



ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY



VIỆN TOÁN ỨNG DỤNG VÀ TIN HỌC
School of Applied Mathematics and Informatics

Kho dữ liệu và kinh doanh thông minh

Nhóm 9 - Chủ đề: Thương mại điện tử Olist

GVHD: ThS. Nguyễn Danh Tú

Nguyễn Thị Vân Anh 20200036

Nguyễn Lương Quỳnh Trang 20206175

Bùi Thanh Tùng 20206181

Hà Nội, Ngày 13 tháng 8 năm 2023

Lời mở đầu

Các công cụ kinh doanh thông minh đã mở ra cơ hội cho doanh nghiệp thuộc mọi quy mô tiếp cận với khả năng phân tích dữ liệu mạnh mẽ. Tận dụng được nguồn tài nguyên dữ liệu dồi dào và tìm ra xu hướng là điều cần thiết để các doanh nghiệp thúc đẩy các quyết định kinh doanh tốt hơn, đưa ra những chiến lược hiệu quả cho phép các tổ chức tăng doanh thu, cải thiện hiệu suất hoạt động và hơn nữa là đạt được lợi thế cạnh tranh so với các đối thủ kinh doanh.

Kho dữ liệu xuất hiện đóng vai trò như một thành phần cốt lõi của kinh doanh thông minh, giúp thúc đẩy và nâng cao hiệu suất.

Trong báo cáo này, chúng em giới thiệu về những khái niệm quan trọng trong kho dữ liệu và kinh doanh thông minh, từ đó áp dụng những kiến thức đã học được để thực hành xử lý và phân tích với bộ dữ liệu của doanh nghiệp. Qua đó đưa ra được những góc nhìn về hoạt động kinh doanh và rút ra được những kinh nghiệm quý báu trong quá trình làm việc với dữ liệu thực tế.

Trong quá trình hoàn thiện và trình bày báo cáo, chúng em xin gửi lời cảm ơn đến thầy Nguyễn Danh Tú, đã dẫn dắt và tận tình giúp đỡ, giải đáp thắc mắc và đưa ra góp ý cho đề tài của chúng em. Nội dung báo cáo được chúng em tìm hiểu, tổng hợp từ nhiều nguồn tài liệu nên không tránh khỏi tồn tại những sai sót. Vậy nên chúng em hy vọng sẽ nhận được ý kiến nhận xét, góp ý từ thầy và các bạn để báo cáo của chúng em được hoàn thiện hơn.

Chúng em xin chân thành cảm ơn!

Nhóm 9.

Mục lục

Lời mở đầu	6
Danh sách hình vẽ	9
Tự đánh giá báo cáo	11
I Tổng quan về Data Warehouse	13
1 Khái niệm kho dữ liệu	13
2 Mô hình dữ liệu đa chiều	16
3 Mô hình thiết kế kho dữ liệu	17
II Tổng quan về BI	21
1 Kinh doanh thông minh (BI)	21
2 Các bước trong quy trình kinh doanh thông minh	21
3 Lợi ích từ các ứng dụng BI	22
4 Công cụ trực quan hóa dữ liệu Power BI	23
4.1 Giới thiệu chung	23
4.2 Các chức năng của Power BI	23
III Ứng dụng Data Warehouse và BI	27
1 Khảo sát	27
1.1 Giới thiệu tập đoàn	27
1.2 Khảo sát nghiệp vụ	28
1.3 Khảo sát hệ thống	30
1.4 Xác định nhu cầu	31
1.5 Quy mô bộ dữ liệu	33
1.6 Entity Relationship Diagram (ERD) OLTP	35
2 Phân tích và thiết kế	36
2.1 Giới thiệu về bộ dữ liệu	36
2.2 Data Exploration	43
2.3 Giới thiệu về ODS	54
2.4 Kiến trúc DataWarehouse	55
2.5 Tiền xử lý ETL	56
2.6 Dimension	69
2.7 Mô hình dữ liệu OLAP	71
3 Xây dựng báo cáo trực quan	73
3.1 Báo cáo về doanh thu	73

3.2	Báo cáo về sản phẩm	75
3.3	Báo cáo về đơn hàng	76
3.4	Báo cáo về khách hàng	77
3.5	Báo cáo về cửa hàng	78
3.6	Kết luận	79
Tổng kết		81
Tham khảo		83

Danh sách hình vẽ

1	Kiến trúc kho dữ liệu một tầng	14
2	Kiến trúc kho dữ liệu hai tầng	15
3	Kiến trúc kho dữ liệu ba tầng	15
4	Mô hình dữ liệu đa chiều	16
5	Lược đồ hình sao	18
6	Lược đồ bông tuyết	19
7	Quy trình kinh doanh thông minh	21
8	Kết nối dữ liệu trong Power BI	23
9	Tiền xử lý dữ liệu trong Power BI	24
10	Các kiểu trực quan hóa dữ liệu trong Power BI	25
11	Trang web thương mại điện tử Olist	27
12	Mô hình Canvas của Olist	28
13	Quá trình đặt hàng	29
14	Quá trình giao hàng	29
15	Sơ đồ luồng dữ liệu	30
16	Sơ đồ chủ điểm phân tích	33
17	Mô hình thực thể liên kết ERD OLTP	35
18	Thống kê dữ liệu từng cột trong bảng Customers	36
19	Chi tiết dữ liệu của bảng Customers	36
20	Thống kê dữ liệu từng cột trong bảng Geolocation	37
21	Chi tiết dữ liệu của bảng Geolocation	37
22	Thống kê dữ liệu từng cột trong bảng Order_items	38
23	chi tiết dữ liệu của bảng Order_items	38
24	Thống kê dữ liệu từng cột trong bảng Order_payment	38
25	Chi tiết dữ liệu của bảng Order_payment	39
26	Thống kê dữ liệu từng cột trong bảng Orders	39
27	Chi tiết dữ liệu của bảng Orders	40
28	Thống kê dữ liệu từng cột trong bảng Products	40
29	Chi tiết dữ liệu của bảng Products	41
30	Thống kê dữ liệu từng cột trong bảng Sellers	41
31	Chi tiết dữ liệu của bảng Sellers	41
32	Thống kê dữ liệu từng cột trong bảng Reviews	42
33	Chi tiết dữ liệu của bảng Reviews	42
34	Thống kê dữ liệu từng cột trong bảng Translation	42
35	Chi tiết dữ liệu của bảng Translation	43
36	Tổng quan về bộ dữ liệu	43

37	Số lượng người bán theo thời gian	44
38	Số lượng người bán theo khu vực	44
39	Top những thành phố, bang có số lượng người bán nhiều	45
40	Số lượng khách hàng theo thời gian	45
41	Số lượng khách hàng theo khu vực	46
42	Top những thành phố, bang có số lượng khách hàng nhiều	46
43	Số lượng đơn hàng theo thời gian	47
44	Số lượng đơn hàng theo thời gian	47
45	Giá trị đơn hàng, cước phí vận chuyển theo từng tháng	48
46	Giá trị theo từng tháng ở các năm	48
47	Top 20 mặt hàng bán chạy nhất 2016-2018	49
48	Top 5 mặt hàng bán chạy nhất 2016	49
49	Top 5 mặt hàng bán chạy nhất 2017	49
50	Top 5 mặt hàng bán chạy nhất 2018	50
51	Top 5 mặt hàng được bán nhiều nhất theo từng tháng năm 2017-2018	50
52	Hình thức thanh toán	51
53	Trạng thái giao hàng và điểm đánh giá	51
54	Điểm đánh giá theo từng trạng thái giao hàng	52
55	Phân tích thống kê các dữ liệu	53
56	Kiến trúc DataWarehouse	55
57	Tổng quan hoạt động ETL	56
58	Xóa cột geolocation_city	57
59	Xóa các cột trong bộ dữ liệu còn lại	58
60	Xóa dữ liệu bị trùng trong bảng Geolocotion	58
61	Xóa dữ liệu bị trùng trong bảng Orders	59
62	Xóa dữ liệu trong bảng Payment	59
63	Chuyển đổi kiểu dữ liệu trong bảng Orders	60
64	Thêm trường dữ liệu shipping_days	61
65	Xóa dữ liệu có shipping_days < 0	61
66	Góm nhóm dữ liệu hình thức thanh toán	62
67	Gom nhóm dữ liệu trạng thái giao hàng	63
68	Gom nhóm dữ liệu loại mặt hàng	65
69	Gom nhóm dữ liệu trạng thái giao hàng	65
70	Gom nhóm dữ liệu khu vực	66
71	Quy trình chuyển đổi fact_payments	66
72	Quy trình chuyển đổi dim_geolocation	67
73	Quy trình chuyển đổi dim_customer	68
74	Quy trình chuyển đổi dim_payment_type	68
75	Các bảng dim và giá trị	70
76	Mô hình dữ liệu logic OLAP	71

77	Mô hình dữ liệu vật lý OLAP	72
78	Báo cáo về doanh thu	73
79	Báo cáo về sản phẩm	75
80	Báo cáo về đơn hàng	76
81	Báo cáo về khách hàng	77
82	Báo cáo về cửa hàng	78
83	Các cảng biển lớn ở Brazil	78

Tự đánh giá báo cáo

Báo cáo của nhóm được chia thành ba phần chính:

1. Tổng quan về Data Warehouse
2. Tổng quan về BI
3. Ứng dụng Data Warehouse và BI vào bài toán phân tích hoạt động sàn thương mại điện tử Olist

Với nội dung được phân chia phù hợp thành các chương lớn như trên, nhóm đã trình bày được những kiến thức cơ bản nhất về các chủ đề tương ứng:

- Đưa ra được khái niệm và các mô hình thiết kế kho dữ liệu
- Trình bày được BI là gì, các bước trong quá trình kinh doanh thông minh và giới thiệu công cụ trực quan hóa dữ liệu Power BI
- Ứng dụng những kiến thức trên vào dự án thực tế: phân tích hoạt động của sàn thương mại điện tử Olist. Cụ thể:
 - Tiến hành khảo sát nghiệp vụ sàn thương mại điện tử cũng như hệ thống, đặt ra được những yêu cầu báo cáo và phân tích: phân tích những chủ điểm nào, với chỉ số nào và trên những khía cạnh nào. Trình bày chi tiết bài toán nghiệp vụ, mô hình luồng nghiệp vụ, luồng dữ liệu ứng với bài toán đó.
 - Trình bày được quá trình làm sạch, xử lý dữ liệu thô để có được bộ dữ liệu phù hợp có thể đưa vào phân tích.
 - Trình bày được quá trình phân tích và thiết kế kiến trúc DataWarehouse, mô hình dữ liệu OLAP.
 - Sử dụng công cụ Power BI để xây dựng báo cáo trực quan. Từ dashboard, nhóm đã trình bày được ý nghĩa các giá trị trên dashboard và giải thích được một số kết quả báo cáo.

Tuy nhiên, những nội dung lý thuyết nhóm chúng em trình bày vẫn còn mang tính tổng quát, giới thiệu chung chứ chưa thực sự đi sâu vào chi tiết. Kết quả sau khi áp dụng mô hình phân tích cũng chưa đủ tốt để phản ánh tính chất, trạng thái hoạt động của sàn thương mại bởi nhiều lý do: bộ dữ liệu chưa đủ tốt, khía cạnh phân tích các chủ điểm đang bị giới hạn,... Trong tương lai, nhóm sẽ cố gắng tìm hiểu, nghiên cứu kỹ hơn để các dự án sau sẽ đem lại kết quả tốt nhất.

Tổng quan về Data Warehouse

1 Khái niệm kho dữ liệu

Kho dữ liệu (DW) là quy trình thu thập và quản lý dữ liệu từ nhiều nguồn khác nhau để cung cấp thông tin chi tiết có ý nghĩa về doanh nghiệp. Kho dữ liệu thường được sử dụng để kết nối và phân tích dữ liệu kinh doanh từ các nguồn không đồng nhất. Chúng lưu trữ dữ liệu lịch sử và hiện tại ở một nơi duy nhất được sử dụng để tạo báo cáo phân tích cho người lao động trong toàn doanh nghiệp. Kho dữ liệu là cốt lõi của hệ thống BI được xây dựng để phân tích và báo cáo dữ liệu.

Cơ sở dữ liệu hỗ trợ quyết định (Kho dữ liệu) được duy trì tách biệt với cơ sở dữ liệu hoạt động của tổ chức. Tuy nhiên, kho dữ liệu không phải là một sản phẩm mà là một môi trường. Nó là một cấu trúc kiến trúc của hệ thống thông tin cung cấp cho người dùng thông tin hỗ trợ quyết định hiện tại và lịch sử mà khó có thể truy cập hoặc trình bày trong kho dữ liệu hoạt động truyền thống.

Kho dữ liệu hoạt động như một kho lưu trữ trung tâm, nơi thông tin đến từ một hoặc nhiều nguồn dữ liệu. Dữ liệu chảy vào kho dữ liệu từ hệ thống giao dịch và các cơ sở dữ liệu quan hệ khác.

Dữ liệu có thể là: Có cấu trúc, Bán cấu trúc và Dữ liệu phi cấu trúc.

Dữ liệu được xử lý, chuyển đổi và nhập để người dùng có thể truy cập dữ liệu đã xử lý trong Kho dữ liệu thông qua các công cụ Business Intelligence, ứng dụng khách SQL và bảng tính. Kho dữ liệu kết hợp thông tin đến từ các nguồn khác nhau thành một cơ sở dữ liệu toàn diện.

Bằng cách hợp nhất tất cả thông tin này vào một nơi, một tổ chức có thể phân tích khách hàng của mình một cách tổng thể hơn. Điều này giúp đảm bảo rằng nó đã xem xét tất cả các thông tin có sẵn. Kho dữ liệu giúp cho việc khai thác dữ liệu có thể thực hiện được. Khai thác dữ liệu đang tìm kiếm các mẫu trong dữ liệu có thể dẫn đến doanh số và lợi nhuận cao hơn.

Ba loại Kho dữ liệu (DW) chính là:

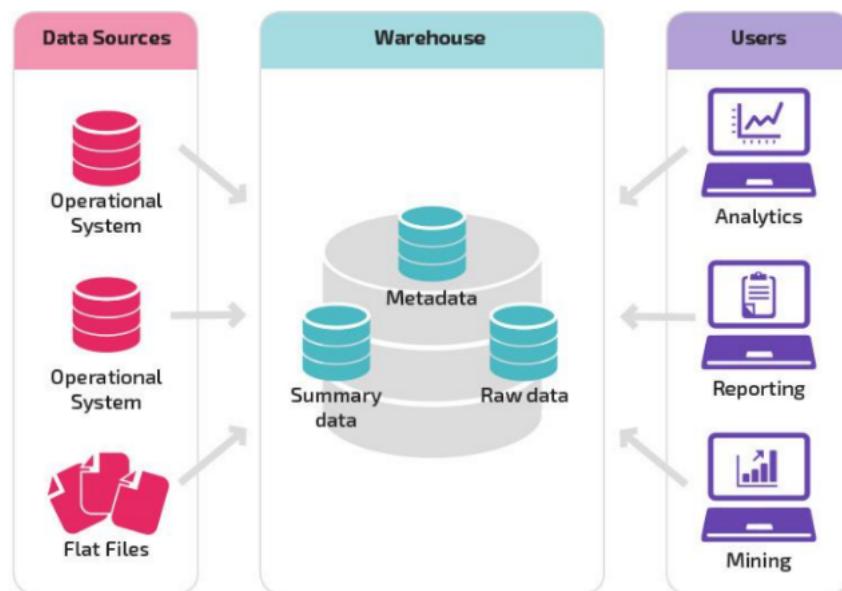
1. Kho Dữ liệu Doanh nghiệp (EDW): Kho Dữ liệu Doanh nghiệp (EDW) là một kho tập trung. Nó cung cấp dịch vụ hỗ trợ quyết định trên toàn doanh nghiệp. Nó cung cấp một cách tiếp cận thống nhất để tổ chức và biểu diễn dữ liệu. Nó cũng cung cấp khả năng phân loại dữ liệu theo chủ đề và cấp quyền truy cập theo các phân chia đó.
2. Kho dữ liệu hoạt động: Kho Dữ liệu Hoạt động, còn được gọi là ODS, không phải là nơi lưu trữ dữ liệu được yêu cầu khi cả Kho dữ liệu và hệ thống OLTP đều không hỗ trợ các nhu cầu báo cáo của tổ chức. Trong ODS, Kho dữ liệu được làm mới theo thời gian thực. Do đó, nó được ưa thích rộng rãi cho các hoạt động thường ngày như lưu trữ hồ sơ của Nhân viên.

3. Data Mart: Data mart là một tập hợp con của kho dữ liệu. Nó được thiết kế đặc biệt cho một ngành kinh doanh cụ thể, chẳng hạn như bán hàng, tài chính, bán hàng hoặc tài chính. Trong kho dữ liệu độc lập, dữ liệu có thể thu thập trực tiếp từ các nguồn.

Trích xuất, biến đổi, tải (ETL) và trích xuất, tải, biến đổi (ELT) là hai cách tiếp cận chính được sử dụng để xây dựng hệ thống kho dữ liệu.

Kho dữ liệu cho phép người dùng ở các chức vụ như nhà quản lý, người ra quyết định thực hiện phân tích với data bằng hệ thống xử lý phân tích trực tuyến (Online analytical processing – OLAP). Ngoài ra kho dữ liệu cũng được sử dụng cho báo cáo, data mining và phân tích thống kê. Như vậy, nếu như cơ sở dữ liệu được so sánh như một cái tủ sách cá nhân, nơi người ta thường xuyên tra cứu, cập nhật, ghi chú, thêm mới hoặc chuyển sách đi, thì Kho dữ liệu lại được so sánh với thư viện quốc gia, nơi các tài liệu kinh điển được đưa đến liên tục để lưu trữ và tham khảo, không ai sửa chữa hoặc chuyển chúng qua chỗ nào khác cả.

Kiến trúc một tầng

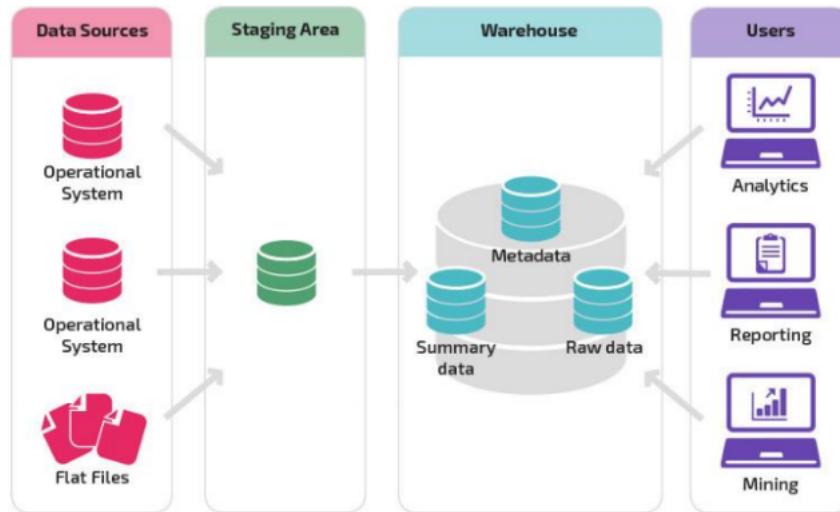


Hình 1: Kiến trúc kho dữ liệu một tầng

Kiến trúc đơn giản của hệ thống Data Warehouse gồm 3 phần:

- **Data Source:** Là nơi dữ liệu từ nhiều nguồn khác nhau được thu thập.
- **Warehouse:** Nơi lưu trữ dữ liệu đã được xử lý, gồm Metadata, Raw Data và Summary Data.
- **User:** Gồm các hệ thống phân tích, báo cáo và Mining. Đây là một kiến trúc đơn giản với phần ETL (extraction, transformation, and loading) đã bị lược bỏ, người dùng cuối truy xuất dữ liệu trực tiếp từ các hệ thống xử lý nghiệp vụ thông qua data warehouse.

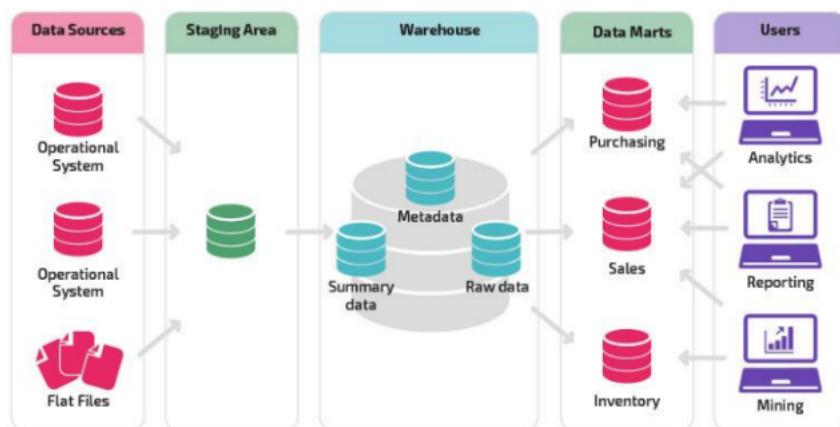
Kiến trúc hai tầng



Hình 2: Kiến trúc kho dữ liệu hai tầng

Tại kiến trúc 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.

Kiến trúc ba tầng



Hình 3: Kiến trúc kho dữ liệu ba tầng

Kiến trúc 3 tầng bổ sung thêm bước ETL, giúp phân Warehouse ra thành các chủ đề nhỏ hơn (Data mart).

Dây là kiểu kiến trúc kho dữ liệu phổ biến nhất hiện tại vì nó tạo ra một luồng dữ liệu được tổ chức tốt từ chố dữ liệu thô đến thông tin chi tiết hữu ích. Tầng dưới cùng thường bao gồm

máy chủ cơ sở dữ liệu tạo ra một lớp trùm tượng trên dữ liệu từ nhiều nguồn, như các cơ sở dữ liệu giao dịch được sử dụng cho các mục đích sử dụng front-end.

Tầng giữa bao gồm một máy chủ Online Analytical Processing (OLAP). Từ nhu cầu của người dùng, tầng giữa này sắp xếp dữ liệu phù hợp hơn cho việc phân tích về nhiều mặt.

Tầng thứ ba và là tầng cao nhất, tầng mà khách hàng sử dụng, bao gồm các công cụ và giao diện API được sử dụng để phân tích, truy vấn và báo cáo dữ liệu cấp cao.

2 Mô hình dữ liệu đa chiều

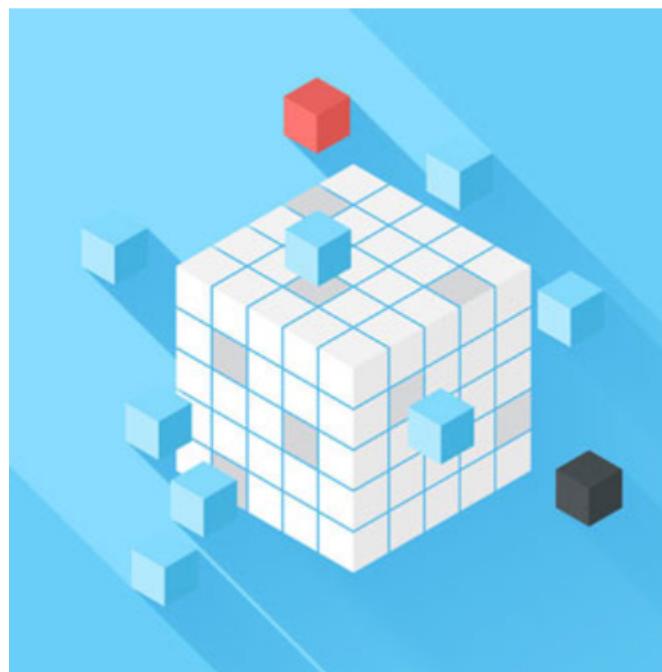
OLAP là gì?

Xử lý phân tích trực tuyến (OLAP) 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. OLAP là viết tắt của Online Analytical Processing.

Mô hình dữ liệu đa chiều



Hình 4: Mô hình dữ liệu đa chiều

Cơ sở dữ liệu được định cấu hình cho OLAP sử dụng mô hình dữ liệu đa chiều, cho phép phân tích phức tạp và truy vấn đặc biệt với tốc độ nhanh chóng. Mô hình dữ liệu đa chiều tương tự như mô hình cơ sở dữ liệu quan hệ với một biến thể là có cấu trúc đa chiều để tổ chức

dữ liệu và thể hiện mối quan hệ giữa các dữ liệu. Dữ liệu được lưu trữ dưới dạng hình khối và có thể được truy cập trong giới hạn của mỗi hình khối. Hầu hết, kho dữ liệu hỗ trợ hình khối hai hoặc ba chiều; tuy nhiên, có nhiều hơn ba kích thước dữ liệu được mô tả bởi khối lập phương được gọi là khối kết hợp.

Các chiều (Dimensions) là quan điểm hoặc thực thể liên quan đến việc tổ chức lưu giữ hồ sơ. Ví dụ: một cửa hàng có thể tạo một kho dữ liệu bán hàng để lưu giữ hồ sơ về doanh số của cửa hàng cho các chiều, mặt hàng và địa điểm. Các chiều này cho phép lưu theo dõi mọi thứ, ví dụ: doanh số bán hàng tháng của các mặt hàng và vị trí mà các mặt hàng đã được bán. Mỗi chiều có một bảng liên quan đến nó, được gọi là bảng chiều (Dim table), mô tả thêm về chiều. Ví dụ, một bảng thứ nguyên cho một mặt hàng có thể chứa các thuộc tính item_name, brand và type.

Mô hình dữ liệu đa chiều được tổ chức xung quanh chủ điểm phân tích (Facts), ví dụ: bán hàng. Chủ đề này được thể hiện bằng một bảng fact .Bảng Fact chứa dữ liệu mà chúng ta muốn thêm vào reports, tổng hợp trên các giá trị trong các bảng dimension. Một bảng fact chỉ có các cột lưu giá trị và các cột khóa ngoại tham chiếu đến bảng dimensions. Kết hợp tất cả các khóa ngoại và khoá chính trong bảng fact. Ví dụ, một bảng fact có thể lưu trữ một số lượng các hợp đồng và số lượng các nhân viên bán hàng từ các danh sách hợp đồng.

Hệ thống OLAP chủ yếu được phân loại thành ba:

- MOLAP (OLAP đa chiều)
- ROLAP (OLAP quan hệ): hoạt động với cơ sở dữ liệu quan hệ
- HOLAP (Hybrid OLAP): cơ sở dữ liệu phân chia dữ liệu giữa lưu trữ quan hệ và lưu trữ chuyên dụng.

