CSC555 Assignment 2
Sarah Cummings

1) From Mining Massive Data Sets:

   a. Exercise 1.3.3: Suppose hash-keys are drawn from the population of all non-negative integers that are multiples of some constant c, and hash function h(x) is x mod 15. For what values of c will h be a suitable hash function, i.e., a large random choice of hash-keys will be divided roughly equally into buckets?

Given our hash function of x mod 15, we will have 15 buckets ranging from 0 to 14 for our hash keys. If c= 3, our set is

x= {3, 6, 9, 12, 15, 18, 21, 24, 27. 30, 33 }

h(x)= {3, 6, 9, 12, 15, 3, 6, 9, 12, 0, 3.. }

This would be a bad choice because only the buckets that are multiples of c will be used. To use all buckets roughly equally, we must choose a c like 1, 16 or 31. A c value that is not a multiple or divisor of 15 would work best and we would end up using all of the buckets.

   b. Exercise 2.2.1 a. : Suppose we execute the word-count MapReduce program described in this section on a large repository such as a copy of the Web. We shall use 100 Map tasks and some number of Reduce tasks. Suppose we do not use a combiner at the Map tasks. Do you expect there to be significant skew in the times taken by the various reducers to process their value list? Why or why not?

In our textbook example, using this process for a repository of documents would likely produce a skew since there could be significant variation in the lengths of the value lists for different keys. In this case, different reducers would take different amounts of time. However, given such a large source as the web, if reduce task are sent to reducer randomly, we can assume that the times would average out and there would be very little skew.

2) Describe a how to implement a MapReduce job for the following scenarios:

   a. Exercise 2.3.1 -a: Design MapReduce algorithms to take a very large file of integers and produce as output the largest integer.

For this MapReduce task, the mapper would take a chunk of integers as input produce the max value of each chunk as a key. The reducer then takes each value and finds the overall maximum as a value.

b. Exercise 2.3.1 -b: Design MapReduce algorithms to take a very large file of integers and produce as output the average.

For this MapReduce task, the mapper must take the integers as input and return an average and a count of the numbers processed for each chunk as a key-value pair. Then, the reducer would take in the key-value pairs and produce a weighted average.

c. For a Student table (ID, Name, Address, Phone), convert

SELECT ID, COUNT(DISTINCT Name)

FROM Student

GROUP BY ID;

For this MapReduce task, ID is the key and Name is the value. The mapper would output the IDs and names, and the reducer would count the number of names associated with each ID.

d. For Employee(EID, EFirst, ELast, Age) and Customer(CID, CFirst, CLast, Address), find everyone with the same name using MapReduce:

