TRƯỜNG ĐẠI HỌC KINH TẾ ĐÀ NẴNG

KHOA THƯƠNG MẠI ĐIỆN TỬ

BÁO CÁO ĐỀ ÁN THỰC HÀNH 2

XÂY DỰNG HỆ THỐNG PIPELINE ETL THU THẬP VÀ PHÂN TÍCH DỮ LIỆU SẢN PHẨM APPLE–SAMSUNG TRÊN AMAZON

Giảng viên bộ môn:	ThS.Nguyễn Văn Chức
Nhóm thực hiện:	05
Thành viên:	1. Hoàng Thị Kim Chi
	2. Nguyễn Đình Khoa

Đà Nẵng, tháng 12 năm 2025

MỤC LỤC

Họ Và Tên	Mã Sinh Viên	Công Việc	Phần Trăm Đóng Góp
Hoàng Thị Kim Chi	211124029107	Viết báo cáo , ETL, Crawl dữ liệu, Video	50%
Nguyễn Đình Khoa	221124029221	Viết báo cáo , ETL, Crawl dữ liệu	50%

LỜI MỞ ĐẦU

Trong bối cảnh toàn cầu hóa và bùng nổ công nghệ thông tin, thương mại điện tử (TMĐT) nổi lên như xu hướng tất yếu, góp phần hình thành nền kinh tế số toàn cầu. Tại Việt Nam, với hơn 100 triệu dân và trên 70% tiếp cận internet, TMĐT đang tăng trưởng mạnh mẽ nhờ hạ tầng mạng cải thiện, phủ sóng 4G/5G rộng rãi, cùng các chính sách hỗ trợ từ Chính phủ như "Chương trình quốc gia về phát triển TMĐT" và "Kế hoạch tổng thể đến 2025". Người tiêu dùng ngày càng ưa chuộng mua sắm trực tuyến nhờ tiện lợi, đa dạng sản phẩm và giá cạnh tranh.

TMĐT mở ra cơ hội cho doanh nghiệp nội địa mở rộng thị trường, giảm chi phí trung gian, nâng cao cạnh tranh quốc tế; đồng thời hỗ trợ SMEs chuyển đổi số, tăng hiệu suất. Nó còn thúc đẩy đầu tư vào logistics, giao thông, kho vận và thanh toán số. Việc áp dụng AI, Big Data, Cloud Computing không chỉ tối ưu hóa quy trình mà còn cải thiện trải nghiệm người dùng. Tóm lại, TMĐT là yếu tố thiết yếu để Việt Nam hội nhập và phát triển trong kỷ nguyên số.

I. Giới thiệu và định hướng Data Engineer (DE)

Data Engineer (Kỹ sư Dữ liệu) là chuyên gia đảm nhận vai trò quan trọng trong việc xử lý và quản lý dữ liệu, giúp biến dữ liệu thô thành nguồn thông tin có giá trị. Công việc của họ tập trung vào việc thiết kế và xây dựng các hệ thống (pipeline) để thu thập, chuyển đổi, và lưu trữ dữ liệu từ nhiều nguồn khác nhau, đảm bảo dữ liệu sẵn sàng phục vụ các mục đích phân tích và ứng dụng thực tế.

Data Engineer là cầu nối giữa nguồn dữ liệu phức tạp và các hệ thống phân tích hoặc ra quyết định. Họ không chỉ cần hiểu về cấu trúc và tính chất của dữ liệu mà còn phải thành thạo các công nghệ xử lý hiện đại, như cơ sở dữ liệu, hệ thống lưu trữ, và các công cụ xử lý dữ liệu lớn (Big Data). Vai trò này ngày càng quan trọng trong bối cảnh doanh nghiệp phụ thuộc nhiều hơn vào dữ liệu để tạo lợi thế cạnh tranh.

1. **Thị trường chung của Data Engineer**

Thị trường **DE** tại Việt Nam đang bùng nổ nhờ kinh tế số dự kiến đạt 45 tỷ USD, với nhu cầu tăng 12% so với 2024, dẫn đầu bởi IT, AI và data analytics (CAGR 9.6% đến 2033). Các công ty outsourcing và tech hub tại TP.HCM, Hà Nội tuyển dụng mạnh, tập trung vào Data/AI/ML Engineers. Chi phí thuê DE dao động 20-50 USD/giờ, với cơ hội thăng tiến cao do thiếu hụt nhân lực skilled.

2. **Nhiệm vụ chính của Data Engineer**

DE thường đảm nhận các nhiệm vụ sau:

- Thiết kế và xây dựng data pipelines để thu thập, xử lý và lưu trữ dữ liệu từ nhiều nguồn.
- Đảm bảo chất lượng dữ liệu (data quality) bằng cách giám sát lỗi, latency và tuân thủ quy định (compliance).
- Hợp tác với data scientists, analysts để tối ưu hóa dữ liệu cho mô hình ML.
- Quản lý cơ sở dữ liệu, triển khai công cụ orchestration và bảo mật dữ liệu.
- Tối ưu hóa hiệu suất hệ thống, đặc biệt trong môi trường cloud và real-time processing.

Để làm DE, bạn cần thành thạo:

- **Kỹ năng cốt lõi**: SQL (query và modeling), Python (scripting, automation), kiến thức ETL/ELT, data architecture, problem-solving.
- **Công cụ chính**: Cloud platforms (AWS, Azure, GCP), Big Data (Apache Spark, Hadoop, Kafka), Orchestration (Airflow, Dagster), Containerization (Docker, Kubernetes), Database (PostgreSQL, MongoDB).

3. **Skill Set, Mindset, Toolset của DE**

- **Skill Set**: Ngoài kỹ năng kỹ thuật, cần kiến thức về data governance, version control (Git) và soft skills như giao tiếp để làm việc nhóm. Đến 2025, kỹ năng AI/ML integration (như vector databases) trở nên thiết yếu.
- **Mindset**: Tập trung vào tư duy hệ thống (systems thinking), chú ý chi tiết (attention to detail) để tránh lỗi dữ liệu, và học hỏi liên tục vì công nghệ thay đổi nhanh (ví dụ: từ batch processing sang real-time). Mindset "declarative" giúp code hóa logic kinh doanh rõ ràng, giảm lỗi.
- **Toolset**: Mở rộng từ cơ bản (SQL/Python) sang nâng cao như dbt (data build tool), Snowflake (data warehouse), và công cụ monitoring (Monte Carlo). Không cần biết hết, mà chọn tool phù hợp với dự án để tránh "tool overload".

Tiêu chí	Fresher (0–1 năm)	Junior (1–3 năm)	Senior (5+ năm)
Năng lực chính	Nền tảng SQL/Python, ETL cơ bản, hỗ trợ pipeline đơn giản. Ít kinh nghiệm dự án thực tế.	Xây dựng pipeline độc lập, tối ưu data quality, làm việc với cloud cơ bản. Hiểu data modeling.	Thiết kế architecture phức tạp, lãnh đạo team, tích hợp AI/real-time. Giải quyết vấn đề ở scale lớn.
Kinh nghiệm	Mới ra trường, dự án cá nhân hoặc internship.	Tham gia dự án thực tế, quen toolset cơ bản.	Lãnh đạo dự án lớn, mentor junior, kinh nghiệm multi-cloud.
Mức lương	15–25 triệu VNĐ (11.5–20 triệu ở Đà Nẵng)	25–40 triệu VNĐ (20–30 triệu entry–mid)	50–100+ triệu VNĐ (70–100 triệu ở TP.HCM)
Cơ hội việc làm	Cao ở outsourcing, cần chứng chỉ (Google Data Analytics).	Tăng trưởng nhanh, chuyển sang specialist.	Lãnh đạo/consultant, remote cho công ty nước ngoài.

II. LÝ DO CHỌN ĐỀ TÀI

1. Mục tiêu của Đề tài

Đề tài hướng đến việc xây dựng một hệ thống pipeline ETL trên nền tảng cloud nhằm thu thập, xử lý và phân tích dữ liệu sản phẩm Apple và Samsung trên Amazon. Cụ thể, các mục tiêu chính bao gồm:

- Thiết kế và triển khai pipeline ETL tự động để thu thập dữ liệu sản phẩm từ Amazon
- Chuẩn hóa và xử lý dữ liệu nhằm đảm bảo tính nhất quán, loại bỏ dữ liệu thiếu hoặc sai lệch, và chuẩn bị dữ liệu cho quá trình phân tích.
- Phân tích so sánh giá bán giữa các sản phẩm Apple và Samsung theo từng loại sản phẩm (điện thoại, laptop, đồng hồ, v.v.), từ đó rút ra xu hướng định giá và chiến lược cạnh tranh.

- Phân tích xu hướng đánh giá khách hàng, bao gồm:
  + Mức độ hài lòng thông qua điểm đánh giá trung bình và số lượt đánh giá.
  + Mối liên hệ giữa giá bán và mức độ đánh giá.
  + Sự khác biệt trong phản hồi khách hàng giữa hai thương hiệu.
- Trực quan hóa dữ liệu phân tích thông qua dashboard hoặc biểu đồ tương tác, giúp người dùng dễ dàng theo dõi sự biến động về giá và đánh giá sản phẩm theo thời gian.

2. Lý do chọn Đề tài

Trong bối cảnh chuyển đổi số và bùng nổ dữ liệu, việc thu thập, xử lý và phân tích dữ liệu đang trở thành yếu tố cốt lõi giúp doanh nghiệp hiểu rõ hơn về thị trường, hành vi khách hàng và xu hướng tiêu dùng. Đặc biệt trong lĩnh vực thương mại điện tử, lượng dữ liệu phát sinh hằng ngày là vô cùng lớn và đa dạng, đòi hỏi những giải pháp kỹ thuật hiện đại để có thể khai thác thông tin có giá trị từ khối dữ liệu đó.

Hai thương hiệu Apple và Samsung là những đối thủ hàng đầu trong ngành công nghệ di động và thiết bị thông minh, luôn thu hút sự quan tâm của người tiêu dùng trên toàn cầu. Nền tảng Amazon.com lại là một trong những kênh thương mại điện tử lớn nhất thế giới, nơi tập trung lượng dữ liệu khổng lồ về giá, doanh số, đánh giá và phản hồi của khách hàng. Việc phân tích dữ liệu sản phẩm của Apple và Samsung trên Amazon không chỉ giúp nhận diện xu hướng thị trường, mà còn góp phần so sánh năng lực cạnh tranh, phản ứng của khách hàng và chiến lược định giá của hai thương hiệu.

3. Phương pháp và Phạm vi Nghiên cứu

3.1 Phương pháp nghiên cứu

Phương pháp nghiên cứu của đề tài "Phân tích Dữ liệu Sản phẩm Apple – Samsung trên Amazon" được thực hiện thông qua quy trình xây dựng và triển khai hệ thống Data Pipeline nhằm đảm bảo dữ liệu được thu thập, xử lý, lưu trữ và trực quan hóa một cách tự động và chính xác.

Trước hết, dữ liệu được thu thập trực tiếp từ trang thương mại điện tử Amazon.com bằng ngôn ngữ Python trong môi trường Google Colab. Thông qua việc sử dụng các thư viện BeautifulSoup và Selenium, hệ thống tiến hành cào dữ liệu về các sản phẩm của hai thương hiệu Apple và Samsung, bao gồm thông tin như tên sản phẩm, giá bán, đánh giá, xếp hạng, ngày cập nhật và các thuộc tính mô tả khác. Dữ liệu sau khi thu thập được lưu ở định dạng CSV, đảm bảo dễ dàng tích hợp vào quy trình xử lý tiếp theo.

