Robert Kaszubski

CSC555 – Assignment 2

**Part 1**

1)

a)

The mapper is going to produce First as the Key and Grade as the Value. Then the reducer will output First as the Key, and Sum(Grade) as the Value for each key which will combine grades for each key.

b)

The mapper is going to produce City\_State as the Key and Name as the Value. Then the reducer will first have to remove duplicate Key, Value pairs because we are trying to find the number of distinct names. Each name can then be set to 1 and the reducer can output City\_State as the Key, and Sum(1) as the Value. The reducer is counting the number of records for each key.

2)

a)

One way to speed up the query you could increase the number of nodes processing the data. Given that it’s a large amount of data then this should improve the performance and run the query much faster. You could also alter the number of map reduce tasks. There are likewise 4 columns that aren’t used and hinder performance.

b)

The NameNode handles failure so it would reattempt to process a block of data on a different node. The map reduce framework would prefer to start from the name node.

c)

i) There would be 3 output files. Each reducer would generate one.

ii) You could use MOD 3 to assign keys to reducers. You would first have to get the hash value of the key. Then when you find the mod 3 value of that hash value, it’ll return either 0, 1, 2. All the 0s go to one reducer, all the 1s to another, and all the 2s to the third reducer. We do MOD 3 because there are 3 reducers.

3)

a) 65 minutes

b) 7 minutes

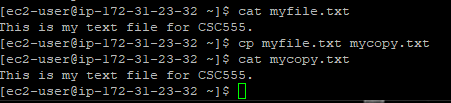
c) 3 minutes

d) 1 minute

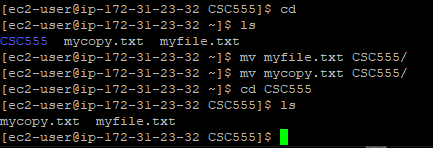
e) None of the answers must change. Replication refers to how many times a block is stored. We are running a mapreduce job on one copy of each block. The results wouldn’t change if we run the same job on 3 copies of each block so we only have to run it on one. Therefore none of our results would change based on replication factor. If we are talking about storing the results of the job, then it would take longer as we would have to store 3 copies instead of 1.

**Part 2)**

Take a screen shot of the contents of your copied file displayed on the terminal screen:



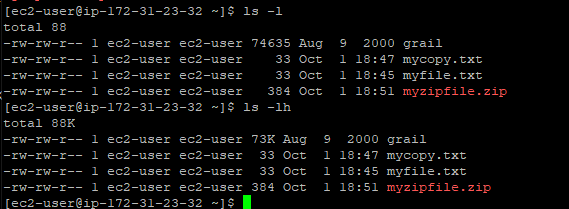
Take a screenshot of the files listed in the CSC555 directory:



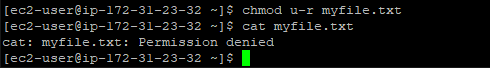
Take a screenshot of the screen after this command (zip):



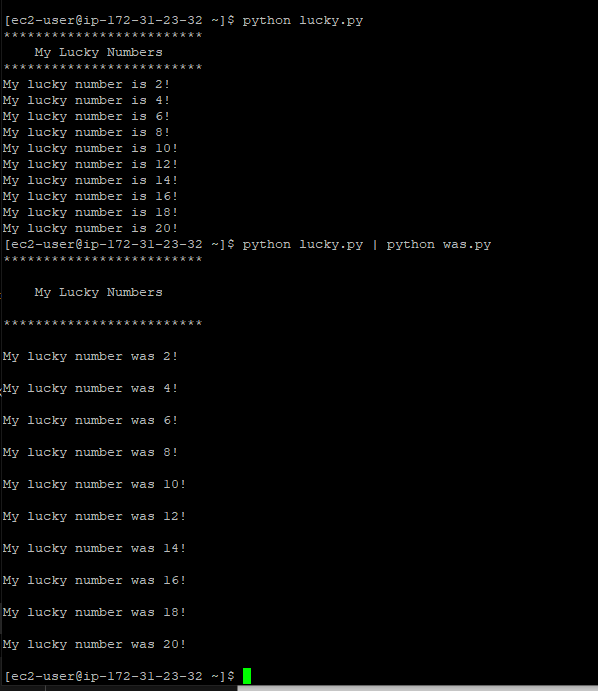
The Size of the grail file: about 73K (kilobytes) or 74635 bytes



