viettel

TẬP ĐOÀN CÔNG NGHIỆP – VIỄN THÔNG QUÂN ĐỘI

BÁO CÁO MINI-PROJECT

Phân tích thông tin dữ liệu thương mại điện tử trên sàn Shopee

TRẦN NGỌC BẢO
tnbao120603@gmail.com.

Chương trình Viettel Digital Talent 2024
Lĩnh vực: Data Engineering

Mentor: _____Nguyễn Anh Dũng_____

Đơn vị: _____Viettel AI_____

HÀ NỘI, 06/2024

Lời mở đầu

Xin gửi lời cảm ơn chân thành đến Ban Tổ chức Chương trình Viettel Digital Talent 2024 và các anh chị trong Lĩnh vực Data Engineering đã tạo cơ hội để em tham gia và học hỏi. Đặc biệt, em rất vinh dự khi được chia sẻ về chủ đề "Phân tích thông tin dữ liệu thương mại điện tử trên sàn Shopee". Đây là một lĩnh vực đầy tiềm năng và thử thách, mở ra nhiều hướng nghiên cứu và ứng dụng thực tiễn.

Sự hỗ trợ và hướng dẫn tận tình của các anh mentor đã giúp em có cái nhìn sâu sắc và toàn diện hơn, từ đó góp phần nâng cao năng lực chuyên môn và định hướng nghề nghiệp. Một lần nữa, xin chân thành cảm ơn các anh chị trong Ban Tổ chức Chương trình Viettel Digital Talent 2024 nói riêng và Tập đoàn Công nghiệp - Viễn thông Quân đội Viettel nói chung!

Tóm tắt nội dung và đóng góp

Chủ đề "Phân tích thông tin dữ liệu thương mại điện tử trên sàn Shopee" nghiên cứu sâu về cách khai thác và hiểu biết dữ liệu từ một trong những nền tảng thương mại điện tử hàng đầu khu vực. Trong bối cảnh thương mại điện tử đang phát triển mạnh mẽ, việc phân tích dữ liệu từ Shopee không chỉ giúp hiểu rõ hơn về hành vi mua sắm của người tiêu dùng mà còn cung cấp các insights quan trọng để cải thiện chiến lược kinh doanh và tiếp thị.

Chủ đề này tập trung vào việc thu thập, xử lý và phân tích các loại dữ liệu khác nhau như số lượng sản phẩm bán ra, đánh giá của khách hàng, xu hướng tìm kiếm và hành vi mua sắm. Bằng cách áp dụng các kỹ thuật Data Engineering, em sẽ xây dựng các mô hình phân tích, từ đó đưa ra các báo cáo và đề xuất chiến lược dựa trên dữ liệu thực tế.

Một số kết quả đạt được trong mini-project: Thu thập dữ liệu từ sàn Shopee; Lưu trữ dữ liệu; Phân tích dữ liệu bán hàng; Các biểu đồ trực quan hóa dữ liệu.

Một số định hướng phát triển mở rộng khi thực hiện đề tài: Mở rộng cụm, triển khai trên hạ tầng thật thay nền tảng ảo hóa, tăng số lượng máy tính tham gia hệ thống.

Trong và sau khi thực hiện đề tài, em đã học được những kiến thức về thu thập, lưu trữ, xử lý và phân tích dữ liệu lớn. Cùng với đó, em cũng được làm quen với rất nhiều công nghệ trong từng bước xây dựng hệ thống: Môi trường phát triển (Ubuntu, Docker); Thu thập dữ liệu (Airflow, Restful API); Lưu trữ dữ liệu (Redis, MinIO, Hive, PostgreSQL); Phân tích và xử lý dữ liệu (Spark-Iceberg, Trino); Trực quan hóa dữ liệu (Superset).

Sinh viên thực hiện

Bảo
Trần Ngọc Bảo

MỤC LỤC

DANH MỤC HÌNH VẼ

CHƯƠNG 1. TỔNG QUAN ĐỀ TÀI

1.1 Giới thiệu chung

Trong bối cảnh thương mại điện tử ngày càng phát triển mạnh mẽ, việc phân tích thông tin dữ liệu từ các sàn giao dịch như Shopee trở nên vô cùng quan trọng. Đề tài "Phân tích thông tin dữ liệu thương mại điện tử trên sàn Shopee" nhằm mục tiêu khai thác, xử lý và hiểu rõ các thông tin dữ liệu từ nền tảng thương mại điện tử hàng đầu này. Thông qua việc phân tích dữ liệu, em mong muốn hiểu rõ hơn về hành vi tiêu dùng, xu hướng thị trường và các yếu tố ảnh hưởng đến hoạt động kinh doanh trên Shopee.

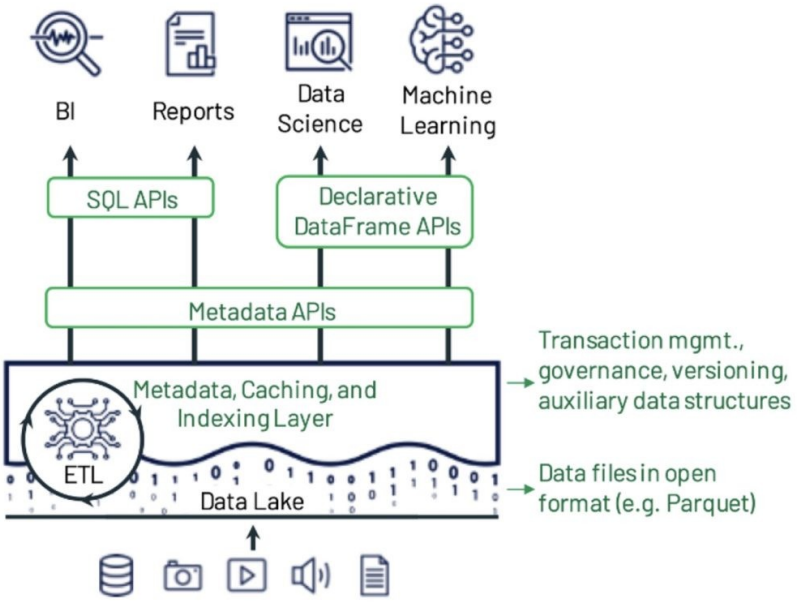
Shopee là một trong những sàn thương mại điện tử phổ biến nhất tại khu vực Đông Nam Á, với hàng triệu người dùng và lượng giao dịch khổng lồ mỗi ngày. Dữ liệu từ Shopee bao gồm nhiều khía cạnh khác nhau như thông tin sản phẩm, đánh giá của khách hàng, lượng bán hàng, và hành vi tìm kiếm của người dùng. Việc phân tích những dữ liệu này không chỉ giúp các doanh nghiệp tối ưu hóa chiến lược kinh doanh mà còn cải thiện trải nghiệm người dùng, tăng cường hiệu quả tiếp thị và tối đa hóa lợi nhuận.

Đề tài này tập trung vào việc thu thập, làm sạch và phân tích dữ liệu từ Shopee. Em sử dụng các kỹ thuật và công cụ tiên tiến trong lĩnh vực Data Engineering để xử lý và trực quan hóa dữ liệu. Qua đó, em xác định được các xu hướng mua sắm, đánh giá hiệu quả của các chiến dịch tiếp thị và hiểu rõ hơn về hành vi tiêu dùng. Những kết quả này không chỉ cung cấp các thông tin chi tiết và chính xác về thị trường mà còn đưa ra các đề xuất chiến lược giúp doanh nghiệp nâng cao năng lực cạnh tranh.

Việc phân tích dữ liệu thương mại điện tử trên Shopee không chỉ dừng lại ở việc khai thác thông tin mà còn hướng đến việc ứng dụng thực tiễn trong kinh doanh. Những insights thu được từ quá trình phân tích sẽ giúp các doanh nghiệp đưa ra các quyết định thông minh và hiệu quả hơn. Đồng thời, nó cũng mở ra những hướng nghiên cứu mới, đóng góp vào sự phát triển bền vững và sáng tạo của ngành thương mại điện tử trong thời đại số hóa.

1.2 Mô hình hệ thống

Mô hình Data Lake là một giải pháp hiệu quả để quản lý và phân tích dữ liệu trong đề tài "Phân tích thông tin dữ liệu thương mại điện tử trên sàn Shopee". Mô hình này được áp dụng để xây dựng một kho lưu trữ dữ liệu linh hoạt và có thể mở rộng, cho phép chúng tôi tích hợp và lưu trữ dữ liệu từ nhiều nguồn khác nhau một cách có tổ chức.



