive报错：

Failed to start database 'metastore\_db'

ERROR XSDB6: Another instance of Derby may have already booted the database



mv metastore\_db/ metastore\_db1/ 删除metastore\_db后再启动即可



6，hive外部表，实际只有表的元数据，即使删除表，也不会删除所关联的数据文件；

7，hive-site.xml配置文件， 这个用户名配错了，应该写自己的账号，改之后，可以删掉hive目录下的的hadoop文件夹，下次启动hive，会自动建立新账号的文件夹，但是库要重建；  
