Hardware and Software for Big Data

University of Naples Federico II

Stock Clustering Analysis Project

Professor: DR. GIANCARLO SPERLI

Group Members:

Seyyed Alireza Khoshsolat

Farshad Farahtaj

First semester of 2023-2024

Abstract: This comprehensive documentation outlines the stock clustering analysis project conducted by Seyyed Alireza Khoshsolat and Farshad Farahtaj as part of the Hardware and Software for Big Data course at the University of Naples Federico II. The project employs Apache Kafka, PySpark, and machine learning techniques to cluster stock data, aiming to facilitate better analysis and decision-making in the stock market. The document provides an in-depth exploration of each project component, from the initial installation to the interpretation of clustering results.

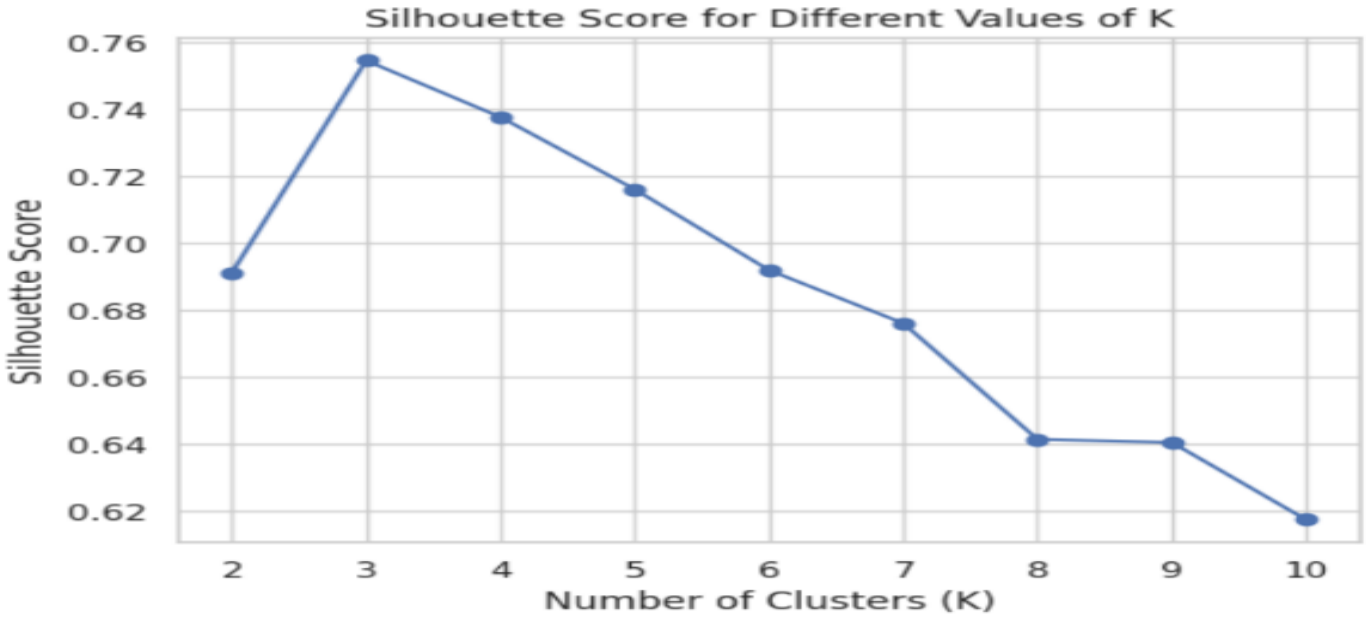
Introduction: Stock clustering analysis is a pivotal aspect of modern finance, aiding investors and analysts in understanding market patterns. This project focuses on leveraging cutting-edge technologies and methodologies to group stocks based on their price behavior. Through Apache Kafka, PySpark, and the K-Means clustering algorithm, the project endeavors to offer a systematic and insightful approach to stock analysis.

