大数据处理综合实验

实验3 MyJoin\_Hive

实验报告

小组成员

尹浚宇 161130118

宋昱豪 171850505

刘扬 171850524

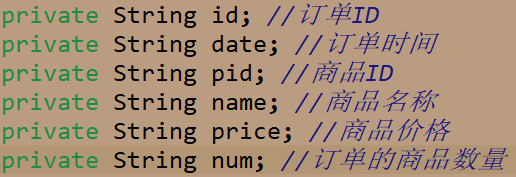
1.实验设计思路与伪代码

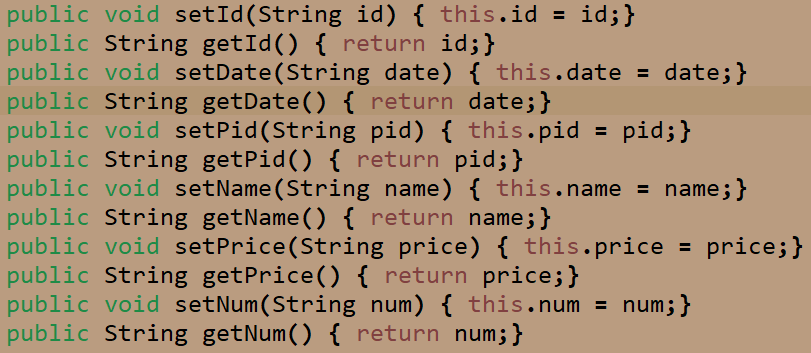
实验要求实现使用MapReduce完成两张表的join操作.

设计思路为: 自定义数据类型OrderBean存储Mapper读入的数据, 并将此数据类型 作为Mapper输出的key和Reducer的输入输出key类型.

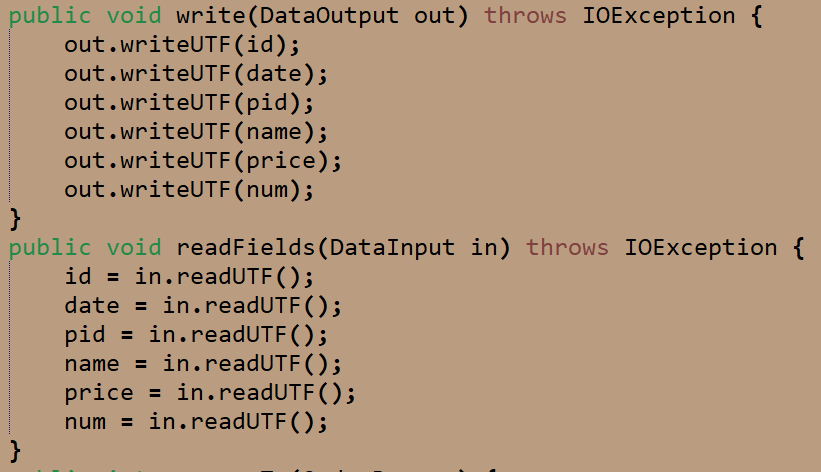
(1)OrderBean设计

OrderBean作为key使用, 需要实现WritableComparable接口, 即重写write readFields compareTo方法. OrderBean用来描述合并后的数据内容, 所以需要存 储6个属性对应的数据, 并为这6个数据设置get set方法.

