TRƯỜNG ĐẠI HỌC NGÂN HÀNG TP. HỒ CHÍ MINH

KHOA HỆ THỐNG THÔNG TIN QUẢN LÝ

----------OOO----------

BÀI BÁO CÁO

MÔN KHO DỮ LIỆU & HỆ HỖ TRỢ RA QUYẾT ĐỊNH

Đề tài:

BÀI BÁO CÁO XÂY DỰNG HỆ THỐNG OLAP CHO BÀI TOÁN KINH DOANH LĨNH VỰC BÁN HÀNG (SALE)

Lớp học phần : DAT701_241_1_D01

Giảng viên : Bùi Hữu Đông

Nhóm thực hiện: 06

MỤC LỤC

BẢNG PHÂN CÔNG NHIỆM VỤ				
STT	**Họ và tên**	**MSSV**	**Nhiệm vụ**	**Mức độ hoàn thành**
1	Đoàn Thị Diệu Linh	030238220106	Thiết kế kho dữ liệu	100%
2	Lê Hồng Phương Quí	030238220203		100%
3	Nguyễn Phương Nhi	030238220166	Thực hiện ETL dữ liệu	100%
4	Nguyễn Phan Hoàng Phúc	030238220192		100%
5	Nguyễn Thị Nhật Liên	030238220105	Thực hiện hệ thống OLAP	100%
6	Lê Tấn Cường	030238220016		100%

GIỚI THIỆU

Trong thời đại số hóa hiện nay, ngành bán lẻ không ngừng phát triển và trở nên cực kỳ cạnh tranh. Các doanh nghiệp bán lẻ phải không ngừng đổi mới và thích ứng để không những tồn tại mà còn phát triển trong một môi trường thay đổi nhanh chóng. Điều này đòi hỏi một sự hiểu biết sâu sắc về hành vi mua sắm của khách hàng, các xu hướng sản phẩm mới nhất, và hiệu quả của các chiến lược bán hàng. Việc lựa chọn đề tài về bán hàng là hoàn toàn phù hợp bởi bán hàng không chỉ là yếu tố then chốt quyết định doanh thu mà còn là nền tảng vững chắc cho sự phát triển bền vững của doanh nghiệp.

Đề tài về **Sales** được chọn vì bán hàng là yếu tố then chốt quyết định doanh thu và sự phát triển bền vững của doanh nghiệp trong bối cảnh thị trường cạnh tranh gay gắt. Với sự thay đổi nhanh chóng của công nghệ tiên tiến, các quy trình và hành vi mua sắm của khách hàng cũng biến đổi liên tục, đòi hỏi các doanh nghiệp phải không ngừng tối ưu hóa chiến lược bán hàng của mình. Điều này không chỉ bao gồm việc phát triển đội ngũ chuyên nghiệp mà còn phải tận dụng triệt để công nghệ để duy trì và mở rộng thị phần. **Sales** đóng vai trò trung tâm trong việc nắm bắt nhu cầu và xu hướng của thị trường, từ đó giúp doanh nghiệp điều chỉnh sản phẩm và dịch vụ sao cho phù hợp nhất, tối ưu hóa sự hài lòng của khách hàng và cuối cùng là đạt được mục tiêu tăng trưởng doanh thu bền vững.

1 Giới thiệu công cụ

1.1 SQL và SQL Server Management Studio (SSMS):

SQL (Structured Query Language) đóng vai trò quan trọng trong việc quản lý, truy vấn, và phân tích dữ liệu trong môi trường Data Warehouse. Ngôn ngữ này hỗ trợ thực hiện các tác vụ phức tạp như truy vấn tổng hợp, sử dụng hàm cửa sổ, và tối ưu hóa dữ liệu để tạo báo cáo phù hợp với mục tiêu kinh doanh.

Dựa trên tầm quan trọng của SQL trong việc quản lý và phân tích dữ liệu trong môi trường Data Warehouse, SQL Server Management Studio (SSMS) là công cụ tích hợp mạnh mẽ để quản lý SQL Server và tối ưu hóa hệ thống Data Warehouse. SSMS hỗ trợ đầy đủ các tính năng cần thiết cho việc:

- Truy xuất và phân tích dữ liệu lớn với hiệu suất cao.
- Thực hiện các tác vụ như ghép nối dữ liệu, lập chỉ mục, và phân vùng bảng.
- Quản lý cấu trúc cơ sở dữ liệu lớn một cách trực quan.

Ngoài ra, SSMS tích hợp tốt với các công cụ phát triển quy trình ETL, cho phép xử lý dữ liệu từ nhiều nguồn khác nhau, chuẩn hóa và tổng hợp dữ liệu trước khi đưa vào Data Warehouse. Tính năng tự động hóa trong SSMS hỗ trợ giảm thiểu lỗi thủ công, đồng thời đảm bảo tính nhất quán và chính xác trong hệ thống.

SSMS cũng cung cấp các công cụ giám sát hiệu suất và tối ưu hóa truy vấn, đảm bảo hệ thống Data Warehouse hoạt động ổn định ngay cả khi xử lý lượng dữ liệu lớn.

1.2 SQL Server Integration Services (SSIS):

SQL Server Integration Services (SSIS) là công cụ chuyên biệt hỗ trợ các quy trình ETL (Extract, Transform, Load) trong Data Warehouse. SSIS cung cấp một nền tảng mạnh mẽ để:

- **Kết nối và di chuyển dữ liệu:** Thu thập dữ liệu từ nhiều nguồn như cơ sở dữ liệu quan hệ, tệp văn bản, dịch vụ đám mây.
- **Chuyển đổi và chuẩn hóa dữ liệu:** Thực hiện các tác vụ như tính toán, phân tách, hợp nhất, và làm sạch dữ liệu.
- **Tích hợp dễ dàng:** Hỗ trợ kết nối với các API, dịch vụ web, và hệ thống lưu trữ đám mây, đảm bảo tính linh hoạt trong việc tích hợp dữ liệu.

SSIS nổi bật với giao diện thiết kế kéo-thả thân thiện, giúp người dùng dễ dàng tạo các workflow xử lý dữ liệu mà không cần viết mã phức tạp. Các tính năng nâng cao như lập lịch tự động, quản lý lỗi, và tối ưu hóa hiệu suất giúp đảm bảo quy trình ETL diễn ra chính xác và đáng tin cậy.

Ngoài ra, SSIS tích hợp mạnh mẽ với các công cụ như SSMS và SSRS (SQL Server Reporting Services), tạo ra một hệ sinh thái đồng bộ để quản lý và phân tích dữ liệu hiệu quả.

1.3 SQL Server Analysis Services (SSAS):

SQL Server Analysis Services (SSAS) là công cụ mạnh mẽ để xây dựng và triển khai các mô hình phân tích dữ liệu trong môi trường Data Warehouse. SSAS hỗ trợ:

- **Phân tích đa chiều:** Xây dựng các khối dữ liệu (cubes) dựa trên mô hình OLAP (Online Analytical Processing), cho phép phân tích dữ liệu theo nhiều chiều như thời gian, địa lý, sản phẩm, khách hàng, v.v.
- **Tối ưu hóa truy vấn:** Xử lý và trả kết quả phân tích phức tạp với tốc độ cao, đảm bảo dữ liệu luôn được cập nhật và chính xác.
- **Dự báo và ra quyết định:** Hỗ trợ các phân tích nâng cao, tìm kiếm mối quan hệ giữa các yếu tố dữ liệu, và dự báo xu hướng để phục vụ mục tiêu kinh doanh.

SSAS tích hợp tốt với các công cụ trực quan hóa dữ liệu như Power BI và Excel, giúp doanh nghiệp dễ dàng tạo các báo cáo và biểu đồ trực quan. Ngoài ra, SSAS hỗ trợ các mô hình dữ liệu phức tạp như mô hình phân cấp hoặc mối quan hệ, cho phép tạo ra các báo cáo chi tiết và phân tích hiệu quả theo nhiều cấp độ.