Sau giai đoạn thu thập, dữ liệu được đưa vào hệ thống Apache NiFi để thực hiện các bước ETL (Extract – Transform – Load). Tại đây, NiFi đóng vai trò tự động hóa quy

trình xử lý, bao gồm việc làm sạch dữ liệu, loại bỏ các giá trị trùng lặp hoặc thiếu, chuyển đổi kiểu dữ liệu, và chuẩn hóa cấu trúc theo định dạng phù hợp để lưu trữ. NiFi cũng hỗ trợ thiết lập lịch trình xử lý định kỳ, giúp hệ thống vận hành liên tục và cập nhật dữ liệu mới từ Amazon một cách linh hoạt.

Sau khi dữ liệu được xử lý, kết quả đầu ra được tải lên Google BigQuery, đóng vai trò là kho dữ liệu trung tâm (Data Warehouse). BigQuery cung cấp khả năng truy vấn dữ liệu nhanh, hiệu quả và có thể mở rộng, tạo nền tảng vững chắc cho các bước phân tích sâu hơn. Tại đây, dữ liệu được tổ chức thành các bảng và schema hợp lý, hỗ trợ cho các truy vấn thống kê và phân tích xu hướng thị trường giữa hai thương hiệu.

Cuối cùng, dữ liệu trong BigQuery được kết nối trực tiếp với Looker Studio để thực hiện trực quan hóa. Các dashboard được thiết kế hiển thị các chỉ số quan trọng như xu hướng giá, độ phổ biến, mức độ đánh giá của người dùng, cũng như sự khác biệt về hành vi mua sắm giữa sản phẩm của Apple và Samsung. Việc trực quan hóa này giúp người xem dễ dàng nhận diện các xu hướng nổi bật và hỗ trợ quá trình ra quyết định chiến lược dựa trên dữ liệu.

Phương pháp nghiên cứu trên không chỉ đảm bảo tính logic và khoa học trong quy trình xử lý dữ liệu, mà còn thể hiện được khả năng ứng dụng tổng hợp của nhiều công nghệ hiện đại — từ thu thập dữ liệu bằng Python, tự động hóa pipeline bằng NiFi, quản lý dữ liệu lớn với BigQuery, cho đến trực quan hóa và phân tích bằng Looker Studio — góp phần xây dựng một hệ thống dữ liệu toàn diện phục vụ cho việc phân tích xu hướng thị trường thương mại điện tử Apple – Samsung trên Amazon.

3.2 Phạm vi nghiên cứu

Đề tài "Phân tích Dữ liệu Sản phẩm Apple – Samsung trên nền tảng Amazon nhằm đánh giá xu hướng thị trường" tập trung vào việc khai thác, xử lý và phân tích dữ liệu liên quan đến hai thương hiệu công nghệ hàng đầu là Apple và Samsung trên trang thương mại điện tử Amazon

Phạm vi nghiên cứu bao gồm:

- **Nguồn dữ liệu:** các sản phẩm thuộc danh mục điện thoại, laptop, máy tính bảng, phụ kiện của hai thương hiệu Apple và Samsung.
- **Thông tin dữ liệu khai thác:** giá bán, xếp hạng, số lượng đánh giá, mức độ phổ biến, ngày phát hành, và các thuộc tính sản phẩm khác có ảnh hưởng đến **quyết định mua hàng** của người tiêu dùng.
- **Thời gian dữ liệu:** giới hạn trong giai đoạn gần đây (tùy theo thời điểm thu thập, ví dụ 2024–2025), đảm bảo phản ánh đúng xu hướng hiện tại của thị trường.

- **Phạm vi kỹ thuật:** tập trung vào việc xây dựng quy trình xử lý dữ liệu (ETL Pipeline) để khai thác, làm sạch, lưu trữ và phân tích dữ liệu Amazon.

Đề tài không mở rộng sang các thương hiệu khác hoặc các nền tảng thương mại điện tử ngoài Amazon, nhằm đảm bảo độ sâu và độ chính xác khi đánh giá sự cạnh tranh giữa hai thương hiệu này.
Kết quả nghiên cứu kỳ vọng mang lại góc nhìn so sánh rõ ràng giữa Apple và Samsung về giá, độ phổ biến, mức độ hài lòng của khách hàng, từ đó giúp doanh nghiệp và người nghiên cứu hiểu rõ hơn động lực thị trường ngành hàng điện tử cao cấp.

III. TỔNG QUAN CHUNG

1. Tổng quan về Thương mại điện tử

Thương mại điện tử (E-Commerce) là hình thức mua bán hàng hóa và dịch vụ thông qua các nền tảng trực tuyến, cho phép người tiêu dùng và doanh nghiệp giao dịch mà không bị giới hạn bởi không gian hay thời gian. Trong kỷ nguyên số, thương mại điện tử đã trở thành một phần không thể thiếu của nền kinh tế toàn cầu, đóng vai trò quan trọng trong việc thúc đẩy tiêu dùng, tối ưu chuỗi cung ứng và tạo ra những mô hình kinh doanh linh hoạt hơn.

Sự phát triển mạnh mẽ của Internet, cùng với hạ tầng thanh toán điện tử và logistics, đã giúp các nền tảng thương mại điện tử như Amazon, eBay, Alibaba, Shopee, Lazada, Tiki... bùng nổ và chiếm lĩnh thị trường. Theo thống kê, doanh thu từ thương mại điện tử toàn cầu đạt hàng nghìn tỷ USD mỗi năm, với tốc độ tăng trưởng trung bình từ 15–20%/năm.

Thương mại điện tử không chỉ thay đổi cách thức mua sắm, mà còn tạo ra một kho dữ liệu khổng lồ, phản ánh hành vi, sở thích và xu hướng tiêu dùng của khách hàng. Việc khai thác, xử lý và phân tích dữ liệu này trở thành yếu tố cốt lõi giúp các doanh nghiệp đưa ra quyết định chính xác hơn, tối ưu chiến lược marketing, giá bán, và phát triển sản phẩm.

2. Thương mại điện tử Amazon

Amazon.com là nền tảng thương mại điện tử lớn nhất thế giới, được thành lập vào năm 1994 bởi Jeff Bezos tại Mỹ. Ban đầu, Amazon chỉ tập trung vào việc bán sách trực tuyến, nhưng hiện nay đã mở rộng sang hàng triệu mặt hàng trong nhiều lĩnh vực như điện tử, thời trang, đồ gia dụng, công nghệ, và đặc biệt là thiết bị di động.

Amazon không chỉ là một nền tảng bán lẻ, mà còn là hệ sinh thái dữ liệu khổng lồ, nơi người tiêu dùng để lại hàng tỷ lượt đánh giá, xếp hạng và phản hồi mỗi năm. Điều này

giúp Amazon không ngừng cải thiện trải nghiệm người dùng thông qua phân tích dữ liệu hành vi và thuật toán gợi ý sản phẩm (Recommendation System).

Ngoài ra, Amazon còn phát triển nền tảng Amazon Web Services (AWS) – hệ thống cung cấp các dịch vụ điện toán đám mây hàng đầu thế giới, được nhiều doanh nghiệp sử dụng để xây dựng, lưu trữ và xử lý dữ liệu lớn. Với lượng dữ liệu phong phú và đáng tin cậy, Amazon trở thành nguồn dữ liệu lý tưởng cho các nghiên cứu về xu hướng tiêu dùng, giá bán và sự cạnh tranh giữa các thương hiệu toàn cầu.

Trong phạm vi đề tài này, Amazon được chọn làm nguồn dữ liệu chính nhằm khai thác thông tin về sản phẩm của Apple và Samsung, giúp phân tích xu hướng thị trường công nghệ tiêu dùng một cách khách quan và toàn diện.

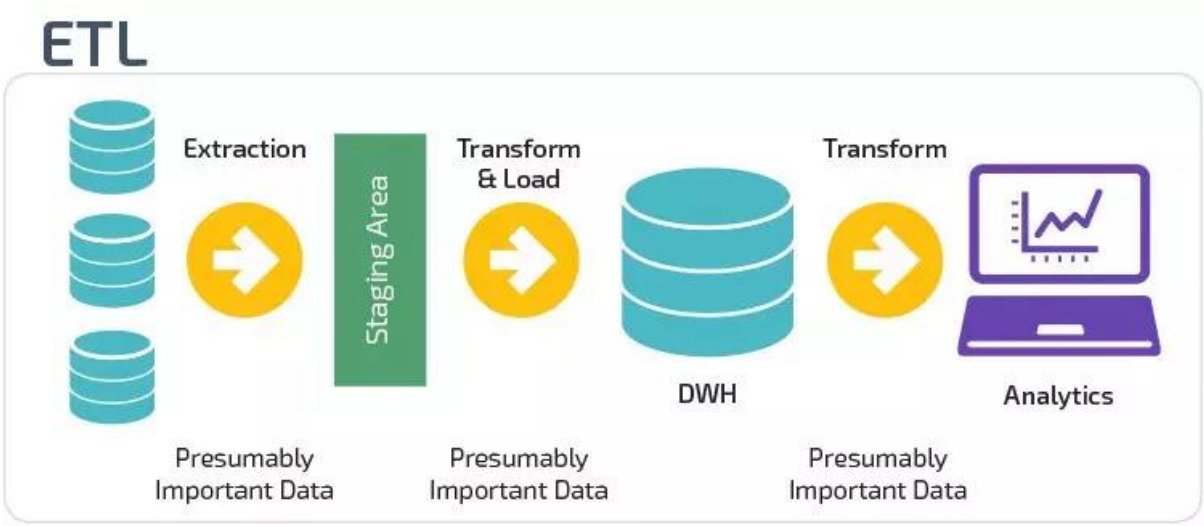
3. Giới thiệu về thị trường Apple & Samsung

Apple và Samsung là hai thương hiệu hàng đầu thế giới trong lĩnh vực điện tử và thiết bị di động thông minh, luôn duy trì vị thế cạnh tranh trực tiếp trong suốt hơn một thập kỷ qua. Cả hai hãng đều có hệ sinh thái sản phẩm đa dạng, bao gồm điện thoại thông minh, máy tính bảng, laptop, đồng hồ thông minh và phụ kiện công nghệ.

Trên nền tảng Amazon, sản phẩm của Apple và Samsung luôn nằm trong top doanh số cao nhất, phản ánh mức độ quan tâm lớn từ người tiêu dùng. Apple nổi bật với chiến lược định vị cao cấp, tập trung vào thiết kế sang trọng, hiệu năng ổn định và hệ sinh thái iOS khép kín. Trong khi đó, Samsung lại hướng đến đa dạng phân khúc, từ cao cấp đến tầm trung, kết hợp công nghệ tiên tiến và hệ điều hành mở Android, giúp mở rộng phạm vi tiếp cận khách hàng.

Sự cạnh tranh giữa hai thương hiệu này thể hiện rõ qua chính sách giá, chiến lược marketing, số lượng đánh giá sản phẩm và mức độ hài lòng của khách hàng. Việc phân tích dữ liệu của Apple và Samsung trên Amazon giúp làm rõ xu hướng lựa chọn sản phẩm, sự biến động giá, và đánh giá của người tiêu dùng theo thời gian, qua đó đưa ra nhận định chính xác về vị thế của từng thương hiệu trên thị trường thương mại điện tử toàn cầu.

