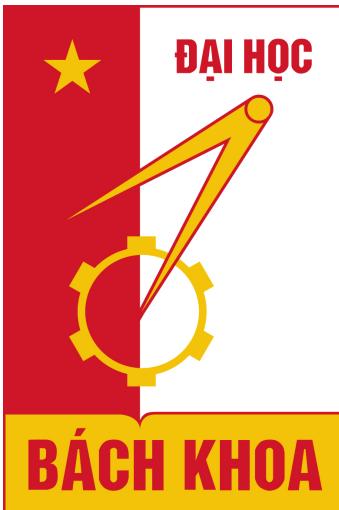


ĐẠI HỌC BÁCH KHOA HÀ NỘI

KHOA TOÁN - TIN



KHO DỮ LIỆU VÀ KINH DOANH THÔNG MINH

CHỦ ĐỀ: THƯƠNG MẠI ĐIỆN TỬ

Nhóm 16

Giảng viên hướng dẫn:	ThS. Nguyễn Danh Tú
Nhóm sinh viên thực hiện:	Đặng Duy Hậu 20216825
	Nguyễn Đình Nam 20216859
	Cao Bảo Nguyên 20216864
	Lê Nguyễn Trường Phước 20210688
	Hoàng Anh Tuấn 20216899

Hà Nội, tháng 06 năm 2024

Lời mở đầu

Lời đầu tiên, tập thể nhóm 16 xin gửi lời cảm ơn chân thành nhất đến thầy, Ths. Nguyễn Danh Tú đã đồng hành cùng tập thể nhóm và các bạn sinh viên trong môn học Kho dữ liệu và Kinh doanh thông minh kỳ 2023.2. Các nội dung thầy giảng dạy trên lớp rất có ý nghĩa đối với quá trình phát triển của bản thân tập thể nhóm với định hướng công việc đã đưa ra. Với nội dung phong phú và cách trình bày gãy gọn của thầy đã giúp tập thể nhóm có thêm nhiều kiến thức về dữ liệu, Kho dữ liệu, Kinh doanh thông minh, các quy trình cơ bản của một hệ thống phân tích dữ liệu,...

Báo cáo nhóm 16 với chủ đề Thương mại điện tử (Electronics Commerce) này đã trình bày lại chi tiết và đầy đủ hơn project của nhóm trong học phần này sau lần một thuyết trình. Mặc dù nhóm đã cố gắng để hoàn thiện các nội dung được thầy góp ý, nhưng chắc chắn rằng báo cáo có thể còn nhiều thiếu sót về nội dung và hình thức, tập thể nhóm rất mong được nhận thêm những góp ý từ thầy và các bạn để báo cáo có thể hoàn thiện hơn.

Cuối cùng, tập thể nhóm 16 xin chúc thầy có nhiều sức khỏe, đạt nhiều thành công trong công tác giảng dạy và nghiên cứu khoa học.

Tập thể nhóm chúng em xin chân thành cảm ơn thầy rất nhiều !

Hà Nội, tháng 6 năm 2024
Nhóm 16

Mục lục

Lời mở đầu	1
Tự đánh giá báo cáo	3
Đánh giá thành viên	5
Danh sách hình vẽ	6
1 Tổng quan về kho dữ liệu	9
1.1 Giới thiệu chung về Data Warehouse	9
1.2 Đặc điểm của Data Warehouse	9
1.3 Phân loại Data Warehouse	10
1.4 Kiến trúc của Data Warehouse	11
1.4.1 Các lớp của Data Warehouse	11
1.4.2 Các thành phần chính của kiến trúc Data Warehouse	13
1.4.3 Các loại kiến trúc Data Warehouse	14
1.4.4 Kiến trúc 2 lớp	14
1.4.5 Kiến trúc 3 lớp	15
1.5 Ưu nhược điểm của Data Warehouse	17
1.6 Ứng dụng của Data Warehouse	17
1.7 Mô hình dữ liệu đa chiều (OLAP)	18
1.7.1 Thành phần của OLAP	18
1.7.2 Cách hoạt động của OLAP	19
1.7.3 Các thao tác trong OLAP	20
1.7.4 Các công nghệ OLAP	20
1.8 Hệ thống dữ liệu OLTP	21
1.8.1 Giới thiệu	21
1.8.2 Các thành phần của hệ thống OLTP	21
1.8.3 Các tính năng của hệ thống OLTP	22
1.8.4 Ưu điểm và nhược điểm của hệ thống OLTP	22
1.8.5 So sánh giữa OLTP và OLAP	24
2 Tổng quan về Business Intelligence	26
2.1 Giới thiệu chung	26
2.1.1 Khái niệm	26
2.1.2 Lịch sử	27
2.2 Các thành phần của hệ thống BI	27
2.3 Ứng dụng của BI	29
2.4 Trực quan hóa dữ liệu với PowerBI	30
2.4.1 Tổng quan về PowerBI	30
2.4.2 Xử lý dữ liệu với PowerBI	32

3	Ứng dụng của DW và BI vào bài toán Electronics Commerce	35
3.1	Giới thiệu về bài toán (Requirement)	35
3.2	Giới thiệu về ODS	42
3.2.1	Data exploration	42
3.3	Kiến trúc Data Warehouse	47
3.4	Tiền xử lý dữ liệu	48
3.4.1	Khái niệm tiền xử lý dữ liệu	48
3.4.2	Tóm tắt quá trình tiền xử lý	49
3.4.3	Thiết kế Data Pipeline	49
3.4.4	Chi tiết quá trình tiền xử lý	50
3.5	Mô hình dữ liệu OLTP	58
3.6	Mô hình dữ liệu OLAP	60
3.6.1	Phân tích các chiều (Dimensions) và chủ điểm phân tích (Facts)	60
3.6.2	Data model logic	62
3.6.3	Mô hình dữ liệu quan hệ OLAP	63
3.7	Các mẫu Dashbroad	63
3.7.1	Dashboard phân tích giao dịch đặt hàng	64
3.7.2	Dashboard phân tích truy cập hệ thống	65
3.7.3	Dashboard phân tích doanh thu	67
4	Mở rộng: Trực quan hóa dữ liệu bằng mô phỏng 3D	70
4.1	Các công cụ sử dụng	70
4.1.1	SketchUp	70
4.1.2	3DBI	70
4.2	Mục đích thiết kế	72
4.3	Các quy trình xây dựng kho dữ liệu 3D	73
4.4	Lợi ích của việc trực quan hóa 3D	76
Tài liệu tham khảo		78

Tự đánh giá báo cáo

Nội dung

- Khái quát về Kho dữ liệu (Data Warehouse).
- Khái quát về Kinh doanh thông minh (Business Intelligence).
- Phân tích nghiệp vụ trong thương mại điện tử.
- Đưa ra các yêu cầu cần thiết khi xây dựng kho dữ liệu.
- Kiến trúc của Kho dữ liệu.
- Sơ đồ quá trình ETL và các hoạt động liên quan. Trình bày được tóm tắt và chi tiết thao tác
- Thiết kế được data pipeline.
- Xử lý và vẽ sơ đồ dữ liệu OLTP.
- Hệ thống các chiêu dữ liệu.
- Đưa dữ liệu vào công cụ phân tích như Power BI và thực hiện OLAP.
- Vẽ sơ đồ dữ liệu OLAP.
- Tạo các dashboard theo chủ đề.
- Phân tích dashboard.
- Nội dung mở rộng: Sử dụng công cụ mới của PowerBI.

Kết quả đạt được

Qua nội dung học phần và báo cáo, tập thể nhóm 16 đã đạt được các kết quả sau:

- Hiểu được những kiến thức nền tảng cơ bản về Kho dữ liệu, Kinh doanh thông minh, phân tích dữ liệu và phân tích kinh doanh. Bên cạnh đó xây dựng được kiến trúc của kho dữ liệu
- Khảo sát quy trình nghiệp vụ, trình bày các yêu cầu phân tích, quy mô hệ thống dữ liệu, sơ đồ ER, luồng dữ liệu, OLTP, yêu cầu phân tích.
- Thiết kế được Mind map nhu cầu phân tích, hệ thống lại bộ dữ liệu bằng data taxonomy.
- Khám phá dữ liệu, kiến trúc DWH, ETL dữ liệu bằng đa dạng công cụ, thiết kế được data pipeline.
- Sử dụng đa dạng công cụ như Power Query, Python, SQL.

6. Trình bày về data model logic, vật lý.
7. Hệ thống chiêu khái niệm.
8. Cây phân tích dashboard.
9. Dashboard đa dạng biểu đồ, có slicer, định dạng có điều kiện, thống kê số lượng dashboard, public lên website để mọi người tiện theo dõi.
10. Phân tích được Dashboard theo các chủ điểm đã đưa ra. Nội dung phân tích chi tiết
11. Sử dụng các tool trực quan hóa dữ liệu mới của PowerBI như PowerBI 3D, trực quan hóa dữ liệu bằng các công cụ khác mới hơn như Sketch up.

Đánh giá thành viên

BẢNG ĐÁNH GIÁ THÀNH VIÊN

MÔN HỌC: KHO DỮ LIỆU VÀ KINH DOANH THÔNG MINH

	HỌ VÀ TÊN:	Lê Nguyễn Trường Phước
	LỚP:	K66 - Toán Tin 01
	NHÓM:	N16

STT	Tên thành viên	Làm tốt phần việc được giao	Liên hệ được khi cần	Khả năng đóng góp sáng kiến, ý kiến cho hoạt động nhóm	Sẵn sàng giúp đỡ	Đóng góp chung vào kết quả của nhóm	Tổng điểm
1	Lê Nguyễn Trường Phước	5	5	5	5	5	25
2	Đặng Duy Hậu	5	5	5	5	5	25
3	Nguyễn Đình Nam	5	5	5	5	5	25
4	Cao Bảo Nguyên	5	5	5	5	5	25
5	Hoàng Anh Tuấn	5	5	5	5	5	25

Danh sách hình vẽ

1.1	Các lớp của kiến trúc DW	12
1.2	Thành phần của một kiến trúc Data Warehouse	13
1.3	Kiến trúc 1 lớp	14
1.4	Kiến trúc 2 lớp	15
1.5	Kiến trúc 3 lớp	16
1.6	Mô hình dữ liệu OLAP	19
1.7	OLTP to OLAP	21
2.1	Business Intelligence	26
2.2	Kiến trúc chung của hệ thống kinh doanh thông minh	28
2.3	Logo Power BI	30
2.4	Tiền xử lý dữ liệu với Power Query	32
2.5	Tạo dashboard để trực quan hóa dữ liệu	34
3.1	Giao diện trang chủ Startup Campus	37
3.2	Canvas Model	38
3.3	Mindmap nhu cầu phân tích	39
3.4	Cây phân tích Dashboard	39
3.5	Nghiệp vụ đặt hàng kèm nội dung sẽ phân tích trên Dashboard	40
3.6	Nghiệp vụ giao hàng	40
3.7	Data flow - Sơ đồ luồng dữ liệu	41
3.8	Phân bố khách hàng theo giới tính	42
3.9	Phân bố khác hàng theo giới tính	42
3.10	Phân bố khác hàng theo nhóm tuổi	43
3.11	Phân bố loại hàng hóa theo doanh thu mang lại	43
3.12	Phân bố sản phẩm theo mùa	44
3.13	Phân bố khác hàng theo giới tính	45
3.14	Phân bố khác hàng theo giới tính	45
3.15	Tương quan phí ship với tổng giá trị đơn hàng và số ngày giao hàng	46
3.16	Tương quan giá trị đơn hàng và số lượng đơn hàng	46
3.17	Tần suất số lượng đơn hàng so với số lượng khách hàng	47
3.18	Kiến trúc datawarehouse	47
3.19	Data Pipeline	49
3.20	đọc các file dữ liệu	50
3.21	Ví dụ thông tin dữ liệu file click_stream	51
3.22	xóa dòng null trong bảng product	51
3.23	kiểm tra và xóa các trùng lặp	52
3.24	Xử lý dạng thời gian	52
3.25	Sau khi xử lý thời gian	52
3.26	Xử lý giới tính trong bảng customers	53
3.27	Định dạng lại 1 số kiểu dữ liệu	53
3.28	Sinh thêm dữ liệu về tuổi và nhóm tuổi	53
3.29	Kết quả sinh thêm dữ liệu nhóm tuổi	54
3.30	Sinh thêm dữ liệu số ngày giao hàng và kết quả	54

3.31	Sinh thêm dữ liệu orders từ product_metadata và kết quả	54
3.32	Xử lý metadata của bảng order	55
3.33	Kết quả xử lý metadata bảng order	55
3.34	Trước khi xử lý metadata của click stream	56
3.35	xử lý metadata của click_stream - lần 1	56
3.36	xử lý metadata của click_stream - lần 2	56
3.37	Kết quả	57
3.38	Trước khi xóa cột thừa của bảng transaction	57
3.39	Lệnh xóa cột thừa	57
3.40	Sau khi xóa cột thừa của bảng transaction	57
3.41	Lệnh xóa cột thừa của customer	57
3.42	Lưu lại các file sau khi đã Transform	58
3.43	Thực hiện đẩy dữ liệu vào MySQL	58
3.44	OLTP	59
3.45	Hệ thống chiều khái niệm	61
3.46	Data model logic	62
3.47	Data model OLAP	63
3.48	Dashboard phân tích giao dịch đặt hàng	64
3.49	Dashboard truy cập hệ thống	66
3.50	Dashboard kết quả kinh doanh	67
4.1	SketchUp	70
4.2	3DBI	71
4.3	Mô hình kho chứa hàng tạo bởi SketchUp và 3DBI	72
4.4	Xác định vị trí của sản phẩm	73
4.5	Bảng Product kèm vị trí đã cấu hình cho kệ hàng với trường locationID	74
4.6	Kết nối PowerBI và SketchUP thông qua extension 3DBI	75
4.7	Dashboard mở rộng ứng dụng kho hàng 3D	75

Chương 1

Tổng quan về kho dữ liệu

1.1 Giới thiệu chung về Data Warehouse

Data Warehouse hay kho dữ liệu là một hệ thống cung cấp một kiến trúc mở (kiến trúc có thể thay đổi dựa vào yêu cầu của hệ thống) và công cụ cho các doanh nghiệp, nhà điều hành nhằm hỗ trợ quá trình ra quyết định, cũng như tập trung vào việc lưu trữ và phân tích dữ liệu.

Data Warehouse còn là một phần của giải pháp kinh doanh thông minh (Business Intelligence). Theo truyền thống, kho dữ liệu được lưu trữ tại chỗ và tập trung vào việc trích xuất dữ liệu từ các nguồn khác nhau. Tuy nhiên, ở thời điểm hiện tại, Data Warehouse có thể được lưu trữ trên một thiết bị chuyên dụng hoặc trên dữ liệu đám mây,... để tối ưu hóa quá trình phân tích.

Khác với Database được thiết kế để lưu trữ thì Data Warehouse vừa lưu trữ vừa hỗ trợ phân tích dữ liệu và báo cáo.

Data Warehouse sẽ có các chức năng :

- Cung cấp một góc nhìn toàn diện về doanh nghiệp.
- Cung cấp đầy đủ thông tin hiện tại và lịch sử của doanh nghiệp, sẵn sàng cho việc khai thác và sử dụng cho việc hỗ trợ ra quyết định chiến lược.
- Đảm bảo thông tin có tính nhất quán.
- Là nguồn thông tin mềm dẻo và có tính tương tác, tức là người dùng có thể lấy các thông tin khác nhau của cùng 1 đối tượng, với nhiều thao tác thay vì trả lại một danh sách tĩnh.

1.2 Đặc điểm của Data Warehouse

Data Warehouse sẽ lưu trữ dữ liệu từ nhiều nguồn khác nhau, dữ liệu trong kho thường mang tính lịch sử, biến đổi theo thời gian, không biến động và tích hợp cũng như hướng chủ đề.

- **Hướng chủ đề** : Data Warehouse cung cấp thông tin phục vụ cho một chủ đề cụ thể thay vì các hoạt động liên tục của toàn tổ chức. Điều đó có nghĩa là quy trình lưu trữ dữ liệu được đề xuất để xử lý theo một chủ đề cụ thể được xác định rõ hơn. Các chủ đề này có thể là bán hàng, phân phối, tiếp thị, ...

- **Tính tích hợp :** Tích hợp có nghĩa là thành lập một thực thể dùng chung để mở rộng quy mô tất cả dữ liệu tương tự từ các cơ sở dữ liệu khác nhau. Dữ liệu cũng được yêu cầu phải được lưu trữ trong các kho dữ liệu khác nhau theo cách được chia sẻ và cấp phép chung. Ví dụ: nếu một hệ thống sử dụng tên vùng như “Bắc Carolina” và hệ thống khác sử dụng các chữ viết tắt như “NC”, kho dữ liệu tích hợp sẽ điều chỉnh tên vùng để tạo ra một hệ thống mã hóa nhất quán.
- **Tính Biến đổi theo thời gian :** Dữ liệu trong kho được duy trì theo thời gian, cho phép phân tích xu hướng, dự báo, AI/ML và báo cáo lịch sử. Đó không chỉ là ảnh chụp nhanh về thời điểm hiện tại mà còn là dòng thời gian dữ liệu cho phép doanh nghiệp nhìn thấy những thay đổi và phát triển theo thời gian. Nó tạo ra các giới hạn thời gian khác nhau được cấu trúc giữa các bộ dữ liệu lớn và được giữ trong quy trình giao dịch trực tuyến (OLTP).
- **Tính không biến động :** Dữ liệu nằm trong kho dữ liệu là vĩnh viễn. Điều đó cũng có nghĩa là dữ liệu không bị xóa khi dữ liệu mới được chèn vào (ngoại trừ bị lấy cắp hoặc bảo trì). Trong đó, dữ liệu ở chế độ chỉ đọc và được làm mới theo các khoảng thời gian cụ thể. Điều này có lợi trong việc phân tích dữ liệu lịch sử và hiểu được chức năng.
- **Tính lịch sử :** Khả năng phân tích của nó cho phép các tổ chức thu được những hiểu biết kinh doanh có giá trị từ dữ liệu của họ để cải thiện việc ra quyết định. Theo thời gian, nó xây dựng một hồ sơ lịch sử có thể là vô giá đối với các nhà khoa học dữ liệu và nhà phân tích kinh doanh.

1.3 Phân loại Data Warehouse

Data Warehouse chia thành 3 loại chính :

- **Enterprise Data Warehouse (EDW - Kho Dữ Liệu Doanh Nghiệp):** EDW đóng vai trò là hệ thống lưu trữ dữ liệu toàn diện cho toàn bộ tổ chức hoặc doanh nghiệp. Nó tích hợp dữ liệu từ nhiều nguồn khác nhau, biến đổi và lưu trữ chúng để hỗ trợ quá trình ra quyết định toàn diện.
 - **Ưu điểm :**
 - ❖ Cung cấp một nguồn dữ liệu chính xác và đáng tin cậy cho toàn bộ tổ chức.
 - ❖ Hỗ trợ việc thực hiện các phân tích phức tạp và báo cáo quản lý.
 - ❖ Đảm bảo tính nhất quán và tính toàn vẹn của dữ liệu.
 - **Nhược điểm :**
 - ❖ Đòi hỏi nguồn lực và ngân sách lớn để triển khai và duy trì.
 - ❖ Việc tích hợp và chuẩn hóa dữ liệu có thể tốn thời gian và phức tạp.
- **Operational Data Store (ODS - Kho Dữ Liệu Hoạt Động):** ODS là một kho dữ liệu trung gian giữa hệ thống ghi chép (OLTP) và Data Warehouse. Nó chứa dữ liệu được cập nhật gần thời gian thực và hỗ trợ các nhiệm vụ hoạt động như giao dịch và cập nhật dữ liệu.
 - **Ưu điểm :**
 - ❖ Cung cấp truy cập nhanh chóng đối với dữ liệu gần thời gian thực

- ❖ Hỗ trợ quá trình ghi chép và nhiệm vụ hoạt động hàng ngày của tổ chức.

- o **Nhược điểm :**

- ❖ Không phải lúc nào cũng phản ánh dữ liệu lịch sử hoặc phân tích.
 - ❖ Có thể dẫn đến sự phức tạp trong việc quản lý và duy trì nếu không được thiết kế cẩn thận.

- **Data Mart (Kho Dữ Liệu Riêng):** Data Mart là một phần của Data Warehouse tập trung vào một lĩnh vực hoặc phòng ban cụ thể của tổ chức. Nó chứa dữ liệu liên quan đến một loại hoạt động hoặc một số lượng nhỏ người dùng cuối.

- o **Ưu điểm :**

- ❖ Dễ dàng triển khai và quản lý vì kích thước nhỏ hơn so với EDW.
 - ❖ Phục vụ nhu cầu cụ thể của một phần của tổ chức hoặc người dùng cuối.

- o **Nhược điểm :**

- ❖ Có thể dẫn đến tính không nhất quán nếu không được tích hợp chặt chẽ với EDW.
 - ❖ Không thể hỗ trợ phân tích toàn diện của tổ chức.

- o **Data Mart có ba loại Data mart chính:**

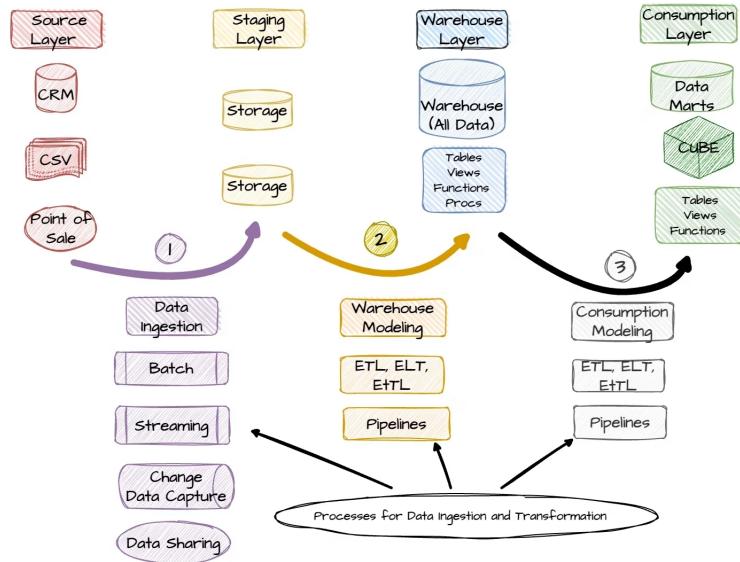
- ❖ Độc lập: Hoạt động độc lập và không phụ thuộc vào Data Warehouse.
 - ❖ Phụ thuộc: Phụ thuộc vào Data Warehouse hiện có.
 - ❖ Kết hợp: Tích hợp dữ liệu từ các nguồn bên ngoài vào một Data Warehouse hiện có.

Mỗi loại kho dữ liệu này có ưu điểm và nhược điểm riêng, và việc lựa chọn loại kho phù hợp phụ thuộc vào nhu cầu và mục tiêu của tổ chức. Thường thì một tổ chức sẽ kết hợp cả ba loại này để đảm bảo có sự linh hoạt trong việc quản lý và sử dụng dữ liệu.

1.4 Kiến trúc của Data Warehouse

Kiến trúc kho dữ liệu là một thiết kế có chủ đích của các dịch vụ dữ liệu và hệ thống con nhằm hợp nhất các nguồn dữ liệu khác nhau thành một kho lưu trữ duy nhất cho hoạt động thông minh kinh doanh (BI), AI/ML và phân tích. Bản thân kiến trúc là một tập hợp các dịch vụ logic tạo nên xương sống của hệ thống kho dữ liệu, cung cấp cách lưu trữ, quản lý và truy xuất lượng dữ liệu khổng lồ có cấu trúc và mạch lạc.

1.4.1 Các lớp của Data Warehouse



Hình 1.1: Các lớp của kiến trúc DW

- **Lớp nguồn :**

Lớp logic của tất cả các hệ thống bản ghi (SOR) cung cấp dữ liệu vào kho. Chúng có thể bao gồm các hệ thống điểm bán hàng, tự động hóa tiếp thị, CRM hoặc ERP. Mỗi SOR nguồn có một định dạng dữ liệu cụ thể và có thể yêu cầu một phương pháp thu thập dữ liệu khác nhau dựa trên định dạng dữ liệu đó.

- **Lớp dàn dựng :**

Khu vực đích cho dữ liệu từ SOR nguồn. Cách tốt nhất để dàn dựng dữ liệu là nhập dữ liệu từ SOR mà không áp dụng logic nghiệp vụ hoặc chuyển đổi. Điều quan trọng nữa là phải đảm bảo rằng dữ liệu dàn dựng không được sử dụng trong phân tích dữ liệu sản xuất; dữ liệu trong khu vực tổ chức vẫn chưa được làm sạch, chuẩn hóa, lập mô hình, quản lý và xác minh.

- **Lớp kho :**

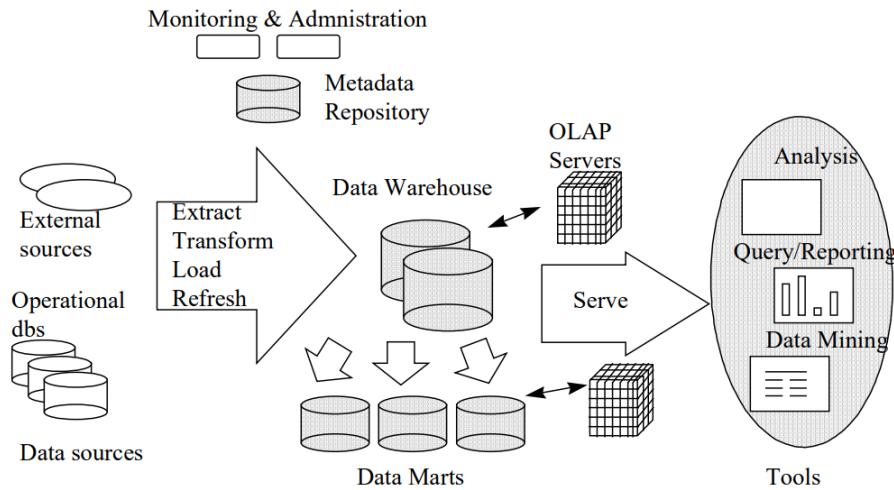
Lớp kho là nơi tất cả dữ liệu được lưu trữ. Dữ liệu kho bây giờ được định hướng theo chủ đề, tích hợp, thay đổi theo thời gian và không biến động. Lớp này sẽ có các lược đồ vật lý, bảng, dạng xem, thủ tục được lưu trữ và các hàm cần thiết để truy cập dữ liệu được mô hình hóa kho.

- **Lớp tiêu thụ :**

Còn được gọi là lớp phân tích, là nơi bạn lập mô hình dữ liệu để sử dụng bằng các công cụ phân tích như ThoughtSpot, nhà phân tích dữ liệu, nhà khoa học dữ liệu và người dùng doanh nghiệp.

Và để có thể cho dữ liệu đi lớp nguồn đến lớp dữ liệu cần 3 quy trình là: **di chuyển, làm sạch và chuyển đổi dữ liệu.**

1.4.2 Các thành phần chính của kiến trúc Data Warehouse



Hình 1.2: Thành phần của một kiến trúc Data Warehouse

- **Data Sources (Nguồn Dữ Liệu) :**

Các nguồn dữ liệu ban đầu từ hệ thống giao dịch, cơ sở dữ liệu, ứng dụng, và nguồn dữ liệu khác. Nguồn dữ liệu này có thể là dữ liệu có cấu trúc (như cơ sở dữ liệu SQL) hoặc dữ liệu không có cấu trúc (như các tệp văn bản, hình ảnh, hoặc dữ liệu từ máy chụp cảm biến). Vai trò của Data Sources là cung cấp dữ liệu nguyên thủy để được xử lý và biến đổi thành dữ liệu có ích cho phân tích.

- **ETL (Extract, Transform, Load) :**

ETL là quá trình trích xuất (Extract), biến đổi (Transform) và nạp (Load) dữ liệu từ các nguồn dữ liệu vào kho dữ liệu. Quá trình này bao gồm việc làm sạch, chuyển đổi định dạng, và tích hợp dữ liệu để nó phù hợp với cấu trúc của kho dữ liệu và khả năng phân tích.

- **Data Warehouse Database (Cơ Sở Dữ Liệu Kho Dữ Liệu) :**

Dữ liệu sau khi đã được ETL sẽ lưu trữ tại đây. Kho dữ liệu thường được thiết kế để hỗ trợ việc truy vấn và phân tích dữ liệu một cách hiệu quả. Cơ sở dữ liệu kho dữ liệu thường có cấu trúc sao cho dễ dàng tạo ra các báo cáo và truy vấn phức tạp. Vai trò chính là lưu trữ và cung cấp dữ liệu cho các ứng dụng phân tích và báo cáo.

- **Data Marts (Kho Dữ Liệu Phụ) :**

Data Marts là các phần con của kho dữ liệu được tạo ra để phục vụ cho các phần của tổ chức cụ thể, chẳng hạn như một bộ phận hoặc một nhóm công việc. Data Marts thường tập trung vào một tầm nhìn cụ thể và được tối ưu hóa cho mục tiêu đó. Data Marts cung cấp dữ liệu tập trung và tối ưu hóa cho một phần của tổ chức.

- **Metadata (Dữ Liệu Mô Tả) :**

Metadata là thông tin về dữ liệu trong kho dữ liệu, bao gồm các mô tả về nguồn gốc của dữ liệu, cấu trúc, nguồn dữ liệu, quyền truy cập và các thông tin liên quan. Metadata giúp người quản trị và người sử dụng hiểu dữ liệu và cách sử dụng nó.

Metadata hỗ trợ quản lý dữ liệu, giúp người dùng hiểu cấu trúc và tính chất của dữ liệu, và quản lý quyền truy cập.

