

SCALABLE BIG DATA ANALYTICS FRAMEWORK FOR FINANCIAL TIME SERIES USING SNOWFLAKE AND PYSPARK AND MACHINE LEARNING

¹Y.P Srinath Reddy, ²Nithin Kodamala, ³Latheesh Kandhi, ⁴Vishnu Madanambeti

¹Assistant Professor, ²Student, ³Student, ⁴Student

¹Department of Computer Science and Engineering (Data Science),

¹Rajeev Gandhi Memorial College of Engineering and Technology, Nandyal, Andhra Pradesh, India

srinathreddycseds@rgmcet.edu.in, 2126nithin@gmail.com,

latheeshkandhi833@gmail.com, vishnumadanambeti@gmail.com

Abstract— Financial trading activities generate massive amounts of time-series data on a daily basis, which creates challenges for conventional data processing and analytical systems. This project presents a cloud-based big data analytics framework designed to efficiently manage and analyze financial time-series data, with particular emphasis on coffee commodity price analysis. The proposed system implements a complete data engineering workflow in which raw financial data is first stored in Amazon S3 and then processed using PySpark for distributed data cleaning, transformation, and aggregation. The transformed datasets are loaded into Snowflake, which functions as a scalable cloud data warehouse supporting fast analytical queries and reporting. A Machine Learning layer is integrated on top of the analytics layer to identify historical price patterns and generate future trend predictions from the processed time-series data. The framework is organized using a layered architecture that includes data storage, ETL processing, cloud warehousing, analytics, and machine learning components. The results show that the framework is capable of handling large-scale financial time-series data efficiently while maintaining scalability, flexibility, and performance. The inclusion of machine learning improves the system's analytical capabilities by enabling predictive insights, and the cloud-based design allows the framework to scale with increasing data volumes. The system also supports future enhancements such as real-time data ingestion, advanced forecasting techniques, and analysis of multiple commodities.

Index Terms— Big Data Analytics, Financial Forecasting, PySpark, Snowflake, Random Forest Regression, Cloud Computing, Time-Series Analysis, Predictive Analytics, Power BI

I. INTRODUCTION

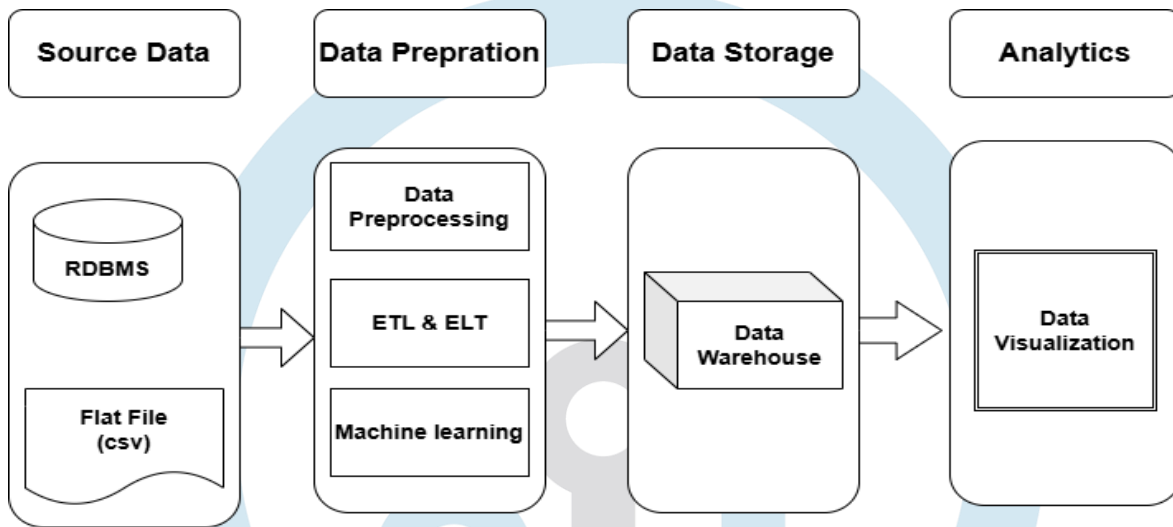
Financial time-series forecasting plays a crucial role in understanding market trends and supporting investment decisions. With the rapid growth of financial data, traditional statistical methods are no longer sufficient to handle large-scale datasets efficiently or capture complex patterns present in the data. In recent years, advancements in cloud computing and machine learning have enabled the development of scalable and efficient forecasting systems. Distributed data processing frameworks such as PySpark allow handling of large datasets, while cloud platforms like Amazon S3 and Snowflake provide reliable storage and fast data access. This paper proposes a scalable big data analytics framework for financial forecasting that integrates distributed processing and machine learning techniques. The system is designed to collect, preprocess, and analyze financial time-series data efficiently. A Random Forest Regression model is used to generate accurate predictions based on historical data patterns. The results are visualized using Power BI dashboards, allowing users to easily interpret trends and make informed decisions. The proposed system improves scalability, processing speed, and prediction accuracy, making it suitable for real-world financial applications.

