**HỌC VIỆN NGÂN HÀNG**

**KHOA HỆ THỐNG THÔNG TIN QUẢN LÝ**

A picture containing text, logo, font, symbol

Description automatically generated

**ĐỀ TÀI: XÂY DỰNG HỆ HỖ TRỢ RA QUYẾT ĐỊNH CHO**

**PHÂN HỆ BÁN HÀNG TẠI CÔNG TY NORTHWIND**

**Giảng viên hướng dẫn: Cô Ngô Thùy Linh**

**NHÓM 9**

**MÃ HỌC PHẦN: 211IS31A02**

|  |  |
| --- | --- |
| **Nhóm sinh viên thực hiện:** | |
| Nguyễn Thị Mai Anh | 23A4040005 – K23HTTTA |
| Lê Hà Chi | 23A4040018 – K23HTTTA |
| Nguyễn Thị Duyên | 23A4040025 – K23HTTTA |
| Đặng Thị Hạnh | 23A4040034 – K23HTTTA |
| Phan Thị Quỳnh Trang | 23A4040149 – K23HTTTA |

**DANH SÁCH SINH VIÊN THỰC HIỆN**

|  |  |  |  |
| --- | --- | --- | --- |
| **Họ và tên** | **Mã sinh viên** | **Công việc** | **Đóng góp** |
| Nguyễn Thị Mai Anh | 23A4040005 | Câu 2, Câu 3, Câu 6 | 20% |
| Lê Hà Chi | 23A4040018 | Câu 1, Câu 4, Câu 7 | 20% |
| Nguyễn Thị Duyên | 23A4040025 | Câu 3, Câu 4, Câu 7, Câu 8 | 20% |
| Đặng Thị Hạnh | 23A4040034 | Câu 2, Câu 3, Câu 4, Câu 5, Câu 8 | 20% |
| Phan Thị Quỳnh Trang | 23A4040149 | Câu 4, Câu 5, Câu 7 | 20% |

**NHẬN XÉT CỦA GIẢNG VIÊN**

**LỜI CẢM ƠN**

Đầu tiên, em xin gửi lời cảm ơn tới ThS.Ngô Thùy Linh – Giảng viên khoa Hệ thống thông tin Quản lý – Học viện ngân hàng đã trực tiếp giảng dạy, hướng dẫn tận tình chúng em bộ môn này. Kiến thức mà cô truyền đạt cùng những cuốn sách mà cô giới thiệu đã giúp nhóm em hiểu rõ hơn về các khái niệm, phương pháp và công cụ hỗ trợ ra quyết định trong kinh doanh. Nhóm đã học được cách đọc tài liệu, thu thập, phân tích dữ liệu, xây dựng các báo cáo vận hành, báo cáo thống kê cùng rất nhiều công cụ khác mà cô giới thiệu. Ngoài ra, cô cũng luôn sẵn sàng giải đáp thắc mắc của sinh viên một cách nhiệt tình và tận tâm.

Môn học Hệ hỗ trợ ra quyết định và kinh doanh thông minh sẽ mang lại cho nhóm em rất nhiều kiến thức và kinh nghiệm quý báu. Chúng em tin rằng những kiến thức này sẽ giúp ích rất nhiều trong công việc và cuộc sống sau này.

Mặc dù trong quá trình tìm hiểu gặp nhiều khó khăn do kiến thức, kinh nghiệm còn hạn chế nhưng nhóm đã cố gắng hoàn thành. Vì vậy, nhóm rất mong có thể nhận được những ý kiến đóng góp của cô để bài báo cáo có thể hoàn thiện một cách tốt hơn.

**MỤC LỤC**

[DANH MỤC VIẾT TẮT 1](#_Toc153916547)

[DANH MỤC HÌNH ẢNH 2](#_Toc153916548)

[DANH MỤC BẢNG BIỂU 8](#_Toc153916549)

[Câu 1. Trình bày bối cảnh công ty mà nhóm đang tìm hiểu: Kinh doanh sản phẩm dịch vụ gì? Đó là một chuỗi cửa hàng hay chỉ một chi nhánh? Bán hàng online hay offline? (0.5 điểm) 9](#_Toc153916550)

[Câu 2. Mô tả usecase của nhóm: có mấy phân hệ nghiệp vụ ở mức vận hành (có mấy nguồn), ở các phòng ban nào? Trình bày mô hình cơ sở dữ liệu ở các hệ thống nghiệp vụ này? (1 điểm) 9](#_Toc153916551)

[Câu 3: Thiết kế 1 số báo cáo vận hành trong ngày (operational reports): câu hỏi và đáp án (1 điểm) 11](#_Toc153916552)

[3.1. Báo cáo những đơn hàng chưa hoàn thành 11](#_Toc153916553)

[3.2. Báo cáo tổng doanh thu theo ngày 12](#_Toc153916554)

[3.3. Số lượng đơn hàng trong ngày 12](#_Toc153916555)

[3.4 Doanh thu theo danh mục sản phẩm/sản phẩm theo ngày 13](#_Toc153916556)

[3.5. Top 10 sản phẩm bán chạy nhất 15](#_Toc153916557)

[3.6. Báo cáo tình trạng đơn hàng 15](#_Toc153916558)

[Câu 4: Liệt kê (dạng câu hỏi) một số báo cáo để phân tích thống kê số liệu mà bộ phận quản lý cần? Nếu chỉ dựa vào phân hệ vận hành thì có đáp ứng được không? Giải thích (1 điểm) 16](#_Toc153916559)

[4.1. Câu hỏi 16](#_Toc153916560)

[4.2. Nhận xét 19](#_Toc153916561)

[Câu 5. Phát biểu sự cần thiết xây dựng Kho dữ liệu trong bối cảnh này? Mô tả kiến trúc kho dữ liệu dự kiến của nhóm? (1 điểm) 22](#_Toc153916562)

[5.1. Sự cần thiết xây dựng Kho dữ liệu 22](#_Toc153916563)

[5.2. Mô tả kiến trúc kho dữ liệu dự kiến 25](#_Toc153916564)

[Câu 6. Xác định nghiệp vụ mà nhóm quan tâm. Xác định độ chi tiết của dữ liệu - dựa vào cơ sở nào để đưa ra được mức độ chi tiết này (1 điểm) 31](#_Toc153916565)

[Câu 7. Thiết kế các bảng chiều (bảng cắt lớp) DIM, giải thích tại sao lại có các bảng DIM này, mô tả chi tiết từng bảng, nguồn của bảng DIM lấy từ bảng nào? (2 điểm) 33](#_Toc153916566)

[Câu 8. Thiết kế bảng FACT. Mô hình lược đồ trên Data mart? Mô hình sao hay bông tuyết? (2 điểm) 45](#_Toc153916567)

[8.1. Cấu trúc bảng Fact\_Sales 45](#_Toc153916568)

[8.2. Ý nghĩa trường dữ liệu 47](#_Toc153916569)

[8.3. Mô hình lược đồ trên Data mart, mô hình sao 47](#_Toc153916570)

[Câu 9: Đổ dữ liệu từ các bảng nguồn vào các bảng DIM, minh họa bằng hình ảnh việc đổ dữ liệu thành công (2 điểm) 51](#_Toc153916571)

[9.1. Bảng Dim\_Customer 52](#_Toc153916572)

[9.2. Bảng Dim\_Product 57](#_Toc153916573)

[9.3. Bảng Dim\_Location 60](#_Toc153916574)

[9.4. Bảng Dim\_Date 66](#_Toc153916575)

[Câu 10: Đổ dữ liệu vào bảng Fact, minh họa đổ dữ liệu thành công (2 điểm) 71](#_Toc153916576)

[Câu 11: Sử dụng kiểu SCD nào để tải dữ liệu (data loading) từ hệ thống NGUỒN vào bảng DIM? Minh họa (1 điểm) 78](#_Toc153916577)

[Câu 12: Thiết kế các báo cáo phân tích thống kê (tương ứng với câu 4 trong bài kiểm tra 1) sử dụng bất cứ tool nào (SQL Server Business Intelligence, Tableau, Power BI, Olap Excel,...) để minh họa các thao tác roll-up; drill-down; slice; dice (2 điểm) 84](#_Toc153916581)

[12.1. Thống kê số lượng đơn hàng theo khu vực địa lý theo sự kiện 84](#_Toc153916582)

[12.2. Thống kê doanh số theo sản phẩm, theo khu vực và thời gian. 88](#_Toc153916583)

[12.3. Thống kê tổng doanh số theo khách hàng, theo thời gian và theo danh mục sản phẩm Bằng Excel 92](#_Toc153916584)

[12.4. Mức tăng trưởng doanh số của sản phẩm theo khu vực so với cùng kỳ năm trước bằng Excel 95](#_Toc153916585)

[12.5. Thống kê tổng doanh thu và tổng số tiền giảm giá theo sản phẩm và theo thời gian bằng Power BI 97](#_Toc153916586)

[12.6. Thống kê doanh thu của khách hàng theo sản phẩm, theo khu vực và theo thời gian 103](#_Toc153916587)

[Câu 13: Sau quá trình thử nghiệm cài đặt hệ thống và khai thác dữ liệu trên Cube, nhóm hãy nêu một số đánh giá về hiệu quả của DW về: khả năng truy vấn, khai thác dữ liệu, ... ? (0.5 điểm). Loại DB nào phù hợp để thiết kế DW? Giải thích? (0.5 điểm) 108](#_Toc153916588)

[13.1. Khả năng truy vấn, khai thác dữ liệu trên Data warehouse 108](#_Toc153916589)

[13.2. Loại Database phù hợp xây dựng Data warehouse 111](#_Toc153916590)

[Câu 14: Khi dữ liệu từ NGUỒN được tải vào Data mart (tại bảng Dim, Fact) thì nhóm em có sử dụng kỹ thuật học máy, học sâu, luật kết hợp... để xây dựng mô hình dự báo, dự đoán (trên tập dữ liệu ở data mart) hỗ trợ ra quyết định mức chiến lược, chiến thuật không? (1 điểm) 114](#_Toc153916591)

[14.1. Tổng quan về Machine Learning (Học máy) 114](#_Toc153916592)

[14.2. Mô hình hồi quy tuyến tính (Linear Regression) 117](#_Toc153916593)

[14.3. Tính cấp thiết trong xây dựng mô hình dự báo số lượng sản phẩm tiêu thụ theo vùng trong thời điểm tới 118](#_Toc153916594)

[14.4. Áp dụng thuật toán Linear Regression để dự báo xu hướng tiêu thụ sản phẩm theo vùng trong thời điểm tới 119](#_Toc153916595)

[Câu 15: Nhóm hãy phát huy trí tưởng tượng và thiết kế kịch bản để case study mà nhóm đang nghiên cứu hướng đến việc triển khai data lake: nêu lý do/thách thức (làm thế nào gây xúc động cho người nghe là được) và trình bày giải pháp (sử dụng dịch vụ nào của nhà cung cấp nào) có minh họa kiến trúc bằng hình ảnh, lợi ích (1 điểm) 128](#_Toc153916596)

[15.1. Lý do đề xuất triển khai Data lake? 128](#_Toc153916597)

[15.3. Giá trị của Data Lake 134](#_Toc153916598)

[15.4. Những thách thức với Data Lake 135](#_Toc153916599)

[15.5. Giải pháp triển khai Data Lake 137](#_Toc153916600)

[15.5.1. Kiến trúc Data lake 137](#_Toc153916601)

[15.5.2. Triển khai các Hồ dữ liệu trên đám mây 138](#_Toc153916602)

[15.5.3. Xây dựng Data Lakes của Northwind trên đám mây với AWS 139](#_Toc153916603)

[TÀI LIỆU THAM KHẢO 144](#_Toc153916604)

# DANH MỤC VIẾT TẮT

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Từ viết tắt** | **Nguyên nghĩa** | **Chú thích** |
| 1 | ODS | Operational Data Store | Kho dữ liệu hoạt động |
| 2 | OLAP | Online analytical processing | Xử lý phân tích trực tuyến |
| 3 | DIM | Dimension |  |
| 4 | OLTP | Online transactional processing | Xử lý giao dịch trực tuyến |
| 5 | POS | Point of Sale | Điểm bán lẻ như cửa hàng tạp hóa |
| 6 | CRM | Customer Relationship Management | Quản lý quan hệ khách hàng |
| 7 | EDW | Enterprise data warehouse | Kho dữ liệu doanh nghiệp |
| 8 | MS | Microsoft |  |
| 9 | SCD | Slowly Changing Dimensions |  |
| 10 | BI | Business Intelligence | Kinh doanh thông minh |
| 11 | R&D | Research and Development | Nghiên cứu và phát triển |
| 12 | CSDL |  | Cơ sở dữ liệu |

# DANH MỤC HÌNH ẢNH

[Hình 2. 1: Diagram của bộ dữ liệu Northwind 10](#_Toc151277175)

[Hình 3.1: Báo cáo những đơn hàng chưa hoàn thành 11](#_Toc151300686)

[Hình 3.2: Báo cáo tổng doanh thu theo ngày trong tháng 5 năm 1998 12](#_Toc151300687)

[Hình 3.3: Báo cáo số lượng đơn hàng theo ngày trong tháng 5 năm 1998 13](#_Toc151300688)

[Hình 3.4: Báo cáo doanh thu theo sản phẩm trong ngày 1 tháng 5 năm 1998 14](#_Toc151300689)

[Hình 3.5: Báo cáo doanh thu theo danh mục sản phẩm trong ngày 1 tháng 5 năm 1998 14](#_Toc151300690)

[Hình 3.6: Top 10 sản phẩm bán chạy nhất trong ngày 1 tháng 5 năm 1998 15](#_Toc151300691)

[Hình 3.7: Báo cáo tình trạng đơn hàng trong tháng 5 năm 1998 16](#_Toc151300692)

[Hình 5.1: Minh họa khó khăn khi lập báo cáo 23](#_Toc151300704)

[Hình 5.2: Giải pháp 23](#_Toc151300705)

[Hình 5.3: Các báo cáo trong kinh doanh 25](#_Toc151300706)

[Hình 5.4: Cấu trúc cơ bản Data Warehouse 26](#_Toc151300707)

[Hình 5.5: Khung kiến trúc Kho dữ liệu doanh nghiệp 28](#_Toc151300708)

[Hình 5.6: Khung kiến trúc Kho dữ liệu hoạt động 28](#_Toc151300709)

[Hình 5.7: Khung kiến trúc xây dựng Data Mart 29](#_Toc151300710)

[Hình 5.8: Cách tiếp cận từ dưới lên của Kimball 30](#_Toc151300711)

[Hình 6.1: Cây phân cấp dữ liệu của công ty Northwind 32](#_Toc151300755)

[Hình 7.1: Sơ đồ bán lẻ sơ bộ Northwind 34](#_Toc151300762)

[Hình 7.2: Các thuộc tính bảng Dim\_Date 37](#_Toc151300763)

[Hình 7.3: Thuộc tính bảng Dim\_Customer 40](#_Toc151300764)

[Hình 7.4: Thuộc tính bảng Dim\_Product 42](#_Toc151300765)

[Hình 7.5: Thuộc tính bảng Dim\_Location 44](#_Toc151300766)

[Hình 8.1: Bảng Fact\_Sales 46](#_Toc151300767)

[Hình 8.2: Mô hình sao Northwind 48](#_Toc151300768)

[Hình 8.3: Star schema versus OLAP cube 49](#_Toc151300769)

[Hình 9.1: Tạo kết nối database 51](#_Toc153923125)

[Hình 9.2: Kết nối với nguồn 52](#_Toc153923126)

[Hình 9.3: Tạo SSIS Packages 52](#_Toc153923127)

[Hình 9.4: Tạo dataflow bảng dim\_customer 53](#_Toc153923128)

[Hình 9.5: Lấy nguồn từ bảng Customers 53](#_Toc153923129)

[Hình 9. 6: Liên kết các thuộc tính liên quan SCD2 54](#_Toc153923130)

[Hình 9.7: Liên kết các thuộc tính liên quan SCD2 54](#_Toc153923131)

[Hình 9.8: Chọn date\_from, date\_to 55](#_Toc153923132)

[Hình 9.9: Kết thúc SCD2 55](#_Toc153923133)

[Hình 9.10: DataFlow bảng Dim\_Customer 56](#_Toc153923134)

[Hình 9.11: Đổ thành công bảng Dim\_Customers 56](#_Toc153923135)

[Hình 9.12: Dữ liệu Dim\_Customer sau khi đổ thành công 57](#_Toc153923136)

[Hình 9.13: Lấy nguồn dữ liệu đổ bảng Dim\_Products 57](#_Toc153923137)

[Hình 9.14: Tạo kiểu SCD 2 Dim\_Products 58](#_Toc153923138)

[Hình 9.15: Chọn thuộc tính thay đổi khi cập nhật SCD2 58](#_Toc153923139)

[Hình 9.16: DataFlow bảng Dim\_Products 59](#_Toc153923140)

[Hình 9.17: Đổ dữ liệu Dim\_Product thành công 59](#_Toc153923141)

[Hình 9.18: Dữ liệu Dim\_Products khi đổ thành công 60](#_Toc153923142)

[Hình 9.19: Chọn thuộc tính địa chỉ 61](#_Toc153923143)

[Hình 9.20: Tạo Snipt Component bảng dim\_location 61](#_Toc153923144)

[Hình 9.21: Tạo Snipt Component bảng dim\_location 62](#_Toc153923145)

[Hình 9.22: Thêm code vào Snipt 62](#_Toc153923146)

[Hình 9.23: Tạo SCD2 bảng dim\_location 63](#_Toc153923147)

[Hình 9.24: Thêm dữ liệu cần cập nhật SCD2 bảng dim\_location 63](#_Toc153923148)

[Hình 9.25: DataFlow bảng dim\_location 64](#_Toc153923149)

[Hình 9.26: Đổ dữ liệu thành công dim\_location 65](#_Toc153923150)

[Hình 9.27: Dữ liệu bảng dim\_location sau khi đổ 66](#_Toc153923151)

[Hình 9.28: Dữ liệu bảng Dim\_Date 71](#_Toc153923152)

[Hình 10.1: Thêm nguồn bảng Fact\_sales 72](#_Toc153926663)

[Hình 10.2: Lookup với bảng Order 72](#_Toc153926664)

[Hình 10.3: Chọn những thuộc tính cần lấy ở bảng Orders 72](#_Toc153926665)

[Hình 10.4: Lookup tới bảng dim\_customer 73](#_Toc153926666)

[Hình 10.5: Chọn những thuộc tính cần lấy bảng dim\_customer 73](#_Toc153926667)

[Hình 10.6: Lookup tới bảng dim\_products 73](#_Toc153926668)

[Hình 10.7: Chọn những thuộc tính cần lấy bảng dim\_products 74](#_Toc153926669)

[Hình 10.8: Lookup tới bảng dim\_date 74](#_Toc153926670)

[Hình 10.9: Chọn những thuộc tính cần lấy bảng dim\_date 75](#_Toc153926671)

[Hình 10.10: Lookup tới bảng dim\_location 75](#_Toc153926672)

[Hình 10.11: Chọn những thuộc tính bảng dim\_location 76](#_Toc153926673)

[Hình 10.12: Xây dựng công thức tính doanh thu 76](#_Toc153926674)

[Hình 10.13: Đích bảng fact\_sales 76](#_Toc153926675)

[Hình 10.14: Nối các thuộc tính 77](#_Toc153926676)

[Hình 10.15: DataFlow bảng fact\_sales 77](#_Toc153926677)

[Hình 10.16: Đổ thành công bảng fact\_sales 78](#_Toc153926678)

[Hình 10.17: Dữ liệu bảng fact\_sales sau khi đổ 78](#_Toc153926679)

[Hình 11.1: Thuộc tính bảng dim\_customer 80](#_Toc153928114)

[Hình 11.2: Cập nhật bảng nguồn Customers 80](#_Toc153928115)

[Hình 11.3 Kiểm tra cập nhật SCD2 dim\_customers 81](#_Toc153928116)

[Hình 11.4: Kết quả dim\_customer 81](#_Toc153928117)

[Hình 11.5: Thuộc tính bảng dim\_products 81](#_Toc153928118)

[Hình 11.6: Cập nhật bảng nguồn Products 82](#_Toc153928119)

[Hình 11.7: Kiểm tra cập nhật SCD2 dim\_products 82](#_Toc153928120)

[Hình 11.8: Kết quả dim\_products 82](#_Toc153928121)

[Hình 11.9: Cập nhật dữ liệu bảng Customers 83](#_Toc153928122)

[Hình 11.10: Kiểm tra cập nhật SCD2 dim\_location 83](#_Toc153928123)

[Hình 11.11: Kết quả dim\_location 84](#_Toc153928124)

[Hình 12.1: Biểu đồ tổng số lượng đơn hàng theo quốc gia theo sự kiện 85](#_Toc153928149)

[Hình 12.2. Biểu đồ tổng số lượng đơn hàng theo thành phố 86](#_Toc153928150)

[Hình 12.3: Biểu đồ tổng số lượng đơn hàng ở Italy, France, Germany trong ngày Lễ giáng sinh và Lễ tạ ơn 87](#_Toc153928151)

[Hình 12.4: Biểu đồ tổng số lượng đơn hàng tại Germany theo sự kiện cụ thể 88](#_Toc153928152)

[Hình 12.5: Biểu đồ tổng doanh số sản phẩm theo quốc gia theo năm 89](#_Toc153928153)

[Hình 12.6. Biểu đồ tổng doanh số sản phẩm theo thành phố trong các năm 90](#_Toc153928154)

[Hình 12.7: Biểu đồ tổng doanh số sản phẩm theo thành phố trong năm 1998 91](#_Toc153928155)

[Hình 12.8: Biểu đồ tổng doanh số sản phẩm theo thành phố Aachen, Brandenburg, Cunewalde của Đức trong năm 1996 92](#_Toc153928156)

[Hình 12.9: Biểu đồ doanh số của khách hàng theo từng thành phố 92](#_Toc153928157)

[Hình 12.10: Biểu đồ doanh số của khách hàng theo từng quốc gia 93](#_Toc153928158)

[Hình 12.11: Biểu đồ tổng doanh số theo từng năm 93](#_Toc153928159)

[Hình 12.12: Biểu đồ tổng doanh số theo từng quý 94](#_Toc153928160)

[Hình 12.13: Biểu đồ tổng doanh số theo từng quốc gia 94](#_Toc153928161)

[Hình 12.14: Biểu đồ tổng doanh số theo quốc gia Arghentina 94](#_Toc153928162)

[Hình 12.15: Biểu đồ tổng doanh số theo sản phẩm Chai ở quốc gia Canada 95](#_Toc153928163)

[Hình 12.16. Biểu đồ mức tăng của doanh số theo quốc gia từng năm 95](#_Toc153928164)

[Hình 12.17: Biểu đồ mức tăng của doanh số theo thành phố từng quý 96](#_Toc153928165)

[Hình 12.18: Biểu đồ tổng doanh thu và tổng số tiền giảm giá theo sản phẩm 97](#_Toc153928166)

[Hình 12.19: Biểu đồ tổng doanh thu và tổng số tiền giảm giá theo danh mục sản phẩm 98](#_Toc153928167)

[Hình 12.20: Biểu đồ tổng doanh thu và tổng số tiền giảm giá theo danh mục sản phẩm 99](#_Toc153928168)

[Hình 12.21: Biểu đồ tổng doanh thu và tổng số tiền giảm giá theo sản phẩm 99](#_Toc153928169)

[Hình 12.22: Biểu đồ tổng doanh thu và tổng số tiền giảm giá theo sản phẩm 100](#_Toc153928170)

[Hình 12.23: Biểu đồ tổng doanh thu và tổng số tiền giảm giá của sản phẩm Chai 101](#_Toc153928171)

[Hình 12.24: Biểu đồ tổng doanh thu và tổng số tiền giảm giá của sản phẩm theo quý 102](#_Toc153928172)

[Hình 12.25: Biểu đồ tổng doanh thu và tổng số tiền giảm giá của sản phẩm Chai trong quý III 103](#_Toc153928173)

[Hình 12.26: Biểu đồ doanh thu của khách hàng theo từng thành phố 103](#_Toc153928174)

[Hình 12.27: Biểu đồ doanh thu của khách hàng theo từng quốc gia 104](#_Toc153928175)

[Hình 12.28: Biểu đồ doanh thu của khách hàng theo từng quý 104](#_Toc153928176)

[Hình 12.29: Biểu đồ doanh thu của khách hàng theo từng tháng 105](#_Toc153928177)

[Hình 12.30: Biểu đồ doanh thu của khách hàng theo sản phẩm, theo khu vực và theo thời gian 106](#_Toc153928178)

[Hình 12.31: Biểu đồ doanh thu của khách hàng ở Argentina 106](#_Toc153928179)

[Hình 12.32: Biểu đồ doanh thu của khách hàng theo từng sản phẩm ở từng quốc gia 107](#_Toc153928180)

[Hình 12.33: Biểu đồ doanh thu của các khách hàng mua sản phẩm Chai ở Sao Paulo trong Quý I 108](#_Toc153928181)

[Hình 13.1: Sự khác biệt giữa CSDL dạng cột và dạng hàng 113](#_Toc153928182)

[Hình 14.1: Dữ liệu nhóm thu thập được 120](#_Toc153928201)

[Hình 14.2: Top 10 sản phẩm có số lượng bán cao nhất trong tháng 5 năm 1998 124](#_Toc153928202)

[Hình 15.1: Minh họa 1 129](#_Toc153928207)

[Hình 15.2: Minh họa 2 130](#_Toc153928208)

[Hình 15.3: Minh họa 3 131](#_Toc153928209)

[Hình 15.4: Minh họa 4 132](#_Toc153928210)

[Hình 15.5: Kiến trúc Data Lake cơ bản 133](#_Toc153928211)

[Hình 15.6: Data Swamp (Đầm lấy dữ liệu) 136](#_Toc153928212)

[Hình 15.7: Kiến trúc Data lake 137](#_Toc153928213)

[Hình 15.8: Kiến trúc và quy trình xây dựng Data Lake trên AWS Cloud 140](#_Toc153928214)

# DANH MỤC BẢNG BIỂU

[Bảng 4.1: Bảng câu hỏi bộ phận quản lý cần để thực hiện báo cáo phân tích thống kê 19](#_Toc151300876)

[Bảng 7.1: Mô tả chi tiết bảng Dim date 36](#_Toc151300889)

[Bảng 7.2: Mô tả chi tiết bảng Dim\_Customer 39](#_Toc151300890)

[Bảng 7.3: Mô tả chi tiết bảng Dim\_Product 41](#_Toc151300891)

[Bảng 7.4: Mô tả chi tiết bảng Dim\_Location 44](#_Toc151300892)

[Bảng 8.1: Mô tả chi tiết bảng Fact\_sales 47](#_Toc151300897)

[Bảng 13.1: So sánh truy vấn dữ liệu ở đích và nguồn 109](#_Toc153928848)

# Câu 1. Trình bày bối cảnh công ty mà nhóm đang tìm hiểu: Kinh doanh sản phẩm dịch vụ gì? Đó là một chuỗi cửa hàng hay chỉ một chi nhánh? Bán hàng online hay offline? (0.5 điểm)

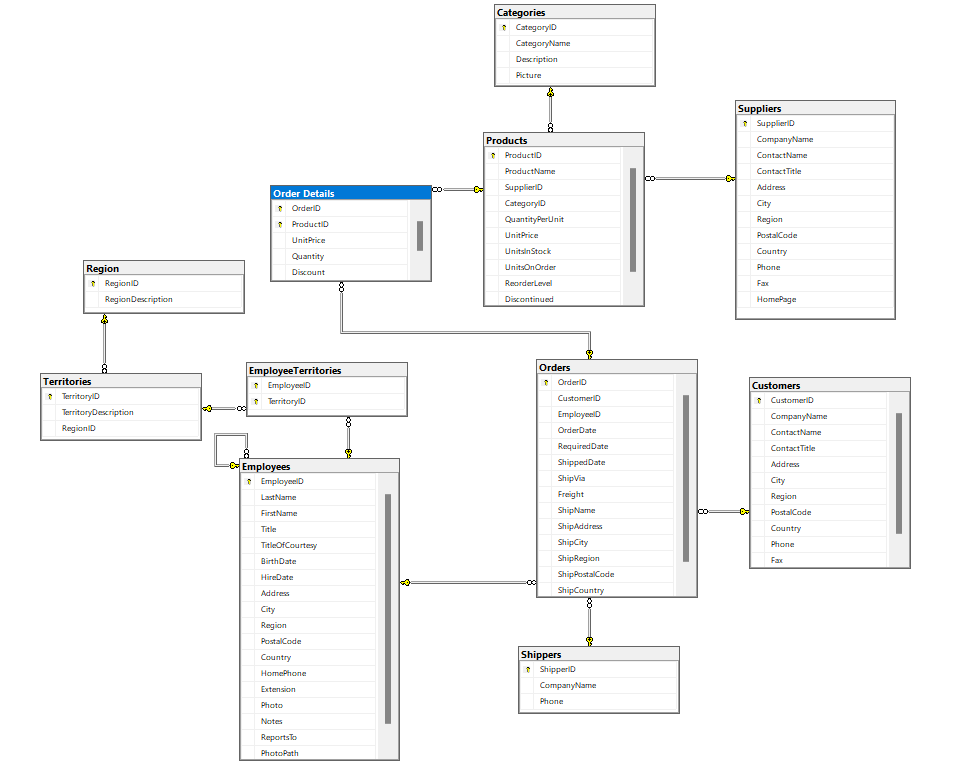
Northwind là công ty cung cấp thực phẩm cho các công ty và doanh nghiệp, gồm một trụ sở duy nhất bán hàng theo hai hình thức online và offline.

# Câu 2. Mô tả usecase của nhóm: có mấy phân hệ nghiệp vụ ở mức vận hành (có mấy nguồn), ở các phòng ban nào? Trình bày mô hình cơ sở dữ liệu ở các hệ thống nghiệp vụ này? (1 điểm)

Cơ sở dữ liệu Northwind chứa các bảng và dữ liệu liên quan đến các khía cạnh khác nhau trong hoạt động bán hàng của một công ty nhập xuất sản phẩm. Tuy nhiên, cơ sở dữ liệu Northwind không có sự phân chia các phân hệ rõ ràng ở cấp độ vận hành. Trong Northwind DB có đủ các bảng của một bài toán bán hàng, ngoài ra còn Nhân viên (Employees), Danh mục sản phẩm (Categories), Nhà cung cấp (Suppliers), Dịch vụ chuyển phát (Shippers)…

* Gồm các nguồn dữ liệu trong SQL Server:
* Các dữ liệu về hóa đơn bán hàng (hóa đơn bán, chi tiết hóa đơn) thuộc bộ phận kinh doanh.
* Các dữ liệu về sản phẩm (nhà cung cấp, danh mục) thuộc bộ phận sản xuất.
* Các dữ liệu về khách hàng (khu vực) thuộc bộ phận chăm sóc khách hàng.
* Các dữ liệu về nhân viên bán hàng thuộc bộ phận quản lý nhân sự.

Các mô hình cơ sở dữ liệu ở hệ thống nghiệp vụ này: Mô hình cơ sở dữ liệu quan hệ, được tổ chức ở dạng các bảng như sau:



Hình 2.1: Diagram của bộ dữ liệu Northwind

Cấu trúc cơ sở dữ liệu của Northwind gồm những thông tin sau:

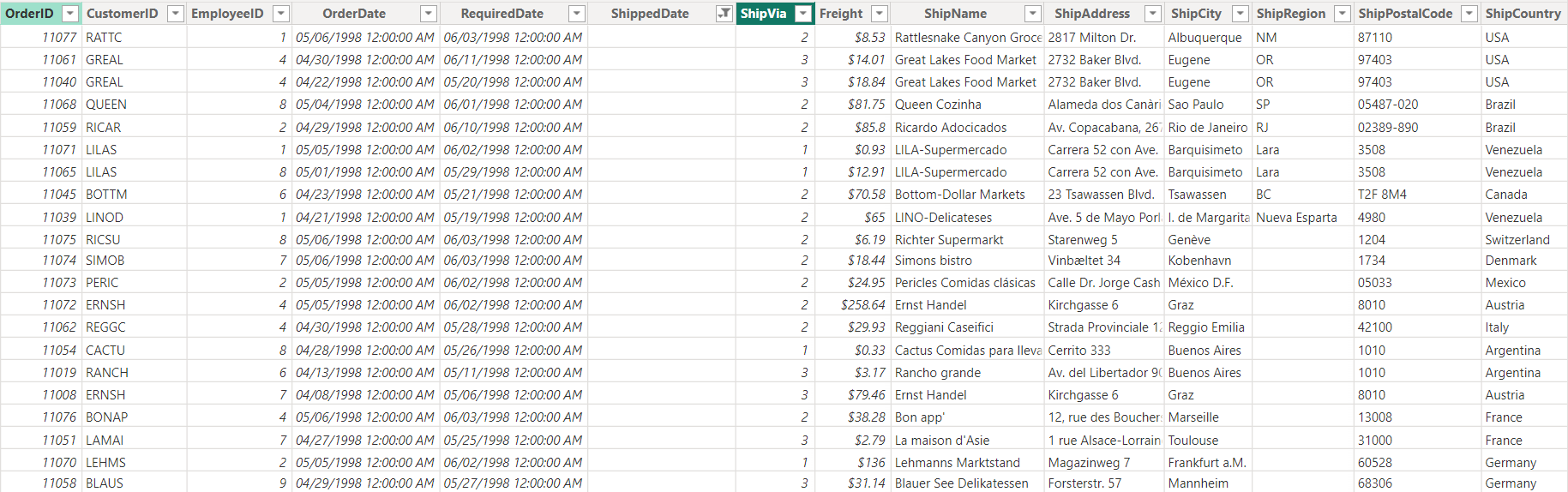
* Về khách hàng:
  + Customers: thông tin về các khách hàng của công ty
* Nhóm nghiệp vụ bán hàng:
  + Oders: thông tin về các đơn đặt hàng từ khách hàng
  + Oders\_details: thông tin về chi tiết đơn hàng
  + Products: thông tin về sản phẩm
  + Categories: thông tin về danh mục của sản phẩm
* Về nhà cung cấp:
  + Suppliers: thông tin về các nhà cung cấp sản phẩm cho công ty
* Về nhân viên:
  + Employees: thông tin về nhân viên xử lý đơn hàng
  + Territories: thông tin về các lãnh thổ mà nhân viên quản lý
  + Region: lưu trữ thông tin về khu vực địa lý mà nhân viên quản lý
  + Employee\_territories: Thông tin hỗ trợ cho quan hệ nhiều nhiều của employees với territories

# Câu 3: Thiết kế 1 số báo cáo vận hành trong ngày (operational reports): câu hỏi và đáp án (1 điểm)

Một số báo cáo vận hành được nhóm thiết kế và xây dựng trên phần mềm MS Power BI:

## 3.1. Báo cáo những đơn hàng chưa hoàn thành

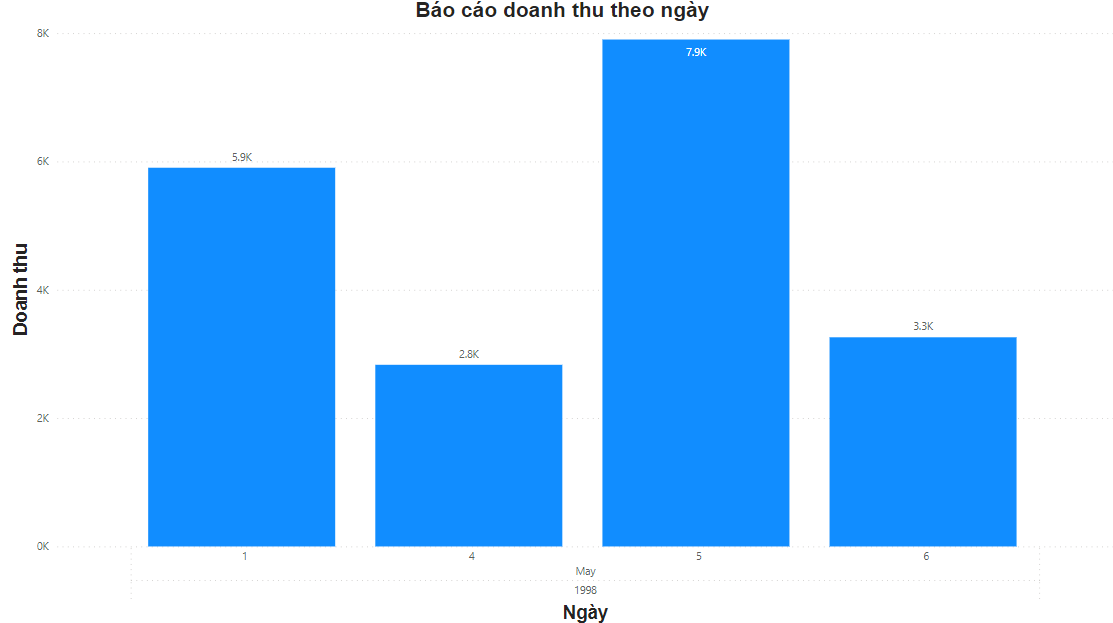
Báo cáo này giúp doanh nghiệp theo dõi được tiến độ thực hiện của từng đơn hàng, từ đó có thể đưa ra các biện pháp điều chỉnh kịp thời để đảm bảo đơn hàng được hoàn thành đúng thời hạn. Thường được lập theo định kỳ, thường là hàng tuần, hàng tháng hoặc hàng quý.

******

Hình 3.1: Báo cáo những đơn hàng chưa hoàn thành

## 3.2. Báo cáo tổng doanh thu theo ngày

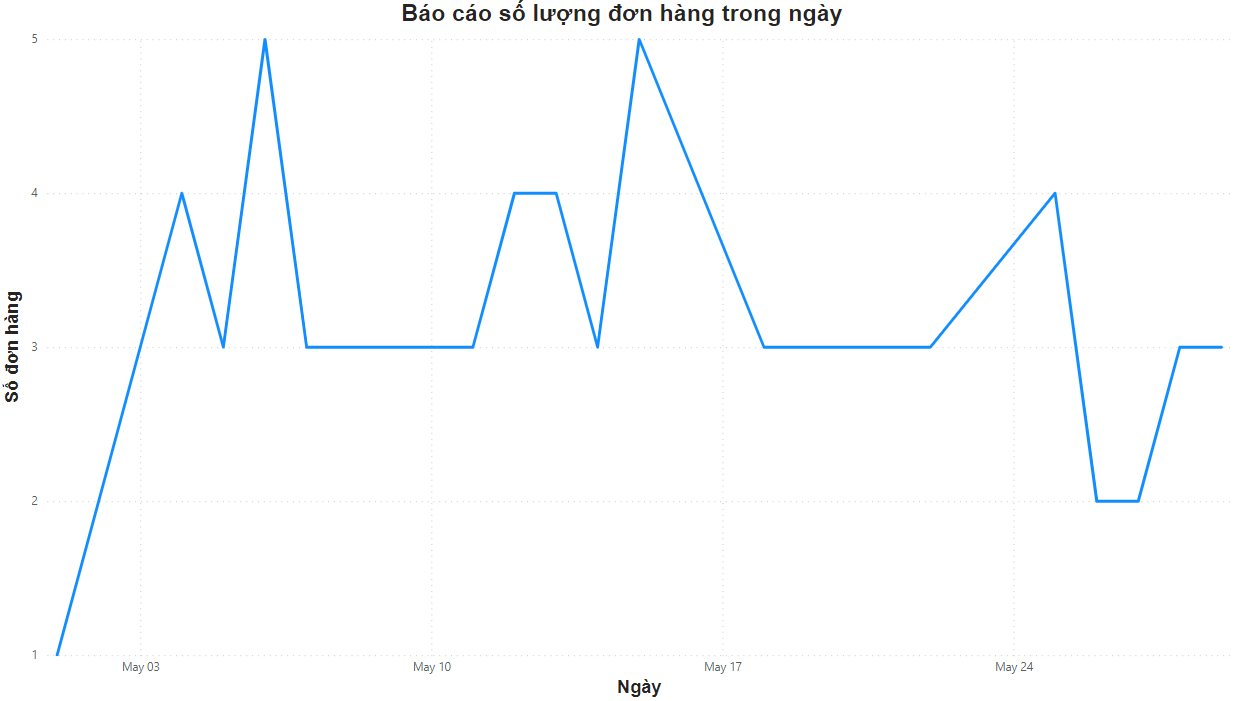
Báo cáo này cung cấp cho doanh nghiệp thông tin về tổng doanh thu của doanh nghiệp trong một ngày cụ thể. Từ đó, doanh nghiệp có thể đánh giá được hiệu quả của hoạt động kinh doanh trong ngày, kịp thời phát hiện và xử lý các vấn đề phát sinh. Thường được lập theo định kỳ, thường là hàng ngày.

