**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA HỆ THỐNG THÔNG TIN**



**BÁO CÁO ĐỒ ÁN MÔN HỌC**

KHO DỮ LIỆU VÀ OLAP

**ĐỀ TÀI:**

THIẾT KẾ KHO DỮ LIỆU NORTHWIND

**GVHD: ThS. Nguyễn Thị Kim Phụng**

Nhóm sinh viên thực hiện:

1. Đặng Huỳnh Vĩnh Tân - MSSV: 21520442

2. Lý Phi Lân - MSSV: 21520319

**TP. HỒ CHÍ MINH, NĂM 2024**

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

*……., ngày……...tháng……năm 2023*

**Người nhận xét**

*(Ký tên và ghi rõ họ tên)*

***MENU***

[CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN 5](#_Toc135768930)

[1. Lý do chọn đề tài 5](#_Toc135768931)

[2. Giới thiệu dataset 5](#_Toc135768932)

[3. Mô tả dữ liệu gốc 6](#_Toc135768933)

[4. Thiết kế kho dữ liệu 7](#_Toc135768934)

[4.1 Lược đồ kho dữ liệu( Lược đồ hình sao) 7](#_Toc135768935)

[4.2 Các Bảng chiều 8](#_Toc135768936)

[4.2.1 Bảng Dime\_Date 8](#_Toc135768937)

[4.2.2 Bảng Dime\_Ship 8](#_Toc135768938)

[4.2.3 Bảng Dime\_Order 9](#_Toc135768939)

[4.2.4 Bảng Dime\_Product 9](#_Toc135768940)

[4.3 Bảng sự kiện 10](#_Toc135768941)

[5. Các câu truy vấn 10](#_Toc135768942)

[CHƯƠNG 2. TÍCH HỢP DỮ LIỆU VÀO KHO (SSIS) 11](#_Toc135768943)

[1. Chuẩn bị công cụ va DataWareHouse 11](#_Toc135768944)

[2. Tạo project SSIS trong visual studio 2022 13](#_Toc135768945)

[3. Làm sạch dữ liệu và tạo các bảng Dim, Fact 14](#_Toc135768946)

[CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU TRONG KHO (SSAS) 33](#_Toc135768947)

[1. TẠO PROJECT SSAS TRONG VISUAL STUDIO 2022 33](#_Toc135768948)

[2. XÁC ĐỊNH DỮ LIỆU NGUỒN (DEFINE DATE SOURCE) 34](#_Toc135768949)

[3. XÁC ĐỊNH KHUNG NHÌN DỮ LIỆU NGUỒN (DEFINE DATE SOURCE VIEW) 37](#_Toc135768950)

[4. XÂY DỰNG CÁC KHỐI VÀ CÁC MEASURE 40](#_Toc135768951)

[5. QUÁ TRÌNH PHÂN TÍCH DỮ LIỆU BẰNG CÔNG CỤ SSAS TRÊN CÁC KHỐI CUBE 44](#_Toc135768952)

[5.1. Liệt kê tổng tiền mua của các thành phố theo từng tháng 45](#_Toc135768953)

[5.2. Cho biết hóa đơn đạt tổng số tiền mua lớn nhất 46](#_Toc135768954)

[5.3.Theo từng tháng, quý, năm liệt kê các tổng số lượng mua của từng sản phẩm và thể loại sản phẩm được mua. 48](#_Toc135768955)

[5.4. Cho biết hóa đơn có tổng số lượng sản phẩm đã mua trên 3 trong tháng 6. 48](#_Toc135768956)

[5.5. Cho biết tên sản phẩm, loại sản phẩm và tổng số lượng sản phẩm đó lớn hơn 3 được đặt theo từng thành phố. 49](#_Toc135768957)

[5.6. Với từng thành phố liệt kê số lượng chi tiết đơn hàng bán ra theo từng tháng, trừ thành phố ADONI và Adyal. 51](#_Toc135768958)

[5.7. Top 3 sản phẩm có số lượng hóa đơn cao nhất 52](#_Toc135768959)

[5.8. Cho biết sản phẩm nào nòa có số lượng mua cao nhất và sản phẩm có số lần mua hàng thấp nhất. 54](#_Toc135768960)

[5.9. Liệt kê những sản phẩm có doanh thu từ nhỏ hơn 500 hoặc 1000 55](#_Toc135768961)

[5.10 Đưa ra doanh thu, số lượng sản phẩm bán ra của từng thể loại sản phẩm. 56](#_Toc135768962)

[5.11. Với mỗi tỉnh đưa ra 3 thành phố có lợi nhuận cao nhất. 56](#_Toc135768963)

[5.12. Đưa ra 10 sản phẩm có số lượng hóa đơn thấp nhất nhất tại tỉnh CHHATTISGARH 58](#_Toc135768964)

[6. QUÁ TRÌNH PHÂN TÍCH DỮ LIỆU BẰNG CÔNG CỤ PIVOT EXCEL 59](#_Toc135768965)

[6.1 Liệt kê tổng tiền mua của các thành phố. 59](#_Toc135768966)

[6.2 Cho biết hóa đơn đạt tổng số tiền mua lớn nhất 59](#_Toc135768967)

[6.3 Theo từng tháng, quý, năm liệt kê các sản phẩm và thể loại sản phẩm được mua 60](#_Toc135768968)

[6.4 Cho biết hóa đơn có tổng số lượng sản phẩm đã mua trên 3 60](#_Toc135768969)

[6.5 Cho biết mã sản phẩm, loại sản phẩm và tổng số lượng sản phẩm đó lớn hơn 3 61](#_Toc135768970)

[6.6 Với từng thành phố liệt kê số lượng chi tiết đơn hàng bán ra theo từng tháng, trừ thành phố ADONI và Adyal. 63](#_Toc135768971)

[6.7 Top 3 sản phẩm có số lượng hóa đơn cao nhất 64](#_Toc135768972)

[6.8 Cho biết sản phẩm nào nào có số lượng mua cao nhất và sản phẩm có số lần mua hàng thấp nhất. 65](#_Toc135768973)

[6.9 Liệt kê những sản phẩm (ID) có doanh thu nhỏ hơn 500 hoặc lớn hơn $1000 65](#_Toc135768974)

[6.10 Đưa ra doanh thu, số lượng sản phẩm bán ra của từng thể loại sản phẩm. 66](#_Toc135768975)

[6.11 Với mỗi tỉnh đưa ra 3 thành phố có lợi nhuận cao nhất. 67](#_Toc135768976)

[6.12 Đưa ra 10 sản phẩm có doanh thu thấp nhất tại tỉnh CHHATTISGARH 67](#_Toc135768977)

[7. QUÁ TRÌNH PHÂN TÍCH DỮ LIỆU BẰNG NGÔN NGỮ MDX 68](#_Toc135768978)

[7.1 Liệt kê tổng tiền mua của các thành phố. 68](#_Toc135768979)

[7.2 Cho biết hóa đơn đạt tổng số tiền mua lớn nhất. 69](#_Toc135768980)

[7.3 Theo từng tháng, quý, năm liệt kê các sản phẩm và thể loại sản phẩm được mua 69](#_Toc135768981)

[7.4 Cho biết hóa đơn có tổng số lượng sản phẩm đã mua trên 3 69](#_Toc135768982)

[7.5 Cho biết mã sản phẩm, loại sản phẩm và tổng số lượng sản phẩm đó lớn hơn 3 được đặt theo từng thành phố. 70](#_Toc135768983)

[7.6 Với từng thành phố liệt kê số lượng chi tiết đơn hàng bán ra theo từng tháng, trừ thành phố ADONI và Adyal. 70](#_Toc135768984)

[7.7 Top 3 sản phẩm có số lượng hóa đơn cao nhất trong tháng 5 71](#_Toc135768985)

[7.8 Cho biết sản phẩm nào có số lượng mua cao nhất và sản phẩm có số lần mua hàng thấp nhất. 71](#_Toc135768986)

[7.9 Liệt kê những sản phẩm (ID) có doanh thu nhỏ hơn 500 hoặc lớn hơn $1000 72](#_Toc135768987)

[7.10 Đưa ra doanh thu, số lượng sản phẩm bán ra của từng thể loại sản phẩm. 72](#_Toc135768988)

[7.11 Với mỗi tỉnh đưa ra 3 thành phố có lợi nhuận cao nhất. 73](#_Toc135768989)

[7.12 Đưa ra 10 sản phẩm có doanh thu thấp nhất tại tỉnh CHHATTISGARH 74](#_Toc135768990)

[8. THỰC HIỆN TRÊN POWER BI 74](#_Toc135768991)

[8.1 Liệt kê tổng tiền mua của các thành phố 75](#_Toc135768992)

[8.2 Cho biết hóa đơn đạt tổng số tiền mua lớn nhất 76](#_Toc135768993)

[8.3 Theo từng tháng, quý, năm liệt kê các sản phẩm và thể loại sản phẩm được mua 78](#_Toc135768994)

[8.4 Cho biết hóa đơn có tổng số lượng sản phẩm đã mua trên 3 78](#_Toc135768995)

[8.5 Cho biết mã sản phẩm, loại sản phẩm và tổng số lượng sản phẩm đó lớn hơn 3 82](#_Toc135768996)

[8.6 Với từng thành phố liệt kê số lượng đơn hàng bán ra theo từng tháng, trừ thành phố ADONI và Adyal. 85](#_Toc135768997)

[8.7 Top 3 sản phẩm có số lượng hóa đơn cao nhất 87](#_Toc135768998)

[8.8 Cho biết sản phẩm nào nòa có số lượng mua cao nhất và sản phẩm có số lần mua hàng thấp nhất. 89](#_Toc135768999)

[8.9 Liệt kê những sản phẩm (ID) có doanh thu nhỏ hơn 500 hoặc lớn hơn $1000 89](#_Toc135769000)

[8.10 Đưa ra doanh thu, số lượng sản phẩm bán ra của từng thể loại sản phẩm. 91](#_Toc135769001)

[8.11 Với mỗi tỉnh đưa ra 3 thành phố có lợi nhuận cao nhất. 91](#_Toc135769002)

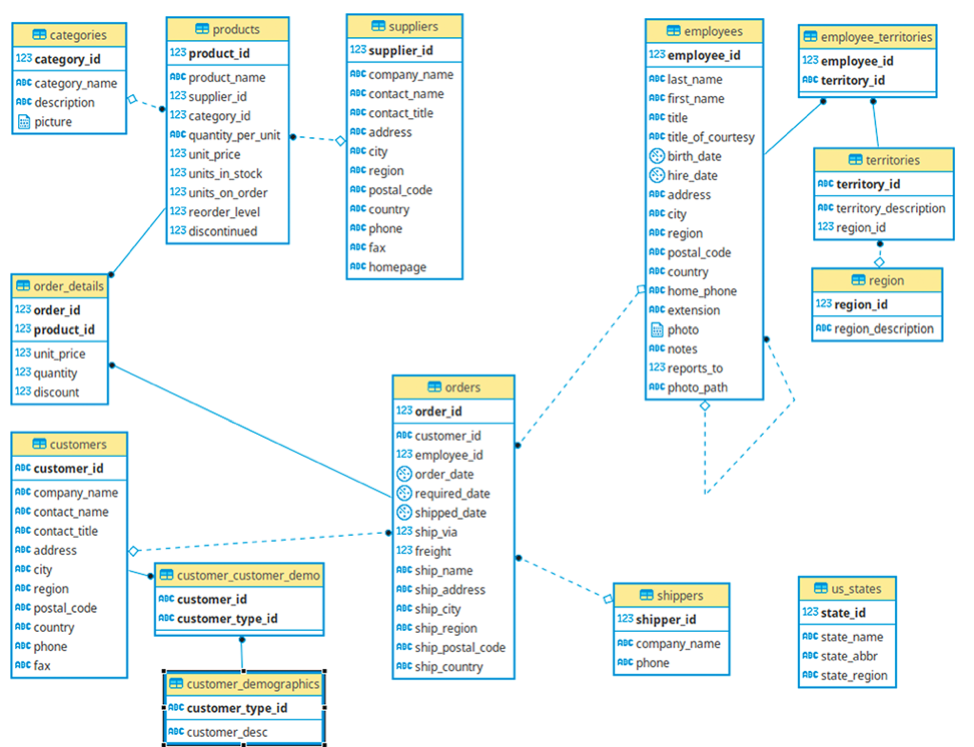
[8.12 Đưa ra 10 sản phẩm có doanh thu thấp nhất tại tỉnh CHHATTISGARH 91](#_Toc135769003)

# CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN

## Lý do chọn đề tài

* Northwind là bộ dữ liệu mô phỏng về một công ty chuyên nhập khẩu và xuất khẩu các loại thực phẩm từ khắp nơi trên thế giới. Dữ liệu đã mô phỏng các thông tin về khách hàng, đơn đặt hàng, hàng tồn kho, mua hàng, nhà cung cấp, vận chuyển, nhân viên,... Vì thế bộ dữ liệu này sẽ trở thành một cơ sở dữ liệu mẫu giúp cho các công ty nhỏ có thể dựa vào nó mà thiết kế cơ sở dữ liệu của chính họ. Tận dụng điều này nhóm quyết định xây dựng một kho dữ liệu – OLAP dựa trên bộ dữ liệu Northwind.

## Giới thiệu dataset

* Tên Bộ dữ liệu: Amazone sale report
* Ngày cập nhật gần nhất: tháng 10/2022
* Cơ sở dữ liệu gồm 14 bảng và 50 thuộc tính
* Link nguồn dữ liệu: [tại đây](https://docs.yugabyte.com/preview/sample-data/northwind/#:~:text=The%20Northwind%20database%20is%20an,%2DMicrosoft%20databases%2C%20including%20PostgreSQL)
* Thiết kế cơ sở dữ liệu:

Hình 1: Mô hình cơ sở dữ liệu Northwind

## Mô tả dữ liệu gốc

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Thuộc tính | Ý nghĩa | Kiểu dữ liệu |
| 1 | Order ID | ID hóa đơn | Nvarchar(50) |
| 2 | Date | Ngày | Date |
| 3 | Status | Trạng thái | Nvarchar(50) |
| 4 | Fulfilment | Phương pháp thực hiện | Nvarchar(50) |
| 5 | Sales\_Channel | Kênh bán hàng | Nvarchar(50) |
| 6 | Ship\_service\_level | Cấp độ của dịch vụ ship | Nvarchar(50) |
| 7 | Style | Kiểu cách sản phẩm | Nvarchar(50) |
| 8 | SKU | Đơn vị quản lý hàng tồn kho | Nvarchar(50) |
| 9 | Category | Thể loại sản phẩm | Nvarchar(50) |
| 10 | Size | Kích cỡ sản phẩm | Nvarchar(50) |
| 11 | ASIN | Mã định danh sản phẩm của Amazon | Nvarchar(50) |
| 12 | Courier\_Status | Trạng thái ship hàng của sản phẩm ở thành phố | Nvarchar(50) |
| 13 | Qty | Số lượng sản phẩm trong hóa đơn | int |
| 14 | Currency | Tiền tệ | Nvarchar(50) |
| 15 | Amount | Tổng tiền của hóa đơn | float |
| 16 | Ship\_city | Tên thành phố cần ship | Nvarchar(50) |
| 17 | Ship\_state | Tên tỉnh cần ship | Nvarchar(50) |
| 18 | Ship\_postal\_code | Mã vùng cần ship (Mỗi một tình tương tứng với 1 thành phố thì có một mã vùng duy nhất) | Int |
| 19 | Ship\_country | Thuộc tính ship trong hay ngoài nước | Nvarchar(50) |
| 20 | promotion-ids | Mã khuyễn mãi | Text |
| 21 | B2B | Giao dich hoạt động kinh doanh giưa hai doanh nghiệp | Nvarchar(50) |
| 22 | fulfilled-by | phương thức đáp ứng được cung cấp bởi Amazon cho người bán | Nvarchar(50) |
| 23 | Unnamed: 22 | Không xác định | Không xác định |
| 24 | Index | Số thứ tự của hóa đơn | Int |

