



STOCK MARKET ANALYSIS USING MAPREDUCE AND PYSPARK

Dr. X. Stellamary

Head & Assistant Professor of Commerce Maria Arts and Science College for Women, Vallioor
(Affiliated to Manonmaniam Sundaranar University, Tirunelveli)

ABSTRACT :

Stock market analysis has become increasingly important for understanding market movements, investment behaviour and financial risk. The rapid growth of financial data has created a need for scalable computational methods capable of processing large datasets efficiently. This study presents an analytical framework that combines Hadoop MapReduce and PySpark for stock volatility analysis, sentiment analysis and stock price prediction. A large stock dataset is examined to identify securities with relatively high and low volatility. MapReduce is used to calculate monthly returns and volatility, while PySpark and machine-learning techniques are applied to examine price movements. News information is also incorporated to study the relationship between market sentiment and stock behaviour. The framework demonstrates how big-data technologies can support financial analysis by processing historical market information and unstructured news data. The study highlights the usefulness of distributed computing and machine learning for analysing complex stock-market information, while recognising that market behaviour is influenced by numerous changing economic and financial factors.

Keywords: Stock Market, Big Data Analytics, Hadoop MapReduce, PySpark, Stock Volatility, Sentiment Analysis, Machine Learning, Stock Price Prediction.

1. INTRODUCTION

The expansion of digital information has significantly increased the volume of data available to financial institutions, businesses and investors. Financial markets generate large quantities of structured and unstructured information through stock prices, trading volumes, economic indicators, financial reports and news publications. Analysing these sources requires computational techniques that can manage large datasets while producing useful information within a reasonable period.

Big-data technologies provide an effective environment for processing such information. Hadoop MapReduce supports distributed processing of large datasets, whereas Apache Spark and PySpark provide faster in-memory computation and machine-learning capabilities. These technologies can be applied to financial datasets to examine volatility, identify patterns and develop predictive models.

The present study develops an integrated approach for stock-market analysis. The framework focuses on three major components: stock-volatility measurement using Hadoop MapReduce, sentiment analysis of financial news, and stock-price prediction using PySpark-based machine-learning methods.

2. PRELIMINARY KNOWLEDGE

2.1 Hadoop MapReduce

Hadoop MapReduce is a distributed programming framework designed to process very large datasets across multiple computing nodes. The input data are divided into smaller portions and processed in parallel by mapper tasks. The resulting intermediate information is then grouped and processed by reducer tasks. This architecture allows large-scale calculations to be performed efficiently and provides mechanisms for task scheduling, monitoring and recovery from processing failures.

For stock-market analysis, MapReduce can process numerous stock files and calculate return-related measures for individual securities. The distributed structure is particularly useful when the dataset contains information for a large number of stocks and trading periods.

2.2 PySpark

PySpark is the Python interface for Apache Spark. It enables users to perform distributed data processing and supports capabilities such as Spark SQL,

DataFrames, streaming and machine learning. Spark can perform computations in memory and is therefore suitable for analytical tasks involving large datasets.

PySpark also provides machine-learning functionality through MLlib. Classification, regression, clustering and other analytical methods can be applied to financial datasets after appropriate preprocessing and feature preparation.

3. STOCK MARKET

3.1 Overview

The stock market represents a network of organised exchanges and other trading venues where securities such as company shares and related financial instruments are bought and sold. Stock exchanges provide a regulated environment through which companies can raise capital and investors can participate in securities trading.

3.2 Market Operations

The primary market enables companies to issue securities to investors, including through initial public offerings. Funds raised through such issues can be used for expansion and other business activities. The secondary market enables investors to buy and sell securities that have already been issued.

3.3 Purpose and Functions

Stock markets perform important economic functions by connecting companies seeking capital with investors seeking investment opportunities. Market liquidity allows securities to be traded more easily. Stock prices are influenced by economic conditions, company-specific information, investor expectations, interest rates, exchange rates and other factors.

3.4 Stock Volatility

Volatility represents the degree to which the price or return of a security changes over a specified period. Standard deviation is commonly used to quantify volatility. A higher volatility value indicates greater variation in returns, while a lower value indicates relatively smaller fluctuations. Volatility is therefore an important measure when analysing the risk characteristics of securities.

4. OBJECTIVES OF THE STUDY

- To analyse stock-price volatility using Hadoop MapReduce.
- To examine the relationship between financial news sentiment and stock movements.
- To develop a machine-learning approach for stock-price prediction using PySpark.
- To evaluate the usefulness of big-data technologies in financial-market analysis.

5. STOCK VOLATILITY ANALYSIS USING HADOOP MAPREDUCE

5.1 Dataset

The analysis uses stock-market information containing records for a large number of securities. The source dataset consists of separate CSV files containing fields such as trading date, opening price, highest price, lowest price, closing price, trading volume and adjusted closing price. The adjusted closing price and date are used for calculating stock returns and subsequent volatility measures.

5.2 Methodology and Implementation

The volatility-analysis process is organised into multiple MapReduce stages. In the first stage, mapper tasks read stock records and associate each observation with its corresponding stock identifier. The reducer groups the observations for each stock and calculates periodic rates of return.

A subsequent MapReduce stage uses the calculated returns to determine the volatility of each stock. The volatility measure is obtained from the dispersion of the periodic returns. After the volatility values are generated, another mapping stage reorganises the data so that securities can be ordered according to their volatility. The final reducer identifies the securities at the lower and upper ends of the volatility distribution.