- **Monitoring and Administration :**

Đây là quá trình theo dõi hiệu suất của kho dữ liệu và thực hiện các điều chỉnh nhằm đảm bảo việc truy vấn và phân tích dữ liệu diễn ra một cách hiệu quả. Như vậy, đây chính là thành phần giúp kho dữ liệu hoạt động ổn định và có hiệu suất tốt.

- **OLAP (Xử lý phân tích trực tuyến)**

- **Data Access Tools (Công Cụ Truy Cập Dữ Liệu):**

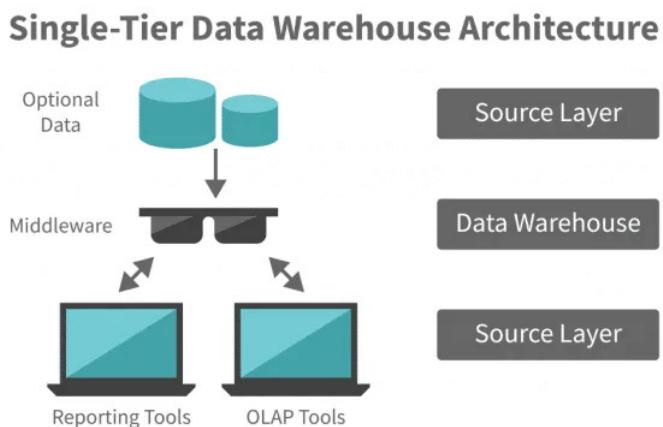
Data Access Tools được hiểu là các ứng dụng và công cụ được sử dụng để truy cập và tương tác với dữ liệu trong kho dữ liệu. Bao gồm các công cụ truy vấn SQL, các ứng dụng phân tích dữ liệu, và các giao diện trực quan cho việc tạo báo cáo và biểu đồ. Cho phép người dùng cuối truy cập, truy vấn, và hiển thị dữ liệu theo cách dễ dàng và linh hoạt.

1.4.3 Các loại kiến trúc Data Warehouse

Như ta thường được biết, kiến trúc Data Warehouse được chia thành 3 loại phổ biến .

Kiến trúc một lớp

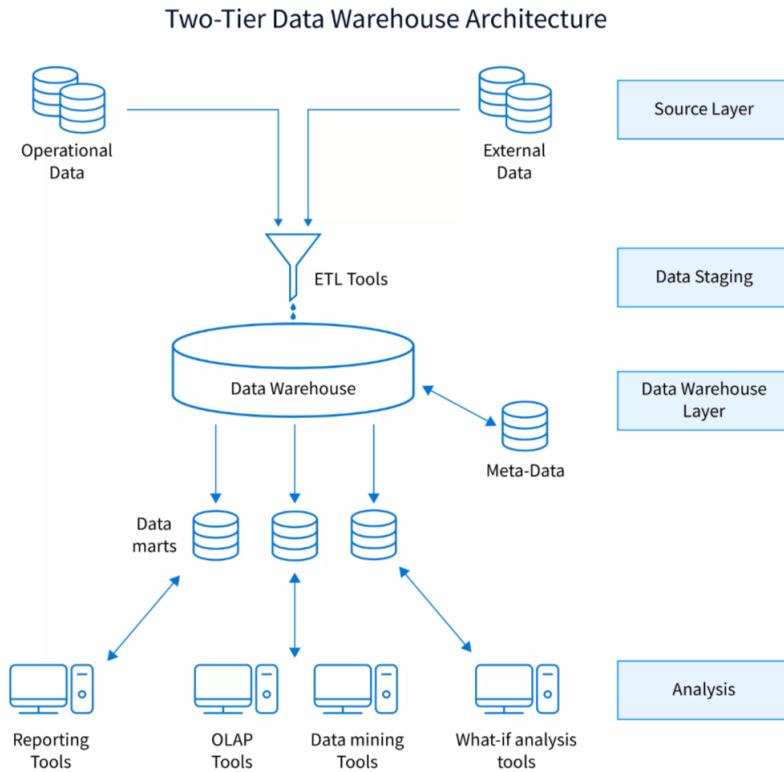
Kiến trúc một lớp không được sử dụng thường xuyên trong thực tế. Mục đích là giảm thiểu lượng dữ liệu được lưu trữ bằng cách loại bỏ những dữ liệu dư thừa. Kho dữ liệu là ảo và được triển khai dưới dạng chế độ xem đa chiều của dữ liệu vận hành. Điểm yếu chính của kiến trúc một lớp là không tách biệt được việc xử lý dữ liệu phân tích và giao dịch.



Hình 1.3: Kiến trúc 1 lớp

1.4.4 Kiến trúc 2 lớp

Trong kiến trúc hai lớp, có sự tách biệt giữa hai lớp: một lớp nguồn dữ liệu và dữ liệu lớp kho. Mặc dù nó được gọi là kiến trúc hai lớp để nhấn mạnh sự tách biệt của hai lớp, nhưng thực tế nó bao gồm bốn giai đoạn luồng dữ liệu: lớp nguồn, tầng dữ liệu, lớp kho dữ liệu và phân tích. Ngược lại với kiến trúc một lớp, trong kiến trúc này có sự tách biệt giữa dữ liệu phân tích và dữ liệu giao dịch.



Hình 1.4: Kiến trúc 2 lớp

1.4.5 Kiến trúc 3 lớp

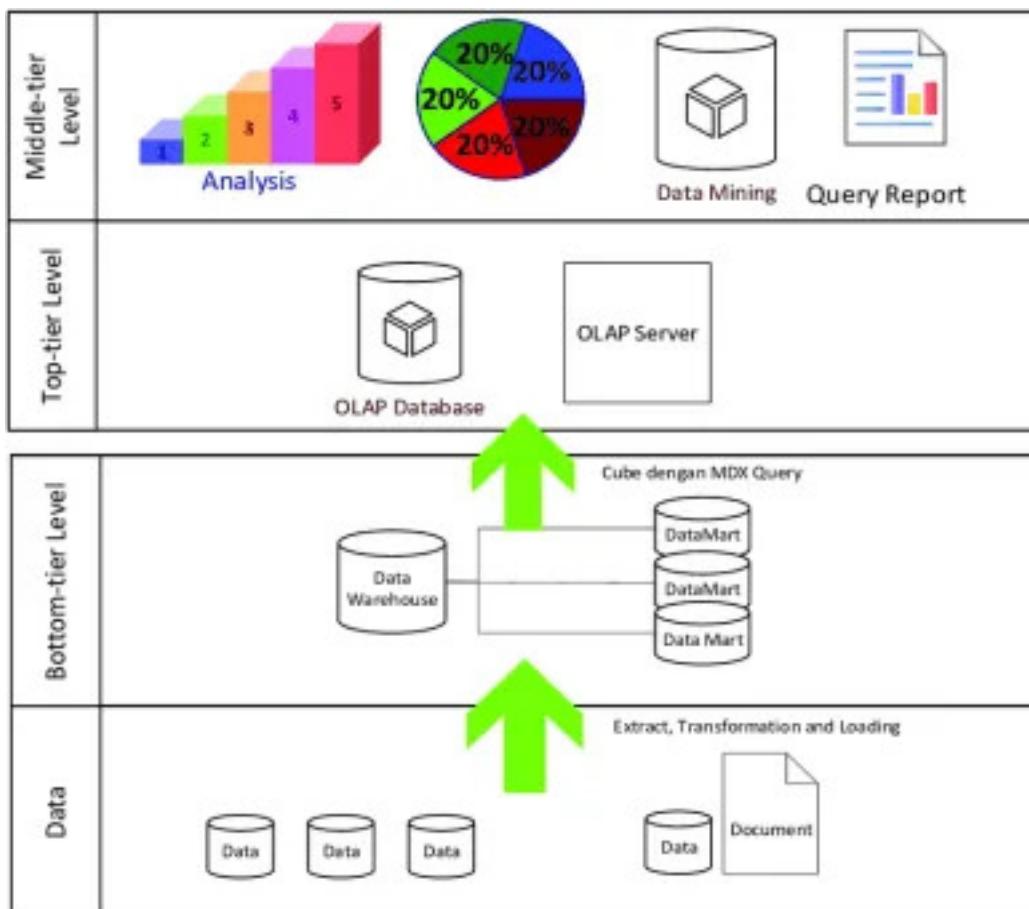
Kiểu kiến trúc kho dữ liệu ba tầng là kiểu thiết kế DWH hiện đại phổ biến nhất vì nó tạo ra luồng dữ liệu được tổ chức tốt từ thông tin thô đến những hiểu biết có giá trị.

Tầng dưới cùng trong mô hình kho dữ liệu thường bao gồm máy chủ ngân hàng dữ liệu tạo ra lớp trùu tượng trên dữ liệu từ nhiều nguồn, như ngân hàng dữ liệu giao dịch được sử dụng cho mục đích sử dụng giao diện người dùng.

Tầng giữa bao gồm máy chủ Xử lý phân tích trực tuyến (OLAP). Cấp độ này thay đổi dữ liệu thành một cách sắp xếp phù hợp hơn để phân tích và thăm dò nhiều mặt từ góc độ của người dùng. Vì nó bao gồm một máy chủ OLAP được xây dựng sẵn trong kiến trúc nên chúng ta cũng có thể gọi nó là kho dữ liệu tập trung vào OLAP.

Cấp thứ ba và trên cùng là cấp độ máy khách bao gồm các công cụ và Giao diện lập trình ứng dụng (API) được sử dụng để phân tích, tìm hiểu và báo cáo dữ liệu cấp cao.

Tuy nhiên, mọi người hầu như không đưa tầng thứ 4 vào kiến trúc kho dữ liệu vì nó thường không được coi là không thể thiếu như ba loại còn lại.



Hình 1.5: Kiến trúc 3 lớp

1.5 Ưu nhược điểm của Data Warehouse

Ưu điểm	Nhược điểm
Chất lượng dữ liệu tốt hơn: Kho dữ liệu tập trung dữ liệu từ nhiều nguồn khác nhau như hệ thống giao dịch, cơ sở dữ liệu vận hành và tệp phẳng. Sau đó, Data warehouse sẽ tiến hành tiêu chuẩn hóa để tạo ra một nguồn dữ liệu duy nhất, chính xác.	Chi phí lớn: Khi các công ty bắt đầu lưu trữ nhiều và mở rộng kho dữ liệu, chi phí đầu tư sẽ trở nên đắt đỏ.
Tăng khả năng phân tích cho doanh nghiệp: Kho dữ liệu cho phép tích hợp dữ liệu, giúp công ty tận dụng toàn bộ dữ liệu vào quá trình phân tích, đánh giá kết quả.	Lỗi đầu vào: Các lỗi đầu vào có thể ảnh hưởng đến tính toàn vẹn của thông tin được lưu trữ.
Ra quyết định thông minh hơn: Data warehouse cung cấp cho các nhà lãnh đạo thông tin chính xác để ra quyết định trong quy trình kinh doanh, quản lý tài chính và quản lý hàng tồn kho.	Tích hợp nhiều nguồn: Việc tích hợp nhiều nguồn có thể dẫn đến sự không nhất quán trong dữ liệu.
Phát triển ưu thế cạnh tranh: Tất cả các yếu tố trên kết hợp lại sẽ giúp doanh nghiệp tìm thấy nhiều cơ hội, điểm mạnh, điểm yếu của thị trường.	

Bảng 1.1: So sánh ưu điểm và nhược điểm của Data Warehouse

1.6 Ứng dụng của Data Warehouse

Hiện nay, mỗi doanh nghiệp đều cần phát triển Data warehouse để kết nối và tổng hợp thông tin từ các nguồn khác nhau. Nguyên nhân là vì kho dữ liệu là yếu tố quan trọng trong việc dự đoán, phân tích, báo cáo, triển khai kinh doanh thông minh và tạo điều kiện cho quyết định mạnh mẽ.

Dưới đây là một số ứng dụng xuất sắc của kho dữ liệu trong các ngành công nghiệp đa dạng.

- **Thương mại điện tử :**

Kho dữ liệu (Data Warehouse) được sử dụng phổ biến trong việc quản lý thông tin hàng hóa, người bán, người mua, tình trạng đơn hàng, các chương trình khuyến mãi.

- **Giáo dục :**

Kho dữ liệu là yếu tố quan trọng giúp ngành giáo dục quản lý thông tin học sinh – giáo viên – công nhân viên của trường, quản lý quá trình học tập, giáo án, bài giảng, kết quả học tập của học sinh...

- **Ngân hàng :**

Data Warehouse được dùng quản lý dòng tiền, quản lý các quỹ đầu tư, cho vay, thời hạn

thanhs toán. Ngoài ra, việc triển khai giải pháp Data Warehouse còn giúp ngân hàng tối ưu hóa quản lý tài nguyên. Điều này cho phép doanh nghiệp có thể kiểm soát thông tin về khách hàng, quản lý nguồn lực theo hướng mà họ mong muốn.

1.7 Mô hình dữ liệu đa chiều (OLAP)

Data Warehouse chủ yếu được các doanh nghiệp kinh doanh lớn sử dụng để phân tích xu hướng kinh doanh và theo dõi lợi nhuận kinh doanh của họ. Các nhà phân tích sử dụng kho dữ liệu để trích xuất thông tin kinh doanh giúp đưa ra quyết định tốt hơn. Loại quy trình ra quyết định tương tác này được cung cấp bởi các công cụ OLAP (Xử lý phân tích trực tuyến). Các ứng dụng OLAP này hầu hết chỉ sử dụng việc đọc dữ liệu để đưa ra quyết định. Các truy vấn phân tích phức tạp theo thời gian thực được trả lời bằng OLAP.

OLAP là một trong những công nghệ mạnh mẽ cung cấp các công cụ tinh vi cho doanh nghiệp để đáp ứng mục tiêu cạnh tranh của mình.

1.7.1 Thành phần của OLAP

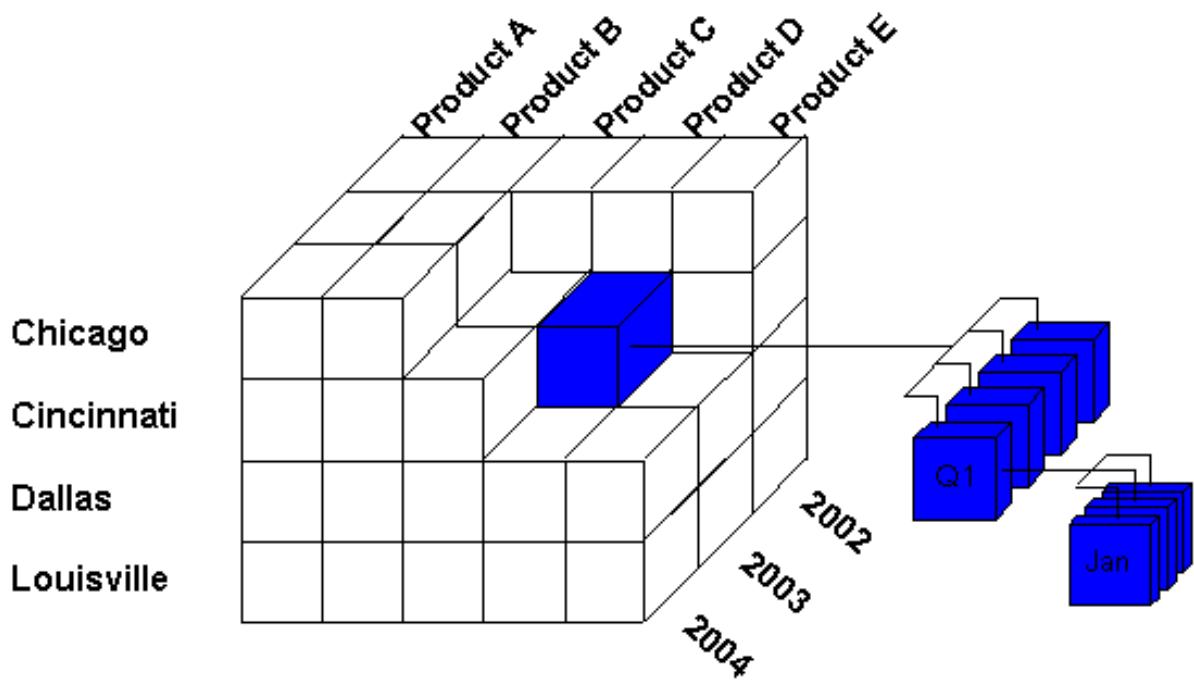
Các hệ thống xử lý phân tích trực tuyến (OLAP) lưu trữ dữ liệu đa chiều bằng cách biểu diễn thông tin dưới dạng hai chiều trở lên hoặc các danh mục. Dữ liệu hai chiều bao gồm các cột và hàng, nhưng dữ liệu đa chiều có nhiều đặc điểm hơn.

Đối tượng chính của OLAP là khối, một sự biểu diễn đa chiều của dữ liệu chi tiết và tổng thể. Một khối bao gồm một bảng sự kiện (Fact), một hoặc nhiều bảng chiều (Dimensions), các đơn vị đo (Measures) và các phân hoạch (Partitions).

Ngoài ra còn có Cây phân cấp và số liệu tổng hợp Mức độ chi tiết của các tiêu chí thể hiện cho người dùng được gọi là mức dữ liệu (data granularity), được quyết định bằng việc kết hợp các mức dữ liệu của từng cắt lớp.

Ví dụ: Người dùng có thể lựa chọn mức độ chi tiết của số liệu:

- **Chiều hàng hoá(Product)**, có các mức : sản phẩm, loại sản phẩm, công nghiệp.
- **Chiều thị trường**, có các mức : khu vực, quốc gia, thành phố, địa điểm.
- **Chiều thời gian**, có các mức : năm, quý, tháng, tuần, ngày.



Hình 1.6: Mô hình dữ liệu OLAP

Số liệu tổng hợp: Việc tổng hợp số liệu xảy ra khi người dùng thay đổi mức chi tiết của dữ liệu lấy ra từ cube, bằng cách duyệt qua cây phân cấp của cắt lớp.

Ví dụ: Nếu cắt lớp Thời gian sử dụng ở mức quý thay vì mức ngày thì doanh số của quý sẽ được tổng hợp bằng phép cộng. Tương tự, dữ liệu ở mức Tất cả được tổng hợp bằng giá trị dữ liệu của tất cả các ngày.

1.7.2 Cách hoạt động của OLAP

Một hệ thống xử lý phân tích trực tuyến (OLAP) hoạt động bằng cách thu thập, tổ chức, tổng hợp và phân tích dữ liệu theo các bước sau:

- 1 Máy chủ OLAP thu thập dữ liệu từ nhiều nguồn dữ liệu, bao gồm cơ sở dữ liệu quan hệ và kho dữ liệu.
- 2 Sau đó, các công cụ trích xuất, chuyển đổi và tải (ETL) làm sạch, tổng hợp, tính toán trước và lưu trữ dữ liệu trong một khối OLAP theo số lượng chiều được chỉ định.
- 3 Các chuyên viên phân tích kinh doanh sử dụng công cụ OLAP để truy vấn và lập báo cáo từ dữ liệu đa chiều trong khối OLAP.

OLAP sử dụng ngôn ngữ truy vấn đa chiều (MDX) để truy vấn khối OLAP. MDX là một truy vấn, tương tự như SQL, cung cấp một tập các hướng dẫn để thao tác cơ sở dữ liệu.

1.7.3 Các thao tác trong OLAP

Để thực hiện phân tích đa chiều một cách nhanh chóng và để có phản hồi truy vấn nhanh hơn, OLAP bao gồm các thao tác cơ bản sau:

- *Roll-Up*: Còn được gọi là tổng hợp trong đó dữ liệu từ cấp thấp đến cấp cao được tổng hợp để cung cấp bản tóm tắt ở cấp cao. Chọn A, B, C, SUM (số lượng).
- *Drill-down*: Cho phép điều hướng dữ liệu từ dữ liệu cấp cao hơn đến dữ liệu cấp thấp hơn.
- *Slicing*: Mô tả việc lựa chọn dữ liệu theo một chiều mà khung nhìn là một bảng.
- *Dicing*: Mô tả việc lựa chọn dữ liệu theo nhiều chiều mà chế độ xem lại là một khối phụ.

Sử dụng các thao tác trên OLAP sẽ đưa ra phân tích đa chiều theo yêu cầu của người dùng. Việc sử dụng tổng phụ của hoạt động Roll-Up có thể được tổng hợp thành tổng cuối, bằng cách sử dụng Drill-down có thể điều hướng từ tổng cuối đến tổng phụ. Bằng cách sử dụng thao tác Dicing, một khối phụ có thể được chọn. Sử dụng Slicing một mặt cắt ngang của khối được chọn tức là có thể chọn một bảng.

1.7.4 Các công nghệ OLAP

Hiện nay OLAP có 3 công nghệ chiếm ưu thế lớn:

- OLAP đa chiều (MOLAP)
- OLAP quan hệ (ROLAP)
- OLAP lai (HOLAP)

● **MOLAP:**

Xử lý phân tích trực tuyến đa chiều (MOLAP) liên quan đến việc tạo ra một khối dữ liệu đại diện cho dữ liệu đa chiều từ một kho dữ liệu. Hệ thống MOLAP lưu trữ dữ liệu được tính toán trước trong siêu khối. Các kỹ sư dữ liệu sử dụng MOLAP vì loại công nghệ OLAP này cung cấp phân tích tốc độ cao.

● **ROLAP:**

Thay vì sử dụng một khối dữ liệu, xử lý phân tích trực tuyến quan hệ (ROLAP) cho phép các kỹ sư dữ liệu thực hiện phân tích dữ liệu đa chiều trên một cơ sở dữ liệu quan hệ. Nói cách khác, các kỹ sư dữ liệu sử dụng truy vấn SQL để tìm kiếm và truy xuất thông tin cụ thể dựa trên các chiều yêu cầu. ROLAP phù hợp cho phân tích dữ liệu rộng và chi tiết. Tuy nhiên, ROLAP có hiệu suất truy vấn chậm so với MOLAP.

● **HOLAP:**

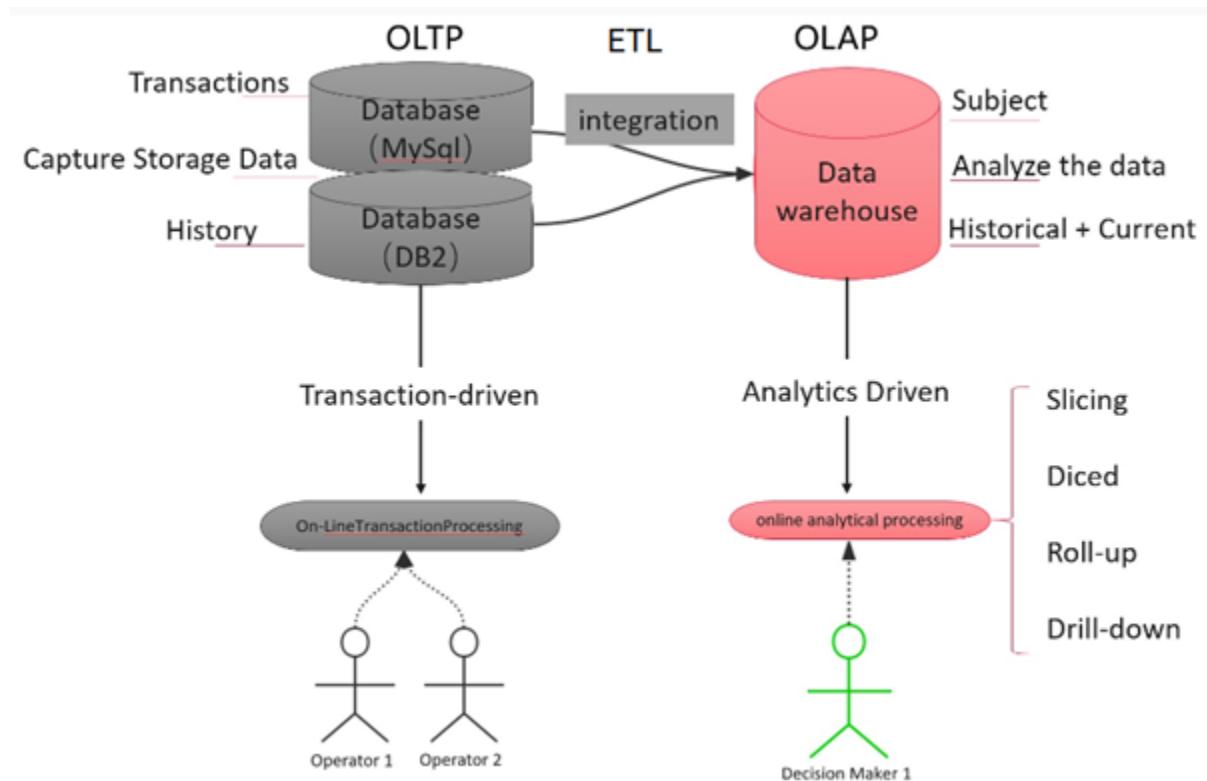
Xử lý phân tích trực tuyến lai (HOLAP) kết hợp MOLAP và ROLAP để mang tới những ưu điểm tốt nhất của cả hai kiến trúc. HOLAP cho phép các kỹ sư dữ liệu nhanh chóng lấy kết quả phân tích từ một khối dữ liệu và trích xuất thông tin chi tiết từ cơ sở dữ liệu quan hệ.

1.8 Hệ thống dữ liệu OLTP

1.8.1 Giới thiệu

OLTP là viết tắt của Online Transaction processing . Các hệ thống OLTP là các hệ thống cổ điển xử lý các giao dịch dữ liệu hay nói cách khác nó thường là một cơ sở dữ liệu. Hầu hết các ứng dụng kinh doanh đều là các hệ thống OLTP. Một ví dụ đơn giản là máy ATM hoặc hệ thống máy tính ngân hàng mà nhân viên sử dụng hằng ngày ghi lại các giao dịch, đây là hệ thống OLTP.

Một số ứng dụng khác của hệ thống OLTP là trong ứng dụng chăm sóc khách hàng, tài chính, bán hàng, quản lý nhân sự,...



Hình 1.7: OLTP to OLAP

1.8.2 Các thành phần của hệ thống OLTP

Hệ thống OLTP gồm:

- Cơ sở dữ liệu: Là nơi lưu trữ dữ liệu giao dịch và được tổ chức theo mô hình quan hệ, giúp truy cập và cập nhật dữ liệu một cách nhanh chóng và hiệu quả.
- Hệ thống ETL (Extract, Transform, Load): Công việc của hệ thống này là chịu trách nhiệm trích xuất, tách dữ liệu từ nhiều nguồn. Đồng thời chuyển đổi dữ liệu, tải dữ liệu đã được xử lý vào kho dữ liệu.
- Mô hình kiến trúc 3 tầng: Bao gồm tầng giao diện người dùng, tầng nghiệp vụ và tầng dữ liệu giúp phân chia rõ ràng nhiệm vụ và tối ưu hiệu suất

- Kho dữ liệu: Nơi lưu trữ dữ liệu được xử lý và chuẩn hóa.

1.8.3 Các tính năng của hệ thống OLTP

- **Xử lý giao dịch thời gian thực (Real-time Transaction Processing):** Hệ thống OLTP xử lý các giao dịch ngay lập tức khi chúng được thực hiện, đảm bảo sự kịp thời và chính xác trong các hoạt động kinh doanh.
- **Tính khả dụng cao (High Availability):** Hệ thống được thiết kế để luôn sẵn sàng hoạt động, với các cơ chế dự phòng và khôi phục tự động để đảm bảo hoạt động liên tục ngay cả khi xảy ra sự cố.
- **Tính toàn vẹn dữ liệu (Data Integrity):** Đảm bảo rằng các giao dịch được thực hiện một cách chính xác và đáng tin cậy.
- **Hiệu suất cao (High Performance):** Tối ưu hóa để xử lý một lượng lớn các giao dịch trong thời gian ngắn. Điều này bao gồm việc tối ưu hóa truy vấn cơ sở dữ liệu, thiết kế hệ thống phần cứng phù hợp và sử dụng các kỹ thuật cân bằng tải.
- **Tính bảo mật (Security):** Hệ thống OLTP tích hợp các cơ chế bảo mật mạnh mẽ để bảo vệ dữ liệu giao dịch khỏi các mối đe dọa từ bên ngoài và bên trong, bao gồm kiểm soát truy cập, mã hóa dữ liệu, và giám sát hoạt động.
- **Khả năng phục hồi (Recovery Capability):** Hệ thống có khả năng phục hồi sau khi xảy ra sự cố, như lỗi phần cứng hoặc phần mềm, thông qua các biện pháp sao lưu và khôi phục dữ liệu.
- **Tích hợp dễ dàng (Easy Integration):** Hệ thống OLTP có khả năng tích hợp với các hệ thống khác trong doanh nghiệp, như hệ thống ERP, CRM, và các ứng dụng web, thông qua API và dịch vụ web.

1.8.4 Ưu điểm và nhược điểm của hệ thống OLTP

Ưu điểm

- **Xử lý giao dịch nhanh chóng và chính xác:** OLTP cho phép các giao dịch được thực hiện gần như ngay lập tức, đảm bảo tính chính xác và kịp thời trong các hoạt động kinh doanh hàng ngày.
- **Tính toàn vẹn dữ liệu cao:** Sử dụng các nguyên tắc ACID (Atomicity, Consistency, Isolation, Durability) giúp đảm bảo tính toàn vẹn và nhất quán của dữ liệu trong quá trình xử lý giao dịch.
- **Hiệu suất cao:** Các cơ chế dự phòng và khôi phục tự động giúp đảm bảo hệ thống luôn sẵn sàng hoạt động, giảm thiểu thời gian ngừng hoạt động và đảm bảo liên tục trong hoạt động kinh doanh.
- **Khả năng mở rộng:** Hệ thống OLTP có thể dễ dàng mở rộng bằng cách thêm vào tài nguyên hệ thống như máy chủ hoặc bộ nhớ, phù hợp với sự phát triển và mở rộng của doanh nghiệp.