Hình 1: Mô hình Data Lake (Nguồn: Internet)

Đầu tiên, Data Lake cho phép em thu thập dữ liệu từ Shopee thông qua các công cụ scraping và API, lưu trữ các tập dữ liệu gốc mà không cần tiền xử lý nhiều. Dữ liệu này có thể bao gồm thông tin sản phẩm, đánh giá của người dùng, lượng bán hàng, và các thông tin khác liên quan đến hành vi mua sắm trên nền tảng.

Tiếp theo, em sử dụng công nghệ lưu trữ MinIO (tương tự Amazon S3) để lưu trữ dữ liệu trong Data Lake. Đây là nơi chúng tôi có thể lưu trữ các tập dữ liệu lớn, bao gồm cả dữ liệu cấu trúc và bán cấu trúc, mà không cần phải chuẩn hóa trước.

Sau khi dữ liệu được lưu trữ, em sử dụng các công cụ và framework như Trino để xử lý và phân tích dữ liệu. Trino cho phép chúng tôi thực hiện các phép tính phức tạp trên dữ liệu lớn một cách hiệu quả, từ phân tích thống kê đơn giản đến xây dựng các mô hình học máy phức tạp.

Mô hình Data Lake không chỉ giúp chúng tôi duy trì tính toàn vẹn và khả năng mở rộng của dữ liệu mà còn cung cấp sự linh hoạt trong việc truy xuất và sử dụng dữ liệu. Điều này rất quan trọng trong việc nghiên cứu và phân tích dữ liệu thương mại điện tử từ Shopee, vì chúng tôi cần có khả năng tiếp cận và phân tích các loại dữ liệu đa dạng và lớn lượng.

Tóm lại, việc áp dụng mô hình Data Lake trong đề tài này không chỉ giúp tối ưu hóa quy trình xử lý dữ liệu mà còn mở ra nhiều cơ hội cho việc nghiên cứu sâu sắc và phát triển ứng dụng trong lĩnh vực thương mại điện tử.

1.3 Phương pháp thực hiện

Để thực hiện đề tài "Phân tích thông tin dữ liệu thương mại điện tử trên sàn Shopee", em đã áp dụng một quy trình nghiên cứu chi tiết và khoa học, bao gồm các bước chính như thu thập dữ liệu, làm sạch dữ liệu, lưu trữ dữ liệu, phân tích dữ liệu và trực quan hóa dữ liệu. Mỗi bước được thực hiện bằng các phương pháp và công cụ tiên tiến nhằm đảm bảo tính chính xác và hiệu quả của kết quả nghiên cứu.

Bước đầu tiên trong quy trình là thu thập dữ liệu. Em sử dụng các kỹ thuật web scraping để lấy dữ liệu từ Shopee. Dữ liệu bao gồm thông tin sản phẩm, đánh giá của khách hàng, số lượng bán ra. Công cụ phổ biến như Apache Airflow được sử dụng để tự động hóa quá trình này, giúp thu thập một lượng lớn dữ liệu một cách nhanh chóng và hiệu quả.

Sau khi thu thập dữ liệu, bước tiếp theo là lưu trữ dữ liệu. Em sử dụng các công cụ như Iceberg, MinIO để lưu trữ dữ liệu tương tự như Amazon S3.

Bước phân tích dữ liệu là trọng tâm của đề tài. Các công cụ như Trino và SQL được sử dụng để thực hiện các phép tính phức tạp, phân tích mô tả. Em tập trung vào việc hiểu rõ hành vi mua sắm của người tiêu dùng, xác định các sản phẩm bán chạy trên Shopee.

Cuối cùng, em trực quan hóa kết quả phân tích để dễ dàng truyền đạt các phát hiện và đề xuất. Công cụ Superset được sử dụng để tạo ra các biểu đồ, đồ thị và dashboard tương tác. Những hình ảnh trực quan này giúp trình bày các kết quả một cách rõ ràng và dễ hiểu, hỗ trợ các doanh nghiệp trong việc đưa ra các quyết định chiến lược dựa trên dữ liệu thực tế.

Quy trình thực hiện đề tài không chỉ đảm bảo tính chính xác và khoa học mà còn tạo điều kiện cho việc ứng dụng kết quả nghiên cứu vào thực tiễn. Những kết quả phân tích từ đề tài này không chỉ giúp các doanh nghiệp tối ưu hóa hoạt động kinh doanh trên Shopee mà còn mở ra những hướng nghiên cứu mới trong lĩnh vực thương mại điện tử.

CHƯƠNG 2. XÂY DỰNG HỆ THỐNG

Sau khi xác định được mô hình hệ thống và phương pháp thực hiện, em tiến hành xây dựng hệ thống phân tích thông dữ liệu thương mại điện tử trên sàn Shopee được tiến hành theo từng bước nhỏ lần lượt: Thu thập dữ liệu; Lưu trữ dữ liệu; Xử lý dữ liệu; Trực quan hóa dữ liệu.

2.1 Thu thập dữ liệu

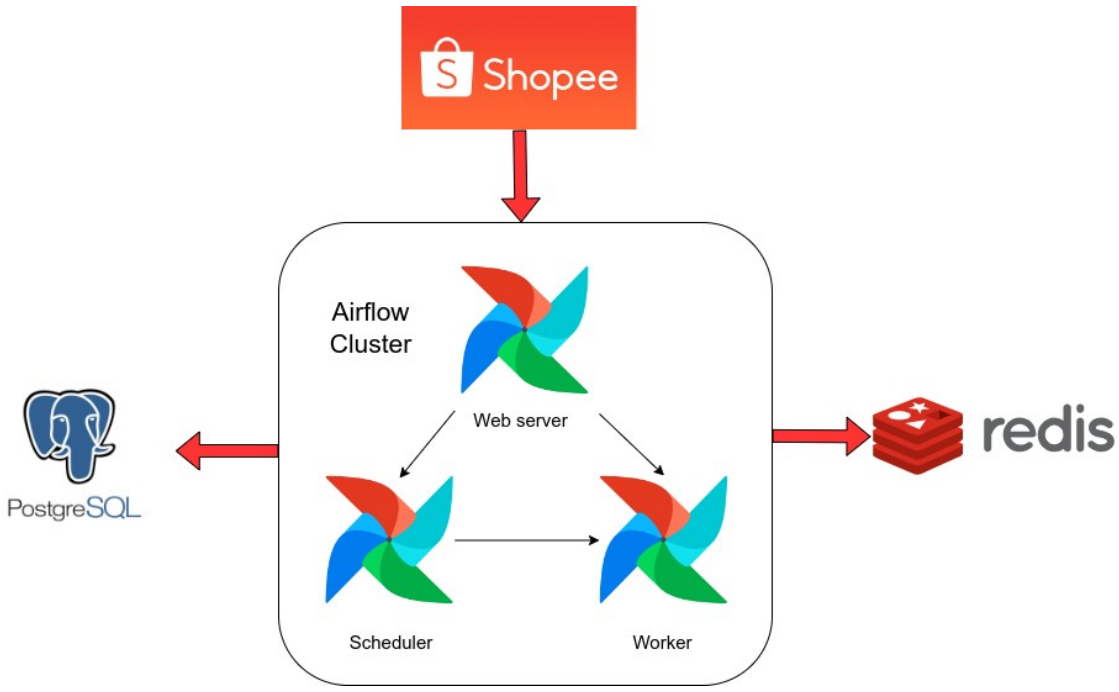
Xây dựng nên một cụm Apache Airflow gồm có 1 node Airflow-Webserver, 1 node Airflow-Scheduler và 1 node Airflow-Worker. Node web server trong Apache Airflow là giao diện người dùng chính để quản lý và giám sát các công việc lập lịch. Nó cung cấp giao diện web cho người dùng để xem và quản lý các lịch trình công việc, xem trạng thái thực thi và xem các logs. Người dùng có thể tạo, sửa đổi và xóa các lịch trình công việc (workflows) thông qua giao diện này. Node web server cũng quản lý xác thực và phân quyền để đảm bảo an toàn và bảo mật trong quản lý công việc.

Scheduler là thành phần quản lý lịch trình và thực thi các công việc theo lịch đã định. Nhiệm vụ chính của scheduler là đọc và lập lịch các công việc dựa trên các lịch trình được định nghĩa bởi người dùng thông qua giao diện web server. Scheduler đảm bảo rằng các công việc được thực thi đúng thời điểm và theo đúng thứ tự đã được định sẵn. Nó kiểm soát và phân phối các công việc đến các worker để thực thi.