II. SYSTEM OVERVIEW

The proposed financial forecasting system is designed as a scalable and integrated framework that combines cloud computing, distributed data processing, and machine learning techniques. The system begins with **data collection** from financial sources such as Yahoo Finance. The collected data is stored in a cloud-based storage system using Amazon S3. This ensures scalability and efficient handling of large datasets. Next, the data undergoes **preprocessing** using AWS Glue with PySpark, where operations such as data cleaning, transformation, and feature engineering are performed. The processed data is then stored in **Snowflake**, which acts as a cloud data warehouse for efficient querying and analysis. A **Random Forest Regression model** is applied to the processed data to generate predictions of future price trends. Finally, the results are visualized using **Power BI**

dashboards, enabling users to easily interpret the output and make informed decisions. This architecture ensures efficient data flow, improved scalability, and better prediction performance.

Fig: System Architecture



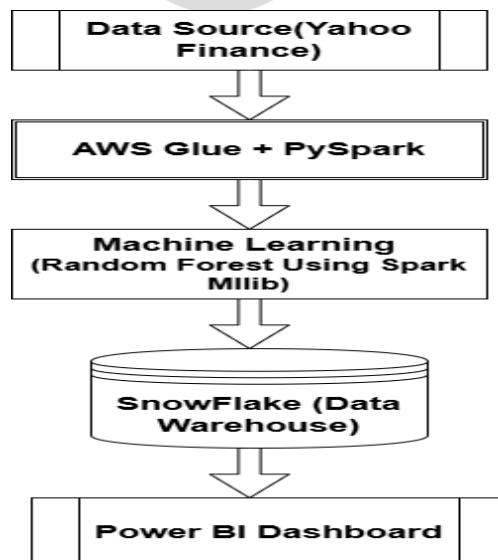
III. METHODOLOGY

The proposed system follows a structured methodology to process financial data and generate accurate predictions. The workflow consists of multiple stages including data collection, preprocessing, feature engineering, model training, and visualization. Initially, financial data is collected from reliable sources such as Yahoo Finance. The collected data is stored in a cloud-based storage system to ensure scalability and efficient data management. In the preprocessing stage, the data is cleaned and transformed using PySpark. This includes handling missing values, removing duplicates, and converting the data into a suitable format for analysis. Feature engineering techniques are then applied to generate meaningful attributes such as previous price values and price differences, which help improve model performance. The processed data is used to train a Random Forest Regression model, which is capable of capturing complex relationships in financial time-series data. The model generates predictions based on historical trends. Finally, the results are visualized using Power BI dashboards, allowing users to analyze trends and make informed decisions.

Fig Data Flow Diagram

IV. PROPOSED SYSTEM

The proposed system is designed as a cloud-based financial forecasting framework that integrates big data technologies and machine learning techniques to improve prediction accuracy and scalability. The system begins with data collection, where historical financial data is obtained from reliable sources such as Yahoo Finance. This data is stored in Amazon S3, which acts



as a scalable data lake capable of handling large volumes of financial data efficiently. In the next stage, the collected data undergoes preprocessing using AWS Glue with PySpark. During this process, data cleaning, transformation, and feature engineering are performed to ensure that the dataset is consistent, accurate, and suitable for analysis. Important features such as previous price values and price changes are generated to capture temporal dependencies in the data. The processed data is then stored in Snowflake, a cloud-based data warehouse that enables efficient data management and fast querying. A Random Forest Regression model is applied to the processed data to predict future price trends based on historical patterns. Finally, the prediction results are visualized using Power BI dashboards, allowing users to easily interpret trends and make informed decisions. Overall, the proposed system provides improved scalability, faster data processing, and higher prediction accuracy compared to traditional financial forecasting methods

Mathematical Model

1. Random Forest Prediction Formula

The Random Forest model predicts the output by averaging the results of multiple decision trees:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N T_i(x)$$

Where:

- \hat{y} = Predicted value
- N = Number of trees
- $T_i(x)$ = Output of the i^{th} decision tree

