Contents

[1 Spin up a new EMR cluster 1](#_Toc35121638)

[2 Connect to master node of cluster 7](#_Toc35121639)

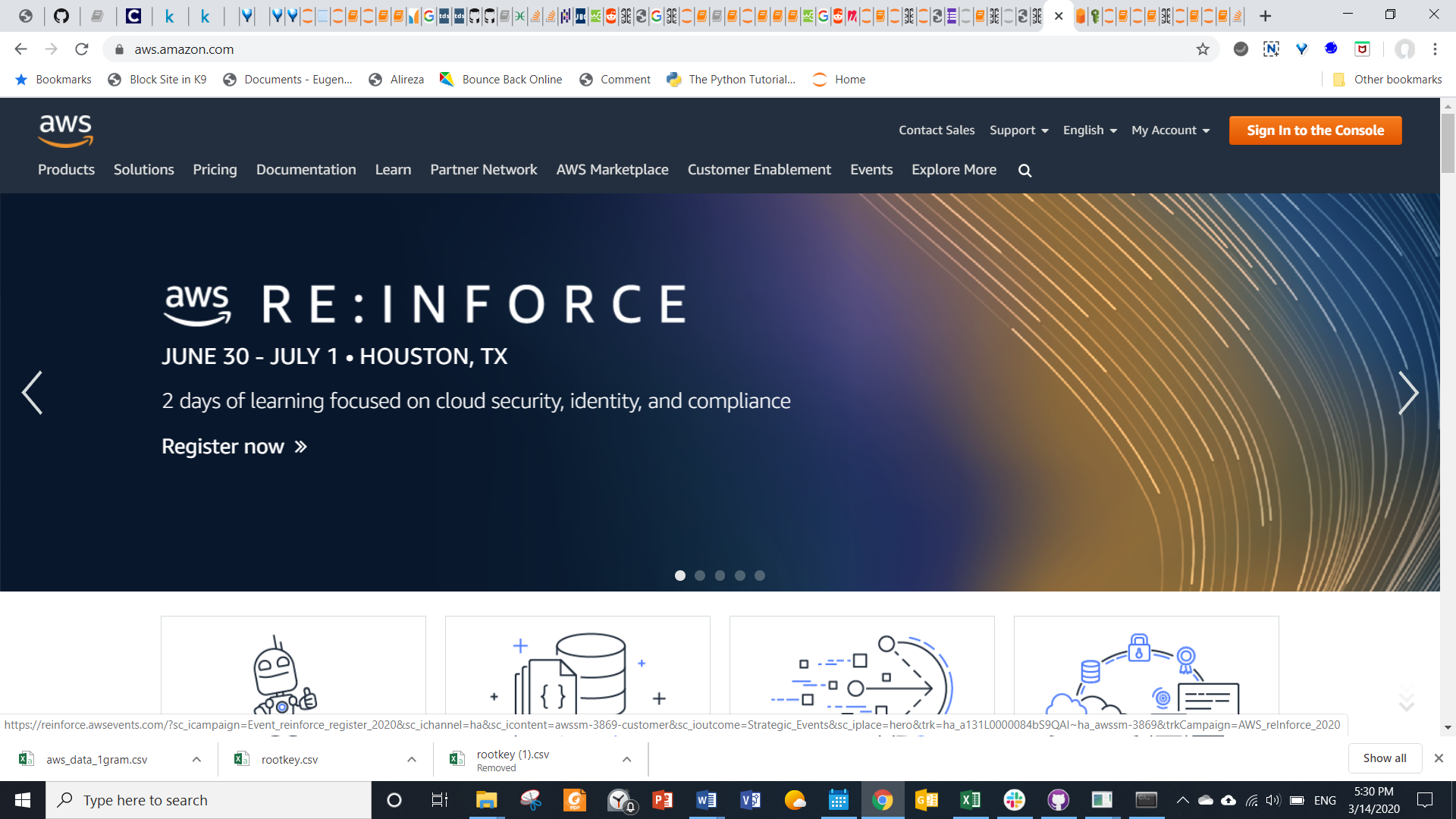
[3 Copy data from S3 bucket to Hadoop directory 8](#_Toc35121640)

[4 Analyze the data with spark 9](#_Toc35121641)

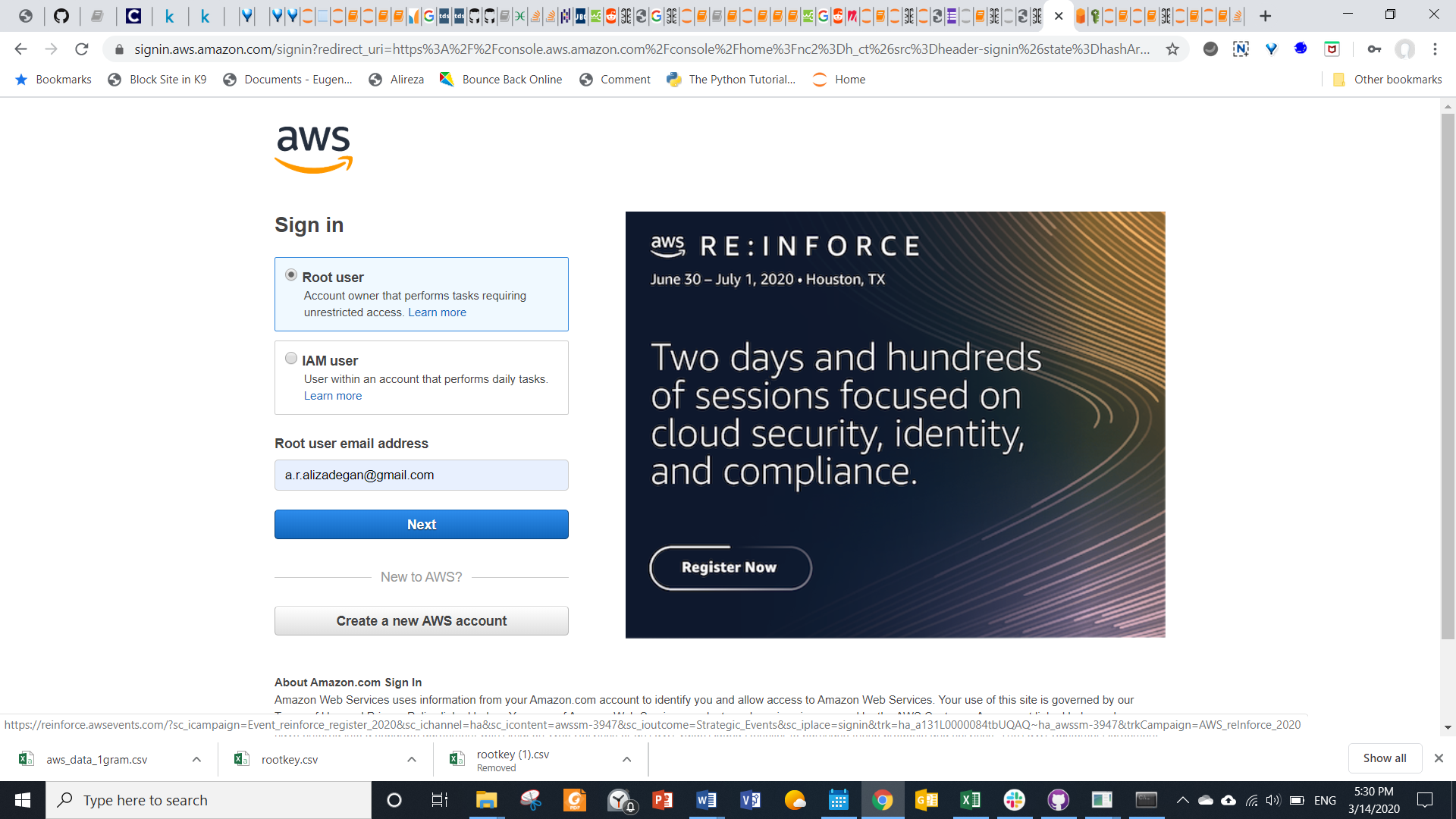
[5 Merge data into master node and copy to S3 bucket 12](#_Toc35121642)

