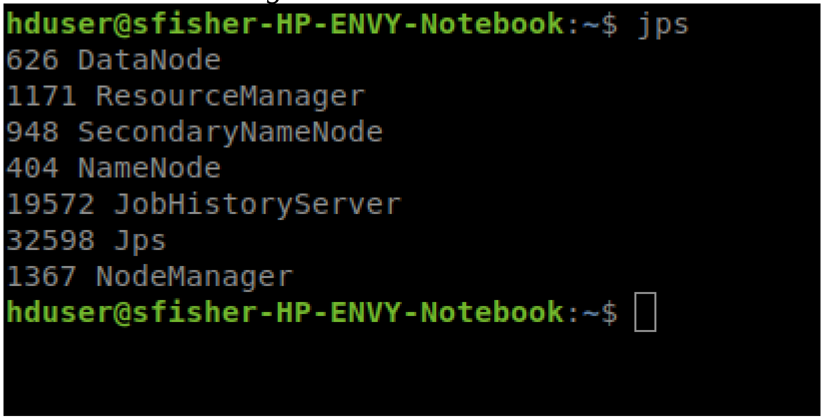
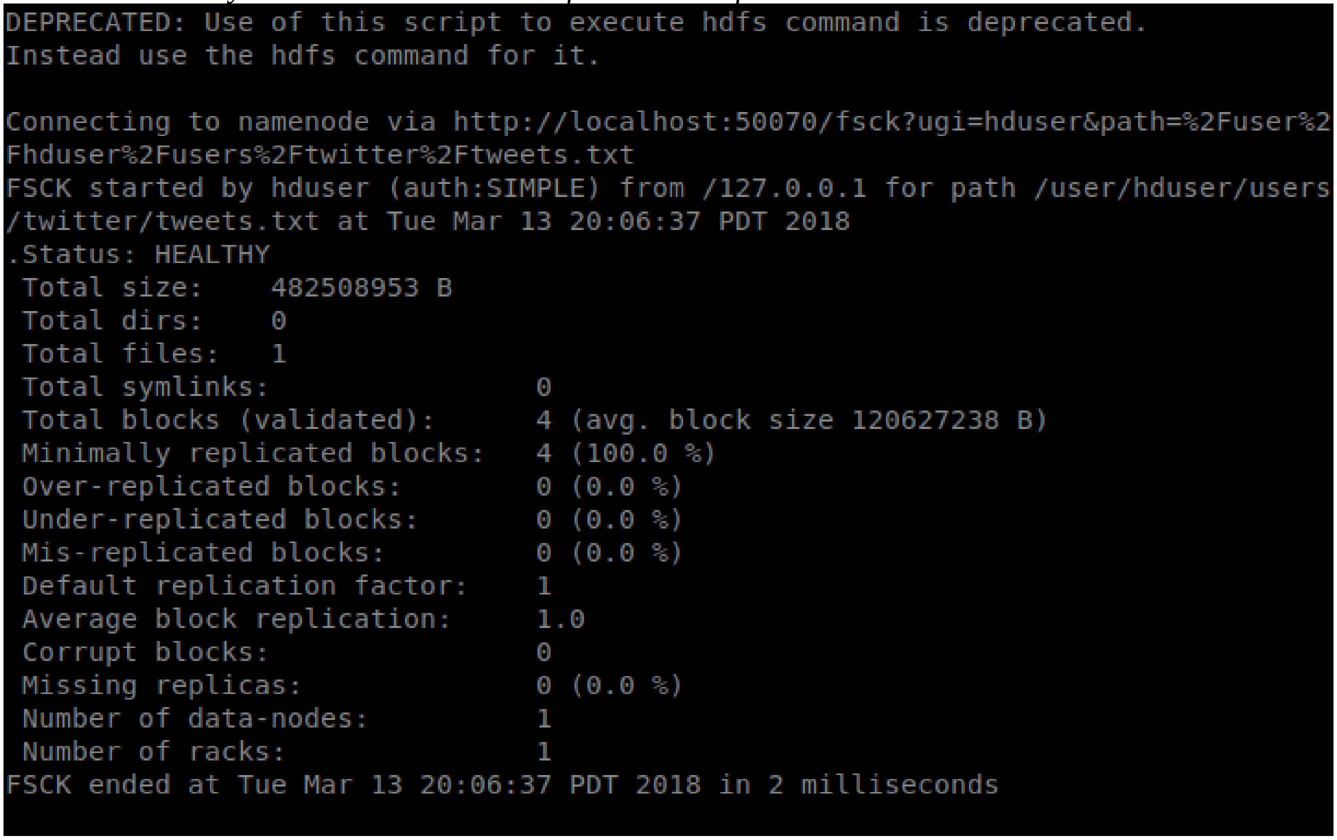
Steven Fisher
HW1
TTH 3:00 PM – 4:15 PM