Worker là các tiến trình thực thi các công việc trong Apache Airflow. Mỗi worker có nhiệm vụ lắng nghe và nhận các công việc được lập lịch từ scheduler. Khi nhận được công việc, worker sẽ thực thi các nhiệm vụ (tasks) được xác định trong lịch trình công việc. Các công việc có thể là các nhiệm vụ ETL (Extract, Transform, Load), tính toán, gửi email, hay bất kỳ tác vụ nào mà người dùng đã định nghĩa.

Bên cạnh đó, em còn dựng lên 1 node PostgreSQL để lưu trữ metadata của cụm Apache Airflow. PostgreSQL được sử dụng để lưu trữ metadata của Apache Airflow, bao gồm thông tin về các lịch trình công việc (DAGs), các tasks, dependencies giữa các tasks, và lịch sử thực thi công việc. PostgreSQL cũng có thể được cấu hình để lưu trữ logs của các công việc được thực thi. Điều này cung cấp một nền tảng ổn định và đáng tin cậy để xem lại và phân tích các logs trong quá trình thực thi công việc.

Cùng với 1 node Redis để lưu trữ dữ liệu trung gian, chuyển tiếp dữ liệu cho phần tiếp theo của hệ thống sau khi cụm Airflow thực hiện xong nhiệm vụ. Redis có thể hoạt động như một message broker cho các worker trong Airflow.



Hình 2: Thu thập dữ liệu

Trong quá trình thu thập dữ liệu, em gặp phải một khó khăn đó là sàn thương mại điện tử Shopee không công khai toàn bộ API của mình. Muốn lấy được tất cả API của Shopee phải đăng ký trên trang Shopee Open Platform và phải thỏa mãn một trong hai tiêu chí: một Shopee Mall hoặc một công ty có đăng ký với Shopee.

Do đó, khi thu thập dữ liệu không thể gọi tự động một cách liên tục các API của Shopee, phải chia ra các khoảng thời gian cách nhau để gọi API từ Shopee và thực hiện thủ công bằng CURL trên máy cục bộ và mount dữ liệu vào node Airflow-Worker để chạy các task từ đồ thị DAG được xây dựng nên bởi code Python.

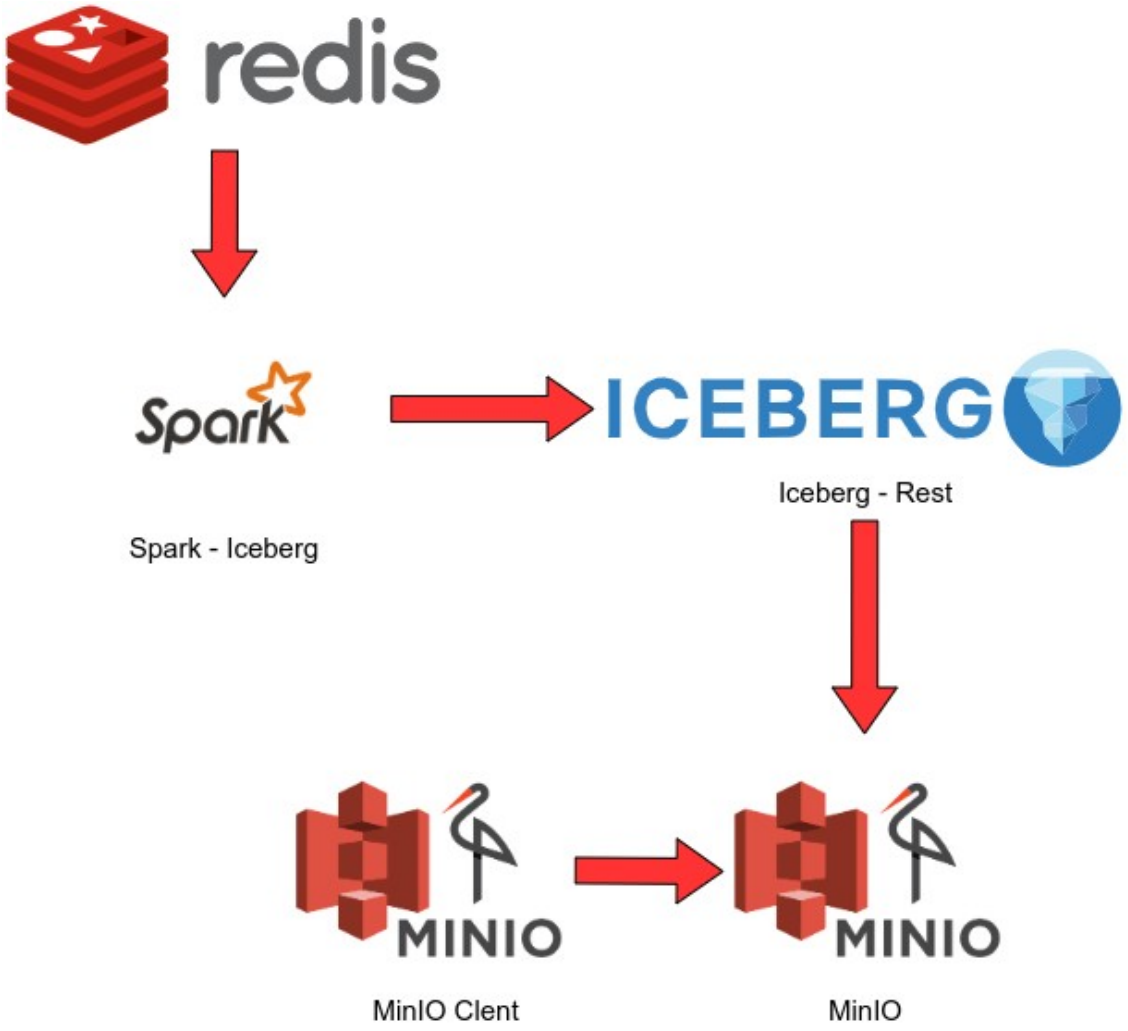
2.2 Lưu trữ dữ liệu

Triển khai một cụm Spark-Iceberg và Iceberg-Rest để tiếp nhận dữ liệu dữ liệu từ Redis và tiếp tục quá trình lưu trữ dữ liệu. Spark-Iceberg là một tích hợp của Apache Spark với dự án Iceberg, một dạng lưu trữ dữ liệu phân tán hiệu quả và đáng tin cậy. Nó cho phép lưu trữ các tệp dữ liệu lớn và quản lý chúng dưới dạng bảng có cấu trúc. Iceberg-Rest là một API REST đơn giản và linh hoạt để truy cập và quản lý dữ liệu lưu trữ trong hệ thống Iceberg. Nó cung cấp các endpoint để thực hiện các thao tác như đọc dữ liệu, ghi dữ liệu, quản lý phiên bản và metadata của bảng.

Đồng thời dụng lên cụm MinIO và MinIO client là thành phần chính trong bước lưu trữ dữ liệu. MinIO là một hệ thống lưu trữ đối tượng mã nguồn mở, được thiết kế để phục vụ nhu cầu lưu trữ đối tượng dữ liệu lớn với tính khả dụng cao và hiệu suất tối ưu. Nó hỗ trợ giao thức lưu trữ đối tượng S3, cho phép các

ứng dụng và dịch vụ có thể lưu trữ và truy xuất dữ liệu một cách dễ dàng và hiệu quả. MinIO được xây dựng để có khả năng mở rộng ngang (horizontal scaling) linh hoạt. Người dùng có thể dễ dàng mở rộng dung lượng lưu trữ và tăng cường khả năng xử lý bằng cách thêm các nodes MinIO vào hệ thống.

MinIO Client (mc) là một công cụ dòng lệnh cho phép người dùng quản lý và điều khiển các hệ thống MinIO từ xa. Nó cung cấp các lệnh để tải lên và tải xuống dữ liệu, quản lý bucket, quản lý các đối tượng, và thực hiện các thao tác quản lý hệ thống như sao chép, di chuyển và xóa dữ liệu. MinIO Client cho phép quản lý nhiều bucket trên nhiều server MinIO khác nhau từ một giao diện dòng lệnh duy nhất, đơn giản hóa quá trình quản lý và sử dụng hệ thống lưu trữ đối tượng.



