Final Project: CS 6350.002 Big Data Management and Analytics

--------------------- Authors ---------------------

Mahsa Eslamialishah - mxe190002

Novonil Das - nxd180022

Pritom Das Radheshyam - pxr180033

--------------------- Files Included ---------------------

CS 6350 Final Project (Zipped File with all the code)

Report.pdf

Readme.txt

LogFile.txt

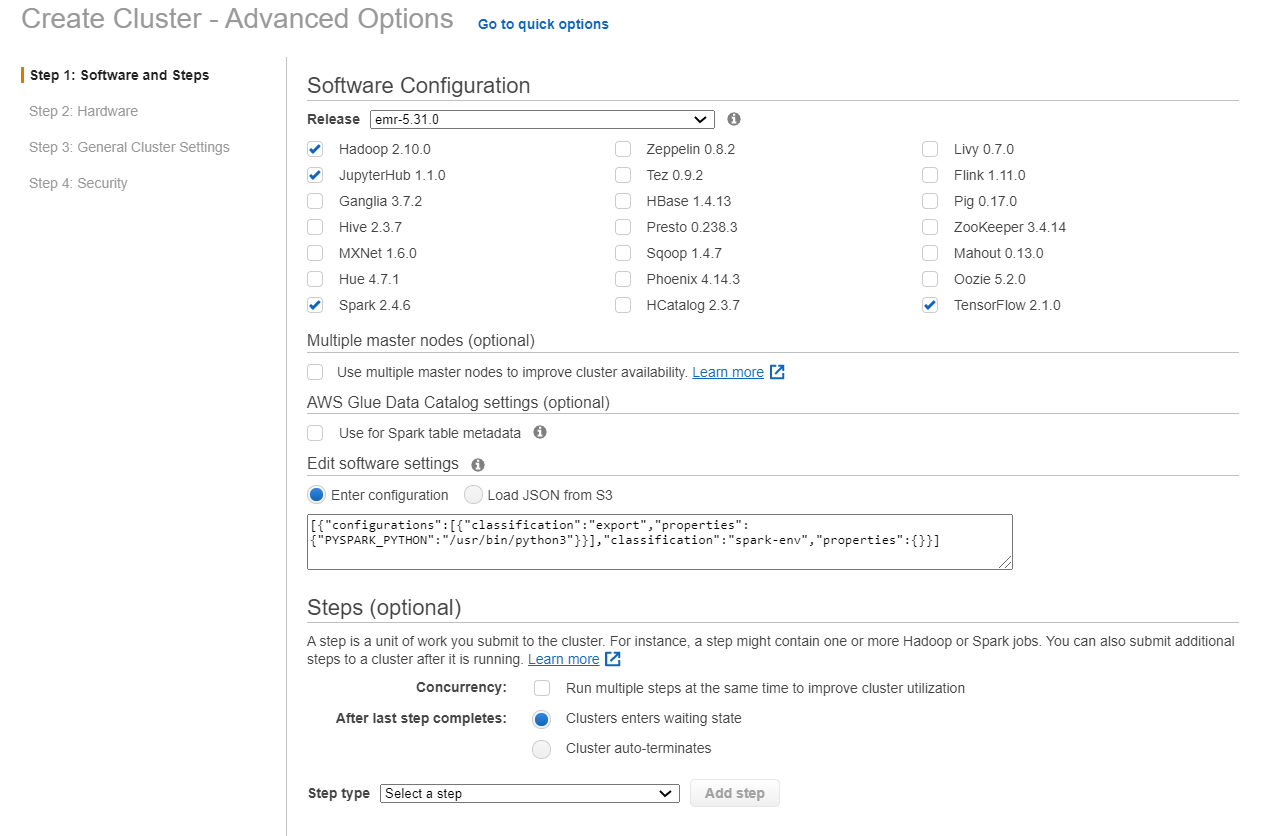