- **Quản lý đồng thời hiệu quả:** Khả năng xử lý nhiều giao dịch đồng thời mà không làm giảm hiệu suất hoặc gây xung đột dữ liệu, đảm bảo các giao dịch được thực hiện một cách mượt mà.
- **Tính bảo mật cao:** Hệ thống OLTP tích hợp các cơ chế bảo mật mạnh mẽ như kiểm soát truy cập, mã hóa dữ liệu và giám sát hoạt động, đảm bảo dữ liệu giao dịch được bảo vệ khỏi các mối đe dọa từ bên ngoài và bên trong.
- **Khả năng phục hồi nhanh chóng:** Các biện pháp sao lưu và khôi phục dữ liệu giúp hệ thống nhanh chóng phục hồi sau sự cố, đảm bảo dữ liệu không bị mất mát và các hoạt động kinh doanh không bị gián đoạn.
- **Giao diện người dùng thân thiện:** Cung cấp giao diện người dùng đơn giản và trực quan, giúp người dùng dễ dàng thực hiện và quản lý các giao dịch một cách hiệu quả.
- **Tích hợp dễ dàng với các hệ thống khác:** Hệ thống OLTP có khả năng tích hợp tốt với các hệ thống khác trong doanh nghiệp như ERP, CRM, và các ứng dụng web, giúp đồng bộ hóa dữ liệu và tối ưu hóa quy trình kinh doanh.
- **Giảm thiểu sai sót và gian lận:** Tự động hóa và kiểm soát chặt chẽ các quy trình giao dịch giúp giảm thiểu sai sót và gian lận, tăng cường độ tin cậy của hệ thống.

Nhược điểm

- **Chi phí cao:**
 - Cài đặt và duy trì: Việc triển khai và bảo trì hệ thống OLTP đòi hỏi chi phí lớn, bao gồm chi phí phần cứng, phần mềm, và nhân lực chuyên môn cao để quản lý và vận hành hệ thống.
 - Cập nhật và nâng cấp: Các chi phí liên quan đến việc cập nhật và nâng cấp hệ thống để đáp ứng nhu cầu tăng trưởng và bảo mật cũng rất đáng kể.
- **Phức tạp trong quản lý:**
 - Quản lý giao dịch đồng thời: Việc xử lý đồng thời nhiều giao dịch có thể gây ra các vấn đề về hiệu suất và đồng bộ dữ liệu, yêu cầu các biện pháp kiểm soát đồng thời phức tạp.
 - Phát hiện và xử lý lỗi: Khắc phục lỗi trong hệ thống OLTP đòi hỏi kỹ năng chuyên môn cao và có thể phức tạp do tính liên tục và sự quan trọng của dữ liệu.
- **Yêu cầu hạ tầng mạnh mẽ:** Hệ thống OLTP đòi hỏi một cơ sở hạ tầng mạnh mẽ với máy chủ, lưu trữ và mạng có hiệu suất cao để đảm bảo xử lý giao dịch nhanh chóng và đáng tin cậy.
- **Tính phức tạp của phần mềm:**
 - Phát triển và bảo trì ứng dụng: Các ứng dụng OLTP thường phức tạp và đòi hỏi nhiều công sức trong việc phát triển, thử nghiệm, và bảo trì.
 - Tích hợp với các hệ thống khác: Việc tích hợp OLTP với các hệ thống khác (như ERP, CRM) có thể gặp nhiều thách thức kỹ thuật.

- **Rủi ro bảo mật cao:** Mặc dù hệ thống OLTP có các cơ chế bảo mật, nhưng do thường xuyên xử lý các giao dịch nhạy cảm, hệ thống này là mục tiêu hấp dẫn của các cuộc tấn công mạng. Việc bảo vệ dữ liệu và ngăn chặn truy cập trái phép đòi hỏi các biện pháp bảo mật phức tạp và liên tục được cập nhật.
- **Khả năng phục hồi sau thảm họa:** Mặc dù có các biện pháp sao lưu và khôi phục, việc đảm bảo phục hồi nhanh chóng và đầy đủ sau các thảm họa lớn vẫn là một thách thức lớn, đòi hỏi kế hoạch và tài nguyên đáng kể.
- **Hiệu suất giảm khi khối lượng công việc tăng cao:** Hệ thống OLTP có thể gặp phải các vấn đề về hiệu suất khi khối lượng công việc tăng đột ngột hoặc trong các giai đoạn cao điểm, đòi hỏi phải có kế hoạch mở rộng tài nguyên và tối ưu hóa hệ thống liên tục.
- **Khả năng quản lý dữ liệu lớn:** Mặc dù OLTP xử lý tốt các giao dịch ngắn và thường xuyên, nhưng việc quản lý và lưu trữ khối lượng lớn dữ liệu lịch sử có thể gặp khó khăn, đòi hỏi các chiến lược lưu trữ và quản lý dữ liệu hiệu quả.

Thách thức

- Quản lý hiệu suất: Xử lý đồng thời và tối ưu hóa truy vấn phức tạp.
- Bảo mật: Bảo vệ dữ liệu và tuân thủ quy định pháp lý.
- Độ tin cậy và khả năng phục hồi: Khả năng phục hồi sau thảm họa và duy trì tính khả dụng cao.
- Quản lý dữ liệu: Quy mô và lưu trữ dữ liệu lớn, tích hợp với các hệ thống khác.
- Chi phí: Đầu tư ban đầu và bảo trì cao, cùng với chi phí cập nhật và nâng cấp.
- Quản lý phức tạp: Phát triển và bảo trì phần mềm, quản lý giao dịch đồng thời.
- Tối ưu hóa hạ tầng: Cân bằng tải và mở rộng hệ thống.
- Đào tạo và nhân lực: Đòi hỏi nhân lực có kỹ năng cao và đào tạo liên tục.

1.8.5 So sánh giữa OLTP và OLAP

	OLTP	OLAP
Ứng dụng	Vận hành: ERP, CRM, ứng dụng truyền thông, tin tức, weblog, ...	Hệ thống quản lý thông tin, Hệ thống hỗ trợ ra quyết định,...
Người dùng chủ yếu	Người dùng, nhân viên	Quản lý, điều hành
Phạm vi xử lý	Hàng tuần/tháng	Hàng năm
Mức độ làm mới	Ngay tức thì	Định kỳ
Mô hình dữ liệu	Entity-relationship, hoặc thích thì có thể BigData, key-value,...	Multi-dimensional
Schema	Normalized (Chuẩn hoá dữ liệu)	Star hoặc snowflake
Nhấn mạnh	Cập nhật	Truy hồi thông tin
Số lượng người dùng	Hàng ngàn tới hàng triệu	Hàng trăm
Dung lượng ổ đĩa	MB tới GB	GB tới TB
Đo lường	Số lượng giao dịch/người dùng đồng thời	Số lượng câu truy vấn và thời gian phản hồi

Bảng 1.2: So sánh giữa OLTP và OLAP

Chương 2

Tổng quan về Business Intelligence

2.1 Giới thiệu chung

2.1.1 Khái niệm

Theo Forrester Research , kinh doanh thông minh (Business Intelligence) là "một tập hợp các phương pháp, quy trình, kiến trúc và công nghệ biến dữ liệu thành thông tin có ý nghĩa và hữu ích được sử dụng để cho phép hiểu biết sâu sắc về chiến lược, chiến thuật và hoạt động cũng như ra quyết định hiệu quả hơn trong kinh doanh."



Hình 2.1: Business Intelligence

Theo định nghĩa này, kinh doanh thông minh (Business Intelligence) bao gồm quản lý thông tin (tích hợp dữ liệu , chất lượng dữ liệu , lưu trữ kho dữ liệu, quản lý dữ liệu thông minh, phân tích nội dung văn bản,...). Do đó, Forrester coi việc chuẩn bị dữ liệu và sử dụng dữ liệu là hai phân đoạn riêng biệt nhưng được liên kết chặt chẽ với nhau trong kiến trúc kinh doanh thông minh.

Một số yếu tố trong kinh doanh thông minh, gọi tắt là BI, bao gồm:

- Kết hợp và phân bổ đa chiều.
- Phi chuẩn, gán nhãn và chuẩn hóa.
- Báo cáo thời gian thực với cảnh báo phân tích.

- Phương pháp giao tiếp với các nguồn dữ liệu phi cấu trúc.
- Hợp nhất nhóm, lập ngân sách và dự báo theo kiểu rolling.
- Suy luận thống kê và mô phỏng xác suất.
- Tối ưu hóa các chỉ số hiệu suất chính.
- Kiểm soát phiên bản và quản lý quy trình.
- Quản lý mục mở.

Forrester phân biệt điều này với thị trường kinh doanh thông minh, vốn "chỉ là các lớp trên cùng của kiến trúc BI, chẳng hạn như báo cáo, phân tích và dashboard" [1].

2.1.2 Lịch sử

Năm 1865, Richard Millar Devens trình bày cụm từ “Kinh doanh thông minh” (BI) trong “Cyclopædia of Commercial and Business Anecdotes.” Ông dùng nó để mô tả cách Ngài Henry Furnese, một chủ ngân hàng, thu lợi từ thông tin bằng cách thu thập và hành động dựa trên thông tin đó trước đối thủ cạnh tranh. Gần đây hơn, vào năm 1958, một bài báo được viết bởi nhà khoa học máy tính của IBM tên là Hans Peter Luhn, mô tả tiềm năng thu thập thông tin kinh doanh thông minh (BI) thông qua việc sử dụng công nghệ.

Kinh doanh thông minh, như được hiểu ngày nay, sử dụng công nghệ để thu thập và phân tích dữ liệu, chuyển nó thành thông tin hữu ích và hành động dựa trên nó “trước khi cạnh tranh”. Về cơ bản, phiên bản hiện đại của BI tập trung vào công nghệ như một cách để đưa ra quyết định nhanh chóng và hiệu quả, dựa trên thông tin phù hợp vào đúng thời điểm.

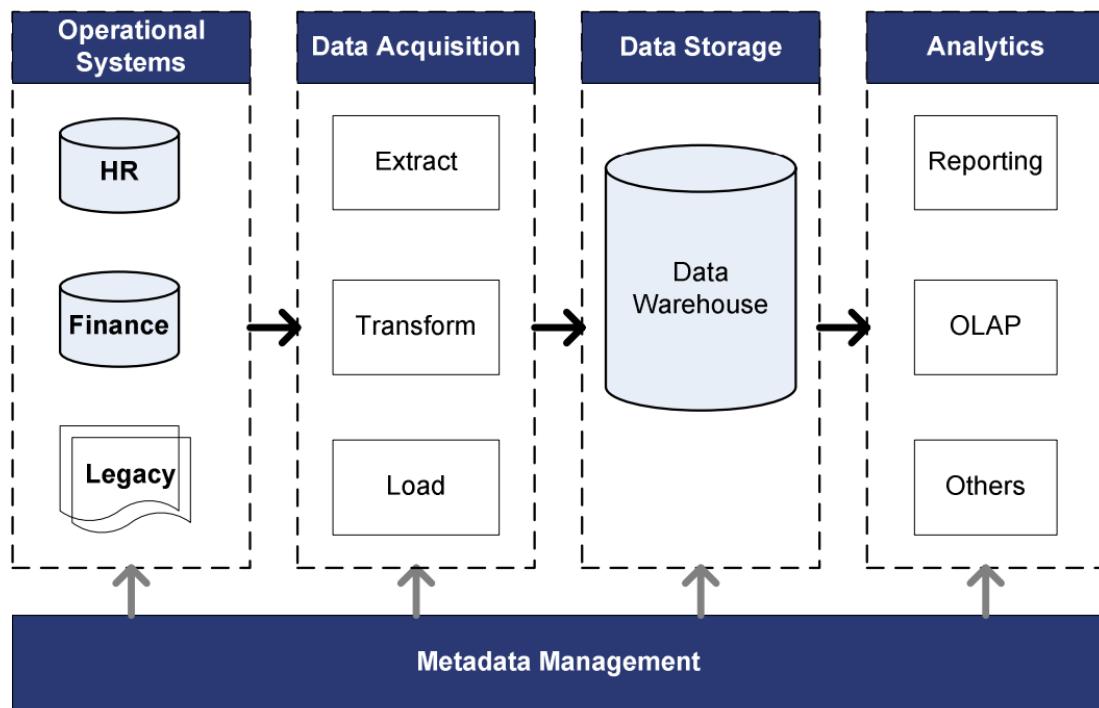
Vào năm 1968, chỉ những cá nhân có kỹ năng cực kỳ chuyên môn mới có thể chuyển dữ liệu thành thông tin có thể sử dụng được. Vào thời điểm này, dữ liệu từ nhiều nguồn thường được lưu trữ trong các kho riêng biệt và nghiên cứu thường được trình bày dưới dạng một báo cáo rời rạc, phân mảnh và có thể giải thích được. Edgar Codd nhận ra đây là một vấn đề và xuất bản một bài báo vào năm 1970, thay đổi cách mọi người nghĩ về cơ sở dữ liệu. Đề xuất của ông về việc phát triển một “mô hình cơ sở dữ liệu quan hệ” đã trở nên vô cùng phổ biến và được áp dụng trên toàn thế giới.

Hệ thống hỗ trợ quyết định (Decision support systems(DSS)) là hệ thống quản lý cơ sở dữ liệu đầu tiên được phát triển. Nhiều nhà sử học cho rằng phiên bản kinh doanh thông minh hiện đại được phát triển từ cơ sở dữ liệu DSS. Số lượng nhà cung cấp BI tăng lên vào những năm 1980, khi các doanh nhân phát hiện ra giá trị của kinh doanh thông minh. Một loạt các công cụ đã được phát triển trong thời gian này để truy cập và sắp xếp dữ liệu theo những cách đơn giản hơn. OLAP, hệ thống thông tin điều hành và kho dữ liệu là một số công cụ được phát triển để hoạt động với DSS.

2.2 Các thành phần của hệ thống BI

Một hệ thống BI điển hình bao gồm bốn cấp độ thành phần và mô-đun quản lý siêu dữ liệu. Kiến trúc chung của các hệ thống BI truyền thống được hiển thị bên dưới. Các thành phần khác

nhanh này hợp tác với nhau để tạo điều kiện thuận lợi cho các chức năng BI cơ bản: trích xuất dữ liệu từ hệ thống vận hành của công ty, lưu trữ dữ liệu đã trích xuất trong kho dữ liệu trung tâm và truy xuất dữ liệu đã lưu trữ cho các ứng dụng phân tích kinh doanh khác nhau [2].



Hình 2.2: Kiến trúc chung của hệ thống kinh doanh thông minh

- **Cấp độ hệ thống vận hành:**

Là nguồn dữ liệu của hệ thống BI, hệ thống vận hành kinh doanh chủ yếu là hệ thống xử lý giao dịch trực tuyến (OLTP) hỗ trợ hoạt động kinh doanh hàng ngày. Các hệ thống OLTP điển hình là hệ thống xử lý đơn đặt hàng của khách hàng, hệ thống tài chính và hệ thống quản lý nguồn nhân lực.

- **Cấp độ thu thập dữ liệu:**

Cấp độ này là thành phần tiền xử lý dữ liệu bao gồm ba giai đoạn: trích xuất, chuyển đổi và tải (ETL). Một công ty thường có các hệ thống OLTP khác nhau tạo ra lượng dữ liệu khổng lồ. Dữ liệu này trước tiên được trích xuất từ các hệ thống OLTP bằng quy trình ETL và sau đó được chuyển đổi theo một bộ quy tắc chuyển đổi. Dữ liệu được chuyển đổi sẽ sạch sẽ, thống nhất và tổng hợp và cuối cùng được tải vào kho dữ liệu trung tâm. ETL là thành phần cơ bản nhất của hệ thống BI vì chất lượng dữ liệu của tất cả các thành phần khác chủ yếu dựa vào quy trình ETL. Trong thiết kế và phát triển ETL, chất lượng dữ liệu, tính linh hoạt của hệ thống và tốc độ xử lý là những mối quan tâm chính.

- **Mức lưu trữ dữ liệu:**

Dữ liệu được xử lý bởi thành phần ETL được lưu trữ trong kho dữ liệu được triển khai chủ yếu bằng hệ thống quản lý cơ sở dữ liệu quan hệ truyền thống (RDBMS). RDBMS được thiết kế để hỗ trợ xử lý giao dịch. Ngược lại, kho dữ liệu là kho lưu trữ dữ liệu tích hợp (Integrated), có định hướng chủ đề (Subject Oriented), thay đổi theo thời gian (Time-variant) và không biến động (Non-volatile)(Inmon 1993). Dữ liệu từ hệ thống OLTP của

công ty được trích xuất, chuyển đổi và tải vào kho dữ liệu dựa trên các lược đồ được xác định trước. Lược đồ hình sao (Star Schema) và lược đồ bông tuyết(Snowflake Schema) là những lược đồ kho dữ liệu phổ biến nhất. Cho dù kho dữ liệu được thiết kế theo loại lược đồ nào thì kho dữ liệu luôn bao gồm hai loại bảng cơ bản: bảng Fact và bảng Dimension.

- **Cấp độ phân tích:**

Dựa trên kho dữ liệu, nhiều loại ứng dụng phân tích khác nhau được phát triển, đại diện cho cấp độ cuối cùng: Phân tích. Hệ thống BI hỗ trợ hai loại chức năng phân tích cơ bản: báo cáo và xử lý phân tích trực tuyến (OLAP). Chức năng báo cáo cung cấp cho người quản lý các báo cáo kinh doanh khác nhau, chẳng hạn như báo cáo bán hàng, báo cáo sản phẩm và báo cáo nhân sự. Báo cáo được tạo bằng cách thực hiện các truy vấn vào kho dữ liệu (DW). Các truy vấn DW chủ yếu là các câu truy vấn được xác định trước do các nhà phát triển DW lập trình. Do đó, các báo cáo do hệ thống BI tạo ra thường có định dạng tĩnh và chứa các loại dữ liệu cố định.

Phân tích BI hứa hẹn nhất là OLAP. OLAP cho phép người quản lý duyệt hiệu quả dữ liệu kinh doanh của họ từ các khía cạnh phân tích khác nhau thông qua các hoạt động Slicing, Dicing và Drilling theo ý muốn. Phân tích Dimension là một góc độ trong đó dữ liệu được trình bày, ví dụ: loại sản phẩm, địa điểm bán hàng, thời gian và khách hàng. So với chức năng báo cáo, OLAP hỗ trợ phân tích dữ liệu đặc biệt, tức là người quản lý có toàn quyền kiểm soát dữ liệu bằng cách chọn các dimension phân tích khác nhau mà họ quan tâm. OLAP dựa trên các mô hình dữ liệu đa chiều (được gọi là lược đồ bông tuyết và ngôi sao).

Ngoài báo cáo và OLAP, còn có nhiều loại ứng dụng phân tích khác có thể được xây dựng trên cơ sở hệ thống DW, chẳng hạn như khai thác dữ liệu, dashboard, quản lý quan hệ khách hàng và quản lý hiệu quả kinh doanh. Về mặt kỹ thuật, các ứng dụng này không nhất thiết phải được xây dựng trên kho dữ liệu. Tuy nhiên, việc tích hợp chúng với các hệ thống DW đã trở thành một thông lệ trong nhiều hệ thống BI thực tế.

- **Quản lý siêu dữ liệu:**

Siêu dữ liệu là dữ liệu đặc biệt về các dữ liệu khác như nguồn dữ liệu, lưu trữ kho dữ liệu, quy tắc kinh doanh, ủy quyền truy cập cũng như cách trích xuất và chuyển đổi các dữ liệu khác nhau. Siêu dữ liệu rất quan trọng để tạo ra thông tin chính xác, nhất quán và bảo trì hệ thống. Nó ảnh hưởng đến toàn bộ quá trình thiết kế, phát triển, thử nghiệm, triển khai và sử dụng hệ thống BI (Caserta 2004; Inmon 2002).

2.3 Ứng dụng của BI

BI có thể được ứng dụng trong nhiều lĩnh vực kinh doanh khác nhau, bao gồm:

- **Marketing:** BI giúp doanh nghiệp phân tích hành vi khách hàng, từ đó đưa ra các chiến lược marketing hiệu quả hơn.
- **Bán hàng:** BI giúp doanh nghiệp phân tích hiệu quả hoạt động bán hàng, từ đó tối ưu hóa quy trình bán hàng.
- **Chăm sóc khách hàng:** BI giúp doanh nghiệp phân tích phản hồi của khách hàng, từ đó nâng cao chất lượng dịch vụ khách hàng.

- **Sản xuất:** BI giúp doanh nghiệp tối ưu hóa quy trình sản xuất, từ đó giảm chi phí và tăng năng suất.
- **Tài chính:** BI giúp doanh nghiệp phân tích tình hình tài chính, từ đó đưa ra các quyết định đầu tư hiệu quả hơn, hỗ trợ quyết định.
- **Kinh doanh:** Phân tích dữ liệu về tương tác của khách hàng với các kênh tiếp thị như email, mạng xã hội, và trang web, đánh giá hiệu quả của các chiến dịch tiếp thị, từ đó tối ưu hóa chiến lược tiếp thị.

2.4 Trực quan hóa dữ liệu với PowerBI

2.4.1 Tổng quan về PowerBI

Giới thiệu chung

Power BI là một công cụ phân tích dữ liệu và trực quan hóa dữ liệu của Microsoft. Công cụ này là tập hợp các ứng dụng, dịch vụ phần mềm và trình kết nối kết hợp với nhau để biến các tệp dữ liệu rời rạc thành thông tin chuyên sâu mang tính tương tác và ấn tượng về mặt trực quan. Power BI có thể hoạt động với các nguồn dữ liệu đơn giản như Microsoft Excel và các nguồn dữ liệu phức tạp như cơ sở dữ liệu trên Cloud, MySQL... Power BI có khả năng dễ dàng kết nối với các nguồn dữ liệu, trực quan hóa, chia sẻ và publish các phân tích của bạn để bất kỳ ai cũng có thể truy cập. Power BI có thể kết nối với file Excel hoặc cơ sở dữ liệu cục bộ một cách dễ dàng. Đây cũng là một công cụ mạnh mẽ, phù hợp cho việc phân tích dữ liệu thời gian thực. Điều này có nghĩa là nó có thể được sử dụng trong nhiều môi trường khác nhau, từ công cụ trực quan và báo cáo cá nhân đến công cụ phân tích và ra quyết định đăng sau các dự án nhóm, bộ phận hoặc toàn bộ tập đoàn. Vì Power BI là một sản phẩm của Microsoft và được tích hợp sẵn các kết nối với Excel nên có nhiều chức năng quen thuộc với người dùng Excel.



Hình 2.3: Logo Power BI

Tính năng chính

Power BI có một số tính năng chính sau:

- Kết nối với dữ liệu: Power BI cho phép người dùng kết nối với dữ liệu từ nhiều nguồn khác nhau, bao gồm:
 - Cơ sở dữ liệu SQL Server
 - Cơ sở dữ liệu Oracle
 - Cơ sở dữ liệu MySQL
 - Tệp Excel
 - Tệp CSV
 - Ứng dụng Web
 - Dịch vụ đám mây
- Phân tích dữ liệu: Power BI cung cấp một số công cụ phân tích dữ liệu mạnh mẽ, bao gồm:
 - Tạo truy vấn
 - Tạo bảng tổng hợp
 - Tạo biểu đồ và đồ thị
 - Tạo mô hình dữ liệu
- Trực quan hóa dữ liệu: Power BI cung cấp một thư viện các biểu đồ và đồ thị trực quan hóa dữ liệu phong phú. Người dùng có thể tùy chỉnh các biểu đồ và đồ thị này để phù hợp với nhu cầu của mình.
- Chia sẻ dữ liệu: Power BI cho phép người dùng chia sẻ dữ liệu và báo cáo với người khác. Người dùng có thể chia sẻ dữ liệu qua web, qua email hoặc qua ứng dụng di động

Ưu điểm nổi bật

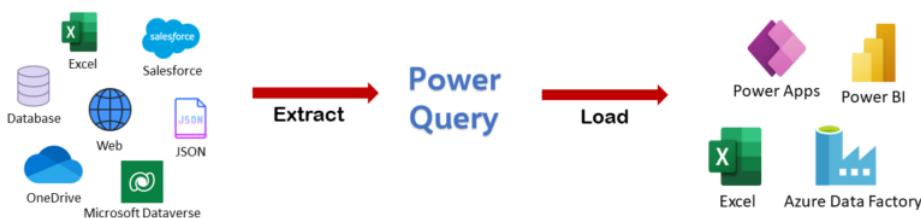
Power BI có một số ưu điểm nổi bật sau:

- Dễ sử dụng: Power BI có giao diện người dùng trực quan và dễ sử dụng. Người dùng không cần phải có kiến thức kỹ thuật cao để sử dụng Power BI.
- Khả năng kết nối với nhiều nguồn dữ liệu: Power BI có thể kết nối với dữ liệu từ nhiều nguồn khác nhau, giúp người dùng dễ dàng truy cập và phân tích dữ liệu.
- Công cụ phân tích dữ liệu mạnh mẽ: Power BI cung cấp một số công cụ phân tích dữ liệu mạnh mẽ, giúp người dùng dễ dàng phân tích dữ liệu phức tạp.
- Tính năng trực quan hóa dữ liệu phong phú: Power BI cung cấp một thư viện các biểu đồ và đồ thị trực quan hóa dữ liệu phong phú, giúp người dùng dễ dàng hiểu và truyền đạt kết quả phân tích dữ liệu.

2.4.2 Xử lý dữ liệu với PowerBI

Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là một bước quan trọng trong quá trình phân tích dữ liệu. Tiền xử lý dữ liệu giúp đảm bảo rằng dữ liệu được làm sạch và chuẩn hóa, sẵn sàng cho phân tích. Tiền xử lý dữ liệu gồm 3 bước: trích xuất, chuyển đổi và tải. Quá trình này trích xuất, chuyển đổi và tải dữ liệu từ nhiều nguồn vào kho dữ liệu. Dữ liệu được thu thập tập trung này giúp phân tích sâu hơn và xử lý dữ liệu dễ dàng hơn. Trong Power BI, tiền xử lý dữ liệu được thực hiện trong Power Query Editor. Power Query Editor là một công cụ tích hợp trong Power BI cho phép người dùng kết nối với dữ liệu từ nhiều nguồn khác nhau, chuyển đổi dữ liệu và tải dữ liệu vào Power BI.



Hình 2.4: Tiền xử lý dữ liệu với Power Query

Có một số công cụ và tính năng tiền xử lý dữ liệu có sẵn trong Power Query Editor, có thể kể đến:

- **Tạo truy vấn:** Truy vấn là một tập hợp các hướng dẫn cho Power Query Editor về cách trích xuất dữ liệu từ nguồn dữ liệu. Khi tạo truy vấn, người dùng có thể sử dụng các công cụ và tính năng tiền xử lý dữ liệu để làm sạch và chuẩn hóa dữ liệu.
- **Thay đổi kiểu dữ liệu:** Power Query Editor cho phép người dùng thay đổi kiểu dữ liệu của các cột dữ liệu. Điều này có thể giúp đảm bảo rằng dữ liệu được xử lý chính xác.
- **Xóa các cột:** Power Query Editor cho phép người dùng xóa các cột dữ liệu không cần thiết. Điều này có thể giúp cải thiện hiệu suất phân tích dữ liệu.
- **Trộn và hợp nhất các bảng:** Power Query Editor cho phép người dùng trộn và hợp nhất các bảng dữ liệu. Điều này có thể giúp tạo ra một tập dữ liệu duy nhất cho phân tích.
- **Thêm các cột:** Power Query Editor cho phép người dùng thêm các cột dữ liệu mới. Điều này có thể giúp bổ sung thêm thông tin cho phân tích dữ liệu.
- **Pivot:** Chuyển đổi các hàng thành cột. Điều này rất hữu ích khi bạn muốn tổng hợp dữ liệu theo một tiêu chí cụ thể. Ví dụ, nếu bạn có dữ liệu bán hàng hàng tháng và muốn chuyển nó thành dữ liệu tổng hợp theo năm.
- **Unpivot:** Chuyển đổi các cột thành hàng. Điều này hữu ích khi bạn có dữ liệu trong nhiều cột và muốn hợp nhất chúng thành một cột với các giá trị tương ứng.
- **Power Query Editor** cho phép bạn chạy các script R và Python để thực hiện các phân tích và biến đổi dữ liệu nâng cao mà không có sẵn qua giao diện người dùng. Điều này rất hữu ích cho các nhà phân tích dữ liệu và nhà khoa học dữ liệu muốn áp dụng các kỹ thuật phức tạp hơn.
- **Data Profiling:** Hiển thị thông tin thống kê về các cột dữ liệu như số lượng giá trị duy nhất, giá trị rỗng, phân phối giá trị, giúp bạn kiểm tra và đảm bảo chất lượng dữ liệu.