Project Workflow:

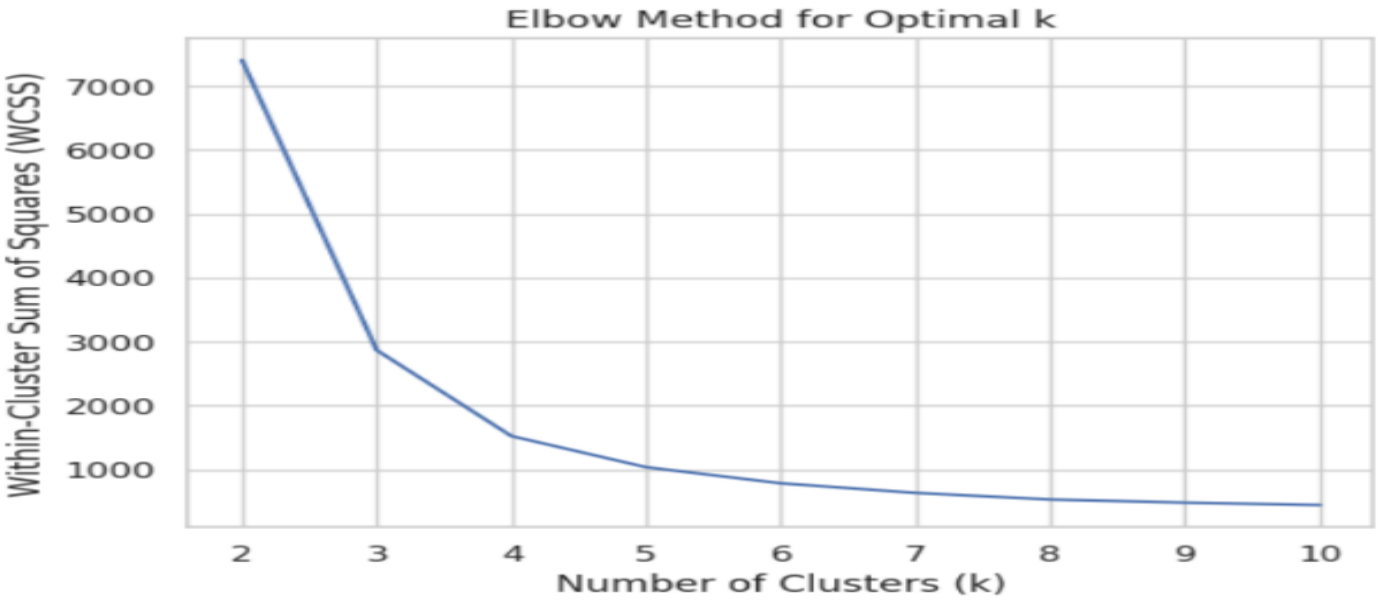
1. **Installation of Kafka and Yahoo Finance:**
    - The initial phase involves the installation of pivotal libraries, such as kafka-python and yfinance, to establish connections with Apache Kafka and extract stock data from Yahoo Finance.
2. **Kafka Setup and Data Streaming:**
    - Apache Kafka is meticulously downloaded, extracted, and initiated with Zookeeper and Kafka server configurations.
    - A curated list of stock tickers is selected, and relevant data is fetched from Yahoo Finance, dynamically streamed to designated Kafka topics.
3. **Exploration and Analysis in PySpark:**
    - The project leverages the robust capabilities of PySpark for distributed computing, facilitating comprehensive analysis of stock data.
    - Key PySpark functionalities include the seamless conversion of data from Pandas DataFrame to Spark DataFrame, correlation analysis unveiling relationships between stock features, and the preparation of data for subsequent clustering analysis.

