

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM

KHOA CÔNG NGHỆ THÔNG TIN



ĐỒ ÁN CÔNG NGHỆ THÔNG TIN

ĐỀ TÀI ĐỒ ÁN

ỨNG DỤNG KHO DỮ LIỆU TRONG PHÂN TÍCH
VÀ TRỰC QUAN HÓA DỮ LIỆU BÁN HÀNG

Giảng viên hướng dẫn: TS. TRẦN NHẬT QUANG

Mã lớp học phần: PROJ215879_05CLC

Sinh viên thực hiện: TRẦN NGUYỄN TRÍ ĐẠT

Mã số sinh viên: 21110162

Thành phố Hồ Chí Minh, Tháng 12 năm 2024

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM

KHOA CÔNG NGHỆ THÔNG TIN



ĐỒ ÁN CÔNG NGHỆ THÔNG TIN

ĐỀ TÀI ĐỒ ÁN

ỨNG DỤNG KHO DỮ LIỆU TRONG PHÂN TÍCH
VÀ TRỰC QUAN HÓA DỮ LIỆU BÁN HÀNG

Giảng viên hướng dẫn: TS. TRẦN NHẬT QUANG

Mã lớp học phần: PROJ215879_05CLC

Sinh viên thực hiện: TRẦN NGUYỄN TRÍ ĐẠT

Mã số sinh viên: 21110162

Thành phố Hồ Chí Minh, Tháng 12 năm 2024

PHIẾU NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

HỌC KÌ I, NĂM HỌC 2024–2025

Mã học phần: PROJ215879

Tên đề tài: Ứng dụng kho dữ liệu trong phân tích và trực quan hóa dữ liệu bán hàng.

STT	HỌ VÀ TÊN SINH VIÊN	MÃ SỐ SINH VIÊN	TỈ LỆ % HOÀN THÀNH
1	Trần Nguyễn Trí Đạt	21110162	100%

Ghi chú:

Tỷ lệ %: Mức độ phần trăm hoàn thành của sinh viên tham gia.

Nhận xét của giảng viên:

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Tp. Hồ Chí Minh – Tháng 12 năm 2024

LỜI NÓI ĐẦU

Trong thời đại số hóa bùng nổ, dữ liệu đã và đang trở thành một nguồn tài sản chiến lược, quyết định sự thành công của mọi lĩnh vực, đặc biệt là trong ngành bán lẻ và kinh doanh. Với sự phát triển mạnh mẽ của công nghệ và nhu cầu ngày càng tăng về trải nghiệm cá nhân hóa, việc thu thập, lưu trữ và phân tích dữ liệu đã không chỉ đơn thuần là một hoạt động hỗ trợ mà còn là trung tâm của quá trình ra quyết định.

Người tiêu dùng hiện nay không còn chỉ mua sắm một cách thụ động mà còn tham gia vào chuỗi giá trị, mong muốn những trải nghiệm được thiết kế phù hợp với sở thích và hành vi cá nhân. Chính vì vậy, dữ liệu về hành vi mua sắm, xu hướng tiêu dùng, và doanh số bán hàng trở thành nguồn thông tin quý giá, giúp các doanh nghiệp hiểu rõ hơn về khách hàng của mình. Từ đó, doanh nghiệp có thể tối ưu hóa chiến lược kinh doanh, cải thiện hiệu quả tiếp thị và nâng cao chất lượng dịch vụ.

Bên cạnh đó, dữ liệu bán hàng không chỉ phục vụ mục tiêu tối ưu doanh thu mà còn đóng vai trò quan trọng trong việc quản lý chuỗi cung ứng, phân tích lợi nhuận, dự báo nhu cầu, và ra quyết định chiến lược. Việc tổ chức và phân tích dữ liệu một cách khoa học giúp các doanh nghiệp không chỉ giảm thiểu rủi ro mà còn nhanh chóng nhận diện và tận dụng các cơ hội kinh doanh mới.

Để đáp ứng những nhu cầu này, kho dữ liệu (data warehouse) nổi lên như một giải pháp không thể thiếu, cung cấp một môi trường lưu trữ dữ liệu tập trung, nhất quán và hỗ trợ các hoạt động phân tích chuyên sâu. Với khả năng phân tích và trực quan hóa dữ liệu, kho dữ liệu cho phép các doanh nghiệp có cái nhìn toàn diện và sâu sắc hơn, từ đó hỗ trợ xây dựng chiến lược hiệu quả.

Xuất phát từ những ý nghĩa quan trọng của kho dữ liệu trong ngành kinh doanh, em lựa chọn đề tài “**Ứng dụng kho dữ liệu trong phân tích và trực quan hóa dữ liệu bán hàng**”. Qua đề tài này, chúng em mong muốn thiết kế và xây dựng một hệ thống kho dữ liệu, đồng thời ứng dụng các công cụ phân tích và trực quan hóa để hỗ trợ doanh nghiệp trong việc quản lý và khai thác giá trị dữ liệu một cách hiệu quả.

MỤC LỤC

DANH MỤC TỪ VIẾT TẮT	5
DANH MỤC CÁC HÌNH	6
DANH MỤC CÁC BẢNG.....	8
1. ĐẶC TẢ ĐỀ TÀI	9
1.1. Mục tiêu đề tài	9
1.2. Dữ liệu, thông tin đầu vào	9
1.3. Mục đích, tính năng của đề tài.....	9
1.4. Kết quả dự kiến	10
2. KẾ HOẠCH THỰC HIỆN	11
3. THIẾT KẾ	12
3.1. Tiền xử lý dữ liệu	12
3.2. Thiết kế cơ sở dữ liệu nguồn	15
3.3. Mô hình hóa kho dữ liệu	18
3.4. ETL dữ liệu.....	21
3.4.1. Package DateDimensionImport	21
3.4.1.1. Control flow	21
3.4.1.2. Data flow	22
3.4.2. Package ETL_Tables.....	23
3.4.2.1. Control flow	23
3.4.2.2. Data flow	23
3.4.3. Package ETL_FactSales.....	27
3.4.3.1. Control flow	27
3.4.3.2. Data flow	28
3.5. Data cube design với SSAS	30
4. CÀI ĐẶT VÀ KIỂM THỦ	33
4.1. Xây dựng dashboard trên Excel pivot table.....	33

4.2. Thủ nghiệm dashboard trên Excel Pivot Table	36
4.3. Xây dựng dashboard trên Power BI	39
4.1. Thủ nghiệm dashboard trên Power BI.....	41
5. KẾT LUẬN	44
5.1. Các kết quả đạt được.....	44
5.2. Ưu điểm, khó khăn và hạn chế của đề tài.....	44
5.3. Định hướng phát triển	45
TÀI LIỆU THAM KHẢO.....	46

DANH MỤC TỪ VIẾT TẮT

STT	Ký hiệu chữ viết tắt	Chữ viết đầy đủ
1	CSDL	Cơ sở dữ liệu
2	SSAS	SQL Server Analysis Services: công cụ phân tích dữ liệu và xử lý trực tuyến
3	SSIS	SQL Server Integration Services: tích hợp dữ liệu và chuyển đổi dữ liệu
4	ETL	Extract, Transform, Load: trích xuất, biến đổi, tải dữ liệu
5	MOLAP	Multi-dimensional Online Analytical Processing: xử lý phân tích trực tuyến đa chiều
6	OLAP	Online Analytical Processing: Xử lý phân tích trực tuyến
7	DF	Data flow: lộ trình dữ liệu di chuyển qua lại giữa các đơn vị bên ngoài, quy trình và kho lưu trữ dữ liệu.

DANH MỤC CÁC HÌNH

Hình 1: Diagram thể hiện mối quan hệ giữa các bảng	15
Hình 2: Mô hình star schema của kho dữ liệu	18
Hình 3: Control flow cho DimDate.....	21
Hình 4: Trích xuất dữ liệu từ nguồn vào stage.....	22
Hình 5: Tải dữ liệu từ stage vào bảng chiều	22
Hình 6: Control flow các bảng DimProduct, DimCustomer, DimDerpartment	23
Hình 7: Trích xuất dữ liệu từ nguồn vào stage Product	23
Hình 8: Tải dữ liệu từ stage vào bảng chiều Product	24
Hình 9: Trích xuất dữ liệu từ nguồn vào stage Customer	25
Hình 10: Tải dữ liệu từ stage vào bảng chiều Customer.....	25
Hình 11: Trích xuất dữ liệu từ nguồn vào stage Department.....	26
Hình 12: Tải dữ liệu từ stage vào bảng chiều Department	26
Hình 13: Control flow cho bảng Sales	27
Hình 14: Trích xuất dữ liệu từ nguồn vào stage Sales	28
Hình 15: Biến đổi và tải dữ liệu từ stage vào bảng FactSales.....	28
Hình 16: Tạo Data Source	30
Hình 17: Tạo Data source view.....	30
Hình 18: Tạo Data cube	31
Hình 19: Phân cấp dữ liệu DimDate	31
Hình 20: Phân cấp dữ liệu DimCustomer	31
Hình 21: Phân cấp dữ liệu Dim Product	32

Hình 22: Phân tích trên Data cube đã tạo bằng SSAS	32
Hình 23: Kết nối đến khối dữ liệu.....	33
Hình 24: Thiết lập kết nối	33
Hình 25: Chọn khối dữ liệu.....	34
Hình 26: Lưu kết nối đến dữ liệu	34
Hình 27: Lựa chọn kết nối đã lưu	35
Hình 28: Lựa chọn hình thức tổ chức dữ liệu	35
Hình 29: Giao diện thực hiện trực quan hóa trên Excel Pivot Table	35
Hình 30: Dashboard tạo bằng Pivot Table	36
Hình 31: Tình hình kinh doanh ở năm 2015, 2016 tại Mỹ	37
Hình 32: Tình hình kinh doanh ở năm 2017 của phòng ban Apparel tại California	38
Hình 33: Kết nối đến cube data.....	39
Hình 34: Thiết lập kết nối	39
Hình 35: Chọn cube dữ liệu đã tạo.....	40
Hình 36: Giao diện thực hiện trực quan hóa trên Power BI.....	40
Hình 37: Dashboard tạo bằng Power BI	41
Hình 38: Tình hình kinh doanh ở Puerto Rico vào năm 2016 tại thị trường Europe.....	42
Hình 39: Tình hình kinh doanh của phòng ban Fitness ở tiểu bang California vào quý 2, năm 2016	43

DANH MỤC CÁC BẢNG

Bảng 1: Kế hoạch thực hiện đề tài	11
Bảng 2: Code thực hiện tiền xử lý dữ liệu	12
Bảng 3: Bảng mô tả các Table trong CSDL.....	16
Bảng 4: Mô hình hóa kho dữ liệu.....	19
Bảng 6: Mô tả Control flow cho bảng DimDate.....	21
Bảng 7: Mô tả data flow cho DimDate	22
Bảng 8: Mô tả Control flow các bảng DimProduct, DimCustomer, DimDerpartment	23
Bảng 9: Mô tả data flow cho DimProduct	24
Bảng 10: Mô tả data flow Customer	25
Bảng 11: Mô tả data flow cho DimDerpartment.....	26
Bảng 12: Mô tả Control flow cho bảng Sales	27
Bảng 13: Mô tả data flow cho bảng FactSales.....	29
Bảng 14: Các bước thực hiện thiết kế Data cube	30
Bảng 15: Các bước xây dựng dashboard trên Pivot Table.....	33
Bảng 16: Các bước xây dựng dashboard trên Power BI	39

1. ĐẶC TẢ ĐỀ TÀI

1.1. Mục tiêu đề tài

Tạo cơ sở dữ liệu tổng hợp: Xây dựng một kho dữ liệu tập trung từ nguồn dữ liệu bán hàng trên Kaggle, bao gồm các thông tin chi tiết như danh mục sản phẩm, doanh số, chi phí, lợi nhuận, thị trường, khách hàng, và các yếu tố ảnh hưởng đến hiệu quả kinh doanh.

Trực quan hóa dữ liệu hiệu quả: Sử dụng các công cụ phân tích và trực quan hóa (như Power BI hoặc Excel Pivot Table) để trình bày dữ liệu một cách dễ hiểu, giúp người dùng nhanh chóng nhận diện các cơ hội kinh doanh, rủi ro tiềm ẩn và các chỉ số quan trọng.

1.2. Dữ liệu, thông tin đầu vào

Đề tài sử dụng dữ liệu bán hàng được lấy từ nền tảng Kaggle, bao gồm các thông tin liên quan đến sản phẩm, khách hàng, đơn hàng, và thị trường. Dữ liệu cung cấp chi tiết về các giao dịch bán hàng như giá cả, số lượng, doanh thu, chi phí, lợi nhuận, cũng như thông tin vị trí địa lý và thời gian giao dịch. Tập dữ liệu này sẽ được làm sạch và chuẩn hóa trước khi tích hợp vào kho dữ liệu, nhằm đảm bảo tính chính xác và đầy đủ, phục vụ hiệu quả cho quá trình phân tích và trực quan hóa.

1.3. Mục đích, tính năng của đề tài

Đề tài hướng đến việc xây dựng một hệ thống kho dữ liệu tập trung, giúp tổ chức và quản lý dữ liệu bán hàng một cách hiệu quả. Hệ thống này không chỉ lưu trữ và tích hợp dữ liệu từ nhiều nguồn mà còn cung cấp các công cụ hỗ trợ phân tích sâu, khám phá xu hướng và đưa ra các quyết định chiến lược dựa trên dữ liệu.