2 Quy trình thực hiện

Quy trình thực hiện phân tích kinh doanh gồm các bước:

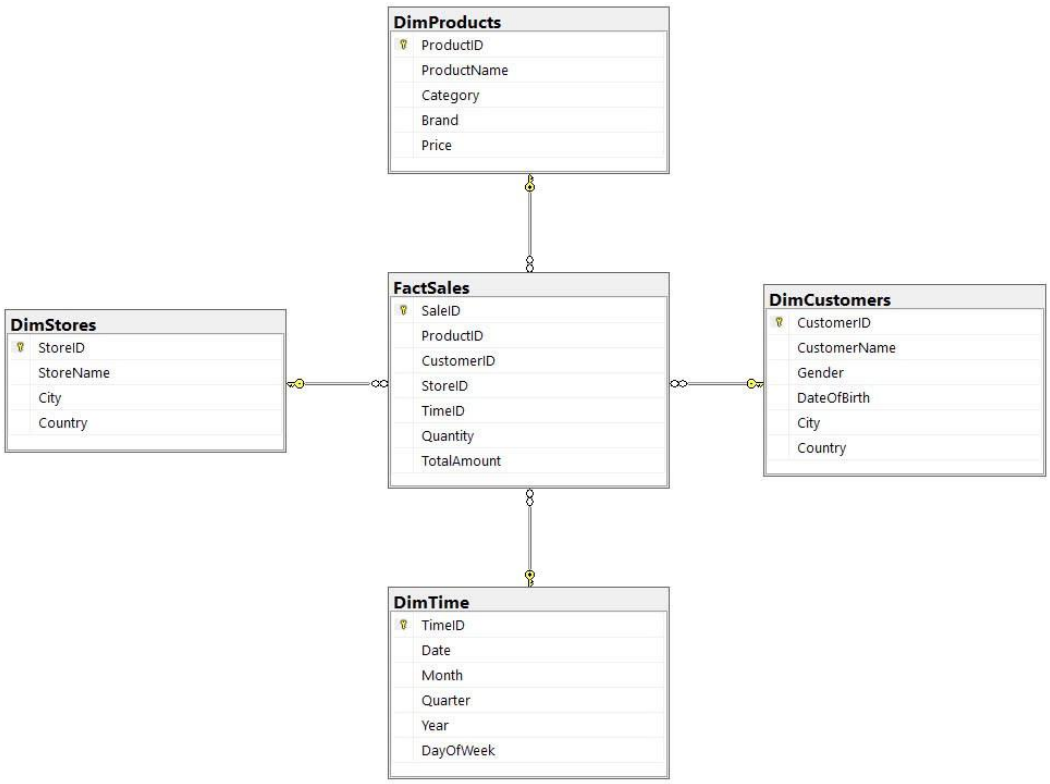
- Thu thập dữ liệu từ các nguồn khác nhau: ở đây là CSV file, nguồn đã được thu thập sẵn.
- ETL: quá trình trích xuất - làm sạch - tải vào kho dữ liệu.
- Data Warehouse: thực hiện lưu trữ dữ liệu vào kho dữ liệu.
- OLAP: dùng dữ liệu để phân tích và dự đoán, ở đây là phân tích và dự đoán kinh doanh

2.1 Thiết kế kho dữ liệu

Để có thể hiểu được ý nghĩa và các mối quan hệ giữa dữ liệu với nhau, chúng ta cần thiết kế lược đồ quan hệ, ở đây lược đồ quan hệ (ERD) được chọn là lược đồ hình sao,

lý do thì cũng đơn giản, dữ liệu không phức tạp và không chồng chéo nhau, lựa chọn lược đồ hình sao là hợp lý.

Lược đồ **Star Schema** (Ngôi Sao), là mô hình phổ biến trong kho dữ liệu ("Data Warehouse"). Mô hình này hỗ trợ hiệu quả cho các hệ thống phân tích dữ liệu và truy vấn nhanh chóng. Trong lược đồ, **Fact Table** ("Bảng Sự Kiện") chứa các dữ liệu giao dịch trung tâm, trong khi đó các **Dimension Tables** ("Bảng Chiều") cung cấp ngữ cảnh bổ trợ chi tiết.



Customers: Bảng chiều khách hàng (DimCustomers) chứa thông tin chi tiết về khách hàng, bao gồm: CustomerID (khóa chính, xác định khách hàng duy nhất), CustomerName (tên khách hàng), Gender(giới tính), DateOfBirth (ngày sinh), City (thành phố cư trú) và Country(quốc gia). Thông qua bảng này, doanh nghiệp có thể theo dõi và phân tích đặc điểm của khách hàng để cải thiện dịch vụ và xây dựng mối quan hệ lâu dài.

Products: Bảng chiều sản phẩm (DimProducts) chứa thông tin về sản phẩm, bao gồm: ProductID (khóa chính, định danh mỗi sản phẩm duy nhất), ProductName (tên sản phẩm), Category (nhóm danh mục sản phẩm như điện tử, quần áo), Brand (thương hiệu) và Price (mức giá). Bảng này giúp doanh nghiệp quản lý và phân tích danh mục sản phẩm hiệu quả, từ đó tối ưu hóa chiến lược kinh doanh.

Time: Thực thể thời gian giúp định danh và theo dõi thời điểm xảy ra các sự kiện trong hệ thống. Bảng chiều thời gian (DimTime) chứa thông tin về thời gian giao dịch, bao gồm: TimeIDl (khóa chính), Date (ngày giao dịch), Month(tháng giao dịch), Quarter (quý giao dịch), Year (năm giao dịch) và DayOfWeek (ngày trong tuần). Bảng này giúp doanh nghiệp phân tích và theo dõi các xu hướng giao dịch theo thời gian, từ đó đưa ra quyết định kinh doanh chính xác hơn.

Stores là một bảng chiều quan trọng trong hệ thống kho dữ liệu, được sử dụng để quản lý thông tin về các cửa hàng. Bảng này cung cấp ngữ cảnh cho các giao dịch bán hàng, cho phép doanh nghiệp phân tích hiệu suất của từng cửa hàng và đưa ra các chiến lược kinh doanh phù hợp theo từng địa điểm. Bảng này bao gồm các trường như StoreID, là khóa chính giúp định danh mỗi cửa hàng duy nhất, StoreName để ghi nhận tên của cửa hàng, cùng với City và Country chỉ rõ vị trí địa lý của cửa hàng. Thông qua bảng DimStores, doanh nghiệp có thể theo dõi và phân tích hiệu suất của từng cửa hàng, từ đó phát triển các chiến lược kinh doanh phù hợp theo từng địa điểm. Việc quản lý thông tin này không chỉ hỗ trợ tối ưu hóa hoạt động bán hàng mà còn nâng cao trải nghiệm khách hàng tại mỗi cửa hàng.

2.1.1 Fact Table - FactSales

Bảng trung tâm của mô hình là **FactSales**, chứa các thông tin giao dịch, bao gồm:

- **SaleID**: Khóa chính (“Primary Key”) duy nhất để xác định từng giao dịch.
- **ProductID**: Khóa ngoại (“Foreign Key”) tham chiếu đến sản phẩm trong bảng **DimProducts**.
- **CustomerID**: Tham chiếu đến khách hàng trong bảng **DimCustomers**.
- **StoreID**: Tham chiếu đến cửa hàng trong bảng **DimStores**.
- **TimeID**: Tham chiếu đến thông tin thời gian trong bảng **DimTime**.
- **Quantity**: Số lượng sản phẩm bán trong giao dịch.
- **TotalAmount**: Tổng số tiền thu được từ giao dịch.

Bảng FactSales tích hợp dữ liệu giao dịch từ nhiều ngốc nhìn khác nhau, cho phép phân tích linh hoạt.

2.1.2 Dimension Tables

Mô hình bao gồm 4 bảng chiều chính, cung cấp thông tin chi tiết liên quan tới các giao dịch trong bảng FactSales:

1. **DimProducts** (Bảng Chiều Sản Phẩm):
    - **ProductID**: Khóa chính, định danh mỗi sản phẩm duy nhất.
    - **ProductName**: Tên sản phẩm.

