TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN - TIN HỌC

DATA PIPELINE FOR STOCK MARKET TREND ANALYSIS AND PREDICTION

SEMINAR KHOA HỌC DỮ LIỆU
CHƯƠNG TRÌNH CHÍNH QUY

Giảng viên hướng dẫn: Th.S Đoàn Thị Trâm
Sinh viên thực hiện:

- Nguyễn Thị Bích Ngọc
- Đoàn Thị Mẫn Nhi
- Nguyễn Thúy Vy

Tp. Hồ Chí Minh, tháng 01/2025

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA TOÁN - TIN HỌC

DATA PIPELINE FOR STOCK MARKET TREND ANALYSIS AND PREDICTION

SEMINAR KHOA HỌC DỮ LIỆU

CHƯƠNG TRÌNH CHÍNH QUY

Giảng viên hướng dẫn: Th.S Đoàn Thị Trâm

Sinh viên thực hiện:

- Nguyễn Thị Bích Ngọc - 21280100
- Đoàn Thị Mẫn Nhi - 21280102
- Nguyễn Thúy Vy - 21280121

Tp. Hồ Chí Minh, tháng 01/2025

Lời cảm ơn

Kính gửi cô Đoàn Thị Trâm cùng quý thầy cô Khoa Toán – Tin học, nhóm chúng em xin gửi lời cảm ơn sâu sắc và chân thành nhất đến cô vì đã tận tình hướng dẫn và đồng hành cùng nhóm trong suốt quá trình thực hiện đề tài **"Data Pipeline for Stock Market Trend Analysis and Prediction"**. Sự hỗ trợ và định hướng quý báu từ cô đã giúp nhóm vượt qua nhiều khó khăn và thử thách trong quá trình nghiên cứu, đồng thời mang lại những góc nhìn mới mẻ và toàn diện hơn về lĩnh vực này.

Những lời động viên, chia sẻ cùng sự tận tâm của cô không chỉ giúp nhóm hiểu rõ hơn về cách xây dựng hệ thống xử lý dữ liệu, mà còn khơi gợi niềm yêu thích và sự say mê đối với lĩnh vực nghiên cứu này. Qua mỗi lời góp ý, cô đã giúp nhóm nâng cao khả năng tư duy, hoàn thiện phương pháp nghiên cứu và không ngừng tiến bộ.

Chúng em thật sự cảm kích và may mắn khi được cô làm người hướng dẫn và rất trân trọng những kiến thức, kinh nghiệm mà cô đã chia sẻ trong suốt hành trình này.

Bên cạnh đó, nhóm cũng xin gửi lời tri ân đến quý thầy cô trong Khoa Toán – Tin học, những người đã giảng dạy, truyền đạt kiến thức và tạo dựng nền tảng vững chắc để nhóm tự tin thực hiện đề tài nghiên cứu.

Nhóm xin kính chúc cô Đoàn Thị Trâm cùng các thầy cô trong Khoa luôn dồi dào sức khỏe, hạnh phúc và đạt được nhiều thành tựu trong sự nghiệp giảng dạy và nghiên cứu khoa học.

Trân trọng,
Nhóm thực hiện

Mục lục

Danh sách hình

Danh sách bảng

Chương 1

Tổng quan đề tài

1.1 Giới thiệu bài toán

Trong bối cảnh thế giới ngày càng phát triển, nền kinh tế toàn cầu trở nên phức tạp và cạnh tranh hơn bao giờ hết. Thị trường chứng khoán, với vai trò là một trong những lĩnh vực cốt lõi của kinh tế, luôn nhận được sự quan tâm đặc biệt từ các nhà đầu tư và tổ chức tài chính. Tuy nhiên, sự biến động không ngừng của thị trường khiến việc phân tích và dự đoán xu hướng trở thành một thách thức lớn, đòi hỏi sự hỗ trợ từ công nghệ hiện đại.

Sự phát triển của công nghệ dữ liệu lớn (Big Data) và trí tuệ nhân tạo (AI) mang lại cơ hội lớn cho việc thu thập, xử lý và phân tích dữ liệu tài chính. Với khối lượng dữ liệu khổng lồ được tạo ra mỗi ngày từ các giao dịch chứng khoán, tin tức kinh tế và các yếu tố thị trường khác, nhu cầu về một hệ thống xử lý dữ liệu nhanh chóng và hiệu quả ngày càng trở nên cấp thiết.

Vấn đề được đặt ra: Là một kỹ sư dữ liệu (Data Engineer) của một tổ chức tài chính hoặc công ty đầu tư, làm thế nào để thiết kế và triển khai một **"đường ống dữ liệu" (Data Pipeline)**, giúp xử lý và phân tích dữ liệu thời gian thực (real-time data) từ thị trường chứng khoán? Hệ thống này cần đảm bảo tính chính xác, hiệu suất cao và khả năng dự đoán xu hướng để hỗ trợ các nhà đầu tư trong việc ra quyết định chiến lược.

Đề tài không chỉ tập trung vào giải quyết bài toán kỹ thuật mà còn góp phần nâng cao hiệu quả đầu tư, giảm thiểu rủi ro và khai thác tối đa giá trị từ dữ liệu tài chính.

1.2 Mục tiêu và nhiệm vụ

Mục tiêu của đề tài **"Data Pipeline for Stock Market Trend Analysis and Prediction"** là xây dựng một hệ thống tự động và hiệu quả để thu thập, xử lý, lưu trữ và phân tích dữ liệu thị trường chứng khoán. Hệ thống này hỗ trợ việc dự đoán xu hướng thị trường, cung cấp thông tin giá trị cho các nhà đầu tư và tổ chức tài chính trong việc ra quyết định.

Đề tài đặt ra các nhiệm vụ chính như sau: Đầu tiên, xây dựng một hệ thống đường ống dữ liệu (Data Pipeline) toàn diện, đóng vai trò trung tâm trong việc tự động hóa quá trình thu thập, xử lý và lưu trữ dữ liệu một cách hiệu quả, đảm bảo đáp ứng khối lượng dữ liệu lớn từ thị trường chứng khoán. Tiếp theo, triển khai quy trình thu thập và xử lý dữ liệu thời gian thực, giúp làm sạch, chuẩn hóa và đồng bộ hóa dữ liệu từ nhiều nguồn khác nhau, từ đó cung cấp dữ liệu đầu vào chất lượng cao cho các bước phân tích. Ngoài ra, áp dụng các mô hình học máy tiên tiến, thống kê phân tích để phân tích và dự đoán xu hướng thị trường, từ đó hỗ trợ các nhà đầu tư đưa ra quyết định chiến lược chính xác hơn. Cuối cùng, hệ thống cần được thiết kế với khả năng vận hành ổn định, đảm bảo tính bảo mật và dễ dàng mở rộng trong tương lai để phù hợp với các yêu cầu kinh doanh ngày càng cao.

1.3 Định hướng thực hiện

1.3.1 Thu thập dữ liệu

- Xây dựng quy trình thu thập dữ liệu từ các nguồn khác nhau như API(Application Programming Interface) thị trường chứng khoán, các trang web tin tức kinh tế, và báo cáo tài chính.
- Thiết lập hệ thống thu thập dữ liệu thời gian thực, đảm bảo khả năng thu thập nhanh chóng và liên tục.
- Đảm bảo dữ liệu được thu thập có chất lượng cao, đúng định dạng và phù hợp với mục tiêu phân tích.

1.3.2 Xử lí dữ liệu

- Làm sạch dữ liệu bằng cách loại bỏ các giá trị không hợp lệ, trùng lặp hoặc thiếu sót.
- Chuẩn hóa và tích hợp dữ liệu từ nhiều nguồn khác nhau để đảm bảo tính nhất quán.
- Xây dựng các bước tiền xử lý dữ liệu cần thiết để phục vụ cho các mô hình phân tích và dự đoán.

1.3.3 Lưu trữ dữ liệu

- Thiết kế hệ thống lưu trữ dữ liệu linh hoạt, đáp ứng khối lượng dữ liệu lớn và yêu cầu truy cập nhanh chóng.
- Sử dụng các giải pháp như Data Lake hoặc Data Warehouse để lưu trữ dữ liệu thô và dữ liệu đã qua xử lý.
- Đảm bảo tính bảo mật, toàn vẹn dữ liệu và khả năng mở rộng hệ thống lưu trữ.

1.3.4 Trực quan hóa dữ liệu

- Phát triển các công cụ và bảng điều khiển (dashboard) để trình bày dữ liệu một cách trực quan, dễ hiểu.
- Sử dụng các biểu đồ và đồ thị phù hợp để mô tả xu hướng, mẫu và dự đoán trong dữ liệu thị trường.
- Cung cấp thông tin trực quan cho người dùng cuối để hỗ trợ quá trình ra quyết định chiến lược.

Chương 2

Trình bày báo cáo

2.1 Cơ sở lý thuyết

2.1.1 Tổng quan về chứng khoán

2.1.1.1 Khái niệm chứng khoán

Chứng khoán là tài sản bao gồm cổ phiếu, chứng chỉ quỹ, chứng khoán phái sinh, trái phiếu, chứng quyền, chứng quyền có bảo đảm, quyền mua cổ phần, chứng chỉ lưu ký. Những loại tài sản này có điểm chung là một bằng chứng xác nhận sở hữu hợp pháp của người sở hữu (gọi chung là nhà đầu tư) với tài sản của doanh nghiệp hoặc tổ chức phát hành.

