Faculty of Computer& Informatics.
Department of Artificial Intelligence
& Data science
Department of Software Engineering

Stock Market Data Analysis using Big Data Techniques

Junior Project

Prepared by

Abdullah Khadem Aljame

Ali Hussain Alaswad

Supervised by

Dr.Eng.Mouhib Alnoukari

Eng.Anas Abdulaziz

Academic Year

2023-2024

SUPERVISOR CERTIFICATION

I Certify that the preparation of this project entitled…… , prepared by ……, was made under my supervision at Faculty of Computer, Informatics Engineering & Artificial Intelligence in partial Fulfillment of the Requirements for the Degree of Bachelors of Software and Information System Engineering.

Name: …… Signature: …… Date: ……

الهدف من المشروع

يلعب سوق الأوراق المالية دورًا محوريًا في الاقتصاد العالمي، حيث يعمل كمنصة لشراء وبيع الأوراق المالية، مثل الأسهم والسندات والسلع.

أحد الجوانب الأساسية لعمليات سوق الأوراق المالية هو تحليل كميات هائلة من البيانات المالية. فهو يمكّن المستثمرين والمؤسسات المالية من تحديد الأنماط والاتجاهات والعلاقات المتبادلة داخل السوق، مما يساعد على اتخاذ قرارات استثمارية فعالة.

كما يساعد تحليل البيانات في تقييم وإدارة المخاطر مما يسمح بتخفيف الخسائر المالية المحتملة .

ولا يمكن أن تحدث هذه العملية دون الاعتماد على تقنيات البيانات الضخمة، حيث تعمل تقنيات البيانات الضخمة على تسهيل معالجة وتحليل كميات هائلة من البيانات المالية في الوقت الحقيقي، مما يتيح اتخاذ القرار السريعة واستخلاص رؤية قيمة للعثور على أفضل أداء بأقل تكلفة تقنية.

ولذلك تم العمل في هذا المشروع على تهيئة بيئة حقيقية على سيرفرين للجامعة السورية الخاصة للعمل على تحليل البيانات في سوق الأوراق المالية حيث تم تثبيت بيئة Hadoop و العديد من تقنيات البيانات الضخمة ك Hive و Power BI ،كما تم التركيز في التحليلات على مؤشر Nifty 100، حيث يوفر هذا المؤشر نظرة شاملة لسوق الأوراق المالية الهندي، ويشمل الشركات من مختلف القطاعات، ويستخدم على نطاق واسع من قبل المستثمرين ومديري الصناديق كمعيار لتقييم أداء السوق وبناء المحافظ الاستثمارية ،لذلك تم انشاء عدد من لوحات القيادة من اجل مساعدة المستثمرين على ايجاد افضل رؤية لمساعدتهم في تقييم وإدارة المخاطر مما يسمح بتخفيف الخسائر المالية المحتملة ،بالإضافة لتوضيح فوائد استخدام أنظمة البيانات الضخمة.

Project Objectives

The stock market plays a pivotal role in the global economy, serving as a platform for buying and selling securities, such as stocks, bonds, and commodities.

One of the fundamental aspects of stock market operations is the analysis of vast amounts of financial data. It enables investors and financial institutions to identify patterns, trends and interrelationships within the market, helping to make effective investment decisions.

Data analysis also helps in assessing and managing risks, allowing potential financial losses to be mitigated.

This process cannot occur without relying on big data technologies, as big data technologies facilitate the processing and analysis of huge amounts of financial data in real time, allowing quick decision-making and deriving valuable insight to find the best performance at the lowest technical cost.

Therefore, work was done in this project to create a real environment on two servers for the Syrian Private University to work on analyzing data in the stock market, where the Hadoop environment and many big data technologies such as Hive and Power BI were installed, and the analysis was also focused on the Nifty 100 index. This index provides a comprehensive view of the Indian stock market, includes companies from various sectors, and is widely used by investors and fund managers as a standard for evaluating market performance and building investment portfolios. Therefore, a number of dashboards have been created in order to help investors find the best vision. To help them evaluate and manage risks, which allows mitigating potential financial losses, in addition to explaining the benefits of using big data systems.

Content list

Figures List

Abbreviations List

Abbreviations	Definitions
HDFS	Hadoop Distributed File System
YARN	Yet Another Resource Negotiator
NSE	National Stock Exchange
OHLCV	Open, High, Low, Close, Volume
ELT	Extract, Load, Transform
SMA	Simple Moving Average
EMA	Exponential Moving Average
ROC	Rate Of Change

Chapter 1: Project Introduction

In this chapter, work was done to give an overview of the project, along with the project problem, its goal, and the method for solving the project problems

1.1. Stock Market

1.1.1. Overview

A stock market is a market where securities assets such as stocks, bonds and derivatives are traded and sold. These shares are traded among investors for the purpose of making profit by buying them at a low price and selling them at a low price. Stock exchanges are a primary place for trading securities, providing an infrastructure for trading. The stock market also provides an opportunity for companies to raise money and finance their activities by selling or issuing shares. The stock market is an important part of the local and global economies, helping finance companies and encouraging overall investment.

1.1.2. Stock Market Analysis

Analyzing stock market data can provide an important overview of market movement, price trends and the performance of various assets. Analysis of securities data usually involves the use of several tools and techniques such as technical indicators, trend analyses, and fundamental analysis. Data analysis is based on numbers and statistics related to securities and market activities, which helps in making informed investment decisions.

Analysis of stock market data may include quantitative studies of recurring patterns in prices and trading volume, which can help identify future investment opportunities. Artificial intelligence and machine learning techniques can also be used to analyze data and derive investment rules and strategies.

By analyzing stock market data, analysts can identify prevailing patterns, predict market changes, determine investment risks, estimate asset performance, and discover promising investment opportunities. Good analysis of stock market data is one of the essential tools for investors and traders in making informed decisions and achieving desired returns.

1.1.3 Benefits of analyzing stock market data

Stock market analysis offers a multitude of benefits for investors, traders, and financial professionals.

Some of advantages stock market analysis:

1. Informed Investment Decisions: By conducting thorough analysis, investors can make more informed decisions about buying, selling, or holding stocks. Understanding market trends, company fundamentals.

2. Risk Management: Analysis allows investors to assess and manage risk effectively. By evaluating factors such as volatility and correlations.

3. Value Identification: Analysis helps in identifying stocks that may be undervalued or overvalued based on their fundamental metrics, relative to their market price.

4. Portfolio Optimization: Analyzing the stock market enables investors to optimize their portfolios by diversifying across different asset classes, industries, and regions.

5. Understanding Market Trends: Stock market analysis provides insights into broader market trends, helping investors respond to changing market dynamics and adjust their strategies accordingly.

1.1.4. Nifty 100

The Nifty 100 index tracks the performance of the top 100 companies listed on the National Stock Exchange (NSE) of India. This broader index includes constituents of the Nifty 50 as well as an additional 50 companies, providing a more comprehensive view of the Indian stock market. The Nifty 100 index offers exposure to a diversified set of stocks across various sectors, making it a valuable benchmark for investors and fund managers. It serves as a barometer of the Indian equity market, allowing for a broader assessment of market performance and trends. The inclusion of additional companies in the Nifty 100 offers investors a more extensive representation of the Indian stock market compared to the Nifty 50, allowing for a broader perspective on market dynamics and diversification opportunities. (Nifty Indices, 2023)

1.2 Big Data

1.2.1. Overview

Big data refers to the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it's not the amount of data that matters. It's what organizations do with the data that counts. Big data can be analyzed for insights that lead to better decisions and strategic business moves. (Tom White , 2015)

1.2.2. Concept of Big Data

The three Vs that categorize the concept of Big Data:

1. Volume: Big data involves large volumes of data. This might be terabytes, petabytes, or even exabytes of data.

2. Velocity: Big data is often created in real time. For example, social media updates, sensor data, and other sources continuously generate new pieces of data.

3. Variety: Big data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, emails, videos, audios, stock ticker data, and financial transactions.

We can include additional Vs, such as Veracity (the quality and reliability of the data) and Value (the ability to extract meaningful insights and value from the data).

Harnessing big data allows organizations to gain insights from new data sources that were previously untapped. This leads to better decisions, operational efficiency, cost reductions, and the development of new products and services.

