

# The Investigation of the Application of Apache Spark in Stock Analysis

Yuxi Liu<sup>a</sup>

Department of Computing, Chengdu University of Technology, Chengdu, China

**Keywords:** Apache Spark, Streaming Data, Resilient Distributed Dataset, Hadoop Distributed File System.

**Abstract:** Due to the complexity and scale of data increasing rapidly in nowadays, and the requirements of effectiveness and accuracy to stock market analysis on processing data. Some ways based on Apache Spark become widely accepted by a lot of financial companies and organizations. This paper summarized two useful and satisfactory methods of how to enhance the accuracy and reliability of stock market forecast. For Nowcasting the financial time series with streaming data analytics under Apache Spark, this method integrates Apache Spark and various real-time data, through monitoring these data, system can recognise the trends, then put the results into model training to get more rigorous model. In Sentiment analysis and machine learning model, through combining Resilient Distributed Dataset (RDD) and Hadoop Distributed File System (HDFS), efficient data preprocessing, feature extracting and model training can be achieved. Furthermore, Resilient Distributed Dataset as the core of Apache Spark, provides memory management and fault tolerance to it. Meanwhile, the Hadoop Distributed File System offers a dependable method for distributed storage of large-scale textual data. The integration of Resilient Distributed Datasets and the Hadoop Distributed File System significantly enhances the accuracy of analytical outcomes. In conclusion, this paper demonstrates how forecasting financial time series using streaming data analytics within Apache Spark, alongside sentiment analysis and machine learning models, enhances the reliability and precision of stock market analyses. These approaches contribute to making the results of stock analyses more trustworthy and accurate for users.


## 1 INTRODUCTION

With the increasing of the financial market and the explosive growth of data, the stock market forecasting has become one of the popular topics in the financial field. Moreover, with the coming of the era of big data, the scale of data is growing fast, and the demand for big data analysis and processing is becoming higher and higher. Apache spark is a fast, versatile big data engine, which has become one of the choices for lots of financial enterprises and organizations.

In recent years, the research of stock market prediction based on big data has gradually increased. Researchers use machine learning algorithms and statistical analysis to analyse historical stock data to predict future stock price movements. However, the traditional data processing and analysis methods often face the problems of low computing efficiency and long processing time in the face of large-scale stock data.

In the past few decades, the research of stock predictive analysis has made progress. Earlier studies mainly relied on traditional econometric methods such as time series analysis and statistical regression. These methods are mainly based on historical price data and use the statistical properties of time series to build forecasting models. However, these methods are often limited in forecasting accuracy in the face of complex and non-linear stock markets (Lo, 2004).

With the rise of big data and machine learning technology, the research of stock predictive analysis has entered a new stage. Researchers are beginning to use large-scale historical data, combined with machine learning algorithms such as Support Vector Machines (SVMS), Random Forest, and deep learning, to build more powerful predictive models. These models can better capture the complexity and nonlinear relationship of the stock market and improve the accuracy of the forecast (Zhang, Liu and Yu, 2019). In addition, some studies have explored hybrid forecasting models that combine fundamental and technical analysis. These models combine a

