BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN CUỐI KỲ
CÁC NỀN TẢNG DỮ LIỆU

ĐỀ TÀI: HỆ THỐNG DATA MANAGEMENT PLATFORM

Nhóm thực hiện: **Nhóm 5**

Trịnh Hà Gia Phú - 20079741

Phạm Hoàng Phúc - 20072291

Phạm Tấn Lan Anh - 20010761

Hoàng Thị Ánh Dương – 20047711

Khóa: **K16**

Lớp: **DHKHDL16A**

Người hướng dẫn: **TS. Nguyễn Chí Kiên**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024

BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN CUỐI KỲ
CÁC NỀN TẢNG DỮ LIỆU

ĐỀ TÀI: HỆ THỐNG DATA MANAGEMENT PLATFORM

Nhóm thực hiện: **Nhóm 5**

Trịnh Hà Gia Phú - 20079741

Phạm Hoàng Phúc - 20072291

Phạm Tấn Lan Anh - 20010761

Hoàng Thị Ánh Dương - 20047711

Khóa: **K16**

Lớp: **DHKHDL16A**

Người hướng dẫn: **TS. Nguyễn Chí Kiên**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024

TÓM TẮT

Đề tài xây dựng hệ thống Data Management Platform (DMP) nhằm phân loại khách hàng cho hệ thống bán lẻ online sử dụng k-means clustering và RFM segmentation tập trung vào việc tối ưu hóa quản lý và phân tích dữ liệu khách hàng để nâng cao hiệu quả kinh doanh. Hệ thống DMP là nền tảng tập trung thu thập, lưu trữ, và quản lý dữ liệu từ nhiều nguồn khác nhau, cung cấp thông tin chi tiết về hành vi và sở thích của khách hàng.

Phương pháp RFM segmentation (Recency, Frequency, Monetary) là kỹ thuật phân tích khách hàng dựa trên ba yếu tố: thời gian gần nhất khách hàng mua hàng (Recency), tần suất mua hàng (Frequency), và tổng giá trị chi tiêu (Monetary). Bằng cách phân tích dữ liệu RFM, doanh nghiệp có thể xác định các nhóm khách hàng có giá trị cao, tiềm năng, và có khả năng tái mua hàng cao.

Kết hợp với K-means clustering là một thuật toán học máy không giám sát, được sử dụng để phân nhóm khách hàng thành các cụm dựa trên sự tương đồng về hành vi mua sắm. Bằng cách kết hợp k-means với RFM, hệ thống có thể phân loại khách hàng một cách hiệu quả, giúp doanh nghiệp hiểu rõ hơn về từng nhóm khách hàng và từ đó phát triển các chiến lược tiếp thị và chăm sóc khách hàng phù hợp.

Hệ thống này không chỉ giúp tối ưu hóa chiến dịch marketing, tăng cường trải nghiệm khách hàng, mà còn góp phần cải thiện doanh thu và sự hài lòng của khách hàng. Bên cạnh đó, việc áp dụng k-means và RFM vào hệ thống DMP còn giúp tự động hóa quá trình phân tích dữ liệu, giảm thiểu sai sót và tăng cường tính chính xác trong việc đưa ra các quyết định kinh doanh.

LỜI CẢM ƠN

Trong quá trình thực hiện đề tài này, chúng em muốn gửi lời cảm ơn chân thành đến Thầy NGUYỄN CHÍ KIÊN đã hỗ trợ chúng em thực hiện đồ án cuối kỳ này.

Đầu tiên, chúng em xin gửi lời cảm ơn chân thành đến TS. NGUYỄN CHÍ KIÊN và KS. TRẦN TẤN THÀNH, giảng viên hướng dẫn và chỉ dạy môn "Các nền tảng dữ liệu". Sự chỉ bảo và phản hồi từ Thầy không chỉ giúp chúng em hiểu rõ hơn về nền tảng dữ , mà còn giúp chúng em phát triển kỹ năng và đam mê trong lĩnh vực này.

Chúng em cũng muốn bày tỏ lòng biết ơn đến các bạn bè và những người đã chia sẻ kiến thức, kinh nghiệm và ý kiến quý báu. Sự hỗ trợ từ thầy cũng như cùng các bạn học tập đã đóng góp quan trọng trong việc nâng cao chất lượng hoàn thành đề tài của chúng em.

Với bài báo cáo cuối kỳ này, chúng em rất mong sự góp ý từ thầy để đề tài trở nên hoàn thiện và có thể trong tương lai không xa sẽ được áp dụng vào thực tế.

Một lần nữa, chúng em muốn thể hiện lòng biết ơn sâu sắc đến thầy TS. NGUYỄN CHÍ KIÊN và KS. TRẦN TẤN THÀNH cùng tất cả những người đã đồng hành và đóng góp vào thành công của dự án này.

TP. Hồ Chí Minh, ngày tháng năm 2024

Tác giả

MỤC LỤC

DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

CÁC CHỮ VIẾT TẮT

DMP	Data Management Platform
DSP	Demand Side Platform
SSP	Sell/Supply Side Platform
CRM	Customer Relationship Management
GDPR	General Data Protection Regulation
CCPA	California Consumer Privacy Act
CTR	Click-Through Rate
DBT	Data Build Tool
ETL	Extract, Transform, Load
ELT	Extract, Load, Transform
RFM	Recency, frequency, Monetary
DAG	Directed Acyclic Graph
GCS	Google Cloud Storage

DANH MỤC CÁC HÌNH VẼ

DANH MỤC CÁC BẢNG

CHƯƠNG 1
GIỚI THIỆU

1.1 Mô tả bài toán

Nền tảng quản lý dữ liệu (DMP) thu thập, sắp xếp và kích hoạt dữ liệu đối tượng của bên thứ nhất, bên thứ hai và bên thứ ba từ nhiều nguồn trực tuyến, ngoại tuyến và di động. Sau đó, nó sử dụng dữ liệu đó để xây dựng hồ sơ khách hàng chi tiết nhằm thúc đẩy các sáng kiến cá nhân hóa và quảng cáo được nhắm mục tiêu. DMP cung cấp các hồ sơ khách hàng ẩn danh này cho các công cụ khác—trao đổi quảng cáo, nền tảng bên cầu (DSP) và nền tảng bên cung cấp (SSP)—để cải thiện việc nhắm mục tiêu, cá nhân hóa và tùy chỉnh nội dung.

DMP là xương sống của tiếp thị kỹ thuật số , cho phép các công ty hiểu khách hàng của họ tốt hơn.

DMP không chỉ theo dõi các đối tượng, khách hàng đã đăng ký thông tin của họ trên các "phương tiện truyền thông và quảng cáo kỹ thuật số" (thư điện tử quảng cáo, quảng cáo trả cho mỗi lần nhấp, tối ưu hóa công cụ tìm kiếm, quảng cáo hiển thị, tiếp thị truyền thông xã hội, tiếp thị nội dung, tiếp thị liên kết…) mà còn theo dõi cả các khách hàng chưa đăng ký. Cụ thể DMP thu thập khách hàng "chưa đăng ký" hay còn gọi là "khách hàng tìm năng" qua các "thẻ ẩn danh" như địa chỉ IP, thiết bị và cookie gắn thẻ các website để theo dõi thông tin về người dùng đã truy cập vào website đó và thời gian họ truy cập là bao lâu. Sau đó DMP cho phép phân nhóm khách hàng dựa vào các đặc điểm hành vi của họ

1.2 Các đặc trưng quan trọng của DMP

1.2.1. Khả năng thu thập dữ liệu :

Đa nguồn dữ liệu: Đánh giá xem DMP có khả năng thu thập dữ liệu từ nhiều nguồn khác nhau như trang web, ứng dụng di động, CRM, mạng xã hội, và các nguồn dữ liệu bên thứ ba hay không.

