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:

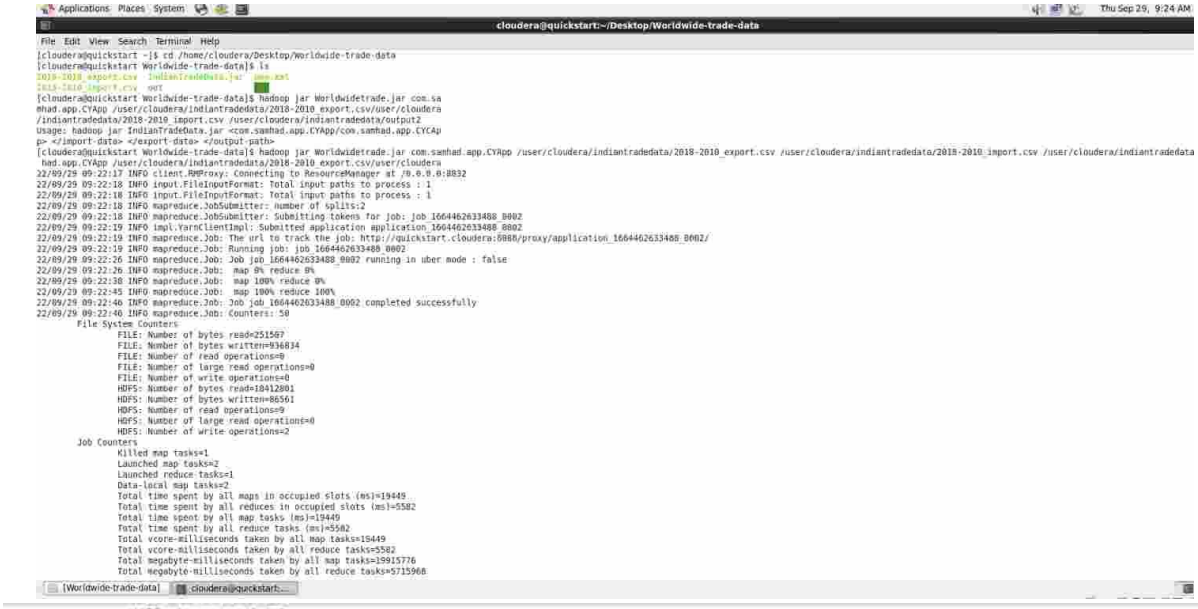
A sample data file has been included in /data directory to support quick validations through the Hadoop streaming mode. However, the file does not contain tweets from @PrezOno and hence, it would be necessary update the user_name for filtering the tweets.

