Xây dựng hệ thống xử lý dữ liệu lớn trong phân tích và dự đoán xu hướng thị trường chứng khoán Việt Nam

Viện Trí tuệ nhân tạo - Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội
Đào Tự Phát

Ngày 20 tháng 10 năm 2025

Tóm tắt nội dung

Thị trường chứng khoán, với hàng triệu giao dịch mỗi ngày, tạo ra một khối lượng dữ liệu khổng lồ (Big Data). Việc áp dụng các công nghệ xử lý dữ liệu lớn để phân tích nguồn dữ liệu này đã trở thành một công cụ hữu hiệu, giúp các nhà đầu tư đưa ra quyết định thông minh và hạn chế rủi ro. Bài tập này trình bày việc mô phỏng một hệ thống Big Data hoàn chỉnh để lưu trữ và xử lý dữ liệu chứng khoán. Dự án tập trung vào dữ liệu giao dịch lịch sử của thị trường Việt Nam. Kiến trúc hệ thống được xây dựng dựa trên hai công nghệ cốt lõi: **Hadoop (HDFS)** được sử dụng cho lưu trữ phân tán và **Apache Spark** được dùng cho xử lý dữ liệu song song. Dữ liệu được thu thập từ API công khai **vnstock**. Dự án minh họa một quy trình hoàn chỉnh từ khâu thu thập, lưu trữ phân tán đến xử lý dữ liệu chứng khoán quy mô lớn.

1 Giới thiệu

1.1 Bối cảnh và tính cấp thiết của đề tài

Thị trường chứng khoán là một môi trường phức tạp với hàng triệu lượt giao dịch diễn ra mỗi ngày, tạo ra một kịch bản không ngừng chuyển động. Khối lượng dữ liệu khổng lồ này (Big Data) vừa là một thách thức, vừa là một cơ hội lớn.

Việc áp dụng các công nghệ xử lý dữ liệu lớn vào phân tích môi trường chứng khoán đã trở thành một công cụ hữu hiệu cho các nhà đầu tư. Các phân tích sâu từ dữ liệu lớn, thay vì chỉ dựa vào trực giác hay các phân tích đơn lẻ, sẽ giúp các nhà đầu tư đưa ra những quyết định thông minh hơn, dự đoán xu hướng và hạn chế rủi ro trên một thị trường đầy biến động.

1.2 Mục tiêu dự án

Bài tập này hướng đến việc mô phỏng một hệ thống Big Data hoàn chỉnh sử dụng các công nghệ cốt lõi là Hadoop và Spark cho hai nhiệm vụ chính:

- **Lưu trữ:** Xây dựng một hệ thống tệp phân tán (HDFS) có khả năng lưu trữ dữ liệu chứng khoán lịch sử quy mô lớn.
- **Xử lý:** Xây dựng một cụm tính toán phân tán (Spark) để thực thi các tác vụ phân tích, thống kê và dự đoán trên tập dữ liệu đó.

2 Kiến trúc hệ thống

2.1 Công nghệ

Hệ thống tích hợp các công nghệ cốt lõi sau:

- **Hadoop (HDFS & YARN):** Được sử dụng làm nền tảng chính. HDFS đóng vai trò lưu trữ phân tán, và YARN chịu trách nhiệm quản lý tài nguyên cụm.
- **Spark:** Được sử dụng làm framework xử lý dữ liệu song song.

2.2 Kiến trúc hệ thống

2.2.1 Cụm lưu trữ HDFS

Đây là thành phần chịu trách nhiệm lưu trữ dữ liệu phân tán.

- **1** `namenode`**:** Quản lý metadata của hệ thống tệp.
- **4** `datanode`**:** Lưu trữ các block dữ liệu thực tế.

2.2.2 Cụm quản lý tài nguyên YARN

Hệ thống sử dụng YARN để quản lý và điều phối tài nguyên (CPU, RAM) cho toàn bộ cụm Hadoop.

- `resourcemanager`**:** Dịch vụ trung tâm, tiếp nhận yêu cầu từ các ứng dụng và cấp phát tài nguyên.
- `nodemanager`**:** Các dịch vụ chạy trên các node worker, chịu trách nhiệm khởi chạy và giám sát các tác vụ theo chỉ thị của ResourceManager.
- `historyserver`**:** Dịch vụ lưu trữ và cung cấp giao diện web để theo dõi lịch sử của các ứng dụng đã hoàn thành.

