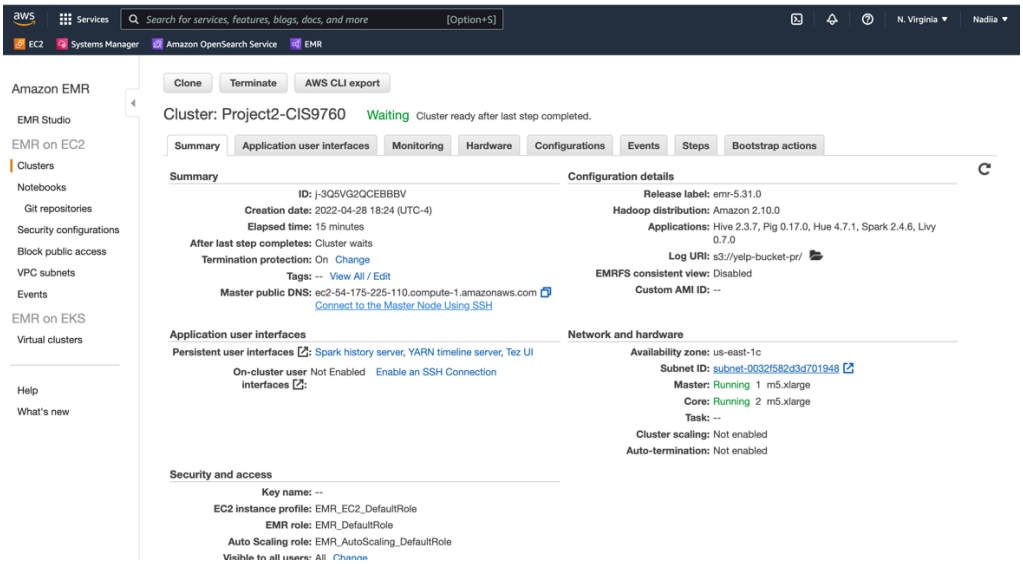
Project 2

Analyzing 10Gb of Yelp's Review Data
- Language: Python (PySpark)
- Libraries: pandas, matplotlib, SciPy, seaborn
- AWS Services: S3, EMR
- Data: Yelp Dataset (https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset)

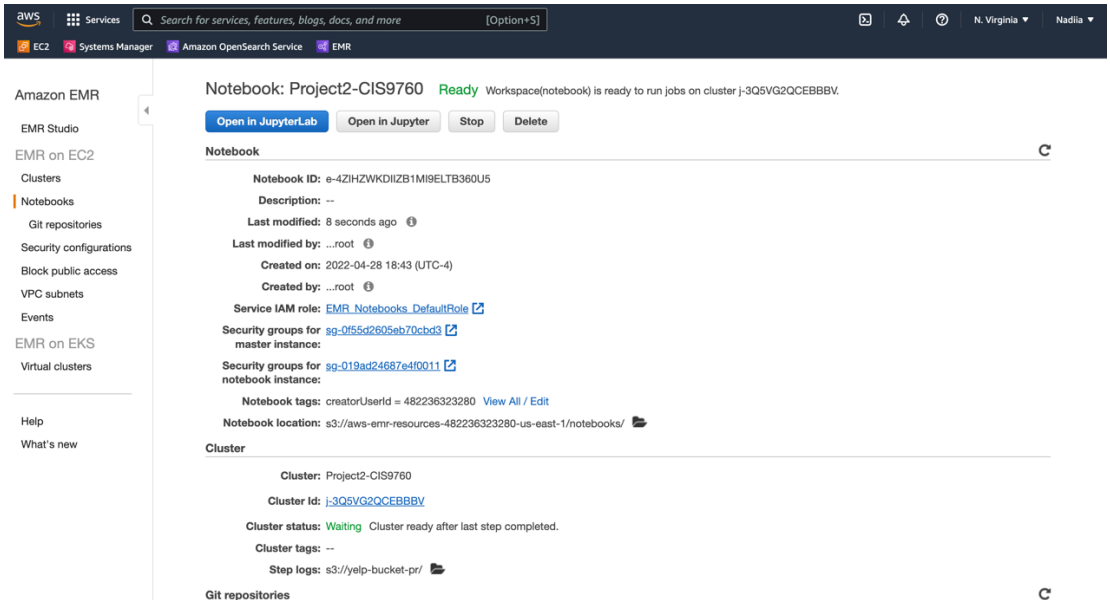
For this project, 10Gb of Yelp's Review Data were loaded into an AWS S3 bucket. This project aims to analyze Yelp's Reviews, Businesses and Users datasets by provisioning a Spark Cluster on AWS EMR and running analysis via Jupyter Notebook.

