**CSC 555 Assignment 2**

**Yiyang Yang**

**Part 1**

a).

Mapper: Year will be used as key, value is the count of the Full names

Reducer: Group all the full names into corresponding years.

b).

Mapper: EID and AID as keys, values are EFirst, ELast, and AFirst, ALast.

Reducer: Show the EFirst and Elast, when EFirst same as AFirst, ELast same as ALast.

c).

Mapper: EID and AID are used to be keys, average of the age is the value

Reducer: Show all the EFirst and average of age, for each EID same as AID, and then group the average of age into the corresponding EFirst.

**Part 2**

a).

1. I think using “Combiners” will be good and reduce the runtime make it faster.

2. I think changing the number of mapreduce tasks, which change the size of blocks of the large files, it will run less MapReduce tasks at same time, and reduce the time of the mission.

b).

I think should restart the job where replica for data maintained.

c).

i). 4 files output generate.

ii). Hash key = 4 MOD 3.

d).

i). True

ii). True

**Part 3**

a).

60 \* 2 = 120 mins

b).

All blocks will be divided into 20 nodes for 3 rounds, the run time of 3 rounds, 3 \* 2 = 6 mins.

c).

All blocks will be divided into 50 nodes for first round, and 15 blocks left for second rounds, the run time of 2 rounds, 2 \* 2 = 4 mins.

d).

All blocks will be divided into 75 nodes for 1 round, the run time of 1 round, 1 \* 2 = 2 mins.

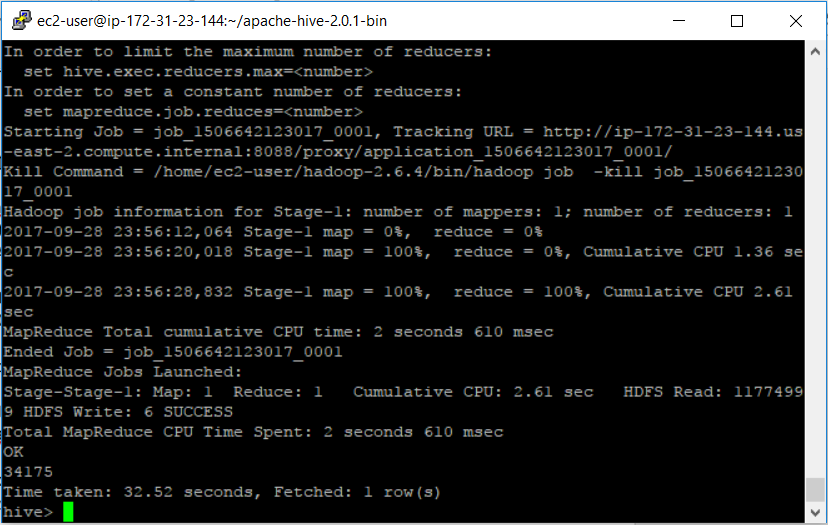
e).

I think run time will not change when replication factor changed to 3.

**Part 4**

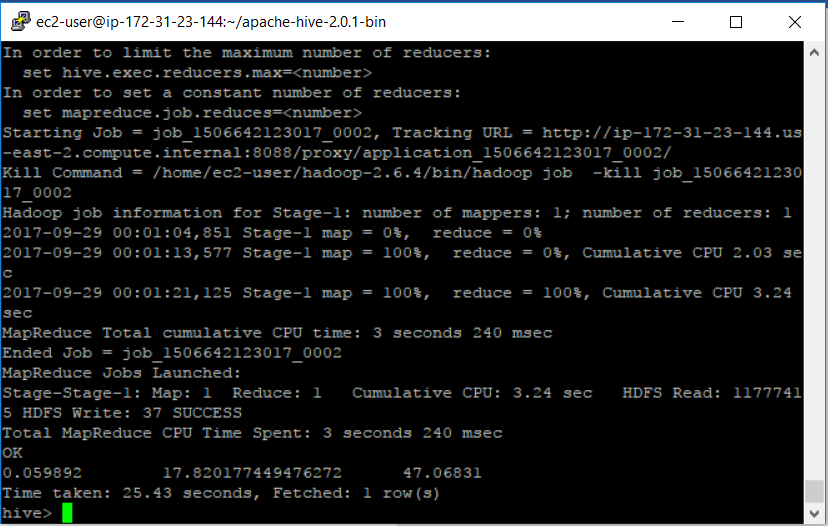
a).

