BÁO CÁO ASSIGNMENT

Đề tài: Triển khai hệ thống phân tích và xử lý dữ liệu lớn

Case study: Stock Price Bigdata

Họ và tên: [Cao Minh Quang]
Lớp: [K68A-AI1]
Giảng viên hướng dẫn: [TS.Trần Hồng Việt]

Ngày 25 tháng 10 năm 2025

Mục lục

1 Giới thiệu **2**
1.1 Đặt vấn đề . . . 2
1.2 Mục tiêu của dự án . . . 2

2 Mô hình hệ thống và Công nghệ sử dụng **2**
2.1 Mô hình hệ thống . . . 2
2.2 Công nghệ sử dụng . . . 2

3 Các bước triển khai xây dựng hệ thống **3**
3.1 Cấu hình và khởi chạy môi trường Docker . . . 3
3.2 Thu thập dữ liệu . . . 3
3.3 Đưa dữ liệu vào HDFS . . . 3
3.4 Xử lý và phân tích dữ liệu . . . 4

4 Kết quả thực hiện **4**
4.1 Kết quả phân tích dữ liệu . . . 4
4.2 Dự đoán xu hướng giá cổ phiếu . . . 8

5 Kết luận **10**
5.1 Kết quả đạt được . . . 10
5.2 Thuận lợi và Khó khăn . . . 10
5.3 Hướng phát triển . . . 10

1 Giới thiệu

1.1 Đặt vấn đề

Thị trường chứng khoán là một môi trường phức tạp với kịch bản không ngừng chuyển động, nơi hàng triệu lượt giao dịch diễn ra mỗi ngày. Với khối lượng dữ liệu khổng lồ và tốc độ phát sinh liên tục, việc phân tích thủ công hoặc sử dụng các công cụ truyền thống trở nên kém hiệu quả.

Trong bối cảnh đó, việc áp dụng xử lý dữ liệu lớn vào môi trường chứng khoán đã trở thành một công cụ hữu hiệu cho các nhà đầu tư. Các phân tích dữ liệu lớn có khả năng xử lý và khai phá thông tin từ tập dữ liệu đa dạng, giúp các nhà đầu tư đưa ra những quyết định thông minh hơn, đồng thời hạn chế rủi ro trên một môi trường đầy biến động này.

1.2 Mục tiêu của dự án

Dự án này hướng đến việc tìm hiểu các chủ đề về phân tích dữ liệu lớn và vận dụng các kỹ thuật phân tích, xử lý dữ liệu liên quan.

Mục tiêu cụ thể của dự án là xây dựng và triển khai một hệ thống dữ liệu lớn mô phỏng với bài toán cụ thể là "Stock Price Bigdata". Hệ thống này sẽ sử dụng các công nghệ tiêu chuẩn công nghiệp là Hadoop và Spark cho việc lưu trữ và xử lý dữ liệu chứng khoán. Cuối cùng, kết quả thực hiện sẽ được kiểm tra và trình bày trực quan trên màn hình giao diện ứng dụng.

2 Mô hình hệ thống và Công nghệ sử dụng

2.1 Mô hình hệ thống

Để giải quyết bài toán đặt ra, một hệ thống mô phỏng dữ liệu lớn được thiết kế bao gồm hai cụm chính: cụm lưu trữ và cụm xử lý.

- **Cụm lưu trữ (HDFS):** Dữ liệu được lưu trữ trên một cụm Hệ thống tệp phân tán Hadoop (HDFS). Cụm này bao gồm:
  - 1 Namenode: Đóng vai trò quản lý siêu dữ liệu (metadata) và điều phối các Datanode.
  - 4 Datanode: Đóng vai trò lưu trữ các khối dữ liệu thực tế.
- **Cụm xử lý (Spark):** Để lấy dữ liệu ra và xử lý, hệ thống sử dụng một cụm Spark. Cụm này bao gồm:
  - 1 Spark Master: Quản lý và điều phối tài nguyên cho các ứng dụng Spark.
  - 4 Spark Worker: Các nút thực thi, chịu trách nhiệm chạy các tác vụ xử lý dữ liệu.

2.2 Công nghệ sử dụng

Hệ thống được xây dựng dựa trên các công nghệ và nền tảng sau:

- **Công nghệ lõi:** Hadoop (cho lưu trữ HDFS và quản lý tài nguyên YARN) và Spark (cho xử lý dữ liệu phân tán).
- **Nền tảng mô phỏng:** Toàn bộ hệ thống mô phỏng được xây dựng và quản lý bằng nền tảng Docker.
- **Docker Images:**

  - Sử dụng các image của **Big Data Europe (bde2020)** để tạo ra các thành phần của cụm, bao gồm namenode, các datanode, dịch vụ YARN, spark-master và các spark-worker.
  - Sử dụng image `pyspark-notebook` của Jupyter để cung cấp môi trường tương tác, cho phép người dùng viết mã Python (PySpark) để demo việc xử lý dữ liệu của cụm Spark.