## Thiết kế kho dữ liệu

### 4.1 Lược đồ kho dữ liệu (Lược đồ hình bông tuyết)

A screenshot of a computer

Description automatically generated with medium confidence

### 4.2 Các Bảng chiều

#### 4.2.1 Bảng Dim\_Date

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **STT** | **Tên trường** | **Kiểu dữ liệu** | **Độ dài** | **Cho phép trống** | **Giá trị duy nhất** | **Khóa chính** | **Khóa ngoại** | **Giá trị mặc định** | **Mô tả** |
| 1 | date | date |  | N | N | Y | N |  | Ngày |
| 2 | month | int |  | N | N | N | N |  | Tháng |
| 3 | quarter | int |  | N | N | N | N |  | Quý |
| 4 | year | int |  | N | N | N | N |  | Năm |

#### 4.2.2 Bảng Dim\_Ship

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **STT** | **Tên trường** | **Kiểu dữ liệu** | **Độ dài** | **Cho phép trống** | **Giá trị duy nhất** | **Khóa chính** | **Khóa ngoại** | **Giá trị mặc định** | **Mô tả** |
| 1 | shipper\_id | int |  | N | N | Y | N |  | Id công ty giao hàng |
| 2 | company\_name | varchar | 40 | N | N | N | N |  | Tên công ty giao hàng |

#### 4.2.3 Bảng Dim\_Customer

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **STT** | **Tên trường** | **Kiểu dữ liệu** | **Độ dài** | **Cho phép trống** | **Giá trị duy nhất** | **Khóa chính** | **Khóa ngoại** | **Giá trị mặc định** | **Mô tả** |
| 1 | customer\_id | varchar | 100 | N | Y | Y | N |  | Id của khách hàng |
| 2 | company\_name | varchar | 40 | N | N | N | N |  | Tên công ty |
| 3 | contact\_name | varchar | 30 | Y | N | N | N |  | Tên người liên hệ |
| 4 | address | varchar | 60 | Y | N | N | N |  | Địa chỉ |
| 5 | postal\_code | varchar | 10 | Y | N | N | Y |  | Mã bưa điện |
| 6 | country | varchar | 15 | Y | N | N | Y |  | Quốc gia |
| 7 | phone | varchar | 24 | Y | N | N | N |  | Số điện thoại |
| 8 | fax | varchar | 24 | Y | N | N | N |  | Số fax |

#### 4.2.4 Bảng Dim\_Location

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **STT** | **Tên trường** | **Kiểu dữ liệu** | **Độ dài** | **Cho phép trống** | **Giá trị duy nhất** | **Khóa chính** | **Khóa ngoại** | **Giá trị mặc định** | **Mô tả** |
| 1 | postal\_code | varchar | 10 | N | Y | Y | N |  | Mã bưa điện |
| 2 | country | varchar | 15 | N | Y | Y | N |  | Tên quốc gia |
| 3 | city | varchar | 15 | Y | N | N | N |  | Tên thành phố |
| 4 | region | varchar | 15 | Y | N | N | N |  | Tên vùng |

#### 4.2.5 Bảng Dim\_Supplier

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **STT** | **Tên trường** | **Kiểu dữ liệu** | **Độ dài** | **Cho phép trống** | **Giá trị duy nhất** | **Khóa chính** | **Khóa ngoại** | **Giá trị mặc định** | **Mô tả** |
| 1 | supplier\_id | int |  | N | Y | Y | N |  | Id của nhà cung cấp |
| 2 | company\_name | varchar | 40 | N | N | N | N |  | Tên công ty |
| 3 | address | varchar | 60 | Y | N | N | N |  | Địa chỉ |
| 4 | postal\_code | varchar | 10 | Y | N | N | Y |  | Mã bưa điện |
| 5 | country | varchar | 15 | Y | N | N | Y |  | Quốc gia |
| 6 | phone | varchar | 24 | Y | N | N | N |  | Số điện thoại |

#### 4.2.6 Bảng Dim\_Product

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **STT** | **Tên trường** | **Kiểu dữ liệu** | **Độ dài** | **Cho phép trống** | **Giá trị duy nhất** | **Khóa chính** | **Khóa ngoại** | **Giá trị mặc định** | **Mô tả** |
| 1 | product\_id | int |  | N | Y | Y | N |  | Id của sản phẩm |
| 2 | category\_id | int |  | Y | N | N | Y |  | Id của danh mục |
| 3 | supplier\_id | int |  | Y | N | N | Y |  | Id của nhà cung cấp |
| 4 | product\_name | varchar | 40 | Y | N | N | N |  | Tên sản phẩm |
| 5 | quantity\_per\_unit | varchar | 20 | Y | N | N | N |  | Đơn vị tính |

#### 4.2.7 Bảng Dim\_Category

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **STT** | **Tên trường** | **Kiểu dữ liệu** | **Độ dài** | **Cho phép trống** | **Giá trị duy nhất** | **Khóa chính** | **Khóa ngoại** | **Giá trị mặc định** | **Mô tả** |
| 1 | category\_id | int |  | N | Y | Y | N |  | Id của danh mục |
| 2 | category\_name | varchar | 100 | N | N | N | N |  | Tên danh mục |
| 3 | desciption | text |  | Y | N | N | N |  | Mô tả |

#### 4.2.8 Bảng Dim\_Employee

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **STT** | **Tên trường** | **Kiểu dữ liệu** | **Độ dài** | **Cho phép trống** | **Giá trị duy nhất** | **Khóa chính** | **Khóa ngoại** | **Giá trị mặc định** | **Mô tả** |
| 1 | employee\_id | int |  | N | Y | Y | N |  | Id của của nhân viên |
| 2 | last\_name | varchar | 20 | N | N | N | N |  | Họ nhân viên |
| 3 | first\_name | varchar | 10 | N | N | N | N |  | Tên nhân viên |
| 4 | home\_phone | varchar | 24 | Y | N | N | N |  | Số điện thoại |

### 4.3 Bảng sự kiện

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **STT** | **Tên trường** | **Kiểu dữ liệu** | **Độ dài** | **Cho phép trống** | **Giá trị duy nhất** | **Khóa chính** | **Khóa ngoại** | **Giá trị mặc định** | **Mô tả** |
| 1 | order\_id | int |  | N | Y | Y | N |  | Id đơn hàng |
| 2 | customer\_id | varchar | 100 | N | N | N | Y |  | Id khách hàng |
| 3 | employee\_id | int |  | N | N | N | Y |  | Id nhân viên |
| 4 | product\_id | int |  | N | N | N | Y |  | Id sản phẩm |
| 5 | order\_date | date |  | Y | N | N | N |  | Ngày mua |
| 6 | shipped\_date | date |  | Y | N | N | N |  | Ngày vận chuyển |
| 7 | ship\_via | int |  | Y | N | N | N |  |  |
| 8 | freight | real |  | Y | N | N | N |  | Phí giao hàng |
| 9 | unit\_price | real |  | N | N | N | N |  | Giá trên một đơn vị hàng hoá |
| 10 | quantity | int |  | N | N | N | N |  | Số lượng |

## Các câu truy vấn

**Câu 1**: Liệt kê tổng tiền mua của các thành phố theo tháng.

**Câu 2**: Cho biết hóa đơn đạt tổng số tiền mua lớn nhất.

**Câu 3**: Theo từng tháng, quý, năm liệt kê số lượng các mã sản phẩm bán được.

**Câu 4**: Cho biết hóa đơn có tổng số lượng sản phẩm đã mua trên 3.

**Câu 5**: Cho biết mã sản phẩm, loại sản phẩm và tổng số lượng sản phẩm đó lớn hơn 3 được đặt theo từng thành phố.

**Câu 6**: Top 3 sản phẩm có số lượng hóa đơn cao nhất trong tháng 5.

**Câu 7**: Cho biết sản phẩm nào có tổng số lượng mua cao nhất và sản phẩm có số lần mua hàng thấp nhất.

**Câu 8**: Đưa ra doanh thu, số lượng sản phẩm bán ra của từng thể loại sản phẩm.

**Câu 9**: Với mỗi tháng đưa ra 3 sản phẩm có lợi nhuận cao nhất.

**Câu 10**: Đưa ra 10 sản phẩm có số lượng hóa đơn thấp nhất nhất tại tỉnh CHHATTISGARH.

# CHƯƠNG 2. TÍCH HỢP DỮ LIỆU VÀO KHO (SSIS)

## Chuẩn bị bộ dữ liệu:

A computer screen with a box and text

Description automatically generated**Bước 1: Khởi chạy Microsoft SQL Server Management Studio và kết nối với Server**

**A screenshot of a computer

Description automatically generatedA screenshot of a computer program

Description automatically generatedBước 2: Thực hiện chạy 2 file NorthWindTable.sql và NorthWindData.sql mà nhóm đã chuyển từ ngôn ngữ PostgreSQL thành ngôn ngữ SQLserver để tạo cơ sở dữ liệu**

## Chuẩn bị công cụ

* **Download công cụ SQL Server Data Tools**

A screenshot of a computer

Description automatically generated

Chọn Data storage and processing và chọn những mục của các công cụ SSDT như hình minh họa.

* **Các bước download công cụ SQL Server Integration Services Projects cho Visual Studio**

**Bước 1: Tạo 1 project trong Visual Studio**

**A screenshot of a computer

Description automatically generatedBước 2: Chọn Extensions → Manage Extensions**

**Bước 3: Tìm kiếm SQL Server Integration Services Projects**

**Bước 4: Visual Studio sẽ tự động giúp ta tải về SQL Server Integration Services Projects phù hợp với phiên bản của nó**

**A screenshot of a computer

Description automatically generatedBước 5: Tiến hành cài đặt SQL Server Integration Services Projects (phải tắt Visual Studio trước khi cài đặt)**

## Chuẩn bị Data WareHouse

**Bước 1: Mở Microsoft SQL Server Management Studio và kết nối với Server**

A computer screen with a box and text

Description automatically generated

A screenshot of a computer

Description automatically generated**Bước 2: Tạo kho dữ liệu mới bằng cách chạy file NorthWindDW**

Kho dữ liệu chúng ta bao gồm NorthWind\_DW (kho dữ liệu chính) và NorthWind\_PT (chứa phần dữ liệu lỗi trong quá trình ETL)

## Tạo project SSIS trong visual studio 2022

**Bước 1: Tiến hành tạo Intergration Services Project:**

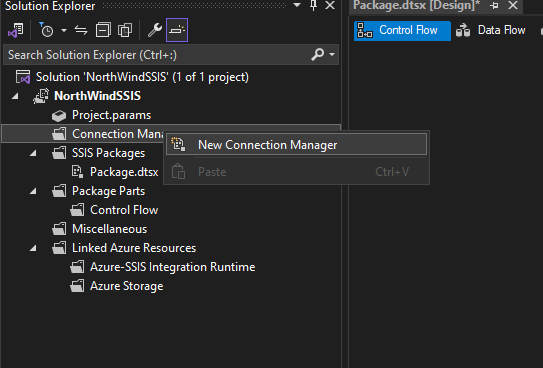
A screenshot of a computer

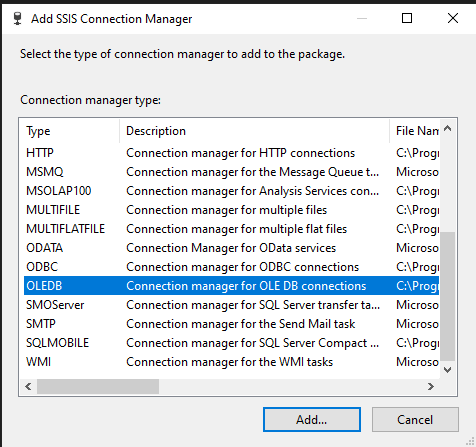
Description automatically generated

A screenshot of a computer program

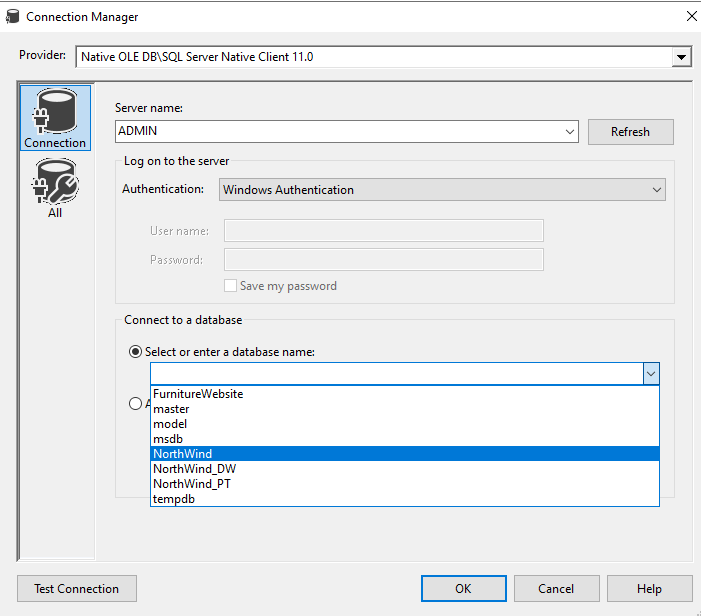
Description automatically generated**Bước 2: Cài đặt thông tin của project để tạo mới**

## Làm sạch dữ liệu và tạo các bảng Dim, Fact

**Bước 1: Tạo kết nối đến các database sử dụng: Chọn Connection Manager trong Solution Explorer → New Connection Manager**

**Bước 2: Chọn loại kết nối OLEDB để có thể kết nối đến SQL Server database. Sau đó chọn Add...**

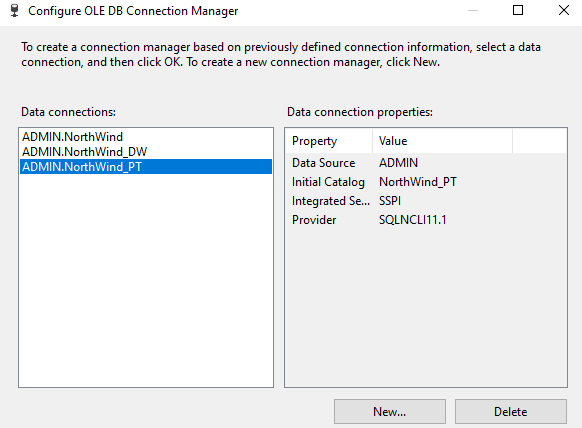
**A screenshot of a computer

Description automatically generatedBước 3: Chọn New → điền tên sever name → chọn database NorthWind**

**Bước 4: Thực hiện tương tự với NorthWindPT và NorthWindDW**

**A screenshot of a computer program

Description automatically generatedA screenshot of a computer program

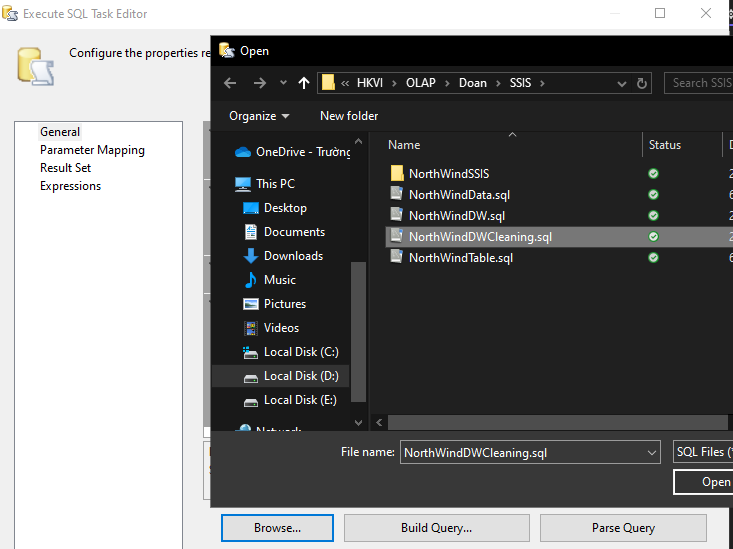
Description automatically generated****Bước 5: Tạo query đặt tên là NorthWindDWCleaning.sql và NorthWindPTCleaning.sql bao gồm các câu lệnh drop tất cả dữ liệu trong các bảng trong NorthWind\_DW và NorthWindPT để làm sạch**