1. *Run "jps" command and show running tasks:*



```
hduser@sfisher-HP-ENVY-Notebook:~$ jps
626 DataNode
1171 ResourceManager
948 SecondaryNameNode
404 NameNode
19572 JobHistoryServer
32598 Jps
1367 NodeManager
hduser@sfisher-HP-ENVY-Notebook:~$
```

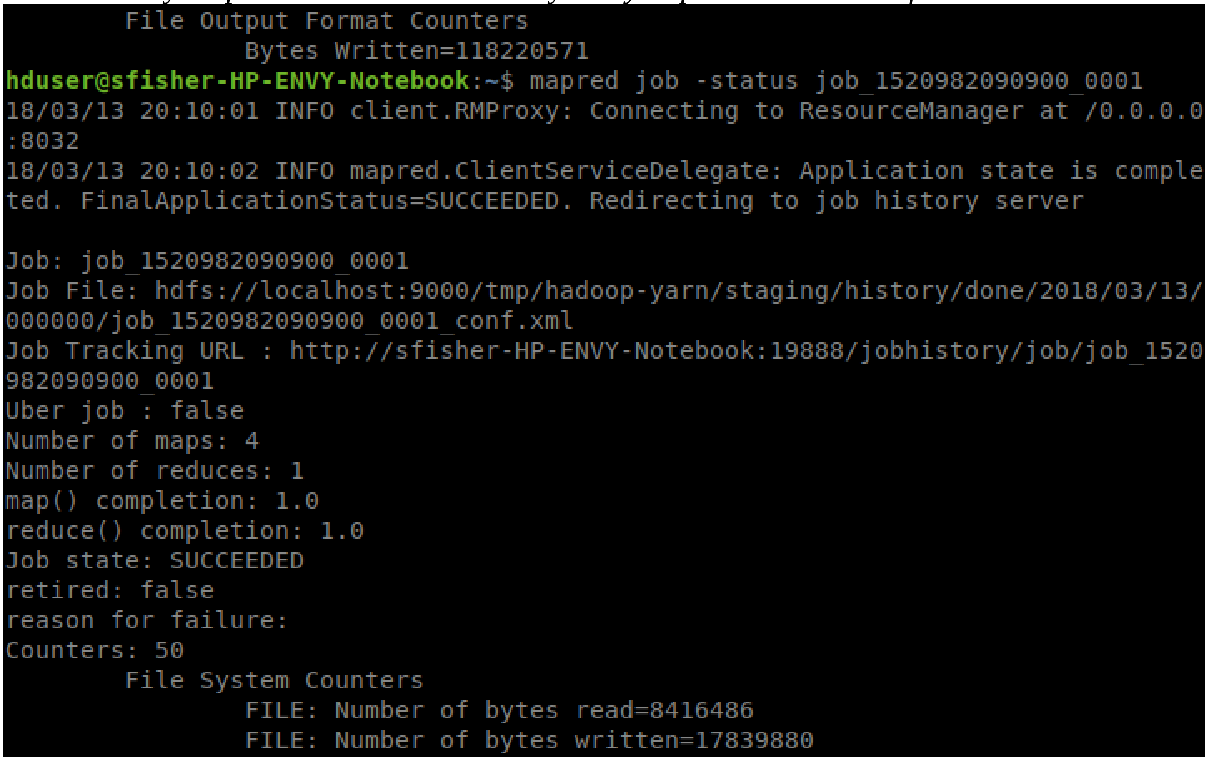
2. *Show how many blocks created in HDFS for "tweets" file:*



```
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Connecting to namenode via http://localhost:50070/fsck?ugi=hduser&path=%2Fuser%2
Fhduser%2Fusers%2Ftwitter%2Ftweets.txt
FSCK started by hduser (auth:SIMPLE) from /127.0.0.1 for path /user/hduser/users
/twitter/tweets.txt at Tue Mar 13 20:06:37 PDT 2018
.Status: HEALTHY
 Total size:    482508953 B
 Total dirs:    0
 Total files:   1
 Total symlinks:                0
 Total blocks (validated):      4 (avg. block size 120627238 B)
 Minimally replicated blocks:   4 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       0 (0.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    1
 Average block replication:     1.0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)
 Number of data-nodes:          1
 Number of racks:               1
FSCK ended at Tue Mar 13 20:06:37 PDT 2018 in 2 milliseconds
```