4. Clustering Analysis with K-Means:

- The project's pivotal clustering algorithm is K-Means, chosen for its efficiency in grouping stocks based on similarities in opening, closing, high, and low prices.
- The optimal number of clusters (K) is meticulously determined through a dual methodology involving Silhouette and Elbow methods.
- The Silhouette method provides an insight into the quality of clustering for different values of K. A higher Silhouette score (closer to 1) signifies better-defined clusters (Plot 1). With Silhouette method we got two values for K that we can consider (K=3 or K= 4). The Elbow method, on the other hand, identifies the 'elbow' point where the within-cluster sum of squares (WCSS) begins to decrease at a slower rate, indicating an optimal K value (Plot 2). With Elbow method we got two values for K that we can consider (K=4 or K=5).



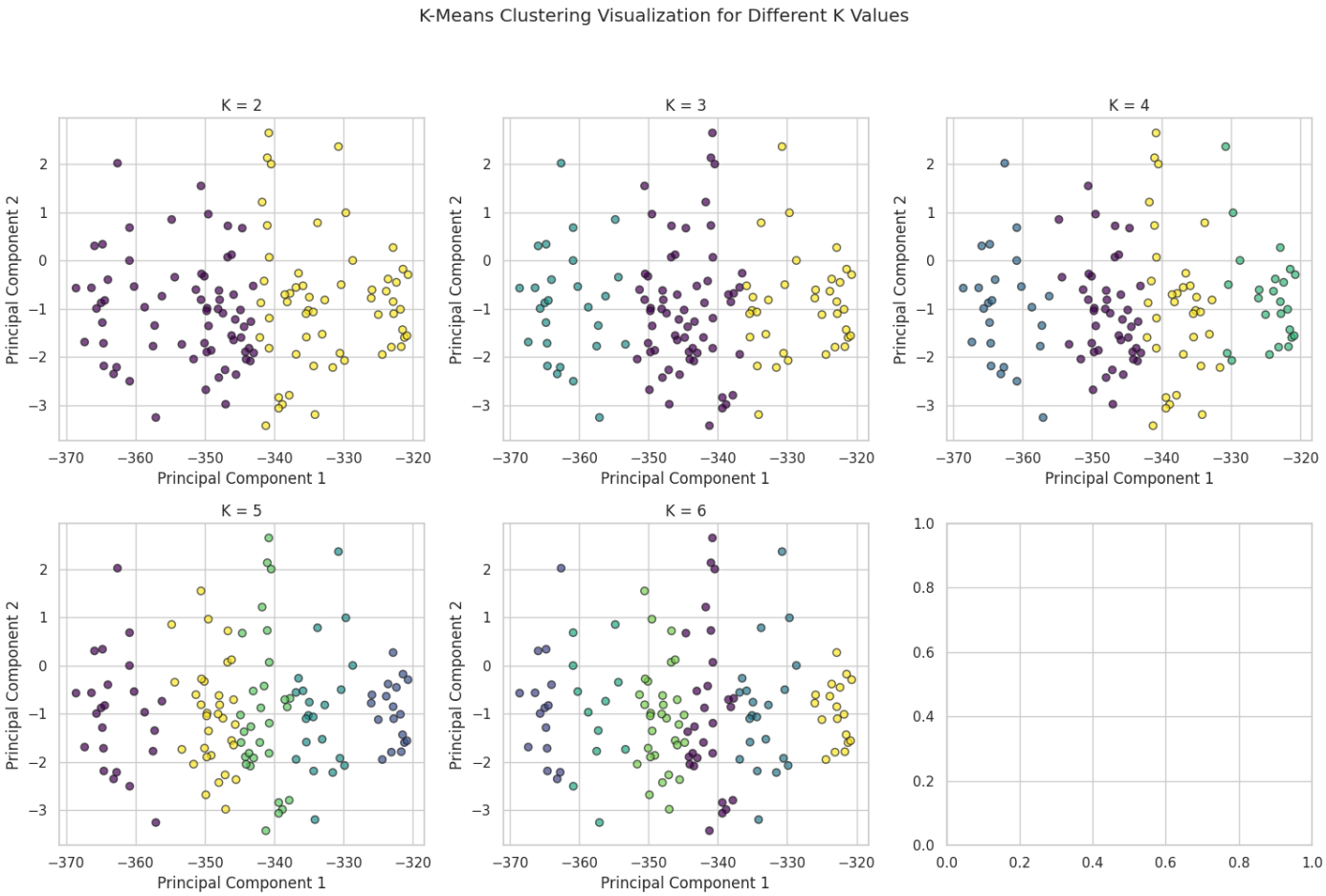
Plot 1
base on Silhoutte score we can consider k = 3 or k = 4