2.2.3 Cụm xử lý Spark

Một cụm Spark độc lập được thiết lập song song để thực thi các phân tích dữ liệu.

- `spark-master`**:** Dịch vụ quản lý của cụm Spark, nhận các tác vụ xử lý từ Jupyter Notebook và phân phối chúng đến các Spark Worker.
- `spark-worker`**:** Dịch vụ thực thi, trực tiếp đọc dữ liệu từ HDFS và thực hiện các phép tính toán, biến đổi dữ liệu.

3 Thu thập dữ liệu

Toàn bộ dữ liệu của thị trường chứng khoán Việt Nam được thu thập từ API của **VnStock**. Dự án này đã thu thập thông tin của 1721 công ty, đối với mỗi công ty, API được gọi để lấy dữ liệu cổ phiếu lịch sử từ ngày giao dịch đầu tiên của công ty đến ngày thu thập dữ liệu (ngày 19/10/2025).

Sau khi thu thập và loại bỏ dữ liệu bị lỗi, còn lại 1685 công ty. Dữ liệu của mỗi công ty được ghi vào một tệp CSV và bao gồm 5 trường: `Date` (ngày giao dịch), `Open` (giá mở cửa), `High` (giá cao nhất), `Low` (giá thấp nhất), `Close` (giá đóng cửa). Các tệp CSV này tuân thủ nghiêm ngặt định dạng dữ liệu thị trường (OHLCV).

Sau khi dữ liệu được thu thập, chúng được gửi đến hệ thống tệp **HDFS** trước khi được xử lý bởi **Spark**.

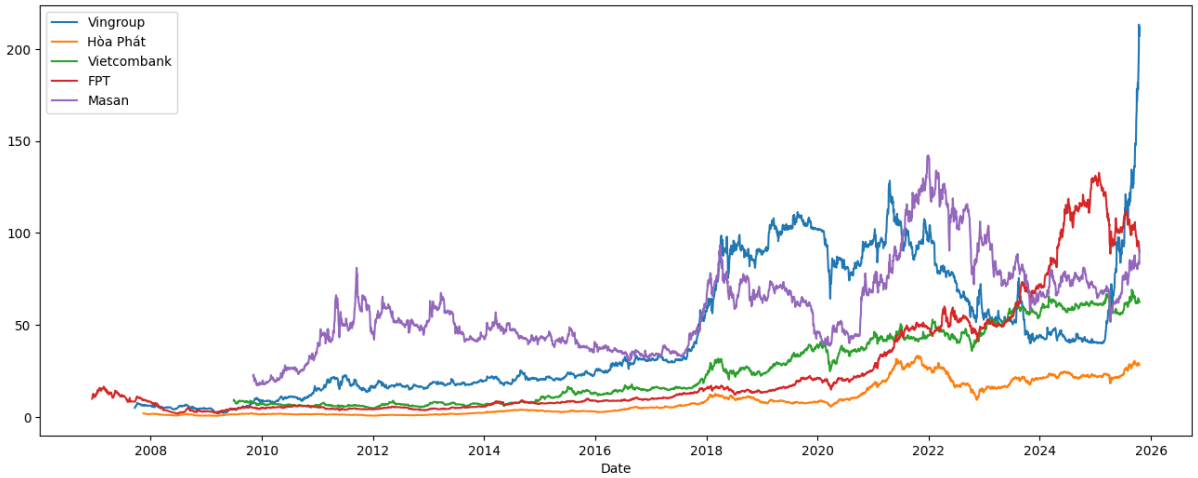
4 Phân tích dữ liệu

Để hiểu rõ hơn về hành vi của thị trường chứng khoán Việt Nam, dự án này đã tiến hành phân tích dữ liệu lịch sử từ 5 cổ phiếu lớn trên thị trường chứng khoán Việt Nam từ các lĩnh vực khác nhau, bao gồm: **Tập đoàn Vingroup (VIC)**, **Tập đoàn Hòa Phát (HPG)**, **Ngân hàng Vietcombank (VCB)**, **Tập đoàn FPT (FPT)**, và **Tập đoàn Masan (MSN)**.

