

/usr/lib/jvm/java-1.8.0-openjdk-1.8.0.312.b07-1.amzn2.0.2.x86_64/jre/bin/java

Part 1.

1.a)

For the blocks, the mapper will prepare key value pair for the reduce output. The key will be Last and the value will be the Minimum Grade. The reducer will group the grades by last name.

1.b.)

For the blocks, the mapper will produce key value pairs where the keys will be City and State with the value being the count of distinct names. The reducer will be the group of names by city and state.

2.a.)

One method to speed up execution is to use a combiner function. A second method would be to adjust the number of map-reduce tasks.

2.b.)

The namenode is where the MapReduce framework would prefer restart the datanode or mapper task. The namenode receives periodic messages from the datanodes that indicate if they are available or have failed. Once a failure is detected, the namenode will then replicate or restart the datanode or mapper process.

2.c.)

- i.) There will be two output files generated.
- ii.) Using the formula $h(k) \text{ MOD } N$, with TimeOfAccess as a Key, then one hash function could be: $\text{TimeOfAccess MOD } 2$.

3.a.)

65 minutes

3.b.)

7 minutes

3.c.)

3 minutes

3.d.)

1 minute

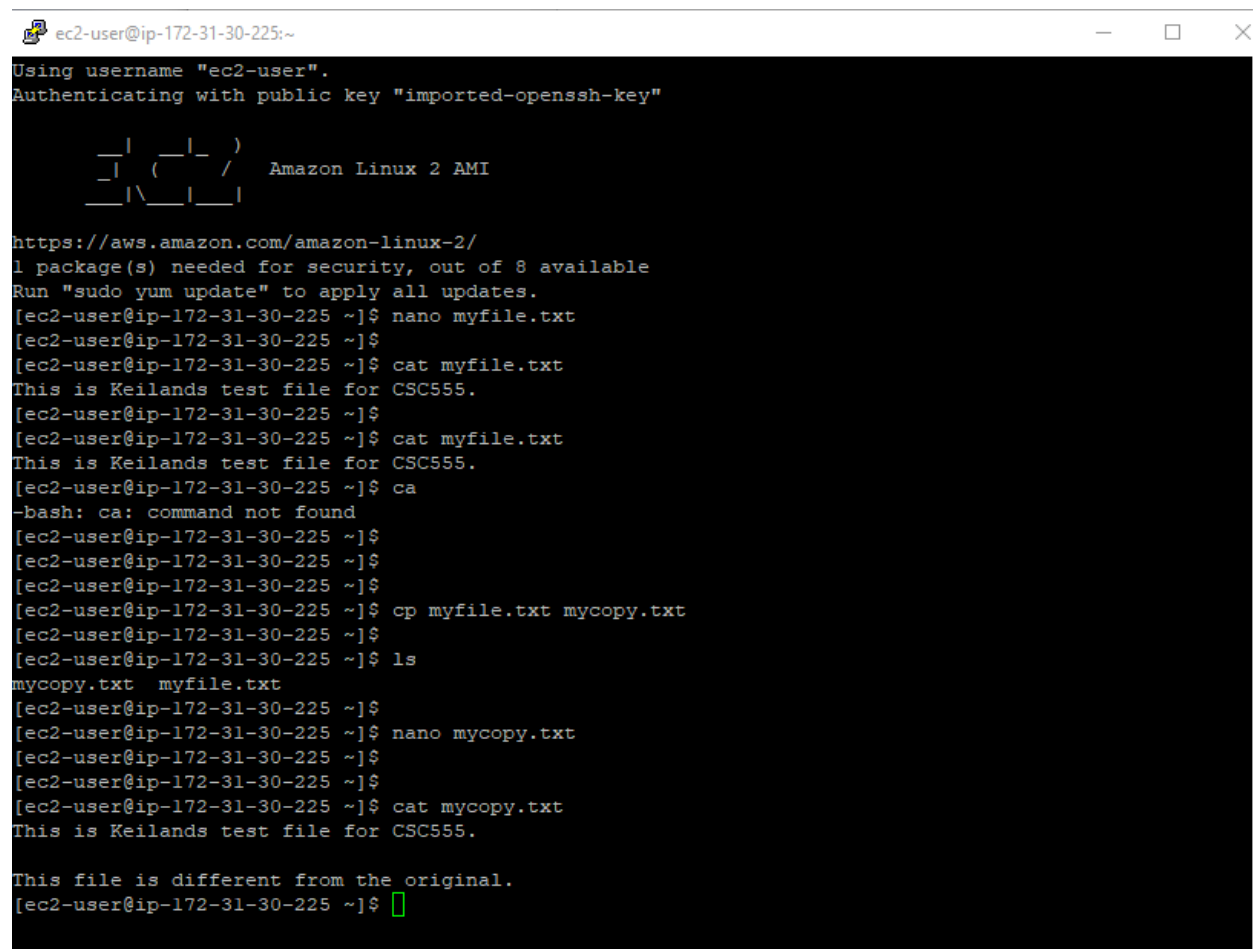
3.e.)

