SAMSUNG

Together for Tomorrow!
Enabling People
Education for Future Generations

Big Data Final Project

SIC 6

Real-Time Stock Market Analysis Using Streaming Processing

Hassan Mohamed ElMahallawy
Mohamed reda mohamed amer
Maya Assem Mohamed Ali
Abdelrahman Hussien Mohamed
Nancy Youssef Zaher Ahmed

Facilitator: Abdelrahman Mahmoud

Samsung Innovation Campus

Agenda

- Introduction
- Problem Statement
- Project Objectives
- Data Understanding
- Pipeline Architecture
- Key Features
- Results & Visualizations

Problem Statement

Data Overload

- Stock markets generate massive amounts of data in real-time, which is difficult to monitor manually

Market Volatility

- Prices and trends change rapidly, and it is crucial to identify trends or react to changes (like a price drop) instantly.

News Impact

- Market-moving news (e.g., company earnings, geopolitical events) directly affects stock prices, and tracking relevant news in real-time is a challenge.

Decision Lag

- Without automated alerts and real-time data analysis, decision-making can be slow, leading to missed opportunities or financial loss

Project Objectives

01 Real-Time Stock Data Tracking

02 Sentiment Analysis of News Articles

03 Generate Real-Time Alerts

04 Data Pipeline Integration

05 Real-Time Visualization and Querying

Project Objectives

Predictive Analytics for Stock Movements

- By analyzing live price data and news sentiment, this solution enables early detection of trends and predictions of market shifts, helping investors make better decisions

Faster Reaction to Market News

- The integration of news sentiment analysis allows traders and companies to react faster to events affecting stock prices.

Real-Time Alerts:

- The system can generate real-time alerts based on specific criteria like price drops, ensuring that investors can act instantly

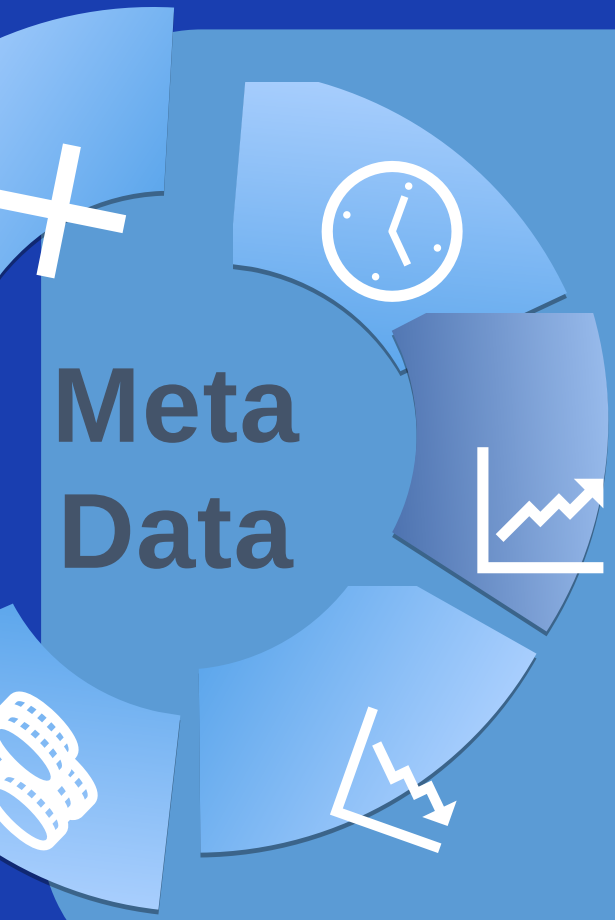
Storage And Visualization:

- Streaming processing algorithms are used to identify trends and patterns in stock price movements. These algorithms analyze data streams in real time

Understand Our Data

Samsung Innovation Campus

Understand Our Data



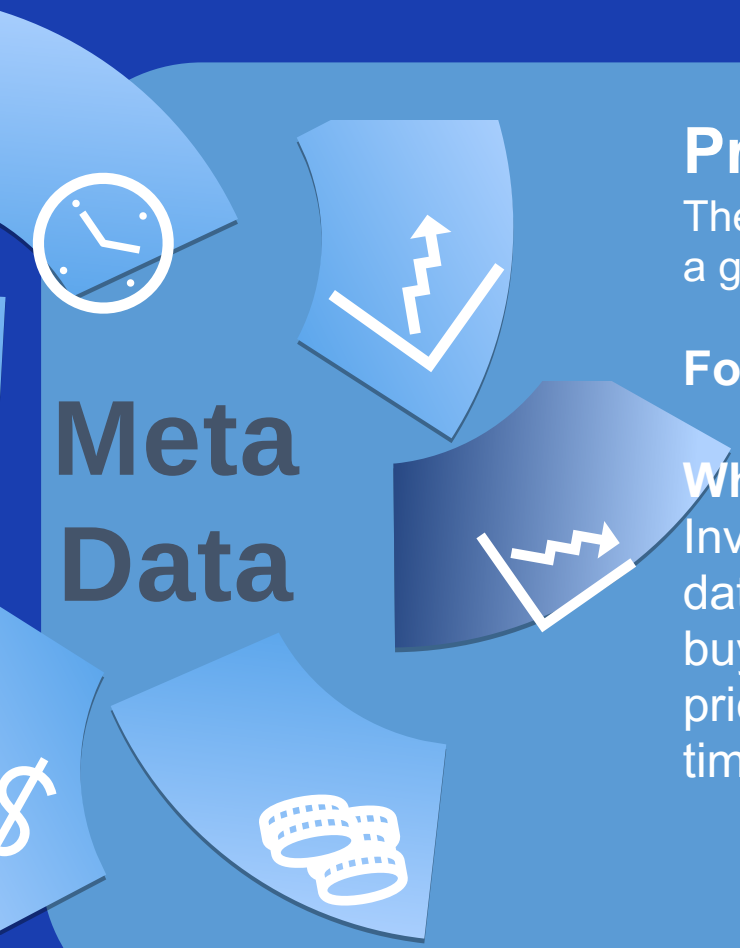
Symbol

Represents the stock ticker symbol (e.g., AAPL for Apple, TSLA for Tesla). This is the unique identifier for each stock.

