Jaeseok Huh
2015005241
14 November 2018

Distributed Computing

Assignment #2

1. Directory Structure

averageRating.jar
build.sh*
com/jaeseok/
  compiled classes
data/
  ratings.csv
doc/
  report.pdf
.gitignore
run.sh*
src/
  AverageRating.java - includes MapReducer
  PairWritable.java - Pair<Double, Int>

2. A Brief Explanation about Code

In the folder src/, there are two classes.

Mapper and Reducer are core elements in Hadoop programming. This program is to get the ratings of each user and finally calculate his/her average rating. For the sake of scalability, Mapper sorts the data by user's ID and Reducer plays a role in calculating the average of grouped data.

First, in AverageRating.java, there are two inner classes, "TokenizerMapper"(line 24) and "AverageReducer"(line 54). They serve as a wrapper class for mapper and reducer respectively.

Second, in "TokenizerMapper", public void map(Object, Text, Context) (line serves as a mapper function and later added to the job(line 75). As stated in the documentation, it maps the data into Pair<Double, Int>, denoting a rating of the user and a count (always one) respectively. Since Hadoop requires to write the fields through Writable Interfaces and also the keys should be comparable, class Pair and its member shall be "WriteComparable" (See PairWritable.java)

In "AverageReducer", it provides reducer(line 58). It iterates through the intermediate lists and get the average value of them. It finally writes to a context(line 67).

In java main, a job which Hadoop shall execute is specified. Since the output values of Mapper and Reducer have different type, you should use job.setMapOutputValueClass() to clarify it.