# pin up a new EMR cluster

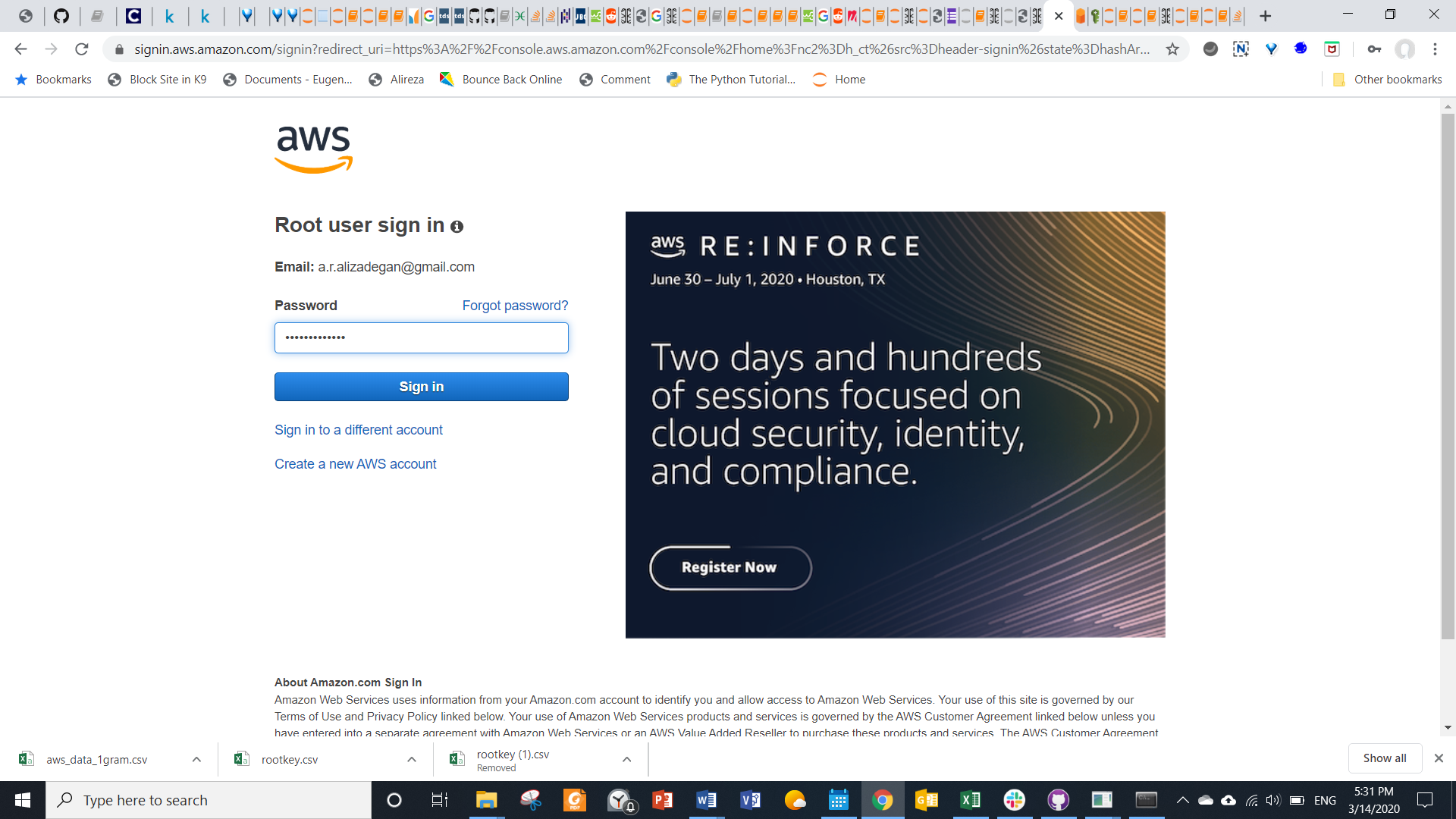
Go to aws.amazon.com and sign in to the console



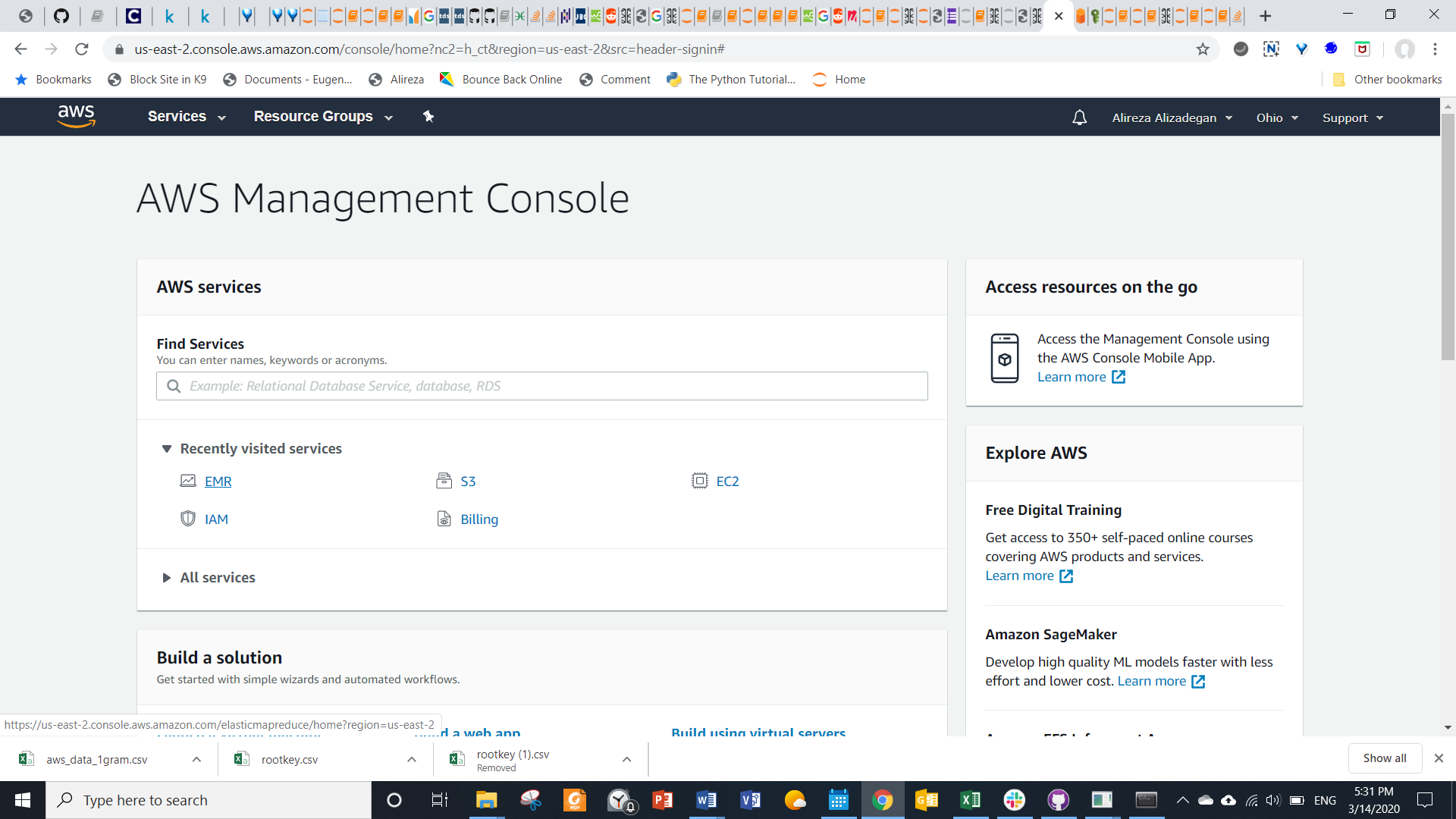
Enter root user email address and click next



