Report: Big Data Stock Price Analysis

Analysis of VNStock Data Using Hadoop & Spark

Data Source: VNStock (2015–2025)

Author: Trinh Tuan Ngoc Bao

Institution: University of Engineering and Technology – Big Data

Date: October 25, 2025

Contents

Abstract

This report presents the development of a distributed Big Data analytics system for analyzing Vietnamese stock market data from VNStock. The project demonstrates an end-to-end data pipeline including data collection, storage, preprocessing, and distributed analysis on a Docker-based Hadoop–Spark cluster. The system efficiently handles thousands of stock records and performs analytical tasks such as price trend tracking and growth rate comparison across major Vietnamese bank stocks.

1 Introduction

The Vietnamese stock market has expanded rapidly in both transaction volume and market capitalization. Consequently, traditional data processing tools struggle to handle the growing scale and velocity of market data.

This project builds a simulated Big Data environment for analyzing stock prices using VNStock as the primary data source. It integrates the **Hadoop Distributed File System (HDFS)** for scalable storage and **Apache Spark** for distributed computation. The approach demonstrates the advantages of Big Data technologies in managing and analyzing financial time-series data efficiently.

2 Theoretical Background

2.1 Big Data Systems

Big Data refers to datasets with massive volume, variety, and velocity, which require distributed systems for efficient storage and analysis. Frameworks such as Hadoop and Spark enable horizontal scalability and high processing speed.

2.2 Hadoop and HDFS

Hadoop provides distributed data storage through HDFS, allowing large files to be partitioned across multiple DataNodes. A central NameNode manages metadata, ensuring fault tolerance and high availability.

2.3 Apache Spark

Apache Spark is an in-memory data processing engine designed for large-scale distributed computation. It performs transformations and aggregations much faster than traditional MapReduce, making it ideal for iterative analytics tasks such as stock trend analysis.

2.4 Stock Data Characteristics

Stock market data are a form of time-series data — values recorded sequentially over time. They are highly dynamic and influenced by numerous market factors, requiring efficient systems for continuous collection and processing.

3 System Design and Implementation

3.1 System Architecture

The system was deployed using Docker Compose and consists of:

- **HDFS Cluster:** One NameNode and four DataNodes to store distributed data.
- **Spark Cluster:** One Spark Master and four Spark Workers for distributed computation.
- **Jupyter Notebook:** An interactive interface for accessing the Spark cluster.

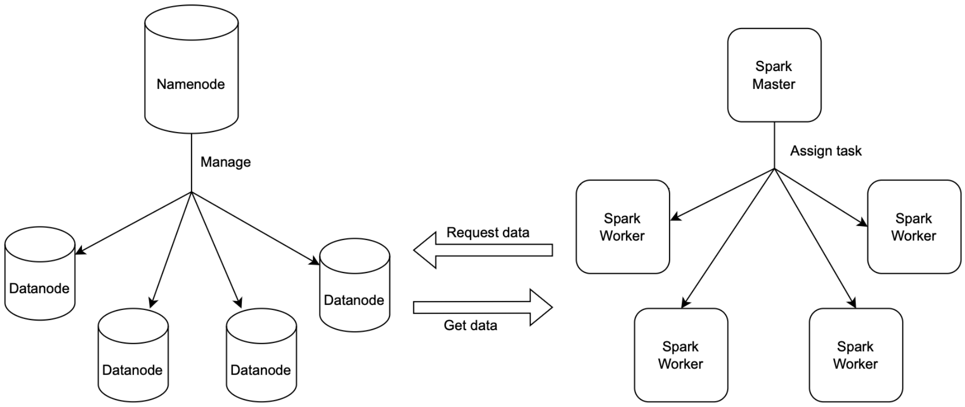


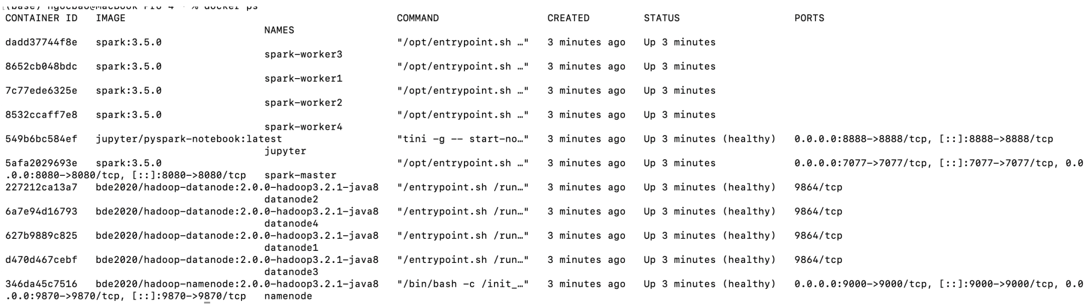
Figure 1: System architecture integrating Hadoop and Spark clusters.

