HOME ASSIGNMENT 2

Roll No: 208W1A1299

Name: MOHAMMAD RIZWANULLAH

PROJECT NAME:Twitter data analysis.

HadoopMapReduce-Twitter

Implementing MapReduce algorithms in Hadoop using the Twitter dataset(schema - https://github.com/episod/twitter-api-fields-as-crowdsourced/wiki)

Question answered:

- 1. What hour of the day does @PrezOno's tweet the most on average, using every day we have twitter data? Directory https://github.uc.edu/loganasr/HadoopMapReduce-Twitter/tree/master/TweetsByHour
- 2. What day of the week does @PrezOno tweet the most on average? Use the same example as in #1 but for days of the week. Directory
 - https://github.uc.edu/loganasr/HadoopMapReduce-Twitter/tree/master/TweetsByDay
- 3. How does @PrezOno's tweet length compare to the average of all others? What is his average length? All others? Directory https://github.uc.edu/loganasr/HadoopMapReduce-Twitter/tree/master/TweetLength

Instructions:

A sample data file has been included in /data directory to support quick validations through the Hadoop streaming mode. However, the file does not contain tweets from @PrezOno and hence, it would be necessary update the user_name for filtering the tweets.

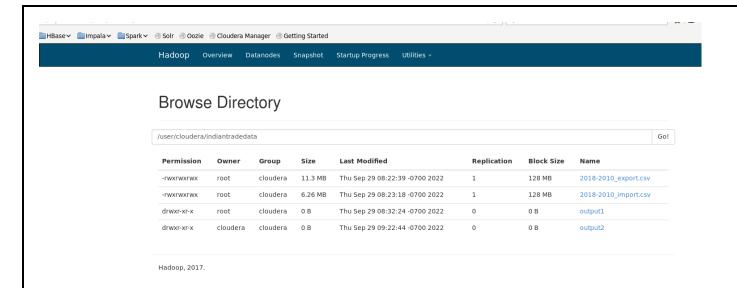
Sample command: cat /data/sample-data | ./mapTweetsByHour.py | sort | ./reduceTweetsByHour.py

To run the map reduce programs in the hadoop cluster, utilize the following command.

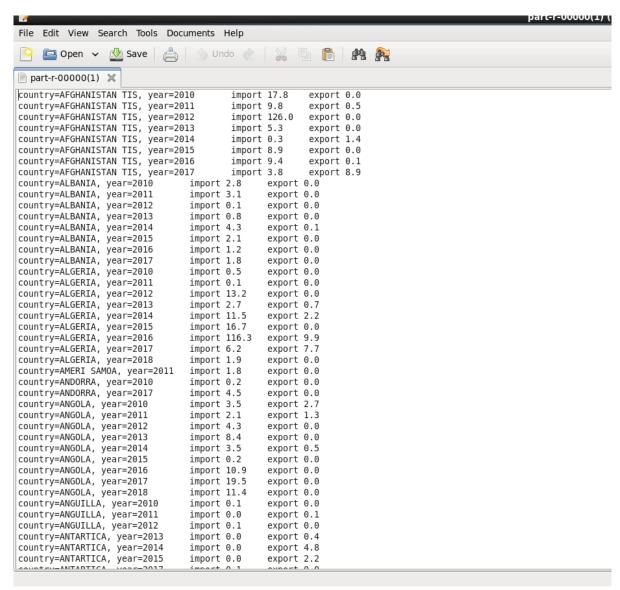
hadoop jar /root/hadoop-2.7.1/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar -input /data/twitter -output myoutput -file *.py -mapper mapTweetsByHour.py -reducer reduceTweetsByhour.py

Execution Screenshots:

```
[ [Worldwide-trade-data] [ ] clouderal-quickstart...
                              Milled map tasks=1
Luunched map tasks=2
                               Launched reduce Lusks=1
                              Launched reduce tusks=1
Data-local map tanks=2
Total time spent by all maps in occupied slots (ms)=19449
Total time spent by all reduces in occupied slots (ms)=5582
Total time spent by all map tasks (ms)=19449
Total time spent by all map tasks (ms)=5587
Total vcore-milliseconds taken by all map tasks=19449
Total vcore-milliseconds taken by all map tasks=19449
Total vcore-milliseconds taken by all map tasks=195776
Total megabyte-milliseconds taken by all map tasks=195776
Total megabyte-milliseconds taken by all reduce tasks=5715960
use Framework
            Map Reduce Framework
Map input records=213149
Map sutput records=9313
Map output bytes=292875
                              Map output materialized bytes=251513
Input split bytes=578
                              Combine input records=8
Combine output records=8
                              Reduce input groups=1677
Reduce shuffle bytes=251513
                              Reduce input records=9311
Reduce output records=1677
                               Spilled Records=18628
                               Shuffled Mapa ≈2
                             Shiffled Maps =2
Failed Shiffles=0
Herged Map Outputs=2
GC time elapsed (ms)=399
CPU time spent (ms)=399
CPU time spent (ms)=360
Physical semory (bytes) snapshot=563458048
Virtual nemory (bytes) snapshot=4519178240
Total committed heap usage (bytes)=391579008
              Shuffle Errors
BAD II=0
                               CONNECTION#0
                               TO ERRORES
                               WRONG LENGTH=0
                               WRONG MAPHD
                              WRONG REDUCE=0
             File Input Format Counters
                               Bytes Read=0
             File Output Format Counters
Bytes Written=86561
loudera@quickstart Worldwide-trade-data($ hadoop fs -Ls /user/cloudera/indiantradedata/output2
W-r-r- 1 clouderw clouders 8 7872-89-29 89:22 /user/clouders/indiantradedata/output2/_SUCCESS
W-r-r- 1 clouders clouders 86561 2922-89-29 89:22 /user/clouders/indiantradedata/output2/part-r-88888
louder@Bquickstart Worldwide-trade-data]s hadoop fs -ls /user/clouders/indiantradedata/output2/
```



Output:



Output: We are able to get the most accurate times where celebrities tweet in which hour and in which day of the week