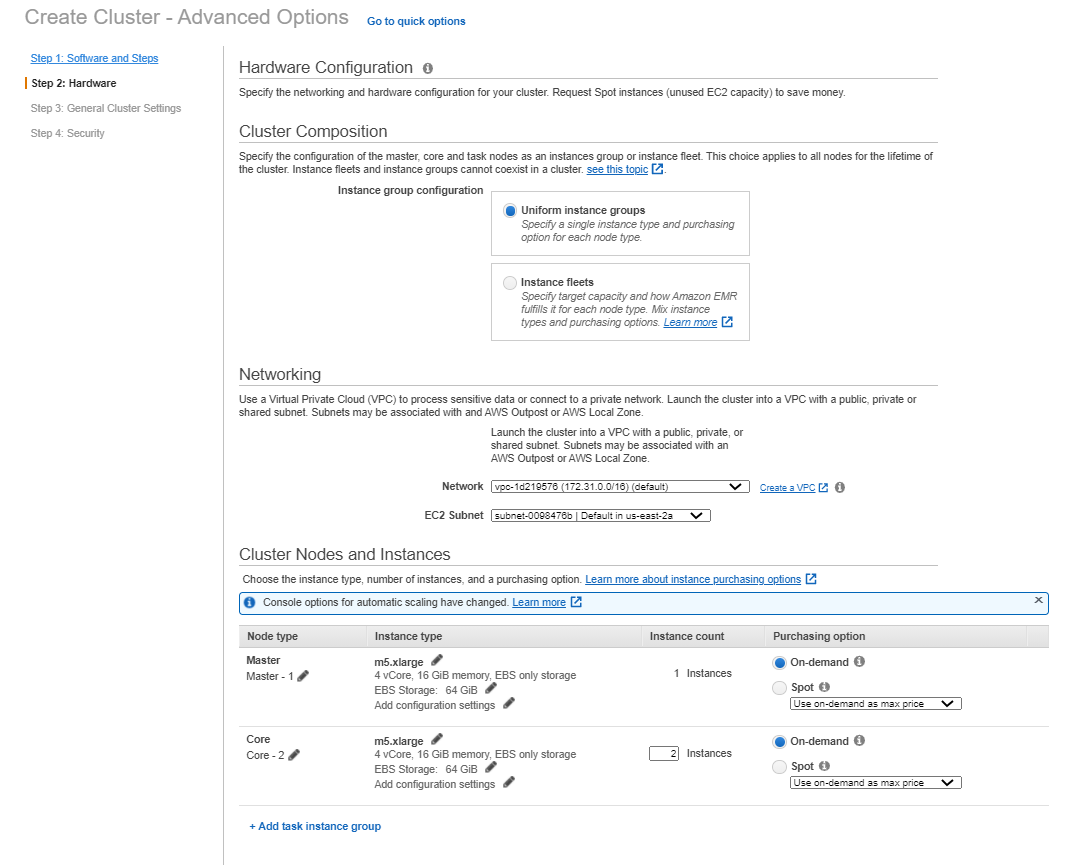
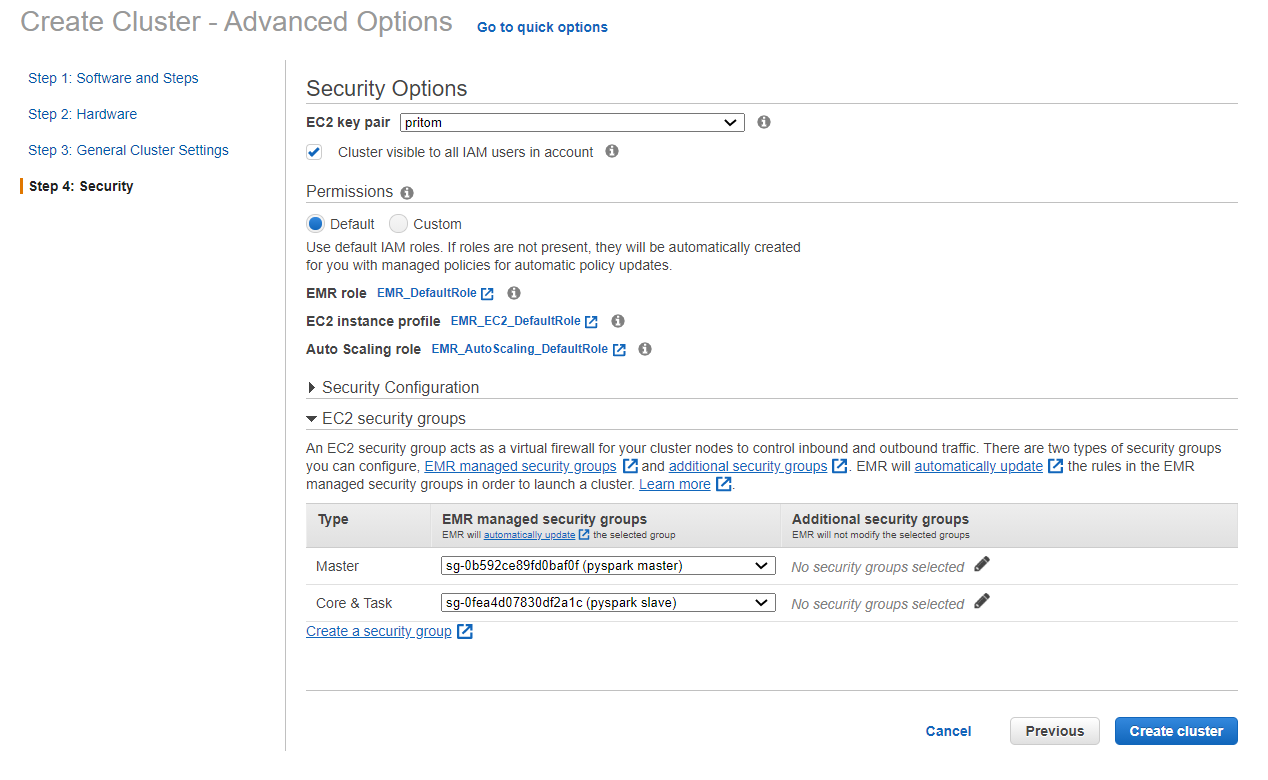
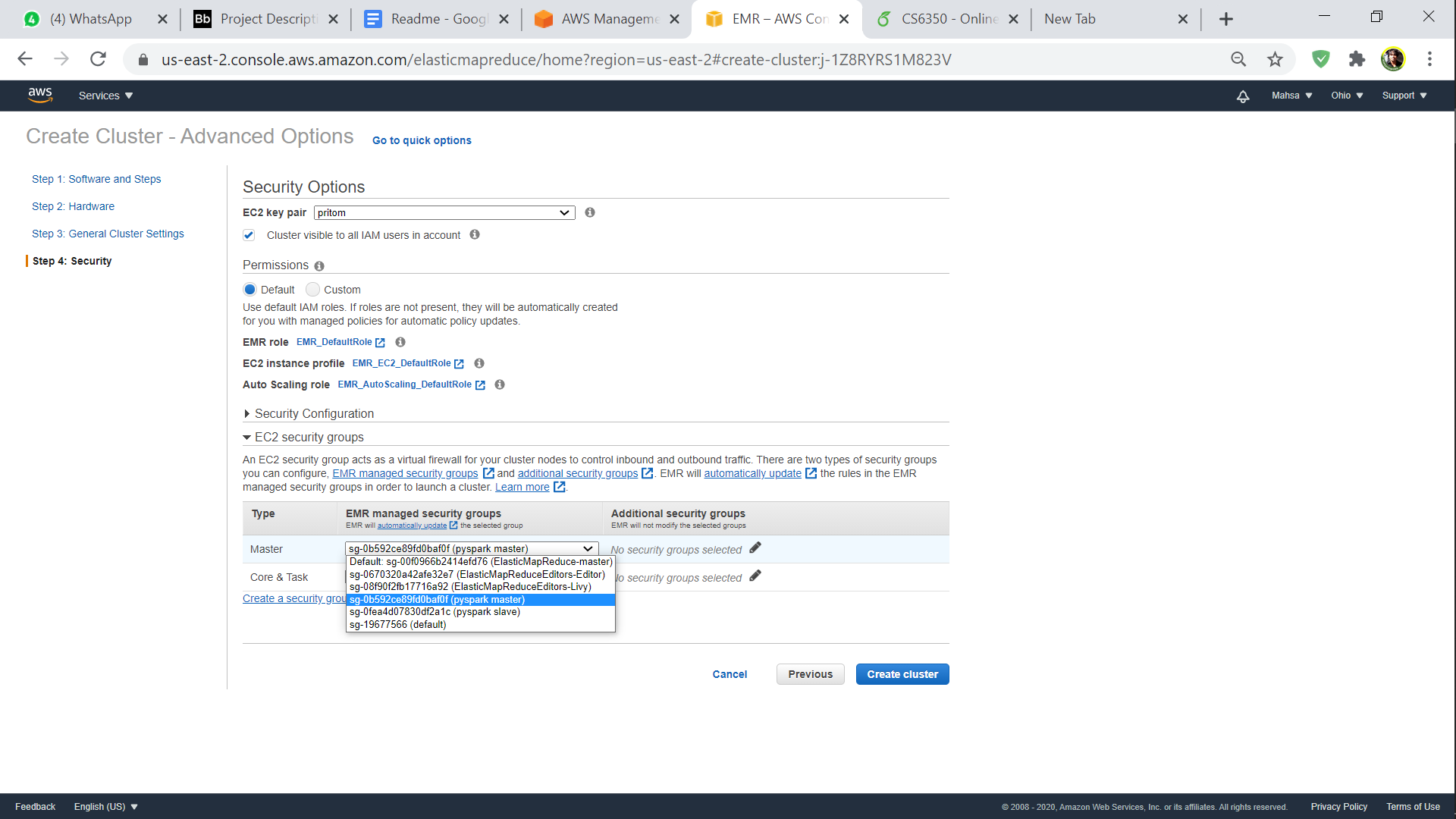
--------------------- DataSet ---------------------