Plot 2
based on Elbow we can choose k=4 or k=5

5. PCA and Comparing Clusters with Different K Values:

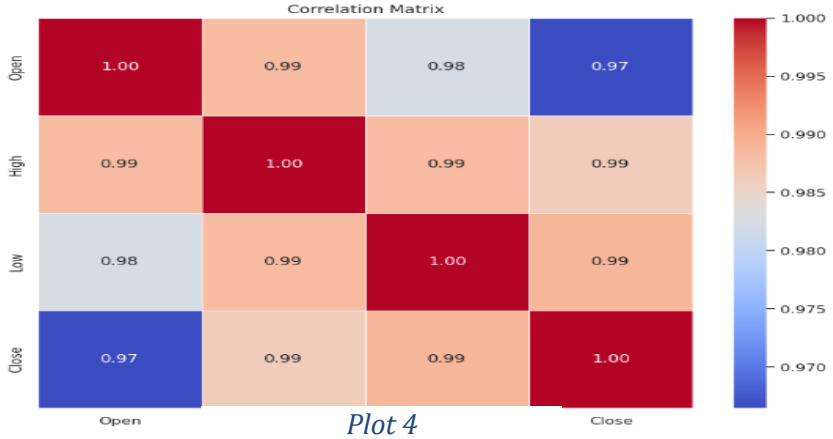
- Applied Principal Component Analysis (PCA) to visualize clusters for different K values between 2 and 6 (Plot 3).
- Conducted a detailed examination and comparison of resulting cluster plots to determine the optimal K value. Based on PCA plots and Silhouette and Elbow methods we determined that the best value for K would be 4 because it shows the best result clustering our stocks.



Plot 3

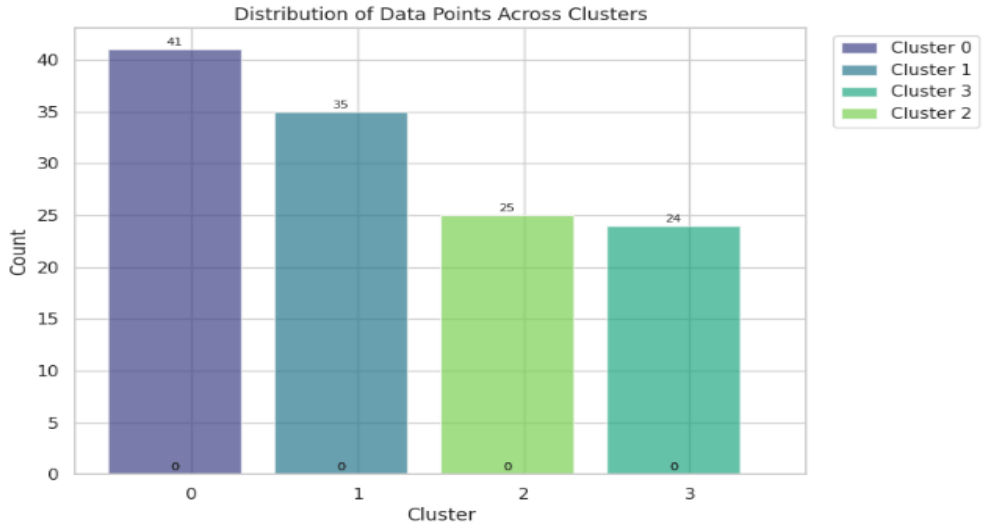
6. Visualization and Interpretation:

- The project places a strong emphasis on visualization for enhanced interpretation of results.
- A correlation matrix heatmap offers a visual representation of the relationships between various stock features (Plot 4).



