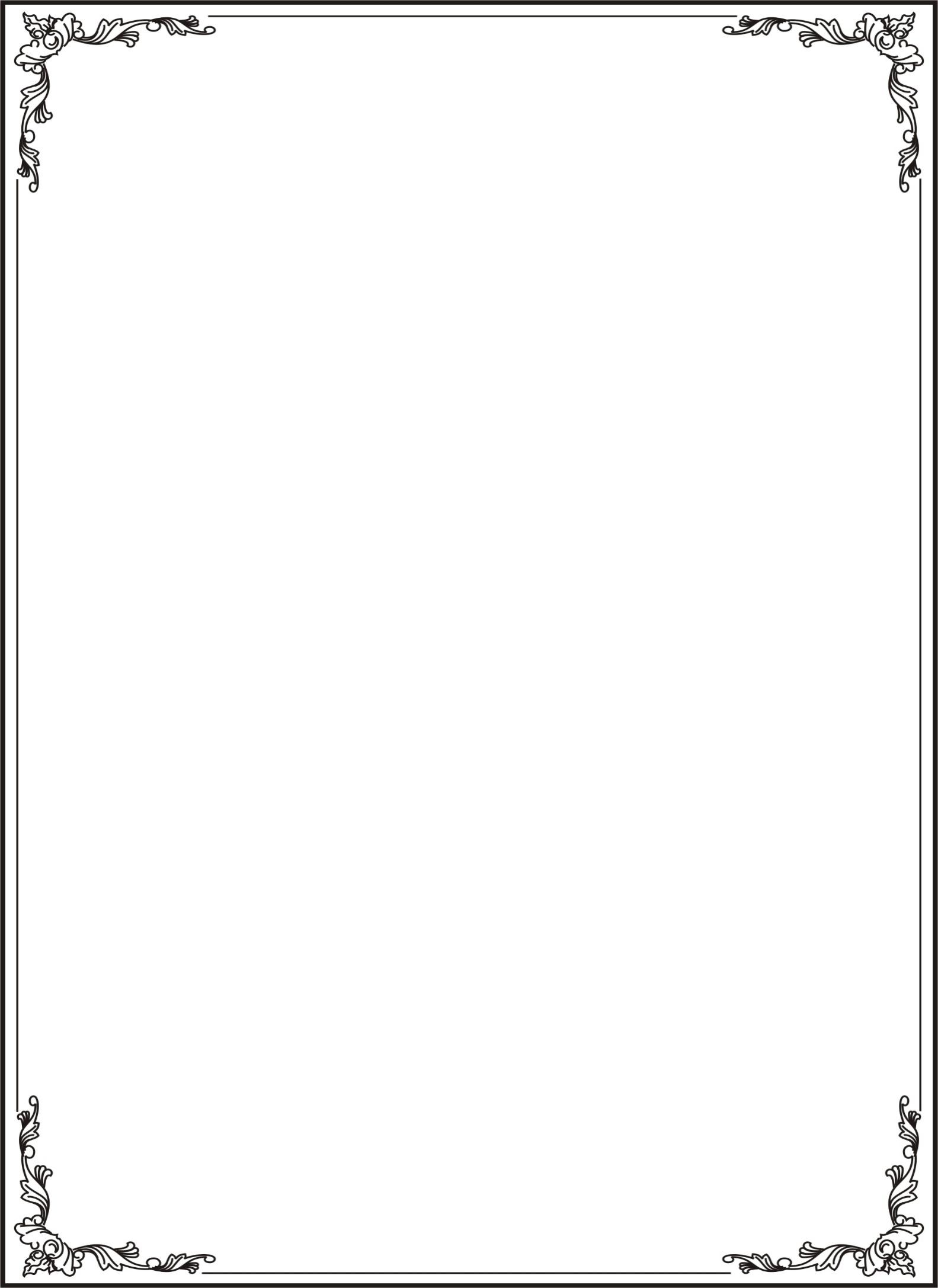
**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**

