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

.gitignore

run.sh*

src/

  AverageRating.java - includes MapReducer

  PairWritable.java - Pair<Double, Int>

2. A Brief Explanation about Code

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.

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 (one) respectively. Since Hadoop requires to access the fields, 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.
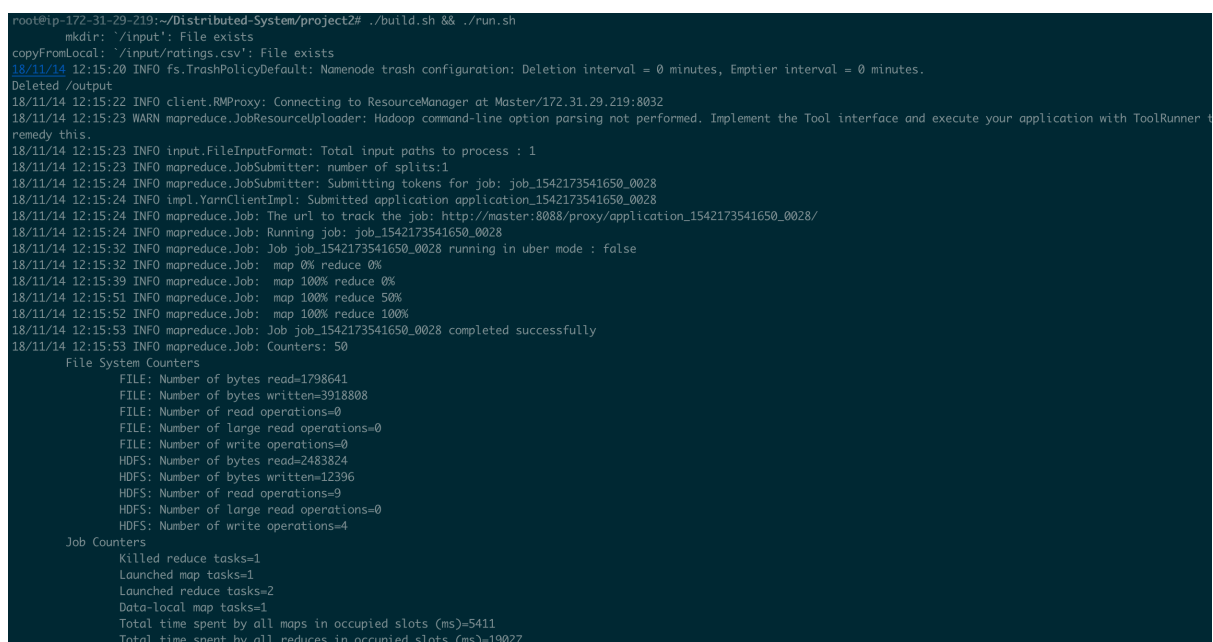
3. How to Build & Run

./build.sh && ./run.sh

We assume that 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

```
                Total time spent by all map tasks (ms)=5411
                Total time spent by all reduce tasks (ms)=19027
                Total vcore-milliseconds taken by all map tasks=5411
                Total vcore-milliseconds taken by all reduce tasks=19027
                Total megabyte-milliseconds taken by all map tasks=5540864
                Total megabyte-milliseconds taken by all reduce tasks=19483648
        Map-Reduce Framework
                Map input records=100837
                Map output records=100836
                Map output bytes=1596957
                Map output materialized bytes=1798641
                Input split bytes=101
                Combine input records=0
                Combine output records=0
                Reduce input groups=610
                Reduce shuffle bytes=1798641
                Reduce input records=100836
                Reduce output records=610
                Spilled Records=201672
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=208
                CPU time spent (ms)=3860
                Physical memory (bytes) snapshot=429408256
                Virtual memory (bytes) snapshot=2389426176
                Total committed heap usage (bytes)=200282112
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=2483723
        File Output Format Counters
                Bytes Written=12396
root@ip-172-31-29-219:~/Distributed-System/project2#
```

```
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     4.439393939393939
108     3.98684210526315.8
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
```