2.1.1.2 Thị trường chứng khoán

Thị trường chứng khoán là một thị trường nơi chứng khoán, chủ yếu là cổ phiếu, được mua và bán. Nó cung cấp một nền tảng cho các công ty huy động vốn bằng cách phát hành cổ phiếu và cho các nhà đầu tư giao dịch những cổ phiếu đó.

Thị trường chứng khoán được phân loại thành thị trường sơ cấp và thị trường thứ cấp. Thị trường sơ cấp là nơi chứng khoán được phát hành lần đầu tiên, còn thị trường thứ cấp là nơi chứng khoán đã phát hành được giao dịch giữa các nhà đầu tư.
Thị trường chứng khoán đóng vai trò là trung tâm của nền kinh tế hiện đại, nơi diễn ra các giao dịch cổ phiếu, trái phiếu, và các công cụ tài chính khác. Đây là nơi phản ánh tình trạng kinh tế của một quốc gia và cung cấp cơ hội đầu tư cho các nhà đầu tư cá nhân cũng như tổ chức.

Đặc điểm của thị trường chứng khoán:

- Tính biến động cao: Giá cổ phiếu thay đổi nhanh chóng và liên tục, chịu ảnh hưởng bởi nhiều yếu tố như kinh tế, chính trị, và tâm lý nhà đầu tư.

- Tính phi tuyến tính: Dữ liệu giá chứng khoán không tuân theo các quy tắc toán học đơn giản, đòi hỏi các phương pháp phân tích phức tạp.

- Khối lượng dữ liệu lớn: Giao dịch hàng ngày tạo ra lượng dữ liệu khổng lồ cần được xử lý và phân tích.

Nhờ vào việc đóng góp của thị trường chứng khoán, các công ty và tổ chức có thể tăng cường vốn để đầu tư vào hoạt động kinh doanh của mình, đồng thời cũng thu hút được sự quan tâm và đầu tư từ các nhà đầu tư trong và ngoài nước. Điều này tạo ra các cơ hội đầu tư mới, đẩy mạnh sự phát triển của các ngành công nghiệp và góp phần vào tăng trưởng kinh tế của một quốc gia.

Với sự phát triển của công nghệ, việc phân tích và dự đoán xu hướng thị trường chứng khoán trở nên khả thi hơn nhờ các công cụ mạnh mẽ trong lĩnh vực khoa học dữ liệu và xử lý dữ liệu lớn

2.1.2 Data Pipeline

2.1.2.1 Khái niệm Data Pipeline

Data Pipeline (đường ống dữ liệu) là một chuỗi các quy trình được tổ chức để thu thập, chuyển đổi, lưu trữ và xử lý dữ liệu từ nhiều nguồn khác nhau đến đích tự động và liên tục.
Mục tiêu của Data Pipeline là tổ chức, xử lí dữ liệu và đảm bảo dữ liệu được truyền tải một cách liền mạch, chính xác và sẵn sàng cho các bước phân tích hoặc dự đoán tiếp theo.

2.1.2.2 Phân loại

Một số loại Data Pipeline phổ biến:

- **Batch Data Pipeline:** được sử dụng để xử lý dữ liệu theo cách đồng bộ và định kỳ. Dữ liệu được gom nhóm và xử lý trong các quá trình chạy hàng loạt.
- **Real-time Data Pipeline:** cho phép xử lý và chuyển đổi dữ liệu trong thời gian thực. Real-time Data Pipeline thích hợp cho các ứng dụng yêu cầu phản hồi nhanh chóng.
- **Streaming Data Pipeline:** được sử dụng để xử lý dữ liệu đến liên tục và không ngừng. Dữ liệu được xử lý theo luồng và được gửi từ nguồn đến đích một cách liên tục.
- **Cloud Data Pipeline:** được triển khai và vận hành trong môi trường đám mây. Cloud Data Pipeline sử dụng các dịch vụ và tài nguyên đám mây để xử lý và chuyển đổi dữ liệu.
- **Hybrid Data Pipeline:** kết hợp của nhiều loại Data Pipeline, kết nối các nguồn và đích dữ liệu trong một hệ thống phức tạp.

2.1.2.3 Kiến trúc Data Pipeline

Một Data Pipeline thường có kiến trúc cơ bản:

1. **Thu thập dữ liệu:** Là quá trình thu thập dữ liệu từ nhiều nguồn khác nhau như API(Application Programming Interface) và chuyển vào luồng Data Pipeline để xử lí
2. **Chuyển đổi dữ liệu:** Là quá trình chuyển đổi và xử lí dữ liệu về định dạng phù hợp với yêu cầu của bài toán phân tích và phù hợp với mô hình được xay dựng sau đó.
   Quá trình gồm các bước:
   - Chọn lọc dữ liệu
   - Chuẩn hóa dữ liệu
   - Xử lý các dữ liệu bị lỗi: dữ liệu bị thiếu, trùng lặp
   - Tính toán các thông số cần thiết

3. **Lưu trữ và quản lí dữ liệu:** Sau khi hoàn thành quá trình xử lí, dữ liệu có thể được lưu trữ trong các hệ thống phù hợp như Data Lake, Data Warehouse, ...

2.1.2.4 Thách thức

Khi thiết kế đường ống dữ liệu, cần phải giải quyết một số thách thức để duy trì hiệu quả của kiến trúc, bao gồm:

- Đảm bảo chất lượng dữ liệu: Việc duy trì chất lượng dữ liệu trong suốt quá trình xử lý là điều cần thiết, vì dữ liệu chất lượng kém có thể dẫn đến thông tin chi tiết không chính xác và dự đoán lỗi.

- Giảm độ phức tạp khi tích hợp: Việc tích hợp các nguồn dữ liệu và công nghệ khác nhau là một công việc phức tạp, đòi hỏi phải lập kế hoạch và phân tích chuyên sâu để tránh các vấn đề về khả năng tương thích và các xung đột tích hợp dữ liệu khác.

- Duy trì bảo mật và quyền riêng tư dữ liệu: Việc bảo vệ dữ liệu nhạy cảm trong suốt quá trình truyền dữ liệu, từ khi thu thập đến khi phân tích, thường là yêu cầu bắt buộc về mặt tuân thủ pháp lý, đòi hỏi phải phân tích và thực hiện cẩn thận liên tục.

- Giải quyết các hạn chế về khả năng mở rộng: Khi khối lượng dữ liệu tăng lên, một đường ống dữ liệu được thiết kế kém sẽ nhanh chóng trở nên không bền vững và khó quản lý. Đảm bảo chất lượng dữ liệu phù hợp và phân phối theo thời gian đòi hỏi phải cân nhắc cẩn thận về cơ sở hạ tầng dữ liệu cơ bản so với khối lượng dữ liệu hiện tại và dự kiến trong tương lai.

2.1.3 Data Lake - Data Warehouse

2.1.3.1 Data Lake

Hồ dữ liệu là kho lưu trữ tập trung thu thập và lưu trữ khối lượng lớn dữ liệu ở dạng ban đầu. Sau đó, dữ liệu có thể được xử lý và sử dụng làm cơ sở cho nhiều nhu cầu phân tích khác nhau. Do có kiến trúc mở và có

thể mở rộng, hồ dữ liệu có thể chứa mọi loại dữ liệu từ mọi nguồn, từ dữ liệu có cấu trúc (bảng, trang tính) đến dữ liệu bán cấu trúc (tệp XML, trang web) đến dữ liệu không có cấu trúc (hình ảnh, tệp âm thanh,...).
Các tệp dữ liệu thường được lưu trữ dưới dạng dữ liệu thô, đã làm sạch và được quản lý — để các loại người dùng khác nhau có thể sử dụng dữ liệu ở nhiều dạng khác nhau để đáp ứng nhu cầu của họ. Hồ dữ liệu cung cấp tính nhất quán của dữ liệu trên nhiều ứng dụng khác nhau như hỗ trợ phân tích dữ liệu lớn , học máy , phân tích dự đoán,...

2.1.3.2 Data Warehouse

Data Warehouse(kho dữ liệu) là một hệ thống lưu trữ dữ liệu được thiết kế để tập trung vào việc lưu trữ dữ liệu theo từng chủ đề và được tối ưu hóa cho việc truy vấn và phân tích dữ liệu, chứ không chỉ tập trung riêng cho giao dịch.
Trong kho dữ liệu, dữ liệu được lưu trữ theo một trật tự sắp xếp và tổ chức rõ ràng, giúp cho việc truy vấn và phân tích dữ liệu trở nên hiệu quả hơn.
Một số chức năng của Data Warehouse:

- Lưu trữ dữ liệu: Data Warehouse cung cấp hệ thống lưu trữ tập trung, nơi tất cả dữ liệu từ nhiều nguồn khác nhau được tích hợp và sắp xếp một cách có hệ thống.
- Phân tích và tổng hợp dữ liệu: Data Warehouse có khả năng phân tích và tổng hợp dữ liệu theo nhiều chiều khác nhau, giúp đánh giá xu hướng từ dữ liệu.
- Hỗ trợ ra quyết định: Nhờ khả năng tổng hợp và phân tích dữ liệu, Data Warehouse giúp người dùng ra quyết định dựa trên thông tin thực tế, từ đó cải thiện hiệu quả của chiến lược.
- Tích hợp dữ liệu từ nhiều nguồn: Data Warehouse cho phép tích hợp dữ liệu từ nhiều nguồn khác nhau như hệ thống CRM, ERP hoặc các ứng dụng quản lý khác.