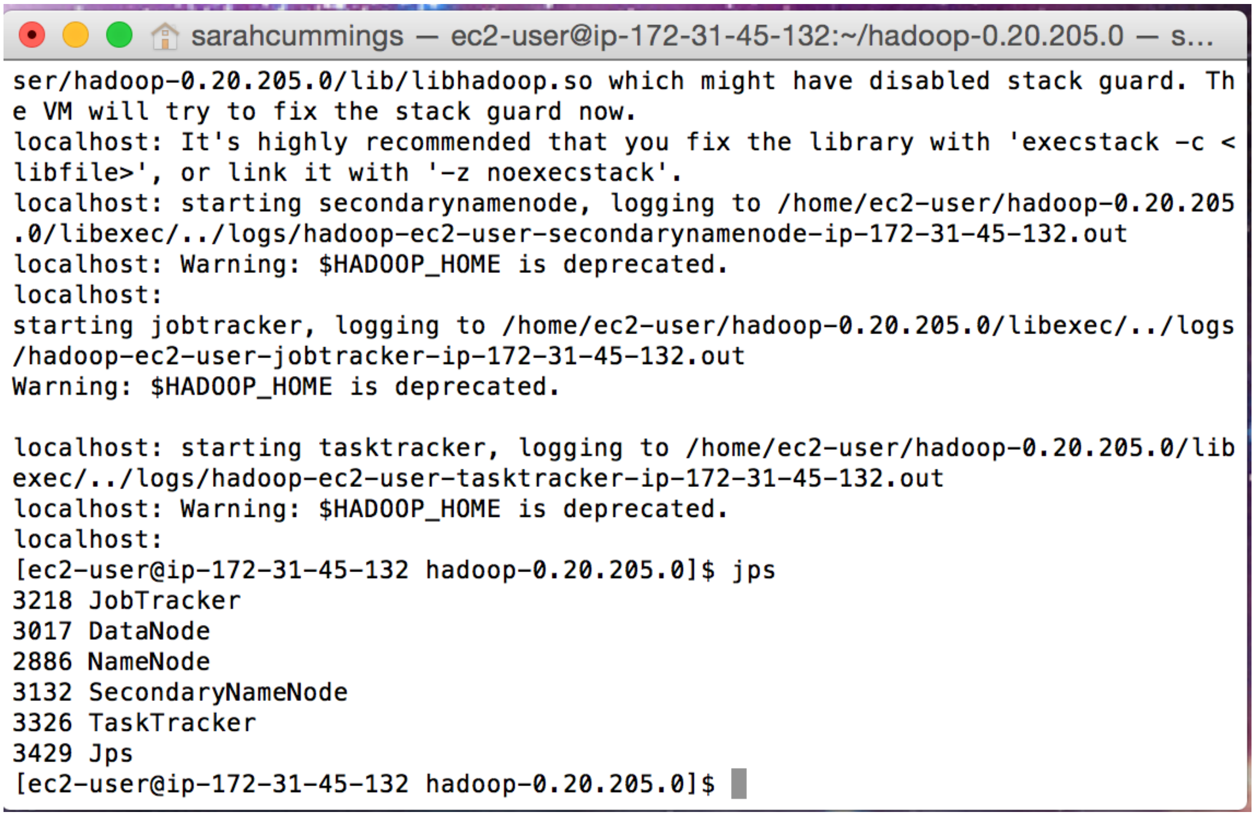
SELECT First, Last, EID, CID, Address

FROM Employee, Customer

WHERE EFirst = CFirst AND ELast = CLast;

For this MapReduce task, the first mapper would take in the First, Last, EID and Adress. The second mapper would take in the First, Last, CID and address. The reducer would then match the keys from each of these mappers where EFirst = CFirst AND ELast = CLast.

3) Installing and setting up hadoop:

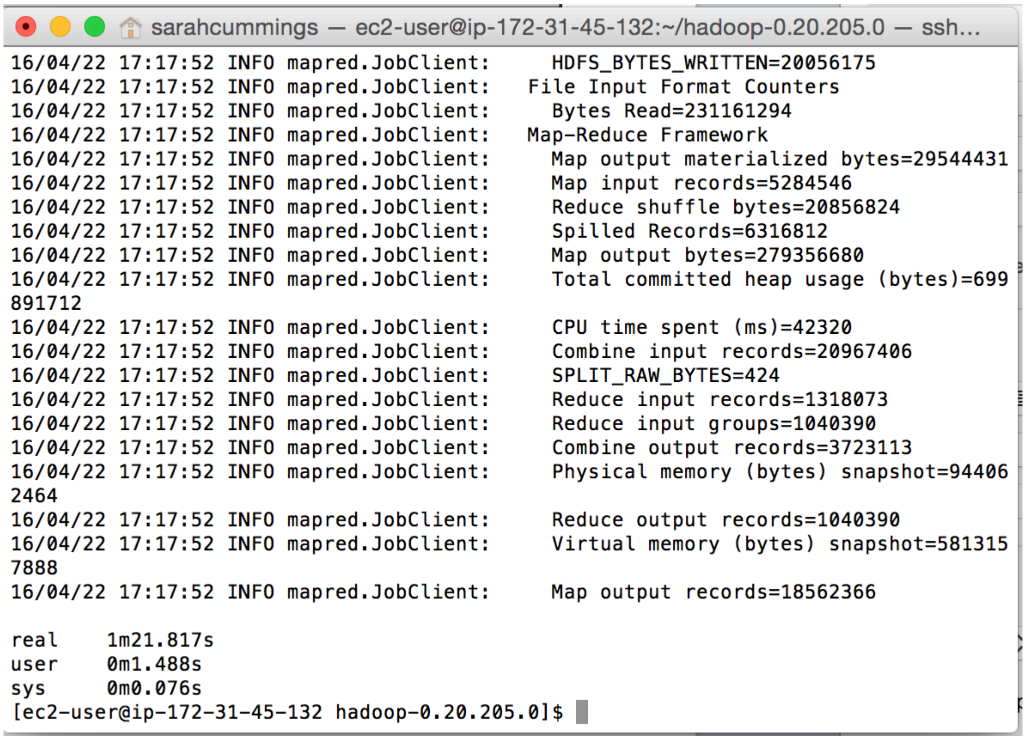


sarahcummings — ec2-user@ip-172-31-45-132:~/hadoop-0.20.205.0 — s…

```
ser/hadoop-0.20.205.0/lib/libhadoop.so which might have disabled stack guard. Th
e VM will try to fix the stack guard now.
localhost: It's highly recommended that you fix the library with 'execstack -c <
libfile>', or link it with '-z noexecstack'.
localhost: starting secondarynamenode, logging to /home/ec2-user/hadoop-0.20.205
.0/libexec/../logs/hadoop-ec2-user-secondarynamenode-ip-172-31-45-132.out
localhost: Warning: $HADOOP_HOME is deprecated.
localhost:
starting jobtracker, logging to /home/ec2-user/hadoop-0.20.205.0/libexec/../logs
/hadoop-ec2-user-jobtracker-ip-172-31-45-132.out
Warning: $HADOOP_HOME is deprecated.

localhost: starting tasktracker, logging to /home/ec2-user/hadoop-0.20.205.0/lib
exec/../logs/hadoop-ec2-user-tasktracker-ip-172-31-45-132.out
localhost: Warning: $HADOOP_HOME is deprecated.
localhost:
[ec2-user@ip-172-31-45-132 hadoop-0.20.205.0]$ jps
3218 JobTracker
3017 DataNode
2886 NameNode
3132 SecondaryNameNode
3326 TaskTracker
3429 Jps
[ec2-user@ip-172-31-45-132 hadoop-0.20.205.0]$
```

