Renat Norderhaug
CS 433
3/6/20

Homework 1: Tweet analysis with MapReduce

1. HDFS daemons

```
rnorderhaug@cs433:~$ ps -ef | grep hadoop | grep -P 'namenode|datanode|tasktracker|jobtracker'
root     23234     1  0 Feb06 ?        01:13:39 /usr/lib/jvm/java-8-openjdk-amd64//bin/java -Dproc_namenode -Djava.net.preferIPv4Stack=true -
Dhdfs.audit.logger=INFO,NullAppender -Dhadoop.security.logger=INFO,RFAS -Dyarn.log.dir=/opt/hadoop-3.2.1//logs -Dyarn.log.file=hadoop-root-na
menode-cs433.log -Dyarn.home.dir=/opt/hadoop-3.2.1/ -Dyarn.root.logger=INFO,console -Djava.library.path=/opt/hadoop-3.2.1//lib/native -Dhadoo
p.log.dir=/opt/hadoop-3.2.1//logs -Dhadoop.log.file=hadoop-root-namenode-cs433.log -Dhadoop.home.dir=/opt/hadoop-3.2.1/ -Dhadoop.id.str=root
-Dhadoop.root.logger=INFO,RFA -Dhadoop.policy.file=hadoop-policy.xml org.apache.hadoop.hdfs.server.namenode.NameNode
root     23396     1  0 Feb06 ?        01:00:59 /usr/lib/jvm/java-8-openjdk-amd64//bin/java -Dproc_datanode -Djava.net.preferIPv4Stack=true -
Dhadoop.security.logger=ERROR,RFAS -Dyarn.log.dir=/opt/hadoop-3.2.1//logs -Dyarn.log.file=hadoop-root-datanode-cs433.log -Dyarn.home.dir=/opt
/hadoop-3.2.1/ -Dyarn.root.logger=INFO,console -Djava.library.path=/opt/hadoop-3.2.1//lib/native -Dhadoop.log.dir=/opt/hadoop-3.2.1//logs -Dh
adoop.log.file=hadoop-root-datanode-cs433.log -Dhadoop.home.dir=/opt/hadoop-3.2.1/ -Dhadoop.id.str=root -Dhadoop.root.logger=INFO,RFA -Dhadoo
p.policy.file=hadoop-policy.xml org.apache.hadoop.hdfs.server.datanode.DataNode
root     23742     1  0 Feb06 ?        00:47:14 /usr/lib/jvm/java-8-openjdk-amd64//bin/java -Dproc_secondarynamenode -Djava.net.preferIPv4Sta
ck=true -Dhdfs.audit.logger=INFO,NullAppender -Dhadoop.security.logger=INFO,RFAS -Dyarn.log.dir=/opt/hadoop-3.2.1//logs -Dyarn.log.file=hadoo
p-root-secondarynamenode-cs433.log -Dyarn.home.dir=/opt/hadoop-3.2.1/ -Dyarn.root.logger=INFO,console -Djava.library.path=/opt/hadoop-3.2.1//
lib/native -Dhadoop.log.dir=/opt/hadoop-3.2.1//logs -Dhadoop.log.file=hadoop-root-secondarynamenode-cs433.log -Dhadoop.home.dir=/opt/hadoop-3
.2.1/ -Dhadoop.id.str=root -Dhadoop.root.logger=INFO,RFA -Dhadoop.policy.file=hadoop-policy.xml org.apache.hadoop.hdfs.server.namenode.Second
aryNameNode
rnorderhaug@cs433:~$
```

2. There are 4 blocks for the training_set_tweets file

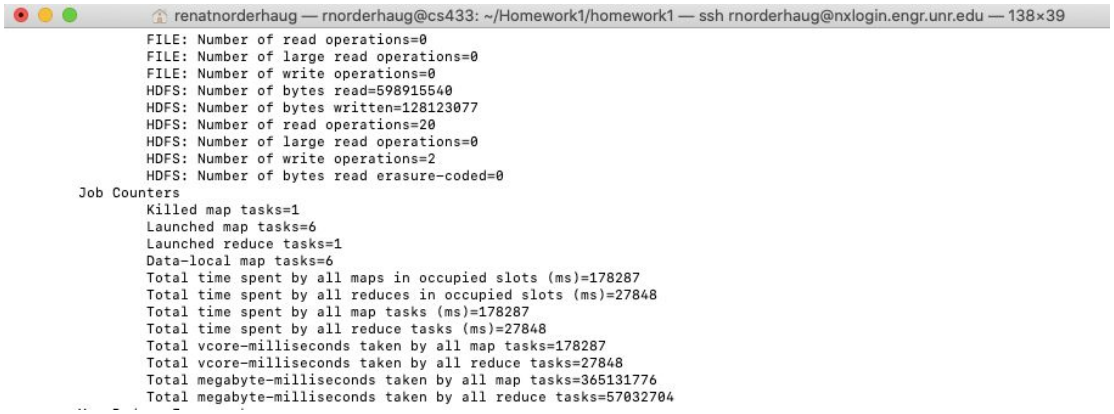
```
rnorderhaug@cs433:~$ hadoop fsck /homework1/training_set_tweets.txt -files -blocks
WARNING: Use of this script to execute fsck is deprecated.
WARNING: Attempting to execute replacement "hdfs fsck" instead.

Connecting to namenode via http://localhost:9870/fsck?ugi=rnorderhaug&files=1&blocks=1&path=%2Fhom
work1%2Ftraining_set_tweets.txt
FSCK started by rnorderhaug (auth:SIMPLE) from /127.0.0.1 for path /homework1/training_set_tweets.
xt at Fri Mar 06 16:13:15 UTC 2020
/homework1/training_set_tweets.txt 482508953 bytes, replicated: replication=1, 4 block(s):  OK
0. BP-1227683103-127.0.1.1-1579204754407:blk_1073741894_1070 len=134217728 Live_repl=1
1. BP-1227683103-127.0.1.1-1579204754407:blk_1073741895_1071 len=134217728 Live_repl=1
2. BP-1227683103-127.0.1.1-1579204754407:blk_1073741896_1072 len=134217728 Live_repl=1
3. BP-1227683103-127.0.1.1-1579204754407:blk_1073741897_1073 len=79855769 Live_repl=1


Status: HEALTHY
 Number of data-nodes:  1
 Number of racks:               1
 Total dirs:                    0
 Total symlinks:                0

Replicated Blocks:
 Total size:    482508953 B
 Total files:   1
 Total blocks (validated):      4 (avg. block size 120627238 B)
```

3. The number of map tasks is dependent on the data volume, blocks size, split size. From the picture below



4. The replication factor is now set to 3 and mapreduce is reran. The launched reduce tasks = 3, this was edited in the .java file and in the commandLine