Mô hình hóa dữ liệu

Về bản chất, mô hình hóa dữ liệu là quá trình xác định cấu trúc dữ liệu, các thuộc tính và mối quan hệ trong một mô hình dữ liệu. Mô hình dữ liệu trong Power BI là một biểu diễn logic của cách dữ liệu được cấu trúc và liên kết trong công cụ này. Nó là một tập hợp các bảng và các mối quan hệ giữa chúng được sử dụng để tạo báo cáo và trực quan hóa. Một mô hình dữ liệu thường bao gồm một hoặc nhiều nguồn dữ liệu, có thể là từ bảng tính Excel cho đến các cơ sở dữ liệu trên nền tảng đám mây và một hoặc nhiều bảng đại diện cho dữ liệu trong các nguồn đó. Các mối quan hệ kết nối các bảng này là nền tảng của mô hình hóa dữ liệu. Có ba khái niệm mô hình hóa dữ liệu: mô hình hóa dữ liệu khái niệm, mô hình hóa dữ liệu logic và mô hình hóa dữ liệu vật lý. Từ trừu tượng đến cụ thể, các khái niệm mô hình hóa dữ liệu tạo ra một bản thiết kế cho cách dữ liệu được tổ chức và quản lý trong một tổ chức. Sự cần thiết của mô hình hóa dữ liệu:

- **Hỗ trợ khám phá dữ liệu:** Mô hình hóa dữ liệu cho phép người dùng tạo các cấp bậc và các đường đi phân tích, điều này hỗ trợ việc khám phá dữ liệu hiệu quả. Người dùng có thể nhanh chóng điều hướng qua dữ liệu để phát hiện các thông tin chi tiết và nhận diện xu hướng.
- **Ảnh hưởng đến hiệu suất:** Cách thiết kế mô hình dữ liệu ảnh hưởng trực tiếp đến tốc độ và hiệu quả của việc truy xuất dữ liệu. Các mô hình dữ liệu thiết kế kém, như những mô hình có mối quan hệ phức tạp và sự dư thừa dữ liệu, có thể dẫn đến hiệu suất truy vấn chậm, gây ra sự chậm trễ trong việc hiển thị báo cáo.
- **Thúc đẩy báo cáo chính xác hơn:** Khi mô hình dữ liệu được thiết kế đúng cách, nó có thể đảm bảo độ chính xác, sự nhất quán và độ tin cậy của dữ liệu sử dụng trong báo cáo. Điều này có thể mang lại những thông tin chi tiết chính xác hơn và cải thiện quyết định.
- **Dễ bảo trì:** Một mô hình dữ liệu được thiết kế tốt đảm bảo rằng các báo cáo có khả năng mở rộng, điều này có thể giảm bớt nỗ lực cần thiết để duy trì báo cáo khi doanh nghiệp phát triển và thay đổi. Việc tạo các thành phần có thể tái sử dụng, cải thiện tài liệu và chuẩn hóa dữ liệu cũng có thể đảm bảo rằng các báo cáo dễ dàng bảo trì và cập nhật theo thời gian.

Các khái niệm mô hình hóa dữ liệu:

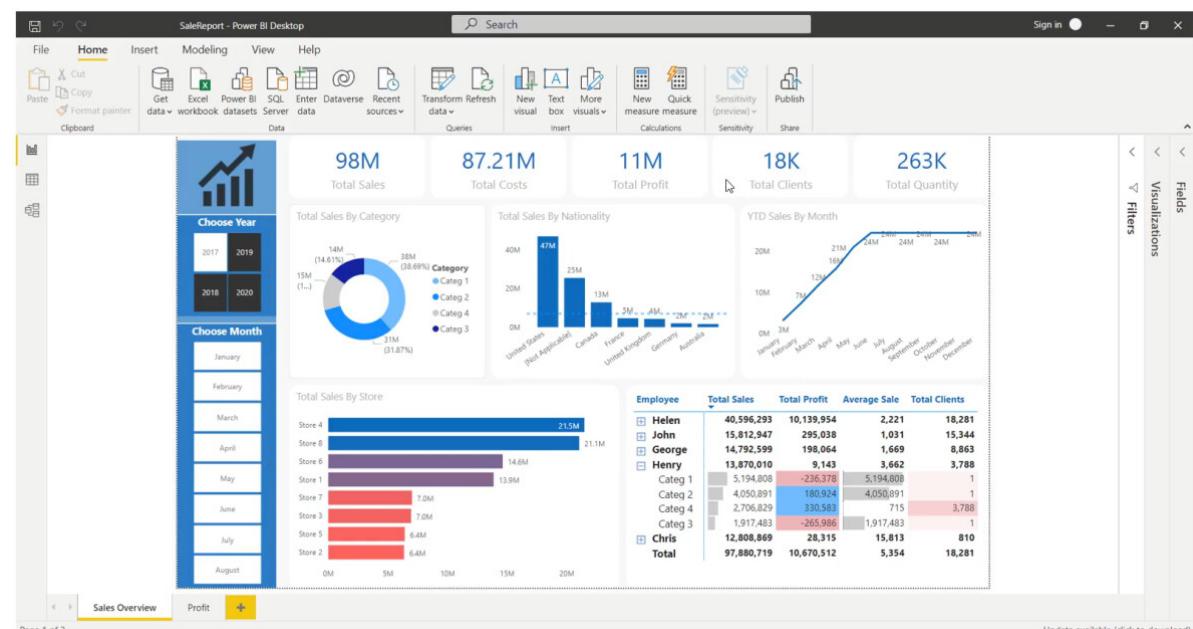
- **Mô hình dữ liệu khái niệm:** Thể hiện cách thế giới kinh doanh nhìn nhận thông tin mà nó sử dụng, tập trung vào bức tranh tổng thể và các mục tiêu chiến lược quan trọng. Nó không cụ thể mà mang tính chung chung, chỉ nắm bắt mức độ sử dụng kinh doanh cao nhất. Mô hình này đi kèm với sơ đồ phân cấp hoạt động hoặc phân rã chức năng để thể hiện các chức năng chính trong khu vực vấn đề kinh doanh.
- **Mô hình dữ liệu logic:** Mô hình dữ liệu logic là diễn giải có cấu trúc hơn của mô hình kinh doanh khái niệm, dùng làm cơ chế giao tiếp trong môi trường kỹ thuật. Nó tập trung vào chi tiết các thực thể và mối quan hệ của chúng. Các bước chính bao gồm:
 1. Đưa mô hình khái niệm vào dạng chuẩn thực thể.
 2. Giải quyết mối quan hệ nhiều-nhiều bằng cách xác định thực thể liên kết.
 3. Xác định các mục dữ liệu duy nhất nhận diện thực thể (Candidate Identifiers).
 4. Lựa chọn các định danh chính từ danh sách các ứng cử viên (Primary Identifiers).

5. Gán thuộc tính cho các thực thể đã được xác định và khóa.
6. Chuyển các khóa chính xuống các thực thể phụ thuộc như khóa ngoại.
7. Định nghĩa các miền của tất cả các thuộc tính và các quy tắc ràng buộc.

- **Mô hình dữ liệu vật lý:** Mô hình vật lý là diễn giải chi tiết hơn và có cấu trúc của mô hình logic, mô tả những thay đổi để phù hợp với môi trường mục tiêu. Nó đi kèm với các sơ đồ đường dẫn truy cập, hiển thị chi tiết các đường dẫn sẽ được thực hiện qua mô hình khi mỗi quy trình được thực hiện. Mô hình này đảm bảo hiệu suất tối đa trong môi trường mục tiêu.
- **Mô hình dữ liệu chiều:** Mô hình dữ liệu chiều là sự triển khai vật lý của cấu trúc quan hệ thực thể đã được chuẩn hóa. Chúng thường được sử dụng trong các kho dữ liệu và các kho dữ liệu nhỏ, được xem như phần của các cơ sở dữ liệu chuyên dụng.

Trực quan hóa dữ liệu

Sau khi tạo mô hình dữ liệu, bạn có thể tạo báo cáo và dashboard. Báo cáo và dashboard là các công cụ trực quan hóa dữ liệu giúp bạn dễ dàng hiểu và truyền đạt kết quả phân tích dữ liệu.



Hình 2.5: Tạo dashboard để trực quan hóa dữ liệu

Ứng dụng của DW và BI vào bài toán Electronics Commerce

3.1 Giới thiệu về bài toán (Requirement)

Vấn đề khảo sát

Khảo sát về hoạt động giao dịch trong thương mại điện tử thời trang (Fashion e-commerce transactional activity) thường tập trung vào việc phân tích và đánh giá các mẫu hành vi mua sắm và giao dịch của người tiêu dùng trong lĩnh vực này. Đây là một lĩnh vực nghiên cứu quan trọng trong ngành thương mại điện tử, nhằm hiểu rõ hơn về cách mà các công ty thời trang và bán lẻ sử dụng các nền tảng điện tử để thu hút khách hàng và tăng cường doanh số bán hàng.

Mục đích khảo sát

- Hiểu thị trường: Điều này bao gồm việc nắm bắt xu hướng mua sắm trực tuyến trong ngành thời trang, những sản phẩm nào đang được ưa chuộng, và cách mà người tiêu dùng tương tác và mua hàng trực tuyến.
- Phân tích cạnh tranh: Điều này giúp các doanh nghiệp hiểu rõ về đối thủ cạnh tranh của mình trên thị trường trực tuyến, từ đó đề xuất những chiến lược cạnh tranh hiệu quả hơn.
- Đánh giá sản phẩm: Khảo sát có thể giúp đánh giá độ phổ biến của các sản phẩm thời trang trên các nền tảng thương mại điện tử, từ đó cải thiện hoặc điều chỉnh các sản phẩm để phù hợp với nhu cầu của thị trường.
- Tìm kiếm ý kiến khách hàng: Cung cấp một cơ hội để thu thập ý kiến phản hồi từ khách hàng về trải nghiệm mua sắm trực tuyến của họ, giúp cải thiện dịch vụ và sản phẩm.
- Dự đoán xu hướng tương lai: Phân tích dữ liệu từ hoạt động giao dịch trực tuyến có thể giúp dự đoán xu hướng tiêu dùng trong tương lai, từ đó giúp các doanh nghiệp thời trang chuẩn bị cho những thay đổi trong thị trường.

Nội dung khảo sát

Các mặt nghiên cứu trong khảo sát này bao gồm:

- Phân tích xu hướng mua sắm trực tuyến: điều này có thể bao gồm các yếu tố như sự tăng trưởng của việc mua sắm qua các thiết bị di động, thói quen mua sắm theo mùa, và các yếu tố ảnh hưởng đến quyết định mua sắm của người tiêu dùng.

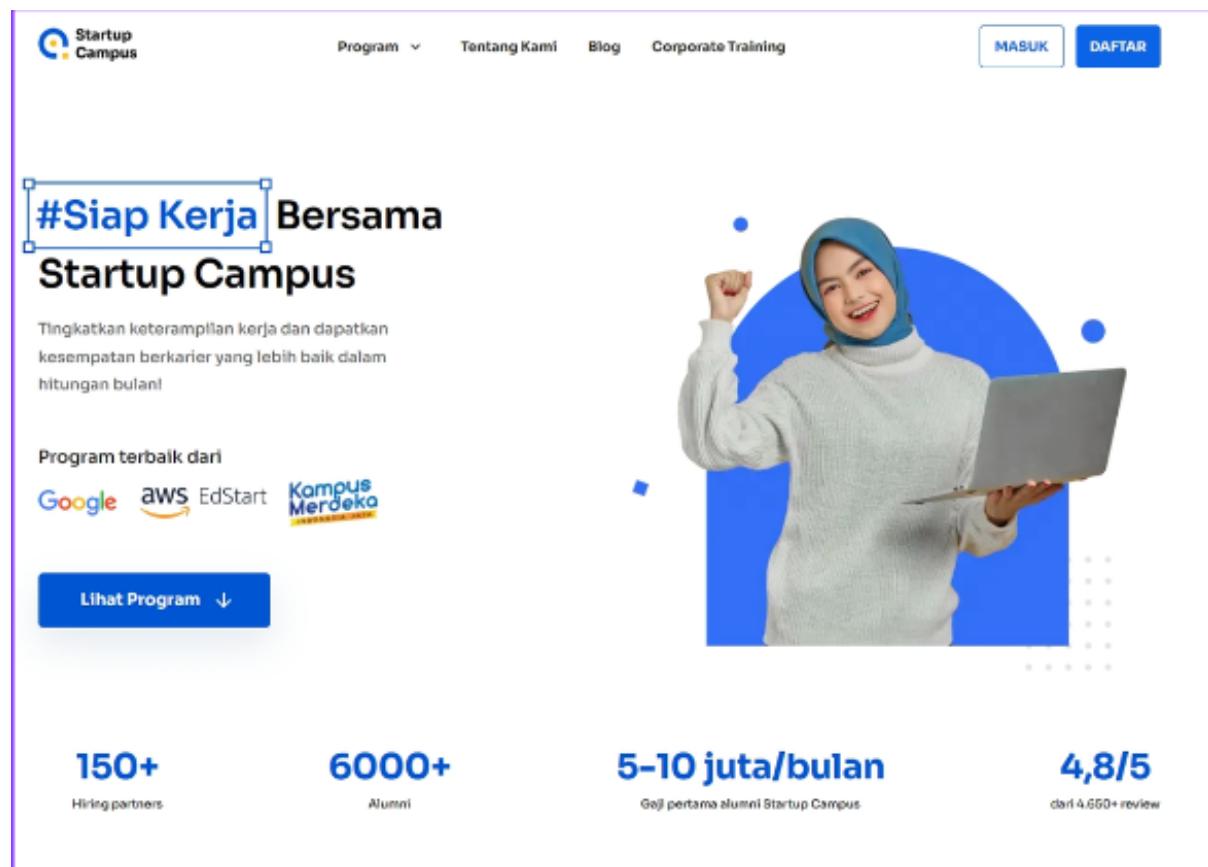
- Đánh giá các chiến lược marketing và khuyến mãi: bao gồm các chiến lược quảng cáo, chương trình khuyến mãi, ảnh hưởng của các sự kiện lớn.
- Phân tích hành vi người dùng trên các nền tảng mạng xã hội: như là sự ảnh hưởng của influencer, các chiến lược chia sẻ, bình luận, và đánh giá.
- Tìm hiểu các đối thủ tiềm năng, phân tích và đưa ra được nhu cầu thiết yếu của thị trường. Từ đó xây dựng các kịch bản để ứng phó với thực tế.

Thị trường thương mại điện tử thời trang lớn nhất được tìm thấy ở châu Á, tiếp theo là Bắc Mỹ. Ở châu Á, doanh thu thời trang trực tuyến dự kiến sẽ đạt hơn 680 tỷ đô la Mỹ vào năm 2027, gần gấp đôi doanh thu dự kiến của thương mại điện tử thời trang ở Bắc Mỹ trong cùng năm đó. Trong tổng doanh số bán lẻ thời trang trên toàn thế giới, gần 21% là giao dịch thương mại điện tử. Ở các khu vực khác như Châu Âu, Châu Đại Dương hoặc Châu Mỹ nói chung, khoảng 30% doanh số bán lẻ thời trang được tạo ra thông qua các kênh trực tuyến. Doanh số bán hàng thời trang ở Châu Phi chủ yếu diễn ra ngoại tuyến, chỉ có 6,5% được tạo ra thông qua thương mại điện tử. Mặc dù Châu Âu không phải là thị trường lớn nhất về thời trang trực tuyến, trong số các công ty thời trang thương mại điện tử trên toàn thế giới có vốn hóa thị trường cao nhất là các nhà bán lẻ thời trang trực tuyến lớn nhất Châu Âu, Zalando, ASOS và About You. Tuy nhiên, với việc châu Á là thị trường thương mại điện tử thời trang lớn nhất toàn cầu, doanh thu thuần của các cửa hàng trực tuyến hàng đầu trong phân khúc thời trang phần lớn đến từ các công ty có trụ sở chính ở lục địa phía Đông. Ngoài ra, trang web thời trang nhanh cực kỳ nổi tiếng của Trung Quốc, shein.com, là trang web thời trang được truy cập nhiều nhất trên toàn thế giới vào tháng 5 năm 2022.

Tổng quan về Startup Campus

Startup Campus là website chuyên cung cấp các khóa học bootcamp trực tuyến với sự hướng dẫn chuyên sâu và dự án thực tế. Startup Campus được thành lập vào năm 2021, được hỗ trợ hoàn toàn bởi Kemendikbud Ristekdikti dưới sự bảo trợ của Kampus Merdeka và được vinh danh là một trong những chương trình Studi Independen tốt nhất trong chương trình MSIB2. Họ cung cấp các khóa học trong nhiều lĩnh vực khác nhau như Data Science, UI/UX Design, và Artificial Intelligence.

Năm 2022, StartupCampus đã được Amazon Web Services (AWS) vinh danh là một trong những công ty khởi nghiệp sáng tạo nhất ở Châu Á Thái Bình Dương. Startup Campus đã hợp tác với hơn 118 đối tác bao gồm các công ty khởi nghiệp, tổ chức giáo dục, công nghệ, dịch vụ và tổ chức xã hội để đào tạo ra những tài năng làm việc trong các công ty, tập đoàn,.....



Hình 3.1: Giao diện trang chủ Startup Campus



Hình 3.2: Canvas Model

Nhu cầu phân tích

1. Khách hàng

- Phân tích dữ liệu liên quan đến đối tượng: thông tin chi tiết chuyên sâu về giới tính và nhóm tuổi
- Xu hướng phát triển ở từng khu vực
- Báo cáo về thói quen của từng khách hàng ở từng khu vực, phân khúc: Sản phẩm quan tâm, loại sản phẩm tìm kiếm, ...

2. Doanh số

- Phân tích doanh số dựa trên từng khu vực cùng với loại sản phẩm cụ thể, nhóm hàng, theo thời gian,...

3. Sản phẩm

- Báo cáo về số lượng sản phẩm, loại sản phẩm cùng với đó là phân bổ sản phẩm theo mùa, theo đối tượng

4. Giao dịch đơn hàng

- Báo cáo về đơn hủy, đơn được giao, địa điểm và thời gian

5. Giao Hàng

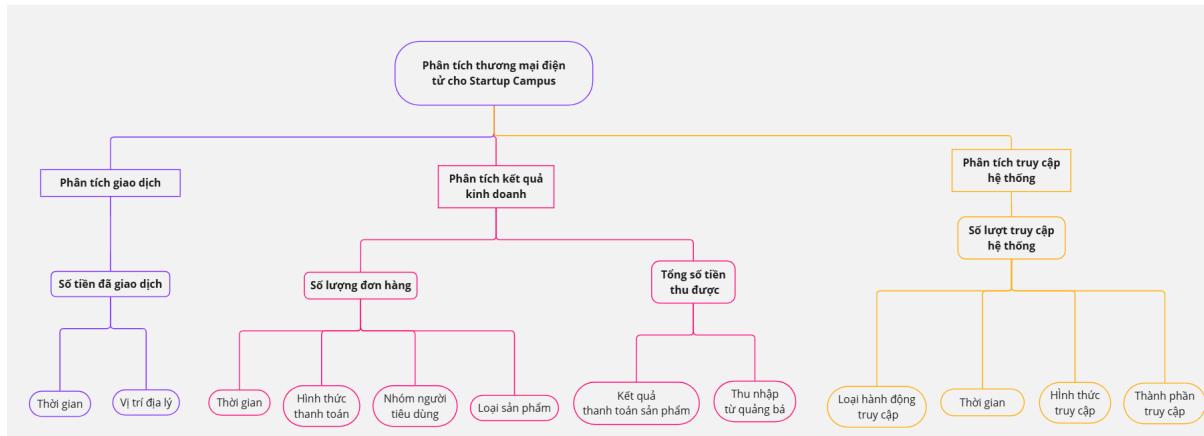
- Báo cáo về thời gian trung bình vận chuyển của từng loại sản phẩm dựa theo khu vực và số lượng

Ta xây dựng được nhu cầu phân tích được minh họa bằng mindmap:



Hình 3.3: Mindmap nhu cầu phân tích

Dựa vào các chủ điểm phân tích, nhóm đã thực hiện xây ra cây phân tích dashboard như sau



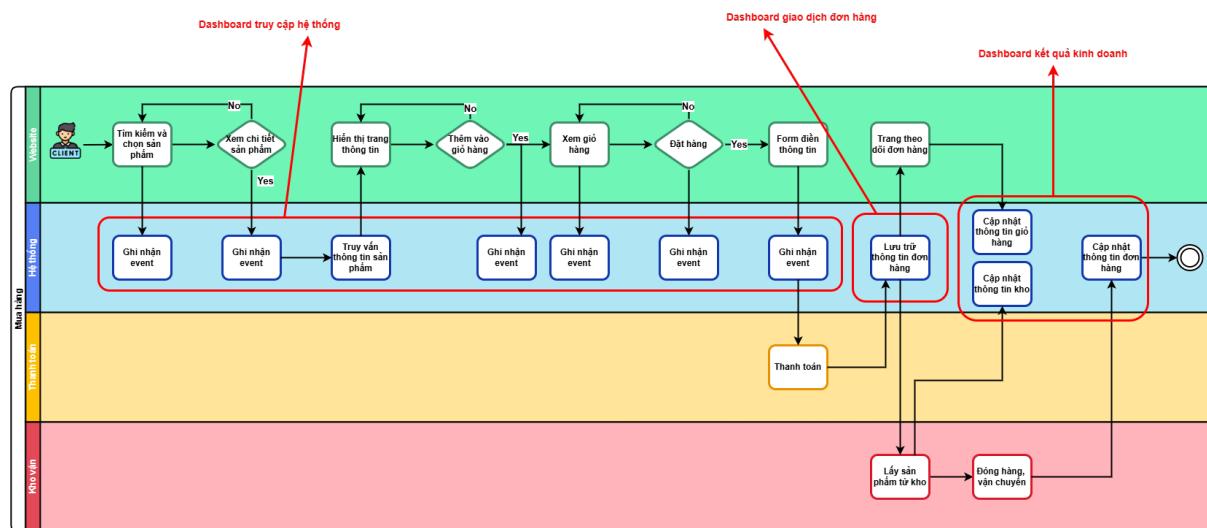
Hình 3.4: Cây phân tích Dashboard

Quy trình nghiệp vụ trên Startup Campus

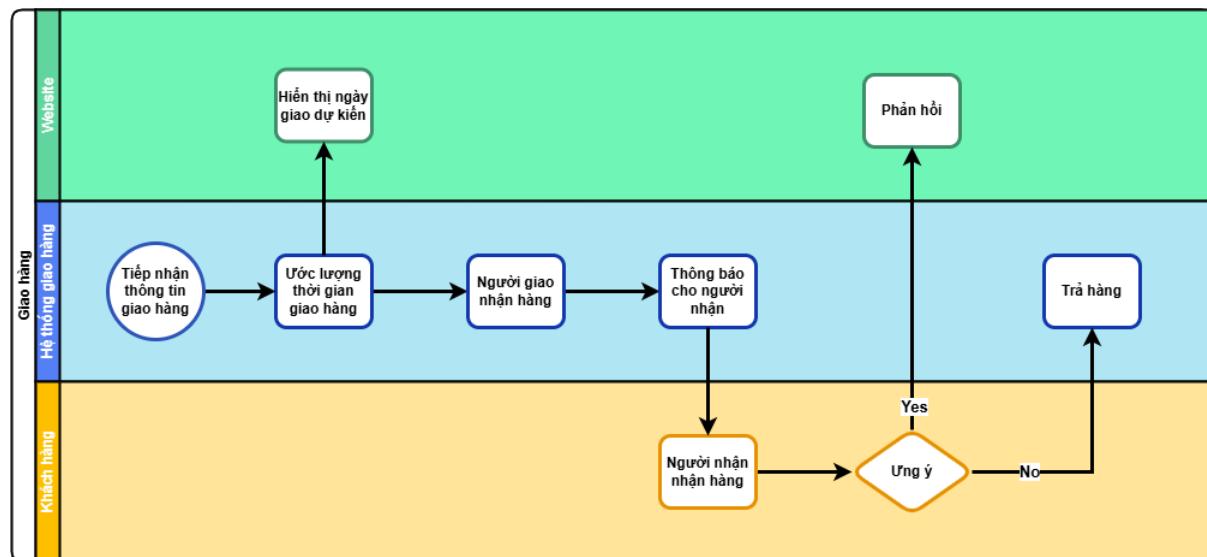
Khi một khách hàng muốn mua sản phẩm trên StartupCampus hay các sàn thương mại điện tử nào khác, họ sẽ quan tâm và cần thực hiện các công việc sau:

- Đăng nhập vào tài khoản cá nhân
- Truy tìm thông tin sản phẩm, tìm hiểu về sản phẩm thông qua các thông tin chi tiết, đánh giá từ khách hàng, Seller,....
- Khách hàng thêm các sản phẩm quan tâm vào giỏ hàng, chuyển đến mục thanh toán

- Nhập các thông tin về địa chỉ giao hàng (nhà ở, số điện thoại liên lạc, tên người nhận,...) và phương thức thanh toán.
- Sau khi đơn hàng được lên thành công, hàng sẽ được vận chuyển đến khách hàng sau một số ngày nhất định thông qua một hoặc nhiều bên vận chuyển thứ ba. Khách hàng có thể theo dõi tình trạng vận chuyển của đơn hàng trên hệ thống
- Sau khi người đặt hàng nhận được hàng, họ sẽ nhận hàng và thực hiện đánh giá sản phẩm.
- Với tất cả các hoạt động của người dùng trên sàn thương mại điện tử, hệ thống sẽ thực hiện ghi lại các thao tác, đồng thời truy vấn đến cơ sở dữ liệu để cho ra các thông tin cần thiết với người dùng



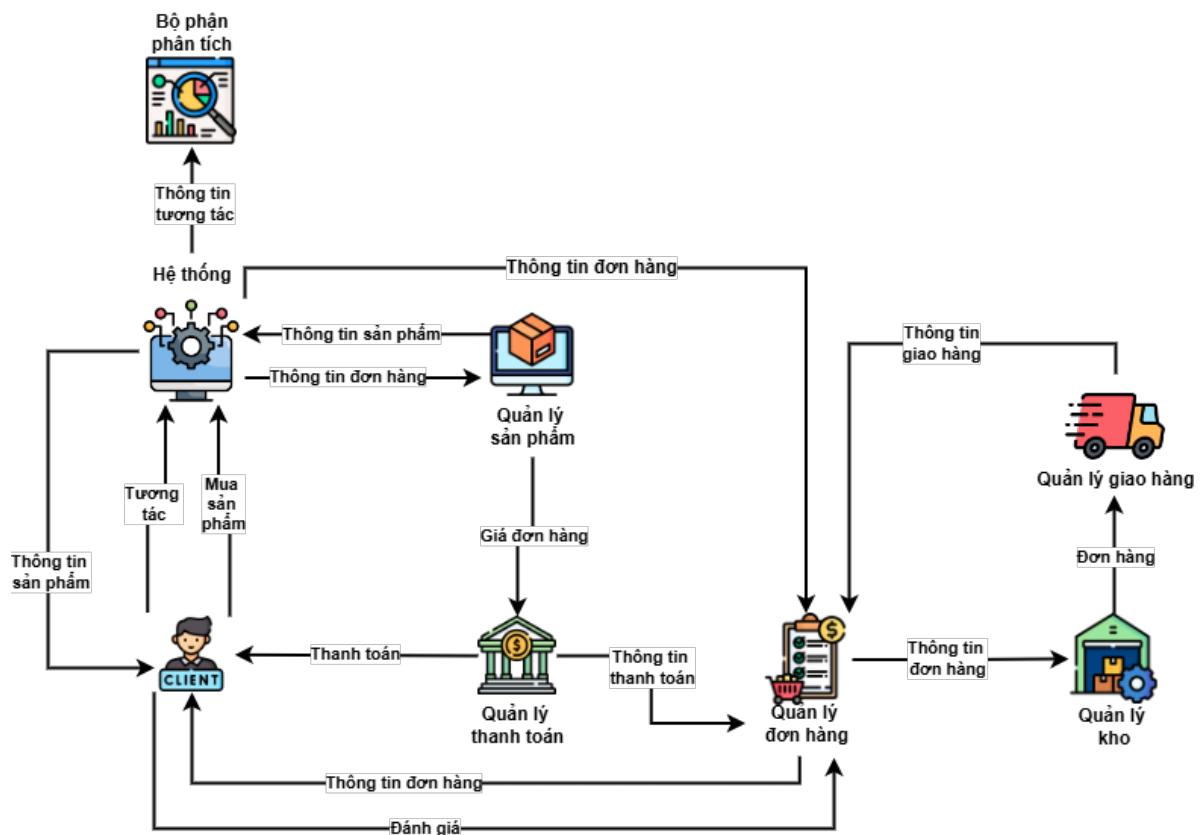
Hình 3.5: Nghiệp vụ đặt hàng kèm nội dung sẽ phân tích trên Dashboard



Hình 3.6: Nghiệp vụ giao hàng

Từ quy trình nghiệp vụ đặt hàng được thể hiện trong sơ đồ luồng nghiệp vụ, ta có thể xác định được các nội dung cần xây dựng trên Dashboard. Cụ thể:

- Các hành động ghi nhận các hoạt động tương tác với hệ thống sẽ cần được phân tích trong **Dashboard truy cập hệ thống**
- Các hành động ghi nhận các đơn hàng được đặt mua từ khách hàngG sẽ cần được phân tích trong **Dashboard giao dịch đơn hàng**
- Các hành động ghi nhận các thông tin về giá trị đơn hàng và đánh giá sẽ cần được phân tích trong **Dashboard kết quả kinh doanh**

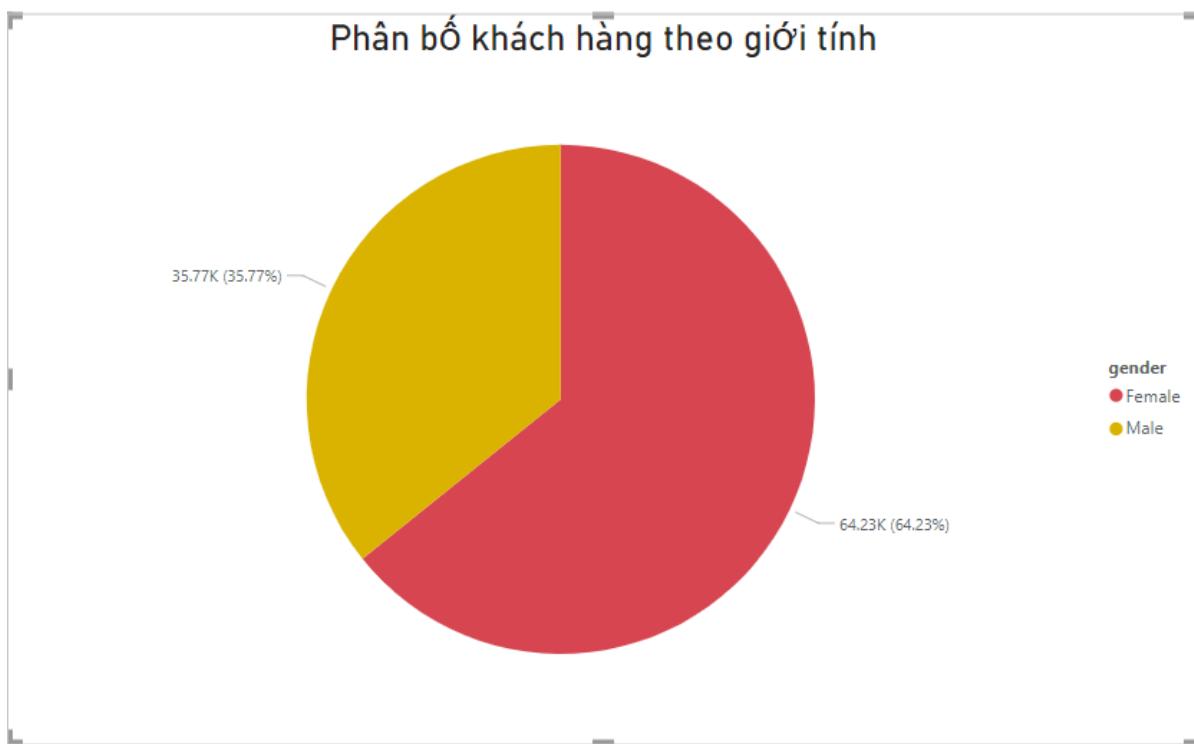


Hình 3.7: Data flow - Sơ đồ luồng dữ liệu

3.2 Giới thiệu về ODS

3.2.1 Data exploration

Phân bố khách hàng theo giới tính



Hình 3.8: Phân bố khách hàng theo giới tính

Từ biểu đồ tròn thấy được rằng phần lớn số lượng khách hàng được cho thấy là nữ chiếm 2/3 số lượng khách hàng cho thấy nhu cầu mua sắm của người dùng tập trung ở nữ.