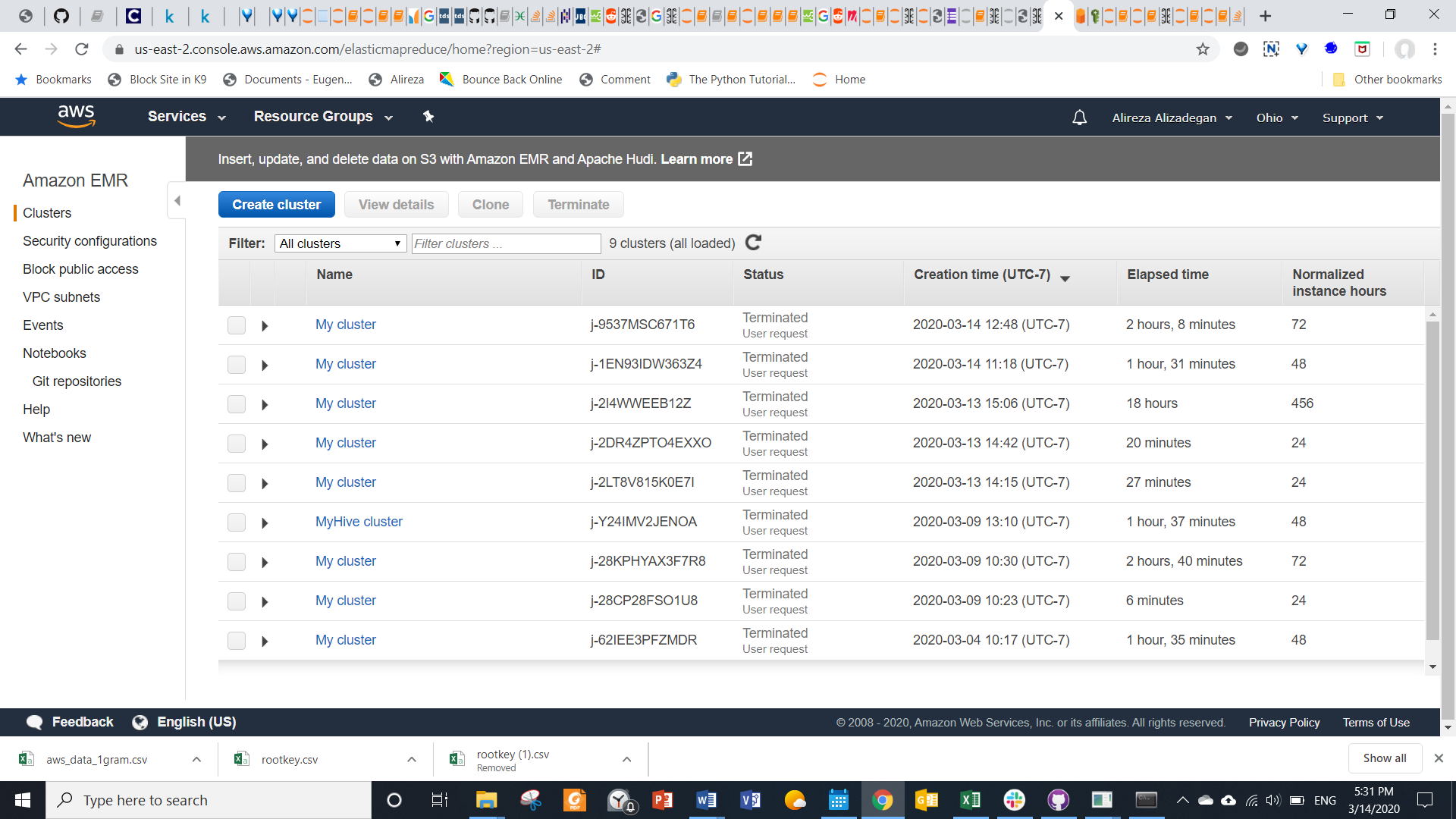
Enter password and sign in



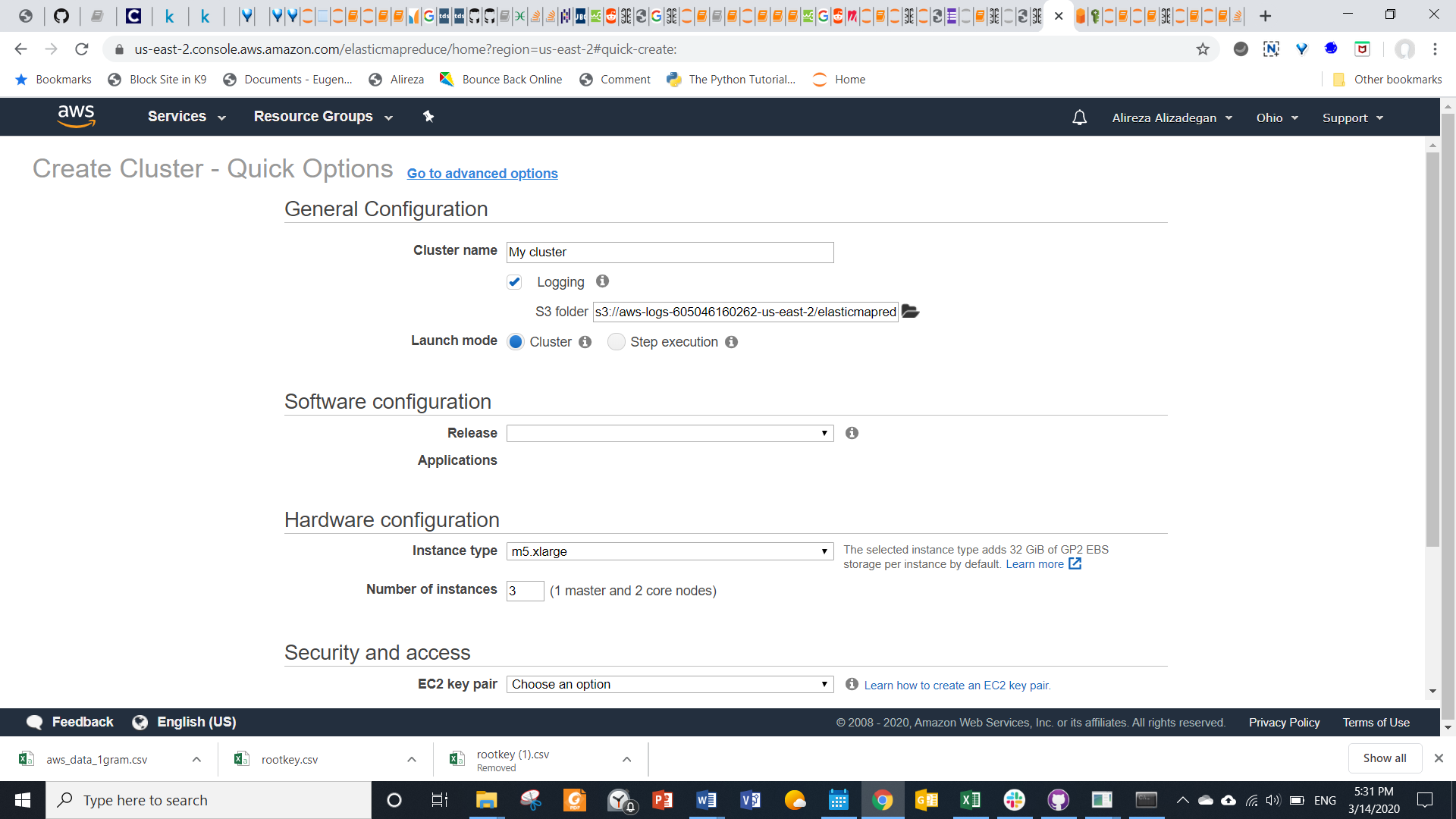
Choose EMR from AWS services in their management console



Click create cluster in EMR clusters page



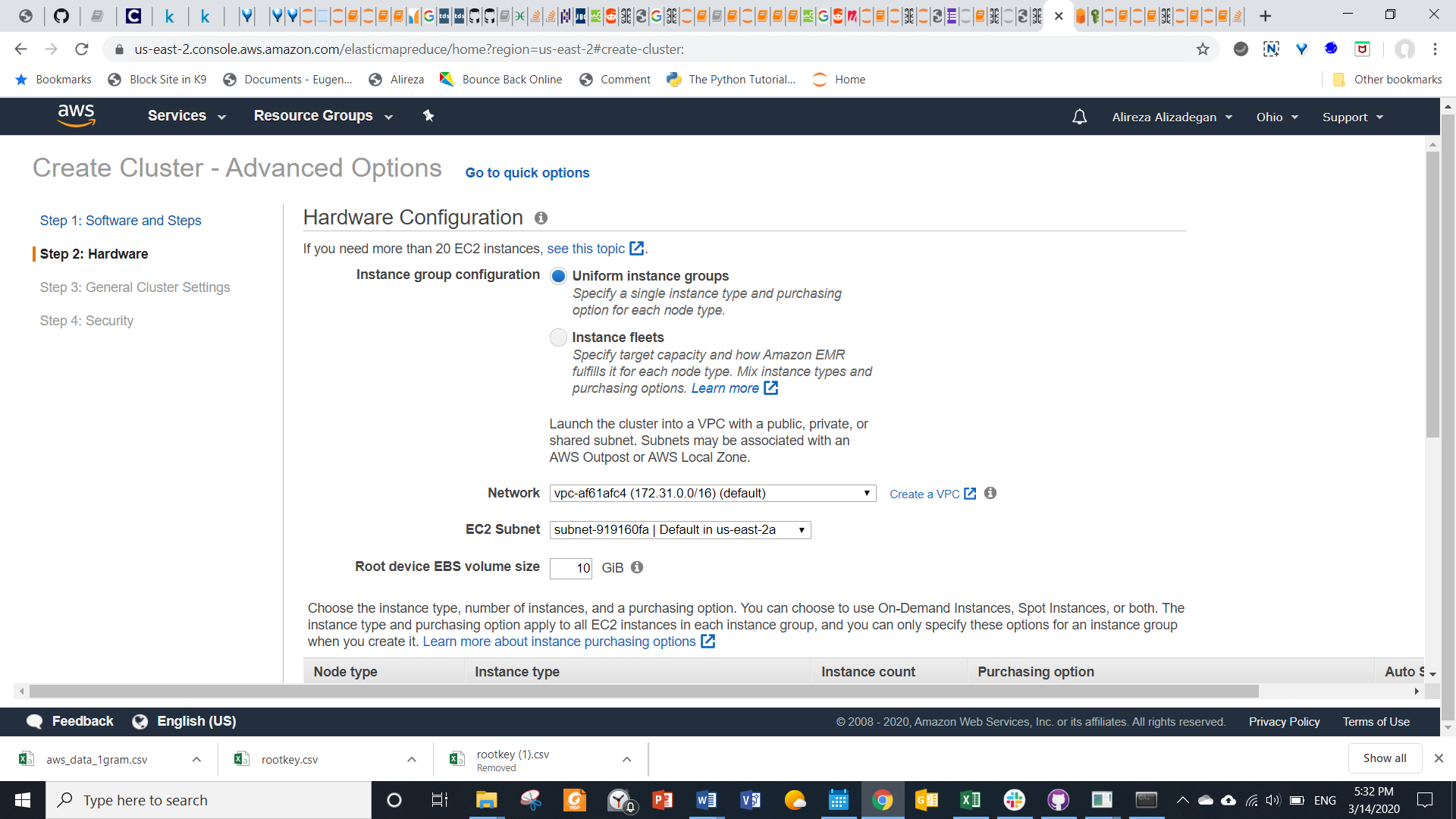
Go to advanced options in the options page