- **Category**: Nhóm danh mục của sản phẩm (ví dụ: điện tử, quần áo).
- **Brand**: Thương hiệu.
- **Price**: Mức giá của sản phẩm.

Ứng dụng: Hỗ trợ báo cáo doanh thu theo sản phẩm, danh mục, hoặc thương hiệu.

2. **DimCustomers** (Bảng Chiều Khách Hàng):
    - **CustomerID**: Khóa chính, xác định khách hàng duy nhất.
    - **CustomerName**: Tên khách hàng.
    - **Gender**: Giới tính.
    - **DateOfBirth**: Ngày sinh của khách hàng.
    - **City**: Thành phố khách hàng sinh sống.
    - **Country**: Quốc gia của khách hàng.

    Ứng dụng: Phân tích hành vi một tập khách hàng dựa trên giới tính, khu vực, hoặc độ tuổi.

3. **DimStores** (Bảng Chiều Cửa Hàng):
    - **StoreID**: Khóa chính, định danh mỗi cửa hàng.
    - **StoreName**: Tên cửa hàng.
    - **City**: Thành phố.
    - **Country**: Quốc gia.

    Ứng dụng: Báo cáo doanh thu và hiệu quả bán hàng theo từng cửa hàng hoặc khu vực.

4. **DimTime** (Bảng Chiều Thời Gian):
    - **TimeID**: Khóa chính.
    - **Date**: Ngày giao dịch.
    - **Month**: Tháng giao dịch.
    - **Quarter**: Quý giao dịch.
    - **Year**: Năm giao dịch.
    - **DayOfWeek**: Ngày trong tuần.

    Ứng dụng: Báo cáo doanh thu theo tháng, quý, năm hoặc xu hướng trong các ngày trong tuần.

2.1.3 Đặc Điểm Của Mô Hình

- **Quan hệ:**

- Các bảng chiều ("Dimensions") được kết nối với Fact Sales bằng khóa ngoại, tạo thành một cấu trúc hình ngôi sao.
- Quan hệ một - nhiều ("One-to-Many") được duy trì giữa Fact Table và các Dimension Tables.

- **Hỗ trợ phân tích:**
  - Mô hình tối ưu cho truy vấn và tổng hợp dữ liệu nhanh chóng.
  - Linh hoạt khi phân tích theo nhiều ngôi ngữ cảnh (sản phẩm, khách hàng, cửa hàng hoặc thời gian).

2.2 Thực hiện ETL dữ liệu

Quá trình ETL (Extract, Transform, Load) là một quá trình quan trọng trong việc xây dựng kho dữ liệu, cho phép các tổ chức thu thập dữ liệu từ nhiều nguồn khác nhau, chuyển đổi dữ liệu đó để phù hợp với nhu cầu kinh doanh và tải nó vào một kho dữ liệu để phân tích. Dưới đây là chi tiết từng bước của quá trình ETL:

2.2.1 Trích Xuất (Extract):

Mục đích: Lấy dữ liệu từ các nguồn đầu vào khác nhau.
Thực hiện: Trong dự án, dữ liệu được trích xuất từ các tập tin dạng .csv, mỗi tập tin chứa thông tin cụ thể như thông tin sản phẩm, khách hàng, cửa hàng, thời gian và các giao dịch bán hàng.

2.2.2 Biến Đổi (Transform):

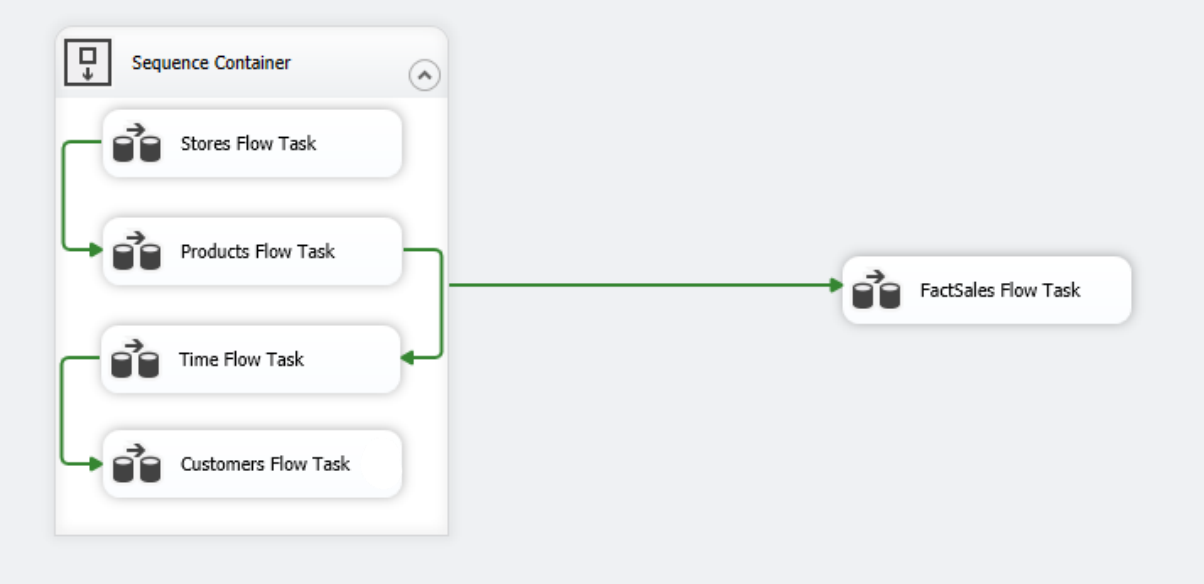
Mục đích: Điều chỉnh dữ liệu để đảm bảo tính nhất quán, chính xác, và phù hợp với mô hình dữ liệu của kho dữ liệu.
Thực hiện:

- **Phân Nhánh Điều Kiện**: Sử dụng để điều hướng hoặc lọc dữ liệu dựa trên các điều kiện nhất định, ví dụ, loại bỏ dữ liệu không đạt yêu cầu hoặc chia dữ liệu thành các luồng khác nhau cho các mục đích khác nhau.
- **Tra Cứu (Lookup)**: Dùng để tham chiếu và bổ sung dữ liệu từ các bảng khác, giúp tăng tính chính xác và đầy đủ cho dữ liệu (ví dụ, kết nối thông tin sản phẩm với thông tin kho hàng).
- **Chuyển Đổi Dữ Liệu và Cột Dẫn Xuất**: Các kiểu dữ liệu được chuyển đổi (ví dụ, đổi định dạng ngày tháng, tiêu chuẩn hóa các chuỗi văn bản), và tạo các cột mới từ dữ liệu hiện có để phục vụ phân tích (ví dụ, tính toán tổng tiền từ số lượng và đơn giá).

2.2.3 Tải (Load):

Mục đích: Lưu trữ dữ liệu đã được biến đổi vào trong kho dữ liệu.
Thực hiện: Dữ liệu được tải vào kho dữ liệu thông qua các kết nối OLE DB, điều này cho phép tải dữ liệu vào cơ sở dữ liệu quan hệ, nơi dữ liệu được tổ chức trong các bảng đã được định nghĩa sẵn trong quá trình thiết kế kho dữ liệu.