Phân bố khác hàng theo địa lý

Sử dụng Map Chart ta thu được kết quả



Hình 3.9: Phân bố khác hàng theo giới tính

Từ biểu đồ cho thấy: Lượng khách hàng tập trung phân bố chủ yếu ở các thành phố lớn nơi và những nơi tập trung phân lõn dân số cùng với công nghệ phát triển.

Phân bố khách hàng theo nhóm tuổi

Sử dụng Matrix Chart ta thu được kết quả:

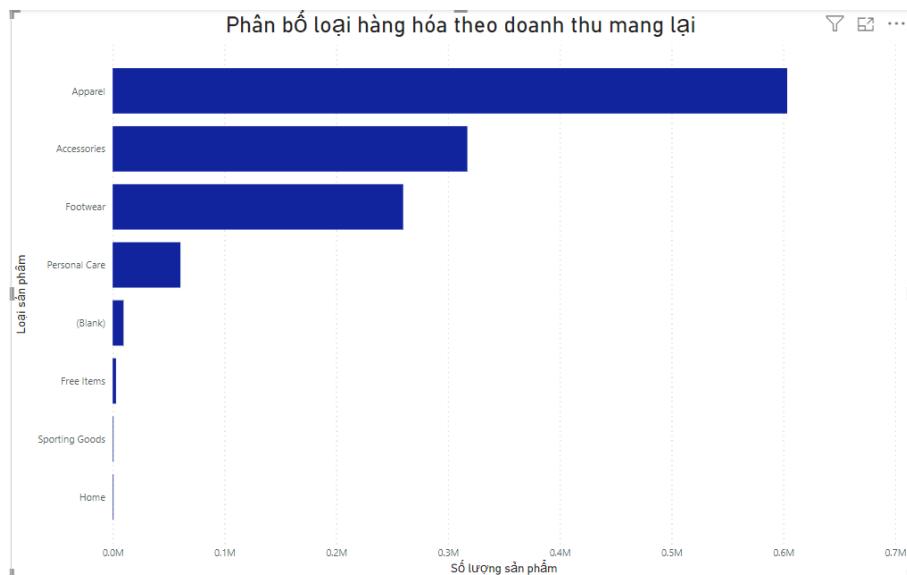


Hình 3.10: Phân bố khác hàng theo nhóm tuổi

Kết quả cho thấy: Cùng với sự phát triển của công nghệ và thế hệ con người mới, rõ ràng có thể thấy qua biểu đồ rằng chiếm phần lớn trong số khách hàng chính là thế hệ trẻ khi mà khả năng tiếp cận và học hỏi nhanh với internet và sự tò mò là vô hạn sau đó mới đến nhóm người trưởng thành.

Phân bố hàng hóa theo khoảng giá trị doanh thu

Sử dụng biểu đồ cột ngang, ta thu được kết quả

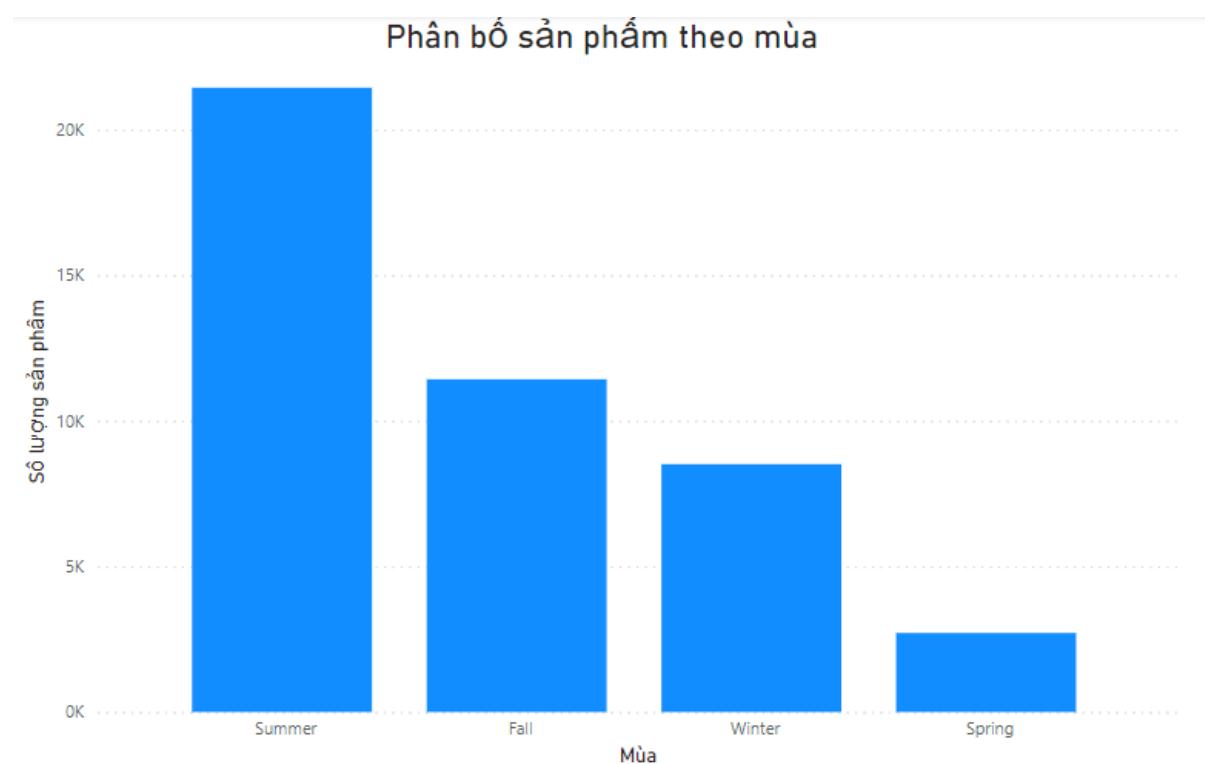


Hình 3.11: Phân bố loại hàng hóa theo doanh thu mang lại

Chiếm phần tiêu thụ nhiều nhất chính là mặt hàng may mặc, nguyên do:

- Nhu cầu tiêu dùng cao: Hàng may mặc là một trong những mặt hàng tiêu dùng cơ bản và thiết yếu của người dân. Vì vậy, nhu cầu mua sắm đồ may mặc luôn cao, đặc biệt là trong một đất nước như Indonesia với dân số đông đúc và một nền kinh tế đang phát triển.
- Dễ dàng sản xuất và kinh doanh: Sản xuất các sản phẩm may mặc có thể được thực hiện với mức độ tự động hóa khác nhau, từ quy mô lớn đến quy mô nhỏ. Điều này làm cho việc khởi đầu kinh doanh trong lĩnh vực này tương đối dễ dàng, đặc biệt là với các nhà sản xuất và thương gia nhỏ. Sau đó là các mặt hàng khác như phụ kiện, giày dép,...

Phân bố sản phẩm theo mùa



Hình 3.12: Phân bố sản phẩm theo mùa

Do Indonesia nằm ở gần vùng xích đạo, nên ở đây chỉ có 2 mùa rõ rệt nhất là mùa mưa và mùa khô. Chính vì vậy khi thời điểm mùa hè và mùa thu ở Indonesia sẽ rơi vào khoảng thời gian mùa khô và cũng là thời gian mà nền kinh tế Indonesia tăng mạnh, chính vì vậy sản phẩm tập trung vào khoảng thời gian này.

Số lượng sản phẩm theo giới tính

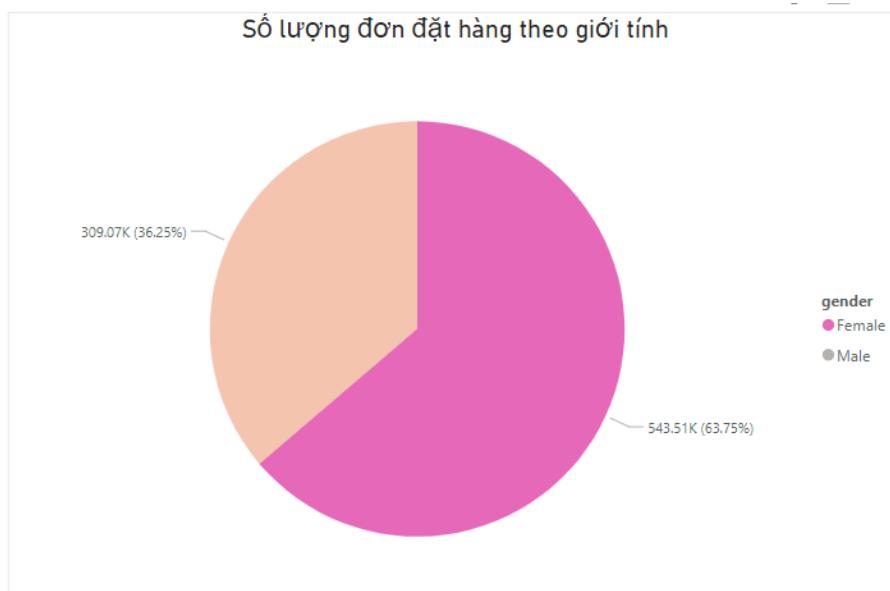
Sử dụng biểu đồ cột ta thu được:



Hình 3.13: Phân bố khác hàng theo giới tính

Dựa vào biểu đồ có thể thấy được rằng các sản phẩm tập trung vào lứa tuổi trẻ và người trưởng thành, do sản phẩm may mặc và các trang thiết bị, phụ kiện là các sản phẩm được mua nhiều nhất.

Phân bố số lượng đơn đặt hàng theo giới tính



Hình 3.14: Phân bố khác hàng theo giới tính

Từ biểu đồ tròn: Có hai nguyên nhân đến việc phần lớn người mua hàng là nữ:

- Khi mà thời điểm hiện nay những người phụ nữ sẽ phụ trách phần lớn việc mua sắm đồ cho gia đình của mình.
- Với việc đa dạng các mẫu mã thời trang, sản phẩm may mặc thì điều đó góp phần lớn việc thu hút các khách hàng là nữ giới, những người yêu thích cái đẹp.

Tương quan phí ship với tổng giá trị đơn hàng và số ngày giao hàng

Sử dụng ma trận hệ số tương quan ta thu được kết quả:

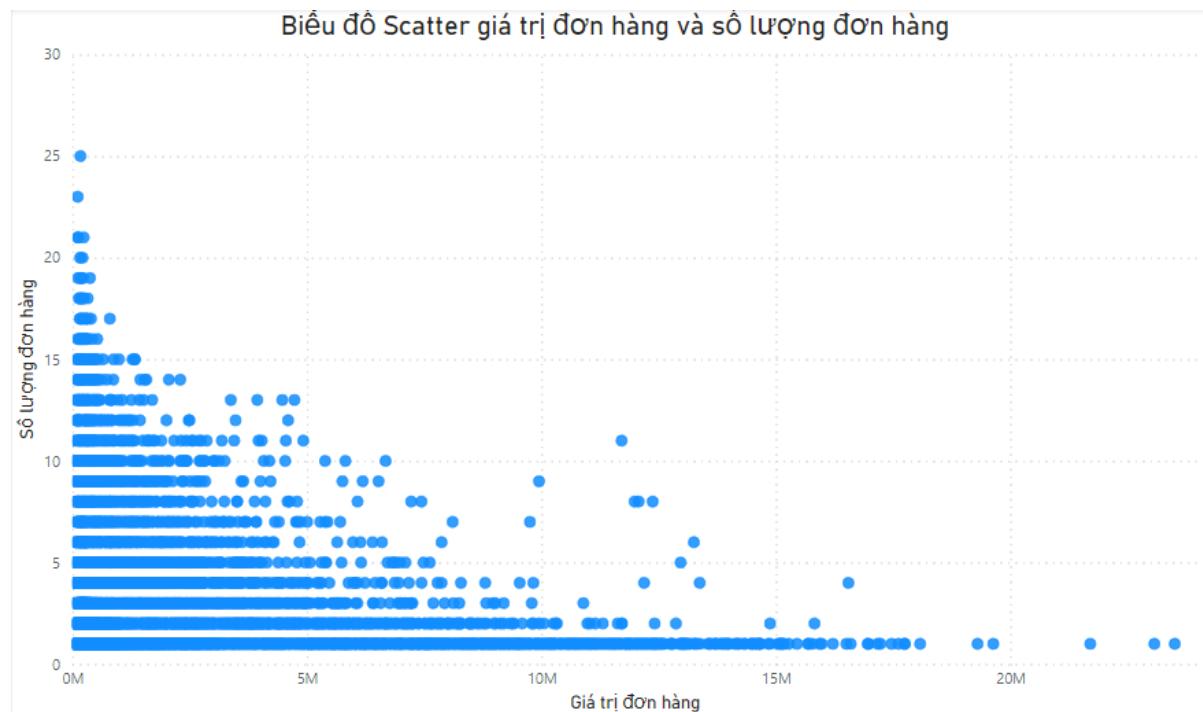
	Phí Ship	Tổng giá trị đơn hàng	Số ngày giao hàng
Phí Ship	1		
Tổng giá trị đơn hàng	0.01231669	1	
Số ngày giao hàng	-0.0026493	-0.000481311	1

Hình 3.15: Tương quan phí ship với tổng giá trị đơn hàng và số ngày giao hàng

Rõ ràng bên cạnh việc công nghệ phát triển mạnh cùng với đó là chất lượng dịch vụ ngày càng cải thiện để thu hút khách hàng thì việc đẩy cao chất lượng giao hàng và rút ngắn thời gian giao hàng gần như được coi là một trong những việc quan trọng nhất đối với thương mại điện tử hiện nay. Chính vì vậy có thể thấy phần lớn các sản phẩm thường có ngày giao hàng ngắn

Tương quan giá trị đơn hàng và số lượng đơn hàng

Sử dụng Scatter Chart ta thu được kết quả:

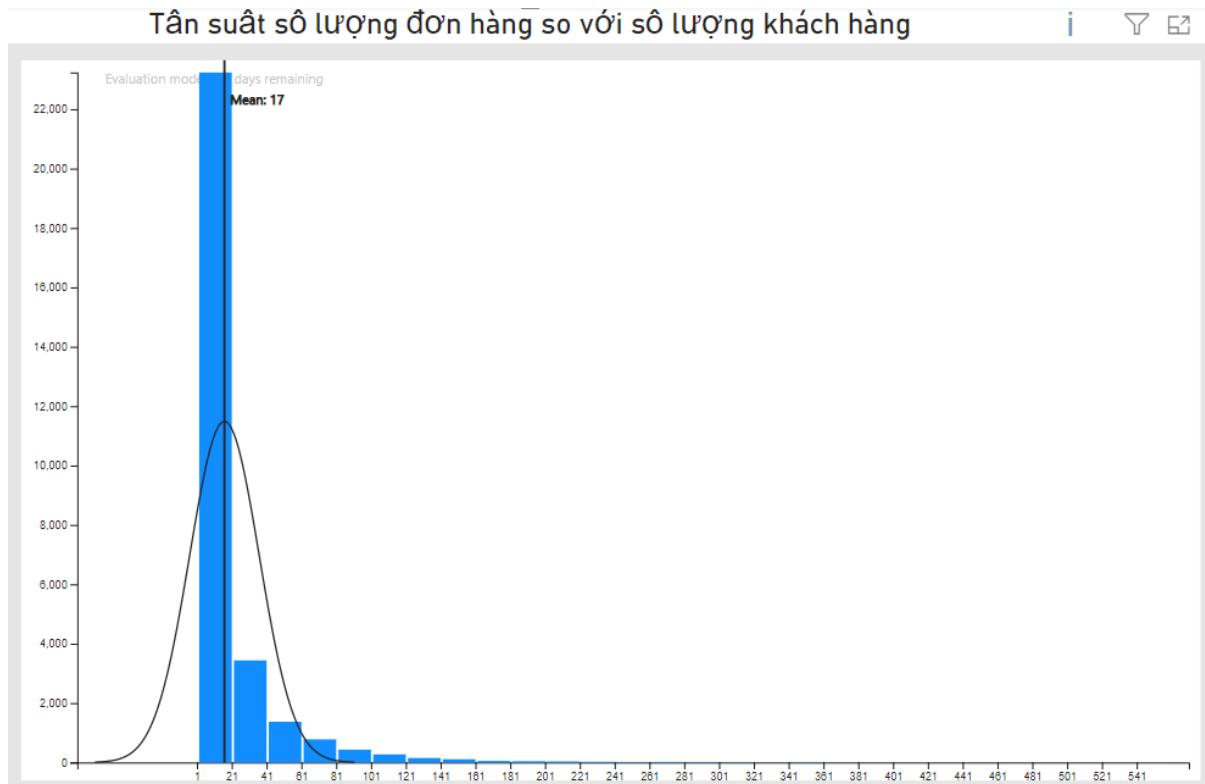


Hình 3.16: Tương quan giá trị đơn hàng và số lượng đơn hàng

Số lượng đơn hàng nhỏ lẻ chiếm phần lớn số lượng đơn hàng do muôn có khả năng cạnh tranh thị trường thì các sàn thương mại điện tử sẽ đánh vào tâm lí người dùng khi tung ra nhiều giảm giá, khuyến mãi và nhiều sản phẩm giá rẻ để thu hút mọi người.

Tần suất số lượng đơn hàng so với số lượng khách hàng

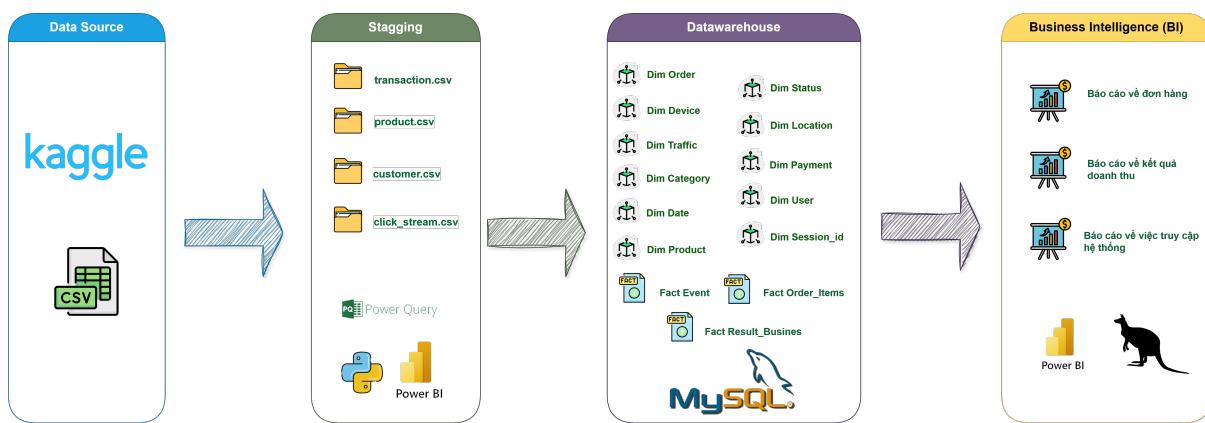
Sử dụng biểu đồ đường cột kết hợp



Hình 3.17: Tần suất số lượng đơn hàng so với số lượng khách hàng

Dựa vào biểu đồ có thể thấy được rằng, rõ ràng những khách hàng chỉ có xu hướng mua đú dùng cho bản thân và gia đình, chưa có quá nhiều sự thu hút để khách hàng quay trở lại mua tiếp nên số lượng đơn hàng không quá lớn.

3.3 Kiến trúc Data Warehouse



Hình 3.18: Kiến trúc datawarehouse

Cụ thể, kiến trúc sẽ gồm 4 phần sau đây:

- Tầng dữ liệu nguồn (Data Source): gồm các file csv được tạo ra từ chương trình thu thập dữ liệu mà nhóm đã nêu ở phần trên.

- Vùng đệm (Staging Zone): Sau khi dữ liệu được thu thập thành công, thông qua quá trình tiền xử lý dữ liệu, ta sẽ thu được 5 file csv mới.
- Tầng kho dữ liệu (Data Warehouse): Thông qua việc tiền xử lý dữ liệu, dữ liệu đã được làm sạch, nhóm sẽ tiến hành phân chia dữ liệu thành các bảng chiều dữ liệu và các bảng chủ điểm để phục vụ cho việc phân tích. Ở đây, nhóm thực hiện chia thành 11 bảng dim và 3 bảng fact được liệt kê trong hình.
- Tầng phân tích kinh doanh: Cuối cùng, nhóm sẽ thực hiện trực quan hóa dữ liệu bằng PowerBI để tiến hành xây dựng các bản báo cáo phân tích. Chúng ta sẽ thực hiện xây dựng và phân tích các Dashboard xoay quanh các bảng fact đã phân tách được ở tầng kho dữ liệu. Dashboard sẽ được trực quan bằng PowerBI và mở rộng trực quan bằng Kangaroo.

3.4 Tiềm xử lý dữ liệu

3.4.1 Khái niệm tiềm xử lý dữ liệu

ETL (Extract, Transform, Load) là một quy trình quan trọng trong việc di chuyển và chuyển đổi dữ liệu từ các nguồn khác nhau vào kho dữ liệu hoặc hệ thống phân tích dữ liệu. ETL giúp chuẩn bị dữ liệu một cách có hệ thống và hiệu quả, làm cho dữ liệu sẵn sàng cho các phân tích và quyết định kinh doanh. Cụ thể trong ETL sẽ có các công việc như sau:

- **Extract (Trích xuất dữ liệu)**
 - * Thu thập dữ liệu từ các nguồn khác nhau như cơ sở dữ liệu, tệp tin, API, v.v.
 - * Ví dụ: Trích xuất dữ liệu khách hàng từ hệ thống CRM và dữ liệu giao dịch từ hệ thống ERP.
- **Transform (Chuyển đổi dữ liệu)**
 1. **Làm sạch dữ liệu (Data Cleaning)**
 - * Loại bỏ hoặc sửa chữa các giá trị thiếu và sai sót.
 - * Chuẩn hóa dữ liệu để đảm bảo tính nhất quán.
 2. **Chuyển đổi định dạng (Data Transformation)**
 - * Chuyển đổi kiểu dữ liệu (ví dụ: từ string sang datetime).
 - * Tổng hợp và tính toán các chỉ số mới.
 3. **Tích hợp dữ liệu (Data Integration)**
 - * Kết hợp dữ liệu từ nhiều nguồn khác nhau.
 - * Xử lý dữ liệu trùng lặp và không nhất quán.
- **Load (Tải dữ liệu)**
 - * Đưa dữ liệu đã chuyển đổi vào hệ thống đích, thường là một kho dữ liệu hoặc cơ sở dữ liệu phân tích.
 - * Phương pháp tải dữ liệu:
 - ❖ *Full Load*: Tải toàn bộ dữ liệu từ nguồn vào hệ thống đích.
 - ❖ *Incremental Load*: Tải chỉ những dữ liệu mới hoặc đã thay đổi từ lần tải trước.

3.4.2 Tóm tắt quá trình tiền xử lý

Với bộ dữ liệu lần này, nhóm đã thực hiện các thao tác như sau:

- Kiểm tra và xóa trùng lặp
- Xóa dữ liệu null
- Xóa các cột không dùng
- Định dạng lại kiểu dữ liệu
- Nhóm dữ liệu
- Sinh dữ liệu mới

3.4.3 Thiết kế Data Pipeline

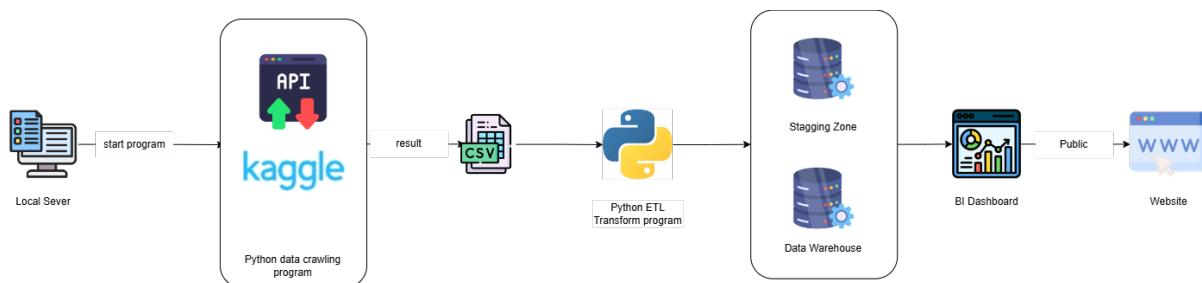
Khái niệm Data Pipeline

Đường ống dữ liệu (data pipeline) được hiểu như dòng chảy của dữ liệu, từ khi là dữ liệu thô cho đến khi lên báo cáo phân tích.

Xử lý đường ống dữ liệu là một phương pháp có cấu trúc để thu thập dữ liệu thô từ nhiều nguồn khác nhau, xử lý thông qua các bước chuyển đổi, và sau đó lưu trữ vào kho dữ liệu như Data Lake hoặc Data Warehouse để phân tích. Quá trình này bao gồm nhiều bước để đảm bảo chất lượng và tích hợp dữ liệu, điều này đặc biệt quan trọng đối với các cơ sở dữ liệu quan hệ có các schema được định nghĩa sẵn. Các pipeline dữ liệu hoạt động như các kênh dẫn cho các dự án khoa học dữ liệu và công cụ kinh doanh thông minh, cho phép dữ liệu được lấy từ các nguồn khác nhau như API, cơ sở dữ liệu và tệp tin. Quá trình này cũng bao gồm việc ghi lại nguồn gốc dữ liệu để theo dõi dữ liệu qua các môi trường và ứng dụng khác nhau.

Data Pipeline của bài toán

Data Pipeline cho dữ liệu lần này được nhóm thiết kế như sau:



Hình 3.19: Data Pipeline

Mô tả thiết kế: Dữ liệu sẽ được thu thập thông qua một chương trình được lập trình bằng Python, kết nối với API của Kaggle và kết nối với đĩa chỉ chứa dữ liệu, kết quả sẽ trả về các file csv. Sau đó, các file csv sẽ tiếp tục được chuyển vào một chương trình tiền xử lý dữ liệu khác cũng được lập trình bằng Python rồi được đưa vào vùng đệm và kho dữ liệu để xây dựng

các mô hình dữ liệu. Sau đó dữ liệu sẽ được trực quan hóa thông qua các Dashboard và được Public lên Website để mọi người dễ dàng theo dõi.

3.4.4 Chi tiết quá trình tiền xử lý

Quá trình tiền xử lý dữ liệu được thực hiện bằng ngôn ngữ lập trình Python và PowerQuerry.

Ngoài ra, chúng ta sẽ lập trình với ngôn ngữ Python trên file **.ipynb**. Định dạng **.ipynb** là định dạng tệp tin đặc biệt có cấu trúc dạng JSON (JavaScript Object Notation). Nó chứa cả mã lập trình (Python, R, Julia, v.v.) và các ô văn bản được định dạng bằng Markdown, cùng với kết quả tính toán trực tiếp như đồ thị, bảng, hoặc văn bản được hiển thị ngay trong cùng một tài liệu. Điểm nổi bật của các file lập trình được lưu ở dạng này đó là mã nguồn lập trình sẽ được lưu trong các ô (Code Cells), lập trình viên có thể thêm nội dung mã nguồn tùy ý độ dài vào mỗi ô và sau đó có thể tương tác với mã lập trình trong các ô và thực thi mã từng ô một để kiểm tra kết quả. Điều này sẽ giúp em dễ dàng nhận biết sự thay đổi trước và sau khi thêm đoạn chương trình tiền xử lý vào, rằng ta có được kết quả đúng như ý hay không.

Ngoài ra, để tiện cho quá trình tiền xử lý chúng ta sẽ sử dụng thư viện pandas và gọi đến gói hỗ trợ làm việc với dữ liệu Dataframe (DF). DataFrame là một cấu trúc dữ liệu hai chiều (2D) dạng bảng, trong đó dữ liệu được tổ chức thành các hàng và các cột. Để có thể thấy DataFrame là một thư viện cực hợp với các file dữ liệu có cấu trúc, đặc biệt là nguồn dữ liệu là file csv của nhóm trong bài toán lần này.