Format: String

Why It's Important: This column acts as a partition key in Cassandra, ensuring that stock data is easily queryable by stock symbol.

Understand Our Data



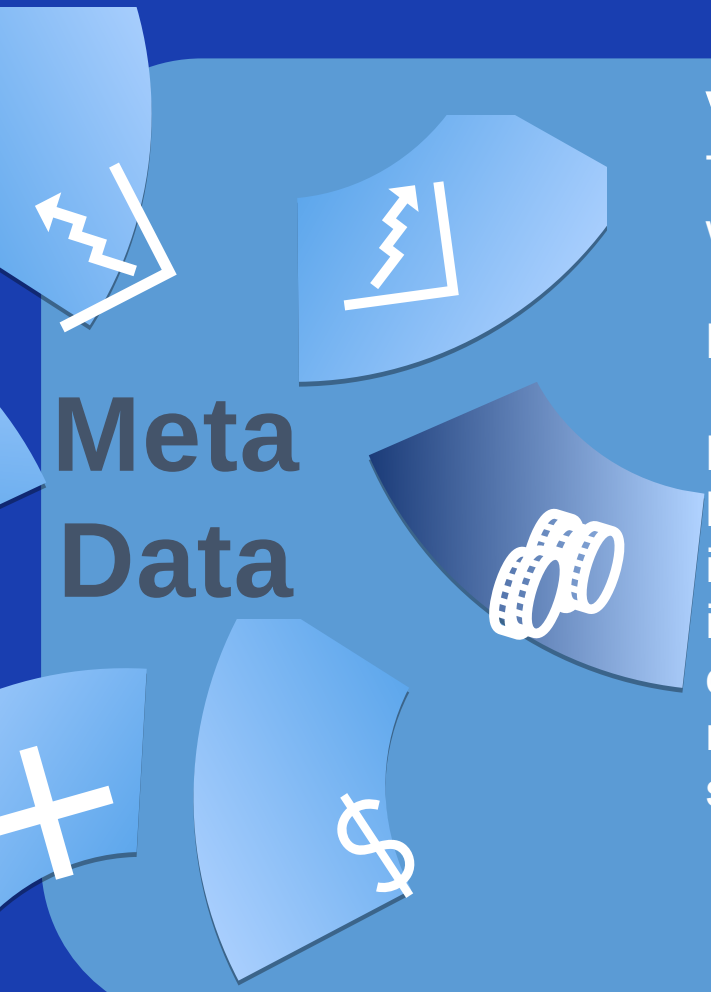
Price:

The Real-Time Price of the stock at a given timestamp

Format: Decimal

Why It's Important:
Investors need the most up-to-date price to make quick buying or selling decisions. The price column is critical for real-time alerts.

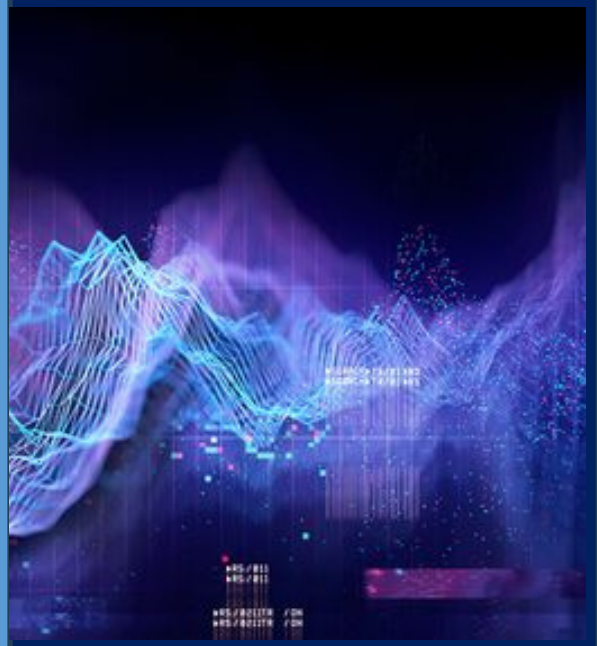
Understand Our Data



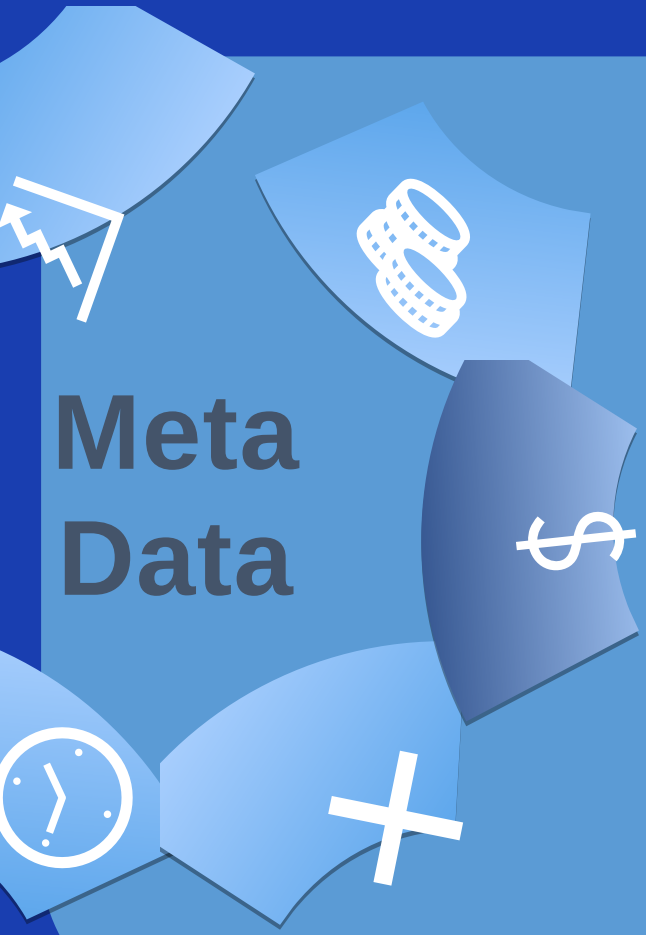
Volume
The total number of shares that were traded during the day.