- **Công cụ điều phối:** File `docker-compose.yml` được sử dụng để định nghĩa và khởi chạy toàn bộ các dịch vụ của hệ thống một cách đồng bộ.

3 Các bước triển khai xây dựng hệ thống

3.1 Cấu hình và khởi chạy môi trường Docker

Bước đầu tiên trong quá trình triển khai là định nghĩa kiến trúc hệ thống thông qua file `docker-compose.yml`. File cấu hình này khai báo tất cả các dịch vụ (services) cần thiết, bao gồm:

- Các dịch vụ Hadoop: namenode, datanode (số lượng 4), và các dịch vụ YARN (resource manager, node manager).
- Các dịch vụ Spark: spark-master và spark-worker (số lượng 4).
- Dịch vụ Jupyter: `pyspark-notebook` để tương tác với cụm.

Sau khi hoàn tất file cấu hình, hệ thống được khởi chạy bằng lệnh `docker-compose up -d`. Docker sẽ tự động tải về (pull) các image `bde2020` và `pyspark-notebook` nếu chúng chưa tồn tại, sau đó khởi tạo và liên kết các container theo đúng định nghĩa.

3.2 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).

Dữ liệu của mỗi công ty được ghi vào một tệp CSV và bao gồm 5 trường: *Time* (thời gian 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**.

3.3 Đưa dữ liệu vào HDFS

Khi cụm Hadoop đã khởi chạy thành công (Namenode và các Datanode đều ở trạng thái sẵn sàng), bước tiếp theo là nạp dữ liệu thô của bài toán "Stock Price Bigdata" vào HDFS.

Quá trình này được thực hiện thông qua giao diện dòng lệnh (CLI) bằng cách truy cập vào container Namenode (ví dụ: `docker exec -it namenode /bin/bash`) và sửdụng các lệnh HDFS. Dữ liệu (ví dụ: các file .csv) được sao chép từ hệ thống tệp cục bộ vào một thư mục được chỉ định trên HDFS (ví dụ: `hdfs dfs -put /data /input`).

3.4 Xử lý và phân tích dữ liệu

Giai đoạn xử lý và phân tích được thực hiện trong môi trường Jupyter Notebook, được cung cấp bởi image `pyspark-notebook`.

Thông qua giao diện web của Jupyter (thường ở cổng `localhost:8888`), người dùng kết nối vào Spark Master của cụm. Mã PySpark được sử dụng để:

1. **Đọc dữ liệu từ HDFS:** Khởi tạo một SparkSession và sử dụng các API như `spark.read` để tải dữ liệu giá chứng khoán đã lưu trữ.

   (ví dụ:`spark.read.csv("hdfs://namenode:8020/input/data.csv")`).

2. **Xử lý và biến đổi dữ liệu (Data Cleansing & Transformation):** Thực hiện các thao tác làm sạch (loại bỏ giá trị null), chuyển đổi kiểu dữ liệu (ví dụ: từ string sang timestamp hoặc float), và trích xuất các đặc trưng (features) cần thiết như giá đóng cửa, khối lượng giao dịch.

3. **Phân tích dữ liệu (Data Analysis):** Vận dụng các kỹ thuật phân tích dữ liệu, thực hiện các truy vấn Spark SQL hoặc các phép toán trên DataFrame để tính toán các chỉ số thống kê (trung bình, min, max), xu hướng giá, hoặc các mô hình dự báo đơn giản.

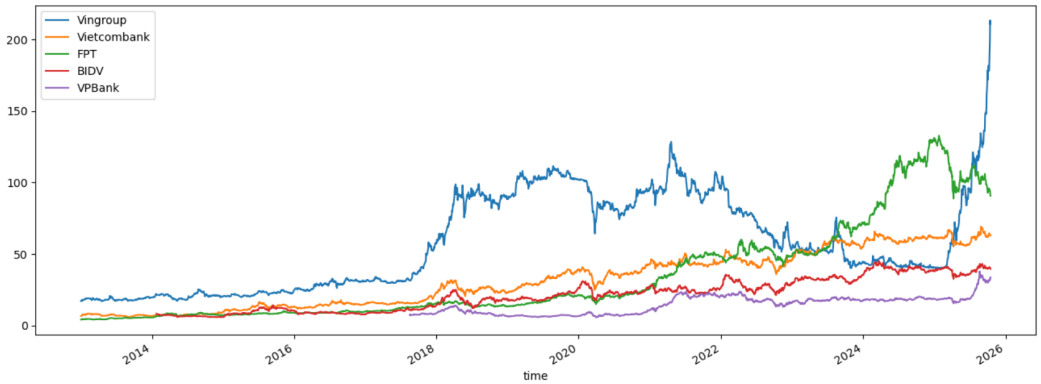
Quá trình này tận dụng khả năng xử lý song song của 4 Spark Worker để tăng tốc độ tính toán trên tập dữ liệu lớn.

