

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN TOÁN ỨNG DỤNG VÀ TIN HỌC



Báo cáo cá nhân
Học phần: Kho dữ liệu và kinh doanh thông minh

Chủ đề: Electronic Commerce

Giảng viên hướng dẫn: ThS. Nguyễn Danh Tú
Sinh viên thực hiện: Vũ Thị Thu Hoài
Mã lớp: 133598

HÀ NỘI, 7/2022

Mục lục

Lời mở đầu	5
Chương 1. Tổng quan về Data Warehouse	6
1.1 Định nghĩa về Data Warehouse	6
1.2 Tính chất của Data Warehouse	7
1.3 Ưu điểm của Data Warehouse	7
1.4 Nhược điểm của Data Warehouse	8
1.5 Kiến trúc của Data Warehouse	8
1.6 Mô hình dữ liệu đa chiều	11
Chương 2. Tổng quan về Business Intelligence	12
2.1 Khái niệm kinh doanh thông minh	12
2.2 Các thành phần chính	12
2.3 Các bước trong quy trình kinh doanh thông minh	13
2.4 Lợi ích từ các ứng dụng BI	14
2.5 Một số công cụ hỗ trợ BI	15
2.6 Liên hệ giữa BI và DSS	15
Chương 3. Ứng dụng Data Warehouse và BI vào vấn đề thực tế	17
3.1 Giới thiệu về bài toán	17
3.1.1 Đặt vấn đề	17
3.1.2 Khái niệm về thương mại điện tử	18
3.1.3 Các loại mô hình thương mại điện tử chủ yếu	19
3.1.4 Mô hình B2C	20

BÁO CÁO CÁ NHÂN

3.1.5	Lý do khiến bạn dùng thương mại điện tử	21
3.1.6	Những trở ngại khi mua trên thương mại điện tử	22
3.2	Phân tích nghiệp vụ quy trình xử lý mua bán trên sàn thương mại điện tử	23
3.3	Requirements	24
3.4	Bộ dữ liệu gốc	26
3.5	Mô tả một số trường dữ liệu quan trọng	27
3.6	Data Exploration	28
3.7	Phân tích và thiết kế	40
3.7.1	Kiến trúc Data Warehouse	40
3.7.2	Quy trình ETL dữ liệu	41
3.7.3	Kiến trúc hệ thống OLTP	47
3.7.4	Phân tích chiều dữ liệu (Dimension) và chủ điểm (Fact)	49
3.7.5	Thiết kế hệ thống OLAP	51
3.8	Xây dựng Dashboard	54
3.8.1	Phân tích Dashboard dựa trên doanh thu: Sales	54
3.8.2	Phân tích Dashboard dựa trên số lượng đơn đặt hàng: Order quantity	56
3.8.3	Phân tích Dashboard dựa trên lợi nhuận: Profit	57
Kết luận		59
Tài liệu tham khảo		60

Danh sách hình vẽ

1.1	Kiến trúc Data Warehouse 1 tầng	9
1.2	Kiến trúc Data Warehouse 2 tầng	10
1.3	Kiến trúc Data Warehouse 3 tầng	10
1.4	Mô hình OLAP	11
2.1	Quy trình kinh doanh thông minh	14
3.1	Hoạt động mua bán trên sàn thương mại điện tử	18
3.2	Khái niệm thương mại điện tử	18
3.3	Quy mô thị trường B2C trong năm từ 2015 – 2020	20
3.4	Lưu lượng truy cập website của 4 sàn thương mại điện tử được truy cập nhiều	20
3.5	Những loại hàng được bán trên thương mại điện tử	21
3.6	Giao diện mua hàng trên thương mại điện tử	21
3.7	Những trở ngại khi mua trên thương mại điện tử	22
3.8	Quy trình nghiệp vụ trên thương mại điện tử	23
3.9	Một số trường quan trọng trong bộ dữ liệu	27
3.10	Dữ liệu đơn đặt hàng theo ngày	28
3.11	Phân bố khách hàng theo vị trí địa lý	28
3.12	Tỷ lệ phân bố khách hàng theo vị trí địa lý	29
3.13	Mật độ khách hàng	29
3.14	Giá trị trung bình đơn đặt hàng theo tháng	30
3.15	Giá trị đơn hàng theo tỷ lệ	30
3.16	Tỷ lệ hình thức thanh toán	31
3.17	Lượng đơn hàng theo số lần trả góp	31
3.18	Tỷ lệ đơn hàng theo danh mục sản phẩm	32
3.19	Xếp hạng danh mục sản phẩm hàng đầu theo mỗi tháng	33
3.20	Phí giao hàng theo danh mục sản phẩm	33
3.21	Doanh số tích lũy bán hàng theo người bán	34
3.22	Phân bố vị trí những người bán hàng chính	34

BÁO CÁO CÁ NHÂN

3.23 Thời gian hàng được giao kể từ khi khách hàng đặt đơn	35
3.24 Phân bố khoảng cách đặt hàng	36
3.25 Thời gian trung bình đơn hàng được chấp nhận theo hình thức thanh toán	36
3.26 Thời gian vận chuyển theo danh mục sản phẩm	37
3.27 Thời gian giao hàng dự kiến và thực tế	38
3.28 Phân bố tỷ lệ đánh giá của khách hàng	39
3.29 Kiến trúc Data Warehouse cũ	40
3.30 Kiến trúc Data Warehouse mới	41
3.31 Sơ đồ quy trình ETL	42
3.32 Remove duplicate	43
3.33 Xóa giá trị Null	43
3.34 Xóa các cột không sử dụng	44
3.35 Thêm cột thể tích	44
3.36 Thêm cột có điều kiện	45
3.37 Merge query	45
3.38 Chính lại trường dữ liệu thời gian	46
3.39 Đặt lại kiểu dữ liệu cho cột	46
3.40 OLTP model	48
3.41 Mô hình RE OLTP	48
3.42 Phân tích Dimension	50
3.43 Mô hình logic	51
3.44 Mô hình OLAP	52
3.45 Mô hình dữ liệu quan hệ ERD OLAP	53
3.46 Dashboard dựa trên doanh thu	54
3.47 Dashboard dựa trên số lượng đơn đặt hàng	56
3.48 Dashboard dựa trên lợi nhuận	57

Lời mở đầu

Kho dữ liệu và kinh doanh thông minh (Data Warehouse & Business Intelligence) là một trong những học phần quan trọng của ngành công nghệ thông tin. Học phần giúp trang bị cho chúng ta những kiến thức cơ bản nhất về cách thiết kế kho dữ liệu phục vụ các hệ hỗ trợ quyết định, xây dựng các hệ thống kinh doanh thông minh. Từ đó vận dụng các kiến thức vào việc phát triển các ứng dụng thực tế đem lại nguồn lợi nhuận cao cho doanh nghiệp.

Bất kì môn học nào thì cũng đều cần phải có sự kết hợp giữa lý thuyết và thực hành. Việc học lý thuyết giúp chúng ta nắm được những kiến thức cơ bản và hiểu rõ bản chất của kiến thức; thực hành giúp chúng ta vận dụng kiến thức học được vào những bài tập, giúp chúng ta nhớ kiến thức lâu hơn. Và qua quá trình học tập cũng như thực hành nhóm với một vấn đề thực tế cụ thể, nhóm chúng em đã làm bài báo cáo này để tổng kết lại nội dung kiến thức lý thuyết và các phần đã làm được khi xây dựng hệ thống Data Warehouse và Business Intelligence.

Mục tiêu của báo cáo là nhằm cung cấp kiến thức cơ sở lý thuyết về kho dữ liệu và kinh doanh thông minh; hiểu được quá trình để xây dựng lên một hệ thống Data Warehouse hoàn chỉnh và từ đó có thể tạo các dashboard phân tích báo cáo; biết cách sử dụng các công cụ xử lý dữ liệu như Excel, Power Query, Power BI, Python.

Nội dung của bài báo bao gồm: Chương 1, trình bày tổng quan lý thuyết về Data Warehouse. Những kiến thức nền tảng cũng như giới thiệu các công cụ hỗ trợ BI sẽ được trình bày ở chương 2. Chương 3, trình bày từng bước để xây dựng một Data Warehouse ứng với vấn đề thực tế (việc quản lý mua bán trên sàn thương mại điện tử). Cuối cùng, những kết luận và bài học học thu được từ báo cáo được trình bày ngắn gọn trong phần kết luận.

Em xin gửi lời cảm ơn chân thành và sâu sắc nhất tới ThS. Nguyễn Danh Tú, Giảng viên bộ môn Toán tin, Viện Toán ứng dụng và Tin học, Trường Đại học Bách khoa Hà Nội đã luôn tận tình giảng dạy, giúp đỡ và hướng dẫn nhóm trong suốt quá trình hoàn thành báo cáo này.

Chương 1

Tổng quan về Data Warehouse

1.1 Định nghĩa về Data Warehouse

Kho dữ liệu (Data Warehouse) được hiểu là một tập hợp các dữ liệu tương đối ổn định (không hay thay đổi), cập nhật theo thời gian, được tích hợp theo hướng chủ đề nhằm hỗ trợ quá trình tạo quyết định về mặt quản lý (W.H. Inmon).

Kho dữ liệu về bản chất là một cơ sở dữ liệu bình thường, các hệ quản trị cơ sở dữ liệu quản lý và lưu trữ nó như các cơ sở dữ liệu thông thường. Tuy nhiên, nó có thể quản lý dữ liệu lớn và hỗ trợ truy vấn. Nên điểm khác biệt giữa kho dữ liệu và cơ sở dữ liệu là ở quan niệm, cách nhìn nhận vấn đề.

Kho dữ liệu là một loại hệ thống quản lý dữ liệu được thiết kế để kích hoạt và hỗ trợ các hoạt động kinh doanh thông minh (BI), đặc biệt là phân tích. Kho dữ liệu chỉ nhằm mục đích thực hiện các truy vấn và phân tích và thường chứa một lượng lớn dữ liệu lịch sử. Dữ liệu trong kho dữ liệu thường được lấy từ nhiều nguồn như tệp nhật ký ứng dụng và ứng dụng giao dịch.

Kho dữ liệu tập trung và tổng hợp một lượng lớn dữ liệu từ nhiều nguồn. 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. Do những khả năng này, kho dữ liệu có thể được coi là "nguồn sự thật duy nhất của tổ chức".

Một kho dữ liệu điển hình thường bao gồm các yếu tố sau:

- Cơ sở dữ liệu quan hệ để lưu trữ và quản lý dữ liệu
- Giải pháp trích xuất, tải và biến đổi (ELT) để chuẩn bị dữ liệu cho phân tích
- Khả năng phân tích thống kê, báo cáo và khai thác dữ liệu
- Các công cụ phân tích khách hàng để trực quan hóa và trình bày dữ liệu cho người dùng doanh nghiệp

nghiệp

- Các ứng dụng phân tích khác, phức tạp hơn tạo ra thông tin có thể hành động bằng cách áp dụng các thuật toán khoa học dữ liệu và trí tuệ nhân tạo (AI) hoặc các tính năng đồ thị và không gian cho phép nhiều loại phân tích dữ liệu hơn trên quy mô lớn

1.2 Tính chất của Data Warehouse

- **Tính hướng chủ đề (Subject - oriented):** Mục đích của Kho dữ liệu là phục vụ các yêu cầu phân tích, hoặc khai phá cụ thể được gọi là chủ đề. Ví dụ với chủ đề phân tích nhân sự thì có thể bao gồm các độ đo về doanh thu của từng người, số ngày nghỉ trong tháng, số dự án tham gia trong tháng, theo các chiều phân tích: thời gian, chi nhánh, sản phẩm,...

Một sự so sánh dễ hiểu, giống như chẩn đoán một bệnh ví dụ bệnh liên quan đến tim, thì bác sĩ cần quan tâm không chỉ một mà một vài chỉ số như các chỉ số liên quan đến máu, chỉ số về huyết áp, nhịp tim, điện tâm đồ. Ngoài ra còn cần theo dõi theo thời gian (có thể là hàng ngày) để xem xét sự thay đổi mà có phương pháp điều trị kịp thời. Trong trường hợp này thời gian được gọi là chiều phân tích. Để chẩn đoán được chính xác thì cần đầy đủ các thông tin về các chỉ số trên, và cũng không cần các chỉ số khác lẩn vào làm nhiễu quá trình chẩn đoán và cũng không cần thiết. Việc tổ chức dữ liệu theo chủ đề này sẽ dẫn đến nhu cầu tổ chức lưu trữ dữ liệu khác với các cơ sở dữ liệu tác nghiệp.

- **Tính toàn vẹn (Integrated):** Giải quyết các khó khăn trong việc kết hợp dữ liệu từ nhiều nguồn dữ liệu khác nhau, giải quyết các sai khác về tên trường dữ liệu (dữ liệu khác nhau nhưng tên giống nhau), ý nghĩa dữ liệu (tên giống nhau nhưng dữ liệu khác nhau), định dạng dữ liệu (tên và ý nghĩa giống nhau nhưng kiểu dữ liệu khác nhau).
- **Tính bất biến (Nonvolatile):** Quy định rằng dữ liệu phải thống nhất theo thời gian (bằng cách hạn chế tối đa sửa đổi hoặc xóa dữ liệu), từ đó làm tăng quy mô dữ liệu lên đáng kể so với hệ thống nghiệp vụ (5-10 năm so với 2 đến 6 tháng như database thông thường).
- **Giá trị lịch sử (time-varying):** Gắn thời gian và có tính lịch sử: Dữ liệu trong kho dữ liệu bao gồm cả quá khứ và hiện tại. Mỗi dữ liệu trong kho dữ liệu đều được gắn với thời gian và có tính lịch sử.

1.3 Ưu điểm của Data Warehouse

Kho dữ liệu cho phép người dùng doanh nghiệp nhanh chóng truy cập dữ liệu quan trọng từ một số nguồn ở mọi nơi. Kho dữ liệu cung cấp thông tin phù hợp về các hoạt động đa chức năng khác nhau.

Nó cũng hỗ trợ báo cáo và truy vấn đặc biệt. Kho dữ liệu giúp tích hợp nhiều nguồn dữ liệu để giảm căng thẳng chỗ nhiều nguồn sản xuất. Kho dữ liệu giúp giảm tổng thời gian quay vòng để phân tích và báo cáo. Tái cấu trúc và tích hợp giúp người dùng dễ sử dụng hơn để báo cáo và phân tích. Kho dữ liệu chỗ phép người dùng truy cập dữ liệu quan trọng từ số lượng nguồn ở một nơi duy nhất. Do đó, nó giúp tiết kiệm thời gian lấy dữ liệu của người dùng từ nhiều nguồn. Kho dữ liệu lưu trữ một lượng lớn dữ liệu lịch sử. Điều này giúp người dùng phân tích các khoảng thời gian và xu hướng khac nhau để đưa ra dự đoán trong tương lai.

Tóm lại, một số lợi ích của Data Warehouse như sau:

- Dữ liệu sau khi đưa vào kho dữ liệu đều tuân theo những quy tắc thống nhất.
- Dữ liệu được tổ chức tạo thuận lợi cho việc truy vấn phân tích và tạo tiền đề để đưa ra những quyết định có ảnh hưởng lớn.
- Cải thiện tính bảo mật và hiệu suất mà không cần tách động tới hệ thống dữ liệu gốc.
- Công việc kinh doanh trở nên thông minh hơn, nâng cao dịch vụ khách hàng.

1.4 Nhược điểm của Data Warehouse

Không phải là một lựa chọn lý tưởng chỗ dữ liệu phi cấu trúc. Kho dữ liệu có thể bị lỗi thời tương đối nhanh. Khó thực hiện thay đổi về kiểu và phạm vi dữ liệu, lược đồ nguồn dữ liệu, chỉ mục và truy vấn. Kho dữ liệu có vẻ dễ dàng, nhưng thực sự, nó quá phức tạp đối với người dùng trung bình. Mặc dù có những nỗ lực tốt nhất trong quản lý dự án, phạm vi dự án kho dữ liệu sẽ luôn tăng. Đôi khi người dùng kho sẽ phát triển các quy tắc kinh doanh khac nhau. Các tổ chức cần dành nhiều nguồn lực chỗ mục đích đào tạo và thực hiện. Tương lai của kho dữ liệu Thay đổi các ràng định có thể hạn chế khả năng kết hợp nguồn dữ liệu khac nhau. Những nguồn khac nhau này có thể bao gồm dữ liệu phi cấu trúc rất khó lưu trữ. Khi kích thước của cơ sở dữ liệu tăng lên, các ước tính về những gì tạo nên một cơ sở dữ liệu rất lớn tiếp tục phát triển. Việc xây dựng và chạy các hệ thống kho dữ liệu tăng kích thước là rất phức tạp. Các tài nguyên phần cứng và phần mềm có sẵn ngày hôm nay không chỗ phép giữ một lượng lớn dữ liệu trực tuyến. Dữ liệu đa phương tiện không thể dễ dàng thao tác dưới dạng dữ liệu văn bản, trong khi thông tin văn bản có thể được truy xuất bằng phần mềm quan hệ hiện có.

1.5 Kiến trúc của Data Warehouse

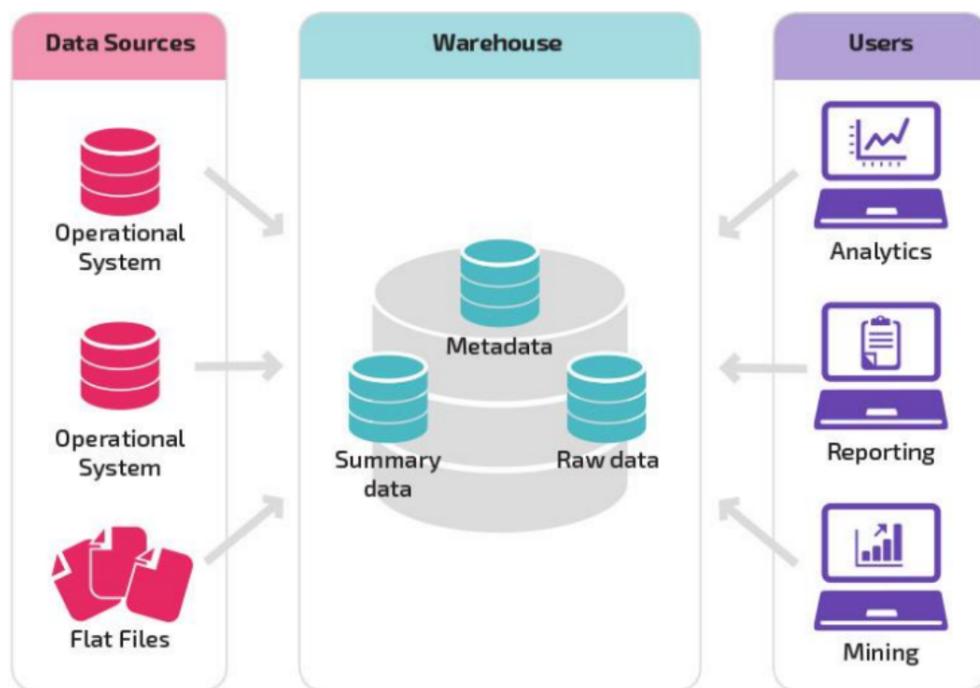
Có 3 loại kiến trúc phổ biến:

- Kiến trúc Data Warehouse cơ bản (kiến trúc 1 tầng).

- Kiến trúc Data Warehouse với vùng dữ liệu Staging (kiến trúc 2 tầng).
- Kiến trúc Data Warehouse với vùng dữ liệu Staging và Data Marts (kiến trúc 3 tầng) - đây cũng là kiến trúc phổ biến nhất trong 3 loại.

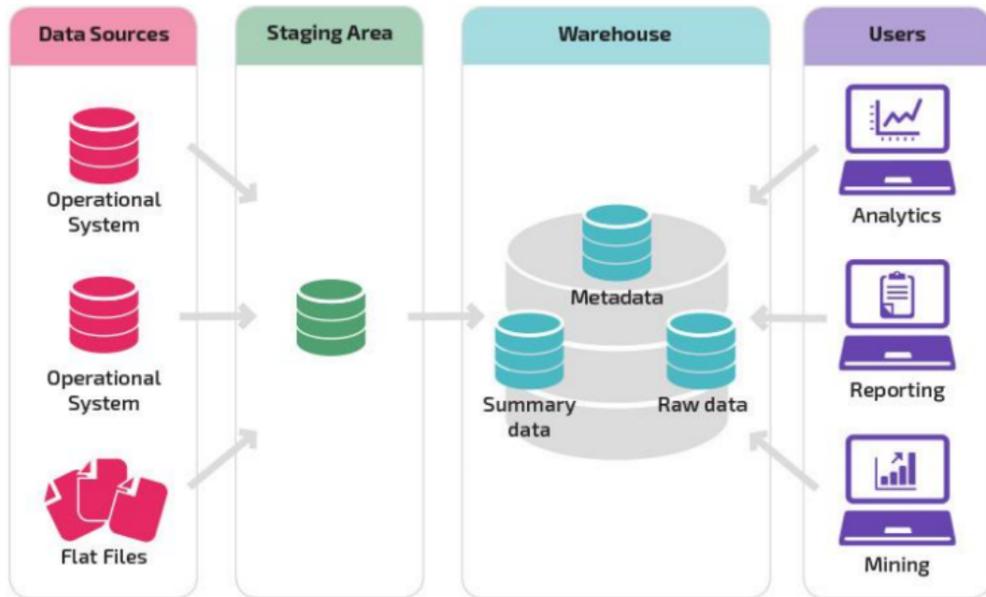
1. Kiến trúc Data Warehouse 1 tầng gồm 3 phần cơ bản:

- Data Source: các nguồn dữ liệu quan hệ hoặc phi quan hệ.
- Warehouse: nơi lưu trữ dữ liệu đã được xử lý, gồm Metadata, Raw Data và Summary Data.
- Users: Gồm các hệ thống phân tích, báo cáo và khai phá dữ liệu.