******

Hình 3.2: Báo cáo tổng doanh thu theo ngày trong tháng 5 năm 1998

## 3.3. Số lượng đơn hàng trong ngày

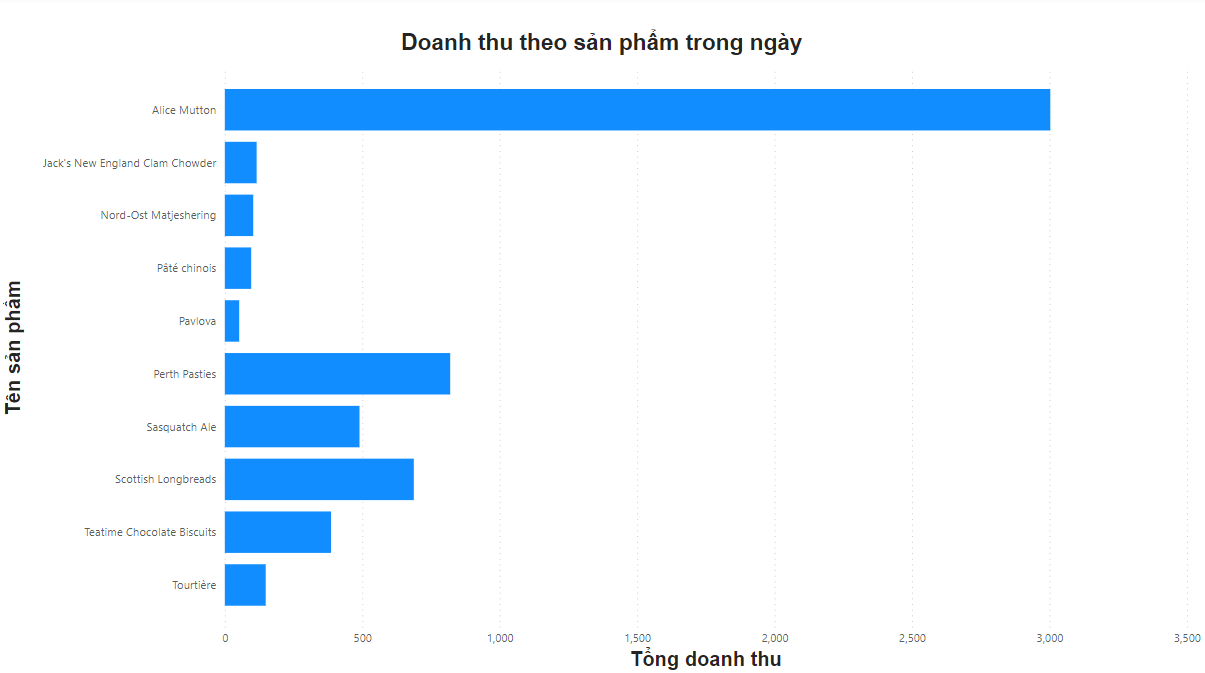
Dựa trên nội dung báo cáo, doanh nghiệp có thể lập kế hoạch sản xuất, kinh doanh phù hợp với nhu cầu của thị trường. Nó giúp doanh nghiệp xác định các cơ hội kinh doanh tiềm năng, chẳng hạn như các sản phẩm, dịch vụ mới được khách hàng yêu thích.

******

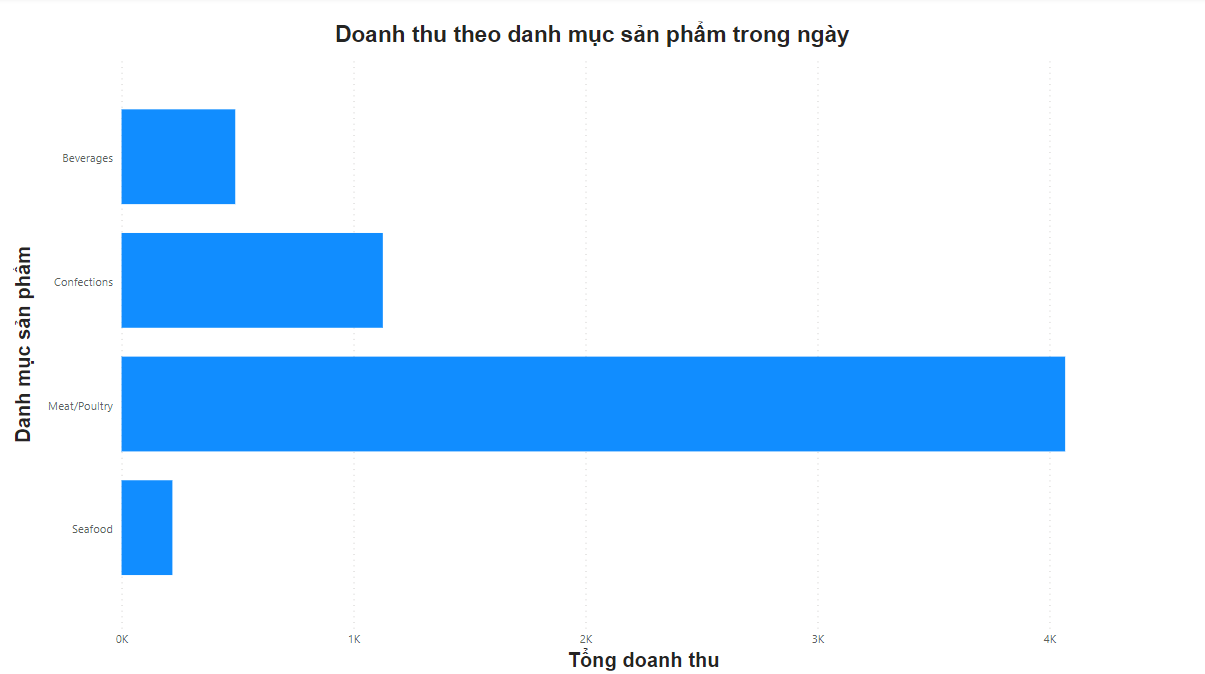
Hình 3.3: Báo cáo số lượng đơn hàng theo ngày trong tháng 5 năm 1998

## 3.4 Doanh thu theo danh mục sản phẩm/sản phẩm theo ngày

Báo cáo này cung cấp cho doanh nghiệp thông tin về doanh thu sản phẩm/dịch vụ bán ra của từng danh mục sản phẩm/sản phẩm. Từ đó, doanh nghiệp có thể đánh giá được hiệu quả của hoạt động bán hàng của từng danh mục sản phẩm/sản phẩm, kịp thời phát hiện và xử lý các vấn đề phát sinh. Đây là một công cụ quan trọng giúp doanh nghiệp nắm bắt tình hình bán hàng, phân tích hiệu quả hoạt động kinh doanh và lập kế hoạch kinh doanh theo danh mục sản phẩm/sản phẩm.

******

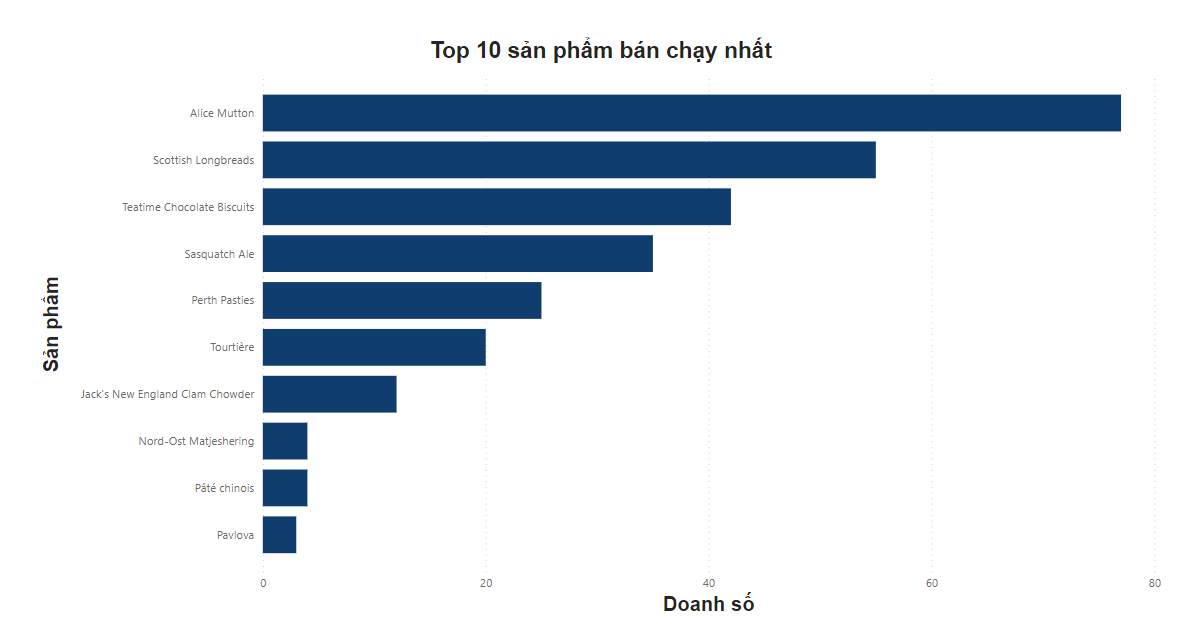
Hình 3.4: Báo cáo doanh thu theo sản phẩm trong ngày 1 tháng 5 năm 1998

******

Hình 3.5: Báo cáo doanh thu theo danh mục sản phẩm trong ngày 1 tháng 5 năm 1998

## 3.5. Top 10 sản phẩm bán chạy nhất

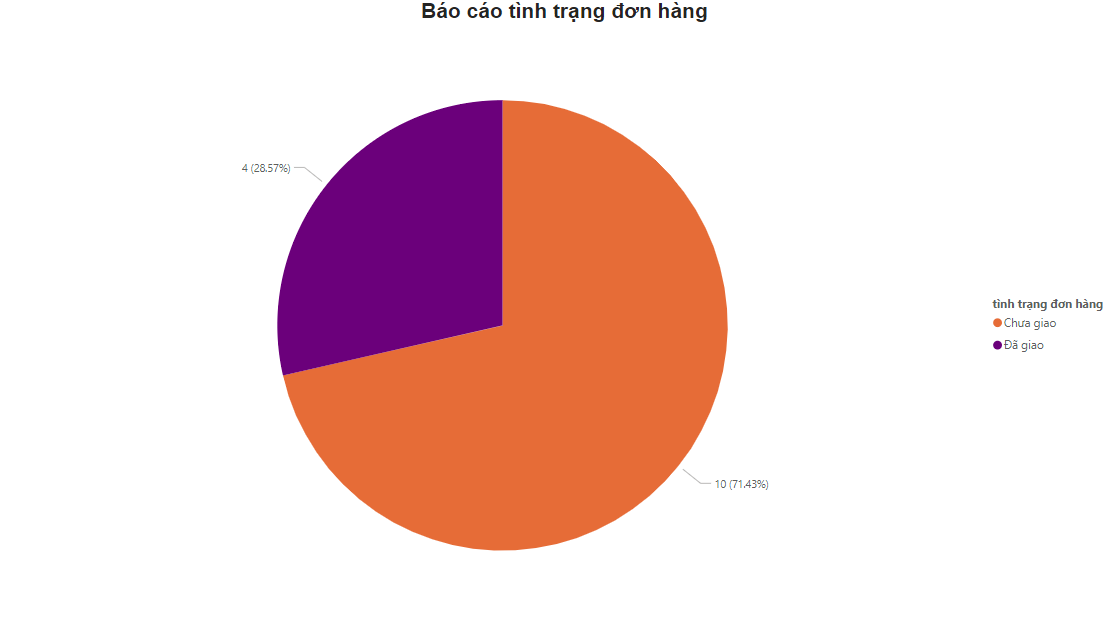
Báo cáo Top 10 sản phẩm bán chạy nhất trong ngày giúp doanh nghiệp nắm bắt tình hình bán hàng, xác định được những sản phẩm mang lại doanh thu cao, từ đó có thể tập trung nguồn lực để phát triển các sản phẩm này.

******

Hình 3.6: Top 10 sản phẩm bán chạy nhất trong ngày 1 tháng 5 năm 1998

## 3.6. Báo cáo tình trạng đơn hàng

Báo cáo này giúp doanh nghiệp theo dõi được tiến độ giao hàng của từng đơn hàng, từ đó có thể đưa ra các biện pháp điều chỉnh kịp thời để đảm bảo đơn hàng được giao đúng thời hạn và xử lý khiếu nại của khách hàng.



Hình 3.7: Báo cáo tình trạng đơn hàng trong tháng 5 năm 1998

# Câu 4: Liệt kê (dạng câu hỏi) một số báo cáo để phân tích thống kê số liệu mà bộ phận quản lý cần? Nếu chỉ dựa vào phân hệ vận hành thì có đáp ứng được không? Giải thích (1 điểm)

## 4.1. Câu hỏi

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Bộ phận** | **STT** | **Nội dung câu hỏi** | **Tầm quan trọng** | **Tính khả thi** |
| Quản lý bán hàng | 1 | Thống kê số lượng đơn hàng theo khu vực địa lý trong một sự kiện cụ thể. | Cao | Khả thi |
| 2 | Thống kê doanh số theo sản phẩm, theo khu vực và thời gian. | Cao | Khả thi |
| 3 | Dự báo xu hướng tiêu thụ sản phẩm theo vùng trong thời điểm tới. | Cao | Không khả thi |
| 4 | Mức tăng trưởng doanh số của sản phẩm theo khu vực so với cùng kỳ năm trước. | Cao | Khả thi |
| 5 | Báo cáo lợi nhuận theo thời gian, khu vực địa lý, danh mục sản phẩm. | Cao | Không khả thi |
| 6 | Thống kê tổng doanh số theo khách hàng, theo thời gian và theo danh mục sản phẩm. | Cao | Khả thi |
| 7 | Thống kê tổng doanh thu và tổng số tiền giảm giá theo sản phẩm và theo thời gian. | Cao | Khả thi |
| 8 | Thống kê doanh thu khách hàng theo sản phẩm, theo khu vực và theo thời gian. | Cao | Khả thi |
| 9 | Số tiền bán hàng trung bình theo đơn đặt hàng và theo thời gian. | Cao | Khả thi |
| Quản lý Marketing | 1 | Doanh số của công ty sau từng chiến dịch marketing. | Cao | Không khả thi |
| 2 | Tỷ lệ bán hàng trên các kênh online so với offline theo ngày, tuần, tháng, quý hoặc năm. | Cao | Không khả thi |
| 3 | Tỷ lệ chuyển đổi từ khách ghé thăm cửa hàng trở thành người mua hàng. | Cao | Không khả thi |
| 4 | Các chính sách kích cầu mua sắm ảnh hưởng như thế nào đến doanh thu? | Cao | Không khả thi |
| Quản lý nhân sự | 1 | Tổng số nhân viên hiện tại và biến động trong quý 3 năm nay? | Trung bình | Khả thi |
| 2 | Tỷ lệ tuyển dụng thành công trong năm nay (số lượng ứng viên tuyển dụng thành công chia cho tổng số ứng viên đã phỏng vấn)? | Trung bình | Không khả thi |
| 3 | Tỷ lệ giữ chân nhân viên trong năm nay (số lượng nhân viên ở cuối năm chia cho số lượng nhân viên ở đầu năm)? | Cao | Khả thi |
| 4 | Số lượng nhân viên nghỉ việc trong quý 3 năm nay và lý do nghỉ việc chính? | Trung bình | Không khả thi |
| 5 | Tỷ lệ phát triển nghề nghiệp trong năm nay (số lượng nhân viên được đào tạo phát triển chia cho tổng số nhân viên)? | Cao | Không khả thi |
| 6 | Tỷ lệ nhân viên được thăng chức trong năm nay so với năm trước? | Cao | Khả thi |
| 7 | Tỷ lệ sử dụng chính sách và quy trình nhân sự trong năm nay (ví dụ: tỷ lệ sử dụng chính sách nghỉ phép, tỷ lệ sử dụng chính sách thưởng, tỷ lệ hoàn thành đánh giá hiệu suất)? | Trung bình | Không khả thi |
| 8 | Phản hồi và đánh giá từ nhân viên về môi trường làm việc, chính sách nhân sự và cơ hội phát triển trong năm nay? | Cao | Không khả thi |

Bảng 4.1: Bảng câu hỏi bộ phận quản lý cần để thực hiện báo cáo phân tích thống kê

## 4.2. Nhận xét

Nếu chỉ dựa vào phân hệ vận hành của Northwind - cơ sở dữ liệu liên quan đến nghiệp vụ bán hàng thì không thể đáp ứng được những báo cáo trên. Những câu hỏi bên trên nhằm xây dựng các báo cáo theo tính định kỳ, phân chia rõ ràng các mốc thời gian, địa điểm… để thuận tiện phân tích, báo cáo. Trong khi đó, phân hệ vận hành của Northwind chỉ cung cấp các thông tin cơ bản về các giao dịch, đơn hàng, hóa đơn, khách hàng nên sẽ cần những câu lệnh truy vấn phức tạp, thiếu thông tin cần thiết và khó có thể lập ra những báo cáo trên được.

Chẳng hạn như muốn trả lời câu hỏi *Thống kê doanh số bán hàng theo sản phẩm, khu vực địa lý và thời gian* (câu 2 Bảng 4.1)*,* ta cần sử dụng dữ liệu thời gian trong bảng Đơn hàng (thông qua thuộc tính Ngày đặt hàng) và được câu lệnh SQL như sau:

--Doanh số sản phẩm năm 1997 tại Oulu

WITH doanh\_so\_theo\_san\_pham AS (

select O.ProductID, SUM(O.Quantity) as Doanh\_so

from OrderDetails O join Orders S on O.OrderID = S.OrderID

where S.OrderDate between '1997-01-01' and '1997-12-31' and ShipCity = 'Oulu'

group by ProductID

)

SELECT \*

FROM doanh\_so\_theo\_san\_pham

*Thống kê doanh thu khách hàng theo sản phẩm, theo khu vực và theo thời gian.* (câu 8 Bảng 4.1) cần có dữ liệu thời gian để tổng hợp, bộ dữ liệu Northwind chỉ lưu trữ dữ liệu thời gian trong bảng Đơn hàng (thông qua thuộc tính Ngày đặt hàng), để lập báo cáo theo tháng thì cần tổng hợp doanh thu theo từng tháng. Nếu sử dụng truy vấn trong cơ sở dữ liệu ta có lệnh SQL sau:

--Doanh thu sản phẩm theo khu vực Oulu trong tháng 6 năm 1997

WITH doanh\_thu\_san\_pham AS (

SELECT ProductID,

UnitPrice,

Quantity,

UnitPrice \* Quantity AS Doanh\_thu

FROM OrderDetails O

JOIN Orders S ON O.OrderID = S.OrderID

WHERE ShipCity = 'Oulu' and OrderDate between '1997-06-01' and '1997-06-30'

GROUP BY ProductID,

UnitPrice,

Quantity

)

SELECT \*

FROM doanh\_thu\_san\_pham

Từ những câu truy vấn trên, nhóm thấy có thể thực hiện được những câu SQL theo nhiều chiều trên hệ thống vận hành, tuy nhiên trong quá trình truy vấn có gặp khó khăn sau:

* **Câu truy vấn phức tạp**: JOIN nhiều bảng để lấy nguồn dữ liệu thời gian, khu vực, sản phẩm.
* **Tốn thời gian, giảm hiệu suất truy vấn**: Do kích thước dữ liệu lớn
* **Khó khăn trong việc quản lý và bảo trì các truy vấn:** Các truy vấn dữ liệu theo nhiều chiều thường phức tạp và liên quan đến nhiều bảng dữ liệu và các mối quan hệ phức tạp. Điều này có thể khiến việc quản lý và bảo trì các truy vấn này trở nên phức tạp và tốn thời gian khi dữ liệu cập nhật.

Hiện tại, khi tạo báo cáo tổng hợp doanh thu từ *hoạt động bán hàng* online, offline, doanh nghiệp vẫn đang phải tích hợp dữ liệu từ nhiều nguồn khác nhau như: các file excel, bộ lưu trữ đám mây, cơ sở dữ liệu quan hệ… một cách thủ công. Điều này gây tốn thời gian và nhiều sai sót thậm chí phức tạp và khó có thể thực hiện được.

Trong trường hợp các báo cáo phải lập một cách thủ công cũng gây khó khăn trong quá trình phân tích mối quan hệ giữa doanh thu và chi phí *marketing* khiến việc ra quyết định về chiến dịch marketing chậm trễ và khó đo lường được hiệu quả và rủi ro.

Về vấn đề *quản lý nhân sự*, cơ sở dữ liệu thông thường không lưu trữ lịch sử dữ liệu nhân sự và phân tích xu hướng theo thời gian khiến việc theo dõi xu hướng và dự đoán về nhân sự trở nên khó khăn và hạn chế.

Do đó chỉ với những truy vấn trong cơ sở dữ liệu vận hành khó có thể phục vụ cho công ty trong quá trình tổng hợp báo cáo, phân tích và quá trình ra quyết định kinh doanh của mình.

# Câu 5. Phát biểu sự cần thiết xây dựng Kho dữ liệu trong bối cảnh này? Mô tả kiến trúc kho dữ liệu dự kiến của nhóm? (1 điểm)

## 5.1. Sự cần thiết xây dựng Kho dữ liệu

**Vấn đề đặt ra:** Liệu chỉ sử dụng OLTP (Online Transactional Processing) trên Database thì có thể đáp ứng tốt trong quá trình ra quyết định của doanh nghiệp hay không?

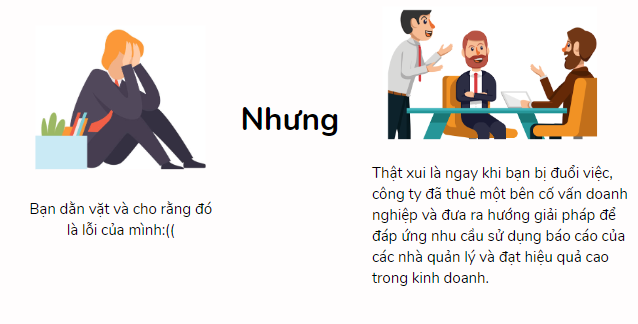
Cơ sở dữ liệu sử dụng Xử lý giao dịch trực tuyến (OLTP) chỉ phù hợp với mô hình dữ liệu, kích thước dữ liệu và truy vấn đơn giản. Trong trường hợp sử dụng OLTP, doanh nghiệp chỉ tạo được báo cáo vận hành thông qua việc chạy các truy vấn dữ liệu.

Là một công ty xuất nhập khẩu thực phẩm lớn trong và ngoài nước, Northwind gồm nhiều các hệ thống con phục vụ quá trình kinh doanh bao gồm: hệ thống kế toán - tài chính, hệ thống bán hàng online, hệ thống POS, phần mềm CRM, hệ thống quản lý nhân sự. Dữ liệu từ các hệ thống này được lưu trữ thành nhiều nguồn khác nhau, có cấu trúc và định dạng khác nhau cùng với khối lượng giao dịch lớn lên đến hàng nghìn bản ghi mỗi ngày khiến việc tích hợp dữ liệu trở lên khó khăn và tốn thời gian. Trong quá trình tích hợp dữ liệu có thể gây sai sót do lỗi nhập liệu, lỗi hệ thống hoặc các yếu tố khác khiến các báo cáo không chính xác hoặc báo cáo có tính linh hoạt thấp, khó thay đổi theo yêu cầu của người dùng dẫn đến không thể khai thác được các giá trị của dữ liệu và gây ảnh hưởng tới việc ra quyết định. Dựa trên một câu chuyện có thật!



Hình 5.1: Minh họa khó khăn khi lập báo cáo

(Nguồn: Nhóm tự design)



Hình 5.2: Giải pháp

(Nguồn: Nhóm tự design)

Và giải pháp mà bên cố vấn đưa ra ở đây đó là xây dựng hệ thống Data Warehouse để lưu trữ tất cả các thông tin được tổng hợp từ nhiều nguồn, phục vụ dễ dàng cho việc quản lý dữ liệu, xây dựng các báo cáo phân tích kinh doanh giúp cho việc ra quyết định được diễn ra nhanh chóng và chính xác hơn.

Từ câu chuyện trên có thể thấy với nhu cầu tạo báo cáo từ những truy vấn phức tạp không dễ để OLTP trên cơ sở dữ liệu có thể xử lý. Đồng thời quá trình trả lại dữ liệu cho các báo cáo phân tích kinh doanh này gây ra vấn đề tranh chấp tài nguyên trên cơ sở dữ liệu do phải thực hiện quá nhiều tác vụ cùng một lúc (lưu trữ và truy xuất dữ liệu phức tạp).

Vậy rõ ràng OLTP không còn phù hợp với tiêu chí của Northwind hiện nay. Do đó doanh nghiệp cần phải thay đổi mô hình dữ liệu sao cho phù hợp nhất với các truy vấn cần thiết cho việc báo cáo và phân tích kinh doanh. Trong trường hợp này, nhóm đề xuất xây dựng kho dữ liệu sử dụng Xử lý phân tích trực tuyến (OLAP) để giải quyết những vấn đề trên của doanh nghiệp.

**Một số lợi ích của Data Warehouse:**

#1: Lưu trữ dữ liệu ở một nguồn duy nhất

Khi tất cả dữ liệu của bạn được lưu trữ và khai thác trong một kho lưu trữ duy nhất, việc quản lý và bảo trì sẽ loại bỏ các vấn đề về chất lượng dữ liệu, dữ liệu không nhất quán, báo cáo dữ liệu không chính xác và hiệu suất truy vấn thấp.



Hình 5.3: Các báo cáo trong kinh doanh

(Nguồn: beads.net)

#2. Hỗ trợ ra quyết định nhanh hơn

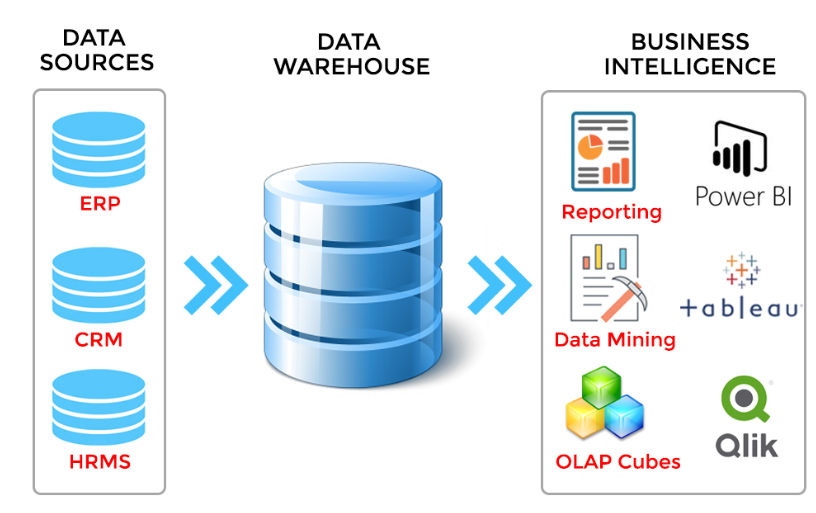
Khi lưu trữ tất cả dữ liệu, báo cáo và phân tích của bạn ở một nơi, bạn có thể dễ dàng phân tích trên khối lượng lớn tập dữ liệu. Data Warehouse không chỉ cải thiện hiệu quả truy cập vào bộ dữ liệu mà còn giúp mọi người trong nhóm dễ dàng đưa ra quyết định nhanh hơn dựa trên thông tin chi tiết có được.

#3. Sự nhất quán về dữ liệu và công thức

Dữ liệu trong kho dữ liệu là “phiên bản duy nhất của sự thật”. Các khối OLAP (lấy dữ liệu từ kho dữ liệu) cung cấp nguồn dữ liệu chung duy nhất đó cho mọi người và do đó sự phù hợp của dữ liệu được duy trì trong khi tạo báo cáo kinh doanh.

## 5.2. Mô tả kiến trúc kho dữ liệu dự kiến

Data Warehouse (kho dữ liệu) là thuật ngữ được khai sinh và nhắc đến lần đầu bởi hai nhà nghiên cứu của công ty IBM là Barry Devlin và Paul Murphy. Tại kho dữ liệu, mọi thông tin đều được mã hóa và ghi lại trên thiết bị điện tử để phục vụ cho mục đích truy vấn, phân tích và báo cáo dữ liệu. Do đó kho dữ liệu là thành phần cốt lõi của kinh doanh thông minh.



Hình 5.4: Cấu trúc cơ bản Data Warehouse

Nguồn: (SinnovaSoft, 2020)

Các thành phần chính trong Data Warehouse gồm:

**Nguồn dữ liệu và trích xuất dữ liệu**: trích xuất dữ liệu từ nhiều nguồn khác nhau, chuyển đổi dữ liệu thành định dạng nhất quán và tải dữ liệu đó vào Data Warehouse. Các công cụ ETL tự động hóa các quy trình này và đảm bảo tính toàn vẹn và chính xác của dữ liệu.

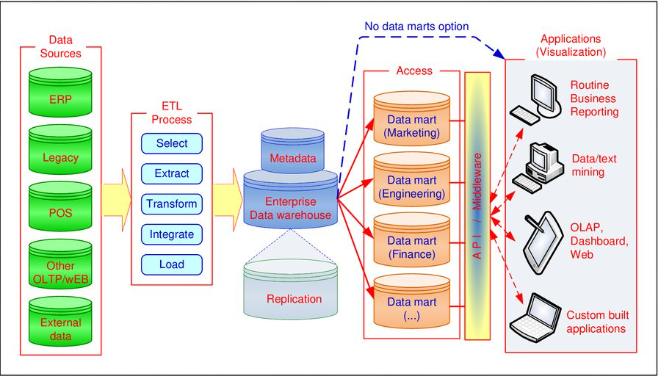
**Quản lý Warehouse**: thực hiện các tác vụ liên quan đến việc quản lý dữ liệu trong kho, đảm nhận các hoạt động như phân tích dữ liệu, tạo các chỉ mục nhỏ và khung nhìn để đảm bảo tính nhất quán, chuẩn hóa và tổng hợp thông tin cùng chuyển đổi và hợp nhất dữ liệu nhiều nguồn.

**Quản lý truy vấn:** là một thành phần phụ trợ nằm trong Data Warehouse. Bộ phận thực hiện tất cả các hoạt động liên quan đến việc tìm kiếm, truy vấn của người dùng. Ngoài ra, hoạt động của trình quản lý này là truy vấn trực tiếp đến các bảng thích hợp nhằm lên lịch thực hiện truy vấn.

**Công cụ truy cập của người dùng cuối:** chia thành năm nhóm riêng biệt gồm báo cáo dữ liệu, công cụ truy vấn, công cụ phát triển ứng dụng, công cụ EIS, công cụ OLAP và các công cụ khai thác dữ liệu.

Mỗi doanh nghiệp đều có một kiểu kiến trúc Data Warehouse khác nhau do nhu cầu sử dụng khác nhau. Hiện nay có 3 loại Data Warehouse phổ biến đó là:

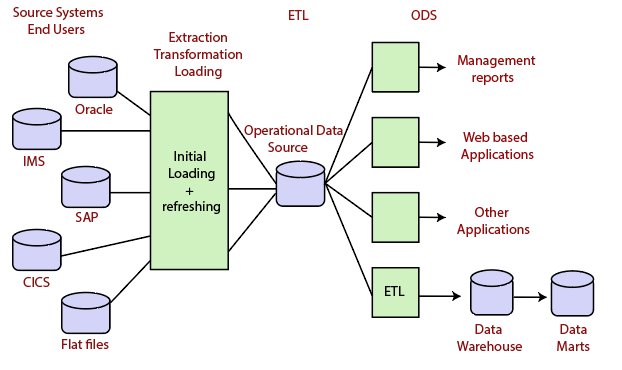
**1. Kho dữ liệu doanh nghiệp - Enterprise Data Warehouse (EDW)**: là một kho tập trung. EDW có thể cung cấp một số lợi ích, chẳng hạn như cải thiện khả năng truy cập và phân tích dữ liệu, nhưng nó cũng có thể tốn kém và phức tạp để triển khai và duy trì. Không phải tất cả các tổ chức đều cần EDW. Do các câu hỏi báo cáo phân tích thống kê của nhóm khá đơn giản nên cần triển khai Data Warehouse nhỏ hơn để tiết kiệm chi phí.



Hình 5.5: Khung kiến trúc Kho dữ liệu doanh nghiệp

Nguồn: (Turban, Sharda, Delen, & King, 2014)

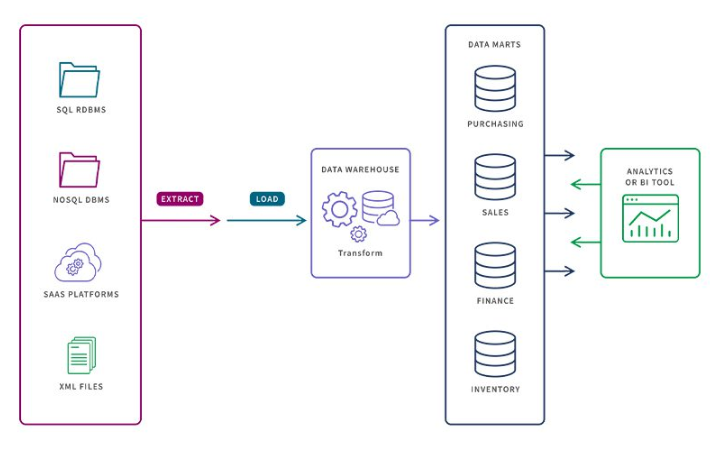
**2. Kho lưu trữ dữ liệu hoạt động - Operational Data Store (ODS)**: là kho lưu trữ dữ liệu hoạt động, chỉ đơn giản là kho lưu trữ dữ liệu cần thiết khi Data Warehouse thông thường và hệ thống OLTP không hỗ trợ báo cáo nhu cầu cho các tổ chức. ODS thường được sử dụng để hỗ trợ các quyết định kinh doanh tác nghiệp, chẳng hạn như báo cáo và phân tích hàng ngày.



Hình 5.6: Khung kiến trúc Kho dữ liệu hoạt động

Nguồn: (INDA, 2023)

**3. Dữ liệu cục bộ - Data Mart**: Một dữ liệu cục bộ được coi là một tập hợp con của kho dữ liệu và thường được định hướng cho một nhóm hoặc ngành nghề kinh doanh cụ thể, chẳng hạn như tài chính hoặc bán hàng. Nó được định hướng theo chủ đề, cung cấp dữ liệu cụ thể cho một nhóm người dùng xác định nhanh hơn, cung cấp cho họ những hiểu biết quan trọng. Tính sẵn có của dữ liệu cụ thể đảm bảo rằng họ không cần mất thời gian tìm kiếm trong toàn bộ kho dữ liệu.



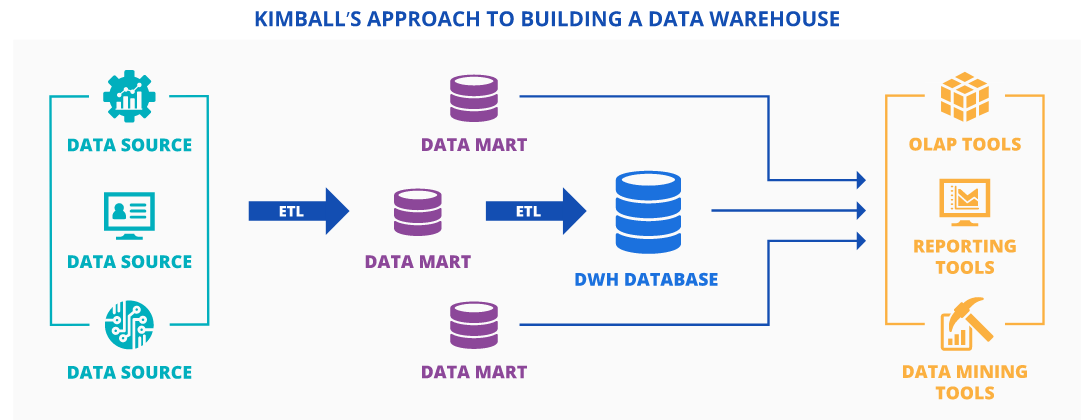
Hình 5.7: Khung kiến trúc xây dựng Data Mart

Nguồn: (Analytics, 2023)

Như vậy, ODS thường được sử dụng để lưu trữ dữ liệu nhỏ hơn, có cấu trúc, chẳng hạn như dữ liệu bán hàng, dữ liệu tài chính và dữ liệu hoạt động. EDW thường được sử dụng để lưu trữ dữ liệu lớn, có cấu trúc và không có cấu trúc, chẳng hạn như dữ liệu khách hàng, dữ liệu sản phẩm và dữ liệu chuỗi cung ứng. Data Mart thường được sử dụng để lưu trữ dữ liệu nhỏ hơn, có cấu trúc, phục vụ nhu cầu của một bộ phận hoặc chức năng kinh doanh cụ thể.

Do dữ liệu của công ty Northwind chỉ lưu trữ từ một nguồn duy nhất trong Excel và được tổ chức theo cách dễ hiểu. Hơn nữa các câu hỏi báo cáo phân tích thống kê của nhóm khá đơn giản nên cần triển khai Data Warehouse nhỏ để tiết kiệm chi phí mà vẫn phục vụ tốt cho nhóm người dùng của bộ phận bán hàng và không cần phải truy cập dữ liệu từ các vị trí khác nhau.

Do đó, dựa vào kích thước, độ phức tạp của tổ chức cùng nhu cầu phân tích dữ liệu của quản lý và ngân sách của công ty nhóm quyết định xây dựng Data Mart cho Kho dữ liệu của mình.



Hình 5.8: Cách tiếp cận từ dưới lên của Kimball

Nhóm chọn cách tiếp cận từ dưới lên của Kimball: Dữ liệu nguồn từ phân hệ vận hành được trích xuất đến Data Staging Area (khu vực nơi mà dữ liệu nguồn được lưu trữ như một dạng backup storage), sau đó sẽ được tải lên Data mart, với phương pháp Bottom-up, Data mart được tạo ra trước để cung cấp khả năng phân tích và báo cáo cho việc kinh doanh, sau khi tạo xong Data mart rồi, mới xây dựng Data Warehouse. Phương pháp tiếp cận của Kimball tập trung vào các yêu cầu dữ liệu của người dùng cuối. Những nhu cầu này thúc đẩy việc sử dụng các công cụ truy cập dữ liệu thích hợp.

1. Ưu điểm:

* Báo cáo được tạo nhanh chóng.
* Có thể cung cấp thêm Data mart, nhờ đó Data Warehouse có thể được mở rộng.
* Tiết kiệm chi phí và tiêu tốn thời gian ít hơn, thiết lập ban đầu cũng nhanh hơn.

1. Nhược điểm:

* Mỗi Data Mart cho thấy góc nhìn hẹp về dữ liệu.
* Có thể xuất hiện dữ liệu dư thừa hoặc không tương thích trong mỗi Data mart.
* Khó bảo trì.

# Câu 6. Xác định nghiệp vụ mà nhóm quan tâm. Xác định độ chi tiết của dữ liệu - dựa vào cơ sở nào để đưa ra được mức độ chi tiết này (1 điểm)

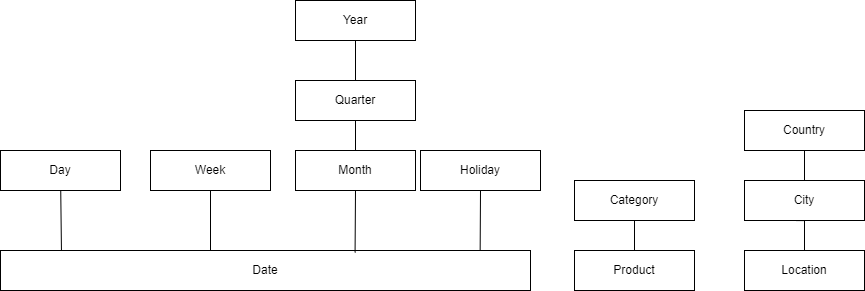
Với bối cảnh của công ty Northwind nhóm chọn xây dựng kho dữ liệu phục vụ cho nghiệp vụ bán hàng. Vì hoạt động bán hàng là một hoạt động kinh doanh quan trọng, đóng vai trò trung gian giữa công ty và khách hàng, giúp cho hàng hóa lưu thông và tạo ra doanh thu và lợi nhuận cho doanh nghiệp cũng như có thể phân tích thị trường và dự báo cho tương lai. Khi doanh nghiệp phát triển và mở rộng thì nhu cầu về nhân lực cũng tăng từ đó cũng tạo thêm việc làm cho người lao động.

Sau khi nhóm xác định chọn nghiệp vụ bán hàng để phân tích, bây giờ cần xác định mức độ chi tiết của dữ liệu, dữ liệu nào sẽ cần có trong bảng dimension. Dựa vào việc phân cấp dữ liệu giúp chúng ta biết được dữ liệu của các bảng cắt lớp chi tiết đến đâu. Trong bảng cắt lớp đôi khi sẽ có thuộc tính này là tập con của thuộc tính khác tức là chúng ta có thể cộng dồn (tổng hợp) thuộc tính này theo thuộc tính khác.