4 Kết quả thực hiện

Nhằm có được cái nhìn toàn diện hơn về xu hướng vận động của thị trường chứng khoán Việt Nam, dự án này đã tiến hành phân tích dữ liệu lịch sử của 5 mã cổ phiếu có quy mô vốn hóa hàng đầu, đại diện cho nhiều lĩnh vực khác nhau trên thị trường. Cụ thể gồm: **Tập đoàn Vingroup (VIC)**, **Ngân hàng TMCP Ưu việt VPBank (VPB)** , **Ngân hàng Vietcombank (VCB)**, **Tập đoàn FPT (FPT)** và **Ngân hàng Đầu tư Phát triển Việt Nam BIDV (BID)**.

4.1 Kết quả phân tích dữ liệu

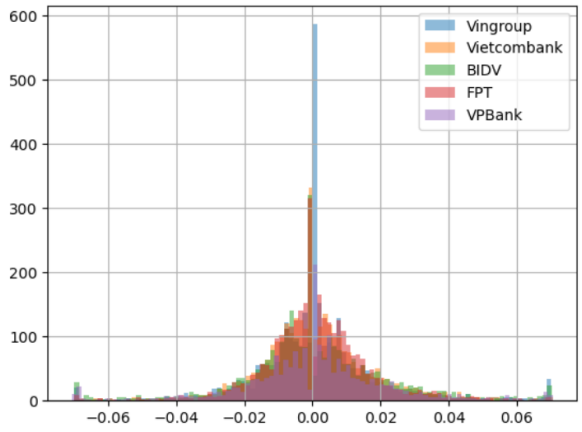
Biểu đồ sau thể hiện diễn biến giá cổ phiếu theo thời gian của năm doanh nghiệp tiêu biểu thuộc các lĩnh vực khác nhau trên thị trường chứng khoán Việt Nam, giúp quan sát rõ xu hướng tăng trưởng dài hạn và những giai đoạn biến động mạnh.



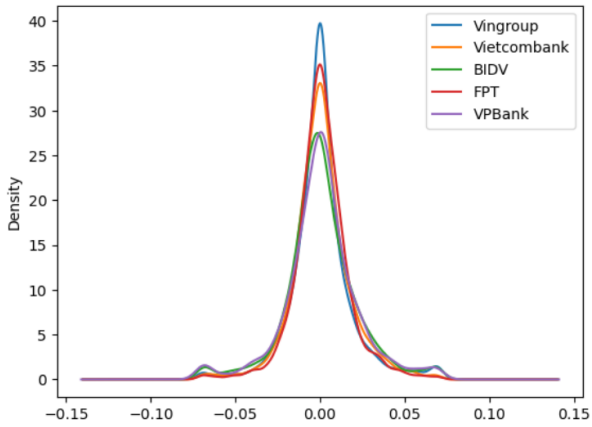
Hình 1: Diễn biến giá cổ phiếu của các doanh nghiệp lớn tại Việt Nam giai đoạn 2013–2025.

- **Tổng quan chung.** Nhìn vào toàn bộ chuỗi, có thể thấy diễn biến giá cổ phiếu của năm doanh nghiệp lớn thuộc các ngành khác nhau trên thị trường chứng khoán Việt thể hiện sự khác biệt rõ rệt về mức tăng trưởng và độ ổn định giữa các mã.
- **Vingroup (VIC):** Biến động mạnh nhất, có nhiều giai đoạn tăng đột biến (đặc biệt trước 2019 và cuối giai đoạn), thể hiện tiềm năng tăng trưởng cao nhưng rủi ro lớn.
- **FPT:** Tăng đều và ổn định hơn VIC, ít dao động cực đoan, phản ánh đặc trưng của doanh nghiệp công nghệ có tăng trưởng bền vững.
- **Vietcombank (VCB):** Đường giá ổn định, tăng chậm nhưng chắc, biên độ dao động nhỏ, phù hợp với nhà đầu tư ưu tiên an toàn.
- **BIDV và VPBank:** Cùng nhóm ngân hàng, có xu hướng tăng tương tự nhưng chậm hơn VCB; BID ổn định hơn VPBank, vốn chỉ tăng mạnh ở giai đoạn cuối.

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.