Dataset for this project is hosted in AWS S3. The following details are the meta data for the dataset used.

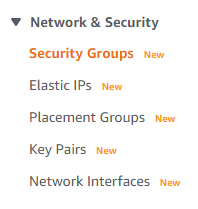
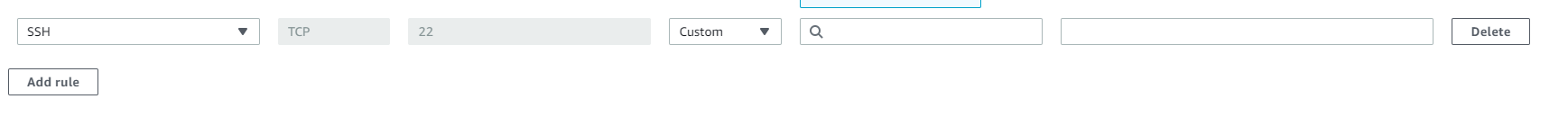
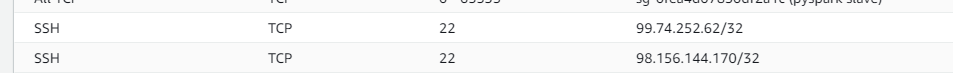
|  |  |
| --- | --- |
| Original Dataset from Kaggle | https://www.kaggle.com/mateuszbuda/lgg-mri-segmentation |
| Dataset hosted in S3 Bucket Name | braintumorproject |
| Dataset S3 URI Link | s3://braintumorproject/Healthcare\_AI\_Datasets/ |

