**TRƯỜNG CAO ĐẲNG FPT POLYTECHNIC**

A logo of a company

AI-generated content may be incorrect.

**BÁO CÁO DỰ ÁN E - COMMERCE**

**Chuyên ngành Xử Lý Dữ Liệu  
<<Phân tích doanh thu và hành vi mua sắm trên Amazon>>**

Lớp:DAT20301 Nhóm: Nghiện Data

Các thành viên:

Nguyễn Phan Hoàng Tùng – PS44334 (Nhóm trưởng)

Lê Thị Tuyết Trinh – PS45187

Dương Bách Sang – PS45279

Lê Tấn Quốc Trung–PS49207

GV HD: Thầy Trần Văn Huy

TP.HCM 11 - 2025

MỤC LỤC

[1.1 Giới thiệu : 4](#_Toc214400991)

[1.2 Yêu cầu của công ty 6](#_Toc214400992)

[1.3 Lập kế hoạch dự án 10](#_Toc214400993)

[2 Phân tích yêu cầu khách hàng 11](#_Toc214400994)

[2.1 Phân tích yêu cầu 11](#_Toc214400995)

[2.2 Câu chuyện dữ liệu 13](#_Toc214400996)

[2.2.1 Đặt vấn đề : 15](#_Toc214400997)

[2.2.2 Xác định câu chuyện 17](#_Toc214400998)

[2.2.3 Xác định rõ đối tượng 17](#_Toc214400999)

[2.2.4 Xác định câu chuyện chi tiết 19](#_Toc214401000)

[2.2.5 Trình bày dữ liệu 20](#_Toc214401001)

[2.2.6 Những điều cần lưu ý 21](#_Toc214401002)

[2.3 Kiến trúc hệ thống 22](#_Toc214401003)

[2.3.1 Kiến trúc 22](#_Toc214401004)

[2.3.2 Giải thích 23](#_Toc214401005)

[2.4 Giải thích về bộ dữ liệu khách hàng 24](#_Toc214401006)

[2.4.1 Các khái niệm 24](#_Toc214401007)

[2.4.1 Các trường dữ liệu 25](#_Toc214401008)

[2.5 Làm sạch và chuyển đổi dữ liệu 27](#_Toc214401009)

[2.6 Chuẩn bị dữ liệu 29](#_Toc214401010)

[2.6.1 Giải pháp lưu trữ dữ liệu 30](#_Toc214401011)

[2.6.2 Giải pháp phân bố dữ liệu 31](#_Toc214401012)

[2.7 Làm sạch dữ liệu 32](#_Toc214401013)

[2.7.1 Các vấn đề ảnh hưởng tới dữ liệu 32](#_Toc214401014)

[2.7.2 Các tiêu chí đánh giá chất lượng dữ liệu 32](#_Toc214401015)

[2.7.3 Các bước làm sạch dữ liệu 32](#_Toc214401016)

[2.8 Chuyển đổi dữ liệu 34](#_Toc214401017)

[2.8.1 Các trường hợp cần chuyển đổi: 34](#_Toc214401018)

[2.8.2 Kỹ thuật chuyển đổi: 34](#_Toc214401019)

[2.8.3 Trình bày các phép chuyển đổi trong dự án 34](#_Toc214401020)

[3 Xử lý dữ liệu 35](#_Toc214401021)

[3.1 Chuẩn hóa dữ liệu 35](#_Toc214401022)

[3.1.1 Trình bày các bước chuẩn hóa trong dự án 36](#_Toc214401023)

[3.2 Mô hình hóa dữ liệu 37](#_Toc214401024)

[3.2.1 Các loại mô hình hóa 37](#_Toc214401025)

[3.2.2 Các tiêu chí đánh giá mô hình dữ liệu 38](#_Toc214401026)

[3.3.3 Trình bày các bước mô hình hóa 38](#_Toc214401027)

[3.3.4 Trình bày các bước tạo bảng dữ liệu 42](#_Toc214401028)

[3.3 Xử lý dữ liệu DAX 42](#_Toc214401029)

[3.3.1 Measure 43](#_Toc214401030)

[3.3.2 Calculated column 43](#_Toc214401031)

[3.3.3 Filter 43](#_Toc214401032)

[4 Trực quan hóa dữ liệu 44](#_Toc214401033)

[4.1 Các kỹ thuật trực quan hóa 44](#_Toc214401034)

[4.2 Các nguyên tắc trực quan hóa 44](#_Toc214401035)

[4.3 Trình bày cách thêm visual mới 44](#_Toc214401036)

[4.4 Trình bày tạo các report cho dự án 44](#_Toc214401037)

[5 Xây dựng báo cáo 45](#_Toc214401038)

[5.1 Dashboard và report 45](#_Toc214401039)

[5.2 Xây dựng báo cáo 46](#_Toc214401040)

[6 KẾT LUẬN 47](#_Toc214401041)

[6.1 Báo cáo 47](#_Toc214401042)

[6.2 Khó khăn 48](#_Toc214401043)

[6.3 Hướng phát triển 48](#_Toc214401044)

**Giới thiệu dự án**

## Giới thiệu :

*Giới thiệu về hiện trạng*

* Trong bối cảnh thương mại điện tử phát triển mạnh mẽ toàn cầu, **Amazon** trở thành nền tảng dẫn đầu trong việc cung cấp sản phẩm tiêu dùng cho hàng trăm triệu khách hàng. Cùng với đó, các nhóm hàng thuộc lĩnh vực **“Chăm sóc cá nhân và gia đình”** (Personal & Household Care) ngày càng thu hút người tiêu dùng nhờ nhu cầu thiết yếu, tần suất mua cao và sự cạnh tranh khốc liệt về giá, thương hiệu và dịch vụ.
* Dữ liệu được thu thập phản ánh **hành vi mua sắm thực tế của khách hàng** trên Amazon, bao gồm thông tin giao dịch, nhân khẩu học, đánh giá sản phẩm, khu vực địa lý và hiệu quả kinh doanh. Điều này tạo điều kiện thuận lợi cho việc **phân tích toàn diện hành trình khách hàng** — từ đặc điểm cá nhân đến yếu tố giá, khuyến mãi và phản hồi sau mua hàng.

*Mô tả về hiện trạng dữ liệu*

#### Hiện trạng dữ liệu

Bộ dữ liệu gồm **51.290 bản ghi với 37 trường thông tin**, phản ánh chi tiết hoạt động kinh doanh và đặc điểm khách hàng. Cấu trúc dữ liệu bao gồm:

* **Thông tin giao dịch:** mã đơn hàng (order\_id), thời gian đặt (order\_date, order\_datetime), số lượng (quantity), doanh thu (sales), lợi nhuận (profit), chiết khấu (discount).
* **Thông tin sản phẩm:** tên sản phẩm (product), mã sản phẩm (product\_code), danh mục (category, subcategory), giá đơn vị (unit\_price), chi phí sản xuất (cost\_per\_product).
* **Thông tin khách hàng:** nhân khẩu học (giới tính, tình trạng hôn nhân, nghề nghiệp, thu nhập, học vấn), khu vực địa lý (city, state, country, region, market).
* **Đánh giá & phản hồi:** điểm đánh giá (rating) phản ánh mức độ hài lòng, tác động trực tiếp đến quyết định mua hàng.
* Dữ liệu này cho phép triển khai **phân tích hành vi mua sắm đa chiều**, xác định **xu hướng tiêu dùng, phân khúc khách hàng tiềm năng** và **các yếu tố ảnh hưởng đến hiệu suất kinh doanh**.

*Công cụ*

Để đảm bảo khả năng trực quan hoá và phân tích hiệu quả, nhóm dự án sử dụng Power BI làm công cụ chính cho việc xử lý và trình bày dữ liệu.  
 Các bước thực hiện bao gồm**:**

1. **Làm sạch & chuẩn hóa dữ liệu (Data Cleaning):**
   * Loại bỏ dữ liệu trùng, thiếu hoặc sai định dạng.
   * Chuyển đổi định dạng ngày tháng (order\_date, datetime) và chuẩn hoá đơn vị đo lường.
2. **Xây dựng mô hình dữ liệu (Data Modeling):**
   * Kết nối các bảng dữ liệu theo quan hệ khóa (Order – Customer – Product).
   * Tính toán các chỉ số KPI: *Doanh thu, Lợi nhuận, Tỷ suất lợi nhuận, Tỷ lệ giảm giá trung bình, Rating trung bình,...*
3. **Trực quan hóa dữ liệu (Data Visualization):**
   * Thiết kế Dashboard Power BI gồm các trang phân tích:
     + Doanh thu & Lợi nhuận theo thời gian
     + Top sản phẩm và danh mục bán chạy
     + Phân bố hành vi khách hàng theo khu vực & nhân khẩu học
     + Ảnh hưởng của rating và discount đến doanh số
4. **Khai thác và phân tích chuyên sâu (Insight Analysis):**
   * Dùng các biểu đồ động (Dynamic visuals), slicers và drill-through để khám phá sâu mối quan hệ giữa các yếu tố.
   * Xác định phân khúc khách hàng sinh lời cao và các chiến lược giá – khuyến mãi hiệu quả nhất.