Phân tích này, được thực hiện bằng PySpark, tập trung vào các chỉ số tài chính chính từ dữ liệu giao dịch hàng ngày (giá mở, cao, thấp, và đóng cửa).

4.1 Phân tích giá trung bình và xu hướng

Để phân tích xu hướng dài hạn, dự án này tính toán giá trung bình hàng ngày ($Mean = \frac{High+Low}{2}$) cho mỗi cổ phiếu để làm mượt các biến động trong ngày.



Hình 1: Biểu đồ đường so sánh giá trung bình (Mean Price) theo ngày.

Biểu đồ đường (Hình 1) thể hiện sự biến động giá của các mã này từ khoảng năm 2008 đến cuối năm 2025.

- **Xu hướng chung:** Cả 5 mã đều cho thấy xu hướng tăng trưởng dài hạn, tuy nhiên với mức độ biến động và quỹ đạo tăng trưởng rất khác nhau.
- **Vingroup (VIC - Xanh dương):** Thể hiện sự tăng trưởng tương đối ổn định cho đến khoảng năm 2017, sau đó bước vào một chu kỳ tăng trưởng mạnh mẽ, đạt đỉnh vào khoảng năm 2021. Sau giai đoạn điều chỉnh, cổ phiếu này có một cú tăng vọt đột biến vào cuối năm 2025, trở thành mã có giá trị cao nhất trên biểu đồ.
- **Hòa Phát (HPG - Cam):** Gần như đi ngang ở mức giá rất thấp trong phần lớn thời gian, trước khi có một đợt tăng trưởng đột biến vào năm 2021. Đây là mã có thị giá thấp nhất trong nhóm được phân tích.
- **Vietcombank (VCB - Xanh lá) & FPT (Đỏ):** Cả hai mã này đều có chung đặc điểm là tăng trưởng chậm, ổn định trong thập kỷ đầu tiên (khoảng 2008-2018). Từ 2019 trở đi, cả hai bắt đầu một xu hướng tăng tốc mạnh mẽ và bền bỉ hơn.
- **Masan (MSN - Tím):** Mã này thể hiện mức độ biến động cao, thường xuyên dẫn đầu thị trường về giá trong giai đoạn 2018-2025, với các đỉnh lớn vào năm 2022 và 2025, nhưng cũng xen kẽ các đợt sụt giảm sâu.

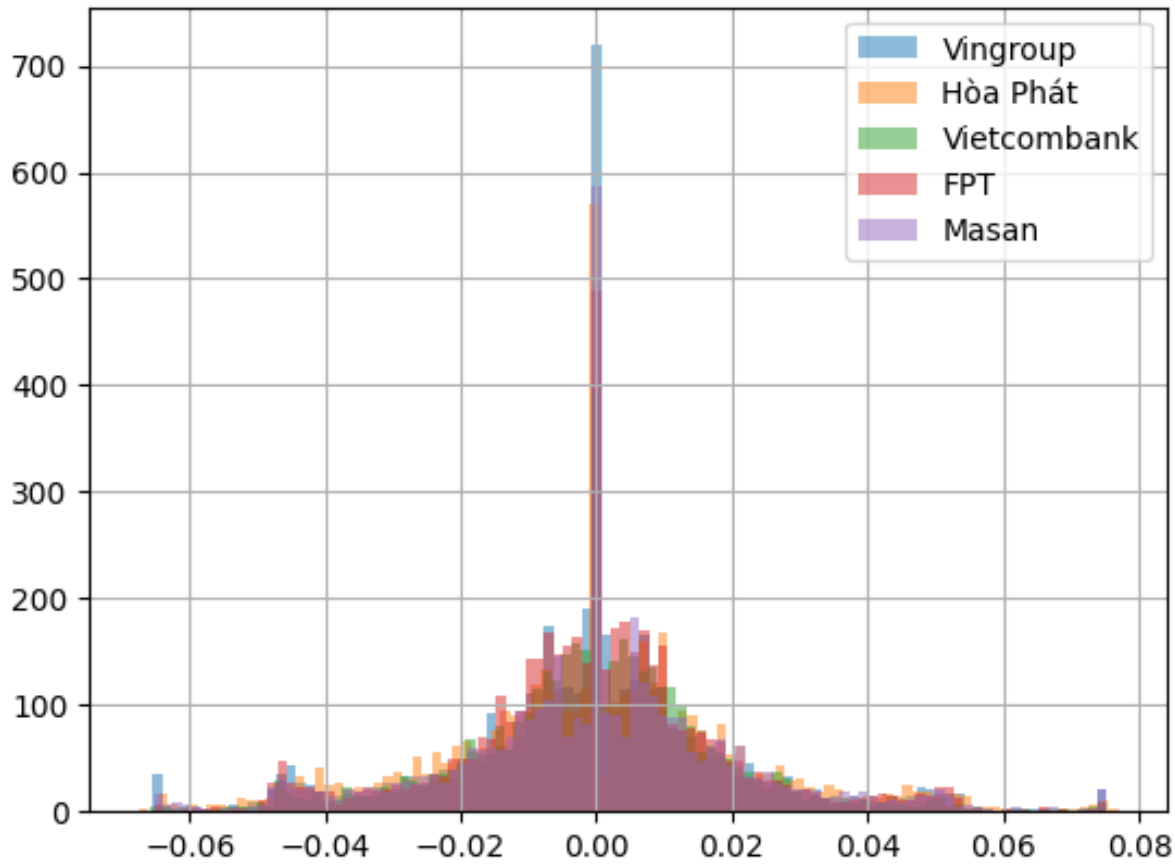
4.2 Phân tích biến động (Volatility)

