**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

**Abstract**

In the fast-paced domain of financial markets, data plays a critical role in decision-making. This project focuses on developing a robust cloud-based data pipeline to integrate stock market data from the Alpha Vantage API with Google Trends data, providing deeper insights into investor behaviour and market movements. By leveraging GCP resources, the pipeline automates data ingestion, transformation, and storage while ensuring scalability and reliability. The methodology involves designing ETL workflows, implementing analytics using Python and SQL, and integrating automated data validation and trend correlation analyses. The pipeline effectively processes large volumes of data, reducing manual effort by 70%, and initial results demonstrate promising insights, enhancing the predictive power for stock market trends.

**List of Figures**

* Fig 1.1: Application Problem With Integration

**List of Abbrevation**

* 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 operate as one of the most dynamic and data-intensive domains, where timely insights are critical for making informed decisions. Among various aspects of financial markets, stock price movements represent a key area of focus, driven not only by economic indicators but also by investor sentiment, global events, and market trends. In recent years, the integration of auxiliary data sources, such as Google Trends, has opened new avenues for analyzing and predicting market behavior. Google Trends, by reflecting the search patterns and interests of millions of users, serves as a proxy for public sentiment and potential investment behavior. Combining stock market data with Google Trends presents a unique opportunity to gain a holistic view of market dynamics and predict future trends more effectively.

Despite its potential, leveraging such diverse data sources in a unified and efficient manner presents several challenges. On the application level, financial analysts and researchers face difficulties in correlating real-time market movements with broader sentiment data, as existing tools often lack the ability to process, integrate, and analyze this information at scale. Manual methods for data collection and analysis are time-intensive, error-prone, and fail to capture real-time trends, leaving significant gaps in actionable insights. For instance, without a streamlined pipeline, correlating stock price fluctuations with user interest in specific keywords or sectors becomes an arduous task. The lack of automated systems for such analyses further hampers the ability to adapt quickly to changing market conditions.

The benefits of solving this application-level problem are substantial. A reliable and automated data pipeline that integrates stock market data and Google Trends can provide stakeholders, including investors, portfolio managers, and financial researchers, with actionable insights in near real-time. This solution can empower users to identify emerging trends, predict stock price movements more accurately, and make data-driven decisions with greater confidence. Additionally, it can open new possibilities for analyzing market anomalies and understanding the relationship between public sentiment and market behavior.



Fig 1.1: Application problem of Integration

From a technical perspective, the challenge lies in managing and processing large volumes of data from disparate sources with varying formats, frequencies, and reliability. Stock market data from sources such as the Alpha Vantage API is structured and updated frequently, while Google Trends data is semi-structured and reflects user interest over time. Designing a system capable of integrating these datasets requires addressing issues such as real-time data ingestion, ETL (Extract, Transform, Load) processes, scalability, and maintaining data integrity. Moreover, technical bottlenecks, such as high latency, unreliable data pipelines, and storage constraints, further complicate the implementation.

To address these technical challenges, this project proposes the development of a cloud-based data pipeline leveraging Google Cloud Platform (GCP) services. The solution automates data ingestion, transformation, and storage processes, ensuring scalability and reliability. Using technologies like Python, the pipeline processes large datasets efficiently, enabling near real-time analysis. The technical design includes implementing ETL workflows to clean and harmonize data, ensuring compatibility between stock market and Google Trends data formats. Additionally, advanced analytics features, such as trend correlation analysis and automated data validation, are integrated to enhance the usability and reliability of the pipeline. By leveraging cloud infrastructure, the solution achieves high scalability and reduces operational costs, making it 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**