* **Thông qua Power BI, toàn bộ quy trình phân tích trở nên trực quan, dễ hiểu và có thể cập nhật theo thời gian thực, giúp nhóm dự án chuyển dữ liệu thô thành thông tin chiến lược có giá trị cho doanh nghiệp.**

*Kết luận hiện trạng*

Từ dữ liệu được thu thập, nhóm dự án nhận thấy xu hướng tiêu dùng tại các thị trường khác nhau có sự khác biệt đáng kể. Những yếu tố như giá bán, chương trình giảm giá, và đánh giá của khách hàng đều ảnh hưởng mạnh đến doanh số và lợi nhuận.

Việc áp dụng Power BI trong dự án không chỉ giúp tự động hóa quá trình tổng hợp dữ liệu, mà còn tạo ra hệ thống dashboard thông minh, hỗ trợ nhà quản lý ra quyết định kinh doanh nhanh, chính xác và dựa trên dữ liệu thực tế.

## Yêu cầu của công ty

* *Nêu yêu cầu*

 **Về dữ liệu:**

* Làm sạch, chuẩn hóa và phân tích dữ liệu bán hàng theo thời gian, khu vực, danh mục sản phẩm.
* Loại bỏ giá trị trùng lặp, xử lý dữ liệu bị thiếu hoặc định dạng sai (đặc biệt các cột order\_date, rating, annual\_income, sales, profit).
* Chuẩn hoá đơn vị đo lường và định danh sản phẩm theo từng **danh mục (Category)** và **phân nhóm (Subcategory)**.
* Phân tách dữ liệu thành **các bảng quan hệ**: *Customer*, *Order*, *Product*, *Region* để thuận tiện cho việc xây dựng mô hình dữ liệu trong Power BI.

 **Quản lý & lưu trữ:**

* Dữ liệu được lưu trữ dưới dạng CSV và xử lý trên nền tảng Power BI, giúp trực quan hóa và ra quyết định.
* Thiết lập **Data Model quan hệ (Relational Model)** gồm các bảng Fact và Dimension nhằm tối ưu hiệu năng khi truy vấn..
* Đảm bảo dữ liệu được **mã hóa và bảo mật**, tránh rò rỉ thông tin người tiêu dùng.

 **Mục tiêu:**

* Hỗ trợ doanh nghiệp **nắm bắt xu hướng thị trường** trong lĩnh vực chăm sóc cá nhân & gia đình.
* Giúp đội ngũ Marketing & Kinh doanh **đưa ra quyết định định giá, khuyến mãi, phân phối và quảng cáo hợp lý**.
* Tăng khả năng **dự báo doanh thu & tối ưu lợi nhuận**, đồng thời nâng cao **mức độ hài lòng khách hàng** dựa trên dữ liệu thực tế.

 **Về phân tích dữ liệu:**

Tiến hành các phân tích chuyên sâu nhằm:

* Đánh giá **hiệu suất kinh doanh (KPI)**: doanh thu, lợi nhuận, tỷ lệ chiết khấu, tỷ suất lợi nhuận, rating trung bình, v.v.
* Xác định **mối tương quan giữa giá – đánh giá – doanh số**, làm rõ ảnh hưởng của chương trình khuyến mãi đến hành vi tiêu dùng.
* Phân tích **hành vi khách hàng** theo giới tính, thu nhập, độ tuổi, nghề nghiệp và khu vực địa lý.
* Phân tích **hiệu quả danh mục sản phẩm**: nhóm hàng bán chạy, sản phẩm có lợi nhuận thấp, và tỷ lệ hài lòng khách hàng.
* *Đánh giá tính khả thi (năng lực, khả năng của mình có thể thực hiện dự án này)*

**Kỹ năng hiện có:**

* Thành thạo **Excel cơ bản**: xử lý dữ liệu thô, thống kê, tạo bảng tổng hợp (PivotTable) và biểu đồ mô tả.
* Có kiến thức về **SQL cơ bản**: thao tác truy vấn dữ liệu, lọc – nhóm – tổng hợp thông tin từ tập dữ liệu lớn.
* Đã **thực hành Power BI ở mức cơ bản**: nhập dữ liệu từ CSV, làm sạch bằng Power Query, thiết kế dashboard đơn giản và tạo các biểu đồ trực quan hóa cơ bản.
* Có kinh nghiệm cơ bản trong **phân tích hành vi tiêu dùng** và nhận diện xu hướng dữ liệu trong thương mại điện tử.
* **Kỹ năng cần học thêm:**
* **Nâng cao kỹ năng Power Query** để làm sạch, chuẩn hóa dữ liệu quy mô lớn, xử lý giá trị khuyết và sai định dạng.
* **Viết DAX nâng cao** để tạo các **Measure và Calculated Column** phục vụ tính toán KPI (doanh thu, lợi nhuận, tỷ lệ chiết khấu, chỉ số hài lòng khách hàng,...).
* Thiết kế **dashboard động, chuyên nghiệp theo tiêu chuẩn báo cáo doanh nghiệp**, có tương tác với slicers, filters, drill-through.
* Nâng cao khả năng sử dụng **Python hoặc Excel Power Pivot** để phân tích sâu dữ liệu trước khi nhập vào Power BI.
* Học cách **kể chuyện bằng dữ liệu (Data Storytelling)** — biến kết quả phân tích thành insight dễ hiểu và trực quan.
  + =>Với năng lực hiện tại, nhóm hoàn toàn có thể **triển khai và hoàn thiện dự án ở mức chuyên sâu**, bao gồm các bước:
* Làm sạch và chuẩn hóa dữ liệu Amazon bằng Power Query.
* Xây dựng mô hình dữ liệu quan hệ trong Power BI.
* Thiết kế hệ thống báo cáo gồm các chỉ số chính: doanh thu, lợi nhuận, tỷ lệ giảm giá, rating trung bình, hành vi khách hàng.
* Trực quan hóa và rút ra insight hỗ trợ quyết định kinh doanh.

## Lập kế hoạch dự án

* *Lập kế hoạch kế hoạch dự án theo mẫu sau*

| **TT** | **Hạng mục** | **Bắt đầu** | **Kết thúc** | **Kết quả** |
| --- | --- | --- | --- | --- |
| 1 | Giới thiệu dự án | 03/11/2025 | 04/11/2025 | Hoàn thành |
| 1.2 | Yêu cầu của công ty | 03/11/2025 | 04/11/2025 | Hoàn thành |
| 1.3 | Lập kế hoạch dự án | 03/11/2025 | 05/11/2025 | Hoàn thành |
| 2 | Phân tích yêu cầu khách hàng | 05/11/2025 | 06/11/2025 | Hoàn thành |
| 2.1 | Phân tích yêu cầu | 06/11/2025 | 06/11/2025 | Hoàn thành |
| 2.2 | Câu chuyện dữ liệu | 06/11/2025 | 06/11/2025 | Hoàn thành |
| 2.3 | Kiến truc hệ thống |  |  |  |
| 2.4 | Giải thích về bộ dữ liệu khách hàng | 07/11/2025 | 09/11/2025 | Hoàn thành |
| 3 | Làm sạch và chuyển đổi dữ liệu | 07/11/2025 | 09/11/2025 | Hoàn thành |
| 3.1 | Chuẩn bị dữ liệu |  |  |  |
| 3.2 | Làm sạch dữ liệu | 09/11/2025 | 10/11/2025 | Hoàn thành |
| 3.3 | Chuyển đổi dữ liệu | 09/11/2025 | 10/11/2025 | Hoàn thành |
| 4 | Xử lý dữ liệu |  |  |  |
| 4.1 | Chuẩn hóa dữ liệu | 11/11/2025 | 12/11/2025 | Hoàn thành |
| 4.2 | Mô hình hóa dữ liệu | 12/11/2025 | 14/11/2025 | Hoàn thành |
| 4.3 | Xử lý dữ liệu DAX | 14/11/2025 | 15/11/2025 | Hoàn thành |
| 5 | Trự quan hóa dữ liệu | 15/11/2025 | 16/11/2025 | Hoàn thành |
| 5.1 | Các kỹ thuật trực quan hóa | 16/11/2025 | 18/11/2025 |  |
| 5.2 | Các nguyên tắc trực quan hóa | 18/11/2025 | 20/11/2025 |  |
| 5.3 | Trình bày cách thêm Visual | 20/11/2025 | 22/11/2025 |  |
| 5.4 | Trình bày tạo các report cho dự án | 22/11/2025 | 24/11/2025 |  |
| 6 | Xây dựng báo cáo | 24/11/2025 | 26/11/2025 |  |
| 6.1 | Dashboard và report | 26/11/2025 | 28/11/2025 |  |
| 6.2 | Dashboard | 28/11/2025 | 30/11/2025 |  |
| 6.3 | Report | 01/12/2025 | 03/12/2025 |  |
| 6.4 | Bookmark | 03/12/2025 | 05/12/2025 |  |
| 7 | KẾT LUẬN | 05/12/2025 | 07/12/2025 |  |
| 7.1 | Báo cáo | 07/12/2025 | 09/12/2025 |  |
| 7.2 | Khó khăn | 09/12/2025 | 11/12/2025 |  |
| 7.3 | Thuận lợi | 11/12/2025 | 13/12/2025 |  |
| 7.4 | Hướng phát triển | 13/12/2025 | 19/12/2025 |  |