Để đo lường rủi ro và mức độ biến động, dự án đã tính toán tỷ suất sinh lợi hàng ngày (Daily Percentage Return) để đo lường hiệu suất và mức độ biến động của từng cổ phiếu. Độ lệch chuẩn (Standard Deviation) của các tỷ suất sinh lợi này được sử dụng làm thước đo định lượng cho sự biến động.

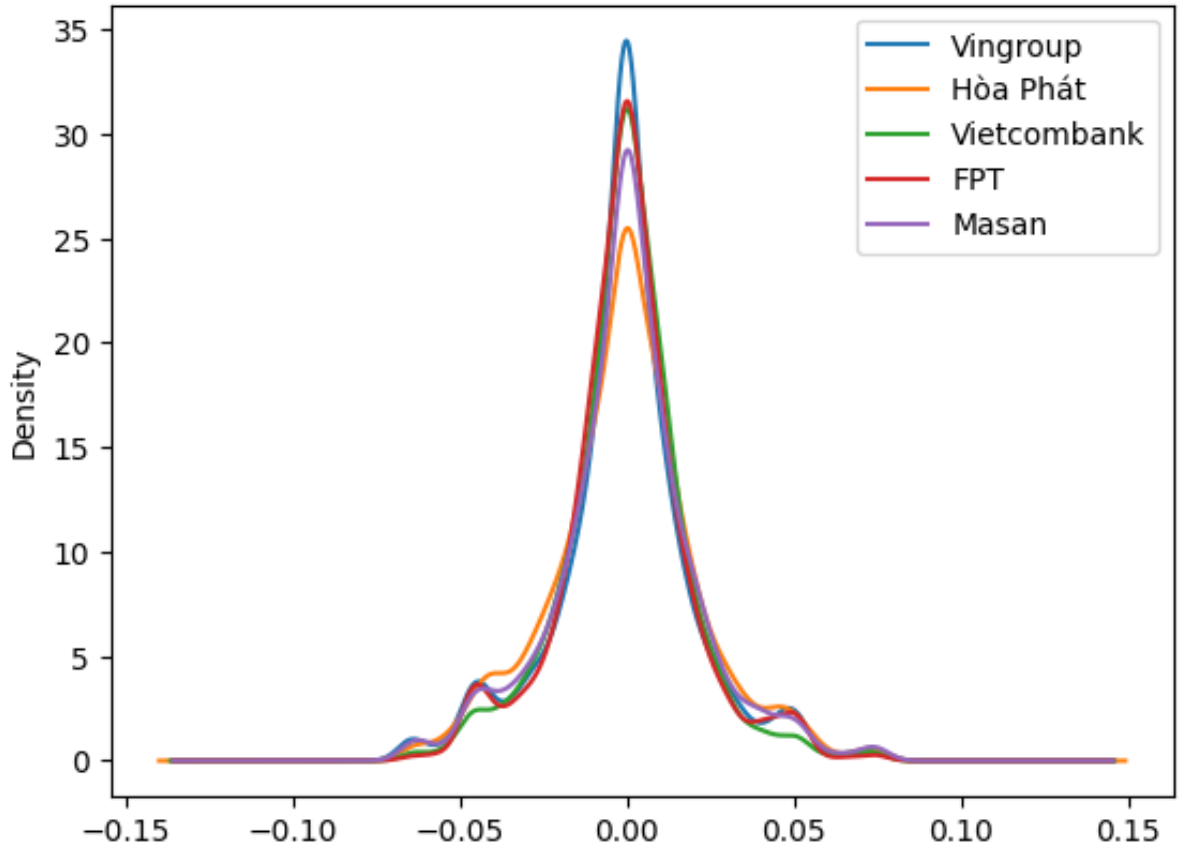
- **Phân phối tập trung:** Cả hai biểu đồ (Histogram tại Hình 2 và Biểu đồ Mật độ tại Hình 3) đều xác nhận rằng phần lớn tỷ suất sinh lợi hàng ngày của cả 5 mã đều tập trung rất cao xung quanh mốc 0. Điều này cho thấy các thay đổi lớn (lãi/lỗ) xảy ra với tần suất thấp, và thị trường phần lớn là ổn định trong ngày.
- **So sánh độ biến động:** Khi so sánh 5 mã với nhau, **Hòa Phát (HPG - màu cam)** thể hiện rõ sự khác biệt.