2.1.3.3 Khác biệt giữa Data Lake và Data Warehouse

Bảng 2.1: Bảng khác biệt giữa Data Lake và Data Warehouse

Tiêu chí	**Data Lake**	**Data Warehouse**
Loại dữ liệu	Cấu trúc, bán cấu trúc, phi cấu trúc	Cấu trúc
Quan hệ dữ liệu	Dữ liêu có quan hệ, phi quan hệ	Dữ liệu có quan hệ
Xử lý	Dữ liệu thô, có thể chưa qua xử lý	Dữ liệu đã được xử lý
Khả năng mở rộng	Dễ dàng mở rộng với chi phí thấp	Khó mở rộng và tốn kiếm chi phí nhiều hơn
Ứng dụng	Học máy. phân tích, dự đoán, ...	Làm báo cáo, BI, xây dựng dashboard,...

2.1.4 Trực quan hóa dữ liệu

Trực quan hóa dữ liệu (Data Visualization) giúp biến các tập dữ liệu phức tạp thành biểu đồ, đồ thị dễ hiểu để người xem có thể phân tích thông tin và hỗ trợ việc ra quyết định.
Các công cụ trực quan hóa phổ biến bao gồm Tableau, Power BI, và Matplotlib. Trực quan hóa không chỉ giúp phân tích xu hướng dữ liệu mà còn làm nổi bật các mối quan hệ và mô hình tiềm ẩn trong dữ liệu.

Hình 2.1: Minh họa Data Visualization

Chương 3

Công nghệ và kiến trúc áp dụng vào Data Pipeline

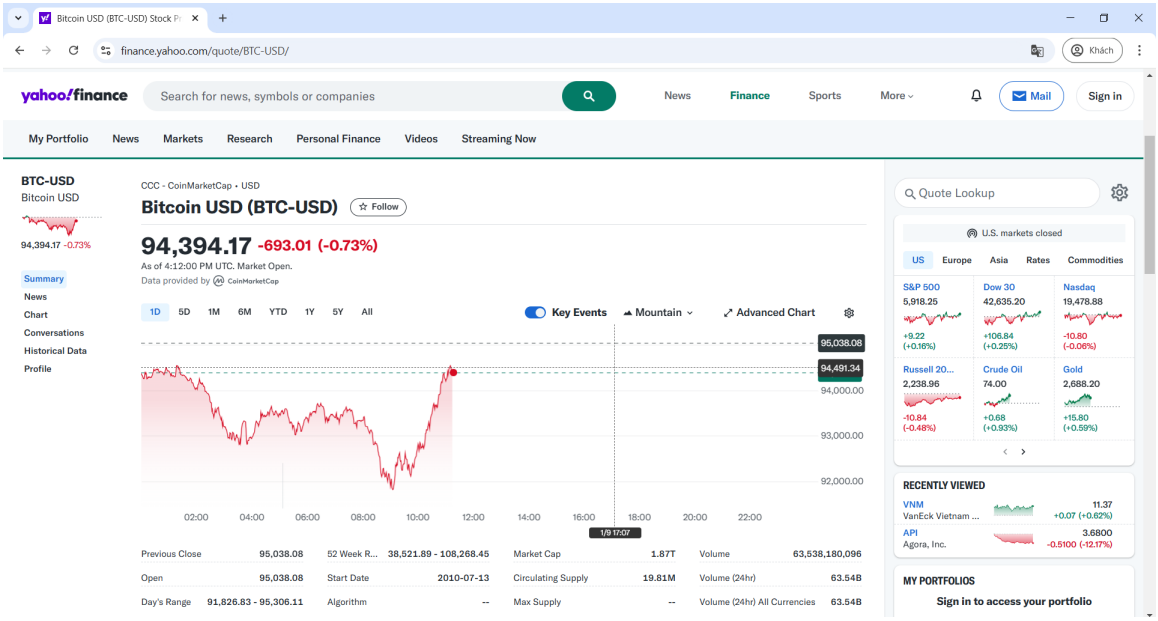
3.1 Yahoo Finance

Yahoo Finance là một trong những nguồn cung cấp dữ liệu tài chính phổ biến và đáng tin cậy nhất hiện nay. Dữ liệu từ Yahoo Finance bao gồm giá cổ phiếu, khối lượng giao dịch, thông tin thị trường, và các chỉ số tài chính khác.

Vai trò

- Thu thập dữ liệu thị trường chứng khoán: Yahoo Finance cung cấp API, cho phép tự động hóa việc thu thập dữ liệu theo thời gian thực.
- Đa dạng nguồn dữ liệu: Bao gồm dữ liệu lịch sử, dữ liệu thời gian thực, và các chỉ số phân tích tài chính khác.
- Kết nối với các hệ thống phân tích: Dữ liệu từ Yahoo Finance thường được sử dụng làm đầu vào cho các hệ thống phân tích lớn hơn như Azure Data Lake và Azure Databricks.

Trong dự án này, Yahoo Finance đóng vai trò là nguồn cung cấp dữ liệu đầu vào cho Data Pipeline. Dữ liệu thu thập được từ nền tảng này là nền tảng để thực hiện các bước phân tích, dự đoán và trực quan hóa.



Hình 3.1: Minh họa Yahoo Finance

3.2 Azure Portal

Portal Azure được hiểu đơn giản là một nền tảng điện toán đám mây với một cổng thông tin duy nhất của Microsoft Azure, cho phép người dùng truy cập và quản lý toàn bộ ứng dụng, dịch vụ và tài nguyên được lưu trữ.

Thông qua phần mềm này, chỉ với một bảng điều khiển duy nhất người dùng có thể theo dõi, cấu hình và triển khai các dịch vụ đám mây, có thể thực hiện việc quản lý dữ liệu cho đến việc xây dựng các giải pháp công nghệ một cách rất nhanh chóng và dễ dàng. Điều này sẽ giúp cho người dùng có thể tối ưu hóa quy trình quản lý và triển khai môi trường điện toán đám mây của mình hiệu quả hơn trên Microsoft Azure.
Phần mềm Portal Microsoft Azure cùng lúc cung cấp tới hơn 200 dịch vụ khác nhau và được chia làm 18 loại chính bao gồm các mảng như máy tính, lưu trữ, mạng, di động, phân tích, di chuyển, trí tuệ nhân tạo, bảo mật... Dưới đây là 03 dịch vụ chính được sử dụng thường xuyên nhất hiện nay.

- **Tính Toán**
  - Azure Virtual Machines: Chỉ trong vài giây thao tác, máy chủ ảo đã được tạo trong hệ điều hành Windows, Linux...

- Dịch vụ điện toán đám mây: Tạo ứng dụng và mở rộng trên đám mây, sau khi ứng dụng này được triển khai, toàn bộ các giai đoạn từ cung cấp, theo dõi đến phân tích kết quả đều có thể thực hiện trên phần mềm Azure.
- Service Fabric: Toàn bộ microservice – ứng dụng chứa nhiều ứng dụng nhỏ sẽ được đơn giản hóa rất nhiều bước.
- Các Hàm: Sử dụng ngôn ngữ lập trình đa dạng bằng các hàm để tạo ứng dụng trên Portal Azure mà không đòi hỏi các yêu cầu liên quan đến phần cứng.
- Azure Kubernetes Service: Giúp người dùng triển khai, quản lý và mở rộng các ứng dụng container của mình một cách nhanh chóng và dễ dàng.
- Azure Batch: Cho phép người dùng thực hiện các công việc tính toán lớn theo lô, với tính khả dụng và khả năng mở rộng linh hoạt.
- Azure Container Instances: Giúp người dùng chạy các container trên đám mây Azure mà không cần đến việc triển khai và quản lý một cụm Kubernetes.
- Azure Virtual Machine Scale Sets: Cho phép người dùng tự động mở rộng các máy ảo dựa trên lưu lượng của ứng dụng.
- ...

- **Kết nối mạng**
  - Azure CDN: Cung cấp toàn bộ nội dung cho người dùng và có thể gửi đến bất kỳ người dùng nào ở bất cứ khu vực nào. Dịch vụ này còn sử dụng một mạng lưới máy chủ để có thể truy cập dữ liệu với tốc độ nhanh chóng và chính xác hơn.
  - Express Route: Cho phép kết nối mạng với bộ nhớ lưu trữ đám mây của Microsoft hoặc với bất kỳ dịch vụ nào khác thông qua kết nối ở chế độ riêng tư.
  - Azure Virtual Network (VNet): Cho phép người dùng tạo ra

các mạng ảo, giúp cùng lúc nhiều dịch vụ Azure có thể kết nối với nhau an toàn và riêng tư.

- Azure DNS: Dịch vụ này cho phép host các DNS domain hoặc domain của hệ thống.
- Azure VPN Gateway: Cung cấp cho người dùng khả năng kết nối mạng an toàn giữa các mạng VNet trong hệ thống Azure và mạng riêng của họ.
- ...

