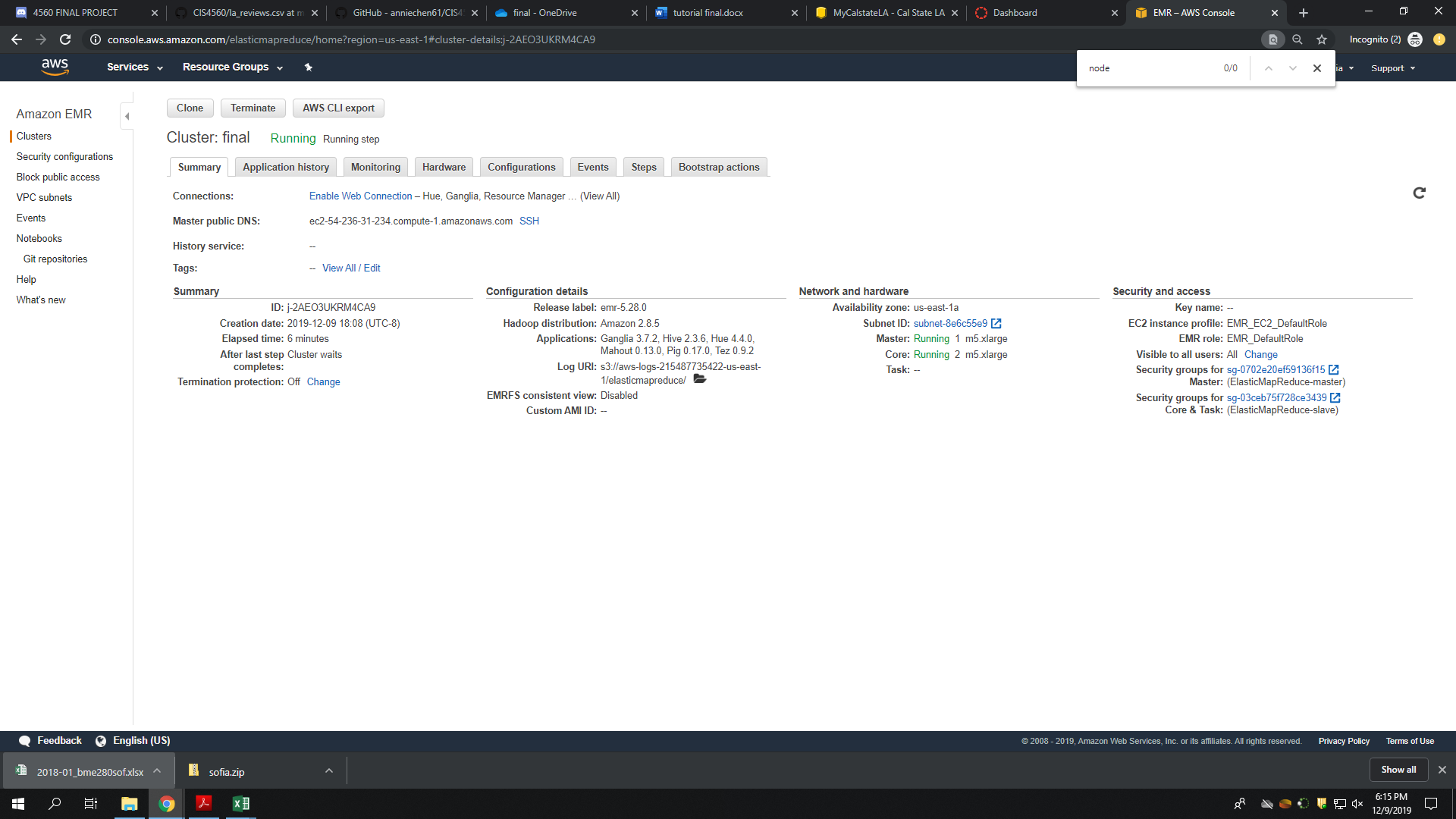
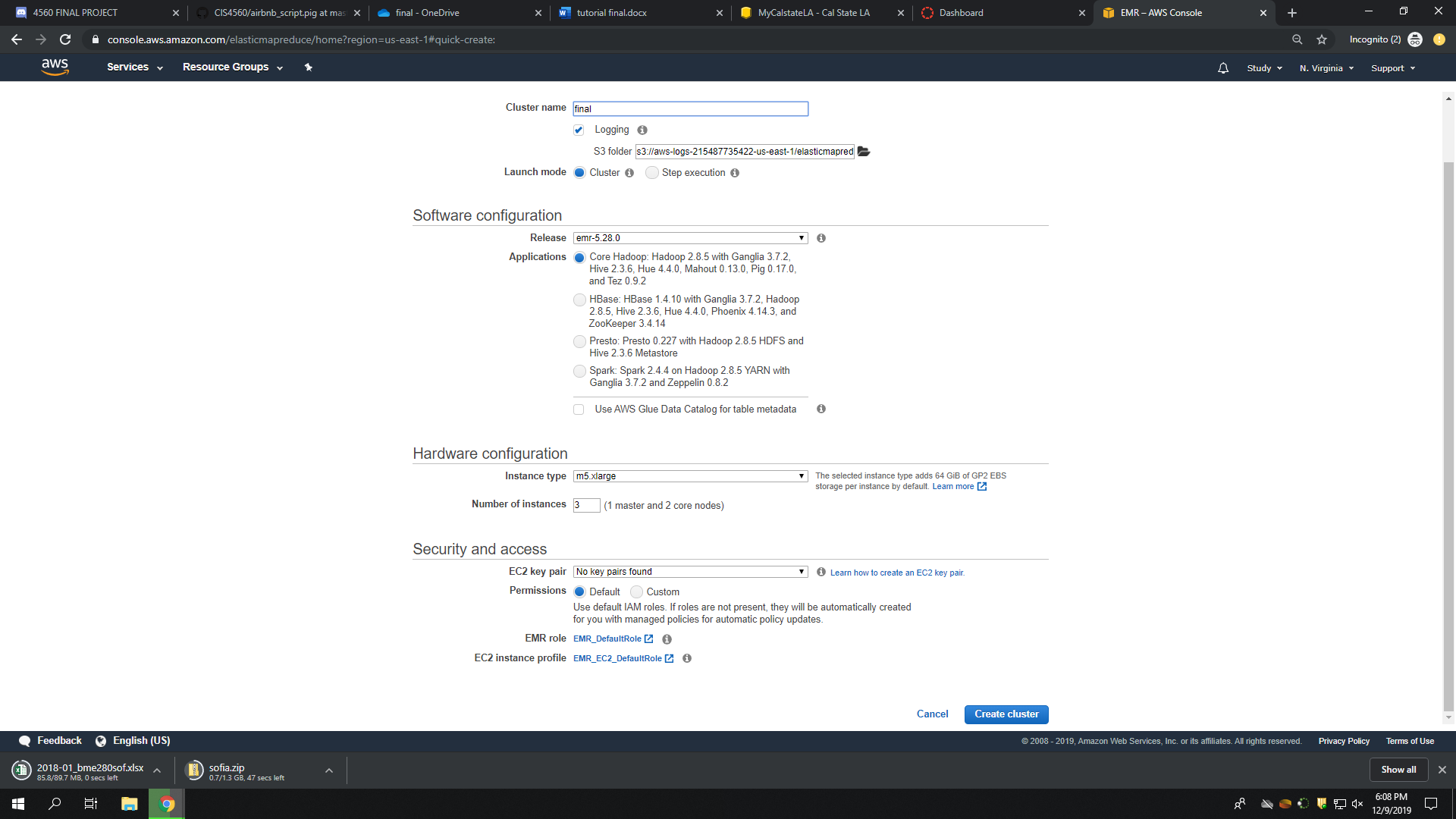
**GROUP 1 TUTORIAL**