For Hadoop, the default replication factor is 3, so there shouldn't be any change. However, there is a possibility that there may be an improvement in performance.

Part 2.

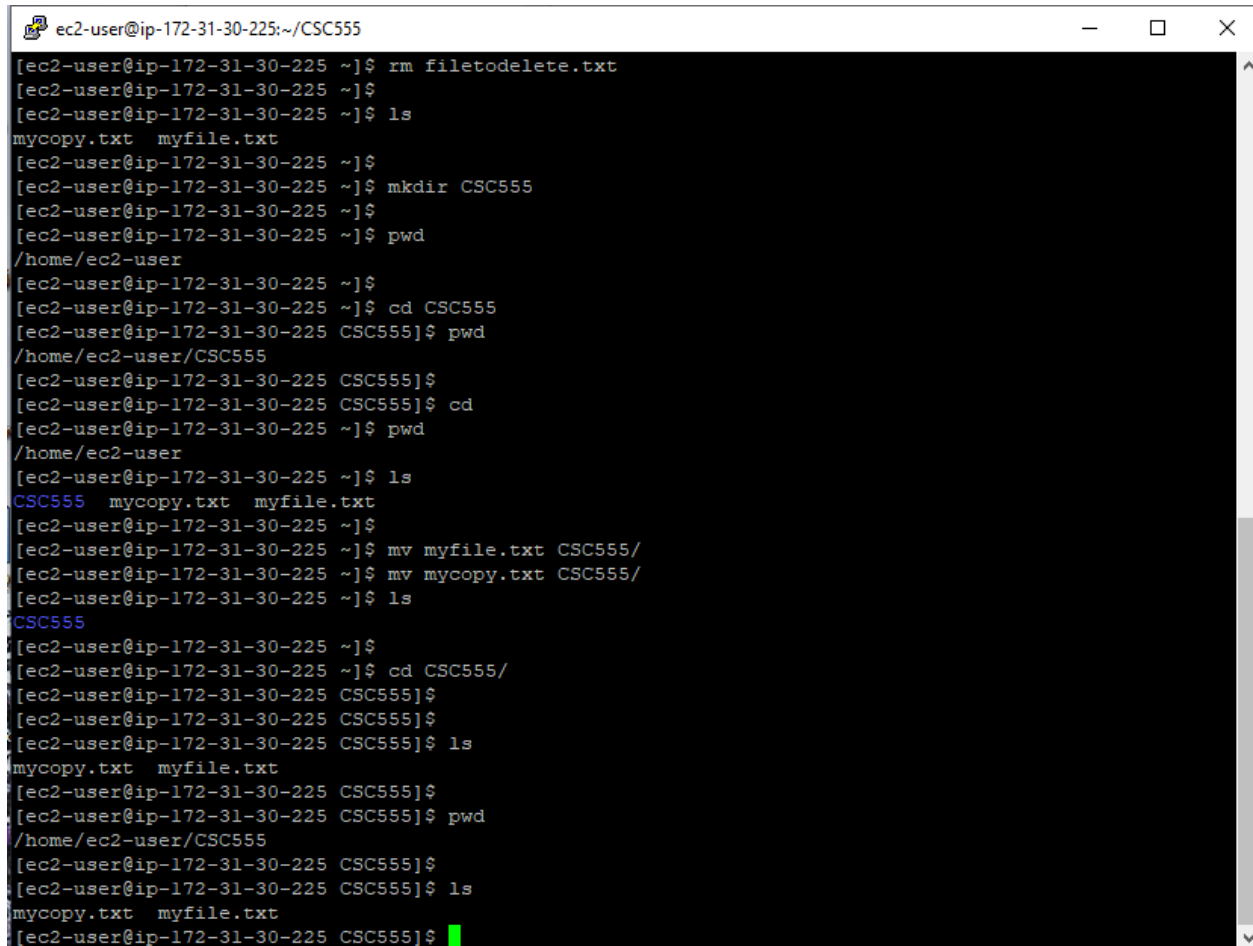
Instance Name: kpullen

Screen shot of contents of home directory:



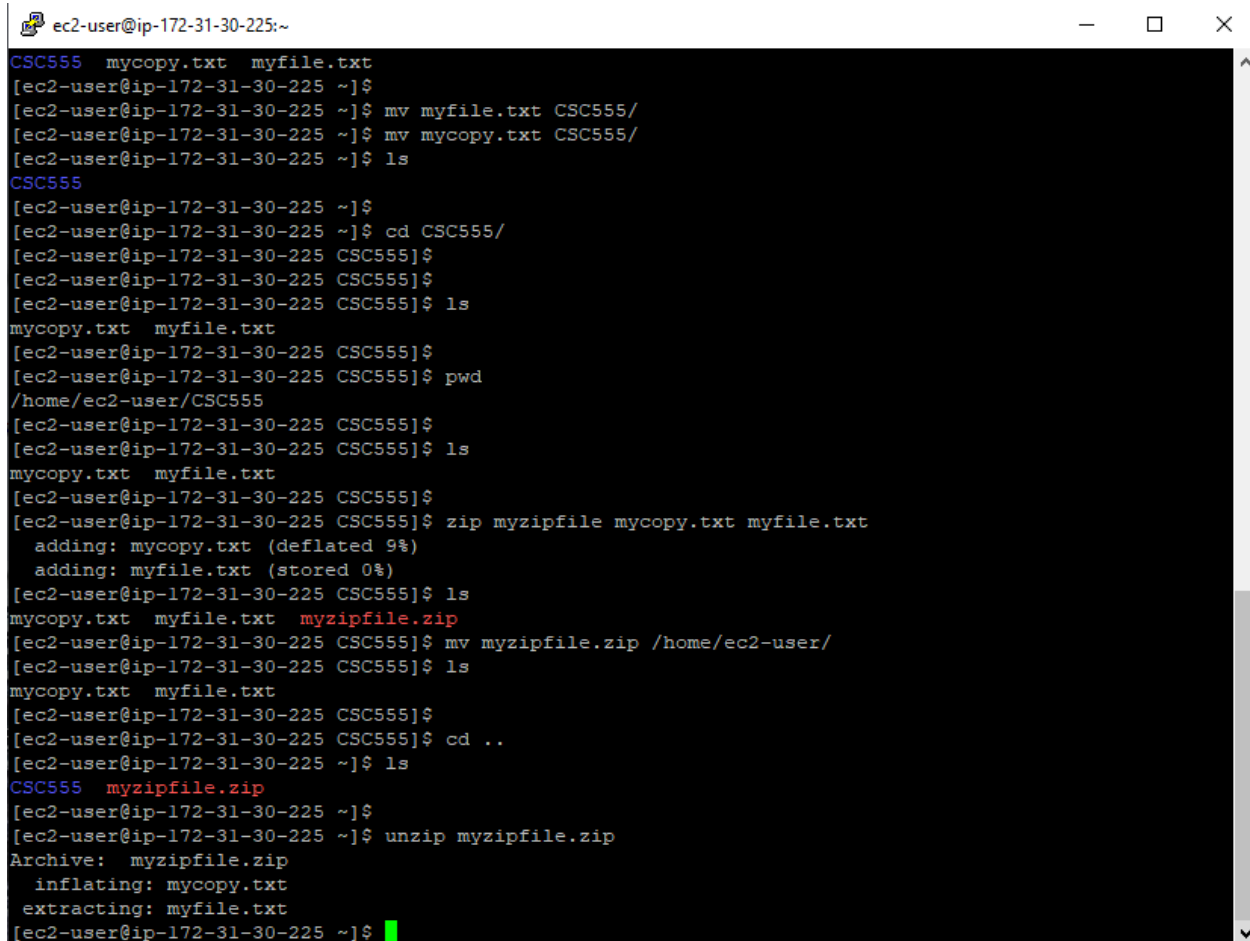
```
ec2-user@ip-172-31-30-225:~  
Using username "ec2-user".  
Authenticating with public key "imported-openssh-key"  
  
  _|_  _|_ )  
  _|_ ( _|_ /  Amazon Linux 2 AMI  
  _|_ \ _|_ |  
  
https://aws.amazon.com/amazon-linux-2/  
1 package(s) needed for security, out of 8 available  
Run "sudo yum update" to apply all updates.  
[ec2-user@ip-172-31-30-225 ~]$ nano myfile.txt  
[ec2-user@ip-172-31-30-225 ~]$  
[ec2-user@ip-172-31-30-225 ~]$ cat myfile.txt  
This is Keilands test file for CSC555.  
[ec2-user@ip-172-31-30-225 ~]$  
[ec2-user@ip-172-31-30-225 ~]$ cat myfile.txt  
This is Keilands test file for CSC555.  
[ec2-user@ip-172-31-30-225 ~]$ ca  
-bash: ca: command not found  
[ec2-user@ip-172-31-30-225 ~]$  
[ec2-user@ip-172-31-30-225 ~]$  
[ec2-user@ip-172-31-30-225 ~]$  
[ec2-user@ip-172-31-30-225 ~]$ cp myfile.txt mycopy.txt  
[ec2-user@ip-172-31-30-225 ~]$  
[ec2-user@ip-172-31-30-225 ~]$ ls  
mycopy.txt  myfile.txt  
[ec2-user@ip-172-31-30-225 ~]$  
[ec2-user@ip-172-31-30-225 ~]$ nano mycopy.txt  
[ec2-user@ip-172-31-30-225 ~]$  
[ec2-user@ip-172-31-30-225 ~]$  
[ec2-user@ip-172-31-30-225 ~]$ cat mycopy.txt  
This is Keilands test file for CSC555.  
  
This file is different from the original.  
[ec2-user@ip-172-31-30-225 ~]$
```