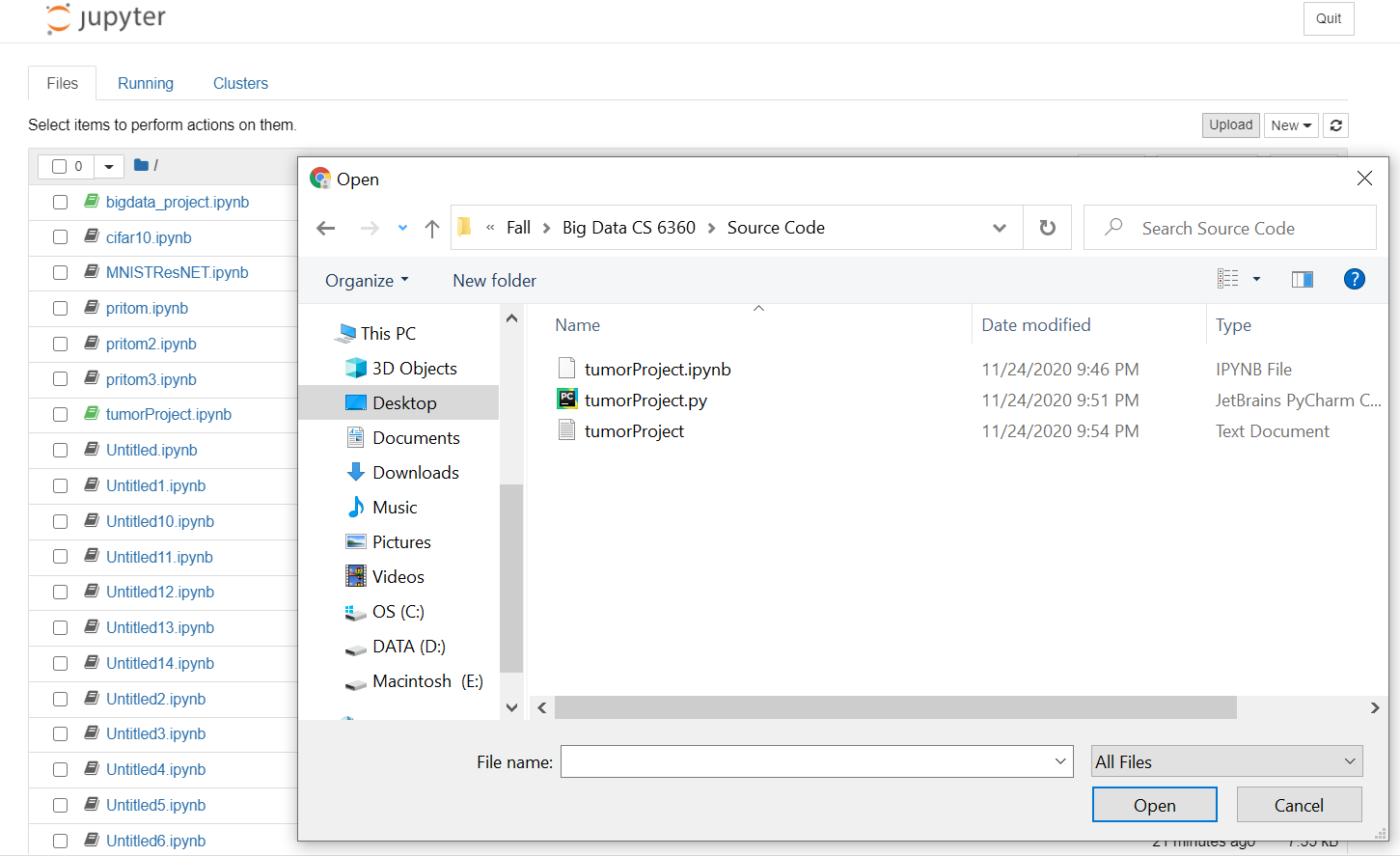
--------------------- Steps to run Project on AWS---------------------

1. The Dataset is hosted on a public AWS S3 Bucket
2. Start an EMR cluster in AWS. Follow these steps:-

* Change region to “US East (Ohio) us-east-2”
* Create an EMR Spark Cluster by following the images below.
* 
* 
* 
* Need to create a EC2 Key Pair if there is a need to SSH into the name node.
* 
* Then create the cluster.

1. To allow connection between the EMR to access the bucket in S3, change settings in “Security Group” which is found in AWS EC2. Follow the steps below.

* Head over to AWS EC2.
* Change region to “US East (Ohio) us-east-2”
* From the Dashboard Panel, select “Security Groups”
* Create Security Groups, one for the name node and another for the data nodes which are being created by the AWS EMR.
* Add Inbound Rule, to allow access.
* We have added two of our machine’s IP address in similar way.

1. Head on to AWS EMR again, and after the cluster is ready, create a “Notebook” and attach to an existing cluster.
2. Open the notebook in “Jupyter” which is hosted over the AWS EMR.
3. Upload “tumorProject.ipynb” which can be found under the folder “Source Code”. 
4. Run all the cells.
5. Under parameter which is denoted as “Params”, change the epoch number as needed. Higher epoch will fetch better results but will need longer time to execute. 