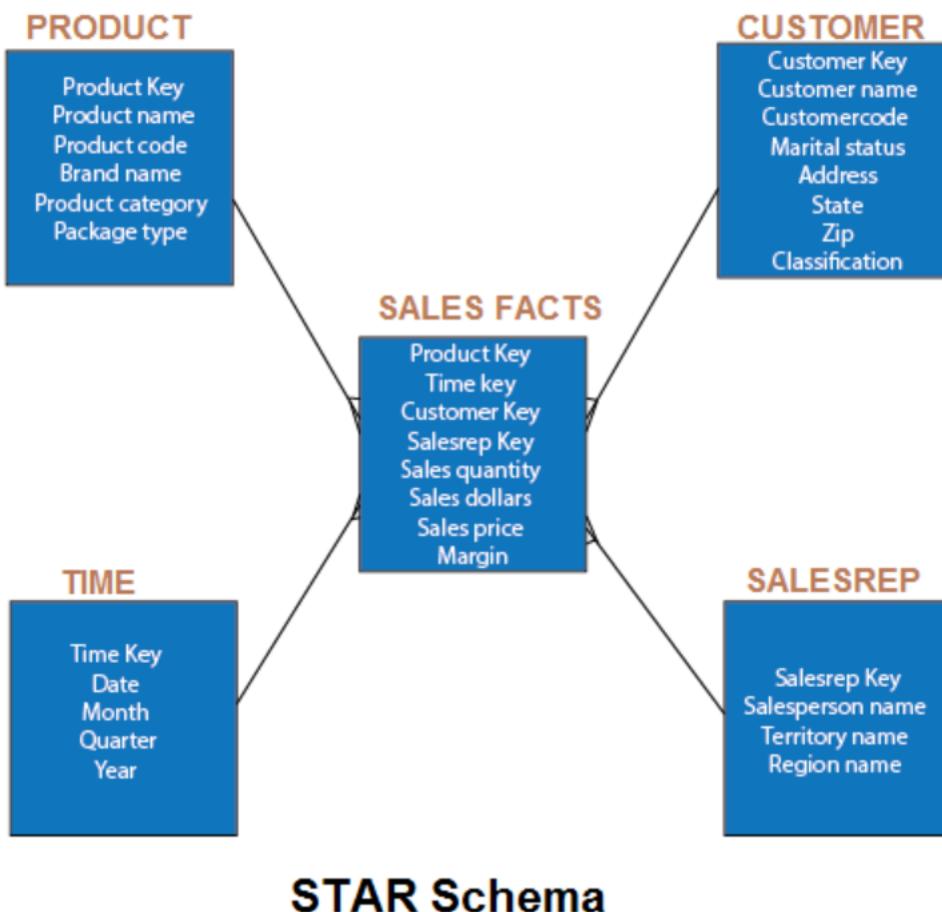
3 Mô hình thiết kế kho dữ liệu

Dựa trên những cách thức lưu trữ dữ liệu khác nhau, mô hình dữ liệu đa chiều thường được lưu trữ theo 3 hướng dưới đây:

- Mô hình OLAP kiểu quan hệ (Relational OLAP – ROLAP). ROLAP lưu trữ dữ liệu trong các cột và hàng (còn được gọi là bảng quan hệ) và truy xuất thông tin theo yêu cầu thông qua các truy vấn do người dùng gửi. Cơ sở dữ liệu ROLAP có thể được truy cập thông qua các truy vấn SQL phức tạp để tính toán thông tin. ROLAP có thể xử lý khối lượng dữ liệu lớn, nhưng dữ liệu càng lớn thì thời gian xử lý càng chậm.
- Mô hình OLAP đa chiều (Multidimensional OLAP – MOLAP). MOLAP sử dụng một khối đa chiều truy cập dữ liệu được lưu trữ thông qua nhiều cách kết hợp khác nhau. Dữ liệu được tính toán trước, tóm tắt trước và được lưu trữ (một điểm khác biệt so với ROLAP, nơi các truy vấn được phục vụ theo yêu cầu).

- Mô hình OLAP lai (Hybrid OLAP – HOLAP). Chế độ lưu trữ HOLAP kết nối các thuộc tính của cả MOLAP và ROLAP. Vì HOLAP liên quan đến việc lưu trữ một phần dữ liệu của bạn trong cửa hàng ROLAP và một phần khác trong cửa hàng MOLAP, các nhà phát triển nhận được lợi ích của cả hai. Trong hệ thống OLAP kiểu quan hệ, dữ liệu đa chiều được lưu trữ dưới dạng bảng qua nhiều hệ, tổ chức theo cấu trúc chủ yếu theo lược đồ hình sao, lược đồ hình bông tuyết, ngoài ra còn có lược đồ ánh sao và lược đồ chòm sao.

Lược đồ ngôi sao (Star Schema)



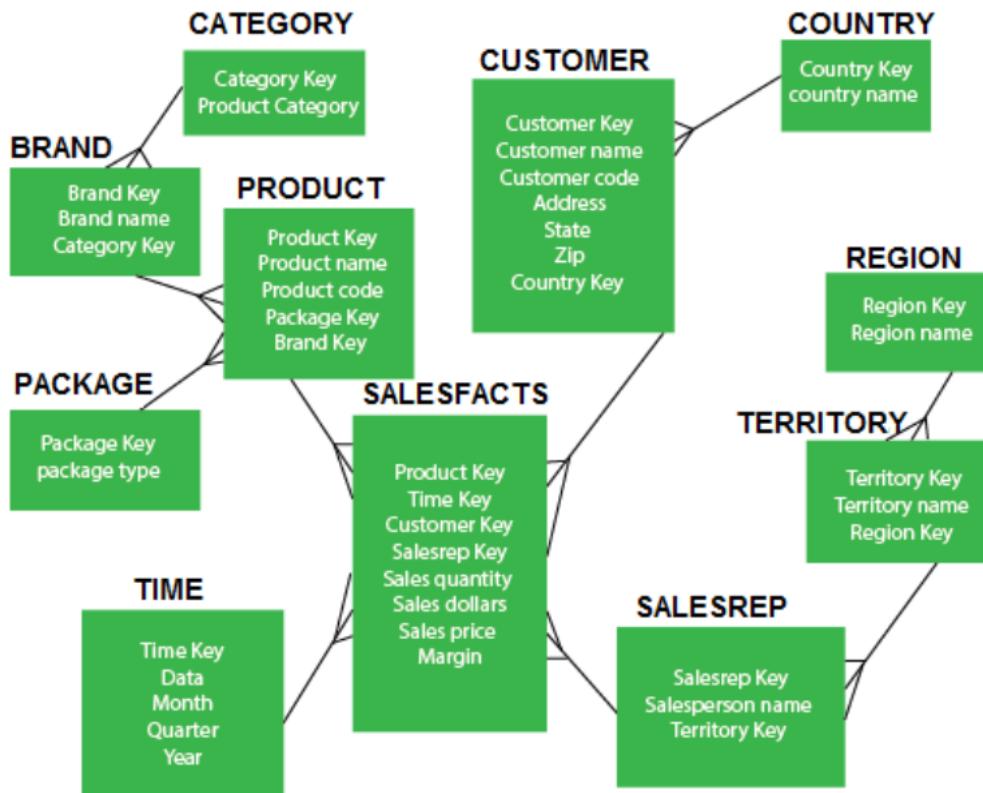
Hình 5: Lược đồ hình sao

Sơ đồ ngôi sao là mô hình đơn giản nhất được sử dụng trong DWH. Bởi vì bảng fact là trung tâm của mô hình với các bảng dimension xung quanh nó, nó nhìn giống như một ngôi sao. Điều này rất rõ ràng khi bảng fact được bao quanh bởi 5 bảng dimension. Một biến thể của sơ đồ ngôi sao là sơ đồ con rết (centipede schema), nơi mà bảng fact được bao quanh bởi số lượng lớn các bảng dimension nhỏ.

Mô hình ngôi sao được sử dụng rộng rãi trong data marts. Chúng ta có thể kết hợp chung trong mô hình top-down. Chúng ta sẽ phân tích mô hình 2 ngôi sao và kết hợp chúng để tạo ra mô hình đơn giản.

Lược đồ bông tuyết (Snowflake schema)

Lược đồ bông tuyết là dạng mở rộng của lược đồ hình sao bằng cách bổ sung thêm các Dim. Bảng Fact giống như lược đồ hình sao, các bảng Dim được chuẩn hoá. Các chiều có cấu trúc rõ ràng. Bảng Dim được chia thành hai chiều là chiều chính hay chiều phụ.



Snowflake Schema

Hình 6: Lược đồ bông tuyết

Ưu điểm: Số chiều được phân cấp thể hiện dạng chuẩn của bảng Dim.

Nhược điểm: Cấu trúc phi dạng chuẩn của lược đồ hình sao phù hợp hơn cho việc duyệt các chiều.

Tổng quan về BI

Trong chương này, nhóm sẽ giới thiệu tổng quan về Kinh doanh thông minh và các công việc có liên quan.

1 Kinh doanh thông minh (BI)

Thị trường kinh doanh đang ngày càng phát triển, mở rộng một cách nhanh chóng, thị trường phát triển đòi hỏi các công ty phải liên tục cập nhật, trau dồi và đổi mới liên tục để phù hợp với thị trường. Với sự tiến bộ của công nghệ, các công ty, các doanh nghiệp có thể dễ dàng tiếp cận đến khách hàng một cách thông minh và nhanh chóng.

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 Các bước trong quy trình kinh doanh thông minh



Hình 7: Quy trình kinh doanh thông minh

-
- Nguồn dữ liệu (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...
 - Kho dữ liệu (Data Warehouse) là cơ sở dữ liệu được thiết kế theo mô hình Online Analytical Processing (OLAP), dữ liệu trong data warehouse chỉ có thể đọc, không được ghi hay xóa mà chỉ được update bởi gói ETL chuyển đổi dữ liệu từ Data Sources vào Data Warehouse.
 - Khám phá dữ liệu (Data Exploration) là bước đầu tiên trong phân tích dữ liệu, trong đó người dùng khám phá một tập dữ liệu lớn theo cách không có cấu trúc để khám phá các mẫu, đặc điểm và điểm quan tâm ban đầu. Khám phá dữ liệu tạo ra các truy vấn, báo cáo, biểu đồ phân tích, thống kê từ mô hình dữ liệu OLAP ở Kho dữ liệu.
 - Khai thác dữ liệu (Data mining) là quá trình phân tích khối lượng lớn dữ liệu để khám phá thông tin kinh doanh giúp các công ty giải quyết vấn đề, giảm thiểu rủi ro và nắm bắt cơ hội mới.

3 Lợi ích từ các ứng dụng BI

Những lợi ích chính mà doanh nghiệp có thể nhận được từ các ứng dụng BI:

- Tăng tốc và cải thiện việc ra quyết định
- Tối ưu hóa quy trình kinh doanh nội bộ
- Phát hiện các vấn đề kinh doanh cần được giải quyết
- Xác định các xu hướng kinh doanh và thị trường mới nổi
- Phát triển các chiến lược kinh doanh mạnh mẽ hơn
- Thúc đẩy doanh số bán hàng cao hơn và doanh thu mới
- Đạt được lợi thế cạnh tranh so với các công ty đối thủ Một số công cụ thông dụng hiện nay
 - QlikView
 - Power BI
 - Spago
 - Pentaho

- SAP BO
- Oracle BI
- IBM Cognos

Trong báo cáo này, chúng em sử dụng công cụ chính là Power BI Desktop để phân tích dữ liệu bán hàng tại nạn của Australia.

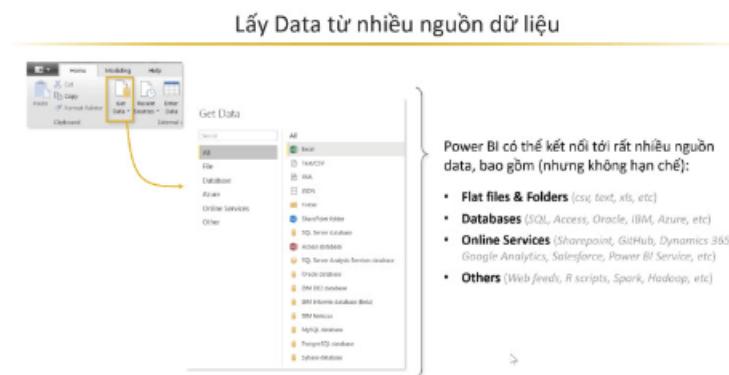
4 Công cụ trực quan hóa dữ liệu Power BI

4.1 Giới thiệu chung

Power BI được ra đời vào năm 2011, được phát triển bởi Microsoft, sau đó nó được đưa vào sử dụng chính thức vào năm 2015. Power BI tập hợp rất nhiều các dịch vụ về phần mềm, các ứng dụng, các trình kết nối hoạt động song song cùng nhau để biến đổi các nguồn dữ liệu từ nhiều nguồn khác nhau thành các thông tin chi tiết liền mạch và trực quan. Power BI được phát triển, sử dụng trên nền tảng Desktop, Website Service và Mobile App, nó hoàn toàn thân thiện và dễ dàng thích ứng với mọi người dùng mặc dù mỗi người đều có những nhu cầu khác nhau.

4.2 Các chức năng của Power BI

Kết nối dữ liệu từ nhiều nguồn



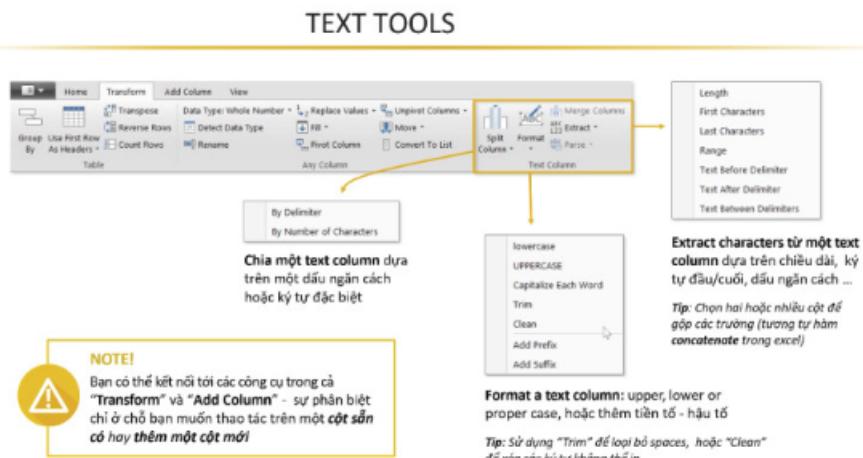
Hình 8: Kết nối dữ liệu trong Power BI

Chúng ta có thể truy cập dữ liệu từ nhiều nguồn khác nhau dựa trên nền tảng Power BI, bao gồm các nguồn như sau:

- File: các dạng Excel, Text/CSV, XML, JSON, Folder, PDF, SharePoint folder.
- Database: SQL Server database, Access database, Oracle database, IBM Db2 database, MySQL, ...
- Power Platform: Power BI datasets, Power BI dataflows, Common Data Service, ...

- Power Platform: Power BI datasets, Power BI dataflows, Common Data Service, ...
- Power Platform: Power BI datasets, Power BI dataflows, Common Data Service, ...
- Azure
- Online Services: SharePoint Online List, Microsoft Exchange Online, Dynamics 365...

Tiền xử lý dữ liệu



Hình 9: Tiền xử lý dữ liệu trong Power BI

Hầu hết trong doanh nghiệp, dữ liệu thu thập được đều đã trải qua quá trình tiền xử lý dữ liệu để có thể sẵn sàng sử dụng tạo ra các báo cáo. Trong quá trình tiền xử lý dữ liệu, các dữ liệu được trích xuất từ một nguồn dữ liệu, sau đó được chuyển đổi, xác thực, chuẩn hóa, sửa chữa, kiểm tra và cuối cùng được tải vào kho dữ liệu.

Quy trình tiền xử lý dữ liệu được thực hiện bởi các ứng dụng như SQL Server Integration Services (SSIS) hoặc các công cụ của bên thứ ba khác. Tuy nhiên, trong một số doanh nghiệp, công việc tiền xử lý dữ liệu được thực hiện ngay trong Excel, được gọi là chuyển đổi dữ liệu. Tuy nhiên, quy trình ETL trong Excel là một quy trình thủ công, mất nhiều thời gian và khó có thể tự động hóa. Vì thế Microsoft đã tạo ra công cụ để có thể làm cho quá trình này trở nên nhanh và dễ dàng hơn nhiều đó là Power Query, Power BI Desktop. Hai công cụ này cung cấp cho người dùng khả năng tự động hóa quá trình nhập, chuyển đổi và tải dữ liệu vào các bảng nội bộ trong Power BI, sau đó có thể được sử dụng làm nguồn cho các báo cáo hay dashboard của Power BI. Vì Power Query duy trì bản ghi từng bước của mọi hành động được thực hiện để nhập, chuyển đổi và tải dữ liệu, các bước này sẽ được lặp lại khi có thêm dữ liệu được thêm vào.

Mô hình hóa dữ liệu (Data modeling)

Mô hình hóa dữ liệu là quá trình tạo ra một biểu diễn trực quan của toàn bộ hệ thống thông tin hoặc các bộ phận của nó để giao tiếp các kết nối giữa các điểm và cấu trúc dữ liệu. Mục đích là minh họa các loại dữ liệu được sử dụng và lưu trữ trong hệ thống, mối quan hệ giữa các loại dữ liệu này, cách dữ liệu có thể được nhóm và tổ chức cũng như các định dạng và thuộc

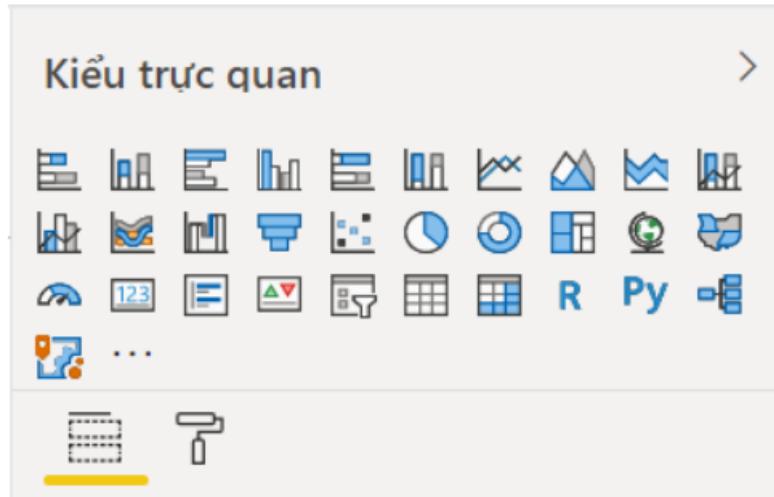
tính của nó. Dữ liệu có thể được mô hình hóa ở nhiều mức độ trừu tượng khác nhau. Quá trình bắt đầu bằng cách thu thập thông tin về các yêu cầu kinh doanh từ các bên liên quan và người dùng cuối. Các quy tắc nghiệp vụ này sau đó được chuyển thành cấu trúc dữ liệu để hình thành một thiết kế cơ sở dữ liệu cụ thể. Mô hình dữ liệu có thể được so sánh với lộ trình, bản thiết kế của kiến trúc sư hoặc bất kỳ sơ đồ chính thức nào giúp hiểu sâu hơn về những gì đang được thiết kế.

Mô hình hóa dữ liệu sử dụng các lược đồ chuẩn hóa và các kỹ thuật chính thức. Điều này cung cấp một cách chung, nhất quán và có thể dự đoán được để xác định và quản lý tài nguyên dữ liệu trong một tổ chức hoặc thậm chí xa hơn.

Quy trình mô hình hóa dữ liệu:

- Xác định các thực thể
- Xác định các thuộc tính chính của từng thực thể
- Xác định mối quan hệ giữa các thực thể
- Ánh xạ các thuộc tính cho các thực thể hoàn toàn
- Gán các khóa khi cần thiết và quyết định mức độ chuẩn hóa cân bằng giữa nhu cầu giảm dư thừa với các yêu cầu về hiệu suất.
- Hoàn thiện và xác thực mô hình dữ liệu

Trực quan hóa dữ liệu



Hình 10: Các kiểu trực quan hóa dữ liệu trong Power BI

Trực quan hóa dữ liệu là biểu diễn đồ họa của thông tin và dữ liệu. Bằng cách sử dụng các yếu tố trực quan như biểu đồ, đồ thị và bản đồ, các công cụ trực quan hóa dữ liệu cung cấp một cách dễ tiếp cận để xem và hiểu các xu hướng, ngoại lệ và mẫu trong dữ liệu. Trong thê

giới của Dữ liệu lớn, các công cụ và công nghệ trực quan hóa dữ liệu là rất cần thiết để phân tích một lượng lớn thông tin và đưa ra các quyết định dựa trên dữ liệu.

Trực quan hóa dữ liệu giúp bạn biến tất cả dữ liệu chi tiết đó thành thông tin kinh doanh dễ hiểu, hấp dẫn về mặt hình ảnh — và hữu ích để phân tích. Bằng cách khai thác các nguồn dữ liệu bên ngoài, các công cụ trực quan hóa dữ liệu ngày nay không chỉ cho phép bạn xem KPI của mình rõ ràng hơn, chúng hợp nhất dữ liệu và áp dụng phân tích dựa trên AI để tiết lộ mối quan hệ giữa KPI của bạn, thị trường và thế giới.

Trực quan hóa dữ liệu làm cho dữ liệu trở nên sống động, khiến bạn trở thành người kể chuyện bậc thầy về những thông tin chi tiết ẩn trong các con số của bạn. Thông qua trang tổng quan trực tiếp, báo cáo tương tác, biểu đồ, đồ thị và các biểu thị trực quan khác, trực quan hóa dữ liệu giúp người dùng phát triển thông tin chi tiết mạnh mẽ về doanh nghiệp một cách nhanh chóng và hiệu quả.

Ứng dụng Data Warehouse và BI

Phần sau đây trình bày ứng dụng Data Warehouse và BI vào bài toán phân tích hoạt động sàn thương mại điện tử Olist.

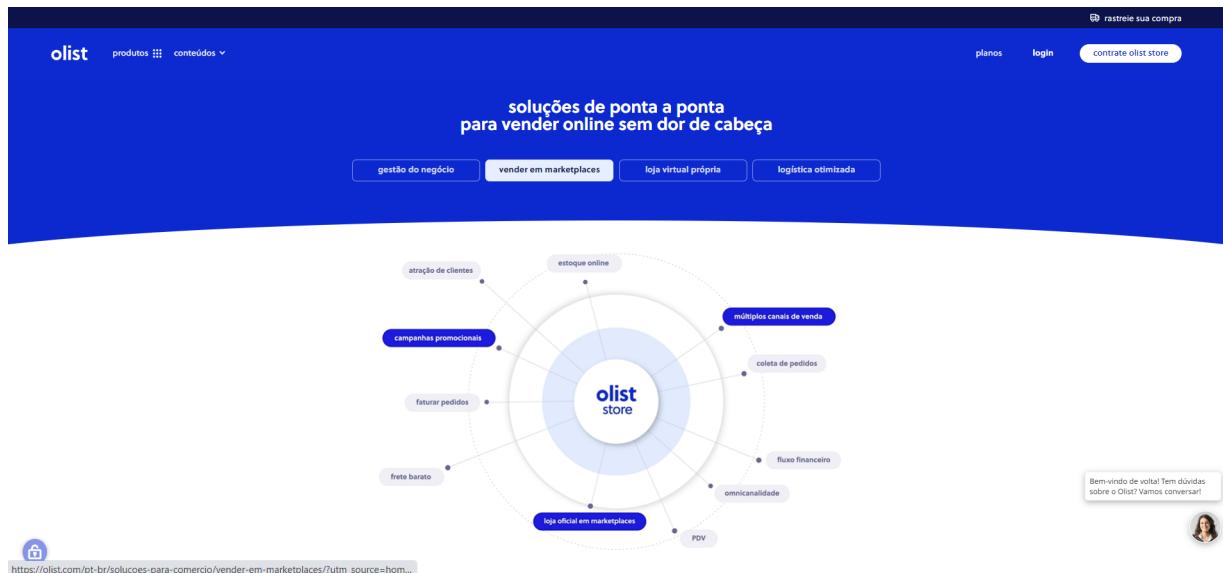
1 Khảo sát

1.1 Giới thiệu tập đoàn



Olist là một tập đoàn công nghệ được thành lập vào năm 2015 bởi Tiago Dalvi, trụ sở đặt tại thành phố Curitiba, bang Paraná, Brazil. Tập đoàn này điều hành trang web thương mại điện tử **Olist** - là nơi kết nối nhiều cửa hàng và sản phẩm của họ với thị trường tiêu dùng Brazil. Olist tạo ra môi trường hoạt động, kinh doanh mua bán cho nhiều đối tượng, từ chủ cửa hàng bán lẻ đến các doanh nghiệp lớn đều có thể tham gia vào hệ thống thương mại điện tử này.