Hình 2: Biểu đồ phân phối tỉ suất sinh lợi



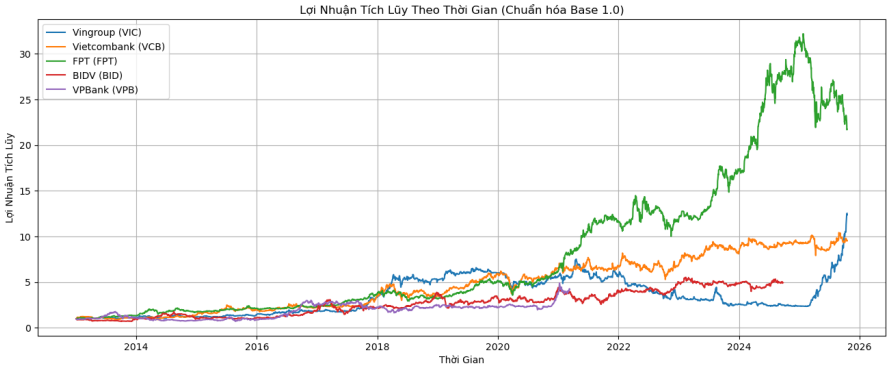
Hình 3: Biểu đồ mật độ tỉ suất sinh lợi

Hai biểu đồ thể hiện đặc trưng thống kê của tỉ suất sinh lợi hàng ngày cho nhóm cổ phiếu gồm Vingroup, Vietcombank, BIDV, FPT và VPBank. Biểu đồ bên trái là phân phối tần suất, cho thấy hầu hết các giá trị tỉ suất sinh lợi tập trung rất cao quanh mức 0, chứng tỏ phần lớn biến động giá mỗi ngày là nhỏ. Tuy nhiên, hai bên phân phối có đuôi dài và dày, nghĩa là đôi khi vẫn xuất hiện những ngày biến động mạnh cả theo chiều tăng và giảm — dấu hiệu của rủi ro cực trị (extreme volatility) trên thị trường. Biểu đồ bên phải thể hiện hàm mật độ xác suất (density plot), cung cấp góc nhìn mượt hơn về cấu trúc phân phối. Tất cả các đường mật độ gần như trùng nhau, chứng tỏ các cổ phiếu có hành vi lợi nhuận tương đồng. Đỉnh nhọn và tập trung quanh 0 cho thấy hiệu suất trung bình ngắn hạn gần bằng 0, trong khi sự khác biệt nhỏ về độ rộng giữa các đường phản ánh độ biến động khác nhau nhẹ — ví dụ, FPT và VPBank có biên độ dao động lớn hơn so với Vietcombank hay BIDV.

Tóm lại, hai biểu đồ cùng khẳng định rằng nhóm cổ phiếu trên có đặc điểm biến động tương đối ổn định, phần lớn các phiên có lợi nhuận nhỏ, nhưng vẫn tồn tại khả năng xuất hiện các phiên dao động mạnh, thể hiện tính rủi ro đặc trưng của thị trường chứng khoán.

Phân tích Lợi nhuận Tích lũy Theo Thời gian

Phân tích lợi nhuận tích lũy (Cumulative Returns) là yếu tố cốt lõi để đánh giá hiệu suất đầu tư dài hạn của mỗi mã cổ phiếu, giả định tái đầu tư toàn bộ lợi nhuận trong suốt giai đoạn nghiên cứu (chuẩn hóa mức cơ sở là 1.0).



Hình 4: Lợi Nhuận Tích Lũy Theo Thời Gian.

- **FPT (Tăng trưởng dẫn đầu):** Mã FPT cho thấy hiệu suất tích lũy vượt trội nhất, với mức tăng ổn định và mạnh mẽ trong suốt giai đoạn. Điều này khẳng định FPT là một cổ phiếu có tăng trưởng bền vững, là lựa chọn lý tưởng cho chiến lược đầu tư dài hạn.
- **VPBank (Tăng trưởng bùng nổ):** VPB, đặc biệt ở giai đoạn sau, đã trải qua những cú tăng trưởng đột biến và vượt qua nhiều mã khác để giành vị trí cao thứ hai về lợi nhuận tích lũy. Điều này phản ánh tiềm năng sinh lợi cao của VPB nhưng cũng đi kèm với rủi ro lớn hơn so với VCB.
- **Vietcombank (Ổn định Bền vững):** VCB duy trì tốc độ tăng trưởng chậm nhưng rất ổn định, tránh được các đợt sụt giảm mạnh. Hiệu suất của VCB cho thấy sự phù hợp với các nhà đầu tư theo đuổi chiến lược bảo thủ.
- **VIC và BIDV (Hiệu suất thấp):** Cả VIC và BID đều có lợi nhuận tích lũy cuối giai đoạn thấp nhất trong nhóm. VIC thể hiện **độ biến động cao** (sharp drops) nhưng lại **kém hiệu quả** trong việc chuyển hóa biến động đó thành tăng trưởng tích lũy dài hạn.