Chất lượng dữ liệu: Kiểm tra chất lượng của dữ liệu được thu thập, bao gồm độ chính

xác, tính toàn vẹn và độ cập nhật của dữ liệu.

1.2.2. Phân tích và sắp xếp dữ liệu:

Sắp xếp dữ liệu: Sau quá trình thu thập, các doanh nghiệp sử dụng Data Management Platform để phân chia và nhóm các dữ liệu tương tự dựa trên các tham số khác nhau.

Phân tích dữ liệu: Đây là giai đoạn Data Management Platform mô hình hóa dữ liệu để xác định thông tin hữu ích. Sau đó thống nhất thông tin này đồng thời liên tục cập nhập các dữ liệu mới nhằm cá nhân hóa và nhắm mục tiêu chính xác.

1.2.3. Phân tích đối tượng mục tiêu

DMP sẽ tiến hành lưu trữ tất cả dữ liệu khách hàng doanh nghiệp ở một nơi duy nhất, giúp hợp lý hóa việc phân tích, cập nhật và điều chỉnh chiến dịch. Bên cạnh đó, nền tảng quản lý dữ liệu liên tục phân tích dữ liệu đối tượng liên quan bằng cách dựa trên các đặc điểm nhân khẩu học, hành vi, thiết bị, vị trí,... tương tự đối tượng mục tiêu. Từ đó xác định thông tin và lập các danh mục khách hàng tiềm năng.

1.2.4. Độ tin cậy và bảo mật của DMP:

DMP có khả năng cung cấp lớp bảo mật mạnh mẽ cho người dùng và ứng dụng. Thực tế, khách hàng thời đại kỹ thuật số ngày càng quan tâm đến quyền riêng tư. Vì vậy, nền tảng tiên phong trong việc tích hợp bảo mật riêng tư sẽ nắm bắt được nhiều lợi thế so với các đối thủ cùng phân khúc thị trường.

1.2.5. Phát triển chiến lược hiệu quả:

Quản lý dữ liệu (DMP) là một công cụ quan trọng trong chiến lược marketing của doanh nghiệp. Nó giúp kết nối trực tiếp với khách hàng mục tiêu trên nhiều nền tảng và thiết bị, nhằm tối ưu hóa doanh thu và lợi nhuận.

Mặc dù có lợi ích, DMP cũng tồn tại một số hạn chế, ví dụ như chỉ phân tích dữ liệu từ các kênh kỹ thuật số và không thể tự quản lý chiến dịch quảng cáo. Do đó, việc tích hợp

DMP với các nền tảng khác là cần thiết.

Xác định hành trình khách hàng là bước quan trọng để tối ưu hóa chiến dịch marketing và thúc đẩy ý thức mua hàng của khách hàng mục tiêu.

1.2.6. Đề xuất các biện pháp quản lý rủi ro và tuân thủ pháp luật:

Với quy định bảo vệ dữ liệu chung (GDPR) và các luật bảo mật khác được áp dụng, các doanh nghiệp cần cẩn thận về cách họ thu thập và chia sẻ dữ liệu khách hàng. Tiền phạt cho việc xử lý sai dữ liệu của công dân EU theo GDPR bắt đầu từ 10 triệu euro hoặc 2% doanh thu của một công ty. DMP có thể giúp doanh nghiệp của bạn tuân thủ các giới hạn của pháp luật, tránh các khoản tiền phạt tốn kém và bảo vệ danh tiếng của bạn

Việc thu thập và sử dụng dữ liệu cá nhân phải tuân thủ các quy định về bảo vệ quyền riêng tư như GDPR ở châu Âu hay CCPA ở California. Vi phạm các quy định này có thể dẫn đến các hình phạt pháp lý nghiêm trọng và ảnh hưởng tiêu cực đến hình ảnh doanh nghiệp

1.3 Phân tích quy trình quản lý dữ liệu DMP

1.3.1. Phạm vi quản lý của DMP

Phạm vi dữ liệu :

Dữ liệu Nhân khẩu học (Demographic Data) Tuổi, giới tính, địa điểm thông tin cơ bản về người dùng.Tình trạng hôn nhân, nghề nghiệp, trình độ học vấn thông tin chi tiết hơn về hồ sơ cá nhân.

Dữ liệu Hành vi (Behavioral Data) :Lịch sử duyệt web ,Các trang web đã truy cập, thời gian trên trang, tần suất truy cập. Tương tác trên mạng xã hội Like, share, comment.Hành vi mua sắm Sản phẩm đã xem, sản phẩm đã thêm vào giỏ hàng, giao dịch mua bán.

Dữ liệu Giao dịch (Transactional Data) Lịch sử mua hàng . Lịch sử thanh toán

Dữ liệu Tương tác (Interaction Data): Mở email, click vào liên kết, hủy đăng ký. click-

through rate (CTR), conversion rate.

Dữ liệu từ các Thiết bị (Device Data) Loại thiết bị , hệ điều hành, trình duyệt: Thông tin kỹ thuật về thiết bị người dùng.

Phạm vi nguồn dữ liệu:

Nguồn Dữ liệu Nội bộ (Internal Data Sources) : Dữ liệu khách hàng, lịch sử tương tác, Dữ liệu về quản lý kinh doanh và tài chính. Dữ liệu về quản lý kinh doanh và tài chính.

Nguồn Dữ liệu Bên ngoài (External Data Sources)

Mạng xã hội: Dữ liệu từ các nền tảng như Facebook, Twitter, Instagram.Dữ liệu từ các đối tác liên kết hoặc mua dữ liệu.Dữ liệu từ các nguồn công khai như các cuộc khảo sát, thống kê.

Nguồn Dữ liệu Quảng cáo (Advertising Data Sources)

Dữ liệu từ các nền tảng mua quảng cáo tự động.Dữ liệu từ các nền tảng cung cấp không gian quảng cáo.

Phạm vi đối tượng Sử dụng:

Chuyên gia Phân tích Dữ liệu , Quản lý Chiến dịch Marketing, Quản lý Bán hàng, Quản trị Hệ thống, Chuyên viên Tích hợp, Quản lý cấp cao , Đối tác Quảng cáo

1.3.2. Phân loại và Xác định Nguồn Dữ liệu

Dữ liệu first-party là dữ liệu được thu thập trực tiếp từ một nguồn, chẳng hạn như trang web hoặc ứng dụng và thường thuộc sở hữu của chính công ty.

Dữ liệu của second-party ban đầu được thu thập bởi một bên, nhưng được chia sẻ với một bên khác thông qua một thỏa thuận. Nó bao gồm những thứ như hồ sơ mạng xã hội và đánh giá của khách hàng, được chia sẻ với bạn thông qua một đối tác đáng tin cậy.