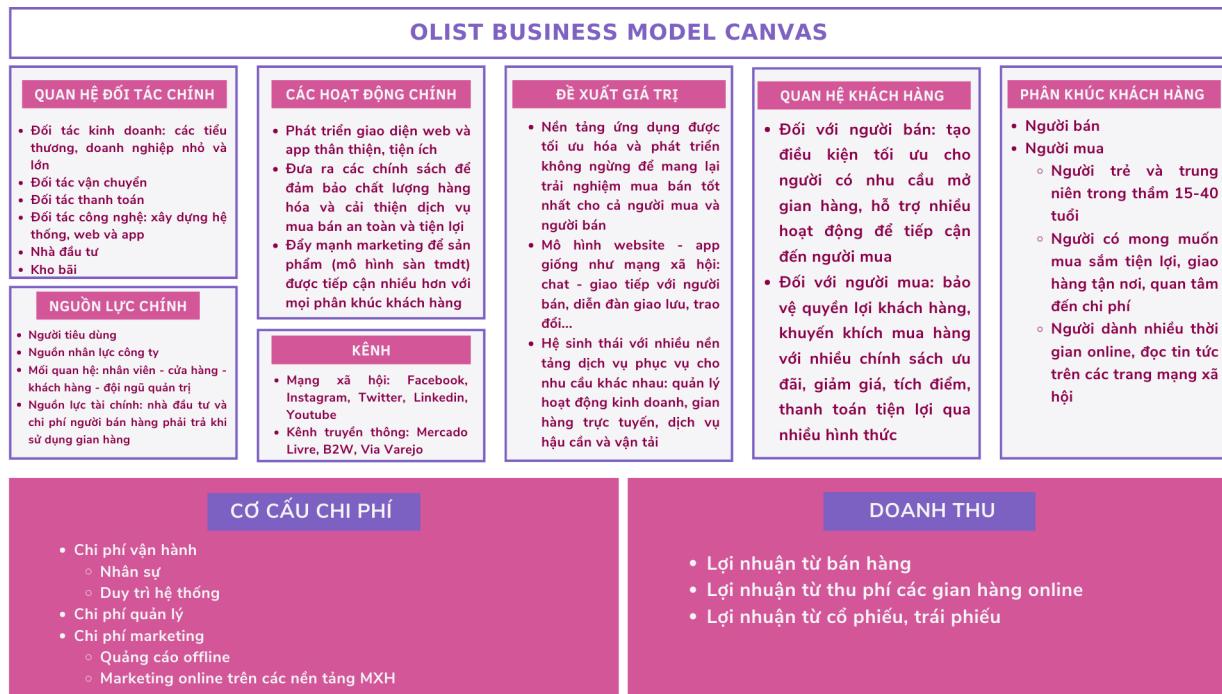
Đến nay, sàn thương mại điện tử Olist đã thu hút được hơn 200.000 người dùng đến từ hơn 180 quốc gia trên thế giới và hơn 45.000 cửa hàng, nhà bán lẻ đăng ký kinh doanh trên hệ thống. Olist được đầu tư phát triển trên nhiều kênh truyền thông cũng như mạng xã hội, tăng khả năng tiếp cận thêm nhiều đối tượng khách hàng thông qua nhiều phương thức khác nhau.



Hình 11: Trang web thương mại điện tử Olist

Nhóm khách hàng chính của hệ thống Olist chính là: **Người mua** và **Người bán - các cửa hàng, doanh nghiệp**. Dựa trên khảo sát mô hình hệ thống cũng như mô tả hoạt động

của Olist, nhóm chúng em đã trình bày mô hình kinh doanh của tập đoàn theo 9 yếu tố: *Phân khúc khách hàng, Đề xuất giá trị, Các kênh truyền thông, Quan hệ khách hàng, Dòng doanh thu, Nguồn lực chính, Hoạt động chính, Đối tác chính, Cơ cấu chi phí* như sau:



Hình 12: Mô hình Canvas của Olist

1.2 Khảo sát nghiệp vụ

1.2.1 Bài toán xây dựng hệ thống

Bài toán xây dựng hệ thống thương mại điện tử:

Khách hàng truy cập vào website của hệ thống thương mại điện tử. Website sẽ hiển thị những sản phẩm đến từ nhiều cửa hàng khác nhau và khách hàng có thể tìm kiếm, lựa chọn sản phẩm mình muốn mua. Nếu người đó muốn thanh toán ngay sản phẩm đã chọn, hệ thống sẽ tiếp nhận yêu cầu mua hàng, tính giá trị đơn hàng và gửi thông tin thanh toán đến khách hàng. Trong trường hợp không mua ngay sản phẩm đã chọn, khách hàng có thể thêm sản phẩm vào giỏ hàng của mình và lựa chọn thanh toán giỏ hàng sau đó.

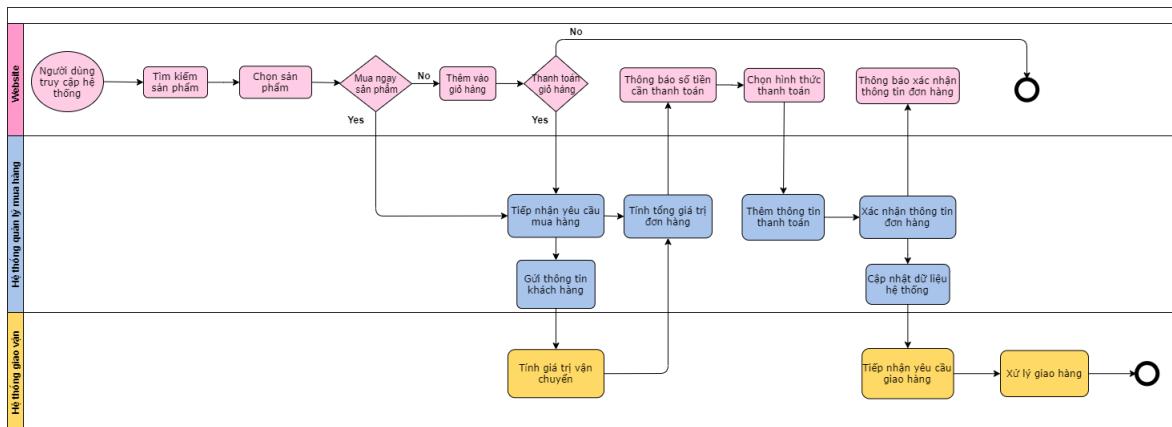
Khi website hiển thị giá trị đơn hàng và thông tin thanh toán, khách hàng sẽ lựa chọn hình thức thanh toán cho đơn hàng. Hệ thống sau khi xác nhận thanh toán sẽ gửi thông báo xác nhận thanh toán cho khách hàng qua website, đồng thời gửi thông tin địa chỉ của khách hàng cho bộ phận quản lý giao vận. Bộ phận này sẽ tiến hành tính toán và gửi thông báo cho khách hàng về thời gian giao hàng dự kiến cũng như chi phí giao hàng. Như vậy, đơn hàng chính thức được xác nhận, tiến vào quá trình xử lý và tiến hành giao hàng.

Trong quá trình vận chuyển đơn hàng đến với khách hàng, bộ phận quản lý giao vận phải liên tục kiểm tra và cập nhật tình trạng đơn hàng cho đến khi khách hàng xác nhận đã nhận

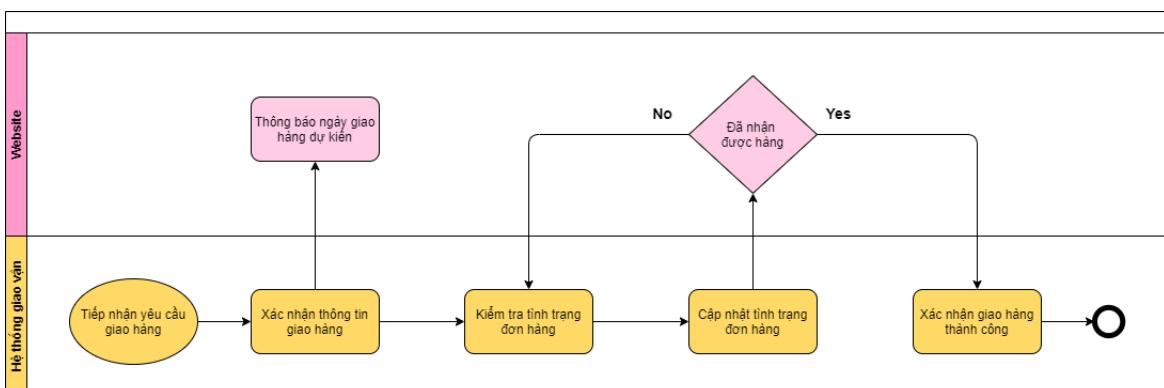
được hàng - đơn hàng được giao thành công. Nếu có vấn đề về giao nhận, hệ thống phải kiểm tra tình trạng đơn hàng và cập nhật thông tin cho khách hàng kịp thời.

1.2.2 Luồng nghiệp vụ

Dựa vào bài toán xây dựng hệ thống thương mại điện tử ở phần trước, nhóm chúng em đã đưa ra mô hình luồng nghiệp vụ của hai quá trình chính: *Quá trình đặt hàng* và *Quá trình giao hàng* như sau:



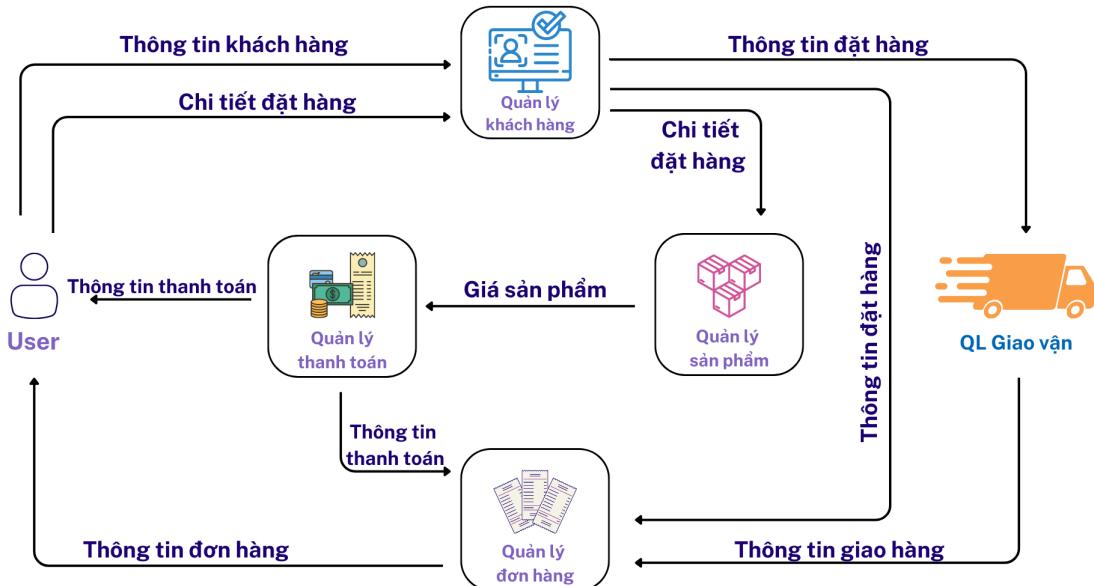
Hình 13: Quá trình đặt hàng



Hình 14: Quá trình giao hàng

1.2.3 Luồng dữ liệu

Hệ thống quản lý mua hàng được thành 4 nhóm hệ thống chính: *Hệ thống quản lý sản phẩm*, *Hệ thống quản lý khách hàng*, *Hệ thống quản lý thanh toán*, *Hệ thống quản lý đơn hàng*. Dựa trên tương tác của khách hàng với hệ thống, sơ đồ luồng dữ liệu được mô tả như hình dưới đây:



Hình 15: Sơ đồ luồng dữ liệu

1.3 Khảo sát hệ thống

1.3.1 Hệ thống hiện tại

Hệ thống xử lý dữ liệu hiện có đáp ứng được các chức năng lưu trữ và quản lý cơ bản:

- Thu thập, xử lý các thông tin: khách hàng, đơn hàng, người bán hàng,...
- Làm các báo cáo đơn giản

Tuy nhiên vẫn còn rất nhiều vấn đề tồn đọng trong hệ thống xử lý này:

- Hệ thống chưa đáp ứng được các yêu cầu về phân tích dữ liệu, đặc biệt là phân tích dữ liệu nhiều chiều, không thể tự động xuất ra các báo cáo đang sử dụng mà cần thống kê lại một cách thủ công.
- Hệ thống hiện tại lưu trữ dữ liệu một cách rời rạc, không thống nhất, khoa học khiến việc truy xuất dữ liệu gặp nhiều khó khăn.
- Các báo cáo dựa trên hệ thống hiện tại đang mang tính sơ lược, còn thiếu chính xác, chưa mang lại hiệu quả phân tích cao.

1.3.2 Yêu cầu hệ thống mới

Từ những hạn chế của hệ thống hiện tại, chúng em đặt ra những yêu cầu cho hệ thống mới để cải tiến chất lượng hệ thống quản trị, xử lý dữ liệu:

- Dễ dàng trích xuất dữ liệu
- Kiến trúc linh hoạt, dễ dàng mở rộng quy mô dữ liệu

- Cung cấp các công cụ phân tích, đưa ra các dashboard, báo cáo trực quan dữ liệu
- Đáp ứng dữ liệu lớn, dễ dàng chuẩn hóa thống nhất dữ liệu theo yêu cầu phân tích
- Hệ thống mới sẽ có chức năng phân tích dữ liệu dựa trên nhiều chiều
- Cập nhật dữ liệu theo định kỳ

1.4 Xác định nhu cầu

1.4.1 Liệt kê báo cáo

Trên cương vị điều hành trang web thương mại điện tử, nhóm chúng em sẽ liệt kê những báo cáo cần phân tích sau đây:

1. Báo cáo doanh thu theo địa điểm (bang, khu vực), theo thời gian (tháng, năm), theo loại mặt hàng, ngành hàng, theo phương thức thanh toán.
2. Báo cáo chất lượng giao vận: Thời gian giao hàng, tỉ lệ giao hàng đúng thời gian dự kiến, chi phí giao hàng, khu vực.
3. Báo cáo số lượng sản phẩm, điểm đánh giá đã bán được theo từng loại mặt hàng, ngành hàng, thời gian, theo địa điểm.
4. Báo cáo số lượng khách hàng, ARPU (doanh thu trung bình trên một khách hàng) theo địa điểm, thời gian, phân khúc khách hàng.
5. Báo cáo số lượng đơn hàng, ARPO (doanh thu trung bình trên một đơn hàng) theo địa điểm, thời gian, ngành hàng, hình thức thanh toán.
6. Báo cáo số lượng cửa hàng, điểm đánh giá cửa hàng, top doanh thu theo địa điểm, thời gian, theo loại mặt hàng, ngành hàng.

1.4.2 Chủ điểm phân tích

Từ danh sách báo cáo đã được liệt kê, nhóm xác định được các chủ điểm cùng với chỉ số, khía cạnh phân tích trong bảng sau:

Chủ điểm	Chỉ số	Khía cạnh
Hoạt động bán hàng	Doanh thu	<ul style="list-style-type: none"> - Địa điểm (bang, khu vực) - Thời gian (năm, tháng) - Ngành hàng - Phương thức thanh toán

Đơn hàng	Số lượng đơn hàng	<ul style="list-style-type: none"> - Địa điểm (bang, khu vực) - Thời gian (năm, tháng) - Ngành hàng - Hình thức thanh toán
Chất lượng giao vận	<ul style="list-style-type: none"> - Thời gian giao hàng (trung bình) - Chi phí giao hàng - Tỷ lệ giao hàng đúng hẹn 	Địa điểm (bang, khu vực)
Khách hàng	<ul style="list-style-type: none"> - Số lượng khách hàng - ARPU 	<ul style="list-style-type: none"> - Địa điểm (bang, khu vực) - Thời gian (năm, tháng) - Phân khúc khách hàng
Sản phẩm	<ul style="list-style-type: none"> - Số lượng sản phẩm - Điểm đánh giá 	<ul style="list-style-type: none"> - Địa điểm (bang, khu vực) - Thời gian (năm, tháng) - Ngành hàng, loại mặt hàng
Cửa hàng	<ul style="list-style-type: none"> - Số lượng cửa hàng - Số sản phẩm đã bán - Điểm đánh giá - Doanh thu 	<ul style="list-style-type: none"> - Địa điểm (bang, khu vực) - Thời gian (năm, tháng) - Ngành hàng, loại mặt hàng

1.4.3 Mindmap

Để tăng tính trực quan hóa, nhóm xin trình bày mindmap biểu diễn các chủ điểm phân tích như hình dưới:



Hình 16: Sơ đồ chủ điểm phân tích

1.5 Quy mô bộ dữ liệu

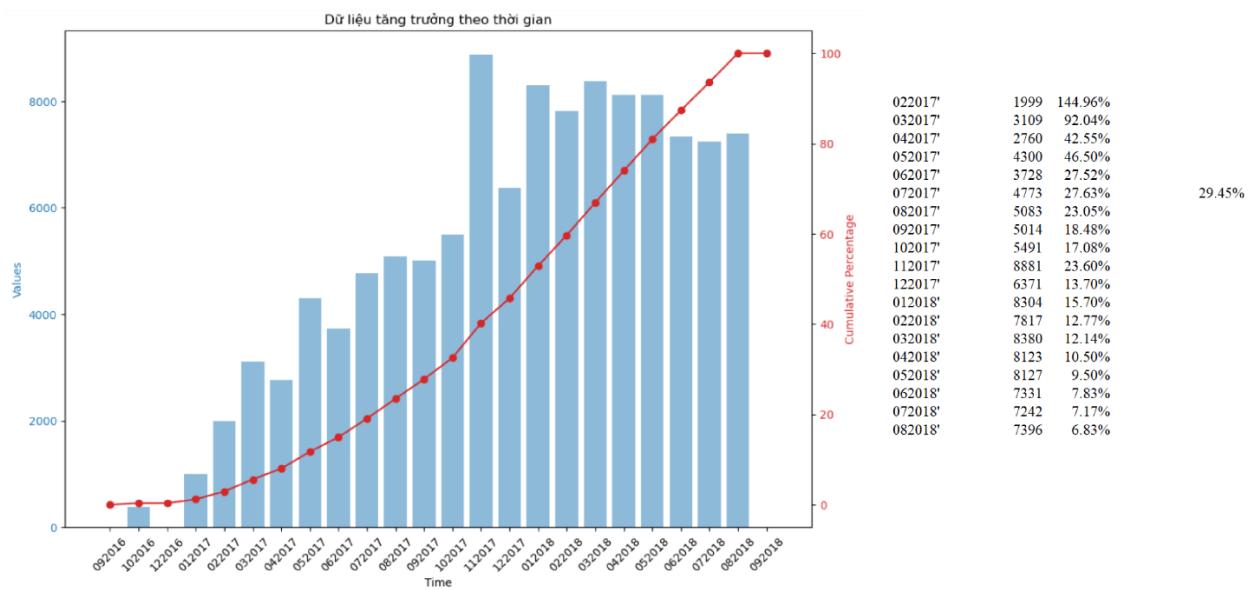
- Tên bộ dữ liệu: Brazilian E-Commerce Public Dataset by Olist
- Nguồn: Kaggle
- Kích thước: 126.19 MB
- 9 files, 52 columns
- Dịnh dạng file: .csv
- Thời gian: 2016-2018
- Nội dung: Các đơn đặt hàng được thực hiện tại Cửa hàng Olist. Bộ dữ liệu có thông tin khoảng 100.000 đơn đặt hàng từ năm 2016 đến năm 2018 được thực hiện tại nhiều thị trường ở Brazil.

Data Explorer

[Version 2 \(126.19 MB\)](#)

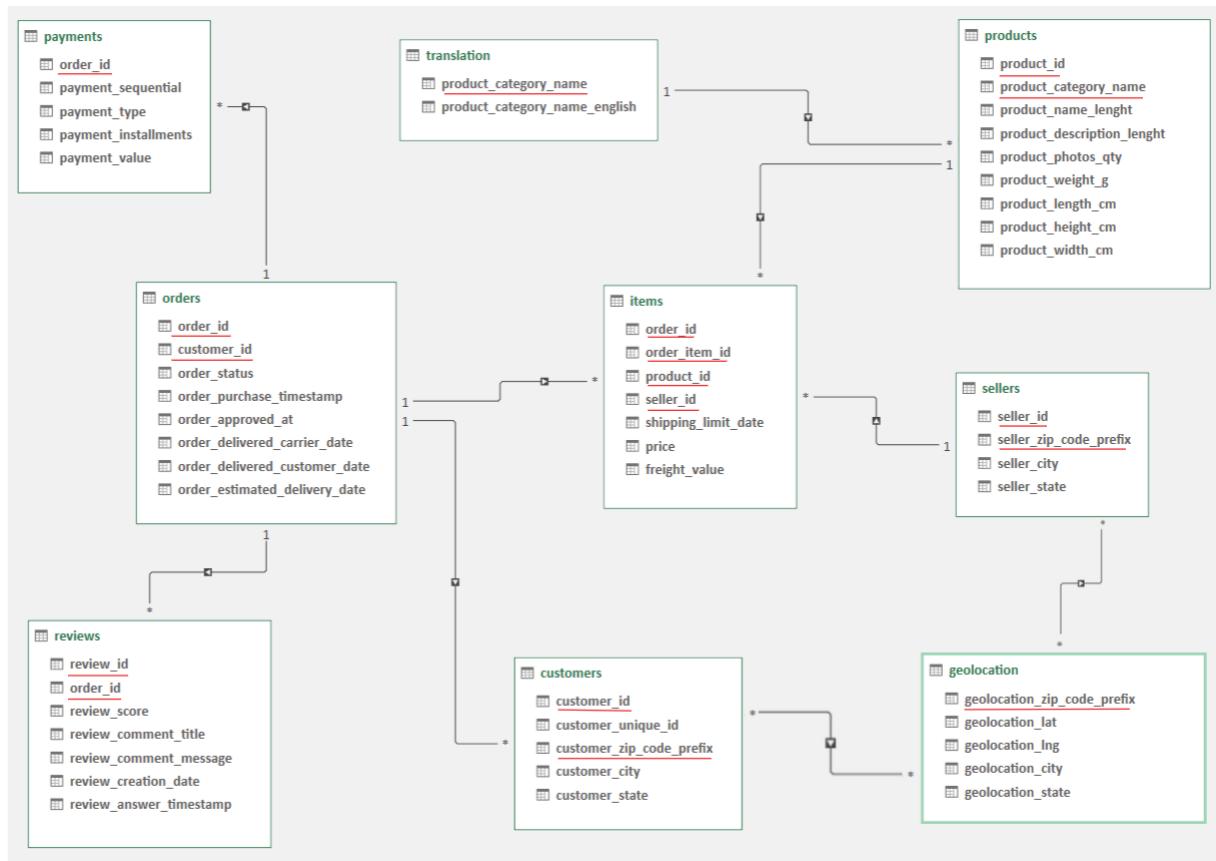
- olist_customers_dataset.csv
- olist_geolocation_dataset.csv
- olist_order_items_dataset.csv
- olist_order_payments_dataset.csv
- olist_order_reviews_dataset.csv
- olist_orders_dataset.csv
- olist_products_dataset.csv
- olist_sellers_dataset.csv
- product_category_name_trans

Tốc độ cập nhật dữ liệu



Để đảm bảo tính khách quan của việc cập nhật dữ liệu. Chúng ta chỉ xét đến những dữ liệu ở giữa, bỏ đi 1 vài giá trị ở đầu và cuối (vì Olist đi vào hoạt động từ 2016 và dữ liệu tháng 9 năm 2018 chưa được cập nhật đầy đủ). Đường màu đỏ thể hiện tốc độ tăng trưởng của bộ dữ liệu, các số liệu về thông tin dữ liệu theo từng tháng được trình bày ở trên. Từ đây ta có thể thấy **tốc độ cập nhật dữ liệu trung bình ~30%**.

1.6 Entity Relationship Diagram (ERD) OLTP

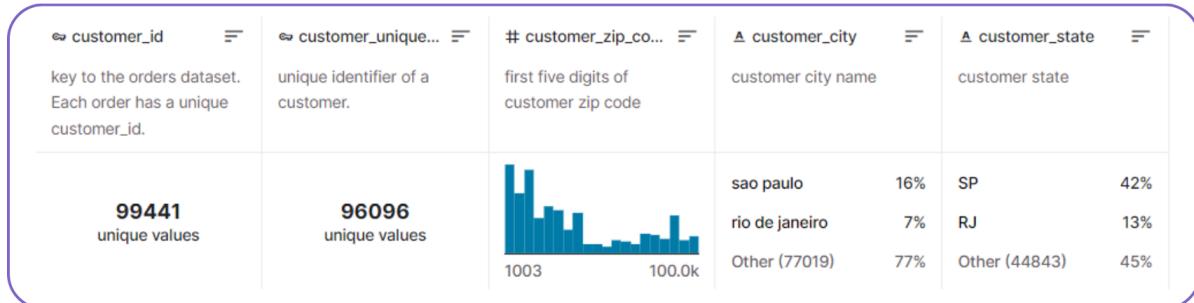


Hình 17: Mô hình thực thể liên kết ERD OLTP

2 Phân tích và thiết kế

2.1 Giới thiệu về bộ dữ liệu

1. Bộ dữ liệu Customers



Hình 18: Thống kê dữ liệu từng cột trong bảng Customers

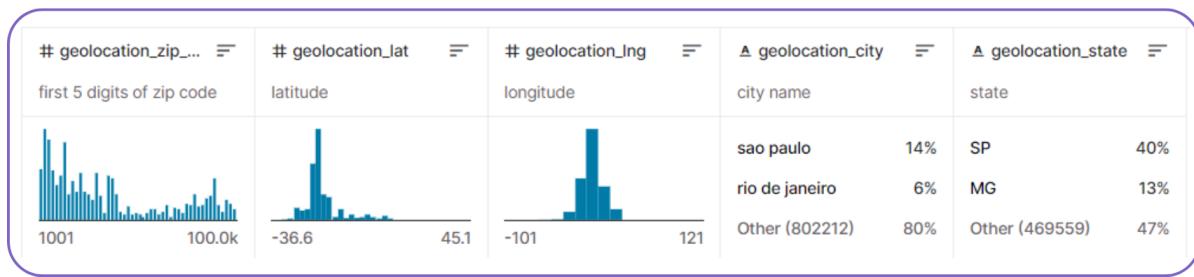
- **Customer_ID:** Mỗi đơn hàng có một customer_id duy nhất.
- **Customer_unique_ID:** định danh duy nhất của một khách hàng.
- **Customer_zip_code_prefix:** năm chữ số đầu tiên của mã zip của khách hàng
- **Customer_City:** tên thành phố của khách hàng
- **Customer_State:** Bang

Data columns (total 5 columns):

#	Column	Non-Null Count	Dtype
---	---	-----	-----
0	customer_id	99441	non-null object
1	customer_unique_id	99441	non-null object
2	customer_zip_code_prefix	99441	non-null int64
3	customer_city	99441	non-null object
4	customer_state	99441	non-null object
	dtypes: int64(1), object(4)		
	memory usage: 3.8+ MB		

Hình 19: Chi tiết dữ liệu của bảng Customers

2. Bộ dữ liệu Geolocation



Hình 20: Thống kê dữ liệu từng cột trong bảng Geolocation

Bộ dữ liệu này có thông tin về mã zip Brazil và tọa độ lat/lng của nó. Được sử dụng để vẽ bản đồ và tìm khoảng cách giữa người bán và khách hàng.