# Phân tích yêu cầu khách hàng

* *Tiến hành phân tích yêu cầu khách hàng theo các bước sau đây*

## Phân tích yêu cầu

#### *Dựa vào bộ dữ liệu công ty cung cấp và các yêu cầu khách hàng gửi, nhóm tiến hành phân tích như sau:*

* ***Dữ liệu:*** *Bộ dữ liệu “Amazon Personal & Household Care Dataset” bao gồm hơn* ***51.000 bản ghi****, cung cấp thông tin chi tiết về:*
  + *Giao dịch bán hàng (order\_id, order\_date, quantity, sales, discount, profit)*
  + *Sản phẩm (product\_code, product, category, subcategory, unit\_price, cost\_per\_product)*
  + *Khách hàng (customer\_id, gender, marital\_status, income, occupation, education\_level)*
  + *Khu vực địa lý (city, state, country, region, market)*
  + *Đánh giá và phản hồi (rating, review behavior)*
* *Bộ dữ liệu này phản ánh* ***hành vi mua sắm và mức độ hài lòng của người tiêu dùng trên Amazon****, cho phép xây dựng* ***phân tích hành vi, xu hướng và yếu tố tác động đến doanh thu – lợi nhuận****.*

#### *Quản lý và lưu trữ:*

* *Toàn bộ dữ liệu được chuẩn hóa và lưu trữ ở định dạng* ***CSV UTF-8****.*
* *Dữ liệu được phân chia theo mô hình* ***Fact – Dimension****, bao gồm:*
* ***Fact\_Sales:*** *chứa các giao dịch bán hàng, doanh thu, chiết khấu, lợi nhuận.*
* ***Dim\_Product:*** *thông tin sản phẩm, danh mục, giá.*
* ***Dim\_Customer:*** *nhân khẩu học và thu nhập khách hàng.*
* ***Dim\_Geography:*** *vùng, quốc gia, khu vực thị trường.*
* *Các bảng này được quản lý, kết nối và mô hình hóa trực tiếp trong* ***Power BI Desktop*** *để dễ dàng xử lý, truy vấn và tạo báo cáo động.*

#### *Công nghệ sử dụng:*

* ***Microsoft Power BI****: công cụ trực quan hoá và phân tích chính.*
* ***Power Query****: làm sạch và chuyển đổi dữ liệu trước khi mô hình hóa.*
* ***DAX (Data Analysis Expressions)****: tạo Measure và Calculated Columns để tính KPI (Doanh thu, Lợi nhuận, Tỷ lệ chiết khấu, Rating trung bình,…).*
* ***SQL / Excel hỗ trợ:*** *dùng để kiểm tra, tổng hợp dữ liệu ban đầu trước khi nhập vào Power BI.*

#### *Quyết định dùng công nghệ Power BI vì sao?*

*Dựa vào yêu cầu của khách hàng về* ***hiển thị trực quan, cập nhật nhanh, dễ phân tích và linh hoạt trong báo cáo****, Power BI được lựa chọn nhờ các ưu điểm sau:*

* ***Trực quan hóa mạnh mẽ:*** *hiển thị dữ liệu thông qua dashboard tương tác.*
* ***Kết nối linh hoạt:*** *hỗ trợ nhập từ nhiều định dạng như CSV, Excel, SQL Server.*
* ***Khả năng mở rộng:*** *cho phép thêm các trường dữ liệu và biểu đồ mới mà không ảnh hưởng mô hình gốc.*
* ***Tự động cập nhật (Auto Refresh):*** *thuận tiện khi cần phân tích dữ liệu định kỳ.*
* ***Phù hợp yêu cầu học thuật và thực tế doanh nghiệp*** *trong phân tích hành vi người tiêu dùng.*

## Câu chuyện dữ liệu

Xây dựng câu chuyện phân tích dữ liệu

* ***Ai đọc báo cáo này?***

*Độc giả chính là* ***ban quản lý kinh doanh và nhóm phân tích thương mại điện tử****. Họ quan tâm đến việc:*

*1.Sản phẩm nào mang lại* ***doanh thu cao nhất****?*

*2.Khách hàng* ***đánh giá thế nào*** *về sản phẩm và người bán?*

*3.Giá bán thấp có thật sự giúp sản phẩm bán chạy hơn?*

* ***Giải thích vấn đề là gì?***

*Trong bối cảnh thị trường Amazon cạnh tranh khốc liệt, doanh thu có thể bị ảnh hưởng bởi nhiều yếu tố: giá gốc – giá thật, chất lượng, niềm tin người dùng (qua rating), và xu hướng mua sắm (qua total\_sold).**Doanh thu và lợi nhuận giữa các khu vực có sự chênh lệch lớn. Một số sản phẩm có lượng bán cao nhưng lợi nhuận thấp do mức chiết khấu và giá giảm sâu. Bên cạnh đó, các lượt đánh giả sản phẩm góp phần ảnh hưởng rõ rệt đến lượng mua và hành vi khách hàng.*

* ***Ở mức độ nào?****Phân tích tập trung vào* ***sản phẩm chăm sóc cá nhân & gia đình****, với độ chi tiết theo* ***thị trường, khách hàng và danh mục sản phẩm****.*
* ***Sử dụng phương pháp gì?***
  + *Làm sạch dữ liệu và chuẩn hóa trong Power Query.*
  + *Xây dựng mô hình dữ liệu trong Power BI (Fact – Dimension).*
  + *Tạo KPI động: doanh thu, lợi nhuận, chiết khấu, rating trung bình.*
  + *Trực quan hóa bằng dashboard:*
    - *Trang 1: Hiệu suất kinh doanh tổng thể.*
    - *Trang 2: Xu hướng giá và rating theo sản phẩm.*
    - *Trang 3: Phân tích khách hàng theo thu nhập, giới tính, khu vực.*
    - *Trang 4: Insight hành vi mua và yếu tố tác động đến quyết định mua hàng.*
* *Phân tích yêu cầu khách hàng được xây dựng xoay quanh mục tiêu* ***hiểu rõ hành vi người tiêu dùng, tối ưu chiến lược bán hàng, và trực quan hoá hiệu quả kinh doanh*** *bằng Power BI. Từ đó, dự án mang lại giá trị thực tế trong việc* ***ra quyết định dựa trên dữ liệu (Data-driven Decision Making)****.*

### 2.2.1 Đặt vấn đề :

* **Mô tả thực trạng:**

Hiện nay, **Amazon** là nền tảng thương mại điện tử lớn nhất toàn cầu, nơi hàng triệu người tiêu dùng thực hiện các giao dịch mua sắm mỗi ngày. Trong số đó, **nhóm sản phẩm chăm sóc cá nhân & gia đình (Personal & Household Care)** chiếm tỷ trọng đáng kể nhờ nhu cầu thiết yếu, đa dạng thương hiệu và mức độ cạnh tranh cao.

Tuy nhiên, sự phong phú về mặt hàng, khu vực và đối tượng khách hàng lại khiến việc **phân tích hành vi tiêu dùng trở nên phức tạp**. Các doanh nghiệp gặp khó khăn trong việc trả lời các câu hỏi trọng tâm như:

* Sản phẩm nào thực sự mang lại lợi nhuận cao nhất?
* Giá bán, chiết khấu và đánh giá của người dùng ảnh hưởng ra sao đến quyết định mua hàng?
* Có sự khác biệt nào về hành vi tiêu dùng giữa các khu vực và nhóm thu nhập?

Trong bối cảnh đó, **việc trực quan hóa dữ liệu mua sắm và xây dựng hệ thống báo cáo thông minh** là cần thiết để doanh nghiệp **hiểu khách hàng hơn, tối ưu chiến lược bán hàng và gia tăng lợi nhuận**.

**Dữ liệu liên quan:**

Bộ dữ liệu được cung cấp gồm **hơn 51.000 bản ghi** phản ánh chi tiết về:

* **Giao dịch bán hàng** (order\_id, order\_date, quantity, sales, discount, profit)
* **Thông tin sản phẩm** (product, category, subcategory, unit\_price, cost\_per\_product)
* **Thông tin khách hàng** (gender, marital\_status, occupation, income, education\_level)
* **Đánh giá và hành vi tiêu dùng** (rating, repeat purchase, khu vực địa lý)

Sau khi xử lý và làm sạch, dữ liệu cho thấy một khối lượng thông tin lớn, dễ gây **quá tải và khó diễn giải bằng bảng số liệu thuần túy**.  
 👉 Do đó, nhóm cần tìm cách **chuyển đổi dữ liệu khô cứng thành insight sinh động, dễ hiểu và có tính thuyết phục cao**.

**Mục tiêu:**

Mục tiêu của phần này là **truyền đạt thông điệp dữ liệu một cách rõ ràng, hấp dẫn và định hướng ra quyết định**.  
 Cụ thể:

* **Trực quan hóa các chỉ số kinh doanh (KPI)** như doanh thu, lợi nhuận, tỷ lệ chiết khấu, mức độ hài lòng.
* **Giúp người đọc dễ dàng nắm bắt mối quan hệ giữa các yếu tố** (giá – rating – doanh số – lợi nhuận).
* **Hỗ trợ doanh nghiệp nhận diện xu hướng và cơ hội thị trường** trong lĩnh vực chăm sóc cá nhân & gia đình.
* Thể hiện thông tin một cách **sinh động, logic, dễ hiểu và có khả năng thuyết phục người ra quyết định**.