2.3 Chi tiết từng vấn đề



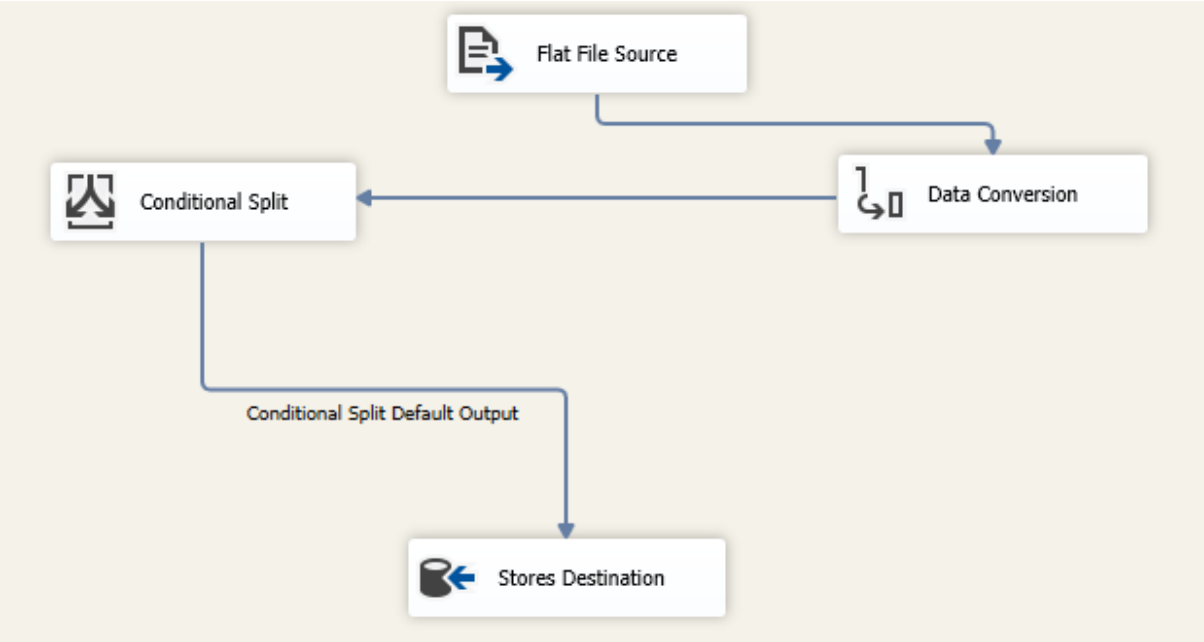
2.3.1 Nội dung của Sequence Container

1. **Store Flow Task**
   Nhiệm vụ này có thể liên quan đến việc xử lý và tải dữ liệu liên quan đến cửa hàng từ các nguồn dữ liệu vào kho dữ liệu.
2. **Product Flow Task**
   Đây là nhiệm vụ để xử lý thông tin về sản phẩm, có thể bao gồm trích xuất, biến đổi và tải dữ liệu sản phẩm.
3. **Customer Flow Task**
   Nhiệm vụ này thực hiện xử lý dữ liệu khách hàng, bao gồm các chi tiết như thông tin cá nhân, mua hàng và lịch sử giao dịch.
4. **Time Flow Task**
   Nhiệm vụ này xử lý dữ liệu liên quan đến thời gian, như ngày mua, ngày giao hàng, v.v.

2.3.2 Liên kết đến FactSales Flow Task

- Mũi tên xanh chỉ ra rằng sau khi hoàn thành tất cả các nhiệm vụ trên trong "Sequence Container," dữ liệu sẽ được đẩy vào nhiệm vụ "FactSale Flow Task."
- "FactSale Flow Task" có thể là nơi dữ liệu được tổng hợp và tải vào bảng "FactSales" trong kho dữ liệu, nơi dữ liệu từ các bảng chiều như sản phẩm, khách hàng, cửa hàng, và thời gian được sử dụng để tạo ra các báo cáo và phân tích doanh số bán hàng.

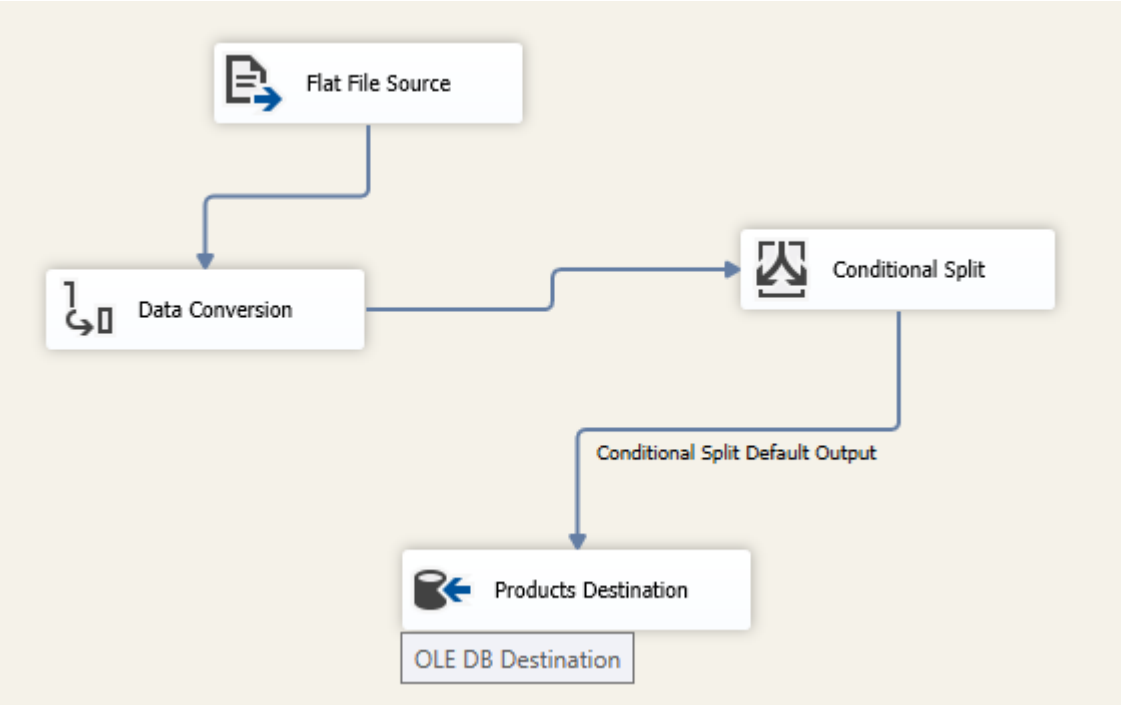
2.3.3 Stores Flow Task



1. **Flat File Source**:
    - **Mục đích**: Đây là điểm bắt đầu của quy trình, nơi dữ liệu được trích xuất từ một tập tin phẳng, ví dụ như CSV hoặc text file.
    - **Chức năng**: Đọc dữ liệu từ stores.csv và chuyển nó vào dòng dữ liệu SSIS để xử lý tiếp theo.
2. **Data Conversion**:
    - **Mục đích**: Chuyển đổi kiểu dữ liệu của các cột trong dòng dữ liệu.
    - **Chức năng**: Dữ liệu khi được từ module “Flat File Source” mặc định sẽ là text, và sau khi chạy qua module thì chuyển đổi kiểu dữ liệu theo nhu cầu.
3. **Conditional Split**:
    - **Mục đích**: Phân tách dòng dữ liệu dựa trên các điều kiện bạn định nghĩa.
    - **Chức năng**: Phân tách dữ liệu NULL của khoá chính.
4. **Stores Destination**:

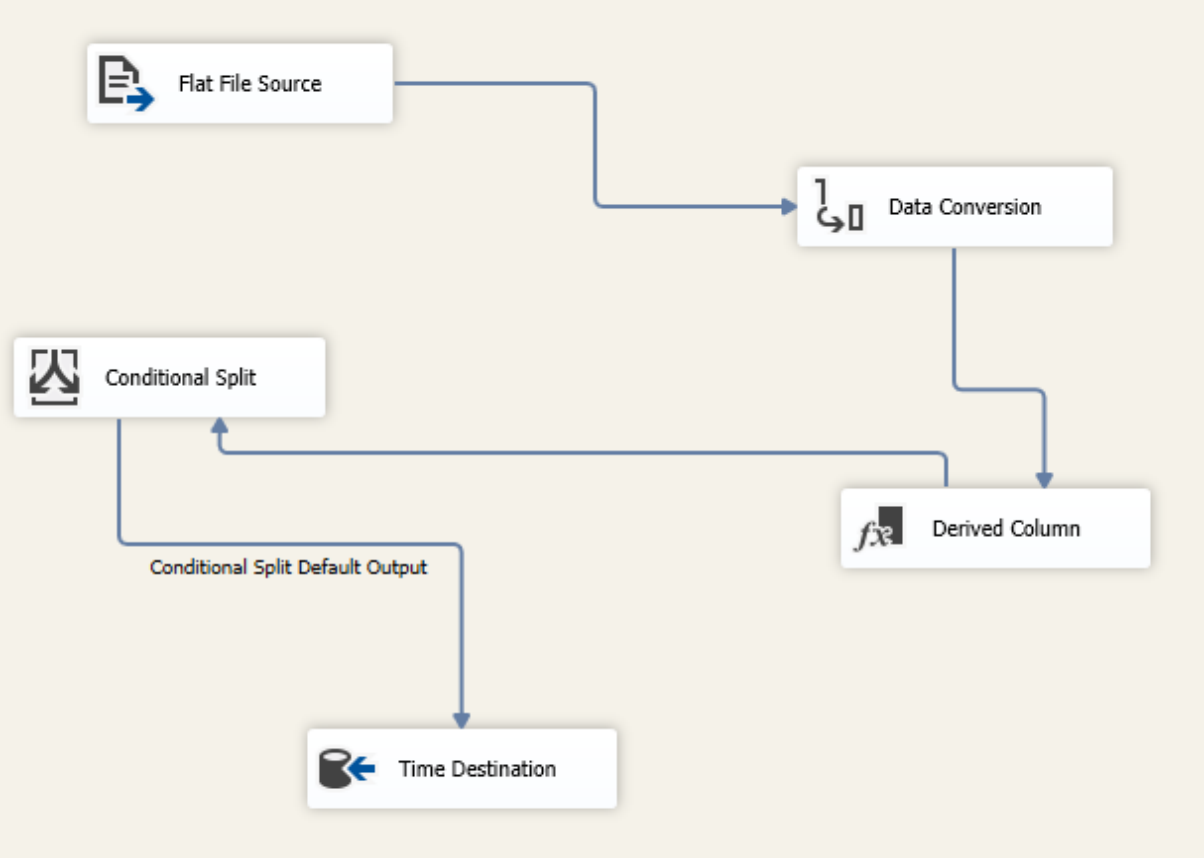
- **Mục đích**: Là điểm đến cuối cùng trong luồng này của quy trình ETL, nơi dữ liệu được ghi vào.
- **Chức năng**: Tải dữ liệu vào kho.

2.3.4 Products Flow Task



1. **Flat File Source**:
   - **Mục đích**: Đây là nơi dữ liệu được trích xuất từ một tập tin dạng phẳng, thường là CSV hoặc text file.
   - **Chức năng**: Đọc dữ liệu từ Products.csv và đưa vào luồng dữ liệu của SSIS để tiếp tục xử lý.
2. **Data Conversion**:
   - **Mục đích**: Chuyển đổi kiểu dữ liệu của các cột trong dòng dữ liệu.
   - **Chức năng**: Dữ liệu khi được từ module "Flat File Source" mặc định sẽ là text, và sau khi chạy qua module thì chuyển đổi kiểu dữ liệu theo nhu cầu.
3. **Conditional Split**:
   - **Mục đích**: Phân loại dữ liệu dựa trên một hoặc nhiều điều kiện.
   - **Chức năng**: Phân tách dữ liệu NULL của khoá chính.
4. **Products Destination**:
   - **Mục đích**: Là điểm đến cuối cùng trong luồng này của quy trình ETL, nơi dữ liệu được ghi vào.
   - **Chức năng**: Tải dữ liệu vào kho.

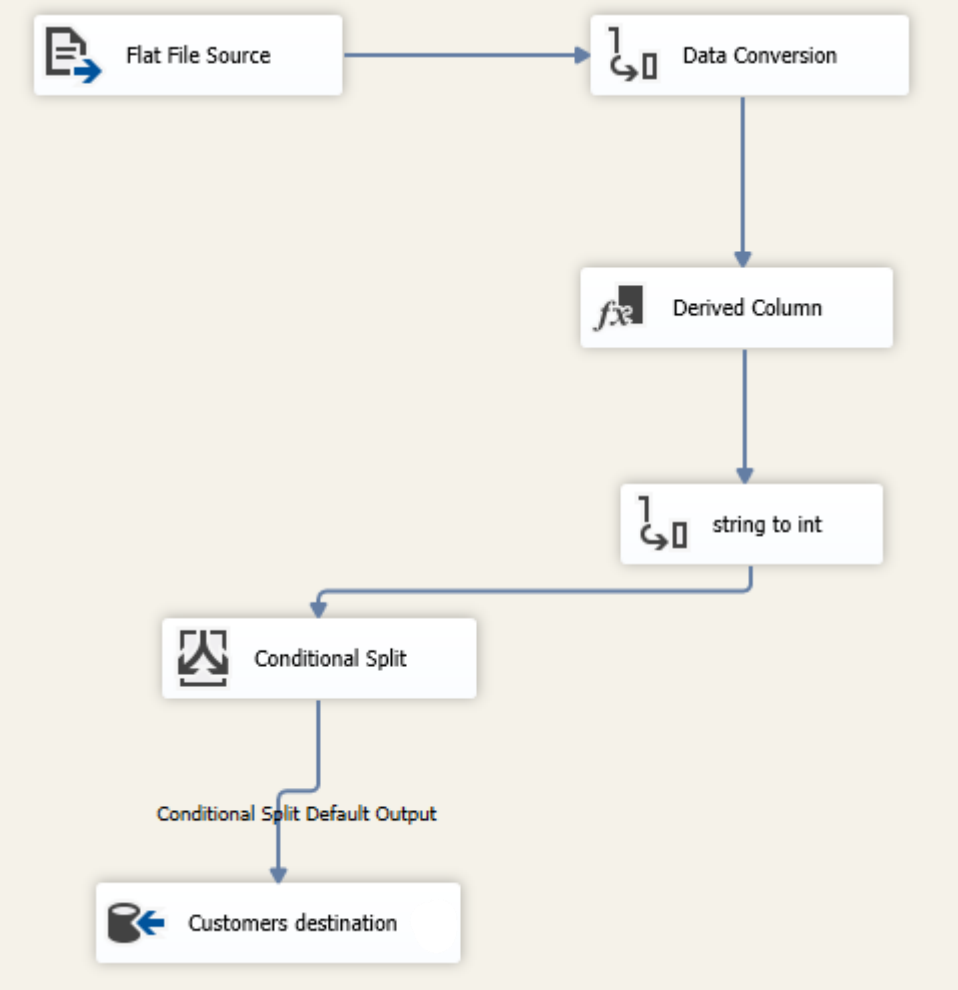
2.3.5 Time Flow Task



1. **Flat File Source**:
    - **Mục đích**: Đây là điểm bắt đầu của quy trình ETL, nơi dữ liệu được trích xuất từ một tập tin dạng phẳng
    - **Chức năng**: Đọc dữ liệu từ tập tin time.csv và đưa vào dòng dữ liệu SSIS để xử lý.
2. **Data Conversion**:
    - **Mục đích**: Chuyển đổi kiểu dữ liệu của các cột trong dòng dữ liệu.
    - **Chức năng**: Dữ liệu khi được từ module "Flat File Source" mặc định sẽ là text, và sau khi chạy qua module thì chuyển đổi kiểu dữ liệu theo nhu cầu.
3. **Derived Column**:
    - **Mục đích**: Thêm cột mới hoặc thay đổi cột hiện tại trong dòng dữ liệu.
    - **Chức năng**: thay thế cột month đã bị lỗi ở trong dữ liệu.
4. **Conditional Split**:
    - **Mục đích**: Phân loại dữ liệu dựa trên một hoặc nhiều điều kiện.
    - **Chức năng**: Phân tách dữ liệu NULL của khoá chính.
5. **Time Destination:**
    - **Mục đích**: Là điểm đến cuối cùng trong luồng này của quy trình ETL, nơi dữ liệu được ghi vào.

- **Chức năng**: Tải dữ liệu vào kho.

