大数据处理综合实验

实验4 SecondSort

实验报告

小组成员

尹浚宇 161130118

宋昱豪 171850505

刘扬 171850524

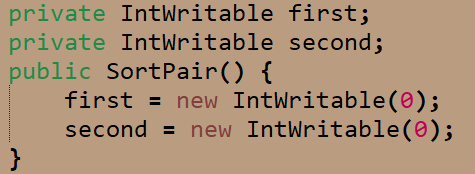
1.实验设计思路与伪代码

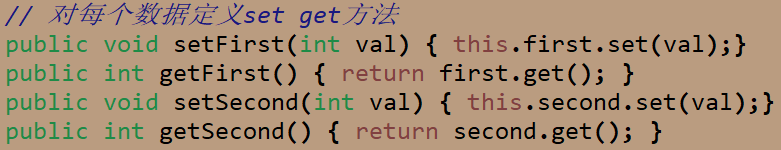
实验要求实现对输入数据按第一列升序排列, 第一列数据相同时按第二列降序排列.

设计思路为: 自定义数据类型SortPair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.

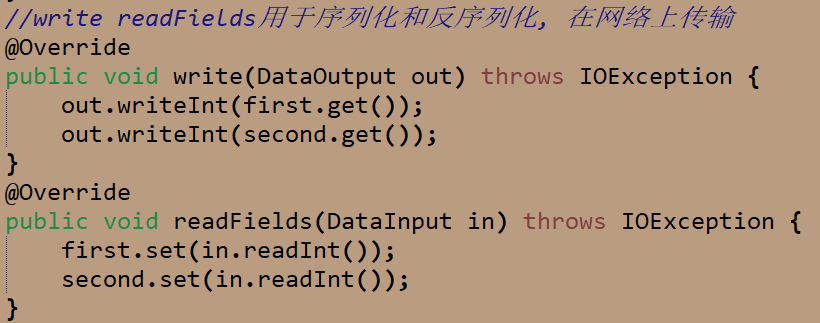
(1)SortPair设计

SortPair作为key使用, 需要实现WritableComparable接口, 即重写write readFields compareTo方法.SortPair用来描述每行两个数据, 所以需要存储2个 IntWritable类型的变狼, 并为这2个变量设置get set方法.

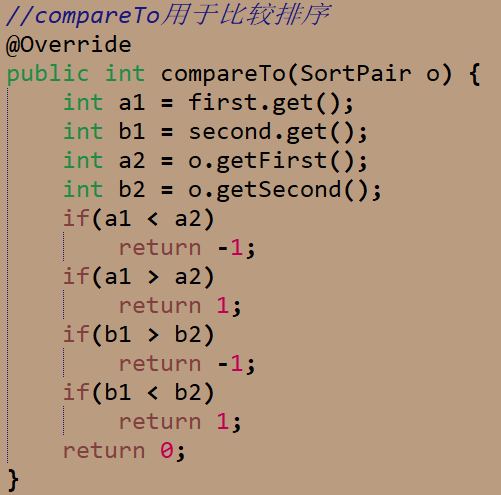




write和readFields的实现比较简单, 只需要分别调用DataInput类的 writeUTF方法和DataOutput类对应的readInt方法即可.



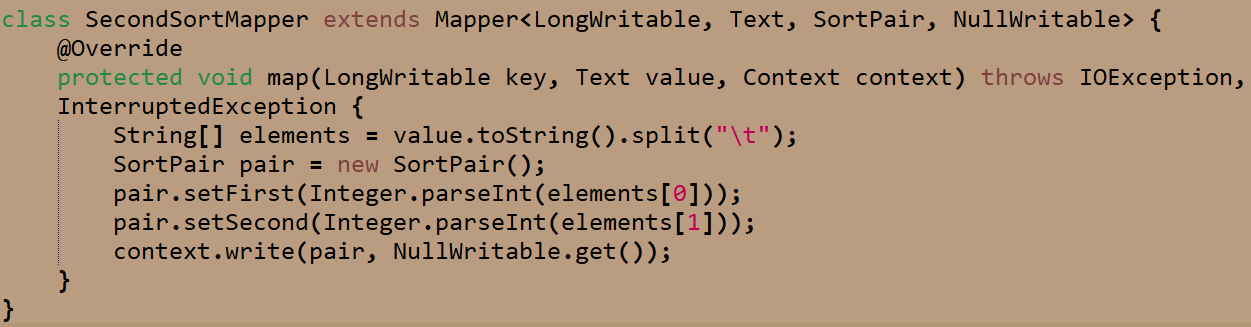
重写compareTo方法进行自定义比较, 该方法会在Reducer内对<key, values 进行排序时被调用. 该方法按照实验要求的排序方式进行实现, 首先按照第一个数据 进行升序排列, 在第一个数据相同的前提下, 按照第二个数据进行降序排列.