This process provides a scalable method for screening a large stock dataset and identifying securities that exhibit comparatively greater or smaller price fluctuations.

6. STOCK SENTIMENT ANALYSIS

Financial news can provide information about events and expectations that may influence investor behaviour. Sentiment analysis converts textual information into measurable indicators that can be examined alongside historical stock data.

6.1 Data Collection

The study combines historical stock-market information with financial news. Stock-price information can be obtained from financial-data providers, while news articles can be collected from recognised news aggregators and financial news sources. The resulting dataset contains both numerical market information and unstructured text.

6.2 Text Pre-processing

News articles require preprocessing before they can be used in a machine-learning model. The main steps include tokenisation, removal of unnecessary terms, elimination of punctuation and other noise, and stemming or normalisation of related words. Domain-specific stop words can also be excluded so that the model concentrates on meaningful financial terms.

6.3 Deep Learning Approach

A recurrent neural-network approach can be used to process sequences of words in financial news. Long Short-Term Memory (LSTM) networks are particularly suitable for sequential text because their gated architecture can retain relevant information over longer sequences. The model can classify news into positive and negative sentiment categories and produce probability values for the respective classes.

6.4 Model Evaluation

The sentiment model can be evaluated using training and validation performance. Monitoring both training and validation results is important because a model may achieve high training performance while failing to generalise to unseen information. Such behaviour is commonly associated with overfitting.

7. STOCK MARKET PREDICTION USING PYSARK

7.1 Data Collection

Stock prediction requires historical observations containing variables such as date, opening price, highest price, lowest price, closing price and trading volume. Additional economic and market indicators may also be incorporated when they are relevant to the prediction objective.

7.2 Data Pre-processing

The collected data are inspected for missing or invalid observations. Records with a small proportion of missing information may be removed when such removal does not materially affect the analysis. The variables required for modelling are then selected based on their relevance and statistical relationships.

7.3 Feature Engineering

Machine-learning algorithms require numerical representations of the input variables. Categorical variables, when present, can be transformed through techniques such as label encoding or one-hot encoding. Numerical features can be assembled into a single feature vector using Spark's VectorAssembler so that the resulting dataset can be supplied to machine-learning algorithms.

7.4 Training and Testing

The dataset is divided into training and testing portions. The training data are used to develop the predictive model, while the testing data are retained for evaluating performance on observations that were not used during model development.

7.5 Machine-Learning Model

A decision-tree-based regression or classification approach can be implemented using Spark MLlib. The model learns relationships between selected market indicators and the target variable. Other techniques, including support vector machines, logistic regression and neural-network models, can also be considered for comparison.

7.6 Model Evaluation

Prediction performance can be assessed using suitable evaluation measures. Mean Absolute Error (MAE) measures the average magnitude of prediction errors, while the coefficient of determination (R^2) indicates the proportion of variation in the dependent variable explained by the model. Using more than one metric provides a broader view of model performance.

8. CORRELATION BETWEEN STOCK PRICES AND SENTIMENT

After the market and sentiment models are developed, the relationship between news sentiment and stock-price movements can be examined. Positive coefficients indicate a positive statistical association between variables, whereas negative coefficients indicate an inverse association. Correlation analysis can therefore provide an initial indication of whether sentiment measures are associated with changes in stock prices. Such associations should be interpreted carefully because correlation does not by itself establish causation.

9. LIMITATIONS OF THE STUDY

- Stock prices are affected by numerous economic, political, company-specific and market-related factors that may not be represented in the model.
- Financial markets are dynamic, and relationships identified from historical observations may change over time.
- Machine-learning models may be affected by data quality, feature selection, parameter settings and overfitting.
- News sentiment may not capture every factor that influences investor decisions or market movements.
- Prediction results should be interpreted as analytical outputs rather than guarantees of future investment performance.

10. FUTURE SCOPE

- Future studies can use larger and more diverse datasets covering longer time periods.
- Real-time market feeds and streaming news can be incorporated into the analytical framework.
- Advanced machine-learning and deep-learning algorithms can be compared using consistent evaluation procedures.
- Additional economic indicators, company fundamentals and market variables can be included as predictive features.
- The framework can be extended to real-time monitoring of market volatility and sentiment.

11. CONCLUSION

The study demonstrates an integrated approach to stock-market analysis using distributed computing, sentiment analysis and machine learning. Hadoop MapReduce can be used to process large collections of stock records and calculate volatility measures, while PySpark provides a flexible environment for preprocessing data and developing predictive models. Incorporating financial-news sentiment adds an additional information source for examining market movements.

The combined framework illustrates the potential of big-data technologies for handling complex financial datasets. At the same time, stock-market prediction remains challenging because prices are influenced by changing and interacting factors. Therefore, future research should focus on robust datasets, careful model validation, real-time information and comparative evaluation of alternative analytical techniques.

12. REFERENCES :

- Kavya, P., Saagarika, S., Subavarshini, R. T., Nivetheni, C. G., Marimuthu, M., & Velvadivu, P. (2021). Stock Market Analysis Using MapReduce and PySpark. *International Research Journal of Modernization in Engineering Technology and Science*, 3(7).
- Apache Hadoop documentation. MapReduce framework and distributed data processing concepts.
- Apache Spark documentation. PySpark and Spark ML documentation.
- Investopedia. Stock market and financial-market concepts.
- Relevant academic literature on financial sentiment analysis, stock-market prediction and machine learning.