Câu hỏi *Thống kê số lượng đơn hàng theo khu vực địa lý trong một sự kiện cụ thể* (Câu hỏi 1 Bảng 4.1), *Thống kê doanh số theo sản phẩm, khu vực và thời gian* (Câu hỏi 2 Bảng 4.1), *Dự báo xu hướng tiêu thụ sản phẩm theo vùng trong thời điểm tới* (Câu hỏi 3 Bảng 4.1), *Mức tăng trưởng doanh số của sản phẩm theo khu vực so với cùng kỳ năm trước* (Câu hỏi 4 Bảng 4.1), *Báo cáo lợi nhuận theo thời gian, khu vực địa lý, danh mục sản phẩm* (Câu hỏi 5 Bảng 4.1), *Thống kê doanh thu khách hàng theo sản phẩm, khu vực và thời gian* (Câu hỏi 8 Bảng 4.1) của bộ phận quản lý cần mà nhóm đặt ra liên quan đến việc thống kê các báo cáo theo khu vực địa lý, theo vùng do đó chúng ta cần biết chi tiết của từng địa điểm theo City (thành phố), Country (Nước).

Hầu hết các câu hỏi của bộ phận quản lý cần mà nhóm đặt ra đều liên quan đến việc thống kê các báo cáo theo thời gian. Điển hình như: *Báo cáo số lượng đơn hàng theo khu vực địa lý trong một sự kiện cụ thể* (Câu hỏi 1 Bảng 4.1)*, Thống kê doanh số theo sản phẩm, theo khu vực và thời gian* (Câu hỏi 2 Bảng 4.1), *Dự báo xu hướng tiêu thụ sản phẩm theo vùng trong thời điểm tới* (Câu hỏi 3 Bảng 4.1), *Báo cáo lợi nhuận theo thời gian, khu vực địa lý, danh mục sản phẩm* (Câu hỏi 5 Bảng 4.1). Do đó chúng ta cần độ chi tiết của thời gian về ngày cụ thể, về tuần trong tháng, về tháng/quý/năm và về sự kiện (ngày lễ).

Dựa vào các câu hỏi: *Báo cáo lợi nhuận theo thời gian, khu vực địa lý, danh mục sản phẩm* (Câu hỏi 5 Bảng 4.1), *Thống kê tổng doanh số theo khách hàng, theo thời gian và theo danh mục sản phẩm* (Câu hỏi 6 Bảng 4.1) thì nhà quản lý cần báo cáo phân tích theo danh mục sản phẩm. Do đó chúng ta cần biết độ chi tiết về Category (Danh mục sản phẩm).



Hình 6.1: Cây phân cấp dữ liệu của công ty Northwind

(Nguồn: Nhóm tự nghiên cứu)

Dữ liệu trong bảng cắt lớp Date (thời gian) cho thấy chúng ta có thể tổng hợp dữ liệu theo week (tuần trong tháng) và week không thể cộng dồn thành tháng nên mức cao nhất của nhánh này là week. Nhánh thứ hai của Date cho thấy chúng ta có thể tổng hợp dữ liệu theo month/quarter/year (tháng trong năm/quý/năm) và có thể tổng hợp dữ liệu theo month thành quarter (quý) và quarter được tổng hợp thành year. Nhánh thứ 3 là holiday (ngày lễ) và đây là mức cao nhất của nhánh, chúng ta có thể tổng hợp báo cáo theo Holiday (ngày lễ) để đưa ra báo cáo phù hợp.

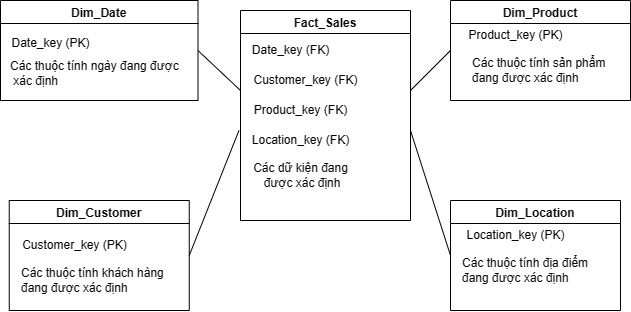
Dữ liệu trong bảng cắt lớp Product (sản phẩm) thì chúng ta có thể tổng hợp sản phẩm theo Category (Danh mục sản phẩm) từ đó có thể đưa ra các báo cáo doanh thu, số lượng bán, số lượng tồn kho,... theo danh mục sản phẩm.

Dữ liệu trong bảng cắt lớp Location (Địa điểm), chúng ta có thể tổng hợp Address (Địa chỉ) theo City (Thành phố) hoặc Country. Tại nhánh này cho thấy dữ liệu về City có thể tổng hợp từ Country.

Việc phân cấp dữ liệu giúp cho dữ liệu dễ dàng quản lý và dễ thao tác thực hiện phân tích tại các cấp độ khác nhau của hệ thống phân cấp dữ liệu. Điều này giúp chúng ta có thể dễ dàng đưa ra các báo cáo mà nhà quản lý cần và từ đó dễ dàng đưa ra các chiến lược kinh doanh phù hợp với từng thời điểm, từng khu vực cũng như từng danh mục sản phẩm.

# Câu 7. Thiết kế các bảng chiều (bảng cắt lớp) DIM, giải thích tại sao lại có các bảng DIM này, mô tả chi tiết từng bảng, nguồn của bảng DIM lấy từ bảng nào? (2 điểm)

Các bảng cắt lớp sẽ được xây dựng dựa trên các báo cáo phân tích thống kê nhóm đã đặt ra ở câu hỏi số 4 ví dụ như: Báo cáo doanh số bán hàng theo khách hàng, theo khu vực và theo thời gian; Báo cáo doanh thu theo sản phẩm và theo danh mục sản phẩm. Với Báo cáo doanh số bán hàng theo khách hàng, theo khu vực và theo thời gian cần dữ liệu về khách hàng, khu vực bán sản phẩm và dữ liệu về thời gian, vậy nên để làm báo cáo này ta cần xây dựng các bảng cắt lớp Dim\_Customer, Dim\_Location và Dim\_Date. Còn với Báo cáo doanh thu theo sản phẩm và theo danh mục sản phẩm, cần dữ liệu về sản phẩm và danh mục sản phẩm mà công ty kinh doanh nên cần xây dựng bảng Dim\_Product chứa cả thuộc tính về Category (danh mục sản phẩm) và Product (sản phẩm).



Hình 7.1: Sơ đồ bán lẻ sơ bộ Northwind

*(Nguồn: Nhóm tự nghiên cứu)*

* Dim\_date

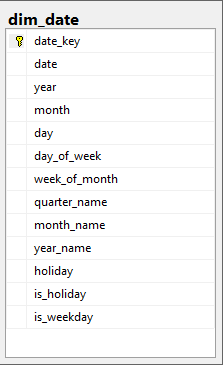
Để trả lời cho các câu hỏi về thời gian theo ngày, tháng, quý, năm thì việc xây dựng bảng cắt lớp thời gian dim\_date là cần thiết. Bảng cắt lớp thời gian là bảng tham chiếu được sử dụng chung trong hệ thống kho dữ liệu để lưu trữ thông tin về thời gian. Thông tin về thời gian được chia thành các mức khác nhau, chẳng hạn như ngày, tháng, quý, năm. Trường dữ liệu thời gian được ưu tiên nhất khi sắp xếp hoặc tạo chỉ mục index trong cơ sở dữ liệu. Bảng cắt lớp thời gian có thể được tạo sẵn, cố định và không thay đổi trong suốt quá trình xây dựng, triển khai và vận hành hệ thống kho dữ liệu.

|  |  |  |  |
| --- | --- | --- | --- |
| Thuộc tính | Mô tả | Kiểu dữ liệu | Giá trị mẫu |
| date\_key | Khóa chính bảng dim date | INT |  |
| date | Ngày | DATE | 1996-07-01 |
| month | Tháng | INT | 07 |
| year | Năm | INT | 1996 |
| day | Ngày trong tháng | INT | 01 |
| day\_of\_week | Ngày trong tuần | VARCHAR(10) | 2 |
| week\_of\_month | Tuần trong tháng | INT | 1 |
| quarter | Quý | INT | 3 |
| quarter\_name | Tên quý | VARCHAR(50) | Quý 3 |
| month\_name | Tên tháng | VARCHAR(50) | Tháng 7 |
| year\_name | Tên năm | VARCHAR(50) | Năm 1996 |
| is\_holiday | Ngày lễ nhận giá trị 0 hoặc 1 | BIT | 0 |
| is\_weekend | Cuối tuần nhận giá trị 0 hoặc 1 | BIT | 1 |

Bảng 7.1: Mô tả chi tiết bảng Dim date

Bảng cắt lớp thời gian trong bộ dữ liệu Northwind được thiết kế để lưu trữ thông tin về các ngày liên quan đến công ty. Bảng này cung cấp thông tin về các ngày và được sử dụng bởi các bảng fact để lưu trữ dữ liệu kinh doanh. Mặt khác, trường ngày cuối tuần giúp người quản trị phân biệt các ngày đặc biệt trong năm từ đó đưa ra các quyết định nhập số lượng hàng phù hợp kết hợp với các hoạt động bán hàng và marketing. Từ đó, doanh nghiệp có thể đưa ra chính sách mua sắm kích cầu nhằm gia tăng doanh thu.

Bảng dim\_date sẽ như sau:



Hình 7.2: Các thuộc tính bảng Dim\_Date

Câu lệnh tạo bảng:

CREATE TABLE dim\_date (

date\_key int NOT NULL IDENTITY PRIMARY KEY,

date datetime DEFAULT NULL,

year int DEFAULT NULL,

month int DEFAULT NULL,

day int DEFAULT NULL,

day\_of\_week varchar(10) DEFAULT NULL,

week\_of\_month int DEFAULT NULL,

quarter\_name varchar(50) DEFAULT NULL,

month\_name varchar(50) DEFAULT NULL,

year\_name varchar(50) DEFAULT NULL,

holiday varchar(200) DEFAULT NULL,

is\_holiday BIT DEFAULT NULL,

is\_weekday BIT DEFAULT NULL,

)

* Dim\_customer

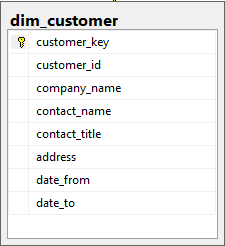
Để trả lời cho câu hỏi về khách hàng như câu hỏi 6: “Thống kê tổng doanh số theo khách hàng, theo thời gian và theo danh mục sản phẩm”, câu hỏi 8: “Thống kê doanh thu khách hàng theo sản phẩm, theo khu vực và theo thời gian.” nhóm đã xây dựng bảng dim\_customer. Các thông tin của khách hàng sẽ bao gồm như id, tên công ty, người liên hệ, chức vụ, … Bảng dim\_customer được xây dựng từ các trường sau đây:

|  |  |  |  |
| --- | --- | --- | --- |
| Thuộc tính | Mô tả | Kiểu dữ liệu | Giá trị mẫu |
| customer\_id | Khóa chỉnh bảng customer của nguồn | VARCHAR(5) | 2 |
| customer\_key | Khóa chính của bảng dim customer | INT | 8 |
| company\_name | Tên công ty | VARCHAR(40) | Maria Anders |
| contact\_name | Tên người liên hệ | VARCHAR(30) | SalesRepresentative |
| contact\_title | Chức vụ | VARCHAR(30) | Owner |
| address | Địa chỉ | VARCHAR(60) | Avda.delaCostitucin 2222 |
| date\_from | Thời gian ghi bản ghi cắt lớp vào cơ sở dữ liệu | DATETIME | 02/09/1995 |
| date\_to | Thời gian bản ghi hết hiệu lực, khi bản ghi còn hiệu lực thì trường dữ liệu này nhận giá trị null | DATETIME | 05/11/2099 |

Bảng 7.2: Mô tả chi tiết bảng Dim\_Customer

Bảng cắt lớp customer mô tả tất cả các khách hàng mà Northwind đã và đang quản lý. Bảng dim\_customer được tạo ra bằng cách bảng customers trong phân hệ vận hành, cộng thêm một trường khóa thay thế (surrogate key) customer\_key làm khóa chính. Với các trường dữ liệu yêu cầu khi có sự thay đổi giá trị thì giá trị mới phải được ghi nhận và đồng thời giá trị cũ vẫn phải được lưu trữ để đảm bảo bản ghi fact ở thời điểm nào sẽ nhận được giá trị cắt lớp tương ứng với thời điểm ấy (theo SCD kiểu 2), nhóm đã thêm vào bảng dim\_customer 3 cột dữ liệu: version, date\_from và date\_to để lưu trữ lịch sử, khoảng thời gian dữ liệu có hiệu lực.

Cuối cùng ta có cấu trúc bảng cắt lớp dim\_customer như sau:



Hình 7.3: Thuộc tính bảng Dim\_Customer

Câu lệnh tạo bảng dim\_customer:

CREATE TABLE dim\_customer (

customer\_key int NOT NULL IDENTITY PRIMARY KEY,

customer\_id varchar(5) NULL,

company\_name varchar(40) DEFAULT NULL,

contact\_name varchar(30) DEFAULT NULL,

contact\_title varchar(30) DEFAULT NULL,

address varchar (60) DEFAULT NULL,

version BIT DEFAULT NULL,

date\_from datetime NULL,

date\_to datetime NULL

)

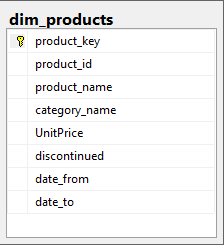
* Dim\_product

|  |  |  |  |
| --- | --- | --- | --- |
| Thuộc tính | Mô tả | Kiểu dữ liệu | Giá trị mẫu |
| product\_key | Khóa chính bảng Dim\_product | int | 8 |
| product\_id | Khóa chính bảng products của nguồn | int | 48 |
| product\_name | Tên sản phẩm | nvarchar(50) | Chartreuse verte |
| category\_name | Tên danh mục sản phẩm | nvarchar(50) | Beverages |
| unit\_price | Giá bán trên mỗi sản phẩm | float | 18 |
| Discontinued | Sản phẩm ngừng bán, nhận giá trị 1 đối với sản phẩm đã ngừng bán, 0 với các sản phẩm đang được kinh doanh | bit | 0 |
| Date\_from | Thời gian ghi bản ghi cắt lớp vào cơ sở dữ liệu | datetime | 02/09/1995 |
| Date\_to | Thời gian bản ghi hết hiệu lực, khi bản ghi còn hiệu lực thì trường dữ liệu này nhận giá trị null | datetime | 05/11/2099 |

Bảng 7.3: Mô tả chi tiết bảng Dim\_Product

Bảng cắt lớp products mô tả tất cả các sản phẩm mà Northwind đã và đang kinh doanh, cụ thể là thông tin về 69 loại thực phẩm đang bày bán và khoảng 8 loại thực phẩm đã ngừng kinh doanh. Để trả lời cho các câu hỏi về thống kê doanh số, doanh thu của công ty theo sản phẩm/danh mục sản phẩm thì bảng dim\_product được tạo ra bằng cách join bảng products với bảng categories trong phân hệ vận hành, cộng thêm một trường khóa thay thế (surrogate key) product\_key làm khóa chính và bỏ đi một số trường dữ liệu không có ý nghĩa với kho dữ liệu là categories\_description, picture, supplierID, QuantityPerUnit, UnitsInStock, UnitsOnOrder, ReorderLevel. Với các trường dữ liệu yêu cầu khi có sự thay đổi giá trị thì giá trị mới phải được ghi nhận và đồng thời giá trị cũ vẫn phải được lưu trữ để đảm bảo bản ghi fact ở thời điểm nào sẽ nhận được giá trị cắt lớp tương ứng với thời điểm ấy (theo SCD kiểu 2), nhóm đã thêm vào bảng dim\_product 3 cột dữ liệu: version, date\_from và date\_to.

Cuối cùng ta có cấu trúc bảng cắt lớp dim\_product như sau:



Hình 7.4: Thuộc tính bảng Dim\_Product

Câu lệnh tạo bảng dim\_product:

CREATE TABLE dim\_product (

product\_key int NOT NULL IDENTITY PRIMARY KEY,

product\_id int NULL,

version bit NULL,

product\_name nvarchar (50) NULL,

category\_name nvarchar (50) NULL,

UnitPrice float NULL,

discontinued bit NULL,

date\_from datetime NULL,

date\_to datetime NULL

)

* Dim\_location

Câu hỏi *Thống kê doanh số theo sản phẩm, khu vực và thời gian* (câu hỏi 2 Bảng 4.1), *Thống kê doanh thu khách hàng theo sản phẩm, khu vực và thời gian* (câu hỏi 8 Bảng …) của bộ phận quản lý bán hàng mà nhóm đặt ra liên quan đến việc thống kê các báo cáo theo khu vực địa lý. Để trả lời các câu hỏi báo cáo về doanh số, doanh thu của Northwind theo khu vực địa lý thì việc xây dựng bảng Dim\_Location là cần thiết. Mỗi đơn hàng của khách hàng đặt về đâu được coi là một địa điểm để tổng hợp. Do đó, Dim\_Location có thể lấy thuộc tính địa lý, chẳng hạn như thành phố, quốc gia từ đó dễ dàng hơn trong việc phân tích, tổng hợp dữ liệu. Nhóm xây dựng bảng Dim\_Location với các thuộc tính như bảng sau:

|  |  |  |  |
| --- | --- | --- | --- |
| Thuộc tính | Mô tả | Kiểu dữ liệu | Giá trị mẫu |
| Location\_key | Khóa chính của bảng | INT | 25 |
| City | Thành phố | VARCHAR(50) | Rio de Janeiro |
| Country | Quốc gia | VARCHAR(50) | Brazil |
| Date\_from | Ngày bắt đầu (Ngày địa chỉ được thêm vào) | DATETIME | 02/09/1995 |
| Date\_to | Ngày kết thúc (Ngày địa chỉ không còn trong bảng nữa) | DATETIME | 08/07/1998 |

Bảng 7.4: Mô tả chi tiết bảng Dim\_Location

Bảng Dim\_Location lấy location\_key làm khóa chính. Thuộc tính địa điểm: City, Country được lấy từ bảng Customer của nguồn. Đồng thời nhóm thêm trường version, date\_from, date\_to để lưu trữ lịch sử, khoảng thời gian dữ liệu có hiệu lực.

Câu lệnh tạo bảng Dim\_Location:

CREATE TABLE dim\_location (

"location\_key" int NOT NULL IDENTITY PRIMARY KEY,

"City" varchar(50) NULL,

"Country" varchar(50) NULL,

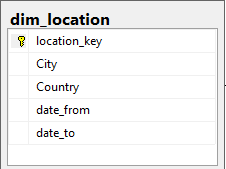
"version" INT NULL,

"date\_from" datetime NULL,

"date\_to" datetime NULL

)

Bảng cắt lớp Dim\_Location:



Hình 7.5: Thuộc tính bảng Dim\_Location

# Câu 8. Thiết kế bảng FACT. Mô hình lược đồ trên Data mart? Mô hình sao hay bông tuyết? (2 điểm)

## 8.1. Cấu trúc bảng Fact\_Sales

Trong quá trình hoạt động các nhà quản lý sẽ có những yêu cầu, đưa ra những sản phẩm theo tuần, tháng, quý có doanh thu cao nhất, thu hút nhiều khách hàng mới nhất… Khi đó các thông tin cần trả lời sẽ được thể hiện trên bảng Fact sales. Trong trường hợp này nhóm thiết kế bảng Fact theo kiểu tổng hợp: Tính toán sẵn các chỉ tiêu có liên quan đến một nhóm dimension nào đó, giúp thuận lợi trong quá trình phân tích và xuất ra báo cáo.

KimBall (2013) cho rằng một bảng fact chứa các số đo được tạo ra bởi một sự kiện đo lường vận hành trong thế giới thực. Trong trường hợp này các thước đo quantity (số lượng), unit\_price (đơn giá) được lấy từ bảng products trong hệ thống nguồn, nhưng thước đo revenue (dùng để xác định giá trị doanh thu) được xác định như sau: Revenue = unit\_price \* quantity.

Như vậy ngoài các thước đo bằng số, bảng fact luôn chứa các khóa ngoại cho từng bảng dim liên quan của nó. Như trường location\_id là một khóa ngoại tham chiếu đến bảng dim\_location, cung cấp thông tin về tên địa điểm, quốc gia, khu vực…; Trường date\_id là một khóa ngoại tham chiếu đến bảng dim\_date, cung cấp thông tin về ngày, tháng, năm…; Trường customer\_id là một khóa ngoại tham chiếu đến bảng dim\_customer, cung cấp thông tin về tên khách hàng, địa chỉ…; Trường product\_id là một khóa ngoại tham chiếu đến bảng dim\_product, cung cấp thông tin về tên, giá cả sản phẩm…

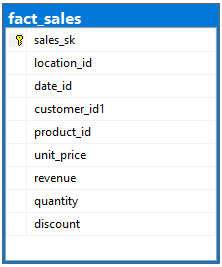
Dữ kiện revenue ở đây thuộc loại additive: có thể được tổng hợp thông qua tất cả các bảng dim trong bảng fact.

Cụ thể trong trường hợp này để trả lời câu hỏi 1 trong bảng 4.1 của bộ phận bán hàng: “*Thống kê số lượng khách hàng được tạo ra theo khu vực địa lý trong một sự kiện cụ thể*”, chúng ta cần câu truy vấn trên trường customer\_id, location\_id và phân theo trường date\_ id.

Đối với câu hỏi 2 trong bảng 4.1 của bộ phận bán hàng: “*Thống kê doanh số theo sản phẩm, khu vực và thời gian”,* báo cáo được trả về dưới dạng biểu đồ được lấy dữ liệu tổng hợp trong các trường của bảng Fact như là: product\_id, location\_id và date\_id.

Trong câu hỏi 7 trong bảng 4.1 của bộ phận bán hàng: “*Thống kê tổng doanh thu và tổng số tiền giảm giá theo sản phẩm và theo thời gian*”, báo cáo được thực hiện dựa trên những dữ liệu tổng hợp từ bảng Dim product\_id, theo date\_id để tổng hợp lại trên trường revenue và discount trong bảng Fact.

Cuối cùng ta có cấu trúc bảng cắt lớp Fact\_sales như sau:



Hình 8.1: Bảng Fact\_Sales

Câu lệnh tạo bảng Fact\_sales

CREATE TABLE fact\_sales (

sales\_sk int NOT NULL PRIMARY KEY,

location\_id int DEFAULT NULL,

date\_id int DEFAULT NULL,

customer\_id int DEFAULT NULL,

product\_id int DEFAULT NULL,

unit\_price money DEFAULT NULL,

revenue money DEFAULT NULL,

quantity int DEFAULT NULL,

discount float NULL,

)

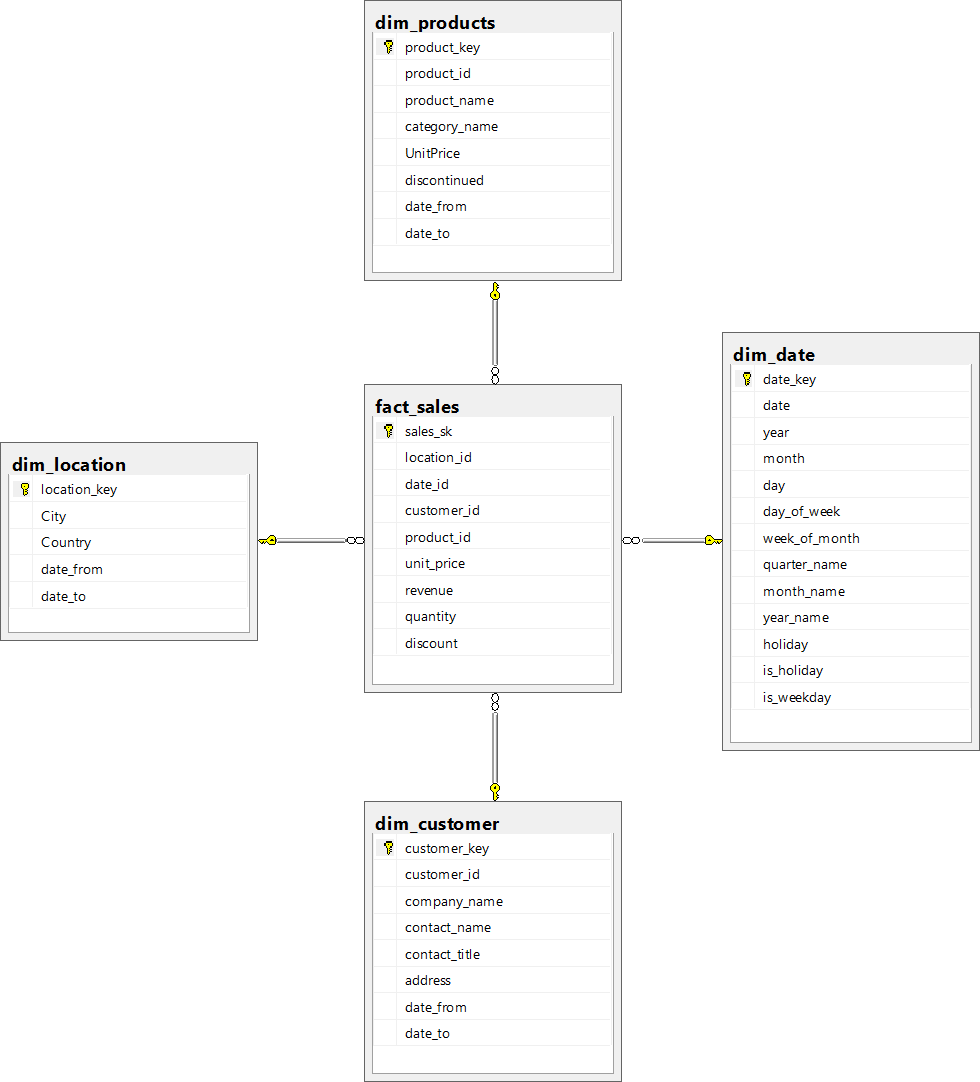
## 8.2. Ý nghĩa trường dữ liệu

|  |  |  |  |
| --- | --- | --- | --- |
| Thuộc tính | Mô tả | Kiểu dữ liệu | Giá trị mẫu |
| sales\_sk | Khóa chính của bảng | INT | 19 |
| location\_id | Mã khu vực | INT | 15 |
| date\_id | Mã thời gian | INT | 28 |
| customer\_id | Mã khách hàng | INT | 14 |
| product\_id | Mã sản phẩm | INT | 6 |
| unit\_price | Đơn giá 1 sản phẩm tính bằng $ | MONEY | 6.15 |
| revenue | Doanh thu tính bằng uni\_price \* quantity | MONEY | 50.10 |
| quantity | Số lượng sản phẩm được bán | INT | 9 |

Bảng 8.1: Mô tả chi tiết bảng Fact\_sales

## 8.3. Mô hình lược đồ trên Data mart, mô hình sao

Từ những bảng Dim, Fact đã xây dựng bên trên thì nhóm chọn kiểu mô hình sao làm lược đồ trên Data Mart.



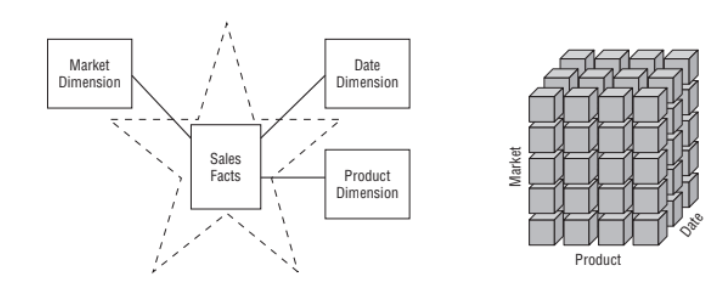
Hình 8.2: Mô hình sao Northwind

(Nguồn: Nhóm tự nghiên cứu)

Mô hình lược đồ này được sử dụng để lưu trữ dữ liệu về doanh số bán hàng của Northwind. Bảng Fact\_Sales chứa dữ liệu định lượng chính, chẳng hạn như số lượng sản phẩm được bán và doanh thu. Các bảng dimension "Dim\_product", "Dim\_customer", “Dim\_location” và "Dim\_date" chứa dữ liệu mô tả về sản phẩm được bán, khách hàng mua hàng, địa điểm bán hàng và thời gian bán hàng.

Đầu tiên, vì lý do hiệu suất cũng như thiết kế đơn giản và nhất quán, nên nhóm sử dụng lược đồ hình sao thay vì lược đồ hình bông tuyết, lược đồ hình sao đơn giản hơn và thống nhất hơn lược đồ hình bông tuyết vì nó chỉ có một cấp độ trong tất cả các chiều. Bởi vậy nên các quy trình ETL sẽ dễ dàng tải dữ liệu vào nó hơn.

Hơn nữa, nhóm đã đề xuất xây dựng kho dữ liệu sử dụng Xử lý phân tích trực tuyến (OLAP), trong chapter 1 cuốn *The Data Warehouse Toolkit Third Edition* của Ralph KimBall có viết đề cập nếu DW/BI gồm lược đồ hình sao hoặc khối OLAP, thì sẽ tận dụng được các khái niệm đa chiều.



Hình 8.3: Star schema versus OLAP cube

Nguồn: (KimBall & Ross, The data warehouse toolkit: The definitive guide to dimensional modeling (3rd ed.), 2023)

Lược đồ hình sao được lưu trữ trong cơ sở dữ liệu quan hệ là nền tảng vật lý tốt để xây dựng khối OLAP và thường được coi là cơ sở ổn định hơn cho việc sao lưu và phục hồi. Vậy nên việc sử dụng mô hình sao để lưu trữ hoạt động của Northwind sau đó dùng OLAP để tổng hợp và phân tích giúp cung cấp một sự cân bằng tốt giữa hiệu suất, khả năng phân tích và tính linh hoạt cho bài toán nhóm đặt ra.

**DANH SÁCH SINH VIÊN THỰC HIỆN**

|  |  |  |  |
| --- | --- | --- | --- |
| **Họ và tên** | **Mã sinh viên** | **Công việc** | **Đóng góp** |
| Nguyễn Thị Mai Anh | 23A4040005 | Câu 9, Câu 11, Câu 14 | 20% |
| Lê Hà Chi | 23A4040018 | Câu 9, Câu 12, Câu 13 | 20% |
| Nguyễn Thị Duyên | 23A4040025 | Câu 9, Câu 10, Câu 11, Câu 13 | 20% |
| Đặng Thị Hạnh | 23A4040034 | Câu 9, Câu 12, Câu 15 | 20% |
| Phan Thị Quỳnh Trang | 23A4040149 | Câu 9, Câu 12 | 20% |

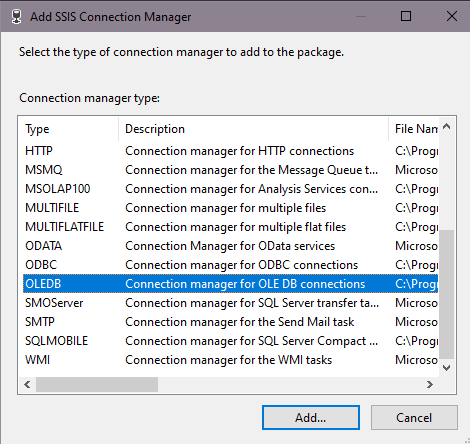
# Câu 9: Đổ dữ liệu từ các bảng nguồn vào các bảng DIM, minh họa bằng hình ảnh việc đổ dữ liệu thành công (2 điểm)

Nhóm thực hiện đổ dữ liệu trong Visual Studio code tích hợp tool SQL Sever Integration Services Projects, các bước thực hiện như sau:

**Bước 1**: Tạo Proect mới: Chọn Create a new project -> Integration Services Project -> Điển tên Project name là “northwind\_biN9” -> Create

**Bước 2**: Tạo kết nối với database trong SQL SERVER

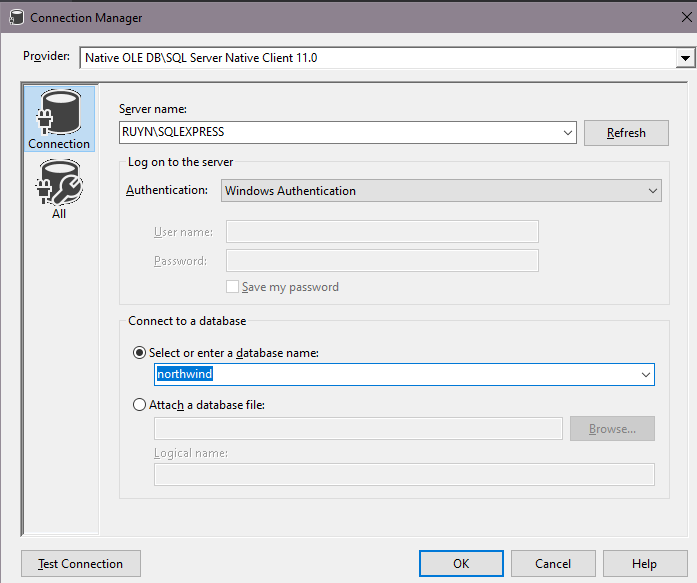
Chuột phải vào “Connection Managers” Chọn New connect Manager



Hình 9.1: Tạo kết nối database

Chọn OLEDB -> Add

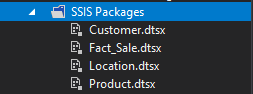
Hiển thị hộp thoại “Configure OLE DB Connection Manager”-> Click New



Hình 9.2: Kết nối với nguồn

Chọn Sever Name -> Chọn cơ sở dữ liệu nguồn có tên “northwind”, tương tự thêm cả nguồn “northwind\_bi” là Đích mà các dữ liệu các bảng Dim đổ đến.

Bước 3: Bên góc phải, click chuột phải -> New SSIS Package để tạo các SSIS Packages là các flow bảng Dim, Fact nhóm thiết lập gồm: dim\_customer, dim\_product, dim\_location, fact\_sale.



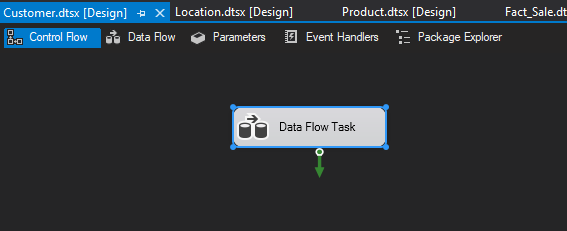
Hình 9.3: Tạo SSIS Packages

Các bước đổ bảng Dim chi tiết như sau:

## 9.1. Bảng Dim\_Customer

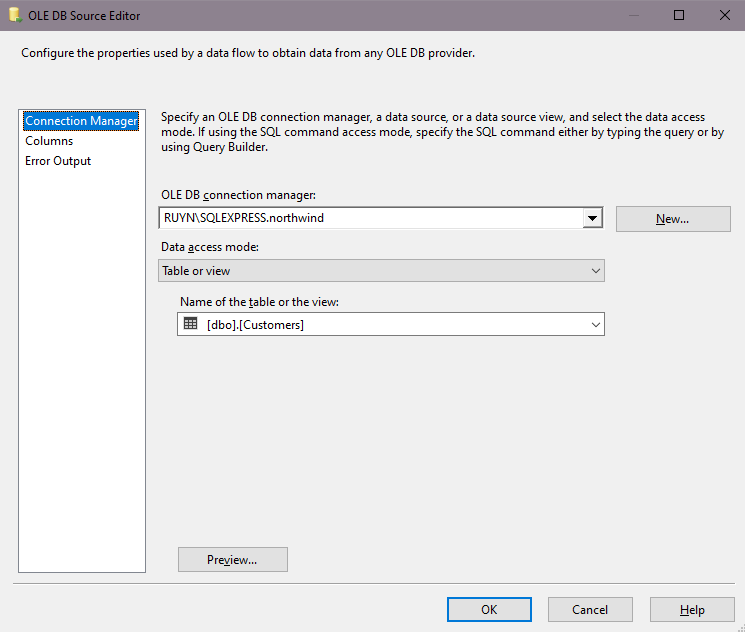
Bảng Dim\_Customer được lấy từ 1 nguồn duy nhất là bảng Customer của Northwind. Các bước cài đặt Slowly Changing Dimention kiểu 2 như sau:

**Bước 1:** Phần Control Flow, kéo thả Data Flow Task -> click đúp Data Flow Task để thực hiện



Hình 9.4: Tạo dataflow bảng dim\_customer

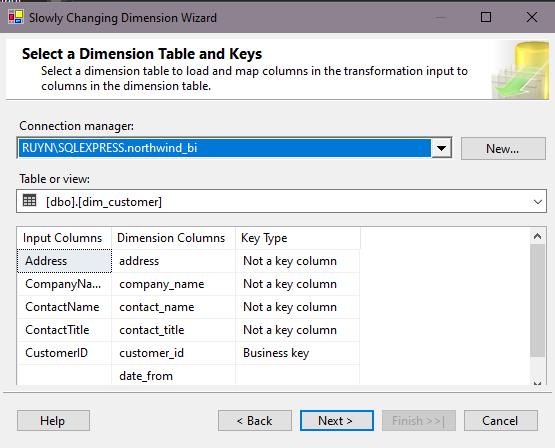
**Bước 2:** Kéo thả OLEDB Source, click đúp và thực hiện: Chọn bảng nguồn là northwind và chọn bảng Customers -> OK



Hình 9.5: Lấy nguồn từ bảng Customers

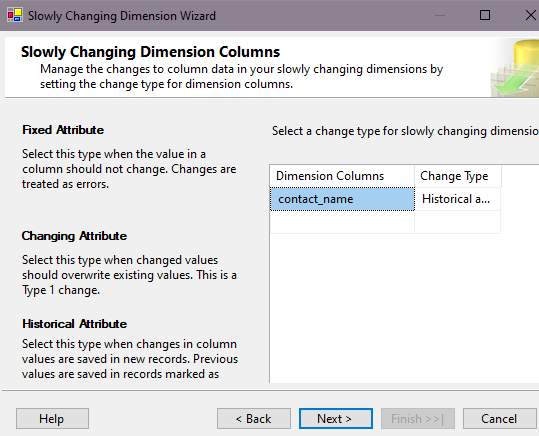
**Bước 3:** Kéo thả Snowly changing Dimension và nối với Source ở trên

\* Click đúp vào và setup theo ảnh sau:



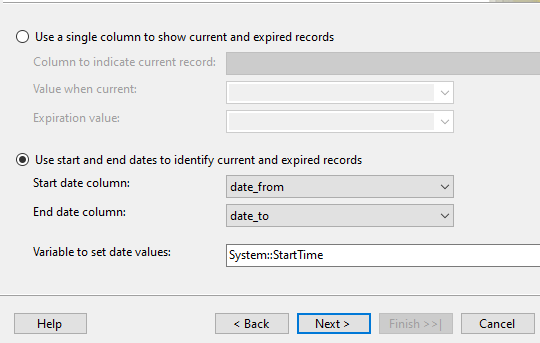
Hình 9. 6: Liên kết các thuộc tính liên quan SCD2

Nối các giá trị liên quan, Chọn Business Key là customer\_id và Next



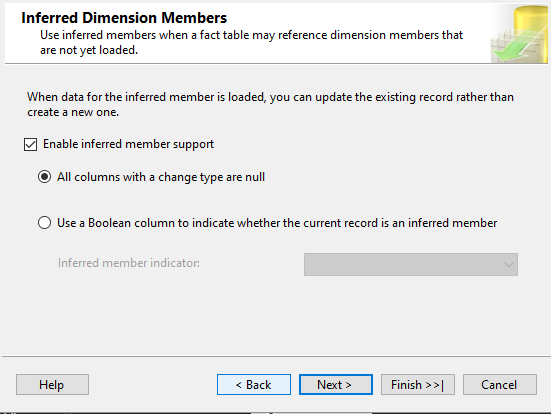
Hình 9.7: Liên kết các thuộc tính liên quan SCD2

Khách hàng của Northwid là doanh nghiệp nên chủ yếu chỉ thay đổi tên khi thực hiện giao dịch, nhóm chọn contact\_name là giá trị để thực hiện kiểu SCD2.



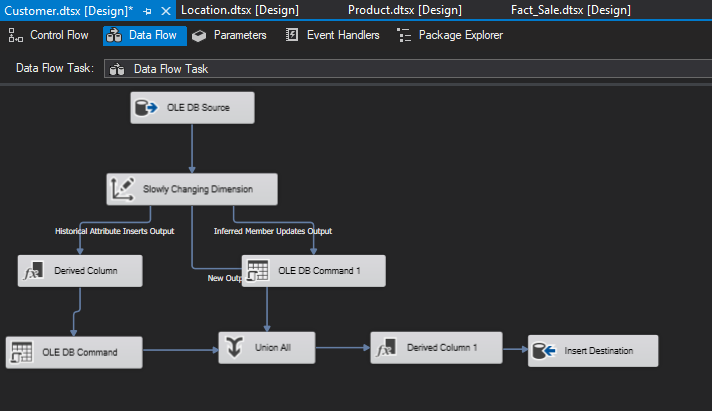
Hình 9.8: Chọn date\_from, date\_to

Chọn date\_from, date\_to và bên dưới chọn StartTime để lưu giá trị thời gian thực.