Dữ liệu third-party là dữ liệu đã được thu thập từ nhiều nguồn và được tổng hợp và bán

bởi một tổ chức bên thứ ba. Trong quảng cáo lập trình, loại dữ liệu này có thể được sử dụng để nhắm mục tiêu đối tượng cụ thể, theo dõi hành vi của người dùng, tối ưu hóa chiến dịch và đo lường kết quả.

Nguồn dữ liệu bên ngoài: Dữ liệu từ các nguồn bên ngoài như mạng xã hội, dữ liệu hành vi từ website, dữ liệu từ đối tác.

1.3.3. Định nghĩa Quy trình Xử lý Dữ liệu

Thu thập dữ liệu: Các phương pháp thu thập dữ liệu từ các nguồn khác nhau.

Xử lý và làm sạch dữ liệu: Xử lý dữ liệu thô, loại bỏ các dữ liệu không hợp lệ, và chuẩn hóa dữ liệu.

Lưu trữ dữ liệu: Xác định cách lưu trữ dữ liệu (cơ sở dữ liệu, data warehouse, data lake).

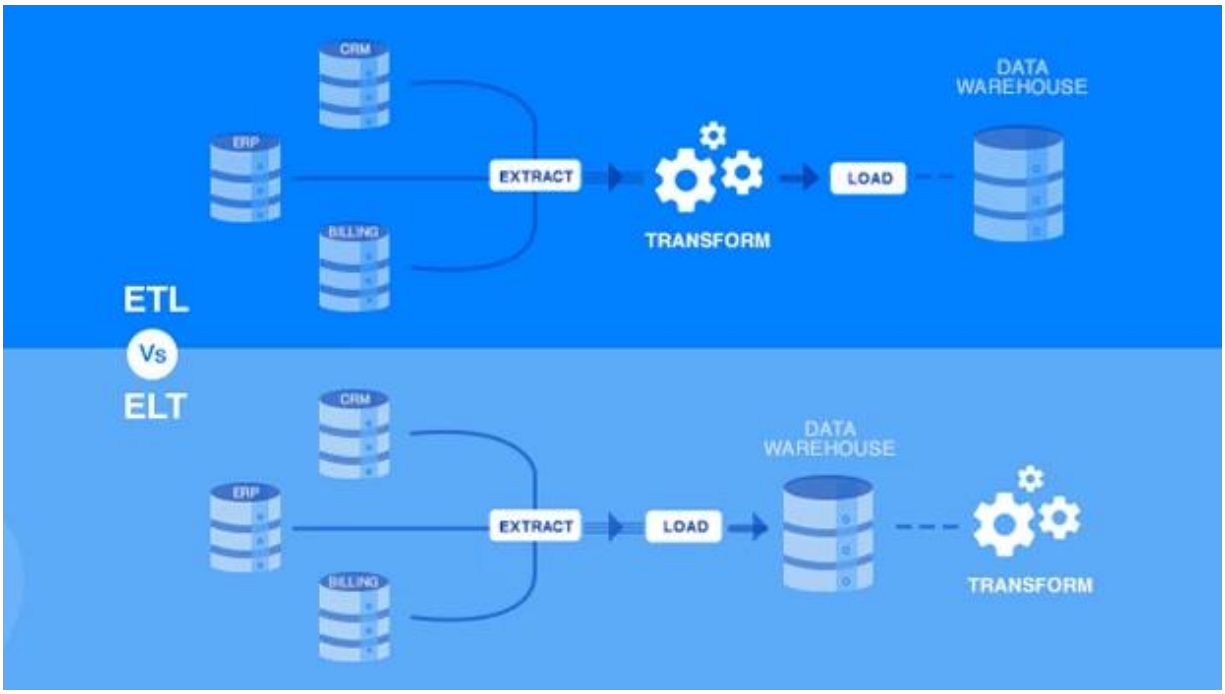
CHƯƠNG 2
CƠ SỞ LÝ THUYẾT

2.1 Các thành phần của DMP

Hệ thống DMP gồm các thành phần cơ bản như:

- Data Ingestion: Thu thập dữ liệu từ nhiều nguồn khác nhau như website, ứng dụng di động, hệ thống CRM (Customer Relationship Management), mạng xã hội, và các nguồn dữ liệu bên thứ ba.
- Data Storage: Lưu trữ dữ liệu dưới dạng cấu trúc và phi cấu trúc trong các cơ sở dữ liệu SQL, NoSQL hoặc hệ thống lưu trữ đám mây.
- Data Processing: Xử lý và biến đổi dữ liệu để làm sạch, chuẩn hóa và tích hợp dữ liệu từ các nguồn khác nhau.
- Data Analytics: Phân tích dữ liệu để khám phá thông tin giá trị, xây dựng báo cáo và dự báo xu hướng.
- Data Activation: Sử dụng dữ liệu đã xử lý và phân tích để thực hiện các chiến dịch tiếp thị, cải thiện trải nghiệm khách hàng và tối ưu hóa hoạt động kinh doanh.

2.2 Quá trình chuyển đổi dữ liệu



Hình 2.2. 1ETL và ELT Pipeline

ETL và ELT đều nhằm mục đích chuyển đổi dữ liệu từ nhiều nguồn khác nhau thành dạng có thể sử dụng cho phân tích và báo cáo. Cả hai đều bao gồm ba giai đoạn chính:

- Trích xuất dữ liệu (Extract)
- Chuyển đổi dữ liệu (Transform)
- Tải dữ liệu (Load).

2.2.1 ETL (Extract, Transform, Load)

ETL là quy trình tiêu chuẩn trong quản lý dữ liệu, được sử dụng để trích xuất dữ liệu từ các nguồn khác nhau, chuyển đổi dữ liệu đó thành định dạng phù hợp, và sau đó tải nó vào hệ thống đích, chẳng hạn như kho dữ liệu.

ETL được sử dụng rộng rãi trong các doanh nghiệp để hợp nhất dữ liệu từ nhiều nguồn khác nhau, giúp cải thiện chất lượng dữ liệu và hỗ trợ các quy trình phân tích.

Các công cụ ETL phổ biến bao gồm Apache Nifi, Informatica, và Talend.

2.2.2 ELT (Extract, Load, Transform)

ELT là một biến thể của ETL, trong đó dữ liệu được trích xuất từ nguồn và tải trực tiếp vào hệ thống đích trước khi được chuyển đổi. Quy trình này thường được sử dụng trong các môi trường big data và cloud, nơi mà các hệ thống đích có khả năng xử lý và lưu trữ mạnh mẽ.