- **Geolocation_zip_code_prefix:** 5 chữ số đầu tiên của mã zip
- **Geolocation_lat:** Vĩ độ
- **Geolocation_lng:** Kinh độ
- **Geolocation_city:** Tên thành phố
- **Geolocation_state:** Định vị địa lý

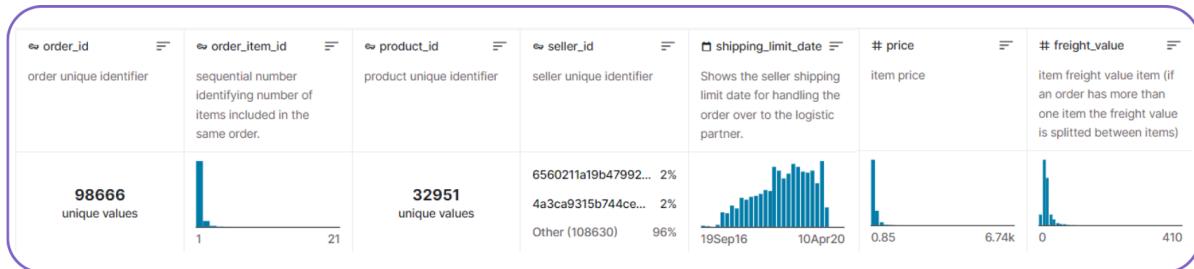
```
Data columns (total 5 columns):
 #   Column           Non-Null Count  Dtype  
 ---  -- 
 0   geolocation_zip_code_prefix  1000163 non-null   int64  
 1   geolocation_lat            1000163 non-null   float64 
 2   geolocation_lng            1000163 non-null   float64 
 3   geolocation_city           1000163 non-null   object  
 4   geolocation_state          1000163 non-null   object  
 dtypes: float64(2), int64(1), object(2)
 memory usage: 38.2+ MB
```

Hình 21: Chi tiết dữ liệu của bảng Geolocation

3. Bộ dữ liệu Order_items

Bộ dữ liệu này bao gồm dữ liệu về các mặt hàng đã mua trong mỗi đơn hàng

- **Order_id:** Mã định danh đặt hàng (duy nhất)
- **Order_item_id:** Số thứ tự xác định số mục được bao gồm trong cùng một thứ tự.
- **Product_id:** Mã định danh duy nhất của sản phẩm
- **Seller_id:** Mã định danh duy nhất của người bán
- **Shipping_limit_date:** Ngày giới hạn vận chuyển của người bán để xử lý đơn đặt hàng cho đối tác hậu cần.



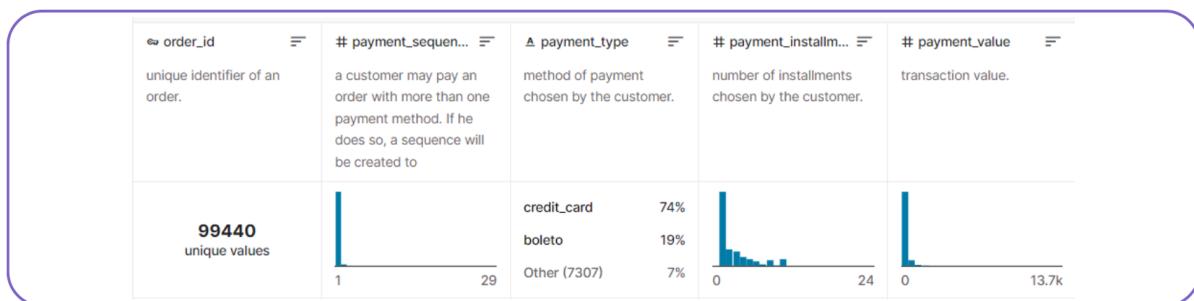
Hình 22: Thống kê dữ liệu từng cột trong bảng Order_items

- **Price:** giá mặt hàng
- **Freight_value:** giá trị vận chuyển mặt hàng

```
Data columns (total 7 columns):
 #   Column           Non-Null Count Dtype  
 --- 
 0   order_id         112650 non-null  object  
 1   order_item_id    112650 non-null  int64   
 2   product_id       112650 non-null  object  
 3   seller_id        112650 non-null  object  
 4   shipping_limit_date 112650 non-null  object  
 5   price            112650 non-null  float64 
 6   freight_value    112650 non-null  float64 
 dtypes: float64(2), int64(1), object(4)
 memory usage: 6.0+ MB
```

Hình 23: chi tiết dữ liệu của bảng Order_items

4. Bộ dữ liệu Order_payment



Hình 24: Thống kê dữ liệu từng cột trong bảng Order_payment

Bộ dữ liệu này bao gồm dữ liệu về các tùy chọn thanh toán đơn đặt hàng

- **Order_id:** định danh duy nhất của một đơn đặt hàng
- **Payment_sequential:** một khách hàng có thể thanh toán một đơn đặt hàng bằng nhiều phương thức thanh toán.

- **Payment_type:** Phương thức thanh toán mà khách hàng lựa chọn
- **Payment_installments:** Số đợt do khách hàng lựa chọn
- **Payment_value:** giá trị giao dịch

```
Data columns (total 5 columns):
 #   Column           Non-Null Count   Dtype  
 --- 
 0   order_id         103886 non-null    object 
 1   payment_sequential 103886 non-null    int64  
 2   payment_type      103886 non-null    object 
 3   payment_installments 103886 non-null    int64  
 4   payment_value     103886 non-null    float64
 dtypes: float64(1), int64(2), object(2)
 memory usage: 4.0+ MB
```

Hình 25: Chi tiết dữ liệu của bảng Order_payment

5. Bộ dữ liệu Orders



Hình 26: Thống kê dữ liệu từng cột trong bảng Orders

- **Order_ID:** Định danh duy nhất của đơn hàng
- **Customer_ID:** Mỗi đơn hàng có một customer_id duy nhất
- **Order_status:** Trạng thái đơn hàng
- **Order_purchase_timestamp:** Hiển thị dấu thời gian mua hàng
- **Order_approved_at:** Hiển thị dấu thời gian phê duyệt thanh toán
- **Order_delivered_carrier_date:** Hiển thị dấu thời gian đăng đơn hàng. Khi nó được xử lý cho đối tác hậu cần.
- **Order_delivered_customer_date:** Hiển thị ngày giao hàng thực tế cho khách hàng
- **Order_estimated_delivery_date:** Hiển thị ngày giao hàng ước tính đã được thông báo cho khách hàng tại thời điểm mua hàng.

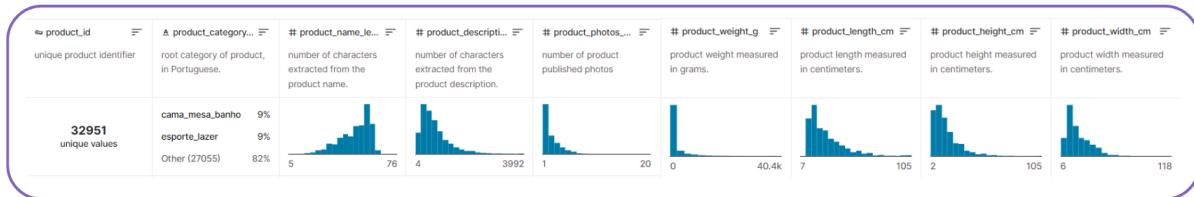
```

Data columns (total 8 columns):
 #   Column           Non-Null Count Dtype
 --- 
 0   order_id         99441 non-null  object
 1   customer_id     99441 non-null  object
 2   order_status     99441 non-null  object
 3   order_purchase_timestamp 99441 non-null  object
 4   order_approved_at 99281 non-null  object
 5   order_delivered_carrier_date 97658 non-null  object
 6   order_delivered_customer_date 96476 non-null  object
 7   order_estimated_delivery_date 99441 non-null  object
 dtypes: object(8)
 memory usage: 6.1+ MB

```

Hình 27: Chi tiết dữ liệu của bảng Orders

6. Bộ dữ liệu Products



Hình 28: Thống kê dữ liệu từng cột trong bảng Products

- **Product_id:** Số nhận dạng của sản phẩm duy nhất
- **Product_category:** Danh mục gốc sản phẩm, bằng tiếng BDN
- **Product_name_length:** Số ký tự được trích ra từ tên sản phẩm.
- **Product_description_length:** Số ký tự được trích ra từ mô tả SP
- **Product_photos_qty:** Số lượng ảnh công bố sản phẩm
- **Product_weight_g:** Trọng lượng sản phẩm tính bằng gam
- **Product_length_cm:** Chiều dài sản phẩm tính bằng cm
- **Product_height_cm:** Chiều cao sản phẩm tính bằng cm
- **Product_width_cm:** chiều rộng sản phẩm tính cm

```

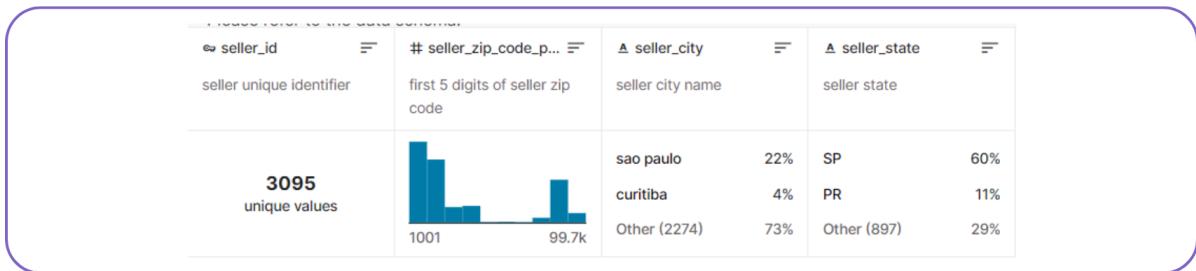
Data columns (total 9 columns):
 #   Column           Non-Null Count Dtype  
--- 
 0   product_id       32951 non-null  object  
 1   product_category_name 32341 non-null  object  
 2   product_name_lenght    32341 non-null  float64 
 3   product_description_lenght 32341 non-null  float64 
 4   product_photos_qty     32341 non-null  float64 
 5   product_weight_g      32949 non-null  float64 
 6   product_length_cm     32949 non-null  float64 
 7   product_height_cm     32949 non-null  float64 
 8   product_width_cm      32949 non-null  float64 

dtypes: float64(7), object(2)
memory usage: 2.3+ MB

```

Hình 29: Chi tiết dữ liệu của bảng Products

7. Bộ dữ liệu Sellers



Hình 30: Thống kê dữ liệu từng cột trong bảng Sellers

- **Seller_id:** định danh duy nhất của người bán
- **Seller_zip_code_prefix:** 5 chữ số đầu tiên của mã zip của người bán
- **Seller_city:** Tên thành phố người bán
- **Seller_state:** Bang

```

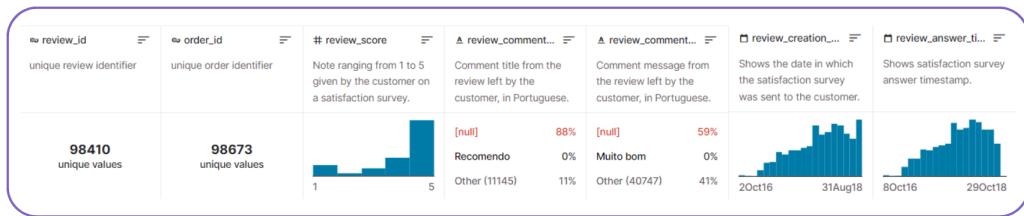
Data columns (total 4 columns):
 #   Column           Non-Null Count Dtype  
--- 
 0   seller_id       3095 non-null  object  
 1   seller_zip_code_prefix  3095 non-null  int64  
 2   seller_city      3095 non-null  object  
 3   seller_state     3095 non-null  object  

dtypes: int64(1), object(3)
memory usage: 96.8+ KB

```

Hình 31: Chi tiết dữ liệu của bảng Sellers

8. Bộ dữ liệu Reviews



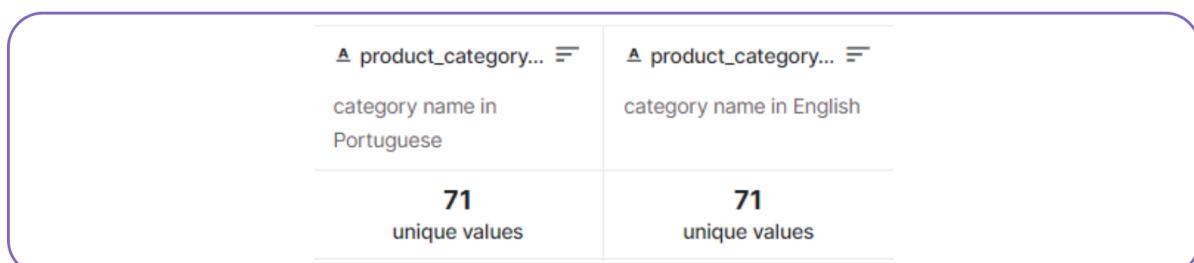
Hình 32: Thống kê dữ liệu từng cột trong bảng Reviews

- **Review_id:** Số nhận dạng đánh giá duy nhất
- **Oder_id:** mã định danh đơn hàng duy nhất
- **Review_Score:** Dánh giá mức độ hài lòng của khách hàng
- **Review_comment_title:** Tiêu đề nhận xét đánh giá
- **Review_comment_messenge:** Thông báo nhận xét
- **Review_creation_date:** Hiển thị ngày gửi bài khảo sát về mức độ hài lòng cho khách hàng.
- **Review_answer_timestamp:** Hiển thị dấu thời gian của câu trả lời khảo sát mức độ hài lòng

```
Data columns (total 7 columns):
 #   Column           Non-Null Count  Dtype  
 --- 
 0   review_id        99224 non-null   object 
 1   order_id         99224 non-null   object 
 2   review_score     99224 non-null   int64  
 3   review_comment_title 11568 non-null   object 
 4   review_comment_message 40977 non-null   object 
 5   review_creation_date 99224 non-null   object 
 6   review_answer_timestamp 99224 non-null   object 
 dtypes: int64(1), object(6)
 memory usage: 5.3+ MB
```

Hình 33: Chi tiết dữ liệu của bảng Reviews

9. Bộ dữ liệu Translation



Hình 34: Thống kê dữ liệu từng cột trong bảng Translation

- **product_category_name**: Tên mặt hàng bằng tiếng Bồ Đào Nha
- **product_category_name_english**: Tên mặt hàng bằng tiếng Anh

```
Data columns (total 2 columns):
 #   Column           Non-Null Count Dtype
 ---  -----
 0   product_category_name    71 non-null   object
 1   product_category_name_english 71 non-null   object
 dtypes: object(2)
 memory usage: 1.2+ KB
```

Hình 35: Chi tiết dữ liệu của bảng Translation

2.2 Data Exploration

1. Tổng quan về bộ dữ liệu

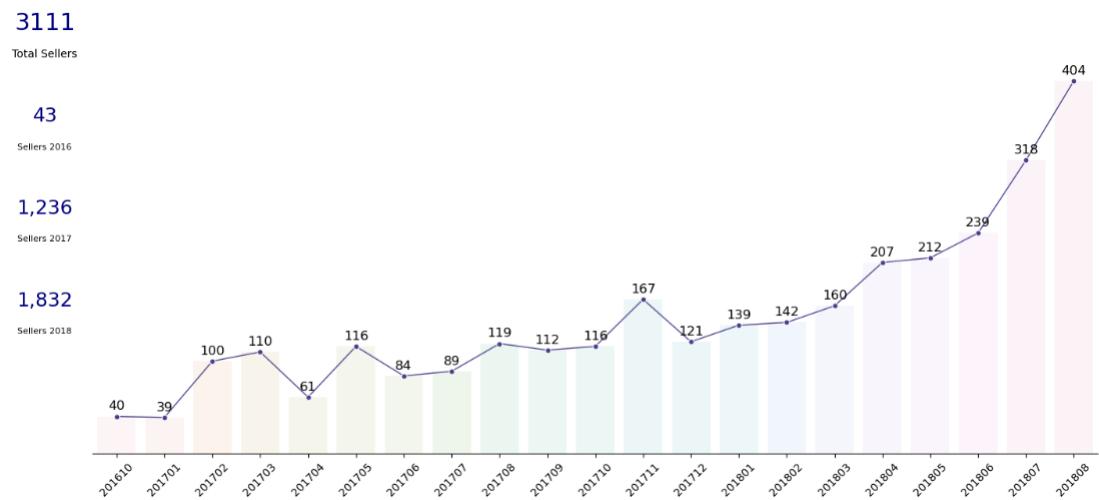
Bảng dữ liệu mô tả đầy đủ thông tin các bảng: Tên bảng, Tên cột của từng bảng, Tổng số dòng, Tổng số cột, Số dòng bị trùng lặp, Số dòng null và Cột chứa dữ liệu bị null.

	Dataset	Columns	Total_rows	Total_cols	Total_duplicate	Total_null	Null_cols
0	Customers	customer_id, customer_unique_id, customer_zip_code_prefix, customer_city, customer_state	99441	5	0	0	
1	Sellers	seller_id, seller_zip_code_prefix, seller_city, seller_state	3095	4	0	0	
2	Reviews	review_id, order_id, review_score, review_comment_title, review_comment_message, review_creation_date, review_answer_timestamp	99224	7	0	145903	review_comment_title, review_comment_message
3	Items	order_id, order_item_id, product_id, seller_id, shipping_limit_date, price, freight_value	112650	7	0	0	
4	Products	product_id, product_category_name, product_name_length, product_description_length, product_photos_qty, product_weight_g, product_length_cm, product_height_cm, product_width_cm	32951	9	0	2448	product_category_name, product_name_length, product_description_length, product_photos_qty, product_weight_g, product_length_cm, product_height_cm, product_width_cm
5	Geolocations	geolocation_zip_code_prefix, geolocation_lat, geolocation_lng, geolocation_city, geolocation_state	1000163	5	261831	0	
6	Category translation	product_category_name, product_category_name_english	71	2	0	0	
7	Orders	order_id, customer_id, order_status, order_purchase_timestamp, order_approved_at, order_delivered_carrier_date, order_delivered_customer_date, order_estimated_delivery_date	99441	8	0	4908	order_approved_at, order_delivered_carrier_date, order_delivered_customer_date
8	Payments	order_id, payment_sequential, payment_type, payment_installments, payment_value	103886	5	0	0	

Hình 36: Tổng quan về bộ dữ liệu

2. Sellers

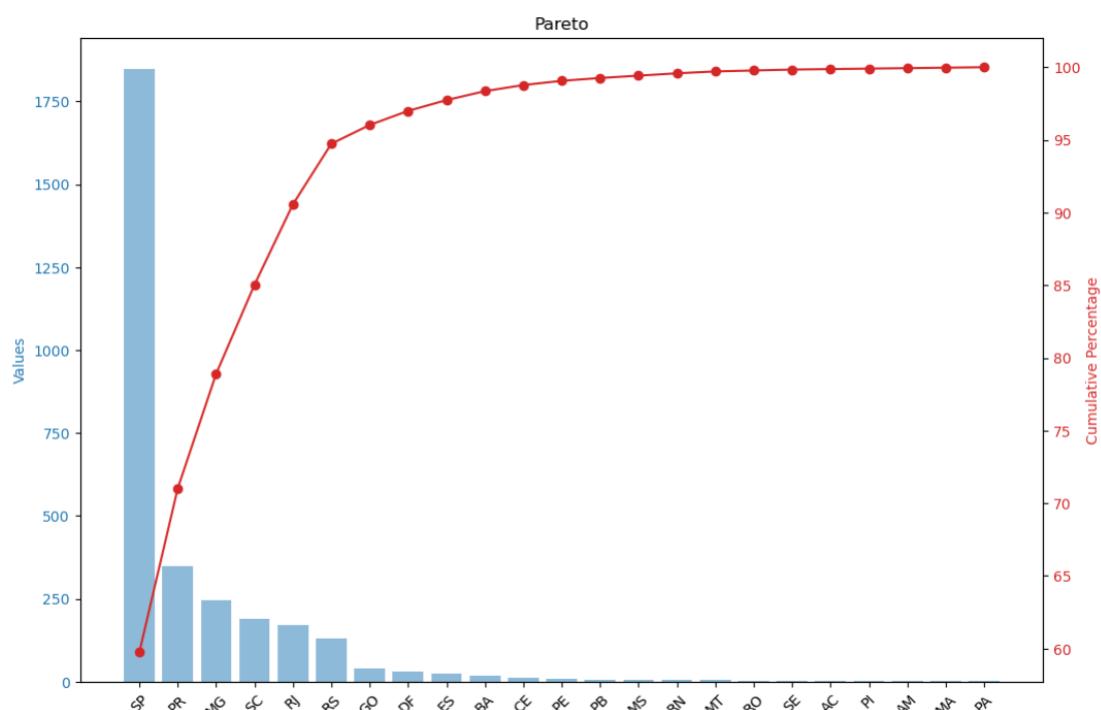
(a) Sự tăng trưởng của số người bán hàng theo từng mốc thời gian.



Hình 37: Số lượng người bán theo thời gian

⇒ Số lượng người bán hàng đã tăng theo từng năm, năm 2018 có sự tăng mạnh rõ rệt. 4 tháng năm 2016 có 43 người, năm 2017 có 1236, với 8 tháng năm 2018 thì đã vượt năm 2017 là 1832 người bán. Tổng số người bán của bộ dữ liệu là 3111 người.

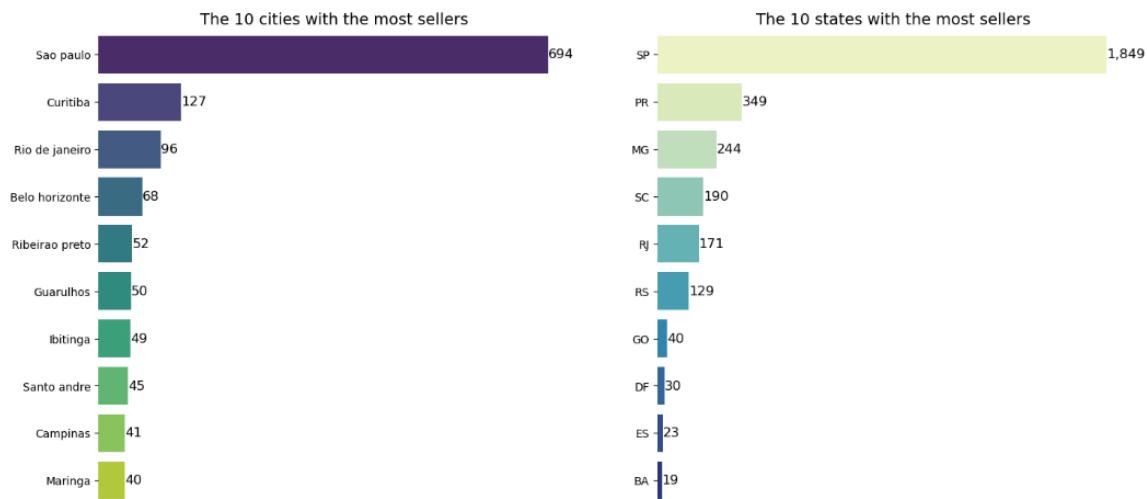
(b) Sự phân bố số lượng người bán theo khu vực



Hình 38: Số lượng người bán theo khu vực

⇒ Số lượng người bán ở bang SP lớn nhất, vượt trội hơn tất cả các bang còn lại. 95% số lượng người bán tập trung chủ yếu ở 6 bang: SP, PR, MG, SC, RJ, RS.

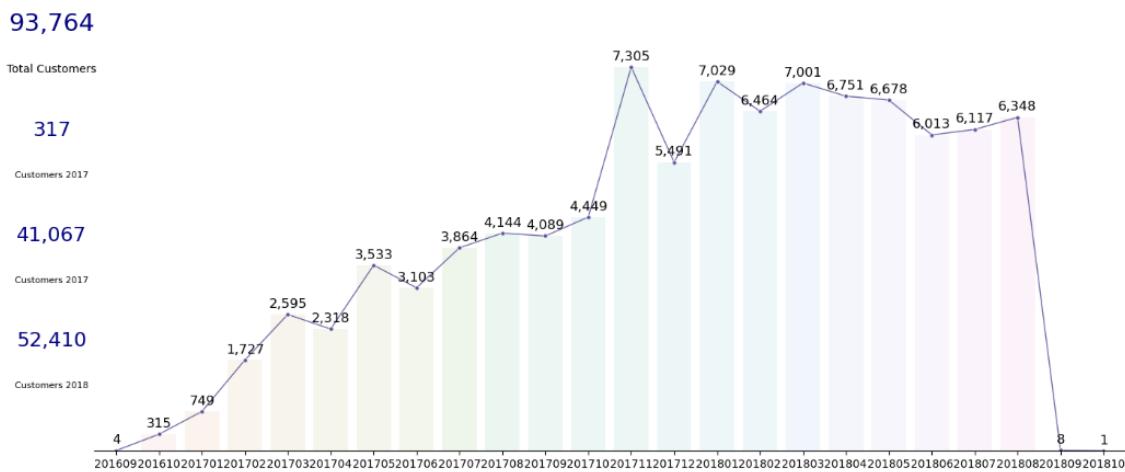
(c) Top 10 bang, thành phố có số lượng người bán nhiều nhất.



Hình 39: Top những thành phố, bang có số lượng người bán nhiều

3. Customers

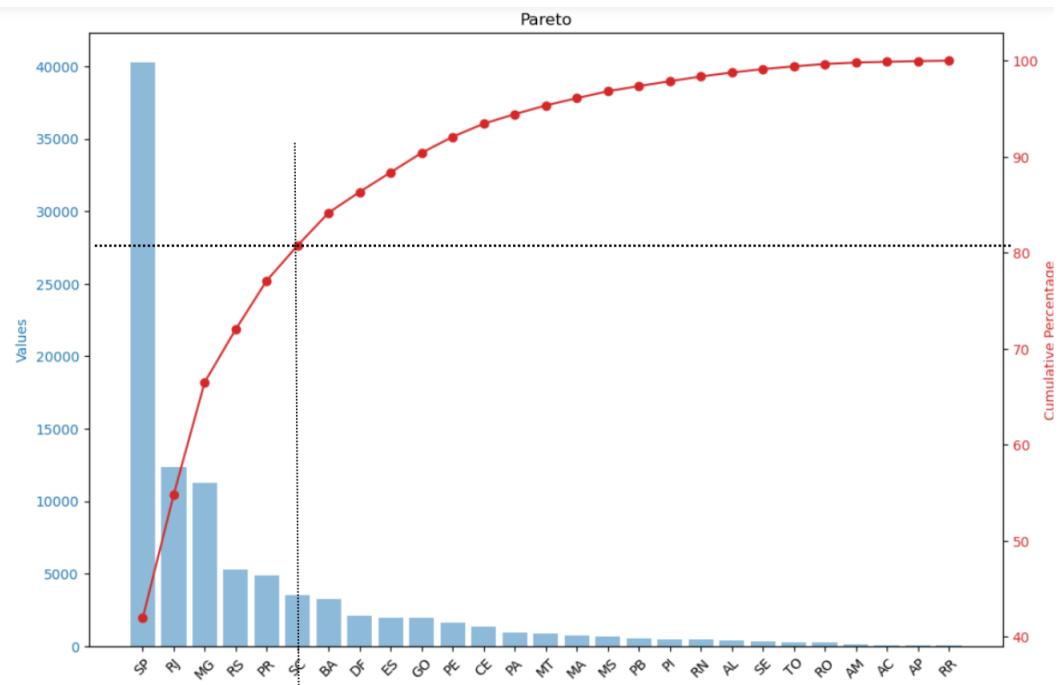
(a) Sự tăng trưởng của số lượng khách hàng theo từng mốc thời gian.