1.3. Bringing the Problem

In the past few years, data scientists have noticed a significant increase in the amount of data added to companies annually, with the inability of traditional analysis techniques to find the real benefit that can be obtained from this data, in addition to their inability to keep pace with the increase in the quantity and speed of data, and the diversity of its sources and forms, in addition to The problem of expensive technical devices that do not accept future expansion and do not enable developers to work on improving the system.

Therefore, data scientists worked to develop a system to solve big data problems while applying the idea of distributed systems to make the system more flexible and scalable.

1.4. Project Goals

- Build a Real Big Data platform in SPU
- Analyzing Data using Big Data Techniques
- Make the environment ready to receive future Big Data projects
- Work on implementing financial analysis for financial companies
- Build a working environment for a Big Data system while applying distributed systems concepts

Therefore, this project worked on creating a real environment to work on analyzing data in the stock market using big data techniques, with a focus on the Nifty 100 index

1.5. Big Data techniques in Stock market analysis

The use of big data technology in analyzing stock market data offers several key benefits.

Firstly, big data technology allows for the processing and analysis of large volumes of data from diverse sources, such as market transactions, news articles, social media sentiment, and economic indicators. This comprehensive analysis enables investors to gain a more holistic understanding of market trends and make more informed investment decisions.

Secondly, big data analytics can uncover valuable insights and patterns that might have been missed with traditional analysis methods. By applying advanced algorithms and machine learning techniques to large datasets, investors can identify correlations, trends, and anomalies that could potentially lead to profitable trading strategies.

Additionally, big data technology can facilitate real-time analysis, allowing investors to respond quickly to market developments and capitalize on emerging opportunities. This agility is crucial in fast-paced and dynamic financial markets where timely decision-making is essential.

Lastly, big data analytics can help investors manage risks by identifying potential market downturns, predicting stock price movements, and improving portfolio diversification strategies. By mapping and analyzing historical data, investors can gain valuable intelligence on past market behavior and use it to inform their risk management strategies.

1.6. Hadoop

In order to solve the problems mentioned in this project, work was done on building a system that runs on Hadoop (Aril Maheshwari, 2019)

1.6.1. Overview

Hadoop is an open-source framework that enables the processing and storage of large datasets across a cluster of computers. It provides a distributed computing environment that allows for the efficient processing of big data (Alex Holmes, 2012)

1.6.2. Hadoop Component

Hadoop consists of several core components that work together to enable its efficient processing and storage of large datasets. These components include:

1. Hadoop Distributed File System (HDFS): This is the primary storage system in Hadoop. It is designed to store huge amounts of data across multiple machines in a distributed and fault-tolerant manner.

2. Yet Another Resource Negotiator (YARN): YARN is the cluster management layer of Hadoop. It is responsible for allocating resources and managing the execution of applications on the cluster. YARN allows different types of applications, such as MapReduce, Spark, and Hive, to run on Hadoop.

3. MapReduce: This is the programming model that enables parallel processing of large datasets in Hadoop. MapReduce breaks down tasks into smaller sub-tasks, distributes them across the cluster, and then combines the results.

4. Hive: Hive is a data warehousing infrastructure built on top of Hadoop. It provides a SQL-like query language called HiveQL, which allows users to query and analyze large datasets stored in Hadoop. (Hanish Bansal & shrey Mehrota)

For data visualization, work was done on Apache Superset

Apache Superset is an open source business intelligence (BI) tool that provides a platform for data exploration and visualization, and is designed to facilitate data analysis and reporting. Apache Superset integrates with Hadoop to extract, transform, and visualize data stored in Hadoop. So they work together to provide a comprehensive data analysis solution.

1.6.3. Benefit of using Hadoop

Hadoop provides a cost-effective solution for handling large volumes of data by leveraging parallel processing across a cluster of commodity hardware. It allows organizations to store, process, and

analyze vast amounts of data in a distributed and scalable manner. Hadoop has become a fundamental technology for big data analytics and supports a wide range of applications, including stock market analysis, by providing a robust and scalable infrastructure.

1.7. Big Data Analytics

There are many big data analysis methods and techniques. Among these four popular methods are:

1.7.1. Exploratory analysis:

It is an exploratory method based on exploring and examining big data with the aim of discovering hidden patterns and connections in it. This is done by using data visualization, statistics, and machine learning tools to better understand the data. With this method, we can extract new insights and discover patterns and connections that improve our understanding of the data.

1.7.2. Predictive analysis:

This method relies on the use of big data, machine learning techniques, and statistical modeling to anticipate and predict future events and outcomes. This is done by analyzing past data and arriving at models that predict the future with the aim of achieving a deep understanding of the data and extracting laws and rules that affect future results.

1.7.3. Descriptive analysis:

Expresses how to summarize, describe, and understand big data using the concepts of descriptive statistics, graphical visualization, and statistical summarization. It aims to provide a comprehensive overview of the data and understand variables and graphical representation. This is done by collecting, processing and

summarizing data in easy-to-understand ways such as graphs, charts and statistical frequencies.

1.7.4. Behavioral analysis:

This method relies on understanding the behavior of users or customers by analyzing big data about interactions and activities. This is done by using techniques such as graphic visualization and statistical clustering to discover common patterns and trends and achieve related goals with the aim of understanding users' behavior and improving their experience.

From these analyses, Predictive analysis and Descriptive analysis were applied

Chapter 2: Theoretical Study

There are many types of big data tools available to process and analyze large amounts of data, but as we mentioned previously and some study of some of the tools used in building a big data system, it turns out that Hadoop is one of the best tools that can be used to build a real system with a medium-state server, as it Hadoop is an open source software framework that allows distributed processing of large data sets across clusters of computers. This means it is highly scalable and fault-tolerant, making it ideal for processing big data.

In a previous study entitled “Dynamic Interaction Between Nifty 50 and Nifty Sectoral Indices: An Empirical Study on Indian Stock Indices” work was done on statistical analyzes and comparison of the rest of the stock market indicators without working on conducting analyzes on the basis of companies. (K Ramya & Bhuvaneshwari D, 2020)

In another study entitled “A Scalable System for Textual Analysis for Stock Market Prediction” Work was done on forecast analysis using Apache Mahout in the stock market on one of the global indices without working on building a data warehouse to do better analyzes (Roy Guanyu Lin & Tzu-Chieh Tsai, 2014)

The research paper, titled "Analyze Stock Data Using Apache Hive," focuses on using Big Data tools, specifically Apache Hive, to predict stock market trends, The study involves analyzing historical data from the New York Stock Exchange (NYSE) to predict future stock market behavior. (Raswitha Bandi, T.Shravani Reddy, K.Nikhila, & Mohd Abdul Javeed, 2018)

The paper is titled “Parallelizing K-means with Hadoop/Mahout for Big Data Analytics” talks about applying the K-means algorithm in a big data system, but the data is different from the one we use in project. (Jianbin Cui, 2015)

Chapter 3: Project Data Set

3.1. Gather, Find and understand the data

The data was searched in many sources such as Yahoo Financ, Alpha Vantage, and Kaggle, and we chose this data because it provides an approximate view of how the project works on the available system. (Kaggle, 2022)

3.1.1. Data source

https://www.kaggle.com/datasets/debashis74017/stock-market-data-nifty-50-stocks-1-min-data

3.2. About data

3.2.1. Overview

This dataset contains historical daily prices for Nifty 100 stocks and indices currently trading on the Indian Stock Market.

3.2.2. Nifty 50 & Nifty 100

The NIFTY 50 index is a well-diversified 50 companies index reflecting overall market conditions. NIFTY 50 Index is computed using the free float market capitalization method.

The Nifty 100 index tracks the performance of the top 100 companies listed on the National Stock Exchange (NSE) of India. This broader index includes constituents of the Nifty 50 as well as an additional 50 companies, providing a more comprehensive view of the Indian stock market. The Nifty 100 index offers exposure to a diversified set of stocks across various sectors, making it a valuable benchmark for investors and fund managers. It serves as a barometer of the Indian equity market, allowing for a broader assessment of market performance and trends. The inclusion of additional companies in the Nifty 100 offers investors a more extensive representation of the Indian stock market compared to the

Nifty 50, allowing for a broader perspective on market dynamics and diversification opportunities.

Data samples are of 1-minute intervals and the availability of data is from Jan 2015 to Feb 2022.

3.2.3. Content of data

The whole dataset is around 66.2 GB (compressed here to 26 GB), and 100 stocks (Nifty 100 stocks) and 2 indices (Nifty 50 and Nifty Bank indices) are present in this dataset.