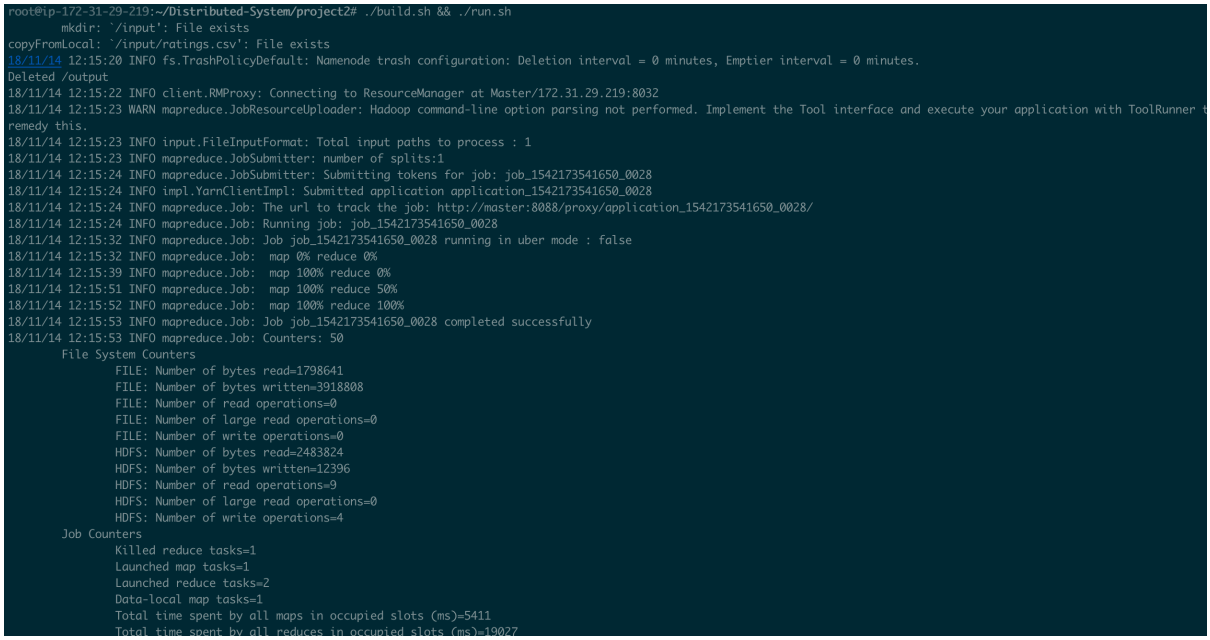
3. How to Build & Run

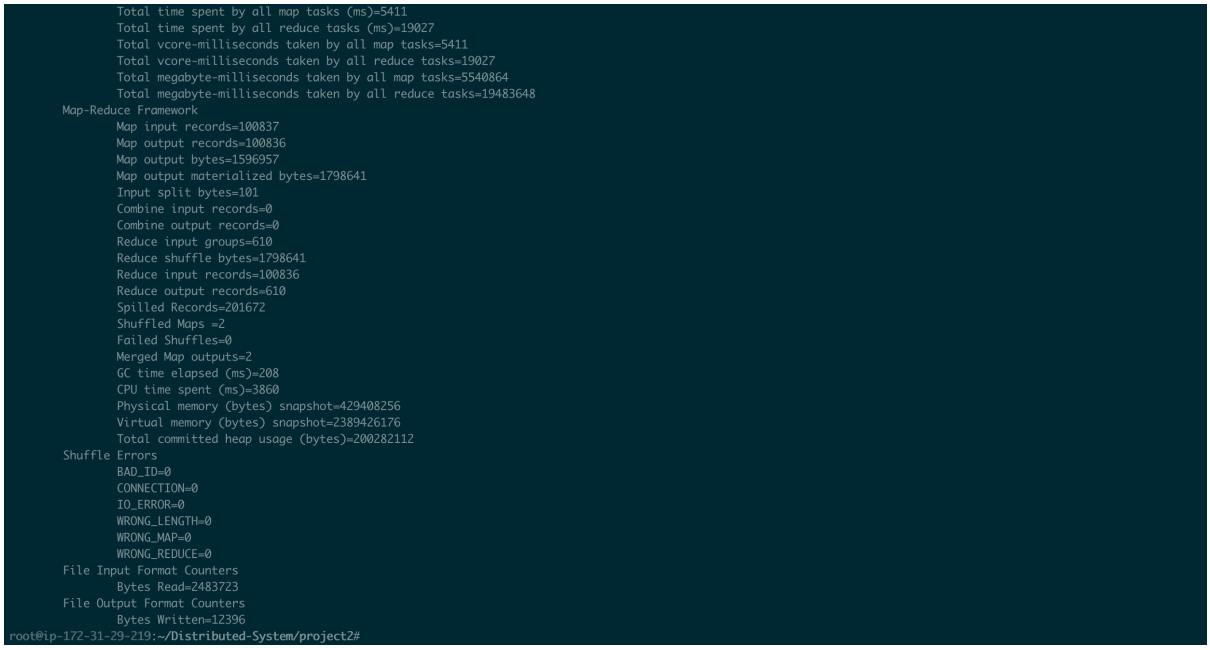
```
$ ./build.sh && ./run.sh
```

We assume that you have followed the instruction and your environment is set as following:

```
export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64/
export JRE_HOME=$JAVA_HOME/jre
export CLASSPATH=.:$JAVA_HOME/lib:$JRE_HOME/lib:$CLASSPATH
export PATH=$JAVA_HOME/bin:$JRE_HOME/bin:$PATH
export M2_HOME=/usr/local/apache-maven-3.3.9
export PATH=$M2_HOME/bin:$PATH
export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:/usr/local/lib
export HADOOP_HOME=/usr/local/src/hadoop-2.6.5-src/hadoop-dist/target/
hadoop-2.6.5
export PATH=$HADOOP_HOME/bin:$PATH
export HADOOP_CLASSPATH=${JAVA_HOME}/lib/tools.jar
```

4. Screenshot





```
root@ip-172-31-29-219:~/Distributed-System/project2# hdfs dfs -text /output^C
root@ip-172-31-29-219:~/Distributed-System/project2# hdfs dfs -ls /output
Found 3 items
-rw-r--r--   2 root supergroup          0 2018-11-14 12:15 /output/_SUCCESS
-rw-r--r--   2 root supergroup       6408 2018-11-14 12:15 /output/part-r-00000
-rw-r--r--   2 root supergroup       5988 2018-11-14 12:15 /output/part-r-00001
root@ip-172-31-29-219:~/Distributed-System/project2# hdfs dfs -text /output/_SUCCESS
root@ip-172-31-29-219:~/Distributed-System/project2# hdfs dfs -text /output/part-r-00000
1       4.366379310344827
10      3.2785714285714285
100     3.945945945945946
102     3.357142857142857
104     3.5073260073260073
106     4.4393939393939394
108     3.986842105263158
111     3.3397832817337463
113     3.6466666666666665
115     3.767857142857143
117     3.3393939393939394
119     4.176744186046512
12      4.390625
120     3.409090909090909
122     4.546232876712328
124     3.99
126     3.289473684210526
128     4.363636363636363
131     3.4420289855072466
```