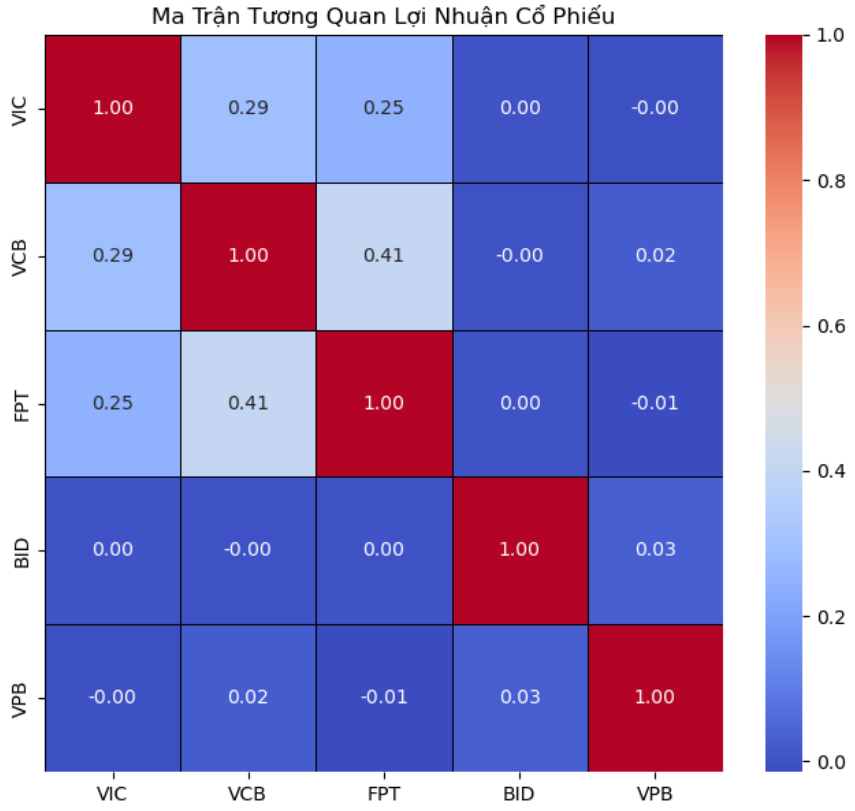
Phân tích Tương quan Lợi nhuận Hàng ngày

Để đánh giá mức độ đồng vận (synchronicity) trong biến động giá giữa các cổ phiếu, ta sẽ tính toán **Ma trận Tương quan (Correlation Matrix)** dựa trên tỷ suất sinh lợi hàng ngày của 5 mã.

Để có thể quan sát chi tiết hơn, ta có thể nhìn vào bảng và ma trận dưới đây:

Bảng 1: Ma trận tương quan lợi nhuận hàng ngày giữa các cổ phiếu

Mã cổ phiếu	VIC	VCB	FPT	BID	VPB
VIC	1.000000	0.293420	0.249702	0.001649	-0.000833
VCB	0.293420	1.000000	0.406149	-0.001471	0.022833
FPT	0.249702	0.406149	1.000000	0.001638	-0.014452
BID	0.001649	-0.001471	0.001638	1.000000	0.025961
VPB	-0.000833	0.022833	-0.014452	0.025961	1.000000



Hình 5: Ma trận Tương quan Lợi nhuận Cổ phiếu VIC, VCB, FPT, BID, VPB.

Sau khi quan sát, ta có thể đưa ra một số nhận xét như sau:

- **Tương quan Ngân hàng (VCB, FPT):** Hai mã cổ phiếu ngân hàng lớn (VCB, BID) và FPT có mức tương quan vừa phải, ví dụ, $Corr(\text{VCB}, \text{FPT}) \approx 0.41$. Điều này cho thấy khi FPT hoặc VCB tăng/giảm, mã còn lại có xu hướng biến động cùng chiều một cách đáng kể.
- **Tương quan Ngân hàng - Khác ngành (VCB, VIC):** VIC có mức tương quan thấp hơn với VCB, ở mức $Corr(\text{VIC}, \text{VCB}) \approx 0.29$. Tương quan thấp này (dưới 0.5) cho thấy biến động của VIC ít bị ảnh hưởng trực tiếp bởi VCB hơn so với FPT.
- **Tương quan Thấp/Gần như không (BID, VPB, VIC):**
  - Lợi nhuận hàng ngày của BID và VPB có tương quan rất yếu, gần bằng 0 (ví dụ: $Corr(\text{BID}, \text{VIC}) \approx 0.0016$).
  - Đặc biệt, $Corr(\text{VIC}, \text{VPB}) \approx -0.0008$, là một mức tương quan gần như bằng 0 (hoặc âm rất nhẹ).
- **Kết luận về Đa dạng hóa:** Mức tương quan rất thấp giữa các mã như BID, VPB và VIC là một tín hiệu tích cực cho các nhà đầu tư muốn **đa dạng hóa danh mục đầu tư (portfolio diversification)**. Việc kết hợp các tài sản có tương quan thấp sẽ giúp giảm thiểu rủi ro tổng thể của danh mục.

4.2 Dự đoán xu hướng giá cổ phiếu

Trong nghiên cứu này, mô hình **CNN + BiGRU + Attention** (Convolutional Neural Network kết hợp với Bi-directional Gated Recurrent Unit và Cơ chế Attention) được sử dụng để dự đoán giá cổ phiếu. Đây là một kiến trúc tiên tiến trong xử lý chuỗi thời gian, được thiết kế để khai thác tối đa cả **đặc trưng cục bộ** và **mối quan hệ phụ thuộc dài hạn** trong dữ liệu giá.