Hình 3: Lưu trữ dữ liệu

Định dạng file Iceberg có rất nhiều ưu điểm nổi bật trong một hệ thống dữ liệu lớn. Iceberg lưu trữ metadata và schema của dữ liệu một cách cấu trúc và chi tiết. Nó cho phép người dùng quản lý các thông tin về cấu trúc dữ liệu, các thuộc tính của bảng và vị trí lưu trữ của các tệp dữ liệu. Iceberg cung cấp các cơ chế tối

ưu hóa như pruning (loại bỏ dữ liệu không cần thiết khi truy vấn), cắt tỉa (pushdown predicate) và tối ưu hóa lựa chọn cột (column selection) để cải thiện hiệu suất truy vấn dữ liệu. Điều này giúp cho việc xử lý dữ liệu trở nên nhanh hơn và hiệu quả hơn. Iceberg hỗ trợ đa nền tảng và có thể tích hợp với các công nghệ lưu trữ dữ liệu phân tán như Apache Hadoop, Amazon S3, Google Cloud Storage, và các hệ thống lưu trữ dữ liệu khác. Điều này làm cho Iceberg trở thành một lựa chọn phù hợp cho các hệ thống phân tán và đa nền tảng.

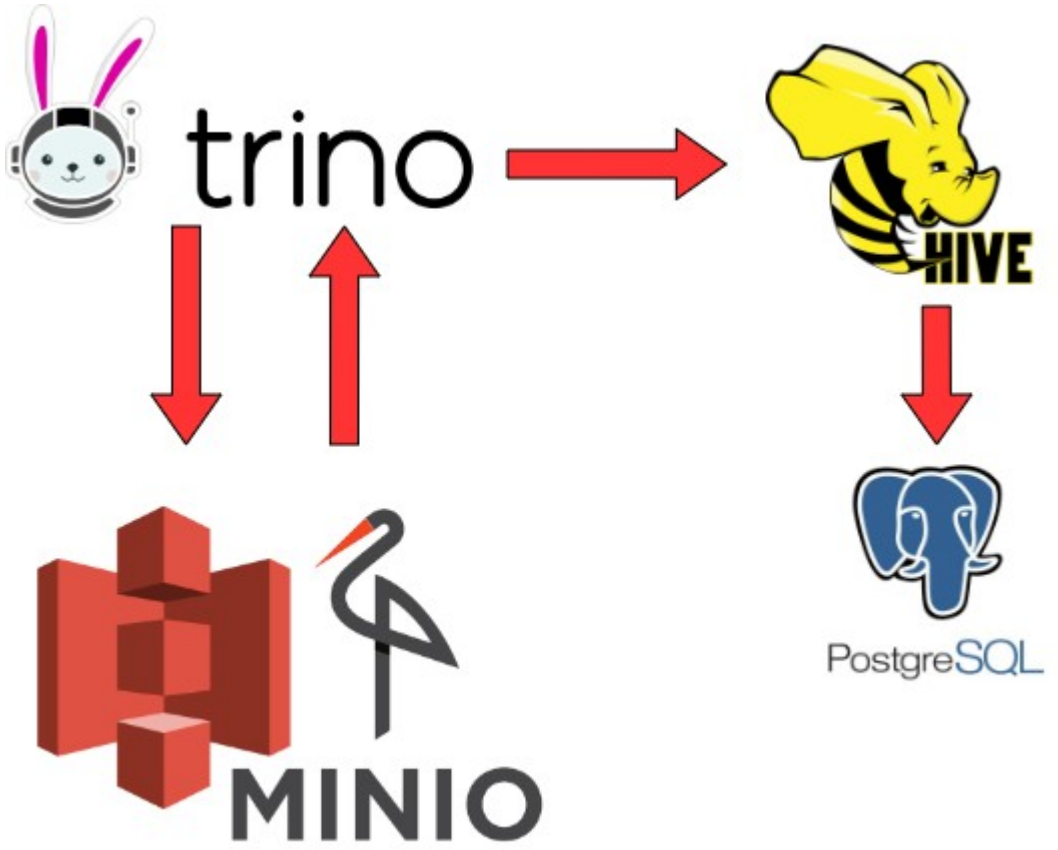
2.3 Xử lý dữ liệu

Sử dụng Trino làm công cụ để phân tích và xử lý dữ liệu. Trino (trước đây được biết đến là Presto SQL) là một hệ thống truy vấn phân tán mở và hiệu suất cao, được thiết kế để thực hiện các truy vấn phức tạp trên dữ liệu lớn, phân tán trên nhiều nguồn khác nhau. Trino cung cấp một cơ chế truy vấn mạnh mẽ và hiệu quả cho việc phân tích dữ liệu. Nó cho phép người dùng thực hiện các truy vấn phức tạp, bao gồm các phép join, lọc, tổng hợp và tính toán trên dữ liệu lớn mà không cần phải di chuyển hoặc sao chép dữ liệu.

Trino hỗ trợ truy vấn và tích hợp dữ liệu từ nhiều nguồn khác nhau như Hadoop HDFS, Amazon S3, Google Cloud Storage, cơ sở dữ liệu quan hệ (MySQL, PostgreSQL, Oracle), cũng như các hệ thống lưu trữ dữ liệu khác như Apache Cassandra, MongoDB và Redis. Trino cung cấp các kỹ thuật tối ưu hóa để cải thiện hiệu suất truy vấn, bao gồm cắt tỉa (pushdown) và sử dụng các chỉ số (index) để giảm số lượng dữ liệu cần quét khi thực hiện các truy vấn phức tạp. Điều này giúp giảm thời gian và tài nguyên tính toán cần thiết cho các hoạt động phân tích dữ liệu.

Đồng thời cũng dựng lên 1 node Apache Hive để làm metastore cho node Trino và thêm 1 node PostgreSQL để làm metastore database cho Hive. Apache Hive được sử dụng như một metastore cho Trino. Metastore trong Hive là nơi lưu trữ các metadata quan trọng về các bảng, các cột, định dạng dữ liệu, vị trí lưu trữ và các tham chiếu khác đến dữ liệu trong hệ thống. Trino sử dụng các thông tin này để hiểu cấu trúc dữ liệu và nơi lưu trữ của dữ liệu khi thực hiện các truy vấn. Hive Metastore là nơi tập trung quản lý metadata của các bảng dữ liệu. Điều này giúp Trino có thể truy vấn và xử lý dữ liệu từ Hive một cách hiệu quả, mà không cần phải biết chi tiết về cách dữ liệu được lưu trữ.

PostgreSQL được triển khai như một cơ sở dữ liệu để lưu trữ metadata của Hive. Hive sử dụng PostgreSQL để lưu trữ các thông tin về cấu trúc dữ liệu, quyền truy cập và quản lý thư mục. Điều này giúp tăng cường hiệu suất và tính nhất quán trong việc quản lý và truy xuất metadata của Hive.



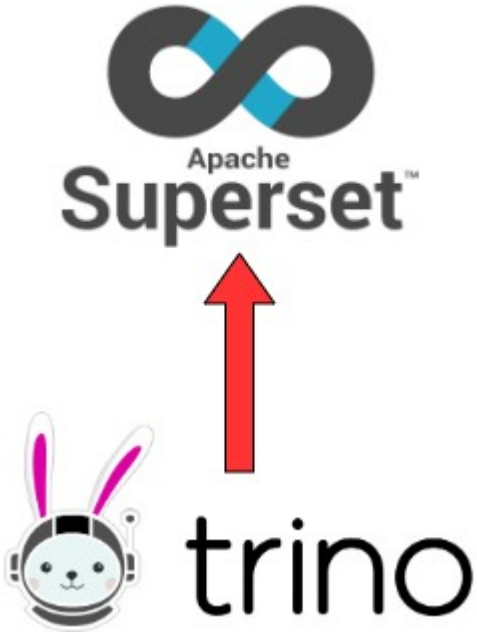
Hình 4: Xử lý dữ liệu

Trino truy vấn dữ liệu trực tiếp trên MinIO và lưu kết quả sau truy vấn ngược trở lại MinIO cùng với Hive và PostgreSQL lưu trữ metastore của Trino bằng cơ sở dữ liệu quan hệ có thể được coi như một Data Warehouse giúp hoàn hiện kiến trúc Data lake đã đề ra.

2.4 Trực quan hóa dữ liệu

Sử dụng Superset kết nối với Trino làm công cụ trực quan hóa dữ liệu để vẽ ra các biểu đồ, các bảng thân thiện, dễ quan sát hơn với người sử dụng. Superset cung cấp một giao diện trực quan và dễ sử dụng để khai thác và trực quan hóa dữ liệu từ nhiều nguồn khác nhau. Người dùng có thể tạo các biểu đồ, đồ thị, bản đồ và bảng điều khiển để hiển thị và phân tích dữ liệu một cách trực quan.