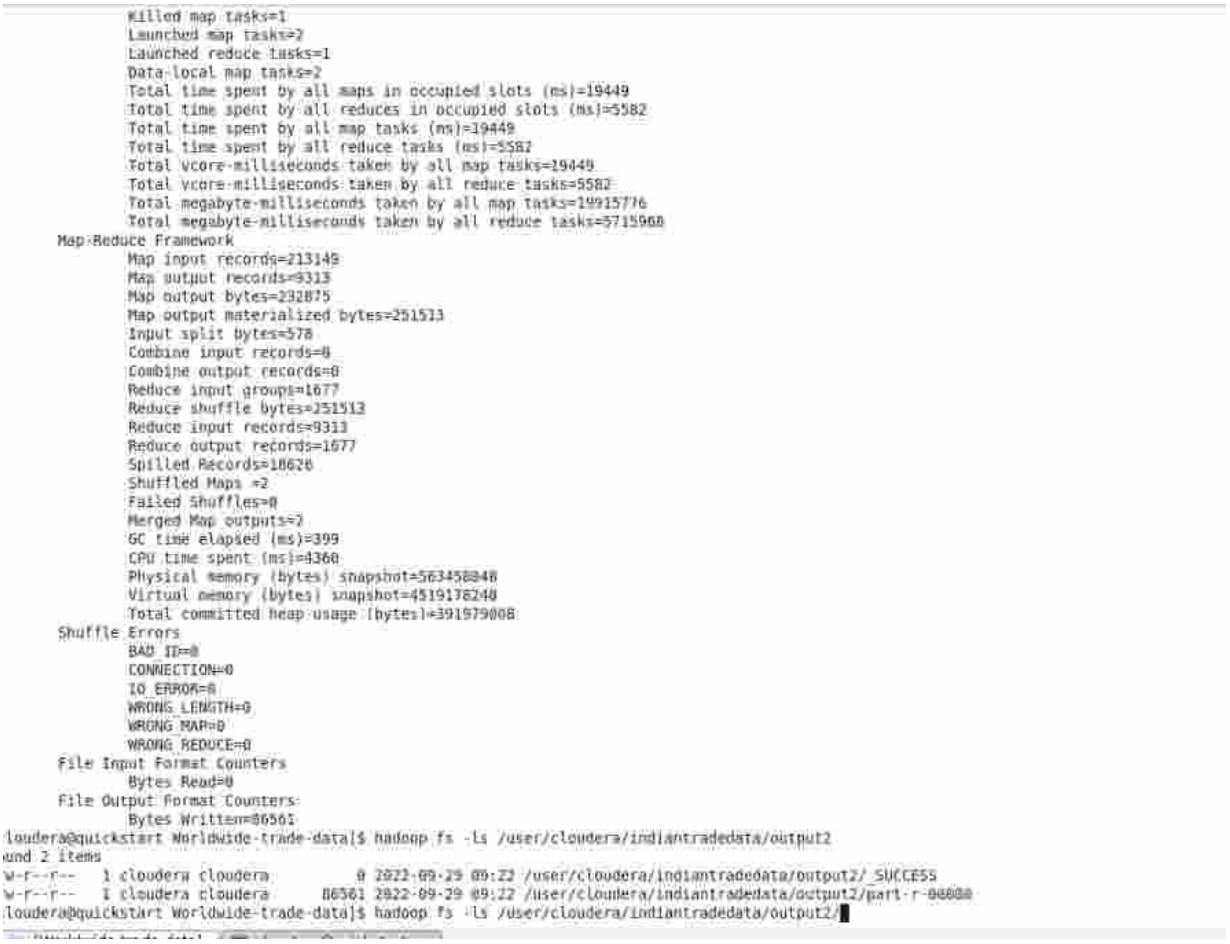
Sample command: cat /data/sample-data | ./mapTweetsByHour.py | sort | ./reduceTweetsByHour.py

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

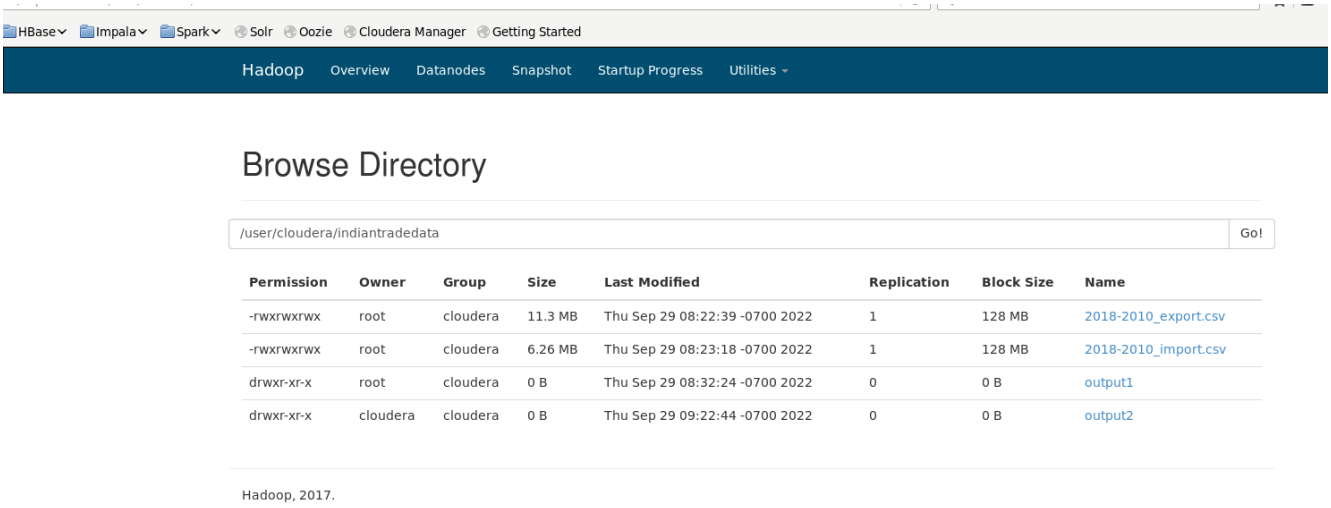
hadoop jar /root/hadoop-2.7.1/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar -input /data/twitter -output myoutput -file *.py -mapper mapTweetsByHour.py -reducer reduceTweetsByhour.py

