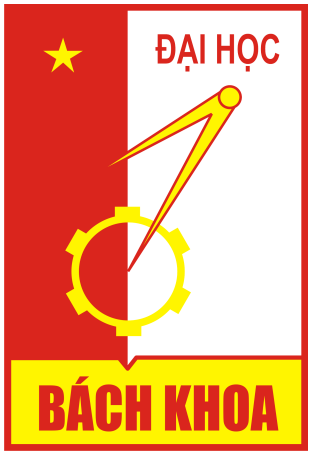
ĐẠI HỌC BÁCH KHOA HÀ NỘI
TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG



QUẢN TRỊ DỮ LIỆU VÀ TRỰC QUAN HÓA - IT5425

BÁO CÁO BÀI TẬP LỚN
ĐỀ TÀI: PHÂN TÍCH VÀ TRỰC QUAN HÓA DỮ LIỆU THƯƠNG MẠI ĐIỆN TỬ CỦA OLIST

Giảng viên hướng dẫn: **TS. Nguyễn Hữu Đức**

Nhóm: 8
Thành viên: Đào Duy Anh - 20235000
Vũ Hải Minh - 20235166
Lê Hoài Nam - 20235173
Hoàng Tấn Phúc - 20235189

Hà Nội, tháng 1 năm 2026

Mục lục

1 Tổng quan đề tài **2**
1.1 Bối cảnh . . . 2
1.2 Giới thiệu về bộ dữ liệu Olist . . . 2
1.2.1 Thông tin chung . . . 2
1.2.2 Cấu trúc . . . 2
1.3 Mục tiêu . . . 3

2 Cơ sở lý thuyết và kiến trúc hệ thống **4**
2.1 Cơ sở lý thuyết và công nghệ nền tảng . . . 4
2.1.1 Apache Spark . . . 4
2.1.2 Kiến trúc Hybrid Cloud . . . 4
2.2 Kiến trúc hệ thống . . . 4
2.2.1 Tầng nguồn dữ liệu . . . 4
2.2.2 Tầng thu thập và lưu trữ . . . 5
2.2.3 Tầng xử lý ETL . . . 5
2.2.4 Tầng kho dữ liệu . . . 5
2.2.5 Tầng trực quan hóa . . . 5

3 Xây dựng luồng xử lý dữ liệu **6**
3.1 Giai đoạn trích xuất dữ liệu (Extraction) . . . 6
3.2 Giai đoạn biến đổi dữ liệu (Transform) . . . 6
3.2.1 Làm sạch dữ liệu . . . 6
3.2.2 Làm giàu dữ liệu . . . 6
3.3 Gia đoạn nạp dữ liệu (Load) . . . 6
3.3.1 Chiến lược ghi theo lô . . . 6
3.3.2 Kiểm soát mức độ song song . . . 7
3.3.3 Đảm bảo tính nhất quán dữ liệu . . . 7

4 Kết quả trực quan hoá **8**
4.1 Trang 1: Executive Overview . . . 8
4.2 Trang 2: Product & Seller . . . 9
4.3 Trang 3: Logistics & Customer . . . 9

5 Kết luận **11**
5.1 Kết luận chung . . . 11
5.2 Bài học kinh nghiệm . . . 11

Lời cảm ơn **12**

1 Tổng quan đề tài

1.1 Bối cảnh

Trong kỷ nguyên số, thương mại điện tử đã trở thành một trong những ngành công nghiệp phát triển nhanh nhất toàn cầu. Cùng với sự tăng trưởng về doanh thu là sự bùng nổ về khối lượng dữ liệu, bao gồm: thông tin đơn hàng, hành vi khách hàng, hoạt động logistic, phản hồi người dùng,...

Tuy nhiên, dữ liệu thô không thể mang lại giá trị nếu không được xử lý và phân tích đúng cách. Các doanh nghiệp thương mại điện tử hiện đại đang đối mặt với hai thách thức lớn:

1. **Sự rời rạc dữ liệu:** Dữ liệu nằm rải rác ở nhiều định dạng (csv, json), nhiều hệ thống khác nhau (local, cloud).
2. **Nhu cầu ra quyết định tức thời:** Các nhà quản lý cần những báo cáo trực quan cập nhật để trả lời các câu hỏi: Tại sao đơn hàng bị giao trễ? Khách hàng đang phàn nàn về điều gì? Sản phẩm nào đang là xu hướng?