- **Lưu trữ dữ liệu**
  - Disk Storage: Cho người dùng chuyển từ HDD hoặc SSD làm tùy chọn lưu trữ với hệ thống máy ảo. phép
  - Blob Storage: Tối ưu hóa lưu trữ dữ liệu bao gồm cấu trúc, văn bản và dữ liệu nhị phân.
  - File Storage: Dịch vụ quản lý file dữ liệu và được truy cập thông qua giao thức SMB.
  - Queue Storage: Dịch vụ lưu trữ số lượng lớn tin nhắn và truy cập bất kỳ vị trí nào.
  - Azure Table Storage: Là dịch vụ lưu trữ dữ liệu không cấu trúc và có khả năng mở rộng, giúp phù hợp cho việc lưu trữ dữ liệu có cấu trúc đơn giản.
  - Azure Data Lake Storage: Là dịch vụ lưu trữ dữ liệu lớn, dùng để giúp người dùng lưu trữ, phân tích và xử lý một lượng dữ liệu lớn, bao gồm cả những dữ liệu có cấu trúc và dữ liệu không cấu trúc.
  - Azure SQL Database: Là dịch vụ cơ sở dữ liệu trên đám mây, giúp cung cấp cho người dùng các khả năng mở rộng, bảo mật cùng với tính khả dụng cao cho dữ liệu có cấu trúc.
  - ...

3.3 Azure Functions

Azure Functions là một dịch vụ serverless của Microsoft Azure, cho phép thực thi các đoạn mã (functions) mà không cần quản lý máy chủ. Đây là công cụ lý tưởng để tự động hóa các tác vụ xử lý dữ liệu trong pipeline.

Tổng quan:

- Azure Functions giúp triển khai các ứng dụng nhỏ, có khả năng mở rộng linh hoạt mà không cần thiết lập cơ sở hạ tầng phức tạp.
- Dịch vụ này được kích hoạt dựa trên sự kiện (event-driven), giúp tối ưu hóa quy trình tự động hóa

Công dụng:

- Xử lý sự kiện: Thu thập và xử lý dữ liệu từ API hoặc các nguồn khác khi có sự kiện xảy ra.
- Tự động hóa quy trình: Thực hiện các tác vụ như thu thập dữ liệu định kỳ, kích hoạt các bước trong Data Pipeline.
- Tích hợp hệ thống: Kết nối với các dịch vụ khác trong hệ sinh thái Azure như Data Lake, DataBricks.

Azure Functions được sử dụng để tự động thu thập dữ liệu từ Yahoo Finance và chuyển đến Data Lake. Nó đảm bảo việc thu thập dữ liệu diễn ra liên tục và chính xác, giảm thiểu can thiệp thủ công.

3.4 Azure Data Lake Gen 2

Azure Data Lake Gen 2 là một dịch vụ lưu trữ dữ liệu phân tán, hỗ trợ các loại dữ liệu có cấu trúc (structured data), bán cấu trúc (semi-structured data), và phi cấu trúc (unstructured data). Đây là nền tảng chính để lưu trữ dữ liệu trong pipeline.
Data Lake Gen 2 được thiết kế đặc biệt cho khối lượng dữ liệu lớn (big data), cho phép lưu trữ, truy xuất và quản lý dữ liệu hiệu quả. Bên cạnh

đó,giúp tổ chức dữ liệu thành các lớp (bronze, silver, gold), từ đó tối ưu hóa quy trình phân tích.

Vai trò trong Data Pipeline

- Lưu trữ dữ liệu thô: Dữ liệu từ Yahoo Finance được lưu trữ tại lớp bronze của Data Lake.
- Quản lý và tổ chức dữ liệu: Dữ liệu được tổ chức thành các lớp (bronze, silver, gold) để dễ dàng xử lý và phân tích.
- Tích hợp với các dịch vụ khác: Azure Data Lake tích hợp tốt với Azure Databricks và các công cụ phân tích như Power BI.

Kiến trúc lưu trữ cơ bản

- Lớp bronze: Lưu trữ dữ liệu thô chưa qua xử lý.
- Lớp silver: Lưu trữ dữ liệu đã làm sạch và chuẩn hóa.
- Lớp gold: Lưu trữ dữ liệu tổng hợp, sẵn sàng cho phân tích và trực quan hóa.

3.5 Azure DataBricks

Azure DataBricks là một nền tảng phân tích dữ liệu lớn dựa trên Apache Spark, được tối ưu hóa để tích hợp trong hệ sinh thái Azure. Nó hỗ trợ xử lý dữ liệu lớn, học máy, và các tác vụ phân tích chuyên sâu.

Công dụng

- Xử lý dữ liệu lớn: Làm sạch, chuyển đổi và phân tích dữ liệu từ các nguồn khác nhau.
- Hỗ trợ học máy: Xây dựng và huấn luyện các mô hình dự đoán.
- Tự động hóa: Chạy các job xử lý dữ liệu định kỳ.
- Tích hợp liền mạch: Kết nối với Azure Data Lake, Power BI, và các dịch vụ Azure khác.

Azure DataBricks được sử dụng để thực hiện quá trình ELT (Extract, Load, Transform), giúp chuyển đổi dữ liệu từ trạng thái thô thành dữ liệu sẵn sàng phân tích. Ngoài ra, nền tảng này còn hỗ trợ xây dựng các mô hình học máy để dự đoán xu hướng thị trường.

3.6 Power BI

Power BI là sự kết hợp hoàn hảo giữa phân tích kinh doanh (Business Analytics) và trực quan hóa dữ liệu (Data Visualization), giúp các tổ chức đưa ra những quyết định hiệu quả và tối ưu hơn trong tương lai. Công cụ này giúp người dùng xây dựng các biểu đồ và đồ thị dựa trên dữ liệu có sẵn để trực quan hóa dữ liệu.

Power BI có khả năng thu thập dữ liệu từ hàng trăm nguồn dữ liệu khác nhau như trang web, cơ sở dữ liệu, mạng xã hội,... và xuất bản các báo cáo một cách an toàn, có tính bảo mật cao cũng như trích xuất thông tin kinh doanh nhanh chóng.

Các thành phần chính của Power BI:

- Power Query: là công cụ chuyển đổi và tổ hợp dữ liệu. Cho phép người dùng khám phá, kết nối, kết hợp và tùy chỉnh các nguồn dữ liệu để đáp ứng cho nhu cầu phân tích. Power Query được cài đặt dưới dạng Add-in (tiện ích mở rộng) cho Excel hoặc có thể dùng như một phần của Power BI Desktop.

- Power Pivot là một kỹ thuật mô hình hóa dữ liệu cho phép người dùng tạo mô hình dữ liệu, thiết lập mối quan hệ và tạo phép tính. Vơi Power Pivot, bạn có thể làm việc với các tập dữ liệu lớn (Big Data), xây dựng quan hệ rộng và tao các phép tinh, từ đơn giản đến phức tạp. Nó sử dụng ngôn ngữ Data Analysis Expressions (DAX) để lập các mô hình dữ liệu đơn giản hoặc phức tạp.

- Power View: công nghệ trưc quan hoa dư liêu có sẵn trong Microsoft Excel, Sharepoint, SQL Server và Power BI. Công cụ này cho phép

người dùng tạo biểu đồ, đồ thị, bản đồ và các hình ảnh trưc quan khac có tính tương tác cao để các dữ liệu trở nên sống động hơn. Ngoài ra, nó cũng có thể kết nối với nhiều nguồn dữ liệu và lọc dữ liệu cho từng phần tử (element), sau đó trực quan hóa dữ liệu hoặc toàn bộ báo cáo.

- Microsoft Power Map: dành cho Excel và Power BI là công cụ trực quan hóa dữ liệu ba chiều (3D), cho phép người dùng xem thông tin theo nhiều cách thức mới lạ và khám phá những điều mà có thể bạn chưa từng nhìn thấy trong các bảng hay biểu đồ hai chiều truyền thống.

- Power QA: Tính năng QA (Question Answer) trong Power BI cho phép người dùng khám phá dữ liệu bằng khả năng trực quan, ngôn ngữ tự nhiên và nhận được câu trả lời dưới dạng biểu đồ hay đồ thị. Nó có thể xuất ra kết quả dữ liệu mà bạn muốn chỉ bằng cách đặt những câu hỏi đơn giản.