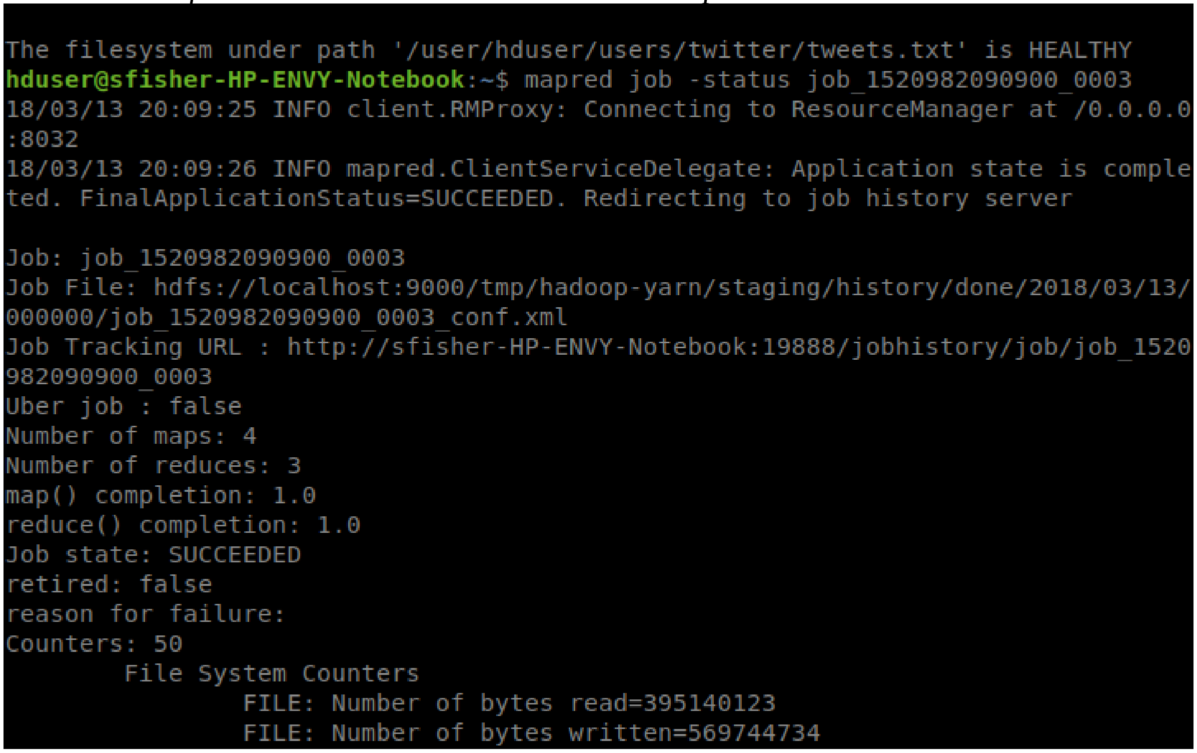
3. *Show how many map tasks are created when you try to process “tweets” file in HDFS*



```
        File Output Format Counters
                Bytes Written=118220571
hduser@sfisher-HP-ENVY-Notebook:~$ mapred job -status job_1520982090900_0001
18/03/13 20:10:01 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
18/03/13 20:10:02 INFO mapred.ClientServiceDelegate: Application state is comple
ted. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server

Job: job_1520982090900_0001
Job File: hdfs://localhost:9000/tmp/hadoop-yarn/staging/history/done/2018/03/13/
000000/job_1520982090900_0001_conf.xml
Job Tracking URL : http://sfisher-HP-ENVY-Notebook:19888/jobhistory/job/job_1520
982090900_0001
Uber job : false
Number of maps: 4
Number of reduces: 1
map() completion: 1.0
reduce() completion: 1.0
Job state: SUCCEEDED
retired: false
reason for failure:
Counters: 50
        File System Counters
                FILE: Number of bytes read=8416486
                FILE: Number of bytes written=17839880
```