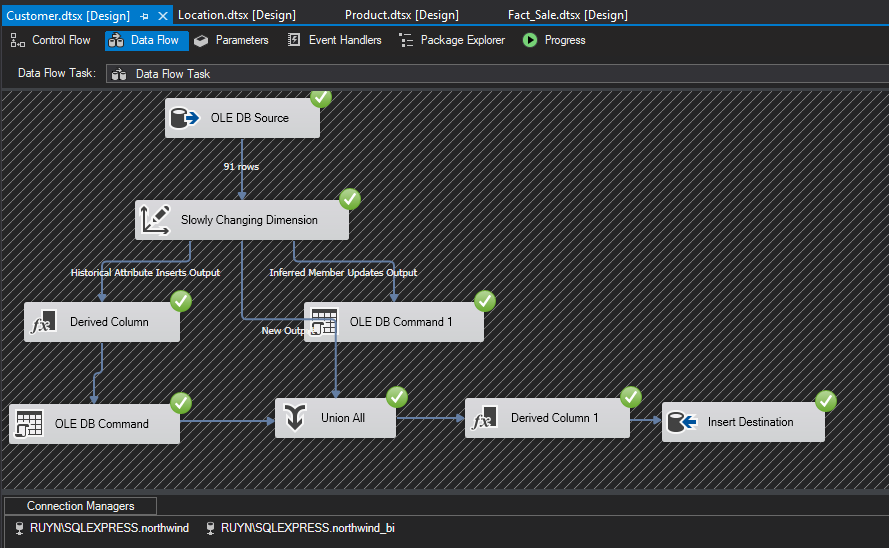
Hình 9.9: Kết thúc SCD2

Ấn Next -> Finish và kết quả như sau:



Hình 9.10: DataFlow bảng Dim\_Customer

Kết quả đổ dữ liệu thành công:



Hình 9.11: Đổ thành công bảng Dim\_Customers

Dữ liệu sau khi đã được đổ:



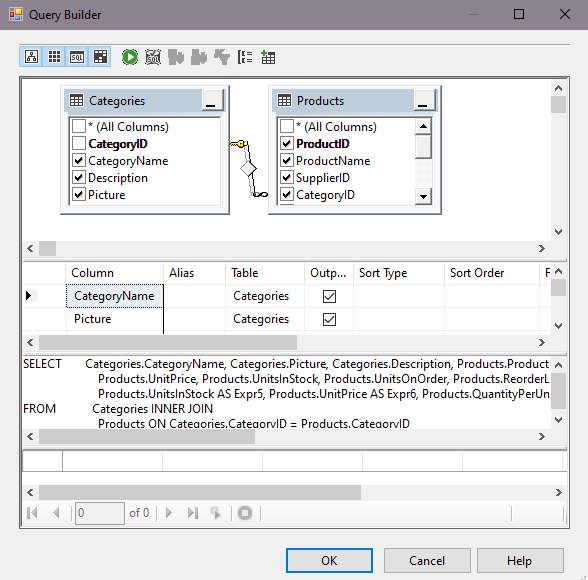
Hình 9.12: Dữ liệu Dim\_Customer sau khi đổ thành công

## 9.2. Bảng Dim\_Product

Dim\_product của Northwind cần dữ liệu từ 2 bảng của nguồn là bảng Products và Categories

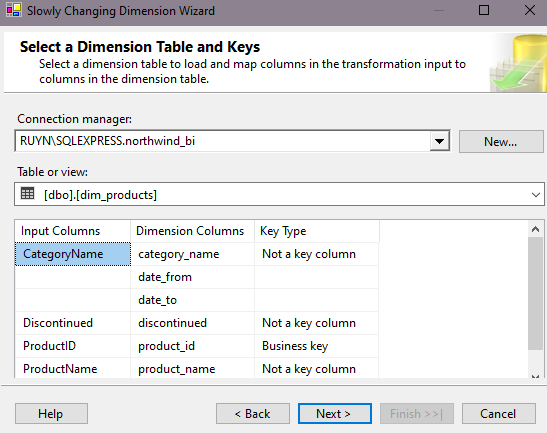
**Bước 1**: Trong phần OLEDB Source phần Data access mode, chọn SQL command - > chọn Build query

Nối 2 bảng, chọn các thuộc tính cần lấy và nhấn OK



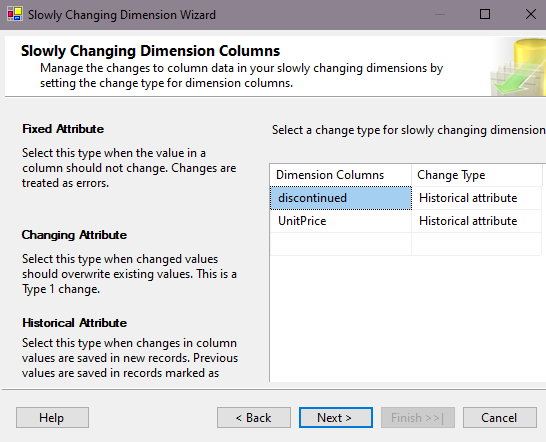
Hình 9.13: Lấy nguồn dữ liệu đổ bảng Dim\_Products

**Bước 2**: Tương tự, cũng kéo thả Snowly Changing Dimension và set up:



Hình 9.14: Tạo kiểu SCD 2 Dim\_Products

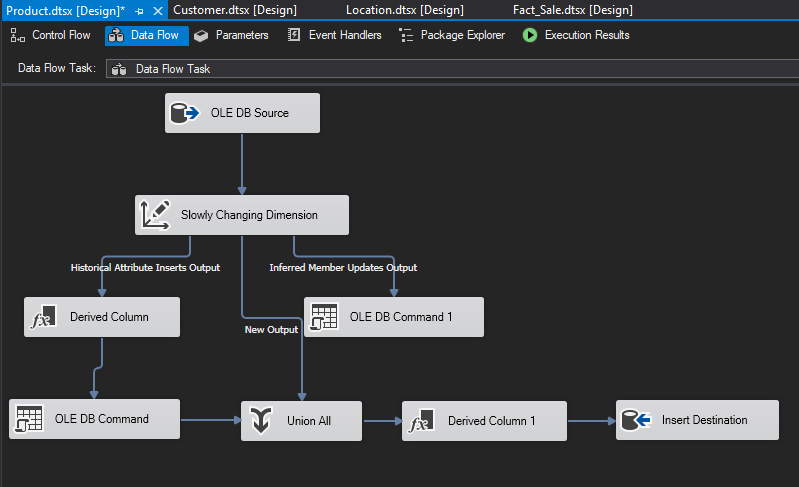
Chọn các thuộc tính cần lấy, chọn Business key là product\_id



Hình 9.15: Chọn thuộc tính thay đổi khi cập nhật SCD2

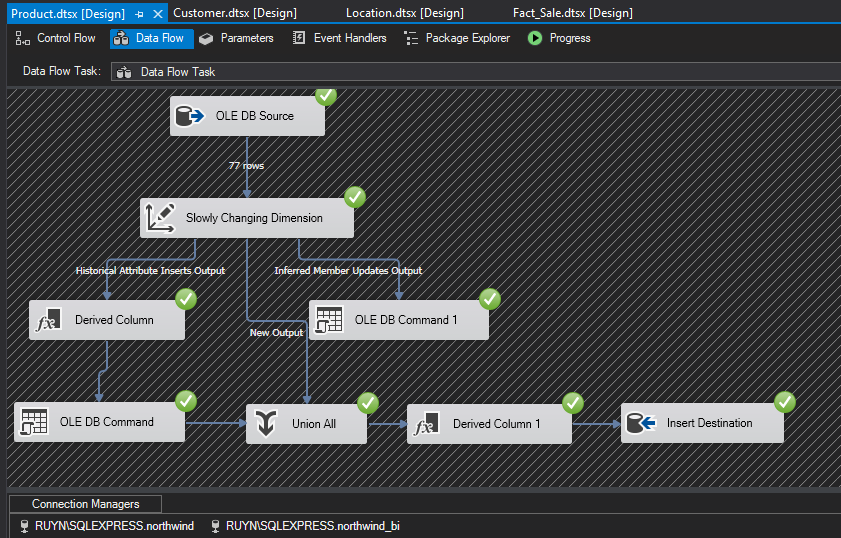
Nhóm chọn discontinued và UnitPrice là giá trị để thực hiện kiểu SCD2.

Nhấn Next, chọn các trường date\_from, date\_to, Starttime -> Next -> Finish và kết quả:



Hình 9.16: DataFlow bảng Dim\_Products

Kết quả đổ dữ liệu thành công:



Hình 9.17: Đổ dữ liệu Dim\_Product thành công

Dữ liệu sau khi đã được đổ:



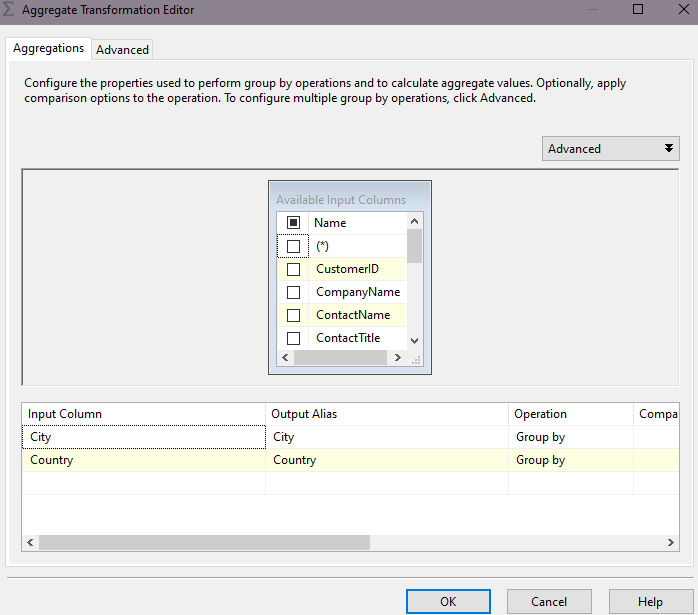
Hình 9.18: Dữ liệu Dim\_Products khi đổ thành công

## 9.3. Bảng Dim\_Location

Dim\_Location lấy dữ liệu về địa điểm của Northwind và lấy dữ liệu từ 1 bảng Customer của nguồn, cách đổ như sau:

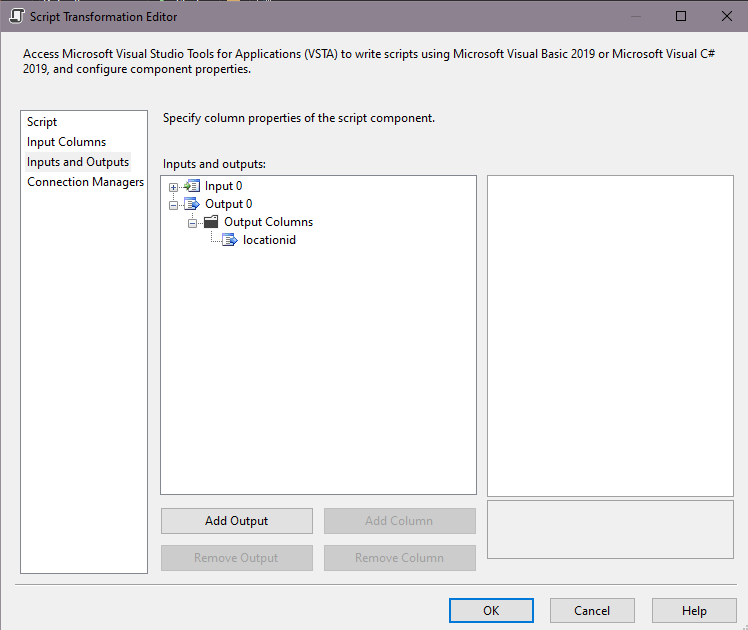
**Bước 1**: Chọn OLEDB Source và chọn bảng nguồn là Customers

**Bước 2**: Kéo thả Aggregate để chọn thuộc tính địa điểm là City và Country



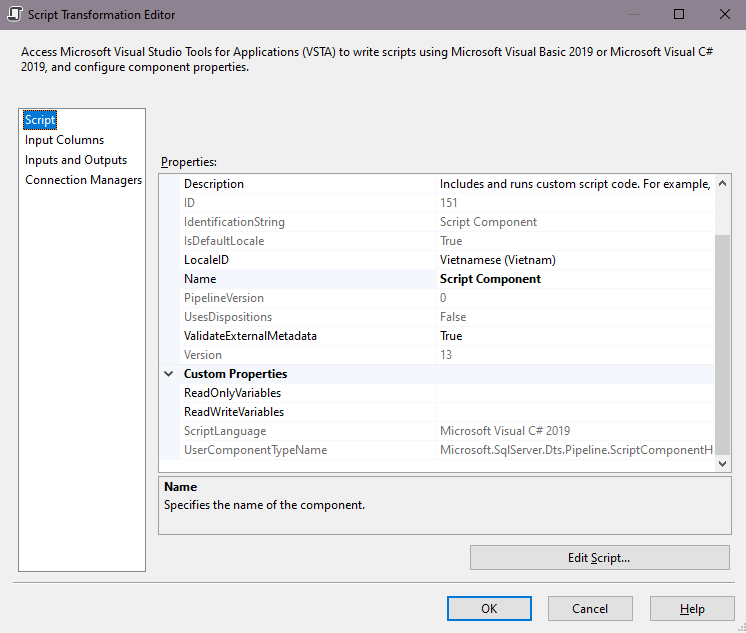
Hình 9.19: Chọn thuộc tính địa chỉ

**Bước 3:** Kéo thả Script Component và thiết lập như sau:



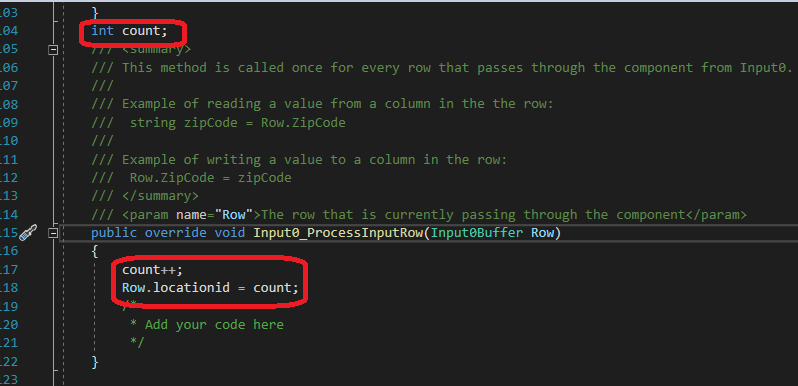
Hình 9.20: Tạo Snipt Component bảng dim\_location

Vào phần Inputs and Outputs và thêm locationid



Hình 9.21: Tạo Snipt Component bảng dim\_location

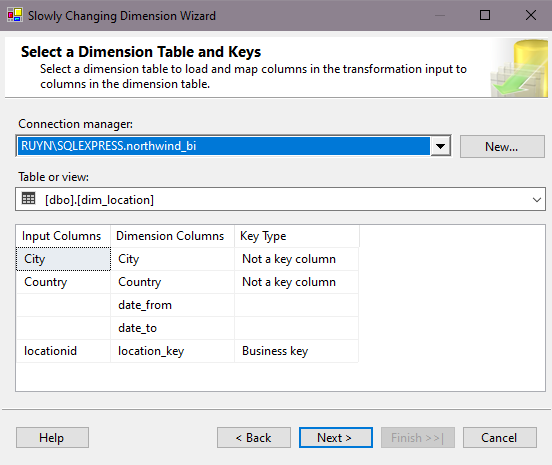
Vào phần Script -> Edit Script để thực hiện các lệnh tiếp theo



Hình 9.22: Thêm code vào Snipt

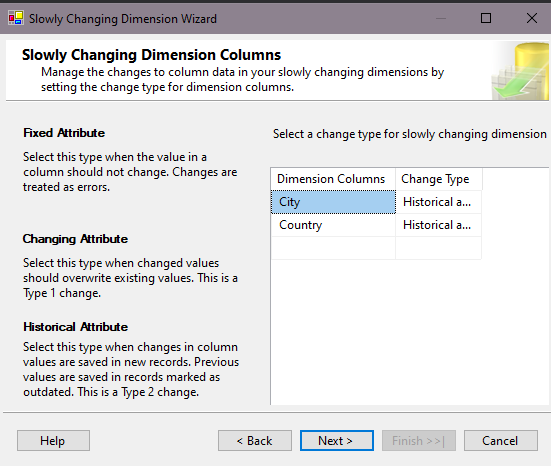
Thêm 2 dòng code trên để thực hiện đếm các giá trị location (City, Country) trùng và sinh ra mã, tránh trường hợp bảng Dim\_Location có bản ghi giống nhau

**Bước 4**: Thêm Snowly Changing Dimesion và thiết lập



Hình 9.23: Tạo SCD2 bảng dim\_location

Chọn các thuộc tính cần thiết và chọn location\_key là Business key



Hình 9.24: Thêm dữ liệu cần cập nhật SCD2 bảng dim\_location

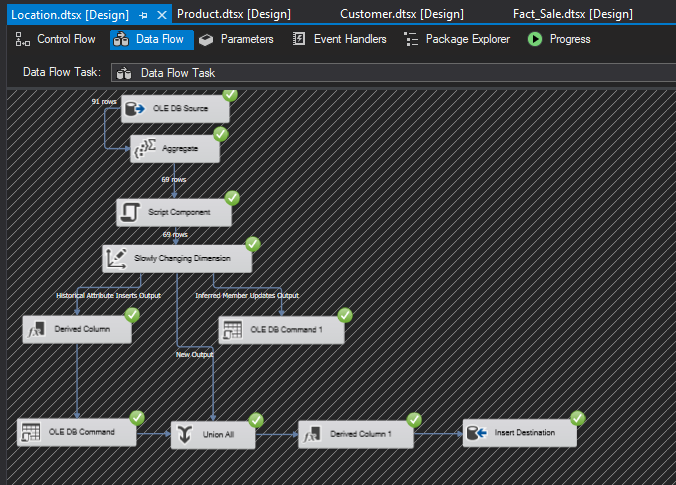
Nhóm chọn City và Country là giá trị để thực hiện kiểu SCD2 khi khách hàng thay đổi địa chỉ

Chọn date\_from, date\_to, StartTime -> Next -> Finish và kết quả là:



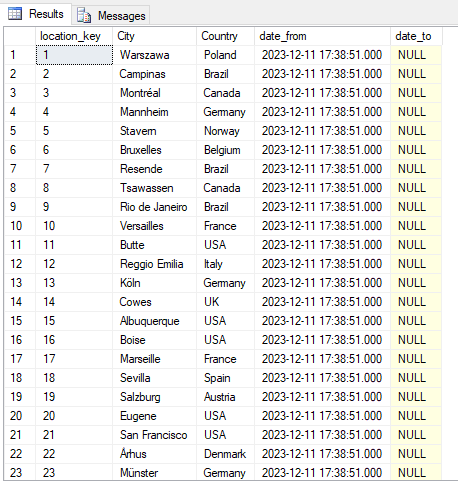
Hình 9.25: DataFlow bảng dim\_location

Kết quả đổ dữ liệu thành công:



Hình 9.26: Đổ dữ liệu thành công dim\_location

Dữ liệu sau khi được đổ:



Hình 9.27: Dữ liệu bảng dim\_location sau khi đổ

## 9.4. Bảng Dim\_Date

Không giống các bảng Dim khác, bảng Dim\_Date được tạo tự động, nhóm đã lấy dữ liệu thời gian từ năm 1980 đến năm 2030 để phù hợp với bộ dữ liệu Northwind. Câu lệnh tạo bảng Dim\_Date trong sql như sau:

SET NOCOUNT ON

DECLARE @StartDate DATE = '1980-01-01'

DECLARE @EndDate DATE = '2030-12-31'

WHILE @StartDate < @EndDate

BEGIN

INSERT INTO [dbo].[dim\_date] (

date\_key,

date,

year,

month,

day,

day\_of\_week,

week\_of\_month,

quarter\_name,

month\_name,

year\_name,

is\_holiday,

is\_weekday

)

SELECT date\_key = YEAR(@StartDate) \* 10000 + MONTH(@StartDate) \* 100 +

DAY(@StartDate),

DATE = @StartDate,

year = YEAR(@StartDate),

month = MONTH(@StartDate),

Day = DAY(@StartDate),

day\_of\_week = DATEPART(dw, @StartDate),

week\_of\_month = DATEPART(WEEK, @StartDate) - DATEPART(WEEK, DATEADD(MM,

DATEDIFF(MM, 0, @StartDate), 0)) + 1,

quarter\_name = CASE

WHEN DATENAME(qq, @StartDate) = 1

THEN 'Q1'

WHEN DATENAME(qq, @StartDate) = 2

THEN 'Q2'

WHEN DATENAME(qq, @StartDate) = 3

THEN 'Q3'

WHEN DATENAME(qq, @StartDate) = 4

THEN 'Q4'

END,

month\_name = DATENAME(mm, @StartDate),

year\_name = DATENAME(yy, @StartDate),

is\_holiday = 0,

is\_weekday = CASE

WHEN DATENAME(dw, @StartDate) = 'Sunday'

OR DATENAME(dw, @StartDate) = 'Saturday'

THEN 1

ELSE 0

END

SET @StartDate = DATEADD(DD, 1, @StartDate)

END

UPDATE [dim\_date]

SET is\_holiday = 1,

holiday = N'Tết dương lịch'

    WHERE month = 1

    AND Day = 1;

UPDATE [dim\_date]

SET is\_holiday = 1,

holiday = N'Ngày lễ tình nhân'

    WHERE month = 2

    AND Day = 14;

UPDATE [dim\_date]

SET is\_holiday = 1,

    holiday = N'Quốc tế Lao Động'

    WHERE month = 5

    AND Day = 1;

UPDATE [dim\_date]

SET is\_holiday = 1,

holiday = N'Ngày Độc lập'

    WHERE month = 7

    AND Day = 4;

UPDATE [dim\_date]

SET is\_holiday = 1,

holiday = N'Ngày Quốc tế Thiếu nhi'

    WHERE month = 6

    AND Day = 19;

UPDATE [dim\_date]

SET is\_holiday = 1,

holiday = N'Ngày Lao động'

    WHERE month = 8

    AND Day = 1;

UPDATE [dim\_date]

SET is\_holiday = 1,

holiday = N'Lễ Tạ ơn'

    WHERE month = 11

    AND Day = 25;

UPDATE [dim\_date]

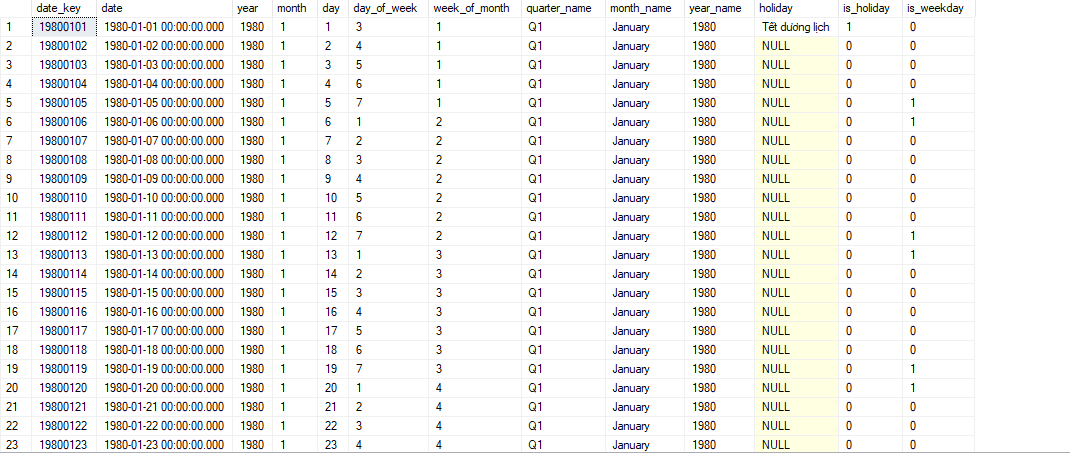
SET is\_holiday = 1,

    holiday = N'Lễ Giáng sinh'

    WHERE month = 12

    AND Day = 25;

Kết quả sau khi insert:



Hình 9.28: Dữ liệu bảng Dim\_Date

# Câu 10: Đổ dữ liệu vào bảng Fact, minh họa đổ dữ liệu thành công (2 điểm)

Bảng Fact\_Sales nhóm xây dựng gồm những thuộc tính về UnitPrice, Revenue, Discount và các thuộc tính khóa nối với các bảng Dim khác, nguồn dữ liệu đổ vào gồm:

* Bảng OrderDetail
* Bảng Order
* Bảng dim\_customer
* Bảng dim\_product
* Bảng dim\_location
* Bảng dim\_date

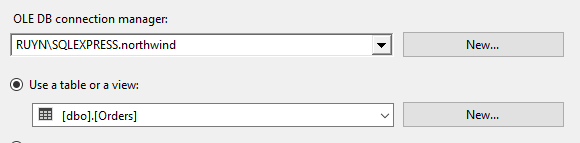
Các bước đổ dữ liệu vào bảng Fact\_Sales

**Bước 1:** Kéo thả OLEDB Source và chọn bảng OrderDetail của DB northwind là nguồn



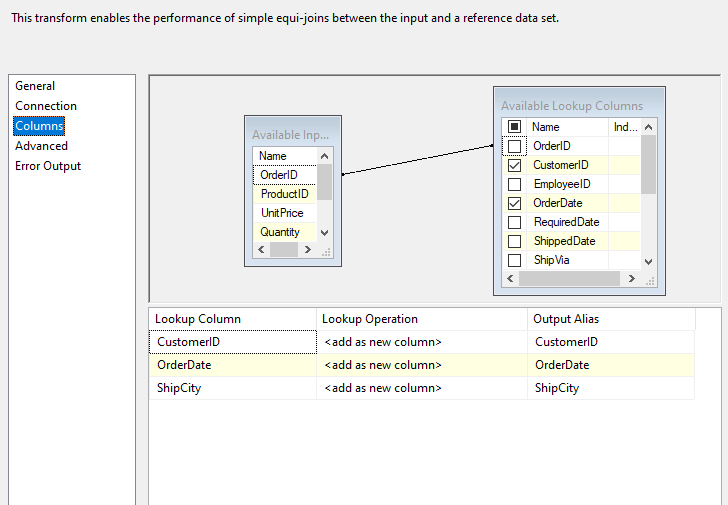
Hình 10.1: Thêm nguồn bảng Fact\_sales

**Bước 2:** Kéo thả Lookup để nối với bảng Order của nguồn



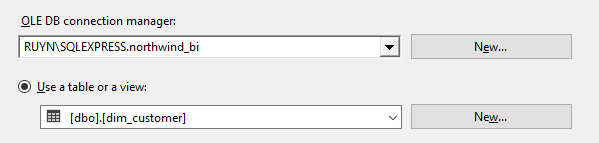
Hình 10.2: Lookup với bảng Order

Phần Column, nối thuộc tính liên quan là OrderID với nhau, chọn dữ liệu mà bảng Fact cần, ở đâychọn CustomeID, OrderDate, ShipCity



Hình 10.3: Chọn những thuộc tính cần lấy ở bảng Orders

**Bước 3:** Chọn Lookup (Lookup match Output) để kết nối với bảng dim\_customer



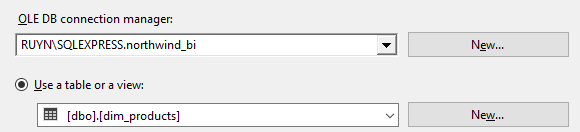
Hình 10.4: Lookup tới bảng dim\_customer

Phần Columns, nối CustomerID với nhau và chọn thuộc tính Customer\_key để đổ vào bảng Fact



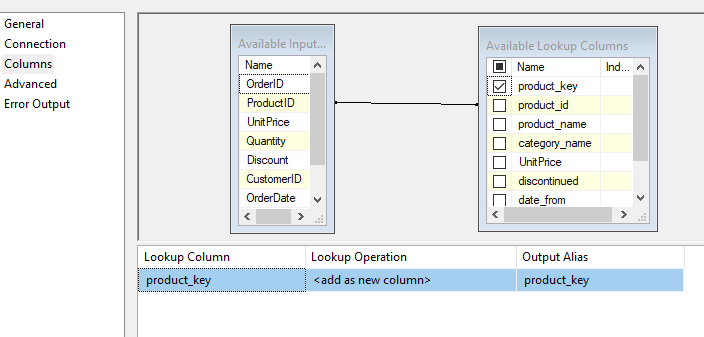
Hình 10.5: Chọn những thuộc tính cần lấy bảng dim\_customer

**Bước 3:** Tiếp tục kéo thả Lookup (Lookup match Output) để nối với bảng dim\_product



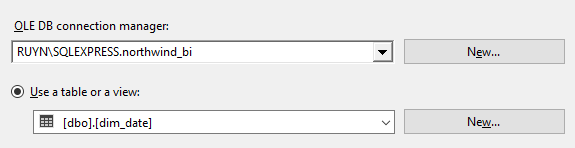
Hình 10.6: Lookup tới bảng dim\_products

Phần Columns, nối Producid với nhau và chọn product\_key để đổ vào bảng fact



Hình 10.7: Chọn những thuộc tính cần lấy bảng dim\_products

**Bước 4:** Tương tự lại kéo thả  Lookup (Lookup match Output) để nối với bảng Dim\_Date



Hình 10.8: Lookup tới bảng dim\_date

Phần Columns, nối OrderDate với Date và chọn date\_key để đổ vào Fact



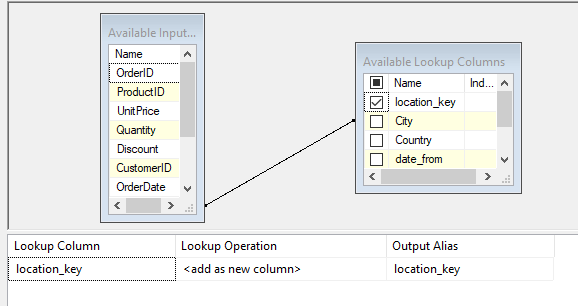
Hình 10.9: Chọn những thuộc tính cần lấy bảng dim\_date

**Bước 5:** Lookup (Lookup match Output) để kết nối với bảng dim\_location



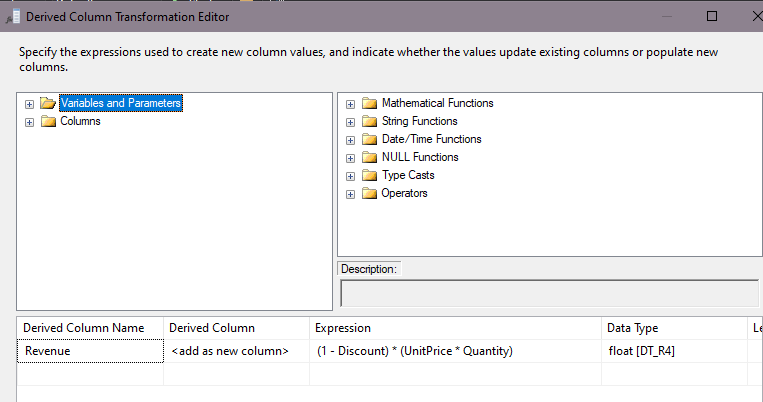
Hình 10.10: Lookup tới bảng dim\_location

Phần Columns, chọn location\_key để đổ vào fact



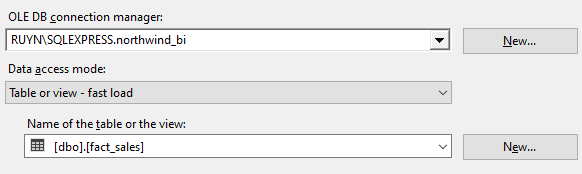
Hình 10.11: Chọn những thuộc tính bảng dim\_location

**Bước 6:** Kéo thả Derived Column để thực hiện tính toán cho thuộc tính Revenue



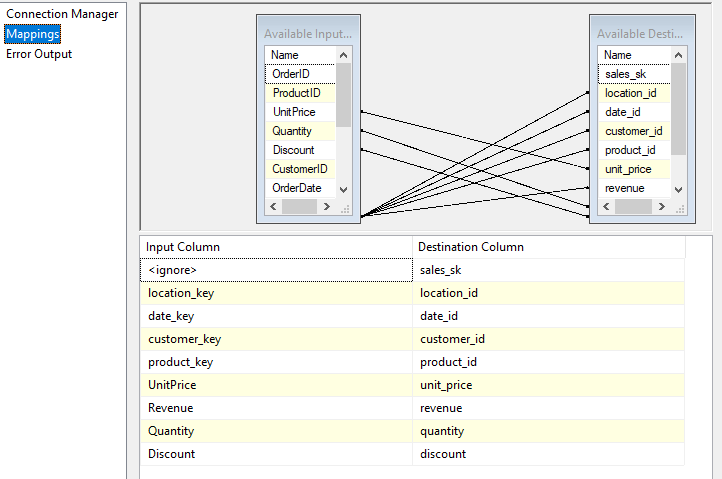
Hình 10.12: Xây dựng công thức tính doanh thu

**Bước 7:** Kéo OLEDB Destination là đích của bảng Fact\_Sales



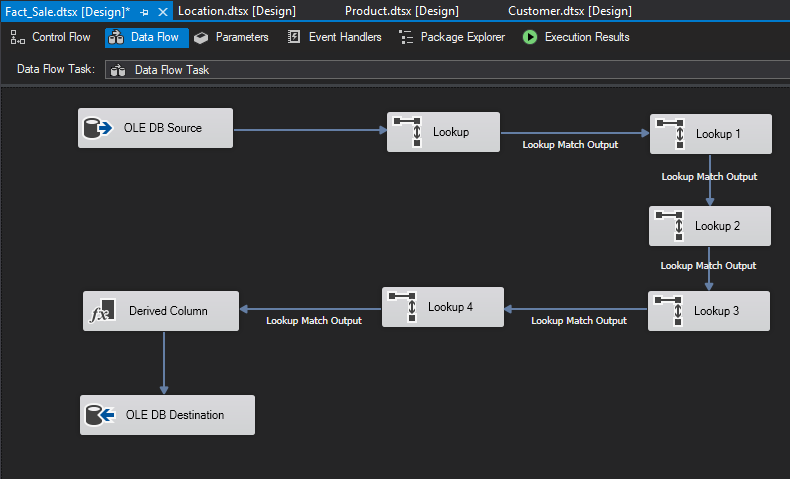
Hình 10.13: Đích bảng fact\_sales

Trong phần Mappings, nối các thuộc tính để đổ vào bảng fact:



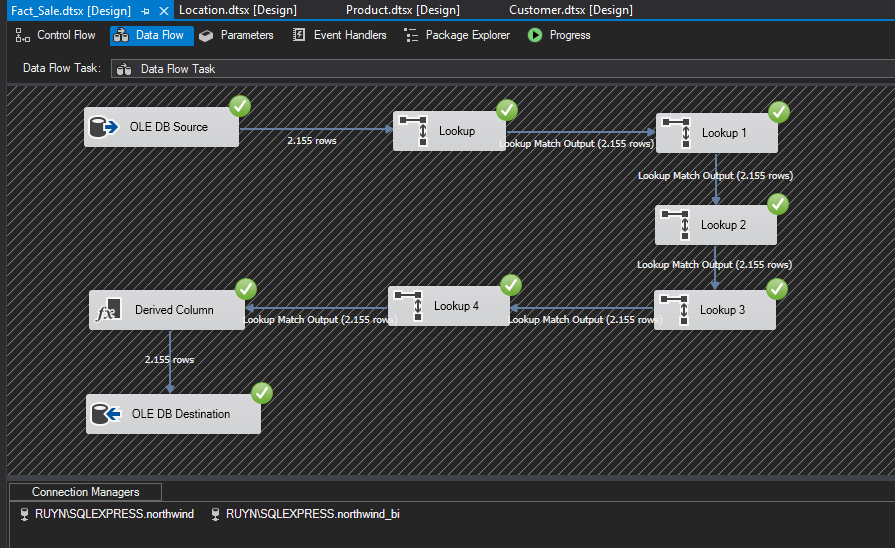
Hình 10.14: Nối các thuộc tính

Data Flow của bảng fact\_sales:



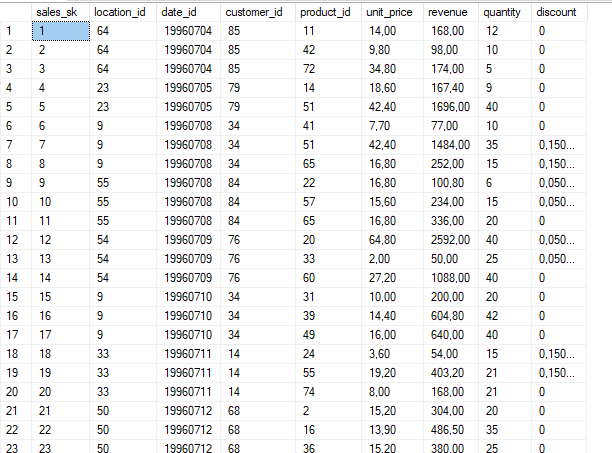
Hình 10.15: DataFlow bảng fact\_sales

Kết quả đổ vào bảng thành công:



Hình 10.16: Đổ thành công bảng fact\_sales

Dữ liệu sau khi đổ vào:



Hình 10.17: Dữ liệu bảng fact\_sales sau khi đổ

# Câu 11: Sử dụng kiểu SCD nào để tải dữ liệu (data loading) từ hệ thống NGUỒN vào bảng DIM? Minh họa (1 điểm)

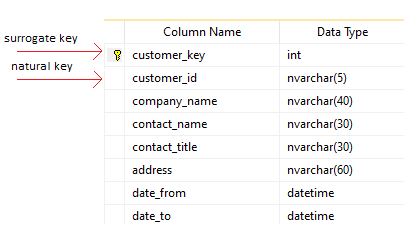
SCD là phương pháp giúp lưu trữ và quản lý chiều dữ liệu (dimension) hiện tại và lịch sử trong kho dữ liệu (DWH). Có 3 loại SCD cơ bản:

* SCD loại 1: Khi có thay đổi, giá trị cũ sẽ bị ghi đè bởi giá trị mới. Loại này không lưu trữ lịch sử thay đổi, chỉ giữ lại giá trị hiện tại nhất
* SCD loại 2: Khi có thay đổi, giá trị cũ sẽ được giữ lại và giá trị mới sẽ được thêm vào với một khóa chính mới. Loại này lưu trữ toàn bộ lịch sử thay đổi, cho phép phân tích xu hướng và biến động theo thời gian
* SCD loại 3: Khi có thay đổi, giá trị cũ sẽ được lưu vào một cột riêng và giá trị mới sẽ được ghi đè vào cột ban đầu. Loại này chỉ lưu trữ một phần lịch sử thay đổi, cho phép so sánh giá trị hiện tại với giá trị trước đó

Trong bài toán của Northwind, nhóm chọn xây dựng SCD loại 2 vì nó sẽ giúp theo dõi lịch sử thay đổi dữ liệu của các khách hàng, sản phẩm, giá. Điều này sẽ giúp các nhà phân tích dữ liệu và người dùng cuối có thể phân tích xu hướng, hành vi, hiệu suất của doanh nghiệp trong những khoảng thời gian khác nhau.