Xuất phát từ như cầu thực tế đó, đề tài **"Phân tích và trực quan hóa dữ liệu thương mại điện tử của Olist"** được thực hiện nhằm mục đích xây dựng một quy trình xử lý dữ liệu khép kín để chuyển đổi dữ liệu thô thành các thông tin chi tiết có giá trị, hỗ trợ quá trình ra quyết định kinh doanh.

1.2 Giới thiệu về bộ dữ liệu Olist

1.2.1 Thông tin chung

Dữ liệu được sử dụng trong đề tài là Brazilian E-Commerce Public Dataset được cung cấp bởi Olist - nền tảng thương mại điện tử lớn nhất tại Brazil, đóng vai trò kết nối các doanh nghiệp nhỏ với khách hàng trên khắp cả nước.

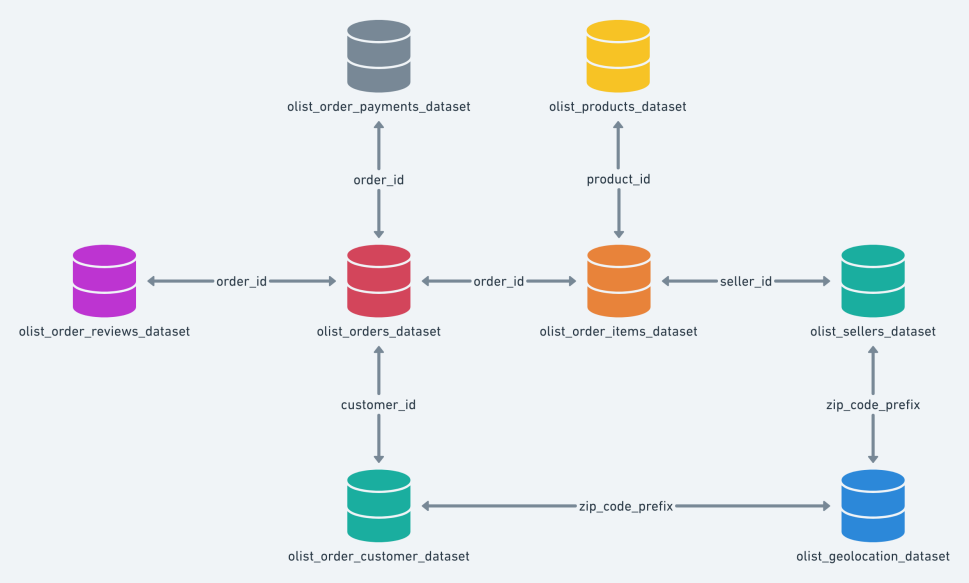
Các đặc điểm của bộ dữ liệu:

- Thời gian: Dữ liệu kéo dài từ năm 2016 đến 2018.
- Quy mô: Khoảng 100.000 đơn hàng thực tế (đã được ẩn danh).
- Cấu trúc: Dữ liệu không nằm trên một bảng đơn lẻ mà được tổ chức thành 9 bảng dữ liệu quan hệ, mô tả toàn bộ hành trình của một đơn hàng từ lúc khởi tạo, thanh toán, vận chuyển cho đến khi khách hàng nhận hàng và đánh giá.

1.2.2 Cấu trúc

- **olist_orders_dataset:** Bảng xương sống của hệ thống, chứa các mốc thời gian quan trọng (ngày đặt hàng, ngày duyệt đơn, ngày vận chuyển, ngày giao thành công) và trạng thái đơn hàng. Đây là cơ sở để tính toán các chỉ số KPI về Logistics (thời gian giao hàng, tỷ lệ trễ hạn).
- **olist_order_items_dataset:** Lưu trữ chi tiết từng sản phẩm trong một đơn hàng, bao gồm giá bán và phí vận chuyển.