**Bước 6: Chọn Execute SQL Task trên SSIS Toolbox vào khu vực làm việc và nhấn đúp chuột vào nó để thay đổi thông tin**

**Bước 7: Thay đổi thông tin về Name, Description và Connection như hình**

**A screenshot of a computer

Description automatically generated**

**Bước 8: Chọn Broswe... để chọn file NorthWindDWCleaning.sql tải query vào task**

**A screenshot of a computer

Description automatically generatedBước 9: Thực hiện tương tự cho NorthWindPT ta được như sau**

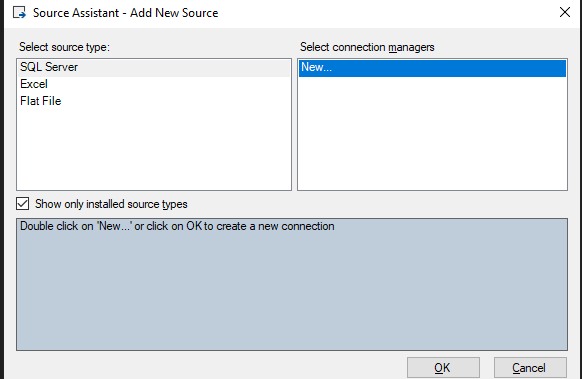
**Bước 10: Chọn Data Flow Task và Sequence Container ở SSIS Toolbox vào khu vực làm việc để bắt đầu quá trình tạo từng bảng DIM**

**A screenshot of a computer

Description automatically generatedBước 11: Ấn vào Data Flow Task để đổi tên thành DIM\_Customer và đặt thiết kế như sau**

**Bước 12: Ấn đúp vào DIM\_Customer để bắt đầu thực hiện việc điều chỉnh đổ dữ liệu**

**Bước 13: Chọn Source Assistant ở SSIS Toolbox để chọn nguồn dữ liệu vào nơi làm việc**

**Bước 14: Chọn New... ở phần Select connection managers trong hộp thoại mới hiện**

**A grey and black text

Description automatically generated**A screenshot of a computer

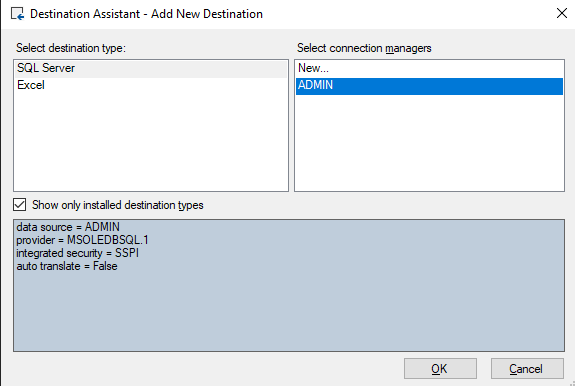
Description automatically generatedA screenshot of a computer

Description automatically generated**Bước 15: Chọn OLEDB Provider và điền tên server  
Bước 15: OLEDB Source xuất hiện, đổi tên thành Customer Table và ấn đúp vào để thay đổi thông tin**

**Bước 16: Chuyển sang Columns và chọn các cột theo thiết kế của bảng DIM\_Customer và ấn OK**

**A grey rectangular sign with black text

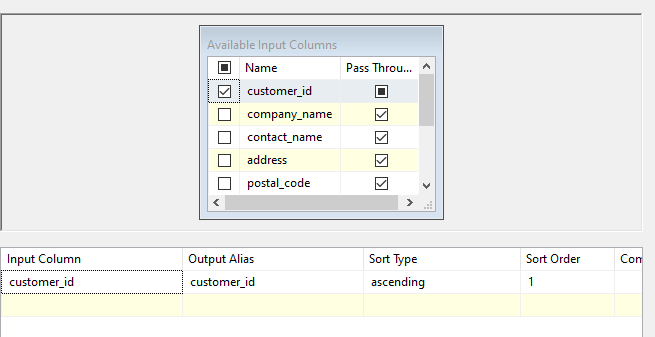
Description automatically generatedA close up of a sign

Description automatically generatedBước 17: Chọn Destination AssistantA screenshot of a computer

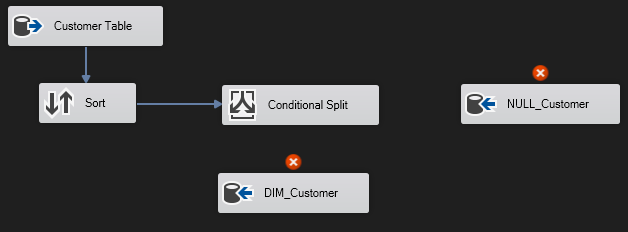
Description automatically generated trên SSIS vào khu vực làm việc, các bước thực hiện gần giống khi chọn Source Assistant để tạo 2 đầu ra dữ liệu**

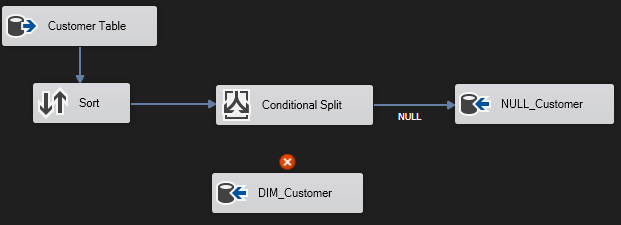
**A screenshot of a computer

Description automatically generatedBước 18: Chọn Sort trên SSIS Toolbox và thực hiện cài đặt như hình**



**A screenshot of a computer

Description automatically generatedBước 19: Chọn Conditional Split trên SSIS và thực hiện cài đặt như hình**

**A screenshot of a computer error

Description automatically generatedBước 20: Kết Conditional Split và NULL\_Location, chọn input phù hợp**

**A screenshot of a computer

Description automatically generatedBước 21: Điều chỉnh NULL\_Location để tiếp nhận dữ liệu: Chọn OLEDB như hình và New... ở phần table để tạo bảng mới để nhận dữ liệu**

**A screenshot of a computer program

Description automatically generated**

**Bước 22: Kết Conditional Split và DIM\_Location và cấu hình tương tự NULL\_Location**

**A screenshot of a computer

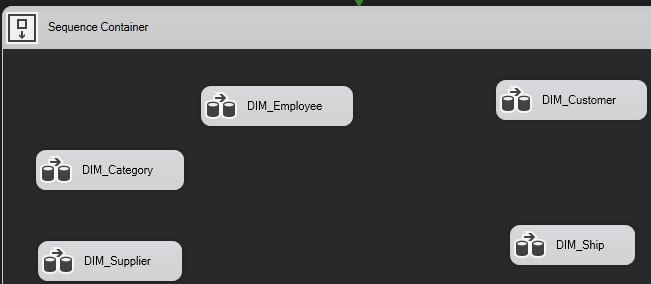
Description automatically generated**

**A screenshot of a computer

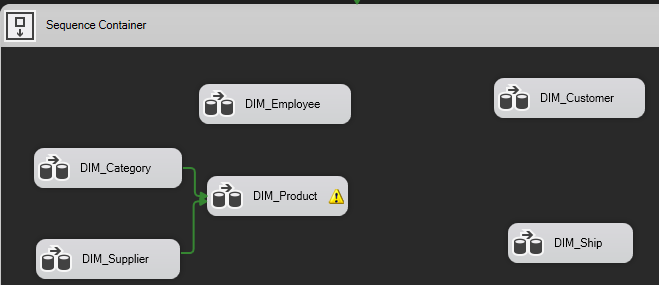
Description automatically generated**

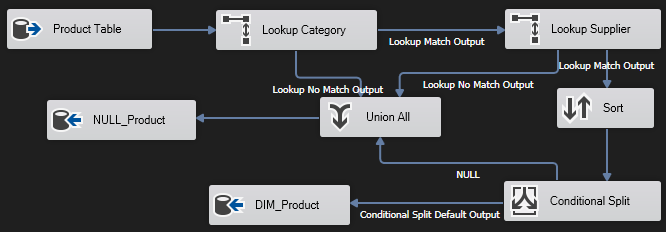
**A screenshot of a computer screen

Description automatically generated**

**Bước 23: Thực hiện cấu hình tương tự lần lượt cho DIM\_Supplier, DIM\_Category, DIM\_Employee, DIM\_Ship**

**Bước 24: Thực hiện cấu hình cho DIM\_Product như sau:**

****

**Bước 25: Cấu hình Lookup dùng để lấy dòng dữ liệu từ nguồn có cột khớp với cột dữ liệu đích nào đó. Bao gồm tạo connection & nối cột so sánh**

****

**A screenshot of a computer

Description automatically generated**

**Bước 26: Cấu hình Union All không cần cấu hình gì thêm**

**A screenshot of a computer

Description automatically generatedBước 27: Thực hiện cấu hình cho DIM\_Date như sau**

A screenshot of a computer

Description automatically generated**A screenshot of a computer

Description automatically generatedA white rectangular object with a black background

Description automatically generated with medium confidenceBước 27: Cấu hình Date Column như sau:**

**Bước 28: Cấu hình Derived Column như sau để trích các loại ngày tháng**

**A screenshot of a computer

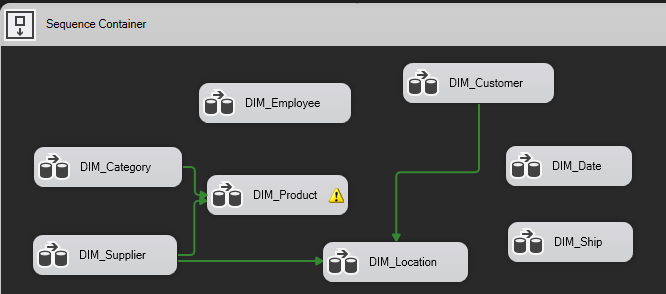
Description automatically generated**

A screenshot of a computer

Description automatically generated**Bước 29: Cấu hình Sort như sau để loại bỏ cột thừa**

**Bước 30: Cấu hình DIM\_Date tương tự với các bảng trước đó**

A screenshot of a computer

Description automatically generated**Bước 31: Cấu hình DIM\_Location như sau**

**Bước 32: Cấu hình Supplier Table và Customer Table: Chọn đúng bảng và chọn cột như sau**

**Bước 33:** A screenshot of a computer

Description automatically generated**Union All không cần cấu hình gì thêm**

**Bước 34: Cấu hình Sort và các thành phần còn lại tương tự các bảng khác**

* **Làm tương tự với các bảng Dim\_Ship, Dim\_Order**.
* **Bước 28**: Kéo data flow task vào màn hình và sửa thành Fact.

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 29**: Edit Fact và kéo OLE DB Source.

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 30**: Edit OLE DB Source chọn connection thích hợp và chọn bảng Amazon\_Clean, và chọn các ô thích hợp.

A screenshot of a computer

Description automatically generated

* **Bước 31**: Kéo chức năng lookup và đổi tên thành Date và edit.

A screenshot of a computer

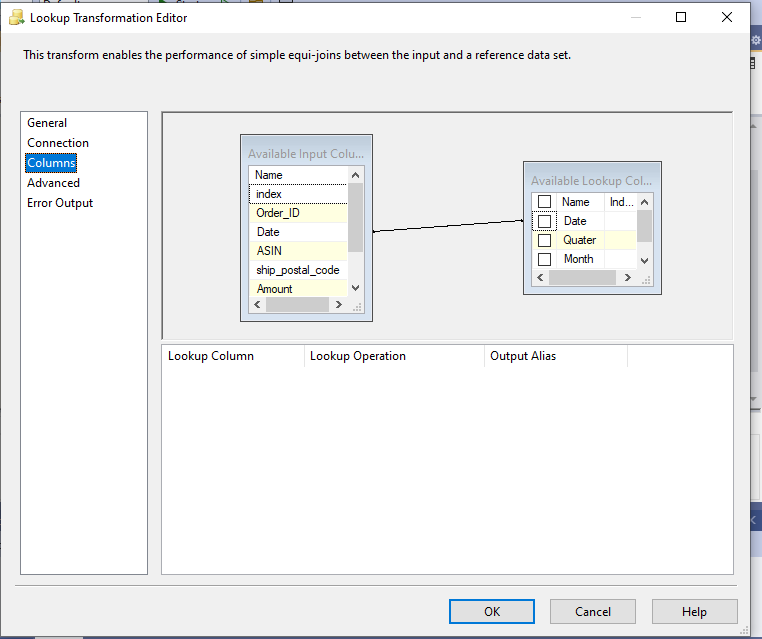
Description automatically generated

* **Bước 32**: Chọn connection thích hợp và con use a table là Dim\_Date.

A screenshot of a computer

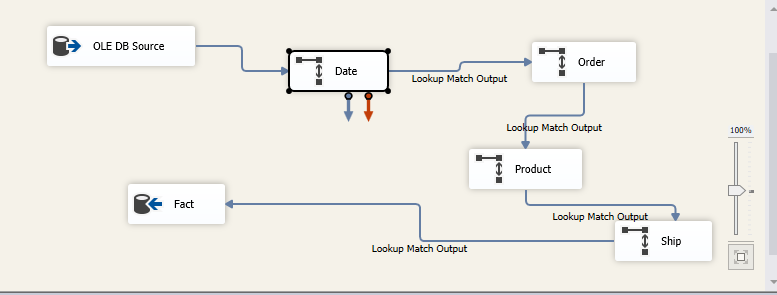
Description automatically generated

* **Bước 33**: Kéo hai cột Date lại với nhau.



**🡪Làm tương tự với các lookup Order, Product, Ship.**

* **Bước 34**: Kéo OLE DB Destination và đổi tên thành Fact.

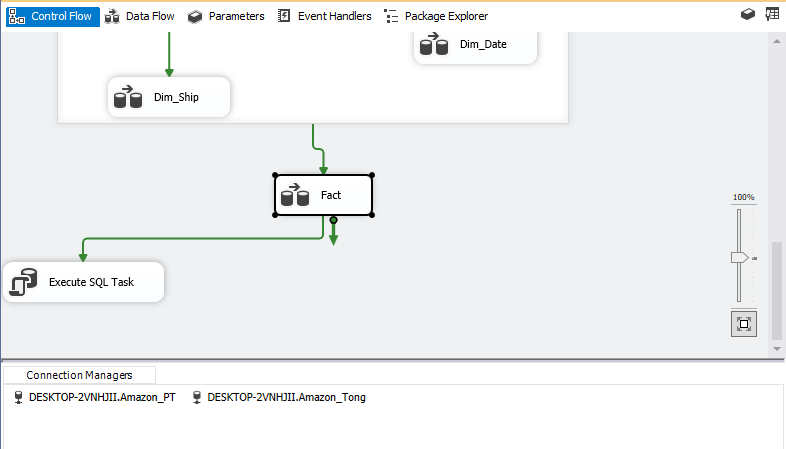


* **Bước 35**: Edit Fact và lựa chọn connection thích hợp, và tạo bảng mới tên Fact.

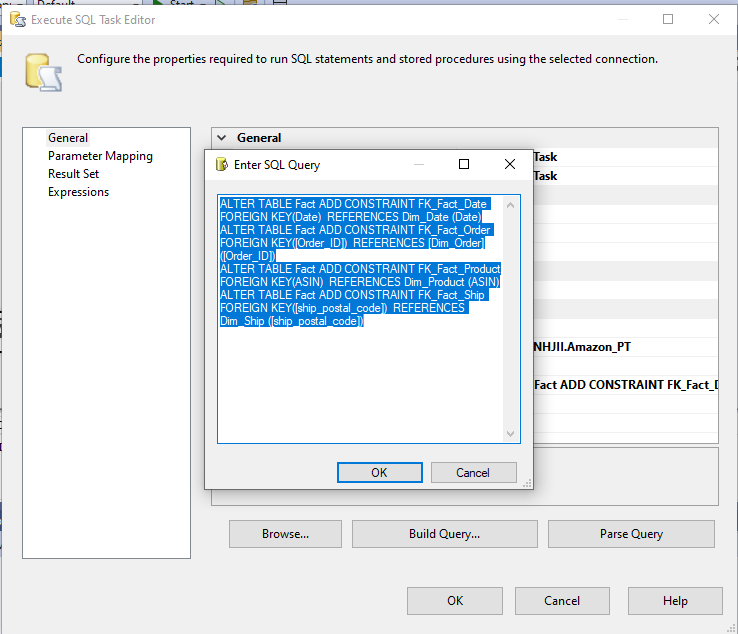
A screenshot of a computer

Description automatically generated

* **Bước 36**: Kéo chức năng excute SQL Task và đổi thành tạo khóa ngoại.