Format: Integer (e.g., 5,000,000)

Purpose: Volume reflects the liquidity of a stock and can indicate the level of interest from investors. High trading volume often accompanies price movements and can signal significant News or market events



Understand Our Data

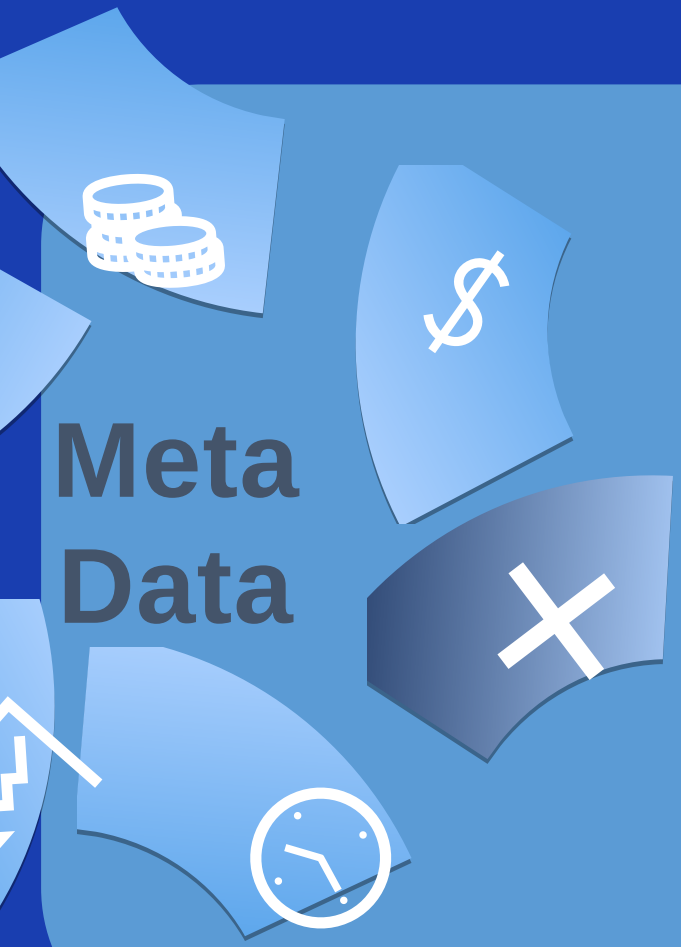


News_source
The source from which the news sentiment data was obtained.

Format: Text

Why It's Important:
Identifying the news source helps evaluate the reliability and potential influence of the sentiment score on the stock's price.

Understand Our Data

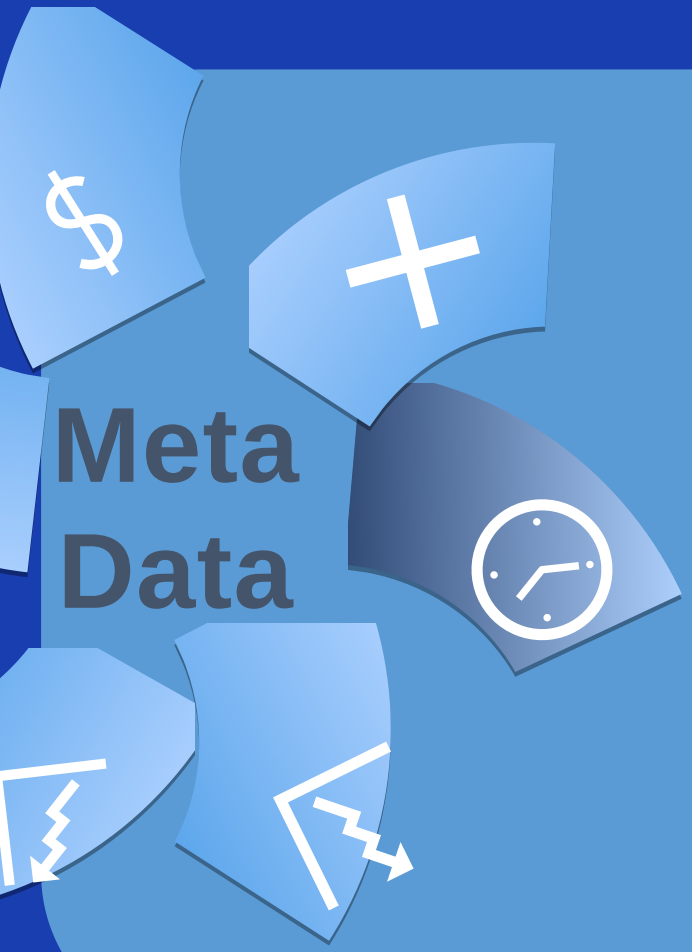


sentiment_score
The score derived from sentiment analysis, typically ranging from -1 (negative) to +1 (positive), representing the sentiment of financial news articles.

