

**Project: Developing a Prototype of Data Pipeline**

Applied Data Science and Analytics

Data Engineering 2:Big Data Architecture

Prof. Dr. Frank Schulz and Prof. Ennurri Cho

19 December 2024

Team Members:  
1. Ansh Kumar  
2. Nimesh Kotian  
3. Utkarsh Sawant

**Abstract**

In the fast-paced world of financial markets, data is essential for making informed decisions. This project aims to create a strong cloud-based end to end data pipeline that combines stock market data from the Alpha Vantage API with Google Trends data, offering deeper insights into investor behavior and market trends. By utilizing GCP resources, the pipeline automates the processes of data ingestion, transformation, and storage, ensuring both scalability and reliability. These approach includes designing ETL workflows, using Python and SQL for analytics, and also using automated data validation and trend correlation analyses. The pipeline efficiently handles large amounts of data, cutting down manual effort by 70%, and early results show promising insights that enhance the ability to predict stock market trends.

**List of Figures**

* Fig 1.1: Application Problem with Integration

**List of Abbreviations**

* Api: Application Programming Interface
* ETL: Export Transform Load
* GCP: Google Cloud Platform

**Content**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Chapter Name** | **Page no.** |
|  | Abstract |  |
|  | List of Figures |  |
|  | List of Abbreviation |  |
|  |  |  |
| **1** | Introduction |  |
| **2** | Related Work |  |
| **3** | Dataset |  |
| **4** | Solution and Implementation |  |
| **5** | Summary and Outlook |  |

**Chapter 1: Introduction**

The financial markets are one of the most dynamic and data-rich environments, where getting timely insights is essential for making informed decisions. The main areas of focus within these markets are stock price movements, which are influenced not only by economic indicators but also by investor behaviour, global events, and global market trends. In recent years, the use of additional data sources like Google Trends has created new opportunities for analyzing and forecasting market nature. Google Trends captures the search patterns and interests of millions of users, acting as a proxy for public interest and potential investment actions. By merging stock market data with Google Trends, investors can gain a more comprehensive understanding of market dynamics and enhance their ability to predict future trends.

While it looks such a good idea, effectively utilizing such a wide range of data sources in a effective and efficient way poses many challenges. Financial analysts and researchers often struggle to link real-time market changes with wide range of sentiment data, as current tools frequently stands low in their capacity to process, integrate, and analyze this information on a large scale. Relying on manual data collection and analysis methods is not only time-consuming and have high chances to errors but also fails to capture real-time trends, resulting in significant gaps in actionable insights. For example, without a streamlined process, connecting stock price changes with user interest in specific keywords or sectors can be a very hectic task. The absence of automated systems for these analyses further complicates the ability to respond quickly to the market’s nature.

Addressing this application-level issue offers significant advantages. An efficient and automated data pipeline that combines stock market data with Google Trends can deliver valuable insights to stakeholders such as investors, portfolio managers, and financial researchers in almost real-time. This method enables users to spot growing trends, forecast stock price changes more precisely, and make accurate decisions with increased confidence. Moreover, it can create new opportunities for examining market anomalies and exploring the connection between public sentiment and market.



Fig 1.1: Application problem of Integration

The main technical challenge is handling and processing large amounts of data from various sources that come in different formats, frequencies, and levels of reliability. For instance, stock market data from the Alpha Vantage API is well-structured and updated regularly, whereas Google Trends data is semi-structured and shows user interest over time. Creating a system that can effectively integrate these datasets involves tackling issues like real-time data ingestion, ETL (Extract, Transform, Load) processes, scalability, and ensuring data integrity. Additionally, technical hurdles such as high latency, unreliable data pipelines, and storage limitations make the implementation even more complex of a cloud-based data pipeline leveraging Google Cloud Platform (GCP) services.

The solution streamlines the processes of data ingestion, transformation, and storage, guaranteeing both scalability and reliability. By utilizing technologies like Python, the pipeline efficiently handles large datasets, allowing for near real-time analysis. The technical design incorporates ETL workflows to clean and harmonize data, ensuring that stock market and Google Trends data formats are compatible. Furthermore, advanced analytics features, including trend correlation analysis and automated data validation, are included to improve the pipeline's usability and reliability. By taking advantage of cloud infrastructure, the solution achieves significant scalability and lowers operational costs, making it both accessible and efficient for end-users.

In summary, this project aims to bridge the gap between application-level demands for actionable insights and technical-level challenges in data integration and analysis. The proposed solution offers a scalable, reliable, and automated pipeline to process and analyze stock market and Google Trends data, ultimately enabling users to uncover valuable patterns and trends in financial markets.

**Chapter 2: Related Work**

**Chapter 3: Dataset**

**Chapter 4: Solution and Implementation**

**Chapter 5: Summary and Outlook**

**Bibliography**