Dự án vì thế hướng tới việc **biến dữ liệu thô thành câu chuyện kinh doanh trực quan** – giúp lãnh đạo và bộ phận marketing **nắm bắt hành vi người tiêu dùng nhanh, chính xác và dựa trên dữ liệu thực tế**.

### 2.2.2 Xác định câu chuyện

* *Bằng cách đặt câu hỏi, hình thành giả thuyết, tổng hợp đào sâu dữ liệu để tìm câu trả lời:*

Trong dự án này, nhóm thực hiện xây dựng **“câu chuyện dữ liệu”** xoay quanh hành vi mua sắm các sản phẩm chăm sóc cá nhân & gia đình trên nền tảng Amazon.   
 Thay vì chỉ trình bày con số khô khan, câu chuyện được hình thành thông qua **hệ thống câu hỏi và giả thuyết** phản ánh mối quan tâm của doanh nghiệp:

* **Bạn đang cố gắng giải thích điều gì từ dữ liệu?** → Nhóm muốn xác định các yếu tố ảnh hưởng đến **hành vi mua hàng và hiệu quả kinh doanh**, bao gồm giá bán, chiết khấu, đánh giá khách hàng, khu vực và thu nhập.
* **Mục tiêu cụ thể của bạn khi kể câu chuyện dữ liệu này là gì?** → Câu chuyện hướng đến việc **làm rõ sự tương quan giữa rating – discount – profit**, qua đó giúp doanh nghiệp hiểu **khách hàng mua vì giá, thương hiệu hay trải nghiệm**.
* **Bạn muốn đề xuất giải pháp gì từ những phân tích này?** → Đề xuất **chiến lược định giá và khuyến mãi hợp lý**, tăng tỷ lệ hài lòng khách hàng và tối ưu doanh thu theo từng thị trường.

### 2.2.3 Xác định rõ đối tượng

* **Tìm kiếm mối tương quan:** So sánh tác động của giá và đánh giá sản phẩm đến lượng bán và lợi nhuận.
* **Tìm kiếm xu hướng:** Xác định nhóm sản phẩm bán chạy theo mùa, theo khu vực, hoặc theo phân khúc thu nhập.
* **Rút ra so sánh:** So sánh hành vi mua sắm giữa các nhóm giới tính, vùng địa lý, và mức thu nhập khác nhau.

#### Hình thành hướng kể chuyện trên Power BI:

Để truyền tải câu chuyện dữ liệu một cách sinh động, nhóm thiết kế các **dashboard tương tác** theo mạch kể sau:

1. **Mở đầu:** Toàn cảnh hoạt động kinh doanh — doanh thu, lợi nhuận, chiết khấu, rating trung bình.
2. **Thân bài:** Phân tích hành vi khách hàng — nhóm khách hàng nào mua nhiều, nhóm nào có rating cao, nhóm nào mang lại lợi nhuận thấp.
3. **Kết thúc:** Đưa ra insight và đề xuất giải pháp — định giá tối ưu, điều chỉnh khuyến mãi và chiến lược giữ chân khách hàng.

Câu chuyện dữ liệu được kể **bằng biểu đồ động, màu sắc tương phản và tương tác drill-down**, giúp người xem dễ dàng nắm bắt thông tin, khám phá chi tiết và hiểu sâu vấn đề kinh doanh.

* **Tổng kết:**

Phần “Xác định câu chuyện” tập trung vào việc **biến dữ liệu Amazon thành hành trình trực quan**, giúp người đọc **không chỉ thấy con số, mà còn hiểu được bức tranh toàn cảnh của hành vi tiêu dùng**.

Từ đó, doanh nghiệp có thể **ra quyết định nhanh, chính xác và mang tính chiến lược** dựa trên insight thực tế.

### 2.2.4 Xác định câu chuyện chi tiết

#### Giải thích dữ liệu rõ ràng và trình tự logic hợp lý

* **Bối cảnh:** Báo cáo được thực hiện trong bối cảnh thị trường thương mại điện tử phát triển mạnh mẽ, đặc biệt ở mảng **Personal & Household Care** – nơi người tiêu dùng ra quyết định mua sắm dựa trên nhiều yếu tố như giá, thương hiệu, chất lượng và đánh giá từ người dùng khác.
* **Ai sẽ xem báo cáo này?** Báo cáo hướng tới các **nhà quản lý kinh doanh**, **bộ phận marketing**, và **chuyên viên phân tích dữ liệu (data analyst)** đang cần **hiểu sâu về hành vi người tiêu dùng** để đề xuất chiến lược tăng trưởng.
* **Mục tiêu truyền tải:** Trình bày một **câu chuyện dữ liệu rõ ràng, mạch lạc**, đi từ tổng quan đến chi tiết:  
  1. Tổng quan tình hình kinh doanh.
  2. Xu hướng tiêu dùng theo khu vực và nhân khẩu học.
  3. Ảnh hưởng của rating và discount đến doanh thu.
  4. Kết luận và đề xuất chiến lược giá, marketing, và sản phẩm.

Câu chuyện được thể hiện trong Power BI qua cấu trúc dashboard đa tầng (Overview → Product → Customer → Insight), giúp người xem dễ dàng **nắm bắt logic phân tích và tìm ra nguyên nhân cốt lõi của vấn đề.**

### 2.2.5 Trình bày dữ liệu

#### Dùng hình ảnh và biểu đồ trực quan để trình bày kết quả phân tích dữ liệu.

* ***Biểu đồ sử dụng trong Power BI:***
  + **Column Chart / Bar Chart:** biểu diễn doanh thu, lợi nhuận, số lượng sản phẩm bán ra theo danh mục.
  + **Line Chart:** thể hiện xu hướng theo thời gian (sales & rating).
  + **Pie / Donut Chart:** phân bố khách hàng theo giới tính, khu vực, nhóm thu nhập.
  + **Scatter Plot:** mô tả mối quan hệ giữa rating – discount – profit.
  + **Map Chart:** trực quan hóa hành vi mua sắm theo quốc gia và vùng địa lý.
  + **Card KPI:** hiển thị nhanh các chỉ số chính (Doanh thu, Lợi nhuận, Tỷ lệ giảm giá, Rating trung bình).
* **Cách trình bày dữ liệu:**
  + Mỗi biểu đồ được thiết kế gọn gàng, có tiêu đề rõ ràng, màu sắc phân biệt theo danh mục sản phẩm.
  + Các **filter và slicer động** (theo thời gian, khu vực, danh mục, thu nhập) giúp người dùng dễ dàng tương tác và khám phá dữ liệu sâu hơn.
  + Dashboard có phần **tóm tắt insight cuối trang**, giải thích ý nghĩa của các biến động và mối quan hệ được phát hiện.

### 2.2.6 Những điều cần lưu ý

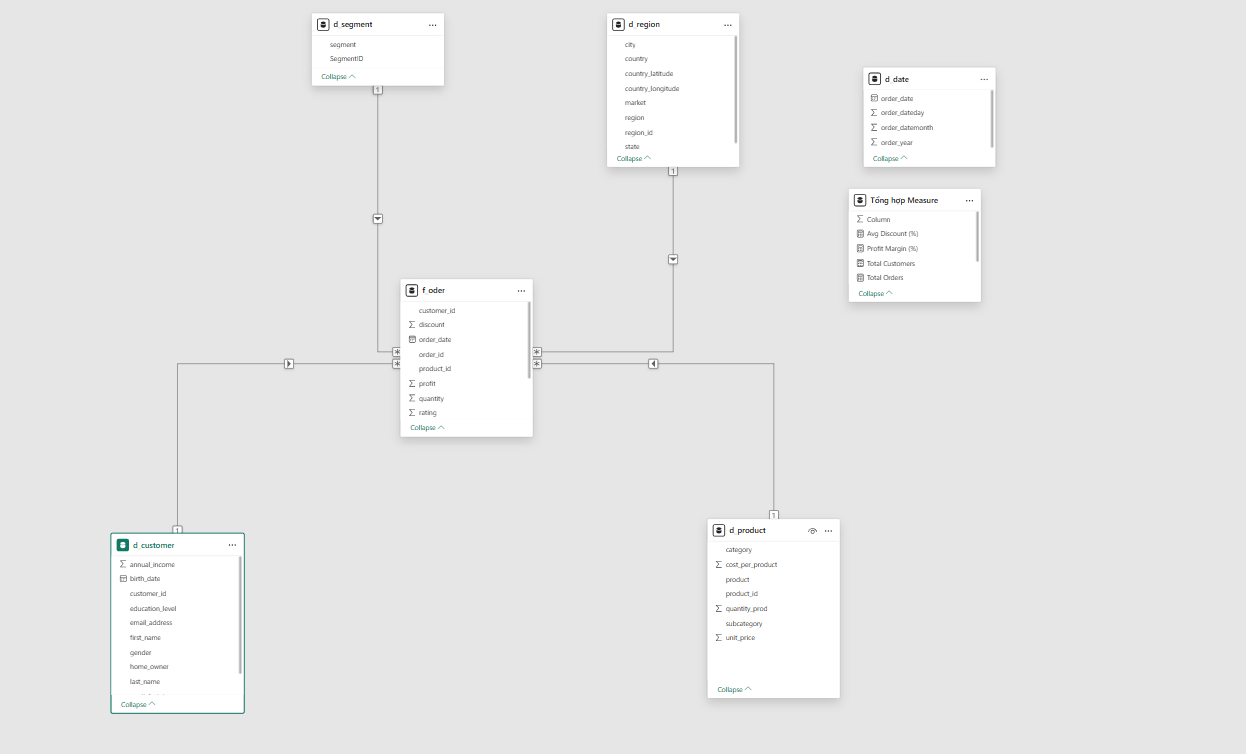
Dự án này có 1 số lưu ý quan trọng để đảm bảo tính mạch lạc của thông điệp truyền tải chính xác và hiệu quả đến độc giả:

* Tập trung vào "Insight" (Sự thật ngầm hiểu) chứ không phải "Data" (Dữ liệu): Báo cáo này hướng đến Ban quản lý kinh doanh. Do đó, câu chuyện dữ liệu sẽ tập trung trả lời các câu hỏi chiến lược (ví dụ: "Sản phẩm nào nên ngừng giảm giá?", "Khu vực nào mang lại lợi nhuận cao nhất?") thay vì trình bày các kỹ thuật xử lý dữ liệu phức tạp.
* Phân biệt rõ **"Tương quan"** và **"Nhân quả"**: Đây là lưu ý quan trọng nhất. Dữ liệu có thể cho thấy rating (đánh giá) cao thường đi kèm với sales (doanh số) cao. Tuy nhiên, điều này không đồng nghĩa rating cao gây ra sales cao. Báo cáo này chỉ trình bày các "mối liên hệ" và "xu hướng" quan sát được, không vội vàng kết luận "nguyên nhân" nếu không có phân tích chuyên sâu hơn.
* **Tránh "Thiên kiến xác nhận" (Confirmation Bias):** Người xem cần cẩn trọng để không chỉ tìm dữ liệu xác nhận cho một giả thuyết có sẵn (ví dụ: "giảm giá nhiều thì bán chạy"). Một sản phẩm có sales cao nhưng profit (lợi nhuận) âm do discount (chiết khấu) quá cao thực chất là một vấn đề kinh doanh. Câu chuyện dữ liệu phải luôn được nhìn nhận toàn diện qua cả 4 yếu tố: **Sales – Profit – Discount – Rating**.
* Bộ dữ liệu gốc (dù đã được làm sạch) vẫn có những hạn chế. Cụ thể, có một số dòng thiếu thông tin gender (giới tính) và thiếu thông tin product (sản phẩm), Vì lí do không đủ dữ kiện hoặc dữ kiện không đủ độ tin cậy nên có các ô trống, tuy nhiên vẫn chấp nhận được vì tính khách quan,đại đa số. Các phân tích liên quan đến các trường này được thực hiện dựa trên phần dữ liệu đầy đủ, nhưng cần được xem xét với một mức độ cẩn trọng nhất định.

## Kiến trúc hệ thống

### 2.3.1 Kiến trúc

* Trong dự án này, nhóm áp dụng kiến trúc **Mô hình hình sao (Star Schema)**. Đây là mô hình tối ưu nhất cho việc báo cáo và phân tích dữ liệu trên Power BI, giúp đơn giản hóa các truy vấn và tăng hiệu suất xử lý.
* Mô hình bao gồm một bảng trung tâm (Fact Table) chứa dữ liệu giao dịch và các bảng vệ tinh (Dimension Tables) chứa dữ liệu tham chiếu.



### 2.3.2 Giải thích

* *Giải thích cách thiết kế.*

Dựa vào sơ đồ quan hệ thực thể (ERD) phía trên, kiến trúc dữ liệu của dự án được tổ chức như sau:

**1. Bảng sự thật (Fact Table) - Fact\_Sales:**

* **Vai trò:** Là bảng trung tâm lưu trữ toàn bộ các sự kiện bán hàng chi tiết. Đây là nơi diễn ra các phép tính toán số học.
* **Dữ liệu chứa:**
  + **Các chỉ số đo lường (Measures):** Sales (Doanh thu), Quantity (Số lượng), Discount (Chiết khấu), Profit (Lợi nhuận), Rating (Đánh giá).
  + **Các khóa ngoại (Foreign Keys):** Customer\_ID, Product\_Code, Region\_Code (hoặc Geo\_ID) để liên kết với các bảng chiều.

**2. Các bảng chiều (Dimension Tables):** Các bảng này chứa thông tin mô tả, giúp trả lời các câu hỏi "Ai?", "Cái gì?", "Ở đâu?", "Khi nào?".

* **Dim\_Customer:** Lưu trữ thông tin nhân khẩu học khách hàng như Customer\_ID, Name, Segment (Phân khúc), Age (Tuổi), Gender (Giới tính).
* **Dim\_Product:** Chứa thông tin chi tiết về sản phẩm như Product\_Code, Category (Danh mục), Subcategory, Unit\_Price.
* **Dim\_Geography:** Quản lý thông tin địa lý nơi đơn hàng được giao, bao gồm City, State, Country, Region, Market.
* **Dim\_Date (Calendar):** Bảng thời gian chuyên biệt giúp phân tích xu hướng theo Year, Quarter, Month, Day một cách liên tục và chính xác.

**3. Mối quan hệ (Relationships):**

* Hệ thống sử dụng mối quan hệ **Một - Nhiều (1-\*)**.
* **Cơ chế:** Một dòng trong bảng Dimension (ví dụ: 1 Khách hàng) có thể liên kết với nhiều dòng trong bảng Fact (ví dụ: Khách hàng đó mua nhiều Đơn hàng).
* Dòng dữ liệu chảy từ bảng Dimension (1) xuống bảng Fact (\*), cho phép lọc và phân tích số liệu theo bất kỳ chiều nào (ví dụ: Lọc Doanh thu theo Quốc gia hoặc theo Năm).

## Giải thích về bộ dữ liệu khách hàng

### 2.4.1 Các khái niệm

Bộ dữ liệu được sử dụng trong dự án này là **Master E-commerce Dataset** – một tập hợp dữ liệu có cấu trúc (Structured Data) toàn diện về hoạt động thương mại điện tử bán lẻ.

* **Nguồn dữ liệu:** Dữ liệu ghi nhận các giao dịch thực tế từ một hệ thống bán lẻ đa quốc gia, bao gồm cả thông tin chi tiết về khách hàng, sản phẩm và kết quả tài chính của từng đơn hàng.
* **Quy mô và Phạm vi:** Bộ dữ liệu bao phủ các thị trường lớn (Market) như LATAM, Europe, US... với danh mục sản phẩm đa dạng từ hàng tiêu dùng nhanh đến nội thất và thiết bị công nghệ.
* **Đặc điểm nổi bật:** Đây là dữ liệu dạng **"Customer 360"**, nghĩa là nó kết nối được ba chiều thông tin quan trọng:
  1. **Who:** Ai là người mua (Nhân khẩu học).
  2. **What:** Họ mua cái gì (Sản phẩm).
  3. **How:** Hiệu quả giao dịch ra sao (Doanh thu, Lợi nhuận, Đánh giá).

Việc sử dụng bộ dữ liệu này cho phép nhóm thực hiện các phân tích sâu không chỉ về hiệu quả kinh doanh (Sales Performance) mà còn về hành vi tiêu dùng (Consumer Behavior).

### 2.4.1 Các trường dữ liệu

Để khai thác bộ dữ liệu này hiệu quả, dự án tập trung vào 3 khái niệm và nghiệp vụ phân tích cốt lõi sau:

**a. Phân khúc khách hàng (Customer Segmentation)**

* **Khái niệm:** Là quá trình chia tập khách hàng lớn thành các nhóm nhỏ dựa trên các đặc điểm chung để tối ưu hóa chiến lược tiếp cận.
* **Nghiệp vụ trong dự án:** Sử dụng các trường dữ liệu như Segment (Phân khúc tiêu dùng/doanh nghiệp), Age (Độ tuổi), Annual\_Income (Thu nhập) để xác định đâu là nhóm khách hàng mục tiêu mang lại giá trị cao nhất (High-value customers).

**b. Phân tích Lợi nhuận thực tế (Profitability Analysis)**

* **Khái niệm:** Doanh thu (Sales) cao không đồng nghĩa với hiệu quả kinh doanh tốt nếu chi phí vốn (Cost) cao hoặc giảm giá (Discount) quá sâu.
* **Nghiệp vụ trong dự án:** Tính toán và so sánh giữa Doanh thu và Lợi nhuận (Profit) trên từng dòng sản phẩm và từng khu vực địa lý. Việc này giúp phát hiện các sản phẩm bán chạy nhưng đang gây lỗ vốn cho doanh nghiệp.

**c. Mô hình RFM (Recency - Frequency - Monetary)**

* **Khái niệm:** Đây là mô hình tiêu chuẩn trong ngành bán lẻ để đánh giá mức độ trung thành của khách hàng.
* **Nghiệp vụ trong dự án:**
  + **Recency:** Dựa vào Order\_Date để biết lần mua gần nhất.
  + **Frequency:** Dựa vào số lượng Order\_ID để biết tần suất mua.
  + **Monetary:** Dựa vào tổng Sales để biết giá trị chi tiêu.
  + *Mục đích:* Nhận diện nhóm khách hàng VIP và nhóm khách hàng có nguy cơ rời bỏ (Churn).