Screen shot of “ls” command in CSC555 directory:



```
ec2-user@ip-172-31-30-225:~/CSC555
[ec2-user@ip-172-31-30-225 ~]$ rm filetodelete.txt
[ec2-user@ip-172-31-30-225 ~]$
[ec2-user@ip-172-31-30-225 ~]$ ls
mycopy.txt  myfile.txt
[ec2-user@ip-172-31-30-225 ~]$
[ec2-user@ip-172-31-30-225 ~]$ mkdir CSC555
[ec2-user@ip-172-31-30-225 ~]$
[ec2-user@ip-172-31-30-225 ~]$ pwd
/home/ec2-user
[ec2-user@ip-172-31-30-225 ~]$
[ec2-user@ip-172-31-30-225 ~]$ cd CSC555
[ec2-user@ip-172-31-30-225 CSC555]$ pwd
/home/ec2-user/CSC555
[ec2-user@ip-172-31-30-225 CSC555]$
[ec2-user@ip-172-31-30-225 CSC555]$ cd
[ec2-user@ip-172-31-30-225 ~]$ pwd
/home/ec2-user
[ec2-user@ip-172-31-30-225 ~]$ ls
CSC555  mycopy.txt  myfile.txt
[ec2-user@ip-172-31-30-225 ~]$
[ec2-user@ip-172-31-30-225 ~]$ mv myfile.txt CSC555/
[ec2-user@ip-172-31-30-225 ~]$ mv mycopy.txt CSC555/
[ec2-user@ip-172-31-30-225 ~]$ ls
CSC555
[ec2-user@ip-172-31-30-225 ~]$
[ec2-user@ip-172-31-30-225 ~]$ cd CSC555/
[ec2-user@ip-172-31-30-225 CSC555]$
[ec2-user@ip-172-31-30-225 CSC555]$
[ec2-user@ip-172-31-30-225 CSC555]$ ls
mycopy.txt  myfile.txt
[ec2-user@ip-172-31-30-225 CSC555]$
[ec2-user@ip-172-31-30-225 CSC555]$ pwd
/home/ec2-user/CSC555
[ec2-user@ip-172-31-30-225 CSC555]$
[ec2-user@ip-172-31-30-225 CSC555]$ ls
mycopy.txt  myfile.txt
[ec2-user@ip-172-31-30-225 CSC555]$
```