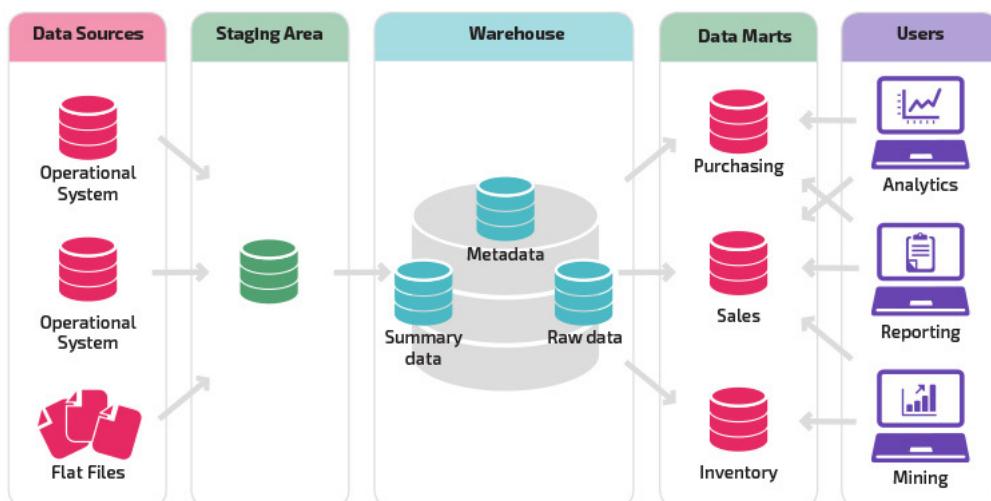
Hình 1.1: Kiến trúc Data Warehouse 1 tầng

2. Kiến trúc Data Warehouse 2 tầng: có thêm bước chuyển dạng và tích hợp dữ liệu. Dữ liệu trước khi đưa vào Data Warehouse, được tập hợp từ nhiều nguồn, chuyển đổi dạng và lưu trữ tại bước Staging Area, người dùng cuối truy xuất dữ liệu trực tiếp từ các hệ thống xử lý nghiệp vụ thông qua Data Warehouse.



Hình 1.2: Kiến trúc Data Warehouse 2 tầng

3. Kiến trúc Data Warehouse 3 tầng



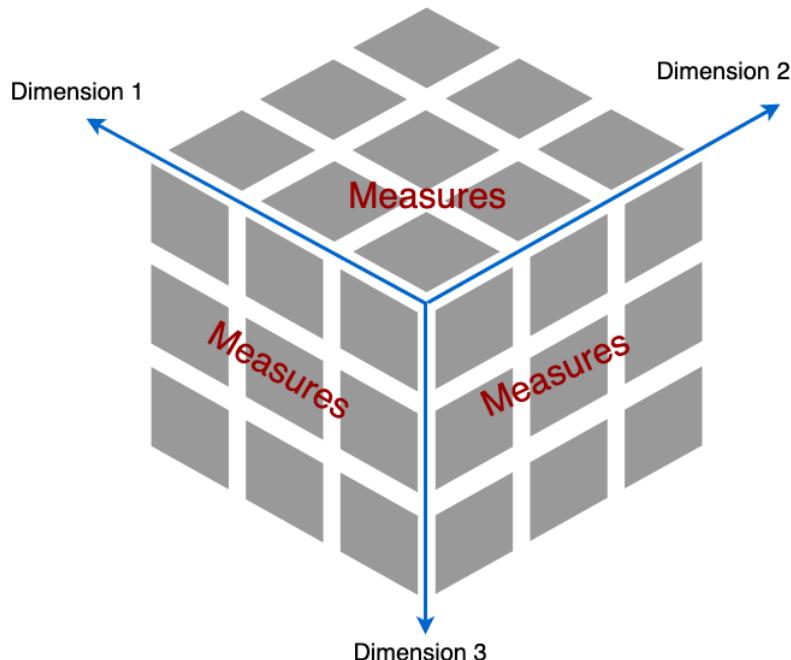
Hình 1.3: Kiến trúc Data Warehouse 3 tầng

Trong kiến trúc Data Warehouse thứ 3, dữ liệu từ các nguồn khác nhau được tập hợp và chuyển đến vùng dữ liệu Staging. Ở vùng dữ liệu Staging, chúng ta sẽ làm các tác vụ chuyển đổi tất cả dữ liệu khác nhau thành các dữ liệu giống nhau về tên, ý nghĩa, và kiểu dữ liệu. Sau đó, tiếp tục load dữ liệu vào Data Warehouse để tiến hành phân tích. Sau khi phân tích với mục đích kinh doanh cụ thể (như là Purchasing, Sales, Inventory ...) dữ liệu sẽ được phân thành các data marts. Ở bước cuối cùng, người dùng có thể tương tác với data marts qua báo cáo, analytics hoặc data mining.

1.6 Mô hình dữ liệu đa chiều

Data Warehouse và các hệ thống OLAP được xây dựng theo mô hình dữ liệu đa chiều (multi-dimensional model). Dữ liệu trong kho dữ liệu được thể hiện dưới dạng đa chiều (Multi Dimension) gọi là khối (cube). Mỗi chiều mô tả một đặc trưng nào đó của dữ liệu (Nếu số chiều dữ liệu lớn hơn 3, gọi là Hyper Cube).

OLAP (Online Analytical Processing) - hệ thống xử lý phân tích trực tuyến là một loại phần mềm cho phép người dùng phân tích thông tin từ nhiều hệ thống cơ sở dữ liệu cùng một lúc. Đây là một công nghệ cho phép các nhà phân tích trích xuất và xem dữ liệu kinh doanh từ các quan điểm khác nhau. Các nhà phân tích thường xuyên cần nhóm, tổng hợp và kết hợp dữ liệu. Các hoạt động này trong cơ sở dữ liệu quan hệ sử dụng nhiều tài nguyên. Với OLAP, dữ liệu có thể được tính toán trước và tổng hợp trước, giúp phân tích nhanh hơn. Cơ sở dữ liệu OLAP được chia thành một hoặc nhiều khối. Các hình khối được thiết kế theo cách mà việc tạo và xem các báo cáo trở nên dễ dàng.



Hình 1.4: Mô hình OLAP

- **Chiều (Dimension):** Chiều cung cấp các thông tin, ngữ cảnh của bảng Fact. Ví dụ về Dimension như: thời gian (ngày, tháng, năm); danh sách các thành phố, quốc gia, khu vực; tên các loại sản phẩm, ...
- **Độ đo (Measure):** Độ đo là đại lượng có thể tính toán được trên các thuộc tính của bảng Fact.

Chương 2

Tổng quan về Business Intelligence

2.1 Khái niệm kinh doanh thông minh

Kinh doanh thông minh (Business Intelligence - BI) là quy trình/hệ thống công nghệ cho phép phân tích và thể hiện thông tin giúp cho các nhà quản lý và người sử dụng của tổ chức đưa ra các quyết định phù hợp.

Kinh doanh thông minh (BI) bao gồm các chiến lược và công nghệ được các doanh nghiệp sử dụng để phân tích dữ liệu, thông tin kinh doanh. Công nghệ BI cung cấp các quan điểm lịch sử, hiện tại và dự đoán về hoạt động kinh doanh. Các chức năng phổ biến của công nghệ thông minh kinh doanh bao gồm báo cáo, xử lý phân tích trực tuyến, phân tích, phát triển bảng điều khiển, khai thác dữ liệu, khai thác quy trình, xử lý sự kiện phức tạp, quản lý hiệu suất kinh doanh, đo điểm chuẩn, khai thác văn bản, phân tích dự đoán và phân tích mô tả.

2.2 Các thành phần chính

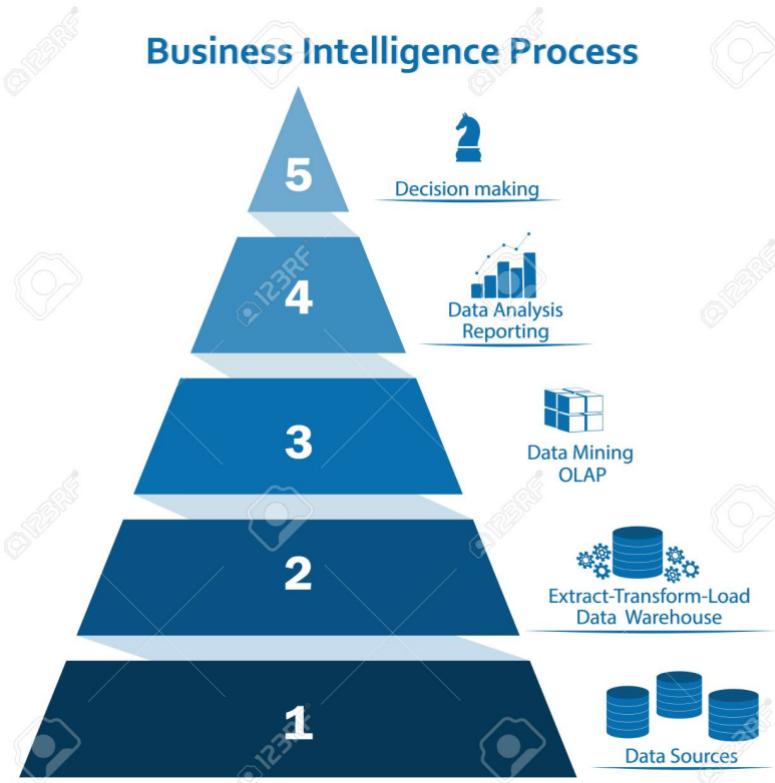
- **Data Source:** là vị trí bắt nguồn dữ liệu đang được sử dụng. Nguồn dữ liệu có thể là vị trí ban đầu nơi dữ liệu được sinh ra hoặc nơi thông tin vật lý được số hóa lần đầu tiên. Tuy nhiên, ngay cả những dữ liệu tinh tế nhất cũng có thể đóng vai trò là nguồn, miễn là một quy trình khác truy cập và sử dụng nó. Cụ thể, nguồn dữ liệu có thể là một cơ sở dữ liệu đến từ các hệ quản trị cơ sở dữ liệu MySQL, SQL, Oracle, MSSQL, một tệp phẳng, các phép đo trực tiếp từ các thiết bị vật lý, dữ liệu web có sẵn hoặc bất kỳ dịch vụ dữ liệu trực tuyến và tinh nào có rất nhiều trên internet ...
- **Data Warehouse:** Là cơ sở dữ liệu được thiết kế theo mô hình khác với cơ sở dữ liệu OLTP thông thường (Online Transaction Processings - OLTP là thiết kế CSDL dành cho việc đọc ghi

thường xuyên, lượng dữ liệu cho mỗi lần đọc ghi ít) và là nơi lưu trữ dữ liệu lâu dài của tổ chức. Dữ liệu của Data Warehouse chỉ có thể đọc, không được sử dụng để ghi hay update bởi ứng dụng thông thường, nó chỉ được cập nhật/ghi bởi công cụ ETL (Extract, Transform, Load) - công cụ chuyển đổi dữ liệu từ nguồn dữ liệu (Data Sources) vào kho dữ liệu (Data Warehouse).

- **Integrating Server:** Chịu trách nhiệm trung gian vận hành công cụ ETL để chuyển đổi dữ liệu từ nguồn dữ liệu vào kho dữ liệu.
- **Analysis Server:** Chịu trách nhiệm thực thi các cube được thiết kế dựa trên các chiều dữ liệu và tri thức nghiệp vụ khôi, chịu trách nhiệm nhận dữ liệu đầu vào từ Data Warehouse và thực thi theo nghiệp vụ định nghĩa sẵn để trả về kết quả.
- **Reporting Server:** Thực thi các báo cáo với đầu ra nhận được từ máy chủ phân tích (Analysis Server). Nơi quản trị tập trung các báo cáo trên nền web, các báo cáo này có thể được đính kèm vào ứng dụng web, hay tầng ứng dụng.
- **Data Mining:** Là quá trình trích xuất thông tin dữ liệu đã qua xử lý (phù hợp với yêu cầu riêng của doanh nghiệp) từ kho dữ liệu (DW) rồi kết hợp với các thuật toán để đưa ra (hoặc dự đoán) các quyết định có lợi cho việc kinh doanh của doanh nghiệp. Đây là một quá trình quan trọng trong BI, thông thường một doanh nghiệp muốn sử dụng giải pháp BI thường kèm theo khai thác dữ liệu (Data Mining).
- **Data Presentation:** Tạo ra các báo cáo, biểu đồ từ quá trình khai thác dữ liệu (Data Mining) để phục vụ cho nhu cầu của người dùng cuối.

2.3 Các bước trong quy trình kinh doanh thông minh

Hệ thống kinh doanh thông minh được thực hiện thông qua quy trình gồm các bước sau: Data sources (Nguồn dữ liệu) → Kho dữ liệu và khối dữ liệu → Business intelligence methodologies (Phương pháp kinh doanh thông minh) → Data exploration (Thăm dò dữ liệu) → Data mining (Khai phá dữ liệu) → Optimization(Tối ưu hóa) → Decisions(Quyết định).



Hình 2.1: Quy trình kinh doanh thông minh

2.4 Lợi ích từ các ứng dụng BI

BI giúp doanh nghiệp kiểm soát thông tin một cách chính xác, hiệu quả từ đó có thể hỗ trợ phân tích, khai thác dữ liệu, dự đoán về xu hướng của giá cả dịch vụ, hành vi khách hàng, phát hiện khách hàng tiềm năng để đề ra các chiến lược kinh doanh phù hợp nhằm gia tăng khả năng cạnh tranh của doanh nghiệp.

Có thể kể đến một số lợi ích thiết thực doanh nghiệp dễ dàng nhận thấy thông qua việc ứng dụng Business Intelligence như:

- Giúp các doanh nghiệp sử dụng thông tin một cách hiệu quả, chính xác để thích ứng với môi trường thay đổi liên tục và cạnh tranh khốc liệt trong kinh doanh.
- Hỗ trợ nhà quản trị tối đa trong việc đưa ra các quyết định kinh doanh nhanh chóng, kịp thời, hiệu quả.
- Xác định được vị thế và khả năng cạnh tranh của doanh nghiệp.
- Phân tích hành vi khách hàng.

- Xác định mục đích và chiến lược Marketing.
- Dự đoán tương lai của doanh nghiệp.
- Xây dựng chiến lược kinh doanh.
- Giữ chân được khách hàng cũ và dự đoán khách hàng tiềm năng.
- Đáp ứng nhu cầu thu thập báo cáo của các bộ phận.
- Cung cấp cái nhìn tổng thể toàn doanh nghiệp.
- Hỗ trợ tối đa công tác điều hành, tiết kiệm thời gian và chi phí cho quản trị.
- Góp phần thay đổi kỹ năng điều hành, phục vụ khách hàng tốt hơn.
- Tạo lợi thế cạnh tranh, gia tăng cơ hội tìm kiếm và nắm bắt các cơ hội kinh doanh.
- Hỗ trợ người dùng nội bộ trong đánh giá, cải thiện và tối ưu hóa khả năng cũng như quy trình hoạt động của tổ chức.

2.5 Một số công cụ hỗ trợ BI

- Power BI
- FineReport
- QlikView
- Sisense
- Tableau

2.6 Liên hệ giữa BI và DSS

DSS (Decision Support System) nghĩa là hệ thống hỗ trợ ra quyết định hoặc hệ hỗ trợ quyết định. DSS được tạo ra với mục đích hỗ trợ trong việc đưa ra quyết định cũng như đưa ra những dự đoán, chiều hướng hành động của một tổ chức hoặc một doanh nghiệp.

DSS hay hệ hỗ trợ ra quyết định là một hệ thống phần mềm tương tác và thu thập mọi thông tin liên quan từ rất nhiều nguồn như vận hành, thu nhập, chi phí, thị trường, xu hướng, mô hình doanh nghiệp. Sau đó DSS sẽ sàng lọc và phân tích những dữ liệu này, tổng hợp lại thành các thông tin một cách toàn diện từ đó có thể sử dụng để hỗ trợ trong việc đưa ra quyết định.

BÁO CÁO CÁ NHÂN

DSS có thể được điều khiển thủ công bởi con người hoặc máy tính hóa hoàn toàn. Trong một vài trường hợp, DSS có thể kết hợp cả hai phương pháp với nhau. Hệ thống DSS lý tưởng sẽ phân tích dữ liệu và tổng hợp thông tin để tăng độ chính xác cũng như tốc độ trong việc đưa ra quyết định.

Liên quan giữa BI và DSS:

- Kiến trúc khá tương đồng - BI sinh ra từ DSS.
- DSS hỗ trợ trực tiếp việc ra quyết định. BI cung cấp thông tin, phân tích - hỗ trợ gián tiếp việc ra quyết định.
- BI thường hướng tới người dùng tác nghiệp - lãnh đạo và quản lý. DSS hướng tới chuyên gia phân tích.
- Các hệ BI thường xây dựng từ các cấu phần (thương mại hay mã mở), trong khi DSS thường được xây dựng từ đầu, chuyên cho bài toán và tổ chức cụ thể.
- Nhiều công cụ trong BI cũng dc sử dụng trong DSS.

Chương 3

Ứng dụng Data Warehouse và BI vào vấn đề thực tế

3.1 Giới thiệu về bài toán

3.1.1 Đặt vấn đề

Công nghệ thông tin và truyền thông ngày càng phát triển và góp phần làm thay đổi diện mạo nền kinh tế, tạo ra lĩnh vực thương mại mới đó là thương mại điện tử. Nhờ sức mạnh của thông tin số hóa mà mọi hoạt động thương mại truyền thống ngày nay đã được tiến hành trực tuyến giúp các bên tham gia vào hoạt động này tiết kiệm được chi phí, thời gian, tăng hiệu suất và nâng cao năng lực cạnh tranh.

Trước kia muốn mua một quyển sách thì bạn đọc phải ra tận cửa hàng để tham khảo, chọn mua một cuốn sách mà mình mong muốn, sau khi đã chọn được cuốn sách cần mua thì người đọc phải ra quầy thu ngân để thanh toán mua cuốn sách đó. Nhưng giờ đây, với sự ra đời của thương mại điện tử, chỉ cần có một chiếc máy tính nối mạng Internet, thông qua vài thao tác kích chuột, người đọc không cần biết mặt của người bán hàng thì họ vẫn có thể mua một cuốn sách mình mong muốn trên các website mua bán trực tuyến như amazon.com; vinabook.com.vn, tiki, shope,...

Bên cạnh đó, một cá nhân tại Việt Nam còn có thể mua được một sản phẩm từ một gian hàng ảo tại Mỹ, hay ngồi tại nhà người đó cũng có thể kê khai các thủ tục hải quan điện tử để tiến hành nhập khẩu sản phẩm. Từ đó ta có thể thấy những lợi ích to lớn của thương mại điện tử đối với nền kinh tế nước ta.



Hình 3.1: Hoạt động mua bán trên sàn thương mại điện tử

3.1.2 Khái niệm về thương mại điện tử

E – Commerce (Electronic commerce) hay còn được gọi là thương mại điện tử – là các hoạt động mua hoặc bán các sản phẩm thông qua dịch vụ trực tuyến. Thương mại điện tử tiện lợi đến mức bạn có thể mua bán sản phẩm trên toàn thế giới ở bất kì thời gian nào. Đây chính là điều mà các cửa hàng truyền thống không thể có được.



Hình 3.2: Khái niệm thương mại điện tử

Một số các ngành nghề sử dụng E – Commerce phổ biến là thương mại di động, chuyển tiền điện tử, quản lý chuỗi cung ứng, tiếp thị qua Internet, giao dịch trực tuyến, trao đổi dữ liệu điện tử (EDI), hệ thống quản lý hàng tồn kho, ...

3.1.3 Các loại mô hình thương mại điện tử chủ yếu

Doanh nghiệp đến doanh nghiệp (B2B)

- Là giao dịch giữa doanh nghiệp với doanh nghiệp, không có sự tham gia của người tiêu dùng. Một số ví dụ về sàn TMĐT B2B như Alibaba.com, Amazon.com, nơi tập trung buôn bán giữa hàng trăm ngàn doanh nghiệp trên toàn thế giới với nhau. Các sàn này giúp kết nối các doanh nghiệp toàn cầu, giúp việc giao dịch, mua bán dễ dàng đồng thời tiết kiệm chi phí tiếp thị và quảng cáo. GoSELL cũng vinh dự là đối tác chiến lược của Alibaba.com tại Việt Nam.

Kinh doanh cho người tiêu dùng (B2C)

- Công ty bán hàng hóa hay dịch vụ trực tiếp cho người tiêu dùng, người tiêu dùng vào các trang web của họ, xem đọc các nhận xét sau đó đặt hàng và công ty chuyển hàng trực tiếp cho họ. Ví dụ về các doanh nghiệp B2C ở Việt Nam như các nhà bán lẻ trực tuyến độc quyền bao gồm Elise, HoangPhuc, Bibomart, Nike, Adidas... Lợi ích mà mô hình này đem lại đối với các doanh nghiệp này chính là tiết kiệm chi phí bán hàng, khi chỉ cần xây dựng một website thương mại điện tử có khả năng tiếp xúc được lượng khách hàng khổng lồ qua internet, không mất tiền thuê mặt bằng, người bán hàng...
- Người tiêu dùng cũng sẽ thoải mái hơn trong việc lựa chọn sản phẩm và thực hiện mua hàng với các thao tác nhanh chóng, sản phẩm được giao tới tận nhà, không mất thời gian đi lại.

Người tiêu dùng đối với người tiêu dùng(C2C)

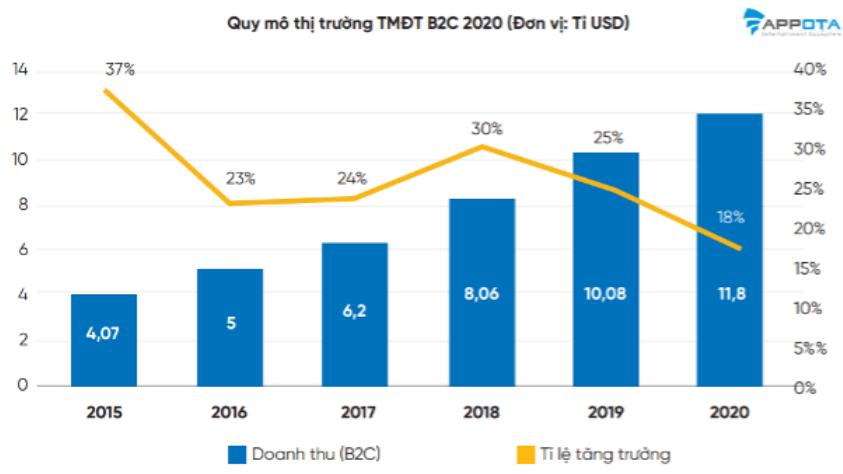
- Hoạt động như các trang trao đổi, mua bán, đấu giá qua internet trong đó người dùng bán hàng hóa cho nhau. Đây có thể là những sản phẩm họ làm ra, chẳng hạn như đồ thủ công hoặc đồ cũ mà họ sở hữu và muốn bán.
- Như vậy có thể thấy mô hình C2C có đại diện phía bên mua và bán đều là các cá nhân và họ thường giao giao dịch trực tuyến với nhau thông qua các sàn thương mại điện tử hay các website đấu giá trung gian. Vd: bạn mua sản phẩm trên Facebook, website của cá nhân nào đó.