<sup>a</sup> <https://orcid.org/0009-0005-7803-9946>

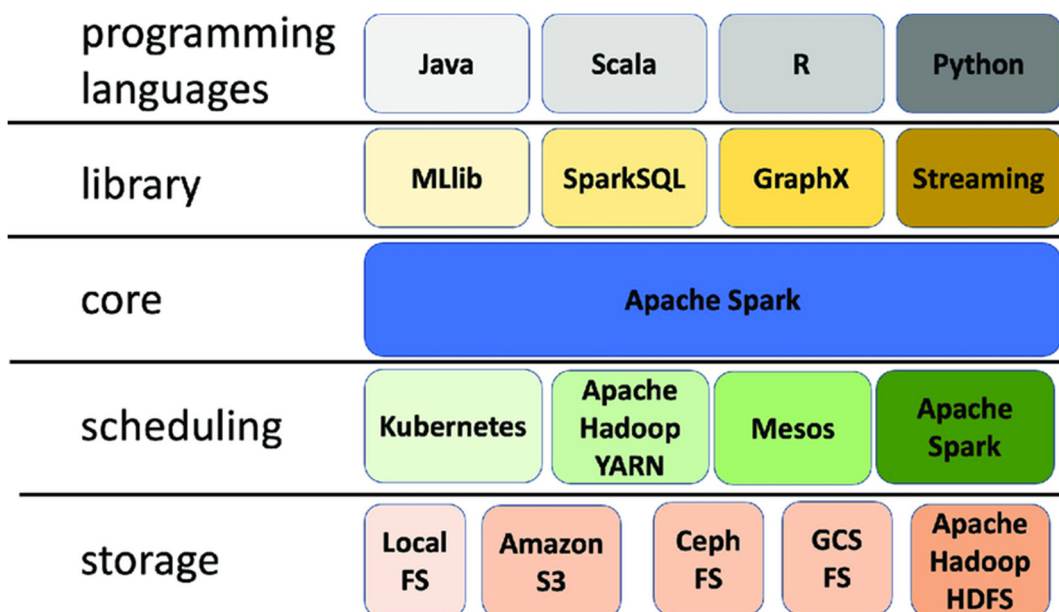


Figure 1: The Apache Spark layered architecture (Chicco, 2023).

company's financial data, market trends, investor sentiment and other multi-dimensional information to provide a more comprehensive predictive analysis. These methods not only improve forecasting accuracy, but also provide investors with deeper market insight (Neely, Rapach and Tu, 2014). Despite the continuous advances in techniques and methods of stock predictive analysis demonstrated in past papers, many challenges and problems remain. The impact of data quality and integrity on the prediction results, the generalization ability of the model, and overfitting problems are still key problems that need to be solved (Liu, Zhou and Xiao, 2021).

Apache Spark is a distributed computing framework, Apache Spark can process massive data, and provides a machine learning library and data processing tools, which provides a new solution for stock market prediction, such as combining Apache Spark with financial time series with streaming data analytics (Khan et al, 2022) and combining Apache Spark with the data that generated from historical price and social events (Seif, Ramzy Hamed and Abdel, 2018).

In this paper, there are two methods Nowcasting the financial time series with streaming data analytics under Apache Spark and Sentiment analysis and machine learning model will be mentioned in section 2, to give more details to readers. What is more, the section 3 is discussion, talking about some current limitation, challenges, possible solutions of limitations and future prospects in stock and Apache

Spark field. Followed by discussion section is conclusion.

## 2 METHODS

In this section, nowcasting with streaming data analysis and sentiment analysis for stock data based on Apache spark were investigated.

### 2.1 Nowcasting the Financial Time Series with Streaming Data Analytics Under Apache Spark

Nowcasting is a term of market analysis, which helps a market predict future state by high frequency data (Khan et al, 2022). In today's big data era, the practicality and relevance have become more obvious. With the rapid generating and updating data of market, to control the dynamics of the market in real time and accurately to make wise invest decision, and Nowcasting is the key technology to meet this demand.

Spark streaming can handle a large amount of live data resultful and spark streaming is the core extension of Spark Application Programming Interface (API), which is suitable to process live data stream like unstructured data, semi-structured data and highly structured data in a high throughput, fault tolerance and highly scalable manner (Khan et al, 2022), which make Spark Streaming become ideal choice of handling real-time in stock market.

In Khan, A. A. M., et al (2022)' s research discussed the approach of how to combine Apache spark data analysis with nowcasting. In this research, Spark Streaming is responsible for collecting data and putting collected data into different batches. Using Lasso, Ridge, RF, GBT and GLM models to test data, researchers found that in the index for Symmetric Mean Absolute Percentage Error, GLM is superior to other methods. Then Spark Streaming come with the results of stream data by batch processing (Khan et al, 2022), which develops the efficiency of processing data, and also more accuracy and more timely information can be provided to investors.

### 2.2 Sentiment Analysis and Machine Learning Model

Seif et al. found that traditional stock analysis only using historical price does not have the ability of emergency handling, in other words, traditional stock market analysis cannot guarantee the accuracy when real market events taken place. So, combining sentiment analysis with machine learning can help improve accuracy (Seif, Ramzy Hamed and Abdel, 2018).

This model is combined with Hadoop Distributed File System, Apache Spark framework and Resilient Distributed Dataset. Hadoop Distributed File System is responsible for storing data, Apache Spark working on processing data and Resilient Distributed Dataset is to parallelize data processing (Seif, Ramzy Hamed and Abdel, 2018). Apache spark mainly works on Information acquisition stage, data storage stage, and data analysis stage. In data Acquisition Phase, Apache Spark uses its ability of data integration, get data from different sources efficiently. And use different models to test data, then single out the most accurate model Random Forest, which accuracy is 90.23, which is higher than traditional stock analysis to predict data. Thanks to Apache Spark can store data in worker's memory, which ensures the integrity and accuracy of data (Seif, Ramzy Hamed and Abdel, 2018).