🙠🙟🕮🙝🙢



**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**

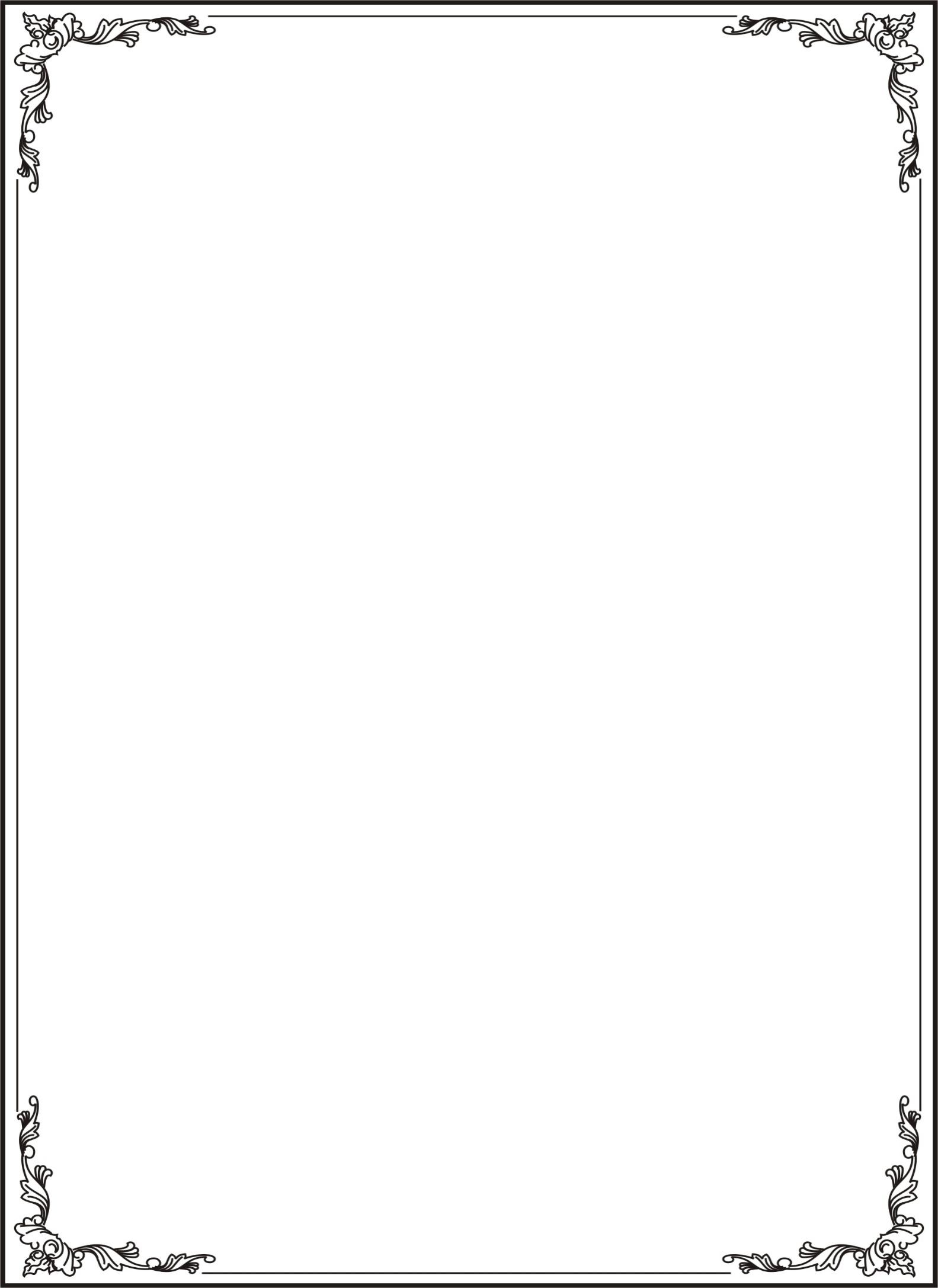
**Giảng viên hướng dẫn: TS. TRẦN NHẬT QUANG**

**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**

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**

🙠🙟🕮🙝🙢

A logo of hands holding a book

Description automatically generated

**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 4](#_Toc183598274)

[DANH MỤC CÁC HÌNH 5](#_Toc183598275)

[DANH MỤC CÁC BẢNG 6](#_Toc183598276)