Execution Screenshots:

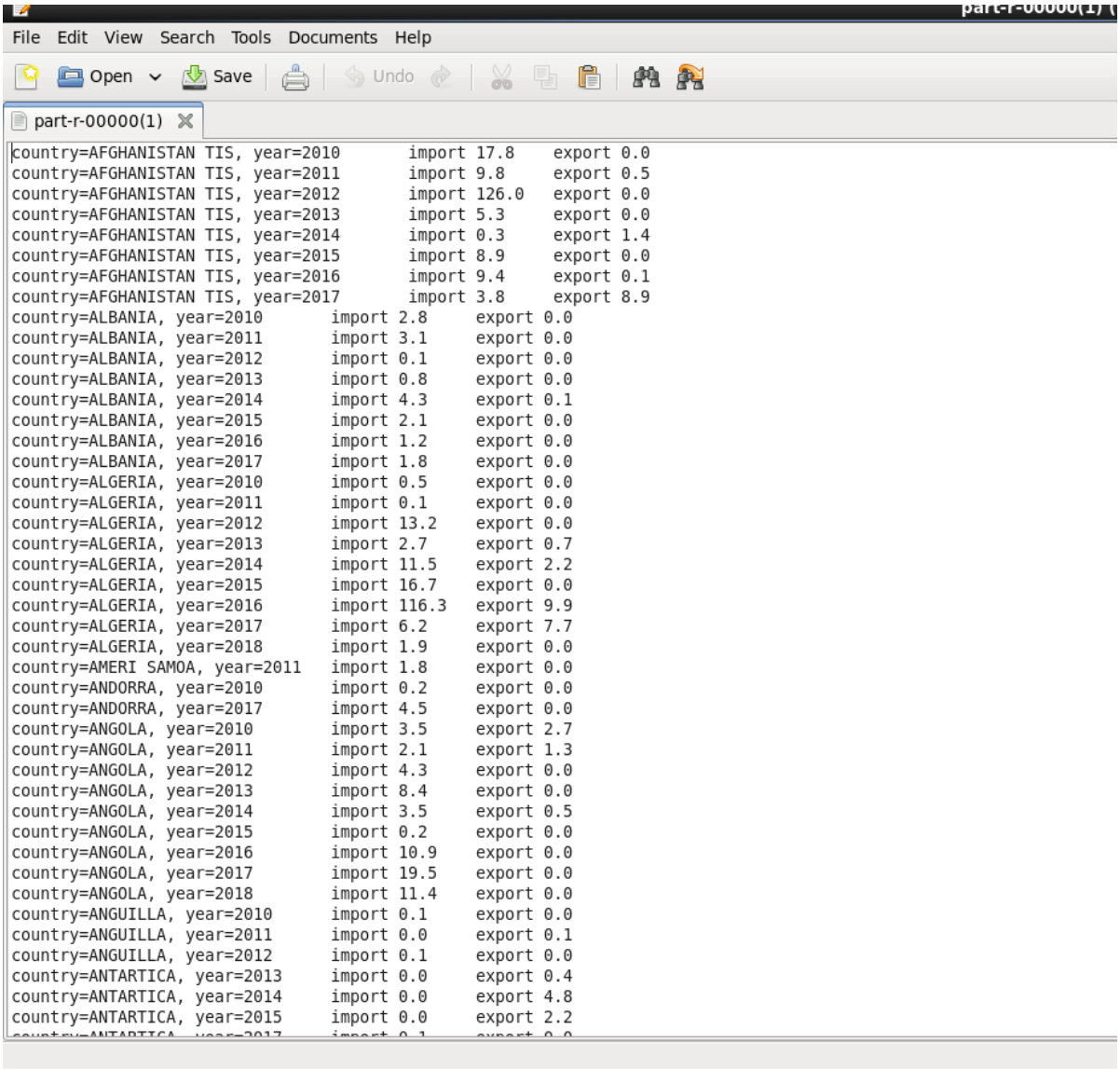




Checking for output file.



Output:



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