3.2 Cluster Setup

Each node is deployed as a Docker container using the following images:

- `bde2020/hadoop-namenode:2.0.0-hadoop3.2.1-java8`
- `bde2020/hadoop-datanode:2.0.0-hadoop3.2.1-java8`
- `spark:3.5.0` (for both Master and Workers)

- `jupyter/pyspark-notebook:latest`



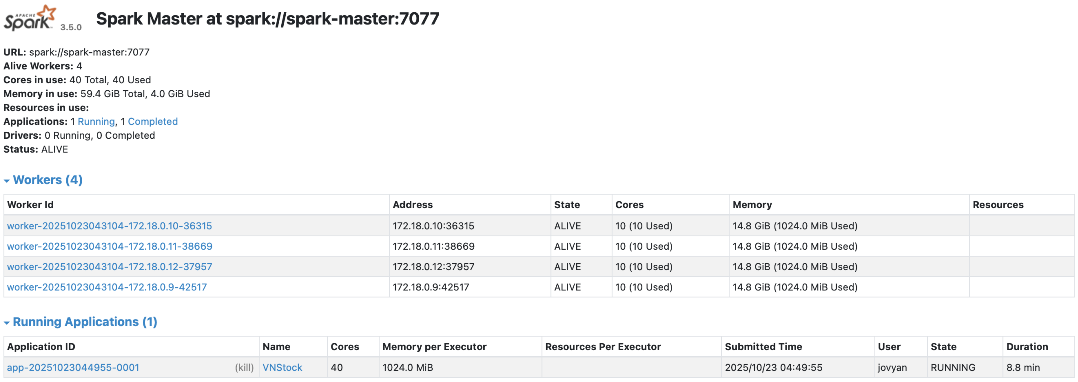
```
(base) ngocbao@MacBook Pro 4 ~ % docker ps
CONTAINER ID   IMAGE                                             COMMAND                  CREATED         STATUS                   PORTS
                                           NAMES
dadd37744f8e   spark:3.5.0                                       "/opt/entrypoint.sh …"   3 minutes ago   Up 3 minutes
                                           spark-worker3
8652cb048bdc   spark:3.5.0                                       "/opt/entrypoint.sh …"   3 minutes ago   Up 3 minutes
                                           spark-worker1
7c77ede6325e   spark:3.5.0                                       "/opt/entrypoint.sh …"   3 minutes ago   Up 3 minutes
                                           spark-worker2
8532ccaff7e8   spark:3.5.0                                       "/opt/entrypoint.sh …"   3 minutes ago   Up 3 minutes
                                           spark-worker4
549b6bc584ef   jupyter/pyspark-notebook:latest                   "tini -g -- start-no…"   3 minutes ago   Up 3 minutes (healthy)   0.0.0.0:8888->8888/tcp, [::]:8888->8888/tcp
                                           jupyter
5afa2029693e   spark:3.5.0                                       "/opt/entrypoint.sh …"   3 minutes ago   Up 3 minutes             0.0.0.0:7077->7077/tcp, [::]:7077->7077/tcp, 0.0
.0.0:8080->8080/tcp, [::]:8080->8080/tcp   spark-master
227212ca13a7   bde2020/hadoop-datanode:2.0.0-hadoop3.2.1-java8   "/entrypoint.sh /run…"   3 minutes ago   Up 3 minutes (healthy)   9864/tcp
                                           datanode2
6a7e94d16793   bde2020/hadoop-datanode:2.0.0-hadoop3.2.1-java8   "/entrypoint.sh /run…"   3 minutes ago   Up 3 minutes (healthy)   9864/tcp
                                           datanode4
627b9889c825   bde2020/hadoop-datanode:2.0.0-hadoop3.2.1-java8   "/entrypoint.sh /run…"   3 minutes ago   Up 3 minutes (healthy)   9864/tcp
                                           datanode1
d470d467cebf   bde2020/hadoop-datanode:2.0.0-hadoop3.2.1-java8   "/entrypoint.sh /run…"   3 minutes ago   Up 3 minutes (healthy)   9864/tcp
                                           datanode3
346da45c7516   bde2020/hadoop-namenode:2.0.0-hadoop3.2.1-java8   "/bin/bash -c /init_…"   3 minutes ago   Up 3 minutes (healthy)   0.0.0.0:9000->9000/tcp, [::]:9000->9000/tcp, 0.0
.0.0:9870->9870/tcp, [::]:9870->9870/tcp   namenode
```

Figure 2: Docker containers running in the virtual Big Data cluster.

This setup allows distributed processing and scalability across multiple nodes in a controlled environment.

3.3 Connecting to the Spark Cluster

We connect and create a Spark application using the kernel token displayed in the Jupyter log. Once connected, the Spark UI confirms active workers and job execution, as shown in Figure 3.