Superset hỗ trợ kết nối và tích hợp với nhiều nguồn dữ liệu khác nhau như cơ sở dữ liệu quan hệ (PostgreSQL, MySQL, Oracle), hệ thống lưu trữ dữ liệu phân tán (Apache Hive, Apache Druid, Presto), và các công cụ lưu trữ đối tượng (Amazon S3, Google Cloud Storage). Superset cho phép người dùng tạo và quản lý các bảng điều khiển (dashboards) để tổng hợp và hiển thị các thông tin quan trọng từ dữ liệu. Các bảng điều khiển này có thể được tùy chỉnh và chia sẻ để cung cấp thông tin phân tích cho người dùng khác.



Hình 5: Trực quan hóa dữ liệu

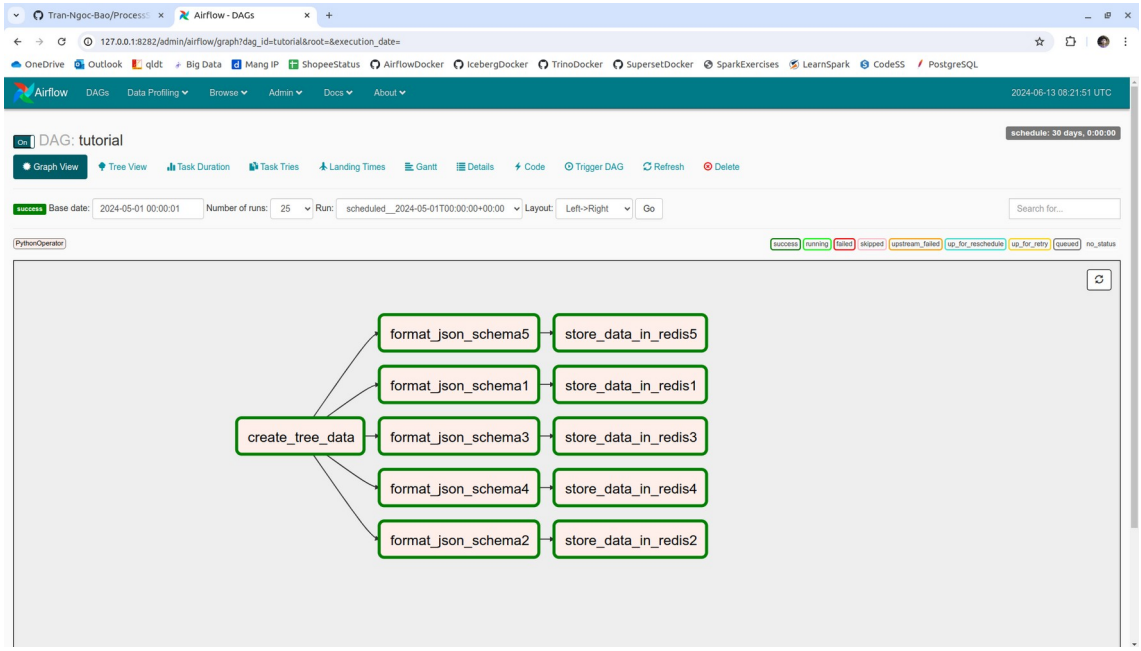
CHƯƠNG 3. KẾT QUẢ THU ĐƯỢC

Trong quá trình xây dựng hệ thống, sau khi thực hiện từng bước trong kế hoạch đề ra đã thu được một số kết quả cụ thể, từng phần. Và cũng đạt được các bảng, biểu đồ yêu cầu đề ra khi bắt đầu đề tài.

3.1 Hoạt động của cụm Airflow

Xây dựng đồ thị DAG cho cụm airflow bằng code Python gồm các bước:

- Tạo cây danh mục của Shopee bằng API, gồm các danh mục lớn và các danh mục nhỏ. Từ các danh mục nhỏ sẽ gọi tiếp API để thu thập dữ liệu các sản phẩm cụ thể.
- Format lại cấu trúc dữ liệu từ các file JSON một các hợp lý hơn, với các file JSON lấy được từ quá trình gọi API thủ công trên local bằng curl để hỗ trợ cho quá trình đọc dữ liệu từ Redis sau này của Spark.
- Chuyển dữ liệu sau khi format tới Redis và lưu trữ dưới dạng key – value.

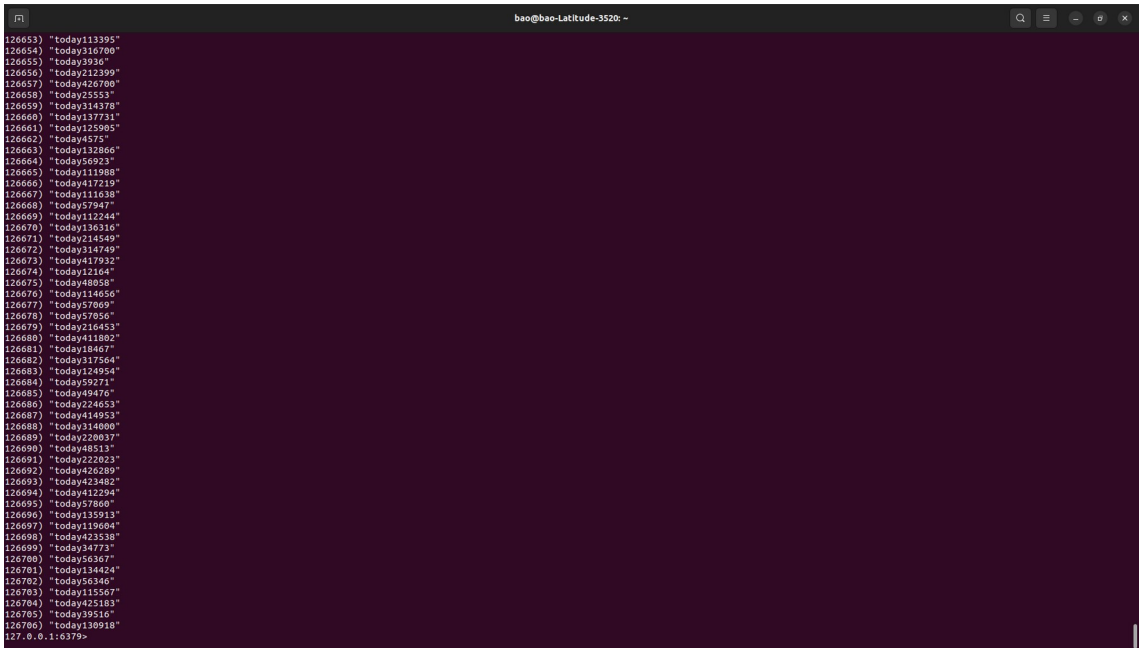


Hình 6: Hoạt động của cụm Airflow

3.2 Lưu trữ dữ liệu trên Redis

Dữ liệu sau khi đi qua cụm Airflow được đẩy tới Redis để lưu trữ dưới dạng key – value. Tương ứng với mỗi key là một value lưu trữ danh sách các sản phẩm cùng với thông tin chi tiết của sản phẩm.

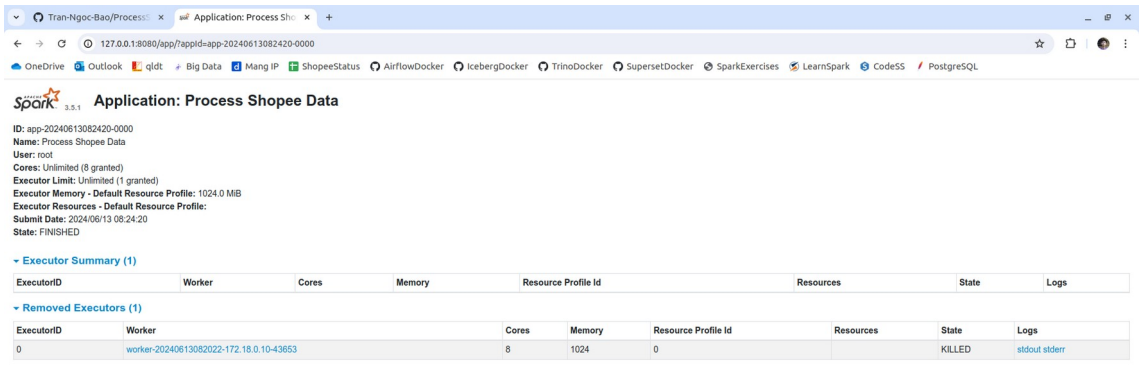
Dữ liệu trong quá khứ cũng sẽ được lưu lại đồng thời với dữ liệu mới và được phân biệt bởi code Pyspark trong hoạt động ở bước sau.



