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 23, 2025

Contents

1 Introduction **2**

2 Theoretical Background **2**

2.1 Big Data Systems 2

2.2 Hadoop and HDFS 2

2.3 Apache Spark 2

2.4 Stock Data Characteristics 3

3 System Design and Implementation **3**

3.1 System Architecture 3

3.2 Cluster Setup 3

4 Data Collection **4**

5 Data Preprocessing **4**

6 Data Analysis **4**

6.1 Price Trend Analysis 4

6.2 Price Growth Rate Analysis 5

7 Conclusion **6**

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 **Hadoop Distributed File System (HDFS)** for scalable storage and **Apache Spark** for distributed computation. The approach demonstrates the benefits 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:** 1 NameNode and 4 DataNodes to store distributed data.
- **Spark Cluster:** 1 Spark Master and 4 Spark Workers for distributed computation.
- **Jupyter Notebook:** interactive interface for accessing the Spark cluster.

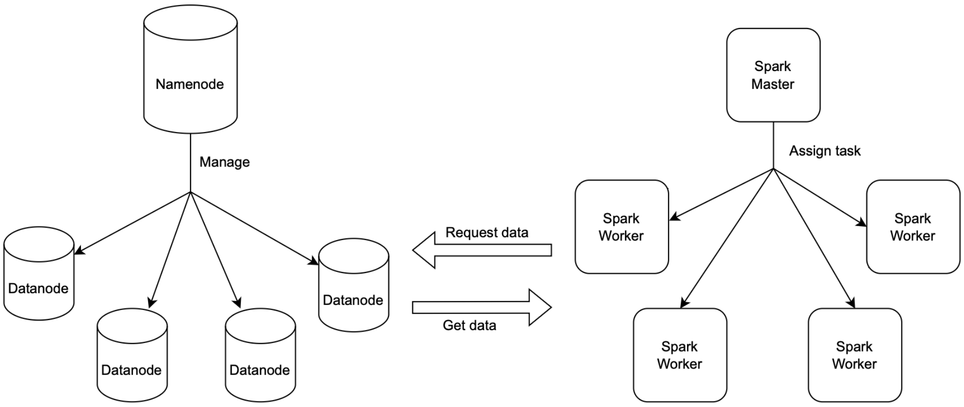

Figure 1: System architecture integrating Hadoop and Spark clusters.

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`

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

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

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

5 Data Preprocessing

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

- Convert date strings into `datetime` format.
- Remove missing or invalid records.
- Normalize stock symbols and price columns.
- Filter 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. The figure below illustrates how selected banking stocks evolved between 2015 and 2025.

Discussion: The analysis reveals that most major bank stocks—particularly VCB, BID a steady increase in price over the long term, despite temporary fluctuations during market corrections. Stocks like VCB and MBB exhibited strong resilience and long-term stability, reflecting solid fundamentals and investor confidence.

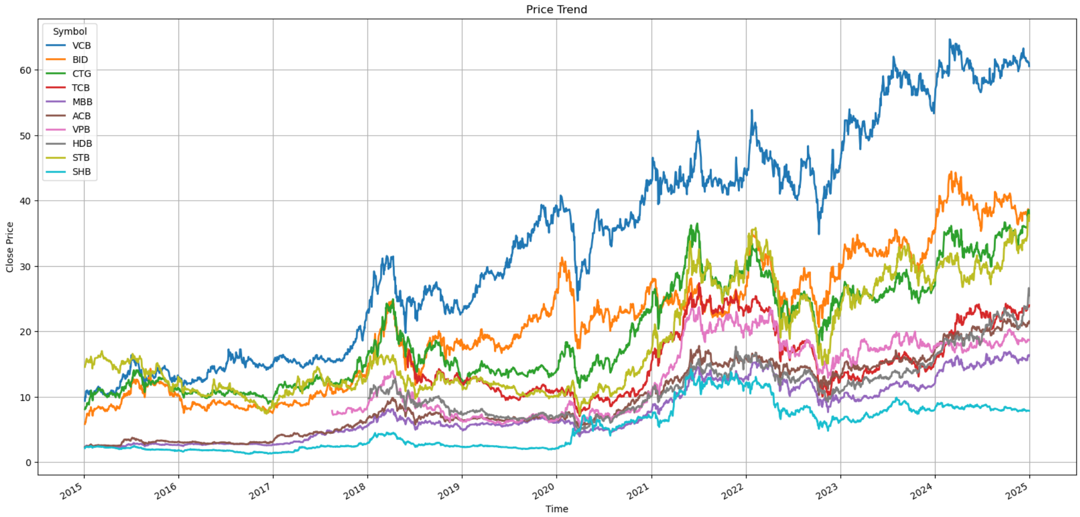

Figure 2: Price trends of major Vietnamese bank stocks (2015–2025).

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.

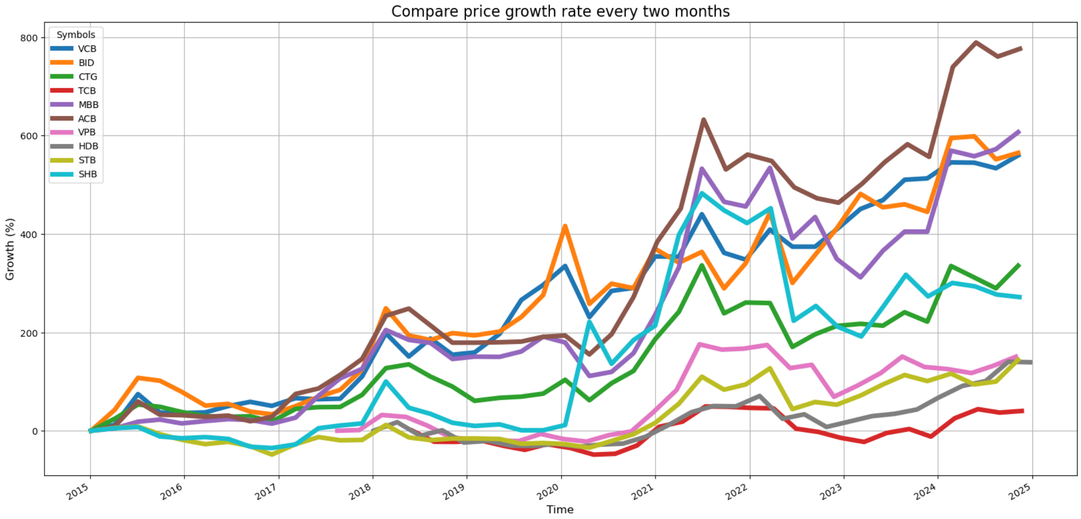

Figure 3: Two-month price growth rate comparison among bank stocks.

Discussion: The growth rate analysis highlights varying levels of volatility. Stocks like ACB and MBB demonstrate high fluctuations, implying higher risk but potential for speculative gains. Conversely, ACB and BID maintain smoother growth trajectories, indicating stable long-term investments. These insights can guide investment decisions based on risk tolerance and growth consistency.

7 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 provide valuable insights for investors and researchers.

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