Steps:
1) Create a cluster on EMR
   - Provisioning the Hardware
   - Configuring Jupyter Notebook
2) Running Spark cluster via Jupyter Notebook
   - Notebook operations and Kernel
   - Loading data into S3
   - Loading data to EMR from S3

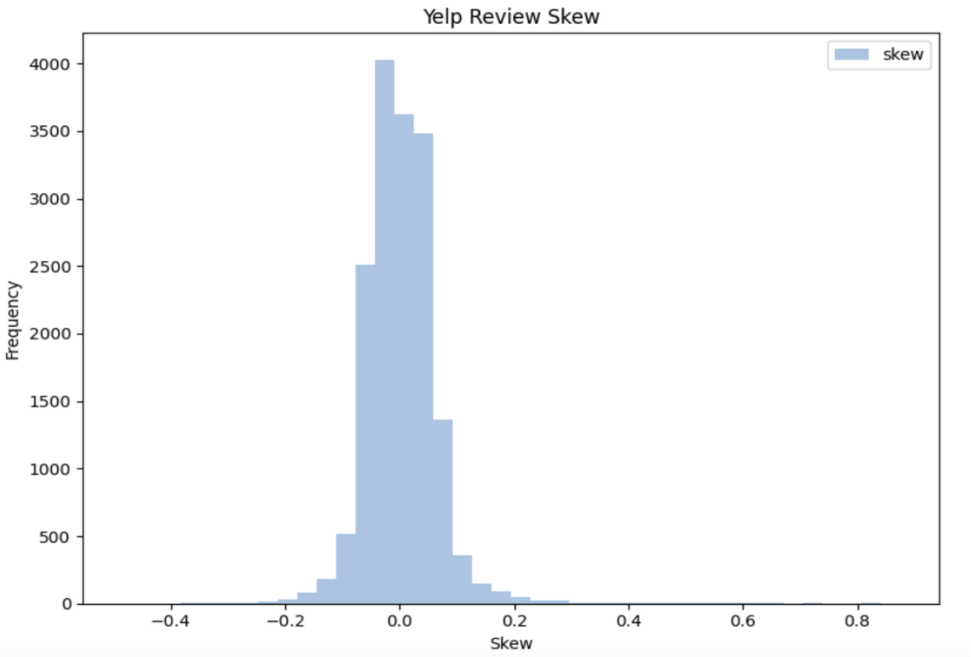
Cluster Configuration



Notebook Configuration

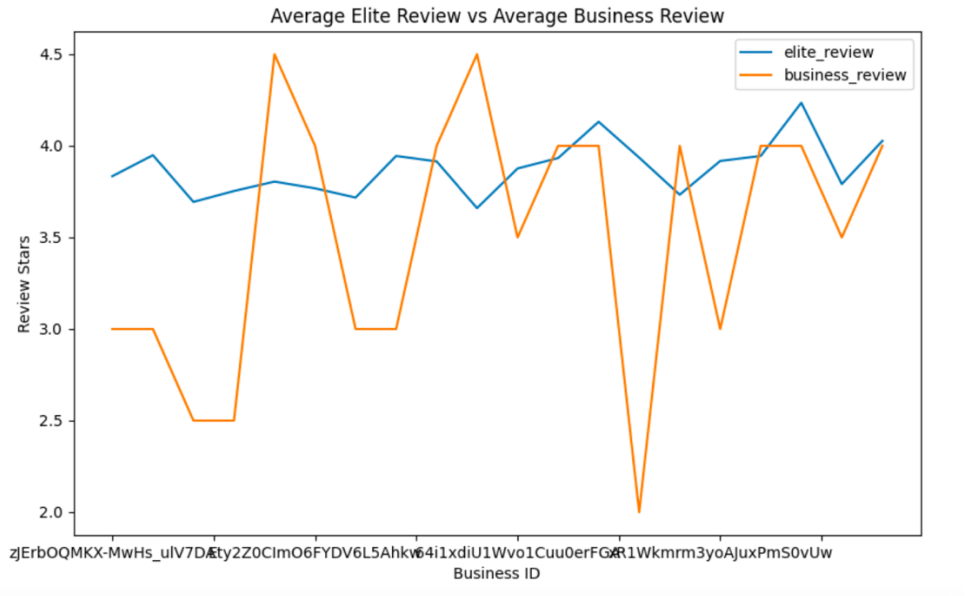


• Analysis 1 – Yelp Review Skew



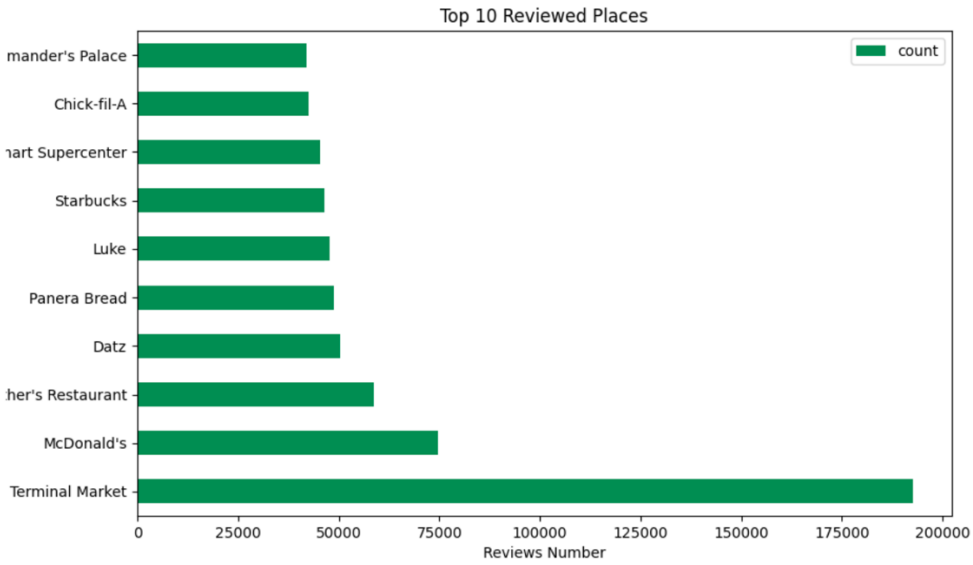
In order to analyze Yelp review skew trend, a histogram was used. This visualization shows that according to the skew, reviewers who left both positive and negative written responses were posted equally on Yelp, without bias for only positive reviews.

• Analysis 2 – Average Elite Review vs Average Business Review



To visualize a trend of average elite reviews and average business reviews, a line graph was used. This visualization shows that average reviews from elite users tend to be more positive. Average business reviews, on the other hand, are much more varied between positive and negative. From this observation, we can conclude that reviews from Yelp's elite users cannot equally be trusted.

• Analysis 3 – Top 10 Reviewed Places



To visualize the top ten reviewed places, a horizontal bar chart was used. This visualization shows that the “Reading Terminal Market,” is leading among the top 10 reviewed places, in which 19,2708 reviews were received. “McDonald's” (74,551) was the second-highest reviewed place. The next highest ranked places were “Mother's Restaurant” (58,724), “Datz” (50,503), and “Panera Bread” (48,794), respectively.