Người tiêu dùng đến doanh nghiệp (C2B)

- Là khi người tiêu dùng bán hàng hóa hoặc dịch vụ cho các doanh nghiệp. Khi người tiêu dùng tạo ra giá trị cho doanh nghiệp, đó là thương mại C2B. Tạo giá trị có thể có nhiều hình thức. Chẳng hạn, C2B có thể đơn giản như một khách hàng để lại đánh giá tích cực cho một doanh nghiệp hoặc một trang web nhiếp ảnh mua hình ảnh từ các nhiếp ảnh gia tự do. Ngoài ra, C2B còn là các doanh nghiệp bán hàng secondhand đôi khi mua hàng hóa từ những người dùng internet bình thường.

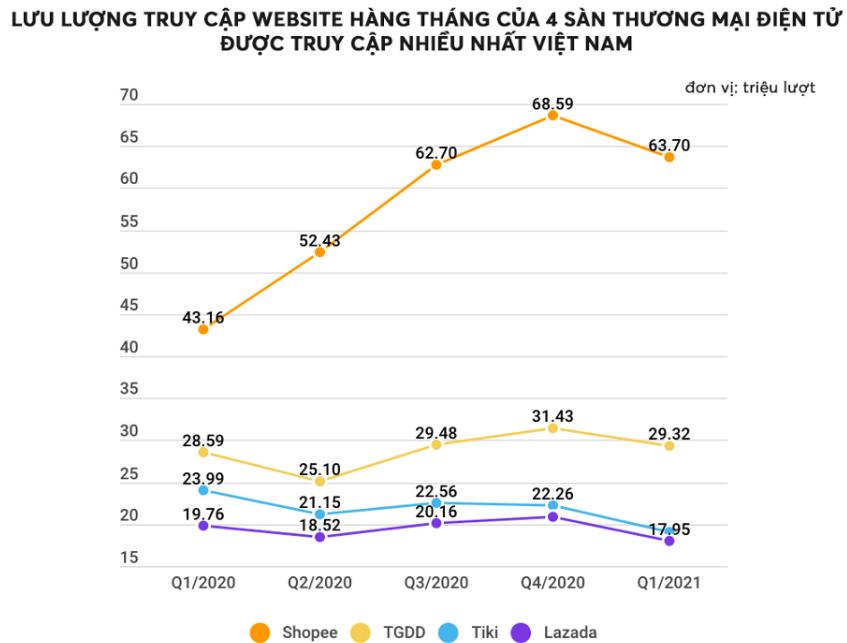
3.1.4 Mô hình B2C

Trong bài báo cáo này, ta xét chủ yếu về mô hình B2C.



Hình 3.3: Quy mô thị trường B2C trong năm từ 2015 – 2020

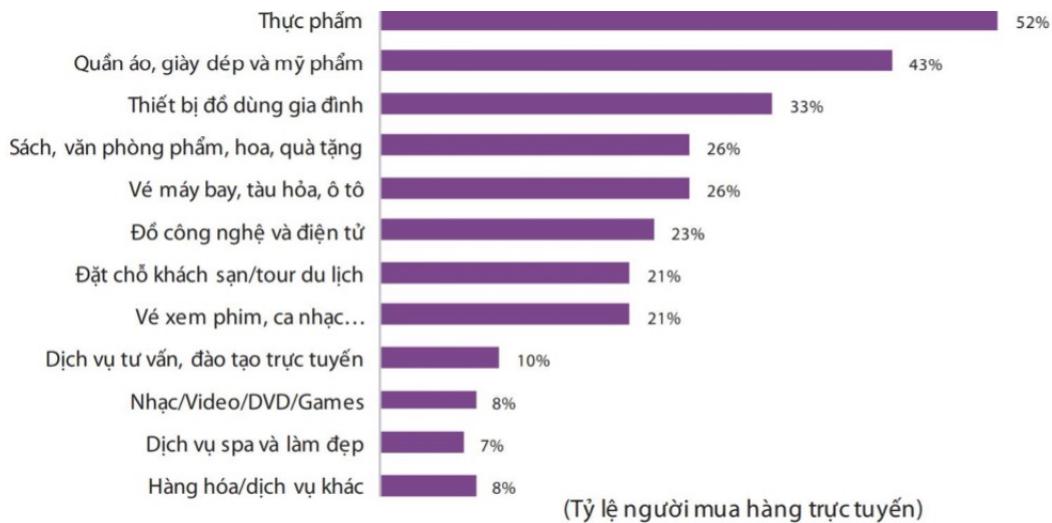
Nhìn vào đồ thị trên, ta có thể thấy doanh thu tăng đều qua các năm và tỷ lệ khá là bền vững qua các năm. Ta có thể thấy mô hình này khá được ưa chuộng và phát triển mạnh mẽ ở Việt Nam. Thị trường thương mại điện tử rất phát triển và được sự tham gia của nhiều doanh nghiệp lớn trong và ngoài nước. Trong đó nổi bật nhất là 5 sàn thương mại điện tử hàng đầu như là Shopee, Lazada, Tiki, Sendo,... Dưới đây ta có biểu đồ lượng truy cập website hàng tháng của 4 sàn thương mại điện tử được truy cập nhiều nhất.



Hình 3.4: Lưu lượng truy cập website của 4 sàn thương mại điện tử được truy cập nhiều

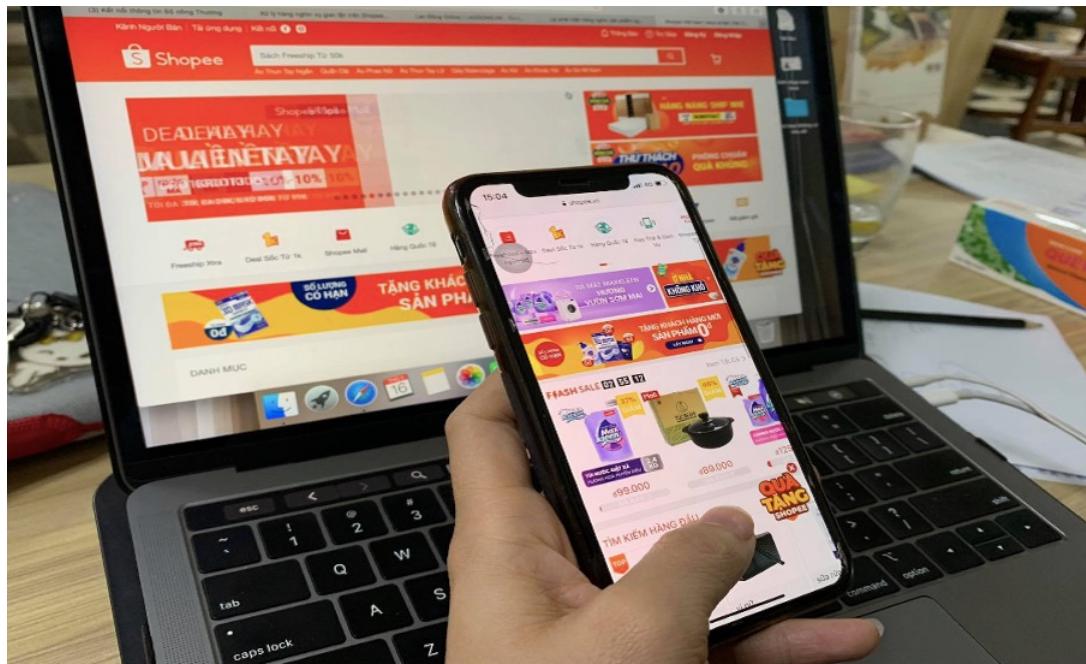
BÁO CÁO CÁ NHÂN

Nhìn vào biểu đồ trên, ta thấy Shopee chiếm lượng truy cập cao nhất, điều này cũng dễ hiểu bởi vì Shopee khá là quen thuộc và đa dạng sản phẩm rồi đến TGDD, Tiki rồi cuối là Lazada. Các mặt hàng trên thương mại điện tử hay được mua như là thực phẩm, thời gian, đồ dùng gia đình,...



Hình 3.5: Những loại hàng được bán trên thương mại điện tử

3.1.5 Lý do khiến bạn dùng thương mại điện tử



Hình 3.6: Giao diện mua hàng trên thương mại điện tử

Một số lý do nên dùng thương mại điện tử:

- Có rất nhiều các voucher ưu đãi trong những ngày dịp lễ. Ví dụ như Shope, Lazada vào những ngày trùng với tháng, giữa tháng, cuối tháng đều có những dịp Flash Sale thu hút rất nhiều người tiêu dùng.
- Tiết kiệm thời gian: Bạn có thể không cần phải di chuyển mà vẫn có thể mua được các sản phẩm cần dùng.
- Sản phẩm đa dạng. Có rất nhiều mặt hàng được mua bán trên trang thương mại điện tử cho bạn lựa chọn.
- Dễ dàng sử dụng: Thông thường các trang web thương mại điện tử thiết kế giao diện rất dễ dàng sử dụng.

3.1.6 Những trở ngại khi mua trên thương mại điện tử

Bên cạnh những lợi ích khi mua đồ trên thương mại điện tử, bạn còn có thể đối mặt với nhiều rủi ro như sản phẩm kém chất lượng do ta không được nhìn thấy trực tiếp, dịch vụ chăm sóc khách hàng kém, vận chuyển bị hỏng đồ hay thông tin cá nhân có thể bị lộ, ...

Các trở ngại khi mua hàng trực tuyến



Hình 3.7: Những trở ngại khi mua trên thương mại điện tử

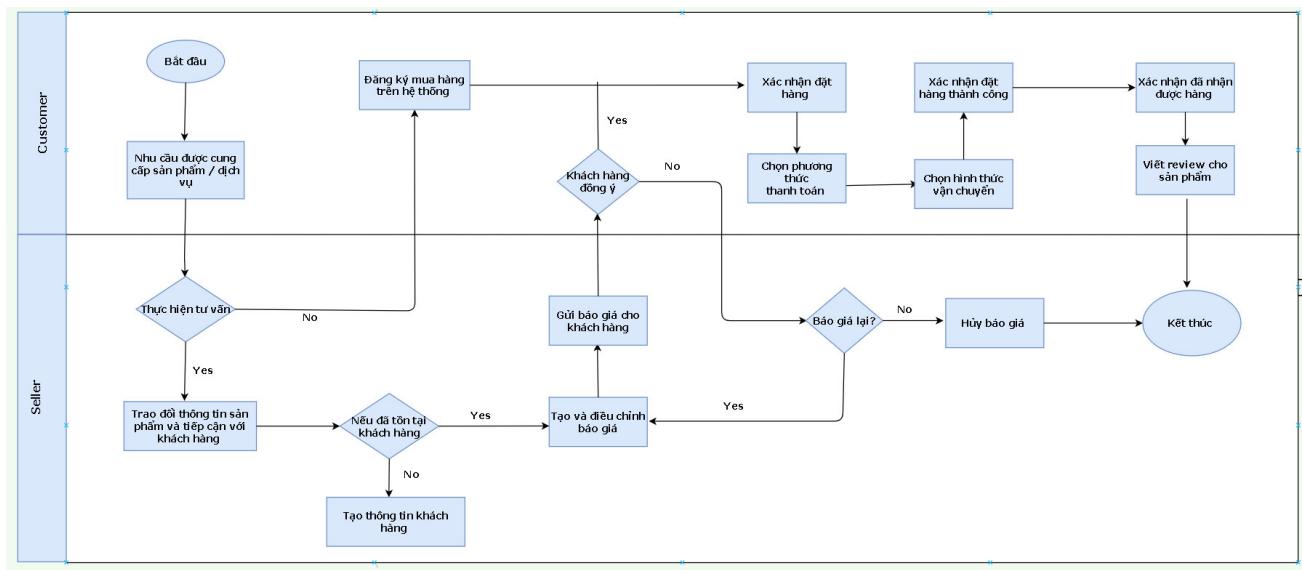
3.2 Phân tích nghiệp vụ quy trình xử lý mua bán trên sàn thương mại điện tử

Khai thác được các thông tin có giá trị từ bộ dữ liệu thu thập được để phục vụ cho việc phân tích dữ liệu. Từ đó, ta đưa ra được những phân tích, đánh giá các kết quả thu được và dự báo trong tương lai nhằm đưa ra phương án phù hợp giúp tăng doanh thu và lợi nhuận của các sàn thương mại điện tử cũng như tránh được những rủi ro.

Khi mua 1 sản phẩm trên trang thương mại điện tử thì thông thường có các bước sau:

- Khách hàng sẽ order đơn hàng.
- Sau đó, khách hàng sẽ thanh toán đơn hàng đó.
- Người quản lý sẽ chấp thuận đơn đó và đưa dữ liệu lên DataWarehouse.
- Sau đó bàn giao cho bên đóng gói và phương thức vận chuyển.
- Người vận chuyển sẽ đưa đến khách hàng.
- Nếu như khách hàng cảm thấy hài lòng thì sẽ đưa ra đánh giá và nếu không hài lòng thì ta sẽ phải tìm cách khắc phục

Dưới đây là quy trình chi tiết cho bài toán nghiệp vụ:



Hình 3.8: Quy trình nghiệp vụ trên thương mại điện tử

Quy trình chi tiết nghiệp vụ bài toán như sau:

- Khi khách hàng có nhu cầu muốn được cung cấp các sản phẩm thì khách hàng sẽ vào các trang web bán hàng để tìm kiếm những sản phẩm mà mình mong muốn, về chất lượng và giá cả sau đó sẽ xem xét có cần tư vấn với bên cung cấp dịch vụ hay không.
- Nếu như khách hàng không muốn thực hiện tư vấn thì sẽ đăng ký mua hàng trực tiếp trên hệ thống này. Lúc này thông tin chi tiết của khách hàng được lưu lại, bao gồm lịch sử đặt hàng, số lượng đơn đặt hàng và phương thức thanh toán. Cuối cùng, đơn đặt hàng của khách hàng được gửi đến kho.
- Còn nếu như khách hàng thực hiện tư vấn thì bên cung cấp dịch vụ sản phẩm sẽ tiến hành tư vấn qua điện thoại hay tin nhắn để trao đổi thông tin sản phẩm cũng như tiếp cận với khách hàng. Nếu như khách hàng chưa mua hàng thì tạo thông tin khách hàng, nếu mua rồi ta sẽ xem xét và điều chỉnh lại giá cho khách quen, báo giá lại cho khách. Nếu khách không đồng ý với báo giá này ta sẽ tiến hành báo giá lại, nếu không thành công ta sẽ hủy báo giá và kết thúc. Nếu khách hàng đồng ý với báo giá này thì lúc này sẽ tiến hành lưu thông tin các bước thực hiện của khách hàng, từ xác nhận đơn hàng, chọn phương thức thanh toán , hình thức vận chuyển. Cuối cùng, đơn đặt hàng của khách hàng được gửi đến kho.
- Hàng trong kho được kiểm tra bởi người quản lý kho và các mặt hàng đều liên tục từ các nhà cung cấp được ghi lại. Nếu hàng tồn kho gần hết, hoặc hết sạch do một đơn hàng lớn, thì một đơn mua hàng sẽ được chuyển đến bộ phận thu mua.
- Đơn đặt hàng được gửi đến bộ phận kế toán, ghi lại là thanh toán bằng tiền mặt hoặc bằng tài khoản ngân hàng. Việc bán hàng được ghi vào sổ cái, một hóa đơn được tạo và gửi cho khách hàng, và khoản thanh toán được ghi lại.
- Dịch vụ vận chuyển của bên thứ ba (hoặc của chính công ty) sẽ giao hàng cho khách hàng. Khách hàng nhận hàng và kiểm tra đơn hàng. Lúc này, khách hàng có muốn đánh giá và review sản phẩm hay không, nếu có hệ thống thương mại điện tử sẽ cho khách hàng xu đánh giá, còn nếu không thì khách hàng sẽ rời đi. Khi đó, đơn đặt hàng được hoàn thành.

3.3 Requirements

Yêu cầu xây dựng Data Warehouse:

- Đáp ứng dữ liệu lớn, đa dạng chuẩn hóa, thống nhất dữ liệu theo yêu cầu phân tích.

BÁO CÁO CÁ NHÂN

- Cung cấp công cụ phân tích và dự báo các xu hướng mua sắm để tăng doanh thu, tạo báo cáo trực quan.
- Dễ dàng trích xuất dữ liệu thông tin về khách hàng, đơn hàng, doanh thu.
- Kiến trúc linh hoạt, dễ dàng mở rộng quy mô dữ liệu, yêu cầu khi tổ chức thay đổi.

Yêu cầu bài toán: Từ bộ dữ liệu về 100 nghìn đơn đặt hàng thương mại điện tử trong năm 2017 - 2018 thực hiện ở nhiều thị trường ở Brazil của công ty Olist, ta sẽ đánh giá được tình hình phát triển thương mại điện tử của công ty Olist theo các khía cạnh sau:

1. Doanh thu: Phân tích doanh thu của công ty Olist dựa trên các chiều năm, tháng, ngày giờ xảy ra, các khu vực vị trí địa lý, theo hình thức thanh toán, theo loại hình sản phẩm, kích cỡ của sản phẩm cũng như tình trạng của các đơn hàng.
2. Số lượng đơn đặt hàng: Phân tích Số lượng đơn đặt hàng của công ty Olist dựa trên các chiều năm, tháng, ngày giờ xảy ra, các khu vực vị trí địa lý, theo hình thức thanh toán, theo loại hình sản phẩm, kích cỡ của sản phẩm cũng như các nhóm trả góp.
3. Lợi nhuận: Phân tích lợi nhuận của công ty Olist dựa trên các chiều năm, tháng, ngày giờ xảy ra, các khu vực vị trí địa lý, theo hình thức thanh toán, theo loại hình sản phẩm, kích cỡ của sản phẩm cũng như các nhóm trả góp.

Input: Dữ liệu về 100 nghìn đơn đặt hàng thương mại điện tử trong năm 2017 - 2018 thực hiện ở nhiều thị trường ở Brazil của công ty Olist.

Xây dựng dashboard có thể trả lời các câu hỏi sau:

- Doanh thu theo thời gian (ngày, tháng, năm) có sự biến động như thế nào từ năm này sang năm khác.
- Doanh thu theo địa điểm (bang, vùng lãnh thổ) có sự biến động như thế nào theo thời gian.
- Doanh thu theo hình thức thanh toán và kích cỡ sản phẩm có sự biến động như thế nào theo thời gian.
- Doanh thu theo loại hình sản phẩm có sự biến động như thế nào theo thời gian.
- Số lượng đơn đặt hàng theo thời gian (ngày, tháng, năm).
- Số lượng đơn đặt hàng theo địa điểm (bang, vùng lãnh thổ) qua từng tháng, từng năm.
- Số lượng đơn đặt hàng theo hình thức thanh toán và kích cỡ sản phẩm có sự biến động như thế nào theo thời gian.

- Số lượng đơn đặt hàng theo loại hình sản phẩm có sự biến động như thế nào theo thời gian.
- Số lượng đơn đặt hàng theo nhóm trả góp có sự biến động như thế nào theo thời gian.
- Lợi nhuận theo thời gian (ngày, tháng, năm).
- Lợi nhuận theo địa điểm (bang, vùng lãnh thổ) có sự biến động như thế nào theo thời gian.
- Lợi nhuận theo loại hình sản phẩm có sự biến động như thế nào theo thời gian.
- Lợi nhuận theo nhóm trả góp có sự biến động như thế nào theo thời gian.
- Lợi nhuận hình thức thanh toán và kích cỡ sản phẩm có sự biến động như thế nào theo thời gian.
- Dự báo tăng số mặt hàng nào trong tương lai, phát triển thêm những khía cạnh nào.

3.4 Bộ dữ liệu gốc

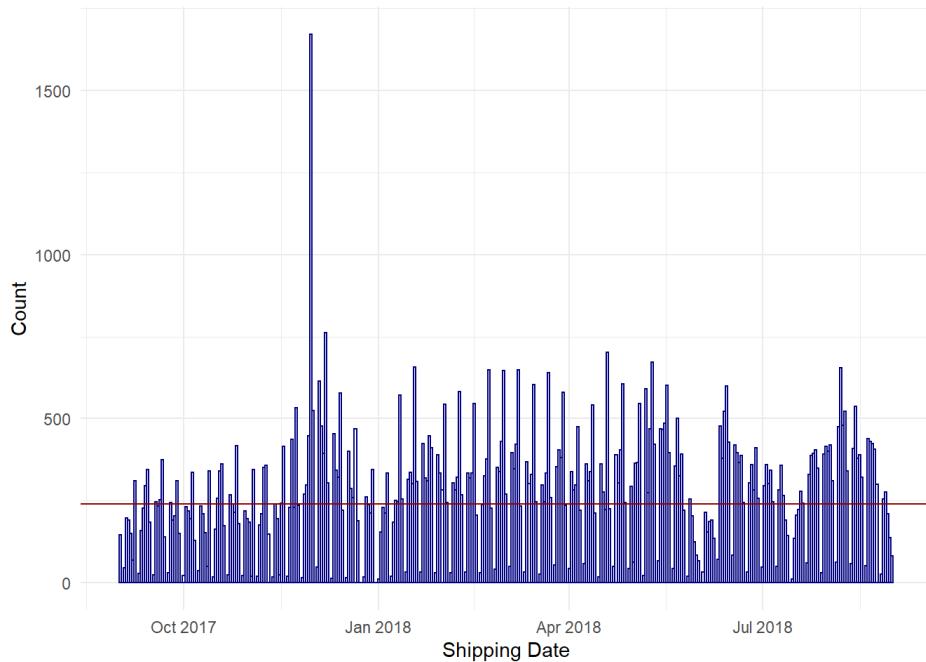
- Tên bộ dữ liệu: Brazilian E – commerce.
- Mô tả: là tập dữ liệu công khai về thương mại điện tử của Brazil về các đơn đặt hàng được thực hiện tại Olist Store.
- Nguồn thu thập: Kaggle.
- Thời gian: 5/1/2017 - 29/8/2018.
- Kích cỡ: 126,19 MB.
- Số bản ghi: 1 565 259.
- Gồm 9 file:
 1. File olist_customers_dataset
 2. File olist_geolocation_dataset
 3. File olist_order_items_dataset
 4. File olist_order_payments_dataset
 5. File olist_order_reviews_dataset
 6. File olist_order_dataset
 7. File olist_products_dataset
 8. File olist_sellers_dataset
 9. File product_category_name