* ***Thực hiện SCD2 bảng dim\_customer***

Đối với bảng dim\_customer, để thực hiện kiểu SCD 2 nhóm đã xây dựng các thuộc tính thêm như date\_from, date\_to và 1 khóa thay thế (Surrogate Key) - khóa này sẽ tăng tự động và đảm bảo rằng khi mỗi thuộc tính thay đổi sẽ thêm một hàng mới với Surrogate Key khác nhau mà khóa tự nhiên (Natural Key) không đổi



Hình 11.1: Thuộc tính bảng dim\_customer

Nhóm đã chọn ***contact name*** là thuộc tính cần thay đổi khi cập nhật lên kiểu SCD2 vì khách hàng Northwind là công ty nên sẽ cử những người đại diện khác nhau đến để mua hàng, câu lệnh cập nhật bảng nguồn trong sql như sau:

UPDATE customers

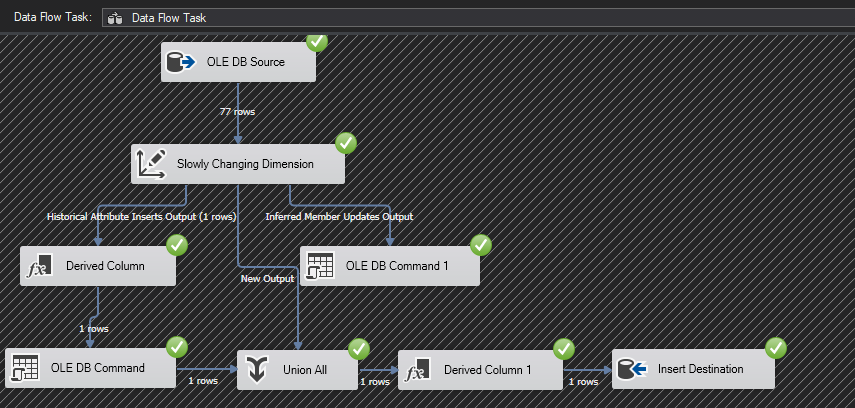
SET contactname = 'Jaekyung'

WHERE customerid = 'ALFKI';



Hình 11.2: Cập nhật bảng nguồn Customers

Kiểm tra dữ liệu khi update:



Hình 11.3 Kiểm tra cập nhật SCD2 dim\_customers

Kết quả:



Hình 11.4: Kết quả dim\_customer

* ***Thực hiện SCD2 bảng dim\_products***

Đối với bảng dim\_product, cũng tương tự, nhóm xây dựng thêm khóa thay thế và các thuộc tính thời gian date\_from, date\_to và lấy ***UnitPrice*** là thuộc tính cần thay đổi khi cập nhật kiểu SCD2:



Hình 11.5: Thuộc tính bảng dim\_products

Câu lệnh cập nhật nguồn:

UPDATE products

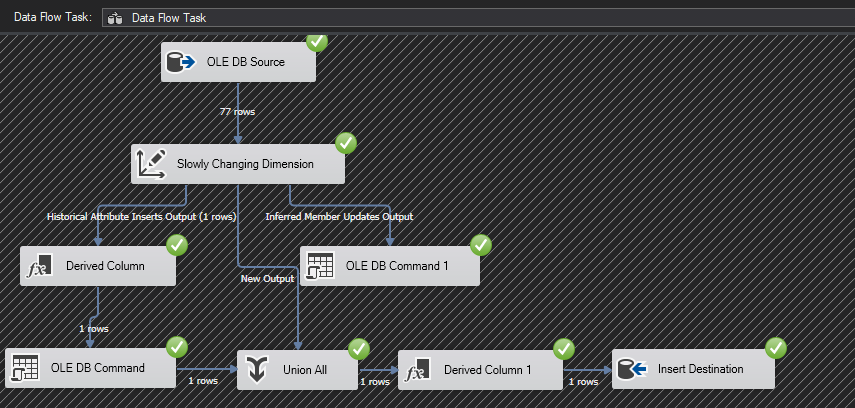
SET unitprice = 20

WHERE productid = 1;



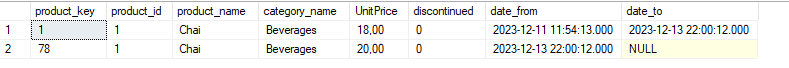
Hình 11.6: Cập nhật bảng nguồn Products

Kiểm tra dữ liệu sau khi cập nhật ở nguồn:



Hình 11.7: Kiểm tra cập nhật SCD2 dim\_products

Kết quả:



Hình 11.8: Kết quả dim\_products

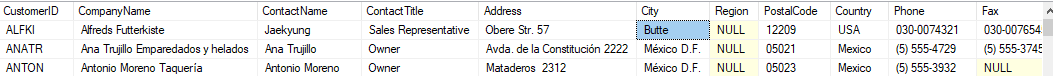
* ***Thực hiện SCD2 bảng dim\_location***

Đối với bảng dim\_location, tương tự nhóm chọn ***city, country*** là giá tri thay đổi khi cập nhật SCD 2, câu lệnh sql ở bảng nguồn:

SET city = 'Butte',

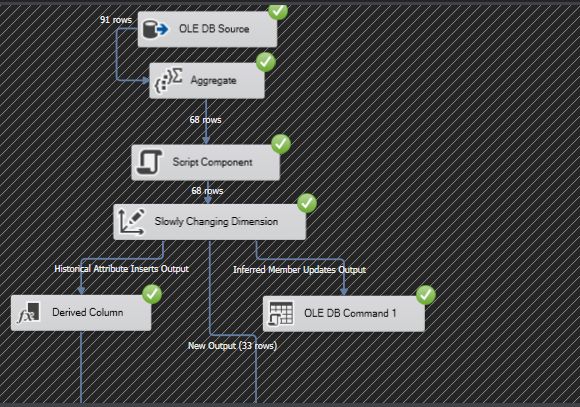
country = 'USA'

WHERE customerid = 'ALFKI';



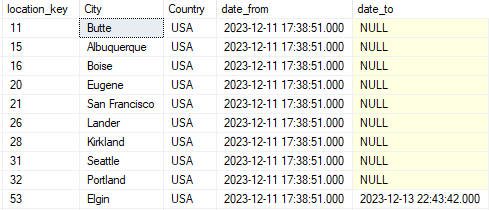
Hình 11.9: Cập nhật dữ liệu bảng Customers

Kiểm tra dữ liệu khi đổ:



Hình 11.10: Kiểm tra cập nhật SCD2 dim\_location

Dữ liệu sau khi thực hiện SCD 2:

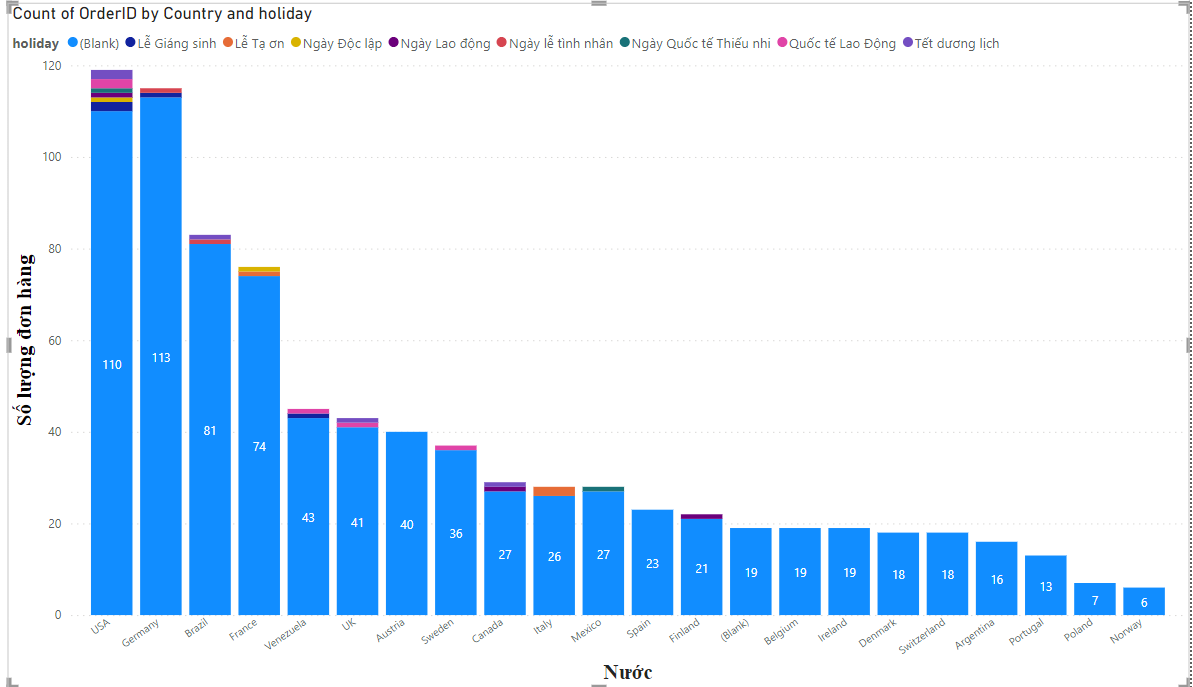


Hình 11.11: Kết quả dim\_location

# Câu 12: Thiết kế các báo cáo phân tích thống kê (tương ứng với câu 4 trong bài kiểm tra 1) sử dụng bất cứ tool nào (SQL Server Business Intelligence, Tableau, Power BI, Olap Excel,...) để minh họa các thao tác roll-up; drill-down; slice; dice (2 điểm)

## 12.1. Thống kê số lượng đơn hàng theo khu vực địa lý theo sự kiện

12.1.1. Thống kê số lượng đơn hàng theo khu vực quốc gia theo sự kiện sử dụng thao tác **Roll-up** trong Power BI.



Hình 12.1: Biểu đồ tổng số lượng đơn hàng theo quốc gia theo sự kiện

**Phân tích biểu đồ:**- Nhìn chung, USA là quốc gia có số lượng đơn hàng lớn nhất là 119 đơn hàng, chiếm 14,46%. Trong đó có 9 đơn hàng là trong ngày lễ và 110 đơn hàng không trong sự kiện ngày lễ

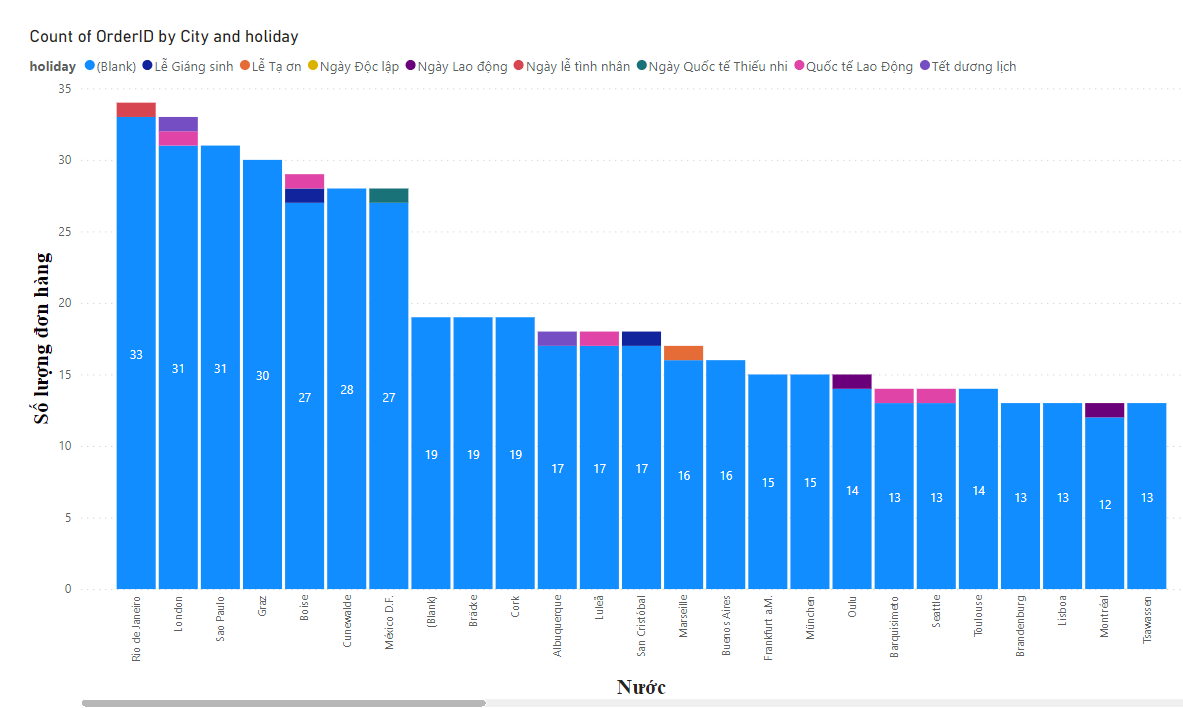
- Các quốc gia có số lượng đơn hàng thực phẩm lớn tiếp theo là Germany, Brazil, France, Austria, Venezuela…

- Các nước Switzerland, Argentina, Portugal, Porland, Norway có số lượng đơn hàng thấp hơn so với mức trung bình

- Cho thấy được sự phân cấp rõ nét trong việc tiêu thụ sản phẩm ở từng đất nước.

12.1.2. Thống kê số lượng đơn hàng theo khu vực thành phố của 1 đất nước sử dụng thao tác **Drill- down** trong Power BI.

Từ biểu đồ hình 12.1 ta sử dụng thao tác Drill-down được biểu đồ như sau:



Hình 12.2. Biểu đồ tổng số lượng đơn hàng theo thành phố

**Phân tích biểu đồ:**- Thành phố có nhiều đơn hàng nhất là Rio de Janeiro, với tổng số 34 đơn hàng trong đó có 1 đơn hàng trong ngày lễ tình nhân

* Các thành phố khác có số lượng đơn hàng đáng kể bao gồm:

London: 33 đơn hàng

Sau Paulo: 32 đơn hàng

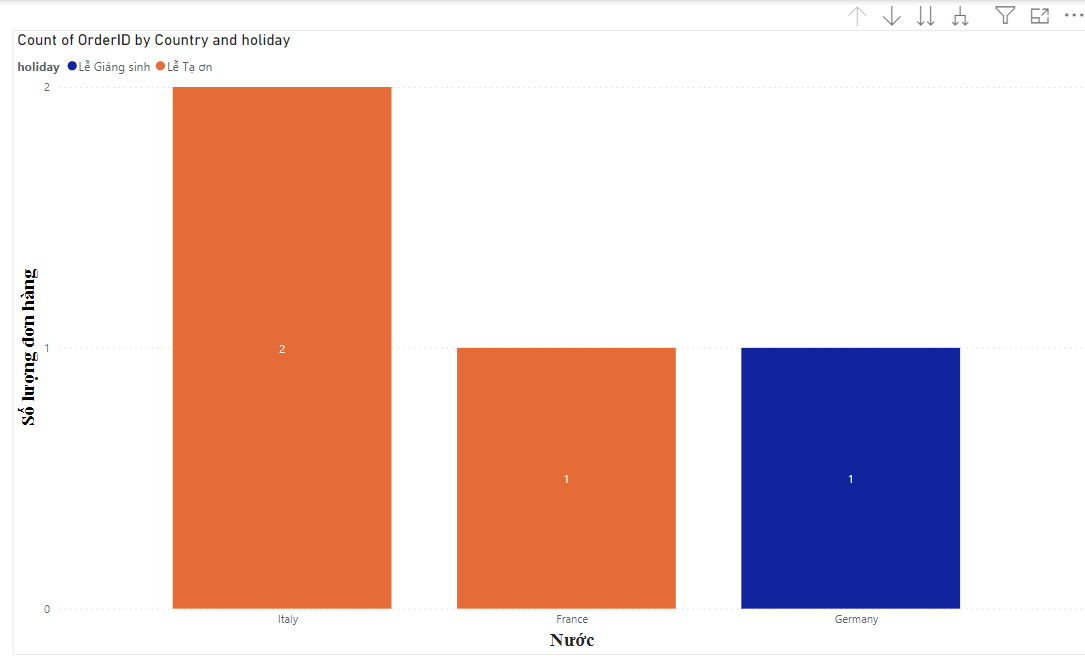
Graz: 30 đơn hàng

* Các thành phố có số lượng đơn hàng ít nhất bao gồm:

Caracas và Walla Walla cùng có 2 đơn hàng

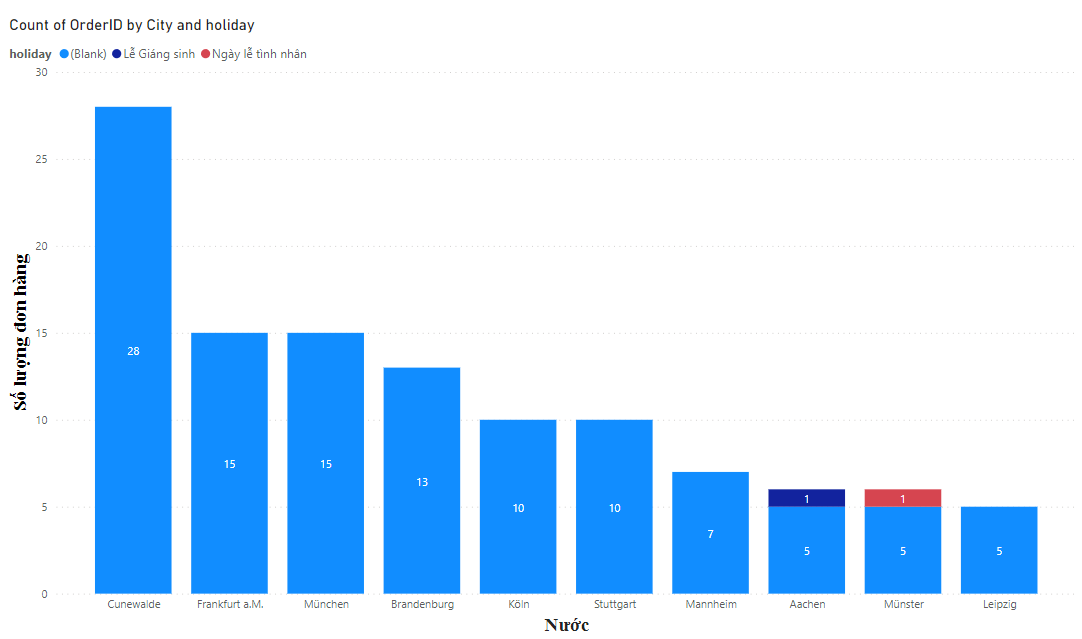
- Các thành phố lớn và đông dân thường có nhiều đơn hàng hơn các thành phố nhỏ và ít dân. Điều này càng chứng tỏ nhu cầu tiêu thụ sản phẩm ở các địa phương là khác nhau, từ đó Northwind cần đưa ra những chính sách và chương trình phù hợp cho từng khu vực.

12.1.3. Thống kê số lượng đơn hàng theo quốc gia cụ thể trong 1 sự kiện cụ thể sử dụng thao tác **dice** trong Power BI



Hình 12.3: Biểu đồ tổng số lượng đơn hàng ở Italy, France, Germany trong ngày Lễ giáng sinh và Lễ tạ ơn

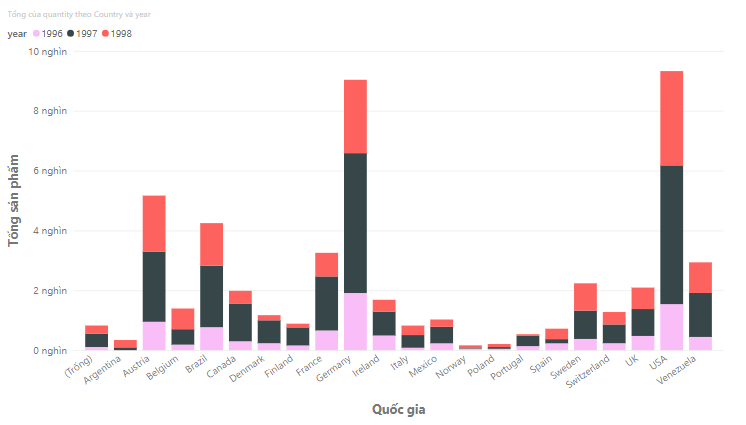
12.1.4. Thống kê số lượng đơn hàng tại Germany theo sự kiện cụ thể sử dụng thao tác **Slice** trong Power BI



Hình 12.4: Biểu đồ tổng số lượng đơn hàng tại Germany theo sự kiện cụ thể

## 12.2. Thống kê doanh số theo sản phẩm, theo khu vực và thời gian.

12.2.1. Thống kê doanh số sản phẩm theo khu vực đất nước và theo năm bằng thao tác **Roll - up** trên Power BI.



Hình 12.5: Biểu đồ tổng doanh số sản phẩm theo quốc gia theo năm

**Phân tích biểu đồ:**

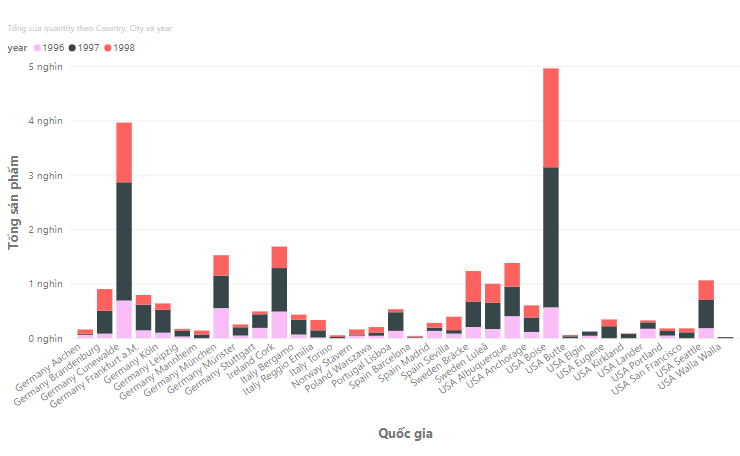
- Nhìn chung, tổng doanh số sản phẩm của các quốc gia đều tăng trưởng trong giai đoạn này. Tuy nhiên, tốc độ tăng trưởng không đồng đều giữa các quốc gia.

- Mỹ có số lượng tiêu thụ sản phẩm nhiều nhất với 9330 sản phẩm, tỉ lệ tiêu thụ tăng dần theo thời gian.

- Các quốc gia có số lượng đơn hàng thực phẩm lớn tiếp theo là Đức, Úc, Brazil và Pháp

- Cho thấy được sự phân cấp rõ nét trong việc tiêu thụ sản phẩm ở từng đất nước.

12.2.2. Thống kê doanh số sản phẩm theo thành phố của Germany và theo quý bằng thao tác **drill - down** trên Power BI.



Hình 12.6. Biểu đồ tổng doanh số sản phẩm theo thành phố trong các năm

**Phân tích biểu đồ:**

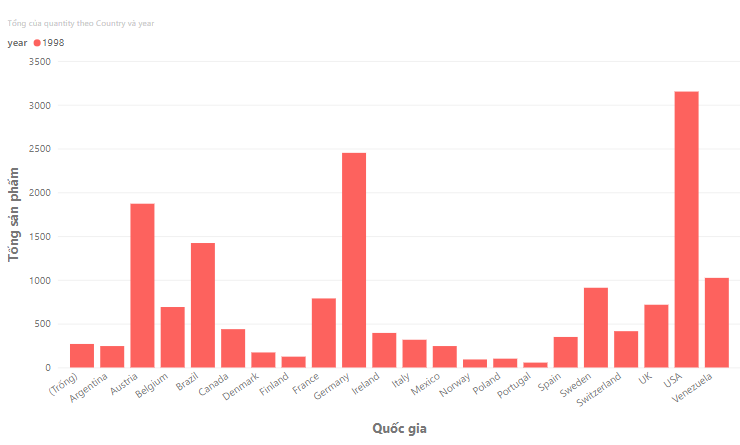
- Nhìn chung, tổng doanh số sản phẩm của các quốc gia đều tăng trưởng trong giai đoạn này. Tuy nhiên, tốc độ tăng trưởng không đồng đều giữa các quốc gia.

- Mỹ có số lượng tiêu thụ sản phẩm nhiều nhất với 9330 sản phẩm, tỉ lệ tiêu thụ tăng dần theo thời gian.

- Các quốc gia có số lượng đơn hàng thực phẩm lớn tiếp theo là Đức, Úc, Brazil và Pháp

- Cho thấy được sự phân cấp rõ nét trong việc tiêu thụ sản phẩm ở từng đất nước.

12.2.3. Thống kê doanh số sản phẩm của các thành phố ở Đức trong năm 1996 bằng thao tác **slice** trên Power BI.



Hình 12.7: Biểu đồ tổng doanh số sản phẩm theo thành phố trong năm 1998

**Phân tích biểu đồ:**

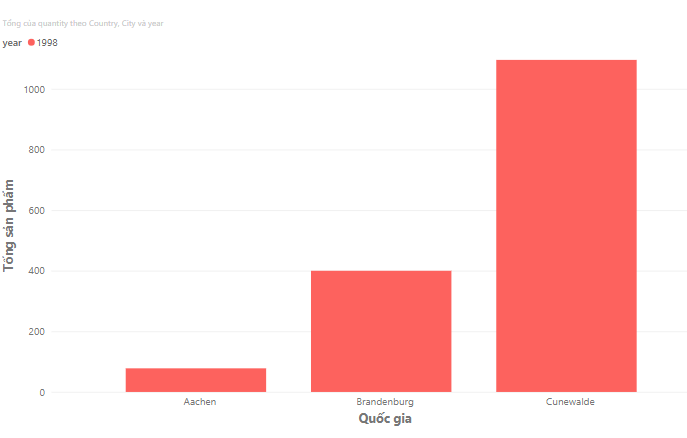
- Nhìn chung, tổng doanh số sản phẩm của các quốc gia đều tăng trưởng trong giai đoạn này. Tuy nhiên, tốc độ tăng trưởng không đồng đều giữa các quốc gia.

- Mỹ có số lượng tiêu thụ sản phẩm nhiều nhất với 3152 sản phẩm, tỉ lệ tiêu thụ tăng dần theo thời gian.

- Các quốc gia có số lượng đơn hàng thực phẩm lớn tiếp theo là Đức, Úc, Brazil và Pháp

- Cho thấy được sự phân cấp rõ nét trong việc tiêu thụ sản phẩm ở từng đất nước.

12.2.4. Thống kê doanh số sản phẩm của các thành phố cụ cụ thể ở Đức: Aachen, Brandenburg, Cunewald trong năm 1996 bằng thao tác **slice** trên Power BI.

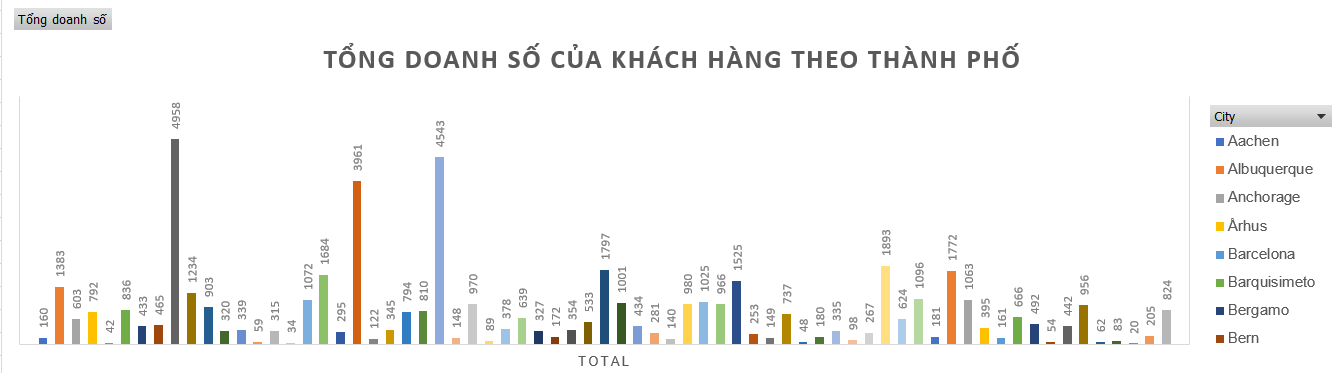


Hình 12.8: Biểu đồ tổng doanh số sản phẩm theo thành phố Aachen, Brandenburg, Cunewalde của Đức trong năm 1996

## 12.3. Thống kê tổng doanh số theo khách hàng, theo thời gian và theo danh mục sản phẩm Bằng Excel

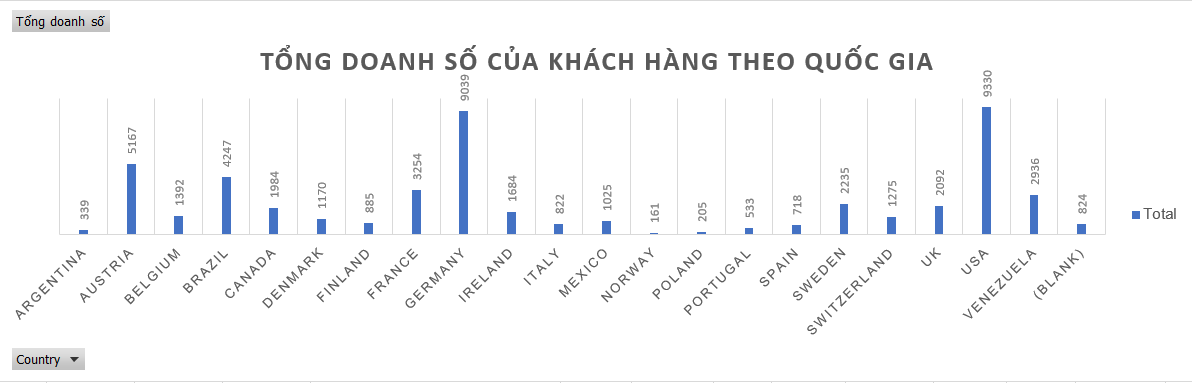
12.3.1. Minh họa thao tác **Roll-up**

Từ thống kê doanh số của khách hàng theo **từng thành phố:**



Hình 12.9: Biểu đồ doanh số của khách hàng theo từng thành phố

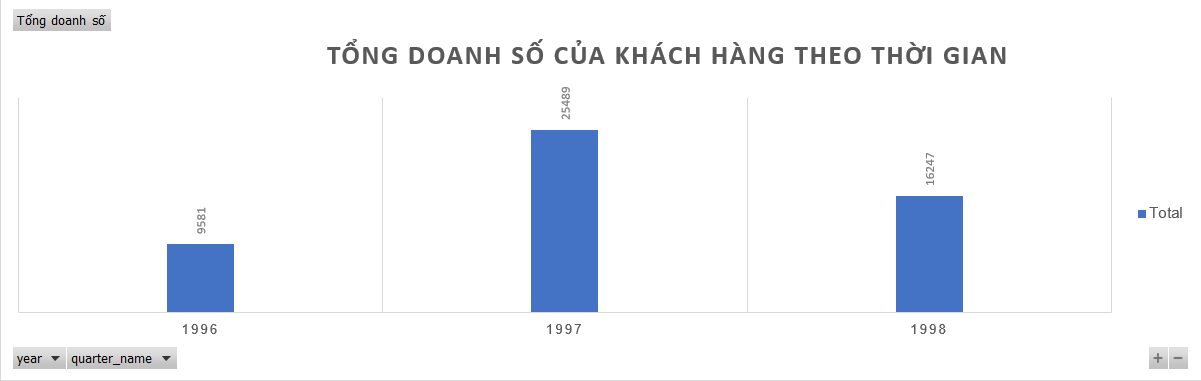
Đưa ra thống kê doanh số của khách hàng theo **từng quốc gia:**

****

Hình 12.10: Biểu đồ doanh số của khách hàng theo từng quốc gia

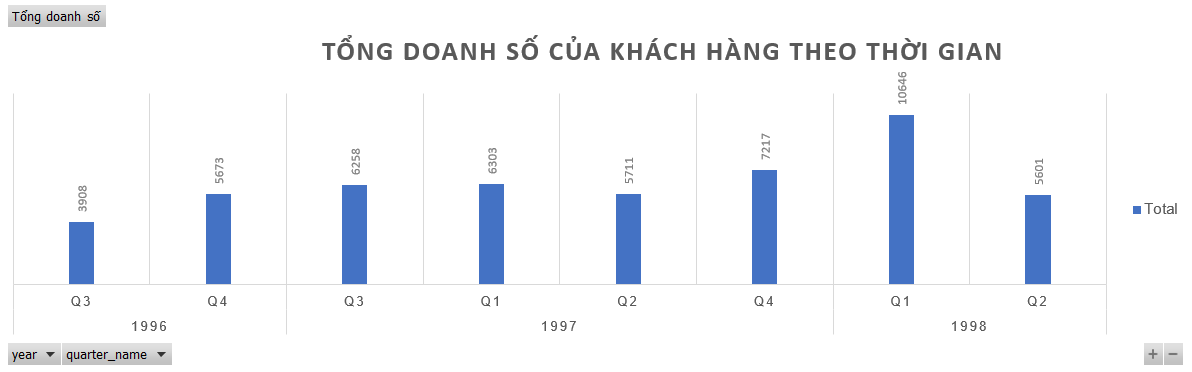
12.3.2. Minh họa thao tác **Drill down**

Từ thống kê doanh số của khách hàng theo từng **năm**



Hình 12.11: Biểu đồ tổng doanh số theo từng năm

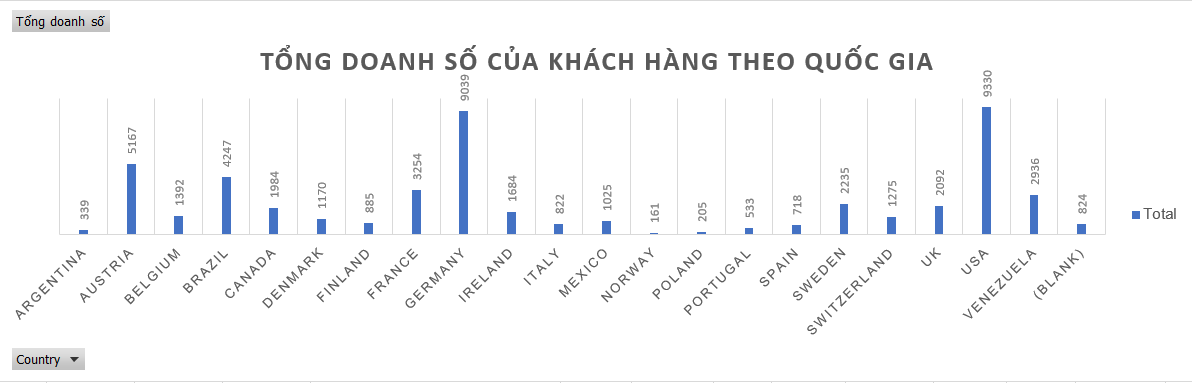
Đưa ra thống kê doanh số của khách hàng theo từng **quý:**



Hình 12.12: Biểu đồ tổng doanh số theo từng quý

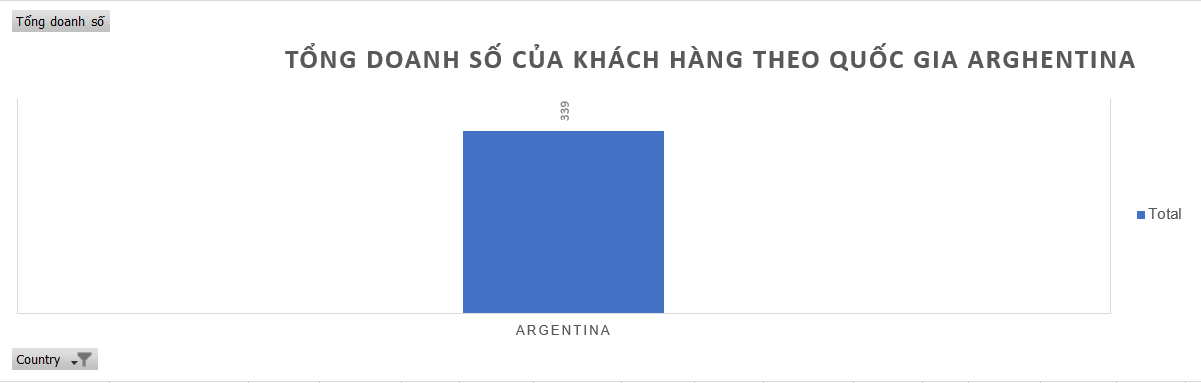
12.3.3. Minh họa thao tác Slice

Từ thống kê doanh số của khách hàng theo **từng quốc gia:**

****

Hình 12.13: Biểu đồ tổng doanh số theo từng quốc gia

Đưa ra thống kê doanh số của khách hàng ở **Argentina:**

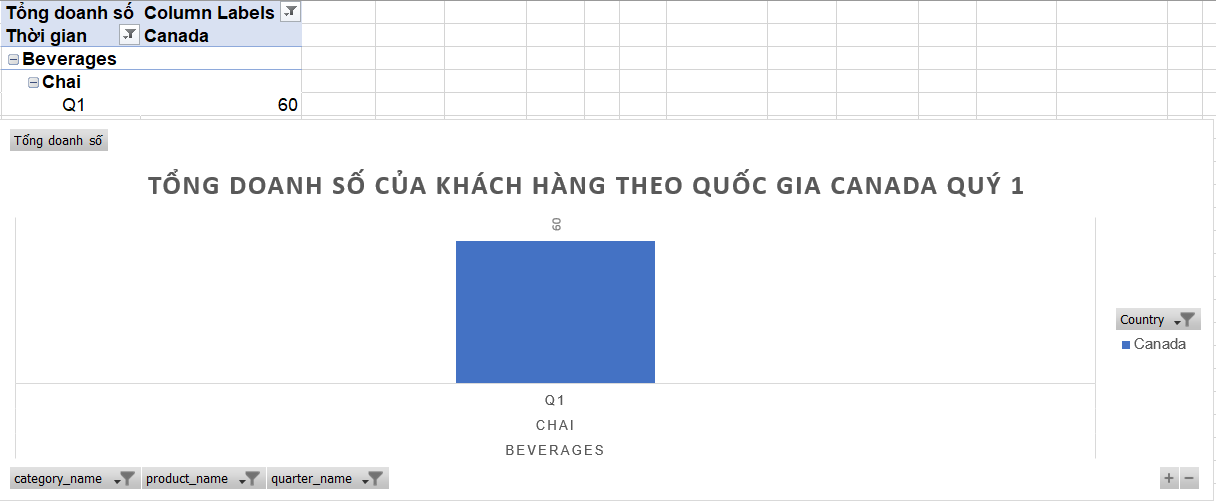


Hình 12.14: Biểu đồ tổng doanh số theo quốc gia Arghentina

12.3.4. Minh họa thao tác **Dice**

Đưa ra thống kê doanh số của khách hàng theo **sản phẩm Chai ở quốc gia Canada**

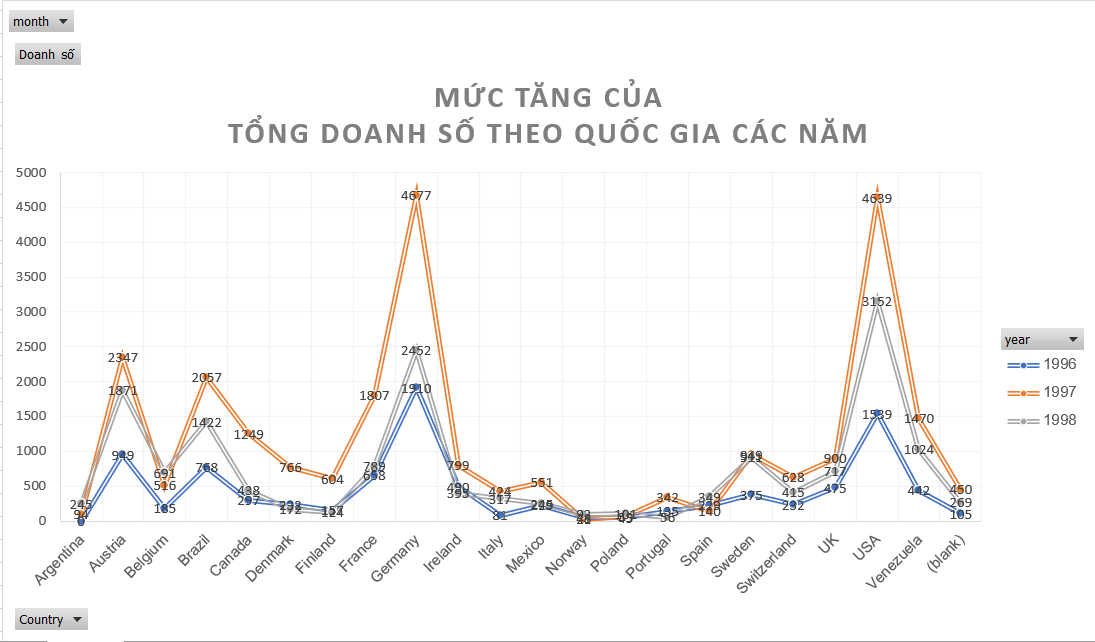
**Quý I**



Hình 12.15: Biểu đồ tổng doanh số theo sản phẩm Chai ở quốc gia Canada

## 12.4. Mức tăng trưởng doanh số của sản phẩm theo khu vực so với cùng kỳ năm trước bằng Excel

12.4.1. Mức tăng trưởng doanh số của sản phẩm theo quốc gia qua các năm sử dụng **Roll up**

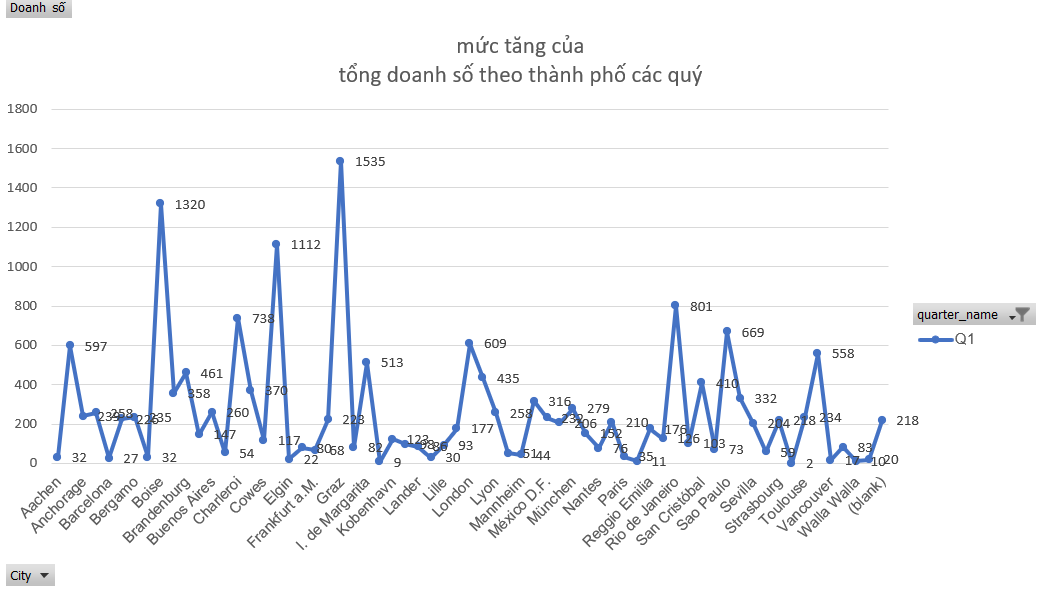


Hình 12.16. Biểu đồ mức tăng của doanh số theo quốc gia từng năm

Nhận xét:

Nhìn biểu đồ ta thấy rằng năm 1996 so với các năm còn lại có mức tổng doanh số tăng cao hơn hẳn. Cụ thể, ở Quốc gia Germany và USA tổng doanh số ở nhóm cao nhất với 4677 và 4689 sản phẩm. Năm 1998, tổng doanh số tiêu thụ là thấp nhất, quốc gia có số lượng sản phẩm tiêu thụ cao nhất ở năm này chỉ là 3162 thuộc về USA và thấp dần trải sang các nước khác, đạt 67,4% so với năm 1996

12.4.2 Mức tăng trưởng doanh số của sản phẩm theo thành phố qua các quý sử dụng **Drill down**



Hình 12.17: Biểu đồ mức tăng của doanh số theo thành phố từng quý

Ta đưa ra một số nhận xét về báo cáo như sau:

* Thành phố Graz có mức tăng trưởng doanh số cao nhất trong Q1.
* Thành phố Baise có mức tăng trưởng doanh số cao thứ hai.
* Thành phố Raigo Emilina có mức tăng trưởng doanh số thấp nhất, mức tăng trưởng này là thấp hơn so với mức trung bình của tất cả các thành phố.