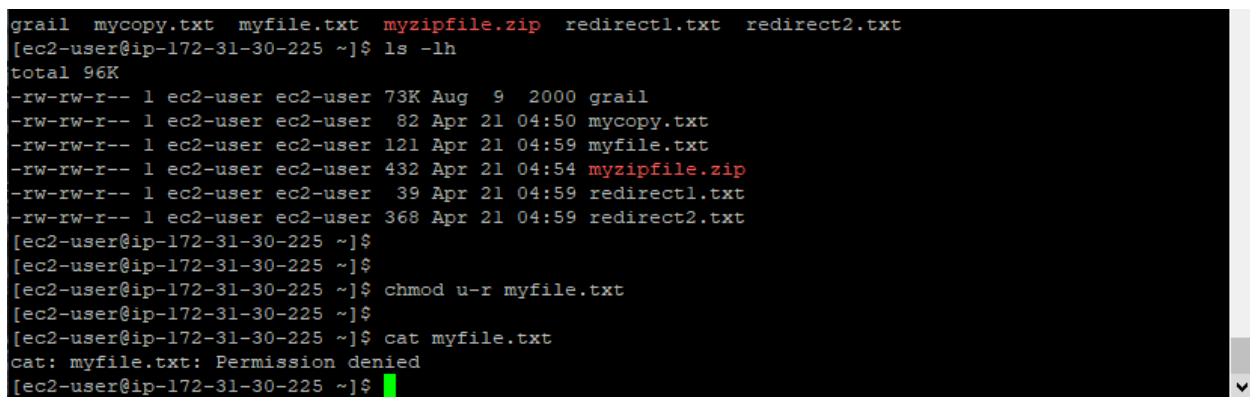
Screen shot of “Unzip” command:



```
ec2-user@ip-172-31-30-225:~  
CSC555 mycopy.txt myfile.txt  
[ec2-user@ip-172-31-30-225 ~]$  
[ec2-user@ip-172-31-30-225 ~]$ mv myfile.txt CSC555/  
[ec2-user@ip-172-31-30-225 ~]$ mv mycopy.txt CSC555/  
[ec2-user@ip-172-31-30-225 ~]$ ls  
CSC555  
[ec2-user@ip-172-31-30-225 ~]$  
[ec2-user@ip-172-31-30-225 ~]$ cd CSC555/  
[ec2-user@ip-172-31-30-225 CSC555]$  
[ec2-user@ip-172-31-30-225 CSC555]$  
[ec2-user@ip-172-31-30-225 CSC555]$ ls  
mycopy.txt myfile.txt  
[ec2-user@ip-172-31-30-225 CSC555]$  
[ec2-user@ip-172-31-30-225 CSC555]$ pwd  
/home/ec2-user/CSC555  
[ec2-user@ip-172-31-30-225 CSC555]$  
[ec2-user@ip-172-31-30-225 CSC555]$ ls  
mycopy.txt myfile.txt  
[ec2-user@ip-172-31-30-225 CSC555]$  
[ec2-user@ip-172-31-30-225 CSC555]$ zip myzipfile mycopy.txt myfile.txt  
  adding: mycopy.txt (deflated 9%)  
  adding: myfile.txt (stored 0%)  
[ec2-user@ip-172-31-30-225 CSC555]$ ls  
mycopy.txt myfile.txt myzipfile.zip  
[ec2-user@ip-172-31-30-225 CSC555]$ mv myzipfile.zip /home/ec2-user/  
[ec2-user@ip-172-31-30-225 CSC555]$ ls  
mycopy.txt myfile.txt  
[ec2-user@ip-172-31-30-225 CSC555]$  
[ec2-user@ip-172-31-30-225 CSC555]$ cd ..  
[ec2-user@ip-172-31-30-225 ~]$ ls  
CSC555 myzipfile.zip  
[ec2-user@ip-172-31-30-225 ~]$  
[ec2-user@ip-172-31-30-225 ~]$ unzip myzipfile.zip  
Archive: myzipfile.zip  
  inflating: mycopy.txt  
  extracting: myfile.txt  
[ec2-user@ip-172-31-30-225 ~]$
```