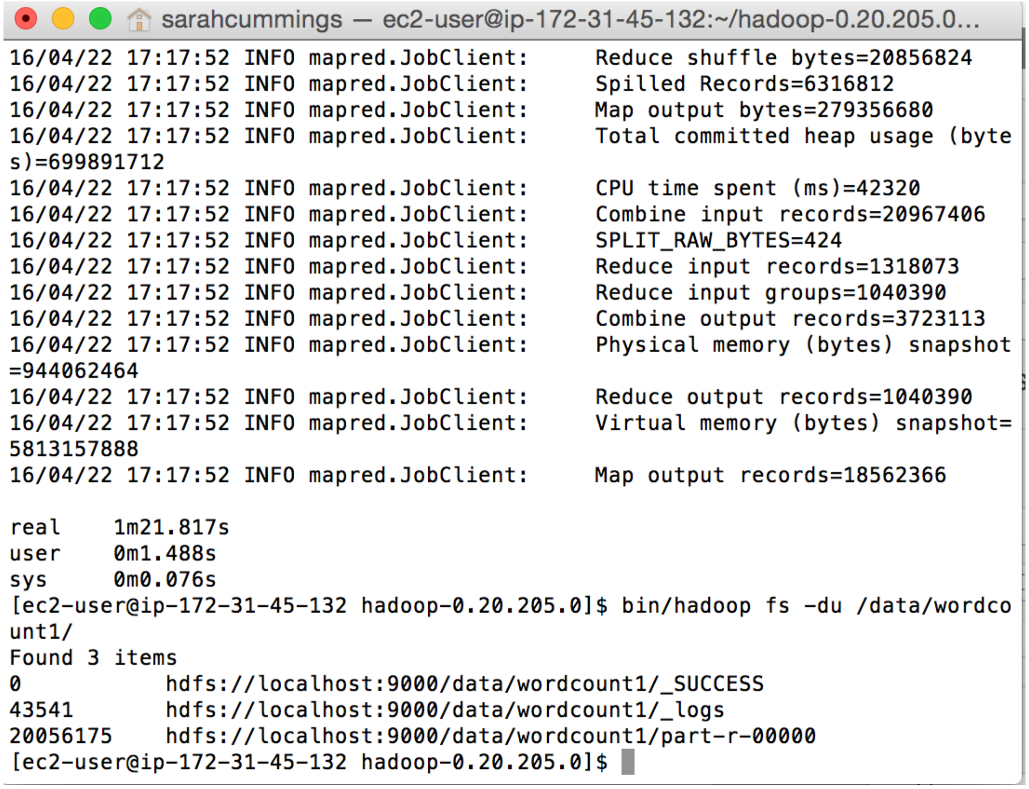
I’m not sure why I get the warning that Hadoop Home is deprecated, but no matter what I do I can’t get it to go away…. hope this isn’t an issue.

4) Downloading file from web. After running the word count— screen shot containing the time it took to execute:

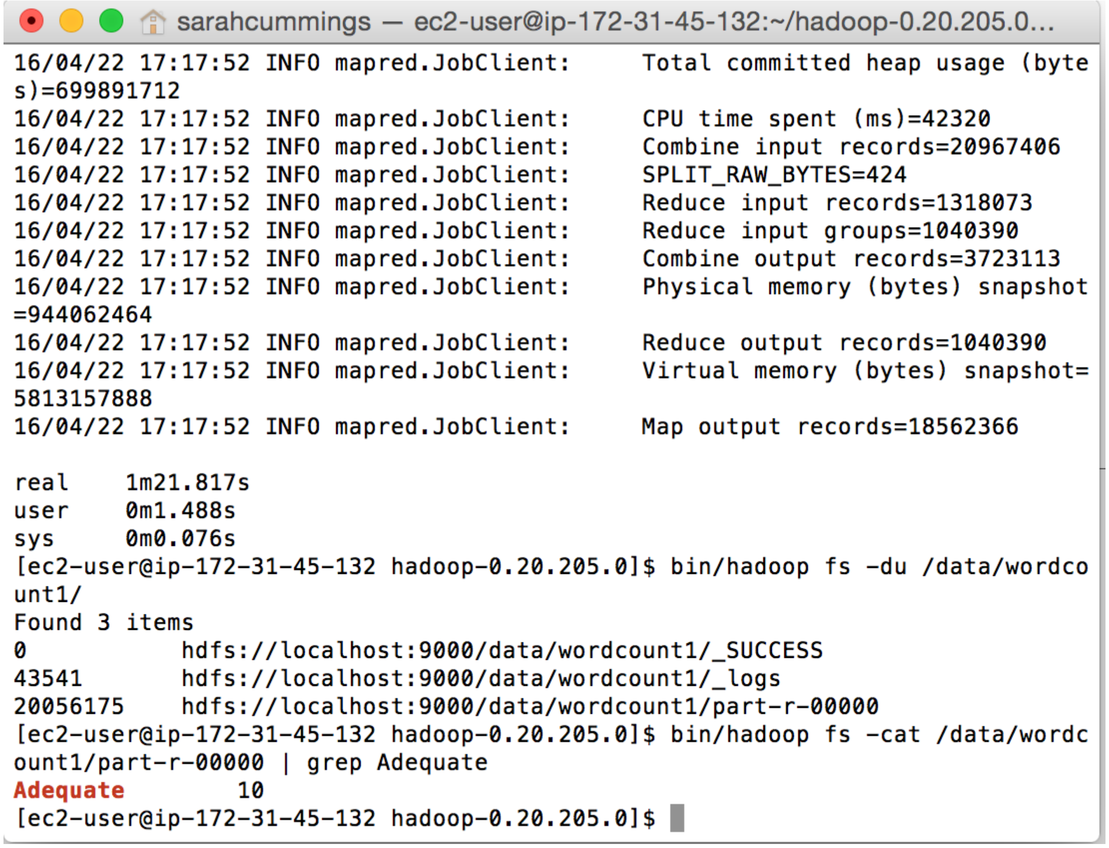

sarahcummings — ec2-user@ip-172-31-45-132:~/hadoop-0.20.205.0 — ssh...

```
16/04/22 17:17:52 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=20056175
16/04/22 17:17:52 INFO mapred.JobClient:   File Input Format Counters
16/04/22 17:17:52 INFO mapred.JobClient:     Bytes Read=231161294
16/04/22 17:17:52 INFO mapred.JobClient:   Map-Reduce Framework
16/04/22 17:17:52 INFO mapred.JobClient:     Map output materialized bytes=29544431
16/04/22 17:17:52 INFO mapred.JobClient:     Map input records=5284546
16/04/22 17:17:52 INFO mapred.JobClient:     Reduce shuffle bytes=20856824
16/04/22 17:17:52 INFO mapred.JobClient:     Spilled Records=6316812
16/04/22 17:17:52 INFO mapred.JobClient:     Map output bytes=279356680
16/04/22 17:17:52 INFO mapred.JobClient:     Total committed heap usage (bytes)=699
891712
16/04/22 17:17:52 INFO mapred.JobClient:     CPU time spent (ms)=42320
16/04/22 17:17:52 INFO mapred.JobClient:     Combine input records=20967406
16/04/22 17:17:52 INFO mapred.JobClient:     SPLIT_RAW_BYTES=424
16/04/22 17:17:52 INFO mapred.JobClient:     Reduce input records=1318073
16/04/22 17:17:52 INFO mapred.JobClient:     Reduce input groups=1040390
16/04/22 17:17:52 INFO mapred.JobClient:     Combine output records=3723113
16/04/22 17:17:52 INFO mapred.JobClient:     Physical memory (bytes) snapshot=94406
2464
16/04/22 17:17:52 INFO mapred.JobClient:     Reduce output records=1040390
16/04/22 17:17:52 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=581315
7888
16/04/22 17:17:52 INFO mapred.JobClient:     Map output records=18562366

real    1m21.817s
user    0m1.488s
sys     0m0.076s
[ec2-user@ip-172-31-45-132 hadoop-0.20.205.0]$
```

Size of the output file and count for the word Adequate:

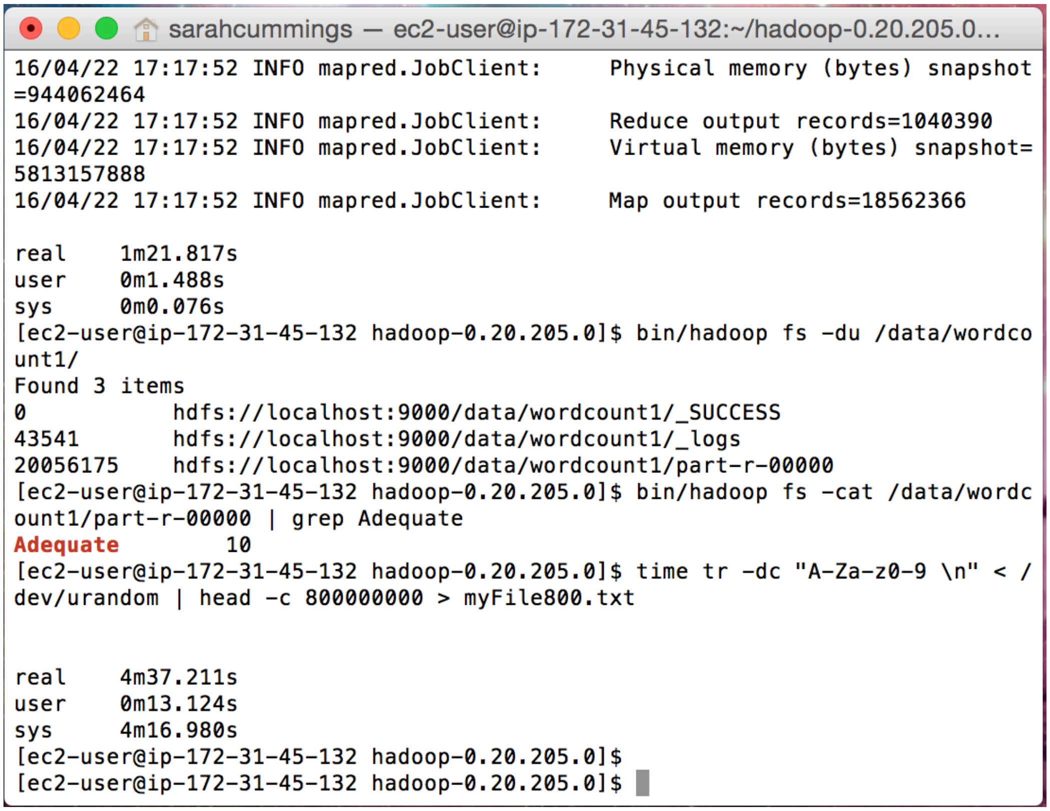

sarahcummings — ec2-user@ip-172-31-45-132:~/hadoop-0.20.205.0...

```
16/04/22 17:17:52 INFO mapred.JobClient:     Reduce shuffle bytes=20856824
16/04/22 17:17:52 INFO mapred.JobClient:     Spilled Records=6316812
16/04/22 17:17:52 INFO mapred.JobClient:     Map output bytes=279356680
16/04/22 17:17:52 INFO mapred.JobClient:     Total committed heap usage (byte
s)=699891712
16/04/22 17:17:52 INFO mapred.JobClient:     CPU time spent (ms)=42320
16/04/22 17:17:52 INFO mapred.JobClient:     Combine input records=20967406
16/04/22 17:17:52 INFO mapred.JobClient:     SPLIT_RAW_BYTES=424
16/04/22 17:17:52 INFO mapred.JobClient:     Reduce input records=1318073
16/04/22 17:17:52 INFO mapred.JobClient:     Reduce input groups=1040390
16/04/22 17:17:52 INFO mapred.JobClient:     Combine output records=3723113
16/04/22 17:17:52 INFO mapred.JobClient:     Physical memory (bytes) snapshot
=944062464
16/04/22 17:17:52 INFO mapred.JobClient:     Reduce output records=1040390
16/04/22 17:17:52 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=
5813157888
16/04/22 17:17:52 INFO mapred.JobClient:     Map output records=18562366

real    1m21.817s
user    0m1.488s
sys     0m0.076s
[ec2-user@ip-172-31-45-132 hadoop-0.20.205.0]$ bin/hadoop fs -du /data/wordco
unt1/
Found 3 items
0           hdfs://localhost:9000/data/wordcount1/_SUCCESS
43541       hdfs://localhost:9000/data/wordcount1/_logs
20056175    hdfs://localhost:9000/data/wordcount1/part-r-00000
[ec2-user@ip-172-31-45-132 hadoop-0.20.205.0]$
```


sarahcummings — ec2-user@ip-172-31-45-132:~/hadoop-0.20.205.0...