Spark Master at spark://spark-master:7077

URL: spark://spark-master:7077
Alive Workers: 4
Cores in use: 40 Total, 40 Used
Memory in use: 59.4 GiB Total, 4.0 GiB Used
Resources in use:
Applications: 1 Running, 1 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (4)

Worker Id	Address	State	Cores	Memory	Resources
worker-20251023043104-172.18.0.10-36315	172.18.0.10:36315	ALIVE	10 (10 Used)	14.8 GiB (1024.0 MiB Used)	
worker-20251023043104-172.18.0.11-38669	172.18.0.11:38669	ALIVE	10 (10 Used)	14.8 GiB (1024.0 MiB Used)	
worker-20251023043104-172.18.0.12-37957	172.18.0.12:37957	ALIVE	10 (10 Used)	14.8 GiB (1024.0 MiB Used)	
worker-20251023043104-172.18.0.9-42517	172.18.0.9:42517	ALIVE	10 (10 Used)	14.8 GiB (1024.0 MiB Used)	

Running Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20251023044955-0001 (kill)	VNStock	40	1024.0 MiB		2025/10/23 04:49:55	jovyan	RUNNING	8.8 min

Figure 3: Spark UI confirming successful cluster connection.

4 Data Collection

The dataset consists of daily stock prices collected from VNStock for 10 major banks:

VCB, BID, CTG, TCB, MBB, ACB, VPB, HDB, STB, SHB

Table 1: Sample Trading Data of MBB (2015)

time	**open**	**high**	**low**	**close**	**volume**
2015-01-05	2.31	2.31	2.29	2.29	338,650
2015-01-06	2.29	2.38	2.29	2.36	2,865,710
2015-01-07	2.38	2.43	2.36	2.38	2,235,000
2015-01-08	2.38	2.43	2.38	2.40	853,040
2015-01-09	2.40	2.54	2.40	2.52	6,690,100

The dataset includes fields such as *time, open, high, low, close, volume*, spanning the period from 2015 to 2025. Each stock's data was saved as a CSV file and uploaded to HDFS for distributed analysis.

5 Data Preprocessing

Before analysis, the data underwent several cleaning and transformation steps:

- Converted date strings into `datetime` format.
- Removed missing or invalid records.
- Normalized stock symbols and price columns.
- Filtered the dataset to focus on the banking sector for consistent comparison.

These preprocessing steps ensure data consistency and reliability before Spark-based processing.

6 Data Analysis

6.1 Price Trend Analysis

The closing price trend for each bank was analyzed to visualize performance over time. Figure 4 illustrates how selected banking stocks evolved between 2015 and 2025.

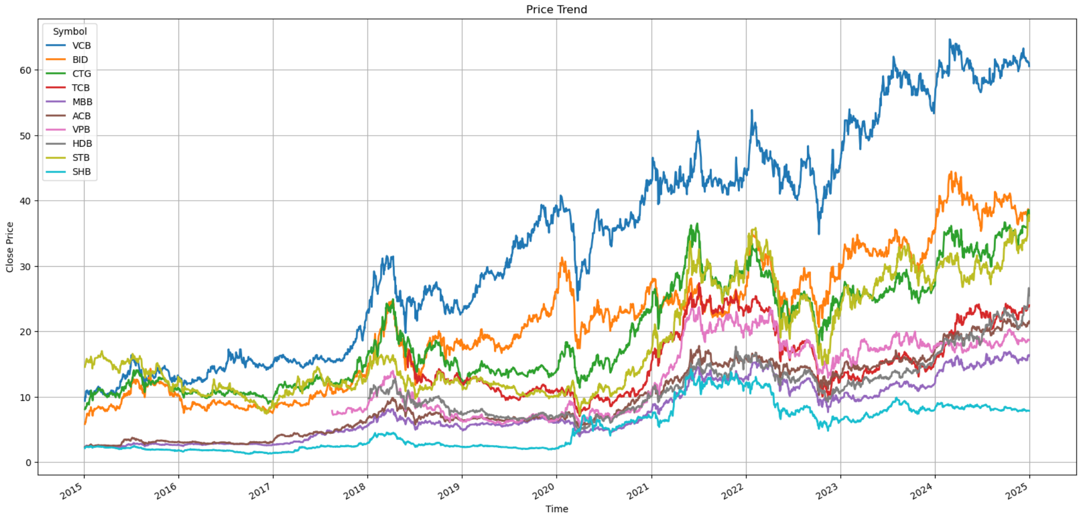


Figure 4: Price trends of major Vietnamese bank stocks (2015–2025).

Discussion: The analysis reveals that most major bank stocks — particularly VCB and BID — experienced steady long-term growth despite short-term market corrections. Stocks such as MBB and ACB demonstrated strong resilience and long-term stability, reflecting solid fundamentals and investor confidence.