- **CNN (Conv1D):** Lớp này trích xuất các đặc trưng cục bộ (local features) trong chuỗi thời gian, ví dụ như các mẫu hình (patterns) biến động ngắn hạn hoặc các mức hỗ trợ/kháng cự.
- **BiGRU (Bidirectional GRU):** Tận dụng lợi thế của mạng hồi quy GRU để học các phụ thuộc theo thời gian, đồng thời xử lý chuỗi dữ liệu theo **cả hai chiều (quá khứ và tương lai)** để có cái nhìn toàn diện hơn.
- **Attention Mechanism:** Cơ chế này cho phép mô hình **gán trọng số khác nhau** cho các bước thời gian đầu vào. Thay vì coi tất cả các ngày lịch sử là quan trọng như nhau, mô hình sẽ tập trung vào những ngày có **biến động mạnh** hoặc **quan trọng nhất** trong việc dự đoán giá tương lai.

Quy trình Huấn luyện và Dự đoán

Mục tiêu chính của nghiên cứu này là xây dựng một mô hình học sâu có khả năng **dự đoán giá trị trung bình (*Mean*) hằng ngày của cổ phiếu CII** trong giai đoạn kiểm thử năm 2025 — một giai đoạn đại diện cho bối cảnh thị trường hiện tại, nơi biến động giá thường xuyên xảy ra do các yếu tố kinh tế vĩ mô và tâm lý nhà đầu tư.

Thay vì chỉ dự đoán giá trị tuyệt đối tại từng thời điểm riêng lẻ, mô hình được thiết kế để **học và nắm bắt mối quan hệ chuỗi thời gian giữa các phiên giao dịch liên tiếp**, từ đó nhận diện được cả xu hướng ngắn hạn và mô hình biến động trung hạn của cổ phiếu. Dữ liệu đầu vào được tổ chức thành các **chuỗi 60 ngày liên tiếp**, phản ánh diễn biến giá trong hai tháng gần nhất, đóng vai trò như một *cửa sổ trượt* (sliding window) chứa thông tin động học của thị trường.

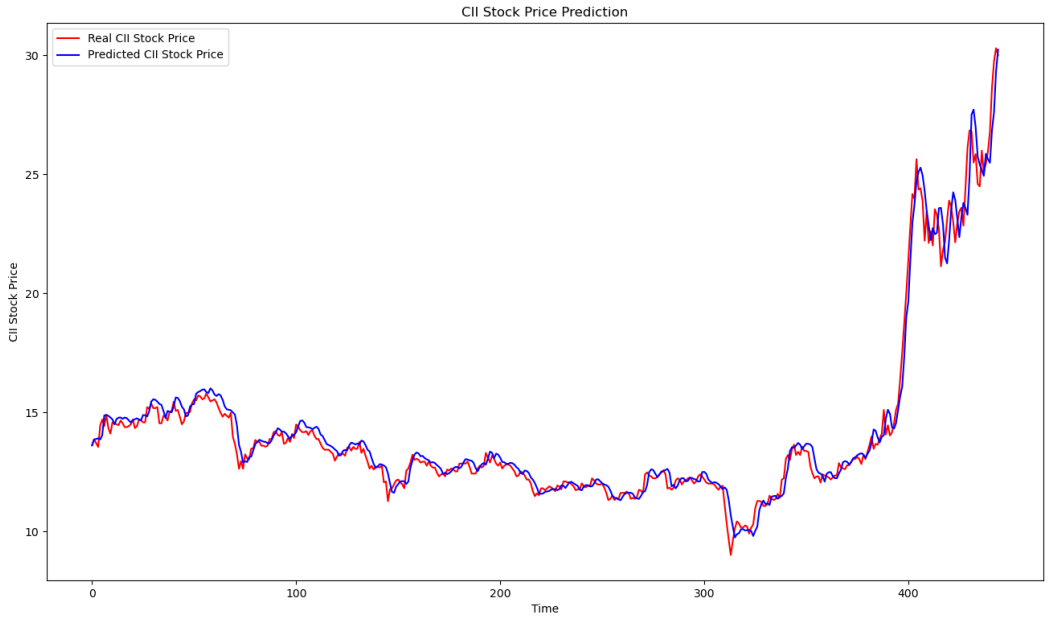
Mỗi chuỗi dữ liệu này bao gồm các đặc trưng biến đổi của giá trung bình qua thời gian, được chuẩn hóa nhằm đảm bảo tính ổn định trong quá trình huấn luyện. Bằng cách khai thác chuỗi 60 ngày quá khứ làm ngữ cảnh đầu vào, mô hình không chỉ dự đoán được giá trị Mean của ngày tiếp theo mà còn **nắm bắt được nhịp điệu và cấu trúc dao động** — yếu tố then chốt trong các bài toán dự báo tài chính.

Cách tiếp cận này cho phép mô hình học được *động lực thị trường* (market dynamics) theo thời gian, thay vì chỉ ghi nhớ các điểm dữ liệu rời rạc. Nhờ đó, mô hình có thể phát hiện sớm các tín hiệu thay đổi xu hướng, giúp dự đoán trở nên **ổn định, mượt mà và mang tính chiến lược hơn** khi áp dụng vào thực tế giao dịch.

