****

大数据系统及应用

大作业设计报告

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 设计题目 | 网站用户行为分析 | | | |
| 完成时间 |  | | | |
| 指导教师 | 沈浩 | | | |
| 小组成员 | 班级 | 学号 | 姓名 | 角色 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

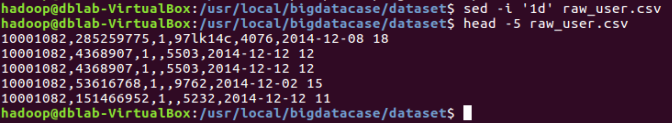
**任务一：本地数据集上传到数据仓库Hive**

**1、数据集的预处理**

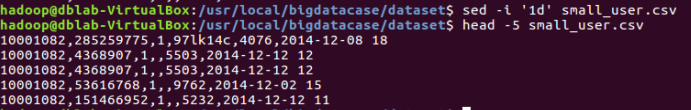
**1.1 删除文件第一行记录(即字段名称)**

raw\_user和small\_user中的第一行都是字段名称，把文件中的数据导入到数据仓库Hive中时，不需要第一行字段名称，因此，在做数据预处理时，需要删除第一行，执行以下命令(Shell命令)：

删除raw\_user.csv文件中第一行：

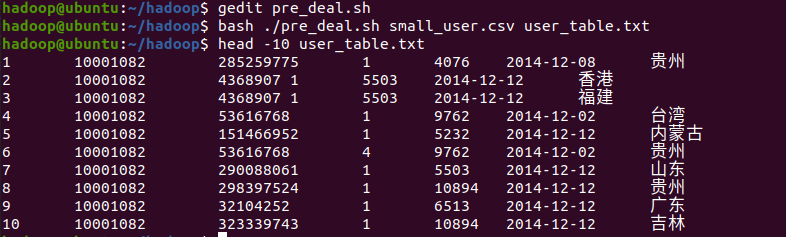


删除small\_user.csv文件中第一行：



接下来的操作中，都是用small\_user.csv这个小数据集进行操作，这样可以节省时间。等所有流程都跑通以后，你就可以使用大数据集raw\_user.csv去测试一遍整个案例。