Kho dữ liệu sẽ giúp dễ dàng theo dõi doanh thu, lợi nhuận, hành vi khách hàng và hiệu quả kinh doanh theo nhiều góc độ khác nhau. Đồng thời, thông qua các báo cáo và biểu đồ trực quan, có thể nhanh chóng nắm bắt thông tin, tương tác với dữ liệu, và phân tích dữ liệu đa chiều.

Đề tài được xây dựng với mục đích là nghiên cứu và phân tích các câu hỏi liên quan đến doanh thu, kinh phí, lợi nhuận của ngành công nghiệp bán lẻ nên phạm vi chỉ dừng lại ở bước phân tích và xây dựng các báo cáo thể hiện dữ liệu một cách trực quan.

1.4. Kết quả dự kiến

Đề tài sẽ xây dựng thành công một hệ thống dashboard trực quan hóa dữ liệu bán hàng dựa trên kho dữ liệu được thiết kế. Dashboard sẽ cung cấp các biểu đồ, bảng số liệu và báo cáo trực quan, cho phép người dùng dễ dàng theo dõi các chỉ số quan trọng như doanh thu, lợi nhuận, xu hướng bán hàng, và hành vi khách hàng.

Hệ thống dự kiến sẽ hỗ trợ việc phân tích dữ liệu đa chiều, giúp người dùng nắm bắt dữ liệu kinh doanh theo thời gian, khu vực, hoặc danh mục sản phẩm cụ thể. Thông qua các tính năng tương tác, người dùng có thể tự điều chỉnh góc nhìn dữ liệu theo nhu cầu, từ đó hỗ trợ quá trình ra quyết định nhanh chóng và hiệu quả hơn.

2. KẾ HOẠCH THỰC HIỆN

Bảng 1: Kế hoạch thực hiện đề tài

STT	Sinh viên phụ trách	Nhiệm vụ (công việc dự kiến)	Phần trăm hoàn thành
1	Trần Nguyễn Trí Đạt	Phân tích yêu cầu đề tài và tìm hiểu các công cụ, kiến thức liên quan. Lựa chọn tập dữ liệu cho đề tài trên Kaggle	100%
2	Trần Nguyễn Trí Đạt	Thực hiện tiền xử lý dữ liệu trên tập dữ liệu được chọn Phân tích các quy trình nghiệp vụ (Business Process) cho tập dữ liệu.	100%
3	Trần Nguyễn Trí Đạt	Phân tích quy trình nghiệp vụ và đưa ra các câu hỏi liên quan. Xây dựng và lựa chọn mô hình cho kho dữ liệu Mô hình hóa dữ liệu	100%
4	Trần Nguyễn Trí Đạt	Thực hiện ETL (Extract, Transform, Load) dữ liệu vào kho dữ liệu. Nhập dữ liệu vào công cụ xử lý phân tích và khai thác dữ liệu (SSAS)	100%
5	Trần Nguyễn Trí Đạt	Trực quan hóa dữ liệu trên các công cụ trực quan hóa dữ liệu (Excel pivot table, Power BI) Trả lời cho các câu hỏi đặt ra cho quy trình nghiệp vụ bằng cách trực quan hóa dữ liệu	100%
6	Trần Nguyễn Trí Đạt	Tiến hành quá trình tổng hợp, kiểm thử, viết báo cáo.	100%

3. THIẾT KẾ

Để có thể xây dựng kho dữ liệu để có thể trực quan hóa dữ liệu cần phải thực hiện qua các bước chính nhằm đảm bảo hệ thống hoạt động hiệu quả và đáp ứng các yêu cầu về phân tích dữ liệu. Đầu tiên, tập dữ liệu sẽ được thu thập và tiến hành tiền xử lý, bao gồm làm sạch, chuẩn hóa và xử lý các giá trị thiếu nhằm đảm bảo chất lượng và tính nhất quán. Tiếp theo, chia tập dữ liệu thành các bảng và được thiết kế vào cơ sở dữ liệu một cách hợp lý, tạo nền tảng cho việc lưu trữ và truy vấn dữ liệu.

Sau đó, kho dữ liệu được mô hình hóa theo kiến trúc star schema hoặc snowflake schema, với các bảng fact và dimension được xây dựng phù hợp để tối ưu hóa cho các phân tích đa chiều. Quá trình ETL (Extract, Transform, Load) được thực hiện bằng công cụ SSIS, giúp trích xuất dữ liệu từ nguồn, chuyển đổi thành định dạng phù hợp và tải vào kho dữ liệu.

Cuối cùng, Data Cube được thiết kế bằng công cụ SSAS để hỗ trợ phân tích OLAP (Online Analytical Processing). Cube cho phép tổ chức dữ liệu theo nhiều chiều, giúp người dùng dễ dàng thực hiện các truy vấn phân tích và trực quan hóa dữ liệu trên các công cụ như Power BI hoặc Excel, tạo ra các báo cáo chi tiết và biểu đồ minh họa trực quan, hỗ trợ việc ra quyết định một cách hiệu quả.

3.1. Tiền xử lý dữ liệu

Sinh viên phụ trách: Trần Nguyễn Trí Đạt

Bảng 2: Code thực hiện tiền xử lý dữ liệu

STT	Code	Mục đích
1	<pre>import pandas as pd data= pd.read_csv('DataCoSupplyChainDataset.csv', encoding='ISO-8859-1') print(data.head()) print(data.shape) data</pre>	Import thư viện Pandas của Python để thực hiện xử lý dữ liệu. Đọc file dataset vào biến data để dễ dàng thực hiện xử lý dữ liệu.

2	<pre>null_values = data.isnull().sum() null_columns = null_values=null_values[null_values > 0] print("Những cột có giá trị null:") print(null_columns) data = data.fillna("Unknown")</pre>	Tìm những cột có nhiều giá trị null và tự động điền giá trị null thành giá trị “Unknown”.
3	<pre>null_values = data.isnull().sum() null_columns = null_values=null_values[null_values > 0] print("Những cột có giá trị null:") print(null_columns)</pre>	Sau khi tự động điền các giá trị null, kiểm tra lại để đảm bảo không còn cột nào có giá trị null.
4	<pre>selected_columns = data[['Customer Zipcode', 'Customer City', 'Customer Country', 'Customer State']] unique_zipcode_data = selected_columns.drop_duplicates(subset='Customer Zipcode') print(unique_zipcode_data) unique_zipcode_data.to_csv('city.csv', index=False)</pre>	Chọn ra các cột có liên quan đến thông tin thành phố, và loại bỏ các dòng bị trùng và xuất ra file “city.csv” để đưa vào cơ sở dữ liệu SQL Server
5	<pre>selected_columns = data[['Customer Zipcode', 'Customer Segment', 'Customer Id', 'Customer Fname', 'Customer Lname']] unique_customer_data = selected_columns.drop_duplicates(subset='Customer Id') print(unique_customer_data) unique_customer_data.to_csv('customers.csv', index=False)</pre>	Chọn ra các cột có liên quan đến thông tin khách hàng, và loại bỏ các dòng bị trùng và xuất ra file “customers.csv” để đưa vào cơ sở dữ liệu SQL Server.
6	<pre>selected_columns = data[['Product Card Id', 'Product Name', 'Category Id', 'Product Status', 'Product Price']] unique_product_data = selected_columns.drop_duplicates(subset='Product Card Id') print(unique_product_data.head()) unique_product_data.to_csv('product.csv', index=False)</pre>	Chọn ra các cột có liên quan đến thông tin sản phẩm, và loại bỏ các dòng bị trùng và xuất ra file “product.csv” để đưa vào cơ sở dữ liệu SQL Server.

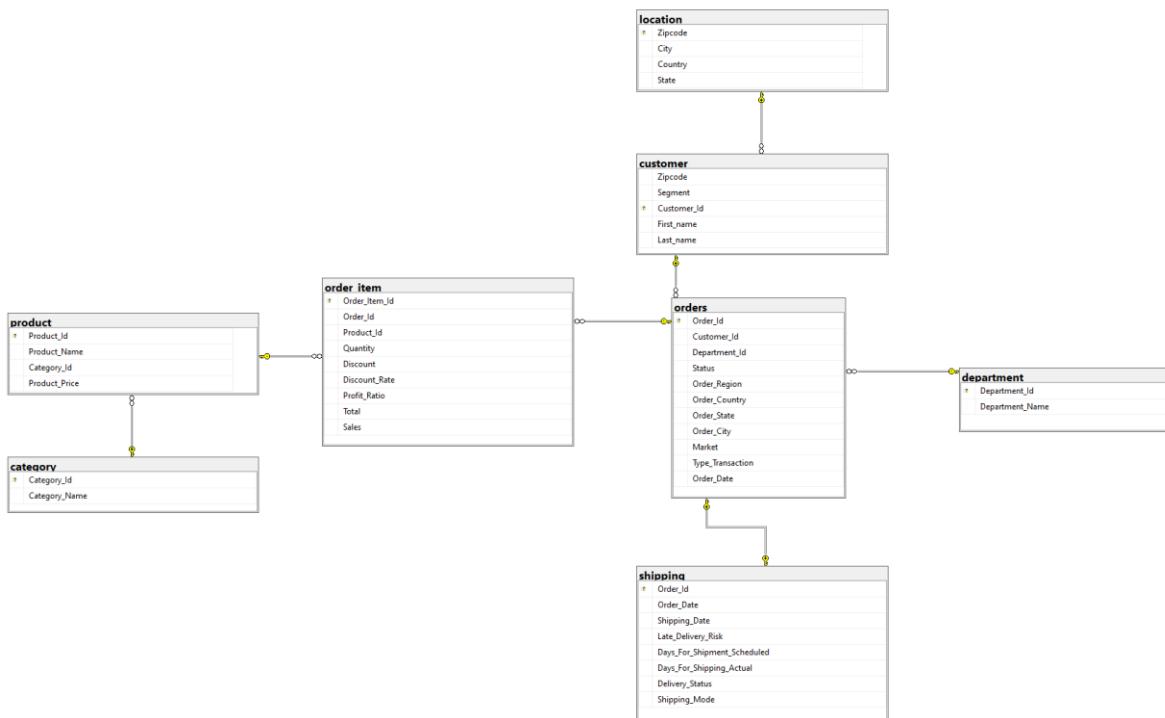
7	<pre>selected_columns = data[['Category Id', 'Category Name']] unique_category_data = selected_columns.drop_duplicates(subset='Category Id') print(unique_category_data) unique_category_data.to_csv('category.csv', index=False)</pre>	<p>Chọn ra các cột có liên quan đến thông tin loại sản phẩm, và loại bỏ các dòng bị trùng và xuất ra file “category.csv” để đưa vào cơ sở dữ liệu SQL Server.</p>
8	<pre>selected_columns = data[['Department Id', 'Department Name']] unique_Department_data = selected_columns.drop_duplicates(subset='Department Id') print(unique_Department_data) unique_Department_data.to_csv('Department.csv', index=False)</pre>	<p>Chọn ra các cột có liên quan đến thông tin phòng ban, và loại bỏ các dòng bị trùng và xuất ra file “Department.csv” để đưa vào cơ sở dữ liệu SQL Server.</p>
9	<pre>selected_columns = data[['Order Id', 'Customer Id', 'Department Id', 'Order Status', 'Order Region', 'Order Country', 'Order State', 'Order City', 'Market', 'Type', 'order date (DateOrders)']] unique_order_data = selected_columns.drop_duplicates(subset='Order Id') print(unique_order_data) unique_order_data.to_csv('order.csv', index=False)</pre>	<p>Chọn ra các cột có liên quan đến thông tin đơn hàng, và loại bỏ các dòng bị trùng và xuất ra file “order.csv” để đưa vào cơ sở dữ liệu SQL Server.</p>
10	<pre>selected_columns = data[['Order Id', 'order date (DateOrders)', 'shipping date (DateOrders)', 'Late_delivery_risk', 'Days for shipment (scheduled)', 'Days for shipping (real)', 'Delivery Status', 'Shipping Mode']] unique_Shipping_data = selected_columns.drop_duplicates(subset='Order Id') print(unique_Shipping_data) unique_Shipping_data.to_csv('shipping.csv', index=False)</pre>	<p>Chọn ra các cột có liên quan đến thông tin vận chuyển, và loại bỏ các dòng bị trùng và xuất ra file “shipping.csv” để đưa vào cơ sở dữ liệu SQL Server.</p>

11	<pre>selected_columns = data[['Order Item Id', 'Order Id', 'Product Card Id', 'Order Item Quantity', 'Order Item Discount', 'Order Item Discount Rate', 'Order Item Profit Ratio', 'Order Item Total', 'Sales']] unique_order_Item_data = selected_columns.drop_duplicates(subset='Order Item Id') print(unique_order_Item_data) unique_order_Item_data.to_csv('order_Item.csv', index=False)</pre>	<p>Chọn ra các cột có liên quan đến thông tin chi tiết đơn hàng, và loại bỏ các dòng bị trùng và xuất ra file “order_Item.csv” để đưa vào cơ sở dữ liệu SQL Server.</p>
----	---	---

3.2. Thiết kế cơ sở dữ liệu nguồn

Sinh viên phụ trách: Trần Nguyễn Trí Đạt

Diagram thể hiện mối quan hệ giữa các bảng:



Hình 1: Diagram thể hiện mối quan hệ giữa các bảng

Bảng 3: Bảng mô tả các Table trong CSDL

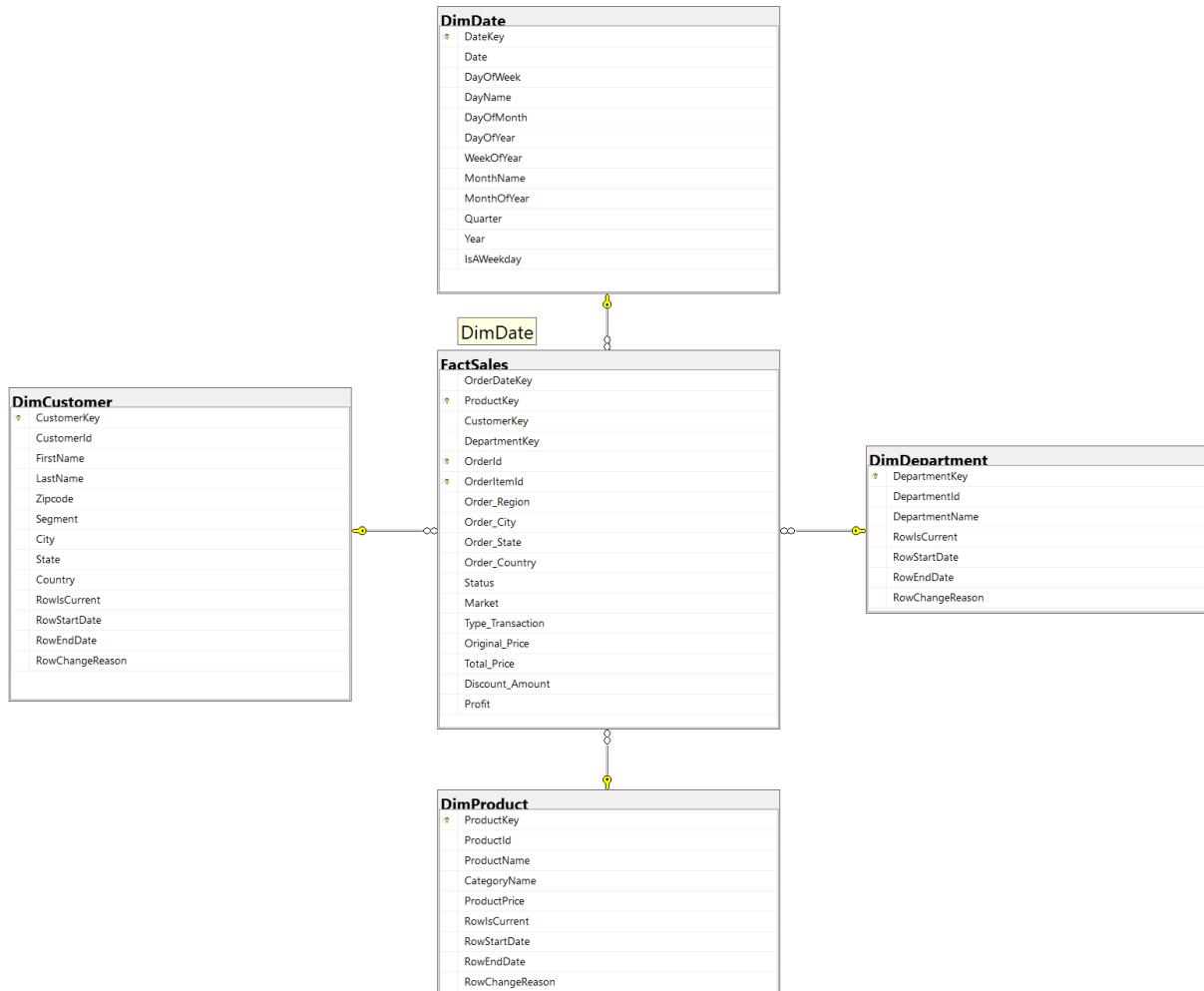
STT	Tên bảng	Tên trường	Kiểu dữ liệu	Mục đích
1	Location	<u>Zipcode</u>	Varchar(6)	Lưu trữ thông tin về vị trí địa lý
		City	Varchar(50)	
		Country	Varchar(50)	
		State	Varchar(3)	
2	Customer	<u>Customer_Id</u>	Int	Lưu trữ thông tin về khách hàng
		Zipcode	Varchar(6)	
		Segment	Varchar(50)	
		First_name	Varchar(50)	
		Last_name	Varchar(50)	
3	Department	<u>Department_Id</u>	Int	Lưu trữ thông tin về phòng ban của cửa hàng
		Department_Name	Varchar(50)	
4	Category	<u>Category_Id</u>	Tinyint	Lưu trữ thông tin về loại hình sản phẩm
		Category_Name	Varchar(50)	
5	Product	<u>Product_Id</u>	Int	Lưu trữ thông tin về sản phẩm
		Product_Name	Varchar(200)	
		Category_Id	Tinyint	
		Product_Price	Float	
6	Orders	<u>Order_Id</u>	Int	Lưu trữ thông tin về đơn hàng
		Customer_Id	Int	
		Department_Id	Int	
		Status	Varchar(50)	
		Order_Region	Varchar(50)	
		Order_Country	Varchar(50)	
		Order_State	Varchar(50)	

		Order_City	Varchar(50)	
		Market	Varchar(50)	
		Type_Transaction	Varchar(50)	
		Order_Date	Datetime	
7	Order_Item	<u>Order_Item_Id</u>	Int	Lưu trữ thông tin về chi tiết đơn hàng
		Order_Id	Int	
		Product_Id	Int	
		Quantity	Int	
		Discount	Float	
		Discount_Rate	Float	
		Profit_Ratio	Float	
		Total	Float	
		Sales	Float	
8	Shipping	<u>Order_Id</u>	Int	Lưu trữ thông tin về vận chuyển
		Order_Date	Datetime	
		Shipping_Date	Datetime	
		Late_Delivery_Risk	Bit	
		Days_For_Shipment_Scheduled	Int	
		Days_For_Shipping_Actual	Int	
		Delivery_Status	Varchar(50)	
		Shipping_Mode	Varchar(50)	

3.3. Mô hình hóa kho dữ liệu

Sinh viên phụ trách: Trần Nguyễn Trí Đạt

Diagram thể hiện mối quan hệ giữa các bảng chiều:



Hình 2: Mô hình star schema của kho dữ liệu

Bảng 4: Mô hình hóa kho dữ liệu

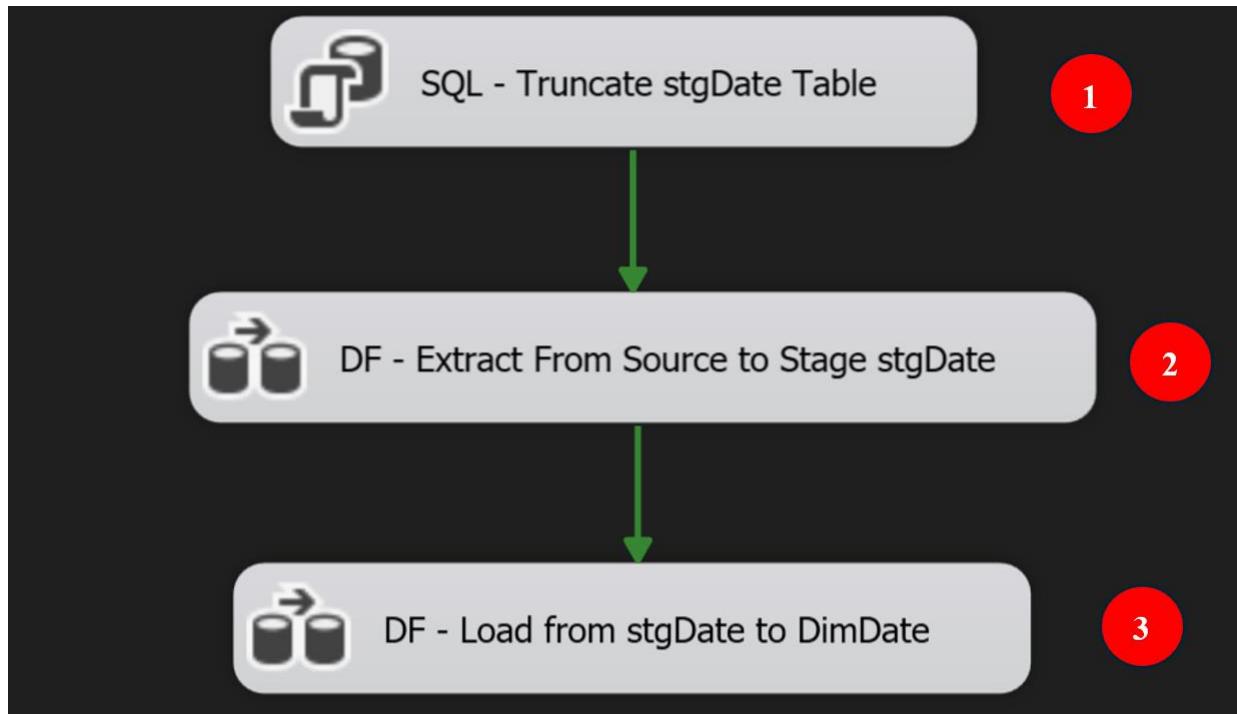
STT	Code	Mục đích
1	<pre>-- Tạo cơ sở dữ liệu CREATE DATABASE DataCo_DWH; GO USE DataCo_DWH; -- Tạo bảng DimDate CREATE TABLE DimDate(DateKey int NOT NULL, Date datetime NOT NULL, DayOfWeek tinyint NOT NULL, DayName varchar(9) NOT NULL, DayOfMonth tinyint NOT NULL, DayOfYear smallint NOT NULL, WeekOfYear tinyint NOT NULL, MonthName varchar(9) NOT NULL, MonthOfYear tinyint NOT NULL, Quarter tinyint NOT NULL, Year smallint NOT NULL, IsAWeekday varchar(1) NOT NULL DEFAULT ('N'), constraint PK_DimDate PRIMARY KEY (DateKey))</pre>	<p>Tạo kho dữ liệu “DataCo_DWH” để thực hiện xây dựng kho dữ liệu.</p> <p>Tạo bảng chiều DimDate lưu trữ dữ liệu thông tin các ngày tháng qua các năm. Dữ liệu của DimDate được lấy từ file excel “Ch3-SampleDateDim” có chức năng tạo script tự động</p>
2	<pre>-- Tạo bảng DimProduct CREATE TABLE DimProduct (ProductKey INT IDENTITY PRIMARY KEY, -- attributes ProductId INT NOT NULL, ProductName VARCHAR(200) NOT NULL, CategoryName VARCHAR(50) NOT NULL, ProductPrice FLOAT NOT NULL, -- metadata RowIsCurrent BIT DEFAULT 1 NOT NULL, RowStartDate DATETIME DEFAULT '12/31/1899' NOT NULL, RowEndDate DATETIME DEFAULT '12/31/9999' NOT NULL, RowChangeReason NVARCHAR(200) NULL,);</pre>	<p>Tạo bảng chiều DimProduct lưu trữ dữ liệu thông tin các sản phẩm có trong tập dữ liệu.</p>
3	<pre>-- Tạo bảng DimCustomer CREATE TABLE DimCustomer (CustomerKey INT IDENTITY PRIMARY KEY, -- Attributes CustomerId INT NOT NULL, FirstName VARCHAR(50) NOT NULL, LastName VARCHAR(50) NOT NULL, Zipcode VARCHAR(6) NOT NULL, Segment VARCHAR(50) NOT NULL, City VARCHAR(50) NOT NULL, State VARCHAR(3) NOT NULL, Country VARCHAR(50) NOT NULL, -- Metadata RowIsCurrent BIT DEFAULT 1 NOT NULL,</pre>	<p>Tạo bảng chiều DimCustomer lưu trữ dữ liệu thông tin các khách hàng có trong tập dữ liệu.</p>

	<pre> RowStartDate DATETIME DEFAULT '12/31/1899' NOT NULL, RowEndDate DATETIME DEFAULT '12/31/9999' NOT NULL, RowChangeReason NVARCHAR(200) NULL,); </pre>	
4	<pre> -- Tạo bảng DimDepartment CREATE TABLE DimDepartment (DepartmentKey INT IDENTITY PRIMARY KEY, -- Attributes DepartmentId INT NOT NULL, DepartmentName VARCHAR(50) NOT NULL, -- Metadata RowIsCurrent BIT DEFAULT 1 NOT NULL, RowStartDate DATETIME DEFAULT '12/31/1899' NOT NULL, RowEndDate DATETIME DEFAULT '12/31/9999' NOT NULL, RowChangeReason NVARCHAR(200) NULL); </pre>	Tạo bảng chiều DimDepartment lưu trữ dữ liệu thông tin các phòng ban có trong tập dữ liệu.
5	<pre> -- Tạo bảng FactSales CREATE TABLE FactSales (OrderDateKey int NOT NULL, ProductKey INT NOT NULL, CustomerKey INT NOT NULL, DepartmentKey INT NOT NULL, -- Attributes OrderId INT NOT NULL, OrderItemId INT NOT NULL, Order_Region VARCHAR(50) NOT NULL, Order_City VARCHAR(50) NOT NULL, Order_State VARCHAR(50) NOT NULL, Order_Country VARCHAR(50) NOT NULL, Status VARCHAR(50) NOT NULL, Market VARCHAR(50) NOT NULL, Type_Transaction VARCHAR(50) NOT NULL, Original_Price FLOAT NOT NULL, Total_Price FLOAT NOT NULL, Discount_Amount FLOAT NOT NULL, Profit FLOAT NOT NULL, -- Constraints CONSTRAINT PK_FactSales PRIMARY KEY (ProductKey, OrderId, OrderItemId), CONSTRAINT FK_FactSales_Date FOREIGN KEY (OrderDateKey) REFERENCES DimDate(DateKey), CONSTRAINT FK_FactSales_Product FOREIGN KEY (ProductKey) REFERENCES DimProduct(ProductKey), CONSTRAINT FK_FactSales_Customer FOREIGN KEY (CustomerKey) REFERENCES DimCustomer(CustomerKey), CONSTRAINT FK_FactSales_Department FOREIGN KEY (DepartmentKey) REFERENCES DimDepartment(DepartmentKey)); </pre>	<p>Tạo bảng sự thật FactSales chứa các thông tin số lượng (measurements) của dữ liệu để thực hiện tính toán theo các bảng chiều để đưa ra dữ liệu trực quan nhất</p> <p>Xây dựng kho dữ liệu theo star schema vì bảng fact là trung tâm của mô hình với các bảng dimension xung quanh nó, nó nhìn giống như một ngôi sao. Bởi vì bảng fact liên quan đến mỗi bảng dimension bởi một quan hệ nên giúp đơn giản truy vấn và giảm thời gian thực thi</p>