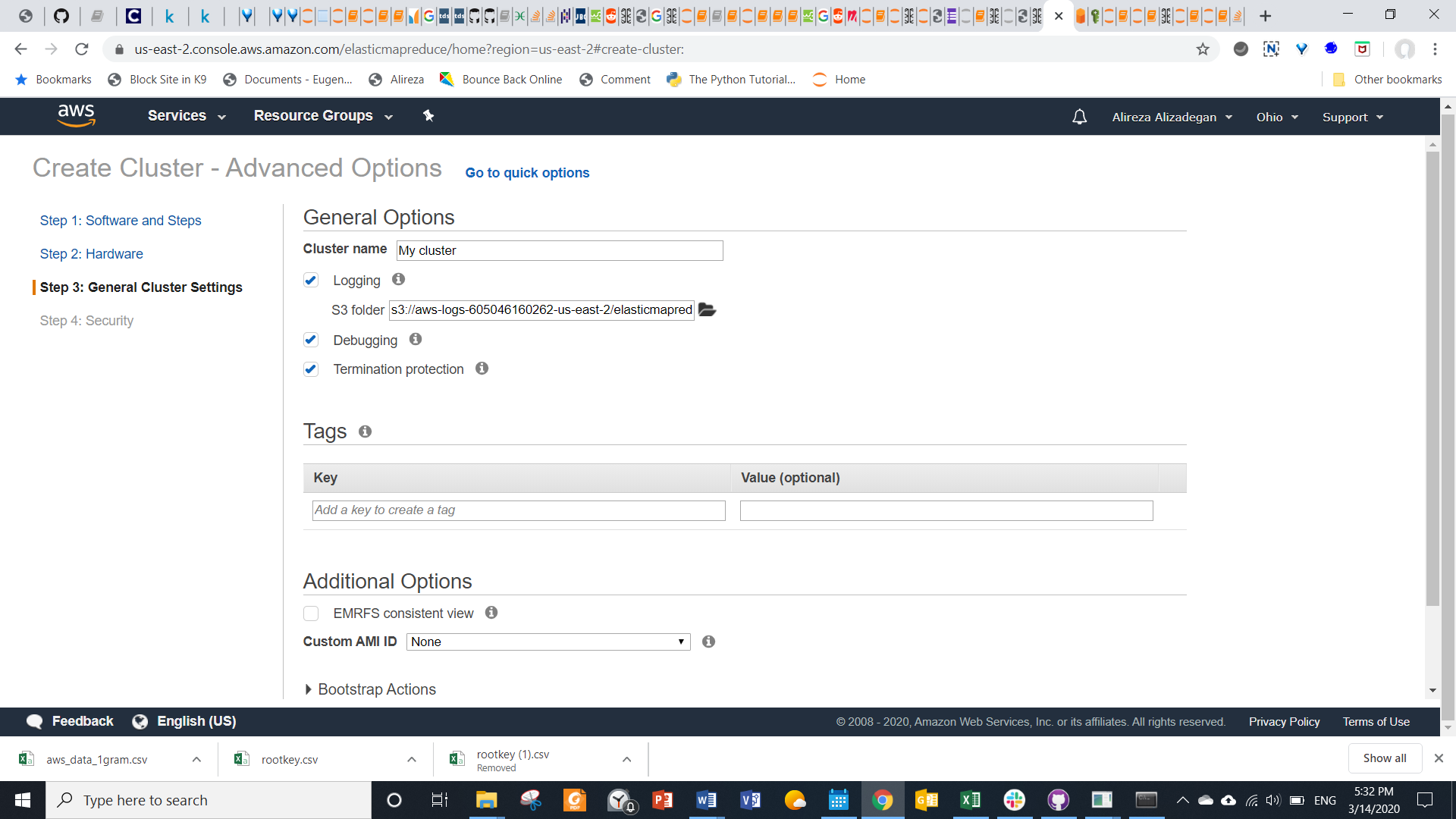
Select Hadoop, jupyterhub, hive, and spark in software & steps step



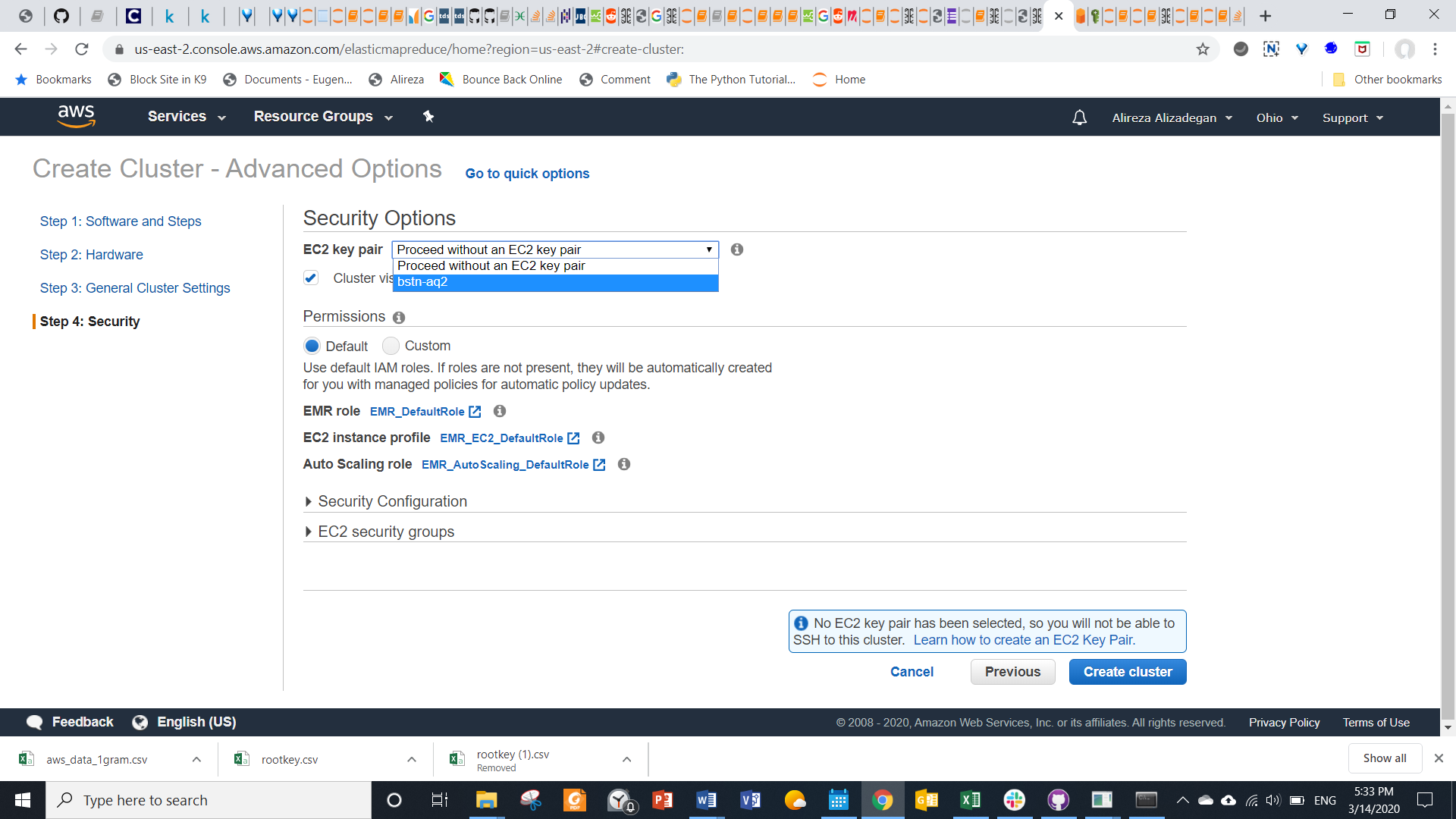
Do not make any changes in hardware step