Hạn chế khi sử dụng Power BI:

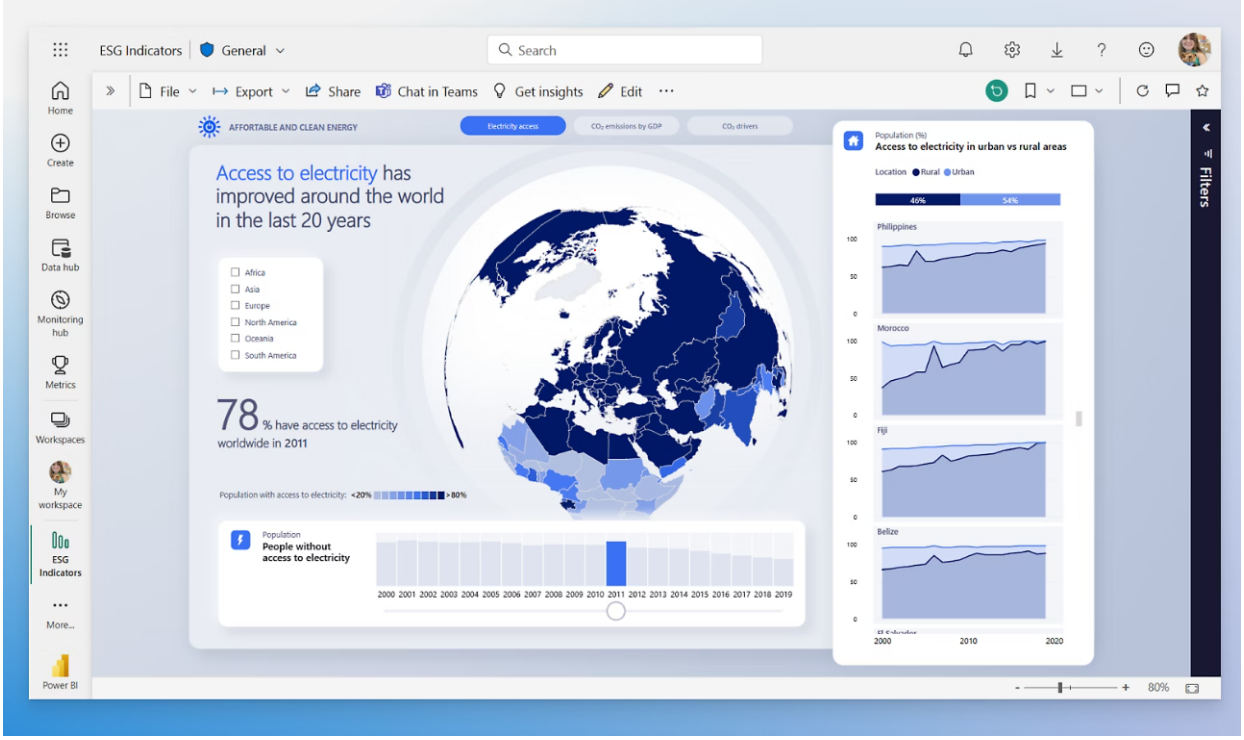
- Liên kết bảng biểu chưa hoàn hảo: Power BI có thể xử lý các mối quan hệ đơn giản giữa các bảng trong mô hình dữ liệu rất tốt. Tuy nhiên, khi có các mối quan hệ phức tạp giữa các bảng, tức là, nếu có nhiều hơn một liên kết giữa các bảng, Power BI sẽ gặp khó khăn trong việc xử lý chúng. Vì vậy, người dùng cần đảm bảo rằng các mô hình dữ liệu bổ sung chỉ có một trường duy nhất để Power BI không bị nhầm lẫn giữa các bảng.

- Cần có kiến thức khi sử dụng công thức DAX: Ngôn ngữ biểu thức được sử dụng để xử lý dữ liệu trong Power BI là DAX và người dùng có thể thực hiện rất nhiều hành động thông qua việc viết công thức DAX. Tuy nhiên, đây không phải là ngôn ngữ dễ sử dụng nhất, vì khi kết nối nhiều hơn hai phần tử (elements) sẽ khiến các câu lệnh bị lồng vào nhau.

- Phức tạp: Về cơ bản, Power BI là một công cụ trực quan và tương đối đơn giản để nhập dữ liệu hay tạo báo cáo. Tuy nhiên, khi mục đích sử

dụng không chỉ để tạo báo cáo trong Power BI Desktop, người dùng sẽ phải tìm hiểu và thành thạo một số công cụ khác như Gateways, Power BI Report Server, Power BI Services,...

- Gặp vấn đề khi xử lý khối lượng siêu dữ liệu lớn: Nhiều người dùng đã báo cáo rằng Microsoft Power BI mất nhiều thời gian hơn thông thường, thậm chí bị treo máy khi phải xử lý khối lượng dữ liệu khổng lồ mà không biết tối ưu chúng cũng như lựa chọn cách thức import phù hợp.

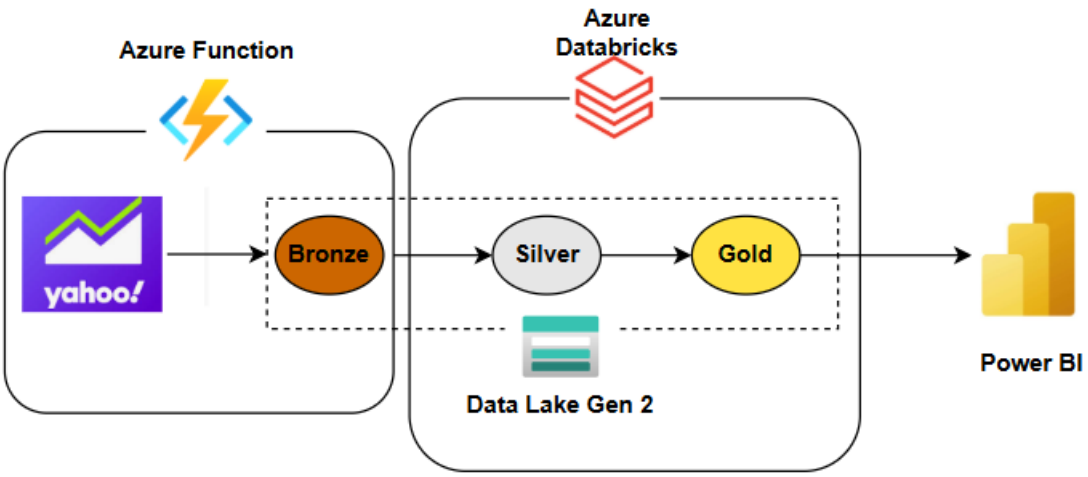


Hình 3.2: Minh họa Power BI

Chương 4

Triển khai Data Pipeline

4.1 Kiến trúc Data Pipeline



Hình 4.1: Kiến trúc Data Pipeline

Mô tả kiến trúc Data Pipeline

Data Pipeline được chia làm 3 quá trình chính là: thu thập dữ liệu, chuyển đổi và xử lí dữ liệu, trực quan hóa dữ liệu

Đầu tiên là quá trình thu thập dữ liệu, dữ liệu sẽ được lấy trực tiếp từ API là Yahoo Finance và đưa trực tiếp vào lớp bronze trong Data Lake Gen 2 bằng việc viết các hàm lấy dữ liệu từ API bằng Azure Function. Tại quá trình này cũng sẽ thiết lập thời gian định kì để dữ liệu được cập nhật vào Data Lake Gen 2.

Ngay sau khi dữ liệu được đưa vào lớp bronze, quá trình chuyển đổi dữ liệu sẽ bắt đầu. Dữ liệu trong bronze sẽ lần lượt được được xử lý các lỗi

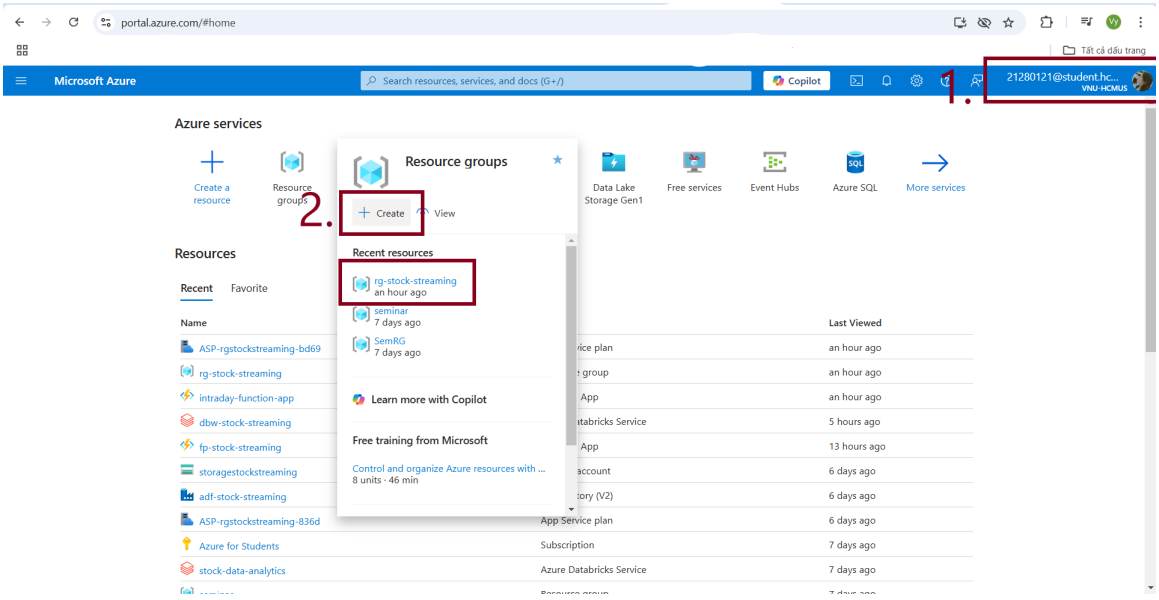
và định dạng lại qua các lớp silver và lớp gold. Mỗi lần thực hiện chuyển đổi dữ liệu từ bronze qua silver và từ silver qua gold được thực hiện bởi Azure DataBricks. DataBricks sẽ được thiết lập thời quan để tự động xử lí dữ liệu mới được thêm vào từ lớp bronze.