In data analysis phase, data is deposited in Hadoop Distributed File System, which provides data redundancy and error tolerance and has the capability to process large data (Shvachko, 2010). Then, achieve efficient parallelize data processing (Seif, Ramzy Hamed and Abdel, 2018). Apache Spark combines machine learning and sentiment analysis to enhance stock market analysis accuracy. In Seif, Ramzy Hamed and Abdel's research, mentioned that in offline database, the accuracy of data mining tools with sentiment analysis compared with without

sentiment analysis were improved by 0.5 on average. In real-time database, the accuracy of data mining tool with sentiment analysis increased nearly 2 on average. The MLlib provides abundant algorithms and tools, which offers support for stock market predictions and decisions (Seif, Ramzy Hamed and Abdel, 2018). What is more, sentiment analysis technology can catch investors' perspective and expectation of market. Through the combination of sentiment analysis and machine learning, the trend prediction of market will be more and more accuracy. Furthermore, Resilient Distributed Dataset also plays an important role in Apache Spark (Seif, Ramzy Hamed and Abdel, 2018). It allows Spark to parallelize processing data on different nodes to guarantee stability and performance (Zaharia, 2012).

### 3 RESULTS AND DISCUSSIONS

This section focuses primarily on analysing the experimental results of methods mentioned in the last section and challenges they face and the future prospects. The two ways, combination of nowcasting techniques and the strong data processing ability of Apache spark and Sentiment analysis and machine learning model, which are scalable and flexible offer advantages for stock market analysis. Investors can make full use of the instantaneity of high-frequency and the ability of batch processing in Spark Streaming, to make the most sensible decision.

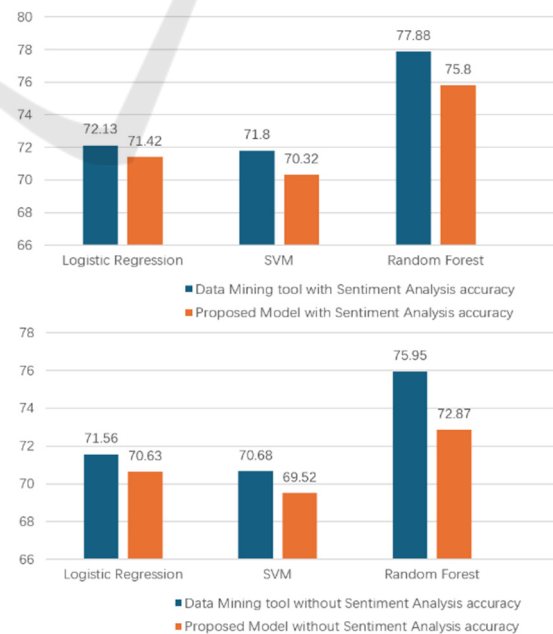


Figure 2: The performance of different models (Seif, Ramzy Hamed and Abdel, 2018).

Furthermore, the models can meet different complexity and scale of stock analysis market requirement. Specifically, for the more complex data come from BSE, NSE and BTC-INR, the SMAPE value of GLM are 0.06754, 0.06497 and 0.04921, which stand for more reliable and closer prediction to actual data, then leads investors can get more valuable information, so as to achieve revenue maximization and stand out in drastic marketing competition.

Based on the Seif's et al. studies (Seif, Ramzy Hamed and Abdel, 2018), the first bar chart compares the accuracy of two models for sentiment analysis using Logistic Regression, SVM, and Random Forest algorithms. The 'Data Mining tool with Sentiment Analysis' shows slightly lower accuracy rates (72.13%, 71.42%, and 77.88%, respectively) compared to the 'Proposed Model with Sentiment Analysis' (71.8%, 70.32%, and 75.8%, respectively). The second chart contrasts the accuracy of the same models without sentiment analysis. Here, the 'Data Mining tool' has higher accuracies for Logistic Regression and SVM (71.56%, 70.63%) but a lower one for Random Forest (72.87%) compared to the 'Proposed Model' (70.68%, 69.52%, and 75.95%, respectively). This demonstrated the importance of sentiment analysis in the domain of stock price prediction.

With the constant development of artificial intelligence and machine learning, people could foresee a more automated and intelligent stock market analysis. Through using the Apache Spark to analyse streaming live data and the combination of the Nowcasting technique and advanced machine learning algorithm (Das, 2024; Qiu, 2024), developers can build more stronger and accuracy prediction model, which provide a deeper market insight and more personalized investment advice. Moreover, the constant evolution of distributed computing frames like the Apache Spark will give more support on processing on the larger scale and more complex dataset.

