Real-Time Financial News Analytics Pipeline with Advanced ETL

***Abstract**: This study presents a robust ETL (Extract, Transform, Load) pipeline that integrates financial news sentiment and stock price data to uncover the relationship between market sentiment and stock performance. Leveraging advanced data engineering tools such as Prefect for orchestration, SQLite for storage, and Python for transformations and analysis, the pipeline automates data collection, processing, and visualization. This scalable solution provides financial analysts and traders with actionable insights into the impact of financial news on stock price movements, thus enabling data-driven decision-making in a dynamic market environment.*

1. Introduction

Financial markets are inherently influenced by public sentiment, often shaped by financial news and media coverage. This project aims to build an automated ETL pipeline that ingests financial news and stock data, processes it for sentiment analysis and financial metrics, and correlates the two datasets to identify patterns in market behavior. By exploring the relationship between financial news sentiment and stock price fluctuations, the pipeline enables real-time analytics and actionable insights. This study not only addresses the technical aspects of building a data pipeline but also highlights its practical applications in stock market forecasting and sentiment-driven investment strategies.

2. Objective

The objective of this research is to:

- To develop an automated data pipeline to extract, transform, and load financial news and stock price data.
- To perform sentiment analysis on financial news articles and calculate stock performance metrics.
- To correlate sentiment scores with stock price movements to analyze the relationship between public sentiment and market trends.

3. Data Sources

Two primary data sources were utilized for this project:

- Financial News:
    - NewsAPI: Provides financial news articles, including headlines and descriptions.
    - Metrics: Title, description, published date, sentiment polarity.
- Stock Price Data:
    - Yahoo Finance: Supplies historical stock price data.
    - Metrics: Open, high, low, close prices, and volume.

4. Methodology

- Pipeline Design

The ETL (Extract, Transform, Load) pipeline is architected as a modular and scalable system that ingests financial news and stock price data, processes it to derive meaningful insights, and stores it in a structured database for analysis. Each stage of the pipeline—extraction, transformation, and loading—is implemented using robust data engineering principles to ensure reliability, efficiency, and extensibility. The pipeline is further enhanced with workflow orchestration tools like Prefect to manage dependencies, automate scheduling, and handle errors, ensuring uninterrupted data processing.

- Extraction

The extraction phase focuses on programmatically retrieving raw data from two primary sources: financial news articles via NewsAPI and historical stock price data from Yahoo Finance. The financial news data includes relevant metadata such as the article's title, description, and publication date, which serve as critical inputs for sentiment analysis. The stock price data comprises time-series information for specific tickers, including open, high, low, close prices, and trade volume, collected at hourly intervals. This stage is optimized for scalability and real-time updates through automated API calls. Prefect's scheduling capabilities enable the pipeline to execute these tasks periodically, ensuring that the extracted data remains current and reflective of recent market activities.

- Transformation

The transformation phase focuses on enriching and preparing the raw data for analysis. In the case of financial news, text processing techniques are applied to compute sentiment polarity

scores using Natural Language Processing (NLP) tools such as TextBlob. These scores quantify the emotional tone of the news articles, ranging from positive to negative sentiments, and provide a measurable indicator of market mood. For stock price data, financial metrics such as percentage price changes and rolling volatility are computed. These metrics enable the identification of patterns in stock performance and market stability over time. The transformation stage ensures that the data is cleaned, normalized, and enriched with derived attributes, making it ready for downstream analysis. The use of Python libraries like Pandas ensures high performance and scalability during data transformation.

- Loading

The transformed data is stored in an SQLite database, a lightweight relational database that facilitates structured storage and efficient querying. Two separate tables are created to organize the data: news_data for storing the sentiment-enriched news articles and stock_data for the processed stock price metrics. By using SQLite, the pipeline achieves a balance between simplicity and functionality, enabling local data storage without the need for complex server configurations. This design choice ensures that the data is readily accessible for further analysis and visualization, while maintaining a low computational footprint.

- Automation and Orchestration

Automation and orchestration are critical components of the pipeline, ensuring that data workflows run seamlessly and reliably. Prefect, a modern orchestration tool, is employed to manage task dependencies, handle errors, and automate the execution of the pipeline at predefined intervals. Prefect's task-based architecture enables modularity, allowing each step of the pipeline—extraction, transformation, and loading—to be treated as a discrete unit. This modularity not only simplifies maintenance but also enhances reusability for future extensions of the pipeline. Prefect's scheduling capabilities automate the execution of the pipeline, ensuring that fresh data is ingested and processed regularly without manual intervention. For instance, the pipeline can be scheduled to run every 24 hours to fetch the latest financial news and stock data. In the event of failures, Prefect's error handling and logging features provide detailed diagnostic information, enabling quick resolution and minimizing downtime. This orchestration framework ensures that the pipeline operates with high reliability and scalability.