Extract

Thông qua thư viện DataFrame, ta sẽ đọc các File CSV vào Python và xem thông tin các file đó

```
df_click_stream = pd.read_csv('archive/click_stream.csv')
df_customer = pd.read_csv('archive/customer.csv')
df_product = pd.read_csv('archive/product.csv')
df_transaction = pd.read_csv('archive/transactions.csv')
```

Hình 3.20: đọc các file dữ liệu

- `df_click_stream.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12833602 entries, 0 to 12833601
Data columns (total 6 columns):
 #   Column           Dtype  
 --- 
 0   session_id      object  
 1   event_name       object  
 2   event_time       object  
 3   event_id         object  
 4   traffic_source  object  
 5   event_metadata   object  
dtypes: object(6)
memory usage: 587.5+ MB
```

Hình 3.21: Ví dụ thông tin dữ liệu file click_stream

Transform

1. Xóa các dòng null:

```
df_product.dropna(inplace=True)
```

Hình 3.22: xóa dòng null trong bảng product

2. **Xóa dữ liệu trùng lặp:** Ta sẽ thực hiện kiểm tra trùng lặp ở trên tất cả các dòng, nếu có ta sẽ các dòng này

```

arr_df = [df_transaction, df_click_stream, df_product, df_customer]

total_duplicates = sum(df.duplicated().sum() for df in arr_df)

print("Tổng số lượng dòng trùng lặp:", total_duplicates)

```

Tổng số lượng dòng trùng lặp: 0

Hình 3.23: kiểm tra và xóa các trùng lặp

3. Định dạng lại kiểu dữ liệu:

- Các cột thời gian đang bị chứa lỗn 2 kiểu dữ liệu là giờ và ngày tháng, với bộ dữ liệu này nhóm sẽ chỉ thực hiện lấy các dữ liệu về ngày tháng.

```

df_transaction.head(2)

      created_at  customer_id  booking_id  session_id  product_metadata  payment_method  payment_status  promo_amount  promo_code  shipment_fee  shipment_date_limit
0 2018-07-29T15:22:01.458193Z        5868  186e2bee-0637-4710-8981-50c2d737bc42  3aba06ce-e320-4e51-9469-d9f3fa28e86  [{"product_id": "54728", "quantity": 1, "item_pr...}  Debit Card  Success  1415  WEEKENDSERU  10000  03T00:00:00
1 2018-07-30T12:40:22.365620Z        4774  caadb57b-e808-4f94-9e96-8a7d4c9898db  2ee5ead1-f13e-4759-92df-7ff48475e970  [{"product_id": "16193", "quantity": 1, "item_pr...}  Credit Card  Success  0  NaN  10000  03T00:00:00

df_transaction[['created_at', 'shipment_date_limit']] = df_transaction[['created_at', 'shipment_date_limit']].apply(pd.to_datetime)

# Loại bỏ phần thông tin về múi giờ (UTC)
df_transaction['created_at'] = df_transaction['created_at'].dt.date
df_transaction['shipment_date_limit'] = df_transaction['shipment_date_limit'].dt.date

df_click_stream['event_time'] = pd.to_datetime(df_click_stream['event_time'])

df_click_stream['event_time'] = df_click_stream['event_time'].dt.date

```

Hình 3.24: Xử lý dạng thời gian

	customer_id	booking_id	session_id	product_metadata	payment_method	payment_status	promo_amount	promo_code	shipment_fee	shipment_date_limit
0	5868	186e2bee-0637-4710-8981-50c2d737bc42	3aba06ce-e320-4e51-9469-d9f3fa28e86	[{"product_id": "54728", "quantity": 1, "item_pr...} Debit Card Success 1415 WEEKENDSERU 10000 2018-08-29						
1	4774	caadb57b-e808-4f94-9e96-8a7d4c9898db	2ee5ead1-f13e-4759-92df-7ff48475e970	[{"product_id": "16193", "quantity": 1, "item_pr...} Credit Card Success 0 NaN 10000 2018-08-30						

Hình 3.25: Sau khi xử lý thời gian

- Các cột giới tính, thay vì để M và F, nhóm sẽ thực hiện để đầy đủ từ giới tính: Male và Female

```
df_customer['gender'] = df_customer['gender'].replace({'F': 'Female', 'M': 'Male'})
```



```
df_customer.head(2)
```

	customer_id	first_name	last_name	username	gender	birthdate	device_type
0	2870	Lala	Maryati	671a0865-ac4e-4dc4-9c4f-c286a1176f7e	Female	1996-06-14	iOS
1	8193	Maimunah	Laksmiwati	83be2ba7-8133-48a4-bbcb-b46a2762473f	Female	1993-08-16	Android

Hình 3.26: Xử lý giới tính trong bảng customers

- Định dạng lại một số kiểu dữ liệu khác:

```
df_customer[['birthdate', 'first_join_date']] = df_customer[['birthdate', 'first_join_date']].apply(pd.to_datetime)

df_product['year'] = df_product['year'].astype(int)
```

Hình 3.27: Định dạng lại 1 số kiểu dữ liệu

4. Sinh thêm dữ liệu:

- Từ thông tin tuổi của khách hàng, ta sẽ thực hiện thêm thông tin về tuổi của khách hàng. Trong cơ sở dữ liệu này, nhóm thực hiện chia thành 5 nhóm tuổi:
 - Từ 0 đến dưới 18 tuổi:** Junivelle (Thanh thiếu niên)
 - Từ 18 đến dưới 30 tuổi:** Youth (lớp người trẻ)
 - Từ 30 đến dưới 45 tuổi:** Adults (Người lớn)
 - Từ 45 tuổi đến dưới 60 tuổi:** Middle_age(Người trung niên)
 - Trên 60 tuổi:** Older (Người già)

```
today = datetime.now()
df_customer['age'] = (today.year - (df_customer['birthdate']).dt.year)
```

Python


```
df_customer = df_customer.drop(columns='birthdate')
```

Python


```
# Giả sử df_users là DataFrame chứa dữ liệu về người dùng, có cột 'age' chứa thông tin về tuổi
# Thêm một cột mới để chứa nhóm tuổi
df_customer['age_group'] = pd.cut(df_customer['age'], bins=[0, 18, 30, 45, 60, float('inf')]),
labels=['juvenile', 'youth', 'adults', 'middle_age', 'older'])
```

Python

Hình 3.28: Sinh thêm dữ liệu về tuổi và nhóm tuổi

df_customer.head(2)													Python
	customer_id	first_name	last_name	username	gender	device_type	device_id	device_version	home_location	home_country	first_join_date	age	age_group
0	2870	Lala	Maryati	671a0865-ac4e-4dc4-9c4f-c286a1176f7e	Female	iOS	c9c0de76-0a6c-4ac2-843f-65264ab9fe63	iPhone; CPU iPhone OS 14_2_1 like Mac OS X	Sumatera Barat	Indonesia	2019-07-21	28	youth
1	8193	Maimunah	Laksmiwati	83be2ba7-8133-48a4-bbc9-b46a2762473f	Female	Android	fb331c3d-f42e-40fe-afe2-b4b73a8a6e25	Android 2.2.1	Jakarta Raya	Indonesia	2017-07-16	31	adults

Hình 3.29: Kết quả sinh thêm dữ liệu nhóm tuổi

- Từ thông tin các ngày tạo đơn và thời gian giao vận được thể hiện trong bảng transaction, nhóm thực hiện sinh thêm dữ liệu số ngày giao hàng và kết quả

# Replace 0 values with NaT in delivered_at and shipped_at columns df_transaction['shipment_date_limit'] = pd.to_datetime(df_transaction['shipment_date_limit'], errors='coerce') df_transaction['created_at'] = pd.to_datetime(df_transaction['created_at'], errors='coerce') # Tính số ngày giao hàng và điều chỉnh 0 cho các ô không thực hiện được hoặc có giá trị 0 delivery_days = (df_transaction['shipment_date_limit'] - df_transaction['created_at']).dt.days.fillna(0) # Replace negative delivery days with 0 df_transaction['num_of_delivery_days'] = np.where(delivery_days < 0, 0, delivery_days)	Python																																	
df_transaction.head(2)																																		
<table border="1"> <thead> <tr> <th>_id</th> <th>session_id</th> <th>product_metadata</th> <th>payment_method</th> <th>payment_status</th> <th>promo_amount</th> <th>promo_code</th> <th>shipment_fee</th> <th>shipment_date_limit</th> <th>total_amount</th> <th>num_of_delivery_days</th> </tr> </thead> <tbody> <tr> <td>ee-3aba06ce- 10-e320-4e51- 81-9469- c42-d9f3fa328e86</td> <td></td> <td>[{'product_id': 54728, 'quantity': 1, 'item_pr...}</td> <td>Debit Card</td> <td>Success</td> <td>1415</td> <td>WEEKENDSERU</td> <td>10000</td> <td>2018-08-03</td> <td>199832</td> <td>5</td> </tr> <tr> <td>7b-2ee5ead1- 94-f13e-4759- 96-92df- fdb 7ff48475e970</td> <td></td> <td>[{'product_id': 16193, 'quantity': 1, 'item_pr...}</td> <td>Credit Card</td> <td>Success</td> <td>0</td> <td>NaN</td> <td>10000</td> <td>2018-08-03</td> <td>155526</td> <td>4</td> </tr> </tbody> </table>		_id	session_id	product_metadata	payment_method	payment_status	promo_amount	promo_code	shipment_fee	shipment_date_limit	total_amount	num_of_delivery_days	ee-3aba06ce- 10-e320-4e51- 81-9469- c42-d9f3fa328e86		[{'product_id': 54728, 'quantity': 1, 'item_pr...}	Debit Card	Success	1415	WEEKENDSERU	10000	2018-08-03	199832	5	7b-2ee5ead1- 94-f13e-4759- 96-92df- fdb 7ff48475e970		[{'product_id': 16193, 'quantity': 1, 'item_pr...}	Credit Card	Success	0	NaN	10000	2018-08-03	155526	4
_id	session_id	product_metadata	payment_method	payment_status	promo_amount	promo_code	shipment_fee	shipment_date_limit	total_amount	num_of_delivery_days																								
ee-3aba06ce- 10-e320-4e51- 81-9469- c42-d9f3fa328e86		[{'product_id': 54728, 'quantity': 1, 'item_pr...}	Debit Card	Success	1415	WEEKENDSERU	10000	2018-08-03	199832	5																								
7b-2ee5ead1- 94-f13e-4759- 96-92df- fdb 7ff48475e970		[{'product_id': 16193, 'quantity': 1, 'item_pr...}	Credit Card	Success	0	NaN	10000	2018-08-03	155526	4																								

Hình 3.30: Sinh thêm dữ liệu số ngày giao hàng và kết quả

- Trong các bảng click_stream và order có chứa các cột có thông tin được gói trong metadata như product_metadata và event_metadata. Sự tiện lợi của metadata nằm ở chỗ: Bằng cách gom các trường thông tin vào một metadata, có thể dễ dàng tích hợp thông tin từ nhiều nguồn khác nhau. Thay vì phải duyệt qua nhiều bảng dữ liệu riêng biệt, chúng ta chỉ cần xem xét “metadata” duy nhất. Nhờ thế nhóm sẽ thực hiện giải phóng gói metadata này để tạo ra thêm nhiều thông tin hơn, chung cách giải phóng cho cả 2 cột.

# Split the 'booking_id' column by '-' and take the first element df_transaction['order_id'] = df_transaction['booking_id'].str.split('-').str[-1]	Python
df_order = df_transaction[['order_id', 'product_metadata']] df_transaction = df_transaction.drop(columns=['product_metadata'])	

order_id	product_metadata
0 50c2d737bc42	[{'product_id': 54728, 'quantity': 1, 'item_pr...}
1 8a7d4c9898db	[{'product_id': 16193, 'quantity': 1, 'item_pr...}

Hình 3.31: Sinh thêm dữ liệu orders từ product_metadata và kết quả

```

import ast
metadata_list = df_order['product_metadata'].apply(ast.literal_eval).tolist()
# Khởi tạo danh sách để lưu kết quả
result = []
# Lặp qua mỗi hàng trong dữ liệu
for idx, row in enumerate(metadata_list):
    # Lặp qua mỗi dictionary trong danh sách metadata
    for item in row:
        # Tạo một tuple chứa giá trị id_order và các giá trị từ từng dictionary trong metadata
        result.append((df_order['order_id'][idx], item['product_id'], item['quantity'], item['item_price']))
# Tạo DataFrame từ danh sách kết quả
df_order_final = pd.DataFrame(result, columns=['order_id', 'product_id', 'quantity', 'item_price'])

```

Python

Hình 3.32: Xử lý metadata của bảng order

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1254585 entries, 0 to 1254584
Data columns (total 4 columns):
 #   Column      Non-Null Count   Dtype  
--- 
 0   order_id    1254585 non-null  object 
 1   product_id  1254585 non-null  int64  
 2   quantity    1254585 non-null  int64  
 3   item_price  1254585 non-null  int64  
dtypes: int64(3), object(1)
memory usage: 38.3+ MB

df_click_stream['event_id'].value_counts()

event_id
9c4388c4-c95b-4678-b5ca-e9cbc0734109    1
b15ad399-9ea0-4ecc-986a-01eaa8955776    1
a2d88daa-e6dd-4a77-a7b8-5e48738af4c8    1
2c3f3a94-09fd-4ef1-bee2-1441cac40654    1
abf3061c-b98e-42fb-8c69-3318ef832706    1
..
a50f54cc-3dcd-4bd9-ad7a-1a88664f5559    1
056296e2-371d-43e4-8f64-43784cff554b    1
0dba504b-4dd7-4846-8aab-87165903457a    1
33450c3e-e59f-447a-a8b6-101a26b80e20    1
3c60d8bc-b3e1-41fa-a87b-f65a9053856a    1
Name: count, Length: 12833602, dtype: int64

```

Hình 3.33: Kết quả xử lý metadata bảng order

Click to add a breakpoint d[10]

✓ 0.0s

	session_id	event_name	event_time	event_id	traffic_source	event_metadata
0	fb0abf9e-fd1a-44dd-b5c0-2834d5a4b81c	HOMEPAGE	2019-09-06	9c4388c4-c95b-4678-b5ca-e9cbc0734109	MOBILE	NaN
1	fb0abf9e-fd1a-44dd-b5c0-2834d5a4b81c	SCROLL	2019-09-06	4690e1f5-3f99-42d3-84a5-22c4d8500a	MOBILE	NaN
2	7d440441-e67a-4d36-b324-80ffd636d166	HOMEPAGE	2019-09-01	88aeaeb5-ec98-4859-852c-8abb483faf31	MOBILE	NaN
3	7d440441-e67a-4d36-b324-80ffd636d166	ADD_TO_CART	2019-09-01	934e306e-ecc6-472f-9ccb-12c8536910a2	MOBILE	{'product_id': 15315, 'quantity': 4, 'item_pri...
4	7d440441-e67a-4d36-b324-80ffd636d166	BOOKING	2019-09-01	9f4767a1-40fa-4c9c-9524-dfad18634d56	MOBILE	{'payment_status': 'Success'}
5	7d440441-e67a-4d36-b324-80ffd636d166	SEARCH	2019-09-01	c952142b-4fe9-4694-ad7f-21a5d1bed9ca	MOBILE	{'search_keywords': 'Dress Kondangan'}
6	7d440441-e67a-4d36-b324-80ffd636d166	HOMEPAGE	2019-09-01	365b3840-9647-4bf5-917f-f0bec3d05332	MOBILE	NaN
7	7d440441-e67a-4d36-b324-80ffd636d166	ITEM_DETAIL	2019-09-01	1a1e3548-108e-4520-bf04-1b01d43a72cb	MOBILE	NaN
8	7d440441-e67a-4d36-b324-80ffd636d166	SCROLL	2019-09-01	f7246095-b094-46c1-b2ac-bcf0d080146c3	MOBILE	NaN
9	7d440441-e67a-4d36-b324-80ffd636d166	ITEM_DETAIL	2019-09-01	74a000dc-b931-4c59-abcc-02e991623bc1	MOBILE	NaN

Hình 3.34: Trước khi xử lý metadata của click stream

```
# Chuyển đổi dữ liệu từ cột 'event_metadata' thành danh sách các từ điển
# Bỏ qua các hàng không chứa từ điển
data = []
keys = set()
for i in df_click_stream['event_metadata']:
    if isinstance(i, str) and i.strip() != "":
        row = ast.literal_eval(i)
        data.append(row)
        keys.update(row.keys())
    else:
        data.append({})
```

Hình 3.35: xử lý metadata của click_stream - lần 1

```
# Mở (hoặc tạo) tệp CSV để ghi dữ liệu
with open('a.csv', 'w', newline='') as csvfile:
    writer = csv.DictWriter(csvfile, fieldnames=fieldnames)

    # Viết tiêu đề cột vào tệp CSV
    writer.writeheader()

    # Duyệt qua từng hàng dữ liệu và ghi vào tệp CSV
    for row in data:
        writer.writerow({key: row.get(key, None) for key in fieldnames})

df_concat = pd.read_csv('a.csv')

# Sử dụng pd.concat() để thêm các cột mới vào DataFrame hiện có
df_event = pd.concat([df_click_stream, df_concat], axis=1)
```

Hình 3.36: xử lý metadata của click_stream - lần 2

df_event.head(10)													
✓	0.0s	session_id	event_name	event_time	event_id	traffic_source	payment_status	quantity	promo_amount	search_keywords	product_id	promo_code	item_price
0		fb0abf9e-fd1a-44dd-b5c0-2834d5a4b81c	HOMEPAGE	2019-09-06	9e4388c4-c95b-4678-b5ca-e9cbc0734109	MOBILE	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1		fb0abf9e-fd1a-44dd-b5c0-2834d5a4b81c	SCROLL	2019-09-06	4690e1f5-3f99-42d3-84a5-22c4cd8500a	MOBILE	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2		7d440441-e67a-4d36-b324-80ffd636d166	HOMEPAGE	2019-09-01	88aaeb5-ec98-4859-852c-8abb483fa31	MOBILE	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3		7d440441-e67a-4d36-b324-80ffd636d166	ADD_TO_CART	2019-09-01	934e306e-ecc6-472f-9ccb-12c8536910a2	MOBILE	NaN	4.0	NaN	NaN	15315.0	NaN	313529.0
4		7d440441-e67a-4d36-b324-80ffd636d166	BOOKING	2019-09-01	9f4767a1-40fa-4c9c-9524-dfad18634d56	MOBILE	Success	NaN	NaN	NaN	NaN	NaN	NaN
5		7d440441-e67a-4d36-b324-	SEARCH	2019-09-01	c952142b-4fe9-4694-ad7f-	MOBILE	NaN	NaN	NaN	Dress Kondangan	NaN	NaN	NaN

Hình 3.37: Kết quả

5. Xóa cột không sử dụng: Ta sẽ kho

df_transaction.head(2)									
product_metadata	payment_method	payment_status	promo_amount	promo_code	shipment_fee	shipment_date_limit	shipment_location_lat	shipment_location_long	total_amount
{'product_id': 1, 'quantity': 1, 'item_pr...	Debit Card	Success	1415	WEEKENDSERU	10000	2018-08-03T05:07:24.812676Z	-8.227893	111.969107	199832
{'product_id': 1, 'quantity': 1, 'item_pr...	Credit Card	Success	0	Nan	10000	2018-08-03T01:29:03.415705Z	3.013470	107.802514	155526

Hình 3.38: Trước khi xóa cột thừa của bảng transaction

```
df_transaction = df_transaction.drop(columns=['shipment_location_lat', 'shipment_location_long'])
```

Hình 3.39: Lệnh xóa cột thừa

# Sau khi xóa										
df_transaction.head(2)										
customer_id	booking_id	session_id	product_metadata	payment_method	payment_status	promo_amount	promo_code	shipment_fee	shipment_date_limit	total_amount
5868	186e2bee-0637-4710-8981-50c2d737bc42	3abaa6ce-e320-4e51-9469-d9f3fa328e86	{'product_id': 54728, 'quantity': 1, 'item_pr...	Debit Card	Success	1415	WEEKENDSERU	10000	2018-08-03T05:07:24.812676Z	199832
4774	caadb57b-e808-4f94-9e96-8a7d4c9898db	2ee5ead1-f13e-4759-92df-7ff48475e970	{'product_id': 16193, 'quantity': 1, 'item_pr...	Credit Card	Success	0	Nan	10000	2018-08-03T01:29:03.415705Z	155526

Hình 3.40: Sau khi xóa cột thừa của bảng transaction

```
df_customer = df_customer.drop(columns=['email', 'home_location_lat', 'home_location_long', 'first_name', 'last_name', 'username'])
```

Hình 3.41: Lệnh xóa cột thừa của customer

Load

Sau khi kết thúc quá trình Transform biến đổi dữ liệu, ta lưu lại 5 file csv mới để sẵn sàng đưa dữ liệu này vào cơ sở dữ liệu

```
df_transaction.to_csv('Data after ETL/transaction.csv', index=False, mode='w')
df_customer.to_csv('Data after ETL/customer.csv', index=False, mode='w')
df_order_final.to_csv('Data after ETL/order.csv', index=False, mode='w')
df_event.to_csv('Data after ETL/event.csv', index=False, mode='w')
df_product.to_csv('Data after ETL/product.csv', index=False, mode='w')
```

Hình 3.42: Lưu lại các file sau khi đã Transform

Thực hiện đổ dữ liệu từ các file csv đã được xử lý vào MySQL:

```
31 # Open the CSV file and create a CSV reader object
32 with open(file_path, 'r', newline='', encoding='utf-8') as file:
33     csv_data = csv.reader(file)
34     firstRow = True
35     for row in csv_data:
36         if firstRow:
37             header = row
38
39             table_name = name
40             columns = []
41
42             for col in header:
43                 data_type = guess_data_type(col)
44                 columns.append(f'{col} {data_type}')
45
46             create_table_sql = f"CREATE TABLE IF NOT EXISTS `{table_name}` ({', '.join(columns)})"
47             drop_table_query = f"DROP TABLE IF EXISTS {_db}.{table_name}"
48             cursor.execute(drop_table_query)
49             cursor.execute(create_table_sql)
50
51             firstRow = False
52             continue
53
54             s = (f'INSERT INTO `{name}` ({", ".join(header)}) VALUES ({", ".join(["%" + "s"] * len(row))})')
55             cursor.execute(s, row)
56
57
58 # Commit the changes
59 mydb.commit()
```

Hình 3.43: Thực hiện đẩy dữ liệu vào MySQL

3.5 Mô hình dữ liệu OLTP

OLTP (Xử lý giao dịch trực tuyến) được đặc trưng bởi số lượng lớn các giao dịch trực tuyến ngắn (CHÈN, CẬP NHẬP, XÓA). Điểm nhấn chính của hệ thống OLTP là xử lý truy vấn rất nhanh, duy trì tính toàn vẹn của dữ liệu trong môi trường đa truy cập và hiệu quả đo bằng số lượng giao dịch mỗi giây. Trong cơ sở dữ liệu OLTP có dữ liệu chi tiết hiện tại và lược đồ được sử dụng để lưu trữ cơ sở dữ liệu giao dịch là mô hình thực thể (thường là 3NF) Ví dụ hệ thống OLTP:

- Trung tâm ATM
- Ngân hàng trực tuyến
- Gửi tin nhắn văn bản
- Đặt vé máy bay trực tuyến
- Thêm sách vào giỏ hàng

Từ bộ dữ liệu ban đầu, chúng em xây dựng được mô hình dữ liệu quan hệ gồm những bảng sau:

- **Bảng event:** Thông tin liên quan đến các sự kiện tương tác với ứng dụng của người dùng như click xem sản phẩm, mua sản phẩm, hủy đơn....
- **Bảng product:** Thông tin liên quan đến sản phẩm như tên sản phẩm, tồn kho, giá tiền, nhà sản xuất...
- **Bảng order:** Thông tin liên quan đến các đơn hàng như người đặt đơn, trạng thái đơn, ngày giao đơn...
- **Bảng transaction:** Thông tin liên quan đến giao dịch như trạng thái giao dịch, hình thức thanh toán, số tiền...
- **Bảng customer:** Thông tin liên quan đến người dùng



Hình 3.44: OLTP

3.6 Mô hình dữ liệu OLAP

3.6.1 Phân tích các chiều (Dimensions) và chủ điểm phân tích (Facts)

Các chiều (Dimensions)

Trong kho dữ liệu, các dim là tập hợp thông tin tham chiếu về một sự kiện có thể đo lường được.

Kho dữ liệu lưu trữ các thuộc tính mô tả dưới dạng các cột trong bảng dim. Ví dụ: các thuộc tính của dim khách hàng có thể bao gồm họ và tên, ngày sinh, giới tính,... hoặc dim trang web sẽ bao gồm tên trang web và URL. Mỗi bảng dim có một khóa chính duy nhất xác định mỗi bản ghi (hàng). Bảng dim được liên kết với bảng dữ liệu bằng cách sử dụng khóa này. Dữ liệu trong các bảng có thể được lọc và nhóm theo nhiều cách, kết hợp các thuộc tính khác nhau. Ví dụ: dữ liệu bán hàng với các dim sản phẩm, cửa hàng và ngày có thể được truy vấn để tìm "số lượng sản phẩm thuộc danh mục 'Điện tử' đã bán tại cửa hàng 'TechMart' vào tháng 3 năm 2023, được nhóm theo ngày".

Nhiều dim chứa một hệ thống phân cấp các thuộc tính hỗ trợ việc khoan lên và xuống. Ví dụ: Dim ngày có thể chứa phân cấp năm > quý > tháng > tuần > ngày. Báo cáo hiển thị số lần đăng nhập trang web cho năm 2009 theo tháng có thể xem chi tiết để hiển thị thông tin đăng nhập theo năm hoặc xem chi tiết để hiển thị thông tin đăng nhập theo ngày.

Kích thước được sử dụng trong các lược đồ hình sao và bông tuyết trong kho dữ liệu, khối OLAP và các ứng dụng phân tích kinh doanh.

Đối với bộ dữ liệu sử dụng trong báo cáo này, nhóm tác giả đã phân tích thành các dim sau:

- Dim Order: Thông tin liên quan đến đơn hàng
- Dim Product: Thông tin liên quan đến sản phẩm
- Dim Category: Thông tin liên quan đến thể loại hàng
- Dim Session: Thông tin liên quan đến phiên hoạt động
- Dim Traffic Source: Thông tin liên quan đến nguồn truy cập
- Dim Device: Thông tin liên quan đến thiết bị truy cập
- Dim Location: Thông tin liên quan đến vị trí
- Dim Date: Thông tin liên quan đến ngày giờ
- Dim Customer: Thông tin liên quan đến khách hàng
- Dim Payment: Thông tin liên quan đến phương thức thanh toán
- Dim Status: Thông tin liên quan đến trạng thái thanh toán

Dưới đây là các giá trị cho các bảng dim, được nhóm và phân loại theo các màu sắc khác nhau:

Dim_device		Dim_Product				Dim_Order		Dim_Category	
	2	4	142	46	8	44	3362	7	45
device_type									
iOS	Men	Shirts	Navy Blue	Casual		1	191247	Apparel	Topwear
Android	Women	Jeans	Blue	Ethnic		4	145526	Accessories	Bottomwear
	Boys	Watches	Silver	Formal		6	135174	Footwear	Watches
	Girls	Track Pants	Black	Sports		2	271012	Personal Care	Socks
	Unisex	Tshirts	Grey	Smart Casual		8	198753	Free Items	Shoes
		Socks	Green	Travel		3	183234	Sporting Goods	Belts
		Casual Shoes	Purple	Party		5	296599	Home	Flip Flops
		Belts	White	Home		7	143913		Bags
		Flip Flops	Beige			12	364776		Innerwear
		Handbags	Brown			10	292052		Sandal
		Tops	Bronze			9	241945		Shoe Accessories
		Bra	Teal			13	297248		Fragrance
		Sandals	Copper			11	191138		Jewellery
		Shoe Accessories	Pink			20	78559		Lips
		Sweatshirts	Off White			14	193870		Saree
		Deodorant	Maroon			19	314243		Eyewear
		Formal Shoes	Red						Glasses

Dim_Session_id Dim_Traffic_Source Dim_Location

Dim_Session_id	Dim_Traffic_Source	Dim_Location
6456	MOBILE	Sumatera Barat
	WEB	Indonesia
		Jakarta Raya
		Nusa Tenggara Barat
		Kalimantan Timur
		Kalimantan Selatan
		Kepulauan Riau
		Bengkulu
		Jawa Barat
		Nusa Tenggara Timur
		Yogyakarta
		Jawa Timur
		Jawa Tengah
		Sulawesi Tenggara
		Lampung
		Bangka Belitung
		Maluku

Dim_Payment	Dim_Status	Dim_date	Dim_Payment	Dim_Customer
5	2	31 12 4 7	5	2 62 5
payment_method	payment_status		payment_method	
Debit Card	Success		Debit Card	Female 28 youth
Credit Card			Credit Card	Male 31 adults
OVO	Failed		OVO	35 juvenile
LinkAja			LinkAja	33 middle_age
Gopay			Gopay	24 older
				34
				19
				29
				37
				42
				38
				26
				32
				23
				25
				21

Hình 3.45: Hệ thống chiêu khái niệm

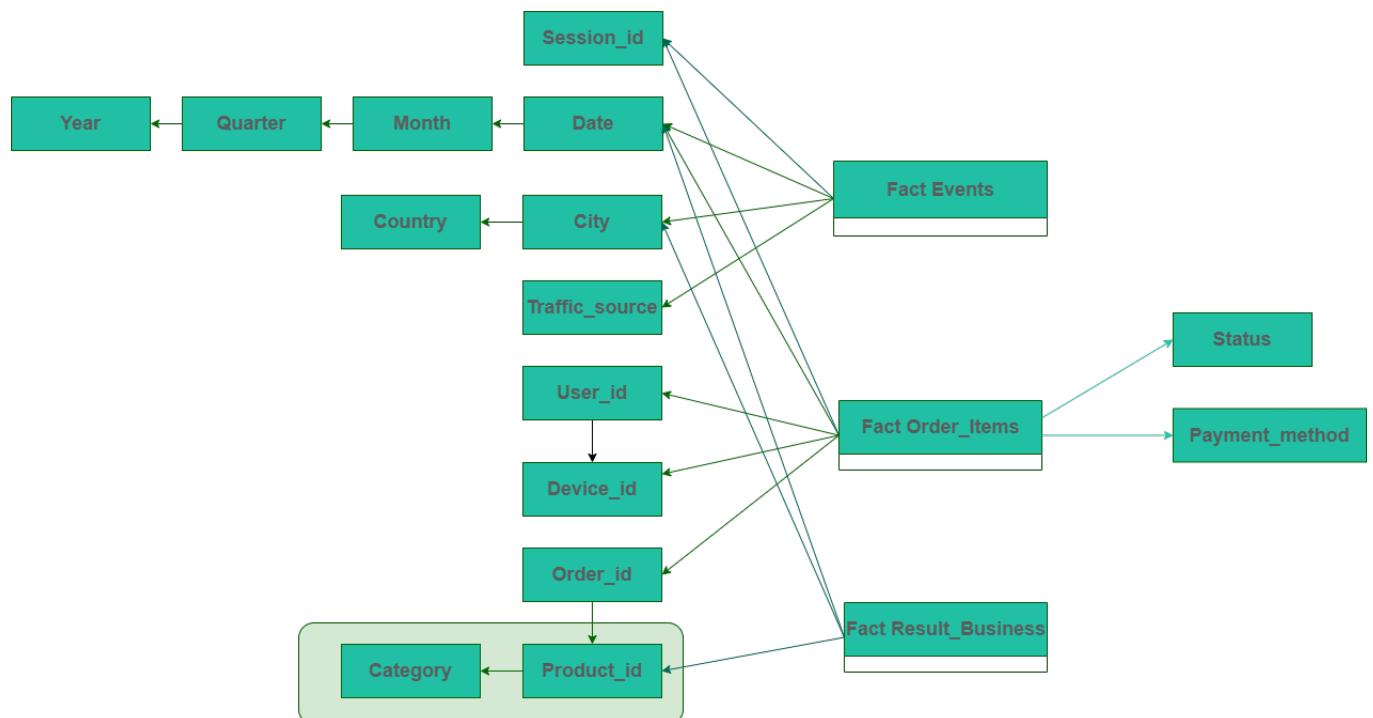
Các chủ điểm phân tích (Facts)

Chủ điểm phân tích là trung tâm lược đồ hình sao của kho dữ liệu. Đây là một khái niệm quan trọng cần thiết cho kho dữ liệu. Chủ điểm cần phân tích lưu trữ thông tin định lượng để phân tích và thường không được chuẩn hóa. Chủ điểm phân tích hoạt động với các chiều (dim) và nó chứa dữ liệu được phân tích và các chiều dữ liệu về cách mà dữ liệu có thể được phân tích. Do đó, một bảng chủ thể phân tích bao gồm hai loại cột. Cột khóa ngoại cho phép kết hợp với các bảng thứ nguyên và cột đo lường chứa dữ liệu đang được phân tích.