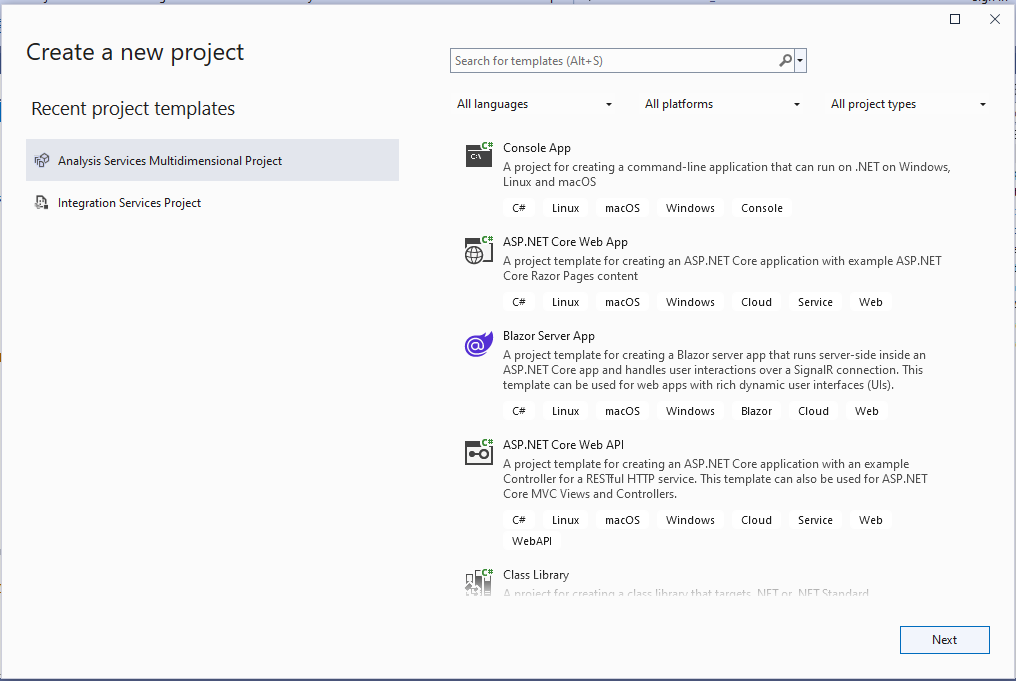
* **Bước 37**: Edit và chọn conneciton thích hợp và điền câu lệnh SQL tạo khóa ngoại.



# CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU TRONG KHO (SSAS)

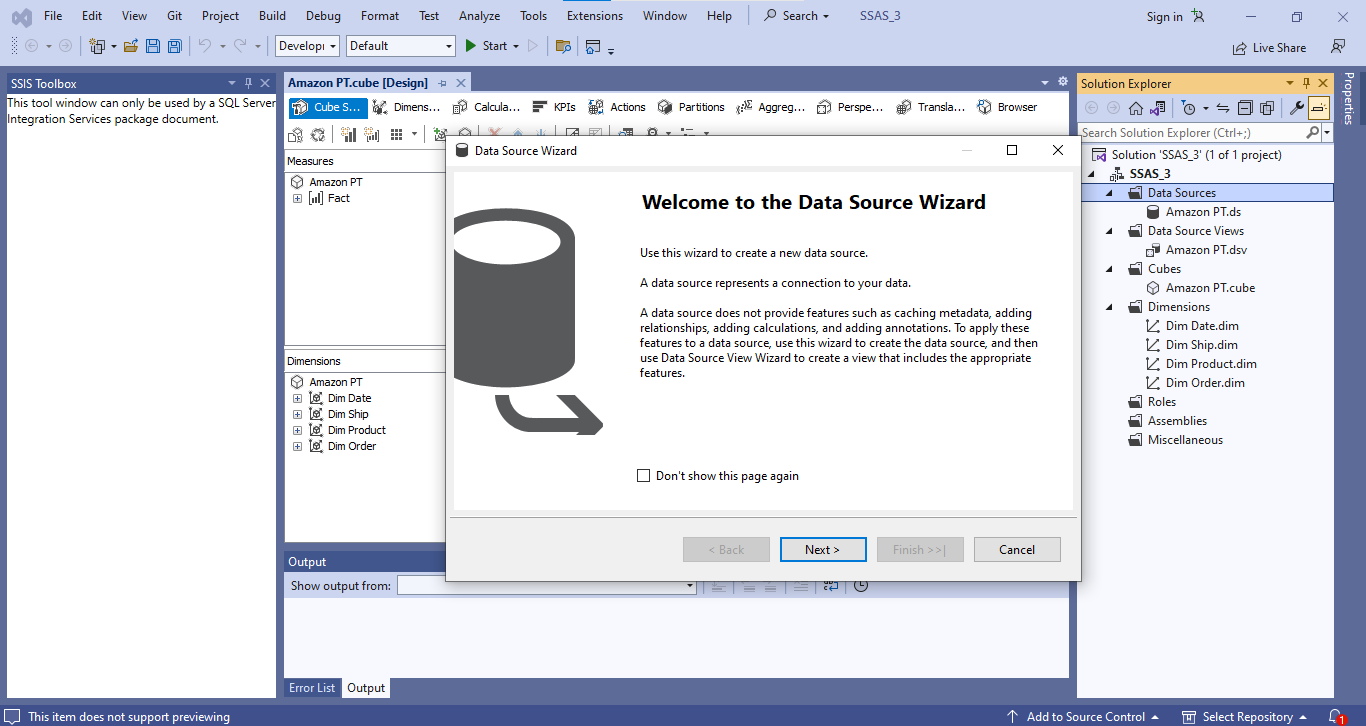
## 1. TẠO PROJECT SSAS TRONG VISUAL STUDIO 2022

* **Bước 1**: Chọn File → New → Project → Chọn Analysis Services như hình:



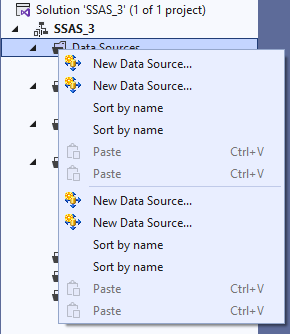
* **Bước 2**: Chọn trong mục Templates chọn Business Intelligence. Sau đó chọn công cụ Analysis

Services chọn đường dẫn để lưu project, đặt tên và nhấn OK. Giao diện quá trình SSAS của công cụ BI hiện ra.

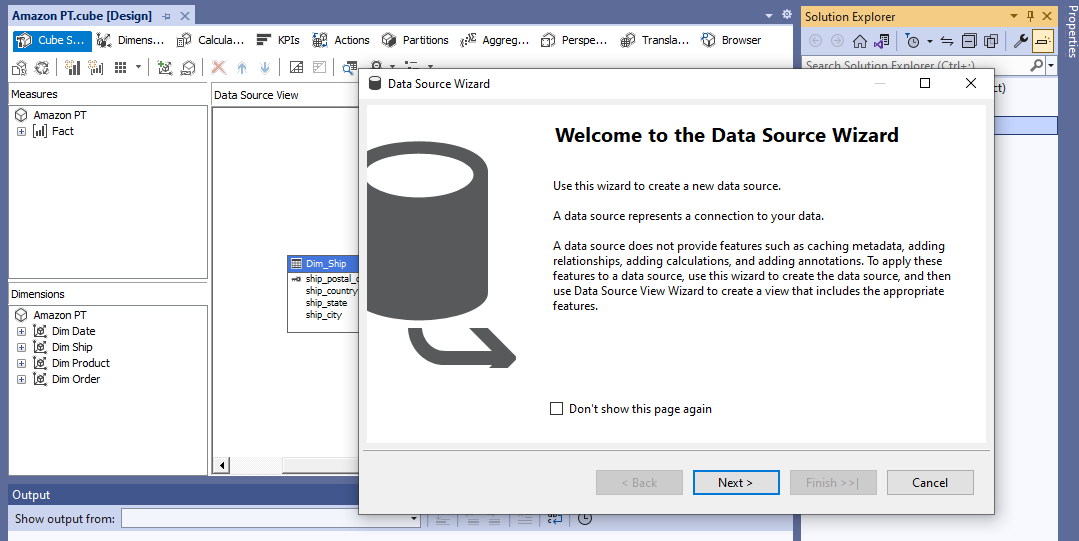


## 2. XÁC ĐỊNH DỮ LIỆU NGUỒN (DEFINE DATE SOURCE)

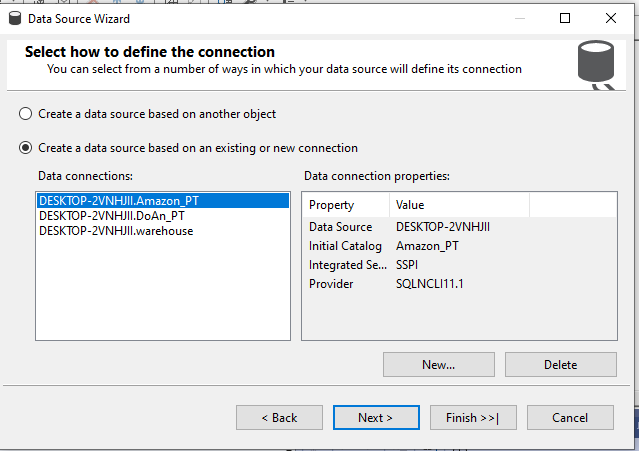
* **Bước 1**: Bên gốc phải phần Solution Explorer nhấn chuột phải vào Data Source.



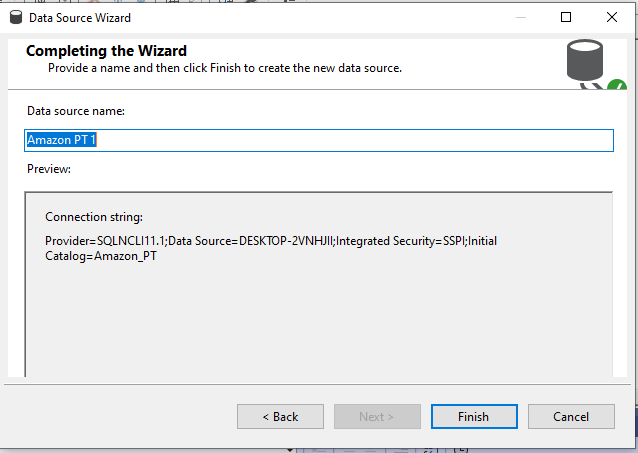
* **Bước 2**: Một cửa sổ hiện ra tiếp tục nhấn Next để tiếp tục.



* **Bước 3**: Chọn kết nối đã tạo từ quá trình SSIS rồi tiếp tục nhấn Next.

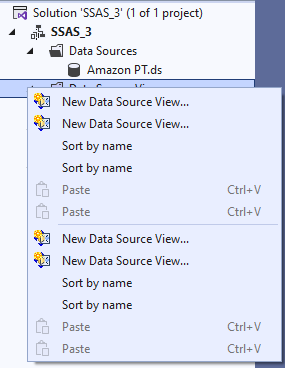


* **Bước 4**: Chọn Use the service account nhấn Next rồi nhấn Finish

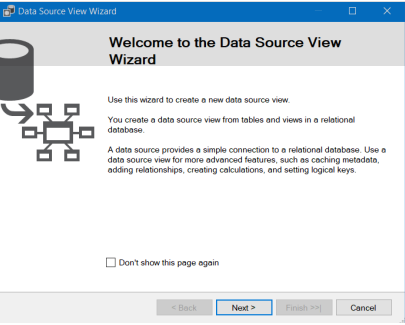


## 3. XÁC ĐỊNH KHUNG NHÌN DỮ LIỆU NGUỒN (DEFINE DATE SOURCE VIEW)

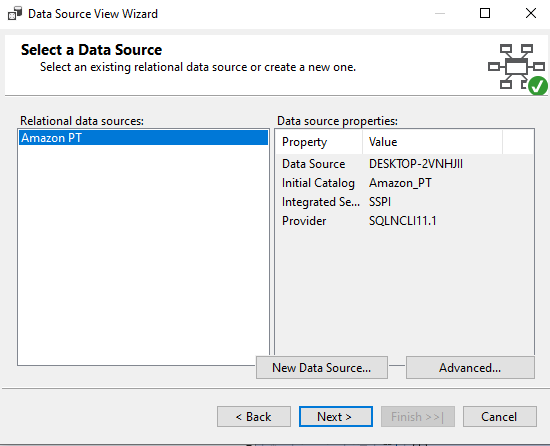
* **Bước 1**: Bên gốc phải phần Solution Explorer nhấn chuột phải vào Data Source View.



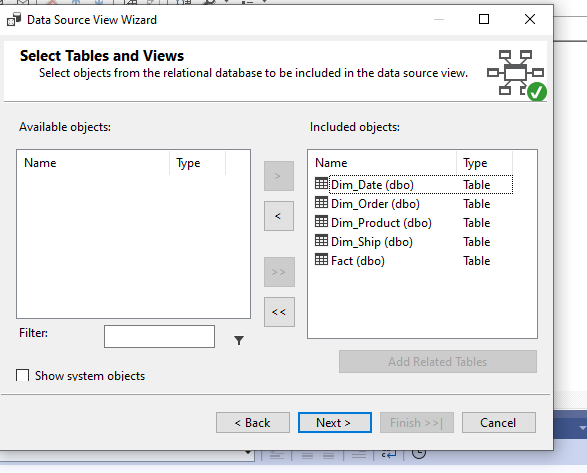
* **Bước 2**: Một cửa sổ hiện ra tiếp tục nhấn Next để tiếp tục.



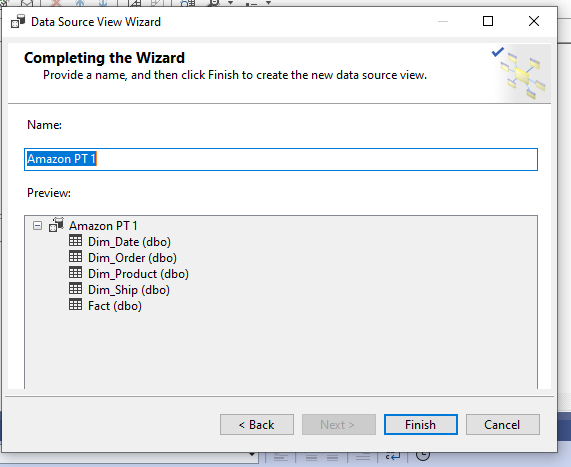
* **Bước 3**: Cho kho dữ liệu Amazon PT rồi nhấn Next để tiếp tục



* **Bước 4**: Chọn bảng FACT và DIM cho quá trình phân tích sau đó nhấn Next.