- **olist_order_payments_dataset:** Cung cấp thông tin về dòng tiền, bao gồm phương thức thanh toán (thẻ tín dụng, voucher, tiền mặt,...) và số lần trả góp.
- **olist_products_dataset:** Chứa đặc tính vật lý của sản phẩm (danh mục, kích thước, khối lượng).
- **olist_customers_dataset:** Thông tin về khách hàng mua (đã ẩn dung).
- **olist_sellers_dataset:** Thông tin về người bán.
- **olist_order_reviews_dataset:** Chứa dữ liệu phản hồi của khách hàng sau khi hoàn tất đơn hàng, bao gồm số điểm đánh giá (từ 1 đến 5 sao) và nội dung bình luận. Đây là nguồn dữ liệu quý giá để đo lường mức độ hài lòng của khách hàng.
- **olist_geolocation_dataset:** Chứa thông tin zip code tương ứng với tọa độ.
- **olist_product_category_name_translation:** Bảng từ điển hỗ trợ tên danh mục sản phẩm từ tiến Bồ Đảo Nha sang tiếng Anh, giúp chuẩn hóa dữ liệu.



Hình 1: Sơ đồ quan hệ dữ liệu gốc của bộ dữ liệu Olist

1.3 Mục tiêu

Đề tài tập trung giải quyết các bài toán trên thông qua 3 mục tiêu cụ thể:

- **Xây dựng Pipline ETL tự động hóa:** Thiết lập quy trình trích xuất, biến đổi và nạp dữ liệu. Sử dụng công nghệ Apache Spark để xử lý lượng lớn dữ liệu, thực hiện làm sạch và tính toán các chỉ số phát sinh.
- **Khắc phục rào cản kết nối Hybrid:** Ứng dụng giải pháp tunneling (Ngrok) để kết nối kho dữ liệu nội bộ (MinIO Local) với nền tảng xử lý đám mây (Kaggle).
- **Trực quan hóa và phân tích chuyên sâu:** Xây dựng hệ thống báo cáo thông minh trên Power Bi.

2 Cơ sở lý thuyết và kiến trúc hệ thống

2.1 Cơ sở lý thuyết và công nghệ nền tảng

2.1.1 Apache Spark

Để giải quyết thách thức về khối lượng dữ liệu lớn và cấu trúc phức tạp của Olist, đề tài lựa chọn **Apache Spark** làm nền tảng xử lý trung tâm. Spark là một engine thống nhất cho xử lý dữ liệu lớn, nổi bật với khả năng tính toán trong bộ nhớ, giúp tốc độ xử lý nhanh hơn gấp nhiều lần so với các mô hình MapReduce truyền thống khi thực hiện các tác vụ lặp đi lặp lại.

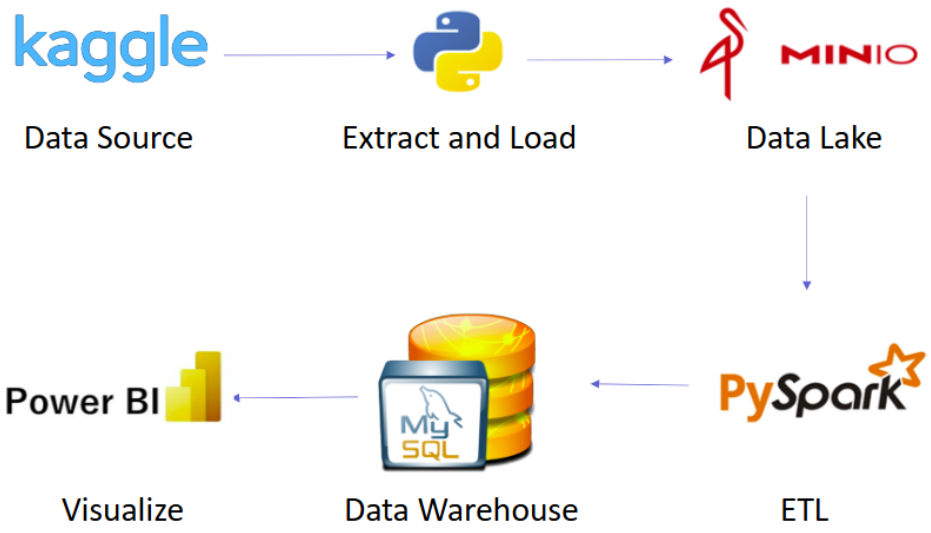
Trong bài tập lớn, thư viện **PySpark** được sử dụng để tương tác với Spark Core. Việc lựa chọn PySpark cho phép tận dụng sự linh hoạt và hệ sinh thái thư viện phong phú của ngôn ngữ Python, đồng thời vẫn đảm bảo hiệu năng xử lý mạnh mẽ của Spark.

