BÁO CÁO DỰ ÁN BIG DATA STOCK PRICE

Stock Price Big Data (Vietnam)

Sinh viên: Đỗ Việt Dũng 23020343

Tháng 10 năm 2025

Mục lục

Chương 1

Giới thiệu tổng quan

1.1 Mục tiêu

Dự án hướng đến việc áp dụng các công nghệ **Big Data** trong thu thập, lưu trữ và phân tích dữ liệu thị trường chứng khoán Việt Nam. Mục tiêu chính:

- Thu thập dữ liệu lịch sử giá cổ phiếu Việt Nam.
- Lưu trữ dữ liệu lớn trên hệ thống phân tán Hadoop.
- Phân tích, trực quan hóa và dự đoán xu hướng giá cổ phiếu.

1.2 Ý nghĩa

Thị trường chứng khoán Việt Nam thay đổi liên tục, với hàng trăm mã cổ phiếu và hàng triệu giao dịch mỗi ngày. Việc áp dụng **phân tích dữ liệu lớn** giúp:

- Nhận diện xu hướng tăng giảm giá cổ phiếu.
- Đưa ra quyết định đầu tư dựa trên dữ liệu.
- Giảm thiểu rủi ro và nâng cao hiệu quả đầu tư.

Chương 2

Mô hình hệ thống

2.1 Kiến trúc

Hệ thống được triển khai mô phỏng bằng Docker gồm:

- **HDFS Cluster:**
  - 1 Namenode quản lý metadata.
  - 4 Datanode lưu trữ dữ liệu.
- **Spark Cluster:**
  - 1 Spark Master.
  - 4 Spark Worker.

Dữ liệu được lưu trên HDFS và xử lý song song trên cụm Spark thông qua **PySpark Notebook**.

Chương 3

Công nghệ sử dụng

- **Hadoop 3.2.1** – lưu trữ dữ liệu phân tán.
- **Spark 3.1.2** – xử lý dữ liệu song song.
- **Docker** – ảo hóa môi trường hệ thống.
- **Python, Jupyter Notebook** – xử lý và trực quan dữ liệu.
- **vnstock** – lấy dữ liệu cổ phiếu Việt Nam.

Chương 4

Thu thập và lưu trữ dữ liệu

4.1 Mã nguồn Python

Dữ liệu cổ phiếu Việt Nam được lấy qua thư viện **vnstock** và lưu vào thư mục `datack`.

4.2 Đưa dữ liệu vào HDFS

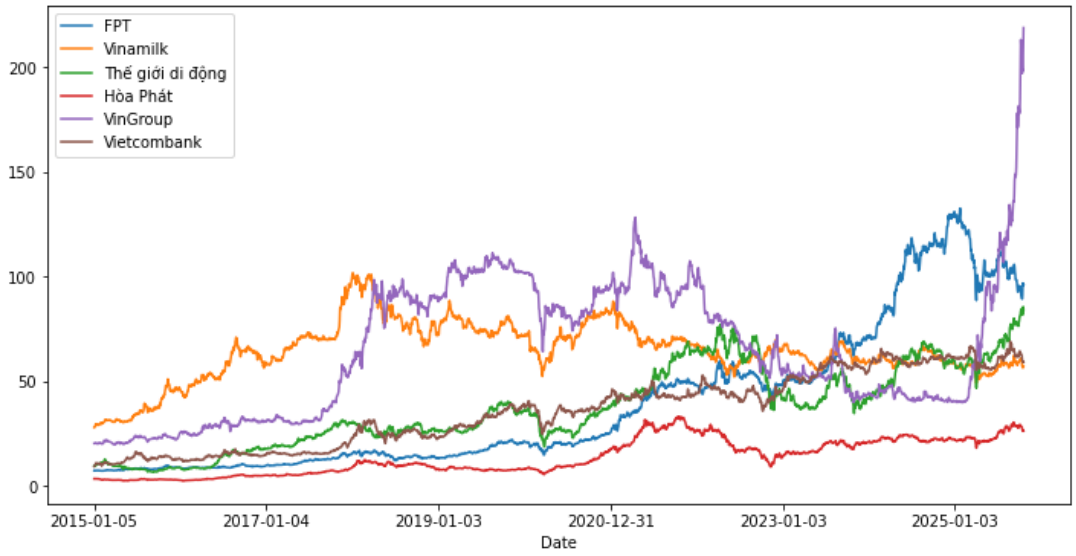
```
docker cp ./datack namenode:/datack
hdfs dfs -put /datack /
```

Sau khi hoàn tất, dữ liệu được lưu trữ và quản lý bởi HDFS.