Hình 7: Lưu trữ dữ liệu trên Redis

3.3 Chuyển đổi dữ liệu thành định dạng Iceberg

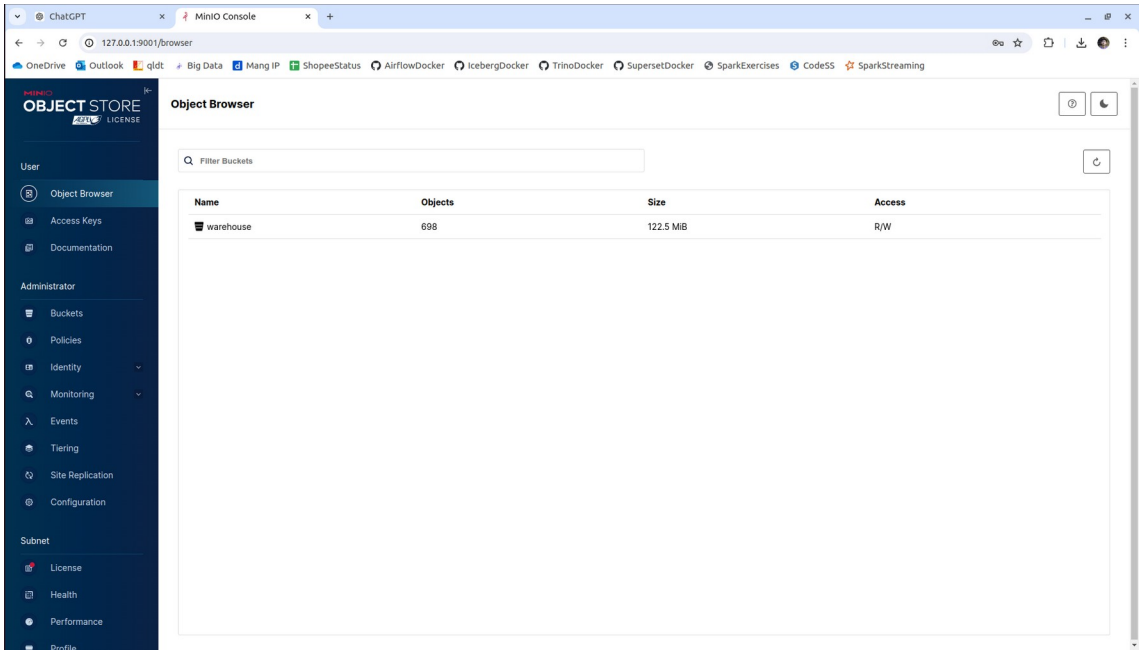
Sử dụng node Spark-Iceberg cùng với node Iceberg-Rest và code Python để chuyển đổi dữ liệu lấy được từ Redis dưới dạng key-value thành định dạng file Iceberg và lưu vào MinIO bằng S3 API nhờ chính node Iceberg-Rest tương tự như Amazon S3. File code Pyspark được submit lên cụm Spark để thực thi.



Hình 8: Chuyển đổi dữ liệu thành định dạng Iceberg

Dữ liệu sẽ được lưu trữ trên MinIO và được tối ưu hóa nhờ định dạng file

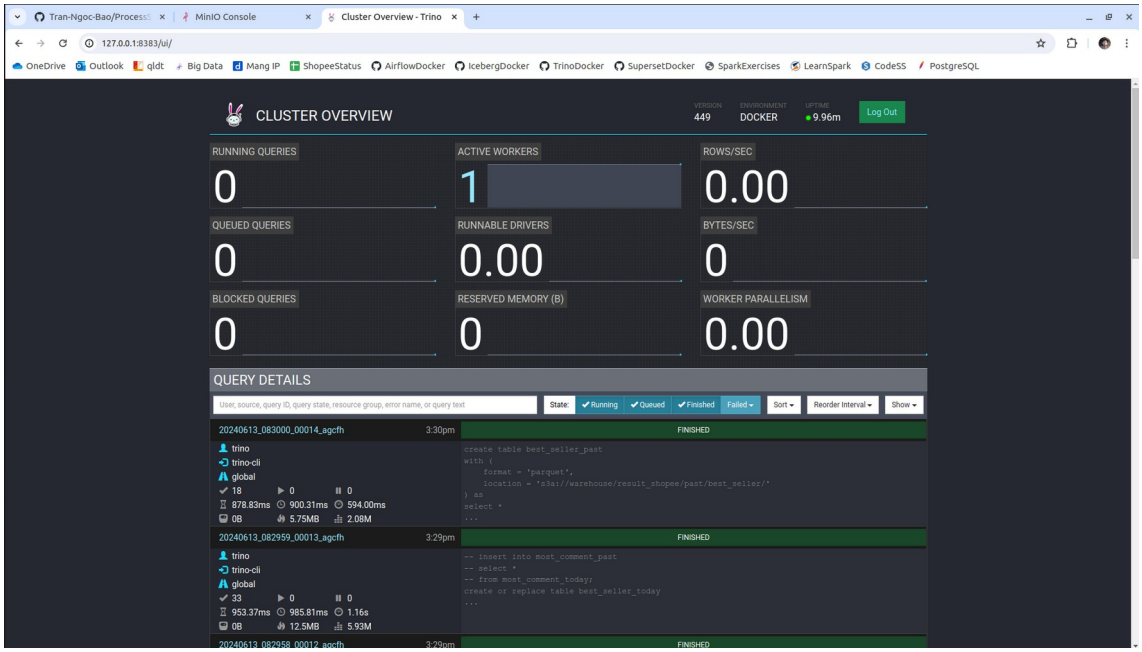
Iceberg khi dữ liệu thô ban đầu thu được từ việc gọi API là 2.2 GiB xuống còn hơn 100 MiB lưu trên MinIO.



Hình 9: Lưu trữ dữ liệu trên MinIO

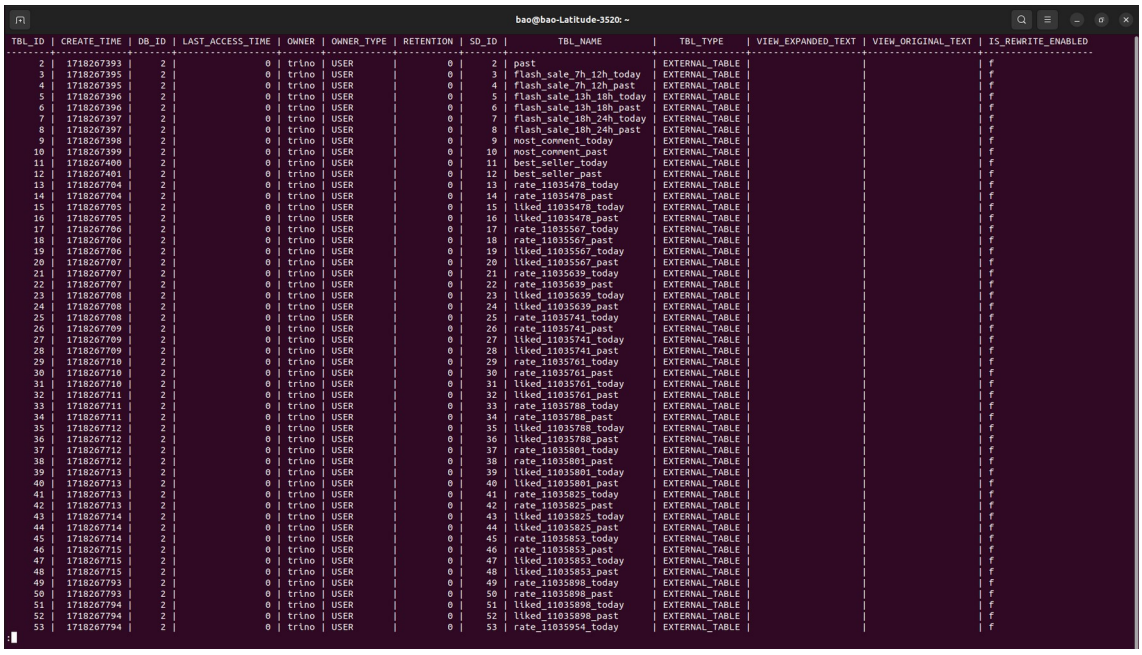
3.4 Truy vấn dữ liệu bằng Trino

Trino truy vấn trực tiếp dữ liệu trên MinIO và lưu kết quả thu được dưới dạng file parquet cũng lên chính MinIO. Trino sử dụng các câu lệnh SQL để truy vấn dữ liệu một cách dễ dàng dựa theo các yêu cầu đề ra.