2.1.2 Kiến trúc Hybrid Cloud

Một trong những thách thức kỹ thuật lớn nhất của đề tài là việc kết hợp tài nguyên lưu trữ cục bộ với sức mạnh tính toán trên đám mây. Cụ thể, dữ liệu được lưu trữ tại máy cá nhân, trong khi việc xử lý tận dụng tài nguyên phần cứng mạnh mẽ của Kaggle.

Để giải quyết vấn đề giao tiếp qua internet giữa MinIO và Kaggle, bài tập ứng dụng kỹ thuật tuneling thông qua công cụ **Ngrok**. Ngrok hoạt động bằng cách thiết lập một đường hầm mã hóa an toàn từ máy cục bộ đến public endpoint của Ngrok, cho phép Spark truy cập dữ liệu nội bộ như thể nó đang nằm trên một Public Cloud S3.

2.2 Kiến trúc hệ thống



Hình 2: Sơ đồ kiến trúc tổng thể

2.2.1 Tầng nguồn dữ liệu

- Thành phần: Bộ dữ liệu Brazilian E-Commerce Public Dataset
- Đặc điểm: Dữ liệu ở trạng thái thô, chưa được chuẩn hóa, chứa các giá trị rỗng, phản ánh nguyên trạng hoạt động ghi nhận giao dịch của hệ thống nguồn.

2.2.2 Tầng thu thập và lưu trữ

- Thành phần: MinIO Object Storage (triển khai trên Docker Container).
- Vai trò: Đóng vai trò là Data Lake.
- Hoạt động:
  - Quy trình thu thập sử dụng Python script để tải dữ liệu từ nguồn vào MinIO.
  - MinIO giả lập môi trường lưu trữ chuẩn S3 ngay tại hạ tầng cục bộ, đảm bảo tính bảo mật và toàn vẹn của dữ liệu gốc trước khi đưa vào xử lý.

2.2.3 Tầng xử lý ETL

- Thành phần: Apache Spark (PySpark) trên Kaggle và Ngrok.
- Vai trò: Là tầng trung tâm hệ thống, chịu trách nhiệm giải quyết bài toán kết nối Hybrid Cloud và thực hiện biến đổi dữ liệu.
- Hoạt động:
  - Thiết lập kết nối: Do MinIO nằm trong mạng nội bộ, Ngrok được sử dụng làm lớp trung gian để thiết lập một đường hầm mã hóa, công khai cổng dịch vụ của MinIO (Port 9000) ra internet.
  - Trích xuất: PySpark kết nối tới endpoint của Ngrok thông qua giao thức s3a để đọc dữ liệu từ Data Lake MinIO lên bộ nhớ đệm.
  - Biến đổi: Sau khi dữ liệu đã nằm trong môi trường Spark, các thuật toán xử lý phân tán được kích hoạt để làm sạch và làm giàu dữ liệu.
  - Nạp: Dữ liệu sau khi xử lý được định hình lại theo mô hình ngôi sao và ghi xuống tầng kho dữ liệu.

2.2.4 Tầng kho dữ liệu

- Thành phần: Aiven MySQL (Cloud Database).
- Vai trò: Vùng dữ liệu sạch.
- Cấu trúc: Dữ liệu tại đây được tổ chức theo mô hình ngôi sao, bao gồm các bảng fact và bảng dim. Đây là dữ liệu đã được tối ưu hóa cho các truy vấn phân tích và đảm bảo tính nhất quán phục vụ cho báo cáo.

2.2.5 Tầng trực quan hóa

- Thành phần: Microsoft Power BI.
- Vai trò: Giao diện tương tác với người dùng.
- Hoạt động: Kết nối trực tiếp tới Data Warehouse để truy xuất dữ liệu. Dữ liệu số liệu được chuyển đổi thành các biểu đồ, dashboard quản trị,..., hỗ trợ việc ra quyết định kinh doanh dựa trên dữ liệu.