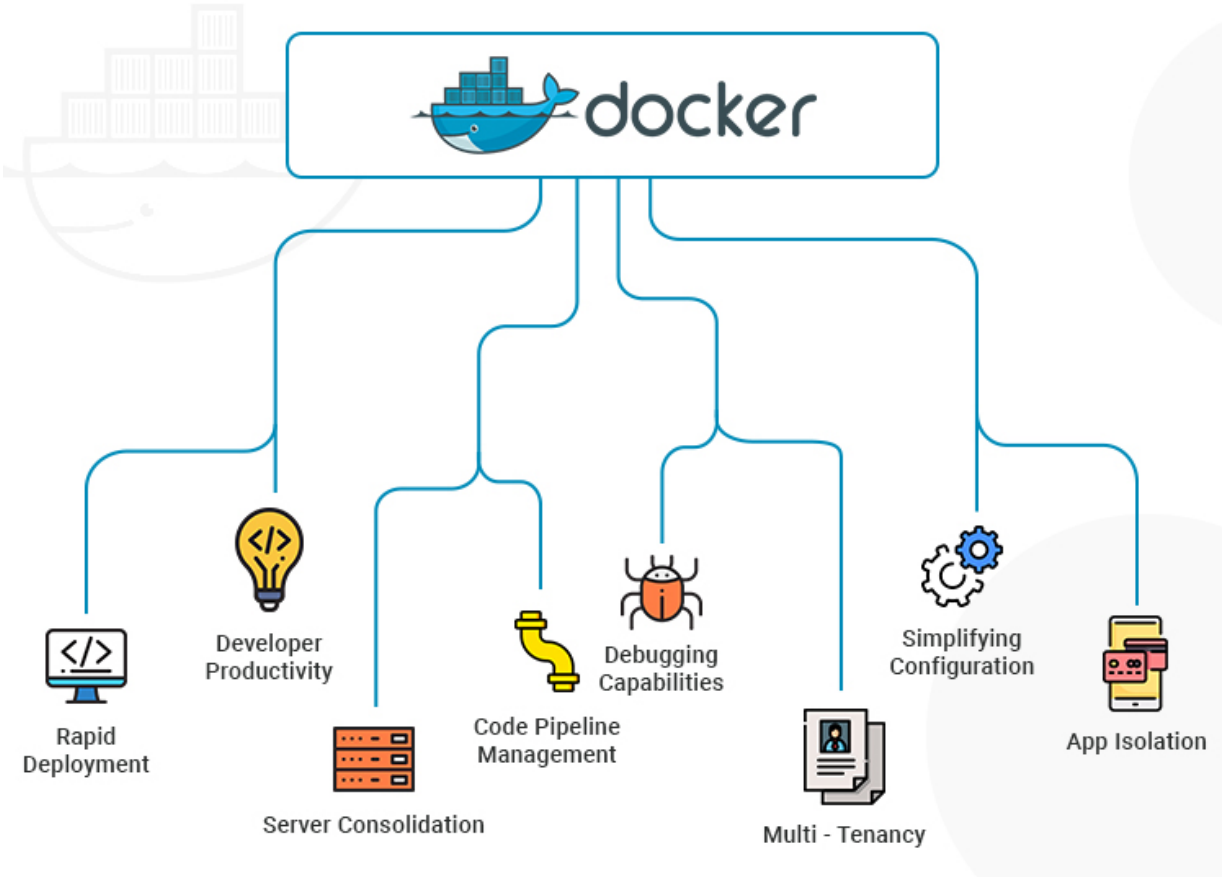
ELT phù hợp với các hệ thống đám mây hiện đại, nơi chi phi lưu trữ thấp và khả năng tính toán cao, giúp giảm thời gian tải và cho phép xử lý dữ liệu linh hoạt hơn.

Các công cụ ELT phổ biến bao gồm Google BigQuery, Amazon Redshift, và Snowflake, những nền tảng hỗ trợ việc xử lý dữ liệu trực tiếp trong hệ thống đích.

Nhóm quyết định sử dụng ELT pipeline là quy trình cho hệ thống Data Management Platform vì nhóm đang được sử dụng Google Cloud Services.

2.3 Các Công Nghệ và Công Cụ Chính trong DMP

2.3.1 Docker



Hình 2.3.1 Ứng dụng của Docker

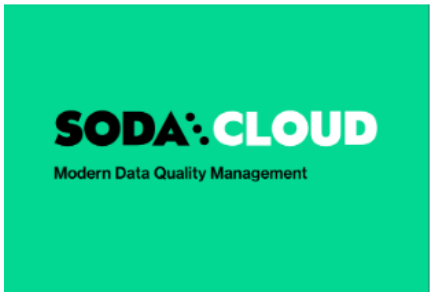
Docker là nền tảng mã nguồn mở cho phép đóng gói ứng dụng cùng với các phụ thuộc vào trong một đơn vị gọi là container. Container này có thể chạy được trên mọi môi trường, đảm bảo tính nhất quán và dễ triển khai.Docker giúp tiết kiệm tài nguyên và tăng tính linh hoạt trong phát triển và triển khai ứng dụng.

2.3.2 Astro CLI

Hình 2.3. 2 Astro CLI

Astro CLI là công cụ dòng lệnh giúp dễ dàng quản lý và triển khai Apache Airflow, một nền tảng điều phối luồng công việc mạnh mẽ. Airflow cho phép tạo, lập lịch và giám sát các quy trình công việc bằng cách sử dụng Python. Astro CLI hỗ trợ phát triển cục bộ và đồng bộ hoá với môi trường Airflow trên cloud.

2.3.3 Soda



Hình 2.3. 3 Soda

Soda là nền tảng quản lý chất lượng dữ liệu, giúp phát hiện và giám sát các vấn đề về dữ liệu trong hệ thống. Công cụ này cung cấp các tính năng kiểm tra tự động, báo cáo và cảnh

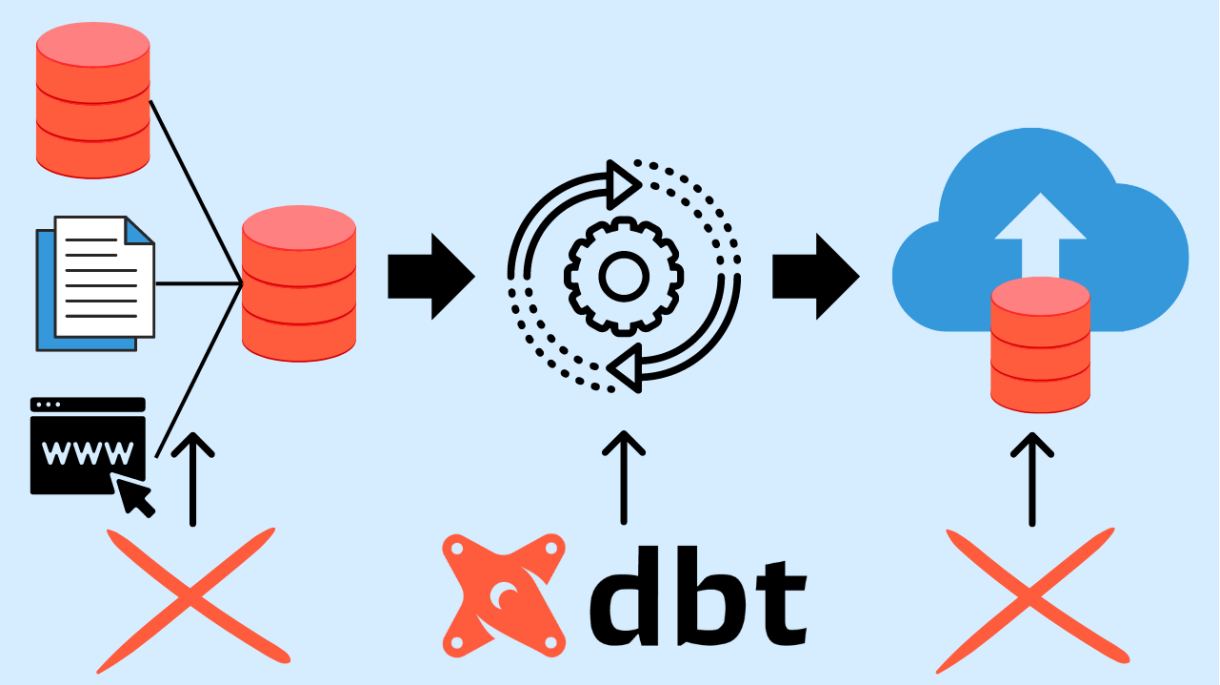
báo khi dữ liệu không đạt yêu cầu. Soda giúp đảm bảo tính chính xác và tin cậy của dữ liệu trong DMP.