You can use AWS EMR to create your cluster.



Connect to EMR cluster by using command

Ssh yourusername@cluserIPAddress

Using wget command to retrieve csv file

Successful result looks similar to this

2018-12\_bme280sof.csv [ <=> ] 63.40K --.-KB/s in 0.02s

2019-12-10 02:58:00 (3.83 MB/s) - ‘2018-12\_bme280sof.csv’ saved [64926]

**Step 1: create directories in haddop for sofia dataset**

1. -bash-4.2$ hdfs dfs –mkdir sofia
2. -bash-4.2$ hdfs dfs -mkdir sofia/2017
3. -bash-4.2$ hdfs dfs -mkdir sofia/2018
4. -bash-4.2$ hdfs dfs -mkdir sofia/2019

It should shows result like this

drwxr-xr-x - thua hadoop 0 2019-12-10 02:41 sofia/2017

drwxr-xr-x - thua hadoop 0 2019-12-10 02:41 sofia/2018

drwxr-xr-x - thua hadoop 0 2019-12-10 02:41 sofia/2019

**Step 2: Put files in the relevant directories**

1. Put 2017 data to sofia/2017. Please do the same with 2018 and 2019 data

Hdfs dfs –put 2017\*\*\*.csv sofia/2017

1. Double check if all files are correctly put in relevant folder

Hdfs dfs –ls sofia/201\*

Result should be something like this

-rw-r--r-- 1 thua hadoop 64926 2019-12-10 03:00 sofia/2018/2018-12\_bme280sof.csv

S**tep 3: Using Pig command**

**Open another terminal and use pig command**

1. **Load dataset (Remember to change your user name**)

grunt> sofia = LOAD 'user/yourusername/sofia/2017/201\*\*\*.csv' USING PigStorage(',') AS ( sensor\_id:int, location:int, lat: double, lon:double, timestamp:datetime, pressure:double, temperature:double, humidity:double);

grunt>sofia = LOAD 'user/yourusername/sofia/2018/201\*\*\*.csv' USING PigStorage(',') AS ( sensor\_id:int, location:int, lat: double, lon:double, timestamp:datetime, pressure:double, temperature:double, humidity:double);

grunt>sofia = LOAD 'user/yourusername/sofia/2019/201\*\*\*.csv' USING PigStorage(',') AS ( sensor\_id:int, location:int, lat: double, lon:double, timestamp:datetime, pressure:double, temperature:double, humidity:double);

1. Describe sofia

grunt> describe sofia;

It should be like this

sofia: {sensor\_id: int,location: int,lat: double,lon: double,timestamp: datetime,pressure: double,temperature: double,humidity: double}

1. Sort data to see the location the 10 highest degree in 2017, 2018, and 2019

hotplace = ORDER location by temperature DESC;

top10hot = limit hotplace 10;

dump top10hot;

1. Sort data to see 10 coldest place in 2017, 2018, 2019

coldplace = ORDER location BY temperature ASC;

top10cold = limit coldplace 10;

dump top10cold;

1. Sort data for 10 most humid place

humidplace= ORDER location by humidity DESC;

top10humid = limit humidplace 10;

dump top10humid;

1. Save all processed file into PigStorage

Step 4. Generalizing 3D map in Excel

Download csv files, then convert to .xlsx file since only .xlsx work with 3D map.

Note: Remember to format the table in excel file before generalizing 3D map

