Explanation for the first part of the assignment

In Map phase, the input is (key, **value**). Each line of the input file users.xml is as **value** for each map. Here call transformXmlToMap function to parse **value** into HashMap<String, String> for us from which the relevant fields (id and reputation) can be extracted. Then use the TreeMap<Integer, String> to store the top ten records, here Integer is reputation, String is "id_reputation". The reason why putting reputation as key is to easily sort in TreeMap. And the reason why storing the top ten records here in each Map is to reduce the burden in Reduce phase.

Then the cleanup method gets called. The top ten records in the TreeMap are output to the reducers in the cleanup method. Here the output (key, values) format is (NullWritable, Iterable<Text> values). And <Text> is text("id_reputation").

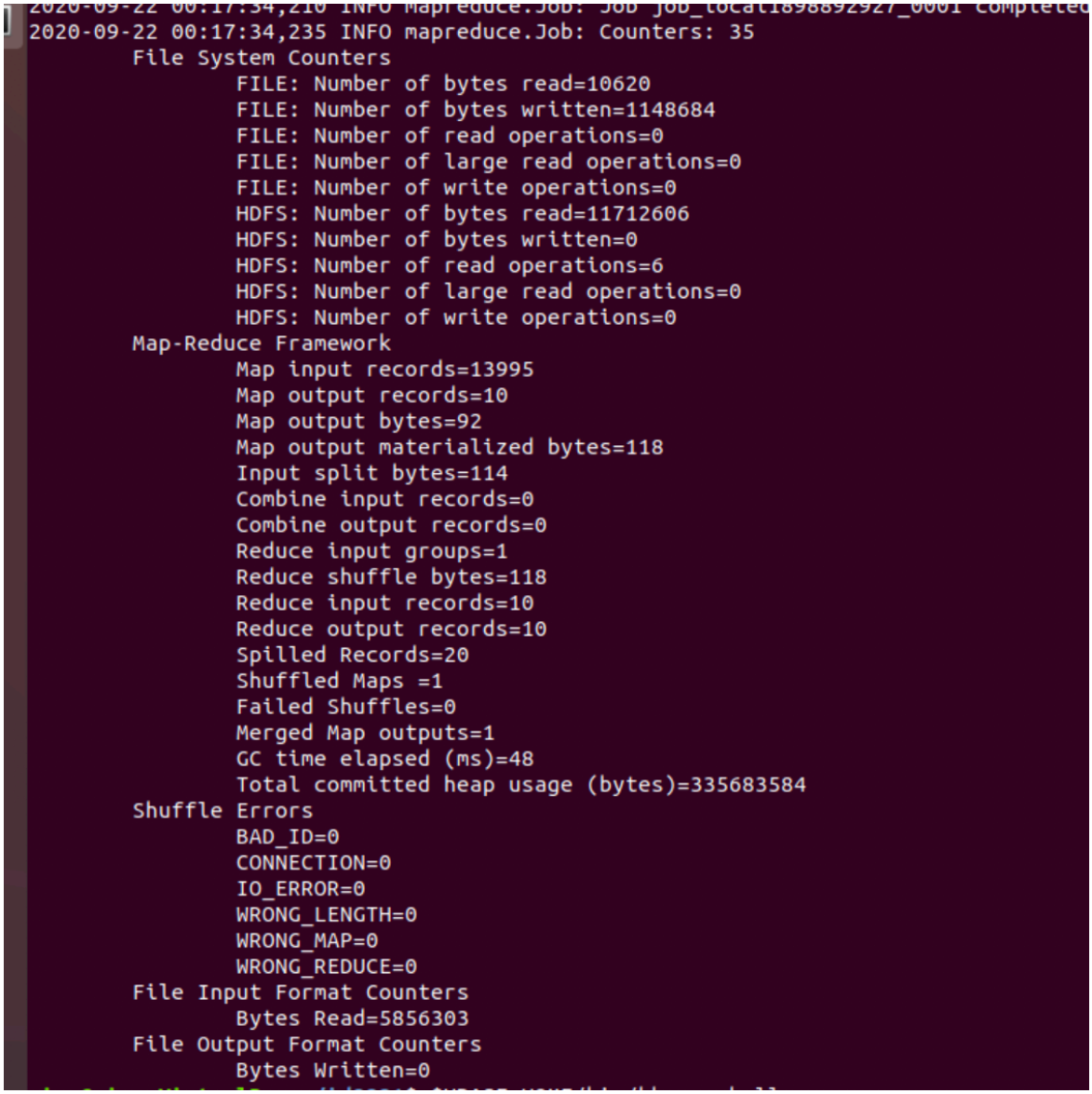
In Reduce phase, according to the input of Reduce (output of cleanup), for each value in Iterable<Text> values, use the TreeMap<Integer, String> to store the top ten records. Here Integer is reputation, String is "id_reputation". Thus, TreeMap<Integer, String> stores the final topten result. Then write the result into Hbase.

In the main function, set all the configurations, create job instance, set map class and Jar class, define input file path parameter and output hbase table, and define scan and column families.

Then run this code. Firstly, upload input file into HDFS and create Hbase table, then compile code and make jar file, and then set input file path, finally run the application.

The result is shown as following screenshots. The first screenshot is the final result in Hbase. The second one is the processing summary of map-reduce framework.

```
hbase(main):001:0> scan 'topten'
ROW                     COLUMN+CELL
 row0                   column=info:id, timestamp=1600726653423, value=836
 row0                   column=info:reputation, timestamp=1600726653423, value=1846
 row1                   column=info:id, timestamp=1600726653423, value=9420
 row1                   column=info:reputation, timestamp=1600726653423, value=1878
 row2                   column=info:id, timestamp=1600726653423, value=108
 row2                   column=info:reputation, timestamp=1600726653423, value=2127
 row3                   column=info:id, timestamp=1600726653423, value=434
 row3                   column=info:reputation, timestamp=1600726653423, value=2131
 row4                   column=info:id, timestamp=1600726653423, value=84
 row4                   column=info:reputation, timestamp=1600726653423, value=2179
 row5                   column=info:id, timestamp=1600726653423, value=548
 row5                   column=info:reputation, timestamp=1600726653423, value=2289
 row6                   column=info:id, timestamp=1600726653423, value=21
 row6                   column=info:reputation, timestamp=1600726653423, value=2586
 row7                   column=info:id, timestamp=1600726653423, value=11097
 row7                   column=info:reputation, timestamp=1600726653423, value=2824
 row8                   column=info:id, timestamp=1600726653423, value=381
 row8                   column=info:reputation, timestamp=1600726653423, value=3638
 row9                   column=info:id, timestamp=1600726653423, value=2452
 row9                   column=info:reputation, timestamp=1600726653423, value=4503
10 row(s) in 0.6120 seconds
```


2020-09-22 00:17:34,235 INFO mapreduce.Job: Counters: 35
File System Counters
FILE: Number of bytes read=10620
FILE: Number of bytes written=1148684
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=11712606
HDFS: Number of bytes written=0
HDFS: Number of read operations=6
HDFS: Number of large read operations=0
HDFS: Number of write operations=0
Map-Reduce Framework
Map input records=13995
Map output records=10
Map output bytes=92
Map output materialized bytes=118
Input split bytes=114
Combine input records=0
Combine output records=0
Reduce input groups=1
Reduce shuffle bytes=118
Reduce input records=10
Reduce output records=10
Spilled Records=20
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=48
Total committed heap usage (bytes)=335683584
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=5856303
File Output Format Counters
Bytes Written=0