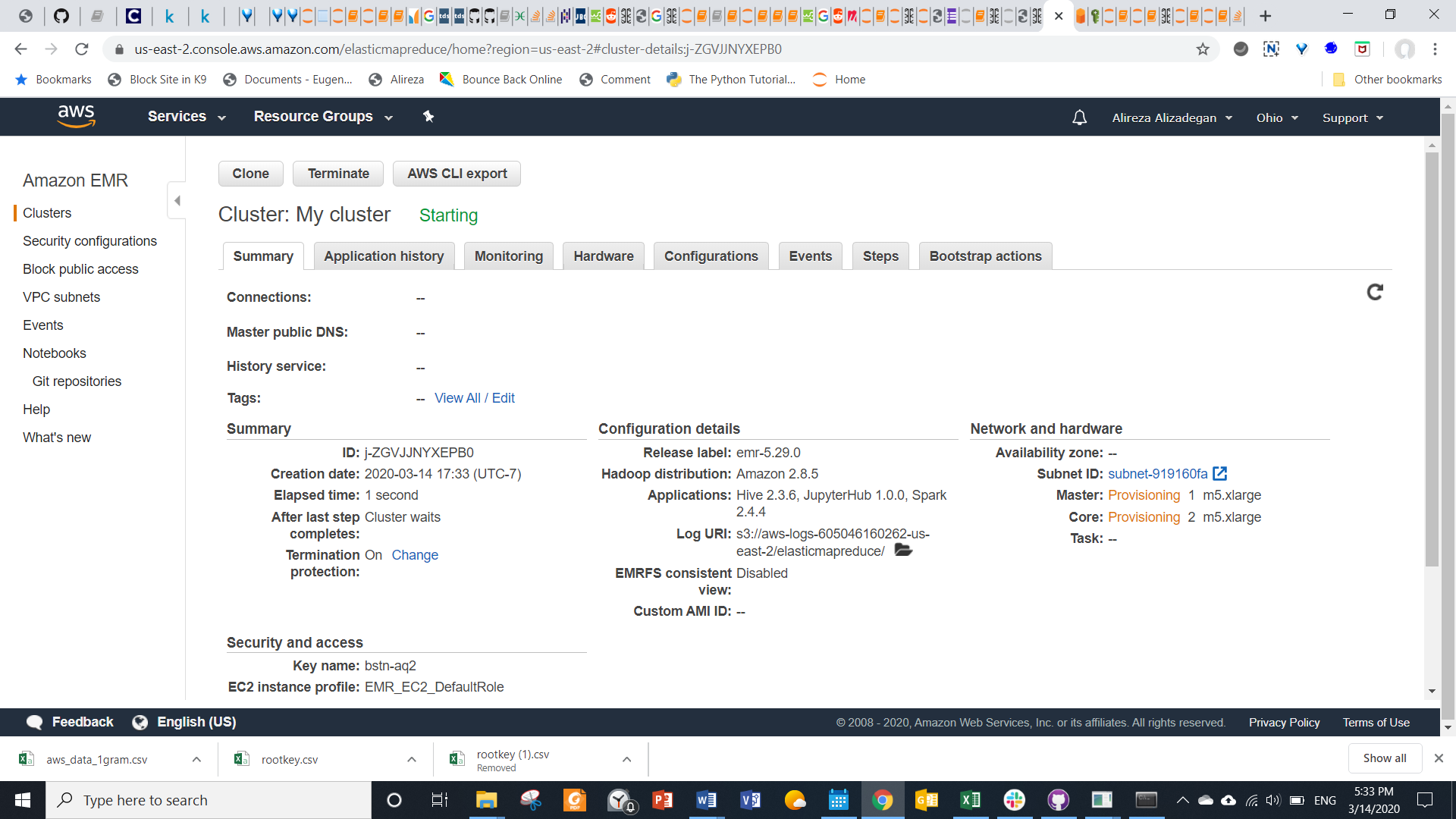
Name the cluster in the general cluster settings step



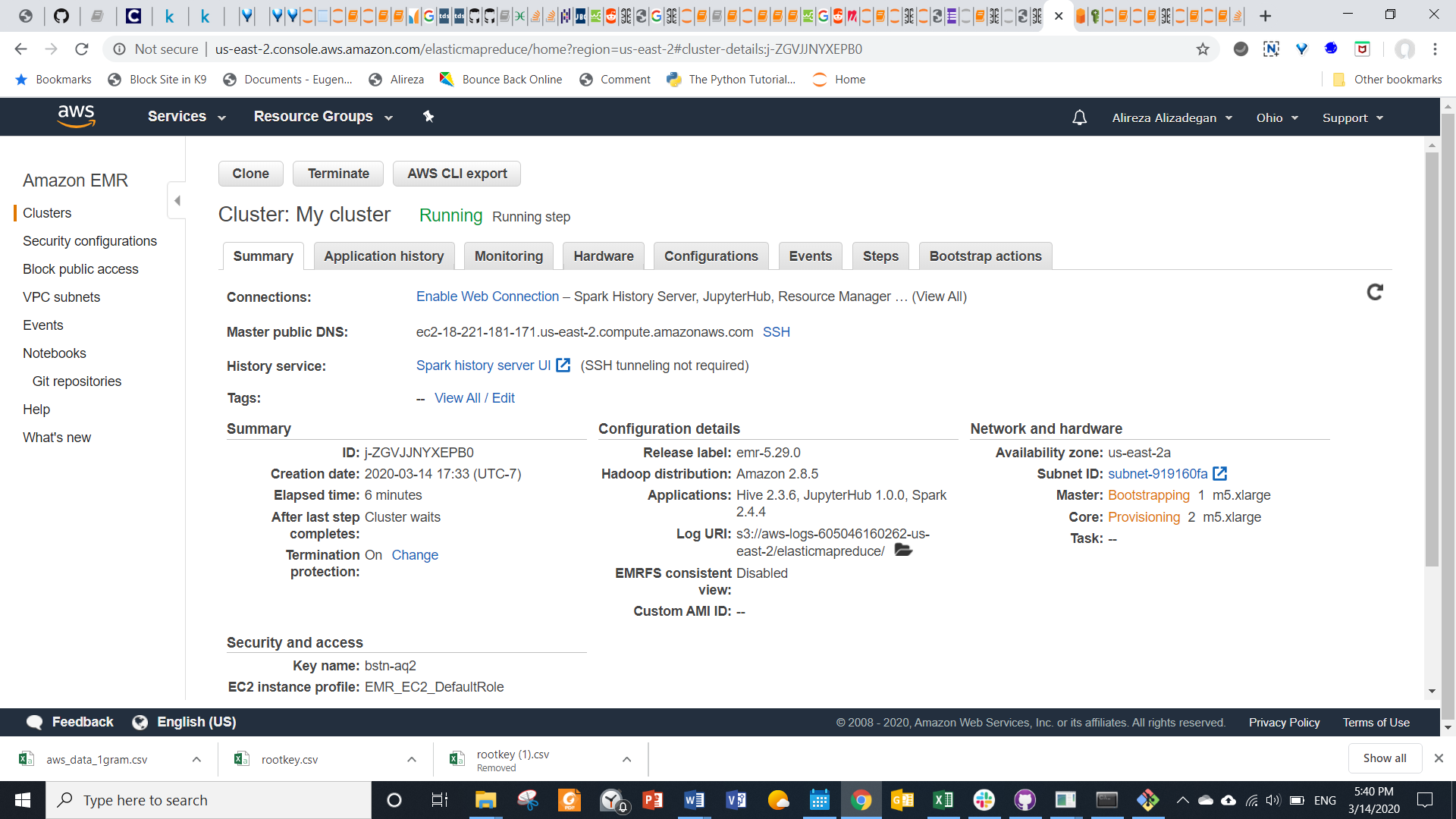
Select the EC2 key pair in security options section of security step and click create cluster



Starting status shows up in green font in front of the cluster name as below



Wait until status changes to running as below then click on SSH link in master public DNS