Format: Decimal

Why It's Important:
This score helps in predicting stock price movement based on external factors such as news, tweets, or press releases. It provides a way to correlate public sentiment with stock performance.

Understand Our Data

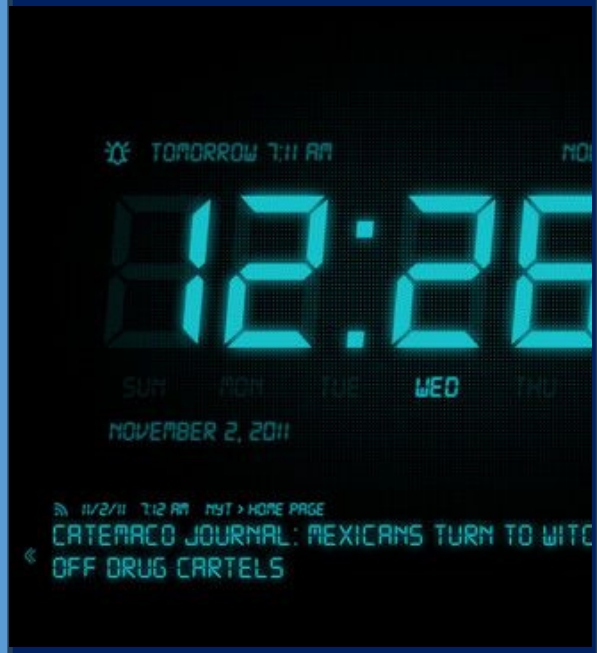


Time-Stamp

The exact time the data was captured (both for stock prices and related news sentiment).

Format: timeStamp

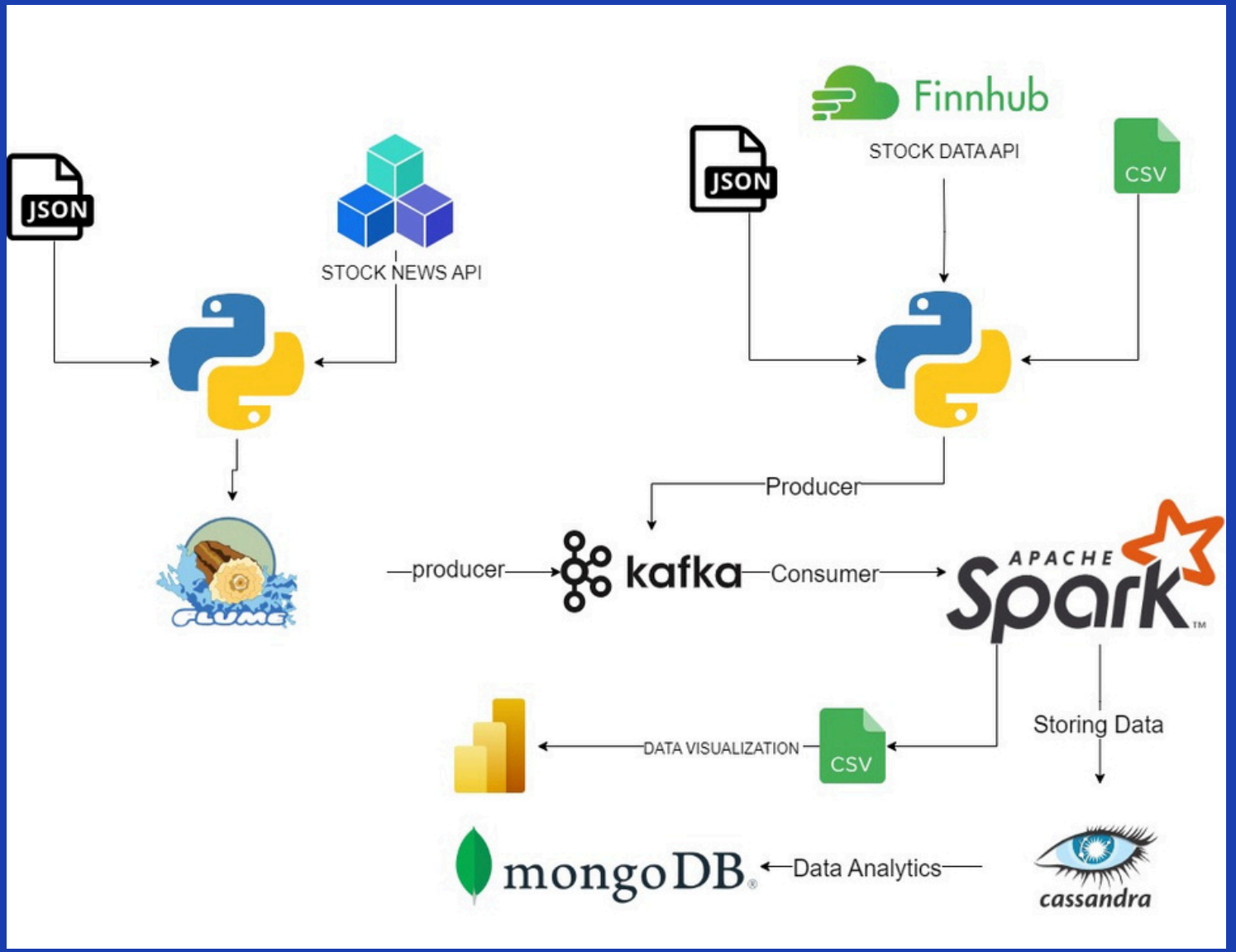
Why It's Important:
This allows tracking stock performance and sentiment over time. It helps identify when a price spike or significant sentiment change occurred.