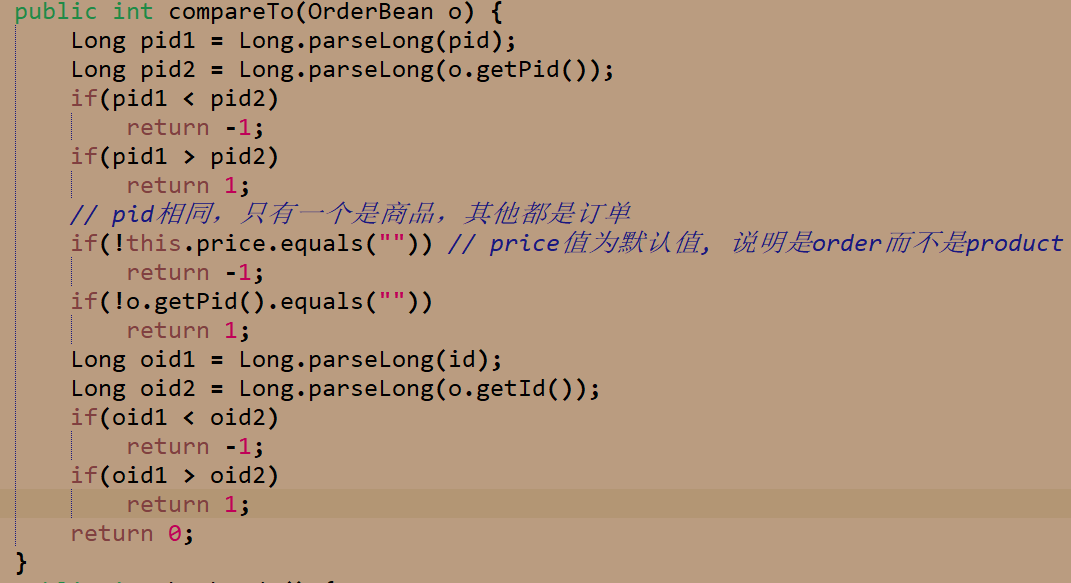




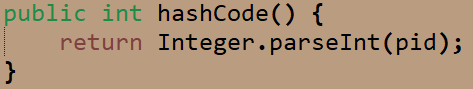
write和readFields的实现比较简单, 只需要分别调用DataInput类的 writeUTF方法和DataOutput类对应的readUTF方法即可.



重写compareTo方法进行自定义比较, 该方法会在Reducer内对<key, values 进 行排序时被调用. 希望将同一个pid的数据放在一起, 并且希望product在最前 面, order在后面. 比较的思路是: 先将pid解释为long型整数进行比较; 若pid 相同, 则判断进行比较的OrderBean哪一个是product, 通过price的值是否为默 认值(空字符串)来判断是否为product, pid相同的OrderBean中有且仅有一个是 product; 若两个OrderBean都不是product, 则将id解释为long型数据进行比 较. 至此比较结束.