Hình 10: Truy vấn dữ liệu bằng Trino

Truy cập vào node PostgreSQL để xem được metastore database của Apache Hive khi lưu dữ metastore của hoạt động truy vấn dữ liệu bằng Trino và các metadata của dữ liệu mà Trino sử dụng. Có thể coi Hive và PostgreSQL như một Data Warehouse của hệ thống.



Hình 11: Hoạt động của Hive và PostgreSQL

3.5 Trực quan hóa dữ liệu

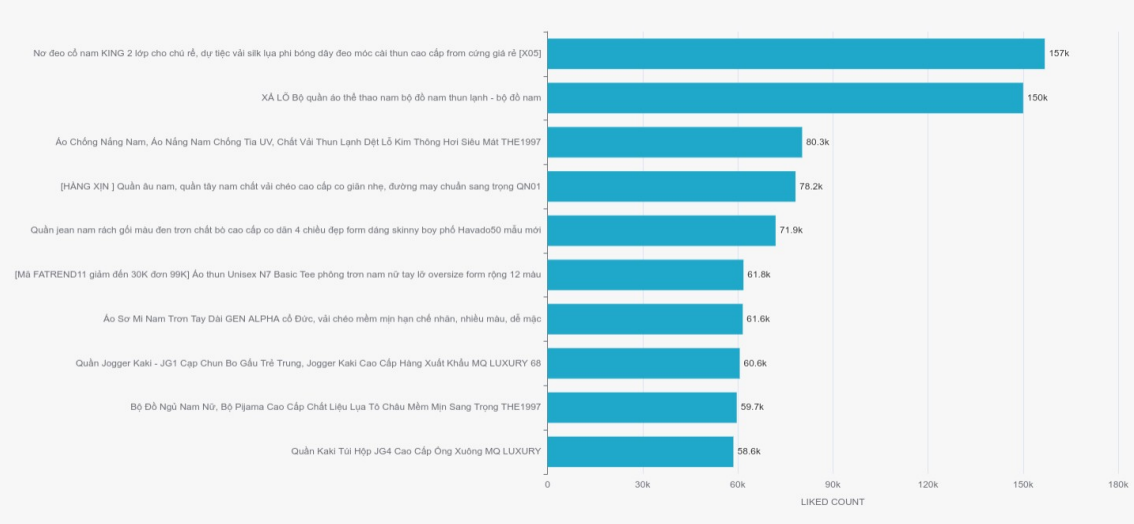
Sử dụng Superset để trực quan hóa dữ liệu theo các yêu cầu đề ra khi nhận đề tài. Thu được các bảng, các biểu đồ từ thông tin dữ liệu sàn thương mại điện từ Shopee. Dưới đây là một số bảng, biểu đồ minh họa:

- Top những sản phẩm được đánh giá tốt theo mặt hàng

name	RATING STAR	RATING COUNT
[Mã FATREND126 giảm đến 50k đơn từ 150k] [DA BÒ THẬT 100%] Ví da nam, Bóp da nam cao cấp, tặng hộp đựng-AL018	4.99	4.58k
(Có Khắc Tên) Bóp Nam Cao Cấp TRẦN THANH HOW Chính Hãng Thiết Kế Nam Tính Chất Liệu Da Nhập Khẩu Chống Cháy Ưu Việt HN02	4.97	3.97k
Túi đeo chéo mini nam nữ nhỏ đi chơi đựng điện thoại đen thời trang chống nước DC01	4.98	3.82k
Balo laptop Arctic Hunter chất liệu Oxford chống nước, có cổng USB - B00120	4.97	2.95k
Balo du lịch Arctic Hunter chất liệu Oxford chống nước, có cổng USB - B00120	4.98	1.6k
Balo laptop Arctic Hunter chất liệu Oxford chống nước, có cổng USB - B00227	4.99	1.54k
Balo học sinh Arctic Hunter chất liệu Oxford chống nước, có cổng USB - B00120	4.99	1.5k
Balo du lịch Arctic Hunter chất liệu Oxford chống thấm nước - B00227	4.99	1.38k
Ví nam đựng thẻ Arctic Hunter thiết kế nhiều ngăn, chống nước, chống xước - Q00013	5	1.24k
Balo laptop Arctic Hunter Arctic Hunter chất liệu Oxford Fabric chống thấm nước - B00530	4.99	1.17k
Balo đựng laptop 15.6 inch GENBAG balo nữ đi học quai đeo có lót đệm trơn basic BL11	4.98	1.14k
ví nam khắc tên theo yêu cầu da safiano thời trang túi ví công sở full box hộp và túi giấy ,thiệp phù hợp làm quà	4.98	1.09k
Balo học sinh Arctic Hunter chất liệu Oxford chống thấm nước - B00227	5	1.08k
ví nam khắc tên theo yêu cầu có in tặng ảnh da saffiano nhập khẩu cosmos Lucaster fullbox hộp và túi bh 12 tháng	4.98	1.08k
Balo laptop Arctic Hunter chất liệu Oxford chống thấm nước, nhiều ngăn tiện dụng - B00477	4.98	1.04k
Balo laptop Arctic Hunter chất liệu chống nước, có cổng USB - B00218	4.98	944
(Mới) Cặp đen học sinh cấp 2 và 3 Kim Long KL035	4.98	857
Balo du lịch Arctic Hunter chất liệu Oxford Fabric chống thấm nước - B00530	5	826
Ví vải CAMELIA BRAND Button Card Wallet	4.97	807
Balo học sinh Arctic Hunter chất liệu Oxford Fabric chống thấm nước - B00530	5	781
Balo học sinh Arctic Hunter chất liệu Oxford chống nước, có cổng USB - B00477	4.99	769
Balo du lịch Arctic Hunter chất liệu Oxford chống nước, có cổng USB - B00477	5	758

Hình 12: Top những sản phẩm được đánh giá tốt theo danh mục Balo & Túi Ví Nam

- Top những sản phẩm được thích theo mặt hàng



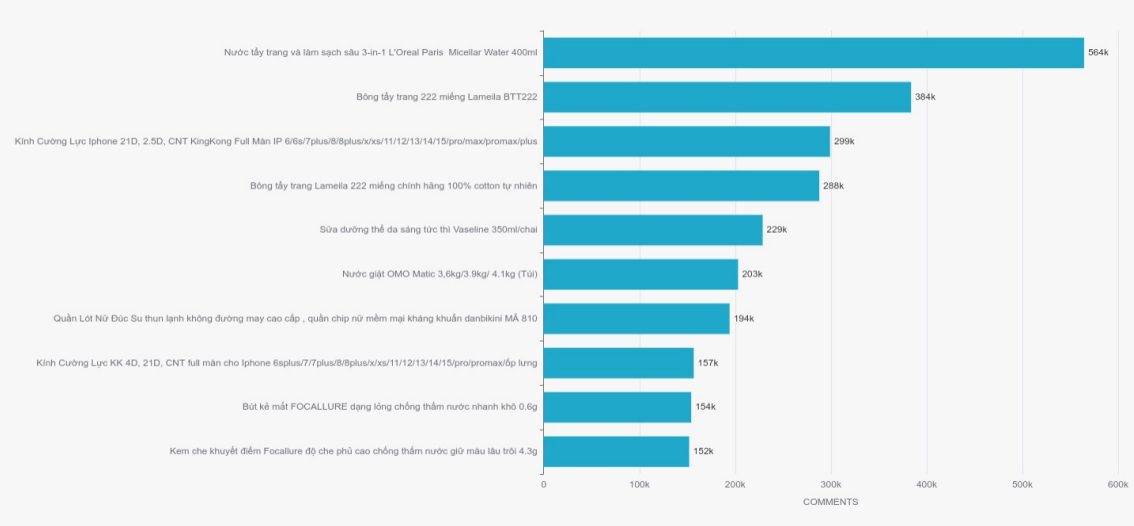
Hình 13: Top những sản phẩm được thích theo danh mục Thời Trang Nam