考虑到数据量过大时, 一个Reducer会降低处理速度, 可以让hashcode方法返回第一个数据值, 这样使用10个Reducer时可以保证文件之间数据的有序性.

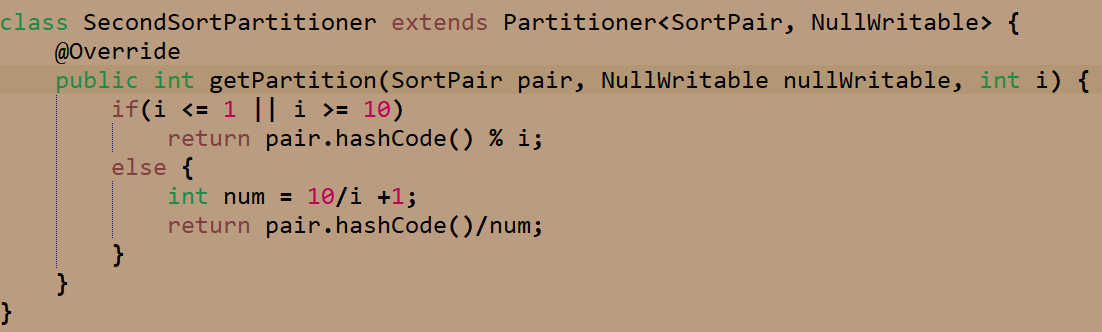
(2)Mapper设计

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. 对每行的内容使用换行符进行分割, 分别得到两个整数, 用于构造一个SortPair变量.



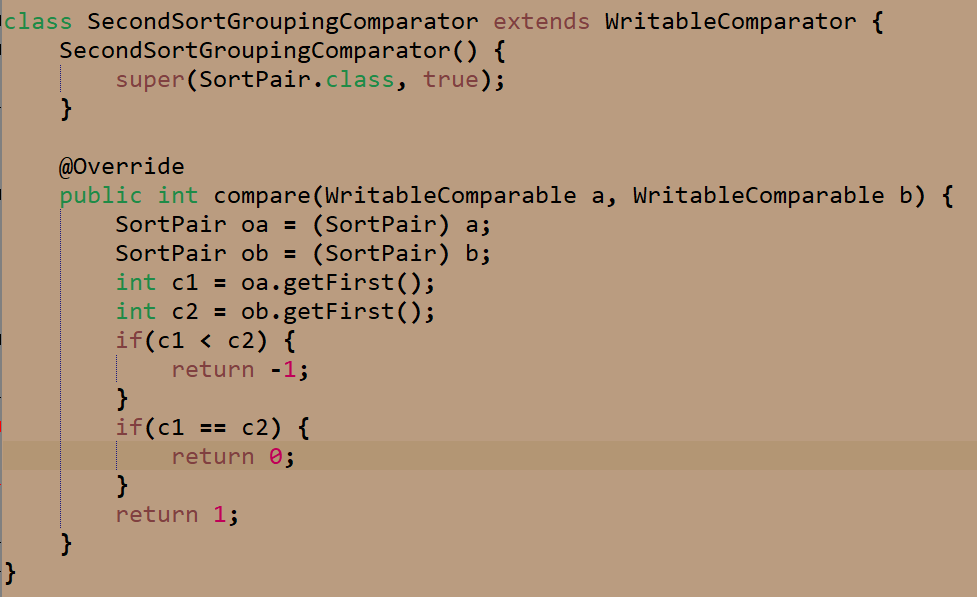
(3)Partitioner设计

自定义分组, 按照第一个数据进行分组. 重写getPartition方法, 针对不同的 Reducer个数分别进行处理.