4. Giới thiệu về hệ thống ETL



Hệ thống **ETL** (Extract – Transform – Load) là quy trình cốt lõi trong kỹ thuật xử lý và quản lý dữ liệu của một Data Engineer. ETL giúp chuyển đổi dữ liệu từ nhiều nguồn khác nhau thành dạng thống nhất, có thể sử dụng cho phân tích hoặc lưu trữ trong kho dữ liệu (Data Warehouse).

- **Extract (Trích xuất):** là quá trình thu thập dữ liệu từ các nguồn như website, API, cơ sở dữ liệu hoặc file CSV. Trong đề tài này, dữ liệu được trích xuất từ **Amazon.com** thông qua Python.

- **Transform (Chuyển đổi):** là bước làm sạch, xử lý và chuẩn hóa dữ liệu – loại bỏ dữ liệu trùng lặp, xử lý giá trị thiếu, định dạng lại kiểu dữ liệu. Bước này được thực hiện bằng Apache NiFi nhằm tự động hóa quá trình chuyển đổi.

- **Load (Tải):** là quá trình lưu trữ dữ liệu đã xử lý vào hệ thống kho dữ liệu – ở đây là Google BigQuery – để phục vụ cho việc phân tích và trực quan hóa bằng Looker Studio.

ETL đóng vai trò trung tâm trong việc đảm bảo chất lượng và tính toàn vẹn của dữ liệu, giúp hệ thống dữ liệu vận hành hiệu quả, tự động và có khả năng mở rộng. Việc ứng dụng ETL trong đề tài này giúp tối ưu hóa toàn bộ quá trình xử lý dữ liệu Amazon, từ khâu thu thập đến khâu phân tích xu hướng.

4.1. Apache NiFi

Apache NiFi là một công cụ mạnh mẽ cho phép quản lý luồng dữ liệu một cách tự động và linh hoạt. Được thiết kế để đáp ứng nhu cầu xử lý luồng dữ liệu phức tạp, Apache NiFi hỗ trợ trích xuất dữ liệu từ nhiều nguồn, xử lý dữ liệu với giao diện trực quan, và đảm bảo luồng dữ liệu liên tục và an toàn.

- Khả năng tích hợp đa dạng: NiFi hỗ trợ hàng loạt các nguồn và đích dữ liệu, từ các file, cơ sở dữ liệu, đến API.
- Dễ dàng cấu hình và giám sát: Người dùng có thể thiết lập các quy trình xử lý dữ liệu chỉ bằng thao tác kéo và thả, giúp giảm thời gian triển khai.
- Bảo mật và kiểm soát luồng dữ liệu: NiFi cung cấp khả năng kiểm soát và giám sát luồng dữ liệu trong toàn bộ quy trình.

4.2. Google BigQuery

Google BigQuery là dịch vụ kho dữ liệu đám mây của Google, cho phép lưu trữ và truy vấn dữ liệu lớn một cách nhanh chóng và hiệu quả. Với khả năng hỗ trợ các truy vấn
SQL, BigQuery giúp nhà phân tích dễ dàng thao tác và phân tích dữ liệu mà không cần lo

lắng về hạ tầng vật lý.

- Khả năng mở rộng linh hoạt: BigQuery có thể mở rộng để xử lý hàng terabyte dữ liệu mà không ảnh hưởng đến hiệu suất.
- Hỗ trợ phân vùng và clustering: Giúp tối ưu hóa quá trình truy vấn dữ liệu, giảm thiểu thời gian phản hồi.
- Tích hợp dễ dàng với các công cụ phân tích: BigQuery có thể kết nối trực tiếp với Google Data Studio để tạo ra các báo cáo trực quan.

4.3. Looker Studio

Looker Studio là một công cụ trực quan hóa dữ liệu mạnh mẽ, hỗ trợ người dùng tạo các báo cáo và biểu đồ trực quan để phân tích và theo dõi chỉ số kinh doanh. Looker Studio giúp biến dữ liệu từ BigQuery thành các biểu đồ, bảng và báo cáo dễ hiểu, hỗ trợ đưa ra quyết định dựa trên dữ liệu.

- Tạo báo cáo động: Cho phép người dùng tùy chỉnh báo cáo theo thời gian thực, thêm các biểu đồ và bảng dữ liệu theo nhu cầu.
- Tích hợp với Google BigQuery: Trực tiếp truy xuất dữ liệu từ BigQuery và các dịch vụ khác của Google.
- Dễ sử dụng và chia sẻ: Các báo cáo có thể dễ dàng chia sẻ với các thành viên khác, giúp mọi người nắm bắt tình hình kinh doanh.

IV. FRAMEWORK

1. Tổng quan Dự án

Dữ liệu được thu thập từ trang web chính thống của cửa hàng Amazon lưu trữ vào các thư mục với các định dạng Excel sau đó dữ liệu được tự động đọc vào Apache Nifi để tiến hành tiền xử lý dữ liệu (làm sạch , chuyển đổi dữ liệu , …) .
Sau đó dữ liệu được tổng hợp thành bảng. Ở bước Load , dữ liệu sẽ được load định kỳ vào Google Bigquery theo bảng được tạo trước . Sử dụng các lệnh SQL đơn giản trên Google Bigquery để trích xuất dữ liệu, dữ liệu có thể được chuyển trực tiếp qua Looker Studio để vẽ các chart trực quan hóa dữ liệu.

2. Mô tả Chi tiết Các Giai đoạn

2.1. Extract (Trích xuất)

Dữ liệu crawl từ trang thương mại điện tử chính thống amazon.com

Nguồn dữ liệu : File Excel
Công cụ: Apache Nifi

2.2. Transform (Chuyển đổi)

Mục tiêu: Chuẩn hóa dữ liệu, loại bỏ dữ liệu không hợp lệ và chuyển đổi dữ liệu sang định dạng phù hợp để phân tích.
Các hoạt động:

- Chuyển đổi dữ liệu sang Json: Đưa tất cả dữ liệu về một định dạng chung để dễ dàng xử lý.
- Loại bỏ dữ liệu null Đảm bảo tính toàn vẹn của dữ liệu.
- Loại bỏ dữ liệu trùng lặp: Tránh tình trạng dữ liệu bị ghi đè lên nhau.
- Chuẩn hóa định dạng giá: Loại bỏ ký tự
- Lấp đầy dữ liệu còn thiếu

Công cụ: Apache NiFi

2.3. Load (Tải)

Mục tiêu: Tải dữ liệu đã được làm sạch vào kho dữ liệu (data warehouse).
Công cụ: Apache Nifi , Google BigQuery

2.4. Visualization (Trực quan hóa)

Mục tiêu: Tạo các biểu đồ, báo cáo để giúp người dùng dễ dàng hiểu và phân tích dữ liệu.
Sử dụng câu lệnh SQL truy vấn dữ liệu cần thiết cho việc vẽ dashboard và dẩy dữ liệu sang Looker Studio
Công cụ: Looker studio

V. CHI TIẾT DỰ ÁN

1. Bộ Dữ liệu

- Crawl trực tiếp từ trang web Amazon
- Thu thập dữ liệu sản phẩm theo từng danh mục cụ thể: điện thoại, laptop, tai nghe, đồng hồ thông minh, máy tính bảng
- Trích xuất được các thông tin cần thiết : **2.473 dòng** và **10 cột**

Tên trường	Mô tả chi
brand	Thương hiệu sản phẩm (Apple, Samsung)
product_id	Mã định danh duy nhất của sản phẩm trên Amazon
title	Tên đầy đủ của sản phẩm
type_product	Phân loại sản phẩm
price	Giá bán hiện tại hiển thị trên Amazon
actual_price	Giá gốc ban đầu của sản phẩm
discount_percentage	Phần trăm giảm giá (nếu sản phẩm đang được khuyến mãi)
rating	Điểm đánh giá trung bình của sản phẩm (từ 1 đến 5 sao)
reviews	Số lượng đánh giá của khách hàng
availability	Tình trạng sản phẩm (còn hàng, hết hàng, hoặc tạm thời không có hàng)
about_product	Mô tả chi tiết về sản phẩm (thông số kỹ thuật, tính năng nổi bật, v.v)

2. Hệ thống Pipeline ETL

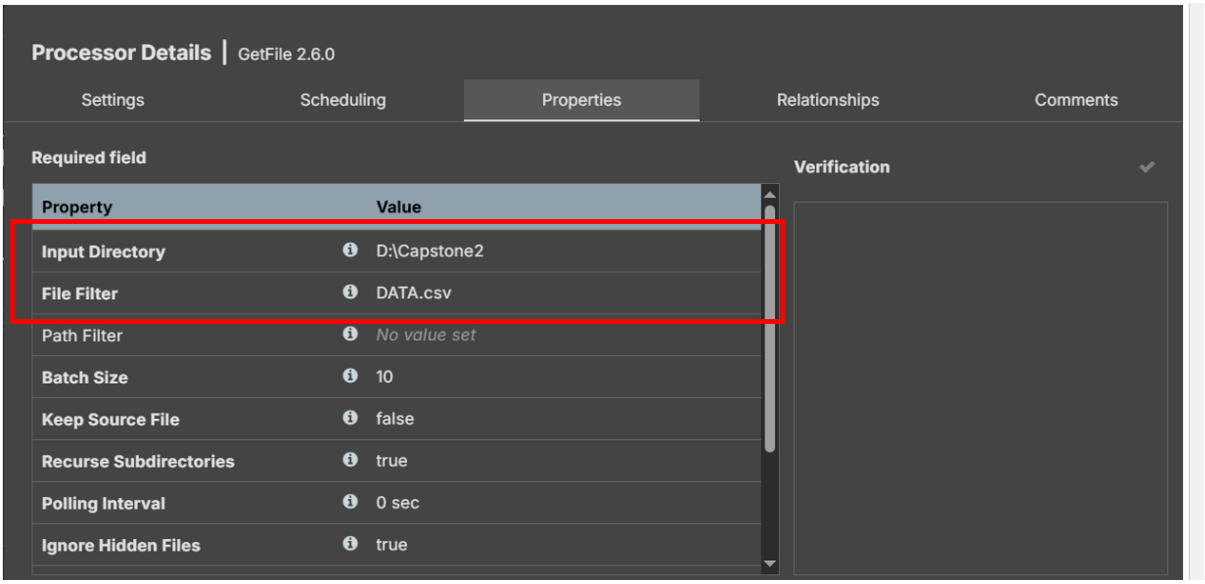
2.1. Extract (Trích xuất)

Nguồn dữ liệu: Dữ liệu sản phẩm được lưu vào google drive sau khi được thu thập (crawl) gồm 2 bộ dữ liệu APPLE và SAMSUNG. Sau đó sẽ gộp thành 1 file có tên là DATA.