Agenda

- Introduction
- Problem Statement
- Project Objectives
- Data Understanding
- **Pipeline Architecture**
- Key Features
- Results & Visualizations

Stock Market Data Pipeline



Samsung Innovation Campus

Stock News Data Pipeline

Key Features

Ingesting Real-Time Stock Market Data

- Utilize **Apache Kafka** to continuously stream live stock prices and trading volume data from multiple market sources, ensuring up-to-the-second insights.

Sentiment Analysis on Financial News

- Leverage **Apache Flume** to capture relevant financial news articles and perform sentiment analysis, helping to assess market sentiment and its potential impact on stock prices.

Simultaneous Data Stream Processing

- Process both stock market and news data streams in real-time using **Apache Spark** Streaming for comprehensive, near-instant analysis of price trends and market sentiment.

Key Features

Dynamic Alerts for Key Market Events

- Automatically trigger alerts based on specific conditions like sharp price fluctuations or unusual trading volumes, allowing for immediate action.

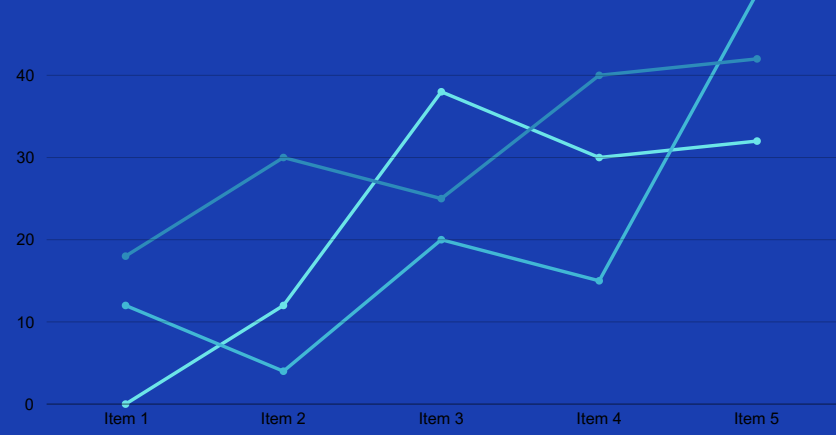
Optimized Data Storage for Real-Time Querying

- Store processed data in **Cassandra**, providing fast, scalable, and efficient querying capabilities to support real-time decision-making.

Interactive Visualization Dashboards

- Display live stock market trends, trading volumes, and sentiment analysis results using **MongoDB** and **Python-powered dashboards**, offering actionable insights through visually engaging charts and graphs.

News Sentiment Analysis



It centers around analyzing the news articles related to stock market activities.
And to understand public sentiment regarding specific stocks and provide actionable insights for investors and financial analysts

Why Random Forest?

- **Robustness:** It improves predictive accuracy and controls overfitting by averaging the results of individual trees.
- **Feature Importance:** The model can provide insights into which features (words or phrases in the news articles) are most significant for sentiment prediction.
- **Flexibility:** Random Forest handles both numerical and categorical data well, making it adaptable to various types of input features that may arise from news articles.

Data Preparation with CSV Files

The purposes:

- ***Training the Model:*** *By applying a train-test split approach*
  - ***Training Set***
  - ***Test Set:*** *vital to avoid overfitting*
- ***Performance Metrics:*** *(accuracy, precision, recall, and F1 score) that inform us how effectively the model predicts sentiment. These metrics are essential for validating the model's effectiveness before deploying it in real-time applications.*

After That

- **Load the model using joblib**
- **Deploy in real time news data**

```
Accuracy: 0.8714285714285714
Classification Report:
              precision    recall  f1-score   support

    negative       0.99      0.59      0.74       115
     neutral       0.88      0.82      0.85       158
    positive       0.85      0.97      0.91       427

    accuracy                           0.87       700
   macro avg       0.91      0.79      0.83       700
weighted avg       0.88      0.87      0.87       700
```

Real-World Benefits:

- **Informed Decision-Making:** Investors can access immediate sentiment insights, helping them make informed decisions.

- **Risk Management:** Understanding sentiment can aid in identifying potential risks. If the sentiment turns negative across multiple news sources, stakeholders can take preemptive actions to mitigate losses

- **Market Predictions:** analysts can predict stock price movements and market trends, enhancing their investment strategies.

- **Competitive Advantage:** Financial institutions can leverage this sentiment analysis tool to enhance their research capabilities, providing clients with better insights and advice, ultimately fostering a competitive edge in the market.