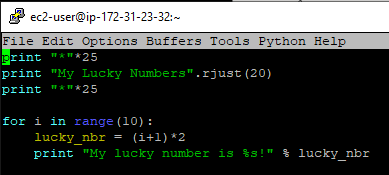
Permission denied error:

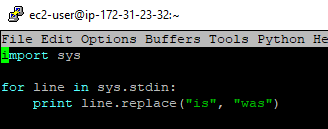


Python output:



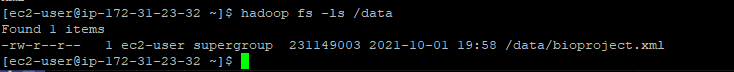
Python Code:





**Part 3)**

Verify file upload to HDFS:



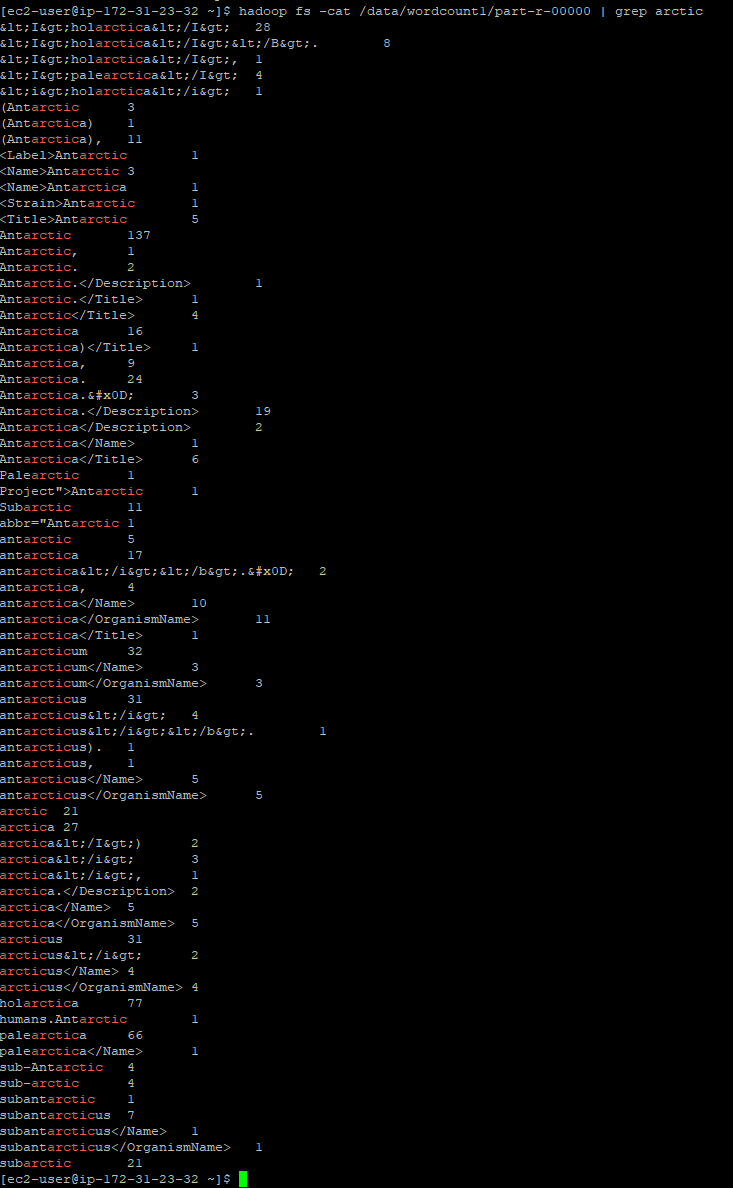
Time to run:



Size:

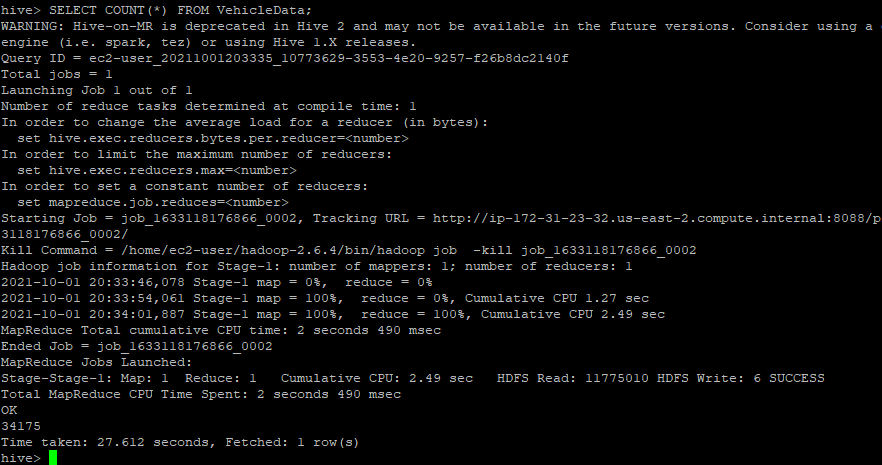


Output (arctic)



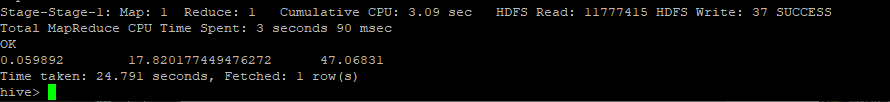
**Part 4)**

Select Count(\*) FROM VehicleData:



34175 rows

SELECT MIN(barrels08), AVG(barrels08), MAX(barrels08) FROM VehicleData;

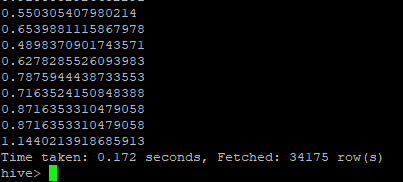


Min was .059892

Avg was 17.8201

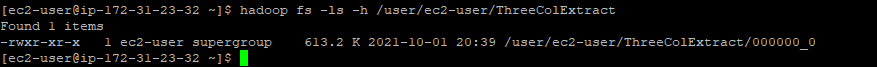
Max was 47.06831

SELECT (barrels08/city08) FROM VehicleData;



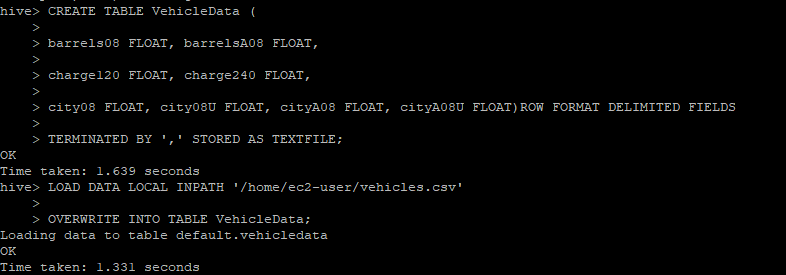
Time taken was .172 seconds

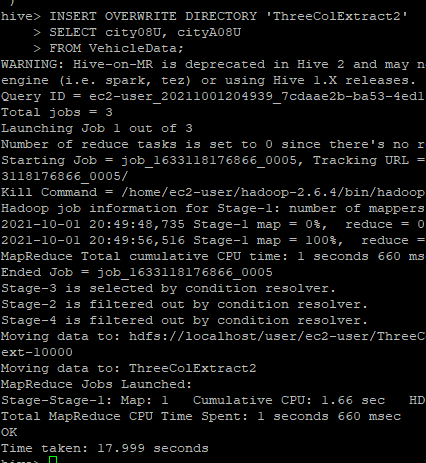
ThreeColExtract file size:



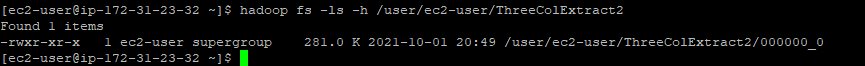
613.2K which is the same as 627873 bytes

New Table:





New file size:



281K