2.3.4 Google Cloud Services – BigQuery

Hình 2.3. 4 Google Cloud Services - BigQuery

BigQuery là dịch vụ kho dữ liệu phân tích lớn trên nền tảng Google Cloud, cho phép xử lý và phân tích dữ liệu quy mô lớn. BigQuery hỗ trợ truy vấn SQL và có khả năng xử lý dữ liệu cực nhanh nhờ công nghệ chia sẻ và phân tán. Dịch vụ này cũng tích hợp tốt với các công cụ phân tích và quản lý dữ liệu khác trên Google Cloud.

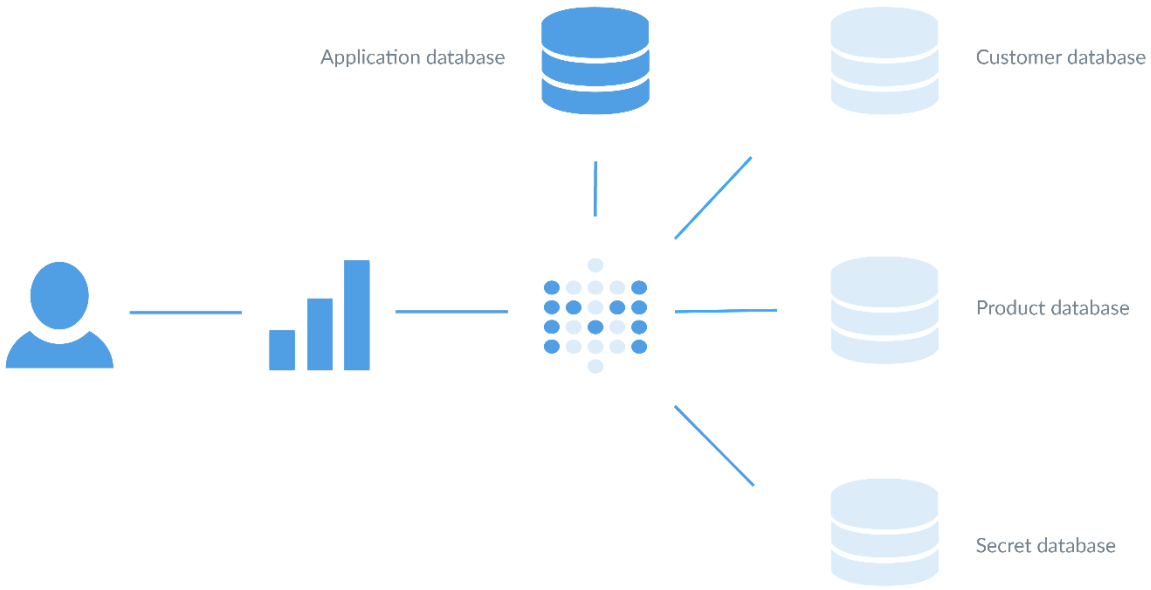
2.3.5 dbt (Data Build Tool)



Hình 2.3. 5 dbt (Data Build Tool)

Dbt là công cụ chuyển đổi dữ liệu, cho phép tạo, quản lý và triển khai các mô hình dữ liệu bằng SQL. Nó hỗ trợ quản lý mã nguồn, kiểm thử và tài liệu hoá dữ liệu, giúp cải thiện chất lượng và tính dễ dàng bảo trì của dữ liệu. dbt dễ dàng tích hợp với các kho dữ liệu như BigQuery và Snowflake

2.3.6 Metabase



Hình 2.3. 6 Metabase

Metabase là công cụ phân tích dữ liệu mã nguồn mở, giúp tạo ra các báo cáo và dashboard một cách dễ dàng mà không cần kỹ năng lập trình. Người dùng có thể trực quan hoá dữ liệu, thiết lập cảnh báo và chia sẻ kết quả với nhóm. Metabase hỗ trợ kết nối với nhiều nguồn dữ liệu khác nhau như SQL databases, Google Analytics và nhiều hơn nữa.

2.3.7 Databrick – Spark

Hình 2.3. 7 Databricks Spark

Databricks là nền tảng hợp nhất cho phân tích dữ liệu và máy học, dựa trên Apache Spark. Spark là hệ thống xử lý dữ liệu phân tán, cho phép thực hiện các tác vụ như ETL, phân tích

dữ liệu và học máy trên quy mô lớn. Databricks cung cấp giao diện thân thiện, hỗ trợ quản lý phiên bản, cộng tác và tích hợp với các công cụ đám mây khác.

CHƯƠNG 3
HIỆN THỰC

3.1 Dữ liệu

3.1.1 Dữ liệu thô

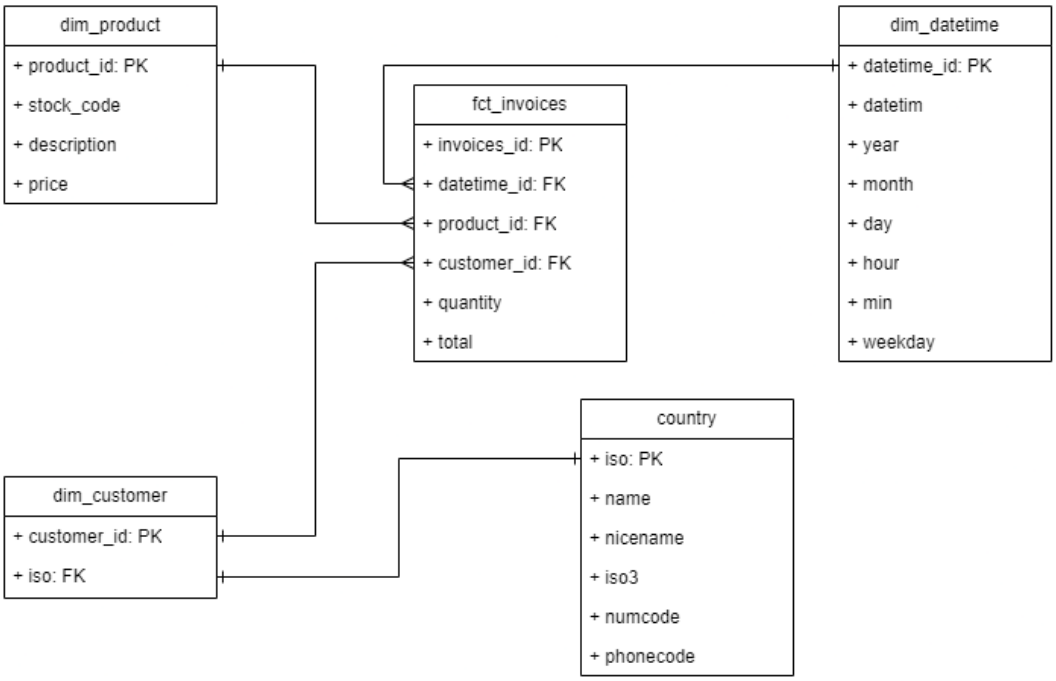
Bộ dữ liệu được cung cấp từ Kaggle, chứa các giao dịch của một cửa hàng bán lẻ trực tuyến đa quốc gia từ ngày 01/12/2010 đến 09/12/2011. Cửa hàng có trụ sở tại vương quốc Anh, khách hàng của cửa hàng là những người bán lẻ.

