**ADTA 5240-Harvesting, Storing and Retrieving Data**

Final Project

TOP 10 US Banks Stock Performance

2015-2023

Executive Summary

In this project, our main objective is to analyze the performance of the top 10 largest US bank stocks for the pre-COVID-19, COVID-19, and post-COVID-19 periods. Our focus will be on the five phases of the data life cycle using different big data tools such as Google Cloud Platform, Hadoop, Big Query, Hive, and Spark.

We obtained the top 10 US bank stock data from Yahoo Finance based on the New York Stock Exchange (NYSE) ticker names. Yahoo Finance has a well-structured data format for bank stocks. The daily stock transactions from January 2015 to November 2023 were obtained using Python API.The obtained data was divided into three:**pre-Covid** *(from January 5, 2015, to 21st January 2020****),* Covid** *(from January 22, 2020, to August 31, 2022),* and **post-Covid** *(from September 1, 2022, to November 24, 2023).*

We performed the Exploratory Data Analysis (EDA) using different Python libraries such as plotly, matplotlib, etc., and different EDA visualization graphs and charts such as histograms, density plots and heatmaps, boxplots, line charts, and time series charts. The EDA analysis helps us understand important trends and patterns. From the EDA, based on descriptive statistics of the data, we had a detailed understanding of how these top banks' stocks performed in the stock market throughout the eight years. Generally, the price movement of all the 10 banks tends to move in the same direction throughout the eight years although some of the banks had minor changes from the others. We calculated the bank returns based on the adjusted close price of each bank. From the EDA, it was that all 10 banks' returns have a very strong to strong correlation. The correlation between Bank of America and JP Morgan Chase(JPM) was very strong with a coefficient of 0.91. In the pre-COVID-19 period, Goldman Sachs(GS) had the highest return of about 9.54% while Capital One(COF) had the lowest return of about -13.12%. In the COVID-19 period, Capital One(COF) had the highest day return of about 18.76% and the lowest day return of -23.87%.In post-Covid-19, Capital One had the highest day return of about 11.78% and Truist Bank had the lowest day return of -16.99% during this period. The volatility plots showed that the most volatile stock in pre-COVID-19 was Capital One Bank, Toronto-Dominion Bank(TD) was the most volatile in the COVID-19 period while the most volatile in post-COVID-19 were Toronto-Dominion(TD) Bank and Capital One Bank.

For this project, we used different big data technologies below:

1. Open Refine:

We used Open Refine to clean the data by dropping the columns and rows, also we used Python for this process as an alternative.

2. Google Cloud Platform(GCP):

We used the GCP cloud storage for bucket creation and Hadoop clusters. We uploaded the csv file containing the bank's Adjusted Close Price to the GCP.

3. BigQuery:

We used BigQuery to get the external files to the GCP.

4. Hive:

We used the GCP DataProc clusters to load the data in the Hive Hadoop Ecosystem so that it is more accessible for running SQL queries. The completion time was 6.3 seconds.

5. Spark:

We used Spark for large-scale data processing and running queries with a completion time of 5.99 seconds. It showed that the completion time was faster than the time in Hive.

Submission by :

Aditya Keshavapatnam

Aeshala Rudra Tapaswi

Ezeanyagu Ifeanyichukwu

Harsha Vardhan Chavvakula

Zakiya Al Jabri