But at the same time, development comes with limitations and challenges, such as interpretability and data skew. For interpretability, it is a one of important indicators to evaluate machine learning model, which makes model more reliable and explains how model works. So, add Shapley Value (SHAP) is a good choice to help ameliorate interpretability (Jia, 2019; Sundararajan, 2020). Shapley Value is an explanatory method based on game theory, which is utilized to measure the weight of each feature to the model analysis results. And it evaluates the importance of feature by calculating the average contribution of a feature across all possible

feature subsets, which makes the results have uniqueness, local accuracy and consistency. Furthermore, data skew is a normal problem in distributed computing environment. In stock analysis, if some stocks trade much more than others, then it is possible to meet data skew. Preprocessing data and adjusting parallelization are two possible ways to figure out this problem. Preprocessing data can filter out some keys that skew the data. Like if there are lot of useless null values, then these values can be filtered out before shuffle. To adjust parallelize, which means to adjust the parallelize of shuffle, the reduce tasks quantity can be increased by doing this, then release the data skew. The future prospects of Apache Spark in stock analysis will be brighter and brighter. The tools in Apache Spark like MLlib, Spark SOL and GraphX make it can handle more complex data in stock market analysis, Apache Spark's potential should be efficiently used in stock market analysis.

## 4 CONCLUSIONS

In this work, in order to improve the accuracy in stock market analysis, two ways under Apache Spark and machine learning are proposed. The Nowcasting the financial time series with streaming data analytics under Apache Spark and Sentiment analysis and machine learning model are two effective ways to enhance the reliability in stock market. Nowcasting technology offset the inadequacy of Apache Spark processing real-time data and by using the strong streaming data analysis ability, more larger data scale can be handled easily and rapidly. The sentiment analysis assists machine learning model to recognize emotional tendency, which provides greater advantage on catch the market by referring the events happening in real-time that may cause volatility to stock market. In the future, we need to optimize the performance of Apache Spark by making research on interpretability and data skew then deal with these kinds of problems which could be barriers on enhancing development of model.

## REFERENCES

- Chicco, D., Ferraro Petrillo, U., & Cattaneo, G. 2023. Ten quick tips for bioinformatics analyses using an Apache Spark distributed computing environment. *PLOS Computational Biology*, 19(7), e1011272.
- Das, N., Sadhukhan, B., Chatterjee, R., & Chakrabarti, S. 2024. Integrating sentiment analysis with graph neural

- networks for enhanced stock prediction: A comprehensive survey. *Decision Analytics Journal*, 100417.
- Jia, R., Dao, D., Wang, B., Hubis, F. A., Hynes, N., Gürel, N. M., ... & Spanos, C. J. 2019. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics* (pp. 1167-1176). PMLR.
- Khan, M. A. A., Bhushan, C., Ravi, V., Rao, V. S., & Orsu, S. S. 2022. Nowcasting the financial time series with streaming data analytics under apache spark. *arXiv preprint arXiv:2202.11820*.
- Liu, Y., Zhou, X., and Xiao, J. 2021. Stock price prediction based on deep learning: A literature review. *Journal of Big Data*, 8(1), 1-23.
- Lo, A. W. 2004. The adaptive markets hypothesis. *Journal of Portfolio Management*, 30(5), 15-29.
- Neely, C. J., Rapach, D. E., & Tu, J. 2014. Forecasting the equity risk premium: The role of technical indicators. *Management Science*, 60(7), 1772-1791.
- Qiu, Y., Hui, Y., Zhao, P., Cai, C. H., Dai, B., Dou, J., ... & Yu, J. 2024. A novel image expression-driven modeling strategy for coke quality prediction in the smart cokemaking process. *Energy*, 130866.
- Seif, M. M., Ramzy Hamed, E. M., & Abdel Ghfar Hegazy, A. E. F. 2018. Stock market real time recommender model using apache spark framework. In *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018)* (pp. 671-683). Springer International Publishing.
- Shvachko, K., Kuang, H., Radia, S., & Chansler, R. 2010. The hadoop distributed file system. In *2010 IEEE 26th Symposium on mass storage systems and technologies (MSST)* (pp. 1-10). IEEE.
- Sundararajan, M., & Najmi, A. 2020. The many Shapley values for model explanation. In *International conference on machine learning* (pp. 9269-9278). PMLR.
- Zaharia, Matei, et al. 2012. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*.
- Zhang, G., Liu, H., & Yu, L. 2019. Stock price prediction using statistical and machine learning techniques: A review and evaluation. *Journal of Financial Markets*, 22(1), 1-28.