Dữ liệu gồm các thuộc tính:

- InvoiceNo: Số hóa đơn, gồm 6 chữ số được gán cho mỗi giao dịch. Nếu số hóa đơn bắt đầu bằng chữ "C" thì đơn hàng này bị hủy
- StockCode: Mã sản phẩm, gồm 5 chữ số được gán cho từng sản phẩm riêng biệt
- Description: Mô tả sản phẩm
- Quantity: Số lượng sản phẩm trên mỗi giao dịch
- InvoiceDate: Ngày và giờ lập hóa đơn
- UnitPrice: Đơn giá. Số, giá sản phẩm trên mỗi đơn vị đồng bảng Anh
- CustomerID: Mã khách hàng
- Country: Tên quốc gia nơi mỗi khách hàng cư trú

3.1.2 Dữ liệu được xử lí

Dữ liệu được xử lí thành các dimension tables trong DataWarehouse:

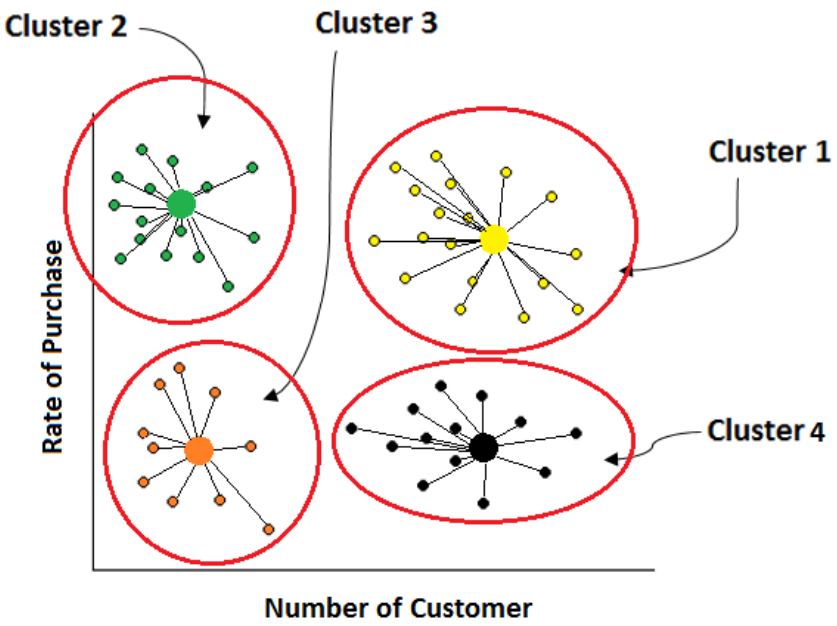


Hình 3.1: Hình Database Diagram

3.2 Các phương pháp phân loại khách hàng

3.2.1 Thuật toán K-means

K-Means Clustering là thuật toán phân cụm, bài toán đơn giản nhất trong Unsupervised learning. Trong thuật toán k-Means mỗi cụm dữ liệu được đặc trưng bởi một tâm (centroid). Tâm là điểm đại diện nhất cho một cụm và có giá trị bằng trung bình của toàn bộ các quan sát nằm trong cụm. Dựa vào khoảng cách từ mỗi quan sát tới các tâm để xác định nhãn cho chúng trùng thuộc về tâm gần nhất. Ban đầu thuật toán sẽ khởi tạo ngẫu nhiên một số lượng xác định trước tâm cụm. Sau đó tiến hành xác định nhãn cho từng điểm dữ liệu và tiếp tục cập nhật lại tâm cụm. Thuật toán sẽ dừng cho tới khi toàn bộ các điểm dữ liệu được phân về đúng cụm hoặc số lượt cập nhật tâm chạm ngưỡng.



Hình 3.2 Mô hình phân cụm K-Means

Dựa tên tập dữ liệu, mô hình Kmeans được sử dụng cho tập dữ liệu qua xử lí, mô hình phân nhóm dựa trên 3 tiêu chí: Recency, Frequency và Monetary.

3.2.2 RFM Segmentation

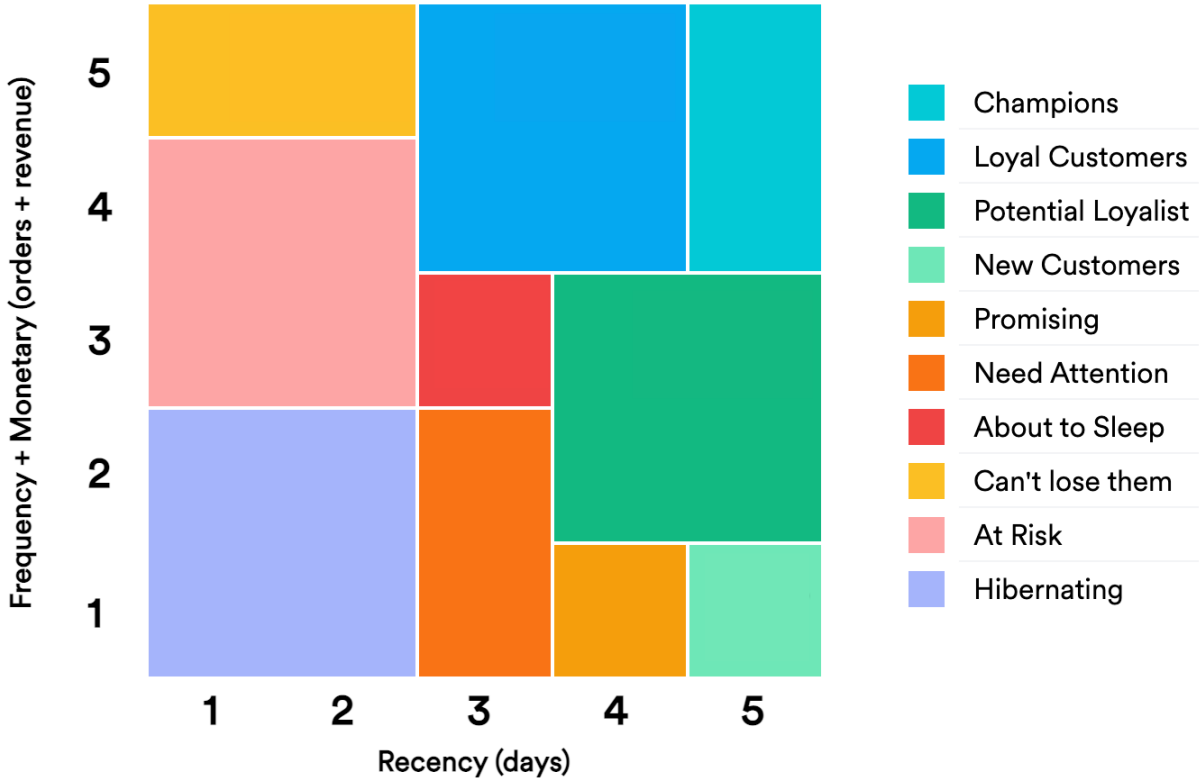
RFM là một phương pháp được sử dụng để phân tích giá trị khách hàng. Thường được sử dụng trong marketing cơ sở dữ liệu (kiểu như dựa vào dữ liệu về khách hàng để tiếp thị sản phẩm) và marketing trực tiếp và đã nhận được sự chú ý đặc biệt trong ngành bán lẻ và dịch vụ.