#### Các trường dữ liệu

Bộ dữ liệu bao gồm các nhóm trường thông tin chính được giải thích chi tiết dưới đây:

**Nhóm 1: Thông tin Định danh & Nhân khẩu học (Customer Profile)** Dùng để trả lời câu hỏi: *"Khách hàng là ai?"*

* **customer\_id**: Mã định danh duy nhất của khách hàng (Primary Key).
* **first\_name, last\_name**: Họ và tên khách hàng.
* **segment**: Phân loại đối tượng khách hàng (Consumer, Corporate, Home Office).
* **gender**: Giới tính (M - Nam, F - Nữ).
* **birth\_date / Age**: Ngày sinh (Dùng để tính độ tuổi khách hàng).
* **annual\_income**: Thu nhập hàng năm (Chỉ số quan trọng đánh giá sức mua).
* **occupation**: Nghề nghiệp hiện tại.
* **education\_level**: Trình độ học vấn.
* **marital\_status**: Tình trạng hôn nhân.
* **home\_owner**: Trạng thái sở hữu nhà (Y/N).

**Nhóm 2: Thông tin Giao dịch & Tài chính (Transaction Data)** Dùng để đo lường hiệu quả kinh doanh.

* **order\_id**: Mã đơn hàng.
* **order\_date**: Ngày thực hiện giao dịch.
* **sales**: Doanh thu của dòng sản phẩm trong đơn hàng.
* **quantity**: Số lượng sản phẩm đặt mua.
* **discount**: Số tiền hoặc tỷ lệ giảm giá áp dụng.
* **profit**: Lợi nhuận ròng (Doanh thu - Giá vốn - Chi phí khác).
* **rating**: Điểm đánh giá sự hài lòng của khách hàng (1-5 sao).

**Nhóm 3: Thông tin Sản phẩm (Product Information)** Dùng để quản lý danh mục hàng hóa.

* **product\_code**: Mã sản phẩm (SKU).
* **product**: Tên chi tiết sản phẩm.
* **category**: Danh mục ngành hàng lớn (Ví dụ: Personal Care).
* **subcategory**: Danh mục con chi tiết (Ví dụ: Skin Care, Lips...).
* **unit\_price**: Giá bán đơn vị.
* **cost\_per\_product**: Giá vốn hàng bán.

**Nhóm 4: Thông tin Địa lý (Location)**

* **market**: Thị trường quản lý (Ví dụ: LATAM, APAC).
* **region**: Khu vực địa lý.
* **country, state, city**: Quốc gia, Tỉnh/Bang, Thành phố giao hàng.

## Làm sạch và chuyển đổi dữ liệu

* *Bao gồm 2 bước:*
  + *Làm sạch dữ liệu*

Mục tiêu là loại bỏ các dữ liệu nhiễu, sai sót hoặc không cần thiết trước khi đưa vào mô hình phân tích.

* **Kiểm tra và thay đổi kiểu dữ liệu (Data Type Validation):**
  + Chuyển đổi các cột order\_date, birth\_date từ dạng Text sang dạng **Date** để thực hiện các hàm thời gian.
  + Đảm bảo các cột tài chính (sales, profit, unit\_price) ở định dạng **Decimal Number** (Số thập phân) hoặc **Currency** để tính toán chính xác.
  + Định dạng cột rating thành số thập phân (Decimal).
* **Xử lý giá trị rỗng (Handling Missing Values):**
  + Kiểm tra độ phủ dữ liệu (Column Quality) của các trường quan trọng như Sales, Customer\_ID.
  + Đối với cột rating (nếu có ô trống), nhóm quyết định giữ nguyên hoặc điền giá trị mặc định là 0 (tùy theo logic kinh doanh đã thống nhất) để không làm sai lệch điểm trung bình.
* **Loại bỏ cột thừa (Removing Irrelevant Columns):**
  + Xóa các cột không phục vụ cho bài toán phân tích để giảm dung lượng file, ví dụ: row\_id (chỉ là số thứ tự dòng), datetime (vì đã có order\_datetime và order\_date), hoặc quantity\_prod (nếu toàn bộ giá trị đều bằng 1).
* **Xử lý trùng lặp (Removing Duplicates):**
  + Kiểm tra trùng lặp dựa trên khóa chính của đơn hàng (order\_id kết hợp product\_code) để đảm bảo doanh thu không bị tính hai lần.
  + *Chuyển đổi dữ liệu*

Mục tiêu là làm giàu dữ liệu (Enrich data) và tái cấu trúc bảng để phục vụ cho Data Modeling (Mô hình sao).

**Tạo cột tính toán mới (Calculated Columns - Feature Engineering):**

**Tính độ tuổi (Age):** Tạo cột mới từ birth\_date bằng công thức: Year(CurrentDate) - Year(birth\_date) để phục vụ phân khúc khách hàng theo độ tuổi.

**Tạo cột Họ và Tên (Full Name):** Ghép hai cột first\_name và last\_name để hiển thị tên đầy đủ trên biểu đồ.

**Phân nhóm đánh giá (Rating Group):** (Tùy chọn) Nhóm các điểm rating thành các mức: "Tệ" (1-2), "Bình thường" (3), "Tốt" (4-5) để dễ filter.

**Chuẩn hóa dữ liệu địa lý:**

Gộp các cột City, State, Country thành một cột Location\_Full (nếu cần) hoặc kiểm tra lỗi chính tả tên thành phố để đảm bảo tính năng Map (Bản đồ) trong Power BI nhận diện đúng tọa độ.

**Tách bảng (Data Normalization):**

Thực hiện kỹ thuật **Reference** hoặc **Duplicate** bảng gốc trong Power Query để tách thành các bảng Dimension (Dim\_Customer, Dim\_Product, Dim\_Location) và bảng Fact (Fact\_Sales) như đã thiết kế ở mục 3.

Sau khi tách, xóa các cột thông tin chi tiết ở bảng Fact (ví dụ: xóa tên khách hàng ở bảng Fact, chỉ giữ lại customer\_id) để tối ưu mô hình.

## Chuẩn bị dữ liệu

Quá trình chuẩn bị dữ liệu cho dự án được thực hiện nhằm đảm bảo dữ liệu thô (Raw Data) từ file CSV chuyển sang trạng thái sẵn sàng để phân tích. Các công việc chính bao gồm:

* **Thu thập:** Tiếp nhận file master\_ecommerce\_dataset\_final.csv từ nguồn cung cấp.
* **Kiểm tra định dạng:** Đảm bảo file sử dụng bảng mã **UTF-8** để không bị lỗi font chữ tiếng Việt (nếu có) hoặc các ký tự đặc biệt trong tên khách hàng nước ngoài.
* **Gia tăng độ phong phú:** Trong phạm vi dự án này, chúng ta sử dụng bộ dữ liệu đơn nhất nhưng sẽ thực hiện kỹ thuật "làm giàu dữ liệu" (Data Enrichment) bằng cách tính toán thêm các cột chỉ số mới (như nhóm tuổi, biên lợi nhuận) ở bước chuyển đổi.

### 2.6.1 Giải pháp lưu trữ dữ liệu

* *So sánh giải pháp nền tảng đám mây (Cloud-based solution) hay* *các ứng dụng tại chỗ (on-premise) .*

#### So sánh giải pháp nền tảng đám mây và ứng dụng tại chỗ

| **Tiêu chí** | **Giải pháp tại chỗ (On-premise)** | **Giải pháp đám mây (Cloud-based)** |
| --- | --- | --- |
| **Định nghĩa** | Dữ liệu lưu trên máy tính cá nhân hoặc server nội bộ của nhóm. | Dữ liệu lưu trên các dịch vụ như Google Drive, OneDrive, AWS S3. |
| **Chi phí** | Thấp (tận dụng phần cứng có sẵn), nhưng tốn phí bảo trì nếu hỏng hóc. | Linh hoạt, thường miễn phí cho dung lượng nhỏ (như dự án này). |
| **Truy cập** | Khó chia sẻ, chỉ truy cập được khi ngồi tại máy. | Truy cập mọi lúc, mọi nơi, hỗ trợ làm việc nhóm tốt. |
| **Bảo mật** | Phụ thuộc vào bảo mật của máy cá nhân. | Được bảo vệ bởi nhà cung cấp dịch vụ lớn. |

#### Quyết định chọn giải pháp cho dự án

* **Giải pháp được chọn:** Kết hợp **Cloud-based (Google Drive)** và **Local Processing**.
* **Lý do:**
  1. **Tính cộng tác:** Nhóm "Nghiện Data" có nhiều thành viên, việc lưu file Master trên Google Drive giúp mọi người cùng truy cập vào một nguồn dữ liệu duy nhất (Single Source of Truth), tránh việc mỗi người dùng một phiên bản cũ/mới khác nhau.
  2. **Chi phí và Hiệu năng:** Dữ liệu dự án (~50MB - 100MB) phù hợp với các gói Cloud miễn phí. Tuy nhiên, việc xử lý (Transform) sẽ thực hiện trên **Power BI Desktop** (Local) để tận dụng sức mạnh RAM của máy tính, giúp thao tác mượt mà hơn.