APPLE

SAMSUNG

Đọc dữ liệu vào Apache Nifi



Cung cấp đường dẫn tới thư mục cũng như tên file cho processor

➢ Mục **File Filter** có thể điền [^\.].* nếu muốn lấy bất kì file nào xuất hiện trong folder thay vì 1 tên file cụ thể)

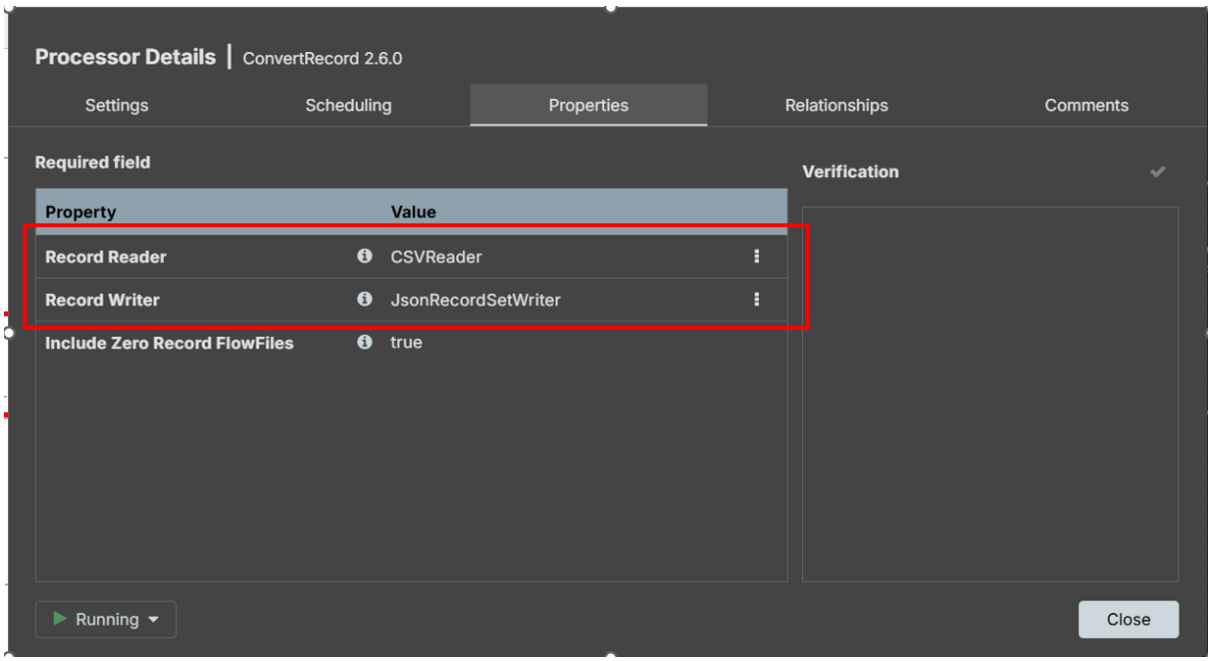
Processor sẽ liên tục kiểm tra sự xuất hiện của file dữ liệu và đọc dữ liệu vào.

2.2. Transform (Chuyển đổi)

Trong giai đoạn này, dữ liệu thô được làm sạch và chuẩn bị để nạp vào Data Warehouse.

a. **Chuyển đổi dữ liệu sang Json:** Đưa tất cả dữ liệu về một định dạng chung để dễ dàng xử lý.

Sử dụng **ConvertRecord** để chuyển đổi file bất kì thành JSON.



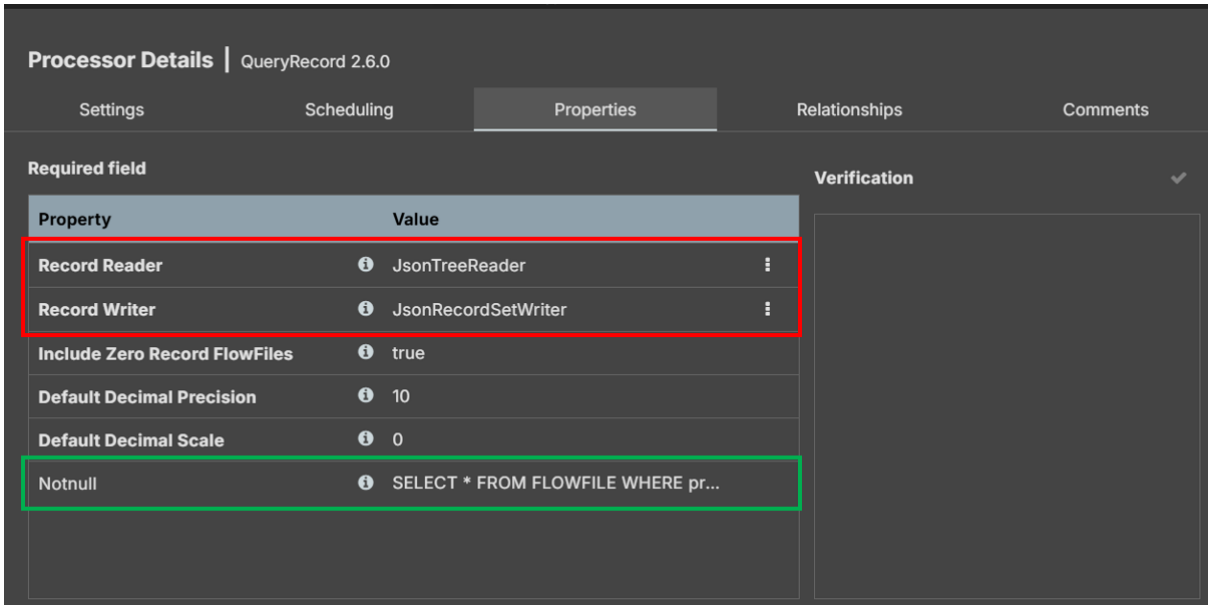
- **Controller Service cần có:** CSVReader và JsonRecordSetWriter

Thực hiện chuyển đổi kiểu dữ liệu từ CSV sang JSON

- ***Record Reader*** : Kiểu dữ liệu đầu vào
- ***Record writer*** : Kiểu dữ liệu đầu ra

b. Loại bỏ dữ liệu null đảm bảo tính toàn vẹn của dữ liệu.

Đối với những dòng dữ liệu bị trống **product_id** tức sản phẩm lỗi → Bỏ qua những
dòng này.

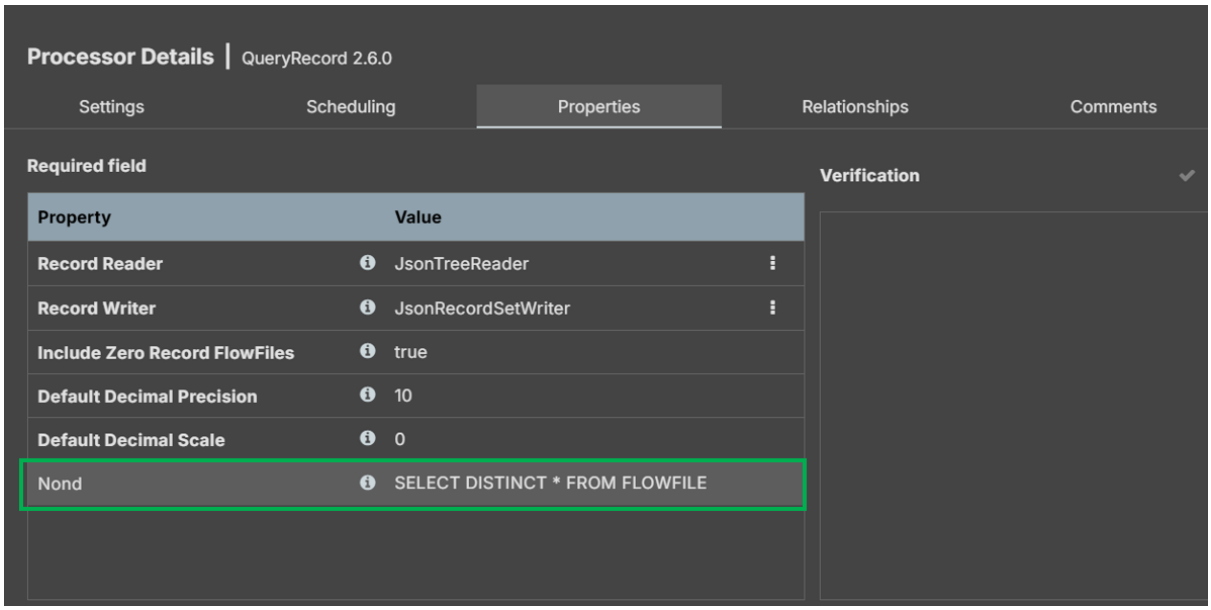


Chọn kiểu dữ liệu đầu vào và ra phù hợp , Tạo 1 ô chứa biểu thức đặt tên phù hợp và sử dụng biểu thức SQL sau : « **Select * from flowfile where product_id is not null** »
Truy vấn này sẽ tiến hành lấy những dòng dữ liệu có đầy đủ cột **product_id**.

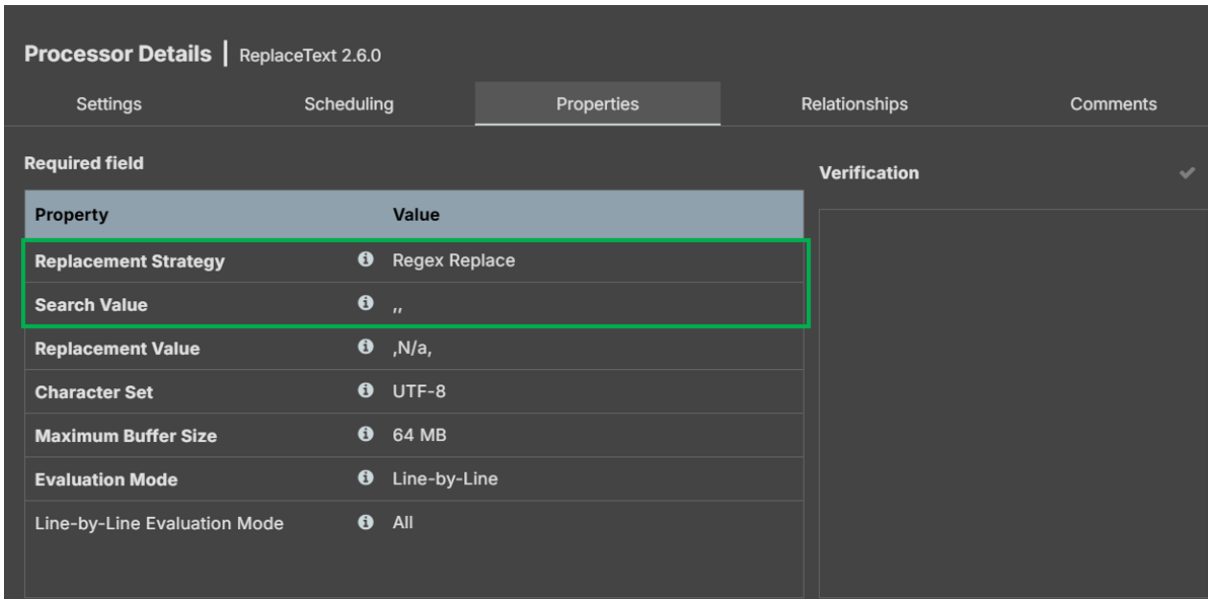
c. Loại bỏ dữ liệu trùng lặp

Đối với dữ liệu trùng lặp, nhóm sử dụng **Processor QuerryRecord** để truy vấn những dòng dữ liệu duy nhất