[1. ĐẶC TẢ ĐỀ TÀI 7](#_Toc183598277)

[**1.1. Mục tiêu đề tài** 7](#_Toc183598278)

[**1.2. Dữ liệu, thông tin đầu vào** 7](#_Toc183598279)

[**1.3. Mục đích, tính năng của đề tài** 7](#_Toc183598280)

[**1.4. Kết quả dự kiến** 8](#_Toc183598281)

[2. KẾ HOẠCH THỰC HIỆN 9](#_Toc183598282)

[3. THIẾT KẾ 10](#_Toc183598283)

[**3.1. Tiền xử lý dữ liệu** 10](#_Toc183598284)

[**3.2. Thiết kế cơ sở dữ liệu nguồn** 13](#_Toc183598285)

[**3.3. Mô hình hóa kho dữ liệu** 16](#_Toc183598286)

[**3.4. ETL dữ liệu** 19](#_Toc183598287)

[**3.5. Data cube design với SSAS** 21](#_Toc183598288)

[4. CÀI ĐẶT VÀ KIỂM THỬ 24](#_Toc183598289)

[**4.1. Xây dựng dashboard trên Excel pivot table** 24](#_Toc183598290)

[**4.2. Xây dựng dashboard trên Power BI** 27](#_Toc183598291)

[5. KẾT LUẬN 29](#_Toc183598292)

[**5.1. Các kết quả đạt được** 29](#_Toc183598293)

[**5.2. Ưu điểm và hạn chế của đề tài** 29](#_Toc183598294)

[**5.3. Định hướng phát triển** 29](#_Toc183598295)

[TÀI LIỆU THAM KHẢO 30](#_Toc183598296)

# **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 |

# **DANH MỤC CÁC HÌNH**

[Hình 1: Diagram thể hiện mối quan hệ giữa các bảng 13](#_Toc183643273)

[Hình 2: Mô hình star schema của kho dữ liệu 16](#_Toc183643274)

[Hình 3: Dashboard tạo bằng Pivot Table 26](#_Toc183643275)

[Hình 4: Dashboard tạo bằng Power BI 28](#_Toc183643276)

# **DANH MỤC CÁC BẢNG**

[Bảng 1: Kế hoạch thực hiện đề tài 9](#_Toc183643296)

[Bảng 2: Code thực hiện tiền xử lý dữ liệu 10](#_Toc183643297)

[Bảng 3: Bảng mô tả các Table trong CSDL 14](#_Toc183643298)

[Bảng 4: Mô hình hóa kho dữ liệu 17](#_Toc183643299)

[Bảng 5: Quy trình ETL dữ liệu 19](#_Toc183643300)

[Bảng 6: Các bước thực hiện thiết kế Data cube 21](#_Toc183643301)

[Bảng 7: Các bước xây dựng dashboard trên Pivot Table 24](#_Toc183643302)

[Bảng 8: Các bước xây dựng dashboard trên Power BI 27](#_Toc183643303)

## **ĐẶC TẢ ĐỀ TÀI**

### **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.

### **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.

### **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.

### **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.

## **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% |

## **THIẾT KẾ**

### **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 | import pandas as pd  data= pd.read\_csv('DataCoSupplyChainDataset.csv', encoding='ISO-8859-1')  print(data.head())  print(data.shape)  data | 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 | null\_values = data.isnull().sum()  null\_columns = null\_values[null\_values > 0]  print("Những cột có giá trị null:")  print(null\_columns)  data = data.fillna("Unknown") | Tìm những cột có nhiều giá trị null và tự động điền giá trị null thành giá trị “Unknown”. |
| 3 | null\_values = data.isnull().sum()  null\_columns = null\_values[null\_values > 0]  print("Những cột có giá trị null:")  print(null\_columns) | Sau khi tự động điền các giá trị null, kiểm tra lại để đảm bảo không còn cột nào có giá trị null. |
| 4 | 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) | 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 | 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) | 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 | 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) | 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 | 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) | 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. |
| 8 | 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) | 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. |
| 9 | 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) | 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. |
| 10 | 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) | 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. |
| 11 | 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) | 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. |