### 2.6.2 Giải pháp phân bố dữ liệu

* *Nhân bản là một kĩ thuật quan trọng và hữu hiệu trong việc phân bố cơ sở dữ liệu (CSDL) và thực thi các Stored procedure. Đây là chủ đề rất rộng, chỉ tập trung chọn giải pháp trong phạm vi dự án.*

#### Ý nghĩa việc phân bố dữ liệu

Phân bố dữ liệu giúp đảm bảo tính sẵn sàng (Availability) và độ tin cậy (Reliability) của hệ thống. Trong môi trường làm việc nhóm, nó giúp đồng bộ hóa thông tin để khi một thành viên cập nhật dữ liệu làm sạch, các thành viên khác đều nhận được thay đổi đó.

#### Trình bày cách phân bố dữ liệu

Dựa trên quy mô dự án phân tích dữ liệu, nhóm áp dụng kỹ thuật **Nhân bản (Replication)** ở mức độ file nguồn:

**Master Data:** File gốc master\_ecommerce\_dataset\_final.csv được lưu trữ tập trung trên Cloud.

**Local Replicas:** Mỗi thành viên trong nhóm sẽ kết nối Power BI đến file trên Cloud (hoặc tải bản sao về máy để xử lý cache).

**Distribution via Report:** Sau khi hoàn thành Dashboard, giải pháp phân bố thông tin cuối cùng là publish báo cáo lên **Power BI Service (Web)**. Điều này giúp người xem (Giảng viên/Manager) có thể xem báo cáo trên trình duyệt mà không cần cài đặt phần mềm hay copy file dữ liệu gốc

## Làm sạch dữ liệu

* **Tầm quan trọng: "Garbage In, Garbage Out" -** Dữ liệu đầu vào rác thì kết quả phân tích cũng là rác. Làm sạch dữ liệu giúp tăng độ chính xác của các quyết định kinh doanh

### 2.7.1 Các vấn đề ảnh hưởng tới dữ liệu

* Dữ liệu bị thiếu (Missing values/Null).
* Dữ liệu trùng lặp (Duplicates).
* Sai định dạng (Data Type mismatch).
* Dữ liệu ngoại lai (Outliers) - ví dụ: Tuổi = 200, Doanh thu < 0.

**Vấn đề đang tồn tại trong dự án này:** Qua bước kiểm tra sơ bộ (Data Profiling) trên bộ dữ liệu E-commerce, nhóm phát hiện:

* Cột Order\_Date đang được hiểu là dạng Text thay vì Date.
* Các cột địa lý (City, State) có thể chứa các ký tự lạ hoặc khoảng trắng thừa.
* Cột Rating có thể chưa đồng nhất định dạng số thập phân.

### 2.7.2 Các tiêu chí đánh giá chất lượng dữ liệu

Nhóm áp dụng 3 tiêu chí chính:

**Tính đầy đủ (Completeness):** Không được thiếu các trường quan trọng như Sales, Customer\_ID.

**Tính hợp lệ (Validity):** Dữ liệu phải đúng định dạng (VD: Ngày tháng phải là dd/mm/yyyy).

**Tính nhất quán (Consistency):** Tổng doanh thu tính theo đơn hàng phải khớp với tổng chi tiết sản phẩm.

### 2.7.3 Các bước làm sạch dữ liệu

* *Quá trình làm sạch dữ liệu bao gồm một số bước xác định và sửa chữa các mục bị lỗi.*

#### Trình bày các bước làm sạch

Nhóm sử dụng công cụ **Power Query Editor** để thực hiện các bước sau (được lưu vết trong *Applied Steps*):

**Promote Headers:** Đảm bảo dòng đầu tiên của file CSV (row\_id, order\_id,...) được nhận diện là tiêu đề cột.

**Xử lý định dạng (Data Typing):**

Cột order\_date, ship\_date $\rightarrow$ Chuyển sang **Date**.

Cột sales, profit, unit\_price $\rightarrow$ Chuyển sang **Decimal Number** (Số thập phân cố định) để tính tiền tệ.

Cột postal\_code $\rightarrow$ Chuyển sang **Text** (vì không dùng để cộng trừ).

**Kiểm tra giá trị rỗng (Null Handling):**

Kiểm tra cột Rating: Nếu phát hiện ô trống (null), nhóm quyết định thay thế bằng giá trị **0** hoặc giá trị trung bình (tùy ngữ cảnh) để không gây lỗi khi tính toán.

**Loại bỏ cột thừa:** Xóa cột row\_id vì đây là số thứ tự vô nghĩa trong phân tích, giúp giảm dung lượng mô hình.

**Xóa trùng lặp (Remove Duplicates):** Chọn toàn bộ bảng và chọn "Remove Duplicate Rows" dựa trên khóa chính tổng hợp (Order\_ID + Product\_Code) để đảm bảo mỗi dòng là duy nhất.

## Chuyển đổi dữ liệu

Chuyển đổi dữ liệu là bước biến đổi dữ liệu đã làm sạch thành định dạng phù hợp nhất cho việc trực quan hóa.

### 2.8.1 Các trường hợp cần chuyển đổi:

* Khi dữ liệu gốc quá chi tiết, cần nhóm lại (Aggregation).
* Khi cần tính toán các chỉ số KPI không có sẵn trong file gốc.
* Khi cần chuẩn hóa mô hình dữ liệu (Normalization).

### 2.8.2 Kỹ thuật chuyển đổi:

* Calculated Columns: Tạo cột mới dựa trên công thức.
* Splitting Columns: Tách một cột thành nhiều cột.
* Conditional Columns: Tạo cột phân loại dựa trên điều kiện IF/ELSE.

### 2.8.3 Trình bày các phép chuyển đổi trong dự án

Dưới đây là các bước chuyển đổi chi tiết được áp dụng cho file master\_ecommerce\_dataset\_final.csv:

**Bước 1: Tạo cột tính toán độ tuổi (Feature Engineering)**

* *Mục đích:* Phân tích hành vi mua sắm theo lứa tuổi.
* *Thao tác:* Tạo cột Age = Year(DateTime.LocalNow()) - Year([birth\_date]).

**Bước 2: Phân nhóm khách hàng (Binning)**

* *Mục đích:* Gom nhóm độ tuổi để dễ vẽ biểu đồ.
* *Thao tác:* Tạo cột Age\_Group dùng hàm IF:
  + Nếu Age < 30: "Young"
  + Nếu Age 30-50: "Middle-aged"
  + Nếu Age > 50: "Senior"

**Bước 3: Tạo cột Tên đầy đủ**

* *Mục đích:* Hiển thị tên khách hàng đẹp hơn trên báo cáo.
* *Thao tác:* Merge cột first\_name và last\_name với dấu cách ở giữa -> Cột Full\_Name.

**Bước 4: Chuẩn hóa mô hình (Star Schema Conversion)**

*Mục đích:* Tối ưu hiệu suất báo cáo.

*Thao tác:*

Tạo bảng Dim\_Customer: Chỉ giữ lại các cột customer\_id, name, segment, city... sau đó xóa trùng lặp trên customer\_id.

Tạo bảng Dim\_Product: Chỉ giữ lại product\_code, category, unit\_price...

Tạo bảng Fact\_Sales: Giữ lại các khóa ngoại (customer\_id, product\_code) và các cột số liệu (sales, quantity, profit).

# Xử lý dữ liệu

## Chuẩn hóa dữ liệu

* *Chuẩn hóa dữ liệu:*
  + *Power Query*
  + *Chỉnh sửa table, column, data type*

**Chuẩn hóa dữ liệu giúp dữ liệu sạch**, đồng nhất và có cấu trúc phù hợp để mô hình hóa và viết DAX.  
Dự án sử dụng **Power Query** của Power BI để thực hiện các hoạt động:

* Chỉnh sửa bảng (table)
* Xử lý và chuẩn hóa từng cột (column)
* Xác định và sửa kiểu dữ liệu (data type)
* Loại bỏ lỗi và chuẩn hóa text

### 3.1.1 Trình bày các bước chuẩn hóa trong dự án

* *Trình bày cụ thể các bước chuẩn hóa dữ liệu bằng Power Query tác động cụ thể đến bộ dữ liệu dự án (từng cột, từng dòng).*
* *Việc này giúp lưu vết các chỉnh sửa thay đổi trên.*

**(1) Promote Headers – Chuẩn hóa tên cột**

* Chuyển dòng đầu tiên thành tiêu đề cột.
* Chỉnh tên cột theo chuẩn không dấu, không khoảng trắng.  
  Ví dụ:
  + “Order Date” → order\_date
  + “Customer ID” → customer\_id

Mục tiêu: thống nhất tên cột, dễ đọc và đúng chuẩn BI.

**(2) Chuẩn hóa kiểu dữ liệu (Data Type)**

Kiểm tra từng cột bằng Column Profile và chỉnh lại đúng loại dữ liệu:

| **Cột** | **Kiểu đúng** | **Lý do** |
| --- | --- | --- |
| Order\_date | Date | Tạo Calendar, phân tích theo thời gian |
| order\_datetime | DateTime | Phân tích theo ngày – giờ |
| sales | Decimal Number | Tính KPI |
| profit | Decimal Number | Tính lợi nhuận |
| discount | Decimal Number | Tính tỷ lệ giảm giá |
| rating | Whole Number | Tính điểm đánh giá |
| annual\_income | Decimal Number | Phân tích thu nhập |
| customer\_id, product\_id | Text | Tránh Power BI tự convert số |

**(3) Xử lý trùng lặp (Remove Duplicates)**

Kiểm tra bằng Column Quality → Remove duplicates dựa trên:

**order\_id + product\_id**

**→ đảm bảo không nhân đôi doanh thu và lợi nhuận.**

## Mô hình hóa dữ liệu

* *Mô hình hóa dữ liệu:*
  + *Data model*
  + *Data modeling*

Sau khi chuẩn hóa, dữ liệu được đưa vào Model View của Power BI để mô hình hóa.