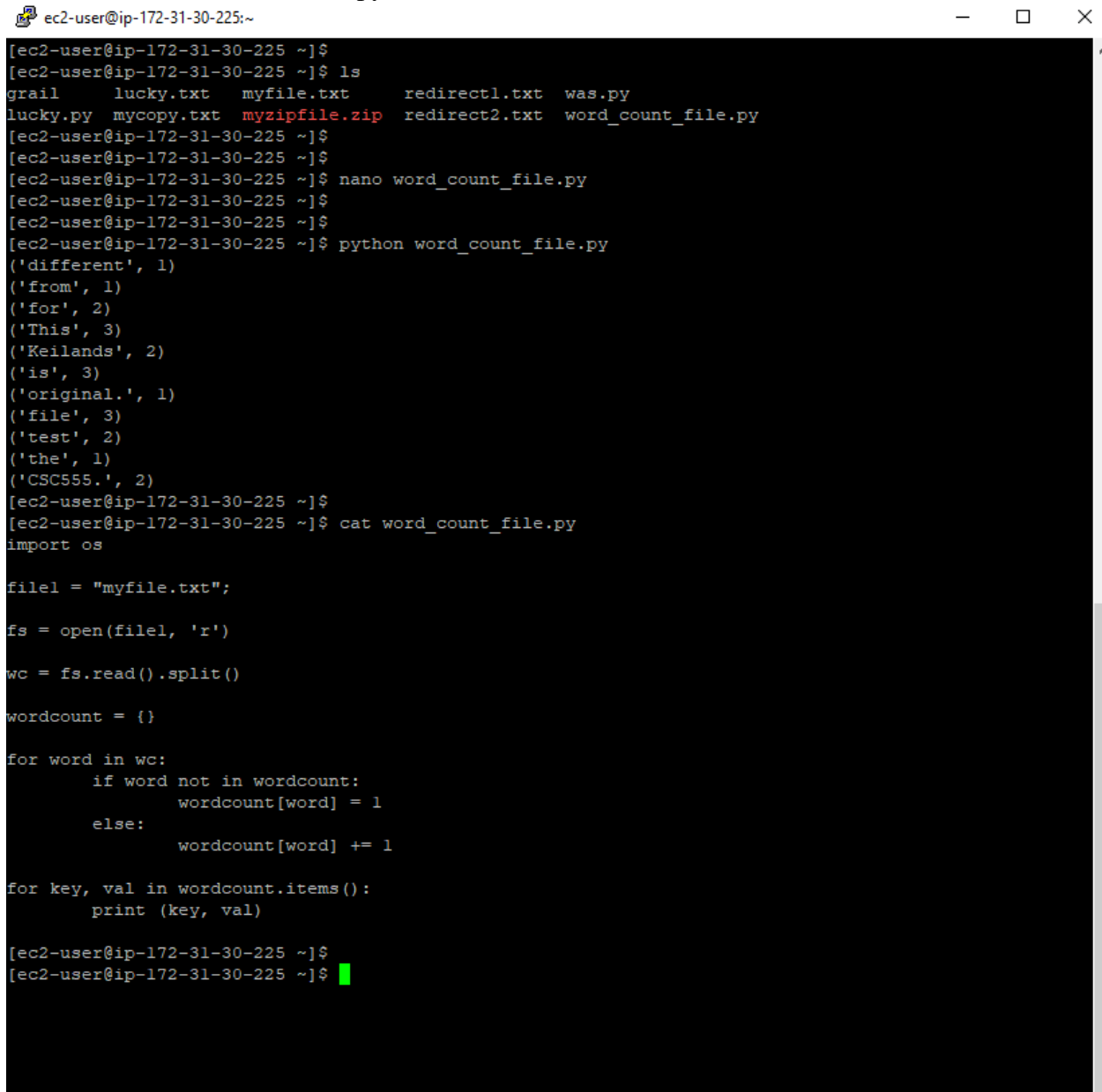
File size for grail is 73K

Screen shot of “Permission Denied” message:



```
grail mycopy.txt myfile.txt myzipfile.zip redirect1.txt redirect2.txt  
[ec2-user@ip-172-31-30-225 ~]$ ls -lh  
total 96K  
-rw-rw-r-- 1 ec2-user ec2-user 73K Aug  9  2000 grail  
-rw-rw-r-- 1 ec2-user ec2-user  82 Apr 21 04:50 mycopy.txt  
-rw-rw-r-- 1 ec2-user ec2-user 121 Apr 21 04:59 myfile.txt  
-rw-rw-r-- 1 ec2-user ec2-user 432 Apr 21 04:54 myzipfile.zip  
-rw-rw-r-- 1 ec2-user ec2-user  39 Apr 21 04:59 redirect1.txt  
-rw-rw-r-- 1 ec2-user ec2-user 368 Apr 21 04:59 redirect2.txt  
[ec2-user@ip-172-31-30-225 ~]$  
[ec2-user@ip-172-31-30-225 ~]$  
[ec2-user@ip-172-31-30-225 ~]$ chmod u-r myfile.txt  
[ec2-user@ip-172-31-30-225 ~]$  
[ec2-user@ip-172-31-30-225 ~]$ cat myfile.txt  
cat: myfile.txt: Permission denied  
[ec2-user@ip-172-31-30-225 ~]$
```

Screen shot of word count and python code:



```
ec2-user@ip-172-31-30-225:~  
[ec2-user@ip-172-31-30-225 ~]$  
[ec2-user@ip-172-31-30-225 ~]$ ls  
grail      lucky.txt  myfile.txt  redirect1.txt  was.py  
lucky.py  mycopy.txt  myzipfile.zip  redirect2.txt  word_count_file.py  