3.4. ETL dữ liệu

3.4.1. Package DateDimensionImport

3.4.1.1. Control flow

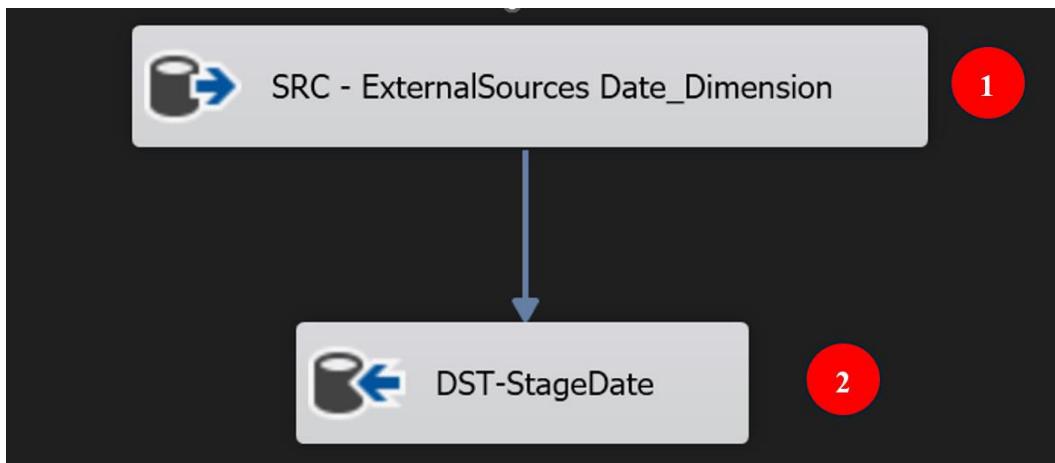


Hình 3: Control flow cho DimDate

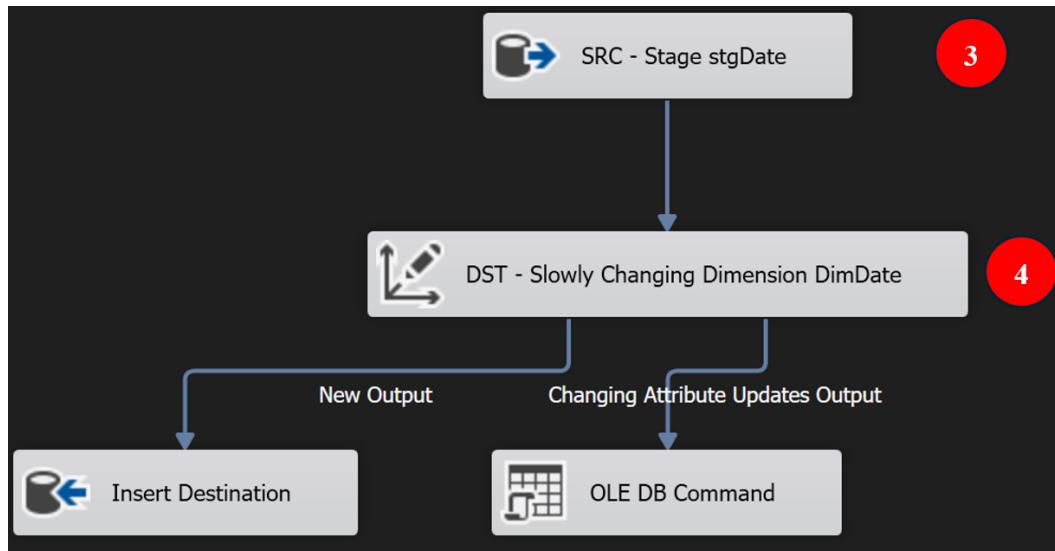
Bảng 5: Mô tả Control flow cho bảng DimDate

STT	Kiểu	Mô tả / Ghi chú
1	Execute SQL Task	Thực hiện xóa dữ liệu trong bảng stage Date
2	Data Flow Task	Trích xuất dữ liệu từ nguồn vào bảng stage Date
3	Data Flow Task	Tải dữ liệu từ stage Date sang bảng chiều DimDate

3.4.1.2. Data flow



Hình 4: Trích xuất dữ liệu từ nguồn vào stage



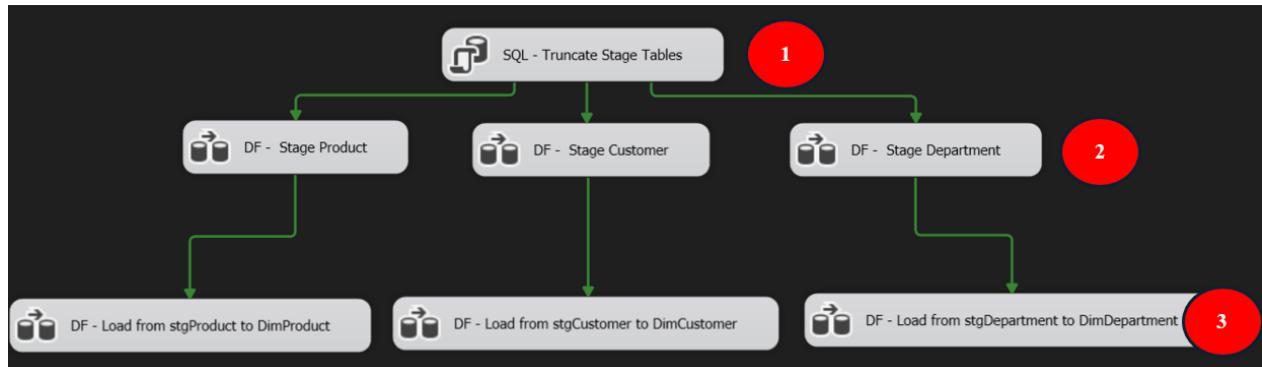
Hình 5: Tải dữ liệu từ stage vào bảng chiều

Bảng 6: Mô tả data flow cho DimDate

STT	Kiểu	Mô tả / Ghi chú
1	Source Assistant	Thực hiện lấy dữ liệu từ bảng nguồn
2	Destination Assistant	Đỗ dữ liệu từ nguồn vào bảng stage Date
3	Source Assistant	Lấy dữ liệu từ bảng stage Date
4	Slowly Changing Dimension	Tải dữ liệu từ bảng stage vào bảng chiều DimDate

3.4.2. Package ETL_Tables

3.4.2.1. Control flow



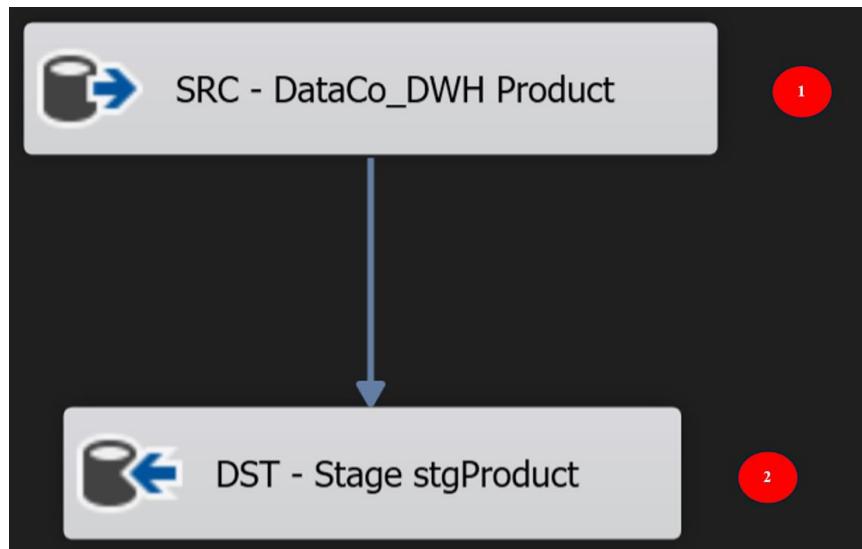
Hình 6: Control flow các bảng DimProduct, DimCustomer, DimDerpartment

Bảng 7: Mô tả Control flow các bảng DimProduct, DimCustomer, DimDerpartment

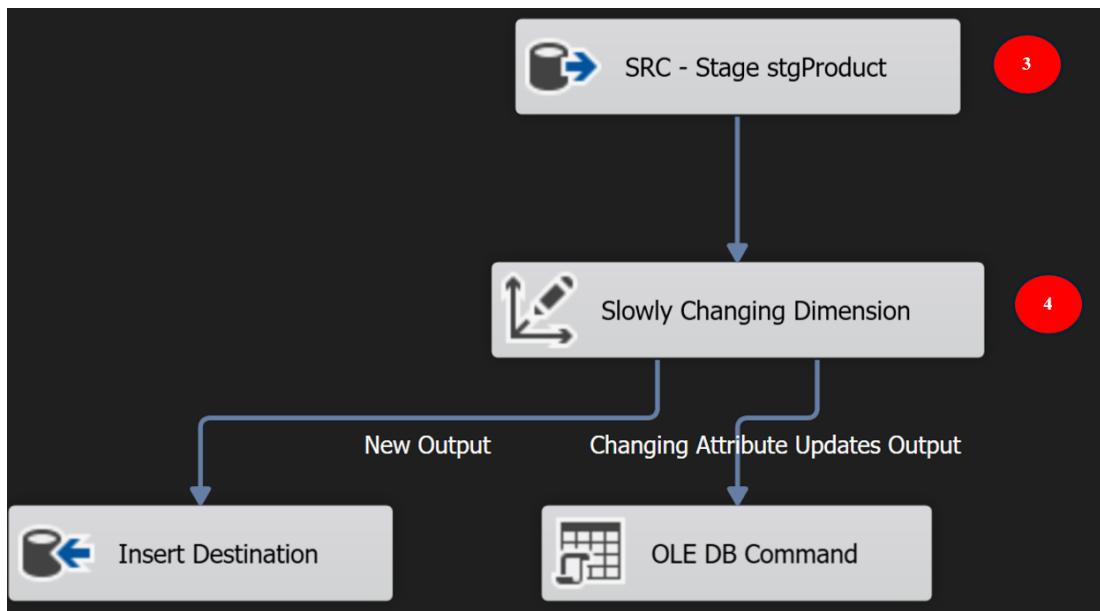
STT	Kiểu	Mô tả / Ghi chú
1	Execute SQL Task	Thực hiện xóa dữ liệu trong các bảng stage
2	Data Flow Task	Trích xuất dữ liệu từ nguồn vào các bảng stage
3	Data Flow Task	Tải dữ liệu từ các bảng stage sang các bảng chiều

3.4.2.2. Data flow

- Product



Hình 7: Trích xuất dữ liệu từ nguồn vào stage Product

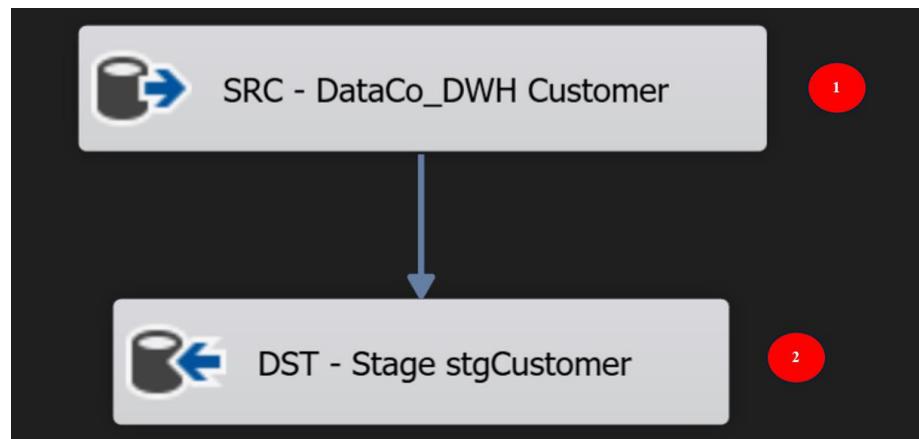


Hình 8: Tải dữ liệu từ stage vào bảng chiều Product

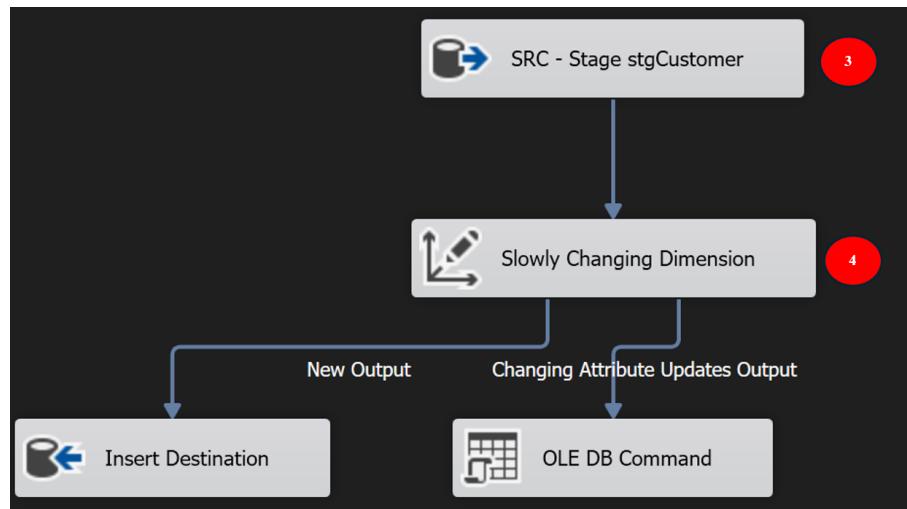
Bảng 8: Mô tả data flow cho DimProduct

STT	Kiểu	Mô tả / Ghi chú
1	Source Assistant	Thực hiện lấy dữ liệu từ bảng nguồn
2	Destination Assistant	Đỗ dữ liệu từ nguồn vào bảng stage Product
3	Source Assistant	Lấy dữ liệu từ bảng stage Product
4	Slowly Changing Dimension	Tải dữ liệu từ bảng stage vào bảng chiều DimProduct