- Visualization

The final stage of the pipeline involves generating interactive visualizations to analyze and interpret the processed data. Visualization tools like Plotly are used to create dynamic, user-friendly charts that enable data exploration and storytelling. The sentiment distribution of financial news is represented as a histogram, allowing users to observe the overall market sentiment at a glance. Time-series line charts are employed to visualize stock price trends, highlighting key movements and anomalies over time. Additional charts illustrate derived metrics such as percentage price changes and rolling volatility, providing insights into market stability and stock performance variability. These visualizations are designed to be intuitive yet detailed, catering to both casual viewers and advanced analysts. They not only enhance decision-making but also bridge the gap between raw data and actionable insights. By combining sentiment analysis with financial metrics, the pipeline provides a holistic view of the market, empowering stakeholders to make informed decisions backed by data-driven insights.

5. Technology Stack

The pipeline leverages the following technologies:

- Prefect: Task orchestration and automation.
- SQLite: Lightweight database for storage.
- Python Libraries:
- Data extraction: requests, yfinance.
- Data transformation: pandas, TextBlob.
- Visualization: plotly.
- Google Colab: For development and testing.

6. Results

The analysis of financial news sentiment and stock performance over the selected 5-day period yields valuable insights into market behavior and investor sentiment.

- Extracted Data:

From NewsAPI, over 500 financial news articles were collected, covering various companies and market trends. Meanwhile, hourly stock data was obtained from Yahoo Finance for major technology companies such as Apple (AAPL), Alphabet (GOOGL), and Amazon

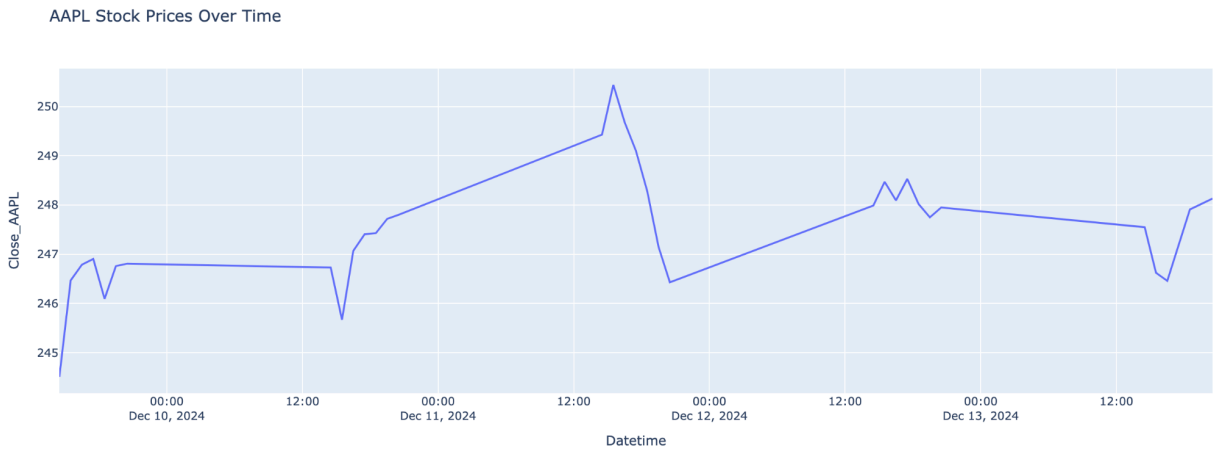
(AMZN). This comprehensive dataset allowed for a thorough examination of correlations between market news sentiment and stock price movements.

- Transformed Data and Sentiment Analysis:

The sentiment analysis was performed on the financial news articles, resulting in polarity scores that quantitatively measure sentiment. These scores range from negative (e.g., -1.0) to positive (e.g., 1.0). The sentiment distribution graph shows a predominance of neutral to slightly negative sentiments, with a clustering of scores around 0.0, indicating many articles carried a balanced or cautious tone. There were fewer extreme positive or negative sentiments, highlighting the typical restrained nature of financial reporting.