3.5 Mô tả một số trường dữ liệu quan trọng

STT	Trường dữ liệu	Mô tả
1	customer_id	Mỗi đơn hàng có một customer_id riêng
2	customer_unique_id	Mỗi khách hàng có một customer_unique_id riêng
3	order_id	Mỗi đơn hàng có một order_id riêng
4	order_item_id	Số lượng sản phẩm của mỗi đơn hàng
5	product_id	Mỗi sản phẩm có một product_id riêng
6	product_category_name_english	Tên mặt hàng bằng tiếng anh
7	payment_type	Phương thức thanh toán
8	price	Giá của sản phẩm có trong đơn hàng
9	freight_value	Phi vận chuyển của sản phẩm có trong đơn hàng
10	payment_value	Giá trị thanh toán theo phương thức thanh toán
11	review_id	Mỗi review từ khách hàng có review_id riêng
12	review_score	Điểm đánh giá của khách hàng
13	order_status	Trạng thái đơn hàng
14	order_purchase_timestamp	Thời gian khách hàng nhấn mua hàng
15	order_approved_at	Thời gian người bán chấp nhận thanh toán
16	order_delivered_carrier_date	Thời gian gói hàng được gửi đến bến vận chuyển
17	order_delivered_customer_date	Thời gian thực tế gói hàng được chuyển đến khách hàng
18	order_estimated_delivery_date	Thời gian dự kiến khách hàng sẽ nhận được gói hàng
19	seller_id	Mỗi người bán có seller_id riêng
20	shipping_limit_date	Ngày giới hạn người bán phải giao hàng cho bến giao hàng

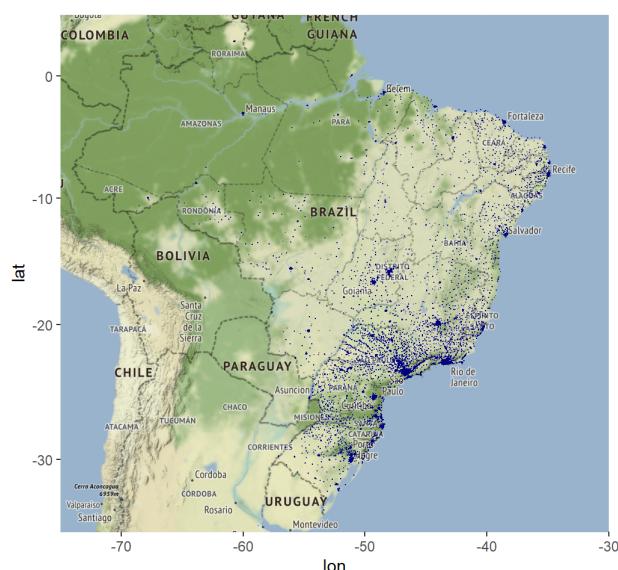
Hình 3.9: Một số trường quan trọng trong bộ dữ liệu

3.6 Data Exploration



Hình 3.10: Dữ liệu đơn đặt hàng theo ngày

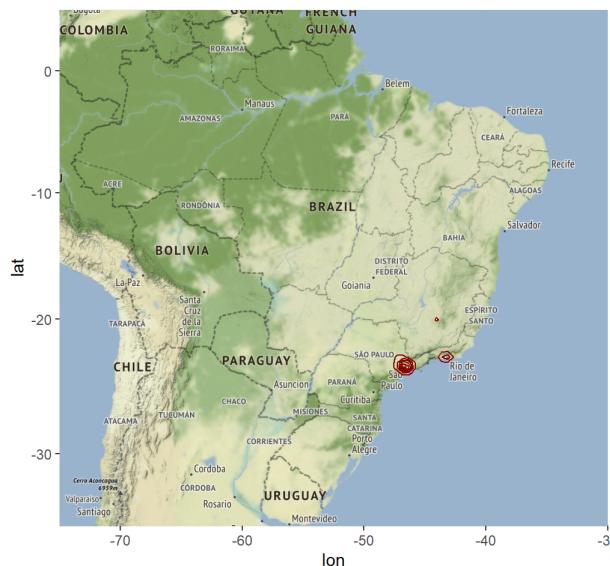
Biểu đồ này hiển thị số đơn đặt hàng được vận chuyển hàng ngày. Số đơn hàng trung bình là 281 đơn mỗi ngày và ta nhận thấy sự sụt giảm thú vị như khoảng thời gian Giáng sinh cuối năm 2017 đầu năm 2018 và một đợt sụt giảm khác vào tháng 6 năm 2018 mà không rõ lý do. Những ngày cuối tuần có thể dễ dàng nhận biết bởi các đợt giảm thường xuyên. Đáng chú ý, không có doanh số bán hàng vào black Friday /thanksgiving. Người quản lý có thể cố gắng khai thác các sự kiện này trong những năm tới.



Hình 3.11: Phân bố khách hàng theo vị trí địa lý

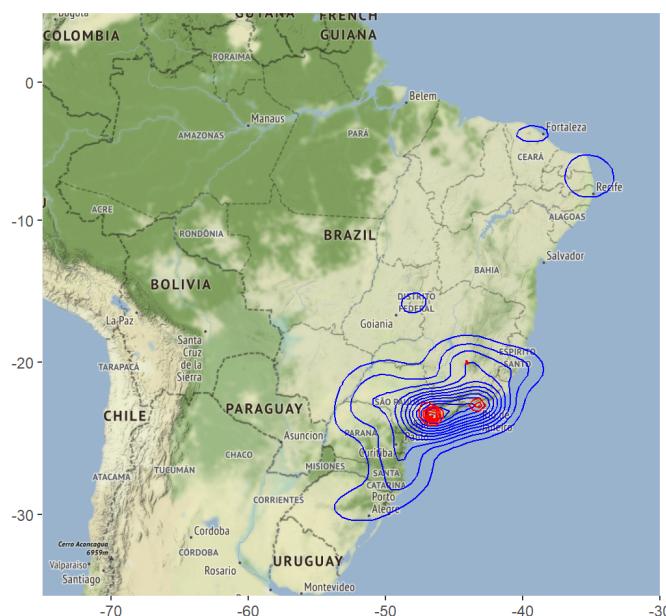
BÁO CÁO CÁ NHÂN

Trên biểu đồ phân tán, ta thấy mỗi chấm màu xanh lam đại diện cho một mã bưu điện từ nơi ít nhất một khách hàng đã đặt hàng. Có thể nhận thấy dân số đông hơn ở phần phía nam của Brazil so với phần phía bắc, trong khi các khu vực phía tây, thuộc rừng nhiệt đới Amazon thì ngược lại.



Hình 3.12: Tỷ lệ phân bố khách hàng theo vị trí địa lý

Bản đồ này cho thấy hình ảnh khách hàng nhưng thay vì các điểm phân tán, thường không chính xác khi xử lý hàng nghìn điểm trùng lặp, ta sử dụng biểu đồ mật độ 2d. Các vòng tròn bên trong cho biết mật độ khách hàng cao nhất. Rõ ràng hơn ở đây là phần lớn khách hàng Olist tập trung ở khu vực Rio và São Paulo.

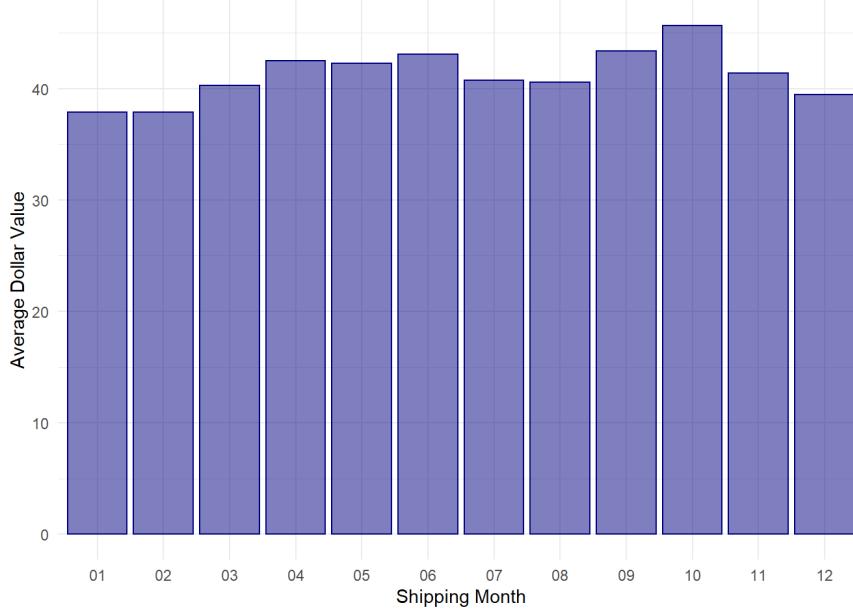


Hình 3.13: Mật độ khách hàng

Ở quy mô quốc gia, ta quan sát thấy mật độ khách hàng màu xanh lam lan rộng hơn so với khách hàng

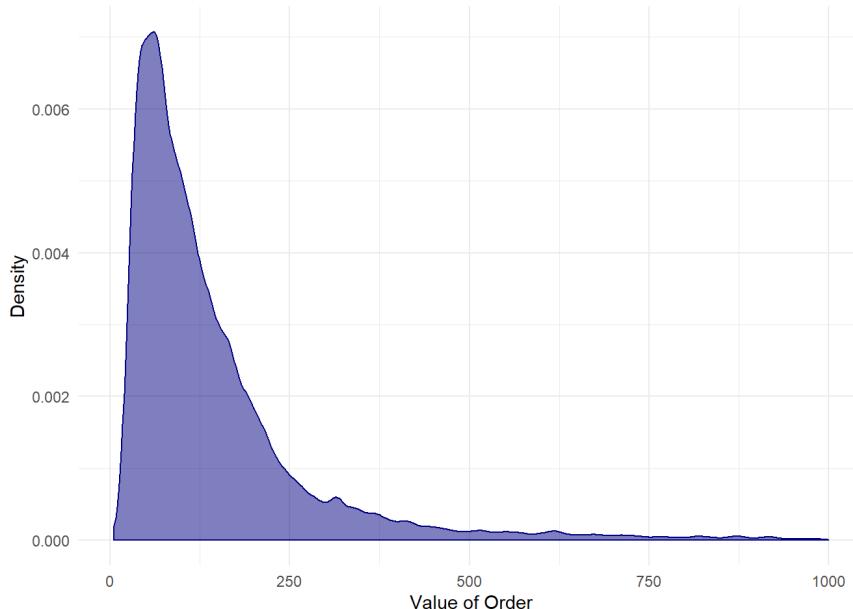
BÁO CÁO CÁ NHÂN

màu đỏ. Có thể giả thuyết những người sống ở vùng nông thôn hoặc bên ngoài các thành phố lớn sẽ cân nhắc mua hàng trực tuyến do tỷ lệ ngân sách có thể chi thấp.



Hình 3.14: Giá trị trung bình đơn đặt hàng theo tháng

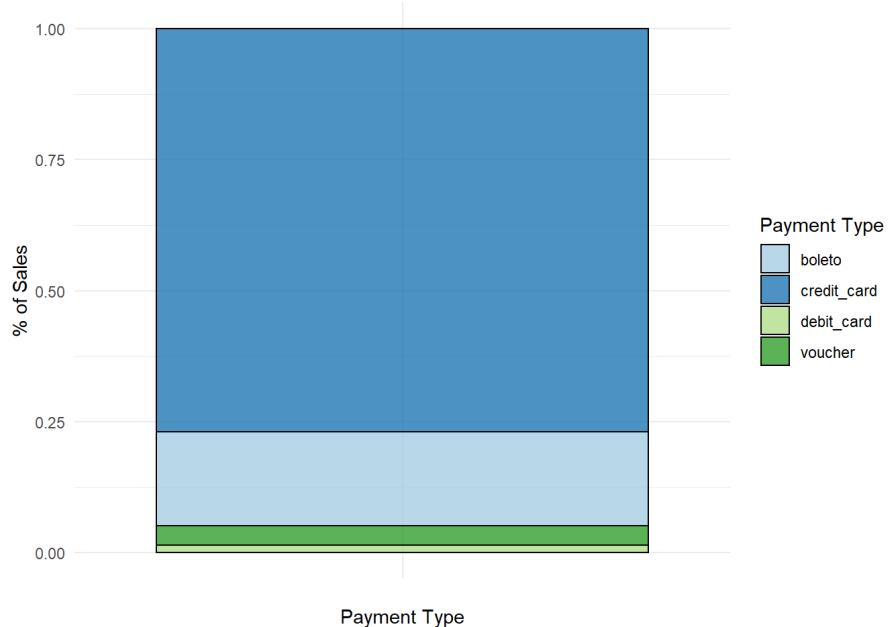
Biểu đồ thanh cho ta thấy giá trung bình của một đơn đặt hàng (một giỏ) trên nền tảng danh sách, giá trị đó là khoảng 40 đô la, bằng khoảng một nửa giá trị đơn đặt hàng trung bình 78 đô la tại Amazon. Xem xét thu nhập của người Brazil, trung bình nhỏ hơn 5 lần so với ở Mỹ, dễ dàng thấy giá đơn đặt hàng trung bình của Olist là tương đối cao. Nhưng để đưa ra một so sánh hữu ích hơn, chúng ta cần tiếp cận thu nhập của khách hàng trong danh sách, được cho là cao hơn thu nhập trung bình.



Hình 3.15: Giá trị đơn hàng theo tỷ lệ

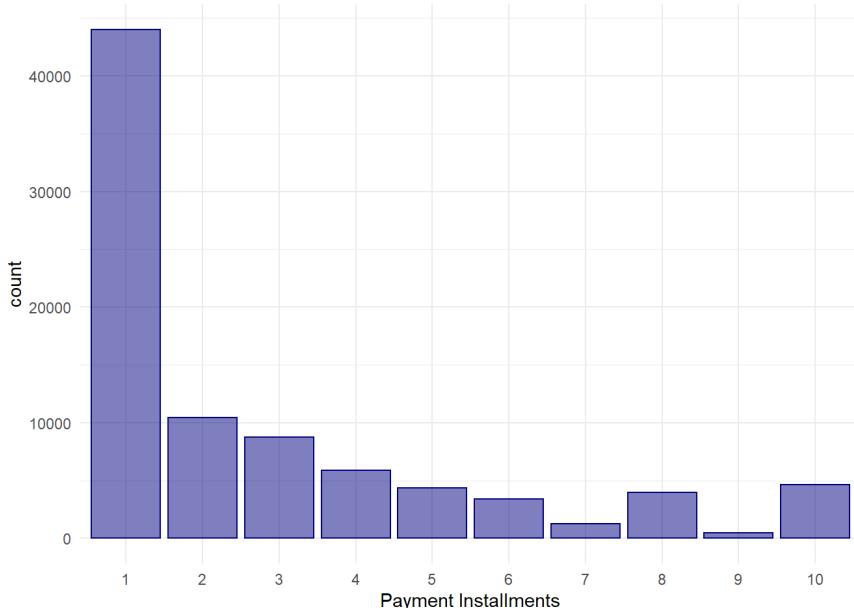
BÁO CÁO CÁ NHÂN

Như có thể thấy trên biểu đồ mật độ này, phần lớn các đơn đặt hàng được phân phối cho 40 đô la (ở đây tính theo nội tệ).



Hình 3.16: Tỷ lệ hình thức thanh toán

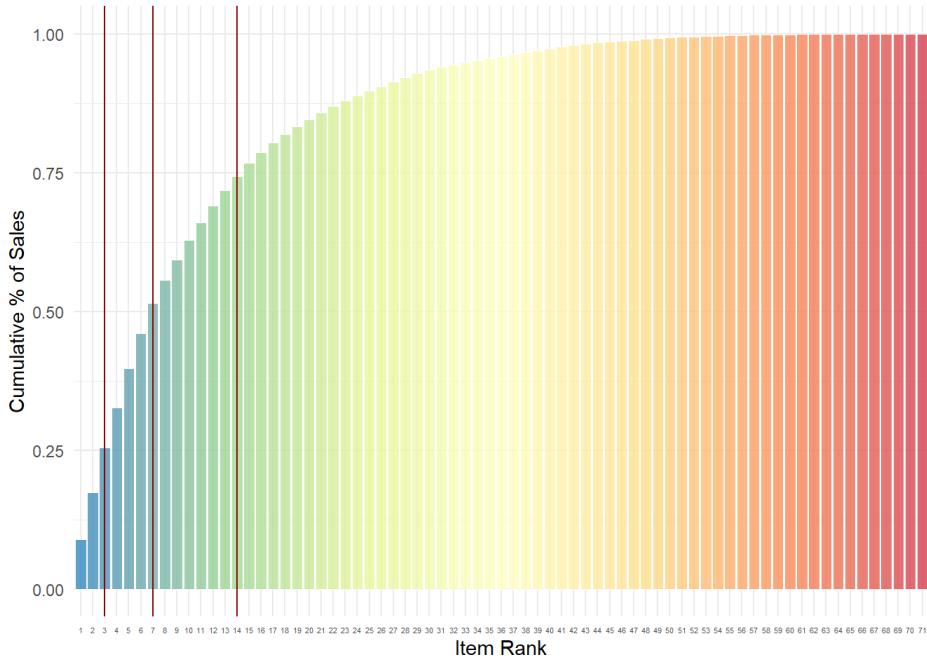
Thẻ tín dụng và boleto chiếm hơn 90% doanh thu. Boleto là một phương thức thanh toán của Brazil yêu cầu người dùng đến ngân hàng gần nhất để thực hiện chuyển tiền. Mặc dù tồn thời gian, phương thức thanh toán này rất phổ biến vì nó được coi là an toàn hơn so với sử dụng thẻ tín dụng. Trên thực tế, Brazil nổi tiếng với tỷ lệ gian lận thẻ tín dụng, cao hơn hầu hết các quốc gia.



Hình 3.17: Lượng đơn hàng theo số lần trả góp

BÁO CÁO CÁ NHÂN

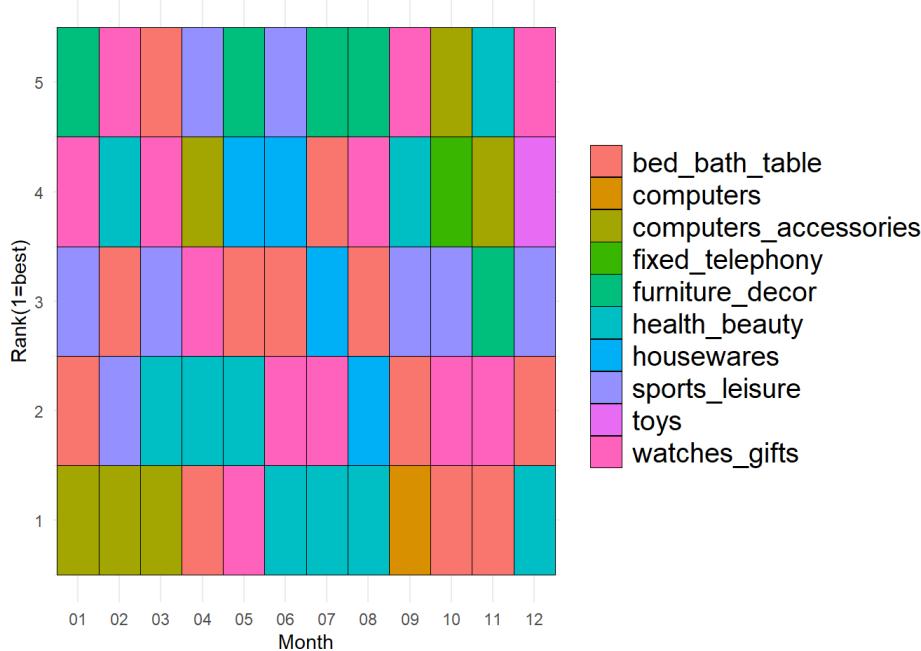
Khách hàng trong danh sách thanh toán bằng thẻ tín dụng có khả năng thanh toán nhiều đợt (có thể là lý do tại sao thẻ tín dụng trung bình thanh toán cho các giỏ hàng có giá trị cao hơn). Không ngạc nhiên khi số lần trả góp có mối quan hệ cùng chiều với giá trị tổng thể của các món hàng đã mua.



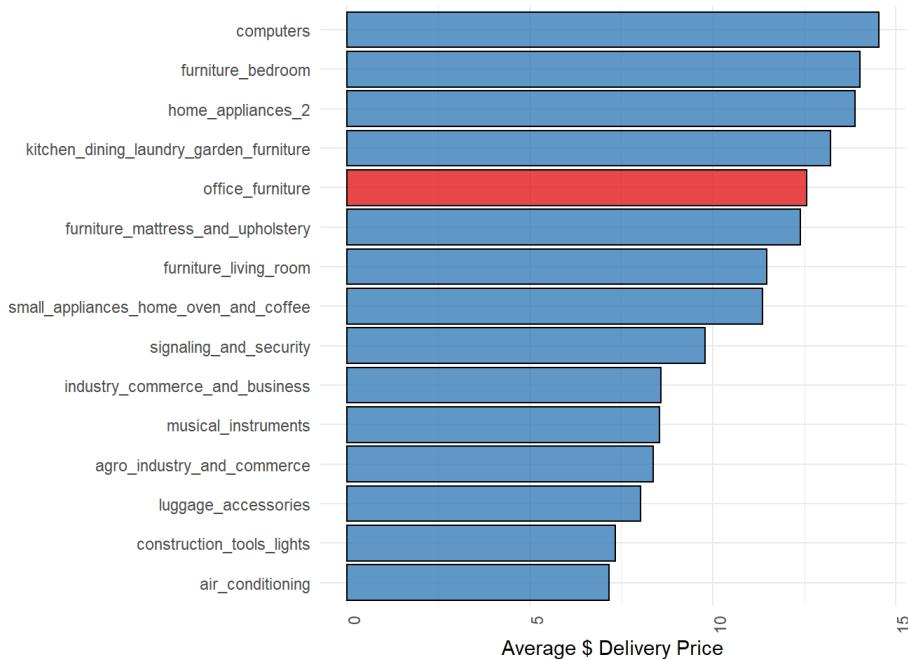
Hình 3.18: Tỷ lệ đơn hàng theo danh mục sản phẩm

Biểu đồ thanh bên dưới cho thấy tỷ lệ phần trăm số lượng đơn đặt hàng giảm dần tích lũy của các danh mục sản phẩm. Ví dụ, ta thấy rằng 75% doanh số bán hàng tập trung trong số 15 danh mục sản phẩm hàng đầu. Biểu đồ thứ hai bên dưới cho thấy 15 danh mục hàng đầu đó. Sẽ rất thú vị đối với các nhà quản lý khi nghiên cứu mối tương quan về doanh số giữa các loại sản phẩm đó và xem liệu có thể đạt được bất kỳ lợi ích đa dạng hóa nào bằng cách phát triển các phân khúc sản phẩm khác hay không, chẳng hạn để tự bảo vệ mình trong trường hợp kinh tế bị sốc. Các biến số khác như độ co giãn theo giá riêng hoặc phí bán hàng cần được các nhà quản lý tính đến khi phản ánh cách thay đổi tổ hợp sản phẩm.

Biểu đồ nhiều màu sắc dưới đây cho thấy 5 danh mục sản phẩm hàng đầu cho mỗi tháng trong năm. Điều thú vị là có rất nhiều doanh thu và ít danh mục vẫn đứng đầu trong suốt cả năm. Chỉ có giường_bếp_mặt và sức_khỏe_đẹp là bán chạy nhất quanh năm. Các danh mục khác chẳng hạn như đồ gia dụng hoặc máy tính_accesories có vẻ theo mùa hơn. Cuối cùng, một số loại sản phẩm thực hiện rất đúng giờ chẳng hạn như máy tính (có thể là đầu năm học).