### **Thiết kế cơ sở dữ liệu nguồn**

Sinh viên phụ trách: Trần Nguyễn Trí Đạt

Diagram thể hiện mối quan hệ giữa các bảng:

A computer screen shot of a computer flowchart

Description automatically generated

Hình : Diagram thể hiện mối quan hệ giữa các bảng

Bảng : Bảng mô tả các Table trong CSDL

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Tên bảng** | **Tên trường** | **Kiểu dữ liệu** | **Mục đích** |
| 1 | Location | Zipcode | Varchar(6) | Lưu trữ thông tin về vị trí địa lý |
| City | Varchar(50) |
| Country | Varchar(50) |
| State | Varchar(3) |
| 2 | Customer | Customer\_Id | Int | Lưu trữ thông tin về khách hàng |
| Zipcode | Varchar(6) |
| Segment | Varchar(50) |
| First\_name | Varchar(50) |
| Last\_name | Varchar(50) |
| 3 | Department | Department\_Id | Int | Lưu trữ thông tin về phòng ban của cửa hàng |
| Department\_Name | Varchar(50) |
| 4 | Category | Category\_Id | Tinyint | Lưu trữ thông tin về loại hình sản phẩm |
| Category\_Name | Varchar(50) |
| 5 | Product | Product\_Id | Int | Lưu trữ thông tin về sản phẩm |
| Product\_Name | Varchar(200) |
| Category\_Id | Tinyint |
| Product\_Price | Float |
| 6 | Orders | Order\_Id | 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 | Order\_Item\_Id | 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 | Order\_Id | 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) |

### **Mô hình hóa kho dữ liệu**

Sinh viên phụ trách: Trần Nguyễn Trí Đạt

Diagram thể hiện mối quan hệ giữa các bảng chiều:

A screenshot of a computer

Description automatically generated

Hình : Mô hình star schema của kho dữ liệu

Bảng 4: Mô hình hóa kho dữ liệu

|  |  |  |
| --- | --- | --- |
| **STT** | **Code** | **Mục đích** |
| 1 | -- 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)  ) | Tạo kho dữ liệu “DataCo\_DWH” để thực hiện xây dựng kho dữ liệu.  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 |
| 2 | -- 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,  ); | Tạo bảng chiều DimProduct lưu trữ dữ liệu thông tin các sản phẩm có trong tập dữ liệu. |
| 3 | -- 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,  RowStartDate DATETIME DEFAULT '12/31/1899' NOT NULL,  RowEndDate DATETIME DEFAULT '12/31/9999' NOT NULL,  RowChangeReason NVARCHAR(200) NULL,  ); | Tạo bảng chiều DimCustomer lưu trữ dữ liệu thông tin các khách hàng có trong tập dữ liệu. |
| 4 | -- 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  ); | Tạo bảng chiều DimDepartment lưu trữ dữ liệu thông tin các phòng ban có trong tập dữ liệu. |
| 5 | -- 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)  ); | 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  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 |

### **ETL dữ liệu**

Sinh viên phụ trách: Trần Nguyễn Trí Đạt

Bảng 5: Quy trình ETL dữ liệu

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Control flow** | **Data flow** | **Mục đích** |
| 1 | Control flow cho DimDate | **Date:**  Extract from Source to StageDate | Thực hiện trích xuất dữ liệu từ cơ sở dữ liệu nguồn (Extract), và biến đổi (Transform) để đổ vào bảng Stage |
| Load from Stage to DimDate | Thực hiện tải (Load) dữ liệu vào kho dữ liệu từ bảng Stage vào bảng Dim |
| 2 | Control flow cho các bảng: DimProduct, DimCustomer, DimDepartment | **Product:**   * Nguồn -> Stage: | Thực hiện trích xuất, biến đổi, tải dữ liệu của bảng Product vào bảng chiều DimProduct của kho dữ liệu |
| * Stage -> Dim: |
| **Customer:**   * Nguồn -> Stage: | Thực hiện trích xuất, biến đổi, tải dữ liệu của bảng Customer vào bảng chiều DimCustomer của kho dữ liệu |
| * Stage -> Dim: |
| **Department:**   * Nguồn -> Stage: | Thực hiện trích xuất, biến đổi, tải dữ liệu của bảng Department vào bảng chiều DimDepartment của kho dữ liệu |
| * Stage -> Dim: |
| 3 | Control flow cho FactSales | **Sales:**   * Nguồn -> Stage: | Thực hiện trích xuất, biến đổi, tải dữ liệu của bảng Sales vào bảng FactSales của kho dữ liệu |
| * Stage -> Fact: |