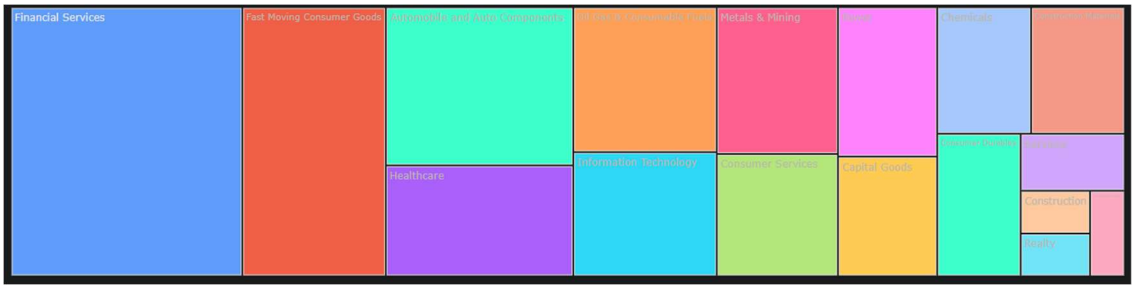

Figure 1-tree map for company By field

Figure 2-tree map for company By name

3.2.4. Details about each file

OHLCV (Open, High, Low, Close, and Volume) data

55 Technical indicator values

The columns contain different technical indicators related to financial market data. Below is a brief explanation of each:

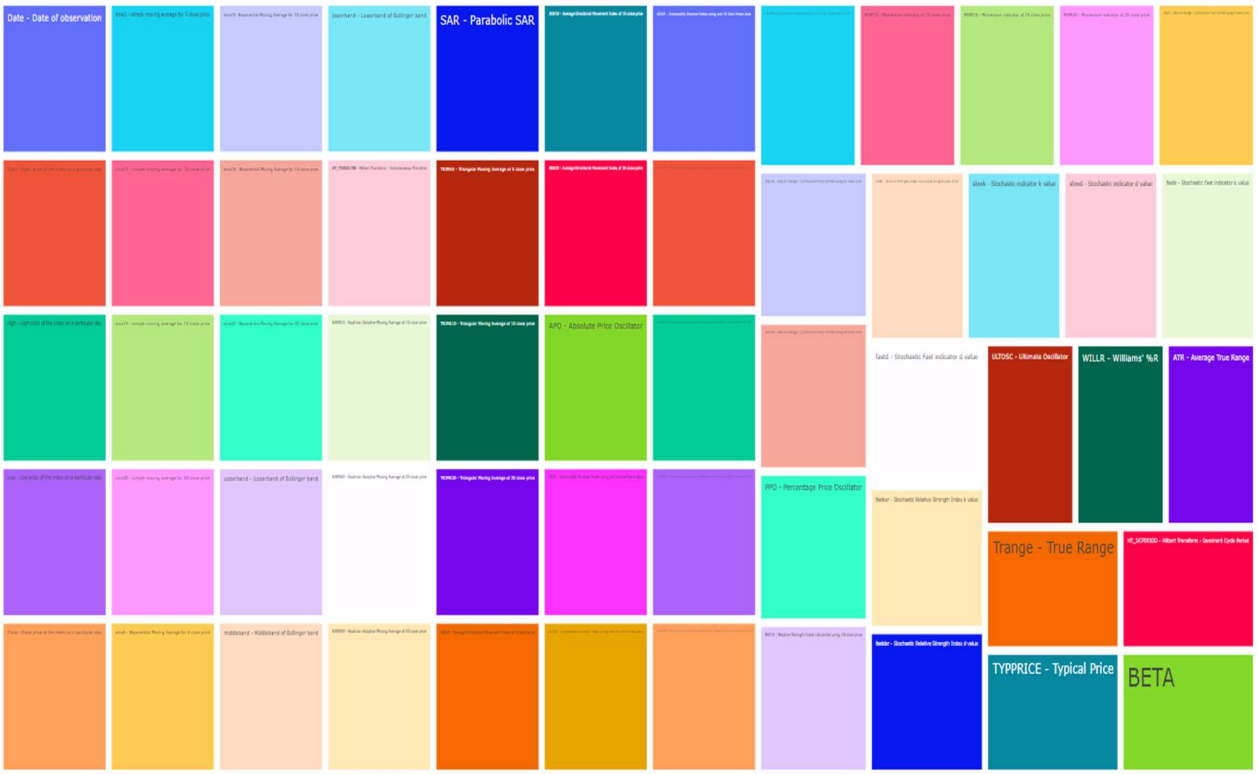

Figure 3-tree map for Technical indictor values

Chapter 4: Project Implementation

4.1 Building environment and tools used

As we mentioned in the introduction, we worked on preparing a system to work using Hadoop

Therefore, we worked on two servers for the university and we built the project on them from the beginning

4.1.1. Available servers & Operating system used

The project was built on two servers of the type each containing 16 GB of RAM, 16 CPUs, and 300 GB of storage.

Ubuntu Desktop 22.04 was used on both servers for Ubuntu's compatibility with Hadoop.

All steps for working and installing the system are in this link:

https://github.com/msoc10/SpuBigData

4.1.2. System architecture

After studying the available servers and the required system, work was done to build the system on the master-slave architecture.

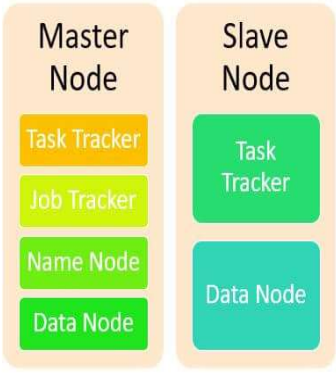

Figure 4-System architecture

Master-slave architecture is a design pattern that involves dividing a system into two main components: master and slave. In this architecture, the master component is responsible for controlling and coordinating the overall system, while the slave component performs specific tasks assigned by the master.

In analyzing stock market data, a master-slave structure can be used to distribute workload and improve performance. The main component can handle tasks such as data collection, processing, and analysis, while subcomponents can perform these tasks in parallel.

For example, a native component can fetch stock market data. It can then distribute the data to dependent components for further processing, such as calculating indicators, performing statistical analysis, or generating reports.

This distributed architecture can enhance scalability and fault tolerance. If the workload increases, more dependent components can be added to handle additional tasks. If one slave component fails, the master can assign its tasks to another available slave component without disrupting the entire system.

4.1.3. Hadoop cluster environment &configuration files

Hadoop environment and configuration files

We used Apache Hadoop 3.3.6 with all initialization files needed to run main components of Hadoop

(Hadoop-HDFS-YARN-Hive-SuperSet)

All configuration files with download steps at the link:

https://github.com/msoc10/SpuBigData

4.1.4. ELT process and Data warehouse build

1-Extract (download and Decompress) Data From Kaggle

2-Load the modify data in HDFS

3-Transformation By using Apache Hive:

Apply some transformation at the data:

- Split the "Date" column that have this format "2015-03-15 00:52:40" that include the date and time to date column and time column
- Add an "Index" column for each file to know the file and the name of the stock market
- Add "Quarter" Column to data

- Add new Column "Month", this column Extract from Date column (In Power BI)

Build Data Warehouse:

(1 fact table, 2 dimension, Star Schema)

- -The first Dimension contain (Index, Stock_names)
- -The Second Dimension contain (Quarters)
- -The fact table contain all measured data (61 columns)

The schema:

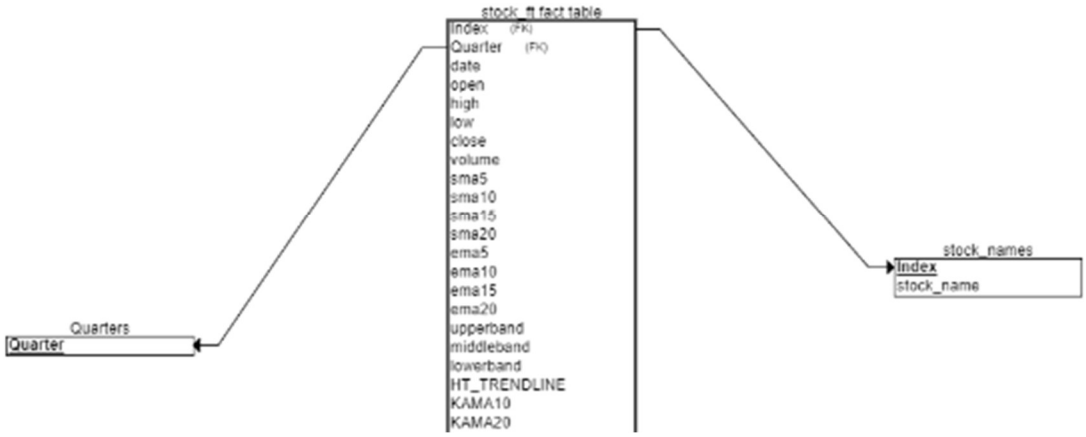

Figure 5-The schema of data

All steps and code of ETL process and data warehouse build are at the link:

https://github.com/msoc10/SpuBigData

4.1.5. Analysis & Dashboards:

Initially, some analyzes were done using Python to clarify the general view of the data