Câu lệnh truy vấn : **SELECT DISTINCT * FROM FLOWFILE**



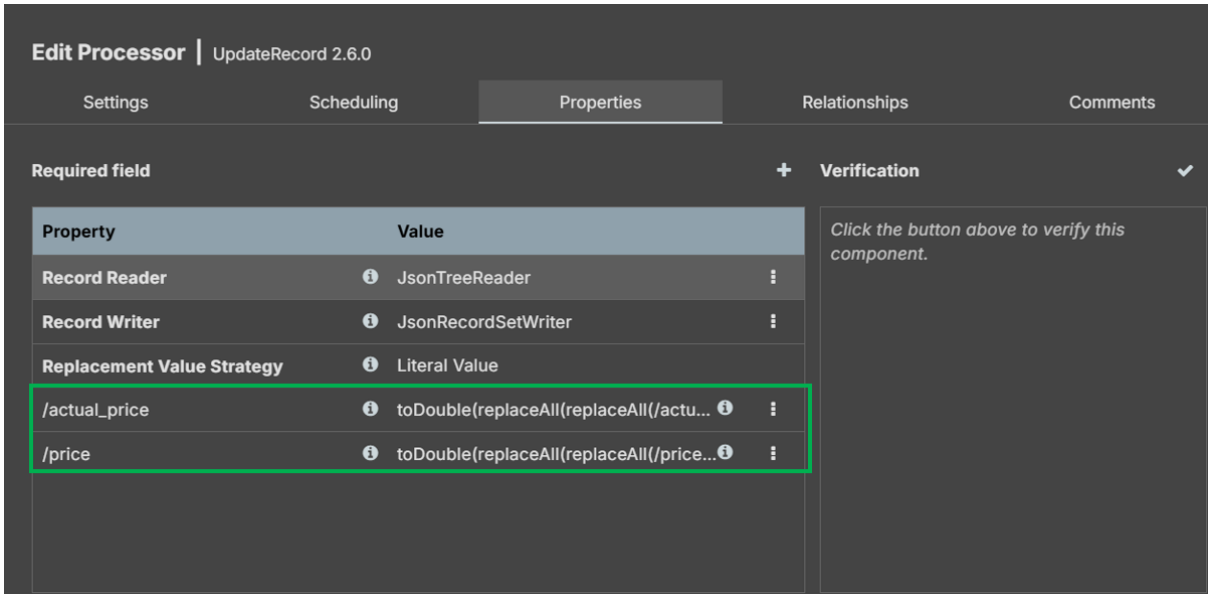
d. Lấp đầy dữ liệu còn thiếu



Sử dụng processor **ReplaceText** để lấp đầy những ô dữ liệu bị thiếu tránh trường hợp lỗi trong quá trình đẩy lên Google Bigquery

e. Chuẩn hóa định dạng giá

Dùng Processor **UpdateRecord** để chuẩn hóa hai trường **price** và **actual_price**.



Mục đích của chuẩn hóa là **loại bỏ ký tự tiền tệ, dấu phẩy và chuyển giá về dạng số (double)** để BigQuery có thể lưu trữ và thực hiện các phép tính như trung bình, min–max, so sánh giá.

Câu lệnh truy vấn:

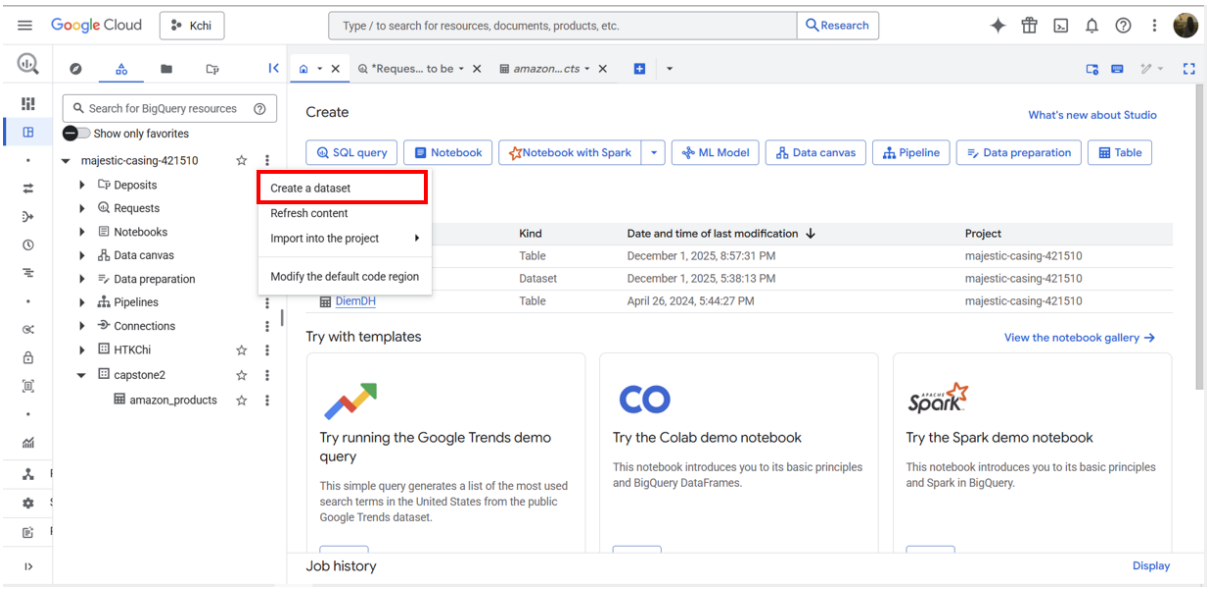
- **toDouble(replaceAll(replaceAll(/actual_price, '\\$', ''), '\\.\\.', '.'))**
- **toDouble(replaceAll(replaceAll(/price, '\\$', ''), '\\.\\.', '.'))**

2.3. Load (Tải)

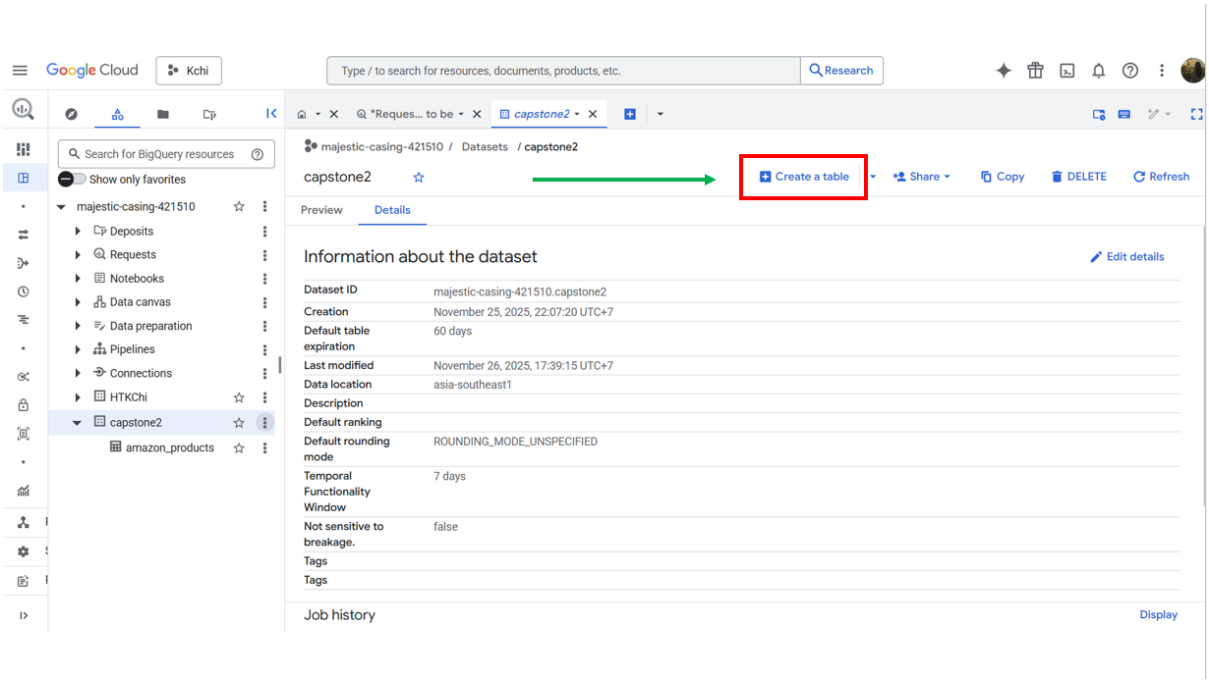
Sau khi dữ liệu được làm sạch sẽ được tổng hợp và load lên Google Bigquery, để làm điều này phải trải qua các bước sau:

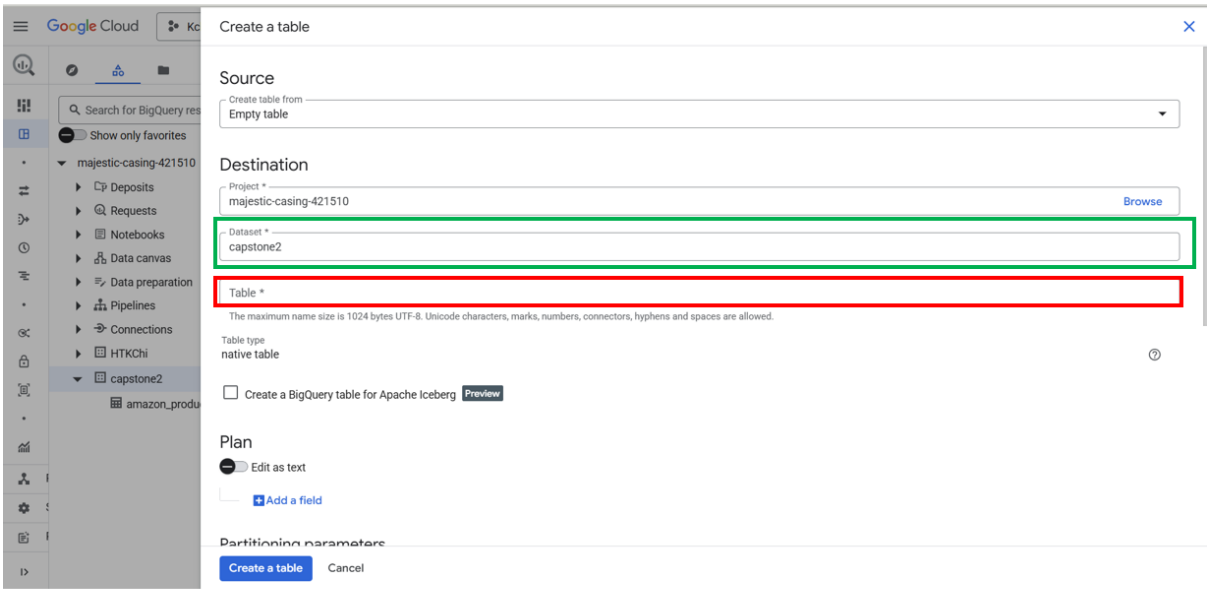
2.3.1. Tạo bảng dữ liệu

Tạo Dataset cho dự án



Sau khi tạo Dataset, tiến hành tạo Table để lưu trữ dữ liệu

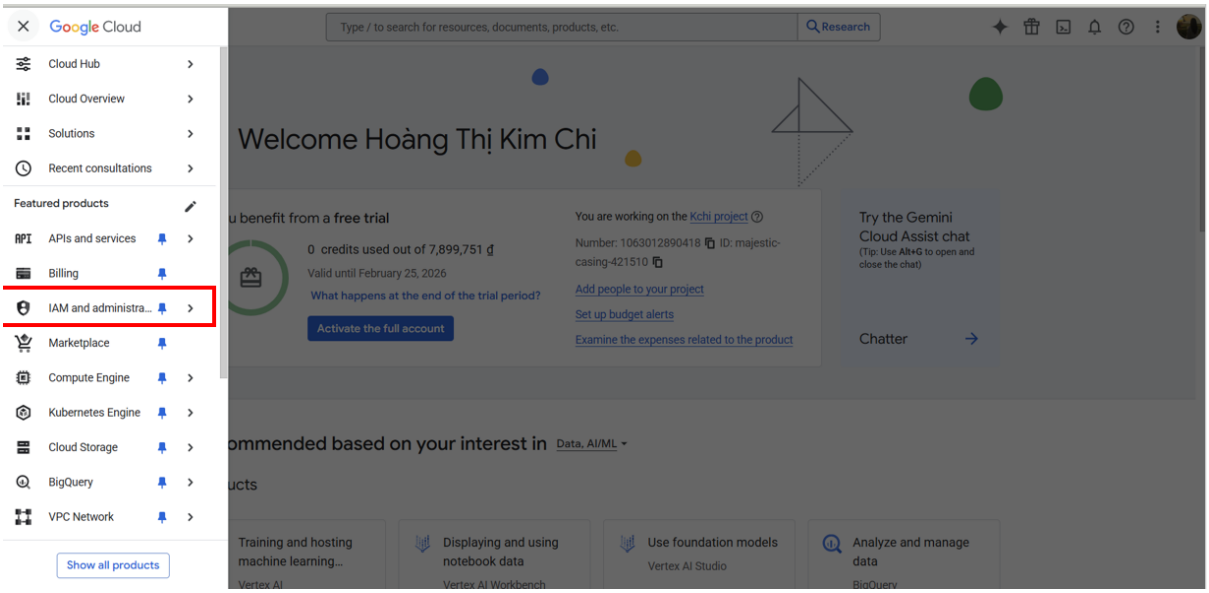




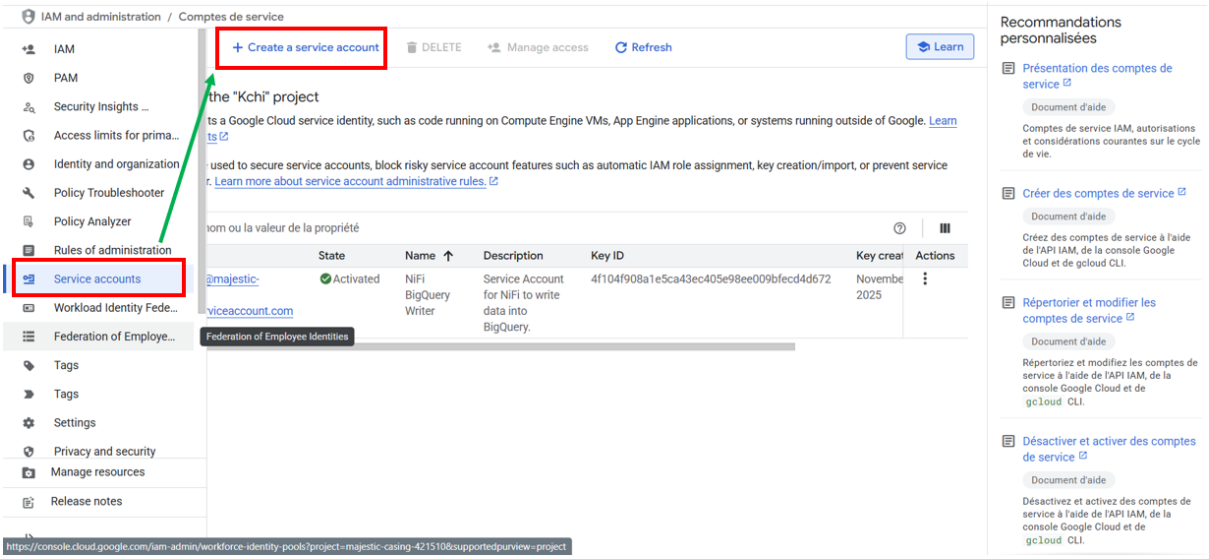
Chọn Dataset đã tạo → Đặt tên cho table

2.3.2. Chọn phương thức xác thực

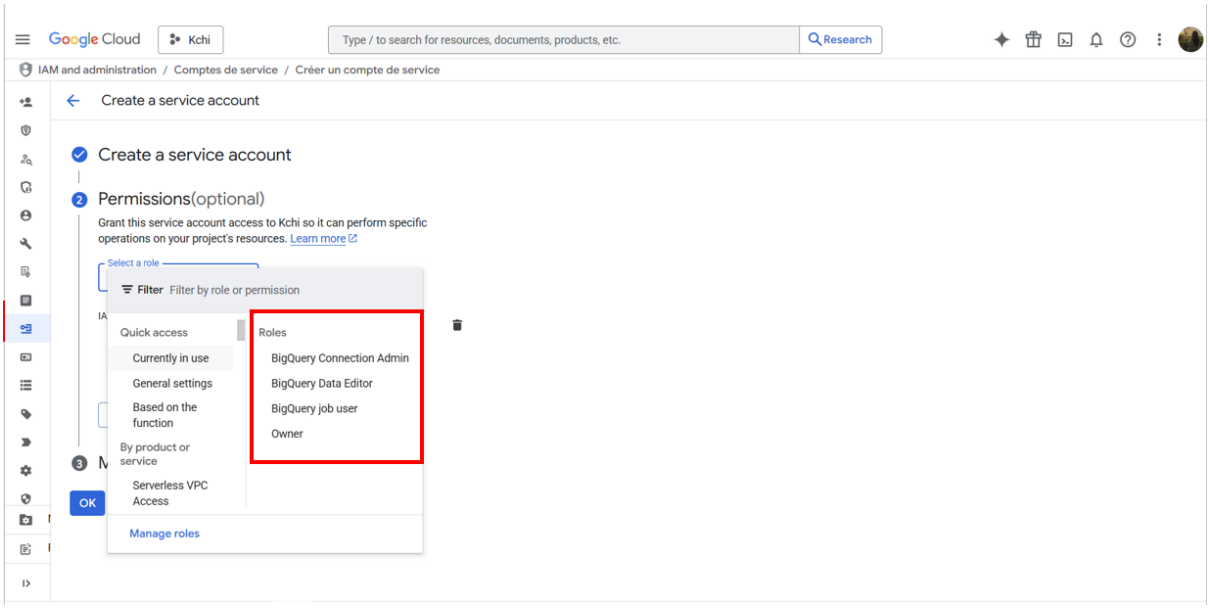
Trên trang console Google cloud, nhấn biểu tượng 3 dấu gạch ngang → chọn IAM&Admin



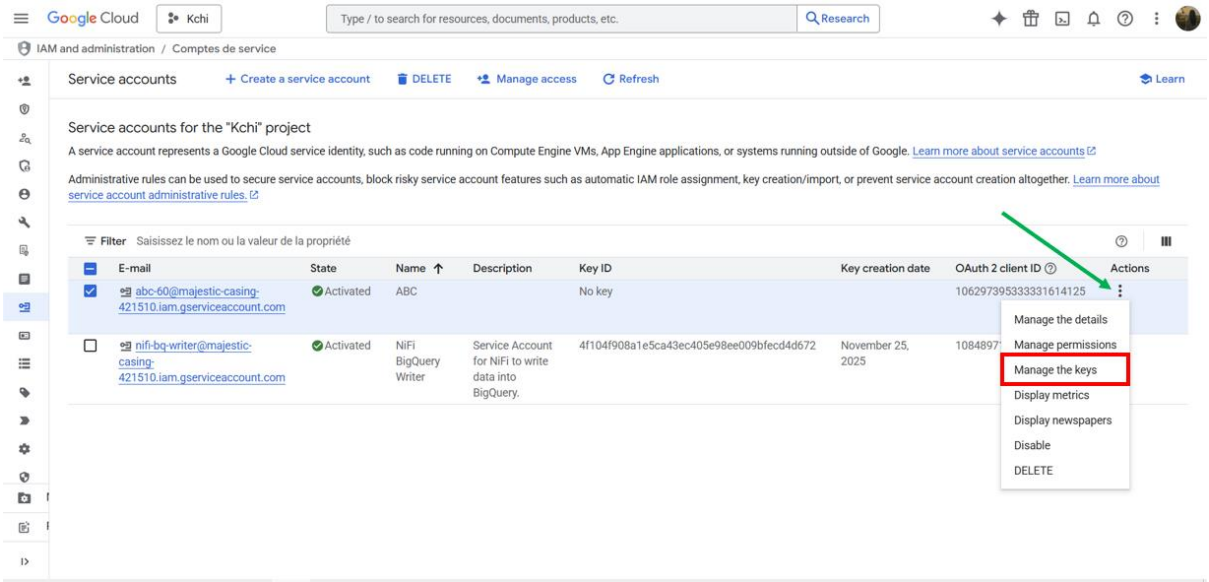
Chọn Service Account sau đó CREATE SERVICE ACCOUNT

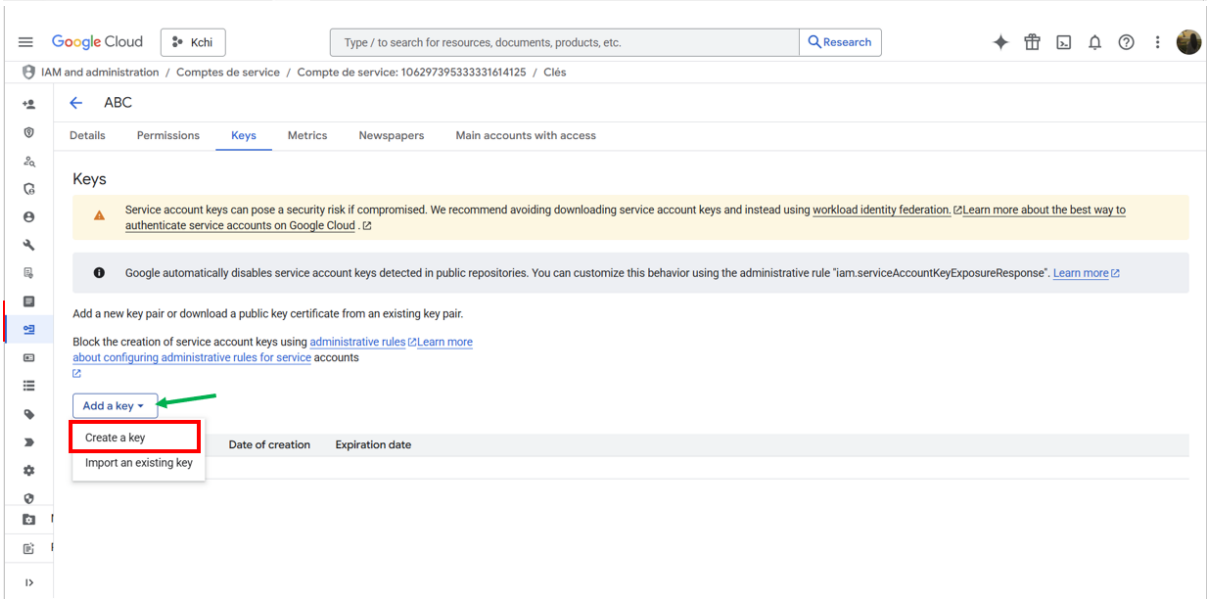


Chọn Role cho account như hình để có quyền ghi dữ liệu



Sau khi tạo account , tiến hành tạo key xác thực.