### **Data cube design với SSAS**

Sinh viên phụ trách: Trần Nguyễn Trí Đạt

Bảng 6: Các bước thực hiện thiết kế Data cube

|  |  |  |
| --- | --- | --- |
| **STT** | **Các bước thực hiện** | **Mục đích** |
| 1 | Tạo data source: | - 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 | Tạo data source view: | - 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 | Tạo data cube cho data source view: | - 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 diễn một mô hình dữ liệu đa chiều, trong đó dữ liệu được tổ chức theo nhiều chiều khác nhau. |
| 4 | Phân cấp dữ liệu:   * DimDate:      * DimCustomer:      * DimProduct: | - 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 |
| 5 | Phân tích trên data cube đã tạo: | - Sau khi hoàn thành Process data cube, chuyển qua tab Browser của data cube để thực hiện **duyệt và phân tích dữ liệu** trong các Cube sau khi chúng đã được triển khai (deployed) và xử lý (processed). |

## **CÀI ĐẶT VÀ KIỂM THỬ**

### **Xây dựng dashboard trên Excel pivot table**

Bảng : Các bước xây dựng dashboard trên Pivot Table

|  |  |  |
| --- | --- | --- |
| **STT** | **Các bước thực hiện** | **Mục đích** |
| 1 | Kết nối đến cube data:  A screenshot of a computer  Description automatically generated | Thực hiện kết nối đến dữ liệu thông qua Analysis Services để kết nối đến cube data đã tạo bằng SSAS. |
| 2 | Thiết lập kết nối:  A screenshot of a computer  Description automatically generated | Cấu hình cho kết nối đến cube dữ liệu thông qua server name để kết nối trực tiếp đến khối dữ liệu. |
| 3 | Chọn cube dữ liệu đã tạo:  A screenshot of a computer  Description automatically generated | 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 |
| 4 | Lưu kết nối:  A screenshot of a computer  Description automatically generated | 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 |
| 5 | Kết nối đến connection vừa tạo để tạo các PivotTable Report / PivotChart:    Chọn tùy chọn theo nhu cầu để tạo Report / Dashboard: | 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. |
| 6 | Trực quan hóa dữ liệu: | 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 Field List bên phải |

Sau khi hoàn thành được dashboard trên Pivot Table để trực quan hóa dữ liệu như sau:

A screenshot of a computer

Description automatically generated

Hình : Dashboard tạo bằng Pivot Table

### **Xây dựng dashboard trên Power BI**

Bảng 8: Các bước xây dựng dashboard trên Power BI

|  |  |  |
| --- | --- | --- |
| **STT** | **Các bước thực hiện** | **Mục đích** |
| 1 | Kết nối đến cube data: | Thực hiện kết nối đến dữ liệu thông qua Analysis Services để kết nối đến cube data đã tạo bằng SSAS. |
| 2 | Thiết lập kết nối: | Cấu hình cho kết nối đến cube dữ liệu thông qua server name để kết nối trực tiếp đến khối dữ liệu. |
| 3 | Chọn cube dữ liệu đã tạo: | 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 |
| 4 | Trực quan hóa dữ liệu: | 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 |

Sau khi hoàn thành được dashboard trên Power BI để trực quan hóa dữ liệu như sau:

**A screenshot of a computer screen

Description automatically generated**

Hình : Dashboard tạo bằng Power BI

## **KẾT LUẬN**

### **Các kết quả đạt được**

Thực hiện đề tài "Ứng dụng kho dữ liệu trong phân tích và trực quan hóa dữ liệu bán hàng" giúp hiểu sâu hơn về kiến thức Data Warehouse, từ việc thiết kế mô hình dữ liệu cho đến triển khai hệ thống thực tế. Đề tài cũng giúp em làm quen và thành thạo hơn trong việc sử dụng các công cụ hỗ trợ quá trình ETL như SSIS để trích xuất, chuyển đổi và tải dữ liệu vào kho dữ liệu, cũng như các công cụ phân tích và trực quan hóa dữ liệu như Power BI Desktop và Excel Pivot.