3 Xây dựng luồng xử lý dữ liệu

3.1 Giai đoạn trích xuất dữ liệu (Extraction)

Quá trình trích xuất bắt đầu bằng việc thiết lập kết nối an toàn giữa Spark Session trên cloud và MinIO dưới local.

Để Spark có thể đọc dữ liệu từ MinIO qua đường hầm được tạo bảo Ngrok, môi trường thực thi được khởi tạo với các tham số cấu hình hệ thống tệp tin s3a đặc thù:

- Định tuyến qua tunnel (fs.s3a.endpoint): Thay vì trỏ đến địa chỉ nội bộ, tham số endpoint được cấu hình trỏ tới url công khai của Ngrok.
- Cơ chế truy cập đường dẫn (fs.s3a.path.style.access): Hệ thống bắt buộc kích hoạt chế độ Path Style Access. Điều này đảm bảo các yêu cầu HTTP đi qua đường hầm Ngrok được định tuyến chính xác đến bucket dữ liệu đích mà không bị lỗi phân giải tên miền.
- Bảo mật đường truyền (fs.s3a.connect.ssl.enable): Kích hoạt mã hóa SSL/TLS cho toàn bộ dữ liệu truyền tải giữa local và cloud, đảm bảo tính bảo mật khi dữ liệu di chuyển qua hạ tầng internet công cộng.

3.2 Giai đoạn biến đổi dữ liệu (Transform)

3.2.1 Làm sạch dữ liệu

- Chuyển tên sản phẩm trong bảng products về tiếng Anh, nếu không có bản dịch thì giữ nguyên tên gốc.
- Chuẩn hóa bảng geolocation: Tổng hợp tọa độ theo phương pháp tính trung bình và định dạng lại tên thành phố (loại bỏ dấu, viết hoa chữ cái đầu).
- Loại bỏ các đơn hàng không giao thành công trong bảng orders.
- Loại bỏ những reviews có ngày tháng không hợp lệ và chuẩn hóa kiểu dữ liệu của các cột ngày tháng trong bảng reviews.

3.2.2 Làm giàu dữ liệu

- Tính thời gian giao hàng và kiểm tra đơn hàng có bị giao muộn không.
- Tính tổng giá trị đơn hàng trong order_items.
- Phân loại các thành phố, bang theo khu vực.

3.3 Gia đoạn nạp dữ liệu (Load)

3.3.1 Chiến lược ghi theo lô

Việc ghi từng bản ghi riêng lẻ qua mạng sẽ gây độ trễ rất lớn. Hệ thống giải quyết vấn đề này bằng cách cấu hình trình điều kiển JDBC với tham số rewriteBatchedStatements=true.

- Cơ chế: Trình điều khiển sẽ tự động gom hàng nghìn câu lệnh insert đơn lẻ thành một câu lệnh insert đa giá trị duy nhất trước khi gửi đến cơ sở dữ liệu.
- Hiệu quả: Giảm thiểu tối đa số lượng round-trip qua mạng, tăng tốc độ nạp dữ liệu lên gấp nhiều lần.

3.3.2 Kiểm soát mức độ song song

Để bảo vệ cơ sở dữ liệu khỏi tình trạng quá tải kết nối từ hệ thống phân tán Spark:

- Giải pháp: Sử dụng kỹ thuật coalesce để giảm số lượng phân vùng dữ liệu xuống một con số tối ưu trước khi ghi.
- Tác dụng: Giới hạn số lượng kết nối đồng thời vào MySQL, đảm bảo quá trình ghi diễn ra ổn định nhưng vẫn tận dụng được khả năng xử lý đa luồng.