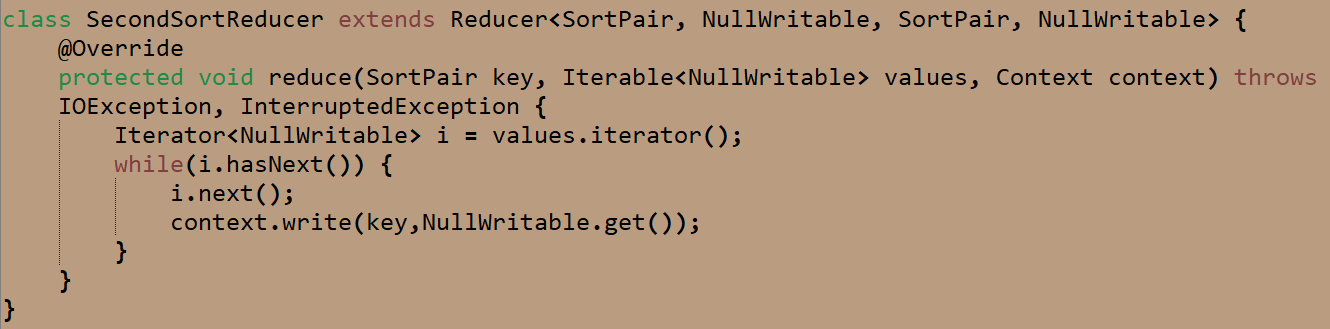
(4)Comparator设计

自定义分组, 将first相同的key划分为一组. 按照first进行比较即可.



(5)Reducer设计

Reducer处理的<key, values>已经排好序. 迭代输出每一个key.

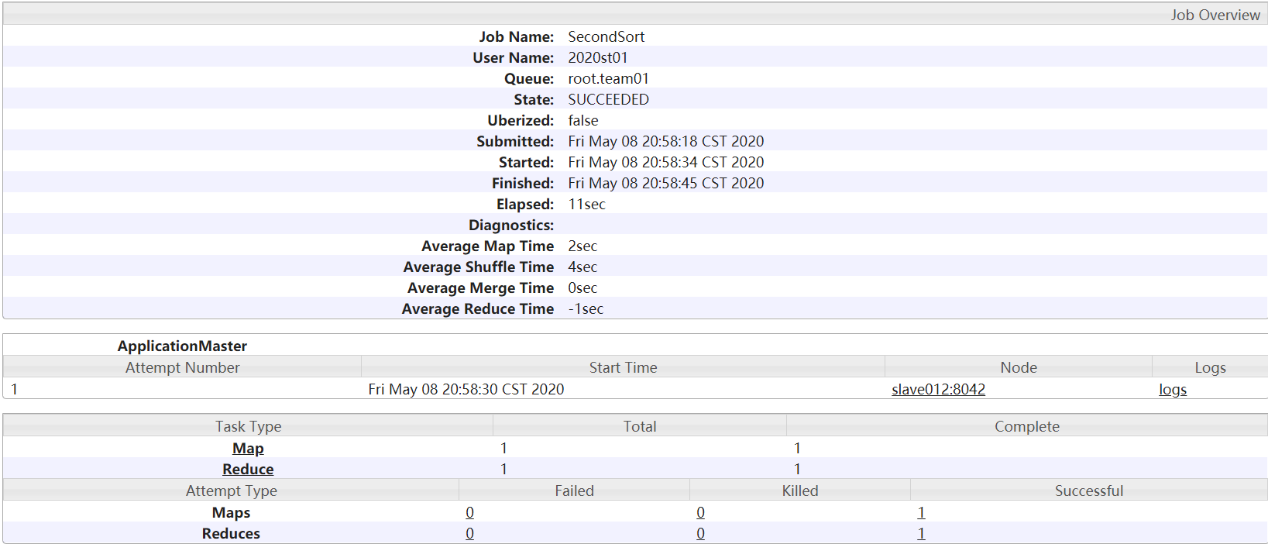


2.输出结果截图



3. 执行报告

实验任务的WebUI执行报告



5.实验反思与不足

实验大体上使用上次实验的解决思路, 并未从新的角度尝试解决问题.一开始我们并未考虑多个Reducer的情况, 后来对getPartition函数进行修改, 使程序能够应对多个Reducer的情况.

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>