Nhờ vào sự kết hợp giữa khả năng trích chọn đặc trưng mạnh mẽ của CNN, khả năng ghi nhớ tuần tự của BiGRU và tính tập trung của Attention, mô hình có thể **nắm bắt được cả xu hướng ngắn hạn lẫn dài hạn** trong biến động giá cổ phiếu. Điều này giúp mô hình không chỉ dự đoán chính xác giá trị trung bình trong tương lai gần, mà còn thể hiện năng lực ổn định khi thị trường xuất hiện các biến động đột ngột hoặc các pha đảo chiều xu hướng.

Kết quả trực quan của quá trình dự đoán được thể hiện trong hình 5, minh họa sự tương quan giữa đường giá thực tế và đường giá dự đoán mà mô hình tạo ra trong giai đoạn kiểm thử.

Nhìn chung, mô hình dự đoán (**đường xanh dương**, P_{pred}) cho thấy **hiệu suất rất tốt** và có độ bám sát cao so với giá thực tế (**đường đỏ**, P_{real}).



Hình 6: Biểu đồ so sánh giá cổ phiếu CII thực tế và giá dự đoán theo thời gian.

1. **Giai đoạn Ổn định và Giảm Nhẹ (Time** $0 \to 300$**):** Trong giai đoạn này, giá cổ phiếu dao động quanh mức từ 10 đến 16. Đường dự đoán (màu xanh) gần như **trùng khớp hoàn toàn** với đường thực tế (màu đỏ), chỉ có sai lệch rất nhỏ tại các điểm đảo chiều ngắn hạn.
    - Ví dụ: Ở khu vực Time $60 \to 80$, mô hình bắt được **cú giảm giá nhẹ** và **phản ứng chính xác với các biến động nhỏ**.
2. **Giai đoạn Tăng Mạnh và Biến động Cao (Time** $300 \to 450$**):** Đây là giai đoạn thể hiện rõ sức mạnh của mô hình Transformer.
    - Mô hình **dự đoán chính xác xu hướng tăng trưởng bùng nổ**, từ mức giá khoảng 10 lên trên 30.
    - Mặc dù xuất hiện một số dao động mạnh ở vùng đỉnh, mô hình vẫn **bám sát rất tốt** xu hướng thực tế, chỉ sai lệch nhẹ ở các điểm cực trị.
    - Điều này cho thấy mô hình có khả năng **bắt kịp các chuyển động phi tuyến và biến động đột ngột** của thị trường.

Quan sát kỹ, sai số (khoảng cách giữa hai đường $|P_{\text{real}} - P_{\text{pred}}|$) có vẻ **tăng nhẹ** ở cuối biểu đồ (khoảng Time $180 \to 200$), đặc biệt là tại đỉnh cuối cùng:

- Đường thực tế (Đỏ) đạt đỉnh khoảng $28 \to 29$.
- Đường dự đoán (Xanh) dường như có xu hướng **vượt nhẹ** qua đường thực tế ở đỉnh cuối cùng (tiệm cận hoặc vượt 29), nhưng vẫn giữ được hình dạng chung của biến động.

Kết Luận về Mô Hình CNN + BiGRU + Attention

Mô hình **CNN + BiGRU + Attention** đã mang lại **kết quả ấn tượng**, vượt trội so với các mô hình hồi quy chuỗi thời gian cơ bản. Sự kết hợp giữa các lớp mạng đã phát huy hiệu quả:

- **Ưu điểm:** Hiệu suất vượt trội, khả năng nắm bắt **xu hướng dài hạn** và **biến động ngắn hạn** với độ chính xác cao. Chỉ số lỗi huấn luyện MSE (Mean Squared Error) đạt được sau 50 epochs là 5.95×10^{-4} (từ kết quả huấn luyện), cho thấy độ chính xác rất cao của mô hình.
- **Hạn chế & Gợi ý cải tiến:** Mặc dù rất tốt, sai số giữa hai đường dự đoán và thực tế vẫn có dấu hiệu **tăng nhẹ** ở cuối biểu đồ (gần Time 200), đặc biệt tại các điểm cực trị (đỉnh và đáy). Việc tinh chỉnh thêm (như tăng số lượng epochs, điều chỉnh tỷ lệ Dropout, hoặc sử dụng các kỹ thuật *Ensemble*) là cần thiết để cải thiện độ chính xác tại các điểm đảo chiều quan trọng của thị trường.

5 Kết luận

5.1 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 tiêu chí 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*), thống kê (tính **tương quan**), và trực quan hóa đã được thực thi phân tán trên 5 mã cổ phiếu vốn hóa lớn (FPT, VCB,...) bằng PySpark SQL.
- **Học máy:** Đã xây dựng và huấn luyện thành công mô hình dự đoán chuỗi thời gian **CNN + BiGRU + Attention** bằng TensorFlow, cho thấy kết quả trực quan **rất tốt** trong việc dự đoán xu hướng giá của cổ phiếu CII trên tập dữ liệu thử nghiệm.