6.2 Price Growth Rate Analysis

The two-month price growth rate was computed to evaluate short-term and medium-term volatility among different bank stocks.

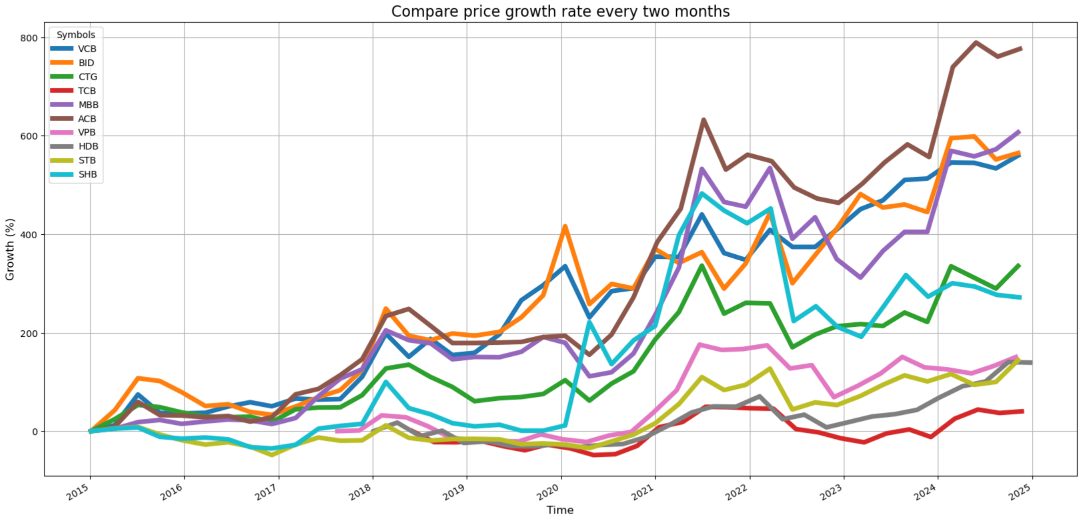


Figure 5: Two-month price growth rate comparison among bank stocks.

Discussion: The growth rate analysis highlights varying levels of volatility. Stocks such as ACB and MBB demonstrate higher fluctuations, implying higher risk but also potential for speculative gains. Conversely, BID and VCB maintain smoother growth trajectories, suggesting stable long-term investment opportunities. These insights can guide investors based on their risk tolerance and return expectations.

7 Prediction

In this section, we present the results of stock price prediction for MBB and other major Vietnamese banks. The prediction model is based on a **Long Short-Term Memory (LSTM)** neural network trained on historical OHLCV data from 2015 to 2025. Data were normalized, sequenced with a 60-day lookback window, and used to predict the next day's closing price.

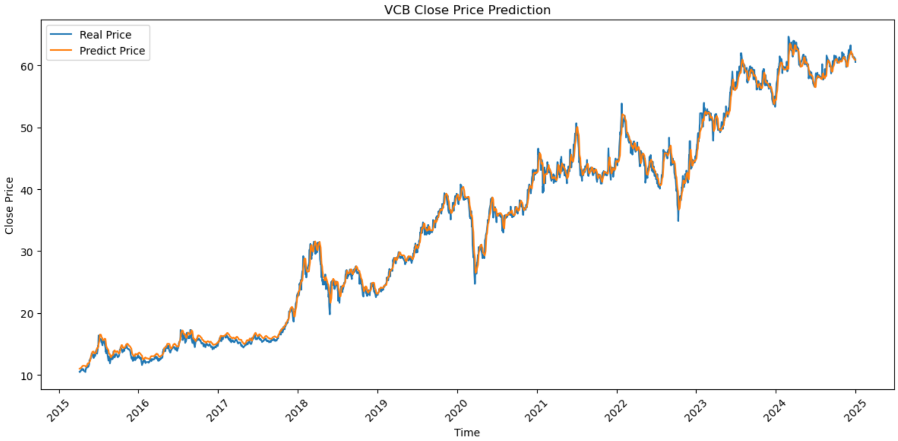


Figure 6: Actual vs Predicted Closing Prices for MBB (LSTM Model).

The model achieves an R^2 score of 0.9968, indicating a strong correlation between predicted and actual prices.

8 Conclusion

This report presents the design and implementation of a Big Data analytics system for analyzing VNStock data using Hadoop and Spark. The system demonstrates efficient data collection, preprocessing, and distributed computation capabilities.

Through analyses of price trends and growth rates, the study identifies differences in market behavior among leading Vietnamese banks. The results confirm that distributed Big Data technologies are effective tools for large-scale financial analytics and can provide valuable insights for both investors and researchers.

GitHub: `https://github.com/ngocbao220/Bigdata-ASM1-StockPrice`