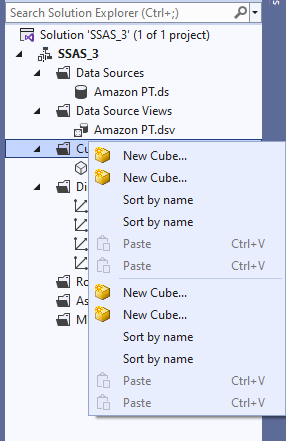


* **Bước 5**: Nhấn Finish để kết thúc quá trình tạo Data View

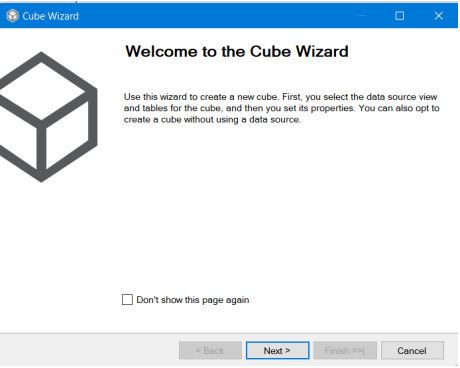


## 4. XÂY DỰNG CÁC KHỐI VÀ CÁC MEASURE

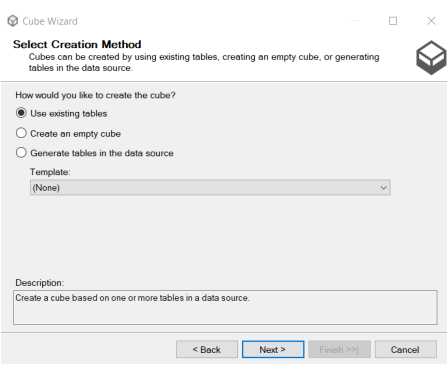
* **Bước 1**: Bên gốc phải phần Solution Explorer nhấn chuột phải vào Cube nhấn tiếp New Cube.



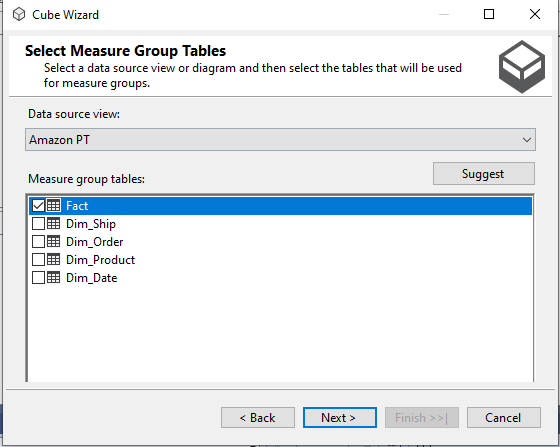
* **Bước 2**: Một cửa sổ hiện ra tiếp tục nhấn Next để tiếp tục.



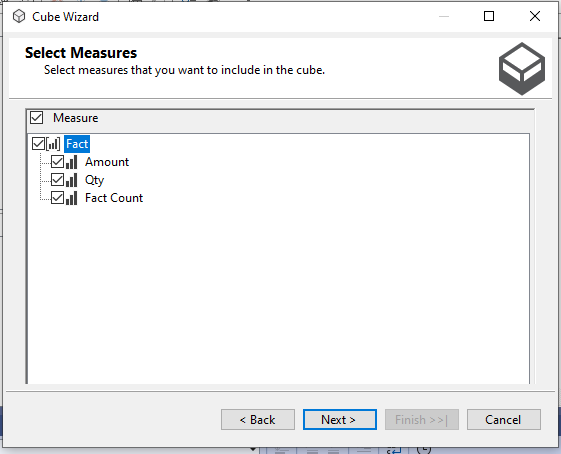
* **Bước 3**: Trong cửa sổ Cube Wizard chọn use an existing table sau đó nhấn Next.



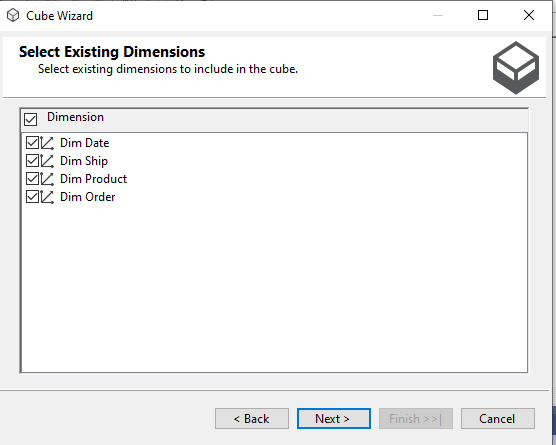
* **Bước 4**: Chọn Fact\_SuperStore để phân chia các measure group



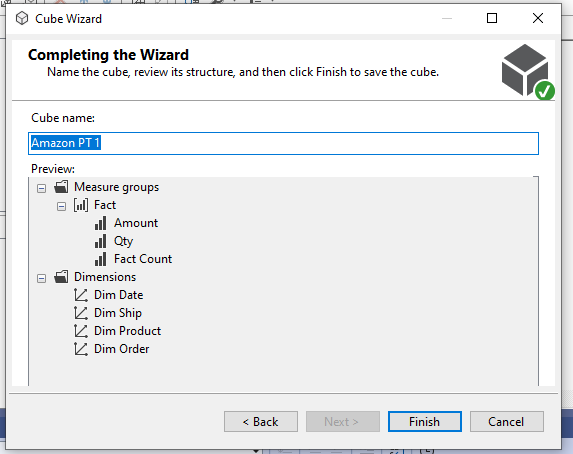
* **Bước 5**: Chọn những thuộc tính định lượng đề xuất. Nhấn Next để tiếp tục.

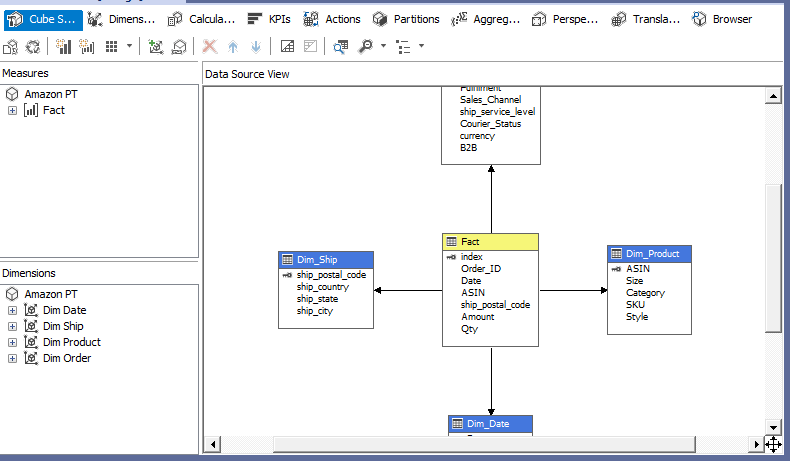


* **Bước 6**: Chọn danh sách Dimension nhấn Next để tiếp tục.



* **Bước 7**: Nhấn Finish để hoàn thành quá trình tạo cube





## PHÂN CẤP TRONG BẢNG CHIỀU

### 5.1 Dim Date

**Bước 1:** Nhấp đúp chuột vào Dim **Date** trong mụcDimensions để tiến hành phân cấp**.**

A screenshot of a computer

Description automatically generated with medium confidence

Bước 2: Thực hiện kéo thả các thuộc tính từ Attributes sang Hierarchies**.**

A screenshot of a computer

Description automatically generated

Bước 3: Chuyển qua tab Atribute Relationships để liên kết các mối liên hệ phân cấp.

A screenshot of a computer

Description automatically generated

Bước 4: Thêm KeyColumns và NameColumns trong Properties của Month.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Bước 5: Thêm KeyColumns và NameColumns trong Properties của Quater.

A screenshot of a computer

Description automatically generated

Bước 6: Thêm KeyColumns và NameColumns trong Properties của Year.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Bước 7: Chỉnh lại Relationship Type thành Rigid.

A screenshot of a computer

Description automatically generated

A picture containing font, screenshot, text, white

Description automatically generated

A screenshot of a computer

Description automatically generated

Bước 8: Nhấp phải chuột vào SSAS\_3 chọn Deploy để tiến hành chạy.

A screenshot of a computer program

Description automatically generated with medium confidence

Thu được kết quả sau:

A screenshot of a computer

Description automatically generated with medium confidence

### Dim Ship

Bước 1: Nhấp đúp chuột vào Dim **Ship** trong mục Dimensions để tiến hành phân cấp.

A screenshot of a computer

Description automatically generated with medium confidence

Bước 2: Thực hiện kéo thả các thuộc tính từ Attributes sang Hierarchies**.**

A screenshot of a computer

Description automatically generated with medium confidence

Bước 3: Chuyển qua tab Atribute Relationships để liên kết các mối liên hệ phân cấp.

A picture containing text, screenshot, line, font

Description automatically generated

Bước 4: Thêm KeyColumns và NameColumns trong Properties của Ship City.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Bước 5: Chỉnh lại Relationship Type thành Rigid.

A screenshot of a computer

Description automatically generated

A close-up of a computer screen

Description automatically generated with low confidence

Bước 6: Nhấp phải chuột vào SSAS\_3 chọn Deploy để tiến hành chạy.

A screenshot of a computer program

Description automatically generated with medium confidence

Thu được kết quả sau:

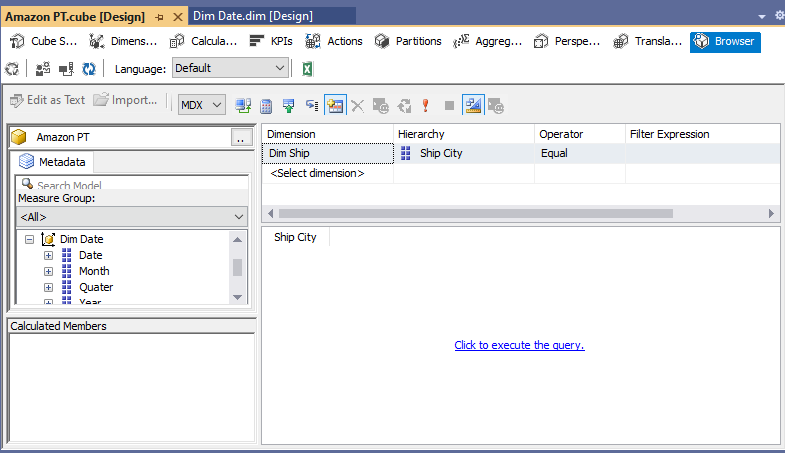
A screenshot of a computer program

Description automatically generated with low confidence

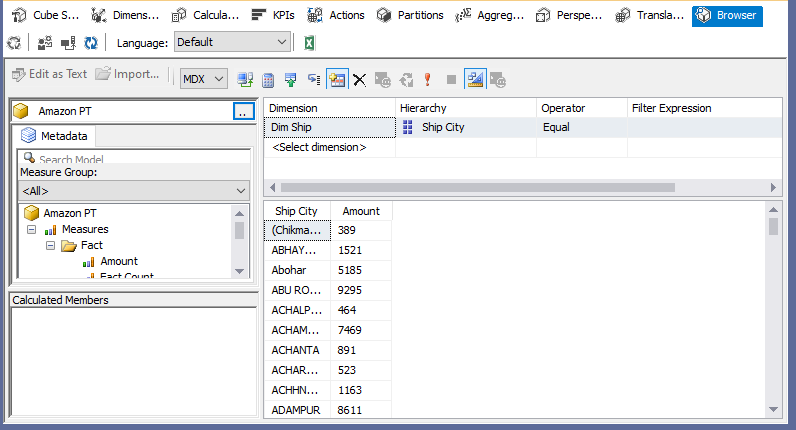
## 5. QUÁ TRÌNH PHÂN TÍCH DỮ LIỆU BẰNG CÔNG CỤ SSAS TRÊN CÁC KHỐI CUBE

### 5.1. Liệt kê tổng tiền mua của các thành phố theo từng tháng

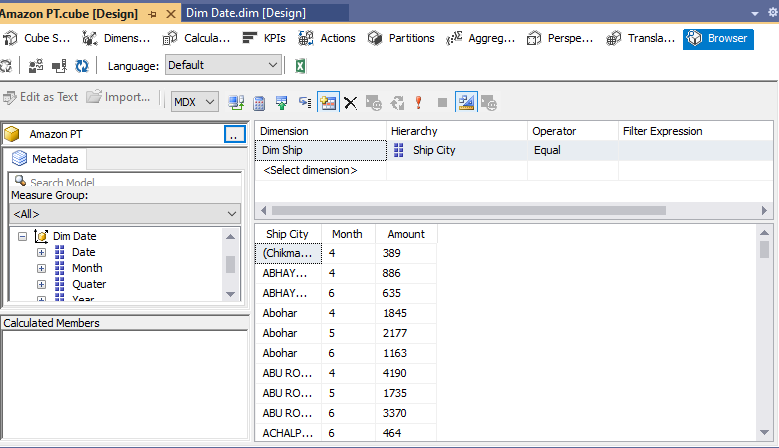
* Ý nghĩa câu truy vấn: Dùng để thống kê doanh số bán hàng ở các thành phố theo từng tháng để biết xem doanh số tăng hay giảm để điều chỉnh chiến lượt kinh doanh tại các thành phố này này.
* Thực hiện quá trình phân tích:
* **Bước 1**: Kéo thuộc tính Ship City của bảng DIM\_Ship sang cửa sổ lọc dữ liệu và cửa sổ thực thi câu truy vấn



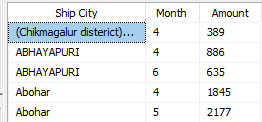
* **Bước 2**: Kéo độ đo Amount của bảng FACT sang cửa sổ thực thi truy vấn



* **Bước 3**: Kéo thuộc tính Year từ bảng DIM\_DATE sang để lấy từng tháng.

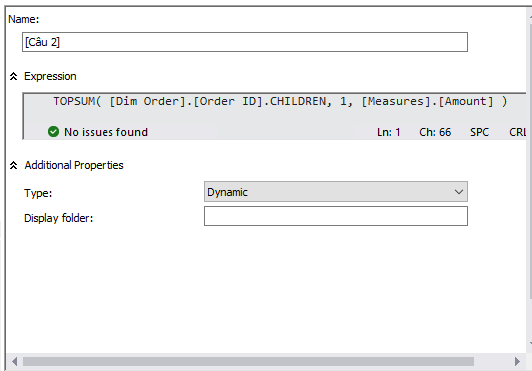


* Kết quả của câu truy vấn

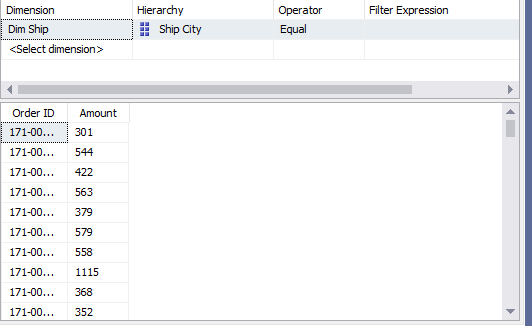


### 5.2. Cho biết hóa đơn đạt tổng số tiền mua lớn nhất

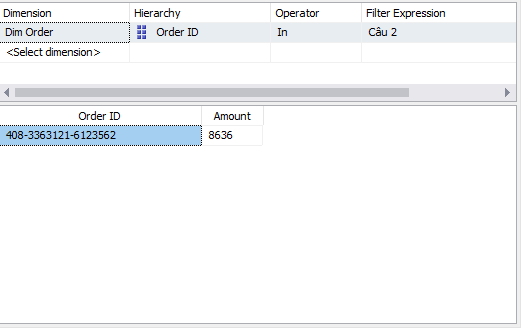
* Ý nghĩa câu truy vấn: Tìm hóa đơn đạt tổng số tiền mua nhiều nhất.
* Thực hiện quá trình phân tích:
* **Bước 1**: Tạo một nameset câu 2 dùng để lấy hóa đơn tổng tiền mua lớn nhất



* **Bước 2**: Kéo thuộc tính Order ID và độ đo Amount sang cửa sổ thực thi truy vấn.