Hình 40: Số lượng khách hàng theo thời gian

⇒ Số lượng người bán hàng đã tăng theo từng năm, về sau thì dần ổn định. 4 tháng năm 2016 có 317 người, năm 2017 có 41067, với 8 tháng năm 2018 thì đã vượt năm 2017 là 52410 khách hàng. Tổng số khách hàng của bộ dữ liệu là 93764 người.

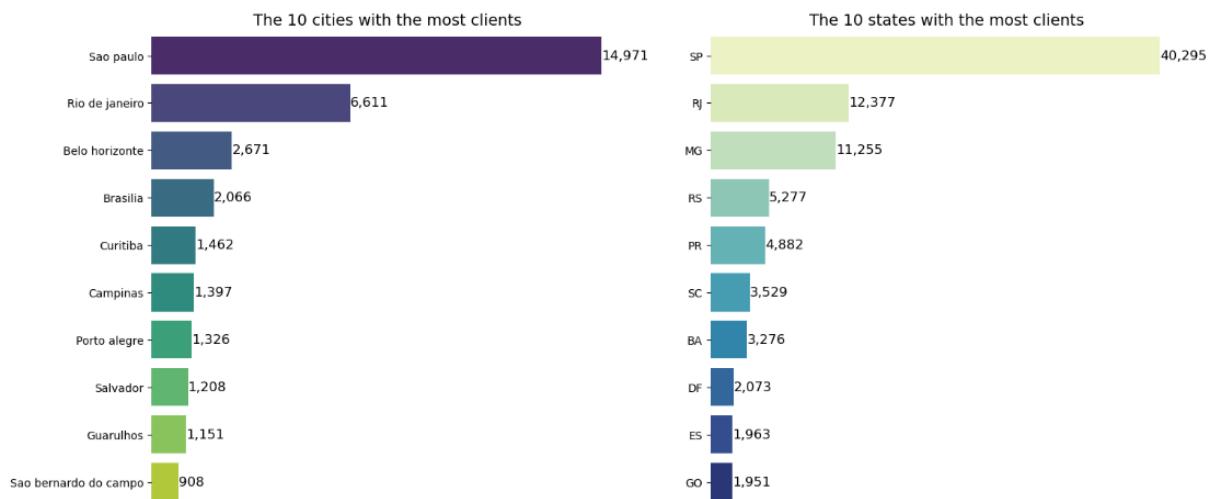
(b) Sự phân bố số lượng người bán theo khu vực



Hình 41: Số lượng khách hàng theo khu vực

⇒ Số lượng khách hàng ở bang SP lớn nhất, vượt trội hơn tất cả các bang còn lại. 80% số lượng người bán tập trung chủ yếu ở 6 bang: SP, PR, MG, SC, RJ, RS.

(c) Top 10 bang, thành phố có số lượng người bán nhiều nhất.



Hình 42: Top những thành phố, bang có số lượng khách hàng nhiều

4. Orders

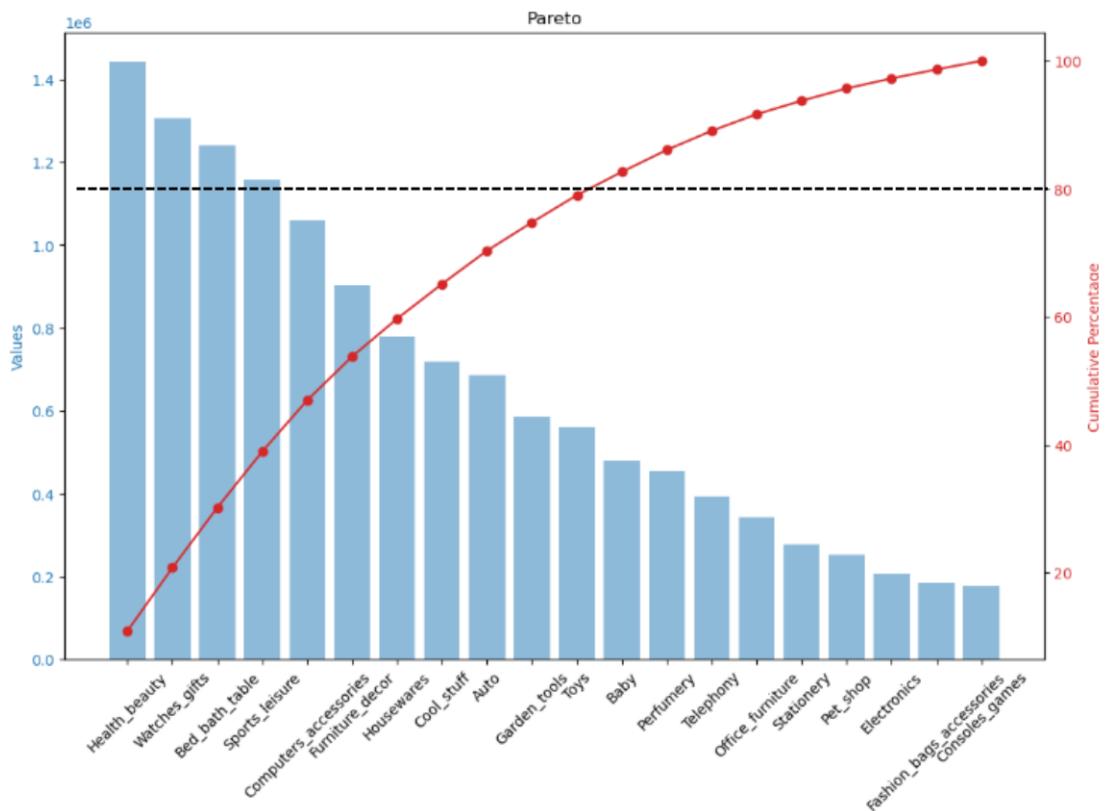


Hình 43: Số lượng đơn hàng theo thời gian

⇒ Số lượng đơn hàng đã tăng theo từng năm, về sau thì dần ổn định. 4 tháng năm 2016 có 329 đơn, năm 2017 có 42697 đơn, với 8 tháng năm 2018 thì đã vượt năm 2017 là 54011 đơn hàng. Tổng số đơn hàng của bộ dữ liệu là 97037 đơn.

5. Payment_value

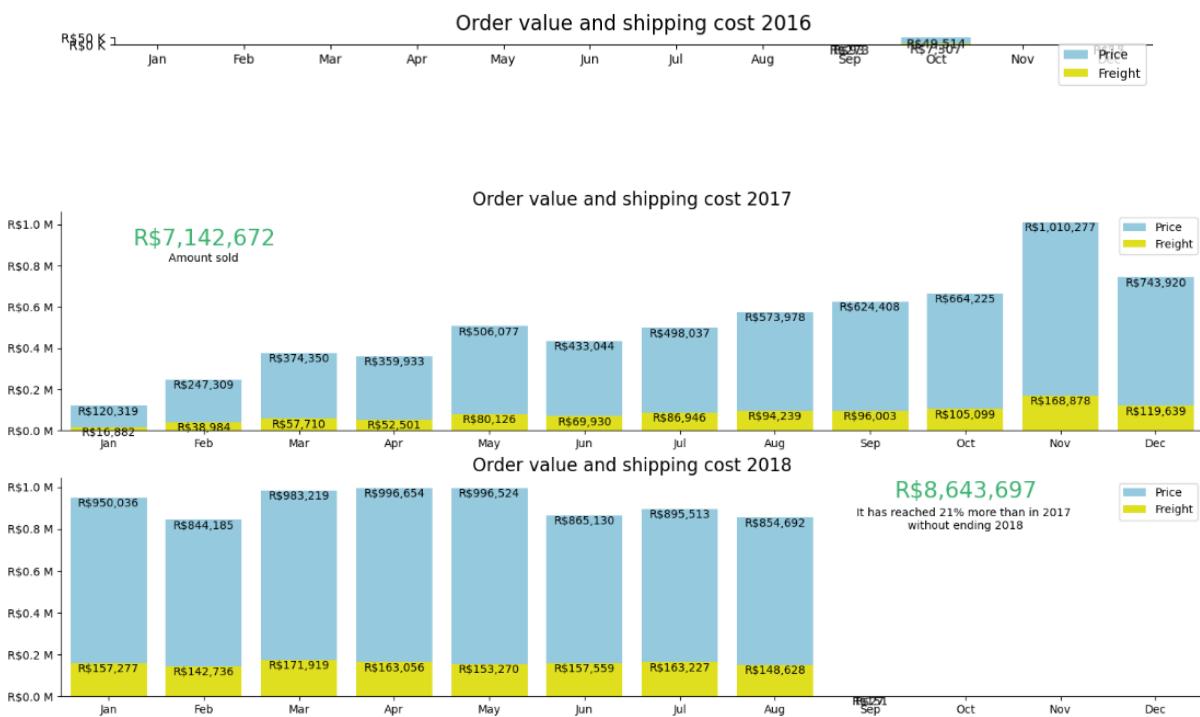
(a) Doanh số theo mặt hàng



Hình 44: Số lượng đơn hàng theo thời gian

⇒ Doanh số ở mặt hàng Healthy _ beauty cao nhất. Doanh số các mặt hàng giảm dần từ trái qua phải, 80% doanh số đạt tại mặt hàng Toys

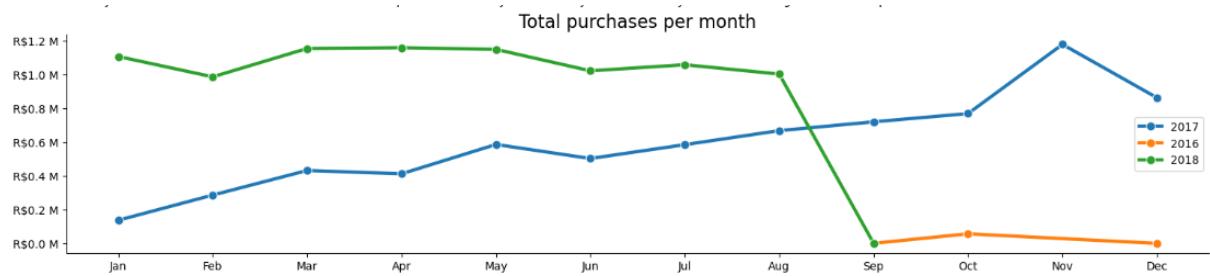
(b) Giá trị đơn hàng, cước phí vận chuyển theo từng mốc thời gian



Hình 45: Giá trị đơn hàng, cước phí vận chuyển theo từng tháng

⇒ Giá trị đơn hàng, cước phí vận chuyển tăng dần theo các tháng từ năm 2016-2017, đến năm 2018 thì dần ổn định và cao hơn so với 2 năm trước.

(c) Giá trị thanh toán theo từng tháng ở các năm

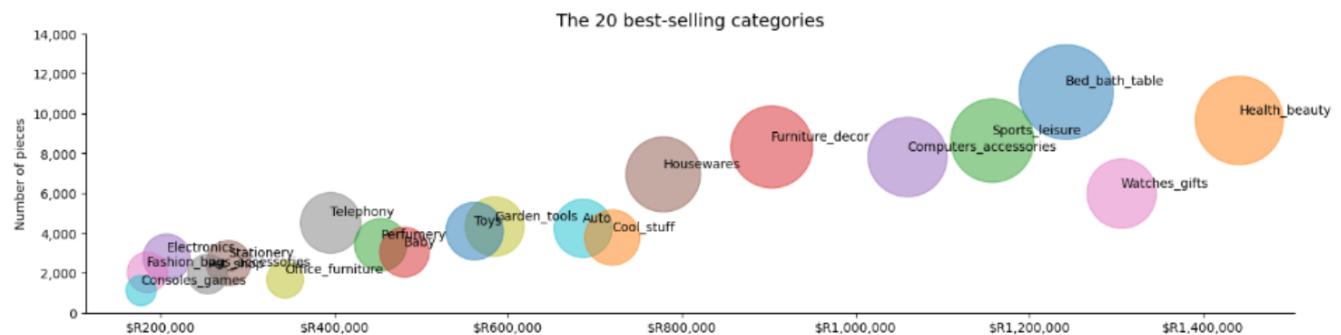


Hình 46: Giá trị theo từng tháng ở các năm

⇒ Ở như biểu đồ này, ta có thể thấy giá trị thanh toán ở các năm phân chia thành 3 phần rõ rệt như nào nào. Năm 2016, rất thấp. Năm 2017, tăng dần theo các tháng. Năm 2018, cao hơn hẳn so với năm 2017.

6. Category

(a) Top 20 mặt hàng bán chạy nhất



Hình 47: Top 20 mặt hàng bán chạy nhất 2016-2018

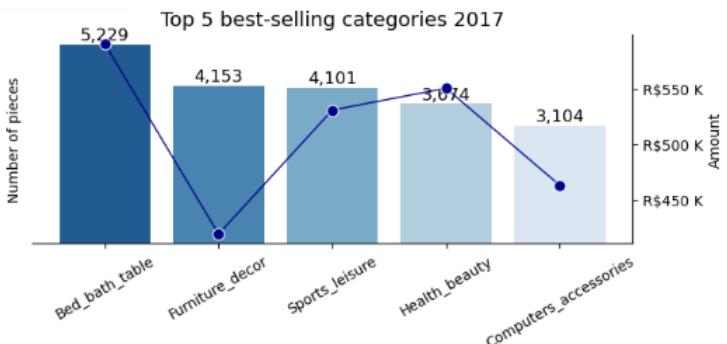
⇒ Healthy_beauty bán được doanh số nhiều nhất, Bed_bath_table bán được số lượng nhiều nhất

(b) Top 5 mặt hàng bán chạy nhất 2016



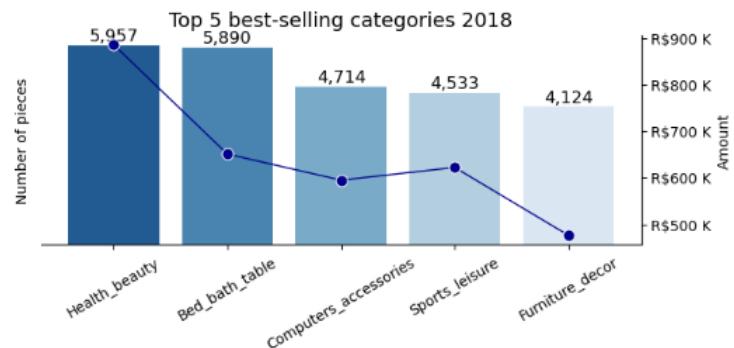
Hình 48: Top 5 mặt hàng bán chạy nhất 2016

(c) Top 5 mặt hàng bán chạy nhất 2017



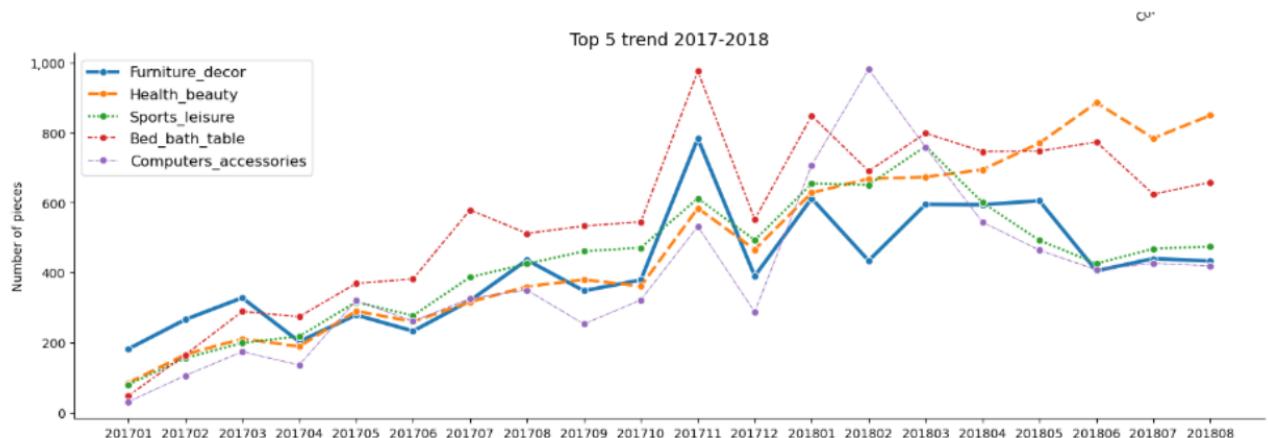
Hình 49: Top 5 mặt hàng bán chạy nhất 2017

(d) Top 5 mặt hàng bán chạy nhất 2018



Hình 50: Top 5 mặt hàng bán chạy nhất 2018

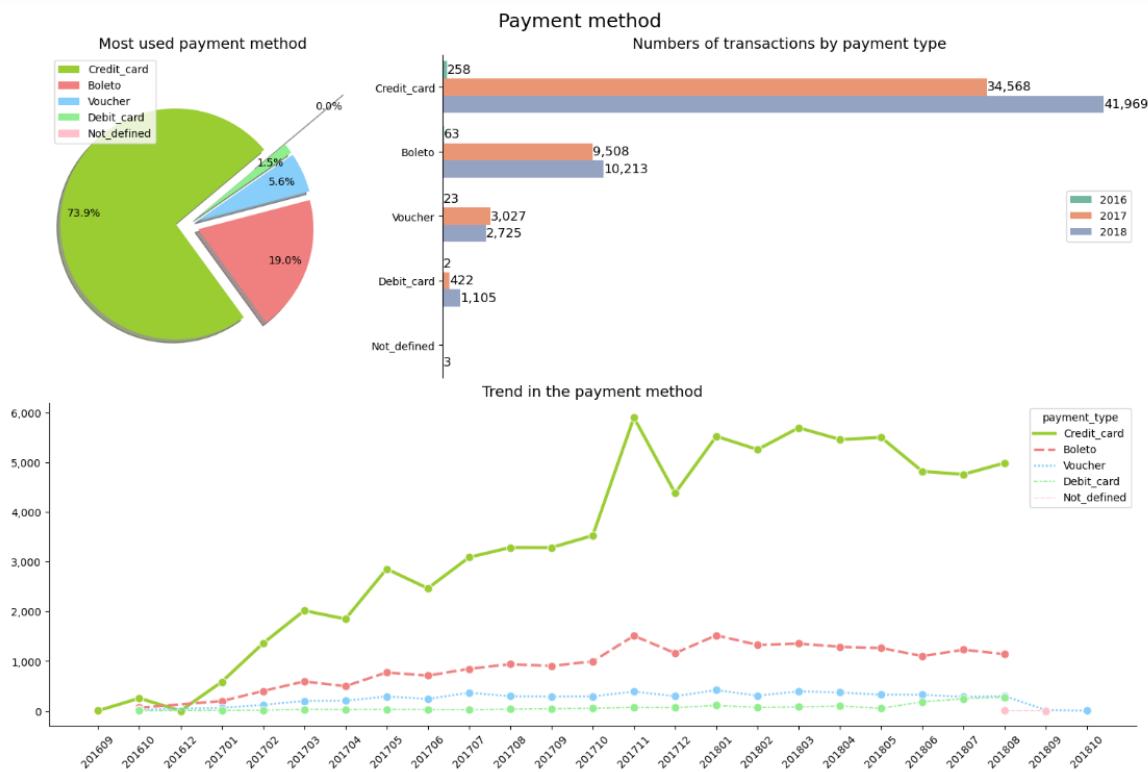
(e) Top 5 mặt hàng được bán với số lượng nhiều nhất theo từng tháng năm 2017-2018



Hình 51: Top 5 mặt hàng được bán nhiều nhất theo từng tháng năm 2017-2018

⇒ Như 2 biểu đồ trước ta có thể được 5 mặt hàng được bán nhiều nhất năm 2017 và 2018 là như nhau. Có thể thấy, xu hướng mua hàng ở 5 loại mặt hàng này đang tăng. Có lúc giảm, nhưng không đáng kể so với mặt bằng chung.

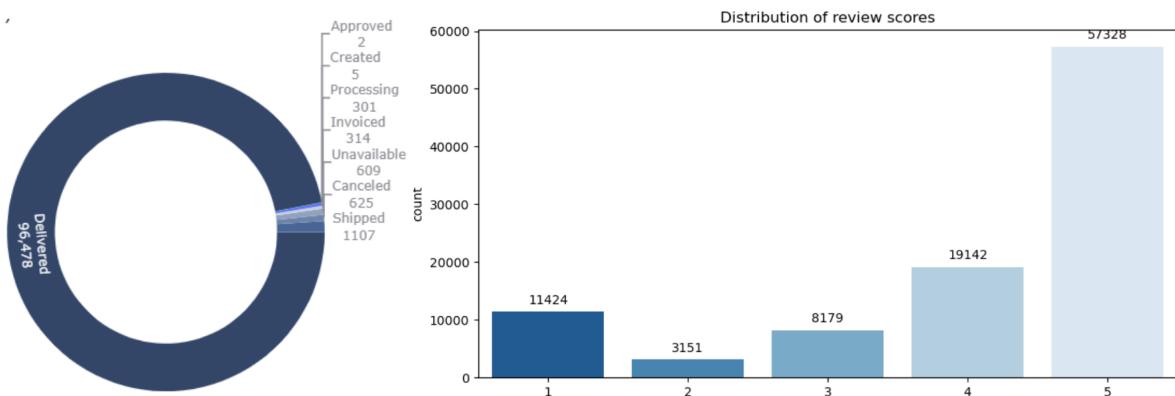
7. Payment _ method



Hình 52: Hình thức thanh toán

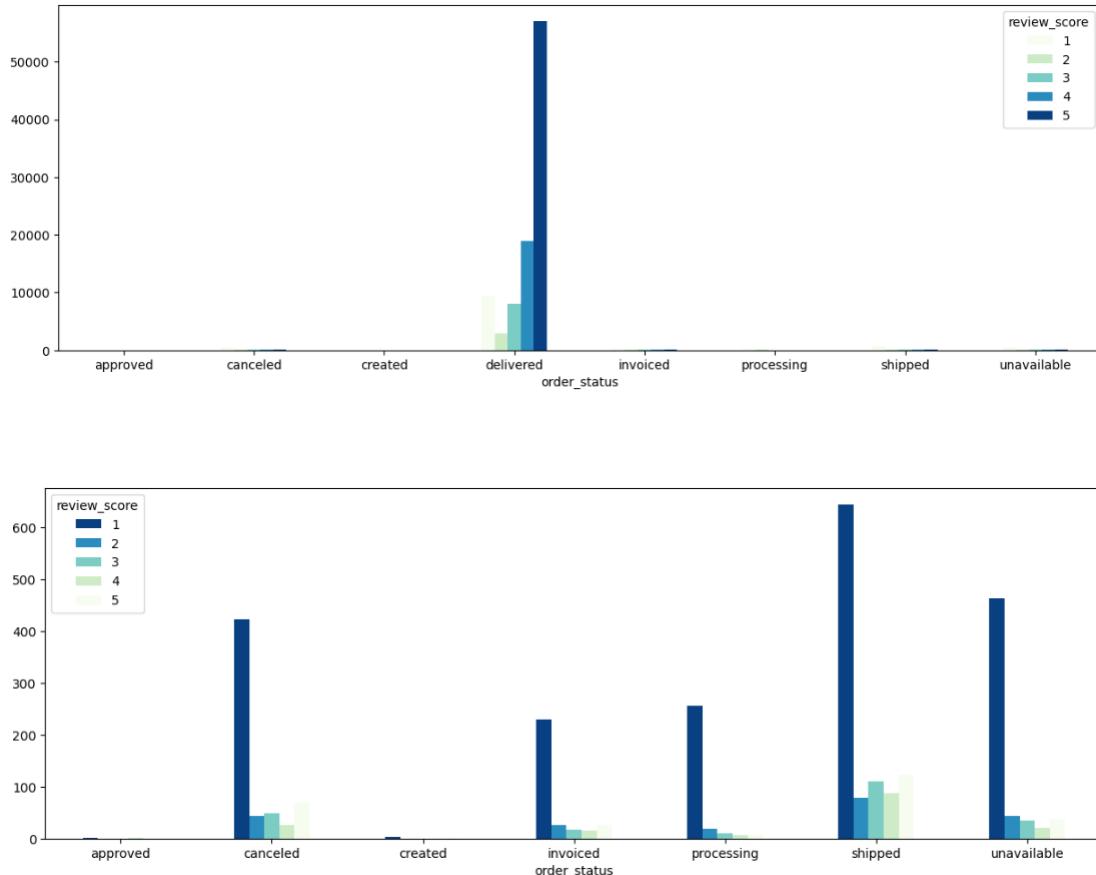
⇒ Nhìn vào biểu đồ ta có thể thấy: hình thức thanh toán bằng CreditCard chiếm 1 tỷ lệ khá lớn gần ~ 74%, tiếp theo đến là Boleto chiếm ~ 19%. Số lượng đơn hàng tăng dần nên số lượng thanh toán cũng dần tăng lên. Xu hướng thanh toán ở CreditCard tăng mạnh, Boleto tăng nhẹ, còn các hình thức khác thì không thay đổi nhiều theo từng mốc thời gian khác nhau.

8. Status & Review Scores



Hình 53: Trạng thái giao hàng và điểm đánh giá

⇒ Số lượng đơn hàng được giao chiếm 1 tỉ lệ khá lớn $\sim 95\%$ còn lại khoảng 5% là các đơn hàng nằm ở các trạng thái khác. Số lượng đơn hàng được đánh giá 5 điểm cao nhất, 2 điểm thấp nhất.



Hình 54: Điểm đánh giá theo từng trạng thái giao hàng

⇒ Ta có thể thấy được mối tương quan giữa điểm đánh giá và trạng thái giao hàng. Các đơn có trạng thái 'delivered' đa số đều được đánh giá 4-5 điểm, số lượng đơn được đánh giá < 3 điểm chiếm 1 tỉ lệ khá nhỏ so với tổng thể. Các trạng thái còn lại thì đa số sẽ bị đánh giá 1 điểm.