RFM định lượng giá trị của một khách hàng dựa trên 3 thông tin chính:

Recency: Khoảng thời gian mua hàng gần đây nhất là bao lâu. Cho biết khách hàng có đang thực sự hoạt động gần thời điểm đánh giá. Chỉ số này càng lớn càng cho thấy xu hướng rời bỏ của khách hàng càng cao.

Frequency: Tần suất mua hàng của khách hàng. Nếu khách hàng mua càng nhiều đơn thì giá trị về doanh số mang lại cho công ty càng cao và tất nhiên giá trị của họ càng lớn.

Monetary: Là số tiền chi tiêu của khách hàng. Đây là yếu tố trực quan nhất ảnh hưởng tới doanh số.



Hình 3.3 Hình phân bố RFM Scores

Phân khúc khách hàng	**Đặc điểm**
Champions	Là những khách hàng mới giao dịch, mua hàng thường xuyên và chi tiêu nhiều nhất. Những khách hàng này rất trung thành, sẵn sàng chi tiêu hào phóng và có khả năng sẽ sớm thực hiện một giao dịch mua khác.

Loyal Customers	Là những khách hàng chi tiêu ở mức trung bình – khá nhưng mua hàng rất thường xuyên.
Potential Loyalist	Là những khách hàng mới có giao dịch gần đây, chi tiêu trung bình khá và đã mua hàng nhiều hơn một lần.
Recent Customers	Những khách hàng mới mua gần đây nhất, giá trị giỏ hàng thấp và không mua hàng thường xuyên.
Promising	Là những khách hàng mới mua hàng gần đây, sức mua lớn nhưng chưa thường xuyên.
Customers Needing Attention	Là những khách hàng có tần suất mua hàng và giá trị giỏ hàng ở mức khá, chưa quay lại mua hàng gần đây.
About To Sleep	Là những khách hàng đã khá lâu chưa mua hàng, trước đó mua hàng với tần suất thấp và giá trị giỏ hàng thấp.
At Risk	Là những khách hàng đã khá lâu không quay lại và đã từng mua hàng rất thường xuyên với giá trị giỏ hàng ở mức trung bình khá.
Can't Lose Them	Là những khách hàng đã rất lâu không quay lại và từng mua hàng thường xuyên, với giá trị giỏ hàng rất lớn. Doanh nghiệp có thể đánh mất những khách hàng này nếu không có hoạt động kích thích họ quay lại.
Hibernating	Là những khách hàng đã khá lâu không quay lại, sức mua yếu (tần suất mua thấp và giá trị giỏ hàng không cao).

Lost	Là những khách hàng đã rất lâu không quay lại, tần suất mua và giá trị giỏ hàng cũng rất thấp. Nhóm này thường là những khách hàng có hành vi mua tìm kiếm sự đa dạng hoặc chỉ mua suy nhất một lần để trải nghiệm và so sánh với các sản phẩm/dịch vụ khác.

Bảng 3.1: Bảng mô tả các phân khúc khách hàng

3.2 Cấu hình phần cứng, phần mềm

Hệ thống chạy trên 1 nền tảng mã nguồn mở Docker được cấp phần cứng:

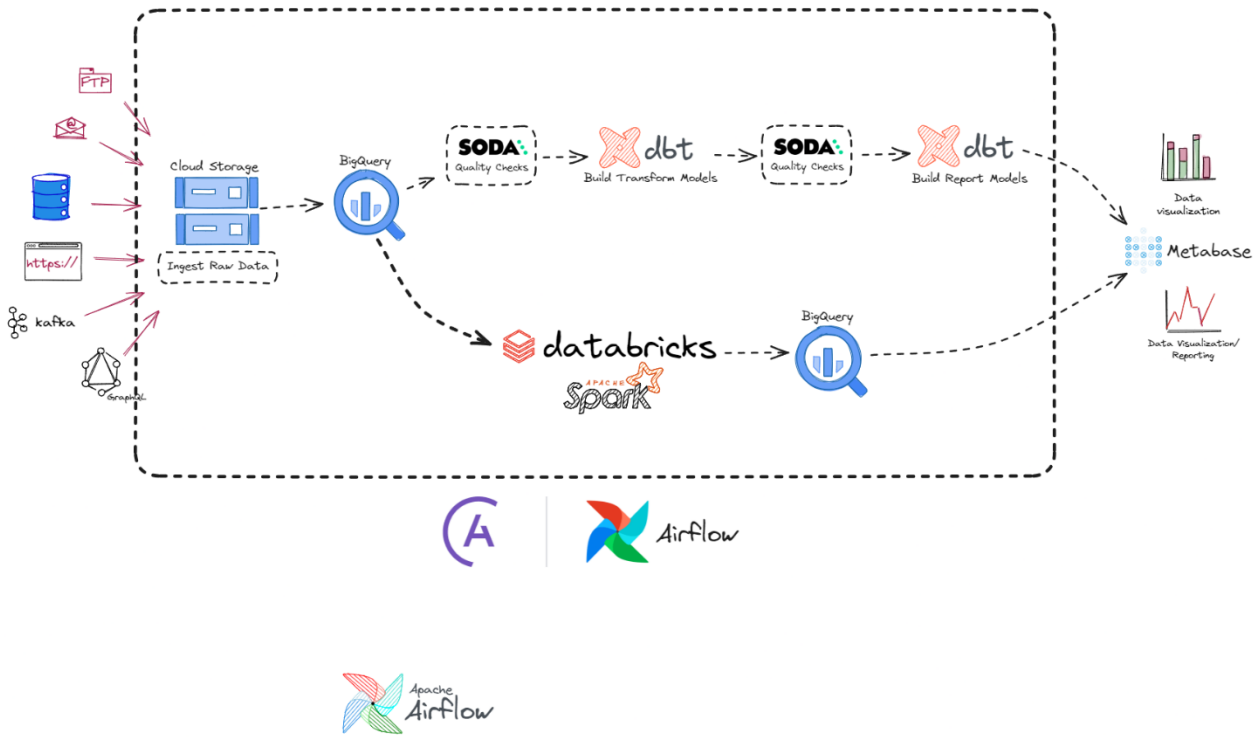
Processor: Intel(R) Core(TM) i5-8250U CPU @ 1.6gHz 8 CPUs

Ram: 7.52GB

Ngoài ra, sử dụng **Google Cloud Service** và **Databricks Cluster.**

3.3 Thực nghiệm

3.4.1 Tổng quan quy trình



Hình 3.4: Kiến trúc Data Management Platform

Data Ingestion Pipeline with DAGs in Airflow: Các tác vụ được xác định trong DAG để tải dữ liệu CSV lên Google Cloud Storage (GCS) và tạo tập dữ liệu trong BigQuery.