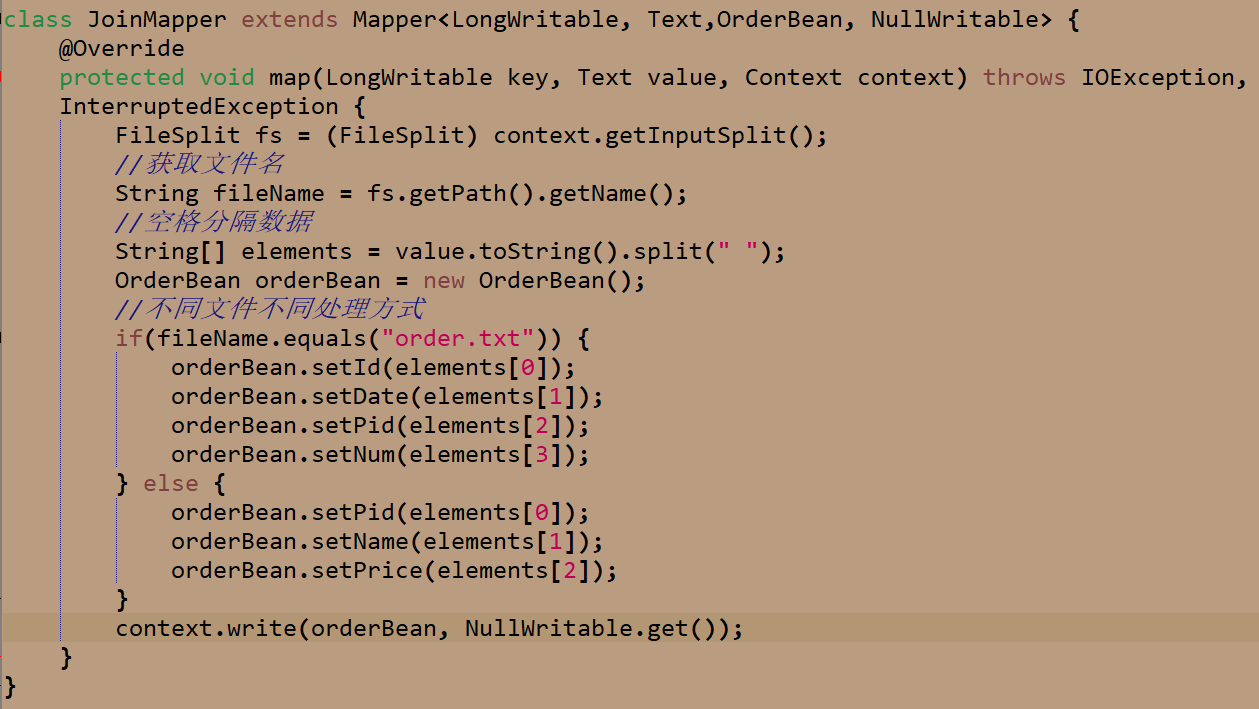


除此以外, 为了保证相同pid的数据能够被分配至一个Reduceri进行处理, 重写hashCode方法, 将pid解释为int型整数作为哈希值返回. 这样便可以利用 默认的HashPartitioner类进行划分.



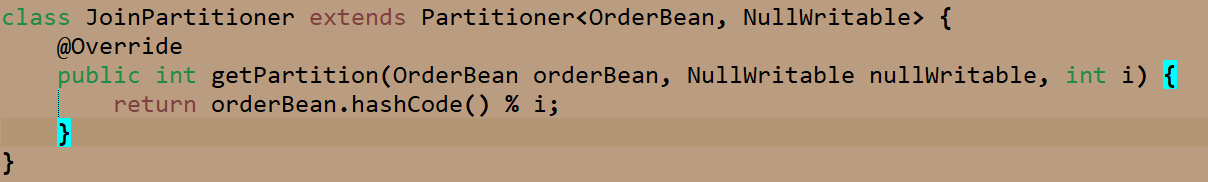
(2)Mapper设计

重写Mapper类的map方法. 指定输入格式为TextInputFormat, 则map方法输 入的key,value分别是文本行偏移和当前行的内容. 获取文件名, 通过文件名判断当 前行的内容属于order还是product, 在分别处理.



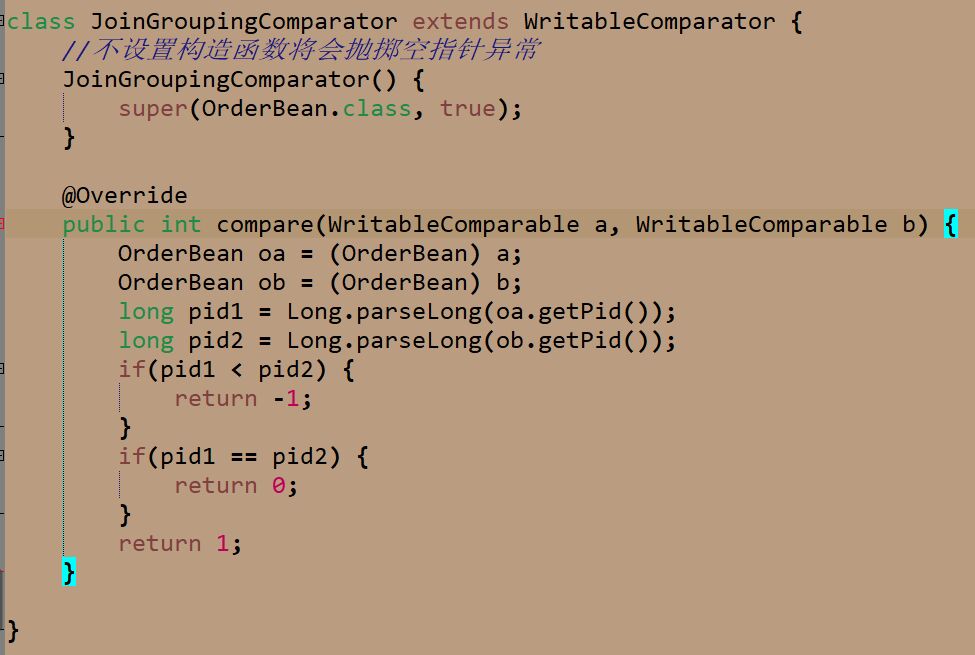
(3)Partitioner设计

自定义分组, 按照pid进行分组. 重写getPartition方法, 返回pid模 Reducer个数的结果即可.