Chương 5

Phân tích và trực quan hóa dữ liệu

5.1 So sánh giá cổ phiếu theo thời gian

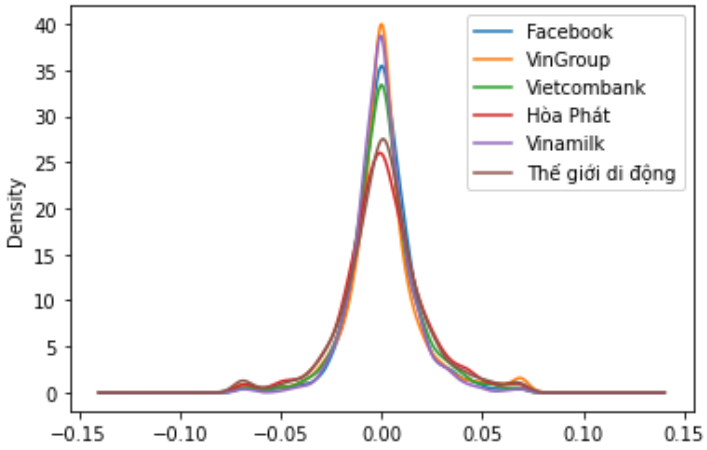


Hình 5.1: Biểu đồ giá cổ phiếu Việt Nam (FPT, VNM, MWG, HPG, VIC, VCB).

Nhận xét: Biểu đồ thể hiện xu hướng biến động giá cổ phiếu của 6 doanh nghiệp lớn tại Việt Nam. Ta thấy:

- **Vingroup (VIC)** có biến động mạnh nhất, tăng mạnh giai đoạn 2025.
- **Thế Giới Di Động (MWG)** và **FPT** duy trì mức tăng ổn định.
- **Hòa Phát (HPG)** giảm nhẹ sau giai đoạn tăng trưởng 2021.

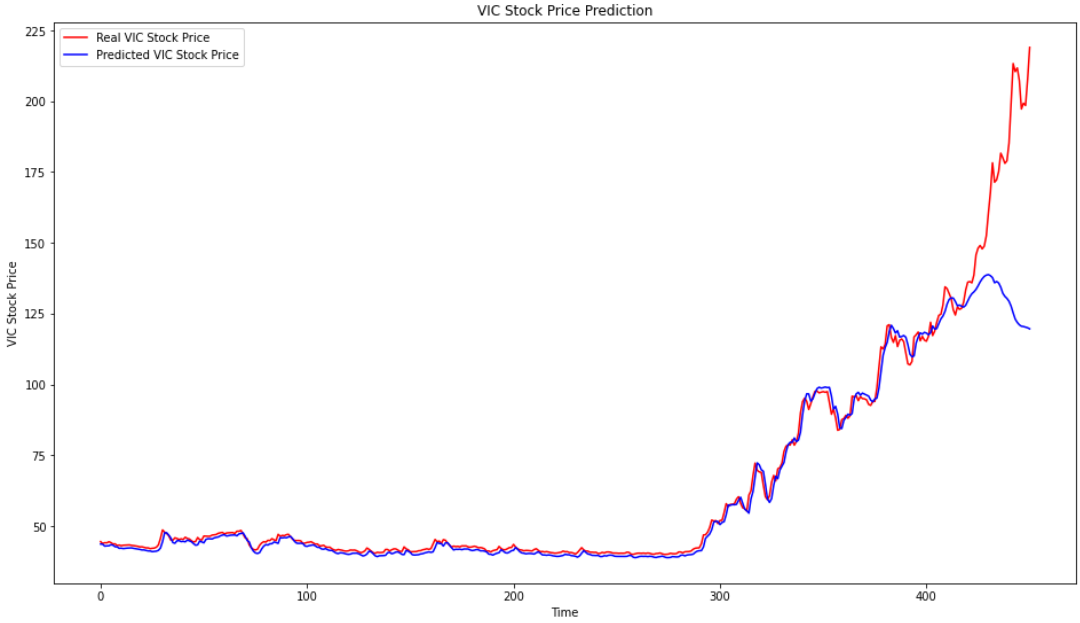
5.2 Phân phối tần suất biến động giá



Hình 5.2: Histogram phân phối lợi nhuận cổ phiếu.

Nhận xét: Các cổ phiếu đều có phân phối gần đối xứng quanh 0, cho thấy phần lớn biến động giá nằm trong khoảng trung bình, ít có đột biến lớn. Điều này phản ánh tính ổn định tương đối của thị trường.

5.3 So sánh giá thực tế và giá dự đoán (VIC)



Hình 5.3: So sánh giá cổ phiếu VIC thực tế và dự đoán.

Nhận xét: - Đường màu đỏ: giá thực tế. - Đường màu xanh: giá dự đoán từ mô hình. Mô hình dự đoán bám sát xu hướng thực tế, chỉ sai lệch tại các điểm biến động mạnh, chứng tỏ độ chính xác tương đối cao. Năm 2025, giá cổ phiếu VinGroup tăng đột ngột, mô hình khó dự đoán chỉnh xác.

Chương 6

Kết luận

Dự án đã thực hiện thành công:

- Mô phỏng hệ thống Big Data bằng Hadoop & Spark.
- Thu thập và lưu trữ dữ liệu cổ phiếu Việt Nam từ năm 2015 đến nay.
- Phân tích, trực quan và dự đoán xu hướng giá cổ phiếu.

Hướng phát triển:

- Áp dụng các mô hình Machine Learning (LSTM, Prophet) để dự đoán chính xác hơn.
- Mở rộng dữ liệu sang các sàn khác như HNX, UPCOM.
- Tự động hóa pipeline thu thập và phân tích dữ liệu.