2. Mean Absolute Error (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Where:

- y_i = Actual value
- \hat{y}_i = Predicted value
- n = Number of observations

3. Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

4. Feature Representation

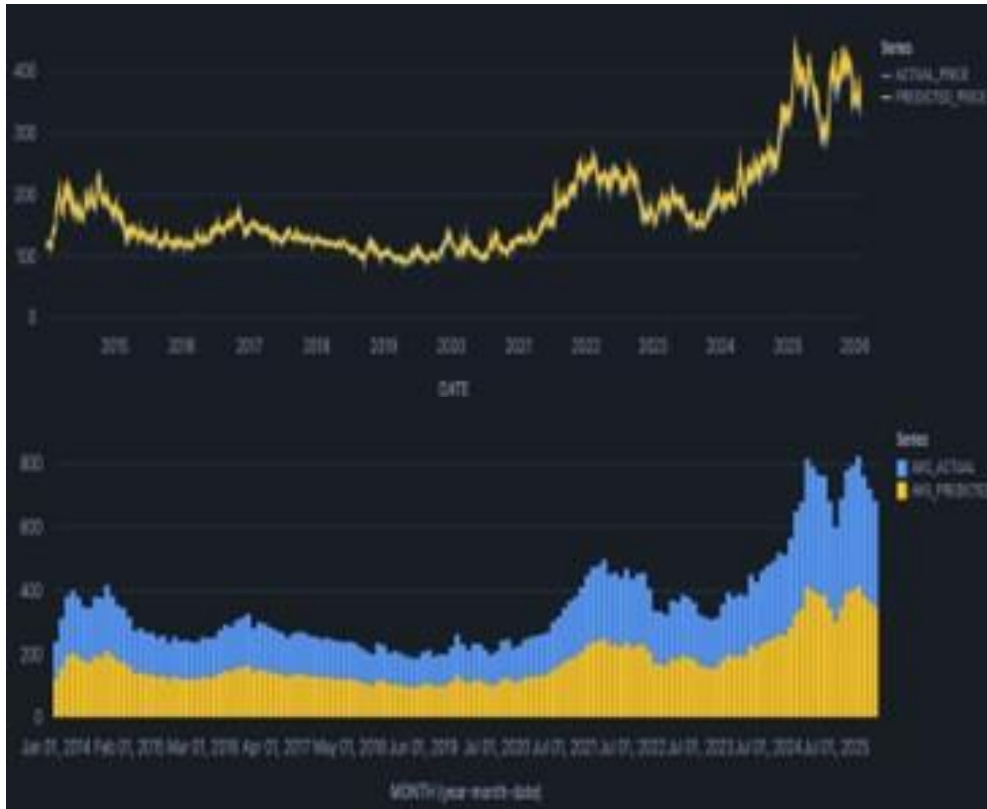
$X = [Open, High, Low, Volume, Previous Price, Price Change]$

$Y = Close Price$

V. RESULTS AND DISCUSSION

The proposed system was tested using historical financial data to evaluate its performance in predicting future price trends. The Random Forest Regression model demonstrated strong performance in capturing nonlinear patterns present in time-series data. The model was evaluated using standard metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), which indicated improved prediction accuracy compared to traditional approaches. The use of PySpark for distributed data processing significantly reduced computation time, making the system efficient for handling large-scale datasets. Additionally, the integration of Snowflake enabled faster data querying and better data management. The prediction results were visualized using Power BI dashboards, which provided clear insights into market trends and forecasted values. These visualizations help users easily interpret the results and support better decision-making. Overall, the system demonstrates high

scalability, improved processing efficiency, and reliable prediction performance, making it suitable for real-world financial forecasting applications.



VI. ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Department of Computer Science and Engineering (Data Science), Rajeev Gandhi Memorial College of Engineering and Technology (RGM CET), Nandyal, for providing the necessary infrastructure, resources, and academic support to successfully carry out this research work. We extend our heartfelt thanks to our project guide, **Y.P. Srinath Reddy**, for his continuous guidance, valuable suggestions, and encouragement throughout the development of this project. His insights and support played a significant role in shaping the direction and quality of this research. The authors also acknowledge the contributions of faculty members and peers who provided constructive feedback and technical assistance during various stages of the project. Their support helped in improving both the implementation and documentation of the work. Finally, we are thankful to all those who directly or indirectly contributed to the successful completion of this research.

REFERENCES

- [1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type," *Phil. Trans. Roy. Soc.*, 1955.
- [2] J. Clerk Maxwell, *A Treatise on Electricity and Magnetism*, Oxford, 1892.
- [3] S. Jacobs and C. P. Bean, "Fine particles and thin films," Academic Press, 1963.
- [4] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [5] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *KDD Conference*, 2016.
- [6] M. Zaharia et al., "Apache Spark: A unified engine for big data processing," *Communications of the ACM*, 2016.
- [7] Amazon Web Services, "Amazon S3 Developer Guide," 2023.

[8] Snowflake Inc., “Snowflake Cloud Data Platform,” 2023.

[9] J. Brownlee, Machine Learning for Time Series Forecasting, 2018.

[10] R. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice, 2018.

[11] Microsoft, “Power BI Documentation,” 2023.

[12] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.