Đối với bộ dữ liệu sử dụng để phân tích báo cáo này, các chủ điểm phân tích là kết quả kinh doanh, tần suất truy cập hệ thống, các giao dịch.

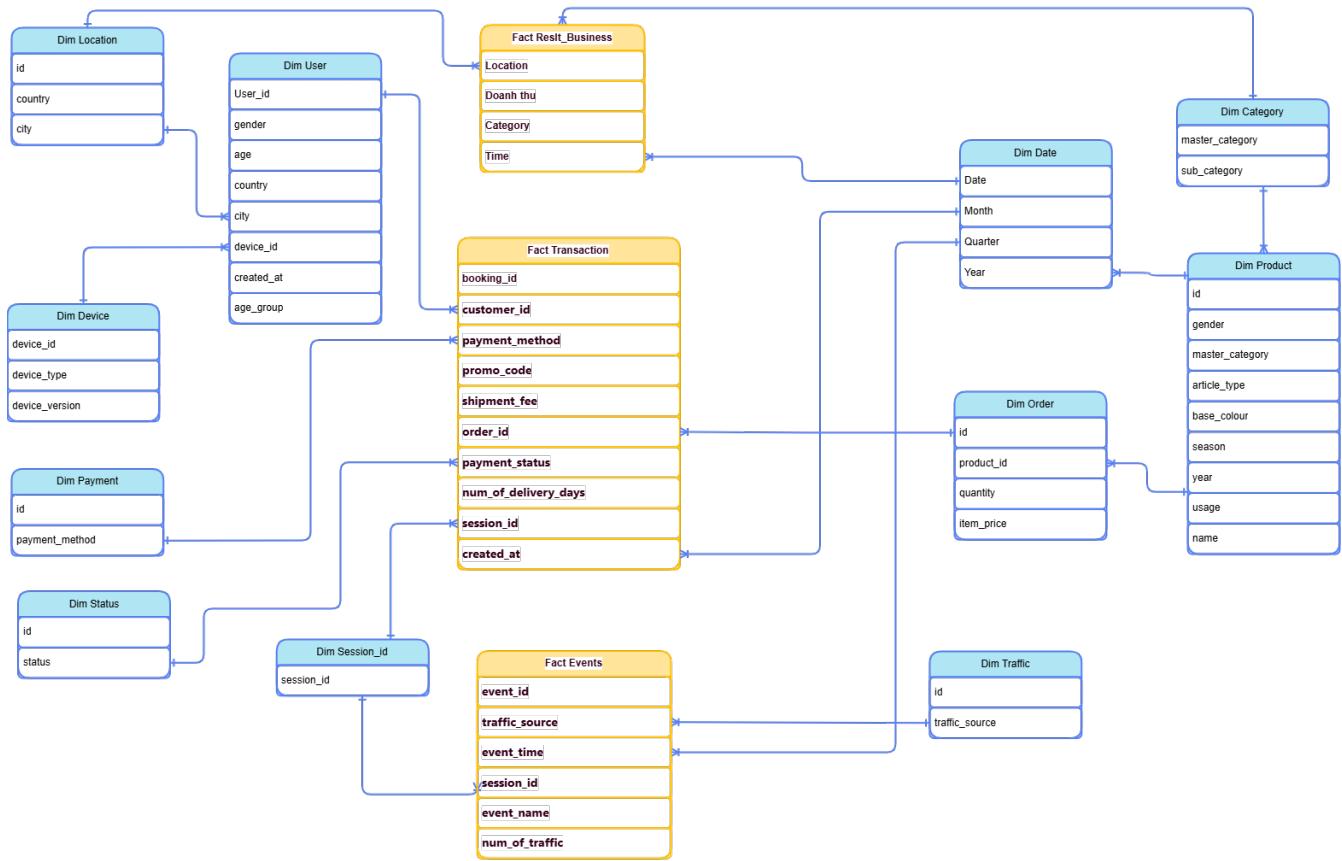
3.6.2 Data model logic

Từ các dim và fact đã xác định ở trên, ta xây dựng được data model logic như dưới đây



Hình 3.46: Data model logic

3.6.3 Mô hình dữ liệu quan hệ OLAP



Hình 3.47: Data model OLAP

3.7 Các mẫu Dashboard

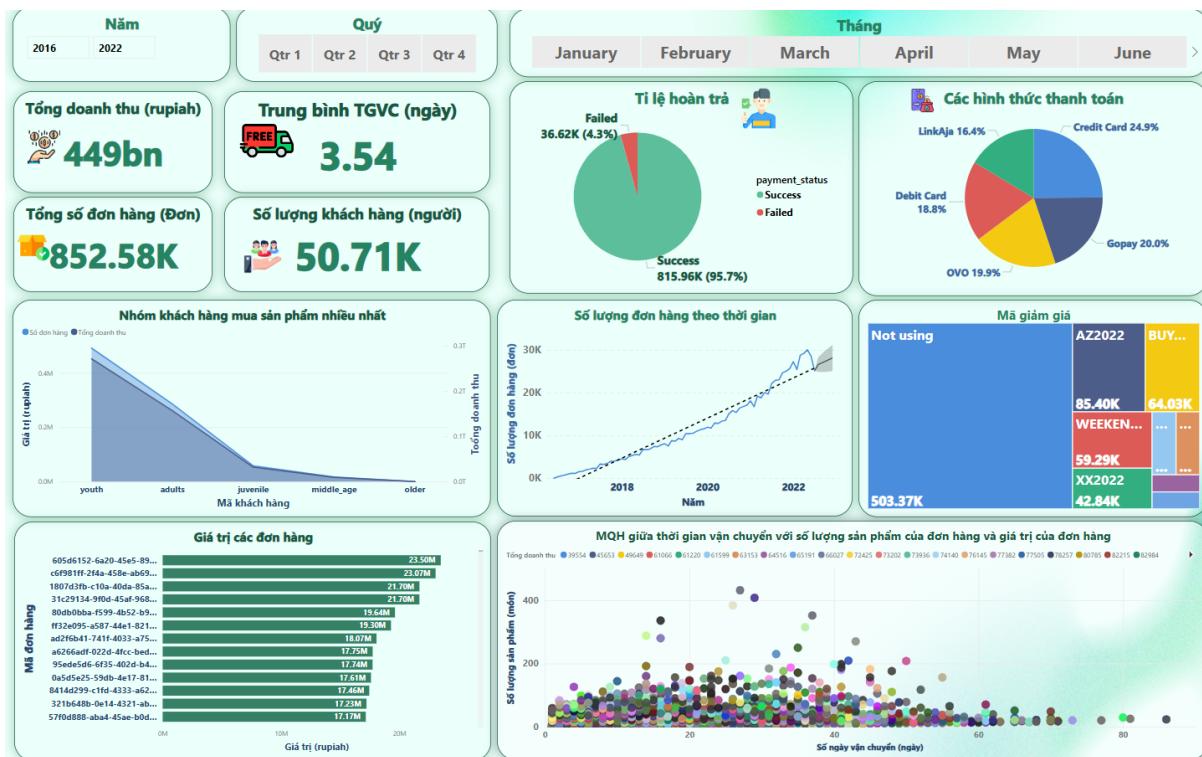
Thông qua việc xử lý dữ liệu, nhóm đã thiết kế được 3 Dashboard bằng công cụ PowerBI, các Dashboard đã được đóng gói đưa lên website để mọi người dễ dàng theo dõi.

Một số thành phần sẽ có trong Dashboard bao gồm:

- **Các Slicer:** chứa các trường thông tin về thời gian (ngày, tháng, quý, năm) hay các trường thông tin về quốc gia.
- **Các Visual Chart:** với đa dạng thể loại (Map Chart, Column Chart, Clustered Column Chart, Bar Chart, Line Chart,..) thể hiện các thông tin cần phân tích.
- **Các Visual Card:** Chứa các thông tin quan trọng cần nhấn mạnh.
- **Hàm DAX:** được thiết lập để cảnh báo cho các trường thông tin đặc biệt.

3.7.1 Dashboard phân tích giao dịch đặt hàng

Thiết kế Dashboard



Hình 3.48: Dashboard phân tích giao dịch đặt hàng

Phân tích Dashboard

Tính từ năm 2016 đến năm 2022, StartUp Campus đã ghi nhận được hơn 852 nghìn đơn đặt hàng, với lượt khách hàng đặt mua khoảng 50.71 nghìn người và tổng giá trị các đơn hàng là 449 tỷ rupiah (đơn vị tiền tệ Indonesia), có thể thấy dù số lượng người mua có phần khiêm tốn so với quy mô dân cư Indonesia nhưng lại có sức tiêu thụ khá lớn.

Cùng với đó, thời gian vận chuyển tính trên một đơn hàng chỉ là hơn 3 ngày (3 ngày và 12 giờ). Điều này cho thấy dù Indonesia là quốc đảo, thế nhưng các khâu làm việc của công ty vận hành khá tốt, từ bên kho nhận hàng cho đến các đơn vị vận chuyển. Đây là một điểm tích cực giúp là điểm cộng lớn với các khách hàng.

Trong hơn 852 nghìn đơn hàng được đặt thì số lượng đơn hoàn trả chỉ chiếm khoảng 4.7%, điều này cho thấy phần đa các sản phẩm đều có chất lượng ổn định, cùng với đó quá trình vận chuyển được diễn ra khá trơn tru để bảo đảm chất lượng sản phẩm. Việc số lượng đơn bị hủy ít vừa cho thấy người tiêu dùng đều xem khá nghiêm túc các sản phẩm, vừa hài lòng với chất lượng sản phẩm họ nhận được.

Trên sàn thương mại này, người tiêu dùng có đến 5 hình thức thanh toán (2 hình thức quốc tế: Debit Card và Credit Card, 3 hình thức thanh toán qua các phương thức nội địa: GoPay, Ovo, LinkAja) và thị phần các phương thức thanh toán cho thấy phương thức thanh toán thẻ quốc tế có phần nhỉnh hơn, nhưng nhìn chung thị phần các phương thức thanh toán là khá sát nhau,

chênh lệch nhau không nhiều. Có thể thấy các phương thức đều đem lại trải nghiệm tốt cho người dùng, giúp người dùng đa dạng phương thức thanh toán. Tiếp tục là một điểm cộng lớn của công ty.

Số lượng đơn hàng theo nhóm khách hàng: Có thể thấy, nhóm lớp người trẻ và người lớn (youth và adults) đóng góp lớn vào số lượng đơn hàng và doanh thu chung. Điều này là dễ hiểu khi nhóm người này là nhóm người có khả năng thích nghi với nhịp độ xã hội cao, dễ dàng tiếp thu các nền tảng công nghệ mới, đặc biệt là khi nền thương mại điện tử phát triển trong thời gian họ có sự nhạy bén nhất và cực kỳ phù hợp với lối sống hiện tại của họ. Do đó doanh thu họ đóng góp lớn là điều không thể phủ nhận và hoàn toàn thực tế.

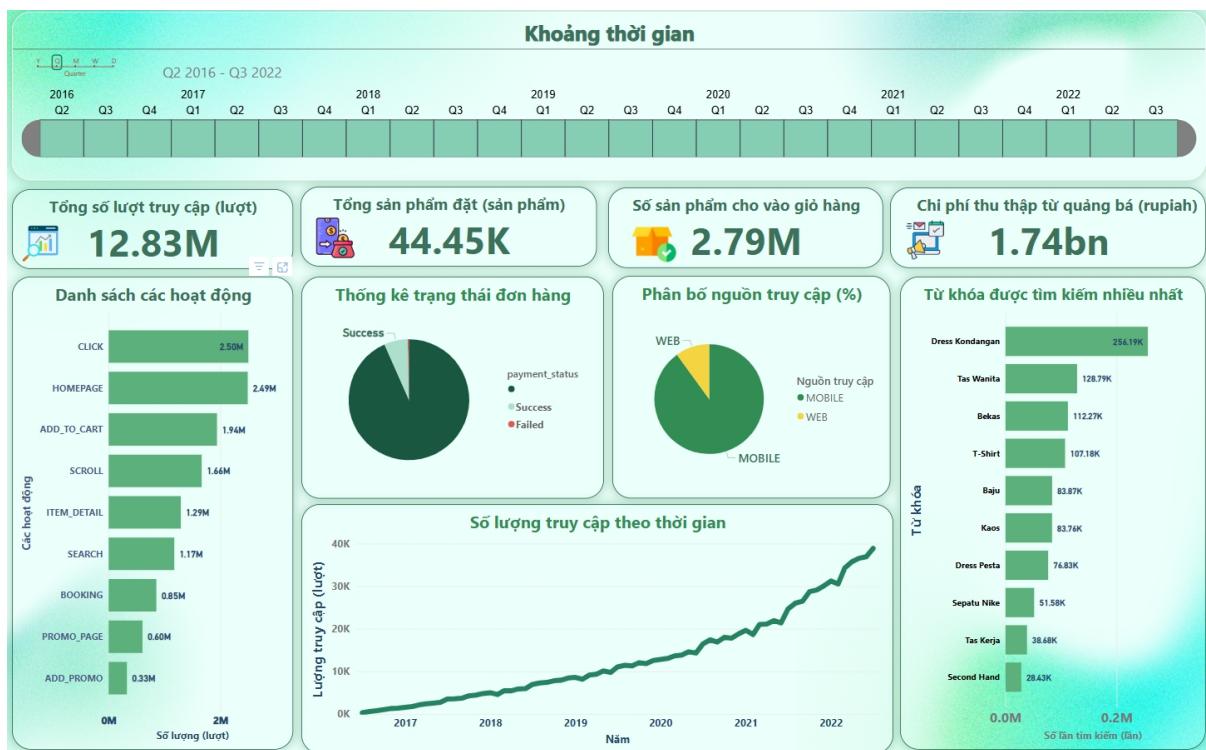
Số lượng đơn hàng theo thời gian: Biểu đồ đường cho thấy số lượng đơn hàng tăng qua từng năm và tiếp tục có dự báo tăng trong các năm tiếp theo. Điều này là dễ hiểu khi xã hội phát triển hơn và sự tiện lợi mà các sàn thương mại điện tử mang lại.

Phân tích sản phẩm theo mã giảm giá: Cố thể thấy phần lớn các đơn hàng người dùng không áp mã giảm giá (Not Using). Có thể lý giải điều này bằng nhiều cách. Một là công ty chưa áp dụng nhiều chiến dịch giảm giá marketing để thu hút người dùng; nếu nhìn ở góc độ này thì số lượng khách hàng khá khiêm tốn là điều có thể hiểu được. Hai là các sản phẩm trên sàn thật sự có giá phù hợp với nhu cầu tiêu dùng của khách hàng, dẫn đến việc họ tái mua hàng nhiều lần trên sàn; điều này cũng có thể là cơ sở để khách định tệp khách hàng của StartupCampus khá vững chắc và họ đang làm tốt việc cân bằng giữa lợi nhuận thu vào và tệp khách hàng trung thành.

Tương quan giữa giá trị đơn hàng và thời gian vận chuyển với giá thành: Scatter Chart cho thấy với giá trị đơn hàng lớn (số lượng sản phẩm trong đơn hàng) dễ kéo theo việc thời gian vận chuyển tăng cao. Điều này là hợp lý khi bên vận chuyển họ sẽ cần thời gian lớn hơn để sắp xếp lại lượng sản phẩm trong thiết bị vận chuyển sao cho đảm bảo chất lượng sản phẩm và tối ưu chi phí. Cùng với đó, các điểm trên chart tập trung ở vùng thấp bên trái, có thể thấy người dùng hay mua số lượng sản phẩm khá vừa phải về cả số lượng và túi tiền của họ.

3.7.2 Dashboard phân tích truy cập hệ thống

Thiết kế Dashboard



Hình 3.49: Dashboard truy cập hệ thống

Phân tích

Trong khoảng thời gian từ 2016 đến 2022, ghi nhận trên hệ thống đã thực hiện 12.83 triệu lượt truy cập, với đa dạng các thao tác như Click, Search, Add Promo Code (thêm mã giảm giá), Add To cart (thêm vào giỏ hàng)... Gần 13 triệu lượt truy cập này dẫn đến gần 45 nghìn sản phẩm được đặt, khoảng 2.8 triệu sản phẩm được tạo ra. Cùng với đó họ thu về gần 1.74 tỉ tiền thu thập từ chi phí quảng bá.

Thống kê truy cập theo thời gian: Lượng truy cập tăng dần theo thời gian dẫn đến việc các sản phẩm được quan tâm cũng tăng. Điều này cho thấy công ty đang giữ chân được người dùng khá tốt.

Phân bổ nguồn truy cập: Có thể thấy nguồn truy cập chủ yếu đến từ các thiết bị di động. Bởi vì sự tiện lợi của các điện thoại di động và việc ngày càng dễ dàng sở hữu cho mình một chiếc điện thoại cũng khiến nguồn truy cập là Mobile trở nên áp đảo là điều không khó dự đoán.

Phân tích danh sách các hoạt động: Với đa dạng các hoạt động, ta lại thấy một điều đáng chú ý là các hoạt động Click và hoạt động Homepage đang chiếm số lượng cực lớn khi tổng 2 hoạt động này chiếm gần 50% tổng số lượt thao tác truy cập hệ thống. Hai hoạt động này tuy không đem lại nhiều giá trị kinh tế, nhưng lại có nhiều ý nghĩa khác. Với hoạt động Click cho thấy, người dùng có xu hướng tìm kiếm rất kỹ các sản phẩm sau đó mới chốt thứ họ muốn mua, đây là hành vi mua sắm tích cực khi giúp cả hai bên biết họ cần gì thực sự và nên bán gì thực sự. Hay hoạt động Homepage là sự tồn tại của người dùng trong hệ thống, cho thấy người dùng hay ra vào lại trang bán hàng, chứng tỏ họ có nhu cầu khá lớn cho việc mua sắm. Ngoài ra,

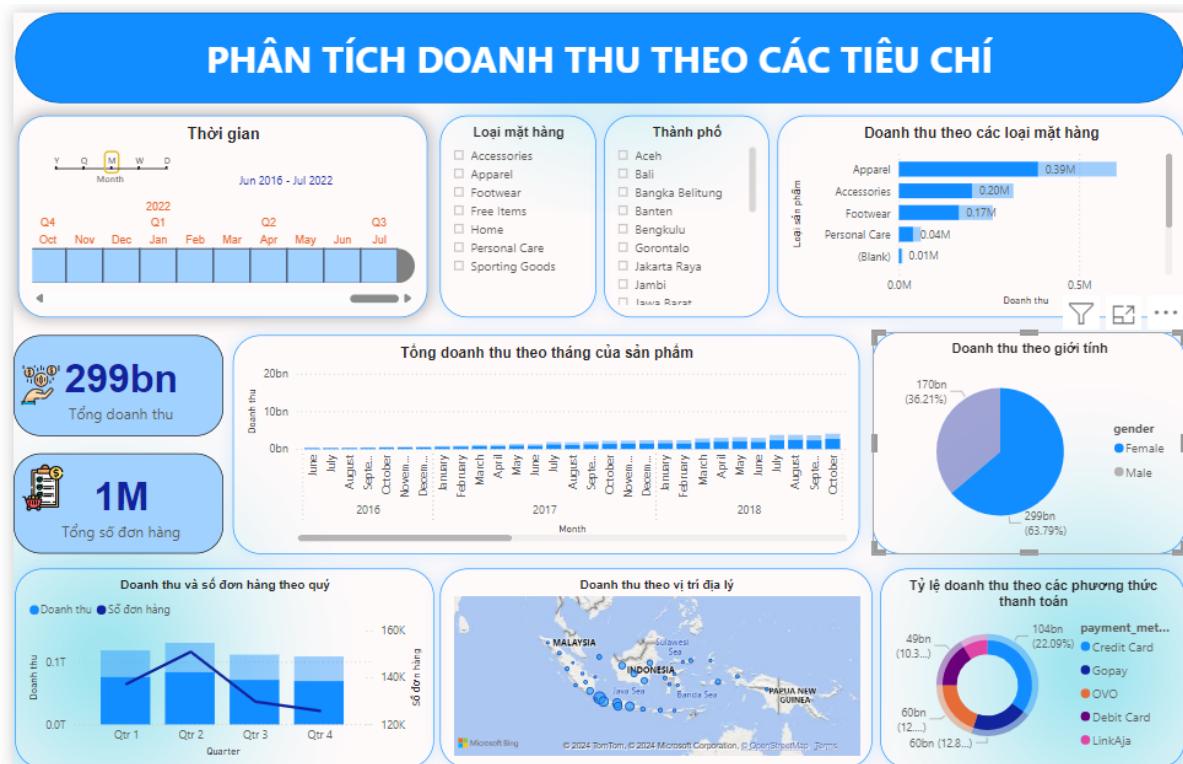
hoạt động Add To Cart (thêm vào giỏ hàng) cũng chiếm số lượng khá lớn, qua đó có thể thấy người dùng họ có nhiều sản phẩm để quan tâm; chỉ với gần 1.8 triệu lượt thao tác nhưng họ cho đến gần 2.8 triệu sản phẩm vào giỏ, chứng tỏ thường một lần họ xem khá nhiều sản phẩm và có được thu hút nhiều hơn 1 sản phẩm mỗi lần họ quan tâm.

Từ khóa được tìm kiếm nhiều nhất: Trong số 1.17 triệu lượt tìm kiếm thì số lượng tìm kiếm sản phẩm "Dress Kongdagan" là lớn nhất. Đây là một mẫu váy khá kinh điển của Indonesia dành cho phụ nữ, là dòng trang phục truyền thống với thiết kế dài nhưng vải không quá dày, khá thoáng và được cách tân hiện đại hoặc truyền thống với đa dạng mẫu mã và dòng váy này cũng có số lượt tìm kiếm nhiều nhất quanh năm; khi liên kết với việc khảo sát dữ liệu - rằng thị phần người tiêu dùng là nữ chiếm nhiều hơn thì điều này là điều phù hợp vì người tiêu dùng phải mua loại sản phẩm phù hợp vs giới tính và nhu cầu của họ. Loại váy này có tính thời đại cao và có thể dùng trong nhiều mục đích, điều này cho thấy dòng sản phẩm này sẽ là nguồn doanh thu chủ chốt. Công ty nên chú ý điểm này để tiếp tục phát triển dòng sản phẩm này trong tương lai.

3.7.3 Dashboard phân tích doanh thu

Mô hình kinh doanh và dòng doanh thu

Chúng ta xây dựng mô hình kinh doanh đơn giản gồm các quy trình theo thứ tự: Xác định vấn đề khách hàng, phân tích các giá trị khách hàng cảm nhận được, nắm bắt các lợi thế cạnh tranh, và lập đội nhóm. Mục đích là khiến người dùng quay lại dùng sản phẩm hoặc dịch vụ của bạn. Với các doanh nghiệp thì họ nhận biết các giá trị giải pháp hữu ích từ phía mô hình kinh doanh chúng ta., nhất là các nhà cung cấp phát triển việc kinh doanh.



Hình 3.50: Dashboard kết quả kinh doanh

Phân tích Dashboard

Doanh thu theo các loại mặt hàng

Từ biểu đồ và căn cứ vào số liệu thống kê được trực quan theo các khoảng thời gian (theo tháng, quý, năm), ta thấy doanh thu của mặt hàng may mặc (Apparel) luôn có số lượng cao nhất trong các loại mặt hàng. Số lượng đơn hàng và tổng doanh thu cũng tăng dần theo thời gian. Từ năm 2016 đến năm 2022 tổng doanh thu tăng từ 3 tỉ Rupiah lên 107 tỉ Rupiah. Điều này phản ánh nhu cầu cần thiết về việc thể hiện bản thân, làm đẹp - mức cao nhất trong tháp nhu cầu Maslow là rất rõ.

Từ đây chúng ta có thể áp dụng các chính sách sinh lợi nhuận để tăng doanh thu trong lĩnh vực may mặc. Nhưng bên cạnh đó cũng cần cải thiện các chiến dịch kinh doanh trong các lĩnh vực khác để doanh thu có thể đạt mức tối ưu. Chúng ta có thể dùng các hình thức như: marketing, giảm giá, khuyến mại, cải tiến chất lượng,...

Tổng doanh thu theo tháng của sản phẩm

Từ biểu đồ và căn cứ vào số liệu thống kê theo tháng của các mặt hàng. Ta thấy các tháng trong những năm gần đây (từ 2019-2022) tổng doanh thu của các loại mặt hàng là tương đối đồng đều. Điều này là do việc tận dụng ưu điểm của Internet để mua hàng trực tuyến, giảm thiểu các chi phí đi lại, hoàn trả, mua hàng,... phức tạp, từ đó triển khai các hệ thống dịch vụ tối ưu kinh phí từ khâu bán ra đến khi mua vào. Lấy số liệu của năm 2022 làm đơn cử, tháng 1 và tháng 6 có tổng doanh thu lần lượt là 15 tỉ Rupiah và 14 tỉ Rupiah.

Cải tiến doanh thu bằng cách báo cáo định kỳ hàng tháng và có kế hoạch hành động hàng tháng để đưa ra tối ưu hóa lợi nhuận.

Doanh thu theo giới tính

Từ biểu đồ và căn cứ vào số liệu thống kê theo độ tuổi của người mua. Chúng ta thấy trong khoảng thời gian từ năm 2016-2022, tỉ lệ nữ giới (64%) mua hàng luôn chiếm nhiều hơn nam giới (36%). Đặc biệt là ở mặt hàng may mặc chiếm đại đa số khi nữ giới mua hàng trên các sàn thương mại điện tử ở Indonesia.

Cải tiến chiến lược:Tạo ra các chiến dịch marketing nhắm vào nhóm giới tính có doanh thu cao để tối ưu hóa doanh thu. Phát triển và cải thiện các sản phẩm dựa trên sở thích và nhu cầu của từng nhóm giới tính. Xem xét các chiến lược giảm giá hoặc khuyến mại cho nhóm giới tính có doanh thu thấp để tăng cường doanh thu.

Doanh thu và số đơn hàng theo quý

Từ biểu đồ và căn cứ vào số liệu thống kê theo đơn hàng với doanh thu tương ứng. Ta thấy trong những năm 2016-2022, chủ yếu trong quý 2 là số đơn hàng đạt số lượng cao nhất so với các quý khác trong năm. Và mặt hàng được ưa thích nhất vẫn là may mặc. Tổng doanh thu tăng khoảng 104 tỉ Rupiah và số đơn hàng tăng từ 11k đơn lên đến 421000 đơn.

Cải tiến chiến lược: Cải thiện trang web và trải nghiệm khách hàng để tăng tỉ lệ chuyển đổi từ lượt truy cập thành đơn hàng. Phát triển chiến lược bán hàng dựa trên phân tích hiệu quả của từng loại sản phẩm và kênh bán hàng. Tăng cường dịch vụ khách hàng để duy trì và tăng cường mối quan hệ với khách hàng hiện tại.