Ngoài ra, quá trình thực hiện còn rèn luyện khả năng tiếp thu và xử lý thông tin đầu vào, phân tích dữ liệu một cách có hệ thống, đặt câu hỏi phù hợp để tìm ra hướng giải quyết vấn đề, và xây dựng các báo cáo trực quan nhằm trả lời các câu hỏi kinh doanh. Đây là nền tảng quan trọng giúp em nâng cao kỹ năng thực hành, tư duy phân tích, và khả năng vận dụng công nghệ vào thực tiễn.

### **Ưu điểm, khó khăn và hạn chế của đề tài**

**Ưu điểm:**

Đề tài mang lại nhiều lợi ích trong việc hiểu và áp dụng kiến thức về kho dữ liệu và trực quan hóa dữ liệu. Thông qua việc sử dụng các công cụ như SSIS và Power BI, không chỉ nâng cao kỹ năng thực hành mà còn phát triển tư duy phân tích dữ liệu, học cách thiết kế và xây dựng một hệ thống kho dữ liệu hiệu quả. Kết quả của đề tài, với dashboard trực quan, giúp tổng hợp dữ liệu phức tạp thành các biểu đồ và báo cáo dễ hiểu, là một thành tựu đáng ghi nhận.

**Khó khăn và hạn chế:**

Quá trình làm việc gặp khó khăn trong việc làm sạch và chuẩn hóa dữ liệu, đặc biệt khi dữ liệu từ Kaggle có thể không đồng nhất hoặc thiếu thông tin. Việc thiết kế kho dữ liệu đòi hỏi nhiều thời gian và sự cân nhắc để đảm bảo tính hợp lý và khả năng mở rộng. Hơn nữa, việc sử dụng các công cụ như SSAS hoặc Power BI đòi hỏi thời gian tự học, nhất là khi chưa có nhiều kinh nghiệm. Một thách thức khác là khả năng chia sẻ dashboard cho người khác, do yêu cầu về môi trường phần mềm hoặc tài khoản cụ thể, gây khó khăn khi muốn triển khai cho nhiều đối tượng sử dụng. Cuối cùng, dự án chỉ dừng lại ở mức mô phỏng và học tập, chưa đủ khả năng để áp dụng trực tiếp vào quy mô doanh nghiệp thực tế.

### **Định hướng phát triển**

Để nâng cao tính ứng dụng và hoàn thiện đề tài, hệ thống có thể được mở rộng theo nhiều hướng. Một trong những định hướng chính là tối ưu hóa quá trình ETL, đảm bảo xử lý được khối lượng dữ liệu lớn hơn và cải thiện hiệu suất khi tích hợp dữ liệu từ nhiều nguồn khác nhau. Việc bổ sung các kịch bản tự động hóa trong việc cập nhật dữ liệu theo thời gian thực cũng là một bước tiến quan trọng, giúp tăng tính linh hoạt và hiệu quả của hệ thống.

Hệ thống dashboard có thể được phát triển để hỗ trợ nhiều người dùng hơn, với tính năng phân quyền và chia sẻ dữ liệu linh hoạt thông qua nền tảng web hoặc các công cụ trực tuyến. Ngoài ra, việc tích hợp các công nghệ mới như trí tuệ nhân tạo hoặc học máy để phân tích dữ liệu chuyên sâu cũng là một định hướng tiềm năng, giúp cung cấp các dự báo chính xác hơn về doanh thu, xu hướng bán hàng, và hành vi khách hàng.

Cuối cùng, hệ thống có thể được nghiên cứu để chuyển từ một dự án học thuật sang ứng dụng thực tiễn trong môi trường doanh nghiệp, đáp ứng tốt hơn các yêu cầu cụ thể của thị trường và nâng cao giá trị của dữ liệu trong việc hỗ trợ ra quyết định.

# **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>