[ec2-user@ip-172-31-30-225 ~]$  
[ec2-user@ip-172-31-30-225 ~]$  
[ec2-user@ip-172-31-30-225 ~]$ nano word_count_file.py  
[ec2-user@ip-172-31-30-225 ~]$  
[ec2-user@ip-172-31-30-225 ~]$  
[ec2-user@ip-172-31-30-225 ~]$ python word_count_file.py  
( 'different', 1)  
( 'from', 1)  
( 'for', 2)  
( 'This', 3)  
( 'Keilands', 2)  
( 'is', 3)  
( 'original.', 1)  
( 'file', 3)  
( 'test', 2)  
( 'the', 1)  
( 'CSC555.', 2)  
[ec2-user@ip-172-31-30-225 ~]$  
[ec2-user@ip-172-31-30-225 ~]$ cat word_count_file.py  
import os  
  
file1 = "myfile.txt";  
  
fs = open(file1, 'r')  
  
wc = fs.read().split()  
  
wordcount = {}  
  
for word in wc:  
    if word not in wordcount:  
        wordcount[word] = 1  
    else:  
        wordcount[word] += 1  
  
for key, val in wordcount.items():  
    print (key, val)  
  
[ec2-user@ip-172-31-30-225 ~]$  
[ec2-user@ip-172-31-30-225 ~]$
```

```
import os  
file1 = "myfile.txt";  
fs = open(file1, 'r')  
wc = fs.read().split()  
wordcount = { }  
for word in wc:  
    if word not in wordcount:  
        wordcount[word] = 1  
    else:  
        wordcount[word] += 1  
  
for key, val in wordcount.items():  
    print (key, val)
```

Part 3.)