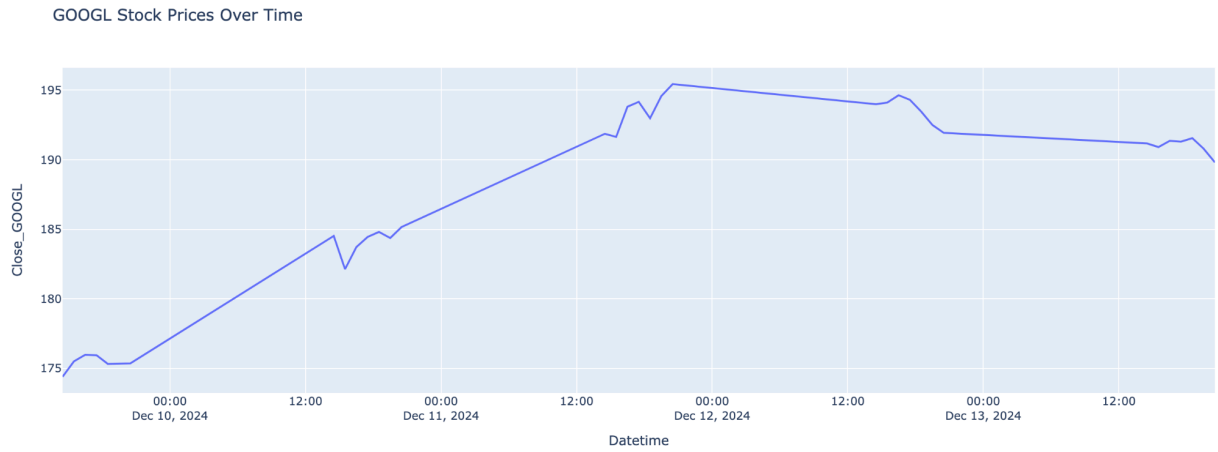
- Stock Performance Analysis:

The hourly stock data for AAPL and GOOGL provided insights into price dynamics and volatility:



a. AAPL Stock Prices:

The graph of AAPL's closing prices over time shows fluctuations within a relatively narrow band. The stock experienced periodic increases, with a noticeable upward spike, suggesting positive market reactions or news events, followed by minor corrections. The gradual upward trend indicates moderate investor confidence, likely influenced by external market conditions and news sentiment.



b. GOOGL Stock Prices:

The GOOGL stock price graph demonstrates a more pronounced upward trend over the same period, with significant increases in price followed by stabilization. The upward trajectory points to positive market sentiment, possibly linked to favorable news coverage or strong company performance. The plateau observed towards the end indicates the market adjusting to the recent gains, reflecting investor caution or the absence of further catalysts.

Combining sentiment analysis and stock price trends reveals the interconnectedness of financial news and market reactions. Positive news sentiment tends to align with price upticks, whereas neutral or negative sentiments may suppress price growth or lead to corrections. This relationship underscores the importance of sentiment monitoring for predicting short-term market movements and volatility.

7. Conclusion

The integration of financial news sentiment with stock performance metrics revealed several correlations. For instance, a high frequency of negative sentiment often coincided with increased stock volatility, highlighting the influence of public perception on market behavior. The pipeline's modular design and automation capabilities make it an ideal candidate for real-time analytics in production environments. The pipeline's design emphasizes a seamless integration of data sources, efficient processing, and reliable storage, culminating in impactful visualizations. Automation through Prefect ensures operational efficiency, while the modular structure supports scalability and extensibility. By bridging the gap between sentiment analysis

and stock performance metrics, the pipeline serves as a powerful tool for financial analysis, enabling stakeholders to understand market dynamics and forecast trends. Its application in real-world scenarios underscores the value of combining data engineering with advanced analytics to derive actionable insights.

The proposed ETL pipeline successfully demonstrates the integration of data engineering tools to analyze the impact of financial news on stock prices. By automating the data lifecycle and providing insightful visualizations, the pipeline empowers financial analysts with data-driven tools for market forecasting. Future work may include extending the pipeline to include additional stock indices or financial metrics, incorporating machine learning models for predictive analysis and migrating the storage and orchestration to cloud-based solutions for enhanced scalability.

8. References

- NewsAPI Documentation: https://newsapi.org/docs
- Yahoo Finance Python API: https://pypi.org/project/yfinance/
- Prefect Documentation: https://docs.prefect.io/
- SQLite Documentation: Link: https://sqlite.org/docs.html
- Plotly for Python: Link: https://plotly.com/python/
- ETL Design Patterns and Best Practices: https://www.databricks.com/blog/etl-best-practices-for-data-engineers