# Connect to master node of cluster

Add location of the private key to provided command in step 2 of instructions



Paste command in git bash and enter Hadoop environment as below



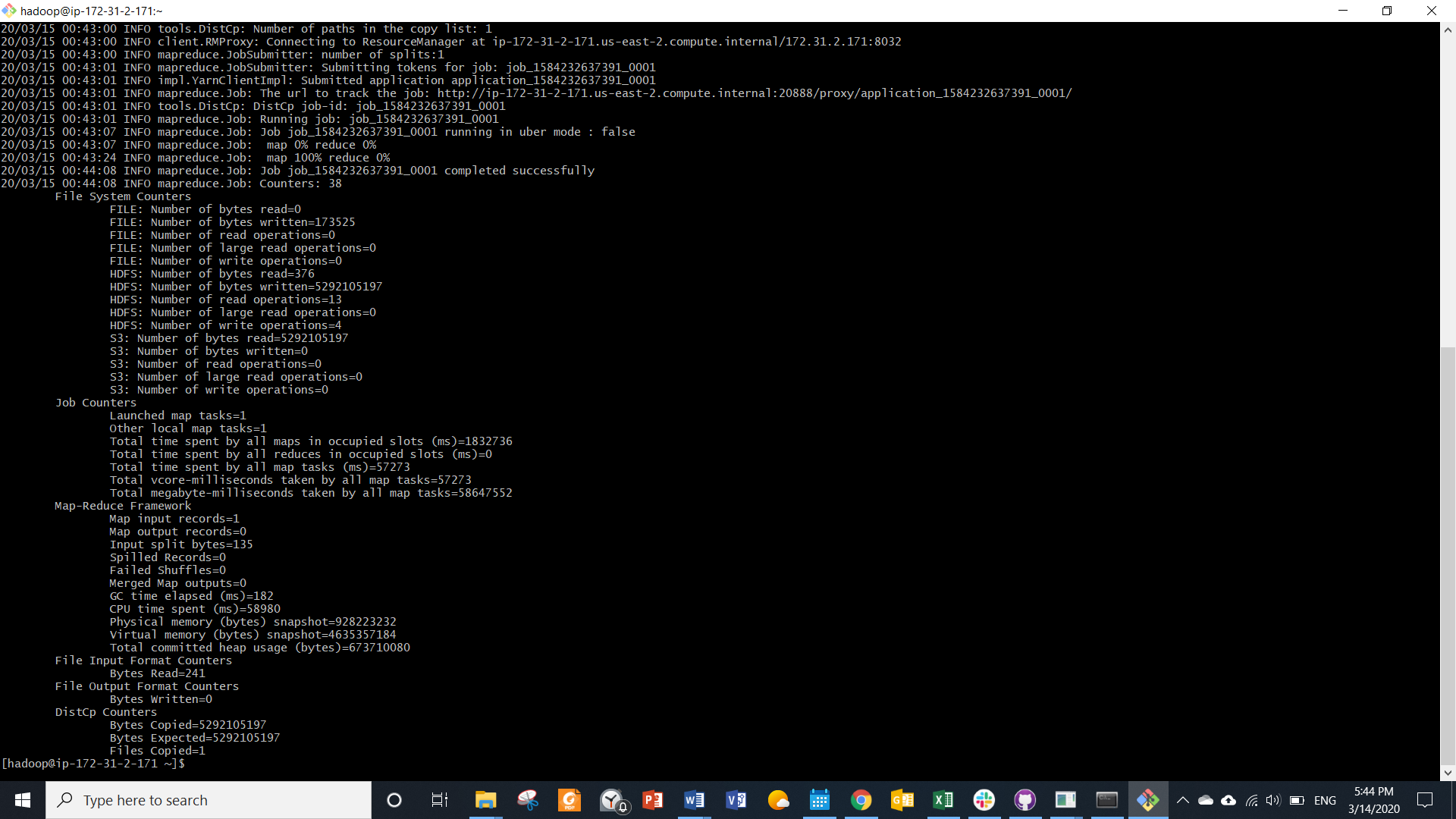
# Copy data from S3 bucket to Hadoop directory

Create a directory in Hadoop environment



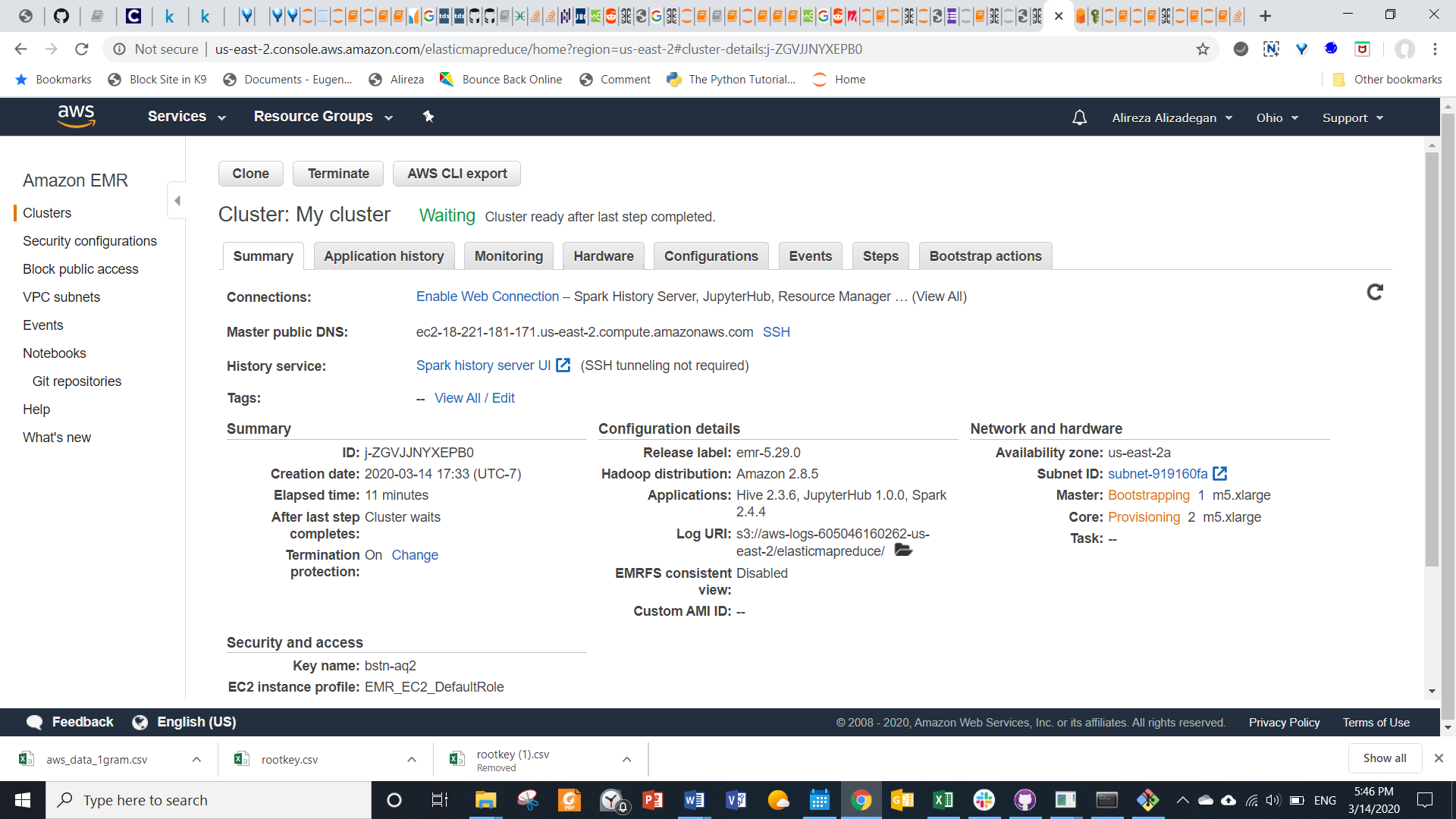
Copy data from brainstation public repository to directory by pasting following command in git bash

hadoop distcp s3://brainstation-dsft/eng\_1M\_1gram.csv /user/hadoop/eng\_1M\_1gram