  - Trên biểu đồ Histogram (Hình 2), cột trung tâm (quanh mốc 0) của HPG thấp hơn đáng kể so với các mã còn lại, cho thấy HPG có ít ngày giao dịch gần như không đổi.
  - Biểu đồ Mật độ (Hình 3) làm rõ điều này: đường cong của HPG có đỉnh (peak) thấp nhất và phần "vai"(shoulders) rộng nhất, minh chứng cho việc HPG có nhiều ngày ghi nhận biến động ở mức vừa phải hơn.
- **Kết luận:** Dựa trên phân tích phân phối, **HPG** được xác định là cổ phiếu có độ biến động (volatility) cao nhất trong nhóm được quan sát. Ngược lại, **Vingroup (VIC - màu xanh dương)** có đỉnh cao và nhọn nhất, cho thấy tính ổn định cao nhất (biến động thấp nhất).

5 Dự đoán xu hướng giá cổ phiếu (LSTM)

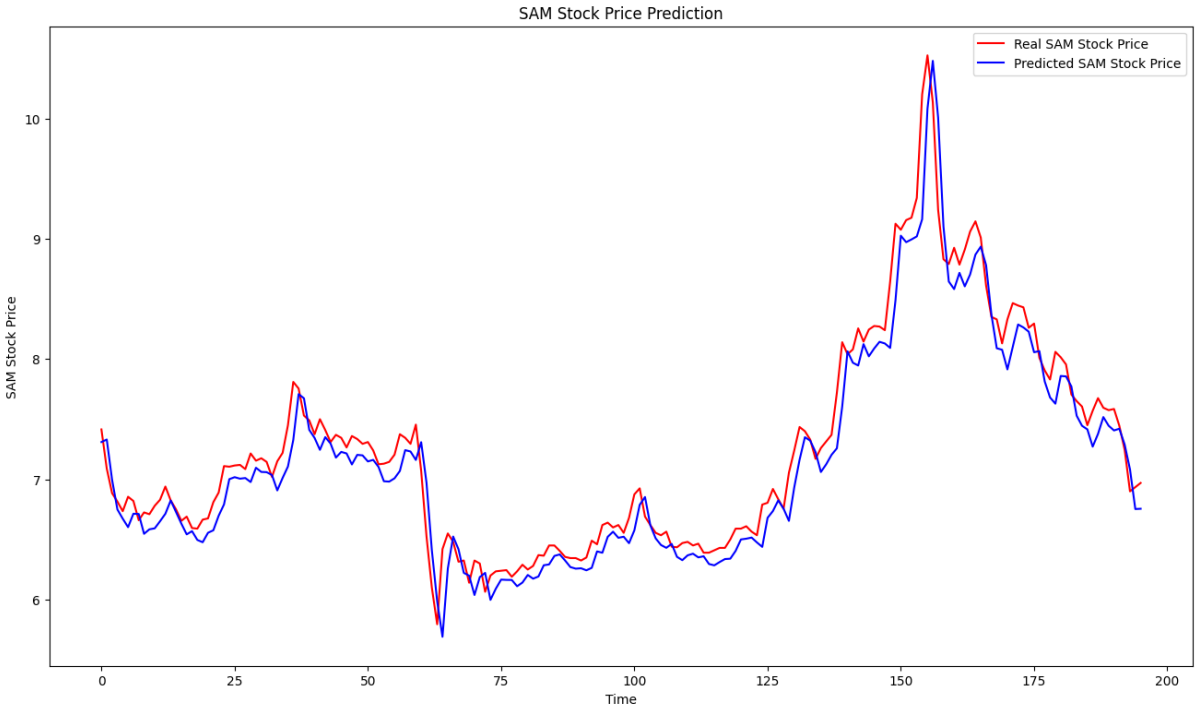
Để kiểm tra khả năng dự đoán, dự án đã xây dựng một mô hình học sâu (Deep Learning) sử dụng kiến trúc **LSTM (Long Short-Term Memory)**. Mô hình này được lựa chọn