9. Data Analyst Tools

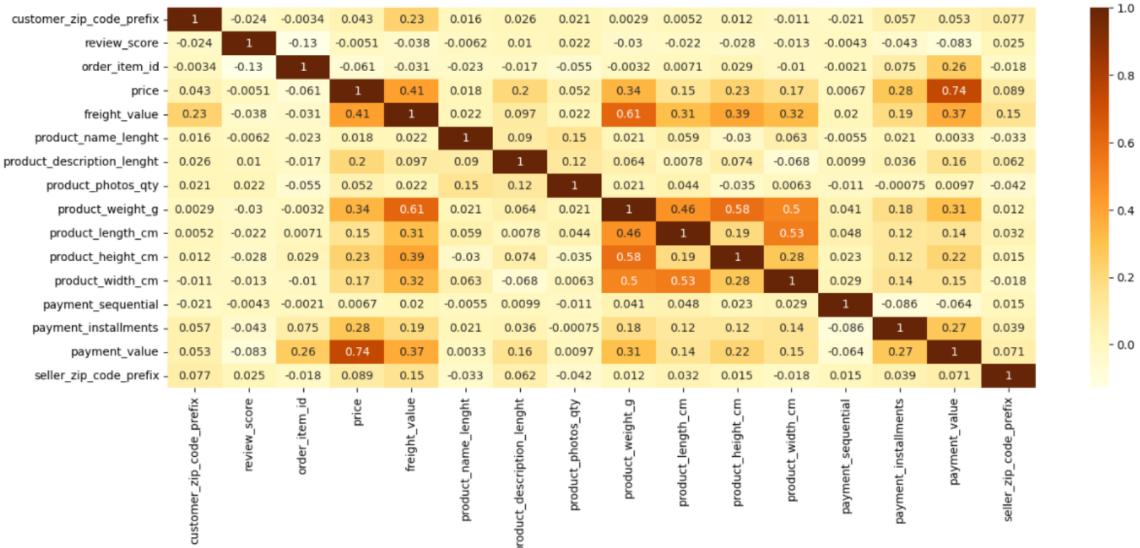
	review_score	price	freight_value	product_weight_g	product_length_cm
count	115609.000000	115609.000000	115609.000000	115608.000000	115608.000000
mean	4.034409	120.619850	20.056880	2113.907697	30.307903
std	1.385584	182.653476	15.836184	3781.754895	16.211108
min	1.000000	0.850000	0.000000	0.000000	7.000000
25%	4.000000	39.900000	13.080000	300.000000	18.000000
50%	5.000000	74.900000	16.320000	700.000000	25.000000
75%	5.000000	134.900000	21.210000	1800.000000	38.000000
max	5.000000	6735.000000	409.680000	40425.000000	105.000000

	product_height_cm	product_width_cm	payment_sequential	payment_installments	payment_value
count	115608.000000	115608.000000	115609.000000	115609.000000	115609.000000
mean	16.638477	23.113167	1.093747	2.946233	172.387379
std	13.473570	11.755083	0.729849	2.781087	265.873969
min	2.000000	6.000000	1.000000	0.000000	0.000000
25%	8.000000	15.000000	1.000000	1.000000	60.870000
50%	13.000000	20.000000	1.000000	2.000000	108.050000
75%	20.000000	30.000000	1.000000	4.000000	189.480000
max	105.000000	118.000000	29.000000	24.000000	13664.080000

Hình 55: Phân tích thống kê các dữ liệu

⇒ Đơn hàng có giá cao nhất là 6735, thấp nhất là 0.85, giá trị trung bình 120.61. Đơn hàng có phí vận chuyển cao nhất là 409.68, thấp nhất là 0 (đơn hàng bị hủy). Số lần thanh toán nhiều nhất là 24, thấp nhất là 0 (đơn hàng bị hủy), trung bình ~ 3 lần. Giá trị thanh toán cao nhất 13664 thấp nhất là 0 (đơn hàng bị hủy), trung bình 172.4

10. Ma trận hệ số tương quan



⇒ Giá trị tương quan của price và payment_value là 0.74, con số này khá cao, có thể thấy được sự tương quan tuyến tính ở đây. product_weight_g và freight_value là 0.61, phí vận chuyển phụ thuộc tuyến tính vào trọng lượng của vật.

2.3 Giới thiệu về ODS

Kho dữ liệu hoạt động (ODS) được sử dụng để báo cáo hoạt động và là nguồn dữ liệu cho kho dữ liệu doanh nghiệp. Nó là một yếu tố bổ sung cho EDW trong bối cảnh hỗ trợ quyết định và được sử dụng để báo cáo hoạt động, kiểm soát và ra quyết định, trái ngược với EDW, được sử dụng để hỗ trợ quyết định chiến thuật và chiến lược.

ODS là cơ sở dữ liệu được thiết kế để tích hợp dữ liệu từ nhiều nguồn cho các hoạt động bổ sung trên dữ liệu, để báo cáo, kiểm soát và hỗ trợ quyết định hoạt động. ODS nằm giữa Data source và Data Warehouse (có thể cùng lớp với Staging, gọi chung là integrated layer). Không giống như một kho lưu trữ dữ liệu tổng thể về sản xuất, dữ liệu không được chuyển đổi lại các hệ thống vận hành. Nó có thể được chuyển cho các hoạt động tiếp theo và đến kho dữ liệu để báo cáo.

Không nên nhầm lẫn ODS với trung tâm dữ liệu doanh nghiệp (EDH). Một kho lưu trữ dữ liệu hoạt động sẽ lấy dữ liệu giao dịch từ một hoặc nhiều hệ thống sản xuất và tích hợp nó một cách lỏng lẻo, ở một số khía cạnh, nó vẫn có định hướng đối tượng, tích hợp và biến thể thời gian, nhưng không có các ràng buộc về tính biến động. Sự tích hợp này chủ yếu đạt được thông qua việc sử dụng các cấu trúc và nội dung EDW.

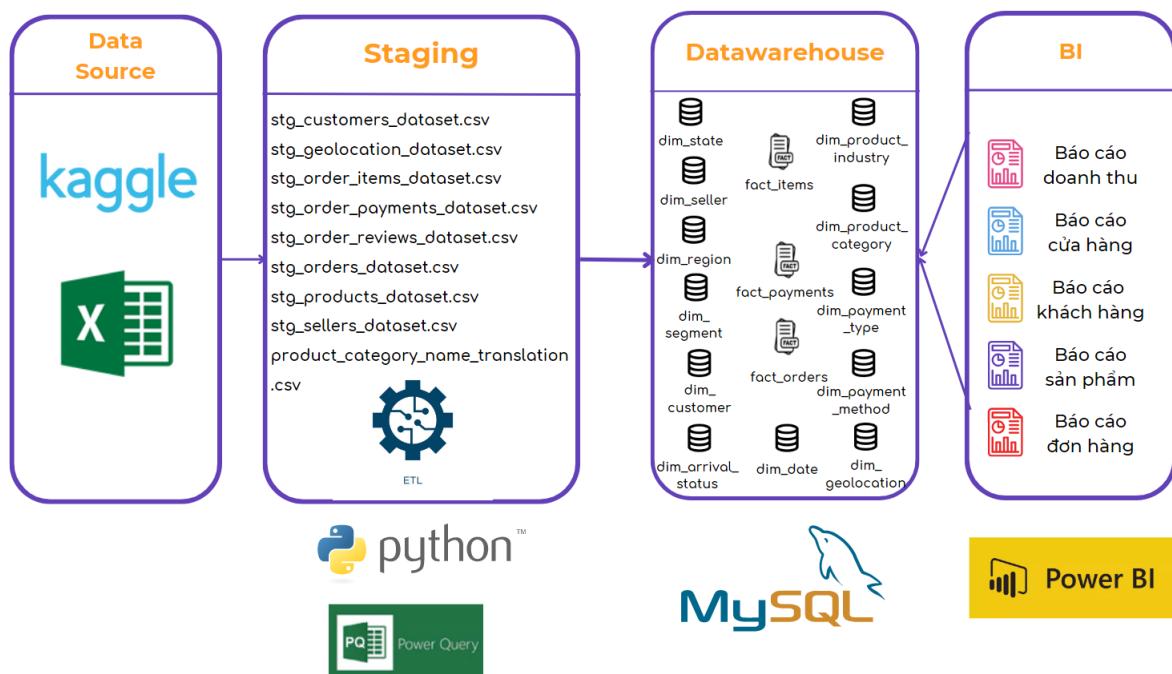
Vì dữ liệu bắt nguồn từ nhiều nguồn nên việc tích hợp thường liên quan đến việc dọn dẹp, giải quyết tình trạng dư thừa và kiểm tra tính toàn vẹn của các quy tắc nghiệp vụ. ODS thường được thiết kế để chứa dữ liệu cấp thấp hoặc nguyên tử (không thể phân chia) (chẳng hạn như giao dịch và giá cả) với lịch sử hạn chế được ghi lại "thời gian thực" hoặc "gần thời gian thực"

trái ngược với khối lượng dữ liệu lớn hơn nhiều được lưu trữ trong kho dữ liệu nói chung ít thường xuyên hơn.

Lợi ích của ODS bao gồm:

- Cung cấp dữ liệu được làm sạch nhưng vẫn giữ nguyên định dạng từ nhiều nguồn vào một chỗ duy nhất.
- Có thể được sử dụng như một vùng tạm thời cho DW
- Được thiết kế cho truy vấn đơn giản trên tập dữ liệu nhỏ trong khi DW thì dùng cho truy vấn phức tạp và tập dữ liệu lớn

2.4 Kiến trúc DataWarehouse

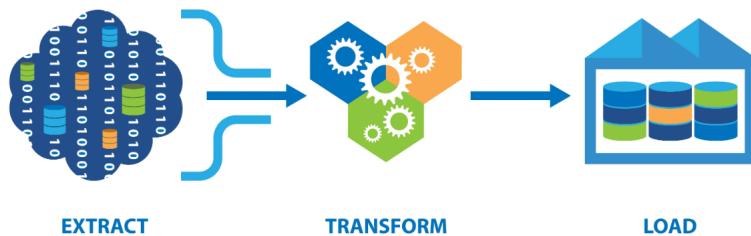


Hình 56: Kiến trúc DataWarehouse

- **Data source:** Nguồn dữ liệu sẽ được lấy Kaggle và được lưu trữ dưới dạng .csv
- **Staging:** Vùng tạm thời sẽ chứa các dữ liệu tạm thời và quy trình ETL dữ liệu. Công cụ được sử dụng chính ở đây là Python và Power Query.
- **DataWarehouse:** Kho dữ liệu sẽ chứa 3 fact và 11 dim như trên hình. Công cụ lưu trữ chính là MySQL.
- **BI:** Gồm 5 báo cáo có các tên như trên. Được trực quan hóa bằng công cụ Power BI.

2.5 Tiền xử lý ETL

Tổng quan về hoạt động ETL dữ liệu



Hình 57: Tổng quan hoạt động ETL

- Extract: Trích rút dữ liệu từ nguồn.
- Transform:
 - Xóa các trường không dùng
 - Xóa dữ liệu bị trùng
 - Xóa dữ liệu bị null
 - Chuyển đổi kiểu dữ liệu
 - Thêm trường dữ liệu
 - Xóa dữ liệu không hợp lý trong cột thêm mới
 - Gom nhóm dữ liệu
- Load: Tải dữ liệu vào trong cơ sở dữ liệu

1. Xóa các trường không dùng đến

(a) Xóa các cột trong bộ dữ liệu Geolocation

```
geolocation_df = geolocation_df.drop(  
    'geolocation_city', axis = 1)
```

Before

```
Data columns (total 5 columns):
 #   Column           Non-Null Count  Dtype  
 ---  --  
 0   geolocation_zip_code_prefix  1000163 non-null  int64  
 1   geolocation_lat            1000163 non-null  float64 
 2   geolocation_lng            1000163 non-null  float64 
 3   geolocation_city           1000163 non-null  object  
 4   geolocation_state          1000163 non-null  object  
dtypes: float64(2), int64(1), object(2)
memory usage: 38.2+ MB
```

After

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000163 entries, 0 to 1000162
Data columns (total 4 columns):
 #   Column           Non-Null Count  Dtype  
 ---  --  
 0   geolocation_zip_code_prefix  1000163 non-null  int64  
 1   geolocation_lat            1000163 non-null  float64 
 2   geolocation_lng            1000163 non-null  float64 
 3   geolocation_state          1000163 non-null  object  
dtypes: float64(2), int64(1), object(1)
memory usage: 30.5+ MB
```

Hình 58: Xóa cột geolocation_city

- (b) Xóa các cột trong các bộ dữ liệu còn lại

```
df = df.drop(['customer_city', 'review_comment_title',
              'review_comment_message', 'review_creation_date',
              'review_answer_timestamp', 'product_name_lenght',
              'product_description_lenght', 'product_photos_qty',
              'product_weight_g', 'product_length_cm',
              'product_height_cm', 'product_width_cm',
              'payment_sequential', 'seller_city'], axis = 1)
```

Before

```
Data columns (total 48 columns):
 #   Column           Non-Null Count Dtype  
 --- 
 0   customer_id      115609 non-null  object  
 1   customer_unique_id 115609 non-null  object  
 2   customer_zip_code_prefix 115609 non-null  int64  
 3   customer_city      115609 non-null  object  
 4   customer_state     115609 non-null  object  
 5   order_id          115609 non-null  object  
 6   order_status       115609 non-null  object  
 7   order_purchase_timestamp 115609 non-null  object  
 8   order_approved_at  115595 non-null  object  
 9   order_delivered_carrier_date 114414 non-null  object  
 10  order_delivered_customer_date 113209 non-null  object  
 11  order_estimated_delivery_date 115609 non-null  object  
 12  review_id         115609 non-null  object  
 13  review_score      115609 non-null  int64  
 14  review_comment_title 13801 non-null  object  
 15  review_comment_message 8606 non-null  object  
 16  review_creation_date 115609 non-null  object  
 17  review_answer_timestamp 115609 non-null  object  
 18  product_id        115609 non-null  int64  
 19  product_item_id   115609 non-null  object  
 20  seller_id         115609 non-null  object  
 21  shipping_limit_date 115609 non-null  object  
 22  price             115609 non-null  float64 
 23  freight_value    115609 non-null  float64 
 24  product_category_name 115609 non-null  object  
 25  product_name_length 115609 non-null  float64 
 26  product_description_lenght 115609 non-null  float64 
 27  product_photos_qty 115609 non-null  float64 
 28  product_weight_g   115609 non-null  float64 
 29  product_length_cm 115609 non-null  float64 
 30  product_height_cm 115609 non-null  float64 
 31  product_width_cm  115609 non-null  float64 
 32  payment_sequential 115609 non-null  int64  
 33  payment_type       115609 non-null  object  
 34  payment_installments 115609 non-null  int64  
 35  payment_value      115609 non-null  float64 
 36  seller_zip_code_prefix 115609 non-null  int64  
 37  seller_city         115609 non-null  object  
 38  seller_state        115609 non-null  object  
 39  product_category_name_english 115609 non-null  object  
dtypes: float64(18), Int64(6), object(24)
memory usage: 36.2+ MB
```

```
df.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 115608 entries, 0 to 115608
Data columns (total 26 columns):
 #   Column           Non-Null Count Dtype  
 --- 
 0   customer_id      115609 non-null  object  
 1   customer_unique_id 115609 non-null  object  
 2   customer_zip_code_prefix 115609 non-null  int64  
 3   customer_state     115609 non-null  object  
 4   order_id          115609 non-null  object  
 5   order_status       115609 non-null  object  
 6   order_purchase_timestamp 115609 non-null  object  
 7   order_approved_at  115595 non-null  object  
 8   order_delivered_carrier_date 114414 non-null  object  
 9   order_delivered_customer_date 113209 non-null  object  
 10  order_estimated_delivery_date 115609 non-null  object  
 11  review_id         115609 non-null  object  
 12  review_score      115609 non-null  int64  
 13  order_item_id     115609 non-null  object  
 14  product_id        115609 non-null  object  
 15  seller_id         115609 non-null  object  
 16  shipping_limit_date 115609 non-null  object  
 17  price             115609 non-null  float64 
 18  freight_value    115609 non-null  float64 
 19  product_category_name 115609 non-null  object  
 20  payment_type       115609 non-null  object  
 21  payment_installments 115609 non-null  int64  
 22  payment_value      115609 non-null  float64 
 23  seller_zip_code_prefix 115609 non-null  int64  
 24  seller_state        115609 non-null  object  
 25  product_category_name_english 115609 non-null  object  
dtypes: float64(18), Int64(6), object(18)
memory usage: 27.8+ MB
```

Hình 59: Xóa các cột trong bộ dữ liệu còn lại

2. Xóa dữ liệu bị trùng

(a) Xóa dữ liệu bị trùng trong bảng Geolocation

```
geolocation_df = geolocation_df.drop_duplicates(
    subset = ['geolocation_zip_code_prefix',
              'geolocation_state'])
```

Before

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000163 entries, 0 to 1000162
Data columns (total 4 columns):
 #   Column           Non-Null Count Dtype  
 --- 
 0   geolocation_zip_code_prefix 1000163 non-null  int64  
 1   geolocation_lat            1000163 non-null  float64 
 2   geolocation_lng            1000163 non-null  float64 
 3   geolocation_state          1000163 non-null  object  
dtypes: float64(2), int64(1), object(1)
memory usage: 30.5+ MB
```

After

```
Data columns (total 4 columns):
 #   Column           Non-Null Count Dtype  
 --- 
 0   geolocation_zip_code_prefix 19023 non-null  int64  
 1   geolocation_lat            19023 non-null  float64 
 2   geolocation_lng            19023 non-null  float64 
 3   geolocation_state          19023 non-null  object  
dtypes: float64(2), int64(1), object(1)
memory usage: 743.1+ KB
```

Hình 60: Xóa dữ liệu bị trùng trong bảng Geolocation

3. Xóa dữ liệu bị null

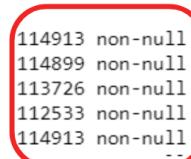
(a) Xóa dữ liệu trong bảng Orders

Trong bảng Orders có các giá trị ngày bị null do có những đơn hàng bị hủy, thất lạc,...

```
orders_df = orders_df.dropna()
```

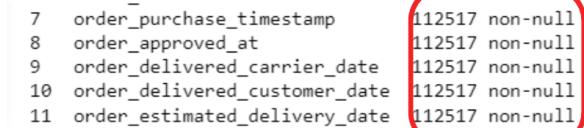
Before

```
7   order_purchase_timestamp      114913 non-null object  
8   order_approved_at           114899 non-null object  
9   order_delivered_carrier_date 113726 non-null object  
10  order_delivered_customer_date 112533 non-null object  
11  order_estimated_delivery_date 114913 non-null object
```



After

```
7   order_purchase_timestamp      112517 non-null object  
8   order_approved_at           112517 non-null object  
9   order_delivered_carrier_date 112517 non-null object  
10  order_delivered_customer_date 112517 non-null object  
11  order_estimated_delivery_date 112517 non-null object
```



Hình 61: Xóa dữ liệu bị trùng trong bảng Orders

(b) Xóa dữ liệu trong bảng Payment

Trong bảng Payment có các giá trị phương thức thanh toán là 'not_defined', xóa các giá trị này.

```
payments_df[payments_df.payment_type == 'not_defined'].index  
payments_df = payments_df.drop(index  
                               =[51280,57411,94427]).reset_index()
```

Before

```
credit_card    76795  
boleto        19784  
voucher        5775  
debit_card     1529  
not_defined      3  
Name: payment_type, dtype: int64
```



After

```
credit_card    76795  
boleto        19784  
voucher        5775  
debit_card     1529  
Name: payment_type, dtype: int64
```

Hình 62: Xóa dữ liệu trong bảng Payment

4. Chuyển đổi kiểu dữ liệu

(a) Chuyển đổi kiểu dữ liệu ngày tháng trong bảng Orders

```
df['order_purchase_timestamp'] = pd.to_datetime(  
    df['order_purchase_timestamp'])  
df['order_delivered_customer_date'] = pd.to_datetime(  
    df['order_delivered_customer_date'])  
df['order_estimated_delivery_date'] = pd.to_datetime(  
    df['order_estimated_delivery_date'])  
df['shipping_limit_date'] = pd.to_datetime(  
    df['shipping_limit_date'])  
df['order_delivered_carrier_date'] = pd.to_datetime(  
    df['order_delivered_carrier_date'])
```

Before

7	order_purchase_timestamp	112517	non-null	object
8	order_approved_at	112517	non-null	object
9	order_delivered_carrier_date	112517	non-null	object
10	order_delivered_customer_date	112517	non-null	object
11	order_estimated_delivery_date	112517	non-null	object
20	shipping_limit_date	112517	non-null	object



After

7	order_purchase_timestamp	114914	non-null	datetime64[ns]
8	order_approved_at	114900	non-null	object
9	order_delivered_carrier_date	113727	non-null	datetime64[ns]
10	order_delivered_customer_date	112534	non-null	datetime64[ns]
11	order_estimated_delivery_date	114914	non-null	datetime64[ns]
22	shipping_limit_date	114914	non-null	datetime64[ns]

Hình 63: Chuyển đổi kiểu dữ liệu trong bảng Orders

5. Thêm trường dữ liệu

(a) Thêm trường dữ liệu Shipping_days

```
df['shipping_days'] = (df['order_delivered_customer_date'].dt.date  
    - df['order_delivered_carrier_date'].dt.date).dt.days
```

Before

```
40 payment_method           112517 non-null int64
41 product_category          112517 non-null object
dtypes: datetime64[ns](5), float64(7), int32(2), int64(10), object(18)
memory usage: 36.1+ MB
```



After

```
40 payment_method           112517 non-null int64
41 product_category          112517 non-null object
42 shipping_days             112517 non-null int64
dtypes: datetime64[ns](5), float64(7), int32(2), int64(11), object(18)
memory usage: 36.9+ MB
```

Hình 64: Thêm trường dữ liệu shipping_days

6. Xóa dữ liệu không hợp lý trong cột thêm mới

- (a) Xóa dữ liệu có shipping_days < 0.

```
df.drop((df[['order_delivered_carrier_date',
              'order_delivered_customer_date']]
         [df.shipping_days < 0]).index, inplace= True)
```

Before

```
41 product_category          112517 non-null object
42 shipping_days              112517 non-null int64
dtypes: datetime64[ns](5), float64(7), int32(2), int64(11), object(18)
memory usage: 36.9+ MB
```



After

```
41 product_category          112464 non-null object
42 shipping_days              112464 non-null int64
dtypes: datetime64[ns](5), float64(7), int32(2), int64(11), object(18)
memory usage: 36.9+ MB
```

Hình 65: Xóa dữ liệu có shipping_days < 0

7. Gom nhóm dữ liệu

- (a) Gom nhóm dữ liệu hình thức thanh toán

Với hình thức thanh toán bằng Voucher thì sẽ gán giá trị là 0, các hình thức khác ta sẽ gán giá trị là 1. Sẽ thực hiện tính tổng số lần thanh toán. Từ đó sẽ biết được khách hàng lựa chọn thanh toán trả góp hay thanh toán 1 lần.

```

def std(x):
    if x == 'voucher':
        return 0
    else:
        return 1
payments_df['std_payment'] = payments_df.payment_type.apply(std)
def method(x):
    if x > 1:
        return 1
    else:
        return 0
payment_df_temp['payment_method']=payment_df_temp.payment_count_sum.apply(method)

```

Before

```

Data columns (total 2 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   order_id         99437 non-null   object  
 1   payment_count_sum 99437 non-null   int64  
dtypes: int64(1), object(1) 
memory usage: 1.5+ MB

```



After

#	Column	Non-Null Count	Dtype	
0	order_id	99437	non-null	object
1	payment_count_sum	99437	non-null	int64
2	payment_method	99437	non-null	int64

dtypes: int64(2), object(1)
memory usage: 2.3+ MB

Hình 66: Góm nhóm dữ liệu hình thức thanh toán

(b) Gom nhóm dữ liệu trạng thái giao hàng

Nếu như số ngày giao hàng sớm /đúng so với ngày dự kiến thì trạng thái sẽ là 'OnTime/Early'. Ngược lại sẽ là 'Late'

```
df['arrival_status'] = (df['order_estimated_delivery_date'].dt.date -  
                         df['order_delivered_customer_date'].dt.date).dt.days  
  
df['arrival_status'] = df['arrival_status'].apply(  
    lambda x : 'OnTime/Early' if x >=0 else 'Late')
```

Before

```
41 product_category          112464 non-null object  
42 shipping_days             112464 non-null int64  
dtypes: datetime64[ns](5), float64(7), int32(2), int64(11), object(18)  
memory usage: 36.9+ MB
```



After

```
44 arrival_status           112464 non-null object  
45 arrival_status_id       112464 non-null int64  
dtypes: datetime64[ns](5), float64(7), int32(4), int64(12), object(20)  
memory usage: 40.3+ MB
```

Hình 67: Gom nhóm dữ liệu trạng thái giao hàng

(c) Gom nhóm dữ liệu loại mặt hàng

```

def classify_cat(x):
    if x in ['office_furniture', 'furniture_decor',
        'furniture_living_room', 'bed_bath_table',
        'kitchen_dining_laundry_garden_furniture',
        'garden_tools', 'home_comfort', 'home_comfort_2',
        'home_construction', 'furniture_bedroom',
        'furniture_mattress_and_upholstery']:
        return 'Furniture'
    elif x in ['auto', 'computers_accessories',
        'musical_instruments', 'consoles_games', 'watches_gifts',
        'air_conditioning', 'telephony', 'electronics',
        'fixed_telephony', 'tablets_printing_image', 'computers',
        'small_appliances_home_oven_and_coffee', 'small_appliances',
        'audio', 'signaling_and_security', 'security_and_services']:
        return 'Electronics'
    elif x in ['fashio_female_clothing', 'fashion_male_clothing',
        'fashion_bags_accessories', 'fashion_shoes', 'fashion_sport',
        'fashion_underwear_beach', 'fashion_childrens_clothes',
        'baby', 'cool_stuff', ]:
        return 'Fashion'
    elif x in ['housewares', 'home_comfort', 'home_appliances',
        'home_appliances_2', 'flowers', 'construction_tools_garden',
        'garden_tools', 'construction_tools_lights',
        'construction_tools_tools', 'luggage_accessories',
        'la_cuisine', 'pet_shop', 'market_place']:
        return 'Home & Garden'
    elif x in ['sports_leisure', 'toys', 'cds_dvds_musicals',
        'music', 'dvds_blu_ray', 'cine_photo', 'party_supplies',
        'christmas_supplies', 'arts_and_craftmanship', 'art']:
        return 'Entertainment'
    elif x in ['health_beauty', 'perfumery', 'diapers_and_hygiene']:
        return 'Beauty & Health'
    elif x in ['food_drink', 'drinks', 'food']:
        return 'Food & Drinks'
    elif x in ['books_general_interest', 'books_technical',
        'books_imported', 'stationery']:
        return 'Books & Stationery'
    elif x in ['construction_tools_construction',
        'construction_tools_safety', 'industry_commerce_and_business',
        'agro_industry_and_commerce']:
        return 'Industry & Construction'

df['product_category'] =
    df.product_category_name_english.apply(classify_cat)

```



Before					
38	product_category_name_english	112517	non-null	object	
39	payment_count_sum	112517	non-null	int64	
40	payment_method	112517	non-null	int64	
	dtypes:	float64(7), int32(2), int64(10), object(22)			
	memory usage:	35.2+ MB			

After					
38	product_category_name_english	112517	non-null	object	
39	payment_count_sum	112517	non-null	int64	
40	payment_method	112517	non-null	int64	
41	product_category	112517	non-null	object	
	dtypes:	float64(7), int32(2), int64(10), object(23)			
	memory usage:	36.1+ MB			

Hình 68: Gom nhóm dữ liệu loại mặt hàng

(d) Gom nhóm dữ liệu phân khúc khách hàng

Ở đây chúng ta sẽ nhóm các phân khúc khách hàng theo giá trị thanh toán. Sẽ nhóm thành 4 phân khúc: '0-500', '500-2000', '2000-10000', '> 10000'

```
my_bin = [0, 500, 2000, 10000, dim_segment.payment_value.max()]
my_label = ['0-500', '500-2000', '2000-10000', '> 10000']
segment['cluster']= pd.cut(segment.payment_value,
                           bins = my_bin, labels= my_label)
```

Before

payment_value
146.87
275.79
275.79
140.61
137.58
...
232.19
426.70
160.46
55.18
100.00

After

payment_value	cluster
114.74	0-500
67.41	0-500
195.42	0-500
179.35	0-500
107.01	0-500
...	...
91.91	0-500
81.36	0-500
63.13	0-500
214.13	0-500
91.00	0-500

Hình 69: Gom nhóm dữ liệu trạng thái giao hàng

(e) Gom nhóm dữ liệu khu vực

Ở đây chúng ta có những liệu 27 bang. Để thuận tiện cho việc phân tích thì chúng ta sẽ nhóm thành 5 khu vực: North, NorthEast, South, SouthEast, CenterWest.