2.3.6 Customers Flow Task

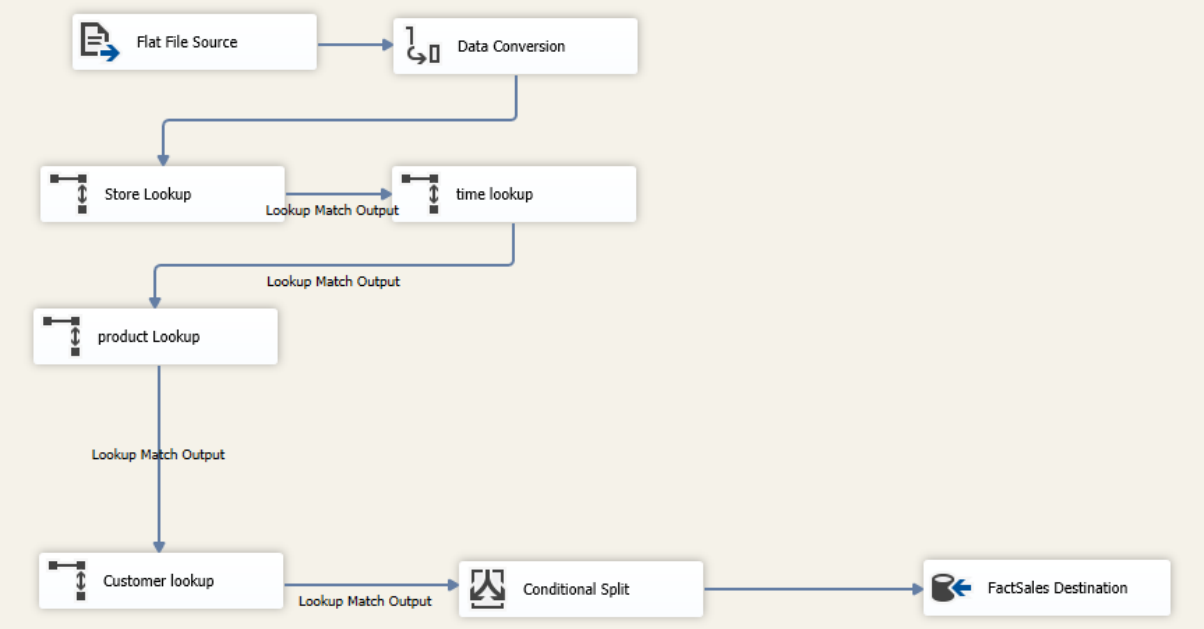


1. **Flat File Source**:
   - **Mục đích**: Đây là nơi dữ liệu được trích xuất từ một tập tin dạng phẳng (ví dụ như CSV).
   - **Chức năng**: Đọc dữ liệu từ Customers.csv được chỉ định và đưa vào dòng dữ liệu SSIS để xử lý.
2. **Data Conversion**:
   - **Mục đích**: Chuyển đổi kiểu dữ liệu của các cột trong dòng dữ liệu.
   - **Chức năng**: Dữ liệu khi được từ module "Flat File Source" mặc định sẽ là text, và sau khi chạy qua module thì chuyển đổi kiểu dữ liệu theo nhu cầu.
3. **Derived Column**:
   - **Mục đích**: Thêm cột mới hoặc thay đổi cột hiện tại trong dòng dữ liệu.
   - **Chức năng**: Sử dụng để chuyển đổi male và female sang 0 và 1.
4. **Data Conversion (string to int)**:

- **Mục đích và Chức năng**: Cụ thể hơn, ở bước này, chuyển đổi kiểu gender từ string to int

5. **Conditional Split**:
   - **Mục đích**: Phân loại dữ liệu dựa trên một hoặc nhiều điều kiện.
   - **Chức năng**: Phân tách dữ liệu NULL của khoá chính.
6. **Customer Destination**:
   - **Mục đích**: Là điểm đến cuối cùng trong quy trình ETL này, nơi dữ liệu được tải vào cơ sở dữ liệu.
   - **Chức năng**: Ghi dữ liệu vào bảng khách hàng trong cơ sở dữ liệu, có thể là SQL Server hoặc bất kỳ nguồn OLE DB nào khác mà SSIS hỗ trợ.

2.3.7 FactSales Flow Task



1. **Flat File Source**:
   - **Mục đích**: Đây là nơi dữ liệu được trích xuất từ một tập tin phẳng.
   - **Chức năng**: Đọc dữ liệu từ Sales.csv và đưa vào dòng dữ liệu SSIS để xử lý.
2. **Data Conversion**:
   - **Mục đích**: Chuyển đổi kiểu dữ liệu của các cột trong dòng dữ liệu.
   - **Chức năng**: Dữ liệu khi được từ module "Flat File Source" mặc định sẽ là text, và sau khi chạy qua module thì chuyển đổi kiểu dữ liệu theo nhu cầu.
3. **Lookup Transformations** (Store Lookup, Time Lookup, Product Lookup, Customer Lookup):
   - **Mục đích**: Lấy thông tin bổ sung từ các bảng chiều (dimension tables) tương ứng như Store, Time, Product, và Customer.

- **Chức năng**: Tham chiếu các khóa ngoại từ bảng Fact đến các bảng chiều, giúp định danh và liên kết dữ liệu của các sự kiện trong bảng Fact với dữ liệu chi tiết có trong các bảng chiều.

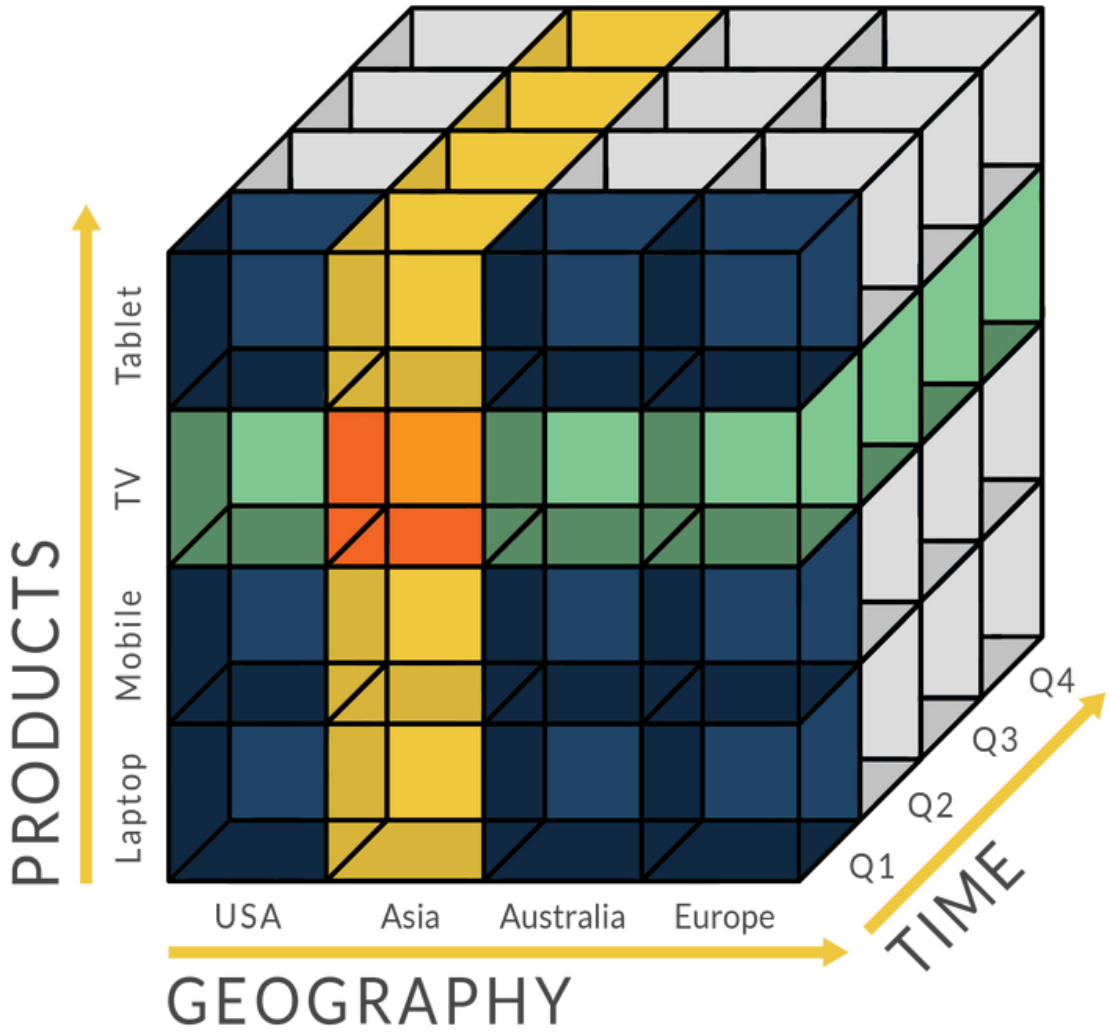
4. **Conditional Split**:
   - **Mục đích**: Phân loại dữ liệu dựa trên một hoặc nhiều điều kiện.
   - **Chức năng**: Phân tách dữ liệu NULL của khoá chính.
5. **FactSales Destination**:
   - **Mục đích**: Là điểm đến cuối cùng cho dữ liệu trong quy trình ETL, nơi dữ liệu được ghi vào bảng Fact của kho dữ liệu.
   - **Chức năng**: Lưu trữ các sự kiện kinh doanh sau khi đã được xử lý và liên kết với các bảng chiều.