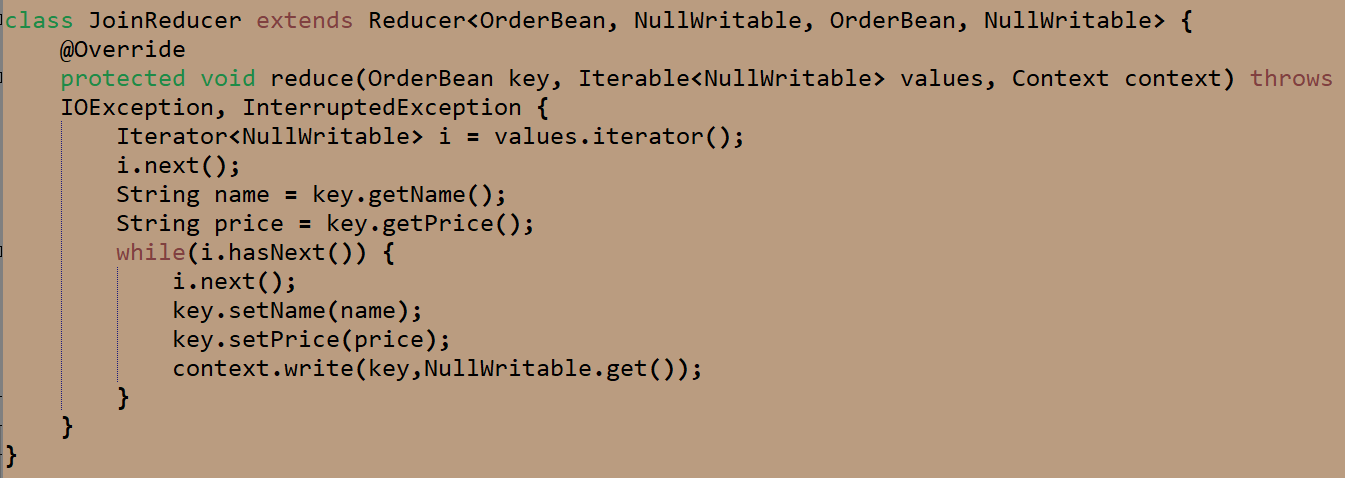
(4)Comparator设计

自定义分组, 将相同pid的key划分为一组, 并且保证product在前. 划分为 一组的key将会被一次reduce方法调用. 首先取出来product的名称和价格, 之后 迭代对每一个key填充name和price属性再输出.



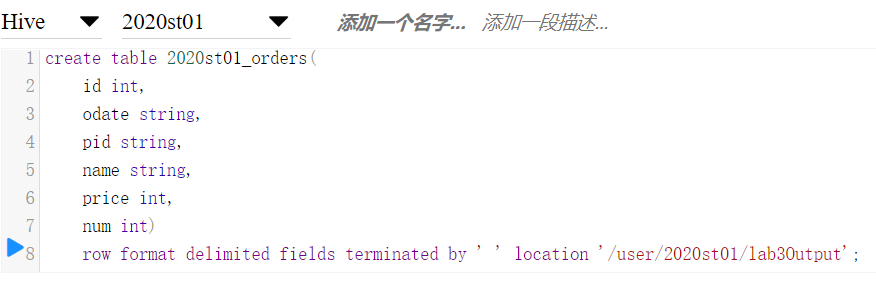
(5)Reducer设计

Reducer处理的<key, values>已经排好序. 对于一组具有相同pid的key, 第 一个key存储的是商品信息, 其他的key存储的是order信息, 保存第一个key的信 息, 对之后的key进行填充, 就可以实现join操作.