- **Customer**



Hình 9: Trích xuất dữ liệu từ nguồn vào stage Customer

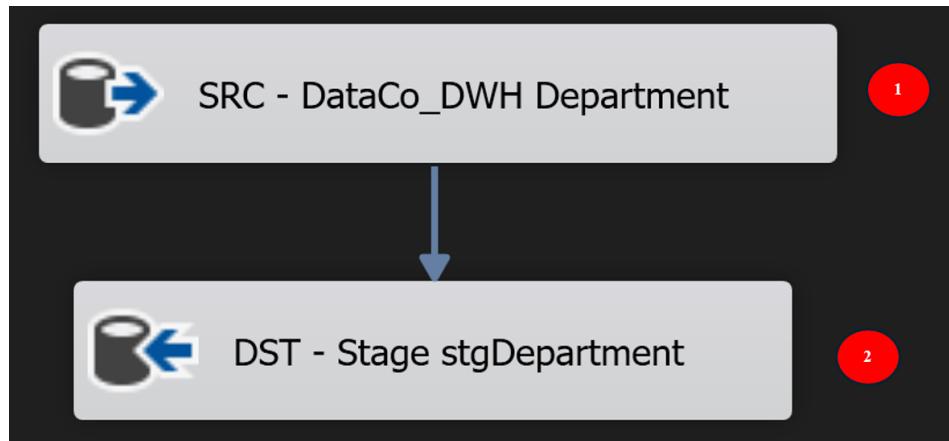


Hình 10: Tải dữ liệu từ stage vào bảng chiều Customer

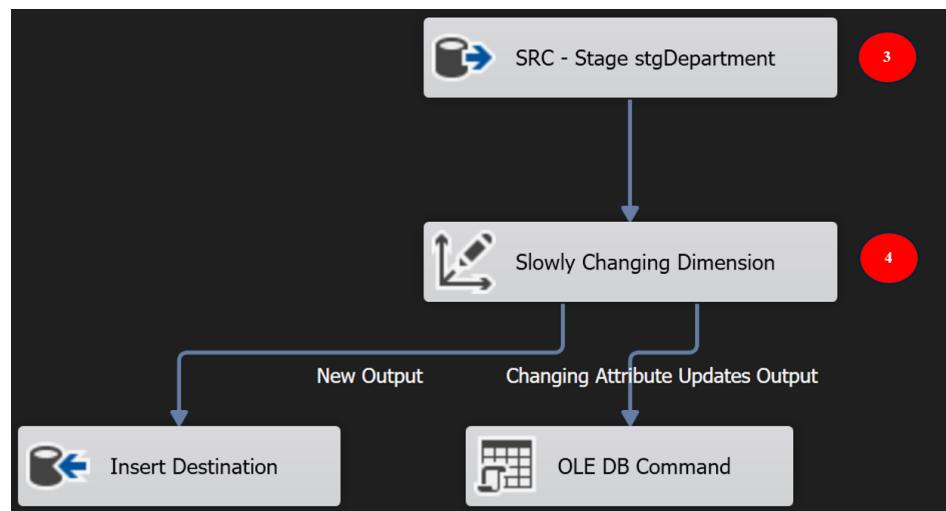
Bảng 9: Mô tả data flow Customer

STT	Kiểu	Mô tả / Ghi chú
1	Source Assistant	Thực hiện lấy dữ liệu từ bảng nguồn
2	Destination Assistant	Đỗ dữ liệu từ nguồn vào bảng stage Customer
3	Source Assistant	Lấy dữ liệu từ bảng stage Customer
4	Slowly Changing Dimension	Tải dữ liệu từ bảng stage vào bảng chiều DimCustomer

- **Department**



Hình 11: Trích xuất dữ liệu từ nguồn vào stage Department



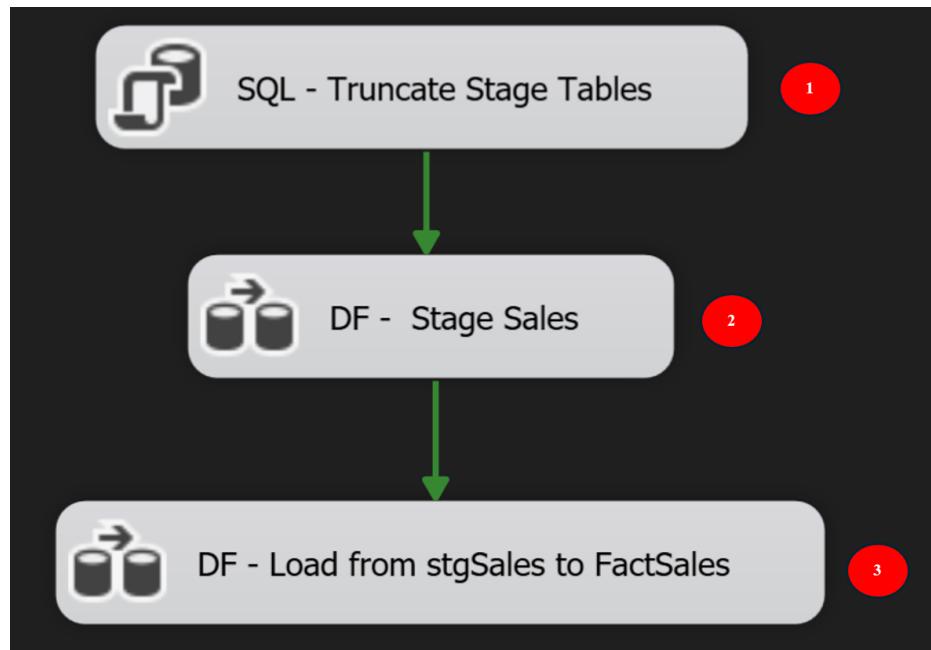
Hình 12: Tải dữ liệu từ stage vào bảng chiều Department

Bảng 10: Mô tả data flow cho DimDepartment

STT	Kiểu	Mô tả / Ghi chú
1	Source Assistant	Thực hiện lấy dữ liệu từ bảng nguồn
2	Destination Assistant	Đỗ dữ liệu từ nguồn vào bảng stage Department
3	Source Assistant	Lấy dữ liệu từ bảng stage Department
4	Slowly Changing Dimension	Tải dữ liệu từ bảng stage vào bảng chiều DimDepartment

3.4.3. Package ETL_FactSales

3.4.3.1. Control flow

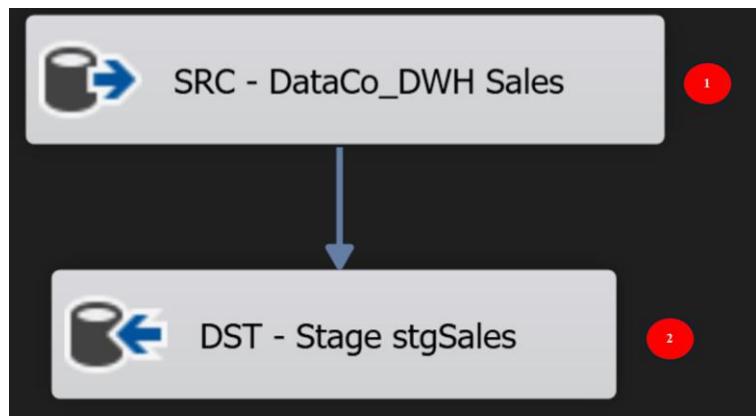


Hình 13: Control flow cho bảng Sales

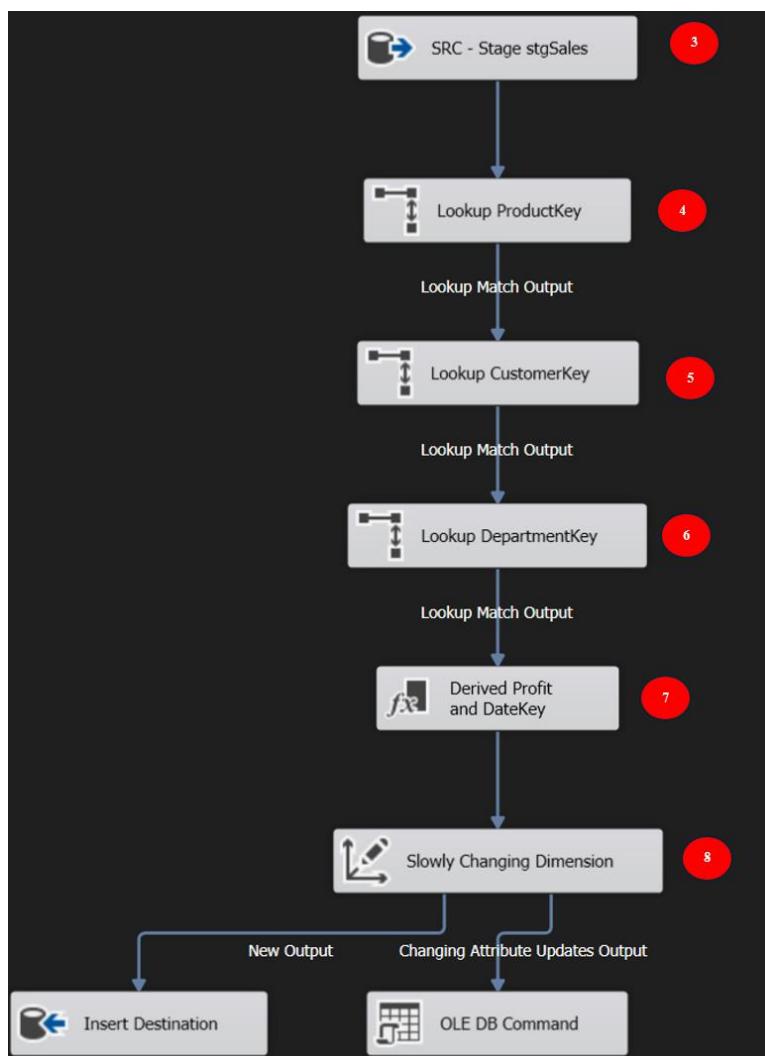
Bảng 11: Mô tả Control flow cho bảng Sales

STT	Kiểu	Mô tả / Ghi chú
1	Execute SQL Task	Thực hiện xóa dữ liệu trong bảng stage Sales
2	Data Flow Task	Trích xuất dữ liệu từ nguồn vào bảng stage Sales
3	Data Flow Task	Tải dữ liệu từ stage Sales sang bảng FactSales

3.4.3.2. Data flow



Hình 14: Trích xuất dữ liệu từ nguồn vào stage Sales



Hình 15: Biến đổi và tải dữ liệu từ stage vào bảng FactSales

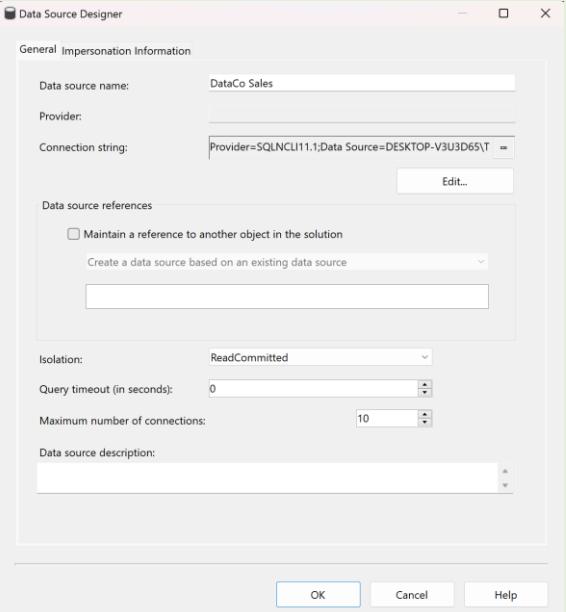
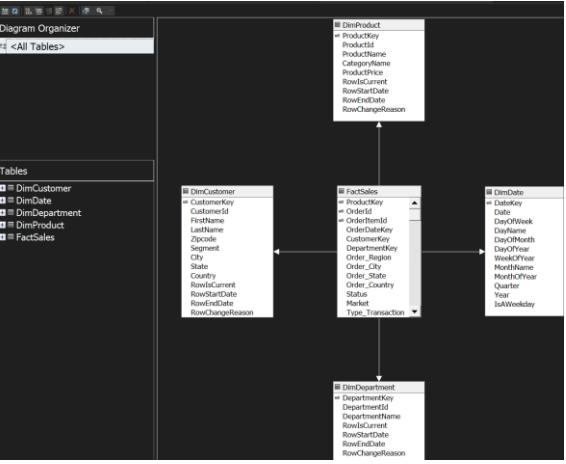
Bảng 12: Mô tả data flow cho bảng FactSales