Simple visualize:

Figure 6-Simple visualize 1

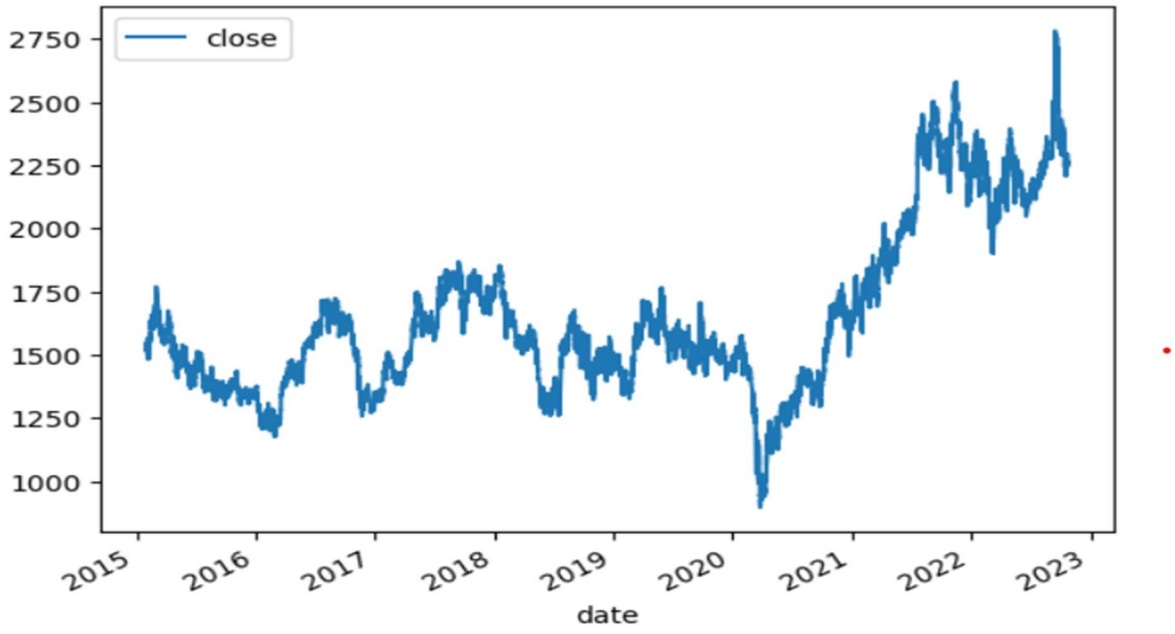

Figure 7-Simple visualize 2

1-Basic Descriptive Statistics: Calculate the mean, median, mode, minimum, and maximum values for each column. This will give you a general overview of the central tendency and range of values.

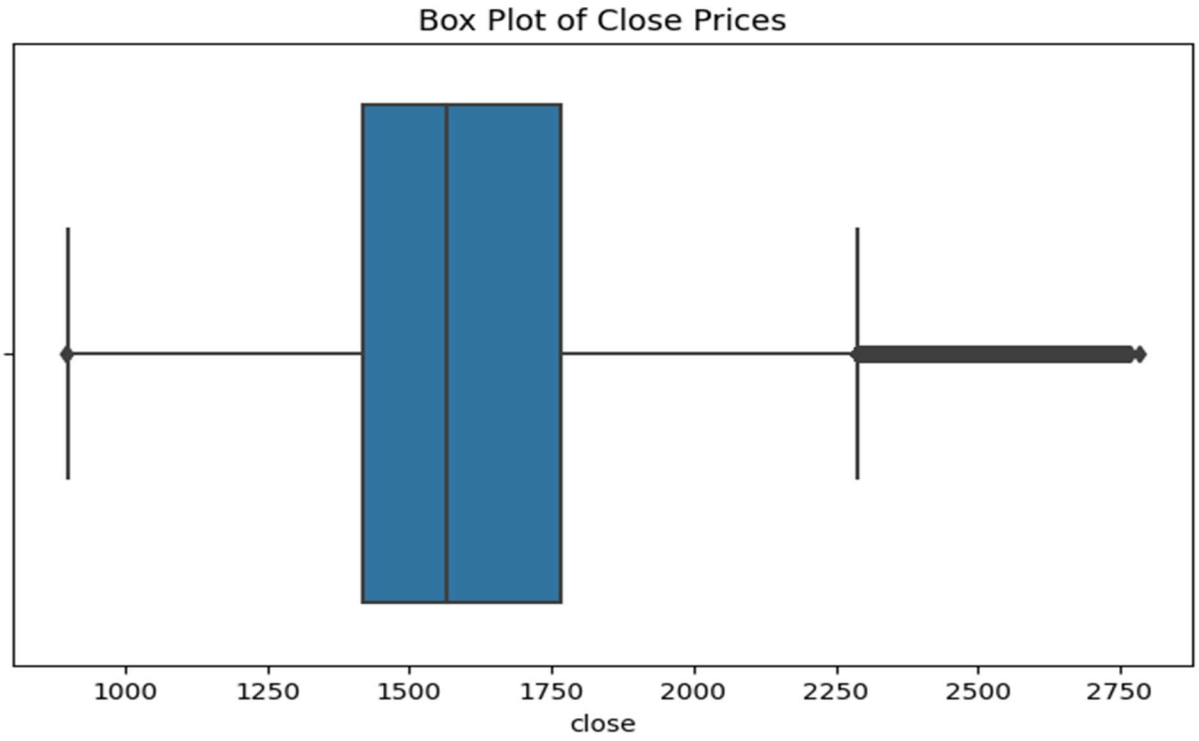

Figure 8-Box Plot of Close Prices

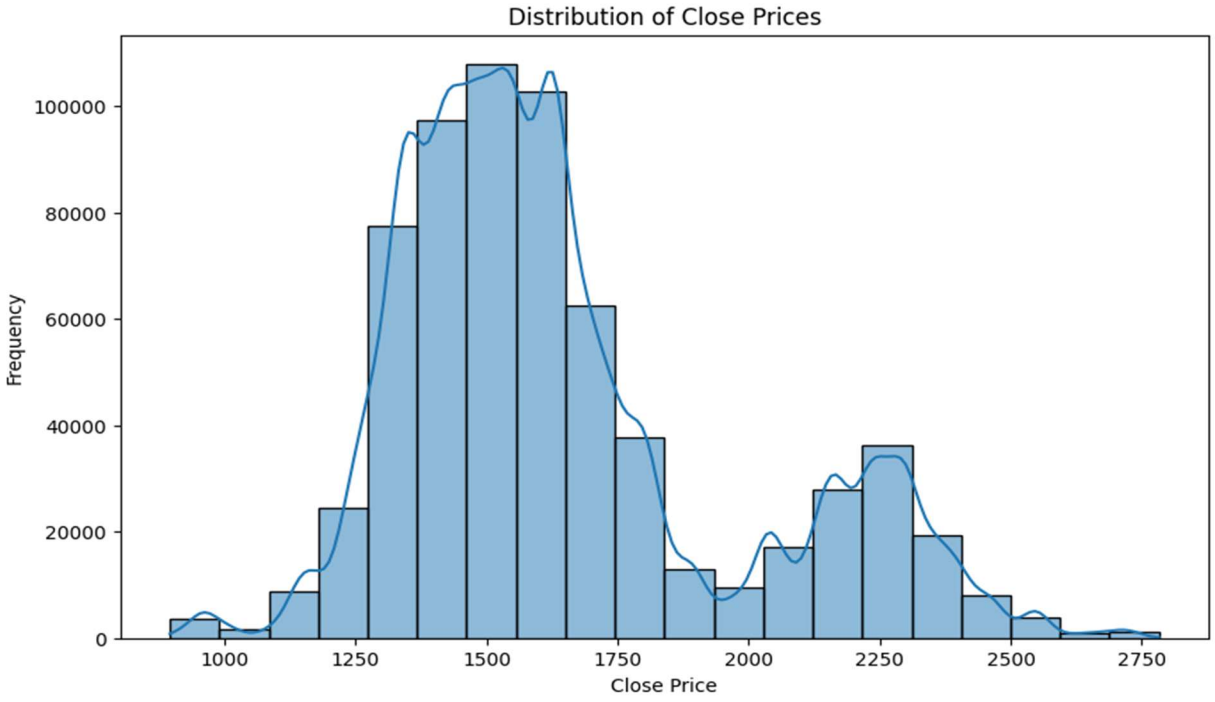

Figure 9-Distribution of Close Prices

The basics Descriptive statistics for close price and Volume:

Year	Close Mean	Close Median	Close Min	Close Max	Close Std	Volume Mean	Volume Median	Volume Min	Volume Max	Volume Std
2015	1447.3036461974552	1416.7	1301.1	1770.0	100.22217765200737	804.4324338066124	352.0	0.0	825479.0	3668.380516361473
2016	1459.6305087843416	1453.0	1175.9	1724.8	157.5215953475246	725.2871359509492	316.0	0.0	210112.0	2145.5224628037136
2017	1622.8521777084056	1659.25	1314.6	1869.65	153.06642862332475	1055.7428141981732	449.0	0.0	500892.0	3661.282795594651
2018	1514.281260082452	1526.55	1260.9	1854.1	134.4344369283616	1445.7640795841548	636.0	0.0	930263.0	5972.3665454859965
2019	1530.8125605887278	1530.35	1326.7	1766.25	85.08687735471864	2108.8020945125186	1083.0	0.0	324126.0	4588.097841293055
2020	1389.9351965125304	1403.0	896.5	1782.5	191.65594827392633	3484.402656685078	1851.5	0.0	586897.0	7400.378674141749
2021	2080.4188201513966	2045.3	1585.85	2581.45	255.6532407983641	1963.7455447189197	850.0	0.0	260083.0	4352.169306647161
2022	2230.5591151993617	2218.95	1901.1	2782.65	132.6511165454724	1768.718770484795	685.0	0.0	962709.0	5286.303846382271

Figure 10-basics Descriptive statistics for close price and Volume

The best selling for each Quarter from 2015 to 2022:

Year	Best-Selling Quarter	volume
2015	2015Q2	18069994
2016	2016Q3	16687525
2017	2017Q3	29382788
2018	2018Q4	39534071
2019	2019Q3	51242118
2020	2020Q4	104609456
2021	2021Q1	61945983
2022	2022Q3	52985210

Figure 11-best selling for each Quarter from 2015 to 2022

Basics Statistics for 2015:

Figure 12-Statistics for Each Quarter in 2015

Distribution for each quarter in 2015:

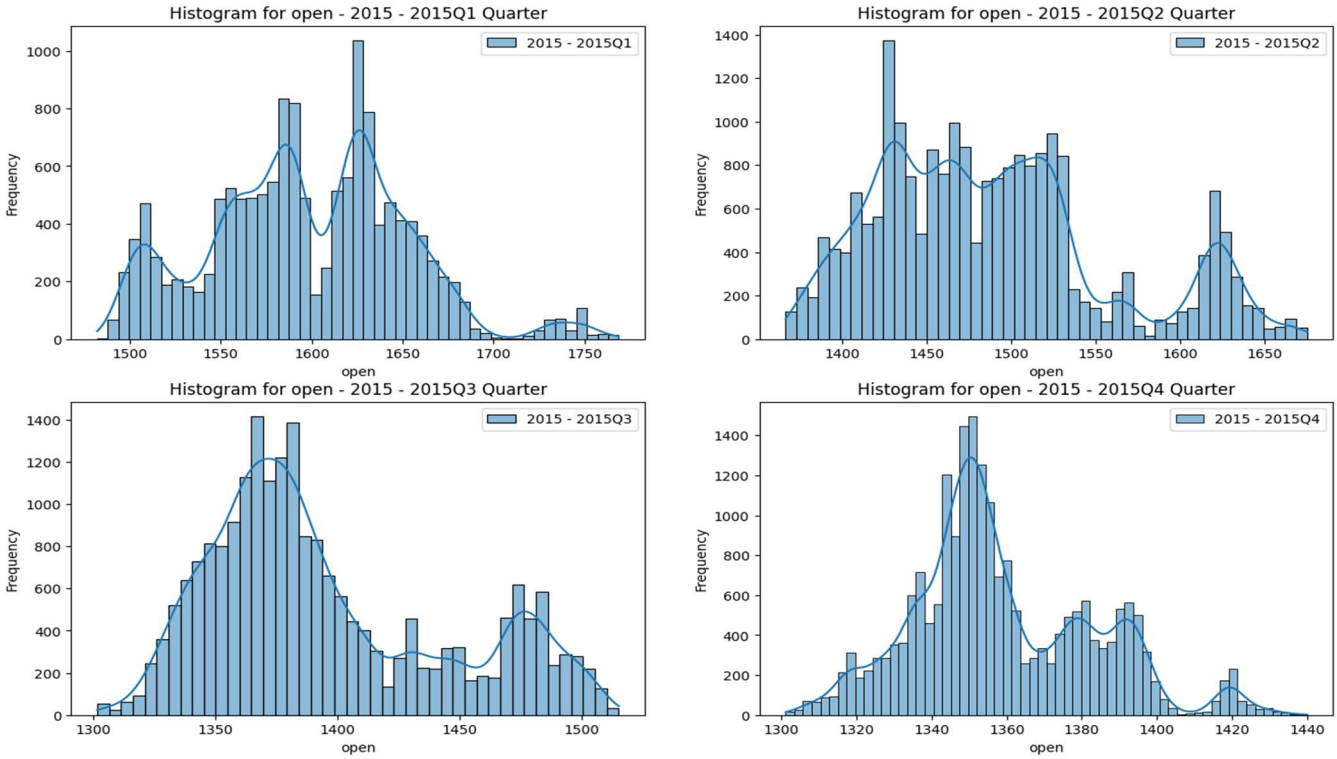

Figure 13-Distribution for each quarter in 2015

2- Trend Indicators:

Analyze the trend indicators (sma and ema columns) to identify trends in the stock prices over different time frames. You can calculate the rate of change (ROC) for these indicators to measure the percentage change in the moving averages.

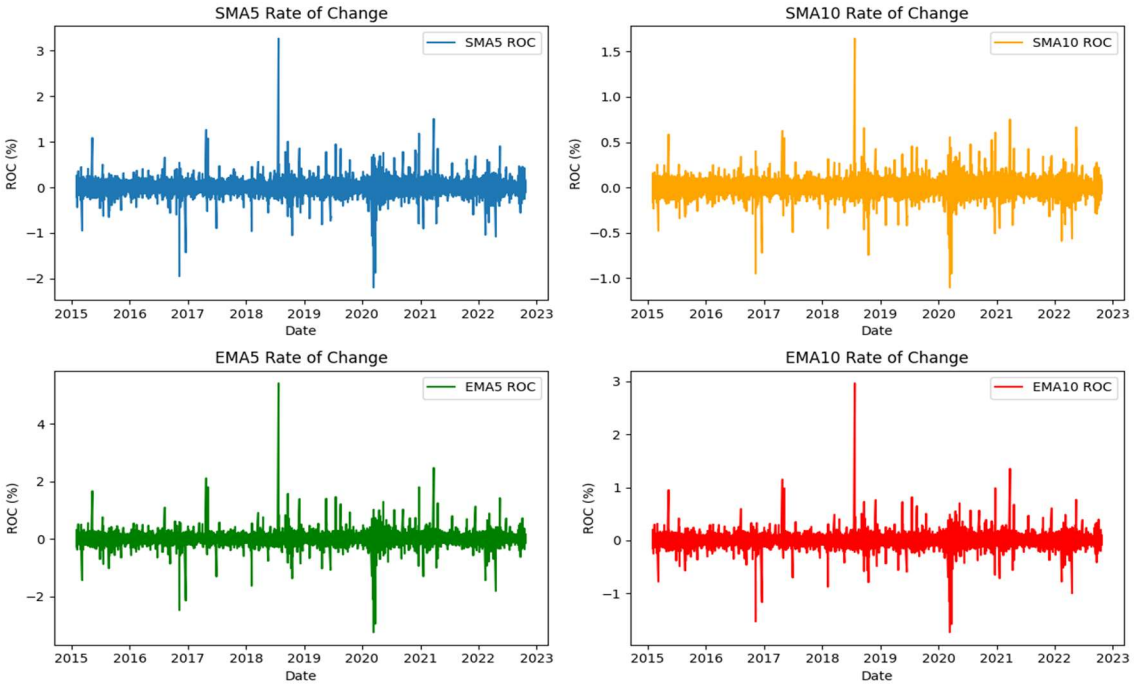

Figure 14-Trend Indicators

3- Bollinger Bands:

Explore the upperband, middleband, and lowerband columns to understand the volatility and potential reversal points in the stock prices

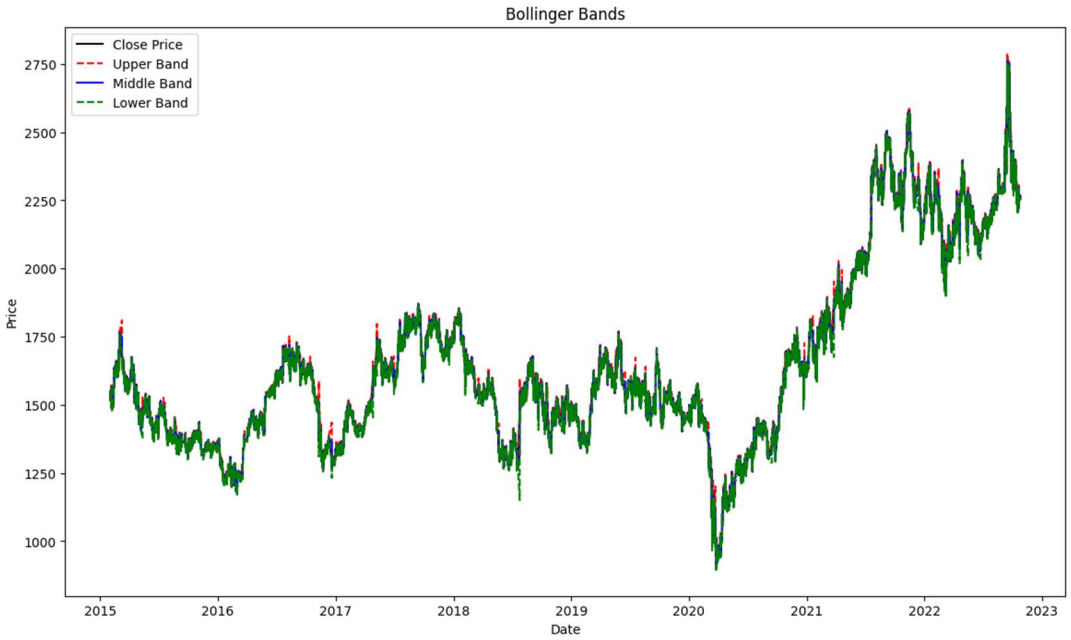

Figure 15-Bollinger Bands

4- Variability:

Compute the standard deviation and variance for columns like Open, Close, High, Low, and the moving averages (sma5, sma10, etc.). This helps you understand the dispersion or volatility in the data.

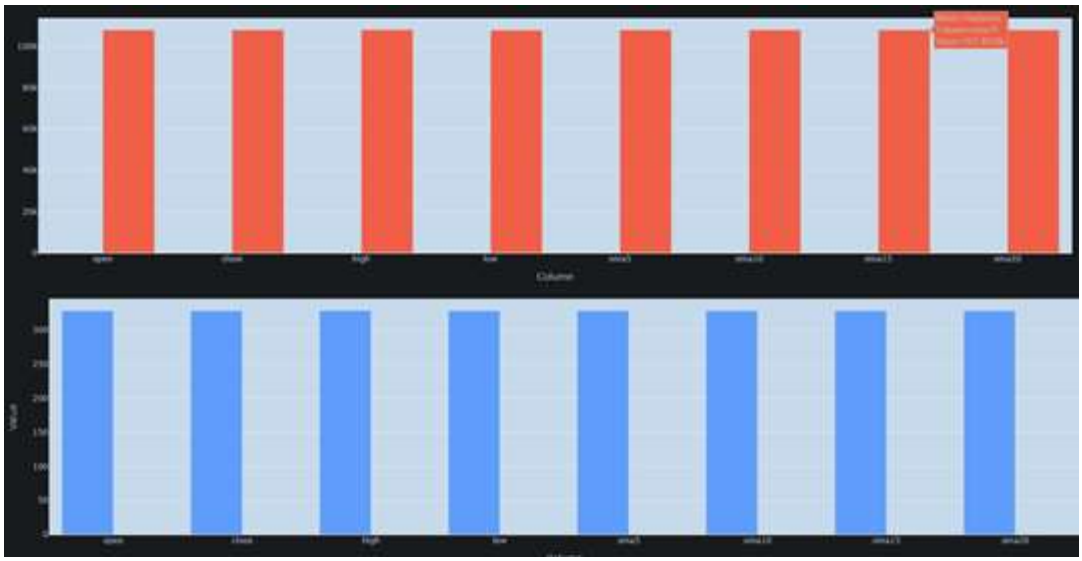

Figure 16-Variability

You can find all the analyzes code at this link:

https://github.com/msoc10/SpuBigData

Then, after building the data warehouse, new **Dashboards** were created for some of the following types of analyses:

Descriptive Analytics

We use the most popular columns to do descriptive statistics (open-close-high-low-volume)

