

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

The screenshot displays the AWS Management Console interface for an Amazon EMR cluster. The top navigation bar shows the AWS logo, a search bar, and the current region (N. Virginia) and user (Nadia). The left sidebar lists various AWS services, with 'Amazon EMR' selected. The main content area shows the cluster 'Project2-CIS9760' in a 'Waiting' state. The 'Summary' tab is active, displaying key information about the cluster. The 'Configuration details' section provides a comprehensive overview of the cluster's setup, including its ID, creation date, release label, Hadoop distribution, applications, log URI, and network configuration. The 'Application user interfaces' section shows that the Spark history server, YARN timeline server, and Tez UI are enabled. The 'Network and hardware' section details the availability zone, subnet, and instance types (m5.xlarge) for the master and core nodes. The 'Security and access' section lists the key name, EC2 instance profile, EMR role, and auto scaling role.

Summary	Configuration details
ID: j-3Q5VG2QCEBBBV	Release label: emr-5.31.0
Creation date: 2022-04-28 18:24 (UTC-4)	Hadoop distribution: Amazon 2.10.0
Elapsed time: 15 minutes	Applications: Hive 2.3.7, Pig 0.17.0, Hue 4.7.1, Spark 2.4.6, Livy 0.7.0
After last step completes: Cluster waits	Log URI: s3://yelp-bucket-pr/
Termination protection: On	EMRFS consistent view: Disabled
Tags: -- View All / Edit	Custom AMI ID: --
Master public DNS: ec2-54-175-225-110.compute-1.amazonaws.com	
Connect to the Master Node Using SSH	

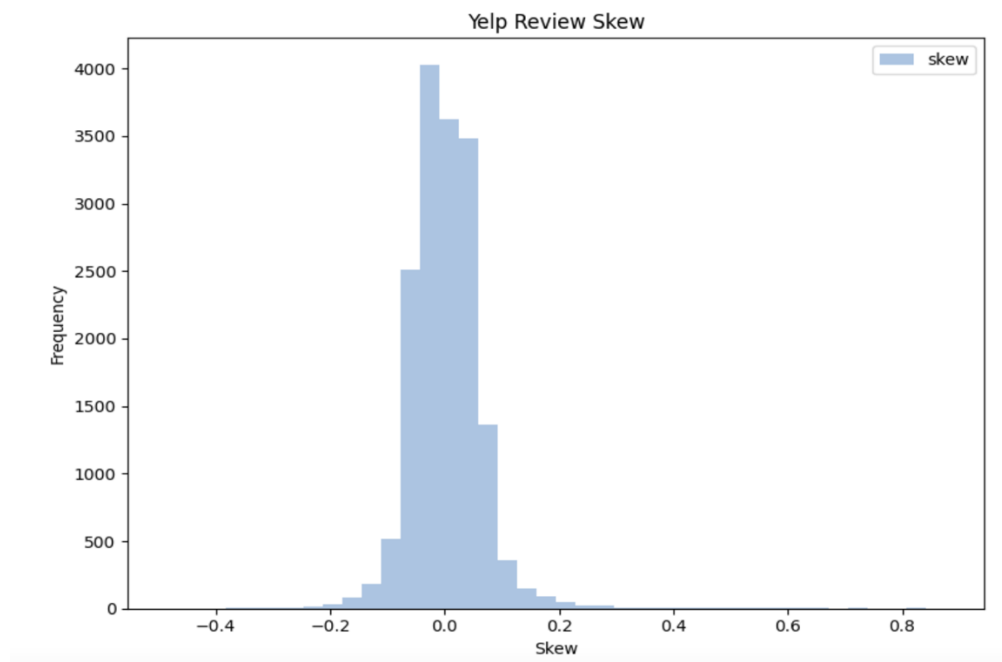
Application user interfaces	Network and hardware
Persistent user interfaces: Spark history server , YARN timeline server , Tez UI	Availability zone: us-east-1c
On-cluster user: Not Enabled Enable an SSH Connection	Subnet ID: subnet-9032f582d3d701948
Interfaces: View All / Edit	Master: Running 1 m5.xlarge
	Core: Running 2 m5.xlarge
	Task: --
	Cluster scaling: Not enabled
	Auto-termination: Not enabled

Security and access
Key name: --
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Auto Scaling role: EMR_AutoScaling_DefaultRole
Visible to all users: All Change

Notebook Configuration

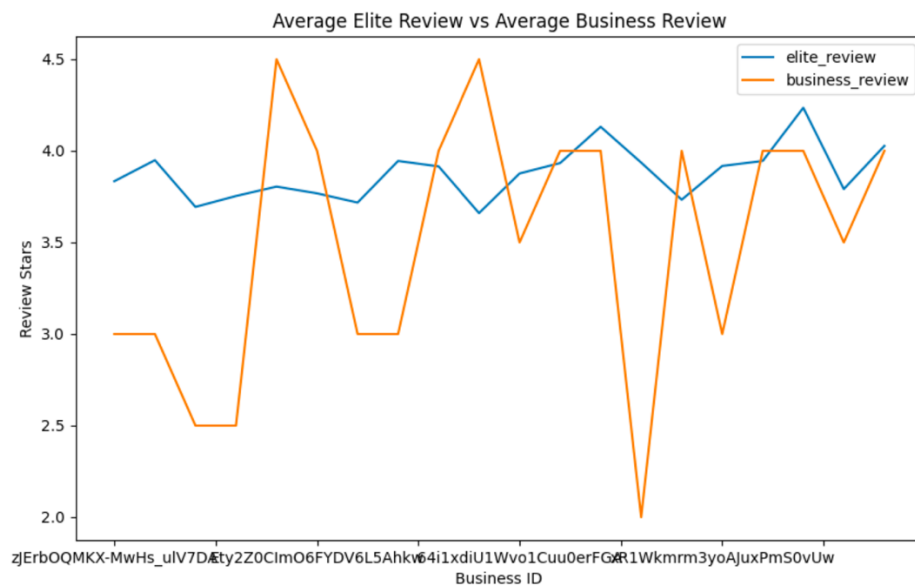
The screenshot displays the AWS Management Console interface for an Amazon EMR Notebook. The left sidebar shows the navigation menu with options like EMR Studio, Clusters, Notebooks, and Security configurations. The main content area shows the configuration for a Notebook named 'Project2-CIS9760', which is in a 'Ready' state. Key details include the Notebook ID, description, last modified time, and the IAM role 'EMR_Notebooks_DefaultRole'. The Notebook is associated with a cluster 'Project2-CIS9760' and a master instance. The Notebook location is specified as 's3://aws-emr-resources-482236323280-us-east-1/notebooks/'. The cluster status is 'Waiting', and the step logs are available at 's3://yelp-bucket-pr/'.

• 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 19,270 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.