Data Quality Checks with Soda: Kiểm tra chất lượng dữ liệu được tiến hành thủ công bằng Soda và theo chương trình trong DAG. Việc kiểm tra xác minh tính toàn vẹn và chất lượng của dữ liệu được trích xuất.

Transform Data Using DBT: Các DBT models được tích hợp vào DAG để thực thi một cách có hệ thống, dẫn đến việc tạo ra dimension tables trong BigQuery.

Data Quality Checks on Transformed Data: Sau khi dữ liệu được chuyển đổi, các kiểm tra chất lượng dữ liệu bổ sung sẽ được thực hiện để đảm bảo rằng dữ liệu được chuyển đổi tuân thủ các tiêu chí đã chỉ định.

Processing data with Spark: BigQuery có thể xử lý data thông qua truy vấn SQL, nhưng việc sử dụng Spark có thể linh hoạt hơn nhiều. Chúng ta có thể viết mã bằng Python, xử lý

và chuyển đổi dữ liệu cũng như lưu trữ vào BigQuery. Thay vì Pandas, Spark thích hợp hơn trong việc xử lý dữ liệu lớn.

Data Visualization and Dashboards with Metabase: Metabase được thiết lập cục bộ và được sử dụng để tạo báo cáo và bảng điều khiển nhằm trực quan hóa và phân tích dữ liệu, sử dụng dữ liệu đã chuyển đổi.

3.4.2 Công nghệ sử dụng

+ **Docker:** là một nền tảng ảo hóa dựa trên container, giúp bạn đóng gói ứng dụng và các phụ thuộc của chúng vào một container duy nhất.

+ **Visual Studio Code (VS Code):** là một trình biên tập mã nguồn mở, hỗ trợ nhiều ngôn ngữ lập trình và tích hợp nhiều tiện ích mở rộng.

+ **Astro CLI - Airflow:** là một công cụ quản lý luồng công việc (workflow) dựa trên Apache Airflow. Nó giúp bạn xây dựng, lên lịch và theo dõi các tác vụ tự động trong dự án của bạn.

+ **Soda:** là một công cụ kiểm tra chất lượng dữ liệu (data quality) trong hệ thống dữ liệu của bạn. Nó giúp bạn phát hiện và khắc phục các lỗi hoặc sai sót trong dữ liệu.

+ **Google Cloud Services - Big Query:** là dịch vụ dựa trên đám mây của Google cho phép bạn truy vấn và phân tích dữ liệu lớn. Nó hỗ trợ SQL và có khả năng xử lý dữ liệu với tốc độ cao.

+ **dbt:** (data build tool) là một công cụ giúp bạn xây dựng và quản lý luồng công việc xử lý dữ liệu. Nó tập trung vào việc biến dữ liệu nguyên thủy thành dữ liệu đã xử lý.

+ **Metabase:** Metabase là một công cụ trực quan hóa dữ liệu mã nguồn mở. Nó cho phép bạn tạo báo cáo, biểu đồ và truy vấn dữ liệu một cách dễ dàng.

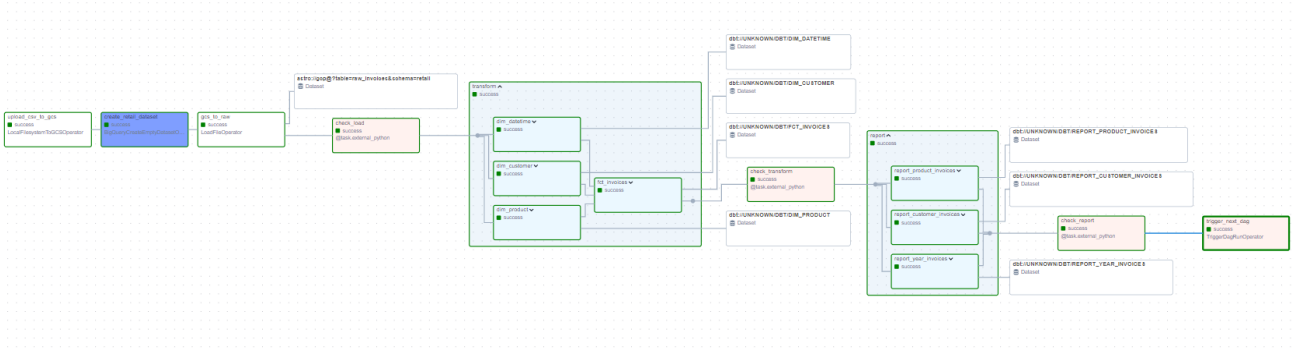
+ **Databricks:** Databricks là một nền tảng phân tích dữ liệu dựa trên đám mây, hỗ trợ Apache Spark. Nó giúp bạn xử lý và phân tích dữ liệu lớn với hiệu suất cao.

CHƯƠNG 4
KẾT LUẬN

4.1 Kết luận

Kết quả đạt được: Tạo Dashboard với 1 số biểu đồ thể hiện nhiều ý nghĩa như phân loại khách hàng, thị trường tập trung, thống kê số lượng sản phẩm đã bán ra, ...

Hình 4.1 Dashboard trực quan hóa dữ liệu cho Marketing



Hình 4.2: Graph Tasks Dag

Hạn chế: Hệ thống Data management platform cần sửu dụng dữ liệu khách hàng là chính, đặc biệt là những khách hàng ẩn danh, liệu về khách hàng còn hạn chế vì tính bảo mật thông tin cá nhân,

Trong 1 số loại dữ liệu mà DMP thường sử dụng là: Dữ liệu giao dịch, Dữ liệu ý định và hành vi khách hàng, dữ liệu Appographic. Nhóm chỉ sử dụng dữ liệu lịch sử giao dịch của khách hàng, chưa tìm được những loại dữ liệu khác.

TÀI LIỆU THAM KHẢO

[1] N. C. Minh, "Data Management Platform (DMP): Hiểu đầy đủ về khái niệm và chức năng chính," 08 01 2024. [Online]. Available: https://seothanhcong.vn/data-management-la-gi/.

[2] Hshan.T, "Exploring Customers Segmentation With RFM Analysis and K-Means Clustering With Python.," 1 11 2020. [Online]. Available: https://medium.com/swlh/exploring-customers-segmentation-with-rfm-analysis-and-k-means-clustering-93aa4c79f7a7.

[3] Connectif, "What Are RFM Scores and How To Calculate Them," 24 05 2022. [Online]. Available: https://connectif.ai/en/blog/what-are-rfm-scores-and-how-to-calculate-them/?fbclid=IwZXh0bgNhZW0CMTAAAR3zJnujMc7LlwVLvi-vpYENK7Iox_Erkq2t4TSlRyGt97K3OpJdKci2jzQ_aem_AbJkn1kdujR8OUm-2ivgbewc5_x5CiR-iQWZPVxTkuqAEUGXE1Eh4ns1HBepOgQkId7w_jvJfOkPRoq3GKIh7avp.

[4] in *Modern DataManagement*, 2023/2024.

[5] "What is a customer data platform (CDP)?," 2023. [Online]. Available: https://www.oracle.com/vn/cx/customer-data-platform/what-is-cdp/#link11.