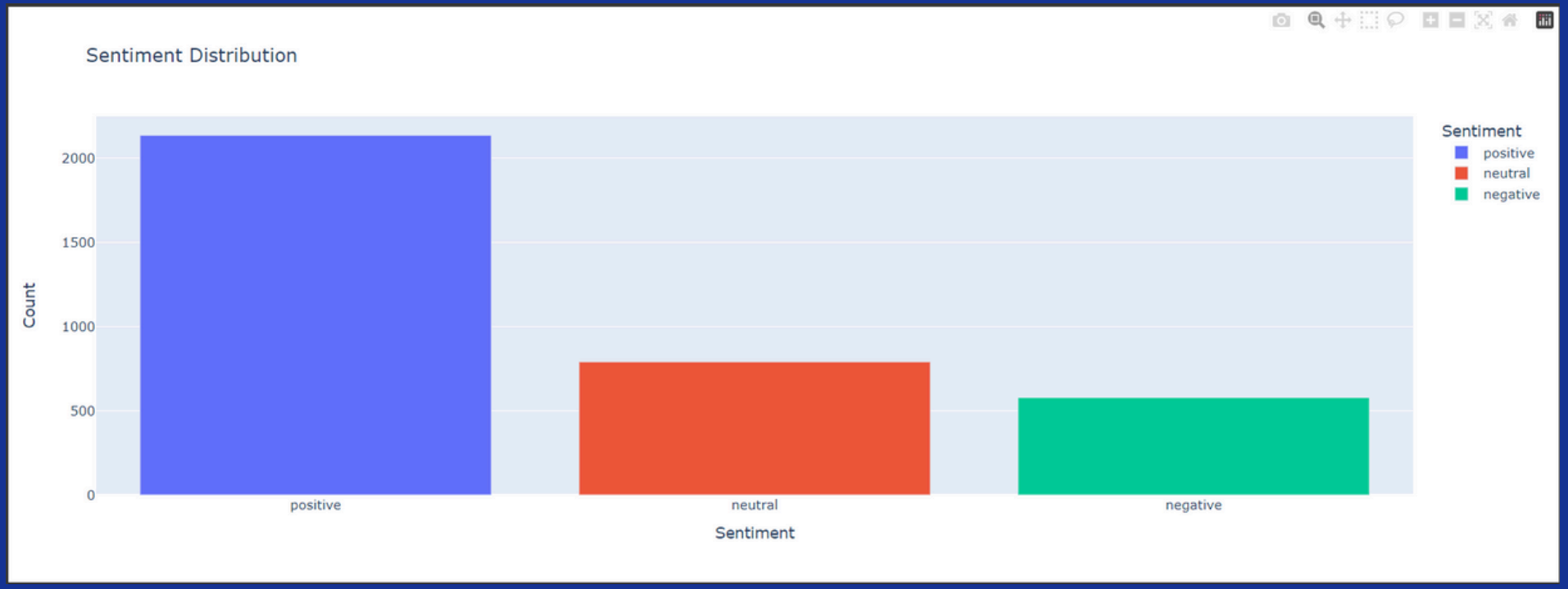
>Visualizations

Trading Volume by Stock Symbol

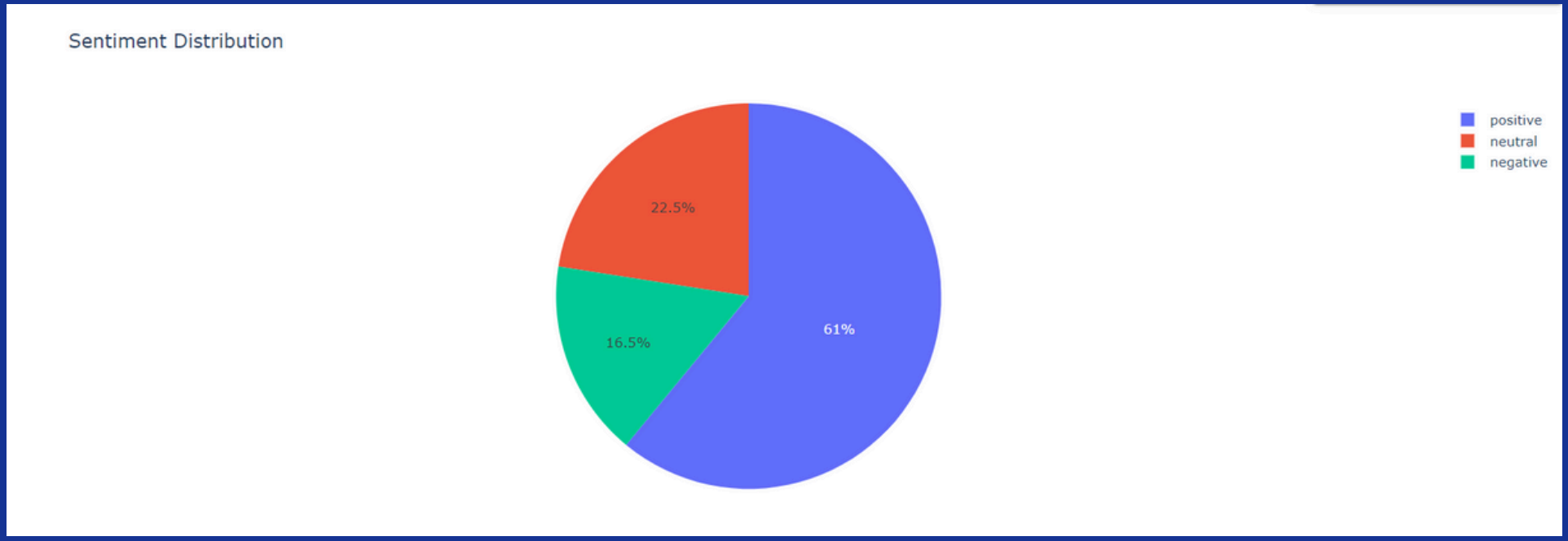
quickly identifying which stocks are experiencing the highest trading activity, which might indicate increased market interest or volatility

stock_symbol	total_volume
AAPL	5,100,000
TSLA	3,450,000
AMZN	2,900,000
MSFT	2,300,000
GOOGL	1,850,000

- This bar chart displays the count of each sentiment category (positive, negative, neutral) in the dataset. It provides a quick overview of sentiment distribution, which can be useful for stakeholders.