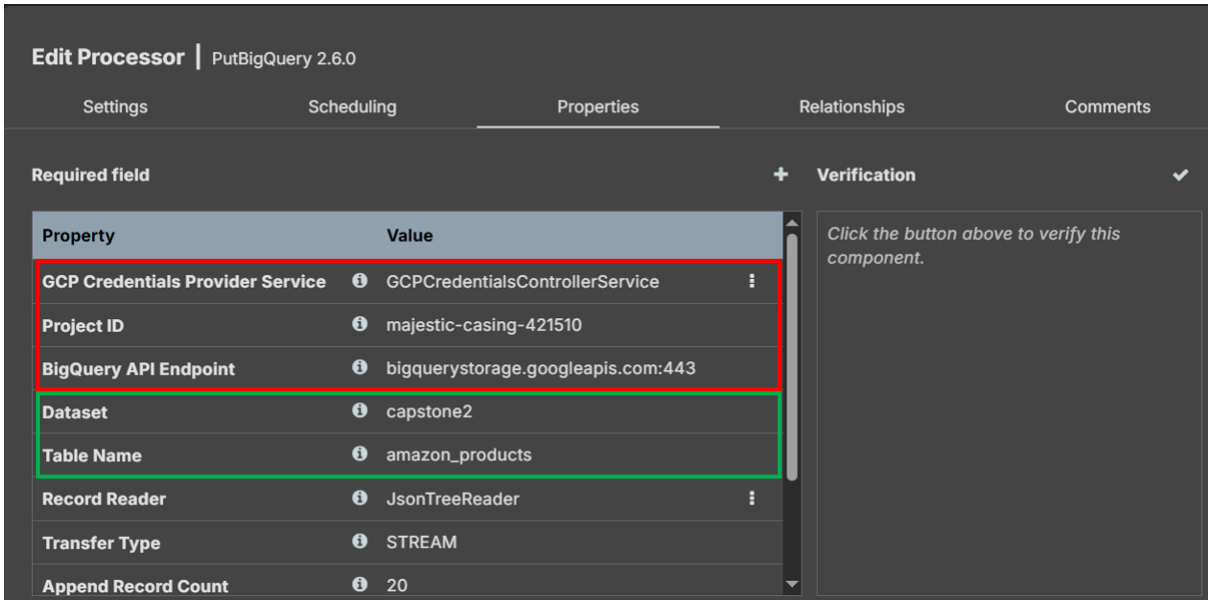
Chọn định dạng Key là Json và lưu lại.

2.3.3. Cấu hình Processor Put Bigquery

Cung cấp các thông tin sau trong cấu hình của processor:

- **Dataset**: Chọn dataset trong BigQuery lưu trữ dữ liệu bán hàng.
- **Table**: Chọn hoặc tạo bảng (table) mà dữ liệu sẽ được tải vào.Có thể tạo bảng mới nếu nó chưa tồn tại.
- **Schema**: Cung cấp schema cho bảng để đảm bảo rằng các trường trong dữ liệu (ví dụ, mã sản phẩm,phân loại sản phẩm, giá bán,đánh giá) phù hợp với các cột trong BigQuery.
- **Write Mode**: Chọn phương thức ghi dữ liệu, có thể là:
    - **Append**: Thêm dữ liệu mới vào bảng hiện tại mà không ghi đè dữ liệu cũ.
    - **Overwrite**: Xóa dữ liệu cũ và ghi dữ liệu mới lên bảng.
    - **Update**: Cập nhật dữ liệu dựa trên các khóa chính (nếu có).

Xác thực: sử dụng key JSON của Google Service Account để cấp quyền cho Apache NiFi tương tác với BigQuery.

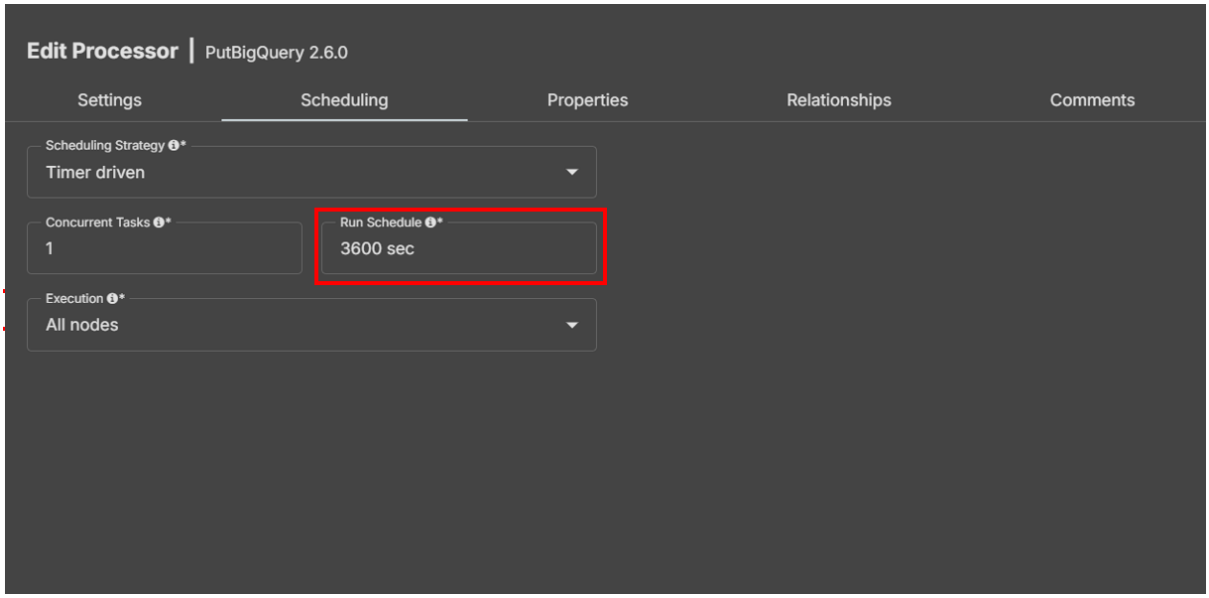


Thiết lập **GCP Credentials**, **Project ID** và **API Endpoint** giúp NiFi xác thực và kết nối được với **Google Cloud Platform**, từ đó cho phép PutBigQuery gửi dữ liệu trực tiếp vào BigQuery

Cấu hình **Dataset** và **Table Name** giúp NiFi biết chính xác dữ liệu sau khi xử lý sẽ được ghi vào đâu trong BigQuery.

2.3.4. Lập lịch load dữ liệu

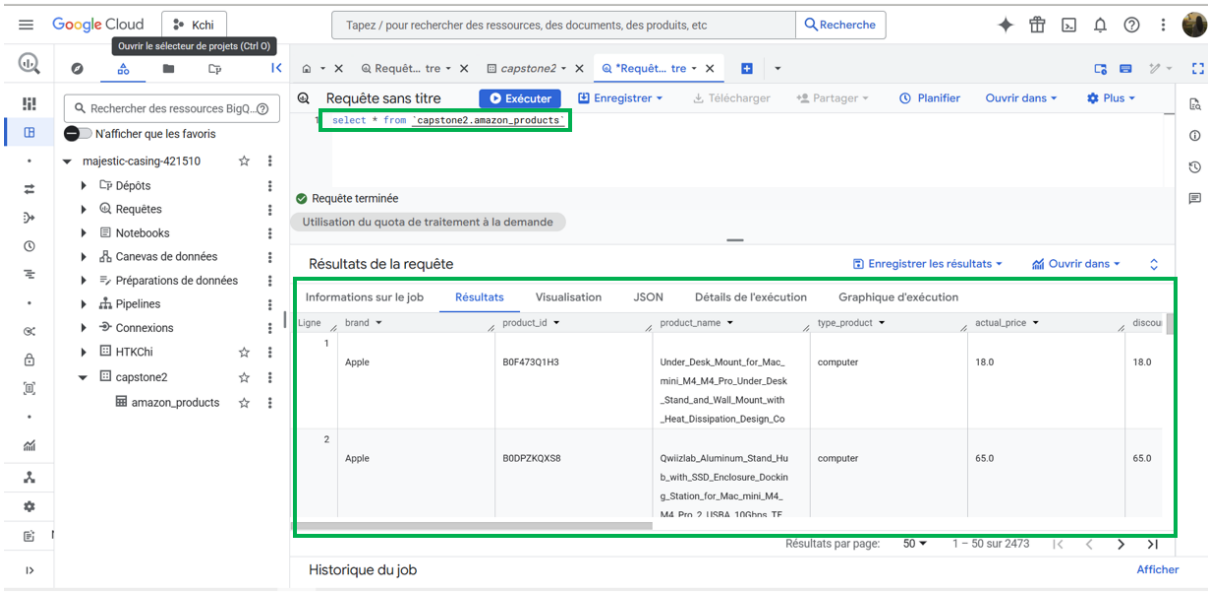
Trong tab SCHEDULING của processor PutBigQuery tiến hành đặt thời gian giữa các lần load dữ liệu , ở đây nhóm tiến hành đặt 60 phút load 1 lần.



3. Trực quan hóa dữ liệu

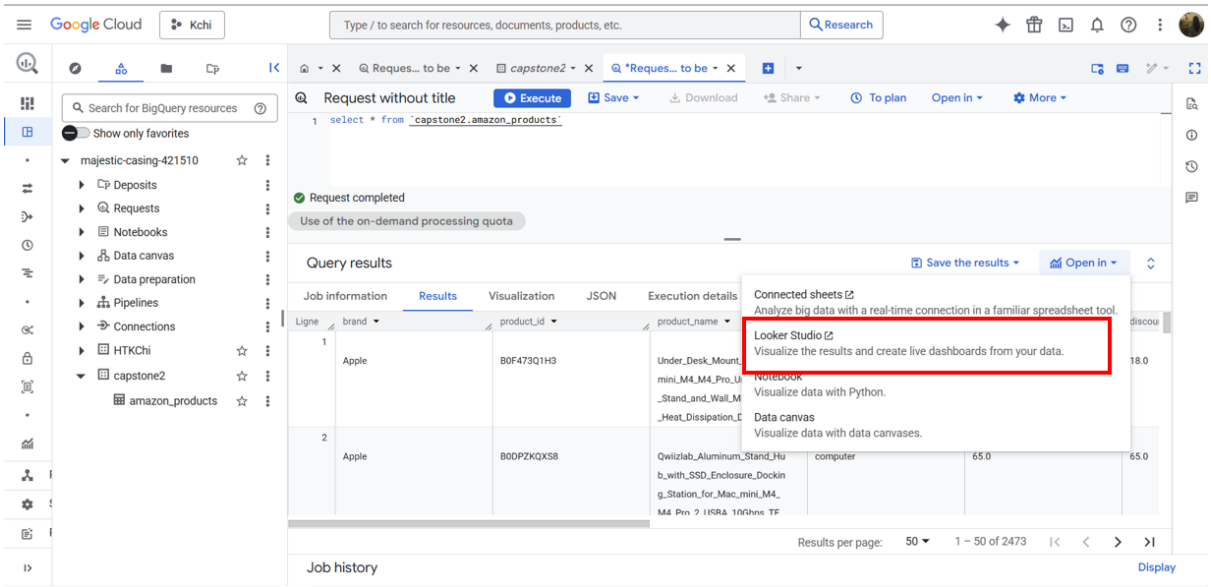
Phân tích dữ liệu:

- Sử dụng SQL để viết các truy vấn lấy dữ liệu cần phân tích.