Hình 3.19: Xếp hạng danh mục sản phẩm hàng đầu theo mỗi tháng

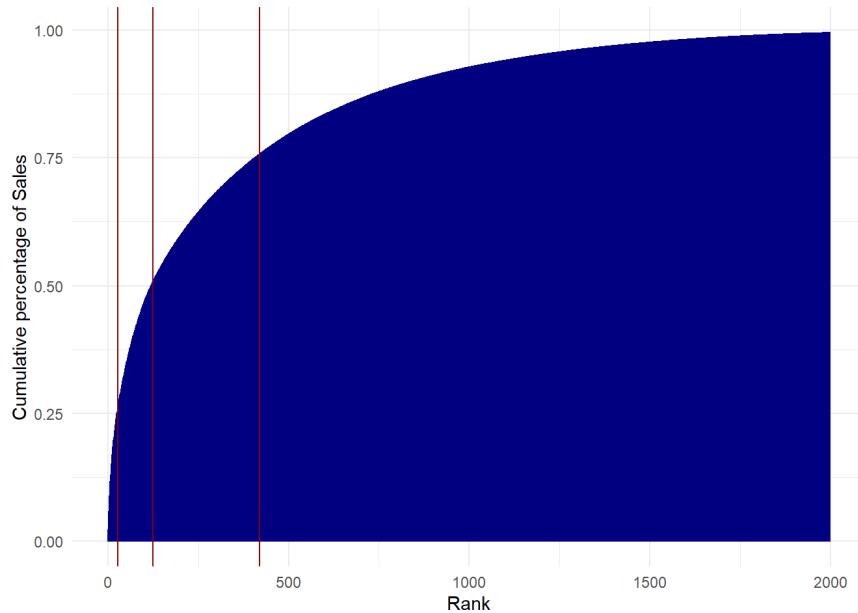


Hình 3.20: Phí giao hàng theo danh mục sản phẩm

Biểu đồ thanh ở trên hiển thị phí giao hàng trên mỗi danh mục sản phẩm và danh mục có tỷ lệ giá trị đơn hàng trên giá trị phân phối cao nhất. Một nhận xét đầu tiên là phí giao hàng có vẻ khá đắt so với mức giá mà chúng ta biết ở các nền kinh tế phát triển, người quản lý có lẽ nên xem liệu có thể đạt được hiệu quả nào trên khía cạnh đó hay không, hoặc xem việc giấu phí giao hàng trong giá sản phẩm sẽ ảnh hưởng đến doanh số bán hàng như thế nào. Ngoài ra, đáng chú ý là một số danh mục sản phẩm có

BÁO CÁO CÁ NHÂN

tỷ lệ phí giao hàng quá đắt. Trong số đó có một số danh mục sản phẩm bán chạy nhất. Các nhà quản lý nên điều tra xem liệu tỷ lệ này có thể được giảm xuống bằng cách nào, để tăng tỷ suất lợi nhuận hoặc doanh số bán hàng tổng thể.



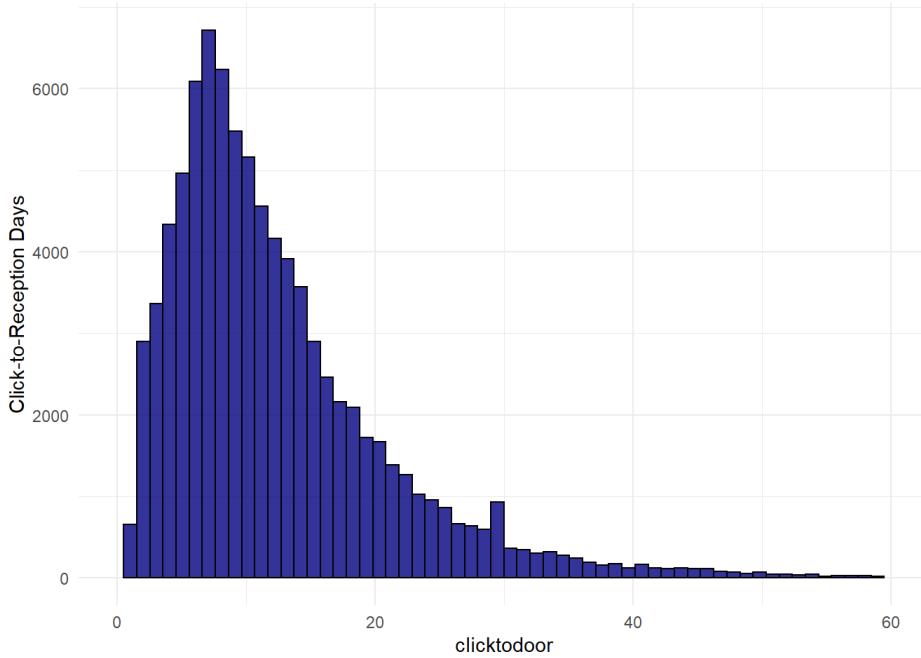
Hình 3.21: Doanh số tích lũy bán hàng theo người bán

Sự phân bổ tích lũy doanh số bán hàng theo từng người bán riêng lẻ này lại dẫn đến sự phân phối không cân bằng độc đáo, với 420 người bán (tổng cộng là 2566 người) kiểm soát 75% doanh số bán hàng. Thật kỳ lạ, tỷ lệ phân bổ đó rất giống với tỷ lệ phân bổ của cải trên thế giới. Chắc chắn không phải là một hiện tượng ngẫu nhiên.



Hình 3.22: Phân bố vị trí những người bán hàng chính

Trên bản đồ trên, chúng ta thấy vị trí của những người bán hàng chính, màu xanh lam là 27 (25%) và màu đỏ là 127 (50%). Hầu hết tập trung ở khu vực phía Nam, nơi hầu hết các hoạt động kinh doanh đang được thực hiện. Một vị trí bên ngoài là nhà bán hàng lớn thứ ba, nằm ở vùng Salvador.

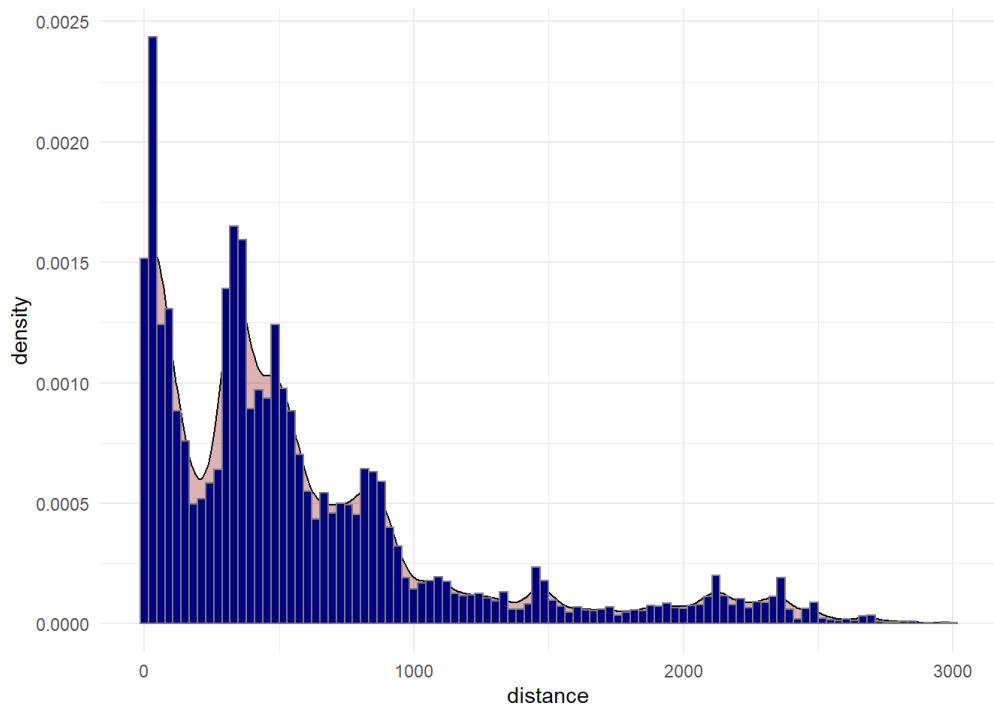


Hình 3.23: Thời gian hàng được giao kể từ khi khách hàng đặt đơn

Một nửa số đơn đặt hàng sẽ đến đích sau 10 ngày kể từ khi khách hàng nhấp vào mua hàng lần đầu tiên. Đây là mức cao hơn nhiều so với Amazon mà trung bình các đơn đặt hàng chỉ đến trong 3,28 ngày. Đây chắc chắn là một kpi mà Olist nên cố gắng cải thiện, nhưng dù sao đây cũng không phải là điểm đáng quan tâm chính đối với tuổi trẻ tương đối của công ty. 3 năm trước, trung bình gói hàng của Amazon mất 6 ngày để đến nơi.

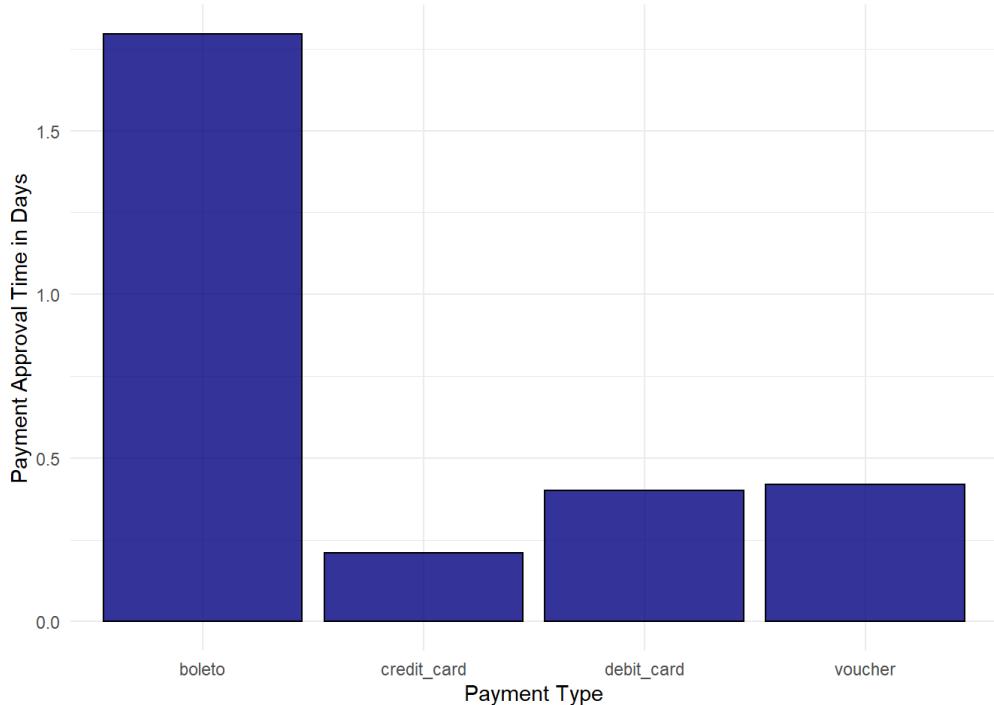
Sự phân bổ thú vị dưới đây của các đơn đặt hàng như một hàm của khoảng cách di chuyển, không suôn sẻ chút nào. Hai đỉnh núi đầu tiên có thể được giải thích theo cách sau: Rio và São Paulo là hai trung tâm hoạt động kinh tế chính và cách nhau khoảng 400 km, khoảng cách giữa các đỉnh. Số lượng lớn thương mại song phương giữa các thành phố đó, những người buôn bán để kết tụ của họ hoặc kết tụ của các thành phố khác, gây ra mô hình bùng nổ và phá sản này. Các đỉnh khác cũng phải là kết quả của một hiện tượng tương tự.

BÁO CÁO CÁ NHÂN



Hình 3.24: Phân bố khoảng cách đặt hàng

Về trung bình, một gói hàng đi được 427 km, nếu chia cho thời gian trung bình mà một gói thực hiện để đến đích thì sẽ cho 42 km một ngày, đường như không nhiều. Một chỉ số hiệu suất quan trọng khác mà các nhà quản lý nên tập trung cải thiện.

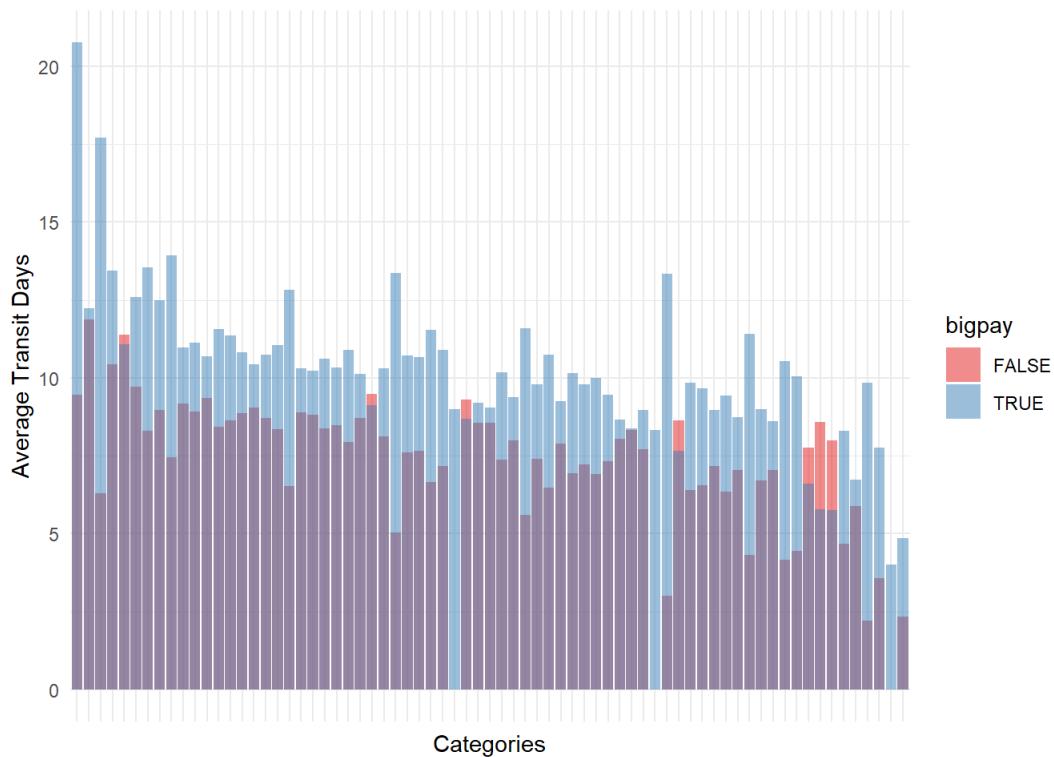


Hình 3.25: Thời gian trung bình đơn hàng được chấp nhận theo hình thức thanh toán

Biểu đồ trên thể hiện trung bình mất bao lâu để một đơn đặt hàng được chấp thuận tùy thuộc vào phương tiện thanh toán. Nó có thể được suy luận một lần nữa, rằng loại hình thanh toán Boleto là một phanh tiêu dùng vì nó làm tăng loại phê duyệt một cách có hệ thống và do đó ảnh hưởng đến khách hàng cần gói hàng của họ một cách nhanh chóng hoặc mua một cách bắt buộc.

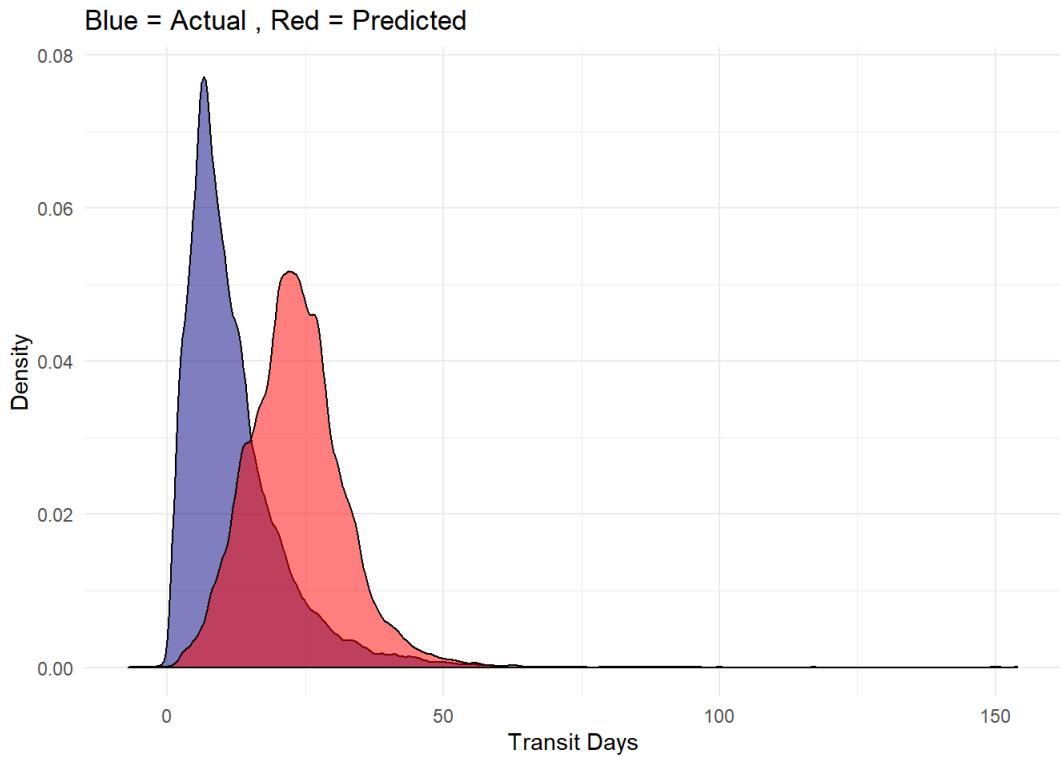
Quá trình phân phối một gói hàng được chia thành 3 phần:

- Khách hàng nhấp vào mua hàng và đợi thanh toán của mình được chấp thuận.
- Sau khi thanh toán được chấp thuận, người bán phải đóng gói đơn đặt hàng và gửi đến nhà vận chuyển.
- Người vận chuyển sau đó nhận đơn hàng và giao hàng cho khách.



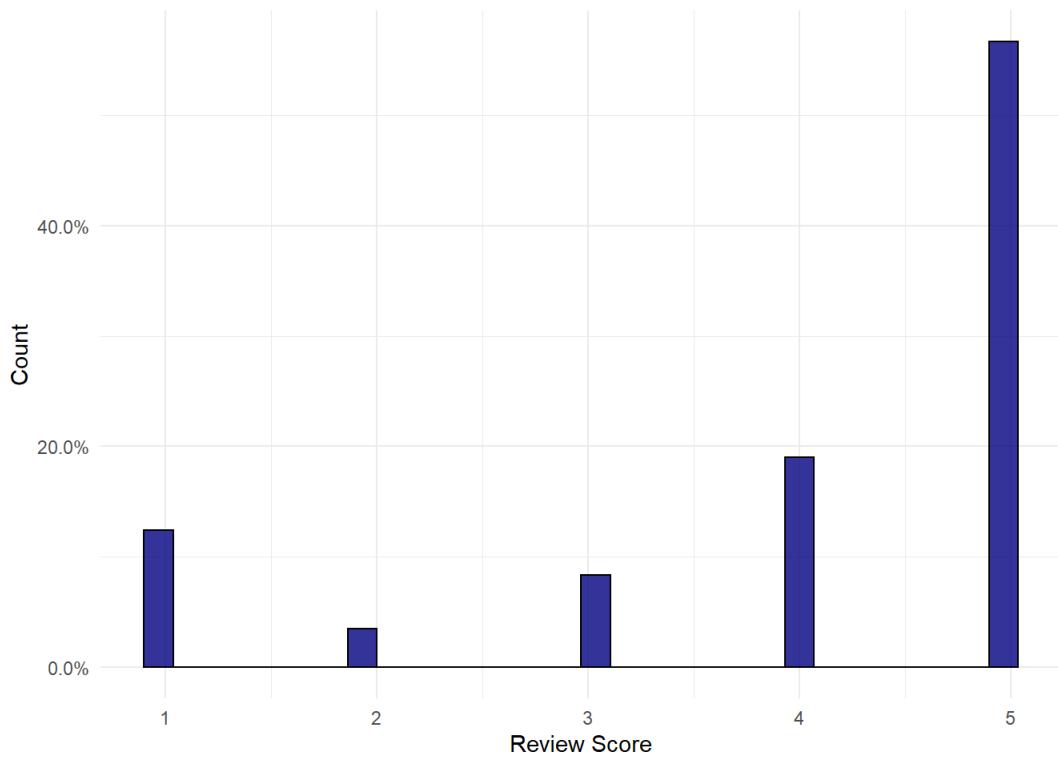
Hình 3.26: Thời gian vận chuyển theo danh mục sản phẩm

Biểu đồ trên có các danh mục sản phẩm trên trục X và số ngày vận chuyển trung bình trên trục Y. Màu đỏ là số ngày vận chuyển trung bình của đơn hàng mà khách hàng đã trả thấp hơn giá giao hàng trung bình và màu xanh lam là những người đã trả cao hơn. Điều thú vị là những khách hàng trả nhiều hơn phí giao hàng trung bình có xu hướng nhận gói hàng của họ muộn hơn, gần như một cách có hệ thống. Sự khác biệt là đáng kể khi kích thước mẫu đủ lớn.



Hình 3.27: Thời gian giao hàng dự kiến và thực tế

Biểu đồ trên hiển thị màu xanh lam những ngày ước tính trước khi giao hàng và màu đỏ là những ngày thực tế. Đáng chú ý là thời gian giao hàng thực tế ít hơn 10 ngày so với thời gian giao hàng dự đoán với mô hình đường cong rất giống nhau (cho đến một điểm). Thứ hai, giao hàng dự đoán bắt đầu sau 2 tuần đối với khoảng cách 0 km và kéo dài đến 3 tuần đối với thời gian giao hàng 200 km. Cuối cùng, đường phân phối dự đoán tăng lên một cách đơn điệu trong khi màu đỏ có hình dạng lõm. Hình dạng lõm có thể được giải thích là do khoảng cách tăng lên hoặc nếu khách hàng sống trên một hòn đảo, người vận chuyển có thể chuyển sang các phương tiện vận chuyển khác, chẳng hạn như máy bay, điều này làm tăng tốc độ di chuyển trung bình của một bưu kiện. Tất cả những yếu tố này chỉ ra một phần mềm lỗi thời để ước tính thời gian giao hàng. Các nhà quản lý nên ưu tiên điều tra điều này. Không phải là một tuyên bố mạnh mẽ khi thấy rằng thời gian giao hàng ước tính là hai tuần (tối thiểu) sẽ ngăn cản nhiều người mua.