STT	Kiểu	Mô tả / Ghi chú
1	Source Assistant	Thực hiện lấy dữ liệu từ bảng nguồn
2	Destination Assistant	Đỗ dữ liệu từ nguồn vào bảng stage Department
3	Source Assistant	Lấy dữ liệu từ bảng stage Department
4	Lookup	Tìm kiếm ProductKey để lấy giá trị từ bảng DimProduct
5	Lookup	Tìm kiếm CustomerKey để lấy giá trị từ bảng DimCustomer
6	Lookup	Tìm kiếm DepartmentKey để lấy giá trị từ bảng DimDepartment
7	Derived column	Tính toán giá trị Profit và DateKey cho bảng FactSales
8	Slowly Changing Dimension	Tải dữ liệu từ bảng stage vào bảng FactSales

3.5. Data cube design với SSAS

Sinh viên phụ trách: Trần Nguyễn Trí Đạt

Bảng 13: Các bước thực hiện thiết kế Data cube

STT	Các bước thực hiện	Mục đích
1	<p>Tạo data source:</p>  <p>Hình 16: Tạo Data Source</p>	<ul style="list-style-type: none"> - Tạo data source kết nối đến kho dữ liệu đã tạo trước đó để tạo một cube dữ liệu để phân tích dữ liệu đa chiều hay còn gọi là Multi-dimensional Online Analytical Processing (MOLAP)
2	<p>Tạo data source view:</p>  <p>Hình 17: Tạo Data source view</p>	<ul style="list-style-type: none"> - Data source view chứa logical model của CSDL (tables, keys, columns, và các constraints) sẽ được sử dụng bởi OLAP database để tạo các data cube.
3	<p>Tạo data cube cho data source view:</p>	<ul style="list-style-type: none"> - Sau khi cấu hình và cài đặt ta sẽ tạo ra được một data cube dựa trên data source view, data cube này dùng biểu

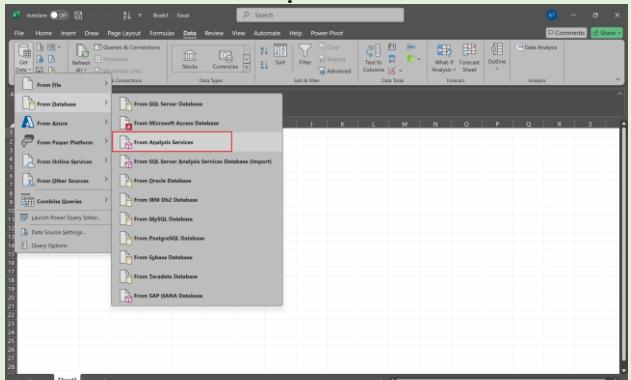
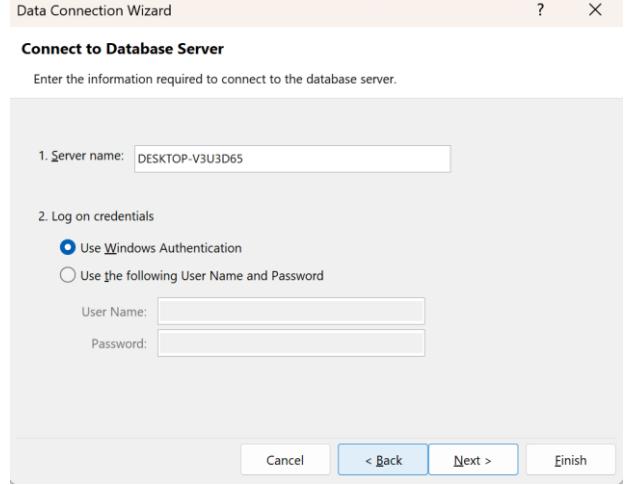
		<p>diễn một mô hình dữ liệu đa chiều, trong đó dữ liệu được tổ chức theo nhiều chiều khác nhau.</p>
4	<p>Phân cấp dữ liệu:</p> <ul style="list-style-type: none"> - DimDate: 	<ul style="list-style-type: none"> - Phân cấp dữ liệu (Hierarchies) trong SSAS được sử dụng để tổ chức các thuộc tính dữ liệu (attributes) theo cấu trúc logic nhằm hỗ trợ việc phân tích, truy vấn, và trực quan hóa dữ liệu dễ dàng và hiệu quả hơn. Sau khi phân cấp dữ liệu thực hiện Process project để hoàn thành tạo cube data
	<p>Hình 19: Phân cấp dữ liệu DimDate</p> <ul style="list-style-type: none"> - DimCustomer: 	<p>Hình 20: Phân cấp dữ liệu DimCustomer</p>

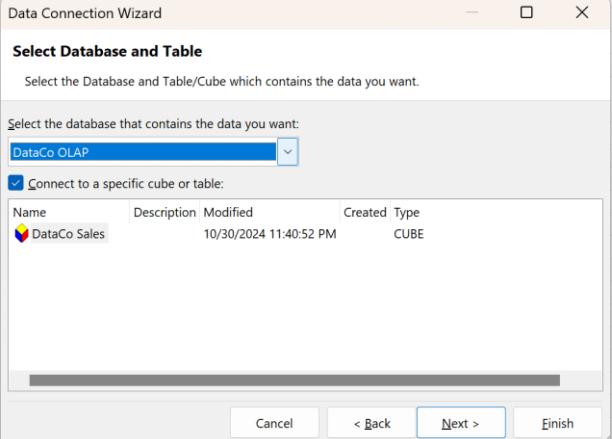
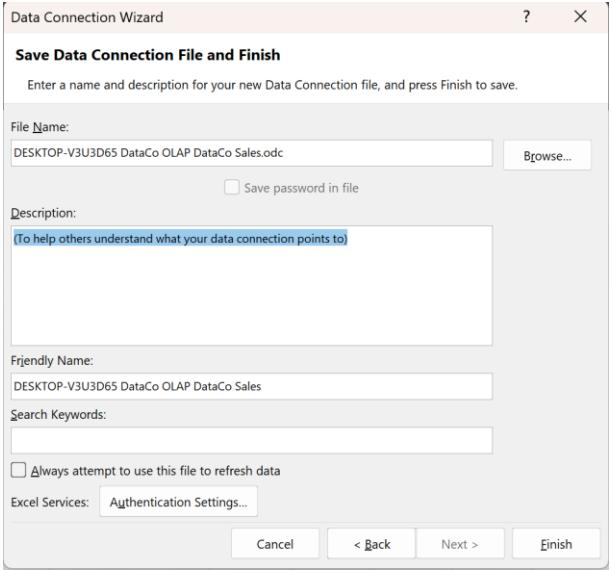
	<p>- DimProduct:</p>	
5	<p>Phân tích trên data cube đã tạo:</p>	<p>- Sau khi hoàn thành Process data cube, chuyển qua tab Browser của data cube để thực hiện đuyệt và phân tích dữ liệu trong các Cube sau khi chúng đã được triển khai (deployed) và xử lý (processed).</p>

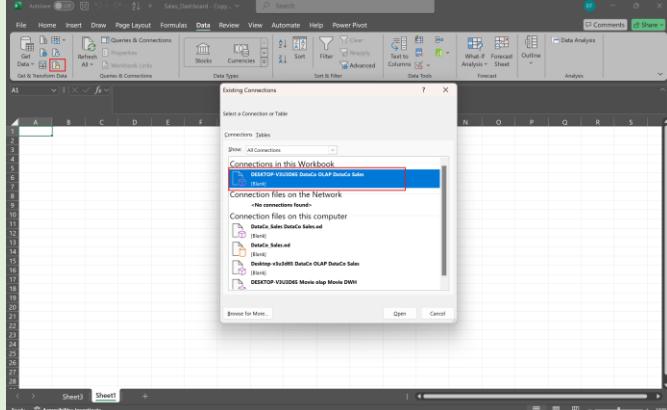
4. CÀI ĐẶT VÀ KIỂM THỬ

4.1. Xây dựng dashboard trên Excel pivot table

Bảng 14: Các bước xây dựng dashboard trên Pivot Table

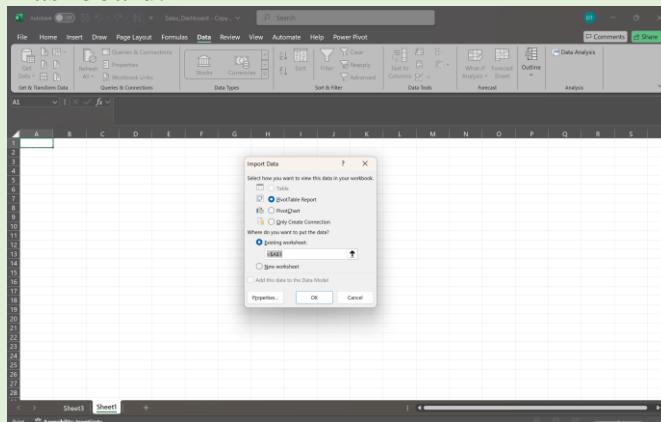
STT	Các bước thực hiện	Mục đích
1	<p>Kết nối đến khối dữ liệu:</p>  <p>Hình 23: Kết nối đến khối dữ liệu</p>	<p>Thực hiện kết nối đến dữ liệu thông qua Analysis Services để kết nối đến khối dữ liệu đã tạo bằng SSAS.</p>
2	<p>Thiết lập kết nối:</p>  <p>Hình 24: Thiết lập kết nối</p>	<p>Cấu hình cho kết nối đến cube dữ liệu thông qua server name để kết nối trực tiếp đến khối dữ liệu.</p>
3	<p>Chọn cube dữ liệu đã tạo:</p>	<p>Sau khi kết nối đến server của Analysis Services, sẽ xuất hiện những data cube đã được tạo, chọn khối dữ liệu cần kết nối đến để thiết lập kết nối</p>

	 <p>Hình 25: Chọn khối dữ liệu</p>	
4	<p>Lưu kết nối:</p>  <p>Hình 26: Lưu kết nối đến dữ liệu</p>	<p>Lưu file kết nối dữ liệu để sau này Excel pivot table tự động kết nối đến cube dữ liệu đó để lấy dữ liệu</p>
5	<p>Kết nối đến connection vừa tạo để tạo các PivotTable Report / PivotChart:</p>	<p>Kết nối đến khối dữ liệu đã tạo trong quá trình SSAS để thực hiện tạo các report hoặc là dashboard để phân tích và trực quan hóa dữ liệu trên Excel Pivot Table.</p>



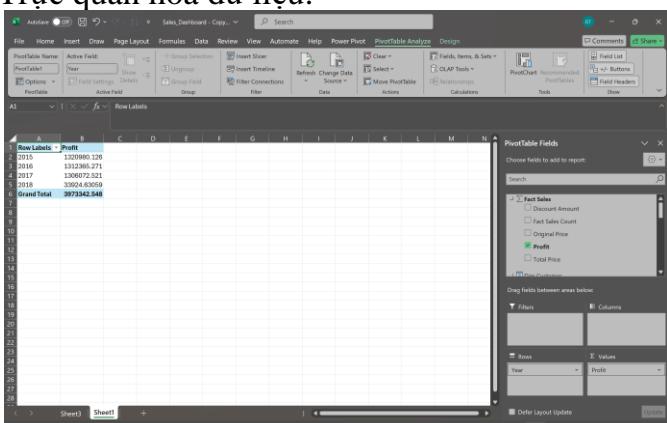
Hình 27: Lựa chọn kết nối đã lưu

Chọn tùy chọn theo nhu cầu để tạo Report / Dashboard:



Hình 28: Lựa chọn hình thức tổ chức dữ liệu

Trực quan hóa dữ liệu:

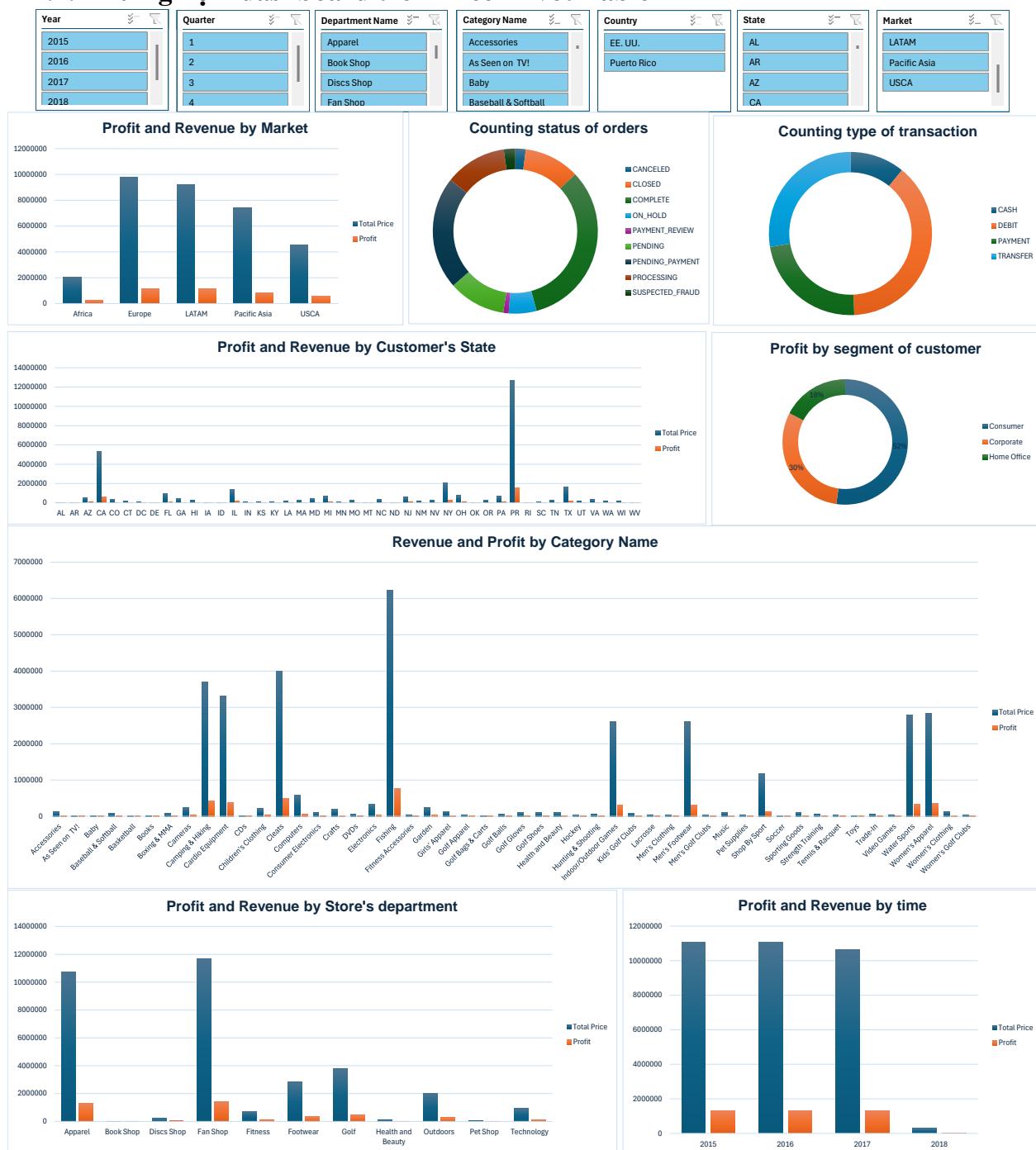


Hình 29: Giao diện thực hiện trực quan hóa trên Excel Pivot Table

6

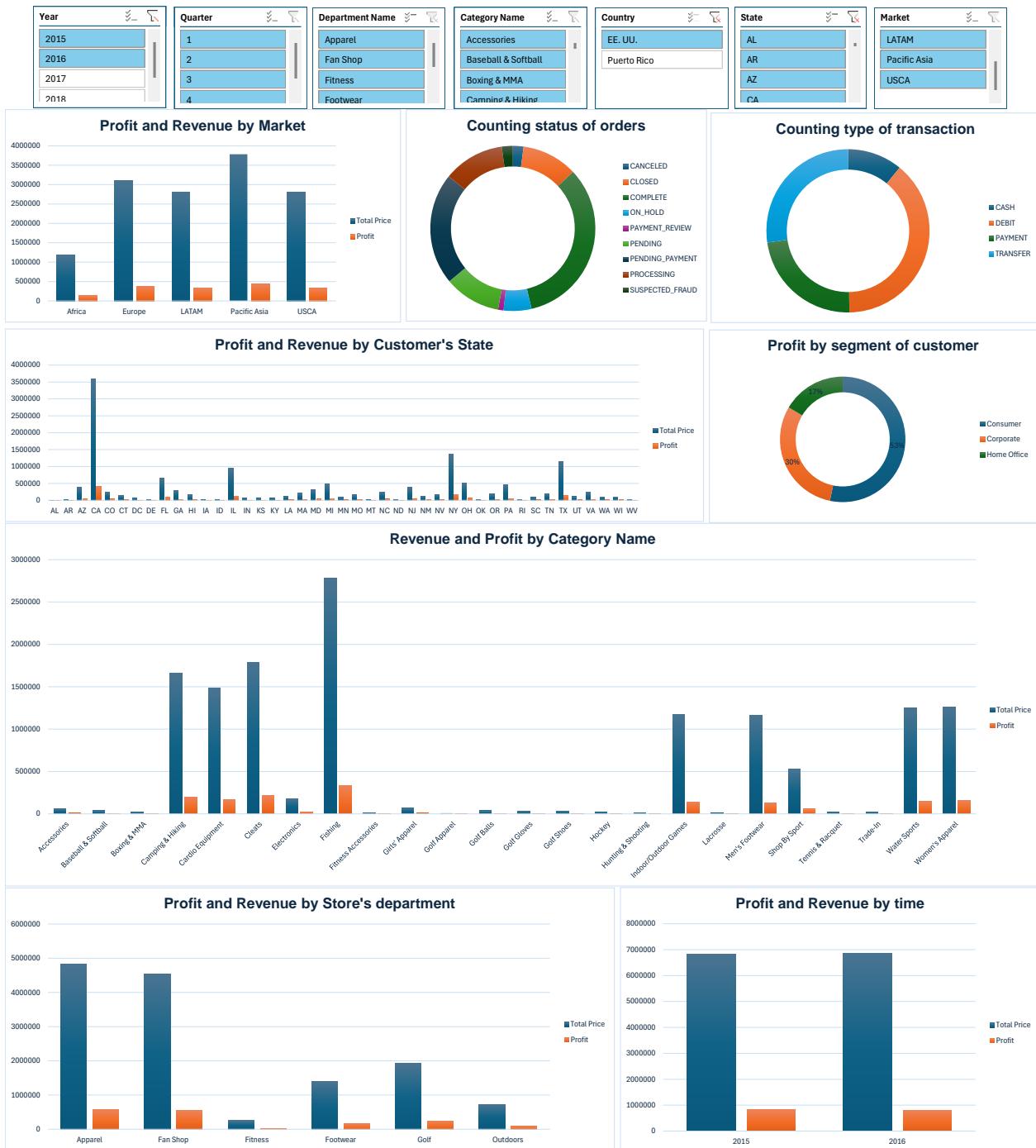
Sau khi hoàn thành kết nối đến khôi dữ liệu, có thể tự do xây dựng các bảng, biểu đồ để phục vụ cho việc phân tích và trực quan hóa dữ liệu thông qua Field List bên phải

4.2. Thử nghiệm dashboard trên Excel Pivot Table

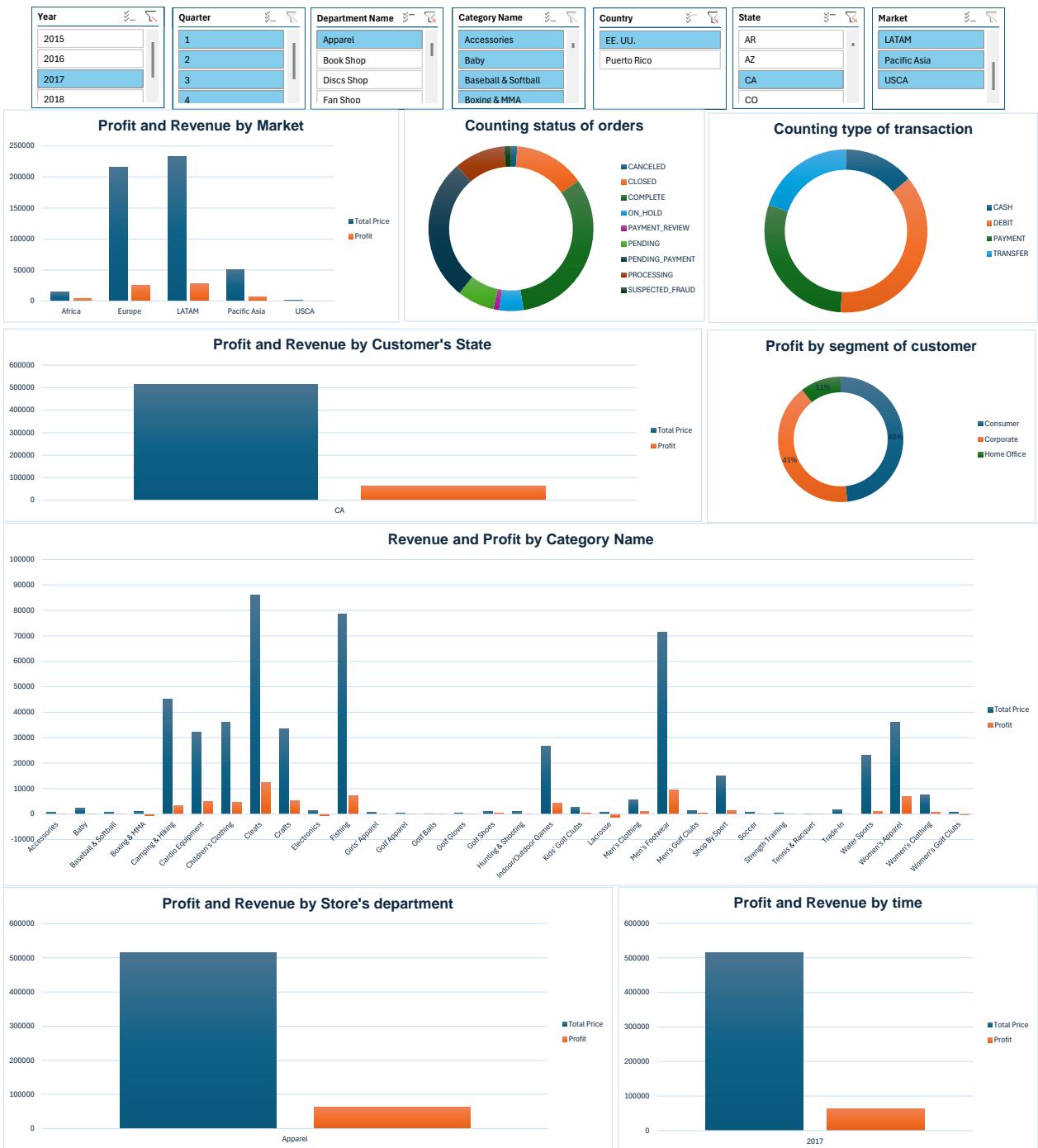


Hình 30: Dashboard tạo bằng Pivot Table

Một số kết quả sau khi sử dụng bộ lọc



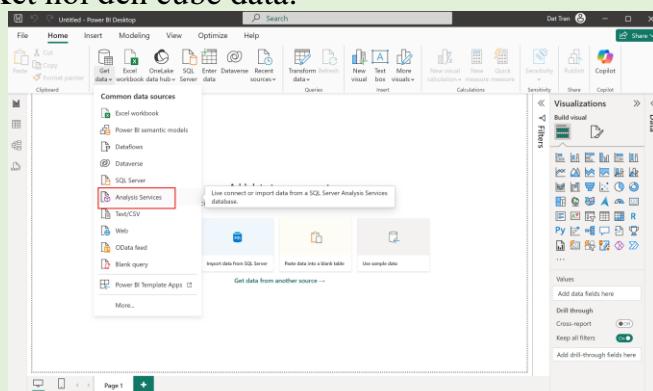
Hình 31: Tình hình kinh doanh ở năm 2015, 2016 tại Mỹ