* **Bước 3**: Dùng name set câu 2. Và sẽ có kết quả như hình bên dưới:



### 5.3.Theo từng tháng, quý, năm liệt kê các tổng số lượng mua của từng sản phẩm được mua.

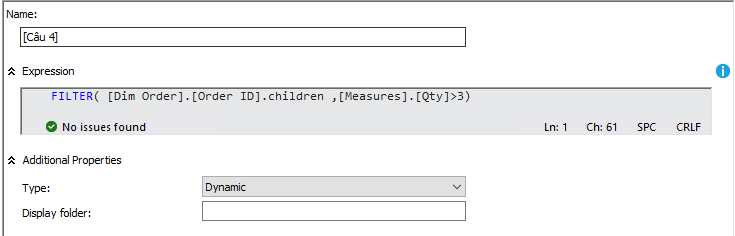
* Ý nghĩa câu truy vấn: Để biết được theo từng tháng quý năm loại các sản phẩm và thể loại sản phẩm được mua để điều chỉnh hướng marketing cho phù hợp.
* Thực hiện câu truy vấn
* Kéo thả các thuộc tính ASIN bảng Dim\_Product và thuộc tính Month, Quarter, Year của bảng Dim\_Date và kéo độ đo Qty sang cửa sổ thực thi truy vấn ta sẽ có được kết quả như hình bên dưới.

A screenshot of a computer

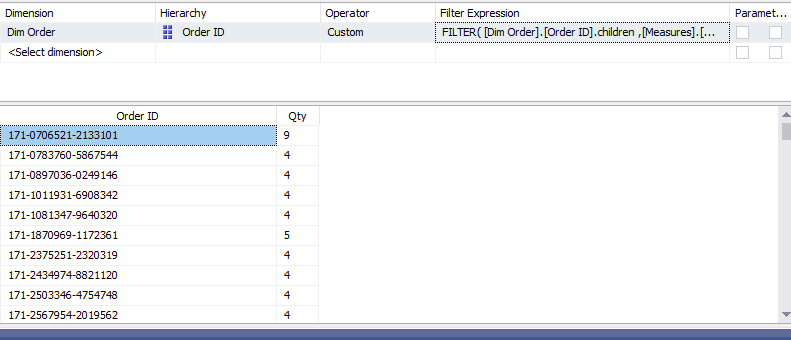
Description automatically generated

### 5.4. Cho biết hóa đơn có tổng số lượng sản phẩm đã mua trên 3 trong tháng 6.

* Ý nghĩa câu truy vấn: Tìm ra những hóa đơn đã mua 3 sản phẩm trở lên.
* Thực hiện quá trình phân tích:
* **Bước 1**: Tạo một nameset câu 4 dùng để lấy tổng số lượng sản phẩm của khách hàng đã mua trên 3 sản phẩm.



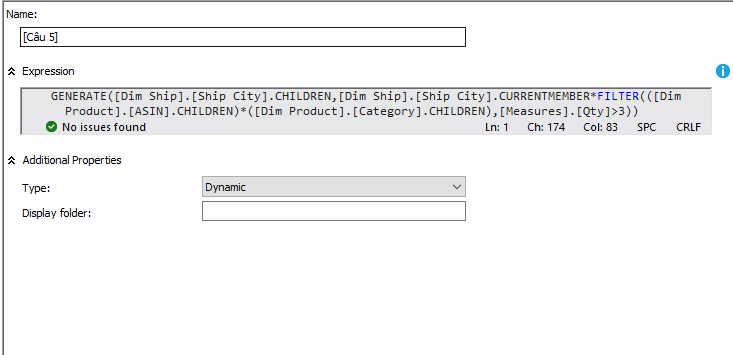
* **Bước 2**: Kéo thuộc tính Order ID và độ đo Qty sang cửa sổ thực thi truy vấn. Và sẽ có kết quả như hình bên dưới.



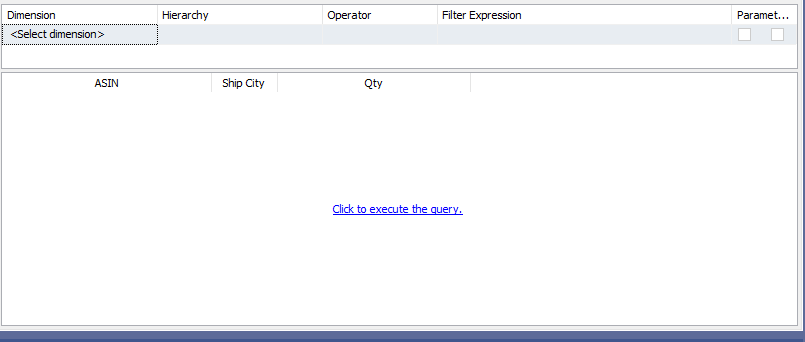
### 5.5. Cho biết tên sản phẩm, loại sản phẩm và tổng số lượng sản phẩm đó lớn hơn 3 được đặt theo từng thành phố.

* Ý nghĩa câu truy vấn: Với từng quốc gia liệt kê những sản phẩm nào đã được bán trên 3.
* Thực hiện câu truy vấn: Với câu truy vấn này không sử dụng công cụ BI và name set bình thường mà kết hợp công cụ BI và một số kiến thức liên quan đến ngôn ngữ MDX.
* Dùng hàm Generate và Current member cho câu hỏi này

GENERATE([Dim Ship].[Ship City].CHILDREN,[Dim Ship].[Ship City].CURRENTMEMBER\*FILTER(([Dim Product].[ASIN].CHILDREN)\*([Dim Product].[Category].CHILDREN),[Measures].[Qty]>3))



* **Bước 1**: Copy hàm trên vào bộ nhớ tạm. Qua brower của công cụ BI kéo thả những thuộc tính cần thiết cho câu truy vấn như hình sau đây:



* **Bước 2**: Chuyển từ Design Mode sang Script Mode bằng cách nhấn chuột vào biểu tượng Design Mode như hình minh họa:

A screenshot of a computer

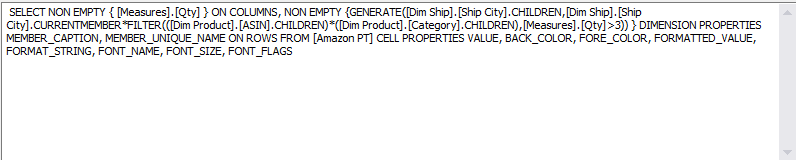
Description automatically generated with medium confidence

* **Bước 3**: Xóa bỏ dòng tô xanh như hình minh họa :

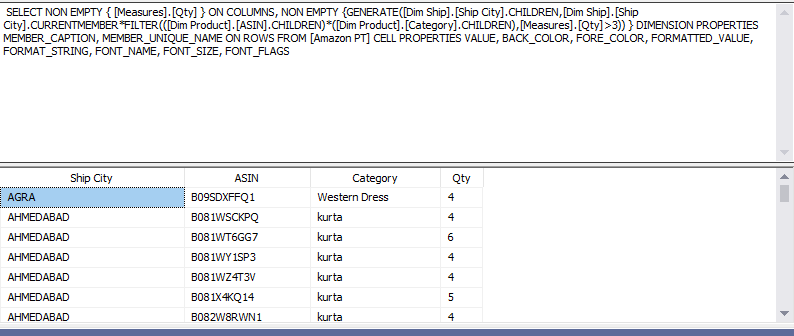
A screen shot of a computer

Description automatically generated with low confidence

* **Bước 4**: Dán đoạn code trên vào dòng màu xanh vừa xóa



* **Bước 5** : Nhấn Execute Query tại cửa sổ thực thi truy vấn ta được kết quả như hình minh họa :



### 5.6. Top 3 sản phẩm có số lượng hóa đơn cao nhất trong tháng 5

* Ý nghĩa câu truy vấn: Dùng để tính tổng lượng hóa đơn của các sản phẩm rồi đưa ra 3 sản phẩm có số lượng hóa đơn nhiều nhất.
* Thực hiện câu truy vấn:
* **Bước 1**: Tạo NameSet [Câu 6] như hình minh họa:

A picture containing text, software, line, font

Description automatically generated

* **Bước 2**: Kéo thả thuộc tính ASIN, bảng DIM\_Product, ShipYear trong bảng DIM\_DATE và độ đo Fact count và lọc theo tháng 5 ta được như hình dưới.

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 3**: Kéo thả Nameset [Câu 6] sang trường lọc ta được kết quả như hình:

A picture containing text, screenshot, font, line

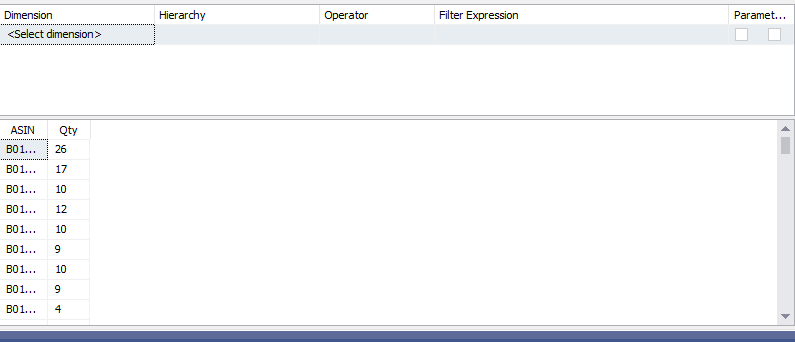
Description automatically generated

### 5.7. Cho biết sản phẩm nào có số lượng mua cao nhất và sản phẩm có số lần mua hàng thấp nhất.

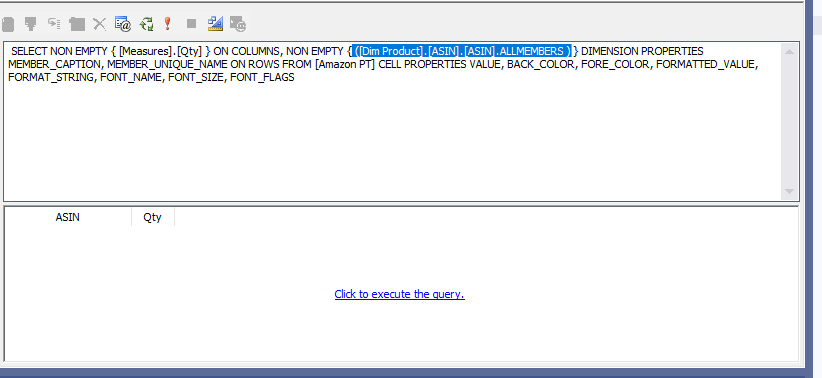
* Ý nghĩa câu truy vấn: Tìm ra sản phẩn được mua nhiều nhất và ít nhất.
* Thực hiện câu truy vấn:
* Dùng hàm Union(TopCount, Bottom Count) cho câu hỏi này



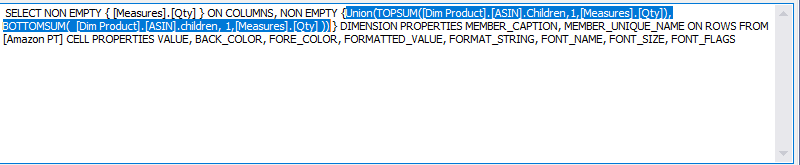
* **Bước 1**: Copy hàm trên vào bộ nhớ tạm. Qua brower của công cụ BI kéo thả những thuộc tính cần thiết cho câu truy vấn như hình sau đây:



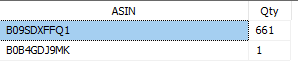
* **Bước 2**: Chuyển từ Design Mode sang Script Mode bằng cách nhấn chuột vào biểu tượng Design Mode
* **Bước 3**: Xóa bỏ dòng tô xanh như hình minh họa :



* **Bước 4**: Dán đoạn code trên vào dòng màu xanh vừa xóa.



* **Bước 5** : Nhấn Execute Query tại cửa sổ thực thi truy vấn ta được kết quả như hình minh họa :



### 5.8 Đưa ra doanh thu, tổng số lượng sản phẩm bán ra của từng thể loại sản phẩm.

* **Bước 1**: Kéo thả các thuộc tính như trong hình.

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 3**: Nhấn execute để hiện kết quả.

A screenshot of a computer

Description automatically generated with medium confidence

### 5.9. Với mỗi tháng đưa ra 3 sản phẩm có lợi nhuận cao nhất.

Thực hiện câu truy vấn: Với câu truy vấn này không sử dụng công cụ BI và name set bình thường mà kết hợp công cụ BI và một số kiến thức liên quan đến ngôn ngữ MDX.

* Dùng hàm Generate và Current member cho câu hỏi này

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 1**: Copy hàm trên vào bộ nhớ tạm. Qua brower của công cụ BI kéo thả những thuộc tính cần thiết cho câu truy vấn như hình sau đây:

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 2**: Chuyển từ Design Mode sang Script Mode bằng cách nhấn chuột vào biểu tượng Design Mode như hình minh họa:



* **Bước 3**: Xóa bỏ dòng tô xanh như hình minh họa :

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 4**: Dán đoạn code trên vào dòng màu xanh vừa xóa

A screen shot of a computer

Description automatically generated with low confidence

* **Bước 5** : Nhấn Execute Query tại cửa sổ thực thi truy vấn ta được kết quả như hình minh họa :

A screenshot of a computer

Description automatically generated with medium confidence

### 5.10. Đưa ra 10 sản phẩm có số lượng hóa đơn thấp nhất nhất tại tỉnh CHHATTISGARH

* **Bước 1**: Xây dựng Name set [Câu 10 ]

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 2**: Deploy Nameset sau đó qua Brower kéo thả những thuộc tính cần thiết cho câu truy vấn 10, và kéo thuộc tính ship state và lọc lấy giá trị CHHATTISGARH ta được kết quả như hình:

A screenshot of a computer

Description automatically generated with medium confidence

## 6. QUÁ TRÌNH PHÂN TÍCH DỮ LIỆU BẰNG CÔNG CỤ PIVOT EXCEL

### 6.1 Liệt kê tổng tiền mua của các thành phố.

* Thực hiện quá trình phân tích dữ liệu:
* **Bước 1**: Kéo thuộc tính Ship city của bảng Dim\_Ship vào cột rows và Amount vào cột values ta được kết quả:

A screenshot of a computer

Description automatically generated

* Nhận xét: Kết quả giống câu 5.1 (Kéo thả bằng công cụ BI)

### 6.2 Cho biết hóa đơn đạt tổng số tiền mua lớn nhất

* **Bước 1**: Kéo các thuộc tính câu 2 vào cột rows và Amount vào cột values như hình:

A screenshot of a computer

Description automatically generated with medium confidence

* Kết quả câu truy vấn

A picture containing text, screenshot, font, line

Description automatically generated

* Nhận xét: Ta thấy kết quả giống với kéo thả 5.2 (BI).

### 6.3 Theo từng tháng, quý, năm liệt kê tổng số lượng các sản phẩm được mua

* **Bước 1**: Kéo thả các hierarchies có tên là time, và thuộc tính ASIN thuộc bảng Dim\_Product vào rows và Time của bảng Dim\_Date vào cột Rows và kéo thả Qty vào values trong pivot table.