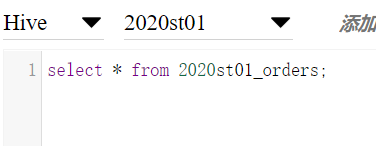


(4)使用Hive建表并查询内容

通过以下指令建立表:



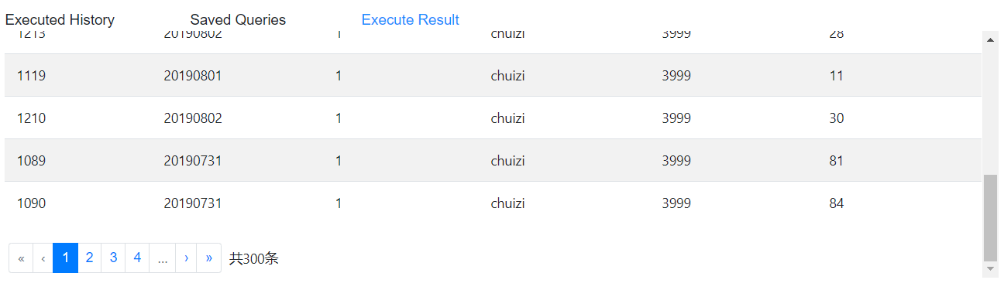
通过以下命令查询表的内容



2.输出结果截图

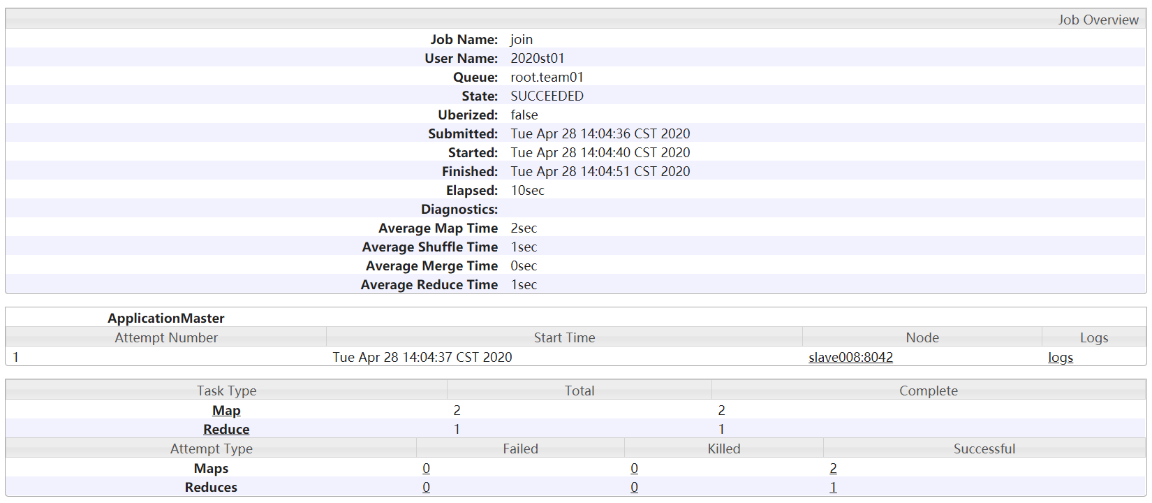


查询结果:



3. 执行报告

实验任务的WebUI执行报告



5.实验反思与不足

最开始我们只定义了OrderBean类, 通过pid进行分区, 之后根据compareTo方法进行排序后, 可以保证进入一个Reducer的<key,values>对符合实验要求的顺序, 结果也是预期的结果. 但是这种方法未涉及分组, 后重写添加了分组的功能.

分组功能需要继承WritableComaprator类重写compare方法. compare方法有三种不同参数, 我们需要重写参数类型为WritableComparable的方法, 重写参数类型为Object的类不会起作用.

按照实验流程可知, 最终的输出结果是按照pid局部有序的. 若实现按照oid进行全局排序, 将会使结果更加直观.

6.实验分工

思路设计是由小组三人共同讨论得出.

代码实现: 刘扬

集群测试: 宋昱豪

实验报告撰写: 尹浚宇 宋昱豪

7.参考资料

<https://hadoop.apache.org/docs/r2.8.2/api/org/apache/hadoop/io/WritableComparator.html>

<https://blog.csdn.net/brightming/article/details/29589249>

<https://zhuanlan.zhihu.com/p/55884610>