*SELECT COUNT(\*) FROM VehicleData;*

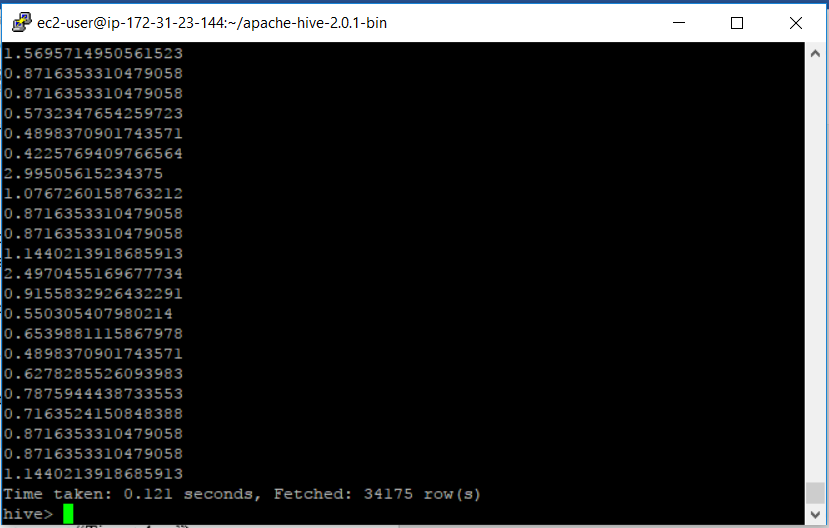


There are 34175 rows.

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

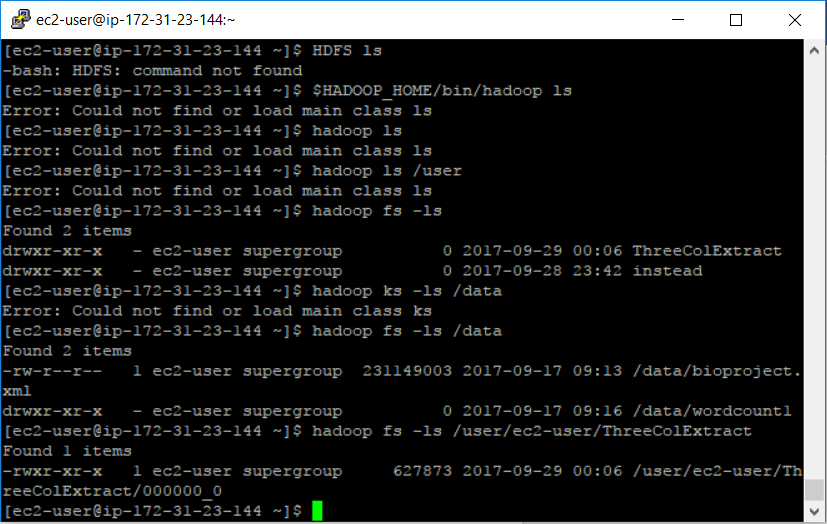


*SELECT (barrels08/city08) FROM VehicleData;*



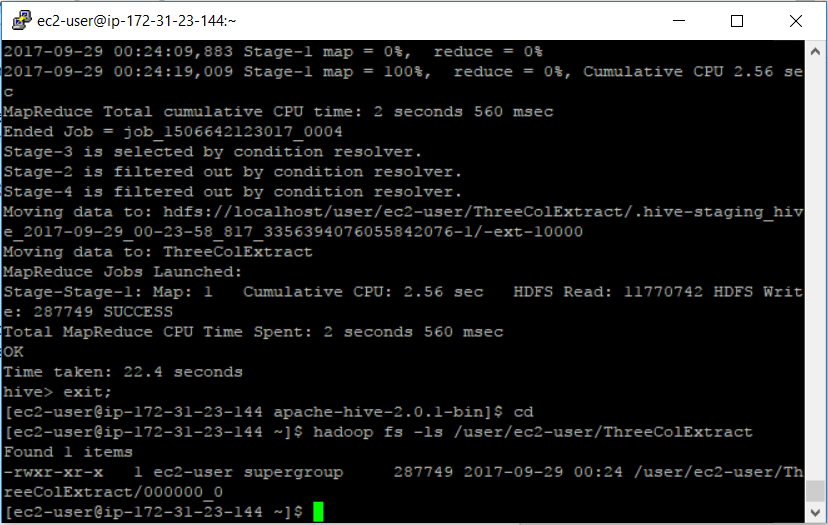
Time taken 0.121 seconds.

Size of the newly created file.



The size of the file is 627873 bytes.

Size of new created file.



The size of the file is 287749 bytes.