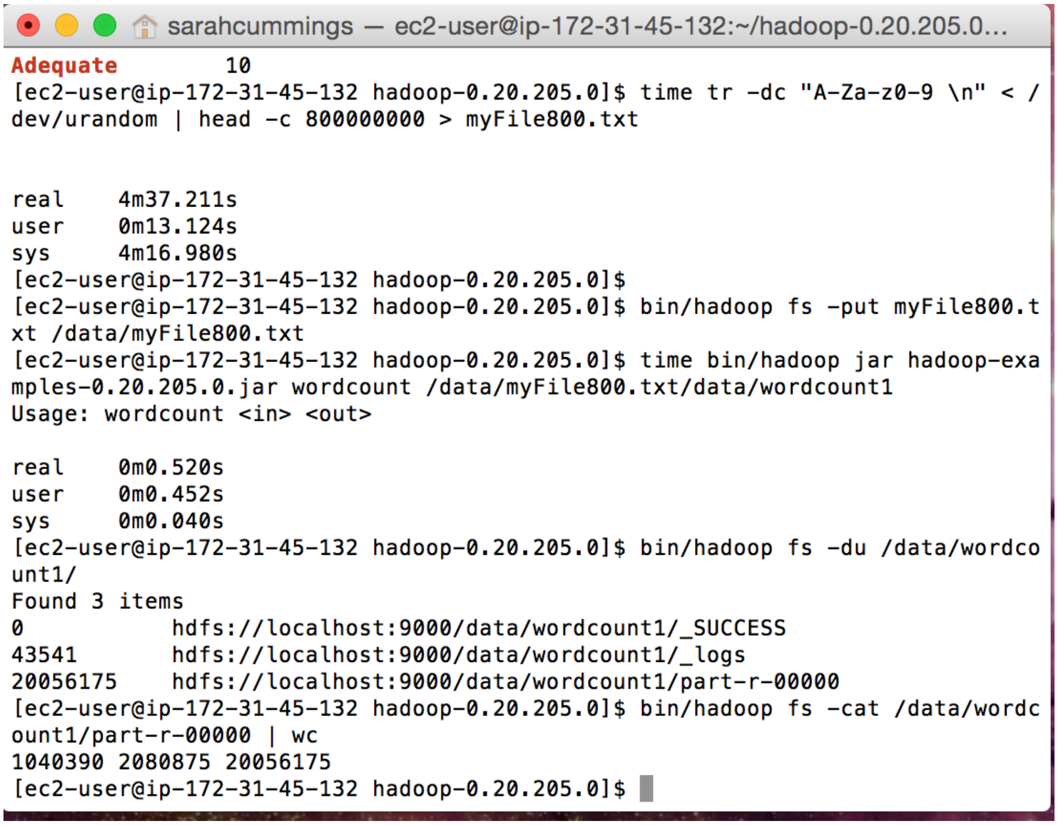
```
16/04/22 17:17:52 INFO mapred.JobClient:     Total committed heap usage (byte
s)=699891712
16/04/22 17:17:52 INFO mapred.JobClient:     CPU time spent (ms)=42320
16/04/22 17:17:52 INFO mapred.JobClient:     Combine input records=20967406
16/04/22 17:17:52 INFO mapred.JobClient:     SPLIT_RAW_BYTES=424
16/04/22 17:17:52 INFO mapred.JobClient:     Reduce input records=1318073
16/04/22 17:17:52 INFO mapred.JobClient:     Reduce input groups=1040390
16/04/22 17:17:52 INFO mapred.JobClient:     Combine output records=3723113
16/04/22 17:17:52 INFO mapred.JobClient:     Physical memory (bytes) snapshot
=944062464
16/04/22 17:17:52 INFO mapred.JobClient:     Reduce output records=1040390
16/04/22 17:17:52 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=
5813157888
16/04/22 17:17:52 INFO mapred.JobClient:     Map output records=18562366

real	1m21.817s
user	0m1.488s
sys	0m0.076s
[ec2-user@ip-172-31-45-132 hadoop-0.20.205.0]$ bin/hadoop fs -du /data/wordco
unt1/
Found 3 items
0           hdfs://localhost:9000/data/wordcount1/_SUCCESS
43541       hdfs://localhost:9000/data/wordcount1/_logs
20056175    hdfs://localhost:9000/data/wordcount1/part-r-00000
[ec2-user@ip-172-31-45-132 hadoop-0.20.205.0]$ bin/hadoop fs -cat /data/wordc
ount1/part-r-00000 | grep Adequate
Adequate        10
[ec2-user@ip-172-31-45-132 hadoop-0.20.205.0]$ 
```