A screenshot of a computer

Description automatically generated with medium confidence

* Kết quả:

A screenshot of a computer

Description automatically generated

Nhận xét: Ta thấy kết quả giống với 5.3 (kéo thả BI)

### 6.4 Cho biết hóa đơn có tổng số lượng sản phẩm đã mua trên 3

* **Bước 1**: Kéo thả thuộc name set câu 4 vào rows và Qty vào values trong pivot table

A screenshot of a computer

Description automatically generated with medium confidence

* Kết quả:

A screenshot of a computer

Description automatically generated

* Nhận xét: Giống kết quả 5.4 (kéo thả BI)

### 6.5 Cho biết mã sản phẩm, loại sản phẩm và tổng số lượng sản phẩm đó lớn hơn 3

* **Bước 1**: Kéo thả name set câu 5 và Amount của Fact như hình dưới:

A screenshot of a computer

Description automatically generated with medium confidence

* Kết quả:



* Nhận xét thấy kết quả giống 5.5( kéo thả BI)

### 6.6 Top 3 sản phẩm có số lượng hóa đơn cao nhất trong tháng 5

* **Bước 1**: Kéo name set câu 6 vào rows và Fact count vào values, cùng với độ là kéo Month trong Dim\_Date vào filters và lọc chỉ lấy tháng 5 ta có:

A screenshot of a computer

Description automatically generated with medium confidence

* Kết quả:

A screenshot of a computer

Description automatically generated with medium confidence

* Nhận xét: Ta thấy kết quả giống với 5.6 (kéo thả BI)

### 6.7 Cho biết sản phẩm nào nào có tổng số lượng mua cao nhất và sản phẩm có số lần mua hàng thấp nhất.

* **Bước 1**: Kéo name set câu 7 vào rows và Qty vào values như hình:

A screenshot of a computer

Description automatically generated with medium confidence

* Kết quả:

A screenshot of a computer

Description automatically generated with medium confidence

* Nhận xét: Kết quả giống 5.7 (kéo thả BI)

### 6.8 Đưa ra doanh thu, số lượng sản phẩm bán ra của từng thể loại sản phẩm.

* **Bước 1**: Kéo Amount, Qty vào values, và ASIN của bảng Dim\_Product vào rows ta có :

A screenshot of a computer

Description automatically generated with medium confidence

* Kết quả:

A screenshot of a computer

Description automatically generated with medium confidence

* Nhận xét: Giống 5.8 (kéo thả BI)

### 6.9 Với mỗi tỉnh đưa ra 3 thành phố có lợi nhuận cao nhất.

* **Bước 1**: Kéo name set câu 9 vào rows và Amount vào values ta có:

A screenshot of a computer

Description automatically generated with medium confidence

* Kết quả:

A screenshot of a computer

Description automatically generated with low confidence

Nhận xét: Kết quả giống với 5.9 (kéo thả)

### 6.10 Đưa ra 10 sản phẩm có doanh thu thấp nhất tại tỉnh CHHATTISGARH

* **Bước 1:** Kéo thả name set câu 10 và kéo Amount vào values, kéo Ship city vào filters để lọc lấy tình CHHATTISGARH.

A screenshot of a computer

Description automatically generated with medium confidence

* Kết quả:

A screenshot of a computer

Description automatically generated with medium confidence

* Nhận xét: Kết quả giống với 5.10 (kéo thả BI)

## 7. QUÁ TRÌNH PHÂN TÍCH DỮ LIỆU BẰNG NGÔN NGỮ MDX

### 7.1 Liệt kê tổng tiền mua của các thành phố.

* Câu lệnh truy vấn bằng ngôn ngữ MDX

A black text on a white background

Description automatically generated with low confidence

* Kết quả câu truy vấn

A screenshot of a computer

Description automatically generated with medium confidence

* Nhận xét: Kết quả giống câu 5.1 (Kéo thả bằng công cụ BI) và 6.1 (Công cụ Pivot)

### 7.2 Cho biết hóa đơn đạt tổng số tiền mua lớn nhất.

* Câu truy vấn MDX:

A picture containing text, screenshot, font, line

Description automatically generated

* Kết quả:



* Nhận xét: Kết quả giống câu 6.2 (Kéo thả bằng công cụ BI) và 5.2 (Công cụ Pivot)

### 7.3 Theo từng tháng, quý, năm liệt kê tổng số lượng các loại sản phẩm được mua

Câu truy vấn mdx:

A picture containing text, font, screenshot, line

Description automatically generated

Kết quả:

A picture containing text, screenshot, number, calendar

Description automatically generated

Nhận xét: Kết quả giống 5.3 (kéo thả), 6.3(excel)

### 7.4 Cho biết hóa đơn có tổng số lượng sản phẩm đã mua trên 3

* Câu truy vấn Mdx:

A screenshot of a computer

Description automatically generated with low confidence

* Kết quả câu truy vấn:

A screenshot of a computer

Description automatically generated with medium confidence

* Nhận xét: kết quả giống với 5.4( kéo thả BI) và 6.4 (excel)

### 7.5 Cho biết mã sản phẩm, loại sản phẩm và tổng số lượng sản phẩm đó lớn hơn 3 được đặt theo từng thành phố.

* Câu truy vấn MDX:

A screenshot of a computer code

Description automatically generated with low confidence

* Kết quả:

A screenshot of a computer

Description automatically generated with medium confidence

* Nhận xét: Kết quả giống 5.5( kéo thả BI) và 6.5 (excel)

### 7.6 Top 3 sản phẩm có số lượng hóa đơn cao nhất trong tháng 5

* Câu truy vấn MDX:

A screenshot of a computer program

Description automatically generated with low confidence

* Kết quả:

A screenshot of a computer

Description automatically generated with low confidence

* Nhận xét: Kết quả giống với 5.6 (kéo thả BI) và 6.6 (excel)

### 7.7 Cho biết sản phẩm nào có tổng số lượng mua cao nhất và sản phẩm có số lần mua hàng thấp nhất.

* Câu truy vấn mdx:

A picture containing text, screenshot, font, line

Description automatically generated

* Kết quả:

A screenshot of a computer

Description automatically generated with medium confidence

* Nhận xét: Kết quả giống 5.7 và 6.7

### 7.8 Đưa ra doanh thu, số lượng sản phẩm bán ra của từng thể loại sản phẩm.

* Câu truy vấn mdx:

A screenshot of a computer

Description automatically generated with low confidence

* Kết quả:

A screenshot of a computer

Description automatically generated with low confidence

* Nhận xét: Kết quả giống 6.8 và 5.8

### 7.9 Với mỗi tháng đưa ra 3 sản phẩm có lợi nhuận cao nhất.

* Câu truy vấn mdx:

A screenshot of a computer code

Description automatically generated with low confidence

* Kết quả:

A screenshot of a computer

Description automatically generated with medium confidence

* Nhận xét: Giống kết quả của 5.9 và 6.9

### 7.10 Đưa ra 10 sản phẩm có doanh thu thấp nhất tại tỉnh CHHATTISGARH

* Truy vấn mdx:

A screen shot of a computer

Description automatically generated with low confidence

* Kết quả:

A screenshot of a computer

Description automatically generated with low confidence

* Nhận xét: kết quả giống 5.10 và 6.10

## 8. THỰC HIỆN TRÊN POWER BI

### 8.1 Liệt kê tổng tiền mua của các thành phố

* **Bước 1**: chọn visual dạng bảng và chọn các thuộc tính Ship city của Dim\_Ship và Amount của bảng Fact để tạo report dạng bảng:

A screenshot of a computer

Description automatically generated

* Kết quả:

A screenshot of a table

Description automatically generated with low confidence

* Nhận xét: Ta thấy kết quả giống 5.1, 6.1, 7.1

### 8.2 Cho biết hóa đơn đạt tổng số tiền mua lớn nhất

* **Bước 1**: chọn visual dạng bảng và chọn các thuộc tính Order ID của Dim\_Order và Amount của bảng Fact để tạo report dạng bảng:

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 2**: Trong cột filters on this visual ta mở rộng ô Order ID và chọn Top N tại filter type

A screenshot of a computer

Description automatically generated

* **Bước 3**: Ta điền số 1 vào Show item và kéo thuộc tính Amount của bảng Fact vào By value

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 4**: Nhập apply filter và ta có kết quả:

A picture containing text, font, screenshot, line

Description automatically generated

* Nhận xét: Ta thấy kết quả giống 5.2, 6.2 và 7.2

### 8.3 Theo từng tháng, quý, năm liệt kê tổng số lượng các loại sản phẩm được mua

* **Bước 1**: Chọn visual dạng ma trận và chọn các thuộc tính ASIN và Category của Dim\_Product, hierarchies có tên là Time của Dim\_Datevà Qty của bảng Fact để tạo report dạng bảng:

A screenshot of a computer

Description automatically generated

Kết quả:

A screenshot of a computer

Description automatically generated with low confidence

Nhận xét: Ta thấy giống kết quả của 5.3, 6.3, 7.3

### 8.4 Cho biết hóa đơn có tổng số lượng sản phẩm đã mua trên 3

* **Bước 1**: Chọn visual dạng bảng và chọn các thuộc tính Order ID của Dim\_Order và Qty của bảng Fact để tạo report dạng bảng:

A screenshot of a computer

Description automatically generated

* **Bước 2**: Ở hàng Filter mở rộng ô Qty

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 3**: Ngay phần Show items ta chọn is greater than và điền số 3

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 4**: Nhấn Apply filter sẽ ra kết quả bên dưới:

A screenshot of a computer

Description automatically generated with low confidence

* Nhận xét: Giống kết quả của 5.4, 6.4, 7.4

### 8.5 Cho biết mã sản phẩm, loại sản phẩm và tổng số lượng sản phẩm đó lớn hơn 3

* **Bước 1**: Chọn visual dạng bảng và chọn các thuộc tính ASIN, thuộc tính Category của Dim\_Product, thuộc tính Ship city của Dim\_Ship và Qty của bảng Fact để tạo report dạng bảng:

A screenshot of a computer

Description automatically generated

* **Bước 2**: Ở cột Filters mở rộng ô Qty-> is greater than-> điền số 3

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 4**: Nhấn Apply filter và sort theo ship city ta có được kết quả:

A screenshot of a computer

Description automatically generated with medium confidence

* Nhận xét: Ta thấy kết quả giống 5.5 , 6.5 , 7.5

### 8.6 Top 3 sản phẩm có số lượng hóa đơn cao nhất trong tháng 5

* **Bước 1**: Chọn visual dạng bảng và chọn các thuộc tính ASIN của Dim\_Product, thuộc tính Month của Bảng Dim\_Date và Fact count của bảng Fact để tạo report dạng bảng:

A screenshot of a computer

Description automatically generated

* **Bước 2**: Tại cột Filters ta mở rộng ô Asin và chọn Top N

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 3**: Điền shows item bằng 3 và kéo thuộc tính Fact count vào by value

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 4**: Mở rộng ô Month và điền tích vào tháng 5

A screenshot of a computer

Description automatically generated with medium confidence

* Sau đó ta có kết quả:

A screenshot of a computer

Description automatically generated with medium confidence

* Nhận xét: Ta thấy kết quả giống 5.6, 6.6 ,7.6

### 8.7 Cho biết sản phẩm nào có tổng số lượng mua cao nhất và sản phẩm có số lần mua hàng thấp nhất.

* **Bước 1**: Chọn Getdata->Analysis Services:

A screenshot of a computer

Description automatically generated

* **Bước 2**: Điền tên server cũng như chọn database thích hợp và điện câu truy vấn MDX ở 7.8:

A screenshot of a computer

Description automatically generated

Kết quả:

A picture containing text, screenshot, font, line

Description automatically generated

Nhận xét: Ta thấy kết quả giống 5.7, 6.7, 7.7

### 8.8 Đưa ra doanh thu, số lượng sản phẩm bán ra của từng thể loại sản phẩm.

* **Bước 1**: Chọn visual dạng bảng và chọn các thuộc tính Category của Dim\_Product và Qty, Amount của bảng Fact để tạo report dạng bảng:
* Kết quả:

A screenshot of a computer

Description automatically generated with low confidence

* Nhận xét: Ta thấy kết quả giống 5.8, 6.8, 7.8

### 8.9 Với mỗi tháng đưa ra 3 sản phẩm có lợi nhuận cao nhất.

* **Bước 1**: Chọn Getdata->Analysis Services:

A screenshot of a computer

Description automatically generated with medium confidence

* Bước 2: Điền tên server cũng như chọn database thích hợp và điện câu truy vấn MDX ở 7.9 ta được kết quả:

A screenshot of a computer

Description automatically generated with medium confidence

Nhận xét: Kết quả giống 5.9, 6.9, 7.9

### 8.10 Đưa ra 10 sản phẩm có doanh thu thấp nhất tại tỉnh CHHATTISGARH

* **Bước 1**: Chọn visual dạng bảng và chọn các thuộc tính ASIN của Dim\_Product và Amount của bảng Fact để tạo report dạng bảng:

A screenshot of a computer

Description automatically generated

* **Bước 2**: Tại cột Filters ta mở rộng cột ASIN, chọn bottom 1 và kéo cột Amount vào by value tao được kết quả:

A screenshot of a computer

Description automatically generated with low confidence

* Nhận xét: Kết quả giống với 5.10, 6.10, 7.10

# CHƯƠNG 4: REPORT

## Report với Power BI

### Liệt kê tổng tiền mua của các thành phố

* **Bước 1**: chọn visual dạng cột và chọn các thuộc tính Ship city của Dim\_Ship và Amount của bảng Fact để tạo report dạng Clustered column chart:

A screenshot of a computer

Description automatically generated

* Kết quả:

A picture containing text, screenshot, plot, parallel

Description automatically generated

### 1.2 Cho biết hóa đơn đạt tổng số tiền mua lớn nhất

* **Bước 1**: chọn visual dạng bảng và chọn các thuộc tính Order ID của Dim\_Order và Amount của bảng Fact để tạo report dạng bảng:

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 2**: Trong cột filters on this visual ta mở rộng ô Order ID và chọn Top N tại filter type

A screenshot of a computer

Description automatically generated

* **Bước 3**: Ta điền số 1 vào Show item và kéo thuộc tính Amount của bảng Fact vào By value

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 4**: Nhập apply filter và ta có kết quả:

A picture containing text, font, screenshot, line

Description automatically generated

* Nhận xét: Ta thấy kết quả giống 5.2, 6.2 và 7.2

### 1.3 Theo từng tháng, quý, năm liệt kê tổng số lượng các sản phẩm được mua

* **Bước 1**: Chọn visual dạng ma trận và chọn các thuộc tính ASINcủa Dim\_Product, hierarchies có tên là Time của Dim\_Date và Qty của bảng Fact để tạo report dạng matrix:

A screenshot of a computer

Description automatically generated with medium confidence

Kết quả:

A screenshot of a computer screen

Description automatically generated with low confidence

### 1.4 Cho biết hóa đơn có tổng số lượng sản phẩm đã mua trên 3

* **Bước 1**: Chọn visual dạng cột và chọn các thuộc tính Order ID của Dim\_Order và Qty của bảng Fact để tạo report dạng line chart:

A screenshot of a computer

Description automatically generated

* **Bước 2**: Ở hàng Filter mở rộng ô Qty

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 3**: Ngay phần Show items ta chọn is greater than và điền số 3

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 4**: Nhấn Apply filter sẽ ra kết quả bên dưới:

A screenshot of a graph

Description automatically generated with low confidence

### 1.5 Cho biết mã sản phẩm, loại sản phẩm và tổng số lượng sản phẩm đó lớn hơn 3

* **Bước 1**: Chọn visual dạng ma trận và chọn các thuộc tính ASIN, thuộc tính Category của Dim\_Product, thuộc tính Ship city của Dim\_Ship và Qty của bảng Fact để tạo report dạng Funnel:

A screenshot of a computer

Description automatically generated

* **Bước 2**: Ở cột Filters mở rộng ô Qty-> is greater than-> điền số 3

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 4**: Nhấn Apply filter và sort theo ship city ta có được kết quả:

A screenshot of a computer

Description automatically generated with medium confidence

### 1.6 Top 3 sản phẩm có số lượng hóa đơn cao nhất trong tháng 5

* **Bước 1**: Chọn visual dạng cột ngang và chọn các thuộc tính ASIN của Dim\_Product, thuộc tính Month của Bảng Dim\_Date và Fact count của bảng Fact để tạo report dạng pie chart:

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 2**: Tại cột Filters ta mở rộng ô Asin và chọn Top N

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 3**: Điền shows item bằng 3 và kéo thuộc tính Fact count vào by value

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 4**: Mở rộng ô Month và điền tích vào tháng 5

A screenshot of a computer

Description automatically generated with medium confidence

* Sau đó ta có kết quả:

A screenshot of a graph

Description automatically generated with low confidence

### 1.7 Cho biết sản phẩm nào nào có tổng số lượng mua cao nhất và sản phẩm có số lần mua hàng thấp nhất.

* **Bước 1**: Get data và chọn analysis.