Trực quan hóa dữ liệu

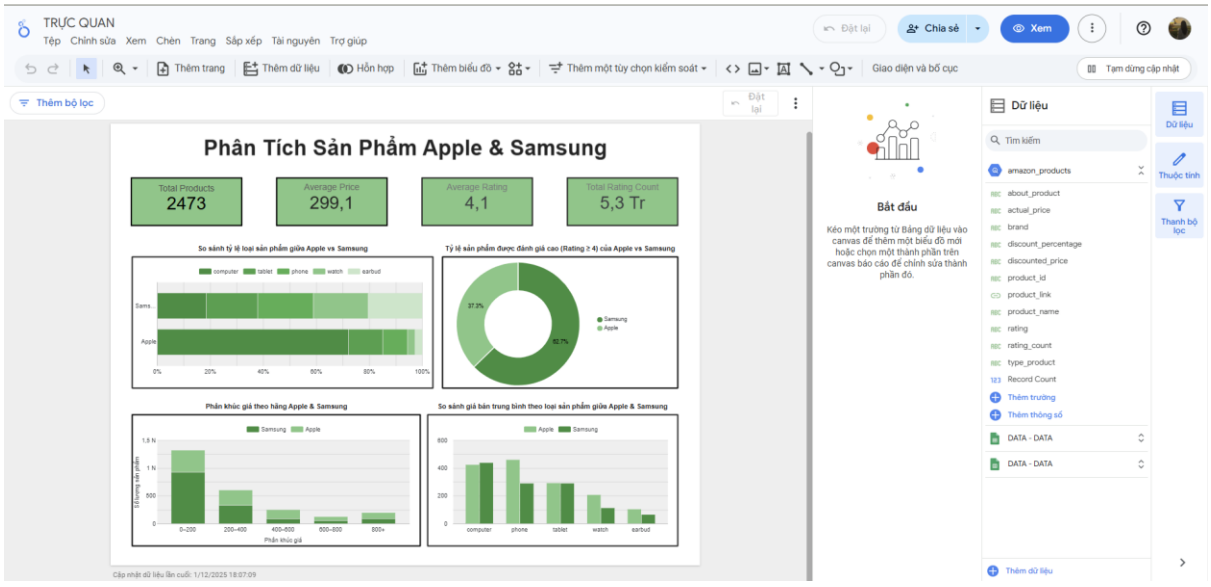
- Liên kết dữ liệu từ Google Bigquery vào Looker Studio
- Tạo các biểu đồ và bảng so sánh để hiển thị các thông tin quan trọng từ dữ liệu sản phẩm của 2 hãng Apple và Samsung



Tùy chỉnh báo cáo:

- **Filter (Bộ lọc)**: thêm các bộ lọc theo thời gian, sản phẩm hoặc khách hàng để người xem có thể tự điều chỉnh dữ liệu hiển thị.
- **Metrics và Dimensions:** Chọn các chỉ số chính (metrics) như tổng doanh thu, số

  lượng bán, và các chiều dữ liệu (dimensions) như thời gian, sản phẩm, khu vực để trực quan hóa.



Nhận xét:

- Trước hết, có thể thấy Samsung sở hữu số lượng sản phẩm nhiều hơn đáng kể. Điều này phản ánh chiến lược mở rộng danh mục theo chiều rộng, cung cấp đa dạng sản phẩm ở nhiều dòng như điện thoại, tablet, đồng hồ, máy tính và tai nghe. Việc trải rộng phân khúc giúp Samsung tiếp cận được nhiều nhóm khách hàng với nhu cầu và khả năng chi trả khác nhau. Ngược lại, Apple có số lượng

sản phẩm ít hơn, nhưng mỗi dòng sản phẩm đều mang tính tập trung và được quản lý đồng nhất về thiết kế và tính năng.

- Về giá bán, sự khác biệt giữa hai hãng thể hiện rất rõ. Ở hầu hết các nhóm sản phẩm, giá trung bình của Apple cao hơn Samsung, đặc biệt ở điện thoại và đồng hồ thông minh. Apple gần như duy trì chiến lược định vị trong phân khúc trung – cao cấp, thậm chí cao cấp, với mức giá ổn định và không có nhiều model giá rẻ. Ngược lại, Samsung phân bổ sản phẩm rất mạnh ở phân khúc 0–200 USD, cho thấy hãng tập trung vào cả thị trường phổ thông – vốn có lượng người dùng rất lớn.
- Mặc dù có mức giá cao hơn, Apple lại chiếm tỷ lệ lớn hơn về số sản phẩm được đánh giá cao (rating ≥ 4). Điều này cho thấy sản phẩm Apple được người dùng đánh giá tích cực hơn về chất lượng, độ bền, trải nghiệm sử dụng và sự nhất quán trong thiết kế. Trong khi đó, dù Samsung có nhiều sản phẩm, nhưng mức độ đánh giá "rất tốt" không cao bằng Apple, một phần do danh mục mở rộng khiến chất lượng giữa các dòng có sự khác biệt.
- Xét về loại sản phẩm, cả hai hãng đều cung cấp đầy đủ các nhóm cơ bản như phone, tablet, watch, earbud. Tuy nhiên, Apple thể hiện sự ưu thế ở các dòng sản phẩm cao cấp như iPhone, iPad, Apple Watch, trong khi Samsung mạnh hơn ở phân khúc phổ thông và tầm trung với các dòng Galaxy A, Galaxy M.

⇒Tổng kết:

- **Samsung**: mạnh về độ phủ thị trường, đa dạng sản phẩm, mức giá linh hoạt từ thấp đến cao → phù hợp nhiều nhóm người dùng.
- **Apple**: tập trung vào chất lượng, trải nghiệm người dùng và phân khúc cao cấp → do đó có tỷ lệ đánh giá tốt cao hơn dù giá bán cao.

VI. KẾT LUẬN

Sau khi hoàn thành dự án **"Xây dựng Hệ thống ETL cho Phân tích Dữ liệu Sản phẩm Apple–Samsung trên Amazon"**, người thực hiện đã đạt được nhiều kiến thức và kỹ năng quan trọng trong lĩnh vực Data Engineering và phân tích dữ liệu. Thông qua quá trình triển khai dự án, khả năng thu thập, xử lý và vận hành một pipeline dữ liệu hoàn chỉnh đã được củng cố một cách thực tế.

Trước hết, dự án giúp hiểu rõ quy trình thu thập dữ liệu web bằng Python, đặc biệt là cách sử dụng Selenium và BeautifulSoup để tự động hóa quá trình crawl dữ liệu từ Amazon. Tiếp theo, việc làm việc với Apache NiFi mang lại kiến thức quan trọng về xây dựng và quản lý luồng dữ liệu (data flow), từ tạo Processor, xử lý dữ liệu, làm sạch, chuẩn hóa cho đến cấu hình pipeline ETL hoàn chỉnh. Bên cạnh đó, kỹ năng thiết kế mô hình dữ liệu và quản lý bảng trên Google BigQuery cũng được cải thiện, đặc biệt trong tối ưu truy vấn và tổ chức dữ liệu. Cuối cùng, dự án giúp nâng cao khả năng trực quan hóa và trình bày dữ liệu bằng Looker Studio thông qua các dashboard mô tả xu hướng giá, đánh giá và mức độ quan tâm của khách hàng đối với sản phẩm Apple và Samsung.

Về **kết quả**, dự án đã xây dựng thành công một pipeline dữ liệu tự động, từ giai đoạn trích xuất dữ liệu thô đến việc phân tích và trực quan hóa. Hệ thống hoạt động ổn định, dữ liệu được xử lý nhất quán và các dashboard tạo ra đã phản ánh khá chính xác xu hướng thị trường trên Amazon. Những phân tích thu được cho phép đánh giá rõ sự khác biệt về giá bán, mức độ đánh giá và mức độ phổ biến của hai thương hiệu lớn, qua đó chứng tỏ tính hữu ích của hệ thống trong việc hỗ trợ nghiên cứu hoặc ra quyết định.

Những thành tựu đạt được:

- Pipeline ETL tự động và ổn định:
    - Hệ thống được thiết kế với khả năng trích xuất dữ liệu từ nhiều nguồn khác nhau, xử lý và chuẩn hóa dữ liệu hiệu quả, đồng thời tải dữ liệu vào kho lưu trữ Google BigQuery một cách ổn định.
    - Quy trình được lập lịch tự động, giảm thiểu can thiệp thủ công và đảm bảo tính liên tục trong việc quản lý dữ liệu.
- Hệ thống phân tích dữ liệu mạnh mẽ:
    - Với các dashboard được xây dựng trên Looker Studio, dữ liệu được cập nhật tự động giúp hệ thống cung cấp các báo cáo trực quan và dễ hiểu và kịp thời, giúp nhà quản lý nắm bắt nhanh chóng các chỉ số quan trọng như giá cả, sản phẩm có nhiều đánh giá cao, và mặt hàng bán chạy của từng hãng.
    - Dữ liệu phân tích mang tính thực tiễn cao, hỗ trợ nhà quản lý trong việc dự đoán xu hướng và tối ưu hóa chiến lược kinh doanh.
- Phát triển kỹ năng và kinh nghiệm:
    - Trong suốt quá trình thực hiện, nhóm đã cải thiện đáng kể kỹ năng làm việc với các công cụ như Apache NiFi, Google BigQuery và Looker Studio.
    - Các kiến thức về quản lý và xử lý dữ liệu lớn, thiết kế hệ thống ETL, cũng như phân tích dữ liệu đã được áp dụng và củng cố một cách thực tế.

Tuy nhiên, dự án vẫn còn một số điểm cần cải thiện. Quá trình crawl dữ liệu từ Amazon vẫn còn phụ thuộc vào cấu trúc HTML thay đổi, do đó cần tối ưu thêm logic thu thập hoặc chuyển sang API nếu có thể. Một số bước xử lý dữ liệu trên NiFi có thể mở rộng để tăng khả năng tự động phát hiện lỗi định dạng. Ngoài ra, pipeline hiện mới hỗ trợ xử lý theo lô (batch), có thể nâng cấp thêm khả năng xử lý gần thời gian thực (near real-time). Việc mở rộng phạm vi sản phẩm hoặc thêm các nền tảng thương mại điện tử khác cũng sẽ làm kết quả phân tích phong phú và toàn diện hơn.

Tổng kết lại, dự án không chỉ mang lại kết quả kỹ thuật rõ ràng mà còn giúp nâng cao tư duy hệ thống, khả năng giải quyết vấn đề và kinh nghiệm triển khai thực tế – những yếu tố quan trọng trong hành trình trở thành một Data Engineer chuyên nghiệp.

TÀI LIỆU THAM KHẢO

https://nifi.apache.org/documentation/v1/

https://www.ibm.com/topics/etl

https://200lab.io/blog/looker-studio-la-gi/

https://kyanon.digital/google-bigquery-la-gi-tim-hieu-ve-google-bigquery/

https://mmcommunications.vn/vi/e-commerce-la-gi-tam-quan-trong-cua-e-commerce-trong-ky-nguyen-so-40-n185?utm_source=chatgpt.com

https://bigdatauni.com/tin-tuc/ung-dung-big-data-trong-linh-vuc-e-commerce-phan-1.html?utm_source=chatgpt.com

https://aws.amazon.com/vi/what-is/etl/?utm_source=chatgpt.com

https://www.tailieu123.org/uploads/thuong-mai-dien-tu/2018/khai-niem-day-du-ve-thuong-mai-dien-tu.pdf?utm_source=chatgpt.com

https://bigdatauni.com/tin-tuc/ung-dung-big-data-trong-linh-vuc-e-commerce-phan-1.html?utm_source=chatgpt.com