**1.2 对字段进行预处理**



**2、导入数据库**

**2.1 启动HDFS**

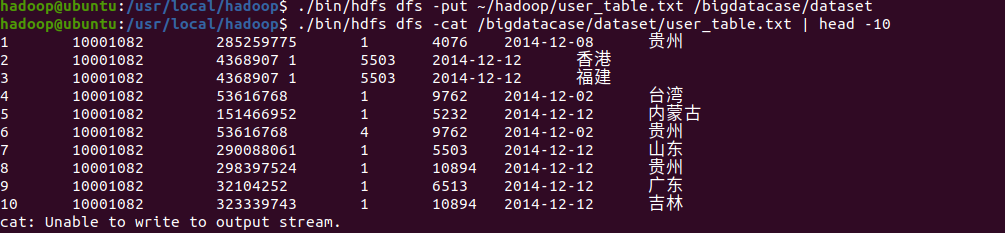
文本

描述已自动生成

文本

描述已自动生成

**2.2 把user\_table.txt上传到HDFS中**



**2.3 在Hive上创建数据库**

文本

描述已自动生成

**2.4 创建外部表**

文本

描述已自动生成

**2.5 查询数据**

电脑萤幕画面

描述已自动生成

文本

描述已自动生成

**任务二：Hive数据分析**

**1、简单查询分析**

文本

描述已自动生成

文本

描述已自动生成

**2、查询条数统计分析**

**2.1 用聚合函数count()计算出表内有多少条数据**

文本

描述已自动生成

**2.2 在函数内部加上distinct，查出uid不重复的数据有多少条。**

文本

描述已自动生成

**2.3 查询不重复的数据有多少条(为了排除客户刷单情况)**

文本

描述已自动生成

**3、关键字条件查询分析**

**3.1 以关键字的存在区间为条件的查询**

文本

描述已自动生成

文本

描述已自动生成

**3.2 关键字赋予给定值为条件，对其他数据进行分析。**

文本

描述已自动生成

**4、根据用户行为的分析**

**4.1 查询一件商品在某天的购买比例**

文本

描述已自动生成

文本

描述已自动生成

**4.2 查询某个用户在某一天点击网站占该天所有点击行为的比例 (点击行为包括浏览、加入购物车、收藏、购买)**

文本

描述已自动生成

文本

描述已自动生成

**4.3 给定购买商品的数量范围，查询某一天在该网站的购买该数量商品的用户id。**

文本

描述已自动生成

**5、用户实时查询分析**

文本

描述已自动生成

**任务三：Hive与MySQL数据互导**

**1、Hive预操作**

**1.1 创建临时表user\_action**



**1.2 将bigdata\_user表中的数据插入到user\_action**

文本

中度可信度描述已自动生成

屏幕上写着字

描述已自动生成

**2、使用Sqoop将数据从Hive导入MySQL**

**2.1 将前面生成的临时表数据从Hive导入到 MySQL 中**

**1) 登录MySQL**

文本

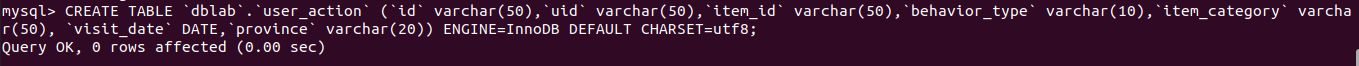
描述已自动生成

**2) 创建数据库**

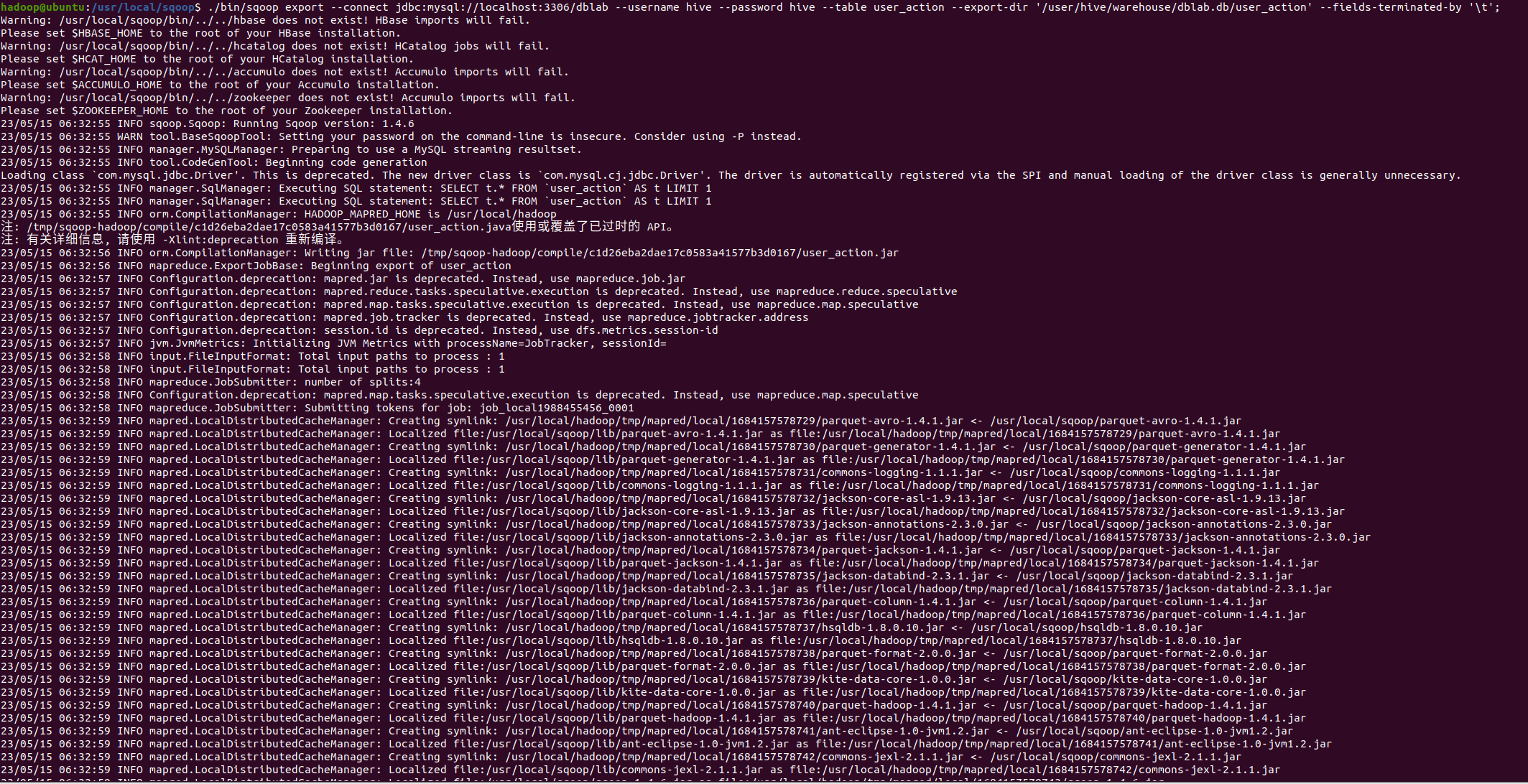
文本

描述已自动生成

**3) 创建表**



**4) 导入数据**



**2.2 查看MySQL中user\_action表数据**

