

# HOME ASSIGNMENT 1

**Roll No:** 208W1A1299

**Name:** MOHAMMAD RIZWANULLAH

**PROJECT NAME :** Twitter data analysis.

**The output of Mapreduce will be in below format**

The output of the data preprocessing will be to identify the day in the weekend and hour in the day where celebrities are tweeting in twitter

**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:**

```
Applications Places System cloudera@quickstart:~/Desktop/Worldwide-trade-data Thu Sep 29, 9:24 AM
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ cd /home/cloudera/Desktop/Worldwide-trade-data
[cloudera@quickstart Worldwide-trade-data]$ ls
2018-2019_export.csv indiantradedata.jar sanhad.jar
2018-2019_import.csv out
[cloudera@quickstart Worldwide-trade-data]$ hadoop jar Worldwidetrade.jar com.sahad.app.CYApp /user/cloudera/indiantradedata/2018-2019_export.csv /user/cloudera/indiantradedata/output2
usage: hadoop jar indiantradedata.jar <com.sahad.app.CYApp /com.sahad.app.CYApp>
p <input-data> <export-data> <output-path>
[cloudera@quickstart Worldwide-trade-data]$ hadoop jar Worldwidetrade.jar com.sahad.app.CYApp /user/cloudera/indiantradedata/2018-2019_export.csv /user/cloudera/indiantradedata/2018-2019_import.csv /user/cloudera/indiantradedata
had.app.CYApp /user/cloudera/indiantradedata/2018-2019_export.csv /user/cloudera/indiantradedata/2018-2019_import.csv
22/09/29 09:22:17 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
22/09/29 09:22:18 INFO input.FileInputFormat: Total input paths to process : 1
22/09/29 09:22:18 INFO input.FileInputFormat: Total input paths to process : 1
22/09/29 09:22:18 INFO mapreduce.JobSubmitter: Number of splits:2
22/09/29 09:22:18 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1664462633488_0002
22/09/29 09:22:19 INFO impl.YarnClientImpl: Submitted application application_1664462633488_0002
22/09/29 09:22:19 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8080/proxy/application_1664462633488_0002/
22/09/29 09:22:19 INFO mapreduce.Job: Running job: job_1664462633488_0002
22/09/29 09:22:26 INFO mapreduce.Job: Job job_1664462633488_0002 running in uber mode : false
22/09/29 09:22:26 INFO mapreduce.Job: map 0% reduce 0%
22/09/29 09:22:38 INFO mapreduce.Job: map 100% reduce 0%
22/09/29 09:22:45 INFO mapreduce.Job: map 100% reduce 100%
22/09/29 09:22:46 INFO mapreduce.Job: Job job_1664462633488_0002 completed successfully
22/09/29 09:22:46 INFO mapreduce.Job: Counters: 59
File System Counters:
  FILE: Number of bytes read=251567
  FILE: Number of bytes written=936834
  FILE: Number of read operations=9
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=10412801
  HDFS: Number of bytes written=86561
  HDFS: Number of read operations=9
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
Job Counters:
  Killed map tasks=1
  Launched map tasks=2
  Launched reduce tasks=1
  Data-local map tasks=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 reduce tasks (ms)=5582
  Total vcore-milliseconds taken by all map tasks=19449
  Total vcore-milliseconds taken by all reduce tasks=5582
  Total megabyte-milliseconds taken by all map tasks=19915776
  Total megabyte-milliseconds taken by all reduce tasks=5715968
[Worldwide-trade-data] cloudera@quickstart:~/Desktop/Worldwide-trade-data
```

```
Killed map tasks=1
Launched map tasks=2
Launched reduce tasks=1
Data-local map tasks=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 reduce tasks (ms)=5582
Total vcore-milliseconds taken by all map tasks=19449
Total vcore-milliseconds taken by all reduce tasks=5582
Total megabyte-milliseconds taken by all map tasks=19915776
Total megabyte-milliseconds taken by all reduce tasks=5715968
```

#### Map-Reduce Framework

```
Map input records=213149
Map output records=9313
Map output bytes=232875
Map output materialized bytes=251513
Input split bytes=578
Combine input records=0
Combine output records=0
Reduce input groups=1677
Reduce shuffle bytes=251513
Reduce input records=9313
Reduce output records=1677
Spilled Records=18620
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=399
CPU time spent (ms)=4360
Physical memory (bytes) snapshot=563458848
Virtual memory (bytes) snapshot=4519178240
Total committed heap usage (bytes)=391579008
```

#### 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=0
```

#### File Output Format Counters

```
Bytes Written=86561
```

```
[cloudera@quickstart Worldwide-trade-data]$ hadoop fs -ls /user/cloudera/indiantradedata/output2
and 2 items
```

```
w-r--r-- 1 cloudera cloudera 0 2022-09-29 09:22 /user/cloudera/indiantradedata/output2/ SUCCESS
w-r--r-- 1 cloudera cloudera 86561 2022-09-29 09:22 /user/cloudera/indiantradedata/output2/part-r-00000
[cloudera@quickstart Worldwide-trade-data]$ hadoop fs -ls /user/cloudera/indiantradedata/output2/
```