4. *Set the number of reduce tasks to 3 and show that Hadoop created 3 reduce tasks:*



```
The filesystem under path '/user/hduser/users/twitter/tweets.txt' is HEALTHY
hduser@sfisher-HP-ENVY-Notebook:~$ mapred job -status job_1520982090900_0003
18/03/13 20:09:25 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
18/03/13 20:09:26 INFO mapred.ClientServiceDelegate: Application state is comple
ted. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server

Job: job_1520982090900_0003
Job File: hdfs://localhost:9000/tmp/hadoop-yarn/staging/history/done/2018/03/13/
000000/job_1520982090900_0003_conf.xml
Job Tracking URL : http://sfisher-HP-ENVY-Notebook:19888/jobhistory/job/job_1520
982090900_0003
Uber job : false
Number of maps: 4
Number of reduces: 3
map() completion: 1.0
reduce() completion: 1.0
Job state: SUCCEEDED
retired: false
reason for failure:
Counters: 50
        File System Counters
                FILE: Number of bytes read=395140123
                FILE: Number of bytes written=569744734
```

5. *Write a MapReduce code to count the occurrences of hashtags and find the most repeated 100 hashtags*.

In our program when utilized the ideas at the following website, to help with the sorting by value.
https://dzenanhamzic.com/2016/09/21/java-mapreduce-for-top-n-twitter-hashtags/

Command used for hashtag count:

```
hadoop jar HashCount.jar HashCount <INPUT FILE> <OUTPUT FILE>
```

Example: hadoop jar HashCount.jar HashCount /user/hduser/users/twitter/tweets.txt
/user/hduser/users/twitter/HashOut

6. *Write a MapReduce code find the most tweeted 5 days. (Tweets are associated with time stamps so you need to count all the tweets posted in same days)*

In our program when utilized the ideas at the following website.
https://dzenanhamzic.com/2016/09/21/java-mapreduce-for-top-n-twitter-hashtags/

Command used to count top 5 most tweeted days:

```
hadoop jar TweetCount.jar TweetCount  <INPUT FILE> <OUTPUT FILE>
```

Example: hadoop jar TweetCount.jar TweetCount /user/hduser/users/twitter/tweets.txt
/user/hduser/users/twitter/TweetOut

7. *Write a MapReduce code to find the most tweeted 10 cities along with the number of tweets ("training_set_users.txt" file has user_id → city relation to extract city information)*

In our code we utilized the ideas at the following websites:
https://www.edureka.co/blog/mapreduce-example-reduce-side-join/
https://stackoverflow.com/questions/2499585/chaining-multiple-mapreduce-jobs-in-hadoop

Command used for city count:

```
hadoop jar MostTweetedCities.jar MostTweetedCities <USERS INPUT FILE> <TWEETS INPUT
FILE> <TEMP OUTPUT> <OUTPUT FILE>
```

Example: hadoop jar CityCount.jar CityCount /user/hduser/users/twitter/users.txt
/user/hduser/users/twitter/tweets.txt /user/hduser/users/twitter/CityTmp1
/user/hduser/users/twitter/CityOut1