Generating random file:


sarahcummings — ec2-user@ip-172-31-45-132:~/hadoop-0.20.205.0...

```
16/04/22 17:17:52 INFO mapred.JobClient:     Physical memory (bytes) snapshot
=944062464
16/04/22 17:17:52 INFO mapred.JobClient:     Reduce output records=1040390
16/04/22 17:17:52 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=
5813157888
16/04/22 17:17:52 INFO mapred.JobClient:     Map output records=18562366

real	1m21.817s
user	0m1.488s
sys	0m0.076s
[ec2-user@ip-172-31-45-132 hadoop-0.20.205.0]$ bin/hadoop fs -du /data/wordco
unt1/
Found 3 items
0           hdfs://localhost:9000/data/wordcount1/_SUCCESS
43541       hdfs://localhost:9000/data/wordcount1/_logs
20056175    hdfs://localhost:9000/data/wordcount1/part-r-00000
[ec2-user@ip-172-31-45-132 hadoop-0.20.205.0]$ bin/hadoop fs -cat /data/wordc
ount1/part-r-00000 | grep Adequate
Adequate        10
[ec2-user@ip-172-31-45-132 hadoop-0.20.205.0]$ time tr -dc "A-Za-z0-9 \n" < /
dev/urandom | head -c 800000000 > myFile800.txt


real	4m37.211s
user	0m13.124s
sys	4m16.980s
[ec2-user@ip-172-31-45-132 hadoop-0.20.205.0]$
[ec2-user@ip-172-31-45-132 hadoop-0.20.205.0]$ 
```

After running my word count, we find the random file has 2080875 words as seen below:



sarahcummings — ec2-user@ip-172-31-45-132:~/hadoop-0.20.205.0...

```
Adequate        10
[ec2-user@ip-172-31-45-132 hadoop-0.20.205.0]$ time tr -dc "A-Za-z0-9 \n" < /
dev/urandom | head -c 800000000 > myFile800.txt


real    4m37.211s
user    0m13.124s
sys     4m16.980s
[ec2-user@ip-172-31-45-132 hadoop-0.20.205.0]$
[ec2-user@ip-172-31-45-132 hadoop-0.20.205.0]$ bin/hadoop fs -put myFile800.t
xt /data/myFile800.txt
[ec2-user@ip-172-31-45-132 hadoop-0.20.205.0]$ time bin/hadoop jar hadoop-exa
mples-0.20.205.0.jar wordcount /data/myFile800.txt/data/wordcount1
Usage: wordcount <in> <out>

real    0m0.520s
user    0m0.452s
sys     0m0.040s
[ec2-user@ip-172-31-45-132 hadoop-0.20.205.0]$ bin/hadoop fs -du /data/wordco
unt1/
Found 3 items
0           hdfs://localhost:9000/data/wordcount1/_SUCCESS
43541       hdfs://localhost:9000/data/wordcount1/_logs
20056175    hdfs://localhost:9000/data/wordcount1/part-r-00000
[ec2-user@ip-172-31-45-132 hadoop-0.20.205.0]$ bin/hadoop fs -cat /data/wordc
ount1/part-r-00000 | wc
1040390 2080875 20056175
[ec2-user@ip-172-31-45-132 hadoop-0.20.205.0]$
```