Cuối cùng khi có dữ liệu đã qua xử lý tại lớp gold, ta sẽ kết nối Power BI với dữ liệu tại lớp gold để tiến hành trực quan hóa dữ liệu.

4.2 Thiết lập các tài nguyên cần thiết trên Azure Portal

Các bước cần thiết trước khi tiến hành thiết kế Data Pipeline:

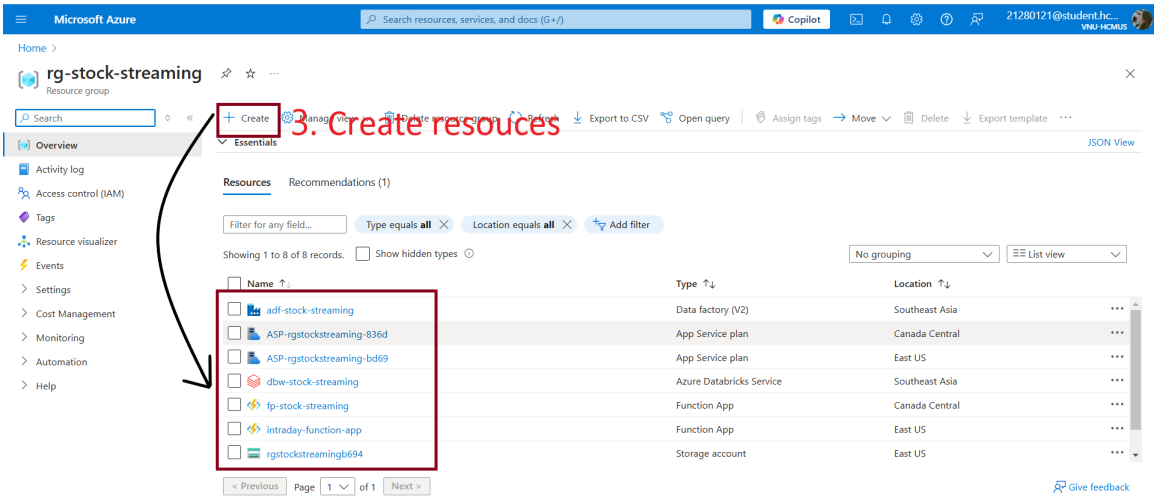
1. Đăng ký tài khoản phần mềm Microsoft Azure Portal.
2. Tạo Resource Group - nơi chứa các dịch vụ cần thiết để tạo Data Pipeline.



Hình 4.2: Minh họa Resource Group

Trong hình 4.2 thể hiện đã đăng kí tài khoản Miccrosoft Azure và tạo Resource Group **"rg-stock-streaming"**.

3. Trong Resource Group, tạo các resource(tài nguyên) như Data Lake Gen 2(Strorage Account), DataBricks, Azure Function và thiết lập các thông số phù hợp với yêu cầu bài toán.



Hình 4.3: Minh họa tạo các Resource

4.3 Thu thập dữ liệu

Giai đoạn thu thập dữ liệu(Data Ingestion) là bước đầu tiên trong quy trình Data Pipeline. Mục đích của giai đoạn này là thu thập dữ liệu từ các nguồn khác nhau, bao gồm:

- Thu thập dữ liệu từ API của Yahoo bằng **Azure Function**.
- Lưu trữ dữ liệu thô vào lớp *Bronze* trong **Data Lake Gen 2**.

4.3.1 Thu thập dữ liệu bằng Azure Function:

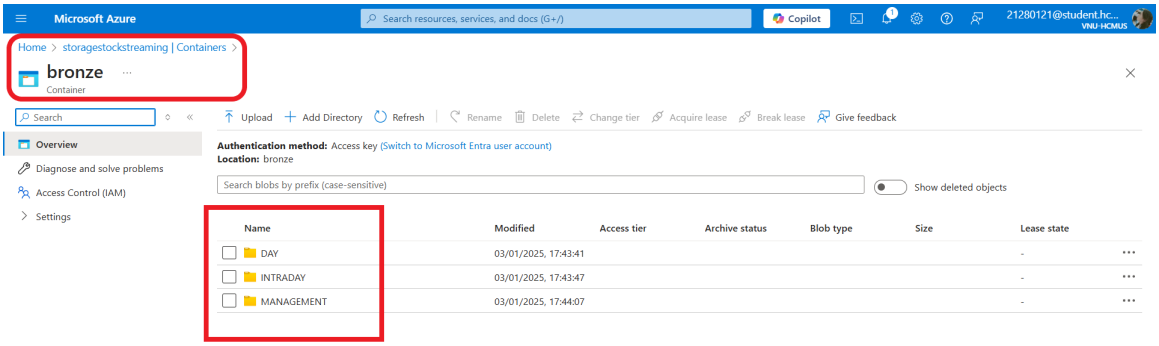
Quá trình này được thực hiện trên máy local bằng việc viết các cripts để thực hiện các chức năng:

- Kết nối Azure Function với Data Lake Gen 2
- Lấy data từ Yahoo Finance
- Cài đặt trigger để tự động lấy dữ liệu mới sau 1 khoảng thời gian 5 phút.

4.3.2 Đưa dữ liệu vào lớp bronze

Việc đưa dữ liệu vào lớp Bronze trong Data Lake là một bước quan trọng trong quá trình thu thập và xử lý dữ liệu. Lớp Bronze đóng vai trò là nơi lưu trữ dữ liệu thô (raw data) mà không thực hiện bất kỳ thay đổi hay xử lý nào, giúp đảm bảo rằng dữ liệu gốc được bảo toàn và có thể sử dụng lại khi cần.

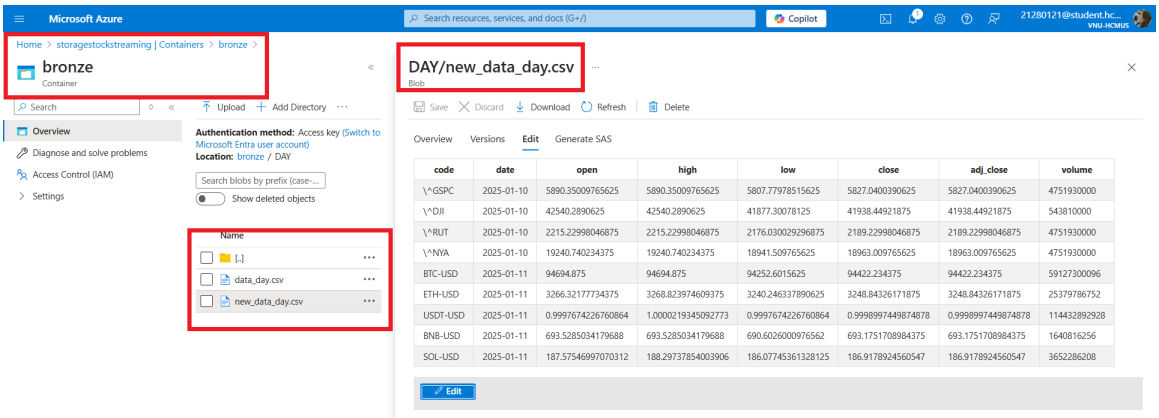
Dữ liệu được đưa vào bronze được chia làm 3 thư mục: **DAY, INTRA-DAY, MANAGEMENT**



Hình 4.4: Các thư mục trong lớp bronze

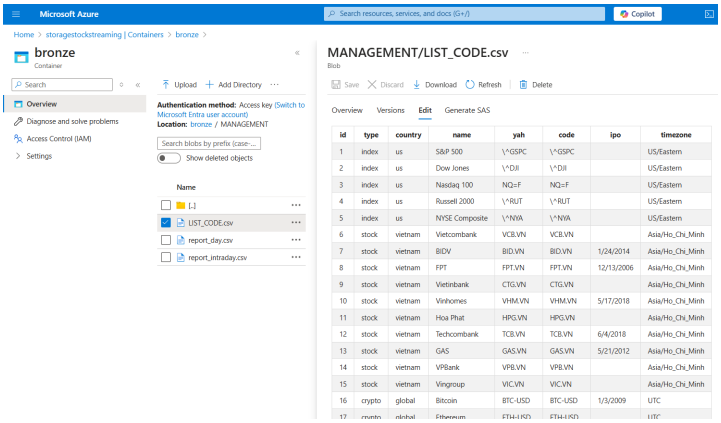
Tại 2 thư mục DAY và INTRADAY có cấu trúc tệp bên trong tương tự nhau: Gồm 2 tập tin là:

- Tập **data_day(data_intraday)**: chứa dữ liệu từ lần lấy thông tin trước đó về xa nhất có thể (total data)
- Tập **new_data_day (new_data_intraday)**: chứa dữ liệu mới được cập nhật trong lần lấy thông tin hiện tại (current data)