- Top những sản phẩm đang sale mạnh

name	SHOW DISCOUNT	PRICE
Bao Cao Su có gai CARE 48 Hương Dâu, Kéo Dài Thời Gian, Mỏng Nhiều Gel Bôi Trơn Hộp 12 Bcs gai li ti	97	139k
Gel Bôi Trơn gốc nước CARE LATEX dạng gói 5ml/ gói, Siêu Trơn Lâu Khô An Toàn Cho Cảm Giác Hương Dịu Nhẹ Hỗ Trợ bcs Care	94	47k
Đồng hồ nam nữ Unisex Led kiểu dáng phi hành gia dây cao su êm tay thời trang cá tính BLACK29	90	28k
Đồng Hồ Điện Tử Đa Năng Chống Rơi Họa Tiết Rằn Ri Phong Cách Thể Thao Mới Cho Nam	90	24.9k
Đồng hồ đeo tay nam W3 dây da viền vàng mặt 40mm lịch lãm sang trọng lịch sự	90	35k
Viên Sủi Bổ Sung Vitamin C Tăng Cường Sức Đề Kháng Barocco Dân Khang Tuýp 10 Viên	90	75k
Đồng hồ nam nữ Unisex Led BLACK kiểu dáng phi hành gia dây cao su êm tay thời trang cá tính DH29	90	25k
Sỉ 100 Cái Khẩu Trang 5D 3 Lớp Cao Cấp Dưa Hấu - Khẩu Trang 5D Chống UV	90	40k
Sỉ 100 Chiếc Khẩu Trang 5D Chống Tia UV , Khẩu Trang Y Tế 3 Lớp Kháng Khuẩn Cao Cấp.	90	35k
[Sỉ 100 Chiếc] Khẩu Trang 5D Cao Cấp 4UKorea Có Lớp Meltblown Kháng Khuẩn, Kháng Bụi Mịn	90	32.5k
Đồng hồ nam dây thép không gỉ phong cách cá tính lịch sự DH22	90	42k
[Combo] Bao Cao Su Kimono Siêu Mỏng 0.01 Nhiều Gel Bôi Trơn Kéo Dài Thời Gian, Hộp 10 Cái Bcs	90	65k
Chuyến phát nhanh Bộ Trang Điểm Đầy Đủ 24 Món Cơ Bản Từ A-Z Set Trang Điểm Cá Nhân Bộ Makeup Nhẹ Nhàng Đi Chơi Cho Nàng	90	345k
Thùng 200 chiếc khẩu trang 5d Thịnh Phát 3 lớp kháng khuẩn [SHIP 2H HOẢ TỐC HCM]	90	69k
Combo bao cao su juncai siêu mỏng 0.01 cao cấp nhiều gel bôi trơn kéo dài thời gian gân gai tăng khoái cảm bcs olo-store	90	59k
Bao cao su Juncai size nhỏ 49mm Siêu mỏng 0.01 Spearmint Hương bạc hà Nhiều gel bôi trơn - Hộp 10 bcs durex_olo_store	90	65k
(Che tên) Bao cao su nam juncai siêu mỏng 0.01 Ultrathin tăng khoái cảm, nhiều gel bôi trơn, bcs kéo dài thời gian	90	69k
[Free Ship] - Thùng 100 Chiếc Khẩu Trang 5D Thịnh Phát 3 Lớp Kháng Khuẩn	90	42.5k
[10 Túi 100 Chiếc] Khẩu Trang 5D 3 Lớp Chống Nắng Chống Bụi Chống Tia Uv	90	22k
Đồng hồ nam W10 dây da bản cao cấp thời trang lịch lãm	90	40k
Thùng 100 Chiếc Khẩu Trang 5D Uni Mask Người Lớn	90	40k
Đồng Hồ Nữ Dây Da Đeo Tay Thời Trang Doukou DH31 Mặt Số Chữ Nhật Đơn Giản Tinh Tế	90	35k

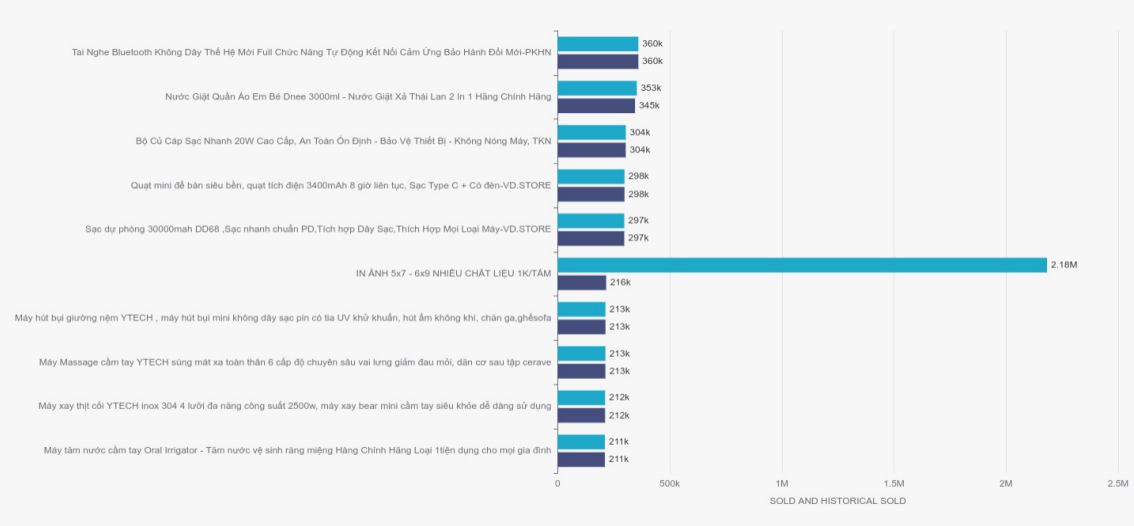
Hình 14: Top những sản phẩm đang sale mạnh từ 13h tới 18h ngày 9 tháng 6

- Top những sản phẩm được comment nhiều nhất



Hình 15: Top những sản phẩm được comment nhiều nhất

- Top những sản phẩm bán chạy nhất



Hình 16: Top những sản phẩm bán chạy nhất

CHƯƠNG 4. KẾT LUẬN

4.1 Kết luận

Như vậy, mini-project đã xây dựng lên một mô hình Data Lake đơn giản hữu ích trong việc phân tích thông tin dữ liệu trên sàn thương mại điện từ Shopee. Có thể giúp các doanh nghiệp trong việc tìm hiểu các mặt hàng được bán chạy nhất, các mặt hàng được yêu thích nhất, được quan tâm nhất và mức độ sale từ các mặt hàng.

Khi giải quyết được vấn đề Open API của Shopee để tăng khả năng tự động hóa của hệ thống, thì hệ thống tương đối khả thi trong thực tế.

Khi thực hiện đề tài em đã được tiếp xúc với rất nhiều công nghệ mới như Airflow, Redis, PostgreSQL, Spark, MinIO, Trino, Hive, Superset cùng với định dạng file Iceberg và các hoạt động xung quanh API.

Đồng thời, khi thực hiện đề tài em cũng tăng được khả năng đọc hiểu tài liệu tiếng anh trên mạng, tham khảo các mã nguồn mở trên Github, Docker Hub và kỹ năng viết báo cáo phục vụ cho việc học tập trên trường và làm việc sau này.

4.2 Hướng phát triển trong tương lai

Sẽ đăng ký với Shopee để sử dụng Open API giúp thu thập một lượng dữ liệu chi tiết và lớn hơn nữa. Qua đó sẽ tối ưu code và cấu hình để tự động hóa toàn bộ hệ thống.

Tích hợp thêm xử lý dữ liệu theo thời gian thực để hỗ trợ việc xử lý dữ liệu sale mạnh của Shopee một cách chính xác hơn.

TÀI LIỆU THAM KHẢO

[1] Trino Software Foudation, https://trino.io

[2] Apache Software Foudation, https://iceberg.apache.org

[3] Apache Software Foudation, https://airflow.apache.org

[4] MinIO, https://min.io

[5] Github, https://github.com/puckel/docker-airflow, 2020

[6] Github, https://github.com/tabular-io/docker-spark-iceberg, 2024

[7] Github, https://github.com/starburstdata/dbt-trino, 2024

[8] Github, https://github.com/dgkatz/trino-hive-superset-docker/tree/main, 2021