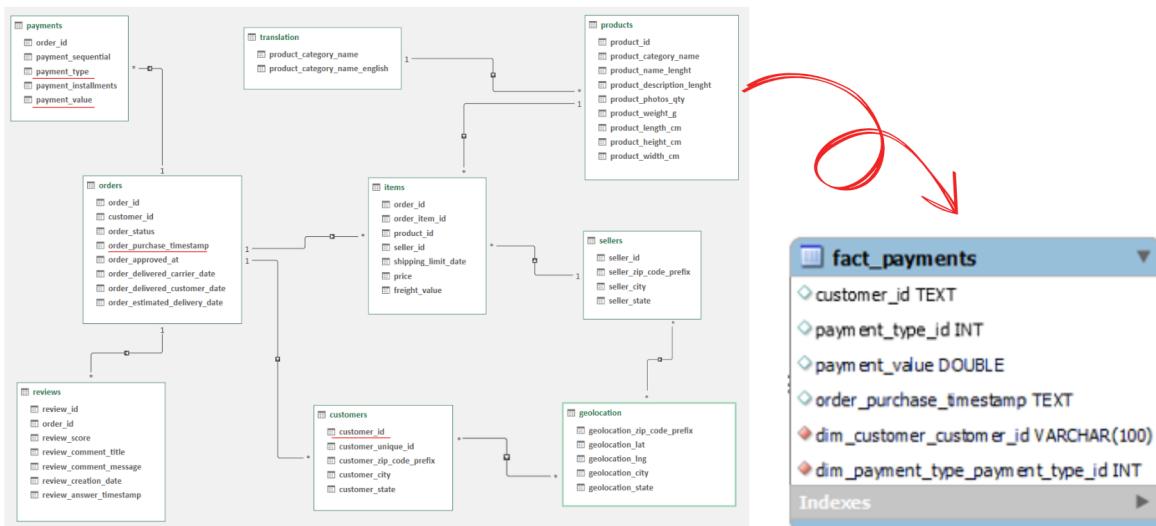
geolocation_state	state	region	geolocation_state	state	region
AC	Acre	North	PB	Paraíba	Northeast
AL	Alagoas	Northeast	PR	Paraná	South
AP	Amapá	North	PE	Pernambuco	Northeast
AM	Amazonas	North	PI	Plaí	Northeast
BA	Bahia	Northeast	RJ	Rio de Janeiro	Southeast
CE	Ceará	Northeast	RN	Rio Grande do Norte	Northeast
DF	Distrito Federal	Center West	RS	Rio Grande do Sul	South
ES	Espírito Santo	Southeast	RO	Rondônia	North
GO	Goiás	Center West	RR	Roraima	North
MA	Maranhão	Northeast	SC	Santa Catarina	South
MT	MatoGrosso	Center West	SP	São Paulo	Southeast
MS	MatoGrosso do Sul	Center West	SE	Sergipe	Northeast
MG	Minas Gerais	Southeast	TO	Tocantins	North
PA	Pará	North			

Hình 70: Gom nhóm dữ liệu khu vực

8. Quy trình chuyển đổi từ OLTP sang OLAP

(a) fact_payments

```
fact_payments =
pd.DataFrame(df[['customer_id',
'payment_type_id','payment_value',]])
```



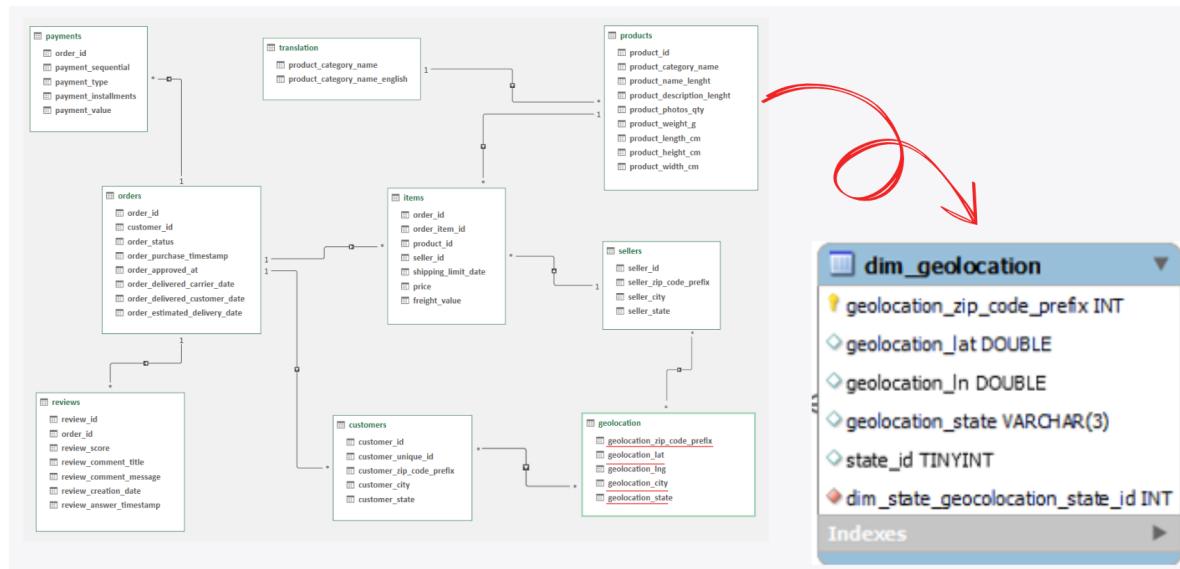
Hình 71: Quy trình chuyển đổi fact_payments

(b) dim_geolocation

```

dim_geolocation = geolocation_df[['geolocation_zip_code_prefix',
                                   'geolocation_lat', 'geolocation_lng', 'state_id']]
dim_geolocation = pd.DataFrame(dim_geolocation)

```



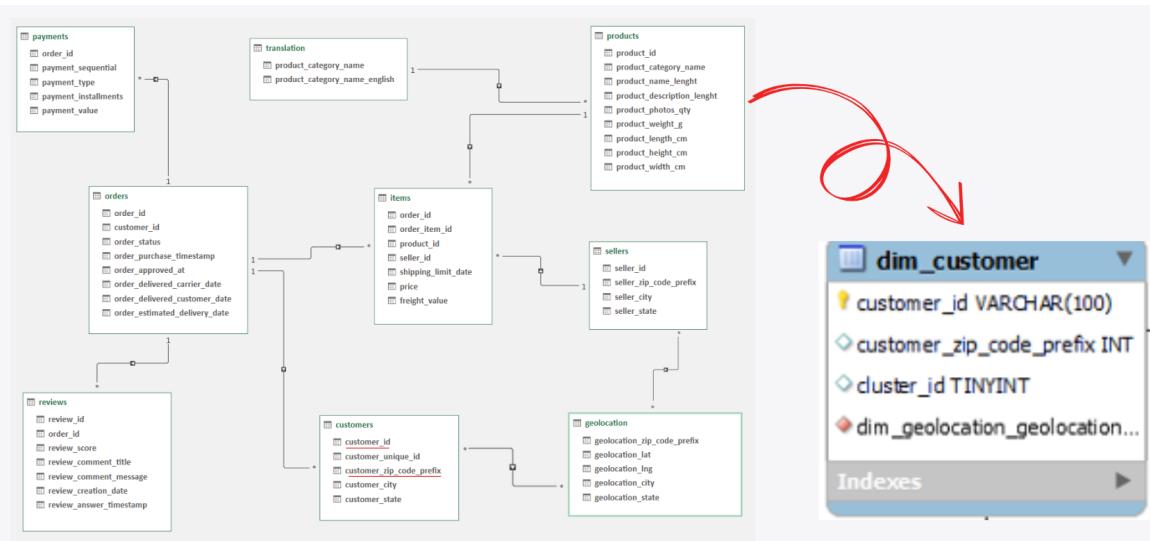
Hình 72: Quy trình chuyển đổi dim_geolocation

(c) dim_customer

```

dim_customer =
    customers_df[['customer_id', 'customer_zip_code_prefix']]
dim_customer = pd.DataFrame(dim_customer).drop_duplicates()

```



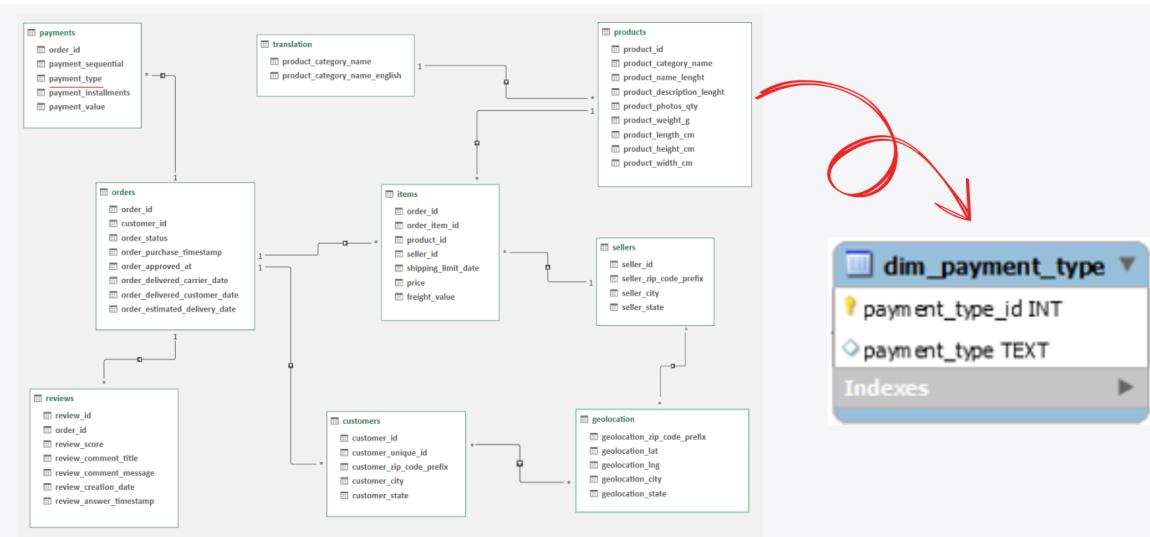
Hình 73: Quy trình chuyển đổi dim_customer

(d) dim_payment_type

```

dim_payment_type = payments_df[['payment_type_id', 'payment_type']]
dim_payment_type = pd.DataFrame(dim_payment_type).drop_duplicates()

```



Hình 74: Quy trình chuyển đổi dim_payment_type

2.6 Dimension

Sau quá trình tiền xử lý, từ mô hình dữ liệu OLTP ban đầu ta thu được 12 bảng dim là các chiều mà ta sẽ dùng đến trong quá trình phân tích.

12 bảng dim gồm có:

- dim_geolocation: chứa thông tin về khu vực địa lý - tiểu khu, trong đó có khóa chính là mã tiểu khu, cùng với các trường kinh độ, vĩ độ, mã bang
- dim_state: chứa thông tin về các bang, gồm có mã bang, tên bang, tên viết tắt và mã khu vực; một bang gồm nhiều tiểu khu ở trong bảng dim_geolocation
- dim_region: gom nhóm dim_state và chia ra thành 5 vùng tương ứng với 5 khu vực địa lý của Brazil, gồm các thông tin mã khu vực và tên khu vực
- dim_seller: thông tin về người bán hàng tham gia sàn thương mại điện tử, gồm các trường mã người bán và mã tiểu khu
- dim_customer: thông tin khách hàng tham gia vào nền tảng thương mại điện tử, gồm có mã khách hàng, mã tiểu khu và mã phân khúc khách hàng
- dim_product_category: thông tin về các loại sản phẩm được giao dịch, gồm có mã loại sản phẩm, tên tiếng Anh và mã ngành hàng
- dim_product_industry: gom nhóm các loại sản phẩm thành 9 ngành hàng lớn, gồm các trường mã ngành hàng và tên ngành hàng
- dim_payment_type: thông tin về các phương thức được sử dụng để thanh toán, gồm mã và tên phương thức
- dim_payment_method: thông tin về hình thức thanh toán Trả một lần hay Trả góp, gồm mã và tên hình thức
- dim_arrival_status: thông tin về trạng thái giao vận là Giao đúng hạn hay Giao muộn, gồm mã trạng thái và trạng thái
- dim_segment: thông tin phân khúc khách hàng, gồm mã phân khúc và phân khúc
- dim_date: thông tin ngày tháng được lấy bằng Min và Max của thời gian đặt hàng lấy từ bảng đơn hàng, phân cấp thành các trường ngày, tháng, năm.

Các bảng dim và giá trị của nó được thể hiện như sau:

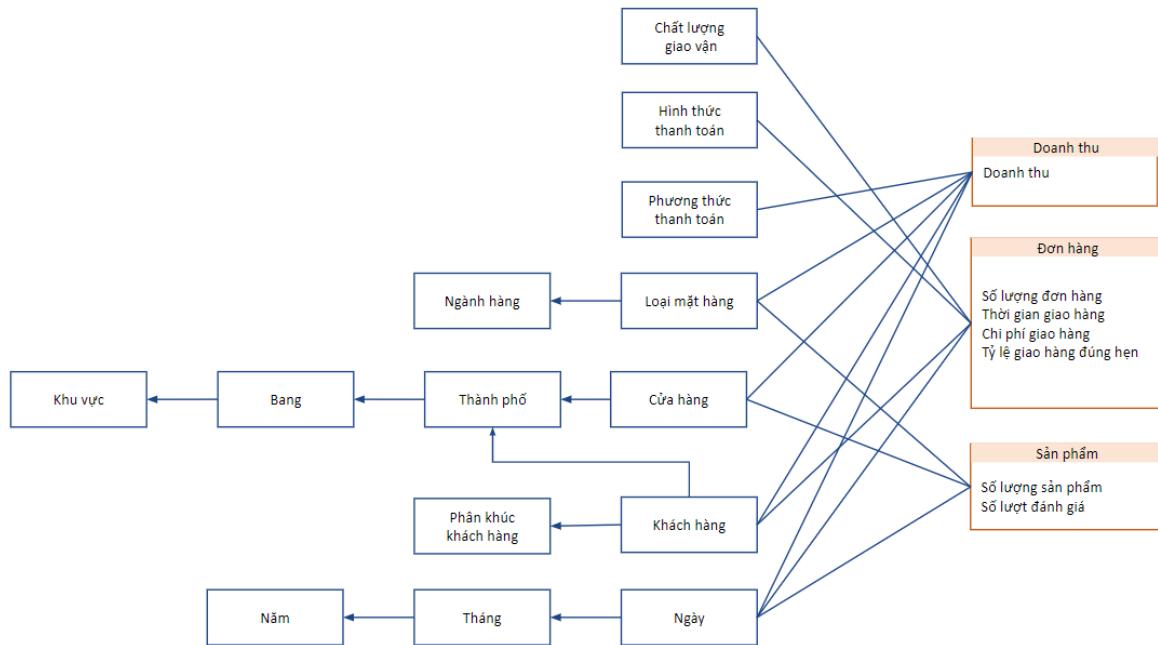
The screenshot displays several dimension tables with their cardinalities (number of values) indicated by callout boxes:

- Region:** 5 giá trị (values: Center West, North, North East, South, South East)
- Segment:** 4 giá trị (values: 0-500, 500-2000, 2000-10000, > 10000)
- Payment Method:** 2 giá trị (values: Installment, One-time Payment)
- Payment Type:** 27 giá trị (values: credit_card, boleto, voucher, debit_card)
- Arrival Status:** 2 giá trị (values: OnTime/Early, Late)
- Product Industry:** 9 giá trị (values: Furniture, Industry & Construction, Entertainment, Electronics, Home & Garden, Beauty & Health, Fashion, Books & Stationery, Food & Drinks)
- State:** 27 giá trị (values: Acre, Alagoas, Amapa, Amazonas, Bahia, Ceara, Distrito Federal, Espírito Santo, Goias, Maranhao, Mato Grosso, Mato Grosso do Sul, Minas Gerais, Para)
- Geolocation Zip Code Prefix:** 19023 giá trị (values: 1001 to 1014)
- Customer ID:** 99441 giá trị (values: 00012a2c... to 00072d033fe2e59061ae5c3aff1a2be5)
- Seller ID:** 3095 giá trị (values: 0015a82c... to 011b0eaba87386a2ae96a7d32bb531d1)
- Product Category:** 71 giá trị (values: agro_industry_and_commerce, air_conditioning, art, arts_and_craftsmanship, audio, auto, baby, bed_bath_table, books_general_interest, books_imported, books_technical, cds_dvds_musicals, christmas_supplies, cine_photo)
- Year:** 3 giá trị (values: 2016, 2017, 2018)
- Month:** 12 giá trị (values: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)
- Day:** 31 giá trị (values: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31)

Hình 75: Các bảng dim và giá trị

2.7 Mô hình dữ liệu OLAP

Ta xác định mô hình dữ liệu logic như sau:



Hình 76: Mô hình dữ liệu logic OLAP

Dưới góc nhìn logic ta có 3 đối tượng chính cần quan tâm:

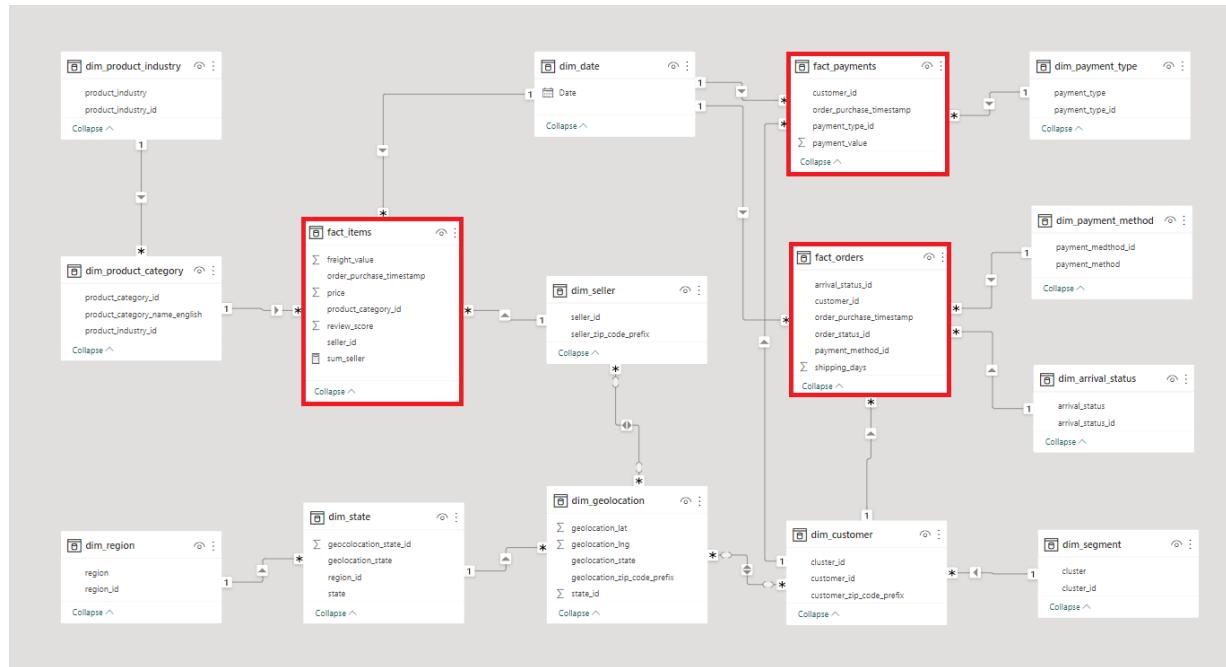
- Doanh thu, với chỉ số chính là doanh thu
- Đơn hàng, với các chỉ số là: số lượng đơn hàng, thời gian giao hàng, chi phí giao hàng và tỷ lệ giao hàng đúng hạn
- Sản phẩm, với các chỉ số: số lượng sản phẩm, số điểm đánh giá

Với các đối tượng này và các chỉ số của nó, các khía cạnh được xem xét đến là:

- Thời gian: chỉ số và sự biến đổi của nó theo ngày, tháng, năm
- Cửa hàng và khách hàng, từ đó gom nhóm lại để đánh giá theo khu vực địa lý là bang, khu vực
- Ngoài ra, khi xem xét dưới khía cạnh khách hàng, ta đánh giá thêm theo phân khúc khách hàng
- Loại mặt hàng và tổng quan hơn là ngành hàng
- Phương thức thanh toán
- Hình thức thanh toán: Trả một lần hay Trả góp
- Chất lượng giao vận

Từ mô hình logic, ta đưa dữ liệu về mô hình vật lý để lưu trữ trong cơ sở dữ liệu và sử dụng để phân tích.

Mô hình dữ liệu vật lý như sau:



Hình 77: Mô hình dữ liệu vật lý OLAP

Mô hình dữ liệu gồm 12 bảng dim và 3 bảng fact. 12 bảng dim trong mô hình tương ứng với 12 dim đã nêu ở phần trước. 3 bảng fact gồm có:

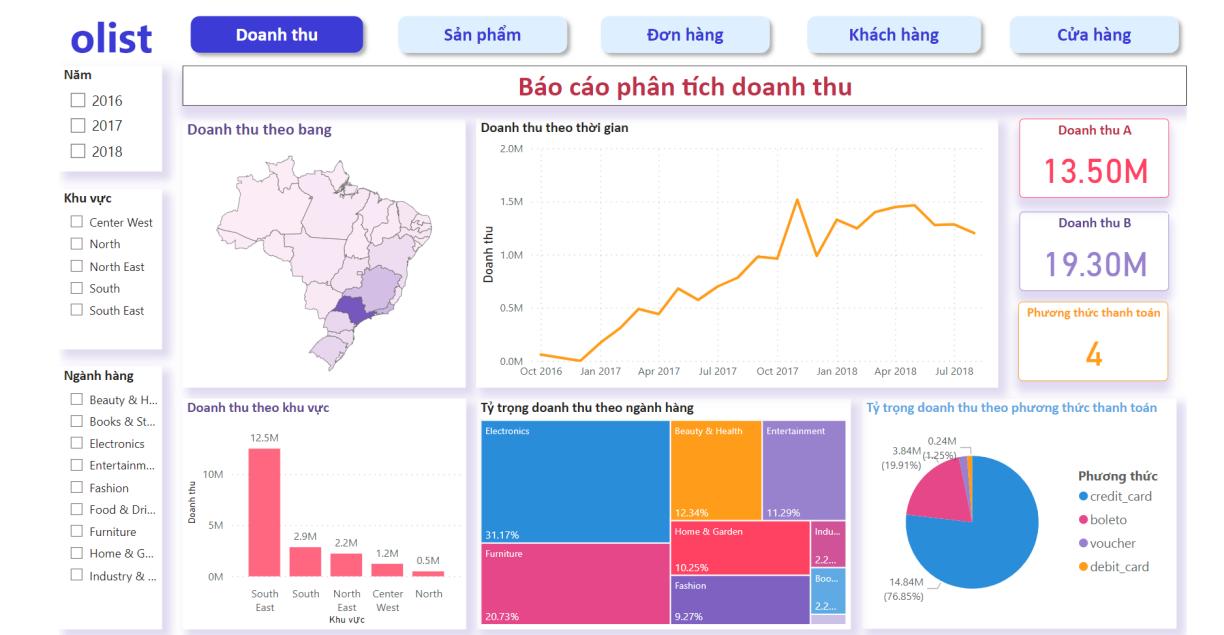
- fact_orders: chứa thông tin về đơn hàng, gồm mã khách hàng, mã trạng thái giao hàng, mã phương thức thanh toán, mốc thời gian đặt hàng và thời gian vận chuyển tính theo ngày
- fact_items: thông tin chi tiết về sản phẩm trong đơn hàng, gồm mã người bán, mã loại mặt hàng, thời gian đặt hàng, giá mặt hàng, chi phí vận chuyển và điểm đánh giá
- fact_payment: thông tin về giao dịch, gồm mã khách hàng, mã phương thức giao dịch, giá trị giao dịch và mốc thời gian.