Hình 3.28: Phân bô tỷ lệ đánh giá của khách hàng

Các bài đánh giá có sự phân bố hình chữ U, theo trực giác có ý nghĩa bởi vì mọi người là những người sống tình cảm và có xu hướng đưa ra những tuyên bố cực đoan, do đó, ví dụ: khi bạn muốn bày tỏ sự thất vọng của mình, bạn sẽ không phải suy nghĩ lâu giữa việc đưa ra hai hay một. Nó thường là tất cả.

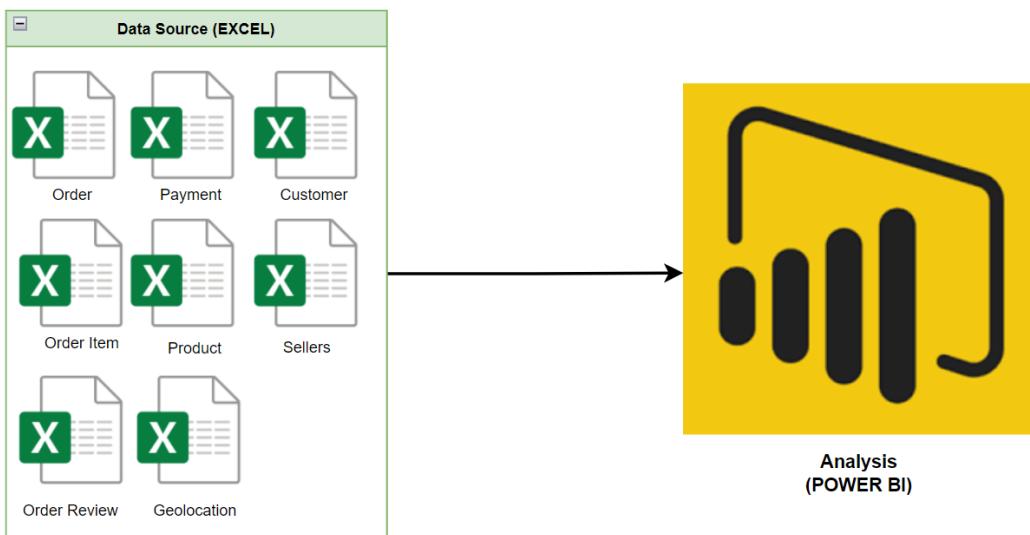
3.7 Phân tích và thiết kế

3.7.1 Kiến trúc Data Warehouse

Kiến trúc cũ

Hệ thống sử dụng dữ liệu từ các file csv, từ đó tạo ra các báo cáo phân tích. Vì vậy hệ thống cũ này còn một số nhược điểm như:

- Làm giảm hiệu năng hệ thống quản lý giao dịch.
- Dữ liệu để phân tích không ổn định.
- Không hiệu quả về mặt tốc độ.

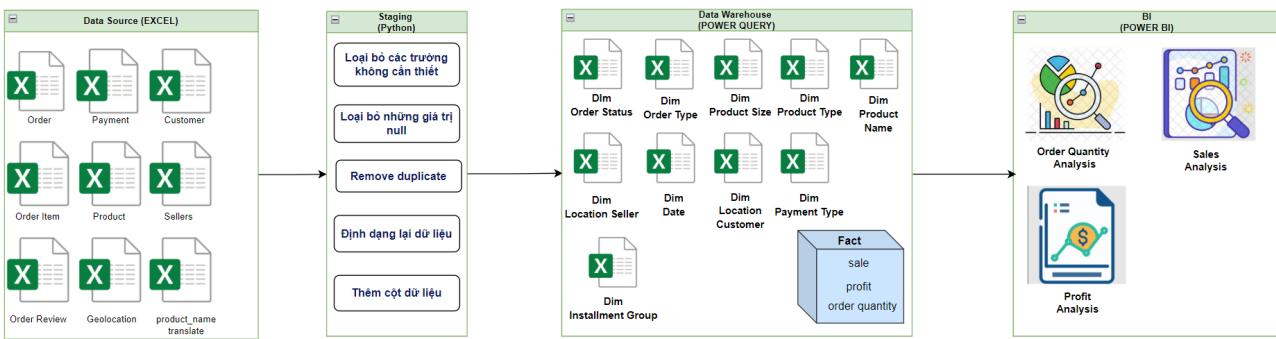


Hình 3.29: Kiến trúc Data Warehouse cũ

Kiến trúc mới

Từ những nhược điểm nêu trên thì việc xây dựng kiến trúc hệ thống mới là vô cùng cần thiết. Dưới đây là kiến trúc của hệ thống mới mà chúng em sẽ triển khai:

BÁO CÁO CÁ NHÂN



Hình 3.30: Kiến trúc Data Warehouse mới

Kiến trúc mới này bao gồm 4 tầng xử lý:

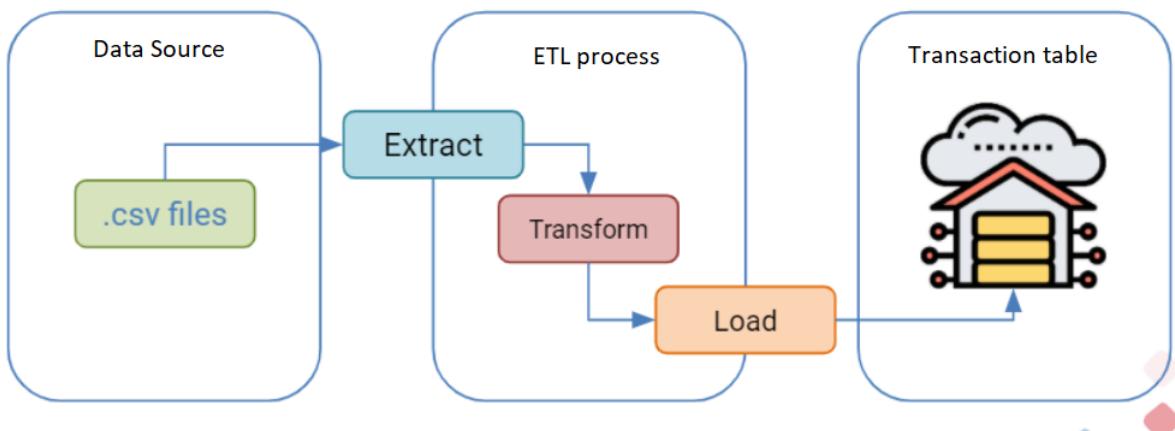
- Tầng dữ liệu nguồn: Bao gồm các bảng có định dạng csv như kiến trúc cũ được đưa vào Excel để xử lý.
- Dữ liệu của tầng dữ liệu nguồn sẽ qua tầng xử lý (trong vùng Staging) để ETL dữ liệu và dữ liệu sẽ được gửi đến DataWarehouse. Công nghệ sử dụng ở đây là Power Query.
- Tiếp đến là tầng DataWarehouse: Ở tầng này sau khi xử lý xong dữ liệu, ta tiến hành phân chia dữ liệu thành các Dim và các fact để phục vụ cho việc phân tích. Ở đây, chúng em chia thành 10 Dim đó là Dim Order Status, Dim Order Type , Dim Product Size, Dim Product Type , Dim Product Name, Dim Location Seller, Dim Date, Dim Location Customer, Dim Payment Type, Dim Installment Group và 3 Fact đó là Fact Sales, Fact Profit, Fact OderQuantity.
- Cuối cùng, sau khi có dữ liệu về các Dim và fact, chúng em sẽ đưa dữ liệu vào Power BI để tiến hành xây dựng các báo cáo phân tích. Ở đây, chúng em tập trung vào các báo cáo phân tích về số lượng đơn đặt hàng, lợi nhuận và doanh thu.

3.7.2 Quy trình ETL dữ liệu

Quy trình ETL (Extract, Transform and Load) bao gồm các bước sau:

- **Extract:** là quá trình đọc dữ liệu từ cơ sở dữ liệu. Trong giai đoạn này, dữ liệu được thu thập, thường là từ nhiều loại nguồn khác nhau.
- **Transform:** là quá trình chuyển đổi dữ liệu được trích xuất từ biểu mẫu trước đó thành biểu mẫu cần có để có thể được đặt vào cơ sở dữ liệu khác. Chuyển đổi xảy ra bằng cách sử dụng các quy tắc hoặc bảng tra cứu hoặc bằng cách kết hợp dữ liệu này với dữ liệu khác.
- **Load:** là quá trình chia dữ liệu thành các bảng với data model liên kết với nhau và đưa vào trong Data Warehouse.

Sơ đồ thực hiện ETL



Hình 3.31: Sơ đồ quy trình ETL

Các nội dung ETL đã thực hiện được

Quá trình ETL dữ liệu có thể được xử lý bằng rất nhiều cách và bằng nhiều công cụ xử lý dữ liệu khác nhau như Power Query, Power BI, Python, ... Trong phạm vi của bài báo cáo này, nhóm chúng em đã sử dụng công cụ lập trình Python để thực hiện các thao tác ETL dữ liệu. Lợi thế của việc sử dụng Python là đem lại khả năng xử lý được với dữ liệu lớn, tốc độ xử lý nhanh, cấu trúc câu lệnh rõ ràng, ...

Một số thao tác ETL dữ liệu với Python đã được nhóm chúng em sử dụng như:

- Loại bỏ những giá trị trùng lặp (Remove duplicate).
- Loại bỏ những giá trị null.
- Xóa cột không sử dụng.
- Thêm cột kiểu custom
- Thêm cột có điều kiện.
- Merge query.
- Chính lại dữ liệu trường thời gian.
- Chính lại kiểu dữ liệu cho các trường.

BÁO CÁO CÁ NHÂN

Chi tiết của các thao tác ETL dữ liệu:

- Đặt lại khóa chính `geolocation_zip_code_prefix` cho bảng Geolocation sử dụng remove duplicate để loại bỏ những giá trị trùng lặp.

The screenshot shows two code snippets and their resulting outputs. The first snippet sorts the DataFrame by the 'geolocation_zip_code_prefix' column and displays the top 10 rows. The second snippet uses the `drop_duplicates` method to remove duplicates based on the same column and displays the resulting DataFrame. A blue arrow points from the 'Before' state to the 'After' state, indicating the transformation.

	df_geolocation.sort_values("geolocation_zip_code_prefix", inplace=True)	df_geolocation.drop_duplicates(subset="geolocation_zip_code_prefix")
geolocation_zip_code_prefix	1001	1001
geolocation_lat	-23.549292	-23.549292
geolocation_lng	-46.633559	-46.633559
geolocation_city	sao paulo	sao paulo
geolocation_state	SP	SP
...
999961	999960	-27.953858
9999933	999965	-28.210845
999974	99970	-28.343232
999924	99980	-28.387693
999864	99990	-28.329472
		19015 rows × 5 columns

Before

After

Hình 3.32: Remove duplicate

- Loại bỏ những giá trị Null. Trong bảng product, có một số mã sản phẩm không có thông tin của sản phẩm như tên sản phẩm, kích thước, ... nên sẽ xóa các cột đó đi.

The screenshot shows two code snippets and their resulting outputs. The first snippet filters the 'product' DataFrame to find rows where any column contains a null value. The second snippet drops rows where any column is null using the `dropna` method. A blue arrow points from the 'Before' state to the 'After' state, indicating the transformation.

	null_data = df_product[df_product.isnull().any(axis=1)]	df_product = df_product.dropna(how='any',axis=0)
product_id	NaN	1e9e8ef04dbcff4541ed26657ea517e5
product_category_name	NaN	perfumaria
product_name_length	NaN	40.0
product_description_length	NaN	287.0
...
32815	NaN	artes
32889	NaN	44.0
32616	NaN	276.0
32772	NaN	esporte_lazer
32852	NaN	46.0
	...	250.0
32946	NaN	bebés
	...	261.0
32947	NaN	utilidades_domesticas
	...	402.0
32948	NaN	moveis_decoracao
	...	45.0
32949	NaN	construcao_ferramentas_iluminacao
	...	41.0
32948	NaN	cama_mesa_banho
	...	50.0
32949	NaN	informatica_acessorios
	...	60.0
32950	NaN	cama_mesa_banho
	...	156.0
	...	58.0
	...	309.0
611 rows × 9 columns	32340 rows × 9 columns	

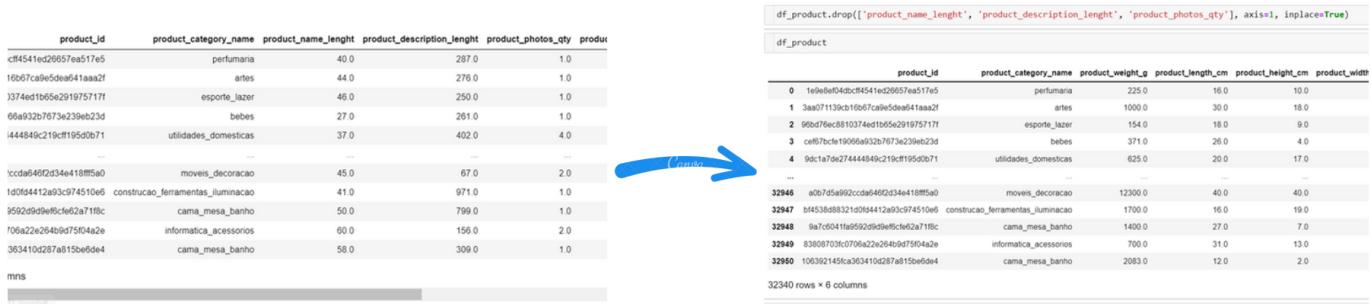
Before

After

Hình 3.33: Xóa giá trị Null

BÁO CÁO CÁ NHÂN

3. Xóa các cột không sử dụng. Trong bảng product, các cột tên sản phẩm, độ dài tên sản phẩm không được sử dụng nên ta sẽ xóa chúng đi.



df_product.drop(['product_name_lenght', 'product_description_lenght', 'product_photos_qty'], axis=1, inplace=True)

product_id	product_category_name	product_name_lenght	product_description_lenght	product_photos_qty	product
cff4541ed26657ea517e5	perfumaria	40.0	287.0	1.0	
16b67ca9e5de6a41aa2f	artes	44.0	276.0	1.0	
374ed1b5e291975717f	esporte_lazer	46.0	250.0	1.0	
66a932b7673e239eb23d	bebés	27.0	261.0	1.0	
444849c219cff195d0b71	utilidades_domesticas	37.0	402.0	4.0	
...	
cccd6469fd34e418ff5a0	moveis_decoracao	45.0	67.0	2.0	
1d0f44121a93c974510e6	construcao_ferramentas_luminacao	41.0	971.0	1.0	
9592df9d9fc6e52a71f8c	cama_mesa_banho	50.0	799.0	1.0	
106a22e264b0d75f04a2e	informatica_acessorios	60.0	156.0	2.0	
363410d287a815be6de4	cama_mesa_banho	58.0	309.0	1.0	

product_id	product_category_name	product_weight_g	product_length_cm	product_height_cm	product_width_cm
0 1e0e8ef04dbcd4541ed26657ea517e5	perfumaria	225.0	16.0	10.0	14.0
1 3aa07113cb1b67ca9e5de6a41aa2f	artes	1000.0	30.0	18.0	20.0
2 96bd7f6c8883210594412a93c974510e6	esporte_lazer	154.0	18.0	9.0	15.0
3 ce97bcf19056a9e52b787e239eb23d	bebés	371.0	26.0	4.0	26.0
4 9dc1a7de27444849c219cff195d0b71	utilidades_domesticas	625.0	20.0	17.0	13.0
...
2946 a0b7d5a992ccda54692d34e418ff5a0	moveis_decoracao	12300.0	40.0	40.0	40.0
2947 b45358883210594412a93c974510e6	construcao_ferramentas_luminacao	1700.0	16.0	19.0	16.0
2948 9a7c0416f929d9e9fc6e52a71f8c	cama_mesa_banho	1400.0	27.0	7.0	27.0
2949 83080703c070fa22e2049d75f04a2e	informatica_acessorios	700.0	31.0	13.0	20.0
2950 106392145ka363410d287a815be6de4	cama_mesa_banho	2083.0	12.0	2.0	12.0

32340 rows × 6 columns

Before

After

Hình 3.34: Xóa các cột không sử dụng

4. Thêm cột thể tích sản phẩm. Trong bảng product, thêm cột thể tích sản phẩm để có thể phân tích đa chiều.



df_product['volume'] = df_product['product_length_cm'] * df_product['product_height_cm'] * df_product['product_width_cm']

product_id	product_category_name	product_weight_g	product_length_cm	product_height_cm	product_width_cm
57ea517e5	perfumaria	225.0	16.0	10.0	14.0
sa641aaa2f	artes	1000.0	30.0	18.0	20.0
391975717f	esporte_lazer	154.0	18.0	9.0	15.0
e239eb23d	bebés	371.0	26.0	4.0	26.0
ff195d0b71	utilidades_domesticas	625.0	20.0	17.0	13.0
...
fe418ff5a0	moveis_decoracao	12300.0	40.0	40.0	40.0
ic974510e6	construcao_ferramentas_luminacao	1700.0	16.0	19.0	16.0
fe62a71f8c	cama_mesa_banho	1400.0	27.0	7.0	27.0
jd75f04a2e	informatica_acessorios	700.0	31.0	13.0	20.0
815be6de4	cama_mesa_banho	2083.0	12.0	2.0	7.0

product_id	product_category_name	product_weight_g	product_length_cm	product_height_cm	product_width_cm	volume
0 1e0e8ef04dbcd4541ed26657ea517e5	perfumaria	225.0	16.0	10.0	14.0	2240.0
1 3aa07113cb1b67ca9e5de6a41aa2f	artes	1000.0	30.0	18.0	20.0	10800.0
2 96bd7f6c8883210594412a93c974510e6	esporte_lazer	154.0	18.0	9.0	15.0	2430.0
3 ce97bcf19056a9e52b787e239eb23d	bebés	371.0	26.0	4.0	26.0	2740.0
4 9dc1a7de27444849c219cff195d0b71	utilidades_domesticas	625.0	20.0	17.0	13.0	4420.0
...
2946 a0b7d5a992ccda54692d34e418ff5a0	moveis_decoracao	12300.0	40.0	40.0	40.0	40.0 4000.0
2947 b45358883210594412a93c974510e6	construcao_ferramentas_luminacao	1700.0	16.0	19.0	16.0	16.0 4864.0
2948 9a7c0416f929d9e9fc6e52a71f8c	cama_mesa_banho	1400.0	27.0	7.0	27.0	5103.0
2949 83080703c070fa22e2049d75f04a2e	informatica_acessorios	700.0	31.0	13.0	20.0	8060.0
2950 106392145ka363410d287a815be6de4	cama_mesa_banho	2083.0	12.0	2.0	7.0	7.0 168.0

340 rows × 7 columns

Before

After

Hình 3.35: Thêm cột thể tích

BÁO CÁO CÁ NHÂN

5. Thêm cột có điều kiện. Trong bảng product, thêm cột loại sản phẩm dựa vào thể tích của sản phẩm.

The screenshot shows two tables side-by-side. The left table, labeled 'Before', contains columns: product_id, product_category_name, product_weight_g, product_length_cm, product_height_cm, product_width_cm, and volume. The right table, labeled 'After', includes the same columns plus a new column 'size_product'. A blue arrow points from the 'Before' table to the 'After' table, indicating the transformation. The 'size_product' column is defined by the following Python code:

```

def addSizeProduct(df_product):
    size_product = []
    for volume in df_product['volume']:
        if volume < 1000:
            size_product.append('small')
        elif volume < 5000:
            size_product.append('middle')
        elif volume < 20000:
            size_product.append('big')
        else:
            size_product.append('very big')
    df_product['size_product'] = size_product
    return df_product

```

Hình 3.36: Thêm cột có điều kiện

6. Merge query. Merge query bảng product và bảng product_category để lấy cột product_type.

The screenshot shows two tables. The left table, labeled 'Before', is the original product DataFrame. The right table, labeled 'After', is the merged DataFrame where a new column 'product_type' has been added. A blue arrow points from the 'Before' table to the 'After' table, indicating the merge operation. The merged DataFrame includes columns: product_id, product_category_name, size_product, and product_type. The product_type values are derived from the product_category_name values using a merge operation:

```

df_product = pd.merge(df_product, df_product_name, how="left", on= "product_id")
df_product.drop("product_category_name_english", axis=1, inplace=True)

```

Hình 3.37: Merge query

BÁO CÁO CÁ NHÂN

7. Chính lại trường dữ liệu thời gian. Do trong quá trình phân tích không sử dụng yếu tố giờ nên ở cột thời gian, nhóm em loại bỏ thành phần giờ và chỉ giữ lại ngày, tháng, năm.

The diagram illustrates the transformation of the 'shipping_limit_date' column. On the left, labeled 'Before', a screenshot of a Jupyter Notebook shows a DataFrame with a timestamp column. A blue arrow points from this 'Before' state to the right, where the code `df_olist_item['shipping_limit_date'] = pd.to_datetime(df_olist_item['shipping_limit_date']).dt.normalize()` is shown, resulting in a new DataFrame where the dates are now represented as strings without time components. This transformed DataFrame is labeled 'After'.

product_id	seller_id	shipping_limit_date	price
4244733e06e7ecb04970a6e2683c13e61	48436dade18ac8b2bce089ec2a041202	2017-09-19 09:45:35	58.90
e5f2d52b602189ee658865ca93d83a8f	dd7dc04e1b6c2c614352b383fe2d36	2017-05-03 11:05:13	239.90
c777355d18b72b67abbee9fd44f00d	5b51032ed0d242ad0c84c38acab88f23d	2018-01-18 14:48:30	199.00
7634da152a4610f1595ef3a32f14722fc	9d7a1d3a5a5052409006425275ba1c2b4	2018-08-15 10:10:18	12.99
ac6c3623068f30de03045865e4e10089	df560393f3a51e74553ab94004ba5c87	2017-02-13 13:57:51	199.90

product_id	seller_id	shipping_limit_date	price
0 00010242fe8c5a6d1ba2dd792cb16214	1 4244733e06e7ecb04970a6e2683c13e61	2017-09-19	58.90
1 00018ff7f2f0320c557190d7a144bd53	1 e5f2d52b602189ee658865ca93d83a8f	2017-05-03	239.90
2 000229ec398224ef6ca065784fc703e	1 c777355d18b72b67abbee9fd44f00d	2018-01-18	199.00
3 00024acbcdf0a6daa1e931b38114c75	1 7634da152a4610f1595ef3a32f14722fc	2018-08-15	12.99
4 00042b26cf59d7ce69dfabb4e55b4fd9	1 ac6c3623068f30de03045865e4e10089	2017-02-13	199.90

order_id	order_item_id	product_id	seller_id	shipping_limit_date	price
112645 mfc94f6ce0a00581880bf54a75a037	1 4aa6014eceb6820779dc4bfebc05b0	b6bc237ba3788b23da09c0f1f3a3288c	2018-05-02	299.99	
112646 mfc0d46f2263f404302a634eb57f7eb	1 32e07f915822b0765e448c4dd74c628	f3c38ab652836d21de61fb8314b69182	2018-07-20	350.00	

Hình 3.38: Chính lại trường dữ liệu thời gian

8. Đặt lại kiểu dữ liệu cho cột. Chính lại kiểu dữ liệu của cột *payment_installments* trong bảng *olist_order*.