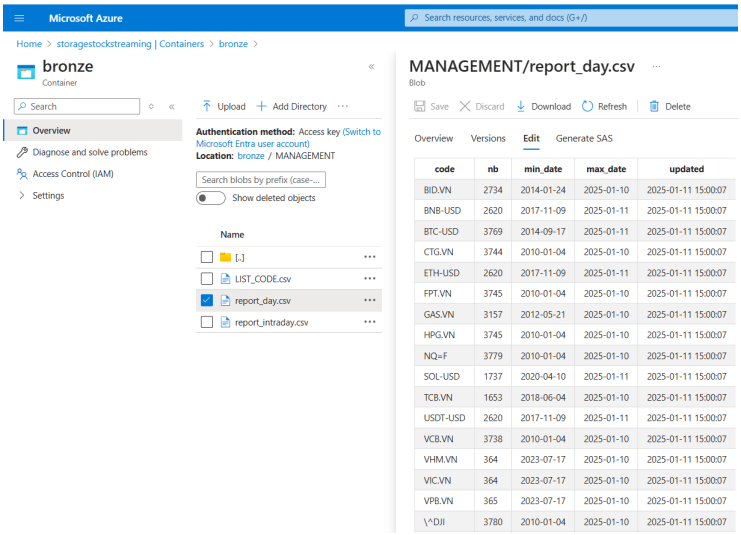
Hình 4.5: Các dữ liệu trong new_data_day

Sau khi cập nhật dữ liệu vào 'current data', dữ liệu sẽ được đưa về cùng định dạng với dữ liệu trong 'total data'. Sau đó, các dư liệu trong 'current data' sẽ được thêm vào 'total data' và bắt đầu các bước chuyển đổi qua các lớp silver và gold sau đó.
Với thư mục MANAGEMENT: chứa các tập **LIST_CODE** có nhiệm vụ quản lí các thông tin cơ về các mã chứng khoán



Hình 4.6: Tập tin LIST_CODE

Và **report_day** và **report_intraday** để quản lí thời gian cập nhật mới nhất của từng mã chứng khoán



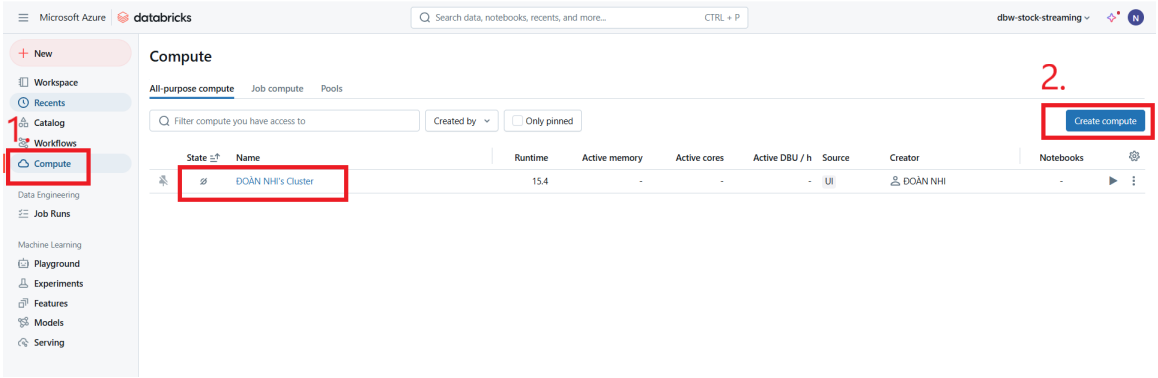
Hình 4.7: Tập tin report_day

4.4 Chuyển đổi và xử lí dữ liệu

Chuyển đổi và xử lý dữ liệu (Data Transformation) là giai đoạn quan trọng để làm sạch, chuẩn hóa và tổ chức dữ liệu trước khi đưa vào phân tích hoặc sử dụng.
Quá trình chuyển đổi dữ liệu được thực hiện trên Azure DataBricks

4.4.1 Cài đặt DataBricks

Sau khi có resource DataBricks trong Resource Group đã tạo trước đó, việc đầu tiên cần thực hiện là tạo 1 cluster mới



Hình 4.8: Tạo cluster

Đầu tiên, vào tab compute và nhấn vào nút *Create Compute* để tạo 1 cluster mới. Sau đó tiến hành đặt tên cluster và thiết lập các thông số phù hợp với đề tài. Theo hình 4.8, đã tạo một cluster là **ĐOÀN NHI's Cluster** để làm việc với DataBricks. Sau khi có được cluster, ta có thể thực thi các notebook để xử lí dữ liệu tiếp theo.

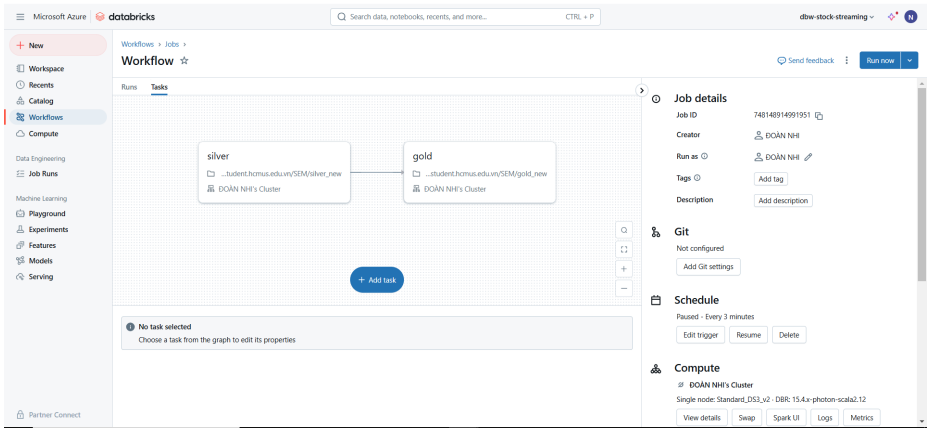
4.4.2 Chuyển đổi dữ liệu từ lớp Bronze sang Silver:

- **Làm sạch dữ liệu**: Xử lý các giá trị bị thiếu (missing values) và loại bỏ dữ liệu trùng lặp, không hợp lệ.
- **Chuẩn hóa dữ liệu**: Chuyển đổi dữ liệu sang định dạng chuẩn để đảm bảo tính đồng nhất.

4.4.3 Chuyển đổi dữ liệu từ lớp Silver sang Gold:

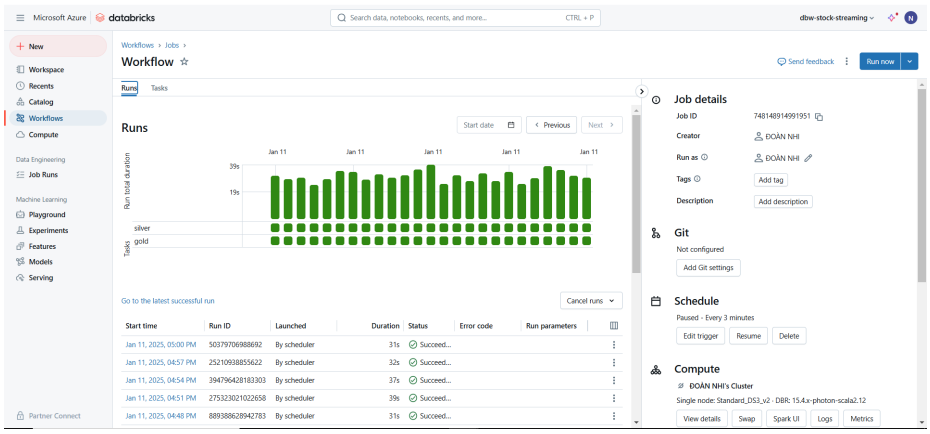
- **Tạo các chỉ số**: Tính toán các chỉ số (metrics) hoặc các biến mới cần thiết cho phân tích.
- **Tối ưu hóa dữ liệu**: Tổ chức và lưu trữ dữ liệu dưới dạng tối ưu (như định dạng Delta Lake) để tăng hiệu suất truy vấn.
- **Lưu trữ dữ liệu phân tích**: Dữ liệu đã qua xử lý được lưu trữ trong lớp *Gold* để sẵn sàng sử dụng cho trực quan hóa hoặc mô hình phân tích.

Cả hai quá trình được thực hiện bằng cách thực thi 2 notebook (silver_new và gold_new) đã được lập trình.