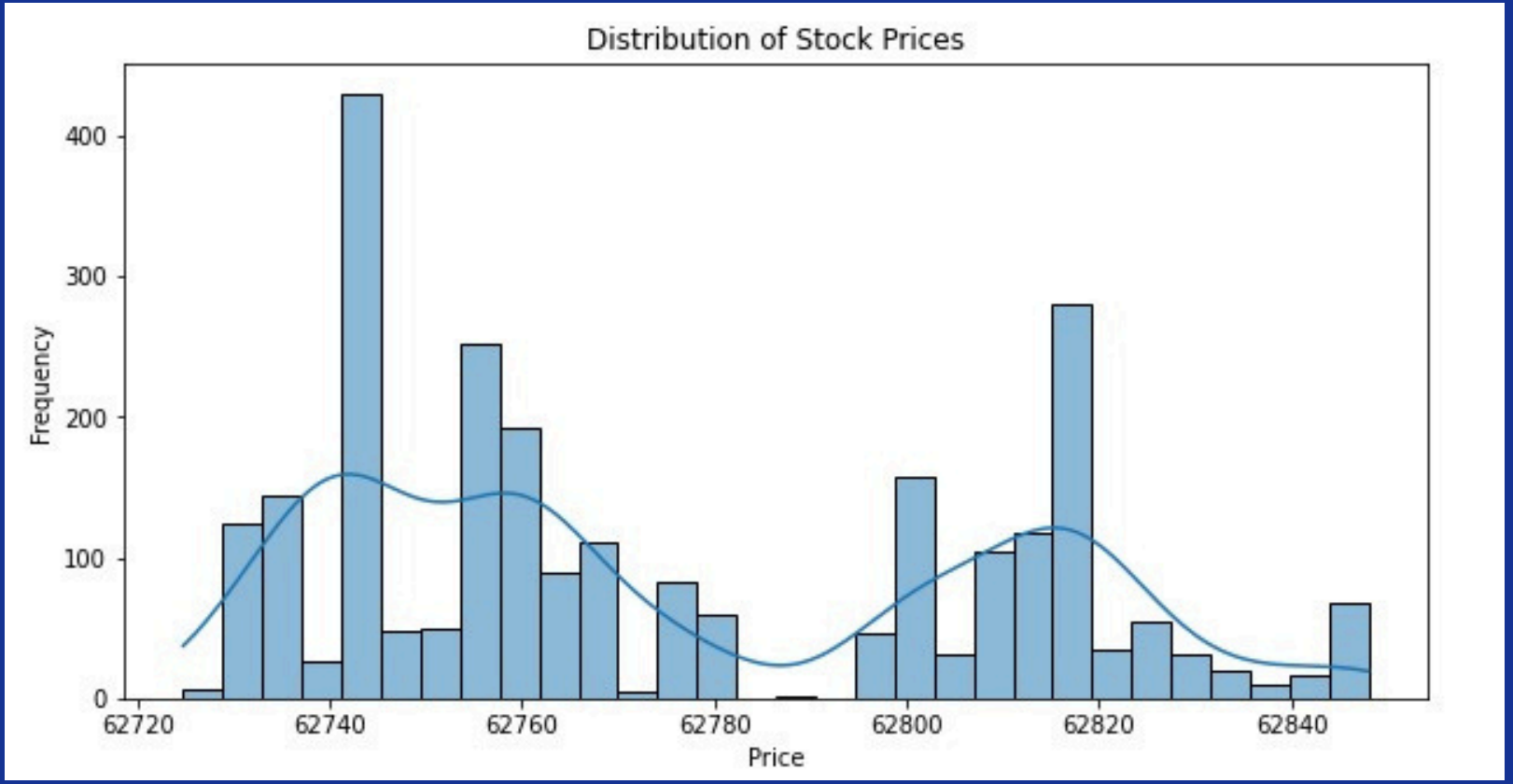


- The pie chart visually represents the proportion of each sentiment category. It helps in understanding the overall sentiment landscape of the news articles.