Hình 2: Biểu đồ phân phối (Histogram) tỷ suất sinh lợi hàng ngày.



Hình 3: Biểu đồ mật độ (KDE) của tỷ suất sinh lợi hàng ngày.



Hình 4: So sánh giá trị thực tế (Real SAM Stock Price - màu đỏ) và giá trị dự đoán (Predicted SAM Stock Price - màu xanh) trên tập Test.

vì đây là một dạng Mạng Nơ-ron Hồi quy (RNN) chuyên dụng, có khả năng học hỏi và nhận diện các xu hướng phụ thuộc dài hạn (long-term dependencies) trong dữ liệu chuỗi thời gian, rất phù hợp với bài toán giá cổ phiếu.

Mô hình được huấn luyện trên dữ liệu lịch sử (trước năm 2025) để dự đoán giá trị trung bình (*Mean*) của cổ phiếu **SAM - Công ty Cổ phần SAM Holdings** trong giai đoạn thử nghiệm (Test - dữ liệu năm 2025).

5.1 Kết quả và Nhận xét

Kết quả dự đoán của mô hình LSTM trên tập dữ liệu thử nghiệm được trực quan hóa tại Hình 4.

Từ biểu đồ, có thể rút ra các nhận xét sau:

- **Bám sát xu hướng:** Đường dự đoán (màu xanh) đã bám sát rất tốt theo xu hướng chung của đường giá trị thực tế (màu đỏ). Mô hình đã học thành công các mô hình (patterns) tăng và giảm dài hạn của cổ phiếu SAM, bao gồm cả cú sụt giảm mạnh (khoảng thời gian 50-60) và giai đoạn tăng trưởng chính (từ 125 đến 155).
- **Độ trễ:** Mô hình thể hiện một độ trễ (lag) nhỏ. Điều này rõ ràng nhất tại các đỉnh và đáy nhọn. Ví dụ, tại đỉnh cao nhất (khoảng thời gian 155), đường dự đoán màu xanh đạt đỉnh hơi chậm và thấp hơn so với giá thực tế.
- **Độ chính xác:** Mặc dù có độ trễ, mô hình dự đoán rất chính xác về *hướng* di chuyển của giá (ví dụ: mô hình đã dự đoán chính xác xu hướng giảm ngay sau đỉnh cao 155). Nó cũng xử lý tốt các giai đoạn biến động phức tạp (ví dụ, từ 160-200), cho thấy khả năng của LSTM trong việc mô hình hóa dữ liệu chuỗi thời gian.