Tóm lại, các thành phố đều có sự tăng trưởng nhưng chưa đồng đều

## 12.5. Thống kê tổng doanh thu và tổng số tiền giảm giá theo sản phẩm và theo thời gian bằng Power BI

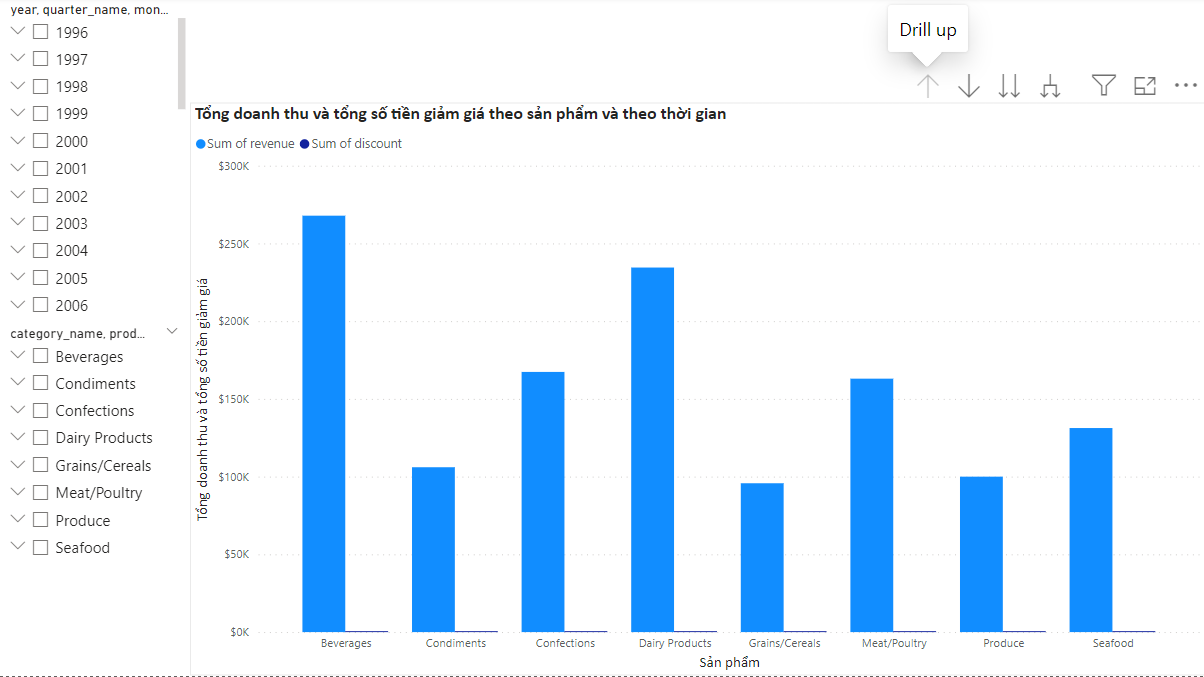
12.5.1. Minh họa thao tác **Roll-up**

Từ thống kê tổng doanh thu và tổng số tiền giảm giá theo từng **sản phẩm**:



Hình 12.18: Biểu đồ tổng doanh thu và tổng số tiền giảm giá theo sản phẩm

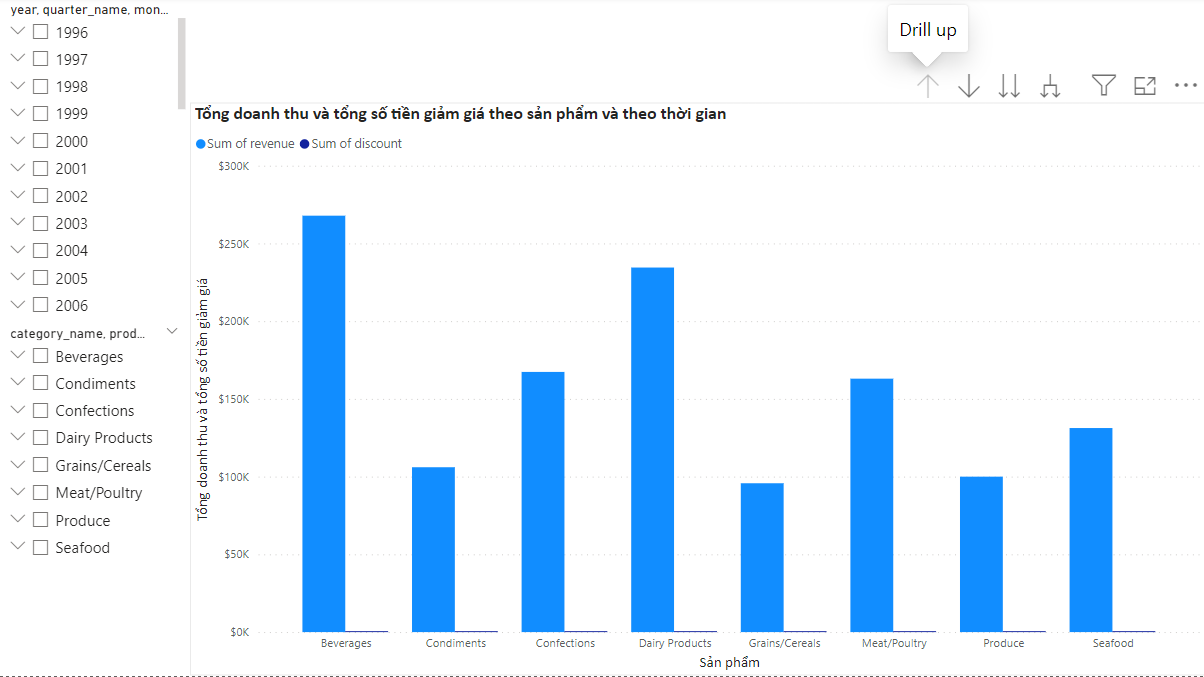
Đưa ra thống kê tổng doanh thu và tổng số tiền giảm giá theo từng **danh mục sản phẩm**:



Hình 12.19: Biểu đồ tổng doanh thu và tổng số tiền giảm giá theo danh mục sản phẩm

12.5.2. Minh họa thao tác **Drill-down**

Từ thống kê tổng doanh thu và tổng số tiền giảm giá theo từng **danh mục sản phẩm**:



Hình 12.20: Biểu đồ tổng doanh thu và tổng số tiền giảm giá theo danh mục sản phẩm

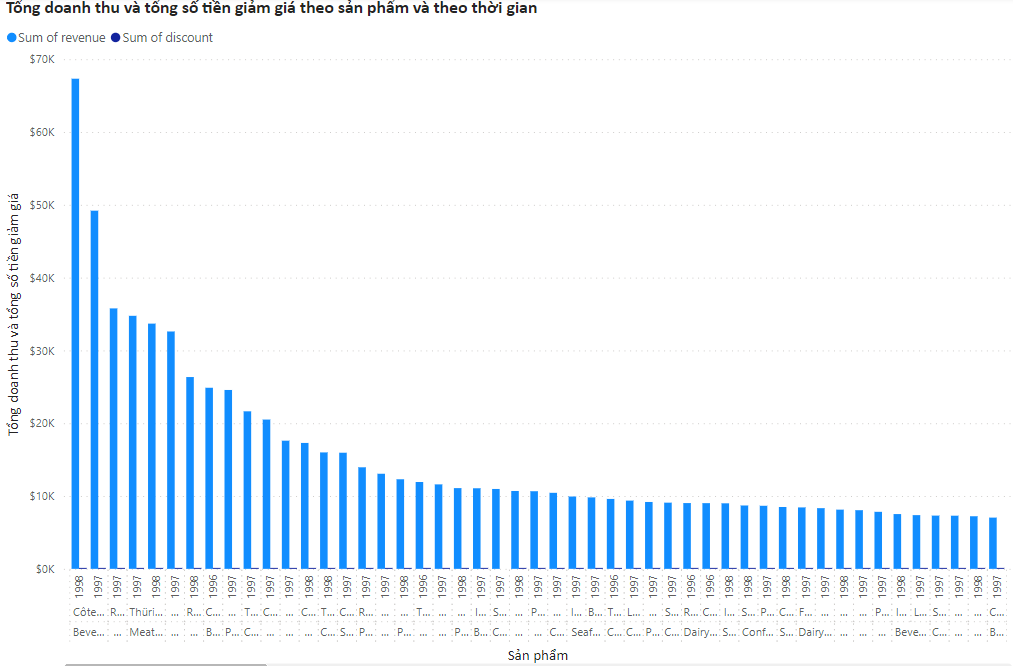
Đưa ra thống kê tổng doanh thu và tổng số tiền giảm giá theo từng **sản phẩm**:



Hình 12.21: Biểu đồ tổng doanh thu và tổng số tiền giảm giá theo sản phẩm

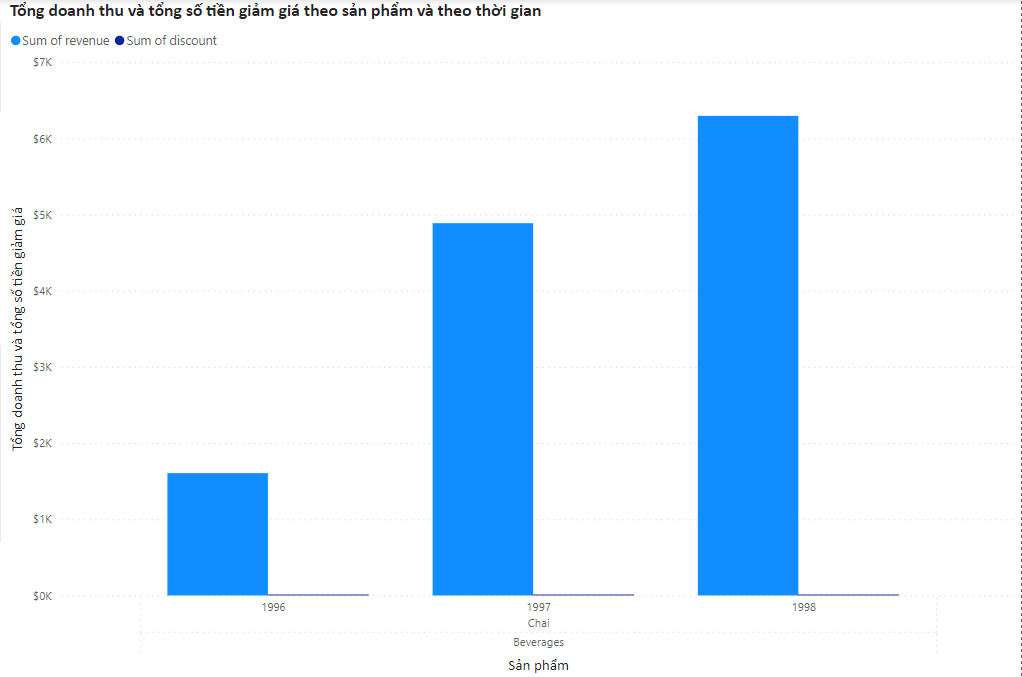
12.5.3. Minh họa thao tác **Slice**

Từ thống kê tổng doanh thu và tổng số tiền giảm giá của **từng sản phẩm theo thời gian**:



Hình 12.22: Biểu đồ tổng doanh thu và tổng số tiền giảm giá theo sản phẩm

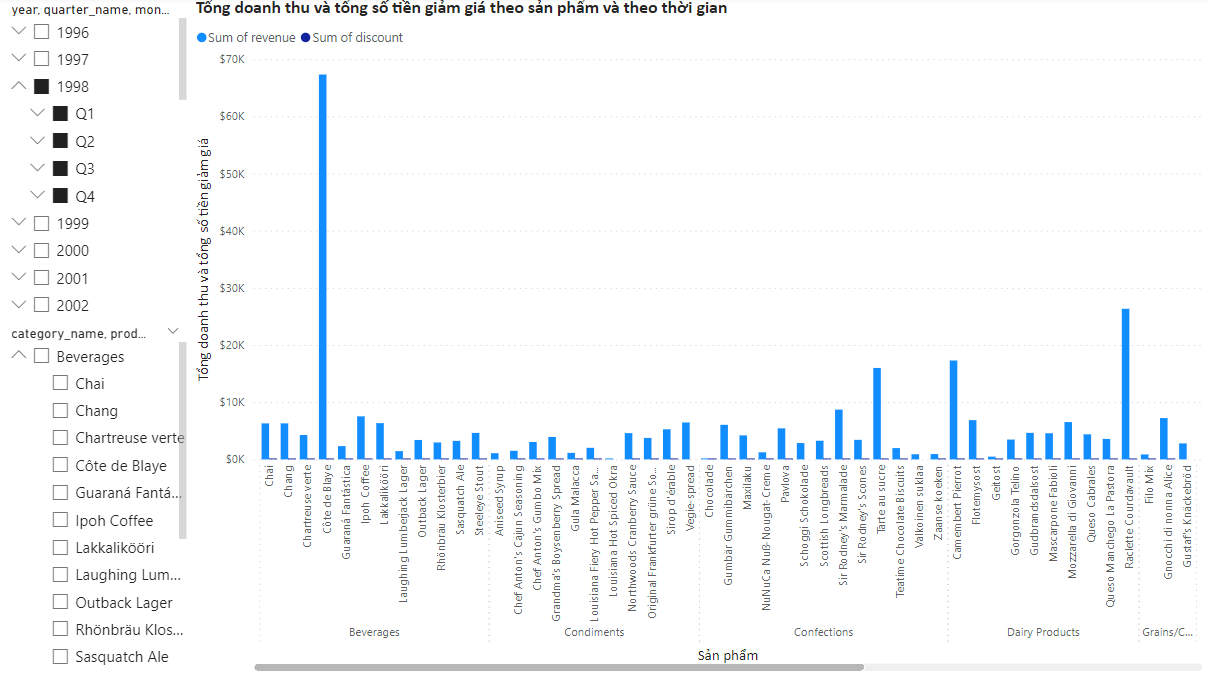
Đưa ra thống kê tổng doanh thu và tổng số tiền giảm giá của **sản phẩm Chai theo thời gian**:



Hình 12.23: Biểu đồ tổng doanh thu và tổng số tiền giảm giá của sản phẩm Chai

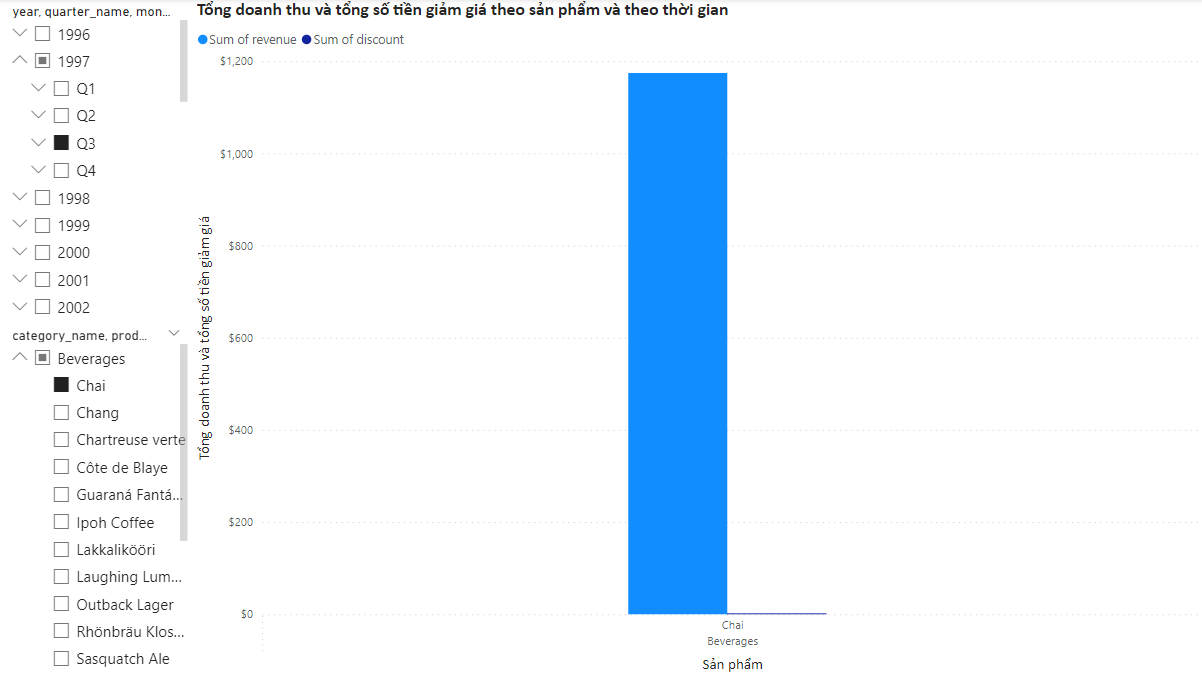
12.5.4. Minh họa thao tác **Dice**

Từ thống kê tổng doanh thu và tổng số tiền giảm giá của **từng sản phẩm theo từng quý**:



Hình 12.24: Biểu đồ tổng doanh thu và tổng số tiền giảm giá của sản phẩm theo quý

Đưa ra thống kê tổng doanh thu và tổng số tiền giảm giá của **sản phẩm Chai trong quý III**:



Hình 12.25: Biểu đồ tổng doanh thu và tổng số tiền giảm giá của sản phẩm Chai trong quý III

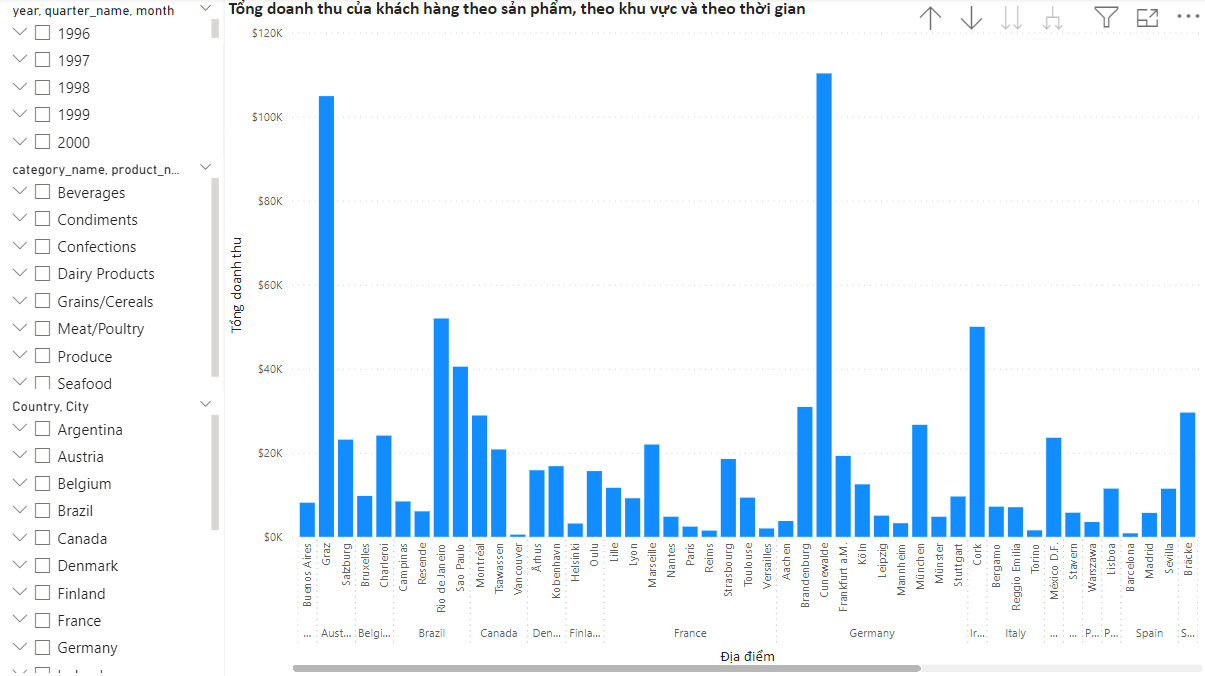
**Phân tích biểu đồ:**

* Nhìn chung, số tiền giảm giá rất nhỏ so với doanh thu của từng sản phẩm.
* Xét theo từng sản phẩm thì Cote de Blaye đang là sản phẩm tạo ra nhiều doanh thu nhất, điều này dẫn đến danh mục Đồ uống “Beverages” là danh mục sản phẩm mang lại doanh thu cao nhất cho Northwind.

## 12.6. Thống kê doanh thu của khách hàng theo sản phẩm, theo khu vực và theo thời gian

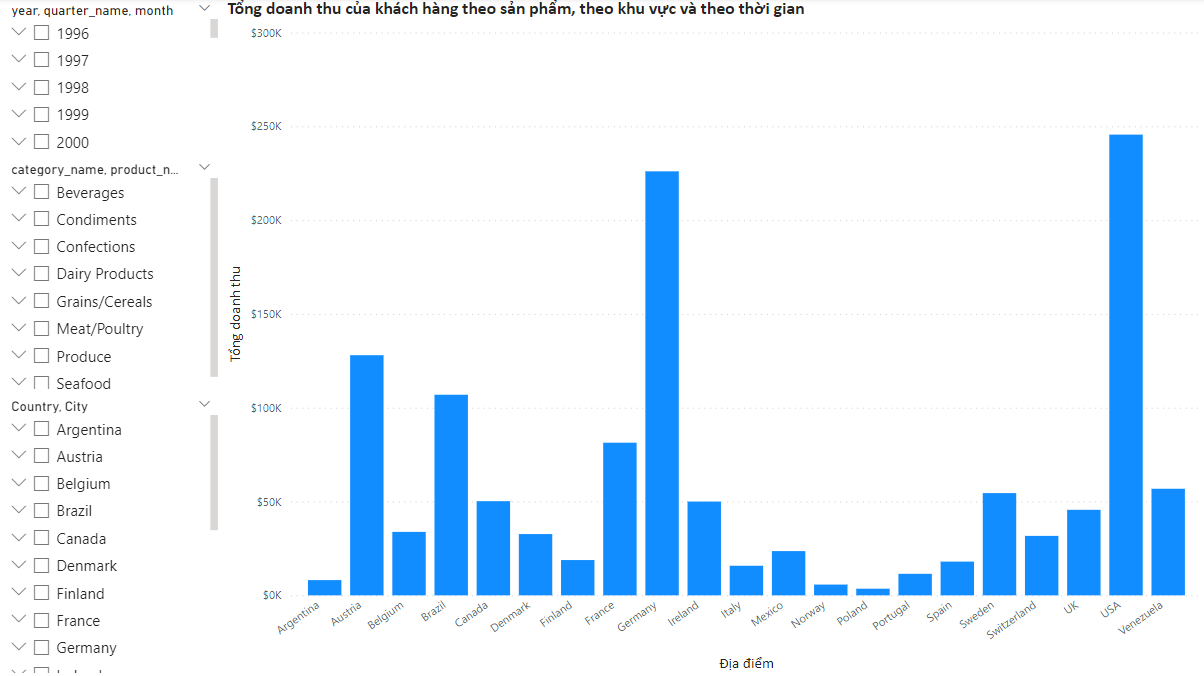
12.6.1. Minh họa thao tác **Roll-up**

Từ thống kê doanh thu của khách hàng theo **từng thành phố:**



Hình 12.26: Biểu đồ doanh thu của khách hàng theo từng thành phố

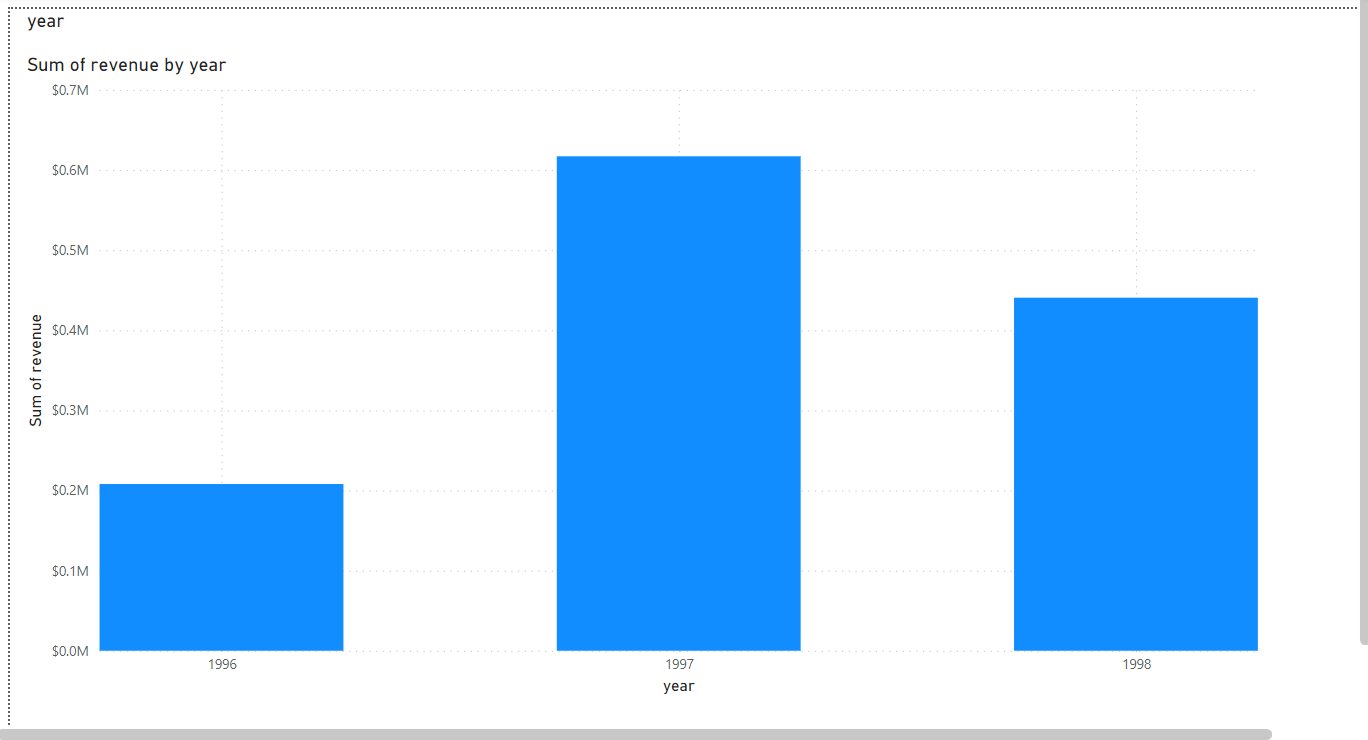
Đưa ra thống kê doanh thu của khách hàng theo **từng quốc gia:**



Hình 12.27: Biểu đồ doanh thu của khách hàng theo từng quốc gia

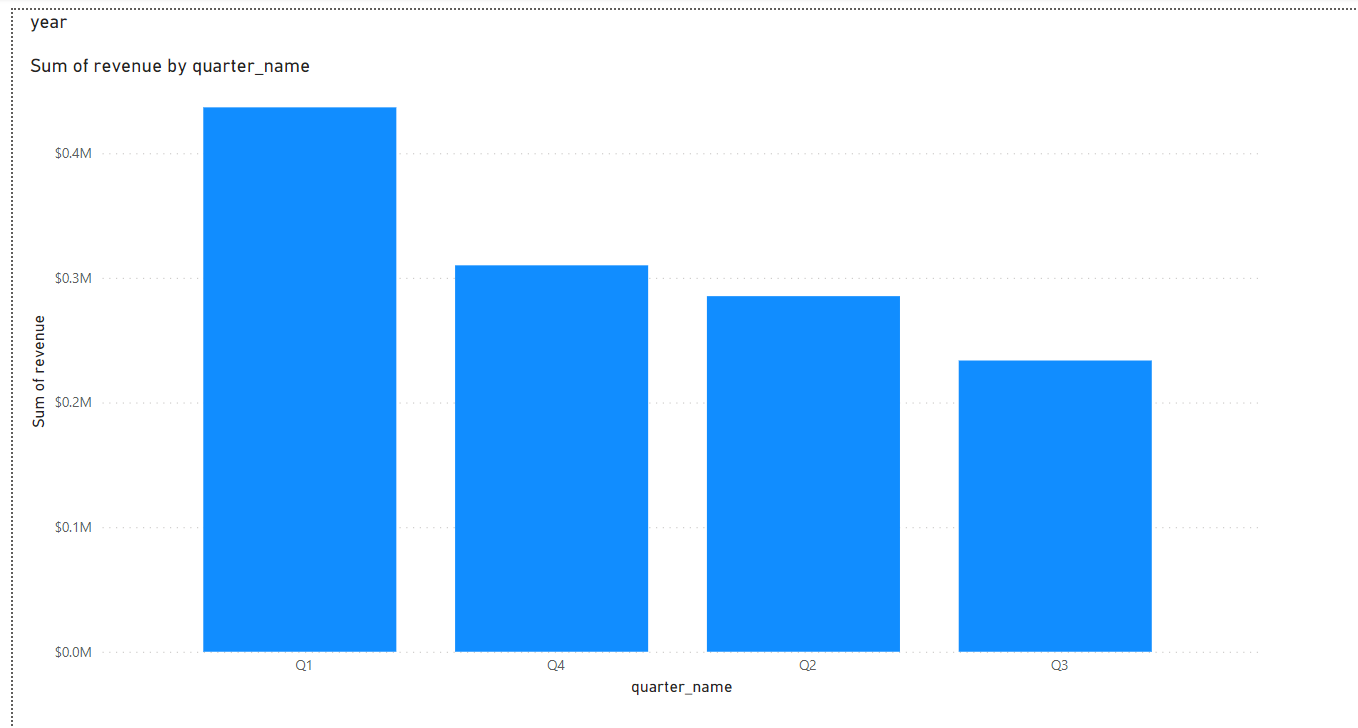
12.6.2. Minh họa thao tác **Drill-down**

Từ thống kê doanh thu của khách hàng theo từng **tháng**



Hình 12.28: Biểu đồ doanh thu của khách hàng theo từng quý

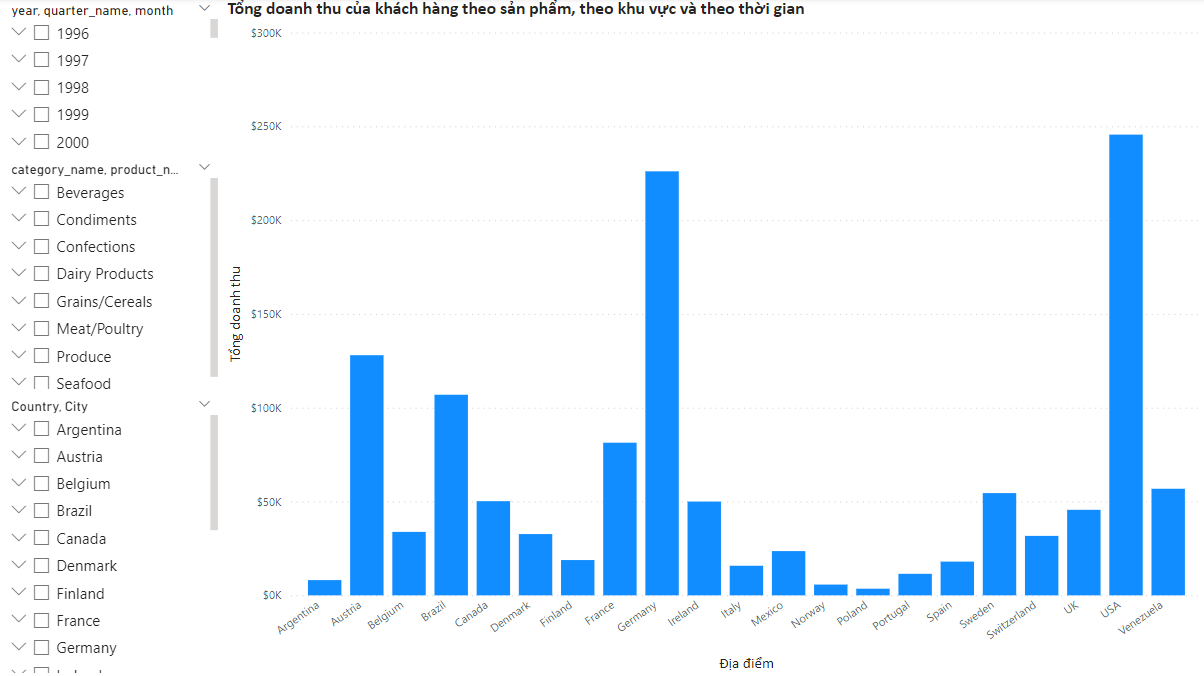
Đưa ra thống kê doanh thu của khách hàng theo từng **tháng:**

****

Hình 12.29: Biểu đồ doanh thu của khách hàng theo từng tháng

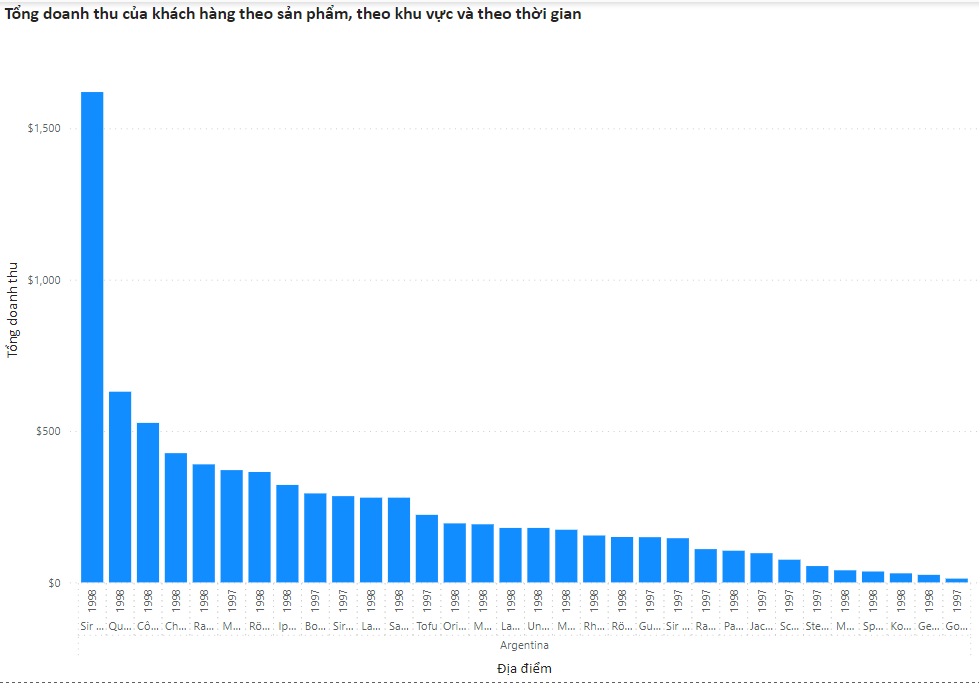
12.6.3. Minh họa thao tác **Slice**

Từ thống kê doanh thu của khách hàng **theo sản phẩm, theo khu vực và theo thời gian:**



Hình 12.30: Biểu đồ doanh thu của khách hàng theo sản phẩm, theo khu vực và theo thời gian

Đưa ra thống kê doanh thu của khách hàng ở **Argentina, theo sản phẩm, theo thời gian:**

****

Hình 12.31: Biểu đồ doanh thu của khách hàng ở Argentina

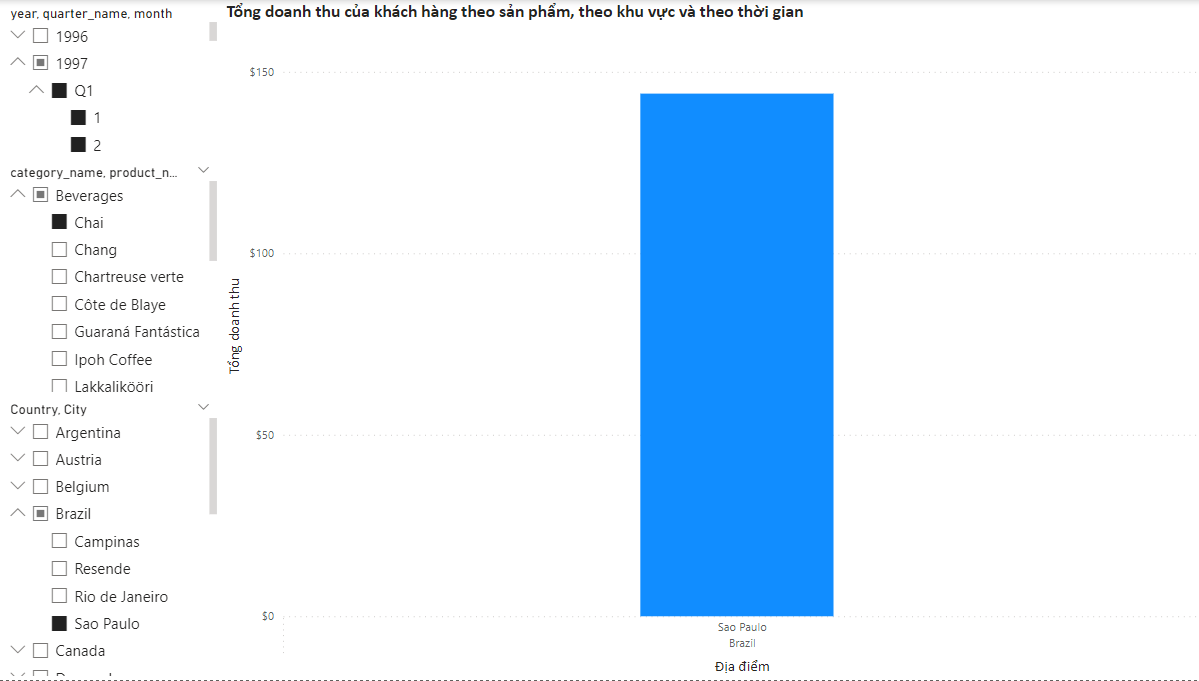
12.6.4. Minh họa thao tác **Dice**

Từ thống kê doanh thu của khách hàng theo **từng sản phẩm ở từng quốc gia:**



Hình 12.32: Biểu đồ doanh thu của khách hàng theo từng sản phẩm ở từng quốc gia

Đưa ra thống kê doanh thu của các khách hàng mua **sản phẩm Chai ở Sao Paulo trong Quý I:**



Hình 12.33: Biểu đồ doanh thu của các khách hàng mua sản phẩm Chai ở Sao Paulo trong Quý I

**Phân tích biểu đồ:**

* Dựa vào biểu đồ ta thấy doanh thu từ các khách hàng ở Cunewalde, Đức là nhiều nhất với hơn 110,2 nghìn đô; tiếp sau đó là Graz, Áo với hơn 104,8 nghìn đô và Boise của Mỹ với hơn 104,3 nghìn đô. Điều này cho thấy tệp khách hàng chính của Northwind tập trung ở Trung Âu và Mỹ.
* Các khách hàng mang lại nhiều doanh thu cho Northwind đều từ danh mục sản phẩm Đồ uống, cụ thể là sản phẩm Côte de Blaye.