Screen shot of hadoop /data directory:

```
2022-04-21 14:01:24 (45.6 MB/s) - 'bioproject.xml' saved [231149003/231149003]

[ec2-user@ip-172-31-30-225 ~]$ hadoop fs -put bioproject.xml /data/
[ec2-user@ip-172-31-30-225 ~]$ hadoop fs -ls /data
Found 1 items
-rw-r--r-- 1 ec2-user supergroup 231149003 2022-04-21 14:01 /data/bioproject.xml
[ec2-user@ip-172-31-30-225 ~]$
```

Screen shot of time to execute wordcount:

```
real    1m28.176s
user    0m4.191s
sys     0m0.210s
[ec2-user@ip-172-31-30-225 ~]$
```

Part 4.)

Screen shot of “Select Count(*) FROM VehicleData, appears to have returned 1 row:

```
hive> SELECT COUNT(*) FROM VehicleData;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider us
ing a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = ec2-user_20220421142445_78fe3c8c-db02-4800-a42c-2c341371460c
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1650549619130_0002, Tracking URL = http://ip-172-31-30-225.us-east-2.compute.internal
:8088/proxy/application_1650549619130_0002/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_1650549619130_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-04-21 14:25:00,552 Stage-1 map = 0%, reduce = 0%
2022-04-21 14:25:11,017 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.28 sec
2022-04-21 14:25:17,823 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 2.53 sec
MapReduce Total cumulative CPU time: 2 seconds 530 msec
Ended Job = job_1650549619130_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 2.53 sec HDFS Read: 11775010 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 530 msec
OK
34175
Time taken: 34.922 seconds, Fetched: 1 row(s)
hive>
```

Screen shot of query: SELECT MIN(barrels08), AVG(barrels08), MAX(barrels08) FROM VehicleData:

```
hive>
>
> SELECT MIN(barrels08), AVG(barrels08), MAX(barrels08) FROM VehicleData:
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider us
ing a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = ec2-user_20220421142719_054f4eae-c062-4794-9910-77eac9fc8a74
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1650549619130_0003, Tracking URL = http://ip-172-31-30-225.us-east-2.compute.internal
:8088/proxy/application_1650549619130_0003/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job -kill job_1650549619130_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-04-21 14:27:29,375 Stage-1 map = 0%, reduce = 0%
2022-04-21 14:27:39,211 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.89 sec
2022-04-21 14:27:46,806 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 3.1 sec
MapReduce Total cumulative CPU time: 3 seconds 100 msec
Ended Job = job_1650549619130_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 3.1 sec HDFS Read: 11777415 HDFS Write: 37 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 100 msec
OK
0.059892      17.820177449476272      47.06831
Time taken: 28.215 seconds, Fetched: 1 row(s)
hive>
```

Screen shot from Query: SELECT(barrels08/ city08) FROM VehicleData;

```
Time taken: 0.197 seconds, Fetched: 34175 row(s)
hive>
```

Screen shot of ThreeColExtract directory and file:

```
[ec2-user@ip-172-31-30-225 ~]$ hadoop fs -ls ThreeColExtract
Found 1 items
-rwxr-xr-x  1 ec2-user supergroup      627873 2022-04-21 14:39 ThreeColExtract/000000_0
[ec2-user@ip-172-31-30-225 ~]$
```

Screen shot of TwoColExtract directory and file:

```
hive> exit;
[ec2-user@ip-172-31-30-225 apache-hive-2.0.1-bin]$ hadoop fs -ls TwoColExtract
Found 1 items
-rwxr-xr-x  1 ec2-user supergroup      287749 2022-04-21 15:01 TwoColExtract/000000_0
[ec2-user@ip-172-31-30-225 apache-hive-2.0.1-bin]$
```