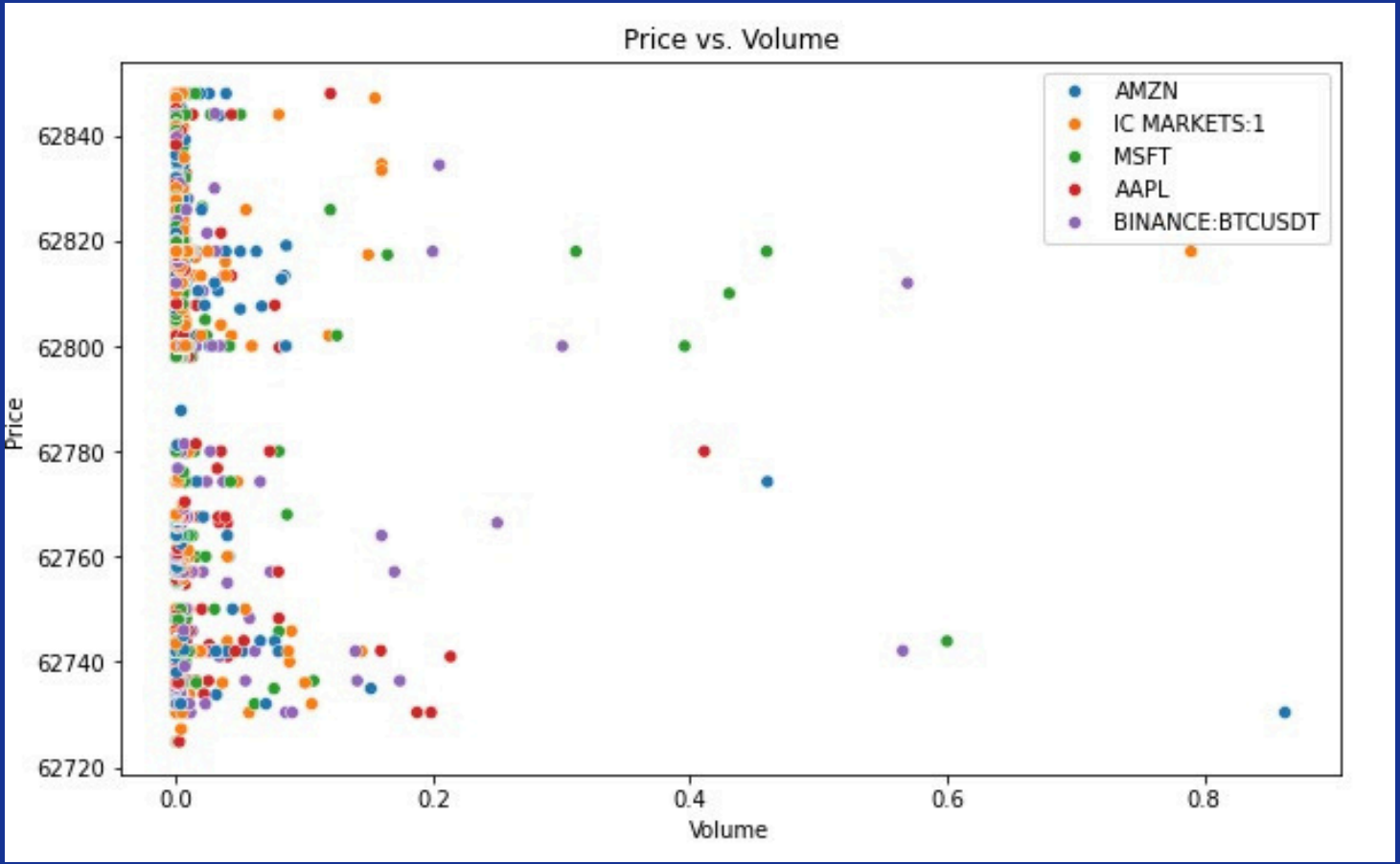
Distribution of Stock Prices

This graph visualizes the distribution of stock prices over a certain period. Peaks in the graph represent common price ranges, while the KDE line gives a smooth estimate of the probability distribution, showing how stock prices are distributed around the central value.



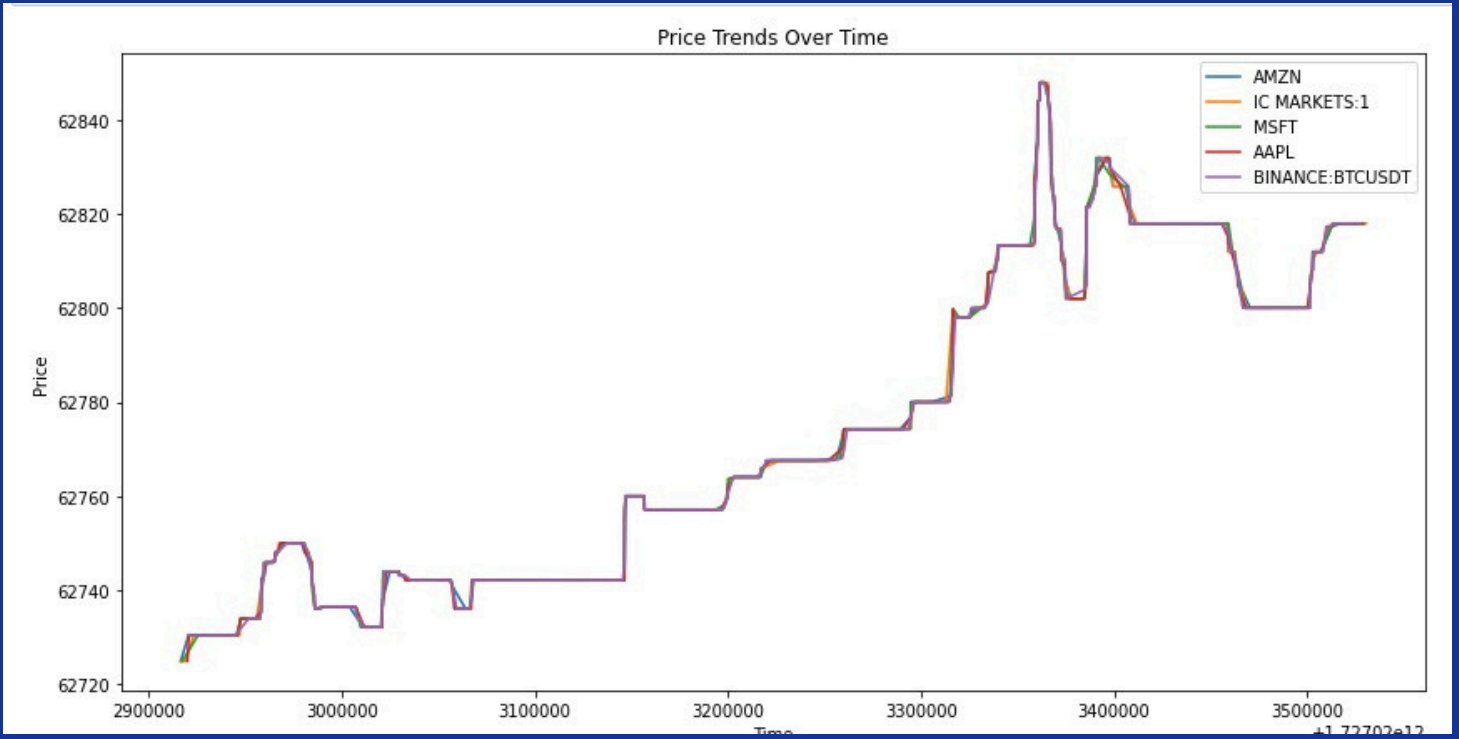
Price vs. Volume for Multiple Stocks

This graph shows the relationship between the price of stocks and their corresponding trading volume.



The price trends of five different assets over time

The graph shows that the prices of all assets have generally increased over time, with some periods of fluctuation



SAMSUNG

Together for Tomorrow!

Enabling People

Education for Future Generations