Mục tiêu mô hình hóa:

* Sắp xếp dữ liệu theo quan hệ
* Tối ưu hiệu năng
* Hỗ trợ viết DAX và xây dựng Dashboard
* Giảm sự phụ thuộc giữa các bảng

Dự án áp dụng **Star Schema (mô hình hình sao)**

### 3.2.1 Các loại mô hình hóa

* *Trình bày các loại mô hình*
* *Mô hình nào đang được áp dụng cho bộ dữ liệu dự án ?*

| **Loại mô hình** | **Mô tả** | **Đặc điểm** |
| --- | --- | --- |
| Flat Table | Mọi dữ liệu trong 1 bảng | Chậm, khó phân tích |
| Snowflake Schema | Nhiều dimension phân tách thêm | Cấu trúc phức tạp |
| Star Schema | 1 fact trung tâm liên kết đến các dimension | Nhanh – Chuẩn BI – Tối ưu |

**→ Dự án sử dụng: STAR SCHEMA ( Ngôi Sao)**

Vì:

* Dễ mở rộng
* Dễ filter và drill-down
* Tối ưu tốc độ xử lý
* Phù hợp với thiết kế KPI trong Dashboard Blueprint

### 3.2.2 Các tiêu chí đánh giá mô hình dữ liệu

* *Trình bày các tiêu chí đánh giá mô hình dữ liệu tốt ?*
* *Mô hình đang dùng trong dự án đáp ứng các tiêu chí nào ?*

**Một mô hình tốt phải đáp ứng:**

1. Đơn giản – trực quan
2. Hiệu năng cao
3. Không có vòng lặp (No Circular Relationship)
4. Quan hệ 1–\* rõ ràng
5. Single Direction Filtering
6. Tách Fact – Dimension rạch ròi

**→ Mô hình hiện tại đáp ứng đầy đủ 6 tiêu chí.**

### 3.3.3 Trình bày các bước mô hình hóa

* *Trình bày cụ thể các bước mô hình hóa dữ liệu bằng Power BI trên bộ dữ liệu của dự án.*
* *Việc này giúp lưu vết các chỉnh sửa thay đổi trên.*

**Bước 1: Reference bảng nguồn để tách bảng**

Từ bảng gốc → Reference → tạo các Dimension:

* d\_customer
* d\_product
* d\_region
* d\_segment
* d\_date

**Bước 2: Xử lý từng Dimension**

Ví dụ:

**a. d\_customer**

Giữ lại:

* customer\_id
* first\_name, last\_name
* gender
* annual\_income
* birth\_date
* education\_level
* home\_owner

Remove duplicates theo customer\_id.

**b. d\_product**

Giữ lại:

* product\_id
* product
* category
* subcategory
* cost\_per\_product
* unit\_price

Xóa trùng product\_id.

**c. d\_region**

Giữ:

* region\_id
* region
* country
* state
* city
* market

Chuẩn hóa text theo bước 3.1.

**d. d\_segment**

* SegmentID
* segment

**Bước 3: Tạo bảng Fact f\_order**

Giữ lại:

* order\_id
* customer\_id
* product\_id
* region\_id
* segment
* order\_date
* order\_datetime
* quantity
* sales
* discount
* profit
* rating

Loại bỏ các cột đã chuyển sang Dimension.

**Bước 4:** Tạo mối quan hệ giữa các bảng

A screenshot of a computer

AI-generated content may be incorrect.

### 3.3.4 Trình bày các bước tạo bảng dữ liệu

* *Trình bày cụ thể các bước tạo bảng dữ liệu bằng Power BI trên bộ dữ liệu của dự án.*
* *Việc này giúp lưu vết các chỉnh sửa thay đổi trên.*

**(1) Import CSV vào Power BI**

→ Power Query Editor

**(2) Chuẩn hóa dữ liệu theo mục 3.1**

**(3) Tách bảng từ bảng gốc bằng Reference**

* Tạo bảng Dimension
* Loại bỏ cột không cần thiết

**(4) Remove duplicates trong từng bảng Dimension**

**(5) Đặt Key cho từng Dimension**

* customer\_id
* product\_id
* region\_id
* SegmentID

**(6) Tạo Fact và giữ các cột giao dịch**

**(7) Load vào Model View**

**(8) Tạo Relationships dạng 1–\***

**(9) Tạo bảng Measure để chứa KPI**

## Xử lý dữ liệu DAX

* *Xử lý dữ liệu bằng ngôn ngữ DAX*
  + *Measure*
  + *Calculated column*
  + *Filter*

### 3.3.1 Measure

* *Trình bày các measure được dùng trong dự án.*
* *Việc này giúp lưu vết các chỉnh sửa thay đổi trên.*

#### Tạo calendar

…

#### Tạo số lượng

…

#### Tạo số lượng

…

#### Tạo measure tiếp theo ……

…

### 3.3.2 Calculated column

* *Trình bày các calculated column được dùng trong dự án.*
* *Việc này giúp lưu vết các chỉnh sửa thay đổi trên.*

#### Tạo cột

…

#### Tạo cột tiếp theo ….

…

### 3.3.3 Filter

* *Trình bày các bộ lọc filter được dùng trong dự án.*
* *Việc này giúp lưu vết các chỉnh sửa thay đổi trên.*

#### Tạo filter chọn

…

#### Tạo filter chọn

…

#### Tạo filter chọn Ngày

…

# Trực quan hóa dữ liệu

## Các kỹ thuật trực quan hóa

* *Trình bày các kỹ thuật trực quan hóa*
* *Kỹ thuật nào đang được áp dụng cho dự án ? Vì sao ?*

…

## Các nguyên tắc trực quan hóa

* *Trình bày các nguyên tắc trong trực quan hóa*
* *Nguyên tắc nào bạn cho rằng quan trọng nhất trong phạm vi dự án này ? Vì sao ?*

…

## Trình bày cách thêm visual mới

* *Các visual trong Power BI đôi khi không đáp ứng được nhu cầu nghiệp vụ mới.*
* *Thêm visual mới là điều tất yếu. Trình bày các bước thêm visual trong dự án này.*
* *Khuyến khích tìm visual mới phục vụ cho business của bộ dữ liệu.*

…

## Trình bày tạo các report cho dự án

* *Tạo các bộ lọc, biểu đồ theo nhu cầu dự án.*
* *Có thể phân chia theo đối tượng trong mô hình.*
* *Có thể phân chia theo chức năng hoặc page.*
* *Trong dự án cho môn học này thì phân chia theo page.*

#### Tạo visual thống kê chi tiết

* *Tạo các visual dùng trong trang thống kê chi tiết.*
* *Việc này giúp lưu vết các chỉnh sửa thay đổi trên.*

#### Tạo visual filter theo

…

#### Tạo visual filter theo

…

#### Tạo visual filter theo

…

#### Tạo visual tiếp theo ….

…

#### Tạo visual thống kê tổng thể

* *Tạo các visual dùng trong trang thống kê tổng thể.*
* *Việc này giúp lưu vết các chỉnh sửa thay đổi trên.*

#### Tạo visual filter theo ngày giao dịch

…

#### Tạo visual thống kê Tổng GTGD

…

#### Tạo visual thống kê Tổng KLGD

…

#### Tạo visual tiếp theo ….

…

# Xây dựng báo cáo

## Dashboard và report

* *Cách tối ưu hóa:*
  + *Dashboard*
  + *Report*

…

## Xây dựng báo cáo

* *Xây dựng báo cáo:*
  + *Dashboard*
  + *Report*

#### Dashboard vs Report

* *Trình bày sự khác nhau giữa Dashboard và Report.*
* *Giải thích vì sao làm dashboard, report trong dự án này ?*

…

*Giải thích:* …

#### Dashboard

* *Trình bày các loại dashboard có trong dự án.*
* *Giải thích vì sao làm các dashboard này ?*

…

*Giải thích:* …

#### Report

* *Trình bày các loại report có trong dự án.*
* *Giải thích vì sao làm các report này ?*

#### Tạo report Detail

*Giải thích:* …

…

#### Tạo report Overview

*Giải thích:* …

…

#### Bookmark

* *Trình bày các loại bookmark có trong dự án.*
* *Giải thích vì sao làm các bookmark này ?*

#### Tạo bookmark slicer

*Giải thích:* ….

#### Tạo bookmark slicer chọn

…

*Giải thích:* …

#### Tạo bookmark slicer chọn

…

*Giải thích:* …

#### Tạo visual tiếp theo ….

…

*Giải thích:* …

# KẾT LUẬN

## Báo cáo

#### Các bước viết báo cáo

* *Trình bày các bước viết báo cáo phân tích dữ liệu*

…

#### Tổng hợp

* *Tổng hợp các câu trả lời trong Câu Chuyện Dữ Liệu ở Phân Tích Khách Hàng, kết hợp với các bước viết báo cáo phân tích, để tạo thành bản báo cáo hoàn chỉnh.*

*Gợi ý:*

…

## Khó khăn

##### Thuận lợi

## Hướng phát triển