Figure 17-Descriptive Analytics1

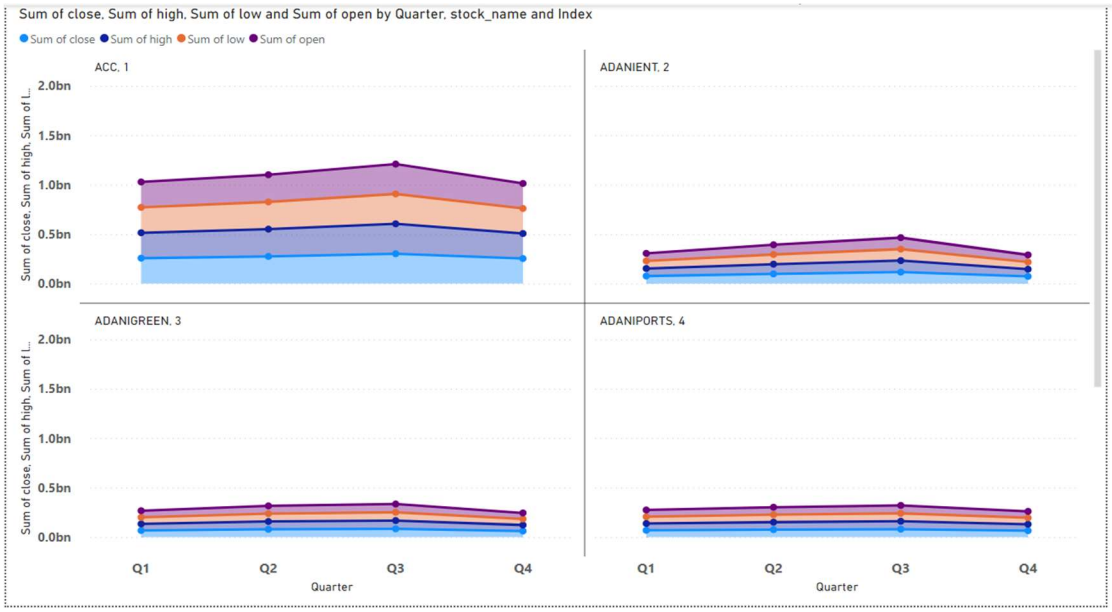

Figure 18-Descriptive Analytics 2

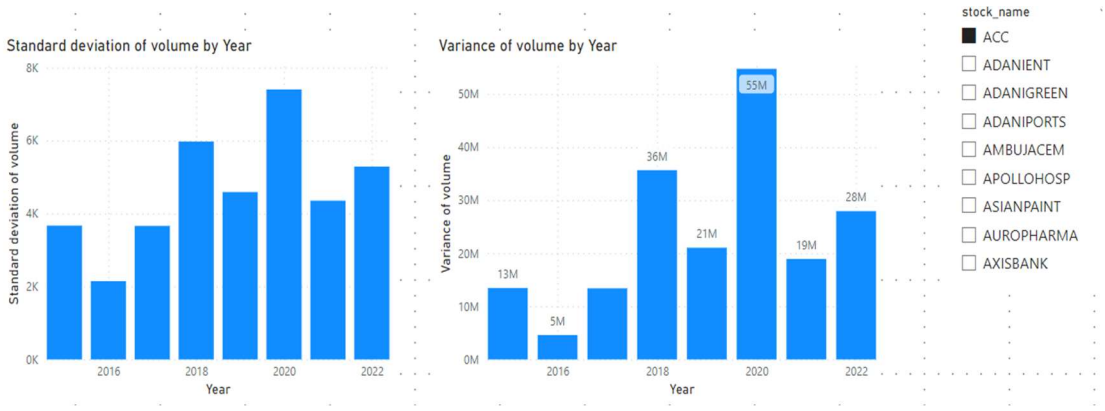

Figure 19-Descriptive Analytics 3

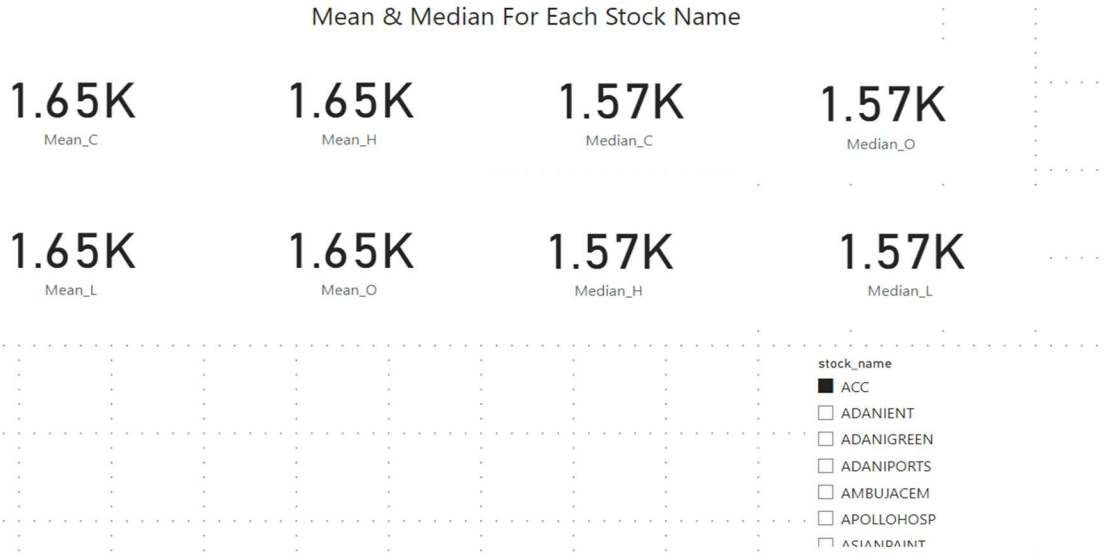

Figure 20-Descriptive Analytics 4

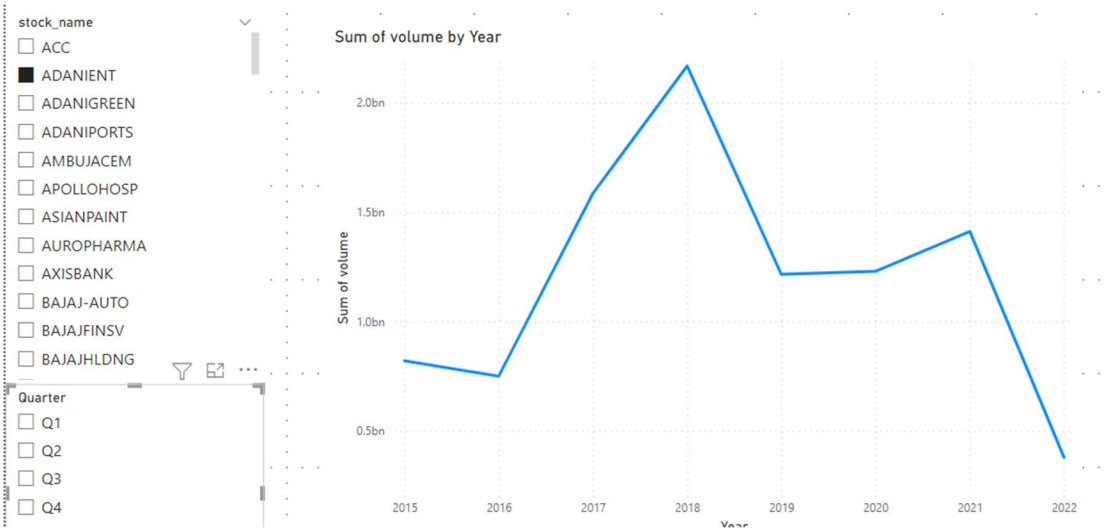

Figure 21-Descriptive Analytics 5

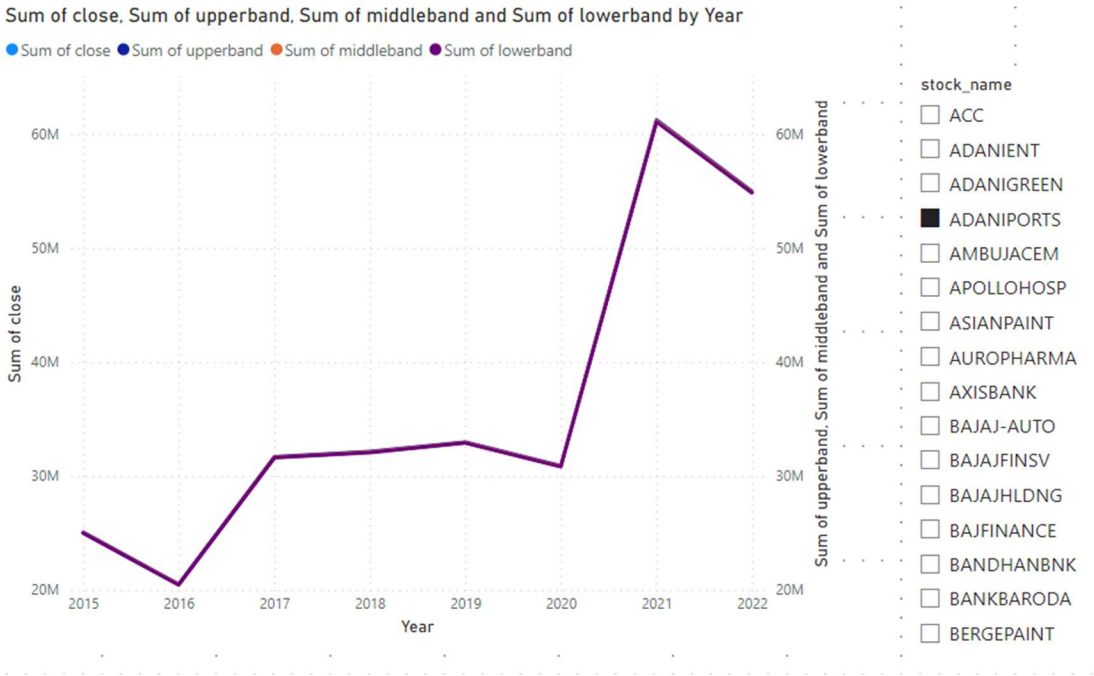

Figure 22-Descriptive Analytics 6

Predictive Analytics

For predictive we use:

(sma5-sma10-sma15-ema10-ema10-ema15) with Year

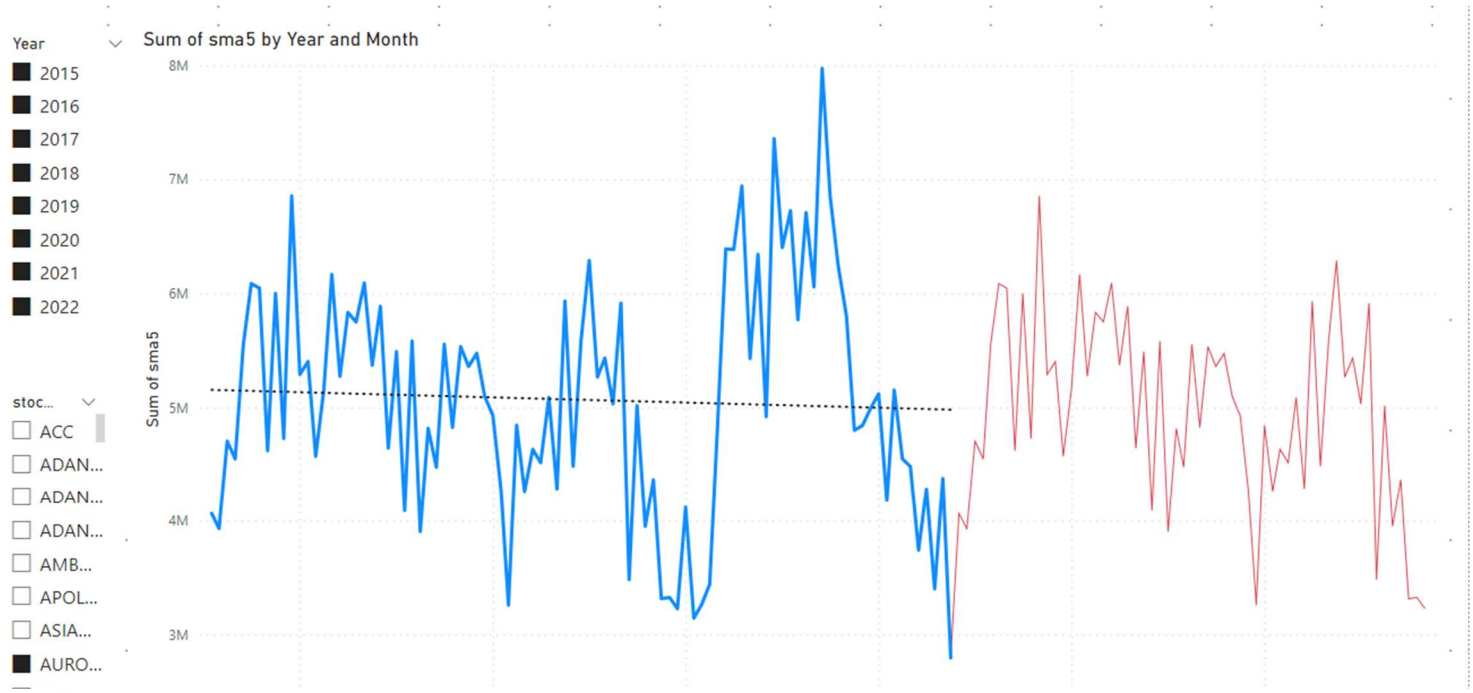

Figure 23-Predictive Analytics

4.1.6. Weekly Project Plan

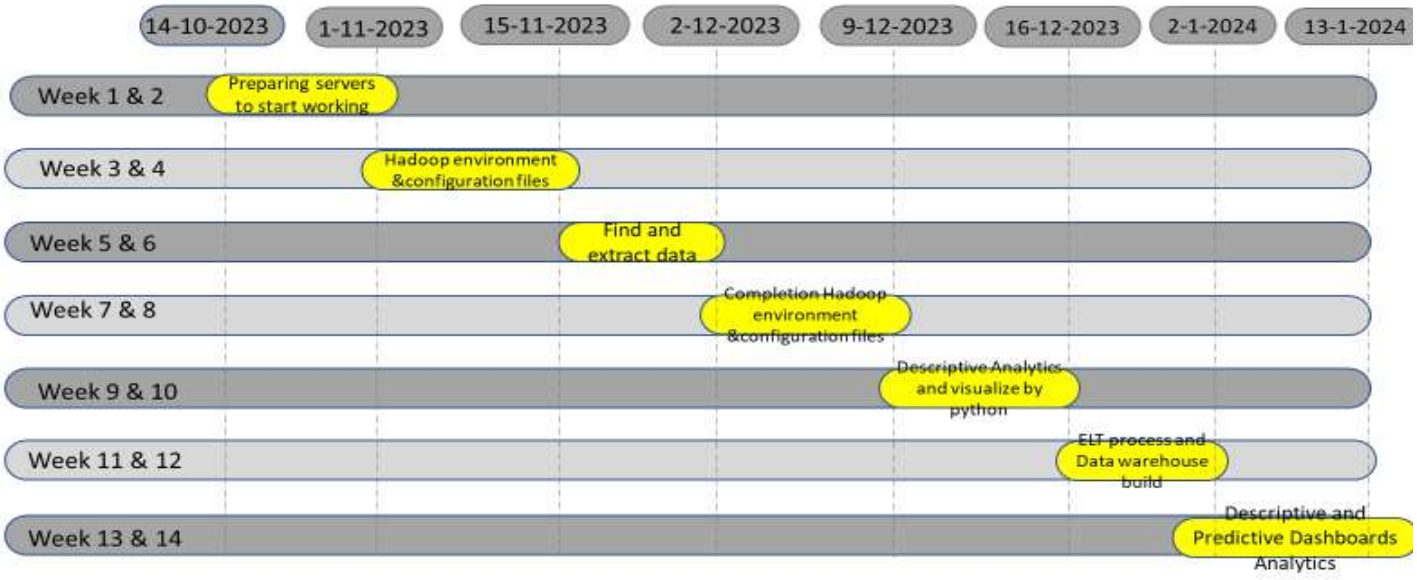

Figure 24-Weekly Project Plan

Week 1 & 2: Preparing servers to start working

Week 3 & 4: Hadoop environment &configuration files

Week 5 & 6: Find and extract data

Week 7: Completion Hadoop environment &configuration files

Week 8 & 9: Descriptive Analytics and visualize by python

Week 10 & 11: ELT process and Data warehouse build

Week 12 & 13: Descriptive and Predictive Dashboards Analytics

Chapter 5: Conclusion and Future Works

5.1. Conclusion

In conclusion, we worked in this project to building a big data system to solve the recurring problems related to data (volume - Variety- Velocity) by applying big data techniques using Hadoop on a real structure within the Syrian Private University in order to carry out many analyzes that cannot be To occur in traditional ways at work

In addition, the project seeks to create an environment for future projects related to big data

Working on the servers, starting from installing the operating system on them, to installing the Hadoop architecture, and from there to installing the Hive and Superset architecture in order to carry out analysis operations and build the data warehouse.

Then, work was done on the Nifty 100 index in order to test the system on large real data, as the data that was worked on contains approximately 70 million lines, which gives an overview of the system's ability to work to apply useful analyzes to companies without the need to purchase expensive devices.

5.2. Future Works

There are many things that can be developed after this project, including:

1. Perform other analyzes on a real-time basis using big data techniques
2. Add some additional servers to make the environment more powerful
3. An additional backup server can be added to make the service more reliable and fault tolerant
4. You can start analyzing new and different types of data, such as images, videos, and audio files, by working with non-relational databases (NoSQL).

Chapter 6: References

6.1. References

Alex Holmes. (2012). *Hadoop in Practice.*

Aril Maheshwari. (2019). *Big Data Made Accessible.*

Hanish Bansal, & shrey Mehrota. (n.d.). *Apache Hive Cookbook.*

Jianbin Cui. (2015). Parallelizing K-means with Hadoop/Mahout for Big Data Analytics. *Brunel University, UK*.

K Ramya, & Bhuvaneshwari D. (2020, July). Dynamic Interaction Between Nifty 50 and Nifty Sectoral Indices: An Empirical Study on Indian Stock Indices. *NMIMS*.

Kaggle. (2022). *kaggle*. Retrieved from https://www.kaggle.com/datasets/debashis74017/stock-market-data-nifty-50-stocks-1-min-data

Nifty Indices. (2023). *Nifty Indices*. Retrieved from Nifty Indices: http://www.niftyindices.com/

Raswitha Bandi, T.Shravani Reddy, K.Nikhila, & Mohd Abdul Javeed. (2018). Analyze Stock Data Using Apache Hive. *IT MLRIT, Dundigal*.

Roy Guanyu Lin, & Tzu-Chieh Tsai. (2014). Scalable System for Textual Analysis of Stock Market Prediction. *The Third International Conference on Data Analytics*.

Tom White . (2015). *Hadoop The Definitive Guide.*