3.3.3 Đảm bảo tính nhất quán dữ liệu

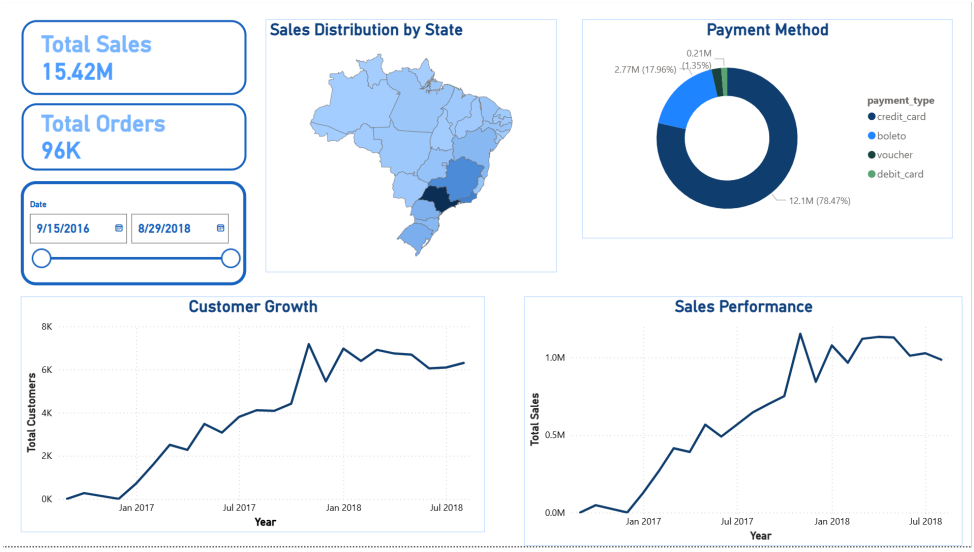
Chế độ ghi đè (mode='overwrite'): Hệ thống áp dụng chiến lược 'full refresh'. Mỗi lần pipeline chạy, bảng dữ liệu cũ sẽ được thay thế hoàn toàn bằng dữ liệu mới nhất. Điều này đảm báo tính bất biến cho hệ thống báo cáo.

4 Kết quả trực quan hoá

Sau khi hoàn tất quá trình xử lý dữ liệu và xây dựng mô hình dữ liệu, nhóm đã thực hiện trực quan hóa dữ liệu trên Power BI. Bộ báo cáo bao gồm 3 trang chính: Executive Overview, Product & Seller & Customer Experience.

4.1 Trang 1: Executive Overview

Trang báo cáo này cung cấp cái nhìn toàn cảnh về tình hình kinh doanh của Olist trong giai đoạn từ tháng 9/2016 đến tháng 8/2018.



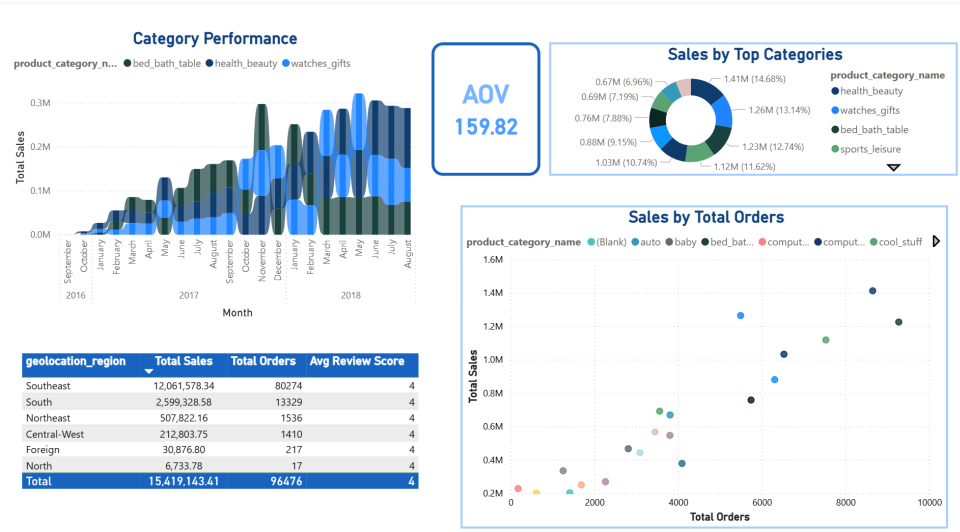
Hình 3: Tổng quan bán hàng

Phân tích các chỉ số chính:

- **Quy mô doanh thu:** Tổng doanh thu đạt 15.42 triệu, với tổng số lượng đơn hàng là 96,000 đơn.
- **Phân bổ địa lý:** Biểu đồ bản đồ nhiệt cho thấy sự tập trung mạnh mẽ của hoạt động kinh doanh tại khu vực Đông Nam Brazil. Bang São Paulo (SP) và các vùng lân cận có màu xanh đậm nhất, đóng vai trò là thị trường trọng điểm.
- **Phương thức thanh toán:** Thẻ tín dụng (Credit Card) là phương thức thanh toán áp đảo, chiếm 78.47% tổng giao dịch. Đứng thứ hai là Boleto (17.96%). Các hình thức khác chiếm tỷ trọng không đáng kể.
- **Xu hướng tăng trưởng:**
  - Biểu đồ *Customer Growth* cho thấy lượng khách hàng tăng trưởng đều đặn theo thời gian.
  - Biểu đồ *Sales Performance* ghi nhận đỉnh điểm doanh thu vào tháng 11/2017 (tương ứng với sự kiện Black Friday), sau đó duy trì mức ổn định cao trong năm 2018.