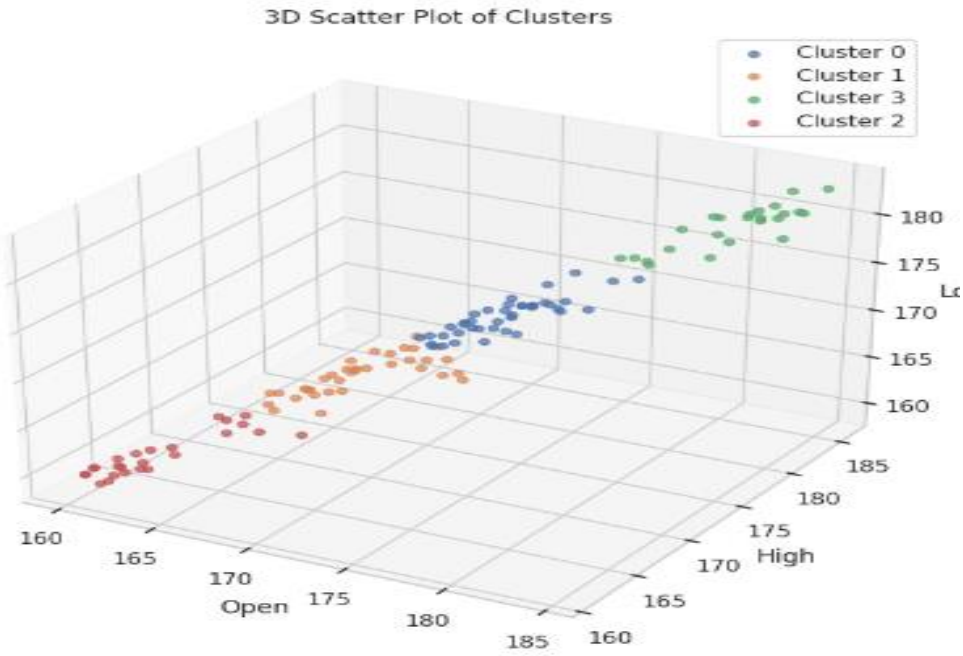
Plot 4

- Silhouette and Elbow plots contribute significantly to the selection of the optimal K value.
- Employed histograms to depict the distribution of data points across clusters (Plot 5).



Plot 5

- 3D scatter plots are crafted to provide an immersive and visually insightful representation of stock clusters in a three-dimensional space (Plot 6).



Plot 6

Results and Findings:

1. Correlation Analysis:

- Correlation matrices are meticulously generated, shedding light on the intricate relationships between various stock features. This analysis offers a nuanced understanding of how different attributes are correlated within the dataset.

2. **Clustering Analysis:**
    - K-Means clustering emerges as the core algorithm for grouping stocks based on similarities in their market behaviors.
    - Silhouette and Elbow methods work in tandem to pinpoint the optimal number of clusters (K=4). The Silhouette method ensures the chosen K value results in well-defined clusters, while the Elbow method validates the optimal point where additional clusters provide diminishing returns.
3. **Visualization:**
    - Visualizations serve as a pivotal component in interpreting the clustering results. Histograms elucidate the distribution of data points across clusters, offering insights into the composition and size of each cluster.
    - 3D scatter plots provide a captivating and multi-dimensional view of the clustered stocks, facilitating a deeper understanding of their spatial relationships.

Conclusion: The Stock Clustering Analysis project aptly showcases the seamless integration of Apache Kafka, PySpark, and advanced machine learning techniques for the purpose of categorizing stocks based on their price behaviors. The findings hold immense value for investors and financial analysts, empowering them to identify intricate patterns and make informed decisions within the dynamic stock market landscape.

Future Work: As the project serves as a robust foundation, potential future enhancements could include:

- Experimenting with alternative clustering algorithms for comparative purposes.
- Implementing automated data streaming and clustering processes for real-time analysis, enhancing the project's scalability and applicability.