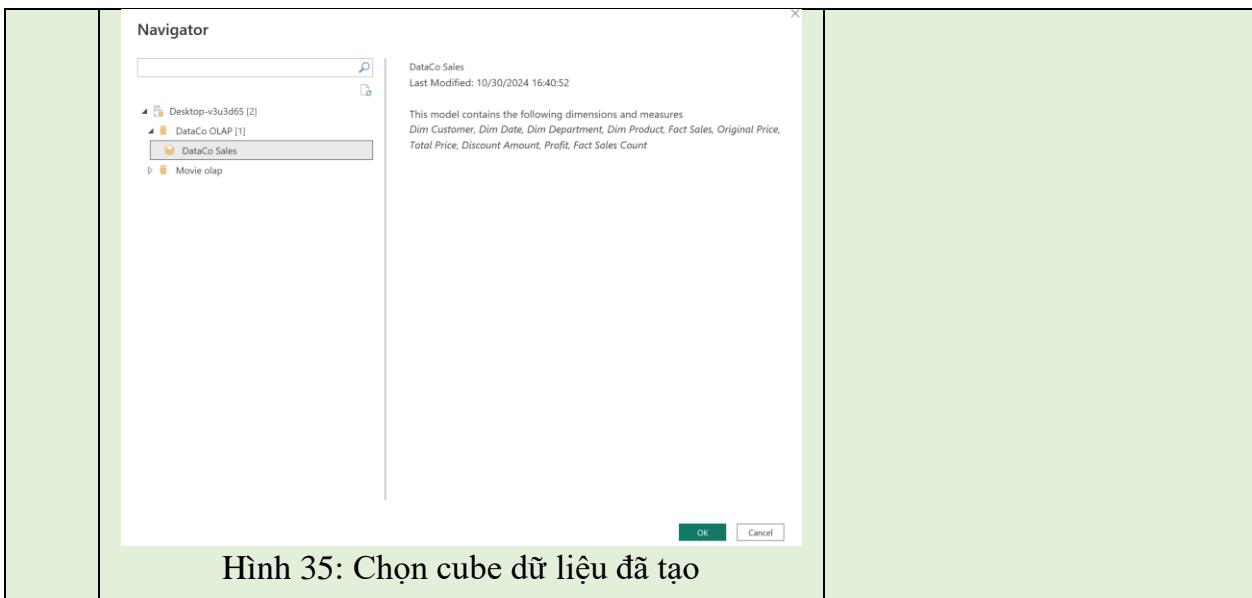


Hình 32: Tình hình kinh doanh ở năm 2017 của phòng ban Apparel tại California

4.3. Xây dựng dashboard trên Power BI

Bảng 15: Các bước xây dựng dashboard trên Power BI

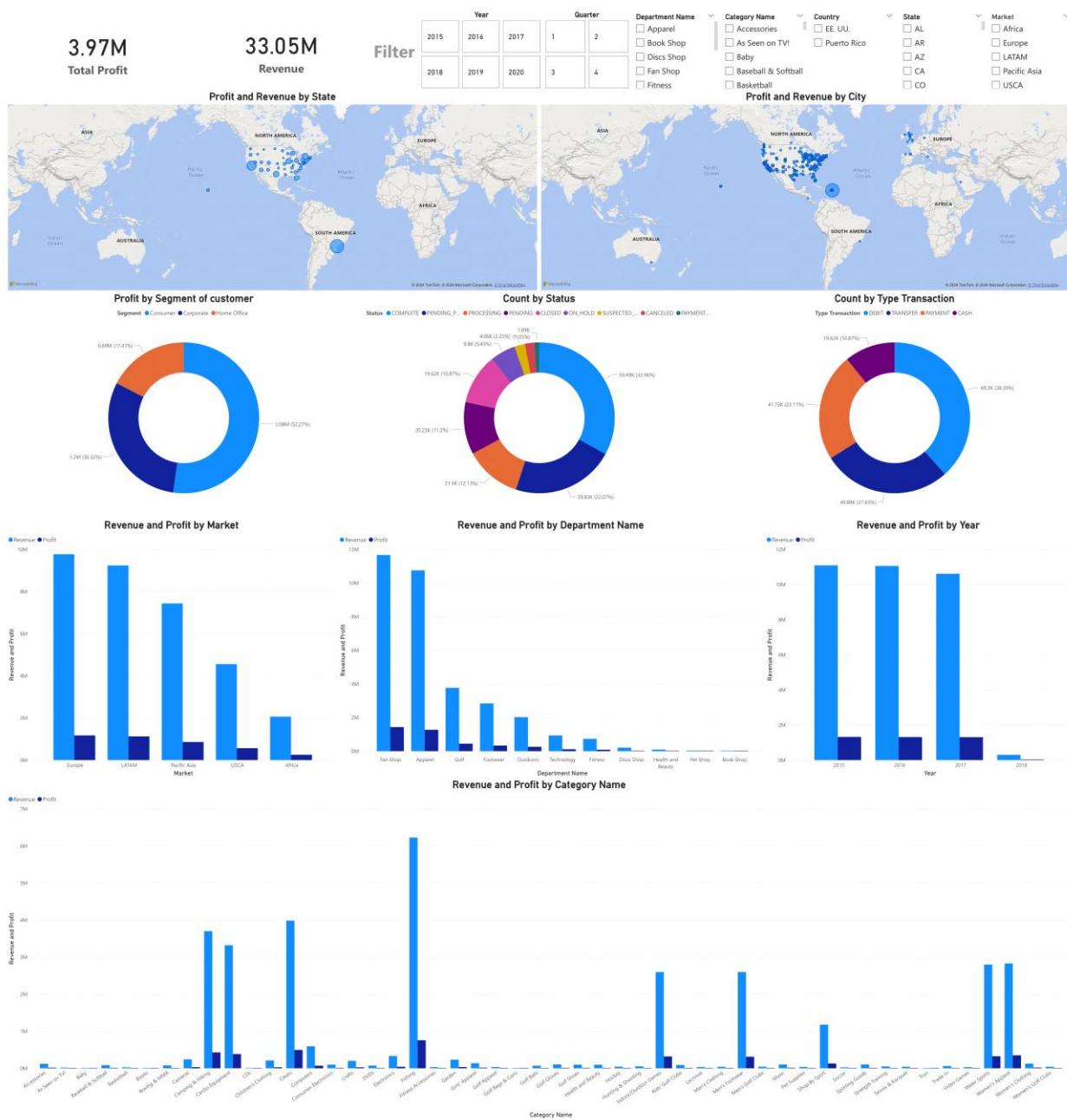
STT	Các bước thực hiện	Mục đích
1	<p>Kết nối đến cube data:</p>  <p>Hình 33: Kết nối đến cube data</p>	<p>Thực hiện kết nối đến dữ liệu thông qua Analysis Services để kết nối đến cube data đã tạo bằng SSAS.</p>
2	<p>Thiết lập kết nối:</p>  <p>Hình 34: Thiết lập kết nối</p>	<p>Cấu hình cho kết nối đến cube dữ liệu thông qua server name để kết nối trực tiếp đến khối dữ liệu.</p>
3	<p>Chọn cube dữ liệu đã tạo:</p>	<p>Sau khi kết nối đến server của Analysis Services, sẽ xuất hiện những data cube đã được tạo, chọn khối dữ liệu cần kết nối đến để thiết lập kết nối</p>



Hình 35: Chọn cube dữ liệu đã tạo

4	<p>Trực quan hóa dữ liệu:</p> <p>Hình 36: Giao diện thực hiện trực quan hóa trên Power BI</p>	<p>Sau khi hoàn thành kết nối đến khối dữ liệu, có thể tự do xây dựng các bảng, biểu đồ để phục vụ cho việc phân tích và trực quan hóa dữ liệu thông qua tab Data bên phải</p>
---	--	--

4.1. Thử nghiệm dashboard trên Power BI

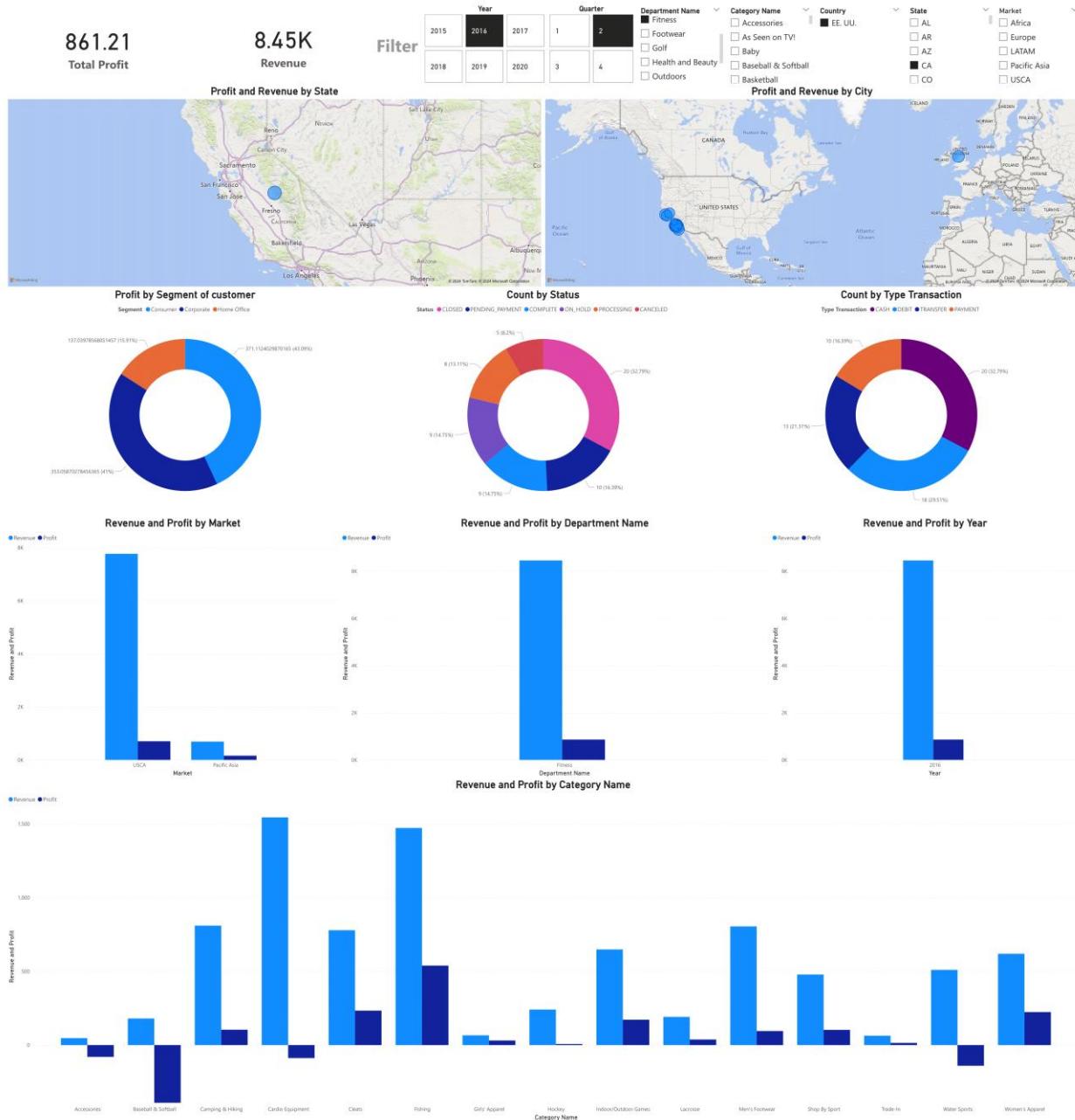


Hình 37: Dashboard tạo bằng Power BI

Một số kết quả sau khi sử dụng bộ lọc



Hình 38: Tình hình kinh doanh ở Puerto Rico vào năm 2016 tại thị trường Europe.



Hình 39: Tình hình kinh doanh của phòng ban Fitness ở tiểu bang California vào quý 2, năm 2016

5. KẾT LUẬN

5.1. Các kết quả đạt được

Về mặt kiến thức

- Hiểu sâu về kiến thức của Data Warehouse, từ thiết kế mô hình dữ liệu đến triển khai hệ thống thực tế.
- Nắm vững quy trình ETL (Trích xuất, Chuyển đổi, và Tải dữ liệu) thông qua công cụ SSIS, và cách xây dựng khói dữ liệu đa chiều bằng công cụ SSAS
- Làm quen và sử dụng thành thạo các công cụ giúp phân tích và trực quan hóa dữ liệu như Power BI và Excel Pivot.

Về mặt kỹ năng

- Rèn luyện được khả năng tiếp thu và xử lý thông tin đầu vào một cách có hệ thống
- Phát triển tư duy phân tích dữ liệu và đặt câu hỏi phù hợp để giải quyết vấn đề
- Nâng cao kỹ năng thực hành với các công cụ hỗ trợ trong lĩnh vực Data Warehouse và BI.
- Tăng cường khả năng vận dụng công nghệ vào thực tiễn.

Về mặt sản phẩm

- Xây dựng kho dữ liệu hoàn chỉnh, từ giai đoạn ETL đến tổ chức dữ liệu trong Data Warehouse.
- Tạo các báo cáo trực quan hóa để trả lời các câu hỏi kinh doanh thông qua Power BI và Excel Pivot.
- Đáp ứng nhu cầu phân tích và hỗ trợ ra quyết định của doanh nghiệp thông qua sản phẩm trực quan hóa dữ liệu.

5.2. Ưu điểm, khó khăn và hạn chế của đề tài

Ưu điểm:

- Đề tài giúp hiểu và áp dụng kiến thức về kho dữ liệu và trực quan hóa dữ liệu.
- Nâng cao kỹ năng thực hành với các công cụ như SSIS và Power BI, hỗ trợ quy trình ETL và phân tích dữ liệu.

- Phát triển tư duy phân tích dữ liệu và khả năng thiết kế, xây dựng hệ thống kho dữ liệu hiệu quả.
- Tạo ra các dashboard trực quan, giúp chuyển đổi dữ liệu phức tạp thành các biểu đồ và báo cáo dễ hiểu, phục vụ tốt cho việc trình bày thông tin.

Khó khăn và hạn chế:

- Khả năng chia sẻ dashboard bị hạn chế bởi các yêu cầu về môi trường phần mềm hoặc tài khoản cụ thể.
- Chưa có giải pháp tối ưu để triển khai hệ thống báo cáo trên nền tảng trực tuyến hoặc các hệ thống phân phối.
- Cân nhắc tích hợp thêm các mô hình dự đoán hoặc trí tuệ nhân tạo để khai thác thêm giá trị từ dữ liệu

5.3. Định hướng phát triển

- Khắc phục các khó khăn và hạn chế để hoàn thiện hơn, giúp các báo cáo thuận tiện hơn và tăng tính thuyết phục
- Đảm bảo xử lý hiệu quả khói lượng dữ liệu lớn hơn, cải thiện hiệu suất khi tích hợp dữ liệu từ nhiều nguồn khác nhau.
- Phát triển dashboard trên nền tảng web hoặc công cụ trực tuyến, khắc phục hạn chế trong việc chia sẻ và triển khai báo cáo.
- Sử dụng trí tuệ nhân tạo và học máy để phân tích dữ liệu chuyên sâu, cung cấp dự báo chính xác hơn về doanh thu, xu hướng bán hàng, và hành vi khách hàng.

TÀI LIỆU THAM KHẢO

1. Nguyen, M. (2024). Tìm hiểu về quy trình ETL (Extract, Transform, Load) và cách chúng được áp dụng trong thực tế. Retrieved from <https://viblo.asia/p/tim-hieu-ve-quy-trinh-etl-extract-transform-load-va-cach-chung-duoc-ap-dung-trong-thuc-te-38X4EPYXVN2>
2. Lucie. (2022). Data Warehouse Là Gì? Tổng Quan Về Kho Dữ Liệu. Retrieved from <https://topdev.vn/blog/data-warehouse-la-gi-tong-quan-ve-kho-du-lieu/>
3. Chugugrace. (n.d.). SQL Server Integration Services - SQL Server Integration Services (SSIS). Retrieved from <https://learn.microsoft.com/en-us/sql/integration-services/sql-server-integration-services?view=sql-server-ver16>
4. Kfollis. (n.d.). Analysis Services core documentation. Retrieved from <https://learn.microsoft.com/en-us/analysis-services/?view=asallproducts-allversions>