4.2 Trang 2: Product & Seller

Trang này tập trung phân tích sâu về các nhóm ngành hàng để xác định động lực tăng trưởng.



geolocation_region	Total Sales	Total Orders	Avg Review Score
Southeast	12,061,578.34	80274	4
South	2,599,328.58	13329	4
Northeast	507,822.16	1536	4
Central-West	212,803.75	1410	4
Foreign	30,876.80	217	4
North	6,733.78	17	4
Total	15,419,143.41	96476	4

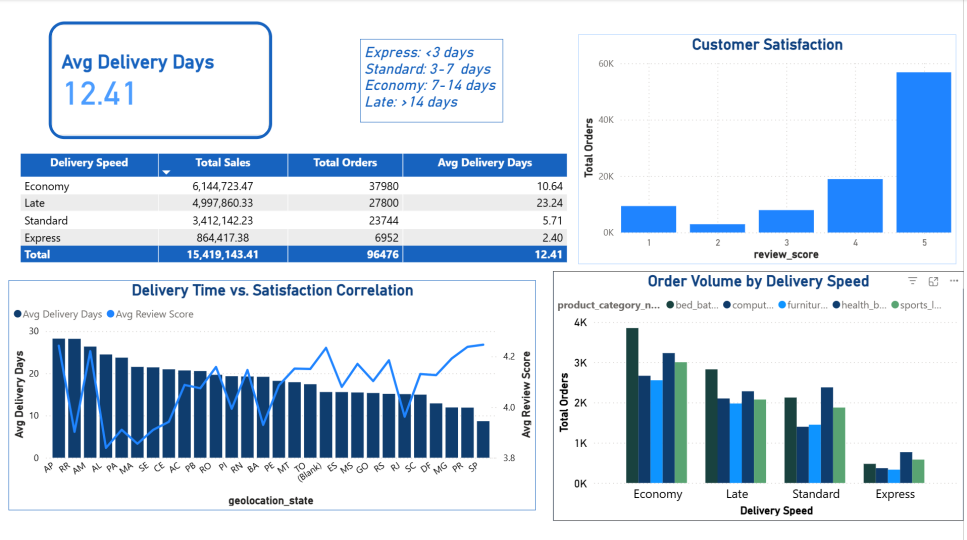
Hình 4: Hiệu suất sản phẩm

Phân tích chi tiết:

- **Giá trị đơn hàng trung bình (AOV):** Đạt mức 159.82, cho thấy mức chi tiêu của khách hàng ở mức khá tốt.
- **Top ngành hàng:** Dựa trên biểu đồ *Sales by Top Categories*, các nhóm hàng đóng góp doanh thu lớn nhất bao gồm: bed_bath_table, health_beauty, watches_gifts.
- **Tương quan Số lượng - Doanh thu:** Biểu đồ phân tán (*Sales by Total Orders*) cho thấy sự phân hóa rõ rệt. Nhóm *bed_bath_table* nằm ở góc trên bên phải, biểu thị cho việc vừa có số lượng bán lớn vừa có doanh thu cao, là nhóm hàng chủ lực của sàn.
- **Thống kê theo vùng:** Bảng dữ liệu chi tiết khẳng định lại sự thống trị của vùng Southeast với hơn 12 triệu doanh thu và hơn 80,000 đơn hàng.

4.3 Trang 3: Logistics & Customer

Trang cuối cùng phân tích hiệu quả của chuỗi cung ứng và tác động của nó đến trải nghiệm người dùng.