2.3.8 Tổng kết

Quy trình ETL này tạo điều kiện cho việc lưu trữ và quản lý dữ liệu một cách hiệu quả trong kho dữ liệu, làm cơ sở cho việc phân tích dữ liệu và ra quyết định trong doanh nghiệp. Mỗi bước trong quy trình đều đóng một vai trò quan trọng, từ việc đảm bảo tính chính xác của dữ liệu đầu vào cho đến việc lưu trữ dữ liệu trong một định dạng thích hợp để dễ dàng truy vấn và phân tích.

2.4 Thực hiện hệ thống OLAP

OLAP (Online Analytical Processing) hay **Xử lý phân tích trực tuyến** là một công cụ mạnh mẽ giúp các doanh nghiệp khám phá, phân tích và hiểu rõ dữ liệu của mình một cách sâu sắc. Nó cho phép người dùng nhìn vào dữ liệu từ nhiều góc độ khác nhau, từ đó đưa ra những quyết định kinh doanh sáng suốt hơn.
OLAP sử dụng một mô hình dữ liệu đa chiều, thường được biểu diễn dưới dạng một khối lập phương. Mỗi cạnh của khối lập phương đại diện cho một chiều (dimension) của dữ liệu, ví dụ như thời gian, sản phẩm, khách hàng, địa lý. Các giá trị đo lường (measures) như doanh thu, lợi nhuận được lưu trữ tại các giao điểm của các chiều này.



2.4.1 Các hoạt động chính trong OLAP

- **Cắt lát (Slice):** Xem dữ liệu theo một mặt cắt cụ thể của khối lập phương.
- **Khoan sâu (Drill-down):** Đi vào chi tiết hơn của dữ liệu bằng cách chia nhỏ các chiều.
- **Cuộn lên (Roll-up):** Tổng hợp dữ liệu lên các cấp độ cao hơn.
- **Xoay (Pivot):** Thay đổi trục của khối lập phương để xem dữ liệu từ góc độ khác.

Đối với data (Sales) mà nhóm sử dụng vào bài làm, sử dụng công cụ SSAS (SQL Server Analysis Services) và Visual Studio để tạo Cube OLAP và sử dụng công cụ Excel để trực quan hóa các hoạt động chính trong OLAP

Quá trình tạo hệ thống OLAP với Analysis Services trên Visual Studio:

1. Chuẩn bị ban đầu:

- Chuẩn bị kho dữ liệu từ các bước ETL trước đó
- Cài đặt SQL Server kèm Analysis Services
- Cài Visual Studio với SQL Server Data Tools (SSDT) và Microsoft Analysis Services Projects trên Extensions.
- Cài Excel (khuyến nghị phiên bản 2013 trở lên)

2. Tạo dự án OLAP:

- Mở Visual Studio
- Chọn File > New > Project
- Chọn Analysis Services Multidimensional Project
- Đặt tên dự án (Analysis_Services_Project.sln)

3. Kết nối nguồn dữ liệu:

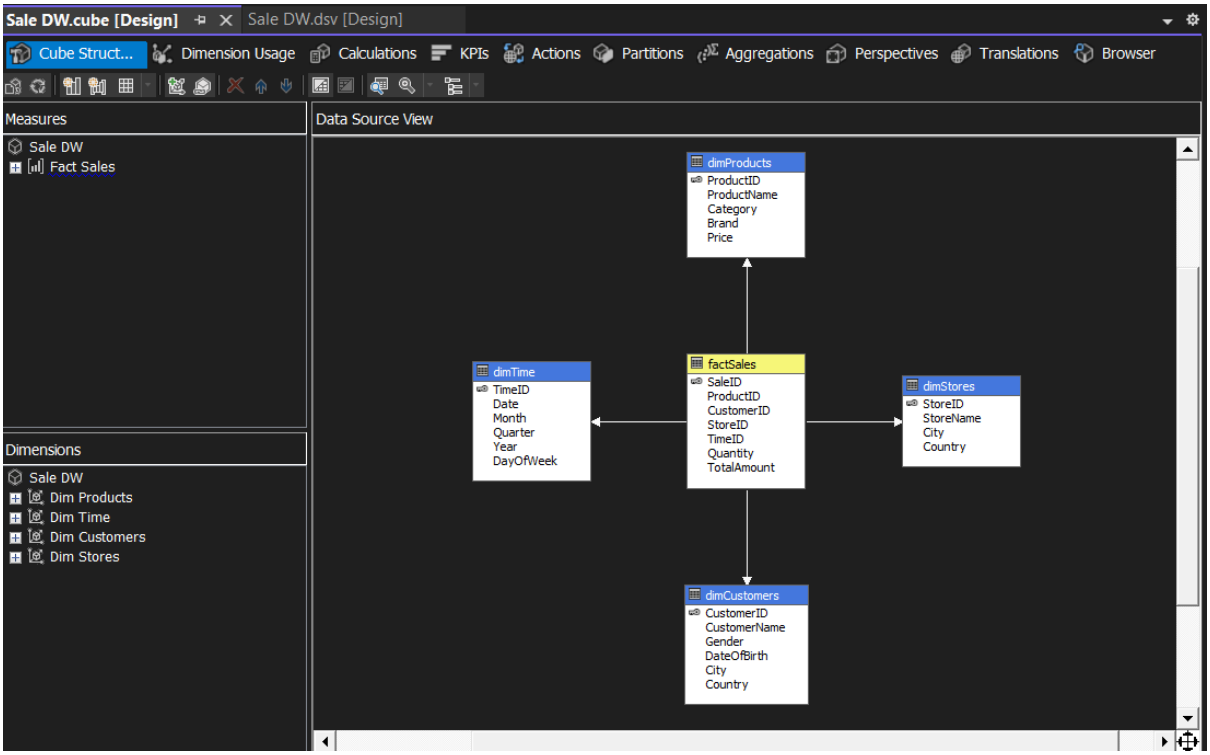
- Thêm Data Source trong Solution Explorer
- Chọn SQL Server làm nguồn dữ liệu
- Nhập thông tin kết nối CSDL

4. Xây dựng Dimensions:

- Tạo mới Data Source View
- Định nghĩa các chiều (Dimensions)
- Thiết lập thuộc tính và phân cấp

5. Tạo Cube:

- Thêm Cube mới
- Chọn bảng fact và dimension
- Cấu hình các thước đo (Measures)



Cấu trúc Cube

6. Xử lý Cube:

- Deploy dự án
- Process cube để nạp dữ liệu

2.4.2 Kết nối và Phân tích:

Sử dụng Excel hoặc SSMS (SQL Server Analysis Services) để kết nối

Thực hiện các phép phân tích OLAP

Để phân tích OLAP, vấn đề đặt ra để ra quyết định: Mặt hàng nào?