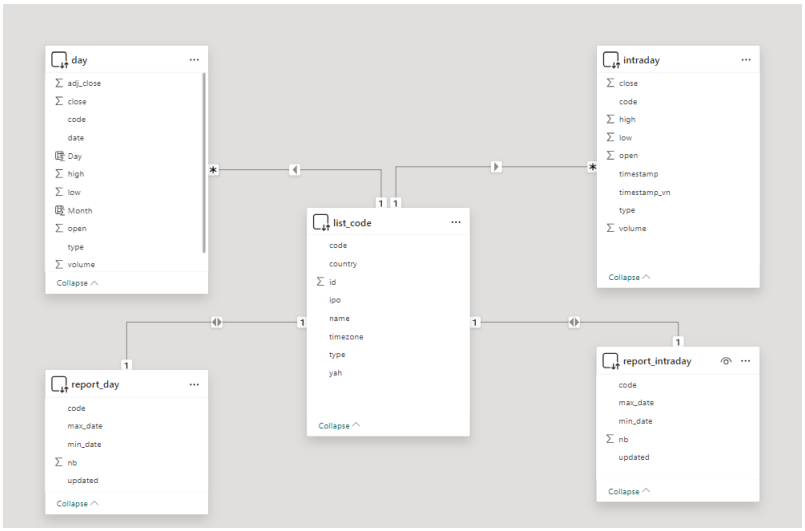
Hình 4.9: Thiết lập workflow để thực thi 2 notebook

Và được thiết lập thời gian thực thi mỗi 3 phút. Việc thiết lập này đảm bảo rằng dữ liệu mới mỗi khi được cập nhật từ API vào bronze mỗi 5 phút đều sẽ được tiến hành xử lí và lưu vào lớp silver



Hình 4.10: Kết quả triển khai workflow

Sau khi kết thúc mỗi lần cập nhật, ta đều có được cơ sở dữ liệu có cấu trúc như hình 4.11



Hình 4.11: Final Database Schema

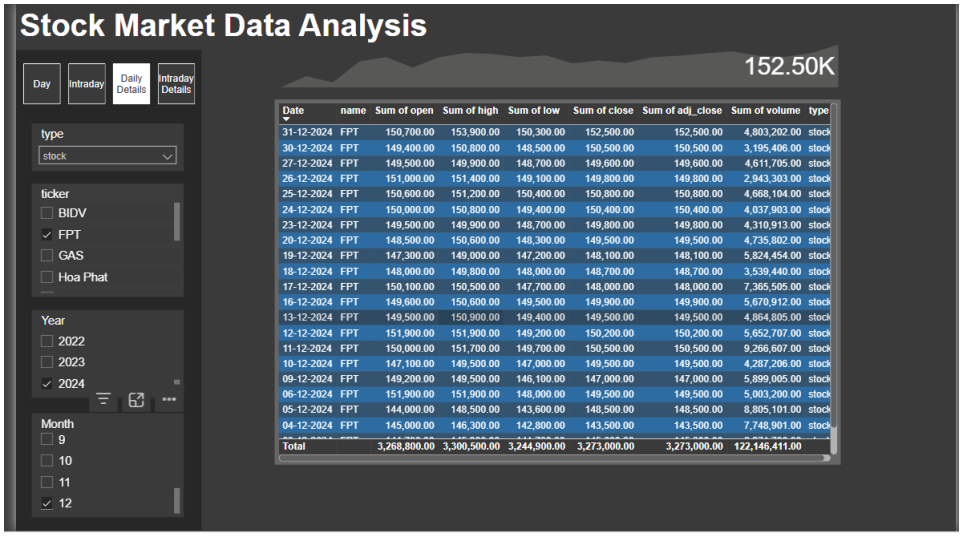
4.5 Trực quan hóa dữ liệu

Trực quan hóa dữ liệu (Data Visualization) giúp người dùng hiểu rõ hơn về thông tin thông qua các biểu đồ và bảng phân tích. Trong pipeline này, **Power BI** được sử dụng để:

4.5.1 Các loại biểu đồ và báo cáo:

- **Biểu đồ đường (Line Chart)**: Hiển thị xu hướng giá theo thời gian.

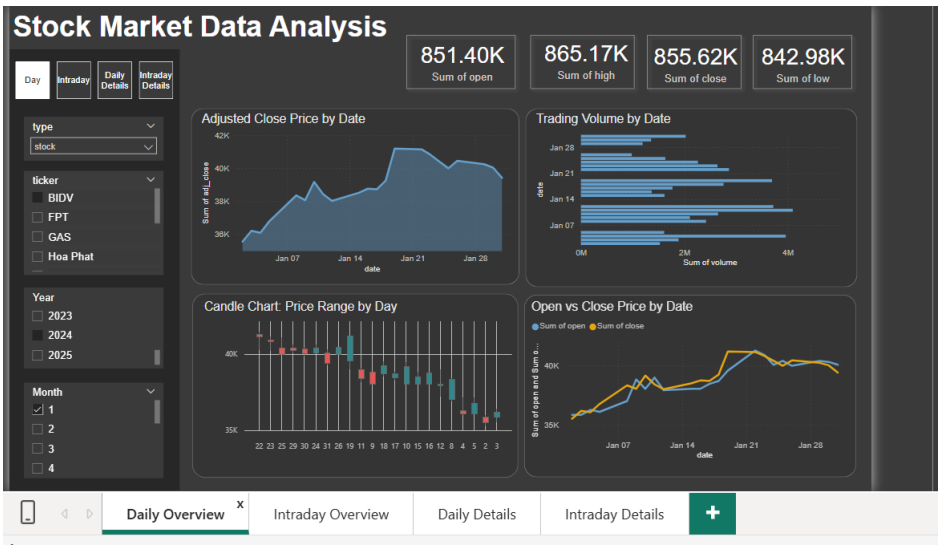
- **Biểu đồ nến**: khối lượng giao dịch của các mã cổ phiếu.
- **Báo cáo tổng hợp**: Cung cấp thông tin tổng quan về hiệu suất của toàn bộ danh mục đầu tư.



Hình 4.12: Báo cáo tổng hợp

4.5.2 Tích hợp Power BI với Data Lake:

- Kết nối trực tiếp Power BI với lớp *Gold* trong Data Lake Gen 2 để lấy dữ liệu mới nhất.
- Cập nhật tự động: Thiết lập lịch trình cập nhật để luôn hiển thị dữ liệu mới nhất.
- Tùy chỉnh dashboard: Xây dựng dashboard phù hợp với nhu cầu của từng người dùng hoặc nhóm phân tích.



Hình 4.13: Trực quan dữ liệu chứng khoán

Chương 5

Thảo luận và Kết luận

5.1 Kết quả

Dự án triển khai Data Pipeline đã đạt được các kết quả sau:

- Xây dựng thành công kiến trúc Data Pipeline với các thành phần chính bao gồm Azure Function, Data Lake Gen 2, Azure Databricks, và Power BI.
- Thu thập dữ liệu tự động từ Yahoo Finance và lưu trữ dữ liệu thô vào lớp *Bronze* của Data Lake Gen 2.
- Chuyển đổi và xử lý dữ liệu từ lớp *Bronze* sang *Silver* và *Gold* để làm sạch, chuẩn hóa và tổ chức dữ liệu.
- Trực quan hóa dữ liệu thành công bằng Power BI, cung cấp các báo cáo và biểu đồ phân tích trực quan, hỗ trợ ra quyết định kinh doanh.

5.2 Hướng phát triển

Dựa trên các kết quả đã đạt được, có thể mở rộng và cải tiến hệ thống Data Pipeline theo các hướng sau:

- Tích hợp thêm các nguồn dữ liệu mới để tăng cường tính đa dạng và độ chính xác của dữ liệu.
- Áp dụng các mô hình học máy (Machine Learning) trên dữ liệu ở lớp *Gold* để dự đoán và phân tích xu hướng.
- Tăng cường bảo mật dữ liệu trong Data Lake Gen 2, đảm bảo tuân thủ các tiêu chuẩn bảo mật và quyền riêng tư.

- Tự động hóa toàn bộ Data Pipeline, bao gồm cả giai đoạn triển khai và giám sát, nhằm giảm thiểu sự can thiệp thủ công.
- Nâng cao khả năng mở rộng của hệ thống để xử lý khối lượng dữ liệu lớn hơn trong tương lai.

5.3 Tổng kết

Dự án này đã cung cấp một giải pháp toàn diện cho việc thu thập, xử lý, và phân tích dữ liệu thông qua Data Pipeline. Các kết quả đạt được không chỉ giúp tối ưu hóa quy trình làm việc với dữ liệu mà còn tạo nền tảng vững chắc để triển khai các ứng dụng phân tích dữ liệu nâng cao trong tương lai. Với các hướng phát triển đã đề xuất, hệ thống có tiềm năng mở rộng và cải tiến, đáp ứng tốt hơn các nhu cầu kinh doanh ngày càng phức tạp. Sự thành công của dự án là minh chứng cho tầm quan trọng của việc xây dựng một hệ thống quản lý và xử lý dữ liệu hiệu quả.

Tài liệu tham khảo

[1] Ansh Lamba. *End-to-End Data Engineering Projects*. URL: https://www.youtube.com/@AnshLambaJSR (visited on 01/17/2025).

[2] Microsoft. *Azure Databricks documentation*. URL: https://learn.microsoft.com/en-us/azure/databricks/ (visited on 01/17/2025).

[3] Microsoft. *Azure Functions documentation*. URL: https://learn.microsoft.com/en-us/azure/azure-functions/ (visited on 01/17/2025).

[4] Microsoft. *Introduction to Azure Data Lake Storage*. URL: https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-introduction (visited on 01/17/2025).

[5] Microsoft. *Power BI documentation*. URL: https://learn.microsoft.com/en-us/power-bi/ (visited on 01/17/2025).

Chương A

Ngữ pháp tiếng Anh

Tiếng Anh	**Dịch nghĩa**
Data Pipeline	Luồng dữ liệu
API - Application Programming Interface	Giao diện lập trình ứng dụng
Dashboard	Bảng điều khiển
Data Lake	Hồ dữ liệu
Data Warehouse	Kho dữ liệu
DAX - Data Analysis Expressions	Biểu thức phân tích dữ liệu