Hình 5: Logistics và Đánh giá khách hàng

Kết quả phân tích:

- **Thời gian giao hàng:** Thời gian giao hàng trung bình toàn trình là 12.41 ngày.
- **Phân loại tốc độ giao vận:** Biểu đồ cho thấy tốc độ giao hàng thuộc nhóm Economy là phổ biến nhất, trong đó:
  - Nhóm *Express* (<3 ngày): Trung bình 2.40 ngày.
  - Nhóm *Standard* (3-7 ngày): Trung bình 5.71 ngày.
  - Nhóm *Economy* (7-14 ngày): Trung bình 10.64 ngày.
  - Nhóm *Late* (>14 ngày): Thời gian trung bình lên tới 23.24 ngày. Đây là điểm nghẽn cần khắc phục.
- **Sự hài lòng của khách hàng:**
  - Phân bố điểm đánh giá cho thấy đa số khách hàng chấm 5 sao. Tuy nhiên, lượng đánh giá 1 sao khá cao, phản ánh một bộ phận khách hàng có trải nghiệm tồi tệ.
  - **Mối tương quan:** Biểu đồ kết hợp giữa *Avg Delivery Days* và *Avg Review Score* chỉ ra mối tương quan nghịch biến rõ rệt. Tại các bang có thời gian giao hàng kéo dài (cột màu xanh cao), điểm đánh giá trung bình giảm thấp (đường màu xanh dương đi xuống).

5 Kết luận

5.1 Kết luận chung

Thông qua quá trình thực hiện đề tài "Phân tích và Trực quan hóa dữ liệu thương mại điện tử của Olist", nhóm đã hoàn thành các mục tiêu đề ra ban đầu:

1. **Xử lý và Mô hình hóa dữ liệu:** Đã làm sạch, chuẩn hóa và xây dựng thành công mô hình dữ liệu dạng hình sao từ bộ dữ liệu thô, đảm bảo tính nhất quán và chính xác cho việc phân tích.

2. **Trực quan hóa hiệu quả:** Xây dựng bộ Dashboard trên Power BI với 3 góc nhìn toàn diện: Tổng quan quản trị, Hiệu suất sản phẩm và Vận hành Logistics.

3. **Kết luận chính:**

    - Xác định được thị trường trọng điểm (Đông Nam Brazil) và các dòng sản phẩm chủ lực (Bed_bath_table, Health_beauty).
    - Chứng minh được mối tương quan nghịch biến mạnh mẽ giữa thời gian giao hàng và sự hài lòng của khách hàng. Đây là cơ sở quan trọng để Olist cải thiện chất lượng dịch vụ.
    - Nhận diện được các điểm nghẽn trong vận hành, đặc biệt là tỷ lệ đơn hàng giao trễ (Late) còn cao ở các vùng xa.

5.2 Bài học kinh nghiệm

Trong quá trình thực hiện đồ án, nhóm đã tích lũy được những bài học quan trọng:

- **Về kỹ thuật:** Nâng cao kỹ năng sử dụng Power BI để tạo các Measure và kỹ thuật Data Modeling để tối ưu hiệu năng báo cáo.
- **Về tư duy phân tích:** Hiểu rõ tầm quan trọng của việc "kể chuyện với dữ liệu" (Data Storytelling). Một biểu đồ đẹp không chỉ cần thẩm mỹ mà phải truyền tải được thông điệp rõ ràng, hỗ trợ ra quyết định.
- **Về quy trình làm việc:** Nhận thấy tầm quan trọng của bước tiền xử lý dữ liệu (ETL). Dữ liệu đầu vào sạch sẽ quyết định 80% sự thành công của dự án.

Lời cảm ơn

Lời đầu tiên, chúng em xin gửi lời cảm ơn chân thành nhất đến **TS. Nguyễn Hữu Đức**. Trong suốt quá trình học tập và thực hiện đồ án môn *Quản trị dữ liệu và trực quan hóa*, thầy đã tận tình giảng dạy, định hướng và đưa ra những nhận xét quý báu giúp chúng em hoàn thiện đề tài này.

Mặc dù đã rất cố gắng, nhưng do kiến thức và kinh nghiệm còn hạn chế nên dự án khó tránh khỏi những thiếu sót. Chúng em rất mong nhận được sự đóng góp ý kiến của thầy để đề tài được hoàn thiện hơn.

Chúng em xin chân thành cảm ơn!