Project 2 - Analyzing 10Gb of Yelp Reviews Data

Data Set

- Yelp's Reviews and Businesses dataset (about 10gb) from **Kaggle**
- Data Frames – Business, Review, User

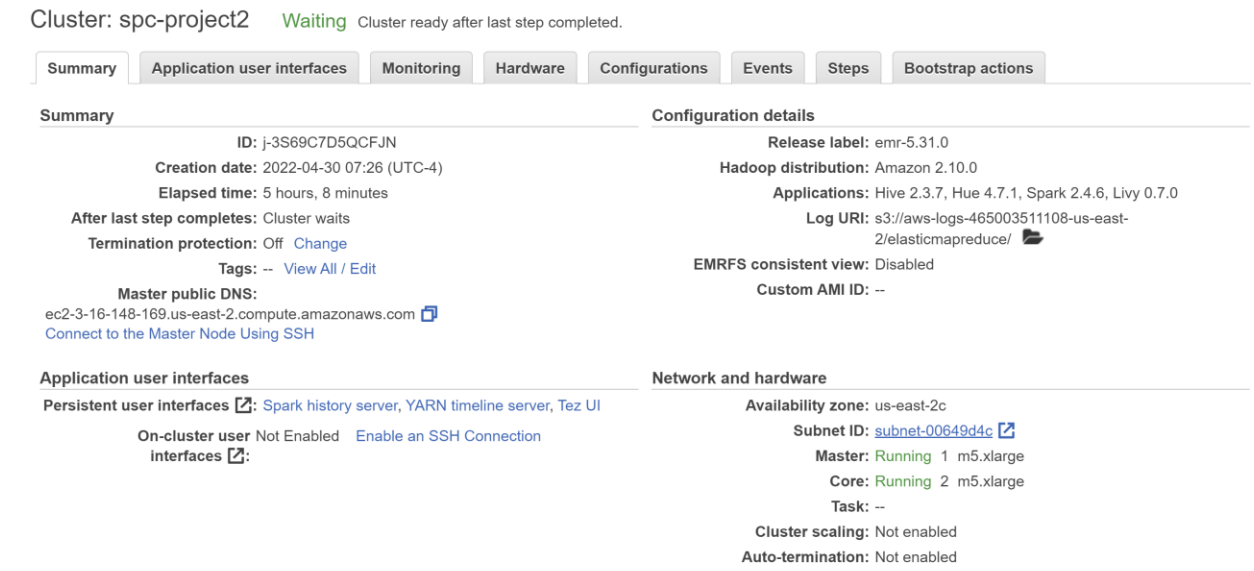
Introduction

- The dataset is loaded onto a S3 bucket, and this URI is used to load them into spark cluster.
- We provision a Spark cluster on AWS EMR for loading and analyzing the above-mentioned dataset.
- The whole analysis is produced in Jupyter Notebook created on that cluster. (Analysis.ipynb)

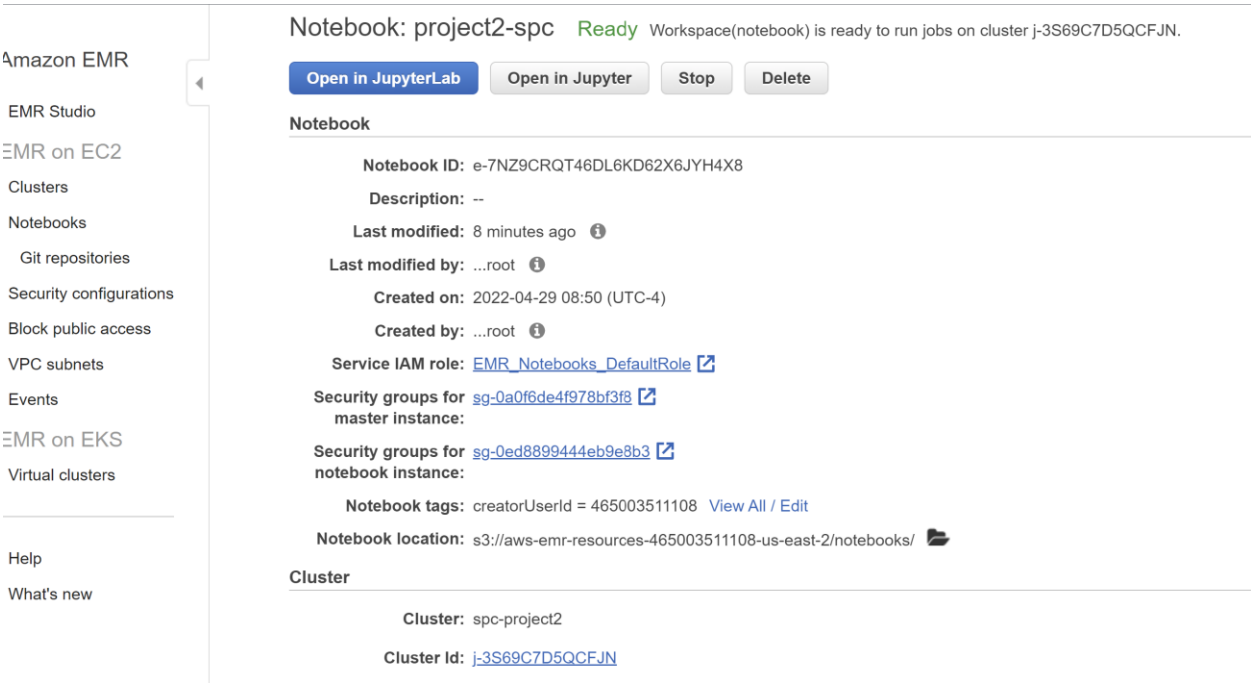
Steps to configure Spark Cluster and Notebook

1) Creating a cluster on EMR
    - Provision a "cluster" – with a single master node and two worker nodes in EMR. This is the hardware necessary to run our Spark jobs.
    - Open AWS EMR page and click on 'Create Cluster'.
    - In the advanced configuration options, select emr-5.31.0 as the Release option. Unselect Pig, Select Spark and Livy.
    - Select the instance types as m5.xlarge for both master as well as the core nodes
    - Give the cluster a relevant name and click on create cluster
    - Refer to the cluster configuration for more information
    - Click on the create notebook option under EMR Notebooks option
    - Give a relevant name to the notebook, choose the previously created cluster.
    - Once the notebook is ready, it can be opened in jupyter host for writing your analysis
    - Refer to the notebook configuration for more information
2) Running Spark cluster tasks via Jupyter Notebook
    - Change the kernel to PySpark before running any jobs
    - Type in `%%info` to check that everything is working as expected

Cluster configuration



Notebook configuration



Packages installed

- Matplotlib 3.2.1
- Pandas 1.0.3
- Scipy 1.7.1
- Seaborn 0.11.2

Analysis

Various questions related to the dataset have been answered using pyspark data frame capabilities.

Utilized various libraries to filter, group, transform, and render visualizations in jupyter notebook.