# Câu 13: Sau quá trình thử nghiệm cài đặt hệ thống và khai thác dữ liệu trên Cube, nhóm hãy nêu một số đánh giá về hiệu quả của DW về: khả năng truy vấn, khai thác dữ liệu, ...? (0.5 điểm). Loại DB nào phù hợp để thiết kế DW? Giải thích? (0.5 điểm)

## 13.1. Khả năng truy vấn, khai thác dữ liệu trên Data warehouse

Trong câu hỏi 4 (4.2), nhóm đã thực hiện truy vấn từ nguồn để lấy dữ liệu doanh thu, doanh số của một khu vực trong khoảng thời gian cụ thể, để đánh giá hiệu quả truy vấn trên đích, nhóm thực hiện so sánh như sau:

|  |  |
| --- | --- |
| **Truy vấn trên nguồn** | **Truy vấn trên đích** |
| Doanh số sản phẩm năm 1997 tại Oulu | |
| WITH doanh\_so\_theo\_san\_pham AS (  select O.ProductID, SUM(O.Quantity) as Doanh\_so  from OrderDetails O join Orders S on O.OrderID = S.OrderID   where year(S.OrderDate) = 1997 and ShipCity = 'Oulu'    group by ProductID  )  SELECT \*  FROM doanh\_so\_theo\_san\_pham | select \* from dim\_location where city = 'Oulu'  select product\_id, sum(quantity) as 'Doanh số' from fact\_sales a join dim\_date b on a.date\_id = b.date\_key  where location\_id = 36 and year   = 1997 group by product\_id |
| Doanh thu sản phẩm theo khu vực Oulu trong tháng 6 năm 1997 | |
| WITH doanh\_thu\_san\_pham AS (   SELECT ProductID,           UnitPrice,           Quantity,           UnitPrice \* Quantity AS Doanh\_thu   FROM OrderDetails O    JOIN Orders S ON O.OrderID = S.OrderID  WHERE ShipCity = 'Oulu' and month(OrderDate) = 6 and year(OrderDate) = 1997  GROUP BY ProductID,           UnitPrice,           Quantity  )  SELECT \*  FROM doanh\_thu\_san\_pham | select product\_id, sum(revenue) as 'Doanh thu' from fact\_sales a join dim\_date b on a.date\_id =b.date\_key  where location\_id = 36 and month = 6 and year = 1997 group by product\_id |

Bảng 13.1: So sánh truy vấn dữ liệu ở đích và nguồn

Như vậy, qua bảng trên nhóm nhận thấy data warehouse phù hợp hơn cho các truy vấn dữ liệu phức tạp, yêu cầu tổng hợp dữ liệu theo nhiều chiều khác nhau. Trong khi đó, database phù hợp hơn cho các truy vấn dữ liệu đơn giản, yêu cầu tính toàn vẹn và khả năng cập nhật dữ liệu. Cụ thể khi khai thác dữ liệu trên đích:

***Tính tổng hợp:*** Dữ liệu trong data warehouse được tổng hợp theo các chiều khác nhau, giúp cho việc truy vấn dữ liệu trở nên nhanh chóng và hiệu quả hơn.

***Tốc độ truy vấn nhanh hơn***: Do dữ liệu được lưu trữ ở dạng tập hợp dữ liệu tích hợp, với các cột được liên kết với nhau theo một cách nhất quán.

***Câu truy vấn đơn giản hơn:*** Cho phép người dùng thực hiện các truy vấn dữ liệu phức tạptheo nhiều chiều mà không cần nối nhiều bảng.

***Khả năng nén:*** Dữ liệu trong data warehouse có thể được nén hiệu quả hơn so với database, giúp tiết kiệm không gian lưu trữ và cải thiện hiệu suất truy vấn.

Còn về Database được thiết kế để lưu trữ và xử lý các giao dịch dữ liệu, vì vậy nó có khả năng truy vấn dữ liệu đơn giản tốt hơn data warehouse. Cụ thể, database có những ưu điểm sau:

***Tính toàn vẹn:*** Dữ liệu trong database phải luôn được đảm bảo tính toàn vẹn, vì dữ liệu này thường được sử dụng để thực hiện các giao dịch tài chính.

***Khả năng cập nhật:*** Database có thể được cập nhật một cách dễ dàng. Điều này giúp cho các doanh nghiệp có thể cập nhật dữ liệu một cách nhanh chóng và hiệu quả.

***Khả năng tích hợp:*** Database có thể được tích hợp với các hệ thống khác một cách dễ dàng. Điều này giúp cho các doanh nghiệp có thể dễ dàng truy cập và xử lý dữ liệu từ các hệ thống khác.

Về khả năng khai thác dữ liệu trên Data Warehouse (DW) thường được đánh giá dựa trên khả năng nắm bắt thông tin mới, hiểu biết sâu sắc về dữ liệu, và khả năng dự đoán xu hướng tương lai. Dưới đây là một số điểm mà Data Warehouse thường có ảnh hưởng tích cực đối với khả năng khai phá dữ liệu:

***Tích Hợp Dữ Liệu***: Data Warehouse tích hợp dữ liệu từ nhiều nguồn khác nhau, giúp tạo ra một tập hợp toàn diện và đa dạng về thông tin. Sự tích hợp này làm cho dữ liệu dễ truy cập và phân tích hơn, giúp người dùng khai phá mối quan hệ và hiểu biết sâu sắc hơn về tổ chức.

***Lịch Sử Dữ Liệu:*** Data Warehouse giữ lịch sử của dữ liệu, cho phép người dùng xem xét các thay đổi và xu hướng theo thời gian. Điều này làm cho việc khai phá dữ liệu trở nên linh hoạt hơn, đặc biệt khi đánh giá sự thay đổi trong các biến số kinh doanh theo thời gian.

***Dữ Liệu Đa Chiều:*** Các mô hình dữ liệu đa chiều như data cubes trong Data Warehouse giúp tạo ra các góc nhìn đa dạng về dữ liệu. Người dùng có khả năng phân tích dữ liệu từ nhiều chiều khác nhau, giúp phát hiện ra các mối quan hệ và xu hướng phức tạp.

***Công Cụ OLAP và Data Mining:*** Các công cụ hỗ trợ như OLAP và Data Mining được tích hợp trong môi trường Data Warehouse giúp người dùng khai phá dữ liệu một cách hiệu quả. Công cụ này cung cấp các phương pháp phân tích, gợi ý và dự đoán thông tin mới từ dữ liệu.

***Chất Lượng Dữ Liệu:*** Việc làm sạch và chuẩn hóa dữ liệu trong quá trình xây dựng Data Warehouse giúp cải thiện chất lượng dữ liệu. Dữ liệu chất lượng cao là yếu tố quan trọng để đảm bảo khả năng khai phá và đưa ra quyết định chính xác.

***Phân Tích Tính Năng Kinh Doanh:*** Data Warehouse thường tập trung vào việc cung cấp thông tin hỗ trợ quyết định kinh doanh. Điều này giúp người dùng khai phá thông tin về hiệu suất kinh doanh, xu hướng thị trường, và nhu cầu của khách hàng.

***Tính Năng Dự Đoán:*** Khả năng sử dụng các mô hình dự đoán và các thuật toán máy học trên dữ liệu trong Data Warehouse giúp người dùng khá

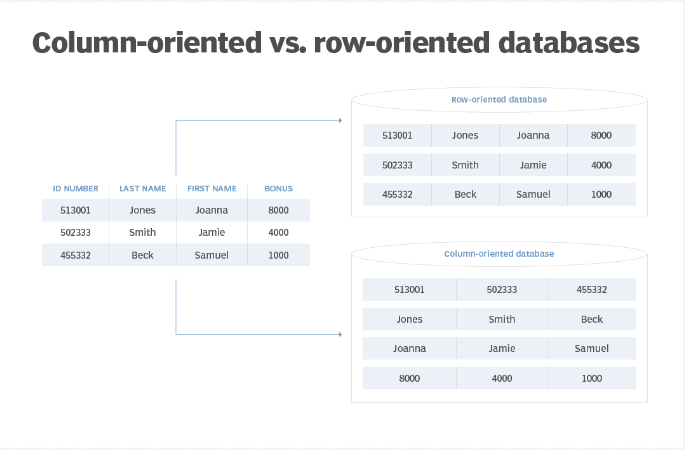
## 13.2. Loại Database phù hợp xây dựng Data warehouse

Ý tưởng kho dữ liệu hay Data Warehouse được ra đời khi doanh nghiệp cần hợp nhất các loại dữ liệu khác nhau nhằm gắn kết thông tin dữ liệu giữa các hệ thống để từ đó đưa ra những báo cáo BI hoàn chỉnh cho doanh nghiệp.

Để xây dựng Kho dữ liệu, trước hết bước đầu tiên luôn là xác định yêu cầu kho dữ liệu như truy vấn trong kho, dữ liệu lưu trữ, tốc độ như thế nào. Tiếp đến là thiết kế kiến trúc kho dữ liệu như là chọn loại Database sử dụng, quy trình tải dữ liệu, quản lý dữ liệu mà trong đó việc chọn loại Database là quan trọng nhất. Một số Database phổ biến hay được sử dụng, đó là Database quan hệ, Database dạng cột, Database MPP, Database NoSQL và giải pháp dựa trên đám mây. Khi chọn loại Database phù hợp với kho dữ liệu, ta cần xem xét đến yếu tố khối lượng dữ liệu, hiệu suất truy vấn, khả năng tích hợp dữ liệu và các công cụ có sẵn hỗ trợ lưu dữ liệu. Do vậy Database dạng cột phù hợp để xây dựng Data Warehouse.

Cơ sở dữ liệu dạng cột là một hệ quản trị cơ sở dữ liệu (DBMS) lưu trữ dữ liệu theo cột thay vì theo hàng. Mục đích của cơ sở dữ liệu dạng cột là để ghi và đọc dữ liệu một cách hiệu quả từ đĩa cứng, nhằm tăng khả năng và hiệu quả truy vấn.

Cơ sở dữ liệu dạng cột thay vì giữ một bản ghi của mỗi cột trong một bảng trong một hàng duy nhất, nó sẽ lưu trữ dữ liệu cho mỗi cột trong một cột duy nhất, hiệu suất nhanh hơn so với cơ sở dữ liệu dạng hàng. Đó là bởi vì nó truy cập ít bộ nhớ hơn để xuất dữ liệu. Vì cơ sở dữ liệu cột có thể lưu trữ nhiều dữ liệu hơn trong một lượng bộ nhớ nhỏ hơn. Và bởi vì việc truy xuất dữ liệu ban đầu được thực hiện trên cơ sở từng cột, chỉ những cột cần được sử dụng mới được truy xuất. Điều này giúp cơ sở dữ liệu cột có thể mở rộng quy mô hiệu quả và xử lý lượng lớn dữ liệu. Database dạng cột cũng được sử dụng cho các mục đích khác như chạy các khối xử lý phân tích trực tuyến OLAP theo nhiều chiều như các báo cáo nhóm đã thực hiện ở câu 12. Do vậy, nó hữu ích cho việc phân tích dữ liệu và thiết kế kho dữ liệu. Ngoài ra, việc tổng hợp dữ liệu của Database dạng cột nhanh chóng có lẽ vì không cần tải dữ liệu trước khi xuất ra như các bản ghi của hàng.



Hình 13.1: Sự khác biệt giữa CSDL dạng cột và dạng hàng

(Nguồn: techtarget)

Kết luận rằng CSDL có đầy đủ yếu tố phù hợp để thiết kế kho dữ liệu như:

* Các truy vấn dạng cột có thể được sử dụng để truy vấn dữ liệu theo cột một cách nhanh chóng và hiệu quả bởi vì việc lưu trữ dữ liệu theo cột giúp hệ thống có thể truy cập dữ liệu nhanh chóng, ngay cả đối với các truy vấn phức tạp liên quan đến nhiều cột dữ liệu.
* Sử dụng cho các ứng dụng phân tích dữ liệu. Các ứng dụng phân tích dữ liệu thường yêu cầu truy vấn dữ liệu theo cột, chẳng hạn như truy vấn để tìm tất cả các khách hàng từ một quốc gia cụ thể.
* Chạy các khối xử lý phân tích trực tuyến ([OLAP),](https://www.techtarget.com/searchdatamanagement/definition/OLAP) lưu trữ [siêu dữ liệu](https://www.techtarget.com/whatis/definition/metadata) và thực hiện phân tích thời gian thực.

# Câu 14: Khi dữ liệu từ NGUỒN được tải vào Data mart (tại bảng Dim, Fact) thì nhóm em có sử dụng kỹ thuật học máy, học sâu, luật kết hợp... để xây dựng mô hình dự báo, dự đoán (trên tập dữ liệu ở data mart) hỗ trợ ra quyết định mức chiến lược, chiến thuật không? (1 điểm)

Sau khi đổ dữ liệu từ nguồn vào Data mart, nhóm chúng em đã áp dụng kĩ thuật Machine Learning (học máy) để xây dựng mô hình “Dự báo số lượng sản phẩm tiêu thụ theo từng vùng trong thời điểm tới”.

## 14.1. Tổng quan về Machine Learning (Học máy)

14.1.1. Khái niệm Machine Learning

Machine Learning (Học máy) là một nhánh của trí tuệ nhân tạo (AI) và khoa học máy tính, tập trung vào việc sử dụng dữ liệu và thuật toán để học tập cách con người học, dần dần cải thiện độ chính xác của nó. (Theo IBM)

Bài toán của Machine Learning thường được chia làm hai loại là dự đoán (prediction) và phân loại (classification).

14.1.2. Các phương pháp Machine Learning

Có 4 phương pháp trong Machine Learning:

**- Học máy có giám sát (Supervised learning):** Máy tính được xem một số mẫu gồm đầu vào (input) và đầu ra (output) tương ứng trước. Sau khi học xong các mẫu này, máy tính quan sát một đầu vào mới và cho ra kết quả. Phương pháp này giúp các tổ chức giải quyết nhiều vấn đề trong thế giới thực trên quy mô lớn, chẳng hạn như phân loại thư rác trong một thư mục riêng biệt từ hộp thư đến của bạn.

- Một số phương pháp được sử dụng trong học máy có giám sát bao gồm:

     + Logistic regression

+ Neural networks

+ Linear regression

+ Naive bayes

+ Random forest

+ Support vector machine (SVM)

**- Học máy không giám sát (Unsupervised machine learning):** Máy tính chỉ được xem các mẫu không có đầu ra, sau đó máy tính phải tự tìm cách phân loại các mẫu này và các mẫu mới

Học máy không giám sát là phương pháp sử dụng các thuật toán máy học để phân tích và phân cụm các tập dữ liệu không được gắn nhãn. Khả năng phát hiện ra những điểm tương đồng và khác biệt trong thông tin của phương pháp này giúp ích cho việc phân tích dữ liệu khám phá, chiến lược bán chéo (cross-sell), phân khúc khách hàng cũng như nhận dạng hình ảnh và mẫu.

Học máy không giám sát cũng được sử dụng để giảm số lượng các tính năng trong một mô hình thông qua quá trình giảm kích thước. Phân tích thành phần chính và phân tích giá trị đơn lẻ là hai cách tiếp cận phổ biến cho việc này.

- Một số thuật toán được sử dụng trong học máy không giám sát bao gồm:

+ K-means clustering

+ Neural networks

+ Probabilistic clustering methods.

**- Học máy bán giám sát (Semi-supervised learning):** Một dạng lai giữa hai nhóm giải thuật.

**- Học tăng cường (Reinforcement Learning)**: Máy tính đưa ra quyết định hành động (action) và nhận kết quả phản hồi (response/reward) từ môi trường (environment). Sau đó máy tính tìm cách chỉnh sửa cách ra quyết định hành động của mình.

14.1.3. Quy trình xây dựng mô hình machine learning

Quy trình xây dựng mô hình machine learning gồm 5 bước:

- Bước 1: Thu thập dữ liệu (gathering data): Thu thập dữ liệu để mô hình học

- Bước 2: Tiền xử lý dữ liệu (Data pre-processing): Xử lý và đưa dữ liệu về định dạng tối ưu, trích chọn đặc trưng hoặc giảm chiều dữ liệu

- Bước 3: Huấn luyện mô hình (Training model): Tại pha này, thuật toán machine learning thực hiện việc học thông

qua các ví dụ đã được thu thập và chuẩn bị từ hai bước trên

- Bước 4: Đánh giá (evaluating):  Kiểm thử mô hình để đánh giá chất lượng của mô hình tốt đến đâu

- Bước 5: Cải thiện (Improve): Nếu kết quả từ bước đánh giá mô hình không khả quan thì machine learning phải được huấn luyện lại. Ta sẽ lặp lại bước 3, bước 4 cho đến khi đạt được độ chính xác như kỳ vọng.

14.1.4. Các ứng dụng của học máy

Học máy có ứng dụng rộng khắp trong các ngành khoa học/sản xuất, y tế, giáo

dục,… đặc biệt những ngành cần phân tích khối lượng dữ liệu khổng lồ.

Một số ứng dụng thường thấy:

- Xử lý ngôn ngữ tự nhiên (Natural Language Processing): xử lý văn bản, giao

tiếp người – máy, …

- Nhận dạng (Pattern Recognition): nhận dạng tiếng nói, chữ viết tay, vân tay, thị

giác máy (Computer Vision) …

- Tìm kiếm (Search Engine).

- Chẩn đoán trong y tế: phân tích ảnh X-quang, các hệ chuyên gia chẩn đoán bệnh

tự động.

- Tin sinh học: phân loại chuỗi gene, quá trình hình thành gene/protein.

- Vật lý: phân tích ảnh thiên văn, tác động giữa các hạt …

- Phát hiện gian lận tài chính (financial fraud): gian lận thẻ tỉn dụng.

- Phân tích thị trường chứng khoán (stock market analysis) Chơi trò chơi: tự động

chơi cờ, hành động của các nhân vật ảo.

## 14.2. Mô hình hồi quy tuyến tính (Linear Regression)

14.2.1. Khái niệm hồi quy tuyến tính

Linear Regression (hồi quy tuyến tính) là một thuật toán học máy có giám sát (supervised machine learning) trong học máy, nó là một phương pháp thống kê dùng để ước lượng mối quan hệ giữa các biến độc lập (input features) và biến phụ thuộc (output target). Linear Regression giả định rằng sự tương quan giữa các biến là tuyến tính, từ đó tìm ra hàm tuyến tính tốt nhất để biểu diễn mối quan hệ này. Thuật toán này dự báo giá trị của biến output từ các giá trị của các biến đầu vào.

14.2.2. Phân loại

Có 2 loại trong mô hình hồi quy tuyến tính đó là Simple Linear Regression (mô hình hồi quy đơn biến) và Multiple Linear Regression (mô hình hồi quy đa biến)

* Simple Linear Regression (mô hình hồi quy đơn biến)

Mô hình này chỉ có một biến độc lập (input feature) mô tả mối quan hệ tuyến tính giữa biến phụ thuộc (output target) và biến độc lập. Phương trình của Simple Linear Regression có dạng:

y = ax + b (trong đó x là biến độc lập và y là biến phụ thuộc vào x)

* Multiple Linear Regression (mô hình hồi quy đa biến)

Mô hình này có nhiều hơn một biến độc lập, biểu diễn mối quan hệ tuyến tính giữa các biến độc lập và biến phụ thuộc. Phương trình của Multiple Linear Regression có dạng: y = a1x1 + a2x2 + a3x3 + … + anxn

Trong đó:

x1, x2, …, xn: là các biến độc lập

y là biến phụ thuộc

14.2.3. Mục tiêu của Linear Regression

Mục tiêu của hồi quy tuyến tính là tìm ra hệ số góc và điểm giao với trục tung sao cho hàm dự đoán tuyến tính đạt được sai số nhỏ nhất. Một trong những cách phổ biến để ước lượng các hệ số là sử dụng phương pháp Ordinary Least Squares (OLS), trong đó chúng ta cần tối thiểu hóa tổng bình phương sai số (sum of squared error).

## 14.3. Tính cấp thiết trong xây dựng mô hình dự báo số lượng sản phẩm tiêu thụ theo vùng trong thời điểm tới

Dự báo số lượng sản phẩm tiêu thụ theo vùng trong thời gian tới mang lại nhiều ý nghĩa chiến lược quan trọng cho doanh nghiệp. Việc này không chỉ giúp xác định đối tượng mục tiêu và nhu cầu thị trường mà còn tạo ra cơ hội để tối ưu hóa chiến lược kinh doanh.

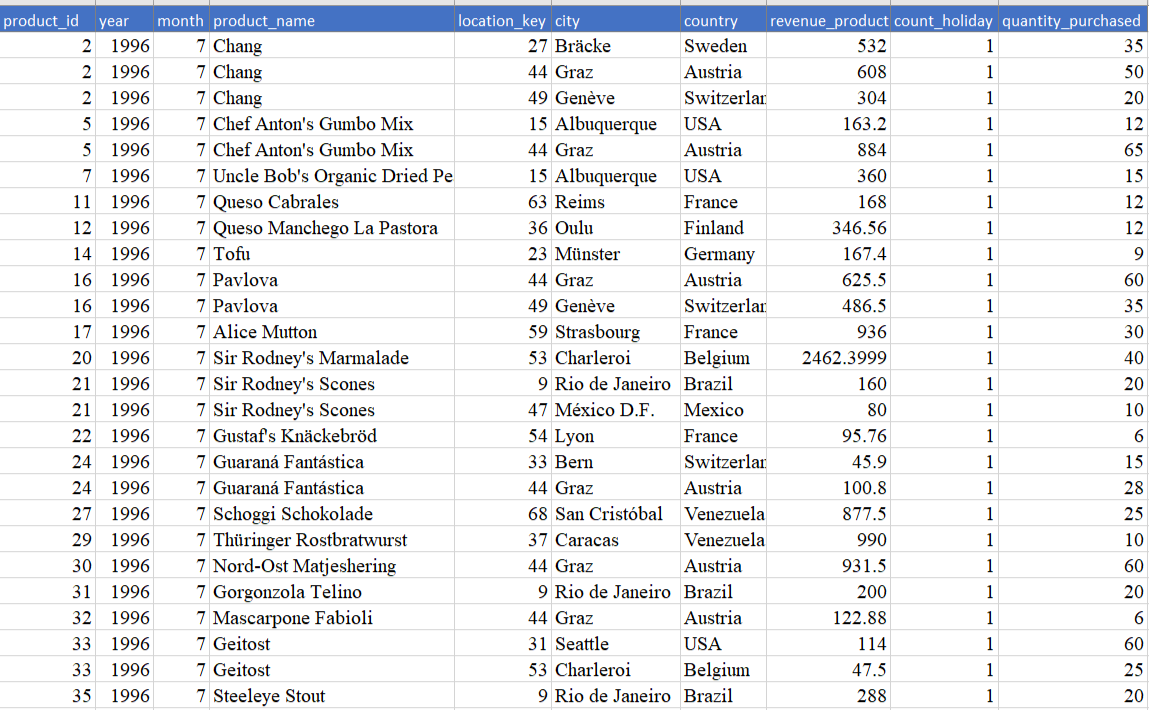
Dự báo sản lượng tiêu thụ theo vùng là chìa khóa để doanh nghiệp hiểu rõ hơn về sự biến động và đặc thù của từng thị trường địa phương. Điều này giúp họ tinh chỉnh sản phẩm và dịch vụ để phản ánh chính xác nhu cầu cụ thể của từng vùng, từ đó tối ưu hóa trải nghiệm của khách hàng và tăng cường sự hài lòng.

  Hơn nữa, dự báo số lượng sản phẩm cũng là bước quan trọng để doanh nghiệp có thể quản lý hiệu quả nguồn ứng. Điều này giúp họ tránh tình trạng thiếu hụt hoặc thiệt hại do dư thừa, từ đó tối ưu hóa chi phí và tăng cường sự linh hoạt trong quá trình quản lý chuỗi cung ứng. Không chỉ vậy, điều này cung cấp cái nhìn toàn diện về sự phân phối địa lý của nhu cầu thị trường, giúp doanh nghiệp tập trung nguồn lực tiếp thị và quảng cáo vào những khu vực có tiềm năng phát triển cao nhất. Điều này không chỉ giúp tối ưu hóa chi phí tiếp thị mà còn tăng cường hiệu suất chiến lược tiếp thị và quảng bá.

## 14.4. Áp dụng thuật toán Linear Regression để dự báo xu hướng tiêu thụ sản phẩm theo vùng trong thời điểm tới

14.4.1. Thu thập dữ liệu

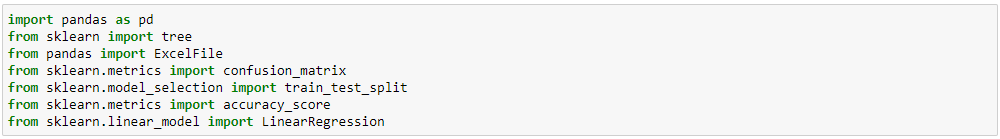
Dữ liệu được nhóm thu thập bên trong data mart bằng cách dùng lệnh truy vấn SQL dữ liệu cần thiết và lưu dưới dạng file Excel như sau:



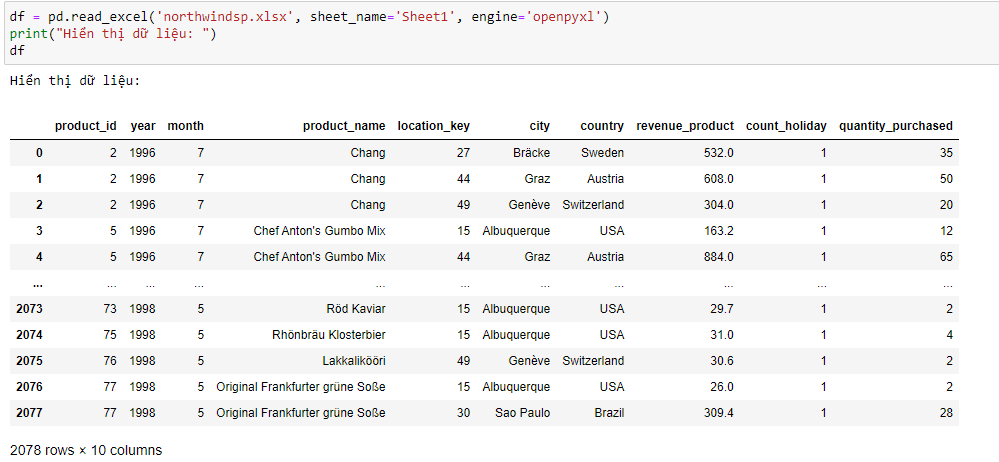
*Hình* 14.1: Dữ liệu nhóm thu thập được

14.4.2. Khái phá dữ liệu

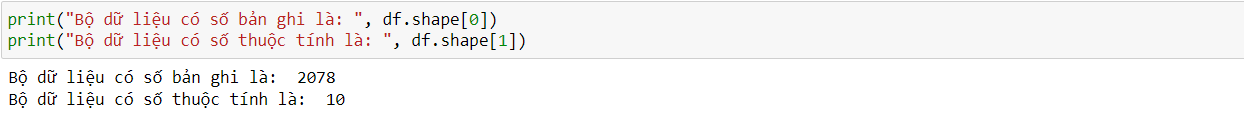
* Khai báo thư viện cần thiết



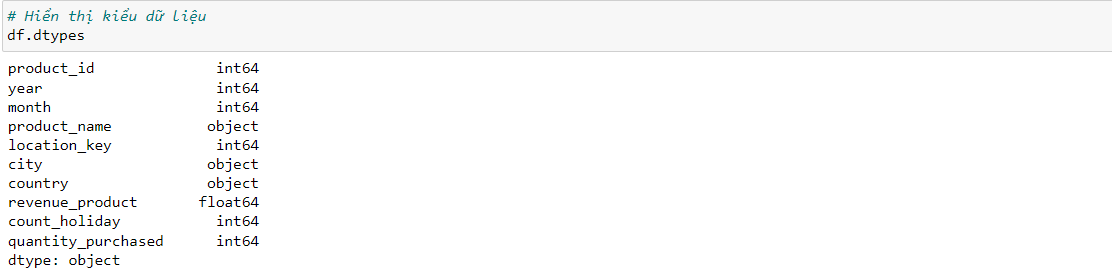
* Đọc dữ liệu:



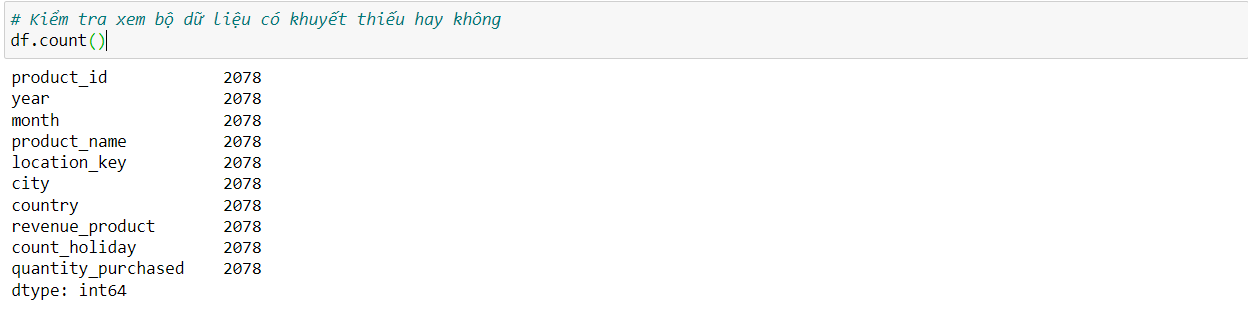
* Hiển thị số lượng mẫu và số thuộc tính



* Hiển thị kiểu dữ liệu:

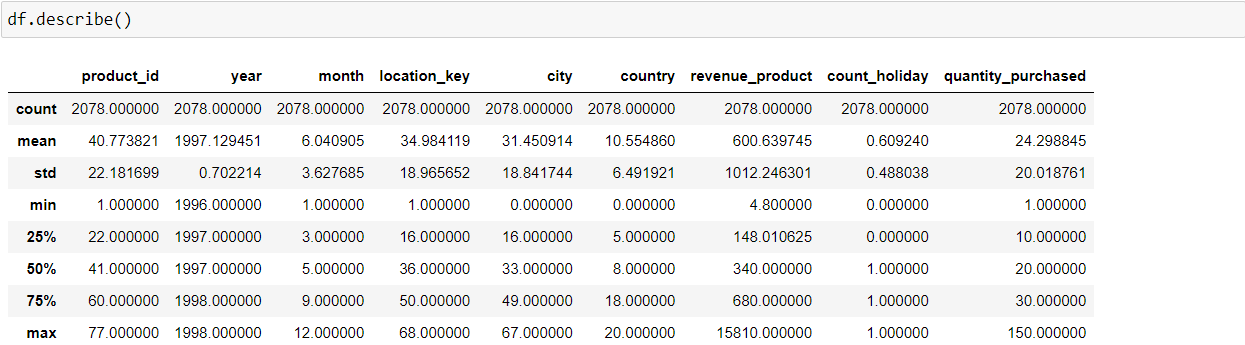


* Kiểm tra khuyết thiếu



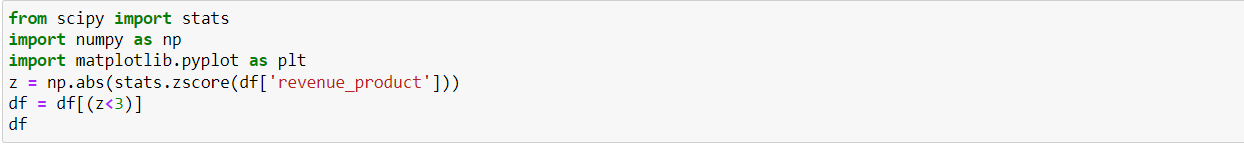
=> Vậy bộ dữ liệu không khuyết thiếu

* Thống kê các giá trị định lượng

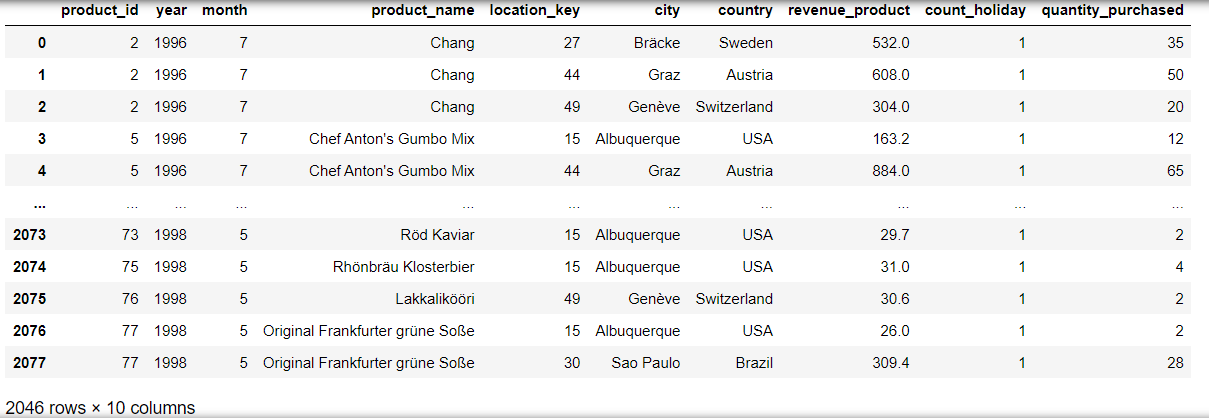


* Tiền xử lý dữ liệu

Khai báo thư viện và xử lý giá trị ngoại lai cho revenune\_product

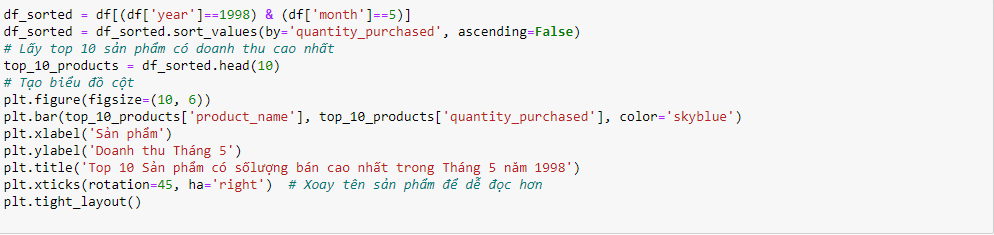


* Kết quả thu được:

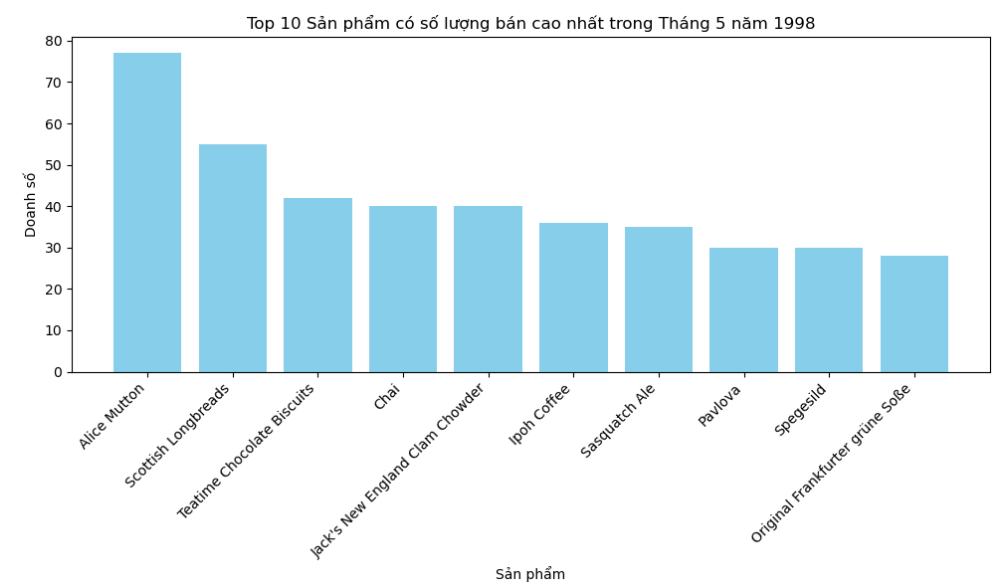


Nhận xét: Như vậy đã có 46 dòng có giá trị ngoại lai về revenue\_product được loại bỏ

* Vẽ biểu đồ thống kê

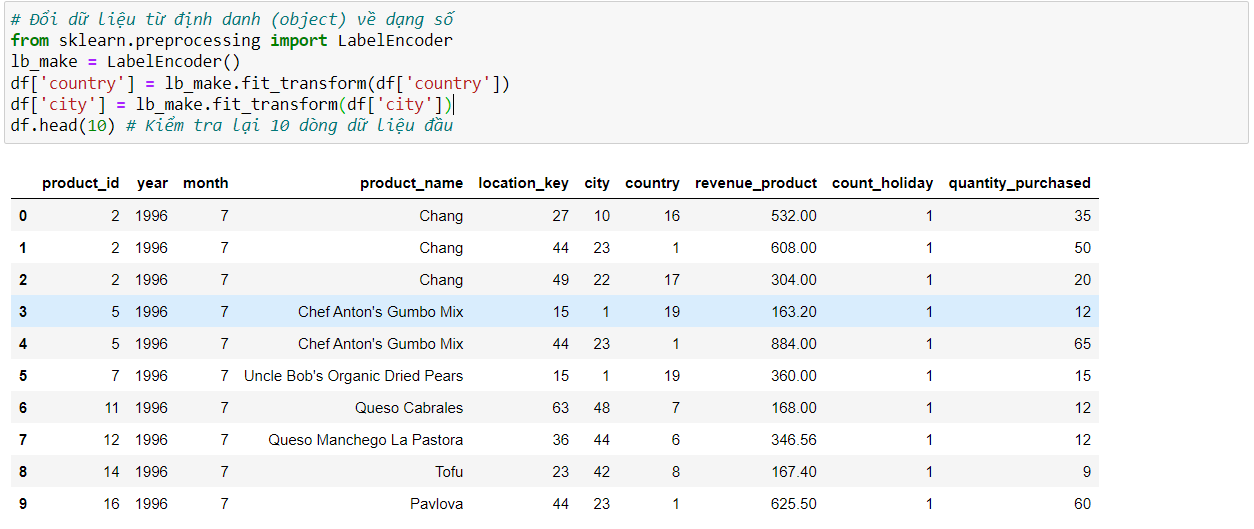


Kết quả thu được:



Hình 14.2: Top 10 sản phẩm có số lượng bán cao nhất trong tháng 5 năm 1998

Đổi dữ liệu từ dạng định danh (object) về dạng số

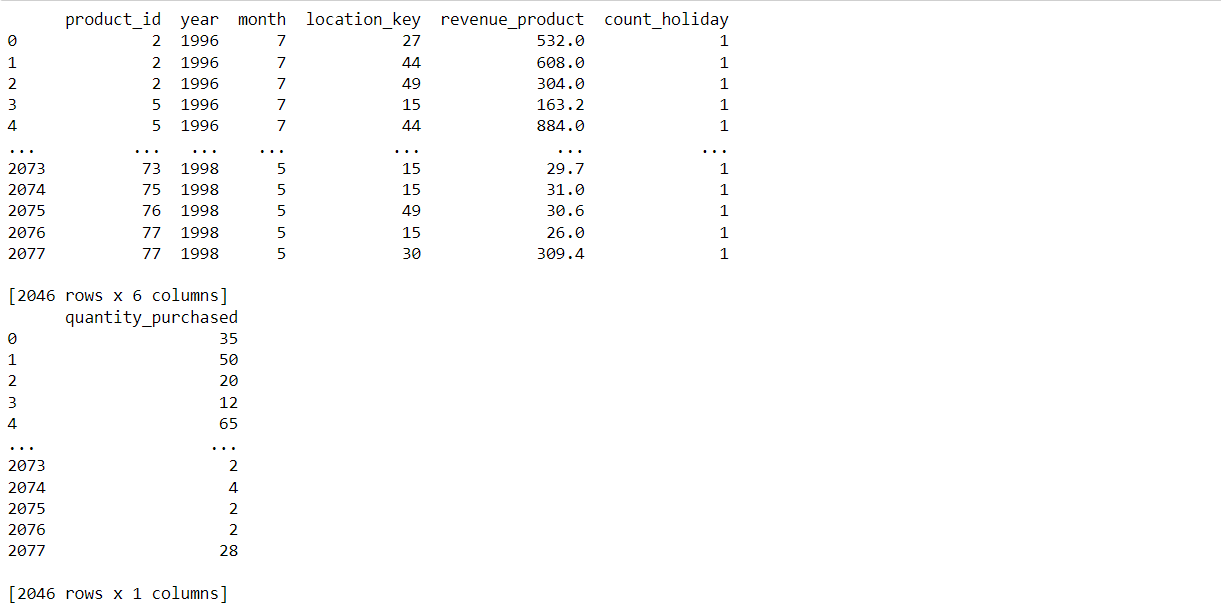


Xác định thuộc tính mô tả và dự đoán

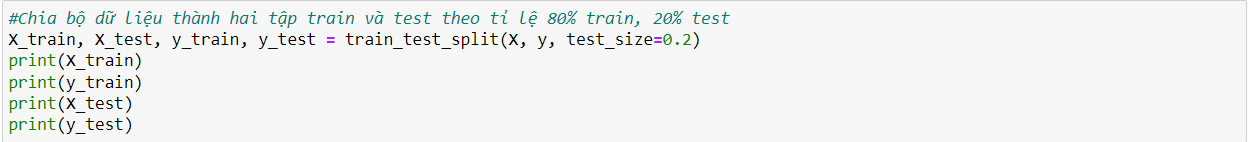
14.4.3. Xây dựng mô hình



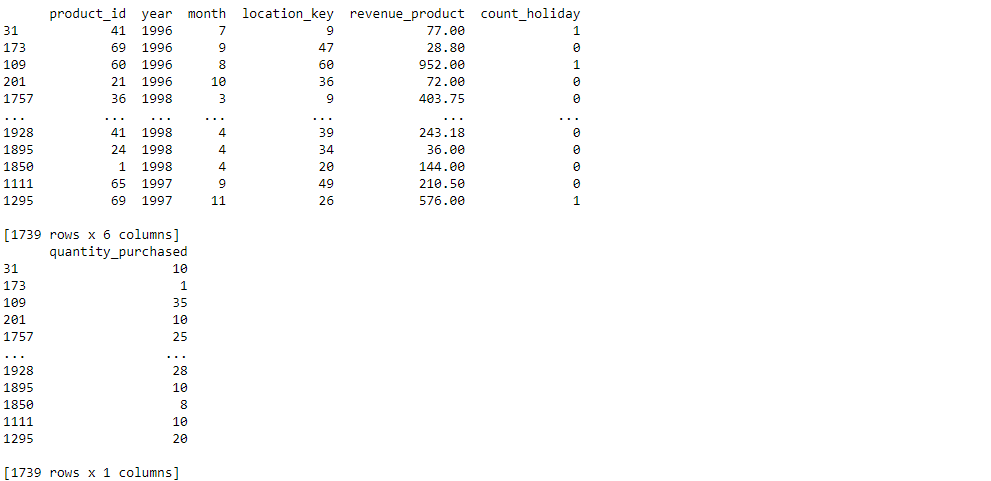
* Kết quả thu được

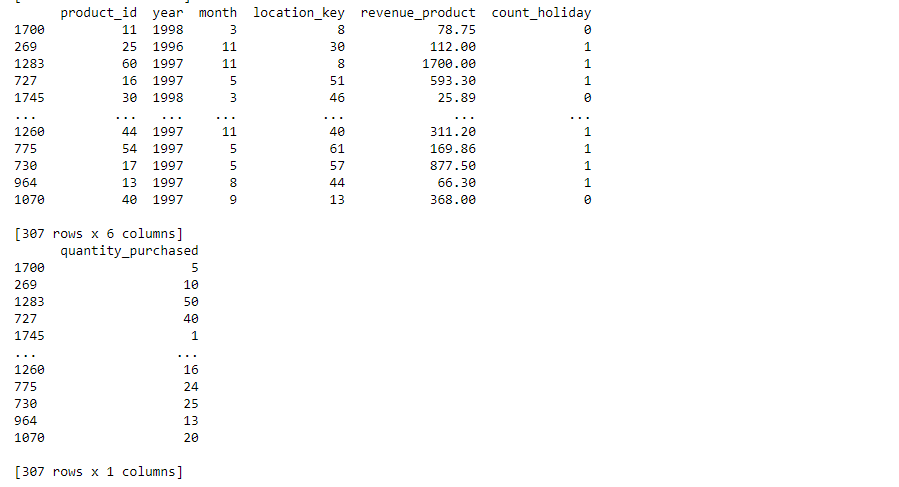


Chia bộ dữ liệu thành hai tập train và test với tỉ lệ Train 80%, Test 20%



* Tập dữ liệu Train:



* Tập dữ liệu Test:
* 

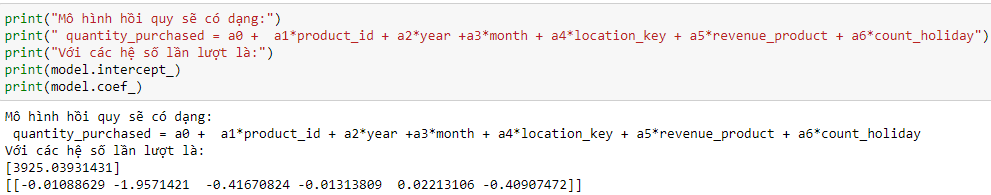
Khai báo mô hình hồi quy tuyến tính:

****

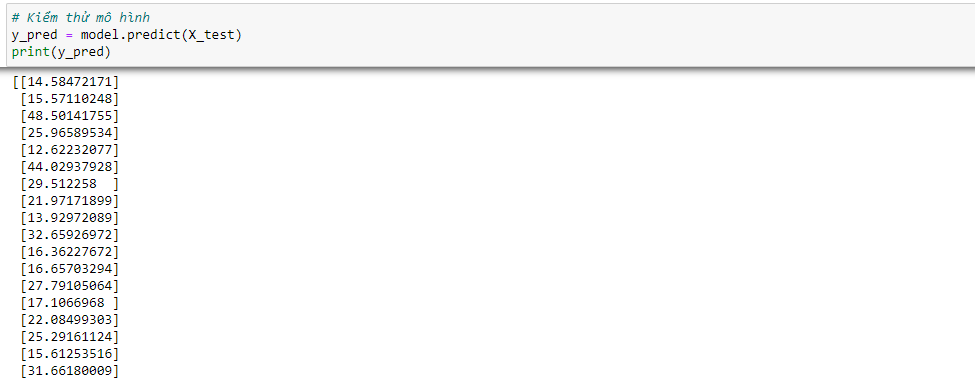
Thực thi mô hình:



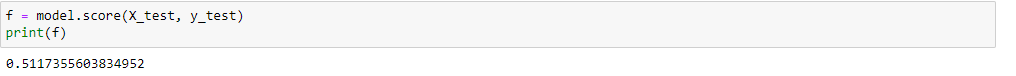
Viết phương trình hồi quy tuyến tính:



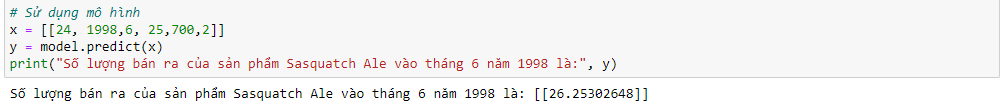
Kiểm thử mô hình:



Độ chính xác của mô hình:



* Độ chính xác của mô hình xấp xỉ 0.51174. Như vậy ta thấy độ chính xác chưa được cao.
* Sử dụng mô hình



Dự đoán số lượng sản phẩm Sasquatch bán ra tại thành phố Aachen của nước Germany vào tháng 6 năm 1998 là: 26 sản phẩm

# Câu 15: Nhóm hãy phát huy trí tưởng tượng và thiết kế kịch bản để case study mà nhóm đang nghiên cứu hướng đến việc triển khai data lake: nêu lý do/thách thức (làm thế nào gây xúc động cho người nghe là được) và trình bày giải pháp (sử dụng dịch vụ nào của nhà cung cấp nào) có minh họa kiến trúc bằng hình ảnh, lợi ích (1 điểm)

## 15.1. Lý do đề xuất triển khai Data lake?

Hiện nay, xu hướng xây dựng Data Warehouse nổi lên như một hiện tượng toàn cầu, mang lại những lợi ích không tưởng đến cho các doanh nghiệp trên khắp thế giới với đủ mọi lĩnh vực ngành nghề. Nắm bắt được những lợi ích đó không ít những ông lớn trong làng kinh doanh sẵn sàng đầu tư thời gian, công sức, tiền bạc để triển khai Data Warehouse và nhận được những kết quả vượt mong đợi. Với đặc tính là một kho lưu trữ dữ liệu lớn, Data Warehouse giúp:

* Tăng cường khả năng truy cập và phân tích dữ liệu: giúp các doanh nghiệp dễ dàng truy cập và phân tích dữ liệu từ nhiều nguồn khác nhau. Điều này giúp các doanh nghiệp hiểu rõ hơn về khách hàng, đối thủ cạnh tranh, và thị trường.
* Tăng cường khả năng ra quyết định: cung cấp cho các doanh nghiệp thông tin chi tiết để đưa ra các quyết định kinh doanh sáng suốt.
* Tăng cường khả năng đổi mới: giúp các doanh nghiệp phát triển các sản phẩm và dịch vụ mới dựa trên dữ liệu.

Tuy nhiên về lâu dài Data Warehouse cũng có một số hạn chế:

* Chi phí: Xây dựng và duy trì Data Warehouse có thể tốn kém, đặc biệt là đối với các doanh nghiệp nhỏ và vừa.
* Thời gian: Xây dựng và triển khai Data Warehouse có thể mất nhiều thời gian, đặc biệt là đối với các doanh nghiệp có lượng dữ liệu lớn.
* Kỹ năng: Yêu cầu kỹ năng cao để xây dựng và quản lý Data Warehouse.
* Hạn chế về dữ liệu: Data Warehouse thường chỉ có thể lưu trữ dữ liệu có cấu trúc. Điều này có thể gây khó khăn cho việc lưu trữ và phân tích dữ liệu phi cấu trúc, chẳng hạn như dữ liệu từ các cảm biến, dữ liệu từ các trang mạng xã hội, và dữ liệu từ các thiết bị di động.



Hình 15.1: Minh họa 1

  (Nguồn: Nhóm tự design)

Là một nhân viên phòng phân tích dữ liệu, Hạnh không xa lạ gì với những chiếc báo cáo được lấy dữ liệu trên Data Warehouse. Công việc phân tích báo cáo, tìm xu hướng diễn ra khá là mượt mà nhưng kể từ khi:



Hình 15.2: Minh họa 2

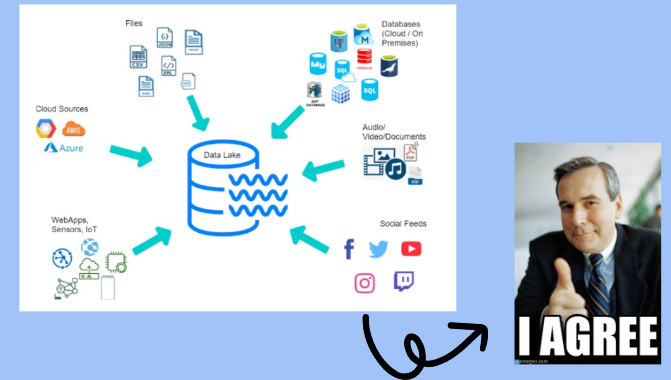
      (Nguồn: Nhóm tự design)

Là một công ty xuất nhập khẩu thực phẩm lớn trong và ngoài nước, dữ liệu phi cấu trúc trong công ty Hạnh bao gồm các loại sau:

* **Dữ liệu từ các cảm biến**: Dữ liệu từ các cảm biến được tạo ra bởi các thiết bị cảm biến, chẳng hạn như cảm biến nhiệt độ, cảm biến độ ẩm, và cảm biến lực. Dữ liệu từ các cảm biến thường ở dạng dữ liệu thô và phi cấu trúc.
* **Dữ liệu từ các trang mạng xã hội**: Dữ liệu từ các trang mạng xã hội được tạo ra bởi người dùng mạng xã hội, chẳng hạn như bài đăng, bình luận, và lượt thích. Dữ liệu từ các trang mạng xã hội thường ở dạng dữ liệu văn bản và hình ảnh.
* **Dữ liệu từ các thiết bị di động**: Dữ liệu từ các thiết bị di động được tạo ra bởi người dùng thiết bị di động, chẳng hạn như lịch sử duyệt web, vị trí, và thói quen mua sắm. Dữ liệu từ các thiết bị di động thường ở dạng dữ liệu văn bản, hình ảnh, và video.

=> Dữ liệu phi cấu trúc có thể cung cấp cho các công ty xuất nhập khẩu thực phẩm nhiều thông tin chi tiết hữu ích, chẳng hạn như:

* **Tình trạng chất lượng của thực phẩm**: Dữ liệu từ các cảm biến có thể được sử dụng để theo dõi nhiệt độ, độ ẩm, và mức độ tồn kho của thực phẩm. Điều này có thể giúp các công ty đảm bảo rằng thực phẩm được lưu trữ trong điều kiện thích hợp và không bị hư hỏng.
* **Nhu cầu của khách hàng**: Dữ liệu từ các trang mạng xã hội có thể được sử dụng để hiểu nhu cầu của khách hàng về các sản phẩm và dịch vụ thực phẩm. Điều này có thể giúp các công ty phát triển các sản phẩm và dịch vụ đáp ứng nhu cầu của khách hàng.
* **Tình hình thị trường**: Dữ liệu từ các trang mạng xã hội và các thiết bị di động có thể được sử dụng để theo dõi tình hình thị trường thực phẩm. Điều này có thể giúp các công ty đưa ra các quyết định kinh doanh sáng suốt.



Hình 15.3: Minh họa 3

      (Nguồn: Nhóm tự design)

Trở lại với câu chuyện của công ty Hạnh. Sau những lần đề xuất, tham vấn của bộ phận phân tích dữ liệu thị trường, công ty đã quyết định thuê một đơn vị hỗ trợ để xây dựng data lake cho công ty. Với mong muốn tối ưu hóa nguồn dữ liệu hiện có trong doanh nghiệp để nâng cao hiệu quả phân tích và dự báo thị trường.

Sau khi triển khai data lake, Hạnh đã có thể dễ dàng truy cập và phân tích dữ liệu từ các nguồn khác nhau. Điều này đã giúp Hạnh tìm ra nhiều xu hướng mới về nhu cầu của khách hàng một cách nhanh chóng và hiệu quả hơn.

Một năm sau khi triển khai data lake, công ty đã tăng doanh số bán hàng lên 15%. Hạnh rất vui mừng vì công việc của mình đã giúp công ty phát triển. Hạnh cũng cảm thấy rất tự hào vì đã góp phần giúp công ty mang lại lợi ích cho công ty.

Như vậy vấn đề của công ty của Hạnh đã được giải quyết. Để lại đó một câu hỏi dành cho các nhà phân tích dữ liệu trong các công ty hiện nay đó là:



Hình 15.4: Minh họa 4

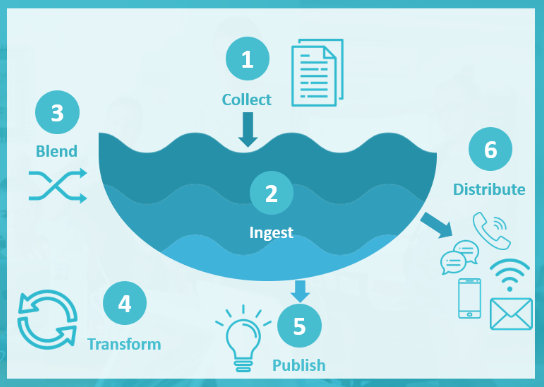
  (Nguồn: Nhóm tự design)

         Tùy thuộc vào các yêu cầu, một doanh nghiệp điển hình sẽ yêu cầu cả kho dữ liệu và hồ dữ liệu vì chúng phục vụ các nhu cầu và trường hợp sử dụng khác nhau.

Trường hợp trên của Công ty Hạnh khá giống với Northwind hiện nay. Mặc dù đã triển khai xong Data warehouse nhưng nhóm vẫn quyết định đi theo xu hướng xây dựng thêm các Data lake, với mong muốn nâng cao khả năng lưu trữ linh hoạt và tận dụng Big data mang lại lợi ích cho công ty.

=> Vậy Data lake là gì? Có những lợi ích và khó khăn gì? Nó có thực sự đáng để thử hay không?

15.2. Data lake là gì?



Hình 15.5: Kiến trúc Data Lake cơ bản

(Nguồn: https://indaacademy.vn)

**Data Lake** - còn được gọi là Hồ dữ liệu, là một nơi lưu trữ tập trung được thiết kế để chứa, xử lý và bảo mật lượng lớn dữ liệu đa dạng, bao gồm dữ liệu có cấu trúc, bán cấu trúc và phi cấu trúc. Nơi này có khả năng lưu trữ dữ liệu ở dạng nguyên thể và xử lý mọi loại dữ liệu khác nhau mà không bị ràng buộc về kích thước. Hồ dữ liệu cung cấp khả năng lưu trữ lượng lớn dữ liệu, nâng cao hiệu suất phân tích và tích hợp dữ liệu gốc, phân tích thời gian thực và học máy để đưa ra các quyết định tốt hơn.

Data Lake giúp doanh nghiệp khai thác dòng dữ liệu mới từ nhiều nguồn hơn trong thời gian ngắn và trao quyền cho người dùng cộng tác để đưa ra quyết định chính xác hơn. Không chỉ vậy, nó cũng có thể làm cho các hệ thống dữ liệu hiện có trở nên hiệu quả bằng cách giảm tải công suất cho cơ sở hạ tầng linh hoạt hơn. Một cuộc khảo sát của Aberdeen cho thấy các tổ chức đã triển khai Data Lake vượt trội hơn các công ty tương tự 9% về mức tăng trưởng doanh thu không phải trả tiền. Những nhà lãnh đạo này đã có thể thực hiện các loại phân tích mới như học máy qua các nguồn mới như tệp nhật ký, dữ liệu từ các luồng nhấp chuột, mạng xã hội và các thiết bị được kết nối internet được lưu trữ trong hồ dữ liệu. Từ đó, giúp các doanh nghiệp thu hút và giữ chân khách hàng, tăng năng suất hoạt động để phát triển kinh doanh.

## 15.3. Giá trị của Data Lake

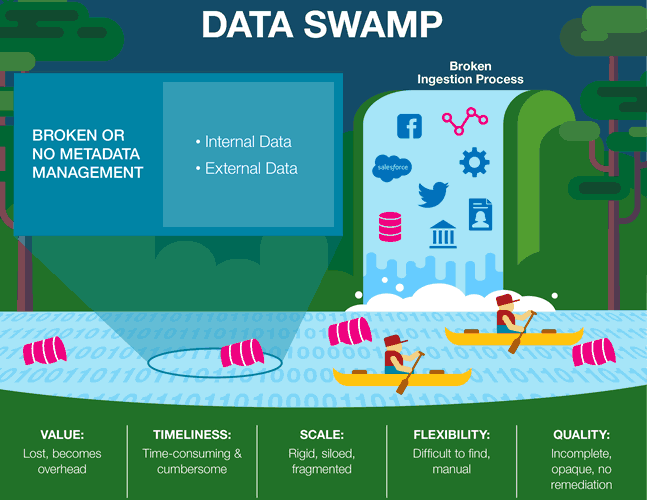
Data Lake có khả năng khai thác nhiều dữ liệu hơn, từ nhiều nguồn hơn, trong thời gian ngắn hơn và trao quyền cho người dùng cộng tác và phân tích dữ liệu theo những cách khác nhau dẫn đến việc đưa ra quyết định tốt hơn, nhanh hơn. Các ví dụ trong đó các Hồ dữ liệu có giá trị gia tăng bao gồm:

* ***Cải thiện tương tác của khách hàng***: Data Lake có thể kết hợp dữ liệu khách hàng từ nền tảng CRM với phân tích phương tiện truyền thông xã hội, một nền tảng tiếp thị bao gồm lịch sử mua hàng và phiếu sự cố để trao quyền cho doanh nghiệp hiểu được nhóm khách hàng sinh lời cao nhất, nguyên nhân khiến khách hàng bỏ cuộc và các chương trình khuyến mại hoặc phần thưởng điều đó sẽ làm tăng lòng trung thành của khách hàng với doanh nghiệp.
* ***Cải thiện các lựa chọn đổi mới R&D (Nghiên cứu và phát triển)***: Hồ sơ dữ liệu có thể giúp nhóm R&D của bạn kiểm tra giả thuyết của họ, tinh chỉnh các giả định và đánh giá kết quả — chẳng hạn như chọn đúng vật liệu trong thiết kế sản phẩm của bạn, dẫn đến hiệu suất nhanh hơn, thực hiện nghiên cứu bộ gen dẫn đến thuốc hiệu quả hơn hoặc hiểu được mức độ sẵn sàng của khách hàng trả tiền cho các thuộc tính khác nhau.
* ***Tăng hiệu quả hoạt động***: Internet of Things (IoT) giới thiệu nhiều cách hơn để thu thập dữ liệu về các quy trình như sản xuất, với dữ liệu thời gian thực đến từ các thiết bị được kết nối internet. Hồ dữ liệu giúp dễ dàng lưu trữ và chạy phân tích trên dữ liệu IoT do máy tạo ra để khám phá các cách giảm chi phí hoạt động và tăng chất lượng.

## 15.4. Những thách thức với Data Lake

Bên cạnh những lợi ích, tạo Data Lake cũng có thể mang đến cả những rủi ro cho các doanh nghiệp:

***Vấn đề kiểm soát truy cập:*** Thách thức chính với cấu trúc hồ dữ liệu là dữ liệu thô được lưu trữ mà không có sự giám sát của nội dung. Để một hồ dữ liệu có thể sử dụng được dữ liệu, nó cần phải có các cơ chế xác định để lập danh mục và bảo mật dữ liệu. Nếu không có các yếu tố này, dữ liệu không thể được tìm thấy hoặc đáng tin cậy dẫn đến "data swamp" (Đầm lấy dữ liệu). Việc đáp ứng nhu cầu của nhiều đối tượng hơn đòi hỏi các hồ dữ liệu phải có khả năng quản trị, tính nhất quán về ngữ nghĩa và các kiểm soát truy cập.



Hình 15.6: Data Swamp (Đầm lấy dữ liệu)

       (Nguồn:smartindustry.vn)

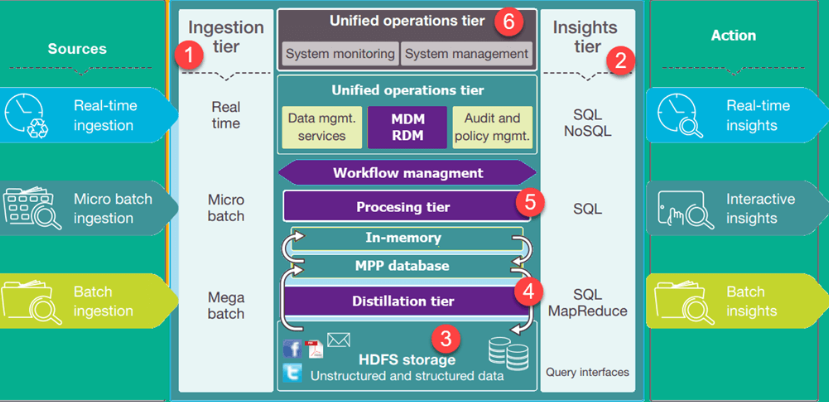
***Ngăn cản việc trích xuất dữ liệu giá trị:***Ưu điểm của Data Lake là cho phép lưu trữ dữ liệu dưới bất kỳ định dạng nào mà không giới hạn số lượng, tuy nhiên, chính điều này làm ngăn cản việc trích xuất giá trị từ dữ liệu bởi Data Lake không có khả năng phân loại hoặc thiết lập hệ thống phân cấp giữa dữ liệu, không cho phép ưu tiên dữ liệu và cách sử dụng dữ liệu.

***Độ trễ dữ liệu:***Nếu các Data Lake cách xa nhau về mặt vật lý, có thể mất nhiều thời gian để truy xuất một phần dữ liệu cụ thể. Tuy nhiên, dữ liệu được sử dụng thường xuyên sẽ bỏ qua hiện tượng này. Dữ liệu chỉ hữu ích nếu nó có thể được sử dụng để đưa ra những quyết định đúng đắn vào đúng thời điểm. Nếu một doanh nghiệp muốn phân tích dữ liệu và phải dành nhiều thời gian để tìm kiếm và chuẩn bị dữ liệu trong Data Lake, thì hiệu quả sẽ giảm đi rất nhiều.

***Khó kiểm soát quyền riêng tư dữ liệu:*** Data Lake có thể làm rò rỉ dữ liệu khách hàng. Việc thiếu các ràng buộc của Data Lake có thể khiến các doanh nghiệp đặt dữ liệu rủi ro ở một nơi không an toàn. Một thực tế đã xảy ra đó là gần 200 triệu cử tri Mỹ đã bị lộ dữ liệu về thông tin cá nhân và thậm chí lộ cả dữ liệu dự đoán về hành vi bầu cử của công dân sau khi Data Lake được cung cấp trên đám mây công cộng.

## 15.5. Giải pháp triển khai Data Lake

### 15.5.1. Kiến trúc Data lake



Hình 15.7: Kiến trúc Data lake

      (Nguồn: <https://www.guru99.com>)

Hình trên thể hiện kiến ​​trúc của Hồ dữ liệu doanh nghiệp. Các cấp thấp hơn biểu thị dữ liệu chủ yếu ở trạng thái nghỉ trong khi các cấp trên hiển thị dữ liệu giao dịch theo thời gian thực. Dữ liệu này chảy qua hệ thống mà không có hoặc có độ trễ rất nhỏ. Sau đây là các tầng quan trọng trong Kiến trúc hồ dữ liệu:

* Cấp nhập (**Ingestion Tier**): Các cấp ở bên trái mô tả nguồn dữ liệu. Dữ liệu có thể được tải vào hồ dữ liệu theo đợt hoặc theo thời gian thực.
* Cấp thông tin chi tiết (**Insights Tier**): Các cấp bên phải đại diện cho phía nghiên cứu nơi sử dụng thông tin chi tiết từ hệ thống. SQL, truy vấn NoSQL hoặc thậm chí excel có thể được sử dụng để phân tích dữ liệu.
* **HDFS** là giải pháp tiết kiệm chi phí cho cả dữ liệu có cấu trúc và không cấu trúc. Đây là nơi hạ cánh cho tất cả dữ liệu đang ở trạng thái nghỉ trong hệ thống.
* Tầng chưng cất (**Distillation tier**): lấy dữ liệu từ bộ lưu trữ và chuyển đổi dữ liệu đó thành dữ liệu có cấu trúc để phân tích dễ dàng hơn.
* Cấp xử lý (**Processing tier**): chạy các thuật toán phân tích và truy vấn của người dùng với thời gian thực, tính tương tác, hàng loạt khác nhau để tạo dữ liệu có cấu trúc nhằm phân tích dễ dàng hơn.
* Cấp hoạt động (**Unified operations tier**): hợp nhất chi phối việc quản lý và giám sát hệ thống. Nó bao gồm kiểm tra và quản lý trình độ, quản lý dữ liệu, quản lý quy trình làm việc.

### 15.5.2. Triển khai các Hồ dữ liệu trên đám mây

Data Lakes là một khối lượng công việc lý tưởng để triển khai trên đám mây, vì đám mây cung cấp hiệu suất, khả năng mở rộng, độ tin cậy, tính khả dụng, một bộ công cụ phân tích đa dạng và quy mô kinh tế lớn. Nghiên cứu của ESG cho thấy 39% người được hỏi coi đám mây là hoạt động triển khai chính của họ để phân tích, 41% đối với kho dữ liệu và 43% đối với Spark. Các lý do hàng đầu khiến khách hàng coi đám mây là một lợi thế của Data Lakes là bảo mật tốt hơn, thời gian triển khai nhanh hơn, tính khả dụng tốt hơn, cập nhật tính năng/ chức năng thường xuyên hơn, tính đàn hồi hơn, phạm vi địa lý cao hơn và chi phí liên quan đến việc sử dụng thực tế.

Một vài đơn vị cung cấp xây dựng Data lake trên Cloud hiện nay:

**Amazon Web Services (AWS)**: AWS có lợi thế về khả năng mở rộng, tính ổn định và độ tin cậy. AWS cung cấp một loạt các dịch vụ lưu trữ dữ liệu, bao gồm Amazon S3, Amazon Redshift, và Amazon EMR. Amazon S3 là một dịch vụ lưu trữ dữ liệu đám mây phổ biến cho các loại dữ liệu khác nhau, bao gồm dữ liệu có cấu trúc, dữ liệu bán cấu trúc, và dữ liệu phi cấu trúc cho phép các doanh nghiệp chạy các công cụ phân tích dữ liệu như Hadoop và Spark.

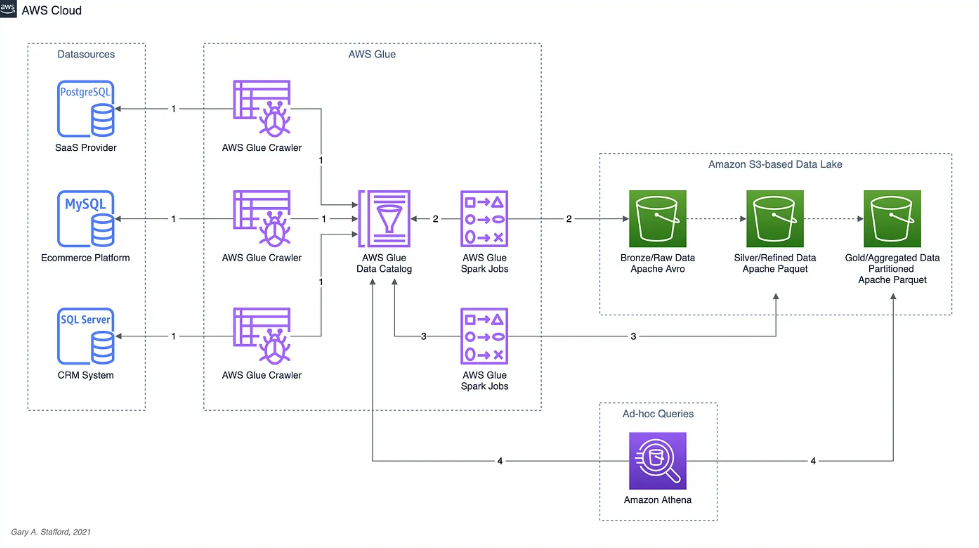
**Google Cloud Platform (GCP):** có lợi thế về chi phí cạnh tranh, tính thân thiện với môi trường và khả năng tích hợp. GCP cung cấp các dịch vụ lưu trữ dữ liệu với chi phí cạnh tranh, bao gồm Cloud Storage, BigQuery và Dataproc. Với ưu thế về các ứng dụng về AI, machine learning.

**Alibaba Cloud**: có lợi thế về khả năng mở rộng, tính linh hoạt và khả năng hỗ trợ ngôn ngữ địa phương. Alibaba Cloud cung cấp các dịch vụ lưu trữ dữ liệu với khả năng mở rộng cao, bao gồm OSS, ODPS và EMR.

### 15.5.3. Xây dựng Data Lakes của Northwind trên đám mây với AWS

AWS cung cấp danh mục dịch vụ an toàn nhất, có thể mở rộng, toàn diện và hiệu quả về chi phí cho phép khách hàng xây dựng hồ dữ liệu của họ trên đám mây, phân tích tất cả dữ liệu của họ, bao gồm dữ liệu từ các thiết bị IoT với nhiều phương pháp phân tích khác nhau bao gồm cả học máy. Do đó, có nhiều tổ chức chạy các hồ dữ liệu và phân tích của họ trên AWS hơn bất kỳ nơi nào khác với các khách hàng như NETFLIX, Zillow, NASDAQ, Yelp, iRobot và FINRA… tin tưởng AWS điều hành khối lượng công việc phân tích quan trọng của doanh nghiệp họ.

Minh họa Kiến trúc xây dựng Data Lake trên AWS Cloud như sau:



Hình 15.8: Kiến trúc và quy trình xây dựng Data Lake trên AWS Cloud

(Nguồn: Gray A.Stafford, 2021)

Chúng ta sẽ xây dựng Data Lake trên AWS bằng cách sử dụng kết hợp các dịch vụ, bao gồm AWS Glue Data Catalog, AWS Glue Crawlers, AWS Glue Jobs, AWS Glue Studio, Amazon Athena, Amazon Relational Database Service (Amazon RDS) và Amazon S3.

AWS Glue Data Catalog:

* Là một thư mục trung tâm lưu trữ thông tin về dữ liệu của bạn trên AWS.
* Nó giúp bạn dễ dàng tìm kiếm, hiểu và quản lý dữ liệu của mình từ nhiều nguồn khác nhau

AWS Glue Crawlers:

* Là các công cụ tự động khám phá và phân tích dữ liệu trong kho lưu trữ của bạn, chẳng hạn như S3.
* Chúng trích xuất thông tin về cấu trúc dữ liệu, chẳng hạn như các trường, loại dữ liệu và định dạng.

AWS Glue Jobs:

* Là các công cụ giúp bạn thực hiện các tác vụ ETL (Extract, Transform, Load) trên dữ liệu của mình.
* Chúng trích xuất dữ liệu từ các nguồn khác nhau, biến đổi nó thành dạng phù hợp, sau đó tải nó vào kho đích, chẳng hạn như Redshift hoặc RDS.

AWS Glue Studio:

* Là một giao diện trực quan cho phép bạn xây dựng và chạy các công việc ETL mà không cần viết mã.
* Nó cung cấp các thành phần kéo và thả để thực hiện các tác vụ ETL, giúp việc xử lý dữ liệu dễ dàng hơn.
* Nó giống như một bộ đồ bếp tiện ích với các thiết bị sẵn sàng sử dụng, giúp bạn nấu ăn mà không cần phải học tất cả các kỹ thuật từ đầu.

Amazon Athena:

* Là một dịch vụ phân tích dữ liệu tương tác cho phép bạn truy vấn trực tiếp dữ liệu của mình trong S3.
* Bạn không cần phải thiết lập hoặc quản lý kho dữ liệu; bạn chỉ cần truy vấn dữ liệu trực tiếp từ S3.

Amazon Relational Database Service (Amazon RDS):

* Là một dịch vụ cơ sở dữ liệu quan hệ cho phép bạn chạy các ứng dụng cơ sở dữ liệu trên AWS.
* Bạn có thể chọn từ nhiều loại công cụ cơ sở dữ liệu phổ biến, chẳng hạn như MySQL, PostgreSQL, Oracle Database và Microsoft SQL Server.

Amazon S3:

* Là một dịch vụ lưu trữ dữ liệu trên nền tảng đám mây cho phép bạn lưu trữ và truy cập dữ liệu của mình từ bất cứ đâu.
* Nó là một kho lưu trữ lớn và linh hoạt, có thể chứa mọi loại dữ liệu, từ tài liệu văn bản đến video và mã ứng dụng.

Chúng ta sẽ lập danh mục và di chuyển dữ liệu từ ba nguồn dữ liệu riêng biệt vào hồ dữ liệu dựa trên Amazon S3. Khi ở trong hồ dữ liệu, thực hiện ETL trên dữ liệu thô - làm sạch, tăng cường và chuẩn bị nó để phân tích dữ liệu. Cuối cùng, thực hiện tổng hợp trên dữ liệu được tinh chỉnh và ghi các bộ dữ liệu cuối cùng đó trở lại hồ dữ liệu. Hồ dữ liệu sẽ được tổ chức xung quanh mô hình hồ dữ liệu gồm dữ liệu đồng (hay còn gọi là thô), bạc (hay còn gọi là tinh chế) và vàng (hay còn gọi là tổng hợp).

**TỔNG KẾT**

Thành công của doanh nghiệp phụ thuộc vào khả năng ra quyết định đúng đắn của cấp quản lý. Quyết định được đưa ra dựa trên nhiều yếu tố, bao gồm cảm xúc, tri thức, thiên kiến cá nhân của người ra quyết định, cũng như các yếu tố bên ngoài như tình hình kinh tế, hoàn cảnh thị trường, chiến lược của đối thủ cạnh tranh, vận hành của công ty, và thay đổi của công nghệ. Do đó, người ra quyết định cần có đầy đủ thông tin cần thiết để đưa ra quyết định chính xác. Công cụ công nghệ thông tin như hệ hỗ trợ ra quyết định đã được phát triển để hỗ trợ người ra quyết định trong việc thu thập và xử lý thông tin, từ đó đưa ra quyết định tốt hơn.

Bản báo cáo này có khả năng thu thập dữ liệu liên quan từ nhiều nguồn khác nhau, phân tích và tổng hợp chúng thành các biểu đồ, đồ thị trực quan. Điều này giúp người quản lý dễ dàng nắm bắt thông tin, phân tích tình hình và đưa ra quyết định chính xác. Các quyết định của người quản lý có ảnh hưởng lớn đến sự thành công của doanh nghiệp. Do đó, việc thu thập và phân tích thông tin liên quan là rất quan trọng.Trong quá khứ, dữ liệu thường được thu thập bằng cách quan sát và ghi chép thủ công. Tuy nhiên, cách thức này có thể dẫn đến sai sót, do sự chủ quan của con người hoặc do các yếu tố khách quan như lỗi hệ thống. Dữ liệu không chính xác hoặc không liên quan có thể dẫn đến ra quyết định sai lầm, gây ra hậu quả nghiêm trọng cho doanh nghiệp. Để khắc phục hạn chế này, phần lớn các công ty hiện nay đang chuyển sang sử dụng công nghệ thông tin để thu thập và phân tích dữ liệu. Công nghệ thông tin có thể tự động hóa các quy trình thu thập dữ liệu, giúp giảm thiểu sai sót và đảm bảo tính chính xác của dữ liệu. Ngoài ra, công nghệ thông tin cũng có thể giúp phân tích dữ liệu một cách nhanh chóng và hiệu quả, giúp doanh nghiệp đưa ra quyết định chính xác hơn.

Dữ liệu được thu thập tự động từ nhiều nguồn khác nhau, bao gồm dữ liệu vận hành, thị trường, tài chính, xu hướng, và mô hình doanh nghiệp. Dữ liệu sau đó được phân tích bởi phần mềm BI để phát hiện các mối liên hệ và xu hướng có liên quan đến vấn đề cần giải quyết. Các kết quả phân tích được trình bày trong các báo cáo trực quan, dễ hiểu, giúp người quản lý đưa ra quyết định chính xác.

# TÀI LIỆU THAM KHẢO

Analytics, M. D. (2023). Được truy lục từ Data warehousing process diagram: https://mastering-da.com/

IBM. (2023). *Sample reports for Operational Reporting*. Được truy lục từ IBM Documentation: https://www.ibm.com/docs/en/order-management?topic=reports-sample-operational-reporting

INDA, H. v. (2023). Được truy lục từ ETL Architecture: https://indaacademy.vn/

KimBall, R. (2013). The Data Warehouse Tollkit: The definitive guide to dimensional modeling (2nd ed.). Wiley.

KimBall, R., & Ross, M. (2023). The data warehouse toolkit: The definitive guide to dimensional modeling (3rd ed.). Wiley.

Osam. (2021). *Data lake là gì? Khai thác hồ dữ liệu trên đám mây như thế nào?* Được truy lục từ Osam: https://www.osam.io/post/data-lake-ho-du-lieu-la-gi

Rainardi, V. (2014). Building a data warehouse: With examples in SQL Server. Apress.

Savjani, P. (2013). *Back to basics: Why do you need OLAP cubes/ Data-Warehouses for enterprise business reporting systems?* Được truy lục từ Insight Extractor - Blog: https://insightextractor.com/2013/06/25/back-to-basics-why-do-you-need-olap-cubes/

SinnovaSoft. (2020). *Data warehouse architecture*. Được truy lục từ sinnovasoft.com

Stafford, G. A. (2021). *Building a Data Lake on AWS*. Được truy lục từ Garystafford: Garystafford.medium.com

Taylor, D. (2023). *What is Data Lake? It’s Architecture: Data Lake Tutorial*. Được truy lục từ guru99: https://www.guru99.com/data-lake-architecture.html

Turban, E., Sharda, R., Delen, D., & King, D. (2014). Data warehouse architecture. Trong *Business Intelligence and Analytics: Systems for Decision Support* (trang 69). Pearson Education.

Vaisman, A., & Zimányi, E. (2014). Data Warehouse Systems: Design and Implementation (Data-Centric Systems and Applications). Springer.