- Có doanh số lớn nhất?
- Trong khoảng thời gian nào?
- Tại khu vực nào?
- Phổ biến đối với những ai?

Drill-down: Cho phép xem xét dữ liệu ở mức độ chi tiết hơn.

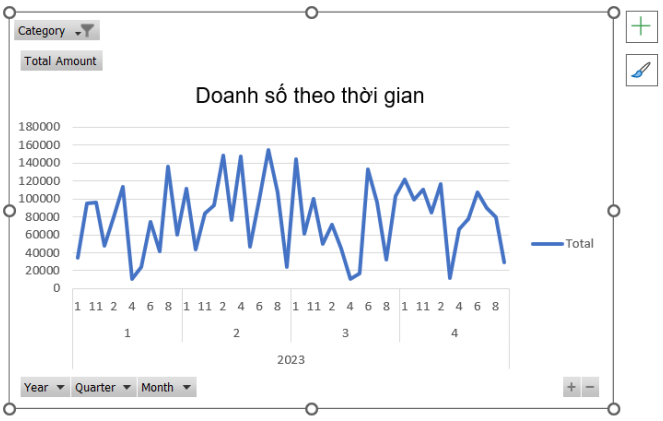
Để trả lời cho câu hỏi trong thời gian nào, công ty này có doanh thu lớn nhất.

→ Sử dụng công cụ pivot-table để trực quan hóa xu hướng doanh số theo từng quý. Từ doanh số theo quý, có thể drill-down để xem được chi tiết doanh số theo từng tháng.

Category	Home

Row Labels	Total Amount
2023	
1	
1	34487
10	95078
11	96181
12	47737
2	79211
3	113010
4	10960
5	23691
6	74374
7	41394
8	135920
9	60063
2	
1	111643
10	43570
11	83754
12	93096
2	148101
3	76925
4	147124
5	47033
6	98676
7	154048
8	106746
9	24477

Biểu diễn xu hướng doanh số theo từng quý và tháng trong năm.



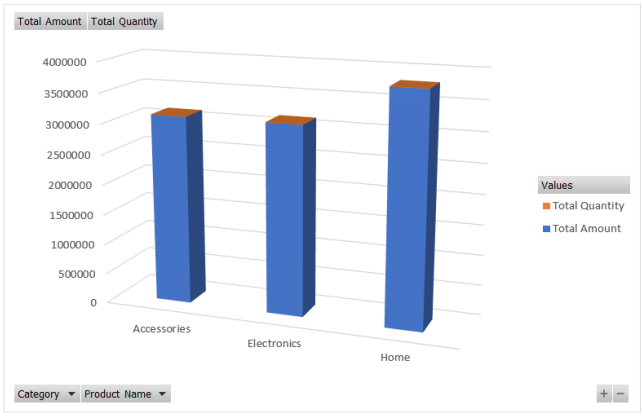
Ta thấy: Quý 4 có doanh số cao nhất trong năm. Đồng thời, tại quý 2 và quý 4 doanh số của sản phẩm gia dụng bán chạy nhất.

Roll-up: Cho phép xem xét dữ liệu ở mức độ tổng quát hơn.

Để mở rộng tầm nhìn dữ liệu ở mức độ tổng quát hơn, roll-up để xem được doanh số theo danh mục sản phẩm.

Row Labels	Total Amount	Total Quantity
Accessories	3123505	3317
Electronics	3128367	3244
Home	3801381	4057
Grand Total	10053253	10618

So sánh doanh số giữa các danh mục sản phẩm như Phụ Kiện, Điện Tử, Đồ Gia Dụng.



Nhận thấy: Các sản phẩm gia dụng có doanh thu chiếm cao nhất.

Slice: Cho phép lọc dữ liệu theo một hoặc nhiều chiều

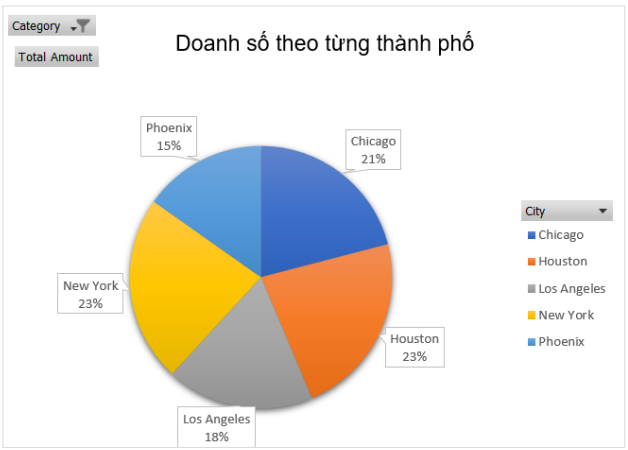
Tiếp theo, để trả lời cho câu hỏi phát triển nhất ở khu vực nào?

→ "Slice" để lọc dữ liệu theo một hoặc nhiều chiều. Cụ thể, nhóm lọc dữ liệu theo khu vực

Category	Home

Row Labels	Total Amount
Chicago	796434
Houston	863172
Los Angeles	690857
New York	872292
Phoenix	578626
Grand Total	**3801381**

Hiển thị doanh số của từng thành phố để so sánh thị phần.



Phát hiện được: Thành phố New York là nơi đạt được doanh số cao nhất nhờ vào việc bán các sản phẩm gia dụng, kế tiếp là Houston và Chicago.

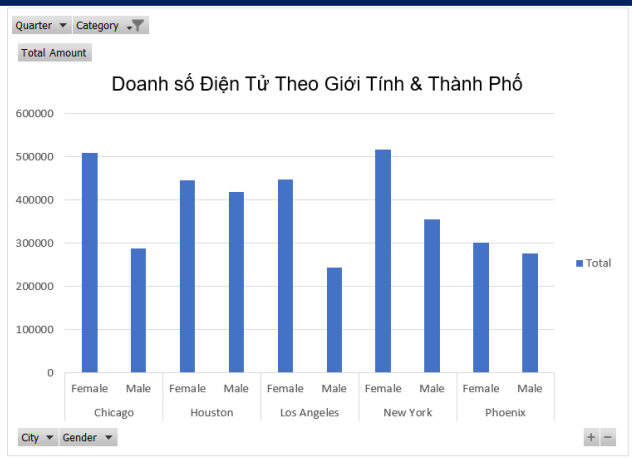
Dice: Cho phép xem xét dữ liệu theo nhiều chiều cùng lúc.

Để chi tiết hóa hơn, sử dụng "dice" để xem doanh số theo danh mục, thành phố và giới tính khách hàng.

Quarter	All
Category	Home

Row Labels	Total Amount
Chicago	
Female	509290
Male	287144
Houston	
Female	444951
Male	418221
Los Angeles	
Female	447688
Male	243169
New York	
Female	517133
Male	355159
Phoenix	
Female	301687
Male	276939
Grand Total	**3801381**

Xem xét doanh số nhóm sản phẩm Điện Tử phân theo giới tính khách hàng và thành phố.



Từ biểu đồ, ta có thể kết luận: Lượng người mua các mặt hàng gia dụng hầu hết là nữ giới và tập trung tại các thành phố như Chicago, Houston và New York.

Từ các hoạt động trên, hỗ trợ đưa ra các quyết định:

- Đầu tư vào các sản phẩm phân loại "Home" (đồ gia dụng)
- Mở rộng cửa hàng tại New York, Houston và Chicago.
- Chiến lược marketing cho nữ giới.

TÀI LIỆU THAM KHẢO

Christian Cote, Matija Lah, Dejan Sarka. (June 30, 2017). *SQL Server 2017 Integration Services Cookbook: Powerful ETL techniques to load and transform data from almost any source.* Packt Publishing.