3 Xây dựng báo cáo trực quan

Dựa trên nhu cầu của doanh nghiệp và quy mô bộ dữ liệu, chúng ta xây dựng được báo cáo trực quan về 5 chủ đề sau đây:

- Báo cáo về doanh thu
- Báo cáo về sản phẩm
- Báo cáo về đơn hàng
- Báo cáo về khách hàng
- Báo cáo về cửa hàng

3.1 Báo cáo về doanh thu



Hình 78: Báo cáo về doanh thu

Theo thông tin từ bộ dữ liệu đã thu thập, ta có thể tính doanh thu theo 2 cách, được nêu ra theo 2 định nghĩa:

- Doanh thu A: Tính bằng tổng số tiền của các sản phẩm được ghi lại trong đơn hàng
- Doanh thu B: Tính bằng tổng giá trị thanh toán đã nhận được, được ghi lại trong thông tin thanh toán

Có sự chênh lệch giữa doanh thu A và doanh thu B ta nhìn thấy trên dashboard, doanh thu B nhiều hơn doanh thu A khoảng 6 triệu Real - đơn vị tiền tệ của Brasil. Nguyên nhân của

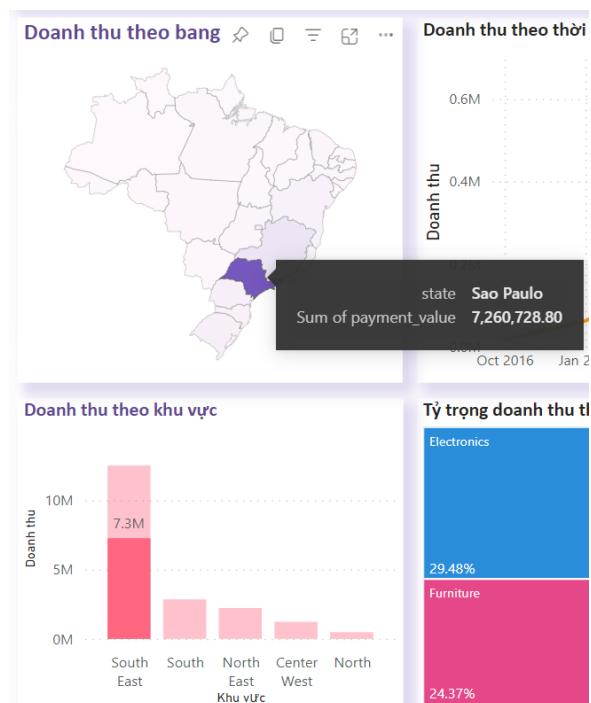
sự chênh lệch này là do doanh thu A chỉ bao gồm chi phí đơn thuần của sản phẩm, trong khi doanh thu B tính cả phí vận chuyển, phí dịch vụ và các chi phí phát sinh khác trong quá trình giao dịch.

Trong báo cáo này, phụ thuộc vào thực tế bộ dữ liệu được ghi lại, doanh thu theo thời gian, khu vực địa lý và theo phương thức thanh toán được tính dựa theo doanh thu B, doanh thu chia theo ngành hàng được tính theo doanh thu A.

Doanh thu theo thời gian là một đường gấp khúc. Tuy có những thời điểm lên xuồng nhưng nhìn tổng quan, ta thấy doanh thu của doanh nghiệp có xu hướng tăng lên theo thời gian. Thời gian đầu khi mới thành lập vào cuối năm 2016, doanh thu của doanh nghiệp là khá thấp và gần như không đáng kể. Khúc đi lên dốc nhất ứng với thời điểm tháng 11 (năm 2017) với doanh thu trong tháng vào khoảng hơn 1.5 triệu Real. Thông tin này cho thấy doanh thu mua sắm tăng vọt vào thời điểm này. Đây là một xu hướng khá phù hợp với thị trường chung trên toàn thế giới bởi trong tháng 11 có sự kiện Black Friday - ngày hội mua sắm được hưởng ứng với nhiều ưu đãi thu hút người mua, đẩy mạnh thị trường tiêu dùng. Sau thời điểm này, vào tháng sau đó doanh thu quay trở lại mức trước đó và dao động không nhiều, tuy vẫn giữ được xu hướng tăng lên.

Nhìn vào bản đồ phân bố doanh thu theo bang, ta thấy có sự khác biệt rõ rệt về màu sắc ở hai khu vực: miền Bắc có màu hồng thể hiện đây là khu vực có doanh thu thấp, vùng Đông Nam có tone màu tím thể hiện là khu vực này có doanh thu cao, trong đó vùng đậm nhất là khu vực São Paulo - trung tâm tài chính sôi động, một trong những thành phố nổi tiếng nhất của Brazil.

Qua đó cho thấy hoạt động mua bán qua sàn thương mại điện tử này đạt được doanh thu chủ yếu từ khu vực Đông Nam đất nước Brazil, là vùng đồng bằng và cao nguyên thấp ven biển, tập trung đông đúc dân cư và có nền kinh tế phát triển. Tổng doanh thu của cả bang São Paulo là 7 260 728 Real. Con số này chiếm hơn 50% doanh thu của toàn khu vực Đông Nam. Ngược lại, khu vực phía Bắc là nơi có lưu vực sông Amazon chảy qua và được bao phủ bởi rừng rậm nhiệt đới, dân cư thưa thớt, có doanh thu thấp.



Biểu đồ cột thể hiện sự phân bố doanh thu chia theo 5 khu vực Bắc, Đông Bắc, Tây Trung, Nam và Đông Nam cho thấy Đông Nam là khu vực có doanh thu cao nhất, như ta đã thấy ở bản đồ phân bố doanh thu, với con số cụ thể là 12.5 triệu Real. Con số này cao vượt trội hơn hẳn so với doanh thu ở 4 khu vực còn lại. Tổng doanh thu ở khu vực phía Bắc là ít nhất với 0.5

triệu Real, một con số không đáng kể so với vùng cao nhất.

Tỷ trọng doanh thu tập trung chủ yếu ở 6 ngành hàng, trong số tất cả 9 ngành hàng được kinh doanh trên sàn thương mại điện tử này. Trong đó 2 ngành hàng chiếm tỷ trọng cao nhất là ngành hàng Điện tử, với 31.17% và ngành hàng Đồ dùng văn phòng, với tỷ trọng 20.73%. 2 ngành hàng này chiếm hơn 50% tổng doanh thu toàn sàn.

Có 4 phương thức thanh toán đã được áp dụng cho các giao dịch được thực hiện qua hình thức mua bán này, đó là: Thanh toán qua thẻ tín dụng, qua thẻ ghi nợ, sử dụng voucher và một hình thức mới được phát triển bởi Ngân hàng Nhà nước Brazil là boleto. Thanh toán qua thẻ tín dụng là hình thức có mức độ phổ biến áp đảo, chiếm tỷ trọng 76.85% tổng doanh thu. Trong số các phương thức thanh toán còn lại, boleto là một phương thức mới khá triển vọng, có khả năng dẫn tiếp cận đến nhiều khách hàng hơn với tỷ trọng là 19.91%.

3.2 Báo cáo về sản phẩm



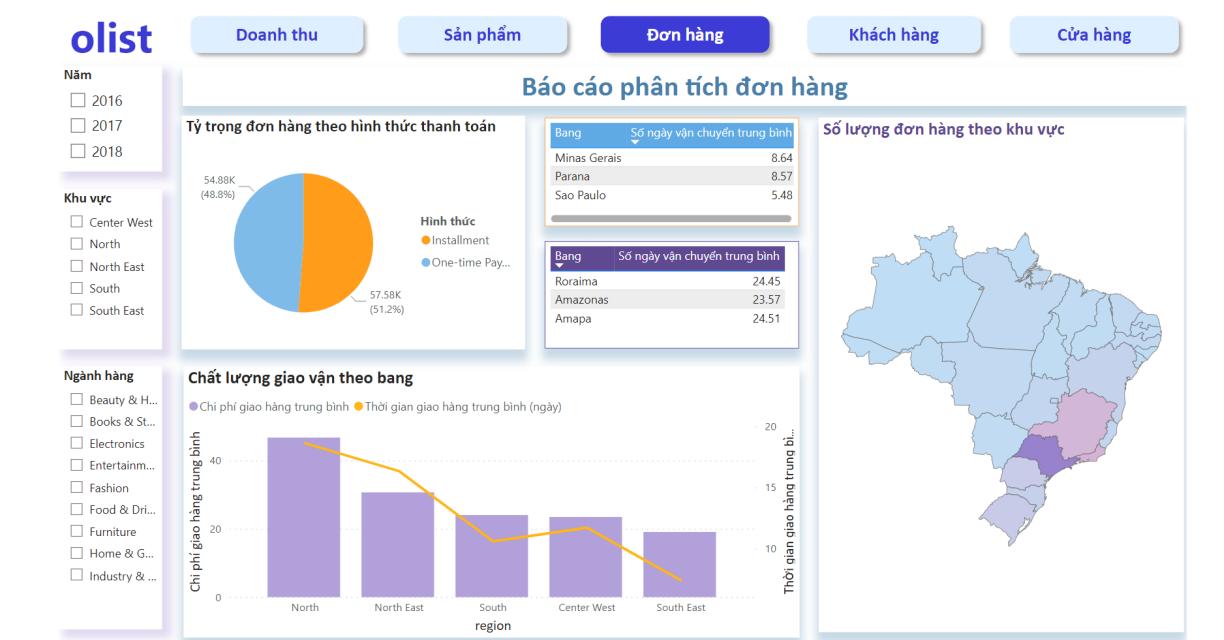
Hình 79: Báo cáo về sản phẩm

Chất lượng sản phẩm dưới góc nhìn của khách hàng được ghi nhận qua đánh giá với các mức điểm từ 1 đến 5. Qua biểu đồ vòng, ta thấy tỷ lệ đánh giá tốt (4 và 5 điểm) chiếm phần lớn số đánh giá, với 76.88%, bước đầu cho thấy người dùng khá hài lòng với chất lượng sản phẩm mua ở đây. Ở biểu đồ thanh phía dưới, ta thấy phân bố các mức điểm đánh giá này vẫn khá đồng đều ở từng ngành hàng, với mức điểm đánh giá tốt chiếm phần lớn.

Biểu đồ pareto về số sản phẩm đã bán theo từng ngành hàng cho thấy số sản phẩm đã bán nhin chung khá tương đồng với doanh thu đem lại: Hai ngành hàng Đồ điện tử và Đồ dùng văn phòng có số lượng sản phẩm bán được lớn nhất, và cũng là 2 ngành hàng duy nhất có lượng đã bán trên 20 nghìn sản phẩm.

Top 5 sản phẩm bán chạy và top 5 sản phẩm có doanh thu cao nhất tuy có chút khác biệt về thứ tự nhưng những sản phẩm có mặt trong top 5 không thay đổi. Ta thấy được những loại sản phẩm được ưa chuộng sử dụng hình thức mua bán online này gồm có: nội thất phòng ngủ, mỹ phẩm và chăm sóc sức khỏe, dụng cụ thể thao, phụ kiện máy tính và quà lưu niệm.

3.3 Báo cáo về đơn hàng



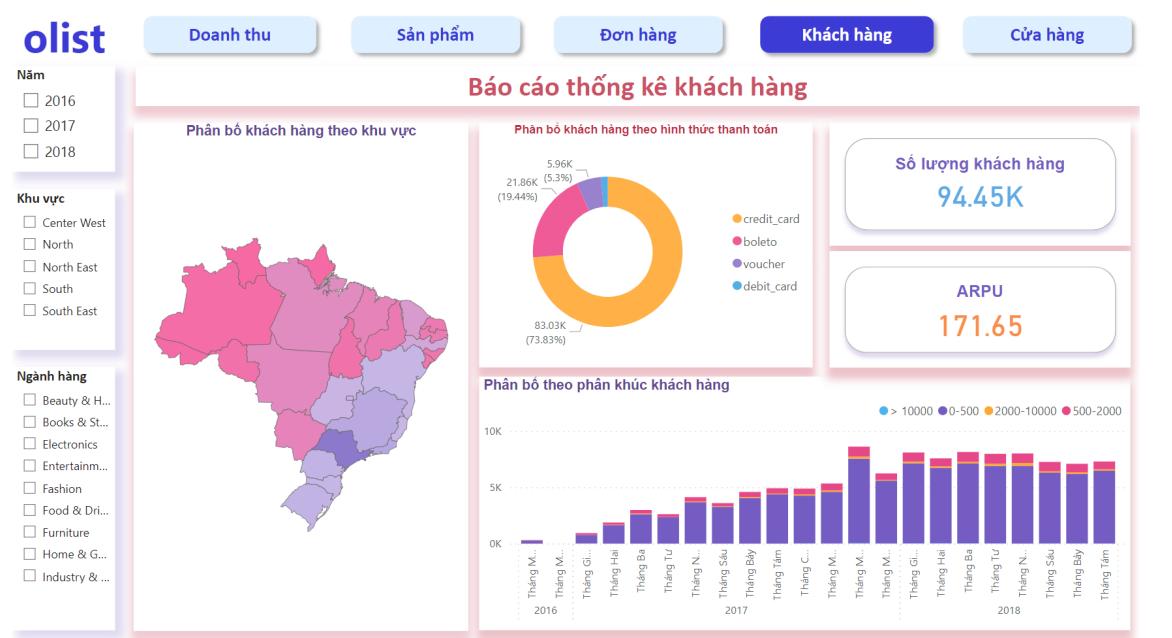
Hình 80: Báo cáo về đơn hàng

So sánh hai hình thức thanh toán Trả một lần và Trả góp, ta thấy khách hàng có xu hướng sử dụng 2 hình thức này tương đương nhau, với tỷ lệ là 51.2% đơn hàng sử dụng hình thức Trả góp và 48.8% sử dụng hình thức Trả một lần. Xu hướng này không có nhiều biến động khi xem xét theo từng khu vực.

Bản đồ phân bố số lượng đơn hàng theo khu vực có cấu trúc khá tương đồng với bản đồ phân bố doanh thu: màu nhạt ở phía Bắc, Tây Bắc và đậm hơn ở khu vực phía Nam, trong đó đáng chú ý nhất vẫn là São Paulo và khu vực các bang lân cận.

Ở các khu vực có nhiều đơn hàng này cũng có thời gian giao hàng và đặc biệt là chi phí giao hàng nhỏ hơn hẳn so với những vùng khác. Ở São Paulo, thời gian giao hàng trung bình là 7-8 ngày cho mỗi đơn hàng, với mức chi phí dao động quanh mức dưới 20 Real. Đối lập với khu vực này là khu vực miền Bắc. Thời gian giao hàng ở đây là khoảng 18 ngày, với chi phí giao hàng trung bình lên đến 46 Real. Nguyên nhân của sự đắt đỏ cả về chi phí và thời gian này, có thể kể đến do đây là khu vực rừng rậm và thượng lưu sông Amazon, địa hình nguy hiểm, không thuận lợi cho việc di chuyển và vận chuyển hàng hóa, cũng như khoảng cách từ khu vực này đến người bán, đến các kho chứa hàng có thể khá xa, do các kho chứa hàng hóa có xu hướng tập trung ở vùng trung tâm ở phía Nam và Đông Nam. Ta có thể kiểm chứng điều này ở dashboard sau khi xem xét đến chủ đề Cửa hàng.

3.4 Báo cáo về khách hàng



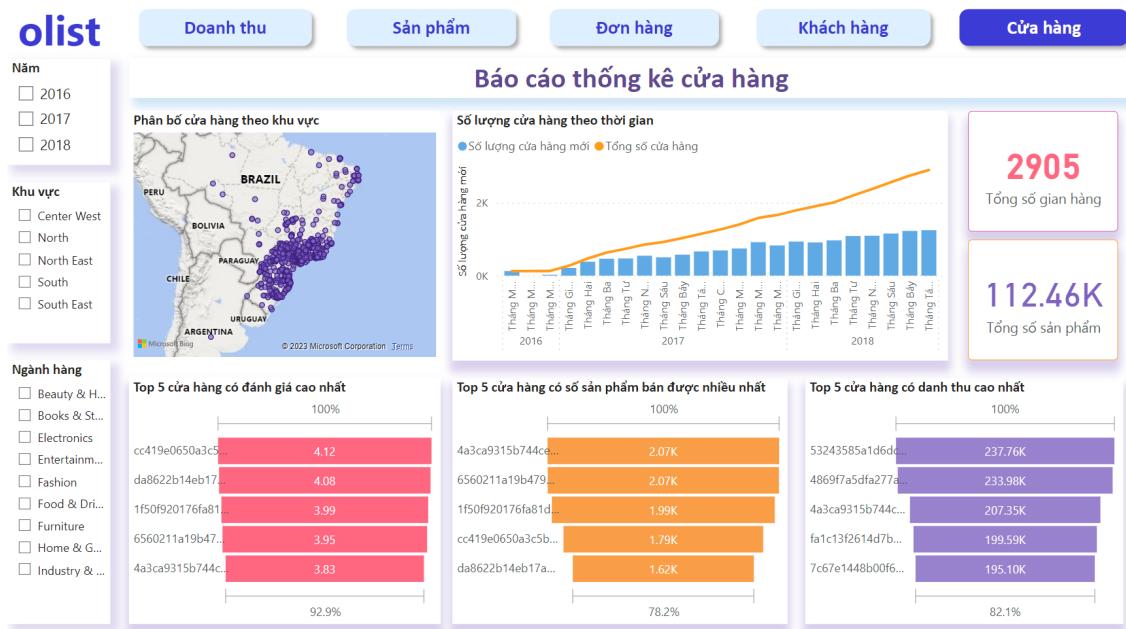
Hình 81: Báo cáo về khách hàng

Sau khoảng 2 năm đi vào hoạt động, nền tảng đã thu hút số lượng khách hàng ấn tượng: 94.450 nghìn khách hàng. Doanh thu trung bình trên một khách hàng là 171.65 Real.

Nhìn vào biểu đồ cột, số lượng khách hàng có xu hướng tăng và biến động đều theo thời gian. Sau một năm hoạt động, số lượng khách hàng ra tăng hằng tháng trên nền tảng này luôn duy trì trên mức 5 nghìn khách hàng mới/ tháng. Thời gian đầu, khách hàng mới chủ yếu nằm ở phân khúc thấp, với mức chi tiêu từ 0-500 Real. Sau thời gian khoảng nửa năm hoạt động, nền tảng dần thu hút được sự tham gia nhiều hơn từ các khách hàng ở phân khúc trung bình - cao với mức chi tiêu 500-2000 Real, và một số ít khách hàng ở phân khúc cao với mức chi tiêu 2000-10000 Real. Ngoài ra hệ thống có ghi nhận khách hàng với mức chi tiêu trên 10000 Real, tuy nhiên số lượng này là rất ít và không thể nhìn được trên biểu đồ. Biểu đồ này cùng với doanh thu trung bình trên mỗi khách hàng cho thấy phần lớn đối tượng khách hàng của doanh nghiệp này nằm ở phân khúc thấp.

Ở hai khía cạnh địa lý và phương thức thanh toán, sự phân bổ khách hàng tương đồng khá lớn đối với sự phân bổ doanh thu đã được phân tích ở báo cáo trước về chủ đề Doanh thu. Về địa lý, lượng khách hàng tiếp tục tập trung đông đúc ở khu vực phía Nam và Đông Nam, trong đó đông nhất là São Paulo và thua thớt hơn về phía Bắc. Về phương thức thanh toán, thẻ tín dụng chiếm phần lớn giao dịch với 73.83%, xếp liền sau là boleto với 19.44%

3.5 Báo cáo về cửa hàng



Hình 82: Báo cáo về cửa hàng

Bản đồ với các chấm tròn màu tím thể hiện sự phân bố các cửa hàng. Phần lớn các cửa hàng tập trung ở các khu đô thị lớn khu vực phía Nam và Đông Nam. Đây cũng là khu vực ven biển phía đông tập trung hầu hết các cảng biển thương mại lớn của đất nước này. Hàng hóa được nhập khẩu và vận chuyển về Brazil qua đường hàng hải đều tập trung ở khu vực này, do đó đây là khu vực tập trung nhiều kho bãi chứa hàng, tạo điều kiện phát triển nhiều cửa hàng, doanh nghiệp nhỏ với nhu cầu phân phối hàng hóa đến người tiêu dùng.

Việc người bán và kho hàng tập trung nhiều ở khu vực này cũng là nguyên nhân khiến cho khu vực này có ưu thế hơn hẳn về mặt vận chuyển, khiến cho chi phí và thời gian vận chuyển ở đây đều nhỏ. Điều này cũng cho thấy giả thuyết được đưa ra khi phân tích về nguyên nhân chi phí và thời gian giao hàng cao ở một số khu vực khác trong phần báo cáo về đơn hàng ở phía trên.

Số lượng người bán có xu hướng tăng đều và ổn định theo thời gian. Ngoài ra, báo cáo đưa ra top 5 cửa hàng theo từng khía cạnh: điểm đánh giá cao nhất, sản phẩm bán chạy nhất và doanh thu cao nhất. Những thông tin này có thể hữu ích khi xây dựng hệ thống gợi ý cho khách hàng khi tìm kiếm sản phẩm. Việc gợi ý những cửa hàng có hoạt động kinh doanh uy tín, chất



Hình 83: Các cảng biển lớn ở Brazil

lượng sản phẩm tốt giúp nâng cao trải nghiệm của khách hàng, từ đó nâng cao chất lượng cũng như hiệu quả kinh doanh của người bán và doanh nghiệp, đồng thời thúc đẩy các cửa hàng kinh doanh chuyên nghiệp và chất lượng.

3.6 Kết luận

Qua việc phân tích 5 báo cáo theo những chủ đề được quan tâm ở trên, ta có thể đưa ra một số kết luận về hoạt động kinh doanh của doanh nghiệp như sau:

- Các con số được đưa ra là hợp lý khi xem xét trong sự tương quan giữa các đối tượng, cũng như tương quan với tình hình thực tế về kinh tế - xã hội - địa lý của quốc gia và thế giới. Qua đó ta có thể thấy được các đối tượng (người bán - người mua - hệ thống) có sự tương tác tốt, với xu hướng tăng trưởng đều và ổn định, dự đoán doanh nghiệp sẽ phát triển ổn định trong tầm nhìn 1-2 năm tới.
- Hoạt động kinh doanh diễn ra sôi nổi và phổ biến ở khu vực miền Nam và Đông Nam, là khu vực đồng bằng và cao nguyên thấp ven biển phía đông, dân cư đông đúc, kinh tế và thương mại phát triển mạnh mẽ, tập trung nhiều thành phố lớn như São Paulo, Rio de Janeiro,... và hầu hết cảng biển. Trong khi đó ít phát triển hơn ở khu vực miền Bắc, khu vực lưu vực sông Amazon và rừng rậm nhiệt đới, ít dân cư. Điều kiện vận chuyển hạn chế và khoảng cách xa các khu đô thị trung tâm cũng khiến cho thời gian và chi phí vận chuyển ở khu vực này cao hơn hẳn so với các vùng khác.
- Các ngành hàng được ưa chuộng với hình thức mua bán này là Đồ điện tử và Đồ dùng văn phòng. Máy móc công nghiệp và Thực phẩm chiếm tỷ trọng nhỏ cả về số lượng và doanh thu, cho thấy khách hàng có vẻ e ngại khi đặt mua online những loại mặt hàng này. Nguyên nhân có thể do sự bất tiện về vận chuyển và khó khăn trong việc bảo quản lâu dài.
- Thanh toán bằng thẻ tín dụng là phương thức thanh toán được sử dụng phổ biến nhất. Phương thức mới do Ngân hàng Nhà nước Brazil phát hành là boleto có triển vọng phát triển và tiếp cận rộng rãi hơn tới khách hàng. Quá trình này có thể được thúc đẩy thông qua các chương trình quảng bá, ưu đãi và giảm giá qua phương thức thanh toán này.
- Chất lượng sản phẩm nhìn chung là tốt. Để cải thiện và nâng cao hơn nữa, có thể áp dụng những chính sách chặt chẽ hơn, đồng thời tận dụng sự trợ giúp của các hệ thống hỗ trợ đưa ra gợi ý, hệ thống kinh doanh thông minh,...

Tổng kết

Qua quá trình hoàn thiện báo cáo, chúng em rút ra được bài học tổng kết như sau:

- Việc xây dựng kho dữ liệu là vô cùng cần thiết cho mục đích phân tích dữ liệu.
- Dữ liệu trong thực tế không phải lúc nào cũng được chuẩn hóa theo một quy tắc và không dữ liệu nào là hoàn hảo. Vì vậy luôn cần phải xử lý theo nhiều cách linh hoạt nhưng chuẩn mực trước khi đưa vào hệ thống.
- Quá trình ETL dữ liệu rất quan trọng và tốn nhiều thời gian nên đòi hỏi cần phải tập trung và quan tâm thực hiện. Cần nghiên cứu bộ dữ liệu một cách tỉ mỉ để xử lý dữ liệu một cách đúng đắn để tránh xảy ra sai sót và mất dữ liệu.
- Dữ liệu có kích thước lớn gây ảnh hưởng đến thời gian chạy của máy tính, cần chọn máy có cấu hình phù hợp cho việc xử lý và phân tích. Có thể chia nhỏ các file để dễ dàng thực hiện.
- Kiến thức nghiệp vụ của mỗi quy trình là vô cùng quan trọng, vì khi hiểu được nghiệp vụ chúng ta mới có thể xây dựng được một hệ thống BI đúng và hiệu quả. Vì vậy, đầu tiên cần phải khảo sát thật kỹ nghiệp vụ của từng quy trình hoạt động.
- Mô hình dữ liệu đa chiều giúp phân tích dữ liệu trên nhiều góc nhìn khác nhau, và có thể phân cấp.
- Các dashboard nên được xây dựng theo hướng chủ đề, phải đảm bảo tính logic và thẩm mỹ.
- Quá trình làm việc nhóm để đạt được hiệu quả cần phải có sự phân công rõ ràng, giúp đỡ lẫn nhau của các thành viên trong nhóm.

Tham khảo

- 1 ThS. Nguyễn Danh Tú, *Bài giảng Kho dữ liệu và kinh doanh thông minh*, Viện Toán ứng dụng và Tin học, 2020
- 2 Kênh Youtube Học Excel cơ bản
- 3 Fanpage Facebook Phân tích dữ liệu
- 4 *Data Analysis Methods and Techniques*, The Datapine Blog
- 5 *What is a Business Analyst?*, International Institute Of Business Analysis
- 6 Lưu Đức Thắng, Ban BI, Phòng Nghiên cứu phát triển - Trung tâm phần mềm viễn thông Viettel, *Data Warehouse và Cách thiết kế Data Warehouse*, SlideShare
- 7 Business Intelligence: What It Is, How It Works, Its Importance, Examples, & Tools, Tableau.