Doanh thu theo vị trí địa lý

Từ biểu đồ và căn cứ vào số liệu thống kê theo vị trí địa lý. Dòng doanh thu của thương mại điện tử ở Indonesia tập trung chủ yếu ở các vùng gần mảng lưỡi sông ngòi hay các vùng ven biển (ở dưới dạng bong bóng, doanh thu càng nhiều thì bong bóng càng lớn). Điều này minh chứng cho việc doanh thu phụ thuộc vào vị trí địa lý. Bong bóng lớn nhất chính là ở thủ đô Jarkarta, đây chính là thủ đô, nơi tập trung các doanh nghiệp lớn, cốt lõi của quốc gia và được

thiết kế cơ sở hạ tầng hiện đại, tiên tiến phục vụ người dân.

Ở phía lục địa, không phải đảo sẽ có bong bóng doanh thu thấp hơn, do địa hình đồi núi hiểm trở, sông ngòi thưa thớt, đầu tư thiết bị, hạ tầng còn yếu kém,...

Minh chứng cho các điều này là bong bóng doanh thu tăng từ năm 2016 đến năm 2022, với độ tăng doanh thu khoảng 104 tỉ Rupiah. Do vậy chúng ta cần có các cải tiến, phân bổ tăng cường vốn đầu tư, cùng với các chính sách khuyến khích, tạo các loại hình kinh doanh mới tận dụng lợi thế của tự nhiên như: du lịch, văn hóa,...

Tỷ lệ doanh thu theo các phương thức thanh toán

Từ biểu đồ và căn cứ vào số liệu thống kê theo các phương thức thanh toán. Chúng ta thấy hình thức thanh toán phổ biến qua các năm (2016-2022) là Credit Card. Điều này cho thấy sự tiện lợi và các cải tiến hình thức thanh toán đa dạng trên các sàn thương mại điện tử.

Chương 4

Mở rộng: Trực quan hóa dữ liệu bằng mô phỏng 3D

4.1 Các công cụ sử dụng

4.1.1 SketchUp

SketchUp là một ứng dụng tạo mô hình 3D trực quan cho phép bạn tạo và chỉnh sửa các mô hình 2D và 3D bằng phương pháp "Kéo thả". SketchUp được sử dụng cho nhiều dự án mô hình 3D như kiến trúc, thiết kế nội thất, kiến trúc cảnh quan và thiết kế trò chơi điện tử, cùng một số công dụng của nó.



Hình 4.1: SketchUp

Lợi ích:

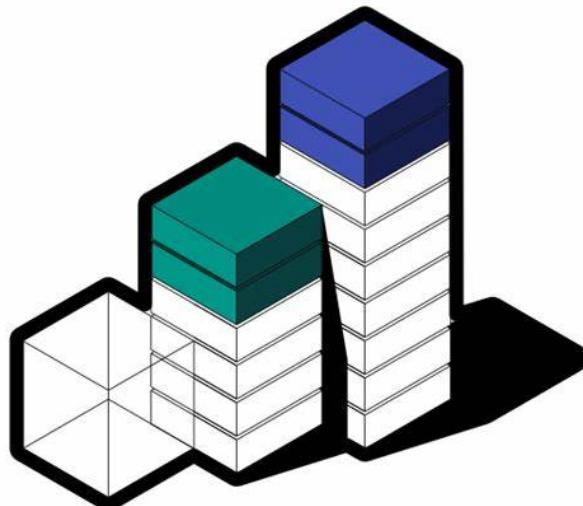
- Các cách làm mô hình: Tạo trên các ứng dụng Desktop, iPad và web.
- Các cách chia sẻ: Quản lý quyền truy cập vào dự án, đánh dấu mô hình và đồng bộ hóa tệp trên đám mây để nhận phản hồi.
- Các cách để xem: Đưa các bên liên quan hoàn toàn vào tầm nhìn của người sử dụng bằng các kế hoạch 2D, hướng dẫn 3D, chế độ xem thực tế mở rộng,...

4.1.2 3DBI

3DBI là công cụ trực quan 3D tùy chỉnh mà Microsoft đã tạo cho Power BI, cho phép bạn trực quan hóa dữ liệu của mình trong bối cảnh không gian của nó.

Các trường hợp sử dụng bao gồm từ BIM (số lượng, lập kế hoạch, ...) đến vẽ đồ thị dữ liệu IoT (màu sắc và lọc các đối tượng theo nhiệt độ, độ ẩm, nồng độ CO₂, công suất sử dụng, độ hao mòn, ...) cũng như Kho bãi (quản lý hàng tồn kho, chọn đơn hàng, ...) và Năng lượng

(biểu đồ sản lượng của các tấm pin mặt trời hoặc trang trại cối xay gió) hoặc bất kỳ lĩnh vực nào khác mà bạn có thể có nhu cầu trực quan hóa dữ liệu trong môi trường 3D.

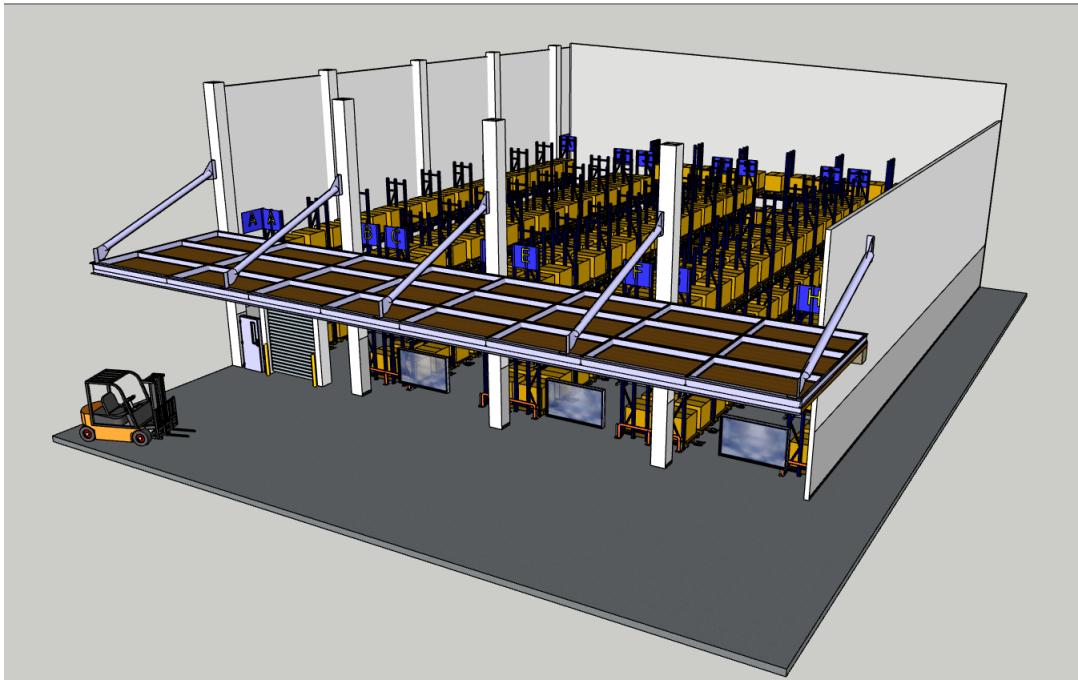


Hình 4.2: 3DBI

Các tính năng độc đáo:

- Dựa trên tệp: việc xuất dữ liệu và tạo hình học phù hợp với Microsoft Power BI được thực hiện cục bộ trên máy của bạn.
- Metadata được xuất dưới dạng tệp .json mở, có thể đọc được. Điều này có nghĩa là việc xuất dữ liệu tương tự cũng có thể được sử dụng trong các gói phần mềm khác có khả năng đọc và giải thích các tệp .json, chẳng hạn như Microsoft Excel. 3DBI dành cho SketchUp xuất tất cả số lượng như khối lượng và diện tích, cũng như tất cả các thuộc tính IFC được đính kèm với các thành phần SketchUp.
- Khả năng độc đáo để bao gồm hình học theo ngữ cảnh tĩnh. Hình học tĩnh luôn hiển thị nhưng không tương tác và dùng để cung cấp ngữ cảnh cho bảng điều khiển trừu tượng. Tất cả các đối tượng được xuất mà không có ID sẽ được coi là hình học tĩnh.
- Bao gồm các khung cảnh (scenes) dưới dạng vị trí camera được cấu hình sẵn.
- Tất cả các tính năng độc đáo đi kèm với hình ảnh 3D tùy chỉnh cho Power BI, chẳng hạn như: ghost object opacity control, shadows, camera FOV,...

4.2 Mục đích thiết kế



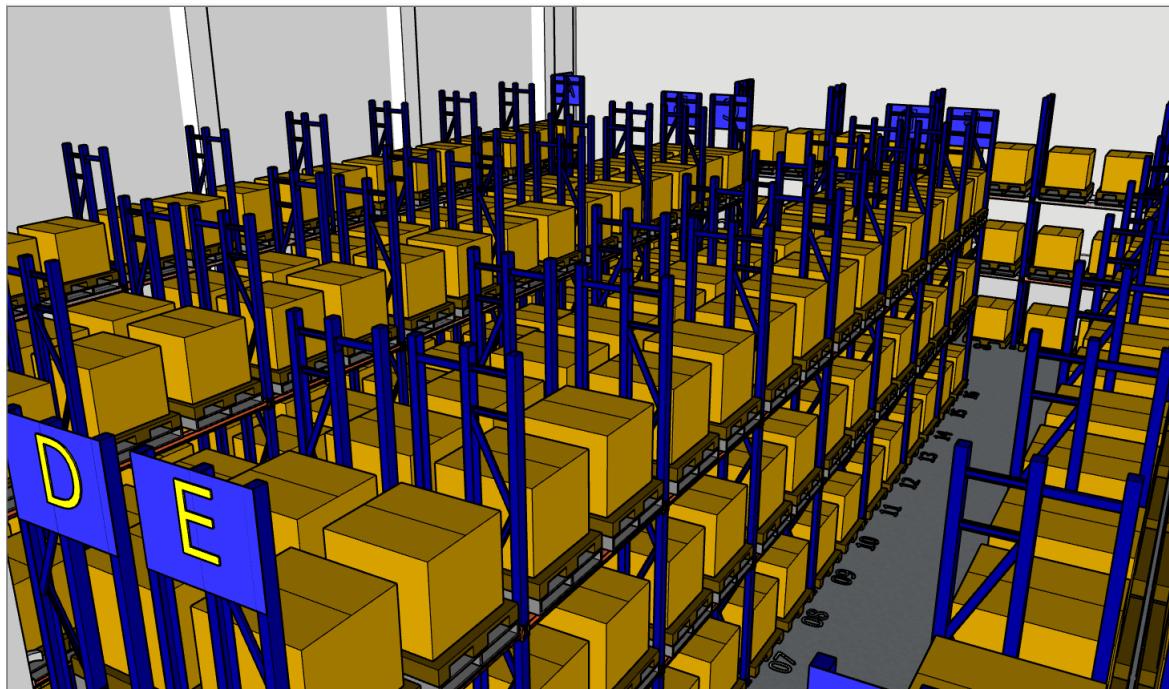
Hình 4.3: Mô hình kho chứa hàng tạo bởi SketchUp và 3DBI

Nhóm thực hiện thiết kế mô hình trực quan hóa dữ liệu, mô phỏng một kho hàng tương ứng trong thực tế. Với các mặt hàng được phân loại theo từng sản phẩm, được thiết kế dưới dạng các hộp màu khác nhau, chứa các thông tin cụ thể về vị trí, thời gian của sản phẩm. Thiết kế kho dữ liệu và trực quan hóa 3D các kệ hàng chứa hộp sản phẩm nhằm cung cấp một giải pháp toàn diện cho việc quản lý kho hàng và tối ưu hóa không gian lưu trữ. Dự án này bao gồm việc tạo ra một mô hình kho ảo sử dụng Power BI và SketchUp, với các mục tiêu chính như sau:

- **Quản lý thông tin sản phẩm chính xác:** Lưu trữ chi tiết về các sản phẩm như mã sản phẩm, tên sản phẩm, số lượng, vị trí trên kệ, ngày nhập kho và hạn sử dụng. Cập nhật thông tin nhanh chóng và chính xác, giúp việc kiểm kê và quản lý hàng tồn kho trở nên dễ dàng hơn.
- **Tối ưu hóa không gian lưu trữ:** Tạo ra mô hình 3D của kho hàng để trực quan hóa vị trí và sắp xếp các hộp sản phẩm trên kệ. Đánh giá và tối ưu hóa cách bố trí kệ hàng nhằm tận dụng tối đa không gian lưu trữ.
- **Nâng cao hiệu quả làm việc:** Giảm thời gian tìm kiếm sản phẩm bằng cách cung cấp thông tin vị trí cụ thể và trực quan. Hỗ trợ nhân viên kho hàng trong việc quản lý và di chuyển sản phẩm hiệu quả.
- **Trực quan hóa dữ liệu:** Sử dụng Power BI để tạo các bảng dữ liệu và báo cáo chi tiết về tình trạng kho hàng. Tích hợp SketchUp và 3DBI để hiển thị mô hình 3D của kho hàng ngay trong Power BI, giúp người dùng dễ dàng theo dõi và kiểm tra.
- **Hỗ trợ ra quyết định:** Cung cấp các báo cáo phân tích, biểu đồ và mô hình 3D giúp quản lý đưa ra quyết định chính xác về việc lưu trữ và phân phối sản phẩm. Dự đoán và lập kế hoạch cho các nhu cầu lưu trữ trong tương lai.

4.3 Các quy trình xây dựng kho dữ liệu 3D

Xây dựng mô hình 3D bằng SketchUp và xác định vị trí của hộp sản phẩm



Hình 4.4: Xác định vị trí của sản phẩm

- **Thiết kế kho hàng:** Sử dụng SketchUp để thiết kế mô hình 3D của kho hàng, bao gồm các kệ hàng và hộp sản phẩm. Đảm bảo rằng mô hình 3D phản ánh chính xác kích thước và vị trí của các kệ hàng và hộp sản phẩm. Tùy chỉnh mô hình 3D bằng việc thêm chi tiết như mã sản phẩm và vị trí trên kệ vào mô hình 3D. Kiểm tra và điều chỉnh mô hình để đảm bảo tính chính xác và thực tế.
- **Gán nhãn sản phẩm:** Cụ thể ta gán nhãn để nhận diện sản phẩm, đồng thời vẽ màu cho hộp sản phẩm.
- **Vị trí:** Các kệ hàng riêng biệt xếp theo nhóm 2 kệ một cạnh nhau, được gán tên kệ bởi các chữ cái A, B, C,... Kệ gồm 3 hàng để xếp sản phẩm, và các cột tương ứng được đánh số bởi các chữ số 1, 2, 3,... Ví dụ A101: là sản phẩm ở kệ A, hàng thứ 1, và ở cột thứ 01 trên kệ.

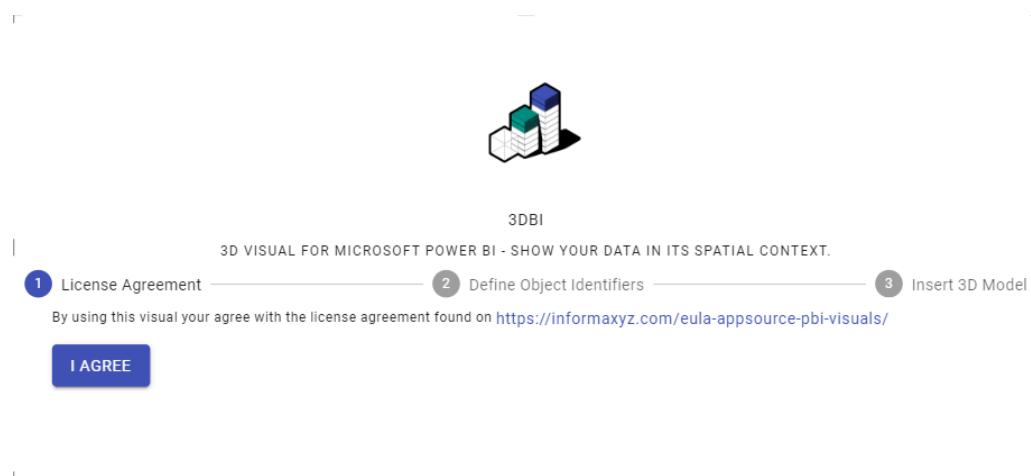
Kết nối Excel Dataset

LocationID	id	gender	masterCategory	subCategory	articleType	baseColour	season	year	usage	productDisplayName
A-101	15970	Men	Apparel	Topwear	Shirts	Navy Blue	Fall	2011	Casual	Turtle Check Men Navy Blue Shirt
C-201	53759	Men	Apparel	Topwear	Tshirts	Grey	Summer	2012	Casual	Puma Men Grey T-shirt
C-202	1855	Men	Apparel	Topwear	Tshirts	Grey	Summer	2011	Casual	Inkfruit Mens Chain Reaction T-shirt
C-302	30805	Men	Apparel	Topwear	Shirts	Green	Summer	2012	Ethnic	Fabindia Men Striped Green Shirt
C-104	26960	Women	Apparel	Topwear	Shirts	Purple	Summer	2012	Casual	Jealous 21 Women Purple Shirt
B-310	12369	Men	Apparel	Topwear	Shirts	Purple	Fall	2011	Formal	Reid & Taylor Men Check Purple Shirts
B-102	42419	Girls	Apparel	Topwear	Tops	White	Summer	2012	Casual	Gini and Jony Girls Knit White Top
C-108	13089	Men	Apparel	Topwear	Sweatshirts	Grey	Fall	2011	Sports	ADIDAS Men Lfc Auth Hood Grey Sweatshirts
E-102	7990	Men	Apparel	Topwear	Tshirts	Navy Blue	Fall	2011	Sports	Fila Men's Round Neck Navy Blue T-shirt
A-105	37812	Men	Apparel	Topwear	Shirts	Navy Blue	Summer	2012	Formal	John Players Men Navy Blue Shirt
A-106	4729	Boys	Apparel	Topwear	Tshirts	Green	Summer	2011	Casual	Disney Kids Boy's Crew Sea Life Sialing Green Teen Kidswear
B-210	56825	Men	Apparel	Topwear	Shirts	Brown	Summer	2012	Casual	John Players Men Brown Shirt
G-104	20099	Women	Apparel	Topwear	Kurtas	Green	Fall	2011	Ethnic	Diva Women Embroidered Green Kurtा
G-108	3954	Women	Apparel	Topwear	Tshirts	Pink	Summer	2011	Casual	Jealous 21 Women's Pink T-shirt
H-103	28690	Women	Apparel	Topwear	Kurtas	Beige	Summer	2012	Ethnic	W Women Printed Beige Kurta
H-104	8580	Men	Apparel	Topwear	Waistcoat	Grey	Fall	2011	Casual	Scullers Men Grey Waistcoat
I-201	9452	Men	Apparel	Topwear	Shirts	Red	Fall	2011	Formal	John Miller Men Stripes White Red Shirts
B-109	45856	Women	Apparel	Topwear	Kurtas	Brown	Summer	2012	Ethnic	Vishudh Women Brown Kurta
E-206	5891	Men	Apparel	Topwear	Tshirts	Black	Summer	2011	Casual	Puma Men's Stripe Polo Black T-shirt
E-207	38630	Women	Apparel	Topwear	Tshirts	Purple	Summer	2012	Casual	Nike Women Purple Polo T-shirt
E-208	4943	Men	Apparel	Topwear	Shirts	White	Summer	2011	Casual	Gini and Jony Boy's Kaleb White Brown Kidswear
E-210	10866	Men	Apparel	Topwear	Tshirts	Red	Fall	2011	Casual	Wrangler Men Motor Rider Red T-Shirts
D-202	17528	Men	Apparel	Topwear	Tshirts	Black	Fall	2011	Casual	Puma Men Scribble Black Tshirts
G-112	9660	Men	Apparel	Topwear	Shirts	Pink	Fall	2011	Casual	Indigo Nation Men Reversible Bling Pink Shirts
G-113	2288	Women	Apparel	Topwear	Tshirts	Black	Fall	2010	Sports	Nike Women Black T-shirt
A-102	23876	Men	Apparel	Topwear	Sweatshirts	Blue	Fall	2011	Casual	ADIDAS Men Blue Sweatshirt
G-204	18237	Men	Apparel	Topwear	Tshirts	Red	Fall	2011	Casual	Manchester United Men Solid Red Tshirt
G-207	49653	Women	Apparel	Topwear	Tops	Green	Summer	2012	Casual	Mineral Women Green Top
G-211	13479	Men	Apparel	Topwear	Tshirts	Pink	Summer	2011	Casual	United Colors of Benetton Men Solid Pink Polo T-shirts
H-207	58513	Women	Apparel	Topwear	Tops	Off White	Summer	2012	Casual	Tonga Women Maroon Top

Hình 4.5: Bảng Product kèm vị trí đã cấu hình cho kệ hàng với trường locationID

- Xác định các yêu cầu dữ liệu:** Thu thập thông tin về các sản phẩm cần lưu trữ, bao gồm mã sản phẩm, tên sản phẩm, số lượng, vị trí trên kệ, ngày nhập kho, hạn sử dụng, kích thước và trọng lượng hộp. Xác định cấu trúc kệ hàng, bao gồm số lượng kệ, số lượng và kích thước các tầng trên mỗi kệ.
- Tạo bảng dữ liệu:** Sử dụng Excel hoặc một công cụ quản lý dữ liệu khác để tạo bảng dữ liệu chứa các trường thông tin đã xác định. Kiểm tra và làm sạch dữ liệu để đảm bảo tính chính xác và nhất quán.
- Xác định cụ thể bảng:** Xây dựng bảng Product chứa các thông tin về các trường như: LocationID, ID, Gender, articleType, season, year, usage,...

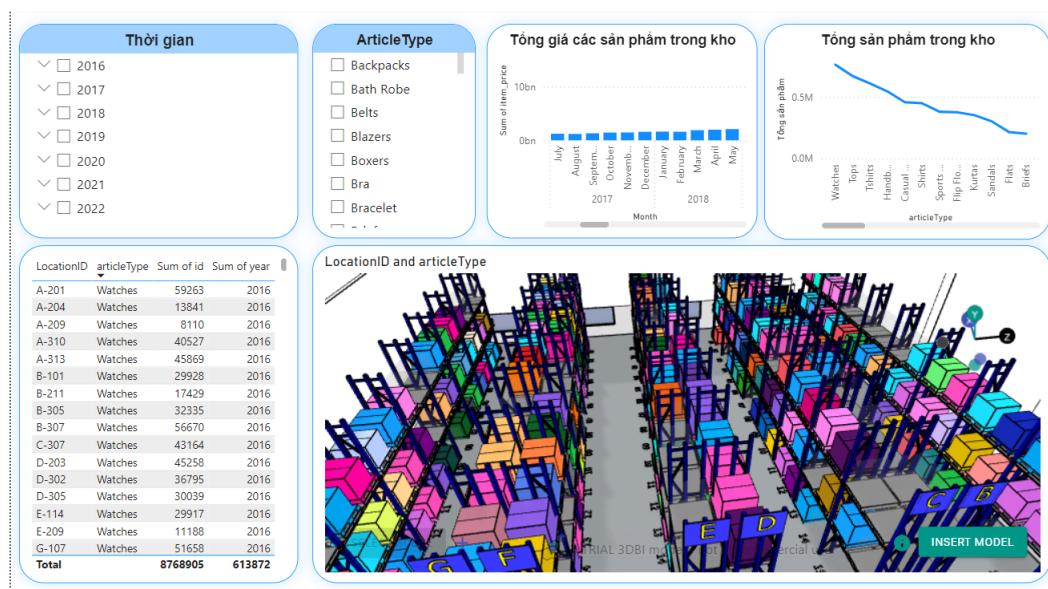
Kết nối Data Set với PowerBI thông qua 3DBI extension và Report



Hình 4.6: Kết nối PowerBI và SketchUP thông qua extension 3DBI

- Nhập dữ liệu vào Power BI:** Tải bảng dữ liệu từ Excel hoặc nguồn dữ liệu khác vào Power BI. Tạo các bảng và mối quan hệ giữa các bảng để tổ chức dữ liệu một cách hợp lý.
- Tích hợp 3DBI:** Sử dụng plugin 3DBI để nhúng mô hình 3D từ SketchUp vào Power BI. Đảm bảo rằng mô hình 3D có thể tương tác được trong Power BI, cho phép người dùng xoay, phóng to, thu nhỏ và xem chi tiết các kệ hàng và sản phẩm.
- Thiết kế các trạng thái kho hàng:** Đề cập đến các chức năng được thực hiện trong mô hình 3D của kho hàng và thể hiện trạng thái của kho hàng như: Mức sử dụng kho, tổng vật liệu, vị trí,...

Xây dựng Dashboard



Hình 4.7: Dashboard mở rộng ứng dụng kho hàng 3D

- **Thiết kế báo cáo:** Tạo các báo cáo và bảng điều khiển trong Power BI để hiển thị thông tin chi tiết về tình trạng kho hàng, bao gồm số lượng sản phẩm, vị trí trên kệ, ngày nhập kho và hạn sử dụng. Sử dụng biểu đồ, bảng và các yếu tố trực quan khác để hiển thị dữ liệu một cách rõ ràng và dễ hiểu.
- **Tích hợp mô hình 3D:** Thêm mô hình 3D vào các báo cáo và bảng điều khiển trong Power BI. Đảm bảo rằng mô hình 3D được cập nhật theo thời gian thực dựa trên dữ liệu trong Power BI.
- **Xây dựng Dashboard chi tiết:** Xây dựng các công cụ Slicer để lọc các dữ liệu theo các trường, phục vụ mục đích kết xuất kho, kiểm tra thông tin nhu cầu nhập/xuất hàng, xem số lượng mặt hàng tồn kho/xuất kho,...Kèm theo đó là mô hình 3D đã thiết kế để theo dõi các tình trạng các mặt hàng một cách trực quan, từ đó xây dựng các luồng công việc hợp lý để xử lý kịp thời.

Kiểm tra và triển khai

- **Kiểm tra hệ thống:** Kiểm tra toàn bộ hệ thống để đảm bảo rằng dữ liệu và mô hình 3D hoạt động chính xác và không có lỗi. Thực hiện các điều chỉnh cần thiết dựa trên phản hồi từ người dùng.
- **Triển khai hệ thống:** Triển khai hệ thống vào môi trường sản xuất, cho phép nhân viên kho hàng và quản lý sử dụng. Đào tạo nhân viên về cách sử dụng hệ thống mới, bao gồm việc nhập dữ liệu, theo dõi tình trạng kho hàng và tương tác với mô hình 3D.

4.4 Lợi ích của việc trực quan hóa 3D

Nhóm thực hiện thiết kế kho dữ liệu (kho hàng) và trực quan hóa 3D bằng 3DBI extension trong Power BI và SketchUp mang lại nhiều lợi ích quan trọng, giúp tối ưu hóa việc quản lý kho hàng và nâng cao hiệu quả hoạt động. Dưới đây là một số lợi ích chính:

- **Quản lý thông tin sản phẩm chính xác:** Cập nhật thông tin theo thời gian thực, dễ dàng theo dõi và cập nhật thông tin sản phẩm như số lượng, vị trí trên kệ, ngày nhập kho và hạn sử dụng. Giảm thiểu sai sót, tăng độ chính xác trong quản lý hàng hóa, giảm thiểu sai sót do nhập liệu thủ công.
- **Tối ưu hóa không gian lưu trữ:** sắp xếp hiệu quả, khoa học, tận dụng tối đa không gian lưu trữ. Dễ dàng bố trí kệ hàng và vị trí sản phẩm để tối ưu hóa không gian khi có sự thay đổi trong kho.
- **Trực quan hóa dữ liệu mạnh mẽ:** Sử dụng Power BI để tạo các báo cáo chi tiết về tình trạng kho hàng, giúp quản lý dễ dàng theo dõi và ra quyết định. Sử dụng mô hình 3D trong Power BI cung cấp cái nhìn tổng quan về kho hàng, giúp người dùng dễ dàng tương tác và kiểm tra.
- **Hỗ trợ ra quyết định:** Cung cấp các công cụ phân tích mạnh mẽ, giúp quản lý dự đoán nhu cầu lưu trữ, lập kế hoạch nhập hàng và phân phối sản phẩm hiệu quả hơn. Dự báo và lập kế hoạch lưu trữ trong tương lai, đảm bảo kho hàng luôn hoạt động hiệu quả.

- **Dễ dàng mở rộng và nâng cấp:** Hệ thống có thể dễ dàng mở rộng để hỗ trợ nhiều kho hàng hơn hoặc thêm các tính năng mới khi cần. Dễ dàng nâng cấp hệ thống để đáp ứng các yêu cầu mới và cải thiện hiệu suất.



Tài liệu tham khảo

1. Nguyễn Danh Tú (2023). *Slide bài giảng Kho dữ liệu và Kinh doanh thông minh*. Khoa Toán - Tin, Đại học Bách khoa Hà Nội.
2. Kimball, R., Ross, M., & Thornthwaite, W. (2013). *The Data Warehouse Toolkit*, 3rd Edition. Wiley.
1. (2023) Nguyễn Danh Tú, *Slide bài giảng Kho dữ liệu và Kinh doanh thông minh*, Khoa Toán - Tin, Đại học Bách khoa Hà Nội
2. Kênh Youtube thầy Nguyễn Danh Tú
3. "Kaggle." <https://www.kaggle.com>.
4. "Step by Step creating 3D Visual Warehouse Report in Power BI using 3DBI ", Yaser Ali Husen, July, 2023 <https://www.youtube.com/watch?v=U4kiRd5KliU&t=308s>
5. "Microsoft Power BI for Beginners Series" YouTube, AlexTheAnalyst, 24 May 2022, <https://www.youtube.com/watch?v=g0m5sEHPUs&list=PLUaB-1hjhk8HqnmK0gQhfmIdCbxwoAoys>.