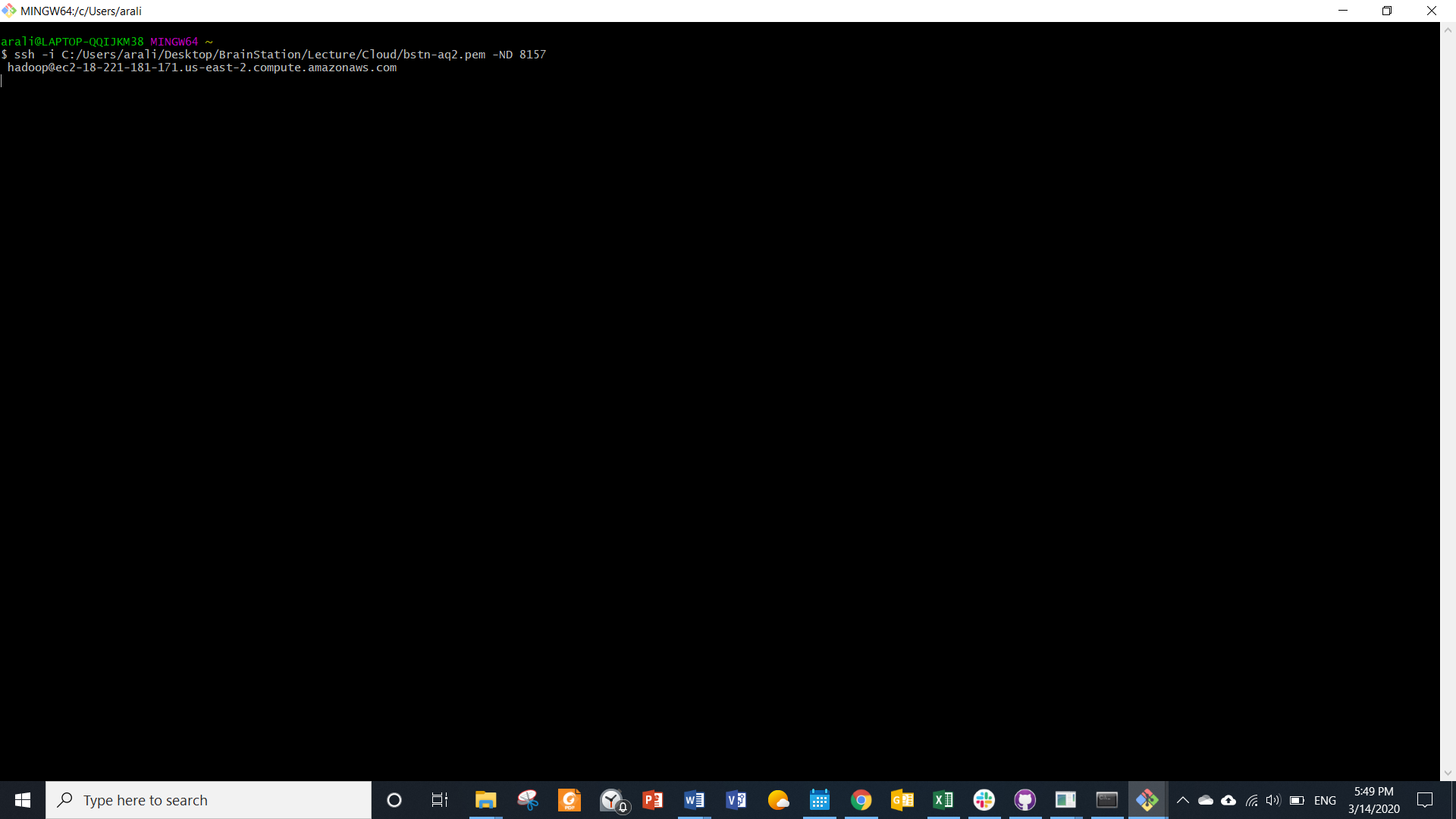


# Analyze the data with spark

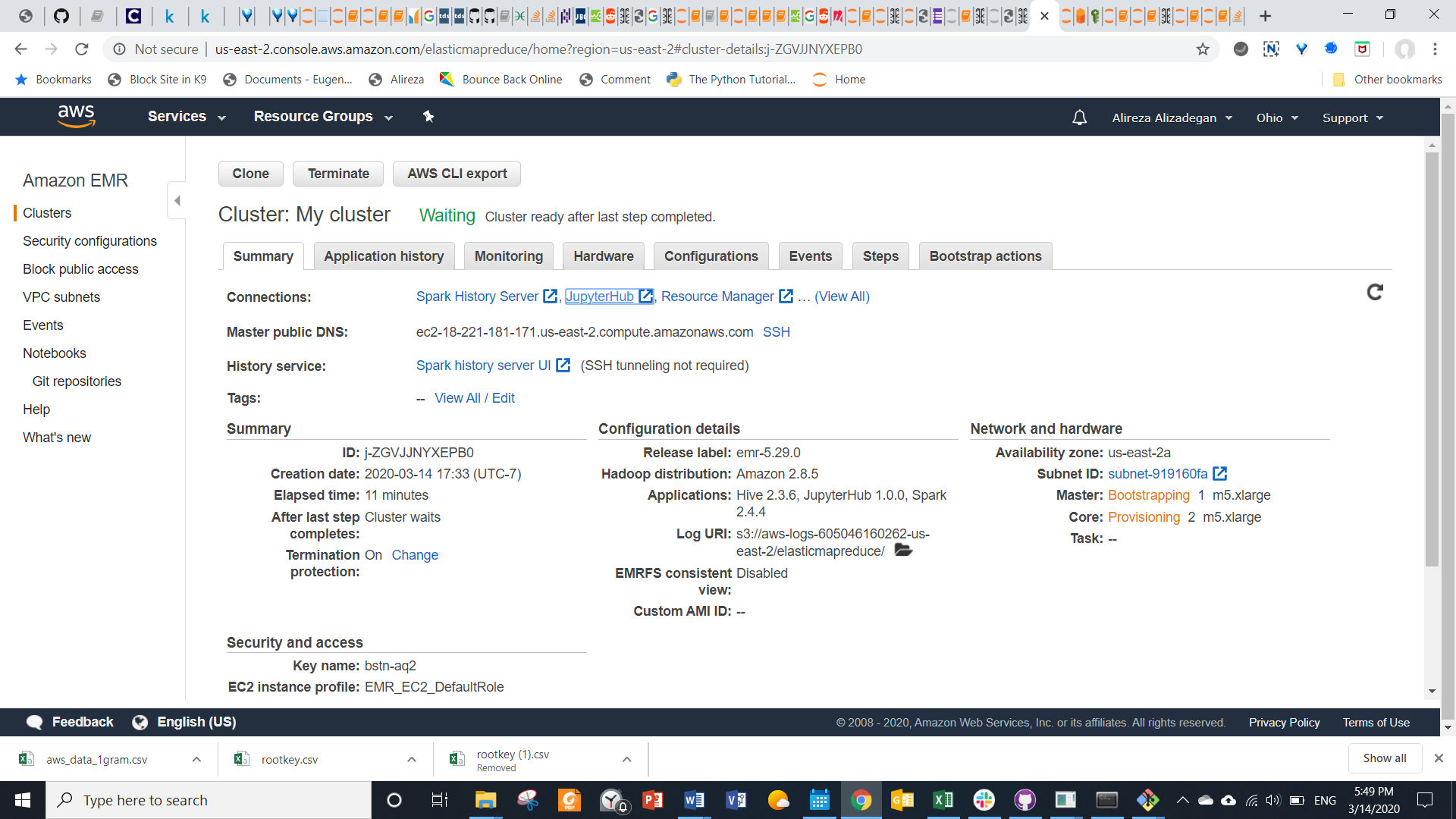
Click enable web connections link in front of connections



Add location of the private key to provided command in step 2 of instructions and paste it in git bash as below



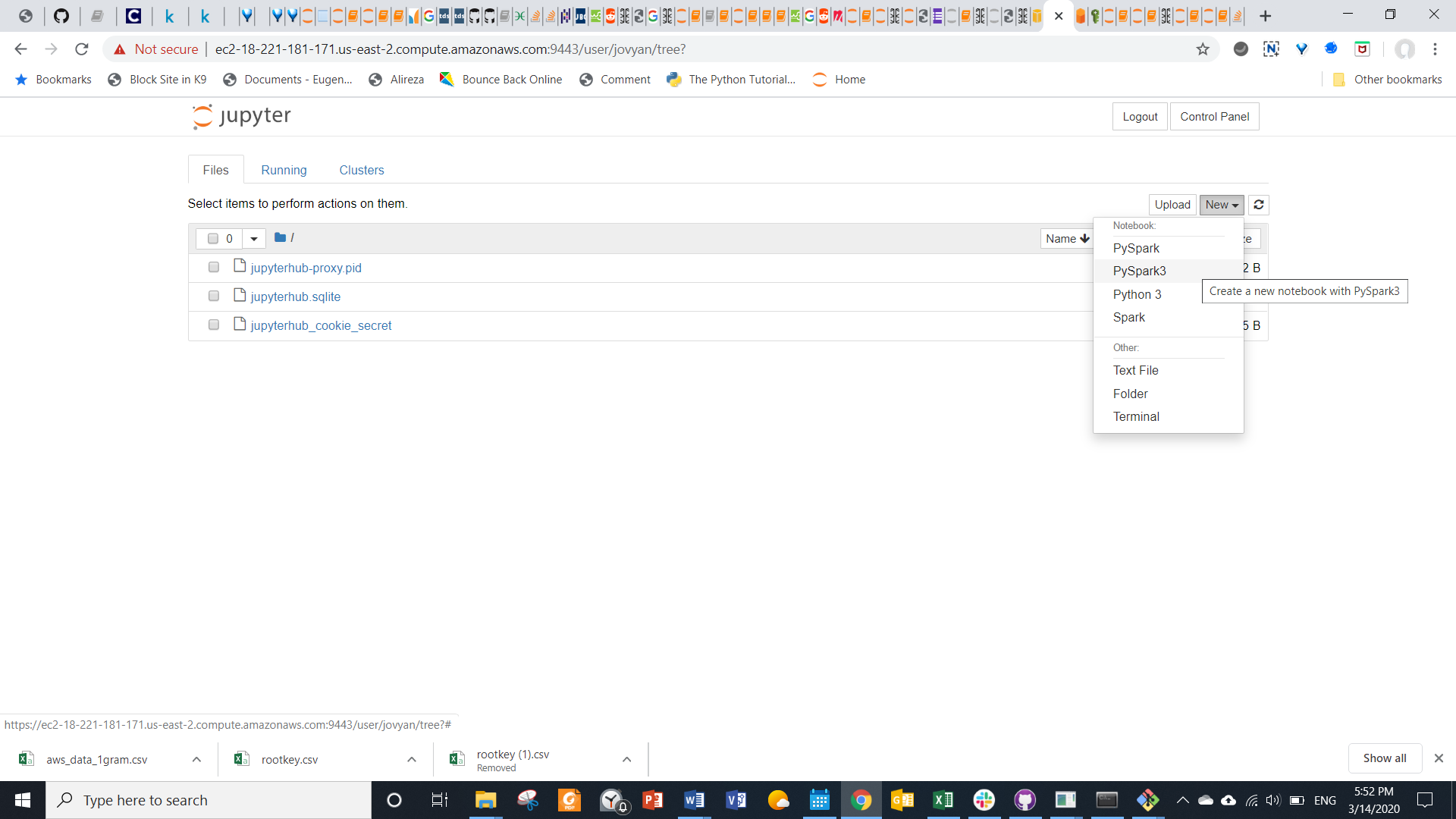
Click the activated JupyterHub link in connections as below