6 Kết luận và hướng phát triển

6.1 Tóm tắt kết quả đạt được

Dự án đã mô phỏng thành công một quy trình (pipeline) Big Data cơ bản để phân tích dữ liệu chứng khoán, đạt được các mục tiêu chính sau:

- **Kiến trúc:** Một cụm Big Data hoàn chỉnh bao gồm HDFS (lưu trữ), YARN (quản lý tài nguyên), và Spark Standalone (xử lý) đã được triển khai thành công và ảo hóa bằng Docker.
- **Thu thập và lưu trữ:** Dữ liệu lịch sử (OHLCV) của thị trường chứng khoán Việt Nam đã được thu thập (sử dụng thư viện `vnstock`) và nạp thành công vào hệ thống tệp phân tán HDFS.
- **Phân tích:** Các tác vụ phân tích dữ liệu như lọc, tạo đặc trưng (Mean Price), và thống kê đã được thực thi phân tán trên 5 mã cổ phiếu vốn hóa lớn (VIC, HPG,...) bằng PySpark SQL.
- **Học máy:** Đã xây dựng và huấn luyện thành công một mô hình dự đoán chuỗi thời gian (LSTM) bằng Tensorflow, cho thấy kết quả trực quan tốt trong việc dự đoán xu hướng giá của cổ phiếu SAM trên tập dữ liệu thử nghiệm.

6.2 Hạn chế của dự án

Bên cạnh các kết quả đạt được, dự án vẫn còn tồn tại một số hạn chế mang tính kỹ thuật:

- **Nút thắt cổ chai (Bottleneck):** Hạn chế lớn nhất là việc huấn luyện mô hình LSTM. Dữ liệu đã được thu thập về Driver node (thông qua `.toPandas()`) thay vì được huấn luyện phân tán. Điều này làm mất đi lợi thế xử lý song song của Spark và không có khả năng mở rộng cho các tập dữ liệu lớn hơn.
- **Đặc trưng:** Mô hình dự đoán (SAM) còn tương đối đơn giản, chỉ dựa trên giá trị 'Mean' lịch sử mà bỏ qua các yếu tố quan trọng khác như Khối lượng giao dịch (Volume) hay các chỉ báo kỹ thuật (RSI, MACD).
- **Phân tích cơ bản:** Phân tích dữ liệu trên 5 mã (VIC, HPG, ...) mới chỉ dừng ở mức thống kê mô tả, chưa khai thác sâu các mối tương quan phức tạp hay các mô hình ẩn của thị trường.

6.3 Hướng phát triển tương lai

Để giải quyết các hạn chế trên và xây dựng một hệ thống hoàn thiện hơn, dự án đề xuất các hướng phát triển trong tương lai, bám sát các đặc tính của Big Data:

- **Huấn luyện phân tán:** Tích hợp các thư viện chuyên dụng như `TensorFlowOnSpark` để giải quyết "nút thắt cổ chai". Điều này cho phép mô hình LSTM được huấn luyện song song trên toàn cụm Spark, tăng tốc độ và khả năng xử lý dữ liệu lớn.
- **Làm giàu đặc trưng:** Tận dụng sức mạnh của Spark SQL để tính toán các đặc trưng kỹ thuật phức tạp (SMA, EMA, ...) và đưa vào làm đầu vào cho mô hình dự đoán.

- **Xử lý dữ liệu tốc độ cao:** Kích hoạt và tích hợp `Kafka` và `Spark Streaming` để xây dựng một pipeline có khả năng phân tích dữ liệu giao dịch gần thời gian thực.
- **Xử lý dữ liệu đa dạng:** Xây dựng một pipeline `Spark NLP` (Xử lý Ngôn ngữ Tự nhiên) để thu thập và xử lý dữ liệu phi cấu trúc (ví dụ: tin tức từ Vietstock). Việc trích xuất phân tích Cảm xúc từ tin tức có thể được sử dụng như một đặc trưng đầu vào quan trọng cho mô hình dự đoán.

Tài liệu

[1] Apache Hadoop. `https://hadoop.apache.org/`

[2] Apache Spark. `https://spark.apache.org/`

[3] PySpark Documentation. `https://spark.apache.org/docs/latest/api/python/index.html`

[4] VnStock. `https://vnstocks.com/docs`

[5] Hochreiter, S., & Schmidhuber, J. (1997). *Long short-term memory.* Neural computation, 9(8), 1735-1780.

[6] Dự án mẫu. `https://github.com/thviet79/Stock-Price`