5.2 Thuận lợi và Khó khăn

- **Thuận lợi:**
  - Việc sử dụng Docker và các image bde2020 giúp đơn giản hóa quá trình cài đặt và cấu hình cụm.
  - Môi trường Jupyter Notebook cung cấp giao diện trực quan để phát triển và demo.
- **Khó khăn:**
  - Yêu cầu tài nguyên hệ thống (RAM, CPU) lớn để chạy đồng thời nhiều container.
  - Khó khăn ban đầu trong việc cấu hình mạng (networking) giữa các container.

5.3 Hướng phát triển

Dự án đã thiết lập thành công một quy trình Big Data cơ bản. Trong tương lai, việc phát triển sẽ tập trung vào ba trụ cột chính: Mở rộng Dữ liệu, Nâng cao Mô hình Học máy, và Tối ưu hóa Kiến trúc hệ thống.

1. Mở Rộng Phạm Vi Dữ Liệu và Đặc Trưng

- **Mở rộng Nguồn Dữ liệu:** Tăng cường thu thập các loại dữ liệu phi truyền thống để làm phong phú tập huấn luyện:
  - Dữ liệu **Sentiment Analysis** (Phân tích cảm xúc) từ các nguồn tin tức, mạng xã hội, và diễn đàn chứng khoán.
  - Dữ liệu **Vĩ mô** (lãi suất, lạm phát, GDP) và dữ liệu các thị trường quốc tế liên quan.
- **Xây dựng Đặc trưng (Feature Engineering) Phức tạp:** Phát triển các chỉ báo kỹ thuật chuyên sâu (ví dụ: Ichimoku Cloud, Elliott Waves) và các đặc trưng liên thị trường để cung cấp tín hiệu dự đoán mạnh mẽ hơn cho mô hình ML.

2. Nâng Cao và Đa Dạng Hóa Mô Hình Học Máy

- **Thử nghiệm Kiến trúc Chuỗi Thời gian Tiên tiến:** Thử nghiệm các kiến trúc mạnh hơn mô hình hiện tại, đặc biệt là mô hình **Transformer** hoặc các biến thể của nó như **BERT** hoặc **GPT** được tinh chỉnh cho chuỗi thời gian, để tăng khả năng học các mối quan hệ phụ thuộc toàn cục (global dependencies).
- **Áp dụng Học Tăng cường (*Reinforcement Learning* - RL):** Xây dựng một tác tử RL để đưa ra các quyết định Mua/Bán (*Trading Agent*) dựa trên dự đoán của mô hình và tối đa hóa lợi nhuận thực tế.
- **Mô hình Hồi quy Đa biến (Multi-variate Regression):** Mở rộng dự đoán từ một mã cổ phiếu sang dự đoán đồng thời nhiều mã cổ phiếu hoặc dự đoán biến động của chỉ số thị trường (VN-Index), có tính đến **ma trận tương quan** đã phân tích.

3. Tối Ưu Hóa Hệ Thống và Triển Khai Thực tế (Production)

- **Tối ưu hóa Tốc độ Xử lý:** Chuyển đổi các tác vụ xử lý của PySpark SQL sang sử dụng Spark DataFrames hoặc RDD thuần túy để tối ưu hiệu suất, đặc biệt khi quy mô dữ liệu tăng.
- **Triển khai Kiến trúc Phục vụ (Serving Architecture):** Sử dụng các công cụ như Kubernetes để quản lý các Docker Container, và triển khai mô hình học máy theo thời gian thực (*Real-time Inference*) thông qua API (REST API), cho phép mô hình dự đoán được truy cập và sử dụng liên tục.
- **Xây dựng Giao diện Trực quan (*Dashboard*):** Phát triển một giao diện trực quan (Streamlit hoặc Dash) kết nối trực tiếp với kết quả dự đoán để người dùng có thể theo dõi hiệu suất mô hình và các tín hiệu giao dịch một cách dễ dàng.

Tài liệu

[1] Big Data Europe (bde2020), *Docker Images for Hadoop and Spark*, https://github.com/big-data-europe/docker-hadoop (Truy cập lần cuối: Ngày 25 tháng 10 năm 2025).

[2] Jupyter Docker Stacks, *pyspark-notebook*, https://jupyter-docker-stacks.readthedocs.io/en/latest/using/selecting.html#jupyter-pyspark-notebook (Truy cập lần cuối: Ngày 25 tháng 10 năm 2025).

[3] Apache Spark, *Apache Spark Documentation*, https://spark.apache.org/docs/latest/ (Truy cập lần cuối: Ngày 25 tháng 10 năm 2025).

[4] Dự án tham khảo, *thviet79-StockPrice*, https://github.com/thviet79/Stock-Price (Truy cập lần cuối: Ngày 25 tháng 10 năm 2025).

[5] Link repo dự án , https://github.com/quang011105/BigdataStockPrice (Truy cập lần cuối: Ngày 25 tháng 10 năm 2025).