The diagram illustrates the transformation of the 'payment_installments' column. On the left, labeled 'Before', a screenshot of a Jupyter Notebook shows a DataFrame with a float64 column. A blue arrow points from this 'Before' state to the right, where the code `df_olist_order['payment_installments'] = df_olist_order['payment_installments'].astype('int64')` is shown, resulting in a new DataFrame where the column is now of type int64. This transformed DataFrame is labeled 'After'.

order_id	payment_sequential	payment_type	payment_installments
0 b61ef226f3fe1789b1e8b2acc339d17	1 credit_card	8	
1 a98610da82917af2d9ae8fd1278f1dcfa0	1 credit_card	1	
2 256ea4e9339665fa0633d708e76cc1bd	1 credit_card	1	
3 ba7699792108cc13738041e913ab953	1 credit_card	8	
4 42fdf880ba16b47b592515dd489d441a	1 credit_card	2	

order_id	payment_sequential	payment_type	payment_installments
103881 0406037ad97744d563a178ecc7a2075c	1 boleto	1	
103882 7b905661d7c825891d6347454ea7863f	1 credit_card	2	
103883 326090bb3d96993c066a6860554a77bf	1 credit_card	1	
103884 b861059626efa996a60be9b09320e10	1 credit_card	5	
103885 28bbeae6599b09d39caa060b747b6632b1	1 boleto	1	

order_id	payment_sequential	payment_type	payment_installments
103886 rows × 5 columns			


```
print(df_olist_order['payment_installments'].dtypes)
float64
```


order_id	payment_sequential	payment_type	payment_installments
103881 0406037ad97744d563a178ecc7a2075c	1 boleto	1	
103882 7b905661d7c825891d6347454ea7863f	1 credit_card	2	
103883 326090bb3d96993c066a6860554a77bf	1 credit_card	1	
103884 b861059626efa996a60be9b09320e10	1 credit_card	5	
103885 28bbeae6599b09d39caa060b747b6632b1	1 boleto	1	


```
df_olist_order['payment_installments'] = df_olist_order['payment_installments'].astype('int64')
print(df_olist_order['payment_installments'].dtypes)
int64
```

Before

After

Hình 3.39: Đặt lại kiểu dữ liệu cho cột

3.7.3 Kiến trúc hệ thống OLTP

Kiến trúc hệ thống

OLTP - Online transactional processing (xử lý giao dịch trực tuyến) được đặc trưng bởi một số lượng lớn các giao dịch trực tuyến ngắn (chèn, cập nhật, 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 được đo bằng số lượng giao dịch mỗi giây. Trong cơ sở dữ liệu OLTP có dữ liệu chi tiết và 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).

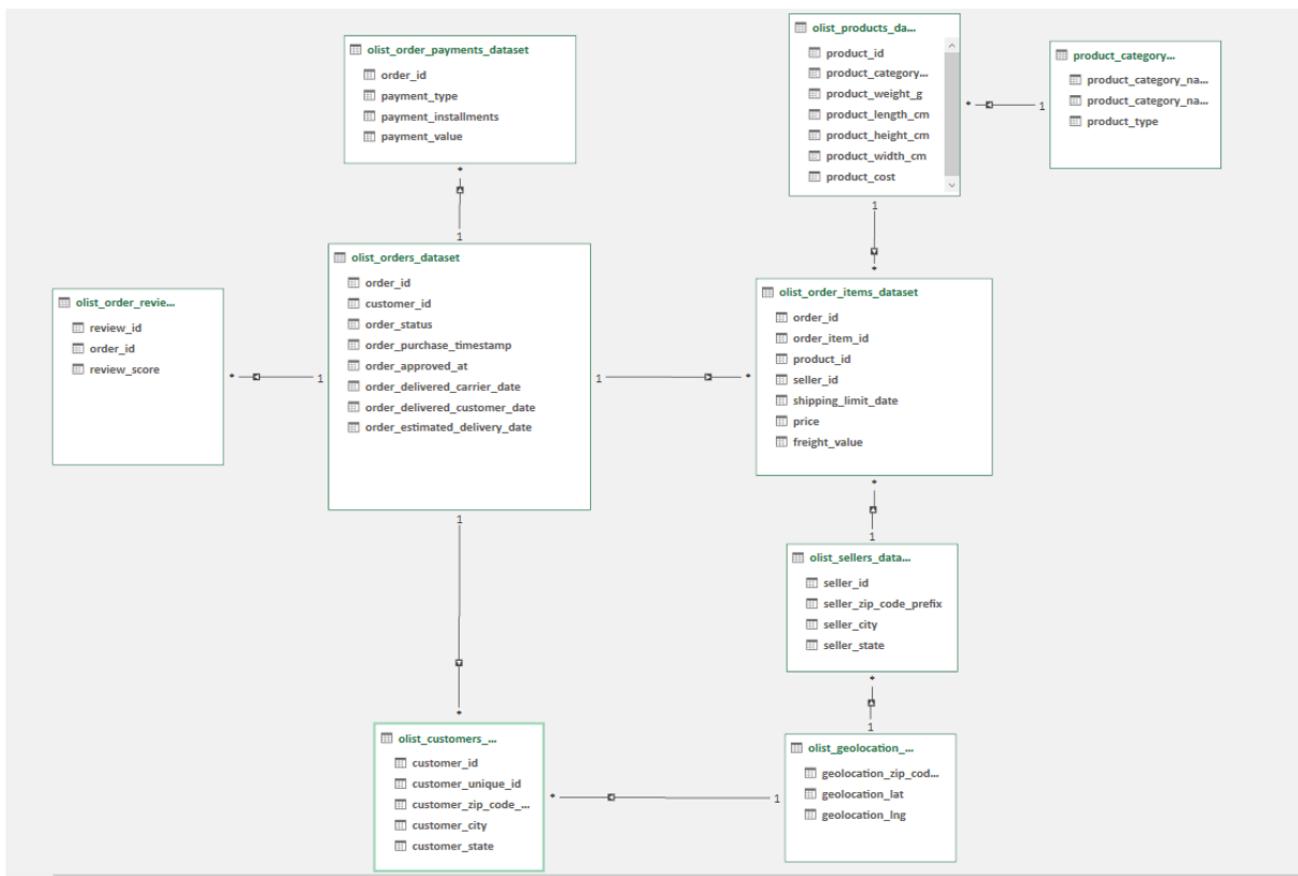
Mô hình dữ liệu quan hệ OLTP

Từ bộ dữ liệu thương mại điện tử của công ty Olist, chúng em xây dựng mô hình dữ liệu quan hệ bao gồm các bảng như sau:

- olist_orders_dataset: thông tin về các đơn đặt hàng.
- olist_order_item_dataset: thông tin chi tiết về món hàng trong đơn hàng.
- olist_products_dataset: thông tin về các loại sản phẩm.
- olist_order_payments_dataset: thông tin về thanh toán đơn hàng.
- olist_order_reviews_dataset: nhận xét và đánh giá của khách hàng về đơn hàng.
- olist_customers_dataset: thông tin của khách hàng.
- olist_sellers_dataset: thông tin của người bán hàng.
- olist_geolocation_dataset: thông tin địa lý về các thành phố ở Brazil.
- olist_product_category: thông tin về tên sản phẩm được dịch theo tiếng anh.

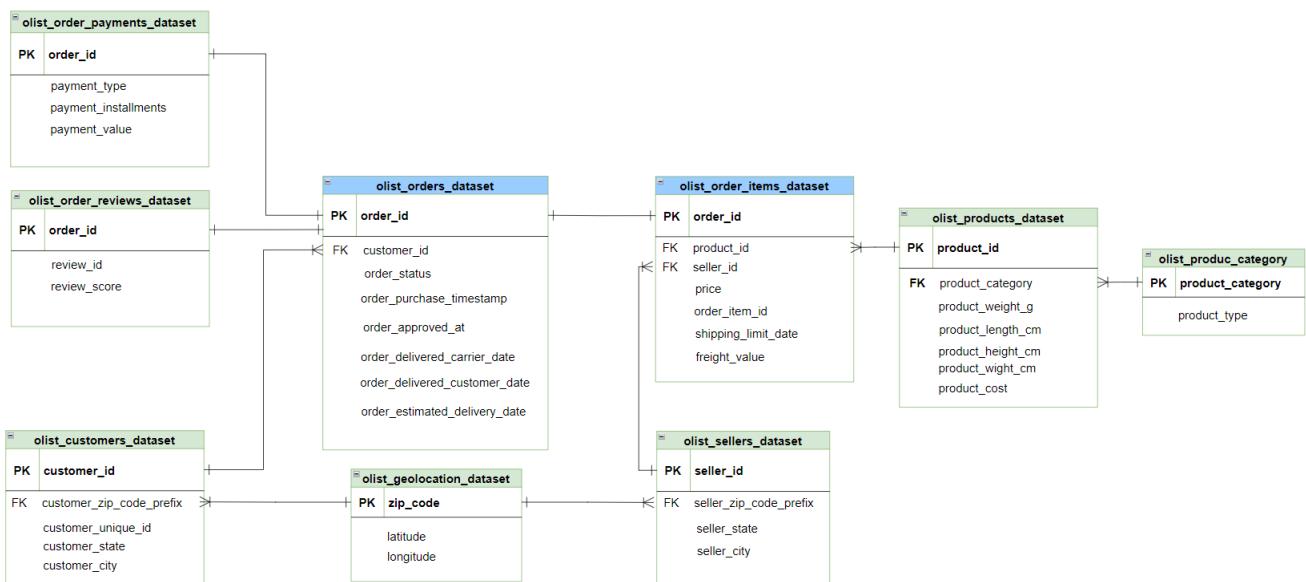
BÁO CÁO CÁ NHÂN

1. Mô hình OLTP



Hình 3.40: OLTP model

2. Mô hình RE OLTP



Hình 3.41: Mô hình RE OLTP

3.7.4 Phân tích chiều dữ liệu (Dimension) và chủ điểm (Fact)

Phân tích các Dimension

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 tổ chức các thuộc tính mô tả dưới dạng các cột trong bảng dim.

Ví dụ về các dimension: thuộc tính của dim khách hàng có thể bao gồm họ và tên, ngày sinh, giới tính, v.v. hoặc dim trang web sẽ bao gồm tên trang web và thuộc tính URL. Bảng dim có cột khóa chính xác định duy nhất mỗi bản ghi dim (hàng). Bảng dim được liên kết với bảng dữ kiện bằng cách sử dụng khóa này. Dữ liệu trong bảng dữ liệu có thể được lọc và nhóm (“cắt nhỏ và cắt hạt lựu”) theo nhiều cách kết hợp thuộc tính khác nhau. Nhiều Dim chứa một hệ thống phân cấp các thuộc tính hỗ trợ việc khoan lén và xuồng. Ví dụ: Dim Ngày có thể chứa phân cấp năm - quý - tháng - tuần - 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 (BI) và kinh doanh (BA).

Đối với bộ dữ liệu thương mại điện tử của công ty Olist nhóm chúng em phân tích thành các Dimension sau:

- Dim_product_name: bao gồm tên của các sản phẩm và có tổng cộng 71 bản ghi.
- Dim_order_type: gồm các loại đơn hàng từ nhỏ (small) đến các đơn hàng rất lớn (very big), có tổng cộng 5 bản ghi.
- Dim_order_status: gồm các trạng thái của đơn hàng như delivered, canceled, ... và có 8 bản ghi.
- Dim_product_size: gồm kích cỡ của các sản phẩm (small, middle, big, very big), tổng cộng 4 bản ghi.
- Dim_product_type: gồm các loại sản phẩm như cosmetic, technology, handmade, ... và có 9 bản ghi.
- Dim_payment_type: gồm các hình thức thanh toán như credit_card, boleto, voucher, debit_card và có tổng cộng 4 bản ghi.
- Dim_installment_group: gồm các loại hình thức trả góp như one_shot, shot_term, mid_term, long_term và có tổng cộng là 4 bản ghi.
- Dim_customer: gồm có 2 trường là customer_city và customer_state.
- Dim_seller: gồm có 2 trường là seller_city và seller_state.
- Dim_Date: gồm có 597 bản ghi là thời gian bán hàng từ 5/1/2017 đến ngày 29/8/2018.

BÁO CÁO CÁ NHÂN



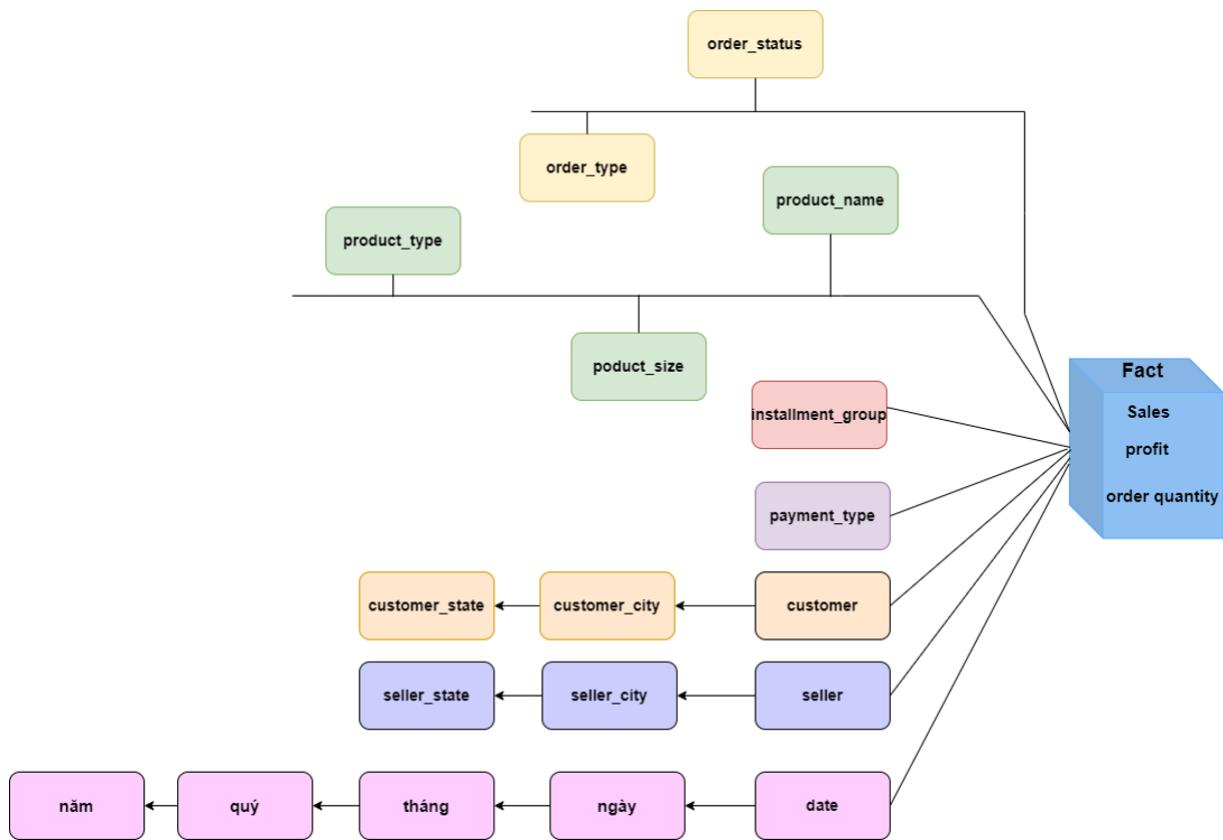
Hình 3.42: Phân tích Dimension

Phân tích chủ điểm Fact

Chủ điểm phân tích là trung tâm trong lược đồ hình sao của một kho dữ liệu. Đây là một khái niệm quan trọng cần thiết cho Kho dữ liệu và Chứng nhận BI. Chủ điểm 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 lưu trữ dữ liệu về các 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 trong báo cáo này, các fact chính được phân tích đó là: doanh thu (Sales), lợi nhuận (Profit) và số lượng đơn đặt hàng (Order quantity).

Mô hình logic



Hình 3.43: Mô hình logic

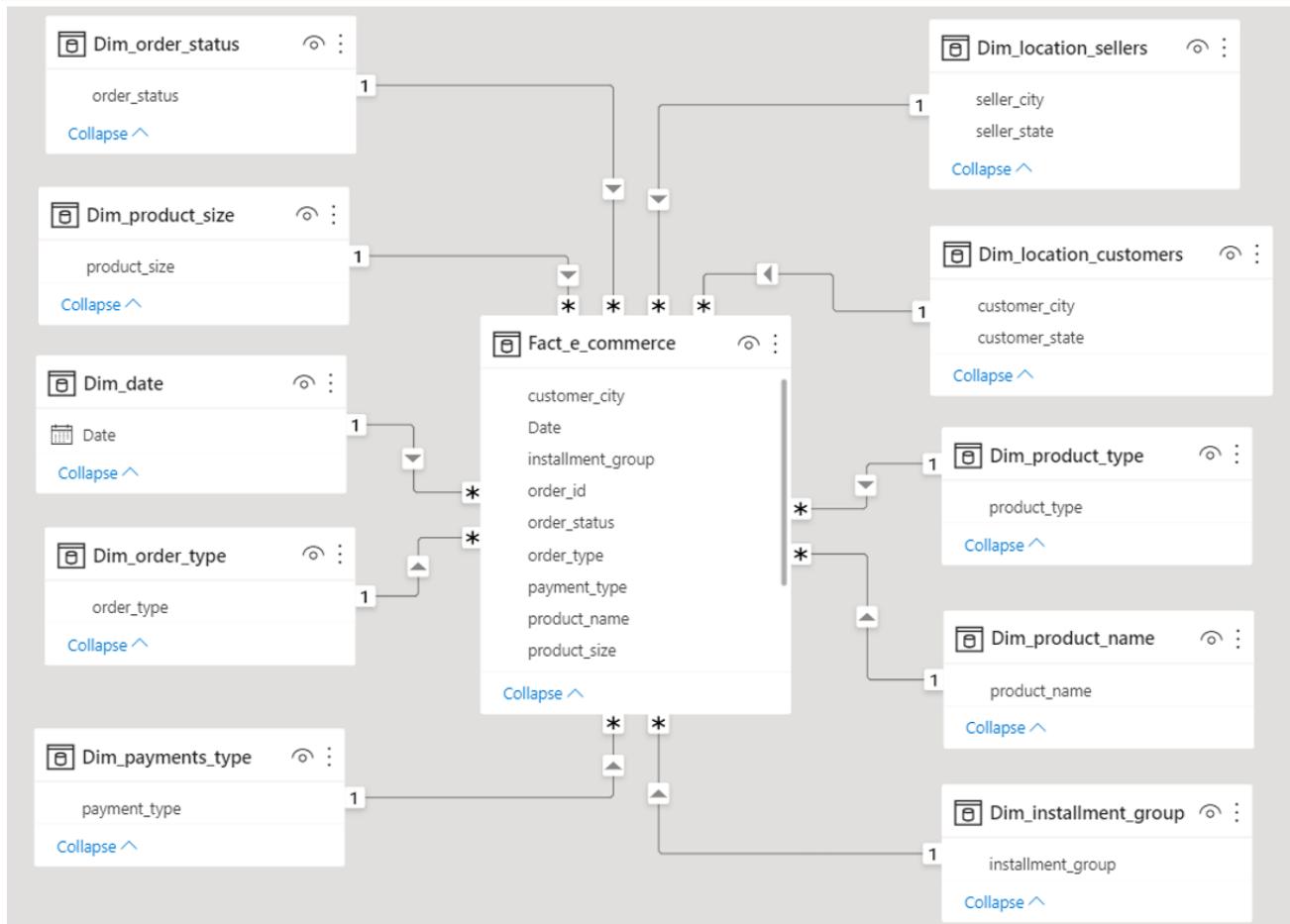
3.7.5 Thiết kế hệ thống OLAP

Mục đích của mô hình OLAP

- Hạn chế tối đa mức độ ảnh hưởng phân tách tài nguyên của quá trình trích xuất, chuyển đổi dữ liệu.
- Cung cấp hệ thống phân phối báo cáo chuyên nghiệp đến người sử dụng cuối.
- Cụ thể trong báo cáo này: phân tích dữ liệu đa chiều theo 3 chủ điểm đã nêu trên.