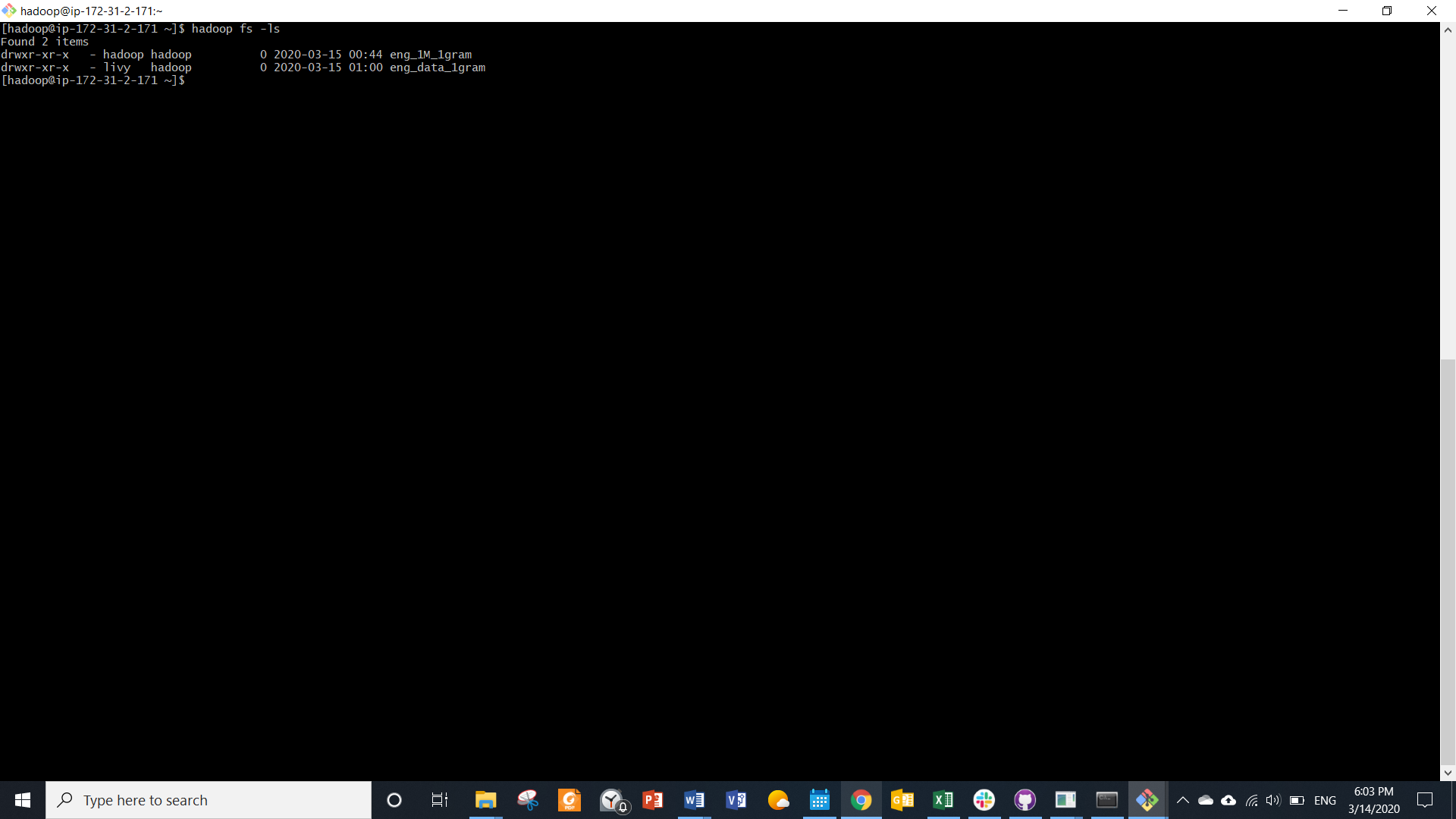
Enter ‘jovyan’ for username and ‘jupyter’ for password and sign in



Create new pyspark3 notebook and run commands as attached jupyter notebook

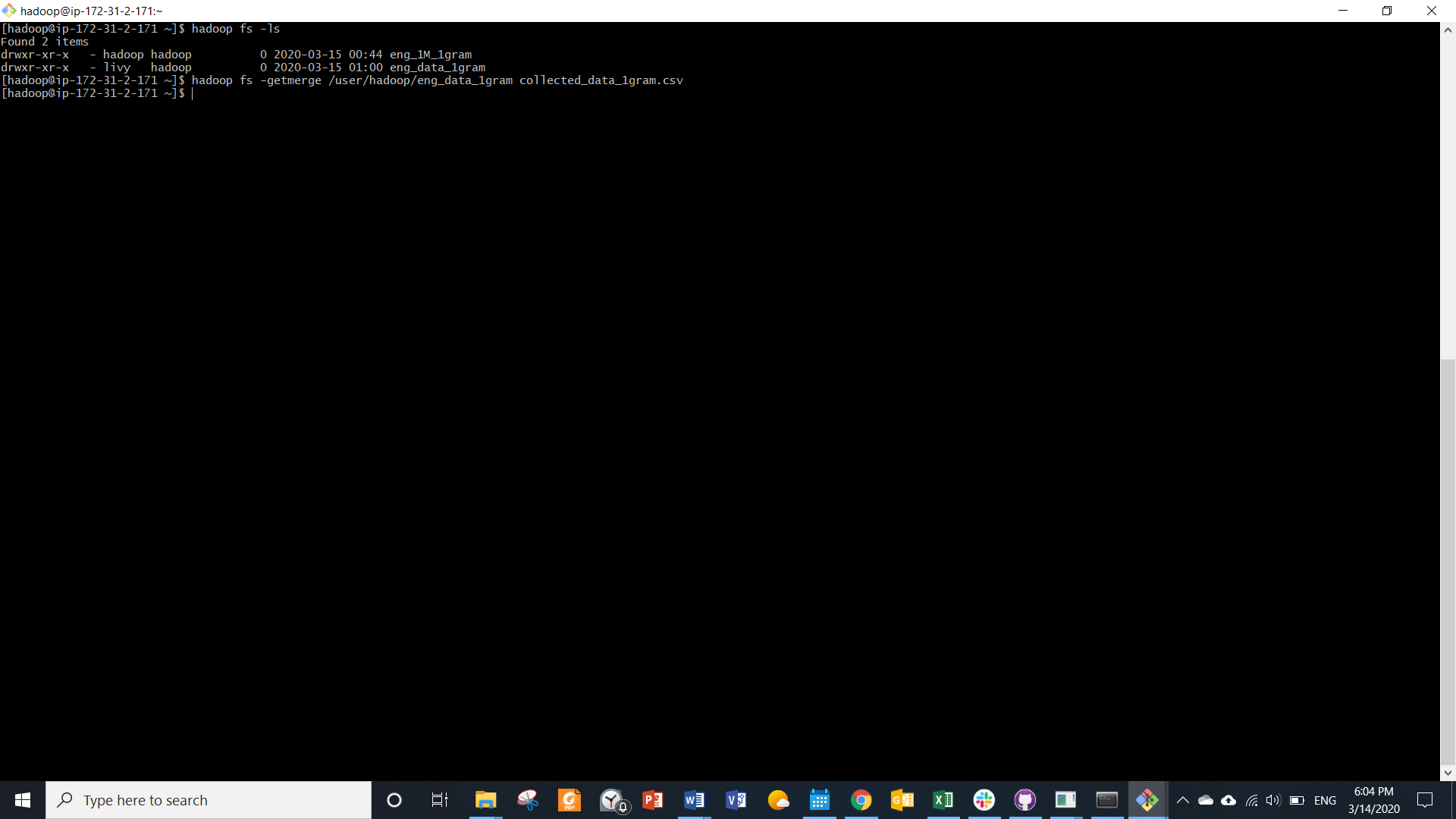


Verify that results of analytics in spark are saved to Hadoop by listing content



# Merge data into master node and copy to S3 bucket

Collect the data to master node as a CSV file



Copy the file from master node to personal AWS S3 bucket ‘lastassignmentbucket’