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 2**: Chọn import và dán câu truy vấn mdx câu 7 vào và chỉnh lại các thuộc tính cột Qty cùng với đó là chọn biểu đồ cột ta có kết quả:

A screenshot of a computer

Description automatically generated

Bước 3: Chọn biểu đồ cột và chọn các thuộc tính như trong hình

A screenshot of a computer

Description automatically generated with medium confidence

Kết quả:

A picture containing text, screenshot, software, display

Description automatically generated

### 1.8 Đưa ra doanh thu, số lượng sản phẩm bán ra của từng thể loại sản phẩm.

* **Bước 1**: Chọn visual dạng bảng và chọn các thuộc tính Category của Dim\_Product và Qty, Amount của bảng Fact để tạo report dạng bảng:
* Kết quả:

A picture containing text, screenshot, plot, number

Description automatically generated

### 1.9 Với mỗi tháng đưa ra 3 sản phẩm có lợi nhuận cao nhất.

* **Bước 1**: Get data và chọn analysis.

A screenshot of a computer

Description automatically generated with medium confidence

* **Bước 2**: Chọn import và dán câu truy vấn mdx câu 7 và chỉnh lại đơn vị của Amount:

A screenshot of a computer

Description automatically generated

Bước 3: Chọn biểu đồ cột và tích câu query ta có kết quả:

A picture containing text, screenshot, font, number

Description automatically generated

### 1.10 Đưa ra 10 sản phẩm có doanh thu thấp nhất tại tỉnh CHHATTISGARH

* **Bước 1**: Chọn visual dạng bảng và chọn các thuộc tính ASIN của Dim\_Product và Amount của bảng Fact để tạo report dạng bảng:

A screenshot of a computer

Description automatically generated

* **Bước 2**: Tại cột Filters ta mở rộng cột ASIN, chọn bottom 1 và kéo cột Amount vào by value tao được kết quả:

A screenshot of a computer

Description automatically generated with low confidence

## Report với Excel

### 2.1 Liệt kê tổng tiền mua của các thành phố.

* Thực hiện quá trình phân tích dữ liệu:
* **Bước 1**: Kéo thuộc tính Ship city của bảng Dim\_Ship và thuộc tính Month của bảng Dim\_Date vào cột rows và Amount vào cột values ta được:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Bước 2: Vào insert, chọn chart thích hợp ta có:

A screenshot of a graph

Description automatically generated with medium confidence

### 2.2 Cho biết hóa đơn đạt tổng số tiền mua lớn nhất

* **Bước 1**: Kéo các thuộc tính câu 2 vào cột rows và Amount vào cột values như hình:

A screenshot of a computer

Description automatically generated with medium confidence

* Kết quả câu truy vấn

A picture containing text, screenshot, font, line

Description automatically generated

### 2.3 Theo từng tháng, quý, năm liệt kê tổng số lượng các sản phẩm được mua

* **Bước 1**: Kéo thả các hierarchies có tên là time, và thuộc tính Time thuộc bảng Dim\_Date vào rows và Category vào cột Columns và kéo thả Qty vào values trong pivot table.

A screenshot of a computer

Description automatically generated with medium confidence

* Kết quả:

A screenshot of a computer

Description automatically generated with medium confidence

### 2.4 Cho biết hóa đơn có tổng số lượng sản phẩm đã mua trên 3

* **Bước 1**: Kéo thả thuộc name set câu 4 vào rows và Qty vào values trong pivot table

A screenshot of a computer

Description automatically generated with medium confidence

Bước 2: Insert vào biểu đồ chart line có kết quả:

A picture containing text, screenshot, font, plot

Description automatically generated

### 2.5 Cho biết mã sản phẩm, loại sản phẩm và tổng số lượng sản phẩm đó lớn hơn 3

* **Bước 1**: Kéo thả name set câu 5 và Amount của Fact như hình dưới:

A screenshot of a computer

Description automatically generated with medium confidence

Kết quả:

A screenshot of a computer

Description automatically generated with medium confidence

### 2.6 Top 3 sản phẩm có số lượng hóa đơn cao nhất trong tháng 5

* **Bước 1**: Kéo name set câu 6 vào rows và Fact count vào values, cùng với độ là kéo Month trong Dim\_Date vào filters và lọc chỉ lấy tháng 5 ta có:

A screenshot of a computer

Description automatically generated with medium confidence

* Bước 2: Vào insert chọn Columns ta có kết quả:

A picture containing text, screenshot, number, plot

Description automatically generated

### 2.7 Cho biết sản phẩm nào nào có tổng số lượng mua cao nhất và sản phẩm có số lần mua hàng thấp nhất.

* **Bước 1**: Kéo name set câu 7 vào rows và Qty vào values như hình:

A screenshot of a computer

Description automatically generated with medium confidence

Bước 2: Chọn insert cột ta có kết quả:

A picture containing text, screenshot, number, plot

Description automatically generated

### 2.8 Đưa ra doanh thu, số lượng sản phẩm bán ra của từng thể loại sản phẩm.

* **Bước 1**: Kéo Amount, Qty vào values, và ASIN của bảng Dim\_Product vào rows ta có :

A screenshot of a computer

Description automatically generated with medium confidence

* Bước 2: Vào insert ta chọn biểu đồ cột ta có:

A picture containing text, screenshot, number, font

Description automatically generated

### 2.9 Với mỗi tháng đưa ra 3 sản phẩm có lợi nhuận cao nhất.

* **Bước 1**: Kéo name set câu 9 vào rows và Amount vào values ta có:

A screenshot of a computer

Description automatically generated with medium confidence

Bước 2: Vào insert chọn biểu đồ cột và chọn các thuộc tính như hình ta có:  
A picture containing text, screenshot, number, diagram

Description automatically generated

### 2.10 Đưa ra 10 sản phẩm có doanh thu thấp nhất tại tỉnh CHHATTISGARH

* **Bước 1:** Kéo thả name set câu 10 và kéo Amount vào values, kéo Ship city vào filters để lọc lấy tình CHHATTISGARH.

A screenshot of a computer

Description automatically generated with medium confidence

* Kết quả:

A screenshot of a computer

Description automatically generated with low confidence

# CHƯƠNG 5: QUÁ TRÌNH DATA MINING

### Chuẩn bị dữ liệu

* Bước 1: Download dataset Amazon Sale Report.csv tại link sau: <https://www.kaggle.com/datasets/thedevastator/unlock-profits-with-e-commerce-sales-data?select=Amazon+Sale+Report.csv&fbclid=IwAR1NEkEIskwi0QyXnnVOIceP1NGWmLT5NyVpcTtbW9wJ_hurh7kNzJPXBjQ>.
* Bước 2: Download file order-rejection-classification.ipynb tại link sau( lưu ý phải để file ipynb và csv chung 1 folder):

### Triển khai data mining

Bước 1: Mở visual studio code lên và mở file order-rejection-classification.ipynb.

A screenshot of a computer

Description automatically generated

Bước 2: Bấm Run All để chạy file.

A screenshot of a computer

Description automatically generated with medium confidence

### Phân tích kết quả

Giải thích các từ ngữ:

Accuracy Score (Độ chính xác): Độ chính xác được tính bằng cách so sánh kết quả dự đoán của mô hình với kết quả thực tế trên dữ liệu huấn luyện.

Precision Score (Độ chính xác của dương tính): Precision Score đo lường khả năng của mô hình nhận diện chính xác các điểm dữ liệu thuộc lớp positive (positive class).

Recall Score (Độ nhạy): Recall Score đo lường khả năng của mô hình tìm ra tất cả các điểm dữ liệu thuộc lớp positive trong tập huấn luyện.

#### 3.1 Logistic Regression Classifier

1. Dựa vào ma trận nhầm lẫn

A picture containing text, screenshot, software, rectangle

Description automatically generated

A picture containing text, font, screenshot, typography

Description automatically generated

Ta thấy:

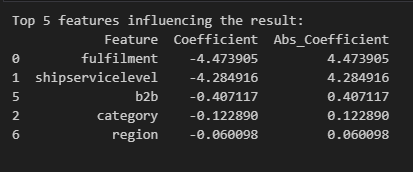
Accuracy Score (Độ chính xác): Trong trường hợp này, độ chính xác của mô hình được tính là 80.89%. Điều này có nghĩa là mô hình dự đoán đúng khoảng 80.89%.

Precision Score (Độ chính xác của dương tính): Trong trường hợp này, Precision Score đạt 94.83%. Điều này có nghĩa là mô hình đánh giá chính xác 94.83% điểm dữ liệu được dự đoán là positive.

Recall Score (Độ nhạy): Trong trường hợp này, Recall Score đạt 65.33%. Điều này có nghĩa là mô hình chỉ nhận diện được 65.33% các điểm dữ liệu thuộc lớp positive.

Các điểm số đánh giá hiệu suất này cho thấy mô hình Logistic Regression đã đạt được một độ chính xác tổng thể tương đối cao trên dữ liệu (80.89%). Tuy nhiên, mô hình có thể cần cải thiện khả năng nhận diện đầy đủ các điểm dữ liệu positive (độ nhạy thấp) trong khi vẫn duy trì độ chính xác của dương tính cao.

1. Top 5 thuộc tính hàng đầu ảnh hưởng đến quyết định đối với việc từ chối đơn hàng dựa trên phân loại hồi quy logistic.



Như ta thấy kết quả trên ta có nhận xét:

Fulfilment và shipservicelevel có giá trị tuyết đối trọng số (coefficient) lớn nên có ảnh hưởng đáng kể đến kết quả dự đoán của mô hình.

1. Luật

A picture containing text, screenshot, diagram, number

Description automatically generated

Từ hình vẽ trên ta có một vài luật đưa ra:

+ Nếu như đơn hàng có thuộc tính Fulfilment có giá trị là Amazon và ship-service-level là Expedited thì đơn hàng có xu hướng bị hủy.

+ Nếu như đơn hàng có có thuộc tính Fulfilment có giá trị là Merchant và ship-service-level là Standard thì đơn hàng có xu hướng bị hủy.

#### 3.2 Random Forest Classifier

1. Ma trận nhậm lẫn

A black background with white text

Description automatically generated with low confidence

Accuracy Score (Độ chính xác): Kết quả là 87.28%, cho thấy mô hình đã dự đoán đúng khoảng 87.28% dữ liệu huấn luyện.

Precision Score (Độ chính xác của dương tính): Kết quả là 96.39%, cho thấy mô hình có khả năng dự đoán chính xác khoảng 96.39% các trường hợp thuộc nhãn dương tính.

Recall Score (Độ nhạy của dương tính): Kết quả là 93.72%, cho thấy mô hình có khả năng dự đoán chính xác khoảng 93.72% các trường hợp thực tế thuộc nhãn dương tính.

Kết quả này cho thấy mô hình Random Forest Classifier đã đạt được mức độ chính xác cao và khả năng phân loại tốt trong việc dự đoán các trường hợp thuộc nhãn dương tính trên tập dữ liệu huấn luyện.

Trên tập test:

A picture containing text, screenshot, rectangle, software

Description automatically generated

A picture containing text, font, screenshot, white

Description automatically generated

Accuracy Score (Độ chính xác): Kết quả là 81.67%, cho thấy mô hình đã dự đoán đúng khoảng 81.67% dữ liệu kiểm tra.

Precision Score (Độ chính xác của dương tính): Kết quả là 95.00%, cho thấy mô hình có khả năng dự đoán chính xác khoảng 95.00% các trường hợp thuộc nhãn dương tính trên dữ liệu kiểm tra.

Recall Score (Độ nhạy của dương tính): Kết quả là 66,87%, cho thấy mô hình có khả năng dự đoán chính xác khoảng 66,87% các trường hợp thực tế thuộc nhãn dương tính trên dữ liệu kiểm tra.

Kết quả trên cho thấy mô hình Random Forest Classifier có khả năng dự đoán đúng một phần lớn các trường hợp trên tập dữ liệu. Tuy nhiên, độ chính xác, độ chính xác của dương tính và độ nhạy của dương tính trên dữ liệu kiểm tra thường thấp hơn so với trên dữ liệu huấn luyện. Điều này có thể cho thấy mô hình có một mức độ overfitting (quá khớp) nhất định trên tập dữ liệu huấn luyện. Để cải thiện hiệu suất của mô hình trên dữ liệu kiểm tra, có thể thực hiện việc tinh chỉnh (tuning) các tham số hoặc sử dụng các phương pháp kiểm soát overfitting như cross-validation.

1. Top 7 thuộc tính hàng đầu ảnh hưởng đến quyết định đối với việc từ chối đơn hàng dựa trên phân loại hồi quy logistic.

A picture containing text, screenshot, diagram, plot

Description automatically generated

Thuộc tính amount, Fulfilment, shipservicelevel có mức độ quan trọng khá cao trong mô hình. Điều này có thể đồng nghĩa với việc 3 thuộc tính này có ảnh hưởng lớn đến quyết định của mô hình phân loại.

1. Luật

Ta có một nhánh của random forest được vẽ bằng cây quyết định:

A picture containing text, font, screenshot, diagram

Description automatically generated

Từ cây quyết định trên ta có thể rút ra được một số luật sau:

* Nếu như đơn hàng có thuộc tính Fulfilment không có giá trị là Merchant và thuộc tính ship-service-level không có giá trị là Standard và giá trị của đơn hàng từ 0.5 trở xuống thì đơn hàng đó sẽ có xu hướng bị hủy
* Nếu như đơn hàng có thuộc tính Fulfilment không có giá trị là Merchant và thuộc tính ship-service-level không có giá trị là Standard và giá trị của đơn hàng lớn hơn 0.5 thì đơn hàng đó sẽ có xu hướng không bị hủy.
* Nếu như đơn hàng có thuộc tính Fulfilment có giá trị là Merchant và Amount lớn hơn 33.5 và nhỏ hơn hoặc bằng 733.5 thì đơn hàng sẽ có xu hướng không bị hủy.

#### 3.3 Nhận xét tổng quan về hai thuật toán

Với kết quả:

A picture containing text, font, screenshot, number

Description automatically generated

Nhận xét:

Cả hai mô hình đều có độ chính xác khá cao, cho thấy chúng có thể dự báo đúng cho một phần lớn dữ liệu.

Logistic Regression cho thấy độ chính xác cao hơn so với Random Forest.

Logistic Regression cũng có độ chính xác dự báo cao hơn, có nghĩa là nó có tỷ lệ sai dương thấp hơn.

Random Forest, ngược lại, cho thấy tỷ lệ bắt được (recall) cao hơn một chút, cho thấy nó có thể xác định được nhiều giá trị dương thực sự hơn.

Tóm lại, dựa trên các đánh giá hiệu suất, Logistic Regression có hiệu suất tốt hơn so với Random Forest.

Ngoài ra, chúng tôi có thể kết luận rằng các đặc điểm quan trọng ảnh hưởng đến quyết định của khách hàng trong việc từ chối đơn đặt hàng hay không là "Amount ","Fulfilment" và "ship-service-level" vì chúng được tìm thấy có ảnh hưởng lớn và nằm trong năm thuộc tính phổ biến nhất, phổ biến ở cả hai kiểu máy.

Các phát hiện có thể là thông tin chuyên sâu có giá trị đối với chủ doanh nghiệp, người bán và các bên liên quan khác, đồng thời khuyến nghị nên tập trung vào các thuộc tính quan trọng này.

# CHƯƠNG 6. PHÂN CÔNG CÔNG VIỆC

|  |  |  |
| --- | --- | --- |
| Thành viên  Công việc | Nguyễn Đức Thuần | Chu Quyết Thắng |
| Chọn dữ liệu | 50% | 50% |
| Lược đồ hình sao | 30% | 70% |
| Viết các câu truy vấn | 50% | 50% |
| Đưa dữ liệu vào kho (SSIS) | 50% | 50% |
| Quá trình phân tích dữ liệu (SSAS) | 50% | 50% |
| Quá trình Report | 60% | 40% |
| Quá trình Datamining | 50% | 50% |
| Tổng kết báo cáo | 50% | 50% |

# CHƯƠNG 7. TÀI LIỆU THAM KHẢO

1. Slide bài giảng và hướng dẫn thực hành do giảng viên cung cấp
2. Tài liệu “Quá trình SSIS”, tác giả: Dương Quốc Cường
3. Tài liệu “Quá trình SSAS”, tác giả: Dương Quốc Cường
4. Tài liệu “Quá trình data mining”, tác giả: Yaipharen P M Potsangbam
5. SQL Server 2018 Tutorials: Analysis Services - Multidimentional ModelingFile