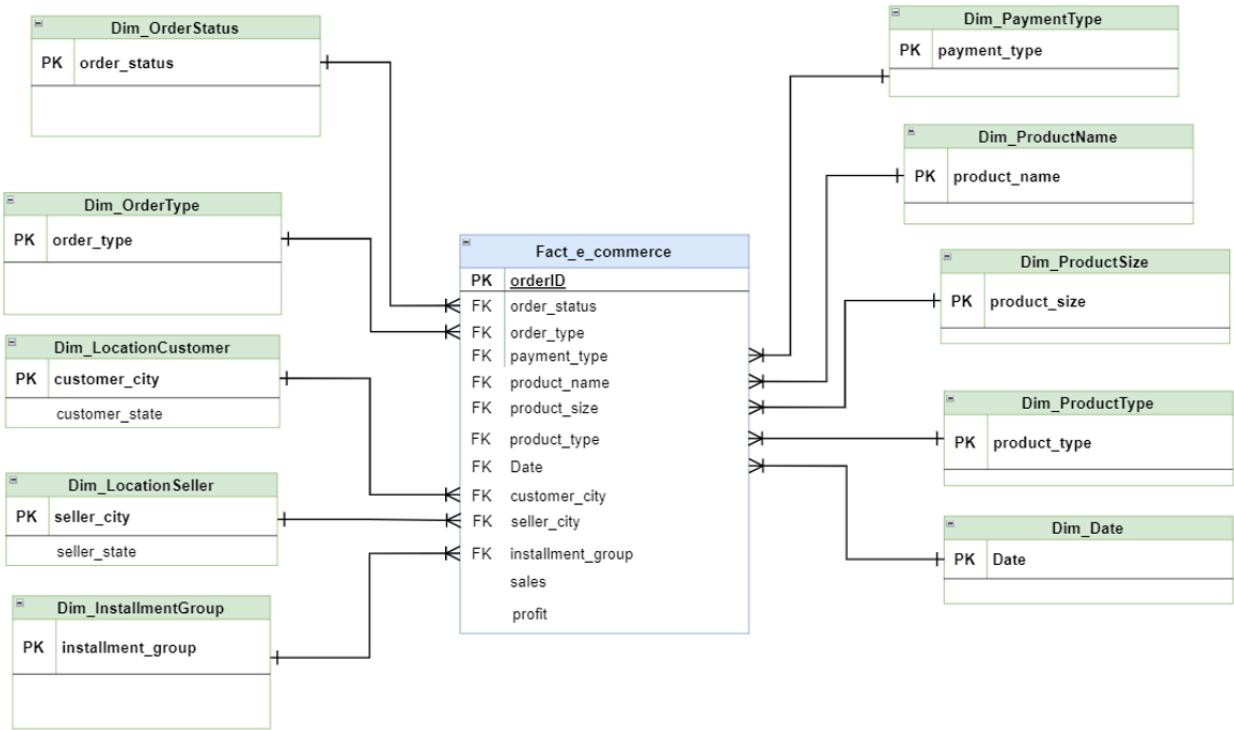
BÁO CÁO CÁ NHÂN

Mô hình OLAP



Hình 3.44: Mô hình OLAP

Mô hình dữ liệu quan hệ ERD OLAP

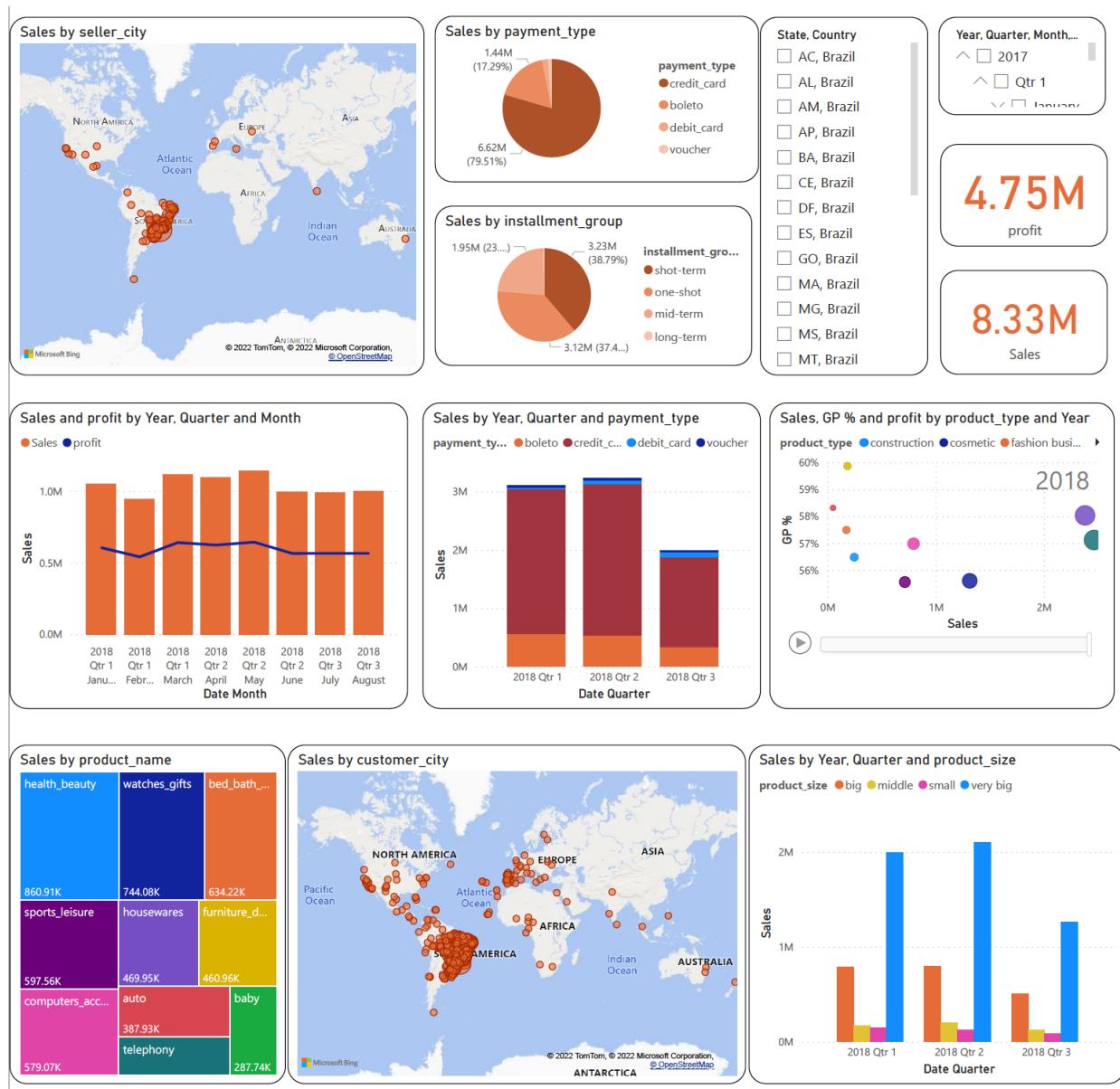


Hình 3.45: Mô hình dữ liệu quan hệ ERD OLAP

3.8 Xây dựng Dashboard

Qua việc phân tích requirements và dimension đã được đề cập ở phần trước. Nhóm sẽ xây dựng dashboard dựa trên bộ dữ liệu Brazilian E-Commerce của chuỗi Olist từ 01/2017 đến 08/2018 theo các chủ đề: Sales, số lượng đơn đặt hàng, Profit.

3.8.1 Phân tích Dashboard dựa trên doanh thu: Sales



Hình 3.46: Dashboard dựa trên doanh thu

Nhìn chung doanh thu các mặt hàng đa số tăng từ năm 2017 đến năm 2018, riêng chỉ có loại hàng trong xây dựng (construction) giảm nhưng GP% lại tăng. Điều này cho thấy mặc dù doanh thu của loại mặt

hàng giảm nhưng lợi nhuận tăng cao, chứng tỏ lợi nhuận từ các loại mặt hàng xây dựng rất lớn. Chính vì vậy, Olist cần đẩy mạnh chiến dịch quảng bá sản phẩm, cũng như các chương trình khuyến mãi giảm giá, tăng sự đa dạng về sản phẩm để giúp người dân có thêm nhiều sự lựa chọn mua hàng.

Nhìn vào biểu đồ ta thấy doanh thu tăng hẳn vào tháng 11 năm 2017 - tháng có ngày lễ siêu giảm giá Black Friday, cũng là tháng trước ngày lễ lớn nhất trong năm - Christmas nên nhu cầu của người dân tăng mạnh và người dân cũng tiết kiệm được chi phí đáng kể khi mua đồ trong tuần lễ siêu giảm giá. Vậy nên Olist cần đảm bảo hệ thống ổn định khi có quá nhiều lượt truy cập, không gây overload,... và đẩy mạnh chiến dịch marketing đến người dân.

Khu vực Sao paulo có doanh thu lớn nhất, đây là điều dễ hiểu khi São Paulo là thành phố lớn nhất của Brazil và cũng là thành phố đông dân nhất ở Nam bán Cầu. Chính vì vậy nhu cầu của người dân rất lớn, Olist cần tập trung đẩy mạnh về sự đa dạng sản phẩm, dịch vụ tại đây.

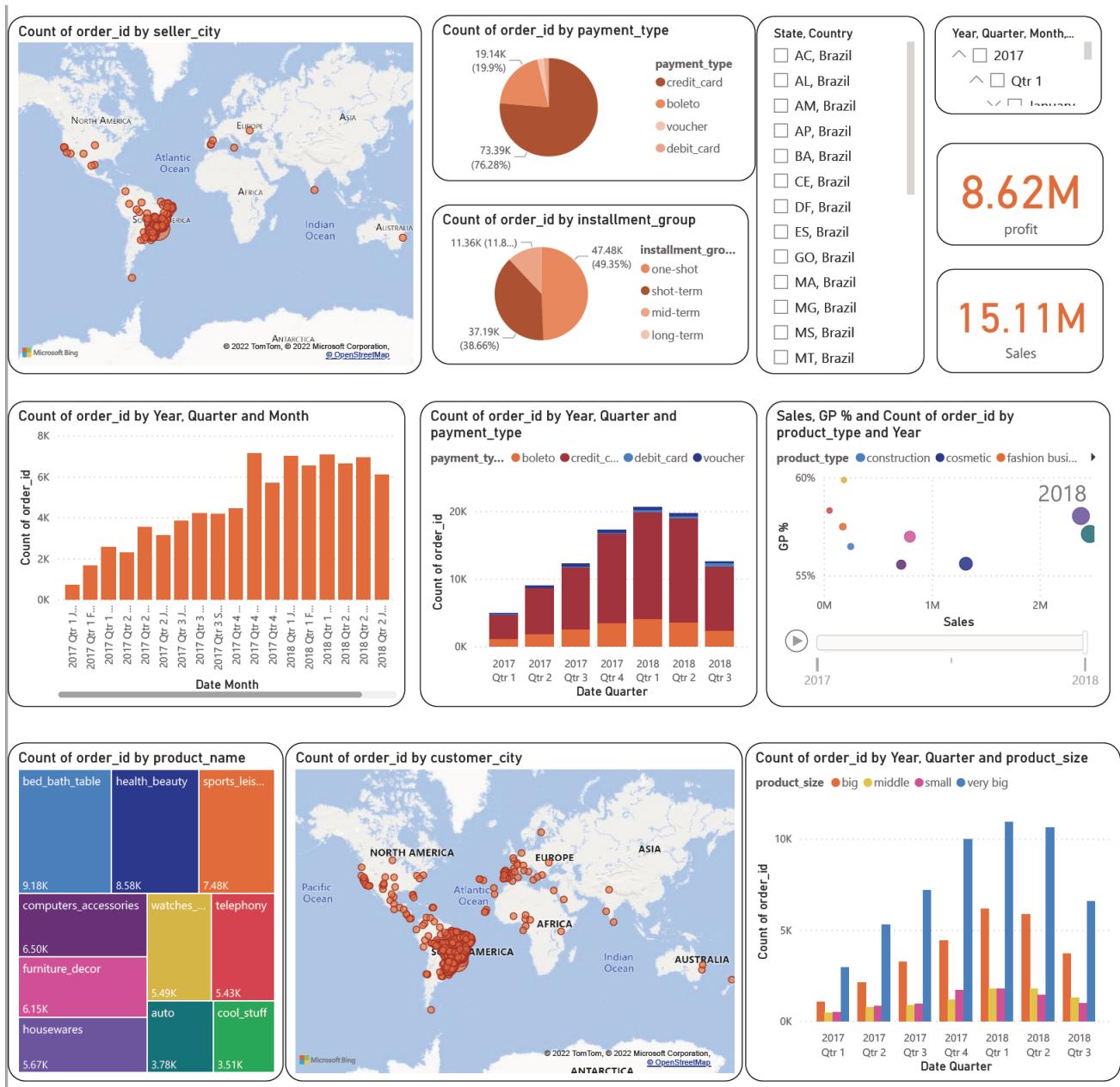
Mặt hàng mọi người mua chủ yếu liên quan đến health & beauty điều này cho thấy người dân chú trọng vào việc chăm sóc sức khoẻ bản thân rất nhiều. Olist cần bổ sung thêm đa dạng sản phẩm, mặt hàng liên quan đến sức khoẻ, làm đẹp đa dạng để phù hợp với dân số có sự đa dạng về chủng tộc như ở Brazil.

Sản phẩm kích thước very big, tập trung chủ yếu vào mặt hàng house hold cho thấy xu hướng người dân thích mua sản phẩm lớn để sinh hoạt thoả mái.

Về phương thức thanh toán, người dân có xu hướng trả bằng credit card chiếm đa số (78.42% vào năm 2017 và 79.51% vào năm 2018) người dân tận hưởng việc mua mặt hàng mình có nhu cầu sử dụng trước và sau đó cuối tháng có lương bù vào sau. Năm 2018, tỉ lệ người dân trả luon một lần chiếm 37.48% đứng sau việc trả góp trong 6 tháng là short-term và mid-term – 12 tháng lần lượt là 38.79% và 23.36%. Còn đối với năm 2017, tỉ lệ người dân trả một lần, trả góp trong vòng 6 tháng, 12 tháng lần lượt là: 36.03%, 36.75%, 26.64%.

Sau phương thức trả bằng credit card, người dân còn trả bằng boleto, boleto là dạng phiếu mua hàng tại Brazil có thời hạn sử dụng. Chính vì vậy người dân tại Brazil cũng thường trả bằng boleto vì nếu không sử dụng thì boleto sẽ hết hạn và không có giá trị sử dụng nữa.

3.8.2 Phân tích Dashboard dựa trên số lượng đơn đặt hàng: Order quantity



Hình 3.47: Dashboard dựa trên số lượng đơn đặt hàng

Nhìn chung số lượng đơn hàng người dân đặt hàng tăng cao vào khoảng tháng 11/2017 và những tháng đầu năm của năm 2018. Có thể thấy nhu cầu mua sắm vào dịp cuối năm và đầu năm của người dân lớn, chính vì vậy Olist cần có những chính sách để đây mạnh nhu cầu mua sắm của người dân hơn nữa để tăng doanh thu. Mặt khác, Olist cần có những kế hoạch về inventory để có thể vừa đáp ứng đủ nhu cầu mua hàng của khách hàng và vừa giảm chi phí lưu kho của công ty.

Về phương thức thanh toán đối với số lượng đơn đặt hàng phần đa là người dân thanh toán bằng credit card và sau đó là Boleto.

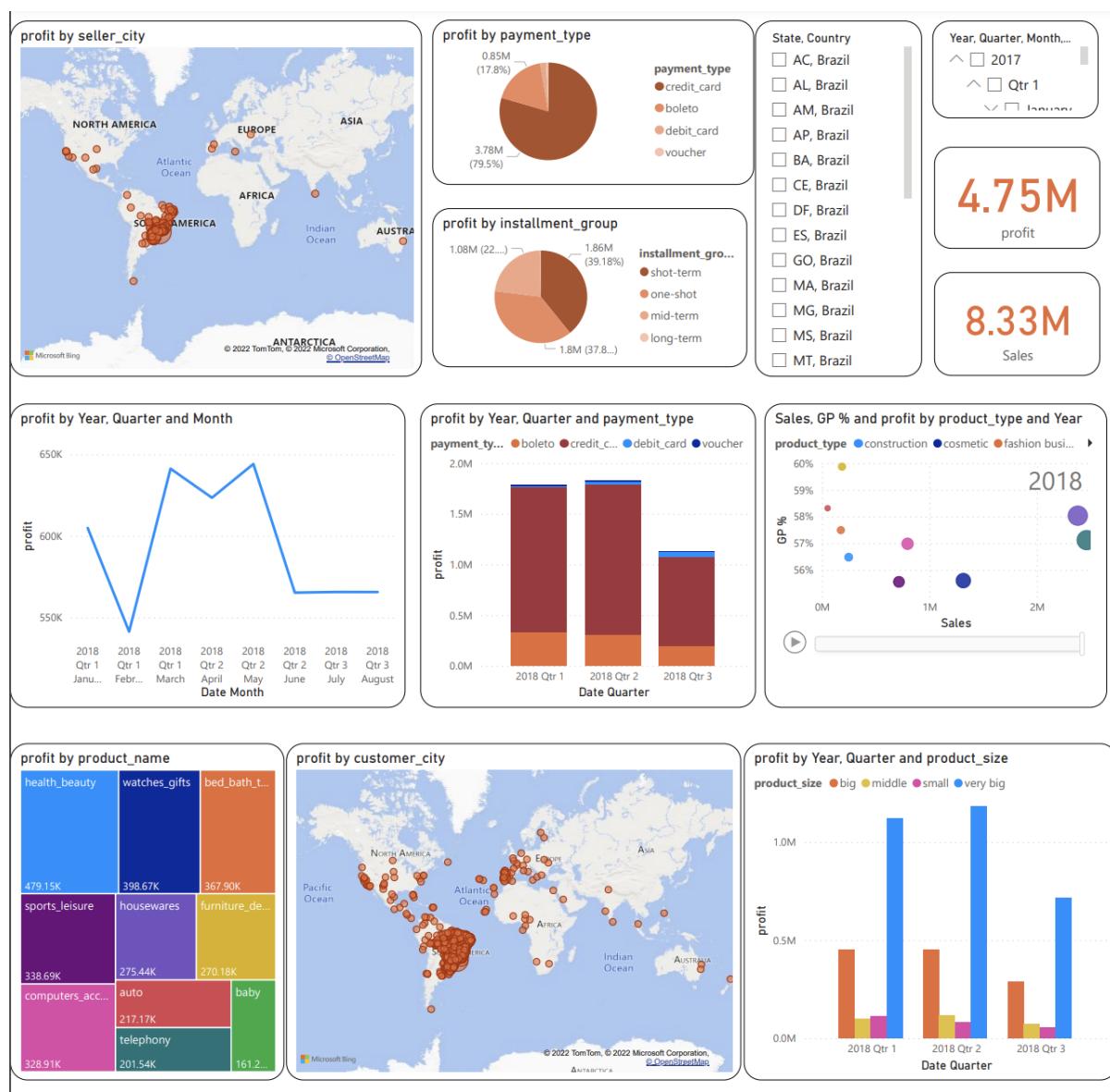
Người dân đặt hàng về bed, bath, table (các mặt hàng liên quan đến household) cho thấy người dân rất

chú trọng về việc mua sắm các dụng cụ phục vụ sinh hoạt chung trong gia đình. Người dân cũng mua chủ yếu các loại mặt hàng có kích thước lớn. Dựa trên những hành vi mua sắm trong suốt thời gian 01/2017 đến 08/2018, Olist cần chú trọng phát triển các mặt hàng về house hold đa dạng để phục vụ nhu cầu của người dân, làm tăng doanh thu.

Đơn đặt hàng cũng tập trung chủ yếu tại São Paulo - thành phố lớn tại Brazil.

Đơn đặt hàng nhiều về household với size sản phẩm lớn, người dân xu hướng trả trong 1 lần duy nhất, trả bằng credit card. Điều này cho thấy rằng thu nhập người dân tăng cao nên có thể chi trả những đồ vật trong nhà theo sở thích và nhu cầu của gia đình và có thể trả trong 1 lần luon.

3.8.3 Phân tích Dashboard dựa trên lợi nhuận: Profit



Hình 3.48: Dashboard dựa trên lợi nhuận

BÁO CÁO CÁ NHÂN

Lợi nhuận Olist tháng 11/2017 và những tháng đầu năm 2018 đạt đỉnh. So sánh với dashboard về số lượng đơn đặt hàng trước đó, ta có thể thấy, số lượng đơn đặt hàng tăng thì lợi nhuận cũng tăng.

Lợi nhuận thu về từ health & beauty chiếm phần lớn, cũng giống như doanh thu của mặt hàng này. Điều này cho thấy, giá trị sản phẩm của những loại mặt hàng này cao và lợi nhuận cũng cao so với các loại mặt hàng còn lại trong bộ dữ liệu này. Đứng ở vị trí thứ hai là về household: bed, bath, table tương ứng với mặt hàng có số lượng đơn đặt hàng nhiều nhất. Từ hai biểu đồ lợi nhuận dựa trên loại mặt hàng, ta thấy lợi nhuận tập trung chủ yếu vào: cosmetic và household. Chính vì vậy Olist cần có chiến lược marketing cụ thể trong thời gian ngắn cũng như nâng cao chất lượng, sự đa dạng sản phẩm để người dân tin dùng mua sắm hơn.

Người dân lựa chọn thanh toán credit card là chủ yếu chiếm 79.03% lợi nhuận, và người dân thường trả góp trong thời gian ngắn - 6 tháng. Với mặt hàng household việc người dân lựa chọn trả góp là lựa chọn khôn ngoan. Chính vì vậy, trong tương lai Olist ngoài việc tập trung phát triển sản phẩm, cần tập trung vào những chính sách hỗ trợ khách hàng trong khâu: mua sắm, tư vấn, phương thức thanh toán, hậu mãi,...

Mặt hàng kích thước very big vẫn được ưu chuộng và đem lại lợi nhuận rất lớn, lớn hơn hẳn lợi nhuận từ cùng loại mặt hàng có kích thước nhỏ hơn: middle, small.

Kết luận

Kết luận tổng thể của báo cáo

Dưới sự làm việc nghiêm túc của em và sự hướng dẫn tận tình của thầy Nguyễn Danh Tú, bài báo cáo đã đạt được những mục tiêu đề ra ban đầu và một số kết quả sau:

1. Trình bày ngắn gọn, súc tích cơ sở lý thuyết của Data Warehouse và Business Intelligence.
2. Trình bày chi tiết được các bước để xây dựng một Data Warehouse ứng với vấn đề thực tế quản lý bán hàng trên sàn thương mại điện tử.
3. Phân tích các Dashboard trực quan, sinh động; dự báo được xu hướng mua bán trong tương lai.

Bài học thu được

Một số bài học mà em thu được sau quá trình học và thực hành nhóm:

1. Việc xây dựng kho dữ liệu là vô cùng cần thiết cho mục đích phân tích dữ liệu.
2. Hiểu rõ về nghiệp vụ chuyên môn là vô cùng quan trọng nên cần khảo sát kỹ để hiểu bộ dữ liệu trước khi phân tích. Kiến thức nghiệp vụ rất quan trọng để tạo ra Data warehouse phù hợp và sử dụng được nó cho các phân tích báo cáo.
3. Dữ liệu trong thực tế không phải lúc nào đạt chuẩn (dữ liệu sạch) vì vậy cần phải có quá trình ETL dữ liệu cẩn thận để biến từ dữ liệu gốc thô thành dữ liệu sạch có thể sử dụng để xây dựng mô hình OLAP.
4. Việc phân tích dữ liệu với mô hình đa chiều OLAP là rất dễ dàng, trực quan, nhanh chóng.
5. Việc vẽ các Dashboard thì với mỗi biểu đồ cần phải sử dụng thêm yếu tố thời gian để có thể phân tích được xu hướng phát triển cũng như dự báo trong tương lai.

Tài liệu tham khảo

- [1] Nguyễn Danh Tú, 2020, *Slide bài giảng Kho dữ liệu và kinh doanh thông minh*, Đại học Bách khoa Hà Nội.
- [2] Oracle website, Data Warehouse Defined: <https://www.oracle.com/database/what-is-a-data-warehouse/>.
- [3] Fanpage Facebook *Phân tích dữ liệu* - <https://www.facebook.com/Phân tích số liệu>.
- [4] Kênh Youtube *Học excel cơ bản* - <https://www.youtube.com/Học excel cơ bản>.
- [5] Kênh Youtube *Power BI of Pavan Lalwami* - <https://www.youtube.com/Power BI Full Online Training by Pavan Lalwami>.